EPIC2018

CONFERENCE PROCEEDINGS · 10–12 OCTOBER 2018

How is evidence created, used, and abused?

EPIC2018 explored what it means to make claims, to demonstrate, and to create new knowledge in a world where it seems like evidence itself has come under fire. What kinds of evidence scratch a research itch? Create value for stakeholders? Make a connection with others? We explored the richness of research methods, tools, and approaches, and considered the innovations possible by combining low tech with high. We discussed ways of mobilizing evidence in humanistic exploration, in decision-making, and in everyday life. And we shared our strategies for creating value through evidence in an increasingly tricky social environment.

EPIC

Advancing the Value of Ethnography

EPIC is a dynamic, global community working to ensure that innovation, strategies, processes, and products are anchored in what matters to people in their everyday lives. We draw on tools and resources from the social sciences and humanities as well as design, business, computer & information sciences, and other fields to realize value from understanding people and their practices.

EPIC people come together year-round on epicpeople.org and at our annual conference to build expertise, share resources, and push each other to improve, experiment, and make change happen. Conference Papers go through a double-blind peer review process managed by an independent program committee, and Case Studies, PechaKucha, and Film submissions are peer-reviewed and selected by jury. Thirteen years of EPIC proceedings are freely available at epicpeople.org/intelligences, as well as published by Wiley Blackwell under ISSN 1559-8918.

Membership Matters

Your EPIC Membership supports the first professional organization committed to advancing the value of ethnography in business, research, and nonprofit settings. As a member, you can access our full program of learning and video resources. We are a volunteer-run, non-proft organization continuously developing new resources, events, and learning opportunities to support our international community. Join us!

Board of Directors

President
Maria Bezaitis
Intel

Secretary
Martha Cotton
Fjord, Chicago

Treasurer
ken anderson
Intel

epicpeople.org

EPIC2018 Sponsors

Platinum

Gold

Silver

Conference Organization

Program Co-Chair
Dawn Nafus
Intel

Program Co-Chair
Tye Rattenbury
Salesforce

Local Chair
Scott Robertson
University of Hawaii at Manoa

Papers Commitee

Hilary Dwyer, LogMeIn
Josh Kaplan
Helen Kolsto, IBM
Ilpo Koskinen, University of Twente
Ian Lowrie, Lewis & Clark College
Anne McClard
Rebekah Park, ReD Associates
Gemma Petrie, Mozilla
Tiffany Romaine, Turnitin
Bill Selman, Mozilla
Hiroshi Tamura, Re:Public Inc.
Josef Wieland, Gemic

Case Studies Committee

Josh Dresner, SPARCK Innovation
Bob Evans, Google
Ryoko Imai, Hitachi America
Gary Gebhard, HEC Montreal
Chad Maxwell, Kelly Scott Madison
Will Monge, Good Research
Simon Roberts, Stripe Partners

PechaKucha Committee

Evan Hanover, Conifer
Robin Kwong, Financial Times
Vyjayanthi Vadrevu, Rasa.nyc
Carrie Yury, Yuryka

Film Committee

Monica Frota, Pontifical Catholic University
Jay Hasbrouck, Filament Insight + Innovation
Sue Faulkner, Intel
Meg Kinney, Bad Babysitter
Hal Phillips, Bad Babysitter

Salons Committee

Julia Haines, Google
Tom Rowley, Stripe Partners

Gallery Committee

Makalé Faber Cullen, LORE, llc
Laura Forlano, IIT Institute of Design
Anja Maerz, Future Cities Catapult
Wafa Said Mosleh, University of Southern Denmark

Conference Support

Networking Coordinator: Matthew Bernius, Implosion Labs
Proceedings Coordinator: Greg Weinstein

Peer Reviewers

Yoko Akama
Elizabeth Allen
Shaheen Amirebrahimi
ken anderson
Ilkka Arminen
Allen Batteau
Sharon Bautista
William Beeman
Masha Bell
Sarah Brooks
Anna Chaney
Jared Cole
Maria Cury
Jennifer Davidson
Elizabeth DeLuca
Rita Denny
Carl DiSalvo
Christina Dunbar
Jan English-Lueck
Brittany Fiore-Gartland
Danielle Flam
Jeff Fleisher
Erica Gibson
Mike Griffin
Julia Haines
David Hallowell
Chris Hammond
Stefan Holmlid
Fumiko Ichikawa
Lilly Irani
Gemma John
Luke Johnson
Dalia Kahn
Sabrina Kang
Shipra Kayan
Christopher Keener
Kathi Kitner
Lisa Kleinman
Laksh Krishnamurthy
Esko Kurvinen
Taylor Leet Otley
Peter Levin
Ed Liebow
Yuebai Liu
Jonah Lowenfeld
Alexandra Mack
Alison Marlin
Mark Marrara
Chris Massot
Ben Matthews
Heather McGaw
Seth Messinger
David R Millen
Robert Morais
Michael Muller
Rob Murray
Dawn Nafus
Kendra Optovsky
Anssi Perakyla
Katie Pine
David Platzer
Frank Romagosa
David Rubeli
Shelley Sather
Nick Seaver
Jamie Sherman
John Sherry
Kate Sieck
Alvaro Soto
Michael Strickland
Jordan Suchow
Sakari Tamminen
Janice Tsai
Stefan Van Der Stockt
Suzanne Wertheim
Jack Whalen
Alan Wiig
Matthew Yapchaian
Joyce Yee
Alexandra Zafiroglu
David Zax
Jing Zhang

Student Volunteers

Jonathan Chinen
Liz DeLuca, New York University
Monica Frota, Pontífica Universidade Católica de Rio de Janeiro
Vivian Giang, University of Alberta
Philipp Jordan, University of Hawai'i Manoa
Suluem Lee Coulter
Ian Lowrie, Lewis & Clark College
Karen McPhail-Bell, University of Sydney
Caleb Newton, Columbia University
Margaret Pulver, University of Hawai'i Manoa
May Reid-Marr
Sara Said Mosleh, University of Southern Denmark
Prateeksha Singh, OCAD University
Ashveer Singh, Stanford
Dale Zigelsky, University of Arkansas

Contents

Papers 5: Methods at the Intersection of Ethnography and Big Data
Curator: Josef Wieland (Gemic)

Papers 6: The Organizational Context of Data
Curator: Dawn Nafus (Intel)

Papers 7: Ethnographic Theory and Practice
Curator: Josef Wieland (Gemic)

Papers 8: Brave New World: Evidence in the Age of AI and Machine Learning
Curator: Ian Lowrie (Lewis & Clark College)

Case Studies 1: Rehumanizing Data
Curator: Chad Maxwell (Kelly Scott Madison)

Case Studies 2: Desigining Hybrid Futures
Curator: Simon Roberts, Stripe Partners

Case Studies 3: Shifting Power & Agency
Curator: Will Monge (Good Research)

Case Studies 4: Balancing Evidence
Curator: Ryoko Imai (Hitachi America)

Case Studies 5: Possibilities and Limitations Moving Forward
Curator: Ryoko Imai (Hitachi America)

PechaKucha 1: Negotiating Positionality
Curator: Robin Kwong (Financial Times)

PechaKucha 2: Whose Story Is It, Anyway?
Curator: Carrie Yury (Yuryka)

PechaKucha 3: Systemic Evidence: Seen and Not Seen
Curator: Evan Hanover (Conifer)

PechaKucha 4: Embodied Perspectives
Curator: Vyjayanthi Vadrevu (Rasa.nyc)

Ethnographic Film
Curators: Jay Hasbrouck (Filament) and Monica Frota (Pontifical Catholic University of Rio de Janeiro)

International Business Ethnography: Are We Looking for Cultural Differences?

YUUKI HARA
Hitachi, Ltd. Research & Development Group

LYNN SHADE
Independent UX Research & Design Consultant

In international business ethnography, clients and subjects don't share the same background. Without an understanding of the underlying factors affecting the subject's behaviors, data can lead to false home-market based assumptions about cause and effect. Where do we as researchers look to detect meaningful findings in international contexts? Drawing on our decades of international fieldwork, we describe how focusing on culture or cultural differences to interpret differences in workflows and attitudes can sometimes hamper accurate interpretation of observations. We describe through case studies how instead, identifying foundation factors shaping behaviors and mindsets such as market forces, government policy, labour markets, and financial schemas can be the key to insight in international contexts.

Keywords: Ethnography, International, Japan, fieldwork, workflow, products and systems, user research, UX, cross-cultural, personas, culture, cultural differences

INTRODUCTION

As corporate ethnographers conducting international research, we study products and work systems designed in one country in another country. The objective of corporate ethnography is gathering, analyzing, and presenting actionable findings about target user workflows and behaviors to client project teams in order to inform product decisions. In such work, where subject and client are separated by context and country, the lack of shared background means the researcher needs to dig deeper for insights and work harder on explanations. Without an understanding of the context and rules making people's actions meaningful (Randall et al. 2007) data can easily lead to false assumptions.

In user research, research models for how to frame international fieldwork are limited. One common school of thought is that 'cultural differences' are the key to understanding local workflows and attitudes (Lee 2007). However, 'cultural differences' as the lens with which to interpret key aspects of local workflows and attitudes does not result in discovering the actionable data we and our clients need. We describe through case studies how a different focus: researching local foundation factors such as business structures, the labor market, market forces, and more, which shape differences in workflows and attitudes in international situations — yields more accurate insights and leads to actionable results. Communicating international research results to clients can be challenging and the paper concludes with aspects of communicating results including why UX personas do not work in international contexts.

A caveat: this paper focuses on researching local workflows and attitudes and their relevant underlying factors. It does not address other topics in international research and design such as language, localization issues, and definitions of culture beyond a working definition.

APPROACH TO INTERNATIONAL RESEARCH & DESIGN

Sources on how to frame international UX research & design frequently discuss the importance of incorporating cultural differences into the product or service (Ferreira 2016, Sun 2012, Nielsen & Del Galdo 1996). International UX commonly references definitions of cultural differences such as Hall's cultural iceberg model where behaviors and beliefs are visible while values & mindsets are hidden below the water line — "Just as nine-tenths of the iceberg is out of sight and below the water line, so is nine-tenths of culture out of conscious awareness" (Hall, 1976) and Hofstede's cultural dimensions theory (Hofstede, 2001) which defines how cultural dimensions such as individualism vs. collectivism affect behavior. However, this focus on 'culture' is not very useful to working researchers and designers looking for practical models on *how to understand* international user needs in order to create products that are useful, usable, and desirable. Furthermore, as we will describe, viewing international workflows, work systems, and attitudes through the lens of cultural differences can at times lead to undesirable outcomes.

Over-focus on cultural differences makes it easy to miss key drivers of behavior and attitudes and can cause errors of attribution

Focus on cultural differences as an explanation for observations can overshadow key underlying factors driving the behavior. For example, Japanese research participants are usually quick to notice visual imperfections and inconsistencies in interfaces. The attitude towards UI imperfection in Japan is, 'if I can see one problem, it indicates many more serious problems hiding underneath.' US research participants tend not to place the same weight on visible imperfections: "Guess you guys didn't get around to that." This could be interpreted as a Japanese cultural trait towards perfection and visual sensitivity. But focusing on this as a cultural difference obscures a key underlying factor shaping attitudes about product development: the difference in how companies ship products in the two countries. Japanese companies do not budget time or money for fixes or version improvements beyond shipping the product; thus products in Japan do not ship until they are complete. By contrast American companies plan ahead for version improvements and dot releases assuming improvements will be needed. The way Japanese companies ship products has trained the Japanese consumer to expect perfection in shipping products. In this case, it is domestic business strategy, not a 'cultural difference', that creates the expectation for visual perfection in the Japanese consumer.

For the client cultural differences are hard to understand and prioritize, and not easily actionable

The objective of corporate ethnology is to detail issues arising from work practices resulting in actionable data that can lead to concrete solutions. Although ethnography excels in describing complex systems to stakeholders, when it comes to international research, researchers tend to both focus on discovering and summing up findings under the umbrella of cultural differences (Graham, 2010, Anderson et al. 2017). This results in three problems for clients. One, the vagueness of the term 'cultural differences' can make it difficult to determine solutions. Two, clients have no way of assessing the relative importance of a 'cultural difference'. How important is this difference? How critical is it that the difference is addressed? Three, cultural differences without a deep understanding of what's underpinning the difference can sometimes be interpreted as malleable or changeable.

For example, research of Japanese document approval workflow reveals the importance of standardized stylized communication beyond please and thank you when passing on documents. These workplace communications could be seen as a cultural difference in politeness: i.e. Japanese people are polite, and that politeness carries into the workplace. But whether or not Japanese culture is polite is beside the point. What is critical is the role the behavior plays in the workplace, what it signifies (reliability and experience), and how it supports and is supported by the underlying corporate structure. Discussion can flounder on discussing a cultural difference. It can be difficult for the client, once grasping the idea of a cultural difference ("they value politeness") to understand that rather than politeness, which seems flexible, what the behavior signifies and the underlying business structure is hard-and-fast and must be accommodated in the product.

In international research, which often takes place in limited bursts, researchers can jump to cultural differences, which are more easily observable and accessible as explanations, to interpret differences in workflows and behaviors. As Mitchell says, in some approaches 'culture' is reified as an explanation and given causal force (Mitchell 1995). We think there is a better way to accurately identify root causes shaping the differences: looking at country and domain-specific foundation factors. What are 'foundation factors'? These are the near-impossible-to-change factors that shape the evolution of work, motivations, and attitudes; what we might speak of as the 'givens' of certain countries and domains. The following are examples of some of the foundation factors shaping workflows, mindsets, the division of labor, what people are rewarded for financially, and more:

- Geographic reality
- Market forces (global and local) such as demand fluctuation
- Government fiscal policy, tax structures, regulations, deregulation
- Labor Market: how fluid is it?
- Social structures: what people are rewarded for socially
- Financial schemas: how business contracts are structured, governance mechanisms such as public-private partnerships
- The design of work: the division of labor, what people are rewarded for financially. Technically, factors above have shaped the design of work, but for simplicity we include the design of work into foundation factors.

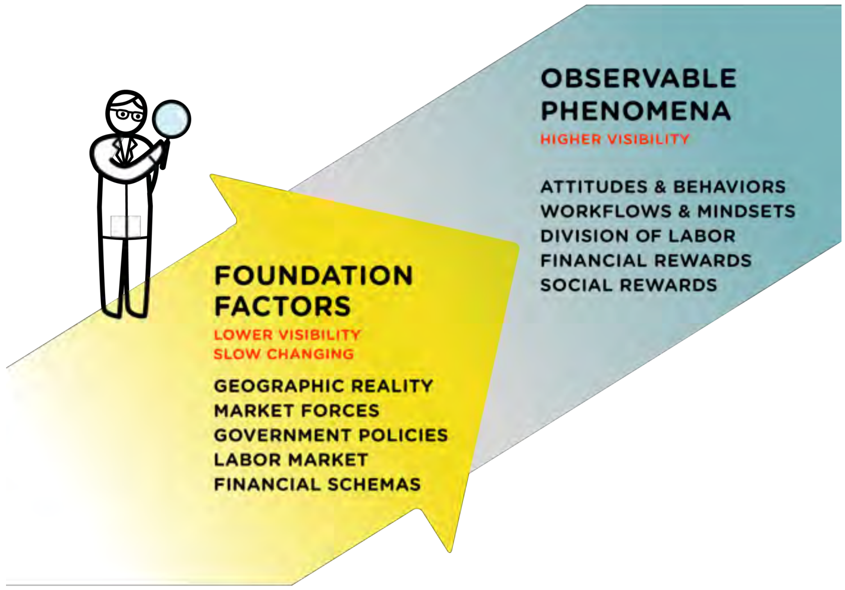


Figure 1. Foundation factors shape behaviors, mindsets and more

Foundation factors shape workflows and behaviors, mindsets and attitudes, and individual attributes such as motivations. Much of the behavior or mindsets that could be categorized as cultural is in reality the logical outcome of the foundation factors shaping the situation. Thus, researching the foundation factors underlying international situations can lead to valuable insights on workflows and mindsets that might otherwise remain opaque. For example, the fluidity of the labor market plays a large and unexpected role in shaping day-to-day knowledge-sharing behavior amongst train depot workers (Case Study 1, below).

Foundation factors are important to bringing into the open hidden assumptions based on home context

Client understanding naturally tends to be based on their home context. In order to communicate the facts of observed differences and have them understood, the international researcher must include descriptions of the shaping factors. Without an understanding of the local foundation factors, it is all too easy to develop false assumptions about cause and effect based on the home context. When stakeholders do not have the same baseline understanding of local factors, interpretation tends to be based on the home underlying factors. What could go unsaid or unexplained or unresearched in a domestic context, must be researched and explained in international research.

Without the explanations offered by foundation factors, some findings can be taken too lightly

In international fieldwork, findings on workflows and behaviors that aren't accompanied by explanations of the key foundation factors shaping the behavior run the risk of being interpreted as open to change. Naturally extrapolating from their own home-country experience, it's easy for clients to think attitudes can be changed; behaviors can be encouraged; workflows can be re-defined. With a clear understanding the underlying forces shaping the situation, the client can see what can be controlled, and what can't.

In the first case study below, the initial client assumption is with the right system improvements and some encouragement, European workers will embrace the computerized system. The researcher's identifying the foundation factors shaping the respective behaviors is key to convincing the client that the behaviors are based on identified intractable forces and the proposed solution will need to be re-thought.

Foundation factors are slow-moving

Particularly in international contexts, it is well worth looking into the foundation factors shaping observable behavior for insight into factors that will continue to affect product plans for years to come.

For example, much has been written about gift-giving practices in Japan (Rupp, 2003, Befu, 1986). But most studies haven't addressed that behavior around travel gift-giving in Japan is heavily influenced by, compared to the USA, a relatively static labor market (people tend to stay at the same company) and relatively static real estate market (people often remain in their locales). When traveling 'away from the group', acknowledging the support of the group by bringing back small gifts for neighbors, friends, and co-workers is important to reinforce these often long-term relationships. When the foundation factors change, the behavior changes, as can be seen in the comparative infrequency of gift-giving between Tokyo neighbors in high-turnover apartment buildings. When the foundation factors remain the same, the same forces will continue affect situations for decades or more. It can be helpful to identify and understand the foundation factors affecting domains in order to know what observable behaviors and attitudes are likely and unlikely to change in the near term.

CASE STUDIES IN IDENTIFYING FOUNDATION FACTORS

The following case studies illustrate how foundation factors plays a key role in illuminating attitudes and mindsets, and explaining how aspects of workflows and behaviors arise.

Case 1: How European train maintenance workers share acquired expertise

While conducting ethnographic research of train vehicle maintenance in a European depot, we observed that one of the key challenges maintainers face is getting the information needed to accomplish repair tasks. Maintenance windows are limited. If they fail to return the train on time the maintenance company must pay a penalty. Past reference cases for identifying faults are stored in a computerized system, and the original client request was for ethnographic research to identify issues with the computerized system. The thinking was if the system were improved, repairs could be accomplished faster.

To fix a fault, the maintainer needs to find and review similar, past cases to determine probable causes. In addition to a hands-on check of the train, the maintainer must identify and reference past cases to assess whether the fault can be fixed at the depot or needs to be escalated to the manufacturer.

Although information on past faults is stored in the computerized system, the maintainers were not relying on the system to obtain the repair work information. In field research we observed that maintainers would seek out fellow maintainers with experience of similar cases, asking the informal network to identify who had encountered a comparable case. Our key finding was this informal knowledge exchange, in reality, was the *primary* method of obtaining needed information, and the key to successfully completing repair tasks on time.

As stated, our client thought that improving the knowledge sharing system would close the knowledge gap amongst maintainers and speed up the repair tasks. However, our research showed this would not work. The maintainers didn't *want* to share their knowledge in a computerized system. Each maintainer believes their value in the railway maintenance labor market is determined by their troubleshooting expertise. In the European system, when an expert maintainer shares their know-how informally with colleagues, their reputation and status is enhanced. Simultaneously, expert maintainers believe that if they share their acquired know-how through a formalized system and their know-how becomes available to all, their market value might be decreased.

The foundation factor shaping this phenomenon is found in the fluidity of the European train maintenance labor market. The career cycle is: work in one company for several years; accumulate achievement & reputation. In order to achieve promotions and raises, maintainers must move to another company. By contrast the Japanese train maintenance labor market is based on lifetime employment. Sharing expert knowledge is regarded as an organizational asset and is a factor in promotions. Because Japanese train maintenance workers benefit, the workers are willing to share their acquired know-how through a formalized system.

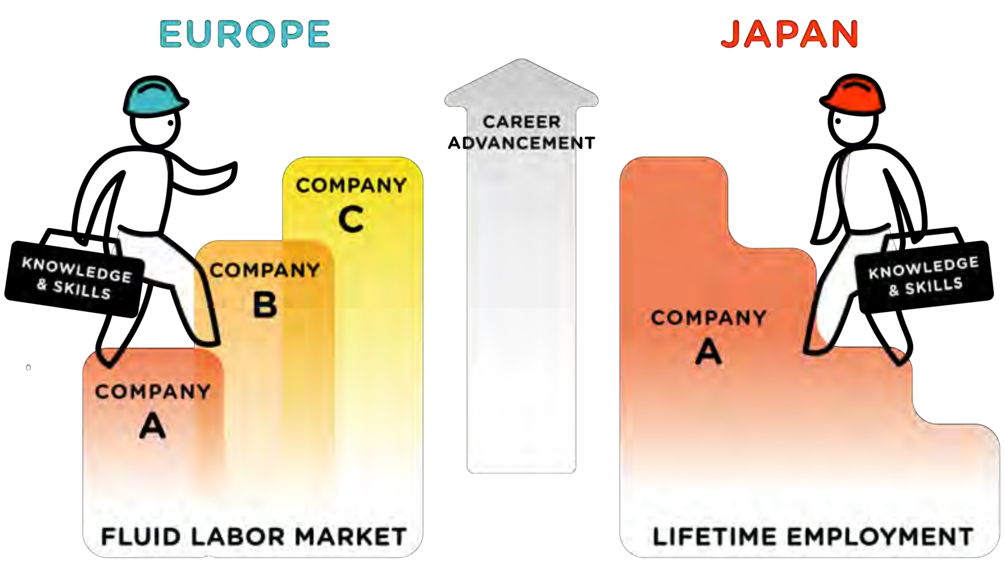


Figure 2. The fluidity of the labor market affects how workers behave to advance their careers

The realities of each country's labor market result in differences in what workers pride themselves on, leading to different attitudes and behaviors in where and how to share knowledge. In the European system the need for employment mobility to get promotions, and subsequent need to establish a reputation for personal expertise act as disincentives for maintainers to share acquired know-how in a formalized system. Describing the differences as cultural would not explain how heavily the behaviors and attitudes towards knowledge sharing practices between European and Japanese train maintenance workers are shaped by the underlying labor market. Lacking information on how the two different labor markets shape career paths and worker rewards, there is a risk that the client might invest in the wrong solution.

Case 2: Why Japanese and European train maintenance workers have evolved different skill sets

When conducting ethnographic research of train repair work, the differences in the focus of the work between Europe and Japan were notable. In the European rail sector, the priority was on repair work efficiency. In Japan, the focus was on planned maintenance. This difference in focus shapes how individual maintainers in each country develop their skills.

For example, European maintainers are experts at adjusting tasks when unexpected repair work appears. In order to effect the repair in time, they flexibly sub in a maintainer on the fly who's familiar with that particular repair, and adjust the order of ongoing tasks to accommodate the sudden repair work. These are all informal skills developed onsite.

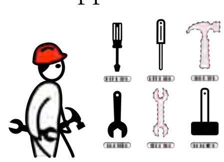


Figure 3. Optimizing tool placement

In Japan veteran train maintainers have expertise in accomplishing planned maintenance work as perfectly as possible through developed know-how such as where to place tools for optimum line efficiency. For Japanese maintainers, the emphasis is on developing skills for planned maintenance work rather than repairs.

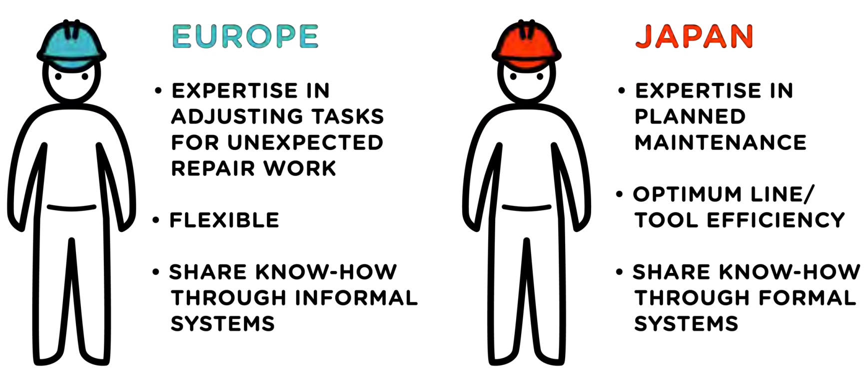


Figure 4. Foundation factors such as financial schemas affected by government policies shape skills workers develop

What is the cause of this difference? It's not simply, for example, that Japanese workers are detail-oriented and order-oriented; neither is it that European workers are flexible and cleverly adjust work tasks to fit the occasion. Behind these work practices is a difference in the financial schemas, leading to a difference in business strategy between Japanese and European rail sectors. Due to deregulation, separation of railway infrastructure and operations are common in European railway sectors. For example, in the European case the rolling stock ownership company, train operating company, and train maintenance company are different business entities. The train maintenance company contracts with the train operator. After maintenance or repair, the maintenance company is obligated to deliver the train at a specific time and place; they face a fine if they fail to meet obligations. Maintenance companies want to avoid the penalties. So rather than focusing on planned maintenance work, more effort is invested in unplanned repair work. Repair work is a primary factor in creating work delays, and it's in their financial interest to avoid the hefty penalties. Therefore, experienced European individual maintainers informally develop situation-specific adaptive know-how in order to complete repair tasks on time. Additionally, European train operating companies approach maximizing profit by prolonging the lifetime of train vehicles with repair work; with older rolling stock, European train maintenance companies face more repair work.

By contrast, in Japanese train sectors there is no separation of railway infrastructure and operations. Train ownership, operation, and maintenance are managed within a single company. Because it's all the same company there are obviously no penalty payments for late repairs. The Japanese approach to maximizing profit includes introducing more new trains in order to minimize repair work. Combined with the focus on planned maintenance, this business strategy is one of the key reasons behind Japanese train punctuality.

The two different business strategies, originating from different business and contractual obligations, directly affect the work practices and the skills maintainers develop. While one can say it is true in general that Japanese people tend to value order and many Europeans in contrast tend to be comfortable with making necessary adjustments on the fly, it would be misleading to chalk up the vastly different work practices to cultural differences. Investigating the foundation factors of the financial and business models — following the money — was the key to understanding this phenomenon.

Case 3: How foundation factors affect the design of Japanese magazines

Observing Japanese magazine staff working on new issues, one notable aspect is how much time and effort the editorial staff put into fresh layout designs for the feature articles. American magazines also change designs for feature articles, but to a far lesser degree. Why does the entire editorial room in Japanese magazines work so hard to create original layouts, breaking their own magazine's grid, issue after issue? It's true that there are many talented Japanese graphic designers. And the readership is responsive to design; perhaps it's an symbiotic pact between magazines and readers. And perhaps Japanese magazine staff are singularly committed to their art, working into the night long past the hour the trains stop running. The determining factor lies not in these cultural differences, but in the underlying business model.

For purposes of comparison, the prototypical American business model for general-interest magazines is based on advertising sales. Magazines with high circulation numbers can sell advertising for high figures. To achieve high circulation, magazines offer yearly subscriptions for an enormous discount — discounts of 80% or more from the newsstand price is typical. Readers are incentivized to renew subscriptions with increasingly inexpensive offers in order to keep circulation high. Additionally, American magazines sell their large circulation lists to other companies so having a great many subscribers, while making no money from subscription fees, yields two-fold benefit to the magazine publisher.

In Japan, the business model is based on direct sales from newsstand competition in addition to advertising income. Unlike in the US, subscription discounts at 5% to 9% are so negligible as to not present a significant incentive. Subscribers are rare. Advertising rates do not jump based on circulation numbers, but advertisers can be impressed with an issue that sold well. The business goal is to sell as many copies on the newsstand as possible. In order to achieve newsstand sales, magazine editorial staff put a great deal of effort into appealing, innovative layout designs. One senior magazine design manager described the goal as creating fresh, surprising designs in order to interest would-be buyers.

This "battle at the newsstand" business model has direct implications for desktop publishing (DTP) software design. A key selling point for DTP software in the USA is the ease of creating, saving, accessing, and automating layout templates. While large parts of Japanese magazines are templatized, improvements to the template feature is not a strong selling point in Japan per se. One president of a magazine company half-jokingly said in order not to hurt sales, he wanted to 'tear the template feature out' and have his designers design everything from scratch. In his view the further his designers got from paper and pencil the more the designs suffered. In the context of the Japanese magazine business model, features that support quicker design experiments and side-by-side comparisons between layout candidates would be gladly received in the Japanese market.

There are additional foundation factors at play. Geographic realities such as the sheer size of the USA make regular, convenient access to newsstand sales difficult for much of the population. The small size of Japan and mature distribution channels which excel at stock management allow for the vast majority of the population easy access to newsstands, bookstores, train station kiosks, and convenience stores.

With creative work in particular, it's tempting to look at design differences and draw conclusions about cultural differences. Some books and workshops on international UI design encourage UX professionals to do just that; a typical exercise is to examine the same company's in different countries and note the visual differences to infer insights into that country's cultural values. While visual differences can provide clues to areas to research, the ethnographer who is overly distracted by the cultural differences inferred from visual treatment differences risks missing the underlying structures shaping the design practices.

To illustrate the point, many Westerners perceive Japanese websites as "busy" in comparison to the current Western ideal of clean design (one current example is www.rakuten.co.jp). There are some cultural differences such as a preference for willingness to read more text, a historical enjoyment of the 'busy marketplace' idea, and a preference for democracy of information all one main page as opposed to hierarchically hidden. And there are other factors such as the fact that most web access is done from mobile devices not PCs. However, far more important and less immediately obvious are two underlying factors. One, the Japanese language is hard to use for text search so Japanese websites tending to be built for browsing. Two, Japanese websites visually reflect the structure of corporate Japan. Each department has a voice in the design of the website, which often translates to each department's areas represented on the main page of the website. The design of work in Japan — internal structure and internal approval processes — and the constraints of the written language directly affect what the consumer sees on the page.

FROM FINDINGS INTO ACTION: INTERNATIONAL RESEARCH AND DELIVERABLES

There are several issues researchers presenting international studies must think through. One is how deep an explanation needs to accompany each aspect of the findings. Often the client has limited understanding of the international market, and doesn't know what they don't know. In international research almost all aspects would benefit from illumination, but as all user researchers know, corporate ethnography operates under extremely tight deadlines.

Client attention spans can be limited and crisp summarizations of key facts and recommended actions are often well-received. The benefit of including a good deal of foundational explanations is that long after the technology or system being studied is gone, research findings that include explanations of foundation factors can remain valid. The understructure is slow-moving compared to directly observable phenomena.

In order to dispel hidden home-market based assumptions, the international researcher must make explicit what is implicit. As other have noted (Cuciurean-Zapan, 2014) there can be assumptions hidden even in target market segmentation. For example 'Professionals using presentation software' likely contains assumptions about work practices and worker motivations when the client is based in the USA. In the client's view, impressive presentation creation and delivery is important to professionals at work. This situation is connected to the very fluid labor market in the USA. This fluidity of the labor market means that professionals have the desire and opportunity to advance their personal career, partially achieved through honed skills and (regarding presentation software) looking polished in front of upper management. Where the labor market is less fluid, the meaning and purpose of presentation software will vary.

Additionally, market segmentation itself has varying validity from country to country, and clients are often unaware that their segments may not work overseas. The international researcher must determine if the right segment is being studied; one way to do this is to cast a wide net in exploratory research. A simple example from a case study in internet café research in China: "Initially, I approached him for help in recruiting small, medium and large-sized iCafes. But he stopped me short saying, "I have different way of categorizing the iCafe market. If I were you, I'd look for 'luxury', 'common' and 'dirty' iCafes." (Thomas & Lang, 2007)

A common situation for international UX researchers is being asked to construct personas. Personas are synthesized composite characters based on real research, combining key aspects of product or service usage into a single 'person' or persona. The personas are named and serve as archetypical shorthand representing different user types in the development process, so stakeholders with different backgrounds can come to consensus during project planning, and so different user needs are represented during development. For example, a persona for a child's car seat might be a single mother handling car-based daycare drop-off and pickup, and their concerns, motivations, and goals. While it's unclear how much personas are actually used, some clients and development teams consider personas a project requirement. However, personas don't work to summarize international findings and contribute to possible mis-direction. Why?

One, as mentioned above, segmentation is often different — and often unrecognized as different — for international markets. International personas are often expected to live neatly inside a framework based on domestic segmentation.

Two, more importantly, personas are by nature a type of shorthand. They rely on a shared understanding of the foundation factors, since goals and motivations are spelled out but not differences in underlying factors. As described above, mindsets, attitudes, and motivations are products of the foundation factors and can't be separated from the underlying context or they lose meaning. In relying on shared context as a shorthand, personas strip away the foundation factors that are essential to understanding the actual situation. Accordingly personas are not of help in generating solution ideas, or providing actionable insight, when the personas and the client stakeholders do not share foundation factors.

CONCLUSION

In international business ethnography, in lieu of focusing on cultural differences to explain behaviors and workflows, looking in-depth at foundation factors can be illuminating and contribute to the long-lastingingness of research results. This paper has described in case studies how researching and understanding the foundation factors can help avoid errors of attribution and false cause and effect, as well as clarify for researcher and client what observed workflows can be changed, and what can't.

Business ethnographers often have very little time to plan, complete, and report fieldwork. We need to be quick and accurate, delivering the kind of quality findings that lead the client to the right next step. In our experience, examining the foundation factors underlying local behavior and mindsets can be the fastest path to actionable insights and goals.

Yuuki Hara is a user researcher at Hitachi, a large multinational company based in Tokyo. Hara manages a team that conducts ethnographic studies of work places in various domains. Through capturing work practices in actual settings, Hara and her team uncover how hidden factors relate to the business goals.

Lynn Shade is a UX research & design consultant who grew up in Japan, worked as a designer at Apple and Adobe, and returned to Japan after 20 years in Silicon Valley. Shade specializes in creating design strategy through discovery research and design.

Illustrations by Akiko da Silva

NOTES

Special thanks to: Wilson Chan, Sheryl Ehrlich, Yoshinori Kikuchi, Priscilla Knoble, Akiko da Silva, Britton Watkins

The views in this paper do not represent the official position of our past or current employers

REFERENCES CITED

Anderson, Ken, Faulkner, Susan, Kleinman, Lisa, and Sherman, Jamie
2017 Creating a Creator's Market: How Ethnography Gave Intel a New Perspective on Digital Content Creators. Ethnographic Praxis in Industry Conference Proceedings Volume 2017, Issue 1

Befu, Harumi
1986 Gift Giving in a Modernizing Japan. Japanese Culture and Behavior: Selected Readings edited by Takie Sugiyama Lebra, William P. Lebra. University of Hawaii Press

Cuciurean-Zapan, Marta
2014 Consulting against Culture: A Politicized Approach to Segmentation, Ethnographic Praxis in Industry Conference Proceedings Volume 2014, Issue 1

Ferreira, Alberto
2016 Universal UX Design: Building Multicultural User Experience, Morgan Kaufmann

Graham, Connor, and Rouncefield, Mark
2010 Acknowledging Differences for Design: Tracing values and beliefs in photo use, Ethnographic Praxis in Industry Conference Proceedings Volume 2010, Issue 1

Hall, Edward T.,
1976 Beyond Culture. Anchor Books

Hofstede, Geert
2001 Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations (2nd ed). Thousand Oaks, CA: Sage.

Ikeya, N., Vinkhuyzen, E., Whalen, J., & Yamauchi, Y.
2007 Teaching organization ethnography. Proceedings of EPIC 2007: The Third Annual Ethnographic Praxis in Industry Conference, (pp. 270-282), Colorado, USA

Mitchell, Don
1995 There's No Such Thing as Culture: Towards a Reconceptualization of the Idea of Culture in Geography Transactions of the Institute of British Geographers, New Series, Vol. 20, No. 1. Wiley-Blackwell (1995)

Nielsen, Jakob, and del Galdo, Elisa
1996 International User Interfaces. Wiley 1996

Randall, Dave, Harper, Richard, and Rouncefield, Mark
2007 Fieldwork for Design: Theory and Practice. London: Springer.

Rupp, Katharine
2003 Gift-Giving in Japan: Cash, Connections, Cosmologies. Stanford University Press

Sun, Huatong
2012 Cross-Cultural Technology Design: Creating Culture- Sensitive Technology for Local Users. Human Technology Interaction Series, Oxford University Press

Thomas, Suzanne, and Lang, Xueming
2007 From Field to Office: The Politics of Corporate Ethnography. Proceedings of EPIC 2007: The Third Annual Ethnographic Praxis in Industry Conference, Colorado, USA

Doing Ethnography in AirSpace: The Promise and Danger of 'Frictionless' Global Research

TOM HOY
Stripe Partners

'AirSpace', according to Kyle Chayka, is the increasingly homogenized experience of the western(ized) business traveller, driven by major tech platforms (including Google, Airbnb and Uber.) As international travellers, ethnographers must account for the impact of AirSpace on their research practice. After delineating the concept of AirSpace the paper posits three dangers ethnographers must negotiate: (1) The cost of control: AirSpace offers researchers control, but can narrow the scope of research (2) The risk of superficiality: AirSpace provides shortcuts to cultural understanding, but can limit deeper comprehension (3) The assumption of equivalence: AirSpace provides shared reference points, but can create the illusion of equivelance with research subjects. By exploring these three dangers the paper invites readers to reflect on their own research practice and consider how to utilize the benefits of these platforms while mitigating the issues outlined.

WELCOME TO AIRSPACE

"[In AirSpace] changing places can be as painless as reloading a website. You might not even realize you're not where you started."
(Kyle Chayka)

It has never been easier to conduct international ethnographic studies. International travel feels more frictionless and familiar than ever. If so inclined it would be possible to land at an airport, hail an Uber, check in to your Airbnb, locate a boutique café on Google Maps, and not be certain if you are in Shanghai, Stockholm, Sydney or Sao Paulo.

Kyle Chayka coined the term 'AirSpace' to describe the increasingly homogenized experience of the international business traveller, driven by major tech platforms: *"the realm of coffee shops, bars, startup offices, and co-live / work spaces that share the same hallmarks everywhere you go"*. AirSpace enables privileged travellers to traverse the world seamlessly, experiencing local culture at a comfortable distance, Flat White in hand, with a taxi never more than 3 minutes away.

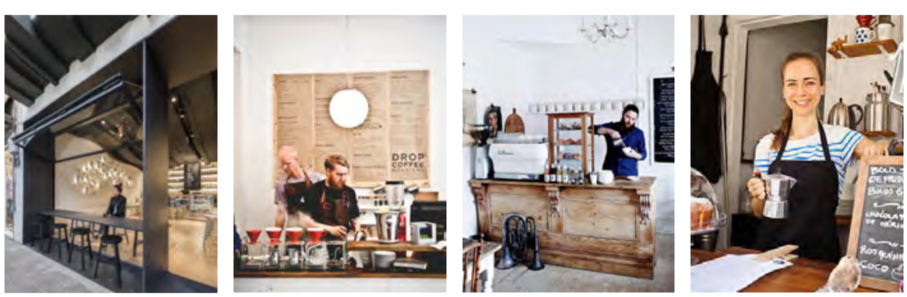


Figure 1. Independent coffee shops in Shanghai, Stockholm, Sydney and Sao Paulo (source: Pintrest)

As privileged travellers, AirSpace will be familiar to many corporate ethnographers. International platforms like Airbnb, Uber, Facebook and Google Maps have all become integral to the logistics and practice of modern fieldwork. From organizing accommodation to recruiting respondents, through to finding a quiet place to write up field notes, these services offer convenience, comfort and efficiency, as well as significant cost savings.

But ethnographers must be mindful. AirSpace represents more than a convenient mode of travel or aesthetic choice. It is part of a deeper technology-driven shift that reframes how we view and experience the world, with important implications for how we gather and martial evidence. From impacting how we frame our studies, subtly shaping our experience of the world, through to making assumptions about the people we are studying, AirSpace has important implications for ethnographic practice.

In this paper I will interrogate the impact of AirSpace on corporate ethnography, drawing on the experiences of practitioners across studies in three continents.

THE USER EXPERIENCE OF MONOPOLY

Before we discuss AirSpace's implications for research, we need to further clarify the term. What makes internet companies like Google, AirBnb, Facebook, Instagram, Uber and DiDi distinctive is not only their capacity to scale their services seamlessly for millions of users, but that the user experience generally improves as the number of users increases. In the words of Ben Thompson, these companies 'aggregate' users by maintaining a virtuous circle in which they "*become better services the more consumers/users they serve — and they are all capable of serving every consumer/user on earth.*" (Thompson, 2015)

And if just one or two companies dominate a specific market then, inevitably, the 'user experience' of that constituency becomes standardised. For many users searching becomes 'Googling'; Facebook *is* social networking; Airbnb becomes our default means for finding holiday accommodation. As we shall see, this is not to claim that these platforms monopolize all users. Or that users experience them in the same way. But it is to point out that the aggregation of a specific groups of users is having a significant impact on the way places and markets are framed, understood and served.

Chayka argues that the effect of these platforms is to privilege specific kinds of products and experiences, driven by the tastes of their most frequent and 'valuable' users: affluent and westernized. This is why many coffee shops – from large chains like Starbucks to independent businesses - look the same around the world. Or why AirBnbs often seem to have the same furniture. Suppliers look at what is succeeding on these platforms and ape the same style and proposition, creating convergence. As Chayka puts it:

> You can hop from cookie-cutter bar to office space to apartment building... You'll be guaranteed fast internet, strong coffee, and a comfortable chair from which to do your telecommuting. What you won't get is anything interesting or actually unique. (Chayka, 2016)

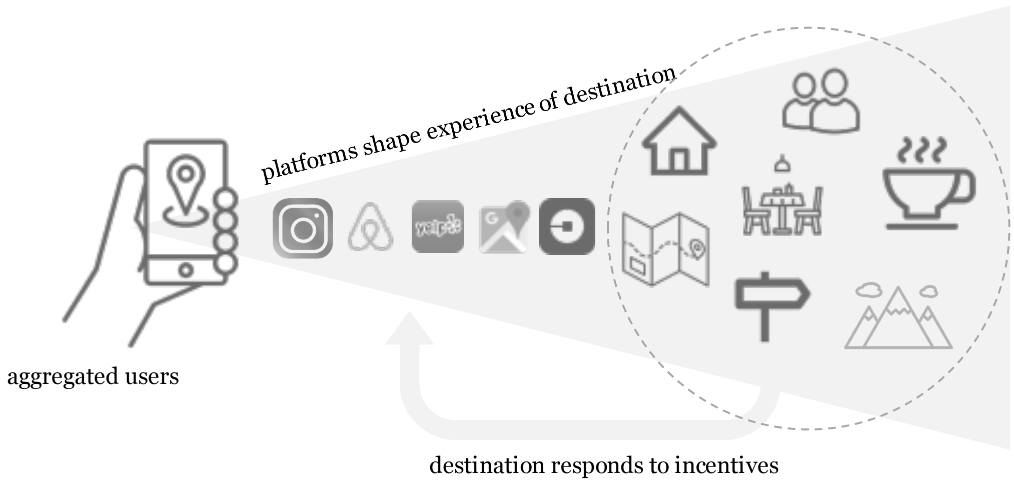


Figure 2.

Let's explore a tangible example. Imagine the same American tourist arriving in Delhi and Sao Paulo. On both occasions the tourist needs a haircut. She gets out her phone and searches for "Hair Salons"

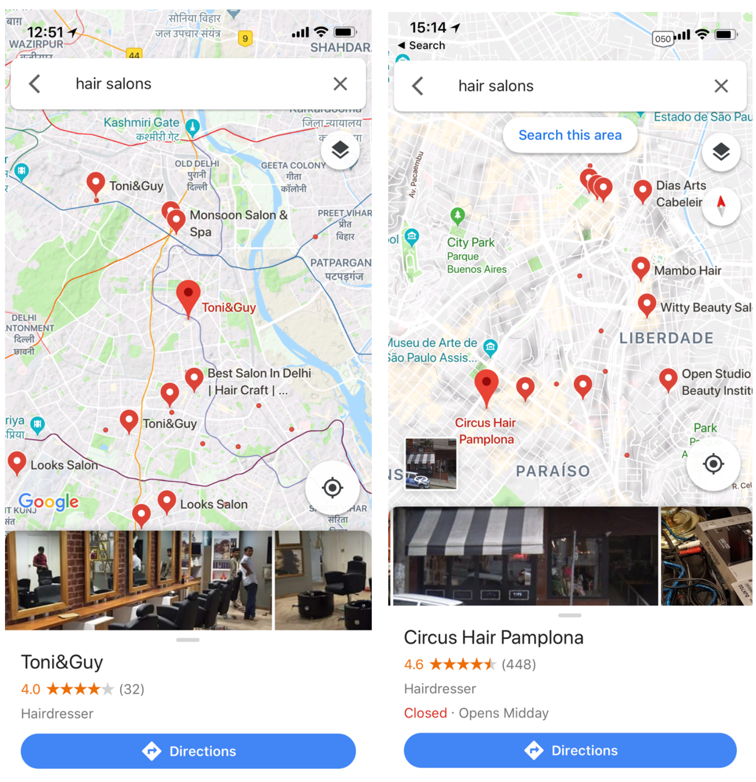


Figure 3. Discovering Hair Salons in Delhi and Sao Paulo through the same Google Maps interface

From a software perspective, the experience is almost identical. Google Maps presents itself as a neutral encyclopaedia that reflects all available options. The visuals and text of each salon are shaped by what the interface prescribes.

But the results are unlikely to be exhaustive. And the best reviewed salons are similar: they tend towards the expensive and westernized. They reflect the specificity of Google's users, and, by extension, the businesses and places that best address their specific needs. Whether the tourist is in Dehli or Sao Paulo, the experience – from the interface to the places she finds – share many commonalities. Of course, as a user of Google Maps this can be highly comforting. Wherever we are in the world we know not just *how* we will find a Hair Salon but also *what* kind of Hair Salon we will find.

AirSpace, then, is the user experience of monopoly. It describes what it feels like when platforms shape how their specific user base is served from discovery to consumption across geographies. This is not to claim that this experience is true for all consumers: most people around the world do not use these platforms frequently. And most places and businesses remain outside of AirSpace. But it is a common experience for many privileged travellers who use these platforms to navigate the world, including ethnographers.

Kyle Chayka and others have explored the aesthetic implications of such convergence. But what does it mean for how we understand the world as researchers? As frequent occupants of AirSpace, there are important implications. I want to highlight three dangers I believe it creates.

1. **The cost of control:** AirSpace de-risks research by making it more predictable (so you can deliver on time and on budget), but can remove opportunities for spontaneity and narrow the scope of research
2. **The risk of superficiality:** AirSpace can make the world more comprehensible, but applies pre-made categories and labels to culture, providing easy shortcuts that can derail deeper understanding
3. **The assumption of equivalence:** AirSpace can provide common reference points with the people we are studying, but encourages us to falsely equate our own experience with their own

But first I want to share a fieldwork story to provide some context for these claims.

ENTERING CHINESE AIRSPACE: A VIGNETTE

Following an 11-hour flight I felt disorientated. As we touched down I fired up the VPN I'd installed the day before, hoping to re-connect. Nothing. As feared, my phone would need a local Chinese SIM card. This made me anxious. Our flight was delayed and I wanted to let our AirBnb host know that we were running late.

With the help of my Mandarin speaking colleague we acquired a SIM at a kiosk at the terminal. I felt uncomfortable handing my iPhone to the clerk as she played with the settings. After several minutes the clerk handed the phone back. WhatsApp and Instagram began populating with updates: a warm feeling of familiarity cast over me as recognizable names appeared on the screen. Once again, I felt in control.

Figure 4. Author with translator and client, Shanghai, 2018

I opened the Airbnb app to contact our English-speaking host. We had chosen to stay in an 'authentic Shanghainese town house' in the French Concession with excellent reviews from other Western travellers (and a few Chinese!) This would be our research base for the week.

Our team was in Shanghai to study the automation of Chinese shopping practices. We had recommended Shanghai to our client because it was a sophisticated retail market, but also because of its accessibility - we knew we could be "in and out" within six days and stay on budget. In preparation for our trip we had used a combination of Google and local contacts to identify new examples of retail automation. (We were later to discover this provided us with a highly superficial understanding of what was going on. Many of the places we identified were either no longer operational or did not represent the cutting edge, forcing us to adapt our plans mid fieldwork.)

My colleague used DiDi (China's answer to Uber) to hail a taxi to the house. DiDi wasn't as quick as getting the train but it meant we wouldn't have to deal with buying a ticket or negotiating the subway. During the ride I played with Google Translate's augmented reality feature to try to figure out the text on the dashboard. "Business Bus Driver" it read. Makes sense, I thought, but is that how Shanghainese people understand it? The poor translation prompted more questions than answers, which, on reflection, I felt was a good thing.

We had an hour or so until our first interviews and we hadn't eaten lunch. I loaded up Apple Maps to see what I could find: a coffee shop was just around the corner. I was tempted to stop by a bustling 'local' restaurant we passed on the way. But a quick check showed there were no English reviews online, and we didn't know how long it would take to get served. Better to not take the risk and make ourselves vulnerable to the unknown.

My translator Rong met me at the coffee shop and we travelled via DiDi to our first interview in a northern suburb of Hongkou. I was glad we travelled together because she explained en route the cultural context of the neighbourhood we were visiting – it was known locally for its excellent schools and therefore attracted wealthy young parents. I started to see the generic apartment blocks we were passing in a different light. I noticed the number of prams being pushed around on the sidewalk. And the surprising number of luxury boutiques. For the first time since I arrived I felt a flicker of insight spread over me. As we stepped into the apartment block I couldn't wait to learn more.

THE COST OF CONTROL

Airbnb, Apple Maps, DiDi, WhatsApp, Instagram: the services highlighted by my experience in Shanghai will be familiar to many corporate ethnographers. The promise of these international platforms is alluring. Never has it been simpler to locate optimal research fieldsites on Google Maps, situate yourself in a 'local' Airbnb, or move seamlessly between interviews in an Uber or DiDi. These generic, familiar interfaces remove the friction of local complexity whilst cutting lead times and costs. In short, being in AirSpace means feeling in control.

The alternative to control is stress. A colleague recently conducted fieldwork in Germany on behalf of a large chemicals company. The study involved speaking to a very specific audience: welders in small manufacturing businesses in rural Germany. Her experience of the fieldwork was fraught because she had little control over its organization and execution.

The researcher was entirely reliant on her client and her client's customers to conduct the fieldwork. Each day she was driven to a field-site (which she had no prior knowledge of) and was then expected to convince unsuspecting workmen to speak to her about their welding practices. On some occasions the workers couldn't spare any time and she would leave empty handed.

"We had no idea who we were going to speak to… there was much more risk to it… no predictability"

What made the project particularly stressful was the time pressure the researcher was under. Budgets and deadlines dictated when the research had to be 'finished' (rather than the satisfactory obtainment of data).

This is nothing new. Corporate ethnographers have always been under pressure to speed the research process up. But it explains the context in which AirSpace platforms are being adopted. The convenience, predictability and control they offer provides new opportunities for making the research process more 'efficient'.

Take my personal experience in Shanghai. It was characterized by the removal of friction by platforms that felt safe and familiar to me. This reduced my stress levels and made me feel in control of the situation. But I chose to use them because they offered comfort and speed, not because I believed they would make me a better researcher.

As another researcher put it, *"when you're doing research you are acting so abnormally because of lack of time… these apps speed things up."* (Practitioner interview 1, 2018)

However, speed is not always conducive to quality research. Many ethnographers feel that negotiating local complexities and the unforeseen situations that they give rise to is often the source of the richest and most lateral insights. Despite the stress of the fieldwork the German welding project proved a success precisely because of its spontaneous, unplanned nature:

> "[The unplanned nature of the project] meant we came across complete gems… there was a worker we came across who turned out to be the star… it was lucky: he was waiting for something so we could speak to him in legitimate downtime… The pros [of messiness] outweigh the cons…as a researcher you need to be more ethnographic… you have to put yourself out there more… you're on the line."

The pressure of the project meant the researcher wanted to make things easier for herself, but this intention, she later recognized, was *"serving myself not the ultimate goal of the project"*. The consequence of this can be to undermine the quality of the research itself: *"you can try and make it frictionless but you can end up with very clinical projects"* (Practitioner interview 1, 2018)

This is not to say that using these platforms is necessarily damaging. Oftentimes making fieldwork more controllable is the right thing to do. For example, saving time on getting around town could mean more time for note writing and reflection at the AirBnb. But in a context in which timescales are being constantly squeezed AirSpace often has the opposite effect: reducing the space for proper ethnography by creating further 'efficiencies'.

In this sense Google, AirBnb, Uber and their ilk facilitate a world in which it is possible to reduce ethnography to no more than 3 hour in-home interviews (Sunderland and Denny, 2013) through the alleviation of the 'friction' that surrounds their organization and execution. In doing so AirSpace enables the narrowing of our frame of reference from people and places to 'users': *"the clean language for describing the messiness of people – deracinated from their contexts"* (Amirebrahimi, 2016).

Time outside of the interview is no longer considered 'research', but 'down time' in which the researcher can return to the comfort of AirSpace. One client researcher demonstrated this last year on another project in China. Rather than venture out of her Chengdu AirBnb to sample the local cuisine she ate from her stash of protein bars, imported with her from the US. This is not to say she would have necessarily learned anything valuable during the meal, but not participating in it certainly precluded the possibility. And AirSpace makes it easier to choose not to participate.

When ethnography is reduced to interviews its perceived value and utility in corporate settings is also undermined. Highly structured fieldwork is less suited to foundational questions that require abductive reasoning. In isolation interviews are good for addressing narrow challenges: rapidly testing and iterating existing hypothesis rather than challenging underlying assumptions. It is in this sense that AirSpace is facilitating what Simon Roberts has termed the 'Uxification of research' (Roberts, 2018) by creating the conditions in which ethnography can be bracketed to interviews.

But the danger of AirSpace extends further: it also informs how we experience the world as researchers.

THE RISTK OF SUPERFICIALITY

In 1995 Marc Augé made a distinction between places and non-places. Places are characterised by their rich cultural content: "complicities of language, local references, the unformulated rules of living know-how" (Auge, 1995). Non-places, in contrast, are characterized by temporary identities and generic labels: passengers in airports, customers in supermarkets, freelancers in coffee shops.

International technology platforms have commodified the most obscure, exotic places by making them superficially accessible. Search Airbnb for Mongolia and the first result is an invitation to stay in a 'Private Room in Yurt' with a 'Mongolian Nomadic Family'. While not as generic as 'customers' or 'passengers', these simple labels make Mongolia feel familiar and available. In Auge's terms, places that would have taken serious time and effort to experience, can now be consumed as non-places through these platforms. Place and non-place are becoming increasingly conflated.

Figure 5. Nomadic lifestyles are now a click away, Airbnb

The role of the ethnographer is to embody, surface and translate the 'living know-how' of place for their audience. But in a world of tight deadlines and shrinking budgets it's easy to cut corners and consume places superficially, utilizing the shortcut AirSpace platforms afford.

Comparing research studies in the US and China, a fellow UK-based researcher described how in the US fortuitous circumstances gave him the opportunity to get his 'bearings' and sense of place by simply becoming familiar with a local bar. In China, by contrast, a lack of time meant he could only experience place superficially.

> "On a recent trip to Cincinnati we happened to be there a few days early and we stumbled upon a bar which ended up being an important piece of the research… in China I never had the time to figure out what was happening around me… with more time I would have gained my bearings… sometimes it takes an hour, sometimes a week" (Practitioner interview 2, 2018)

When there is no time "to figure what is going on around me" we naturally look for shortcuts. AirSpace is an increasingly attractive substitute to a deeper understanding because it presents the world as instantly accessible and consumable. The danger is that as researchers we take these pre-existing labels as read and don't probe them further.

In a recent trip to Zurich a colleague and I were conducting a study on 'knowledge workers' for a technology firm. We wanted to understand their lifestyle outside of work so we decided to eat at a 'local' restaurant for our first evening in the field. Instinctively I pulled up Yelp to find a restaurant, figuring if it was popular on Yelp it would be popular with local people.

When we arrived at the top recommendation we noticed the signals Chayka describes as characteristic of AirSpace: raw wood tables, Edison bulbs, industrial furniture. And the rucksacks and accents dotted around other tables made it clear that were surrounded by other short-term visitors. We quickly realised we wouldn't learn much about the experience of local workers here.

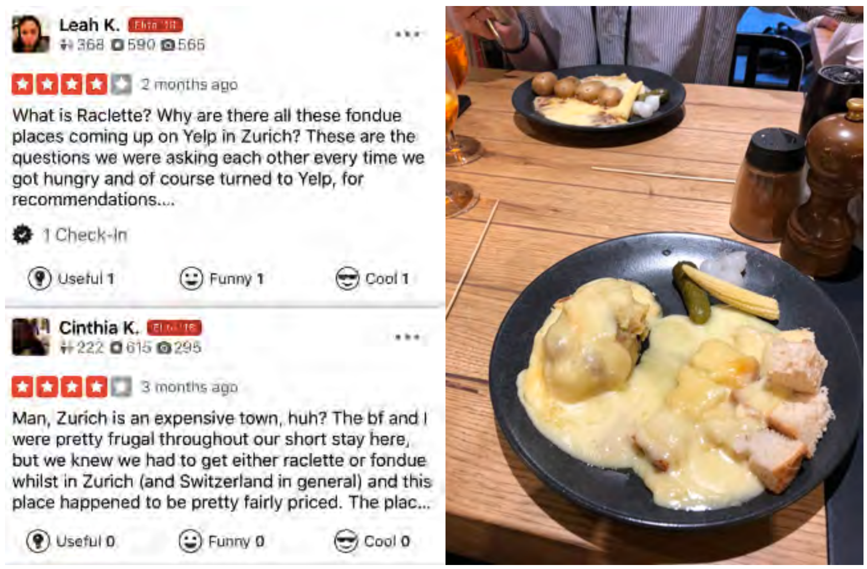


Figure 6. Discovering 'local' cuisine through Yelp

The mistake we'd made, of course, was to assume Yelp provided a 'local' perspective when in fact the people who use the app skew a specific way: American, white, middle class. Unlike Lonely Planet or other self-described tourist guides Yelp's stated purpose is neutral "to connect people with great local businesses." But what is considered 'great' reflects the cultural specificity of its user base, not necessarily a universal or even local perspective. Like the other coffee shops and bars we later encountered in Zurich, this restaurant had clearly been tailored to the tastes and needs of the western business people and tourists.

It is in 'breakdown' moments when experience does not meet expectation, that the edges of AirSpace appear and we can re-orientate ourselves as researchers. In the earlier Shanghai vignette, the fact that Google Translate provided an off-kilter translation of the local taxi service made me question how local people think about taxis: it made me more aware of how much I *didn't* know.

Figure 7. Google Translate reveals the 'edges' of AirSpace

These imperfect experiences make us better researchers by surfacing the limitations of our understanding; revealing the borders between our own worlds and the worlds we are seeking to understand. But how much longer until these imperfections are ironed out and we mistake AirSpace for the world we are seeking to understand? Because culture is being labelled and translated with growing fidelity.

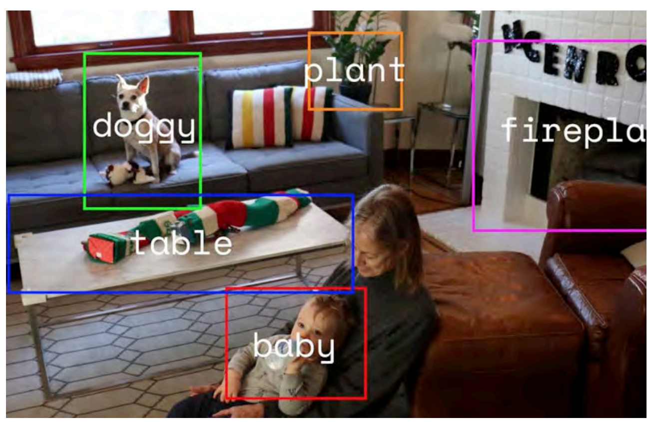


Figure 8. Google's visual recognition makes every object 'readable'

Google continues to 'organize the world's information' through visual and audio recognition technologies. At the most recent I/0 conference, CEO Sundar Pinchai demonstrated Google Assistant's new capacity to book appointments at local businesses. This technology will be used, in part, to harvest information that is currently 'off grid' because the local business has not shared that information with Google pro-actively (for example opening and closing times, what's on the menu etc).

Given this technological trajectory it is conceivable we will find ourselves in a situation in which every minor aspect of global culture – from buildings to market stalls to people in the street - are instantly recognizable, objectified and labelled by these platforms, our perceptions shaped by how they are presented to us.

As our cultural topography is mapped with growing fidelity there will be a temptation for researchers to consume culture superficially through these easy shortcuts to understanding without considering the tacit perspectives and agendas that frame them.

But interrogating these biases is easier said than done. As the fidelity and ubiquity of AirSpace intensifies, so its borders become increasingly difficult to disaggregate.

THE ASSUMPTION OF EQUIVALENCE

"It is difficult to see what is always there. Whoever discovered water, it was not a fish"
(Clifford Geertz)

Tom Boellstorff argues that we make a fundamental mistake when we frame the physical world as 'real' and the digital world as 'unreal' (Boellstorff, 2016). He suggests that both the digital and physical are equally capable of incorporating real and unreal phenomena. In his sense, losing money gambling online is as real as losing it offline. While daydreaming at work is unreal as daydreaming in the middle of a video game.

	PHYSICAL	DIGITAL
REAL	A physical and real	B digital and real
UNREAL	C physical and unreal	D digital and unreal

Figure 9. The Digital Reality Matrix (Boellstorff, 2015)

At the same time, this does not mean the digital and physical are two sides of the same coin. They remain distinctive phenomena capable of generating separate realities, and should not be assumed equivalent.

> For instance, if someone living in Chicago posts on Facebook, it is misleading to assert that this posting is located "in Chicago," even though that is where the poster's physical body is located... If 15 friends responded with comments, their activity would not be located "in Chicago"; these friends may not even live in Chicago. They could be posting while walking on a street with a mobile device or even while using a tablet on an airplane at 30,000 feet. In the sense of social action, these activities occur "on Facebook." The online sociality is real not in an exhaustive or privileged sense but in a perspectival sense." (Boellstorff, 2011)

What makes AirSpace distinctive is that it straddles the digital and physical worlds, shaping our experience of both. You are never clearly 'in AirSpace' in the sense you can be ‘on Facebook.’ Think about the earlier example of searching for hair salons in Shanghai and London using Google Maps. Both the digital aspects (discovering the salon) and physical aspects (the appearance of the salon itself) of the ‘user experience’ are influenced by the content and aesthetics that are rewarded or sanctioned by the platform. Compared to Boellstorff’s Facebook example (a relatively bounded digital world) the impact that Google Maps has in this case is at once more extensive and subtle.

This has serious implications for research. The unified experiences these international platforms offer tacitly shape our research choices: what places sound interesting; the easiest route to navigate; how to read a particular neighbourhood. But despite the tacit claims to universality the representation is always particular and usually reflective of a specific affluent user base. As Chayka explains

> AirSpace creates a division between those who belong in the slick, interchangeable places and those who don’t. The platforms that enable this geography are themselves biased: a Harvard Business School study showed that Airbnb hosts are less likely to accept guests with stereotypically African-American names. (Chayka, 2016)

Ethical concerns notwithstanding, this has serious implications for how we understand the world as researchers. EPIC contributor Tricia Wang recently encountered ads for the Google Maps Explore feature in Brooklyn, and reacted strongly against what she deemed the creeping cultural ‘colonialism’ of tech companies.

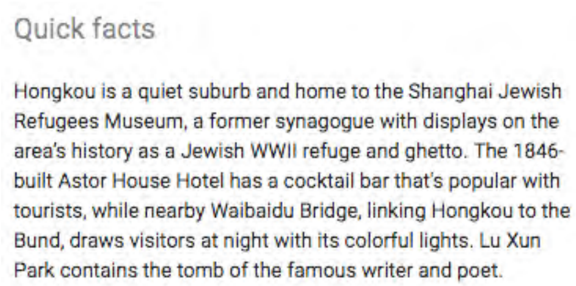
Figure 10. Google Maps labels (and defines) local cultures through its new ‘Explore’ feature, Brooklyn

Wang contrasts the perspective offered by Google Maps with the richer perspective of people who try to understand the world through “direct experience” by saying “hi to people”. She rejects the “colonizing” attempt of Google Maps to read her neighbourhood neutrally when, in fact, it promotes a situated perspective aimed at new residents and non-resident visitors / tourists.

As a resident of Bed-Stuy, Wang can delineate between these perspectives. But as a visitor seeking "direct experience" can make us feel vulnerable, and occupying AirSpace is certainly more convenient.

As the fidelity of AirSpace improves and the 'breakdown' moments that trigger self-awareness are ironed out, our capacity for reflexivity is diminished. And when we are no longer reflexive we assume equivalence: my reality is the same reality as the people I am studying.

Let's return to the Shanghai vignette at the beginning of the paper. When I was introduced to the district of Hongkou by a local translator she pointed out specific aspects to me (relevant to my objective of understanding local retail practices) – the prevalence of young affluent families, good schools and luxury children's boutiques. Alternatively, had I used Google Maps as an introduction to the neighbourhood I would have focused on entirely different aspects:


Quick facts

Hongkou is a quiet suburb and home to the Shanghai Jewish Refugees Museum, a former synagogue with displays on the area's history as a Jewish WWII refuge and ghetto. The 1846-built Astor House Hotel has a cocktail bar that's popular with tourists, while nearby Waibaidu Bridge, linking Hongkou to the Bund, draws visitors at night with its colorful lights. Lu Xun Park contains the tomb of the famous writer and poet.

Figure 11. Google Maps description of Hongkou, Shanghai

The point here is not to say that Google Maps is wrong to focus on these aspects of the neighbourhood. The content reflects the interests of the majority of its users. The issue is that, unlike self-described tourist guides, it presents the information as neutral, up-to-the-minute and totalizing. And as a researcher it is easy to start to make unguarded assumptions: that these features of Hongkou are important for local people too.

But the danger of equivalence extends further still. Not only can we project the reality of these platforms onto non-users, we can also wrongly assume that users themselves are equivalent. In a study of Facebook usage in Trinidad, Miller demonstrates *"where the potential for gossip and scandal (and generally being nosy) is taken as showing the intrinsic 'Trinidadianess' of Facebook."* (Horst and Miller, 2012) He argues that Trinidanian Facebook is a distinctive reality compared to Facebook in other cultures, even if the generic interface is precisely the same. Just because two people use the same interface doesn't mean they interpret or use it in the same way, even if usage seems equivalent at a surface level.

This is not to descend into a relativist vortex and claim that no two experiences can be equated, but rather to warn that our job as researchers is being made more complex by AirSpace. The assumption of equivalence occurs when we mistake the representations of AirSpace as shared by the people we are studying. And when we believe that our "reality" of technology usage is equivalent because, on the surface, it looks the same.

LOCATING THE 'WORLD AROUND HERE'

"…there is no such thing as pure human immediacy; interacting face-to-face is just as culturally inflected as digitally mediated communication, but, as Goffman (1959, 1975) pointed out again and again, we fail to see the framed nature of face-to-face interaction because these frames work so effectively"
(Horst and Miller, 2012)

Our response to AirSpace should not be to retreat from it and return to an 'unmediated' 'pre-digital' time. As Miller and Horst explain, such a time has never existed. And the sheer utility and convenience offered by these platforms ensures they will remain crucial to ethnographic practice in the future.

Which means as the experience of AirSpace becomes all-encompassing we need to work harder to reveal its borders and reflect on the contingencies of our own experience. As Geertz puts it *"…no one lives in the world in general. Everybody, even the exiled, the drifting, the diasporic, or the perpetually moving, lives in some confined and limited stretch of it – "the world around here."* (Geertz, 2004)

By focusing intensely on 'the world around here' we can mitigate the dangers posed by AirSpace.

1. We can question whether the feeling of control AirSpace offers is undercutting the quality and scope of our research. When does removing friction from fieldwork help the work and when does it hinder the work?
2. We can situate and interrogate the simple labels and categories that AirSpace provides to push beyond a superficial understanding.
3. And we can reflect on how AirSpace promotes a particular perspective parading as universal, and how our experience of technology is the same or different from the people we are seeking to understand.

The stakes are high. When we fail to account for AirSpace we not only fail as ethnographers, we can become agents of a broader totalizing phenomenon. These are international platforms that present themselves as neutral but in fact interpret and shape the world from a particular point of view. Our job is to reflect on and problematize their role in impacting our experience of the world, as well as the world itself, while leveraging them for their undoubted value.

REFERENCES

Amirebrahami, Shaheen
2016 The Rise of the User and the Fall of People: Ethnographic Cooptation and a New Language of Globalization. *Ethnographic Praxis in Industry Conference Proceedings*

Auge, Marc
1995 Non-Places: Introduction to an anthropology of Supermodernity. Verso

Boellstorff, Tom
2015 For Whom the Ontology Turns: Theorizing the Digital Real. Current Anthropology 57, no. 4 387-407

Boellstorff, Tom
2012 Rethinking Digital Anthropology. In H. Horst and D. Miller, *Digital Anthropology* Bloomsbury

Chayka, Kyle
2016 Welcome to AirSpace. *The Verge.* https://www.theverge.com/2016/8/3/12325104/airbnb-aesthetic-global-minimalism-startup-gentrification

Chayka, Kyle
2016 Same old, same old. How the hipster aesthetic is taking over the world. *The Guardian.* https://www.theguardian.com/commentisfree/2016/aug/06/hipster-aesthetic-taking-over-world

Geertz, Clifford
2004 "Afterword." In K. Basso and S. Feld, Place. Tourism and Applied Anthropologists: Linking Theory and Practice.

Horst, Heather and Daniel Miller
2012 The Digital and the Human: A Prospectus for Digital Anthropology. In H. Horst and D. Miller, *Digital Anthropology* Bloomsbury

Knox, Hannah and Walford, Antonia
2016 "Is There an Ontology to the Digital?." Theorizing the Contemporary, *Cultural Anthropology* website March 24, 2016. https://culanth.org/fieldsights/818-is-there-an-ontology-to-the-digital

Roberts, Simon
2018 The Uxification of Research. Stripe Partners Website. http://www.stripepartners.com/the-ux-ification-of-research/

Sunderland, Patricia and Rita Denny
2013 Opinions: Ethnographic Methods. *Journal of Business Anthropology* 2(2): 133-167

Thompson, Ben
2015 Aggregation Theory. Stratechery Website. https://stratechery.com/2015/aggregation-theory/

Can I Get a Witness? The Limits of Evidence in Healthcare Quality Evaluation Systems in American Hospitals

LINDSAY FERRIS
Ad Hoc, LLC

NICHOLE CARELOCK
Ad Hoc, LLC

"I got verbals, but verbals don't hold up in court….I need it in black and white."

After Sheila submits hospital quality data to the Center for Medicaid and Medicare Services (CMS), reports indicate that her data hasn't been received. She makes countless calls to the CMS Help Desk to get answers. They reassure her numerous times that they have her data, yet Sheila is insistent that she needs to see the change explicitly stated in the report. Sheila makes it her personal crusade to obtain material evidence because only written testimony will prove that her data has been submitted successfully and protect her facility from CMS penalties.

At a time when we are becoming increasingly reliant on data and technology as the ultimate bearers of truth, Sheila exemplifies how people become stewards of evidence in service to these technical systems. As she moves her facilities' data through CMS' error-ridden reporting system, the burden of proof is on her to provide the type of evidence acceptable to demonstrate her facilities' compliance with federal quality of care standards.

Throughout our paper, we explore the different practices that hospital employees and vendors take to demonstrate their facility's quality of care to CMS, identifying key elements of materiality, evidence and moral obligation. By weaving together their narratives of a responsibility to prove their truth to a capricious, data-driven system, with theoretical concepts of "bearing witness" and governmentality, we reveal the ways in which digital data falls short of being sufficient evidence and the dangers inherent in shifting blame from a body of government onto the body of an individual.

INTRODUCTION

There's a waste crisis happening Ala Ajagbusi, a small village in rural Nigeria. In the absence of a public health infrastructure in the area, women find themselves sifting, sorting and managing waste for their household and community. One participant scoffs at the line of people defecating under the villages' only power line,

"They are not even shy about it[...] But we don't want cholera here."(Abdulwakeel; Bartholdson, 2018). So she walks for over 3 miles with bags of detritus and human waste and burns the trash. Those who don't adhere to the cleaning and waste protocols are deemed *bad housewives.* How does it come to be that morality, government policies and personal action become conflated in these strange arenas?

Similarly, in the pacific northwest, Sheila works for a hospital system where she is the sole employee responsible for submitting data on quality of patient care for 11 hospital facilities into QualityNet, a tool maintained by the Center for Medicaid and Medicare Services (CMS). She feeds the data from the facilities' Electronic Health Records (EHR) into

QualityNet, a in which she cross-references emails and CSV reports to ensure that all of her files have been accepted by the CMS system and that there are no discrepancies between her files and the data in the CMS warehouse. She sifts and sorts through data and organizes it in a way so the delivery is acceptable to the system. She takes personal moral victory in navigating the complex multistep process.

These two women, on opposite sides of the world, are actually engaged in quite similar processes of Public health management. While the village woman takes pride in her visibly clean surroundings and long walks as proof of her compliance to complex protocols. Sheila creates and compounds similar visual evidence and proof of her facilities adherence to CMS strictures.

At a time when we are becoming increasingly reliant on data produced by technological systems for validation, Sheila exemplifies how "data and evidence" become tools by which people imbue themselves with moral and social authority over government policies and practices. Through this paper, we will examine these practices by Sheila and others, borrowing the concept of "bearing witness" from Anthropology of Religion to describe the knowledge production processes of hospital employees and vendors. We ultimately situate this witnessing within Michael Foucault's governmentality framework, with CMS and the quality data reporting system they've developed as the invisible hand. By examining these themes through the joint theoretical lens of "witnessing" and governmentality, our paper will serve two masters. First, we will bring a thoughtful cross-section of critical theory and Anthropology of Religion to the EPIC community. Our hope is that by demonstrating how these theories fit within American healthcare quality, we will empower ethnographers within the community to identify similar patterns and view the labor of the individuals they work with in a new light. Finally, we will also challenge assumptions around what constitutes "evidence" within data-driven technological frameworks. By acknowledging individuals' labor and materiality as playing a key role in proving compliance with quality of care, we reveal the ways in which digital data falls short of being sufficient evidence and the dangers inherent in shifting blame from capricious government processes onto the body.

METHODS

This article took a qualitative, multi-layered approach to data collection and analysis. The two researchers conducted a series of remote, semi-structured interviews with the two major populations involved in reporting quality of care data to CMS: hospital employees and data vendors hired by hospitals to submit data on their behalf. Beyond ensuring a diverse sample of both vendors and hospitals, the researchers strived to include participants from a variety of types of hospital facilities. With corporate healthcare systems (HCS) on the rise nationally, it was necessary to get the perspective of these corporate centers and employees from their associated facilities (Kaufman, Hall and Associates, 2016). However, the researchers also recognized the unique resource and staffing challenges of independent hospitals, particularly those of critical access facilities or rural, community hospitals. The researchers took a two-tiered approach to recruiting the diverse sample type needed for this study. Thanks to an existing relationship with CMS, the researchers sent emails out to a set of listservs that are used by CMS and the contractors that maintain QualityNet to communicate with hospitals and vendors. Those that were interested in participating filled out a survey aimed to aid the researchers in recruiting a diverse sample, filtering potential participants based on their role, type of facility, location and technical ability. In complying with ethics practices, all participants signed consent forms indicating their willing participation, and agreeing to allow their data to be used freely by the research team. Ultimately, this article brings together findings from 15 participants from 10 different organizations across the U.S., including four data vendors, four hospital systems and two small, independent hospitals.

The researchers began the interviews with broader questions about participants' role within their facilities and involvement with quality reporting programs. Then the researcher dug in deeper, probing participants on their practices for handling quality data and recent experiences submitting data to CMS to understand the full lifecycle of participants' interactions with quality care data, from the collection of this data to their final confirmation that their facilities have successfully completed the program requirements. By asking about participant experiences, they collected "thick descriptions" (Geertz, 1937) about their experiences and begin to gather emerging themes amongst participants. The researchers then augmented these traditional interview techniques with digital ethnographic methods that focus on observing themes in human interactions with internet-based technologies (Hine, 2000; Hsu, 2014). Researchers observed participants moving through QualityNet, the portal for complying with quality requirements and communicating with CMS, probing them on their experiences within and adjacent to the system. Finally, the researchers walked through a series of artifacts vital to demonstrating compliance with the CMS reporting program, such as reports and email communications with the CMS Help Desk.

Researchers captured and transcribed interview recordings and video of participants moving through QualityNet during the data collection process. Through analysis of these interviews and videos, researchers began to explore central themes of how these experiences shape hospital employees perceptions of their responsibilities and of themselves. The researchers took a grounded theory approach, splitting up the interviews and conducted the first coding cycle independently using a combination of in-vivo and thematic coding central to this approach (Given, 2008; Saldaña, 2013). The researchers shared the results of their initial round of coding to discuss their findings and check themes before exchanging interviews to then verify their counterparts' codes. This code cross-checking approach allowed the researchers to triangulate between sources and analysts to build trustworthiness in the qualitative data research and analysis process (Patton, 1999; Suter, 2012).

FINDINGS

Seeing is Believing

One of the methodologies hospitals workers and vendors utilize to signify compliance with quality care measures are through complicated and varied *witnessing* techniques. We define witnessing to include the hospital workers and vendors' accounts of personal responsibility, the imperatives around seeing data with one's *own eyes,* and the imperative of holding documentation in one's own hands as demonstrations how these actors "bear witness" to their facilities' quality of care. We borrow witnessing from its origins in Anthropology of Religion to highlight our participants moral relationship to the data and the materiality of evidence quality workers user to avoid penalties in the system.

As our participants described reporting data to CMS, reports produced by the QualityNet system were cited as key visual evidence throughout the data submission process. These reports indicate whether their large zip files have been accepted and received by the CMS data warehouse without errors. Several of our participants described these reports as a "receipt," "confirmation" or "proof" that they've successfully met program requirements and are done with the process. During the 2017 submission period for Electronic Clinical Quality Measures (eCQMs), one of the CMS reports was not updated to reflect new program changes. While some reports indicated that data had been received, the Submission Status Report — the primary source of "proof" according to our participants — indicated that facilities hadn't met the reporting requirements. One section of the report bluntly states: *"Completed 2017 eCQM Reporting Requirements: No."* Upon reading this, hospital employees spring to action. Sheila describes the labor she endures as a result to get the confirmation she needs:

> *"What are you talking about no? Did you get them or not? I have reports that say you got them, but you're saying we haven't satisfied the requirements...So here's where a lot of phone calls had to be made to [the CMS Help Desk] and they were confirming for me multiple times, because I would wait a week and it wouldn't be showing and that's a lot of money to be risking. But they kept telling me, yes, it was fine."*

For hospital workers and data vendors working with hospital clients, reports serve as confirmations that their data has been accepted and they've successfully met program requirements. Notice here also that Sheila's need for particular documentation after her verbal confirmations speak to clear delineations on the levels of input that could "make the cut" and count as evidence. Bill, a submitter at a data vendor, describes a similar experience with this report error.

> *"None of our submissions are actually complete yet because that report is outstanding. It's extremely frustrating because it's a report with one line. And we will be held to that as the submission vendor for our client because we don't have the final report telling them that they're done."*

The confirmation that hospitals have successfully completed requirements hinges solely on the materiality of reports and the *visual* confirmation that they provide. Verbal confirmation in itself it not enough. Despite ongoing reassurance that the mistake is within QualityNet and her facilities have met program requirements, Sheila is insistent. She won't believe it until she sees it explicitly stated in the report. She describes these interactions as "verbals," but, according to her: "*verbals don't hold up in court….I need it in black and white."* By negating "verbals" Sheila illustrates the ways in which hospital employees and data vendors must bear witness to their facilities' quality of care in very particular ways. It is only through seeing this evidence with their own eyes that they feel reassured that their responsibilities are completed. Below is a sample Submission Status Report with visual cues vendors and hospitals look for to signal completion.

R530 Submission Status Report

Page 1 of 3

Report Run Date: 10/09/2017

EHR Hospital Reporting – eCQM Submission Status Report

Submitter: [redacted]

Provider: All

Discharge Quarter: Q1 2017

Data As Of[1]: 08/16/2017
Submitter: [redacted]
Provider: [redacted]
Discharge Quarter: Jan 01 - Mar 31, 2017

Discharge Quarter eCQM Count:
EHR Incentive Program : 8
IQR-EHR: 8

Program Year Successful eCQM Data Submission:
EHR Incentive Program : No
IQR-EHR: No

Measure ID	Domain	Submission Status[2]	Last Submission Date/Time
AMI-8a	Clinical Process/Effectiveness	Zero Denominator Declaration	08/15/2017 12:35
CAC-3	Patient and Family Engagement	Not Submitted	N/A
ED-1	Patient and Family Engagement	Not Submitted	N/A
ED-2	Patient and Family Engagement	Not Submitted	N/A
ED-3*	Care Coordination	Not Submitted	N/A
EHDI-1a	Clinical Process/Effectiveness	Not Submitted	N/A
PC-01	Clinical Process/Effectiveness	Not Submitted	N/A
PC-05	Clinical Process/Effectiveness	Not Submitted	N/A
STK-2	Clinical Process/Effectiveness	Zero Denominator Declaration	08/15/2017 12:35
STK-3	Clinical Process/Effectiveness	Zero Denominator Declaration	08/15/2017 12:35
STK-5	Clinical Process/Effectiveness	Zero Denominator Declaration	08/15/2017 12:35
STK-6	Clinical Process/Effectiveness	Zero Denominator Declaration	08/15/2017 12:35
STK-8	Patient and Family Engagement	Zero Denominator Declaration	08/15/2017 12:35
STK-10	Care Coordination	Zero Denominator Declaration	08/15/2017 12:35
VTE-1	Patient Safety	Not Submitted	N/A
VTE-2	Patient Safety	Zero Denominator Declaration	08/15/2017 12:35

1 The Data As Of Date is the date of the last time data was updated for the Provider.
2 A status of "Not Submitted" indicates there is no data on file for this measure. A status of "Submitted" indicates the Provider has records that were successfully accepted. A status of "Zero Denominator Declaration" indicates that the Provider's Zero Denominator Declaration has been successfully completed. A status of "Case Threshold Exemption Declaration" indicates that the Provider's Case Threshold Exemption has been successfully completed.
* Indicates eCQM is not applicable for the Hospital IQR Program.
**To meet eCQM requirements for the IQR-EHR and EHR Incentive programs, the same 8 eCQMs must be submitted for at least two calendar year quarters. All quality reporting program requirements, in addition to eCQM requirements, must be met to ensure payment.

Figure 1. Sample eCQM Submission Status Report

By situating the work of Hospital Quality workers within discussions of witnessing, our goal is to also hint at the deeply moral imperatives placed on workers in the creation of "Quality" Data. The book of Acts in the Christian Bible is all about a particular type of ethical work. Before Jesus ascended into heaven, his last words were, "You shall be witnesses to me in Jerusalem, and in all Judea and Samaria, and to the end of the earth" (Acts 1:8, NKJV). The followers of Jesus testified, "we cannot but speak the things which we have seen and heard" (Acts 4:20, NKJV). Rooted in the Christian notions of testimony, and of the body, sight and experience as vehicle of knowledge production, witnessing is a deeply persuasive Western cultural form. Witnessing has long played an important part in rights advocacy. Its use grew in the 1990s, when testimonies proliferated in multiple genres and arenas in human rights advocacy. Organizations like Médecins Sans Frontières have created precise and specific methodologies around witnessing that grew out of moral obligations to testify to human rights violations. Redfield (2006) calls these instances of witnessing a kind of "motivated truth" toward socio-political ends. The significance of morally motivated witnessing becomes more apparent when the frame of reference shifts from advocacy causes such as the advancement of human rights to fully operational bureaucratic endeavors, like Hospital Quality Reporting. Here, the "motivated truth" focuses around the materiality of the evidence involved, and the cognitive load of excessive if not pointless process management. These aspects of reframing combine to create a different context for action, one constructed around bodies and paper trails as much as words.

> *"I have learned to keep track of everything on a system folder so should I get hit by a bus or win the lottery and leave, they can follow the crumbs that I left behind me."*

Sheila moves the data through CMS's system, while perennially working to prove compliance with quality of care standards. However, the system itself is fraught with technical errors, so she creates a path of documentation that demonstrates every one of her actions within QualityNet. She makes sure that every email she's received and report she's run and errors she's encountered throughout the submission process are available for anyone to see. Her hope is that in the event that a facility is audited, anyone can take her place and produce their own the documentation needed to prove compliance with the law. Sheila's system folder demonstrates the ways in which hospital employees and vendors are instrumental in creating evidence. "Evidence" of quality of patient care is not simply reflected through the data alone, but is also demonstrated through the documentation process she manages and its potential to be followed by others. By introducing Sheila's folder of documents, we begin to detect the types of knowledge production that hospital employees and data vendors engage in to create evidence of their facilities' quality of care and how this type of knowledge production ultimately culminates in their roles as "witnesses" on behalf of their facilities.

Some participants, such as Judy, a data vendor who submits for hospitals around her state, saves a physical copy of every single one of these reports "*historically, just for defense.*" Through her system of archiving, Judy explains that she is creating a "paper trail" to slowly make the case that her clients have met requirements, with the literal "paper" within this trail emphasizing the creation of materiality being necessary evidence of compliance. Sheila also describes her self-named "Sheila system" of tracking CMS documents as part of her need to leave a trail. She explains: "*I keep proof of each step that I did and that each step is satisfied, with the thought in mind that if someone had to reinvent what I did, can they follow that trail?*" In both these cases, participants emphasize the need to elevate reports through archiving and documentation into a certain type of evidence.

As Elizabeth, a nurse and quality improvement specialist at an independent hospital in the midwest, tracks which of her files have been rejected by the CMS warehouse, she explains: "*I actually print it off and attach it to that other report that I was talking about where it shows I have a rejected case. So that should I ever be asked, I can say: look I ran this report that says I have a rejected case. Here is my Submission Detail Report and here I have circled which case it was.*" Through adding her own touch to these reports, she creates evidence that is again, only meaningful based on her relation to it. However, printing the reports in particular represents their transformation into materiality. She even takes a step back after going through the papers. "*Wow, I didn't know my file was that big,*" she exclaims. She creates a type of evidence that could not only hold up in a court of law, but is also very tangible and physical. Participants across the sample integrated the necessity of materiality into their own knowledge production systems, describing printing out reports, emails and combining them with their own annotations, spreadsheets and physical individual actions. Judy tracks the communication from QualityNet on the status of her files through these extremely manual means:

> *"If there's one that doesn't [get processed], I have the paperwork on my desk that I know I need to follow up on a batch that we sent. And then if I don't get an email back that says it's processed, that paperwork stays on my desk until I get an email that says I can go ahead and look at the reports. It's a manual process on my end that says I didn't get any feedback on this so it hasn't processed yet."*

Like Elizabeth, Judy is taking *digital* information on her files from QualityNet and prints them and physically moves them from one place to another so she knows where to follow up. It is only through this system that she actually knows where her files are in the process. In both these instances, hospital employees and vendors describe moving information that has meaning within QualityNet outside the digital system. They are bringing it into the physical world and attaching personalized practices to this information that makes it relevant and meaningful to them. In this way, they are taking the base "data" and turning it into the material knowledge that can be written about, that can be pointed at, and, most importantly, that can start to exist as an external thing.

However, the relationship between *evidence* and *proof* is a bit more complex. The data from QualityNet isn't truly "information" or "knowledge" until these personalized practices and materiality is applied. While we heard a litany of terms from our participants to describe the information they archive from CMS, participants frequently used terms such as "evidence," "proof" and "defense" to describe their archiving process. Ann, a quality director for a hospital system with seven facilities in the midwest keeps a physical "book of evidence" with all the reports. According to Ann, *"the value of the report is really the protection of the facility, and CMS is known to be unforgiving unless you have documents to prove that the error is on their side and not on the facility."* She keeps this book of evidence *"so if we're ever in a payment year where we're receiving a reduction, we have this to show that the error is on CMS' end."* Similarly, Judy explains that: *"If we had a situation where a hospital was denied payment and said it was because we didn't submit their data. We have the paper trail."*

In both of these instances, hospitals and data vendors are imploring their employers and client to witness what they've seen, to look at the evidence themselves to see with their own eyes. It is through the labor of hospital workers and vendors, as they run, track, save and, finally, show rather than tell, that these reports become evidence. These reports only have power in their relation to the participants' interactions with them. While the need to see and show certain CMS documentation as crucial to their role demonstrates participants "bearing witness" to their facilities' quality of care, the power that these reports have in relation to hospital employees and data vendors begins to demonstrate how these actors are instrumental in creating this evidence. While the methods through which our participants witness their facilities' quality of care may seem varied, personalized and even tied to a sense of moral obligation, every aspect, from data collection to final submission is in service of an inscrutable system of penalties dictated by CMS requirements. The introduction of penalties and their conflation with morality and governmental policy, houses most of the actions we've discussed in a framework of governmentality.

Believing is Doing

Foucault's original essay on governmentality emerged from a lecture series that he presented at the College de France in the 1970s, which was concerned with tracing the historical shift in ways of thinking about and exercising power in certain societies (Elden 2007). Here, Foucault highlights the emergence of a new rationality of rule in early-modern Europe. Crucially, he introduces the term "biopolitics" to draw attention to a mode of power, which operates through the administration of life itself – meaning bodies both individually and collectively (Foucault 2003: 202). In doing so, Foucault illuminates an 'art of governing' that involves modes of practices and precise strategies that deputize citizens in service and execution of government policies . In addition, he articulates a mode of political government more concerned with the management of the population than the management of a territory per se (Jessop 2007). Consider this street sign (Figure 2) for a public place in Savannah, Georgia and this typed sign (Figure 3) that recently made national news in a Dunkin Donuts.

Figure 2

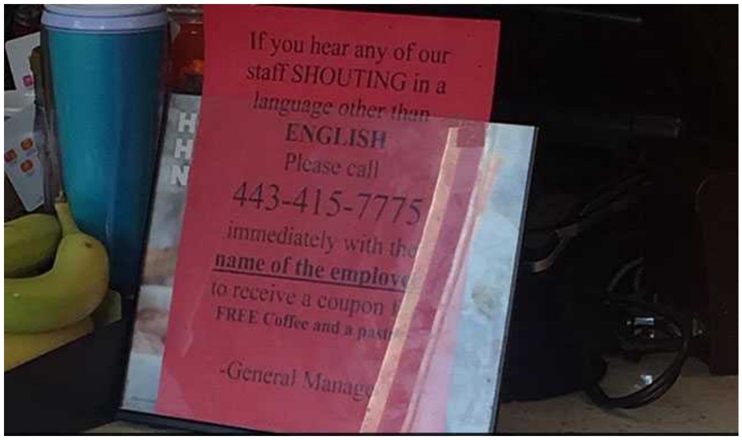


Figure 3

In both images, we see citizens being asked to enforce local policy and intervene on the lives of others or themselves for the benefit of the state. We posit that hospital quality workers have internalized formally external processes of evidence gathering for the sake of quality measure alignment. We examine how CMS structures and shapes the field of possible action of hospital personnel and vendors by giving meaning to these disparate evidence producing techniques.

In a public video recapping the events of the Quality Workers Conference, CMS suggests that quality workers be *relentless* in their commitment to quality reporting.

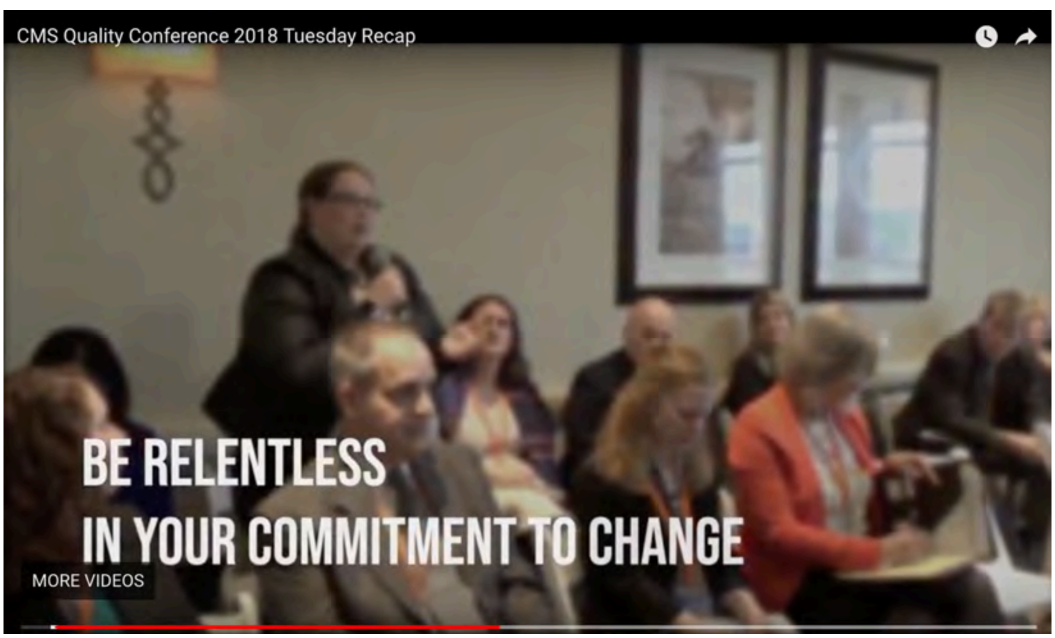


Figure 4. CMS Quality Conference 2018 Recap Video

We see here how policy needs get slowly transformed into moral and ethical imperatives. The anthropological field on ethics—conceived as explicit codes of conduct—is well rehearsed from Ruth Benedict's *Patterns of Culture* (1934), Richard Brandt's *Hopi Ethics* (1954), and Gregory Bateson's *Naven* (1958) to Clifford Geertz's *Religion of Java* (1960). Anthropologists have written about the strong moral codes that regulate Russian understandings of social networks and public assistance (Caldwell, 2004: 86) or the moral worth acquired by Nepalese women who live within the gendered restrictions of their villages (McHugh, 2004: 590).

According to Foucault, ethics can be understood as the actions or practices on the self with the aim of making, developing, or transforming oneself to reach a particular state of being (*The Ethics of the Concern of the Self,* 291). In other words, ethics involves the relationship of the self with the self and the activities that create and develop identities (Foucault, *On Genealogy* 263). Understood in this manner, ethics is not only a certain set of rules but rather consists of practices of self-transformation which may or may not be in relation to universal moral codes. Foucault describes these practices as technologies of the self, the activities which individuals undertake on their own bodies and souls, thoughts, conduct, and way of being, so as to transform themselves in order to attain a certain state of happiness, purity, wisdom, perfection, or immortality (*Technologies of the Self, 225*). From Foucault's approach, then, one becomes a moral person not by following universal rules or norms, but by training oneself in a set of certain practices (Widlock, 2004: 59). We place the quality worker's "witnessing" work in service to this Foucauldian ethical work that enables governmentality.

Beyond creating materiality, these tracking practices mentioned above also begin to sow the seeds of an incredibly moral personal ownership of CMS processes. Another participant describes her tracking mechanism as a "Sheila system." She manages a giant spreadsheet where she's matched the Batch IDs and facility names and then continually updates based on the status of the uploads, whether or not there were errors, when the errors got fixed and final acceptance. By describing it as a "sheila system," she has internalized this myriad of processes as her own, literally enshrining it with her name. While many of the tracking systems participants created are highly varied and convoluted, they are all in service of CMS and QualityNet. The system itself has attached a lot of meaning to a Batch ID, but this number is simply created by QualityNet and only has value within the system as they work to troubleshoot errors with the CMS Help Desk or figure out whether their files have been accepted. It doesn't reflect the facility name, or information that hospital employees and vendors use to organize their own backend systems. Thus the different actions participants take to make this information meaningful to them actually serve as an example of the way CMS guides their actions toward compliance with measures and the complex technological system used for reporting these measures. The fact that Sheila has internalized this practice as her own further demonstrates elements of governmentality as she conflates state practices of compliance with her own personal practices. While she sees it as a "Sheila" system, it is in many ways a CMS system.

Foucault incited this theory of governmentality by examining the modern growth of procedures to produce and continue the life of the nation. Rather than focusing his analysis on concentrated power or sovereignty, Foucault was more concerned with how power is affected by disciplinary and governmental techniques that regulate and order the actions of people (Foucault, 1991). Even the most general definition of governmentality suggests that governance takes place from a distance as the power to influence the actions of others.

CMS's language around Hospital Quality measures, though increasingly mandatory, still claims to respect hospital choice. For example, CMS allows hospitals to make their own decisions about aspects and implementation of quality reporting. However this type of governmentality works from a distance *through* the encouragement (or direction) of 'free conduct'. The extensive witnessing processes however aren't producing true metrics of quality care nor do they protect facilities from financial penalties, low ratings or audits. They are instead simply in service to a punitive system of hospital measurement alliance.

Lucas Introna in his 2015 work analyzing the plagiarism site Turnitin.com talks about the dangers inherent in large data sorting and algorithmic processes and the types of "calculative " practices that spring up around them. He finds Governmentality the best model to understand there new models of understanding practices that stem from sorting, managing and compiling data:

> Governmentality allows us to consider the performative nature of these governing practices. They allow us to show how practice becomes problematized, how calculative practices are enacted as technologies of governance, how such calculative practices produce domains of knowledge and expertise, and finally, how such domains of knowledge become internalized in order to enact self-governing subjects.

As we house these specific practices within a tradition of witnessing, we recognize the moral weight that Quality workers put on their internalized systems of knowledge production and the preference for evidence one can touch and see. Like delicate nesting russian dolls, we use Foucault's notion of Governmentality then to highlight the socio-political uses of such internalized complex practices. For example, the 2018 quality conference CMS officials remind quality workers that their goals create the system

Figure 5.

Despite the morality and personal responsibility that participants described having toward their role, all of the knowledge that they have produced, from reports to emails and spreadsheets showing the status of file submission, are ultimately in service of them building their defense for their facility . While participants placed immense value on the materiality of CMS reports and witnessing this proof firsthand, these reports don't necessarily reflect reality. It's possible to have successfully completed requirements and only get confirmation through a payment. The reports themselves don't reflect evidence or compliance with CMS' rules.

A hospital's financial security relies on a full reimbursement for their Medicare claims. As Sheila's experience illustrates, confirmation that facilities will receive reimbursement is synonymous with this report and discrepancies between what the QualityNet report tells them is true and their own knowledge about their submission throws a wrench of uncertainty into budgeting, causing panic amongst hospital employees. This unease about where one stands in the system is an essential prerequisite to the type of herculean ethical work vendors and hospital systems begin to employ. Built around the pretense of mitigating risks, hospitals create elaborate and often wholly personal and internal processes of knowledge production. These elaborate evidence gathering practices the ethical work of knowledge production serve the same purpose as the panhandling sign mentioned above. They both work to enlist the citizen in enforcement of regulatory policies with the facade of choice and free will. Part of Bill's job as a data vendor is to assuage hospital clients of this fear when system errors arise.

> *"I have to play the good guy and say: no that's really not how this works. There's a process and we just have to follow and adhere to it and try to talk it up because they need to have something to tag onto as a sense of comfort because of what's at stake here. That annual payment update. Once you start talking about that, people pay attention. And you know if you're talking to a hospital that's going to lose 2% of their reimbursement, it's not going to be a pleasant conversation."*

In the face of all of this *wrong* proof, what is believing? If their reporting system itself is faulty, what does it mean to gather evidence? What do these practices mean? Proof it turns out, does not require truth as a necessity in the case of quality reporting.

CONCLUSION

The complex and varied practices by which hospital employees and data vendors manage their interactions with CMS represent a key role of their work as witnesses. These intricate knowledge production systems reflect these actors means of actually creating the evidence of their facilities' quality of care -- a type of evidence that is real in a very tangible and corporeal sense. Two of the most consistent practices described by participants as necessary to the their roles in hospital quality were: archiving and documentation of every interaction with CMS which ultimately culminated in knowledge production. We house these practices of knowledge production in the theoretical framework of Foucault's Governmentality. We focused on how quality workers' behavior is affected and framed by technologies of discipline, and order. We highlight how governmental policies, data technologies and individual techniques shape, regulate, and order the behavior of individuals toward the proliferation of government power. It is not new or novel that people take pride in their jobs or find personal fulfillment in doing things according to protocol. It is the fact that this pride and fulfillment can easily become economic or political weapons against other methodologies of data management. In the context of the Ala Ajagbusi village, women in the households socially organized themselves into an informal institutional arrangement to manage waste, but they also began to morally police women who didn't comply. We hear echoes of quality workers explain away a mercurial system by questioning the thoroughness and systematicity of other workers processes. Power is at its peak when perfectly diffuse. Foucault (1980) argues "Its success is proportional to its ability to hide its own mechanisms". The manipulation of work ethic is one mechanism by which power masks itself in this particular case. Making that which is constraining, tedious and potentially pointless-this process of witnessing we outlined above, or dragging human waste for three miles up a dusty road -appear positive and desirable, becomes the slight-of-hand we all must be vigilant of as we task ourselves to become better workers and better citizens

REFERENCES

Saheed Adebayo Abdulwakeel, and Orjan Bartholdson
2018 "The Governmentality of Rural Household Waste Management Practices in Ala Ajagbusi, Nigeria." *Social Sciences* 7, no. 6: 95.

Bateson, Gregory
1958 *Naven : A Survey of the Problems Suggested by a Composite Picture of the Culture of a New Guinea Tribe Drawn from Three Points of View.* 2nd ed. Stanford, Calif.: Stanford University Press.

Benedict, Ruth
1934 *Patterns of Culture.* Mentor Book. New York: New American Library.

Caldwell, Melissa L.
2004 *Not by Bread Alone [electronic Resource] : Social Support in the New Russia.* Ebook Central. Berkeley: University of California Press.

Center for Medicare & Medicaid Services
2018 "CMS Proposes Changes to Empower Patients and Reduce Administrative Burden." April 24. https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2018-Press-releases-items/2018-04-24.html

Center for Medicare & Medicaid Services
2017 "IQR-FY-2019 CMS Measures Final." November 1. https://www.qualityreportingcenter.com/wp-content/uploads/2017/11/IQR_FY-2019_CMS-Measures_Final_11.1.2017.508c.pdf.

Center for Medicare & Medicaid Services
2018 "Meaningful Measures Hub." May 18. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/MMF/General-info-Sub-Page.html.

Csordas, Thomas J.
2004 "Evidence of and for What?" Anthropological Theory 4, no. 4: 473-80.

Dave, Naisargi N.
2014 "WITNESS: Humans, Animals, and the Politics of Becoming." *Cultural Anthropology* 29, no. 3: 433-56.

Derrida, Jacques., Thomas. Dutoit, and Outi. Pasanen
2005 *Sovereignties in Question : The Poetics of Paul Celan. Perspectives in Continental Philosophy*. New York: Fordham University Press.

Derrida, Jacques., and Marie-Louise. Mallet
2008 *The Animal That Therefore I Am*. Perspectives in Continental Philosophy. New York: Fordham University Press.

Foucault, Michel, James D. Faubion, and Robert. Hurley
1997 *Power*. Foucault, Michel, 1926-1984. Works. Selections. English.; v. 3. London: Penguin, 2002.

Geertz, Clifford
1960 *The Religion of Java*. Glencoe, Ill.: Free Press.

Geertz, Clifford.
1973 "Thick description: Toward an interpretive theory of culture." In *The Interpretation of Cultures [electronic Resource] : Selected Essays,* 3-30. New York: Basic Books.

Given, Lisa M., Editor, and Gale Group
2008 *The Sage Encyclopedia of Qualitative Research Methods.* Gale Virtual Reference Library. Thousand Oaks, Calif.: Sage Publications.

Hine, Christine
2000 *Virtual Ethnography*. London: SAGE.

Hirschkind, Charles
2001 "Civic Virtue and Religious Reason: An Islamic Counterpublic." *Cultural Anthropology* 16, no. 1: 3-34.

Hsu, Wendy F.
2014 "Digital ethnography toward augmented empiricism: A new methodological framework." *Journal of Digital Humanities* 3, no.1, 3-10.

Introna, Lucas D, and Malte Ziewitz
2016 "Algorithms, Governance, and Governmentality: On Governing Academic Writing." *Science, Technology, & Human Values* 41, no. 1: 17-49.

Introna, and Hayes
2011 "On Sociomaterial Imbrications: What Plagiarism Detection Systems Reveal and Why It Matters." *Information and Organization* 21, no. 2: 107-22.

Laidlaw, James
2002 "For An Anthropology Of Ethics And Freedom." *Journal of the Royal Anthropological Institute* 8, no. 2: 311-32.

Li, Tania Murray
2007 "Governmentality." *Anthropologica* 49, no. 2: 275-81.

Mahmood, Saba
2003 "Ethical Formation and Politics of Individual Autonomy in Contemporary Egypt." *Social Research* 70, no. 3: 837-66.

McHugh, Ernestine
2004 "Paradox, Ambiguity, and the Re-Creation of Self: Reflections on a Narrative from Nepal." *Anthropology and Humanism* 29, no. 1: 22-33.

Medicare Prescription Drug, Improvement, and Modernization Act, Section 501b (2003).

Park, Crystal L.
2005 "Religion as a Meaning-Making Framework in Coping with Life Stress." *Journal of Social Issues* 61, no. 4: 707-29.

Suter, W.
2012 "Qualitative data, analysis, and design" In *Introduction to Educational Research: A Critical Thinking Approach.* 2nd ed. Thousand Oaks: SAGE Publications.

Redfield, Peter
2006 "A Less Modest Witness." *American Ethnologist* 33, no. 1: 3-26.

Robbins, Joel
2004 *Becoming Sinners [electronic Resource] : Christianity and Moral Torment in a Papua New Guinea Society.* Ethnographic Studies in Subjectivity ; 4. Berkeley, Calif.: University of California Press.

Saldaña, Johnny
2009 *The Coding Manual for Qualitative Researchers.* Los Angeles: Sage.

Widlok, Thomas
2004 "Sharing by Default?: Outline of an Anthropology of Virtue." *Anthropological Theory* 4, no. 1: 53-70.

Human-Centered Data Science: A New Paradigm for Industrial IoT

MATTHEW YAPCHAIAN
Uptake

Few professions appear more at odds, at least on the surface, than ethnography and data science. The first deals in qualitative "truths," gleaned by human researchers, based on careful, deep observation of only a small number of human subjects, typically. The latter deals in quantitative "truths," mined through computer-executed algorithms, based on vast swaths of anonymous data points. To the ethnographer, "truth" involves an understanding of how and why things are truly the way they are. To the data scientist, "truth" is more about designing algorithms that make guesses that are empirically correct a good portion of the time. Data science driven products, like those that Uptake builds, are most powerful and functional when they leverage the core strengths of both data science and ethnographic insights: what we call Human-Centered Data Science. I will argue that data science, including the collection and manipulation of data, is a practice that is in many ways as human-centered and subjective in nature as ethnographic-based practices. I will explore the role of data, along with its generation, collection, and manipulation by data science and ethnographic practices embedded within organizations developing Industrial IOT software products (i.e. Department of Defense, rail, wind, manufacturing, mining, etc.).

FIGURE DRAWING: OBSERVE AND REFINE

Relating ethnography and data science practices begins in the studio of an artist with a metaphor; figure drawing. An exercise that requires a clothed or nude person (most common), figure drawing is a traditional fine arts practice used by an artist throughout his/her career to continually develop foundational drawing skills. More importantly, figure drawing is an exercise about *refinement of observation*.

An artist drawing a model only glances at their sheet to mark an observation during their session; most of the artist's time is spent observing the model. Drawing is a process of refinement: observe, refine/mark; repeat. The goal is not to render the image in a single pass; rather it is to develop the image over a defined period of time. An artist's gaze fixed too long on the drawing as he/she draws, reveals a practice that results in an image describing what the artist *imagines* the subject to be rather than how he/she exists.

Evidence of a drawing's development (i.e. observations) can be found within the drawn artifact. Each of the figures in Figure 1. and Figure 2. describes a unique pose by a single model that was held over a specific duration of time (i.e. "5 minute pose"). Figure 1's knee is formed over the course of at least 3-4 observations; progression is most visible with this figure in by the number of lines required by the artist to define the knee. Early marks are visible that were refined over time. Development of a drawing is more visible in Figure 2. With the number of marks defining each part of the human form layered over each other.

Observe, refine/mark; repeat. In the context of building and iterating through repeated observations, an artist's practice is similar to that of an ethnographer and data scientist. Not all marks are correct, but they create a set of knowledge that is directional, leading to an accurate solution. All three practices begin with observations and broad marks that are iterated on over time, resulting in a meaningful artifact that is revealed slowly. [1] Images by Tony Cheng

Figure 1. Figure Drawing

Figure 2. Figure Drawing (ran out of time)

INTRODUCTION

The author of this paper is user experience researcher at Uptake, a company specializing in artificial intelligence in the space of Industrial IoT. Within Uptake, software development leverages core resources throughout an engagement, including data science and user experience research. Due to this ongoing working relationship with the data science team, Uptake's user experience researchers have gained a unique understanding of data scientists and their everyday practices.

This view provides Uptake's UX research team an empathetic understanding of and appreciation for the manual, tedious and often-invisible process occupying most of a data scientist's time. Their effort begins with a problem defined by a client, one self-identified or surfaced through field research. Then an ongoing cycle of observing and refining their subject by identifying a source(s) for the required machine data; generating and logging that data, engaging with it, confirming that it is the correct data, and ultimately crafting purposeful models to yield actionable insights.

Data—quantitative and qualitative, supporting data science or UX—in its raw forms is an observation, not evidence. A machine's gears are grinding, the temperature reading is 400° F, a bus will not start, an asset's fluids levels are x, a component is on/off, etc. A single observation is just as a mark on a sheet a paper that contributes to the rendering of a figure.

Actionable insights emerge from this data when rich contextual and historical data is introduced. For example, an alert for a rail engine's loss of horsepower will be triggered when a certain threshold is met. Observation.

This alert is observed in the data during each run at the same location along a specific route. Does the engine need to be serviced? No. Evidence is forged from quantitative and qualitative observations. Upon further inspection, the train passes through a tunnel that has limited circulation, forcing the engine's intake center to be temporarily overwhelmed by the exhaust system, triggering an alert for loss of horsepower. Horsepower returns to normal levels immediately after exiting the tunnel.

Raw sensor readings coupled with qualitative context becomes evidence supporting—in the example above—a decision not to service the engine and ignore all alerts of this type generated at this location. Action may be taken if the same fault occurs at a different location along the same route.

Observe, refine/mark; repeat.

Industrial IoT's soul is large data sets and real world context that require technical excellence and human curiosity to generate meaning from it through an ongoing cycle of observing, refining and marking.

UPTAKE DATA SCIENTISTS AND USER EXPERIENCE RESEARCHERS

A typical figure drawing session is organized as a ring of people, each with a pad of paper and a drawing instrument around a small rise; at the center of this group is a single model posing on the rise. At the conclusion of the pose, each person has rendered an image of the same model from a unique *view*.

Individuals drawing during these sessions are usually fixed to a bench or easel for the duration of the session as a matter of convenience. Models will change poses and their orientation to the group throughout a session, allowing each person to "see" different views of the same model. An alternative approach is to have a model hold a long pose as a group of artists (or single individual) rotate around the model in timed increments to see the same subject from multiple points of view.

A drawing session is an effort to understand a figure through individual poses and views of the model. An artist restricting his or her a view of the model to only front, side or back poses limits their understanding of how the complete form exists in space.

Poses can last for durations of 30 seconds, 1 minute, 5 minutes, etc. Or a pose may be continuous for hours across multiple sessions. No matter the length of time, the initial marks made a on sheet of paper situate the figure within the page's space. These marks do not commit the artist to a final image, that image develops from continuous observation of the model and piecemeal refinement of the drawing.

A project begins at Uptake with each role positioned around the project's challenge. Uptake's data scientists need access to large sets of data—generated by man and machine. Data is not always received in an organized and tidy package ready to be acted on. Data scientists, working with subject matter experts (SME), begin their effort by understanding the data available to them. (Including the asset(s) generating data.) From this initial view, they can determine what additional data and quantity of it is needed. If the required data does not exist, they develop a plan to generate and log it.

Observe, refine/mark; repeat.

As data scientists engage machine data, user experience researchers are understanding the context surrounding the data through preliminary research and SME engagement. This is critical learning ahead of any fieldwork due to the nature of Industrial IoT sites. They are often dangerous spaces with unique obstacles, including: intense security/regulatory requirements (Department of Defense, rail), remote locations that are costly to access (mining), limited number of candidates to speak with (manufacturing), and a required escort for the duration of a visit (all).

Observe, refine/mark; repeat.

Similar to the initial broad strokes of an early figure drawing, these first steps of discovery are to situate the roles before committing to courses of action. Data sources and end users may change over time just as the figure's final image emerges through a series of observations and marks.

INDUSTRIAL IoT 2018

General Overview

Consumer, enterprise and industrial software products have been produced and sold to meet an array of intended and unintended needs for decades. Industrial operations of all types leverage a suite of these solutions to lead their daily operations. Data science driven products leveraging AI and machine learning within Industrial IoT are new tools being developed today to be included in an organization's current suite of tools. These products provide previously unavailable line of sight into operations down to asset components, enabling organizations to develop predictive maintenance practices that lead to increased efficiency, uptime and cost savings.

Ganesh Bell, the president of Uptake, describes an emerging Industrial IoT market by contrasting enterprise and industrial IoT software: "In the last several decades, enterprise software was all about humans entering data and automating business processes. Now we are in a world where machines are generating data; robots and drones are increasingly at work with wearables and humans augmented by machines are generating more data. We believe we will be in a paradigm where it will be about automating decisions versus automating business processes."

Industrial IoT is a critical space because of the volume of data generated (and potential to be generated) that can be leveraged to drive value across industries from transportation to mining, rail to the Department of Defense. Data science and user experience research are core practices in the development of the tools driving value.

I will refer to large industrial organizations that operate fleets of heavy machine assets (rail engines, construction vehicles, components of an oil & gas plant, wind turbines, transit buses, etc.) as "Industrial IoT." Within each fleet, Industrial IoT's current state of technology is an intentional or unintentional mix of "connectivity states" that describe an asset's ability to generate and collect data.

- **Connected** assets generate, collect and act against copious sums of machine data. Many rail engines are embedded with sensors by an original equipment manufacturer (OEM). The generation, logging and use of data are deliberate choices made by an asset's owner.
- **Enabled** assets are generating data but are not configured to collect that information. Example: A transit bus OEM will embed assemble a bust with equipment that generates data by default but the customer may not be logging it.
- **Unconnected** assets are without a sensor-based infrastructure to generate data. An older machine in a manufacturing environment may require the addition of sensors or an edge device to generate and log data.

Many of Uptake's clients—including wind farms, mining and rail operations—have sites located in remote regions across the globe. A parts technician or a maintenance team needing to service a site can expect a commute of several hours to a day to access many of these places. Further complicating these visits is lack of sight into component status on the sites they service. Without visibility of an asset's health, routine maintenance is not prioritized by immediate need.

Uptake's software enables line of sight into a site's operation and, at a granular level, down to an asset. For organizations (mining, manufacturing, etc.) that are in continuous operation with limited scheduled downtime, knowing what component might fail, when that failure will occur and the probability of those insights is a critical advantage to an organization. Annual preventative maintenance (PM) is a current practice used to service assets, but asset failures do not wait for PM appointments.

Industrial IoT's problems are ideal opportunities for data science and user experience research to collaborate in an effort to produce meaningful products that will help shape how industrial problems are solved.

Industrial IoT products are designed for an audience of end users smaller than those of consumer or enterprise products. Within an organization, Industrial IoT software can enable a glimpse into operations for Executives, empower engineers and managers to make decisions with confidence and speed that affect operations, help supply managers curating stock rooms know exactly what they need and for when, and empower the maintenance workers on the floor turning wrenches, performing everyday firefighting and maintenance to be more efficient by taking on problems when they are smaller in scale.

While these industries are slow to change, automation and robotics are transforming the shape and size of workforces on the floor across different types of industrial operations. Individuals present on the floors of companies Uptake engages with recognize machine data will be a core tool of their future practice. Many people occupying these roles are embracing data's emerging presence at work because it enables them to take on problems previously too complicated for them to solve with available toolsets, such as Excel-based applications. These are often complicated tools authored by a single person.

End users receiving insights desire to learn more about their systems by understanding the data behind an insight. Some insights are packaged as a "stop light" (red, yellow and green to signal priority) with the ability for the user to drill down into the data driving the insight. End users, not just data scientists and researchers are continually refining the image before them through a practice of observation and marking

Moving Forward

Industrial IoT SaaS products driving significant financial and operational outcomes from machine data will ultimately lead to broader adoption of solutions offered by Uptake and similar companies by organizations throughout industries.

Uptake's data science and user experience teams are two critical practices found at the point where an insight and an end user meet. Data science produces a model that yields an actionable insight based on a set of conditions, user research must identify which role needs that information and at what time and where in that role's workflow to introduce it for it to provide value.

Data science and user experience research teams dedicate a majority of time to understanding differences between observation and evidences through practices of ongoing refining of their observations and mark making. The challenges of Industrial IoT software development is an array of poses that both roles are continually approaching from complimentary views. Some poses are brief in duration, while other extend over time—inherent in SaaS software. Always observing and refining.

THE TRUTH, THE WHOLE TRUTH, NOTHING BUT THE TRUTH

In a world of "big data," the designed, built, engineered nature of machines—specifically industrial machines—in concert with the rigorous mathematical, computer-based nature of data science, can easily lead customers and end users to believe that data and insights derived from data are objective in nature. In reality, insights are the result of a deliberate process of observing, refining and bringing together quantitative and qualitative data so that the probability of an insight's accuracy is high.

Consider an "edge device" (sometimes also called an "event recorder" or "data logger") on freight locomotive. The purpose of such a device is to collect readings (vibrations, pressures, speeds, temperatures, etc.) from an array of sensors located throughout the locomotive, and to relay those back to computer servers where they can be analyzed to identify signs of impending breakdowns. Sensor readings themselves are measured by precision instruments, and they are relayed automatically without human intervention or hand-offs. This would seem like the ideal set-up for a fully reliable transmission of "objective" information.

Not quite.

Since storing and transmitting data can be expensive, sensors on the locomotive only take measurements under specific operating conditions, for example, when the locomotive is running in a particular throttle position. In turn, the edge device transmits small samples of the data at 15-minute or longer intervals, relying on a set of programmed rules to determine what data to send over, at what granularity (anything from just one to several hundred readings per second), and whether in raw or aggregated form. If for any reason the edge device cannot communicate with the servers, for example due to absence of a cell signal, it will accumulate unsent data on its small hard drive until this hard drive is full at which point it will simply delete some data according to programmed rules to write new data.

The parameters governing these behaviors were at some point designed, programmed and informed by a collection of human beings whose job it is to ensure a certain level of reliability of the locomotive as a whole, itself determined whether contractually or incidentally, by yet another collection of human beings and corporate entities driven by a combination of personal and financial incentives. Considered through this lens, the collection and transmission of information from a machine does not seem so objective at all.

This process of deciding which observations to collect and how to collect them is the result of observing and refining a subject, these are the lines of a figure drawing over time forming a hip or elbow, placing it in a space.

An Uptake data scientist looking to build an algorithm to predict locomotive breakdowns based on this data will engage in an iterative and creative process quite similar, in practice, to that of an ethnographer interviewing a human subject. The data scientist needs to understand the conditions and constraints under which the edge device was designed and built, and interpret the data she receives from it accordingly. Alone at her computer, this task can prove inscrutable, which is why she will spend time talking with those people whose work informed the data collection mechanism with all its quirks and idiosyncrasies. Armed with an ethnographer's toolbox, she can be exponentially more effective.

Point of Origin

Uptake's data scientist may or may not be aware that she is using tools from the ethnographer's box. By looking at an early stage of an algorithm's design and development, we can observe how the two practices depend on discovery, iteration, trial and error, and direct engagement with people, including subject matter experts (SMEs), end users, stake holders, etc..

An algorithm's origin is initiated through numerous possible actions, including a client's request to solve a specific problem, emerging organically through a data scientist pursuing her own curiosity, an internally driven effort, previous experiments, or a lead from ethnographic field research. From here, next steps are a series of questions about quantity and quality of available data—if it exists at all—that are critical in the overall development of an algorithm. This is an iterative process that can include a large number of people from the internal and client teams.

At Uptake, the data scientist assigned to a project will ask if data exists, and is the data provided to her the correct data? How much data is available (volume and duration over time)? What is the source of the data? How was it collected? What is the data's quality? Is additional data required? How much "cleaning of the data" is required? How accessible is the desired data? (Accessibility can be contingent on many variables, including: an asset's network connectivity, security, physical location on an asset, etc.) Uptake's data scientist describe the time required to prepare data so that it will be in a state to have math performed on it as "80% time."

Observe, refine/mark; repeat

Sometimes, in this process of focusing in on the correct data, the pursuit of one data source will reveal itself to be less meaningful than initially thought of, challenging the data scientist to look at other sources. This process often involves ongoing conversations with the client's SMEs.

If data is unavailable from an asset, there are two questions to ask. (1) Is data currently being generated by an asset and not logged? A bus may have the capabilities to generate data streams with built-in sensors, but a logging device has not been connected to collect the streaming data. Or (2) are the hardware/software capabilities that facilitate generation and logging of data from an asset absent? An bus' OEM may not have installed assets components with sensors. For both questions, determining methods and cost required to obtain the necessary data will influence the algorithm's design.

For the later asset types, those that do not generate data of any type, what are the potential solutions to enable generation and logging of data? Is an Edge device required? If it is, will a custom device need to be designed or can an off-the-shelf solution work? Who will design and manufacture a custom device? (An in-house IoT team, third party, etc.?) How does the device's design affect data collection? Observe, refine/mark; repeat

Building out that challenge more, for components that need to be monitored but do not have any type of direct output channel to stream data, the component must be observed indirectly. For example, a wheel can be monitored through rotation count; from this value, information related to a wheel's health might be inferred. Identifying the correct peripheral signals to collect and analyze in an effort to gain insight on a component requires (often) engagement with a subject matter expert (SME) and trial and error.

Decisions related to data that will be generated, logged, cleaned and represented as objective evidence are iterated over time, manually. This process produces a complementary data set that remains invisible—information that will not be generated and logged; a deselection set.

How much data will be transmitted to the cloud and in what intervals? Each decision point in the process of determining the appropriate data to use and why, shapes an algorithm's design and output. Similar to ethnographic and design approaches, data science's process begins with a discovery phase that transitions to an iterative process, then toggles between the two as a point of view is refined. Through trial and error— observe, refine/mark, repeat—an algorithm (a human designed object) develops a specific agenda built on a pastiche of human decisions. Decisions later obscured in a veneer of objectivity by users of a product when encapsulated in "data from the machine."

Data-driven AI Industrial IoT software applications are tools that enable engineers, maintenance workers and other roles within industrial operations to observe, refine/mark; repeat.

Data Scientist and Ethnographers: A Future View

Data scientists and social science based researchers are inherently a motley crew of curious individuals. While the roles "data scientist" and "researcher" can be shorthand for a generic job, each discipline is composed of professionals from a variety of backgrounds and interests. A person may be a weather geek with exceptional math skills or an artist passionate about the intersection of people and technology. These backgrounds and curiosities are fundamental attributes driving these practices, rendering them valuable to an organization.

When organizations recognize the complimentary nature of these two practices and foster synergy between them, more robust products will emerge.

Similar to a studio of artists positioned around a figure model, each person practicing a similar practice of learning and building as they render an accurate but different image of a pose; data scientists and user experience researchers have opportunities to be in the field and office together, experiencing the same client, the same site, and problem from very different (though complimentary) perspectives. Through this mode of collaboration, creativity and understanding flourishes, ultimately contributing to more robust and meaningful practices and products.

Observe, refine/mark; repeat.

NOTES

Thank you Chiara Piccinotti and James Lamb for meaningful conversations and ideas.

1. Figure drawings by Tony Cheng. Images have not been altered other than scale and were found on Flickr. Figure 1.: https://flic.kr/p/G7rUj , Figure 2.: https://flic.kr/p/FhLis . Both images are under Creative Common license with some rights reserved: https://creativecommons.org/licenses/by-nc/2.0/

Towards an Archaeological-Ethnographic Approach to Big Data: Rethinking Data Veracity

SHAOZENG ZHANG
Program of Applied Anthropology, Oregon State University

BO ZHAO
Program of Geography, Oregon State University

JENNIFER VENTRELLA
Program of Mechanical Engineering and Program of Applied Anthropology, Oregon State University

For its volume, velocity, and variety (the 3 Vs), big data has been ever more widely used for decision-making and knowledge discovery in various sectors of contemporary society. Since recently, a major challenge increasingly recognized in big data processing is the issue of data quality, or the veracity (4th V) of big data. Without addressing this critical issue, big data-driven knowledge discoveries and decision-making can be very questionable. In this paper, we propose an innovative methodological approach, an archaeological-ethnographic approach that aims to address the challenge of big data veracity and to enhance big data interpretation. We draw upon our three recent case studies of fake or noise data in different data environments. We approach big data as but another kind of human behavioral traces in human history. We call to combine ethnographic data in interpreting big data, including problematic data, in broader contexts of human behaviors.

Key Words: Big Data, Data Veracity, Human Behavioral Traces, Archaeology, Ethnography

INTRODUCTION

The digitalization of ever more things, although not truly "everything" yet, has led us into an unprecedented era of big data. For its volume, velocity, and variety (the 3 Vs), big data has been widely used for decision-making and knowledge discovery in various sectors of today's society. However, those researchers who are more cautious or critical warn of the potential risks in the hubris or even fetishization of big data analytics (Barnes 2013; Lazer et al. 2014). A major challenge increasingly recognized in big data processing is the issue of data quality, or the veracity (4th V) of big data (Claverie-Berge 2012; Hall 2013; Lukoianova and Rubin 2014) and it is yet being heatedly debated and far from being solved by now (Geerts et al 2018). This paper proposes an innovative methodological approach to big data, an archaeological-ethnographic approach that addresses the issue of big data veracity in particular. With methodological inspirations from archaeology (Cooper and Green 2016; Jones 1997; Kintigh et al 2015; Wesson and Cottier 2014) and ethnography (Moritz 2016; Snodgrass 2015; Wang 2013) among other related fields, our proposal challenges the truth or falsity dichotomy fundamental to big data processing today and approaches big data as human behavioral traces and situational evidence for (re-)contextualized interpretation. We draw upon our three recent case studies of "corrupted" or noise data in different data environments as pilot experiments with this new approach (Ventrella et al. 2018; Zhang and

Zhao 2018; Zhao et al 2018). This paper does not provide a methodic prescription ready to use for data "cleaning;" it is an invitation for epistemological redefinition of big data and methodological reformulation of big data analysis with the hope to appropriate the value (5^{th} V) of big data in more reliable and rewarding ways.

THE COLLAPSE OF TRUE-OR-FALSE DICHOTOMY IN DATA CLEANING

We agree with a simple but crucial observation that big data are "non-standard data" generated by various sensors and digital device users (Gitelman 2013; Schroeck et al 2012). This is why that the problem of poor quality is prevalent in big data of different sources, large databases or on the Web (Saha and Srivastava 2014), or even, as some statistician suggests, that most of the data is just "noise" (Silver 2012). What makes things even worse is what some called the snowballing or butterfly effect of problematic data (Sarsfield 2011; Tee 2013): noise, uncertainties and corruption in raw data can accumulate and be amplified, and therefore compromise the value of big data in both academic and applied fields. Thus, researchers-users of big data have warned the danger of ignoring the data quality issue and urged the establishment of big data veracity before drawing interpretation (Hall 2013; Lukoianova and Rubin 2014; Schroeck et al 2012). The common practice to establish veracity has been data cleaning which simply removes data unpalatable to pregiven rules or algorithms, data described as dirty, noisy, inconsistent, uncertain, or corrupted and so on. According to recent reports, "In most data warehousing projects, data cleaning accounts for 30-80% of the development time and budget for improving the quality of the data rather than building the system" (Saha and Srivastava 2014).

Systematic approaches to data cleaning have been emerging from a range of fields, data science, linguistics, and media studies among others. In our observation, many of those converge at what we would call a context-based approach, although their specific methods and applicability vary. Pregiven data quality rules have been questioned, and context-specific strategies proposed, for example, by combining or corroborating data from multiple sources (Saha and Srivastava 2014; Schroeck et al 2012: 5). Drawing upon Paul Grice's philosophy of language and information, Mai (2013) and others aim to build a new conceptual framework that treats information as a semiotic sign in conversational context and hence addresses information quality as situational and located in context. Building on similar theoretical ground, Emamjome (2014) promotes a new conceptualization of information quality for more context-specific models targeted at big data from social media in particular.

Since more recently, sophisticated models of context-based data cleaning have been developed especially towards automated solutions using software and algorithms (Lukoianova and Rubin 2014; Søe 2018; Storey & Song 2017). For instance, Lukoianova and Rubin reason that high quality big data is "objective, truthful, and credible (OTC)," whereas low quality big data is "subjective, deceptive and implausible (SDI)" (2014). They further argue that data objectivity-subjectivity (or OTC-SDI) variation in many ways depends on its context (Hirst, 2007; Lukoianova and Rubin 2014). They propose to quantify the levels of objectivity, truthfulness, and credibility (OTC) and thus calculate a big data veracity index by averaging OTC levels (Lukoianova and Rubin 2014). In order to assess big data quality and identify false data (e.g. rumors) in social media, Giasemidis et al use over 80 trustworthiness measures including contextual measures such as Tweet authors' profile, past behavior, and social network connections (2016). They develop and train machine-learning classifiers over those measures to generate trustworthiness scores and then filter social media data in an automated manner (Giasemidis et al 2016).

However, in this paper, we challenge the true or false dichotomy in the methodological assumption about big data in current practice of data cleaning. Existing (proposals of) solutions to big data veracity, as discussed above, share the basic methodology of assessing data trustworthiness and then removing those data deemed as false and thus polluting. This methodology, as Søe (2018) points out, follows the ancient philosophical quest for "the truth" which we think is fair enough. However, in current practices in data cleaning, this quest is reduced to an unquestioned assumption that in big data some are simply true and thus ready for knowledge extraction and decision-making consultancy, whereas the rest simply false and only for removal. Or simply put, a true or false dichotomy (Søe 2018). In this paper, we do not take this dichotomous assumption for granted and instead we suggest to first rethink the ontological nature of data nowadays. Scholars from various fields trace the historical and linguistic origin of the concept of data (e.g., Gitelman 2013). For instance, after examining the origination of crowd-sourced geospatial big data, GIS scientists observe that the question of data quality has shifted away from the traditional survey/mapping-based concept to a more human-centric one (Flanagin and Metzger. 2008; Goodchild 2013). However, the traditional focus on truth and falsity disregards the human aspects of data, which is especially problematic in the data environment today. The connections and differences between facts, data and evidence, as delineated by historian Rosenberg in ontological and epistemological terms (2013), provide a unique perspective to reveal the inapplicability of the true or false assumption in big data veracity. Facts have to be true, because facts proven false would cease to be facts; but the existence of data is independent of any consideration of corresponding ontological truth, because "the meaning of data must always shift with argumentative strategy and context" (Rosenberg 2013: 37). With the human contextual aspects of data increasingly taken back into consideration, for example, the human intention as the key aspect in distinguishing misinformation and disinformation, the true or false dichotomy simply "collapses" (Søe 2018). Rosenberg further stresses to "make no assumptions at all about (data) veracity" in mobilizing data for our epistemological process (2013: 37). Inspired by these critical reflections, in this paper, we suggest to suspend this true or false dichotomous assumption and to treat big data as neutral materials or "evidence" signaling their sources.

TOWARDS AN ARCHAEOLOGICAL-ETHNOGRAPHIC APPROACH TO BIG DATA

Guided by methodological inspirations from anthropology, we suggest reinventing the contextual approach for the analysis of big data—including those problematic data—as neutral evidence left behind by human behaviors and situated in broader reality. Although we are no longer preoccupied with the task of judging and removing "false" data as a separate step before data analysis, that doesn't mean we would ignore the troubling veracity issues that have been raised, falsity, uncertainty, biases, incompleteness, spikes and so on. To the contrary, we aim to confront these issues head on, rather than hoping to simply shirk them off. We do so by continuing using the contextual approach that has been evolving as introduced earlier. It has been widely recognized that these troubling issues, such as biases, are intrinsic to big data because after all, data are human creations (Crawford 2013; Gitelman

2013). Drawing upon linguistic theories, Mai (2013) and Søe (2018), as mentioned above, call to approach social media content generation as information behavior in specific conversational contexts. Also focused on social media content, Berghel (2017) argues that fake news should be examined as speech acts in bigger communicative structures and political contexts, for example the online info-wars during the Brexit and U.S. presidential campaigns in 2016. However, way before the recent outburst of interests in post truth on social media, earlier attempts had been made to adapt archaeological and ethnographic perspectives to computer-mediated communication or data, if the term big data was not as widely used yet (Brachman et al 1993; Jones 1997; Paccagnella, 1997).

Archaeological research methodology has been adopted since the 1990s to tap in to the fast accumulating digital data both online and offline. For instance, Brachman et al (1993) aim to develop a methodic system to support data archaeology that digs into digital databases, such as corporate databases, as rich sources of new and valuable knowledge. In their vision, data archaeology is an interactive exploration of knowledge that cannot be specified in advance, and doing data archaeology is an iterative process of data segmentation and analysis (Brachman et al 1993). Nolan and Levesque view the Internet as a giant data graveyard expecting forensic data archaeologists to "sift through memories for past fragments" (2005). For the practical cause of data curation, Goal (2016) promotes the data archaeology approach to recover data encoded or encrypted and data stored in obsolete formats or damaged media. Many others dig deeper into the richness of data and develop interpretive approach to data. For instance, Jones (1997) presents a theoretical outline of a cyber-archaeology approach to online data as "cyber-artifacts" generated and left behind by virtual communities in the Internet. Zimbra et al apply Jones's cyber-archaeology approach to the study of social movement and demonstrate the potentials of this approach in "overcom(ing) many of the issues of scale and complexity facing social research in the Internet" (2010). Akoumianakis and his collaborators have been developing a more sophisticated archaeological approach to Internet-based big data for the discovery of business intelligence among other kinds of knowledge (Akoumianakis et al 2012a; Akoumianakis et al 2012b; Milolidakis et al. 2014a; Milolidakis et al. 2014b). Unsatisfied with existing data archaeology's concentration on excavations of "semantics-oriented properties" of big data, they re-emphasize classic archaeology's commitment in analyzing artifacts in situ so as to evoke particular understandings of the culture within which these artifacts exist (Akoumianakis et al 2012a). In other words, it is not enough to confine the analytical scope to the given semantic content of data. Now treated as digital traces, data are archaeological "evidence" of the activities of particular groups of actors and that of their community culture (Milolidakis et al. 2014a).

Archaeologists have also mobilized themselves to "embrace" big data and, in doing so, have encountered new challenges. On the one hand, the accumulation of archaeological evidence of the traditional kinds, thanks to technological advancements and other historical causes, has been building up big datasets of unprecedented volume and complexity that demands new data tools as well as new methodological strategies (Cooper and Green 2016; Kintigh et al 2015; Wesson and Cottier 2014). On the other hand, we have also seen the recent development in archaeological approach to new kinds of big data, such as online user/crowd-generated, as digital remains of human behaviors and material culture (Cooper and Green 2016; Newman 2011). Amid the recent engagement with big data as such, archaeologists reaffirm the disciplinary tradition and skills in "appreciating the broader interpretative value of 'characterful' archaeological data," data that have "histories," "flaws" and even "biases" (Cooper and Green 2016; Newman 2011; Robbins 2013). Nonetheless, in the data landscape today, the interpretive capacity of archaeology is bounded by the types of data accessible and the tools available for data extraction, analysis and synthesis (Akoumianakis et al 2012a; Kintigh et al 2015). For example, in Akoumianakis et al's research (2012a), Youtube users' demographics or Youtube insight data, which can be very useful and informative, are not entirely available from the Youtube Data API (Application Programming Interface). A related set of challenges is to combine online digital traces with offline activities in order to (better) reconstruct and understand the broader contexts and cultural processes (Akoumianakis et al 2012a). Therefore, facing these challenges, archaeologists call for revolutionary transformation to turn archaeology into a more integrative science which integrates data, tools and models from work in a wide range of disciplines (Cooper and Green 2016; Kintigh et al 2015).

While it is obviously beyond our job and our capability to revolutionize archaeology, we follow these archaeologists' integrative strategy to incorporate methodological wisdoms from another subfield of anthropology—ethnographic wisdoms—into big data research. In an early attempt to adopt ethnographic methods in the study of virtual communities, sociologist Paccagnella (1997) explores the great potentials in integrating the "deep, interpretive" ethnographic research methods with new tools for collecting, organizing and analyzing voluminous online digital data. Informatics scholar Nardi has been interested in tracing Massively Multiplayer Online Role-Playing Game (MMORPG, such as World of Warcraft) playing to its offline sources and uses ethnographic methods to reestablish and understand the social-cultural settings of online gaming behavior (2010). Since recently, drawing upon experience with data collection and processing in and beyond anthropology, researchers trained in ethnographic study have been more openly critical of the fast rising practice of big data analytics (e.g. Bell 2011; Wang 2013), and many of these critiques raise fundamental questions to big data veracity (e.g. Crawford 2013; Snodgrass 2015; Moritz 2016). First, ethnographers are trained to be careful about accepting informants' representation of themselves at face value, due to the potential of people's misrepresentation or even deception especially in computer-mediated contexts (Snodgrass 2015). Second, "ethnographers often take a cross-cultural approach in data collection and analysis because simple words like family, marriage and household in collected data can mean different things in different contexts" (Moritz 2016), and this variation of meaning by contexts is not derived from people's accidental misrepresentation or intended deception but is fundamental to data analysis and interpretation. However, while computational technologies record and make available massive amounts of data, much of these data are "decontextualized and free-floating behavioral traces" (Snodgrass 2015). Moreover, after all, big data are only subsets of behavioral traces left by subsets of people in the world that happen to be captured in the big data sets (Moritz 2016), therefore, however big big data are, they are incomplete and often unrepresentative.

Taking into consideration these concerns and more, "using Big Data in isolation can be problematic," as Wang calls out in her well-read article (2013). Problematic, yet tempting. For their great abundance, read-made streams, and often numeric forms, big data are easy to access, to manipulate using automated programs, and to draw stunning conclusions. In comparison, ethnographic data are often based on a small number of cases, more in qualitative than quantitative/numeric forms, and time consuming to produce and manipulate. Moritz calls "the streetlight effect" this tendency of researchers to study what is easy to study, dubbing the well-known joke of the drunk who searches for his lost wallet at night under the streetlight (2016). There have been pioneering calls and efforts to break the problematic tendency of "using big data in isolation." Amid the overwhelming rise of big data especially in the business world, Honig defends "small data" and calls for refocusing on "the diversity of data available" (2012). From a slightly different perspective, Burreal develops a guide for ethnographers, or the "small data people," to understand and hopefully work with big data (2012). Wang has been a strong advocate for "Thick Data"—extending the term "Thick Description" that anthropologist Clifford Geertz (1973)used to refer to ethnographic methodology—and for the complementarity between big data and thick data (2013). Thick data, although often small in quantity, are good at this fundamental job of rebuilding the social context of and connections between data points so that researchers could uncover "the meaning behind big data visualization and analysis" (Wang 2013).

We take up the pioneering calls and efforts as introduced above and aim to develop a more integrative strategy combining archaeological and ethnographic approaches to big data in the new data landscape today. The so-called big data revolution has been widely debated—celebrated by many, and questioned by some including those on data quality and veracity issues (Barnes 2013; Honig 2012; Lazer et al. 2014; Silver 2012). We agree that it is indeed a revolution. But we take it as revolution not simply for the abundance and easy availability of data. More importantly, we take it as a new data regime that demands methodological innovation. We would not be as pessimistic to disregard most of big data as "noise" (e.g. Silver 2012). Actually, we believe it is unfair for big data to have been accused of serious veracity issues while having being embraced, celebrated, butchered and exploited. What needs to be interrogated, deconstructed, and reinvented is the mainstream methodological approach to big data, including the true-or-false dichotomous judgement and screening before data analysis. We want to reiterate this simple but fundamental observation that big data are "non-standard data" (Gitelman 2013; Schroeck et al 2012)—they are not traditional scientific data produced in chemistry labs or in geology fieldwork following established methodic principles of modern science. Data generated by users on the internet or by sensors installed in people's life, big or small, all are raw and incomplete digital traces of people's behavior and life in the recent past—"naturally occurring social data" as Snodgrass called (2015). Thus, big data, including the discriminated noisy or corrupted data, all can be valid and valuable resources, "cultural resources" to be more accurate (Gitelman 2013). Beyond the easily available big datasets, concerned anthropologists have called big data researchers to get out their labs and do first-hand research by "engaging with the world they aim to understand" (Moritz 2016; Snodgrass 2015). In this new data regime, we take an archaeological approach to the existence and value of big data. As discussed above, we take big data as digital traces of human behaviors and use them as archaeological evidence that should be processed and analyzed along with data from other sources, especially contextual data such as ethnographic data. We believe our innovative methodological approach has the potentials in addressing big data veracity challenge and enhancing big data interpretation.

We explore the potentials of this integrative archaeological-ethnographic approach to big data in our three recent case studies that are presented in the next sections of this paper. These case studies focus on topics and datasets from different fields, location spoofing in mobile online gaming (Zhao and Zhang 2018), fake location-based posts on social media (Zhang et al. 2018), and noise data in sensor-based monitoring of humanitarian technologies performance (Ventrella, MacCarty and Zhang 2018). Nonetheless, they draw upon largely the same methodological approach in development with a few specific methods used in slightly different ways or to different extents. The ethnographic components in the first two case studies rely primarily on virtual ethnography, or online ethnographic fieldwork including specific research activities such as online user profile collection, online post collection, online community participant observation, and online anonymous informal interviews. The ethnographic component in the third case study relies on on-site ethnographic fieldwork in rural communities in Guatemala, Honduras and Uganda including specific research activities such as participant observation, community survey and semi-structured interviews. All these ethnographic research activities were carried out by the co-authors of this paper with our local collaborators' assistance in the third case.

CASE STUDY I: LOCATION SPOOFING IN POKÉMON GO

The worldwide surge of the Location-based game Pokémon Go since mid-2016 has raised wide debates in and beyond online gaming communities. Our study focuses on the unique phenomenon of location spoofing that has been less discussed in these debates but has critical implications in and much beyond this game. Location spoofing has been defined as "a deliberate locational inconsistency between the reported location and actual geographic location where a specific network communication is made to location-based game or other kinds of Internet applications" (Zhao and Sui 2017). Location spoofing has been often simplely considered as generating fake locational data and cheating in gaming. Overall, there is yet rather limited understanding of user-generated spatial data from location spoofing, compared to the well-examined systematic error, outliers, and uncertainty in spatial data. To fill this gap, our study approaches to this proliferating phenomenon as a unique case to engage the fundamental issue of data veracity or quality in the era of big data today. In order to understand the motivations and grasp the associated contexts of location spoofing, we conducted empirical research combining different kinds of data. We collected a big data set of Pokémon Go from the database Pokémapper.co that is the largest one of this kind and the most acknowledged by the Pokémon Go players community. Databases as such are crowdsourced timely by individual players: once a wild Pokémon is sighted, the player voluntarily reports this new finding to the database. Using the API of Pokémapper.co, we collected a dataset of 77,445 Pokémon records on October 21, 2016. These Pokémons were sighted by players from July 10 to October 21, 2016. Beyond that, we also acquired substantial contextual information about the game by being an observing participant in this game and discussing gaming experience with local and online fellow players. In addition, we also used demographic data of New York City and geographic information of downtown Tokyo to contextualize the geographic distribution of Pokémon resources.

Location-based game (Wetzel, Blum, and Oppermann 2012) is a type of digital game in which the physical location of a player in the real world is set to be identical to the location of the player's avatar in the virtual space of the game. Since such game is installed and played in mobile devices, most commonly smartphones, tablet, wearable devices, the physical location of a player can be determined through the positioning system of the mobile device that the player carries. The positioning system in most mobile devices as such can read a series of radio frequencies, including GPS, cellular, crowdsourced WiFi, and possible others (Sommers and Barford 2012). In Pokémon Go, players can locate, catch, train, and level up a virtual creature, called Pokémon, in the game space and, at the same time, projected to the real world. In this way, Pokémon Go merges the real world and the game frame via player's location (Ejsing-Duun 2011; Rao and Minakakis 2003). Reported by yet few observations and discussions, it is not uncommon for players to conduct location spoofing in this game for various purposes, to name a few, downloading the game app, participating in remote battles, catching rarer Pokémons, or levelling up Pokémons (Alavesa et al. 2016; Lee and Lim 2017; Martins et al. 2017; Wang 2017). A few location spoofing techniques or tools have been used, including GPS spoofing apps, VPN spoofing, drones, and dogs. Among these tools, GPS spoofing apps might be the most economic, powerful and popular one. A GPS spoofing app can take over the GPS chipset of a mobile device and report a designated location instead of the real one. By this means, players can virtually visit anywhere as they personally desire and digitally designate. Usually, a GPS spoofing app as such is free or inexpensive, and can be downloaded from Apple Appstore or Google Play Store. This technique of location spoofing enables gamers to engage in remote activities by using simulated, or "falsified", locational information without the gamers physically being out there. Therefore, location spoofing has been largely considered, or rather condemned, as a threat to the underlining fairness of the game and thus to the social order of both online gaming communities and the real world. We argue that the various involved actors—the game players (including spoofers of course), the game company, spoofing bots/apps, drones and dogs, create a new and evolving spatial assemblage and we call it a hybrid space (Althoff, White, and Horvitz 2016; LeBlanc and Chaput 2016).

The spatial distribution of Pokémon resources displays unique patterns and suggests social-economic differentiation. We overlay New York City map with the spots of sighted Pokémons (as from the Pokémapper.co database). The resulted maps (see Figure 4 in Zhao and Zhang 2018) show that most Pokémons clustered at main parks, such as Central Park and Marcus Garvey Park, and famous landmarks such as World Trade Center and Time Square, whereas only few scattered around the suburban areas. This contrastingly uneven distribution of Pokémons makes the game unplayable in suburban and rural areas, as many players reported and an earlier research on Pokémon Go also observed (Colley et al. 2017). We also aggregated choropleth maps of Manhattan with census tracts. These maps indicates that Pokémons are more likely to appear in the neighborhoods with a larger share of white residents (mainly in southern and central Manhattan) than in black neighborhoods (mainly in Northern Manhattan) (see more details in Figure 4 in Zhao and Zhang 2018). This race or ethnicity difference was also found in other cities such as Chicago (Colley et al. 2017). In an even finer scale, we also examined the distribution of Pokémon Go game facilities, such as PokéStops (where players can recharge new times) or gyms (where teams of players battle with each other). These facilities were set up at local businesses as a marketing strategy to lure foot traffic and stimulate local consumptions. With McDonald's as a major sponsor of Pokémon Go in Japan, Pokémon Go has converted local stores of McDonald's into PokéStops or gyms (Yang and Wenxia 2017). To corroborate this strategic association in media report, we count the number of McDonald's local stores converted into gyms in the Chiyoda Ku (aka County) of Tokyo. We found all the McDonald's local stores on Google Map, and then labelled those gyms using Pokémon-radar.net (another online database showing the locations of sighted Pokémons, PokéStops and gyms). As a result, there were 18 McDonald's local stores in Chiyoda, among which 10 were gyms (see Figure 5 in Zhao and Zhang 2018). Obviously, it is a shrewd strategy to turn McDonald's in the real world into

Pokémon gyms in the hybrid space, and it also contributes to the uneven distribution of game resources.

It is in this context of the uneven distribution of Pokémons and game facilities in the hybrid space, we further examine the players' gaming behavior, especially the motivations behind their action of location spoofing. To help players overcome the geographic limitations, Pokémon Go actually offers an alternative option that is buying Pokécoins. Players can buy and use Pokécoins to avoid or reduce the trouble of moving around for capturing and training Pokémons. However, Pokécoins cost real money; and not every player is able to afford or willing to invest. Opposite to its supposed aim of helping players to overcome the uneven distribution of Pokémon resources, Pokécoins have turned out to be another socio-economic mechanism of unequal accessibility and thus aggravated t many players' frustration. Therefore, players have been motivated in multiple ways to manipulate their locational information with various spoofing techniques. For most location spoofing players, their motivation lies in the satisfaction of catching more valuable Pokémons and competing with others in a more time-efficient way. For others including those who are also hackers and inventors of location spoofing bots/apps, they gain especially strong intellectual and emotional satisfaction from their newly developed spoofing techniques to challenge the game rules and even to resist the social-economic inequality and unjustness that they perceived in this game. Our investigation and interpretation thus far advances the understanding of people's gaming behaviors and potentially informs the design, delivery and marketing strategies in the gaming industry.

Our contextualized analysis of location spoofing in this study demonstrates how the human factors—behavioral, social, economic, and emotional among others—give shape to the big data sets that are eventually available for people to conveniently access and use. In this study, we do not make any moral judgement on Pokémon Go players' location spoofing behaviors; nor do we deny or disregard the "falsified" locational data generated through location spoofing behaviors. We take a neutral methodological approach to data inconsistency as in spoofed locational data in this case. Instead of rushing to judge spoofing behaviors as moral or not, we acknowledge the factuality in spoofed or "falsified" data and reveal the rich meanings and underlining logics in inconsistent (and inconvenient) data. By doing so, we advocate for the methodological importance of falsified or corrupted data that often get discarded in data cleaning. We argue that data cleaning by simply screening and ridding inconvenient data runs the risks of losing valuable components of big data sets and threating the integrity of the entire data sets. This case study is meant to be an exploratory and demonstrative experiment with our new approach to big data, including spoofed or "falsified" data, as real data in the sense that they are digital traces of real human behaviors embedded in broad social contexts. It also suggests that big data should not be taken at face value, as their rich values lie in, and thus can only be appropriated in, the social-technological contexts in which the specific big data sets are generated.

CASE STUDY II: FAKE LOCATIONAL DATA IN SOCIAL MEDIA

While big data generated by internet users have been unanimously celebrated and increasingly drawn upon in and beyond both the academia and the high-tech industries for over a decade by now, "post truth" has seemed to strike us by surprise since 2016 especially in social media and been univocally condemned as some blasphemy to today's digital age.

Our second case study seeks to engage the ongoing debates surrounding post truth by examining a collective cyber-protest movement on location-based social media. In late 2016, with the hope to support the local protests against an oil pipeline in construction to pass through the region, tens thousands of Facebook users from worldwide locationally identified themselves to the Indian reservation at Standing Rock, North Dakota using location-based features, mainly check-in and location review. As a result, this online protest movement generated massive volume of fake locational information. In this study, we examine both the locational data and textual content of the "fake" check-ins and location reviews as digital traces of online protests. We reveal the geographical distribution of Facebook protestors and the social-technological network of the involved actors (including Facebook recommendation algorithms) as broader contexts for the interpretation of the fake locational data. This study demonstrates our effort to develop a contextualized approach to the discovering and understanding of fake locational data and broadly post truth in online environments today. This study also combines multiple methods of data collection and analysis and uses data of multiple forms and sources. We built a python program to collect and geocode the check-in and location review posts (the ones made accessible to the public) and then store them in a MongoDB database. Additional information collected and used comes from online and traditional news media, the pipeline company, and government agencies. Moreover, we also conducted a few interviews online and offline with Facebook users who participated in the online protest.

The Dakota Access Pipeline (DAPL) is an underground crude oil pipeline built from June 2016 to April 2017 passing right next to the Standing Rock Indian Reservation. DAPL was strongly opposed by environmental activists and local Native Americans. They deeply worried about the future risks that the local water supplies would be polluted and that the spiritual space of the natives irredeemably stained. Therefore, they had swarmed into Standing Rock and formed several protest camps near the planned DAPL route since early 2016. The on-site protest soon expanded to the cyberspace with sympathizers and participants worldwide, known by the hashtag #NoDAPL in popular social media especially Facebook and Twitter. Our study focuses on the geolocational information streams in this online protest movement (referred to as "the #NoDAPL Movement" henceforth), especially during its peak time at the end of October 2016. Starting from October 30, 2016, a large number of Facebook users expressed their concerns with the pipeline and their supports to this protest in the form of online posts, mainly check-ins to and location reviews of Standing Rock. By the afternoon of October 31, 2016, the number of check-ins went viral from 140,000 to more than 870,000 (Levin and Woolf 2018). Moreover, we also captured 11,915 reviews (out of the approximately 16,000 reviews in total) posted on the profile page of Standing Rock. As clearly stated in many of these posts, most of the Facebook check-in participants and location review authors did these posts without physically being at Standing Rock. Nonetheless, their posts consequentially generated inconsistent locational information in Facebook datasets.

Our mixed-method analysis traces the geographic origin and social formation of the Facebook users' reveals the motivations of the remote check-ins and location reviews. As shown by the time series (see Figure 3 in Zhang et al. 2018), over 99% of the location review posts were posted during the two days of October 30th to 31st, 2016. After geocoding these reviews, we plot the global distribution of the Facebook reviewers of Standing Rock (see Figure 5 in Zhang et al. 2018) and found most of the reviewers were not physically located there around those days. People outside the U.S. also joined the protests both online and offline, and turned the #NoDAPL movement into a global issue. Overall, social media not only gave people the platform to project their concerns and feelings, but also became the virtual bridge connecting geographically disconnected people into a global network of collective actions both online and offline. Initial qualitative analysis of these posts reveals the primary themes and motivations of these posts, including the fact of no physical presence behind most these posts. A word cloud (see Figure 4 in Zhang et al 2018) gives a basic overview of some main terms appearing in check-in and review posts. The high frequency or popularity of key words like "hope", "love", "peace", "human", "water", "solidarity" shows the major sentiments around this online movement. Words like "calling", "people", "EVERYONE", "join", and "share" reveal the grassroots feature of this social media movement. Not as popular but no less important key words like "defeating", "deceived" reveal one of the main motivations behind many of these posts, that is to confuse and overwhelm the police system with their fake check-ins.

Further analysis of the post content, combined with interview data, identifies four major types of participations in the #NoDAPL movement. The first is derived from the popular belief that the local police department and their intelligence program was screening through Facebook's locational data sets to compile a list of protesters and track them down. Therefore, as mentioned above, fake check-ins were meant to collectively flood a stream of potential intelligence for police with voluminous false information, and thus to confuse the police about the number and identity of those actually protesting on site. However, more participants in the #NoDAPL movement did not believe that the police was using Facebook data to track protestors or that they would be able to confuse the police with their fake check-ins even the police was doing so. With that in mind, most of the #NoDAPL movement participants were simply demonstrating their moral and political supports to the on-site protest without the intention to create false locational data or to confuse anyone. Remote check-in, or technically fake check-in, turned out to be a very convenient and highly visible way for them to show their support by virtually "standing" with Standing Rock. Examining the textual content of these posts, we found many Facebook participants were fully honest about their action of online protest through "fake" check-ins. For example, one participant said, "We can't all be at Standing Rock, but we can check in as being there." In thousands of circulated fake check-in posts, the authors clearly stated their stance and motivation as such using similar, if not as succinct, phrases. Third, many Facebook users checked in to or reviewed Standing Rock without clear aims though. After seeing friends' posts or randomly recommended posts indicating an ongoing trend, they simply followed the trend by some harm-free mouse clicks. We can tell this from their posts saying "confused", "not sure", "don't know", "because of the beautiful videos of Standing Rock". Many did not really know what was going on, but still took action out of social media network peer pressure (Cho, Myers, and Leskovec 2011; Seidman 2013) as getting involved with the social network interaction. Nonetheless, their participation did consequentially add to the momentum of the movement, the public pressure on the pipeline project, and the amount of fake locational information. Forth, some other Facebook users "joined" the #NoDAPL movement, but the contents of their posts are completely unrelated to the Standing Rock issue except using the trendy hashtags such as #NoDAPL. They incorporated these trendy hashtags only to increase the visibility of their topically unrelated posts by taking advantage of Facebook's recommendation algorithms. Such participation is not irrelevant to the movement or to our research interest here though; the increased use of the trendy hashtags as such algorithmically amplified the popularity of these hashtags and thus enhanced the visibility and influence of the #NoDAPL movement. These four main kinds of participation were confirmed with responses from our interviews with online protesters.

This case study suggests four tentative arguments. First, our analysis of the fake locational data and the motivations in generating these data poses fundamental challenges to the morally charged description of remote check-ins and reviews as deception or cheating. The second and third types of participations in the #NoDAPL movement described above did not have any intention to deceive anyone. The first and forth types meant to deceive or confuse the police system's data processing programs and the Facebook recommendation algorithms, but not other social media users who would see and read their posts with human eyes. Second, this study provides a unique case of new mode of information generation and diffusion by ordinary people, or namely used- or crowd-generated. In existing works including non-academic debates on post truth and fake news, ordinary people are unanimously treated as passive recipients and consumers of information produced by politicians and mass media. We challenge this elitist approach, and we see ordinary people as actors or agents in information creation and dissemination as well, if not equally powerful. As our case study reveals, fake information could be strategically created by ordinary people and turn out to be bottom-up challenges to or even manipulations of political or technological authorities. Third, our focus on the fake locational data proves once again the rich values and methodological significance of the supposed untrue and unuseful data in big data sets. Our contextualized analysis of data generated by these remote check-ins and reviews provokes us to rethink the true-or-false dichotomy assumed in the currently mainstream practice of data cleaning. In this case, there are obviously inconsistent (locational) data. But among them, only some were intended to be false and deceiving, others not; moreover, those were intended to be false and deceiving only to computerized programs and algorithms, not to human individuals, as in the first and forth motivations described above. New data environments like this are forcing us to rethink our definition of true and false data and to reformulate our methodological approach to big data veracity. Forth, this case study brings forward a unique pattern of interaction between social media users and recommendation algorithms. Many of the involved Facebook users wanted to confuse the police system's locational data screening programs and Facebook's recommendation algorithms, or even more proactively to take advantage of the recommendation algorithms (by using the popular hashtags) to promote their posts and their agenda which were not necessarily related to the protests. Based on this study, we suggest rethinking towards human centered design of algorithms in a new data landscape. Although as non-human actors, algorithms play vital role in the network of social interactions of human beings. In this location spoofing case, the recommendation algorithm, as an invisible function, shaped people's activities. Because of the bias-based preference, social media users are possibly feasted with news illusion. Mainly in response to the phenomena of post-truth, Facebook has recently been testing filtering algorithms to detect and reduce misinformation in the big data generated through social media. Based on this case study, we would point out that social media users have challenged the use of algorithms and call for the integration of human dimensions in algorithm design.

CASE STUDY III: SPIKES IN SENSOR-BASED DATA

This case study is a methodological reflection upon our development of a sensor-based monitoring system and our interpretation of data recorded by this system in the field of development or humanitarian technologies. It experiments with our integrative methodological approach to sensor-based raw data as digital recordings of human behaviors embedded in specific environmental, social and cultural contexts. We report the challenges in traditional automated approach to data cleaning using algorithms and the benefits from contextual ethnographic data for more informed data processing and more accurate data interpretation. Specifically, this case study focuses on the design and testing of a novel sensor system invented to measure the fuel consumption and cookstove use of people in low-income countries, specifically Guatemala, Honduras and Uganda by now in this project. Since its initial stage, this sensor system project employs an integrative or mixed method approach, incorporating ethnographic methods for both sensor system design and sensor data interpretation. This project has harvested 24-hour data from over a hundred sensors in the summers of 2017 and 2018 and it is planned to multiply the sensor installation and the data harvest in the following years. The ethnographic methods used in this study include community meetings, participant observation, informal interviews, focal follow, and semi-structured surveys conducted all on site by our co-author Ventrella, her OSU engineering lab collaborators and local assistants. The ethnographic data turn out to be essential especially in (re-)contextualizing and interpreting spikes and noises in the sensor-recorded data sets. This case study demonstrates how our integrative methodological approach to big data avoids uninformed removal of data outliers and associated misinterpretation of data sets. In other words, this methodological approach helps to maximize the values of sensor-based big data for the design and performance of development technologies. This leads to our suggestion of data diversity and our discussion on less costly data saturation in and beyond the industry of global development.

Today 40% of the global population continues to rely on traditional open fires for cooking and heating (Bonjour et al 2013). To mitigate the harmful health and environmental impacts of this common practice, engineers have designed improved cookstoves with a variety of fuel types to increase the efficiency of heat transfer and combustion. However, adoption and performance of these devices have been found to vary greatly, depending on the design and its ability to meet user requirements. Stakeholders in this new field of development or humanitarian engineering, including academic researchers, non-government organizations (NGOs), funding organizations, climate financing institutions and energy technology industries, have called for better monitoring tools to quantify the adoption and technical performance of these devices. To meet this need, a group of researchers at Oregon State University, including our co-authors, developed the Fuel, Usage and Emissions Logger (FUEL) system, an integrated logging load cell and temperature sensor that measures fuel consumption and cookstove usage in households and schools. Since its very beginning, this project adopted a mixed-method ethnographic approach to understand user context, build empathy, and inform the design process.

Post-analysis of the initial observational data showed the central importance of fuel in both the cooking process and as an indicator of multiple metrics of cookstove performance, which in turn inspired the concept of logging fuel weight. Upon evaluation, it was decided that the logging load cell to weigh a household fuel supply was most likely to meet stakeholder requirements, be technically feasible, and measure the most indicators. It was hypothesized that a logging load cell could be used to determine: (i) the frequency of fuel collection events and amount of fuel collected per event, (ii) fuel consumption per cooking event, (iii) duration of cooking events and number of events, with temperature as a backup measure, (iv) emissions. It was also hypothesized that the fuel holder could: (v) be connected to the load cell could be operated in tension or compression, (vi) double as a carrier during fuelwood collection. After choosing the initial concept of a load cell to measure the indicators of stove performance, the system components were designed. The initial prototype system including the load cell, electronics, thermocouple and carrier is shown in Figure 1 below (see also in Figure 3 in Ventrella, MacCarty and Zhang 2018). a) Fuel weight measurement and storage: A load cell that could accommodate up to 50 kg of fuel was selected. b) Temperature measurement: A thermocouple port was chosen to be flexible in length and accommodate high temperatures. c) Electronics, Data Storage & Transmission: Circuitry design and manufacturing was outsourced to Waltech Systems, a small company in rural Oregon that specializes in custom electronics. Two 1.5 V C batteries were selected to power the logger due to wide availability. To meet the requirement of accurate and remote monitoring, the system was designed to collect data for at least 30 days at a time. Although various modes of wireless transmission were considered, the initial prototype logs data to SD cards, which are reliable, inexpensive, and familiar to field staff. d) Data Analysis: To translate raw weight and temperature data into metrics of cookstove performance and usage, a simple algorithm was developed to integrate reductions in mass over time. Mass changes are also corroborated with temperature to verify an actual cooking event. Fuel consumption and cookstove temperature are then used to report cookstove usage and duration of a cooking event, and be extrapolated to emissions, carbon credits, and averted Disability Adjusted Life Years (aDALYs) (Ventrella and MacCarty 2018).

As mentioned earlier, our new methodological approach turned out to be especially powerful and rewarding at the stage of data processing and interpretation. After receiving data from a field study of 100 sensors in Uganda in 2017, the project team began the procedure of processing, cleaning and analyzing the raw data to output useful and accurate metrics of fuelwood use. As a member of this project team, an electrical engineering student wrote an algorithm designed to remove quick, linear spikes (see Figure 2 below) in the weight data above a certain threshold and applied this algorithm to the data set from Uganda. Outlier data points were attributed to noise or accidental human interaction and removed from logging history using this algorithm. However, another lab member on the team offered some field evidence to explain these spikes in sensor-recorded data. She had been using the FUEL sensor to measure fuel consumption of an institutional sized cookstove in a school dormitory of over 100 girls. The dormitory cook who would be interacting with the FUEL sensor had safety concerns about leaving the sensor out in the dormitory and asked if he could hang it up only to quickly weigh the wood before cooking, and then store it away in a safe space for the remainder of the day. This kind of interaction with the FUEL sensor resulted in brief, linear spikes that represented a rapid but intentional addition and then removal of the wood. These linear spikes were not noise to be removed; they were usable and invaluable data points that could only be explained and accounted for using ethnographic data. This contextual evidence was imperative to more accurately interpret and process the sensor-recorded data. Without the contextual evidence from ethnographic fieldwork, the FUEL project team would still be using data that are not fully representative of actual human activity and are therefore incomplete and not nearly as meaningful. The challenge with the spikes was not an isolated incidence though. Survey and informal interviews identified various concerns or problems that the users reported to have with the FUEL sensor system: fear of LED on the sensor, thermocouple burnt up, the need to chop wood into smaller pieces, minor injury, and fuel holder moving during use. Participant observation helped to re-establish the habitual and deep-rooted process of firewood collection, storage, and the following meal preparation and cooking in different household size and home space layout. Women would often multitask, which created additional complexity in measuring time and determining what was being done with any spare time. The ethnographic methods used also illuminated aspects of the lifestyle, gender relations, and daily rituals that further defined the context in which the problem of more accurately measuring impact metrics was situated.

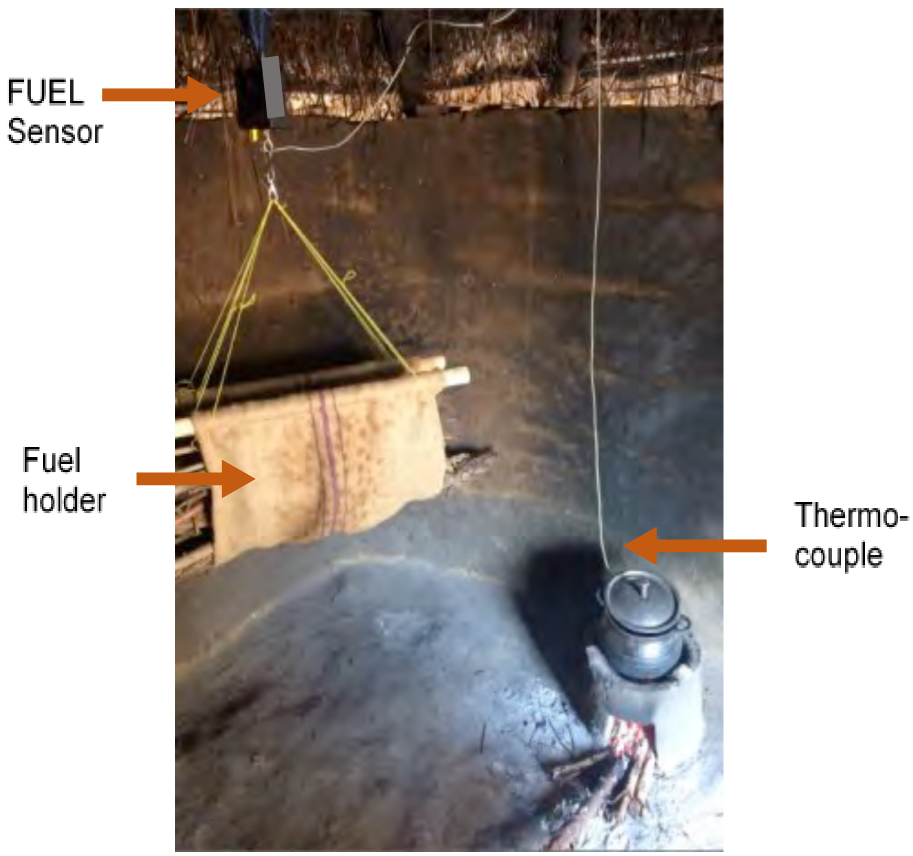


Figure 1: an installed FUEL prototype system in a rural household in Uganda

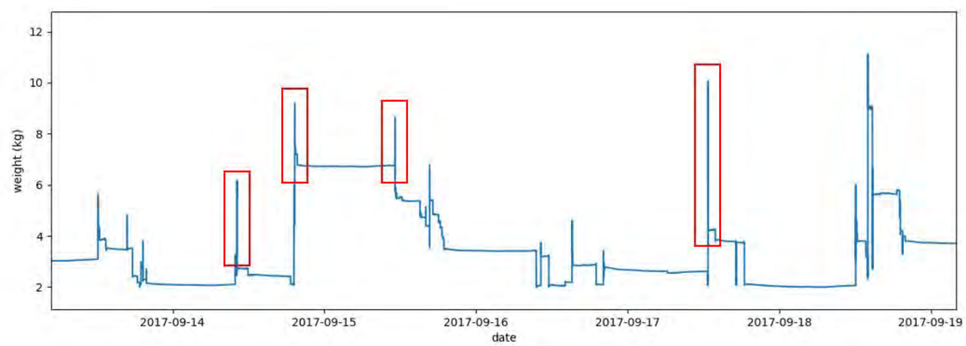


Figure 2. FUEL sensor data, with linear spikes

This case study of a still evolving and expanding project has already turned out to be a very rewarding experiment for us in multiple terms. First, the initial results from this mixed method project so far have demonstrated the essential needs to incorporate ethnographic methods and contextual data. On top of that, we want to point out that digital data recorded by sensors installed in people's real life and data generated in traditional science labs are essentially different. Therefore, we should treat the former differently. We suggest a new methodological approach to sensor-based big data, not as data from controlled or well-defined environment, but as digital recordings of human behaviors embedded in broader social-cultural contexts. Second, this case study successfully demonstrates the power of this experimental approach in turning sensor-based data, especially the outlier data points in our case, into accurate and actionable data. It thus informs and advances the design process, for example by minimizing, if not entirely ridding, the so-called "pro-innovation bias". Beginning with the most simplified version of a product and adding additional functionality only after the context is better understood with our new approach to data can be a more effective method than creating initial complex solutions. This finding speaks to the concept of pro-innovation bias, which theorizes that engineers and designers are biased towards creating new, disruptive innovations instead of implementing more stable changes (Rogers 1983). Third, this is not to underplay the value of big data or to suggest less use of big data. To the opposite, we applaud and embrace the increasing use of big data, sensor-based or not, especially in the fields of humanitarian technologies and global development which have been primarily relying on traditional "small" data (e.g. survey-based) to monitor and evaluate product performance and project impacts. Much beyond the level of individual products or projects, big data can be further used to establish and improve technology standards and to increase accountability and transparency of these fields. Forth, our methodological experiment also seeks to engage the question of data saturation, or simply, how big is big enough. Bigger data usually do have the potential for more objectivity, but that would definitely entail higher costs of human labor and physical resources, especially in the case sensor-based big data. To this question, we fall back to data diversity and advocate for our integrative approach to data—big or small, convenient or inconvenient, true or "falsified"—

all as human behavioral traces that need to, and can, be re-contextualized for more informed processing, more accurate interpretation and eventually more cost-effective use of data.

DISCUSSION AND FUTURE RESEARCH

In this paper, we propose an innovative methodological approach to big data, an archaeological-ethnographic approach that aims to address the challenge of big data veracity and to enhance big data interpretation. It draws upon our three recent case studies of "falsified" or noise data in different big data environments: one on locational data falsified by Pokémon Go gamers using location-spoofing techniques and bots, another on worldwide fake locational check-ins on social media in support of the Standing Rock pipeline protests in North Dakota, and the other on noise data resulted from everyday life activities in sensor-based big datasets in global development projects of improved cookstove. The rich findings from our mixed-method analysis of these problematic data fundamentally challenge the current common practice of data cleaning by simply removing inconvenient data in big data management and analytics. In more fundamental terms, we point to a new data landscape today, made up of hybrid space, decentralized online social movement, and sensor installed in people's life as illustrated above, as the context of the generation or production of big data. We suggest to rethink the epistemological nature of big data as essentially different from data generated in controlled environment of traditional science labs. Hence, we integrate an archaeological approach to big data as but another kind of human behavioral traces in the era of ubiquitous computing in human history. With that, we reason big data should not be taken for its face value, and the inconvenient data should not be simply discarded as unusable noise before analysis. Instead, we suggest to incorporate an ethnographic approach in the processing and interpretation of big data, including inconvenient data, in broader social, cultural, economic and historical contexts of human behaviors. Reusing our case studies, we demonstrate the advantages of our archaeological-ethnographic approach in discovering the values (5^{th} V) in big data. At last, we call to rethink what ethnography and anthropology can offer to other fields, such as data science, tech industries and global development, for more informed knowledge discovery and decision-making in today's world of big data.

Shaozeng Zhang conducts research in applied anthropology and Science and Technology Studies (STS). He received his PhD in anthropology from the University of California, Irvine in 2014. He is currently an Assistant Professor of anthropology at Oregon State University. His email address is shaozeng.zhang@oregonstate.edu

NOTES

Acknowledgments – Thanks to the reviewers and curators of EPIC 2018 for their thoughtful commentaries, which really helped shape this paper.

REFERENCES CITED

Aipperspach, R., Cohen, E., & Canny, J.
2006 Modeling Human Behavior from Simple Sensors in the Home. In K. Fishkin, B. Schiele, P. Nixon, & A. Quigley (Eds.), Pervasive Computing (pp. 337–349). Dublin. https://link.springer.com/content/pdf/10.1007%2F11748625.pdf

Akoumianakis, Demosthenes, Ioannis Kafousis, Nikolas Karadimitriou, and Manolis Tsiknakis.
2012a Retaining and exploring online remains on YouTube. In Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on, pp. 89-96. IEEE.

Akoumianakis, Demosthenes, Nikolas Karadimitriou, George Vlachakis, Giannis Milolidakis, and Nik Bessis.
2012b Internet of things as virtual settlements: Insights from excavating social media sites. In Intelligent Networking and Collaborative Systems (INCoS), 2012 4th International Conference on, pp. 132-139. IEEE.

Alles, M., & Vasarhelyi, M. A.
2014 Thick data: adding context to big data to enhance auditability. International Journal of Auditing Technology, 2(2), 95. https://doi.org/10.1504/IJAUDIT.2014.066237

Asay, M.
2014 8 Reasons Big Data Projects Fail - InformationWeek. Retrieved March 22, 2018, from https://www.informationweek.com/big-data/big-data-analytics/8-reasons-big-data-projects-fail/a/d-id/1297842

Bell, Genevieve.
2011 Big data as a person." In Web Summit 2.0 Conference, October, pp. 17-19. 2011.

Berghel, Hal.
2017 Alt-News and Post-Truths in the" Fake News" Era. Computer 4: 110-114.

Bernard, H. R.
2006 Research methods in anthropology : qualitative and quantitative approaches. AltaMira Press.

Blok, A., & Pedersen, M. A.
2014 Complementary social science? Quali-quantitative experiments in a Big Data world. Big Data & Society, 1(2), 205395171454390. https://doi.org/10.1177/2053951714543908

Bonjour, Sophie, Heather Adair-Rohani, Jennyfer Wolf, Nigel G. Bruce, Sumi Mehta, Annette Prüss-Ustün, Maureen Lahiff, Eva A. Rehfuess, Vinod Mishra, and Kirk R. Smith.
2013 Solid fuel use for household cooking: Country and regional estimates for 1980-2010," Environtal Health Perspectives, vol. 121, no. 7, pp. 784–790.

Brachman, Ronald J., Peter G. Selfridge, Loren G. Terveen, Boris Altman, Alex Borgida, Fern Halper, Thomas Kirk, Alan Lazar, Deborah L. McGuinness, and Lori Alperin Resnick.
1993 Integrated support for data archaeology. International Journal of Intelligent and Cooperative Information Systems 2, no. 02: 159-185.

Burleson, Grace, Nordica MacCarty, Kendra Sharp, and Bryan Tilt.
2018 An Interdisciplinary Mixed-Method Approach to the Evaluation of a Water Treatment Technology in Eastern Uganda.

Burrell, Jenna.
2012 The Ethnographer's Complete Guide to Big Data: Small Data People in a Big Data World. Ethnography Matters. http://ethnographymatters.net/blog/2012/05/28/small-data-people-in-a-big-data-world/

Chatti, D., Archer, M., Lennon, M., & Dove, M. R.
2017 Exploring the mundane: Towards an ethnographic approach to bioenergy. Energy Research and Social Science. https://doi.org/10.1016/j.erss.2017.06.024

Clasen, T., Fabini, D., Boisson, S., Taneja, J., Song, J., Aichinger, E., … Nelson, K. L.
2012 Making Sanitation Count: Developing and Testing a Device for Assessing Latrine Use in Low-Income Settings. Environmental Science & Technology, 46(6), 3295–3303. https://doi.org/10.1021/es2036702

Claverie-Berge, Isabelle
2012 Solutions Big Data IBM. http://www05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf

Cohen, J.
2006 The new world of global health. Science, 311(5758), 162–167. https://doi.org/10.1126/science.311.5758.162

Cooper, Anwen, and Chris Green.
2016 Embracing the complexities of 'Big Data'in archaeology: the case of the English Landscape and Identities project. Journal of Archaeological Method and Theory 23, no. 1: 271-304.

Crawford, Kate
2013 The Hidden Biases in Big Data. Harvard Business Review. http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

Creswell, J. W., & Creswell, J. D.
2018 Research design: qualitative, quantitative, and mixed methods approaches. (H. Salmon, Ed.) (5th ed.). Los Angeles: SAGE Publications.

Durix, L., Carlsson Rex, H., & Mendizabal, V.
2016 Contextual Design and Promotion of Clean Biomass Stoves. Washington, D.C.: LiveWire. Retrieved from https://openknowledge.worldbank.org/handle/10986/25129

Emamjome, Fahame.
2014 A theoretical approach to conceptualize information quality in social media. ACIS.

Fields, D. A., Quirke, L., Amely, J., & Maughan, J.
2016 Combining Big Data and Thick Data Analyses for Understanding Youth Learning Trajectories in a Summer Coding Camp. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education - SIGCSE '16 (pp. 150–155). New York, New York, USA: ACM Press. https://doi.org/10.1145/2839509.2844631

Frechtling, J., Sharp, L., & Westat, I. (Eds.)
1997 User-Friendly Handbook for Mixed Method Evaluations. Retrieved from http://utminers.utep.edu/mtcortez/downloads/mixed methods evaluation handbook from nsf.pdf

Garbarino, S., & Holland, J.
2009 Quantitative and Qualitative Methods in Impact Evaluation and Measuring Results Issues Paper. Retrieved from http://epapers.bham.ac.uk/646/1/eirs4.pdf

Geerts, Floris, Paolo Missier, and Norman Paton.
2018 Special Issue on Improving the Veracity and Value of Big Data. Journal of Data and Information Quality (JDIQ) 9.3: 13.

Geertz, C.

2003 Thick Description: Toward an Interpretive Theory of Culture. In Turning Points in Qualitative Research: Tying Knots into a Handkercheif (pp. 143–168). Retrieved from https://books.google.com/books?hl=en&lr=&id=y3AHb4dJy30C&oi=fnd&pg=PA143&dq=Thick+description:+toward+an+interpretive+theory+of+culture&ots=PVaz12mKr1&sig=E1Rh6T_BDZd9xi-pSIlHohsv9n0#v=onepage&q=Thick description%3A toward an interpretive theory

Geertz, C., & Darnton, R.
1973 The interpretation of cultures: selected essays (3rd ed.). New York: Basic Books.

Gitelman, Lisa. (Ed.)
2013 "Raw data" is an oxymoron. Cambridge, MA: MIT Press.

Gopal, Krishna.
2016 Data curation: the processing of data. Journal of Information Management 3, no. 1: 71-79.

Gunn, W.
1989 Design Anthroplogy. Journal of Chemical Information and Modeling (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Guyer, J. I., & Lambin, E. F.
1993 Land Use in an Urban Hinterland: Ethnography and Remote Sensing in the Study of African Intensification. Source: American Anthropologist, New Series, 95(4), 839–859. Retrieved from http://www.jstor.org/stable/683020

Hall, Kathleen
2013 Data quality more important than fixating over big data, says Shell VP. Computerweekely.com. http://www.computerweekly.com/news/2240186887/Data-quality-more-important-than-fixating-over-big-datasays-Shell-VP

Hirst, Graeme
2007 Views of text-meaning in computational linguistics: Past, present, and future. In G. Dodig-Crnkovic & S. Stuart (Eds.), Computing, philosophy, and cognitive science: The Nexus and the Liminal. Newcastleupon-Tyne: Cambridge Scholars Press.

Honig, Caribou.
2012 In Defense Of Small Data. The Forbes. https://www.forbes.com/sites/ciocentral/2012/03/30/in-defense-of-small-data/#55ebf2aaef29

Islam, M. S., Granger, S. P., Wright, R., Ram, P. K., Hitchcock, D., Jones, T., … Luby, S. P.
2010 Is Structured Observation a Valid Technique to Measure Handwashing Behavior? Use of Acceleration Sensors Embedded in Soap to Assess Reactivity to Structured Observation. The American Journal of Tropical Medicine and Hygiene, 83(5), 1070–1076. https://doi.org/10.4269/ajtmh.2010.09-0763

Jeffery, S. R., Alonso, G., Franklin, M. J., Hong, W., & Widom, J.
2005 A Pipelined Framework for Online Cleaning of Sensor Data Streams. Berkeley. Retrieved from https://pdfs.semanticscholar.org/2dcb/7a97c2fe1184ed4ec2b83e984214a908d0a2.pdf

Johnson, R. B., & Onwuegbuzie, A. J.
2004 Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational Researcher, 33(7), 14–26. https://doi.org/10.3102/0013189X033007014

Jones, Q.
1997 Virtual-communities, virtual settlements, and cyber-archaeology: a theoretical outline. Journal of Computer-Mediated Communication, 3(3).

Kintigh, Keith, Jeffrey H. Altschul, Ann P. Kinzig, Timothy Kohler, W. Fredrick Limp, William K. Michener, Jeremy A. Sabloff, Edward J. Hackett, Bertram Ludäscher, and Clifford A. Lynch.
2015 Cultural Dynamics, Deep Time, and Data: Planning Cyberinfrastructure Investments for Archaeology. Advances in Archaeological Practice 3(1): 1-15. DOI: 10.7183/2326-3768.3.1.1

Latour, B.
1988 The pasteurization of France. Harvard University Press. Retrieved from https://books.google.com/books/about/The_Pasteurization_of_France.html?id=J26KoKtyTxkC

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D.
2012 The whole is always smaller than its parts' – a digital test of Gabriel Tardes. The British Journal of Sociology, 63(4), 590–615. https://doi.org/10.1111/j.1468-4446.2012.01428.x

Law, J.
2009 Actor Network Theory and Material Semiotics. In B. S. Turner (Ed.), The new Blackwell companion to social theory (p. 618). Wiley-Blackwell. Retrieved from https://books.google.com/books?hl=en&lr=&id=Ku51GZC6FwcC&oi=fnd&pg=PA141&ots=RzbSNp4yhv&sig=oHYIgZTcNKLViTnl-_B5e-YN8nY#v=onepage&q&f=false

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani.
2014 The parable of Google Flu: traps in big data analysis. Science 343, no. 6176: 1203-1205.

Lukoianova, Tatiana, and Victoria L. Rubin.
2014 Veracity roadmap: Is big data objective, truthful and credible?. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1065&context=fimspub

Madsbjerg, C.
2017 Sensemaking: the power of the humanities in the age of the algorithm. New York: Hachette Book Group, Inc.

Mai, Jens-Erik.
2013 The quality and qualities of information. Journal of the American Society for Information Science and Technology, 64(4), 675-688. doi: 10.1002/asi.22783

Maxwell, C. R.
2016 Accelerated Pattern Recognition, Ethnography, and the Era of Big Data. In B. Jordan (Ed.), Advancing Ethnography in Corporate Environments (pp. 175–192). Routledge. https://doi.org/10.4324/9781315435459-16

Milolidakis, Giannis, Demosthenes Akoumianakis, and Chris Kimble.
2014a Digital traces for business intelligence: A case study of mobile telecoms service brands in Greece. Journal of Enterprise Information Management, Vol. 27 Issue: 1, pp.66-98, https://doi.org/10.1108/JEIM-09-2012-0061

Milolidakis, Giannis, Demosthenes Akoumianakis, Chris Kimble and Nikolas Karadimitriou.
2014b Excavating Business Intelligence from Social Media. In Encyclopedia of Business Analytics and Optimization, ed. John Wang, 897-908, accessed July 25, 2018. doi:10.4018/978-1-4666-5202-6.ch084

Moritz, Mark.
2016 Big data's 'streetlight effect'_ where and how we look affects what we see. http://theconversation.com/big-datas-streetlight-effect-where-and-how-we-look-affects-what-we-see-58122

Mosse, D.
2004. Cultivating Development: An Ethnography of Aid Policy and Practice. Retrieved March 17, 2018, from https://ebookcentral.proquest.com/lib/osu/reader.action?docID=3386658&query=

Muraskin, W.
2004 The Global Alliance for Vaccines and Immunization: Is It a New Model for Effective Public–Private Cooperation in International Public Health? American Journal of Public Health, 94(11). Retrieved from http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.94.11.1922

Naumoff, K.
2007 Quantitative Metrics of Exposure and Health for Indoor Air Pollution from Household Biomass Fuels. University of California, Berkeley. Retrieved from https://pqdtopen.proquest.com/doc/304900697.html?FMT=AI

Newman, Martin.
2011 The Database as Material Culture. Internet Archaeology 29. https://doi.org/10.11141/ia.29.8. Accessed 3 July 2018.

Nolan, Jason, and Michelle Levesque.
2005 Hacking human: data-archaeology and surveillance in social networks. ACM SIGGROUP Bulletin 25, no. 2: 33-37.

O'Reilly, K., Louis, E., Thomas, E., & Sinha, A.
2015 Combining sensor monitoring and ethnography to evaluate household latrine usage in rural India. Journal of Water, Sanitation and Hygiene for Development, 5(3), 426. https://doi.org/10.2166/washdev.2015.155

Paccagnella, L.
1997 Getting the seats of your pants dirty: Strategies for ethnographic research on virtual communities. Journal of Computer-Mediated Communication, 3(1).

Ponterotto, J. G.
2006 Brief Note on the Origins, Evolution, and Meaning of the Qualitative Research Concept Thick Description. The Qualitative Report, 11(3), 538–549. Retrieved from http://nsuworks.nova.edu/tqr

Robbins, K.
2013 Balancing the scales: Exploring the variable effects of collection bias on data collected by the Portable Antiquities Scheme. Landscapes, 14(1), 54–72.

Roden, C. A., Bond, T. C., Conway, S., Osorto Pinel, A. B., MacCarty, N., & Still, D.
2009 Laboratory and field investigations of particulate and carbon monoxide emissions from traditional and improved cookstoves. Atmospheric Environment, 43(6), 1170–1181. https://doi.org/10.1016/J.ATMOSENV.2008.05.041

Rogers, Everett M.
1983 Diffusion of innovations. Simon and Schuster.

Rolland, B.
2016 Data Sharing and Reuse: Expanding Our Concept of Collaboration - Team Science Toolkit. Retrieved March 18, 2018, from https://www.teamsciencetoolkit.cancer.gov/public/ExpertBlog.aspx?tid=4

Rosenberg, Daniel.
2013 "Data before the fact." Chapter 15 in Lisa Gitelman (Ed.) "Raw data" is an oxymoron. Cambridge, MA: MIT Press.

Rosenthal, J., Balakrishnan, K., Bruce, N., Chambers, D., Graham, J., Jack, D., … Yadama, G.
2017 Implementation science to accelerate clean cooking for public health. Environmental Health Perspectives. https://doi.org/10.1289/EHP1018

Ruiz-Mercado, I., Canuz, E., & Smith, K. R.
2012 Temperature dataloggers as stove use monitors (SUMs): Field methods and signal analysis. Biomass and Bioenergy. https://doi.org/10.1016/j.biombioe.2012.09.003

Saha, Barna, and Divesh Srivastava.
2014 Data quality: The other face of big data." In Data Engineering (ICDE), 2014 IEEE 30th International Conference, pp. 1294-1297.

Schroeck, Michael, Rebecca Shockley, Janet Smart, Dolores Romero-Morales, and Peter Tufano.
2012 Analytics: The real-world use of big data. IBM Global Business Services 12 (2012): 1-20.

Silver, Nate.
2012 The Signal and the Noise. New York: Penguin.

Simons, A. M., Beltramo, T., Blalock, G., & Levine, D. I.
2014 Comparing methods for signal analysis of temperature readings from stove use monitors. Biomass and Bioenergy. https://doi.org/10.1016/j.biombioe.2014.08.008

Smith, Heather
2013 Big data FAQs – a primer. biblog.arcplan.com/2012/03/big-data-faqs-a-primer/

Snodgrass, Jeffrey G.
2015 Ethnography of online cultures. pages 465-496 in Bernard, H. Russell, and Clarence C. Gravlee, eds. Handbook of methods in cultural anthropology. Rowman & Littlefield

Søe, Sille Obelitz.
2018 Algorithmic detection of misinformation and disinformation: Gricean Perspectives. Journal of Documentation, 74(2), pp.309-332.

Sovacool, B. K.
2014 What are we doing here? Analyzing fifteen years of energy scholarship and proposing a social science research agenda. Energy Research & Social Science, 1, 1–29. https://doi.org/10.1016/J.ERSS.2014.02.003

Stanistreet, D., Hyseni, L., Bashin, M., Sadumah, I., Pope, D., Sage, M., & Bruce, N.
2015 The role of mixed methods in improved cookstove research. Journal of Health Communication. https://doi.org/10.1080/10810730.2014.999896

Stern, P. C.
2014 Individual and household interactions with energy systems: Toward integrated understanding. Energy Research & Social Science, 1, 41–48. https://doi.org/10.1016/J.ERSS.2014.03.003

Thomas, E. A.
2017 Beyond broken pumps and promises: Rethinking intent and impact in environmental health. Energy Research & Social Science, 25, 33–36. https://doi.org/10.1016/J.ERSS.2016.12.006

Thomas, E. A., Barstow, C. K., Rosa, G., Majorin, F., & Clasen, T.
2013 Use of remotely reporting electronic sensors for assessing use of water filters and cookstoves in Rwanda. Environmental Science and Technology. https://doi.org/10.1021/es403412x

Thomas, E., Zumr, Z., Graf, J., Wick, C., McCellan, J., Imam, Z., … Fleming, M.
2013 Remotely Accessible Instrumented Monitoring of Global Development Programs: Technology Development and Validation. Sustainability, 5(8), 3288–3301. https://doi.org/10.3390/su5083288

Thomson, P., & Koehler, J.
2016 Performance-oriented monitoring for the water SDG – challenges, tensions and opportunities. Aquatic Procedia, 6, 87–95. https://doi.org/10.1016/j.aqpro.2016.06.010

Ventrella, Jennifer and Nordica MacCarty
2018 Development and pilot study of an integrated sensor system to measure fuel consumption and cookstove use," in 2018 International Design Engineering Technical Conference.

Ventrella, Jennifer, Nordica MacCarty and Shaozeng Zhang
2018 A Mixed-Method Approach: Design of a Novel Sensor System to Measure Cookstove Usage and Fuel Consumption. IEEE GHTC 2018

Wang, T.
2016 Why Big Data Needs Thick Data. Retrieved March 22, 2018, from https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7

Wesson, Cameron, and John Cottier.
2014 Big Sites, Big Questions, Big Data, Big Problems: Scales of Investigation and Changing Perceptions of Archaeological Practice in the Southeastern United States. Bulletin of the History of Archaeology 24.

Wolaver, B. D., & Sharp, J. M.
2007 Thermochron iButton: Limitation of this Inexpensive and Small-Diameter Temperature Logger. Groundwater Monitoring & Remediation, 27(3), 127–128. https://doi.org/10.1111/j.1745-6592.2007.00158.x

Zhao, Bo, and Shaozeng Zhang.
2018 Rethinking Spatial Data Quality: Pokémon Go as a Case Study of Location Spoofing. The Professional Geographer: 1-13.

Zhang, Shaozeng, Yuanyuan Tian, Bo Zhao and Shenliang Chen
Upcoming Stand with #StandingRock: Location-based social media in the post-truth era. The Annals of the American Association of Geographers.

Zimbra, David, Ahmed Abbasi, and Hsinchun Chen.
2010 A cyber-archaeology approach to social movement research: Framework and case study. Journal of Computer-Mediated Communication 16, no. 1: 48-70.

Physicalizations of Big Data in Ethnographic Context

JACOB BUUR
University of Southern Denmark

SARA SAID MOSLEH
University of Southern Denmark

CHRISTINA FYHN
University of Southern Denmark

With the Big Data hype, making digital data accessible and relatable for non-data experts is becoming an increasing challenge. In this paper we suggest Data Physicalization as a novel approach to facilitate conversations with collaborators about the meaning of data. While this challenge has been approached with data visualizations before, we propose that Data Physicalizations can bring stronger engagement and even more depth to the data than visualizations can offer. Based on four design examples, we investigate how data, when made physical, can be used to instigate valuable reflections between field researchers and non-data experts, and can serve as a platform for design. In all cases we worked with Data Physicalization to enhance our own understanding of the field, to engage collaborators in sense-making, and to move towards design proposals. In the concluding section we discuss the intricate process of developing such Data Physicalizations and suggest traits that make them successful.

INTRODUCTION

Our work with Data Physicalization draws on Design Anthropology, in particular as physicalizations seem to help with two challenges: to move beyond the knowledge that is accessible through observation and interview, and to engage our collaborators in making sense of their daily practices and exploring such potential changes that design may bring about.

Gunn, Otto and Smith (2013) argue that anthropology, although aiming for social change, is lacking the tools and practices to engage people in forming their own futures. Gunn and Donovan (2013) encourage anthropologists to adopt the design qualities of materiality (e.g. prototypes, collages, sketches) that can support people in articulating their experiences of the world and their practices. Kelly (2013) showed how the use of tangible interactions in processes of understanding and designing can help uncover deep insights. Design Anthropology aims to close the gap between observing and understanding, thus making ethnography part of an unfolding design practice (Gunn and Donovan 2013). Drawing on Tim Ingold's work they argue that enabling people to make sense of their own experiences through materials, shifts perception of people from passive consumers to skilled practitioners during processes of enactment. Design Anthropology, while keeping a critical stance towards design, is occupied with making things tangible to enable people to make sense of technologies, systems and plans when engaging in everyday activities. Through interventional forms of fieldwork, involving tools such as videos, scenarios, mock-ups,

2018 EPIC Proceedings, ISSN 1559-8918, https://www.epicpeople.org/intelligences

props, prototypes and tangible interactions, knowledge is created through constant iterations of reflection and action (Gunn, Otto and Smith, 2013).

In our work, we have come to understand the power of Data Physicalizations as an entry point to learning how people reflect on their experiences. We invited people to make sense of their every day practices by interacting with their data in material form. These interactions challenged both researchers' and participants' understandings of the data. This paper inquires into Data Physicalization as a phenomenon, rather than taking a method focus. Through four examples of data physicalizations from library and welfare cases we investigate how and why they work to engage collaborators in 'thinking with their hands'.

The short history of data physicalizations

Jansen et al. (2015) coined the term *Data Physicalization* to describe constructs designed to represent (big) data and help people explore, understand, and communicate data – as we humans explore the world around us with all of our senses. Data physicalizations may be static or interactive but have in common that they afford physical manipulation. They may convey (digital) data from systems or allow people to add or construct data about their own experiences. Physicalization is a way to invite individuals into reflective processes. Huron et al (2014) made a similar point with their 'Constructive Visualizations' that enable individuals to express themselves through adding or removing data tokens. Houben et al.'s (2016) proposal of a human-data design approach links data physicalization to learning: when individuals "create, share and use data through tangible and physical visualizations" they learn more about themselves and their environments. Knowledge is continuously constructed and deconstructed through the interactions we have with the world around us (Ackermann 1996; Kafai, 2006). This resonates with the way designers and architects work. They have over many years developed material practices, like model making and prototyping, to gain insights about how people experience the world. Hull and Willett (2017) suggest how data visualization take inspiration from architects.

Within data visualization, extensive research has been made on the aesthetics of "beautiful data" (Steele and Iliinsky 2010; Perer 2010; Wattenberg and Viégas 2010; Lima 2011), but McCosker & Wilken (2014) criticize that focusing on the end result of data visualization misses the opportunity of knowledge creation in the process. They argue that it is the creation of such diagrams, including all of the steps of planning, mapping, drawing and illustrating that generates understanding. This becomes relevant for Data Physicalization, as materiality affords manipulation and expression for active engagement.

Within ethnography, anderson et al. (2009) showed how data visualizations can be designed to involve participants in making sense of their own data, and thus diminish some of the authority that participants tend to give to the "objectivity" of the data. They claimed that this made participants more comfortable at providing explanations of the data, as they could see how some of the collected data could be misinterpreted. One quality to look out for in designing Data Physicalization is thus how they challenge the "objective" look of numbers and graphs.

In human-centred design research it has become popular to utilize materialization to ease the conversation between designers and 'users'. For instance the generative tools of Sanders and Stappers (2014) (collages, Velcro prototyping) and the tangible business models (Mitchell and Buur 2010) both use design materials to surface memories and stories that otherwise can be tacit and difficult to put into words. In the same way Data Physicalizations can be understood as *boundary objects* (Star 1989) that enable people to work together and make sense of the data, even if they have different ways of understanding it. A Data Physicalization can become an object, which "sits in the middle of a group of actors with divergent viewpoints".

Our research is exploratory, we investigate how Data Physicalizations support learning by allowing tangible manipulation, how they open up the process of analysis, how they challenge the 'objectivity' of data, and how they engage participants as boundary objects. In the following we analyze four experiments where participants co-created meaning with different Data Physicalizations.

TWO PROJECT CASES

Our data physicalizations were designed as part of two design projects, one situated in a library, the other relating to chronic pain.

Our first case was a collaborative project with the university library of University of Southern Denmark in Odense. Conditions for libraries have been drastically changing over the past 10 years, and university libraries increasingly find themselves providing groupwork spaces for students, rather than a repository of knowledge (McKay and Buchanan 2014). The library management encouraged us to use the data they collected through their various systems to better understand the practices of their users, and to point out where they might improve their services or the organization of the library. The project was initiated by an engineering department in our university that hoped to understand how human behaviors (and energy saving attitudes) relate to indoor climate. Our study was conducted in collaboration with 20 graduate design students. The data sets came from the gates, self-service machines, counter registrations, printer logs, Wi-Fi access points and environmental sensors, and they were recorded during two particular weeks in the Fall of 2017. After first creating visualizations of these data sets in various combinations, we made field observations in the library and interviewed library users and staff about the patterns we and they might see. As we realized that more was required to engage the librarians in meaningful conversations about their practices, we developed a set of data physicalizations to be used in design workshops with our collaborators, but also for a second round of field studies.

In our second case we studied chronic pain with arthritis patients based on self-tracking data from an App. Self-tracking has become a popular practice, in particular in the Quantified Self (QS) movement where individuals collect data about themselves and become both project designers, data collectors and critical sense-makers (Nafus and Sherman, 2014). Fiore-Silfvast & Neff (2013) report that within e-health, the discourse has shifted from what data means to how it can perform socially or organizationally - as catalysts for behavioral change. This is important for what role data physicalizations play in reflecting on behaviors. Sauvé et al. (2017) used physical artefacts as data representations with the argument that it makes data seamlessly available for the person trying to make sense of it – even if it does not capture the same amount of data as an App visualization. The chronic pain project was conducted by the third author of this paper. We worked with five patients with rheumatoid arthritis, who used a particular App, RheumaBuddy, to keep track of their pain levels day by day. The participants were asked to self-register their pain, fatigue, stiffness and mood on a scale from 1-10 over a period of six to nine weeks. They gave us permission to retrieve their data from the App developer. Through several iterations we developed data physicalizations to trigger conversations about their experience of pain. The company partner DAMAN develops digital healthcare solutions and collaborates with Danish Health Authorities to provide patients with tools to manage their diseases.

Analyzing the variety of data physicalizations designed in the two projects, we realize that they tend to fall into four categories in terms of how they scaffold conversations with our collaborators. In the following we will scrutinize archetypes for each of these categories.

1. TOUCHING DATA ENRICHES CONVERSATIONS

First, we focus our analysis on how physicalizations enable people to 'touch the data' and what kind of insights this triggers with our collaborators. The empirical base for our research is video documentation from one of the workshops, of how five groups in turn engaged with the data physicalizations to make sense of the practices in the library. The groups typically had 4-5 members: librarians, library researchers, design researchers and library users (students). We have transcribed the recordings and analyzed the video with a view to handling, gestures and verbalizations.

3D printed curves of noise and CO2

One of the library physicalizations consisted of 3D-printed shapes, where the jagged edges represented the levels of, respectively, noise on one and CO2 on another object, Figure 1. They were based on 4,031 data points from each of four environmental sensors (IC meters), which collected data on noise levels, temperatures and CO2-levels over the course of a week in different study areas. The data showed how temperatures stayed rather constant, while the noise and CO2 levels were quite high during day compared to during night. These parameters, which we received in Excel sheets, said very little at a first glance. However, filtering the data with respect to time of the day and the area from which it was collected, helped in creating an initial framing of how the data could become useful. We wanted to gain insights about how users are currently using the different spaces of the library, and how these insights could help inform potential re-design of the spaces to better support the activities of the users.

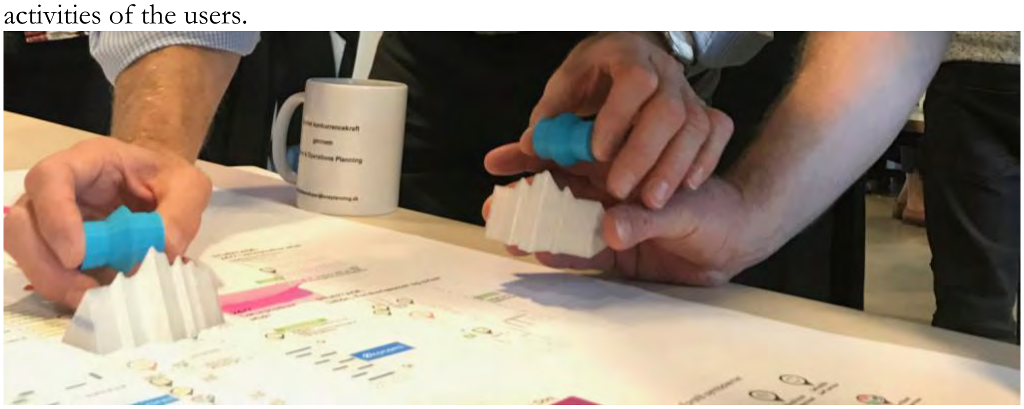


Figure 1. Workshop participants compare 3D printed curves of CO2 levels (blue) and noise levels (white) to find out from where on a library layout plan they might have been collected.

The 3D printed physicalizations imitated trivial data graphs, but being tangible, the data could be touched, moved around, rotated and handed over to others. The artefacts each represented data from one 24-hour day starting and ending at midnight. In the design workshop, the facilitators laid out a large blueprint of the library on the table, then invited each group of participants to discuss how they saw the noise and CO2 levels. This enabled the participants to reflect upon the data in terms of something tangible and relatable.

People make noise and CO2 – but not necessarily at the same time

We look at one illustrative situation: The facilitators hand one physicalization to each participant and ask them to guess where on the map they were recorded. This triggers intense discussions of what the curves mean, and whether they are different at all: "It's the same" – "No, there's two [spikes] here. There's more noise here." The participants begin noticing small details in the curves and in the urge to establish explanations, they start suggesting theories of their own about CO2 and noise:

> H The CO2 level will depend on how many people there are, and how much ventilation there is.
> S I would prefer to move this one *(moves blue CO2 physicalization)* to a smaller room, because you have a big increase in CO2 over a short time. I think that room is too big.
> MK So, from the CO2 point; *(turns both blue curves in hand)* when humans produce CO2 you'll have this kind of spike, whereas this looks more like a place where people are sitting close by and the CO2 comes more times.
> (H is a design researcher, S a library user, MK is a software engineer)

In one of the groups there is a breakthrough, when participants realize that noise and CO2 levels do not necessarily follow each other, as one might think: More people means more noise and more breathing and hence more CO2:

> B OK, the most noisy one in the entrance area. Or what? *(places a white noise curve on the layout plan near to the main library entrance)*
> K *(puts a CO2 curve next to it)*
> H So from the entrance area and upstairs. Of course there'll be noise here. But they won't stay for long, so the CO2 is not there.
> (B is a library researcher, K is a library user, H is a design researcher)

The library entrance area may be very noisy, but has a low CO2 level, as people only move through. On the other hand, the silent reading space is often quite full, but completely silent, so the CO2 level would be high, while noise is low. To an indoor climate engineer such explanations of how human behavior can be read in environmental data will sound banal, but to the participants they help better understand where and when they may hear complaints from library visitors.

Touching data shifts ownership

The physicalizations enabled the participants to move the data around the map, without any permanent changes being made to the data. Assumptions about the spaces and how they were used surfaced as the participants needed to agree upon which area on the map the data might represent, and why. Doing this, the data physicalizations worked as tools for collaborative sense-making of what the data meant in the context of the library space.

With these data physicalizations, comparisons played a central role in making sense of the data. We saw participants constantly matching two against each other, to understand the finer details that make them different. We also observed that often the physicalization got used as a 'talking stick' - participants raised the artifact to be heard in the conversation.

2. JUXTAPOSING DATA SHIFTS PERSPECTIVES

As we saw in the first physicalization example, the act of comparing data is crucial to sense-making. This is well-known from data visualization: it is the difference between curves rather than the curves themselves that help disclose insight. The juxtaposition of data sets suggests new perspectives. This is even more pronounced in the following example.

Laser-cut graphs of chronic pain

For each chronic patient we received 160-250 data points. The data was laser cut in acrylic material in the shape of graphs of different colors, representing pain, fatigue, stiffness and mood, Figure 2. The idea was to enable participants to freely combine and compare the user-generated data from the mobile App.

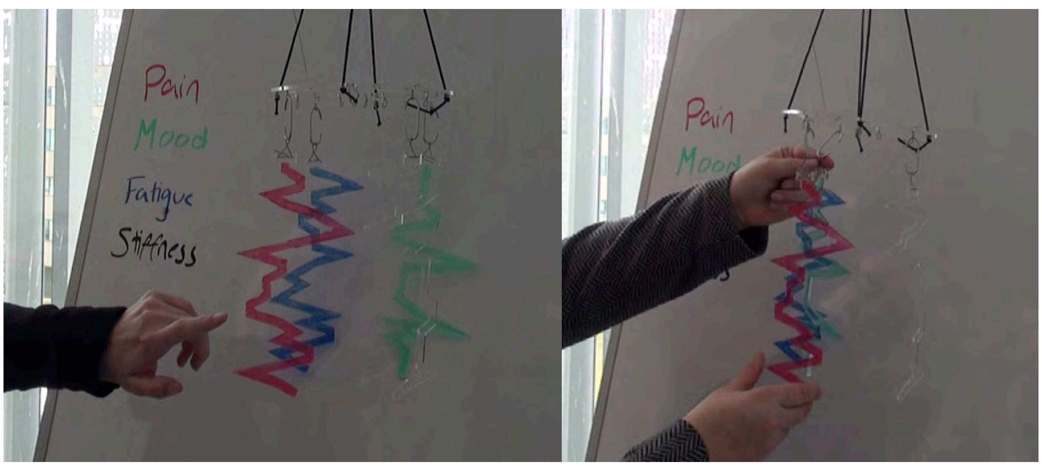


Figure 2. Self-tracking data of chronic pain patients physicalized as acrylic graphs in different colors. To the left the interview participant points out what parts she relates to her story. To the right the researcher places the acrylic graphs on top of each other as encouraged by the participant.

The graphs were hanging from strings and used in interviews with the chronic pain patients: "This is your data from the RheumaBuddy app, where you have been tracking your pain, fatigue, stiffness and mood. I have changed your raw data into these physical graphs, in order for us to see what we can uncover from the data".

Multifaceted pain – and a chronic pain 'filter'

The first participant, suffering from posttraumatic chronic pain, is intrigued to see the graphs on top of each other, so she can get an overview of how the different data relate to each other:

> A It is nice to see that there is a connection between it [the parameters of pain, fatigue etc.] and that sometimes my mood is affected by my pains. And the other way it is nice to see that my mood can be affected just because it is raining, just like everyone else.
>
> *Why do you think there are peaks in the graph?*
>
> A It can be if I have been with many people, or there are many people around me, for instance during a trip to the mall. That can be very tiring. (...) The pain of course comes from me walking around, but that pain would also come if I was just taking a walk. These two *(points at the blue [mood] and green [fatigue] graphs)* are because it can be quite stressing to walk around having to constantly be aware that no one bumps into you [at the mall].
>
> *Do you think having fatigue affects your pain?*
>
> A Yes, because I can't keep up the same filter. I can't ignore the pain. So I don't know whether that could be why I feel more pain when I am tired.

The interaction with the graphs help us – along with the patient – understand more of the complexities of chronic pain, that the pain experience is highly situated. What is interesting in this dialogue is how juxtaposing the acrylic graphs encourages reflection on how pain depends on many other aspects. The concept of a 'filter' helps us understand better the effort it takes to live with pain, and how socializing and fatigue may influence the experience of pain.

Figure 3. To the left the participant gestures through the physicalization how the four graphs can be seen to relate. To the right her friend turns the base plate to examine different combinations of curves.

For an interview with another patient the researcher slotted the acrylic graphs into a plate, so the construct could be picked up, rotated and seen from diverse angles. The chronic pain patient in this interview is accompanied by her friend, who herself suffers from fibromyalgia. The patient explains how she thought there was a certain hierarchy to the parameters, but that after arranging the graphs in one line she can see that there could be certain reciprocal connections between the data, Figure 3.

> F Wasn't this when you were sick?
> B I can't really remember, but I know that I worked a lot in that period, and had a lot of building work. My dad came with new floors. So it was a little stressful to have to prepare for that. So I know this (pointing at the peak) is a result of something. I actually think this was exactly the building period. Or wait, this must have been before, because my mood seemed to be good then. So it must have been while I was working a lot.

The participant first tries to explain the peak in pain level with the stress of having to renovate the floor, but then changes her mind as she sees it in relation to her mood. She then says that it must have been work related, as her mood was good in that period. Obviously the patient experiences pain, but the juxtaposing shows that the pain expression cannot stand alone. The experience of pain is much more than suffering; it affects other parts of the patient's life.

Juxtaposition of curves invites new perspectives

Like we saw in the previous case, the comparing of data sets is instrumental in bringing about theories of what human practices lie behind the data. By turning the square base plate, the participant could choose to compare any two of the graphs – "so now I can look at them like this". "This way you can see these two together". Each combination provided a new perspective on the data. At one point the friend also picked up the physicalization and rotated it, Figure 3 (right):

> F But it doesn't really make sense. This [fatigue] peaks out way more right here compared to this one [stiffness]. So you are much more tired than you are stiff? (she rotates it again) This is actually a little fun. You are actually a very happy person. Well I knew that, but it makes me all happy inside to see it.

The participants explored several angles and relationships, reflecting deeper on how the graphs related to each other. The physicalization turned into a boundary object (Star 1989), inviting collaborators to engage and communicate something that is usually difficult to express. The friend was able to relate to the participant's experience of chronic illness and could help her contextualize the data and bring forward stories and explanations.

3. STACKING DATA HELPS IMAGINE "WHAT COULD BE"

While our first two examples of data physicalizations investigated how people are able to touch data and juxtapose different views of data, in both cases data as such was constant, presented as recorded. In our third example, we explored how participants are able to reconfigure (or de-construct) data in an attempt to both make sense of *what is* and to discuss *what could be*. It also drew on the library project but was based on another data set. The library has 4 service counters with very different 'loads' of queries from the library users. In an attempt to analyze what occupies the librarians at each counter, management installed a self-registration procedure for tracking query types. For the past year, the librarians on duty had been asked to register every query they responded to at each of the four desks by ticking a box on an iPad. The screen had 13 pre-defined categories, for instance:

Question about a book title
Find way to a shelf
Referral to a colleague
Help with pickup of reserved material
Help with book return

Curiously, *'Find way to the toilet'* had its own category, perhaps because the signage in the building was not so very clear. The list led us to ponder about library practices and library design: What activites are important for librarians to manage? Should some of these seemingly mundane tasks rather be shifted to library users to free up time for more 'valuable' work? How is the library lacking in terms of enabling visitors to use the library independently?

Figure 4. The physical bar graph in a librarian design workshop. A physicalization of the 670 queries the librarians respond to at the main service desk during a week, split into 13 color coded categories.

Physical bar graphs of customer queries

We received 1,151 data points spread over four desks during a full week. We chose to focus the physicalization on the main counter with 670 queries during that week. They were represented in the form of a physical bar graph, which proved very successful in the design workshop, Figure 4. The bar graph was a stack of 67 wooden tokens that each represented 10 queries of the same category. Each category had its own color coding. Thus, the physicalization showed the total number of counter queries in a week divided into categories by color. Further, for each participant in the workshop, there was a board with two squares to place tokens on: One said 'Librarian', the other 'Student'. The idea was that for each of the query categories, one could split the stack of tokens into two towers. One for how many of those specific queries 'ought to be' the responsibility of the students, the other for the librarians to handle. This allowed participants to rearrange the data and add new meaning, while maintaining the "numbers".

Like in the 3D curve case, the data physicalization was deployed with five groups of participants in a design workshop. The librarians, library researchers and potential users of the library (students) each received 2-3 categories of tokens and a board and was asked to negotiate with a co-participant how the data tokens should be divided between "responsibility of the student" and "responsibility of the librarian".

I guess they haven't tried

To indicate how participants created new meaning with the data tokens, here follows an example of how two participants (a librarian G and a design researcher H) negotiate how to split a stack of 8 dark green tokens '*Help finding a specific book title*', Figure 5. Should this be the librarian's or the students' responsibility? Is this part of librarians' work?

G I would like something there *(points to the 'student' field on board)*
I guess they haven't tried, or perhaps they have trouble. (…)
H But you would like them to try?
G I would like them to try, yes. Some time we should put them [all]there *(points to 'students')*
H *(shifts 2 tokens out of 8 to 'students')*. Figure 5(left).
G Perhaps one more! *(shifts one more) (laughter)*. Figure 5(mid).
H So it's kind of loose this way. Ideally you would want all of them here *(lifts the rest of the stack towards 'students' and back to 'librarians')*. Figure 5(right).
H In an ideal world, would that be the ratio you'd like, or would you like more like this? *(7 tokens on students, one on librarians)*
G Yes!
H But in the real it is less *(indicates 2 on 'students', 6 on 'librarians')*.

Figure 5. A librarian and a design researcher negotiate how many tokens ('Help finding a specific book title') should go on the 'student' side of the board, and how many stays on the 'librarian' side.

The librarian suggests that some of the current queries could be shifted to the students, if this is simply about asking how to find a specific book title. At least if they 'haven't tried' themselves. They gradually increase the stack of 'student responsibility' until H suggests that they should be looking towards an 'ideal world', where the librarian would 'want all of them here'. But realistically, they agree, only a small part (1 token = 10) of those 80 queries a week can be shifted. An interesting ethnographic finding of this part of the study was that – challenged by the task to split queries between 'Librarian's Work' and 'Student Self-Service' – the librarians formed an agreement about the importance of user contact for their work; that even a book return at the counter, or a trivial question about where to find the toilet, may lead to valuable conversations about books!

De-constructing data sets encourages experiments

The toying with data tokens facilitated dynamic discussions about what the data tell about current practices, and it encouraged the participants to speculate about how things might be different in the future. Different stack configuration experiments could suggest different meanings. Interestingly, the physical bar graph became a tool for the less knowledgeable workshop participant to interview the experienced participant about the current and potential future practice: Do you do this? Would you like that? This pattern was confirmed in other groups too. But then, analyzing the videos, we realize that this is as much a product of the workshop facilitation: When the token stack was 'owned' by ones less knowledgeable about practices, they would use the tokens as an interview tool. If the stack was placed in front of a librarian, *they* would take initiative, and their co-participant rather confirm the arguments and actions without touching the tokens.

4. CO-CONSTRUCTING DATA PROVIDES DEPTH

Our last example takes the physical manipulation of data even further, to the co-construction of data with collaborators in the field. The data came from the same library case: how heavy is the wi-fi use in different parts of the library, and what might that tell us about visitor use of the library layout? We started out with 23,000 data points collected from 45 access points during a single 24-hour period. Armed with a map of the transmitter locations, we first filtered the data to find the areas with the highest numbers of wi-fi connections, assuming that those must also be the areas with most people. Further, the data was filtered with respect to time, narrowing down to the period from 12 to 5 pm. Our aim was to understand why these areas were most commonly used, and to investigate if the spaces were appropriately designed.

Tracing wi-fi use diagrams in the library

For our field studies in the library we printed graphs of the wi-fi data, but we additionally prepared physical boards with yarn, pins and sticky notes to invite participants to construct their own data physicalization of how they perceived their wi-fi use throughout a day, Figure 6. With the pins, one would be able to create a 'graph' of highs and lows against the hour axis, and the sticky notes would help explain.

In the field, we started out observing those areas, which we assumed were used by most students in the selected time span. We quickly learned that our initial assumption was wrong: there simply were not as many students as the data would have us believe. Explanations were found in extra self-service machines connected to wi-fi and a full cafeteria downstairs. Realizing that the students were elsewhere, we searched for the popular areas and found interviewees: a female student, two young male students and a group of three young students (one male, two females). With a starting point in the data visualizations, we invited them to discuss how they were using wi-fi in the library, and then challenged them to show us in a yarn diagram.

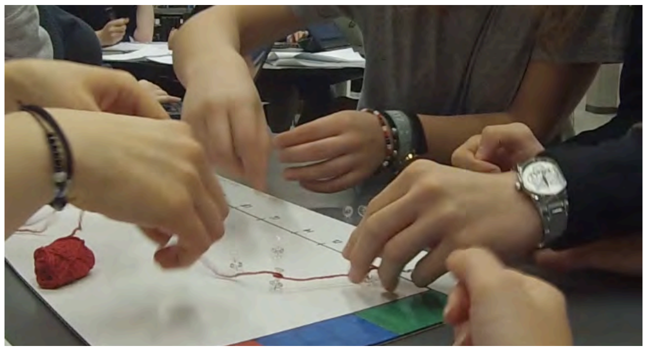


Figure 6. A group of interviewees co-create a diagram with yarn and pins of how they perceive their use of wi-fi through a full day at the library.

Importance of wi-fi – or concentration

As the timeline of the board runs for a full 24 hours, the interviewees start their graph in the morning, before going to university. The female student explains how "I hear music when I get ready, so I probably use a little bit", while she pulls the yarn up slightly on the board. The two males conversely put their yarn all the way up on the wi-fi scale in the morning:

> S2 Checking [on Facebook] that everything is okay in the world (...) because I have nothing else to do. When the alarm goes off I am not one to jump out of the bed. I will be lying there anyways, and if I have no internet I will be pissed off probably.

When asked about the period in the library, the female interviewee stretches out the yarn to a straight line fairly low on wi-fi use explaining "here I am studying". In the same period, the two males bring the yarn up completely to the top stating, "because here we have to be productive, which requires us to be really focused". And the three students discuss how "after this period we are probably tired", while making a straight line with the yarn followed by a sudden low. Towards the evening time, the female student puts the string up high to the red area stating:

> S1 Well it is because I think I watch too much Netflix. And I use the wi-fi much more here. Over here *(pointing to the day time)* it is much more constructive work, where it is like I have to use the internet, while here *(evening time)* I could be doing something else.

The two males put the string very low by the end of the day with the explanation "I will be sitting alone at dinner, maybe scrolling through my phone".

What is important to notice is how wi-fi and the use of the internet is conceptualized very differently from what is measured. While the data on wi-fi was derived from the system based on the number of connected devices at any time, the data physicalizations constructed by the interviewees showed an entirely different story. The two males made a high wi-fi peak in the morning explaining the *importance* of having wi-fi. They also put the wi-fi high when they had to work, because it required them to be *focused*, while the female put it rather low with the explanations that she would not be using wi-fi as she had to be *concentrated*. She consequently put a high peak in the evening, because it represented the *guilt* associated with using wi-fi to watch too much Netflix, while the other participants put it low with the argument that using wi-fi in the evening is not that *stressful*. Observations confirmed that group work in the library would commonly mean less use of wi-fi, adverse to our assumption of 'more wi-fi, more activity'.

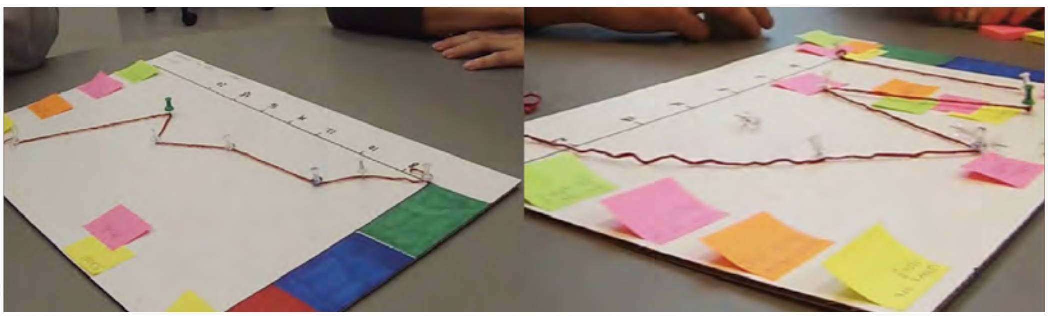


Figure 7. One physicalization shows a high peak in wi-fi use by the end of the day to represent the "guilt" associated with Netflix consumption, while the other explains the low by the end of the day with how not stressful it is.

Co-constructing curves provides depth

Data physicalization became a tool for discussing something the interviewees did not put much thought to. It was not exact to a particular scale but provided a depth of insight, which cannot be derived from the system data. The simple materials (yarn, pins) allowed the participants to change their mind and the placement of the yarn while reflecting. This would not have been possible, had we asked them to draw the graph on paper. Inviting participants to co-construct their own data physicalizations, helped us bring new meaning to the Big Data in our possession.

DISCUSSION

Based on our experiments, we reflect on the design of data physicalizations and how they can add depth and engagement to the understandings of data. In the library case, the data physicalizations quite clearly seemed to empower the librarians to verbalize their practices and values, and to think along in how practices might change. We learned, for instance, that *concentration* is still a core quality of being at the library, it just cannot be adequately expressed in noise measurements or wi-fi use figures. Habits play a strong role in how users approach librarians – but core to the librarians' work is really the personal contact, which can come about in many ways. With our engineering colleagues, the sessions helped challenge the assumption that the data from the 'system' can easily be used as basis to understand human behaviors and even suggest improvements to the interior and services.

Similarly, in the chronic pain case, the physicalizations to the relief of the patients provided a richer means of expressing their experiences than just talk, or pain scales. With the patients we learned that the experience of pain is highly situated and multifaceted; that the momentary pain experience will be filtered or masked through for instance moods and fatigue. The outcomes of the data physicalization sessions – along with those of a cultural probes study and autoethnography - helped us design artefacts for our collaborators to use themselves to convey how they feel to relatives and doctors.

How does data physicalization work?

Our experiments show that it is indeed possible to bring data like the sets used here on a physical form where it can successfully support conversations about human practices and experiences. But what have we learned about how the physicalizations work more precisely?

Touching Data – touch seems to convey responsibility and ownership: Participants feel compelled to do something with the data, explain what they might mean, where they belong. Also, the physicalizations can work as 'talking sticks', bestow the right to talking turns.

Juxtaposing Data – the easy comparison of different sets of data against each other provides new perspectives. It asks questions about human practices that need answering. Comparison is not just a question of visible differences, it may as well engage touch or other senses.

Stacking Data – the de-construction of data sets is a powerful encouragement to perform small experiments and discuss how practices might be different in the future. We have seen how this de-constructing may serve as interview prompts between the participants.

Co-constructing Data – when challenged to express their experiences in numbers and curves, participants reflect about their practices. This provides a depth of insight seldom achievable through interview and observation.

How 'big' data can be physicalized?

Physical representation brings along restrictions when they are based on large data sets (Big Data), as they naturally cannot capture the entire data sets. Did we indeed relate to Big Data, and is physicalization of Big Data at all possible? In the traditional definition of Big Data as 'data sets that are so big and complex that traditional data-processing application software are inadequate to deal with them' (Wikipedia), even our 23,000-point set hardly qualifies as Big Data. On a human scale, however, they are well above what we can deal with. "Big" may be understood differently in our four examples, stressing how the applications of data physicalizations differ. In most of the examples we had to reduce raw data points available to be able to physicalize the data. We will argue with Noraisas & Karpfenm (2014) and Nafus (2016) that the reduction of data must in itself be seen as a process of creating comprehensible meanings. The process of reducing data poses relevant questions for deciding which data to omit and, as we saw in the cases, sometimes the questions lead to assumptions that prove wrong. While physicalizations may require reduction of data, their real strength lie in how they encourage manipulation, expression and questioning, and elicit rich stories and co-constructed theories – something Big Data or data visualization can fail to do.

Is Data Physicalization promising method?

None of the participants prior to the workshops had been exposed to data physicalizations. Therefore, we used shapes familiar to them: curves, bar graphs, making it more relatable for the participants. As data physicalization develops, we expect more vivid designs and technologies to support those designs. To spread this approach as a general method, we need many more examples and design principles to help transform data into meaningful physical designs. This is a worthwhile challenge for (Tangible) Interaction Design research.

Data physicalization has a potential of contributing to the expanding practices of turning field observations into 'material' that can serve as basis for participatory sensemaking (Buur & Sitorus 2007), as has been underway with Cultural Probes (Gaver et al 1999; Mattelmäki 2006), video material (Buur et.al 2000; Raijmakers et.al 2009) and Provotypes ('provocative prototypes', Boer et.al 2013). In their simplest form we might understand data physicalizations as interview prompts, to be used to challenge informants to tell us about their practices and experiences. However, along with the recent developments in Design Anthropology, we rather like to see them as tools in interventional fieldwork to engage collaborators in making sense of their own practices and forming their own futures.

Jacob Buur is a professor of user centered design researching the role of design materials and video in participatory innovation and interventional fieldwork. *buur@sdu.dk*

Sara Said Mosleh is an interaction designer completing her graduate studies in IT Product Design. *sasmo@sdu.dk*

Christina Fyhn Nielsen is a multimedia designer and research assistant in computational thinking. cfni@sdu.dk

ACKNOWLEDGEMENTS

We would like to thank the IT Product Design students at University of Southern Denmark for their creative ideas for data physicalizations in the library case, and Mary Karyda, Aalto University, for inspiring them. Also, thanks are due to librarians, library users and chronic pain patients, who accepted to collaborate with us.

REFERENCES

Ackermann, Edith
2004 Constructing Knowledge and Transforming the World. A Learning Zone of One's Own: Sharing Representation and Flow in Collaborative Learning Environments Part 1: 15–37.

Aipperspach, Ryan; Rattenbury, Tye Lawrence, Woodruff, Allison; Anderson, Ken; Canny John. F; Aoki, Paul
2006 Ethno-Mining: Integrating Numbers and Words from the Ground Up. (EECS-2006-125.pdf).

Anderson, Ken; Dawn Nafus; Rattenbury, Tye, and Aipperspach, Ryan
2009 Numbers Have Qualities Too: Experiences with Ethno-Mining. *In* EPIC 2009 Proceedings Pp. 123–140.

Boer, Laurens; Donovan, Jared, and Buur, Jacob
2013 Challenging industry conceptions with provotypes. CoDesign 9(2), Taylor and Francis: 73-89.

Buur, Jacob; Binder, Thomas and Brandt, Eva
2000 Taking Video beyond 'Hard Data' in User centered Design. *In* Proceedings of Participatory Design Conference, New York, NY, 21-29.

Buur, Jacob, and Larisa Sitorus
2007 Ethnography as Design Provocation. *In* Ethnographic Praxis in Industry Conference Proceedings Pp.

Curran, John
2013 Big Data or 'Big Ethnographic Data'? Positioning Big Data within the Ethnographic Space. Ethnographic Praxis in Industry Conference Proceedings 2013(1): 62–73.

Fiore-Silfvast, Brittany, and Gina Neff
2013 What We Talk about When We Talk Data: Valences and the Social Performance of Multiple Metrics in Digital Health. Ethnographic Praxis in Industry Conference Proceedings 2013(1): 74–87.

Gaver, Bill, Tony Dunne, and Elena Pacenti
1999 Design: Cultural Probes. Interactions: 21–29.

Gunn, Wendy, and Jared Donovan
2012 Design and Anthropology. Wendy Gunn and Jared Donovan, eds. Pp. 1-16. Surrey: Ashgate.

Gunn, Wendy, Ton Otto, and Rachel Charlotte Smith
2013 Design Anthropology. Wendy Gunn, Ton Otto, and Rachel Charlotte Smith, eds. Pp. 1-29. Madison: Bloomsbury Publishing Plc

Gwilt, Ian, and Nick Dulake
2017 Understanding the Needs and Desires of Service Users in the Design and Creation of Meaningful Physical Data Developed by the Authors and Is Presented in The. *In* The ACM SIGCHI Conference on Designing Interactive Systems. Brian O'Keefe and Augusto Esteves, eds. Edinburgh: ACM. Sheffield Hallam University.

Haines, Julia Katherine
2017 Towards Multi-Dimensional Ethnography. Ethnographic Praxis in Industry Conference Proceedings

Houben, Steven, Connie Golsteijn, Sarah Gallacher, et al.
2016 Physikit: Data Engagement Through Physical Ambient Visualizations in the Home. *In* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16) Pp. 1608–1619.

Hull, Carmen, and Wesley Willett
2017 Building with Data: Architectural Models As Inspiration for Data Physicalization. *In* Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems Pp. 1217–1264.

Huron, Samuel, Sheelagh Carpendale, Alice Thudt, Anthony Tang, and Michael Mauerer
2014 Constructive Visualization. *In* Proceedings of the 2014 Conference on Designing Interactive Systems - DIS '14 Pp. 433–442.

Jansen, Yvonne, Pierre Dragicevic, Petra Isenberg, et al.
2015 Opportunities and Challenges for Data Physicalization. *In* Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15 Pp. 3227–3236.

Kafai, Yasmin B.
2006 Chapter 3: Constructionism. In S *The Cambridge Handbook of the Learning Sciences*, awyer, R. Keith (Ed.) Pp. 35-46. Cambridge University Press

Mattelmäki, Tuuli
2005 Applying probes – from inspirational notes to collaborative insights. CoDesign 1(2). Taylor and Francis: 83-102.

Lanclos, Donna, and Andrew D Asher
2016 "Ethnographish": The State of the Ethnography in Libraries. Weave: Journal of Library User Experience 1(5): n.p.

McCosker, Anthony, and Rowan Wilken
2014 Rethinking "big Data" as Visual Knowledge: The Sublime and the Diagrammatic in Data Visualisation. Visual Studies 29(2). Taylor & Francis: 155–164.

McKay, Dana, and George Buchanan
2014 On the Other Side from You: How Library Design Facilitates and Hinders Group Work. *In* OzCHI '14 Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design Pp. 97–106.

Mitchell, Robb, and Jacob Buur
2010 Tangible Business Model Sketches to Support Participatory Innovation. *In* Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design Pp. 29–33.

Nafus, Dawn
2014 This One Does Not Go Up to 11 : The Quantified Self Movement as an Alternative Big Data Practice 8: 1784–1794.

Nafus, Dawn
2016 The Domestication of Data: Why Embracing Digital Data Means Embracing Bigger Questions. Ethnographic Praxis in Industry Conference Proceedings (2014): 384–399.

Norvaisas, Julie Marie, and Karpfen, Jonathan "Yoni"
2014 Little Data, Big Data and Design at LinkedIn. Ethnographic Praxis in Industry Conference Proceedings 2014(1): 227–236.

Raijmakers, Bas; Sommerwerk, Andreas; Leihener, Julia and Tulusan, Indri
2009 How sticky research drives service design. Proceedings of Sercice Design Network conference, Madeira

Sanders, Elizabeth B N, and Pieter Jan Stappers
2014 Probes, Toolkits and Prototypes: Three Approaches to Making in Codesigning. CoDesign 10(1). Taylor & Francis: 5–14.

Sauve', Kim, Steven Houben, Nicolai Marquardt, et al.
2017 LOOP: A Physical Artifact to Facilitate Seamless Interaction with Personal Data in Everyday Life. *In* Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems - DIS '17 Companion Pp. 285–288.

Star, Susan Leigh
1989 Chapter 2 – The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving. *In* Distributed Artificial Intelligence. Les Gasser and Michael N. Huhns, eds. Pp. 37–54. Elsevier Inc.

Steele, Julie, and Iliinsky, Noah
2010 Beautiful visualization: looking at data through the eyes of experts. O'Reilly Media, Inc.

Wattenberg, Martin, and Viégas, Fernanda
2010 Beautiful Visualization: Looking at Data Through the Eyes of Experts. *In* Beautiful history: Visualizing wikipedia. Pp. 175-91. O'Reilly Media.

Empathy Is not Evidence: 4 Traps of Commodified Empathy

RACHEL ROBERTSON
Shopify

PENNY ALLEN
Shopify

Product teams, including our own, often interpret empathy as evidence. However, in practice, empathy is actually something that drives us to seek evidence. By observing and evaluating various examples within Shopify, we have identified 4 traps that are common in the way empathy is manifested. We modelled the relationship between empathy, problems, evidence, and decisions to provide strategies for how to use empathy effectively while being sympathetic to its limitations. Since empathy drives us to seek evidence, and thus cannot be considered evidence itself, empathy must be used at an appropriate level of abstraction throughout the product decision-making process in order to influence good decisions.

INTRODUCTION

Intentionally gaining empathy for users has become an essential way of developing better products or services. Empathy, or getting as close as possible to users to understand their needs, is now a prolific term in the field of user experience (UX). It has become a constant theme in books, articles, conferences, events, career postings, and is a core concept in user-centred design.

At Shopify, empathy is at the very core of our business. As a commerce platform for entrepreneurs (we refer to them as 'merchants'), Shopify's mission is to 'make commerce better for everyone'. We have an organisational culture team dedicated to inspiring us to care about our users and being 'merchant obsessed' is encoded as one of our internal company values. The idea of empathy is ever-present in our work at Shopify—we even specify it as a required skill in our recruitment of UX practitioners. Yet in UX, we have not critically explored the relationship between empathy and the way we make decisions about the products we build and the people for whom we build them. Through the research presented in this paper, we uncover how empathy has influenced Shopify's approach to product development and decision-making.

Empathy is subjective and context-dependent

Currently, in UX, empathy has been positioned as a singular characteristic: we aim to 'have empathy for users'. Crucial as it may be in motivating us to solve user problems, developing this kind of emotional connection is a complex process. People experience and feel empathy in different ways depending on the individual or the subject that elicits the empathetic responses. Empathy, therefore, creates a paradox. While we use it to design for people who are not necessarily like us, there are various influences at play that can make us experience empathy differently depending on the context. This effect could, therefore, make empathy a problematic tool for making decisions if it is used in a way that is not sympathetic to its potential to create bias.

Empathy is a process involving two parts. Firstly, cognitive empathy allows us to understand another person's context, emotions, goals, and motivations. Secondly, emotional or affective empathy is the ability to respond to those things with the appropriate emotions (Hodges and Myers, 2007; Esser, 2018). So empathy is rooted in subjective experience. Research shows that people generally feel more empathy for those closest to, and most like, themselves (Arbuckle et al., 2012; Bloom, 2016; Gutsell et al., 2010).

Our capacity for empathy can also be affected by scale. In other words, there is an adverse correlation between the size of a group and the extent to which we are empathetic towards them. (Slovic, 2007; Cameron, 2017). Cameron and Paine (2011) suggest that this is because empathy can be emotionally exhausting. Even when we engage with it in small doses, we expect the needs of large groups to be potentially overwhelming. This makes us regulate our emotions to prevent these feelings, thus diluting the effects of empathy. As a result, empathy can become more potent with individuals and magnifies when we can personally relate to them.

Differences in how we experience empathy can also depend on factors such as cognitive conditions (Decety & Moriguchi, 2007; Baron-Cohen et al., 2015; Bora et al., 2008; Tone & Tully, 2014) and genetics. Warrier et al. (2018) showed that our predisposition to be empathetic can be influenced by our biology. They found that as much as 10% of our empathetic potential can be determined by genetics. This study also showed that on average, women are more empathetic than men. This is believed to be based on non-genetic biological factors such as socialization.

UX research, empathy, and evidence

UX researchers have the potential to succumb to the influences of empathy in how we make decisions. We have the skills to control for bias when representing information. We adopt research methods to avoid skewing results, and where we cannot control for that, we report on the impact it could have. But researchers are not immune to the bias that can come from developing empathy for users. When we recruit participants into research, we see a name and a profile and we start to paint a picture. It is what our brains are best at: filling in the gaps. This increases when we connect with the person, have something in common and empathize with their story. Due to this connection, researchers could overemphasize the hardships of a certain user they relate to, wanting to solve their problems over those of other users.

UX researchers also have the responsibility for evangelizing and developing user empathy among stakeholders and teams. When teams care deeply about users, they are better equipped to build the right solutions, in the right way. UX research helps to bridge the gap between product teams and users by generating empathy strategically. Researchers hold an incredible amount of responsibility in how we collect, analyse, and disseminate user information to product teams. We leverage empathy by annotating our insights with things like video clips, photos, quotes, and storytelling. This creates emotive responses, which help teams care about users. We should be considerate of this responsibility and the way in which we facilitate empathy within teams and the impact that can have on how teams make decisions on which problems to solve. For this paper, we are going to be talking about how we make decisions, specifically with regard to solving user problems, and the role empathy plays in this process.

As a product team, we don't just use our level of emotional connection to users to decide how to act on our research, we also look to prove a problem is true before choosing to solve it. This is because time wasted on solving the wrong problem, or only symptoms of a problem, rather than the root cause can result in wasted resources. To establish a problem as 'true' we need evidence that accurately measures its impact from valid and reliable information. But like empathy, evidence is a complex concept. In a general sense, it is comprised of facts or information that indicates whether a belief or proposition is true. This makes evidence appear easy to judge, as it uses facts to define truth. However, this is complicated by views from prominent theorists that challenge the possibility for objectivity when interpretation is inherently biased. Donna Haraway (1991), for example, claims that all knowledge is 'situated' and that no perspective can be prefered over another. Our view of the world can cloud how we interpret and understand information, which impacts whether we believe something to be true or not.

This tells us that evidence can be taken out of context depending on the person interpreting it. In order to view evidence objectively, it must be tightly coupled with its original goal and the context in which the information was gathered. We see this in journalism often: a blindly called-out percentage or misused quote to reinforce a point. These are weak forms of evidence because they are deliberately taken out of context to influence belief about a subject in the desired way. This means that evidence is also problematic for decision making if not used responsibly.

Ultimately, an awareness of where knowledge comes from reinforces our approach to UX research. For one, demanding that evidence is robust in validating a problem means that relevant information must come from studies that rely on observable behaviours to bolster the conviction that further research is unlikely to change the degree of confidence we have in our findings (Hodgson, 2017). Furthermore, during evaluation, we must be aware of the ways in which our biases can limit the positive impact of our research. The very nature of the abductive reasoning process we use to develop products makes it unlikely we will have complete data informing our decisions when we gather evidence. As a result, we make decisions as we go, based on our current information, to constantly assess our level of certainty in the evidence we have. Empathy presents a particularly challenging component of this assessment because it can bias the way in which we interpret information as evidence and therefore our notion of what is important.

Since we can never be fully aware of a user's lived experience, recognizing bias means acknowledging that this can be a limitation of our research. A study carried out by Dan Ariely and colleagues (2012) demonstrates empathy and decision making in practice and informs our approach in understanding commerce experiences. It shows the difference between judgments made based on your own experience versus observing that of another. In this research, Ariely and colleagues studied the IKEA effect with origami. Builders made origami creations while buyers observed them. When complete, both were asked how much they would pay for the creation. Across the sample, buyers were willing to pay 5 times less than what builders would pay themselves. The study indicates that people assign more value to the things they put time into and build themselves. It is also an example of the differences in experience between employees at Shopify who observe merchants creating businesses, versus the entrepreneurs who are building and creating these businesses. We often believe that exposing ourselves to the lives of these users will enable us to make good decisions for them, but this study shows that this is a potential trap. Being an observer is not the same as being the user. We did not build their business ourselves and we are unlikely to make the same decisions for it as they would for themselves.

EMPATHY AT SHOPIFY

Today over 600,000 businesses in 175 countries use the Shopify platform. While many entrepreneurs seek the same goal of having a successful business, there is an increasing amount of differences in their needs, contexts, and motivations in pursuit of that goal. This diversity makes our mission statement 'make commerce better for everyone' a much bigger challenge for Shopify. As Shopify has scaled, grown, and gained success, we've begun to view this challenge as an opportunity to innovatively reimagine the commerce landscape and meet our mission statement in new ways.

With scale comes an abundance of information, and parsing all that information to make decisions can be challenging. When you add the effect empathy can have on our judgments, this could be problematic in making good decisions. This is especially noticeable when we don't use empathy responsibly. By admitting that we will never have adequate evidence to make objective decisions, we are actually getting closer to responding to our users' real needs. It can be argued that since we are so embedded in our particular contexts, all knowledge is a product of our environment. However, by questioning the methods by which we seek evidence, we can get closer to a more rigorous approach to creating better products. In this research paper, we look at the effect of empathy on decision making and discuss how to further develop the UX industry definition of how to use it in a more responsible way.

At Shopify, empathy is an ever-present theme. It is encoded in our company values, and there are ample opportunities for employees to develop it for the entrepreneurs who use the platform. It is positioned as a way to help us make good product decisions for our users by prioritising the achievement of their success as a way to ensure our own.

Examples of how developing empathy is encouraged at Shopify:

- New hires design and build a new online commerce store to better understand how the platform works and how it supports entrepreneurship.
- Employees can shadow support calls to hear firsthand the challenges and pain points of users.
- All employees are encouraged to regularly do on-site visits (i.e. in a Shopify user's retail location) to observe how they run their business and use Shopify.
- Communications and company-wide events always feature stories from users and the businesses they run.
- Quarterly internal hackathons are inspired by and revolve around solving particular user problems that are not currently being addressed.
- Common frustrations coming from users via customer support get automatically posted to team Slack channels.
- Team members of all disciplines are encouraged to observe and take notes during usability tests, validation, and interviews.

Further, Shopify embeds researchers and data scientists within product teams to use their expertise to ensure learnings and user insights are shared more readily. Research is integral to understanding needs and contexts, as well as ensuring products are developed in a way that solves real problems. As researchers, we focus on both generative and feature development research, depending on team needs and priorities. We are enabled to choose the best methodologies based on our research questions. We bring the team closer to the users they design for, rather than being a bottleneck to information. Everyone is involved in research and builds empathy through direct exposure to our users.

The opportunities for teams to build empathy for users come in many forms so that teams effectively understand user needs and their environments. However, as our user base grows, so too do the needs and contexts to support. Building understanding of our users is a constant necessity that will never be completely fulfilled. Empathy is something we continuously develop and we've created an ideal environment to facilitate this, but it's currently focused on gaining empathy generally and not necessarily how to apply it.

METHOD

Our method for investigating empathy and the ways it is applied in decision making at Shopify involved interviewing employees from across the organisation. We selected 13 participants based on their varying tenure and exposure to different projects. We also explored viewpoints from different disciplines—namely UX, Engineering, Data Science, Human Relations, and Culture—to assess the impact of empathy from 2 different angles: how we apply it to building products and how we instil it as a company ethos.

We interviewed each person one-on-one for 45 minutes about their experiences in product teams and how empathy had played a role. Those interviews were crafted around the following research questions:

- What are the definitions of empathy and evidence based on different individuals experiences and disciplines?
- What is the relationship between empathy, evidence, and decision making?
- How does empathy inform and influence their day-to-day?
- What are teams' decision-making processes in how they approach solving user problems? What impact does empathy have on this?
- What types of projects or initiatives cause empathy to be used successfully or unsuccessfully in decision making?

We also observed how and when our own team used empathy and evidence. Specifically, we reviewed project documentation, sat in project meetings, team conversations, discussions with stakeholders, show-and-tells, and UX team rituals. We wrote down the information that was used to make decisions or inform direction and how that information was presented. We conducted observations over a period of 6 weeks. For the first 4 weeks, we observed teams not connected to our own products in order to control for any bias we would have in being decision makers in the group. We collated the data and came to some thematic groupings for how empathy impacts decision making. We then spent the final 2 weeks using this as a framework for evaluating within our own team. One example of a UX team ritual we observed was called 'Angry Thursdays'. It is a half hour per week focused on user struggles. During this session, one of the UX researchers highlights user frustrations from recent research and/or support calls. The goal is for the team to regularly stay close to and maintain empathy for users. We observed how information was disseminated to the team and how the team used that to make decisions.

In addition, we documented when empathy was used at the company level as a way to engage and motivate employees. We observed broader UX events, such as Shopify's internal UX conference. Empathy was the subject of numerous talks. One talk's key message was that empathy is a secret superpower and will allow you to have exponential impact through your work. At the company level, we looked for use of the term empathy, a focus on highlighting our users in some way, and the method by which the information was presented.

From all our observations and interviews, we collated the findings and analysed them by impact and effectiveness on decision-making. We noted the ways that empathy had not been sympathetic to its potential bias based on our original research, such as empathy being treated as a singular characteristic. This brought us to 4 thematic traps.

FINDINGS

Product decision-making process

Our findings uncovered a structure in the way in which product teams, in theory, make decisions on which problems to solve and how. We modelled this process to contextualise our findings for the impact of empathy. The product decision-making process has three stages: problem, evidence, and decision. We found empathy traps manifest in some or all of these stages.

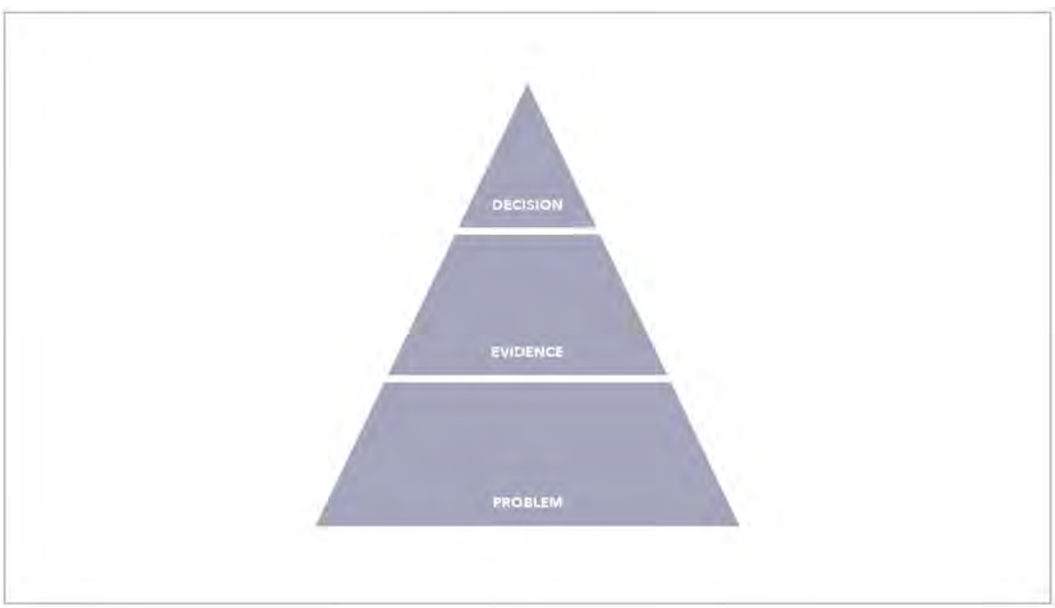


Figure 1. Product decision-making process

Working out which problems to solve and how to prioritize is a complex process that has to take into account many different factors. The foundation of decision-making is the problem itself. The 'problem' here refers to a hypothesis we have about what we are trying to solve, which is based on any of the information we have access to. It is a hypothesis because we don't necessarily know the details of the problem and therefore whether it is true. Before we can make good decisions, we must validate the problem through evidence.

The second stage of the process is evidence. Information becomes evidence when it answers our question or validates the hypothesis about a problem. At Shopify, evidence could be derived from qualitative or quantitative information from research, data science, and/or customer support, but not exclusively from these sources. Evidence can come from information that is either already collected, or it can be gathered and generated within projects. If the information contradicts the hypothesis, this likely means that, for the current belief about the problem, we don't have enough information, the problem is different than we expected, or it turns out not to be a problem at all.

The top stage of the process is the decision. The decision relates to what we will do about the problem now that we have evidence to validate that it is true and assess its impact. While this process is not an exhaustive account of decision-making, it is important to remember that, in addition to evidence; opinions, expertise, business goals, and technical constraints should come into consideration in making decisions about a problem.

4 empathy traps in decision making

Our success in building empathy at Shopify has brought us a unique challenge. While empathy's ubiquity in how product teams approach solving problems is encouraging, through our research we uncovered 4 traps that can happen in decision-making processes when empathy has been commodified: creating fake empathy, unbalanced use of empathy, using empathy to force decisions, and superficial empathy for show. For each, we have visualised the trap's problematic relationship between empathy and the decision-making process to highlight where it presents challenges.

Trap 1: Creating fake empathy – Creating fake empathy is a trap in how empathy manifests itself in the problem stage of decision making. This trap is prevalent in the problem stage when someone experiences a problem through created experience and then applies the experience as an understanding of the user's perspective. But because an individual experience is fundamentally different from person to person, each individual will have a unique perspective on the same situation. Ignoring this trap could mislead our understanding of a problem or misidentify a problem where none actually exists.

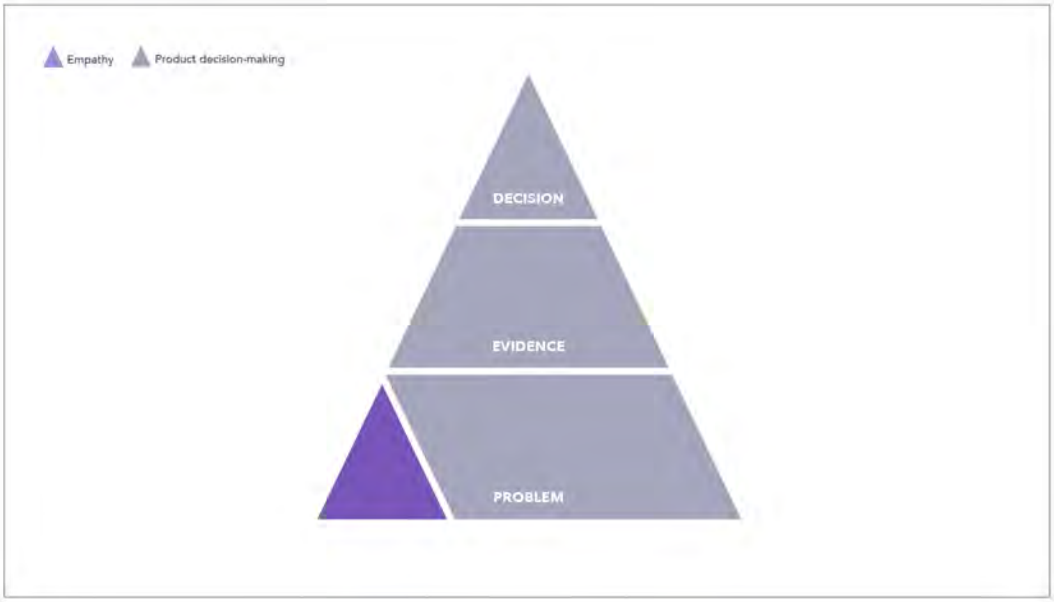


Figure 2. Trap 1: Creating fake empathy

One example of this trap was observed through an initiative encouraging employees to be entrepreneurial and start their own businesses using Shopify. As is common with companies of all kinds, Shopify wants its employees to 'dog-food' the product. This is the practice of using the product in order to experience how it works. While not being exclusively targeted at building empathy, according to our participants it was an impactful by-product. But employees have different information and motivations than the entrepreneurs who use our platform.

> "Although well-meaning, it misses the mark. The stakes aren't real when you have a steady paycheck coming in, when you have access to more information or resources than the general public, and there are no real consequences if your business doesn't succeed. That's not really realistic because the context is so different." Data Scientist

Unlike our user base, Shopify employees are not relying on their business to support their livelihoods. Employees also have extra context from knowing the platform deeply and that information impacts their approach to using it and how they go about starting a business.

> "Someone on my team runs a store and will use themselves as the archetype of who we're building for. Problem is, they have so much tech and product context that users who don't work at Shopify don't, that they are a bad template for users. Their context is not the same." Developer

This program was an incredibly positive exercise for building product understanding along with identifying the kinds of challenges users might be up against. But it does not make us entrepreneurs in the same way our users are. In fact, we found the pain points that employees experienced were often spotlighted, creating a bias for seeing and assigning more weight when those things arose in research and/or customer support data.

> "The idea of needing to put ourselves in the shoes of users is not the best illustration of what empathy is. Kids playing with aluminum foil pretending to be an astronaut doesn't make them understand what it's like to actually be an astronaut. Pretending to be entrepreneurs is projecting our own biases in that scenario. Fake empathy makes you make bad decisions [based] on assumptions and stereotypical emotions." Design Lead

Context is an important part of an experience causing emotions to be subjective. Projecting our own experiences as a true reflection of those of our users will incur inaccuracies from this subjectivity, rendering the 'empathy' we feel fake, as it's based on a personal and biased view. We could miss a key detail or highlight something as important to us that isn't as important to our users. There is not enough information from these pretend scenarios alone to translate those feelings into good product decisions.

Creating fake empathy conflates information from one's own experience with the experience of actual users, mistaking it for a true representation of a user problem. Misguided decisions can come from the belief that, because the problem experienced is the same on the surface, the feelings are also the same, and therefore judgments made will be equivalent.

Trap 2: Unbalanced use of empathy – Unbalanced use of empathy happens people put more weight on information that elicits a strong empathetic response over other pieces of information. This trap causes every step of the product decision-making process to be overshadowed by, and optimized for, the experience of select users or more salient parts of their experience. It is the challenge when encoding the information too quickly goes from experiencing cognitive empathy (understanding) to emotional empathy (assigning emotion to that understanding). Emotional empathy clouds judgements by spotlighting the experience due to a bias created by increased emotion. It tricks people into believing some information is more important than other pieces of information.

One participant in our study sample gave an example of this trap that they experienced when their team was developing a filters feature. In research, one user had demonstrated frustration in not knowing where to begin in selecting and applying filters. The team ended up overusing this qualitative account in deciding how to develop the feature, without considering the size or nuances of this problem throughout the population. In subsequent testing with a larger group, they found the proposed solution actually created friction for most users. When making the decision about what to solve and why, the team misinterpreted the problem by putting too much emphasis on the empathetic impact of one story and optimized for removing a negative experience of one individual.

> "People will take anecdotal evidence and assume it's the full picture. We're only starting to learn to build user empathy, but not necessarily what to do with it." (Developer)

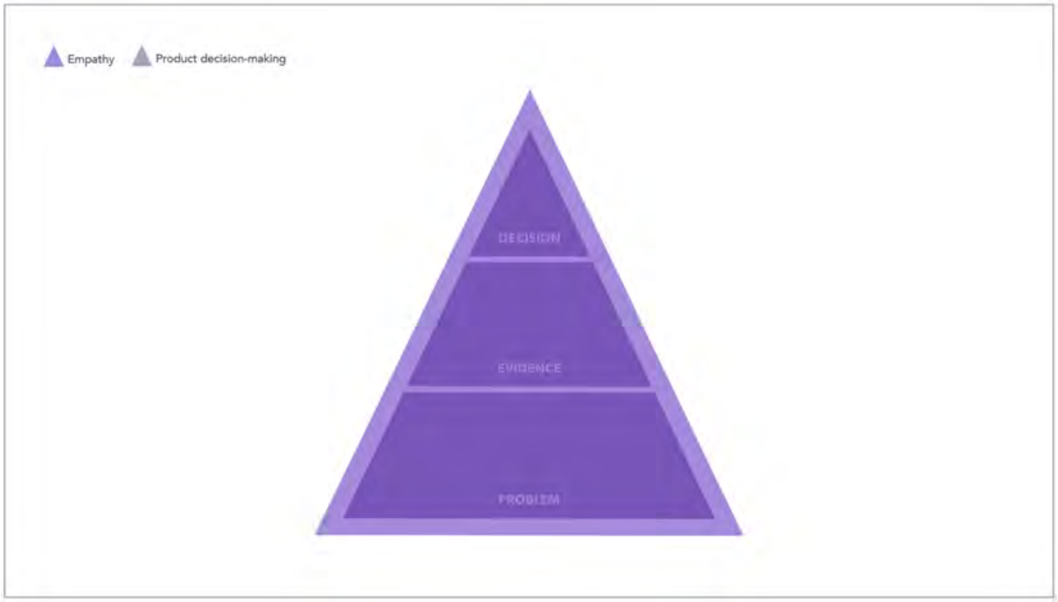


Figure 3. Trap 2: Unbalanced use of empathy

Another example was when a product team made the decision to redesign the workflow of a key section of the product. The solution at the time worked well for a specific segment of users, but it did not scale to the majority of use cases. The team used empathy for that smaller group and decided not to take action for a length of time. The decision to redesign the functionality eventually came from prioritising the needs of the larger group, even though it would be a significant adjustment for the users it currently served well.

Unbalanced use of empathy can also show up deeper in the details of a problem when people rely on salient parts of information without enough context. This happens when teams select certain aspects of information and make decisions based on a misunderstanding of its significance. People have more empathy for the parts of the experience on the fringes because they stand out and more easily create an emotive response. People, therefore, find it harder to connect with issues that fall in the middle, which are more nuanced and difficult to categorise clearly. This overemphasis on salience can lead to acting on things with a biased understanding.

One participant gave us a project example related to an older version of the first page users see when they log in to Shopify (a content feed of information related to a user's business). Shallow understanding of significant business milestones, such as getting to a first sale, meant celebratory messages were created in abundance and shown when such milestones were reached. This was problematic because, despite the milestones being achievements, the messages themselves lacked the appropriate language to recognize that reaching them can also be incredibly stressful. These milestones did not always signify celebration in the minds of users. These messages were based on a conclusion drawn from the salience of highly empathetic pieces of information related to those milestones. While this page still celebrates these moments with users, due to gathering a deeper understanding of these experiences from further pieces of research, as a company we have adjusted our tone to relate more to the context of the milestone.

> "We almost glorify empathy as the way to represent our own assumptions of a merchant's success without understanding the full perspective or considering all the factors. We built a part of our product as a celebration screen, highlighting how much revenue a merchant has made. But what if it's zero dollars all the time and they've put their life savings into this?" Design Lead

We also observed this misuse of empathy in our internal communication channels. Shopify is an open and collaborative company, so any employee can highlight a problem with our product to the relevant team. Along with this, customer support can ask product team members real-time questions coming directly from users via these channels to get help finding answers and troubleshooting. Although it is a great way to provide support and keep teams close to user problems, it also creates salience in the team's mind because they know this is a real user struggling with this problem right now. As a result, there is a prolonged effect of wanting to prioritize and solve symptomatic problems over those that are systemic, and sometimes more impactful.

> "I've definitely seen people amplify a human frustration they've witnessed, even if the data says otherwise. And I've also seen projects that experiment data said should be shut down persist because there were a handful of stories about how users 'really like it.'" UX Researcher

Spotlighting specific experiences based on empathy leads to information being insufficiently framed by its context. This can cause short-term misguided decisions. Due to the level of empathy salient information can create, teams are more motivated to fix these problems. However, that motivation should be channelled into gathering context and deeply understanding the complexity of the problem to be solved. From our examples, we observed the impact of empathy in this trap was to actually cause lower empathy overall because decisions were often lacking a true understanding of users' full experiences.

We found that this trap was less prevalent for participants with longer tenure. They tended to be less reactive to this kind of information and wanted to seek further context by building up a deep understanding of the complexities of running a business.

***Trap 3: Using empathy to force decisions* –** We observed individuals leveraging empathy in team decision-making scenarios, such as team project alignment sessions and reviews, to influence one decision over another. It is important to note that this is not a malicious activity, rather, these team members cared and wanted to improve the platform for users. This trap builds on the previous trap, unbalanced use of empathy, as it is often those singular, more salient experiences being used as evidence and causing too much influence at the point of making a decision. This trap can actively stunt other individuals' contributions and cause important viewpoints and information to be left out.

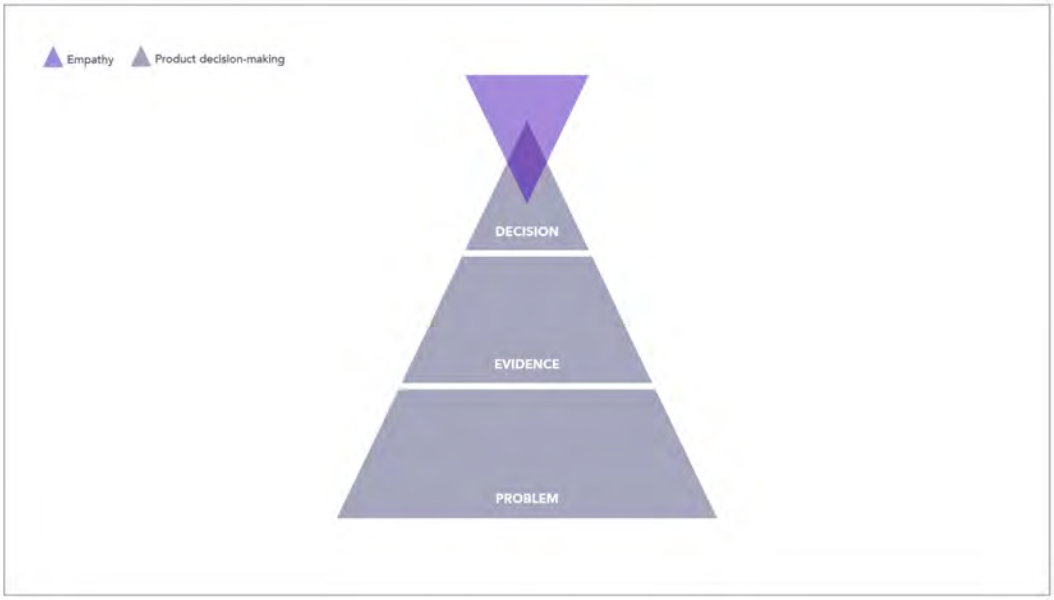


Figure 4. Trap 3: Using empathy to force decisions

> "In the absence of enough information, everyone says things like 'as a user, I would do x' to get what they want, or talk about one user's experience to represent most of our user base, which actually shows a lack of empathy." Developer

An example of this can occur after employees do an on-site visit to learn about a user's business and how they use Shopify. Employees will often come back from these unique visits deeply motivated by empathy to address a specific problem they just witnessed. They recount their observations to the team to inspire more empathy and motivate a desire for action. This pressure can also manifest itself later in prioritising problems, as references back to these visits to influence decisions. But the information obtained should be understood and treated as general findings that broaden the team's worldview or as a starting point for additional research, not as evidence on which to base decisions.

To complicate things even more, if a team member argues against information that is underpinned by empathy, especially when it relates to the success or failure of a user's business, that person can appear unfeeling. In seeking to avoid this, individuals feel discouraged from contributing to a conversation where empathy is being used to force or influence a decision. This happens even though everyone wants the same outcome: building the best product experience and making good decisions on how to get there. However, if individuals fail to contribute other viewpoints, this potentially important information will be left out of the decision-making process.

> "Empathy is a good way to shine a light on what you know nothing about, makes you aware of what you don't know, and can debunk our assumptions about situational use cases. Empathy is good for awareness, but definitely not for decision making." Developer

Another example of this trap is when employees have decided on an idea and implicitly use it as a frame of reference for seeking evidence to support it. This behaviour has been observed when colleagues reached out to UX researchers for specific user quotes and other insights at the final push to get things started or committed to. The information requested is often more emotionally influential, rather than broad and quantitative.

Empathy should not be a singular driving force of decision making. Using empathy to force decisions flips the product decision-making process upside down to be about the decision coming first. As a result, many of the opportunities to understand the problem, along with gathering the evidence required to support good decisions, can be missed.

Trap 4: Superficial empathy for show – Superficial empathy for show happens when it is popularised without a deep understanding of how to gain and apply it. Having empathy becomes a catch-all concept that you can check off once you have 'done empathy'. There is good intention behind this, but it's based on a simplified definition of gaining empathy. It can quickly make empathy a vanity metric for 'good product development'. Teams end up doing it because they *should* care, not because they necessarily *do care*. Empathy becomes the act of doing 'research' and doing 'research' in its practice alone is seen as 'having empathy'.

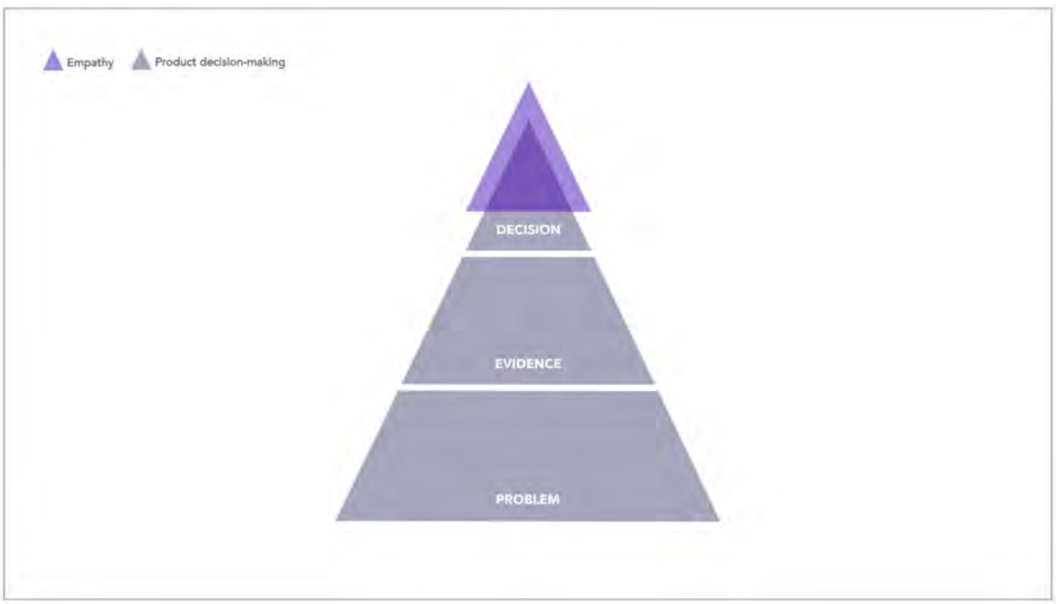


Figure 5. Trap 4: Superficial empathy for show

> "Empathy can be used as a checklist, which is not authentic. People can grab a couple of things and think 'yes, I did the empathy part of this'. But what if you grab the wrong items? Empathy, after all, is not about finding one use case. It's about taking cues from multiple experiences." Developer

As an example, we often heard things like, 'Have you spoken to a user about this?' after conclusions had already been reached. This act suggested that having empathy was in seeking a user's seal of approval and was enough evidence to proceed with a decision. This way of speaking about research as a proxy for having empathy made it a separate entity and something that was disconnected from the process leading to decisions.

We also saw examples of this trap through confirmation bias in how people sought evidence: making a decision and then validating it with evidence after the fact. This is seeking a problem for a solution and has too much bias in how the information is filtered and framed.

> "Empathy is a term that's thrown around. We've done testing and talked to users, so we have empathy, but it almost feels like we're already working towards that. Is that empathy if we're already heavily leaning towards a decision and not looking at things critically?" Design Lead

The impact of this trap on decision making is that it isolates empathy as an item on a checklist to complete separately. Using empathy as an amendment to a decision already made creates bias in how we seek and interpret information, which misleads teams to conclude that information is evidence when it isn't.

Summary of empathy traps

The challenge with all these traps is not the intent to seek and have empathy, but how we interpret the feeling, assign it value, and translate it into decisions. It can be intense enough that it becomes evidence in people's minds, creates weight, and biases decisions. It then inhibits careful interpretation and objective balancing of all the information available.

Here is a summary of the four empathy traps we identified:

1) Creating fake empathy by believing we feel the same as our users when engaging in empathy-building exercises. It happens when we transfer our experiences as accurate representations of the experiences our users feel. Direct decisions on this information lack context and are skewed to our beliefs.
2) Unbalanced use of empathy by over-indexing on it, like spotlighting specific experiences and assigning more weight to some information over others. It happens when we use shallow but salient information based on the empathy it evokes. This can cause short-term and misguided decisions without enough context.
3) Using empathy to force decision-making can stunt others' contributions and cause the omission of other important perspectives. People leverage empathy to elicit an emotional response in others to sway the outcome of decision making in a specific direction.
4) Superficial empathy for show is a simplistic interpretation of empathy as an item on a checklist. It isolates empathy and causes it to be an afterthought. Decisions are often already made and empathy becomes a confirmation bias activity to justify that decision.

In UX at Shopify, empathy is about caring for users and wanting to develop solutions to enable them to run a successful business. This is a big responsibility which, in and of itself, can evoke emotion among our teams. Shopify is dealing with entrepreneurs' livelihoods and therefore the decisions made can have serious consequences. Good decisions are made from synthesizing various pieces of information and viewpoints. Pieces of information that create empathy are valuable, but cannot be used in isolation.

DISCUSSION

Discovering these 4 traps is not to discount the role of empathy in how we work, it is a powerful tool in getting teams to care about users and is a vital first step toward attentive action.

***Empathy is not evidence, it is a driver to seek evidence.* –** In industry settings, empathy can successfully impact a variety of product decisions, starting as an effective motivator for teams to care about users. Throughout our examples, it was this motivation that drove action. But channelling this motivation is crucial. Rather than acting on our emotions, empathy is more beneficial if we harness the drive it creates in order to learn more about our users and their needs.

Empathy provides a strong signal, revealing the need for more information and a deeper context. With empathy for our users, it can give us a glimpse into their point of view when we do not have much information. This works on the basis that we do not trick ourselves into believing that shallow information is enough, or that empathy is an item on a checklist and separate from our responsibilities.

This signal often presents itself as a moral feeling. When someone presents an idea that you feel uneasy or unsure about but don't necessarily know why, that is likely empathy at play, giving you an implicit sense that something doesn't feel right. Use this feeling to go and gather more information, and be empathetic to the fact that we might not have enough understanding yet.

For empathy to act as a signal to gather more context, we first need to build a deep understanding of our users. We saw in our sample that it was the participants who had been with the company the longest who had this sense, but they also pointed out that we cannot get complacent and think we know everything. This is a continuous process to keep gaining empathy and adapting the way we approach using empathy, especially as we scale.

Empathy Decision Model

In order to understand the right relationship between empathy and decisions, we created the Empathy Decision Model. The mapped relationship in this model shows that empathy is a layer on top of each stage of the decision-making process, with more influence at the bottom, and less as you move to the top. As a layer, it is also no longer isolated as an afterthought but is instead a requisite part of each stage.

The main role of empathy is that it is a signal to seek more information and helps us be critical about what we believe to be true. At the problem stage, it should be highly influential to encourage us to dig further into what we know about a problem, our users, and their context. It should hold a mirror up to our understanding and make us critically reflect on any assumptions we might have and/or our lack of information. It drives a conscientious approach that stops us from pretending we know and moves us to an informed place where we know as much as possible. Drawing assumptions about our users from not enough information can incur poor decisions, which is ultimately not an empathetic act.

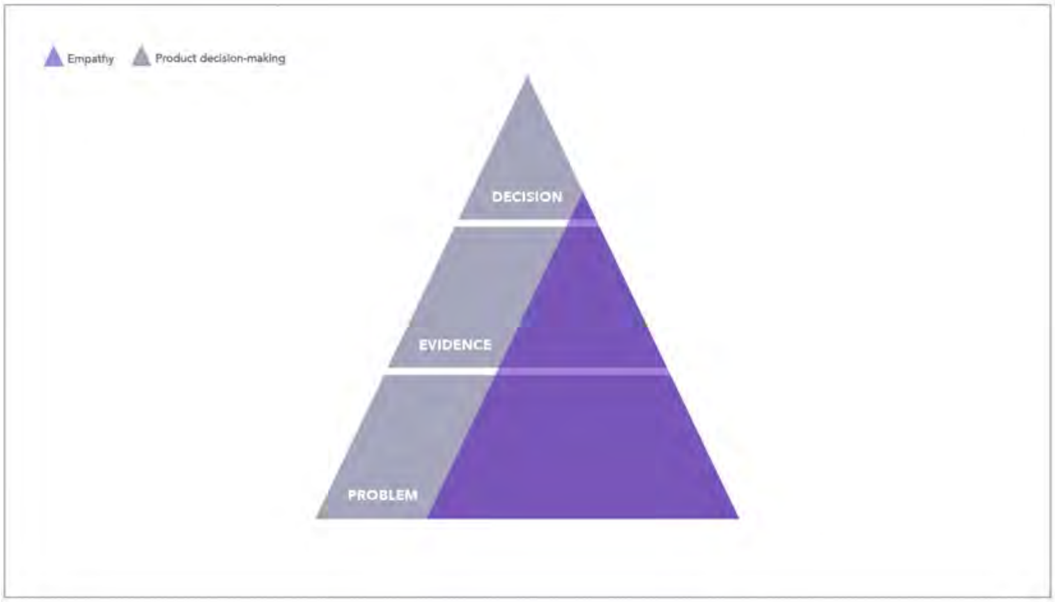


Figure 5. Empathy Decision Model

At the evidence stage, it provides a critical lens through which to evaluate the information we have as evidence of the problem, helping us determine if it reliably represents aspects of the problem or our users' context related to that problem. The best way to do this is to review information through a research question or hypothesis as a counter to any potential bias you might have in what you look for and interpret as evidence.

When making a decision, empathy is less directly influential. At this stage, it should have been contextualised, rationalised, and carefully considered in the evidence we have. It is key we no longer go directly from empathy to decisions. For good decisions, we need to have empathy, but at a level of abstraction. Empathy never goes away, so when decisions are made, it is implicit because it has been baked in throughout the process.

Empathy at the decision stage can also play an explicit role as a final check in decision making. It likely shows up as a doubtful feeling about a decision. This happens when empathy from your deep understanding triggers your brain to tell you to stop, take a step back, and review what we know; that we should seek more information and critically evaluate the evidence. Ultimately, having 100% perfect information, and therefore evidence, is improbable. Most of the time decisions come from the best information we have at the time, which is weighed against other details like technical constraints, business goals, effort, time, and resources.

Weighing a decision based on risk is a critical step. We assess risk based on the potential impact of a decision and how confident we are with the information we have. Some decisions are riskier than others depending on how much of a negative impact the decision could have. Risk is a continuous scale, and even mistakes let us learn and adjust. As we learn more, the potential to review and adjust our decisions will increase.

Following up on your decisions is where empathy also plays a role. It should drive the need to measure and follow up on our decision to ensure it is moving us towards our goal.

This follow-up is also useful for our model. It feeds our problem stage with more information, and the cycle continues.

How we apply the Empathy Decision Model

***Be mindful of how you collect and connect the team with evidence* –** As UX practitioners, we play an integral part within the bottom two stages of our model: identifying a problem, and determining if the evidence supports it. Therefore, we have an incredible amount of responsibility both in how we carry out our research and how we disseminate insights to the team. UX researchers should use empathy to seek a deeper understanding and clearly identify when we know enough to make good decisions. At Shopify, we also have projects which we call 'loose threads'. Loose threads are issues, needs, or problems that are repeatedly found through various pieces of research that can appear small at first because they are not necessarily specific to our research questions. They generally spark a feeling of empathy and a concern that they could be indicators of something important that warrants further research. We have the autonomy to channel our empathy and embark on these projects to gather more context for better understanding of the issue. Some of these loose threads have turned into high impact projects at Shopify, or at the very least created implications within existing projects.

Building empathy in teams comes with the duty to manage how that empathy manifests itself in the process of product development. At Shopify, we have found decoupling decision making from the dissemination of findings to be helpful in allowing product teams to go away and reflect on insights. We then review how the findings fit into what we know to mitigate the risk of making rash decisions based on empathy. This distinction means decisions will be made only after considering all the information we have and how much weight to assign to it.

Another responsibility is framing information in order to make sure it isn't taken out of context after a decision is made. Evidence is based on using information that effectively measures the existence and impact of a problem, and this framing is also important for how this evidence might be referenced later. Commonly, this framing is the way in which the information was gathered. Any discipline that seeks and synthesises information has a responsibility not to let that information be misused or taken out of context, therefore maintaining the frame in which it is relevant. To keep the integrity of the evidence and its context, we encourage our UX practitioners when presented with 'evidence' to ask questions like 'what information is that based on?', 'how was that information collected?', and 'how do you know that to be true?'.

***Where possible, use multiple pieces of evidence to make good decisions* –** At Shopify, we've found that successful decisions are those that incorporate various pieces of evidence elicited from different types of information. As facilitators of information, UX researchers, Data Scientists, and Customer Support Data Analysts work together to answer questions. Not all information holds the same weight, but where possible, we seek multiple sources of information.

This collaborative approach has provided us with the depth of information we need to make good decisions, along with useful decision-making criteria when a clear decision is not present. For example, a quantitative experimental approach may result in no significant differences between the current state and the changed experimental state of a user interface (UI). We review this finding with other information from research and customer support to decide whether making the change is qualitatively a better experience. If so, we will make that decision, because it is the best decision, taking into consideration all the evidence we have. Working together means having the opportunity to rationalise one set of information with another.

***Embrace the impact of empathy* –** We are predisposed to feel different levels of empathy and are conditioned to feel empathy for different things at different degrees, and that's okay. Work with empathy, not against it. What we should stop doing is expecting empathy to be absolute or objective. It is a bias to think we do not have bias.

As our audience grows, so should our approach to understanding it. One way we have embraced the impact of empathy more successfully is by being intentional about the diversity of our teams—both through hiring and staffing projects. We think about diversity of experience, background, culture, and other dimensions. As the organisation scales, we endeavour to offer more solutions to a growing group of users. It will, therefore, be harder to stay close to all users, contexts, and needs because the effort to gather information about them multiplies. Having a more diverse team doesn't replace research practices but it increases the chances of including wider viewpoints that should be considered when designing for users.

Another consideration is to put safeguards in place to check in on our biases. In knowing that bias exists, we can influence how we think about information by having tools that sense check our reactions to that information. A method developed by one of our UX researchers was a decision tree that people can use after visiting a business.

This decision tree gives our teams a route to think more deeply about what they observed and stops them going directly from empathy to a decision. It does not try to control or dismiss empathy, rather it works with it to rationalise what has been observed while gathering information.

NEXT STEPS

This is an internal view from our company of the traps that can occur in the decision-making process when empathy is commodified. How our company has developed with empathy at the core of the business is overall beneficial for the work we do, so this is a good problem to have.

If your organisation does not have enough empathy for users, that is the first step. When you have empathy, it must be used at the right level of abstraction in the decision-making process. This is our view, and we expect you, as UX practitioners, to take the model away and gather experiences from inside your own environment. You can then develop, pivot, and reinforce our understanding of empathy in order to use it effectively and responsibly in your organisation.

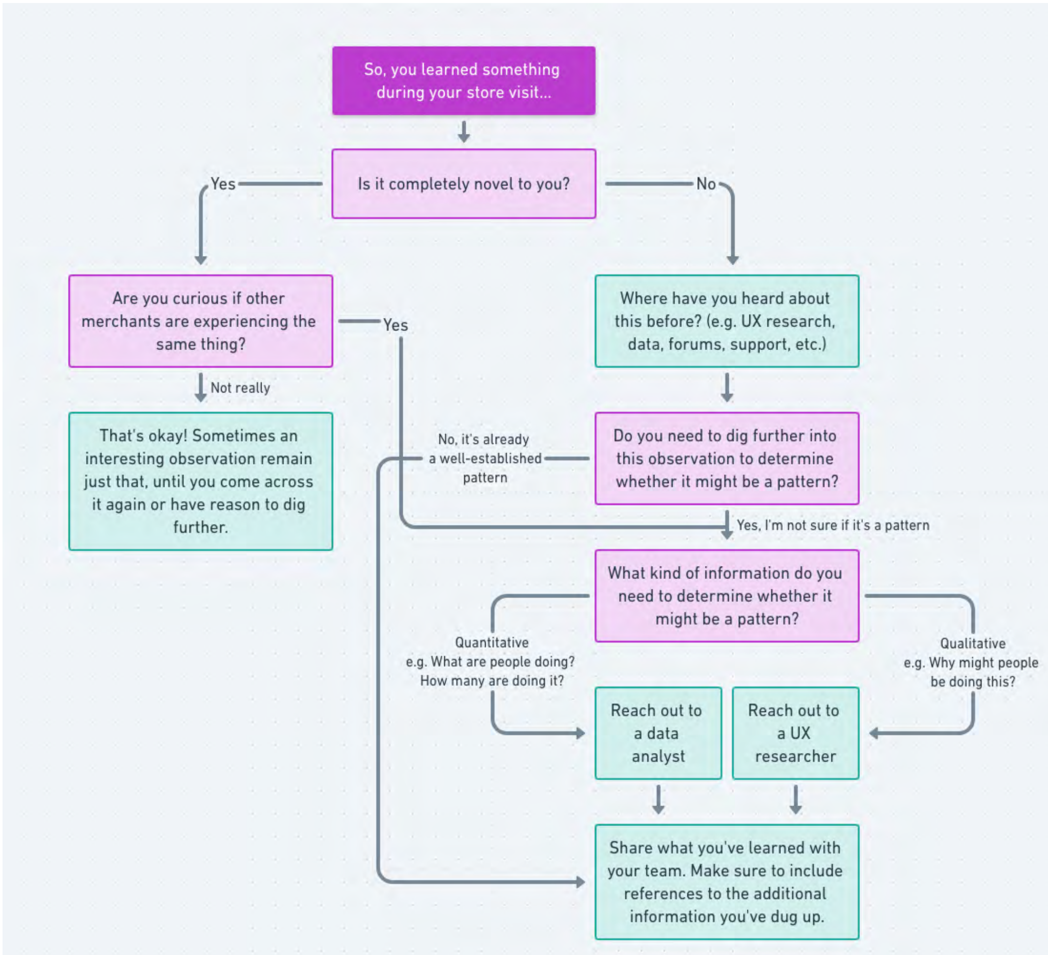


Figure 6. Shopify business visit decision tree.

Rachel Robertson is a UX Lead at Shopify focused on creating quality product experiences at the intersection of technology and commerce. She has over 10 years of experience in communication design, software product design, and UX management. She is passionate about building inclusive teams, working together to deeply understanding users' needs and solving complex design problems. rachel.robertson@shopify.com

Penny Allen is a Senior UX Research Lead at Shopify. She has been a Researcher in product for over 10 years working for various companies including the BBC and The Guardian. With a background in psychology, she's always been motivated to learn deeply about users and work on products that solve real human problems. penny.allen@shopify.com

NOTES

We would like to thank Alëna Iouguina, Ariel Chernin, Austin Dains, Ben Watts, Cassie Kaiser, Cynthia Savard Saucier, Dalia El-Shimy, Dave Newton, David Lennie, Emma Craig, Henry Hammond, Ian Mortier, James Bull, Jen Chow, Michael Patten, and Serena Ngai (Shopify) for their contributions to this paper. With special thanks to Sarah Ebbs, Rebecca Pasiak, and Hilary Dwyer for helping us edit this paper.

REFERENCES CITED

Arbuckle, N. L. & Cunningham, W. A.
2012 Understanding Everyday Psychopathy: Shared Group Identity Leads to Increased Concern for Others among Undergraduates Higher in Psychopathy. Social Cognition: Vol. 30, No. 5. pp. 564-583.

Baron-Cohen S., Bowen D. C., Holt R. J., Allison C., Auyeung B. & Lombardo M. V.
2015 The 'Reading the Mind in the Eyes' test: complete absence of typical sex difference in ~400 men and women with autism. PLoS One; 10: e0136521.

Bloom, P
2016 Against Empathy: The Case for Rational Compassion. Chapter 3: Doing Good.

Bora E., Gökçen S. & Veznedaroglu B.
2008 Empathic abilities in people with schizophrenia. Psychiatry Res; 160: 23–29.

Bose D., Segui-Gomez, M. & Crandall, J. R.
2011 Vulnerability of Female Drivers Involved in Motor Vehicle Crashes: An Analysis of US Population at Risk. Am J Public Health. 2011 December; 101(12): 2368–2373.

Cameron, C. D.
2017 Compassion collapse: Why we are numb to numbers. In E. M. Seppälä, E. Simon-Thomas, S. L. Brown, M. C. Worline, C. D. Cameron, & J. R. Doty (Eds.), Oxford library of psychology. The Oxford handbook of compassion science (pp. 261-271). New York, NY, US: Oxford University Press.

Cameron, C. D., & Payne, B. K.
2011 Escaping affect: how motivated emotion regulation creates insensitivity to mass suffering. Journal of Personality and Social Psychology, 100, 1–15.

Decety J & Moriguchi Y.
2007 The empathic brain and its dysfunction in psychiatric populations: implications for intervention across different clinical conditions. Biopsychosoc Med; 1: 22.

Eisenberg N, Eggum ND.
2009 Empathic responding: sympathy and personal distress. In: Decety J, Ickes W, editors. The Social Neuroscience of Empathy. Cambridge: MIT Press; 2009. pp. 71–83.

Esser, P.
2018 https://www.interaction-design.org/literature/article/empathy-how-to-improve-your-designs-by-deveoping-empathy-for-your-target-group

Foucault, M.
2002 The Order of Things: An Archaeology of the Human Sciences. New York: Routledge.

Gutsell, J. N. & Inzlicht, M.
2010 Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups. Journal of Experimental Social Psychology Vol. 46, Issue 5, Pages 841-845

Haraway, D. J.
1991 Animal Sociology and a Natural Economy of the Body Politic: A Political Physiology of Dominance. In: Simians, Cyborgs, and Women: the Reinvention of Nature: Routledge, pp. 7–20.

Hodges, S. D. & Myers, M.l W.
2007 Empathy. In R. F. Baumeister and K. D. Vohs (Eds.), Encyclopedia of social psychology (pp. 296-298). Thousand Oaks, CA: Sage.

Hodgson, P.
2017 https://www.userfocus.co.uk/articles/strength-of-evidence.html

Nortona, M., Mochon, D. & Ariely, D.
2012 The IKEA effect: When labor leads to love, Journal of Consumer Psychology: Vol. 22 (3): 453-460.

Slovic, P.
2007 "If I look at the mass I will never act": Psychic numbing and genocide. Judgment and Decision Making, 2, 79-95.

Tone, E. B. & Tully, E.C.
2014 Empathy as a 'risky strength': a multilevel examination of empathy and risk for internalizing disorders. Dev Psychopathol; 26: 1547–1565.

Warrier V., Toro R., Chakrabarti, B., Børglum, A., Grove, J., Hinds, D., Bourgeron, T., & Baron-Cohen, S.
2018 Genome-wide analyses of self-reported empathy: correlations with autism, schizophrenia, and anorexia nervosa. Translational Psychiatry: Vol 8, 35.

Map Making: Mobilizing Local Knowledge and Fostering Collaboration

NORA MORALES
UAM Cuajimalpa, México

SALOMON GONZALEZ
UAM Cuajimalpa, México

Participatory mapping – the production of maps in a collective way – is a common activity used for planning and decision making in urban studies. It started as a way to empower men and women, usually from rural vulnerable communities threatened by climate change, degradation of their landfills or any other conflict related to access to their land. It has been considered a fundamental instrument to help marginal groups represent and communicate their needs within the territory and augment their capacity to protect their rights. (FIDA, 2011). Why is it that in some cases participatory mapping works and in others fails? Why do these initiatives not trigger local action? Or even end up being counterproductive, when authorities use the map made by locals, to validate their points, causing conflict instead of negotiation?

As a research team of designers and social scientists involved in the creation of participatory mapping workshops, our goal was to analyze the process and resources and different outcomes of some participatory cartographic projects, including one developed by us for three small communities of the original settlements of the West Mountain Region of Mexico City.

Our findings outline three main principles to consider when pursuing a community mapping project whether using low-end or state of the art technology, in order to involve a community, validate their knowledge acquired from the mapping practice, and foster collaboration and organized action.

MAPS AS PRACTICE INSTEAD OF A REPRESENTATION

In the western world, we are very used to maps. We use them all the time: they help us get from one place to another; they help us find our way through unfamiliar terrain, locate people, objects and events. They have helped us plan, predict and understand our world. What if maps are more than what cartography has been telling us they are?

Historically we think of maps from a certain perspective. This maintains the hierarchical attitude about who makes the map, who owns it, and the power that comes with. This attitude delegates the power of the map to an elite group of experts, who are in charge to capture and portray spatial data accurately. By default, these structures reinforce the power of hegemonic institutions over local people, lands and other resources.

We often forget that the conventional appearance of maps mirrors social and political decisions and contain a bias: they express a point of view and indicate property lines, postal districts and enterprise zones. They are always biased in the sense that "*they project the interests of their creators*," as Wood states in "The Power of Maps" (Wood and Fels 1992).
Maps are not ontologically secure representations nor neutral products of science, they carry a long tradition of conventions and principles that reinforces the dichotomy of the cartographer as a skilled professional who make judgments that privileged discourses and

relies on the map as a container of the "truth", subjugating other kinds of knowledge. And then we have the user or interpreter of the map who is responsible for its depiction by his limited skills and knowledge.

If you think about it, maps are never fully formed, nor complete. They are more like living documents, transitory and fleeting. They are contingent, relational and highly context-dependent. A map is about spatial practices enacted to solve relational problems.

Take for instance this hypothetical example:

> Imagine a group of social scientists that have been given the task by a government agency of reporting on the distribution of population in informal settlements in Mexico City between 2010 and 2018. Given the spatial nature of the problem, producing what is commonly understood to be a map provides one viable solution over a variety of potential solutions to this problem. They have to construct a spatial representation using available data that conform to the agreed standards and conventions and which effectively communicates the pattern of population change.
>
> Starting from a position of having specialized tools and resources, and certain degree of knowledge, experience and skills, they work toward the process of mapping. The map *emerges* through a set of iterative practices of employing certain techniques that built on other's previous works or standardized forms of representation. This process is choreographed to a certain degree, shaped by the scientific culture of conventions, standards, rules, techniques and philosophies but is not determined and essential. The map is contingent and relational in its production through the decisions made by the team with respect to what attributes are mapped, their classification, the scale, the orientation, the colour scheme, labelling, intended message, and so on. The fact that the construction is enacted through affective, reflexive, habitual practices that remains outside cognitive reflection. The team "plays" with the possibilities of how the map will become, they experiment with different colour schemes, different forms of classification, and differing scales to map the same data. Making maps then is inherently creative and maps emerge in process.
>
> While all this decisions and actions might seem trivial, this culmination of a set of practices– creates a spatial representation that they understand as a map and believe that others will accept as a workable map based upon their knowledge and experience of what constitutes a map. Finally, when this spatial representation that the group understands as a map is printed to show to the government agency required, we would argue that their creation is not complete, although it has the appearance of what Bruno Latour (1986) calls an "immutable mobile" with its knowledge and message fixed and portable, and it can be read by anyone understanding maps language, it remains mutable, remade every time it is employed. Their creation is not ontologically secure as a map because it is being transformed by the inhabitants living those places that continue to grow in those settlements, or the state worker that would take a decision regarding a new policy based on that data. Individuals transform the spatial representation created by the team, into a map. Each person engaging with a spatial representation brings a different map into being, framed by their individual's knowledge, skills and spatial experience. For someone who does not understand the concept of thematic mapping or classification schemes, again the map will be bought into being differently to people who do, who will ask different questions of the data and how it is displayed. There is variability in the ability of people to mobilize the representation and to solve particular problems. Moreover, the recognition of the map generates a new, imaginative geography for each person.

In this paper we share how embracing this point of view of maps as processual instead of representational devices and reimagined them more like unfolding activities brought into practice in an embodied, social and technical way. Kitchin, Perkins and Dodge, (2011), provide a mix of creative, playful and tactile tools to support a participatory mapping initiative, considering the experience knowledge and skills of local inhabitants of three communities in the West Mountain Region of Mexico City. The experience of creating a map as a collective, made them conscious of their memories and heritage, linked to their natural resources, made them reflect on their own identity and knowledge and mobilized toward more organized and purposeful actions.

The West Mountain Range of "Las Cruces" comprises one of the last forest areas surrounding the Mexico City megalopolis. Five original indigenous[1] towns settled formally in 1860 as the District County of *Cuajimalpa* and have been living in a direct relationship with their environment and close to their traditional knowledge and religious rituals. Their main activities were agriculture (collected mushrooms and wood from the forests surrounding them), a locally cultivated "pulque"[2] alcohol, and coal production. Unfortunately, various modern pressures on their environment including real estate development, different agrarian regimens, forest closures and corruption has resulted in significant deforestation and river pollution.

The government has attempted partial solutions but have not gained much traction. There is a long history of negligence, inefficient water collection systems and unsustainable programs for river and forest protection. A lot of these settlements are informal and do not have sufficient drainage channels, and the ones that do, are poorly maintained, causing redundant water leaks that bring untreated waste to the underground soil and rivers. There is a partial conservation plan that very few know about or respect. There are also different levels of land tenure, meaning that ownership issues have forced inhabitants to sell their lands to private property, or even relocation to more dangerous areas.

This is the common context that these original towns find themselves in today that constitute the "peri-urban forest belt". The reality is that they are slowly being consumed by urban expansion. The natural resources from these forests are diminishing, yet are still essential livelihood for many, though cared for only by a few, causing many disturbances and conflicts between these communities.

In this context, our academic research team of designers and social scientists worked in collaboration with the Environmental and Territorial Ordinance Procurator's Office (PAOT). We decided to co-create a Natural Resources Catalogue within three of these communities considered original towns in the city: *San Pablo Chimalpa* with 151,127 inhabitants; *San Mateo Tlaltenago* a communal agrarian organization with 14,168 inhabitants (80% of their territory is part of "*Desierto the Los Leones*" National Park); and *Santa Fe Town*, (one of the original hospital-towns founded by missionary Vasco de Quiroga in 1583, it is actually a collection of neighbourhoods belonging to Alvaro Obregon's district). All these original towns preserve a strong traditional representation of the people in front of the governmental structures we call "Delegations", which responds to the urban logic of the city and its government nowadays. The relationship between the socio-religious-parental structure and traditional forms of community governance groups called "*comunas*" had been historically very strong and opposes new ways of government planning and authority which has generated slow processes of transformation in these towns (Portal and Sanchez Mejorada 2010). Since there are very few formal land planning records of the territory from the authorities, and the sources of existing spatial information have many discrepancies and are not public, we decided to pursue a participatory cartographic approach with the purpose of starting to develop a common planning and decision-making activity in the city.

Participatory cartography has been considered a fundamental instrument to help marginal groups represent and communicate their needs within the territory and augment their capacity to protect their rights. (IFAD, 2009). We decided to carry out this initiative as an alternative to empower the community, who have vulnerable backgrounds, are threatened by landfills and other conflicts related to land access and ownership. We also realized that community mapping initiatives can easily fail or even end up being counterproductive, especially when government agencies use the map made by locals, as evidence to favor their points without really considering community needs or voices, causing even more conflict rather than using them to pursue intelligent and balanced negotiation and planning.

Our goal was to include as many stakeholders as possible to better integrate local knowledge and direct experiences from the lived community environments. For this we needed to engage relevant stake-holders, like the community leaders or the women in charge of managing the rural laundries or "*lavaderos*" who have lasted from the colonial period and still represent a traditional way to use natural resources for the community. If we had only relied on tools like geographic information systems technology (GIS), we might be limiting the ratio of engagement to just a few people besides our research team and some specialists (Canevari-Luzardo et al. 2017). We needed to pursue a novel way to target and co-produce local knowledge, so we undertook a study about walking trails that incorporated interviews with the community leaders and developed a generative toolkit (Sanders and Stappers 2012) that supported map making through workshop sessions with a variety of participants and a followed them up with an observational guidebook to be distributed among other neighbours that could not attend the workshops. The design and planning of these generative methods are based on a relational design approach to human agency, called Agency Sensitive Design (ASD) which is based on Actor-Network Theory (ANT) and Activity Theory (AT) framework to develop a relational understanding of built environments. ASD approach suggest a new pragmatic design practice where a more inclusive mind set prevails in favour of more emergent and fluid actions over the prescriptive and controlled control attempt to predict actions. (Kocaballi et al. 2011). Baki Kocaballi and colleagues (2011) suggest six qualities that characterize this relational design approach, built on their analysis of recent developments with situated and embodied perspectives in interaction design field, which we considered to develop the tools for participatory mapping toolkit. We were also inspired by the Argentinian collective *Iconoclasistas,* which have achieved great activists results for Latin American communities. (Ares and Risel, 2015). Within their own style, they have approached mapmaking as a practice, adapted a set of tools to organize mapping workshops, concentrating their efforts in the creation of a set of questions, and icon templates to easily identify major problems.

The study offered an opportunity to test how a processual approach to mapping using generative tools designed with consideration of relational aspects of agency, could trigger different levels of collaborative action in the context of participatory cartography. This new approach helped the three communities to increase legitimacy of the mapping process and led to incorporation of local actions. Interestingly, for two of the communities in the study, those with a more cohesive social and land organization with strong hierarchical and patriarchal distribution of labour, this approach gave women more voice and recognition in the decision-making process within the working groups.

The information generated by the community supported the decision-making process grounded in participation as well as encouraged better cooperation in knowledge co-production between scientists, societal actors and decision-makers. It also informed the designers with a more inclusive way to understand and categorize natural resources and was a key aspect for the interface design of the platform, which is being worked on with some younger members of the community.

Approaching participatory mapping for the West Region communities

Participatory mapping also known as community-based or cultural mapping had its origins around the late 1970's and the beginning of the 1980's, and at its broadest definition involves the creation of maps by local communities often with the involvement of supporting organizations, either governments or non-governments organizations (NGOs). It emerged from participatory rural appraisal (PRA) methodologies, created by Robert Chambers, in 1983 a fellow at the Institute of Development Studies (United Kingdom), and spread widely throughout the development community, emphasizing transparency and inclusiveness of all community members in an event, most often related to a development initiative or some form of community-based decision-making process.

Participatory maps provide a valuable visual representation of what a community perceives as its place and the significant features contained within it. They could include a depiction of natural physical features and resources as well as socio cultural ones. It is significantly different from traditional cartography map-making through the process by which the map is created and the uses to which they are put. They focus on providing skills and expertise for community members to create the maps themselves, to represent the spatial knowledge of community members. (Corbett, et. al. IFAD 2009). The process attempts to make visible the association between land and local communities by using the commonly understood language of cartography, we could argue that, "the power of the map" is assumed as a given, a sentiment first broadly articulated in the by Denis Wood and other participatory mapping pioneers in the 1990s.

Making people collaborate in the construction of a map assumes that the representation has to present some spatial information at various scales, it can depict detailed information of the village infrastructure (rivers, road transport, or individual houses) or a large area (extent of the common natural resources, the distribution of territory and its boundaries) but it could also illustrate intangible things, like important and cultural aspects of their historical and local knowledge. Participatory maps differ from mainstream maps in content, appearance and methodology because they usually represent a socially and culturally distinct understanding of the landscape and include information that is excluded from mainstream maps, which usually represent the views of dominant sectors of society.

West Region Mountain communities and selection of mapping methodology

Our University established in The Santa Fe district in 2004, as part of the West Mountain Region. This area had a long history of controversies, starting from being one of the big garbage dumps for Mexico City in the 1950s when it had a population of 2000 inhabitants. Due to a "modernization" project run by Mayor Carlos Hank Gonzalez from the early 1980s to the mid 1990s has been transformed into a large-scale urban corporate and commercial zone (Moreno-Carranco 2013). The area has surpassed its growth capacity and now land prices have been increasing, forcing local communities, especially the original towns which, use a different organizational structure and land ownership, regulated and recognized by their traditional knowledge and their agrarian origins as 'commons' by the government of the City. They have been forced to sell these lands and move towards other areas currently within conservation districts.

The story of the original towns (settled before the Mexican colonial age) have many tensions and conflicts between communal leaders who had been losing political representation, and the change of their common status due to changes with land use. Many disturbances begin with uncertain property boundaries and excessive urban growth, which places critical pressures over critical natural resources, mainly forests and water. An additional mass migration of outsiders in part due to the earthquake of 1985, poverty from other states like Tlaxcala, Puebla, Veracruz, Chiapas, Michoacán and those displaced from other landfills have initiated informal settlements to take root in the region, threatening one of the most emblematic Natural Resources like "*Desierto de los Leones*" National Park and hundreds of acres of forested preserved areas in the region.

Besides a few isolated efforts there is no Natural Resource Management plan implemented in the region. Communities are fragmented, and usually isolated due to border or land conflicts and corruption. In terms of their hydrology the main rivers like *Río Borracho* and *Rio Atítla* are highly polluted despite the natural occurrence of springs, which inhabitants hide from authorities out of fear of being displaced. Inappropriate agricultural practices and rapid growth of informal settlements have led to water pollution; decline in river flows, and accelerated soil erosion. Combinations of these factors, along with deforestation, are principal the main causes of local environmental degradation.

Part of our mission as a public university in the region is to facilitate the transfer and propagation of knowledge and promote connections between the inhabitants of the original towns with their neighbours, corporate and government agencies. The university is also interested in developing a long-term mind set of sustainable management for natural resources in the area.

We co-created a current catalogue of natural resources that is meant to be a continuously evolving mapping activity reflecting the interests of the community. These artefacts are designed to help reflect the interests of the community, and function as a container that documents and protects local knowledge and communal experience. We saw the co-creation of a Natural Resource Catalogue with the community as an opportunity to facilitate the gathering of information about natural resources in the region. We hypothesized these living documents could increase the ability of the original towns to express their own traditional knowledge and land-related rights. It has helped them share their collective experience through partnership with scholars and reinforces social networks to other nearby towns.

We can recommend participatory mapping as highly effective for indigenous or marginal communities, in particular, where elders share traditional place names and stories with other members of the community. It can also generate interest in the local knowledge, especially among the youth.

One of the functional advantages of GIS technologies is that they convey a sense of unbiased authority making them a valuable tool for advocacy and for influencing land-related decision-making with other stakeholders. From our scientific perspective we needed to use these technologies to store, retrieve, map and analyse geographic data but we also needed to integrate, and layer local knowledge and data generated for the community to use. So, we decided to assume an intermediary or facilitator role for technology and assumed a "partial

participatory GIS approach" a process where all the computerized aspects of GIS are undertaken by a technical expert (Canevari-Luzardo et al. 2017); in our case, the main designer and geographer of the research team.
Our process consisted of 4 stages:

Stage 1. Diagnosis and delimitation of the area of study and community approach – Our University arrived and delimitated their area of influence. We established preliminary contact with community leaders. In this stage we germinate the concept of a co-production of a Natural Resources Catalogue with the communities would help strengthen the relationship between the University and the communities. We also detected that in order to support their traditional knowledge and legitimize their decision-making processes it was crucial to involve them in participatory mapping initiative, since there was no spatial information generated from the community perspective and their memories and practices linked to their local natural resources were being lost.

Stage 2. Fieldwork and information gathering with the community – We conducted some trials and interviews with people from the community using Geographic Positioning Systems (GPS) and recorded photographic and video material of salient resources of their environment like springs, plants and trees, and settlements. The material generated in this stage was crucial for planning the participatory mapping workshop in later phases. We designed a fieldwork guide for the research team, to conduct a natural resources audit, document natural resources specimens, their location and the stories and popular uses to the inhabitants. In this stage we developed the materials for the workshops and prepared the communities for the mapping activity in the next phase.

Stage 3 Community Mapping Sessions – In this stage we facilitated and implemented the workshops with each community using the generative toolkit. We introduced the communities to a short explanation of the purpose of mapping, and the range of tools available to them. We also explored the potential use of the map as part of the catalogue with participants. Participants developed 11 maps. Following the workshop, we provided an observational guide for other members of the community that couldn't attend the mapping exercise or that thought they could complement the information later from their homes. After this stage we collected 10 guidebooks from the communities and integrated this content into the maps.

Stage 4 Evaluation monitor and map use phase – We analysed and evaluated all the community developed maps and integrated the data into general themed maps for each community including three categories: natural resources (water, flora and fauna and forestry), land-use regimen boundaries, and environment impacts. We integrated these into a natural resource platform proposal. As a mechanism of feedback, we took the maps and the proposal for the platform and arranged for feedback questions and interview sessions with the community. All the suggestions and changes were integrated as the platform was designed.

Materiality and the activity of mapping

The notion of co-production of a natural resource catalogue involves a collective activity applied across the process of map making and the interaction of the participants with material-based agency.

In our experience as researchers and practitioners we have seen that co-creation practices requires to use the design process as a means to enable a wide range of activities for different stakeholders in order to collaborate (Burns et al. 2006). We needed to change our researcher's perspective of the participant roles in the map-making exercise. We transformed the people from the community from passive objects of study to active and willing collaborators that need to acquire certain spatial skills for knowledge production working with an expert group of researchers. Our end goal was to bring knowledge from theory into practice in a way that understands the technology of mapping as well as respecting local knowledge.

Mapping is a collective activity where participant's roles get mixed. The person who eventually is going to use the map is given the position of being the "expert of his/her experience" and plays a larger role in overall knowledge development. Evidence for this statement happened during a particular trial in *Desierto de los Leones* National Park with Mr. Juan Esparza, one of the community leaders of San Mateo Tlaltenago. At one point of the trial Mr. Esparza stopped in an open valley to explain our location and other important issues in the territory, he grabbed a stick and started eloquently to explain by drawing with it in the ground. That moment was crucial evidence that Mr. Esparza had vast and critical spatial knowledge of the landscape and location of natural resources in his territory.

Mr. Esparza's explanation reminded me of a story Bruno Latour (2003) uses in his book "Science in Action" to analyse how specific inventions like cartography, helped people to construct facts for an argument. Latour's focuses specifically on how someone persuades someone else to take a statement. To illustrate his point, he refers to La Pérouse travels through the Pacific, for Louis XVI in 1788, with the specific mission of bringing new knowledge to their civilization. Landing in a particular place he encounters aborigines, to his surprise they show him they understood geography quite well, by answering La Pérouse question of where they are, they draw a map of the island on the sand with the scale and details needed by Pérouse to understand. Another, who is younger, sees that the rising tide will soon erase the map and picks one of the explorer's notebooks and draw the map again in pencil.

Latour questions the difference between the 'savage geography' and the 'civilized' or 'scientific mind.' Both actors in this encounter are able to think in terms of a map and navigation, strictly speaking they both have the ability to draw and visualize based on the same principle of projection, first on the sand and then on paper. What he tries to explain in his example is that knowledge is relative, if we analyse the situation closely, the purpose of that drawing changes for each of the persons that generates it. From the side of the aborigine, and in our case from Mr. Esparza, there is no doubt that he and his team know their territory quite well, there is no problem if the drawing fades away, it can be redrawn at any time. In the contrary for La Pérouse or our research team's perspective: the drawing represents a core part of the mission, we need to be able to establish document and pass the location of those places and species and bring them back to people who expect certain documentation. These people expect 'a map' as evidence to determine the contents and locations of this part of the world and if it is worth another visit for claiming new natural resources for their future interests and exploitation. La Pérouse and our team's exclusive interest in this representation relies on some specific attributes that let us incorporate it as projection, writing, archiving and or computing. This capacity somehow needs to hold and endure the journey back to the place where people await that information. The critical information needs to be stored for later use and discourse. Our team and La Pérouse interests "hold on" to a long tradition of knowledge and practice that has been constructed through manipulation of paper, prints and images accumulated through our own culture.

According to Latour the difference between us as researchers and Mr. Esparza's team as actors in this situation is in the strategy, in the power provided by the semiotic material which is inscribed in the object we call a map, how is it that this particular inscription results in something convincing. In other words, Latour focuses on the mechanism used by the artefact to sum up "groups of allies".

Thinking about our project aim, we needed to come up with useful tools for the community in the process of mapping and understand their familiar inscriptions and materials that would facilitate their creative activity. In sum, we needed to invent objects, which have the properties of being mobile, but also immutable, presentable, readable and combinable with one another. (Latour, 1986 p. 21) From our participant's perspective, we needed them to be able to translate their local knowledge and memories related to their natural resources through conversation and pour them directly onto a collective map, we needed to think maps as "immutable mobiles".

Concept of inscription in Design

The concept of inscription is crucial for the design activity because any designer aims to create, modify, enable and or constrain some capacities of action through the designed artefact. Akrich (1991) explains the notion of inscription as: a vision, value, program of action, or prediction about the world that the designer ascribes in the technical content of a new object.

The strength of an inscription may vary from being very strong, that is, imposing on a particular inflexible program of action, to the very weak, offering many flexible programs of action, according to Kocaballi. Strong inscriptions belong to a design perspective of design that aims to predict, prescribe and control the kind of relations between humans and technologies and the ways in which their interaction unfolds.

By letting a few groups of specialists control the technology and resources in the process of map-making of a community instead of facilitating it, we are characterizing the human-technology interaction shaped by strong inscriptions in that situation. This is not suitable in situations where we need appropriation, personalization and adaptation or when exploration is needed. Participatory mapping is a process that does not benefit from the assumptions of agency as predictable and fully controllable phenomenon. On the contrary, to acknowledge and develop sensitivities to manage relationally for designing new technologies of mapping we must formulate design solutions that can deal with the unexpected situations in the various cases of participatory mapping that inevitably arise.

Co-creative approach and generative tools for the mapping activity – The generative tool kit to support the participatory mapping workshops with the communities had the following components:

- A visual presentation explaining the objective of the catalogue, maps and activities involved in the workshop
- A base map that could be a satellite image or a topographic layout of the community
- Translucent or clear paper to overlay the base map, Post-it notes and colour labels with different shapes.
- Photographic images of relevant species and places that we found during the trials (plants, trees, springs, places, animals and short video recordings.
- Glue and Velcro tape, small wood cubes, color markers and cards
- The observational guidebook was a follow-up activity for potential participants after the workshop.

For the purpose of explaining our methodology, we are considering all the components of the toolkit mentioned above as mediating devices that influence the individual experience of the participants in the mapping activity. This focus on artefacts is borrowed from activity theorist Wartofsky, who describes an artefact as being useful for creative thinking. He emphasizes the activity of representing with a purpose, that human beings create their own means cognition, signalling the existence of tertiary artefacts which "transcends the more immediate necessities of productive praxis," giving freer rein to imagining "possible worlds" (Wartofsky 1979). And such possible worlds function as models, embodiments of purpose and at the same time instruments for carrying out such purposes. Based on this, we argue that all the maps generated by each of the collectives in the workshops, specify future object-oriented activities. They serve far beyond their immediate environment, propelled by the creative activity of the collective group.

Co-creative approaches to solve complex problems and identified future opportunities do not belong to a particularly discipline or domain. In fact, very similar approaches exist under the umbrella of Participatory Design that combine the expertise of designers and researchers and the situated expertise of the people whose work is to be impacted by a change. All these approaches are currently in use by academics, designers, international development and social the sciences. (Sanders and Stappers 2012)

Generative design approaches empower everyday people to generate and promote alternatives to their current situation and is based on the motto: "all people are creative". The name "generative tools" refers to the creation of a shared language that researchers and other stakeholders use to communicate visually and directly with each other. The design language is generative in the sense that with it, people can express an infinite number of ideas through a limited set of stimuli. The generative tools approach in our case aims to provide simple and tangible materials to help participants communicate knowledge and memories linked to their natural resources through the exercise of mapping. We looked for inexpensive materials that required no professional or special verbal skills, low spatial expertise, and low effort to construct a tangible artefact.

The selection of the materials, colours and icons should encourage the expression and reflection of past memories and previous experiences. They should be designed to facilitate the process of participation of people unfamiliar with your goals. Participants select materials, point, draw colour and build artefacts and explain to others what relevant information they are aware of. These individual artefact creating activities are a way of harvesting a collective wisdom into a layered and integrated whole.

The Generative Design (GD) approach as well as Agency Sensitive Design (ASD), are two approaches that helped us to develop and design the tools in our sessions. We see many parallel principles between the two approaches for ideation and expression for the mapping exercise. GD sees generative tools as a methodology within design research, focusing on materials or objects of creation for non-designers, through a shared design language that researchers use to visually communicate between the parties involved in a project. With these material objects, people generate and promote alternatives to their current situation and allow people to express their visions, wishes and expectations about the future.

ASD, in comparison, supports a relational nature of human agency (Kocaballi *et al.* 2011) where agency is neither an attribute of the subjects nor the objects, but an ongoing reconfiguration of the world and ultimately an effect of a heterogeneous network of human and non-human actors. (Latour 2005). These new approach to the concepts of agency has been very helpful for projects that need to rethink how technology (e.g. artefacts, tools, objects or things in general) interacts with human intentions and social structures. We argue that this approach is complementary to the participatory perspective. Instead of trying to control, predict or design actions and relations for the user, designers may look for more emergent and fluid relations in the situation they envision. Kocaballi (2012) sets out six different qualities: *Relationality, Visibility, Multiplicity, Accountability, Duality and Configurability.* The majority of these qualities are relevant in creating conceptual lenses for designers to gain a relational understanding of a situation and increase their awareness to accommodate the diversity and richness of human agency and to perform a more responsible and ethical design practice.

Both approaches need to promote alternatives to a specific current situation based on the notion that all people are creative and have knowledge, they become “experts of the experience”. They value the local expertise of the people who inhabit the environment and challenge the existing power structures that exist between dominant organizations.

One mapping session was held for each community through the period of March until May of 2017. The collection of tools the participants used were planned as a common language to lead the participants through conversations where they could communicate their stories, feelings and ideas while constructing the map of their territory. One of the greatest strengths of this initiate relied on the ability to bring mapping process to community members and share together ideas and visions, which can contribute to building community cohesion (Alcorn, 2000). We will use GD and ASD framework to explain some components of the mapping exercise.

Relationally – The quality of relationally refers to the connectedness and relatedness of human and non-human actors or socio-material arrangements where they co-constitute each other through their interactions.

To reconcile the relational character of our capacities for actions, the constructed nature of subjects and objects and the corporeal grounds of knowing an action, we designed the activities and materials for the sessions considering three sensitivities:

1. Understanding of mutual influence, shaping and co-constitution of actors and artefacts.

2. Embracing and supporting emergent and improvised actions.
3. Consider the mapping activity as an assemblage of actors, artefacts and collective hybrids.

For example, the images used for each session were collected from the trials with each community and they represented local places familiar for them, each session was conducted with the same structure but with small differences. In the Santa Fe's session, participants were more used to mobile technology and wanted to incorporate their own images for their maps, so we marked that with posit it notes, and then asked them to send them so we could incorporate them in final maps. We did not make a specific sample criterion for participants but encouraged elders as well as women. In one of the workshops, children were brought to the sessions and we were happy to work with all of them.

From all the generative materials for mapping we should like to stress that the printed photographs and iconography taken from their own environment, helped the flow of actions for all the participants by supporting emergent and improvised actions (Figure 1). The easy manipulation inscribed in simple objects enabled each participant to pick an image and attach it to the map, by doing so they could easily describe memories and stories related to their natural resources or express their concerns that threaten their environment.

Figure 1. A group of participants working with materials to build their map at Chimalpa's Workshop. Nora Morales. Photograph © Nora Morales.

Visibility – The quality of visibility is closely related to qualities like multiplicity and accountability. It involves making visible invisible work, human and non-human actors, infrastructure and interaction during the mapping activity. Visibility not only facilitates the overall awareness of human actors of themselves and others, but also helps the performance of more responsible practices.

This quality was very important to support user appropriation by making resources publicly available. The observation guidebook (Figure 2) given to some participants after the mapping activity is a great example of the quality of visibility, it allowed the people that participated in the workshops to continue collecting information related to their memories and activities linked to natural resources. In some cases, it was given to some neighbours and allowed us to combine more information and integrate more participant knowledge into the project. The layout was designed following a rough sketch style, with a lot of white space encouraging the participants to fill in the space in a more creative way, by pasting their own personal images. In one of the cases a woman showed her family tradition of recollecting mushrooms "*hongeros*" by placing pictures from their family album and imprinting a detail inscription regarding each mushroom type and classification.

Figure 2. Observational guidebook from a participant from *Chimalpa* using family pictures to describe different types of mushroom and activities related to their recollection. Photograph © Nora Morales.

Multiplicity – This quality refers to multiplicity in ways of knowing and representing, which entail participation and heterogeneous sources of influence in the mapping process. In the workshop we had to overcome the traditional dichotomy of scientific/indigenous, expert/layman, men/women embrace knowledge diversity rather than our own traditions focusing on hierarchies.

We established mixed teams of participants and some of the elders couldn't write or read, so each team freely developed roles some of the participants acted as tellers, and others as writers. We also distributed a set of categories that we encouraged to be broken into different classification systems of their resources. There was an instance initiated by a female participant in one of the teams in *Chimalpa*, who tried to explain what she thought were root causes regarding their natural resource's issues using the "4 element classification of nature" (Water, Air, Fire and Earth) from ancient Greece. For example, she ascribed the problem of air pollution under the element of Air, and linked to various causes: burning trash, automobile pollution, lack of ecological culture, garbage in the streets, and animal waste. Those were common practices and situations from the inhabitants of the community. This enunciation later provided fertile ground to develop solutions they could implement.

This example is evidence of how the quality of the mapping activity was able to engage the group in the making of a *rich* map, describing themselves and their particular forms of practice. It also helped them explain to us the complex relations and incorporated multiple points of view of how they see their natural resources.

Accountability – Organized action can be observable and reportable. The materials and activities of the mapping session provided the participants with information about their own activities by dividing them in four or five groups (depending on the number of assistants). When finished, we asked them to explain their work to other groups in a plenary session. The participants were required to relate their position and perspective from other actors taking responsibility for their own perspective and partial knowledge.

Duality – Our designs invite particular kinds of actions, while inhibiting certain others (Latour 2005). This quality is strongly related to the idea of inscription of values into an object or technology. By using the kind of iconography with a "sketchy look" for the maps and the guidebook and the use of paper and tactile materials for the workshops we engaged regular people like woman and children who are more familiar with crafting activities. An important gender note: by prioritizing crafting technics, we de prioritized the formality around written and oral speech that is usually ascribed to male dominant formats. Men in these contexts are usually the ones acting as local leaders, acting as "commissioners of the commons," especially when they negotiate with government delegations.

Configurability – The design process does not stop after the map production phase but the actual use of the map for the community, so our research team also developed a continuous organization of activities with the community to continue the integration process of information between technology and human actors transforming the data we collected into useful knowledge they could use. That is why the observational guidebook was key to the interaction with participants for the community from which we had very good results. We also are planning an open structure for the platform that will let participants continue adding information from their terrain of natural resources on an on-going basis, just like the mapping project.

After the workshop with each community each team generated a collage-map with the type of evidence that we were interested to connect with spatial data, which relates more to community memories and their shared understanding of the problems as well as local knowledge linked to their natural resources. This type of information is usually avoided in formal cartographic scientific maps.

MAPS AS SOCIAL CONSTRUCTS OF KNOWLEDGE

After long sessions of analysis of all the materials, we were able to identify some tangible indicators as result of the mapping activity within the three communities at different scales:

Generation of thematic maps for the platform and validation.

The exercise let us identify which natural resources were more important to the community, San Pablo *Chimalpa* and *San Mateo Tlaltenago* showed a particular interest in their springs and trees, while in Santa Fe's town they were worried about natural landscape and trees as supports if their home's infrastructures since they are settled in the hillside near *Rio Tacubaya*. In the first two communities we noticed interest of recovering old pre-hispanic practices and language, they even refer to some places by referring to pre-hispanic names, they are also proud of speaking native languages and some individuals are keen on excavating their lands looking for archaeological object. With this information we developed thematic maps that will be part of the catalogue digital platform and had a preliminary session of feedback with the communities. As a result of the sessions we need to adjust some boundaries depicted on the collage maps and validate it with other neighbouring communities we are also thinking on generating a map with ancestral boundaries and pre-hispanic names.

The mapping activity trigger local action from the *individual scale to collective actions*. Some examples referred to the individual level: In one session some men from Chimalpa said they were willing to start their own pulque production after they socialized to their team members. There was a group that recovered traditional knowledge from an elderly woman that used to be a "*leñera*" women that carried in their backs (in absence of animal ownership), logs and wood chips from the sawmill in the forest to their towns to make a living. That woman couldn't read or write, but orally accounted for the way she and her co-workers managed to transport the logs, she even remembered some prehispanic names the trees and places in the forest that were called "*parajes*" and corresponded to natural boundaries, she even remembered the names of the tools they use for carrying that they made with their own hands. Stories like these are evidence of how knowledge could be transferred from the elders to the youth within the community through the mapping process. The information generated at the workshop enabled participants to assume different roles while they communicate their ideas to others and supported equality in the decision-making processes,

The use of tactile and picture-based materials for map making was especially helpful for women of two communities in particular. These communities have a more cohesive social organization based on a 'communal commissioner' of the land, who holds a more traditional hierarchical and patriarchal division of labor within it. The objects used for the activity were made familiar to them through their materiality and encouraged them to speak their voice, especially for the elders in their group. In these situations, dominant men are the ones who usually tend to speak while others remain quiet, respecting their leader position for their group.

From individual to a collective activity

The processes of communication and coordination between individuals engaged in the collective activity of mapping has evolved between the participants and created some relationships binding them one to another, while we were doing follow up interviews with some of the participants we learned that some actions and initiatives have started from some of the groups, like women from the community of San Pablo, started a plant and herb recipe book which now is being worked with collaboration from students from our university.

From an interview to a community leader from *Chimalpa,* we learned that the community is meeting during the weekends to clean their rivers and also have initiated some consensus-based management that asks for the owners of private lands to tell the community leaders first, if he or she is planning to sell their land, so the land could stay within the community first.

In San Mateo, there has been some changes in the role of the community leaders and the politics of representing land tenure, that might have influence within their broader regions, some participants of the workshop are thinking about producing conservational areas maps to influence government land decision making in the construction of the inter-urban train coming from Toluca-Mexico.

CONCLUSIONS

The approach of mapping as a practice helped our academic team to overcome the high level of complexity that can marginalize poor communities from sharing their knowledge and beliefs of their territory. Our method provided the three groups with tangible information artefacts that let each person be aware of the spatial knowledge they possessed and how he or she can make immediate use of it. But more importantly it led beyond a common knowledge situation, where one or more people not only know something, but also all the others know, that they know. In other words, led to evidence their local knowledge among them.

By offering different ways to communicate their ideas, we also mobilized the information flow from one person to another and touch the community to a certain level, where the data endowed provided them with relevance and purpose regarding their natural resources. Specially by giving voice to women, an let them express their organizational routines, processes and practices in regard to their natural resources led men to recognize their primordial role in local knowledge.

All the memories evoked through the process of map making and embedded in the maps, made them reconnect with their personal and family history and understand their heritage regarded their natural resources. It empowered them by making them aware of their main risks and threats in their territory and established a hierarchy of resources that they needed to respond through collaborative and coordinated activities towards a positive change, instead of waiting for the government to do something.

Each community started actions according to their situation; in *Chimalpa* they are organizing weekend activities to clean the rivers, and a catalogue of traditional medicine and mushrooms in collaboration with students, In *Tlaltenango* they are looking for new alternatives to communicate their boundaries to the government, and in the Town of Santa Fe they are starting to communicate among neighbours.

The emergent voice of the community was pronounced and ascribed to different forms of acknowledgement regarding natural resources management.

There are still strong barriers to overcome when applying GIS technologies to participatory mapping with rural or indigenous communities in developing countries, mainly to limited financial resources and lack of technical skills that automatically sets an unbalanced situation, from the community to the few experts, that automatically positions local inhabitants at a disadvantage within a power position.

There are major problems to attend to regarding the map authorship of these initiatives and the unintended negative consequences of exposing the information generated by the community. Either purposefully or inadvertently we will end up with private or otherwise valuable community knowledge. In our case it took the form of the location of informal settlements. Depending on how we showed it, or to whom, and how it will be managed could mean further harm and/or marginalization of exploitation of the land to the disadvantage of these people.

Perhaps these kinds of initiatives can help apply more pressure and inform official decision-making processes, and location-based technologies will one day be more accessible to everyone, or maybe we might settle down into something more widely useful, but for now it is still essentially a well-intentioned technological mess and remains unclear how this technology will actually help address these issues.

The absence of best practices and standard methodologies it becomes crucial to assess the validity and credibility of mapping processes within the context of the purpose and its use. If we believe that the ultimate purpose of maps is to support the general prioritization of actions or to increase the adaptive capacity within a community, we need to come up with new methods that help communicate or narrow the gap between local people and government.

We might not have one unique answer for achieving a successful participatory mapping project, but we believe we have revealed a methodology to balance the intended purpose of a map, the available resources, capacity within the community, the duration of the commitment to the project, and finally a way to trigger action and reflection through the process of map making. We might not yet be able to free the power of maps to just a few groups, but we might be pushing to keep all the voices of the stakeholders in a territorial project within the same level. Perhaps even these are merely the preliminary stepping stones to help us rethink new technologies from a different perspective and a step forward to achieve a successful participatory mapping initiative.

We can agree with Kitchin's statement that maps are never fully formed, and their work is never complete: they are transitory and fleeting, relational and context-dependent. My question to you would be how we would to establish a monitoring mechanism, and strategies to adapt to these constantly changing activities? The experience of this project is evidence of the truly fragile status of a place and these inhabitants. They face enormous pressures. At the moment I am narrating this story, these lands and natural resources of the three communities are being threatened and transformed by a giant infrastructural issue the construction of the Inter-urban train *Toluca-Mexico* which has caused the felling of many trees and species, as well as destruction of natural habitats near water springs land. It is doubtful to us that these projects used anything like a participatory map-making process as depicted in this paper. With this transformation the human activities and practices of people in this territory is being transformed too.

NOTES

1. This communities have pre-hispanic origins, some of the regions are even mentioned in pre-hispanic codex as part of the villages in the outskirts of the mountains, there are still residents that remember their parents speaking indigenous tongues like: Nahuatl and Otomi.

2. A Mexican alcoholic beverage made by fermenting sap from the "maguey" a variety of fleshy-leaved agave plant known as "Century plant".

REFERENCES

Akrich, Madeline
1991 The De-scription of a Technical Object. In Shaping Technology/Building Society: Studies in Sociotechnical Change in Bijker et al, eds. Shaping Technology/Building Society MIT Press, 205-224.

Alcorn, Janis
2000 Keys to unleash mapping's good magic. Participatory Communication. International Institute of Environment and Development *PLA Notes* 39:10–13. accessed March 15 2018 http://pubs.iied.org/pdfs/G01906.pdf

Ares, Pablo, and Risel, Julia
2015 *Manual de Mapeo Colectivo: Recursos Cartográficos Críticos para Procesos Territoriales de Creación Colaborativa.* 2nd Edition. Buenos Aires, Argentina: Tinta Limón.

Burns, Colin, Cottam, Hillary, Vanstone, Chris. and Winhall, Jennie
2006 Transformation design. Red Paper 02, Design Council: UK.

Canevari-Luzardo, Laura, Joan Bastide, Isabelle Choutet, and Diana Liverman
2017 Using Partial Participatory GIS in Vulnerability and Disaster Risk Reduction in Grenada. Climate and Development 9(2): 95–109.

Corbett, Jon, and International Fund for Agricultural Development (IFAD)
2009 *Good Practices in Participatory Mapping: A Review Prepared for the International Fund for Agricultural Development* (IFAD). Rome: International Fund for Agricultural Development. http://www.ifad.org/pub/map/PM_web.pdf, accessed may 4, 2018.

Chambers, Robert
2006 Participatory Mapping and Geographic Information Systems: Whose Map? Who Is Empowered and Who Disempowered? Who Gains and Who Loses? The Electronic Journal of Information Systems in Developing Countries 25(1): 1–11.

Kitchin, Rob, and Martin Dodge
2007 Rethinking Maps. Progress in Human Geography 31(3): 331–344.

Kitchin, Rob, Chris Perkins, and Martin Dodge
2011 Thinking about Maps. *In* Rethinking Maps: New Frontiers in Cartographic Theory Pp. 1–25. 1. New York, N.Y.: Routledge, Taylor and Francis Group.

Kocaballi, A. Baki, Petra Gemeinboeck, Rob Saunders, and Andy Dong
2011 Towards a Relational Approach to Design Process. *In* .

Kocaballi, A. Baki, Petra Gemeinboeck, Lian Loke, and Andy Dong
2012 Embracing Relational Agency in Design Process. 2012 1(DeSForM 2012Meaning, Matter, Making,): 99–109.

Moreno-Carranco, María
2013 Global Mexico under Construction: The Santa Fe Megaproject in Mexico City. *In* Transbordering Latin Americas. Liminal Places, Cultures, and Powers. Clara Irazába. Taylor And Francis Group. New York: Routledge.

International Fund for Agricultural Development (IFAD)
2009 Good Practices in Participatory Mapping: A Review Prepared for the International Fund for Agricultural Development (IFAD). Rome: International Fund for Agricultural Development (IFAD).

Lapenta, Francesco
2011 Geomedia: On Location-Based Media, the Changing Status of Collective Image Production and the Emergence of Social Navigation Systems. Visual Studies 26(1): 14–24.

Latour, Bruno
1986 Visualization and Cognition: Drawing ThingsTogether. Knowledge and Society Studies in the Sociology of Culture Past and Present 6(0). Jai Press: 1–40.

Latour, Bruno
1987 Science in Action: How to Follow Scientists and Engineers through Society. Milton Keynes: Open Univ. Pr.

Latour, Bruno
2003 Science in Action: How to Follow Scientists and Engineers through Society. 11. print. Cambridge, Mass: Harvard Univ. Press.

Latour, Bruno
2005 Reassembling the Social: An Introduction to Actor-Network-Theory. Clarendon Lectures in Management Studies. Oxford; New York: Oxford University Press.

Miettinen, Reijo
1999 The Riddle of Things: Activity Theory and Actor-network Theory as Approaches to Studying Innovations. Mind, Culture, and Activity 6(3): 170–195.

Portal, Mariana, and Cristina Sánchez Mejorada
2010 Estrategias culturales, estructuras tradicionales y gestión social en el pueblo urbano de San Pablo Chimalpa. Nueva Antropología 23: 119–146.

Sanders, E. B.-N.
2000 Generative Tools for Co-Designing. *In* Collaborative Design. Stephen A. R. Scrivener, Linden J. Ball, and Andrée Woodcock, eds. Pp. 3–12. London: Springer London. http://link.springer.com/10.1007/978-1-4471-0779-8_1, accessed March 19, 2018.

Sanders, Liz, and Pieter Jan Stappers
2012 Convivial Design Toolbox: Generative Research for the Front End of Design. Amsterdam: BIS.

Wartofsky, Marx W.
1979 Models: Representation and the Scientific Understanding. Boston Studies in the Philosophy of Science, v. 48. Dordrecht, Holland; Boston: D. Reidel Pub. Co.

Wood, Denis, and John Fels
1992 The Power of Maps. Mappings. New York: Guilford Press.

Eye Tracking in Medical Ethnography: Evaluating Evidence for Perception, Action, and Collaboration in Healthcare Professions

LARRY S. MCGRATH
Design Science Consulting, Inc.

Using eye tracking in ethnographic research poses numerous theoretical and practical challenges. How might devices originally intended to record individuals' vision of two-dimensional planes be useful in interpersonal contexts with dynamic visual interfaces? What would the technology reveal about collegial environments in which different levels of knowledge and expertise come together and inform decision-making processes? Why would pupil movement show us anything that conventional ethnographic methods could not? In this paper, I argue that these challenges are not intractable. When tailored to specific questions about perception, action, and collaboration, eye trackers can reveal behaviors that elude ethnographers' gaze. In so doing, the devices enrich the observational and interview-based methods already employed in ethnographic studies of workplace dynamics.

Hospitals are a fruitful context in which to test the value of eye-tracking evidence. Healthcare professionals look, interpret, and act on wide-ranging streams of visual information. The focus of this paper is to explore the possibility of carrying out potential fieldwork in pulmonology suites. I put forward a hypothetical eye-tracking framework to interrogate: 1) the different perceptual skillsets exhibited by pulmonologists; 2) the collaborative decision-making practices by which medical teams interpret monitors' bronchoscopic imagery and perform lung biopsies. According to this model, optical evidence (i.e., fixation durations, scan paths, and heat maps) could demonstrate not only how healthcare professionals look at visual interfaces; the evidence could be especially useful to show when and where vision runs aground. Such gaps in eye-tracking data illuminate additional sensory systems which facilitate information consumption and exchange. When brought to bear on potential fieldwork, the eye-tracking model offers an incisive point of departure for ethnographers to investigate, observe, and present questions about peoples' perceptual expertise.

Far from opening a glimpse "inside" peoples' heads, eye trackers add one perspective among others. When accommodated to the mobile conditions of immersive research, the technology offers a modest, though useful, contribution to ethnographers' toolkit.

INTRODUCTION

"Feel my hand," says a young pulmonology fellow as she extends her right arm toward me. Her fingers quiver; they cup inward, bearing the trace of a grip held tight for over an hour. We are in a major research hospital located on the West Coast of the United States and the fellow has just completed a bronchoscopic lung biopsy on an elderly man suspected to have cancer. I asked what it was like to carry out the procedure. Her hand tells the whole story. Bronchoscopes are 115cm-long tubes with a small probe at the end. It generates a live image of the airways on a monitor. The tube slides down a patient's throat and snakes through the bronchi thanks to the guidance of a handle grasped in the pulmonologist's hand. Manipulating a bronchoscope demands a high degree of manual dexterity. Awkward twists of the wrist and shifts of the body rotate the instrument's direction while the thumb and index finger depress to flex the probe and extract tissue samples. Now the fellow is exhausted. The chief pulmonologist, under whom she is studying, handed her the reigns to conduct the biopsy from beginning to end. Although it's not her first time, she's hardly experienced by medical standards. Behind her are four years of medical school and another four of residency. The pathway to mastering the technique will involve many more throbbing hands over the next three years of the fellowship until she learns to loosen her grip, feel the shape of the airways' anatomy, and rely ever so slightly less on the monitor's visualizations before her eyes.

Experienced pulmonologists often take for granted just how difficult it can be to cultivate the necessary skillset. Monitors' streams of imagery compete for attention with the physical demands of bronchoscopes. How does one strike the right balance? The skills feel innate after years of practice. Such was the line of thought conveyed by the chief pulmonologist. She reflected: the more you do, the more you forget what was hard about it when you first started; over time that becomes seamless and you don't remember until you're trying to teach someone. Teaching fellows entails an intensive mentorship, which is made no easier when inculcating what comes across as an inscrutable knack. How might such an exacting but crucial skillset for bronchoscopy be made transparent, intentional, and helpful for those still developing their practice?

The scenario is ideal for eye tracking. The technology renders implicit practices explicit, which is key to understanding the perceptual skillsets that distinguish experts from novices. This is especially the case in light of the established theory that rational reflection plays a minimal role in expert perception.[1] Eye trackers have been used to study, for instance, how professional pianists' vision hovers between the keyboard and sheet music; both remain in view. By contrast, the tyro's gaze frequently jumps back and forth from notation to the keys.[2] Similarly, psychologists have tracked the pupil movement of elite cricket players and found that they do not follow the ball throughout its flight when batting. Their vision is predictive; it fixates on the path that the ball is likely to follow.[3] Yet, ethnographers like ourselves who study workplace dynamics often shy away from eye trackers. Although the devices could be used to analyze perceptual skillsets and make their constitutive elements intelligible, disciplinary reasons stand in the way. Anthropology examines public behaviors in shared cultural contexts.[4] Apart from winking, pupil movement is a private activity (or at least it appears so at first blush). Eye tracking might thus seem alien to the immersive and interactive environments in which ethnographic inquiry unfolds. In the sections that follow, I argue the contrary is true. I set about establishing principles to adapt eye tracking to potential ethnographic studies. When addressed to specific questions about looking, feeling, and acting, eye tracking can be used to transform private practices into public knowledge.

Exploring perceptual behavior is not new to workplace ethnography. Charles and Marjorie Goodwin showed how airport personnel look at airplanes to make quick decisions about transferring baggage. Despite their brevity, baggage loaders' momentary glances are structured by vast organizational practices of information sharing.[5] The study used cameras to record personnel from multiple angles while they interacted with airplane code sheets and identified haphazardly parked airplanes. Lucy Suchman carried out similar video studies. She recorded airport operations rooms in which people coordinate plane schedules using computer screens, radio frequencies, and telephone lines.[6] Eye trackers add another camera to the mix.

My argument is that eye tracking stands to enrich ethnographers' observational and interview-based methods so long as the technology is adapted to anthropological methods and applied to specific problems at the intersection of perception and action. The task is not innocent. It involves theoretical and practical challenges. Eye trackers were originally designed to evaluate stationary observers' gaze of two-dimensional surfaces. Effectively incorporating the devices into field research hinges on whether ethnographers can accommodate them to the dynamic conditions of three-dimensional environments. Ultimately, such challenges are not insurmountable. An aim of this article is to present methodical insights from psychology, philosophy, and anthropology, which help to wrest eye tracking from its conventional uses in superficial visual studies.

At Design Science, we brought these insights to bear on field research in American hospitals. Our goal was to understand how diverse levels of expertise inflect perception and action, particularly where consuming and communicating dense flows of visual information is a routine aspect of work. In hospitals, effectively interpreting and promptly acting on networks of monitors, charts, x-rays, and other imagery is of vital concern to patient safety. The stakes heighten given the fact that healthcare professionals' perceptual skillsets often diverge widely. Nurse practitioners exchange metrics and orders with new interns; advanced radiologists interpret fluoroscopy scans alongside fresh technicians. Yet, all parties engage with the same visual interfaces.

The focus of our model is pulmonologists. After observing them in pulmonology suites, we reconstructed their perceptual repertoire in an effort to create educational resources for the less experienced. Learning to wield a bronchoscope involves trial and error. Understanding skilled perception in action can make the process more manageable.

Moreover, pulmonology suites are a useful context in which to test the limits of eye tracking. Even though the purpose of bronchoscopy is to visualize the airways for the sake of diagnostic or therapeutic interventions, we found it remarkable how integral sensory systems other than vision are to the procedure. One pulmonologist noted that with practice, he cultivated the tactile ability to know where he was in the airways without seeing them clearly on the bronchoscopic monitor. Pulmonologists see less than they perceive. The models used to interpret their pupil movement should, therefore, point beyond the visual field. More generally, the limits of eye-tracking evidence could reveal opportunities to delve deeper into peoples' perceptual skillsets.

EYE TRACKING TWO DIMENSIONS IN TRANSITION

Although fast-paced workplace environments like hospitals are ripe to be examined through the lens of eye trackers, the technology remains beholden to lingering methodological assumptions. These have persisted since its initial application over 70 years ago.[7] In 1947, Paul Fitts and his colleagues conducted pioneering eye-tracking studies in airplane cockpits.[8] A mounted camera filmed the pupil movements of 40 pilots as they sat and scanned the array of controls, buttons, and dials while flying C-45 planes. The study's objective was to determine how perceptual practices diverged among pilots with various skill levels. Two consequences followed. First, experienced pilots tended to dwell on instruments for less time. They exhibited shorter fixation durations, which Fitts took to mean that the experienced pilots retrieved information quicker than did their less experienced counterparts. Second, he calculated the frequency with which pilots shifted their gaze among instruments and divided the frequency by the total number of transitions. The result revealed which instruments were linked in the pilots' visual field. By evaluating divergences in perceptual expertise, Fitts and his team optimized the arrangement of instruments in cockpits to facilitate smooth decision-making processes. The study laid the foundations of modern usability research but also predisposed eye-tracking methods to two-dimensional visual fields.

Eye tracking has come a long way. Fitts and his colleagues manually classified pilots' eye movement after the fact using film reels. Mobile eye-trackers now sit atop observers' heads and automatically generate fixation data. The technology at work is pupil center corneal reflection. An infrared light illuminates the eye to bring about visible reflections. A small camera captures these reflections, which identify the light source on the cornea and in the pupil. Software calculates the vector of the angles of the reflection patterns between the cornea and pupil; the vector is then used to determine the direction of the observer's gaze. Discrete time stamps with x/y coordinates are assigned to these movements—what are called "gaze points."

Unto themselves, gaze points offer little information. They become meaningful once rendered fixations. When an observer looks over a visual plane, his or her eyes move rapidly before settling on certain points. That moment when the eyes halt their movement and achieve relative stability is a fixation. Fixations are quick. They last between 50 and 600 milliseconds. They're also a useful unit of analysis because fixations serve as indications of information consumption. They mark the minimum threshold of ocular stability necessary for an observer to notice visual content. Eye-tracking systems include a fixation filter, which uses an algorithm to translate gaze points into fixations. What's left out are saccades: the eyes' motion between one fixation and the next. The average duration of a saccade is 20 to 40 milliseconds.

Fixations are made useful when correlated with the space where they're directed. Researchers take a sample image of the visual interface and divide it into areas of interest. Each corresponds to a discrete region of the surface where one's gaze might fall. An auto-mapping program is then used to calculate the sequence, frequency, and duration of fixations across the various areas of interest. What results is a bank of data falling under categories such as fixation duration, visit count, average visit duration, and time to first fixation.

Although technologies have evolved since the time of Fitts' study, much remains the same. Eye-tracking evidence did not directly entail either of Fitts' conclusions. He inferred the meaning of pilots' perceptual activity from their pupils' movement. Fitts correlated fixation duration with the difficulty of information extraction and fixation frequency with the importance of the area seen. Today as well, eye-tracking evidence is only as valuable as the inferences we draw. When contexts shift, our categories do as well. While exploring an art museum, for example, my lingering gaze would likely be indicative of the paintings' appeal—not their complexity. Eye-tracking metrics do not speak for themselves.

Two-dimensional eye tracking is useful in certain aspects of medicine and healthcare. At Design Science, the technology allows us to collect ocular data and to draw inferences about the usability of medical devices. We use the Tobii Pro Glasses 2. This mobile eye tracker rests on a study participant's nose and ears like normal glasses (Figure 1). The glasses weigh 45 grams and include four sensors emitting infrared light to the retinas at a 100-Hz sampling rate. Along the bridge is a high-definition scene camera directed outward to the visual field; its recording angles reach 82° horizontally and 52° vertically. The infrared reflections are superimposed on the camera footage to generate our eye-tracking metrics.

It's worth noting that the recording angles as well as the pupil center corneal reflection technology are artificial apparatuses. Their purpose is to focus on the center of the visual field. The ocular movement tracked is foveal. This is the region of the eye extending one degree from the center (the macula), which contains cone photoreceptors governing directional vision of objects. Eliminated from the eye-tracking sensors are para-foveal vision (extending to five degrees from the center) as well as peripheral vision (occupying the retina beyond five degrees). By isolating foveal vision from the natural spectrum of vision, eye trackers can be used to analyze separate fixations, each of which lands on identifiable content.

Figure 1. IFU usability study with simulated task completion.

As an example of our usability research, we use the Tobii Pro Glasses 2 to examine and help optimize Instructions for Use (IFUs). Medical devices and products are sold with IFUs, which inform users and lay caregivers about proper uses, risks, and benefits. In the United States, the Food and Drug Administration requires that all instructions and labels pass usability testing to certify that they are easily readable.[9] This is a service we offer. During IFU studies, we observe users as they read an IFU and carry out simulated tasks. Perhaps the participant administers an injection. Our researchers observe participants with an eye to use errors. We then ask follow-up questions to ascertain their root cause. Arriving at the fundamental causes underlying errors is crucial to finding design remedies that would correct possible future use errors. With eye tracking, we can plunge even deeper and ask questions whose answers often elude traditional root-cause analysis. Which elements of an IFU stand out to users? Do they ignore, read, or skim certain elements? Do users rely more on images or text to comprehend the IFU? (Figures 2 and 3 below indicate the latter.) What do users observe when performing critical steps? Our reiterative research process serves to evaluate and compare multiple versions of IFUs. It thus becomes easier to eliminate design flaws and to reorganize IFUs' visual architecture for intuitive reading.

After conducting IFU comprehension studies, we use Tobii Pro Lab's automapping program to generate a wide array of eye-tracking data. First, a sample image of the IFU is divided into areas of interest, each corresponding to distinct tasks (and further divided into images and text). Second, the program calculates correlations between each area and participants' fixations. We've found three kinds of data to be useful: fixation duration metrics (indicating the distribution of fixations on each area), gaze pathways (visualizing the sequence in which users read IFU sections), and heat maps (showing which areas solicit users' fixations).

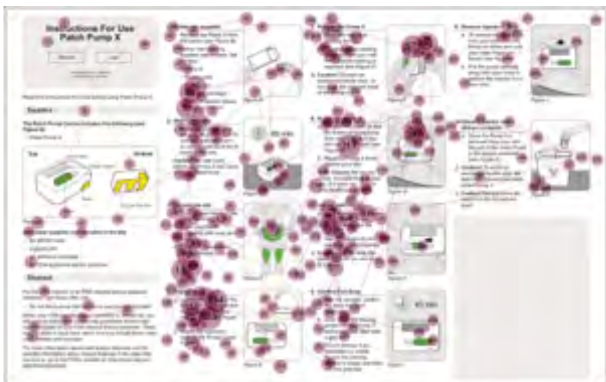


Figure 2. IFU gaze pathway

Figure 3. IFU heat map

Context is key. The data reveal little by themselves. By focusing on participants' intentional tasks, such as preparing the skin for injection, our researchers draw probabilistic inferences from eye-tracking data. Whether a participant fixates on certain IFU areas, and – more importantly – passes over others, is more likely to be indicative of perceptual acts of attention – and not, for instance, curiosity or distraction. Addressing goal-oriented ocular activity allows us to deliver tailored design recommendations. We might reorganize texts and images, add visual prominence at critical steps, or highlight overlooked areas. The end products are more intuitive IFUs that meet FDA guidelines.

TRACKING EYES IN IMMERSIVE CONTEXTS

It may seem that eye tracking is best suited for narrow scenarios in which the space of action and perception comprises only two planes. I'd like to suggest, instead, that the technology is pliable. Its utility hinges on the methods and settings brought to eye-tracking studies. Adapting eye trackers to ethnographic fieldwork turns in large part on researchers' willingness to eschew the assumptions of two-dimensional user research. The interfaces examined in much of that research is accessible at a glance – as if the interface were projected from a stable point. Participants also remain stationary while observing. This setup is alien to the immersive and interactive contexts in which ethnographic inquiry unfolds. Those contexts are three-dimensional and unstable. Therein, perception operates via active exploration of the environment and not via passive observation of a surface. People do not look *at* the visual field. A goal of eye tracking in ethnographic research is to understand how people situate themselves *in* the field.

Adapting eye trackers to immersive contexts is not seamless. The technology has proved to be congenial to two-dimensional interfaces in large part because researchers rely (often implicitly) on two interlocking theories of perception. The first is that perceptions derive from sense data; the second is that perceptions constitute mental representations. According to the first theory, the sensible qualities of objects explain the content of perceptual experience.[10] The shape of the font on an IFU or the color of dials in a cockpit all comprise sense data. One is directly aware of each datum. Moreover, sense-data theory counts among perceptual experience only those qualities directly sensed. That I don't see the reverse side of the IFU or a lever hidden by the dial means that whatever lies occluded before my eyes is not a sense datum.[11] The theory accords so well with the experimental setup of much usability research because automapping programs measure individual fixations in so far as they correlate with what is immediately accessible to vision. Just as sense data are interposed between objects in the world and visual experience, so too do areas of interest on a sample image serve to bridge the visual surface and the participant's perceptual activity.

According to the second theory, perceptions are episodes in the mind which correspond to (and thereby represent) a state of affairs. It follows that when we perceive objects in the world, we perceive them by way of representations. A representation is typically thought to be a sensory state. For the weak version of the theory, this sensory state is a relation one has to objects – my representation, for example, is of the white color in the IFU's background and of the black font in the foreground.[12] For the strong version, a sensory state is a mental item with semantic properties – e.g., my visual perception "that there is a fourth and fifth step in the IFU."[13] Both versions construe representations as pictures projected before vision. And much like a picture, representations include contents which one can describe. The theory nicely accommodates two-dimensional eye-tracking research because fixations function like representations. Researchers use fixations to infer referential acts of attention from the pupils' movement.

Plenty of ink has been spilled contesting both sense-data and representational theories of perception. Resolving those disputes is not my aim. Rather, I'd like to identify and elaborate on theories of perception more appropriate to the methodologies and set up unique to eye tracking in ethnographic research. Sense-data and representational theories inform much of current eye-tracking studies because those theories construe perception in analogous terms. Both take visual perception to function like a camera. For the sense-data theorist, pictures are projected onto our field of view. For the representational theorist, vision is akin to looking at pictures. In either case, visual perception is figured as if it occupied a plane orthogonal to an observer's point of view. Ethnographers would find this approach unbefitting to their work. When studying immersive contexts, our subjects do not see what is projected from a specific place; they see what is accessible in an entire space.

Over the past half-century, psychologists, philosophers, and neuroscientists have contributed to a constellation of theories which posit visual perception as an active engagement with the environment. Taking as their point of departure that perceivers are embedded in an interactive context, these theories claim that vision unfolds in two directions. On the one hand, people rely on sensorimotor and cognitive skills to perceive a space visually. The eyes' vision is of a piece with the body's movement. As James Gibson succinctly wrote, "we must perceive in order to move, but we must also move in order to perceive."[14] On the other hand, space is made accessible only partly due to vision. Such is obviously true in the trivial sense that a variety of sensory systems (i.e., tactile, auditory, and proprioceptive) also facilitate spatial navigation. More importantly, one often subdues some visual elements in order to perceive others in the visual field effectively. Maurice Merleau-Ponty put it thus, "To see the object, it is necessary not to see the play of shadows and light around it."[15] In both cases, vision does not happen *to* someone; people make the environment visible. In neither case does perceptual activity take place, as it were, between the ears.

If we take seriously the idea that visual perception is a form of action in the environment, then a few implications follow. First, what people see depends on where their bodies stand. Given that vision reflects a local perspective in space, how someone positions his or herself is integral to the constitution of the visual field. Whereas superficial eye-tracking studies control participants' perspective by stabilizing their position, ethnographic eye-tracking research is an inescapably perspectival affair. We might, therefore, consider eye trackers to record only a part of vision—namely, the trace left by peoples' embodied movement. Second, people move their bodies in order to act on the environment. In making use of the environment, it appears not simply as a stable source of sensory stimuli, but as a malleable resource which serves our purposes. Visual perception is goal-oriented in the sense that it consists of more than what is actually visible. It also includes what is potentially visible. One moves through space sometimes to gather information, but more often to change the space. Eye-tracking video could therefore be used to reveal how subjects actively transform the visual field. Third, the utility found in the visual field depends on the perceptual skills brought to it. What is useful might not be immediately visible. Objects' color, shape, or location often obfuscate vision. How people alter the visual field to their advantage hinges on a repertoire of habitual, cognitive, and sensory techniques. Alva Noë puts it thus: "Perceptual experience just is a mode of skillful exploration of the world."[16]

Eye trackers offer one tool for visualizing the multi-faceted skillsets with which people explore the environment—particularly in medical facilities where visual displays often require years of training to comprehend. Visual perception is incomplete. It involves competencies and sensory systems that lie beyond what can be seen. Yet, that is not at all a reason to dispense with eye tracking. The technology can supplement the tools already available to ethnographers, including close observation, behavior mapping, and participant interviews. Eye tracking might be most useful, in fact, where vision cedes to other modalities of spatial navigation.

Consequently, some of the data categories used in conventional eye-tracking research are inadequate to ethnographic fieldwork. Since they are not stable, immersive contexts cannot easily be divided into fixed areas of interest. So, tracking fixations' duration and frequency is often a baffling task.

Nonetheless, gaze pathways, heat maps, and stability tracking can offer useful evidence. Gaze pathways reveal the sequence in which people make use of visual resources. Visual experience does not come together at once. Tracing how people gradually maneuver a space over time can shed light on the activities by which they make the environment useful. Heat maps can also be useful when the visual field remains relatively stable. These visualizations indicate not only where vision, but also where the body, moves in space. In addition, researchers can track the stability of pupils' gaze, which often represents the level of expertise involved in visual perception. Those familiar with the environment exhibit smooth ocular movements. By contrast, erratic ocular movements tend to be found among perceivers less skilled in discerning, retrieving, and anticipating visual resources.[17] Data for neither category are generated automatically. Gaze pathway and stability offer two frames of reference through which to analyze eye-tracking recordings.

Far from revealing a hidden stratum of subjective experience, eye tracking actually offers ethnographers another context to observe, describe, and interpret. The inferences drawn from the evidence are only as accurate as the tools ethnographers already bring to their fieldwork. As I suggest in the following section, eye tracking offers a fruitful context when it shows not only peoples' visual navigation of space, but also the limits of their visual perception.

PERCEPTION AND ACTION IN BRONCHOSCOPY

The setting is a pulmonology suite in a research hospital in the Northeast United States. It's a busy environment. Below, webs of wires are taped to the floor so that nobody trips. An anesthesiology workstation sits in the far left corner; adjacent is a mobile nursing station with two screens on which patients' vital signs flow from left to right; cabinets of medical supplies line the walls; a back table where instruments are prepared extends across the right side; near us stands a four-foot tall bronchoscopy power station. It runs wires to a large wall-mounted monitor. This is the focal point of the room. Everyone's gaze is directed at it. The monitor visualizes the interior of the respiratory passage. It is the portal through which the medical team will navigate their way to a cluster of lesions in the patient's upper left lung, which are suspected to be cancerous.

About 230,000 Americans are diagnosed with lung cancer each year.[18] A bronchoscopic lung biopsy is a standard diagnostic procedure. But it is not straightforward. Bronchoscopes are long, flexible devices inserted through the throat. The device includes a white-light probe, which produces a live image of the airways on the monitor. Today, the image will help orient the pulmonologist to the lesions, which are challenging to reach at the edge of the mediastinum (the central region of the thoracic cavity containing the heart, trachea, and lymph nodes). She will proceed to insert a 22-gauge fine needle aspirate through the bronchoscope's inner tube and remove small tissue samples from the lesions. They can be moving targets. When the patient coughs, his lungs move – making the procedure no easier. Finally, the samples will be handed over to a cytopathologist, whose job is to analyze them under a microscope and detect malignancies.

The objective of our observations is to understand the pulmonologist's navigational technique. Guiding a bronchoscope through the airways is a delicate process. It demands not only that the pulmonologist perceives visual imagery on the monitor but also that she manually contorts the bronchoscope's handle, twisting her wrist and shifting her body, while feeling tactile impressions of the bronchial walls. The pulmonologist is trying to see inside the patient's airways in order to help him. But she doesn't rely on sight alone. Pulmonologists develop a muscular memory of the airways' anatomy. The monitor is there to offer verification.

Understanding this perceptual skillset would allow us to make bronchscopic expertise transparent, particularly for the sake of those learning the practice. In this respect, ethnography offers a useful educational tool, which imparts concepts to the pulmonologist's otherwise immediate perceptual process—what Hubert Dreyfus calls "absorbed coping." Learners' apprenticeship is cognitive; they rely on explicit rules to comprehend new techniques. Experts don't. Dreyfus elaborates, "although many forms of expertise pass through a stage in which one needs reasons to guide action, after much involved experience the learner develops a way of coping in which reasoning plays no role. Then, instead of relying on rules and standards to decide on and justify her actions, the expert immediately responds to the current concrete situation."[19] Our aim is to develop a hypothetical eye-tracking model that could disentangle the mélange of perceptual skills. What feels natural to the expert pulmonologist might thus be made deliberate for the fellow.

The procedure is about to begin. An elderly man lying on his back in a mobile patient bed is rolled into the room through doublewide doors. Casual banter with the pulmonologist sets him at ease while the nurse anesthetist administers moderate sedation. Semi-conscious, the patient is intubated and the bronchoscope soon makes its way into his mouth, over his tongue, and down the trachea. An endoscopic image appears on the wall-mounted monitor. It depicts bronchial rings sliding downward as the bronchoscope advances toward the left main bronchus. The pulmonologist's eyes remain glued to the monitor. She confirms that the left upper lobe is near, about two centimeters past the main carina.

Outside her view, the pulmonologist's hands do all of the work. They rotate left then twist right to orient the bronchoscope's tip. At one point, she shifts her body 90 degrees; her right shoulder now faces the monitor. The pulmonologist's own body is the vehicle by which she comprehends that of the patient.[20]

Once the lesion comes into view, its time to activate the ultrasonic probe. The bronchoscope is equipped with endobronchial ultrasound (EBUS), which visualizes structures beyond the airways walls. This second image allows the pulmonologist to see what the white-light probe can't: inside the lesion. She passes a needle into the bronchoscope and attempts to drive it through the side of the airway and into the lesion. But the cartilage lining the airway is thick. The needle deflects. Feedback appears on the monitor, distorting the ultrasonic image (Figure 4).

Figure 4. EBUS monitor with ultrasonic image in the center (with feedback) and endoscopic image in bottom left

The respiratory technician rushes over. He holds the base of the bronchoscope, just above the patient's mouth, helping the pulmonologist to secure the ultrasonic probe in place. Four hands now grip the device. Their manual contortions give rise to what appears in the monitor's dual images. The imagery connects the movement of the medical team with that of the patient. The situation makes literal Merleau-Ponty's words: "there is an immediate equivalence between the orientation of the visual field and the awareness of one's own body as the potentiality of that field."[21] Although the two images on the screen are of the patient's bronchus, they also lay bare the teams' collaborative corporeal efforts to drive through it.

Puncturing the lesion is the most fragile part of the procedure. Pushing the needle too far could result in trauma. So, the pulmonologist gently feathers the needle while verifying its placement on the monitor. The task demands immense attention. She shifts her gaze between two images: the ultrasonic image of the lesion in the center and the endoscopic image of the airway in the bottom left. Her eyes also look elsewhere.

Our simulated eye-tracking model, generated thanks to Tobii Pro Lab's automapping program, reconstructs a triangular pattern: fixations transition between the dual images and outward beyond the left edge of the monitor (Figure 5).

Why would a pulmonologist look away from the monitor, especially in moments that demand heightened attention? What is taking place when the gaze in the model repeatedly shifts outward?

Following the model, the optical data are valuable for what they reveal as much as conceal. Although the heat map indicates the regions of the visual field where the pupils fixate, those regions also indicate sensory systems other than vision. Our inference is that the gaze would shift away from the monitor as a pulmonologist shifts her sensory emphasis from visual to tactile perception. She uses her own haptic impression of the bronchi to guide the needle through the cartilage and into the lesion. It's not so much the she looks at nothing. She feels something. She directs her attention to the feeling of the bronchial anatomy in her grip of the bronchoscope's handle.

Figure 5. Hypothetical model of heat map over EBUS monitor

More specifically, the eyes' trajectory aligns with the body's movement. The pulmonologist shifts her stance left. The hypothetical gaze pathway reconstructed in Figure 6 illustrates this spatial realignment. Gaze plots on the left mark where she would turn laterally, moving her body toward the left of the monitor. To be sure, my interpretation does not obtain on the basis of the gaze pathway alone. I also observed the pulmonologist in my role as ethnographer. She looked at the two-dimensional image in front of her while moving in the three-dimensional space around her. Both are perceptual skillsets essential to performing lung biopsies.

Figure 6. Hypothetical model of gaze pathway over EBUS monitor

What results is a multi-sensory circuit activating visual and tactile perception. Sliding her focus between either, the pulmonologist orients her own body via two spatial configurations: those of the room and of the patient's airways. In the model, my naked eye supplements the eye-tracking data; both offer an occasion to pull apart the pulmonologist's entangled perceptual skillsets. In a follow-up interview, she elaborated that the skillset involves a combination of tactility and vision. What matters when initially performing bronchoscopy is one's ability to know how the device works and to know how one needs to move to make it go where it needs to go. One uses both the visual cues and the hand's movement. In short, it's seeing and feeling.

Touching and seeing flow together. Yet, eye-tracking evidence makes apparent that they do not always flow proportionately. The evidence illuminates when and where pulmonologists' deploy sophisticated kinesthetic competencies, shifting perceptual systems back and forth. Conventional ethnographic methods of observation and participant interviews fill in the blanks, as it were, and reveal the pathway to mastery behind the medical art. Another pulmonologist tells me that all of the incoming first-year fellows go through a two-hour session about how to hold and manipulate the bronchoscope, and how to watch the monitor. Everyone learns to use his or her control arm and support arm. And in the beginning, the fellows usually grip the device too hard because they're overly focused on the monitor. Vision shrouds feeling.

Pulmonologists develop supple techniques over time before they learn to collaborate with medical teams. The intricate skillsets at work in a lung biopsy is a strong reason why the procedure likely could not be carried out by machines. Eye tracking offers one tool to help ethnographers document these skillsets. The findings provide insights for people and organizations (e.g., device manufacturers, medical schools, and value analysis committees) to improve the ensemble of coordinated practices involved in bronchoscopy.

CONCLUSION: PUSHING THE BOUNDS OF ETHNOGRAPHIC EYE TRACKING

Although challenging, adapting devices meant for recording individuals' vision of two-dimensional planes to interpersonal and immersive contexts with three-dimensional visual interfaces is not only feasible; it's also useful for ethnographers. Workplace ethnography broaches many questions. Eye trackers are not dispositive. In fact, they're incomplete when used in insolation. But when tailored to address specific questions about practices of looking, feeling, and acting, the technology offers a launching pad from with to hone ethnographers' lines of inquiry. I hope to have offered one ensemble of methods and a modest model for doing so. We continue to refine our model at Design Science in order, eventually, to integrate eye trackers in actual medical facilities.

The contexts where ethnographic fieldwork takes place are quite unlike the cockpits in which Fitts and his team had originally studied pupil movement in 1947. They configured the cockpit along two planes while participants remained seated. Fast-paced workplaces are nothing of the sort. That ethnographers might be lucky enough to find people willing to wear eye trackers while going about their business is hardly given. The devices are cumbersome when worn by busy healthcare professionals communicating information and making quick decisions in medical facilities. Nevertheless, the questions that Fitts had posed half a century ago – namely, how does expertise inflect perception and action? – remain just as relevant today. Across collegial environments, peoples' disparate perceptual skillsets contribute to collaborative activities.

Eye trackers offer an additional perspective in such contexts. The devices visualize what escapes ethnographers' naked eyes. One of our aims has been to show how eye trackers could also be useful for their limitations, which expand opportunities to probe perspectives other than the eyes'. Clarity and opacity are both valuable sources of evidence. Ethnographers might thus delve into and analyze expert practices whose sensory modalities remain caught, as it were, in peoples' blind spots.

NOTES

Acknowledgments – An invaluable interlocutor throughout this article's revisions was Rebekah Park; I am also indebted to Christina Stefan for her editorial acumen and to Ranjan Nayyar for his persistent insights into human factors research.

1. See Hubert L. Dreyfus and Stuart E. Dreyfus, *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (New York: Free Press, 1988).

2. "What Does a Pianist See?," *Fractal Media: Function.* Published Mar. 18, 2017. https://www.youtube.com/watch?v=GVvY8KfXXgE.

3. D.L. Mann, W. Spratford, and B. Abernethy B. "The Head Tracks and Gaze Predicts: How the World's Best Batters Hit a Ball," *PLoS ONE* 8, no. 3 (2013): e58289.

4. Consider Clifford Geertz's pithy yet canonical quip, "Culture is public because meaning is." *The Interpretation of Cultures* (New York: Basic Books, 1973), 12.

5. Charles Goodwin and Marjorie Goodwin, "Seeing as Situated Activity: Formulating Planes," in *Cognition and Communication at Work*, ed. Y. Engeström and D. Middleton (Cambridge: Cambridge University Press, 1996), 61-95.

6. Lucy Suchman, "Centers of Coordination: A Case and Some Themes," in *Discourse, Tools, and Reasoning: Essays on Situated Cognition*, ed. L.B. Resnick, R. Säljö, R., C. Pontecorvo, and B. Burge (Berlin: Springer Verlag, 1996), 41-62.

7. See E. Sumartojo, A. Dyer, J. Garcia Mendoza, and E. Gomez, "Ethnography through the digital eye: what do we see when we look?" in *Refiguring Techniques in Digital Visual Research*, eds. Edgar Gómez Cruz, Shanti Sumartojo, Sarah Pink (Switzerland: Springer, 2017), 67-80.

8. Paul Fitts, et al., "Eye movements of aircraft pilots during instrument-landing approaches," *Aeronautical Engineering Review* 9, no. 2 (1950): 24-29.

9. See U.S. Department of Health and Human Services: Food and Drug Administration, "Guidance on Medical Device Patient Labeling; Final Guidance for Industry and FDA Reviewers" (2001), page online: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm070801.pdf

10. Sense-data theory has its roots in the work of G.E. Moore; see *Some Main Problems of Philosophy* (London: George, Allen and Unwin, 1953).

11. One may, of course, be aware that the IFU has a reverse side. Sense-data theorists argue that one sees it indirectly – that is, in virtue of the side that one sees directly. See Frank Jackson, *Perception: A Representative Theory* (Cambridge: Cambridge University Press, 1977).

12. For the classical formulation, see Franz Brentano, *Psychologie vom empirischen Standpunkt* (Leipzig: Duncker & Humblot, 1874).

13. See Fred Dretske, *Perception, Knowledge and Belief* (Cambridge: Cambridge University Press, 2000); Jerry Fodor, *Representations* (Cambridge, MA: The MIT Press, 1981).

14. James J. Gibson, *The Ecological Approach to Visual Perception* (Hillsdale, NJ: Lawrence Earlbaum, 1979), 223.

15. Maurice Merleau-Ponty, *The Primacy of Perception: and Other Essays on Phenomenological Psychology, the Philosophy of Art, History and Politics*, trans. James Edie. (Evanston, IL: Northwestern University Press, 1964), 167.

16. Alva Noë, *Action in Perception* (Cambridge, MA: The MIT Press, 2004), 194.

17. Shane L. Rogers, et al. "Using dual eye tracking to uncover personal gaze patterns during social interaction," *Scientific Reports* 8, no. 4271 (2018). Page online. DOI: 10.1038/s41598-018-22726-7.

18. American Cancer Society, "Key Statistics for Lung Cancer." Page revised Jan. 4, 2018. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html

19. Hubert L. Dreyfus, "Overcoming the Myth of the Mental," *Topoi* 25 (2006): 47.

20. A procedural guide distinguishes the airways' anatomy according to the tactile impression transmitted via the bronchoscope. For example, "The posterior part of the trachea is vertically structured (resembling long vertical plies) and easy to distinguish from the arch-shaped cartilage structures (the horizontal support of the bronchial tree)." See, Felix Hirth, et al. "Endobronchial Ultrasound-guided Transbronchial Needle Aspiration." *Journal of Bronchology* 13, no. 2 (2006), 13.

21. Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. Colin Smith (London: Routledge, 1962), 239.

Evidence outside the Frame: Interpreting Participants' "Framing" of Information when Using Participatory Photography

TABITHA STEAGER
Pacific AIDS Network

This paper discusses the benefits and challenges of participatory photography as ethnographic evidence and how as researchers we can "read" the evidence our participants create. Drawing on examples from an ethnographic study examining concepts and constructions of community on Salt Spring Island, British Columbia, I examine how we can interrogate photographs as data rather than factual evidence. Adages such as "the camera doesn't lie" support the view of photography as a purveyor of truth. Photos accompanying journalistic dispatches from far-flung outposts around the world are seen as authentic evidence of real-world situations. Amateur videos of people's life experiences are filmed on smart phones and then posted to YouTube to be taken as authentic representations of life events. Early ethnographic uses celebrated photography as the ultimate tool for showing that anthropologists had actually "been there," displaying the exoticism of other cultures in factual black and white. However, photography has never been a simple representation of the truth—it is not cameras that make photos but people, with all their personal quirks, cultural beliefs, and subjective experiences in tow. Photographs always provide at least two kinds of evidence—what is inside the frame and what is outside the frame. As researchers working with participatory photography, one of our roles is to determine the importance of what is outside the frame. We must ask whether this unseen evidence is as valid—or more so?—than what participants keep inside the frame. In the age of Snapchat, Instagram and other social media, researchers need to interrogate participant-created photography carefully and methodically. We must question how we interpret photographic evidence that has been manipulated by its creators and how that manipulation affects our interpretations of the evidence. Participant-created photographs add valuable depth and complexity to ethnographic research but we need to ask how participants may conceptualize their photographic creations and how context—culture, socioeconomic status, gender, location, etc—impacts the evidence participants create. And in turn, how those same contexts influence our interpretations of participatory photographic evidence.

INTRODUCTION: A (VERY) SHORT HISTORY OF PHOTOGRAPHY IN ANTHROPOLOGY

Early ethnographic uses of photography celebrated it as the ultimate tool for showing that anthropologists had actually "been there," displaying the exoticism of other cultures in factual black and white. Anthropologists like Franz Boas and his student Margaret Mead saw photographs as a way to capture factual records of cultures that were going extinct. Visual recordings, whether still photographs or film, could salvage and preserve for posterity a cultural ceremony or way of doing something, such as a demonstration of Torres Strait fire making or Balinese parenting styles. These visual records would then "act as a template for the process, allowing it to be reproduced, rather like following an instruction manual. Visual recording 'saved' the event in some reified sense" (MacDougall 1997:282). This viewpoint has changed dramatically and anthropology's relationship with photography and film, and the continued evolution of visual anthropology, has been well addressed across anthropological and sociological literature (see, for example, Edwards 2015, Pink 2007, and MacDougall 1997). For my purposes here, what is important to remember is that photography has never been a simple representation of the truth—it is not cameras that make photos but people, with all their personal quirks, cultural beliefs, and subjective experiences in tow. Photographs are evidence of how people feel, what they think about and how they think about it, as well as how they make their way throughout their worlds (Edwards 2015).

Before I go farther it is important to note why I limit my discussion here to photographs rather than photographs and film. As I noted above, both photographs and film have a long history in anthropological research. The reasons I focus on photography are not due to some deep theoretical belief but instead merely practical—while I admire and am fascinated by the use of video and film in ethnographic research, I have not actually done it myself. I have analyzed films and used them to teach and discuss anthropological concepts and theories but I haven't used the medium myself in my research. I choose to focus on photography because that is where my experience lies. For broad discussions on film and video in ethnographic research that are well beyond what space and my understanding allow, see, for example, MacDougall 1997 and particularly Barbash and Taylor's 1996 conversation with Judith and David MacDougall about ethnographic film methodology.

INTERROGATING PHOTOS AS DATA VERSUS FACTUAL RECORDS

Consider this iconic photograph (Figure 1) of one of anthropology's founding fathers, Bronislaw Malinowski, during field work in the Trobriand Islands in 1918. For years anthropology students and others likely took this image at face value—evidence of the eminent anthropologist in the field, studying the "exotic" Trobrianders. Later images like this would be critiqued as proof of colonialism and its associated evils. Even critiques like this take the photo at face value, as evidence of a particular truth and that the image is a *factual* record.

As Alex Golub so succinctly put it in a short analysis of this photo, "If there's one picture that epitomizes White Guys Doing Research, it's this one. The canonical author of the canonical book, naked black people, white guy in white clothes being White" (Golub 2017). However, if we take a step back and instead of considering the photograph as a factual record of some kind but instead data, an entire world of inquiry opens up. Rather than taking it as a record of some truth (an example of colonialism of early anthropology or a record of authentic Trobriand Island drinking vessels) we can interrogate not only the data contained in the photograph but also the data that is outside the frame—in what circumstances the photo might have been made and what Malinowki's thought process might have been. Although we can no longer ask Malinowksi about the photograph and the making of it, we are fortunate to have Malinowki's diaries and field notes, as well as interviews with field research assistants and others who knew him to give us insight into the man and what he might have been thinking when crafting a photograph like this. As Golub further reflects,

Figure 1. Bronislaw Malinowski and Trobriand Islander men. Courtesy University College London archives.

> The thing that most people don't get about this picture is that at least 30% of it is *cosplay*. What surprises me about this image is that many people view it without any sense of irony—as if it had not been posed, as if Malinowski didn't notice the difference between himself and 'the natives', as if Malinowski was unaware of what his lime spatula looked like (emphasis in original, Golub 2017)

When viewed from this perspective, the photograph becomes much more than "factual" evidence but evidence we can examine for the maker's thought process, the positioning of White male researchers in 1918, Malinowski's personality, and more. We can also interrogate it for evidence of contemporary thought and reactions to photographs made during the peak of colonialism. When we think of an image in these terms, a photograph is worth more than 1,000 words. Of course, photographs have never been simply factual records, even though some anthropologists at one time may have thought of them as just that. They've always been data as well as a record that someone was in the place and time with the people they said they were. Indeed, we interrogate old photographs as well as contemporary Instagram posts for evidence in our research on a regular basis. However, there is, I think, an important distinction between the conclusions we make when reviewing photographs in absence of their makers versus in concert with their makers. When the makers (our research participants) themselves explain their photographs and the motivations and meanings behind them to us, we very clearly and deliberately put the research participant—the knowledge holder—at the center of the frame in our inquiries.

PARTICIPATORY PHOTOGRAPHY AND ETHNOGRAPHIC EVIDENCE

Ethnography, if we break the word down, literally means writing about people, so it might make sense to wonder how we can do "ethnography" without words, but with pictures instead. The fact is, ethnography is about observation, and our vision plays a vital role in that observation. However, when it comes time to analyze our observations, we turn to words and theories about what we have seen, or at the most micro level we explore theories about words. However, this kind of analysis belies the fact that our observations in many cases were made with our visual faculties. How, then, to share what our eyes took in and our brain and psyche processed? How do we know if what we see is the same as what someone else sees? Is the blue of the sky I see the same as the one the woman with the long black hair sitting across the way sees? Consider an argument my mother and I had over several years about the color salmon. I argued that it was a pinkish, rose-hued color. She argued it included much more orange than that. She is an artist, so she had artistic color theory on her side, but I just couldn't understand how she could say the color salmon was orange-ish. It's the color of a piece of salmon, which is definitely not orange; I wondered how she could possibly make the argument when we were looking at the same color. Well, a few years later she had eye surgery and lo and behold, she came to me and said that I was right, salmon was in fact pinkish and didn't have any orange in it at all. It turns out we were seeing a different color when standing side by side looking at it—the early cataracts she had had created a yellow film over everything she saw. It is my contention that although we assume, when reading, that we are reading "the Truth," or at the very least the author's truth, there is always a sense of fiction within any account. The brain and memory is a complex, fickle, and creative thing. As readers, whether we are reading words or images, we impose our own interpretations and our own vision, if you will, on the author's creation. So, I can tell my story but each person who reads it is going to "see" that story differently based on their own experiences and opinions.

At its best, ethnography attempts to get at that very personal, internal, embodied experience of both participants and researchers. Incorporating participatory visual research methods into research is one way to get at those deeper, hidden meanings. By using visual materials—in the case of this paper, photographs—produced by participants we create a richer, more layered story than if research data is gleaned only from interviews and written observations. If Willy Wonka's smell-a-vision actually existed, that would be another fantastic enrichment to ethnographic research, but alas, no one seems to have developed it yet, so photographs will have to do. Photographs not only bring more richness and depth to our research evidence but they also facilitate relationships with people, helping to bring us closer to our research participants through shared visual stories (Edwards 2015:248). Photographs can bridge time and space and help us place ourselves into our research participants lives and experiences in a deeper way than only words could.

Details of My Participatory Photography Project

Most of the photographs I reference in this paper were made as part of a larger ethnographic study of community, place, and identity on Salt Spring Island, British Columbia conducted between June 2011 and August 2012. I went to Salt Spring to try to understand how people define, imagine, and create place, and why creating place matters. On an island known for its physical beauty as well as an "alternative" approach to social relations and politics, Salt Spring is also a place where economic development competes with environmental preservation, affluence intersects with poverty, and residents grapple with concepts of insider versus outsider. Definitions of Salt Spring Island are complex and often contested and competing, both between residents and between residents and outsiders. Evidence I gathered from traditional participant observation as well as that created directly by my participants with their photographs helped me to explore these contestations of place and examine how people mobilize various means to define Salt Spring Island for themselves.

My goal with the participatory photographic aspect of data collection was to recruit additional participants who were not part of my core group of participants, not only to widen my pool of research participants but to also test whether or not participatory photographic methods brought forward different types of data than simply traditional participant observation and interviewing. Initially, I put a call out on the Salt Spring Exchange, an electronic message board used on a regular basis by many residents, for participants, asking for people who wanted to tell, in pictures, their story of Salt Spring. I had ten people respond to my call. I sat down with each person who responded for about an hour initially, explaining my project and interviewing them about Salt Spring, how they came to live there, and their impressions of the island. I included these interviews in the final data for the research project, however, not all people who responded ended up participating fully in the photography aspect of the project. Two people took photos but I was not able to successfully schedule a second meeting with them to review their photos. Two people took part in the initial interview but opted not to participate further. Ultimately six participants completed photographs and agreed to share them with me for this research. I asked these participants to take photographs as if they were going to tell someone who had never been to the island about Salt Spring, their experiences on the island and what the island is like for them. The majority of my participants took their photographs over several months in late winter 2011/early spring 2012. After our initial meeting, I sent participants away—all but one with their own cameras—to take the photos on their own time. One participant didn't own a camera, so I supplied her with a disposable one.

In visual ethnography literature this approach—having participants provide pictures of a particular subject—is generally described as photo voice. However, that can encompass photographs that participants have made outside of a specific research project—for example, all the pictures they've taken of their dog since they day they adopted her rather than pictures of a specific research question, such as "can you document your dog and how your dog impacts your life?" Therefore, I like to think of it a bit differently and use the term participatory photography. Participatory to me implies an active role on the part of the research participant, which is not always the case with photo voice. Rather than imposing my presuppositions on the experience and telling my participants what photos they should make, I wanted the participants to lead the process, not only in what and how they chose to photograph but also within the interview process itself when we discussed their photos, so that they were working with me to build a shared understanding, of a shared experience, of place. As Sarah Pink notes,

> when informants take photographs for us the images they produce do not hold intrinsic meanings that we as researchers can extract from them… they are derived from photographic moments that were meaningful to the people who took the photographs …

> when our informant-photographers discuss these photographs they place them within new narratives and as such make them meaningful again (2007:91)

Thicker Description: The Benefits of Participatory Photography

The photos that make up my visual ethnography of Salt Spring are not just pretty pictures, however, but a concerted effort on my part to build a better, and to borrow from Clifford Geertz, thicker ethnography. An ethnography should feel real somehow, it should transport the reader to that place and time, much like a good novel. By incorporating my participants' photographs, I believe I created a richer, thicker description of Salt Spring than would exist without the photographs my participants made and the meaning we made from them. Some research shows the part of the brain responsible for processing visual information developed evolutionarily before the parts that process verbal data, meaning that "images [could] evoke deeper elements of human consciousness than do words; exchanges based on words alone utilize less of the brain's capacity than do exchanges in which the brain is processing images as well as words" (Harper 2002:13). Reflective of this hypothesis, the work with my research participants elicited what I would describe as more layered, more deeply thoughtful conversations when we were reviewing their photographs than when simply talking with no visual cues.

It also made a difference to my participants that they were sharing photographs they had made themselves, in an order and within a story they were driving. They all shared that they felt intimately involved in the research and that it was more interesting for them than had I just sat down and asked them questions. They also felt more in control of their story than had I been leading things more directly. Other researchers have experienced the same benefits with their research participants when using participatory methods. Elizabeth Faulkner and Alexandra Zafiroglu note that their video "participants experience a more heightened engagement in sharing their experiences with us than they do when we film, as they take an active role in constructing how they will be portrayed… They are creating intentional representations of who they want others to imagine them to be" (2010:117). In his work studying the Danish concept of *hygge*, Jonathan Bean found that bringing participants into his data collection process shifted the power dynamic and created an alliance of sorts to document evidence when he and his participants shared the task of filming their homes while discussing *hygge*. He notes that he "found the mere act of asking for assistance with data collection to subtly shift the power dynamic between researcher and participant. Upon reflection, I found that the researcher and participant became allied in their task of documentation" (2008:106). The fact that participants get to participate in the research process and shape evidence that is about their own experiences and the meaning they imbue those experiences with is an important one. Adding their voices in such a direct way lends credibility to research evidence because it is first-hand and participatory, not one-sided and viewed solely through the researcher's lens.

This shifting of power dynamics can be one of the greatest benefits of using participatory photography, if a researcher uses the method with that in mind and truly allows participants to drive the research process, especially during the interviews about their photographs. As Josh Packard discovered in his participatory photography work with homeless men in Nashville, going through every one of a participant's photographs and letting the participant drive the interview process allowed for information he hadn't expected to come out, information that gave more depth to his research (2008:68). I approached my interviews the same way and had the same experience of unexpected information arise. One of my participants, Tanya, didn't own a camera or smartphone so she used a disposable camera I provided. She included images of natural beauty and popular areas in town as many of my other participants did but she also some photos that were out of the ordinary, including one of the gas station.

Figure 2. By Tanya. Co-Op gas station, Ganges, 2011. Used with permission.

I was surprised to see the photo because I knew Tanya didn't drive but I still expected her to discuss something about the price of gas on the island and how people are always waiting in line at the gas station during busy times, but I left the space open for her to tell me why she took the photo and what it meant to her. Leaving that space open for her to drive the interview opened up an entire discussion about people who had access to cars on the island and who didn't, the hitchhiking tradition on the island and the various people she'd met and experiences she had had while hitchhiking. Tanya felt a sense of kinship with many islanders who also had to hitchhike to get around the island but it was a feeling, she said, that was becoming more and more fraught for her as time went on. As the island became more affluent, the gas station didn't just represent an area of life she chose not to participate in (having a vehicle) but was coming to represent a feeling of isolation and being an outsider on an island she had lived on for more than thirty years. Tanya lived in a small apartment and worked several jobs, fighting hard to make ends meet, a situation common to many people on the island, including, she said, her friends who worked at the gas station. She went on to describe the thin line between getting by and not being able to afford food that many people face on Salt Spring and how things had changed. Tanya described a past where this struggle felt more communal, where people shared extra food from their gardens on a regular basis and there were less people who fell through the cracks. As property values increased on the island and more new, affluent people moved in, Tanya felt less welcome and less at home. In fact, Tanya ended up moving away from the island shortly before I finished my research and has not returned. All of this information, contained in hours of conversation, came from a single photograph of the gas station. It was unexpected, rich, valuable evidence of Tanya's intimate, personal experiences and feelings that I never would have gotten had I gone into the interview with a specific set of questions that I was set on asking. By letting Tanya drive the conversation and create meaning for me, sharing her feelings of belonging and exclusion, struggle and empathy, as well as frustration and anger, I turned over power to her to tell her story in a way that made sense for her.

Some Challenges in Participatory Photography

Participatory photography has many benefits as an ethnographic method, but it is not without a variety of challenges. First and foremost from my perspective is the issue of confidentiality. Confidentiality is always of concern when doing any kind of research but when you introduce a camera into the mix it only becomes messier. When we make photos and videos as researchers, we think through the ethical implications of who and where we are photographing or filming. We have been trained to keep confidentiality at the forefront of our minds when doing research—we should know instinctively when it could be inappropriate to take a photo or when we need to ask more than once, to make sure people are completely comfortable and informed when giving their consent to be photographed. However, when we transform our participants into participatory research participants who are actively producing material we will use as data, we need to pay special attention to ethical considerations like confidentiality and appropriateness of who and where they are making photographs. We need to ask whether our participants might be taking photographs of people without those subjects knowing or whether we might be exposing something photographic subjects do not exposed, even unknowingly. Children are photogenic and often the most approachable of our subjects, but photographs of minors hold even more ethical dilemmas than those of adults. Before embarking on a participatory photography project it is important to have a well-defined set of ethical guidelines and to go through them in detail with each participant to make sure they understand them and will honor them. A discussion of the ethics involved in ethnographic photographs is a long and complex one but suffice to say here that I tend to make it a practice to err on the side of caution—if I'm unsure about how a participant would feel about the inclusion of a photograph or if I don't know explicitly that a person has given permission for their image to appear outside of the privacy of an interview, I do not share it. I include these practices in guidelines for my participants who are making photographs and ask them to sign an agreement stating they've understood and will follow the guidelines. Another approach to ethical guidelines is to develop them in concert with your participants. This can be extremely useful in making sure that people comply with ethical guidelines because they have had a hand in creating them and thus feel a stronger sense of agency and ownership in the research project. In addition, ethical considerations are not simply about protecting participants and other people they may include in their photos but also about the construction of knowledge. The selection of photographs, and which ones are safe to include, adds an additional layer to how meaning is constructed and who constructs that meaning.

Other challenges inherent in participatory photography include consistency and follow-through with your participants. As with most qualitative research, there are always going to be participants who do not show up for a scheduled interview or who drop out of contact. Participatory photography is no different and therefore it is important to make sure you determine the minimum number of participants you want for your study and recruit more people in case you lose some through attrition.

While losing participants along the way and then not having enough participants can be a challenge, so too can the sheer volume of photographs people make and want to share. With smart phones and digital cameras, participants are able to take huge numbers of images and can find it difficult to limit the number they want to use to tell their stories. If this is the case, it does offer an opportunity for exploring with participants why they have so many images but it is still important to try to put parameters around how many photos you want each participant to take so that you can realistically go through each one in a single interview, realizing that they may not stick to those numbers.

FINDING MEANING INSIDE AND OUTSIDE THE FRAME: INTERPRETING PARTICIPANT-CREATED PHOTOGRAPHS

Photography has a powerful claim to realism. Despite what we now know about photo retouching and editing, we still tell ourselves 'what you see is what you get,' especially when the photographs seem to show something as mundane as sheep in a field, the contents of a refrigerator or a parking lot filled with cars. As Ball and Smith write,

> photographs of people and things stand as evidence in a way that pure narrative cannot. In many senses, visual information of what the people and their world looks like provides harder and more immediate evidence than the written word; photographs can authenticate a research report in a way that words alone cannot (1992:9).

When a photograph has obviously been edited, with something like a Snapchat filter, we know to interrogate it further and to ask why someone chose that filter. However, when a photograph is not so obviously manipulated, we tend to take it at face value. Photographs serve as hard evidence and as researchers we use them to indicate that our research is authentic and an accurate reflection of reality. However, from a perspective such as this, photographs are not interrogated as data themselves, but rather as factual representations of time, space, and place. It is this unconscious use of photographs that I wish to question here, because our research participants (and as researchers creating photos ourselves) we choose to frame or crop in a particular way (and with easy photo editing apps and filters, we can also dramatically change a photo to reflect a wide array of feelings and perspectives). We choose one photo over another to help us tell our story and we place it in a specific way, in a specific order within the story we recount. As Ball and Smith so cogently remind us, "it is people and not cameras who take pictures…photographs are not unambiguous records of reality: The sense viewers make of them depends upon cultural assumptions, personal knowledge, and the context in which the picture is presented" (18).

Photographs are not direct, "truthful" records of reality and we cannot interpret them as such. They are instead someone's visual statement about the world they live in and their experience (Worth 1980:20). Images, particularly the photographs we make and share, uncover what is culturally and socially significant for us (Ball and Smith 1992; Worth 1980).

As we view images, it is useful to employ all the empathetic skills we have as ethnographers (and human beings) at our disposal. We must attempt to think about how we read the visual "story" our participants are presenting us with, even as we listen to their telling of it. We can go even further by thinking about how our personal "reading" of these images might be read by others and what that does to the story and the evidence or answers to a particular question it contains. The idea is a bit like Alice going down the rabbit hole, I grant you, but one that applies to all ethnography, I think, for ethnography is only as good as the story teller's ability to express in words their very personal, visceral experience and the reader's ability to hear and see that story themselves. However, when I sit down to write a paper or report or create a presentation to share my analysis, it is *my* construction and *my* organization of the photographs, so my mark is more firmly and patently on the story that follows than any one of my participants.

Sarah Pink argues that researchers "should attend not only to the internal 'meanings' of an image, but also to how the image was produced and how it is made meaningful by its viewers" (2003:186). For my purposes what is most important is to look at how my participants imbue their photos with meaning and, in turn, how those images are further imbued with meaning by my reading of the images and the stories behind their production. Like Sarah Pink, I do not see the written word as the only or as a superior form of ethnographic representation (2007:4). Images hold a very real power on their own, just as the written word does.

One of the major themes in both the participatory photography data collection and traditional ethnographic observation of my research on Salt Spring was the island's natural beauty. The island is, without a doubt, incredibly beautiful. It is a powerful natural beauty, one that attracts visitors from around the world and seduces people so that they never want to leave. During my conversations with my research participants, as they were telling me their stories of the island through their photographs, natural beauty came up time and time again and in photograph after photograph. Participants constructed a place where beauty was at the foundation of everything. Even while talking about people who seriously struggle to make a living, their descriptions of Salt Spring were still infused with adjectives of beauty and the notion that it is the natural physical beauty of the place that keeps people on the island even when life is difficult. The siren song of Salt Spring's natural beauty is like the mythical Siren's song in that it has a dark side as well. It is not simply natural, untouched beauty but its beauty is also a commodity to be consumed and used by locals and visitors alike.

Figure 3. By Dusty. Ruckle Park, 2011. Used with permission.

As Dusty told me, her hand moving to touch her heart and her voice a bit breathless,

> It's just beautiful. It's just the most beautiful place I've ever been. I can have a bad day and then I go for a walk and I'm reminded just how lucky I am to live here. I mean, look at this place. Just look. It's amazing. People come here and they see, I think the beauty, it, well, it somehow…the nature, it heals people, you know? I've lived here 30 … oh, wow, more than 30 years … and I'm still struck by it. It's part of who we all are here (Dusty, 14 June 2012).

My own field notes, even when describing the challenges of living on a small, isolated island, reference the wild and powerful beauty of the place:

> There have been wind warnings for the entire week. Life on an island – sailings get cancelled, people get stranded, either on this island or on Vancouver Island or on the mainland or vice versa because ferries can't sail because of the wind. Driving home tonight from book club, the power's out at the north end of the island, just north of the cinema. Giant trees have torn from the ground, taking down power lines in their path to fall across the road and fire trucks are out in force with their lights flashing. They're the only lights in what is otherwise complete darkness. The moon is new so there is no moonlight to guide you. The powerful wind has actually swept away the clouds that lined the sky earlier, so there's only starlight to see by, an inky black ceiling covered with a blanket of stars. I had to drive home along a different route one of the firemen directed me to, back along a tiny, narrow bumpy road filled with potholes and bumps from tree roots, slowly navigating my way around branches that fell, wondering if a tree was going to fall on me next. Living on an

> island teaches you to never underestimate the power of Mother Nature. (Field notes, 24 November 2011, Southey Point)

Dusty and I both were emotionally affected by the beauty of the island—one of the primary attractions glossy travel profiles and tourism advertising highlight—and it forms an essential part of our definitions of Salt Spring. While I was and continue to be aware of the irony in this and my role in reproducing Salt Spring within a dominant discourse, Dusty was not in her reflections of her photographs. For her, as for several other photography participants, the natural beauty of the island was so great that it almost stood outside any possible negative aspects of the island. She went on to describe her personal challenges living on the island with a husband suffering from dementia and how isolating it could feel, but that the natural beauty she woke up to every morning outside her bedroom window acted as a balm and calmed her even on the worst days. I could hear it in her voice, that breathlessness and a sigh as she described it and relaxed a bit more into her chair, drawing solace from the beauty in the photographs she was showing me and then gesturing towards the window, pausing to look outside for a moment before taking a deep breath and moving to the next photo.

The beauty of the island is a given and therefore not something that all but one of my participants questioned. Every single one of them cited the island's natural beauty as one of the primary drivers for settling on the island. However, only Julia recognized the irony in the seductiveness of the island's beauty, telling me,

> Everybody's like … Salt Spring's heaven, well it's not heaven, really, except in small bits. But when it is heaven, it is, totally…We used to live in Ganges and we overlooked a bit of the harbour and we would see people, walking back and forth to the Harbour House, walking back and forth to the Market and you could always tell, you'd be looking out the window and there'd be a couple, standing by the side of the road, looking at the harbour, and you could just tell what they're saying, 'I would love to move here. Look at that view. We could move here. Yes, we could.' Ohhh, my god, they're gonna be calling the realtor in five minutes, right? You could just tell. And that's what this does, right? Sucks you in, oh, it's beautiful here, it's beautiful. (Julia, 17 August 2012)

Julia almost resented the beauty of the island and that people used it as an excuse for putting up with things like poorly maintained roads or expensive groceries—all prices to pay for living in such a beautiful place. But at the same time, she loved it and noted that she never tired of seeing sheep in the meadows or hearing tree frogs singing in the spring.

Figure 4. By Julia. Sheep, 2011. Used with permission.

These photographs my participants made of natural beauty, though, are by and large absent one thing—people (see Figures 3, 4, 5, 6). As I saw these photos, I wondered what this absence of evidence, the lack of people in the frame, meant to my participants. I thought to myself that perhaps they were constructing an idea of untouched, "pure" natural beauty free from human impacts. As I spoke with my participants it became clear that it was partly that, but also that they saw natural beauty as a kind of character or personality, anthropomorphized as a female being who takes shape as the island and exists in concert with as well as separate from the people who live there.

Duncan exemplified Julia's story in that he and his wife, Emma, came to Salt Spring as vacationers and fell in love with the beauty and slower pace of life on the island. He and Emma moved their Internet business to the island and worked very hard over several years to tame their five acres of land so that they could grow as much food as they possibly could, wanting to be as self-sustaining as possible. He showed me photographs of their garden and spoke of how much work it was, "more than we ever could have possibly imagined. It's like a beast, sometimes, that you can't turn your back on or you'll have blackberry vines and broom taking over everything before you know what hit you" (Duncan, 21 September 2012). Despite his descriptions of fighting back nature, he also embraced it, noting that being able to walk along the beach or hike up the mountain every morning was something he had come to cherish and didn't think he'd be able to ever find anywhere else.

Figure 5. By Duncan. Beach Life, 2012. Used with permission.

When we sat down to look at her photographs, Emma set the stage when she told me,

> They say that the island holds onto the people she wants, the ones that need to be here and the others she spits out. If it's too much of a struggle, then maybe you're not meant to be here, you know what I mean? It's not always easy being here. It's beautiful and the people are lovely but it's also very hard work. I've replanted that apple tree three times until it found a place it was happy. We've taken out truckloads of broom, and we're still fighting it. But I wouldn't want to live anywhere else now. We've got a good life. I know we're lucky… [laughing] I guess the island wanted us (Emma, 13 June 2012).

Figure 6 is notable for another piece of evidence that we cannot see. Because I knew where my participant, Emma, had taken Figure 6, I also knew that it was framed in such a way that it excluded the pulp mill across the water in the town of Crofton. When I asked Emma about this, she admitted that she had deliberately framed her photo not to include the mill, saying, "it's so ugly and it's not part of Salt Spring. It's like another world over there" (Emma, 13 June 2012). "It's like Mordor," her husband Duncan chimed in from across the room, comparing it to J.R.R. Tolkien's land of evil, death, and horror (Duncan, 13 June 2012). Both Emma and Duncan deliberately removed images of the Crofton pulp mill because it does not fit with their definition of Salt Spring as a place of pristine, natural beauty. When comparing Crofton to Mordor, Duncan then added, after a pause, "It's not like everything is perfect here. We have our problems, of course, but it's pretty idyllic a lot of the time. And it truly is one of the most beautiful places in the world" (Duncan, 13 June

2012). Figure 7 shows the Crofton pulp mill from a slightly different angle from Emma's photo of Vesuvius each, taken when I was crossing on the ferry to Salt Spring. Vesuvius beach is across and to the south from the pulp mill and all of its industrial pollution and ugliness.

Figure 6. By Emma. Vesuvius Beach, 2012. Used with permission.

VALIDITY AND PARTICIPATORY PHOTOGRAPHY

Photographs and video are not truth any more than words are. They are instead someone's visual statement about the world they live in and their experience (Worth 1980:20). Images, particularly the photographs we make and share, uncover what is culturally and socially significant for us (Ball and Smith 1992:32). As celebrated ethnographic filmmaker David MacDougall notes in an interview about his work,

> there is always an ambiguity about the way a film implies something. It suggests, it draws possible connections, it creates reverberations and harmonics. But this is also one of its strengths, because that sort of complexity is also characteristic of much of our social experience. Something with a meaning in one context will have a different meaning in another, but it will nevertheless drag overtones of its other meaning into the new context (Barbash and Taylor 1996:374).

Substitute still photographs for film in MacDougall's quote and I think we could make the same argument. Participatory photography provides an advantage in this respect because it explicitly acknowledges the role of human observation, interpretation, and construction of meaning through images. What do we do, though, when we know as researchers that our participants are being especially, subjectively human and framing their photos to deliberately leave something out, as Emma did when she deliberately framed her photo not to include the Crofton pulp mill? Can we consider what is outside the frame evidence even though we can't see it and is that evidence as valid as what is inside the frame? I think the answer to both of those questions is a qualified yes: if we ask our participants to tell us about what is outside the frame and then further delve into why they left it out, then it becomes valid evidence. I had an advantage because I knew that Emma's picture could have included the pulp mill, but I deliberately asked her about it and why she had excluded it. Excluding something ugly and industrial and clearly unnatural from an image that she was using to describe the natural beauty of the island was equally strong evidence as what she included—I just made sure to document it. How our participants frame their photographs matters. They make deliberate choices about what to include and what not to include and if we gather data through asking them why, evidence outside the frame is valid and equally important as that which is readily apparent. It is our job as researchers, though, to ask about what we don't see just as much as what we do see and to find out the why behind the image.

Figure 7. The Crofton Pulp Mill, aka Mordor. © 2018 Tabitha Steager.

Ephemeral Evidence, Modification, and Social Media

To further complicate matters around what we see and don't see, when we begin to think about images posted to social media sites like Snapchat or to Instagram stories, where images are temporary and disposable due to the limited timeframes they exist within these apps, the evidence we may be interrogating is ephemeral and exists only in the short term. It is beyond the scope of this paper to delve more deeply into images and video posted to social media (often after being highly edited through an app's built-in lenses, filters, and special effects) than a brief discussion here. It is important to mention, though, because it is something researchers using photography (and video) as evidence need to consider. The ephemeral nature of Snapchat and other time-limited social media poses challenges not only for data collection but also analysis—we need to ask what might these temporary, very often highly manipulated images say about the lives and experiences of their makers and what kind of stories people are telling about themselves through these platforms *as well as* what kinds of evidence the ephemeral nature of the images provide.

Some questions we may ask include: What would a participatory study using Snapchat look like and how might we interrogate the nature of evidence and what is "real" within a platform that is inherently artificial? Further, how does evidence produced in one platform differ from that produced in another and what kind of evidence does the use of one platform over another provide? What might the *absence* of a particular kind of evidence mean within the frame of specific research questions? When we consider the use of lenses, filters, and editing apps we need to consider conceptions of authenticity and reality and what those look like to our participants. We need to tease out whether someone considers a selfie they modified using an app to soften lines, make their cheekbones and jawline sharper, and slightly increase the size of their eyes and lips an expression of (modified) reality but no less authentic. Or maybe they consider it *more* authentic because it reflects what they feel is their best self. Perhaps it's aspirational, expressing what they hope to be. Then there is the contrast of #nofilter and its use, meanings, and value—where could that fit into the equation?

While the type of participant-generated evidence that arises out of social media will likely look vastly different from, say, participant-generated photos that are unedited and more "documentary" in style (like those my participants on Salt Spring made), I think we need to be careful not to place them in a completely distinct realm of their own. When participants create imagery—still photos with #nofilter, heavily edited Instagram images, or Snap photos with cat face lenses placed overtop—it is a reflection of their "truth" and their experiences of the world, no matter how artificial that imagery might appear visually. We cannot say that they are less valid evidence of the reflections of reality as our participants experience, or express, it. As I have discussed, our work as ethnographers is to tease out the very personal, embodied, individual experiences of our participants' lived experiences and then make sense of them as a whole. We look for shared patterns as well as distinct differences, that little piece of evidence that engenders questions we hadn't even thought of when we started.

Lived experience today, for many of our research participants, includes day-to-day interactions with their smartphone cameras, often posting the images they create, either unedited or not, onto social media. It is important to pay attention to the differences, of course, between the images generated and whether they've been manipulated or not, but what remains at the very core of our interrogation of those images, as with any ethnographic enquiry, is the *why*: Why did they choose to make that photo when and where they did? What can the evidence contained within the photo tell us as well as what is not included to answer that why? Why did they use a specific filter or not? Why did they choose one photo over another? What does sharing their images with a researcher mean to them and why? These kinds of questions apply even though the types of mediums are increasing and changing at what can feel like record speed. As ethnographers we need to adapt our inquiry to fit, of course, but this is something that ethnographers do particularly well.

CONCLUSION

To return to the beginning, photographs are not factual records any more than words are. They are imbued with context and subliminal and unspoken meanings. When our participants create photographs based on our directions, the sense they make of the photographs as *creators*, and in turn what we make of them as *researchers*, is grounded in cultural beliefs, personal experiences, and context—both of where and when the photo was made and where, when, and how we discuss it with our participants. My participants, and my own experiences, taught me that the creation of place is one fraught with conscious and unconscious arguments about who has the right and the power to define a place. The notion of power, and who holds it, is always at the center of any ethnographic enquiry but becomes particularly noticeable when it is research participants themselves who are creating evidence through material they make, such as photographs. When participants create evidence and direct how we as ethnographers are to understand this evidence, the power shifts from the researcher leading the charge to a more equitable partnership that is more reflective of how knowledge is produced. Using participatory photography actively acknowledges that ethnographic evidence is co-created and knowledge is built together with our participants.

Of course, as viewers of these images we construct a story of space and time that will differ from every other viewer; we can't ever know exactly what someone else is seeing, just like that salmon color my mother and I argued about. When working with participants and the photos they make, it is important to keep this framing in mind—participants recount to us what their photos mean to them, how they went about taking them, and what the process of making the photos means. However, as researchers we cannot help but see those photos in a different way than our participants. We bring additional analysis and interpretation to layer on top of that of our participants, mixing images and text to create ethnographic evidence that we then present to other audiences. For example, Figure 8 is a montage of images I've chosen to put together, all photographs of a Salt Spring ferry, three made by my participants and one of them made by me. I include this montage not only because it shows the importance of ferries to people on Salt Spring, but more importantly because it epitomizes participatory photography for me—it tells a story about Salt Spring co-created by my participants and me. I would argue that including their photographs alongside mine helps to lend credibility to my work. I didn't just make a picture of the ferry and then provide *my* interpretation of how it contributes to Salt Spring as a place. My research participants also made photos of the same experience, and in showing it to me and discussing it together, we co-created ethnographic evidence. Having participants actively involved in the research process through methods like participatory photography lends our ethnographic work credibility. This paper is an argument for incorporating the method into all of our research projects where images are used, to help to bring our participants' knowledge and experiences to the forefront, which is where we hope to be in any ethnographic endeavor.

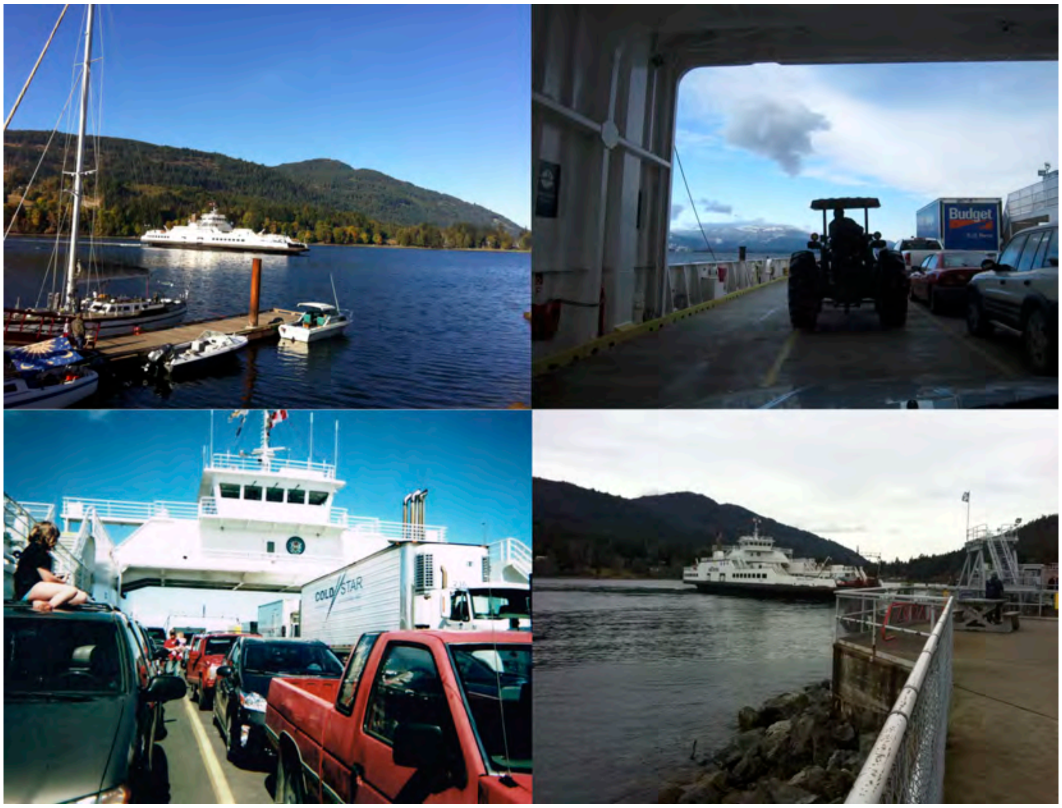


Figure 8. Salt Spring Ferry. Clockwise from top left: Duncan, 2012; Dusty, 2011; Tanya, 2011; Tabitha Steager, 2011. All used with permission.

Tabitha Steager is an anthropologist with interests in place and community, food, visual ways of knowing, and Indigenous rights. She received her PhD in Interdisciplinary Studies from the University of British Columbia. She has conducted research in Canada, the United States, Mexico, England, France, Italy, and with First Nations across British Columbia.

NOTES

Thank you to the EPIC reviewers and curators for their instructive, thoughtful comments, especially Rebekah Pak for her insightful comments, all of which helped shape this paper. Please note that the views expressed in this paper do not reflect the official position of Pacific AIDS Network and the content is based solely on my own research, not that done under the auspices of Pacific AIDS Network.

REFERENCES CITED

Ball, Michael S. and Gregory W. H. Smith
1992 Analyzing Visual Data. Newbury Park, CA: Sage.

Barbash, Ilisa and Lucien Taylor
1996 Reframing Ethnographic Film: A “Conversation” with David MacDougall and Judith MacDougall. American Anthropologist 98(2):371-387.

Bean, Jonathan
2008 Beyond Walking with Video: Co-Creating Representation. In Ethnographic Praxis in Industry Conference [EPIC] 2008, 104-115.

Edwards, Elizabeth Edwards
2015 Anthropology and Photography: A long history of knowledge and affect. Photographies 8(3):235-252.

Faulkner, Susan A. and Alexandra C. Zafiroglu
2010 The Power of Participant-Made Videos: intimacy and engagement with corporate ethnographic video. In Ethnographic Praxis in Industry Conference [EPIC] 2010, 113-121.

Geertz, Clifford
1988 Works and Lives: The Anthropologist as Author. Stanford, Calif.: Stanford University Press.

Golub, Alex
2017 Bronislaw Malinowski: Don't Let the Cosplay Fool You. Savage Minds website, May 15. Accessed July 23, 2018. https://savageminds.org/2017/05/15/bronislaw-malinowski-dont-let-the-cosplay-fool-you/

Harper, Douglas
2002 Talking About Pictures: A Case for Photo Elicitation. Visual Studies 17(1):13-26.

Jacknis, Ira
1988 Margaret Mead and Gregory Bateson in Bali: Their Use of Photography and Film. Cultural Anthropology 3(2):160-177.

MacDougall, David
1997 The Visual in Anthropology. *In* Rethinking Visual Anthropology, Marcus Banks and Howard Morphy, eds. Pp. 276-295. New Haven; London: Yale University Press.

Packard, Josh
2008 "I'm gonna show you what it's really like out here": the power and limitation of participatory visual methods. Visual Studies 23(1):63-77.

Pink, Sarah
2007 Doing Visual Ethnography: Images, Media and Representation in Research (rev. and expanded 2nd ed.). London: Sage.
2003 Interdisciplinary agendas in visual research: re-situating visual anthropology. Visual Studies 18 (2):179-192.

Worth, Sol
1980 Margaret Mead and the Shift from Visual Anthropology to the Anthropology of Visual Communication. Studies in Visual Communication 6(1):15-22.

Human Sensemaking in the Smart City: A Research Approach Merging Big and Thick Data

ANNELIEN SMETS
imec-SMIT, Vrije Universiteit Brussel

BRAM LIEVENS
imec-SMIT, Vrije Universiteit Brussel

This paper aims to contribute to the debate on the integration of ethnography and data science by providing a concrete research tool to deploy this integration. We start from our own experiences with user research in a data-rich environment, the smart city, and work towards a research tool that leverages ethnographic praxis with data science opportunities. We discuss the different key components of the system, how they work together and how they allow for human sensemaking.

THE BIG-THICK CHALLENGE

Both in industry and academia, we witness an increasing focus on human behavior and individual experiences. In this era of cheap data storage, this interest translates into the collection of as much as data as possible of someone's actual behavior. In digital environments for example, each keystroke and each click are precisely being logged, creating high volumes of data. By means of different computational methods these digital traces are then being transformed to gain a deeper understanding of users' behaviors in order to improve products or services. With the emergence of the Internet-of-Things (IoT), the increasing number of physical objects that are being connected, we observe the same phenomenon in environments of our daily lives that originally were not perceived as digital.

Within urban environments, especially, this connectedness is already resulting in an abundance of available data originating from sensors deployed throughout the entire (smart) city. However, the complexity of the urban environment adds additional challenges to the process of transforming data into meaningful information and actionable knowledge.

First, one needs to know what to measure. This includes safeguarding the balance between predefined research variables and the level of serendipity, as well as identifying what should be measured at which moment in time. Second, the entanglement of multiple domains in the city, results in a high variety of data streams that require different expertise knowledge. We need a thorough and unambiguous understanding of what is measured and what this exactly means (Hey et al. 2009). For example, air quality is often represented as the amount of parts per million, but for a layman this is meaningless data.

This links up to the third challenge: contextualization. A great deal of the existing data originates from individualistic systems that have no sense of the collective: "it might be big, but not very useful unless it is set in a wider context involving other data" (Batty 2016, p.323). This means that, in order to really understand the data, we do not only have to look at the data itself, but also take into account its broader context. To stick to the air quality example; to interpret the data of an air quality sensor, one has to take into account parameters such as traffic density and weather conditions, as they influence the measurements and therefore are important to get an understanding of the situation.

The fourth and final challenge, we call it human sensemaking, relates to providing an additional layer to the data, which captures what this data means to the citizens. Again, in case of air quality, we want to add a human interpretation to each of the different quality levels and identify if these interpretative categories align with the different categories resulting from the knowledge domain; is "low air quality" actually perceived as unpleasant?

This human interpretation is necessary as it contributes to a better understanding of experiences and human behavior. Similar to the example of using digital traces to improve services, we need these insights to address urban challenges. Smart cities aim at improving the overall quality of life in the city and following the progress made by merely technological solutions, we are now heading towards smart city solutions that target a change in human behavior (de Oliveira et al. 2015). However, to design smart city solutions that are able to bring about this change, we need to connect the data resulting from the city with in-depth insights of people's behaviors and experiences. This means that we are now dealing with situations where we have to determine if citizens are aware of the smart lightning system in their street, study how they experience a smart traffic light or examine how safe they feel on the road. Can we infer this information from the data the smart city is providing us? And more importantly, how can we do this?

In this paper, we address this sensemaking challenge by exploring the creation of a research tool that leverages ethnographic praxis with data science opportunities. In this way, we aim to contribute to the current call from research to provide actual tools to analyze and work with data (Churchill 2017). The desire to integrate these two disciplines is not only motivated by common interests of ethnographers and data scientists (Curran 2013), but also by the additional value generated by integrating both disciplines: large-scale in-depth insights. While an ethnographic approach allows for an in-depth understanding of people's beliefs and behaviors in their context (thick data), data science is able to detect patterns in data points collected at a large scale (big data) (Wang 2016). Although most of the discussions on this integration remained conceptual, they have already led to numerous valuable insights. Consequently, this paper mainly focuses on the research tool that we are currently developing and the main insights from the research leading up to this.

TOWARDS HUMAN SENSEMAKING

The term "sensemaking" knows many interpretations and, ironically, in its broadest sense it could be described as "getting an understanding or attributing meanings to something" (Kari 1998). In our sensemaking approach, we start from Dervin's assumption about individuals' sensemaking in the way that "they experience and observe their world differently and need to create meaning or make sense of their world" (Dervin 1992, p.62). The goal of our human sensemaking approach is to being able to capture these different senses (let it be in the forms of meanings, experiences, etc.). More specifically, within our smart city research, we use sensemaking to refer to the process by which a participant (most often a citizen) gives meaning to his or her experience related to the interaction with a service, technology or an urban environment.

Dervin argues that sensemaking can be considered as "behavior, both internal (cognitive) and external (procedural), which allows the individual to construct and design his or her movement through the time-space context" (Du 2014, p. 29). In order to understand this behavior and its outcomes we need to be able to study it. This comes, however, with some challenges. First, we need to know the context in which this behavior occurs and second, as it is time-space dependent, we need to be able to capture the experiences throughout different contexts.

The research tool presented in this paper, addresses these challenges by bridging the gap between big and thick data and thereby enables studying experiences within their context. Not only does the tool reinforce the relationship between big and thick, it also adds meaning to each data type individually by providing a greater context. To this end, the research tool relies on four sensemaking strategies, which arose from previous experiences in working with big and thick data and the different needs we identified: time-space dependency, pattern detection in large data-sets, gathering subjective experiences and combining those with objective data to create meaning. We define the following sensemaking strategies:

- **Contextualization**: as Rato (2000) pointed out, being able to determine the 'right' context within which to situate one's analysis, is one of the key elements in both ethnographic and sensemaking research. This strategy is in line with the Living Lab contextualization phase as defined by Pierson & Lievens (2005) and the grounded theory approach. In our approach, contextualization relies on various big data collection methods (such as logging data, tracking data) in order to get a profound insight in one's behavioral context.
- **Semantics**: the semantic strategy builds further on these collected datasets and provides a first meaningful layer to the gathered data points. By means of a semantic framework, raw data is being translated into meaningful data objects. For example geolocation data is being translated into amount of times being present in a certain place, the duration of a person at a given spot.
- **Analysis**: similar to the semantics, the analysis strategy supports researchers to get meaningful information out of the data points. This strategy allows to detect patterns in large data-sets as well as summary statistics that allow researchers to put the data into perspective. Another activity in this strategy is the identification of user profiles (i.e. clustering groups of users).
- **Human interpretation:** whereas the previous strategies are still a construct of the research, the human interpretation strategy allows for a direct sensemaking process by the subject itself within a particular context. By means of using experience sampling techniques it is possible to capture the experience at the time and space of its occurrence. This strategy thus allows to add another additional layer to the data consisting of experiences, emotions, motivations, etc.

These strategies could broadly be divided in two categories where one of them is more concerned with dealing with big data whereas the other deals with gathering the thick data. Nevertheless, these categories are non-exclusive and there is a continuous interaction among them. A specific example of the realization of these strategies and how they work together is provided later in this paper.

In the remainder of this paper, we describe a research toolkit that allows to put these four sensemaking strategies in practice. We position this toolkit within a Living Lab research approach, since this is our primary way of working and hence the requirements of the toolkit find their origin in our previous experiences with Living Lab research.

THE URBAN LIVING LAB

The increasing focus to study human behavior in urban environments, strengthens the need to perform contextual research at large scale and integrate big and thick data within smart city research. At the same time, we observe smart cities having the potential to actually accomplish this integration: their vast amount of available contextual data allowing for exploratory data analysis, reinforces the ethnographic approach by evoking a greater level of serendipity (Rivoal & Salazar 2013). Access to this data and means to turn this into information are obtained by perceiving the city as a Living Lab (Coenen et al. 2014). The city, as an urban innovation ecosystem, thereby acts like a dynamic open experimentation environment involving its citizens (Veeckman & Van der Graaf 2015). Juujärvi & Pesso (2013, p.22) define the Urban Living Lab as follows:

> "a physical region in which different stakeholders form public-private-people partnerships of public agencies, firms, universities, and users collaborate to create, prototype, validate, and test new technologies, services, products, and systems in real-life contexts".

By approaching the smart city as a permanent Urban Living Lab, it internalizes ethnographic characteristics such as naturalism, understanding and discovery and enables studying users' behavior and experiences in the wild (Pierson & Lievens 2005). Additionally, the Urban Living Lab approach provides a framework to co-operate with community partners and establish a method to cross-validate the observed patterns in the data (Kontokosta 2016) as well as to complement these with their human interpretative counterpart.

These characteristics are of big importance in studying the sensemaking process and its outcomes. First of all, the ability to study behaviors and experiences in the wild results in a more genuine sensemaking process that is not (or at least less) influenced by the research itself. Additionally, the Urban Living Lab provides a semi-controlled environment, which allows the researchers to take into account the time-space context of the observed behavior and experiences. At the same time, this time-space context can also be used as an entry-point to steer parts of the research (e.g. only ask questions when the subject is in a particular context). Moreover, the Urban Living Lab allows for big data collection exceeding the bounds of the recruited participants, while it also facilitates the collection of thick data by having dedicated interactions with the set of participants.

Urban Living Labs do not only hold great promises to facilitate the integration between big and thick data, they also need tools that allow for this integration. Currently, we notice two main problems that would benefit from it. First, there is what we call the problem of unavailability. Within the current Urban Living Lab projects, we notice that there is a lack of objective data. To gather data about the interaction with a digital service, researchers can (and do) rely on logging data. However, with regard to the context or the actual behavior of citizens in the urban environment, researchers need to rely on reported data by the participants. This often done by means of traditional qualitative methods such as observations, interviews, focus groups and diaries, whilst -especially in the smart city context- there are possibilities to directly obtain this contextual and behavioral data.

The second problem is more related to the actual integration of big and thick data and can be described as the problem of *asynchronicity*. By asynchronicity, we mean that there is data available on the behavior or the context of the citizens, but this data is only thickened some days (or even weeks) after the behavior occurred. This results in a significant delay between the generation of the data and the moment at which it is being interpreted and gets a meaning. Although this time span can have beneficial outcomes supporting self-reflection, it does also result in a potential recall bias and the experience might be distorted over time. The latter could be due to repeated experiences, but also due to external influences. In our opinion, the time delay results in post-experiences rather than experiences in the moment. These post-experiences are definitely relevant as well, however, within our research scope, we are looking for thickening strategies within the moment itself.

Having observed these two main problems in the last years, we recently piloted some case studies where we tried to augment current Urban Living Lab practices by using digital trace data and at the same time overcoming this problem of asynchronicity. In the next section, we will briefly describe two cases and pinpoint our lessons learned, which have been the major drivers to continue our work on human sensemaking in the smart city. Our experiences with these case studies have also greatly defined the requirements for the research tool that is described in this paper.

Thick Understanding in the Smart City

The two case studies that are described in this section are part of City of Things[1]. This is a Smart City (Urban) Living Lab and IoT testbed located in Antwerp (Belgium) that aims to bridge the gap between the quadruple helix (government, research, industry and citizens) by bringing them together and let them collectively test and validate innovative solutions within a real-life environment (Latré et al. 2016). Currently, different projects are ongoing within this Urban Living Lab focusing on mobility, air quality and traffic safety. Except from the various technological challenges in these projects, as user researchers we are challenged to apply proper methods and tools to capture the needs, requirements and the real-life experiences of the various stakeholders (mainly citizens) to gain sufficient insights to steer and evaluate the design process. The two projects that we will describe, have been set up to explore the possibilities of the technological infrastructure in City of Things and to what extent this can benefit our user-centered, ethnographic research steps. The first project, Citizen Bike, mainly focused on how we could use big data in our research process, while the second project, Be-Und, explored the use of contextual data to trigger in-situ interactions with the participants.

Citizen Bike - In this use case we studied how the cycle experience in the city could be improved. In our approach, we first wanted to identify the current cycling experiences throughout the city and our goal was to do this by means of geolocation data (to capture the movements) and data from different sensors in the city (to capture the context) (see Boonen & Lievens (2018) for a thorough description). Instead of using an off-the shelf smartphone application that tracks one's location and movements, we chose to develop a customized device that allowed for continuous tracking and provided real-time data without any intervention of the end-user (Figure 1). This device also contained two buttons that allowed participants to share their experiences while cycling. The use of this device enabled to gain a deeper insight in the experiences of the cyclists by combining big and thick data methods.

Figure 1. Citizen Bike device mounted on a user's bike.
© Boonen & Lievens (2018), used with permission.

During the field study the device provided us two types of data in real-time while the participants were cycling throughout the city: a continuous stream of sensor data and user experience data (as the users were asked to push one of the buttons when they experienced a positive or negative situation while cycling). We analyzed these data points and integrated some of them (Figure 2). Although this analysis provided us some general insights, it lacked subjective data from the respondents. Subsequently, based on these first insights, we performed additional qualitative research to get a deeper (and hence ticker) understanding of their cycling experiences.

Figure 2. Visualization of users' mobility (left) and their cycling experiences combined with different levels of noise (right).
© Boonen & Lievens (2018), used with permission.

The Citizen Bike case was hence one of our first attempts to use digital data from sensors to capture one's experience and make meaning of it. In terms of methodological approaches, one could say that we used the experience sampling method to capture the cyclists' experiences, because we allowed them to indicate their positive or negative experience in the moment by using the interaction buttons. As mentioned before, we used these data points as entry points for our qualitative interactions with the participants and this turned out to be a very valuable approach. As mentioned in Boonen & Lievens (2018, p.210):

> "(...) this helped us to stimulate users' recollection. For instance, as soon as we confronted one participant with a certain push on the button, she remembered the whole route and could provide more information on why she pushed the button whilst reflecting on her general use."

Being able to present the captured data to the user turned out to not only be useful to get a deeper understanding of the experiences they did share, but it was also helpful to identify why they did not share something. For example, the noise levels at the location of the participant might have been very high, yet this situation was not marked as a negative experience; why not? This insight taught us that having this continuous stream of data on contextual and/or behavioral events is important to allow the researcher to look for interesting situations or patterns by him- or herself, rather than solely relying on the input of the participant. Moreover, during the qualitative research steps, we also learned that being able to interact with the participants in real-time (at a greater extent than interaction buttons) would have been more beneficial because the real experience and the subjective feelings that are associated with it, are often hard to recall and this could have been overcome by being able to interact in the moment.

Be-Und - The objective of this second use case (see Smets et al. (2018) was to explore how we could map and understand current behaviors of citizens, as a first research step within a broader behavior change research approach as defined by Spagnoli et al. (2017). To study the use of digital methods and objective data on both the behavior as well as its context, we first developed a generic behavioral understanding method, which we then applied to a specific use case: commuting behavior.

Whereas the Citizen Bike project was able to capture cycling behavior (route), some contextual parameters and a thin, in-the-moment description of the cycling experience (positive or negative), this use case aimed for a greater in-situ interaction with the users to explore their experiences. To this end, we set-up a research design that relied on the context-aware experience sampling method, which improves the traditional experience sampling method "by using sensing technologies to automatically detect events that can trigger sampling and thereby data collection" (Massachusetts Institute of Technology 2008). This implied that we, as researchers, were able to interact with the participants and ask questions about their experiences based on their context.

This approach us to properly investigate which contextual factors influence one's decision to take either the car or the bike to go to work. For a two-week period, we equipped 6 citizens of Antwerp with a context-aware experience sampling tool that allowed us to track their location. This data provided us information when they are leaving at home and when they arrived at work. This tool sent a notification to the researcher's smartphone when the participant arrived at a particular location (in this case work). This allowed us to immediately, and hence still in the moment, send a tailored questionnaire to the participant to acquire a thick description of the commute and what influenced the commuting decision (car or bike).

To enrich and validate the collected data, an additional focus group was organized after the two-week period, to confront the users with the data and insights. Moreover, in contrast to Citizen Bike, the ability to interact with the participant in the moment (i.e. when they arrived at work) turned out to be particularly interesting since we were able to inquire really specific details that might be forgotten when time has passed. Nevertheless, we also experienced some difficulties with the actual implementation of the context-aware experience sampling and listed some requirements for a future tool: real-time accessible data, integration with contextual data, user-friendly content management system, pre-scripted triggering rules and question sets that can be send automatically and to each participant individually (Smets et al. 2018).

CITIZEN TOOLBOX

Taken into account the learnings from the above cases (Boonen & Lievens (2018) and Smets et al. (2018), we designed a toolbox that would help us to overcome some barriers and hence satisfy our need to be able to conduct user research within an Urban Living Lab while taking advantages of the available digital contextual and personal data. The main goal of this research tool is to provide user researchers an easy and accessible instrument to collect objective, big data and use this data to interact with the end-users to make sense of this data by combining it with thick data. We first present a high-level description of this tool indicating the key concepts and data flows. Next, we dig deeper in the specific components, their configurations and usage. Thereafter, we describe four ways in which the Citizen Toolbox encourages sensemaking.

The development of the Citizen Toolbox was approached by means of an iterative and agile design process (including user story mapping, prototyping, etc.) involving all stakeholders within our organization (user researchers, project managers, database managers, software developers and user involvement coordinators amongst others).

The tool allows to unlock various contextual and personal data streams and, based on these data, provides the ability to directly interact with participants to gain a thicker understanding about their behavior and experiences. More in particular, it consists of (1) a personal tracking system that monitors someone's behavior, (2) a database containing contextual sources (such as weather and traffic) and (3) a rule-based experience sampling method that enables contextual inquiry based on the input of (1) and (2). The main strength of this set-up is that it allows to perform contextual research at a large scale and to integrate big and thick data, thereby augmenting the overall value of the data. Next to this overall research goal, we also identified some additional requirements that we perceive as crucial for the Citizen Toolbox to be used as a tool to support user researchers (Table 1).

Table 1. Requirements and corresponding features in Citizen Toolbox.

Requirement	Feature
Short time-to-research	Fixed set-up and limited scoping Pre-configured devices
Independent of technical development	End-user programming Library (data sources) Standard reporting
Reusable in different projects across different domains	Library (rules, data sources) Standard reporting Stream framework Multiple data input modalities (plug & play)
Safeguard serendipity	Flexible system configuration
Activity monitoring	Various logging data

For the design of the Citizen Toolbox we chose for a solution that consists of a pre-configured set-up in terms of devices and database configurations. This does not only allow us to roll it out quickly, but also allows the tool to be managed and configured by user researchers rather than developers. Moreover, there is the need to safeguard serendipity as it is inherent to ethnographic praxis. Consequently, the Citizen Toolbox allows for flexible system configurations, meaning that not everything should be defined upon the start of a project, rather a researcher can still make changes during the project based on empirical evidence.

> "Ethnography is an inductive science, that is: it works from empirical evidence towards theory, not the other way around. This has been mentioned several times already: you follow the data, and the data suggest particular theoretical issues." (Blommaert & Jie 2010, p.14)

Building Blocks

A schematic representation of the research tool can be found in Figure 3. In general, one could identify three major types of components, which we will extensively describe below: big data input, processing components and thick data capturing.

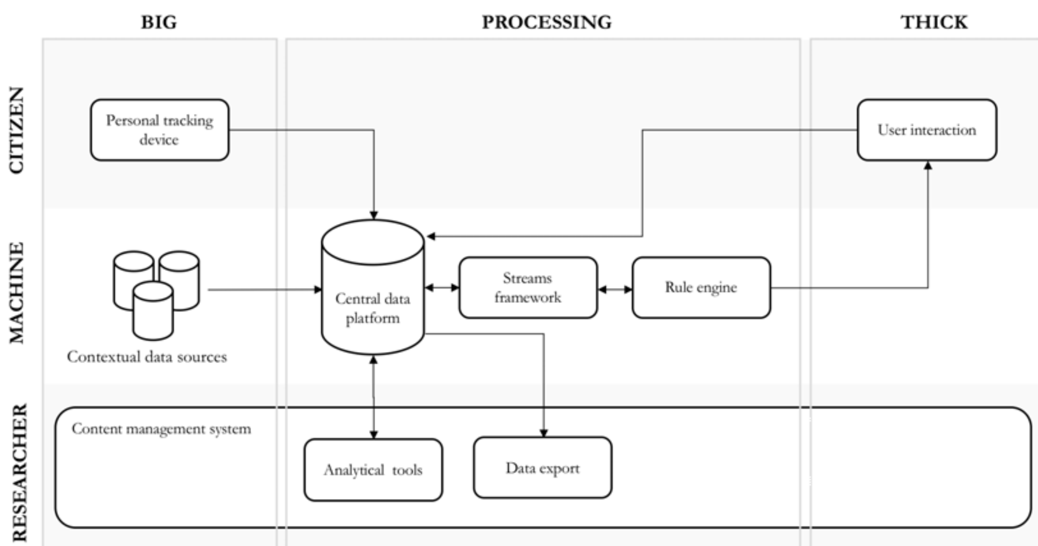


Figure 3. Schematic representation of the different components of the Citizen Toolbox. © Authors' illustration.

***Big data input* -** The Citizen Toolbox foresees two big data sources: a personal tracking device and contextual data sources. The former allows to capture behavioral data, which could be one's location, movements or physiological data such as heartbeat rate or galvanic skin response. In the first version of the Citizen Toolbox, the personal tracking device keeps track of the geolocation of the participants, using a GPS sensor. This is our crucial component to be able to identify the context of the participants because their geolocation can be connected with contextual data from external data sources.

These external, or contextual, data sources allow to gather additional data on contextual parameters. To this end, we foresee a library of various existing, external data sources that are integrated in the tool and can be used in a specific research project. For our research within the smart city, we think of existing external data sources such as weather data, air quality, traffic, number of available parking lots, etc.

Processing components - The big data input is centralized to a central data platform. This platform automatically takes care of the required data processing activities (e.g. extracting the data from its source, cleaning, transforming into the correct data types and loading it into the database). In addition to the big data input sources, the data platform also stores all inputs from other components in the system.

Next, there is the streams framework responsible for the semantic translation of the raw data into data objects that are meaningful to the researchers. The main rationale of this component is the idea that we need some sort of aggregated variables to continue our sensemaking process rather than the individual data points coming from the tracking device and/or external data sources. For example, we want to learn from the GPS data when someone is at home, however, this geographical point 'home' will contain multiple coordinates which we are not interested in, but only in their aggregated variable 'home'.

The rule engine is the core processing component. It is responsible for the automatic handling of predefined actions. These actions could be diverse like writing variables to a database or initiating the calculation of an aggregated variable. This kind of actions are however rather useful in practical terms and what we consider to be the most valuable functionality of this rule engine is its ability to trigger questions to the user. The triggering of these questions is hence based on meeting a condition consisting of one or more variables. At the moment this condition is satisfied, the specific question is sent to the interaction module (see below) and the response given by the participant is stored in the data platform.

Finally, the Citizen Toolbox also contains a content management system (CMS) that allows researchers (and other stakeholders such as project managers and panel managers) to configure the system, view the data and monitor the project's current state. The two most important modules of the CMS are the analytical tools and data export tool.

The analytical component represents a set of analytical tools that allow the researcher to analyze the data. The tools allow the researcher to generate some first, standard insight reports, that act like a dashboard for the researcher. Additionally, these tools provide the ability to perform profound statistical analysis of which the results can be re-inserted in the central data platform. Apart from the descriptive analysis, the main focus will be on the detection of patterns and deviations, as we want to be able to thicken these insights with end-user input, through the experience sampling or as input for more in-depth qualitative research.

At last, the data export module allows researchers to download raw data files, but also provides the possibility to download standard reports.

Thick data capturing - Together with the rule engine, the interaction module represents the core of the thickening capabilities of the Citizen Toolbox. After all, this interaction module allows to add an additional layer to the data with the users' interpretation and/or

experiences. This interaction with the user is supported by a smartphone application that is installed on the participant's own device. The choice to install this application on their own device rather than on an additional device that is provided by us, is deliberately made based on our previous experiences with user research. After all, our goal is to minimize the response time (i.e. time between sending a question and receiving a response) in order to maximize the number of in-situ interactions. Therefore, we chose to install the interaction module on their own device as they are more likely to see a notification on this device compared to a device that they only have to use for the purpose of the research.
This smartphone application sends a notification to the participant when he or she receives a question (triggered by the rule engine). The participant is then able to answer it in the application.

Four Times Sensing

So far, we have described the different components of the Citizen Toolbox and their main functionalities. However, we still need to demonstrate how the configuration of the toolbox allows for the four sensemaking strategies as described above. Figure 4 represents these sensemaking strategies and how they are enabled in the Citizen Toolbox configuration. To illustrate them, we will briefly describe a case study and highlight the activities that allow to thicken the data.

The Story - A high-school located in the center of a large city wants to stimulate its students to come to school by bike. A first survey among their students indicated that first-graders are less likely to bike to school than older students. One of the major barriers that these first-graders face, is the traffic near the school, which makes them feel unsafe. The school wants to support these first-graders to take the bike to school, but they also want to get a better understanding of other cycling motivations of other students because they feel that getting such an understanding might benefit all students in the long-run. Given these needs, a research project was set up, involving the use of the Citizen Toolbox.

Contextualization - The first strategy allows to integrate multiple data sources and consequently makes it possible to see the data in its wider context. In the school example, this could be the integration of the GPS data from the personal tracking devices and information on the traffic (that could be an indicator of road safety). However, the mere integration of these datasets could not be very useful yet, although, in the Citizen Toolbox, it can be directly used as input for other sensemaking strategies to turn it into actionable insights.

Semantics - The semantic strategy heavily relies on the streams framework since this component is responsible for the semantic translation of the raw data into data objects that are meaningful to the researchers. In addition to that, the streams framework also provides some inherent, predefined attributes of these data objects. For example, the streams framework identifies a *trip* as a sequence of individual geolocation data points that has a meaningful start and end point (what we call *touchpoints*). From this trip, the streams framework automatically calculates attributes such as its distance and duration. This information is then sent back to the central data platform and stored such that the researcher

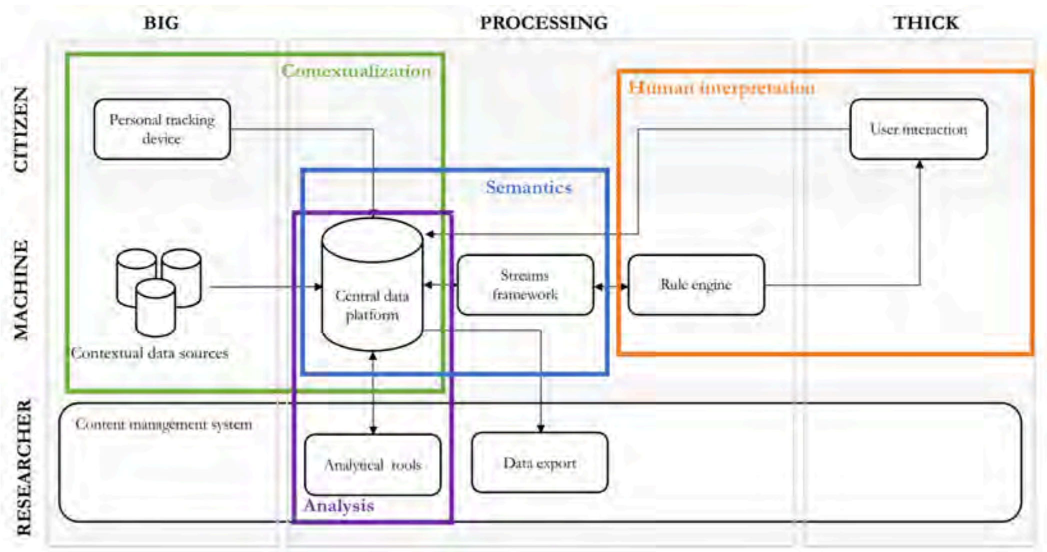


Figure 4. Representation of how the Citizen Toolbox allows for sensemaking in four ways.
© Authors' illustration.

can work with these meaningful objects and their attributes. In our case example, we could identify the school and a student's home as *touchpoints* and consequently identify one's *trip* from home to school.

Analysis - Analyzing the available data in the central data platform is the third sensemaking strategy supported by the Citizen Toolbox. Depending on the aim of the researcher, data analytics could be used either as exploratory or confirmatory. The former tries to find patterns and relations in the data, whereas the latter relies on statistical techniques to validate a hypothesis. An important aspect here, is that the results of the analysis can be added to the existing data. For example, when we segment the students based on the average duration of their trip to school, we can add this segmentation information to their profiles.

Human interpretation - Finally, we come to the real human sensemaking, allowing to add a human interpretation layer to the data consisting of experiences, emotion, motivations, etc. The rule engine allows to act upon all the previously identified data objects with their corresponding attributes by means of rules that trigger an action. For example, when we want to investigate if roads with heavy traffic are perceived as more unsafe, we could define the following rule: if (traffic = heavy) and (transport_mode = bike) and (location = school) then send_question. This means that if the condition is satisfied, the student will receive a question through the application installed on his or her smartphone upon arrival at school: "Did you feel unsafe on the road today?". The student can answer this question by simply choosing for yes or no and send the response.

We hope that it has become clear to the reader that these four sensemaking strategies in the Citizen Toolbox are not independent of one another and the result of one could be the input for another one. In this way, the thickening process becomes a cumulative one where the big data gradually becomes thicker.

CONCLUSIONS AND THOUGHTS

In this paper we described the Citizen Toolbox as a research tool to combine big and thick data as to enable human sensemaking. This toolbox is a supporting tool for user researchers and allows them to collect objective, big data and use this data to interact with the end-users to make sense of this data by enriching it with thick data. This empowers researchers by providing them a better understanding of users' experiences and allowing them to act upon the insights resulting from the integration of big and thick data. The additional value generated by this integration has been discussed in the field for a while and the benefits are clear: large-scale in-depth insights. Thick data allows for an in-depth understanding of people's beliefs and behaviors in their context, while big data allows to detect patterns in large sets of data points.

The core strength of this tool is its ability to combine different data sources and automatically act upon them by means of the rule engine and the interaction module. The tool allows these interactions to be triggered within a specific context, which empowers the researcher in collecting the thick data. The Citizen Toolbox hereby supports researchers to cope with the asynchronicity problem that we have identified in several Urban Living Lab projects. Moreover, the toolbox allows for four types of sensemaking: contextualization, semantics, analysis and human interpretation. These strategies interact with each other and thereby create an additional value in terms of understanding. In other words, these strategies allow to thicken the big data. However, the question is how thick did it become? How can we assess this and when do we reach the limits of the big data feeding the thickening process and do we eventually need to turn to other methods anyway? These are still open questions to us.

We developed the Citizen Toolbox with the goal of being a supportive tool for research, which might have resulted in less attention to the actual research approach. However, we believe that the configuration of the toolbox allows for an implementation to support multiple research approaches in various types of research projects. It is up to the researcher to decide how and when the Citizen Toolbox can be used. From our point of view, we consider the use of the Citizen Toolbox to be relevant to gain more in-depth insights from end-users in end-product development and evaluation, behavior mapping and understanding and the evaluation of behavior change interventions. Moreover, the modularity of the data input allows the Citizen Toolbox to be used in multiple research domains, such as for example the health domain where geolocation might not be the most important personal data input, but rather physiological data such as heart rate or galvanic skin response.

We want to end this paper with two final matters of concern related to ethnographic research: the level of serendipity and the balance between privacy and intrusiveness. First of all, the Citizen Toolbox is designed to be serendipity enabled, because it allows to induce additional insights from the data and act upon it in a later stage. However, at the same time, the set-up of the toolbox does require some initial scoping and pre-definition of rules and variables. This does not only threaten serendipity, it also risks to cause single-mindedness by the researcher to only look for confirmation of the hypotheses. To overcome this, we value the combination of the use of the Citizen Toolkit with other qualitative research methods, such as focus groups or in-depth individual interviews.

Finally, we have to take into account the balance between privacy and intrusiveness. On the one hand, the continuous stream of data coming from the personal tracking device, allows us to use a less intrusive approach, because we can monitor a lot of information without having to ask the participant everything we are interested in. However, on the other hand, this raises some justified questions regarding participants' privacy. Although the toolbox includes measures to protect participants' privacy, it remains a sensitive matter and potentially more complicated, because of new EU regulations. The question is where we have to keep the balance and if it is independent of the research project. For us, this remains an open question for the time being and finding this balance will continue to be a challenge in many research projects involving human subjects and their personal data.

Annelien Smets studies the relationship between society and data-driven innovations. She holds a master's degree in Information Management and Artificial Intelligence from KU Leuven and currently works at imec-SMIT, Vrije Universiteit Brussel as a smart city researcher.

Bram Lievens is a senior user researcher at imec-SMIT, Vrije Universiteit Brussel since 2002. He is involved in various research projects investigating the interplay between technology, society and humans, from a user-centered perspective. He has a bachelor's degree in social and cultural work and a master's degree in communication sciences.

NOTES

1. More information on the City of Things project in Antwerp (Belgium) can be found online at https://www.imec-int.com/en/cityofthings.

REFERENCES CITED

Batty, Michael
2016 Big data and the city. UK: Alexandrine Press.

Blommaert, Jan and Jie Dong
2010 Ethnographic fieldwork: A beginner's guide. UK: Multilingual Matters.

Boonen Michelle and Lievens Bram
2018 The use of live-prototypes as proxy technology in smart city living lab pilots. *Human-computer interaction international conference* 2018.

Churchill, Elisabeth
2017 The Ethnographic Lens: Perspectives and Opportunities for New Data Dialects. EPIC Perspectives. https://www.epicpeople.org/ethnographic-lens

Coenen, Tanguy, van der Graaf Shenja and Walravens Nils
2014 Firing up the city-a smart city living lab methodology. *Interdisciplinary Studies Journal* 3(4), 118.

Curran, John
2013 Big data or 'big ethnographic data'? Positioning big data within the ethnographic space. *Ethnographic praxis in industry conference proceedings* 2013, 62-73.

de Oliveira, Álvaro, Campolargo Margarida and Martins Maria
2015 Constructing human smart cities. *Smart Cities, Green Technologies, and Intelligent Transport Systems,* 32-49.

Dervin, Brenda
1992 From the mind's eye of the 'user': The sense-making qualitative-quantitative methodology. *Qualitative research in information management,* 61-84. Englewood, CO: Libraries Unlimited.

Du, Jia Tina
2014 Library and Information Science Research in Asia-Oceania: Theory and Practice. IGI Global.

Hey, Tony, Tansley Stewart and Tolle Kristin M.
2009 The fourth paradigm: data-intensive scientific discovery. Redmond, WA:Microsoft research.

Juujärvi, Soile and Pesso Kaija
2013 Actor roles in an urban living lab: What can we learn from Suurpelto, Finland?. *Technology Innovation Management Review* 3(11), 22-27. doi: 10.22215/timreview742

Kari, Jarkko
1998 Making sense of sense-making: From metatheory to substantive theory in the context of paranormal information seeking. *Nordis-Net workshop: (Meta) theoretical stands in studying library and information institutions: individual, organizational and societal aspects*. Oslo, Norway.

Kontokosta, Constantine E.
2016 The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation. *Journal of Urban Technology* 23(4), 67-84. doi: 10.1080/10630732.2016.1177260

Latré, Steven, Leroux Philip, Coenen Tanguy, Braem Bart, Ballon Pieter and Demeester Piet
2016 City of Things: an integrated and multitechnology testbed for IoT smart city experiment. *Smart Cities Conference (ISC2), IEEE International*, 1-8. doi: 10.1109/ISC2.2016.7580875

Massachusetts Institute of Technology
2008 Context-aware experience sampling. http://web.mit.edu/caesproject/index.htm

Pierson, Jo and Lievens Bram
2005 Configuring living labs for a 'thick' understanding of innovation. *Ethnographic Praxis in Industry Conference Proceedings* 2005, 114-127

Ratto, Matt
2000 Producing users, using producers. *Participatory Design Conference New York* 28(11).

Rivoal, Isabelle and Salazar Noel B.
2013 Contemporary ethnographic practice and the value of serendipity. *Social Anthropology* 21(2), 178-185.

Smets, Annelien, Lievens Bram and D'Hauwers Ruben
2018 Context-aware experience sampling method to understand human behavior in a smart city: a case study. *Measuring Behavior 2018 Conference Proceedings*, 323-329.

Spagnoli, Francesca, van der Graaf Shenja and Ballon Pieter
2017 Changing behavior in the digital era: a concrete use case from the health domain. *World Academy of Science, Engineering and Technology International Journal of Psychological and Behavioral Sciences* 11.

Veeckman, Carina and van der Graaf Shenja
2014 The city as living labortory: A playground for the innovative development of smart city applications. *Engineering, Technology and Innovation (ICE),* 1-10. doi: 10.1109/ICE.2014.6871621

Wang, Tricia
2016 Why Big Data needs Thick Data. Ethnography Matters. https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7

The Perfect uberPOOL: A Case Study on Trade-Offs

JENNY LO
Uber Technologies

STEVE MORSEMAN
Uber Technologies

One of Uber's company missions is to make carpooling more affordable and reliable for riders, and effortless for drivers. In 2014 the company launched uberPOOL to make it easy for riders to share their trip with others heading in the same direction. Fundamental to the mechanics of uberPOOL is the intelligence that matches riders for a trip, which can introduce various uncertainties into the user experience. Core to the business objective is understanding how to deliver a 'Perfect POOL'—an ideal situation where 3 people in the vehicle are able to get in and out at the same time and location allowing for a more predictable and affordable experience. This case study argues that, for a reduced fare and a more direct route, riders are willing to forego the convenience of getting picked up at their door in exchange for waiting and walking a set amount to meet their driver.

This case study explores the integration of qualitative and quantitative research to understand user trade-offs. Methods utilized were in-person interviews and two large-scale surveys: a maxdiff and a conjoint, each with a different purpose. The study started with a multi-city qualitative research study designed to understand how users make trade-offs among their transportation options, suggesting key characteristics of a 'Perfect POOL.' The team followed up with a maxdiff survey to validate these characteristics and identify the factors most important for riders' decisions. A customized conjoint survey was then built to study what values each product feature contributes to maximize rider opt-in to the 'Perfect POOL' product. The team subsequently explored ways to translate the trade-offs revealed by the conjoint survey back into the product experience. This case study will discuss the conjoint survey's outcomes and implications that directly confirmed the hypothesis that riders are willing to make experiential trade-offs. Learnings from this multi-phase research led to the initial Beta-launch of Express POOL in November 2017.

INTRODUCTION

Uber started with a simple concept—press a button and request a ride to your destination. Founded in 2009, Uber is an on-demand transportation platform that enables individuals to get a ride using their mobile phone. What started as a luxury ride service quickly became a global logistics platform changing how people move around. Today, Uber is available in more than 600 cities around the world, transporting riders in hundreds of languages across dozens of countries. Uber is committed to making transportation safer and more accessible, reducing the congestion impact of urban transportation by getting more people into fewer cars, and creating economic opportunities for people to work on their own terms. Core to realizing this mission is promoting carpooling at scale to enable everyone to afford the experience of Uber.

UberPOOL was originally launched in August 2014 as a service that makes it easy for people headed in the same direction to share their journey. The overall benefit was lower costs for riders and a higher volume of paying passengers for drivers. Since its launch, uberPOOL has become a popular carpooling service for riders worldwide and has served over one billion rides. Moreover, uberPOOL constitutes a large portion of the company's overall business. As such, the service satisfies customer needs, strategically grows Uber's business, and benefits cities by improving the usage of each car on the road.

That said, prior user research identified some critical rider experience pain points on uberPOOL, particularly around routing and affordability. Poor matches cause significant routing detours and prices are not affordable enough. Combined with the goal of delivering a more efficient and enjoyable service, the team focused on creating the 'Perfect uberPOOL,' an ideal situation where all 3 people in a vehicle are able to get in and out at the same time, leading to a more predictable and affordable experience.

The research team, comprised of user researchers and data-scientists, devised a multi-phase research study approach to investigate what and how to create the 'Perfect uberPOOL' from the riders' point of view. The research started with an in-depth qualitative study across multiple cities to understand how riders make trade-off decisions when choosing transportation. Findings from this study informed the core characteristics of the 'Perfect uberPOOL.' As a result, the Product, Engineering and Research teams formulated a set of refined hypotheses:

- Riders are willing to wait for a short period and walk to an optimal pickup location in exchange for a cheaper price and a faster, more direct route to their destination;
- Riders are willing to trade a predictable pre-trip experience for a lower price and a higher quality on-trip experience with fewer detours.

The dual purpose of this paper is to demonstrate collaborative research processes and to assist readers in executing similar research methods. Each stage of the research is discussed, detailing how the team collaborated in continuous and interdependent knowledge building. Details on the methodological approach and execution are intentionally included to support similar types of inquiry.

Section one introduces the mechanics of the original uberPOOL, beginning with an overview of how the product works and the basic unit economics underlying it. Included is a walk-through of the current interface design of uberPOOL. Section two discusses the motivations to improve uberPOOL and the existing concerns across the three main actors: Uber as a company, riders and drivers. Section three focuses on the in-depth qualitative study conducted to understand how riders make trade-off decisions when choosing their transportation. The research suggested key characteristics in creating a 'Perfect POOL.' Section four is devoted to the design of a maximum differentiation ("maxdiff") analysis survey that was constructed to validate the most important factors for a rider when choosing uberPOOL. Outcomes from the maxdiff survey helped the team focus on a core set of features. Section five discusses the conjoint survey that was then built to understand the value of each product feature in order to maximize rider opt-in. The conjoint survey design with its implementation and analysis will be discussed in detail. Section six is an overall summary of the research outcomes and business learnings that led to the launch of the resulting product named Express POOL in November 2017.

DYNAMICS OF UBERPOOL

The fundamental objective of uberPOOL is to promote the efficiency of the service by filling as many available seats in a car as possible, while ensuring an enjoyable and delightful experience for both riders and drivers. With more riders in a vehicle, the costs of riding are shared across more individuals. A lowered cost of transportation provides greater access for a broader set of riders and can unlock new use cases. UberPOOL is also strategically important for Uber's growth as it allows the company to service more rides with more paying passengers in the car for drivers. In essence, uberPOOL can create a holistically beneficial scenario wherein riders benefit from lower fare costs; drivers' time and vehicles are optimally utilized; and Uber is able to fulfill more trips.

The following is a walkthrough of the user experience of ordering an uberPOOL as of September 2017. When a rider chooses uberPOOL in the Uber rider app, they will be shown an upfront fare along with an estimated time of arrival ("ETA"). Riders can also see alternative service choices, such as uberX (a solo-ride with no additional riders[1]), as well as compare cost and time estimates. Once a rider inputs a destination and requests an uberPOOL, he or she will be connected to a driver and be given that driver's name, estimated pick up time ("EDT") , license plate number, vehicle description, and the names of other riders in the carpool. In the Uber app riders can also see the driver's route and the location where other riders will be picked up and dropped off.

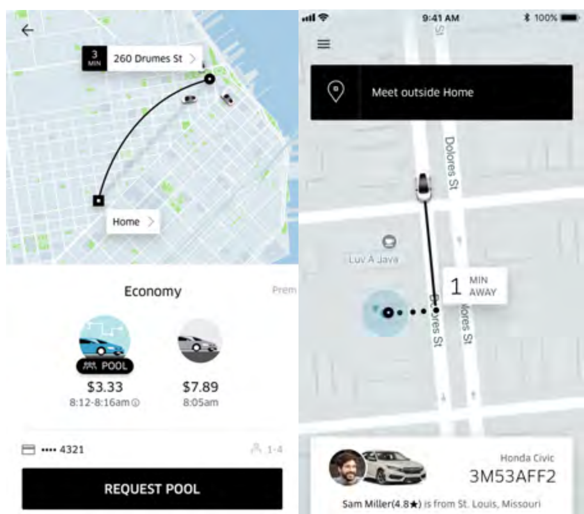


Figure 1. Interface design uberPOOL in the Uber rider app. Designs as of mid-2017.

This popular carpooling service is available at a lower price because riders are able to split the cost of the trip across multiple riders. With uberPOOL, a rider can pay as much as 50% less than uberX, depending on the city, making it most often Uber's cheapest service. In terms of the costs of providing the trip, it includes payout to the driver and Uber. For illustrative purposes, a trip going from point A to point B might cost $12 in total. A rider's cost for that trip will depend on both the number of riders splitting the cost, and the amount of trip overlap among them. Figure 2 illustrates the unit economics of an uberPOOL across 3 simplified scenarios:

***Solo Trip (uberX)* –** In this scenario, there is only one rider going from point A to point B. If the cost of the trip is $12, then he/she will be responsible for the entire cost of the trip. This would have the highest rider cost out of the three scenarios. In terms of Uber's ecosystem, if every rider requested an uberX, then the platform would need a significant number of drivers to fulfill all the requests. If Uber cannot ensure that there are enough drivers, then riders will have to experience longer wait times, resulting in a less ideal experience along at a higher cost.

***An Imperfect POOL* –** For the two riders in this scenario, one might be going from point A to point B, when the second rider is picked up halfway to the destination. The cost of the trip at $12 will be split unevenly between the two. That is, the first rider will be charged $8 for the entire trip and the second rider will pay $4 for only riding halfway. As a result, the first rider receives a small discount paying $8 instead of $12 for riding alone. However, he/she will likely experience added detours and time delays due to picking up the second rider on the way to point B. For Uber, if the second rider is out of the way for the first rider, the cost of the trip for Uber will actually exceed $12, as the trip takes more time and is of greater distance.

***A Perfect POOL* –** This is a situation when a vehicle is transporting three people and they are able to get in and out at the same time and location. Accordingly, a perfect POOL allows riders to have an experience closer to that of a solo trip, without additional pickups or detours along the way. Each rider will pay a much lower price of $4 per person. For Uber, the cost of the trip will not exceed $12, since there are no additional detours. Moreover, the three rider requests are satisfied with one vehicle, providing the most efficient use of the vehicle and the driver's time. Therefore, cars on the roads can be utilized more efficiently while providing riders with a more affordable option.

To summarize, these three scenarios illustrate the intricate connection across riders, drivers and Uber as a company, as well as the efficiencies of uberPOOL and its concomitant costs and experiential consequences.

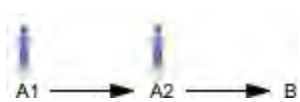
Type of Trip	Illustration	Cost of Trip	Rider Price
Solo Trip (UberX)		$12	$12
Imperfect POOL		$12	$8 + $4
The Perfect POOL		$12	$12/3 = $4

Figure 2. Unit Economics of uberPOOL in 3 Theoretical Scenarios. 'Cost of trip' consists of the driver payout for providing the service. 'Rider price' is the price that the rider will have to pay. Numbers across the 3 theoretical scenarios are for illustrative purposes only.

Since its launch, uberPOOL has made significant improvements; nevertheless, cases of 'imperfect POOLs' still exist and impact the experience and economics of the product. With uberPOOL contributing more than 20% of all Uber trips in cities where it is available, poor experiences lead to major consequences to riders, drivers and the future of the business at a massive scale. Therefore, the opportunity to redefine uberPOOL can help set Uber up for long term success by eliminating sub-par and unsatisfying trip experiences for both drivers and riders, while attaining sustainably optimal and viable unit economics.

The following section discusses the team's motivations to improve uberPOOL and the existing concerns affecting the three main actors: Uber as a company, riders and drivers.

Removing inefficiencies – uberPOOL represents a significant portion of Uber's business; results of inefficiencies (like those inherent in imperfect POOLs) lead to a higher cost for riders, drivers and Uber. While the minimum efficiency reduction is defined by the overall business, the Product team needs to identify how to realize this - part of which is to identify what the system needs in order to be technologically efficient while meeting the needs of customers.

Flywheel effect - uberPOOL is based upon a shared ride concept, by which the experience and economics only get better as more people use it. For example, the greater the volume of riders, the higher chance of finding the best pairing of others going to the same destination. Such a pairing would also translate to shorter rider wait times. For the business, it can create a virtuous cycle of riders wherein the product market fit and economic accessibility allows the company to scale along with the density of usage. This in turn creates a flywheel effect that can enable the product to improve.

Affordability - Achieving perfect POOL trips significantly lowers the trip cost of each ride and rider's price. This helps Uber reach a broader set of riders and unlock new use cases for riders. Past research has shown affordability to be a key factor leading to greater rider adoption. As riders prefer to use Uber as a service more, it may be cost-prohibitive on on a routine and consistent basis. Moreover, this situation could be further exacerbated when riders need the service during times that are busier and, hence, more expensive. As such, perfect POOL trips should be able to reduce the cost of the ride and make the service more financially accessible for a broader set of riders.

Streamlined Trips - Drivers typically complain about the way imperfect POOLs require more time and effort to complete trips, which is mentally taxing for them. Moreover, drivers have mentioned emotional discomfort and burden from interacting with riders upset about the added time with POOL. In contrast, streamlined trips following an organized plan result in simplified drivers experiences that drives describe 'magical,' particularly when the coordination of pickups and routes is straightforward and efficient for everyone.

Riders also complain that picking up co-riders can make trips unpredictable and inconsistent. A rider from New York City said "I will never take POOL when I need to be somewhere at a specific time" - a common sentiment. On the left of Figure 3 is an example of an imperfect POOL trip in New York City with inefficient co-rider pickups and dropoffs characterized by suboptimal zigzagging and looping. On the right of Figure 3 is a photo of a rider sketching an ideal route along main street arteries in New York City. This rider is not alone in preferring simplified routing. While riders enjoy the cost savings of imperfect POOLs compared to the price of a solo trip, poor matching can cause them undue frustration.

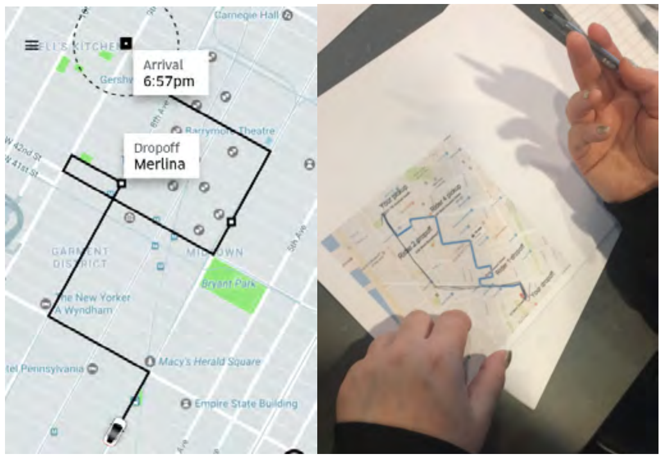


Figure 3. Example of an imperfect POOL trip in New York City (Left). Rider shares opinion of an ideal route on uberPOOL (Right).

Research Design

With problems and opportunity defined, the team pursued a 'Perfect POOL' solution that would meet the needs of customers and business objectives of Uber alike. To do so, a multi-phase research plan was devised to layer learnings to drive actionable insights. Using a novel approach, the team integrated both qualitative and quantitative methods uniquely suited to revealing the optimal technical directions and potential product features of the 'Perfect POOL.' A guiding principle of the research was to ensure that learnings derived from one stage of research would build upon the next to attain the most comprehensive and actionable findings.

Figure 4. Main research stages prior to product launch. Post-launch research studies will not be discussed in this paper.

First and foremost, the team wanted to understand how and why people make the transportation decisions they do. The team started with a qualitative user study utilizing the Jobs-to-be-Done framework (Christensen, 2004). Through in-home rider interviews and ride-alongs, trade-offs in riders' travel choices were revealed. This allowed the team to identify the travel decision criteria that would evolve into a comprehensive list of considerations most relevant to rebuilding uberPOOL.

To complement the qualitative, the team next applied a quantitative approach, using the list of travel decision criteria into a maxdiff survey. The maxdiff used a best-to-worst scaling approach to identify, narrow and validate the relative importance of trip attributes for both user and product. This instrument resulted in an initial set of product levers.

A subsequent conjoint survey populated with the final list of product levers was used to populate a designed to understand the likelihood that riders would consider waiting and walking for a cheaper ride. This unique and novel method simulated realistic purchase scenarios for respondents. The team subsequently explored ways to translate the trade-offs revealed by the conjoint survey back into the product experience. While this paper primarily focuses on the first three phases of research mentioned above, the final section will discuss the important considerations and usability findings needed to translate conjoint insights into product design decisions. The sum of these learnings guided the final ideation of the new product experience.

IN-HOME QUALITATIVE RESEARCH

Everyday people travel to different places for a purpose. Whether an individual is on their daily commute to work or a family is traveling for a vacation, everyone has a specific purpose for their journey. To fully understand an individual's goal or 'job,' it is important to understand the progress they are trying to make under particular circumstances (Christensen, 2004). When a customer buys products or services, they are 'hiring' them to complete a specific job. Customers return if the job is well done. If not, customers will replace the product or service, and look for alternatives that can better satisfy their goal. Therefore, knowing a customer's diverse set of needs and what they are trying to accomplish in a given circumstance explains why customers choose what they use today.

For the purpose of this study, the research team approached the work of understanding 'jobs' through a three-step process. First, the team identified the customers' goals by paying close attention to the context and circumstances that shape the customers' thinking. Second, the team took into account all the functional, emotional, and meaning-based dimensions that govern a transportation choices. This requires knowing what factors constitute each dimension and how customers think through these factors. Third is knowing how customers reason through and evaluate the type of trade-offs they are willing to make. Given a set of choices, customers evaluate and select available options against their 'jobs' specific to the circumstance. Holistically, this process is key to knowing why customers stay with their existing choice, or change to alternate choices.

Qualitative Research Logistics

The research team started the study with a series of in-home interviews with riders in different locales. The study included 23 users across various neighborhoods in Chicago and Washington D.C. The cities were selected based on city density, rider diversity, product performance and business priority. Participants included a mix of prospective riders, new riders and tenured riders spread across specific predominant use cases. For example, the team categorized the main use-cases as: commute to or from work, airport or business travel, social outings and family errands. The team screened for participants exhibiting behaviors within these key uses cases to ensure that their travel experiences were within the realm of those that Uber supports.

Each participant session lasted 2.5 hours and was conducted at their house. The session was comprised of three main sections: a general travel-mapping exercise to understand the rider's lifestyle and travel occasions; a job exploration section to understand the factors and decision-making process; and finally, a ride-along section to capture the context and nature that govern a top key job.

The interview portion of the study is focused on understanding participants' feelings towards travel and a brief overview of their approach to travel. The team employed a travel mapping exercise as a grounding document to anchor and catalog all their travel occasions. This mapping exercise provides a systematic framework for soliciting discussion points. Each item on the travel map would indicate a particular occasion. For example, one participant mentioned taking their child to school every day as part of their daily routine. Therefore, dropping their child at school is one of their top travel occasions. As the discussion progresses, each occasion is built out with greater detail, uncovering details on the who, when, and what of that particular circumstance. This process of documentation provides an initial overview of the types of occasions that riders have as part of their travel.

The second portion of the research study is an in-depth discussion of each travel occasion to understand the complete breadth of jobs that are associated with each occasion. In this study, Ulwick's eight fundamental process steps were utilized to guide the discussion of each occasion. The steps were to: define, locate, prepare, confirm, execute, monitor, modify and conclude (Ulwick, 2016). The first step, 'Define,' requires understanding the participant's main objective. Each following step is a slight progression of their process, demonstrating how participants make calculated trade-offs between various needs when considering their transportation choices.

Qualitative Research Learnings

This research surfaced the *breadth* and *interaction* of the various functional, emotional and meaning-based factors that arise in users' travel decisions. This summarization framework is a simplified adaptation from Maslow's Hierarchy of Needs, a common framework used for organizing human motivations (Maslow, 2013). According to Ulwick, a 'functional' job is the core task that has to be accomplished. An 'emotional' job is defined as the way customers want to feel or want to avoid feeling during the process (Ulwick, 2017). And finally, a 'meaning-based' job is the self-actualization thought process of how the customer wants to be perceived by others. A crucial learning from this qualitative study was recognizing not only the magnitude of functional factors that govern users' transportation decisions, but also the emotional factors that can play a significant role in a user's decision process.

In this study, identified 'functional' jobs include factors such as price, efficiency and vehicle size. Riders are oftentimes much more vocal and aware of these functional factors because they are the core tasks that have to be accomplished in their particular travel circumstance. For example, a common rider task might be to ensure arriving on time at a particular destination. In this study, a participant noted their responsibilities as a mother, where "the school bus arrives at 7:20am, so at 6:45am I will need to drive [my child] down the street to the babysitter, where he will wait and then board the school bus." As the participant shared this particular occasion, she explained how the current travel arrangement is ideal for her work schedule, but it would be beyond her budget if she had to continue this arrangement due to the cost of the school bus. In that instance, she described the functional goal of minimizing the cost expenditure of their travel option to stay within bounds of their budget. In this case, due to the participant's price constraints, she has to trade convenience and efficiency for price.

Efficiency is a key component of uberPOOL and through this study, the team is able to understand how riders talked about this important concept. Riders mentioned topics such as route planning, trip duration, amount of waiting time, and arrival time variability - all of which are aspects of efficiency in riders' travel choices. For example, a rider said "the way I approach travel...I don't know if this is unique but I always make sure that I know more than one route that I can take in case there's traffic." In this case, the rider is concerned about traffic affecting her trip duration and, therefore, paid more attention to route planning. One of this rider's functional goals is to identify the best possible route to her destination, but 'emotionally' she is also trying to increase confidence in her overall travel plan by creating a secondary plan. As such, this illustrated to the team how efficiency trade-offs do not live in isolation but are interconnected with other factors.

Moreover, efficiency factors, such as trip duration, can be interpreted as actual or perceived. For example, a rider described using other sources of travel information. The rider mentioned that by entering the "time you want to be there, [the app] will say 'traffic is usually heavy around this time.' So if you leave at this time, this is how long it's going to take you. So I use [the app] to let me know so I can leave far enough in advance." This would be a case where a rider has actual time predictions that inform them on the efficiency of their trip. In other cases, a rider might believe that a particular route will take longer based on past experience. This terminology around efficiency added a new way to frame of how riders evaluate perceived vs actual efficiency differences in their travel choices.

In summary, riders make trade-offs on trip attributes across all three 'functional,' 'emotional' and 'meaning-based' dimensions. In Figure 4, a participant walks through how his job as a police officer is mostly urgent, unstructured and unplanned - namely that "nothing is the same every day." With his varying destinations and schedule, the participant emphasized the need for a more time-efficient but spontaneous travel arrangement. Meanwhile, he also noted strict values against drinking and driving, concerns around DUI and the need for travel decisions to include how to "help keep people more safe." This example illustrates how 'functionally' time-efficiency is key; 'emotionally' the user needs control, and "meaning" where it supports his personal values, which is on individual and community safety. Considered together, this frames a rider's decision making model for travel choices.

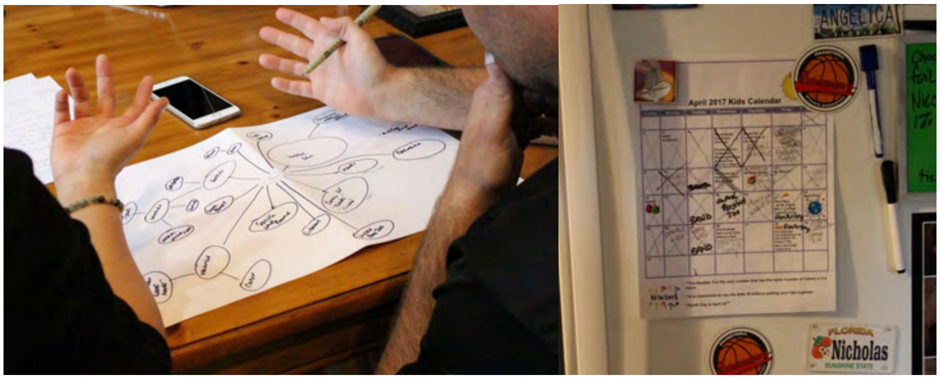


Figure 5. A rider explains the functional, emotional, and meaning-based trade-offs through the travel-mapping exercise (Left). A mother refers to her child's school schedule to guide her travel-mapping exercise (Right).

Understanding Trade-offs in the Context of uberPOOL

The research team recognized how riders' travel decisions cut across functional, emotional and meaning-based dimensions. But as first step in building the perfect POOL, the team had to narrow down the list of factors they had more direct influence and control over, such as the matching intelligence and efficiency attributes of uberPOOL. To do so, they proposed first utilizing a maxdiff survey and then a conjoint survey. After launching early algorithmic changes, the team would return to additional qualitative research assessments for a holistic re-evaluation of the tradeoffs, including an assessment of how users' functional, emotional and meaning-based dimensions interact across the product. Accordingly, the team strove to understand the intricacies of efficiency attributes and how they should manifest in uberPOOL's matching algorithm.

This also means that emotional needs, such as safety or concerns in sharing a car with strangers, are important factors but given that these factors are less controllable from a product perspective and more on a policy perspective, the research team proposed first focusing on understanding factors that the team has more direct influence over and supplement future work on understanding these emotional needs. This work was conducted post-launch but will not be discussed in this paper.

The uberPool algorithm that matches multiple riders with a driver introduces uncertainties to both user groups. When riders request an uberPOOL, they are only provided the bare essential information such as upfront price, approximate time of the driver's arrival, and an estimated time of arrival at their destination. Even though they are told that uberPOOL is a carpooling product, they cannot be certain whether additional riders will actually join them. Riders also do not know ahead of time, when, where, and with whom this sharing will happen. This is because matching decisions happen in real-time. Meaning that users are only provided that information after the decision has been made. Even after the original matching has been made, the result may entail picking up two additional riders, adding 10-minutes to the original rider's on-trip time. Hence, part of the uberPOOL experience involves coping with these uncertainties around the types of inconveniences they will experience during their ride.

An important distinction discovered from the qualitative study is how riders might have control over the types of inefficiencies that they will experience, compared to unknown inconveniences embedded in an uberPOOL experience. For example, a rider driving in their own vehicle but stuck in traffic, will have more control on what route planning to take. However, in the case of an uberPOOL, these routing decisions are decided by Uber, not the rider. As such, the rider will also not know the route that is designed before selecting uberPOOL. Therefore, the qualitative study sought to capture the added nuance of 'ownership' and 'transparency' of what inconveniences to be experienced by riders.

Illustrative Example - Sarah is a teacher at a school in the Hunters Point neighborhood of San Francisco; she commutes from her home 1-hour away. On a particular day, she needs to carry some class equipment and get to school by 8am. She has a transportation budget of around $10 per trip. She wakes up at 6:30am, considers her options, and then decides whether to request an uberPOOL. The Uber team understands that riders like Sarah have to evaluate transportation options against set requirements. Evaluating uberPOOL against these factors can be challenging for riders because they are not privy to complete information about the experience ahead of time, such as how many riders they will share their ride with, how much added time the additional matches will add to their on-trip time, and the overall quality of the route that may also affect the time of arrival.

As a result, the team questioned whether communicating some of these inconveniences that were traditionally unknown to them might be valuable to riders' decisions. For example, riders do not know ahead of time the number of additional riders joining the trip or know where the defined route will take them. The team asks whether these efficiency factors need to be communicated upfront to riders. However, providing such information would come at a cost to Uber and can only be justified if riders deem it valuable in their travel choice. Therefore, the team conducted a maxdiff survey to validate top factors that shape riders' decisions on uberPOOL and whether certain efficiency factors, if communicated upfront, would significantly alter their travel choice.

MAXIMUM DIFFERENTIATION SURVEY (Maxdiff)

The in-home qualitative research with uberPOOL riders provided the research team with insights about key needs informing their existing transportation choices and what they considered the perfect POOL. In order to test and scale the interview findings, the next phase of research included two broad-based online surveys measuring the relative value of their choices. The following section defines the first of the surveys - a maximum differentiation ("maxdiff") analysis - and discusses its design, results and impact on the perfect POOL.

Moving to an online survey allowed the researchers to overcome some common issues found in small sample sized interview methods, namely selection bias (Collier & Mahoney 1996). Interviewers often encounter selection bias in their panelists due to geographic, behavioral, and economic factors. The lower commitment costs of an online survey somewhat mitigate selection bias due to less demand on respondents' time. Employing a survey was also beneficial due to stratified sampling & survey quota capabilities, which allow for a more representative rider population than small-n studies. This ensured a rough approximation of the underlying population (Qualtrics Quotas). The survey send was stratified by rider tenure, city, & product mix (uberX and uberPOOL usage) to ensure that the overall results were reflective of the target market and that the data could be segmented by these rider dimensions.[2]

Understanding Rider Preferences

A non-trivial issue when studying consumer preferences is creating meaningful separations across various product characteristics. Researchers that ask survey respondents to stack-rank desirable product attributes may get only minimal differentiation in the resulting dataset (Epstein 2018). These stack-ranked questions are often coded with ordinal values, and, in many cases, the mean value across all respondents is reported as the primary summary statistic (Lovelace & Brickman 2013). Typically the stacked-rank approach results in point estimates with overlapping confidence intervals. As such,there is no statistically significant difference between the estimates. Hence, these figures often have little difference in value, and, with the exception of extreme outliers, provide little incremental insight to the researcher.

For this study a maxdiff survey was implemented as a better approach to derive the underlying value of each product feature. A maxdiff forces respondents to choose the option they like most and least out of a larger set of answer choices. Even if a respondent values two items nearly identically, a maxdiff will force them to think critically about the pair and order one over another by preference. This encourages differences to emerge that might not be possible with the simple ranked-choice method. The resulting dataset forms the basis of the maxdiff analysis. Running a logistic regression on a transformed version of this dataset will provide coefficient values that can then be rescaled and used to gain insights around what consumers care about most and least (Hess 2014).

There are no strict rules when designing a maxdiff survey, as both the number of answer choices and question sets used in the survey can vary. Best practices in research design dictate that a maxdiff should minimize the number of questions, have answer choices appear at least three times, and ensure that each answer choice is shown to the respondent an equal number of times (Porter & Weitzer 2004, Sawtooth Software 2013). Following this approach minimizes the survey fatigue of respondents and increases the overall survey completion rate.

In designing the maxdiff survey, the research team limited the set of product attributes to those gained from the qualitative research. Some attributes were removed from the maxdiff due to the lack of actionability given the existing marketplace; these included car type and number of co-riders. Other attributes like price were eliminated from the survey because they were already selected for inclusion in the follow up conjoint analysis survey due to their immense importance.

The maxdiff survey consisted of a six question-balanced incomplete block design, with four-answer options where each choice appeared exactly three times in the research design (Graham & Cable, 2001).[3] An incomplete block design (IBD) is an experimental framework used when it is not possible to include all treatments (product features) in every block (survey question). IBD may be necessary due to financial, labor, or cognitive limitations on the survey respondent or organization conducting the study. Uber researchers chose an IBD in order to facilitate the forced trade-offs between subsets of product features in a balanced way.

All of the maxdiff questions were constructed in the same fashion, with each question asking the respondent, "Of the following four options, which is the most important and which is the least important factor for you when deciding to ride uberPOOL?" Of the four answer choices, respondents are required to select one option that is the most important and one that is the least important when considering whether or not to use the uberPOOL product. Figure 6 shows a sample question used by the Uber research team in the maxdiff survey.

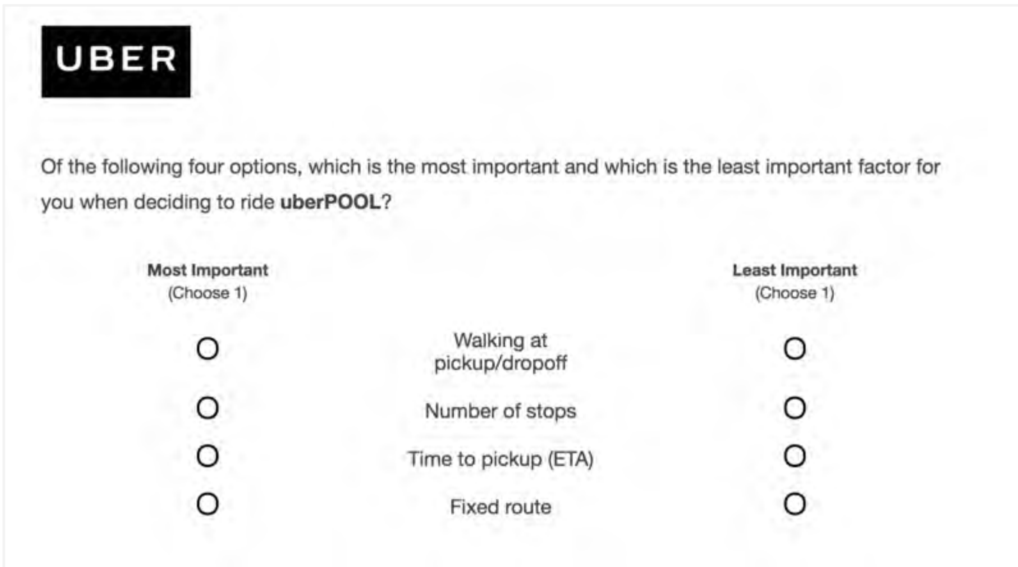


Figure 6. Example maximum differentiation survey question

The researchers implemented the maxdiff survey on the Qualtrics survey platform. They randomized both the order of the questions and the answer choices within each question, using this functionality to minimize bias introduced through survey respondents who use simplification strategies, such as selecting the first answer choice for every question (Qualtrics Question Randomization). In addition to the maxdiff questions, researchers also asked about respondents' current uberPOOL usage, commuting habits, as well as open-ended questions around their existing transportation options and future needs.

The survey was sent to Uber riders who took an uberPOOL trip in the 30 days prior to the survey send in the metropolitan regions of Boston, Washington DC, New York City, Chicago, and the San Francisco Bay Area. These markets were selected due to the high prevalence of uberPOOL usage and the density of Uber trips. As such, these geographies made ideal product launch markets and were a natural choice to use for the survey audience. Five-thousand, 30-day active uberPOOL riders from each market were selected at random to receive the survey by email, resulting in 3,000 completed surveys (for a completion rate of approximately 12%). Conditional upon finishing the survey, the mean-time to completion was 22-minutes and the median was 5-minutes.

Logistic Regression

By assuming that riders maximize their utility when answering questions, researchers are able to use discrete choice models, such as logistic regression, to estimate the value of each product feature (Hosmer & Lemeshow, 2010).[4] The logistic regression model can be described by the following equation:

$$E[Y_i \mid X_i] = p_i = logit^{-1}(\beta X_i) = \frac{1}{1+e^{-\beta * X_i}}$$

The equation gives us the probability of a specific product feature *(Y)* being selected by a survey respondent *(i)* conditional upon specific covariates values *(X)*. This is equal to the inverse logit function of the product between respondent *i*'s covariates and the coefficient values *(β)* of each variable (Marley & Pihlens, 2012). The logistic function, or the natural logarithm of the odds, takes any real number input and outputs a value between zero and one. It is a particularly useful tool when computing the probability that an event occurs.

The raw Qualtrics survey data needed to be reconstructed to be used in a logistic regression model.[5] The independent variables in the equation are dummy variables representing each answer choice in the research design and are coded "1" if the respondent thinks the answer choice is best and as "-1" if it was worst. The coefficient values from the fitted logistic regression can then be directly compared - or normalized - to estimate the relative preference of each answer choice (Marley 2018).[6]

The main findings from the analysis indicate on-trip duration to be riders' most important product attribute when deciding whether to use uberPOOL; duration was followed by the estimated time to arrival (ETA) and on-demand availability of Ubers. Last came fixed routes, number of stops on trip, and finally walking to and from the pickup/dropoff location.

The most important factor - that riders care most about efficiency - corroborated the insights from the interviews. Specifically the maxdiff showed walking to be riders' least important attribute when deciding to use Uber. Walking is also the least expensive feature to implement within our existing matching & routing algorithm. Coupled with the being able to coalesce riders' pickup locations together means that Uber would be able to provide the greatest possible gains in efficiency with the lowest amount of additional inconvenience to the rider.

Another important insight for the research team was that having a fixed route, or limiting the number of stops, was a relatively unimportant product feature to consumers. This complemented the qualitative research that identified poor quality routing as a factor stunting future uberPOOL adoption. Just because riders view trip-routing as a issue doesn't mean they believe that having a fixed-route is the solution. This fixed-route insight allowed the team to shift gears and focus on alternative product features that would be less expensive on Uber's dispatch-matching algorithm and result in a more enjoyable experience for our riders. In short the maxdiff results gave the team a better understanding of what product features constitute the perfect pool.

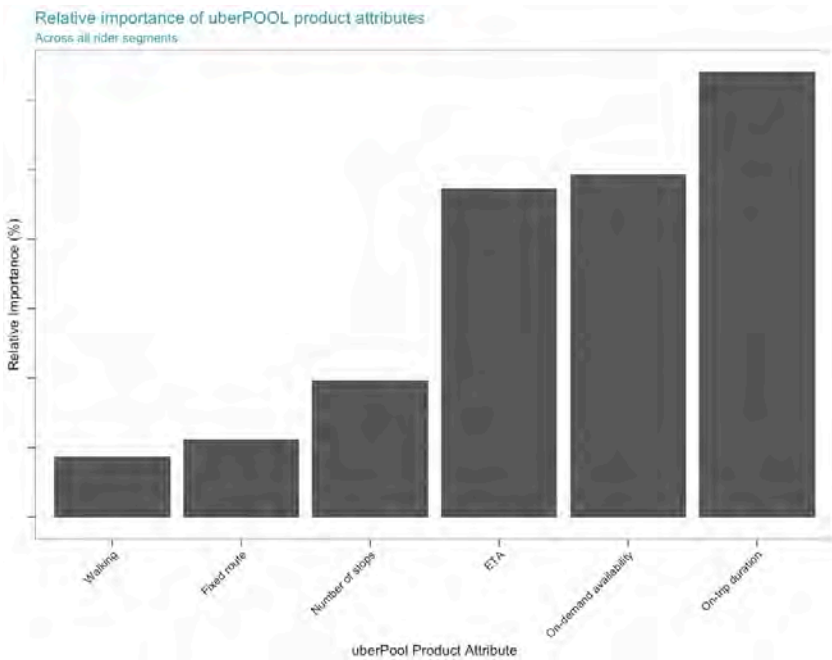


Figure 7. This plot shows the relative difference in the importance of feature utilities, as derived from the maxdiff analysis. Exact percentages have been abstracted to protect business insights, but all the features in the plot add to 100%.

CONJOINT ANALYSIS

The qualitative interviews gave insights to which features riders wanted in an improved uberPOOL product and the maxdiff found which of these features they considered the most important. The next step to understand the perfect pool was to determine what values each of these features should take on. For example, the maxdiff identified walking as an important feature but researchers needed to understand how much walking riders actually find acceptable. To unlock these insights, researchers turned to a conjoint analysis as their methodological tool.

A conjoint analysis is a survey based method used by market researchers to determine how people value different attributes that make up a whole product or service. They can then use these insights to create new products or tweak existing ones to increase market share or optimize profits.[7] Like the maxdiff technique, the conjoint analysis is composed of forced choice questions that researchers analyze using statistical models to determine the underlying value of these product features (Hess 2014).

Conjoint survey questions give respondents a variety of hypothetical products and asks them to choose the option they would be most likely purchase in real life.[8] Each hypothetical product is made up of a bundle of attributes and the value, or level, of each attribute varies randomly from question to question. Conjoint questions will typically include an opt-out option for respondents to more accurately mimic a purchasing situation, where a consumer always has the choice to walk away. Having a “none” option is important because it allows researchers to understand the baseline threshold at which people are willing to buy something. For the uberPOOL study, this threshold corresponds to the baseline utility a rider gets from continuing to use their existing commute option. Riders will only alter their commute when they derive more value from hypothetical uberPOOL product than do they from their status quo situation (personal vehicle, trains, buses, etc). By having the product bundles change randomly from question to question, researchers are able to estimate the value of any individual product feature-level, the interaction effects that may exist between various attributes, and their opt-out threshold.[9]

Good conjoint questions mimic the purchasing decision process as closely as possible, with the respondent having clear context for the choices being made. Conjoints can suffer from a variety of biases, and a good survey design should try and mitigate them using concise and clear questions to avoid cognitive overload on the respondent. These biases can be introduced by questions that are too mentally taxing, suboptimally formatted, or lacking enough “skin-in-the-game” to feel like real simulated purchasing decisions. This may cause respondents to resort to satisficing strategies, which may result in poor quality data (Rossi & Allenby 2009). If these obstacles are overcome, then a well-constructed survey allows researchers to create an artificial marketplace where they can predict a new product’s performance relative to market competitors and other status quo options, as well as give insights regarding a company’s potential cannibalization of existing products.

Creating Testable uberPool Scenarios

The goal of the Uber research team was to create a conjoint that uncovers what constitutes the perfect pool while avoiding the common pitfalls and biases of conjoint surveys. To this end the conjoint developed by the team was more sophisticated than a typical static design, and the questions dynamically changed based on user inputs to create a fully customized survey that closely relates to their existing commuting experience.

Table 1. Conjoint Features and Levels - Research Design Matrix

Feature	Level 1	Level 2	Level 3	Level 4	Level 5
Estimated time of arrival	Request now and wait 5 mins	Request now and wait 10 mins	Book 15 mins ahead	Book 30 mins ahead	
Walking	No walking	Walk 1-block	Walk 2-3 blocks		
Trip length multiplier	1x	1.1x	1.2x	1.3x	1.4x
Trip variance multiplier	1.1x	1.2x	1.3x		
Discount multiplier	Very low	Low	Medium	High	Very high

Table 1 gives the final product attributes and levels selected for study by the research team. Some of the variables were chosen because they were the three highest valued product features discovered in the maxdiff study: time to destination, on-demand availability, and estimated time of arrival of the driver to pick-up the rider (ETA). In addition to these variables, the Uber research team included a price discount relative to UberX in order to recreate a purchasing experience with the same information available to riders in-app.

The team chose to use a fully randomized research design for the conjoint survey. This design resulted in every possible combination of product feature-levels, resulting in 900 potential product packages for use in the survey.[10] By opting to not have any restrictions on product combinations that may be deemed unrealistic, researchers were able to study both the value of each individual product feature-level as well as estimate interactions among them without violating independence assumptions that would cause the team to systematically over or under predict the importance of model estimates. To maximize the number of profiles considered by respondents while minimizing survey fatigue, the team choose to include 7-conjoint questions in the survey and gave respondents three choices for per question, with two of the choices coming from the 900 possible product packages and the other being an opt-out option.. No assumptions were made about respondents' default commuting option and could be anything; including personal vehicle, train, light rail, bus, uberX, other rideshare products, walking, scooting, a combination of these or none of the above.

The survey was implemented in Qualtrics and made use of a backend server to run the experimental product package randomization and create custom on-trip time to destination & pricing estimates for respondents. Question templates coded to accept variables from the server were used to provide programmable questions that could request server data with the web service functionality in Qualtrics (Web Service - Qualtrics Support).

The conjoint questions differed from the verbatim values found on table 1, which took the form of a customized trip itinerary containing the same set of information that riders see in the Uber app when confirming a pickup for a trip. This information includes an upper and lower bound for the estimated time on-trip, the upfront price of the trip, the amount of walking to get picked up, and the estimated time of arrival of the driver. These custom estimates were configured by asking the survey respondents how long it takes them on average to drive door to door from home to work during their commute hours. An example question as rendered to a respondent can be seen in Figure 8.

Self-reported home-to-work driving time estimates were sent to the server where the product packages were randomly chosen. Each rider was randomly assigned 14 different product packages from the 900 total options;[11] these 14 profiles were then randomly paired together to create the question sets for each respondent. After selecting the product profiles, the attribute-levels were translated into an estimated driving time and price for an UberX trip, after which the discount multiplier associated with the package was applied to the UberX price and reported as the cost of the uberPOOL trip. The lower bound of the on-trip time to destination was calculated by taking the estimated door-to-door driving time and applying the trip length multiplier to that estimate, then configuring the upper bound by applying the trip variance multiplier to the lower bound.[12] The extra effort involved creating these customized conjoint questions helped contextualize the purchasing decision for survey

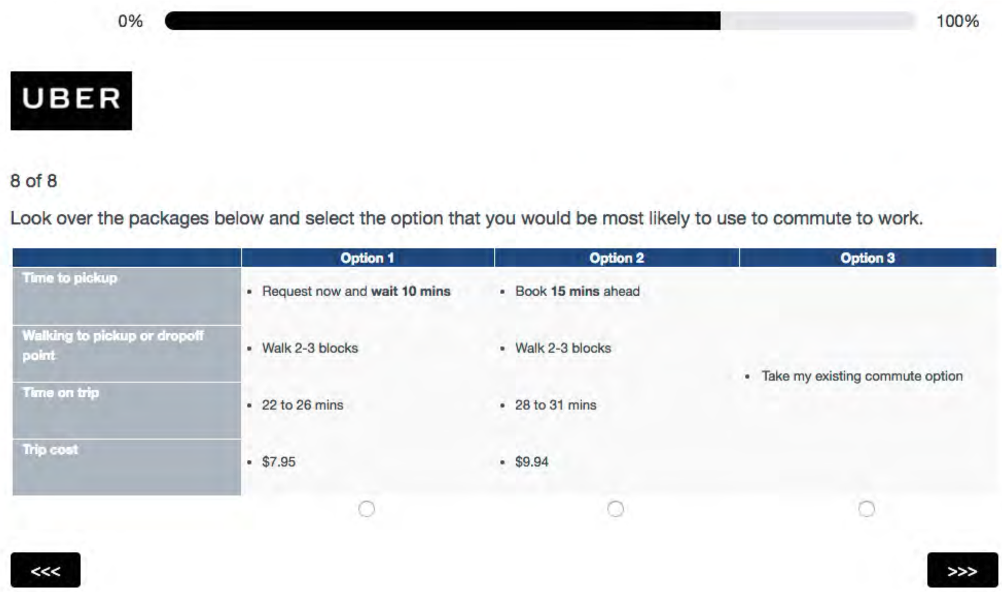


Figure 8. Example conjoint question.

respondents, thus giving researchers higher confidence that the insights from the study are valid and bringing the picture of the perfect pool into greater focus.

The research team conducted multiple rounds of user-testing on the conjoint survey itself to get feedback regarding the ease of comprehension and cognitive overhead. After ensuring that the survey was easily comprehensible by respondents it was sent out to 18,000 riders via email on August 3, 2017, resulting in a total of 1,934 completed surveys and yielding an overall email response rate of 10.7%. The median time to finish the entire survey was 5.5 minutes. Respondents were incentivized to complete it with an opportunity to win one of five $500 Amazon gift cards.

Estimating Feature Utilities

The goal of a conjoint analysis is to obtain the value associated with each feature level. As in the maxdiff analysis, the Uber researcher team assumed that survey respondents choose the hypothetical product package that maximizes their utility. As a result, the part-worth estimates can be seen to represent the value associated with each feature by riders (Rossi & Allenby 2009). Part-worth estimates enables the researcher to compare unrelated features on a common scale and gain insights that would not be possible using purely qualitative approaches. For instance, in the interviews riders indicated that walking and time on trip to destination were both very important to their purchasing decision - yet the conjoint part-worth estimates allowed researchers to precisely quantify the relative worth of these features against each other, or put them directly into a dollar value.

The utilities in this case study were estimated using a hierarchical bayesian multinomial logistic regression. Hierarchical refers to the model estimating utility values at the individual respondent level that can then be aggregated to obtain an overall result, rather than just estimating the aggregate level coefficients for all respondents. This property is useful because it allows for the creation of a mini-marketplace where each person has heterogeneous preferences (Rossi & Allenby 2009). This diversity of opinion can be directly modelled and used to gain deeper understanding for how people interact with the Uber app. Bayesian multinomial logistic regression is a generalized version of the logistic regression covered in the maximum differentiation section that is estimated via simulations for generalized choice data (Rossi 2017).[13,14]

The model was estimated using the bayesm package in the R programming language (Rossi & Allenby 2009).[15] The model can be described by the following specification:

$$\begin{gathered} y_i \sim Multinomial(Pr(X_i, \beta_i)) \\ i = 1, \ldots, \ n \ units \\ \beta_i \sim \Delta' z_i + u_i \\ u_i \sim N(0, V_\beta) \end{gathered}$$

Where the probability of choosing product y for respondent i is distributed multinomially as a function of covariates X_i and β_i. The part-worth estimates (β_i) are distributed logit parameters over respondent units with mean $\Delta' z_.$, with $\Delta' z_{.i}$ being a matrix containing mean-centered control variables for each respondent, with errors (u_i) that are normally distributed with variance V_β (Rossi & Allenby 2009). The posterior distribution of β_i is used to determine the overall utility of each product feature and is the main quantity of interest of the analysis. Researchers used the model's log-likelihood as a measure of goodness of fit, which converged successfully after 100,000 simulations.[16,17]

Control variables used in the analysis include historic Uber usage data, such as the home city of the respondent, rider tenure in days since signing up for an account, lifetime billings, as well as survey-based variables such as the time it took a respondent to complete the survey and demographic information. No other behavioral features were used in the analysis, per Uber's policy of respecting the privacy of user data (Privacy Policy - Uber).

Conjoint Results

Figure 9 shows the expected opt-in rates for feature-levels based on the part-worth utility estimates for each product-feature. Hypothetical products are constructed by selecting a level from within each product attribute and calculating the sum of the part-worth utility estimates. These packages can then be compared against the "none" option that represents the baseline utility a respondent gets from their status quo commute option. Researchers can then construct a simulated marketplace where riders make discrete choices between choosing a new product or not based on this calculus. In practice, the performance of new products within this discrete choice framework are best taken as directional rather than an indication of actual opt-in rates if the product were developed and launched to the public.

Studying the slope of each attribute as levels increases gives powerful insights to the relative worth of each feature to consumers. Unsurprisingly researchers found that price was the most important attribute, with steeper discounts of the uberPOOL product providing positive utility to respondents. All other features represent some degree of inconvenience to the rider and as such have negative utility, with trip variance having the smallest negative impact and walking the largest. The non-linear relationship in ETA utility provided valuable insights regarding respondents' preferences towards waiting longer for a trip, with breaks after 5 and 30 minutes ETAs. This implied a clear need for continuing with an on-demand ridesharing product, while also indicating that riders have an ETA threshold after which their utility decreases substantially. This provided evidence that Uber could increase the efficiency of POOL by making riders wait longer upfront for a larger discount on the trip. The improved efficiency of matching riders with others also results in decreased on-trip duration through improved routes leaving riders and drivers better off.

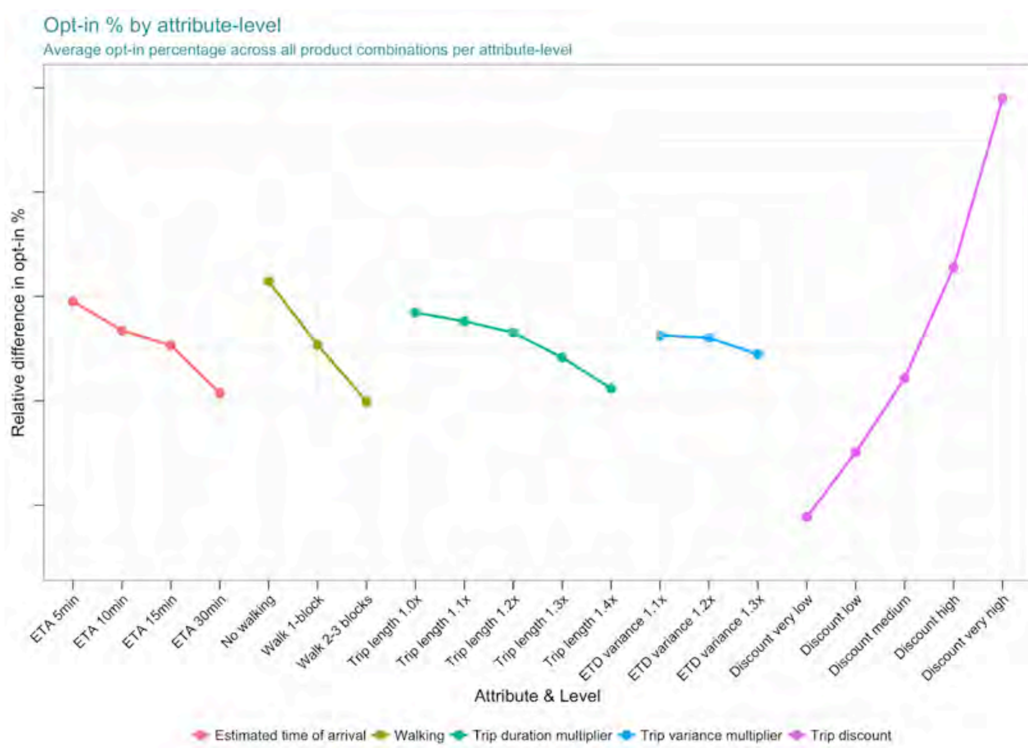


Figure 9. Conjoint results - Overall POOL opt-in by product feature-levels. This plot shows the relative difference in opt-in for each product feature-level, holding all other features constant. Exact percentages have been abstracted to protect business insights.

Segmenting respondents by their commute characteristics and tenure on Uber allowed for more granular insights. Tenured riders that signed up 2+ years ago are more discerning consumers, and are less likely to opt-in than new riders, while newer riders are more price-sensitive with larger increases in utility coming from higher discounts. Segmenting opt-in rates relative to a rider's commute shows that shorter commutes make you more likely to opt in but also more sensitive to walking than riders with a longer commute. These segmented insights provide the framework for creating tailored products that best suite the needs of different market segments.

BUILDING THE PERFECT POOL

This multi-phase research study utilized both qualitative and quantitative methods in building out incremental knowledge towards understanding and building the perfect POOL. Starting with the in-depth qualitative research, the team's approach towards rider trade-offs became more advanced and precise in terms of understanding, implementation and communication. This section outlines the main takeaways from each research phase, and how the team stayed true to rider preferences during product implementation.

Validating Product Concept

The qualitative and quantitative research demonstrated that riders can and will make trade-offs between the inconveniences and benefits of uberPOOL. Riders will accept certain levels of inconvenience such as extended trip length, trip variability, waiting, and walking in return for a lower price and a more direct route. This is because these factors are not foreign concepts and are common across rider's existing travel experience. For example, riders might already walk a few blocks when ordering an Uber for a more convenient pickup. Since aspects of this were already evidence on Uber and other transportation services, the concept was not a far departure from rider's current reality. Therefore, the crucial takeaway was in defining what rider's might expect as an acceptable price in order for them to accept such inconveniences.

Originally, the team was concerned that riders would not be willing to make upfront trade-offs for an improved on-trip experience, such as having a more direct route, but this was not the case. Research demonstrated that some riders do weigh upfront costs, such as effort in pre-planning, to ensure that they have a more ideal travel experience. The team was able to further validate this take-away, showing riders stated preference to wait and walk for a lower price and a faster trip.

The conjoint helped confirm that users find the new value proposition compelling. This confidence was crucial to help align the team on the concept that it was worthwhile to undergo such a massive engineering effort to change the existing uberPOOL product. The mixed-method research approach also provided the team in-person experience with riders, understanding how existing and potential users might perceive the concept. Overall, this grounded team members on the most likely product challenges and helped capture and address concerns of user adoption.

Forming the Narrative on Affordability

The conjoint demonstrated that rider utility increases with lower cost more quickly than disutility increases with inconvenience. Unsurprisingly, it showed how lower price is one of the most important benefits. This was a key strategic piece of evidence for prioritizing the lowest possible cost with uberPOOL that the team was mandated to own and lead efforts towards affordability, driving down price through innovative efficiency solutions.

This evidence not only helped the team identify the significance of price discounts in driving opt-in, but also helped put the magnitude of this business goal into perspective. The conjoint was used repeatedly during offsites, vision exercises, and planning sessions to demonstrate the ideal level of price discounts that the team needed to accomplish. This provided an early signal inspiring the team on the potential growth that could be unlocked. This excitement was shared across various Rider teams, product orgs, and executives on the future of uberPOOL, establishing buy-in and alignment from across the company.

This research demonstrated the power of lower prices, which led the team to put affordability front and center in marketing the product. The team utilized this learning to create a series of marketing claim tests to identify the best messaging. The theme of affordability proved to resonate the most with customers, and the team iterated on numerous concepts to help emphasize this benefit. As such, Express POOL was launched with the focus on savings with the final product tagline 'walk a little, save a lot' to communicate the slight trade-off as evidenced through our research. Media outlets described it as "Uber Express Pool offers the cheapest fares yet in exchange for a little walking." (Hawkins, 2018)

Pricing Decisions

The team was able utilize the conjoint estimates on product opt-in to set realistic, rider-driven pricing and product targets during the development phase and beyond. The team used the conjoint to 'simulate' different configurations of waiting and walking and identified what was an acceptable price point to offset the additional inconvenience.. The team utilized these findings to sanity check pricing to ensure they were not offering an unbalanced product-market fit. For example, with certain levels of walking and waiting, the team utilized the conjoint to get a rider's perspective whether prices would need to go up or down to get a more compelling Express POOL opt-in.

Prior to conducting this research, walking and waiting were previously discussed mostly through the lens of how it affects Uber's ecosystem. Experiential concerns and user metrics, such as opt-in or user feedback, were often delayed inputs measured after product launch. With this new approach the team was able to include users' preferences when setting prices and product parameters Therefore, the conjoint provided estimated rider elasticities that enabled the team to configure the initial launch of the product while also informing subsequent experimentation. With this input, the team was able to discuss product parameters and prices with a more balanced and user-centered approach.

Non-Walking Option

Through this research, the team acquired a nuanced understanding on walking as a trade-off. Segmentation analysis on the conjoint results validated the team's hypotheses that riders have varying sensitivities around walking and waiting, which are influenced by travel conditions and alternatives. As such, walking is not considered uniformly at the same cost across all riders. Rather, its importance fluctuates depending on the rider and the context. The uberPOOL product team recognizes that walking is one of the biggest parameters to balance, because it can provide meaningful efficiency gains to the product experience, but needs to be considerate towards riders with varying walking capability and desires.

This insight around walking ability influenced the team on uberPOOL's product strategy and decision to maintain a non-walking option available to riders at launch. The team identified a subset of riders for whom walking was a great burden and the qualitative research showed the importance of having riders who are motivated to walk and wait, but also able to complete the walking task. Both research inputs identified how divisive walking can be for users, and as a result, the product team believed it was critical to maintain a non-walking shared rides option at product launch.

Translating Conjoint trade-offs into the Product Experience

The team wanted to be faithful to the survey method and translate the trade-offs into the product experience. In a conjoint survey, trade-offs are explicitly described in textual format. However, it is challenging when translating this trade-off into a product experience. As such, the product and design team iterated on numerous ways to communicate the walking and waiting trade-offs throughout the product design.

In terms of walking distance, the conjoint survey utilized a 'blockwise' terminology to indicate the amount of walking that the rider might be expected to do. The respondent was presented choices of 'walk 1 block,' 'walk 2-3 blocks,' or 'no walking' in the conjoint survey. However, the product and engineering team believed that engineering requirements to visually create and communicate such specificity was rather complex. As a result, designers brainstormed on numerous versions and went through rounds of usability testing for potential designs to communicate walking.

An important consideration in going through this design process is assessing the 'specificity' and 'usefulness' that the product and engineering team is able to provide for riders. A circular radius, as depicted below in several design explorations, aimed to communicate the potential walking radius with the potential pickup points to meet their driver. These designs portrayed the spatial area of walking for the rider, did not prove to be useful when selecting their product. In many cases, riders perceived the walking trade-off to be much larger than reality or misinterpreted information about pickup locations.

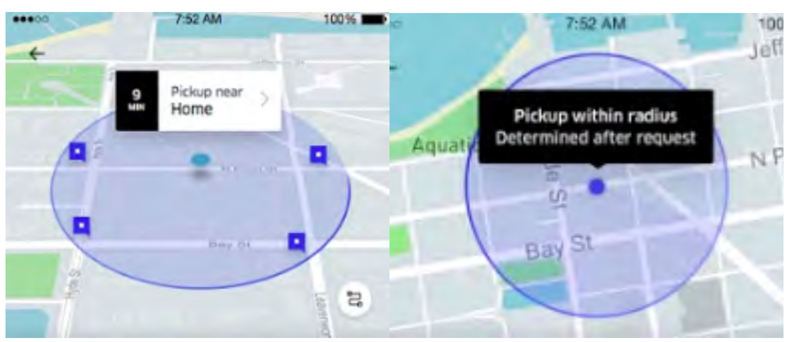


Figure 10. Sample design explorations to communicate spatial trade-offs for walking.

Designers and engineers ideated and created a more engineering complex approach, but believed in investing to better communicate spatial trade-offs for users. The team believed that walking, a critical piece of the user experience, should be useful in helping riders make their trade-offs. As such, the team finalized on a 'bounding box' design that best illustrated the realistic spatial trade-offs for a rider, that received great results from usability tests and after product launch.

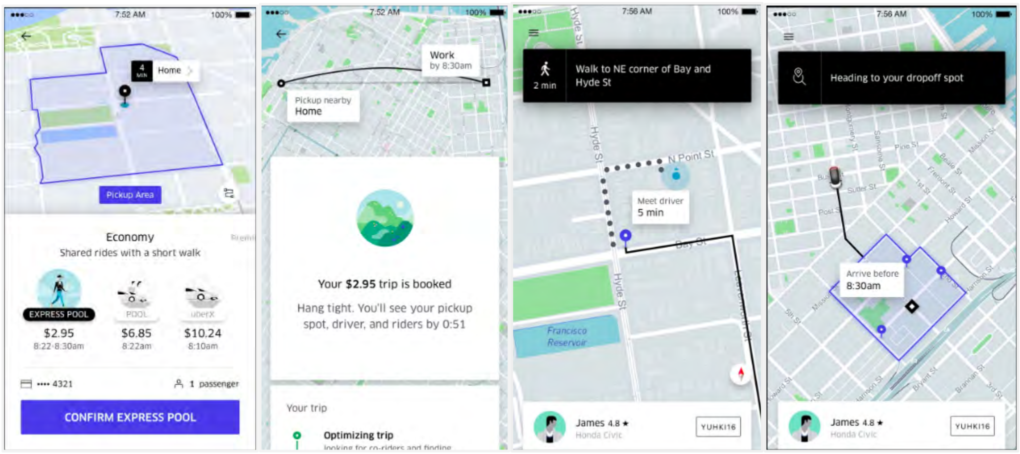


Figure 11. End-to-end Designs of Express POOL - Uber's new shared ride experience with walking and waiting. Designs as of Nov 2017.

Product Impact and Business Outcomes

The research team launched the Express POOL product on November 6th 2017 in San Francisco California and Boston Massachusetts. The rollout of the Express POOL product was spearheaded by Uber's Shared Rides product team in San Francisco and augmented by local operations specialists. A minority of riders in these markets initially qualified to take Express POOL, but after checking the metrics to validate that the product was delivering against expectations it was rolled out to all riders. The product validation was based on the assumption that Uber could improve the uberPOOL experience for riders, drivers, and the company.

The team used trip cancellation rate as one of the many key performance indicators to determine the success of the Express POOL product launch. Trip cancellation rate is defined as the percent of trip requests made by the rider or driver that was cancelled before the driver arrived at the pickup location to start the trip. The metric is experiential, with clear associations between a lower rate and a better rider experience. Other important metrics studied at the product launch include Express POOL opt-in shared-rides rate, driver efficiency & earnings, rider inconvenience, rider earnings, support ticket rate, and more.

Figure 12 shows the Express POOL cancellation rate from the launch of the product in early November 2017 through mid-March 2018. The plot shows that in the period immediately after the launch both riders and drivers had relatively high cancellation rates, but the rates came down significantly as they adjusted to the new experience. A high cancellation rate is expected for new product launches, but is typically following by a decrease as people come up a learning curve of varying steepness. Even though the changes made to the Express POOL product request flow were substantial the team began to see a decline in cancellation rate two-weeks after launching.

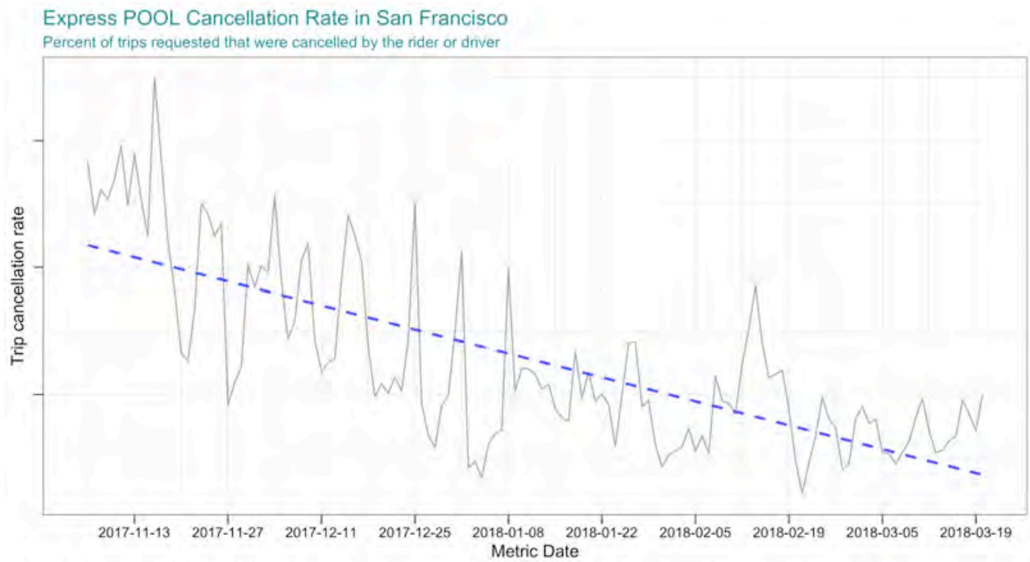


Figure 12. Express POOL Cancellation Rate

The improved rider experience, supported by a falling cancellation rate, may be due to a variety of factors. High cancellation rates may be a symptom of curious riders exploring the new product in their app, and then make a trip request purely to investigate the new experience. These cancelled trip requests naturally decay over time as the novelty factor on the product begins to wear off post-launch. Riders may also gain a better understanding of the mechanisms of Express POOL and decide not to cancel due to uncertainty when waiting for a driver match, apprehension about walking to the pickup-location, safety concerns, or another issue. In sum there are many factors that might drive the falling cancellation rate, but overall it is indicative of good product-market fit and having improved the existing uber POOL experience.

The new product was also able to drive value to Uber's bottomline. In the first month after launch riders that requested an Express POOL only waited for 40-seconds longer on average than riders taking the original POOL option. This 40-second delay was intentional, as Uber made riders wait on the trip request screen to get batched with other riders on their trip. In return for this delay, the match rate for the Express POOL product was 3.6% higher relative to the existing POOL product. Match rate is defined as the total trip requests that were matched divided by the total number of outstanding trip requests; it is a leading indicator of business performance and product experience. The improved experience associated with an increased match rate translated to riders taking 4.6% incremental shared rides trips 1-month after launch.

The success of the November 2017 launch of Express POOL in San Francisco and Boston lead to the product being rolled out in many other domestic and international markets. In late February 2018 Uber launched Express POOL in Los Angeles, San Diego, Denver, Philadelphia, Washington DC, and Miami. These markets experience similar positive effects from Express POOL that were observed in the original launch markets, and by expanding in these cities a few months after the original launch the team was able to synthesize learnings from San Francisco and Boston to better execute on the rollout strategy in these second wave locales. The third wave of domestic cities to get the new product was Chicago, Seattle, Atlanta, Las Vegas, and the New Jersey area in mid-May 2018. Finally, Express POOL went international with the launch of Paris, Sydney and Melbourne in August 2018. Attractive expansion markets exists across the globe and Uber hopes to bring the Express POOL product to all existing shared rides markets.

CONCLUSION

In conclusion, the research efforts to create the perfect POOL required close collaboration between the Uber user research and data science teams to understand rider preferences and the trade-offs they make when evaluating their transportation options. The ability to integrate both research methods enabled the team to provide compelling data to business leaders that was ultimately the single biggest input in developing the next iteration of uberPOOL. The resulting success of Express POOL provides a good example of how cross-pollination between disparate research methods can lead to positive business outcomes.

Jenny Lo is a User Research manager at Uber. Jenny specializes in the study of quantitative research methods and information technology in developing countries (ICTD). She received her Masters of Information Management and Systems from the School of Information at University of California, Berkeley and Bachelors from Wesleyan University. Email: jlo@uber.com

Steve Morseman is a Data Scientist at Uber. His research focuses on rider acquisition, engagement, and product development. Steve received his Masters in Political Science from the University of California, Los Angeles and his Bachelors from the State University of New York at New Paltz. Email: morseman@uber.com

NOTES

We would like to thank Lisa Renery Handalian for going above and beyond for helping us work through multiple drafts of this paper. Chad Maxwell also provided invaluable feedback to improve the content for the EPIC audience.

1. uberX is a solo-trip in a typical sedan, with no carpooling with other riders; whereas uberBLACK is also solo-trip, albeit taken in a premium vehicle with a professional driver.

2. The time since a rider created their account.

3. Uber researchers created the question set utilizing the OptBlock function in the AlgDesign package in R language for statistical programming.

4. Discrete choice models consist of the outcome variable (Y_i) taking on a binary value of 0 or 1. Values are 0 in the absence of an event or 1 when an event occurs.

5. This required the dependent variable to take on a value of 0 or 1 to indicate the occurrence or absence of an answer choice selected in the survey. The data transformation requires turning a -1, 0, 1 dependent variable to 0 or 1, so if a respondent selects any answer choice the response variable will be coded as 1, or a 0 if the respondent didn't select the answer at all.

6. Normalizing the coefficient estimates involves mean-centering the coefficients, exponentiating the mean-centered values, and renormalizing the exponentiated values by their sum.

7. A classic example of a product attributes is a computer. Features include hard drive size, processor speed, memory, screen size, etc. Each of these features can take on a variety of levels, such as having the choice of a 32gb, 64gb, or 128gb of memory when buying a new computer.

8. Purchasing refers to the process of opting to buy, subscribe, or otherwise spend money on a product or service. For the specific example of Uber, choosing to purchase the product means opting to take a given trip on the Uber platform.

9. The model used in this study is a Hierarchical Bayesian Multinomial-Logistic Regression Model.

10. 4 ETA levels * 3 Walking levels * 5 trip length levels * 3 trip variance levels * 5 discount levels = 900 possible profiles.

11. Sampled without replacement, meaning that each profile is only eligible to get selected once when randomizing each respondent's question set.

12. The product profile randomization, question assignment, estimated-time-to-destination, and pricing calculations were completed utilizing a python script called from a php endpoint hit by the Qualtrics web service functionality. The server then returned a json-encoded string to Qualtrics with both the product package identifiers and text to render to the respondent.

13. For the maxdiff the data was binomial, where the dependent variable takes on a value of 0 or 1, but the multinomial distribution is a generalization of this data where the response variable can belong to one of the many different potential groups. The data generated in the conjoint is multinomial because they could have chosen product A, product B, or having opted out (product C).

14. Bayesian refers to the Markov Chain Monte Carlo process used to arrive at the estimated utility estimates, which uses a simulation approach that updates the information for each model run based upon the existing prior results from the previous iteration.

15. The rhierMnlRwMixture function was used to estimate the model. This function uses a hybrid sampler for hierarchical multinomial-logit with a mixture of normal priors.

16. For each iteration of the Markov-Chain, coefficients are estimated for each conjoint product feature-level for each respondent. In addition to these utility estimates, the posterior distribution of the Markov-Chains allow for the study of any control variables included in the respondent matrix.

17. Estimates were calculated over the second ⅔ of their Markov-Chain simulations, allowing for a burn-in period prior to estimates stabilizing, then averaging over the respondent level estimates to arrive at the global averages for each Uber product feature-level.

REFERENCES CITED

Cavusgil, Erin
2005 Seeing What's next: Using the Theories of Innovation to Predict Industry change. Harvard Business School Press, Boston, MA, 2004. 312 Pages. *European Journal of Marketing* 39 (11/12): 1389–1390

Christensen, Clayton M., Scott Cook, and Taddy Hall
2016 "Marketing Malpractice." In *The Clayton M. Christensen Reader*, 45-64. Harvard Business Review Press, Boston, MA, 2016. 224 Pages.

Collier, David, and James Mahoney
1996 Insights and Pitfalls: Selection Bias in Qualitative Research. World Politics 49(01): 56–91. http://www.jstor.org/stable/25053989

Epstein, Liana, Colette Des Georges, and Eric Van Susteren
2018 Comparing Isn't Always a Piece of Cake: Ranking vs. Rating Explained. https://www.surveymonkey.com/curiosity/rating-vs-ranking/, accessed September 26, 2018

Graham, Mary E., and Daniel M. Cable
2001 Consideration of the Incomplete Block Design for Policy-Capturing Research. *Organizational Research Methods* 4(1): 26–45

Hawkins, Andrew J.
2018 Uber Express Pool Offers the Cheapest Fares Yet in Exchange for a Little Walking. The Verge. http://www.theverge.com/2018/2/21/17020484/uber-express-pool-launch-cities, accessed September 26, 2018

Hess, Stephane, and Andrew Daly
2014 Handbook of Choice Modelling. Cheltenham: Edward Elgar

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant
2013 Applied Logistic Regression. Hoboken, NJ: Wiley

Lovelace, Matthew, and Peggy Brickman
2013 Best Practices for Measuring Students' Attitudes toward Learning Science. CBE—Life Sciences Education 12(4): 606–617

Marley, Sarah
2018 Maximum Difference Scaling (MaxDiff). https://select-statistics.co.uk/blog/maximum-difference-scaling-maxdiff/, accessed September 26, 2018

Marley, Tony, and Pihlens, David
2012 Models of Best–Worst Choice and Ranking among Multiattribute Options (Profiles). *Journal of Mathematical Psychology* 56(1): 24–34

Porter, Stephen R., Michael E. Whitcomb, and William H. Weitzer
2004 Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research* 2004(121): 63–73

Qualtrics Support
2018 Question Randomization. https://www.qualtrics.com/support/survey-platform/survey-module/block-options/question-randoization/, accessed September 26, 2018

Qualtrics Support
2018 Quotas - Qualtrics. https://www.qualtrics.com/support/survey-platform/survey-module/survey-tools/quotas/, accessed September 26, 2018

Rossi, Peter
2017 CRAN - Package Bayesm. The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/bayesm/index.html, accessed September 26, 2018

Rossi, Peter E., Greg M. Allenby, and Robert McCulloch
2009 Bayesian Statistics and Marketing. Chichester: Wiley

Sawtooth Software
2013 The MaxDiff System Technical Paper. http://www.sawtoothsoftware.com/download/techpap/maxdifftech.pdf, accessed September 25, 2018

Uber
2018 Uber Privacy. http://privacy.uber.com/policy, accessed September 26, 2018

Ulwick, Anthony
2016 Jobs to Be Done: Theory to Practice. s.l.: Idea Bite Press

Wheeler, Bob
2014 CRAN - Package AlgDesign. Algorithmic Experimental Design. https://cran.r-project.org/web/packages/AlgDesign/index.html, accessed September 25, 2018

Yap, Justin
2017 FlipMaxDiff - GitHub. https://github.com/Displayr/flipMaxDiff, accessed September 26, 2018

Contextual Analytics: Towards a Practical Integration of Human and Data Science Approaches in the Development of Algorithms

MILLIE P. ARORA
ReD Associates

MIKKEL KRENCHEL
ReD Associates

JACOB McAULIFFE
ReD Associates

POORNIMA RAMASWAMY
Cognizant

As algorithms play an increasingly important role in the lives of people and corporations, finding more effective, ethical, and empathetic ways of developing them has become an industry imperative. Ethnography, and the contextual understanding derived from it, has the potential to fundamentally change the way that data science is done. Reciprocally, engaging with data science can help ethnographers focus their efforts, build stronger and more precise insights, and ultimately have greater impact once their work is incorporated into the algorithms that increasingly power our society. In practice, building contextually-informed algorithms requires collaboration between human science and data science teams who are willing to extend their frame of reference beyond their core skill areas. This paper aims to first address the features of ethnography and data science that make collaboration between the two more valuable than the sum of their respective parts; second, to present a methodology that makes collaboration between the two possible in practical terms; and third, to generate critical discussion through an examination of the authors' experiences leading and working within joint teams of ethnographers and data scientists.

INTRODUCTION

The term "Big Data" has gone out of fashion. Some posit that it collapsed under the weight of lofty expectations – access to massive datasets is not, as business leaders found out, a digital panacea.[1] Others contend it was simply refashioned into less fetishizing terms like "data-driven"[2] or replaced by references to its real-world applications like machine learning.[3] Most agree, however, that big data – though no longer deserving of capitalization[4] – lives on in the ever-increasing relevance of its mission: some 80% of any company's data remains unstructured and unused, and businesses strive to make that information accessible to those with the necessary skills and tools to harness its value.[5] Increasingly, this means looking beyond the traditional methods of data analysis.

Social scientists, too, have sought to shed new light on deep and dark data lakes by combining big data analytics with qualitative observation in the form of ethnographically-inflected "thick data."[6] In the proceedings of EPIC alone, discussion around collaboration between the social and data sciences has ranged from new opportunities for mixed methods research,[7] to exploring the lifecycle of a data source,[8] to understanding decision-making within data analytics processes.[9] Moreover, as Kate Crawford describes in her 2017 Keynote to the Conference on Neural Information Processing Systems, data scientists themselves are now seeking input from people with a deep expertise in human issues.[10] For example, in response to recent high-profile privacy scandals, companies like Facebook are advertising the need to look past trace data and towards "the contextual knowledge that computers lack."[11] Data and human scientists are well agreed on the opportunities of combining their respective methodologies. The question, then, is no longer whether we need to combine big and thick data – the question is how best to do it.[12]

The danger of ignoring this practical question – and simply presuming a commonsense relationship between data and human science – lies in the failure to consciously identify how these two methodologies might actually influence one another. As Nick Seaver writes, the big and thick data conversation fits within a long lineage of "neatly opposed methodological moieties."[13] In relation to methods of standardizing behavioral analysis – of which big data is far from the first – human scientists continue to rehearse established scripts about how "renewed attention to the blood and sex of daily life" might rescue or regulate formal analysis.[14] Seaver argues for pushing beyond these familiar scripts to attend more carefully to how these specific methodologies might actually interact. With Seaver, in this paper we will explore how an understanding of social and data science processes might more meaningfully inform collaboration between the two. Consequently, we will attempt to give more "texture and specificity" to the practices of big data and interrogate the "coherence and self-evidence" of ethnography in the process.[15]

Over the last two years Cognizant – a leading technology services company with deep capabilities within Data Analytics and AI – and ReD Associates – a consulting firm that helped pioneer the use of applied ethnography – have developed a suite of offerings that leverage the best of both data and human science in the service of real client problems. We've had hundreds of conversations with Fortune 500 executives around potential services, but it has become clear that the one area where the integration of human and data science is most demanding – and potentially most impactful – is in the creation and refinement of algorithms. These algorithms increasingly make up the backbone of many businesses by automating their interactions with customers, employees, and stakeholders. Concerns about ethics and efficiency, however, have led these same businesses to seek new insights around their algorithms. In response, ReD and Cognizant developed a methodology called Contextual Analytics: a project process for uniting data analysts and social scientists under the mandate of building more effective and credible algorithms.

Simply putting data analysts and social scientists in a room together is not enough to ensure a better algorithm. Rather, our experiences point to the need to design projects in a fundamentally different way in order to overcome the methodological and philosophical challenges of integration. This paper will begin with a short overview of existing efforts to integrate big and thick data before turning to a description of the three phases of the Contextual Analytics methodology. After this description, we will use examples from four recent projects to describe what the three phases look like in practice. Drawing from insights gained while working with clients, the paper will ultimately suggest new ways of achieving meaningful cooperation between the human and data sciences.

INTEGRATING DATA AND HUMAN SCIENCES

A Brief History of Big and Thick Data

Most attempts to define big data focus on the "3 Vs" – volume, velocity, and variety – to demonstrate the sheer size and scope of data sources available today.[16] Other definitions include "veracity" and "value" or look to different letters altogether,[17] but the general concept remains the same: big data is the phenomenon of having massive datasets on human behavior drawn from millions of touchpoints between businesses and organizations and their customers and clients, tracked in a variety of formats. Data analysts use statistical techniques to analyze datasets, and in doing so create new ways of interpreting and classifying relationships between human and non-human actors.[18]

When the concept of big data was still at its nadir, social scientists recognized its potential complementarity with their own discipline. Their initial attempts to promote the use of "thick data" – a term borrowed from Clifford Geertz and used to describe deep insights into human behavior derived through ethnographic method – failed to garner much interest among data scientists.[19] In recent years, however, thick data has become an industry imperative due to shifting priorities around data privacy and ethics. Since the advent of machine learning – a fully-automated tool designed to find relationships in massive pools of data – dozens of studies have called into question the opaque decision-making of their resulting "black box" algorithms.[20] Moreover, business leaders increasingly struggle to accept the results of automated algorithms that lack a persuasive account for their findings.[21] Now that economic incentives around performance are aligned with ethical imperatives to make more transparent algorithms, interventions on behalf of social scientists for "thickening" and "socializing" the process of data analysis are increasingly cited by politicians, pundits, and industry leaders.[22]

The Evolution of Integrated Approaches

Mixed-methods approaches often structure their research in the ways most familiar to them – social science first or data science first – and use the other method as a supplement or corrective.[23] In presenting learnings on Contextual Analytics, however, we will join others who attempt a more even-keeled integration of the two. For example, while we contend that social scientists and data analysts must also participate in and internalize each other's research, we join the authors of the "blended" model by extending their recommendation that the merging of insights be rapid, iterative, and done before analysis is fully finished so that one method may help guide the other.[24]

Moreover, "Living Labs" models – which use coordinated community workshops to test technologies and services in a real-world context – tell us that interaction between teams can start well before the workshop with the analysis of available trace data, subsequently enrichened through participant interaction.[25] Like Living Labs researchers, our studies suggest that, when possible, it is beneficial to start a project with "exploratory" data analysis that finds new patterns to inform subsequent ethnographic research protocols.[26] We also draw on "Collaboratory" models: one-time workshop events where ethnographers and industry data analysts come together to share their knowledge and experiences.[27] While Collaboratories focus on concept development rather than implementation, these workshops demonstrate the benefits of using ethnographic research to slow down traditional data analysis in order to explore more open-ended ways of solving a problem.[28]

In outlining an integrated approach to algorithm development, we hope to build on existing work as well as present new learnings for what teams must do differently in a corporate setting. Drawing on this dual focus, we aligned social and data science workstreams into a process called Contextual Analytics, to which we will now turn.

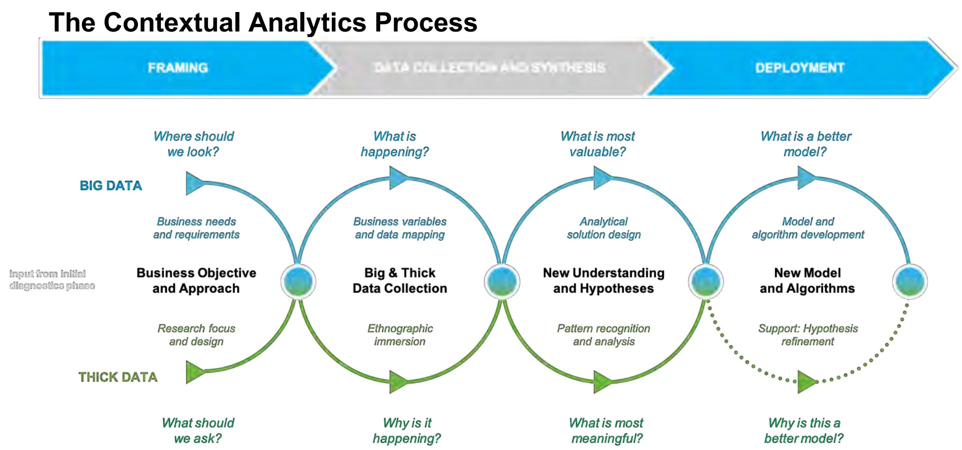


Figure 1. The 12AI Contextual Analytics Process. ReD Associates and Cognizant, 2018. Used with permission.

Contextual Analytics was designed by ReD and Cognizant to allow data analysts, social science researchers, and client data and business teams to work together to address well-established analytics gaps. The process leverages quantitative and qualitative analytical techniques across three project phases – Framing, Data Collection and Synthesis, and Deployment – to build actionable insights for a company's current situation and produce improved algorithms for generating new insights going forward. Over the past year, we have applied this methodology to high-stakes corporate problems in various industries. Across four projects in particular, we learned several lessons pertaining to the central guiding principle of Contextual Analytics: how can we ensure truly integrated, balanced, and collaborative human science and data science workstreams? Overall, our work so far suggests that there are three critical elements for integrated teams to consider:

1. Specificity in Framing – While strong framing is critical for any research project, there is a greater need for epistemological and logistical specificity in the planning stages when combining data science and social science teams. In order to acquire the right data sets in a timely manner and deploy teams to the right locations, ethnographers need to get specific about what they really need to know and how data analysts can help. Data analysts, in turn, need to be clear about what it will take to acquire and operationalize the necessary data sources. Doing so in a way that avoids hypothesis-driven thinking and allows for agile redirections of the project frame requires practitioners on both sides who are experienced enough to foresee potential problems in team communication and data access.

***2. Forced Immersion in Data Collection and Synthesis* –** The complementarity of social science and data science is most beneficial when a single team conducts research and analysis while in constant conversation, collaborating to share their findings and align on next steps. To do so, practitioners from each of the workstreams must understand how the other practically conducts their projects and what this means for their own work. Thus, to collect and present their data in a way that is mutually beneficial, data scientists and social scientists must briefly, but meaningfully, step into the methodological shoes of the other.

***3. Integrated Storytelling in Deployment* –** The customs around good storytelling vary widely between (and among) ethnographers and data scientists. To get the most out of an integrated approach a common language is needed that borrows the best from both disciplines. Thus, integrated teams must not only produce two separate deliverables – a highly persuasive story behind insights and a model with statistically improved results – but also create meaningful links between those two deliverables. To do so requires practitioners to move out of their respective black boxes and towards a common set of mid-level theories: data analysts by opening their algorithms to tell the stories that the data represents, and ethnographers by pushing beyond grand narratives alone to connect them to testable and sizeable hypotheses.

When these elements are taken into account throughout project design and practice, the resulting collaborative efforts are not necessarily faster or cheaper, but allow for teams to jointly create algorithms that are more grounded, that have better results, and that people in the C-suite can understand.

Project Context for Contextual Analytics

Our paper will now discuss Contextual Analytics in practice, with learnings drawn from four case studies: modeling fraud, increasing the hit rate for prospect targeting, retaining skilled employees, and improving call center operations.

Case 1: Modeling Fraud – The business impact of credit card fraud is high. Banks and retailers lost $16 billion to credit card fraud in 2016, and scammers claimed two million more victims than in 2015.[29] Worse, fraudsters move almost ghost-like through the dark web; financial services providers and credit-card facilitators see only the traces that perpetrators leave in data. They know little about what motivates them, or what might scare them away. We wondered what financial institutions might learn if data scientists could gain a firsthand understanding of what the world of credit card fraud looks like. Could they build better fraud detection algorithms?

The question triggered months of study. We met, conversed with, and observed actual fraudsters in New York and Boston and garnered surprising results. Credit card thieves see their line of work as hustling to survive, not a get-rich-quick play. Longevity requires that they move fast and spend small sums of money to stay below the radar. It's labor-intensive work. That reality opens a different window on how to deter fraudsters: make their work a bit more difficult, and suddenly fraud is a less attractive option.

By taking into account human insights about how fraudsters conduct their work, banks can sharpen their fraud-detection algorithms. For example, knowing that fraudsters send their online orders to vacant houses, banks have the ability to scrape real-estate databases "for-sale" listings and cross-tab them against credit-card transactions to flag potentially suspicious purchases.

Case 2: Increasing the Hit Rate for Prospect Targeting – A global financial services company prides itself on developing advanced algorithms to determine which small-to-medium businesses (SMBs) are best primed for new credit offers via phone or mail. Yet in their own words, one of their blind spots was an understanding of *why* those SMBs would be receptive to offers. As a result, their current models resulted in an astoundingly low response rate – a mere 3.5% – and in the process left countless business owners feeling spammed by irrelevant offers. Our challenge was to develop ethnographic insight into why SMBs respond or fail to respond to offers at particular times, and to develop and translate those insights into data implications and proxies for building a new, more precise, targeting model.

Through multi-day immersions with over twenty SMBs in the New York Metro area, we uncovered a range of financial challenges that owners face when running their businesses. For the most part, they aren't business experts, Indeed, owners are somewhat overwhelmed trying to figure out how to run a successful business and often seek personal help in doing so.

We were able to translate our insights into a new model of expanded responsiveness that detailed six components of an "Owner's Ethos" that makes them more or less likely to respond to an offer. For example, SMB owners with a local mindset are less likely to respond to offers since they are keen to support the community and therefore seek out community banks or local providers instead. From insights such as these, we developed over 170 data proxies, and developed a new algorithm of responsiveness that is set to launch at the start of 2019.

Case 3: Retaining Skilled Employees – The problem of employee churn is a ripe subject for big data analysis. We met with a technology solutions company that was using rich data – from exit interviews to Glassdoor reviews – to identify the segments of their workforce who had a higher historic likelihood of quitting. The company HR department had a few hunches about why particular segments were most at risk, but lacked the human data that would explain why the model marked certain employee groups and departments as "attrition hotspots."

In order to gather insights on drivers of attrition (and retention) and build a predictive model from them, our approach entailed four collaborative workstreams: the continued quantitative investigation of attrition hotspots and related data proxies; deep dives with employees and those around them, including managers, colleagues, and family / friends; discourse analysis of data gathered with and from human resource teams and in employee town halls; and stakeholder interviews that clarified the cost of attrition and the benefits to solving the issue.

By layering human insight on top of quantitative work, the team developed a values framework that articulated the key rationales for pursuing a career with the company, and how certain values are under threat from a variety of internal and external factors. Each value was then correlated to range of data proxies and weighted for different employee segments – the balance seeker, the risk taker, etc. — to incorporate into the predictive model.

Case 4: Improving Call Center Operations – A leading Nordic energy provider runs a contact center that typically fields calls from two types of customers: those looking to invest in new energy solutions, and those struggling to pay their bills. The provider worried that they weren't catching a significant amount of the investor population. Meanwhile, employee hours were clogged with calls to from concerned customer discussing their hefty bills. To solve both problems, the provider desired a new predictive algorithm that could streamline operations by suggesting when employees could pre-emptively contact these different customer segmentations.

To identify the markers of each customer segmentation, the ReD and Cognizant team asked what moments matter most to homeowners when it comes to their energy use. Through interviews with homeowners and by shadowing customer calls, ethnographers that people with a "warm" relationship to their energy searched for new energy uses and investments, and those with a "cold" relationship defaulted to the cheapest option and struggled to pay their bills.

These new insights led the integrated team to create an event-based algorithm that indexed customers based on the energy health of their home. They designed proxies to predict whether homeowners are "struggling" or "investing," and thus could benefit from a call, or if they were "sleeping" and best left alone.

Using examples from these four case studies, our paper will now discuss Contextual Analytics in practice, outlining how the methodology can be applied in each of the three phases.

PHASE ONE: FRAMING

In the first phase of a Contextual Analytics project – Framing – teams are designed to balance quantitative and qualitative methods to better direct the course of data analysis. Ideally, a project is framed so that quantitative data informs the start of an ethnographic investigation by focusing its traditionally broad scope through the statistical identification of promising markets and segments. Ethnographic immersion, on the other hand, generates rich contextual information through open-ended inquiry, and so broadens the data analytics framing process which traditionally relies on hypotheses generated from easily available data sets.

We've learned that – despite the best intentions of both teams to work collaboratively in framing – their efforts can be impeded if they don't specifically account for process silos and problems with data access. To account for the former, a strong project management team needs to ensure frequent communication between both social and data scientists so the two don't return to old ways of conducting their research independently. For the latter, teams must consider the practical constraints of data access; a particularly relevant challenge given new data protection and privacy laws like those implemented across the European Union in May of 2018. The result: well-framed projects can enrich research and speed up raw data collection, but teams must spend more time planning them upfront.

Contextual Analytics in Practice: The Project Framing Phase

In the framing phase, project goals are two-fold. First, develop research questions to guide the inquiry process: what to ask and why? Second, develop research protocols: who to meet and where? Typically, social scientists and data analysts approach these same goals from different epistemological starting points. Social scientists, with theoretical foundations in ethnographic and grounded theory approaches, frame research in a purposefully broad manner in order to approach a phenomenon without potentially misguided assumptions.[30] In this way, framing is abstract and leaves room for serendipity: go after the biggest questions, and narrow the scope based on what shows up as important along the way. Data analysts, however, follows a scientific method approach by starting with a set of hypotheses that could explain the business problem. The input for these hypotheses often comes from business leaders who provide their perspective on the issue or from prior models which explain the situation for different use cases. Thus, data analysis typically begins with a deep understanding of existing data sources and how they've been used in the past, along with the preparation of any data that hasn't been used or analyzed previously.

Despite social scientists and data scientists having different starting points when framing projects, a key value added by the Contextual Analytics process is that the integrated team continuously shares, discusses, and debates any prior analyses, hypotheses, and unknowns they may have before and during the course of the project. By integrating their analytical techniques in response to the specific demands of a given case, an integrated team has the potential to dramatically strengthen their starting point for analysis and algorithm development.

Take, for example, our project on prospect targeting for a large financial services company. Before ReD and Cognizant intervened, the financial company's data team had embarked on a project to develop new models that would increase the response rate of SMBs to credit card offers received over the phone or via mail. Working in a silo – removed from any insights that a sales team may have had from actually interacting with SMBs – the client data team had relied on a combination of intuition and evidence to frame their analysis. Their intuition led the team to start with analysis of historical data that showed predictability in offer response-rate for other populations. Then, after searching for those same correlations in their lists of SMBs, subsequent data analysis showed a small uptick in the model which gave them the evidence to justify their data use. One data analyst on the project told the ReD and Cognizant team that this framing constituted a "big creative leap" in which their hypothesis was based not on "what would cause credit card conversion now?", but rather on "there was conversion for this type of problem in the past"; all correlation, with no proof of causation.

Enter Contextual Analytics. After identifying this limitation of the existing model, the ReD and Cognizant team planned their project to start with social science research in order to move past a narrow framing based on existing knowledge – what does the data tell us about the phenomenon? – and towards framing based on open-ended analysis – what does the phenomenon tell us to look for in the data? Instead of going after existing correlations of data, our ethnographic team planned a backwards-engineered segmentation to find the underlying drivers for how financial decisions are made in SMBs; they gathered and studied businesses in different industries, with different revenue and employee size, and then searched for common factors that most affect their probability of responsiveness. This blank-slate segmentation approach allowed the whole team to think creatively around different types of data to start with, including examples of proxies like local mentality: if the business has closer ties to other institutions in their neighborhood, they were less likely to respond to an unsolicited offer. Through their combined efforts, the integrated team was able to envision a sound frame for the project that used social science to make up for a lack of guiding insights.

Where, ideally, a data analysis stream would first develop certain research protocols by sharing hypotheses from past models, in this case those hypotheses couldn't be trusted without their validation in the population that was actually being studied. That being said, machine learning is powerful, and can find corollaries in troves of data that would confound human analysis. Ethnographers learned to pay special attention to areas where data analysis said the business problem was most pressing, and where parts of research needed to happen. If framing is to succeed, conversations like the above have to happen in the context of each project, based on an honest assessment of limitations and strengths.

Key Success Criteria for the Framing Phase

A proper framing process for Contextual Analytics sets up data and human science for success by customizing interaction between the two workstreams. Yet even the best laid plans of human and data scientists are subject to practical tests within a corporate setting. We will now present two learnings to take into account when framing integrated projects.

***Align project management teams to avoid process silos* –** Without establishing a clear process to guide the mandate of working collaboratively, we've seen that teams have a tendency to retract into silos: a data science team working in isolation with a hypothesis dictated by opportunistic data, or a social science team identifying a phenomenon and lines of inquiry unguided by existing analyses. These two silos acutely demonstrate the need to understand and account for one another's epistemologies when framing a project.

In one project, we saw the firsthand the effects of limiting project framing to a single method. Before starting a project on predicting credit card fraud, our team met with employees of a large bank in the United States that was using their data – from financial transactions and payment records to credit history and call-center interactions – to get a good predictive picture of their clients. Building on their success, the bank sought to use these same datasets to locate fraudsters, assuming that fraud was detectable as deviation from well-established spending patterns. Yet the bank found that, even with the most advanced predictive algorithms, their efforts at combating fraud remained as dismally ineffective as the rest of the country.[31] By basing their new hypothesis on recycled data, their analysts simply weren't looking in the right place.

The above example demonstrates that, even for enterprises with virtually unfettered access to some of the richest troves of user data available, data-driven analysis alone is, ironically, insufficiently capacious to sort through its own information. With loads of unstructured and unused data, those teams are pressured to recover the sunk costs of data collection by quickly finding value in its use.[32] Thus, when deciding what questions to ask, overloaded data science teams will typically turn to the low-hanging fruit of "opportunistic" data; sources that are easily available for speedy use. Yet the output of data analysis is only as good as its input; the algorithm was weakened because it couldn't infer how fraudsters operate without data from the actual fraudsters.

Without a conscious framing process, opportunistic data-use is often the default for data-analysis. This use of old and existing data can be useful when it comes to framing for straightforward or predictable problems – such as inferring a customer's preferred method of communication with the bank – but it is much less helpful in developing a research question when the scope of that question itself is at issue.[33] Nevertheless, as we saw in the fraud project, data analysts are often pushed towards recycling old and opportunistic data sources when faced with challenges. Indeed, in our experience, this tendency to revert to methodological silos threatens both workstreams; it is often easier to take refuge in comfortable framing methods rather than at ways to adapt and change one's methods. Thus, having a strong project management team across workstreams is critical to ensure that both attempt to actually communicate and build upon one another.

A successful example of an aligned framing effort – where managers take into account how one process can best help the other – came in our project on retaining skilled employees for a technology solutions company. Ethnographers faced a daunting framing task: explore staff loyalty in the broadest sense within a massive organization, and come up with a perspective on what would make their employees more or less likely to leave. With traditional ethnographic framing, researchers would spend several weeks conducting stakeholder interviews and client research in order to find the right place to invest their time and effort. Thanks to the foresight of management and client teams, however, the company agreed to share their internal data beforehand to facilitate the statistical identification of research areas. The data team used this data to pinpoint specific areas where the company found it difficult to keep employees – what they called "attrition hotspots" – and sent ethnographers to figure out why that was the case. In addition, correlations around attrition from the data analysis such as "frequency of communication with supervisors" and "time on bench" helped frame particular aspects of the ethnographic field guide. Quantitative findings thus allowed qualitative researchers to have a less time-intensive process of recruitment and research.

Start early by thinking concretely about data access – Attempts at collaboration often run up against the logistical challenges of data access, timing, and quality. While ethnographers expressed surprise at lengthy data delays, it is well-known among data professionals that they are likely to spend the majority of their time (nearly 70% of it) just getting the data and putting it into a clean, usable format. Moreover, following a spate of governmental programs on the fair and anonymized use of aggregated data, corporations are cautious when it comes to sharing their internal and private data.[34] The above considerations suggest that integrated teams must plan carefully in order to account for issues around data access.

All four of the projects cited in this paper encountered some internal issues getting to the data, but data access was perhaps most at-issue in the credit card prospect targeting case. Concerns about sharing customer-level data with an external group led to the client's requirement of a lengthy and unreasonable onboarding process for the ReD and Cognizant team, including drug tests and a four-month application processing time. This delay meant that the team could not access the financial organization's massive amounts of internal information, such as neighborhood-level credit card spending, prior prospect responsiveness, and other rich customer-touchpoints. Lack of data had effects on both workstreams. The data analysis team felt limited – or, as one described, "impotent" – since they were unable to contribute in validating and pointing towards new findings. From the human science side, a lack of "quantitative backup" meant social scientists felt they had to be more risk-averse when finding correlations, and were less confident judging feasibility of insights based on an understanding of readily available datasets.

In theory, an integrated approach is one where workstreams are set up to begin their data collection and analysis on an equal footing. But, in reality, that is not always possible, and it may not always be the most adaptive or prudent approach. The data is not always there to start, so quantitative and qualitative tracks will often run at different times. It is important, however, to orchestrate data collection and synthesis so that both tracks can still guide and direct each other as much as possible. In the prospect targeting case, for instance, data analysts made up for lack of internal data by looking to publicly available data they believed the corporation *may* have, and creatively considering what it told them about new and existing proxies. Social science teams, if delayed, can take advantage of data analysis identification of relevant markets and segments to start putting down boundaries around future research.

PHASE TWO: DATA COLLECTION AND SYNTHESIS

The integration of data and human science workstreams is most critical in the second phase of Contextual Analytics, Data Collection and Synthesis. Data collection takes place in two parallel workstreams: ethnographers observing a small sample size of respondents in the field and data analysts tracking correlations in large datasets. The former is meant to add to and explain quantitative findings, while the latter lends statistical significance to qualitative observations. We've learned, however, that synthesis cannot take place during a single event following research. Rather, data collection and synthesis occur simultaneously and iteratively so that each of the workstreams can use their findings to guide the other to new areas of inquiry. This reciprocity is especially important as data analysts guide researchers on which types of data are actually accessible.

When data collection is done independently – each team taking responsibility for only their own realm of expertise – the representation of social-science based insights in the algorithm is put at risk. Thus, collaborative projects require an element of "forced immersion" on behalf of both data analysts and social scientists, where each practitioner purposefully takes an extra step beyond their normal research. Social scientists need not become data experts (or vice versa), but we've learned that taking the time and effort to learn how the other side conducts their analysis prepares each team to pre-emptively set up their own work in ways that make it easier to merge, translate, and present findings.

Contextual Analytics in Practice: The Data Collection and Synthesis Phase

Contextual Analytics crucially relies on the use of social science research to ground possible correlations from data analysis and to identify hypotheses found outside the margins of available datasets. Take, for example, our project on improving call center performance. Due to the team's lack of data access, data collection began with open-ended ethnographic inquiry: how do homeowners relate to their energy? By spending time with homeowners across a wide demographic range, researchers found that the energy provider could benefit from a more general understanding of the events in the life of a homeowner. These events most often correlated with energy consumption and smart-meter data, leading the team to suggest proxies for which the home – rather than the homeowner – was the locus of analysis.

By providing proxies early and often in data collection, ethnography can help jumpstart the quantitative process of data analysis. In Contextual Analytics projects, ethnographic teams collect insights in ways that roughly map onto potential data sources. Thus, after only a few weeks of observing customer calls with the energy provider and setting up post-call interviews, ethnographers had filled their field notes with new and promising correlations. The team focused on event-based proxies – like the purchase of a new home or flat which primed "warm" homeowners to invest in energy – because they believed such insights could be tracked in easily-available data. Yet while opportunistic data sources may risk showing correlation without causation, rudimentary ethnographic findings like the above run the opposite risk. After hunting for and hinting at various correlations from the field, then, qualitative findings need "quantitative backup" to prove their long-term value.

The term "quantitative backup," coined by one of ReD's researchers on the credit card response project, refers to the process of having a data analyst check possible proxies against available data in order to demonstrate sufficient frequency and help prove that a particular correlation in the thick data isn't anomalous. After identifying hypotheses from the field and potential proxies for those hypotheses, the process of checking correlations requires a data analyst to assess the feasibility of translating that insight quantitatively using their knowledge of existing data sources.

By checking in with the data analysts embedded within their team, researchers studying the call center found that many of the events they identified – the installation of a solar panel, the exchange of heating information, a recent move, financial distress – had corresponding variables within existing data sources. Even without access to internal data at the time, the data analyst had the foresight and expertise to inform the team that much of the data they anticipated using existed in a large number of systems, or in datasets that they couldn't trust due to missing or incomplete fields. Thus, validation proved helpful in pointing the team towards finding new proxies to replace their rudimentary attempts.

Key Success Criteria for the Data Collection and Synthesis Phase

Quantitative and qualitative methodologies offer ways to both explore new and validate existing correlations during data collection and synthesis, but only if the two are working in concert. We will now demonstrate the necessity of data and social scientists understanding each other's processes, and the element of forced immersion that ensures proper presentation and translation of insights between workstreams.

Ethnographers need to take a stance on the data – As one of our data analysts cautioned the prospect targeting team, ethnographers need to be careful when choosing which findings to give to the data team for testing, and what format they choose to deliver them in:

> "We have [a data team] for 4-8 weeks, and that is the time we have. Once we have these resources, we have to be specific about what we can deliver and in what time. It will depend on how difficult it is to scrape and how difficult it is to get creative with the data. It's not a

> big deal to just go out and buy the data, that's easy. It's the creativity and thinking that takes time."

Ethnographers tend to underestimate the amount of time and effort it takes for data analysts to translate qualitative insights into variables that can be meaningfully included in the model. The more creative the finding, the harder it will be to unpack it. Given their limited time frame, ethnographers cannot rely on the data analysts to undertake this lengthy translation process by themselves. This problem calls for ethnographers to take a stance on their data by delivering not just insights, but rudimentary proxies that take accessible data sources into account.

Ethnographers can develop better proxies during data collection if they take a step beyond their own human-centric stories and gain a better understanding of the data that analysts use to capture them. On the credit card project, for example, the social science team became very familiar with the various prospect targeting models that the client had developed based on primarily firmographic data. Having this view allowed the ethnographers to better prioritize new hypotheses and data proxies for extending and improving the targeting model using different types of proprietary data that the client had already collected.

***Data analysts need to embrace the intangible* –** Unless data science teams strive to understand the underlying insights delivered by the ethnographic team, and not just the accompanying data proxies, the representation of those insights in the algorithm is put at risk. During the course of the credit card prospect targeting project, for example, we saw firsthand how a lack of qualitative understanding can disrupt the implementation of insights. The ethnographic team surveyed over 20 SMBs and found that those with a strong "community mindset" were much less likely to respond to an unsolicited credit card offer. In order to track this insight, the ReD and Cognizant team delivered the finding along with a list of proxies they believed would indicate a business would fall into that category: proximity to local bank, number of credit cards owned, change in geo-location field as a sign of frequent relocation, etc. Upon presenting their findings, however, the client told researchers that turning those insights to variables would take several months. The client didn't understand that ReD and Cognizant were proposing new uses for old data fields, which gave them new meanings and situations based on fresh insights. As a result, they delayed model development in order to test each individual proxy for completeness and validity.

In order to make the leap from qualitative insight to quantitative proxies, data teams must get familiar with high-level insights that cannot be measured in a direct 1:1 relationship with the data sources. That is, the use of qualitative observation requires data analysts to think of the underlying human behaviors and actions behind an insight, and how both new and existing data sources can capture them in creative ways. Thus, the best data analysts are not just great mathematicians, but also have what C. Wright Mills called a "sociological imagination": the ability to pull their thinking away from the technical problem at hand, and to understand the interactions that are actually taking place between the points within a dataset and the contexts into which they will be applied.[35]

To cultivate sociological imagination, data analysts working with ethnographers likely need to take some part in the ethnographic research process when developing the first concepts of what the proxies will be. A data analyst was able to fulfill this role in the prospect targeting case. He joined the team on workday tagalongs, and left feeling that he grasped where some of the insights came from and thus what the essential point to identify through the proxy really was. For example, ethnographers found that SMB owners who were plugged into "expert networks" like trade and industry associations or LinkedIn were more likely to turn to those networks and less likely to respond to offers. While the client's data science team could not initially conceptualize any feasible proxies to track this insight for the first model, this particular data analyst was able to push the client's team to see if any patterns in transaction data could indicate this type of behavior. Expert data analysts are up to the task of creatively translating social-science based insights, but only after taking a methodological leap of faith in order to understand the contexts from which those insights actually emerge.

PHASE THREE: DEPLOYMENT

Central to the final phase of Contextual Analytics is the development of two critical deliverables: 1) the final model with a proof point that demonstrates the impact of social-science hypotheses on the algorithm, and 2) the overarching "story" that is able to explain why that algorithm actually works.

Neither the story nor the statistics, however, are particularly novel on their own. Rather, the power of a Contextual Analytics project is that collaboration between data and social scientists allows the two deliverables to be linked through a set of mid-level theories designed to balance the specificity of big data proxies and the high-level inspiration of ethnographic narratives.

Contextual Analytics in Practice: The Deployment Phase

Neither big nor thick data can be used as the sole end-point of algorithm development. Qualitative insights are often thought too small to be reliable on their own. Numbers alone, on the other hand, lack the persuasive power to enact change within an organization. Thus, after aligning on proxies, integrated workstreams show their value to the client through two mutually informative deliverables – numbers made emotional, and stories given statistical weight

According to a 2017 Cognizant report, most corporate decision-making remains "gut-based" despite the infusion of new data collection, management, and analytical technologies.[36] This is because data analysts find it difficult to keep on top of business needs while juggling multiple algorithm development projects.[37] As a result, data teams are often technically well-prepared, but ill-equipped to demonstrate the overarching business impact of their "black box" algorithms. In response, social scientists focus on providing the "why" behind new correlations in the model. In the project on curbing attrition and retaining skilled employees, for example, the data science team created a list of over 200 proxies for predicting attrition. This level of granularity meant little to the human resource officers who sought to understand why that attrition happened in the first place. Simultaneously, then, the social science team created a values framework to map onto the proxies that predict attrition and, more importantly, describe the motivations and needs of employees when it comes to their work-life ambitions. This values framework ultimately gave the HR team the explanation they needed in order to identify key employee rationales behind their choosing to stay with or leave the company. The ethnographic "big story" provides buy-in and articulates need on behalf of big data.

As applied ethnographers quickly learn while working in a corporate context, however, scale is often valued over story. Moreover, insights based on ethnographic research can run the risk of seeming anecdotal. As an ethnographer on the improving call center operations project put it, even if the team had a great story to tell, that story is at its strongest when coupled with statistical proof: "It's that magical number that we are chasing. You need the numbers, and you need them to concretely say that we improved the model by this much, and the client operations improved this much because of it."

Key Success Criteria for the Deployment Phase

The creation of two separate, though mutually informing, deliverables provides the client with the technical tools and persuasive stories to build effective and convincing algorithms. When the integrated team separates to build different products. they again tend to focus on the type of product that comes easiest to them: ethnographers developing a governing thought and data scientists focusing on specific proxies. Through all the cases, however, we saw the importance of keeping the two workstreams aligned by pushing each practitioner to embrace mid-level theories in their work.

Focus on mid-level theory – A concept first developed by Robert K. Merton, mid-level theory is an approach to sociological theory construction that balances a focus on the overarching big idea and the particularities of various lower-level hypotheses. As Merton defines the term:

> "...what might be called theories of the middle range: theories intermediate to the minor working hypotheses evolved in abundance during the day-by-day routine of research, and the all-inclusive speculations comprising a master conceptual scheme."

In the case of Contextual Analytics, mid-level theory functions as a method by which both data and social scientists embrace the needs of the other in order to build more well-rounded deliverables.

For social scientists, embracing big data methods means achieving a level of specificity in insights that might be uncommon and uncomfortable for many ethnographers. Given rapid advances in analytics, it is no longer helpful to produce only a single, abstract overarching theory to sum up an entire project. Rather, ethnographers need to meet in the middle with data analysts by splitting their theories into more granular and focused recommendations custom-tailored for different versions of a model. This specificity is critical to ensure that the stories ethnographers tell are internalized and utilized to their maximum capacity across a company's data portfolio.

Aligning on mid-level theories also pushes data scientists to go beyond what can be meaningfully quantified in the data. For example, while building another iteration of the model in the prospect targeting case, data analysts presented several proxies that hadn't yet passed the client's compliance muster, and therefore pushed the boundaries for what the client considered technically feasible in the model. Nevertheless, using examples from ethnographic work, the data team persisted in their argument to include these proxies and convinced the client that they delivered some "incremental value." By elevating their thinking past the technical, they ensured that the next version of the model would stay true to deep human insights.

Ensuring the continued use of integrated social science and data science methodologies is not as easy as checking two separate boxes. Rather, by baking mid-level theory into the way that they produce deliverables, teams ensure that the collaborative spirit of Contextual Analytics extends through the project's resolution.

CONCLUSION

Extending Frames of Reference

In practice, the building of contextually-informed algorithms calls for qualitative and quantitative studies that go beyond merely observing the complementarity of statistical analysis and ethnographic research.[38] Indeed, as we have seen, simply putting the practitioners of big and thick data together does little to enable their practical collaboration. Both social scientists and data analysts must make attempts to better understand one another's project processes in each phase – Framing, Data Collection and Synthesis, and Deployment – in order to facilitate the integration of their findings. We have learned that whether through frequent communication, forced immersion, or mid-level theory, a Contextual Analytics project relies on practitioners who are willing to move beyond their original frames of reference. We encourage continued conversations around the integration of big and thick data workstreams and the practical collaboration of their practitioners, and hope that future discussions will further engage the practicalities of Contextual Analytics processes in corporate settings.

NOTES

1. Oremus, Will (2017, October 16). "How 'Big Data' Went Bust. And what comes next." Slate.

2. Dykes, Brent (2018, August 22). "Build a Data-Driven Culture, One Meeting at a Time." Forbes.

3. Burton, Besty and Hank Barnes. (2017, August 11). "2017 Hype Cycles Highlight Enterprise and Ecosystem Digital Disruptions: A Gartner Trend Insight Report." Gartner.

4. Boelstorff, Tom (2015). "Making Big Data: In Theory." Boelstorf and Maurer (Eds.), *Data, Now Bigger and Better!* (87-108). Chicago, IL: Prickly Paradigm Press.

5. Gartner (2017) "Forecast: Internet of Things - Endpoints and Associated Services, Worldwide, 2017." *Gartner.;* PricewaterhouseCoopers (2015): "Seizing the Information Advantage: How organizations can unlock value and insight from the information they hold." A PwC report in conjunction with Iron Mountain.

6. Wang, Tricia. (2013, May 13). "Big Data Needs Thick Data." Ethnography Matters;

7. Rattenbury, T. and D. Nafus (2018): "Data Science and Ethnography: What's Our Common Ground, and Why Does It Matter?" *Ethnographic Praxis in Industry 2017*.

8. Haines, J. (2017): "Towards Multi-Dimensional Ethnography." *Ethnographic Praxis in Industry Conference Proceedings 2017*.

9. Kolsto, E. (2017): "A Researcher's Perspective on People Who Build with AI." *Ethnographic Praxis in Industry*

Conference Proceedings 2017.

10. Crawford, Kate (2017): "The Trouble with Bias." *Conference on Neural Information Processing Systems (NIPS),* Keynote Address 2017.

11. See, for example, the Facebook-funded post on the New York Times: (2018) "Artificial Intelligence: How We Help Machines Learn." The New York Times Paid Post.

12. Boelstorff (2015)

13. Seaver (2015): 37.

14. Ibid.

15. Ibid. 44

16. Salganik, Matthew (2017) *Bit by Bit: Social Research in the Digital Age.* Princeton University Press: 18

17. Lupton, Deborah (2015, May 11) "The Thirteen Ps of Big Data," *This Sociological Life.*

18. Donoho, David. (2015) "50 Years of Data Science." In *Princeton NJ, Tukey Centennial Workshop.*

19. Wang (2013); Geertz, Clifford. (1973). *The interpretation of cultures: selected essays.* New York: Basic Books.; Madjsberg, Christian (2017) *Sensemaking: The Power of the Humanities in the Age of the Algorithm.* New York: Hachette Books.

20. The data scientist Cathy O' Neil exposes some of the biased data behind opaque algorithms used to predict patterns in policing and felony – algorithms which, through their aura of ostensible objectivity, "create the reality they purport to describe." O'Neil, Cathy (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* Crown Publishing / Random House. See also Osoba, O. and W. Welser IV (2017): "An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence." Santa Monica, Calif.: RAND Corporation, RR-1744-RC.

21. Brynjolfsson, Erik and Andrew McAfee (2018): "The Business of Artificial Intelligence: What it can - and cannot - do for your organization." *Harvard Business Review.*

22. Storey, Margaret-Anne (2016, May 4) Lies, Damned Lies and Software Analytics: Why Big Data Needs Thick Data [Webinar].

23. Lepri, Bruno et al. (2017) "The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good" in *Transparent Data Mining for Big and Small Data*: 3 – 24

24. Ibid.

25. One such example is the "Laboratorio Urbano" in Bogota, Colombia, where city government, citizens, architecture firms, and academic institutions collaborated to test out pilot plans for new mobility and safety policies in the rapidly changing city. (The Tyranny of Big Data, 16) For other creative uses of public data see: Nafus, Dawn (2016). "The Domestication of Data: Why Embracing Digital Data Means Embracing Bigger Questions." Ethnographic Praxis in Industry Conference Proceedings. 2016. 384-399.

26. Smith, A. (2018). "Smart cities need thick data, not big data." *The Guardian.* Accessed April 18, 2018

27. Following an Anthropology of the Contemporary (ARC) model, Collaboratories examine the tactics of experts in order to interrogate their traditional methods of knowledge production. Lepri, Ruppert, Penny Harvey et al. (2015) "Socialising Big Data: From concept to practice." CRESC Working Paper Series. CRESC, The University of Manchester and the Open University.

28. Ibid.

29. Pascual, Al et al. Javelin Strategy and Research. (2017, February 1) *Identity Fraud: Securing the Connected Life*.

30. See Creswell, J. W. & Poth, C. N. (2017). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.)

31. Almost half of all Americans have had an experience with credit card theft in the last five years, and financial losses to credit from fraud are on track to reach $22 billion annually by 2022. Knieff, Ben. Universal Payments Aite Group. (2016): "2016 Global Consumer Card Fraud: Where Card Fraud is Coming From" Boston, MA: 7.

32. Koslo (2017)

33. Madjsberg, Christian. (2018, February 23). "Thanks, Robots! Now These Four Non-Tech Job Skills Are in Demand." *Fast Company*.

34. Crook, Izaak. (2018, Apr 06). "How GDPR Will Affect Data Science." Dataconomy.

35. Mills, C. Wright (1959) *The Sociological Imagination*. New York: Oxford University Press.

36. Krishnamurthy, Karthik, et al. Cognizant. (2017). "Analytical Storytelling: From Insight to Action." Cognizant White Paper: 4.

37. Kolsto (2017)

38. Venturini, Tommaso and Bruno Latour (2010) "The Social Fabric: Digital Traces and Quali-quantitative Methods." Proceedings of Future En Seine.

REFERENCES CITED

Blok, A. and M.A. Pedersen
2014 "Complementary social science? Quali-quantitative experiments in a Big Data world." *Big Data & Society* 1(2):1–6.

Boelstorff, Tom
2015 "Making Big Data: In Theory." Boelstorf and Maurer (Eds.), *Data, Now Bigger and Better!* (87-108). Chicago, IL: Prickly Paradigm Press.

Bornakke, Tobias and Brian L. Due.
2018 "Big-Thick Blending: A method for mixing analytical insights form big and thick data sources." *Big Data and Society*: 1: 1-16.

Bowker, G. C.
2013 "Data Flakes: An Afterward to 'Raw Data' Is an Oxymoron", in L. Gitelman (eds) *"Raw Data" Is an Oxymoron*, MIT Press, Cambridge, MA: 167-171.

Boyd, D. and Kate Crawford
2012 "Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society,* 15 (5): 662-79.

Burton, Besty and Hank Barnes
2017 "2017 Hype Cycles Highlight Enterprise and Ecosystem Digital Disruptions: A Gartner Trend Insight Report." Gartner. August 11, 2017.

Crawford, Kate
2017 "The Trouble with Bias." *Conference on Neural Information Processing Systems (NIPS),* Keynote Address 2017.

Creswell, J. W. & C.N. Poth
2017 *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Sage Publications, 2017.

Crook, Izaak
2018 "How GDPR Will Affect Data Science." Dataconomy. Retrieved from: https://dataconomy.com/2018/04/how-gdpr-will-affect-data-science/. April 06, 2018.

Facebook, The New York Times Paid Post.
2018 "Artificial Intelligence: How We Help Machines Learn." The New York Times Paid Post. Retrieved from: https://paidpost.nytimes.com/facebook/artificial-intelligence-how-we-help-machines-learn.html?tbs_nyt=2018-feb-nytnative_apmod&cpv_dsm_id=189293206.

Frank, Malcolm, Paul Roehrig, and Ben Pring
2017 *What To Do When Machines Do Everything: How to Get Ahead in a World of AI, Algorithms, Bots, and Big Data.* New York: Wiley, 2017.

Geertz, Clifford
1973 *The interpretation of cultures: selected essays.* New York: Basic Books.

Haines, Julia Katherine
2017 "Towards Multi-Dimensional Ethnography." *Ethnographic Praxis in Industry Conference Proceedings 2017.*

Kolsto, Ellen
2017 "A Researcher's Perspective on People Who Build with AI." *Ethnographic Praxis in Industry Conference Proceedings 2017.*

Knieff, Ben. Universal Payments Aite Group.
2016 "2016 Global Consumer Card Fraud: Where Card Fraud is Coming From." Boston, MA. Retrieved from: https://www.aciworldwide.com/-/media/files/collateral/trends/2016-global-consumer-card-fraud-where-card-fraud-is-coming-from.pdf

Krishnamurthy, Karthik, Melissa Morello, and James Jeude. Cognizant.
2017 "Analytical Storytelling: From Insight to Action." Cognizant White Paper. Retrieved from: https://www.cognizant.com/whitepapers/analytical-storytelling-from-insight-to-action-codex2475.pdf

Lepri, Bruno, Jacopo Staiano, David Sankokoya, Emmanuel Letouze and Nuira Oliver.
2017 "The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good" in *Transparent Data Mining for Big and Small Data.* 3: 3 – 24.

Lupton, Deborah
2015 "The Thirteen Ps of Big Data," *This Sociological Life.* Retrieved from: https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/. May 11, 2015.

Madjsberg, Christian
2017 *Sensemaking: The Power of the Humanities in the Age of the Algorithm.* New York: Hachette Books.

2018 "Thanks, Robots! Now These Four Non-Tech Job Skills Are in Demand." *Fast Company.* February 23, 2018.

Margolis, Abby
2013 "Five Misconceptions about Personal Data: Why We Need a People-Centered Approach to 'Big' Data." *Ethnographic Praxis in Industry Conference Proceedings* 2013.

Mills, C. Wright
1959 *The Sociological Imagination*. New York: Oxford University Press.

Nafus, Dawn
2016 "The Domestication of Data: Why Embracing Digital Data Means Embracing Bigger Questions." *Ethnographic Praxis in Industry Conference Proceedings 2016*.

O'Neil, Cathy
2016 *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing & Random House.

Oremus, Will
2017 "How 'Big Data' Went Bust. And what comes next." Slate. Retrieved from: http://www.slate.com/articles/technology/technology/2017/10/what_happened_to_big_data.html. October 16, 2017.

Osonde A. Osoba and William Welser IV
2017 "An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence." Santa Monica, Calif.: RAND Corporation, RR-1744-RC.

Pascual, Al, Kyle Marchini, and Sarah Miller. Javelin Strategy & Research
2017 "2017 Identity Fraud: Securing the Connected Life." Retrieved from: https://www.javelinstrategy.com/coverage-area/2017-identity-fraud. February 21, 2017.

PricewaterhouseCoopers
2015 "Seizing the Information Advantage: How organizations can unlock value and insight from the information they hold.*" A PwC report in conjunction with Iron Mountain.*

Rattenbury, Tye and Dawn Nafus
2018 "Data Science and Ethnography: What's Our Common Ground, and Why Does It Matter?" *Ethnographic Praxis in Industry 2017*.

Ruppert, Evelyn, Penny Harvey, Celia Lury, Adrian Mackenzie, Ruth McNally, Stephanie Alice Baker, Yannis Kallianos, Camilla Lewis
2015 "Socialising Big Data: From concept to practice." CRESC Working Paper Series. CRESC, The University of Manchester and the Open University.

2015 "A Social Framework for Big Data." CRESC Working Paper Series. CRESC, The University of Manchester and the Open University.

Salganik, Matthew
2017 *Bit by Bit: Social Research in the Digital Age*. Princeton & Oxford: Princeton University Press.

Seaver, Nick
2015 "Bastard Algebra" Boelstorf and Maurer (Eds.), *Data, Now Bigger and Better!* (27 - 46). Chicago, IL: Prickly Paradigm Press.

Smith, Adrian
2018 "Smart cities need thick data, not big data." *The Guardian*. Retrieved from: https://www.theguardian.com/science/political-science/2018/apr/18/smart-cities-need-thick-data-not-big-data. May 18, 2018.

Stirrup, Ashley
2018 "The New "V" for Big Data: Virtue," *Business 2 Community*, 18 January 2018. Retrieved from: https://www.business2community.com/big-data/new-v-big-data-virtue-02004798. January 31, 2018.

Storey, Margaret-Anne
2016 Lies, Damned Lies and Software Analytics: Why Big Data Needs Thick Data [Webinar]. In *ACM SIGSOFT Webinar Series*. Retrieved from: https://www.slideshare.net/mastorey/lies-damned-lies-and-software-analytics-why-big-data-needs-rich-data. May 4, 2016.

Venturini, Tommaso and Bruno Latour
2010 "The Social Fabric: Digital Traces and Quali-quantitative Methods." *2010 Proceedings of Future En Seine.*

Wang, Tricia
2013 "Big Data Needs Thick Data." Ethnography Matters. Retrieved from: http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/. May 13, 2010.

Wilson, William J., and Anmol Chaddha
2010 "The Role of Theory in Ethnographic Research." *Ethnography* 10(4): 549-564.

Who and What Drives Algorithm Development: Ethnographic Study of AI Start-up Organizational Formation

RODNEY SAPPINGTON, Ph.D.
Acesio Inc.

LAIMA SERKSNYTE, Ph.D.
Acesio Inc.

The focus of this paper is to investigate deep learning algorithm development in an early stage start-up in which edges of knowledge formation and organizational formation were unsettled and contested. We use a debate by anthropologists Clifford Geertz and Claude Levi-Strauss to examine these contested computational forms of knowledge through a contemporary lens. We set out to explore these epistemological edges as they shift over time and as they have real practical implications in how expertise and people are valued as useful or non-useful, integrated or rejected by the practice of deep learning algorithm R&D. We discuss the nuances of epistemic silences and acknowledgments of domain knowledge and universalizing machine learning knowledge in an organization that was rapidly attempting to develop algorithms for diagnostic insights. We conclude with reflections on how an AI-Inflected Ethnography perspective may emerge from both, data science and anthropology perspectives together, and what such a perspective may imply for a future of AI organizational formation, for the people who build algorithms and for a certain kind of research labor that AI inflection suggests.

Keywords: AI, Deep Learning, Algorithm R&D, Epistemology, Domain Knowledge

SETTING THE SCENE

We are in a 3,000 square foot office space and from a small board room we hear "that's domain knowledge—we'll get subject matter experts for that!" Team members debate over hiring radiologists and hiring expertise outside of algorithm development. They are puzzling over the usefulness of radiologist knowledge in developing deep learning algorithms that are to serve radiologists in their diagnostic interpretations. We have arrived at an early stage deep learning startup in Silicon Valley near a strip of park land where cyclists and joggers stream by an area that once fell into neglect and was reborn into a corridor of fusion restaurants and tech companies.

We find ourselves among decades of tech venture capital infusion set against university and residential growth. You can smell chlorine from Olympic sized pools and spas, coconut-crusted shrimp from a high-end bistro and musty ash from recent wild fires. Flocks of feral parrots sound off in the trees high above, escapees of nearby ranch-style living rooms. We have arrived at what social network theorist Mark Granovetter has termed an "innovation cluster" (Ferrary & Granovetter 2009) of artificial intelligence (AI) or what Andrew Ng has termed the "the new electricity" (Ng 2018). The term refers to algorithms that could be described as "rocket engine[s]" in which "huge amounts of data" can be processed lifting the rocket ship of machine intelligence to new heights (Garling 2015). Jeff Dean, the head of AI at Google assesses Deep Learning algorithms as keys to AutoML, machines that learn to learn and are foundational for early detection of a wide range of diseases (Dean 2018). Such start-up companies that Peter Diamandis describes as "smaller" and "nimble" have evolved out of a linear industrial era into an "exponential" era, a period of market disruption, unpredictability and 100x algorithmic growth (Koulopoulos 2018).

Based on our real experience of such a start-up, the concepts and possibilities of our case study of Deep Learning algorithm development and new organizational formation emerge out of this exponential petri dish.

STRUCTURE OF PAPER

This paper explores original research set within deep learning algorithm development in an early stage start-up organization, from 2014-2015. At the time I (Rodney Sappington) served as a Senior Data Scientist focused on developing algorithms for the early detection of disease. Together with my collaborator we analyze features of deep learning algorithm development which include observations and interactions with strategic partners, clinical leaders and radiologists. We also refer to experiences of highly regarded team members with whom I was involved and had the pleasure in part to lead in algorithm and clinical development. Not the least we have developed our perspective out of references across the social and behavioral sciences and discussions with colleagues outside the walls of algorithm development.

The structure of this paper includes conceptual and case-study analyses. We contextualize our research by way of briefly introducing a passionate debate between Clifford Geertz and Claude Levi-Strauss on the fear and embrace of computational forms in ethnographic practice. This debate is not a gentle or dated conceptual tug of war between renowned anthropologists but instead a contemporary lens through which we can investigate contemporary algorithm development and messy problems of building algorithms with medical diagnostic capabilities. We briefly explore the appearance of deep learning algorithms in Silicon Valley. We turn to early stage start-up features and move to the complexities of a case of a health insurance company that offers large amounts of data and a problem for algorithm development with a twist. We go inside the organization and examine hiring practice and the role of domain knowledge in a machine learning start-up with a scene at a company retreat. We explore the problem of positionality of being a data scientist and anthropologist in the field of applied machine learning. We attempt to open up how the diagnostic patient is being conceptualized. In conclusion we provide observations on epistemological tensions in algorithm development which we term *AI-inflected* ethnography. We are not suggesting a methodology but instead pointing to a way of inhabiting these tensions and silences at the core of algorithm development, an approach that opens up a view into how organizations and organizational members get constituted and sometimes unravel.

FEAR-EMBRACE OF COMPUTATIONAL FORMS

There could not have been two more different scholars studying human behavior and culture than Clifford Geertz and Claude Levi-Strauss. For Geertz field research was immersive, forged in dust, blood and side-bets surrounding Balinese cock-fights. For Levi-

Strauss field research was forged as a product of field work in conceptions of binary order and kinship system. For them a certain kind of intelligence and mind of the ethnographer was at stake, it was almost at once human-driven and machine/system-driven:

> *Society is, by itself and as a whole, a very large machine for establishing communication on many different levels* (Levi-Strauss 1953).

> *Lévi-Strauss has made for himself...an infernal culture machine. It annuls history, reduces sentiment to a shadow of the intellect* (Geertz 1973).

On one side Geertz viewed ethnography as a practice of interpreting human speech and gesture in the wild and intentionality in everyday life, which he called "deep play." On another side Levi-Strauss viewed ethnography as a scientific practice of interpreting universal codes, totems and patterns across society, it was "structural." The gist here was a tension: a type of human perception and cognition emerging *in* local everyday life versus a type of human perception and cognition emerging as universal patterns *across* everyday life. For Levi-Strauss information science held a central place in creating and interpreting culture which implied both human and non-human forms. For Geertz information science suggested an infernal (hellish) culture machine.

There was another kind of legacy that was epistemological. It drove their passion and still fuels passions today in machine learning. As anthropologists and data scientists will still largely live and work in this legacy, categories of "hard" and "soft" knowledge, universality and particularity, probabilities and possibilities of quantified judgment and human intuitive judgment, structured and unstructured data and embodied and cognitive forms of intelligence. I take the Geertz-Levi-Strauss debate as a struggle for thinking how we build, imagine and fashion algorithms for human benefit and machine automation.

DEEP LEARNING ALGORITHMS EMERGENCE IN SILICON VALLEY CONTEXT

Andrew Ng and others in 2012 built high level features using deep learning to recognize and classify cat videos at a 70% improvement over previous "state-of-the-art" networks (Quoc et al 2012). Using computational power, 16,000 computer processors, the deep learning network was presented with 10 million digital images found in YouTube videos. This was a breakthrough and supplied proof that certain deep neural networks could perform (automatically learn without human hand-coding) across complex image sets. This was the beginning of the successful use of convolutional neural networks (CNNs). Two years later overutilization of medical imaging in healthcare made radiology ripe as a testing ground to apply some of the lessons learned from 2012. New early stage organizations began to take shape to productize these findings. The context was Silicon Valley in which a line-up of similar innovations had set an aspirational and venture capital stage for deep learning algorithm development. Expectations were high. From Apple, Uber, Lyft, Google, Airbnb, Salesforce, Tesla and Twitter to name a few, today nearly 30 technology companies are so-called "unicorns" or near "unicorn" status totaling close to $140B in value (Glasner 2018). It has also brought together venture funds. Venture funded machine learning start-ups has recently almost defined the San Francisco-Silicon Valley region in terms of company valuations. As the global economy is expected to grow by 3.5% GDP, venture backed AI startups have had an expected growth rate of over 40% by 2020 with a U.S. market valuation of AI by 2035 of $8.3T (Faggella 2018). Along with this growth came warnings that we "must be thoroughly prepared—intellectually, technologically, politically, ethically, socially—to address the challenges that arise as artificial intelligence becomes more integrated in our lives" (Faggella 2018).

The new start-ups are often described fondly in the industry as "Moonshots." Singularity Ventures founded by Ray Kurzweil also terms them "exponential startups." These are organizations that claim to hyper-scale across person, transaction and global impact. Founders within them are typically referred to as super smart. They exemplify that the social good of machine learning goes beyond a mere transactional machine or consumer recommendation system or ad placement. As the perspective goes these are exceptional people building exponential machine learning products with exceptional resourcefulness. What they know and how they apply what they know has become a global phenomenon that others try to emulate.

This smaller-nimble organizational type has been described by Ferrary and Granovetter as one "node" in "a durable" assembly of organizations that has an almost magical capacity to "anticipate, learn and innovate in order to react to major internal or external changes" (2009). This is a lot to ask of people building the new global electricity with industry expectations to touch every human life on the planet.

DEEP LEARNING EARLY STAGE START-UP FEATURES

How did these expectations measure against our real-life start-up experience?
Features of our early stage start-up were less than magical and included product discovery instead of product development, immediate press coverage and AI hype, fear of replacing radiologists by robots, extreme leadership pains, and intense VC oversight and intervention at the expense of a vulnerable organizational culture. It comprised of a mix of possibility, struggles over knowledge to be brought into the organization and algorithm architecting and iteration. Trial and error were essential ingredients to gaining replicable results and to rapidly building an early stage start-up.

In daily operations, the team's approach to capturing disease was not at first to understand it but to start with a particular, by identifying the features of lungs, their shapes, edges, anomalies. Our team had to first segment thousands of lungs before we could begin to achieve any results with our algorithms in identifying lung nodules. This meant using chest x-ray and chest CT's pixel data and annotations from publicly available data sets that were by no means perfect data and required preprocessing and extensive labeling. The lungs are a vascular world onto themselves. Millions of veins spread out into the lungs and different types of scar tissue can be present that obscure cancerous nodules until late stage becomes incurable. In building our algorithms we were working with convolutional neural networks (CNNs) that have the ability to auto discriminate image features.

This was an exciting time to work at an early AI start-up. Deploying and testing CNNs was a creative endeavor in 2014 when these algorithms were not in wide application for medical imaging analysis. It was a time in which venture capital investors were smitten with deep learning algorithms and needed only a good team and a good idea to trigger an investment. It was a time in which the horizon of what was algorithmically possible in medicine was at an inflection point but the practical application and proof of good performing algorithms were sometimes daunting to demonstrate. Additionally, this excitement came with the price of reducing human/patient complexity to the purity of an all-powerful algorithm that could be generalized across medical contexts.

Ideas and conflicts abounded. They were worked through when we took 'one-on-one' walks. We walked along grassy walk ways puzzling through who to bring in as consultants or full-time employees (FTEs). We white-boarded approaches. As I walked with the lead scientist we often considered bringing in oncologists, radiologists, primary care physicians. Social scientists were seen as "not useful." The social sciences and ethnographic knowledge were a hinderance to successful algorithm development at this time. Clinical expertise was considered but walled off, kept as consultants, advisors, reduced to domain knowledge. In other words, the highly skilled area of radiology could not travel well beyond radiology but data science could travel and exceed radiology workflows, image interpretation, disease classification, sub specialties and the human. What we faced during this period was how to formulate problems and how to formulate a diverse team who held particular forms of knowledge appropriate to the problems we were trying to solve. It was difficult and slippery discussions. No one seemed to have a magic-bullet answer.

When faced with a lack of diversity of knowledge around algorithm development Jeff Dean, Head of AI at Google has stated:

> I am personally not worried about an AI apocalypse" but "I *am* concerned about the lack of diversity in the AI research community and in computer science more generally" (Dean 2016).

Not only he was concerned about encouraging people from different backgrounds to build algorithms, he was thoughtful that certain forms of thinking may not get into algorithm development. Experts from the Google Brain Residency program, which would have been a feeder for such diversity recruitment were composed of "physicists, mathematicians, biologists, neuroscientists, electrical engineers, as well as computer scientists" (Dean 2016). These were largely STEM practitioners. This range of diversity did not include unexpected perspectives. Diversity appeared narrow, bounded.

The fear-embrace of computational forms could be viewed as an epistemological tension between those whose knowledge contributed to algorithm development and those whose experience and knowledge was viewed as consultative. Such consultative knowledge was typically referred to as "domain expertise" and could be reduced to a kind of consulting artifact. As indicated in Dean's statement domain expertise may not have gone completely unacknowledged, it was called for but not followed through with as evidenced in Dean's listing. A STEM defined in this way could screen out social sciences, physicians, policy experts, artists, ethicists, community members and patients to name a few. Such screening out came in the form of particular/local knowledge that was perceived as not algorithmically scalable across industries. Could ethics scale across industries? Could a patient's experience of navigating and overcoming a deadly cancer and a fractured healthcare system scale? We are not questioning Google's idea of inclusiveness or diversity. We are pausing on what/who gets persistently divided up as contributory and held up as core knowledge in algorithm development in an organizational context.

When it comes to marginal, diverse or unexpected perspectives not held in high regard in algorithm development, two words come to mind – *problem formation*. The kinds of problems that get identified and privileged for algorithm development are shaped by the kinds of people who are brought together to identify and attempt to solve those problems. Problem formation is as valuable as problem solving. A red light is as valuable as a green light for a specific algorithm project. When diversity of perspectives, experiences and background is a slogan-only proposition, acute problems simply are invisible to machine learning innovation or worse, they are seen as exciting problems that have positive social value when they instead have potential negative societal consequence. Different kinds of problems drive different outside engagement practices.

PROBLEM FORMATION COMING THROUGH THE DOOR

Machine learning problem formation could come through the door from strategic partners and data providers. Problems were not always defined from behind the organizational door among team members. Locating and defining a problem that fit the capabilities of algorithm development was sometimes called "product-market fit" but in terms of our case it was also an effort to locate large data sets and then allow the problem to emerge from the data. An attractive offering of large data sets could supersede a more sober problem formation process. The team around the table and their backgrounds and training often determined which problem got selected that could set the organization down a developmental road for months and years.

An Example

It was a bright brisk northern California day, the sun bouncing off pavement at a nearby corporate square. It couldn't be more gorgeous. For months we had been negotiating the terms of a partnership with a large health insurer that promised near-term revenue for us. The health insurer team arrived outside the front of our building and discussed things before collectively announcing themselves and walking in. They greeted us in sharply dressed suits and strong handshakes. We all entered a meeting room and settled around a glass table with a white-board of diagrams of medical imaging archiving, workflows and model layers. I turned the white board around and they began by introductions and then launching into the billions of transactions they do each day. One of their team members a man in his mid 50s hair combed straight back who appeared to be a 1970's version of a suburban executive, call him J., began describing their value in terms of transactional data and the possible extent of a strategic partnership with them, "as you know, we have all the data on the patient journey, how the patient is treated and if they go to X hospital to Y pharmacy, rehabilitation center, pharmacy - you know, we have the whole thing." He punctuated his brief description by "we have more patient data than we even know." It was billons of transactions. The project as they described it, was for us to build algorithms using this massive transactional data set.

"We want to eliminate unnecessary [insurance] audits, get rid of them all together if we can," he said.

"You mean a percentage of the audits you conduct" I responded.

"Yes, they're a waste of time."

Their machine learning lead engineer, a man in his mid-thirties in a plaid shirt began to lay out some ideas around predictors and unsupervised learning strategies that could work together to assist in this direction. The goal was to identify the probability of medical fraud and reduce unnecessary insurance audits. They were looking for precision in what was known as "fraud detection."

Algorithm development was to provide a means to rank medical practices in terms of probability of committing fraud not in terms of actually committing fraud. On one level such algorithms could save private practices the back-breaking process of unnecessary insurance audits that could cause mountains of paperwork, anxiety, and administrative time. They wanted our algorithm development to share in a moral victory for physicians and physician practices, we would be the good guys, less audits less onerous oversight less unnecessary phone calls and less legal expenses for the private medical practice. It was a hero's problem and we could be the heroes to solve this problem.

To get the problem in focus I paraphrased it for our meeting to make sure I understood. "You want us to help you build predictive models based on transactional and temporal data that would save millions of dollars of wasted effort in wrongful audits of potentially thousands of private practices."

"Not wrongful, but yes to help save millions in unnecessary audits." He said, they would invest and provide all the data we needed.

Around the table I could feel a kind of moral victory flag being raised. But something was not quite right.

I asked if their mission of reducing audits was all they were intending with our algorithms. He said it was "hard to determine future uses" and technology was "moving so fast" but this was absolutely their "main use case."

"If we are building algorithms to assist you in identifying *less* unnecessary audits couldn't these same algorithms be repurposed to help you identify *more* audit opportunities? You want us to help you reduce audits and I understand this, but couldn't you just as easily increase your audits with our technology? Couldn't you become an audit powerhouse in some way?

"That would not be good for business" he said and then defensively mentioned "we're looking for the right team, it's a great opportunity." I felt our CEO shift in his chair. We needed the data.

Then J. said something more interesting. "In reality", we "rarely commit unnecessary audits, our approach to audits even if they do occur are over years." He took a drink of water as if to refuel and then said, "for example, we never audit a practice twice in the same year and most practices never get audited at all."

My own experience was different. Coming from a private surgical practice background I had been part of exactly such unnecessary insurance audits and their impact on staff. These effects were fresh on my mind. In fact, my practice was audited twice in the same year by this same insurer with no upcoding or wrong doing of any kind found. I was the one who actually experienced this back-breaking administrative work first hand. I experienced late hours, certified letters, operation reports and patient records reassembled daily based on changing requests from the insurance auditing department. How could he have known that I as a data scientist could possibly have experienced this same insurance auditing process. I kindly responded.

"I think that may be inaccurate, my practice was audited twice by you last year and we came away with [our] claims in order."

He quickly shot back. "That's very unusual, we don't conduct audits this way. The name of your practice?"

The point was that they did indeed operate this way and audited more than once in the same year and unnecessarily. From their perspective such algorithm development was a win-win saving time and money for their insurance company and saving human distress and labor for private medical practices across the country. However, "future uses" were to be determined.

Digging deeper, algorithms that were designed to decrease audits could be weaponized to increase audits across millions of medical practices. Physicians could learn quickly to avoid complex patients that might have presented a risk for a billing error or they might experience another consequence of the ongoing threat of audit, and bail out of independent practice altogether and become an employee of the local hospital leaving billing responsibility and legal exposure to the hospital. The ongoing threat of medical audit could reshape networks of private medical practice ownership. No doubt medical fraud has been a key challenge in U.S. healthcare with the justice department in July 2018 announcing 1.3 billion in fraudulent claims across doctors and treatment facilities. A large number of these were for over-prescription of opioids and false billings. On the other hand, delinquency aside, the impact of unnecessary and aggressive insurance audits known as "fraud detection" could collapse a resource-strapped medical practice, drive physicians from owning their own practices, encourage consolidation of medical practices by large health systems and breed a culture of reimbursement fear. There was irony in an early stage machine learning organization asked to take on such a project that could accelerate the destruction of the very kind of organizational ethos it holds so dear: entrepreneurism.

In the current climate of insurance audits it was not the evidence of fraud but the evidence of a mistake that could trigger an audit. A single mistake could trigger a multi-year audit over hundreds or thousands of patient encounters. In my previous experience a misspelled word in an operation report could trigger a process that would lead to a claw-back of hundreds of thousands of dollars. It could cause crushing legal fees, employee burn out and an ever-present anxiety of the next audit always around the corner. Just the threats of audits could be the quickest means for driving out independent ownership of medical practices and the quickest way of controlling (reducing) the complexity of patients that a physician accepts. The more complex the patient the more likelihood of a mistake, no matter how small or administratively mundane.

In our meeting the question was who/what would set the criteria for such fraudulent probabilities? Would criteria be set by this specific insurer? By the insurance industry, by Medicare/Medicaid or those holding federal office? Features could be identified and built into algorithms and customized over time. What defined an outlier billing event could shift, and such outliers could be categorized as suspicious. Such "suspicious" billings events could tilt towards criminalizing practices.

Thus, problem formation coming through the door could hold both, human benefit and harm. We discovered the negative impact of fraud detection at this time not because we were skeptical of health insurance companies but because specific experience was at hand that gave dimensionality to the problem they offered and scope to its downstream cascading possibilities. This expertise was domain specific but was central in making a decision for the entire organization. However, herein laid a problem regarding domain expertise which most typically was disregarded in machine learning start-ups.

DOMAIN EXPERTISE AND HIRING PRACTICES

When we considered problem formation it was best to examine expertise as it was being recruited into the organization that had the potential to shape such problems. Our recruitment process had cognitive, technical, behavioral, and problem-solving dimensions. All the tests on potential new hires were offered in a 'one-on-one' context with a third person taking notes. Criteria for ranking candidates was subjective and based on who conducted the test and on the testing criteria although the criteria itself was viewed as objective with the same problems and same questions asked of each candidate. Additionally, a percentage of candidate testing was partially conducted over the phone or Skype.

We recruited and tested for what we called T-shaped skills. We were hiring for depth of skills, which represented the vertical part of the T and ability to collaborate in cross-functional teams, which represented horizontal part of the T. Data scientists/engineers were all supposed to possess T-shaped skills.

Nevertheless, the overall hiring process ran against this T-shaped approach often identifying skills only useful for coding, and setting aside other skills as consultative, domain specific or non-T-shaped. This blind spot in our hiring was rarely acknowledged although we urgently needed skills focused on disease classification, high mortality cancer and a complex healthcare system, all outside of a typical coding experience and seen as "other" skills. These other skills were not T-shaped but rather singular, particular, walled-in domain knowledge. Broadly, we needed T-shaped radiologists, business leaders, nurses, patient advocates, IT managers and operational managers, we needed a T-shaped organization. Instead we hired data scientists. What this did to us and to our organization was to miss and not encourage people in roles who could translate forms of clinical and technical knowledge across product, business and algorithm development, we missed a translational T-shaped organizational culture.

How did this "T-shaped" hiring process that missed good cross-domain knowledge got expressed in the process of algorithm development?

Domain Knowledge: Universality and Particularity

We arrived at a company retreat by a northern California lakeside surrounded by redwood and pine forests. Swimming pool and cluster of cottages were set among a small gravel parking lot and hiking trails that ascended the hillsides. The retreat was called in somewhat a haste and before we knew, we were all gathered together to discover improved teamwork and sharing of ideas. All our activities were constructed to give everyone a voice, to ask questions and make suggestions on how we could improve the company. But a strange pattern emerged. As machine learning engineers and data scientists held more and more air time, business development, operations, a few clinical consultants had less and less time to make suggestions. Throughout the retreat the problem increased with less and less time allotted for cross-functional expertise outside of coding to offer opinions. The emerging organization was getting locked into its own limited self-perception.

On the last day of the retreat with only 20 mins left a colleague raised his hand and questioned why so few suggestions were asked or offered outside of engineering. The room got quiet. The CEO said "that's not true, everyone has been included." Then my colleague went further. I will call him Paul.

"You've been soliciting suggestions from everyone in your mind, but in this room it has been with only maybe half the company."

The CEO expressed disbelief: "no one's been excluded!" And then, "it's everyone's responsibility to speak up."

The actuality was different. When someone spoke up about problems or solutions outside of coding, the Q&A would move quickly on to the next person. The function of not having enough time to be democratic during a company retreat was not the point. It was a function of certain forms of knowledge that were recognized and other forms of knowledge left unrecognized or invisible. A group company retreat was in fact a fractured retreat of what one colleague said "ideas to be considered" and "ideas that were awesome," and "creative" and "productive". In this context the ideas that received the greatest attention concerned a narrow focus on machine learning algorithms, medicine as strictly a data problem at scale, machine learning expertise as universal knowledge, and team development and cohesiveness around these issues.

What was relegated to the "water cooler" conversations were ideas regarding lack of product direction, building and recruiting a clinical team with cross-functional responsibilities, leadership issues, a diverse organization at scale, work/life balance and low team morale.

Let us briefly unpack these ideas as they were expressed. The "awesome" ideas regarded machine learning as scalable, central to organizational development, around which all other forms of knowledge revolved. Outside this sphere clinical and business knowledge were consultative or being driven by algorithm development. Machine learning knowledge was the epistemological glue that bounded all other operational forms of technical and organizational mastery.

The second-tier ideas revolved around the organization and its people with their knowledge and experience. Disregarded clinical, cultural, retrospective feedback on mistakes and product direction could in fact act as a key to the emotional and professional bonds that reached inside and outside work-life balance. We believe the seeds of innovation were contained here, in this second tier of ideas that could not be well heard. The organization went on retreats and participated in team exercises but their contributions were not well captured nor well understood. The idea of what was universal and what was particular (domain-specific) knowledge were typically translated into concepts that were framed as "useful" or "not useful."

Read or Experience

The following conversation between a team member and a radiologist, and later the reporting of their conversation to the team may serve to expand on this universalizing-particularizing tension in our machine learning start-up.

"How do you see that [spiculation]?" - a colleague who we will call Tim.

"I compare it [nodule] to others I've seen" - the radiologist who we will call Don.

"It's what you have seen or what you know from what you've read?"

"Years [of] experience" - Don explains.

In our team meeting comprised of product and engineering this conversation was translated by Tim as a certain kind of data-driven opportunity for algorithm development.

"You guys, we can detect nodules by edge characteristics. Radiologists have diagnostics in their head and I think we can get the same diagnostics from imaging." Tim's comment became clearer and got reinforced in the following.

"So, we hire radiologists to annotate and we build, annotate and build, right" - says the CEO.

"Not that simple but that's the idea, we bring in domain experts."

"We plug them in or we hire them full time?

"I'm not sure we'll know what to do with them full time, the best would be consultants." Tim speculates.

Domain expertise has been established here as a contributory source of expertise. It provided a certain kind of fuel to algorithm development but was not valued as central to such development. It could be read and learned without having to be experienced. Radiology expertise was reduced to an "industrial designer: everything involved in a situation, any kind of material, must be calibrated for its contribution as evidence" stated anthropologist Marilyn Strathern (2004). The radiologist as a domain expert was seen as delivering a snap shot of knowledge and yet providing knowledge gained over years of experience of disease identification and classification. Except, that experience was disregarded as "useful." This artifact had to be used as evidence that could be codified into algorithm development. Such expertise was not seen as evolving and aspirational, it was seen "useful" only as evidence for creating clinical labels for algorithms. The team's excitement of using radiologists as consultants and not knowing what "to do with them full time" had reduced radiologists' value into a kind of automating piece work.

Contemporary radiology was often seen as a medical science of piece work. The radiologist has been well acquainted with holding a consultative role across medical disciplines with the widespread use of high-speed broadband and three decades of teleradiology. Medical images travel instantly, as radiologist interpretation travels remotely to hospitals and clinics, new imaging technologies continue to develop. Radiologists have been participants in a somewhat similar experience of reducing knowledge and their own knowledge to a commodity. Speed, distribution of radiology report and harsh efficiency have long attended upon radiologists. At first sight, it did seem to be a perfect collusion between algorithm development and teleradiology practice. However, the piece-working of radiology practice as a discipline into teleradiology practice has been different from its use as domain knowledge in a machine learning start up. The difference was radiologist's autonomy and agency to practice their craft that has always been held in high regard and to develop and perfect their skills has been an ongoing commitment. In an algorithm development context, the opportunity to perfect radiologist skills went one direction, toward algorithm building and scale, radiologist-consultants were plugged in and the race to gain new diagnostic insight has been transformed into a race to produce diagnostic evidence useful for algorithm iteration.

One of radiologists' key concerns at the turn of the 20th century was how to enable physicians to quickly reason with diagnostic insight while accompanying patients at bedside. One of the key medical concerns today has a twist: how radiologist's insight is to collaborate with/aid machine intelligence to make life saving diagnostic judgments across billions of data points in an instant.

ANALYTICAL EYE/I

In this section we wish to describe shifting role boundaries in a startup. As a Senior Data Scientist my role included tasks across disease classification, deep learning research, business development and data wrangling for model training. I often occupied a 'wear-all-hats' position in which there were more tasks than people and more roles than could be filled.

Trained both as data scientist and anthropologist the borders of my skills have never been settled. This meant that my ways of thinking and valuing people's insights were not fixed in terms of algorithm development. On one level psychically (a debate in my head) and outwardly (a debate with colleagues and product agendas), a kind of slippery battle waged across domains of knowledge that were valued among one group and not valued among another group. The value I placed on knowledge and expertise was always open for revision depending on the tribe I was working with. This led to occupying various roles even when my title stated a very bounded one. Carrying out various tasks, often contradictory ones within certain roles has been a professional tension that has persisted. For example, I often went from data scientist to product and partnership development taking back from partnership meetings client suggestions that were not recognized until only much later or too late. I caught myself speaking ethnographically instead of with the brevity of an engineer with a definable problem and outcome. Observational thinking was occasionally valued, but often not, and sometimes even scorned. The anthropologist's seeing with an eye towards levels of knowledge and a data scientist that must get "shit done" did not always rest easy together. Ethnographically stepping back and technically stepping into highly productive algorithm development was exciting but it could also be exhausting. One had to be elastic and thick skinned.

Authenticity was a sticky matter. I had to have constantly revisionary set of understandings on how data science, anthropology and machine learning specifically worked together and for whom. I lived and worked among revisionary sets of ideas that were never wholly laid out and never fully given over to my colleagues. Being authentic as a trained data scientist, deep learning researcher and anthropologist was not, however, a negative affair, it was one that had no user manual and no clear-cut boundaries. It did require summoning creative parts in oneself. It was a skeptical stance towards algorithms, and an enthusiastic stance towards their possibility.

Role boundaries and formal positions within our start-up could crash into product sprints, team meetings, and patients and physicians we were supposed to serve. This was our company's way of operating professionally. It was a specific way of working and acting upon knowledge that was cherished and knowledge that was left opaque. This was my direct experience of the organizing potentiality and growth of data science *thinking*, and not just its *application*.

THE CONCEPT OF THE "PATIENT"

One other role, the role of the patient, also eluded the makeup of our early start-up.

Patients as abstract beneficiaries, radiologists as domain experts, hospital administrators as users shaped our evolving and tenuous company culture. An embodied sense of what a patient was or who such a patient could be was often abstracted and would appear lost to busy organizational members. However, the patients that were to be served and the radiologists whose diagnostic skills were to be enhanced by algorithms were not really lost but instead were not shared among team members. 'One-on-one' walks could not get at shared understandings and would veer off into my or your medical story. There were many "patients" many "radiologists" many ideas of how algorithms would "enhance" the diagnostic encounter. The patient was always elusive and suggested a different kind of ethnography in order to render up these clinical-technical nuances that were embedded in such a machine learning organization, algorithm development and ideals and organizational members under intense pressure to build, scale and commercialize.

The problem of getting the patient in view was mixed with the problem of getting our models in view for productization. We remained far off from an actual product. We struggled with what kind of patients would benefit from our work, were these younger/older patients, smokers/non-smokers, male/female, those who only could afford sophisticated algorithms applied to their chest CTs or MRIs, were they local or in the developing world, were they near-term or years away from benefiting from our work?

A team member brought this confusion into relief when he read from a patient letter he had received. He was the one practicing MD on the team. The patient's letter was heartfelt and delivered to his home and he brought it in that day to bring through the door an embodied sense of a patient who had undergone both, good and poor diagnostic experience. He had been his patient. This is a paraphrase of the letter.

> Dear ______
>
> I wish to thank you for what you have done for me and my family. I'm not sure if you remember me but I was the difficult one who kept asking questions and you were the one doctor I could rely on during my time that tried to answer them. My wife said you were in and out most of the time but I felt you were at my side. I'm not sure how you found the tumor when others missed it, I guess I don't know how doctors cannot see things that can kill you and someone like you can see it. I really don't understand medicine but I know you were there and helped us and I'm grateful to you.
>
> I'm delivering this to your home because I wanted to make that effort to come to you as you must have come to me at my bed[side]. I think you saw me as a person not as another patient, I want to believe this and will hold onto it.
>
> Again I and my family wish you much happiness and success in the future and again thank you for your professionalism and care.
>
> Sincerely,
>
> ____________

This was delivered in a company-wide meeting. It was an intervention of sorts in an attempt to ground our efforts in an ethics of patient care and an embodied picture of an actual patient who was spared by good diagnostic care and who was grateful and alive to attest to their care. It was also a way to open up a discussion on the consequences of diagnostic error which our algorithms were to correct and reduce. Questions were asked about this patient but then we moved abruptly into an investor meeting. This quick move away from the reality of this letter was indicative of knowledge and experience that *could not* be taken in or absorbed. A few months later doctor colleague brought up this incident with a sense of shock that any notion of a patient just floated and could not be grounded in our push to build diagnostic algorithms that would help to save such person's lives. Between the care of algorithm development and diagnostic care of patients a huge gulf existed. The perceptual limitations of algorithm builders were well on display.

Upon reflection the awkwardness of the team's responses or non-responses to this patient's letter was of a different order. It conjured up many patients, many diseases, and many possible projects that brought out the ambition in team members to assist doctors in avoiding missing a critical diagnosis. The letter paradoxically, paralyzed and mobilized the team. Our team was not trained to take on and manage patient suffering, they were not trained to hold threats of mortality or the threat that mortality may come anytime around the corner by a missed diagnosis of an aneurysm or a missed adenocarcinoma (most common form of lung cancer). The team was not emotionally capable to both, take on the threat of diagnostic error, or to accept the overwhelming gratitude that comes from having another chance at life. The work of algorithm development and the work of giving more years to a person's life were at professional and experiential odds.

As a data scientist the teams I have worked with have fallen along a spectrum of well-integrated to chaotic. Our team was well-integrated at times and fell apart into chaos at others. Had I been an ethnographer coming in and out of the organization these moments may have been missed and I would have come away with a completely different picture of team dynamics. For example, after this letter was delivered we floated an idea of a "patient committee" inclusive of patients who had experienced lung cancer and who perhaps had a missed or under diagnosis. Everyone had a different idea of what a patient was, who a patient was and what kind of patients we could recruit. The topic of patient insights being integrated into algorithm development eventually went by the wayside and were reduced to domain knowledge, sent to the periphery of the organization and put on the shelf as patient consultants, when needed. The conception of the patient was marginalized even when it was emotionally charged and required for a deeper understanding of doctor-patient interaction, diagnostic product workflow development and understanding of a medical professional's insight and error. Even when we clearly needed to hear the voices of patients as part of our R&D, somehow we could not accommodate those voices.

Our operations manager put it well: "we all have been patients but we can't imagine their needs when we're here. It's like we stop thinking of ourselves as whole people, here we are only parts of our selves. It's strange, my father died of lung cancer, you have extensive experience in the field but the team has a hard time mapping these experiences to their work." I asked her what she was really getting at.

"It's hard to build what we're building [algorithms] on pain and suffering of others."

I reminded her that we were building algorithms to avoid such suffering through enhancing radiologist perception, to save lives. The patient's letter was a "success story." She shook her head as if to say, *that* was only part of the story when we were outside the organization, on retreat, but not accepted inside the organization where the "real" work got done. Emotions and patients with all their suffering and pain were messy, algorithms were to protect us from that messiness.

Revisiting Geertz and Levi-Strauss

One of the key concerns of Geertz and Levi-Strauss was how knowledge traveled, was taken up and applied to reveal everyday forms of life including human potential and its limits. The hesitancy-embrace of the computational could be seen in this light as a tension in realizing human potential to see, categorize and celebrate every day cultural forms and cultural others inside ourselves and outside in organizational life. It was on one level a debate on the uses of domain expertise of their era verses the uses of universalizing cybernetic systems. On another level it suggested a contemporary tension of universalizing and particularizing forms of knowledge that must live side by side in practice, and in our case, inside an early stage machine learning start-up. Computerized data is not only organized into training data for machine learning algorithms but is *organizing* of resources and people with all their complexities. Such data also organizes *possibilities* between computerized agents that "think" and "act" in the space between medical problem formation, forms of knowledge and hiring, and algorithm testing and validation. This universalizing-particularizing epistemic adjacency if embodied and used within contemporary organizations would expand upon and not foreclose upon possibilities of algorithm development.

Levi-Strauss and Geertz were not struggling over the present but rather struggling over a certain kind of everyday future in which human possibilities were circumscribed and discoverable by thinking machines. For Levi-Strauss social life was composed of "universal laws", for Geertz social life was brought forth by the freedom of mind of the ethnographer that had to negotiate intelligent agents that had potentiality to act, feel and displace the knowledge of the ethnographer. Levi-Strauss looked for universal patterns in everyday encounters, Geertz looked for hidden meaning in everyday encounters that was always being found, taken-up again and contested by local inhabitants. As often as they disagreed they both were looking for patterns and universal attributes grounded in small hard-to-access local forms of human life. They shared a focus on making visible possibilities of humanly discoverable evidence in field research, which held different challenges and different opportunities. How local knowledge was shared and how it moved among and between local inhabitants and among ethnographers was crucial to their notion of other forms of thinking and behavior. At a base level they were interested in the ethnographic mind in the context of automation and universalizing computational systems, sometimes posed as a threat, sometimes an embrace, always rubbing up against each other but never losing sight of a human who feels.

CONCLUSION: PUZZLING THROUGH EPISTEMIC EDGES IN ALGORITHM R&D

We have examined dimensions of problem formation in data science and the splits in certain forms of expertise and knowledge that could have practical implications for the shaping of an early stage machine learning start-up. When we thought about the procedures for defining the borders and scope of our case study we became increasingly aware of the shifting role boundaries, problems in-process, and people and their expertise being marginalized and yet continuing to push ahead with algorithm development.

A case study that was dynamic with multiple lines of flight took us back again and again to puzzling through the human messiness of data science knowledge and what got put outside or was made as other in machine learning knowledge today and in the foreseeable future. Machine learning knowledge did not travel alone without being enmeshed in people, organizations, bodies, data and patients, and caught up in certain denials and acknowledgments of human pain and suffering. In other words, we acknowledged the observable tensions and looked for ways to advocate for silences lurking underneath the obvious.

Methodologically, capturing algorithms required daily ethnographic note taking, mundane but consistent participation, and an interpretative stance similar to Geertz's that allowed one to depict the speech, gestures and uncertainties that came from everyday organizational challenges. We suggest, such analysis drew from a knowledge in data science at least an appreciation of how data was not just wrangled, handled or acquired but was *organizing*. To work with computerized data (in our case healthcare data and medical image data) occasioned being organized by its constraints, possibilities and by strategic partnerships as we have seen with our health insurance company focused on fraud detection. Such consistent participation in the life of an early stage start-up allowed us to gain a larger sense of the textures of data and human patterns similar to the spirit of Levi-Strauss that took us out of the particular and into a generalizing conception through algorithm development.

As we have seen, machine learning could be considered universal viewed with cross-industry application irrespective of specific expertise that has shaped those industries. In our case, a universalizing dimension was applied to algorithms, a domain specific boundedness was applied to the content creators of those industries. Thus, one of the negative outcomes we encountered was epistemic splitting, instead of epistemic inclusiveness, parsing knowledge into plug-ins and universalizing machine learning concepts above other forms of knowledge that weakened organization and compromised quality and flow of ideas that ran through algorithm development. Splitting knowledge into bits also split people and eroded organizational bonds that could have forged professional curiosity across disciplines. Domain expertise suggested a depth of knowledge in a particular field, however it might have also suggested an ongoing attempt to be pollinated by and extended into other areas of expertise. It had a potential to be a truly T-shaped enterprise.

A Note on AI-*Inflected* Ethnography

The idea of an AI-*Inflected* ethnography emerged at this time. It came together for us as a way to bring qualitative analysis to silences and to the downstream impact of algorithm development at the moment of conception and problem formation.

We believe an AI-*Inflected* ethnography has a potential to focus on the silences, intentionalities, gaps, aspirations and conceptions that may be in one moment accepted and in another moment forgotten or rendered invisible. An AI-*Inflected* ethnography we suggest, is not about what has been built but what has not yet been built and the reasoning or emotions that go into and are fought over to arrive at the product or algorithmic problem worthy of engineering time and worthy of gaining sometimes expensive data resources.

When we think about AI-*Inflected* ethnography we should not conjure pictures of methodological reasoning. Perhaps such reasoning will come over time and over further case studies but here we are marking active algorithm development today and its implications on people in a start-up that struggled and often failed to inhabit different points of view. An "*inflected*" ethnography focused on AI development does not offer up suggestions for research tools, recording procedures or discussion on synchronous (real-time) or asynchronous (non-real-time) of research data capture and technique in the field. Instead we have chosen to examine dimensions of problem formation in data science and epistemic splits in certain forms of expertise that have had practical implications and anxieties within an early stage machine learning start-up.

This type of analysis we believe is best located in moments when we can tease out the collaborative opportunities and imagination in algorithm development in an everyday when diversity of thought may be up for grabs and when algorithm problem formation may be held open. This form of analysis is unstable and in some ways an outcome of role murkiness between data scientist and ethnographer. But such analytical instability comes as a benefit. It is at these formative moments when certain epistemic events can be made visible through gaps in idea generation.

One of the strengths of *AI-Inflected* ethnography is paying close attention not only to the content of knowledge, but also to the processes of knowledge development focused on shifting edges of algorithm R&D. These edges of organizational knowledge and practice are not simply learned, known and then applied ideas but are locally contested and puzzled through. Deciphering epistemic borders suggests that everyone in an early stage start-up is engaged on one level or another with processes of the formation, co-creation and development of algorithmic knowledge and outcome even when these outcomes are very uncertain and distant. The process of building algorithms paradoxically engages and silences organizational members. The epistemological glue that represents machine learning knowledge does not draw together all forms of knowledge as domain knowledge or as a binary of usefulness/non-usefulness. Instead this glue provides a kind of organizational coherence as much as it limits the flow of forms of knowledge and fashions types of algorithms and types of organizations. These inflection points in AI development are tensions and realities not to fend off and avoid but to identify and transform.

It is important to keep in mind that what characterizes an *AI-Inflected* ethnograph*y* is not consultative attendings but an organizational embeddedness of an ethnographer/researcher that helps capture gaps and slippages within an organizational environment operating as a kind of disjointed body pulling together its internal fortitude to take on and build algorithms that can be integrated into software and into people's lives. The odds of early stage start-up success are daunting, the odds of their failure are well over 90%.
Could these odds be improved upon with a true T-shaped organization and a true diversity of perspective?

Where Do We Go From Here? – What we gained access to was a deeper image of an organizational group of people trying to genuinely integrate a patient's 'thank you' letter into their thinking but unable to seize upon its message of gratitude and transform this message into actionable algorithm development. We gained access to the potential consequences of radiologist's insight as consultative input and domain knowledge. We gained a sense that the generalizing and particularizing aspects of machine learning were open and unsettled. There were forms of knowledge and emotional life that algorithm development resisted. We believe there may have been other forms of knowledge and diversity of knowledge that has not yet been organizationally imagined to produce very different algorithms for very different outcomes. We do not yet know how these forms will appear in the future but we can as researchers prepare for their emergence by laying the groundwork for the conceptualizing of more malleable and inclusive algorithms.

One of the most disembodied technologies in development today requires embodied and embedded research and engagement. AI calls for us as researchers to not only step back into the social but to step into the daily grind of AI's silences and epistemic threads that are constantly being shredded and mended to transform organizational coherence.

If we use algorithm development to screen out the pain and suffering of vulnerable people like patients, instead of finding ways to integrate their journeys, then we may also find ever-narrowing algorithm development populating or even taking over our everyday lives.

It is an extensive kind of labor to evaluate, question, chart, chronicle how a start-up and AI development together forge forms of thinking and acting across algorithm, people, product and clients. It seems to be a worthy endeavor. However, if this kind of research needs to get done who is going to do it? Who will listen and make use of it? What kind of organization will it produce? Most importantly, what kind of commitment will it take to produce it?

Rodney Sappington Ph.D. is a data scientist and ethnographer who leads machine learning strategy and product development. His research focuses on the intersections of knowledge, human-machine relations and behavior, and societal outcomes from algorithm development. Rodney is CEO and Founder of Acesio Inc. *rodney@acesio.org*

Laima Serksnyte Ph.D. has 15+ years of experience in advanced research, consulting and coaching in the areas of executive leadership and organizational psychodynamics. Her research focuses on psycho-social, behavioral and societal levels of analysis in areas of AI development, healthcare, education and organizational member advancement. Laima is Head of Behavioral and Organizational Research at Acesio Inc. *laima@acesio.org*

NOTES

Acknowledgements – Thank you to the reviewers and curators of EPIC, especially to Dawn Nafus for her continued insights and feedback. A big thank you to colleagues in applied deep learning and data science for careful consideration of the concepts in this paper.

1. A "unicorn" company is a start-up that reaches $1B in revenue.

2. "Moonshots" are typically defined as ambitious and aspirational projects and companies reaching to develop bold and future-oriented products and services.

3. As a cautionary note, our educational and psychological systems have to keep up with AI projected development otherwise fall behind and in disuse, creating an elite group or risk a population whose knowledge is woefully behind, not relevant, or worse considered a danger to AI development. Bladerunner and many other dystopian societal images come to mind.

4. Claims (insurer) transactional data can have multiple uses in machine learning. Claims data is standardized and medically coded and covers a wide area of the patient's journey. For example, prescription and behavioral trends can be captured across pain medications (opioids) and cholesterol lowering drugs (statins) and this same data can also be used to track physician practice claims, types of claims and time points when outlier claims could indicate probability for fraud.

5. Spiculated margins of a lung nodule are uneven edges that can indicate a higher risk of cancer. "Most nodules that are not cancer have very smooth or rounded margins or look like several rounded nodules together (also called "lobulated"). See Lung Cancer Alliances explanation "Understanding Lung Nodules: A Guide for the Patient."
https://lungcanceralliance.org/wpcontent/uploads/2017/09/Understanding_Lung_Nodules_Brochure_dig.pdf

REFERENCES CITED

Dean, Jeff
2018 Deep Learning to Build Intelligent Systems by Jeff Dean. AI NEXTCon Silicon Valley 2018, April 10-13, Santa Clara.
https://www.youtube.com/watch?v=4skC-kGCYRA

Dean, Jeff
2016 The head of Google's Brain team is more worried about the lack of diversity in artificial intelligence than an AI apocalypse. *Recode.* August 13, 2016.
https://www.recode.net/2016/8/13/12467506/google-brain-jeff-dean-ama-reddit-artificial-intelligence-robot-takeover

Faggella, Daniel
2018 Valuing the Artificial Intelligence Market, Graphs and Predictions. *Techemergence.* Updated September 16, 2018
https://www.techemergence.com/valuing-the-artificial-intelligence-market-graphs-and-predictions/

Ferrary, Michel and Mark Granovetter
2009 The role of venture capital firms in Silicon Valley's complex innovation network. In *Economy and Society.* 38:2, 326-359.
Geertz, Clifford
1973. *The interpretation of cultures.* New York: Basic Books.

Garling, Caleb
2015 Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines. *Wired.* May 5, 2015.
https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines

Glasner, Joanna
2018 The Unicorns Behind San Francisco's Burdensome Startup Success. *Crunchbase News.* April 18, 2018
https://news.crunchbase.com/news/unicorns-behind-san-franciscos-burdensome-startup-success/

Koulopoulos, Thomas
2018 According to Peter Diamandis and Ray Kurzweil, These Are the Most Dangerous and Disruptive Ideas. *Inc.* January 19, 2018
https://www.inc.com/thomas-koulopoulos/according-to-peter-diamandis-ray-kurtzweil-this-is-how-you-build-future.html

Levi-Strauss, Claude
1953 *An Appraisal of Anthropology Today.* Eds., Sol Tax and others. Chicago, IL: University of Chicago Press.

Ng, Andrew
2018 AI is the new electricity. *Startup Grind.* March 6, 2018.
https://www.youtube.com/watch?v=fgbBtnCvcDI

Quoc V. Le, Marc Aurelio Ranzato, Andrew Ng, Jeff Dean et al
2012 Building High-level Features Using Large Scale Unsupervised Learning
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/38115.pdf

Ruiz, Rebecca R.
2017 U.S. Charges 412, Including Doctors, in $1.3 Billion Health Fraud. *New York Times.* July 13, 2017
https://www.nytimes.com/2017/07/13/us/politics/health-care-fraud.html

Strathern, Marilyn
2004 The Whole Person And Its Artifacts. *Annual Review of Anthropology.* 2004. 33: 1-19

How Modes of Myth-Making Affect the Particulars of DS/ML Adoption in Industry

EMANUEL MOSS
CUNY Graduate Center / Data & Society

FRIEDERIKE SCHÜÜR
Cloudera Fast Forward Labs

The successes of technology companies that rely on data to drive their business hints at the potential of data science and machine learning (DS/ML) to reshape the corporate world. However, despite the headway made by a few notable titans (e.g., Google, Amazon, Apple) and upstarts, the advances that are advertised around DS/ML have yet to be realized on a broader basis. The authors examine the tension between the spectacular image of DS/ML and the realities of applying the latest DS/ML techniques to solve industry problems. The authors discern two distinct ways, or modes, of thinking about DS/ML woven into current marketing and hype. One mode focuses on the ***spectacular capabilities*** *of DS/ML. It expresses itself through one-off, easy-to-grasp marketable projects, such as DeepMind's AlphaGo (Zero). The other mode focuses on DS/ML's* ***potential to transform*** *industry. Hampered by an emphasis on tremendous but as of yet unrealized potential, it markets itself through comparison, in particular the introduction and adoption of electricity. To the former, data is a mere ingredient, a current, but not a necessary, requirement for the training of smart machines. To the latter, data is a fundamental enabler, a digital, always-giving resource. The authors draw on their own experiences as a data scientist and cultural anthropologist working within industry to study the impact of these modes of thinking on the adoption of DS/ML and the realization of its promise. They discuss one client engagement to highlight the consequences of each mode, and the challenges of communicating across modes.*

INTRODUCTION

In a midtown Manhattan conference room, the audience is nodding along to the presenter's slides. Artificial intelligence seems so accomplished and yet so straightforward, from Google DeepMind's Go-playing AI agent AlphaGo (and successors) and Carnegie Mellon's poker-playing Liberatus AI to Sunspring, a short film based on a movie script, replete with dialogue and stage directions, that was written by a neural network. Let two computers play Go against each other and let them learn from their mistakes until they get better than human Go grandmasters. Feed a neural network with movie scripts until it writes one of its own. The artificial intelligence industry has long been adept at foregrounding the "magic" of AI systems (Elish & boyd 2017). On that day, the audience in the conference room was comprised of employees from an entertainment media company who were identified prior to the event as key stakeholders in how the company collected, analyzed, and utilized data across the many lines of their business. They were interested in using data science and machine learning (DS/ML) for their organization and had sought the help of DS/ML experts to do so. Specifically, they were interested in a "data strategy", a set of project, people, and process recommendations designed to help them harness the potential of

DS/ML. The day began with a recognition of the many magical things that data science and machine learning (DS/ML) is capable of. Over the course of the day, conversations shifted towards practical applications of DS/ML and the conditions that allow DS/ML to succeed within organizations.

This paper grapples with the ways in which the contrasting narratives that surround the development of data science, machine learning, and artificial intelligence present different, at times seemingly opposed, paths forward as enterprises develop strategies and make tactical decisions around these emerging technologies. We identify and examine two contrasting narratives for the emergence and development of these technologies. In analyzing the myths and metaphors that attend to the discursive production of DS/ML, we follow Sturken and Thomas (2004), who observe that "metaphors about computers and the Internet are constitutive; they determine how these technologies are used, how they are understood and imagined, and the impact they have on contemporary society". So too do these metaphors determine how businesses strategize around DS/ML.

STORIES AND MYTH-MAKING

The history of technological development is populated with spectacular demonstrations designed to hasten the development of and increase public demand for new products. The spectacular demonstrations of electrical lightning at World Fairs, Centennials, and other grand exhibitions of the late 19th century were designed to increase consumer adoption of the light bulb and serve as a (literally) shining proponent of the potential uses of electricity (Nye, 1994). Similarly, prominent players within DS/ML build and promote spectacular demonstration projects. These demonstration projects take a range of forms; they highlight an emerging capability (e.g., the capability of generating new text; natural language generation) or are engaging in ways that generate press coverage (as when a machine defeats a human expert). These demonstrations serve a range of functions; they establish their producers as serious players in the industry, they promote existing products and services offered under the same brand, they burnish the resumes of those who work on them. Primarily, they perpetuate excitement, and investment, in DS/ML.

AlphaGo Zero and the Modular Myth

AlphaGo (Silver et al. 2016) and its successor AlphaGo Zero (Silver et al. 2017) are algorithmic systems built by Google's DeepMind that spectacularly defeated reigning human world champions of the board game Go. Go is a complex game, with millions of possible moves and billions of possible board configurations. In their release notes of AlphaGo, DeepMind foregrounded the complexity of the game itself and the remarkable achievement of building an agent that can learn to cope with that complexity from human gameplay data. In announcing AlphaGo Zero, DeepMind's promotional materials foregrounded the ability of the algorithmic system to learn from self-play: AlphaGo Zero learns on its own by playing against itself. In the process, it learns strategies that resemble strategies of human Go players, as well as a few novel others (Silver et al., 2017). The accomplishments of AlphaGo and AlphaGo Zero appear as evidence that AlphaGo must be very intelligent since Go is commonly understood as a complex game that only the most intelligent humans can learn to play well. And while AlphaGo still needed some human help, in form of human game playing data, AlphaGo Zero freed itself from this requirement, this dependence on human expertise and labor.

The Conditions of Success for AlphaGo (Zero)– Under scrutiny, we discover that, as Andrej Karpathy put it, “AlphaGo is a narrow AI system that can play Go and that’s it” (Karpathy 2017); the success of AlphaGo is grounded in several conditions or “conveniences” of the game Go (see Table 1).

Table 1. Conveniences of Go (adapted from Andrej Karpathy [2017])

Deterministic	The rules of Go describe possible game states without any randomness or noise.
Fully Observed	Each participant knows everything about the current state of the game simply by looking at the board.
Allows only discrete actions	There are a quantifiable number of different moves that are possible without gradations between these moves.
Is simulatable	It is easy to simulate a game of Go and this simulation will be identical to the game itself.
Is short	Each game of Go lasts approximately 200 moves.
Has a clear outcome	There is a clear definition of what constitutes a ‘win’ or ‘loss’.
Is well-documented	There are hundreds of examples of human gameplay to supercharge the initial knowledge that AlphaGo begins learning from (AlphaGo Zero, of course, freed itself from this condition).

Few ‘real-world’ problems, problems that one is likely to encounter in industry, share these conveniences with the game Go. Real-world problems are full of imperfect information, vaguely defined in terms of a success metric, rare enough that trainable examples are hard to come by, or they involve continuous phenomena rather than discrete moves that allow for gradations between possible states. Arguably, most real-world problems are more complex than the game of Go (see also, Elish & boyd 2017); in DeepMind’s promotional material and the paper detailing the algorithms that power AlphaGo (Silver et al., 2016) and AlphaGo Zero (Silver et al., 2017) complexity is defined in terms of combinatorics, the number of possible board configurations, a narrow definition of complexity. Furthermore, most real-world problems are only simulatable through deliberate decisions about what is and is not part of a system that do not come close enough to approximating reality to be good representations of the problem at hand. The weather does not affect the outcome of a game of Go, yet it is likely to be relevant for algorithms that steer self-driving cars; most real-world problems tie into dynamics part of the world that require us to make decisions about what is relevant and what is not when we model the system. As Karpathy concludes his analysis of AlphaGo (and AlphaGo Zero), it demonstrates not so much the power of DS/ML, but rather shrewdness in choosing a tractable yet impressive problem, as well as the power of Google to devote its resources to create a system that can tackle such a difficult, if singular, problem (Karpathy 2017).

AlphaGo and AlphaGo Zero are only two of the more recent spectacular projects, of course, that demonstrate the dramatic promise of DS/ML. Carnegie Mellon's poker-playing Liberatus AI beat humans in games of Texas Hold 'Em, for example.

The Modular, Bolt-On View of DS/ML – Spectacular demonstrations of emerging DS/ML capabilities solve isolated and isolatable challenges without recognition of the conditions critical to the success of these demonstration. They encourage a modular, bolt-on view of DS/ML. This view encourages us to see DS/ML as an add-on with high interoperability: conditions do not matter. The view suggests that DS/ML can be added to existing software without reconfiguration; existing processes can be complemented, or augmented, by DS/ML without transformation. The modular, bolt-on view of DS/ML suggests that DS/ML can be deployed as a layer that sits atop or replaces existing products and processes, as neatly as desktop word processors seemed to replace typewriters.

But, even in the transition from typewriter to word processor, problems of translation required halting and stepwise adjustments from one to the other. The dot-matrix printer, TrueType font libraries, and skeuomorphic user interfaces (Laurel 2013) all filled in the gaps between how people had designed their work processes around the typewriter and the new, unique affordances of desktop publishing. The modular bolt-on view of DS/ML draws attention away from the very particular conditions that enable DS/ML successes as well as the expertise and labor that is required to conceive of, develop, and refine these technologies over time and often over many rounds of trial and error; this is at the heart of what we call the modular, bolt-on view of DS/ML. Spectacular demonstrations have broad appeal because of the human tendency to misunderstand what constitutes a computationally difficult problem, to see proof of technological capability as proof of pragmatic capability. Arguably, this human tendency is exploited in the choice of demonstration project to achieve reach and effect. AlphaGo Zero, for example, is a dramatic proof of concept of reinforcement learning (Silver et al. 2017). But, it is not a persuasive proof that reinforcement learning can accomplish tasks we as humans understand to be on the same order of complexity as the game of Go. Indeed, human intuitions for what are easy or difficult problems to solve do not map on to computational difficulty. This problem has long bedeviled AI researchers who have struggled to explain, for example, just how hard comprehension tasks are for computers, when they seem so 'easy' to humans. This human tendency to misunderstand what constitutes a computationally difficult problem allows spectacular demonstrations, like AlphaGo Zero, to recast a range of problems that seem 'easier' from the perspective of human intelligence as within grasp of being solved by the technologies that are showcased by the demonstration project.

DS/ML as the New Electricity

Because of the prominence of DS/ML as modular and bolt-on, exceptions to this narrative are worth examining. One such narrative that stands in exception to the modular addition of capabilities story is told by Andrew Ng, formerly of Baidu, now Co-Chairman of Coursera and an Adjunct Professor at Stanford University. He describes the challenges of adopting DS/ML as an emergent technology in terms of the challenges that faced industry around the turn of the 20th Century as the emergent technology of electricity began to replace steam power. At that time, electricity was far from the omnipresent and almost invisible commodity that it is today. Few aspects of the technology had been standardized, from voltages to outlet plug shapes, and ensuring that a new investment in electrification would pay off was far from certain. In Ng's telling, *"a hundred years ago, electricity was really complicated. You had to choose between AC and DC power, different voltages, different levels of reliability, pricing, and so on. And it was hard to figure out how to use electricity: should you focus on building electric lights? Or replace your gas turbine with an electric motor?" According to Andrew Ng, "thus many companies hired a VP of Electricity"* (Ng 2016).

Similarly, DS/ML is, today, "really complicated". Data can be local or distributed in the cloud. It is difficult to know whether and why to use a random forest algorithm or a neural network, or how to evaluate the success of any particular implementation. Furthermore, it is difficult to anticipate the costs of a project; the reliability and cost of machines, data storage, and engineering talent vary widely. And it is difficult to know where to focus one's efforts; should one build an audience segmentation model first or a churn model?

While most commentators gloss Ng's story under breathless headlines like "Artificial Intelligence Is the New Electricity?" (Eckert 2016), the story that Ng tells is more nuanced than one of simple metaphor-making when one focuses on the importance of the "VP of Electricity" to Ng's narrative. It is also more nuanced than the modular addition of capabilities. Through this lens, his story is one of complexity in emerging technologies that requires dedicated expertise to construct new interfaces that mediate between the different needs of the different parts of an organization. For DS/ML, this means preparing data in a way that is easily ingestible, and constructing tools that simplify the underlying complexity but offer affordances for making use of tools that had been previously beyond the reach of non-experts. AlphaGo Zero is presented by its creators as a persuasive proof of reinforcement learning and its capacity to solve 'complex' problems.However, reinforcement learning works best on problems that have been adequately abstracted to sufficiently resemble the kind of closed problems that reinforcement learning can solve. That is, real-world problems must be made sufficiently deterministic, observable, discrete, simulatable, short, evaluable, and well-documented before they can be addressed by the emergent technologies embedded in AlphaGo Zero, as Karpathy points out above. Furthermore, these emergent technologies must also be reshaped to accommodate real-world problems, even in their abstracted conditions, as inputs. This work of abstraction and accommodation is drastically different than the work of software development attuned to understanding DS/ML as the modular addition of capabilities. And yet, most promotional materials for DS/ML tend to further this narrative, evoking the sense of magic that Elish and boyd identify in their work (2017). In business settings, these narratives fulfil specific functions; modular capabilities are easier to sell as products, and they are easier to explain to customers as discrete technologies. Furthermore, they lend themselves to the very same spectacular demonstrations that we have discussed above. These spectacular demonstrations are countered by Ng's metaphor, which argues that until DS/ML can be utilized as easily as a lamp can be plugged into a standard wall outlet, a dedicated form of expertise will be required to make it have any value for an organization.

CONSEQUENCES FOR PATTERNS OF ADOPTION

Metaphors matter, they guide adoption of emerging technology (Sturken & Thomas, 2004). And, they shaped how the audience of stakeholders communicated with the DS/ML expert consultants that had gathered in that Midtown Manhattan conference room. Those stakeholders and expert consultants were gathered to develop and implement a data strategy (see above) for the entertainment media client. The goal of a data strategy is to help companies realize the potential, and potential value, of their data for their organization. Recently, over the past couple years or so, companies have started offering consulting services to help craft such data strategies, responding to a need in the market thereby acknowledging the difficulty of translating the spectacular successes of DS/ML into industry applications, from Amazon's ML Solutions Lab [1] and Google's ML Advanced Solutions Lab [2] to the startup Element AI [3] (to mention the more prominent players).

As a producer of original content, from written text to short-form video, for a variety of different audiences, the client was interested in natural language generation, from de novo generation of content, from text to video, to the automatic tailoring of existing content to appeal to different sets audiences. In addition, they were interested in internal-facing chat bots to increase operational efficiency (e.g., a bot that suggests to re-publish existing content). This set of projects, while feasible, suggests a modular view of DS/ML. Furthermore, in preliminary meetings prior to the workshop, there was little to no concern for the conditions that allow DS/ML to succeed within organizations, yet another hallmark of the modular view of DS ML.

Over the course of the day, the consultants met with business stakeholders, content creators, software and data engineers, and data analysts. They started the onsite with a presentation on data science, machine learning, and artificial intelligence (AI) designed, one the one hand, to define a common language, and one the other, to set realistic expectations for what can be accomplished with DS/ML and the work it takes to achieve these possible accomplishments. In particular, the presentation was designed to create awareness for the conditions that need to be created for a long-term, successful, in-house DS/ML practice that could develop text generation algorithms and smart bots for internal efficiency.

Throughout this paper, we use the following definitions of data analytics, data science, machine learning, and artificial intelligence. We do not claim that our definitions are better than their alternatives, there are many competing definitions, in part because definitions suggest and drive the particulars of the adoption of new technology. Our definition of AI, for example, sidesteps a thorny issue (the definition of "intelligence", which is highly political). Here, we merely define our use of terms for the purposes of this paper to avoid confusion.

The Destruction of Modular Myths

The presentation defined, first, terms such as data analytics, data science, machine learning, and AI. There is a lot of confusion about these terms, in part, because their definition is shifting. Looking to attract talent, companies have started rebranding their data analyst positions as data science positions, for example. Artificial Intelligence is a particularly confusing term; founded as an academic discipline in the 1950s, it has been rebranded several times over the past decades with an emphasis on goals (mimicking intelligence

Table 2. Definition of Terms

Data analytics	Data analytics is the craft of counting. Data analysts count “daily active users”, for example, to inform the business about its performance. In doing so, they make use of descriptive statistics (medians, means, variation, etc.).
Data science	Data science is the craft of making predictions using and surfacing patterns from data. Data scientists use machine learning, from supervised (e.g., classification) to unsupervised (e.g., clustering) techniques in addition to descriptive statistics.
Machine learning	Machine learning is a set of tools, from supervised (e.g., classification) to unsupervised (e.g., clustering) techniques including techniques such as deep learning and reinforcement learning.
Artificial intelligence	Artificial intelligence denotes a set of capabilities or behaviors, from object recognition to goal-oriented decision making to (natural, human) language understanding and generation, that appear, to an observer, to demonstrate some kind of intelligence. Generally, these capabilities are displayed, and behaviors performed, by systems that take a set of inputs and produce outputs guided by internal states, a kind of memory.

behavior), tools (machine learning, logic, etc), or as what is just outside the grasp of current technological capability. The consultants introduced data analytics as “the craft of counting”, data science as “the craft of making predictions using data and surfacing patterns from data”, and machine learning as “a set of tools used by data scientists” to yield insights and to contribute to products that may display (seemingly) smart behavior. To the client, they suggested to leave artificial intelligence out of the day’s conversations, to focus on data analytics, data science, and machine learning, to limit potential for confusion (see Table 2).

The consultants proceeded with a review of popular, celebrated accomplishments in the field of machine learning including AlphaGo, AlphaGo Zero, Sunspring, etc.. The presentation was designed to refer to accomplishments in the field that some audience members may have heard about to first introduce the reasons why there is much excitement in the field of DS/ML. Second, they introduced these examples to then explain, at the high level, the technologies that enabled these feats, the limitations of these technologies, and the conditions they need to work (seamlessly). In Sunspring, for example, many of the protagonists express lack of knowing: *“I don’t know.”, “I don’t know what you’re talking about.”, “What do you mean?”.* The consultants explained how the algorithms that underlie the Sunspring movie script, written by a computer trained on movie scripts, led to these kind of patterns. The key intention of this was to highlight the conditions and circumstances that allow these algorithms to succeed, and consequently, the limits within which they can successfully operate: the consultants used the spectacular feats of DS/ML, and respectfully deconstructed them, to guide the client towards a view of DS/ML as an emergent capability that requires expertise to being into new business contexts. While the presentation was well received, it did not have the intended effect, as we found out later and discuss below.

Empowering Data Teams

The consultants talked to the clients in-house data analytics and data engineering team, two data analysts and one data engineer. They were joined by their current, interim manager (who did not have a data science background). In conversation, it became apparent that the data analytics team was overwhelmed by creating reports requested by the business or creatives on the performance of the business or content. Requests were fulfilled in an ad hoc manner, each one custom based on the specifics of the request. The data engineer worked on making data accessible where needed to satisfy requests. The data analysts were eager to develop self-serve approaches, dashboards that could communicate to the business performance metrics on demand, however, ad-hoc requests took priority and occupied the majority of their time: there was little to no time to build this functionality.

This situation is not uncommon. Data analytics teams tend to struggle to handle their workloads often due to the very specific nature of the requests they are asked to handle and short timelines. To remedy the situation, data analytics teams need to log and monitor incoming requests to identify common themes. They then can build self-serve dashboard for on-demand delivery of data insights around those common themes that will cover a range of frequently asked questions. In doing so, data analytics teams, tasked, due to their function, "to count" need to define "what to count"? They need to answer questions such as "What is a daily active user?" or "For how long does someone need to visit a website, watch a video, or interact with content to qualify as a content consumer?"

Within organizations, there tends to be a variety of definitions of terms such as daily active user or content consumer. Often, differences go unrecognized and unacknowledged. They surface when the data analytics team is tasked to count: they need to translate daily active user into a set of instructions (e.g., a SQL query) that demands specificity. Lacking specificity, data analysts tend to borrow details from their own, sometimes idiosyncratic, definitions of these terms. This practice has several consequences. First, asked to count daily active users, different data analysts tend to produce different answers. What is more, even the same data analyst may give different answers depending on the definition, if available, of daily active user passed on by the stakeholder. This discrepancy in answers is, at many companies, gradually eroding trust in data. Second, idiosyncratic definitions prevent data analysts to build on-demand, self-serve dashboards and other tools. At the extreme, every request becomes custom because every request demands a different way of counting a similar, often seemingly same, concept.

To remedy the situation, data analysts need to be empowered, in collaboration with the business, to define what to count. As they receive requests, they are in the best position to record definitions in current use and to consolidate definitions. To do so, they need to set aside time to work on recording requests and consolidating terms. Working with the client, the consultants received significant pushback to these suggestions, despite a clear opportunity to consolidate terms (it was suggested and requested by members of editorial and creative). According to the client, the core function of the data analytics team was to respond to ad hoc requests first, not to define or redefine them, and then, as time permits, to build on-demand, self-serve tools. There was lack of recognitions of the impossibility of accomplishing the latter task without a say in the consolidation of terms. Disempowered to establish the conditions for their own long-term success, the data analytics team was seen as a mere service function to the detriment of the organization.

Understanding DS/ML as modular, bolt-on solutions de-emphasize the importance of data readiness and interferes with deriving value from data for increased efficiency or novel products. By contrast, understanding DS/ML as an emergent capability emphasizes that a robust in-house data analytics capability is the foundation for successful in-house DS/ML projects and products; it portrays data as a resource. Like any resource, data needs to be harvested and managed. Data analysts interact with the data, build an understanding of the data, in counting they establish concepts, such as daily active user, that find use often as labels in the predictive algorithms of data scientists and machine learning engineers: e.g., the success of a piece of content may be measured in how many content consumers it attracted. In the DS/ML as modular mode, the client, surprised by our suggestions, rejected them. Coming from the DS/ML as emergent technology mode, we were surprised by the client reaction. Each mode leads to a different set of expectations, suggestions, and ultimately strategy.

Successful Data Science Requires Data Analytics –The client company had let go of their only data scientist a couple month prior to our engagement after a tenure of less than one year; the data scientist had failed to make an impact. The client's failures in data science is grounded in the their approach to data analytics. Without a robust analytics function, data science cannot succeed. Data science depends on definitions of what to count as well as data quality and access. Lacking data analytics, data science roles tend to morph into either data analytics roles, the data scientist helps fulfill ad hoc stakeholder requests or does data quality assessment, or helps build the pipelines for better data quality and access. Thes patterns can be exacerbated by lack of a clear distinction between a data science and a data analytics team, as was the case at the client company. Without a robust data analytics function, data science cannot succeed. In such situations, data scientists tend to leave or, as the more expensive members of the data team, are asked to leave, as happened in this case. Understanding DS/ML as modular deemphasizes data readiness, prevents data science from having an impact within organizations; it does not highlight, much less create, the conditions for successful data science within organizations.

Protection of Editorial and Creatives

In the run-up to the engagement, the consultants were advised by the technology and product side of the business to "tread lightly" so as not to upset editorial and creatives who may fear about changes in or loosing their job. During conversations, they found editorial and creatives to be eager to hear about our work, solutions, and possible externally or internally facing data products; they freely talked about their work. They encountered healthy skepticism, not fear. In many ways, editorial and creatives were more receptive to our suggestions and eager for adoption than the technology and product side of the business.

Viewed as spectacle, DS/ML offers modular, bolt-on solutions to add new products or business functions or to replace existing ones; it promotes self-sufficient, mostly autonomous systems. It de-emphasizes the importance of conditions and context. It de-emphasizes the importance of data readiness and the contributions to data and data readiness by people across the organization, from the data analytics team to editorial and content creators. It paints a picture of users as collaborators with machines but on the machine's terms. Workers are to assist the machines, to be tasked with the edge cases that machines can't handle, providing the glue between the complex work environment and its simplified version that allows machines to succeed. This view fosters fear of replacement by machines; AlphaGo pitted the machine against the human. AlphaGo Zero excluded humans from training machines.

Viewed as a an emergent capability akin to the emergence of electricity, DS/ML is a potential, fueled by its resource: data. It not only emphasizes the need to harvest and manage this resource, it encourages us to think of applications not in terms of add-or-replace model but in terms of an open-horizon model: electricity enabled humankind to build entirely new kinds of products; it gave us superpowers, in many ways. We tamed electricity, and it has enabled us to build products that to many were unimaginable prior to their invention. Our lives changed alongside these inventions, we adapted. The view of DS/ML as emergent technology emphasises the potential of DS/ML without giving it concrete form. It emphasizes that we can change, as technology changes around us by our own actions; the adoption of DS/ML becomes less of a zero sum game with winners (machines, technologists, STEM) and losers (humans, humanities).

In the DS/ML as modular mode, the technology and product side of the business were concerned about editorial and creatives and their reaction to the arrival of consultants at the company and their suggestions. There was a big difference between the expected and the actual situation. Viewed as modular, DS/ML devalues conditions and context and with it interaction with teams across companies, especially outside the technology teams, a potential explanation for this discrepancy. It devalues the importance of knowledge of teams outside technology groups, it can lead a kind of "benevolent paternalism". Editorial and creatives, on the other hand, were aware of inefficiencies in their work and were keenly aware of what questions they would like to have answered by data. With confidence in their work, they were looking to DS/ML as an enabler, potential partner, more in line with seeing DS/ML as the new electricity; different lines of business may be more susceptible for one way of thinking about DS/ML with consequences for communication across business lines.

Vendor Strategy

For some of their DS/ML solutions, the client relied on vendors; they paid a companies for a data product or DS/ML service. In one case, the client shared their data with a vendor company for product/service delivery. It was considered to be a "good deal" since the client company was not charged by the vendor (the payment, of course, is in the form of data).

The DS/ML as modular view sees data as a mere requirement for data products and DS/ML services (as AlphaGo Zero showed us, not even a necessary one). As long as you get a product in return for your data, it is a "good deal". The DS/ML as emergent technology view promotes the idea of data as a resource, an enabler. Data enables an entire suite of data products; sharing your data in exchange for one data product becomes a "bad deal" especially if data sharing enables your competition. Most vendors work with multiple organizations often in the same line of business. Data sharing, via such vendor, can remove competitive advantage that increasingly lies in data, as the DS/ML as emergent technology view emphasizes. The DS/ML as modular view deemphasizes data as a resource, a valuable asset that is best protected; it can lead to decisions with negative consequences for the competitiveness of the business in the long term.

ETHNOGRAPHIC LESSONS LEARNED

The Role of Expertise

Taking as a starting point the "simple premise that expertise is something people do rather than something they have" (Carr 2010), it becomes possible to see this case study as revealing the tensions and misunderstandings that arise from the differing sets of practices that are called upon in the shift towards DS/ML within business enterprises. The two motivating myths presented above constitute DS/ML as two different kinds of capabilities. One myth presents a modular capability that can be added instrumentally to existing practices, the other presents a transformative capability that requires the reshaping of existing business processes to new, sometimes custom, interfaces of the emerging and still unstable technology. Each of these two kinds of capabilities, then, entails a different set of interactions between data, personnel, products, and tools, and therefore a different set of practices through which expertise functions. By understanding these two myths as motivating different forms of expertise, the positioning of various actors in the case study presented above becomes more legible.

What expert practices are motivated by the modular myth? The modular myth lends itself to picking and choosing amongst instruments to be deployed, and expertise in this context would be constituted by performing knowledge about these available tools. Such performances might include conducting cost-benefit analyses on available vendor solutions, performing knowledge about the available packages and implementations, and situating DS/ML development as the stepwise incorporation of such modular tools into existing architectures. Indeed, we observed a reliance on such performances of expertise in the reaction of some in the case study presented above. And in particularly power-laden ways, this exercise of expertise was able to repress challenges posed by alternate forms of expertise (see below) by leveraging existing control of economic resources to prioritize one set of priorities (vendor solutions) over others (reorganizing the DS/ML team).

The expert practices mobilized by the modular myth also draw strength from a particular conception of objectivity mobilized by the modular myth. As a historically- and socially-constituted value, objectivity (Daston and Galison 2007) can take many forms. The modular myth contributes to a form of "mechanical objectivity" that sees human judgement as failable, whereas algorithmic systems can stand in for human actors who may introduce "bias, inefficiency, and discrimination" (Christin 2016). Trusting algorithmic systems over human actors allows those who exert control over the use of such systems to participate in this form of objectivity as a further practice of their own expertise. However, sources of mechanical objectivity, whether they be crime scene photographs or brain scans (Dumit 2004) tend towards a situation in which the products of these tools themselves require further expertise in order to be translated for lay audiences or integrated into other sociotechnical systems. The ability to do so constitutes a form of objectivity Christin identifies as "trained judgment". It was precisely this form of objectivity that we highlighted by introducing the story of DS/ML as the "new electricity" through the transformative myth presented above. By demonstrating the ways in which trained judgment could form a "hybrid entanglement of human and machine expertise" (Christin 2016), we show that there was a great deal of human ingenuity still required to craft DS/ML solutions for the particular problems the client was facing. How to make those problems legible to the machine were very human questions, and their outcomes were uncertain, so our best recommendations centered on empowering that form of expertise.

What expert practices are motivated by the transformative myth? The transformative myth lends itself to precisely those practices of expertise that constitute trained judgement, but a trained judgement that extends beyond that which might evaluate between several similar products offered by a vendor. These practices include engaging in forms of collaboration and experimentation that treat DS/ML not as a stable product, but as a set of open, unresolved questions from which meaningful solutions might emerge. Specifically, the expertise of a data science lead (or a VP of electricity, for that matter), is entailed by fostering different lines of communication between disciplinary silos, for example by enacting a process in which data analysts work with data scientists to craft key performance indicators that are useful for machine learning experiments. This form of expertise is also entailed by wielding economic resources to engage in experiments that may be fruitless, but also may produce useful insights or products for further development.

What other expert practices are at stake? The creative team, who in the planning stages of the on-site workshop were to be insulated from any hints that their roles could be automated, was revealed during the workshop to have their own expert practices that actually positioned them to be promising collaborators for the DS/ML team. Indeed, they were central to the business offering at the company, but they also were able to position their work as primarily valuable because they were the ones who 'crafted' new content for the media company. By foregrounding this aspect of their work and downplaying the routinized labor they performed, they could have pragmatic conversations about how to automate the routine work without their central expert practices being compromised. The DS/ML team could potentially be given broad latitude in building systems for the curation of past content, summarization of aggregated content, and the monitoring of dashboards without threatening the practices that constituted creative expertise.

In reflecting on "the pervasive sense that technologies transform us in irrevocable ways means that idealistic concepts of technology are always accompanied by the anxiety that they will also promote some kind of loss - loss of connectivity, of intimacy, of desire, of authenticity in some way." (Sturken & Thomas 2004) we were surprised to realize that this was a far more active concern for those whose expertise depended on control over the technologies that were the subject of the workshop, and not those who were most central to the production of the content offered by the company. This points towards two key findings from the engagement. The first is that where there is resistance to recommendations for a move away from modular solutions and towards transformative capabilities, sensitivity to different enactments of expertise are key. Unless existing expert practices can be reshaped or otherwise adapted to the kinds of practices entailed by a focus on transformative capabilities, a defensive, dismissive, or destructive reaction is possible from those like the CTO, whose existing expertise will be subsumed by such a shift.

The second key insight is not all that different from the lessons Latour drew from examining the history of the pasteurization of France (Latour 1993). While singular inventions and modular capabilities may sometimes be identified as transformative in their own right, they are not enacted or brought to bear on the world without a broad accommodation of the social sphere to the technological apparatus, and of the technical apparatus to the existing practices within the social sphere. As little as Louis Pasteur could accomplish in France on his own, just as little could be done by any one person in the offices of the client in our case study. Rather, ground must first be laid across the organization to accommodate the kinds of changes that any particular form of DS/ML might take. This groundwork can be done purposefully, but requires the active participation of the entire range of actors likely to be impacted by such changes. It also requires working against the emotional grain produced by spectacular demonstrations of DS/ML. The ways in which such spectacles mobilize the sublime are quite persistent, and effectively immunize against alternate understandings of the technologies as anything but modular.

The Modular Myth

The myth of the modular addition of capabilities contributes in concrete ways to the the emergence of "technology" as a "hazardous concept" (Marx 1997) that refuses interaction with anyone besides experts (conditions to not matter). The hazardous concept, in Leo Marx's analysis, is that of technology as an "singular noun" capable of acting as an agent in history. In his telling, it is the technology that affects people's lives and reshapes the possibilities for human existence, not the field of individual actors who comprise the sociotechnical system in which technology is embedded. AlphaGo (Zero), and other spectacular demonstrations, mark the unfolding of DS/ML as a succession of particular inventions, recapitulating in miniature the sweeping narratives of human progress that are marked by key inventions — stone tools, fire, the wheel, gunpowder, semiconductors, perceptrons, reinforcement learning — that have played active roles in human history. Marx goes on to point out that narrow conceptions of technology as constituted by discrete objects like the steel plow or the steam engine are "merely one part of a complex social and institutional matrix", that is entailed by large scale technosocial system. This understanding informs an understanding of technology as a constitutive force that shapes society as a whole, but particularly reshapes the institutions, including corporations, that are intimately bound up with developing, employing, and deploying new technologies. In the context of this paper, technology may sometimes be seen as an active force in the constitution of the corporation.

***The Modular Myth and the Technological Sublime* –** In American Technological Sublime, David Nye (1996) discusses spectacular presentations of technology as participating in an experience of the sublime. The sublime, in this context, is not a "self-conscious aesthetic theory" but rather a "cultural practice of certain historical subjects" that continually produces "new sources of popular wonder and amazement", in Nye's analysis. The modular myths of DS/ML development gain their mythological status from the technological sublime, and considering these spectacles as such through the lens Nye provides is instructive. The sublime, in it's Kantian, Enlightenment-era sense has both a 'mathematical' and 'dynamic' aspect. The mathematical sublime pertains to an experience of scale that produces wonder in a human subject. The Grand Canyon, the vastness of space, and the Great Wall of China all exist at scales dwarfing the normal realm of human experience, and produce, according to Kant, a sense of the mathematical sublime. The dynamic sublime is more closely associated with a sense of terror, as when a crowd gathers for a skyscraper demolition or to watch a passing storm from a safe distance.

The spectacular, modular myths of DS/ML participate in both these forms of the sublime, and indeed are key to understanding these cultural objects as spectacles. The scales at which an algorithm may run are constantly foregrounded in promotional materials, as in a documentary about the defeat of Go champion Lee Sedol by AlphaGo, which tells us that "a game of Go has more possible configurations than there are atoms in the universe". The number of petaflops a computer is capable of, the number of cores and GPUs brought to bear on a computational problem, the nearly infinite permutations of possible outcomes, are all made clear to an audience in order to produce a sense of the mathematical sublime. There is terror in these spectacles, too. Even setting aside the many terrifying scenarios of an "AI Apocalypse" in which machines actually attack humanity (see Dowd 2017), in many ways DS/ML mythmaking points towards a world that doesn't need human subjects at all, self-driving cars, efficiently optimized factories, and flawless recommendation systems sketch out a world in which the human is mostly incidental. Like Niagra Falls, it will keep churning, oblivious to our existence, and that such a world is possible induces a sense of the dynamic sublime.

While these Kantian forms of the sublime are certainly at play in the modular myths of DS/ML, they are also legible as an iteration of the electrical sublime that Nye presents, in which spectacles moved beyond the realm of the natural world, and were developed specifically for celebrations of industry, nationalism, and amusement. These modular myths, as spectacle, make invisible technologies visible. Electricity was made visible through lighting displays, just as AlphaGo (Zero) makes algorithms and data streams visible, as events that pit a human master of a game against a computer: AlphaGo defeated Lee Sedol in front of a human audience.

The Myth of the "New Electricity"

A crucial point Andrew Ng's "VP of Electricity" metaphor makes is about the complexity of emerging technologies and the necessity of expertise to adequately grapple with that complexity. Because of the siloed nature of divisions in modern corporations (Rumelt 1974), expertise is not easily distributed across an organization. Supporting DS/ML expertise in any one part of a company will not necessarily translate to other parts of that company, unless they are empowered to make changes beyond their own division. And as the DS/ML experts will not be able to influence business practices outside their own division, it becomes difficult if not impossible to transform those practices in ways that integrate well with the DS/ML projects they work on. By placing a DS/ML expert at the executive level, or by explicitly designing processes for distributing that form of expertise across existing divisions, the complexities of the emerging technology can be addressed in a coordinated, rather than piecemeal, fashion. An expert in DS/ML can approach these capabilities as resource-driven, capable of using data to transform existing products and processes in ways that a modular, bolt-on approach cannot.

The tendency of the discourse around DS/ML towards narrating the emerging technology as a modular addition of capabilities rather than as resource-driven is highlighted by the way Ng's story was bent towards a metaphor of AI as "the new electricity" (Eckert 2016). Portraying it as such is a subtle rhetorical move that foregrounds the power of the new technology eliding the challenges that remain in building practices around it whilst pointing to future, as of yet unrealized potential. The power of electricity is readily visible to any audience that hears that "AI is the new electricity", even if not all listeners connect AI, machine learning, and the data that drives it with the role electricity has played as a public utility (as opposed to the private commodity data currently is). Indeed, the challenges that were present in the early days of electrification, however, have receded to the background. It has become infrastructural, visible only when it fails (Star 1999). According to Ng, DS/ML share its eventual invisibility and great power.

CONCLUSION

The algorithms that are powered by data participate in their own set of metaphors. Some are 'intelligent', while others are merely 'smart'. The use of games like Chess or Go as demonstrations of DS/ML perpetuate this metaphor. Such games have long been proxies for human intelligence (Ensmenger 2012), foregrounding certain human skills like foresight, planning and concentration over others like sensitivity, compromise, or even deception. But the use of these games in AI research remain an abstraction of human cognition that fails to capture the entire gamut of human intelligence. These are distinctly human capabilities that set algorithms on an even playing field with people who may feel more threatened than enhanced by their presence in the workplace. This tension between human and machine becomes more acute when DS/ML is described as superhuman, either in terms of being hyper-rational, hyper-vigilant, or omniscient. In some cases, DS/ML is imbued with capabilities bordering on the clairvoyant, as in breathless headlines like, "Google's AI Can Predict When A Patient Will Die" (Tangerman 2018). Framing the capabilities of DS/ML as on par with, or even as surpassing, human capabilities places it in competition with the humans who must be full participants in any integration of DS/ML into a company. However, this participation is frequently fraught due to an inadequate consideration of the "affective relationship to the product or system, that is, how someone feels about the technology at stake" (Elish & Hwang 2016).

The metaphors of big data tend to treat data as a resource from which value can be extracted. The metaphors of DS/ML tend to treat machines as somehow more than human, which is to say they have many of the strengths of humans (intelligence, anticipation) but few of the weaknesses (inattention, exhaustion). Both of these sets of metaphors elide the uncertainties inherent in the metaphors they employ. Resource extraction is not a linear processes, it involves the failure of exploratory wells, infrastructural costs to move minerals to markets, and shifting price and demand curves relative the costs of extraction. Neither is human intelligence a completely predictable process, particularly where the development of science and technology are concerned.

In this paper, we have discussed the how the prevailing stories that highlight the emergence of data science and machine learning tend towards an understanding of DS/ML as a modular capability. These stories fail to promote transformative practices that might reshape existing business problems into ones that the emerging capabilities of DS/ML can currently address. To do so would require an attention towards data not as an ingredient, but instead as a means through which other things become possible, but also requires a different set of expert practices than those that are currently incentivized by many technical teams, which was particularly true in the case study laid out above. By understanding expertise as sets of practices that can be encouraged and rewarded, rather than as an object that can be possessed by individuals (Carr 2010), we point the way towards undertaking broad shifts in overall business practices by seeking transformative changes that are not siloed within individual departments, but rather have the opportunity to reshape existing practices broadly in pursuit of interfaces that match the underlying technical capacities of DS/ML with the specific, measurable business needs of an organization.

NOTES

Acknowledgements - The authors would like to thank Cloudera Fast Forward Labs; without their support this work would have been impossible. All conclusions represent the work of the authors, and should not be interpreted to represent the position of Cloudera Fast Forward Labs or any of its employees or officers. The authors would also like to thank Dawn Nafus for her generous notes and Jan Philipp Balthasar Müller for reading a late draft. Emanuel Moss is grateful to the CUNY Graduate Center, Data & Society Research Institute, and the Wenner Gren Foundation for their support.

1. https://aws.amazon.com/ml-solutions-lab/

2. https://cloud.google.com/asl/

3. https://www.elementai.com/

REFERENCES CITED

Carr, E. Summerson
2010 "Enactments of Expertise." Annual Review of Anthropology 39 (1): 17–32.

Christin, Angele
2016 "From Daguerreotypes to Algorithms: Machines, Expertise, and Three Forms of Objectivity." ACM Computers & Society 46 (1): 27–32.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa
2011 "Natural Language Processing (Almost) from Scratch." Journal of Machine Learning Research 12: 2493–2537.

Daston, Lorraine, and Peter Galison
2007 Objectivity. Brooklyn, NY: Zone Books.

Dowd, M
2017 Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse. Vanity Fair. April.

Dumit, Joseph
2004 Picturing Personhood: Brain Scans and Biomedical Identity. Princeton, NJ: Princeton University Press.

Eckert, Daniel
2016 "Artificial Intelligence Is the New Electricity?" Vox Creative.

Elish, M. C., and danah boyd
2017 "Situating Methods in the Magic of Big Data and AI." Communication Monographs, 1–24.

Elish, M.C., and Tim Hwang
2016 An AI Pattern Language. Data and Society. http://autonomy.datasociety.net/patternlanguage/.

Ensmenger, Nathan
2012 "Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm." Social Studies of Science 42 (1): 5–30.

Fei-Fei, Li, and Pietro Perona
2005 “A Bayesian Hierarchical Model for Learning Natural Scene Categories.” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) 2: 524–31.

Ford, Heather
2014 “Big Data and Small: Collaborations between Ethnographers and Data Scientists.” Big Data & Society 1 (2).

Karpathy, Andrej
2017 “AlphaGo, in Context.” Medium.

Koch, Gregory
2015 “Siamese Neural Networks for One-Shot Image Recognition.” University of Toronto.

Latour, Bruno
1993 The Pasteurization of France. Cambridge, MA: Harvard University Press.

Laurel, Brenda
2013 Computers as Theatre. New York: Addison-Wesley Publishing Company.

Marx, Leo
1997 “‘Technology’: The Emergence of a Hazardous Concept.” Social Research 64 (3): 965–88.

Mearian, Lucas
2016 “Google’s DeepMind A.I. Can Slash Data Center Power Use 40%.” Computerworld.

Moravčík, Matej, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling
2017 “DeepStack: Expert-Level Artificial Intelligence in Heads-up No-Limit Poker.” Science 356 (6337): 508–13.

Ng, Andrew
2016 “Hiring Your First Chief AI Officer.” Harvard Business Review. https://hbr.org/2016/11/hiring-your-first-chief-ai-officer.

Nye, David E.
1996 American Technological Sublime. Cambridge, MA: MIT Press.

Popper, Karl
2005 The Logic of Scientific Discovery. New York: Routledge.

Rumelt, Richard P.
1974 Strategy, Structure and Economic Performance. Ann Arbor: University of Michigan Press.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al.
2017 “Mastering the Game of Go without Human Knowledge.” Nature 550 (7676). Nature Publishing Group: 354–59.

Star, S. L.
1999 “The Ethnography of Infrastructure.” American Behavioral Scientist 43 (3): 377–91.

Acting on Analytics: Accuracy, Precision, Interpretation, and Performativity

JEANETTE BLOMBERG
IBM Research

ALY MEGAHED
IBM Research

RAY STRONG
IBM Research

We report on a two-year project focused on the design and development of data analytics to support the cloud services division of a global IT company. While the business press proclaims the potential for enterprise analytics to transform organizations and make them 'smarter' and more efficient, little has been written about the actual practices involved in turning data into 'actionable' insights. We describe our experiences doing data analytics within a large global enterprise and reflect on the practices of acquiring and cleansing data, developing analytic tools and choosing appropriate algorithms, aligning analytics with the demands of the work and constraints on organizational actors, and embedding new analytic tools within the enterprise. The project we report on was initiated by three researchers; a mathematician, an operations researcher, and an anthropologist well-versed in practice-based technology design, in collaboration with a cloud services go-to-market strategy team and a global cloud sales organization. The analytics were designed to aid sellers in identifying client accounts that were at risk of defecting or that offered opportunities for up-sale. Three-years of sales revenue data were used to both train and test the predictive models. A suite of analytic tools was developed, drawing upon widely available algorithms, some of which were modified for our purposes, as well as home-grown algorithms. Over the course of this project important lessons were learned, including that the confidence to act upon the results of data modeling rests on the ability to reason about the outcomes of the analytics and not solely on the accuracy or precision of the models, and that the ability to identify at-risk clients or those with up-sell opportunities by itself does not direct sellers on how to respond as information outside the models is critical to deciding on effective actions. We explore the challenges of acting on analytics in the enterprise context, with a focus on the practices of 'real world' data science.

INTRODUCTION

There is a pervasive view that data analytics can lead to better managed organizations and enhanced organizational performance as employees and their managers are guided to make more informed choices (Davenport, 2007, 2010; Gillon, et al. 2014). Analytics, it is argued, can assist with hiring decisions (Levenson, 2011), targeted sales (Megahed et al. 2016a, 2016b; Wixom et al., 2013), market opportunities (Chen et al., 2012), supply chain management (Gunasekaran, et al., 2017) and shop floor scheduling (Zhong et al., 2017) to name a few. However, the impact of analytics on organizations is governed by the ability to align the outcomes (i.e. predictions, optimizations) with the everyday requirements of the work, and the ability of organizational actors to make sense of the analytics and be positioned to take action informed by them (Sharma et al., 2014).

In part, what is driving enthusiasm for enterprise analytics are the increasing number of organizational processes that generate digital information about the execution of these processes. Zuboff (1985, 1988) was one of the first to recognize the potential of these new sources of information to generate insights about a company's operations and how to improve upon them. More recently there has been renewed excitement in tapping into a company's internal databases, both more structured, so-called *systems of record*, and what Moore (2011) has called *systems of engagement* which generate decentralized information, including interactions in real-time through mobile and social technologies. These internal company data sources are claimed to offer competitive advantages for those organizations able to mine them for insights.

Underpinning these claims are assumptions about the availability of useful data, either sourced internally or externally available. While it might seem straightforward to gain access to internal company data, this is not always the case. Data may be scattered throughout the organization in private or personal databases that in theory are available, but in practice the effort involved in centralizing the data in a single repository may be prohibitive unless a long-term payoff can be clearly and confidently defined. Even internal data that are kept in central locations can present problems for their use in situations where the way the data are 'produced' has varied as boundaries between organizational entities are redrawn or the work processes and policies are redefined, changing the 'meaning' of the data over time. For example, a product offering in the portfolio of one organization may be moved to a newly created organization's portfolio, making it difficult to make machine learning predictions about future sales without making assumptions about the stability or transformation of the data in the new organization. Likewise, a change in policy, such as the point at which sales representatives are required to get price approval, can modify what is recorded in a sales database. Again, if the data are to be used, analysts and data scientists will be required to make assumptions about the importance of such changes and how best to account for them in their analyses.

'Detective' work is often needed to uncover organizational changes that must be accounted for to understand the data and to interpret the outcome of the analytics. This means that measures of analytic precision and accuracy which are often used as quality checks on the analytics, must be measured against the confidence that the data represent meaningful organizational phenomena (Hovland, 2011). In the above examples, the analytics may indicate an organizational or policy change and not a change in selling behavior or future sales opportunities.

Additionally, organizational problems must be framed as ones that the available data are well suited to address. It is not always the case that the most important questions are ones the data in hand are able to shed light upon. Opportunities to upsell may have more to do with information not readily available such as personal relationships between the seller and client or recent contacts clients have had with competing vendors. The analytics team must be realistic about what can be learned from the data available given its limitations. They must assess if the data in hand is adequate to address issues of concern or if they need to invest in acquiring additional data. This situation reminds us of the well-known adage that what can be measured is not always what is worth measuring (Muller, 2018:3). The most important issues may not be those that are addressable by the data available.

Furthermore, numbers have little organizational power unless they can be understood and trusted by organizational actors (Power, 1997)) who themselves are caught up in structures of accountability often outside their immediate control (Barley and Tolbert, 1997). Even when the results of analytics suggest particular courses of action, workers may not be in a position to take such action. For example, in an earlier study by the first author, predictive analysis showed that hiring additional people would increase the throughput of an organizational process and in the end offer financial benefit to the organization, but it was not acted upon because the power to make hiring allocations laid outside the responsibility of the process owners. As the excitement surrounding the potential of advanced analytics confronts the reality of acting upon the analytics within the enterprise, it is becoming increasingly clear that 'explainability' of outcomes will gate the usefulness of the analytics (Abdul et al., 2018; Miller, 2017; Ribeiro et al., 2016). Organizational actors are unlikely to act upon the analytics if they do not trust the outcomes, feel confident in the rationale behind the analytics, and understand the limitations, strengths and weakness of the analysis.

THE CASE

Our case reports on a two-year project to develop sales analytics for an internal group of global cloud IT infrastructure-as-a-service (IaaS) sellers and their managers. Cloud IT infrastructure services are a relatively new type of service that provides computing resources over the internet. Cloud services differ from traditional 'fixed duration IT service' contracts where modifications to a contract can only occur by agreement of the client and the provider and under circumstances clearly outlined in the contract. The new cloud service offerings primarily are sold based on a consumption model: the more the client consumes of the service the more they pay. So the amount of a service consumed, such as number of virtual servers or the amount of storage used, can go up or down depending on the client's needs without a change in the contract. Our project aimed at developing sales analytics to provide insights into client buying and consumption behavior for these new IT infrastructure-as-a-service offerings.

The research team included a machine learning mathematician with prior experience working with the type of data used to build our predictive models, an operations researcher who had developed analytics to predict win-rates for IT infrastructure service contracts (Megahed et al., 2015), and an anthropologist with many years of experience studying organizational work practices, including the work of those who deliver IT infrastructure services. The three researchers worked with a business unit strategy team tasked with helping improve the go-to-market or selling capabilities of the cloud services organization by providing training, sales tactics, and cross-team communication support. We engaged the go-to-market strategy team and the global cloud sales leadership to ascertain the potential value of predictive sales analytics and later directly with sellers and their managers to assess the usefulness of the analytics and how our predictions could be of benefit in their daily practices.

The cloud organization was global and consisted of several business divisions, each with a different set of service offerings in its portfolio. We focused most of our efforts on two geographies, Europe and North America; and two business divisions, one selling 'on premise' cloud services[1] and the other 'public' cloud services[2]. During our project there were two realignments in the cloud organization which resulted in some cloud service offerings being moved from one division of the organization to another.

We used three years of ledger data that recorded revenue for the cloud services organization to develop the predictive models. These data included the name of the client, the offerings sold, the business unit credited with the sale, and the revenue realized. Our aims were to help sellers prioritize sales opportunities, reduce churn and defections, target particular cloud service offerings for expansion, and improve sales productivity overall. The sellers we worked with were members of the direct sales team who had responsibility for specific sales territories and particular clients or client types.

Our overall mission was to enable the cloud sales organization to become a leader in enterprise cloud solutions by providing them with the analytic tools to grow the cloud business, including basic reporting and advanced analytics. Our case study reports on the initial stages of the implementation of a longer-term vision (see Figure 1), where we initially focused on sales leaders responsible for specific geographic territories (geo leaders), sales managers, and sellers as our users. We developed a starter set of analytics that included risk of defection (e.g. customers likely to terminate their contract) and growth or shrinkage of client and offering revenue. Our initial data sources were ledger data and client registration data. In the longer term, we envisioned enabling others in the company to use our 'platform' to add new data sources, analytics, and users.

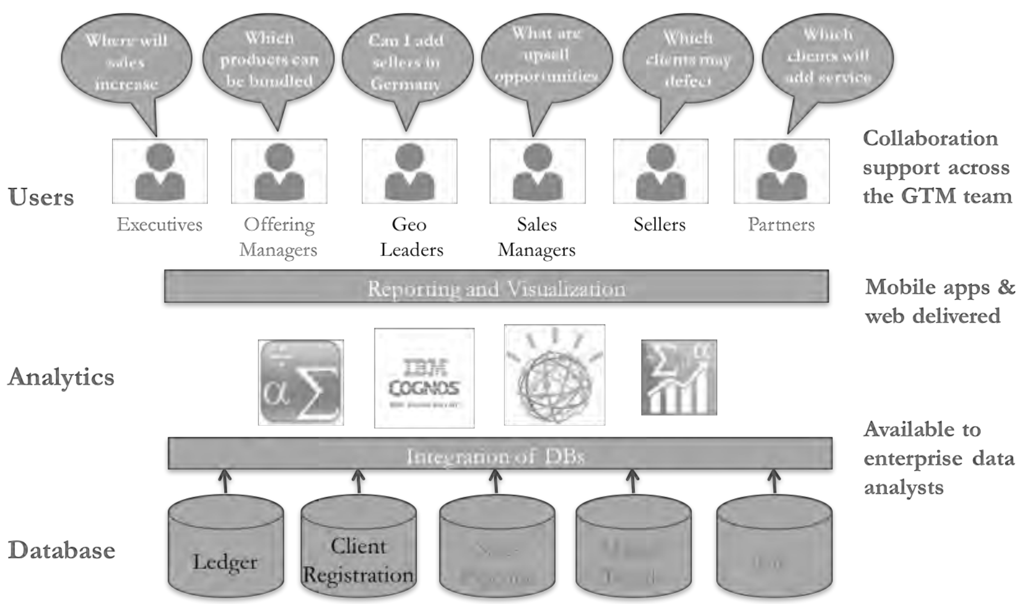


Figure 1. Cloud Sales Analytics Long Term Vision

THE PRACTICES OF DOING DATA ANALATICS

Organizations today have heightened expectations about the contribution advanced machine learning approaches can make to their performance, whether improving internal processes, more successfully connecting with clients, or planning for the future (e.g. hiring, resource allocation, expansion of operations, etc.). However, organizations are only beginning to understand what is required to 'unlock' the secrets of the data sequestered in internal corporate databases. We outline some of our experiences doing data analytics in the enterprise, specifically developing sales analytics for the cloud organization where we highlight practices implicated in transforming 'data' into insights to drive actions within the enterprise. These practices include, data sourcing and cleansing, selecting algorithmic options, troubleshooting 'errant' outcomes, and iterating on analytic models and output.

Data Sourcing and Cleansing

It goes without saying that one of the first tasks required to turn data into insights is gaining access to data, in our case sales ledger data and client registration data. This involved obtaining many approvals where we had to argue for the importance of our project and also demonstrate how we were going to protect the security of this highly confidential data. Once we were granted access to the data we had to identify people in the organization who understood how the ledger database was structured, for example, in tables of various kinds. We then had to write scripts to query the database and export just the data we needed for our analyses. Since these data needed to be updated monthly, we later automated this process to keep the data up-to-date as last month's analyses, while useful, were not nearly as valuable as those that included the most recent revenue figures.

We also found that the ledger data needed to be aggregated to reduce the number of data points used in the analysis. By aggregating the revenue data by month we were able to run our computations faster, facilitating both experimentation and debugging of the algorithms, and eventually the time needed to routinely create up-to-date reports. In developing the algorithm to predict risk of defection we experimented with aggregating monthly data by calendar quarter to reduce some of the noise found in the monthly data where revenue recorded for one month might later be moved to a prior month based on new information. Previous experience with the ledger data showed that calendar quarter data was much less noisy than monthly data which was in part because at the quarter close additional actions were mandated to validate the accuracy of the entries. However, based on feedback from sellers where they expressed a desire to have monthly updates to our predictions, we experimented with a three-month moving average where, somewhat to our surprise, we found the predictive power of our algorithms was not significantly diminished. We finally settled on aggregating the data by a three-month moving average enabling us to update our predications monthly.

Another issue we had to deal with was resolving differences in how entities (clients and service offerings) were named in the data corpus. Entity recognition and resolution is a near universal problem in data analytics and we too had to decide, for example, whether to combine all client accounts at the highest recognizable corporate level. Since the ledger revenue data is based directly on billing operations, it was not surprising to find accounts assigned to billing addresses and not a single 'corporate' address associated with a chain of franchises of the same corporate brand. And for large, complex organizations there might be global subsidiaries of the 'same' company with somewhat different names. Should they be treated as unique entities or combined as a single entity? These distinctions are very hard to recognize programmatically and required data cleansing efforts that were far from trivial. We ultimately arrived at a method for addressing these naming issues knowing we could have made different choices. There was no a priori 'right' way to aggregate and name entities, but any choice made had consequences for our predictions, the interpretation of the results, and how best to target interventions. While more experiments likely would have enabled us to better understand the impact of our choices, we settled on a strategy of client name resolution feeling pressure to get our results to the sellers for their feedback on the usefulness of the predictions.

Algorithmic Options

Early interactions with the cloud services go-to-market strategy team led us to focus our initial analytics on predicting risk of defection, growth and shrinkage in account revenue, growth and shrinkage in service offering revenue, and cross-sale opportunities[3]. Our algorithms deployed supervised machine learning approaches, where we focused on developing models (or patterns in the ledger data) to identify which client accounts[4] were at risk of defection. For this analysis we used three years of revenue data aggregated by month for each client in a given country (e.g. Global Fin Company in France recorded $100K in revenue in April of 2015, $110K in May, $110K in June, etc.).

Through machine learning experimentation we discovered that a single analytic feature that we called the 'quotient' was a good predictor of accounts that were likely to defect in the following six-month period. The quotient uses nine months of revenue data for the prediction and outputs a short list of accounts at risk of defection. Our analysis showed that roughly half the accounts on the list would defect within six months unless action was taken. The quotient (Q) is calculated using a relatively simple formula which takes the current three months of revenue (C3) and divides it by the average revenue over the prior six months (A6) divided by two. $Q = C3/(A6/2)$. The list of accounts at risk of defection is sorted by the geography and country, and ranked by a relative quotient score between 0% and 100%. The relative score considers likelihood of defection as output by the model (Figure 2).

Although this 'simple' algorithm yielded useful results, with precision metrics in the 50% range, we wanted to explore more advanced machine learning methods to see if we could improve the precision and accuracy of our predictions. For this second effort we focused on predicting the growth and shrinkage of the average revenue. We wanted to know how likely it was that revenue by client or by offering would grow or shrink by X% in the next six-month period compared to the average revenue for the current three-month period. For account predictions, revenue was aggregated for all the offerings sold to any given account in a given country. For offering predictions, revenue was aggregated for all the accounts that sold a given offering in a given country. We experimented with different baseline classifiers (Abhinav et al., n.d.) and found gradient boosting machine (GBM) classifier (Chen and Guestrin, 2016) yielded the best results for accuracy. To achieve this metric, we divided the historical labeled dataset into training and testing (80% training and 20% testing). In our case the data consisted of three years of cloud sales revenue data. The model was trained using the training data doing k-folds cross-validation, where the training dataset is divided into k

Account name	Country	Geo	Business division	Score	Revenue		
					Dec 2016	Dec 2017	Feb 2017
Global Fin	France	Region 2	BMDiv 1	100	$11,000	$3,000	$50
Global Ind	US	Region 1	BMDiv 2	100	$250,000	$1,300.000	$6,000
Telco Lite	US	Region 1	BMDiv 2	93	$38,000	$39,000	$3,000
Digital Exc	Canada	Region 1	BMDiv 2	83	$67,000	$81,000	$13,000
Media High	US	Region 1	BMDiv 1	80	$129,000	$96,000	$24,000
…							
…							

Figure 2. Risk of Defection Report

folds, and the model is trained k times on k-1 fold and tested on the held-out fold. This was done to avoid over-fitting, which might result in the model being too good for the training data, but not for the new testing data. Then, a final trained model was run on the testing data to evaluate it on a number of metrics, including precision and accuracy. We experimented with multiple classifiers and chose the ones that gave the highest accuracy on the testing dataset. This model was then used for our predictions of future data points where the outcomes are not yet known. We further developed the model for precision maximization at a minimum recall and solved it using Gaussian optimization. This new model, we called GOPT, directly maximizes precision to yield more actionable results while still maintaining a high degree of accuracy (Abhinav et al., n.d.). Features of both models included revenue for the past nine months (3 quarters), country of the client, business division, and several constructed features not recorded directly in the ledger data (Figure 3).

Features	
Client ID	Backward Ratio
Client Name	Forward Ratio
Offering ID	Log (Backward Ratio)
Offering Name	Log (Forward Ratio)
Geography	Sign (Backward Ratio)
Business Division	Sign (Forward Ratio)
Quarter	# of Positive Quarters
Revenue for Quarter Q	
Revenue for Quarter Q-1	
Revenue for Quarter Q-2	

Figure 3. Features Used for Growth and Shrinkage Predictions

The output of the model was a list of accounts (and offerings) that were predicted to grow or shrink in the next six-month period by X% over revenue for the last three months. The percentage of growth or shrinkage could be set to between 0% (defection) and 100% (double revenue). For our initial reports we set the percentage to 50% growth or shrinkage. The results were sorted by geography and country and ranked by a relative score between 0% and 100% which considered the likelihood of growth or shrinkage as output by the model and the average revenue for the last three months (Figures 4 and 5). We chose to include in our ranking the average revenue for the last three months to prioritize (higher on the list) those accounts or offerings with the most potential revenue gain or loss in absolute dollars.

The values for precision and accuracy differed depending on which growth or shrinkage percentages were used, however, these figures were consistently higher than for the simpler risk of defection quotient model. When our model was tuned to maximize precision over accuracy, the precision of the growth and shrinkage models was over 90% while holding accuracy to over 80% (Abhinav et al., n.d.).

Growth					
Account name	Country	Geo	Business division	Score	Revenue Jan 2016
Banco One	US	Region 1	BMDiv 1	100	$365,000
Gov Works	US	Region 1	BMDiv 1	96	$360,000
Air Run	Germany	Region 2	BMDiv 2	25	$120,000
Max Tech	Canada	Region 1	BMDiv 2	7	$30,000
Shrinkage					
Account name	Country	Geo	Business division	Score	Revenue Jan 2016
Media Two	France	Region 2	BMDiv 1	12	$290,000
Tech Max	US	Region 1	BMDiv 1	5	$97,000
Home report	Germany	Region 2	BMDiv 2	4	$69,000
Best Consult	Canada	Region 1	BMDiv 2	3	$77,000
…					

Figure 4. Accounts Predicted to Grow/Shrink by 50% Report

Growth					
Offering name	**Country**	**Geo**	**Business division**	**Score**	**Revenue Jan 2016**
Top services	US	Region 1	BMDiv 1	42	$,900,000
Cloud open	Canada	Region 1	BMDiv 1	27	$1,120,000
Portfolio X	Italy	Region 2	BMDiv 2	9	$470,000
Local serv	Germany	Region 2	BMDiv 2	8	$320,000
Shrinkage					
Offering name	**Country**	**Geo**	**Business division**	**Score**	**Revenue Jan 2016**
Self manage	US	Region 1	BMDiv 1	42	$,900,000
Open bridge	Canada	Region 1	BMDiv 1	27	$1,120,000
Encrypt serv	Italy	Region 2	BMDiv 2	9	$470,000
Package app	Germany	Region 2	BMDiv 2	8	$320,000
…					

Figure 5. Offerings Predicted to Grow/Shrink by 50% Report

Troubleshooting Errant Outcomes

Our interactions with sellers and sales managers was critical to our ability to debug our analyses, make course corrections in our methods and algorithms, and understand how our predictions could be useful in their everyday work. But before we shared the results with sellers, as their time was limited and we did not want to introduce any unnecessary concerns about the accuracy of our analytics, we first reviewed the output of our models to spot errors. Some errors were relatively easy to identify even by someone without domain knowledge. For example, we found an error in the early growth and shrinkage predictions where the same client was on both the list of accounts whose revenue was predicted to grow by 50% and also shrink by 50%. Once pointed out, a 'bug' in the code was quickly found and corrected. While this error was relatively easy for us to identify, it raised questions about the possibility that 'bugs' with a subtler impact on the predictions might go undetected. Since there is no 'ground truth' regarding which client accounts will grow or shrink, we had to rely on sellers and other users to identify problems with the data, the cleansing processes, the code that implemented the model, and even the measurements (e.g. accuracy and precision) that expressed confidence in the predications.

Errors only detectable by someone familiar with the domain of global IT cloud services and specific client or service offering required our ongoing interactions with sellers and sales managers. For example, in a couple rare cases completely distinct customers were confused in our cleansed data. Our method for resolving entity names created erroneous combinations of unrelated customers. These glitches were identified by the sellers who knew the clients better than we did and recognized that a client appearing on our defection list had no reason to be there.

In a somewhat different example, a sales executive pointed out to us that some offerings (e.g. professional services) were by design fixed duration contracts, even though they were sold by the cloud organization, and we should expect the revenue for these offerings to end without suggesting there might be a problem with the account. We queried the sellers to find out what offerings should be excluded from our analysis. While we always applauded the sellers when they pointed out anomalous results, we also knew this was a double-edged sword, as too many such errors could ultimately undermine their confidence in our analysis.

Iterating on Analytic Models and Output

Beyond their role in helping us detect errors and anomalies in the analytics, the sellers also made suggestions about how our analysis could be more useful to them. As discussed above the analytics were based on aggregating monthly revenue by quarter to smooth out fluctuations in the monthly revenue that detracted from our predictions. However, the sellers' temporal rhythms (how often they contacted clients, checked on the account status, were held accountable by their management) made it more useful to see changes in the predictions on a monthly basis. Because quarterly reports were not as valuable to them we re-wrote the code to aggregate the monthly data based on a rolling three-month period so each month new results were available while leaving in place the benefits to the accuracy of our predictions derived from the use of a three-month aggregation of the revenue data.

LESSONS LEARNED

Our experiences developing analytics for cloud services sellers taught us several important lessons. First, measures of analytic accuracy and precision, by themselves, do not govern the usefulness of the analytics. Second, the ability to predict outcomes with a high degree of confidence does not necessary suggest what actions should be taken in response.

Trading Precision and Accuracy for Interpretability

The first lesson is best demonstrated by comparing the reception received by our two different analytic models – risk of defection and growth/shrinkage prediction. The risk of defection model followed an easily communicated formula that produced the 'predictions' about which client accounts were likely to defect in the next six months. The sellers could inspect the monthly ledger data and posit a reason certain accounts were on the risk of defection list. While they would see fluctuations in the revenue numbers (there was not always a linear decline in revenue from month-to-month), they surmised that this pattern of fluctuating revenue was found in the historical data of those accounts that later defected. They could assess the reasonableness of the outcomes. This allowed a certain level in trust that the analytics were identifying client accounts with a greater likelihood to defect and as such, they should pay attention to these accounts.

What proved to be challenging for some sellers was to understand what was meant by the statement that accounts on the list had a 50% chance of defecting in the next six months. At first glance for some it seemed like a coin toss to say 50% of the accounts on the list would defect and 50% would not. But this list represented only a small fraction of all the accounts in the database and for the entire set of accounts the percentage that would actually defect was quite small. This example points to the importance of considering the denominator in interpreting the meaning of percentages (Guyer, 2014). Our report significantly narrowed the number of accounts that the sellers were advised to investigate why they appeared on the list, and for some, intervene to change the predicted outcome.

Contrasting the relatively simple risk of defection analytics with the growth and shrinkage analytics that used sophisticated machine learning algorithms, we found it more difficult for users to 'intuitively' reason about why some accounts or some offerings were on the list of those predicted to grow or shrink by 50%. The growth and shrinkage model took into account multiple features, producing highly accurate and precise predictions. However, the results of these models were difficult to reason about as some of the features were abstract and not easily mapped on to the sellers' everyday experiences (see Figure 3) and the math behind the algorithms was complex. Inspection of the revenue data could be confusing as it was difficult to see the direct link between the revenue data and the predictions, and impossible to explain in everyday language exactly how the model arrived at the predictions. In our reports to sellers we opted to show only the last month's revenue, because we feared showing more historical data would run the risk of confusing the sellers. The analytics found patterns in the data that humans could not 'see' requiring a level of 'blind' trust on the part of the sellers.

Our growth and shrinkage model had a variable that could be tuned, namely a percentage (between 0% and 100%) by which the client account or offering revenue would grow or shrink. We experimented with setting the percentages at different values and debated about such things as whether it would be more useful to set growth percentages higher or lower than shrinkage percentages. Originally, we planned to let the sellers set the percentage, but we soon realized this likely would introduce more confusion and realistically we knew the sellers had little time or expertise to experiment (even with our help) with how best to set this variable. The potential confusion stemmed, in part, from the fact that accounts could show up on the list as predicted to grow by 30%, but not the list predicted to grow by 20%. While we understood why this was possible as each prediction was discrete, we were concerned this would be difficult to explain to the sellers and this might lead them to distrust the results. From the model's point of view these results were explicable, but it was counter intuitive from the sellers' perspective, as they reasoned if an account is predicted to grow by 30% surely it also would be predicted to grow by 20%. In the end, we set the percentages for both growth and shrinkage at 50% with the view that over time we would tweak these percentages to be most useful to sellers as they began to trust the predictions and have a better understanding of what it meant for an account or offering to be on the growth or shrinkage list.

Due to a change of leadership in the cloud services organization and a reprioritization of resources our project ended before the sellers had a chance to fully engage with the growth and shrinkage predictions. However, our hope was that once the sellers 'experienced' the accuracy of our predictions and their value to them, it would be less important that they were able to reason about the patterns identified by the model. That said, in the initial stages of deployment, we believed it would be important that the models were 'explainable' in a language the sellers understood. This would be an important first step in allowing enough confidence in the results to act upon them.

Prediction is not Prescription: Differences Between Knowing and Acting

The second lesson we learned arose when sellers were faced with deciding what action should be taken when a client account was predicted to be at risk of defection in the next six month period. Even for the risk of defection model where sellers could reason about why an account was on the defection list, the model did not say why revenue was fluctuating in a pattern that predicted defection. As Lycett (2013) makes clear analytic tools can find patterns in the data, but without understanding the reasons for the patterns it is difficult to know what should be done to improve the situation. There were potentially many factors outside the model that were influencing a possible defection. Had there been a reorganization at the client company? Had the client started to use a competitor's services? Was there a recent major service outage? In addition, knowing what courses of action would best address the client situation were not informed by the model. For example, the client's organizational context or business climate, new or improved offerings in the vendor's portfolio of services, resources available to offer incentives to the client, and so on. The analytics pointed sellers to at risk clients, but they did not tell them what action to take.

In addition, sellers did not always understand why an account was no longer on the list for risk of defection. What actions had been taken, if any, to turn the account around? In an email exchange a sales executive expressed uncertainty about how to interpret changes in what accounts were on the defection list from one month to the next, asking *"Does this mean that the X accounts from the old list which don't show up again, did 'heal' themselves (either as it was planned variation in consumption or the account team managed to get it up again without defection analysis trigger) or the contract did end?"* The analytics gave no insight into what might have occurred to result in an account no longer being at risk of defection. Not only did this sales executive want to understand what might have cause a change in a client's risk profile, we too wanted to know what actions sellers might have taken and the impact of their actions had on changes in revenue. We explicitly tried to get this kind of information from the sellers, asking them about steps they had taken to address the risk of defection concern or other changes in the account status that might account for the improvement. Accompanying our risk of defection reports, we included questions to help us tune our model and in the future to provide recommendations about useful steps sellers might take to correct the situation (Figure 6).

Seller Responses				
Does the risk of defection make sense?	Was risk of defection previously known?	If known, what is the reason for possible defection?	Is there a recovery plan to stop defection?	Please describe the recovery plan, if one exists.

Figure 6. Questions to Sellers about Accounts on the Risk of Defection List

Regrettably, because there were no existing work practices that included recording such information, we received few responses back from sellers. Our hope was that if we could get this kind of information from sellers, overtime we would be able to improve our model and provide recommendations for successful strategies to turn around accounts.

ORGANIZATIONAL REALITIES

Organizational realities, outside the purview of analytic models, constrain the influence of analytic outcomes on actions taken by organizational actors. There is a myriad of factors that influence particular courses of action and at times these can run counter to the recommendations of the analytics. We confronted a number of these, including differing stakeholder priorities and organizational changes that had an effect on both funding for our project and the meaning of the historical data.

The Politics and Dynamics of Enterprise Analytics

As mentioned earlier one of our first tasks was securing access to cloud sales data. Extra safeguards were needed when using sales revenue data, requiring secure servers with tightly controlled access to them. There were also issues regarding who was allowed to see client data beyond the revenue figures. Non-disclosure agreements were in place for some clients restricting who was allowed to know about their relationship with the IT service vendor, what services had been purchased, or even that they had a contract at all. Fortunately for us, it usually did not include sellers from the vendor organization. Nonetheless we had to be careful about how we made our risk of defection and growth and shrinkage reports available, limiting access so that only those who had permission to see the information got access to it. As we got closer to making our reports routinely and more widely available, this issue became ever more pressing, with more scrutiny from within the company.

In addition, there were internal stakeholders (not necessarily the sellers) who had reasons to restrict how widely known it was that certain accounts or certain geographic regions risked, for example, declining revenue or opportunities for growth in particular services. Our predictions could potentially have an impact on career opportunities, how marketing dollars were spent, prompt additional oversight on sellers' activities, and so on. We could not know in advance or control all potential unintended consequences of our analytics, but if past experience was a guide, we knew we would likely hear from those who had concerns. In these cases, our ability to move forward on routinely producing reports would require weighing the benefits of doing so against possible risks. And if the risks were to powerful organizational actors, the evidence for the benefit would have to be very strong.

Organizational Change

During the course of our project there were significant organizational changes occurring in the cloud organization, including the departure of our initial executive sponsor. While we welcomed visibility for our project as it increased the likelihood of continued funding, we also had concerns that this visibility might put us in the cross-hairs of accountability for the performance of this business unit. At one point our executive sponsor measured the investment being made in our project against future cloud sales. From previous projects we understood how difficult it would be to show that the outcome of our analytics led directly to actions that reduced churn or increased sales. But if performance pressures mounted, we would need to be able to do so.

In a somewhat different way, changes to the cloud organization had effects on the meaning of our analytics. The cloud organization was 'carved out' of existing groups and the boundaries between the legacy groups and the new cloud organization was unstable. This meant that offerings were moving between organizations as was the responsibility and mandate to sell them. In this email (Figure 7) from our team, we inquired about one of these changes.

From: [redacted]
To: [redacted]
Cc: [redacted]

Date: 08/05/2016 07:53 PM
Subject: History of Accounts/Offerings in [redacted] BMDiv

Hi [redacted],
I'm working on an analytics project sponsored by [redacted] where we are developing predictive analytics for defection, growth, shrinkage and up-sell opportunities for our Cloud business. We have been asked to run the analytics on [redacted] data, but have discovered that [redacted] was recently created so we don't have enough historical data for our models. We are looking to trace the [redacted] accounts back historically so need to know the BMDivs where the [redacted] business/accounts lived prior to the creation of [redacted]. In talking with [redacted] and [redacted] today, they thought you might be able to provide this information. If it's easier to explain over the phone, let me know and I will set up a call.
Thanks in advance.
[redacted]

Figure 7. Email Query Regarding Business Division Alignment

These recent changes (creation of a new business unit) had an impact on where revenue was realized for particular offerings and for the relation between named client entities in the two different business groups. While we tried our best, realistically we could not keep up with all the organizational changes that were occurring and likely would occur in the future. Instead we had to 'assume' that these changes were not significant enough to undermine our analyses, taking some comfort in the continuing precision and accuracy of our analytics. However, this suggests that data scientists will have to stay connected to their analytics so they can make adjustments to data cleansing strategies, entity resolution schemes, and algorithmic choices that are responsive to organizational change.

GOING FORWARD

This case study should not be read only as a recounting of the challenges of enterprise analytics, but as a call for reflexivity among all those who participate in transforming data into organizational insights and action. At each step along the way from data curation to intervention there are choices that must be made, accountabilities that must be acknowledged, and consequences that must be considered. While we do not offer a set of 'best practices' for doing data analytics as each project will have its own exigencies, proficiencies, constraints, and timeframes; awareness of and reflection on the particular choices in play will contextualize results and make them more likely to deliver the desired impact.

Our project ended for a number of reasons before we were able to assess the full value of the analytics. Organizational changes meant that we had to (re)socialize our work and its value and convince our new stakeholders that it was feasible for our analytics to keep up with future organizational changes that inevitably would come. In addition, to ensure the analytics were responsive to these organizational changes there would need to be people inside the cloud organization with the expertise to carry on when the research team moved on to other projects. In the end, the calculation was made that these potential challenges outweighed the immediate and short term benefits of the analytics.

That said, those involved in this project, including the research team, learned a great deal about what it takes to deliver actionable analytics for the enterprise. Data analytics are often portrayed as offering ready-to-hand solutions for those with data and the expertise to put it to work. But our recent experience has humbled us and exposed us to a myriad of challenges, even obstacles, that must be navigated to realize the potential of enterprise analytics.

First and foremost, it will be necessary to design and align enterprise analytics with organizational 'sense making' (Hoy, 2018; Madsbjerg, 2017; McNamara, 2015; Weick, 1995) and employees' sphere of action. In this regard we also must consider, following Hovland (2011:33), that "organizational structures […] infuse […] numbers with power." The authority of numbers, in part, comes from establishing relationships among socio-material entities and as Power (1997) advises 'rituals of verification' that imbue analytic outcomes with their force.

It is also critical to recognize that enterprise analytics and measurement systems more generally (Muller, 2018; Strathern, 2000) are not neutral – they have real consequences for the lives of organizational actors. Furthermore, as Tallon et al. (2013) caution more appreciation is needed for how the increasing utilization of enterprise analytics will affect internal governance structures and accountabilities within organizations.

Finally, we must temper hype with organizational realities. As in our case, even for a company that develops and sells analytic systems and services, there are challenges to adopting them internally. Realizing the full potential of data analytics requires awareness of the technical and organizational complexity of acting on analytics in the enterprise.

Jeanette Blomberg is Distinguished Research Staff Member at the IBM Almaden Research Center in San Jose, California. Known for her research on ethnography in design processes, she currently is focused on organizational analytics where she considers linkages between human action, digital data production, data analytics, and business or societal outcomes.
blomberg@us.ibm.com

Aly Megahed is a research staff member at IBM's Almaden Research Center in San Jose, CA. With a background/expertise in operations research and analytics, his current research interests span over building analytical tools for complex service engagements, cloud computing, and IoT, and advancing research in AI and operations research.
aly.megahed@us.ibm.com

Ray Strong is IBM researcher emeritus. He is trained as a mathematician with a PhD in mathematics from the University of Washington. His research interests include machine learning and AI, most recently applied to the domains of IT services and energy management.
hrstrong@hotmail.com

NOTES

Acknowledgments: We wish to thank our sponsors, collaborators, and the cloud service sellers who went on this journey with us.

1. 'On premise' cloud services are hosted within the client organization's own data center, utilizing the organizations hardware and software rather than a remote facility such as a server farm.

2. 'Public' cloud services use a standard cloud computing model defined by the service provider or vendor and available to the public over the internet.

3. In the end, we decided not to pursue 'cross sale' analytics due to time and resource constraints.

4. We use the term *client*, *account*, and *client account* interchangeably to denote specific clients, with each client having an associated portfolio of specific service offerings currently being received from the vendor.

REFERENCES CITED

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli.
2018 Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems ACM.

Abhinav, M., A. Megahed, R. Strong, J. Blomberg and A. Elwany
n.d. Optimizing Predictive Precision for Actionable Forecasting of Revenue Change from Clients Receiving Periodic Services. *European Journal of Operational Research*.

Anderson, K., D. Nafus, T. Rattenbury and R. Aipperspach
2009. Numbers Have Qualities Too: Experiences with Ethno-Mining. In *Proceedings of EPIC*. Washington, DC: American Anthropological Association.

Barley, S. R., & P. S. Tolbert
1997 Institutionalization and structuration: Studying the links between action and institution. *Organization studies*, 18(1), 93-117.

Blomberg, J.
2008 Negotiating Meaning of Shared Information in Service System Encounters. *European Management Journal* 23:213-222.

boyd, d., and K. Crawford
2012 Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15.5: 662-679.

Cefkin, M., J. Owensby and J. Bomberg
2007 The implications of enterprise-wide pipeline management tools for organizational relations and exchanges. *Proceeding of Group*, Sanibel Island, FL, US, ACM.

Chen, H., R. H. Chiang and V. C. Storey
2012 Business intelligence and analytics: from big data to big impact. *MIS Quarterly* 1165-1188.

Davenport, T. H. and J. H. Harris
2007 *Competing on Analytics: The New Science of Winning*. Harvard Business School Press, Boston, MA.

Davenport, T. H., J. H. Harrisand R. Morison
2010 *Analytics at Work: Smarter Decisions, Better Results*. Harvard Business Press, Boston, MA.

Elish, M. C. and d. boyd
2018 Situating methods in the magic of Big Data and AI. *Communication Monographs*, VOL. 85, NO. 1, 57–80. https://doi.org/10.1080/03637751.2017.1375130

Fiore-Silfvast, B. and G. Neff
2013 What We Talk about When We Talk Data: Valences and the Social Performance of Multiple Metrics in Digital Health, *EPIC 2013 Proceedings*, pp. 48-62, American Anthropological Association.

Gillon, K., S. Aral, C-Y. Lin, S. Mithas and M. Zozulia
2014 Business analytics: radical shift or incremental change? *Communications of the Association for Information Systems* 34(13), 287–296.

Gunasekaran, A., et al.
2017 Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research* 70: 308-317.

Guyer, J. I.
2014 Percentages and perchance: archaic forms in the twenty-first century. *Distinktion: Scandianvian Journal of Social Theory* 15:2, 155-173.

Hovland, J.
2011 Numbers: Their Relation to Power and Organization. In *The Mutual Construction of Statistics and Society*. Saetnan, A. R., Lomell, H. M. and Hammer, S. (eds.), Routledge, NY.

Hoy, T.
2018 Book Review: The Paradox of Sensemaking, *EPIC Perspectives*. https://www.epicpeople.org/book-review-paradox-of-sensemaking/

Levenson, A.
2011 Using targeted analytics to improve talent decisions. *People and Strategy*, 34(2), 34.

Lycett, M
2013) 'Datafication': making sense of (Big) data in a complex world. *European Journal of Information Systems* 22(4), 381–386.

Madsbjerg, C.
2017 *Sensemaking: The Power of Humanities in the Age of the Algorithm.* Hachette Books, UK.

McNamara, L.
2015 Sensemaking in Organizations: Reflections on Karl Weick and Social Theory. *EPIC Perspectives.* https://www.epicpeople.org/sensemaking-in-organizations/

Megahed, A., Ren, G.-J., and Firth, M.
2015 Modeling business insights into predictive analytics for the outcome of IT service contracts. In *Services Computing (SCC), 2015 IEEE International Conference,* 515-521.

Megahed, A., Yin, P., and Motahari Nezhad, H. R.
2016 An optimization approach to services sales forecasting in a multi-staged sales pipeline. In *Services Computing (SCC), 2016 IEEE International Conference*, 713-719.

Megahed, A., Gajananan, K., Asthana, S., Becker, V., Smith, M., and Nakamura, T.
2016 Top-Down pricing of IT services deals with recommendation for missing values of historical and market data. In *International Conference on Service-Oriented Computing*, 745-760.

Miller, T.
2017 Explanation in artificial intelligence: insights from the social sciences. arXiv preprint arXiv:1706.07269.

Muller, J. Z.
2018 *Tyranny of Metrics*. Princeton University Press. Princeton, NJ.

Moore, G.
2011 Systems of Engagement and the Future of Enterprise IT: A Sea Change in Enterprise IT. AIIM Whitepaper.

Power, M.
1997 *The Audit Society: Rituals of Verification.* Oxford, UK: Oxford University Press.

Ribeiro, M. T., S. Singh, and C. Guestrin,
2016 Why Should I Trust You?: Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1135–1144, 2016.

Sharma, R., S. Mithas, and A. Kankanhalli
2014 Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems* 23.4: 433-441.

Strathern, M.
2000 *Audit Culture: Anthropological Studies in Accountability. Ethics and the Academy*, London: Routledge.

Tallon, P. P., R. V. Ramirez and J. E. Short
2013) The information artifact in IT governance: toward a theory of information governance. *Journal of Management Information Systems* 30(3), 141–177.

Weick, K. E.
1995 *Sensemaking in organizations.* Vol. 3. Sage.

Weick, K. E., K. M. Sutcliffe, and D. Obstfeld
2005 Organizing and the process of sensemaking. *Organization science* 16.4: 409-421.

Zhong, R. Y., C. Xu, C. Chen and G.Q. Huang
2017 Big data analytics for physical internet-based intelligent manufacturing shop floors. International journal of production research, 55(9), 2610-2621.

Zuboff, S.
1985 Automate/informate: The two faces of intelligent technology. *Organizational dynamics* 14(2), 5-18.

Zuboff S.
1988 *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, New York.

Regarding the Pain of Users: Towards a Genealogy of the "Pain Point"

DAVID PLATZER
Berggruen Institute

This essay offers an analysis of the "pain point," a commonplace figure of speech in UX design and contemporary business contexts more broadly. By situating this everday trope within a wider discourse of pain, and its politiciztion in the United States, I seek to problematize the modes of relationality and forms of care entailed in the practice of design research. Ultimately, I will argue, while the "pain point" can be an effective tool for communicating with stakeholders and fomenting alignment about research objectives, it also implicates the more troubling ethical dimensions of applied practice. Through a narrative account of an innovation focused ethnographic research project conducted within the design unit of a major tech company, I argue that questions of solidarity, and its contemporary aporias, can be obscured by the humanitarian rhetoric of contemporary design praxis; a rhetoric of which the "pain point" is a prime example.

INTRODUCTION

The "pain point" is a commonplace figure of speech in UX design and contemporary business contexts more broadly. In some corners of the design world, it is nearly ubiquitous; invoked so readily that it can betray a near naturalized quality. But what does "pain point" mean? In its most general sense, a "pain point" refers to a customer or user problem, one that a new product, service, or feature set might profitably "solve" or "resolve." Given their complexity, within many design research contexts, ethnography, in its inductive, open-ended orientation, has been conceptualized as an exemplary means of identifying "pain points" in all their knotty ambiguity. Through its "generative" or qualitative form of inquiry, ethnographers, so the thinking goes, can "uncover" user pain that surveys and other quantitative methods cannot. Consequently, within the field of UX design in particular, this popular trope is often implicated in the ways that ethnographic research projects are articulated, evaluated, and deliberated by researchers and stakeholders alike.

In what follows, I juxtapose a critical, genealogical analysis of the "pain point" with a narrative account of a six-month ethnographic research project conducted within the design unit of a major tech corporation. In this "innovation" focused study, the "pain point" served as a boundary object uniting stakeholders in product management, marketing, and design research, allowing for mutual intelligibility about the scope, method, and underlying objectives of the study (Star and Griesemer 1989). As a periodically conflictual project, the "pain point" also became a weathervane for disputes about research technique and the kinds of findings it generated. At first a study of American workplace collaborative practices, and the lives of the "digital workers" who perform them, by the end it had become an investigation of conflict amongst coworkers. However, in unsuccessfully uncovering any actionable "pain points," this ambitious study was in the end a failure. Nonetheless, in this essay, such failure provides an opportunity to reflect on the concept of "pain point" itself and its implications for our practice.

In devoting such concentrated attention to the "pain point" herein, my aim is to interrogate the modes of mediated *relationality* that UX researchers enact between "users" and stakeholders. Given the theme of this year's conference, it is also a meditation on the evidentiary forms upon which such relations are based (for instance personas, interview snippets quoted in PowerPoint decks and photographs of users). Ultimately, by situating the "pain point" within a broader discourse of pain, and its politicization in the United States, I will argue that the "pain point," though an effective tool for design researchers, can also obscure the aporias or impasses of social solidarity they face. By perpetuating the notion that product design, development, and marketing constitute a form of care, indeed a peculiar mode of welfare, through a language that also draws on a humanitarian rhetoric of aid, the "pain point" can perpetuate the subtle forms of estrangement by which the socio-economy produces and reproduces itself. In an analysis that draws on Susan Sontag's *Regarding the Pain of Others* (2003), a work which questions the evidentiary status of humanitarian photography, as well as the mass mediated mode of alienated sociality which it instantiates, I argue that this ostensibly throwaway figure of business slang thus also encompasses a more troubling set of ethical problematics. In its conclusions and orientation, this essay thus offers a different take on Amirebrahimi' excellent "The Rise of the User and the Fall of People: Ethnographic Cooption and a new Language of Globalization" (2016) and Madsbjerg's 2014 EPIC keynote, "Happy Birthday, Now Grow Up," a talk that provocatively suggested EPIC "divorce design."

THE VALUE OF THE PAIN POINT

For ethnographic design researchers, the "pain point" can be a vital resource in navigating organizational politics. In promoting the value of their practice to stakeholders often suspicious about the utility of qualitative research and its "anecdotal" or "subjective" forms of evidence, the "pain point" can be performative of their particular mode of expertise. In context, it can highlight the business objectives of ethnographic praxis and its unique role within the broader product design and development lifecycle. In this sense, in a perlocutionary fashion, its invocation can serve to differentiate ethnography, and qualitative research more broadly, from other common research methodologies. It does so by locating problems within the user's own lifeworld, in turn ensuring they will prove lucrative ones to solve via new products, features, or updates to an existing application. By contrast, other more overtly evaluative UX research methods – for instance, usability testing, surveys, and heuristic design auditing – are techniques whose procedural, etic form can more readily lend them an aura of objectivity. However, by the same measure, their highly-scripted approach can be limited in finding valuable "new" user pain, pain that is often obscured by its obviousness.

The "pain point" is thus for many ethnographic UX researchers a favored figure of speech. It encapsulates the value proposition of their approach by presenting it as a mode of inquiry uniquely suited to revealing problems that are critical for the user, if easily made opaque by their everydayness. In this sense, the "pain point" can often serve as an elementary form of a wider legitimizing or justificatory discourse, a discourse of which EPIC is an institutional example. At the same time, the "pain point" is a term of art, a figure of speech that in practice allows ethnographers to communicate with stakeholders about the methodology, objectives, and deliverables of the studies they are pursuing. In so doing, it works to enact a mode of relationality between the user and stakeholder (the designer or product manager, product marketing manager, and so on); relations whose management is thus understood as the remit of the UX researcher.

By way of illustration, the following reflections are from a series of interviews I conducted with ethnographic UX researchers in the SF Bay area, including those who work in-house within the design research organizations of major tech companies as well those who work at smaller consultancies symbiotic or adjacent to them. The first interview snippet is from an ethnographer working as a UX researcher in the design unit of a prominent SF start-up. In a conversation about the ways ethnography and qualitative research functions within his collaborative workflows, I asked this researcher whether he uses or encounters the "pain point" in his professional life:

> Let me think about my use of "pain point." I'm pretty sure I used it in a two-line description of an upcoming research project just today. Actually, come to think of it, I use it all the time: to mean "a bad experience for the user," "something unnecessarily painful," and, importantly, "something that we should fix." There is a call to action in the phrase, a movement towards improvement, an intention to resolve. It's also a useful shorthand for demonstrating the usefulness of UX Research: here is your friendly local researcher, who is able to identify bugs (bad experiences or "pain points") in the user experience that we did not previously know about. And once we know, we can fix them, thereby improving the user experience and, by extension, the product. It's a phrase that I and my colleagues throw around quite often, to explain what we are doing and why it is useful. I think the word "pain" is the crucial part, because everyone can agree that pain is not something we want our users to experience, if we can possibly help it.

In this sense, as suggested in the introduction, and as my interlocutor demonstrates in his reflections, the "pain point" not only serves a legitimizing role for research, it also acts as a boundary object (Star and Griesemer 1989) within collaborative workflows. In allowing for mutual intelligibility between UX researchers and their stakeholders, the fact that "everyone can agree…pain is not something we want our users to experience" helps to ensure that cross-disciplinary teams are aligned on research objectives and mutually recognize its value to the product. In other words, boundary objects entail forms of strategic ambiguity, whose "plastic" quality allows for "interpretive flexibility" and a "tacking back and forth" between different levels of "structuration" (Star 2010). In the case of product design, identifying, naming, and resolving to fix "pain points," mediates disciplines, helping teams to move between business objectives (which may be well defined, or highly structured) and some concept of the user; a category whose underlying heterogeneity and ambiguity entails a more diffuse and elusive kind of structure (Amirebrahimi 2016). In this sense, the "pain point" both points to something ostensibly concrete or "real," while doing so with a level of abstraction or vagueness that allows for it to be meaningful within the multi-perspectival framework of a team and its often-divergent assumptions about functionality, usability, and the underlying value proposition of their product.

However, for some researchers, despite such utility, they find themselves expressing reservations about just how pervasive the "pain point" has become and the ways in which they themselves invoke it. The following snippet comes from an interview conducted with an ethnographer and UX researcher working within a small SF design consultancy. Reflecting on the value of the "pain point" in her working life, she nonetheless expresses ambivalence about its implications:

> We do a lot of journey mapping, which means "pain points" come up frequently: little barriers to enjoyment or efficiency, knots of frustration and confusion, along a longitudinal experience. We often mark pain points on pink post-its (closest to red possible), and prioritize them with round red stickers. We have sometimes overemphasized them with little frowny or angry faces drawn on the post it or sticker. You could easily argue that a term like "pain" in UX conversations is excessive. Is it really "painful" to have to enter your birthdate in an app using a clumsy menu? Or to take a click or two extra to purchase a product online? But there is something productively visceral about using "pain" that commands stakeholders' attention in a way that other terms don't. Maybe it's a subtle technique to engender empathy - the natural reaction of seeing someone in pain is to help them. Likewise, the idea that it's a "point" - a sharp little jab - implies that it can be blocked, or parried, or absorbed...fixed. This helps rally resources to the cause, but also masks types of "pain" that are chronic and systemic.

Much as with the first interview, this interlocutor also highlights the performative, persuasive quality of the "pain point." In eliciting a "natural reaction" of sympathy from stakeholders, it is an effective tool to "rally resources." But this interlocutor also worried that in labeling all user problems painful, certain forms of more "chronic and systemic" pain can be "masked." In other words, if all user problems are understood as painful, how can we distinguish between different kinds of unpleasant experience and through more precise scales of severity, significance, and economic opportunity? In this light, as this interviewee herself implies, "pain" can thus be a barrier to a more variegated language of affect and problematization. However, by the same measure, there is a structural bind implicit in her concerns. Discussing user "irritation" or user "frustration," for example, might not have the same "visceral" quality and thus might not "command stakeholders' attention" in the way pain indubitably can. In other words, presenting evidence of "pain" (a pink post-it note approximating a blistering red) is always going to be a more effective tool for cultivating empathy than a more nuanced, if less immediate language of feeling and problem. A pale-yellow sticker marking user impatience simply doesn't make the same kind of claim on a stakeholder as bright pink pain.

At this juncture, we should pause to question a key concept that this interviewee raises; one that, like the "pain point," is also pervasive in UX design. Namely, "empathy." What exactly does she mean when she says "empathy" and what kind of relation is entailed in a form of it mediated by pink post-it notes? Within academic anthropology, empathy has long been a term out of favor. For instance, critiqued for its projective, universalizing assumptions by Geertz (1984), amongst contemporary anthropologists, empathy has been understood as analytically weak because its fundamental ambiguity. It has also been challenged for the way it can presume an immediacy to intersubjective experience and the circulation of feelings between individuals, an immediacy that anthropological analysis soon shows to be suspect. (Throop and Hollan 2008).

However, by contrast, within the design world, this very ambiguity, as well as the intimation of immediacy is strategic and useful. In this context, "empathy" marks the forms of "shared feeling" between a designer (or product manager) and their inscribed user, but a kind of sharing that carries a strictly delimited set of relational demands. After all, the obligation to "empathy" only goes so far as those problems (like wonky menu bars) that can be solved through design. In this sense, user "empathy" is thus only ever partial and contextual. And as this empathy is typically mediated by the work of a UX researcher in their interactions with "real" users, it is the obligation of the researcher to present "pain" in such way that it elicits a specific kind of reaction, one with concrete operational ends. In this sense, pain and empathy alike are produced by the researcher, creating a kind of experience for the designer that, in turn, inspires them to create a kind of "caring" or "empathetic" experience for the user. In this sense, pain and empathy alike become key figures in a series of terms that link the developers of a product to their consumers. But how does such a concept of empathy (and the demonstrable pain points a researcher produces to cultivate it) fit within a broader American vocabulary of feeling, intersubjectivity, and care? In other words, how does the micro-culture of UX design in the United States relate to the larger social contexts within which it is situated. To answer this question, we will have to turn more directly to the problem of pain and its recent American history.

EVIDENCE OF PAIN

In *Pain: A Political History* (2014), the historian Keith Wailloo examines how, beginning with the Reagan revolution of the early 1980s, pain became a key signifier in national debates about the role of the state in American society. Noting the widespread conservative derision that, as a candidate, Bill Clinton encountered in 1992, when he famously said, "I feel your pain" in the second presidential debate, Wailloo demonstrates that this was but one memorable example of a far more encompassing politicization of pain and questions about the "sympathy" it can elicit. In fact, as Wailloo argues, "the question of pain…refined the very meaning of 'conservative' and 'liberal' as keywords in the American political vocabulary (203). While conservatives, following the example of Reagan, articulated their principles of governance via a skepticism about suffering, and what Waillo calls a "cold objectivity" about claims to state care made on its basis, liberals argued that "social policies should be driven by compassion towards…pain" (6). On the one hand, "liberals" articulated their social agenda through a language of care, one premised on the notion that all subjective pains are valid. On the other, for conservatives, governmentality founded on a principle of compassion erodes the self-determination and freedom they viewed as constitutive of the American project in its industrial zeal and entrepreneurial values. As such, "problems of pain and social welfare came to define American politic theater" (203)

It is in this sense that Wailoo argues that the very meaning of "liberal or conservative became ideologically solidified around the problem of pain" (203). Such political consolidations, indeed polarizations, as he demonstrates, were consequent to the way pain became a symbol of welfare and thus a critical site for debates about its legitimacy in relation to the social contract. In the crossfire of such disputes, not only did pain become a site of political struggle, but it also gave rise a new set of questions about the "truth" of pain and the problem of its verification, evaluation, and representation. In other words, the pain in question generated a kind of politicized skepticism.

As Wailoo notes, "Pain's reality – which pains were real, which were false, and who was the best judge" of their veracity became questions of national import. What followed was a range of policies enacted to "measure distress and…define the right to relief" (7). In a similar vein, beyond the concrete development of policy, questions about "which kinds of pain are real, which warrant sympathy, and whose plight is deemed legitimate…provoked constant political spectacle (203). For conservatives, the figure of the "disabled person in chronic pain became a symbol of much that was wrong with liberalism – it's gullibility, its support for government dependence, and its embrace of welfare at the cost of hard work" (ibid). In associating welfare with dependency, and its erosion of the principles of autonomy upon which America was ostensibly founded, pain became a battleground for a debate about our collective values. Not only a principle of public policy, welfare became central to questions about the obligation of citizens to work. Phrased otherwise, pain became central to the problem of the social and our obligations to one another as members of a national community.

In *From Manual Workers to Wage Laborers: Transformations of the Social Question* (2003), Robert Castel demonstrates how, over the course of the twentieth-century, the notion of welfare became central to questions about the role of the state in a capitalist economy of labor. Through a genealogical analysis of the institution of wage labor, one that reaches back to the dawn of industrial modernity, Castel excavates the way in which welfare become "the form — if indeed the mutable form — assumed by the compromise between the economic dynamism demanded by the quest for profits and the sentiment of protection necessary for social solidarity" (193). In the waning years of the twentieth century, such questions of solidarity, and their limits, were increasingly arbitrated by what Castel calls the discourse of "handicapology." By "handicapology," Castel indicates the constellation of administrative procedures and moral technologies that aimed to distinguish those who rightfully demanded the assistance of the state because they were unable to work, versus those who volitionally "chose" not, thus meriting their rejection. In this way, through the diffusion of such discourse and its operationalization in policy, the question of care became linked to a form of procedural, bureaucratic skepticism about pain.

In the United States, the Clinton administration's comprehensive overhaul of the American welfare system sought a "third way" through such polarization; or what Wailoo has described as our collective "political spectacle" of pain. In a sublation of left progressivism, and its recognition of all pain as valid, with conservative suspicion about its underlying reality and threat to self-determination, the Personal Responsibility and Work Opportunity Reconciliation Act (1996) created a new model of "handicapology." Limiting the scope of state care, it nonetheless did so in a reconstituted language of pain and compassion. The new moral vocabulary that ensued rearticulated state restrictions on welfare in an erstwhile "liberal" language of sympathy and care. In this sense, it served as a precursor and hinge to the "compassionate conservatism," which, in unusually crisp turn of phrase from the otherwise tongue-twisted mouth of George W. Bush, was described as follows in remarks he delivered to a San Jose audience in April 2002:

> Government cannot solve every problem, but it can encourage people and communities to help themselves and to help one another. Often the truest kind of compassion is to help citizens build lives of their own. I call my philosophy and approach compassionate conservatism. It is compassionate to actively help our fellow citizens in need. It is conservative to insist on responsibility and results. And with this hopeful approach, we will make a real difference in people's lives… We are using an active Government to promote self-government… The measure of compassion is more than good intentions; it is good results.

Through the rhetoric of Clinton and Bush, care, in the form of "compassion," was no longer indexed to the provision of services and state assistance to the poor and infirm. Rather, it became a rationale for policies that, in "helping" to facilitate work for all, promoted the universal dignity of American citizenry by making work a "right" worth protecting through enterprise. In this sense, solidarity took on a new valence, one in which pain, no longer polarized, instead became an expression of the moral assault occasioned by unemployment and dependency more broadly. As a result, capitalism was no longer conceptualized as a threat to the social, but rather its ethical guarantor.

There are several points worth flagging from this extended digression into recent American political history. First, that pain, beyond its status as a feeling, has also been a symbol of the relationship between the state and those in need or those who are suffering. And in this way, it has become a byword for questions about the social and their impact on interrelated problems of sympathy, compassion, relief, and self-determination. While in the 1980s, sympathy and self-governance were understood as mutually exclusive options, at the turn of the millennium it became possible to see pain relief and free enterprise as co-constituting. That is not to suggest that every time a design researcher describes a user "pain point," they are implicitly taking a side in this political spectacle, but rather that underlying notions of sympathy and relationality have been conditioned by this broader context. Notions like empathy, compassion, sympathy, and pain, as well as he social imaginary they index are discourses that influence how we make sense of our world historically. In this sense, while a UX researcher who labels a research finding or user perception a "pain point" clearly doesn't mean to stake a claim on welfare or the status of the state, the very meaning of pain in our discourse helps to shape the moral logics entailed. We solve a "pain point" to return the worker to their state of productivity. We resolve pain to promote the autonomy and self-determination of the consumer.

The Birth of the Pain Point

So how and why did the pain point emerge as a figure of speech in design and business contexts? Perhaps unsurprisingly, given the speculative social history I have been sketching herein, the "pain point" began to appear at the same moment Bush was delivering his 2002 remarks on "self-government" in San Jose. Originally, an esoteric term from the discipline of pain management, in the early 2000s, "pain point" became a figure of speech in the business world, just as, under the leadership of Bush, the resolution of social pain became, in its new guise, both an ethical and capitalist imperative. However, such a history of pain, and its semantic evolution, tells only one part of the story. The "pain point" also emerged as a reaction to economic factors specific to the tech industry and its excess valuation at the turn of the millennium. But understand this history, however, will demand further speculation.

As mentioned earlier, the pain point is a term that originates in physiology. In medical discourse and practice, it refers to the physical locus of an undesirable sensation caused by intense, distressing, or harmful stimuli. For the physician, locating the pain point, or the point of pain, is a diagnostic activity. However, in seeking to understand the nature of the offending stimulus and, if an external entity, like a splinter, to remove it, or if an endogenous one, like an impinged nerve, to treat it, the ultimate objective of such diagnostic work is to heal. A pain point or point of pain in physiological terms is where pain is either most intense or the place in the body from which it emanates. Its identification is diagnostic; a form of understanding whose aim is to alleviate physical suffering.

How does this physiological definition, or etymology, relate to the more figurative expression that one finds today in UX and the corporate world more broadly? In the late 1990s, e-commerce marketers began using both "pain point" or "point of pain" as a way of articulating the problems that a customer or software user encountered in their working life. It was a relatively rare figure of speech, used intermittently and irregularly, and appearing in business publications and its associated discourse only infrequently. Starting in the early 2000s, however, and picking up in speed around 2004, "pain point" became an increasingly common trope both in practical contexts and in the business press, through articles in journals like the *Harvard Business Review*. Why did this happen? And why then?

As suggested, answering this question requires a bit of speculation or what our colleagues at ReD call abductive reasoning. My inferential hypothesis is that this spread was coterminous with the popularization of "user experience" discourse and symptomatic of the same underlying economic factors that spurred its popularization. A term first coined by the cognitive psychologist Don Norman, "UX" shifts the focus in software development from technological capabilities to concrete use through the development of an easeful interactivity via "empathy" enabled design. Its attendant research practices, and the implicit reconceptualization of technical and economic innovation it entails (innovation as a result of "solving" customer problems), emerged as a reaction to a moment of crisis. "UX" was not invented in the wake of the crisis, as Don Norman himself would be the first to insist, but rather became widespread because of them.

So, why did customer problems become so important at this point in the early 2000s? Although this would demand research outside the scope of this conference presentation, herein I argue that it was in response to the first dot-com bust and an industry that grew on the basis of technical innovation (the internet) unanchored to consumer desires, behaviors, or needs. In reconstituting itself in the wake of excessive speculation and the demise of many once promising enterprises, a new spirit of market cautiousness became normative. While the bubble was fueled by the availability of a new technology and excitement about its commercial potential, the industry that emerged in the wake of the bust paired the technical with the practical, recognizing that new "affordances" were not always economic opportunities if they didn't "solve" an explicit customer problem. In short, people will not purchase a new product unless it does something for them, no matter how impressive the technological novelty.

In this fashion, the user or the customer (or rather projective formulations thereof), became the arbiter of an idea or product. In this new discursive and practical milieu, one careful about the uncertain profitability of technical innovation, and thus now anchored to a consecration of "use-value" and usefulness, the "pain point," a relatively rare term hitherto, became a way of evaluating and articulating utility or presumed utility. In this sense, it became a figure for problematization, or the articulation and identification of real world, market worthy problems; part of an assemblage of practices and symbols oriented towards to the production of value through scales of usefulness and concrete practicability.

In the following decade, as a once eclectic group of academics, designers, and engineers, like Don Norman and Alan Cooper, found their "UX" framework becoming mainstream within the tech industry, the "pain point" too became pervasive. It did so because, as suggested earlier, it was a kind of boundary object allowing for different disciplines (product management, product marketing, UX design, business units) to share a common (if generatively inconsistent) understanding of the "user." Once more, for those unfamiliar with the term, a boundary object is one whose structural ambiguity allows for both divergence and convergence in discourse and interpretation (Star and Griesemer 1989). Divergence in the sense that though we may use the same term, say "pain point" or "user" we may have different referents and models implicit in our invocation. Convergence in that there is nonetheless enough semantic and referential overlap that such divergences do not stand in the way of mutual intelligibility and the practical "alignment" it allows. In short, a "boundary object" is discourse that allows for difference in perspective while still promoting collaborative consensus.

Around 2004-2005, when the "pain point" became more common in design contexts, asking about user or customer "pain" served as a way of checking untested or unvalidated assumptions about the end-user and their needs. Such questions set the ground for the increased viability of research practices forged in the service of such validation. It was no longer enough that a new product, service, or feature should be "cool," impressive, or technically sound, it must also "relieve" user "pain," because if not, the product would not be commercially viable. Or if successful, only because of precarious speculation founded on the basis of technology not consumer behavior. In this sense, developers, product managers, marketers, designers, and researchers became physicians of a sort; translating technologic novelty into the development of digital analgesics. These new commercial therapeutics were expressly created to relieve pains occasioned by the modern world and the increasingly complex ecosystem of tools its immaterial knowledge workers had to master to flourish within it.

In this light, the "pain point" today is most commonly invoked the context of work-related software contexts. For consumer level goods and services, my interviews with UX researchers revealed, it does not appear as frequently. For instance, while at Adobe, where I worked for two years, it would be difficult to spend more than an hour without hearing "pain point" in a meeting. At Facebook or Twitter, I've been told that while it may be invoked occasionally, it does not have nearly the same ubiquity. In this sense, this is precisely how "pain point" framed in the popular and influential *Value Proposition Design: How to Create Products and Services Customers Want* (2014) Here a "pain point" is linked to problems explicitly related to work and the successful performance of one's professional responsibilities: "Pains describe bad outcomes, risks, and obstacles related to customer jobs." (9) Thus, asking, "but what's their pain point" became a way of saying, "is this useful and if so how? Does this solve a problem in the user's working life?" Such questions affirmed the value of design research practices and at the same time became an idiom that allowed for researchers to define their role and place within the product development lifecycle. For this reason, as suggested earlier, the "pain point" has consistently been popular amongst researchers who find it useful in communicating with stakeholders and articulating the goals of a study. In turning now to a case study from my own work as a UX research, I hope to interrogate the themes of what has preceded through a detailed example.

The End of the Road

Towards the end of a long, languid Bay Area summer, the remaining members of the team congregated by videoconference to have one last conversation before the project finally concluded. On the line were three product managers, an external marketing consultant, a mixed-methods user researcher, and myself, the ethnographic lead on the project and the primary field researcher. By that point, we were a battered bunch, as the project had taken many twists and turns over the previous six months of work. Personnel had shifted several times and the original three-month timeframe had doubled by this, the last day we were assigned to the project before long-delayed vacation. Some of this exhaustion was a legacy of the project's early phases, when contradictory approaches and discordant objectives had made fruitful collaboration amongst the team a recurrent challenge. Most of this discord was structural and principled. Thankfully, exceedingly little was personal or temperamental. Even after the most heated debates, interactions remained cordial and professional. Yet despite such civility, tensions remained persistent.

The team was principally comprised of a novice ethnographic researcher, fresh from a PhD program in anthropology (the author), and a more seasoned corporate researcher, whose background was in psychology, information architecture, and mixed-methods user research. They were paired with a product manager and product marketing manager who represented the interests of their key stakeholders. Each researcher represented a distinctive methodological orientation. And they each employed different field and interview techniques. Neither entirely approved of the other's approach. As an ethnographer, I resisted the more structured, procedural UX approach of my colleague, Chris (a pseudonym). By the same token, Chris found my ethnographic orientation overly personal and poorly equipped to uncover the explicitly technical problems our company could solve. In the first weeks of research, our divergent approaches quickly mapped on to an ongoing disagreement between the marketing and product teams who were our primary stakeholders. In short order, I quickly made allies with a product marketing lead assigned to the project and Chris aligned with the product team he had worked with for many years. While marketing was interested in "motivation," product by contrast was concerned with what they alternately called "needs" and "pain points." Sides were drawn, and as research began, our daily debriefs continually served to arbitrate these disputes.

Initially the design research team of which Chris and I were part had been commissioned to provide persona level insights for a large-scale segmentation study that quantitative market researchers were launching later in the year. This project was spearheaded by a product marketing team who had wanted an observational, ethnographic method. But before the work started, a product team asked for the research to also include "innovation" or "new opportunity" findings for a well-established tool. Product claimed method agnosticism, though the key stakeholders many times said their goal was to explicitly "find new pain points that aren't obvious." Via negotiations whose inner machinations were never made entirely apparent to the researchers, management had decided these two projects could indeed be consolidated into one and pursued in tandem. Researchers were left to mediate the resulting tension.

Thus, a circuitous itinerary ensued. What was the project about? Alas, details need to remain opaque to protect privacy and confidentiality. But at a broad level it involved the way American "digital workers" collaborate with one another in (and out of) offices. To understand this problem, and in doing so to develop both persona-level insights and discrete pain points, a first phase comprising fifty 60-90-minute interviews was conducted by phone and Skype. Following a month of analysis, this was followed by a second phase that entailed two, week-long, in-person workplace ethnographic studies I conducted independently. The first study was at a major health focused non-profit in Washington DC. The second was with a large financial services corporation in the Midwest.

After both phases, nobody on the team was confident that our findings were immediately actionable, even if the final ethnographic study had generated a few promising leads. That is, while we had found various aches and pains, nobody involved was convinced they were the elusive "pain points" that senior product leadership had explicitly asked for. However, in the final days of our collaboration, we put aside our differences and trepidations, and collectively developed a taxonomy of "use cases," framed through a series of explicit "pain points." We hoped they would prove sufficient for the innovation work that a design team would lead subsequently, while we were on vacation recuperating.

In our final meeting as a team, I asked my colleagues to briefly reflect on the project that was concluding:

"If we're finished talking about everything we need to discuss today, can I ask you all a favor?" I said. "I'm doing a paper for a professional conference, EPIC, and…well…can I ask you all to define what you mean when you say 'pain point.'? It has been such a central term for us throughout. Now that we're done, I'm still not sure I totally understand what it means. Can you all help me?"

A few smirks appeared on my screen in response and then utter silence. Finally, a junior product manager responded:

"Oh wow."

"Why don't I start with you, then," I responded jocularly.

"Let me think then," he said, laughing. "Actually, I think I've spoken too much already. Someone else go ahead first"

"Don't over think it," I said. "Just explain what a pain point means to someone who doesn't know what a pain point means, like me sort of, since I'm still new to this world, but somebody who had never even heard it before," I added, followed by more silence.

"I don't know how to do it concisely," another, senior product manager said eventually.

"Don't worry about concise, then, we still have ten minutes," I responded.

"Ok, um. I guess it's something that takes a lot of effort. Or, uh…Yeah, maybe it takes a…" he hesitated and paused for a few moments before continuing. "It takes a lot of effort or prevents me from being, um, efficient at doing what I need to get done."

"And why do we call that 'pain'?" I asked in response.

"Well, that's why I was trying to, um…" he paused again. "I don't know, I think it's an emotional thing. It's painful because I'm doing something that I believe doesn't need to be done or could be done in a better way."

"Really interesting. Thank you. Alright, you're next," I said to the marketing consultant.

"Oh, so when I think about a pain point it is something that takes, it takes…it's an inefficient way of doing something for me. Like there is a problem. It…it makes me uncomfortable. It gets me, it gets me…It makes me unproductive. And it's something that I'm looking to solve for. Like it's a problem that occupies my mind and my work and I want to find a solution to it. To me, that's a pain point."

"Awesome, next," I said to the junior product manager, who had originally passed the question off to his colleagues.

"I think a pain point is something I don't want to do or don't like doing but I have to do. And I'm very open to trying out new things or new ideas in order to have it either completely eliminated or mitigated."

"Great, alright. Excellent. Fellow researcher, you're last. Do you have any thoughts?" I said to Chris, my mixed-methods user research colleague from design research.

"Yeah, I think of a pain point as something that blocks me from achieving my goal and leads to an unpleasant experience. I usually think of them as things that cause people some kind of discomfort. And, yeah, pain is kind of an extreme word. But if I was talking to someone outside the company, I'd say a pain point is something that blocks someone from achieving their goal. It's kind of the opposite of delight. You know, designers are always like, 'oh, we want to design a delightful experience.' Well, pain is the opposite. It's like, this is really unpleasant. But the problem is that pain point doesn't really have a level of severity associated with it. Like I'd like to say a 'pain point' is a problem that triggers someone to look for a solution. The reality is that we see lots of people with inefficiencies and what we would call pain points. But they seem to be OK with that. Because they either don't know there's a better solution. Or they're just too busy. It's not important enough of a problem for them. There's not enough at stake. So, I don't know, I guess I just think it's something that's preventing people from achieving their goals. And it's probably, I don't know guys, what do you think? For me, it's kind of like our perspective looking at them. Like, I would love to say, pain points are usually self-acknowledged pain points. But I feel like, sometimes we look at what people are doing and say 'wow, I can't believe they are not complaining about that, that seems like a real pain point to me, since it's so inefficient.' But to them it's fine. So, pain point is how we describe a problem, not necessarily how the user describes it to themselves."

"So, the last question for the whole group then: what role did pain points play in this project? Over the last six months of work, were we looking for pain points after all? That's how product presented the project at least."

"I think so," responded the senior product manager after a few moments pause. "Because I think what we're trying to do is look for opportunities to provide a solution and a new solution is to relieve pain points. And to the earlier point, some of those pains may be things that customers already recognize as bothering them, and other things might be things they don't already recognize. But once they see a solution that does relieve that, they recognize, 'oh yeah, I do want that.' So, solution opportunities come from relieving pain points."

"I think from my point of view," Chris, my fellow researcher responded, "the language I've tried to use for the past several years is 'problems worth solving,' By 'worth solving' that means there is some kind of significant consequence to it. So 'pain point' is something I don't tend to use for my own purposes. I tend to just think of 'problems worth solving.' Because those are problems for users, so they're looking for a solution, it's causing a problem for them. But pain points, I feel like, they are almost, like, tactical things within this big problem that we're solving. Like 'I really struggle to do this, coordinate the work necessary to make this work successful.' That's the problem we're solving, but there are all these pain points along the way which, some of which, we have discovered in this project."

"That's very much how I see pain points," the junior product manager added. "It's almost like thinking of Venn diagrams, they don't intersect, problems and pain points, but you can imagine a circle of a problem or problems and there are different circles inside that that are obviously smaller which form those pain points, right? And the way we try to mitigate or alleviate those pain points is how we take steps towards solving the larger problem or sets of problems."

"Yeah, right, cause it's got that word 'point' in it," my co-researcher added. "So, when you think 'pain point,' you think 'wow, that's hurting me right here.' It's not like my whole body hurts, just that point. And when you see someone hurting, you have to help."

And with that, the project, and our work together ended.

The Micropolitics of the Office

Nearly six months prior, during our first weeks of work together, the team faced its first major dispute. In Skype and phone interviews conducted with office workers throughout the United States, I began my slate of these sessions by attempting to build rapport. I asked the interviewees about their jobs in open-ended terms and tried to get a sense, however limited, of the lifeworld they occupied. I asked whether coworkers went to lunch together and what their breakroom looks like. I would ask what was on their desks or playing on their headphones, before moving into a more explicitly technical set of questions about software preferences and daily workflows. Listening in on these sessions, some of my colleagues were perplexed. Why focus on such irrelevant information? In fact, in a tense exchange with a lead product manager on the team, I was told "Don't waste time with all this stuff, just get straight to the problem and ask them about their pain points." While a marketing manager on the team tried to explain that this was how ethnography works, the tension was recurrent. My co-researcher took the side of product management, and through various means, including escalating to management, tried to compel to ask users more directly about their pain. Refusing to weigh in, management thus left us to our own devices. In this sense, we all agreed that we were misaligned.

However, such misalignment proved to appropriate to our findings. As we moved further into the interview phase of the research, and in learning more about how Americans work with one another, it soon became clear that conflict and relations of power were what our interlocutors themselves wanted to discuss. These conflicts could express themselves in information sharing practices or the way documents were stored on shared drives. Or these conflicts could be expressed in how emails were answered (curtly) or not answered at all (a widely-replicated form of passive-aggression). While we were searching for inefficiencies in software and the kinds of problems that a new feature in a suite of business application might ease, the pain we were encountering was more often about people feeling stuck in the jobs; about the inability of coworkers to find a way to work together; or the sense that the complexity of bureaucratic environments made professional successes impossible. We learned about how subtle forms of domination were enacted at work and the effects of such experiences on the individual worker. Much as in our own disputes about research technique, our interlocutors were suffering -- if indeed suffering was the right term – because of their failures to relate one another or because of the sense of their impatience with their responsibilities.

In this way, our study ultimately became about the micropolitics of the office, entailing a set of problems poorly amenable to representation through a language of "pain points." These were problems about whether and to what degree work was a source of personal satisfaction and how the pressures of the environment impacted their self-understanding. Throughout, prodded to steer conversations back to software, we found that discussions of, say, productivity tools, quickly swerved into gripes about coworkers. And as we proceeded, though we were able to translate some of these complaints into technical problems that we could present to design teams as problems we could solve as a tech company, the real problems encountered were more relational. And thus, the evidence we prepared for the summary decks and thirty-minute read-outs presented to broader stakeholders only pallidly approximated the real, often purgatory "pain" we had been encountering. Rarely a sharp sensation, instead we were observing dull thuds of dissatisfaction and alienation.

Regarding the Pain of Others

It will be noticed that the title of the present essay is an homage of sorts to Susan Sontag's final work, a book-length essay on war photography. A sequel to her earlier *On Photography* (1977), *Regarding the Pain of Others* (2003) analyzes the relationship between representations of atrocity, those who produce them, and their reception within broader public sphere. Through her analysis, Sontag suggests that such images suggest a kind of intimacy between the viewer and subject, but that, as "totems of causes" that marshal action, they forestall the complexity involved in actual relationality. In this sense, they produce a *feeling* of sympathy, but on that makes morality too simple and action too tidy.

David Platzer, PhD is a recent graduate of the anthropology program at Johns Hopkins. He is currently completing a book manuscript *Autism at Work: Labor, Reason, and the Value of Difference* based on dissertation research. Beginning in 2019, as a Berggruen Fellow at UC Berkeley's Center for New Media, he will start a new project research project on artificial intelligence in society.

NOTES

Acknowledgments: A deep debt of gratitude to Paul Rabinow for his tireless mentorship, encouragement, and inexhaustible intellectual energy. Many thanks to Jane Guyer, Anand Pandian, William Stafford, and Megan Steffan for their generous comments on early drafts of this paper. To Josef Wieland, who read many drafts of this essay and patiently tried to shape into something more recognizably EPIC, I can only offer my apologies for this circuitous final product. All the same, I thank you for your friendship and encouragement throughout. Thanks also to ken anderson for helpful comments at an early stage.

REFERENCES CITED

Amirebrahimi, Shaheen.
2016 "The Rise of the User and the Fall of People: Ethnographic Cooption and a New Language of Globalization" *Ethnographic Praxis in Industry Conference Proceedings*. Vol. 2016. No. 1

Koopman, Colin
2003 *Genealogy as Critique: Foucault and the Problems of Modernity*. Bloomington: Indiana UP

Geertz, Clifford
1984 "From the Native's Point of View: On the Nature of Anthropological Understanding." In *Culture Theory: Essays on Mind, Self, and Emotion*. Richard A. Shweder and Robert A. LeVine, eds. Pp. 123–136. Cambridge: Cambridge University Press.

Osterwalder, Alexander et. al.
2015 *Value proposition design*. Campus Verlag.

Sontag, Susan
2003 *Regarding the Pain of Others*. New York: Farrar, Straus, and Giroux.

Star, Susan Leigh, and James R. Griesemer
1989. Institutional ecology, translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, *19*(3

Wailoo, Keith
2014 *Pain: a political history*. Baltimore: Johns Hopkins University Press.

Screenplay, Novel, and Poem: The Value of Borrowing From Three Literary Genres to Frame Our Thinking as We Gather, Analyze, and Elevate Data in Applied Ethnographic Work

MARIA CURY
ReD Associates

MICHELE CHANG-MCGRATH
ReD Associates

Applied ethnography still struggles with the fundamental challenges of (1) framing research to obtain 'thick' data, (2) making sense of data in teams and with clients, and (3) making a convincing case with data in challenging environments. We have observed that borrowing from literary genres can be effective in addressing these challenges. We therefore argue that in an age of data science, it is just as important to draw from the literary arts when gathering, analyzing, and elevating evidence to inspire change in applied ethnographic work. We raise three specific applications of literary genres to distinct project phases, to improve how data is collected and analyzed, and how data travels. In this paper we show: (1) how the screenplay can help solve challenges in research framing, to obtain thicker data; (2) how the novel can help solve challenges in analysis, to turn data into meaningful evidence; (3) how poetry can help solve challenges in the opportunities-development phase of a project, to turn evidence into action.

Keywords: methodology, writing, literary, fiction, framing, research, analysis, impact

INTRODUCTION

Unlike other forms of data collection and analysis – from physics to market research – ethnography has the unique ability to render a deep understanding of everyday human experience in situ and it's many underling meanings. As the field of applied ethnography explores new intersections with data science, we as practitioners should not forget these 'thick descriptions' (Geertz 1973) that set our work apart and give it distinct value. Indeed, it may be the combination of big data with our thick descriptions that increases the relevance and impact of both data science and ethnography (see for example the concept and practicalities of big and thick data working together, explored by Trisha Wang 2013 and Arora et.al. 2018). Yet we still grapple with three fundamental challenges in doing the work that defines applied ethnography: framing our research to obtain thick data when we are in the field, making sense of that data in teams and with clients, and making a convincing case with data in challenging environments, to mobilize change. (Some of these perennial challenges and the reasons behind them – such as demands to do our work faster and cheaper, and client perceptions that we only offer consumer research and not strategy, are explored in Lombardi 2009 and Hou & Holmes 2015.)

In our daily work on projects across various sectors, we have observed that drawing on literary genres to guide our thinking has greatly helped us grapple with these fundamental challenges. We therefore argue that in an age of data science, it is just as important – if not more important – to hone our skills in the humanities, drawing from the literary arts when gathering, analyzing, and elevating evidence to inspire change in applied ethnographic work. Only by continuing to explore innovative ways to do the 'basics' of applied ethnography can we take our field in new directions (like intersections with data science) while upholding our core approach and value.

Using examples from our own ethnographic work, in conversation with critical thinking in applied and academic anthropology, this paper will explore why and how three literary genres – screenplay, novel, and poem – can be impactful across three distinct phases of applied ethnographic work. Incorporating these genres promises to make our research framing more humanistic, our analysis deeper, and the impact of our insights more strongly felt. We will discuss how the screenplay can help solve challenges in research framing to obtain thicker data, how the novel can help solve challenges in analysis to turn data into meaningful evidence, and how poetry can help solve challenges in opportunities-development to turn evidence into action. Experimentation with literary genres should not mean the abuse of raw ethnographic data, therefore risks and careful considerations will be discussed in each literary genre's section in the paper.

ETHNOGRAPHY AND FICTION: A COMPLICATED RELATIONSHIP

We will begin here with a discussion of existing academic and applied thinking on the connections between ethnography and creative writing. More broadly, the Ethnographic Praxis in Industry Conference (EPIC) has continually explored how other disciplines and practices can strengthen our own. Some examples include: borrowing from negative space drawing in the visual arts to help clients broaden their perspectives and sense opportunities beyond their initial objectives (Chang & Lipson 2008); bringing the ethos of repetition found in yoga to generate new insights from past ethnographic research (Thomas 2010); connecting the empathy, nuance, symbolism, and lyricism found in Indian classical dance to the basic aims of good ethnographic practice (Vadrevu 2017); thinking like curators to mediate ideas, perspectives, and discourses for our clients (Powell 2016). As these examples show, we are an interdisciplinary community driven by both a desire to uphold the standards of ethnographic practice set forth in academic anthropology, and a curiosity to experiment with how ethnography flexes and morphs in applied contexts filled with opportunities and constraints not typically found in academia. Thinking through how literary genres can inform our work builds on both this drive to do good ethnography and this curiosity to improve and expand what we do in our unique circumstances.

Rick E. Robinson started the discussion within EPIC on literature's applicability to our work in his paper (2009) on the need to elevate writing and style in applied ethnography. Robinson draws primarily on James Woods' *How Fiction Works* (2008), a short book meant to help literary enthusiasts, writers, and a general audience understand what makes a work of fiction so resonant. Robinson introduces applied ethnographers to Woods' concept of "lifeness" – "life brought to a different life by the highest artistry" that renders "truthfulness to the way things are" (Wood 2008:247). Robinson draws a connection between the lifeness found in literature and ethnography's own preoccupation with truthfulness across all phases, from observation to interpretation (Robinson 2009:94). He suggests that we can do more of the artistry needed to bring out this truthfulness in our projects, "doing the hard work of communicating not just with clarity and fidelity, but with some of the flair, imagination, and voice of the best in fiction" (Robinson 2009:95). Robinson argues for this perhaps counterintuitive approach especially because our audience is different from that of academic ethnographers:

> Style, I think, requires that we do not bracket the passions we find in our work; that when we are stirred, when we observe the stirring, we make space for it in how we write. Writing for an academic audience removes, implicitly, the opportunity to create characters and implies that the authorial viewpoint is an objective one, a scientific one, rooted in description, and shying away from the explicit expression of values, or the imagination of futures. The first move in developing styles for our space then, is to considerably broaden the notion of who our readers might be. (Robinson 2009:42)

Drawing inspiration from Robinson's call to action, this paper delineates how, concretely and specifically, we might introduce more of the literary arts at particular phases of our work. We address the tension that Robinson points to – that it takes artistry (perhaps more so than hard science) to convey the deep truthfulness ethnography aspires to convey.

Indeed, the three literary genres we discuss in this paper – screenplay, novel, and poem – all bring something quite dangerous to applied ethnographic praxis: fiction. The danger of fiction, of course, stems from ethnography's promise to reveal lived realities. As Dawn Nafus and ken anderson (2006) point out, we have built a "real people brand" for ourselves, pitching how "'we' can help businesses figure out what 'real people' want, or otherwise what they do with products. Such 'real people', are always at some distance, a shifting horizon to which the ethnographer goes and returns" (245). And this brand recognition has grown so strong that, as the authors argue, it shortchanges the real work of analysis (Nafus & anderson 2006:249). So what business does fiction have in the 'real people' work that we do? Would it not undermine the truth we promise to deliver?

Originally published in 1986, *Writing Culture: The Poetics and Politics of Ethnographic* (Clifford & Marcus, eds. 2010) launched an ongoing conversation within academic anthropology around the implications of writing ethnographies – what does it mean that there is an author, not just a fieldworker or researcher or scientist, but an author, who ultimately translates experience into text? With the realization that "writing has emerged as central to what anthropologists do both in the field and thereafter" (Clifford 2010a:2), the anthropological eye turned inward to an analysis of the anthropologist's own context, power, and biases. Anthropologists also began to consider their relationship with literature, noting how '[l]iterary processes – metaphor, figuration, narrative – affect the ways cultural phenomena are registered, from the first jotted 'observation,' to the completed book […]" (Clifford 2010a:4). Rather than approach this realization as a problem within the discipline, these scholars were curious to explore both the implications and possibilities of the literary spirit in anthropology. In fact, academic anthropologists today are exploring with literary formats like poetry and graphic novels to reach broader audiences, convey the very intense experiences of participants they've met in the field in more resonant ways, and collaborate across projects to draw broader conclusions (Jackson 2010, Hamdy & Nye 2018).

Even in applied contexts, still what we do is write – memos, reports, presentations, recommendations – from our experiences in the field. Though "we opened ourselves up to ethnography being seen as natural observation" we are not in fact "butterfly collecting or trainspotting" (Nafus & anderson 2006:249, 252). We may have given our clients this impression, but since *Writing Culture* anthropologists have grappled with the dilemma that we are not simply making objective observations. We write, in many ways like playwrights, novelists, and poets write. Like the scholars of *Writing Culture,* we seek in this paper to acknowledge this reality in our applied work – that there are elements of the literary in the ways we analyze and articulate – and to carefully explore its possibilities.

But first, let us establish some fundamental differences between fiction and ethnography. Anthropologist Thomas Hylland Eriksen, outlining the relevance of fiction in anthropology, writes:

> Fictional accounts […] present persons and events which have been invented by the writer. Anthropological texts try to present *a few aspects of* social reality as accurately as possible, taking account of the limitations entailed by fieldwork, 'cultural translation' (or, if one prefers, cultural reduction) and attempts at linguistic representations of society. Lies and deliberate misrepresentations are banished from anthropological scholarship, which should additionally – unlike fictional writing – try to present empirical material systematically and comprehensively and distinguish between description and analysis so that the reader may draw his or her own theoretical conclusions. (1994:168-9)

Fiction, though evocative and vivid, can lack accuracy, comprehensiveness, and comparability – virtues we value in ethnography (Eriksen 1994:193). We write and read ethnographies primarily to understand something about a subset of the world, and this is not typically the case with fiction. Wood points out that although we gain this understanding from reading fiction, this is not what we set out to do when reading it, "[w]e read fiction because it pleases us, moves us, is beautiful, and so on – because it is alive and we are alive" (Wood 2008:170). Ethnography and fiction have different starting points and expectations.

However, there are some fundamental similarities between ethnography and fiction. Clifford goes so far as to say that although "[t]o call ethnographies fictions may raise empiricist hackles…[e]thnographic writings can properly be called fictions in the sense of 'something made or fashioned'" (Clifford 2010a:6). In good ethnography, as in good fiction, there is the careful and deliberate attempt to render a world. Ethnography and fiction both create reductions of social reality (Eriksen 1994:192-3) – we must acknowledge that as ethnographers we do not provide a full account of the worlds we venture into, but an account of what we deem relevant to the questions at hand. Both fiction and ethnography search for higher truths. In ethnography the truth of particular interest is what goes beyond the individual, to what applied anthropologist Suzanne L. Thomas describes as "a truthful social performance, one that enacts social and cultural dynamics not isolatable facts of individual behaviors…we deal in identities, narratives, symbols and artifacts, and seek truthfulness on this scale" (Thomas 2010:241).

Moreover, ethnography and fiction both have a deep concern for empathy. For ethnographers, empathy is what helps build rapport in the field, empathy is what makes the 'other' of traditional anthropological study ultimately comprehensible, understandable, relatable. Good ethnography creates empathy in readers, clients, or stakeholders. Wood points out that in good fiction empathy is a byproduct too, what he calls "fiction's true mimesis: to see a world and its fictional people truthfully may expand our capacity for sympathy in the actual world" (Wood 2008:171-2).

There are some unique circumstances of applied ethnography, in particular, that render the literary even more useful to our field. We are ultimately trying to convince and persuade – we want our clients or the organizations for whom we do applied ethnographic work to be convinced that we have to say is true and insightful, and we want to persuade those clients or organizations to change, sometimes in fundamental ways, according to the implications of our truths and insights. In *Writing Culture*, Vincent Crapanzano draws a comparison between Hermes, the Greek patron god of literature, and the ethnographer, and the comparison may be especially apt for the applied ethnographer in particular:

> Hermes was a trickster: a god of cunning and tricks. The ethnographer is no trickster. He, so he says, has no cunning and no tricks. But he shares a problem with Hermes. He must make his message convincing. It treats of the foreign, the strange, the unfamiliar, the exotic, the unknown – that, in short, which challenges belief. The ethnographer must make use of all the persuasive devices at his disposal to convince his readers of the truth of his message, but, as though these rhetorical strategies were cunning tricks, he gives them scant recognition. (Crapanzano 2010:52)

We could borrow a technique or two from the literary to write persuasively and convincingly, and to think insightfully, for our clients and organizations. These are audiences (often actually collaborators) who need more than the knowledge of descriptions – they need eureka moments, persuasion to believe something matters enough to act on it, and plans for the future. As Robinson argues, "[f]iction is the narrative imagining of invented worlds. We are not in the business of inventing data, but we are in the business of imagining futures every bit as much as we are in the business of representing realities" (2009:101). New products, new services, new spaces, new structures – these require imagination, the stuff of literature. Our need to convince, persuade, and imagine futures makes a familiarity with the literary especially relevant.

But despite the profound impact of *Writing Culture,* scholar of literature and anthropology Oscar Hemer still observes that "[w]hereas writing style is crucial for a literary writer or a journalist, in academia it is not only strikingly subordinated; it is even met with suspicion, as if eloquence were a way of concealing a meagre academic content" (2016:161). His observation that "writing is not usually associated with methodology" (Hemer 2016:161) is not simply pointing towards a problem in technique, but a bigger problem in thinking. Writing as methodology is not just how we express what we have found in the field – it is also how we think throughout our projects. In her foreword to the 25th anniversary edition of *Writing Culture*, anthropologist Kim Fortun reflects that ethnographers should experiment with ways of writing, dappling in different genres like the novel, because "[e]xperimentation, here (as in the sciences) is as much about constraint as freedom. Writing then, is not only about representation or even evocation but a way to generate insight" (2010:xi). We will now proceed to outline how we can draw inspiration from writing in literary genres, not just to create reports, memos, or presentations, but actually to actually frame our thinking and approach across three distinct phases of a project.

THE SCREENPLAY IN A PROJECT'S FRAMING PHASE

When we frame a project in applied ethnography, we typically develop a field guide or protocol to structure our time in the field. This encompasses questions we want to ask of our research participants, places we want to explore, activities we want to observe. Under intense time, resource, and client pressures, we've observed that this guide risks becoming a survey or questionnaire for research participants, flattening the data we collect.

For example, in a recent study on a chronic health condition, a team at ReD Associates set out to meet research participants at various stages of the condition, from pre-diagnosis to decades of chronic health management. The team assembled in a room to prep for going to the field. Capturing, in one concise field guide, the breadth of experiences they wanted to study was proving to be a challenge. Flipping to the section of the field guide that focused on building initial rapport with the participant, the team read through the suggested first questions: 'How do you imagine your future panning out?' 'Are there any big changes on the horizon?'

A burst of laughter rang out: "I honestly can't imagine asking that to a complete stranger who's just walked into my home." The team shifted uncomfortably in their seats, and proceeded to review the rest of the themes, topics, and activities, as originally drafted in the field guide. A new consultant asked: "So, I realize I'm new to this, at least in terms of there being a client involved…but I can't quite shake the feeling that this is incredibly stiff. It feels like an interrogation."

The team continued to go through the field guide, section by section. The project manager began to realize the reason for all the hesitation, the discomfort of what on paper appeared too personal for researchers to see themselves asking when they got to the field. Based on the way the field guide was drafted, they had all been imagining two individuals sitting across from each other on a couch or at a table for eight hours. No one had moved, and only one person had asked questions. It was a one-scene act, set in a living room, one actor playing the interrogator. Seeing it in this way made her realize the research framing was all wrong, and they were not going to get the thick descriptions they needed. "Okay, I get it, this totally reads like a police procedural. Let's think about this differently. What kind of screenplay are we writing?"

The team set out on their revision and started to consider 'setting' –to see where participants took their treatment, where they went day-to-day and how their condition might interfere with those outings. They thought of the 'props'– to know what treatments patients took and what remedies or solutions they carried with them on-the-go. They tried to envision 'dialogue' – to glean the perspectives not just of the participants but also of their families, and to ensure the participants would learn about the researchers in the process too. And they factored in timing – how to ensure, to the best of their abilities, they might encounter everything they wanted to explore in the field. In retrospect, what the team had really done in their revision was to borrow from the concept of a screenplay to develop a field guide that fostered deep ethnographic experiences. The team had moved away from an interview questionnaire and towards setting the scene for a dynamic interaction.

Having a playwright's ability to a set a scene and a director's eye for what to capture – the exercise our team conducted – is akin to what Thomas describes as "documentary finesse," a key skill she hires for in her applied ethnographers:

> It is not a fluke when we, as ethnographers, succinctly capture a moment in image, photo, video clip or tale. The freshness and vitality of Faulkner and Zafiroglu's video clip, I argue, is evidence of their documentary finesse. It is also evidence of a well-prepared and well-conducted field engagement. Both knew what they were looking for, how best to document it and how to finesse the appropriate social environment in which the woman would remember the third mobile phone, pull it out of her tight back pocket and talk to a team of strangers in an impossibly small Shanghai apartment while Faulkner's camera rolled. All of these, from the preparation to the finessing of the social environment to the tedious documentation comprise the art. (Thomas 2010:239)

Though at face value one could read Thomas's description of documentary finesse as an overly contrived and controlled fieldwork experience, we argue that it is more about intuition and openness – being able to imagine, ahead of time, how fieldwork might happen in the field, and planning out an experience that is not just open to, but actively encourages, the twists and turns that fieldwork can take. There is ultimately an artistry in what Thomas is describing, rooted in having a literary sensibility for how interactions – with people, material culture, and physical surroundings – could play out in the field in engaging ways.

While thinking with screenplays in ethnographic contexts might seem radical, both applied and academic scholars have explored this before. In his guide to new writing practices for novice applied ethnographers, H.L. Goodall Jr. describes what is essentially a cinematic, screenplay-like approach to verbal exchanges in the field, asking the ethnographer to consider factors such as the frame or context, the action taking place, not just what is being said but how it is spoken, and where the researcher is in the "scene" (Goodall 2000:106-7). Clifford, too, alludes to the literary and cinematic in his observation that when we go into the field we do not simply float passively in our environments, but rather "[t]he fieldworker presides over, and controls in some degree, the making of a text out of life" (Clifford 2010b:116). We should use this inherent element of ethnography to help foster deeper engagements in the field.

How, then, to have a screenplay-like approach to framing the research design before fieldwork? We here propose five screenplay elements to consider in field guide development:

1. **Setting**: Where do we want the interactions with participants to take place? Consider moving across the settings the participant typically inhabits, being careful not to assume what these places might be ahead of time. For example, the participant may have a bladder condition, but perhaps they keep their medication in the kitchen, not the bathroom. Consider, also, how to progress from setting to setting. Perhaps asking to go for a neighborhood walk does not make sense to do right away, and first the researcher and participant should get to know one another in the living room.

2. **Props**: What material culture do we want participants to share with us? Consider the objects that trigger memories and stories, like photo albums or home videos stored on phones. Consider objects that challenge or add further granularity to a participant's account. For instance, looking over medical records (shown by consent) could delineate a different treatment timeline than the participant recalled, and discussing those discrepancies together could be insightful.

3. **Dialogue**: How do we foster dynamic conversations? Consider the different "actors" of interest beyond the single participant, and how those individuals might be brought in to the interactions. Consider, also, what types of information, stories, and reflections the researchers themselves might be able to bring to the conversation, so that the participant doesn't feel the experience is one-sided (because indeed it should not be). A researcher divulging his own struggles in communicating with loved ones might help a participant share that his car is where he feels most comfortable broaching difficult topics with his kids.

4. **Action**: If we are participant-observers, what do we actually want to observe and participate in, beyond just having a conversation? Consider the activities that would intersect with the phenomenon of study in both obvious and obscure ways, and how to arrange those potential interactions ahead of time, assuming the participant is open to the ideas. This requires collaboration with the participant, and discussing with them ahead of time what day-to-day life is like. Perhaps the researcher should meet with the participant on a Tuesday rather than a Wednesday, to accompany her as she goes to the pharmacy to pick up her medication.

5. **Timing**: How do we flow through the settings and the discussion topics, to ensure there is time for everything we want to cover? Consider what is a need-to-have, and what might be sacrificed if the activities in the field go in an unexpectedly fruitful direction. How long, roughly, should each "scene" take?

Thinking of the field guide as a 'screenplay' can be problematic. We do not want to force situations. We ultimately want to develop the overall structure for the types of scenarios we want to encounter and conversations we want to have, but the key is to be open-ended rather than prescriptive or presumptuous, following how the participant interprets the 'screenplay.' We also must not force participants into situations they do not feel comfortable with – instead co-developing the activities and the spaces the ethnography will occupy, on the participant's terms. The exercise of thinking-like-a-playwright is not, fundamentally, about crafting a beautifully-written field guide, but about ensuring we get as much a sense of the "lifeness" (Wood 2008) of the phenomenon of study as we can, by planning ahead. Fortun describes it as orienting without over-determining: "Texts need to be imagined as we move through the field, directing our attention to the kinds of material we will need to *perform* an analysis. This means that we must also imagine narration and argument as we go, even while remaining open to the field's beckoning…" (Fortun 2010:xii). Outlining ways we can be carefully inspired by the screenplay genre may help us avoid common research-framing challenges in applied contexts, to obtain thicker data with which to do our analyses when we return from the field.

THE NOVEL IN A PROJECT'S ANALYSIS PHASE

After fieldwork, we often collaborate in teams and with stakeholders to develop insights. We come back with pages upon pages of field notes, thousands of photos, and connections that have not been fully formed yet. We typically need to communicate our experiences with other researchers who went to different field sites or with clients who have not been in the field themselves. With overwhelming content, and the pressure to figure it all out efficiently – what does it mean? what are the implications? – we have seen teams struggle conveying the richness of their data, and making sense of it.

For example, a few years ago we had a study on home renovations. Researchers met with participants across field sites in Europe, spending a day with each participant in their kitchen and home, walking through past, current, and future renovation plans, and observing how the participants use their kitchens and other rooms in the home. Upon returning from the field and after an hour of hearing teammates discuss a research participant – Jeppe – we remembered nothing. Who was Jeppe, really? Why did he want to renovate his kitchen, and what was it all for? There were 29 more participants to learn about, and we were grappling with how we would make sense of it all.

Then it was time to discuss another participant – Sybil. Sybil had a story: she was the woman who, after 59 years of what she described as thinking of everyone but herself, was finally indulging in her dream kitchen. The twist: she never cooks. The big reveal: a double-door cabinet with multiple wicker compartments, space for all her unused spices, and drawers that silently closed so she could get her early morning tea without waking anyone up, and have a moment to herself. The team could remember her like a character in a novel. Through Sybil we saw the beginning of a larger pattern and story emerging in our data set. As our thinking developed around this particular type of 'forever' renovation (in contrast to the 'just-for-now' light-touch renovations we also saw in in the field), Sybil was more than just a data point in a sample of thirty – her experiences became evidence, a telling detail in a narrative. When we shared our findings with our clients, Sybil helped to carry one of the key threads of the broader story, illustrating particular motives and drives in high-cost home renovations.

Looking back, we struggled to draw insights from the field until we started thinking like novelists in how we synthesize and communicate our data: with only a brief time to share our findings with teammates and clients, we thought of ways to best describe the people we met as compelling 'characters'; we thought about 'narrative arc' and the climax and denouement of our fieldwork experiences; we considered tone – to whom were we telling our story, and what voice would be most resonant to both the data and the audience? We moved from scattered descriptions to the beginnings of a meaningful 'plot.'

A good novel gets us to understand the people we encounter in its pages. As Robinson argues, "[i]sn't this what we purport to offer to our clients, to our audiences? A level of understanding the subject that is so close as to 'inhabit' their way of being in the world? […] Allowing them to consider it, know it, and ultimately, to value it, respect it, even as we offer to change it?" (Robinson 2009:96) As much as we claim ethnography to be an account of the 'real,' there is a craft of including (determining the essential) and excluding (determining the non-essential) information to get to a deeper meaning, just as a good novel provides details that are relevant, and leaves out the extraneous. Academic anthropologists have observed this aspect of ethnography well: "[t]he best ethnographies, Clifford reminds us, are systems and economies of truth, and are structured accordingly. They convey, convince, and enroll because they select and exclude – drawing out, literally, through content and form, particular relationships" (Fortun 2010:xiii).

Thomas (2010) outlines another skill she looks for in applied ethnographers she works with, in addition to documentary finesse:

> Over the last five years, I stopped hiring ethnographers for a report. I now hire for the ability to build grounds-up a symbolic and narrative fluency amongst my team and my key stakeholders. The raw field notes, the half-baked field reports and the weekly meetings where we debate the significance of an elderly woman's loss of eyesight and her memory of reading a favorite novel – these disciplinary practices extend the longevity of the ethnographic project. The tale of the elderly woman's failing eyesight did not make it into the final report, but our debate shaped how I told and now re-tell the story of the ethnographic project. (Thomas 2010:238)

The skill Thomas is here describing around symbolic and narrative fluency is essentially the skill of a novelist in crafting a story. Sybil in our home renovation study was much like the elderly woman in Thomas's account. As with thinking like a playwright in research framing, thinking like a novelist in analysis is not about creating a vividly written story for a client or organization. It is about using literary approaches to help build ideas and think through the analysis itself, going deeper than a list of observations.

How, then, can the novel genre inspire applied ethnographers to draw out the most from their ethnographic data, upon returning from the field? We here propose three novel elements to consider in post-fieldwork analysis:

1. **Characters**: Who did we meet in the field, and how do we convey them to our teammates and stakeholders in ways relevant to the study? Consider 'sketching' each research participant in terms of their motivations, aspirations, challenges, and habits, and how these intersect with the key topics of the study. Consider what each participant helps illuminate about the project's core research questions as well as the client's assumptions, and how their experiences compare – if the participants were all in a room, how might they relate to one another? Consider how unique details feed into a broader narrative, much like how Eriksen (1994) notes that the good ethnographer, like the good novelist "tries to fuse the universal with the particular and thus accounts for individual idiosyncrasies, as well as structural and cultural defining characteristics of the different situations" (172-3). Sybil's silently closing cabinets are an idiosyncrasy, but together with other data points from other participants, help point towards a bigger insight around kitchens as a space for me-time.

2. **Plot and narrative arc:** A good work of applied ethnography is able to elucidate gripping points of tension. At one level, we describe the tensions in the lives of the participants we meet in the field – how what they say doesn't match what they do, how their aspirations don't line up with reality, how life is full of contradictions. At another level, we consider the tensions between the lives of the research participants and the (mis)understandings that clients or organizations have of them. In a study on beer and bars, we several times came across what one bar owner dubbed the "box of crap" – filled with promotional items that manufacturers, like our client, had given him, thinking these items were preciously used and constant reminders wonderful brands. Such misalignments are like dramatic irony, and also have significant business implications. How do we ensure 'aha' moments for clients and stakeholders through plots which help them come to terms with the dramatic gap between their investments and their customers' lives? These moments do not arise on their own, but rather require Robinson's definition of style, "the control and expression of ironic tension" (2009:96). It is in the analysis phase of a project that this focus on dramatic irony, plot, and a narrative arc come most into play. Moreover, the arc – if well-conceived – should naturally lead applied ethnographers to compelling solutions and opportunities.

3. **Tone**: How do we communicate our story, once we know what that story is? This is inextricably tied to the questions, "who did we meet?" and "who is our audience?" What tone best captures, and respects, the lives of the participants we met in the field? What tone will have the most resonant impact with the client or organization? Aligning or going against the expected tone should be a deliberate decision. In a study on first-time parenting and baby products, we found that a tone that alternated between preciousness and frenzy best encapsulated the whirlwind of parenting experiences we observed in the field, and the way the parents we met with spoke about their experiences themselves – at times with gentleness and awe, at times with humor and exasperation. This tone surprised our clients, who had focused solely on the preciousness of parenting, and who were only just beginning to explore humor in their positioning and messaging.

To say we should create characters out of participants, dramatic twists out of field interactions, risks being reductive. However, we argue that in applied ethnography we already run the risk of flattening the depth of the people we meet in the field. Robinson describes the typical creation of "personas" that are "shallow, simplified, and static when compared with the imperfect messiness of just plain folks […][,] characterized by a lack of multiple registers and a decided absence of tension" (Robinson 2009:102). Nafus & anderson (2006:247-9) and Chang & Lipson (2008:197) describe the problematic predominance of decontextualized photos and pull-quotes in our work, that both hide the behind-the-scenes analysis separating ethnography from naturalistic observation (Nafus & anderson 2006:252) and leave too much for the audience to fill-in with their own assumptions. Doing the work of thinking like a novelist ensures that we go beyond standalone quotes and bring the nuance of the people we met to life for our clients in lucid and relevant ways. The novel genre can inspire teams struggling to process and share complex qualitative data in a short amount of time, so that data becomes evidence – it has meaning in a plot.

Our suggestions for how to take a novelistic approach to analysis come with a bright neon warning label, though, to proceed with the utmost care and caution. We cannot fabricate or alter details from the field to fit a narrative we want to tell our clients or organization. Moreover, and more subtly, we cannot take the creative liberty of speculating on our participants' inner lives: "'*Direct introspection'* is deemed unacceptable in social anthropology. If one ventures to consider the inner states of persons, one must always refer to acts or statements as evidence" (Eriksen 1994:192). As ethnographers making sense and meaning out of people's everyday real life experiences, we must uphold ourselves to the highest truth-telling standards and use creative approaches to guide our thinking only insofar as the approaches help to convey those realities – not fabricate new ones.

THE POEM IN A PROJECT'S RECOMMENDATIONS PHASE

The latter part of an applied ethnographic project usually entails translating data into opportunities and directions. The specifics vary greatly depending on the context the applied ethnographer is working in – embedded as a branch of an organization or contracted by a client, working in private sector or public sector, working for R&D or strategy or marketing, and so forth. Despite the differences, we share the struggle of getting our evidence to 'stick' within an organization and lead to action. This could be for a mix of reasons, not least because of an organization's inability to fully grasp and internally communicate the implications of the study.

For example, we were at a 'directions workshop' where our team was conveying to our clients the big idea for the first time. This was for a healthcare study about respiratory health. Researchers met with participants across field sites in the US, spending a day with each participant in their homes, with their friends and family, and engaging in activities outside the home to understand how chronic respiratory issues affect people in different social and environmental contexts.

We had spent the evening before the workshop finessing our points, making our case for the big misalignments between the pharmaceutical industry's prioritizations and the needs of patients. We were hopeful that our ideas would resonate with our clients, but previous conversations with them led us to believe the insights were not quite sticking, or were not helping them to see new solutions.

The next morning, at the directions workshop, our clients' reception to our slides started off contemplative and silent. We had a slide on the screen, and a few different ways of articulating our central point – a photograph, a quote from a research participant, a bulleted list of needs. In passing, someone from our team added, "…It's really a shift from keeping patients out of hospitals, to getting people out of their homes." And our client, who had been leaning back in her chair squinting at the screen, suddenly looked as if she understood. Ninety minutes and forty slides later, she summed up her key takeaways: "I really liked what you said, it's about that shift from keeping patients out of hospitals, to getting people out of their homes." Something in the simplicity and resonance of that statement, which serendipitously had the cadence of a line of poetry, made an idea stick.

When we developed the second iteration of the presentation, we wrote this line down, and used it to develop a model illustrating how the worlds of the people we met had become smaller and smaller, due to their condition. Thinking poetically helped us in two ways: poetic rhythm helped us create a value proposition that encapsulated our recommendations in a memorable, succinct way, and metaphors guided our development of a visual model to not only explain, but further think through, a key insight.

One might argue that in being too vivid and evocative in our communication we lose the seriousness of being anthropologists. Indeed, Eriksen warns that "anthropologists should probably resist the temptation to indulge in the rich and evocative language of creative writing" to avoid turning our work into "travel writing" (1994:194). But, as Robinson points out, our work as applied ethnographers is a bit different: "Were we only reporting results, there would be good reason for such a limitation. But we aren't only reporting. We are creating alternative interpretations, opening up ways of thinking differently, imagining new futures" (2009:34). He notes that we already use metaphors constantly in the 'journeys' and 'cycles' that we create as deliverables, and perhaps applying these metaphors more broadly to encapsulate insights can help communicate the truths we witness in the field, and the potential futures (2009:104). If some of the deliverables we are asked to develop are, themselves, poetically framed (albeit tepidly – there is hardly anything poetic about a consumer journey anymore) should we not harness that poetic sticking power in fresh ways?

Thomas (2010) puts it succinctly when she writes, "Corporations require one thing from ethnography: the ability to effect change. If we do not change corporate practice (from product definition to sales and marketing practices), we fail" (2010:249). Change, we would argue, comes from newfound understandings, and the metaphorical language of poetry helps us to both make sense of the unfamiliar and to see new potential in what we think we already know – something explored in *Writing Culture* as the ethnographer's task of "render[ing] the foreign familiar and preserv[ing] it's very foreignness at one and the same time" (Crapanzano 2010:52). What Crapanzano describes is a process of disorientation and reorientation, and poetic devices like the metaphor help to achieve just that through juxtapositions. One of the beauties of poetry is that helps us to appreciate what we might overlook, and the other is that it helps us to articulate experiences we struggle to understand – much like ethnography, though often in far fewer words.

How, then, can the poetry genre inspire applied ethnographers to develop impactful ideas, concepts, visualizations, and recommendations that 'stick'? We here propose two poetic elements to consider in the recommendations phase of a project:

1. **Poetic (and economical) language**: How do we crystallize exactly what we want our stakeholders to take away from our findings, in a way that carries across the organization and continues to make sense even when we are not there to explain the finer points of the fieldwork? What evokes the broader story, that the details from the field will then properly unfold in full? As Clifford observes, "'Poetry' is not limited to romantic or modernist subjectivism: it can be historical, precise, objective" (2010a:25-6), and it is precision, not romanticism, that we are after here. Poetry can be florid, but can also be incredibly economical. Thinking in metaphors and similes can do the work of "speeding us, imaginatively, toward new meaning" (Wood 2008:204). Consider thinking in hyperboles, alliterations, personifications, and repetition of imagery to help condense insights and point towards future concepts and imaginaries. For instance, in a study on a chronic illness, we noticed that our clients were caught up in a narrative that patients never took initiative to seek help because they felt too embarrassed about their condition. The reality we saw in the field was that patients didn't take initiative because they had developed coping mechanisms in lieu of medical help they could not access. As a shorthand for our client to understand that their patients were not wallowing in embarrassment but rather going about their lives with the strategies they had discovered on their own, we referred to these as their "superpowers" early in our narrative – it was a hyperbole, and we complicated the concept further into the story, but it helped jolt our clients and get them out of their familiar frameworks.

2. **Metaphoric models and visualizations:** How do we extrapolate from dozens of observations a more abstract model, concept, or image that conveys a higher truth or a deeper meaning behind the experiences of the field? Poetry can help us to de-familiarize and re-familiarize, so that we draw contrasts and comparisons with freshness – whether it is to make sense of something unfamiliar and new, or to see anew something thought to be banal or commonplace. Applied anthropologist Anne Line Dalsgaard describes an extended metaphor she used to encapsulate two different types of organizational cultures – the wolf pack and the beehive, and how

> this helped her stakeholders make sense of their everyday office contexts and determine what worked in company culture and what needed to change (Dalsgaard 2008:151-2, 157).

This is not an argument for distilling thick description into zingers and one-liners, but rather for using poetic form as inspiration in making ideas stick in complex, slow, and indecisive environments. We must be careful not to get carried away by the loose whims of poetry: "What fiction gains from its vividness, freedom to experiment and evocative techniques, it loses in its lack of accuracy, empirical comprehensiveness and attempt to establish interesting comparative dimensions" (Eriksen 1994:193). We must not mistake jargon, business buzzwords, and industry taglines for poetry – overuse of bloated but ultimately empty words like 'connectedness' and 'patient-centered' and 'care' runs rampant in the industries we work for and in. We must strive for precision, specificity, and the imagining of new potentials. Fortun, in her call for ethnographic experimentation with form, writes, "It is not about writing beautifully, or rendering people and events poignantly. And poetics don't license texts without order. The challenge is to put things together inventively, to order them in a way that reorders reader's experience" (2010:xii). As with the screenplay and the novel, drawing inspiration from poetry is less about the written words, and more about the tools for thinking in fresh ways.

CONCLUSIONS AND FUTURES

At the heart of this paper is the belief that as our work in applied ethnography moves further into contexts that demand us to be practical, efficient, and to-the-point, and to delve into the scientific, the quantitative, and the mechanical, it is actually our connection to the humanities that can give us ingenuity of thought and impactful outcomes. As Thomas describes, our work has the potential to be long-lasting:

> There is a longevity to the ethnographic arts: a report referred back to over the years, a photo that captures a critical moment and still resiliently fresh truth, a chart of a common practice that renders it momentarily foreign and, as a result, suddenly intelligible. In cruder words, ethnographic analysis has a longer shelf life than, say, traditional market research. The latter requires tending, updating, refreshing to keep the demographic or other categories replete with a fresh cast of characters. The former is distinguished by a methodological discipline that keeps it fresh and truthful without the necessity of being, for only the moment, a truth. (Thomas 2010:237)

We believe that literary genres can help us to further the resonance in our work and extend its shelf life. Imagining we are playwrights or documentary filmmakers when we frame the interactions we want to have in the field can help ensure that we gather the thick descriptions we seek. Thinking like a novelist when we analyze our fieldwork experiences and the lives of the individuals we meet can help to better make sense of our qualitative data, especially when working in teams or with other stakeholders, to generate insights. Flexing our skills, however limited, in poetic language can help us to mobilize our data and insights into recommendations that stick. These genres can help address challenges particular to applied ethnography – field guides often turning into survey questionnaires, insights that do not crystalize or go deep enough, and recommendations suffering from a lack of sticking power. These of course are the challenges that come with the pressure to do our work faster and cheaper, and the need to deliver on goals and timelines set by stakeholders.

Our attempt here is to keep applied ethnography grounded in its humanistic roots even as forays into data science – the theme of this year's EPIC conference – entice us with new potentials. But it is worth making some speculations about how these literary genres could actually help applied ethnographers to better work with data science. Using a creative writing lens could perhaps help data scientists consider the 'why' or 'how' behind their quantitative data, and consider possibilities or gaps that can be explored through qualitative research – in short, thinking like a novelist could help data scientists realize the blind spots of context in their data. For instance, in their case study at this year's EPIC conference, Wachmann et. al. explore the challenge of integrating big data to ethnographic insights on why people visit a theme park in the UAE. The 'data lake' available was on individuals, rather than groups, making it difficult to further analyze how people bond, or don't, at a theme park – a key qualitative insight in the study (Wachmann et. al. 2018). If data scientists could look at their data patterns and data collection frameworks early on in a literary way – thinking of how the numbers might represent the actions or motivations of characters in a story, for instance – it might help them imagine the contexts of their data, and what further knowledge is needed to really make sense of that data. Are people dragged to the theme park because they have to be, à la *National Lampoon's Vacation*? Is it a romantic place of discovery à la *Adventureland*, or a daring escape, à la *Zombieland*? Such imaginaries – the story behind the numbers – could help expand the thinking beyond funnels of efficiencies, pointing to relevant further studies on underlying motivators, needs, and challenges, and pointing to new metrics (such as measuring groups rather than individuals) that could be collected.

In the reverse direction, thinking like a data scientist might help applied ethnographers ensure their poetic concepts are ultimately grounded in concrete observations. For instance, in a recent global study on the future of shopping, we at ReD Associates used the concept of 'fluid living' (Bauman 2005) to make sense of, and elevate, several observations about the changing structures of contemporary life that were impacting shopping habits. When we then integrated a large quantitative study to further develop and size our insights, we needed to make sure that the evocative language was still pinned to concrete, specific behaviors that could be measured. This helped us stay grounded even as we sought abstract connections and concepts that made our ideas memorable for our clients. Obviously these two ways that a literary approach can work with ethnography and data science – imagining new contexts to probe at the 'why' behind numbers, and staying concrete and specific to size and measure abstract concepts – are initial suggestions explored very briefly, but we hope this helps move the conversation forward on combining humanities and social science with data science.

As we've outlined throughout the paper, when taking inspiration from literary genres we must use them to stick to – and to see new – truths from the field, rather than to feel we can take creative liberties with the data. Data, even when it is quantitative, can be misinterpreted and misconstrued – so this is a warning for any kind of cross-disciplinary work we do in ethnography, not just with the literary. We must, as Clifford states, "tell stories we believe to be true" (Clifford 2010b:121). Thinking like a playwright or a documentary filmmaker in the framing phase of a project does not mean prescribing what participants will do and say in the field. Thinking like a novelist upon return from the field does not mean making up characters that do not exist, filling in motives or introspections we have no proof of, or fabricating experiences from the field. Thinking like a poet in developing concepts and recommendations that stick should not mean flattening or reducing truths and human experiences to empty palatable terms. There is a "fundamental separatedness" between ethnography and the literary that we cannot forget as we experiment – "[t]he *methodological* challenge when using creative writing, whether as a means of exploration or as a knowledge source, is to bear in mind the distinction between fact and fiction" (Hemer 2016:177). Failing to keep in mind these distinctions, and to keep in mind the aims of ethnography, can lead to pitfalls – such as proscriptive fieldwork experiences, fabricated stories, and hollow terms – that tarnish the value we bring to the industries we work in.

As applied ethnographers, we often struggle with being neither-here-nor-there – straddling disciplines, organizational departments and divisions, or roles in a team. But this can often be our strength, as Reese et.al. highlight when they describe the "hyper-skilling" of the ethnographer and how a "road to practical effectiveness (and professional fulfillment) lies in recasting the role of the ethnographer as inter-disciplinary mediator" (Reese et.al. 2010:3). We suggest here that the shapeshifting, nebulous task of doing and writing ethnography in applied contexts can too benefit from leaning in to this interdisciplinary ethos. Indeed, "[e]thnography is hybrid textual activity: it traverses genres and disciplines" (Clifford 2010a:25-6). Playing with literary genres, like the screenplay, the novel, and poem, to reshape our thinking in projects can perhaps bring us more effectiveness, fulfillment, and impact in the work that we do.

Maria Cury is a manager at ReD Associates. She has extensive experience applying ethnography to various industries, such as technology, healthcare, and consumer goods. Cury received an MSc in Visual, Material, and Museum Anthropology from the University of Oxford, and a BA in Anthropology with a certificate in Visual Arts from Princeton University. *mcu@redassociates.com*

Michele Chang-McGrath is Partner and Director of ReD Associates' UK practice. In her 15-year career as an ethnographer, her research has focused on humanizing technology when advising clients in retail, financial services, and life science. Chang-McGrath received her Masters in Interactive Telecommunications from NYU and has a BA in Art History from Reed College. *mcm@redassociates.com*

REFERENCES CITED

Arora, Millie P. et. al.
2018 Contextual Analytics: Towards a Practical Integration of Human and Data Science Approaches in the Development of Algorithms. *Ethnographic Praxis in Industry Conference Proceedings* 2018: forthcoming.

Bauman, Zygmunt
2005 Liquid Life. Cambridge: Polity.

Chang, Michele Frances and Matthew Lipson
2008 Drawing from Negative Space: New Ways of Seeing Across the Client-Consultant Divide. *Ethnographic Praxis in Industry Conference Proceedings* 2008:192-201.

Clifford, James
2010a "Introduction: Partial Truths." In *Writing Culture: The Poetics and Politics of Ethnography, 2th Anniversary Edition* (2nd ed.), edited by James Clifford and George E. Marcus, 1-26. Berkeley: University of California.

2010b “On Ethnographic Allegory.” In *Writing Culture: The Poetics and Politics of Ethnography, 2th Anniversary Edition* (2nd ed.), edited by James Clifford and George E. Marcus, 98-121. Berkeley: University of California.

Clifford, James and George E. Marcus, eds.
2010 Writing Culture: The Poetics and Politics of Ethnography, 25th Anniversary Edition (2nd ed.). Berkeley: University of California.

Crapanzano, Vincent
2010 “Hermes’ Dilemma: The Masking of Subversion in Ethnographic Description.” In *Writing Culture: The Poetics and Politics of Ethnography, 2th Anniversary Edition* (2nd ed.), edited by James Clifford and George E. Marcus, 51-76. Berkeley: University of California.

Dalsgaard, Anne Line
2008 *Verfremdung* and Business Development: The Ethnographic Essay as Eye-opener. *Ethnographic Praxis in Industry Conference Proceedings* 2008:146-59.

Eriksen, Thomas Hylland
1994 “The Author as Anthropologist: Some West Indian Lessons About the Relevance of Fiction for Anthropology.” In *Exploring the Written*, edited by Eduardo P. Archetti, 167-196. Oslo: Scandinavian University.

Fortun, Kim
2010 “Foreword to the 25th Anniversary Edition – Of Writing Culture, 2020.” In *Writing Culture: The Poetics and Politics of Ethnography, 2th Anniversary Edition* (2nd ed.), edited by James Clifford and George E. Marcus, vii-xxii. Berkeley: University of California.

Geertz, Clifford
1973 “Thick Description: Toward an Interpretive Theory of Culture.” In *The Interpretation of Cultures*, 3-30. New York: Basic Books.

Goodall, Jr. H.L.
2000 Writing the New Ethnography. Lanham: Altamira Press.

Hamdy, Sherine and Coleman Nye
2018 “Drawing Culture, or Ethnography as a Graphic Art: The Making of Lissa.” *American Anthropologist* website, June 7. Accessed September 15, 2018. http://www.americananthropologist.org/2018/06/07/drawing-culture-or-ethnography-as-a-graphic-art-the-making-of-lissa/.

Hemer, Oscar
2016 “Writing and Methodology: Literary Texts as Ethnographic Data and Creative Writing as a Means of Investigation.” In *Methodological Reflections on Researching Communication and Social Change*, edited by Norbert Wildermuth and Teke Ngomba, 161-182. Cham: Springer International Publishing.

Hou, Carolyn and Mads Holme
2015 From Inspiring Change to Directing Change: How Ethnographic Praxis Can Move Beyond Research. *Ethnographic Praxis in Industry Conference Proceedings* 2015:190-203.

Jackson, Michael D.
2010 “‘Doing Justice to Life’: A Conversation with Social Anthropologist Michael D. Jackson.” *Harvard Divinity School* website, March 17. Accessed September 16, 2018. https://hds.harvard.edu/news/2010/03/17/doing-justice-to-life-a-conversation-with-social-anthropologist-mich#.

Lombardi, Gerald
2009 The De-Skilling of Ethnographic Labor: Signs of an Emerging Predicament. *Ethnographic Praxis in Industry Conference Proceedings* 2009:41-49.

Nafus, Dawn and ken anderson
2006 The Real Problem: Rhetorics of Knowing in Corporate Ethnographic Research. *Ethnographic Praxis in Industry Conference Proceedings* 2006: 244-258.

Powell, Michael G.
2016 Media, Mediation, and the Curatorial Value of Professional Anthropologists. *Ethnographic Praxis in Industry Conference Proceedings* 2016:2-15.

Reese, William, Wibke Fleischer and Hideshi Hamaguchi
2010 Hyper-Skilling: The Collaborative Ethnographer. *Ethnographic Praxis in Industry Conference Proceedings* 2010:2-13.

Robinson, Rick E.
2009 'Let's Bring It Up to B-Flat' – What Style Offers to Applied Ethnographic Work. *Ethnographic Praxis in Industry Conference Proceedings* 2009:92-107.

Thomas, Suzanne L.
2010 The ~~Martial~~ Ethnographic Arts. *Ethnographic Praxis in Industry Conference Proceedings* 2010:237-249.

Vadrevu, Vyjayanthi
2017 Indian Classical Dance: The Foundational Element in My Practice of Ethnography. *Ethnographic Praxis in Industry Conference Proceedings* 2017. *EPIC* website, 2017. Accessed September 16, 2018. https://www.epicpeople.org/indian-classical-dance-foundational-element-practice-ethnography/.

Wachmann, Jacob et. al.
2018 Below the Surface of the Data Lake: An Ethnographic Case-Study on the Detrimental Effect of Big Data Path Dependency." *Ethnographic Praxis in Industry Conference Proceedings* 2018: forthcoming.

Wang, Tricia
2013 "Big Data Needs Thick Data." *Ethnography Matters* website, March 13. Accessed 18 September 2018. http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/.

Wood, James
2008 How Fiction Works. New York: Farrar, Straus and Giroux.

Designed for Care: Systems of Care and Accountability in the Work of Mobility

ERIK STAYTON
Nissan Research Center – Silicon Valley; Massachusetts Institute of Technology

MELISSA CEFKIN
Nissan Research Center – Silicon Valley

In this paper we explore the idea of a system of care *through a city transit system. We argue that a systematic orientation to care is central to what makes a transit system work for people. Further, we suggest that this care orientation is recognized as such, even though it is not apparent in typical modes of systems management. Care is what knowing in this system is for. We examine how participants in the system navigate different epistemic bases of their work, focusing on how care work and information work intertwine. How is this work practiced and known? And how could we, as design researchers, use these practices to design systems of care? In service of these goals, we expand the notion of care work toward care of non-human actors as well as that of people. We focus particularly on the roles of automation and the risks automation presents for care. In a moment of increased automation in the workplace, what happens to the care of people and things? We argue that the systemic aspect of care, operating at multiple scales toward people and things alilke, is important for maintaining the goods the organization seeks to produce. And we propose a list of critical questions to ask when designing new systems to shape their orientation toward care.*

> In the logic of care exchanging stories is a moral activity in and of itself. But moral activities do not restrict themselves to talk, to verbal exchanges. They also come in physical forms. (Annemarie Mol, *The Logic of Care,* 77)

INTRODUCTION

The radio crackles to life with the voice of a bus operator: "I slid tryin' to make my turn onto Horsetooth and now I've got the median in front of me. I need assistance to back up."

Inside a low set of office buildings and garages painted in tan and beige, several people sit in a control room in front of telephones, radios, and screens. These bus dispatchers are responsible for the operation of all the TransitService buses for this small city in the Western United States. It is clear to everyone present there that the call from this bus operator is potentially serious. But no one in the dispatch center has a clear idea of what has actually happened out at the intersection.

After a few beats of silence, the staff spring into action. One dispatcher summons a road supervisor to a special radio channel. Another asks the operator for more details, finding out that the bus is blocking two whole lanes. The dispatchers work to sort out the situation and find help. Twelve minutes later the situation is resolved. Once a spotter arrives on the scene it takes the bus about 10 seconds to back up and return to its normal route. But the process by which this result was achieved is deeply instructive for how the system operates.

During these 12 minutes, the road supervisor started a long trek across town in rush-hour traffic to get to the scene. Another bus operator who had just returned to base after finishing his shift was sent back out toward the incident in a company car with the thought that he might get there sooner than the supervisor. A dispatcher called the police to inform them of the incident. The dispatchers and the bus operator deliberated about whether it was safe to put out cones and to let a passenger off the bus while it was sitting in that location. The very reason the bus could not back up was itself for the safety of others: operators in this system are not allowed to back up a bus without a spotter, not because it is impossible, but because it is considered too risky for others in the vicinity. In the end it was yet another bus operator returning home from his shift who was first on the scene and provided the decisive help. And that was not by chance—a dispatcher knew this operator, and knew when, where, and how he would be returning home. That dispatcher called him directly to ask for his assistance.

As this event demonstrates, TransitService, like many service organizations, is distributed in space and deeply interdependent with infrastructures beyond its sole control. Corporations today are increasingly designing large-scale service systems—from ubiquitous electronic communications to expanding ride-sharing—that are interlocked with infrastructures managed by other parties. And service work within these systems is increasingly automated or eliminated, often displaced onto the users themselves. But we see in this particular mobility system a counter-argument to the prevailing logics of efficiency.

The overarching motivation for solving this problem was clearly to get the bus moving again. But throughout, there was a marked attention to the care of employees, equipment, and members of the public, all legitimate and important work that service organizations perform that is not directly *business* labor.

We selected the dispatch center as a site for our research with the objective to understand how distributed mobility systems operate, and how they might operate using increasingly automated vehicles. We were especially interested in the role and function of centralized control centers, here, Dispatch. Orchestrating the movement of people and vehicles around a city requires gathering information from a variety of perspectives: from the ground-level viewpoints of buses to the overview visions of dispatchers, to everything in between. The data that are gathered and tracked allow a mobility system to operate in a world that can be known only imperfectly and at a distance, mediated through recording devices, radios, and employees. But we found that the self-contained logic of *information systems* that reinforces practices of measurement and data collection—feeding the system with yet more information—is not enough to ensure that the goals behind the system can be met. In other words, information is not enough for mobility to be achieved. Another system is also necessary, and serves as the lifeblood of mobility services provided to people and publics. We describe this parallel and intertwined system as a *system of care*.

The system of care strongly contributes to the achievement of mobility: it organizes and shapes the purposes of knowing and measuring in this site. Actors seek to know the system in order to care for it and the people it serves. And this care overflows traditional regimes of measurement, prompting questions about how care itself can be known, assured, or documented (Strathern 2011, Adams 2015). Ethnographers expect to find ways, both helpful and harmful, in which work practice does not match up with regimes of measurement (Suchman 1995, Cefkin, Thomas and Blomberg 2007). But we extend this perspective to show that these practices may in fact become part of a systematic organizational approach to carrying out activity in the world, with implications for the design of increasingly automated service systems.

For us as researchers, seeking to conceptualize how systems to perform mobility-work could be automated, the systemic nature of care confounds traditional ideas of how this labor could be formalized into computational procedures.

This paper is a call to ethnographic researchers, designers and strategists to take seriously the multifaceted system by which caring for people and things happens. In this paper we raise specific questions about what it means to design for care. Our particular interest lies in ensuring that increasing automation does not erase these desirable properties of caring systems.

DATA AND METHODS

The authors conducted one week of intensive fieldwork in the bus system of a medium-sized city in the Western United States. We collected fieldnotes in the dispatch center, during ride-alongs on buses, and in a number of staff meetings. We interviewed key staff members. And we collected photographs and hours of video and audio recordings per day in the dispatch center and on bus rides. Our overall engagement with the site was somewhat longer, including several calls in advance to scope the observation and identify useful sites to observe, and some follow-up to collect more documents to help us understand the service and its activities.

We analyzed the material with attention to how workers in this site come to know about the world and guide their actions as part of the bus system, with particular focus on the situated purposes and meanings of these actions. What types of knowledge are valued over others? Why are certain kinds of information collected? What is the affective engagement of employees in the work? Additionally, we analyzed the work-practice dimensions of the material, particularly how employees work together to solve problems. Methodologically, the authors transcribed and coded relevant selections of conversations for these themes, and performed video and interaction analysis using collected audiovisual data. To supplement our own data collection we were given the Operations Manual to study, and have returned to it to put what we saw in context.

LITERATURES

In the ethnographic literature, care work appears as processual and relational work between people. Our notion of the system of care builds from Annemarie Mol's concept of the *logic of care* (Mol 2008) and extends the way that care work is thought about in large, distributed organizations. In healthcare (Mol's subject matter), choice appears under the guise of individual decisions and care is achieved over time by informing, providing support, cajoling, working with, or otherwise providing assistance to patients. Mol and others have found that care work is commonly not acknowledged by professional standards and measurement systems even though it is critical to the functioning of organizations (Nardi and Engeström 1999, Star and Strauss 1999, Bowker and Star 1999). Furthermore, care labor often occurs outside the market frame, and is uncompensated, situated in the home, and provided primarily by women (Cowan 1983), though some of this typically household labor is increasingly corporatized and built into business organizations seeking to provide for greater portions of their employees' needs (English-Lueck and Avery 2014).

Service labor in the market sphere is often rendered invisible or otherwise "veiled" (Kreeger and Holloway 2008). Excavating the real content of the work day unearths all manner of labor that is illegible to official reports and not present in job requirements (Suchman 1995). But making it visible in order to gain official recognition is also a fraught enterprise,[1] and measuring its value may be difficult or even impossible (Bryson et al. 2003). The services literature shows one dark side to making this work too visible: the routinization of service work. While standardized practices provide a sometimes-welcome buffer for employees from genuine interactions of care, they concomitantly reduce the space for individual problem solving and judgment (Leidner 1993).

We see the system of care as operating at multiple scales, for people and objects alike. Relational care work is closely tied to the work of customization, maintenance and repair (Star 1999, Darr 2008, Jackson 2014, Anand 2015). And all of these may be part of care work oriented toward larger agglomerations of people and things alike: from users to communities and publics that may be more diffuse and therefore less visible (Amirebrahimi 2006). Numerous studies have shown how transplanted technical systems may worsen rather than improve conditions if not maintained (for instance Mavhunga 2014), suggesting an emphasis on simplicity and reparability of technical interventions (De Laet and Mol 2000). Maintenance is therefore a key entry point for practices of care toward non-humans, which may yet expand beyond material repair into responsible use, preemptive action, and continued monitoring.[2] Within infrastructure studies (Parks and Starosielski 2015, Larkin 2013) it is maintenance that allows technical systems to "just work" and be a largely stable and invisible background to a "modern" life (Edwards 2002, Anand 2015). Care and maintenance are particularly interesting here when they involve stewardship of a third space like the roadway, which is beyond the direct responsibility of either an organization or its customers.

HOW SYSTEM OF CARE OPERATES FOR DIFFERENT ROLES

Mol argues that the logic of choice is a neoliberal, modernist logic that puts the responsibility for the fate of the individual squarely on his or her own shoulders. So while it aligns with commonly held Western cultural values, it also represents a systematic abrogation of responsibility for individual people, communities, and networks of people and things. The logic of care, on the other hand, involves sustained relationships of responsibility, even if via some level of force. We see a similar dichotomy within the service organization, between care and efficiency. Efficiency, like choice, sounds appealing, but falls short by a similar abrogation of responsibility.

Many important facets of the service and infrastructure-related work described below are not easily amenable to measurement, and measurement tends to distort labor over time to focus more and more on what is measured rather than what is not (Adams 2015). Since so much care of people and things is hard to measure, it seems to face a great risk of being left by the wayside by a focus on automation and efficiency.

To refocus the discussion on the values of care within an organization requires us to look for instances of care in the process of work. What are the signs that this care is happening? How do people in this site know care? Care may be written into policies, procedures, and mission statements, but unless these are lived and practiced they might remain as just so many words. While regimes of accountability seek to provide quantitative evidence for certain kinds of care, from daily checks of bus tires to multi-year infrastructure projects, the caring properties of the system depend on the performance of values that are difficult or perhaps fundamentally impossible to account for by procedural means. Here we attempt to account for them ethnographically, by looking at the meaningful performances and narratives of employees in various roles within the service system: bus operators, field personnel, dispatchers, and support staff.

Bus Operators

> *Hold on, hold on please!* A loud beeping commences as the bus's ramp lowers into place.
> *Hey Walter, how you doin'? You get to go first!* says the driver, as he gets out of his seat to assist.
> An elderly man with a walker gets on Bus 5 and moves toward priority seating.
> *Haven't seen you in a couple weeks.*
> *You good?* The operator asks as the man gets settles in his seat. *Does this [the walker] have brakes?* The operator finds and sets the brakes.
> *Where you getting' off? … Just ridin' around?*

In many ways, driving is not the core activity of a bus operator. As our informants suggested to us, "anyone can be taught to drive a bus." This is not to say that safe operation does not require high levels of skill and attention. And indeed the operator's manual tells that "The most important element of your job as a transit operator is Safety First" (Operations Manual, p. 159). While safe driving is of paramount importance when on the road, even at the cost of other metrics like on-time performance, it is not seen in practice as the distinguishing feature of the operator within this organization.

As routine interactions like the one above show—interactions which happen day-in and day-out for TransitService operators—the role of bus driver is as much about making people safe and comfortable as it is about driving. Operators are responsible for the critical work of aiding customers with mobility issues, such as helping passengers with walkers or wheelchairs board the bus and get secured into the appropriate restraints. This labor combines physical assistive activities that might be automated by a specially-designed vehicle, and communicative work that would be far more difficult, especially when customers themselves have issues communicating.

In the hiring practices of the organization itself, the position of bus operator is primarily a customer service role. The critical skilled task for the operator is interacting with people, a skill that at least some of our informants believed was innate and un-teachable. So they would pass up applicants with existing commercial driving licenses in favor of people with more service experience who would nevertheless require training to be certified as a vehicle operator. Operators must be able to attend to others, chat, cajole, assist, maintain order, and defuse situations. They, like the assistants in a store or the instructors in a classroom, are the stewards of the people who enter their space.

These interactional and service roles set up the operator as more than just a driver. She is a local information source, a help line, and a custodian of public trust, and in that way she joins the ranks of care workers. Operators are instructed by procedures to perform some tasks even when the bus is done for the day, such as to transport stranded passengers if their destination is on the "deadhead" route back to the garage.

Figure 1. Safety is so important that even the restrooms are not an escape from messages about safe driving. But much of what operators do on the road is properly customer service. (Here we have obscured this operator's face and the service's logo.)

Moreover, the orientation toward care configures other kinds of work that operators end up doing, much of which is not required by official procedures. For instance, operators may call in to dispatch to request welfare checks or emergency services for incapacitated people at stops or by the side of the road. They also report issues with roadway infrastructure such as street lights and traffic signals that are out, detour signs down in the roadway, and signage issues at bus stops. The operator-bus unit acts as a kind of roving information system, oriented toward service to the community. This kind of informal effort and attention contributes to making the world outside the bus a safer place.

Operator roles blend the care of people on and around their vehicles with the care of infrastructures and vehicles themselves. They are looking and listening for problems or potential problems, even ones that do not directly impact their own position or even the TransitService organization. In practice, the operators' commitments to care appear to run deeper than what is regulated or required, and at times occur outside of existing regimes of measurement. As the sensory organs of the transit system, they sometimes do not know the significance of the things they report. They do not see the whole picture. But their attention and conscientious reporting are a key part of how the rest of that system, including road supervisors and dispatchers, knows what is going on.

Road Supervisors and Field/Training Coordinators

Dispatch calls out: *500 to an available Road Supervisor, Admin 1 please.*
309 Admin 1, Lincoln and Linden, the RS responds, giving his location.
Shit, he's on the wrong side of town, the dispatcher says as an aside, off the radio. *Do we have a TSO out there?* She turns to her computer to look.

Another dispatcher picks up the radio: *309, Route 12 says she slid trying to make her turn onto Horsetooth from Taft and needs assistance backing up. She's currently blocking traffic at the moment. What's your 10-20?*
I'm at Lincoln and Linden, her 10-20? says the RS, requesting the operator's location.
Horsetooth and Taft.
. . . There is a ten second pause.
OK I'll head that way. I'm a bit out, but I'll head that way, Taft and Horsetooth.

Bus operators are not the only workers who function as eyes and ears for the Dispatch Center. Road Supervisors and Field/Training Coordinators (FTCs) also roam the field to track of what is happening across the area serviced by TransitService. Piloting white vans stocked with essential tools that bus operators would not normally carry, they carry out simple field repairs: un-jamming fareboxes, opening and closing bus stops, and performing accident investigations. They check capacities at stations, deal with problem customers, perform welfare checks on bus passengers, and pull DVRs from buses in the event that accident footage is needed immediately.

As we see in the radio chatter above, the Road Supervisor was the first line of contact when a bus got stuck and needed help to back up, the subject of our opening vignette. Road Supervisors and FTCs perform a range of such ad hoc tasks intended to keep the system on track, such as driving ahead of a bus that is running late ("running down" in TransitService parlance) to pick up passengers at select stops and thereby help the bus get back on time. They may also pick up passengers stranded by a bus that has broken down or missed a stop, or was unable to pick passengers up due to a lack of space on-board. Or, they may drive to bus stops and close them down during a detour, or re-open them after a detour ends, to make changes in service apparent to users trying to catch a bus.

Road Supervisors have a central role in making the physical world match up to the organizational, scheduled world of bus routings. And maintenance and monitoring of the built infrastructure that service depends on are a major part of their responsibilities. Field/Training Coordinators, who work with the Road Supervisors, are in part responsible for producing this organizational culture, which values the respect that operators show to the roadway and to others. This attention comes through in their educational tasks. From continued driver training to leading the bus operator staff in reviewing videos of "near misses" and "good moves" (pulled from bus DVR footage) that demonstrate and dramatize good driving, FTCs are a key part of the maintenance of a safety culture that protects the organization as well as the wider public. Road Supervisors and FTC's are prototypical examples for how humans' flexibility and adaptation helps to keep systems operating, and they are essential elements of the transit system of care.

Dispatchers

Unit 46 to 700. A bus operator radios in on the Maintenance channel.
473, a dispatcher responds with his operator ID, identifying himself.
Uh, I just got into the South Transit Center off of Fairway, my low coolant light came on, and the bus has officially, uh, shut down.
Copy, so you are still on Fairway, or you actually made it into the transit center? The dispatcher pulls the bus up on the real-time service map.
Uh, I am still on, uh, Fairway just past the old pine going into the Transit Center.
Um, copy, um, probably what they'd have you do first is just shut down the bus, um, just turn it off completely. We'll just see. I'll get a hold of 700 and they'll come on to the channel.

Copy. The operator responds, and works on restarting the bus.
The dispatcher picks up the telephone: *Hey Mike this is Rob. We have unit 46 on Fairway, um, said that he had a low coolant light come on, buzzers and everything, and then the bus shut down. So it sounds like he's possibly dead in the water.*

Dispatchers are the control center for the transit organization. They track the buses, anticipate problems, collect information coming in from different sources, make sense of what is happening outside in the space of the street, and, as in the vignette above, connect people in the field who need help with resources that can assist them. Most critically, they are the source of ground truth for the organization. As one dispatcher described their role: "We are here to remember." In the situation above, the dispatcher answered the operator's call, got maintenance on the phone and then on the radio, coordinated in person with maintenance personnel as to where they would find the bus, and sent out an operator in a replacement bus. At the same time, she also filed incident reports, told the other operator where to put the bus once he got it started again, logged the operator onto the new bus when it arrived to speed the departure process, and used CCTV cameras to check the progress of the bus swap. So dispatchers are responsible for many kinds of communication and administrative work. They attend to and filter the details coming from different sources and media, and know when the answer to a question needs to change. The dispatch center, the dispatchers told us, is "a big giant filter." It is responsible for knowing, and informing others: "at this moment, what is true?"

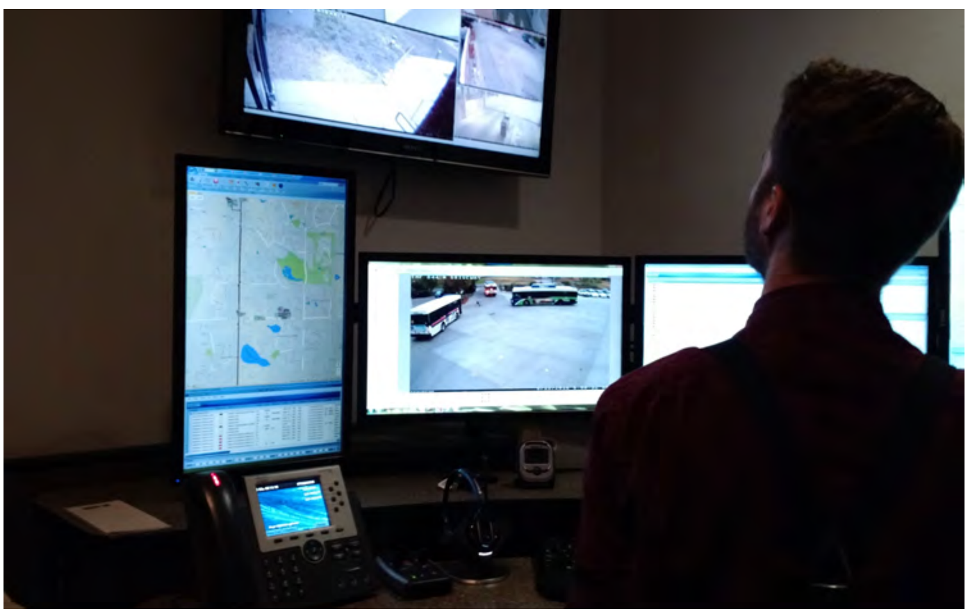


Figure 2. Dispatchers manage multiple streams of information. And they use tools like CCTV creatively to address and preempt problems, and to care for the service as a whole.

Dispatch is organized around a core of adaptable workers, who are oriented toward caring for the organizational system and those that it serves. The work of dispatch requires a kind of *bricolage*, using tools in nontraditional ways and making the best of what is available in order to provide service. We do not wish to suggest that dispatch has no set processes or standards—indeed they are the prime process-setting and standards-keeping body for the rest of the operational personnel, ensuring that bus drivers, road supervisors, and others are following the rules and meeting expectations. But, at the same time, the labor of dispatch is not at its heart about procedures or metrics.

Dispatchers' work is therefore a kind of techno-service work, in which they may be managing information technologies one minute, doing administrative work another minute, and giving instructions to an operator at the check-in window the next minute. Dispatchers use a mix of computer systems, phones, radios, printed materials and both informal and formal information channels to sort through challenges. In several instances, public data sources were involved. We observed Dispatch reading Twitter and thus finding out about a potential disturbance at a busy bus stop, a prize drawing set up there as part of a local festival. In another instance that was reported to us, after a bus operator informed Dispatch of slow traffic, the dispatcher used one of the 4 computer screens in front of her to pull up and scan Twitter, finding reference to an accident at a major intersection near that location. In a different kind of situation, an operator reported an unexpectedly early street closing, due to a downtown special event, over the radio. This prompted a cascade of fact finding in Dispatch about what options were available, whether prior information was missed, and how to direct numerous routes around the streets.

This practice of flexible, bricolaged service work includes using tools in nontraditional ways. When we asked about their use of the security cameras to check on the status of the broken-down bus in the opening vignette for this section, one dispatcher noted that they aim to use the cameras carefully, looking only at TransitService property to protect the privacy of others in the community. But she also told us a story suggesting how important these cameras are as a general use tool. Once, when a bad snowstorm came through, most of the routes were out of service. Dispatch could page transit centers via a one-way audio link to make announcements but had no return channel for communications except for the security cameras. So they asked passengers at each station to raise their hands as to which direction they wanted to go, and then used the cameras to manually count customers. They then allocated and dispatched the appropriate vehicles to get the passengers to their destinations.

There are all kinds of metrics implicated in this story, but the service criterion that was most clearly valued by dispatchers as they discussed their decision-making process was not to leave people behind. That desire not to strand passengers extends to directing Road Supervisors or Field/Training Coordinators out in vans to pick up passengers and take them to their destinations.

And all the reports that operators and Road Supervisors or FTCs make about infrastructural problems outside of the jurisdiction of TransitService do actually go somewhere. We witnessed dispatch reporting infrastructure issues to third parties, such as calling a local police department to let them know about problems with traffic lights. The desire and indeed sense of responsibility to report such things appears to us to be a recognition of the organization's dependence on public infrastructure, and its shared use of the roadway. It is a way of giving back to a shared infrastructure the organization needs in order to survive, as well as a way of making things safer for others. This sort of care work is tied up with TransitService's sense of good organizational citizenship.

Maintenance, Customer Service, and Management

Rounding out our description of the key organizational departments and employees of the bus system are the maintenance and shop personnel, management, and customer service departments.

Working behind the scenes to keep buses in good condition, the maintenance and shop personnel are a key part of TransitService operations. Of course, they are responsible for doing responsive repair work when bus operators identify issues in their pre- and post-trip checks. But preventative maintenance tries to anticipate issues that may crop up later, and requires flexibility and judgment. Conversations within Dispatch and between dispatchers and repair personnel made clear that buses have distinct personalities. They are not mass-produced, identical, and interchangeable. As dispatchers told us, "the worst thing in the world is a brand new bus," a bus is "no good until it has been in service for a year." These buses are hand built, and are all slightly different, so they will also fail differently. The teams of mechanics need to learn the quirks of each bus, and find ways to diagnose specific kinds of problems: one bus has a tendency for the front marquee to get stuck, and the best fix is to hit it from the outside with a broom handle. No manual could ever have provided that information!

There are also all sorts of business-oriented personnel working behind the scenes to keep TransitService running smoothly. Management is in charge of making the bus acquisition decisions, Analysts and IT track buses and passenger loads, Customer Service Representatives answer service questions, and Service Planners identify new routes to better serve the community. Many of these tasks are tied together by the data that make it possible to determine the value of current service. In a tech-focused culture in which machine learning and big data analytics consultants promise almost perfect knowledge of the world via the collection and use of ever increasing amounts of data, it is little surprise that managers at TransitService want detailed knowledge of what their vehicles are doing, how many people they serve, what times are busy and what times are not, etc. And yet, the organization does not collect everything it could collect given the tools it has. Though they collect lots of vehicle data already, they chose not to purchase an engine telematics tool offered by their service management software provider as they felt it would be to "surveillancey" for their drivers' comfort. How data is used, or perceived to be used, is perhaps more consequential than what the data is (Nissenbaum, 2009). And these decisions about how to use the available resources rely on human judgments about role, responsibility, and relationships that are not amenable to programmatic treatment.

Beyond employee care, customer service is primary in the organization's vision and mission statements. The first things in the operation's manual after the table of contents are two pages of vision, mission, and goals. Among them is a page of verse, written about service to passengers, which is again repeated on page 182, in the middle of the bus operator's section of the manual. This litany goes:

> "Our Passengers
> **Are the most important people in our business**
> Our Passengers
> **Are not dependent on us, we are dependent on them**
> Our Passengers
> **Are not an interruption of our work, they are the purpose of it**
> …

Are not cold statistics; they are human beings with feelings and emotions just like you and me …"

While we are accustomed to thinking of vision and mission statements as aspirational texts that may have little impact on the day-to-day operations of the organization, it appeared to us that TransitService really did attempt to live by these goals. At one of the shift meetings we observed, staff announced that they had "lost one of [their] long-term passengers in a very unfortunate way," and commemorated the deceased customer with photos on a section of the whiteboard. We observed these mission, vision, and customer statements performed in the everyday comportment of the organization, in ways not fully accounted for by the reductive sets of functional measures and efficiency metrics that an accounting of the organization is likely to stress. The sense that the organization needs to be a "responsible steward of transit resources" comes through in care of communal transportation infrastructures.

THE SYSTEM AS A WHOLE

These individual situations of care work appear to form a system that goes beyond just the sum of the parts. Individually, a lone employee caring for passengers or equipment will make little difference to the operation as a whole. But care in this organization is systemic: it occurs at all levels within the organization, over varied timescales from days to years, and in ways that support each other across these categories through both formal procedures and informal, individual action. Buses will always have mechanical failures and break down. But multiple ways of attending to the problems of equipment—staggering bus acquisitions in time, replacing vehicles promptly at end-of-service-life, doing preventative maintenance regularly, performing pre- and post- checks daily, informal learning of how to attend to an individual bus's unique problems, and dispatching appropriate buses to appropriate routes—sum together to balance the needs of buses against the needs of the organization and the publics that it serves. Nothing is run on the ragged edge. And service continues.

The idea that systems require a delicate balance of labor to maintain them is nothing new. But what we show is that much of this work, whatever else it may be, is also care work, and caring work. It involves taking care of people and things. And it is performed with a sense of meaning and responsibility, rather than just by a procedure. We have talked about the system of care in terms of the roles of its constituent people. But it is also possible to turn this view around, and look at what properties and things are cared for. This helps makes clear the systemic and mutually reinforcing nature of the care practices involved.

As vehicles are cared for through fueling, daily checks, preventative maintenance and repair (and the vehicle fleet is cared for through the timing of acquisitions), so too are employees cared for through training, safe working conditions, responsible scheduling and information sharing. Customers are cared for via accurate transit information, safe and clean vehicles, on-time service, and welfare checks. Broader publics are cared for through avoidance of obstructions or delays caused by buses, by road safety, by reporting incidents with other infrastructures, and by responding to public desires.

And above all, these separate sub-systems of care work put together serve to maintain and allow for the movement of people and vehicles. Movement is maintained by coordination and timely communication within the vehicle system, and between the vehicle system and other municipal services. Movement is also maintained through the care of other constituents: the vehicles, employees, customers, infrastructures, and publics for and by which the vehicle system operates. Only because of attention to all the facets of the vehicle system can people continue to move through the system; without this attention to care: buses break down, employees get into accidents, customers are injured, the public is obstructed, and infrastructures degrade and become unusable.

Our research does *not* show that the system would cease to function as a system without any single one of these components. There is clearly flexibility in the practices of the organization. There are other kinds of work that are not done by this system, and one could have a successful mobility system with only some of these labors and not others. But take away enough of these activities, we think, and the system would collapse like Aramis, abandoned by Matra for a lack of love (Latour 1996). Either its ultimate effectiveness, or at least its qualitative character of competence and professionalism, would be compromised. By saying this is a system of care, we are saying that individual judgments are involved, but something is happening at a level above the individual, in a collective of multiple components that all pitch in to keep the system running.

DESIGNING FOR CARE

> "You just kind of take it all at the same time and just kinda sort it out, which one's more pressing. Yeah I dunno, you just kinda do it."
> (A dispatcher describing how they manage incoming information)

Our purpose in examining TransitService was to learn about what it takes to move a fleet of vehicles around a city, picking up and dropping off passengers, in order to learn how a mobility service could be automated. Above, we have tried to describe the work of mobility, and the way that that work often overlaps and intersects with types of care. So our next question is: what implications does this have for the design of increasingly automated service systems? How can we use the system of care formulation as material for design?

TransitService stands as an example of the centrality of care to successful service work within larger worldly infrastructures. Someone needs to be watching out for slow-moving, dispersed, or emerging problems, even those that affect other systems outside the purview of the organization itself. What the care lens shows us in this case is that the official procedures of care are not the entirety of the organization's care work. Within the grey areas beyond and around written procedures, the success of that work also depends on small decisions that are within the realms of individual's decision authorities. This labor becomes visible to others via co-presence and collaboration—being part of a shared culture over the radio, in the Dispatch Center, and in staff meetings.

The situations in which care work is needed, and the flexible kinds of labor that are in evidence in these situations, make it likely impossible to strictly define and proceduralize all of this work. Real world decision situations present complex choices between often unknown alternatives, and actors are guided not just by procedures but by judgment and values. Increasing amounts of IT-based tools have helped Dispatch manage the chaos of the external world—it was much more disorganized, we were told, back when it was run using Excel spreadsheets rather than purpose-built fleet management software. But Dispatch was not fully automated: dispatchers who are trusted and expected to act with good judgment and autonomy are still a critical part of the organization.

And simply involving a few isolated humans is not enough. The uncaring strictures of human bureaucracies are unpleasant, and powerful, because of their indifference or antipathy to the people they affect (Herzfeld 1992; Graeber 2015). "Just following the rules" leads to helplessness in the face of a system that was not designed for, but against you. What TransitService shows is that care work is successful when humans are backed up by an organizational ethos, a culture, that gives them that autonomy and provides a field of cultural resources by which to orient themselves and from which to draw in order to solve the problems they encounter. As we saw at TransitService, an ethos of responsibility, customer service and risk-management is baked into the procedures, meetings, trainings, and everyday activities. Designing for care means creating a culture that fits the operations, as well as balancing human and machine roles.

Here, thinking about the system as a system of care overlaps with many of the social critiques of capitalism: the problem is that caring so often requires attending to the externalities of the system (in this case, the negative ways that buses wear down roads, or annoy neighborhoods, and the positive ways that they can be used as a platform for community health by reporting people who may be having medical issues near the roadway). And these externalities are not often captured in the value structures of the system because they are too fuzzy, obscure, or far-off to seem to matter.

So the risk associated with the design of increasingly automated service systems is that the positive, caring characteristics of the organization outlined here will be lost. Our research suggests five key questions that emerge as part of a careful design process for automated service systems.

First, is the design investing in skill and autonomy or trying to eliminate it? This intersects with the long debate between Intelligence Augmentation (IA) and Artificial Intelligence (AI). Are human skills being augmented and human autonomy increased, or is work being routinized and labor replaced wholesale? The care lens suggests investing in IA and aiding accountable autonomy. It suggests giving people better tools to handle the uncertainties in the world rather than hiding uncertainty behind a veneer of objectivity, or steamrolling it by inflexible process. As Mol suggests, care can involve restricting autonomy. But design for care needs to be very deliberate about the cases in which it does so, and should do so in the service of larger, processual accountabilities between the system and those around it.

Second, is the design attending to where work will shift? Automation does not eliminate work, it shifts it around. So where does work go, and what will happen if it falls off the map? If bus operators are computerized, who assists the passengers? If Dispatch becomes an automated load-balancing process, who is left to make judgment calls, find a lost item, or inquire for a customer about another service? It is tempting to simplify the system and cut costs by declaring some tasks out of scope, but that work does not disappear simply because the designers of the system wish it would. Design for care should account for where actual work-in-practice, not just theoretical work-on-paper, shifts.

Third, is the design creating a structure that will do the extra little bit to serve people well? Service, even of machines, has a human touch. So how will an automated system go the extra mile for the customer? Design for care needs to figure out a structure around automation that can make exceptions, address problems, and provide facilities for long-term restructuring. For instance, whether a route is technically possible does not mean that it is organizationally feasible, given TransitService's values. We observed clear concern for not having buses turn around in neighborhoods, and not taking the buses onto streets where the heavy vehicles would cause undue wear. The system needs to be able to behave appropriately on the margins where numbers and procedures are not enough.

Fourth, is the design measuring the right things? And is it using them appropriately, and leaving room for uncertainty and judgment within reasonable limits? Care should be taken as to what is measured, and how it is used. To try to measure everything is tantamount to producing the Borghesian map, as complicated and unwieldy as the full world that it represents. And metrics are always open to being twisted or gamed. Vehicle location is a caring sort of automation when it is then used to track bus times, provide them to the public, and thereby better the overall service. But it has the potential to be a very uncaring one toward operators if it is used to surveil. Certain types of accountability may even be caring toward one group and uncaring toward another at the same time. These are difficult balances to be thought-through specifically, not left to chance. So design for care needs to pick its measures carefully, and combine measures with accountable judgment about their applicability and validity.

Finally, is the success of the design accountable to the unmeasurable goods that it is hoped to produce? If designing for care is ultimately about providing better service, we must recognize service has many meanings and definitions. It may mean different things to different people. And primarily it is going to be defined not by a number of people served, or a length of time in a trip, but by the values that structure the work. So design for care needs to ask: who keeps track of the success of things that seem important, but do not have a quantity or procedure attached to them? Caring means reserving a space for doing the things that are qualitatively "right," for performing values of the organization, even when the results cannot be measured. And it means determining what those values are, what counts as "right" in a situated context. Here is a place for intervention via "reflective tools" to account for and shift organizational processes in a participatory way (Cefkin 2011). This is a method of keeping the organizational unit attending actively to qualities of its service that are experientially apparent but difficult to quantify.

These questions are not the end of the conversation. They are a starting point. But they should push us to think about care in organizations at an ecological level. As we have shown, care is about things as well as people, infrastructures as well as human relationships. Care is about flexibility, judgment, and values. Care is about balancing assistance and augmentation, autonomy and responsibility, in increasingly automated service work. In designing service organizations as systems of care, we need to take up long-term, responsible visions of the organization's place in a wider world and ensure that automation and optimization do not wipe these out. Care is a process, not a product: it is in the relationships of living- and working-with that care happens.

Erik Stayton is a PhD candidate in the Program in History, Anthropology, and Science, Technology and Society at MIT. He investigates human interactions with AI systems, and currently studies the values implicated in the design, regulation, and use of automated vehicle systems. He also interns at the Nissan Research Center.

Melissa Cefkin is Principal Researcher and Senior Manager of the User Experience group at the Nissan Research Center in Silicon Valley. She has had a long career as an anthropologist in industry, including time at the Institute for Research on Learning, Sapient, and IBM Research.

NOTES

Acknowledgments – We thank our colleagues and collaborators in our lab whose ongoing projects and open questions prompted and shaped this research. We also appreciate the detailed commentary and guidance from our EPIC reviewers. But we especially want to extend our heartfelt thanks to all the staff of TransitService, especially the operators we rode with, the dispatchers we observed in action, the road supervisors, managers, FTCs, and everyone else who spoke with us during this project, and was so willing to share about their work and endure the ethnographic gaze. We deeply appreciated the chance to see what you all do, and to render it into this text.

1. "Making work visible" is another way to talk about measuring or accounting. In the background of the literature and the rest of this paper, accounting and measurement are the joint fulcrum that both allows for the automation of work and threatens to overwhelm valuable and difficult to measure aspects of labor by prioritizing the parts that are easier-to-measure. An emphasis on quantification is especially distorting of the kinds of value that can be accounted for (Porter 1995), and it is in part for this reason that an ethnographic perspective is used to uncover these aspects of care work.

2. We think it is critical to expand the definition of care work, in this context, to include the care of nonhuman actors. Maintenance has procedural aspects, but competent maintainers use all their faculties and flexible judgment to anticipate, prevent, diagnose, and repair problems. They also need to maintain social networks at the same time as they maintain machines (Orr 1996). Furthermore, we find in our site a similar affect in the attention to objects and infrastructures as to people. In excess of regimes of formal measurement (such as maintenance schedules), maintenance proceeds by guesses, reasoned judgment, and intuition, motivated by values such as care.

REFERENCES CITED

Adams, Vincanne, ed.
2015 *Metrics: What Counts in Global Health*. Durham, NC: Duke University Press.

Amirebrahimi, Shaheen
2016 The Rise of the User and the Fall of People: Ethnographic Cooptation and a New Language of Globalization. *Ethnographic Praxis in Industry Conference Proceedings* 2016.

Anand, Nikhil
2015 Leaky States: Water Audits, Ignorance, and the Politics of Infrastructure. *Public Culture,* 27, 2:305-330.

Bowker, Geoffrey and Susan Leigh Star
1999 *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Bryson, John, Peter Daniels and Barney Warf
2013 *Service Worlds: People, Organizations, and Technologies*. London: Routledge.

Cefkin, Melissa
2011 Comment on Jane I. Guyer: Blueprints, Judgment, and Perseverance in a Corporate Context. *Current Anthropology* 52, S3 (Supplement to April 2011):S17-S27.

Cefkin, Melissa, J. Thomas and J. Blomberg
2007 The Implications of Enterprise-wide Pipeline Management Tools for Organizational Relations and Exchanges, *GROUP '07*: 61-68.

Cowan, Ruth Schwartz
1983 *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave*. New York: Basic Books.

Darr, Asaf
2008 The Rise of the Techno-Service Sector: The Growing Inter-Dependency of Social and Technical Skills in the Work of ERP Implementers. *Ethnographic Praxis in Industry Conference Proceedings* 2008.

Edwards, Paul N.
2002 Infrastructure and Modernity: Scales of Force, Time, and Social Organization in the History of Sociotechnical Systems. In *Modernity and Technology*, Thomas J. Misa, Philip Brey, and Andrew Feenberg eds. Cambridge, MA: MIT Press.

English-Lueck, J.A. and Miriam Lueck Avery
2014 Corporate Care Reimagined: Farms to Firms to Families," *Ethnographic Praxis in Industry Conference Proceedings* 2014.

Graeber, David
2015 *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy.* Brooklyn, NY: Melville House.

Herzfeld, Michael
1992 *The Social Production of Indifference: Exploring the Symbolic Roots of Western Bureaucracy.* Chicago, IL: University of Chicago Press.

Irani, Lilly
2015 Justice for 'Data-Janitors. *Public Books*, January 1, 2015. www.publicbooks.org/justice-for-data-janitors/

Jackson, Steven
2014 Rethinking Repair. In *Media Technologies: Essays on Communication, Materiality and Society*, T. Gillespie, P. Boczkowski, and K. Foot, eds. Cambridge, MA: MIT Press.

Kreeger, Lisa and Elizabeth Holloway
2008 Now You See It and Now You Don't: Consequences of Veiling Relational Work. *Ethnographic Praxis in Industry Conference Proceedings* 2008.

De Laet, Marianne and Annemarie Mol
2010 The Zimbabwe Bush Pump: Mechanics of a Fluid Technology, *Social Studies of Science*, 30,2: 225-263.

Larkin, Brian
2013 The Politics and Poetics of Infrastructure, *Annual Review of Anthropology* 42, 1: 327-43.

Latour, Bruno
1996 *Aramis, or, the Love of Technology.* Cambridge, MA: Harvard University Press.

Leidner, Robin
1993 *Fast Food, Fast Talk: Service Work and the Routinization of Everyday Life.* Berkeley, CA: University of California Press.

Mavhunga, Clapperton
2014 *Transient Workspaces: Technologies of Everyday Innovation in Zimbabwe.* Cambridge, MA: MIT Press.

Mol, Annemarie
2008 *The Logic of Care: Health and the Problem of Patient Choice.* London: Routledge.

Nardi, Bonnie and Yrjö Engeström
1999 A Web on the Wind: The Structure of Invisible Work, *Computer Supported Cooperative Work* 8: 1-8.

Nissenbaum, Helen
2009 *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford, CA: Stanford University Press.

Orr, Julian
1996 *Talking about Machines: An Ethnography of a Modern Job.* Ithaca, NY: Cornell University Press.

Parks, Lisa and Nicole Starosielski
2015 *Signal Traffic: Critical Studies of Media Infrastructures.* Urbana, IL: University of Illinois Press.

Porter, Ted
1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton, NJ: Princeton University Press.

Star, Susan Leigh and Anselm Strauss
1999 Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work, *Computer Supported Cooperative Work* 8:9-30.

Star, Susan Leigh
1999 The Ethnography of Infrastructure, *American Behavioral Scientist,* 43: 377-391.

Strathern, Marilyn, ed.
2011 *Audit Cultures: Anthropological Studies in Accountability, Ethics, and the Academy* London and New York: Routledge, 2011.

Suchman, Lucy
1995 Making Work Visible, *Communications of the ACM*, 38:56-68.

Reading the Tea Leaves: Ethnographic Prediction as Evidence

CLAIRE MAIERS
WillowTree, Inc.

Those who work in research know that we live in a world that is strongly influenced by what Tricia Wang has called the quantification bias. More so than other forms of information, numbers have incredible formative power. In our culture, numbers are seen as trustworthy representations of reality that are strongly associated with objectivity and untainted by human bias and shortcomings. Recently, data science, big data, algorithms, and machine learning have fueled a new wave of the quantification bias. One of the central fascinations of this wave has been the promise that humans now have the power of prediction at their fingertips. In this paper, I reflect on what it means to make predictions and explore the differences in how predictions are accomplished via quantitative modeling and ethnographic observation. While this is not the first time that ethnographic work has been put in conversation and in contrast with quantified practices, most theorists have framed the role of ethnography as providing context to that quantified work. Here, I argue that ethnographers produce predictions in their own right. I begin by discussing what it means to predict something, focusing on its function. This is followed by a discussion of the ways in which predictions are constructed through both machine learning and ethnographic work. In the course of this discussion I show the commonalities that exist between ethnographic work and machine learning, and I outline methodologies that claim that ethnographic work can make generalizable and accurate statements about the world, including predictive claims. I also point to some of the challenges in using machine learning as a means of producing predictions. This discussion is not meant to discredit these practices, but to demystify the process as a means of loosening quantification's authority, contextualizing its best applications, and putting the two approaches to knowledge production on even footing. Finally, I discuss circumstances in which qualitatively produced predictions may be most valuable, such as when dealing with emerging phenomena and unstable contexts.

INTRODUCTION

As ethnographers in industry, our work is increasingly combined with or compared against the perceived power of big data and data science. Anybody who works in research understands that we live in a world strongly influenced by what Tricia Wang has called the quantification bias (Wang 2016). More so than interpretive work, theoretical concepts, or narrative, numbers have incredible formative power. In our culture, numbers are seen as trustworthy representations of reality (Espeland and Stevens 2008) that are strongly associated with objectivity and untainted by human bias and shortcomings (Daston 1992; Jasanoff 2005). Data science, big data, algorithms, and machine learning not only fit neatly into an epistemological view in which numbers and metrics are seen as taken-for-granted representations of reality (Beer 2016; Espeland and Stevens 2009; Poovey 1998), but they also have fueled a new wave of the quantification bias.

One of the central fascinations of this wave has been the promise that humans now (finally) have the power of prediction at their fingertips. According to the tales told through the public discourse of big data, two key developments have delivered on this promise of science. First, the proliferation of data points provided by the expansion of digital sensors has gifted us the ability to measure and capture the dynamics of a complex world without human interpretation and distortion. Second, the process of machine learning in general, and unsupervised machine learning in particular, has freed knowledge production from human-generated theories and concepts. Together, the narrative goes, these developments have made prediction a reality and revitalized our value of all things quantified.

To be sure, a great deal of this revitalized enthusiasm for numbers is inspired by material changes in our ability to record and create data, our capacity to store and move data, increased processing power, and greater ease of access to the tools to complete these tasks. But the enthusiasm for big data and prediction that stems from the narrative described above has generally outpaced, or at least out-performed, discussions of the epistemological reality of big data predictions among both the public in general and key decision-makers, such as chief marketing officers, policy makers, or even research directors, in particular.

As others have made clear, this new emphasis on data in recent years has provided both an opportunity to reflect on the distinctive value that qualitative and ethnographic work offers in contrast to data science (Wang 2016) and to draw some lessons on what we as qualitative researchers can learn from the practice of data science (Nafus 2016). In this paper, my intention is to add to this conversation by reflecting on what it means to make predictions and to explore the differences in how predictions are accomplished via quantitative modeling and ethnographic observation.

While this is not the first time that ethnographic work has been put in conversation and in contrast with the new quantified practices associated with datafication (van Dijck 2014), most theorists have framed the role of ethnography as providing context to that quantified work. In this paper, I make a slightly different argument, showing that ethnographers produce predictions in our own right. I begin by discussing what it means to predict something, focusing on its function. This is followed by a discussion of the ways in which predictions are constructed through both machine learning and ethnographic work. In the course of this discussion I elaborate on the methodologies that allow us to claim that ethnographic work can generate generalizable, causal, and accurate statements about the world, including predictive claims. I also discuss some of the shortcomings of using machine learning as a means of producing predictions. This discussion is not meant to discredit these practices, but to demystify the process as a means of loosening quantification's authority, contextualizing its best applications, and putting the two approaches to knowledge production on even footing.

WHAT ARE PREDICTIONS?

In order to make claims about the role that ethnographic work plays in generating predictions, we first need to come to terms with what we mean when we use the word "prediction." I discuss both colloquial and technical definitions and then suggest that we utilize a definition that focuses on the function of predictive claims in practice.

Colloquially, we think of a prediction as a claim about an event or a state that will occur in the future. It is this very general, and yet powerful, conception that most us rest upon when using the term. Those more versed in statistical practices, machine learning, or data science may have a nuanced definition in mind: a prediction includes an assessment of the likelihood of such states or events actually manifesting. For example, SAS, a company that is in the business of producing predictive analytics, describes prediction as "the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data" (SAS Institute Inc). This statement of likelihood takes the form of mathematical measurements, such as confidence intervals.

Neither of these definitions serves us well for thinking about the possibility of ethnographic predictions. The colloquial definition says nothing about the origins or production of prediction, while the definition provided by SAS is already infused with the assumptions of statistical sampling methods and inferences. Although references to ethnographic and qualitatively-based predictions can be found in discussions of ethnographic methodology (Burawoy 1998; Small 2009), technical definition for these kinds of predictions are not usually part of these discussions.

However, if we observe the production and application of predictions, we can construct a working understanding. First, the kinds of predictions we talk about with regard to research are linked to empirical data. Second, in applied settings such as marketing agencies, hospitals, consulting firms, or government, predictions are used to support decision-making. At this point, we could say that predictions are empirically-supported claims that are believed to reduce uncertainty about future events or states that are used to buttress decision-making. This definition encompasses a variety of empirical approaches without implying a particular method for generating these claims, and points toward their use in applied settings.

However, careful observation of the problems to which predictions are put and the ways in which they influence decisions allows us to refine this definition further. In my own observations of predictive medical algorithms (Maiers 2017), nurses and doctors used predictive risk scores designed to predict the likelihood of future infection. Despite the theoretical indication of the future onset of infection, these predictions were used to determine if the patient needed antibiotics immediately, suggesting that the prediction factored into the decision-making process by affecting the clinicians' assessment of the patient's current condition.

As I discuss in more detail below, using predictive claims to better recognize current states is a common practice with regard to statistically and machine-learning derived predictions. In order to continue to keep the application of machine-learning predictions within our definition, we must alter it slightly: *predictions are empirically-supported claims believed to reduce uncertainty about current states, future events, or future states that are used to buttress decision-making.* This aspect of the definition is not simply an accommodation for the sake of giving ethnographic work some authority over the realm of predictive research. Rather, it is based on both the function and application of claims that are already called "predictions" in the context of big data and machine learning practices. By this definition, ethnographers and qualitative researchers frequently engage in predictive work. We produce descriptions and claims about the world that help our clients and stakeholders make better decisions. These claims may not always be explicitly future oriented.

HOW PREDICTIONS ARE MADE

To show how ethnographers make predictions in their own right, the remainder of this paper will touch on methodological aspects of both ethnography and machine learning. I focus on machine learning because it is an analytical technique associated with big data and because of its growing application as a quantitative approach to generating predictions. After pointing to some of the commonalities between ethnography and big data and machine learning practices, I outline the process for generating predictions through machine learning. Through this discussion, I hope to arm readers with a better understanding of these practices and the ability to ascertain when and where they work best. I then discuss the scientific status of ethnographic work. Predictions are often seen as belonging to the territory of positive science and seemingly depend upon definitively measured phenomena and the development of covering laws or models. As a result, the interpretive endeavor of ethnographic work may appear to preclude the possibility of prediction. In an effort to show how ethnography can be used in predictive work, I share an alternative framework upon which to base the accuracy of ethnographic claims in general.

Common Ground

When it comes to social data, the processes by which predictions are made in machine learning and big data are similar to the ethnographic process in its basic approach. The value of ethnographic observation is our ability process and synthesize a complex set of data points and relationships. Similarly, the advantage of new data collection practices and the proliferation of sensors is their ability to capture a wide range of data points. On this front, qualitative and quantitative work are increasingly in conversation as data scientists and ethnographers collaborate and utilize new digital tools in the research process (Rattenbury and Nafus 2018; Anderson, Rattenbury, and Nafus 2009).

Although the kinds of data points created through big data and ethnography take different forms, they often describe something similar, namely behavior in context rather than the lab. This is a relatively new application for statistical inference. When it comes to data related to the social world, much of the data that fueled quantitative analysis of social and behavioral data in the past was sourced through surveys. Its application to behavioral data has been expanded thanks to the many sensors and practices that leave "digital sweat," or record of human behavior (Gregg 2015). Whether it is social interactions, unintended uses of technology, purchasing patterns, or twitter traffic, both ethnography and big data work with representations of human behavior in context (Golias 2017; Ladner 2014).

From an analysis of this data, both ethnographers and data scientists extrapolate generalizable claims that help us to better understand phenomena and the relationships between phenomena, thereby reducing our uncertainties about the world. To be sure, the process for extrapolating those claims differs a great deal in process. I now want to walk in a little bit more detail through those processes.

How Machine Learning Makes Predictions

The following section, I describe, in very basic terms, the process by which machine learning produces predictions. This description draws upon my work as a sociologist of knowledge, in which I studied the cultural and epistemological dynamics that both promote and result from quantified practices of knowledge production. I used observations of and conversations with data scientists to examine the cultural assumptions about how legitimate knowledge is produced and the ways in which various methods and claims interact with these cultural assumptions. As part of that work, I frequently asked data scientists to describe the process of machine learning. The resulting description is based, in part, on those conversations.

In the simplest terms, machine learning is a process that allows computers to develop methods for making predictions and inferences. The first step is to provide the computer with a data set. This is often called training data. The learning process may be supervised, in which case the computer is given a data set with labeled or classified phenomena, such as a collection of photos of pets that have been labeled as either "cat" or "dog" and told to develop a method for telling those phenomena apart. Note that this requires the human work of assigning labels to phenomena at some point in the data collection process. Or it may be unsupervised, meaning that the computer defines the categories by which data are described. When it comes to pictures of pets, an unsupervised process could result in categories that humans find meaningful, such as brown pets versus spotted pets, but it might also develop categories that are less salient or even noticeable to humans, such as a mathematical relationship between tail length and ear shape. Once there is an algorithm or model for identifying which images are of cats and which are of dogs, the model will be tested. Often it is tested on a subset of the original data set which has been intentionally set aside for these purposes. If the model fails to successfully predict the known outcomes, the model can be adapted and tested again in an iterative process.

Despite this appeal, there are some limitations in this process worth noting. First, even though the resulting model may be great at predicting which pictures are cats and which are dogs in the original training data set and the test data sets, it may not be very good at making similar predictions on new data sets. In other words, the model may not be very generalizable to new settings and contexts. It is impossible to know in advance how the model will perform in truly novel settings that may have slightly different variables at play. When it comes to social data, this is a particularly difficult problem given that social data are endlessly complex and shaped by both macro structures and local contexts. Furthermore, once these models are applied to novel settings, it can be almost impossible to evaluate their accuracy. The only way to know if predictions are correct is to measure them against the real outcomes, and in many cases that may not be possible.

In order to better clarify this problem, let's consider some cases in which it *is* possible to compare predictions to actual outcomes. The infamous case of the Google Photos algorithm from 2015 that identified and labeled people of color as gorillas is one such instance. This offensive and problematic misrecognition by the algorithm may have stemmed, in part, from a data set trained on photos with an insufficient amount of diversity, bringing attention to issues surrounding bias in data sets and algorithms. It also shows, more generally, how algorithms can fail to be accurate when released from their testing environment on the boarder world. However, we were only aware of the prediction's flaws because we could see and compare the alogrithm's prediction to our own assessment. Similarly, in my own work with predictive medical algorithms, I watched intensive care unit (ICU) clinicians develop a critical assessment of algorithmic predictions (Maiers 2017). Over time, they were confronted with the corporeal reality of their patients in contrast to the algorithm's claims. They saw that the algorithm tended to be successful in some cases and less reliable in others, allowing them to build rules of thumb for contextualizing and sometimes discounting these predictions. However, in many cases, predictions will be used to make crucial decisions long before their accuracy beyond a testing environment can be assessed.

Furthermore, the predictions themselves may be "performative," meaning that the very act of predicting shapes the outcomes that are observed (Callon 1998). This makes it difficult to know what the outcomes would have occurred in absence of such a prediction. Think, for example, of credit scores which are used to assess the likelihood of someone defaulting on a loan. Given that these scores preclude many individuals from taking out a loan in the first place, the algorithm is shaping the very outcomes which it aims to predict, making it ever more difficult to know if the assessments of one's likelihood to default on a loan was accurate in the first place.

The fact that predictive algorithms work best when applied within the same system or domain in which they were trained and tested leads to a second issue. Predictive algorithms are less well-suited for dealing with emerging phenomena, rare events, and unprecedented events. As the definition from SAS reminds us, machine learning predictions are dependent on historical data. This eliminates the possibility of novel events or factors from being included in the model and greatly reduces the chances that the model will sufficiently account for rare occurrences. In addition, depending on the chosen model and method, rare events may be labeled as "outliers" and intentionally eliminated during the data cleaning process. This means that although machine learning may be great at making predictions with stable systems and conditions, it is more likely to mis-predict outcomes in unstable or changing contexts such as social systems or globalizing markets.

Finally, throughout this section of the paper readers may have noticed that the example I used was not about future states at all, but about estimating the likelihood of current states. Though the language of prediction is used to talk about these processes, the actual results are far from our colloquial definition of predictions: they are not about the future. In fact, the same webpage from which I quoted the technical definition of prediction provides many examples of the kinds of predictions machine learning can provide. The first of these predictions is fraud detection. This is not a prediction about the future at all, but an assessment of the likelihood that certain claims *are* fraudulent. In other words, it is reducing uncertainty about a *current* state. This is also the case with the predictive medical algorithms that I mentioned earlier. These predictions are used to identify patients that *are* developing blood infections. In each of these cases, the "prediction" is not about a future state, but about current states. This is not to say that predictive algorithms are not put to uses that are about the future. Models for setting ticket prices, assigning credit scores, or determining how to stock store shelves are future oriented. The point I want to convey is that in their application and function, machine learning predictions are sometimes about the present.

The point of the previous paragraphs has not been to undermine the legitimacy of machine learning claims. Indeed, data scientists have methods in place to mitigate some of the issues discussed here. My hope instead is to demystify how these algorithms work in order to lay the foundation for claiming that ethnography is also a legitimate way to reduce uncertainty about the future.

Foundations for Making Ethnographic Claims

Most of the work that we do in industry is aimed at reducing uncertainty when making decisions (Dourish and Bell 2014). In reviewing her work on the home in several European countries, Genevieve Bell (2001) describes the job of her team as "understanding people and their daily practices with an eye toward finding new users and uses of technology." Not only does this task suggest that her team's work was aimed at reducing uncertainty, their search for the new suggests a future orientation in which they aim to identify which technologies and applications might be developed and successfully adopted by consumers. Given this demand for reducing uncertainty, ethnographers have employed novel qualitative methods designed to better derive insights about potential futures (Dourish and Bell 2014, Forlano 2013, Lindley, Sharma and Potts 2014).

But what is the epistemological framework that allows us to claim that our thick descriptions and qualitative inquiries can reduce uncertainty about consumers and how they will behave and react to products? While some ethnographers might take this capability for granted, a suspicion of qualitative and interpretive work is part and parcel of the quantitative bias in our culture. Depending on the home discipline of the ethnographer, she may not have had to question the validity of her sampling, her analytical methods, or her conclusions based on their epistemological legitimacy before entering into industry. Luckily, this is not the case in my home discipline of sociology, where quantitative sociologists sit on the dissertation committees, editorial boards, and grant committees that review and therefore pass judgement on ethnographic and qualitative work.

In the following paragraphs, I draw on the work of qualitative sociologists who have explored the methodological foundations for claiming that ethnographic work is as justly suited for reducing uncertainty as quantitative research. Though there are many issues I could cover here, I focus primarily on the question that colleagues and clients most frequently ask me when I present my work: how can we be sure that our findings are generalizable? This is also closely related to our ability to establish correlations or causal connections between observed phenomena and consumer or user behaviors. By exploring the epistemological foundations of our work, my hope is to bolster our faith in ethnographic claims and to help put the predictions of machine learning and ethnography on even footing.

Mimicking Statistical Inference or Rejecting Generalizable Claims - The nature and status of ethnographic and qualitative work is one that has been greatly debated within the social sciences (Reed 2011; Pugh 2013; Vaisey 2009). Following Geertz (1973), many of us have stressed the value and legitimacy of interpretation as a type of knowledge production. We have seen first-hand how these thick descriptions illuminate everything from the inner workings of high energy physics (Knorr Cetina 1999) to seemingly paradoxical political phenomena (Hochschild 2016). While this function, in and of itself, is sufficient for work that aims to add to the accumulation of knowledge or to provide better understandings of our fellow humans, it is not a sufficient perspective for those aiming to make broad empirical claims or predictions about entire regions, markets, or segments of consumers or users. This is the case for two reasons. First, interpretation often fails to hold authority in a world saturated in quantification bias and the epistemological mental models that accompany this bias. This is due, in part, to the common complaint that qualitative work fails to meet the standards of statistical representativeness and inference. Second, the ways in which qualitative and interpretative work have become associated with methodological programs that emphasize local particularities over generalized patterns or the creation of shared understanding over relational claims makes it difficult to extend the claims of qualitative work beyond its immediate cases and contexts. Both of these can be overcome through an exploration of the methodological approach to qualitative work.

There have been a variety of attempts to remedy this situation. First, we might try to solve this problem by mimicking the assumptions of positivist quantitative work. The idea is that in choosing the right case we can mimic the assumptions of statistical representativeness upon which many quantitative claims are based. We look for cases and locations that best represent a broader population or we do comparative ethnography as a way to isolate and identify causal relationships and to "control" for confounding variables. This is inevitably problematic. As Small (2009) makes clear, this process often mistakes the concept of representativeness with that of averages. For example, a study of social engagement in a mid-sized town in the mid-west which matches the national average of income or education levels cannot somehow represent this process across most American communities, even though it may statistically resemble the nation as a whole. Furthermore, by intentionally excluding rare or unique cases, we miss out on the opportunity to learn about emergent and developing phenomena or to observe the effects of interactions that may be difficult to observe in average cases.

As an alternative, many of us have been trained in a perspective of interpretive empiricism (as discussed in Reed 2012), in which social knowledge is created through inductive processes that stress the locality of social investigation. Under this epistemological framework, we are not trying to make generalizable, objective claims about covering laws or models at all, but to best explain the social world by articulating the dynamics of the particular. This focus on locality, alongside a resistance to theory and macro social constructs, makes it all but impossible to generalize the findings from one case out to a broader population. It is particularly a problem for industry work in which we hope to use in-depth qualitative analysis of a small sample to inform decisions about entire markets or populations. Is there a way forward in which interpretive and qualitative analysis can make generalized inferential claims without trying to wedge itself into the assumptions of statistical inference?

Finding Our Own Footing - First, we should recognize that statistical inference, and its reliance on large, representative samples, is not the only way to generalize claims. In examining the extended case method (Burawoy 1998), Mario Small (2009) suggests that ethnographers use logical inference instead. This means that the inferences refer to situations rather than populations. As Small explains, in a statistical inference, we hypothesize that populations with a given set of characteristics will display the same set of corresponding characteristics or properties observed in a sample. An example is to say that active adults in the Charlottesville, Virginia are more likely to purchase a gym membership if their household income is over $60,000. These kinds of claims require some sort of instrument for establishing representativeness and therefore the accuracy of claims. With logical inference, the focus shifts to processes and mechanisms of a situation. We might hypothesize that when offered a free trial at a new gym, the consumer's decision to book or ignore the offer depends partially on perceptions of cultural fit between the gym and the potential customer. This statement is based on our ability as ethnographers to observe a chronology of events and therefore make causal links between behaviors, decisions, feelings, and events.

This kind of logical inference is particularly good for making what Small calls "ontological statements" or "the discovery of something previously unknown to exist" (2009: 24). In the example I have given, logical inference allows me to make a claim about the relationship between cultural fit and gym membership purchases. This is a great advantage of ethnographic work. Big data does not capture what it does not measure. In addition to the challenges presented in measuring emotions and things like "perception of cultural fit," a quantitative study could not take such phenomena into account without someone determining it was a variable worth measuring. All phenomena must be known, at least in the form of measurable data points, ahead of the machine learning process.

Another option offered by Small is to take a different approach to sampling. Rather than look for a representative case, we use "case study logic" to sample. With case study logic, instead of relying on the representativeness of a sample, each additional iteration of investigation brings the researcher closer to an accurate understanding of the area under investigation. As such, this is a sequential process that ends only when the researcher is able to accurately predict the dynamics of the next case and no new phenomena or relationships have emerged. Rather than validating a claim or relationship by statistically showing that our sample would be highly unlikely to contain such a correlation or causal relationship when there is not one in the population, the hypothesis is validated through continual testing that challenges and refines the claim. Interestingly, the iterative nature of case study logic as a means of validation is somewhat similar to the process used in machine learning. Where ethnographers return again and again to the field to test and refine hypotheses, machine learning processes also refine models and algorithms through iterative testing with test data sets. Though the processes may look quite different, it is through repeated exposure to data that both data scientists and ethnographers gain confidence in their conclusions.

So far, I have discussed the ability of ethnography to make accurate and generalizable claims that reach beyond the immediate location of our observations. In the course of this discussion, I have also suggested that we can identify causal connections between phenomena through observation. These are important pieces in understanding why ethnographic work can make predictions. Our work is predictive insofar as it is used to reduce the uncertainty about future states and events, such as changes in markets or the reactions and decisions of users.

As I indicated at the start of this section, ethnographers in industry regularly engage in work that serves the function of prediction. We also use analytical and sampling methods that are similar to those offered by Small. In the next portion of the paper, I talk through an example, pointing out ways in which we might frame the epistemological legitimacy of our predictive work to stakeholders along the way.

APPLICATIONS OF AND ARGUMENTS FOR ETHNOGRAPHIC PREDICTIONS

Ethnographic work is particularly apt at reducing uncertainties for several reasons. As others have pointed out (boyd and Crawford 2012; Seaver 2015; Wang 2016), ethnographic and qualitative work captures different information than that captured by quantitative data sets. Ethnography and qualitative work is particularly advantageous for dealing with the connection between meaning and behavior (Reed 2012), for unearthing meta-feelings and cultural schema (Pugh 2013), and for illuminating subjective experiences (Ladner 2012). These are aspects of the human experience that are difficult to capture and surface in quantitative data sets.

In addition to these well-known advantages, ethnography is well-suited to predicting the emergence and implications of new phenomena. A widely cited example of this is Tricia Wang's (2016b) ethnographic work on technology usage in China where she observed that low-income individuals were eager to gain access to the technology afforded by smart phones. Framed slightly differently, I would suggest that her research allowed her to *predict* that low-income consumers in China would purchase affordable smartphones. As such, she advised Nokia to move their business model in that direction. As she describes, her insights fell on deaf ears; Nokia was not convinced by her argument. Their orientation toward quantified metrics and the epistemological models that accompany quantification left them unreceptive to ethnographic data. In particular, she tells us that they resisted her findings because they were not from a large enough sample to suggest representativeness and reliability and that they could not corroborate her findings with their large quantified data sets.

As Wang rightly puts it, "what is measurable isn't the same as what is valuable." This is one of the values of ethnographic work. We are able to observe and capture phenomena for which there are no quantified metrics and for which there may not be a name or label. We are also able to predict outcomes that are related to and intertwined with emerging phenomena. Recall that this is one of the challenges of relying upon machine learning for predictive claims; Predictive algorithms cannot account for forces about which they do not know. This means that we do more than provide context for quantified information. Ethnographers can quickly make adaptations to what they are observing and actively generate and test hypotheses about these dynamics during their work to accommodate rare and unprecedented events. In other words, ethnography is well suited to reducing uncertainties in systems and contexts that are changing and unstable. There is an enormous amount of business value in capturing the emerging needs of smartphone users in the changing digital environment of the gig economy, for example.

In addition, ethnographers employ alternative methods for hypothesis generation and validation, thereby drawing insights from contexts and areas of inquiry where large samples and data sets may not be possible or practical. This means that the best qualitative studies do not simply mimic statistical inference, but are conducted according to validation techniques appropriate for qualitative data. With regard to objections to the small sample from Wang's Nokia study, we might point out that she was not using a statistical sampling logic and that the accuracy and inferential power of her claims are not dependent upon sample size. Instead, she likely used something closer to case study logic to sample, noting that her sample was saturated by consistent and predictable patterns. Instead of identifying relationships between phenomena through statistical inference, Wang may have used logical inference to develop and then verify her hypothesis. She may have seen that in the context of a changing technological and economic landscape, individuals were shifting their financial priorities. To be clear, I do not know the details of how Wang came about her sampling process or her choice of method for conducting analysis. However, many ethnographers and qualitative researchers work with processes that closely resemble the methodologies of case study logic and logical inference.

CONCLUSION

The discussion presented in this paper is meant to give researchers a foundation for arguing that ethnographic and qualitative work are as equally capable of reducing uncertainty surrounding future-oriented decisions as quantitative methods. In doing so, I have not intended to delegitimize research that relies upon machine learning or big data, but to provide a very basic understanding of these practices as a way to demystify them and show qualitative researchers where there is room and need for qualitative observations. As we work to collaborate with data scientists and computer scientists, qualitative researchers would be well served by deepening their understanding of these processes beyond what I have described here in order to make such partnerships more fruitful.

In addition, I have suggested that there are instances in which ethnography may be particularly well-suited to predictive work. Instances in which we do not yet know what categories and phenomena will be relevant may be less well-served by quantitative work which depends upon known categories and mechanisms being in place to measure such phenomena. Similarly, ethnographic work has a particular advantage in unstable systems and contexts where new phenomena and patterns may be emerging.

Finally, I have shared an alternative framework on which to base the epistemological authority of ethnographic claims. Our stakeholders often evaluate and question our work based on the assumptions and processes of statistical inference. Though these criteria do not apply to ethnographic work, we can base our claims and ability to generalize in logical inference and case study logic. By giving these processes a name and educating our stakeholders in their assumptions and applications, we loosen the hold of the quantitative work and be free to employ the best method or mix of methods for the problem at hand.

Whether or not we should also use the term "prediction" to describe our research that is aimed at reducing uncertainty and supporting decision-making remains an open question. However, my hope is that the discussion provided here has made it clear that we have as much of a case for claiming that we do predictive work as those who use statistical inference or machine learning to produce knowledge.

NOTES

1. I want to emphasize that the definition of prediction that I construct here is not meant to be normative or to precisely represent formal definitions of prediction that might be found in mathematical or philosophical treatise. Entire books could and have been written on the topic. Instead, as an observer of knowledge practices, this definition is based on what people actually do when they make a prediction, the expectations they have of this construct, and the uses to which it is put.

2. Small takes the idea of case study logic from Yin (2002) and adapts it for interview-based research.

REFERENCES CITED

Anderson, Ken, Dawn Nafus, Tye Rattenbury, and Ryan Aipperspach
2009 "Numbers Have Qualities Too: Experiences with Ethno-Mining." *Epic*: 123-140.

boyd, danah and Kate Crawford
2012 "Critical Questions for Big Data." *Information, Communication & Society* 15(5): 662–79.

Burawoy, Michael
1998 "The Extended Case Method." *Sociological Theory* 16 (1): 4-33.

Callon, Michel, editor
1998 *The Laws of the Markets*. London: Blackwell Publishers.

Daston, Lorraine
1992 "Objectivity and the Escape from Perspective." *Social Studies of Science* 22(4): 597–618.

Dourish, Paul and Genevieve Bell
2014 "Resistance is Futile": Reading Science FictionAlongside Ubiquitous Computing." *Personal and Ubiquitous Computing* 18 (4): 769–778.

Espeland, Wendy Nelson and Mitchell L. Stevens
2008 "A Sociology of Quantification." *European Journal of Sociology* 49 (3): 401–36.

Forlano, Laura
2013 "Ethnographies from the Future: What can ethnographers learn from science fiction and speculative design?" *Ethnography Matters*. Accessed July 30, 2018. http://ethnographymatters.net/blog/2013/09/26/ethnographies-from-the-future-what-can-ethnographers-learn-from-science-fiction-and-speculative-design/

Geertz, Clifford
1973 *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Gregg, Melissa
2015 "Inside the Data Spectacle." *Television & New Media* 16 (1): 37–51.

Golias, Christopher A.
2017 "The Ethnographer's Spyglass: Insights and Distortions from Remote Usability Testing." *Ethnographic Praxis in Industry Conference Proceedings*, 247-261.

Hochschild, Arlie Russell
2016 *Strangers in Their Own Land: Anger and Mourning on the American Right*. New York: The New Press.

Jasanoff, Sheila
2005 *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton, NJ: Princeton University Press.

Knorr Cetina, Karin
1999 *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge and London: Harvard University Press.

Ladner, Sam
2012 "Ethnographic Temporality: Using time-based data in product renewal." *EPIC 2012 Proceedings*, 30-38. Accessed June 21, 2018. https://anthrosource.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-8918.2012.00005.x

2014 *Practical Ethnography: A Guide to Doing Ethnography in the Private Sector.* Walnut Cree: Left Coast Press.

Lindley, Joseph, Dhruv Sharma, and Robert Potts
2014 "Anticipatory Ethnography: Design Fiction as an Input to Design Ethnography." *Ethnographic Praxis in Industry Conference Proceedings,* 237-253.

Maiers, Claire
2017 "Analytics in Action: Users and Predictive Data in the Neonatal Intensive Care Unit." *Information, Communication & Society* 20: (6): 915-929.

Nafus, Dawn
2016 "The Domestication of Data: Why Embracing Digital Data Means Embracing Bigger Questions." *Ethnographic Praxis in Industry Conference Proceedings*, 384–399.

Nafus, Dawn and Jamie Sherman
2014 "Big Data, Big Questions| This One Does Not Go Up To 11: The Quantified Self Movement as an Alternative Big Data Practice." *International Journal of Communication* 8(0):1784–94.

Poovey, Mary
1998 *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. 1st edition. Chicago: University Of Chicago Press.

Pugh, Allison J.
2013 "What Good Are Interviews for Thinking about Culture? Demystifying Interpretive Analysis." *American Journal of Cultural Sociology* 1(1):42–68.

Rattenbury, Tye and Dawn Nafus
2018 "Data Science and Ethnography: What's Our Common Ground, and Why Does It Matter?" May 7, 2018. Accessed June 1, 2018. https://www.epicpeople.org/data-science-and-ethnography/

Reed, Isaac Arial
2011 *Interpretation and Social Knowledge: On the Use of Theory in the Human Sciences*. Chicago; London: University of Chicago Press.
2012 "Cultural Sociology as a Research Program: Post-positivism, Meaning, and Causality." *The Oxford Handbook of Cultural Sociology*. Jeffrey C. Alexander, Ronald N. Jacobs, and Philip Smith (eds). Oxford and New York: Oxford University Press.

SAS Institute Inc.
"Predictive Analytics: What it is and Why it Matters." Sas.com. Accessed June 4th, 2018. https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Seaver, Nick
2015 "The Nice Thing about Context Is That Everyone Has It." *Media, Culture & Society* 37(7): 1101–9.

Small, Mario
2009 "'How Many Cases do I Need?': On Science and the Logic of Case Selection in Field-based Research." *Ethnography* 10 (5): 5-37.

van Dijck, J.
2014 "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." *Surveillance & Society* 12(2): 197-208.

Vaisey, Stephen
2009 "Motivation and Justification: A Dual-Process Model of Culture in Action." *American Journal of Sociology* 114(6): 1675–1715.

Wang, Tricia
2016 "The Human Insights Missing from Big Data." TEDxCambridge. Accessed June 3rd, 2018. https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data

Wang, Tricia
2016b "Why Big Data Needs Thick Data." *Medium.* Accessed June 2nd, 2018 https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7

Yin, R.
2002 *Case Study Research*. Thousand Oaks, CA: Sage.

The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care

MADELEINE CLARE ELISH
Data & Society Research Institute

The wide-spread deployment of machine learning tools within healthcare is on the horizon. However, the hype around "AI" tends to divert attention toward the spectacular, and away from the more mundane and ground-level aspects of new technologies that shape technological adoption and integration. This paper examines the development of a machine learning-driven sepsis risk detection tool in a hospital Emergency Department in order to interrogate the contingent and deeply contextual ways in which AI technologies are likely be adopted in healthcare. In particular, the paper bring into focus the epistemological implications of introducing a machine learning-driven tool into a clinical setting by analyzing shifting categories of trust, evidence, and authority. The paper further explores the conditions of certainty in the disciplinary contexts of data science and ethnography, and offers a potential reframing of the work of doing data science and machine learning as "computational ethnography" in order to surface potential pathways for developing effective, human-centered AI.

INTRODUCTION

"The problem with sepsis is that we just don't know," says an ER physician who is explaining the process of treating sepsis in her Emergency Department, "Everything is risk versus benefit… we're hoping that Sepsis Watch will help us get to the right point faster." Sepsis Watch is a machine learning-driven system that assesses a patient's risk of developing sepsis, an extremely deadly syndrome that involves a body's over-response to infection and likely results in organ damage or death if untreated. The system, or "the tool" as the clinicians and computer scientists describe it, is one of the first machine learning models for detecting sepsis to be deployed in an Emergency Department (ED).[1] After years of development, Sepsis Watch is in the process of being integrated in the context of emergency care at Duke University Hospital, a large, urban research hospital.

In this paper, I bring an anthropological perspective to bear on the development of Sepsis Watch. My focus is on articulating the epistemological and social entanglements that characterize the emergence of this technology, and analyzing the implications of these entanglements for conceptions of trust, evidence, and authority in clinical care. I begin by providing an overview of the project background, and then move into a discussion of the technologies at stake. I then describe and discuss a set of tensions that emerged from research during the development and planning stages of the tool. In the final section, I draw together the themes of trust, evidence, and authority with the disciplinary groundings of machine learning and ethnography, and propose the idea that machine learning may be productively understood as "computational ethnography." My aim is to explore a set of tensions that arise around certainty, explainability, process, and method, and in turn to offer a potential reframing of the work of doing data science and machine learning in order to surface potential pathways for developing effective, human-centered AI.

This research is based on an ongoing collaboration with a multi-disciplinary team at Duke Health and Duke Institute for Health Innovation (DIHI) working together to design, develop, and implement the tool. The observations described in this paper are based on approximately 20 hours of interviews with technologists, clinicians, and administrators, and approximately 15 hours of observation in the Emergency Department and Cardiac Intensive Care Unit at Duke. In writing this paper I also draw on over three years of participant observation of machine learning technologies in a range of sectors as a researcher at the Data & Society Research Institute.

BACKGROUND

The problem: Sepsis

Sepsis is a widespread and grave problem in healthcare. Some research concludes that over 3.1 million hospitalizations annually are due to severe sepsis, and the U.S. Center for Disease Control (CDC) reports that 1.5 million people develop sepsis annually, and about 250,000 people die from sepsis within the United States each year (CDC 2018). The CDC reports that one out of every three patients who die in a hospital have sepsis.

Sepsis is a syndrome characterized by a body's extreme response to an infection, and without treatment leads to tissue damage, organ failure, and death. Although the condition is more likely to develop in populations with reduced immune responses, sepsis may affect anyone. Not only is the condition life-threatening, it also develops rapidly, often in a matter of hours. Early detection and treatment are critical for survival (Levy et al. 2010). However, there does not exist one test to confirm the onset of sepsis.

Clinicians attempt to diagnose sepsis by detecting an underlying infection that can be tested through blood samples in a lab. However, these tests are not always reliable and may take too long; for instance, blood culture results are updated at 24 and 48 hour intervals. Although the timely diagnosis of sepsis is extremely challenging, once it has been diagnosed, treatment is usually straight-forward and based on existing protocols of antibiotics, known as bundles, that have proven effective in treating sepsis.[2]

Sepsis is a particularly relevant disease to discuss in the context of categories of evidence and certainty. Sepsis is extremely difficult to diagnose in large part because its causes and progression are very poorly understood from a biomedical perspective. As one emergency doctor explained, "Sepsis is very difficult. … how do you know when someone has it? 'I just know' isn't good enough as a diagnostic tool — it's not like an x-ray, where we can see, 'oh that's broken.'" Diagnosing sepsis, in practice, is often ultimately left to "gut instinct" as another clinician told me.

Taken together, these characteristics make sepsis an ideal case for machine learning diagnostics, "prime for tech" as one clinician told me. Sepsis is widespread, with profound severity for people's lives—and also for the allocation of hospital and insurance resources. One emergency doctor explained how "sepsis is always on the front burner." First and foremost, he explained, there's "the moral and ethical imperative" to save people's lives, and at the same time, "there's the brick and mortar bottom line." A technological tool that could facilitate diagnosis is very needed and would be very welcome across the board.

The technology: Machine learning and AI

Most of my informants and collaborators use the term "machine learning" or "deep learning" when referring to Sepsis Watch and their field of research or work. Several doctors I spoke with emphasized that it was best to avoid terms like AI or machine learning when talking to doctors, favoring terms like "predictive analytics." Nonetheless, these terms exist within a constellation of technologies implicated in the umbrella term of AI, a term that is both over-hyped and unmistakably compelling. As the latest technology buzzword to enter common parlance, part of AI's rhetorical power is in its slipperiness, wherein everyone has a notion of what AI is—but everyone's notion is different (Elish and Hwang 2015).

Even as AI is a nebulous term—more marketing than technical—machine learning does refer to a specific set of computer science and statistical techniques that refer to a type of computer program or algorithm that enables a computer to "learn" from a provided dataset and make appropriate predictions based on that data. Computer scientist Tom Mitchell's definition makes clear what "learning" means in this context: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997: 6). Learning, in this case, is narrowly defined and refers essentially to the capacity for an algorithm to recognize a defined characteristic in a dataset in relation to a defined goal and to improve the capacity to recognize this characteristic by repeated exposure to the dataset.

For the purposes of this paper, it is relevant to call attention to precisely what is "intelligent" in machine learning and current formulations of AI. Until relatively recently, the intelligence at stake in "AI" predominantly referred to procedural, logic-based reasoning and the capacity to manipulate abstract symbolic representations – now also known as "Good, old fashioned AI" (Hoagland 1995). These systems formed the commercial technologies within the "expert systems" of 1970s and 1980s, which were eventually critiqued as "brittle" and too limited (Dreyfus 1972; Forsythe 1993, 2002; Suchman 2007). This stands in contrast to the "intelligence" at stake in current conceptions of AI, which are rooted in machine learning techniques, in which logic and abstract representational meaning are beside the point; intelligence is derived through detecting patterns across vast amounts of data and predicting outcomes based on probabilistic statistics. In other words, the "smartness" of AI comes from a system's ability to process and analyze huge amounts of data, beyond the scale of any individual human, in order to predict or automate certain activities. The datasets and models used in these systems are not objective representations of reality. A machine learning algorithm can be said to "know" something only in the sense that it can correlate certain relevant variables accurately. These different paradigms of intelligence within AI research have deep implications for the construction of knowledge, truth, and fact as epistemic categories and the ways in which these categories can be leveraged in social practice.

DEVELOPMENT AND DEPLOYMENT

The team: Duke Health and DIHI

The interdisciplinary team working on Sepsis Watch is led by both physicians and computer scientists and is split across Duke Medicine and the Duke Institute for Health Innovation

(DIHI). DIHI is a multidisciplinary institute that draws together both Duke University and Duke Medicine. According to their website, DIHI "promotes innovation in health and health care through high-impact innovation pilots, leadership development, and cultivation of a community of entrepreneurship" (DIHI 2018). With a staff of approximately fifteen, DIHI functions as a kind of collaboratory, leading or facilitating pilot projects across Duke, involving clinicians and students at Duke Health as well as faculty, staff, and students at Duke University. While each of the existing twenty projects is unique, all are "grassroots projects" in the sense that they were first proposed by external collaborators and then developed in partnership with the team at DIHI.

In the previous section, I described the problem that Sepsis Watch has been developed to address, and then outlined the relevant technological context. This order reflects the ways in which the Sepsis Watch team working on this project approaches their work. It is notable that this is the opposite of dominant norms around developing new AI interventions. In the current climate of AI hype, it is common for new companies or projects to be technology-driven, as opposed to problem- or community-driven. As ethnographers, we know the perils of this approach. The DIHI team is also wary of tech-solutionism, and prioritizing clinical problems and institutional goals is structured into their project conception and development process. For instance, potential projects need to start from a real problem with a specific goal – not an idea for a model that can predict something and see if it will work in real life.

The tool: Sepsis Watch

The origins of Sepsis Watch date back several years to a Duke hospital initiative aimed at improving patient outcomes and decreasing costs of care. However, the project in this iteration began in 2016 when two physicians wrote a small proposal to work with DIHI on the project. Once initiated, the first twelve months of the project involved obtaining and cleaning data from a database of Electronic Health Records (EHRs).

The road to deployment was rockier and slower than anticipated. While the team worked closely with The Epic Systems Corporation, simply referred to as Epic, the leading provider of EHRs to US hospitals, to extract data and explore potential development paths, the model was not able to be deployed within Epic software for the duration of the pilot. Another set of unexpected problems and delays emerged around integration of the tool within existing IT infrastructures. In this context, it is interesting to contrast the Silicon Valley ideal of disruption with the reality of large, legacy healthcare systems, which I discuss in further detail below.

The tool itself leverages real-time data drawn from EHRs, a deep learning model, and a graphical user interface (GUI) in order to predict the risk of a patient developing sepsis before it occurs. The training dataset for the model consisted of 51,697 inpatient admissions at Duke Hospital spanning the course of 18 months stored in Duke's Epic HER (Futoma et al. 2017: 6). This patient data was considered to be representative of the hospital's patient population and clinical settings, and 80% of the data was used for training the model, with the remaining 20% held out for multiple stages of verification. The output of the model produces a score and the score determines whether a patient is "no risk," "low risk," "high risk," or "septic" based on streaming data including lab results, vitals, and medications (Futoma et al. 2017). The GUI allows a clinician to keep track of several patient scores at a time, and when desired, pull up a specific patient's score over time, as well as preceding treatment, and what and when the last lab, vital, and medication were taken. The clinician, after reviewing the scores, can send the patient to "a watch list," and monitor the watch list, which keeps track of all the treatments (completed or uncompleted), chart reviews, and clinical encounters. The application will be accessed through a tablet or other handheld device, a design requirement from the beginning of the project that derived from the need for nurses to remain mobile and able to move around the hospital

Acknowledging limits: "Where can we have a real effect?"

Healthcare, and in particular hospitals, have historically been slow to adopt new technologies was emphasized to me again and again during interviews, and is often a talking point in healthcare industry conferences and op-eds. This is a point of frustration for many, who see the use of big data and machine learning in healthcare as providing immense opportunity to improve patient outcomes. Both industry and academic researchers are focused on developing new models or tools. Using big data in healthcare, for instance in problems like identifying sepsis may seem like a low-hanging fruit, but as one informant put it: "It's a low hanging fruit, but the fruit has a thick stem. You can't really hit it."

An important part of the development process for the team has been learning to work in different timeframes, not just for developing a data pipeline or building a model, but also testing and implementing the tool in the emergency department; most previous projects had been one year long and discreet pilot projects, but this project was intended to operate a different scale of complexity and deployment. The team, perhaps because they are situated within a research institute and teaching hospital, is aware and open to the potential that the tool may not work or be accepted by clinicians. The incentives for the team are aligned not around profit or technical success, per se, but rather, around patient outcomes and basic research. In order to give the tool the best chance of formal success, the team carefully articulated what they thought was a feasible goal for the project: to manage the care of sepsis better – not necessarily to decrease mortality. In addition, when planning for the tool's integration into existing workflows and adoption, the team was careful to explicitly bound the capabilities and role of the tool. In reading a previous version of this paper, one of the leads on the project commented that

> Our experience testing the model during the silent period…has reinforced this. All that we know now is that the model works at predicting the first episode of sepsis in patients admitted to Duke University Hospital, but before transfer to an Intensive Care Unit. Going to a different hospital or expanding into different inpatient settings will all require additional work to validate that the tool continues to perform as well as we expect it.

The choice of the word "tool" to describe the technology was unintentional, but nevertheless underscores the idea of *augmenting* existing clinical practice – not replacing.

SOCIOTECHNICAL ENACTMENTS

In this section, I discuss three salient dimensions of developing Sepsis Watch as they emerged from interviews and ethnographic research and that bring into focus the epistemological implications of introducing a machine learning-driven risk assessment tool into a clinical setting: trust, evidence, and authority. While on the surface these categories may seem relatively self-evident, it is precisely in the articulation of just what these concepts mean that the cultural work and social infrastructures of deploying machine learning systems emerge. I follow Anne-Marie Mol's articulation of "enactments" in clinical settings in order to draw attention to "the techniques that make things visible, audible, tangible, or knowable" (2002: 33). Things, in this case, can be diseases or bodies or any manner of object, and enactments allow us to talk about the ways that realities are multiple but rooted in praxis. That realities are enacted – not *constructed* or *performed* in the senses in which these terms have become overloaded in STS (Science and Technology Studies) literature -- "enacted" draws attention to the particular reconfigurations of the world that are not given or even self-evident but rather emerge as significant at particular times and places among specific actors. In this context, I discuss the ways in which trust, evidence, and authority are talked about and enacted alongside and through the machine learning tool. Far from simple categories, the analysis demonstrates that the tool, as a machine learning technology, requires careful negotiations and interpersonal and institutional reconfigurations even before it can be fully deployed.

Trust: "The answer to trust is not a technical solution"

The nature and development of "trust" was a recurring theme in my conversations with all members of the team. All the team leaders knew that establishing "trust" was an essential foundation upon which everything else would rest. Only if a technology is trusted will it be used. They judged that the successful deployment of Sepsis Watch would rely not only within the clinical interactions in the emergency department, but also in the various formal and informal social networks that intersect, shape, and use or are responsible for the use of the tool. These networks would set the stage for whether people would trust and accept the technology. While the lead physicians were attuned to the importance this "change management," the machine learning researchers perhaps underestimated the extent to which actually deploying the tool would all too often feel like a Sisyphean task.

How does one build a technology that can be trusted? A primary strategy of the team was to "loop in stakeholders" from the very beginning of developing the tool. This included conversations and meetings with not only with hospital leadership, but also physicians, nurses, and other front-line workers who would be using the tool, as well as all the departments that would eventually need to be involved, including the hospital IT department—"the guys who keep the lights on." For each of these stakeholders, it was important to tread carefully and not be or *be seen as* telling other people how to do their work. At the same time, emphasizing the ways in which the tool fit into existing standardized guidelines for care and also could potentially improve patient outcomes was a generative form of engagement.

This strategy was of course informed by previous work and established best practices. As one physician put it:

> When you're in charge of a life, I think this level of distrust is an important one to have until it's validated. … Any time you are adopting new technology which is not validated, I think there is some amount of trust building that has to go along with the project and that comes from working with an engagement right from the beginning.

But in the lead up to the testing, no one knew how to ensure that trust would be established and this caused anxiety.

Another significant concern revolved around the risk of "alarm fatigue." A previously implemented early warning system (Bedoya at al. 2018) to identify patients at risk of cardiac arrest, unplanned ICU admission, and death based on the National Early Warning Score (NEWS) (Smith et al 2013) within the Duke University Health System resulted in 63.4% of alerts triggered to be dismissed by the care nurse who was notified (Futoma et al. 2017). This previous system was a much less sophisticated and precise system than Sepsis Watch. Still, concerns about "alarm fatigue" were a common theme in my interviews with clinicians. Alerts, like pop-up windows on a personal computer, are often experienced as more annoying than helpful, like the "update software and restart" alerts that seems to show up constantly in the corner of a screen on Windows and Mac operating systems. The tricky but necessary balance to achieve, a doctor explained, speaking about alert systems generally, is that "at some point, someone has to write a rule to reach a threshold [to trigger the alert]. If it's too low, it'll get ignored cause it'll alert too often. And if it's too high, you're risking harm to a patient." Establishing trust around a tool is not only about building inter-personal relationships but also about aligning desired behaviors to existing norms and expectations.

Evidence: "Our machine learning is easy to call a black box—but the human body is a black box!"

Key to having meaningful interactions with stakeholders was demonstrating evidence that the tool would be effective. What could constitute evidence or even efficacy was simultaneously central and variable. What kind of evidence matters? Where does it come from and who can interpret it? An important insight of the team leading up to testing was that various types of evidence are salient for different stakeholders: Hospital administrators and managers are convinced by numbers and statistical trends. Front-line clinicians and middle managers are more convinced through anecdotal evidence and discussions of specific cases and patient outcomes. This is not just because it's a compelling story, but because through relating the specifics clinicians have the possibility to identify with a mistake or an oversight they themselves might have made along the way. Telling these stories and going over cases are built into recurring department meetings and are part of student training.

In addition to the *types* of evidence that were salient, the team also had discussions about the *extent* of evidence necessary. Doctors, for instance, people explained to me, are trained to look beneath the surface and understand cause and effect; how much will doctors need or want to look inside the black box of a machine learning tool?

While it is tempting to speak generally about one category of ER physician, for instance, the distinct nature of different types of hospitals makes this problematic. During a tour I was given of the Emergency Department, my guide told me laughing, "If you've seen one academic hospital, you've seen one academic hospital." In the community hospitals that are part of the Duke network but are less prestigious, sometimes with more limited resources, and also with different communities for whom they provide care, physicians and nurses were thought to have different requirements. Existing patterns of knowledge diffusion mean that "the Duke brand" plays a large role in how much people may trust something coming from the research hospital. Duke's reputation and history profoundly influence how people will perceive a technology coming out of a Duke research center.

Still, the working understanding was that many physicians will only trust a machine learning model if they have proof that it works. The researchers felt there was an important distinction between *proving* that a model works and proving *how* it works. They felt, in fact, that because the model was uninterpretable they had to be "even more rigorous" in how the model was tested. The team published technical papers and spent significant time trying to demonstrate that their model performed "better than the status quo" within and beyond the hospital. The technical lead on the project stated, "If our model didn't perform better than every comparable model on our held out sets, there's no amount of trust we could have tried to build via relationships." Both demonstrations of efficacy and enforcing social networks were deemed necessary to establish evidence, a classic theme in the history of science and technology (Shapin and Schaffer 1985).

Interestingly, this troubles a growing emphasis on explainability and interpretability in technical and social science research communities.[3] What does it mean to look into the black box, if everyone has different conceptual lenses through which to see what's inside? Making a model "technically interpretable," the machine learning researchers emphasized, does not equate to make the technology interpretable or trusted by doctors. As one researcher put it: "I think the issue is that interpretability is about understanding the *causation*. That's the key thing that people push for, but instead they would say, 'I want to be able to interpret the model.'" The Sepsis Watch model is "totally uninterpretable" but their development process focused on the trust that can be built from technical demonstrations of efficacy embedded within existing social relationships.

Sepsis is a particularly interesting syndrome in which to think about tradeoffs around interpretability. As described earlier, the causes of sepsis are poorly understood. Other ethnographic work by Maiers (2017) about automated decision aids for detecting sepsis in infants describes the diagnosis of sepsis as being built on "a gut feeling." Moreover, treatment is not "path dependent;" that is, once a patient has sepsis the treatment is not dependent on how sepsis developed. However, this is not the case for all conditions. For example, treatment for cardiogenic shock, a condition in which the heart cannot pump enough blood, is dependent upon what caused the shock. During one conversation a researcher exclaimed, "Our machine learning is easy to call a black box--but the human body is a black box!" Sepsis is like a black box, inside another black box.

Other physicians voiced the opinion that if it seemed to be working by an agreed upon metric, in the case of Sepsis Watch, improving care for septic patients, the inner workings of the model didn't matter. Following one conversation, a researcher directed me to one of his favorite TED talks by Ziad Obermeyer (2017) on the subject, answered the question, "If a machine could predict your death, should it?" with a resounding yes. In this talk, Obermeyer emphasized that the implications for living life well and dying a good death—and the limitations and failures of existing healthcare—were too profound to be discounted when judging the risks and benefits of using a technology without understanding how precisely it works.

Authority: Who's responsible? Who's accountable? Who needs to be communicated with, and who needs to be informed?

As voiced above, the distrust that clinicians have of new technologies are well founded and reasonable. The life of a patient is in a clinicians' hands. Trust has been placed in them and their judgement, and it is their responsibility and professional duty to ensure the patient receives the best care possible.

Lines of authority became relevant on multiple levels during the development process. On one level, the team needed to understand who held ultimate authority to allow the implementation of the tool in the first place. In the context of a vast and slow-moving industry, with intricate policies around data privacy and security, the sources of authority were multiple and not always easy to see.

On another level, the everyday practices of clinicians might be variously enhanced, threatened, or destabilized in the face of a new machine learning tool. One physician emphasized the fear that many doctors have that machine learning and AI will threaten their "autonomy." She explained,

> A lot of this predictive models have the perception of taking away some decisional timing, like you're doing algorithmic or robot-based medicine. Machine learning itself, the term, you're not supposed to use it as much when talking to physicians because it's got a negative connotation.

She said the preferred term right now when talking to clinicians is "predictive analytics." Physicians were also worried about how the jobs of nurses would change. One concern was that nurses would need to learn to interpret and work with the tool in ways that they had not been trained for. Previous studies have shown mixed results (Guidi 2015) when implementing diagnostic aides or automated decision tools in clinical settings. Maiers (2017) argues that nurses in a Neonatal Intensive Care Unit (NICU) incorporate a predictive analytics tool into sepsis diagnosis and care through interpretive processes that combine multiple forms of evidence, including experiential and embodied knowledge. And while in some cases, nursing staff have experience the health information technology tools as empowering and enhancing their abilities to do their work (de Vries et al. 2017), other studies have found the emergence of "data silos" (Leslie et al. 2017), which we might understand as coinciding with what Fiore-Gartland and Neff (2015) termed as "data valences," calling attention to the distinct interpretations and expectations about the same data that different actors may have in healthcare settings.[4] Supporting this, when asked about the potential utility of machine learning tools to assist in diagnosis, a nurse stated concisely, "Your numbers are only as good as the one who's interpreting them." This nurse also felt that such tools would probably be more useful to newer nurses, and less useful for those with more patient care experience.

While it is essential that we explore how new modes of sense-making are emerging, and creating new epistemic paradigms, we must also examine the new over-reaches and blind spots that accompany such shifts. In this final section, I discuss the themes above and propose that the disciplinary evolution of anthropology and ethnography might be leveraged to reframe data science and the ways in which new "intelligent" technologies are deployed.

REFRAMING THE WORK OF BUILDING AI

Machine Learning as Alchemy

One of the largest gatherings of machine learning researchers is a conference called NIPS, the conference on Neural Information Processing Systems. During the NIPS 2017 conference, which was attended by over 7,000 people (Gershgorn 2017), a long-standing and well-respected member of the NIPS community and current research scientist at Google named Ali Rahimi gave one of the most talked about presentations. During his acceptance speech for an award given for a lasting contribution to the field , he provocatively argued that "machine learning has become alchemy" (Rahimi 2017). This indictment of the field relied on mobilizing the wide-spread understanding of alchemy as a "pseudo-science" – an ancient art built on occult knowledge and superstition. Although the history of alchemy is more complex (Moran 2006), alchemy is commonly perceived as the antithesis of the modern principles of science and reproducible experimentation.

To call a roomful of computer scientists with advanced degrees "pseudo-scientists" was quite a blow. The proposed parallel was that many of today's machine learning models, especially those that involve the use of neural nets or deep learning, are poorly understood and under-theorized; the outputs are correct even if the mechanisms work are unknown.[5] Advances tend to occur more through trial and error than theoretical developments.[6] Rahimi emphasized that while sometimes it may not matter much, when such systems are in charge of people's lives and livelihoods in domains like healthcare and criminal justice, this may be unacceptable. Rahimi concluded: "I would like to live in a world whose systems are built on rigorous, reliable, verifiable knowledge, and not on alchemy."

While Rahimi's talk was widely discussed and well-received, it was not without its detractors. In a Facebook post responding to Rahimi, Yann LeCun, Chief AI Scientist at Facebook, wrote,

> It's insulting, yes. But never mind that: It's wrong! … Sticking to a set of methods just because you can do theory about it, while ignoring a set of methods that empirically work better just because you don't (yet) understand them theoretically is akin to looking for your lost car keys under the street light knowing you lost them someplace else. Yes, we need better understanding of our methods. But the correct attitude is to attempt to fix the situation, not to insult a whole community for not having succeeded in fixing it yet. (LeCun 2017)

The two sides represented by Rahimi and LeCun draw out a fundamental tension in science and engineering: must theory come before practice? Is one more valuable than the other? These are questions with deep epistemological implications: how do we know what we know? What claims to truth are we able to make? Why does it matter?

In the context of healthcare, these questions also have life or death implications. Reflecting on these issues, a Sepsis Watch researcher observed,

> There is nothing in medicine comparable to Newton's Laws of Motion that has stood the test of time for approximately 350 years. What has stood the test of time is the Hippocratic Oath, which concerns ethics, not knowledge. The state of knowledge is constantly evolving and even knowledge that seems reliable and verifiable at one point is rapidly debunked.

As the TED talk by Ziad Obermeyer (2017) referenced early asked, "If a machine could predict your death, should it?" If it meant you could live a better life, do you need to know why?

Just as calls for algorithmic transparency gave way to calls for explainable and interpretable machine learning after critical communities realized that transparency itself is not an end goal, but a means to a goal, so it seems explainable and interpretable machine learning must also be thought of a means, not a goal in and of itself. Members of the Sepsis Watch team voiced the opinion that when people talk about "interpretability" what they really are talking about is causality. The Sepsis Watch machine learning model is not interpretable, and while they have developed the GUI to display particular readings that the model indicates are out of the ordinary, causality is not indicated. Recall: the precise causes of sepsis are still unknown. As discussed above, the team considered the role of trust as key to addressing the motivations behind the interest in interpretability; the tool was developed to be in close alignment with federal guidelines around sepsis treatment, and also over years in collaboration with prominent Duke physicians who will also be overseeing its testing and deployment.

But proponents of the need for explainable and auditable AI and machine learning raise important considerations. The implications of black box algorithms of all kinds for legal due process and accountability (Citron 2008; Crawford and Schultz 2013; Pasquale 2015) are troubling and leave open the door for intentional as well as unintentional unfair discrimination and unequal opportunities to resources and care. Checks on the expectations of machine learning systems to assess, recommend, or decide may be grappled with on specific technical teams, as has been the case with Sepsis Watch, but broader understandings of these limits need to be developed and established more widely.

Machine Learning as Computational Ethnography

Data science and the knowledge it produces are often asked to play the role of objective quantifier (Beer 2016), presenting the cold, hard facts. This perception of immutable truth is a surface to productively crack. Elsewhere, danah boyd and I have proposed that one way to ground the universalizing claims of data science would be to develop a rich methodological reflexivity like that at the heart of ethnography, embracing the partiality and situatedness of data science practice (Elish and boyd 2017). We proposed that machine learning could be seen as a form of computational ethnography. The comparison of anthropology and data science is not as odd as it might seem; Like ethnographers, data scientists immerse themselves with data ("a field site"), selecting data points or patterns from what Bronislaw Malinowski, a founding figure of ethnographic methods, once termed, "the imponderabilia of actual life" (Malinowski 1984: 18). They select from a plethora of data a smaller subset that they find significant based on their intuition and training, and then iteratively develop models and frameworks to fit or explain their findings. Over decades, anthropology as a discipline has developed a foundation of methodological reflexivity, confronting the limits of its own knowledge production, ranging from the articulation of research agendas and areas of focus (Asad 1973; Faubion & Marcus 2009; Hymes 1974) to the cultural and geographical delineations of those areas (Gupta & Ferguson 1997), to the very modes of representation and engagement at stake in ethnographic research (Cefkin 2010; Clifford & Marcus 1986; Taussig 2011). The invocation of ethnography is a means to open up the possibilities of contextualizing what it means to produce knowledge about the world and developing a discipline that can grapple with an iterative and interpretive way of knowing.

Expecting Uncertainty

Contextualizing the insights of machine learning systems as situated and partial is key not only to developing the field of computer science *but also* to facilitating effective integration into everyday work practices. This is especially true in the healthcare context. In a review of studies examining unintended consequences of machine learning in medicine published in the *Journal of the American Medical Association*, Cabizta (2017) argues that machine learning-based decision support systems may be problematic because they "bind empirical data to categorial interpretation" (E2), and require "considering digital data as reliable and complete representations of the phenomena" (E1). In the race to optimize and personalize, the variability of interpretations and the diverse contexts of health data and health care may get lost (Ferryman and Pitcan 2018). He draws on one study to demonstrate the ways in which a loss of context produced a technically valid data model but one that incorrectly predicted mortality risk (Caruana et al. 2015); the model predicted (and it seemed counter-intuitive but was judged to be accurate at the time) that patients with both pneumonia and asthma were at a lower risk of death from pneumonia than patients with only pneumonia. Ultimately, the researchers realized this was the case because patients who had a history of asthma and came to the hospital with pneumonia were usually admitted directly to an intensive care unit, which led to better outcomes. When the model was built, this institutional context and behavior was not represented in the model. Cabizta concludes:

> Users and designers of ML-DSS [machine learning decision support systems] need to be aware of the inevitable intrinsic uncertainties that are deeply embedded in medical science. Further research should be aimed at developing and validating machine learning algorithms that can adapt to input data reflecting the nature of medical information, rather than at imposing an idea of data accuracy and completeness that do not fit patient records and medical registries, for which data quality is far from optimal. (E2)

Like sepsis, the extent of the problem with a data model might not reveal itself plainly at the beginning. This warrants the need for acknowledging limits, iterative development, and vigilant testing throughout a model's deployment. Embracing the contingencies and incompleteness of a machine learning model is a way not only to make a better tool, but also to fit its development within its future use contexts.

CONCLUSION

From Startup to Endups

This paper has highlighted a set of interwoven stories that are implicated in the integration of a machine learning risk-detection tool. All are related to the shifting grounds of evidence and certainty in the context of machine learning systems. One storyline has been about the development of a machine learning technology in the clinical context of an Emergency Department at Duke University Hospital. The development and eventual integration of Sepsis Watch in the ED has brought bring into focus the epistemological implications of introducing a machine learning-driven tool into a clinical setting by analyzing shifting categories of trust, evidence, and authority. These shifts speak to the complex set of social practices that are necessary to ensure effective clincial care even as a machine learning system introduces an automated "intelligent" actor into the clincial setting (Lustig et al. 2016).

Another storyline has about the growth and development of disciplinary evidence. This story, more abstract and historical in nature, brings to bear the history of ethnography on the development of data science. While machine learning and ethnography are often thought of as at odds, they share many orientations toward collecting data and inductively piecing together coherent wholes from minute particulars. What would it mean for machine learning, and data science more generally, to adopt a reflexive and situated epistemic posture? How might this posture better take into account the social contexts within which machine learning technologies are deployed, and the ways in which they might be most effectively integrated into existing work practices?

Taken together, these stories offer insight into the changing nature of evidence, how it is constituted, and how it is recognized as meaningful. Through an analysis of how trust, evidence, authority, and the limitations of technologies are "enacted" (Mol 2002) during the development of a machine learning technology, this paper hopes to have contributed to understanding how machine learning will alter the field of health care and medical expertise, and also how machine learning, as a mode of knowing, will alter how we make sense of the world.

While so many resources and so much energy are focused on creating, finding, or investing in the right start-ups, we need more people to be thinking about "end-ups" (Maeda 2013). How do technologies move from prototype to institutional process? This is the real and hard work of building AI and machine learning technologies. As the case of Sepsis Watch demonstrates, the path is far from straight-forward. Focusing only on a technical functionality is not enough. In addition to technical research that focuses on robust functionality and clinical research that focuses on patient outcomes, *socio-technical* ethnographic research is necessary to understand and plan for the ways in which technologies will be disruptive or effectively integrated into society.

M.C. Elish is a Research Lead and co-founder of the AI on the Ground Initiative at Data & Society Research Institute, an independent non-profit research institute focused on the social implications of data-centric technological development. She received her PhD in Anthropology from Columbia University, and an M.S. in Comparative Media Studies from MIT. mcelish@datasociety.net

NOTES

Acknowledgments – This research is indebted to Mark Sendak and the DIHI Sepsis Watch team for insightful conversations and sharing their time with me. I am grateful to Ian Lowrie, Patrick Davison, and Kadija Ferryman for their feedback on previous drafts. This research was conducted with generous financial support from the AI Ethics and Governance Fund. All oversights, errors, and omissions are entirely my own.

1. During a recent machine learning for healthcare conference, a tweet from the conference organizers put it plainly: "We cannot underemphasize that this is a concrete end-to-end deployment of an accurate deep learning clinical predicion [sic] model for real-time patient monitoring. NO ONE ELSE IS DOING THIS YET! #MLHC2018 #ml4healthcare."

2. Bundles in healthcare refer to a set of predetermined clinical elements of care based on evidence-based practice guidelines and which have been demonstrated to work effectively in concert. The treatment of sepsis with bundles, and the adoption of related protocols as formal federal and state regulation was largely driven by the global health initiative, "Surviving Sepsis Campaign" (2002) which began in 2002.

3. For instance, major technical conferences like NIPS and ICML now have devoted symposia and sessions for interpretable AI, and the explainable AI has been a growing focus of inter- disciplinary conferences FAT* (Fairness, Accountability, and Transparency) conference and the law and technology conference, WeRobot. The Defense Research Projects Agency (DARPA) has opened a program for Explainable AI research.

4. See also previous older work from EPIC 2012 about the implications of implementing EMRs for divisions of labor within clinical settings (Vinkhuyzen, Plurkowski, and David 2012).

5. For an overview of the controversy and differing viewpoints at stake, see Peng (2017).

6. This is not to say that machine learning and data science researchers do not draw on existing theories or attempt to develop what they consider rigorous knowledge practices (Lowrie 2017).

REFERENCES CITED

Asad, Talal
1973 Anthropology & the Colonial Encounter. Ithaca, NY: Ithaca Press.

Beane, Matthew
2018 Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. *Administrative Science Quarterly* 33(January), 000183921775169–37.

Bedoya, Armando, Meredith Clement, M. Phelan R.C. Steorts, Cara O'Brien, and B.A. Goldstein
2018 Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration. *American Journal of Respitory and Critical Care Medicine* 197:A4295.

Beer, David
2017 The Social Power of Algorithms. *Information, Communication & Society*, 1–13.

Burrel, Jenna
2016 How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1-12.

Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini
2017 Unintended Consequences of Machine Learning in Medicine. *JAMA* 318 (6), 517–2.

Caruana R, Lou Y, Gehrke J, et al.
2015 Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.

CDC
2018 Sepsis > Data & Reports Center for Disease Control and Prevention. https://www.cdc.gov/sepsis/datareports/index.html (Aug 24, 2018).

Cefkin, Melissa
2010 *Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations*. New York: Berghahn.

Clifford, James, and George E. Marcus
1986 *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley, CA: University of California Press.

Citron, Danielle
2008 Technological Due Process. *Washington University Law Review* 85, 1249-1313.

Crawford, Kate and Jason Schultz
2013 Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55, 93-128.

de Vries, Aisha, Jos M T Draaisma, and Joris Fuijkschot.
2017 Clinician Perceptions of an Early Warning System on Patient Safety. *Hospital Pediatrics*, September, 0138–10.

DIHI
2018 Home page. Duke Institute for Health Innovation. https://dihi.org/ (Aug 24, 2018).

Dreyfus, Herbert
1972 *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Dummett, B Alex, Carmen Adams, Elizabeth Scuth, Vincent Liu, Margaret Guo, and Gabriel J Escobar.
2017 Incorporating an Early Detection System Into Routine Clinical Practice in Two Community Hospitals," December, 1–7.

Elish, M.C. and danah boyd
2017 Situating Methods in the Magic of Big Data and AI. *Communications Monographs* 85:1, 57-80.

Elish, M.C. and Tim Hwang
2016 *An AI Pattern Language*. New York: Data & Society Research Institute.

Ferryman, Kadija and Mikaela Pitcan
2018 *Fairness in Precision Medicine*. New York: Data & Society Research Institute.

Fiore-Gartland, Brittany, and Gina Neff
2015 Communication, Mediation, and the Expectations of Data: Data Valences Across Health and Wellness Communities. *International Journal of Communication*, May, 1466–84.

Faubion, James and George E. Marcus
2009 *Fieldwork Is Not What It Used To Be: Learning Anthropology's Method in a time of Transition*. Ithaca, NY: Cornell University Press.

Forsythe, Diana
2002 *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press.

1993 Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23(3), 445-477.

Futoma, Joseph, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara Obrien
2017 An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. *Proceedings of Machine Learning for Healthcare* 68 (August), 1–12.

Gershgorn, Dave
2017 The world's biggest and most important AI conference. Quartz.com, Dec 12. https://qz.com/1152156/inside-the-worlds-biggest-and-most-important-ai-conference/ (Aug 24, 2018).

Guidi, Jessica L, Katherine Clark, Mark T Upton, Hilary Faust, Craig A Umscheid, Meghan B Lane-Fall, Mark E Mikkelsen, et al.
2015 Clinician Perception of the Effectiveness of an Automated Early Warning and Response System for Sepsis in an Academic Medical Center. *Annals of the American Thoracic Society* 12 (10): 1514–19.

Gupta, Akhil & Ferguson, James
1997 *Anthropological Locations: Boundaries and Grounds of a Field Science.* Berkeley, CA: University of California Press.

Haugeland, John
1985 *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

Hymes, Dell
1974 *Reinventing Anthropology.* New York: Vintage.

LeCun, Yann
2017 My take on Ali Rahimi's "Test of Time" award talk at NIPS. Facebook.com/yann.lecun. Facebook post. https://www.facebook.com/yann.lecun/posts/10154938130592143 (August 24, 2018).

Leslie, Myles, Elise Paradis, Michael A Gropper, Simon Kitto, Scott Reeves, and Peter Pronovost
2017 An Ethnographic Study of Health Information Technology Use in Three Intensive Care Units." *Health Services Research* 52 (4), 1330–48.

Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, Schorr C, Artigas A, Ramsay G, Beale R, et al.
2010 The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. *Intensive Care Medine*, 36, 222–231.

Lowrie, Ian
2017 Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society* (Jan-June), 1-13.

Lustig, Caitlin, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig
2016 Algorithmic Authority. *CHI* , 1057–62.

Maeda, John
2013 Startups are great, but we can learn a lot from "end-ups," too. Gigaom.com, Feb 3. https://gigaom.com/2013/02/03/we-all-love-start-ups-and-sometimes-forget-we-can-learn-a-lot-from-end-ups-too/ (Aug 24, 2018).

Maiers, Claire
2017 Analytics in Action: Users and Predictive Data in the Neonatal Intensive Care Unit. *Information, Communication & Society* 20 (6), 915–29.

Malinowski, Branislaw
1984 *Argonauts of the Western Pacific.* Prospect Heights, IL: Waveland Press.

Mitchell, Tom
1997 *Machine Learning.* New York: McGraw Hill.

Mol, Annemarie
2002 *The Body Multiple: Ontology in Medical Practice.* Durham, NC: Duke University Press.

Moran, Bruce
2006 *Distilling Knowledge: Alchemy, Chemistry, and the Scientific Revolution.* Cambridge, MA: Harvard University Press.

Obermeyer, Ziad
2017 If a Machine Could Predict Your Death, Should it? TEDxBoston Oct 24. https://www.youtube.com/watch?v=jeGJax4SLP0 (Aug 24, 2018).

Pasquale, Frank
2015 *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, MA: Harvard University Press.

Peng, Tony
2017 LeCun vs Rahimi: Has Machine Learning Become Alchemy? *Synced: AI Technology & Industry Review.* https://medium.com/@Synced/lecun-vs-rahimi-has-machine-learning-become-alchemy-21cb1557920d (Aug 24, 2018).

Rahimi, Ali. 2017. Ali Rahimi's talk at NIPS (NIPS 2017 Test-of-time award presentation). Online video, https://www.youtube.com/watch?v=Qi1Yry33TQE.

Rhee, Chanu, Raymund Dantes, Lauren Epstein, David J Murphy, Christopher W Seymour, Theodore J Iwashyna, Sameer S Kadri, et al.
2017 Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014." *JAMA* 318 (13), 1241–49.

Smith, G.B., D.R. Prytherch, P. Meredith, P.E. Schmidt, and P.I. Featherstone
2013 The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 84(4): 465-70.

Shapin, Steve and Simon Schaffer
1985 *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life.* Princeton, NJ: Princeton University Press.

Stevens, Hallam
2017 A Feeling for the Algorithm: Working Knowledge and Big Data in Biology. *Osiris* 32(1), 151–174.

Suchman, Lucy
2007 *Human-Machine Reconfigurations: Plans and Situated Actions, 2nd ed.* New York: Cambridge University Press.

Surviving Sepsis Campaign
2002 "History" http://survivingsepsis.org/About-SSC/Pages/History.aspx (Oct , 2018).

Taussig, Michael
2011 *I Swear I Saw This.* Chicago: University of Chicago Press.

Vinkhuyzen, Erik, Luke Plurkowski, and Gary David
2012 Implementing EMRs: Learnings From a Video Ethnography. *EPIC Ethnographic Praxis in Industry Conference*, February, 235–48.

Mattresses and Moneyboxes: Cultural Affordances for Microfinance in Jordan

ZACH HYMAN
EPAM Continuum

This case study will present how a multicultural and multidisciplinary team from EPAM Continuum, the global innovation design firm, gathered, analyzed, and presented back different forms of "evidence" to satisfy the complex set of client and customer needs for a Jordanian microfinance bank with 30 branches and 65,000 clients. The team navigated cultural and linguistic barriers as they sought to provide stakeholders and their customers the evidence they needed to confidently design a new "mobile payment service" for their microloan customers. Over the course of the engagement, the firm's team strove not only to research, design, and prototype a new service to hand off to a local development team, but also to (1) use a combination of deliverables and in-field accompaniment to train microfinance bank staff in their process; (2) present evidence demonstrating the deep customer understanding that can result from pairing ethnographic research and human-centered design; and (3) create evidence that the firm's process was both effective and replicable by bank staff.

> "The mobile payment service will be one of the first – if not, the first – for the microfinance sector in the Middle East. As such, despite the fact that mobile penetration has reached approximately 200 percent in Jordan, clients may regard it cautiously."
> (USAID/National Microfinance Bank of Jordan's jointly-written Request for Proposal)

INDUSTRY BACKGROUND

Founded in 2006, Jordan's National Microfinance Bank (NMB) was the third-largest microfinance bank in the Hashemite Kingdom of Jordan as of 2016 when they began their engagement with EPAM Continuum. NMB is a private shareholding institution that finances income-generating projects for underserved segments of society, serving 65,000 clients and 30 branches across the country with "an array of financial and non-financial products and services."

As of the end of 2011 NMB held 12% of the market share of total microfinance clients in Jordan, behind Tamweelcom (25%) and the Microfund for Women (31%). Besides the four largest non-profit players (holding 74% market share), there are also several for-profit microfinance banks with smaller marketshares, and an increasing number of commercial banks are entering the space with their own microfinance products. (Jordanian Ministry of Planning and International Cooperation, 2012)

The Request for Proposal (RfP) laid out the high-level vision for the engagement as follows:

> A user-centered approach will be used to develop the mobile payment service. It will explore clients' behaviors, thoughts, needs, and wants related to financial management and mobile applications. Understanding how clients manage finances and why they prefer one potential

solution instead of another will guide the development of a service optimized to fit their lifestyle–which will increase the probability that they will trust, value, embrace, and adopt it.

The RfP was strongly rooted in language of human-centered design. In writing the RfP, USAID had collaborated with Felipe Cabezas, the Product Development Consultant on long-term appointment to NMB, and the person who would be EPAM Continuum's main client contact within the bank. Cabezas' expertise and interest in human-centered design shone through in the phrasing of the "Program Background" and "Objective" in the project's RfP, which included numerous such phrases as "service design," "user-friendly prototypes," and "user interface (UI) and experience (UX)," and also stated that, "While any and all strategies are welcome to be considered, one-on-one interviews should play a central role" when describing how the field research be conducted.

BUSINESS CHALLENGE

Stakeholders

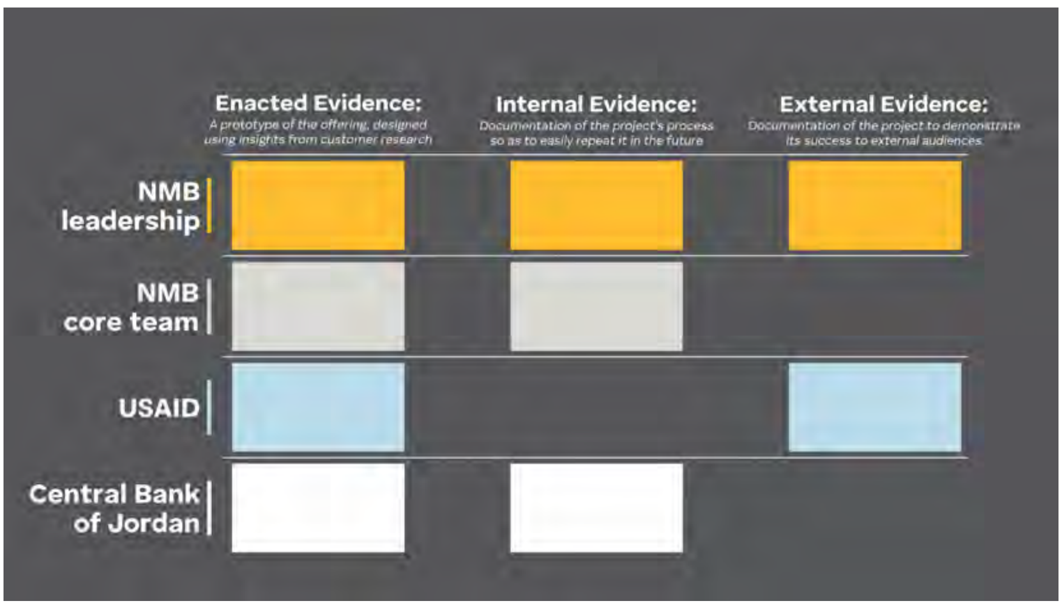


Figure 1. Visual summary of intersecting stakeholder goals. Image © EPAM Continuum, used with permission.

The novelty of an ethnographic approach to understanding people and the use of human-centered design, combined with the significant promise of the project's result (if the aspirations laid out in the proposal were realized) meant that there was a significant amount of both interest and expectation from all of the major involved client parties, of which there were four:

First, the leadership of NMB, which included the General Manager of NMB, the Regional Manager and Head of Non-Financial Services, and the Head of IT, all wanted evidence of the bank's customer-centricity to show the non-profit donors they worked with that they kept their customers in the forefront of their minds. To further raise the stakes, both the Regional Manager and Head of Non-Financial Services as well as the Head of IT were personally involved with the project, and saw the smooth execution of both the process and the result as partially within their responsibility. From the EPAM Continuum team's initial meeting with him, the General Manager wanted "something that could sit on his desk" as proof that the bank had done strong, foundational work to understand what their customers wanted. Speaking broadly, these leaders wanted a successful project to yield a competitive new offering, but part and parcel with that, they were also concerned with good "optics" and positive PR for the organization.

Secondly, EPAM Continuum's primary internal client, Felipe Cabezas, the bank's Product Development Consultant, wanted the project to be a successful reflection of the bank's body of work with non-profit donors (and in particular the United States Agency for International Development (USAID), the main sponsor of the project). In addition to this "organizational" aspiration for the engagement, on a "personal" level, NMB's Cabezas also wanted to further build his skills with the human-centered design process as applied to a complex business challenge.

Thirdly, the donor funding this project, USAID, was just as interested in a successful result for the project as they were in obtaining the proper sequence and collection of reports and artifacts that would enable them to prove to any external oversight committees that the services they had paid EPAM Continuum to render (problem identification, user research, designing wireframes, prototyping a service) and NMB to implement had been fully carried out. Ideally this engagement would also serve as a reflection of the good use to which their funding went, in the event USAID were to request more. At the minimum, USAID wanted proper accountability for tax and auditing standards, and this meant requiring the EPAM Continuum team to fill out and submit a variety of forms, along with examples of the work that was conducted in each project phase. In addition, the NMB/EPAM Continuum team discovered that, upon attending a meeting at USAID on the morning of their second day in the country that they were also expected to have a USAID representative accompany the team to an interview to be able to witness the ethnographic interview process firsthand.

Fourthly, the project team also experienced the somewhat unexpected addition of another critical stakeholder partway through the project: The Central Bank of Jordan (CBJ). The NMB leadership had been in touch with senior officials at the CBJ from the start of EPAM Continuum's fieldwork, and on a day without interviews scheduled the core project team (consisting of the Head of Non-Financial Services, the Head of IT, Felipe Cabezas and the entire EPAM Continuum team) traveled to the Central Bank Building to meet with the Executive Manager for Payment Systems & Domestic Banking Operations and Financial Inclusion Department. Once there, NMB's Head of Non-Financial Services presented the progress of the work to date, after which Stefano Bianchini, the EPAM Continuum team's lead Service Designer and the manager of the project, shared some of the photos from fieldwork and introduced the core ideas behind human-centered design that the team had been employing to date in the project. The CBJ representative was very enthusiastic about both the goal of the project and the novel (for this context) methodology the EPAM Continuum team was using to carry out the work, and wanted to figure out how to have some of her employees gain exposure to the methods of human-centered design as it was being practiced by the project team.

Balancing and ensuring that the day-to-day activities of the project team satisfied the diverse needs of these stakeholders was a consistent point of focus for Bianchini as he tried simultaneously to protect the integrity of EPAM Continuum's process and ensure that stakeholders all felt acknowledged.

NMB's Past Innovations

Even before its collaboration with EPAM Continuum, NMB had a record of innovation that had set it ahead of its competitors in its ability to serve clients. When the EPAM Continuum team initially spoke with the Regional Manager/Head of Non-Financial Services and the Head of IT, they did not frame their innovations in terms of human-centered design, but rather in terms of the bank's internal metrics, like capturing market share and processing clients' microloan applications more quickly. The RfP highlighted a previous tech-led effort, the use of tablets by loan officers to collect loan application information from clients at their homes, instead of clients having to gather documents and travel to a branch to fill out an application. The NMB team was understandably proud that this had, as outlined in the current RfP, "decreased the application-to-disbursal period from 72 to less than 24 hours – freeing up resources to allow loan officers to serve 10 to 15 percent more clients." In the EPAM Continuum team's early conversations with these two NMB employees and Cabezas, NMB's Product Development Consultant, the EPAM Continuum team encountered an unspoken tension; while the two full-time NMB employees spoke of NMB being the first microfinance banks in Jordan to employ tellers in their branches instead of making customers repay their microloans at other bank branches (an advancement that NMB's competition soon copied), the RfP seemed to highlight NMB's recent technology offerings as a means of obviating the need for as many human employees in favor of relying more upon technology.

Having already leveraged mobile technology to realize efficiencies around the microloan application process and broaden their client base, NMB's leadership decided that the next natural step would be to focus on the other key part of the microloan process from the bank's perspective: repayment. While repayment rates were satisfactory overall, the bank's leadership wanted to understand why some microloan customers consistently repaid on time, while others (sometimes in nearly identical circumstances) struggled to do the same. Furthermore, through initial conversations, these variations in repayment rates did not appear to strongly correlate to geography or branch location, loan type, household size, or other metrics that NMB's leadership had previously considered when deciding how to plan a new effort to improve operations. By gaining a deeper understanding of the factors that influenced how people repaid, NMB's leadership hoped the EPAM Continuum team could design and test a solution that solved for the root causes of what prevented struggling microloan customers from repaying on time. Without a more nuanced understanding of the factors influencing repayment, NMB's fast-expanding share of the microloan market could turn from an asset into a vulnerability; too many loans to clients unable to repay could over-leverage the bank and place it in a vulnerable position.

METHODS AND INTERVIEW APPROACH

Identifying and Mapping the "NMB Microloan Journey"

In planning out the methods and stimuli for interviews, the NMB/EPAM Continuum team had to create an interview where respondents felt comfortable, so that respondents could share honestly about their past and current microloan experiences and give candid feedback on the various sketches of low-fidelity ideas the research team presented.

To familiarize themselves with the loan process, the EPAM Continuum team spoke via teleconference with the three core members of the NMB team, interviewing them in depth about all details and stages of the loan process in their eyes: how people typically discovered NMB's services (most often word-of-mouth, or a referral from a friend or relative); the part of the loan process people complained about most (needing to provide two references on a loan application, various identifying documents, and proof of their income as shown on a bank statement); the reasons people most often provided for why they couldn't successfully make a repayment on a given month (unanticipated expenditure wiping out money they'd saved up for their repayment); and more questions aimed at trying to uncover the "human side" of an NMB microloan without the benefit of being on the ground to discover microloan customers' most pressing painpoints through firsthand ethnographic inquiry.

Through these interviews, the EPAM Continuum team came up with the six high-level steps that made up the loan journey:

1. "I realize I need a loan"
2. "I research the best options"
3. "I sign up for my loan"
4. "I get the money I need"
5. "I make my repayments"
6. "I make my final repayment"

The EPAM Continuum team consciously chose a "first-person" framing for the journey steps, because even though initially this made speaking about the steps somewhat cumbersome, it also forced both EPAM Continuum and NMB team members to acknowledge the highly personal nature of getting a loan, and the individuality of each journey.

After receiving approval and buy-in around this as a valid way to break down the microloan process from NMB stakeholders, the EPAM Continuum team adopted this as their general frame for thinking about all potential new ideas, offerings, and processes going forward.

Interview Sequence

The in-depth review of each individual's loan process became what the EPAM Continuum team placed at the beginning of each of their sixteen generative interviews, which they hoped would establish respondents' "expertise" about their loan and make them feel more comfortable sharing their candid thoughts as the team proceeded through the remainder of the interview. After establishing rapport, the lead interviewer would present the set of six "loan journey" cards to respondents, asking them to talk through their most recently completed or current microloan and placing small, colored tiles with either happy, indifferent, or unhappy faces on them on to each of the six journey cards to create a visual record of the respondent's feedback and thoughts around each step of their loan journey. (Sanders, 2014)

Figure 2. The six "journey step" cards, representing the major stages of an NMB microloan customer's journey, along with tiles placed by the respondent indicating their emotions in each step. Image © EPAM Continuum, used with permission.

The research team knew that the majority of respondents would be female, and so decided that the "protagonist" of the loan journey stimuli cards that would be used to help guide respondents should also be female, since through conversations with NMB stakeholders it was revealed women were often the true controllers of the household's finances (even if loans were often applied for in the male head of household's name if there was one).

Once the EPAM Continuum team had settled upon the primary, high-level steps of the journey, they began placing all of the various details, painpoints, and touchpoints between the client and NMB into the journey, incorporating their "understanding from a distance" of some of clients' challenges uncovered through discussions with NMB about the loan process beneath each of the major steps of the loan journey. After standing back to appraise this collage of quotes, client-bank contact points, and conspicuous gaps, the EPAM Continuum team began placing an additional layer of ideas on top of this foundation: loose, basic sketches of ideas that could potentially simplify or improve the client's loan experience. Although EPAM Continuum was tasked with creating a "mobile payment service," they did not want to lose sight of the other business goals NMB had in mind as well—serving more clients, developing desirable and competitive financial products, and becoming Jordan's preferred microfinance institution.

EPAM Continuum categorized the ideas they generated based upon the section of the loan process they affected, and shared these with their broader group of stakeholders at NMB to collect feedback. The research team was careful to emphasize their intention of placing "early stage" sketches of ideas in front of clients, not intending to commit NMB to building the ideas customers reacted most strongly to, but rather to understand the unmet needs that made those ideas resonate with their customers.

Figure 3. Several of the "early stage" idea cards laid out on the floor of a respondent's home for their appraisal. Image © EPAM Continuum, used with permission.

FIELD APPROACH

Stakeholder Management: Balancing Data Quality, Stakeholder Satisfaction, and Client Education

In discussions throughout the Alignment phase of the project, the EPAM Continuum and NMB teams had discussed extensively the "ideal" number of participants to bring along to an interview. While EPAM Continuum was accustomed to bringing a client along to a given respondent interview for the helpful empathy, context, and skills it could build, the team also sought to limit the number of interview attendees to make respondents more comfortable by having fewer strangers in their home.

For this engagement, a typical interview would have a separate note taker and lead interviewer from EPAM Continuum, a translator/fixer, and an employee of NMB for a total of four people attending each interview. In addition, the research team also needed to account for additional guests for certain interviews—an employee of the Central Bank of Jordan interested in learning about human-centered design, and an employee of USAID who was both interested in human-centered design and also wanted to observe interviews to be able to report back that the EPAM Continuum team had in fact delivered upon their promised process.

This made for challenges to the EPAM Continuum team's usual process, as they struggled to balance accommodating multiple stakeholders' representatives that they wanted to attend the interview with the need for intimacy, due to the sensitivity of the conversations that the teams would be having with NMB clients about their finances. (Taylor, 2013) From their past work, the EPAM Continuum team was distinctly aware of the potential decrease in candidness (and therefore quality of ethnographic data) that came with having more strangers in the interview environment. To get around this, the team decided to "over-recruit" and run additional in-context ethnographic interviews around the capital, Amman, (where it would be most convenient for the representatives from both USAID and the Central Bank of Jordan to attend) and be prepared to account for those interviews not yielding the same quality of data as interviews where it was only the core NMB and EPAM Continuum team members in attendance.

Recruiting and Employee Interviews

To recruit respondents for the project's first round of generative interviews, the NMB/EPAM Continuum team worked directly with bank branch managers in the different cities and towns they planned upon visiting for research, chosen for locations with NMB branches, and to give the research team a diverse sample of different population densities (urban *and* rural) and geographic regions. Although the team was aware of the potential trade-offs that would come with the respondent knowing that NMB was the bank about which the team was interested in learning, the team decided that it would be too difficult for the research team's fixer/translator, Shereen Zoumot, to reach out to a respondent independently of NMB and build the trust necessary to get invited into their homes.

To limit the bias that would be inherent if the local NMB branch employee that the respondent personally knew were in the same room during the interview, the team designed the protocol to include the moment where, after making the introduction between the research team members and the respondent, and reminding the respondent to share their honest opinions and answer the team's questions candidly, the local NMB branch employee who arranged the interview would step out of the room. For the purposes of limiting bias, any additional NMB stakeholders from NMB's headquarters observing the interview would not be identified as such, instead being identified either as part of EPAM Continuum's team, or "assisting" the team in some way. The research team felt this method still allowed the respondent to share their candid thoughts about their microloan, their personal finances, and any weaknesses they saw in NMB's current microloan experience, while also making them feel comfortable enough to speak with the research team after having been introduced by a trusted person in their social network.

Finally, to understand microloans as comprehensively as possible, the team interviewed local branch employees (typically the manager in charge of selecting and recruiting customers for the team) at each of the NMB branches that helped the researchers recruit from their customer base. Through conversations with branch employees across the country, the research team understood how the formal mechanisms of credit (background checks into whether an applicant has had past loans, whether they are currently in debt, etc.) are supplemented by less formal "secondary sources," such as when branch employees ask multiple NMB customers with strong ties to the bank and who live in the same neighborhood as the microloan applicant various cross-referencing questions (whether they know the applicant, whether the applicant is in good standing within the community, whether they're a gambler or owe others money, or if the applicant has a background of saving through participation in community savings clubs).

Defining the Social Protocol of an In-home Ethnographic Interview

Despite EPAM Continuum and NMB's efforts to minimize the number of attendees in an interview, oftentimes the surprise would come from the respondents themselves, for whom there was no prior social protocol for an in-home or at-work interview. From previous work, ethnographic interview respondents often did not know how to prepare to accept a group of four or five strangers into their home, and so it often became a sort of "hosting" experience, with respondents offering drinks or snacks, and sometimes invitations to stay for a meal following an interview. Particularly for interviews in smaller urban and rural settings outside of Amman, the NMB/EPAM Continuum interview often became a "social" event involving the respondent's fellow employees, friends, relatives, neighbors, and others. In the first such encounter, the team walked into what they were expecting to be an interview with a 40-year-old female microloan customer in the smaller northern city of Irbid, a place with a population of around 500,000 near the border with Syria. After being greeted by her husband, the team was shown into the family room, where the team was joined by the respondent, her husband, a neighbor, and, at various times, the couple's four children, ranging in age from six- to twelve years old. Sitting in the car afterwards, Zach Hyman, a Design Strategist on the EPAM Continuum team, raised the question, "Do we think this is a problem? Having all those other people in the room [besides the respondent] during the interview?" To which NMB's Cabezas countered, "There wasn't much of an option in that case. We are guests in their home, after all—it'd be too rude to ask a visitor to leave." Zoumot, the team's fixer/translator agreed, saying: "People here don't know what to do with this kind of an interview, so they treat it almost like a party, and imagine how you would feel if someone showed up to your party and asked to you to make your friends leave." EPAM Continuum's Bianchini said: "What if we have a plan in place for next time, if we feel that there are too many extra people and that the number of people is hurting the conversation quality?" The team agreed to develop a plan in case encountering a similar situation in the future, which the team ended up needing to enact several days later, in the village of Deir'Alla.

After being welcomed into the respondent's home in Deir'Alla, that of a 48-year-old woman who ran a small convenience store adjacent to her home, the team sat down and began the interview as normal. About 30 minutes in, once word had traveled around the small neighborhood that there were visitors, a neighbor showed up bearing a silver platter brimming with steaming cups of tea — one for each of the people in the room, plus herself — and proceeded to join into the conversation, despite not having any prior experience with microloans or interactions with NMB. Hyman and Zoumot struggled to keep the conversation focused upon the firsthand experiences of the NMB customer as her neighbor shared various observations and anecdotes — stories of the risks of selling things on credit to customers, or the challenges of trying to assess whether a wholesaler was taking advantage of her. Eventually the research team members made eye contact with one another, and set the previously agreed-upon plan into motion; as the lead interviewer (and ostensibly the one with control over the conversation), EPAM Continuum's Hyman stood up and said, "Ah, I forgot something important in the car, can we take a short break so I can go out and get it?"

After Zoumot translated this, Hyman and NMB's Cabezas walked out to the car, where Cabezas explained to Taha (the team's NMB-appointed driver who was familiar with the roads all across Jordan, and who typically waited in the car while the team conducted interviews) that it would be a great help if he could come in and have a friendly, separate conversation with the talkative guest while Cabezas (who was a strong Arabic speaker) wrote down notes. While a slight deviation from protocol, this still managed to create a positive outcome for all parties; the guest would feel like she was having her opinion heard, and Cabezas would help advance the research by asking her questions about money and her financial life. The team's "bonus" respondent also signed an NDA to make things feel as "official" as possible. An unforeseen positive outcome from Cabezas' conversation with the bonus respondent was that he and Taha were able to direct and control the flow conversation with her; instead of leaving open the opportunity for her to add commentary as she liked to the research team's conversation with the NMB client. This way, Cabezas could assemble a meaningful set of observations and takeaways through his structured conversation with the guest.

The team decided to use the spare audio recorder to capture the conversation, so that it would feel as if the separate conversation she was having with Cabezas was no less important than the conversation that the EPAM Continuum team was having with the "intended" respondent. Signing an NDA also meant that both her privacy and the integrity of the research team's data and methods would be protected. Since the team only brought a limited amount of incentive packages along, the team was unable to give a separate incentive package to the bonus respondent.

Incentives

One element that defined the NMB/EPAM Continuum team's ethnographic interviews and might have caused them to be interpreted as more of a "social" event was the team's choice of interview incentive. After the EPAM Continuum team landed in Jordan and met with Zoumot, their fixer/translator, for the first time, they reviewed their intended approach to how interviews would ideally run. When Hyman asked Zoumot what she thought a reasonable amount to pay respondents would be for what would be around a two-hour conversation in their home, her response was "Oh, no — paying them cash would be considered very rude." The team had encountered what Jan Chipchase covers in detail in *The Field Study Handbook,* when he begins by sharing how "sometimes non-monetary incentives are a better choice." (Chipchase, 2017). The research team considered factors such as whether refrigeration would be available for all respondents (deciding to assume it would not be) and what sorts of products and brands would both be most desirable and would confer the most status upon the respondent as their recipient.

In assembling the box of incentives, Zoumot recommended a value of approximately $50 US equivalent as appropriate to award a respondent for their participation in the two-hour interview, and together Zoumot and the EPAM Continuum team wandered the aisles

Figure 4. Two incentive packages for the day's interviews. Image © EPAM Continuum, used with permission.

of a grocery store located nearby their neighborhood-based pop-up studio in the Jabal al-Weibdeh neighborhood of Amman. After much deliberation, the team decided to include in the box:

- almonds
- dried apricots
- three flavors of Lindt dark chocolate bars
- a jar of Nutella
- a jar of Honey
- a box of dried dates
- a bottle of organic olive oil
- a bottle of Vimto (a type of fruit cordial)

Figure 5. Contents of an incentive package. Image © EPAM Continuum, used with permission.

All of the items were non-perishable, did not require refrigeration, and conferred status through an appearance of elevated taste amongst the recipients. There were also items that would appeal to any children in the household.

CULTURAL CONTEXT

Before the research team could understand what drove successful repayment, they began with a broader inquiry into how people in Jordan thought about money, technology, and any overlaps between the two. The team gained this understanding through:

- stakeholder interviews with both senior bank leadership who understood the systems of money broadly
- local level, community-based NMB branch employees who interacted with customers on a daily basis
- secondary research into both technology usage and microfinance in Jordan
- an initial round of sixteen in-depth, two-hour interviews with microloan customers of different ages and loan types/sizes, spread across five Jordanian cities and villages

The Centrality of Cash

In Jordan, cash is vastly preferred over other means of payment in respondents' daily lives. That said, the Jordanians who the research team interviewed each acknowledged the inherent drawbacks that came with carrying cash.

> "I'm used to carrying around cash, but it still worries me when I do."—Rema, 64, F, convenience store owner

For Rema, one of the first respondents the team interviewed, when it came time in the two-hour interview to broadly discuss whether she ever imagined somehow storing money on her phone, her reaction was quick: "No, never." The team let the answer stand, with the lead interviewer looping back to discuss it in more depth after Rema had had the chance to see some of the stimuli the team had arrived with, and that she was encouraged to indulge in the not-yet-possible. When re-asked in a similar way, Rema replied: "Yes, I could see storing money on my phone, but I would want to put it in myself, and then I would want to be able to have the cash [be dispensed directly out] of the phone whenever I wanted."

While technically infeasible, Rema's design requirement spoke to both her dedication to and comfort with cash, as well as her lack of trust in being able to invisibly "send" her money to someone or somewhere else using only her phone. Rather, when she needed money stored on her phone, she would have preferred to dial in the amount required and then hand the physical cash to the recipient. Whatever this sort of object might be called, one certainty was that it certainly did not behave like any wallet (or "e-wallet") that the research team had ever heard of.

The high standards Rema held for her digital device to hypothetically function as a means of storing and safely carrying physical cash stood out to the team, and from that early interview, the research team made it a key component of subsequent interviews to explore precisely what it was about cash that made people reluctant to put it in one of the several "e-wallets" offered by Jordan's mobile network operators.

The Importance of In-Person Interaction, and the Connection to *Wastah*

As the research team explored Jordanians' preference for cash more deeply in subsequent interviews, it became clear that cash was something that, for all of its perceived risks and weaknesses, was important for bringing people together, into the same time and place, to physically exchange value (Maurer, 2013). The comfort with cash seemed to go part and parcel with people's preference for in-person interactions with their friends and shopkeepers.

Rema, the woman who would have preferred to place money "into her phone" provided she could print it out directly from the phone at her convenience, shared with the research team how she enjoyed the act of repaying her loan in person at the bank branch. When the team probed into what she enjoyed about the journey to the local branch, both a time- and effort-consuming act for a 64-year-old woman suffering from mobility issues, she said, "they're my friends [at the branch], and I enjoy speaking with them. It wouldn't be the same to just use a phone [to repay a microloan]."

As the team asked subsequent respondents about this topic, they learned more about the relationships formed through the in-person interactions and exchanges of money and favors, and how they helped build what respondents referred to as *wastah*. Building up one's *wastah,* or 'personal connections,' are ways that respondents make the various processes of daily life work more easily for them. For Wafa, a 39-year-old hairdresser, when referring to the local NMB branch where she made her repayments, she did not use the phrase 'bank branch,' instead specifically referring by name to Noor, the teller who she sought out each time she visited the bank. Each interaction contributed to building her *wastah* with Noor, and if there was ever a mistake or misunderstanding, "I can show up in person and yell about it!" Likewise, if Noor (or one of her friends) ever needed a last-minute haircut appointment, Wafa would go through whatever pains necessary to arrange one.

Besides cash being useful to build *wastah* between those people exchanging it, the physical nature of cash also helped people underscore the seriousness of what they were committing to, such as when Mohammed, a 36 year-old car mechanic from Amman, shared that while he would be open to the idea of using his phone to help repay his microloan,

> "If I were ever to loan money to a friend, I wouldn't trust using a phone to do it. I'd want to make the loan using cash, so that I could look [my friend] in the eyes as I handed them the money... [this interaction] shows it is important to both of you, that [this personal loan] is a serious matter."—Mohammed, 36, M, mechanic

The flip side of this recurring theme of reliance on cash was an aversion to almost every other form of storing value, including debit and credit cards issued by commercial banks and phone-based "e-wallets" offered by Jordan's three main telecom providers: Zain, Umniah, and Orange.

The main driver for the lack of trust of credit cards originated from either firsthand negative experiences or secondhand stories about banks "tricking" customers who had used such cards. Mohammed, the car mechanic from Amman, shared a story of how he used his credit card to withdraw cash multiple times from various ATM's over the course of a day while he ran various errands. He learned at the end of that month that he'd run up significant fees on that day for reasons he did not understand, and promptly threw his credit card into the discarded motor oil-powered furnace that he used to heat his car repair shop.

To underscore the country's reliance upon cash, while the research team was in Jordan, Amman was one of the few cities in the world where Uber was piloting allowing riders to pay a driver for a ride using cash instead of with a credit card linked to the app (as with almost everywhere else in the world). Although credit cards were technically available in Jordan, relatively few Jordanians chose to have one (compared to the 55% of Americans who are credit-card holders). (Statista.com, 2018)

The Familiarity of Community Savings Clubs

One common 'financial product' that had roots going back into ancient history was the 'community savings club'. A group of typically between eight and twelve people (most often women, according to NMB/EPAM Continuum's field research) would meet each month and all the members would pool their money to pay a set, small sum to one of the members. The group would rotate who receives each month's total payment until all members had been a recipient of the group's pool of multiple small payments, at which point they would decide whether to start over. For the respondents the research team interviewed, there was no interest accrued on the sum they received from the other members of the group, and the savings clubs functioned primarily as means of beating participants' temptation to spend, and building *wastah* between the participants.

Multiple respondents also liked the idea of money 'staying in the community,' rather than going to an 'outside' entity like a bank (even if they trusted the bank's employees, brand, and process). Feda, a merchant who was the head of her community savings club, when asked to compare the experience of repaying a microloan to making a payment into her community savings club, described the latter as a "reverse loan," saying: "When I make a repayment for my savings club, I feel good because I know it will be coming back to me later. When I repay NMB, I feel sad because I know [the money] is going 'away' from me."

With her savings clubs, Feda felt there was less pressure for her to invest in a productive asset that would bring financial returns, because participating in savings clubs did not incur any interest charges. Savings clubs made Feda feel more comfortable choosing to invest in non-productive assets like gold.

The Challenges of Distance

The team's research also revealed differences between rural and urban dwellers when it came to how they actually repaid. For some rural microloan customers, the nearest bank branch might be 45 minutes away by car. If the microloan customer did not own or have easy access to a vehicle, to travel such a great distance meant either waiting on the side of the road and trying to catch a ride into town with a passing driver (while carrying the entirety of one's loan repayment with them in cash), scheduling transportation to the bank branch (which could cost significant monetary or social capital), or giving one's loan repayment amount and bank ID card to a (hopefully sufficiently) trustworthy friend who was headed into town and trust that they would stop to make the repayment at the bank branch on the client's behalf.

Even for urban dwellers, who typically lived closer to the bank, making repayments at the branch could still be a costly or troublesome endeavor. For Maha, a food seller who suffered from back problems after decades of working over a stove, her lack of mobility meant she would have to take a relatively expensive 4 dinar (~US $5.50) taxi ride both to and from the bank branch each month for the times when only option was repaying in person herself. Over time, she picked up workarounds for this expense, including calling her daughter who lived nearby to come and take the repayment to the bank for her, or giving the repayment to her downstairs neighbor, who is an NMB employee, for him to carry with him to work and repay on her behalf.

ANALYSIS & SOLUTIONS

Cultural Affordances in Practice

Upon arriving back in Milan studio from Jordan to begin the analysis and sensemaking process, the research team printed out, rewrote, and delved into the thousands of discrete pieces of data that they had gathered over the dozens of hours spent in conversation and observation in people's homes and businesses (Madsbjerg, 2014). In doing so, the team sought to understand peoples' lives broadly, but were particularly interested in what shaped financial behavior, how that translated those behaviors and beliefs into their relationship with NMB, and how that influenced whether they repaid their microloan. Looking across all sixteen respondents' lives, the team sought to understand the innate behaviors and motivations that drove some to consistently make their monthly repayments and not others. The research team's goal at the end of this two-week phase was to identify strong themes that appeared across conversations and throughout the research, and then invite the core NMB team members to Milan for a several-day workshop where the EPAM Continuum team would introduce the themes and explain the observations that each was drawn from, based upon what the combined team members observed together while immersed in in-context conversations and observation across Jordan. From there, attendees would work together to transform the observed themes into actionable opportunities with which to move forward into the Envisioning and Prototyping phase.

To start, the EPAM Continuum team began by examining the relationship between respondents' financial lives and their digital lives. All respondent households that the researchers visited owned at least one cellphone, a multifaceted object that played roles in people's lives ranging from entertainment to communication. During each interview, the research team brought up a local mobile money service offered by the country's largest mobile network operator, Zain, that allowed one to store money on one's phone after having deposited it at a local Zain shop. None of the sixteen respondents interviewed had ever used the service (nor had most heard of anyone who had), and the majority of respondents dismissed such a service as either unnecessary ("I've already got one wallet that I carry, why do I need another in my phone?") or risky, as respondents cited fears of losing their cellphone, or children breaking a phone as they played with it, or a child accidentally giving away all of their money (as phones were often shared household devices, as with many other countries where mobile phones were still an emerging phenomenon) (Von Bayer, 2017). Respondents vastly preferred their physical wallets (and the cash inside of it) over a phone-based "e-wallet" (a term used by Jordan's three primary mobile networks), particularly when compared with their current money storage solutions, which included hiding money under their mattress or storing it at a trusted relative or neighbor's home to keep it safe from immediate family members looking to spend it. The research team found that this analog approach to saving money spoke to a more familiar mental model for respondents (Young, 2011).

One strong theme that emerged across the lives of those who reliably repaid their microloans monthly was their reliance upon a common object: a *hassalah* (the Jordanian equivalent of a piggy bank). Whether in the form of plastic-molded orbs with a slot for money, or aluminum cans wrapped with adhesive paper covered in colorful anime cartoons, *hassalah* were key enablers of the most successful microloan customers' repayments, as they helped build and enforce regimented, daily microsaving. These objects were universally recognizable, and of the four respondents the research team interviewed who ran small convenience stores, all of them sold various shapes and sizes of *hassalah.*

Although the phone was equally ubiquitous in Jordanians' daily lives, its open-ended and fast-evolving role stood in sharp contrast with the static, focused utility of the *hassalah.* Seeing the pivotal role *hassalah* played in many Jordanians' social practice of saving money, the research team considered whether and how to incorporate the *hassalah* into the service they were designing in the form of a "cultural affordance." John Payne, Co-founder and Managing Director of Moment Design, coined the phrase "cultural affordance" to describe a "culturally specific, iconographic image" that connects a new offering in a market to a familiar, pre-existing set of social practices by associating the novel offering with objects and ideas that are already embedded within people's daily lives and behaviors (Payne, 2015).

Designed properly, cultural affordances ground an unfamiliar new product or service within familiar mental models.

The EPAM Continuum team believed that the *hassalah,* reframed as a cultural affordance within the new service, could be a powerful behavior-changing force for explaining the value of the novel new service to Jordanians. The team hoped it would link to a behavior of saving money to repay microloans, which respondents said they aspired to but found difficulty doing. The team believed that a new service which adopted the mental model of a *hassalah* on one's phone would speak more meaningfully to microloan customers than that of an "e-wallet". The team's hope for the following Envisioning and Prototyping phase was to test whether NMB's microloan customers would react differently to a service that worked more like a *hassalah* (an object designed to help people save money) rather than a wallet (an object designed to help people carry and spend money).

While the team hoped to change how NMB's customers thought about storing money on their phones through shifting their mental model from phone as "e-wallet" to phone as "e-*hassalah,*" they wanted to preserve as much as possible the set of human behaviors that enabled the familiarity and centrality of *wastah* that was so central to many NMB customers. EPAM Continuum's proposed service prototype sought to strike the balance between simplifying the microloan repayment process through a digital element, while at the same time letting customers continue to broaden and deepen the *wastah* enabled by in-person exchanges of money and generosity.

Analysis workshop: Building Credibility in the Evidence

At the Analysis workshop in Milan, attended by the larger group of senior NMB stakeholders as well as the members of the research team, the EPAM Continuum team. With the help of the NMB employees who had attended particular interviews and could speak to what they learned from the respondent, the attendees discussed notable observations and learnings that had been gathered from each individual respondent. This not only transferred the knowledge that was in the heads of those who were in the field into the heads of other attendees, but also placed the presenting attendees from NMB on a pedestal of expertise, as they were able to speak firsthand about the in-home and in-business interviews they attended, further buying both them and their NMB audience members into the human-centered design process's "way of knowing" based upon firsthand human experience. To help with this process, NMB and EPAM Continuum team members presenting about the key takeaways of different interviews, and referred to print-outs of the interview debriefings pinned to the walls.

The EPAM Continuum team and Cabezas co-presented nine themes to the attendees from NMB, with each theme focusing upon a different observation from the field—from how microloan customers were (informally) using new digital channels like Whatsapp to gather the documents needed for a loan application and send them to NMB loan officers, to the many different informal workarounds people used to keep track of the repayments they had made to NMB and the total number of repayments remaining, to the different means people used to help themselves save each month and repay their microloans on time.

The second day began with an open-ended conversation around the nine themes, with the NMB attendees going around the table and talking about which themes they most strongly believed in, and which ones they were more skeptical of. From there, the

NMB/EPAM Continuum team guided the NMB attendees through a handful of different potential offerings that could be prototyped and tested in the Envisioning and Prototyping phase, along with how the different prototyped offerings would connect to the different themes.

By connecting particular themes to the various proposed ways of prototyping offerings in the following phase, the senior NMB attendees could speak to which prototyped offerings they felt most comfortable bringing to life from an organizational capability perspective. This was critical to the EPAM Continuum team's goal of setting the project up for the highest chance of success after their direct involvement ended. The EPAM Continuum team made the conversation around feasibility and logic for pursuing a particular prototype both open and participatory. The goal was to create an atmosphere where the various senior NMB stakeholders in attendance would feel able to ask for any support and resources necessary from colleagues or superiors to make the project a reality if they were tasked with implementing part of it. This would increase the likelihood of successfully scaling the project from a prototype to an in-market offering following the Prototyping phase. From his past experience, EPAM Continuum's Bianchini shared that letting the stakeholders who would potentially be taking control of the project down the road voice their opinions on the prototype's feasibility was a critical component of the project's eventual success. By having say over what could be the beginning of a new product or service, NMB stakeholders would feel both more accountable and in greater control of the project outcome knowing they would eventually be taking responsibility for it.

At the conclusion of the workshop, the EPAM Continuum team presented attendees with copies of a full-color, printed, and bound book of all of the interview debriefings from the generative phase of research, summaries of the nine themes bubbled up from those interviews, and a collection of photographs from the field that documented the team's process and approach. The printed book was important for the senior NMB stakeholder attendees to carry back physical evidence of both the workshop and the project to date. By creating an inherently sharable object to present to their teams, they would be the de facto experts about—and advocates for—the past and future of the work. The EPAM Continuum team felt the book would also play an important role both for NMB's General Manager, who sought physical evidence (in the form of "something to put on [his] desk") about NMB's ability to listen to and design for customers. Finally, EPAM Continuum held on to several copies of the book to pass along to other key stakeholders so they could have evidence of the process they were involved in and feel more invested in the outcome, in particular the project's partners at the Central Bank of Jordan and USAID (for whom the book would be a compelling physical piece of evidence for responsibly-spent funding).

CORE SOLUTION: E-HASSALAH

The solution the senior NMB stakeholders aligned upon with the EPAM Continuum team in the Analysis workshop was to experiment with a mixed digital/analog service that would afford two new types of flexibility to NMB's microloan customers.

The first type of flexibility was the physical location where NMB's microloan customers physically repaid their loans. The prototype service that NMB/EPAM Continuum wanted to test in the Envisioning phase would enable NMB customers to repay their loans at the shops of trusted local merchants in their community, rather than having to arrange costly travel to a bank branch to make their repayments.

The second type of flexibility would be how microloan customers repaid their loans. Rather than limiting clients to repaying their full repayment amount on (or before) the day their monthly repayment was due, the prototype would simulate allowing respondents to make smaller repayments at a pace similar to how they would put money into a *hassalah* to save it: a small amount each day. The prototyped service would enable customers to repay in whatever denominations they chose over the course of a month rather than in one lump sum (as long as their cumulative small repayments added up to be the full repayment amount).

The research team saw these two flexibilities as interdependent. Individually, they were certainly novel service features for any microfinance bank, but together, the team believed they mutually re-enforced one another to make a truly novel service that could make the lives of NMB's microloan customers significantly easier, and make NMB's microloan offering significantly more attractive as compared to its competitors.

At the heart of the digital component of the prototyped service, and the enabler of these two types of flexibility, was the cultural affordance of the 'e-*hassalah*' As microloan customers made multiple small repayments on their microloan over the course of the month, the sum of the repayments they had made to date for that month since the last repayment would be reflected in their e-*hassalah*—an in-app visualization of a *hassalah* accompanied by graphs showing their previous months' successful repayment, and their progress towards repaying their full amount for this month. The team felt the e-*hassalah* feature would help meet the needs that multiple respondents raised throughout the initial round of generative interviews:

> "If I could make smaller repayments two or three times a month, it would make me feel like repayment was in easier reach... maybe I could more easily share the burden of repayment with my parents."—Aseel, 20, F, student

> "I think if someone has repaid as many loans as I have, the bank should let you choose how you want to divide your monthly repayments up over the course of a month... paying all at once is a big burden."—Bilal, 35, shopkeeper

ENVISIONING & PROTOTYPING: DESIGNING A SERVICE, PRESENTING EVIDENCE

Prototyping Interview Process

For the final stage of the project, Envisioning and Prototyping, the team agreed that it would be critically important to understand the proposed service from the perspective of both "users" and "providers"—in this case, both microloan customers with loans to repay, and the trusted local merchants who would be accepting those customers' repayments. To understand about what respondents liked about the prototype and what needed to change, the team planned on recruiting for and running twelve evaluative service prototype interviews, where they would interview a total of four local merchants (two urban and two rural), and speak with two loan customers at each of the four shops.

The team started by reaching out to bank branches in a both urban and rural areas to seek out local merchants with whom the local branch had particularly close relationships.

The research team first interviewed the merchants, having one of the NMB team members, Saif Al Khalili, pretend to be a microloan customer coming in to make a repayment at their store using the new service. As NMB's Al Khalili stood on one side of the counter holding a phone that had been loaded with the "customer-facing" digital wireframes, Hyman and Zoumot would be standing on the other side of the counter next to the merchant, guiding them through the "merchant-facing" digital wireframes, as Cabezas and Bianchini supported with note-taking and video-recording. Both customer- and merchant-facing wireframes had been programmed into older-model iPhones dedicated to prototyping.

Figure 6. Prototyping components, clockwise from top left; physical paper receipt to gauge receptivity towards physical versus digital receipts, physical sign made of foamcore identifying store as an official "NMB repayment point", paper prototype created for feature phone users, digital wireframes made displayed on a prototyping phone using InVision for smartphone users. Image © EPAM Continuum, used with permission.

After testing the prototype with each merchant and gathering feedback about what they liked and would want to change about the service, the NMB/EPAM Continuum team built enough trust and rapport with the merchant for them to be comfortable with having the team 'take over' their store for the several hours needed to run the set of subsequent customer interviews.

For the pair of customer interviews that followed each merchant interview, an NMB microloan customer would simulate the journey of traveling to the merchant's store, where the interview would start with collecting feedback on the sign greeting them outside of the shop, identifying it as an NMB Microloan Repayment Point. In the initial round of generative interviews, the NMB/EPAM Continuum team asked in depth questions about what would make it easier for respondents to trust a given service, and the answer came in the form of official-looking signage or stickers that made it clear that a given store was an officially sanctioned provider of a given service.

By prototyping the service in a way that would resemble reality as closely as possible, the team understood what users liked and disliked about the service as it could appear in the future. To accomplish as life-like a prototype as possible required the analog and digital components of the service to be of equal fidelity (and credibility). The goal was to have merchants and customers believe that the service was 'real' enough to believe it could actually exist, but still in a form where they would feel comfortable giving suggestions and sharing candid feedback on how it could be improved.

After sharing their opinions on the several examples of official-looking signage, the microloan customer would be standing in front of the merchant's counter inside, where they would be handed the iPhone with the customer-facing set of wireframes. From that point, the remainder of the interview would consist of Zoumot and Hyman guiding the microloan customer through the prototyped repayment process, continually collecting feedback and gauging their level of comfort with the process. Throughout the prototype interview, the "merchant" the microloan customer would be speaking with was actually Al Khalili, the NMB employee who was pretending to be another merchant working in the store and sharing the space behind the counter (sometimes a comically small amount of space) with the store's actual proprietor. As Zoumot and Hyman helped guide the customer through the prototype's wireframes, Al Khalili would be navigating in parallel through the merchant-facing wireframes, right down to the moment of when the microloan customer actually 'repaid' the loan by handing Al Khalili some play Jordanian currency that the customer had received at the start of the interview. The prototyping interview concluded with Al Khalili handing the client a mockup of a physical paper receipt that reflected the repayment that they had just made, while a digital copy of the very same receipt automatically appeared in the prototype's screen. The final part of the microloan customer interview sought to understand from customers whether both a physical and digital receipt were necessary, and if not, why the client would be sufficiently comfortable with just one of the two.

For the backstage of this e-*hassalah*-centric service, the research team calculated that several times per month, a local NMB branch would dispatch an employee to a designated neighborhood in their territory to collect the money that the neighborhood's microloan customers had repaid to the merchants in the community whose shops were certified as official repayment points. After double-checking on the amount collected back at the NMB branch, the neighborhood's microloan customers' balances would be updated accordingly in each of their files based upon the amount they repaid, all without microloan customers needing to travel to the branch and stand in line to repay (or trust someone else to do so on their behalf).

Final Presentation and Hand-off

Following the Envisioning phase, the EPAM Continuum team was set to make their final presentation to the leadership of NMB, as well as delegates from the Central Bank of Jordan and USAID over the course of an hour-long meeting. To most compellingly share with them the evidence around the prototype that was gathered during testing, EPAM Continuum and NMB agreed that it would be important to show both merchants and microloan customers "speaking their own truths" as evidence of how they felt about the prototype. EPAM Continuum's Bianchini cut together a highlight reel showing both merchants and customers talking about what they admired about the prototyped service, and why they were excited to see it become a reality. To give the audience a glimpse into the mechanics of the prototype in a more approachable way than just sharing static wireframes, EPAM Continuum's Hyman used Apple Keynote to develop a narrated walkthrough of the wireframes, showing both the customer- and merchant-facing digital wireframes suspended against a changing background of photographs taken during testing, and accompanied by a narration the app's key features and explaining how they were tested with actual merchants and NMB microloan customers.

Following a successful presentation to both the extended team of stakeholders and NMB's Board of Directors, the EPAM Continuum team spent the day working alongside the NMB core team to align upon the project's implementation. While the original RfP that NMB shared requested that the responding bid include the full-fidelity development of all of the new service's digital and physical touchpoints, EPAM Continuum team members advocated for a different approach; if EPAM Continuum led the development of the app from their headquarters in Boston, the process would risk being expensive, time-consuming, and not the most efficient use of NMB's or EPAM Continuum's resources. Instead, the EPAM Continuum team proposed to help NMB write a separate, additional RfP to distribute to local software development companies. By having a local Jordanian company collaborating with NMB to code and implement the full-fidelity version of the service's digital touchpoints, the costs to NMB would be far lower, and any questions the software company had about the development process could be answered quickly and in-person.

On their final days in Jordan, the EPAM Continuum team worked closely with NMB to develop User Stories and a detailed work plan to help better guide and inform the local software developers who would be responding to NMB's technical RfP, which was co-written with guidance from EPAM Continuum. NMB eventually chose to work with a Jordan-based software developer with a track record of developing digital services for several other major financial institutions in Jordan.

IMPLEMENTATION AND OUTCOMES

In April of 2018, NMB released a digital app that enabled customers to repay their microloans in a more flexible way by letting them choose the amount of money they paid out of the app's associated e-wallet. NMB partnered with the Central Bank of Jordan to integrate the NMB app with the Central Bank's "Mahfazti e-wallet," a system developed by the Central Bank and run by the mobile network operator Umniah. Although technically enabling microloan customers to repay their loans remotely using their phone, the service requires the microloan customer to have first placed money into a Mahfazti e-wallet account, which can be done by visiting a participating local Umniah mobile vendor.

While the local software development company chose not to draw upon the cultural affordance of the e-*hassalah,* and NMB has not yet fully scaled the ability for customers to pay in cash at a local merchant and have that amount be considered in the same way as making a repayment at an NMB branch, the current service is a strong middle step along the way through enabling gradual repayment by paying money out of the Mahfazti e-wallet. Released in the spring of 2018, the Android version of the app (Android is the dominant mobile operating system amongst NMB's customers) has been downloaded more than 1,000 times in the first three months it was on the market.

Multiple features of the app in its current form can be traced back to the evidence that the NMB/EPAM Continuum research team uncovered in the field and presented to the extended NMB team during the Analysis workshop. Informed by the findings of how customers struggled and improvised to keep track of their loans, the app lets users seeing both their total amount repaid on their current loan, as well as number of remaining repayments (both pieces of information that the research team heard respondents wish they had at their fingertips over the course of their field research). The app also presents NMB customers with a log of digital receipts that show all of their past successful payments, creating a "digital paper trail" to assuage the fear that some respondents in NMB/EPAM Continuum's generative research shared, which was that money could 'disappear' within their smartphone without a trace. NMB's service also lets users apply for a new microloan by using their phone to submit key documents and information, instead of forcing them to bring in physical copies of the documents (or informally use text messages to submit documents to NMB's local branch employees, which the research team heard some respondents tell they did). These features, based upon the ethnographic data that the research team shared with NMB stakeholders throughout the process, speaks to the potential impact of the right form of evidence presented in the right way—even within organizations that haven't previously relied upon the use of ethnography to understand and design for their customers.

Zach Hyman is a Design Strategist at EPAM Continuum who has previously worked across China, Myanmar/Burma, Jordan, Italy, Viet Nam, Thailand, Denmark, England, and the US designing both products and services for retail, IoT, healthcare, transportation, and education. He conducted Fulbright research into resource-constrained creativity across China, and blogs at SquareInchAnthro.com

NOTES

Support for EPAM Continuum's engagement with the National Microfinance Bank of Jordan was generously provided by the United States Agency for International Development through their Jordan LENS Local Enterprise Support Project. This case study does not represent the official position of EPAM Continuum, USAID, or the National Microfinance Bank of Jordan.

REFERENCES CITED

Chipchase, Jan, and Lee John Phillips
2017 *The Field Study Handbook*. San Francisco: Field Institute

Jordanian Ministry of Planning and International Cooperation
2012 *Jordan Microfinance Market Study.* ProDev.

Madsbjerg, Christian, and Mikkel B. Rasmussen
2014 *The Moment of Clarity: Using the Human Sciences to Solve Your Hardest Business Problems.* Boston, MA: Harvard Business Review Press

Maurer, B., T.C. Nelms and S.C. Rea
2013 "'Bridges to Cash': Channelling Agency in Mobile Money," *Journal of the Royal Anthropological Institute*, 19 (1): 52-74

Payne, John
2015 *Getting Beyond Use: Designing for meaning with cultural affordances.* Presentation at Midwest UX 2015, Pittsburgh, PA.

Sanders, E. B., and Stappers, P. J.
2014 *Convivial toolbox: Generative research for the front end of design.* Amsterdam: BIS.

Statista: The Statistics Portal
2017 *"Number of credit card holders in the United States from 2000 to 2016, by type of credit card (in millions)"* *Statista* website. Accessed June 4th, 2018. https://www.statista.com/statistics/245372/number-of-cardholders-by-credit-card-type/

Taylor, E.B. and H.A. Horst
2013 *"From street to satellite: Mixing methods to understand mobile money users."* Proceedings of the Ethnographic Praxis in Industry Conference, London, September 15-19, pp. 63-77.

Von Bayer, Sarah L.
2017 *"Thinking Outside the Camp": Education Solutions for Syrian Refugees in Jordan*, EPIC 2017 Proceedings.

Young, Indi
2011 *Mental Models: Aligning Design Strategy with Human Behavior*. Brooklyn, New York: Rosenfeld Media.

When Customer Insights Meet Business Constraints: Building a Go-to-Market Strategy for a Smart Home Offering in a Regulated Space

ELIZABETH A. KELLEY, PHD
ILLUME Advising, LLC

AMANDA E. DWELLEY
ILLUME Advising, LLC

This case draws on work in the energy efficiency industry where many utilities rely on data-driven insights and decision-making to encourage consumers to adopt energy-saving products and behaviors. In this highly regulated industry, utility staff must show value through big data, and studies often rely exclusively on quantitative data analytics to create behavioral models to explain or predict behavior. However, purely data-driven research often fails to answer questions about why customers behave a certain way, and what product or program managers and marketers can do about it. In this case study, the team from ILLUME Advising LLC (ILLUME), a research consultancy in the clean energy industry, illustrates how their cross-functional team paired qualitative and quantitative research on residential home energy use. The case study draws on an exploratory market and segmentation study for an electric utility interested in engaging customers through a smart home product offering. The team used a mixed-methods, hybrid research approach, cycling between quantitaive and qualitative methods, and refining the project concept and hypotheses before each stage of research. This approach set the stage for market and consumer insights that showed opportunities beyond the clients' original concept.

INTRODUCTION

This case study explores how an energy efficiency research team used a hybrid approach that included qualitative and quantitative research practices to refine the product offering and go-to-market strategy for a large public utility interested in entering the smart home market. Operating in a regulated environment, the utility had a preliminary concept for a product offering, and the research team's brief was to conduct exploratory research for a concept that was already encumbered by numerous constraints. The team set the stage for a choice-based segmentation study to refine the product offering by socializing a series of market and consumer insights, including customer data mining, competitive market assessment, and in-home ethnography around energy and the smart home. In-home ethnography provided preliminary customer insights around the utility's energy app and customers' motivations and interest in managing different aspects of their home. This shaped a large-scale quantitative segmentation survey by ensuring that the concepts and language matched consumer cognition and expression.

This case study highlights the value of layered and iterative research, particularly in new markets and stakeholder environments constrained by pre-existing business processes, capabilities, or regulatory requirements. It also explores how qualitatively-grounded and experimentally rigorous quantitative research (in this case, choice-based segmentation) can be used to shape a product offering in addition to its more common use of defining the audience.

This case study presents:

1. The stakeholder and regulatory landscape, including unique considerations for developing a consumer product offering within a regulatory framework where products and services are required to demonstrate energy savings
2. An overview of the research, highlighting the layering of qualitative and quantitative methods that not only strengthened the research but socialized findings gradually, with each phase setting expectations for what the next might uncover or confirm
3. A discussion of the strategic recommendations and business outcomes, including trade-offs between addressing customer needs and meeting regulatory requirements in the utility context

PROJECT CONTEXT: GOING TO MARKET WITH A SMART HOME OFFERING THAT SAVES ENERGY

Stakeholder Landscape

The authors represent the ILLUME Advising team; ILLUME is a research consultancy in the energy industry. Our researchers work with electric and gas utilities to enhance their energy efficiency offerings, with the goal of reducing energy usage and demand to meet regulatory goals. Figure 1 introduces the other key characters in this story.

In Spring 2017, the team embarked on a marketing contract with a public utility partner to support their development of Home Energy Management (HEM) offerings in the smart home space. For this engagement, the research team partnered with (a) a marketing firm to develop communications and engagement materials for a "home energy information" app-based product, and (b) two specialized research partners to design a complex pricing and marketing segmentation experiment to inform the structure and positioning of a nascent smart home offering.

The client is a large public utility providing electric and gas service. Like many public utilities, they offer commercial and residential programs to help customers save energy and meet state regulator (top box of Figure 1) mandates to reduce the energy use of their customer base by a certain percentage each year. All utility energy efficiency programs are funded by ratepayers through a per-unit surcharge leveraged on energy bills. As such, utilities are required to demonstrate that any programs or offerings using ratepayer funding deliver energy savings (i.e., reduce consumption) cost-effectively.

The utility partnered with a vendor to introduce an energy analytics app that provides customers with trends and tips on their home energy use. Customers with the app can install hardware – an energy hub – that enables real-time updates to the app. This hardware can also serve as a smart home hub by facilitating communication with other devices, illustrated in Figure 2.

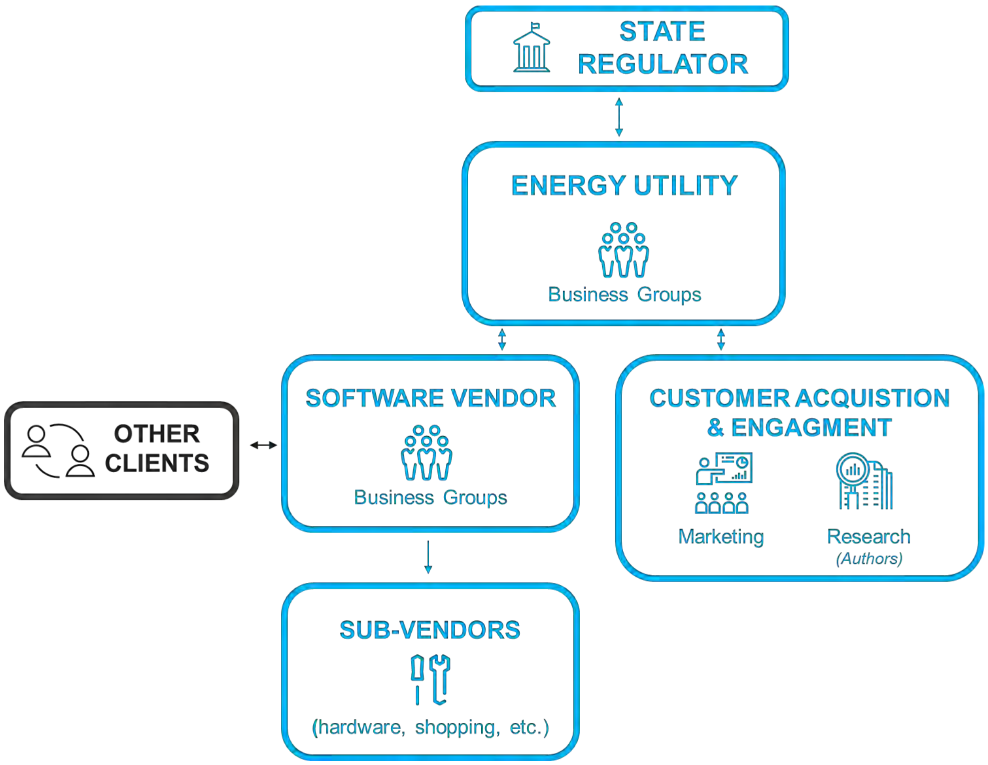


Figure 1. Stakeholder Relationships

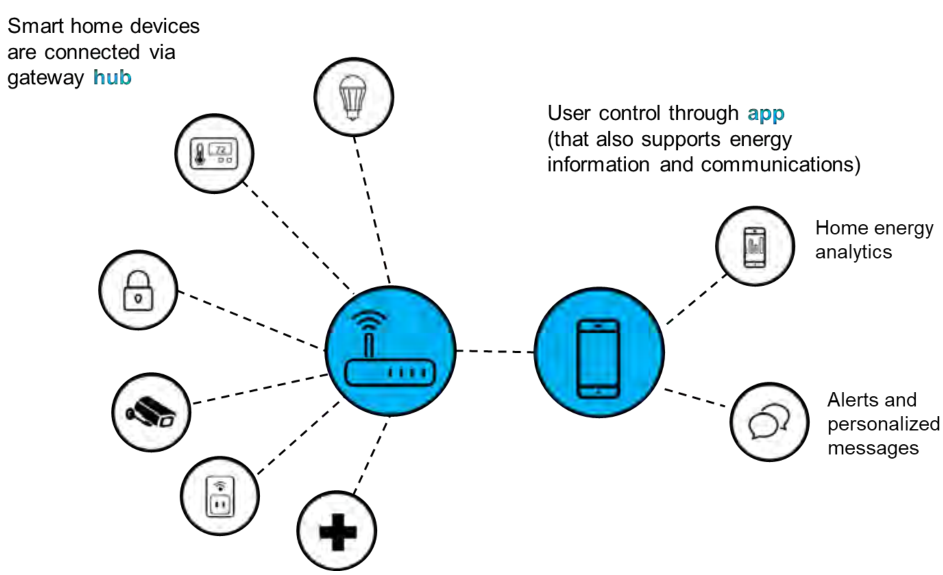


Figure 2. Concept for Smart Home with Embedded Home Energy Management

Initial Offering Concept

Seeing the growth in the smart home space, and the entry of mass-market players like Samsung, Amazon, and Apple, the utility was interested in testing their own offering. In addition to the energy analytics app, the utility offers rebates on smart thermostats and smart LED lighting. Knowing that energy analytics, thermostats, smart plugs, and lighting may not compel their customers to engage with the app (and reduce energy consumption), the utility wanted to develop a smart home offering that could pair with other, non-energy-saving devices.[1] By offering smart locks, sensors, cameras, and smoke detectors in addition to energy efficient products and services, they believed they could increase customer engagement and adoption. While the ultimate goal of the pilot, from a regulatory perspective, was energy savings, the utility was also interested in customer engagement.[2]

The initial offering could be tested as a pilot under looser regulatory and cost-effectiveness requirements, but, ultimately, it would need to deliver cost-effective energy savings before scaling. When the researchers joined the project, the utility energy efficiency group had positioned the smart home pilot as a HEM pilot, with the vision of using broader customer interest in the smart home to generate engagement with HEM (i.e., energy analytics and appliance management). The utility brought in the ILLUME team as they were transitioning to a new version of the energy information app to conduct the pilot.

In addition to the initial concept of using a broader range of smart home devices to encourage engagement with HEM, the software, energy hub hardware, and compatible devices were in development. The energy analytics software vendor had development-phase software that connected their app to a limited set of smart home devices. Nearly all the compatible devices were white label smart home products and the software was not compatible with better-known, mass-retailer brands. These compatible devices offered more control and reliability for the vendor and utility at a much lower cost. Additionally, the software was not yet compatible with a few categories of smart home security devices. While the utility and vendor were open to a long-term strategy that might reveal a different ideal product set, in the short-term, the offering was shaped by these capabilities.

The client had a concept for product pricing at the outset of the research engagement. They preferred a subscription-based service with monthly fees, largely driven by billing capabilities and pre-existing assumptions about customer preferences.

The Role of Customer Research

The utility was considering a range of product configurations (i.e., bundles of devices and features) and pricing strategies when the research team joined the project. These considerations were based on (a) prior customer insights from related energy efficiency offerings (e.g., the energy analytics app), (b) then-current capabilities of the utility and vendor such as online shopping and fulfillment functionality, and (c) business and regulatory objectives. Like many consumer product offerings, the model needed to meet customer expectations and, at minimum, cover its costs. Unique to the utility environment, the offering also needed to meet regulatory requirements for delivering cost-effective energy savings.

Ultimately, the utility wanted to know whether the entry into the smart home space could generate enough engagement in HEM to increase energy savings and recover costs.[3]

The client needed the ILLUME team to quantify and characterize their customers' interest in the smart home and their willingness to purchase those devices from a utility rather than other brands. The client also wanted to test what pricing structure might work best in the market, including options for recurring monthly fees, upfront payment, or a combination.

The utility requested market research that included a competitive market assessment and customer receptivity to, and willingness to pay for, potential offerings. This differed from standard product development research because the utility was enhancing a regulated offering – an app with real-time energy analytics – and needed to demonstrate energy savings.[4] The client was interested in market research for multiple reasons, including to:

- Validate their initial concept for a smart home offering
- Estimate potential market share of the offering (to inform a business model)
- Determine the appropriate pricing model and price level (to optimize customer acceptance while covering costs)
- Inform messaging and positioning of their offering (based on customer needs and language)

Findings from this research would be used to shape the product offering and go-to-market strategy within the business constraints listed above. In other words, even if the utility could not develop or market the ideal product for their customers, they needed to understand the available market for what they *could* offer.

RESEARCH OVERVIEW

ILLUME developed a staged approach for assessing market potential and determining the go-to-market strategy. This approach see-sawed between qualitative and quantitative research, using findings from each phase to socialize potential refinements to the product concept, and develop hypotheses to test in the subsequent phase. In this case, as in Anderson, Faulkner, Kleinman, and Sherman (2017), an iterative research process allowed the research team to socialize findings with the client team over several months.

Figure 3 illustrates how these research efforts came together. The approach was iterative and built on previous research by the ILLUME team and the utility. The team used qualitative ethnographic research in the discovery phase to (a) uncover and socialize customer needs and personas, (b) set expectations for what might emerge from quantitative research (e.g., validating or invalidating the initial concept), and (c) uncover customer language, preferences, and decision-making processes to strengthen survey research. An initial competitive assessment helped set expectations for what quantitative survey research might reveal in terms of customer preferences. After discovery-phase research and prior to survey research, the team presented customer and market insights to set the stage for the quantitative survey.

The following sections detail each stage, including a sampling of customer insights. Like many research findings, these results are specific to the geography, time frame, research context, and client context, and may not be generalizable to other utility service territories or to smart home customers in general.

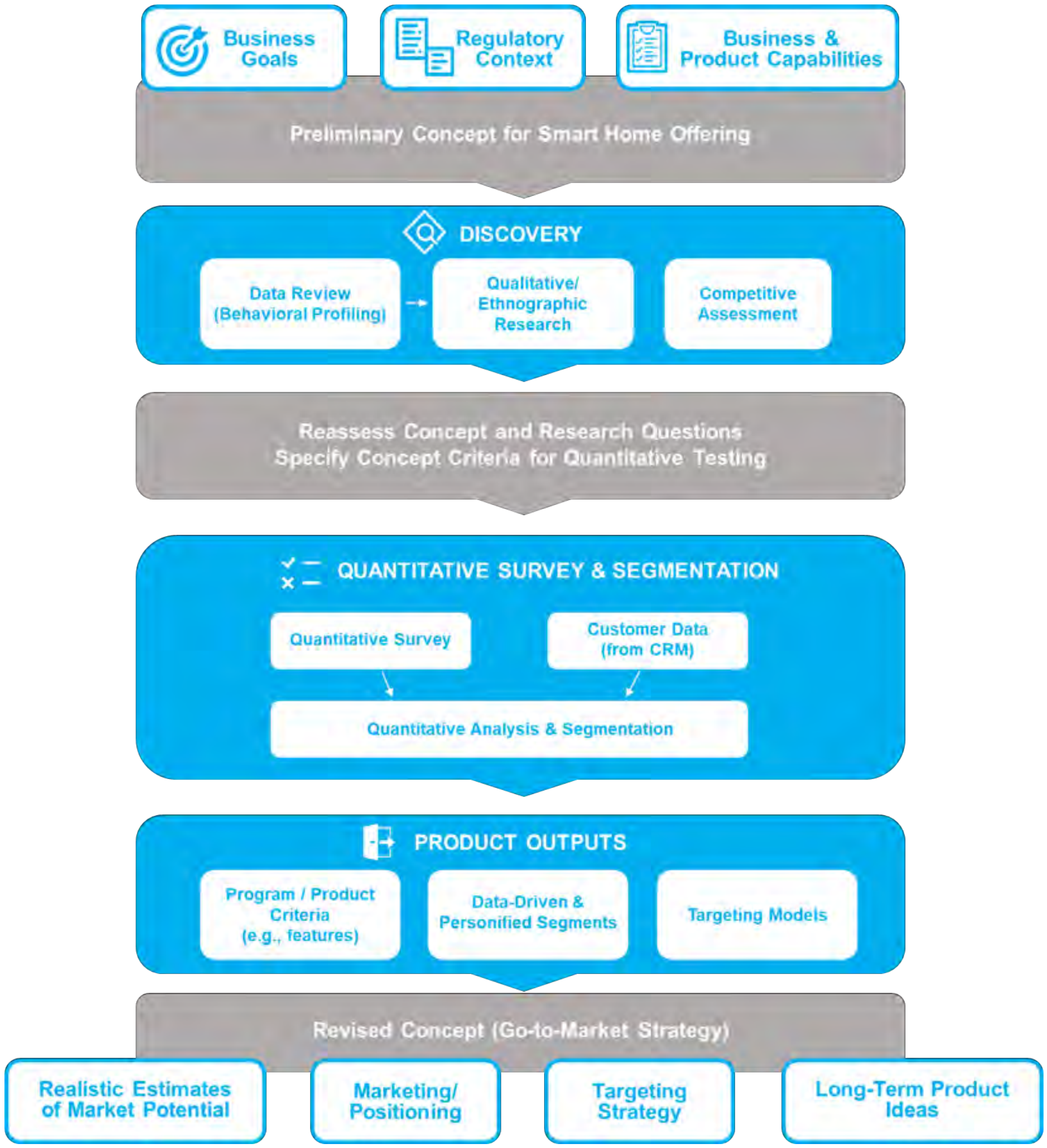


Figure 3. Research Process Flow

PREPARATORY RESEARCH

Competitive Market Assessment

At the time of this study (mid-2017), the smart home market was crowded with small and large product and service providers, including offerings from consumer products manufacturers; telecommunications and home security providers; online and brick-and-mortar retailers; and energy management start-ups. Business and pricing models varied widely, from pay-per-device packages with no recurring fees to subscription packages. Few consumer research or market studies were available to illuminate which of these offerings were most attractive to customers and where – or why – market shares were growing.

As part of the discovery-phase research, the team conducted a competitive review of smart home offerings. A key finding was equipment-based pricing: Most smart home products and services are sold as equipment with a one-time cost, and any associated apps or services are provided at no additional cost. Based on this research, only a subset of providers – primarily security and telecommunications companies – offered subscription-based models and these received mixed consumer reviews.

The prevalence of lighting and home security products in competitors' starter kits and smart home messaging suggested their importance as entry-points into the market. Marketing positioned smart lighting as convenient, easy, and fun, and home monitoring as a tool to keep your family and home safe. In the utility regulatory context, where energy savings are critical, the offering cannot focus on security, even if customers are most interested in these products. Monitoring and security devices do not deliver energy savings – hence, based on the competitive assessment, the research team and stakeholders began to see that a utility-sponsored smart home offering might need to find a different niche than mass market retailers were filling.

A secondary role of the market assessment was to inform the latent class discrete choice (LCDC) experiment embedded in the quantitative survey. As discussed below, the LCDC presented hypothetical smart home shopping decisions to customers, and asked them to make trade-offs between brands, devices, and services. As such, the LCDC experiment needed to represent the market as realistically as possible, including leading brands and common devices.

Customer Data Mining

Next, the research team performed quantitative analysis of consumer behavior around the energy savings app, and customer characteristics of segments identified as having (a) higher engagement, or (b) higher energy savings through the app. The client team and their external evaluators had previously identified several target segments from among the Mosaic® lifestyle segments, which cluster customers based on sociodemographic characteristics. The team leveraged rich customer data from the utility to characterize the target segments in terms of demographics, past utility program participation, communication and engagement with the utility, and the frequency and recency of app use.

ETHNOGRAPHY

The team built on the learnings from the market assessment and customer data mining with ethnographic interviews. The client team wanted to learn more about the people using the app and their experiences to feed into messaging campaigns. They also wanted to ensure that, as they brought a smart home package to market, it met customers' needs and desires.

To learn more about how the app fit into people's lives and their concerns, the team conducted 4 phone interviews and 15 ethnographic, in-home interviews. Interviews were conducted across the utility service territory, followed a semi-structured guide, and were audio recorded. During these interviews, the researchers observed people using the app on their phone or tablet while explaining their experience and how the app fit within other household concerns or priorities.

The sampling approach focused on customers pre-identified to belong to specific Mosaic segments that the client and their external evaluators had hypothesized were good targets for their app based on past program participation data.[5] The team attempted to recruit these segments, but in practice, the team found that these demographic segments were not always a good predictor of behavior or use of the app and observed some discrepancies between the segment demographics and actual participant demographics. These discrepancies led the team to abandon the Mosaic segments as an organizing structure for presenting research. The following section presents some key findings through the story of one of the interview participants.

At home with Melissa:

Two researchers, Liz and Allison, sat at the high-top table in Melissa's dining room. It was summertime, and Melissa's three kids, ranging in age from 2-12, wandered in and out of the room – the two-year-old joining them for much of the interview. As she discussed her approach to household maintenance, Melissa explained that she had had trouble paying her bills in the past, but now kept things in order using a binder system for recurring bills. When a utility bill arrived, she would divide it by four and would send a check out each week. Melissa's affect around these three-ring binders was joyful. When the team asked her about changes or repairs she was thinking of making to the home, she immediately retrieved another binder. This one was for her kitchen renovation and contained her mood-board-style assortment of images of dream kitchens, along with print-offs from websites with specific items and their costs. Her current kitchen was original to the house and had a heavy cast-iron sink and a linoleum floor that was peeling at the edges. She explained that she had chosen everything for her new kitchen already and was little-by-little making the necessary purchases, using the binder to keep track. She had most of the floor tiles already in the basement, along with the new cabinets and countertops. She was waiting on the sink – once she was able to buy it – after enlisting some help to carry out the old sink, she was going to complete the renovations.

When the conversation turned to her experience with the utility's energy savings app, her face lit up with delight – she had been using the app for several months and had recently installed the hub showing her usage in real time. When they moved into the house, she explained, the first bill was $400 or $500 – well beyond what she could afford. She made it her goal to bring the bill down. Maintaining a close watch on the bill was particularly important to her because she had once had her power shut off for being in arrears. She never wanted to be in that situation again, which is why she had started paying her bills in weekly installments. She also had taken steps, such as unplugging devices and appliances when not in use, to lower her bill. She explained that every autumn she would set aside a day to seal the windows with plastic and to weather-strip the doors for the winter; she showed us on her calendar when she planned to do that this year.

As concerned as she was about bringing down her bill, she was more concerned about her family's safety. Although she knew that unplugging devices and chargers when not in use could lower her bill, she did not like her children touching the plugs for fear of an electric shock. Instead, after her kids were in bed, she would go around the house unplugging any of their devices to stop electricity "vampires" from driving up the bill. Concern for her children's safety motivated other changes as well: she and her husband removed a ceiling fan over their son's bed that she worried would fall in the night.

In discussing the app, she liked that she could see how much energy they were using at any given time, but she wanted to see more details about their use. She was also interested in sensors or other features that would help her understand what was going on at home when she was at work.

Key Themes from Ethnography

The team's interaction with Melissa provided several key insights around the app and the potential market for smart home products. The "thicker" data that emerged from the ethnography added nuance and individual stories to the data projections based on Mosaic segments or utility data analytics (see Wang 2013). For example, traditional propensity models would likely have excluded Melissa from a target group given her preference for paper bills and paying by check instead of online. As the team learned, this analog system provided Melissa with control over her bills and a real-time knowledge of what she owed. Melissa supplemented her binders with information from tracking applications such as the utility energy app. Often considerations of self-tracking behaviors, such as considerations of the Quantified Self movement, presume biosensors and other technical apparatuses (see for example, Nafus 2016). Melissa was interested in closely monitoring her household expenditures and usage; she did so through a mix of technologically-enabled tracking apps and her paper binders.

A general theme that emerged during the ethnography was around the value the app provided its users in bill control and visibility. This value of control was not uniform. For some people, it was linked to a confidence around their ability to pay; for others, it had more to do with managing their home remotely (for example, someone who traveled for work enjoyed the ability to ensure his home was using minimal electricity while he was away).

A desire for control and a granular knowledge of what was happening in the home was similarly important in the context of what kinds of features and smart home devices people had installed or were hoping to install. The team found that security – both financial and physical –was a key interest and driver. Many of the participants were looking to install smart door locks, the video-camera doorbell, or other security cameras, along with smart smoke detectors or carbon monoxide (CO) sensors. These products enabled greater control and awareness, suggesting a substantial market for a range of security features, including sensors, smart locks, and cameras. Importantly, the motivation to install these security features was much more often about understanding what was happening in the home than protection of the home. Where safety was a concern, it was more often expressed in terms of family safety (did my children come home from school?) than in terms of fear of crime or intrusion.

The team also spoke with several individuals who had already installed smart home features. While these customers saw the value in the app, several noted that the current functionality and user experience was not on par with larger brand names. These individuals had also already installed many of the features that might be offered in the smart home pilot – from smart thermostats, to lights, to open-close sensors on windows and doors. Therefore, the team hypothesized that these early adopters were not an ideal market.

Ethnography: Implications and Next Steps

The ILLUME team shared with the client that their customers were looking for home security products alongside other smart home products. Although security features had emerged as an area of customer interest in the comparative market research, the utility initially saw this as a supply rather than demand issue. Just because other companies were offering these products did not necessarily mean their customers were likely to purchase them.

After the team presented the qualitative research findings, highlighting the ways in which security and control were so often linked, the client team agreed to consider security features in the next research phase. Some of these features were not compatible with the current version of the hub, and security devices do not have an obvious energy-saving application. However, the team agreed it was important to test them in the LCDC survey to get an accurate image of what smart home features people are interested in purchasing.

LATENT CLASS DISCRETE CHOICE (LCDC) SURVEY

The team conducted an LCDC survey with the objective of (a) segmenting the utility's customers based on their smart home purchase preferences, and (b) understanding overall preferences for products, features, and pricing models. LCDC is a type of stated preference trade-off analysis that combines elements of latent class analysis and discrete choice analysis into a single framework. ILLUME worked with two research partners to design the experiment and analyze results. Members of this combined research team conducted an LCDC experiment for LED bulb preferences in California when LEDs were an emerging technology. The methodology was based on that study and the report, which contains a detailed methodology (Opinion Dynamics and StatsWizard 2012).

Survey Approach

ILLUME deployed the LCDC experiment to utility customers as part of a smart home survey containing attitudinal, behavioral, and demographic questions, in addition to the experiment. The team used results from 1,047 completed surveys for segmentation, segment profiling, and marketing insights.

The survey presented each customer with nine hypothetical shopping scenarios (stores) where they selected between smart home packages with different attributes, shown in Table 1. The packages contained different brands and quantities of smart home devices, including home security, monitoring, lighting, and thermostats, offered at different pricing levels with different support options.

Table 1. Attributes Tested in Embedded LCDC Experiment

Brands	Six total brands representing different product/service providers
Products	• Thermostat: Smart thermostat, learning thermostat, or no thermostat • Lighting: Dimmable smart bulbs, white smart bulbs, or no bulbs • Smart plugs (or none)[i] • Smart light switches • Sensors: Open/close sensors, motion sensors, smart door locks, a "full package", or no sensors • Cameras: Smart camera, doorbell camera, or no camera • Other: Smart smoke/CO sensor, water leak sensor, or none
Features	• Home energy information: Basic home energy feedback or advanced feedback (real-time, with tips, targets, and alerts; appliance "health" monitoring), or none • Voice control (via Home Assistant) • Care plan
Pricing	• Pay upfront or spread costs over 12 months (total price determined based on market prices, plus a +/- 10% discount or premium for experimental purposes) • Monthly service fees ranging from $0-$19.99 (in increments)

The goal of discrete choice experiments is to represent the complexity of consumer decision-making (Boomer 2014). Given the many emerging products and features in the smart home space, the team chose an LCDC experiment to uncover customer preferences by allowing them to make trade-offs between different hypothetical product sets. In contrast with "direct" stated preference questions, this method presents customers with "all-in-one" product and feature bundles where they may like one element of one package, and another element of another package, but must choose one or none based on what features matter most.[7]

Each respondent saw nine experimentally-designed stores and selected one of four smart home packages or "None of These". The team tested two "blocks" of stores in two sub-groups of the sample to ensure that the full range of trade-offs was tested. Figure 4 shows just one example of a store from the survey. Before customers began the experiment, they read instructions about the shopping scenarios and saw an example store. The team elected to represent the brands and attributes in text rather than images to minimize visual emphasis on any one feature and allow respondents to use or develop their own mental models of each feature (Hurtubia et al 2015).

Building on Qualitative Findings

The findings of the qualitative research – including nuances around how customers described their home and expressed interest in smart home features – informed the survey design. This approach required alignment with customer language to produce valid results. The in-home interviews were critical to understand how consumers perceive their homes, how they think about the smart and connected home market, and what features they are interested in. Given customers' interest in security features, the team sought to represent a range of security and monitoring devices, even if they were not compatible with the client's

	1	2	3	4	5
	Provider 1	Provider 2	Provider 3	Provider 4	**None of these**
	PRODUCTSINCLUDED				
	LEARNING Smart thermostat	Smart thermostat (does not learn behaviors)			
	8 dimmable Smart bulbs (white light only)	2 dimmable Smart bulbs (white light only)	8 dimmable Smart bulbs (white light only)	2 dimmable Smart bulbs (color changing)	
		2 Smart plugs			
			2 Smart light switches	2 Smart light switches	
		FULL Smart security package (door locks, motion sensors & open/close sensors)	Smart door locks		
	Smart camera (outdoor or indoor)			Smart camera and Smart doorbell camera	
	Smart water-leak sensor		Smart smoke/CO2 detector	Smart smoke/CO2 detector	
	HOME ENERGYINFORMATION				
		Advanced Energy Feedback*	Energy Feedback		
Notes:	**Energy Feedback: Hourly energy use and energy costs, tips, targets and tracking via smart phone**				
	**Advanced Energy Feedback: REAL-TIME energy use & energy costs, tips, targets and tracking, appliance use and appliance "health" monitoring via smart phone*				
	VOICECONTROLS				
	Voice controlled		Voice controlled		
	Voice control: Controlled by Amazon Alexa, Google Home or similar Smart Assistant (sold separately)				
	SUPPORTAND PRICING				
	Care plan: Help with installation, configuration, and tech support				
		Care plan		Care plan	
Service fee and device cost:	$9.99/month	None	$9.99/month	$4.99/month	
	$51/month for 1 year	$32/month for 1 year	Up-front price of $305	$41/month for 1 year	

Figure 4. LCDC Survey Store Example

offering. In real-life scenarios, customers might be weighing options with these devices. The team also spoke with several people who mentioned they used color-changing smart bulbs in their home for ambience, for fun, to simulate morning or night, and to signal different things to their children (e.g., reading time). From these lighting stories, the team hypothesized that lights may be an important entry point into the smart home, and color-changing bulbs (not then-compatible with the client's offering) may have experiential value beyond controllable standard white bulbs.

The qualitative interviews and market assessment also provided guidance about what language to use to describe technical features. For example, to represent the smart home packages' compatibility with Home Assistants, the team used the term "Voice Control", as the term "home assistant" did not seem commonly-used. The interviews and market assessment also showed that HEM and home feedback are not typical components of smart home offerings, and that the term "Home Energy Management" does not resonate with customers. Energy management is typically a stand-alone product offered by specialized companies or utilities. Through interviews, the team learned that many customers did not have pre-existing concepts of what "home energy information" or "home energy management" was; therefore, the experiment provided extra explanations of this feature.

Analytic Results

Analysis of the Stated Preference Survey provided two results: (a) A rank-ordered set of feature/pricing preferences among all customers and by segment, showing the biggest drivers, and (b) smart home customer segmentation for the utility, organizing customers into five distinct groups with different product/feature preferences, affinity for the utility, and likelihood to purchase.[8] The segmentation was a means to understand overall feature/pricing preferences, as well as characterize the utility's best target.

Overall Results and Preferences

The ILLUME team used the LCDC survey to uncover smart home and HEM preferences specific to the product concepts the utility was considering. The first analysis step was to identify customer segments based on their preferences; overall preferences were derived after segmentation by calculating weighted average preferences across segments. Figure 5 shows the results across all segments. Overall, brand was the leading driver of decisions. Those who preferred to purchase through the utility perceived it to be reliable, accessible and convenient, because it was a local company that had maintained many long-term relationships.

Smart thermostats were not a strong driver of package selection, nor was home energy analytics, one of the key differentiators of the clients' offering. While smart lights were popular, and are energy-efficient LEDs, the quantities offered in smart home kits are typically not large enough to drive energy savings. Furthermore, emerging research on whether controlling lighting through a smart home platform actually saves energy is mixed (Efficiency Vermont 2016). The relatively low rankings of potentially energy-saving devices highlights the challenge of developing an energy-saving smart home offering and finding the niche of customers most likely to use HEM features. Can the utility attract enough customers with an offering differentiated by HEM, or should they offer something more similar to competitors and hope that a sufficiently large slice of customers begin using the energy-saving features?

In the survey, the team included open-ended questions asking customers what they see as the benefits or drawbacks of purchasing smart home packages through their utility. Responses revealed that the client has a strong reputation for service and reliability, which could be a driver for choosing the utility for a smart home package – despite their lack of a track record or reputation in the space:

- Reliability and trust: "I've been with [UTILITY] for years and it's a trusted and known Company"; "Reliable company that can come out and provide service if needed"
- Expectations: "I would expect the package to work more seamlessly from [UTILITY]."; "Local company for faster service, billable via my energy bill."

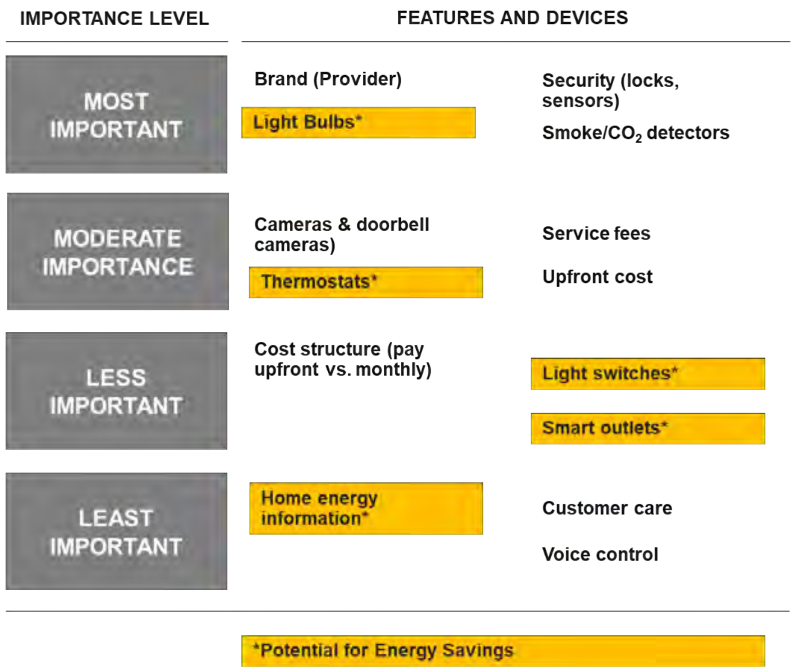


Figure 5. Relative Preferences for Features and Devices from the LCDC Experiment. The importance levels speak to whether differences in a feature or device between packages drove overall preferences for that package. If a feature or device appears as "unimportant" in this rank-ordering, it does not mean it is unimportant in general. Customers may take the feature for granted (e.g., voice control) such that they were not concerned about differences. Additional data points and customer insights are needed to understand why an item ranked lower. Note that the exact order of customer preferences has been obscured to protect client confidentiality.

In return for placing their trust in the utility, customers expect to be able to call easily for help. The value of strong customer service is meaningful in the smart home space, as the earlier ethnographic research revealed several reliability and usability issues and complaints with the utility's then-current app and hardware solution.

Negative customer comments about purchasing through the utility also highlighted skepticism about the utility's ability to produce a viable offering in this space: "Does [UTILITY] have sufficient experience and expertise?"; "To me UTILITY does not have a track record or earned reputation as a retail service like Amazon." This highlights the challenges of offering a product suite through a utility, where the individual products may not carry the brand reputation of market leaders.

Finally, the experiment showed a preference for bundles (e.g., a starter kit) rather than a la carte offerings, particularly if offered by a trusted brand. The research team's interpretation was that smart home technology is relatively new and perceived as complex. Therefore, customers are looking for sets that they can be confident work well together and use brand as a shortcut for quality. This finding also aligns with decision-making theory around cognitive burden and decision fatigue – the value of curating consumer choices to facilitate making any decision and improving satisfaction with decisions (See Iyengar and Lepper 2001). This finding is particularly important for marketing and communications considering the segment findings that highlighted that the "mid-market" customers may be the best target. These customers are eager to make their home a smart home but may not be as comfortable with compatibility and set up.

Segmentation Results

Throughout both the qualitative and quantitative research, the ILLUME team worked with the Mosaic segments the utility client already purchased. In the qualitative research, the Mosaic segments did not map easily or reliably onto observed or reported behaviors or preferences. As Flynn, Lovejoy, Seigel, and Dray (2009) and Cuciurean-Zapan (2014) have suggested, when customer segmentations are divorced from behavioral inputs or become overarching, their usefulness is frequently curtailed (even when their application may not be). Given this limitation, the team recognized the opportunity to use layered research to not only understand customer preferences, but to identify a potential market on the basis of behaviors alongside stated preferences. The team used relative preferences from the LCDC experiment (e.g., more interest in security vs. energy management features; importance of brand vs. pricing) and covariates (other survey questions) to develop five cluster-based customer segments.

A valuable output of the segmentation research was confirmation that the utility's core target for their HEM offering was not early adopters. The ethnography team interviewed a few of these tech-adept customers. They had invested considerable effort into installing and networking an array of emerging smart home products and apps, often overcoming compatibility and connectivity issues. The profiles of this highly-engaged segment presented a tempting target for the client. At that time, the team discussed some of the potential drawbacks, including the segment's already-formed brand preferences and expectations, particularly for name-brand devices that were not compatible with the client's offering. The survey segmentation identified these "technophiles", who seem attractive due to their smart home purchase propensity and acceptance of service fees. However, they are also price-sensitive, have high expectations for products, compatibility, and service (prefer buying bundles), and perceive the utility as an unknown in this product space, suggesting that many may choose other providers.

Instead, two mid-market segments emerged as promising targets for the utility's offering, shown in Figure 6. Target Segment B: with moderate income, middle age range, and kids at home, and who were already buying smart home equipment at a solid rate, and Target Segment A: older, interested and engaged in smart home, in particular, easier-to-understand security products like smart locks and smoke detectors, and a high comparative level of comfort with service fees. This was unique among the segments, making them an attractive target for the utility's offering, which may include a service fee.

The quantitative data allowed the team to characterize these segments not just in terms of their relative preference, but in their distance between each other (i.e., degree of difference), and relative size. The relative importance of brand, price, and smart home features allowed the team to position the five segments in terms of their likelihood to adopt, and their relative preference for the utility as the service provider. The team selected these two dimensions to summarize the segments as the question of whether they would purchase

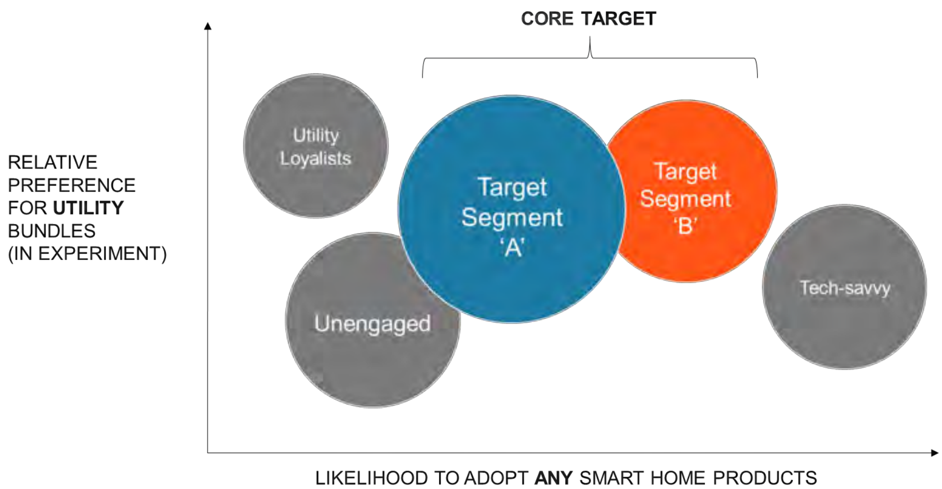


Figure 6. Example Segment Positioning: Likelihood to Adopt Smart Home vs. Relative Preference for Utility

from their utility is an overarching consideration given the growing strength of mass-market players.

Bringing Segments to Life

The research team examined each segment's preferences for the smart home devices, features, pricing, and service options, and developed profiles of their relative preferences. While these segments were defined by the relative importance of smart home features, they could be described by the contextual data gathered in the survey. For example, their attitudes toward the utility, self-report of existing smart home equipment, early adopter profile, and demographics. Further, the team used relational and behavioral customer data from the utility, such as past participation in energy efficiency programs, online account access, and online billing preferences. Using these three layers of data – experimentally-determined preferences; self-reported attitudes, behaviors, and demographics; and utility relational data – the team created a rich profile of the five segments, illustrated in Figure 7.

In Figure 7, the customer characteristics and their smart home and device preference provide a starting point for the product development and marketing team to refine their positioning. For example, highlighting capabilities rather than brand differentiation may be more relevant to a segment that typically prefers packages, is not tech-savvy, and trusts the utility. Emphasis on "why buy from us" may be less important than showing how this technology might fit into their family life. As is frequently the case in more quantitative segmentation, decontextualized demographic details do not always provide adequate understanding of individual concerns or drivers. Moreover, while knowing that individuals in this segment are most interested in smoke/CO_2 sensors and security packages might allow marketers to create their own story of a segment members' life and concerns, that story may not align with actual individual's experiences in that region, market, or segment.

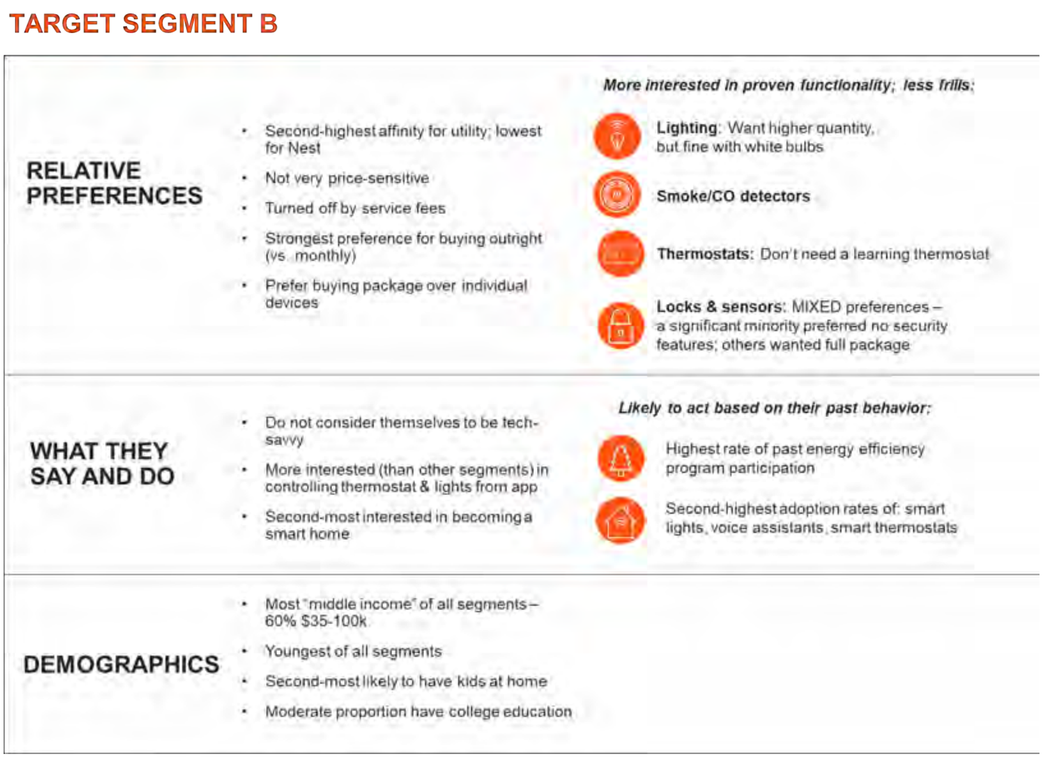


Figure 7. Data-driven Segment Findings – Target Segment B Example (from survey and segmentation analysis only)

To make sure product and marketing staff interpreting findings could envision a living, breathing customer, the team coupled the generalized, data-driven profiles with examples from ethnography. For instance, in the team's conversations with Melissa, she described her concerns around her children and electrical outlets, a safety dimension not addressed by the smart home devices tested. Her interest in smart home devices to monitor the house, like open/close sensors, was not focused on neighborhood safety but her children's physical safety when interacting with their home. Integrating details such as these into the data-driven segment findings allowed for a more nuanced and robust segment depiction.

Recognizing that safety and security may manifest in multiple ways, the team used individual stories to disrupt assumptions about a uniform interest in security, and to inspire consideration of other angles for positioning the same set of products. For instance, highlighting family security within the home in contrast to a neighborhood or crime-oriented approach to security.[8]

Estimating the Market Opportunity

Based on the two target segments, the team developed business models for the utility offering. The models leveraged a market simulator tool that calculated likely market shares of a utility-sponsored smart home package given a specific set of devices, features, and pricing vs. competitors – i.e., entering specific configurations comprised of individual elements tested in the LCDC. The team used the market simulator tool to estimate relative market share of realistic utility packages compared with other providers' most heavily-promoted starter kits. Because the LCDC experiment only represented a sample of competitors, and the survey respondents likely had a higher affinity for the utility than the general market by virtue of completing a somewhat onerous utility survey, the analysts developed adjustments for the utility's likely market share using secondary data. Based on this analysis, and assumptions about the product offering and pricing, the team concluded that the utility had set ambitious first-year participation goals. The utility plans to re-visit the market sizing model soon after the start of the pilot, once enrollment begins.

BUSINESS OUTCOMES

The team presented several strategic findings with the goal of providing concrete feedback to shape the product offering and go-to-market strategy and quantifying the market size and opportunity for the utility. These included:

1. *Mid-market target:* The team suggested a focus on mid-market adopters with higher utility affinity as the base market, rather than the most affluent early adopters. This finding emerged out of the ethnographic research, where the team had observed that many of these more tech-adept individuals had already installed smart home features offered by competitors. This finding was validated in the LCDC where the tech-savvy segment was found to be highly likely to adopt smart home features but unlikely to do so using a utility-branded package. Mid-market adopters in the target segments are identifiable through supervised machine learning (e.g., random forest) that leverage LCDC data, and the research team developed propensity scores to use in recruitment campaigns.
2. *Security and lighting as entry-points:* The team recommended using security features and messaging as an entry-point with inspirational and capabilities-driven illustrations that evoke feelings of family and home security rather than a products-first message. Similarly, the team recommended that the offering contain and promote smart lighting as a gateway to the smart home, focusing on curating the home's atmosphere.
3. *Product bundling:* The team suggested that the product offering include an option to purchase products and features in bundles in addition to a la carte to lighten the decision-making burden. This was particularly helpful to reach less tech-savvy customers and align with competitors' starter kits. Central to this objective is developing education and marketing that focuses on what you can do with a smart home rather than individual products.

While this paper was drafted, the pilot was nearing launch. The research team worked with other stakeholder groups including the product vendor to fine-tune the pilot structure. In parallel, the marketing team drew on the research findings to develop educational and recruitment emails, and online and in-app content on smart home capabilities. The team's research – both quantitative and qualitative – shaped the marketing messages, highlighting what products can do in the home, specifically for security and lighting. The team was previously focused on messaging around smart thermostats (a favorite of energy efficiency programs), and the team decided to de-emphasize thermostats based on (a) overall attribute importance in the LCDC, (b) the ethnographic research showing lighting and security as potential entry points, and (c) relatively low penetration of thermostats compared with other smart home devices. Still, the offering must deliver energy savings, and the team is counting on HEM analytics and messaging, additional thermostat savings from home automation, and LED lighting upgrades to generate savings.

Some findings from customer research cannot be implemented in the short-term. These include offering devices and features as bundles, which the current online shopping application does not support, and an upfront pricing structure rather than a fee-based model. As mentioned in the Initial Offering Concept section, institutional precedent for monthly fee-based pricing established by IT and billing processes as well as the pricing structure in place for the energy analytics hub makes it difficult for the utility to offer upfront pricing in the short-term.

These trade-offs between customer needs and preferences, business constraints, and regulatory objectives highlight the challenges of conducting product development research in a regulated environment. In this case, some of the devices and features customers value most do not save energy (a regulatory requirement). Therefore, the utility is faced with the decision of whether to develop an energy-focused niche offering that may not attract many customers or an offering with more mass market appeal and an embedded HEM offering and hope customer engagement and education encourages HEM adoption and use to save energy within the general population.

CONCLUSIONS

This case study tells two stories. The first is about conducting iterative and adaptive research that layers qualitative, quantitative, and secondary research. An iterative and mixed-methods approach was needed in part because of the product space (an emerging market), and because of the stakeholder and regulatory context (the timing of the study within the product/concept development timeline, and the regulatory requirement to save energy).

The second story focuses on the stakeholder and regulatory context: the challenge of creating a product offering in a constrained regulatory context, layered onto a stakeholder environment in which product and service concepts had become entrenched due to a combination of (a) assumptions about customers and their needs from prior research, and (b) assumptions about organizational capabilities that might limit innovation. The two stories are intertwined – the research approach was developed to respond to the interests, assumptions, and constraints of the stakeholder team within the energy efficiency regulatory context. Through an iterative approach, the research team was able to share challenging insights gradually, creating multiple opportunities for communication.

The research toggled between identifying customer interests and motivations from an exploratory perspective, and testing concepts or assumptions constrained by the utility or vendor's existing capabilities, and regulatory guidelines. When the research suggested that the products and features valued by most customers were not well-aligned with saving energy, the team had to find pathways to a viable market strategy. The team used customer insights to adapt to these constraints, such as (a) finding pockets of customers who may be more interested in a utility-sponsored offering, (b) recommending bundles where energy-saving features can be combined with products that have more customer appeal, like home security sensors, monitoring, and color-changing light bulbs (despite the lack of energy savings), and (c) identifying ways to message monthly service fees that align with customers' brand perceptions of reliable and available utility service.

Amanda Dwelley is Director of Quantitative Research at ILLUME. She leads data mining, predictive modeling, segmentation, and evaluation efforts for energy efficiency programs, including behavior-based and Smart Home programs. She champions mixed methods to quantify "what happened?", understand customer engagement and help improve customer experiences. *amanda@illumeadvising.com*

Elizabeth Kelley is Director of Qualitative Research at ILLUME. As a sociocultural and linguistic anthropologist (Ph.D. University of California, Berkeley), she leads ethnographic and qualitative research studies that explore how people (in their homes or businesses) think about, use, and rely on energy. *liz@illumeadvising.com.*

NOTES

Acknowledgments – The authors would like to thank the ILLUME team (Anne Dougherty, Jason Turner, Jon Koliner, Sharon Talbott, Allison Musvosvi, Shefije Miftari, Jes Rivas, Robert Jamie, and Erin LaVoie), as well as our research partners Jennifer Mitchell-Jackson and Mary Sutter (of Grounded Research LLC), and George Boomer (StatWizards LLC), for their instrumental contributions to the research process and the development of this paper. The team would also like to thank all the individuals who participated in the research, particularly those who briefly shared their homes with us.

1. "Smart" thermostats are those that are WiFi-enabled and programmable, allowing people to change the temperature of their home remotely through their phone. "Learning" thermostats are similar, with the added feature of automatically adjusting to the homes' occupancy norms and comfort preferences without resident input. Smart plugs allow users to turn on or off any appliance plugged into a given outlet.

2. Many of the research team's utility clients are trying to increase brand relevance and customer engagement. This effort is intended to keep lines of communication open for new and emerging business models and maintain strong reputations within their communities to help with regulatory cases and discussion. For the past few decades, many public utilities have come to use their refined energy efficiency offerings as a customer communications and engagement platform. However, not all customers engage with utility-sponsored energy efficiency offerings in their current form. With more consumer products focusing on energy (e.g., smart thermostats and lighting), and non-utility energy service companies expanding offerings into the utility space (e.g., solar and electric vehicles), some utilities are looking for more avenues of customer engagement. The smart home is one of many market opportunities that utilities are eyeing to strengthen their relationship with customers. this!

3. Aside from fulfilling state requirements, there are independent business rationales for utilities to support efficiency programs that result in energy savings as this lightens the demand on the electric grid and may enable them to delay building new plants or other costly infrastructure. Similarly, many utilities are exploring new revenue models with the recognition that alternative energy sources (e.g., renewables), non-utility service providers, and smart/connected products are disrupting the relationship between utilities and their customers.

4. A regulated offering refers to a program with energy savings goals that is regulated by the state (through a Public Utilities Commission) and where energy savings are evaluated by a third-party evaluation team. In the early phases of this research, the ILLUME team also worked with another external partner, the third-party evaluator, who had been conducting research on customer satisfaction with the app, usage behavior, and subsequent energy savings.

5. As an energy-efficiency program, the app had been evaluated by a third-party company. Their analysis used the Mosaic segments to understand patterns in usage and associated savings. The segments with the greatest savings were identified as "target" segments for the research that the ILLUME team conducted.

6. In the emerging smart home market, crowded with large and small, and well-known and lesser-known providers, a fully representative experiment would have 20+ brands or providers. To derive meaningful results from the experiment, the team chose to limit the number of brands to a representative set that included one well-known player from different entry points of the market. This allowed the experiment to test the product and pricing features, at varying levels, to identify the ideal service offering and pricing. One limitation this presents in analysis is potentially inflating each brand's market share, since not all brands were represented. However, the team felt this trade-off was warranted to get a better read on general consumer preferences and price sensitivity.

7. The team determined segments by an expectation-maximization (EM) algorithm that calculates the probability of membership in each segment. The methodology was developed by Jay Magidson and Jeroen Vermunt (2004). A detailed description of the approach, as applied to a similar LED lighting study, can be found in Opinion Dynamics (2012). The analyst establishes the number of segments to mode and starts with random guesses for the segment (latent class) membership probabilities for each respondent. For each latent class, the analyst develops a conditional logit model and obtains a maximum-likelihood estimate of the segment-level parameters. As the preference data from the experiment (and covariates) is added to the logit models, the software computes new segment-membership probabilities, and customers are re-arranged into more similar segments based on these probabilities. The modeling follows an iterative process of re-estimating the logit models until the improvement in likelihood is marginal (maximum likelihood).

8. In previous research efforts, The ILLUME team has conducted segmentation research where the data-driven segments are derived and defined prior to qualitative research, then segment members are identified in the customer database, and recruited for interviews. This approach can be effective in filling in specific insights and nuance that the team needs for the project. However, waiting until after the quantitative research to conduct qualitative research has trade-offs, in particular, losing the opportunity to use qualitative research to inform the survey development.

REFERENCES CITED

Anderson, Faulkner, Kleinman, Sherman.
2017 Creating a Creators' Market: How Ethnography Gave Intel a New Perspective on Content Creators. 2017 *Ethnography in Praxis Conference Proceedings*. 425-443

Boomer, George.
2014 "Building Choice Models that Forecast Well." StatWizards LLC.

Cuciurean-Zapan, Marta
2014 Consulting Against Culture: A Politicized Approach to Segmentation. *Epic 2014*. P. 133-146

Flynn, Lovejoy, Seigel, and Dray
2009 Name that Segment: Questioning the Unquestioned Authority of Numbers. *Epic 2009*. P. 81-91.

Hurtubia, Ricardo, Guevara, Angelo, and Donoso, Pedro.
2015 "Using Images to Measure Qualitative Attributes of Public Spaces through SP Surveys." *Transportation Research Procedia* 11 (December 29, 2015): 460-74. doi:https://doi.org/10.1016/j.trpro.2015.12.038.

Iyengar, Sheena & Lepper, Mark.
2001. When Choice is Demotivating: Can One Desire Too Much of a Good Thing?. *Journal of personality and social psychology*. 79. 995-1006. 10.1037/0022-3514.79.6.995.

Magidson, Jay & Jeroen Vermunt.
2004 "Latent Class Models." Statistical Innovations LLC.

Nafus, Dawn (ed)
2016 Quantified: Biosensing Technologies in Everyday Life. Cambridge, MA and London, England: The MIT Press.

Opinion Dynamics and StatWizards LLC
2012 The Southern California Edison (SCE) Advanced Light Emitting Diode (LED) Ambient Lighting Program Customer Preference and Market Pricing Trial. Prepared for Southern California Edison, December 2012.

Wang, Tricia
2016 "Why Big Data Needs Thick Data." *Medium* website. January 20. Accessed [October 01, 2018]. https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7.

*Re*Humanizing Hospital Satisfaction Data: Text Analysis, the Lifeworld, and Contesting Stakeholders' Beliefs in Evidence

JULIA WIGNALL, MA
Seattle Children's Hospital

DWIGHT BARRY, PhD
Seattle Children's Hospital

Declining clinician engagement, increasing rates of burnout, and stagnant patient and family experience scores have led hospital leadership at Seattle Children's Hospital to submit requests to a data scientist and an anthropologist to identify key themes of survey comments and provide recommendations to improve experience and satisfaction. This study explored ways of understanding satisfaction as well as analytic approaches to textual data, and found that various modes of evidence, while seemingly ideal to leaders, are hard pressed to meet their expectations. Examining satisfaction survey comments via text mining, content analysis, and ethnographic investigation uncovered several specific challenges to stakeholder requests for actionable insights. Despite its hype, text mining struggled to identify actionable themes, accurate sentiment, or group distinctions that are readily identified by both content analysis and end users, while more insightful ethnographic results were sometimes discounted for lack of quantitative results or perceived implementation difficulty. Unfortunately, institutional contexts and preferences for specific types of data can lead to unnecessary requests and wasted efforts. Through including the subjective lifeworld, how an individual's lived experiences impact interactions with others, the authors were able re-humanize satisfaction. Cross-discipline collaboration can enhance the quality, validation, and advocacy of evidence from both qualitative and quantitative data. Co-developing a "Return on Method" (ROM) of satisfaction data can help improve analytic requests and expectations by end users. Ultimately, a lifeworld-informed combination of data science and ethnography can provide contextual and culturally situated insights that are both meaningful and actionable.

INTRODUCTION

Patient experience—in both quality of care and satisfaction—is one of the major improvement goals across every dimension and domain of healthcare in the United States, from neighborhood community health centers to major, internationally-known hospitals (Institute of Medicine 2001). Driven by changing regulations and increasing competition, the healthcare industry believes that more satisfied patients will lead to more revenue, suggesting that a service-oriented model to attract and retain patients will be required to remain competitive. The Affordable Care Act already provides strong financial incentives to improve satisfaction: 25 percent of adult hospital reimbursements from Medicaid are based on how well a hospital performs on the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey. Further, because clinician burnout and engagement are linked to patient satisfaction scores (Bodenheimer & Sinsky 2014), there is a simultaneous desire to measure and improve satisfaction of employees (Mahan 2016; Baer et al. 2017; Azam et al. 2017; West et al. 2018). These shifts have impacted how leaders prioritize improvements to their processes and services, and hospitals are feeling strong pressure to emphasize satisfaction while maintaining or improving their already high standards of safety and public accountability (Balik et al. 2011).

At Seattle Children's Hospital, a hospital, surgery, and specialty care system serving a population of 2.8 million children across Alaska, Idaho, Montana, and Washington, operational leaders at all levels are tasked with improving experiences and satisfaction. These stakeholders include busy, information-burdened physicians, nurses, and administrative leaders, across inpatient, outpatient, and business settings. When experience and satisfaction key performance indicators (KPIs) don't move as desired, they look to survey comments or other forms of qualitative feedback to help decide how and where to implement improvement efforts. Recently, these leaders have started asking for quantitative analysis of survey comments, in hopes of providing those data with the same level of confidence they obtain from their KPIs. In spite of the methodological and epistemological difficulties inherent to such a mix (LaVela & Gallan 2014), hospital leaders are now asking for in-depth quantitative analysis of qualitative data that they previously felt was unmeasurable or too complex to tackle.

The authors are an anthropologist in patient and family experience who works on projects involving human-centered design, ethnographic assessments, and survey analysis, and a data scientist who works on projects primarily in neurology and critical care, but who is often asked to consult on advanced analytics topics across the enterprise. Recently, both analysts independently received requests from stakeholders about text mining survey comments to find key areas for improvement for A) two of Seattle Children's day-surgery locations and B) employee satisfaction with a non-clinical, enterprise-wide support department. Upon comparing their requests, the authors realized that these instances raised fundamental questions with implications for business revenue and improvement strategy: What data analysis methods can best inform and improve patient and staff experiences? Can textual data provide meaningful insights efficiently?

In addition, these data requests put into question the overall approach to satisfaction and experience. In their management roles, leaders have little time to dissect the minutiae of experience, instead seeking evidence-based best practices and interventions from other institutions with the belief that such efforts will work across cultures, systems, and organizations. However, this approach minimizes the effects of the *lifeworld* on perceptions of care. Husserl (1936) introduced the lifeworld concept as he became increasingly concerned that quantitative measures "ignored qualities of the human experience" (Hemingway 2011: 3). The lifeworld is the whole of our lived experiences and is always at play when interacting with others, yet it is often taken for granted in our day-to-day interactions (Husserl 1936; Ekra et al. 2015; Barry et al. 2001). Dahlberg (2009) identifies that patient-centered or patient-led care often denotes an economic (patient as consumer) or political emphasis (patients as citizens). Instead, Dahlberg recognizes that patients want to be seen as people with both illness *and* well-being, agency *and* vulnerability, and can "feel unmet by interactions that emphasize one or the other" (p. 266). Thus, a framework of well-being that encompasses several realities and the tensions of everyday life may be critical to creating a healthcare experience where an individual feels "met" as a whole person (p. 269). By prioritizing satisfaction scores over patient's lifeworld-led experiences, parents are seen simply as consumers, rather than individuals with diverse needs and agency. The lifeworld is also relevant to staff and providers who feel unfulfilled by a medical system that struggles to recognize them as both individuals and institutional members.

To address these text analysis requests, the authors recognized that they needed to acknowledge the underlying beliefs and assumptions about textual data analysis and provide clear examples of the methods and modes of evidence inherent to those methods. In both cases, the authors felt that leaders were not aware of the usefulness and drawbacks of each method, as well as the work effort involved relative to other approaches. The authors sought to address these two requests, while at the same time providing leadership with a more nuanced view of the power and problems of text mining, and of methods combining qualitative and quantitative approaches.

By sharing examples of delivering meaningful and digestible evidence from family and staff satisfaction, the authors discuss the implications of making textual evidence more useful in healthcare's data-saturated context. A clear understanding of methodological approaches, including insights and actionable recommendations, as well as an approach to experience-based improvements that encompass the lifeworld will help decision-makers make meaningful operational choices, though perhaps not in the way leaders will expect. This case study begins to explore the larger questions for our organization: How does one make strong inferences from an "n" of one, or conversely, humanize an "n" of millions of data points? What are the best ways to connect stakeholders with evidence—both statistically and ethnographically—to support their daily work and improve the decisions they must make?

THE BUSINESS CHALLENGE

Hospitals rightfully prioritize safety and accountability. The intense regulation of healthcare means that hospitals must react to rapidly changing and sometimes ambiguous requirements from local, state, and federal government, accreditation agencies, and insurance companies, making long-term strategic planning difficult. Often, improvement efforts are a reaction to changing industry standards, compliance requirements, and competitive growth. While innovation, employee satisfaction, and customer experience are of increasing importance, it has been difficult for those priorities to compete with other long-standing needs within this constantly shifting management context.

Pediatric hospitals are currently exempt from satisfaction-based funding mandates from the federal government because child Medicaid is state funded. Yet value-based purchasing—which would include patient satisfaction—is now being considered for reimbursement by private payors, and this may soon expand into children's Medicaid as well. Significantly, most pediatric hospitals receive more than 50% of their revenue from Medicaid (Lagasse 2017), with nearly as much from private payors. Changes in either payment approach would have wide-ranging effects on hospital operations regardless of funding source. Meanwhile, provider burnout and staff dissatisfaction have led to higher turnover and presumed lower KPI performance, both of which impact the bottom line (Azam et al. 2017; West et al. 2018).

When considering what to do to improve experience and satisfaction, leaders are usually tasked with improving Likert-score based KPIs. When those indicators don't rise, they seek meaning and insight from free text comments and complaints that accompany survey results to try to improve patient satisfaction scores. Historically, leaders have two approaches to assessing evidence from survey comments: read it themselves or request a content analysis. When deeper insights are needed, leaders have occasionally requested ethnographic studies.

Modern analytics can now apply complex algorithms to textual data, so leaders now have a fourth option, text mining, which is believed to yield high quality evidence at a low labor cost. These four methods were applied to address two business problems: improving HCAHPS-based KPIs for two day-surgery sites and assessing employee satisfaction with one non-clinical, company-wide support department using the two-question (one Likert, one free text) Net Promotor Score (NPS) approach (Reichheld 2003).

For the past two years, the HCAHPS survey's patient satisfaction KPI (proportion scoring 9s or 10s on a 0-10 "overall satisfaction" Likert scale) has been in the mid 80s for the Olympic location and the low/mid 90s for the Cascade location, both of which compare similarly with many pediatric hospitals in the US. The employee NPS survey was to acquire a baseline for future comparisons; the initial NPS rating was below zero, a customer-service score below the industry average. Because hospital revenue is indirectly linked to patient satisfaction scores (and may soon be directly linked), and provider and employee actions and incentives can be directly linked to patient satisfaction, there is strong interest in improving these scores.

Even when given quantitative analysis of satisfaction KPIs, our stakeholders frequently ask, "What does this *really* mean?" While the Likert-type scores identify areas for potential improvement (e.g., "provider explained clearly"), there is no clear indication of what needs to be improved (e.g., explanations of treatment, medication, expectations) or why an improvement is needed. An implicit presupposition behind this business challenge is that by quantifying the textual responses as well as KPIs, the hospital may be able to understand that hidden meaning more clearly and use that information to subsequently improve satisfaction scores.

Quantified or not, different analytic methods can lead to different results, and consequently yield recommendations that can lead to very different business interpretations. There are two intertwined issues at play: mathematical properties of KPIs necessarily create incentive structures that can implicitly predefine certain decision strategies, and textual data analysis may not create enough explanatory evidence to help decision makers determine what interventions and improvements to support. Further, the reality of satisfaction being scored by a single KPI is often met with skepticism, given that, for example, a patient's experience of a full week in the hospital with changing units and rotating care teams is ultimately distilled into a single satisfaction score.

The business challenge is not to simply tell the satisfaction and experience stories elicited from both qualitative and quantitative data, but also help the organization understand how different methods might lead them to different outcomes, with the goal of becoming more informed requestors and consumers of analytic techniques for text data. Because of the implicit bias towards quantitative results (regardless of ultimate usefulness), the authors believed that an exploration of *Return on Method* (ROM) would be an essential first step towards understanding eventual *Return on Investment* (ROI). The authors thus used these two business requests to create an initial ROM for the analysis of text-based data, results, and evidence.

As lessons learned about deriving evidence from text data were largely the same across the results for both business requests, this case study presents only a few selected examples from each, focusing primarily on the day-surgery comments, as that is closer to our mission.

RESEARCH DESIGN AND METHODS

Research Design

To address the issues of identifying actionable insights and increasing satisfaction KPIs, the authors decided to test approaches to data. By first testing data approaches, the authors developed their organization's understanding of such methods, and what each can offer to interpretation, actionable insights, and benefits—as well as limitations—of such approaches. The authors compared quantitative and qualitative methods that a) stakeholders had requested, and b) that the authors felt might provide better answers for those requests. Thus, the authors compared three different ways to analyze survey comments: This entailed a detailed comparative analysis of two different satisfaction surveys:

1. Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) patient and family experience survey results from day-surgery encounters at the Cascade and Olympic locations (anonymized), collected between October 2015 and March 2018 (2.5 fiscal years); and
2. Net Promoter Score (NPS) survey of employee satisfaction with a non-clinical support department collected in January 2018.

The HCAHPS survey included 1,004 unique free-text comments from nearly 7,000 surveys. The survey can be completed in either English or Spanish. Caregivers are eligible for survey requests up to four times per year. The survey was completed by approximately 28% of visits during this time period, and 14% of survey respondents left a free-text comment. Both surveys contain the same open-ended question, *"What else would you like to say about your experience?"* This is the comment field stakeholders believed would hold the key to their improvement ideas.

The internal employee NPS survey included 495 free text comments from 658 survey responses, a survey response rate of just over 10%. The free-text question was, *"How likely is it that you would recommend [department name] to a friend or colleague?"* Again, the working idea was that these free-text comments could help identify and define service improvement areas, ideas, and solutions.

In addition to content analysis and text mining, the authors wanted to test two additional approaches to see what they might contribute to understanding and actionable insights for satisfaction: end user reading of survey comments and ethnography. The authors sought to understand if end user reading of survey comments (reading by senior and/or local leaders) was accurate compared with other, more formal, analytic methods.

While ethnography is a very different approach to data collection and analysis, the authors felt that comparing ethnographic results to survey comment analyses could illustrate how different approaches to data can yield different results, which in turn can have a real business impact. As Wang (2013) describes, "thick data", through its rigorous sampling, analytic, and storytelling approach, provides inspirational and emotionally-based insights that can impact many aspects of business, including experience and satisfaction. Thus, a comparison of these four methodological approaches might provide insights into how evidence is constructed, including exposing the inherent issues to interpretation, actionable insights, and benefits of each approach.

Confounding our analyses are several limitations, particularly with respect to the HCAHPS survey. The vendor—and thus survey format—changed at the end of fiscal year 2017 from paper to phone surveys. Further, the earlier version of the survey contained 70+ questions, whereas the new version contains only eight. Both surveys contain the same open-ended question, however, though survey fatigue in the paper version is certainly a concern. Further, like any survey, we expect that our results suffer from survey bias, where particular (unknowable) combinations of family socio-economic status, patient outcomes, and specific within-care patient and family clinical and administrative experiences all contribute to the likelihood of whether a survey will be completed and submitted. The NPS survey was confounded by the use of the free-text question's wording, which prompted specific comments from several respondents along the lines of "this doesn't make any sense—we have no alternative to [department name]'s services." Limitations notwithstanding, leaders expect answers, so the authors explored the data as if it were representative of the populations the hospital serves.

Ethical Considerations – This study was approved by the Institutional Review Board of Seattle Children's Hospital (IRB #1138).

Methods

Content/Thematic Analysis – An inductive analysis was performed where a series of codes and definitions were created from the text to help characterize the meaning from the parents' or employees' perspectives. Comments were classified as positive, negative, mixed, or neutral sentiment. The codes were validated by a second coder. Once the codes were validated, they were placed into parent themes to characterize types of improvement and, for the surgery comments, at what point in the process of care. A code count (frequency) table was generated to understand if there were consistent trends over time, by location, and across locations.

End User Reading – For the employee satisfaction survey, the data scientist suggested that leaders read the comments directly and do some self-directed, casual analysis to indicate whether a comment reflected more positive or negative sentiment about the department and assign a theme of their choice to that same comment (such as "customer service" or "managerial support"). In keeping with how this might play out in practice, no formal instructions were given on how to best accomplish this task.

Text Mining – For both comment sets, exploratory text mining techniques were used to assess whether these approaches could provide useful insights, particularly because they are fast and easy to do. These included word and *n*-gram frequency assessment, lexicon-based sentiment analysis, word correlation analysis, and structural topic modeling using latent Dirichlet allocation. Results from employee satisfaction comments (NPS question 2) were further linked to the Likert scale results (NPS question 1) to explore how each quantitative response compared with the associated qualitative comment. All statistical analyses were performed using R 3.4.3 (R Core Team 2017).

Ethnography – The anthropologist observed and interviewed more than 300 patients and families, providing ethnographic insights to 35 clinical areas across 9 sites of care. The context of this work has been in small, clinical settings that are focused on improving some element of experience for their specific patient populations. For day-surgery, the anthropologist used empathy mapping, workshops with staff, and a focus group of 4 family advisors along with internal stakeholders, to look at location-specific issues.

Data Analysis and Methods Evaluation

Analysis of text mining was performed by the data scientist, while analysis of content analysis and ethnography were performed by the anthropologist. Both practitioners evaluated their prospective stakeholders' ability to conduct an end user analysis by reading the comments. The authors reviewed the results together and evaluated each method by the following factors: accuracy of insights, reliability of consistent results, validity of approach to data, and if the methods met our stakeholder's expectations of being fast, quantifiable, scalable, meaningful, and actionable.

THE VOICE OF THE PATIENT/CUSTOMER: COMPARING VOICES ACROSS METHODS

Content Analysis

Content Analysis: Topics/Themes – Qualitative content analysis of the day-surgery data identified 42 distinct codes related to possible areas for improvement in a total of 329 comments that included either negative sentiment. In addition, there were 31 codes that identified strengths out of 668 positive comments. Positive comments overwhelmingly included general statements like "thank you, great experience", often identifying specific nurses and doctors as kind, helpful, and professional.

A benefit of content analysis is the ability to break down a theme, like communication needs, into specific content or moments throughout the experience (Table 1). However, when looking at a single unit or location like a day-surgery center, the themes become difficult to trend over time. For example, communication issues, while important to families, span many different stages, roles, and functional challenges. Stakeholders are still left wondering what to do to improve communication and when attempting to create specific recommendations, trends are quickly spread thin, leaving many options on what to improve with little prioritization. (Table 2).

Table 1. Example content analysis themes from day-surgery comments, including the broader category, more specific content, sentiment, and theme definitions.

Parent Theme	Child Theme	Sentiment	Definition
Communication	Front staff	negative	Did not show respect or courtesy, not friendly, not helpful
Communication	Changes to team	negative	Different surgeon than expected, switching providers or nurses too much, residents, changes not explained well
Communication	Communicating expectations	negative	When procedure lasts much longer than expected, lack of regular updates
Communication	Parent anxiety & stress	negative	Did not feel supported or prepared
Communication	Concerns not heard	negative	Staff or providers did not listen, not sensitive to needs, did not address pain
Communication	Teamwork/communication	positive	All worked together to make comfortable and explain, consistent information
Communication	Communication with doctor	positive	Easy, quick, timely communication with doctor (texts within minutes, photo updates, etc.)
Communication	Interpreters	negative	Quality of interpretation
Communication	Post-op doctor communication	negative	Took longer than expected
Communication	Post-op get well card	positive	Appreciated the thought, felt cared for by team

Table 2. Example content analysis themes and sentiment code counts for the Cascade day-surgery center, by fiscal year.

Child Theme	Sentiment	FY 2016	FY 2017	FY 2018*
Front staff	negative	2	1	2
Changes to team	negative	2	7	0
Communicating expectations	negative	2	1	0
Parent anxiety & stress	negative	0	1	1
Concerns not heard	negative	0	2	0
Teamwork/communication	positive	19	7	7
Communication with doctor	positive	5	4	1
Interpreters	negative	0	0	1
Post-op doctor communication	negative	0	2	0
Post-op get well card	positive	11	13	5

**FY 2018 data encompassed 6 months of responses.*

While there were many codes that represented improvement opportunities, very few were consistent across fiscal years. Only three codes trended highly: wait times/delays on the day of surgery, being with their child in the recovery room, and patient comfort. Delays on the day of surgery were of acute concern to parents who felt that their child was NPO (nothing by mouth) longer than necessary. In addition, the length of time parents waited was often much longer than what they were told to expect—sometimes twice as long—causing an inability to plan for things like additional meals, entertainment, babysitting for other children, or time off work. Thus, communication and setting expectations before the day of surgery might help but how these fit into the current process is unclear.

The recovery room offered its own arrangement of topics, primarily that parents were upset or concerned that they weren't with their child when they woke up from surgery. Parents were distressed about their child waking up in an unfamiliar place without anyone they recognize. In addition, some patients have adverse reactions to waking up from anesthesia, and parents wanted to be there to help manage it. While the Cascade surgery center allows parents to come to the recovery room when they are waking up, the Olympic location only does so on a case-by-case basis.

In both the positive and negative comments, "patient comfort" became a key insight, where parents reflected on how their child felt throughout the process. Given that only three themes trended consistently in the negative feedback, it can also be helpful to compare how parents describe when something was done well and when it could be improved. Consider the following comments:

> They did an excellent job. The nurses especially were fantastic. I think my son actually thought it was a fun experience. They were excellent with distractions and assessing the needs of my child (and his parents). I was worried, but they made it so much easier to go through the first anesthetic procedure of my son's life. They also sent a card after. He loved the Paw Patrol stickers on it.
>
> My son & I had specific wants for my child going under procedure. I ask that no laughing gas be used for IV insertion, the IV be placed in his AC instead of hand and asked if mom could be there as he went under. I was able to go back, however they put the IV in his hand, gave him laughing gas which made him dizzy and then told me it was time to go before he was out...

In the positive feedback response, the parent identifies elements of experience that were important to them and their child, such as distraction, easing concerns by addressing needs, and sending a get-well card afterwards. However, the specific behaviors that made the team "excellent", the nurses "fantastic", and the experience "fun" during surgery are largely unknown, besides sending a get-well card afterwards.

When patient comfort was a negative experience, parents similarly talked about Children's staff not addressing specific needs or easing concerns they had for their child. These comments illustrate that communication effectiveness and meeting both the parents' and the patients' lifeworld-led concerns are important. Yet, content analysis struggles to recommend specific interventions to access the lifeworld and improve satisfaction scores.

Content Analysis: Sentiment – As expected based on the KPI scores, content analysis found a majority of HCAHPS comments displaying positive sentiments towards their experiences. Over the 2.5 fiscal years of the data, 75% of Cascade's comments were positive, while Olympic's comments were 56% positive, both of which remain fairly constant over time. This suggests that Cascade does provide a better experience than Olympic, something already inferred by managers from the KPI scores.

Content Analysis: Comparing Locations – Content analysis offered some distinct insights into what may contribute to Cascade's higher satisfaction scores. There were consistently fewer comments for Cascade that complained about wait times and day of communication (Table 3). In addition, parents provided more positive details about what the staff did to ensure comfort, set expectations, and allow them to be with their child in the recovery room. Finally, many parents from Cascade commented about the card sent to their child at home post-surgery. These subtle switches in language suggest that some of these improvements could influence the KPI, however further context about the differences between the two locations (such as populations served, facilities, staff engagement, service delivery, and flow) would also need to be considered to make such claims. In short, the comparative content analysis provided some general insights into what could be improved and why, but by itself does not provide enough *actionable* insights to improve the KPI.

Table 3. Code counts and trends for wait time and get-well card themes for each day-surgery location. Olympic has consistently higher complaints about delays and wait times on the day of surgery.

Code	Location	FY 2016	FY 2017	FY 2018*
Wait time	**Cascade**	6	3	7
	Olympic	9	28	15
Get-well cards	**Cascade**	11	11	5
	Olympic	0	0	0

**FY 2018 data encompassed 6 months of responses.*

End user results

The authors wondered if leaders might be able to read the comments themselves to inform decisions, given that they have contextual understanding and are aware of information from numerous data sources. In a post-analysis meeting with Olympic location's day-surgery leadership, it was apparent that they had indeed read their survey comments in detail. Without hesitation, and without having yet seen the analysts' results, a local leader listed off the top themes identified in the content analysis as areas they needed to prioritize and discussed possible interventions. In the case of the internal department satisfaction survey, the leader was also able to accurately identify sentiment and content areas to improve. Yet, when speaking with many of these leaders, they often feel that they can't do anything about the issues identified by patients, families or staff in survey comments. They note that the issues are too large and above their scope or that they have too many other competing demands. More generally, there is often a suspicion of analyses done by leaders and not analysts, because of the common assumption that such work will result in unseen bias and casual rather than scientific analysis, leading to purposeful spin. Hence, their data requests: an attempt to bring *evidence* to light by which they can obtain strong, high-level support for the complex changes that may be needed.

Text Mining

The Cascade location had 538 comments, and the Olympic location had 466 comments. Within these 1,004 comments, there were 1,157 sentences, containing just over 41,600 total words, 3,035 of which were unique. With stopwords removed ("a", "an", "the", "of", numbers, and so on), there were 11,335 words, 2,493 of which were unique. These comments also contain 20,472 unique bigrams (2-word pairs) and 30,969 unique trigrams (3-word pairs) when stopwords are retained.

Text Mining: Topics/Themes – Word clouds—which quantify word counts by sizing the most common words according to total usage—are often used to summarize a collection of text, with the idea that a busy decision-maker could scan the image to get a sense of the topics/themes and tone of those comments. Data scientists are often disparaging of word clouds, partially because the frequency-based sizing is relative (e.g., compare panels A & B in Figure 1), but largely because they only reveal patterns that are already obvious. That said, people like them. Word clouds can help provide a glimpse of overall trends or patterns, even if this glimpse is, practically-speaking, useless for decision making.

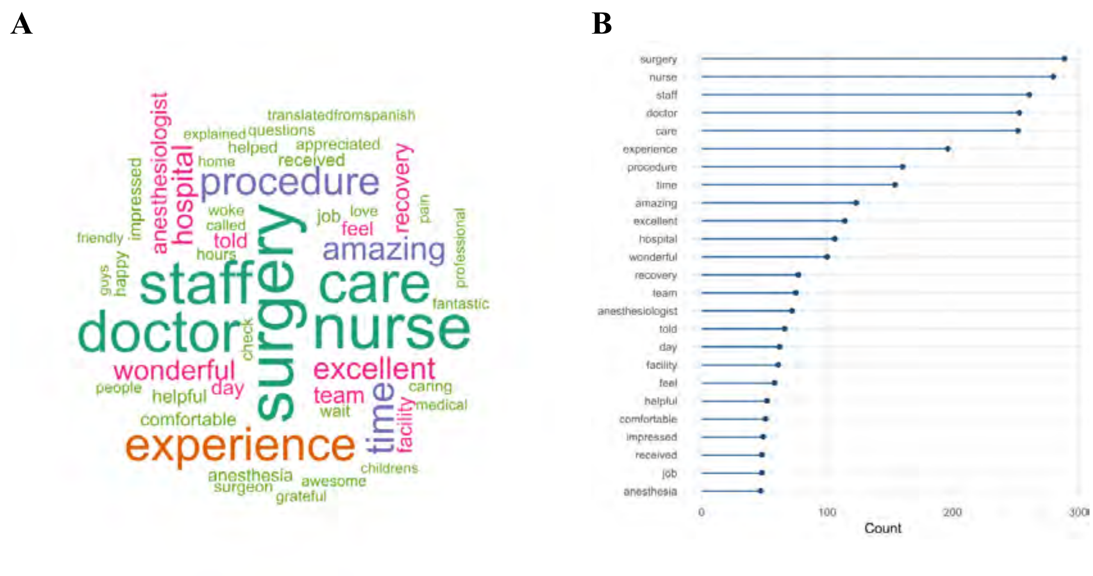


Figure 1. Pretty but vague (A) as contrasted with ugly but accurate (B) word frequencies in Cascade and Olympic locations' day-surgery survey comments between October 2015 and March 2018.

Mining the comments for themes using structural topic modeling applies considerably more sophisticated methods to achieve similarly useless results. Unsupervised topic modeling explores text data to try to identify latent structures. Giving the algorithm a topic count of $k = 14$ (k is the estimated optimal number of topics based on word probability distributions) reveals many of the words seen in the word cloud as specific indicators of each latent topic (Figure 2; Table 4). While this can sometimes indicate particular themes, other times it can also create distinctions where none actually exist. For example, while topics 3 and 6 are fairly different (see Table 4 for some example comments), topics 6 and 14 cover basically the same sorts of negative comments associated with wait times. Topic 9 could be associated with recovery room issues, and 10 could be associated with scheduling, though upon inspection of comments in those topics, neither holds very strongly. The remainder of the topic groups could be reasonably lumped into a single broad topic expressing positive emotions and experiences.

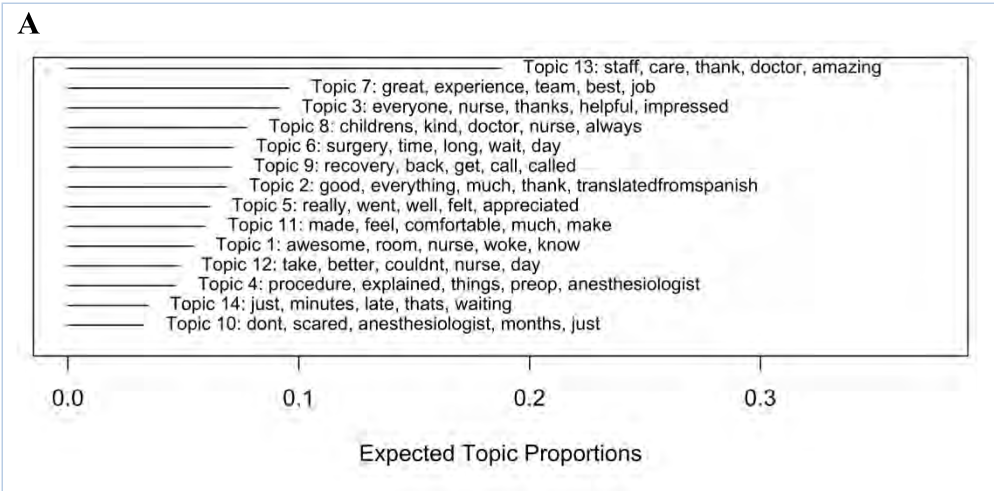


Figure 2. Structural topic model results of the day-surgery comments, given an input of *k*=14 topics, ordered by expected frequency of those topics across all comments. Words listed with each topic are the top 5 representative words of that topic.

Text Mining: Sentiment Analysis – Another common use of text mining is sentiment analysis using pre-made lexicons to attempt to classify a given comment as positive, negative, or neutral. Each comment is broken into individual words, which are in turn assigned a sentiment or sentiment score depending on the lexicon, the sum of which is meant to provide an overall sentiment score for that comment as a whole.

Using three different major sentiment lexicons, overall sentiment scores showed similar general patterns between lexicons—higher at Cascade—as well as some subtle differences between locations when viewed over time (Figure 3).

Using each comment's net sentiment score from FY18 (n = 170) comments only to predict whether the comment will be a "top-box" score—which means to get at the question of whether negative sentiments are related to a respondent's likelihood of rating the hospital highly—shows a clear but weak relationship (Figure 4 A&B; odds ratio = 1.17 [95% CI: 1.09, 1.28]), suggesting that while positive net sentiment is related to high ratings, the relationship is more ambiguous with negative sentiment. The content analysis-based sentiment category (using only "negative" or "positive" classified comments) was a better indicator than text mining for whether the respondent would rate the hospital as a 9 or a 10 (Figure 4 C&D; positive sentiment odds ratio = 7.1 [95% CI: 4.0, 13.7]), but it also did well quantitatively linking negative comments to the "not top-box" score (0s-8s) category (negative sentiment odds ratio = 0.21 [95% CI: 0.04, 0.32]).

Table 4. Examples of representative day-surgery comments for appreciative (3 and 7) and wait time-based (6 and 14) topics.

Topic Number	Topic Keywords	Location	Example Representative Comment
3	everyone, nurse, thanks, helpful, impressed, caring, friendly	Cascade	*Nurse ____ was so nice to our child and to us. She is a great nurse! Nurse ____ was great too. Our son was so welcomed and having fun, he forgot he was at a hospital. I wish all dr's and nurses (even for adults) were as great as you are. Thanks!!!*
		Olympic	*Everyone there was extremely kind and helpful, I had all my questions answered, they cared for my child extremely well, I couldn't ask for a better children's hospital.*
6	surgery, time, long, wait, day, hours, took	Cascade	*My not even two year old had to go quite a long time before surgery. I know that this is a requirement when you're having surgery but I guess he was booked on a day that our surgeon was only doing afternoon surgeries, so he was not able to eat food from four am until … close to 12:30 and this … is a really long time without food.*
		Olympic	*It would be nice not to have to wait so long between check in and the procedure. It was a bit of challenge to occupy a thirteen month old who had an empty stomach. It is a bit rough on parents who are nervous about the procedure. But also understand that it probably can't be helped.*
7	great, experience, team, best, job, work, recommend	Cascade	*I could go on forever. I've never witnessed such a competent, cohesive, perfectly ran organization! Nurse ____ - Amazing! Dr. ____ - Brilliant! Thank you Seattle Childrens!*
		Olympic	*The amount of care and compassion from the preop/post-op nurse, otolaryngologist and anesthesiologist were beyond my expectations. I am so glad I decided to choose Seattle Children's and would definitely continue to recommend them to others.*
14	just, minutes, late, thats, waiting, nurse, doctor	Cascade	*They were running almost two hours late, we were not told in fact at the time of her procedure we were told we would be back in five minutes. It was frustrating to say the least.*
		Olympic	*We were called an hour before check in and told the surgery was going to be delayed. Our infant had already been 'fasting,' and to have to delay this by another hour (last minute!) made for a very hungry baby and stressful situation for parents.*

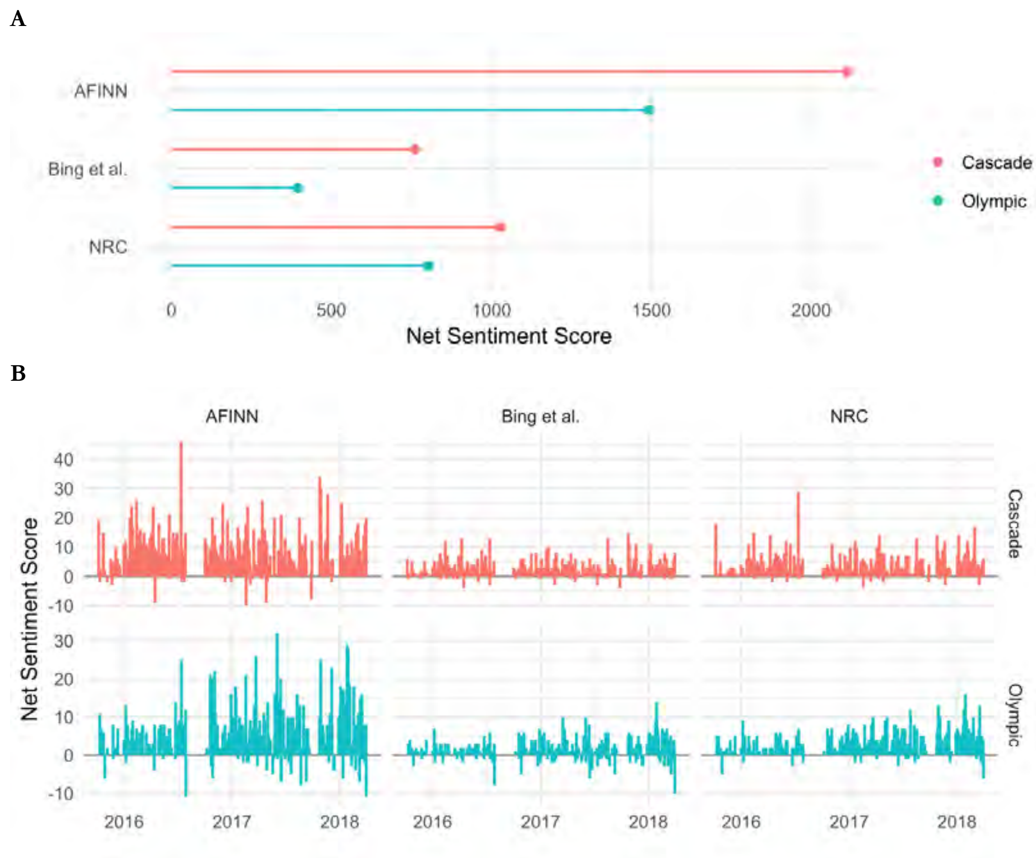


Figure 3. Day-surgery comments' sentiment scores by location and lexicon. A) Overall sentiment score for the entire time period of the comments, Oct 2015-March 2018. B) Trends in sentiment score over time.

For the NPS-based employee survey, net comment sentiment only vaguely trended with scoring group, where the median score of detractors was lower than that of the promoters (Figure 5). However, nearly all comment sentiment scores suggested net positive ratings, even for the detractor group, for which 75% of the distribution was scored as neutral to positive sentiment.

Text Mining: Comparing Locations – Word clouds are a more common way to visualize ungrouped word frequencies, but differences between word frequency by a grouping variable—in this case, day-surgery location—can also be visualized in a single plot. Figure 6 shows the relative frequency of each word, relative to a baseline of identical frequency (the 1:1 line) between groups. Words that appear above the diagonal are more frequently seen in comments from the Cascade location, while those below the diagonal more frequently occur in the Olympic location's comments.

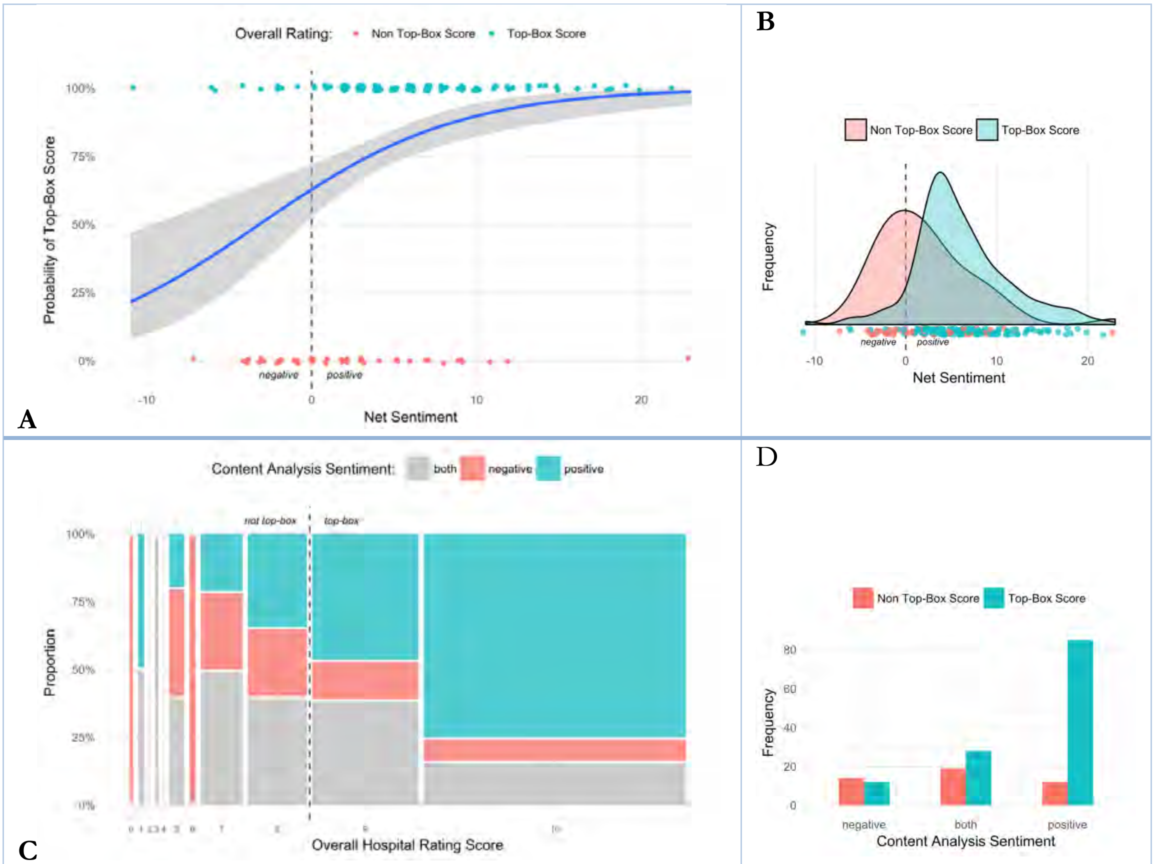


Figure 4. A) Logistic regression of AFINN-based net sentiment score to the probability of a respondent giving the hospital an overall satisfaction rating score of 9 or 10 ("top-box" score) for day-surgery comments received in fiscal year 2018. B) Distribution of net sentiment score by overall rating group. C) Mosaic plot of overall satisfaction score to content analysis-based sentiment category for day-surgery comments received in fiscal year 2018. D) Distribution of content analysis-based sentiment score relative to top-box category. These results were already obvious to end users in a qualitative sense: satisfied patients and families (who respond to the survey) respond positively with high KPI scores.

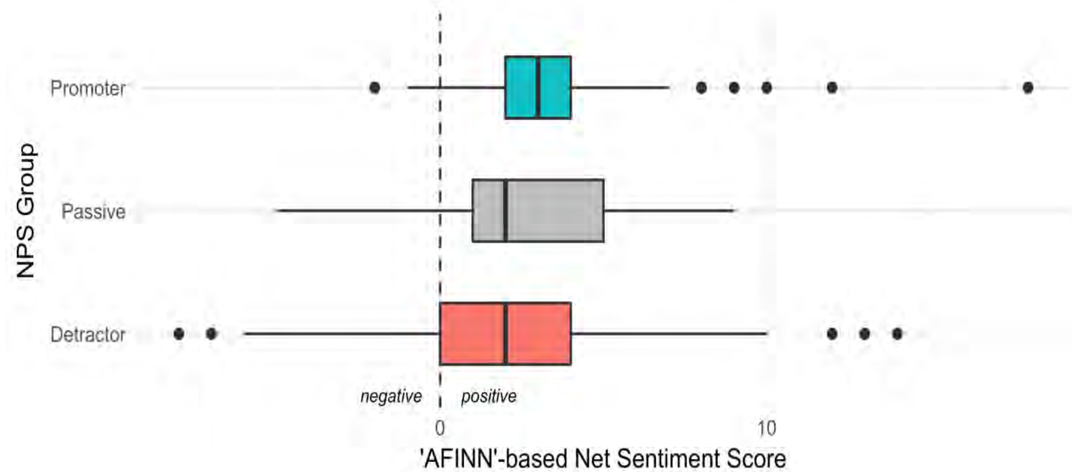


Figure 5. Distribution of comment sentiment scores for employee satisfaction with a non clinical support department, by NPS grouping.

Another way to explore word frequencies between groups is by looking at differences between term frequency-inverse document frequency (*tf-idf*). This method looks at both frequency of occurrence across the groups generally (the *tf* part), as well as how common these same words are across documents within each group (the *idf* part). Basically, it suggests how important a given word is to the given grouping, or similarly provides sets of words that distinguish one group of text from another. Using this method, top terms can be visualized against each other, as seen in Figure 7, which shows the top 10 *tf-idf* terms by location.

Cosine similarity is another way to quantify the degree of difference between two text groups, values closer to 1 reflect similarity in word use between groupings, and values closer to 0 reflect strong differences. The cosine similarity for comments grouped by location is 0.94, indicating that the word distributions within comments from each location are nearly identical.

As with all of the day-surgery comments' text mining results, none of the results shown above came as any surprise to senior or local leaders.

Ethnographic results

Ethnography: Topics/Themes – Ethnographic investigation identified how much distress the pager and communication process during surgery were causing to both parents and staff. The delay in communication was a very nerve-racking experience for all caregivers; several parents relayed stories of being paged much earlier than expected, even only 4 hours into an 8-hour procedure. After receiving the page, families proceed through an extensive process with multiple delays to speak with a provider. In each case when caregivers spoke about this process and lag in communication, they mentioned that while the outcomes for their child were positive, the stress and tension they felt when left waiting was burdensome.

Caregivers also reflected on feeling lost in the system on the day of surgery, amid managing their child's fear and anxieties as well as their own. Once leaving their child in the surgical suite, parents are left alone in the back halls of the hospital: unsure where to go, what do to, or who to go to if they have a question. Family members often wouldn't leave

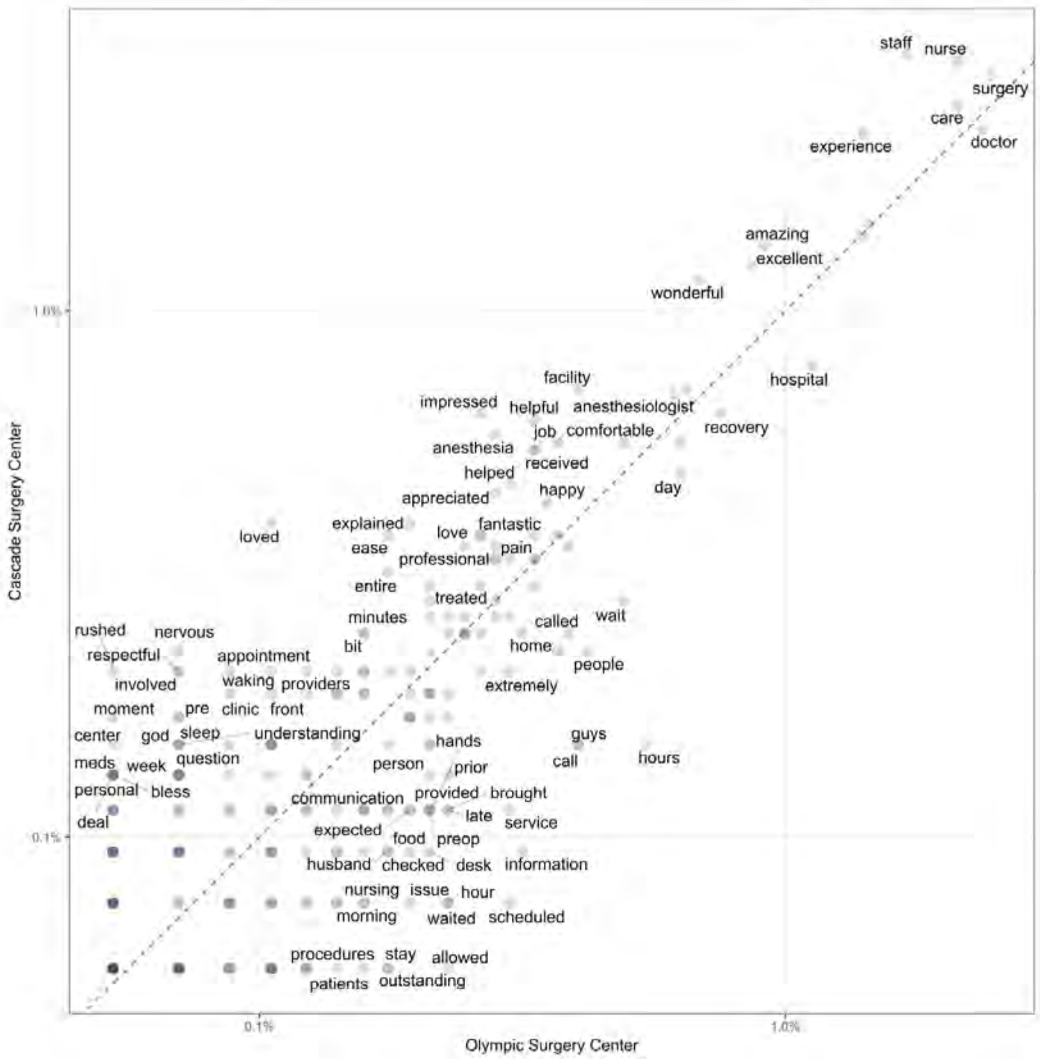


Figure 6. Differences in word frequency between day-surgery locations. For purposes of clarity, not all term labels are plotted. The axes are log-scaled occurrence frequency, the 1:1 line represents the line of equal frequency, and the darkness of the point represents the number of unique words that have the same frequency + location characteristics. The further from the diagonal a word is, the more frequently it appears in one location's comments relative to the other location.

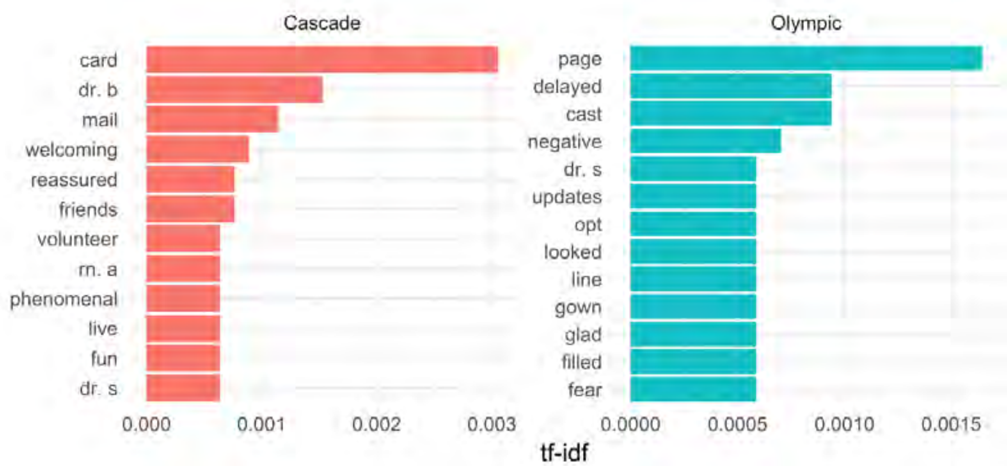


Figure 7. The top 10 highest *tf-idf* words by day-surgery location. Nurse and doctor names are anonymized.

the surgery waiting area, fearful that the pager wouldn't work, or someone wouldn't be able to get a hold of them, even though they need a place to decompress after saying goodbye to their child.

In turn, staff struggled to know when or what was communicated to families, usually relying on nursing documentation that was frequently delayed because documentation necessarily follows actual care events. Staff deeply wanted to address family concerns, but acutely felt the competing needs of patient care, process flow, and team and family communication. These insights were not revealed with any other method and findings from this work created clear understanding into the problem at hand and the motivation to seek workable solutions.

Ethnography: Sentiment – Ethnographic work in the hospital setting has also revealed a tension about delivering constructive criticism to providers and the organization. Often, caregivers are so grateful for the services and positive outcomes that they are hesitant to speak negatively about any aspect of the service. Thus, ethnographic work in hospitals must recognize the vulnerable position of its informants, relying heavily on observations, qualitative interpretation, and family advisors who are in the role of advocating for better experiences. Given the delicate relationship between caregivers and clinicians, work that can capture the trials and suffering of patients and caregivers becomes all the more critical.

In the case of day-surgery, caregivers retrospectively described strong negative emotions, including distress, feeling lost, lack of control, and fear, despite the fact that the outcome for their child was positive. Ethnographic work uncovered how the communication process contributed to these negative emotions: while ethnographic sentiment was not quantifiable or identified as inherently positive or negative, the connection to deeper human emotions created a stronger sense of urgency to solving this problem. By understanding the trauma unintentionally generated through the communication process, prioritization and actionable insights to make improvements were clearly identified.

Ethnography: Comparing Locations – Ethnographic work throughout the hospital and different day-surgery locations has led to findings that span many settings and directly impact patient, family, and staff experiences. First, clinical team dynamics, including shared physical space, interpersonal relationships, and staffing support, all play a central role in their perceived effectiveness by patients and families. Often, changes made to improve patient flow (e.g., medical assistants sitting together so they can efficiently room patients instead of with the rest of the clinical team) impeded staff relationships by creating a less cohesive environment. Awareness of the trade-offs in improvement efforts and their impact on patient and staff experience may need re-evaluation, especially given staff burnout from initiative fatigue and constantly shifting priorities (Briody 2018).

In addition, patients and parents wanted their providers and teams to "know them" and feel "met" where they are, not where institutional habits placed them. Needing to be known and met arose in a variety of interactions: understanding their medical and social history, personal preferences, navigating the medical system, individualized resources and treatments plans, as well as confidence that staff at every step knew what they needed so they didn't have to keep retelling their story. From these insights, ideas about how to incorporate the lifeworld not only into clinical but also administrative systems were identified. Observations showed how the available tools and systems continually impacted the clinical team's ability to address the lifeworld. For example, one clinic's providers were consistently overbooked and scheduled out too far to see acute patients in the needed timeframe. The providers were frequently recommending that patients get on a waiting list to be seen for timely treatment, only to discover much later that such a list did not actually exist. Thus, the deeper insights from ethnography shed light on assumptions, contextual nuances, and social dynamics that directly impacted satisfaction and experience. By understanding such insights, actionable recommendations for local leaders become possible, meeting the stakeholder's requests for meaningful evidence.

DISCUSSION

Different analytic methods necessarily lead to different insights. While both qualitative and quantitative analytic professions hold that as a truism, it is less well appreciated or understood in management. As seen above, each methodological approach provides different views and interpretations of the survey comment results, which can subsequently result in different business interpretations and thus different action recommendations. Essentially, stakeholders need to understand best use cases for each method; the critical role of an analyst is to help them explore the trade-offs before too much time, money, and effort are invested in an approach that doesn't meet leaders' primary needs. A side-by-side evaluation of each method relative to the survey comment data (Table 5) helps illustrate how different approaches to satisfaction problems and data might not deliver the attractive results stakeholders were looking for and could even lead to misinterpretation of patients' and caregivers' key needs, with corresponding impacts on the business and bottom line.

Text mining struggles to identify actionable themes, accurate sentiment, or group distinctions that are readily identified by both content analysis and end users

It is clear from the results that text mining fails to achieve what leaders were hoping for—a quick, quantifiable way to justify prioritizing satisfaction improvements. Instead, it was unable to provide any insights not already obvious to the end users.

Table 5. An example of how different methods can lead to different conclusions and business implications from day-surgery satisfaction comment analysis results.

	Content Analysis	End User Reading	Text Mining	Ethnography
Insight Differences	Wait times, being with child in recovery room and comforting the patient should all be improved at the Olympic location.	Wait times, recovery room, and comforting the patient.	Work on wait time, but both locations seem the same.	Communication process including pagers, team communication and visibility into family communication.
Business Interpretation	Olympic should focus on improvements Cascade has made – e.g., more TV monitors for status updates, volunteer program, & get-well cards.	Find new ways to improve wait times. Allow parents into recovery room.	Find new ways to improve wait times.	End-to-end process redesign focused on creating less traumatic experiences, while also improving communication amongst staff and families.
Best use case	When there are <1,000 comments. To confirm what leaders/ stakeholders already know, unless nothing is yet known.	When there are <1,000 comments. To identify opportunities and how they relate to other sources of data and evidence.	When there are 1,000+ to tens of thousands of comments and a need to classify or identify the most common themes.	When it is unclear what to prioritize, to identify how human behavior, values, or emotions in specific contexts influence experience. When needing innovation or inspiration.

The example of sentiment analysis in particular shows that the problem can be deeper than lack of insight production, as well. While there was a weak trend in sentiment level for the employee NPS survey that showed more positive sentiment linked with higher scores, there were enough notable exceptions that showed the peril inherent to assuming that the aggregate results indeed reflect the underlying sentiments of the respondents. For example, one comment scored up among the most positive sentiment scores of the entire survey. The text of that comment belies the positivity of that score:

> … staff on night shift are not very helpful or nice. They act like we are bothering them rather than being there to help. Frequently they don't help because they can't figure out what problem you have. It used to be called the "help desk" and when someone suggested calling the "help desk" we would all laugh. Just being nice would help out considerably. Customer service...

In fact, this respondent had given the department an overall NPS score of 0, the worst possible rating. The reason the sentiment was rated so positively by the algorithm makes sense when one recalls that the sentiment score is based on the sum of the sentiment of the individual words. Thus, words like "helpful", "nice", and "laugh" all contributed positive values (given the lexicon used), while words like "bothered" and "problem" did not subtract from that overall score. This highlights the importance of understanding not just the data but how lexicons are built and applied to understand whether a sentiment analysis is accurate.

As this example shows, automating sentiment analysis is extremely problematic. Do the quantitative sentiment values accurately reflect the qualitative assessment of the same data source? How would one know? Conversely, is it worth the time to have a human-run sentiment assessment done to get a sense of positive versus negative comments frequency, or does the satisfaction KPI score provide enough information? In short, the idea of sentiment analysis brings to light an important but hidden consideration: strictly from an ROI point of view, are the comments *actually* important?

From an incentive point of view, improving a summary score for 0-10 Likert scale-based KPIs would likely focus on moving 7s and 8s up into the 9s and 10s. More people, with fewer strong complaints (if any) are in this category than those who score <7, and it would certainly be easier to address those respondents' needs as a means to improve KPI scores, if only to nudge them up a single point in their ratings. For example, in the employee satisfaction survey, the most common complaint amongst the 8s involved requests for modernizing their tools, while those who scored <7 focused on deep, difficult problems such as pay scales, skill sets, siloed responsibilities, and gatekeeping issues. While the best improvements to actual conditions would be focus on those <7s, it's far easier to do things like updating tools, which is more likely to create discernable changes in subsequent KPI scores. Additionally, the conceptual constructs behind the surveys' scores may not correlate at all with the complaints voiced in the free text responses, leading to differing implications on what should be done—raising survey scores and addressing complaints may be mutually exclusive activities under resource-limited conditions.

Content analysis cannot fully account for the lifeworld, and priorities should be on meaning, not substitutes for thinking

While content analysis yielded much more meaningful results than text mining for day-surgery feedback across the board, the anthropologist still had reservations as to whether it was providing accurate or useful results for leaders to act upon. Satisfaction comments can relay elements of what an individual liked or didn't like about a service or experience yet meaning and emotions play a key role in the contextual analysis of that experience (Sword et al. 2017).

The subjective nature of an experience itself coupled with the "inexhaustible capacity of language to describe affect" makes analyzing such comments using content analysis fraught with gaps in evidence (Sword et al. 2017:4). In the following comment, "beg" and "blown off" illustrate how adverse the lack of communication and concern were for this parent:

> The doctors and some nurses were amazing. We were scheduled for surgery @ 3:45 pm yet didn't have surgery until 7:30 pm. We had to beg for information. We were blown off a few times until a wonderful nurse, not assigned to us, took over and stayed with us. The surgeon was great.

Yet, when coded, these striking metonyms are hidden by bland terms such as "wait times", "communication", and "addressed concerns", which leaves out the powerful words that could be the motivation and understanding a team needs to act. In the case of content analysis, the lifeworld is understood by the coders but is easily lost when translated into institutionally-understood categories.

Consequently, while an inductive analysis is useful, coding is also rife with subjectivity that forces the coder to prioritize organizational understanding over emotional complexity. This is not to say that content analysis does not have its place. If general but operationally useful categories of feedback are needed, content analysis would be perfectly suited to provide that information and content analysis provided some lifeworld-led concerns for further exploration. However, considering the requests from stakeholders that include asks for not only the *what,* but also the *why* and *how* they should improve, other methods like ethnography and direct input from patients and families are better suited to begin answering such questions. An analyst would be rightly hesitant to recommend that interventions should be created solely based on evidence from content analysis results.

The low yield of meaningful or hidden insights from both content analysis and text mining made the authors wonder, what was the underlying reason for these analytics requests? Since stakeholders felt that these were both valid requests, the question becomes whether monitoring and analyzing data simply because it is available is a good use of time and money. Given that end user reading can provide the highest likelihood of increasing KPI for the least amount of time and expense, leadership's belief that text analyses can provide meaning and context becomes more apparent. Based on identifying this assumption, the authors recommended that stakeholders should prioritize deep meaning and sensemaking rather than focusing energy on poor substitutes.

Ethnographic work found deeper meaning yet failed to strongly influence decision making

For pediatric populations, more understanding and integration of social, economic, and familial context than adult populations is required for clinicians to make recommendations that both treat the condition as well as satisfy the patient and their family. In fact, a study of diabetic children showed that they want to be more involved in their care—often with the support of a parent or guardian since diabetes care is complex. However, the degree of these children's interest in involvement varies, and they did not want a one-size-fits-all approach (Ekra et al. 2015). Regardless, clinicians are asked to operate within a system that seeks to standardize everything from treatments and clinical notes to safety measures. Until the entire operational context—from appointment to billing—can also incorporate lifeworld-led care, clinicians will continue to face obstacles that impede their ability to treat their patients, continuing the cycle of provider "moral injury", decreased staff engagement, and decreased patient satisfaction (Talbot and Dean 2018).

Ethnographic studies of the patient experience at Seattle Children's yielded useful and important results. Yet, since insights based in empathy and emotions encourage innovation and potentially fundamental changes to current business models, they are often seen as disruptive. In addition, patients and families respond through the context of their lived experiences, rather than through a "controlled experience" (i.e., objective feedback about singular elements of an encounter) that the teams would prefer. Indeed, the issues that present themselves from ethnographic work often require teams to approach their work differently all while leveraging limited resources. When this occurs, there has been a tendency to dismiss the qualitative evidence for various reasons, not dissimilar to what ethnographers in other industries have experienced, such as n-size, sampling strategy, questioning the validity of interview questions and responses, arguing the information is too complex to be useful or suggesting that the issues are too large to tackle (Dourish 2006; Flynn & Lovejoy 2008; Ladner 2014). Tellingly, these ethnographic projects have been least successful when the findings do not have a direct impact on immediate business goals, such as findings that are difficult to implement immediately, such as changes to the medical records system. Additionally, teams might not be prepared to act on the data. For example, a team was prepared to address patient expectations but would need administrative support and escalation to address technology issues. Finally, ethnographic findings struggle to take hold when new organizational initiatives supersede the initial request (Briody 2018). A bias towards quantitative evidence is implicit, with its clean, easily understood format that can be neatly placed in improvement work and measured (regardless of the reality underpinning that belief), despite end users continuing to ask for *meaning* in analytic results. Yet, ethnographic findings provide the innovative solutions that will be necessary to reorient healthcare's service delivery model.

Data analysis requests in silos can lead to poor insights

While the content analysis suggested that Cascade was better at meeting lifeworld-based needs through interactions with staff, setting clearer expectations, and sending a follow-up note to the patient after surgery, the leadership team also cited several recent changes—like putting up a surgery progress monitor in the waiting rooms and having volunteers available to stay with families—that they felt had improved their score (see Table 2). This perception led the analysts to initially believe that *perception* of wait times was better at Cascade because of these changes and was therefore increasing their KPI scores. In fact, Olympic surgery center leadership has been adopting each of these practices based on this perceived success. A quick review of surgery start times found that Olympic's starts have been about 30 minutes later on average over the same time period as the comments, and that neither location shows any change in actual wait times. While Olympic does see a more acute patient population, this delay remained even when accounting for case acuity. In fact, Cascade and Olympic's start times trends only diverged for the *lowest* case acuity level. Thus, whether the other interventions were useful or not, Olympic has other issues that cause delays, perhaps separate from any actual or perceived improvement based on implementing Cascade's interventions. The implication here is that relying on survey comments alone can cause inaccurate interpretation and resources spent inefficiently.

Finding return-on-investment (ROI) likely requires exploring return-on-method (ROM) first

The case of sentiment analysis in particular shows some of the problems associated with these methods: text mining is fast and easy but inaccurate, while content analysis is more accurate but takes considerably more time. In regard to themes, content analysis can only provide partial insights, leading to misinterpretation. While text mining is comparatively more objective, it struggles to find any but the most superficial meaning from this type of homogenous comment data and can even be outright wrong. Further, even presumably objective quantitative analyses in text mining are exposed to subjective whims of the analyst, such as the threshold at which to set word correlations, the lexicon used for sentiment scoring, or the number of latent topics to assign an algorithm to find.

When considering stakeholder needs and the complexity of their roles as leaders, highlighting the ways various methods may or may not inform their needs is critical to helping them make productive decisions. With the internal satisfaction survey, the data scientist was able to show stakeholders through simple examples how text mining the comments would *not* produce the requested results. Based on their shared experience addressing their two requests, the authors compiled a list of qualities each method contains, so that stakeholders can constructively consider what forms of evidence might be valuable to them (Table 6; Figure 8).

Table 6. Comparison of four approaches to eliciting meaning for text data.

Qualities of Methods	Text Mining	Content Analysis	Ethnography	Reading (by end user)
Fast	Yes	No	No	Yes
Degree of quantification	High	Medium	Low to none	Low to none
Probabilistic	Yes	Can be	No	No
Quantity of data	Large (10^1-10^9)	Small/Medium (10^1-10^4)	Small (10^0-10^1)	Small/Medium (10^0-10^2)
Ability to scale	Yes	No	No	No
Ability to automate to end user	Yes	No	No	Yes
Frequency of access	Continuous	1x/year	Dependent on length of fieldwork	Continuous
Ease of results interpretation	High	Medium	Low	High
Resource / time burden	Analyst	Analyst	Analyst + end user	End user
Lifeworld	No	Low	High	Medium
Insights	Doubtful	Some	Yes	Casual
ROM	***Low***	***Medium***	***High***	***Medium***

Table 6 (above) provides some specific attributes of each methodological approach, with particular attention to the qualities that make text mining seem desirable. Figure 8 (below) provides a more general sense of the trade-offs between these approaches; much like the "project manager's triangle", you can't simultaneously maximize all desired analytic qualities.

These heuristics helped the authors articulate qualities of interest for stakeholders, as there are clearly significant trade-offs in requesting certain attributes for analytic response. For example, text mining can be done very quickly, but will not likely provide useful insights for this type of data; when the spread of topics and themes is minimal (as is typical in survey-based free-text responses), content analysis will be more accurate yet more expensive than text mining. Comparatively, ethnography can provide the depth of knowledge and actionable results that stakeholders are usually looking for, but at the cost of considerably greater time and effort. Finally, having the end users read the comments can perhaps accomplish much of what leaders need to understand from that data source, albeit only if they can set aside a bias towards quantification and be honest about their own subjective contexts (Gilovich 1991; Croskerry 2002; Kahneman 2011; Mannion & Thompson 2014).

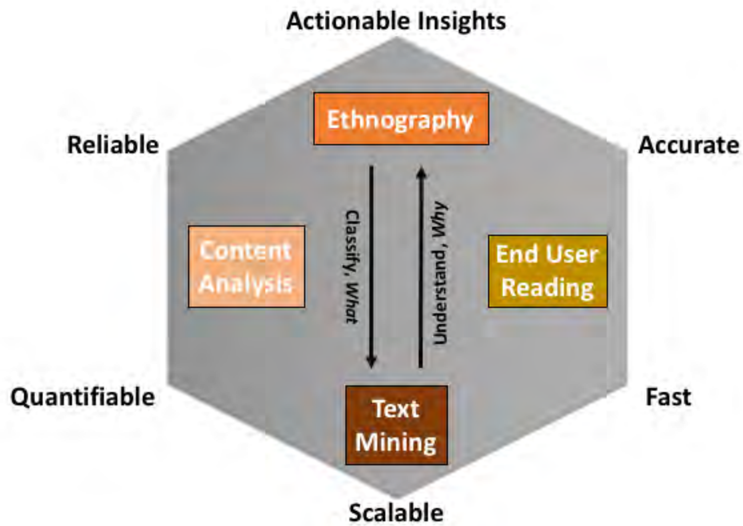


Figure 8. A conceptual summary of the tradeoffs between the various methodological approaches to understanding patient satisfaction comment data.

When the goal is to get a general sense for what patients, families, or staff are saying, the authors have recommended that the end users should just read the comments themselves. Given that they are content experts for their areas of leadership, they should readily understand how this source of feedback relates to the multitude of data sources they review. Because ethnography takes time and survey response analyses are not a high priority for that method given other hospital needs, having the end user read the comments provides the most insight (that's relatively gleanable from survey comments) for the least amount of effort, time, and cost to the organization. The caveat to this is that it does require a shift of the time and critical thinking burden back to leadership; if leadership instead delegates that task to other managers to find information within comment data, they would simply be replacing trained content analysis with untrained content analysis, where results could be left open to possible inadvertent bias, political spin, or siloed interpretations.

Data or information requests are often given to analysts in a vacuum, with little time or scope to seek additional information that could provide better context and understanding. When the results do not match what stakeholders expect, either duplicative requests are made to see if differences can extract more meaning, or the ability of the analyst is put in question. Thus, analysts and ethnographers alike are left grasping at straws to deliver results and evidence by a deadline. It is up to analysts and ethnographers to push back and leverage ROM to explain why certain sources of data or analysis methods will not provide the evidence stakeholders need. To usefully achieve ROI, analysts should hold stakeholders accountable for their requests and provide them with an understanding of the possible risks of such labor: wasted time and effort for both the analyst and the improvement teams, potentially misguided recommendations, and low staff engagement if/when these efforts fail.

*Re*Humanization: Lifeworld-led care and the search for meaning

While researchers of the lifeworld concept have focused on the interaction between medical care and the lifeworld, by offering linkages between these two sets of incongruent meaning systems or by suggesting implications for broad public health policy, very little is understood about the downstream effects of the lifeworld on operational, administrative, and business decisions (Barry et al. 2001; Lo 2010; Ekra et al. 2015; Dahlberg 2009; Hemingway et al. 2011, 2012). As society compels a shift from the traditional medical model of illness and treatment, to concepts such as personalized medicine, upstream impacts on social determinants of health and communities, as well as new sources of automated data tracking, lifeworld-led care must be increasingly recognized to ensure experience and satisfaction don't decline. Yet, the ability to communicate effectively with the lifeworld is impeded when the logics and rationality of the medical system (in this case, organization and hospital culture) are normalized and supersede other ways of communicating (Habermas 1984). Busy stakeholders may feel that they do not have time for the lifeworld, that a claim to statistical significance is enough to make decisions, despite the serious inferential problems with that approach (Simmons et al. 2011; Ruzzo 2014). Ethnographers are often asked to prove their findings by emphasizing the quantity of their sources whether in terms of number of individuals who were interviewed and contributed to data or the volume of data collected (Kelley & Buchanan 2017; Ladner 2014), so a bias towards frequencies and statistical significance in prioritizing information still prevail as key decision-making strategies.

Currently, the complexity of the US healthcare system and the long-ingrained decision-making hierarchy in medicine enforces behaviors that prioritize institutional knowledge over the experiences of patients, families, and staff, which in turn can lead to stagnant or decreased satisfaction. From the patient, family and clinician perspective satisfaction survey data is met with skepticism given that a week in the hospital with rotating care teams are distilled into a single satisfaction score. Experience and satisfaction analysis work, rather than taking the approach of KPI monitoring, needs to be grounded in the lifeworld concept, explicitly including empathy for patients, families, and staff. Human-centered methods that are focused on putting patients at the center of the process, have been recently introduced in healthcare, and may provide a lens through which to include the lifeworld (McCreary 2010).

Addressing patient experience and satisfaction requires dealing with samples of one as well as samples of millions. We can sometimes humanize single comments, and even occasionally topics and themes, by linking those with our own experiences and expectations. However, making strong inferences from small n's necessarily requires an approach that counterbalances big data with “thick data” while also maintaining a view of the lifeworld (Wang 2013; Habermas 1984). Transferring the lifeworld to ever-increasing amounts of text data, which by necessity become *de*humanized through aggregation, is a challenge for which there are not yet any algorithmic solutions. Arguably, there may never be. So *re*humanizing massive amounts of personal data may also require collaboration with ethnographers who can turn stories into evidence all while eliciting unstated assumptions and biases in stakeholders' beliefs to achieve the results leaders expect. Therefore, ethnographic approaches will need to provide windows into patients as well as leaders' lifeworlds.

To create sustained KPI improvement, addressing deep-rooted issues in the medical system are necessary—and ethnographic findings are perfectly situated to support such efforts. It is also possible that ethnographic insights in combination with text mining could eventually provide leaders with both meaning and scores, but this ideal has yet to be realized in any industry. In healthcare, understanding lifeworld-led care in the context of an industry that is increasingly paid for—and expected to be—in the business of service delivery will be critical to influencing long term experience and satisfaction outcomes.

CONCLUSION

This case study exemplifies how collaboration between a data scientist and an anthropologist helped test different methods to address the best way to tackle a tricky problem—improving healthcare satisfaction. Through discussions of their methods, they learned about the inherent biases of each other's approaches and can speak to a wider range of different disciplines' approaches to text analysis. The benefit of shared knowledge means that they can be clear from the beginning about what sorts of results stakeholders can expect to see, saving time and energy otherwise spent on analyses that may not yield the expected results.

This effort required that they also unravel their organization's cultural seams to understand why and how decisions are made, as well as how their work is impacted by organizational assumptions. Using an ROM approach, analysts can make informed recommendations to busy stakeholders and allow them to choose which qualities of data they are willing to compromise (or not). The recognition of where biases in data and analyses occur, as well as understanding the impact on insights, allowed the authors to carefully consider their recommendations to leaders, as well as their own approaches in developing those recommendations.

Analysts internal to an organization are typically expected to follow through on data requests with little-to-no say about the validity in doing so. While they may be aware of internal biases, it can be difficult to articulate their impact on day-to-day operations, especially to busy stakeholders who may see push-back as contesting their leadership. An awareness of organizational assumptions and prioritization allowed the authors to carefully construct a narrative that is less threatening to any individual (or the organization) but still highlights the strengths and weakness of each method. Consequently, the internal practitioner's role is not only to perform the requested assignments but to also advocate for the modes of evidence that will be most accurate, meaningful, and useful. Analysts are not victims of their stakeholder's requests—they must push back on bad assumptions and hold leaders responsible for their decisions in both developing KPIs and in making resource-intensive analytic requests.

As the healthcare industry moves towards a value-based care focus, requests for data around satisfaction and experience will only increase. Work meant to improve experience and satisfaction for hospital patients, their families, and employees must consider how each individual's contextually-grounded lived experiences inform their perceptions of any encounter. In short, the command-and-control methods used in healthcare to save lives will not likely be successful in improving experience and satisfaction; essential safety tools like checklists and standardization necessarily require prioritizing institutional support, language, and meaning (Conley et al. 2011) over the voice of the lifeworld (Habermas 1984). Data analysis that also prioritizes the system's biases towards statistical significance, data science hype, and reporting structure silos will similarly not provide useful evidence for improving experience and satisfaction. A shift from efficiency-based analyses for improvements to human-centered efforts will require organizations to better understand their patients and staff through carefully collected empathy and cautious, well-informed consumption of both qualitative and quantitative evidence.

Julia Wignall, MA, is an Anthropologist at Seattle Children's Hospital and a Qualitative Researcher at Seattle Children's Research Institute in Seattle, Washington, USA.

Dwight Barry, PhD, is a Principal Data Scientist at Seattle Children's Hospital in Seattle, Washington, USA.

NOTES

Acknowledgments – We'd like to thank Katelyn McLaughlin, as well as the rest of our patient experience and clinical analytic teams, for their contributions and permission to be part of this project. Chad Maxwell of Kelly Scott Madison; Aaron McAdie, Libby Montague, Andy Cooper, PhD, Ryan Sousa, Harris Baden, MD, Stephanie Hillman, Jenny Davidson, Joy Gehner and Jenn Scott of Seattle Children's Hospital; Tami Barry, PhD, of Western Washington University; and Evan Smith, DMA and Seattle Children's IRB all provided valuable feedback that improved this paper considerably. The views expressed in this paper are those of the authors, and do not represent the official position of Seattle Children's Hospital.

REFERENCES CITED

Azam, Kamran and Anwar Khan and Muhammad Alam
2017 Causes and Adverse Impact of Physician Burnout: A Systematic Review. *Journal of the College of Physicians and Surgeons Pakistan,* 27(8), 495-501.

Baer, Tamara and Angela Feraco, Selin Sagalowsky, David Williams, Heather Litman, and Robert Vinci
2017 Pediatric Resident Burnout and Attitudes Toward Patients. *Pediatrics,* 139(3), e20162163.

Balik, Barbara and Jim Conway, Lori Zipperer, and Joanne Watson
2011 Achieving an Exceptional Patient and Family Experience of Inpatient Hospital Care. *IHI Innovation Series white paper*. Cambridge, Massachusetts: Institute for Healthcare Improvement.

Banks, George and Haley Woznyj, Ryan Wesslen, and Roxanne Ross
2018 A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App). *Journal of Business and Psychology,* 33(4), 445-459.

Barry, Christine Ann and Fiona Anne Stevenson, Nicky Britten, Colin Bradley, and Nick Barber
2001 Giving Voice to the Lifeworld. More Humane, More Effective Medical Care? A Qualitative Study of Doctor-Patient Communication in General Practice. *Social Science & Medicine,* 53, 487-505.

Bodenheimer, Thomas and Christine Sinsky
2014 From Triple to Quadruple Aim: Care of the Patient Requires Care of the Provider. *Annals of Family Medicine,* 12(6), 573-576.

Briody, Elizabeth
2018 The Woes of Implementation Practice: Getting Caught by the "Program of the Month". *Journal of Business Anthropology,* 7(1), 98-132.

Conley, Dante and Sara Singer, Lizabeth Edmondson, William Berry, and Atul Gawande
2011 Effective Surgical Safety Checklist Implementation. *Journal of the American College of Surgeons,* 212(5), 873-879.

Croskerry, Pat
2002 Achieving Quality in Clinical Decision Making: Cognitive Strategies and Detection of Bias. *Academic Emergency Medicine,* 9(11), 1184-1204.

Dahlberg, Karin and Les Todres and Kathleen Galvin
2009 Lifeworld-Led Healthcare is more than Patient-Led Care: An Existential View of Well-Being. *Medicine, Health Care and Philosophy,* 12(3), 256-271.

Dourish, Paul
2006 Implications for Design. *CHI 2006 Proceedings*. Montreal, Quebec, Canada.

Ekra, Else Mari Ruberg and Tora Korsvold and Eva Gjengedal
2015 Characteristics of Being Hospitalized as a Child with a New Diagnosis of Type 1 Diabetes: A Phenomenological Study of Children's Past and Present Experiences, *BMC Nursing* 14(4), 1-10.

Flynn, Donna and Tracey Lovejoy
2008 Tracing the Arc of Ethnographic Impact: Success and (In)visibility of Our Work and Identities. *Ethnographic Praxis in Industry Conference Proceedings,* 2008(1), 238–250.

Gilovich, Thomas
2008 *How We Know What Isn't So*. New York: Simon and Schuster.

Habermas, Jürgen
1984 *The Theory of Communicative Action: Reason and the Rationalization of Society*. Boston, MA: Beacon Press.

Hemingway, Ann
2011 Lifeworld-led Care: Is it Relevant for Well-being and the Fifth Wave of Public Health Action? *International Journal of Qualitative Studies on Health and Well-being,* 6(4), 1-7.

Hemingway, Ann
2012 Can Humanization Theory Contribute to the Philosophical Debate in Public Health? *Public Health,* 126(5), 448-453.

Husserl, Edmund
1936 *The Crisis of European Science and Transcendental Phenomenology*. Evanston: Northwestern University Press (reprinted edition, 1970).

Institute of Medicine
2001 *Crossing the Quality Chasm: A New Health System for the 21st Century. Committee on Quality of Health Care in America*. Washington, DC: National Academies Press.

Ioannadis, John
2005 Why Most Published Research Findings are False. *PLoS Medicine,* 2(8), e124.

Kahneman, Daniel
2011 *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kelley, Beth and Jennifer Buchanan
2017 The Lifecycle of a Washing Machine: Transforming the Customer Experience for a Home Appliance Manufacturer. *Ethnographic Praxis in Industry Conference Proceedings,* 2017(1), 458-476.

Ladner, Sam
2016 Practical Ethnography: A Guide to Doing Ethnography in the Private Sector. Routledge, NY: Taylor & Francis Group.

Lagasse, Jeff
2017 As Chronic Illnesses Drive Up Healthcare Spending, Solutions Focus on Wellness, Interventions. *Healthcare Finance* website. October 24, 2017. Accessed [July 9, 2018]. https://www.healthcarefinancenews.com/news/chronic-illnesses-driving-healthcare-spending-solutions-focus-wellness-interventions

LaVela, Sherri and Andrew Gallan
2014 Evaluation and Measurement of Patient Experience. *Patient Experience Journal,* 1(1), 28-36.

Lo, Ming-Chen Miriam
2010 Cultural Brokerage: Creating Linkages Between Voices of Lifeworld and Medicine in Cross-Cultural Clinical Settings. *Health,* 14(5), 484-504.

Mahan, John
2017 Burnout in Pediatric Residents and Physicians: A Call to Action. *Pediatrics,* 139(3), e20164233.

Mannion, Russell and Carl Thompson
2014 Systematic Biases in Group Decision-Making: Implications for Patient Safety. *International Journal for Quality in Health Care,* 26(6), 606-612.

McCreary, Lew
2010 Kaiser Permanente's Innovation on the Front Lines. *Harvard Business Review,* 88(9), 92–97.

Nuzzo, Regina
2014 Scientific Method: Statistical Errors. *Nature,* 506 (7487), 150–152.

R Core Team
2017 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reichheld, Frederick
2003 The One Number You Need to Grow. *Harvard Business Review,* 81(12), 46-55.

Simmons, Joseph and Leif Nelson and Uri Simonsohn
2011 False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science,* 22(11), 1359-1366.

Sword, Helen and Marion Blumenstein, Alistair Kwan, Louisa Shen, and Evija Trofimova
2017 Seven Ways of Looking at a Data Set. *Qualitative Inquiry,* 24(7), 1-10.

Talbot, Simon and Wendy Dean
2018 Physicians Aren't 'Burning Out.' They're Suffering from Moral Injury. *StatNews,* July 26, 2018. Accessed [August 12, 2018]. https://www.statnews.com/2018/07/26/physicians-not-burning-out-they-are-suffering-moral-injury/

Wang, Tricia
2013 Big Data Needs Thick Data. *Ethnography Matters,* May 13, 2013. Accessed [September 9, 2018]. http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/

West, Colin and Liselotte Dyrbye and Tait Shanafelt
2018 Physician Burnout: Contributors, Consequences, and Solutions (Review). *Journal of Internal Medicine,* 283(6), 516-529.

Getting from Vision to Reality: How Ethnography and Prototyping Can Solve Late-Stage Design Challenges

BRADY SIH
Kaiser Permanente

HILLARY CAREY
Winnow Research

MICHAEL C. LIN
Kaiser Permanente

In 2014, Kaiser Permanente began implementing a next-generation medical office model that reimagines the outpatient care experience, combining new architecture, workflow, and technology to create a more convenient experience for patients and a connected, efficient experience for staff and care teams. As the first next-gen facilities were being built, challenges emerged as teams across a variety of disciplines attempted to translate the model's vision into reality. Teams were making design and operational decisions in parallel, without the ability to see how their decisions impacted the overall user experience.

To resolve these challenges, our innovation team at Kaiser Permanente used a hybrid make-and-observe method of prototyping and ethnography. Employing a co-creation mindset (Bødker and Grønbaek 1991), we engaged staff and patients to help us bring the future state of these next-gen clinics to life in a minimally viable way. We created a full-scale prototype experience at an existing clinic during working hours that included functional proxies of the future space; iterative techniques that simulated automation, i.e., Wizard of Oz-ing (Kelley 1984); rewritten workflows; and recast staff roles. With this model we were able to mock up the future user experience, test assumptions in the field, surface latent needs, and resolve unanticipated conflicts.

This case study provides an example of how ethnography, in combination with working prototypes of the future, can be essential tools for clarifying radical changes to new services. Using these methods, we were able to provide leaders with tangible evidence of why some elements of the initial vision needed to change in order to ensure a better user experience. Here we outline lessons learned and discuss limitations to our approach.

BACKGROUND

Kaiser Permanente is a leading health care provider in the United States, currently serving 12.2 million members in eight states and the District of Columbia. Kaiser Permanente is an unusual U.S. health care system, as we are made up of three distinct but interdependent entities — a not-for-profit health plan and hospital system, plus a network of regional medical groups — that work together to deliver care.

This case study describes a series of experiences that our innovation team at Kaiser Permanente worked on from 2015 to 2016 as the organization implemented its next-generation medical office model. This model influenced the architecture, created new workflows and roles, and leveraged new technologies to deliver an entirely different and personalized outpatient experience.

Typical in many "reimagining" projects, ethnography and design were used in the initial vision-setting phases of this multi-year initiative. Once the vision was set, functional teams (staffing, IT, architecture, facilities) worked in silos, and decisions were made by experts, without the aid of empirical, observation-based insights. However, as we implemented a new model for medical operations in a new building, these teams did not have enough experience with the new model of care described in the vision to make confident decisions, so they were open to an ethnographic approach to implementation.

Our approach involved creating full-scale, low-fidelity mock-ups of spaces, as well as rehearsing in an actual medical building during normal business hours, so that we could test real users completing real tasks and get immediate feedback. This combination of technological prototypes and ethnography in context is similar to the "Living Labs" that Pierson and Lievens describe as "to embed complex product ideas and prototypes in an environment that resembles as much as possible the context and everyday life setting." (Pierson and Lievens 2005) We used this make-and-observe approach to understand how best to use new technology to support the envisioned user experiences. These rehearsals allowed us to observe and predict what user behaviors might be in the new spaces before expensive systems and structures were finalized. While this robust ethnographic prototyping cannot fully predict what the user experience will be in the actual space, our work led the organization toward the most user-centered and efficient way of delivering care within a vision that was already established. Our application of prototyping and ethnography successfully catalyzed the architects, builders, building owners, and tech teams who were poised and ready to act on our findings immediately.

This paper focuses not on how we used ethnography at the conceptualization phase, nor to investigate or redefine the fundamentals of health care delivery, but rather our application of ethnography to refine the implementation of a previously agreed upon vision. We present this case as an example of how a make-and-observe approach can resolve gaps in understanding and disagreements around next steps when bringing to fruition a dramatically new vision of a complex service.

SOLVING LATE-STAGE DESIGN CHALLENGES

When our innovation team engaged with the project, the first facility's construction was nearly completed. Within the constraints of not being able to change the size and shape of the space, we identified three discrete experiences where combining ethnographic research with prototyping was essential to resolve questions about how to deliver on the patient and care team experience outlined in the initiative's vision:

- **Patient check-in:** The patient declares that they have arrived for an appointment and payment is exchanged if needed.
- **Patient waiting:** Because appointments rarely start precisely on time, the patient is invited to enjoy the space, indoors or outside, until the nurse informs them that they are ready to bring them into the exam space.
- **Care team coordination:** New workflows and technologies needed to be identified and designed to enable care teams to work in the new collaborative open spaces and shared exam rooms.

A New Patient Check-in Experience: Surfacing Incorrect Assumptions

The vision of the next-gen medical office was one in which patients never had to wait in line. For our leadership team, this felt like an emblem of modern, user-centered service delivery. They imagined patients walking through the door, wandering toward an inviting setting, and then being greeted by a roaming receptionist with a tablet computer, who could check them in at the patient's convenience. Based on this assumption, no space was allocated at the entrance for a queue.

To understand the future vision for check-in, we followed a familiar set of steps we now typically use in these projects:

1. Observe the current user and worker experience.
2. Bodystorm with minimal props and in a simulation space (Oulasvirta 2003).
3. Transform a current, working medical facility into the format of the future buildings. We brought in new furniture, signage, and technology to the building and coached the staff to work in new ways.

Figure 1. Bodystorming in a simulation space where staff could fail without worry of being seen by customers. Photograph © Kaiser Permanente, used with permission.

In this way, we practiced new roles and explored multiple solutions before we worked up to pressure-test our best ideas with staff who had urgent jobs to do and patients who were most concerned with getting to their appointments on time.

Our testing of multiple ideas revealed a clear conclusion: When there is an urgent task at hand, such as getting to the doctor on time, patients felt more comfortable in queues. When patients entered the facility, they were primarily worried about missing their time with their provider. They wanted the efficient, fair, and orderly aspects of a queue. Instead of removing the queue, we explored ways to enhance it. In the end we defined new roles for service reps, new supporting furniture pieces, and new architecture layouts. All of these elements helped to clearly indicate to patients where check-in can happen within the new and different space.

As we delivered the findings about the new features and functions of the check-in experience, those in leadership who were planning openings for larger facilities grew concerned about scaling the designs with many more patients moving through the entrance. To address these concerns, we repeated our methodologies at three locations where patient check-in volumes were 100 patients per hour instead of per day. Through our experiments we proved that the model could scale successfully with some modifications to the facility space design, staff role assignments (how many), and the use of self-service technologies.

Untethering Patients from the Waiting Room: Identifying Hidden Barriers

An important aspect of the vision called for untethering patients from waiting rooms and engaging them with healthy activities in a common waiting area. To explore this concept, we temporarily took over a single-story clinic and offered different activities for patients to engage in. However, the untethering did not work. There was a hidden barrier preventing the patients from enjoying the full space.

Through our observations, we found that patients were uncomfortable leaving the clinic module for fear of missing their appointment. Despite our efforts to entice them away, patients stayed where they knew they would be called, expressing concern about missing the nurse. Additionally, the nurses expressed annoyance at not being able to quickly call patients from the doorway.

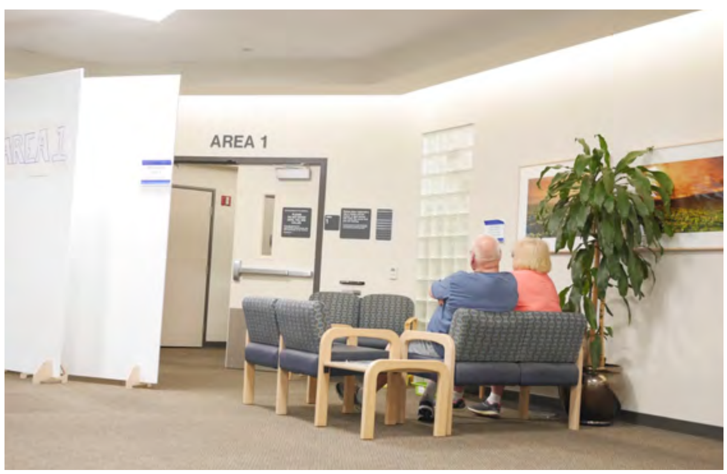


Figure 2. Patients sit close to the door, prioritizing waiting for the nurse's call over enjoying the amenities of the waiting area. Photograph © Kaiser Permanente, used with permission.

One proposal from our team was to give patients a sense of their wait time, so that they could feel confident about where and how to spend that time. If we could not do that, the entire vision was at risk, because patients would feel the need to stay close to the door. However, the idea of providing patients with estimated wait times was met with immediate resistance from leadership across many departments. Medical executives worried about publicly showing this information if a clinic fell behind schedule. Providers worried about delays being associated with their names. And everyone, IT included, thought it would be impossible to accurately calculate wait times in such a dynamic, unpredictable environment.

Yet our prototyped models spoke for themselves. Leadership could see the vision was not being achieved in the current state, so they agreed to let us pursue the concept. After brainstorms and concept iterations, we built several communication techniques for wait-time notifications, and clinical experts were able to draft a way of approximating wait times based on existing provider data that we could test in a live clinic. This allowed us to watch how patients responded to the information about wait time and ask how they felt about delays. Through three days of testing we demonstrated that:

1. Contrary to beliefs, it was possible to accurately calculate wait times, something that had never been done before at Kaiser Permanente.
2. Patients appeared to be less agitated by any delays in appointments when the information from the prototype was available to them. When we inquired about it, patients said that they knew appointments were likely to be delayed, so the proactive communication was a pleasant surprise. We observed people stepping outside to make a phone call or visiting the bathroom without the worry of missing the nurse at the clinic door.
3. Unexpectedly, providing wait times reduced the front-office staff's workload. Said one service representative, "We didn't get anybody asking us about how much longer until they will be called today!"

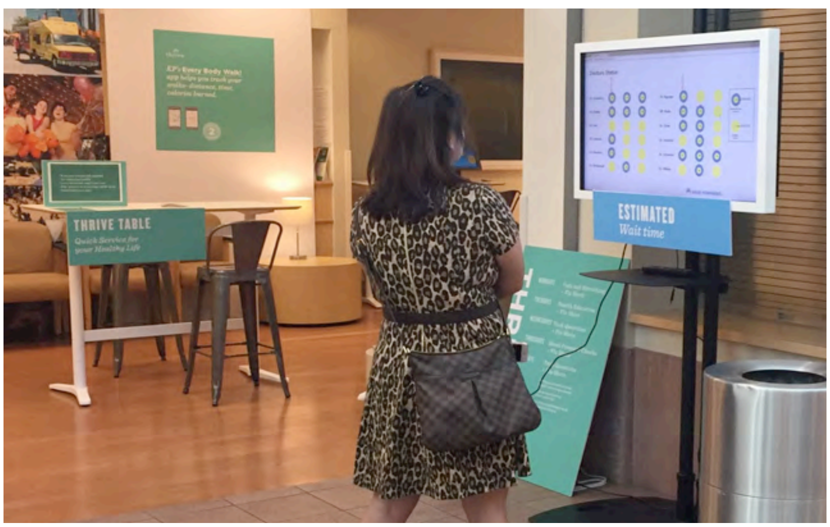


Figure 3. A patient checks her wait time on a prototype of a screen displaying an estimated time until they will be called to the clinic, for patients who have checked-in. Photograph © Kaiser Permanente, used with permission.

The ability to watch patients engage with a wait-time notification board and act differently because they knew they had more than a few minutes to wait convinced medical leadership to sign off on the concept. This was an exciting moment for us where we had built a prototype that seemed impossible and proved to leadership that patients valued it. From there, IT automated the algorithm we had created and continued prototyping to refine a full-fledged product that is now in use and standard equipment in all next-generation medical offices, enhancing the patient and staff experience.

Care Team Coordination: New Collaborative Tools

The third aspect addresses the way physicians, nurses, and other care team members work together. The new vision proposed a more efficient workflow and an open-plan back-office space, with shared exam rooms and shared support staff—a model completely different from the existing one—a less centralized model in which each physician has a dedicated office, exam rooms, and support staff. This new model was predicated on the use of technologies not yet built.

One tool we created through conversations and observations with staff was a working "outpatient dashboard" prototype. On each staff member's tablet computer, the dashboard displayed each patient's physical location once they were roomed in an exam room, what stage of the back-office process they were in, and how long they had been in that stage. During one field test in a modified back office, we observed the reoccurrence of an impromptu huddle where the staff referenced material on their tablet computers together and then quickly split up to tackle the work. Based on that observation and inquiry into that interaction, overnight the IT team created a larger touchscreen display of the outpatient dashboard. The next morning the care team immediately gravitated to it and conducted all their huddles at the new display, interacting and updating it as a team. This touchscreen display was not identified or defined prior to our field testing. It was created as a direct result of combining ethnographic methods with rapid prototyping and is now a standard component of all next-gen medical offices for Kaiser Permanente.

It was crucial to the development of this new technology that IT developers were on hand to witness the patient and care team needs. Otherwise this significant change in the technology roadmaps might not have happened. With first-hand observations of staff behaviors, they understood and were motivated to make rapid changes to their prototypes in order to improve the care team experience.

IMPACT AND VALUE IN A LARGE ORGANIZATION

The typical approach to implementing large-scale changes in our organization involves functionally specific teams (e.g., IT, facilities, operations) executing the vision over long periods of time. Implementing the next-gen medical office model required a different, more integrated approach, especially during the later stages, where these teams were racing to complete their deliverables on time and on budget for the opening of the new clinics. When teams did not have insights into what other teams were doing, it created blinders for each team. The impact and value of combining ethnography and prototyping not only helped remove these blinders but also enabled leaders to refocus their teams on the most critical issues. Simultaneously it accelerated adoption of new ways of working because staff could preview it in action and watch patients respond. Essentially, we moved the conversations from conjecture to observations.

Influencing Leadership Decisions

In "Design as Sociopolitical Navigation," Clark describes the importance of action-oriented techniques that engage decision makers in order to influence leadership's perspective on the needs, scope, and funding of a design project (Clark 2007). We carefully engaged the extended executive teams across functions at strategic times throughout the project so that they could observe the developing prototypes in action and offer input and ideas about how to evolve the idea. This was essential in informing their visions and changing their minds late in the implementation process. A senior vice president in our IT organization described the value this way: "You can make bad assumptions about what people need until you see it in action and find out you're very wrong."

Leaders could see what worked, what didn't work, and what was missing. They asked questions and proposed new solutions based on the evidence we presented. We tested those new ideas so that they could see the results for themselves. And together we further refined the concepts. These observations as evidence helped these leaders shift their thinking. In this way, the value of ethnography within a prototype is not just for the researchers, designers, and engineers working through the iterative cycles of testing concepts. Perhaps more significantly, the crafted performance of real patients and in situ staff conducting real work with mock-ups of space, tools, and roles provide the ability for leaders and stakeholder teams to witness first-hand how their systems would come together in the future (Halse and Clark 2008). This helped us to deliver on our goal of shifting previous planning conversations from "subject matter expertise" to observation-based discussions of how patients and staff were responding to proposed solutions in use.

Acceptance and Adoption of Change

People respond better to changes in workflow when they are engaged in it. Our process inherently did this by involving all impacted roles in the design of their work (responsibilities, technology, spaces, workflows). By engaging providers and clinical staff in the process, they took ownership of the solution and championed the new technologies, facilities, and workflows to their peers across the region. The old way of working was no longer a viable option for those involved in our projects. An administrative leader described her amazement at the change with this statement: "The transition from 'We can't do it.' to 'Now I don't want to give it up.' was a great surprise."

In addition to changing mindsets about the model, our approach also paved the way for smoother medical office building openings than ever before. After the opening of the first three new facilities, an executive director from the Facilities Planning organization remarked about the ease and reliability of the new systems: "The challenge in the past [with new facility openings was] that technology didn't quite work as planned... Because of the field testing, it works exactly like we intended."

Through our experiences experimenting in live clinics we understood the pain points, bottlenecks, and biggest obstacles to change. With that knowledge we were able to conduct staff trainings based on a strong understanding of the new buildings, technologies, workflows, and roles. One care team commented on their first day of operations that it felt like they had been practicing there forever because the training simulations had been so realistic.

CONCLUSION

As innovation transforms healthcare and many other industries, more and more technologies and services are being tried. However, in creating something to make a difference we often end up creating something that actually makes people's jobs harder because we are not able to see how the product actually fits within the context of the environment. And in late-stage development, every issue and question, from high level needs to detailed needs, become equally critical to the completion of the project. We can get so focused on meeting timelines and budgets that we don't have the opportunity to observe and assess the entire experience and discern what is important.

What this project has shown us as an organization—and as design and ethnography practitioners—is that with a growing emphasis on merging digital and physical experiences, there is also a growing need to use ethnography and prototyping to ensure user needs are front and center at every stage of facility design and implementation. These methods to study user experience are particularly important during implementation, when multiple systems are being resolved and information is needed about how they will influence each other.

We propose that teams leading capital and construction projects consider using our method of usage-focused ethnography and full-scale prototyping of future experiences in order to make critical adjustments to space, workflow, and technology as the project develops and before it is too costly to make changes.

Brady Sih is a Kaiser Permanente IT principal consultant who focuses on technology experience, design, and delivery. brady.sih@kp.org

Hillary Carey is an independent design researcher and innovation consultant focused on service design and systems change. Hillary@WinnowResearch.com

For a decade, **Michael C. Lin** helped Kaiser Permanente develop its health care design practice, contributing original knowledge to the field of service design. Mike is currently vice president of innovation at Centura Health. mikelin@centura.org

NOTES

We would like to acknowledge the Kaiser Permanente leaders who made this work possible. Jodie Lesh, senior vice president, National Delivery System Strategy, Planning and Design, developed the vision for the next-gen medical office, and also supported our work throughout, including our use of ethnography and prototyping in late-stage implementation. Wendy Lee, senior vice president, and Ann Sherry, executive director, Digital Corporate Services & Solutions championed the innovation team's work—and gave us the bandwidth and resources to do it. We would also like to thank Shannon Coughlin, communications manager, for her editorial guidance on this case study. Please note: The authors of this case study represent their own perspectives and are not representing the official position of Kaiser Permanente.

REFERENCES CITED

Bødker, Susanne and Kaj Grønbæk
1991 "Design in Action: From Prototyping by Demonstration to Cooperative Prototyping." In *Design at Work: Cooperative Design of Computer Systems*, edited by J. Greenbaum and M. Kyng, 197-218. Hillsdale: Lawrence Erlbaum Associates.

Clark, Brendon
2007 *Design as Sociopolitical Navigation: a performative Framework for Action-Oriented Design*. Mads Clausen Institute. Sønderborg, University of Southern Denmark. PhD dissertation.

Halse, Joachim and Brendon Clark
2008 "Design Rituals and Performative Ethnography." *Ethnographic Praxis in Industry Conference Proceedings* 2008:128-145.

Kelley, J. F.
1984 "An Iterative Design Methodology for User-friendly Natural-language Office Information Applications." *ACM Transactions on Office Information Systems*, 2 (1): 26-41.

Oulasvirta, Antti, Esko Kurvinen, and Tomi Kanjaunen
2003 "Understanding Contexts by Being There: Case Studies in Bodystorming." *Personal Ubiquitous Computing*, 7: 125–134.

Pierson, Jo, and Bram Lievens
2005 "Configuring Living Labs for a "Thick" Understanding of Innovation." *Ethnographic Praxis in Industry Conference Proceedings* 2005:114-127.

Suchman, Lucy, Jeanette Blomberg, Julian E. Orr, and Randall Trigg
1999 "Reconstructing Technologies as Social Practice." *The American Behavioral Scientist*; 43 (3): 392-408.

Humans Can Be Cranky and Data Is Naive: Using Subjective Evidence to Drive Automated Decisions at Airbnb

STEPHANIE CARTER
Airbnb

RICHARD DEAR
Airbnb

How can we build fairness into automated systems, and what evidence is needed to do so? Recently, Airbnb grappled with this question to brainstorm ways to re-envision the way hosts review guests who stay with them. Reviews are key to how Airbnb builds trust between strangers. In 2018 we started to think about new ways to leverage host reviews for decision making at scale, such as identifying exceptional guests for a potential loyalty program or notifying guests that need to be warned about poor behavior. The challenge is that the evidence available to use for automated decisions, star ratings and reviews left by hosts, are inherently subjective and sensitive to the cross-cultural contexts in which they were created. This case study explores how the collaboration between research and data science revealed that the underlying constraint for Airbnb to leverage subjective evidence is a fundamental difference between 'public' and 'private' feedback. The outcome of this integrated, cross-disciplinary approach was a proposed re-envisioned review flow that clearly separates public and private-to-Airbnb feedback with a single binary question. If implemented, it should allow Airbnb to collect additional evidence from hosts that can be utilized to make automatic decisions about whether guests need warnings or whether they have met an exceptional quality bar for a potential loyalty program.

SETTING

"Would you recommend this guest to other hosts? Describe your experience."

These are the first two questions of the review flow for hosts that has existed on Airbnb since January 2011. At that time, Airbnb had only been around 3 years and had around thirty thousand listings and under a hundred thousand guests who had stayed at Airbnbs. A review system was designed for the purpose of highlighting issues that occurred during stays and establishing a fair dynamic between hosts and guests when they review each other. Every time someone stays on Airbnb, the host of the place reviews the guest and the guest reviews the host. That information is aggregated and displayed for all future potential guests and hosts to see; together, it forms Airbnb's reputation system.

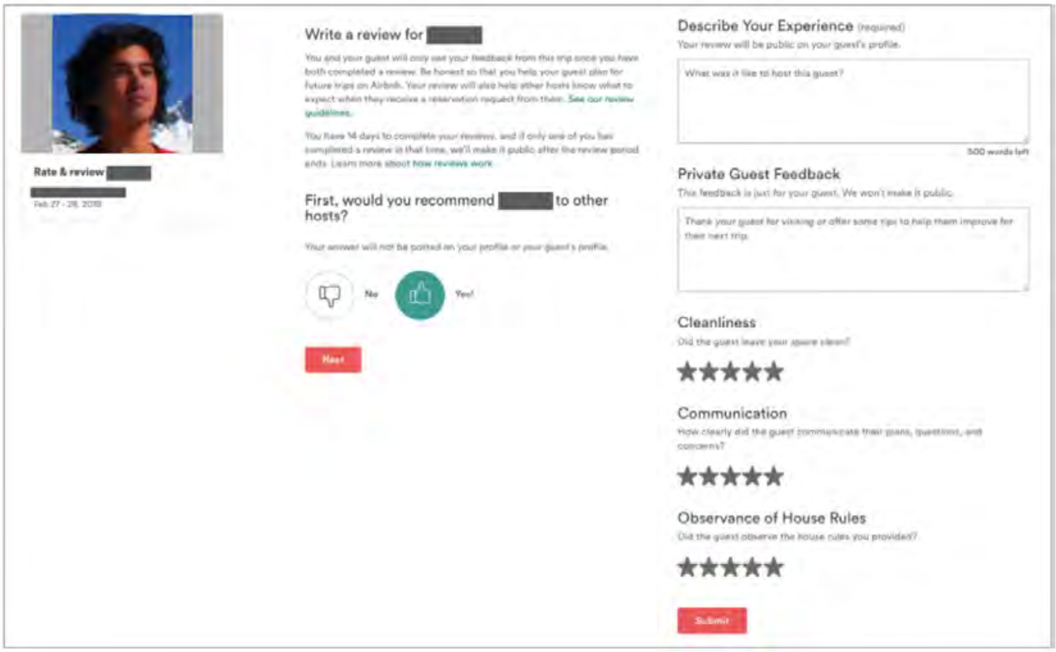


Figure 1. Existing review flow for a host to review a guest.

Nearly eight years later, Airbnb has grown into a community that generates millions of stays each year. People can also book hosted experiences around the world. Airbnb now has over 5 million listings worldwide, in more than 191 countries, across 81,000 cities. There are over 400 million guests who have stayed in Airbnb's apartments, villas, B&Bs, treehouses and many other types of inventory. There were 3 million people who stayed in Airbnbs the night of New Year's Eve (2017) alone. Each day, over a hundred thousand guests get reviewed on Airbnb.

At this scale and maturity, there are two very interesting business challenges related to the guest community.

1. How might we automatically identify which guests need to be warned about poor behavior? Airbnb has a high standard for the quality of its guest community. Yet, in 100 million stays, even a tiny (<0.1%) rate of poor guest behavior becomes a hit to our community and we take that very seriously. We want to do whatever we can to prevent guests and hosts from having less than perfect experiences. Manually investigating potential issues raised in reviews to identify guests that should receive a warning becomes a heavy operational cost.

2. How might we make fair and automated decisions of which guests would qualify as 'exceptional' for a guest loyalty program? Airbnb is exploring the idea of a guest loyalty program and identifying truly exceptional guests (according to the hosts they've stayed with) would be a valuable component of this program. In a community of over 500 million guests arrivals all time, we are at a scale where we need to be able to make these decisions in an automated fashion.

These two business challenges both rely on the evidence of guest behavior that hosts provide in the review flow after a guest stays with them.

However, the idea of using reviews as evidence for loyalty program rewards or to issue guests warnings presents unique challenges. Hosts and guests may perceive the same events differently, so making a decision using host reviews means letting algorithms make automatic judgements based on evidence that is inherently subjective.

Our mental model for how to differentiate guests was limited. Imagine the plight of a traveler if she lost her status at a hotel chain because of a clash in personality with the front desk receptionist, or wasn't able to stay at that hotel again because of a mismatch in communication style. At Airbnb, this challenge is multiplied by the cross-cultural and cross-language nature of the interactions between our hosts and guests, and our own biases as English-speaking Americans designing systems for global interactions.

Figure 2. The original mental model for how we might be able to differentiate guests using ratings & reviews.

ACT I

To use subjective, human evidence of host reviews in potential automated decision-making systems, we knew we'd need holistic research across multiple disciplines (data science, research, etc.). Our first research question was a natural one: how well does the current review system work?

In technology companies, research and data science both typically have a few standard approaches for beginning investigation into a new problem space: in this case, it made sense for research to begin with interviewing hosts and guests as well as reviewing inbound feedback to understand the nuances of the experience of the current review system, and data science to follow these learnings with an opportunity analysis to estimate the scale of any potential user problems.

1. Research Methods: 1:1 interviews & review of in-bound feedback

Research began with one of the standard approaches to understanding a complex problem space: guiding open-ended discussions on the topic with groups of stakeholders (in this case, both guests and hosts) while also looking at pre-existing in-bound feedback about the review flow (in this case, reports submitted through a pop-up widget on the review flow).

We learned that hosts were uncertain about the 'right' way to review guests; they were applying different norms to how they shared feedback. For example, one host said, "If I feel disrespected, if rules weren't followed, I'll share everything publicly. It's my home." Yet another host said, "Recently someone kind of conned me, I didn't review him because I didn't want him to review me, I felt blackmailed." Sometimes the uncertainty of the 'right' way to review became such a barrier that they did not leave any review at all which left Airbnb with an unreliable, incomplete picture of people's experiences.

We also heard from hosts that providing feedback could feel repetitive and there was a desire to reduce the cognitive overhead of free text reviews. As one host said, "It's too time consuming and confusing. I see the same things again and again. If I could just select those things it would save me time and effort." We saw an opportunity to introduce structured options to simplify the flow.

2. Data Science Method: Opportunity analysis

Opportunity analysis is about taking an intuitive notion of a problem and quantifying 'why should we work on this?'. In the interviews, we heard concerns about the uncertainty, inconsistency, and cognitive overhead of the review systems, so if we wanted to use these reviews as evidence for automated systems we needed to understand how often these problems were occurring. The first place to look was the aggregated review data.

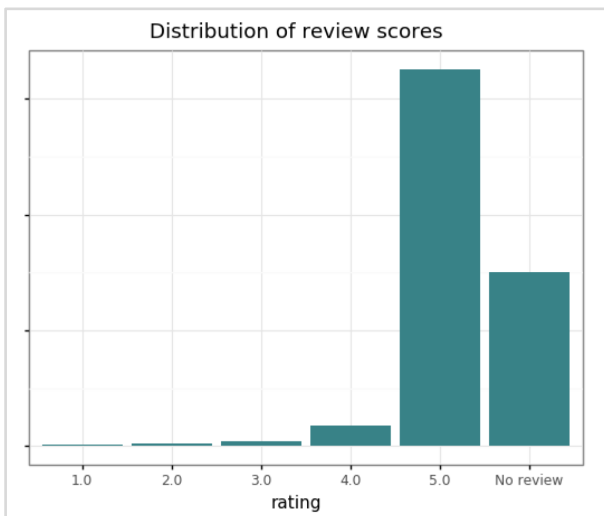


Figure 3. Ratings of guests left by hosts skew largely to 5 stars (1 to 5 stars, 5 is the best).

The vast majority of reviews were 5-star, and a substantial portion of guest stays also had no review at all. This was a challenge -- how many of those stays with no reviews were actually "less than ideal" stays, where the host felt uncertain so they left no review at all? To use statistical terminology, the 1:1 interviews made us pretty sure that these missing data were not missing at random, rather they were intentionally not completed. When comparing to 'guest reviews of stays' (the inverse type of review in the system), previous research showed that of the similar percentage of stays were left unreviewed. Follow-up research on the unreviewed stays indicated that indeed there was a portion who didn't review because they had a less than ideal experience but they didn't want to damage the hosts' reputation.

While the dynamic of guest versus host reviews are a bit different (hosts generally have more at stake), we expected that there was probably some similar behavior happening on the host side of things.

Furthermore, it was interesting that where reviews were left, the vast majority were 5-star. Research has shown that the extremely high ratings of hosts on Airbnb leads to loss of informative value for the guest; a host's reputation has a subsequent diminished effect on things like listing price of booking likelihood (Ert, Fleischer & Magen 2016). Yet, we knew that even though reputation systems are wildly inflated, they do matter in decision-making. Research has shown that reputation systems can significantly increase the trust between dissimilar users and there is an inverse relationship between risk aversion and trust in those with positive reputations. Having a high reputation is actually enough to counteract homophily. Specifically, research on 1 million requests-to-stay by guests on Airbnb data has shown a higher tolerance for individuals at farther social distances between guests and their selected hosts as the reputation of the host got better (Abrahao, Parigi, Gupta & Cook 2017).

At this point, we weren't sure exactly how to interpret the reality behind the inflation of the review ratings, but we would soon realize this was a hint of a deeper, fundamental problem of human psychology.

The combination of concerns from the qualitative research and the quantitative snapshot of the problem's scale made us concerned that the evidence about guest behavior collected in the current host review flow might be unreliable or incomplete. At an industry level, we knew review systems were imperfect but we wanted to see if we could go beyond face value and get a better signal. We suspected we might have to change the review flow -- but how?

3. Hybrid Method: Human judgements of analytically-sampled reviews

To move from identifying a problem to researching solutions, we realized we needed to answer a deeper, fundamental question: 'what is the difference between a problematic and a perfectly reviewed guest?' In trying to answer this, we landed on a new, hybrid methodology that was only possible with the combined skills of our disciplines. The hybrid method began with the realization that our review data was a unique data set that included both structured data (the rating), and unstructured data (the review text).

We agreed that when working with subjective evidence, in this case a review by a host, the 'ground truth' of what the evidence means has to be a human judgement. In this case, we can make that human judgement by closely reading the unstructured review text. Our goal was to compare our human judgements of 'problematic' and 'perfectly reviewed' guests from the unstructured text data with patterns in the structured data of the star rating.

Because the vast majority of Airbnb reviews of guests were 5-star, data science suggested we focus on two particular patterns of the structured data:

1. 'Perfectly reviewed guests' -- guests with at least 10 reviews, all of which had a perfect 5-star rating.
2. 'Potentially problematic guests' -- guests with at least 10 reviews, of which at least two reviews were only 1-star or 2-star.

Data science identified guests that fit these patterns, and then drew a sample of the *next* review received by each guest. For both 'perfectly reviewed' and 'potentially problematic' guests, we had examples of their next review being another 5-star, or a 1-or-2-star. We could then separately apply human judgement to four categories of reviews.

		Current Structured Evidence	
		5 stars now	1-2 stars now
Historical Structured Evidence	All 5 stars in the past	*Unstructured Evidence:* **Good** reviews from **'perfectly reviewed'** guests	*Unstructured Evidence:* **Bad** reviews from **'perfectly reviewed'** guests
	Two or more 1-2 stars in the past	*Unstructured Evidence:* **Good** reviews from **'potentially problematic'** guests	*Unstructured Evidence:* **Bad** reviews from **'potentially problematic'** guests

Figure 4. Four categories of reviews were sampled for analysis.

We printed out a few thousand samples, divided them into the four categories, and manually read through hosts' review text one by one, hand coding our observations. The results of this process of comparing how structured evidence related to human judgements of what a 'perfectly reviewed' or 'potentially problematic' guest is surprised us.

When we compared the one star reviews from the 'potentially problematic' guests with 1-star reviews from the 'perfectly reviewed' guests, we found there were clearly two types of one star reviews: 'actually problematic guests', whom it was clear should receive warnings if not be removed from Airbnb altogether; and 'potentially unlucky guests', who happened to run into a careless incident or what sounded like an overly-sensitive host. In a great deal of cases, these two types of reviews both received a 1-star rating.

> Actually Problematic: "Hosts beware of ___! ___ and her 3 friends stayed in my place. She is a very rude, entitled and ungrateful person… They all showered, left the heating on and did not say one word of complaint. Two days later she contacts me demanding a full refund despite the fact that she and all her friends used my place. If she was not happy she could have left and I would have refunded her."

> Potentially Unlucky: "I'm afraid I cannot recommend these guests to other hosts. They are polite girls but their lack of respect and care put us and our home at very real risk of fire. Somehow it appears a towel was left over a lamp that was on....the towel burned through and the lamp fortunately just melted as it was fire resistant. They informed me of some damage, paid to replace items and apologised. It's an experience I wouldn't want repeated."

It was not a hard-and-fast rule that all ‘perfectly reviewed’ guests with a 1-star review were only ‘potentially unlucky’, but it was certainly more common. If we were to make automated judgements about whether guests were ‘problematic guests’ based on the structured data of only one or two reviews, we would be very prone to unfair decisions based on incomplete data.

This discovery with respect to the 1-star reviews was mirrored with the *5-star* reviews. The vast majority of reviews were 5-star, regardless of whether the text of the review suggested the guest was ‘consistently positively reviewed’ or ‘a truly exceptional, once-in-a-lifetime, would-invite-to-my-wedding personal connection.’ The ‘perfectly reviewed’ guests were somewhat more likely than the ‘potentially problematic’ guests to receive what we considered an ‘exceptional’ review, but it again was not clear-cut: many of their reviews were also just fine and wouldn’t put them in an exceptional category. Even some of the ‘potentially problematic’ guests also received what our human judgement considered were ‘exceptional’ reviews on occasion.

> Positively reviewed: “short but nice stay ... polite and nice guy.”
>
> Exceptionally reviewed: “We are very lucky that we could meet ___ and ___. They are very interesting and friendly couple and it was so much fun to be around them. We are already missing our conversations. We were so impressed to find our place so clean and shiny after they left. You can not ask for better guests than ___ and ___. Our only regret is that their stay was far too short. Ps. ___, I really enjoy reading your book.”

The fundamental problem had become clear: How can we trust this evidence of who is an ‘exceptional’ and a ‘potentially problematic’ guest in an automated system, if it takes so much of our nuanced human judgement to make these decisions? Since this problem was clear even among guests who had a long history of past evidence (10+ past reviews), we knew that for all our guests who had so far received only one or two reviews, there was simply not enough evidence to make a fair judgement.

Thanks to this hybrid research method, we now had a clearer mental model of the range of reviews in the system. Research has argued that the bonding power of the interactions have been diminished by the development of the online reputation systems in a sort of “disenchantment” created by technology (Parigi & State 2014). Despite this “disenchantment” amdist the inflated ratings, we still saw nuance and detail coming through in the review text. But to use the reviews we collected from hosts in automated systems, the reviews would have to provide evidence which could fairly distinguish between five different types of guest behaviors (below) -- and the current system barely even distinguished between two.

This led us to a fourth step in our methodology: designing prototypes of a new review flow that could provide more detailed and fair evidence, and putting these prototypes in front of hosts to gather feedback.

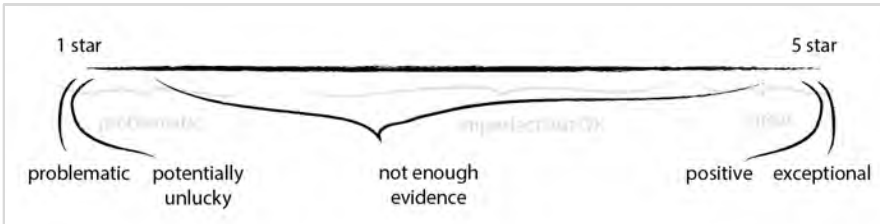


Figure 5. The updated mental model for how we might be able to differentiate guests using ratings & reviews.

4. Research Method: Prototype testing & participatory design

Applying our new understanding, we arrived at two key hypotheses for how we could prototype a better review flow:

1. Firstly, to help the star rating better reflect multiple types of guests (not just '5 stars' or 'not 5 stars'), we moved the point where we asked hosts for the star rating from up front, to the end of the review flow, after first asking hosts to relate the objective facts of their story. Our hypothesis was that this would lead to the star ratings being more spread-out, and thus better capture our more nuanced view of guest behavior.
2. Secondly, to further clarify the 'problematic' vs 'potentially unlucky' distinction, we added a question to ask how responsibly guests acted after any issues that arose. From the many samples we had reviewed before, we thought that any guest who responded responsibly after an issue had a high chance of being just 'unlucky', and not really a 'terrible' guest that shouldn't be on the platform. We also asked a question about how severe the issue was.

We designed two new interactive prototypes of the review flow that we believed would address the challenges of the existing review flow, and we invited hosts to share feedback.

Hosts were interviewed in pairs to stimulate discussion through disagreement and shared stories that would remind each other of their history of guest interactions. After discussing the key issues and reviewing the prototypes, we invited the hosts to share their ideas by drawing and explaining their own proposed review flow.

We realized were were way off track after our conversations with hosts. We thought we could create nuance and accuracy in the reviews through question wording, ordering and structured data capture, but there was a more fundamental issue at play: Hosts told us they were intentionally not sharing their true opinions of guests, because there's little incentive to review someone poorly. They inflate because of fear of retribution or a sense of guilt that they'd be individually responsible for any consequences to the guest (e.g. the guest won't get accepted in the future).

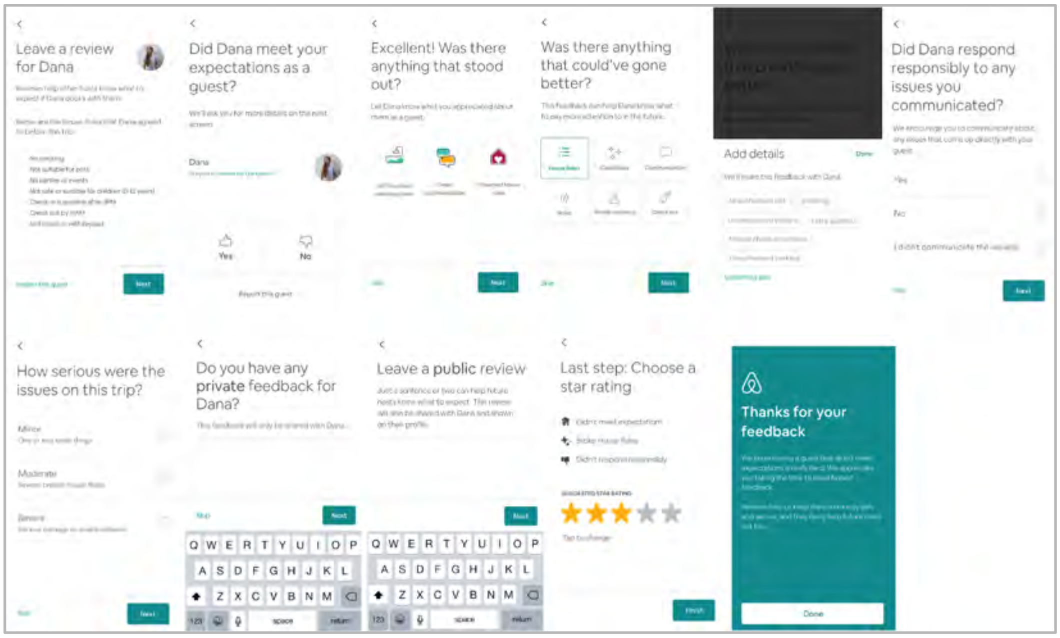


Figure 6. Screens from the interactive prototype of a mobile review flow designed for testing with hosts.

> "It's not the setup. It's about the guilt to say something nasty or unwillingness to grade or fear of retribution… The people I'd leave bad reviews for are the ones that might come back at you… and I don't know where they live."

> "I want to maybe note not to host them again but I don't want to ding her."

We began to form a new hypothesis: the underlying problem here is an intentional mismatch between hosts' public opinions and private opinions of guests. And if this is true, we can't solve it just by tweaking the order of review questions or asking them to go into more detail with structure content. We needed to design a system that would capture the intentional mismatch between public and private reviews.

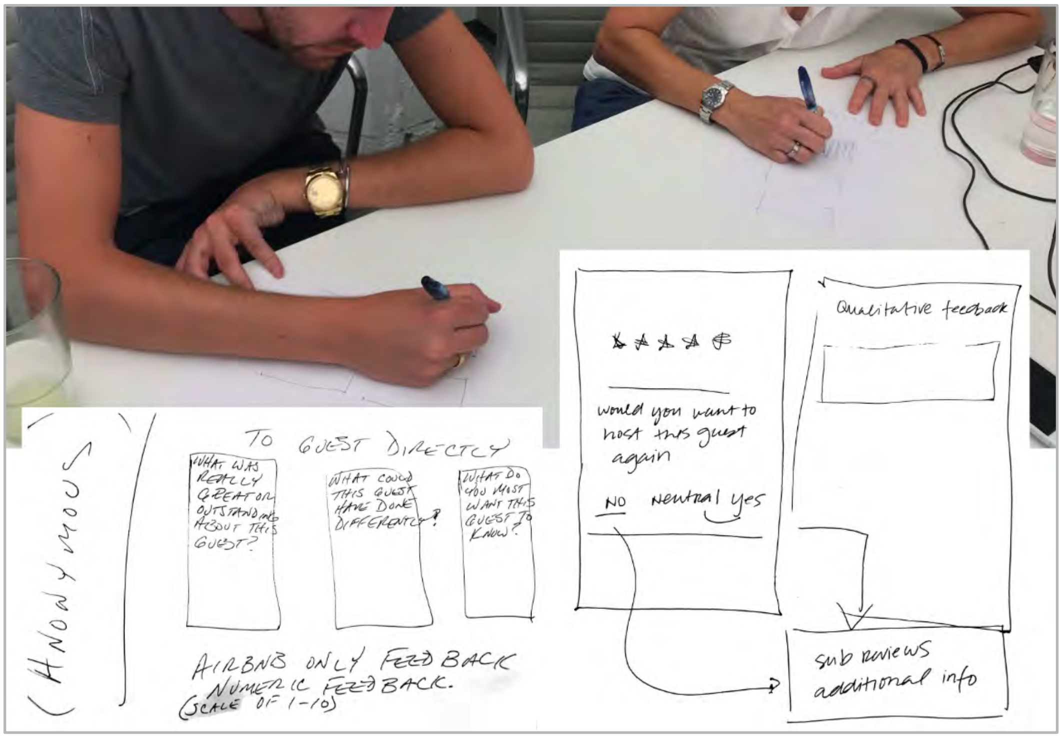


Figure 7. Pairs of Airbnb hosts sketch out their visions for an optimal guest-review flow.

ACT II

We had a hypothesis: review ratings were not consistent with our human judgements of what happened during the stay, because hosts were intentionally not telling us their true opinions. We needed a methodology by which we could investigate this hypothesis further, and more fully understand this mismatch in what hosts were saying publicly with how they sometimes really felt. At this point we were basing our hypothesis largely on the self-stated views of a handful of hosts. Yet this was a topic with so much opportunity for bias; we needed a way to tease apart the public-private distinction.

This is when we realized we had a data source we hadn't yet used: in the past, Airbnb used to give hosts an additional option when leaving reviews to also give private open text feedback to Airbnb. Even though we no longer ask this question, we still had past data we could utilize. Our hope was that if hosts were indeed often intentionally leaving a public review that didn't represent their true opinion, they would at least sometimes give genuine feedback in the private review to Airbnb. Because we had records of both public and private parts of the reviews, we could look for these mismatches.

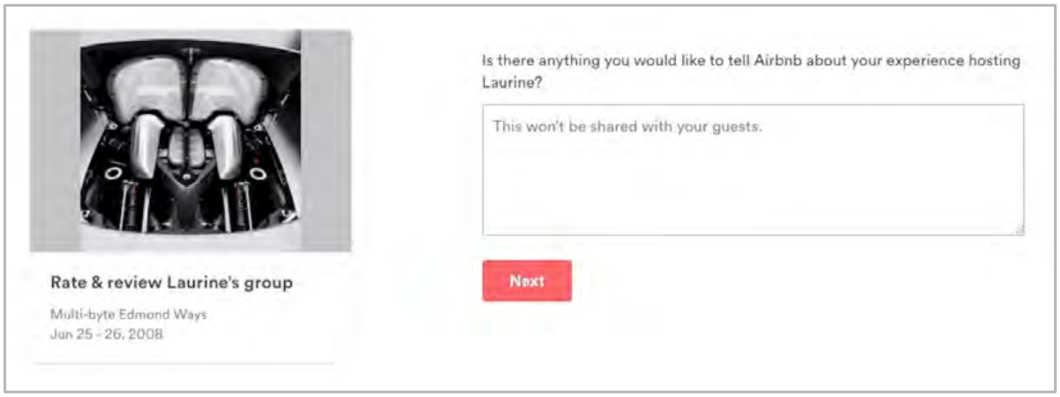


Figure 8. Final question in an old review flow that gave hosts the option to share private feedback with Airbnb.

5. Data Science Method: Sentiment classification

Our goal was to figure out if the mismatch between some hosts' public review and private feelings existed at scale in our review data. One data science technique was ideally suited to this challenge: 'sentiment classification', meaning to classify reviews as 'positive' or 'negative' based on the unstructured written text. If we could identify reviews as positive or negative based on the text alone, we could automatically identify cases where the public review was positive but the private review negative, and estimate how often this was occurring.

The trick to building any classification model, including a sentiment classifier for positive/negative text, was to have the right training data. In our case a unique opportunity presented itself, in that we had public review text that also matched to public review ratings. Our process was as follows:

1. Label public reviews that came with a perfect 5-star rating as 'positive', and public reviews that came with a 1-star or 2-star rating as 'negative'. In fact, we found we actually had to add a third category of 'unsure' to capture uninformative private reviews, such as when the host simply wrote 'No,' meaning, 'No I don't have any private feedback to give.'
2. Train a sentiment classifier on the public review text using the positive and negative labels. Because vast majority of the public reviews are positive, we downsampled the positive training data to create a balanced dataset. We used traditional sentiment classification methods where the public review texts are represented by the collection of meaningful words in them and their frequencies, thus the nuances of word order and grammar are ignored.
3. Confirm the accuracy of the classifier on a separate time period of public review data and ratings.
4. Use this classifier to classify the private review text (that lacks any rating label) as positive or negative. Then, count the percentage of the time that positive public text is written alongside negative private text. This is the 'public/private mismatch rate'.

The accuracy of the model when tested on the public reviews was 91%, with slightly higher accuracy at correctly identifying positive reviews (96%) than negative (83%). This result gave us high confidence that the model could correctly identify whether a review without a known rating label -- i.e. the private reviews -- were positive or negative.

The result we found was that there were enough reviews with 'positive sentiment' public review text that had 'negative sentiment' private review text that we knew the hosts were indeed not always willing to share their true feelings publicly.

Table 1. Stand-out examples of public-private mismatch

Public review	Private review to Airbnb
Easy communications and reliable guest. No problems	I was not there at the time of his visit but the cleaner reported the house to have been left "untidy" and with dirty dishes in the sink. They did pay a cleaning fee so perhaps they assumed that was acceptable.
[ ] is a decent responsible lady. The apartment was well-kept which is very important for us. I hope [ ] had a comfortable stay at ours.	Honestly I'd better not have such a guest next time… There was a plenty of dirty plates all over, shoes were worn inside the flat which is not common in Ukraine.
[ ] and her group were great guests. They were easy going and left the house nice and tidy.	They were a bigger party of people than expected... They failed to wash up properly leaving dried up food on plates, cups and cutlery. The oven was messy and I spend half hour cleaning it… Unfortunately, they didn't leave the house secure.
[ ] has a very friendly personality.	This woman is VERY HIGH maintenance! She has no boundaries or filters. One request after another, some of which were out of line. She would be more suited to stay in a hotel rather than someone's private home.
[ ] was very nice and clean.	[ ] asked to stay extra days and wanted to pay cash... He also completely ignored the checkout time and then needed to leave a lot of belongings at my place after he 'checked out' several hours later... He also is generally just a strange person who made me feel very uncomfortable and unsafe at times. I think he is probably nice but I would not host him ever again.
[ ]'s parents took care of my apartment. All went well. Thank you !	I won't rent to her again. Poor communication. Everything was very hard to deal with. First, I needed to give her the keys for the apartment. She was upset because no one was going to be there to receive them. She asked me to leave the keys at her house, which I did. She was still upset, no reason. After a few days, I asked her how things were going. She told me that Internet had been down for the last two days. I don't know why she wouldn't let me know right away. I called the Internet company and they told me that there was a problem in the whole building, not just my apartment. I shared the information with her, she was still upset with me. I ended up offering a very generous refund of 15% Still upset... I would define her as a crazy customer. Trouble maker.

To see that apparently-positive reviews actually had negative issues revealed in private was a striking discovery. Especially given that such a small percentage of reviews have negative public ratings -- this assured us that our public evidence of the quality of the guests was definitely underestimated.

This application of a data science modeling technique had succeeded: we'd proven that the evidence from the user interviews was showing up at scale. Intentional public/private mismatches existed, and with our quantitative estimate, we knew that these intentional mismatches should be the focus of our attention rather than the mere inaccuracy of the public review questions asked.

This also made us think back to our earlier observation that the vast majority of reviews being 5-star seemed too high. It was too high, because hosts weren't always publicly sharing what they really thought -- and thanks to modern computational methods, we had the evidence to both prove and quantify how often this was happening.

6. Research Method: Remote prototype testing interviews

The final step of research was a set of remote interviews with hosts to gauge their willingness to report honestly using a new private feedback method. We designed a prototype of a new review flow that simplified the number of questions asked (drawing out nuance publicly was superfluous) and included a final question that we had heard almost every host previously interviewed speak to in one way or another: "Would you host 'this guest' again?" Hosts repeatedly had said that they didn't want to publicly thumbs down some people but they really didn't want to personally host them again. We decided that was the perfect question to ask that would yield the most honest answer. The final screen in the new review flow asks hosts this key question and indicates it will not be shown publicly.

Figure 9. The final question of the new review flow asks for a private, binary rating; this should help unlock the public/private mismatch.

Hosts liked the addition of this question. They talked about the "gray area" where guests aren't terrible but also aren't great, and they would be more willing to share this honest assessment using the private question of 'hosting again.' Interestingly, they took as a given that if they said they wouldn't host someone again, Airbnb would never allow that guest to book with them again, even when the version of the prototype they saw didn't explicitly promise this.

The callout on the review flow that "Airbnb will only take action if multiple hosts are not willing to host this guest again" is intended to protect guests who have unfortunate, isolated clashes with hosts, whether they are cross-cultural or personality-based. The hope is that these issues won't repeat themselves if they are indeed isolated, so the guests won't be unduly punished for single offenses.

We believe this structured "would you host again" question will yield useful data that we can act on at scale because it will begin to paint the picture of our updated mental model of guests (problematic - potentially unlucky - not enough evidence - positive - exceptional). In the past, when we had a private feedback question, it didn't direct hosts to focus on the key question of whether they'd host this guest again. It also couldn't reliably be acted on at scale because it was a free form text box. If this new review question gets implemented, future data should indicate whether we are able to use the answer to this question, aggregated over many reviews, to identify guests that need to be warned about poor behavior as well as identify exceptional guests for a potential loyalty program.

OUTCOME

The outcome of this integrated, cross-disciplinary approach was an understanding of the underlying constraints behind Airbnb being able to leverage subjective reviews to make automated decisions. The fundamental difference between 'public' and 'private' feedback is at the crux of this challenge. With this new knowledge, we were able to re-envision a review flow that separates and utilizes the public and private components differently. If implemented, we hope this will allow hosts to share feedback without guilt or fear of retribution while Airbnb can still collect reliable structured evidence from them. This can be fairly used to make automated decisions that will enable two key business goals: identifying guests that need to be warned about poor behavior or identifying exceptional guests for a potential loyalty program.

Our process of getting to the heart of the challenge was not direct; it included mishaps alongside major 'aha' moments, and each step revealed something unique that was essential to informing the solution.

1. *Research*: Our first finding revealed the inconsistency in hosts' approaches to leaving reviews and their desire to reduce the cognitive load of lots of free text.
2. *Data Science*: An opportunity analysis showed that the majority of reviews had 5 stars which seemed even higher than we expected. Almost a third of stays had no review; we had no way of knowing if they weren't reviewed because the guest was problematic or some other reason.
3. *Hybrid*: We used structured data (the rating), and unstructured data (the review text) to try and classify what is a 'problematic' from a 'perfectly reviewed' guest. We realized that human judgement was required to distinguish 'problematic' from 'potentially unlucky' and 'positively reviewed' from 'exceptionally reviewed.'

4. *Research*: We designed a review system that focused on drawing out the nuance of a rating with structured questions but, in doing so, discovered we were missing the big picture; hosts were actually intentionally not sharing their true opinions of guests because of guilt or fear of retribution. No level of nuanced questioning would get them to be more honest so long as the review was shared publicly. The current review system wasn't capturing their private opinions, which sometimes included not wanting to host a specific guest again.
5. *Data Science*: After realizing the public/private mismatch was the crux of the challenge, we dove into old data from when we used to ask hosts for private open-ended feedback on guests. We used semantic classification to identify if reviews were positive or negative and we were able to identify that there was indeed a mismatch in a decent percentage of the apparently-positive reviews, which was striking, given how few reviews actually get negative public ratings.
6. *Research*: Finally, we landed on a new, simple review flow that asks an additional key question privately of hosts: "Would you host this guest again?" If implemented and then aggregated over many reviews, the new data should allow us to both identify guests that need to be warned about poor behavior as well as identify exceptional guests for a potential loyalty program.

DISCUSSION

Reaching this outcome was only possible through a deep synthesis of research and data science. First, we applied human intuition when reading review ratings which revealed that our judgements differed from the recorded data. We saw that the aggregated stats about this data has major consequences at scale. Second, in-depth user tests of a new approach led to a hypothesis that we had misidentified the underlying problem. Data science techniques reaffirmed that our new understanding of the problem was correct. Third, we confirmed that a new approach could solve this problem by scaling an in-person UI test to a statistically significant sample. In reviewing our process, we recognize two generalizable principles that could inform future collaborations where a synthesis of qualitative and quantitative methods is paramount.

Principle 1: Shared artifacts

The first principle is to look for sources of information that can be accessed by multiple disciplines each in unique ways. In our case, we found shared artifacts in the samples of reviews; the review text was both a record of nuanced and layered human expression, and a natural language dataset that could be analyzed at scale for semantic patterns and linked to quantitative ratings. It was applying our different expertise to explore this complex and unique dataset together that generated some of our most important insights.

Principle 2: See-sawing

The second principle we arrived upon is a process of iterating back and forth between our disciplines — we've started calling this see-sawing. Multiple times, one of our approaches seemed to reach a roadblock, but by switching focus to the other discipline a new way forward opened up. For example, our first quantitative analysis of ratings suggested this evidence was too abstract to be useful (almost all 5-star), but 1:1 interviews revealed the nuance that led to a new design to capture more accurate review data. However, when this design was tested with hosts, it seemed like we hit a different wall - hosts did not *want* to give us genuine feedback, and for good reasons. Yet, here a more quantitative way of thinking helped us move forward in seeing the public/private mismatch as a pattern that could be teased out, modeled, and classified. In retrospect, it's not always that the specific skills of the other discipline were necessary to move forward, but rather that the contrasting way of thinking helped shine a light on a new direction.

The conditions under which these principles are most useful are for problems that present puzzles of human psychology and emotion that vary greatly between individuals. In such cases, a purely quantitative approach is unlikely to reach useful conclusions through analysis, yet a purely qualitative approach will be limited in knowing how well individual stories and emotions can be generalized to be wider population. Such cases confront both disciplines with questions that they are ill-equipped to answer. In combination, however, research and data science display a remarkable capacity to combat each other's shortcomings, and reveal new insights about the complex patterns of human experience.

Stephanie Carter leads the Guest Experience Research team at Airbnb. Prior to Airbnb, she worked as a researcher for Facebook and for a med device design incubator; as a Resident Naturalist the jungles of Peru; and as an Urban Planner in post-Katrina New Orleans. She's driven by research's ability to understand and improve people's interactions with their built, natural, and digital environments. She studied Environmental Sciences & Art at Northwestern University and received her MFA in Design at Stanford. *stephanie.carter@gmail.com*

Richard Dear is a data scientist at Airbnb focusing on guest retention and loyalty. Prior to Airbnb he designed mobile games at two startups in China. He studied physics and philosophy at the Australian National University, though he remembers stressing about giving a TEDx talk on cross-disciplinary research more clearly than his actual courses. *richard.dear@airbnb.com*

REFERENCES CITED

Parigi Paolo. and Bogdan State
2014 Disenchanting the World: The Impact of Technology on Relationships. In: Aiello L.M., McFarland D. (eds) Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol 8851. Springer, Cham.

Ert, Eyal, Aliza Fleischer, and Nathan Magen
2016 Trust and reputation in the sharing economy: The role of personal photos in Airbnb. Tourism Management 55 (2016), 62–73.

Abrahao, Bruno, Paolo Parigi, Alok Gupta, and Karen Cook
2017 Reputation offsets trust judgments based on social biases among Airbnb users. PNAS 114 (37): 9848-9853.

Designing for Interactions with Automated Vehicles: Ethnography at the Boundary of Quantitative-Data-Driven Disciplines

MARKUS ROTHMÜLLER
School of Architecture, Design and Planning, Aalborg University Copenhagen, Denmark and Shift Insights & Innovation Consulting

PERNILLE HOLM RASMUSSEN
School of Architecture, Design and Planning, Aalborg University Copenhagen, Denmark

SIGNE ALEXANDRA VENDELBO-LARSEN
School of Architecture, Design and Planning, Aalborg University Copenhagen, Denmark

This case study presents ethnographic work in the midst of two fields of technological innovation: automated vehicles (AV) and virtual reality (VR). It showcases the work of three MSc. Techno-Anthropology students and their collaboration with the EU H2020 project 'interACT', sharing the goal to develop external human-machine interfaces (e-HMI) for AVs to cooperate with human road users in urban traffic in the future. The authors reflect on their collaboration with human factor researchers, data scientists, engineers, experimental researchers, VR-developers and HMI-designers, and on experienced challenges between the paradigms of qualitative and quantitative research. Despite the immense value of ethnography and other disciplines to collectively create holistic representations of reality, this case study reveals several tensions and struggles to align multi-disciplinary worldviews. Results show the value of including ethnographers: 1) in the design and piloting of a digital observation app for the creation of large datasets; 2) in the analysis of large amounts of data; 3) in finding the potential of and designing e-HMI concepts; 4) in the representation of real-world context and complexity in VR; 5) in the evaluation of e-HMI prototypes in VR; and finally 6) in critically reflecting on the construction of evidence from multiple disciplines, including ethnography itself.

> So (...) perhaps for the first time there is a dataset, where the ontology is with the subjects in it rather than anyone else trying to project things onto that. So that appears to be quite a new thing for data science. We still don't understand what benefits it will have, but it's an unusual with a dataset that has been conceptualized in the subject and terms of ethnographic principles, rather than the experimental way. (...) it seems as if it should be a good thing, but we don't know yet exactly why.
> (Data scientist from University of Leeds reflecting on the collaboration with the techno-anthropologists, 2018)

INTRODUCTION AND PROBLEM SPACE

The development of AVs is complex and demanding in regards to technical feasibility and social acceptance among road users. The successful integration of AVs will depend on the ability of the technological system to cope with the social and contextual complexity of mixed-traffic interactions in the future (Vinkhuyzen and Cefkin, 2016.). Therefore, research on human negotiations in such interactions will be crucial for AVs to safely navigate through urban spaces and to predict intentions and future maneuvers of human road users. To be truly cooperative, though, AVs will have to be able to communicate with human road users, which is why the development of e-HMI is one of the currently hottest topics regarding the safe, efficient, and socially accepted use of AVs in urban spaces.

Designing external Human-Machine Interfaces for Autonomous Vehicles

To design human-machine interfaces for future interactions, we need to perceive the interaction through the eyes of the human and through the eyes of the machine and analyze their shared understanding (Suchman 2007, p. 123). Additionally, we have to understand the hybrid relation between the human, the technology and the world they interact in (Verbeek, 2015). To achieve this, research has to investigate traffic interactions from several perspectives: the one of the pedestrian (or other human road user) making sense of driver- and vehicle behavior today; the one of the driver making sense of human road user behavior today; and the one of the AV making sense of human road user behavior in the future. Representing these perspectives and the in real-time interrelated and fluid decision-making processes of each interaction participant, situated in a real-world context, creates the basis for the design of e-HMI concepts.

To start with, research needs to find out what information pedestrians seek for when scanning their surroundings and what information they use when interacting with drivers and vehicles to understand whether it is safe to cross a road or not, today (Schieben et al., 2018). So far, interaction studies have shown that in many situations pedestrians mainly focus on the speed and distance of the approaching cars to inform their crossing decision (Rasouli & Tsotsos, 2018; Dey & Terken 2017; Merat & Madigan, 2017; Clamann et al., 2017; Yannis et al., 2013; Kadali & Perumal, 2012; Cherry et al., 2012). The generalized suspicion is that in many situations pedestrians will not require any additional signals provided by e-HMI, as little to no interaction will take place anyway.

Conversely this means, the first step is to find out when and how additional e-HMI signals might actually add value to the interaction between AVs and pedestrians. One natural starting point is to investigate today's traffic interactions between pedestrians and drivers, the interplay of which might indicate potentials for e-HMI to substitute for the missing driver in the future. The important point is to understand how the perspectives and decision-making processes of drivers and pedestrians inform each other.

Once situations where pedestrians make use of additional information from drivers (e.g. where they look, hand gestures, etc.) are identified, the value of e-HMI could be to substitute the information from the missing driver in AVs in the future (Schieben et al., 2018; Nathanael et al., 2018; Wilbrink et al., 2018; Chang et al., 2017: Mahadevan et al., 2017; Merat et al., 2016; Parkin et al., 2016; Langström & Lundgren, 2015). To truly design for interactions with AVs and to provide e-HMI signals in the right moments, however, we need to understand how the AV's artificial intelligence (AI) will make sense of these situations in the future. It is important to know what information the AV detects from a traffic situation and interacting human road users and how it processes this information, as this will be the basis for AI-based decision-making and real-time communication of signals (Drakoulis et al., 2018, p.9; Wilbrink et al., 2018, p.11).

Therefore, it is important to represent the perspective and decision-making process of pedestrians and drivers in interactions today, as well as the perspective and decision-making process of AVs in the future. The real challenge, however, is not to understand these perspectives and decision-making processes individually. The real challenge is to find proper ways to represent these perspectives and to then align these representations as a form of holistic evidence providing the basis to design and development e-HMI. This is particularly challenging when AVs have neither been fully developed yet nor tested in urban spaces. It makes the investigation of human-AV interactions purely based on today's interactions and speculative best guesses of how these interactions might turn out in a future with AVs (Cefkin & Stayton, 2017).

Evaluating e-HMI Concepts in Virtual Reality

In addition to understanding traffic interactions and designing e-HMI concepts, it is important to evaluate their effects on pedestrians' decision-making processes when interacting with AVs in the future. Since the evaluation in naturalistic city traffic poses a risk to traffic safety and is still restricted by legal frameworks in most places, simulations in VR are an often preferred alternative method of investigation. VR simulations are not only cheaper and safer (Blissing, 2016; Sobhani & Farooq, 2018), but offer an ideal platform for experimental research (Wilson & Soranzo, 2015), which then makes VR an obvious choice to evaluate and measure the effect of e-HMI. Just as any other research method, though, VR experiments have their limitations. As anthropologists we see two profound challenges. The first is that experiments follow a reductionist approach, breaking down the complexity of real-world contexts and social phenomena such as interaction behavior into simpler elements which in turn should enable an understanding of cause and effect. In praxis this means that the complexity of naturalistic traffic situations gets reduced to a considerably simplified representation of reality built in VR. To properly evaluate the effect that e-HMI causes in simulated future traffic interactions, then, all other influence factors that could be found in naturalistic traffic interactions today get reduced to a minimum as they would make the analysis of cause and effect increasingly complex. And even though VR simulations are typically only used as early indicators for whether a design concept works or not, they still do influence decisions in the development of technological products being eventually used in complex real-world contexts. Hence, the challenge here is how to infuse 'important elements' of the complexity and context of naturalistic settings in VR that have been identified as the very basis for why e-HMI might be valuable in the first place.

The second challenge with experimental evaluations of e-HMI in VR is related to cultural and local differences influencing traffic interactions. Even if we manage to infuse features of real-world complexities in VR experiments, the ecological validity of results from these experiments remains highly questionable in terms of applicability to multiple, socioculturally varying real-world environments (Rasouli & Tsotsos, 2018; Schieben, 2018; Stayton, Cefkin & Zhang, 2017; Turkle, 2009). Although some virtual experiments have argued that behavioral differences between virtual- and real-world environments were little to not at all noticeable (e.g. Bhagavathula et al., 2018), one of the most comprehensive surveys on pedestrian behavior studies just recently concluded that contradictions in such generalizations root in *"variations in culture (...) and interrelationships between the factors"* being studied in isolation (Rasouli & Tsotsos, 2018). In simpler words, e-HMI will have to work in several traffic cultures and it is particularly difficult to receive results from VR experiments that can be scaled up and seen as generally valid for different cultures and contexts. Nevertheless, ecological validity and applicability to the real-world complexity and context, needs to inform technological development just as much as cause and effect studied in experiments under simplified conditions of reality.

Multidisciplinary Representations of Reality

The two previous subsections described the problem space in the prototyping of e-HMI and outlined some of the major challenges in regards to representing reality, both, in the quest of understanding today's and future traffic interactions as well as in the pursuit of methods for evaluating e-HMI prototypes in VR experiments. This subsection introduces some of the complexity and challenges of multiple research disciplines collaborating in this problem space. As the visualization below shows, we first collaborated with interACT's human factor researchers, engineers, data scientists, and human behavior modelers in order to investigate today's traffic interactions in naturalistic settings, and then with experimental researchers, engineers, HMI designers and VR developers to investigate simulated future interactions to evaluate our e-HMI concepts in VR.

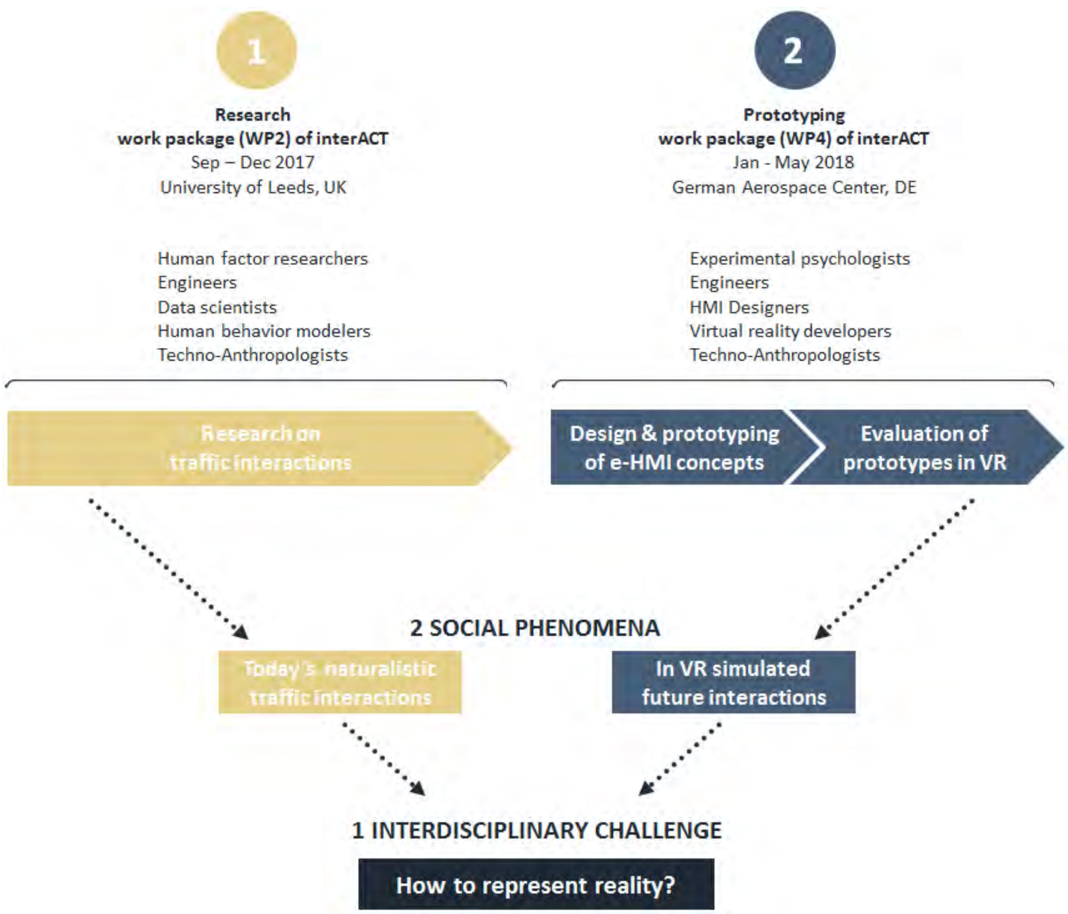


Figure 1. Two collaborations, two phenomena, one central interdisciplinary challenge

Hence, we conducted research on two social phenomena at the core of our engagement with interACT. And even though this research took place at two entirely different locations in different periods with different project teams of interACT, we experienced the same profound interdisciplinary challenge, which was how to represent reality when investigating traffic interactions.

Over the past decades, the often discussed qualitative vs. quantitative dichotomy has opened up for more tolerance between the two paradigms (Sale, Lohfeld & Brazil, 2002). This, however, does not imply that the production of more holistic, interdisciplinary forms of evidence to describe a social phenomenon is an easy task. The human factor researchers, data scientists, engineers, and experimental psychologists that we collaborated with, for example, knew little to nothing about ethnography or anthropological research outside of the realm of adventurous research studying native tribes on tropical islands. This is not necessarily surprising, since anthropologists and ethnographers, are still rarely seen in research directly informing the technical development of AVs[1]. In contrast, the fact that psychology-based human factor research has developed strong links to engineering disciplines over the past decades, roots in the compatibility of representing reality through a similar quantitative lens that proves useful to engineering and product design (Stanton et al., 2013). And even though human factor researchers and experimental psychologists do also work qualitatively, the concept of qualitative induction or ethnography is typically not part of their training (McNamara et al., 2015). In our case, we experienced that, although studying the same phenomena, our ways of producing evidence to describe the phenomenon at hand were inherently different to the ones of our collaborators. Of course all of us referred to the same phenomenon (traffic interactions), yet we did so by following different paradigmatic assumptions (Sale, Lohfeld & Brazil, 2002). Eventually, one could say we talked about different things - different representations of reality. This roots in the fact that both paradigms follow two profoundly different worldviews (ontologies) in their attempts to represent reality (ibid.). The quantitative worldview believes in one single objective reality that can be observed and described independent of the perspective one chooses to perceive this reality (ibid.). The qualitative worldview, on the other hand, argues that reality is constructed by the very perspective one chooses to look at reality. Thereby each perspective on reality essentially represents an own (part of) reality. In our case, the real challenge of developing e-HMI concepts, thus, was how to represent reality qualitatively and quantitatively, from the insider's and outsider's perspective[2].

More broadly, this addresses the challenge that the development of human-machine/human-robot interfaces does not only depend on 1) the investigation of the perspectives of each interaction partner (e.g. pedestrian and driver today changing to pedestrian and AV in the future) and their relation to each other as well as to the context in which the interaction takes place, but also on 2) how research represents the reality from an outsider's and the insider's perspective of each interaction participant in, both, real- and virtual-environments. For example, researchers investigating the reality from an outsider's perspective only, might entirely reject the idea that we need to understand interactions through the insider's perspective of the pedestrian, the driver and the AV, as there is only one objective reality anyway and that is best to be understood from an external perspective. The central question, thus, was: How to represent reality?

HOW TO REPRESENT REALITY

To investigate today's traffic interactions, interACT ran two studies: one using human observations, LiDARs and video recordings in Athens, Leeds and Munich, and one using eye-tracking hardware on drivers in Athens to analyze driver-pedestrian interactions. We, techno-anthropologists, supported the large-scale observational study of interACT in Leeds, and decided to run an in-depth ethnographic study using observations and semi-structured interviews in parallel in the same location. The goal was to understand pedestrians' decision-making processes when interacting with drivers in more detail. As the visualization below shows, each study had its own focus and, thereby, represented reality from a different perspective.

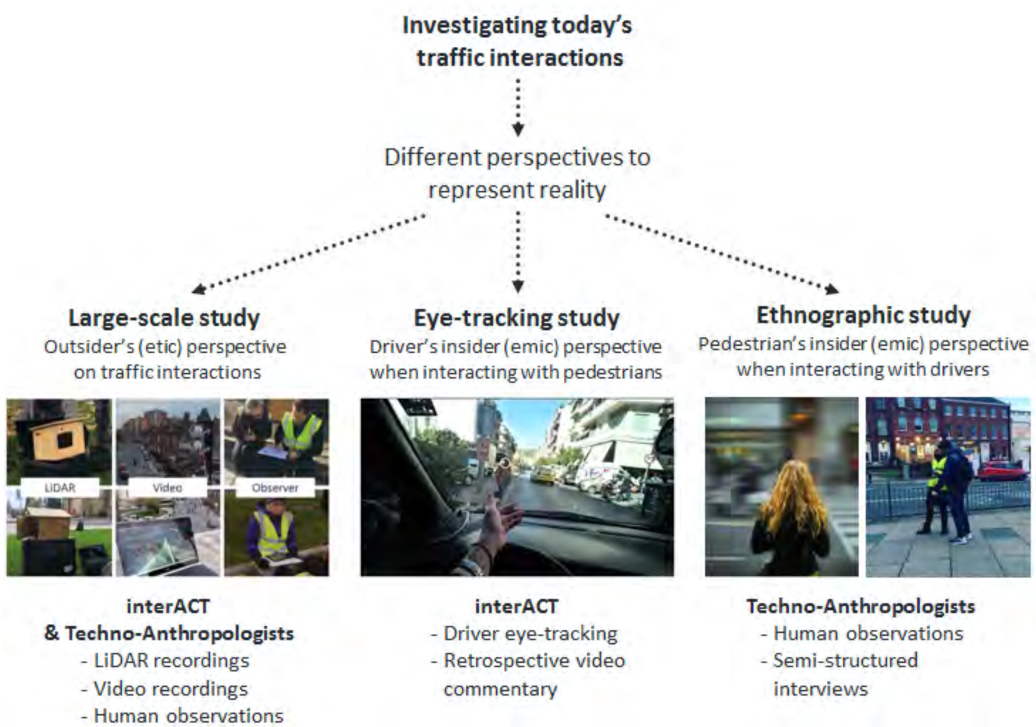


Figure 2. Different perspectives to represent today's traffic interactions (source: 2nd picture on the right: Adobe Spark-Free Pictures; middle: Nathanael et al., 2018; rest: ©interACT2017, used with permission)

The cross-cultural, large-scale study using LiDARs, video cameras and human observations provided an outsider's perspective on traffic interactions. The study applying eye-tracking and retrospective video commentary provided an insider's perspective on driver decision-making when interacting with pedestrians and our ethnographic study offered an insider's perspective on pedestrian decision-making when interacting with drivers. In theory the combination of all these perspectives would be exactly what we need in order to progress the development process of e-HMI, as they provide all the perspectives we described as being needed to design for future interactions. As can be seen, though, not only did each study view reality from a different perspective, but also did they follow completely different methodologies to represent these perspectives. Before aligning any of the results, it is crucial to reflect on the production of evidence of each approach first. As we will show, the devil lies in the detail.

In the following, we will reverse-engineer the process of producing and utilizing evidence to inform the development of e-HMI concepts and active vehicle controllers. We will present the boundaries and achievements of interdisciplinary work, and through this process address limitations, so we can reflect upon these and offer a baseline for discussions on how ethnographers, data scientists, engineers, human factor researchers and experimental psychologists can collaborate in creating more holistic forms of evidence to inform technological innovation in the future.

Representing the Outsider's Perspective: Data Scientists Using Game Theory Models on a Human Observations Dataset

While interACT is still working on combining the different datasets of each approach, the large-scale observation dataset has already been used for different types of analysis. For example, it currently serves data scientists as naturalistic data to train game theoretic models, which will eventually be implemented in active real-time vehicle controllers of AVs. This is not directly related to the interACT project but runs as a parallel spin-off project led by a data science team at the University of Leeds, UK. They view *"pedestrians as active agents having their own utilities and decisions"* which need to be *"inferred and predicted by AVs in order to control interactions with them and navigation around them"* (Camara et al., 2018). Drawing on game theoretic models, they perceive interactions as a competition for road-space between pedestrians and AVs (ibid.). They used the human observation data *"from real-world human road crossings to determine what features of crossing behaviors are predictive about the level of assertiveness of pedestrians and of the eventual winner of the interactions"* (ibid.). To achieve this, they followed a reductionist approach of *"decomposing pedestrian-vehicle interactions into sequences of independent discrete events"* and applied probabilistic methods (logistic- and decision tree regression) as well as *"sequence analysis (...) to find common patterns of behavior and to predict the winner of the interaction"* (ibid.). How this eventually looked like is visualized in the following pictures.

The analysis process of the data scientists (picture 3) led to the creation of two types of evidence: top-10 n-grams (picture 4) and a decision tree for pedestrian-vehicle interactions (picture 5). N-grams (motifs) are the outcome of a process called motif selection which was used by the data scientists to identify *"common short sequences of events which tend to occur together"* (Camara et al., 2018), meaning they tried to define the ten most often co-occurring combinations of behavioral elements in the observed interactions between pedestrians and vehicles. Decision trees were then used to find out which of the single behavioral elements and motifs are informative to predict the winner for each interaction, while providing a *"visualization helpful for human interpretation, and a fast method for real-time systems such as AVs to make decisions based on a few variables"* (ibid.).

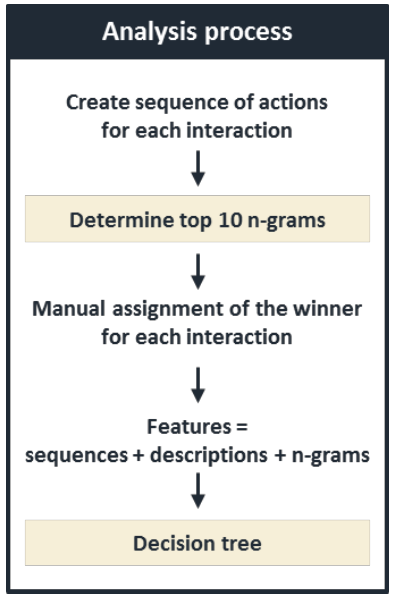


Figure 3. Analysis process of the data scientists (source: Camara et al., 2018)

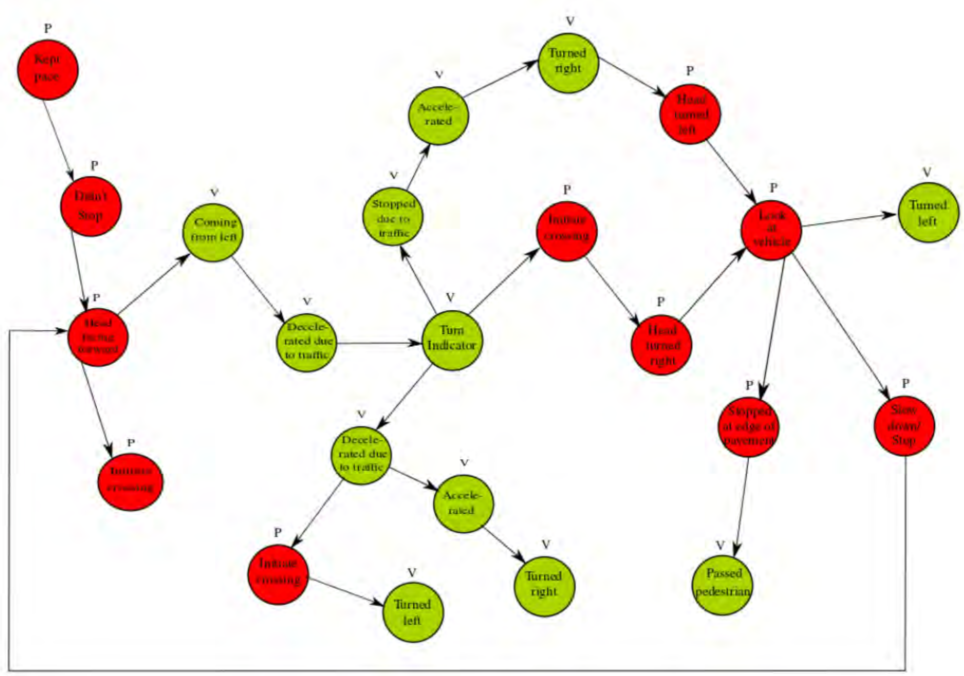


Figure 4. Example evidence of data scientists: N-grams (source: Camara et al., 2018)

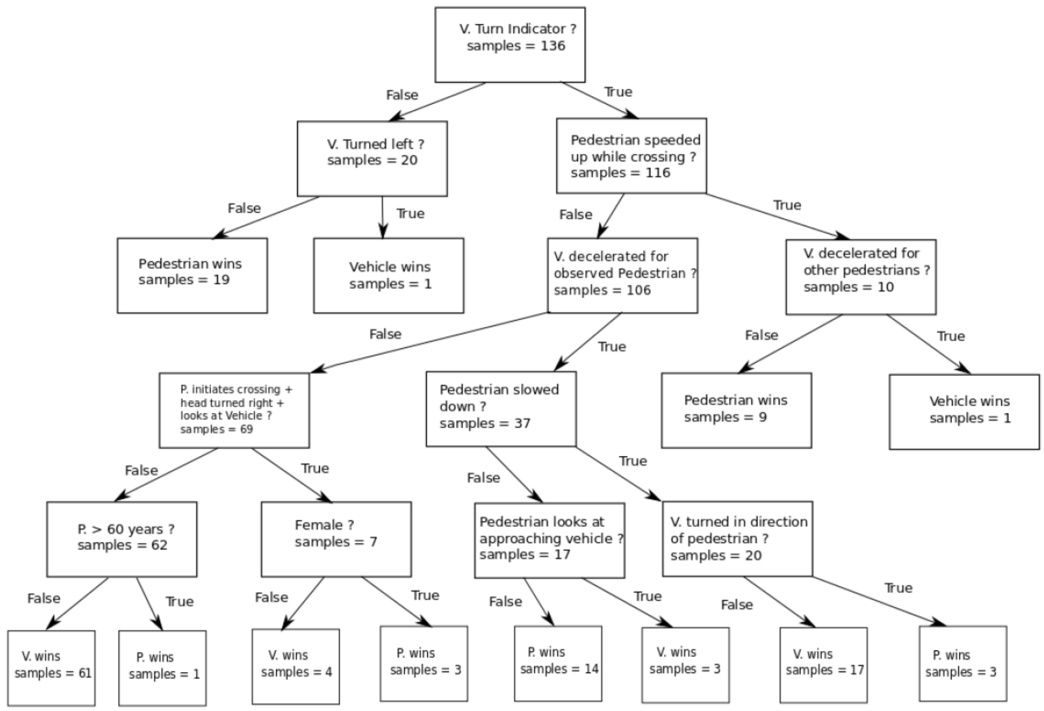


Figure 5. Example evidence of data scientists: Decision Tree (source: Camara et al., 2018)

As stated in the last quote, the described examples of data science evidence built the basis for human interpretation, which led the data scientists to conclude that pedestrians - generally - *"(...) seek for cues in the vehicle's motion [but] not in eye contact with or gestures by the driver"* (ibid.) since:

> (...) the data collection observers themselves did not record much information about the driver gestures because they were difficult to see. So eye gaze by the pedestrian is important, but eye contact with the driver or AV is not (...). These findings are important for AV design as they suggest that AVs should also be designed to communicate simply via their position on the road (...) but maybe not needing artificial face, eye, or gesture substitutes; and that they do need to detect and process pedestrian faces and eyes in order to inform their interactions. (ibid.)

We present this quote, because it introduces the demand for critical reflections on the construction and interpretation of evidence, which this case study attempts to provide in the first place. The line of argument shows the interpretation of evidence from the data science approach that pedestrians - generally - do not seek cues provided by the driver, since we, observers, could often not see the driver from our position, and therefore noted 'pedestrian looked at the vehicle' rather than 'pedestrian looked at driver'. In the following we will show why this interpretation-based generalization was more of a misconception of what the data could actually tell, rather than any sort of objective representation of reality.

The Construction of Evidence: What Data Tells and What Not

To collect comparable observation data in three different cities in a replicable manner, interACT decided to use a standardized observation protocol. Additionally, interACT decided to develop an app to run through the observation protocol on a tablet, which allowed us to transform the expected large amounts of data into digital formats that could directly be used for different types of analysis later. As trained ethnographers with previous experience in conducting observational studies, our role was to support the development of these digital standard observation protocols. Therefore, interACT provided us with a first draft based on else-where published research and hypotheses on what would be important for the agenda of interACT including the one of the data scientists.

Figure 6. Development of the protocols with pen and paper. Picture 7. Observations with the digital protocols on a tablet. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2017)

Our goal was to run a pilot study - a proof of concept of the first protocol draft - following an exploratory observation approach, which should eventually avoid forcing interactions into predefined categories. At first, we struggled to design a 'one-size-fits-all' observation protocol, but ultimately managed to provide the observer with a bit more flexibility by including a space for sketches and additional notes. This enabled the observer to describe interactions beyond the limitations of predefined categories only.

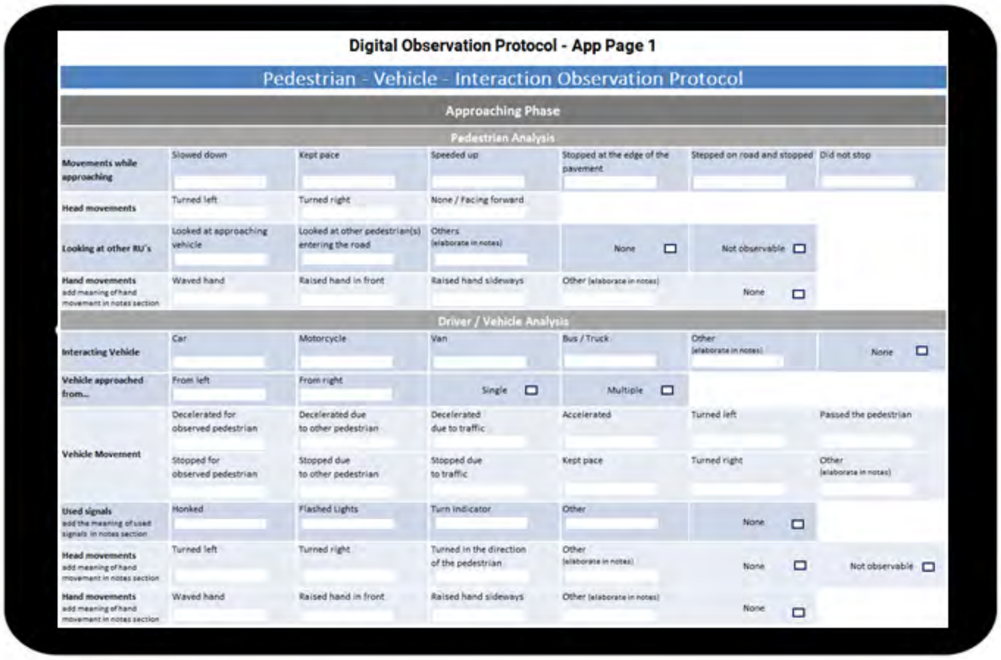


Figure 7. App-mockup of digital observation protocol – Section for interaction behavior. (©interACT2017, used with permission)

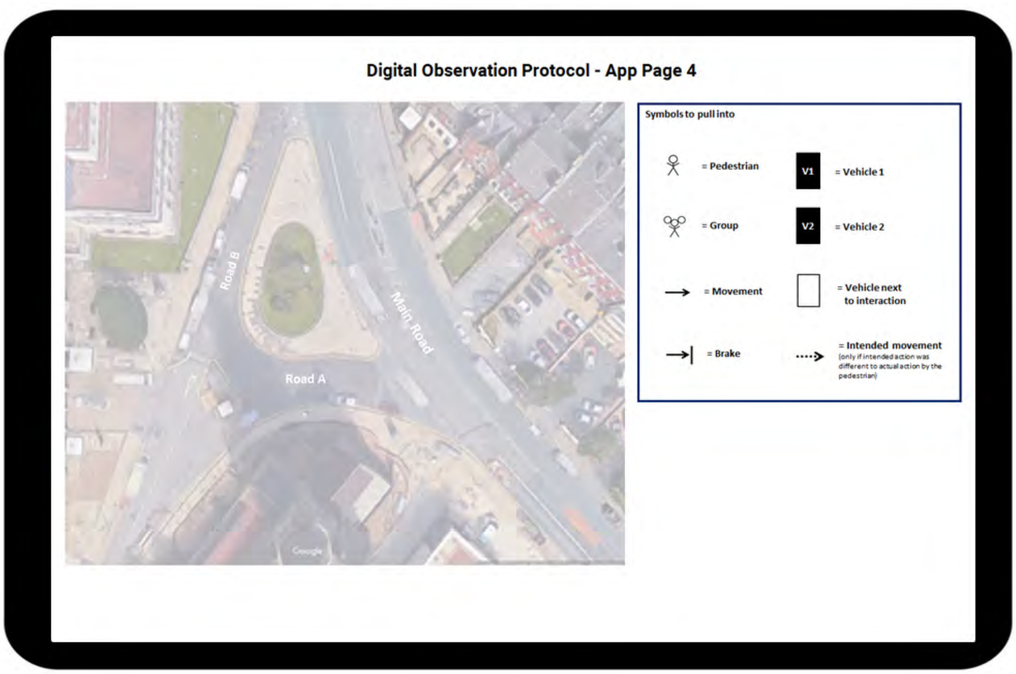


Figure 8. App-mockup of digital observation protocol – Section for sketching. (©interACT2017, used with permission)

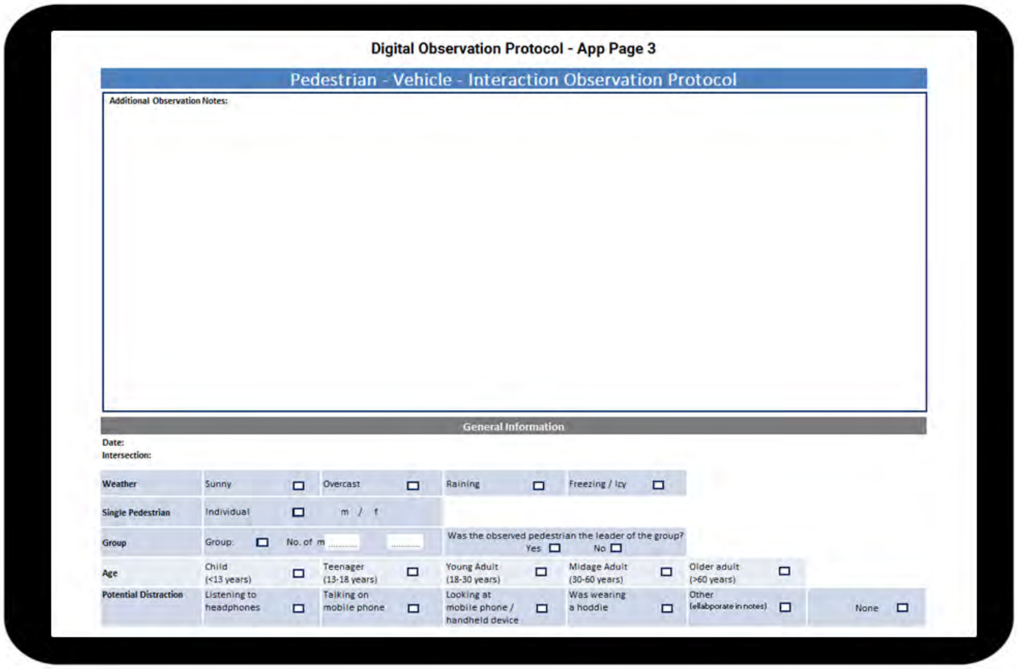


Figure 9. App-mockup of digital observation protocol – Section for observation notes. (©interACT2017, used with permission)

This exploratory pilot study led to a final app that allowed us and other observers from interACT to create one of *"the largest and most detailed dataset of its kind"* (Camara et al, 2018), representing nearly 1000 traffic interactions, each through 62 individual behavioral elements (e.g. behavior expressed through motion, signals, head or hand movements) combined as a sequence, and 12 environmental influence factors. Each human observation of interactions in Athens, Leeds and Munich was, therefore, represented by the exact same selection of behavioral elements - which was important for cross-cultural comparability and quantitative analyses such as the ones of the data scientists.

So in that sense, interACT succeeded in following a replicable observation approach, which, nevertheless, did not come without limitations and implications. Even though, the pilot study was successful in enabling the protocol to capture most of the behavior in interactions, it still reduced the complexity of real-world interactions to a simpler representation of reality. First of all, the fact that the protocol primarily focused on one-to-one interactions, considering other actors in the environment as an environmental factor rather than another interaction participant, was one of the first steps to reduce complexity. Secondly, any form of interaction behavior that was not included or sufficiently well described as a category in the protocol had to be added in the notes field, which has not been part of any analysis so far. Thirdly, the fact that these observations only provided an outsider's perspective on the reality of which the observed pedestrian, vehicle and driver were part of means the dataset does not allow any inference of internal perceptions of these interaction participants. In relation to this we had multiple discussions with our collaborators about what is observable, what is an objective observation, and what a subjective interpretation. These very discussions defined the way we, observers, described interactions, and thereby predefined the dataset's representation of reality.

It makes the whole observational study a subjective choice of how much of reality should be described in the dataset. As much as one would like to promote objectivity by controlling or excluding subjective interpretations, we argue that any decision on how much of the perceivable reality should be represented in an observation is just as subjective. Because, isn't the choice of representing certain elements of reality in a limited way just as influential for the construction of evidence, as the choice to represent certain elements of reality as an interpretation of trained observers? Eye contact, for example, was genuinely excluded from the protocol, as this was categorized as not observable. And it's true; how could you know that a pedestrian had eye contact with a driver, if you are not either that very pedestrian or driver? Of course we could have asked them whenever we had the suspicion, but we didn't. Instead, it was decided to observe whether the pedestrian looked at the driver, or looked at the car, and whether the driver looked at the pedestrian. In many cases, these observations were based on the direction of head turns of pedestrians, which made it still quite difficult to know whether they actually looked at the driver or the vehicle. This essentially means that we only noted down that pedestrians looked at the driver whenever this was very clear to us, which in turn does not mean that pedestrians who we observed as having looked at the vehicle might not also have look at the driver. Thus, the earlier shown conclusion of the data scientists' publication that eye gaze from the pedestrian is important but from the driver or AV is not, and that eye contact and the substitution of this through e-HMI, therefore, is generally not needed was rather a misinterpretation of what the data could actually tell. We could not know whether eye contact or where the driver looked at was important to pedestrians because we did not include their perception in the data collection. This means that we, ethnographers, could not only really contribute to piloting the very instrument to collect data for data scientists, but also to explain and reflect on what the collected data actually tells.

Understanding the dataset and meaning of single data points is as essential as being able to structure and analyze it. In fact, this seemed to have kick-started a reflection process in the broader research work of our collaborators. The data scientists, for example, told us that our contribution to understanding the dataset correctly, also made them question the representation of reality from other datasets they worked with.

Though the need for objective observations was to some extent met; it is important to be aware that any decision taken in relation to how data gets collected eventually influences how reality gets represented. The fact that the dataset merely represented the observer's perception of reality from an outsider's point of view (classically termed as 'passive observation' within the field of ethnography, leading to an etic[2] representation of the phenomenon), meant that any conclusion being based on constructions of evidence from this dataset, needed to reflect these limitations. Missing awareness of these limitations in the analysis, means the interpreter of constructed evidence acts as subjective as the interpreter of observable behavior.

Nonetheless, the collaboration showed that piecing together a team of ethnographers, human factor researchers, and data scientists can indeed go beyond the limits of either one of the disciplines alone:

- Ethnographers can be useful to make sure that hypothesized representations of traffic interactions can ultimately represent reality in an appropriate and applicable level of detail without forcing the dataset-structure onto reality, but rather shaping the structure in a way that it can (at least partly) represent real-world complexities even in a standardized large-scale data collection.
- The human factor researchers, provided an important link to manage the necessary limits of controlling the complexity of real-world representations in a way that the dataset remains useful for the reductionist approach of them and data scientists when analyzing it.
- And the data scientists were useful to link these representations of reality to the technical domain of AI-based technology and analyze the data in a way that enables it to train models for active vehicle control of AVs.
- Ultimately, this collaboration showed that ethnographers, as the creators of dataset structures or whole datasets, should also be involved in the analysis process as a sparring partner to analyze and to reflect on what data actually means when drawing conclusions from constructed evidence. In the following we will argue why ethnography can yet offer an entirely different worldview to understand the reality of traffic interactions.

Representing the Outsider's Perspective: Using Sequence Diagrams to Design Future Communication Strategies

In the quest of designing future communication strategies for AVs when interacting with pedestrians, interACT's work package 4 team worked with so called 'sequence diagrams' - another form of evidence based on the human observation dataset. As can be seen below, it is an easy to read visualization of commonly observed sequences of interaction behavior exchanged between the driver and the pedestrian.

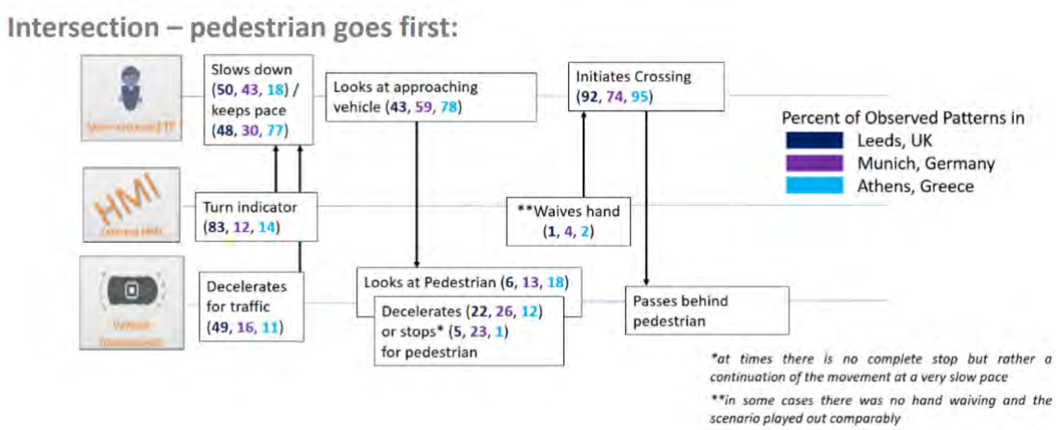


Figure 10. Sequence diagram. (source: Dietrich, 2018; Dietrich et al., 2018; Wilbrink et al., 2018)

These sequence diagrams were used in two ways: 1) to represent one single interaction, 2) to represent a whole dataset of interactions. The upper visualization for example shows a part of the total dataset. It quantitatively visualizes how often (in percentages) a particular element of an interaction occurred in the given sequence throughout the three different field sites. The power of these sequence diagrams is that they show behavioral elements in combination to former and later occurring behaviors, and thereby produce a great basis to design for similar situations with AVs in the future. This means in turn that they lose their power once this combination is not clear anymore. Presenting a part of the dataset in percentages appears to be a competent foundation upon which to debate e.g. that we do not have to design e-HMI signals for drivers looking at pedestrians because this only occurred in 6% of all cases, or for hand gestures from the driver as this only occurred in 1% of these interactions.

The big question, though, is what does this form of evidence actually tell us, and who interprets it? There are at least two entirely contradicting ways these numbers could be interpreted: First, one researcher could argue since the percentages of drivers looking at pedestrians and given hand gestures are so low, we do not need to include these in the development of future communication strategies at all. The other researcher might argue that it is exactly the other way around; no matter how often pedestrians look for clarifying information from the driver today, if this is the potential of e-HMI, and if pedestrians would actually benefit from having this information all the time, then these are the situations we have to focus on. So even though it might be a minority in the quantitative representation of reality, it might be that very minority that we need to understand deeply. This is where the outsider's perspective reaches its limits to contribute to the design of future communication strategies. What becomes more important to understand then, is what value these signals actually do provide to the pedestrian, and therefore research needs to go beyond what is observable to include the pedestrian's and driver's insider perspective in detail.

Representing Reality from The Outsider's AND Insider's Perspective

To understand when and why e-HMI can be valuable to pedestrians, we, techno-anthropologists, conducted an additional, iterative ethnographic study with two goals: first, to investigate the decision-making process of pedestrians as broad and detailed as possible and second, to investigate why pedestrians sometimes look for more information from the driver and how this information then contributes to the decision-making process.

In contrast to the team of interACT, who selected relevant traffic scenarios based on *"a step-wise process of intensive discussions within the consortium"* (Wilbrink et al., 2017) to come up with relevant scenarios for all three European countries in the initial phase of the project, we followed bottom-up explorations grounded in the context and complexity of our field site. While running the pilot of the observation protocols for interACT, we realized that there was another highly interesting crossing scenario in the same intersection, in which explicit pedestrian-driver interactions occurred more regularly. This was jaywalking in dense and slow-speed traffic situations - a behavior that not only showed to be mundane for pedestrians in this intersection, but is in general not prohibited in the UK as we were informed by pedestrians. Hence, our assumption is that jaywalking might, indeed, be a more common traffic practice for British people, which would increase the relevance of this scenario for other intersections and cities in the UK as well. Thus, to investigate the pedestrian's decision-making process when jaywalking, we conducted semi-structured interviews with pedestrians directly after they were involved in an interaction. This provided us with an understanding of their insider's perspective of the interaction with drivers, which we then combined with our outsider's perspective as observers. We structured our study in three phases: 1) a *Pilot Phase* for testing and refining our research approach, 2) an *Open Exploration Phase* to find as many influence factors on the pedestrian's decision-making process as possible, and 3) a *Design Problem Orientation Phase* to specifically focus on the relation between pedestrians and drivers as an important basis to identify potentials for valuable e-HMI.

Our approach resembled the process of a developer team using Scrum to iteratively run through 2-4 weeks coding sprints, while embracing a process of daily reflections.

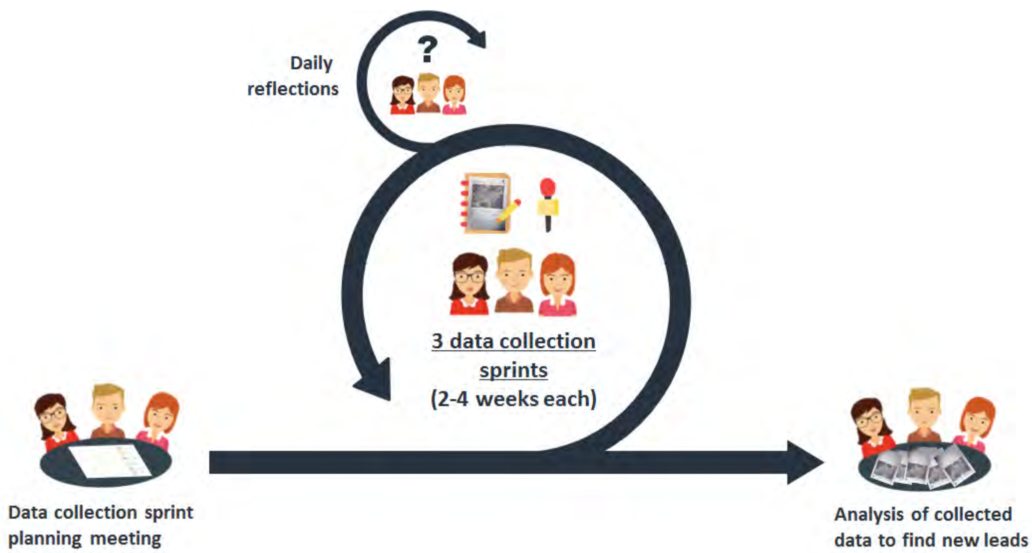
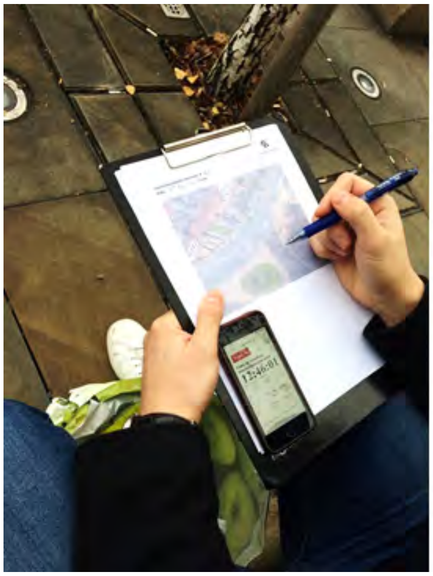


Figure 11. Scrum-like iterative ethnographic research approach.

In between each phase, we preliminarily analyzed the collected data to define the focus of the following phase. We applied Grounded Theory Method (e.g. Glaser & Strauss, 1967) to find emerging patterns in our interview data, which helped us iterate on our research questions and find new leads to explore. In addition to interview recordings, our datasets consisted of observational notes and rough sketches of the situation (see example below).

The combination of sketches and observational notes with the interview data provided the in- and outsider's perspective, and thereby a more holistic representation of the traffic interaction at hand. In comparison to the large-scale observational study of interACT, focusing on what and in which order behavioral elements occurred, our dataset consisted only of 34 interactions but added some depth to *why* and *how* pedestrians made decisions. Each data collection sprint resulted in theoretical models representing an emic[2] understanding of pedestrian's decision-making processes when interacting with drivers. The framework below, for example, shows influence factors that we found within the first data collection sprint, and how these interrelate with each other (Rasmussen, Rothmüller & Vendelbo-Larsen, 2017). Even though this is only a simplified representation of reality as well, it attempts to address more than just the momentary observable influence on pedestrian crossing decisions, and relates the influence of former experiences in different



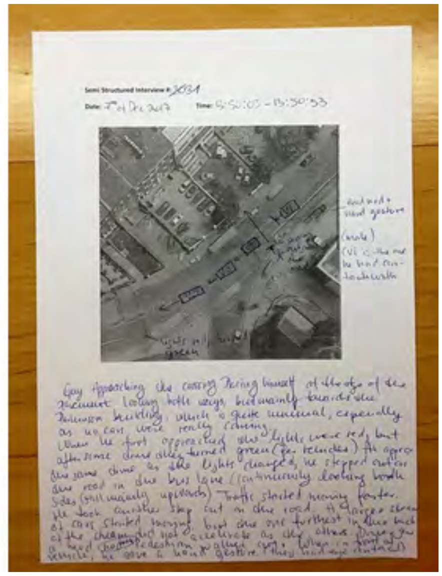


Figure 12. Observer jotting down notes and sketches. Picture 13. Example of a final sketch with observation notes. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2017)

traffic cultures and the knowledge about the intersection to tendencies for specific crossing practices, interpretations and expectations when pedestrians assess a given traffic situation.

This example shows of one of the core elements in our process of constructing evidence. This framework built the basis for the second and third data collection sprint, in which we saturated the categories of the framework with then freshly collected interview data. The visualization below exemplifies how we used the model to understand the individual tendency for a certain crossing practice of one of our participants.

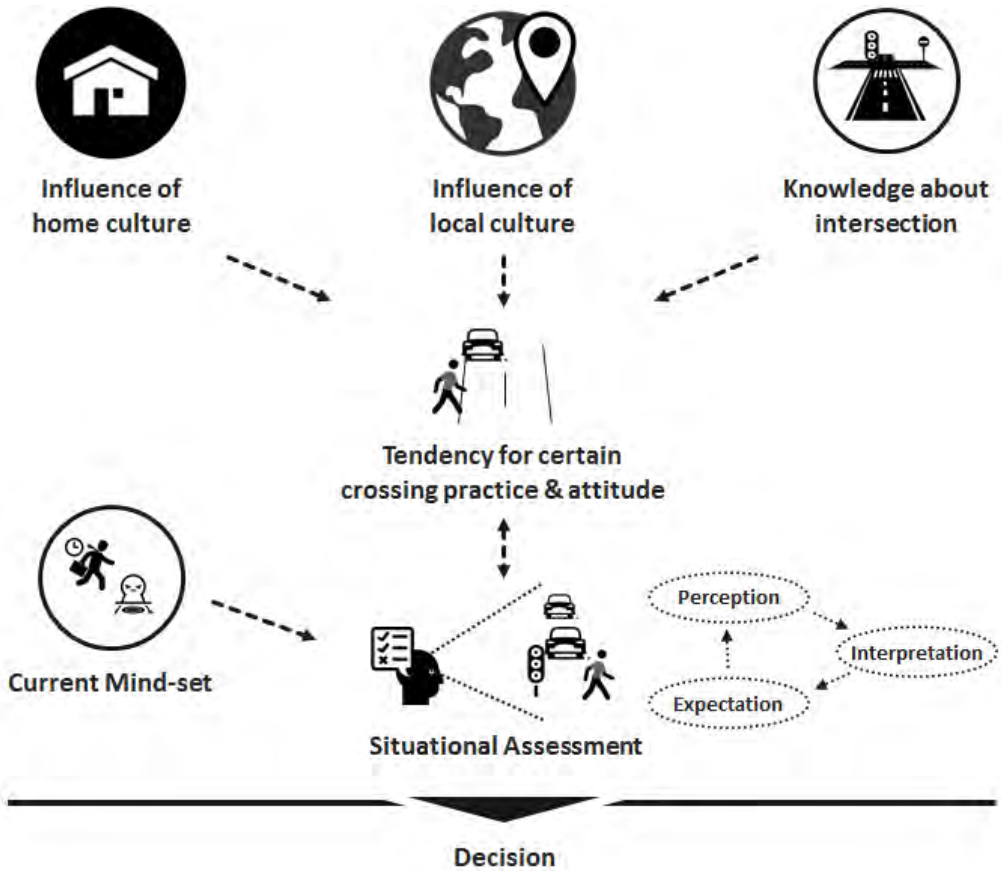
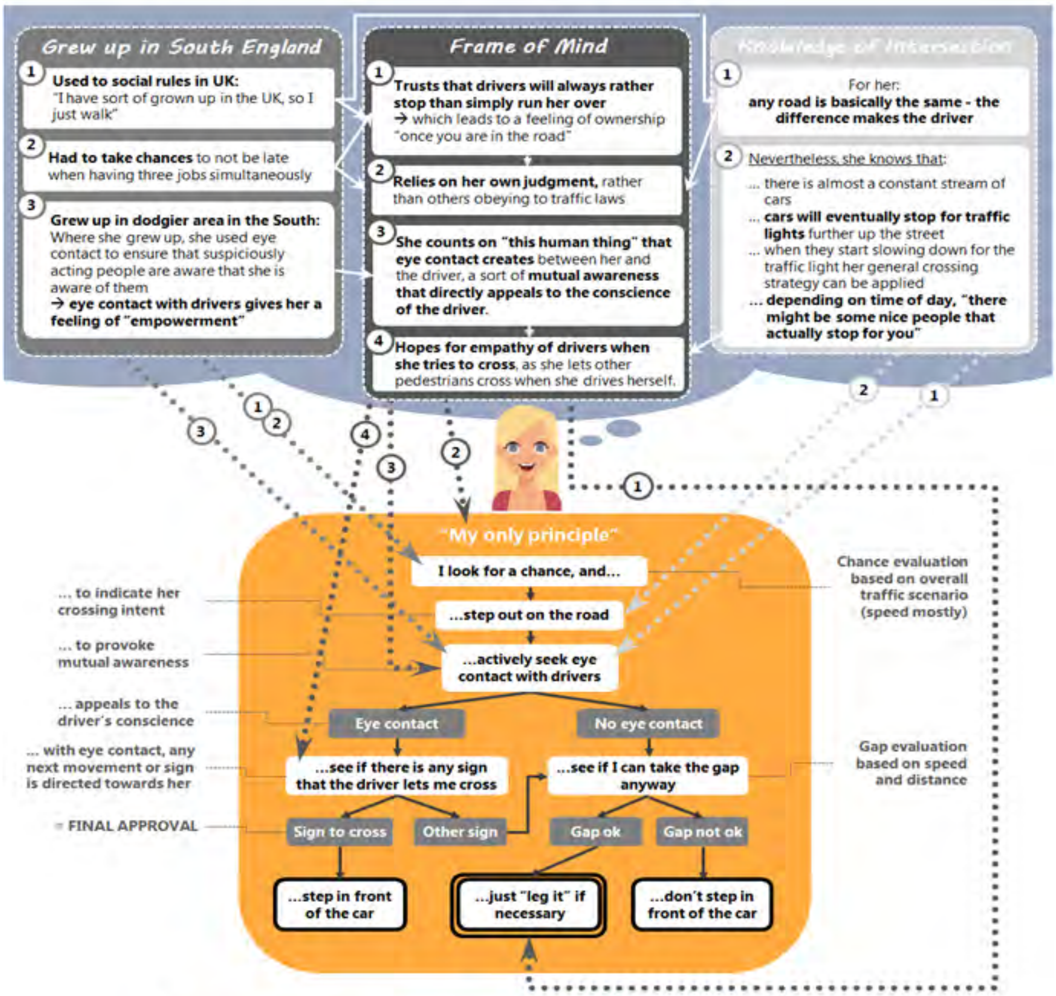


Figure 14. Theoretical framework of pedestrians' decision-making processes. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2017)



"When I cross the road, I sort of hope that someone is gonna make eye contact with me, so that they feel that I need to cross the road (…)
Are you gonna get that from a driverless car?
Is there gonna be the empathy there?"

Figure 15. Example of applied theoretical framework of pedestrians' decision-making process. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2017)

The next model shows one of the elements of the upper framework: the situational assessment model (ibid.).

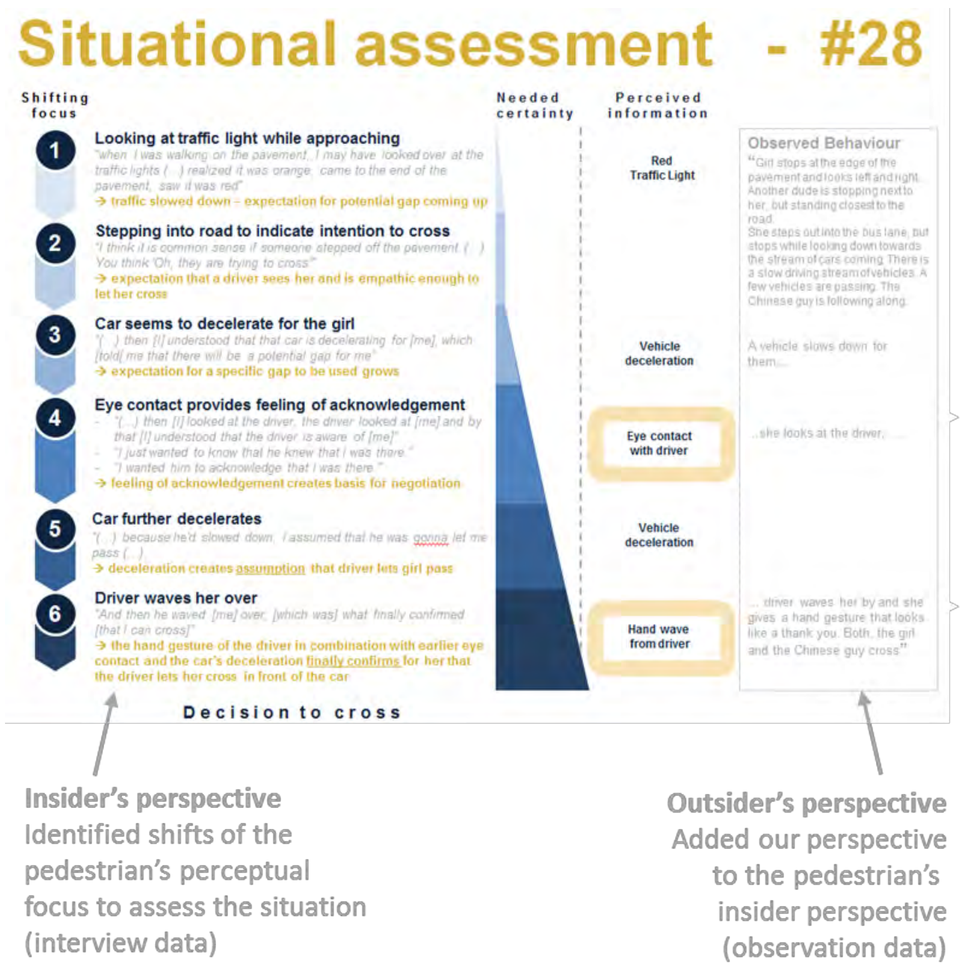


Figure 16. Example of a situational assessment model. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

This example of a situational assessment model shows six stages in this pedestrian's decision-making process from one of the interactions of the third data collection sprint. The six stages are essentially six focal points of the pedestrian, which we identified in the interview data (the pedestrian's insider perspective on reality). Each focal point evoked a certain expectation, and whenever an expectation got confirmed, the pedestrian's certainty to cross increased. It visualizes how we combined the insider's perspective to the for us observable outsider's perspective. The yellow highlighted elements 'Eye contact with driver' and 'Hand wave from driver' were the important interaction behaviors that e-HMI could potentially substitute in the future.

Identifying Potentials for Valuable e-HMI Concepts

As shown in the situational assessment model, expectations are central in the decision-making process of pedestrians (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018 & 2017). For example, we investigated a case where a pedestrian nearly had a traffic accident, as he decided to cross in front of a car that he expected to stop. Not surprisingly perhaps, we found out that e-HMI could be of most value in situations where pedestrians are uncertain about what to expect, or need a confirmation of what they expect, to make the decision to cross. In relation to this, we found out that deceleration alone in dense, slow-speed and close-contact interaction scenarios might not always be interpreted as deceleration for the pedestrian. We noticed early on in the study that *awareness* seemed to be one of the central elements in decision-making processes (ibid.). Not only were pedestrians aware of all sorts of human and non-human actors in the environment, such as approaching cars, traffic lights, other pedestrians, cars driving faster or slower than the general speed of the stream, etc. but also were they really good in judging whether drivers were *aware of them* by e.g. seeing where the driver looked in the environment (ibid.). If they did not see the driver, this also provided a piece of information but did not help them much to cross the road. Traffic interaction is not only about adapting movements to avoid colliding with others, it is about the pedestrian's expectation of what drivers might do next (ibid.). It is about the relation between the pedestrian, the driver, the vehicle, and the real-world context. It is about empathy from both interaction participants to anticipate the other's intention, situated in the complexity and context of the given situation (ibid.).

After having focused on the influence of *awareness* in more detail, we learned that the pedestrian's certainty of drivers being aware of them helped them understand whether a driver was decelerating for them or something/someone else. This seemed to be applicable across cultures; due to the intersection's proximity to the University of Leeds, our study participants came from all sorts of different traffic cultures: Asian, African, American, small villages, cities, metropolis, megalopolis, all indicating this need.

Something very similar was found by interACT's eye-tracking study in Athens. Even though they ran an entirely different study, looking at the driver's insider perspective on interactions with pedestrians, they found similar elements in traffic to create the very basis for e-HMI to be valuable. In *"high density unsignalised urban crossing"* scenarios with a high level of *"proximity and low speeds"*, the team concluded that body movements of the pedestrian alone are not always sufficient to clarify an interaction (Nathanael et al., 2018). As the visualization below shows, pedestrians looking towards the driver's vehicle, led to significantly more effective clarifications of interactions (ibid.).

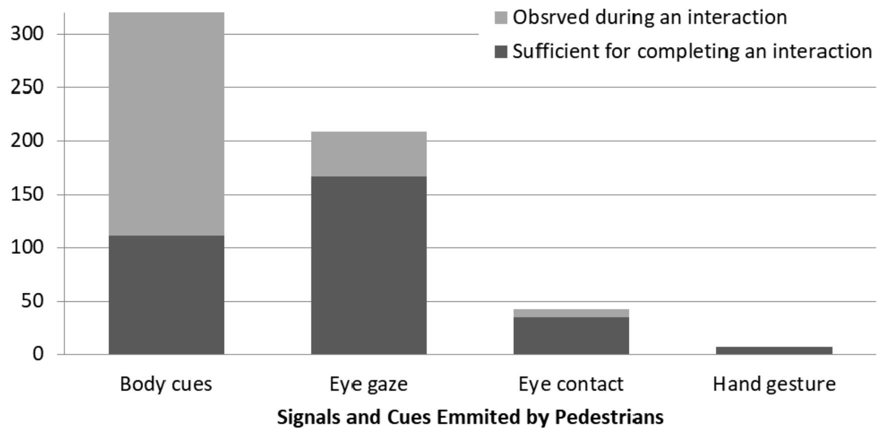


Figure 17. Results on the effectiveness of different cues to resolve an interaction. (source: Nathanael et al., 2018)

Another conclusion from both studies was that in situations where the uncertainty is so high that mutual awareness does not resolve the interaction alone, this awareness then builds the very foundation for further forms of explicit communication such as hand gestures, head nods, or further decelerations to communicate a final approval to the other interaction partner (ibid.; Rasmussen, Rothmüller and Vendelbo-Larsen, 2018). Thus, we identified the potential for valuable e-HMI to communicate a combination of awareness and a final approval, to effectively increase the pedestrian's certainty in the decision-making process (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018).

The relevance of this finding is not only related to today's needs of pedestrians or drivers, but also to the problem of public acceptance of AVs. Throughout our ethnographic work on traffic interactions, we realized that there are two types of prejudices about AVs getting implemented in social spaces. First, some of our participants addressed their trust in scientific research to adequately test and evaluate any form of failure of AVs before they are released into urban traffic. They stated that they would trust AVs more than today's drivers to handle complex traffic situations safely, as they would detect everything around them, and by that would never run someone over. For this group, we see a clear issue of overtrust in future technology. Already today, AVs are tested in urban spaces, which just recently resulted in a lethal situation (Crosley, 2018; Martyn, 2018). Our knowledge about the limitations of sensor technology from prior research confirms this issue (Sangari et al., 2016). At the moment, AV sensor technology simply does not detect everything in the environment (ibid.). For those over-trusting in AV-technology, we expect that e-HMI can effectively reduce risks in future traffic interactions by communicating awareness (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018).

The other prejudice of people was their fear that AV-technology might not work flawlessly in the future, as the process of building this technology is eventually also only a process involving humans. And humans - just as human errors happen in traffic - happen to miss a thing or two in technological innovation (Norman, 2013). This group of people, therefore, would be afraid that the AV might not have detected them. e-HMI communicating awareness would therefore also serve this group of people to know whether they are detected, and in combination with a final approval inform them whether they can safely cross the road (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018).

In the last two subsections, we presented how we produced and used evidence to represent the pedestrian's insider perspective when interacting with drivers. Surprisingly, our results were very similar to interACT's eye-tracking study representing the driver's insider perspective when interacting with pedestrians, which we consider as an important lead that our findings - even though small in the sample and grounded in only one intersection - could apply to other sociocultural environments as well. Understanding what sort of context creates the basis for potentially valuable e-HMI is just as important as understanding which specific elements of interaction behavior from drivers might need to be substituted by e-HMI. The theoretical framework and models we created through our iterative ethnographic study, did then also help us to understand what additional information from the driver means to pedestrians, how they interpret it, why and in which contexts they need it as well as when this plays a role in their decision-making process.

Ethnography-based Design of e-HMI Concepts

In addition to the former described insights for when e-HMI could potentially be valuable, we learned a lot from being engaged in the design process of work package four at interACT. For example, that it is considered a challenge for e-HMI concepts whether the AV should communicate with one specific pedestrian only, or all the pedestrians around the car (Wilbrink et al., 2018). If the car communicates in a way that all pedestrians could interpret the signal as the allowance to go, this could pose a risk for those pedestrians not having been detected, or worse, for those pedestrians who were detected but actually endangered by other cars not having been considered in the AV's decision-making process.

We decided to focus on side-based communication, meaning that the AV only shows on which side it detected pedestrians, while neither communicating how many were detected, nor addressing only one out of a group for example (which we believe to be not only technologically infeasible for proper prototypes but already proved difficult to simulate in one of interACT's VR studies).

The following describes the first design concept of us techno-anthropologists[3]. *e-HMI Design 1 - Side-based Detection and Approval* communicates the detection of pedestrians on one or both sides of the vehicle through a pedestrian symbol lighting up on the side of the car. The choice of a pedestrian symbol was due to our intention to design for a cross-cultural application of our design. Pedestrian traffic lights are one of the already existing symbols to communicate an allowance to go for pedestrians and exist in nearly every traffic culture. The symbol, we believe, will be intuitively understood. For cyclists a cyclist-symbol would need to be added to our design. In combination with detection, our e-HMI communicates the final approval to cross via an LED bar lighting up in a motion that indicates the direction a pedestrian can cross in (inspired by a hand gesture. Thereby it shows pedestrians from one or both sides of the car that a cross can be safely performed. (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

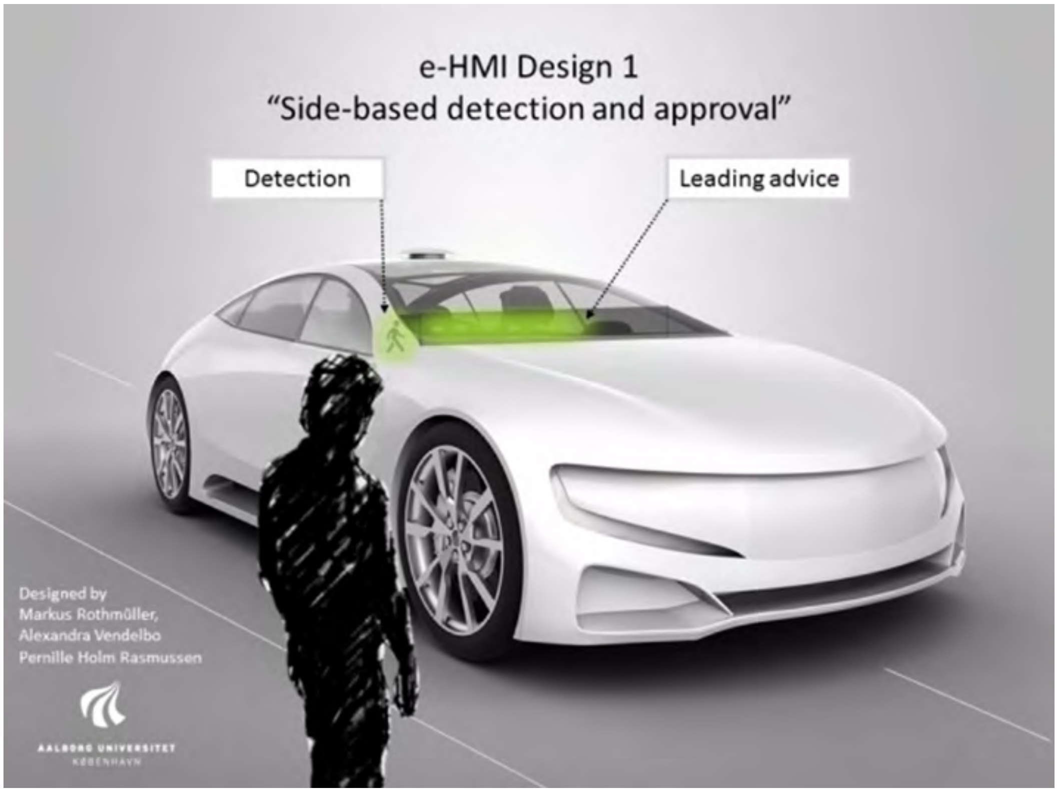


Figure 18. e-HMI Design #1 - 'Side-based detection and approval to cross'. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

While the first holistic design was a very active form of communicating, the second holistic concept was designed in a way to accommodate a more passive form of communication. The basic thought with *e-HMI Design 2 - 360° Awareness and Location-based Tracking* was to empower the pedestrian to decide whether it is safe to cross without the AV communicating an actual advice to cross. In this design, the AV first communicates the readiness to cooperate with the environment through a 360° light band, and then shows the pedestrian(s) around where the car currently detects someone. Once a pedestrian was detected the light band tracks the pedestrian's movement and communicates continuous awareness of the pedestrian's location with a differently colored light stripe following the pedestrian. In the picture below, this tracking functionality was designed for the case of pedestrians crossing in front of the car. (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

Figure 19. e-HMI Design #2 - '360° Awareness & location-based tracking'. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

Our fieldwork-based insights towards the value of e-HMI were, indeed, confirmed by a virtual reality (VR) study in which we evaluated the two e-HMI design concepts. The results of this study suggested that e-HMI communicating the detection of pedestrians and a final approval to cross in situations of high uncertainty for the pedestrian significantly increased the certainty and comfort to cross. In our master thesis "*we argue that e-HMI can show a positive impact on traffic safety, as well as trust and public acceptance of AVs when being integrated in urban traffic"* (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018). A more detailed evaluation of our designs *"in terms of placement, color, motion and the combination of signals and vehicle behavior"*, showed *"the importance of understanding these factors when assessing the effect on decision-making times, the feeling of certainty and comfort to cross when interacting with AVs"* (ibid.). Nevertheless, we experienced yet another clash of disciplines when building the VR scenario to evaluate our e-HMI concepts, which in its core is rooted in different worldviews on how to deal with the representation of reality. This will be the focus of our reflection in this final chapter of our case study. (Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

Representing Real-World Contexts and Complexity in Virtual Reality & Mixed-Methods Evaluation of e-HMI Prototypes

When using VR to evaluate our e-HMI concepts, our goal was to represent specific elements of the real-world context and complexity which we identified as the very basis for why e-HMI might potentially be valuable (dense, slow-speed, close contact traffic situations with high levels of uncertainty for pedestrians).

As stated earlier in the problem space, VR serves as a perfect platform for experimental research, which in contrast to our goal, strives for controlled environments and simple scenarios to effectively verify hypotheses about cause and effect. Our colleagues at the German Aerospace Center, who thankfully were very patient and open to discuss our different worldviews, suggested to reduce the from us desired complex representation of real-world settings to a scenario with minimal influence factors. We, on the other hand, asked: What use is VR when it is unable to represent naturalistic traffic situations? How could we validate the effects of our e-HMI if we could not test it against the influence factors that set the framework for the need of e-HMI in the first place? Hence, we wanted the full package: multiple other pedestrians, cyclists, traffic streams from both sides, a traffic light making the streams slow down and accelerate again, mixed traffic with vehicles including drivers and AVs, different AVs with e-HMI and without, haptic feedback in the room: a sidewalk to step down from, and possibly the risk of getting hit by a pillow if a car virtually collides with the participants. Each of the previously mentioned elements describes a single influence factor which we would have to test our e-HMI design against; resulting in one thousand and fourteen experiment rounds (with a sample of one). How long would that take, we asked. Decades but this is how experimental psychology works, they answered.
To overcome this challenge, we first had to understand the worldview of our collaborators, to then try to merge it with ours. A process of expressing ourselves in hypotheses, dependent and independent variables, and experimental measures helped us getting closer to a statistical controlled experimental design. Eventually, we managed to create a virtual world, which decently satisfied our need for a representation of naturalistic settings, while implementing our colleagues' suggestion for a controlled environment and comparison of e-HMI, without additional influence factors. Therefore, we focused on creating uncertainty for each of our study participants by running through an initial adjustment of gap sizes, setting the baseline for the study, which we hoped would enhance ecological validity (McNamara et al., 2015) - as long as we represented a dense, slow-speed, and close-contact traffic situation in which they had to cross (see picture below).

Our study was a hybrid form of experiment and ethnography-inspired user test. To investigate the outsider's perspective, we jotted down observational notes and measured experimental metrics in video recordings from, both, the participant's in-VR perspective, and a room-perspective recording the behavior of the participant outside of VR (see picture below).

Figure 20. Example of traffic set-up in VR with the goal to evoke a feeling of uncertainty to cross. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

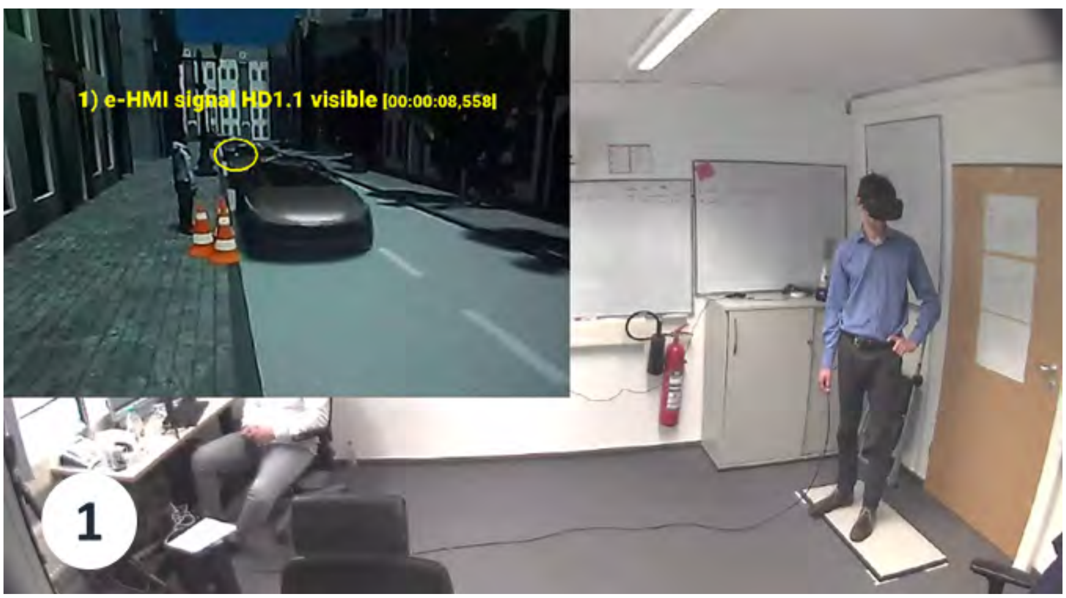


Figure 21. Example of measurements in VR – Visibility time of e-HMI signal HD1.1. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

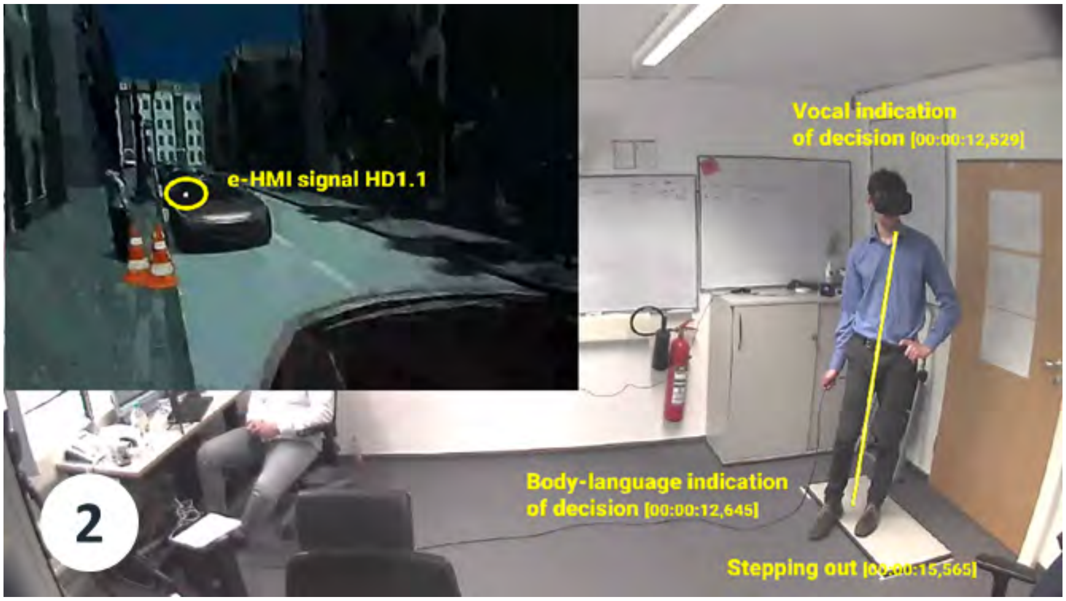


Figure 22. Example of measurements in VR – Indicators of crossing decision. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

Figure 23. Example of measurements in VR – Visibility time of e-HMI signal HD1.2. (source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

Figure 24. Example of qualitative evaluation in VR – Vocal reaction to e-HMI signal.
(source: Rasmussen, Rothmüller & Vendelbo-Larsen, 2018)

To investigate the participant's insider perspective, we combined questionnaire-based ratings of certainty and comfort to cross as well as of the meaning and intuitiveness of our designs with semi-structured interviews, in which we explored a variety of topics such as: the experience of being in a virtual world and its comparability to naturalistic traffic settings; the experience of uncertainty; the experience of receiving an e-HMI versus not receiving any; the meaning of our e-HMI designs; how to improve them; etc.

One of the essential insights from conducting the study in this specific way was that it was the hybrid of merging experimental research and ethnography that provided us with moments of clarity. Multiple times in our analysis we found that the questionnaire based ratings suggested a specific result which was contradicted by the quantitative measures and qualitative insights. Similarly, though, sometimes the ratings of certainty and comfort were really low in alignment with qualitative insights from verbal reactions and interviews, while the rating of intuitiveness and understanding the e-HMI was really high, which could only be described by utilizing the quantitative measures. Similar to the investigation of naturalistic interactions, it was again the combination of representing (virtual) reality from the outsider's and insider's perspective through a variety of methods, which allowed us to create different layers of evidence, finally providing us with a more holistic understanding when evaluating our e-HMI designs.

FINAL REFLECTIONS: CONTRIBUTIONS, IMPACT AND CHALLENGES

Our collaboration with experimental researchers at the German Aerospace Center did not only teach us about proper experimental research, but also about the limits of our own discipline. In fact the whole work with interACT taught us a lot about the limits of applying pure ethnographic work to inform technological development. Nevertheless, we also learned how to overcome some of the long discussed but nevertheless still existing challenges in interdisciplinary collaborations between the paradigms of qualitative and quantitative research. Retrospectively, our work with interACT really enabled 1) multiple forms of hybrid representations of today's traffic interactions to design e-HMI concepts for future interactions between pedestrians and AVs, as well as 2) hybrid representations of virtually simulated future interactions to evaluate our e-HMI designs. The collaboration proved to be a very promising attempt to tackle many of the challenges we introduced in the beginning of this case study, and showed that there is indeed space for ethnographers to navigate in technological development:

Contributions

1. Our ethnographic pilot study of the standard observation protocols helped interACT to develop a digital observation app which could represent naturalistic traffic interactions in a structured way; being closer to the complexity and context of reality while still providing a basis of control and comparability to make the dataset useful for the work of data scientists and human factor researchers.

2. This also showed that ethnographers as creators and shapers of datasets can be valuable sparring partners in the quantitative analysis process to reflect on what the collected data can actually tell when being transformed to multiple constructions of evidence.

3. Ethnographers as experts of immersion in different cultures and groups can provide the insider's perspective of interaction partners to represent the reality of interactions from multiple perspectives which need to be aligned and understood from a more holistic viewpoint rather than used individually when designing HMI for future interactions.

4. When using virtual reality simulations to evaluate interaction prototypes, we have presented a promising approach of ethnography to infuse some context and complexity from real-world environments, which was first identified to be the basis for why HMI might be valuable in the first place. This infusion in VR can then not only serve to verify the identified use-context for HMI to be valuable, but also lead to results that are ecologically valid for other sociocultural environments where this use-context exists equally.
5. Last but not least, combining mixed methods from ethnography and experimental research to represent the study participant's insider and the researcher's outsider

perspective on the investigated phenomenon in VR has proved to be just as valuable as the hybrid representation of perspectives in naturalistic interactions.

Impact

1. The impact of our work of infusing our worldview in the environment of interACT has yet to reveal itself entirely, as the analysis of merging the different perspectives is still an ongoing process. As one of our data science colleagues put it, maybe one day we will all find out why a subject-led ontology for the dataset is a good thing, but at first sight it seems to be of value. This very case study attempts to contribute to this understanding.

2. Nevertheless, our worldview has certainly led to critical reflections for some of the researchers at interACT which is the first step to enter discussions on:

   a) how to align multiple perspectives on the reality of interactions to inform the development of AI-based technological systems enabling valuable means of future human-technology cooperation, and to enter

   b) discussions on how to achieve ecological validity in the conduction of experimental research-based evaluations of human-machine interfaces to develop technological systems that work in the context and complexity of naturalistic environments.

Remaining Challenges

1. One of the remaining challenges is that ethnography, or anthropology in the broader sense, still lacks visibility and establishment as valuable part of any technological development process. More commonly, ethnography and anthropology should be considered as a must-have rather than a nice to add. To align multiple perspectives of reality when designing for future human-technology cooperation, our suggestion for future collaborations is to involve anthropologists and similar disciplines practicing ethnography from the very beginning in the study design phase to include their perspective in the discussion on creating datasets.

2. Another remaining challenge is to evaluate whether the ethnographic way of creating and shaping a large-scale dataset is actually better in terms of applicability for different sociocultural contexts. Since ethnography is really good in understanding the context and complexity of a certain field-site the question is to what extent using ethnography for the building process of datasets is actually better then directly relying on the ontology of former research and hypotheses.

Markus Rothmüller holds a BSc in International Business & Engineering as well as a MSc in Techno-Anthropology, and focuses on exploring and shaping future human-technology cooperation. His goal is to combine his engineering and anthropology background to innovate product development in collaboration with data science, AI and automation.

Pernille Holm Rasmussen holds a BSc in Technologies & Humanities as well as a MSc in Techno-Anthropology and is passionate about the study of human-machine cooperation. She aims at combining ethnographic methods with other disciplines such as experimental research and data science to find new ways to evaluate human-machine interfaces.

Alexandra Signe Vendelbo-Larsen holds a BSc in Market & Management Anthropology as well as a MSc in Techno-Anthropology. She focuses on a wide range of ethnographic research in both the field of human-technology interaction as well as consumer goods.

NOTES

Acknowledgements - We gratefully thank Natasha Merat and Anna Schieben for enabling and supporting our collaboration with the interACT project funded by the European Union's Horizon 2020 research & innovation program under grant agreement no 723395. Additionally, we would like to thank: Ruth Madigan for her great supervision and an amazing facilitation of our work during the data collection at the University of Leeds; André Dietrich for his great leadership of interACT's data collection and his trust in our skills to design and structure the digital observation app; Charles Fox for his introduction to data science in the realm of AV development; Janki Dodyia for her patience as sparring partner when designing and setting-up the virtual reality study; and Alexander Tesch for his great support in building the virtual world and our e-HMI prototypes in VR. We also want to thank Natasha, Anna, André, Ruth, Charles and Janki for their availability to retrospectively reflect on the collaboration in detail. Another special thanks belongs to Torben Elgaard Jensen, Melissa Cefkin, our families and everyone else who have always supported us in this long and intense journey. Thank you!

1. Next to few anthropologists, like Melissa Cefkin, being directly involved in the technological development of autonomous vehicles, anthropologists are underrepresented in this field of innovation. More generally anthropologists are still just in the uprise of being included in the technological development process, which can be concluded by the fact even latest publications have to point out the value of anthropology for technological innovation (e.g. Hartley, 2017; Madsbjerg, 2017; Roth-Lobo, 2015).

2. Emic/Etic: The emic or insider's perspective explains behavior *"in terms of the actors' self-understanding—terms that are often culturally and historically bound"* (Morris et al., 1999). It is the native's point of view. The etic or outsider's perspective explains behavior in correlation with external factors and is *"more likely to isolate particular components of culture and state hypotheses about their distinct antecedents and consequences"* (ibid.).

3. While having been working for interACT at the German Aerospace Center, we were allowed to build up our own virtual reality study and prototype our own e-HMI designs. These designs and prototypes informed the work of interACT but are not directly part of interACT, and have not been decided by interACT as such.

REFERENCES CITED

Bhagavathula, R., Williams, B., Owens, J., & Gibbons, R.
2018 *The Reality of Virtual Reality: A Comparison of Pedestrian Behavior in Real and Virtual Environments.* Proceedings of the Human Factors and Ergonomics. Vol 62, Issue 1, pp. 2056 - 2060

Blissing, B.
2016 *Driving in Virtual Reality: Investigations in Effects of Latency and Level of Virtuality.* Linköping Studies in Science and Technology Licentiate Thesis No. 1759

Camara, F., Giles, O., Madigan, R., Rothmüller, M., Rasmussen, P.H., Vendelbo-Larsen, A,. Markkula, G., Lee, Y.M., Garach, L., Merat, N., & Fox, C.W.
2018 *Predicting pedestrian road-crossing assertiveness for autonomous vehicle control.* White Rose Research Online Available at: http://eprints.whiterose.ac.uk/135432/

Cefkin, M., & Stayton, E.
2017 *Speculating about Autonomous Futures: Is This Ethnographic?* Available at: https://www.epicpeople.org/speculating-about-autonomous-futures/

Chang, C.-M., Toda, K., Sakamoto, D., & Igarashi, I.
2017 *Eyes on a car: An interface design for communication between an autonomous car and a pedestrian.* International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2017, pp.65-73

Cherry, C., Donlon, B., Yan, X., Moore, S., & Xiong, J.
2012 *Illegal mid-block pedestrian crossings in China: gap acceptance.* International Journal of Injury Control and Safety Promotion, vol. 19, no. 4, pp. 320 - 330.

Clamann, M., Aubert, M., & Cummings, M.L.
2016 *Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles.* Traffic Research Board.

Crosley Law Firm
2018 *New Insights From The Uber Self-Driving Crash Investigation.* Crosley Law Firm. Available at: https://crosleylaw.com/blog/new-insights-uber-self-driving-crash-investigation/

Dey, D., & Terken, J.
2017 *Pedestrian interaction with vehicles: roles of explicit and implicit communication.* International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2017, pp. 109–113.

Dietrich, A.
2018 *Interaction in Urban Traffic—Insights into an Observation of Pedestrian-Vehicle Encounters*, Connected Automated Driving Symposium, https://www.interact-roadautomation.eu/wp-content/uploads/interACT_Webinar_Andre-Dietrich_180509_final-1.pdf

Dietrich, A., Madigan, R., Portouli, E., Nathanael, D., Ruenz, J., Bengler, K., Merat, N., Giles, O., Lee, Y. M., Markkula, G., Camara, F., Fox, C., Rothmueller, M., Rasmussen, P. H., Vendelbo-Larsen, A., Schieben, A.
2018 *interACT Work Package 2 – How Do Traffic Participants Interact in Current Urban Scenarios and How This Helps when Designing Automated Vehicles,* Poster session presented at the Transport Research Arena 2018, Vienna

Drakoulis, R., Drainakis, G., Portouli, E., Althoff, M., Magdici, S., Tango, F., & Markowski, R.
2018 *interACT Deliverable 3.1 - Cooperation and Communication Planning Unit Concept.* https://www.interactroadautomation.eu/wp-content/uploads/interACT-WP3_D3.1_CCPU_Concept_v1.1_DRAFTwebsite.pdf

Glaser, G. B., & Strauss, A.L.
1967 *The Discovery of Grounded Theory - Strategies For Qualitative Research.* (Renewed, 1995, Reprinted 2006) Aldine Transaction, A Division of Transaction Publishers, Rutgers-The State University, 35 Berrue Circle, Piscataway, New Jersey

Hartley, S.
2017 *The Fuzzy and the Techie: Why the Liberal Arts Will Rule the Digital World.* Houghton Mifflin Harcourt

Kadali, B., & Perumal, V.
2012 *Pedestrians' gap acceptance behavior at mid-block.* International Journal of Engineering and Technology, vol. 4, no. 2, pp. 158 - 161

Langström, T., & Lundgren, V.M.
2015 *AVIP – Autonomous Vehicles Interaction with Pedestrians: An Investigation of Pedestrian-Driver Communication and Development of a Vehicle External Interface.* Masters Thesis, Chalmers University of Technology, Gothenborg, Sweden.

Madsbjerg, C.
2017 *Sensemaking: The Power of Humanities in the Age of the Algorithm.* Hachette Books

Mahadevan, K., Somanath, S., & Sharlin, E.
2017 *Communicating awareness and intent in autonomous vehicle-pedestrian interaction.* University of Calgary, Tech. Rep., 2017

Martyn, A.
2018 *Toyota executive says self-driving car technology is overhyped.* Article at Consumer Affairs, 21st of September, 2018 [Accessed 15.10.2018] Available at: https://www.consumeraffairs.com/news/toyota-executive-says-self-driving-car-technology-is-overhyped-092118.html

McNamara, L., Cole, K., Haass, M.J., Matzen, L.E., Morrow, J.D., Steven-Adams, S.M. & McMichael, S.
2015 *Ethnographic Methods for Experimental Design: Case Studies in Visual Search.* In: Schmorrow D., Fidopiastis C. (eds) Foundations of Augmented Cognition. AC 2015. Lecture Notes in Computer Science, vol 9183. Springer, Cham

Merat, N., & Madigan, R.
2017 *Human Factors , User Requirements , and User Acceptance of RideSharing In Automated Vehicles.* Discussion Paper No. 2017-10 in International Transport Forum

Merat, N., & Madigan, R., Louw, T., Dziennus, M., & Schieben, A.
2016 *What do Vulnerable Road Users think about ARTS?* in CityMobil2 final conference, Donostia / San Sebastián, Spain.

Morris, M.W., Leung, K., Ames, D. & Lickel Brian
1999 *Views from inside and outside: Integrating emic and etic insights about culture and justice judgment.* Academy of Management Review 1999, Vol. 24. No. 1781-796. Available at: http://www.columbia.edu/~da358/publications/etic_emic.pdf

Nathanael, D., Portouli, E., Papakostopoulos, V., Gkikas, K., & Amditis, A.
2018 *Naturalistic Observations of Interactions Between Car Drivers and Pedestrians in High Density Urban Settings*, Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). IEA 2018. Advances in Intelligent Systems and Computing, vol 823. Springer, Cham

Norman, D.
2013 *The Design of Everyday Things.* Revised and Expanded Edition, Basic Books, a Member of the Perseus Books Group New York.

Parkin, J., Clark, B., Clayton, W., Ricci, M. and Parkhurst, G.
2016 *Understanding interactions between autonomous vehicles and other road users: A Literature Review.* Project Report. University of the West of England, Bristol. Available from: http://eprints.uwe.ac.uk/29153

Rasmussen, H. P., Rothmüller, M., & Vendelbo-Larsen, A.
2017 *Constructions of Individual Crossing Practices - A Novel Approach To Understanding Pedestrians' Decision Making Processes in Crossing Situations to Inform the Design of e-HMI Solutions in Autonomous Vehicles.* School of Architecture, Design and Planning, Aalborg University Copenhagen

Rasmussen, H. P., Rothmüller, M., & Vendelbo-Larsen, A.
2018 *Designing Future Interactions between Humans and Autonomous Vehicles - From Techno-Anthropological Field Observations to Prototypes of External Human-Machine Interfaces Tested in Virtual Reality.* Master's thesis, MSc Techno-Anthropology, School of Architecture, Design and Planning, Aalborg University Copenhagen, Denmark (confidential)

Rasouli, A. & Tsotsos, K. J.
2018 *Autonomous Vehicles that Interact with Pedestrians: A Survey of Theory and Practice*. IEEE Transactions on Intelligent Transportation Systems

Roth-Lobo, H.S.
2015 *What Anthropology Brings to Innovation: John Sherry / A Profile*. Ethnographic Practice in Industry Conference. Available at: https://www.epicpeople.org/john-sherry-what-anthropology-brings-to-innovation/

Sale, J.E.M., Lohfeld, L.H., & Brazil. K.R
2002 *Revisiting the Quantitative-Qualitative Debate: Implications for Mixed-Methods Research*. Quality & quantity 36.1 (2002): 43–53. PMC. Web. 6 Oct. 2018.

Sangari, A.Z., Taniberg, A., Retoft M.F., Rothmüller, M., Rasmussen P.H.,Vendelbo-Larsen, S.A.,
2016 *A Socio-Technical Perspective on Autonomous Vehicles*. School of Architecture, Design and Planning, Aalborg University Copenhagen

Schieben, A.
2018 *Management Summary, International eHMI workshop 19th of April 2018, Vienna,* https://www.interactroadautomation.eu/wpcontent/uploads/interACT_ManagementSummary_eHMIWorkshop.pdf

Schieben, A., Wilbrink, M., Kettwich, C., Madigan, R., Louw, T., Merat, N.
2018 *Designing the interaction of automated vehicles with other traffic participants: design considerations based on human needs and expectations*, Cogn Tech Work, Springer Online, https://doi.org/10.1007/s10111-018-0521-z

Sobhani, A., & Farooq, B.,
2018 *Impact of Smartphone Distraction on Pedestrians' Crossing Behaviour: An Application of Head-Mounted Immersive Virtual Reality*. Journal published: Transportation Research Part F: Traffic Psychology and Behaviour, Acceptance date: June 17, 2018 https://arxiv.org/ftp/arxiv/papers/1806/1806.06454.pdf

Stanton N. A., Salmon, P. M. Rafferty L. A., Walker G. H., Baber C., & Jenkins D. P.,
2013 *Human Factors Methods: A Practical Guide for Engineering and Design*. Human Factors Methods. London: CRC Press. https://www.taylorfrancis.com/books/9781317120162

Stayton, E., Cefkin, M., & Zhang, J.
2017 *Autonomous Individuals in Autonomous Vehicles: The Multiple Autonomies of Self-Driving Cars*. Ethnographic Praxis in Industry Conference Proceedings, ISSN 1559-8918, epicpeople.org/intelligences

Suchman, L.
2007 *Human-Machine Reconfigurations - Plans and Situated Actions*. [2nd edition] Cambridge University Press

Turkle, S.
2009 *Simulation and Its Discontents*. Cambridge, MA: MIT Press.

Verbeek, P.P.
2015 *Beyond Interaction: a short introduction to mediation theory*. Available at: http://interactions.acm.org/archive/view/may-june-2015/beyond-interaction [Accessed 5 December 2016]

Vinkhuyzen, E. & Cefkin, M.
2016 *Developing Socially Acceptable Autonomous Vehicles,* Ethnographic Praxis in Industry Conference Proceedings, p. 522–534, ISSN 1559-8918, https://www.epicpeople.org

Wilbrink, M., Scheiben, A., Markowski, R., Weber, F., Gehb T., Ruenz, J., Tango, F., Kaup, M., Willrodt, J.H., Portouli, V., Merat, N., Madigan, R., Markkula, G., Romano, R., Fox, C., Althoff, M., Söntges, S., Dietrich, A.
2017 *interACT D1.1 Definition of interACT use cases and scenarios*. Available at: https://elib.dlr.de/116445/

Wilbrink, M., Schieben, A., Kaup, M., Willrodt H.J., Weber, F., & Lee M.Y.
2018 *interACT deliverable 4.1 - Preliminary Human Vehicle Interaction Strategies V1.* Available at:: https://www.interact-roadautomation.eu/wpcontent/uploads/interACT_WP4_D4.1_Preliminary_Human_Vehicle_Interaction_Strategies_v1.0_draftWebsite.pdf

Wilson, C.J., & Soranzo, A.
2015 *The Use of Virtual Reality in Psychology: A Case Study in Visual Perception.* Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2015 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4538594/pdf/CMMM2015-151702.pdf

Yannis, G., Papadimitriou, E., & Theofilatos, A.
2013 *Pedestrian gap acceptance for midblock street crossing*. Transportation Planning and Technology, vol. 36, no. 5, pp. 450 - 462.

Revitalising Openness at Mozilla: A Mixed Method Research Approach

RINA TAMBO JENSEN
Mozilla

This is a case about how Mozilla, the open source browser company, set out to reconnect with 'collaborating in the open' to regain its competitive advantage. This case describes how a multi-disciplinary research team used ethnographic, market, and data analysis to articulate and clarify the problem, and build a strategy towards revitalizing Openness at Mozilla. It will aim to prove that the subsequent change achieved could only have been accomplished by a mixed method research approach. And importantly show, how the team used data to prove the distribution of findings, coupled with ethnography to shine light on the why and how of those findings. The case study will do this by discussing the key insights and how these fueled recommendation and subsequent change in the organisation.

The project presented many problems: from convincing stakeholders of the need to fully explore the problem, to connecting widely different research methods and gleaning insights that built strongly on all strands of research. Overcoming these issues together, the team managed to outlay the key problems and opportunities for Mozilla and build a strategy which has affected real change in the company and is now being implemented by a team of 20 employees. We call this strategy Open by Design. Which, to us, means being open, purposefully.

> David (not his real name), a contributor to Firefox through 4 years, was getting ready for his day. The sun was shining outside, spring was clearly in the air. This morning, as all mornings, he went to his small study with his coffee and turned on his computer. David enjoyed his morning routine immensely. The hour or two, before heading to work, where David could plug in and work on some of the projects *he* loved. This was his time, no one could tell him what to do and he decided what mattered. Mostly over the last 4 years it had been working on Firefox, but many other open source projects had from time to time taken his attention, but he always came back to the browser. The people there were his friends and his peers. He knew many of them personally. This morning should be like the rest, but as he sat down David felt different. Lately he had begun to reflect on the work; increasingly the core contributors he had worked with had left the project, overall there were more staff and fewer contributors, and it had become increasingly hard to get things through and see your code in the final product. Most of the time it felt like he was banging at an increasingly closed door. He remembered fondly the old days. Actually, he thought to himself, he wished things were different, but he was starting to question his ability to change them. He starred at the blinking white cursor on his dark terminal screen for a while, an empty window waiting for his command. He looked solemnly around the room, and instead closed his computer. For the first time, in 4 years, David got up from his chair without checking in on IRC or Github, and went straight to work. This morning definitely felt different.
>
> David, is one of many contributors who over time left the Firefox project. Open Source projects have a natural churn but this was different, Mozilla was feeling like it was experiencing a decline in contributors. And this at a time where it needed them more then ever.

BACKGROUND

When in 1998, Netscape made its code open source, it was not just a fun gimmick, it was groundbreaking. Essentially the company had just 'given away' their source code. (Festa, 1998). Back then a hopeful software engineer commented to CNET journalist Paul Festa:

> "You're starting out with an industrial-strength browser and now you have many qualified developers working from it. I imagine Netscape's developer teams can't compete with Microsoft in terms of resources, but with Netscape working with outside developers, I think you're going to wind up with a very good product" - interview from 1998 (Festa, 1998).

He would be proven right in his prediction. While the first years of Netscape, and then subsequently the Mozilla project, were tumultuous, a small group of employees and contributors battled through and after 6 years they released Firefox 1.0. (Shankland, S., 2018)[1].

When Mozilla released Firefox 1.0, it was a better, simpler browser - and the only alternative to the then current monopoly holder; Internet Explorer. The success of the project was not immediate, but over time, it grew, and once momentum took off, there was no going back. The browser market changed forever[2], and in so doing, Mozilla had proven that Open Source, or open collaboration, could work at scale.

It was clear to anyone who watched the subsequent growth of Firefox (Levy, Feb 2008) that open source was Mozilla's main competitive advantage. Throughout the browser wars, Mozilla's Firefox consistently prevailed over the competition through a deep focus on open practices and democratised ways of building technology. Contributors (volunteers on the project) translated the browser into over 124 languages[3]; a key browser feature was invented by a contributor - the Tab[4] - allowing users to easily work across many sites at once. Contributors tested the product and provided crucial support to users with questions. They multiplied the strength of the few Mozilla employees and made the browser the free, open browser that millions of users use today. The success story of Firefox had been turbocharged by open collaboration and open source.

The success of Firefox and Open Source inspired many upcoming entrepreneurs in the tech industry, including some of the major players today. As such, by 2015, open source was no longer a radical movement, having grown industrial in scale (Newman, 1999). Companies like Google, Facebook and Microsoft (Nomas, Marts, 2018), currently use open source licenses to release their code (as an example Google currently has over 2,000 open source projects released).

While we may not think much about it in our daily lives, many of the key platforms we all rely on today are open source; from Android to Wordpress. Companies adopted open source widely and, in so doing, explored new and different configurations of what open source could be, thereby further developing open source structures and archetypes. (Kripalani, 2017, Open Tech Strategies, 2018)

Over time, Mozilla had integrated open source as a key part of its identity as well. However, as the company grew, things increasingly became open source in name more so than in action and while code / documentation was released, there was less and less focus on engaging the communities around the projects. Slowly, the company began building things internally and with dwindling focus on building the communities around the projects (Discourse Post, 2017).

No one person at Mozilla set out to minimise collaboration outside the company; many factors contributed to the shift. While open source had proved great for smaller projects; the company was now larger, and the projects were mature products, with a substantially increased complexity of code and hundreds of millions of users.

> ”The general feeling in the community is that Mozilla has lost its way and doesn’t care about community anymore. This may impact influencers more. Which is really not good as these are the people who recommend what to install. We have lost a lot of those people to Chrome.” - Sr. Product Manager - Firefox.

The stories of how the company had won through open source still fueled the company culture (Briody, 2015), but the stories and narratives were no longer prevalent, and more crucially, they were hidden away deep within certain pockets of the company.

> “We now have a significantly large enough population of folks at Mozilla that don’t have that (open) history, that don’t know the history of Mozilla being built by contributors... ” - Sr. Staff Engineer - Firefox

The sense that Mozilla had lost its focus on its main competitive advantage - collaborating in the open - was becoming an increasing concern to employees and leaders.

This change presented a very real conflict for the organisation. While the company had grown in size from when it first released Firefox, it was obvious that it would never be big enough to directly compete with its main competitors on resources nor funding.

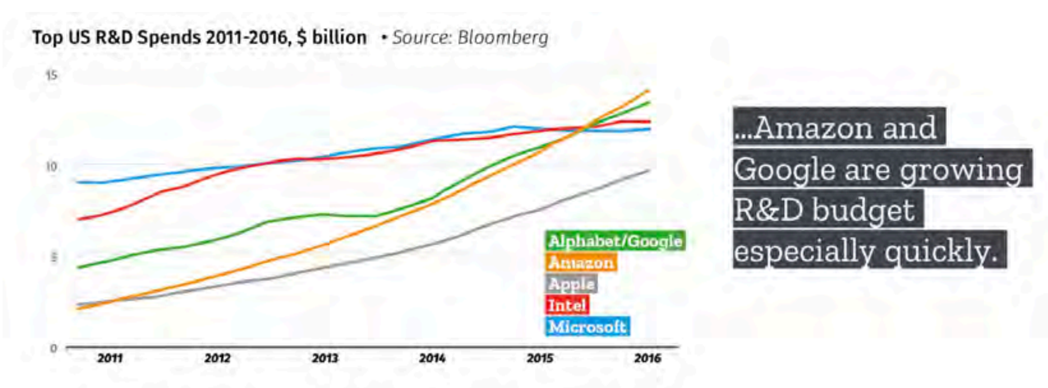


Figure 1. Top US R&D spending 2011-2016. Source: Bloomberg.

It was clear to leadership that there was no single winning strategy that didn’t include collaborating with the ecosystem of the open web, the very same ecosystem that had helped Mozilla succeed before.

In 2016, the leadership took a stance and set a clear goal for the company; Revitalize Openness at Mozilla. To deliver on this ambitious goal, CEO Chris Beard brought in long term board member Katharina Borchert.

Knowing what questions to ask

The goal of revitalizing openness at Mozilla meant identifying and implementing opportunities for Mozilla to make 'collaborating beyond company borders' a competitive advantage again.

While the overall impact of the problem was clear to leadership, the cause and effects were less so. Over time, the projects and people focusing on open collaboration had shifted and moved. As a result no one had a clear overview of how much, who, where, or why, open collaboration was happening on projects across the organisation; nor why or if it wasn't being as effective as it used to be.

Identifying new opportunities first meant that some exploratory questions had to be answered. The group narrowed down to three key research questions:

- What is the current state of open collaborations at Mozilla?
- What, if anything, has changed overtime?
- What are other successful examples of open collaboration that Mozilla can be inspired by?

Choosing the right lenses

Based on the goal at hand, it was clear the problem needed to be understood from many different perspectives. The team wanted the work to have a solid grounding, so a multi-disciplinary team were brought together to deliver a mixed method exploration; consisting of ethnographic, market and data analysis researchers.

METHODOLOGY

In the end, four streams of research were initiated to answer the project's research questions (Edwards, 2010, Greene, J.C., 2006).

1. Internal Interviews & Observations - The in-house ethnographic research included 38 interviews with staff across Mozilla. Questions focused on: the interviewee's role in the company and their work with contributors, the perceived value of external collaboration or "working open" to Mozilla, and potential opportunities for 'open collaboration' within their field. The interviews were conducted on site, where possible, and remotely over video conference software, where not. All interviews were recorded, transcribed and archived. This stream of research was initiated before the others, in order for the team to validate the key research questions and build stakeholder support for the project from the get-go. This early round of stakeholder input enabled the team to move beyond the 'researcher mindset' and ensure the problem space was validated organisation wide (Swanson, R. & Holton, E., 2005, p 10-18).

2. Competitor Research: Competitor analysis of leaders in the space - An early finding from the staff interviews was a general feeling that Mozilla had fallen behind on open source ways of working and size of community. It therefore became clear there was a need to understand external companies' success in open collaboration. Competitor analysis and interviews were done with 7 leaders in open collaboration to explore how they were strategically using open methods in their business. In total 30 interviews were conducted with staff from across the companies. The seven companies were: Automattic, Arduino, Kubernetes/CNCF, 23andMe, NASA, Aleph Objects and Sage Bionetworks. The interviews focused on their internal structures, motivations for external collaboration and the resulting impact. The interviews were conducted in collaboration with Copenhagen Interaction Institute of Design (CIID).

3. Contributor Survey - Focusing on our existing contributor base, and ensuring their perspectives were represented, was a core focus for the team. To this end, a survey instrument was employed. The team had in 2016, through interviews and observations, explored motivations behind contributions. This research was used as a basis for the survey design, thus ensuring that the instrument was built on well-grounded findings[5]. The survey was designed to explore what projects contributors were working on, how they engage with Mozilla and why. The survey target audience consisted of three segments: active contributors, past contributors, and never contributed (lurkers). To reach all three, the survey was distributed widely among the communities and had over 1700 respondents, of which, 1019 were complete (n=1019).

4. Data analysis of past contributions over 16 years - Getting to understand the historic data of contribution was the real challenge for the team as it had never been done before. Luckily one of the teams working directly with contributors had recently begun early explorations with a outside company named Bitergia. Together they started to measure code contributions on the platforms Mozilla primarily used (Bugzilla and Github). Together with Bitergia, an ambitious project was initiated to see if this prototype system could provide contributor data on all Mozilla hosted projects. In the end, the team succeeded in providing data on code contributions across Mozilla's many projects spanning the previous 16 years and sorting out employee data from non-employee data. The subsequent quantitative analysis of the data sources focused on the development of Mozilla products and technology, including: source code repositories, issue tracking systems and asynchronous communication systems. The study was structured across several areas of inquiry: activity (volume of activity of contributors), community (people contributing), processes (efficiency of dealing with issues), attraction / retention (how the project attracted and retained its contributors), levels of contribution, and gender diversity. In order to cover most contribution areas, public data sources from code (git, github), bug tracking (bugzilla, github issues) and discussions (mailing lists and discourse) were also used[6].

Breaking down the methods by research questions

To ensure the team had a clear overview going into the research project each method was mapped out against the key research questions. Each stream of work had its own research design documentation [including sub research questions].

Table 1. Research Questions by Method

Research Activities / Key Research Questions	**What is the current state of external collaborations at Mozilla?**	**What, if anything, has changed overtime?**	**What are other successful examples of open collaboration?**
Internal Interviews & Observations	What are staff perspectives on external collaboration at Mozilla? What motivates or blocks staff from engaging today?	What are staff perspectives on external collaboration at Mozilla?	What opportunities and challenges do staff see for external collaboration?
Data analysis of past contributions over 16 years	What is our current state of affairs (who does what, where)?	How have contributions actually evolved over time?	
Competitor Research			What benefits and challenges are realised by external collaboration?
Contributor Survey	What is our current state of affairs (who does what, where)? Who are contributors are and why they contribute?	Understand what projects people are working on today and why?	Who are the contributors and why do they contribute?

Timeline

Below shows when the different research projects were conducted in relation to each other.

Figure 2. Timeline of research activities.

Tiburon Workshop

To bring the streams of research together, a week-long workshop was held in Tiburon, CA. The participants in the workshop consisted of the Mozilla project team and CIID as well as selected key stakeholders from across the organisation who had been involved or interested in the work throughout the research.

ANALYSIS

The analysis of all the different research streams presented its own challenge - as different researchers had been on different projects, pulling together the findings to identify commonalities or variations was crucial in order to tell a coherent story and to gain the required insights.

The Tiburon Workshop was viewed as the team's chance to pull together insights from across the research streams and to build the final strategy recommendations. A great deal rested on the team's ability to deliver in this single week.

In preparation for the workshop, each research stream was tasked with pulling together an 'interim results report' as well as some key learnings / recommendations as pre-reads.

Every workshop participant also had to write their own summary of insights, going into the week, therefore allowing all participants to have detailed understanding of findings across the reports. Another smaller, but highly beneficial tactic the team employed, was getting a large space for the group and covering the walls in floor to ceiling foam boards with the key research findings, allowing the team to physically see the results throughout the week.

Table 2. Tiburon Schedule

Monday	Tuesday	Wednesday	Thursday	Friday
Setting the week up. Review and synthesis of research begins.	Continued work, including review and synthesis. Insight areas deep dive.	Connecting the insights from different research streams into one set of findings with multiple sources.	What does it mean: From insight to Opportunity.	Strategy deep dive. Agreeing on recommendations and ways forward.

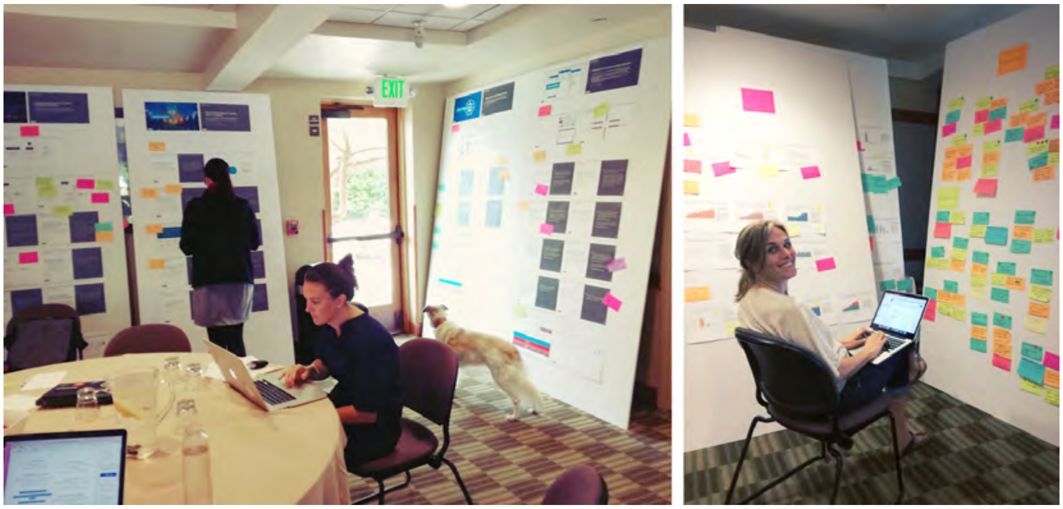


Figure 3. Analysis week.

By the end of Wednesday, it became clear that two days was not long enough for the team to cover opportunities and recommendations. As a last minute change, a small group were assigned individual insight areas and asked to write up suggested actionable next steps and recommendations. These would act as a foundation for the team's discussions the next day. This last minute change in program meant more work for the team members who worked through the evening on building the draft recommendations, but, as the team came back together Thursday it was clear that this work had paid off. The draft recommendations became a great foundation for discussions and, subsequently, most of Thursday was spent reviewing, and further building upon them, as a team. This ensured that, by end of Friday, the team had successfully been able to work from insights in to opportunities and then through to strategic recommendations that outlined how we could achieve the final goal; Revitalising Openness at Mozilla.

While synthesis workshops are always notoriously hard to structure, a few key lessons were learned: firstly, ensure you have enough time to wrap up findings. Secondly, ensure there is broad agreement to recommendations, allowing all to have input[the goal is to build consensus]. Thirdly, allowing stakeholders to start talking about findings immediately after the workshop. This helped build momentum for the work and - even if not perfectly presented - can become a change making catalyst, in and of itself, within the organisation.

Figure 4. Analysis week.

FINDINGS

By September the team was able to present the *first ever,* in depth look on open collaboration and contribution across Mozilla. The findings were intriguing and helped break through several mental glass ceilings that had existed within Mozilla for years.

While the total findings include more than can be shared in this case study; what follows will cover a few of the key highlights and how they influenced the final strategy and recommendations.

Understanding the History and Status Quo

Contribution at Mozilla is steadily growing - The first thing the team identified was how some of the prevailing assumptions across Mozilla (and within the research group) were simply not true. The internal interviews showed a general acceptance that Mozilla had become worse at working in the open but the data analysis, however, showed that contributions, over time, had actually increased slightly. Breaking this mental model of declining external participation was one of the most important insights from the work as it

Key Finding
The question needed to be asked differently; instead of asking 'why did we lose our external collaboration', Mozilla needed to understand 'how do we ensure all projects, across Mozilla, 'learn from the projects which are doing well? '

allowed the conversation within Mozilla to change from one of hopelessness and negativity, in to one of opportunity.

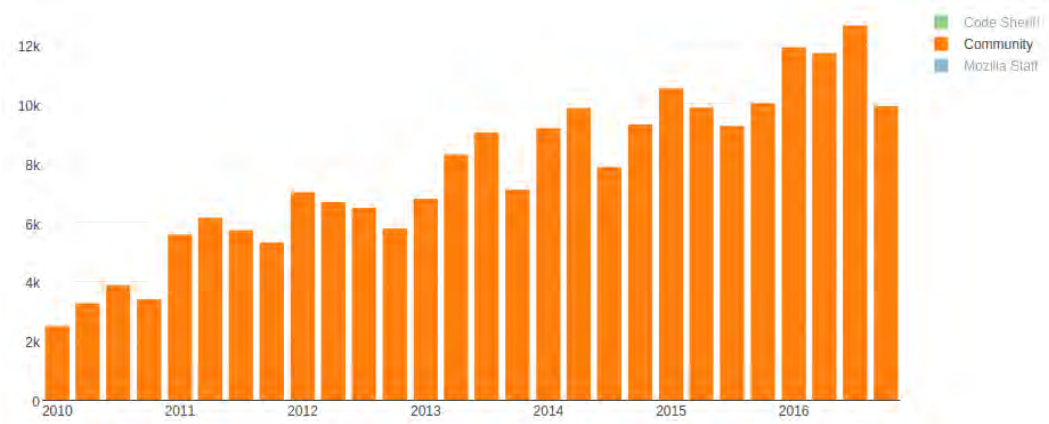


Figure 5. Chart: commits by non-employees over time, per quarter.

The perception of a declining trend had, however, not come from nowhere; Fewer projects were working in the open across Mozilla, leading staff to experience an overall downtrend. Only 6 projects now accounted for over half of contributions by non-employees. Contribution, it seems, had not been declining, instead it had concentrated around fewer projects, with some of those projects doing much better than others.

Mozilla is not one community - The findings from the internal interviews had shown a mental model of contributors as a singular network; 'the Mozilla community'[7].

However, the research showed us that there is not one 'singular' Mozilla community. In fact contributor activity on projects tends to be siloed. We were able to identify 6 distinct siloed communities that had little to no overlap between the contributors.

Their differences were not related simply to one specific project nor skillset, but also to motivations, operational norms, social capital, feelings of affiliation, and more.

Key Finding
Understanding the differences between contributor communities and the reasons behind their respective behaviors is fundamental to improving engagement, retention, and providing collaboration opportunities of mutual benefit.

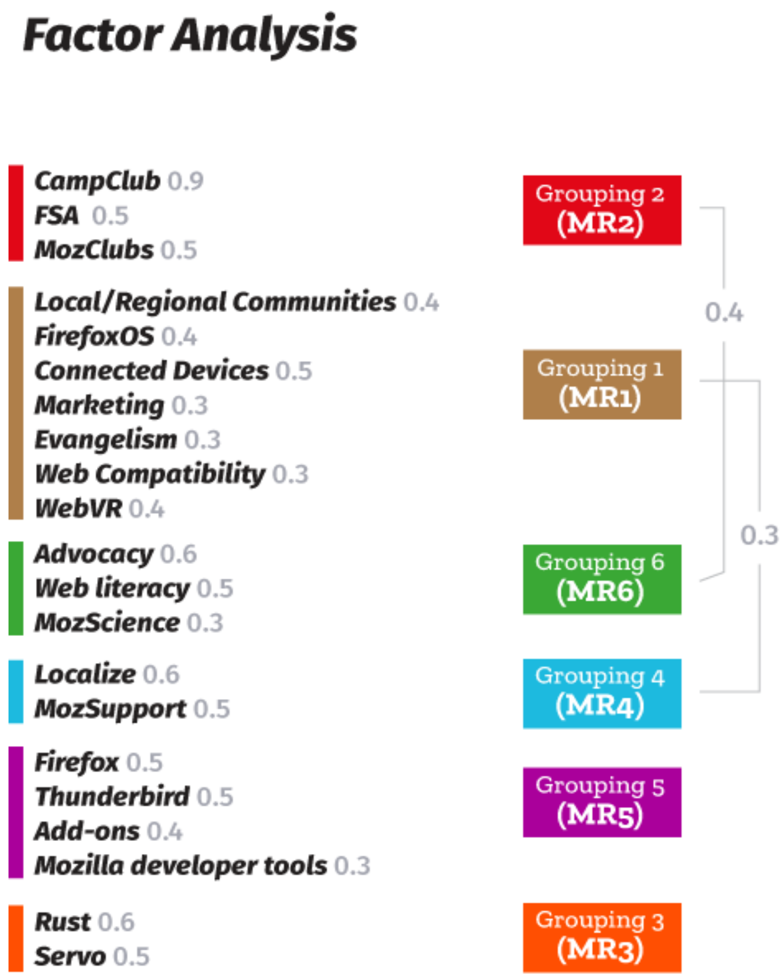


Figure 6. Factor Analysis of Project Contribution / Affiliation (self reported) showing 6 distinct groups. Source: Communities & Contributors Qualitative Survey, 2017

An internal hunger for Open is paired with a sense of skepticism - The organisational research identified a real internal interest for open ways of working - something that had been hypothesised to no longer be part of the Mozilla culture. Out of the 38 interviews conducted internally, 35 spoke positively about the opportunities and experiences of external collaboration. One Staff Engineer expressed it best:

> "The benefits of external collaboration are that it brings in bold experiments - and much more diverse sources of input than any company could do themselves: people from different countries, socioeconomic backgrounds, different vocations... There's an open source saying that "with lots of eyeballs, all bugs are shallow". The raw numbers brought to bear on fixing problems -- and on discovering problems that can thus be fixed."

However, findings also showed that Open Source practices had become increasingly decentralised in the organisation and had, subsequently, lost internal status and recognition.

And although the overall data showed an increasing trend of contribution, the staff interviews highlighted many things, internally, that no longer worked, as well as new opportunities that had never been taken advantage of before:

- Product decisions made behind closed doors does not allow us to take advantage of the diverse community. A staff engineer commented: “Firefox ships to many people, but highly technical people are making decisions within Mozilla that may not resonate with a wider field of users.”
- A move from open to closed tools (discourse, 2017); the integration of slack was highly beneficial for staff but took staff off IRC and into closed networks (slack) the contributors couldn’t easily access.
- Short term goals (quarterly) of staff did not leave room to think about community or ecosystem engagement and led to misalignment with community interest.

The hunger for open was clear, but mental models showed a fair deal of skepticism around the reality of open source: Open collaboration was never an easy fix. A Sr Engineer commented:

> “There are people who make an enormous number of bug contributions, but most of them are just noise… yet there are people who show up once every three months with five lines of code to fix something that would have cost us a lot of man hours.”

Identifying new opportunities of open collaboration

Key Finding
To succeed in revitalizing open at Mozilla, the team would have to answer: How do you find the value in the noise?

The research didn’t just shine a light on the current status of contribution and openness at Mozilla, it also highlighted areas of improvement for the company. These are some of the immediate opportunities that were discovered.

Internal mental models of contributors relate to individual contributors, at the exclusion of thinking about the partners and companies we can collaborate with... - Consequently, we focus on intrinsic motivations for contribution, rather than extrinsic and economic motivations for partners. Partnering with companies and organisations has allowed a company like Arduino to evolve products to meet needs of its diverse users. And for both, Automattic and Kubernetes, the contributing engineers are to a high degree employed by partners - therefore, ensuring wide implementation amongst the set of contribution organisations, allowing the projects to build a strong case for becoming a standard technology or product. Furthermore, focusing on partnerships was a way for all of the companies interviewed to minimise cost and spread risks.

> “NASA wants companies to form and generate these products because we don't want to generate them ourselves and we get an advantage if they're commercialised and the price

> goes down,…we had to shut down Space Shuttle because we couldn't sustain the custom companies (red: internally)" - Deputy Director – CoECI NASA

Different projects attract different profiles of contributors and lend themselves to different forms of contribution - The findings highlighted big differences between type of contribution, motivational factors, and demographics across projects. This, along with the insight that contributor experiences should be designed for the target audience, made it clear to the team that Mozilla's previous 'one size fits all' way of doing open source, or open collaboration in general, did not work. The community is simply too diverse and differs too greatly across projects.

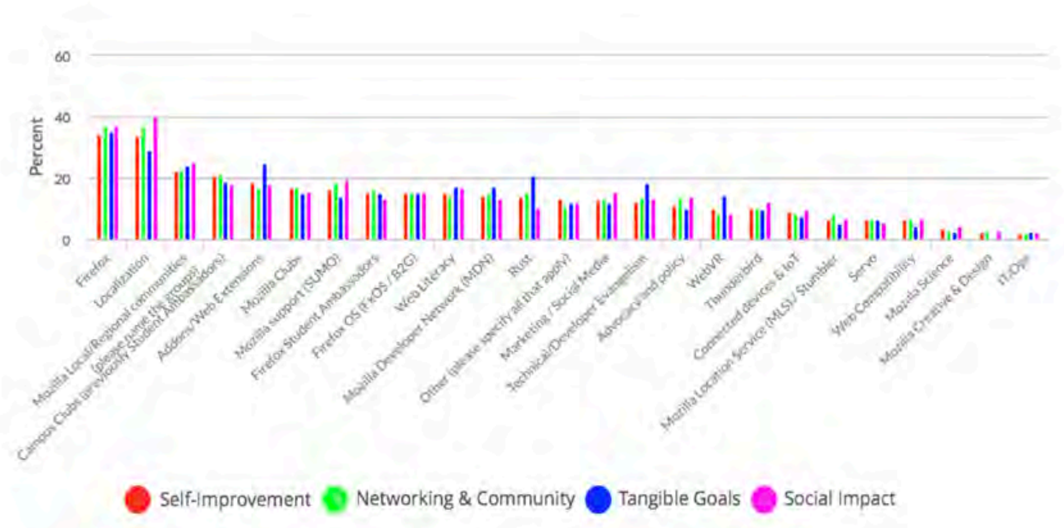


Figure 7. Project Affiliation by Motivations for active contributor

Designing for community diversity supports ecosystem growth - Ensuring that a diverse set of people can be invited to participate in projects turned out to be a key focus for other companies. The companies had instigated many methods to improve this - from a welcoming presence for new contributors, through to allowing contribution across a range of skill sets and ensuring that their community managers focused on spreading the message of diversity and tolerance in online and offline forums.

> Matt believes in open source democracy, but not open democracy. He's not excited about tolerating trolls ... he's not so committed to openness that he would let it affect the organisation or culture. - Simon Phipps, Managing Director Automattic – Meshed Insights / Open Source Thought Leader.

When the team compared this insight to our data analysis findings, the results did not look good. Especially on gender, Mozilla performed poorly. Gender diversity is a known issue in open source, but the Mozilla community was found to be performing worse than others (who were already performing badly).

Table 3. Comparison of female contributors in source code repositories for several FOSS projects[10]

	OpenStack	Linux Kernel	Hadoop Ecosystem	Mozilla
All history	839 (10.63%)	1,150 (8%)	129 (7.5%)	822 (5.5%)
Last year of analysis	422 (11.53%)	330 (9.9%)	71 (8.5%)	282 (6.5%)

Key Finding
Focus on diversity and inclusion as a way to ensure open source community health

Focus on the community experience - The external research showed how other companies had built open practices into their processes by, building experiences for contributors, thinking modularity into their processes, and understanding the value of the crowd in the contribution experience.

From the very beginning, Arduino has focused on enabling people to realize their ideas. Over the years their determination to continually improve the user experience and make it accessible for 'newbies' has created an inclusive system that appeals to a larger audience. Continual and assiduous community engagement, through teaching and rapid prototyping, has been crucial in developing their design & feature strategy.

Key Finding
Design the contribution experience to ensure retention and engagement.

OUTCOMES

This project provided the first ever exploration into the true value of contribution at Mozilla. Over 16 years of historical data was analysed, as well as a wide survey of existing contributors and staff interviews. For the first time, it was possible to build a full picture of 'contribution' whilst also allowing contributors to tell their stories. This research helped steer Mozilla away from making decisions based on assumed mental models. It also gathered employees around the call to action to re-invent 'open collaboration' and usher in a new era of Open at Mozilla based on deep research.

The findings formed the basis of a new strategy for Mozilla towards open collaboration; dubbed Open by Design. It steered the company to focus on 4 areas of improvement based on the research:

1. Build out processes for engaging partners as well as individual contributors.

Compared to other Open Source organisations, Mozilla's internal mental models were shown to skew towards thinking of contributors as 'individuals' at the cost and exclusion of thinking about the ecosystem and partners we can influence. This led to a new program being created, now referred to as the Open Source Strategy program. Its purpose is to identify external partners fostering collaboration on Mozilla's open source projects.

2. Engage in open practices across the product life cycle, with a deep focus on experience design.

The findings showed a current lack of systematic evaluation of community input, as well as a lack of focus on how we design for engagements with contributors. This has led to the creation of a Service Design team currently working across open source and crowdsourcing projects to design the experience touchpoints for external collaboration.

3. Introducing an in-house process for identifying what models of openness to implement, depending on audience and needs.

A key finding from the research showed that different projects attract different profiles of contributors, which leads to different forms of contribution. This drove the Open Source strategy program to start exploring how Mozilla thinks more differently about how it structures new Open Source programs. Following on from this, the team recently launched the Open Source Archetypes[8] work which is now being employed by our internal R&D group.

4. Introducing best practices in Open Innovation and Open Source

Including diversity and inclusion for Open Source group health metrics and community development best practices. The findings underlined that designing for community diversity supports growth and health in the communities. It also showed that there is a general lack of success metrics for open source projects. The community development team has, in recent months, focused on solving for this; including, partaking in CHAOSS[9] efforts to deliver health metrics for Open Source and focusing on D&I best practices, including a monthly community call to share the results.

SUMMARY

This case study aims to show some of the unique benefits of combining research methods, such as data analysis and ethnographic research. It further aims to outline that a team of data analysts and ethnographers can benefit from collaboration, not only with each other, but also with business strategists and market researchers.

The team would never have got the buy-in it needed (as proven by the ethnographic research conducted in 2016) with only one lens on this problem. Data was needed, but data alone could not tell the full story.

Achieving the diversity of skills may not always be possible but, when we strive to, it can bring the exploration of business problems to new levels of impact.

LEARNINGS

Leading a large research project such as this was challenging for all involved. Not only was it the first time that Mozilla had invested so much time and money into understanding its communities, it was also the first time that the team worked with such a diverse set of expertise and backgrounds.

Ensuring we had C-level buy in for such a large-scale project took time and, in hindsight, it would have been beneficial to spread the research over a longer period.

This project is a great example of how an interdisciplinary research team can achieve true company change through the application of mixed method research as well as, employing data science and ethnography to show, not just the distribution of findings, but also the why and how. It helped the team and the wider company to see the value of mixed method, by looking beyond the data and crucially explore the deeper reasons and connections behind it.

Rina Tambo Jensen is an experienced design researcher and service designer working for Mozilla in San Francisco.

The team

Rina Tambo Jensen • Lead Researcher
Patrick Finch • Strategy Lead
George Roter - Community Development Lead
Susy Struble • External & Internal Project Manager
Pierros Papadeas • Open Source Expert
Rubén Martin • C&C Project Manager
David B. Schwartz • Researcher

External Partners

CIID (Copenhagen Institute of Interaction Design) • External Research
Bitergia • Data Platform Analysis

Mozilla's Contributors

And as always to our great volunteers all across the world!

Photo credit: William Quiviger / CC BY-NC-SA

Photo credit: William Quiviger /CC BY-NC-SA

Photo credit: Irayani Queencyputri / CC BY-NC

Photo credit: Mozilla India / CC BY

NOTES

1. https://www.youtube.com/watch?v=yyyoz3yGqrw

2. https://upload.wikimedia.org/wikipedia/commons/7/74/Timeline_of_web_browsers.svg

3. https://pontoon.mozilla.org/projects/firefox/

4. https://allthatiswrong.wordpress.com/2011/01/14/opera-did-not-invent-tabbed-browsing/

5. The initial contributor story presented in this article is also taken from that work

6. Note that some sources couldn't be gathered for this study and are not considered in this report: Mercurial repos not mirrored in github (mainly l10n), contributions through MDN, SUMO or AMO sites and Marketing tools. Firefox and Gecko code were considered as the same group for this version of the report, we will look into separating them for the next iteration.

7. https://wiki.mozilla.org/Community

8. https://blog.mozilla.org/wp-content/uploads/2018/05/MZOTS_OS_Archetypes_report_ext_scr.pdf

9. https://www.linuxfoundation.org/blog/chaoss-project-creates-tools-to-analyze-software-development-and-measure-open-source-community-health/

10. Female contributors is the fraction of developers identified as female, leaving out those identified as male, or unknown.

Definitions

- Active: To have contributed to Mozilla or related project in the last year.
- A Contribution: Is also self reported and could be reported via our different contribution
- Community: Used to refer to the survey population.
- Contributor: The case of the survey a self selection definition. I.e. if people have said they have contributed we assume this is the case and ask them about it. There may therefore be some self selecting bias in this group. We've mitigated for the bias by cleaning up Actives who have not noted any contribution areas or types.
- Contribution Area: The area within Mozilla or related project they contribute too
- Contribution Type: The activity they contribute with to that area. Full list seen in contribution types.
- Inactive (sometimes Past): Have contributed to Mozilla or related project longer than one year ago.
- Never Active: People who have signed up for information but never interacted with Mozilla or related project.

REFERENCES CITED

Andersen, K.
2009. Ethnographic Research: A Key to Strategy. HBR.

Bourdieu, P & Wacquant J.
1992 An invitation to Reflective Sociology. Washington University.

Swanson, R. & Holton, E.
2005 Foundations and Methods of Inquiry - Research in Organisations. Berrett-Koehler Publishers.

Greene, J.C
2006 Toward a methodology of mixed methods social inquiry . Research in the Schools, 13(1), 93-99.

Brewer, J., & Hunter, A.
1989 Multi-method research. Thousand Oaks, CA: Sage

Edwards, Gemma
2010 Mixed-method approaches to social network analysis. Discussion Paper. NCRM.

Newman, N.
1999 The origins and future of open source software.

West, J., & Gallagher, S
2006 Patterns of open innovation in open source software. Open Innovation: researching a new paradigm, 235(11).

Thakker, D. et al
2017 Tracking the explosive growth of open-source software. TWC. April 7 2017.

Shankland, S.
2018 Mozilla's radical open-source move helped rewrite rules of tech. CNET. 2018, March 29.

Discourse Post, Kairo,
2017 Thoughts on Mozilla using Closed-Source Software. Discourse. https://discourse.mozilla.org/t/thoughts-on-mozilla-using-closed-source-software/19000/9

Bertel King, Jr
2016 March 28, 2016, Makeuseof.com: “Is Android Really Open Source? And Does It Even Matter?

Raúl Kripalani
2017 An analysis of the licenses used by 75+ Open Source projects across 35 companies

Jim Hamerly and Tom Paquin with Susan Walton
1999 Open Sources: Voices from the Open Source Revolution”. O’Reilly

Steven Levy
2012 The Browser Wars” 09 Feb 2008 - https://www.wired.com/2008/09/the-browser-wars/

Mozilla
2018 Open Tech Strategies - https://blog.mozilla.org/wp-content/uploads/2018/05/MZOTS_OS_Archetypes_report_ext_scr.pdf

Michelucci, P, & Dickinson, J. L.,
2016 The power of crowds. SCIENCE.

Paul Festa
1998 Net elated over Netscape giveaway - January 22, 1998 - https://www.cnet.com/news/net-elated-over-netscape-giveaway/

Oshri, I., de Vries, H. J., & de Vries, H.
2010 The rise of Firefox in the web browser industry: The role of open source in setting standards. Business History, 52(5), 834-856

Wang, T. et al
2005 The Revival of Mozilla in the Browser War Against Internet Explorer. - https://www.researchgate.net/publication/221550584_The_revival_of_Mozilla_in_the_browser_war_against_Internet_Explorer

Swift, M.
2011 Browser wars: Mozilla’s Firefox challenges Microsoft’s IE, Google’s Chrome. MecuryNews. 2011, Mar 12

Briody, E., Meerwarth Pester, T., & Trotter, R
2012 A story's impact on organizational-culture change. Journal of Organizational Change Management, 25(1), 67-87.

Gino, F.
2015 The Unexpected Influence of Stories Told at Work. - 2015, Sep 15 HBR

Operationalizing Ethnographic Research to Grow Trust in Digital Financial Services

SOREN HEITMANN
International Finance Corporation

SINJA BURI
International Finance Corporation

GISELA DAVICO
International Finance Corporation

FABIAN REITZUG
International Finance Corporation

Trust motivates people's uptake and use of digital financial services (DFS). Understanding the socio-cultural determinants of DFS trust are needed to scale financial access and drive financial inclusion. These are core components of international development strategies, such as the Sustainable Development Goals (SDGs) or Universal Financial Access (UFA2020). The IFC-Mastercard Foundation Partnership for Financial Inclusion (the Partnership) conducted ethnographic research to understand factors that impact people's attitudes and perceptions of DFS. Nine months of field work each in Cameroon, DRC, Senegal and Zambia were conducted, in collaboration with local research institutes' Anthropology departments and the African Studies Center at the University of Leiden. The results of the ethnographic research produced a framework for understanding drivers and barriers to people growing trust in digital financial services.

This paper analyzes the quantitative and qualitative application of the framework for understanding digital financial services perceptions and uptake propensities for a specific market segment: micro, small to medium-sized enterprises (MSMEs). The framework guided the design of focus groups and survey instruments, whose data provide a classification of issues business owners face for adopting DFS and the degree to which the socio-cultural determinants of these issues may drive or hinder DFS adoption. Further, the survey data provide ground-truth observations that are applied to transactional big data sets toward the development of a model to identify business from individual customers. Identifying MSMEs enables providers to deliver business-specific DFS services at scale and further enable financial inclusion.

It was found that MSMEs could be successfully identified using the big data analytics model, and that the classification of small-medium (and micro) segments could be further identified by application of the survey data. Using the framework throughout all stages of the analysis further demonstrates the business value that can be obtained by employing the ethnographic approach. Doing so ensured that salient DFS issues relevant to business owners were inherent in survey questions ex-ante, thereby yielding meaningful results that can be used to engage MSMEs, design relevant products and advance financial inclusion.

INTRODUCTION

The IFC-Mastercard Foundation Partnership for Financial Inclusion is a $37.4 million joint initiative of the International Finance Corporation (IFC) and the Mastercard Foundation to advance Digital Financial Services (DFS) in Sub-Saharan Africa. The seven-year program was launched in January 2012, and works with microfinance institutions, banks, and mobile network operators to develop and test innovative business models for financial inclusion. Financial inclusion supports individuals and businesses to have access to useful and affordable financial products and services that meet their needs—transactions, payments, savings, credit and insurance—delivered in a responsible and sustainable way. This specifically targets underserved market segments, reaching individuals who disproportionately lack access to formal financial services, such as women, rural households and poorer income demographics. To-date, the Partnership with 14 African financial services providers has achieved an increase of 11.4 million new digital financial services users on the continent, 45,000 new banking agents, and $300 million in monthly transactions ("Digital Access: The Future of Financial Inclusion in Africa", IFC 2017).

Digital financial services have expanded rapidly across Sub-Saharan Africa over the past decade, allowing millions of people who previously did not enjoy access to formal financial services the ability to make payments, save, borrow, and sometimes become insured with a few clicks on a mobile phone or a finger's swipe at an agent's point-of-sale device. However, new technologies always embody changes in a society's economic relations; the way a society embraces new digital technologies and services, or resists them, depends on socio- and cultural factors that are specific to that society. These factors influence the consumer's trust in DFS and ultimately their overall use and level of activity.

To gain a better understanding of African DFS users and capture a sense of the deeper fabric of the emerging markets, the Partnership commissioned "An Ethnographic Study Into the Perceptions and Attitudes Towards DFS in Cameroon, the Democratic Republic of Congo (DRC), Senegal, and Zambia." (De Bruijn and Butter, 2017). The four countries were chosen to encompass both anglophone and francophone markets at the early stages of DFS development. Beginning in September 2015, a research team at the African Studies Center Leiden, the University of Leiden, gathered data over the course of one year to gain a better understanding of the African DFS user and capture a sense of the more profound fabric of the emerging market. One of the central questions in this study was whether DFS will be able to allow for increased financial inclusion of people in global economies. To assess this, the researchers used ethnographic research methods to examine the socio-economic, cultural, and political factors that contribute to the uptake of DFS.

An Ethnographic Study of People's Perceptions and Attitudes Toward DFS

"The Ethnographic Study into the Perceptions and Attitudes Towards DFS in Cameroon, DRC, Senegal, and Zambia" aimed to capture and document these socio-cultural relationships, framed by a central question to understand what digital financial inclusion means in different African contexts in relation to historical, cultural and social factors. Researchers sought to examine the mobility of digital money through mobile phones in relation to societal power dynamics, existing modes of access to wealth, and social relationships; and to contextualize this understanding in terms of traditional forms of money and cultural perspectives on the mobility of people, goods, services and transfers of value. The research teams spoke to both users and non-users to learn what cultural or socio-economic drivers and barriers to uptake there may be, and also how the introduction of DFS may alter or influence social relations and cultural norms. Hence, the researchers followed various actors in their use of a variety of services.

While DFS providers often employ both quantitative and qualitative market research to better understand the customer and the opportunities and challenges of the market, such research tends to focus primarily on existing demand, pricing, and product design. In contrast, using an ethnographic approach to understand what digital financial inclusion means in different African contexts brings nuance to understand how sociocultural factors influence what it means to be financially included or excluded in peoples' own words. A primary goal of the study was to amplify the voices of DFS users and thereby provide insights into why people are motivated to use DFS; and, equally, why people may not seek to use such services. The study contextualizes these motivations through a cultural lens to articulate specific socio-economic and political contexts in which mobile money plays an essential role in the decisions that consumers make.

These motivations are in turn distilled into a framework that categorizes the salient socio-cultural parameters that serve as drivers or barriers for individuals to use digital financial services. It particularly aims to distill the voices of individuals who lack access to formal financial services and who may be most served by greater financial inclusion. It provides a structured approach to understanding customer market segments by articulating a framework of enquiry whose answers can help inform marketing strategies, product development, or the placement and roles of agents by better understanding the needs and perceptions of the individuals these products are intended to serve.

To synthesize the study's findings in a way that translates easily to practitioners, the Partnership for Financial Inclusion published "A Sense of Inclusion: An Ethnographic Study of the Perceptions and Attitudes to Digital Financial Services in Sub-Saharan Africa." (Heitmann et al, 2017). The goal of this report was to focus on the core framework that was identified by the original study to support digital financial service providers and implementing partners with operationally-focused insights on how these factors may influence the decisions consumers make.

As IFC is working to scale digital financial services in Sub-Saharan Africa through its advisory services operational projects, this presents a unique opportunity to incorporate the results of the ethnographic study directly into an operational context. This case showcases the experience of operationalizing ethnographic findings in IFC's advisory services project with a financial service provider to expand access to digital financial services in an emerging Sub-Saharan African market.

Background and Summary of the Ethnographic Research

"The Ethnographic Study into the Perceptions and Attitudes Towards DFS in Cameroon, DRC, Senegal, and Zambia" aimed to broaden the scope of DFS research in a way that accounted for the complex ecosystems in which customers who use these services live. Using qualitative data to assess the decisions consumers make by truly amplifying their voices, researchers attempted to gain a more nuanced understanding of the formal and informal economies in which DFS operate. Painting a bigger picture of these economic ecosystems through the words of DFS users and non-users is extremely valuable when it comes to comprehending the status of financial inclusion in the four African countries evaluated in the study. Drawing connections between DFS and access to wealth in the research sites, researchers aimed to extract themes surrounding the various factors that influence the uptake of DFS and the ability of these services to increase financial inclusion. Furthermore, ethnography helped discover the regulations, both formal and informal, power differences (gender, age, social/political hierarchies), access and appropriation of technologies surrounding DFS users and non-users. The following section summarizes the study's approach and includes excerpts of the research methodology and findings.

Within the overall research methodology of ethnography, the research team employed a variety of methods throughout different phases of the study: interviews, focus-group discussions, observation, participation, life histories, visuals, archival analysis, keeping a logbook and diary, emic discourses, and thick description. About 430 interviews and 13 focus-group discussions were conducted across the four countries analyzed.

A TOOL TO UNDERSTAND TRUST TOWARDS DFS: THE SIX-FACTOR FRAMEWORK

The primary research takeaway is a framework that provides crucial insight into users and non-users' usage, perceptions, and attitudes towards mobile money in four study countries. This socio-cultural framework articulates critical factors that may drive or inhibit use and trust towards digital financial services. In order to understand the society and ecosystem in which DFS is introduced, it is important to look at the historical roots of monetary transactions as well as the current mobility of people and money, and peoples' motivations for moving money. This provides a context with which to interpret peoples' perceptions of digital financial services, particularly in contrast with the new technologies with which these services are delivered. Further, informants articulated feelings for how they do or do not "belong" to such new technologies, and how economic stratifications within society play a role in this dynamic and perceptions of digital financial services are affected by these factors.

There are therefore six themes included in the framework: historical roots of monetary transactions; the mobility of people and money; technological appropriation; perceptions of regulation and consumer protection; networks of belonging; and economic hierarchies. Trust emerges as an overarching theme and determinant of peoples' willingness to take-up digital financial services. This feeling of trust is articulated through a holistic understanding of these six themes. An element of application of the study and its framework is communications, for which additional materials are developed. The key framework graphic is shown in Figure 1: infographic icons are arranged on a leafy vine. The vine is intended to represent linkages between these framework topics, while also embodying

Figure 1. Framework Graphic

the organic nature of ethnographic research and the fluidity of socio-cultural elements across different contexts and demographics.

Historical Roots of Monetary Transactions

The research indicated that past experiences with financial institutions and their historical performance influences people's perceptions of the financial sector and financial services, generally. The two major factors that emerged in the four countries were historical perceptions and experiences with financial institutions, as well as historical reliance on social networks. The data showed that an individual's trust in DFS was heavily correlated with the historical confidence, or lack thereof, that people had regarding banks and financial institutions in their country.

Nowadays, there are many options for people to use formal financial services. On the one hand, historical familiarity with remittances, domestic or international, often serves as a driver for DFS adoption. On the other hand, the perceptions based on personal negative experiences or socialized memory of those experiences about money may present barriers that are passed from one generation to the next one; these barriers can inhibit consumers to adopt new digital financial services today, even if such financial risks are no longer present.

The Mobility of People and Money

Cultures of mobility emerge from social phenomena such as sending economic remittances to spreading social networks, creating economic opportunities in a household, and social contact over distance. Domestic remittances were the launch use-case for DFS in Africa, providing a sought-after solution to send money safely, quickly and cheaply. An existing cultural understanding and need for the mobility of money helped drive this adoption. Historically, this culture of mobility often originated in stories of migration and iterant laborers who needed to send back goods and wages.

Social networks have long played an important role in monetary transfers in the research countries. For example, people have trusted intermediary money handlers to send money to support relatives back home in their villages. This reliance in social networks to move money can compound reliance and trust in traditional financial services, as adopting alternatives can mean breaking away from these social networks. This phenomenon is evident in the continued popularity of semi-formal credit unions and savings clubs in some of the research countries.

When this need for money and people to move is recognized by the society as a cultural norm, this works as a driver for use and adoption of DFS and permits the new services to integrate into familiar norms.

Technological Appropriation

The uptake of DFS is strongly linked to experiences with the technology and knowledge of how the technology works, as well as fears of what the new technology may bring. In a broader sense, technological appropriation is not only related to skills and technical literacy, but also how the technology itself may change social customs or habits.

While the spread of DFS in the four study countries has in part been due to the reach of Mobile Network Operators (MNOs, i.e. cellular phone companies), network coverage remains a challenge for customers who wish to use these financial services. When cell network coverage is not always reliable, and agents may not be present, DFS users not only struggle to make transfers, but also sometimes lose money in the digital void (or may attribute "lost" money to technology problems, whereas in reality the issue may be due to not understanding services fee deductions, for example).

In markets where trust in the financial sector is historically low, or where financial literacy and awareness is low, a single bad experience can damage the reputation of providers overall. There were also problems with money being sent unwittingly to the wrong number, especially during MNO promotions. Furthermore, some participants recounted stories of being unable to recover the money they lost due to inadequate information and the inability to access an agent. Additionally, some agents end up running out of float—namely, cash with which to complete transactions—which forces customers to either return to the booth later or find another agent, harming the overall reputation of DFS providers as customers perceive the broader serve to be unreliable.

Another issue that arises with technological appropriation is a lack of access to information, something which is particularly important for first-time adopters. Increased promotions and more effective television, radio, and print advertisements are necessary to motivate consumers and help the population understand how DFS work. However, informants overwhelmingly expressed the need for more adverts providing step-by-step instructions and for agents to engage more directly with customers until they become comfortable with the use of mobile money technology. These advertisements need to be displayed in local languages and accessible to lower-income consumers; moreover, ineffective advertisement also discourages potential adopters.

Related to messaging, informants articulated how the core value proposition of many mobile money services (the ability to send money quickly and easily right from your phone) can present disincentives for uptake for fear that the technology can chance social norms. Specifically, that the technological chance can deny socially-acceptable excuses to say 'no' to family and friends that ask to borrow a little money by saying sending it is too difficult. In this sense, technological appropriation can present a barrier to uptake, motivated by the concern of how technology can change family and social relationships.

Additionally, there could be negative perceptions based on cultural norms. For example, in communities where group-based financial relationships and face-to-face interaction are prevalent, DFS brings technology that delivers individual empowerment, and this might pose social disruption.

Experience with technology and how it works and its tendency for individual empowerment and social disruption can be either a driver or a barrier to embracing digital financial services.

Perceptions of Risk and Customer Protection

The data indicates that many informants are skeptical toward DFS due to a lack of information about regulatory aspects of the financial system. A lack of information on risk in the financial system in general, and for DFS in particular, has created distrust towards DFS in some markets. This is especially true for unclear information regarding consumer protection.

These sentiments can diffuse throughout social networks and create barriers, even in cases that are not factually based. Even when policies have been introduced and rights are known to customers, some informants complained that in reality there are no guarantees that the policies are enforced. Informants in the research countries reported money disappearing, getting "lost in the system".

The spread of information about risk in the financial system can also create opportunities to build trust around new solutions technology brings. In a given a context of distrust on the traditional banking sector, DFS are perceived as a more secure option, moving people to trust and uptake the new technologies and services.

Networks of Belonging

In African markets, semi-formal social financial networks have often played a central role in the absence of formal financial services. These networks support and guide many of the informants in this study, whether the networks are among family members, friends, co-workers, or other social groups. Community is a widely shared value in the four countries surveyed, and DFS have the power to impact and transform peoples' social networks in various ways. Belonging to a financial community, like a semi-formal credit club or savings society, provides social security in a way that commercial financial services providers may not. Digital Financial Services can be seen to challenge existing norms and identities but can also opt to affirm social relationships to drive DFS usage and social trust.

Networks of belonging and the social aspects of money transfers serve as large factors in the decision-making of users and non-users of DFS. Additionally, informants praised the anonymity of DFS, as it allows them to send money while maintaining privacy (that is, not being observed dealing with money in a public setting by being able to conduct transactions independently through a mobile device). DFS may help to circumvent the use of social networks for monetary transfers, hinting at a change in society related to the possibilities of anonymity and individualization offered by DFS.

The researchers found that while many informants would use DFS for payments, they would turn to established structures for savings and credit, including family and friends. In many cases DFS are unable to establish a similar social solidarity that may be found in informal credit unions and money-saving groups that exist in the four countries surveyed.

Economic Hierarchies

While it is clear that some individuals felt excluded from DFS due to their financial status and position within existing societal economic hierarchies, researchers wanted to examine this idea further by looking at the possibility of DFS disrupting these hierarchies at a more systematic level. Compared to traditional banks, Digital Financial Services are often seen as more accessible and affordable alternatives for people who have long been excluded from formal financial services. Historically, banking services were perceived as available only to a small elite (and often this perception was factually sound, such as by having a pensioned job an as eligibility requirement to open a bank account, for example).

Currently, factors such as self-exclusion, lack of resources, and superstition contribute to a feeling among lower-income individuals that they are not affluent or successful enough to use DFS. In some cases, informants tended to equate banks with large amounts of money—and therefore only for people with large amounts of money—giving rise to an oft-reported sentient of 'that's not for people like me'

Data showed that DFS does in fact have the potential to help lower-income individuals move up the economic ladder in several ways. For example, DFS offers a tool for small entrepreneurs to monitor business growth without financial expertise. It can help facilitate business transactions where traditional infrastructure is less developed. DFS do have the potential to bridge the gap between informal and formal banking and improve peoples' livelihoods. DFS can serve as a stepping stone to formal financial services, providing opportunities to entrepreneurs who find themselves starting out with low budgets.

It is important to understand how Digital Financial Services are viewed in relation to the socio-economic status within a given market; that is, if it is generally perceived as accessible or inaccessible by the unbanked. On the one hand, economic hierarchies can work as a boundary for people to adopt DFS; on the other hand, the improved economic status can be aspirational and therefore, DFS could be seen as an enabler to access to a more prosperous and prestigious position in the community.

Trust as an Overarching Theme

One of the main findings was that trust is a crucial and overarching theme. Trust is the overall outcome of the six themes in the analytical framework above, creating a unique fundamental context in each market. Providers who fail to understand and relate to this context may struggle to launch and grow DFS services to financially underserved market segments.

Numerous interviews revealed that a feeling of security is vital to the uptake of DFS in the four countries surveyed. Beginning with the history of banking and government interventions, informants trust of DFS was largely correlated with their perceptions of finance industries and the history of these industries in their country. It was clear that these historical perceptions have the ability to transcend the past and affect current-day decision-making among users.

Trust was also intimately tied to the social aspects of money transferring among communities. For example, informal credit unions provide much more than financial services, giving participants a community that they can rely on while also giving them the flexibility to save small amounts of money and still gain interest. In areas where these unions remain popular, they serve as major barriers to DFS services, as participants trust them more than they do DFS.

Understanding the safety of mobile-money transactions also requires an element of trust from users of DFS. Especially in areas of low digital literacy, it can be difficult to trust that DFS will safely store or transfer one's money. This is exacerbated by network issues and examples of people who have lost their money by using DFS, often due to the poor network connection in rural areas (which can cause transactions to appear to fail, prompting users to resend and try, try again, only to effect multiple debits unintentionally and create a sense of money lost). When users are unable to recover their money, trust is lost, and these news spreads quickly among social networks—and this in turn inhibits the uptake of DFS by new users. Many non-users have in fact heard numerous negative stories about the services, and they attribute their non-use to these discouraging reports.

Through observing the way people handle and manage money across all four of the selected countries, researchers gained great insight into the feelings of security and trust that are tied to exchanges of money among social networks. The social element also remains vital to an understanding of the mechanisms of trust surrounding mobile money and its possible impact on financial inclusion. To succeed, DFS needs to provide value to users while also working to socialize trust in these types of services.

BRINGING THE FRAMEWORK TO THE PRACTITIONER'S ARENA

The ethnographic framework presented in this study is designed to be used. Each theme provides a category for exploring historical, social and cultural DFS drivers and barriers at the overall market level. As such, it may help providers identify areas of specific concern to increase DFS uptake or activity. This six-factor framework provides an understanding of cultural perceptions in a target market by giving practitioners the categorical questions to ask.

As presented in "A Sense of Inclusion," each of these six framework themes may constitute either a driver or barrier toward trust in digital financial services and therefore willingness for an individual to take-up usage of these services. Importantly, as the underlying ethnographic study reports, the respective socio-cultural norms in each of the four study countries may be characteristically different. Synthesizing these differences according to the six-point framework help to characterize overall trust perceptions toward digital financial services. This is presented in the following matrix diagram.

Table 1. DFS Drivers and Barriers

Theme	Cameroon	DRC	Senegal	Zambia
Historical Roots	History of traditional monetary systems that still remain very popular today	Use of banks has never been wide spread and a series of banks collapsed in the past	Traditional ways of sending and saving money are still very present today	Domination of copper industry led to labor migation and remittances
Mobility of People and Money	History of mobility of money and people due to work migration	Movement over large distances requires secure methods of sending money	Emphasis on international remittances due to emigration to Europe	High levels of labor migration due to copper boom in the past
Technological Appropriation	The DFS technology has not yet reached rural areas, remaining urban	Strong usage of mobile phones, but unreliable infrastructure	Low technological and financial literacy	Ineffective advertisements discouraged potential users
Risk Perceptions	Poor consumer protection; people prefer more reliable means to send and save	Different DFS accounts with a perceived difference in consumer protection	A strong, recognized regulatory environment of banks and DFS providers	Low trust in banks creates trust opportunities for DFS and international MNOs
Networks of Belonging	Strong feelings of social exclusion in the informal sector	Anonymity of DFS transfers	DFS strengthens social ties by enabling fast, daily transfers	Social exclusion of formal financial services due to self-perception
Economic Hierarchies	Existing socio-economic inequalities reinforced by DFS	Informal sector embraces DFS technologies, which reduce economic hierarchies	DFS provide informal sector a stepping stone to formal financial services	DFS orientated toward formal sector, reinforcing existing norms

Barrier Driver

A high-level interpretation of this matrix therefore suggests that an operator implementing digital financial services in Cameroon, for example, might face more socio-cultural barriers to uptake as compared to Senegal, where service providers might find more inherent socio-cultural drivers for a similar service targeting financial inclusion.

"A Sense of Inclusion" illustrates how the framework may be used to characterize and describe the financial inclusion market broadly, in terms of customers willingness to trust the service and value proposition. More specifically, the report encourages providers to apply the framework directly as a method for designing better products and engagement strategies that understand the underlying socio-cultural factors that may either stimulate or inhibit customer uptake of digital financial services.

Potential Uses of the Framework

The framework may help to identify historical, social or cultural pain points that hinder DFS uptake, such as a historical mistrust of the financial sector in the DRC. These may be turned into value propositions instead, similar to how disruptions have driven DFS uptake in some markets.

- Acknowledge and understand drivers and barriers in each market: There are existing historical, social and cultural drivers and barriers to DFS uptake in every market. Providers would be wise to leverage the drivers and to ensure not to impose any further barriers. It is equally important to acknowledge that these factors differ from market to market and that one factor may be a driver in one market but a barrier in another.

- Perform market risk assessments: When examined together, the six themes provide a high-level risk assessment for a given market in terms of which markets or segments face higher or lower barriers for entry with respect to the socio-cultural context.

- Adopt a customer-centric approach to DFS: this framework brings the voice of users and non-users and their perceptions and attitudes towards new technologies and digital services at stake. Therefore, this tool can help to design new products with customer centricity in mind.

- Craft more compelling customer value propositions: Customers interpret the DFS value proposition in terms of their socio-cultural perceptions. This framework can be used to identify what people value. The result will be stronger value propositions for digital financial services and products that reflect and effectively address end-users' concerns and preoccupations.

- Design targeted marketing strategies: This framework is a helpful tool to design marketing strategies more sensitive to cultural complexities and marketing materials that inspire trust and move customers to use DFS.

Implementation of the framework

The IFC-Mastercard Foundation Partnership for Financial Inclusion works with partners and clients to scale digital financial services in Sub-Saharan Africa and has therefore opportunities to directly apply results of the ethnographic study in an operational context.

The following sections look at a first use case of the ethnographic framework: its application and results in another African DFS market. Activities aim to further refine lessons through practical usage in an operational project. This is done with a higher-level goal of using the results of the study to build replicable tools or methods that have been guided by the ethnographic process to support clients and IFC operational projects to better achieve their objectives of expanding reach and deployment of digital financial services to underserved markets. Moreover, the goal is to refine the survey by incorporating the lessons learned from this first case study and then to build standard survey modules to apply them to further research in the region.

In the African DFS market for this case study, key gaps remain that IFC and its Client, a Financial Service Provider (FSP), are tackling to develop DFS products to meet the needs of low-income populations. While individual consumers can access digital financial services, relatively speaking, the business sector of micro-small-medium-sized enterprises (MSMEs) is far less supported in the market. Critically, MSMEs are recognized as a core driver for economic growth and job opportunities, which is why supporting the sector is a focus for many development strategies, including the UN's Sustainable Development Goals (SDGs) (Ke Liu, 2018). In order to target MSMEs better with DFS and to be able to respond to their needs, the project aims to build a predictive model to (1) identify MSME clients based on their financial transaction behavior in order to (2) better design product offering to their respective needs.

The market research and data analytics component of the IFC project attempts to address the challenge of identifying MSMEs among a financial service provider's customer base and to understand their needs, to help engage the market and to provide them services they want to use. As well as to understand the customer segment and their reticence or affinity to use DFS services that a business offers, such as merchant payments. Moreover, understanding this customer sentiment is as applicable to the businesses themselves, as for many of the smallest businesses, the owner and the business are one and the same. The Partnership's Applied Research and Learning team is tackling these challenges by incorporating the six-point ethnographic framework into operational surveys with these businesses, whose results inform will inform a big data analytics project that aims to segment and predict MSMEs in the Provider's customer base.

Research Design

The predictive analytics research study is designed on the hypothesis that within the FSP's network, business owners are using digital financial services to support their business using individual service subscriptions; and that the usage patterns of these individual business users are characteristically different from those of individual consumers. The analytic challenge is therefore firstly to segment MSMEs from the individuals who are generating gigabytes of transactions every month. Then, to create a statistical profile of how businesses versus individuals use these services. Finally, to use this understanding to refine DFS products and engagement strategies to help businesses and customers come together in the digital ecosystem.

A ground-truth understanding of who businesses and individuals are is critical to this approach. Where patterns exist in the data, to understand if these are expected to meaningfully correlate with how these segments characterize their own use cases (or lack thereof). The six-point ethnographic framework was operationalized for this purpose by incorporating questions into structured surveys and focus groups, whose results aim to advise exploratory analysis of telephony and mobile money usage patterns within the transactional datasets on the analytic side. And once results are achieved, to help interpret probability scores generated by these segmentation models and guide operational engagement strategies through a better qualitative understanding of why business owners may (or may not) wish to take-up digital financial services that are specifically designed for business needs.

The method of incorporating the six-point ethnographic framework into ground surveys is presented in the following diagram:

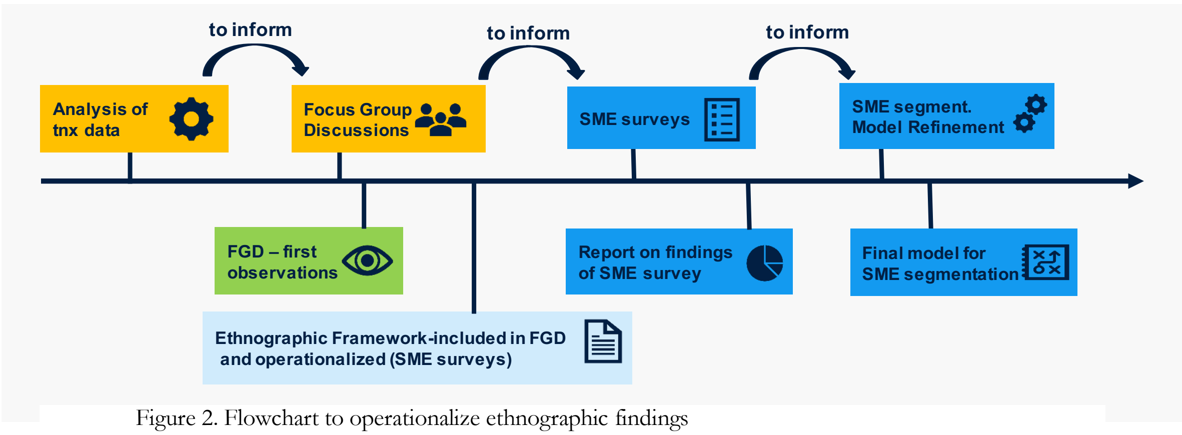


Figure 2. Flowchart to operationalize ethnographic findings

Implementation

In April 2018, researchers facilitated 6 focus group discussions with digital financial services users (enterprises, micro-entrepreneurs and consumers). The focus group discussion (FGD) guide question design was informed by the ethnographic framework. By understanding ex-ante broad socio-cultural drivers (or barriers) to DFS usage, the researchers and focus group facilitators are better able to bring attention to salient factors. Furthermore, the framework is employed again following the focus groups, to aid interpreting participant's feedback and distilling their voices into specific issue questions that can be incorporated into large-scale market survey instruments that in turn provide ground-truth data used for predictive modelling.

Focus Group Discussions

Preliminary results obtained during FGDs shed light on users' perceptions and attitudes towards mobile money in the given country. As in the case of the four countries of the Ethnographic Study, the overarching theme—trust—is very relevant to this country context. Using the approach articulated in "A Sense of Inclusion" a high-level summary of the FGDs are presented as follows:

Table 2. Qualitative Results Summary of Focus Group Discussions

Theme	Historical Roots	Mobility of People and Money	Technological Appropriation	Risk Perceptions	Networks of Belonging	Economic Hierarchies
Results from FGD	Cash payments still predominant	Cross-border DFS trans-actions are perceived as easier and less costly than traditional services or couriers	Advertisements resonate with consumers and have created DFS awareness	Apprehension of the security risks that can affect mobile money technology and platform	Word of mouth or reference from friends, colleague and family as a key driver for uptake	Mobile Money savings are perceived as for the wealthy

Barrier (Historical Roots, Risk Perceptions, Economic Hierarchies) Driver (Mobility of People and Money, Technological Appropriation, Networks of Belonging)

In the case of consumers, in the past, people used to rely on Money Transfer Organizations (MTOs) or courier services to move money. Currently, users prefer mobile money services. They consider the latter more affordable and more accessible to use than traditional forms of transferring money. Users understand there is an inherent risk of transferring money digitally; moreover, it is one of the biggest consumer fears regarding mobile money, pointing to growing dependence on mobile money and the impact of the disruption to normal daily life. Networks of belonging influence people's perceptions on DFS, as word of mouth or reference from friends, colleagues, and family has also played a key role in driving DFS uptake in the studied market.

For MSMEs, the larger the size of the firm, the more frequently firms seem to be willing to adopt DFS. For MSMEs, formal and informal, there is an awareness of the advantages the use of DFS brings vis-a-vis the use of cash. However, only the largest companies appear to have the ability to entirely switch to digital payments using mobile money accounts instead of continuing depending on cash or the traditional bank accounts. For smaller firms, cash is the substitute when DFS are not available. Besides, the larger firms with higher levels of formalization are demanding a higher level of sophistication concerning DFS product offering, for example, loans on the DFS account, to increase the use. Moreover, during FGDs respondents argued that with an adequate digital product offering there would be no incentives to use banks.

In our market of study, MSMEs respondents value the privacy of transactions agents bring compared to a bank, as going to banks also carries a risk of being robbed when leaving the bank with money that is visible to onlookers. For informal MSMEs, respondents from the informal sector believe that digital financial services have created platforms for employment and helped in securing payment of personal bills.

Market Surveys

Building on the insights obtained during the focus group phase, the research team designed surveys to further segment MSMEs and to use these results to advise exploratory analysis and pattern identification in the significant data analytics modeling process.

The field surveys included the operationalization of the six-point ethnographic framework, by building off salient issues raised in the FDGs. In the questionnaire, researchers introduced questions on past experiences with banks and informal savings or loan groups to get a sense of people's perceptions of traditional formal and informal ways of banking. There is a block of questions comparing respondents' attitudes towards DFS vis-à-vis banks and touching issues of economic hierarchies and self-exclusion. Another module measures the role of social networks in the adoption DFS, explicitly asking respondents to recall the first time they transacted with DFS, and who supported them during the first interaction with the new services. In some cases, questions addressed more than one factor, e.g., networks of belonging and technological appropriation, as the following example shows:

Table 3. Example Question for the Factor 'Networks of Belonging'

The first time you transacted with Mobile money services, how did you know how to transact?	1	A relative/friend explained how to use mobile money
	2	Mobile Money agent explained it.
	3	A supplier/customer helped you transact.
	4	Called a customer service number to get instructions.
	5	Read marketing materials/explanatory video/print/ radio advert.
	6	Self-explanatory
	97	Other (Please specify________________________)

In the survey there are questions measuring the perception of risks consumers may have of agents operating in the streets as well as what would be the outcome if hypothetically the money transacted gets "lost in the system", one preoccupation respondents expressed in several opportunities during the FGDs.

The survey sample conceived more than 1200 interviews of DFS users. By design, the sample is exclusively focused on MSMEs and individual entrepreneurs that are using DFS. Out of the 1200 interviews planned, about 200 were sampled through a list of intensive DFS users identified through analyzing transaction data performed in the first step. The rest of the customers were randomly selected in commercial areas using a filter questionnaire to ensure the inclusion of active DFS users that regularly use DFS for their businesses.

First predictive modelling was done using initial results of the qualified FGD results combined with hypotheses of what a would-be business's transactional behavioral pattern might look like (such as higher volumes compared to individuals, or concentrations during normal business hours). The results of the survey then yield ground truth data on whether activity correlates with an actual business or an actual individual. Further, the aggregated results enable refinement by helping to affirm the initial analytic conjectures used to draw expected MSMEs used for the survey, and to help interpret the respective activity patterns in the context of their survey results.

Exploratory Analysis

The first approach to segmenting businesses and individuals is to apply conjectures of what a business behavior might look like to the dataset and see what patterns emerge, and then refine. With a subscriber base exceeding several million accounts and hundreds of millions of individual transactions, the dataset is too big for traditional analysis or to simply look at each individual transaction separately. The approach was to aggregate the data according to some behavioral metrics and then identify statistical thresholds that would indicate abnormal transactional behavior. The underlying hypothesis being that businesses behave characteristically differently, which would anticipate different' behavior patterns compared to users that use DFS only for personal purposes.

Identifying businesses with a survey or typical qualitative approach at this magnitude would not be feasible and ensure the sample was large enough to achieve a model that is statistically significant. Nearly 3 million surveys would need to be conducted to achieve a significance at the 95% confidence interval. Such a sample is not operationally possible for "in person" interviews, not realistically cost effective and not possible within time constraints.

The data centric segmentation approach allowed the project to get information and data insights to review in weeks instead of months, after which the initial segmentation could be tested against intricate in-country knowledge and operation field experience. These insights were then used for selecting a focused sample of the true positive outcomes of the analysis, instead of a blanket approach in selecting the candidates for the focus group discussions. As the focus group discussion guide included talking points around the ethnographic framework, this helped to weave their DFS perceptions into the patterns emerging from the transactional data analysis. It also presented an opportunity to apply the ethnographic framework to a new market, beyond the four study countries of the original field work, to examine its fit in an additional country context in Africa. In this respect, the framework very much held.

Applying the framework as a lens for interpreting quantitative patterns held as well. The focus group discussions proved the above limitations in the analysis immediately, as there were found that some of the analytical positives were in fact not MSME's owners but employees in MSMEs. Something that would never have been identified in the data, nor characteristics that the initial conjecture-based could have ever identified independently. This enabled the team to fine tune the survey sampling frame, to eliminate some of the survey candidates within the first questions, as well as during telephonic interviews accelerating the results obtained from the survey. Finally, with the ethnographic framework articulated within the survey instrument, it will also be possible to quantify how the broader market assessment can be characterized in terms of barriers and drivers to uptake and compare quantitative survey results against qualitative assessments established by the focus group data.

The survey results will be used as ground-truth data to validate and tune the analytic segmentation model for identifying who is an MSME from the multi-million user base. Moreover, interpreting these segments in terms of transactional characteristics and themes of DFS perceptions will add depth to segmented groups and propensity scores generated by the predictive model. At a high level, to compare themes of drivers and barriers to uptake between businesses and individuals (and, equally, small business owners who are themselves individuals who hold perceptions about DFS). Within segments, to look at how these themes may compare across micro-small-medium sized enterprise groups; or equally across individual entrepreneurs where users that may, for example, be more regular active users of services might evidence more trust in DFS compared to latent users. Or perhaps to identify if segments that evidence higher levels of churn or dormancy might emphasize specific elements of the framework.

The results of analytic segmentation models are discrete. In this case: business vs non-business users. More nuanced, the models will also aim to segment the size of the business; however, this classification is also discrete: micro, small, or medium. The surveys give these segments "personality" to help give life to these groups. The ethnographic framework helps to understand how perceptions drive trust in digital financial services. The quantified statistics of several million individuals may be imbued with a voice by interpreting transactional patterns through the lens of the survey instruments that have embedded the ethnographic framework.

Preliminary Survey Results

The full survey started in June 2018 and extended until August 2018. Almost 90% of the interviews took place in the commercial city center of the project market. As of mid-July 2018, about 50% of respondent entrepreneurs had been interviewed. Slightly more than half of the businesses in the sample were formal and registered.

For businesses, the main reasons to start using mobile money were convenient pricing and speed of services. The survey further collected information regarding each of the factors pertaining to the Ethnographic Analytical Framework.

Historical Roots of Monetary Transactions

In the past, people trusted intermediaries to send or receive money. Respondents recall that, before using mobile money as their most preferred way of transferring money, they relied on money transfer services or trusted persons who travelled long distances with cash.

Most of the respondents in the sample are experienced mobile money users, more than half of them have been using mobile money for more than 3 years already. When mobile money services are used cash transfers via trusted person become less frequent. However, as anticipated during the FGDs, the latter is still a prevalent practice, and this reliance on social networks can work as a barrier to grow trust in DFS. Familiarity with remittances can help customers to trust new mobile money services.

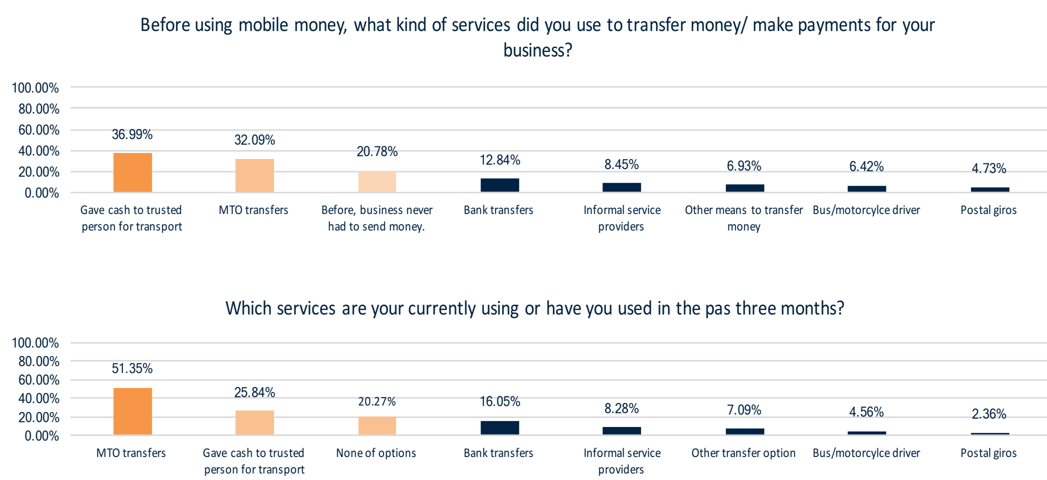


Figure 3.

Mobility of People and Money

Individual trust in DFS seems correlated with the historical confidence or lack thereof, that they have in banks and financial institutions in the country. In the project's study country, the level of trust in the ecosystem seems to be high. Hence, the extent of trust respondents have in FSPs is high with about three quarters of respondents that trust or even highly trust Banks and MFIs and even more than 80% of interviewed entrepreneurs trusting or highly trusting MNOs in the country. Trust in the ecosystem underpins trust in new services and works as a driver for their adoption.

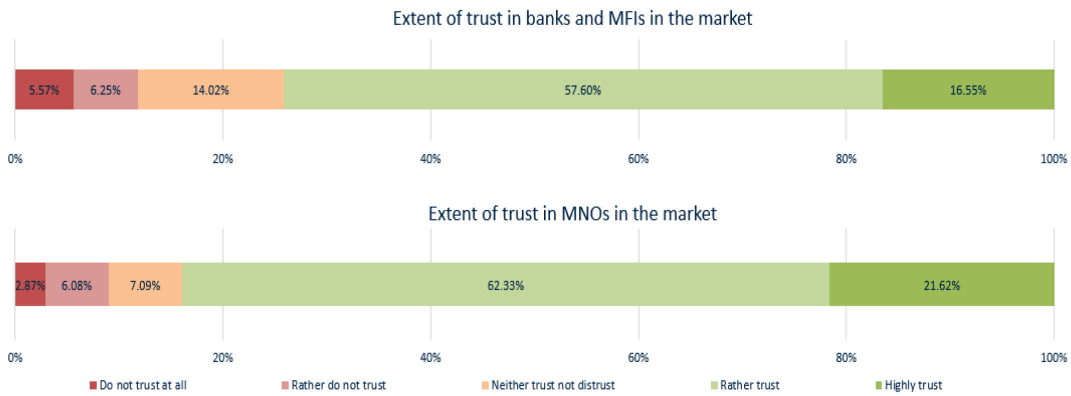
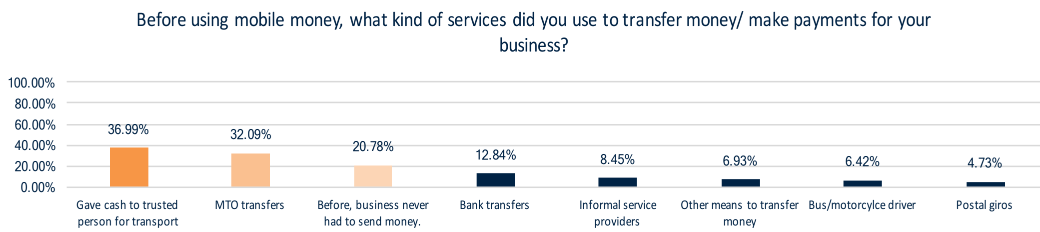


Figure 4.

Convenience matters. The most preferred way of transferring money is using MTO services, such as Western Union, showing a shift in users' preferences when transferring money towards using electronic means such as electronic transfers. This way of transferring money is perceived as more convenient concerning pricing and speed, compared to traditional ways of sending or receiving money.

Local networks and connections with friends, family and neighbors explain most transfers. Among interviewed individuals, domestic transfers are most prevalent when sending and receiving money via mobile money services. When transferring money this way, interviewed entrepreneurs send money above all to family members as well as to friends. Those sending money abroad via mobile money services account for less than 10% of respondents. Recipients of mobile money transfers received money as well in a large majority of the cases from family members and to a smaller extent from friends. Only 13% of the respondents received mobile money transfers from abroad.

As people recognize the need for money and people to move, new digital services can integrate into familiar norms, and work as a driver for the use and adoption of mobile money. As already expressed during the FGDs, those who used mobile money services to make cross-border transactions, already perceived the advantages of mobile money over traditional remittances or courier services.



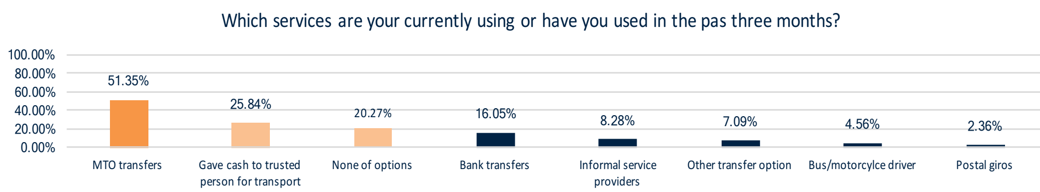


Figure 5.

Technological Appropriation

It is worth to clarify that the sample was, by design, focused on entrepreneurs. More than half of them work in registered companies and in the formal sector. Therefore, respondents may have a higher level of education and technological savviness than the average DFS user in the country. Almost 80% of interviewed entrepreneurs use the Internet. 95% of them access it at home or at work (in more than 70% of the cases), and almost all of them on a smart phone that they own.

The uptake of DFS is strongly linked to experiences with technology and knowledge how to use it, as well as the fears on how the new technology will develop in the future and how it may change social norms and habits.

The experience with new technology, how it works, and the fears of changes brought about by technological advances, can work as either a driver or a barrier to embracing digital financial services. During the survey, about two fifth of users explained that in the past month they had not experienced any issues when using mobile money.

The top 3 challenges experienced with the use of mobile money services were poor geographic coverage; poor customer services, and; failing transactions. Hence, poor infrastructure (poor geographic coverage; power shutdowns at agents) or unreliable technology (failed transactions; missing SMS/ receipts) are affecting users' experiences. Other difficulties that respondents are experiencing are the lack of float or liquidity at agents which result in poor customer experience and reputational risks for providers of digital financial services.

The self-perception on the level of understanding of mobile money services seems to be high among respondents, as two thirds of them Disagreed or Strongly Disagreed that "Mobile money services are difficult to understand." Moreover, almost three quarters of

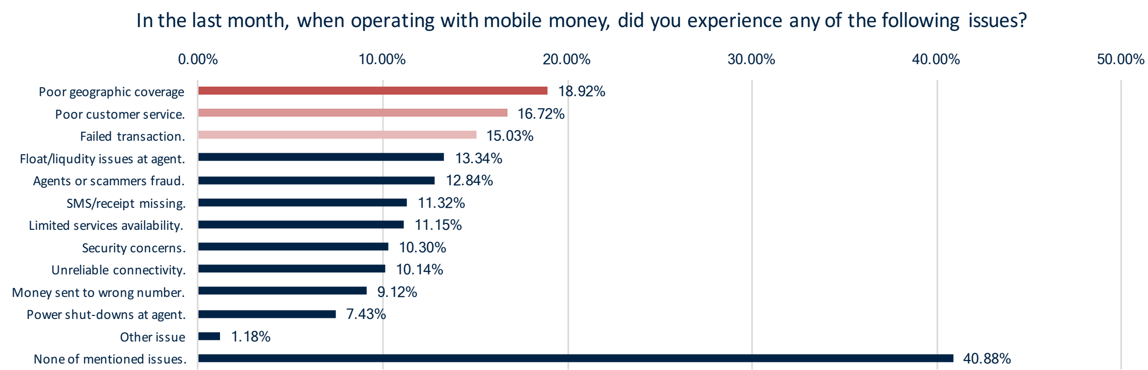


Figure 6.

Figure 7.

entrepreneurs Agreed or Strongly Agreed that "mobile money adverts/ marketing materials clearly explain how to activate your account." This perception was already anticipated during the FGDs and differs from what has been perceived by users in other countries of the ethnographic study, where adverts were not perceived as clear enough by interviewed users.

Technological appropriation does not seem to be a barrier for interviewed entrepreneurs to adopt and use mobile money. Mobile money marketing material is mostly perceived as clear and understandable giving instructions on how to set up mobile money; how to do transactions, and; how to react when transactions fail.

Perceptions of Risk and Customer Protection

The data indicates that only about a third of respondents believe that "mobile money services are not safe" (Agree/Strongly Agree with the statement). Further, 22% of respondents state that they or one of their family members lost money while saving. Surprisingly, the largest proportion (one third) of affected persons lost money with a mobile money provider. About 20% lost savings with a savings group or when saving with a friend. Despite the high percentage who claimed having lost savings with mobile money providers, this fact does not seem to affect respondents level of distrust towards these services. Only slightly more than 10% of respondents in the sample have respectively experienced fraud or scams within the last month or have security concerns when using agents.

Not only trust, but also customer expectations are high. Data collected during the survey on risks perceptions align with the apprehension of the security risks that can affect mobile money technology and platforms to be perceived as a barrier for their adoption and usage. Compared to other countries, users' skepticism towards DFS is lower. If money gets lost in the system, more than half of respondents believe the issue will be solved in less than 2 days. This trust in the system could also be related to the fact adverts are perceived as clear and as providing enough information to successfully use mobile money.

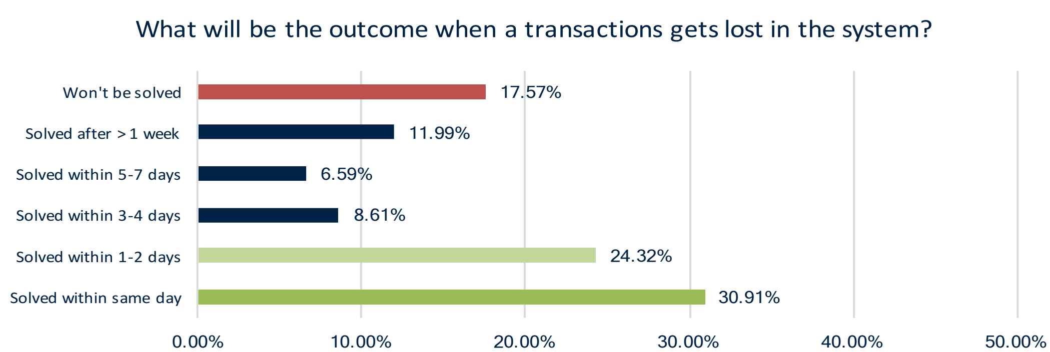


Figure 8.

Networks of Belonging

Digital Financial Services can challenge existing social norms and identities, especially regarding the security and sense of belonging provided by semiformal credit clubs or savings

groups. In that respect, slightly more than half of informants believe that using mobile money agents means less privacy.
While agents remain the most important contact when interacting for the first time with mobile money services, about a third of respondents learned about how to transact for the first time from friends or family members. During the focus groups, interviewees stated word

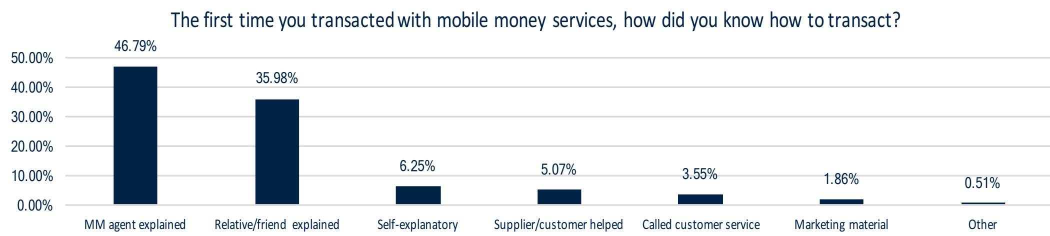


Figure 9.

of mouth references or reference from friends, colleague and family as a key driver for uptake. This is relevant, firstly, because users seem to value human interaction and identity groups when trying to understand and adopt new technologies, and; secondly, because networks of belonging are extremely relevant factor in the decision-making of users of mobile money.

While trust in networks of belonging and word of mouth are pivotal for uptake of new financial services, perceptions of trust towards established financial institutions matter as well. It was found that while many interviewed entrepreneurs would use mobile money for transfers, they would turn to established structures for savings, such as banks. 58% of the respondents stated that the safest place to save is a bank, whereas 34% of interviewed individuals believe the safest place to keep money is a mobile money account. The least safe places to save money are saving with friends and with family members.

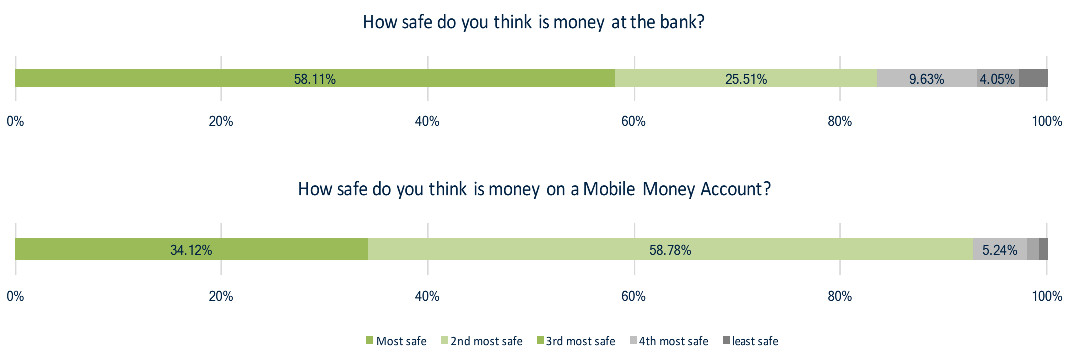


Figure 10.

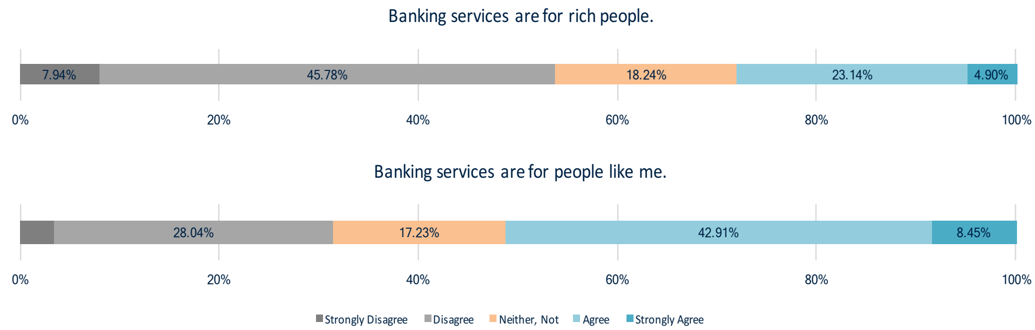


Figure 11.

Economic Hierarchies

DFS have the potential to provide users a more accessible and affordable option to be financially included. Survey results show that over 54% of entrepreneurs disagreed or strongly disagreed with the idea that banking services are for rich people. Indeed, they perceived that banking services are for people like them. More than half of them agreed or strongly agreed with this statement. In this case, respondents exclude themselves through their perceptions, but they don't 'feel' self-excluded from formal financial system as in other countries of the study. This might be due to the fact that the project surveyed entrepreneurs only, which may also have higher income levels than the average mobile money user.

Findings do not suggest that mobile money is disrupting economic hierarchies at a systemic level. Trust in mobile money services is confirmed and they position as an alternative to traditional banks. More than half of respondents agreed or strongly agreed that mobile money services are meant for business like theirs. Mobile money is perceived as a non-discriminative option to transfer money. Almost 55% of the respondents characterized them this way (see following chart).

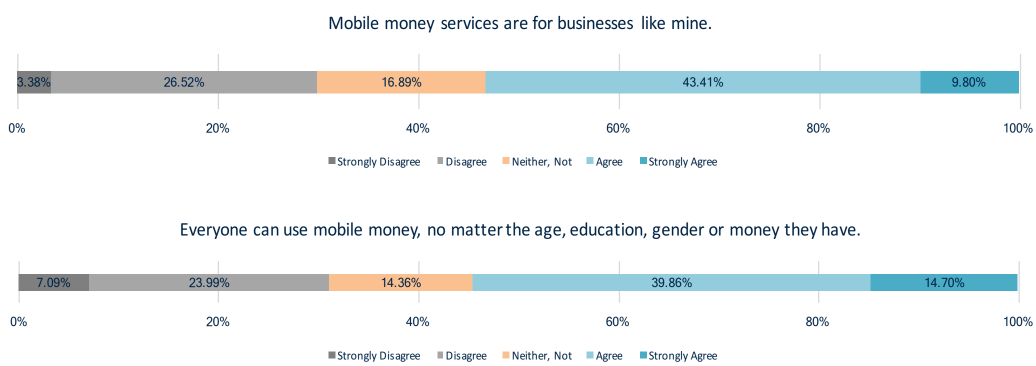


Figure 12.

Quantifying Socio-Cultural Drivers and Barriers

The next step in the analysis is to quantify the extent to which the six themes of the framework present a driver or barrier for the survey participants to trust and use DFS in our country of interest. To this effect, the project created indices for every theme, where each factor index draws on the set of constituent survey questions. Six indices are designed, each of them as an aggregate combined measure from several variables defining the respective themes.

For this first attempt to quantify the framework, no weights are assigned to the variables that are combined to indices, neither are indices weighed differently in the overall framework. We stress this point, as one of the immediate take-aways of this approach and in comparing previous qualitative synthesis is that individual survey questions could belie "big" or "small" motivators for individuals that should be weighted accordingly when developing the index. This observation also amplifies the ongoing need for some qualitative research (such as focus groups) since it's critical to understand if a target market segment might find issues like network reliability, for example, as a 'really big' barrier to uptake versus a minor nuisance that a prospective user just needs to live with if they take-up the service.

All indices are normalized so that their values all range between 0 and 1. This makes the indices comparable across the factors of the framework and facilitates interpretation.

- Index values closer to 1 indicate that the respective factor constitutes a **driver** of trust in and use of DFS;
- whereas lower values, closer to 0, indicate that the factor is a **barrier** that can prevent growth of trust in DFS in the given context.
- When values are higher than 0.4 but lower 0.6 we consider them as **neutral**, meaning that the factor in question is neither a driver nor a barrier.

These gradients help to highlight how factors are positively or negatively influencing the trust and use of DFS for the sample of interviewed respondents, despite the absence of weights.

Table 4 shows the index values we calculate based on survey data and compares them to the findings across factors that came out of the focus group discussions.

For the main drivers, both FGDs and survey results point out in the same direction: mobility of people and money. In this case, the quantitative analysis provides a better understanding of the importance of this factor, which is the strongest driver (score = 0.89) for the uptake of digital financial services such as mobile money.

Albeit to a lesser extent, with a resulting score of 0.68, technological appropriation also qualifies as a driver towards trusting mobile money. During the FGDs participants highlighted their perceptions on advertisements, valuing them as clear and successful in creating awareness on mobile money. The survey asked more detailed questions regarding ownership of devices (feature phones, computers, smart phones, etc.) as well as understanding of how devices and new technologies work. This distinction was acknowledged when building the index. The finding is also interesting when compared to the original ethnographic study, where the technological appropriation factor tended to be a strong barrier across most segments. Whereas in this case, targeting businesses owners, it is clear that they are more technologically savvy; delivering technology-based services is a driver for uptake – as seen through both the qualitative and quantitative lens, and therefore different than what might be expected when engaging more general consumer market segments.

Table 4. Quantitative Results vis-à-vis Qualitative Results from FGDs

	Mobility of People and Money	Technological Appropriation	Historical Roots	Economic Hierarchies	Risk Perceptions	Networks of Belonging
Interpretation survey results	0.89	0.68	0.58	0.58	0.39	0.28
	Driver	Driver	Neutral	Neutral	Barrier	Barrier
Focus group discussions	Cross-border DFS transactions are perceived as easier and less costly than traditional services or couriers	Advertisements resonate with consumers and have created DFS awareness	Cash payments still pre-dominant	Mobile Money savings are perceived as for the wealthy	Apprehension of the security risks that can affect mobile money technology and platform	Word of mouth or reference from friends, colleague and family as a key driver for uptake

Barrier Driver

Framework factors ordered from highest to lowest, driver to barrier

In the case of historical roots of monetary transactions, the effects appear to be more neutral than the focus groups would suggest. Variables capturing the trust in banks, microfinance institutions and mobile networks operators were taken into consideration to create the index. During the focus group discussions, the participants expressed that cash payments were still predominant. When these perceptions are compared with the results obtained after the survey, it seems to be that the prevalence of the cash culture is not as strong as initially inferred during the FGDs. The factor shows a neutral (score=0.58) influence in trust towards new mobile financial technologies. This also presents an interesting comparison given the survey's focus on business owners. Payments are a two-way street: while the businesses themselves might have a more neutral outlook to technology-driven payment services, it must be remembered that it is individual consumers who are coming into shops to buy goods and services. The interplay of broader cultural attitudes toward trust in financial services generally is therefore likely to play a factor (possibly barrier) when considering payments, even if businesses owners may be more neutral (or favorable) on this factor.

A second set of factors initially categorized as barriers are related to economic hierarchies: defined as neither a clear driver nor a clear barrier (neutral) given an index score of 0.58. While focus group discussions convey that mobile money savings are perceived as exclusive to the wealthy, and thus constitute a barrier for uptake, the index would moderate these conclusions - just as in the case of the historical roots.

By contrast, perception of risk and consumer protection is a clearer barrier to trust in mobile money (score=0.39). On one hand, during the focus groups, participants expressed apprehension towards the security risks that can affect mobile money technology and platforms. On the other hand, the survey results show that having lost savings with mobile money providers does not prevent mobile money users from using the services. With that respect, this case shows that it is important to introduce weights in the indices, and to assess the relative relevance that different user experiences have in forming perceptions about risk. The strongest barrier for trusting mobile money (score=0.28) is that of networks of belonging. The corresponding index is a measure of the intensity and diversity in the use of mobile money as a way of understanding the density of networks of belonging and its influence in trust in mobile money. During the focus group discussions, participants pointed out that closest kin's opinions (such as those of relatives, friends or colleagues) act as a strong driver for using mobile money. At the same time, in the survey, it was found that people prefer institutions for saving their money, instead of trusting their networks of belonging (see page 22). The divergence equally raises questions as to whether the survey is adequately capturing cultural issues that pertain to this factor as well. This presents a clear area for more considered refinement of the survey instrument.

This is only a first attempt to quantify the extent to which factors of the framework constitute either drivers or barriers for DFS. The work is subject to ongoing revisions and future iterations and adaptations may still change the exact calculations of indices and therefore also the resulting scores for framework factors. Potential modifications such as dropping or adding variables to the calculation of individual indices or the introduction of weights are part of ongoing discussions.

Nevertheless, these preliminary results and the contrast between results of qualitative and quantitative analysis illustrate the value of combining ethnographic and quantitative approaches to complement each other, allowing to show nuances of peoples' subjectivities. Their general alignment with qualitative analysis is encouraging, and the ability to view factors in terms of gradients highlights value in its own right. That factors can be more clearly deemed 'neutral' is a prominent modification of the analysis, as well as the ability to rank factors from high-to-low, as they're presented in table 4.

Challenges and Lessons Learned From Combining Ethnographic and Empirical Research Approaches

The described research work is the first attempt to apply and operationalize the ethnographic framework for identifying and describing drivers and barriers of mobile financial services in a quantitative survey. Bringing ethnographic and quantitative empirical research approaches together allows new insights and gives a structured way to capture the context of DFS usage and adoption in a country. However, combining these different approaches also requires overcoming challenges.

Convincing stakeholders that are not familiar with ethnographic research about the added value of applying the framework to an analysis can be one challenge. Good communication and showing for example the use cases and operational value of using the framework to a provider can help overcome preconceptions. In this case, the FSP did not question the usefulness of integrating questions that are related to the ethnographic framework into the survey questionnaire. On the opposite, the company was eager to learn more about the differences in the level of trust in financial services offered by different types of financial institutions (MNOs, banks, MFIs, etc.). The project client also provided inputs and suggested the inclusion of additional questions to understand better what digital financial services customers value, how they perceive and use these services and what their considerations are behind behaving in a certain way. The framework was therefore positively received and validated as a way to help test customer appetite for new products.

Another challenge is selecting the right questions for each dimension of the ethnographic framework to capture the essence of that dimension through only a limited number of quantitative questions. Questions should be kept to the necessary minimum to not make the survey and interviews longer than they need to be. Precisely capturing qualitative concepts such as trust and risk perception in a quantitative survey is not easy. In addition, there is no one set of questions that will work for each survey. Finding the right questions for documenting personal and individual information such as perceptions and attitudes towards mobile financial services is highly dependent on the content of a study as well as the country context and the culture of communication.

No one set of questions fits all purposes. Nevertheless, a potential next step for operationalizing the ethnographic framework is to define guidance and lists off potential questions for each dimension of the framework to give researchers in the future a pool of ideas and questions to choose from and to adapt for different contexts in order to capture the drivers and barriers of digital financial services. Such guidance can build on our first experiences and insights from applying the framework by refining and generalizing questions we used for this study, by adding additional questions as well as by exploring more ways to apply them (such as asking respondents if they agree with statements, letting them rank different aspects, directly asking them about concepts or using show cards and images and capturing reactions).

CONCLUSION

We find that it is possible to translate findings of ethnographic research into tangible results that inform deepening financial inclusion and changing mindsets on the use and applicability of ethnographic methods to address business needs. This project contributed to changing mindsets across arenas: in the industry, the DFS provider we collaborated with used it for better segmentation of its customer base; among donors; and internally at IFC. Ultimately, the research has made DFS user and non- user voices more audible, helping to crystalize issues that typical users face in terms of narratives that providers can apply to their service offering.

Applying qualitative research findings to articulate high level market insights around potential customer uptake and usage is challenging. As this report identifies, it is simply not possible to apply methods based on individual engagements, interviews or thick ethnographic description to a potential customer base of millions. Yet, a high level finding of the original ethnographic study demonstrates the value of understanding how complex and nuanced socio-cultural norms deeply drive perceptions, emotions and attitudes about core elements of digital financial services, such as technology, money, and social trust. For providers, scaling these services to reach new market segments and achieve better financial inclusion requires understanding what may drive or inhibit individuals in these targeted market segments to trust the service and entrust their money to it.

This paper identifies the value of using the ethnographic approach to identify these factors of trust drivers and barriers within societies and articulates a six-point framework for probing these issues. It identifies the challenge of extrapolating “thick description” to generalized insights for broader market segments and outlines an approach to meet this challenge by using this framework to design focus group discussions and survey instruments. Furthermore, the framework provides a lens to qualitatively interpret focus group discussions in a coherent and consistent manner (Table 2). Structured surveys provide a quantified understanding of trust factors in the targeted market segment of small business owners (Table 4). Ultimately, this approach provides a consistent and structured way to categorize and aggregate people's perceptions in a way that provides actionable business intelligence.

We also see this in our case study in an emerging DFS market. The operational project's focus on business owners and entrepreneurs provides an opportunity to look at a synthesis of drivers and barriers of that specific group, while also permitting a structured comparison to market segments more generally. As merchant services speak to the interplay between buyer and seller, an ability to structurally compare issues that are salient (whether driver or barrier for one group or the other) to these respective groups and to assess possible alignments or divergences should also help to anticipate issues for developing merchant-specific services.

The approach brings its own set of challenges, such as efficiently selecting the minimum number of key questions to understand the issue, without overwhelming survey respondents. Or, in interpreting survey responses, how to index the magnitude of an issue (i.e., do responds feel very strongly about an issue that is a 'big' driver or barrier versus a 'small' one). Refining survey instruments and structuring interpretation is part of next steps for ongoing research as the research team aims to develop a more formalized tool. The results to-date, as presented herein, show promise, with the ability to meaningfully segment and interpret socio-cultural trust perceptions. Comparing Tables 2 and 4 illustrate how the ethnographic framework can provide a consistent lens for both qualitative and quantitative interpretation. Furthermore, this paper articulates how quantified survey results are also being applied to big data analytics to further marry the lens of the ethnographic framework to other datasets.

A further step, after purely analyzing the survey data along the different dimensions of the ethnographic framework, the following step in the analysis process would be to link this data with mobile money transaction data to better understand how customers' transaction behavior is associated with their perceptions and level of trust in DFS. Linking those data sources allows to explore how individual perceptions about mobile money and the banking system reflects in their use of mobile financial services.

Soren Heitmann is an Operations Officer at the International Finance Corporation (IFC). He leads the Applied Research and Learning (ARL) Program under the Partnership for Financial Inclusion, a joint initiative of IFC and Mastercard Foundation in Sub-Saharan Africa.

Sinja Buri is an Operations Analyst at IFC, based in Dakar, Senegal.

Gisela Davico is a Research Consultant at IFC, based in Washington, DC.

Fabian Reitzug is a Consultant and Carlo Schmid Fellow at IFC, based in Dakar, Senegal.

REFERENCES CITED

Buri, Sinja, Robert Cull, Xavier Giné, Sven Harten, and Soren Heitmann

2018 Banking with Agents: Experimental Evidence from Senegal," *World Bank Policy Research Working Paper*, April 2018.

De Bruijn, Mirjam and Inge Butter

2017 An Ethnographic Study on Mobile Money Attitudes, Perceptions and Usages in Cameroon, Congo DRC, Senegal, and Zambia: Final Ethnographic Narrative Report. African Studies Center Leiden, University of Leiden.

Harten, Sven and Soren Heitmann

2014 Find the Gap: Can Big Data help to increase Digital Financial Services Adoption? Field Note 4. The Partnership for Financial Inclusion. December 2014.

Heitmann, Soren, Anna Koblanck, and Gisela Davico

2017 A Sense of Inclusion, An Ethnographic Study of the Perceptions and Attitudes to Digital Financial Services in Sub-Saharan Africa". The Partnership for Financial Inclusion. November 2017.

IFC

2018 Digital Access: The Future of Financial Inclusion in Sub Saharan Africa. The Partnership for Financial Inclusion. May 2018. Koblanck, Anna, ed.

Ke Liu, Clark

2018 The Role of Micro-Small and Medium Enterprises in Achieving SDGs. Policy Brief, Global Symposium on the Role of Micro-, Small- and Medium- Enterprises (MSMEs) in the achievement of the Sustainable Development Goals (SDGs). United Nations, New York, 2018.

Gaming Evidence: Power, Storytelling and the "Colonial Moment" in a Chicago Systems Change Project

NATHAN HEINTZ
Systems Change Consultant

In 2016 The Chicago Community Trust ("The Trust"), a local Chicago foundation, partnered with Roller Strategies ("Roller"), an international professional services firm, to deploy an innovative mixed-methods approach to community-driven social change on the South Side of Chicago. This partnership convened a diverse group of stakeholders representing a microcosm of the social system, and launched a project with the aim of developing resilient livelihoods for youth aged 18-26 in three specific South Side neighborhoods. Roller designed and facilitated a process through which the stakeholder group scoped, launched, piloted and prototyped community-driven initiatives. While innovative and successful by some metrics, the project had its challenges. The convening institutions and their staff were often perceived as "outsiders" and "experts" without intimate local knowledge of the social challenges they were attempting to address. This dynamic played out in complex power maneuvers across groups in the system. The cultural narratives and interests already at play in the system were employed by individuals and groups at all levels to shape the landscape of agency and power in the system, while attempting to retain the methodological and narrative legitimacy of the publicly-facing project. This case study will explore the narratives and power dynamics at play within the system, look into the causes of these dynamics, and explore the impact they had on the effectiveness of the project as a whole.

> In order to collect "accurate data," ethnographers violate the canons of positivist research; we become intimately involved with the people we study.
> (Philippe Bourgois, *In Search of Respect,* p.13)

INTRODUCTION

Convening, Scope, Organizations

Grove3547 was convened as a response to the question, *How can we work together to support young people in Chicago to develop resilient livelihoods?* This focus was developed during the course of the initial research and community outreach that marked the beginning of the project. As a place to start, The Trust was primarily concerned with the issues of systemic racism and gun violence in Chicago, but these issues are difficult to define, deeply contentious across political, racial and economic lines, and not specific enough to be actionable. As the demographic and ethnographic research at the beginning of the project was conducted, the project team began to focus in on a more specific and actionable framing of the challenge: developing resilient livelihoods for youth aged 18-26 from three specific neighborhoods in an area of the South Side called Bronzeville. This new focus reflected language that would appeal to people across the political and economic spectrum, would be possible to measure, and was seen as plausible (Bronzeville was chosen because it was both replete with social issues and home to a rich history and culture of entrepreneurship and social activism). This new focus also looked at the issues of race inequality and gun violence from a systemic perspective: If youth in this age group were able to develop rich, meaningful and sustainable livelihoods, the project team believed, race inequality and gun violence would decrease.

The project, Grove3547 or "The Grove", was conceived as a strategic approach to addressing complex social challenges called a "Social Lab." The Grove was organized as a series of workshops, a "Kickoff" workshop and three "Design Studios". The Kickoff workshop would be an opportunity for the whole project to do a deep dive into the social system through interviews and learning journeys, begin to map the system and its dynamics, and based on the emerging system map, brainstorm potential leverage points and sites of interventions that could have the most impact. Teams formed around the most relevant leverage points, and began developing and prototyping potential interventions. The individual prototyping teams would meet weekly, coordinating activities in the field, and then the whole project team would meet once a month for a "Design Studio", to present and review their work, share learnings with other teams, get feedback from the hosting team and stakeholders from other parts of the system, and to plan the next phase of their work.

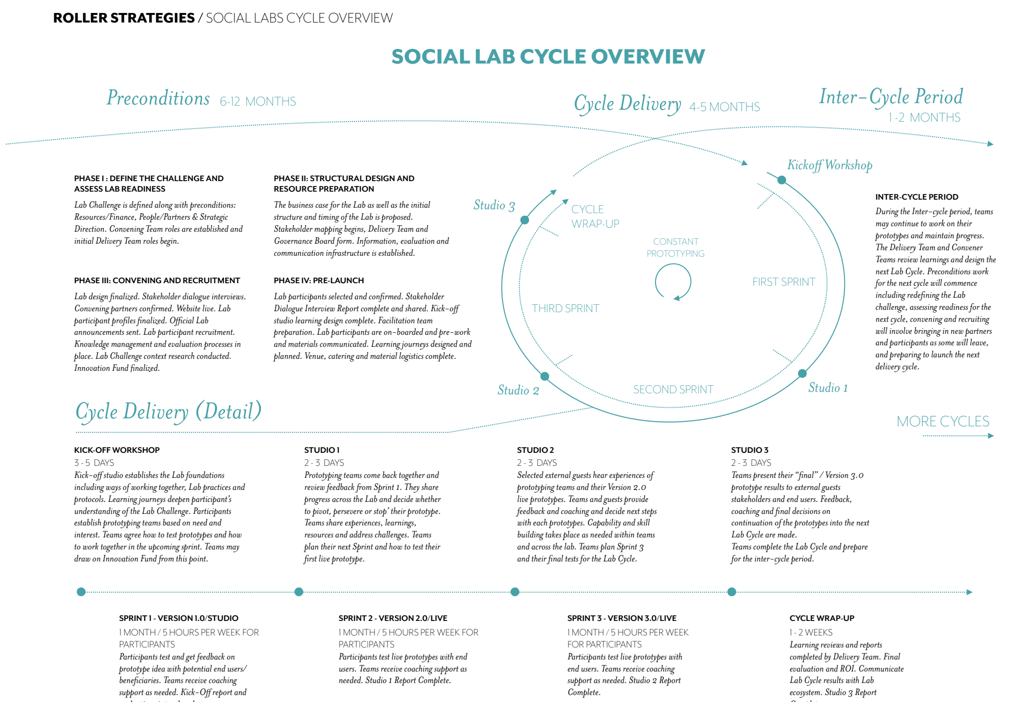


Figure 1.

Social Labs, which draw from a variety of methods, processes and tools, are loosely characterized by Zaid Hassan as being *social*, *experimental*, and *systemic* in nature (Hassan, 13).

The ***social*** aspect of a Social Lab refers to the diversity of the group that is doing the work. In Grove 3547, the convening team worked to ensure that our participant group was *horizontally diverse*, including people from all sectors impacted by the challenge, as well as *vertically diverse*, including people from all levels of the social hierarchy.

The ***experimental*** aspect of a Social Lab refers to the necessity for the project to have a rigorous design culture at all levels marked by trial and error, iteration, and ongoing improvement over time. The prototypes of a Social Lab start small and grow in scope and scale over time, as they can demonstrate effectiveness.

The ***systemic*** aspect of a Social Lab means that the project is trying to impact the *whole system* at the level of *root causes*. The social challenges at the heart of the work are seen as embedded in context and a part of an ecology of intersecting economic, political and cultural forces that make intervention a delicate and difficult matter. Participants are asked to think differently about, and to look deeply into the system they want to change, including conducting their own ethnographic research as part of the project.

The project as a whole was loosely divided into five overlapping, and loosely defined teams: The Roller team, the convening team, the hosting team, the participant team, and the prototyping teams.

The Roller team consisted of those on Roller Strategies' 'core team' who were working on The Grove, including one full-time team member who was "on the ground," living on the South Side.

The convening team consisted of high-level leadership and program officers at The Trust, and high-level leadership and project management level staff at Roller.

The hosting team consisted of most of the people on the convening team, as well as local facilitators, communications professionals and filmmakers, community liaisons, organizers, and support staff.

The participant team was made up of those who were invited to participate in the project through a process of broad and inclusive community outreach. This team included local activists and nonprofit leaders, program managers and innovation professionals from the Trust, local business leaders, small business owners, youth and residents of the South Side at all levels of social strata.

The prototyping teams were five teams that self-organized out of the participant team according to the themes and challenges that emerged during the project's kickoff workshop.

Chicago In Context: The Loop And The South Side

Chicago is a city of almost 3 million people, with 22 miles of coastline along the shores of Lake Michigan. It is a very diverse city with about ⅓ of its population being African American, Latino and Caucasian respectively. However, Chicago is also one of the most segregated cities in the US, the South Side being predominantly African American, the West Side being predominantly Latino, and the North Side being predominantly white.

This segregation is not random but is the consequence of Chicago's long history of settling, relocation, migration, displacement, and housing policy. Chicago history is marked by an effort to intentionally segregate the city along racial lines by powerful white elites in the late 19th and early 20th Century.

Chicago's downtown area is known as "The Loop" because it is surrounded by a loop of elevated rail trains, as well as the Chicago River that geographically mark it as separate from the surrounding neighborhoods or "Community Areas," of which there are 77 in the city.

The Loop is known as the center of power of Chicago. It's where the money and political power are geographically situated, home to Chicago's government, large banks, exchanges and financial institutions, dozens of foundations, a number of universities, and the city's largest commercial districts. "The Loop" is often used synonymously with power and influence in Chicago.

At 225 North Michigan Avenue, in the northeastern section of the Loop, you can find the Chicago Community Trust, a 100 year old Chicago foundation that funds local programs, arts, culture and education across Chicagoland, including the project that this case study explores in detail.

South of the Loop, and far south of The Trust's offices, can be found a large group of about 40 community areas collectively and loosely referred to as the "South Side". The South Side of Chicago is overwhelmingly African American with the exception of Hyde Park and some neighborhoods on the South West Side, which are very diverse.

The South Side is also home to a tragic epidemic of gun violence. While living in Chicago, members of the Roller team witnessed gun violence first hand, including gunshots heard in close proximity daily, frequent arrests, and pervasive police and emergency services presence, especially during the summer months. The team also heard first hand stories of the widespread impact of gun violence and policing, and saw and heard that the youngest residents of the South Side bear the weight of the epidemic.

While this narrative of gun violence is true, it is not the only narrative of the South Side. The South Side varies greatly from neighborhood to neighborhood in terms of safety, affluence, and culture and is home to long-standing communities, generations of families, churches, arts organizations, universities and businesses. Many South Side neighborhoods are marked by thriving local economies. The South Side is also home to a very large number of small, non-profit organizations locally referred to as Community Based Organizations (CBOs). The Roller team encountered a local spirit of activism, community and support that was in many ways the most visible aspect of South Side culture.

Even so, the stark segregation between the North and South Sides in Chicago is striking. To some newcomers to Chicago on the Roller team, the distribution of race and capital in Chicago resembled a form of apartheid. As we will see below, the stark difference between North and South in Chicago plays out at multiple scales.

Chicago's Contested Narrative Landscape: Violence in the Windy City

There are two common and contested Chicagoan cultural narratives that preceded and informed the arrival and acclimatization of Roller's team in Chicago: those regarding violence and politics.

1) Cultural narratives about Chicago focus on violence: *Chicago is a place where people get shot*. Thousands of people every year. The majority of people who are shot in Chicago are young people of color, who live on the South and West Sides.

One person the project team encountered in his mid-twenties from the South Side, noted that many of his closest friends from high school had passed away, victims of gun violence. He went into the military as a means of accessing opportunity and getting out of the South Side. He said he felt it was safer to go to Afghanistan than to stay home, and that surviving into one's mid-twenties in Chicago was a feat to be proud of.

The news stories about gun violence in Chicago pile up daily, weekly and monthly, feeding incomprehensible statistics. These are mostly very short, unemotional news stories that present the simple facts: Who was shot, why they were shot and by whom (if known), where on their body were they shot, where they were when they were shot, if they made it to a particular hospital or not, and whether they were murdered or survived the shooting. These short snippets often come in lists, one news story covering multiple shootings.

The stories about—and impact of—violence in Chicago are everywhere. However there is also a sense among residents that the narratives about the South Side's violence are driven and entrenched by news media rooted in the political and economic power center of the Loop. Political will and economic support from the Loop are seen as limited, with policies often undercutting and underfunding important work that is being done to address the challenge of violence. Meanwhile violence is sensationalized through media coverage, helping to entrench racial bias across the city. This reinforces cultural narratives of the South Side as a desolate and broken community stuck in a cycle of violence. While the South Side is a dangerous place, and parts of it are in dire need, South Side communities are working tirelessly to address the challenges in their neighborhoods. Meanwhile, frustrating and widespread racialized narratives of violence still dominate the larger Chicagoan discourse.

2) Cultural narratives about Chicago focus on place-based politics: *Chicago is a city of ruthless local political culture.* From politics and government, to the nonprofit sector, business, the culture of gangs, and even friend groups, Chicago is known as a place where people are engaged in politics. This political culture is part of the reason that Chicago is nicknamed the "Windy City", referring not only to the weather but also to how much people advocate for their interests.

From one vantage point, this political culture is viewed as a culture of collaboration in which words are rich with meaning, and gestures of shared agency and cooperation can be seen as a kind of currency. However, when miscommunication happens, or when people refuse to collaborate, share agency, or co-create narratives, actors often resort to power and influence, attempting to control which narratives are at play at which particular points of agency in the system.

Before the project began, the Roller leadership was warned by players inside and outside of Chicago to "be careful", and was wished "good luck" navigating the complex political landscape of Chicago, which to some degree shaped their mindset at the outset. The Roller team was careful with its words, entering into a political culture seeking to temper or influence narrative even before arriving in Chicago. Expectations can shape realities. Prophecies can self-fulfill.

However, this narrative of Chicago as a place of intense political culture and contested narratives, is *itself* contested in Chicago. One member of the Roller team experienced this culture of contested narrative as "hidden in plain sight". It's a culture of collaboration and partnership, accompanied by its opposite: competition, secrecy, and political maneuvering. All agree to cooperate, while all play for position. Depending on the context, speaking plainly about power in Chicago can be taboo.

In general, the narratives that are attributed to the South Side—and Chicago as a whole—are contested, reflecting the landscape of contested political and economic power across the city. The North/South dynamic of the Loop and the South Side plays out at all scales along cultural and economic lines.

In this context, to reshape the narratives of Chicago, of North and South, of race, place and belonging—to forge new alliances across boundaries—is to reshape the political landscape of what's possible in the city.

PROCESS: A MIXED METHODS APPROACH TO ADDRESSING COMPLEX SOCIAL CHALLENGES

In carrying out the first phase of the Social Lab, the partnership employed a number of methods including demographic and statistical analysis, network mapping, ethnographic and qualitative research, systems thinking, Agile project management, Lean product development, rapid prototyping, and participatory design.

Demographic and Statistical Analysis

Roller worked with researchers in their network to develop a set of initial data, documenting the intersection of very broad trends across Chicago over time. Data was found and compared regarding income disparity, welfare, race, population, urban mobility, housing prices and other factors. This research, while sourced responsibly, and effective at painting an overall picture, was employed by Roller as evidence of the necessity for the Social Labs approach. The data set was branded *Diverging Chicago: 10 Social Trends Defining the City*, and presented as evidence of a set of urgent and impending crises in Chicago necessitating intervention. Participants in the Grove and some of the courses delivered as part of the project, appeared wary of this branded data. They questioned the sources, and asked where the research had come from and who had conducted it. Some participants dismissed it outright, while others appeared to be very interested in what the data implied.

There was a sense that participants had a healthy skepticism regarding the data. There was a tension between the perceived "actual" data and the perceived "presentation" of the data. Chicagoans were protective of the narratives of their city, defensive against cliches, projections and generalizations, especially coming from outsiders. For example, one participant contested the idea that these social trends "defined the city", as the name implies. Still, participants were interested in data that would deliver insight into how to more effectively address the challenges facing Chicago, and willing to look past a certain amount of rhetoric in the interest of social impact.

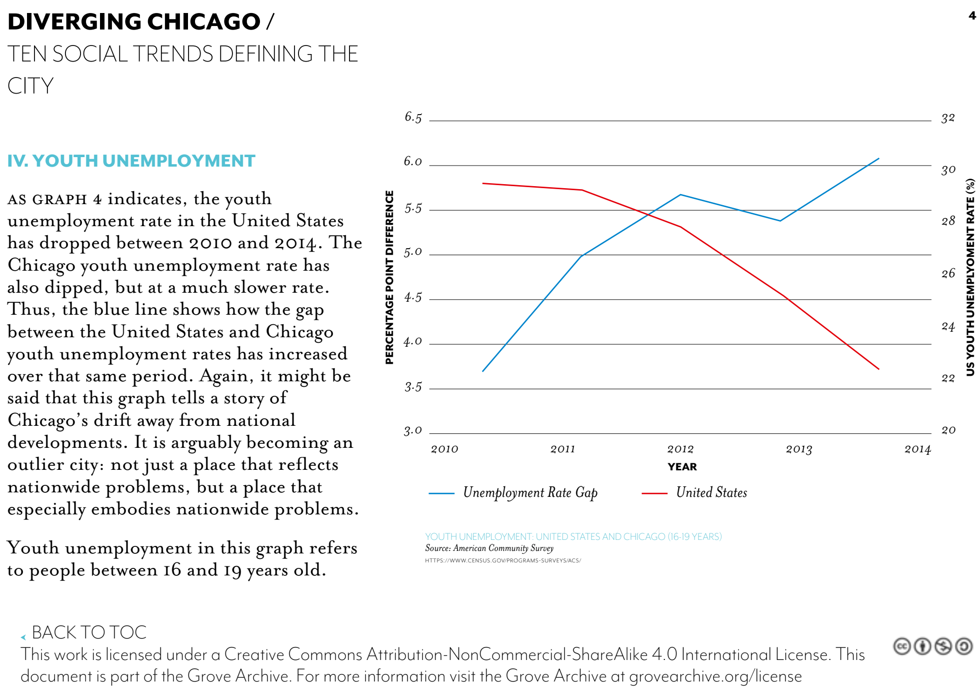


Figure 2.

Network mapping

During the convening phase of the project, Roller and the Trust engaged in an extensive network mapping exercise to analyse the reach and diversity of the Trust's local extended networks. Because the project aimed to have a systemic impact, shifting dynamics in the system across sectors and demographic boundaries, it was necessary to look at the reach and constitution of their network. Using an online tool called Kumu, Roller and the Trust mapped it's networks, starting with grant recipient and partner organizations and other foundations, and then expanding the maps by degrees of separation, and by sector.

This network mapping process was then updated and integrated as the ethnographic interviews added data points to the map. Interviewees were added to the map, and they often offered the project team access and introductions to their own networks, acting as gatekeepers to expertise or demographics missing from the diversity of the project.

The network maps helped inform the project in a number of ways. They allowed the project team to assess where network blind spots were, informing outreach, marketing, and recruitment efforts. They were also able to show where overlapping connections were, enabling more-effective networking, as well as informing the storytelling approach to recruitment. Because of the at-times delicate political culture of Chicago, it was necessary to be careful what was revealed to whom at what time. Recruitment, invitations, and messaging was an important part of the process, and who knows whom would be an important factor informing the convening process.

Multi-stakeholder convening

In order for a Social Lab to represent a truly systemic perspective and have a systemic impact, convening a diverse stakeholder group is crucial. For this reason, one of the core aims of a Social Lab during its initial phase is to convene a *microcosm of the system*. A participant team that is a microcosm of the social system means that all of the different sectors, groups and levels of power and agency in the system are represented.

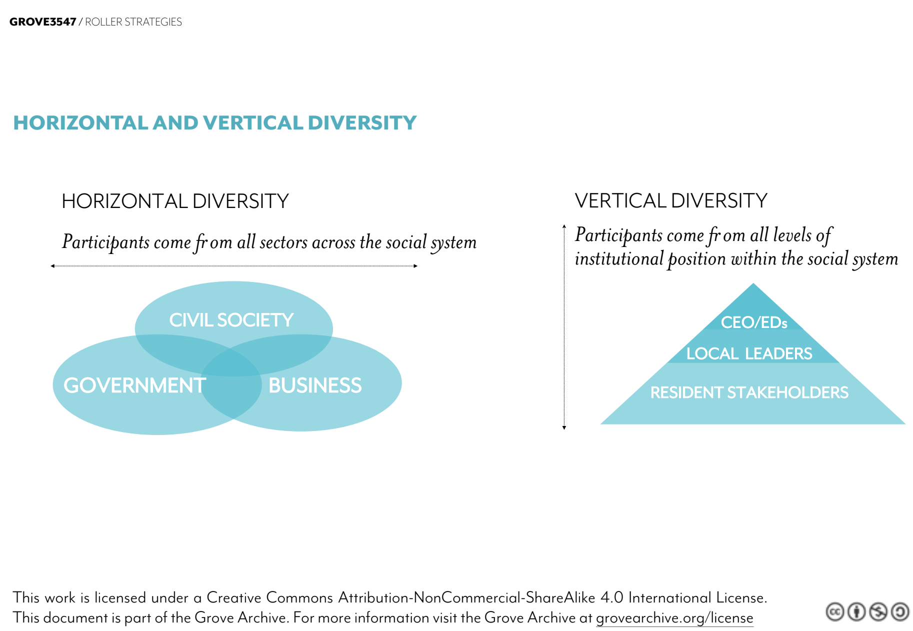


Figure 3.

This ensures that the insights, conversations and outputs of the projects are not able to be co-opted by any single perpective, thereby running into unexpected resistance later on. In other words, by including the full range of perspectives present in the system in a participatory process, the prototypes and outputs are pre-vetted. The system works out some of its challenges and tensions in the microcosm of the Lab so that when initiatives are piloted and scaled in the system as a whole, they have already been subject to a broad range of scrutiny. This process is amplified by inviting guests, allies, advisors, and challengers from an even broader range of actors into the Lab to vet and give feedback on lab team prototypes.

Semi-Structured Dialogue Interviews

Over the course of two months, the project team conducted 42 semi-structured biographical interviews with a horizontally and vertically diverse group of Chicagoans. The interviewee selection process was connected to the network mapping process. Interviewees were chosen first from the Trust and its grantees, then from its wider network. Interviewees from these initial interviews were asked to suggest people whose perspective should be included in determining the course of the project. As the project team continued this process, the pool of interviewees expanded to include a very diverse group of Chicagoans far removed from the Trust and its network.

Each individual interview was structured as follows: After briefly and generally introducing the scope and aims of the project and the Dialogue Interview process, interviewees were invited to tell their life stories, starting at the beginning. The introduction of the project and interview process would generally provide enough context that interviewees would structure their stories to content that was relevant to the project. They would keep their interviews broad enough, however, to include details of their personal and cultural experiences living through and engaging with the context, challenges, and nuances that the Social Lab was seeking to address and navigate.

Going through this process with such a diverse group of interviewees provided a very broad and very detailed set of data to work with. Once they were all complete, the interview teams went through two inductive "interview processing sessions", in which relevant and anonymous quotes were gathered en-masse and then grouped according to themes and issues.

This process yielded a broad and subtle cultural and historical landscape of the issues facing Chicago and its citizens. The findings were distilled into an extensive report, which the Trust ultimately decided not to publish publicly because of the gravity and sensitivity of some of the content.

Participatory Ethnography and Qualitative Research

Ethnographic research is a central element in the Social Lab that is present throughout the process, clumped into activities called "sensing work". In addition to conducting Dialogue Interviews, the project integrated "participatory ethnography" into the process, training Lab participants in semi-structured interviews and participant observation. Participants interviewed each other, pairing with people from different levels and sectors of the social system, to leverage the diversity of the group and expose participants to perspectives different from their own. "Learning Journeys" were also organized, giving participants the opportunity to conduct site visits to various points of interest throughout the system. Participant groups conducted participant observation at a Juvenile Detention Center, City Hall, a non-profit working with youth on civic engagement, an "L" train going all the way from the South Side to the North Side, and visited a local cultural historian in Bronzeville.

By building participatory ethnographic research into the Social Lab, the project team aimed to bridge the divide between researchers and subjects, and offer a deeper level of transparency and agency to the participants in the project. The project team reasoned that those embedded in the social system would have a deeper level of access and legitimacy and would be able to conduct the most effective research and outreach. Rather than "outsiders" doing the research, the project team itself would engage deeply with the system.

Systems Thinking

On the second day of the kickoff workshop, participants went through an extensive systems thinking process. This began with a conversation about the issues facing the South Side of Chicago, followed by a creative brainstorm of possible social interventions and solutions addressing the question: *How can we create resilient livelihoods, in Douglas, Oakland and Grand Boulevard?* (These were the three neighborhoods in Bronzeville, and the question driving the project).

The brainstorm was then distilled into focus areas through a clustering process, yielding a map of the social challenges facing the South Side and a diverse set of potential ideas for solutions. Participants further refined their ideas by voting for the ideas with the most urgency, potential impact, and what they had the most energy to do. This process also yielded ideas for how issues could be approached from a number of different angles at once, and how ideas might overlap and combine to have systemic impact.

This exercise and it's outputs in part determined the course of the project as a whole, as participants formed their prototyping teams around these issues and ideas for intervention.

Agile project management

Agile project management was used to manage and inform the logistics of the project. Agile is a way of working that can support multiple loosely defined, coordinated teams, working in planning cycles or "sprints". Sprints can have varying durations depending on the needs and context of the project. These activities and teams are coordinated so that the strategic thinking, decision making and tactical operations are coordinated, and allow channels of communication and feedback at every level.

Agile is also marked by review and iteration. The teams hold meetings to assess the effectiveness of their work towards project goals, as well as the effectiveness of their Agile planning process, and then to plan their work based on the learnings from their review process, enabling adaptation over time.

Agile as a project management system was generally helpful to Roller. Considering the many moving parts, multiple overlapping teams, and complex sets of relationships in the project, having a system for managing workflows proved invaluable to those inside Roller. However, those on the participant team and in the Trust did not see as much value in the use of Agile. On the contrary, some people found the Agile language of "sprints" and "workstreams" as unneeded added complexity and confusion. Some of this was due to the (aesthetically painful and confusing) way in which the Agile process was represented:

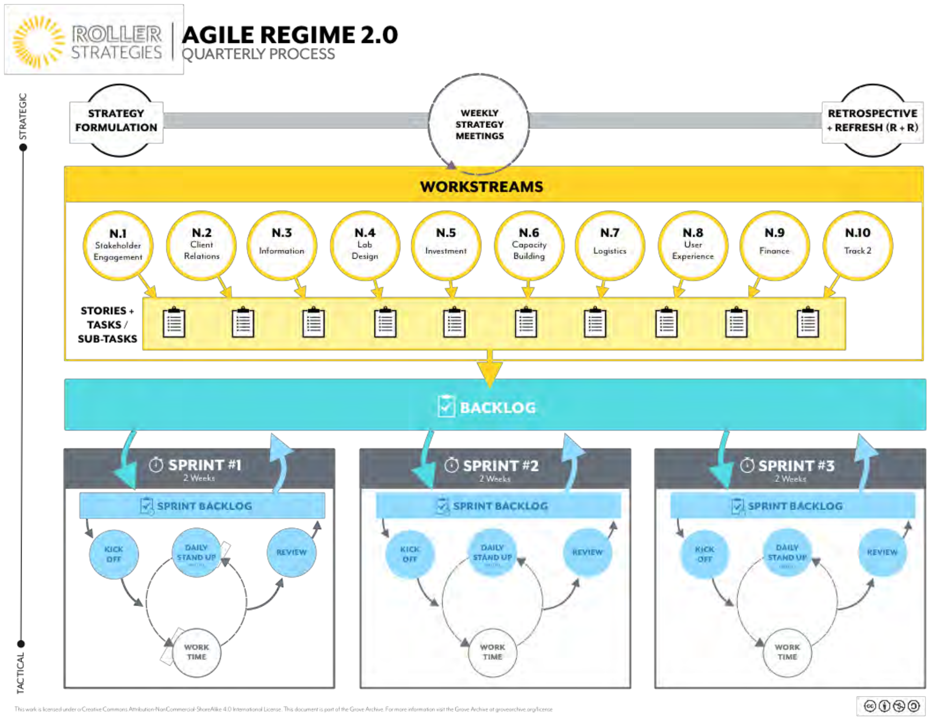


Figure 4.

Rapid prototyping & Participatory design

The project's use of design thinking took shape in a two specific applications. Rapid prototyping was used during the kickoff workshop to generate and refine a large of number of potential ideas for social interventions, and a more ongoing participatory design practice was initiated as the prototypes were launched and engaged further with community.

During the Kickoff workshop, the prototyping teams met and began to explore through model-building what they might actually do together to support the livelihoods of young people in Bronzeville. This model-building process, involved using Legos, pipe-cleaners, playdough and other materials to make models representing the real-world programs or institutions they intended to prototype. Once the models were built, each team had an opportunity to present their model to the rest of the group, answer clarifying questions and receive feedback on their models. They then made changes based on the feedback and developed new iterations of their prototypes.

As the project teams developed their ideas into subsequent workshops, they utilized participatory design principles, taking their ideas out into the community and engaging with the target demographic of their unique project. Depending on the response, some teams took more time to refine their ideas, try to new ones, and do further research with the community, while other teams began to launch small pilot versions of their projects.

Throughout the course of the project, teams utilized participatory design and sought user-input, through ethnographic and qualitative research, community outreach, online ad testing, and principles from design thinking and Lean product development.

Social Labs Framing

The above methods as well as other tools and approaches were conceptually organized according to the Social Lab's top level categories, or "stacks": ***innovation*** (the facilitated processes through which teams learn, prototype and respond creatively to the challenges of the social system), ***governance*** (the organization of the project as a whole and management of power, accountability and transparency), ***information*** (internal and external communications, including knowledge sharing and collaboration infrastructure, and project-wide storytelling), and ***capacity*** (training staff and participants in the theory, methods, tools and skills related to the delivery of a Social Lab).

While capacity building processes took place throughout the events of the lab, one of the core ways that the Social Lab approach was supported in the system as a whole was through the delivery of a series of courses or "masterclasses". The masterclass was a strategic-level course on the theory and practice of Social Labs, offered primarily to local leaders and stakeholders who were close to the Trust's professional networks.

THE "COLONIAL MOMENT" OF SYSTEMS CHANGE: NORTH/SOUTH & INSIDER/OUTSIDER DYNAMICS

> Development anthropologists reinforce ethnocentric and dominating models of development. Moreover, these practitioners disturbingly recycle, in the name of cultural sensitivity and local knowledge, conventional views of modernization, social change, and the Third World. (Escobar, 658)

This section will examine specific points of contact between institutional, identity and cultural groups in the project, and how those points of contact are informed by interest and cultural narrative.

Some of the most poignant challenges that arose during the course of the project came about because of perceived similarity and difference and the formation of groups around both formal and informal lines. Formally, these dynamics played out between the different teams of the project. Informally, groupings and tensions emerged around "localness" and proximity to "the community", socioeconomic status, race, age, institutional affiliation, nationality, and "expertise". At the points of contact between these various groupings and identities, complex dynamics played out as actors and groups vied for power and agency in regards to the project and the broader social system.

These points of contact and the complexities and tensions at their intersections contributed to an "insider/outsider" dynamic that played out at different scales of the project, as well as a tendency to mobilize evidence for the purposes of group interests. This had a strong impact on the trust and strategic alignment of the project as a whole.

"Localness" And Agency

Power in the project was distributed along both economic and cultural lines, with high social capital accompanying the "localness" of a person or institution. To be "born and raised" in Chicago is an emblem of belonging and status in Chicago, and even those who have been there for decades would hesitate to say they were "from Chicago", instead saying they were "relatively new" to the city. On the contrary, to be an "outsider" seems to bring an air of unfamiliarity and even illegitimacy. To be an outsider in Chicago is to be questioned as to why you're there.

Another expression of the power of "localness" was "proximity to community." In this case, "community" generally referred to those actors and institutions that were situated in Bronzeville, and specifically those that were working in the social sector. One of the only points of unanimous agreement among the Convening team was that the project was meant to be "community-driven".

That the institutions funding and setting the overall strategic direction of the project were nowhere near the South Side, was recognized as ironic and problematic by participants, as well as some Roller and Trust staff.

While community agency was the agreed-upon goal of the project, the organizations leading the work vied for power to define and control what that would mean in terms of agency. This played out in disputes and power plays between the Roller team and the Trust in regards to the scope, timeline, composition, access to information and resources, and the process and structure of the project as a whole. So while the whole project team agreed that agency should be situated in "the community," plays for power were often justified with degrees of "localness" (who is "more local" than whom), and/or claims of representing the community's interests. One of the Trust's high-level staff commented:

> The consultants would say "the community knows best", but they didn't believe that the community institutions knew best. They over-rode our decisions every time. If we had had more time, we could have built a local team. We did not put community first.

While the Trust sees itself as a "local" institution and therefore in a legitimate position to diagnose and address the social ills present throughout Chicago; on the South Side the Trust represents the power and influence of the Loop and is seen by some as an "outsider" to local structures and systems. On the other hand, Roller saw itself as holding expertise in grounding collaborative work in the agency of community, judging that they could actually help the Trust do a better job of leading "community driven" work.

Foundation Culture Vs. Community-Driven Social Impact

This section explores the power dynamics and cultural tensions between the foundation funding the work and the local South Side communities and organizations involved in the project.

The Chicago Community Trust is situated on the 22nd floor of a massive skyscraper in the Loop. It's a smooth, black cubic structure, beautiful and almost frightening in size. The building smells like power. 225 N Michigan is guarded by a set of automated gates with a barcode scanner, preventing anyone without an invitation and a photo ID from accessing the elevator. Those who work there on an ongoing basis may get clearance, visit the security offices for a photo, and get a permanent key card giving them access to the building. An infrastructure of security protects the building, seemingly keeping the world out, and also keeping the office workers in, embedded in their world. While the gates would frequently malfunction, giving off a loud, horn-like alarm sound, there was still a clear "inside" and "outside" to the Trust. Even on the 22nd floor, a visitor needs their keycard to enter the offices themselves, or they must resort to waving down a passing office worker through the huge glass wall that keeps out those who don't have access.

The Trust, like many foundations, also has a corporate culture, with a strict dress code (casual Fridays are in place if one cares to wear jeans), and a tight-knit bureaucracy of grant requirements and programming, carried out by a large team working out of cubicles, private offices, and co-working space, depending on one's position in the organization. These cultural and organizational rules and channels essentially act as filters and gatekeepers for those seeking to do business with the Trust. Those that interact with the Trust do so on the Trust's terms.

The Roller team was critical of this, wondering if 18-26 year olds from Bronzeville would be welcome in the Trust's offices as they were, or if instead would they be required to fit in to a culture of bureaucracy, follow a dress code, and show that they meet grant requirements? In other words, the culture of the Trust itself begged the question:

> In the context of power and resources that are centralized in the Loop, and guarded tightly by a securitized and bureaucratic corporate culture that goes so far as to manage dress, how can a diverse, community-driven, participatory project directly and effectively engage 18-25 year old youths on the South Side?

In this way the Trust was experienced as "overly-secure" and "inaccessible to the outside world" by members of the Roller team. This experience was cultural, physical and relational. Although the Roller team was warmly welcomed, and had access to the people and the offices of the Trust, they nonetheless often felt out of place, like visitors in a very particular brand of corporate culture, trying to "do things right" and laughing about how difficult it was. If even the Roller team felt out of place, what about the young people our project would engage with on the South Side? Was this cultural dynamic a Chicago/Outsiders thing, a Loop/South Side thing, a Funders/Recipients thing, or all of the above?

The "North/South" dynamic between funders and recipients proved to be a complex one, in which proximity to the institutional power and resources of the Trust was both a currency for, and a hindrance to locality and community-driven agency.

In one instance, members of the project team were canvassing South Side neighborhoods recruiting individual and organizational participants. They entered a local CBO without calling ahead, described the project and explained that they were looking for participants and collaborators. The team was met with a string of curt and very direct questions, including "Who's funding your project?" ("The Chicago Community Trust."), "Are you looking for funding?" ("No."), "Is the project already funded?" ("Yes."), and "Is there funding available for participants?" ("Yes."). Upon hearing the answers to these questions, and within 15 minutes of having walked in the door, the team was invited into a back room to sit down with the second-highest person in the organization's management, who proceeded to offer a great deal of support and recommend a number of their organization's young leaders for participation in the project. Proximity to the Loop and its resources acted as valuable social currency, allowing the team access to networks and people that would have been very inaccessible otherwise.

While proximity to the Loop and the Trust acted as a kind of currency, there was also an air of mistrust between these different cultural and geographic spaces. At times the project team heard participants remark that local institutions and foundations (including the Trust) would make promises they couldn't keep, and had a reputation for de-funding effective programs. Some participants expressed doubt that the Grove would continue to be funded beyond a few months, anticipating that their work would be cut short. There was both great enthusiasm, as well as a general air of skepticism among participants from the beginning.

The problematic dynamic and narrative of the Loop as a source of power and funding and the South Side as a recipient of funding and programming, is also played out in the relationship between Foundations and CBOs. Leaders of CBOs that were part of the Grove were both frustrated by the bureaucratic requirements of grant-giving organizations as well as driven by the scarcity of funds available. That CBOs were forced to use large portions of their scarce resources on grant-writing and meeting the requirements of the funding organizations (rather than on programming), was an oft-cited point of contention. CBO leaders expressed frustration that they were further scrutinized as to how much of their budget went toward programming rather than management and overhead. There was a two-way squeeze in order: requirements to meet cost-intensive grant requirements, and to prove cost-effectiveness, efficiency, and impact of the CBO's operations. CBOs were in some ways sacrificing their impact in order to meet the "nonprofit bottom line" of funding requirements and impact assessment.

Part of the Social Labs model—aimed at combating the power discrepancy regarding funding—is to make funds available "up-front" so that participants and prototyping teams have access to resources without strings attached. The innovation fund was a pot of money that would be set aside for the prototyping teams to access freely to fund their prototypes. $100,000 total was to be split up and/or spent by the teams as they decided best. It was in their control.

But when Roller requested of the Trust that they transfer the $100,000 to a bank account that would be outside the Trust's control, a key point of dispute became visible: The Trust's internal policies would require a detailed budget, with very specific costs breaking down exactly what the funds were to be spent on, as well as detailed reporting essentially "proving" that the planned budget was working to impact the challenge.

The culture of the foundation was one of tracking, planning, and linear thinking. The point of the Social Lab is to navigate the constantly changing terrain of a complex social challenge with agility, creativity and collaboration, maintaining the freedom to change course rapidly. A process combining ethnography and design would allow project direction to shift immediately with prototype feedback, and would allow project teams to use funds try out new and different ideas that might learn from failure. The whole point was to provide an open space to try creative ideas. The funding had to come first.

This was a pivotal moment for the project. The Trust's leadership were up against their own internal mechanisms and policies and some were skeptical as to the necessity of having funds come first.

The Trust, displaying a profound commitment to the project and a willingness to challenge its own boundaries, convened a special meeting of their leadership and board, and changed their policies to allow for a new type of funding that would allow teams to prototype. This was seen by most everyone involved as a heroic leap in the right direction. This moment was marked by a great deal of discomfort, excitement, and energy on all sides. Some at the Trust, however, remained skeptical of this exceptional decision until mid-way through the first design studio.

A number of the Teams were preparing to set up their weekly meetings and were confirming everyone's contact details. One team member, however, was unable to participate because he didn't have a phone, and another was out of minutes. There was no way that the teams could coordinate without a way to contact all of the team members. Someone suggested that the team simply dip into their budget and buy their team member a phone. This was a very sticky issue. "Shouldn't the funds be spent on the prototype?" "Are we allowed to spend the money this way?" The decision was challenging a few core assumptions, all of which related to power and what was "inside" or "outside" the scope of the project.

Some of the assumptions being challenged were: That the money available still belonged to the Trust and not to the project team; That the social system (and the challenge) was "out there" and so funds should only be spent "out there"; and that the "private" and "public" lives of participants should be kept strictly separate. Ultimately the participant team came to understand that it truly had agency over the direction of the project, and that the challenge was operating at a level of depth close to home. The decision was made to buy the phone.

At this moment, something shifted for a number of participants, including Trust staff. There was a visceral understanding that entrusting the participant teams with resources *without* accounting for every budget line item actually *encouraged* the accountability and collective agency of the entire project by entrusting all of the participants equally with the power to make decisions with fiscal repercussions. The Trust, by giving up its power to control and make final decisions, entrusted the participant group with that responsibility and enabled a culture of shared power and agency. Within an hour, everyone had a phone. The implications of this fact for the project as a whole were staggering. See a need? Meet it. The project realized itself as truly community-driven, and responsive to the immediate situation on the ground. If this was possible at the drop of a hat, what was possible at scale over the course of a years-long lab?

In these ways, the project was experimental at every level. The Trust was experimenting with new funding models, and the participants involved were experimenting with new ways of impacting the social fabric. But the project was not without its divides.

Formal Structural Agency

The practical concentration of agency in the project was in part shaped by the formal structure of the project. As mentioned above, the project was organized into teams. These teams notably and problematically corresponded to different levels and different forms of agency and power.

The convening team, for example, consisted of strategic players at the Trust and in Roller who collaborated to design the research, timeline, facilitation agenda, convening, staffing and composition of the project.

The hosting team, while not formally involved in setting the strategic direction of the project, was involved in designing the agendas for the specific workshops and events that composed the project, as well as actually facilitating the workshops. Being in charge of facilitation also meant being able to change and adapt the course of the workshop and thereby the direction and timeline of the project.

The participant team and the prototyping teams that they formed were not directly involved in the workshop design process or overall strategy, but they did have a great deal of agency in terms of the content of the project and the nature of the work being done in the social system. These teams were the ones "on the ground" and "in the community" doing the actual work of the Social Lab, and they had unilateral control over the content of the project, that is to say, over the design, delivery and strategy of the actual prototypes and projects that were launched and adapted during the project.

The agency of the prototyping teams was not "anything goes" however. There were criteria for the prototypes (designed to ensure quality and impact), that were conceived and agreed upon at the highest level of the project. The Convening Team was basically in charge of the *container* and the *process* - designing and facilitating an effective process that would allow for the launch of scalable community-driven social interventions that would have maximum impact. The participants and the Prototyping Teams were in charge of the *content* of the project's outputs: conceiving, designing, launching and iterating the social interventions in collaboration with community.

While it's not clear whether there was any inequality in the *amount* of power or agency available to these different tiers of the project, the *quality* of that agency was definitely different. The Social Lab acted as a bridge between "the local" or "the community" on the one hand, and the "moneyed institutions" or "the system" on the other. It acted as a tiered bridge between the Loop and the South Side. In this way, the Social Lab's job was essentially to ensure that power was concentrated in "the community" even as resources flowed from the center of power. This was easier said than done.

"Expertise" Vs Local Knowledge

While the process of the overall Lab was determined "from above", the prototyping teams also had agency in regards to process, it was just a specific part of the process: the part that engaged with community. While this agency was real, the convening team had designed the process and the parameters of success independently of the participant team, contributing to an experience by participants in the project that there was "regulation from above" in regards to what was a viable direction to their work.

This regulation from above took the form of codified methods and the presence of outsiders who were presented as experts. For example, at one point during the process Roller invited an outside consultant to teach "Lean" product development as a model for developing and testing initial ideas for prototypes in the field. One participant expressed frustration, because as a professional and entrepreneur, he was quite familiar with Lean, prototyping and design thinking. He expressed that he was totally capable of applying Lean to his work in the community without the need for expensive consultants. He wondered what, if anything, original the "Social Lab" would actually contribute. He wanted to know *what exactly is a Social Lab?*

Unfortunately that question is a bit hard to answer, because a Social Lab is essentially a practice, rather than a method. The "Social Lab" is not intended to be a defined, codified way of doing things. It's supposed to be an approach to strategy that draws on *principles* more than methods. A Social Lab is really any project that's participatory, creative and iterative, led by a horizontally and vertically diverse team, that seeks to have a systemic impact. Whatever tools and processes will be most effective at realizing these principles can be drawn from.

However, Roller as an organization *did* codify the practice of Social Labs in a number of ways. One way is directly by publishing the *Social Labs Fieldbook*, and *The Social Labs Revolution*, as well as by marketing Social Labs as a practice that requires *expertise.* In other words, the practice of Social Labs is designed to be inclusive, participative, and open. But if expertise, codification, and technical knowledge are required in order for a Social Lab to be impactful, this will at times conflict with its inclusivity and participatory intention. It risks becoming a form of "colonial" power, expert knowledge form imposed from "above". There's a conflict between the participatory and open principle of the Lab and the fact that it requires experts to facilitate it at the strategic level. If codification implies that the rules and ways of the Social Lab must be *learned,* then can it really be said to be open and participatory?

This problem also has a subtle cultural dimension. In the words of Arturo Escobar: "Communities [...] have to adopt organizational forms and project designs that the donor can recognize if they are to have access to project funds, even if those forms may not reflect community traditions." (Escobar, 674) In other words, at points of contact between those inside and outside the Loop and its institutions, there is a cultural power move at play in which the "recipient" is required to comply with the cultural and institutional forms of the moneyed institution.

In this case, it comes down to how this challenge is addressed when it surfaces in a specific context. From the beginning of the Kickoff workshop, the participant group was invited to give their input and feedback, to tell the hosting team if something wasn't working. They were assured that it was *their project.*

But it was also the Trust's project, and it was the Trust's project *first.* And the Trust needed the project to have an air of security and legitimacy. The Trust *needs* the work to be grounded in expertise. They need a certain degree of codification because in the formal professional context of the Loop (read "North"), codification denotes expertise and expertise denotes authority and legitimacy.

But in the world of community engagement and nonprofit work on the South Side, codification and expertise can easily be seen as a cover for patronization and control. "The community" already knows how to do its business. On the South Side, it is local knowledge and a proven history of community engagement that speak to legitimacy. In this way, the Social Lab is an intermediary seeking to translate and connect those on opposite sides of an institutional power gap. It's inevitable in this context that conflict and tension will emerge.

The role of the Social Lab in this context is to co-facilitate a process that respects institutional and community interests simultaneously, by holding rigorous standards in regards to both institutional legitimacy and community agency. This requires striking a delicate balance, an open center that facilitates input, dialogue, listening and understanding in both directions.

While this approach is clearly well-intentioned, the danger is that a new form of control replaces the old. Even the idea that the Trust and Roller were enabling or "giving" agency directly to the community is patronizing. At best, these institutions are drawing down their own hegemony, slowly challenging the bureaucratically intensive and hierarchical policies of the philanthropic world. At this constantly moving edge, new forms of community engagement continue to hide and reproduce power inequalities in a rhetoric of participation, while continuing to control and dictate how projects are run. This new form of disguised control allows content to come from the participant team and community, while dictating the processes and forms that define the space of the project more broadly. This means that while agency is shared, the backstop of power and the allocation of resources remains with the Trust and its international network of experts.

While power appears to be "more-shared" in the context of a Social Lab, and therefore constitutes a move in the direction of shared agency, power inequalities are still divided along familiar categories: race, geography, institutionality and access to resources.

Race & Nationality

The hosting team was a melange of a few different groups of people: The Roller team, staff from the Trust, and a number of local facilitators and support professionals. While the hosting team was quite diverse, the Roller team was not, consisting of 90% white people. While more diverse than the Roller team, the team at the Trust was also more white than the participant team. This racial dynamic played out in problematic ways and contributed to an experience of the Roller team and the Trust as outsiders. One participant asked why there were "outsiders coming from England to tell us how to do what we're already doing."

Additionally, the participant group was dismayed that so much had been done on the project by the convening team before bringing it to the community. The fact that a great deal of "preconditions" work had been done, including interviews and defining the challenge prior to bringing the whole group together, was received by some participants negatively. One participant expressed that they felt like they were being invited into something that was developed by those on the other side of a power differential.

Furthermore, branding of the project as a "Social Lab" was not well-received by the participant group, to say the least. During the kickoff workshop, one participant said, "Why is it called a Lab? Are we being experimented on by white people?" The Participant team explained that the term "Lab" was off-putting and even offensive for a number of reasons, including the ring of social experimentation, and the history of racist socio-economic policy and political inequality in the city.

The question of "experimentation" brought a degree of discomfort to the Roller team, in part because there was a grain of truth to it. While the Roller team and the Trust were not "conducting experiments" on the participants of Grove, the project itself was a first; it was an experiment. While members of the Roller team had decades of experience in social change projects and Social Labs among them, they had not yet run a project together. Roller as an organization was quite young, and this was its first full-scale project. It was a real-life example during which the team could try working together, see what works and what doesn't, and prototype their own practices in client and community engagement, facilitation and process design at an almost unprecedented scale.

This "experimental" nature of a Social Lab being run by a new team is not necessarily problematic until you bring in the range of inequalities that are present in the context that the project itself is trying to address. An unseasoned team of non-local, mostly white, international expert-actors; partnered with a powerful Loop-centric financial institution; enter an almost entirely African American community with a long history and rich experience of social engagement; and plan to design and facilitate a process of social change.

In this context, it's difficult to see how the project would *not* reproduce racial inequality and problematic power dynamics reminiscent of the "colonial moment".

For Profit Vs Nonprofit Bottom Lines

Needless to say, there was a palpable tension between people working at the Trust, in Community Based Organizations, and the Roller team. This was also expressed in terms of perceived differences in motivation and culture between for-profit and nonprofit organizations.

For example, there was a sense on the part of the Trust that Roller wanted the project to go ahead at a faster pace than the Trust was initially willing to go. This was at least partially true. Roller leadership expressed concern that if the project didn't get off the ground expediently, political conflict, disagreement, or drawn out decision making processes could take over, sabotaging the project and its potential impacts. Furthermore, the sooner community members were more deeply involved in the project, the sooner agency could be concentrated at that level of the system. If the project didn't get off the ground, this would be bad for a number of different actors in the system and for a number of different reasons. Roller would no longer have a project, but perhaps more importantly, a new community-driven project would not have a chance to succeed. While it is not true that profit drives Roller's core agenda, its survival requires meeting a bottom line, which in turn fuels an agenda to produce work that is effective, timely and visible, and to demonstrate impact and value. At a minimum, for-profit organizations have an interest in getting things done at a pace and with a set of outcomes that enables them to survive as organizations.

This tension resulted, however, in some people in the Trust having an experience of being pressured to move forward, or to move at a pace that was faster than the social system warranted. One of the Trust's high-level staff commented:

> When we brought the Social Labs concept in, we moved too fast. We did not take time. Roller was driven by an agenda to get something off the ground. [...] We allowed the goals of the institutions and the consultants drive the project and it ultimately undermined the whole thing.

Another way that this played out was during negotiations over contracts and budgets. There was an ongoing sense that the Trust was questioning the Roller budget and the Roller team in terms of their operating costs. The project budget was large. Part of the Social Lab approach is to mobilize resources commensurate with the size or cost of the challenge being addressed. This doesn't mean that the budget of the project should be equal to the cost of the challenge, but that it should somehow correspond, be proportional.

Part of the problem is that the nonprofit world is often funded commensurate to the costs of running organizations or programs, not addressing social challenges. This means that there is often literally no relationship between the aims and goals of an organization with respect to social change, and the funding that the organization or project gets. This leaves nonprofits underfunded, but is in part due to the culture of the nonprofit world and its relationship to the philanthropic world.

It's as if the power relationships in the world dictate relationships. Organizations are like people. They have character, interest. Nonprofits obviously want to get funded. Foundations want to ensure that their gifts are used wisely. For profit businesses want to make money. These are the bottom lines. The whole world has a bottom line. But that's not actually the whole story. While these interests to some degree dictate the relationships between these organizations and sectors, they often align on what they want. They all want to have an impact in the world. It was in these spaces of alignment of mission that the project found its bright spots, and its impact.

OUTPUTS, RESULTS AND IMPACT

The Five Teams of Grove

This section will explore in detail the five teams and prototypes that emerged from the process, and how they changed over the course of the project.

Bronze Bridge

Bronze Bridge was a pop up recording studio and incubator for musicians and sound artists, aimed at supporting creative livelihoods for youth. This prototype, originally called "Bronzeville Arts Collective", wished to support emerging artists in Bronzeville. They discussed equipment rental for film makers and designers, co-working spaces, educational programs, art studio cooperatives, art mentorship programs, computer labs and music recording spaces. Eventually, they decided to focus on music as a place to start. The team's mission stated:

> Bronze Bridge aims to support emerging artists in Bronzeville. Young people who graduate from high-school aged arts programs do not have a place to work and refine their skills. Bronze Bridge seeks to create space, resources and tools for emerging artists to turn their craft into a profession. While the overarching goal of the team is to support all artists, Bronze Bridge is starting off with a recording space supporting musicians, singers, spoken word artists and other artists who work with sound. This decision was made in alignment with the team's immediate skill set and connections in the community to a number of emerging artists in need of space to record.

Bronze Bridge's next iteration was to set up a pop-up recording studio at the Harold Washington Library and open up their offering to the public. A number of local musicians showed up to the prototype. The team also developed their brand, building a website and setting up advertisements to track interest in different concepts for the project.

Over time, while the prototype showed great promise, this team stopped working together due to interpersonal challenges, disagreements and time constraints on the part of participants.

Bronzeville Live

Bronzeville Live began by connecting young adults and CBOs in Bronzeville. They began by looking at existing community based organizations in the area and their relationship to the young adult they seek to serve. Their team held one-to-one meetings with young men in the neighborhood and representatives of community based organizations and learned two things: 1) Young adults who are unemployed, underemployed and not in school have clear aspirations for their lives, and 2) CBOs are interested in hearing from young adults to inform program design.

They then convened a small group of young adults and community based organizations, inviting them to reflect on their goals, challenges and aspirations. They named this convening "The Re-Write." Their objective was to identify concrete and immediate opportunities to support young adults in reaching their goals and achieving their full potential.

Bronzeville Live's mission stated:

> Bronzeville Live is a team of residents and community organizations working together to address the challenges facing our community. We are developing and launching a co- design process that brings young adults and community based organizations together to come up with ways that each can support young adults in reaching their goals and achieving their full potential.

This team's long-term aspiration was to take the project to scale, helping to connect CBOs to young people across Bronzeville.

Bronzeville STEAM

The Bronzeville STEAM team began as "Mentor Mingle", inspired to help young people in Chicago connect to mentorship opportunities by creating a mentorship app. They began engaging with youth and discovered that while there wasn't a lot of excitement for a mentorship app, there was a sense among youth of disconnection from place. They found that young people in Bronzeville didn't have a sense of ownership or involvement in the rich cultural history of Bronzeville.

They found that the geography of meaning for young African Americans in Bronzeville was often marked by the territories of gangs, and locations where violent acts had been committed, rather than the rich cultural history (e.g. of basketball and the Black Panthers) that was also present in the neighborhood. Bronzeville STEAM set out to organize cultural tours of Bronzeville for youth, led by elders in the community.

As their idea evolved, they began interviewing people from Bronzeville who were considered "torch-bearers". They connected deeply with the "old heads" in Bronzeville who had been involved in the rich cultural history of the area throughout their lives, and found that the history of Bronzeville was too rich and diverse to be encompassed in a cultural tour.

Instead of bringing the youth on a tour of Bronzeville history, they decided to try and inspire youth to participate in making history themselves.

This team's prototype thus evolved into a project to simultaneously document the cultural history of Bronzeville and empower a next generation of documentarians. They connected youth with elders in Bronzeville, and equipped them with filmmaking and audio gear, giving them the chance to document the old stories of the neighborhood and connect to an intergenerational perspective and sense of place. Participants in their prototype became filmmakers and documentarians, and had a role in creating the cultural history in the neighborhood they belong to.

Bronzeville Surge

This team, originally called "Bronzeville Voice", initiated their project by agreeing to focus on youth leadership. They began with the premise that if they were to support young people in Bronzeville to create resilient livelihoods, their efforts should be guided and led by young people themselves. They began by creating a survey engaging with youth throughout Bronzeville and asking about the core issues they face in their community and the ideas they have for addressing them.

They next set up a one-day meet-up to engage directly with some of the youth they met when they distributed their survey. During this phase they also attended workshops on youth leadership and facilitation, and met with a number of Community Based Organizations to understand their needs and strategies in engaging with young people. They overwhelmingly heard from young people and CBOs that young people need local, youth-led, safe physical spaces where they can learn, engage with community in a safe way, and feel at home.

At this stage they decided to pivot or radically shift their idea. They decided to launch a physical space to support young people and the other grove teams. They named their new project team "Bronzeville Surge".

Bronzeville Surge got wind that the historic YMCA building in Bronzeville was up for rent for community-oriented programs, and immediately expressed interest. They did a tour of the space and began meeting there to get a feel for the space and brainstorm what was possible. They played with the idea of fundraising for their prototype and begin launching and testing new programs in the community space. The Bronzeville Surge Team also hoped to host next cycles of the Grove, as well as expand established community initiatives that had already proven effective.

Over time, this team pivoted again, changed personnel, and landed on a more modest idea: a bike shop for youth called "Grove Eco". One of the team members had experience working in a bike shop and knew the impact that building your own bike can have on youth on the South Side of Chicago.

Many young people in Chicago remain within their immediate neighborhoods, never seeing other parts of the city due to safety concerns and lack of transportation and resources. Building or fixing one's own bike can be an empowering experience, enabling young people to see the results of their own work and instill a sense of possibility and pride. The bike shop itself provides a safe, focused space with access to skill-building, mentorship and creative community. Bicycles also provide transportation, enabling young people to explore parts of the city to which they didn't have access.

Restore Bronzeville

Restore Bronzeville, previously known as the Justice/Just Us team, started out with a number of different ideas for how to approach the issue of safety. These ideas came directly out of the Grove's brainstorming process during the kickoff workshop. One idea was a restorative justice center, another was an art-space and project to melt down guns used in controversial police shootings and turn them into medallions and a clothing brand. Other ideas were collaborations between police and young people including park clean-ups, community gardening, and a game night with basketball and board games. As they narrowed down their ideas, they decided to host a dialogue between youth and police to vet their ideas.

They hosted an event with a number of police officers, mentoring adults, and youth. They all shared stories with one another about what has shaped their lives over dinner, and then explored ideas for how to build a new relationship and community practice around the issue of safety. This team also shared the ideas they had been working on and asked for feedback from the community.

Participants in the event were overwhelmingly supportive of the group's ideas, but in particular was drawn to the idea of a space and dialogue committed to restorative justice. Community members and the police officers present overwhelmingly agreed that a new approach to criminal justice was needed to heal the relationship between police and the community and to build more safety and trust in the neighborhood.

As the project evolved, this team planned to build a youth-led restorative justice hub. Young people that were engaged by Grove participants would have the opportunity to visit other restorative justice hubs in Washington DC and Oakland, CA to learn best practices and insights from organizations already doing the work. The team would then support those youth to create a local restorative justice hub sensitive to the particularities of Bronzeville's culture and social issues. The aim of the project as a whole was to explore and build possible alternatives to a broken punitive justice system that fuels the incarceration of black youth.

Measuring Impact: Multiple Capitals

While the impact of the prototypes is self-evident in some respects, measurement and valuation of impact is not easy. For this reason Roller used the "Six Capitals" approach to integrated reporting. The Six Capitals are:

Human Capital: New teams & capabilities
Social Capital: New networks, relationships, and collaborations
Intellectual Capital: New knowledge and information
Financial Capital: New stocks, flows and distributions of money
Physical Capital: New products, services and infrastructure
Natural Capital: Ecosystem resources and access thereto

REFERENCES CITED

Public Documentation Archive

Full public documentation for the project can be found at https://www.grovearchive.org/

Project and Interview Reports

Integrated Report
Interview Synthesis Report
Kickoff Workshop Report
Studio One Workshop Report

[Studio Two Workshop Report]
[Studio Three Workshop Report]

Literature

Bourgois, Phillippe
1996 *In Search of Respect: Selling Crack in El Barrio.* Cambridge: Cambridge University Press.

Escobar, Arturo
1991 "Anthropology and the Development Encounter: The Making and Marketing of Development Anthropology." *American Ethnologist* 18 (4): 658–82.

Hassan, Zaid
2014 *The Social Labs Revolution.* San Francisco, CA: Berrett-Koehler.

Malinowski, Bronislaw
1929 "Practical Anthropology." *Africa: Journal of the International African Institute* 2 (1): 22–38.

Mitchell, Timothy
2002 *Rule of Experts: Egypt, Techno-Politics, Modernity.* Berkeley: University of California Press.

The Root Cause is Capitalism… and Patriarchy

LINDSEY WALLACE, PhD
Adobe Design Research & Strategy Team

JENNA MELNYK
Adobe Design Research & Strategy Team

The authors used anthropology and other design research methods to develop a new kind of study to capture the world of professional creatives and the people they work with. To uncover core collaboration challenges for professional creatives the authors asked asked them to walk through past projects, who they interacted with at different points, and discuss their affective experiences.

Critical collaborative problems for participants in this study stemmed from two factors: ever-increasing corporate demands to do more with less, and concurrent attempts to automate feminized administrative coordination tasks. To communicate actionable findings, the authors balanced systems-level thinking with the identification of the kinds of problems Adobe could and would solve. While large scale social change was outside the scope of actionable recommendations for a product design team, the implications of social structures on individual experience provided insights that widened our lens beyond the individual.

> What else does the history of ideas prove, than that intellectual production changes its character in proportion as material production is changed? (Marx 1848:25)

INTRODUCTION

In 2013, Adobe entered the world of software as a service by introducing Creative Cloud and transforming creative tools such as Photoshop and Illustrator from boxed to subscription. From the beginning, Adobe knew collaboration would be important for the success of a cloud offering. The initial release of Creative Cloud included features like shared file-storage and libraries functionality that attempted to make collaborative work easier for creatives. Users could now send their work through links that could be opened by anyone they wanted to share with— even those who previously lacked the Adobe software to open a design file. With a broadening of the range of people interacting with Adobe Creative Cloud, it was no longer enough to consider the solo creative user behind the screen. Instead, the company needed to understand and design for nuanced collaborative workflows and new audiences of stakeholders involved in the creative process.

In late 2016, the authors were working on different product initiatives but both realizing the importance of understanding individuals within their organizational context. The authors observed that while product teams needed to be laser focused on their individual products, as researchers they had the unique opportunity to take a horizontal perspective by understanding the effects of creatives' ecosystems on their individual experiences across products.

With all of this in mind, the authors teamed up to understand the dynamics between not just the different types of creatives working together, but all of the people involved in requesting, approving, and producing creative work in collaborative environments. They proposed a study to create a taxonomy of the roles involved in creative projects that would move the company's understanding beyond the user/stakeholder dichotomy to a model that would articulate the way creatives engage with a variety of roles active in the design process. They also sought to identify explicit and implicit collaboration pain points for creatives and stakeholders in hopes of finding areas where Adobe might add new value to the Creative Cloud offering.

METHODS

Participants

Target roles for this project were determined based on previous case-study and collaboration research that the authors carried out with creative teams at Small to Medium Businesses (SMB) and Enterprises.

Participants included marketers, project managers, designers, copywriters, content strategists, developers, creative ops people, design-managers, assets and rights managers, executives, producers, program managers, as well as legal and IT consultants. In addition to understanding the different needs of individuals involved in creative work, they recruited participants from a wide range of organizations to capture cultural and organizational differences that might shape or constrain collaborative practices, such as; security policies, how project teams were formed, who sat and worked together, and other aspects of our participants lived-experience. Companies recruited represented a mix of industries including banking, healthcare, food and beverage, apparel, and technology. Understanding whether working relationships were primarily in-person or remote, and whether teams were made up of people with a shared skill set or organized cross-functionally allowed us to understand how participants communicated during the process and revealed the challenges people faced around communication and translation of expertise in more depth. Participants were recruited through a mix of snowball-sampling, social networks, and corporate partnerships. The team conducted over 60 interviews with people in roles that interacted with creative projects. The research spanned multiple design disciplines, including; graphic design, branding, advertising, packaging, UX, apparel, and industrial design.

Procedure

The research team took a two-pronged approach in order to understand both the individual experiences of creative professionals and their team members, and also the collective challenges teams face when working together. The primary research method used was in-depth interviews and site-visits with sole participants across many companies in a range of roles. The researchers sat with participants physically or digitally and asked them to give in-depth tours of their daily digital and (when feasible) physical work environments. Walking through participants' processes in fine-grained detail and delving into the relationships that structured their days provided the researchers with the opportunity to learn the thick (Geertz) details of these participants life-worlds and conduct interpretive work to understand the meanings and motivations that underlay their practices. They complimented this approach with data gathered using a retrospective case study method to help uncover unarticulated organizational challenges that could only be seen when looking across individuals working together on the same project. Together, these efforts provided both depth and breadth to the work. Although the researchers used methods outside of the standard anthropology toolbox, the focus on deeply understanding participants daily life-worlds beyond the narrow lens of product usage brought it into the realm of the ethnographic.

Table 1. Recruiting Breakdown by Role:

Role	**Count**
Designer	10
Design Manger	10
Marketer	10
Project Management/Operations	16
Physical Production	3
Developer	9
Copywriter/Creative Strategist	4
Other	5

Independent User Interviews: Roles & Process – The 60+ participants the authors interviewed came from different companies. The interview style was flexible, participant-led, and not tied to a single project. During conversations over video conference, in participants' workspaces next to their desks, in conference rooms and even in quiet corners during Adobe's 12,000 person conference, Adobe MAX, participants recalled how they had personally contributed in the making of a past piece of creative work. Although not all of the user sessions were ethnographic (some were by phone), all research participants were asked to let the researchers into their professional life-worlds by screen sharing the digital places they inhabited every-day in which they worked, communicated and struggled.

Keeping the interviews semi-structured and user-driven allowed the researchers to learn over the course of the conversation which topics were most important to different user types, rather than imposing a standardized framework of topics on the users. Many participants also showed the researchers how they participated in creative projects through diagramming their process, team members, or the tools they used for their most important tasks. Through these different forms of communication, the team was ultimately able to understand the unique motivations and needs of different roles during creative projects. This method also allowed the team to uncover the ways that participants' position within an organization shaped their definition of their work, revealing the ways that what counted as a project were contested and fluent based on position within the organization.

Retrospective Creative Project Case Studies - The team conducted retrospective project case studies with 3 key enterprise customers in order to understand the nuances of team dynamics and unspoken challenges. Participants were selected because they had collaborated on a recently finished creative project together. The team spoke with in-house staff (such as marketing managers, creative directors, copywriters, and digital producers) as well as staff from partner agencies that collaborated on the projects (designers, project managers, account managers, etc.).

The research team conducted 1-on-1 in-person and remote sessions with each of these participants using a retrospective emotional journey mapping method adapted from the work of Evangelos Karapanosa, Jean-Bernard Martensb, and Marc Hassenzahlcwhen (2012). The journey diagram and line of questioning used guided participants to talk about real events that had happened during the project and directed them away from generalizations. It helped them recall information and got them to talk about their emotions when they might otherwise not have.

Retrospective Journey Template:

Figure 2. The retrospective emotional journey map tool, with a space for key events at the top (timing from right to left) and a space to plot remembered emotional state on the bottom (from high to low).

The researchers also followed a line of questioning to probe deeper into the highs and lows of the experience to identify pain points in key topic areas. Following the mapping exercise, participants were asked to show concrete examples from the project and discuss the tools and workflows they had described using. Later, during synthesis of the data from each project, the mapping method helped the researchers follow a "single story" and connect the dots between the different accounts told by each participant from a given project to also uncover unarticulated pain points.

The benefit of exploring past experiences in this way versus running a longitudinal ethnographic study had to do with a number of factors. For one, the retrospective emotional journey method allowed the research team to collect data quickly while maintaining a wide scope of information. This method also alleviated privacy concerns for participants who were often working on confidential initiatives. The focus on past work allowed participants to show more of their work as projects had often already shipped and the participants felt less constrained in speaking about their processes and experiences than they would walking outsiders through work in progress.

RESEARCH FINDINGS AND INSIGHTS

As the researchers began to synthesize the vast and varied qualitative data they were collecting, they realized they couldn't understand how the different project roles interacted without understanding the forces and incentives that shaped how the individuals and teams behaved. They saw two key themes emerge from the narratives participants shared:

1. Constant struggles about organization and coordinating projects to make sure everyone knew what they were doing and did their part correctly.
2. People being asked to do more with less on tighter and tighter timelines in a relentless drive for efficiency that left little space for creative unpredictability and exploration.

Regardless of their role, most of the informants described an intense pressure to work as fast as they could, feeling constantly stressed and overwhelmed by many small kinds of disorganization including tight deadlines, missing materials, and not having what they needed to do their work at critical times.

To understand why, the researchers found themselves in the surprising position of using critical feminist theory to find business insights. The framework that emerged from their data was built on the foundation of theory from Lindsey's time as a feminist scholar and women's studies instructor, as well as Jenna's deep understanding of business needs and processes and the lived reality of producing creative work under late capitalism.

Lack of coordination as a symptom of gender ideology and gendered work

Across the large companies and agencies researched, project management positions involved unacknowledged care work. One agency project manager interviewed described how to get a powerful man on her team to provide feedback on images, she had to lay them out on a table, physically walk him over from his desk and sit with him while he chose the images. She described sending the creatives on her team email invites to meetings as the first step, one often followed by individually reminding them to go to their meetings, either through a digital reminder on Slack or email, or by stopping by their desk. Without her, nothing happened. She explained, "If I need the creative director to review images for a project, I have to go lay out all the images on the conference room table. Then I have to walk with him from his office to the conference room and sit with him until he's done selecting images, otherwise it will never happen."

Similarly, at one large corporation, the project manager's primary role was managing requests and feedback to design. They controlled who could ask for design work, made sure requests were done correctly, and managed the flow of creative work and feedback, protecting the designer's time and attention, recording information for them and putting it where the teams could find and access it. While this work was critical, it was often under-acknowledged, and the labor of coordination, including meetings and complying with project management guidelines, was almost universally despised and many individual participants (outside of the project management discipline) dreaded it and avoided it at all costs.

Significantly, this research revealed that administrative labor was universally coded as "project management" although it included a large and unspoken element of clerical and secretarial work. Women first started working in offices as "typewriters" during the civil war, and by 1950 secretaries, stenographers and typists made up the largest group of American working women (Berebitsky 2012:9). These tasks are firmly coded as feminine in mainstream American culture. These feminine jobs have historically been targeted for automation and replacement with technology, beginning with the typewriter and continuing with the replacement of secretarial and project management coordination work with project management and task tracking software.

However, the research done for this project revealed that coordination and what Hardt called "affective labor" (Hardt 1999:90) could make or break whether creative projects got delivered successfully, and revealed the ways that the (almost exclusively female) project managers interviewed not only provided information, but also did exhausting affective labor for their teams. Affective labor, while hidden and intangible, is central to material labor, and is part of the costs of production. The replacement of these roles with technology represents not only the feminization of these roles, but provides an example of how companies externalize costs onto their workers. In this case, the cost is the investment in time and emotional labor needed to coordinate successful work across disorganized groups made up of fragmented organizational structures. Through controlling timing, resourcing and scheduling, project managers controlled the conditions of production for creative teams. However, they did not have autonomy; on the contrary, they were constantly squeezed between executive and client demands for efficiency and creatives' need for unpredictability and desire for exploration time.

Lack of coordination caused the biggest headaches for the participants in this study, regardless of project roles. The depriortization of care work manifested in late and missing requirements, shifting timelines, delays in getting feedback, and lost creative work, among other things. Understanding the feminized history of care work and secretarial labor allowed the researchers to unpack the seeming paradox between the centrality of project management to the success of creative projects, and the explicit devaluing of that same work expressed by many of the participants during the study. Participants viewed tasks of coordination as separate from the creative work, although good and caring project management was often the difference between success or failure.

"Do more with less" and the drive for efficiency

Before this study, the researchers hadn't questioned the devaluing of project management and care work. Like most modern office-workers, they viewed secretarial and administrative staff as a luxury, nice to have but not essential. This research led them to the opposite conclusion, that the lack of administrative staff and the devaluing and shifting of their work led to pervasive problems across creative projects and organizational roles. This insight demanded further analysis to understand what was causing this apparent contradiction.

The answer came from thinking beyond the scope of the individual roles, projects, teams, or organizations. The researchers found themselves faced with understanding the capitalist imperatives for efficiency. This study revealed environments where professional creatives and their collaborators work in a context where their baseline M.O. is rushed, overwhelmed and overworked. Participants described being incentivized to focus on executing as quickly as possible and thus de-emphasizing collective goals. Because of this, "official" processes and tools are often ignored as individual contributors look for ways of maximizing their own productivity (new features, tips, and tricks) vs. the productivity of the organization as a whole.

No student of capitalism would be surprised to learn that the interests of the workers and the interest of capital end up at odds, this insight dates back to Marx's foundational 19th century work (Marx 1848). These opposing forces rarely need to be called out in product research. These conflicts surfaced because the researchers took an ethnographic approach that encompassed the professional life-worlds and affective experiences and work of their users.

This central misalignment of interests caused a number of productivity problems for the participants of this study and reduced their perception of the quality of their creative work. Most of the challenges people faced during their work came down to the self-service model of automated project management combined with organizational drives for efficiency that continually tightened timelines and starved them of critical organizational support. These were problems software could not solve. The logic of capitalism creates a focus on execution over strategic work that limits the effectiveness of creative teams.

The research team saw this play out on the micro-level of daily life in a number of ways. An interview with an in-house developer revealed the challenges he faced because of the collapsed timelines his team worked under. When asked about what he needed from design to start building and what design "handed off," he revealed that because of their timeline his team had to start building before the received designs. As a result they were only able to incorporate them in an approximate way. In this situation he received unfinished drafts of wireframes and work-in-progress from designers that he used as a guide to begin his work. However, since the creative work was unfinished, what he was building was continually in flux and he had to do a lot of redundant work revising the product as the designs and specs changed overtime. He was very uncomfortable with this situation and the developers on his team felt wrong about building in that way. However, because of the production timeline, they needed to start building while the design team was working. This overlap of design and development timeframes meant that they had to make time for extensive meetings to review changes with the design team and discuss how to fix things that they had built incorrectly. He expressed embarrassment during the interview as he discussed his process with the researcher. For this developer, the drive for efficiency that demanded they started building before the designs were complete meant extra work and time spent, as well as an incredibly stressful work experience. However, from his company's perspective, this seemed like an efficient way to "save time" by having development begin on schedule even in the face of design delays.

Companies try to reduce complexity by narrowing responsibilities of end users and focusing them on specific specialized tasks. However, end users need to understand the broader context they are working within in order to avoid unnecessary changes later. There is a related conflict between corporate transparency and accessibility of assets and end users needs to have for work-in-progress (WIP) spaces. In the end, the research articulated different relationships and workflows involved in the creative process and described nuanced differences between how these different roles relate to and interact with one another.

IMPACT & CONCLUSION

The product development work informed by this research is ongoing and can only be shared in a limited way. However, the overall impact of the work at Adobe can be discussed in more depth. This research took approximately 6 months of additional presentation and socialization to be adopted by the core product audiences. Product teams that were used to focusing on narrow user groups needed help learning how to think horizontally and view people in the broader context of their professional lives. The amount of data in this research made it challenging to present. The team solved this problem through creating two large presentations based on their findings from this research:

- An overview of collaboration challenges and recommendations that spanned almost 100 slides.
- A taxonomy of key project roles that described the roles involved in creative work as well as the challenges these people faced individually and working together.

These artifacts went into Adobe's internal database, where they "went viral" and have been downloaded over 300 times. Although the research team did this program of study for design and product needs, this framework is also being adopted by other teams. Adobe stakeholders across the company, from executive leadership, to business-development and marketing, have found these reports and reached out to the research team for help adopting the frameworks and incorporating the findings into their business initiatives. This work has spurred not only new ways of thinking about people involved in creative work externally, but also improvements in internal processes and organization at Adobe, which is of course also a creative enterprise. At a high-level, here are some of the ways this research is re-shaping how Adobe thinks about enterprise collaboration and creative work:

Understanding the misalignment of interests between businesses and their workers under modern capitalism and how that manifested in day-to-day creative work pushed the company to think more critically about users and how they differ from customers. Teams ask what are the corporate "customer" needs vs. employee "end user" needs.

This research also spurred reflection about the macro impacts of technology and discussions and questionings about the social and personal impact of certain product development decisions. That conversation is ongoing and ties in with the larger movement of ethical design in technology (Montiero 2017).

One profound impact of this research was organizational rather than manifest in product. This research showed the value of in-depth ethnographic studies that encompassed entire teams and creative processes. Since this research was not focused on how people interacted with a specific product, it could be picked up and used by many teams across the company.

Through this study, the research team has also been able to demonstrate the predictive strength of exploratory research that can often be written off as "too general." Since the release of these reports, multiple prototype evaluation studies have been conducted, resulting in findings that the authors predicted based on what they had learned about creative teams during this ethnographic work.

This work has also contributed to changing the way Adobe product and design teams work together. Research is agnostic and can play a key role in aligning teams early on through exploratory research. Although the researchers had hypothesized that things like workplace culture, industry, and team dynamics would shape people's needs, they learned that the impact of social and economic forces outside of individual organizations had much deeper and broader-reaching impacts on how work was accomplished than they had imagined at the outset of the project.

Lindsey Wallace, PhD is a Researcher on the Adobe Design Research and Strategy Team. Email: Lwallace@adobe.com.

Jenna Melnyk is a Senior Researcher on the Adobe Design Research and Strategy Team. Email: Melnyk@adobe.com

REFERENCES CITED

Karapanos, Evangelos, Jean-Bernard Martens, and Marc Hassenzahl.
2012 Reconstructing experiences with iScale. *International Journal of Human-Computer Studies*, Elesevier.

Tzanali, Rodanthi
2011 (Dis)organized Capitalism in *Encyclopaedia of Consumer Culture*, Publisher: Sage/CQ Press, Editors: Dale Southerton.

Marx, Karl.
1861 "Formal and Real Subsumption of Labour Under Capital. Transitional Forms." From. *Marx's Economic Manuscripts of 1861-63*, pp. 93 121.

Marx, Karl & Frederick Engels.
1848 *Manifesto of the Communist Party* Marxist Internet Archive

Montiero, Mike
2017 *A Designer's Code of Ethics* Mule Design

Hardt, Michael.
1999 "Affective Labor." In *Boundary 2, Vol 26.2 1999* Duke University Press

Wallerstien, Immanuel
1979 *The Capitalist World-Economy* Cambridge University Press

The Transformative Power of Singular Stories: Making the Case for Qualitative Evidence in Healthcare Contexts in Colombia

JULIANA CARDONA
A Piece of Pie

JULIANA SALDARRIAGA
A Piece of Pie

MARIA FERNANDA ESTUPIÑAN
A Piece of Pie

PAULA GAMBOA
A Piece of Pie

In this case study we describe how we collaborated with a Colombian healthcare provider company and enabled its decision makers to understand the power of stories and other types of qualitative evidence in healthcare contexts. The stories became a tool for recognizing singularities in a complex, massive system, where individuals were constantly reduced to social security numbers. We describe the qualitative methods implemented, such as in-depth interviews, projective techniques, shadowings and observations, explain the difficulty in explaining the value of our qualitative evidence and mention some of the lessons learned throughout the project. We also discuss the project's outcomes, such as understanding the difference between user perception and user experience, the impotance of healthcare providers to go beyond healthcare and using stories as input for measuring quality of the service. We also recall the transformative power of stories: stories allowed the client to emotionally connect with the individual behind the singular story. This empowered them to generate their own initiatives, taking our qualitative evidence as a starting point.

INTRODUCTION

This is a story about stories. It is a tale about the transformative power of stories in medical and healthcare contexts seen through the experience of a research team working with an insurance company operating in Colombia's complex healthcare system, and how, when taken as valid evidence, stories can be complemented with quantitative research and be a tool for decision making.

The case study will also evidence the transformation of an organization from a process-centered focus to a people-centered vision and how qualitative research in healthcare contexts is beginning to be recognized. Appreciating a qualitative focus is happening at a slower pace in developing countries such as Colombia, where public and private health systems rely on quantitative data due to government demands based on the performance of health care institutions. These systems are struggling with limited infrastructure, a growing need for chronic care when systems are designed to provide acute care, among other challenges, that prevent them from recognizing the value of other types of data. Thus, there is an opportunity for ethnographers, to show these organizations the importance of working for people, which is a step towards granting memorable experiences, beyond efficiency and fund management.

The transformations that occurred as a result of the project can be understood as follows: first, the client's mindset was transformed, in the sense that they understood not only the relevance of listening to stories, but of collecting them as well. Second, the user experience was transformed, transcending a list of social security numbers and seeing the wholesome, complex human beings that composed it. Finally, stories also transformed the team, in the sense that they tested its capacities to adapt and to remain resilient in the face of challenges and failures. According to Maria Giulia Marini (Fiorencis 2016), narrative medicine is placed side by side with evidence-based medicine (EBM). On narrative medicine there is an approach to care more focused on human beings: the patients, the healthcare professionals and the relatives allowed to understand the culture, the values, the needs, the passions and the personal and professional projects.

The case study is divided into several sections. First, the context will be described, which consists in a thorough description of Colombia's health system and its changes and challenges. Afterwards the methodology that was implemented throughout the project will be mentioned, explaining why the team chose the methods it chose. In the outcomes section three key findings will be mentioned, which include the importance of understanding the different between user perception and user experience, the importance of healthcare providers to go beyond healthcare and using stories as input for measuring quality of the service. Then the impact will be presented, focusing on the power of stories to transform the mindset of the client and the team's capabilities. Finally, a few conclusions will be drawn.

A CONTEXT OF REFORMS AND MORE THAN A MILLION USERS

During the 90's, most countries in Latin America reformed their healthcare systems to widen coverage and ensure access for patients, and Colombia was not an exception. This country had one of the worst systems in the region, characterized by low coverage (just 30% of the population had access, according to the Ministry of Health), high rates of inequality, old technology and deficient medical practices. So, in 1993, the government introduced an ambitious reform package, described in the "Law 100" of that same year, which stated health was a right of all citizens and called upon the decentralization of health services (Barrera 2014). Since the central government was unable to ensure universal coverage, healthcare would be provided by a series of private and public companies called "Empresas Prestadoras de Salud" or EPS (healthcare provider companies in Spanish)". EPSs do not provide medical services, but they manage public resources from workers contribution and taxes in order to promote affiliation to the social security system. People join EPSs to be treated in clinics and hospitals called "Instituciones Prestadoras de Servicios – IPS (service provider institutions in Spanish)," which do provide medical services but do not charge patients for them. EPSs would start providing access to quality healthcare, and in return the government would pay them periodically. Thus, insurance became the main instrument of healthcare, and each company would have a limited number of users to look after.

Nowadays, health experts and decision makers have reached a consensus on the positive evolution in coverage and access after this reform was implemented (particularly with respect to the poorest), however, it is also true that these companies are still facing challenges that have shown resistance to solution, as well as new challenges that did not exist during the

1990s. For example, in low and middle-income countries like Colombia, chronic diseases now surpass infectious diseases as causes of death, due to the significant increase in life expectancy that characterizes societies in transition (Forman & Sierra 2016). As many other Latin American health systems, Colombia's insurance companies and healthcare providers are not prepared for these new patterns of disease, since they were designed to provide episodic or acute care rather than chronic care. Additionally, besides a growing aging population that requires health services more often, the total amount of users in the system has also increased by 15 to 20% each year. To relieve these companies, the 1993 reform has been accompanied by a notorious rise in public spending, but the quality of the service begun to decrease, and the system's debt still surpasses its returns (Barrera 2014).

It was in this context that one EPS became interested in providing its users a positive user experience even though it was a very process-centric organization. Motivated by the high number of complaints and claims related to the service, the EPS decides to work on improving the experience of its clients understanding that in health contexts, people are exposed to emotions such as fear, anxiety and confusion, elements which had never been taken into account in their evaluations and (key performance indicator) KPIs, which could have great impact on the patients' experience. Due to confidentiality agreements, the name of the company cannot be disclosed, for this reason it will be called company POP.

POP is on the top 5 EPS in the country, and it is an organization that is extremely sensitive to cost. Any adjustment, however minimal it may seem, can have a very high impact on the total cost. Unlike other industries it is not possible to assign a monetary value for what each person pays for the service and is not directly proportional to the profitability of the total of members. Precisely, by being part of the health system, its income is determined by a value called "Unidad de Pago por Capitalización – UPC (Payment Unit for Capitalization)" which is the annual value that the Health Ministry recognizes for each of the members of the social security system and must ensure coverage and quality according to the "Plan Obligatorio de Salud – POS (Mandatory Health Plan)" for all its members.

Although POP had previously made some efforts to improve the patient's experience, these solutions had been designed based on internal company hypothesis, thus the projects did not have the expected success and it seemed impossible to generate profitability by increasing the value of the service due to the UPC metric. For these reasons, the EPS decided to do it as a separate project, choosing a team that could dedicate 100% of its time and be able to make the necessary adjustments to the experience in order to diminish a potential reputational risk, which could affect other business units caused by the unification of the brand.

Initially, a POP team with the understanding of the processes was chosen to develop the project, and it was believed that their focus would be on adjusting these processes. However, the team decided to work from a different perspective, they had heard about "customer-centric" strategies and looked for an ally to work with hiring a research team composed of designers, sociologists and anthropologists, a team that quickly acknowledged the company was venturing into unknown territory since they were accustomed to quantitative results rather than qualitative insights. Most companies and healthcare providers in Colombia are concerned about the operation and do not consider the ways in which they are creating a positive or negative impact on their users' experience. This, plus the fact that the number of users exceeded the million according to official government data (Datos abiertos 2017), implied this project would be challenging, but quite invigorating too.

RESEARCH METHODS

Making the case of qualitative research in a quantitative world

As ethnographers and designers working with business organizations, a challenge constantly faced is proving the relevance of qualitative research methods and results and explaining how this relevance is different and equally valuable to that of other types of evidence. In the service sector, decision makers want evidence that allows them to evaluate their organization internally and externally, which is why the KPIs, satisfaction metrics and other quantitative customer experience indicators are permanently being implemented. But what happens with healthcare organizations? How can evidence obtained by qualitative means complement this need for constant evaluation, while generating additional value?

This situation was the first challenge faced by the research team: unlike a quantitative approach, the outcomes of a qualitative project were not going to yield immediate returns. This meant convincing POP, a health insurance company trapped in a struggling and financially unstable system, was going to be a difficult task. However, from a more optimistic perspective, this situation was also an opportunity to start changing the mindsets of stakeholders in the system: was increasing public spending really the only way to guarantee adequate healthcare? What if companies considered a human-centered approach as an alternative to surveys, polls and statistics to improve health services? This, of course, meant thinking about users as whole and complex beings, and not as digits in a seemingly endless list of social security numbers.

Before heading into the field, key stakeholders within the EPS such as managers and people that participated in previous explorations were interviewed. They shared their vision of the company, the outcomes of the previous projects carried out, positive results and failures, problems already identified from the quantitative data, and expectations about this new project. Other immersions were also held to learn about the functioning of the Colombian health system, the role of the EPS and its different services to which patients could access. At first, it was decided that the research would not be focused on users dealing with specific services, since the objective was to understand user experiences in different services. Afterwards, when fieldwork was concluded and the team started designing solutions, those services in which the major deficiencies were found were given more importance: for example, the ER service and a new model for vulnerable chronic diseases. These services were chosen according to the results obtained in the satisfaction metrics: IPSs, General Practitioners (GP), Specialized Physicians, Labs, Diagnostic Imaging Centers and Pharmacy would be the first approach for the project. The team knew the first step towards designing a positive experience was understanding and mapping the current one.

In-depth interviews and projective techniques

As a starting point, it was essential to choose the fieldwork participants very well, therefore POP provided a database that was already divided into 5 patient segments, based on the types of service they used with more frequency and demographic variables such as the patient's age (client had already done a few things on its own, so the team was not starting from scratch). After the first analysis, this database was thorough and updated so it was possible to understand what services were actually being used by each patient, allowing the team to have a closer look on the different elements to be investigated throughout the research phase.

Based on the database given by the organization, the research team selected 40 patients, 8 per each segment. In-depth Interviews were conducted mainly in their homes and work places where participants could express themselves in a trustworthy and comfortable environment, and where they could feel that POP was concerned about listening to them, something very different from what happens in a customer service center in which the patient must be the one who takes the initiative to be heard. This situation was usually a negative experience for the patient because if he tried to communicate through a telephone line, he could hardly communicate with an advisor and if he approached the place in person, he rarely received a timely and empathetic solution to his problem.

To understand the perceptions and patient's experience with POP and its different services, the interviews followed a semi structured approach where questions were established as a frame of reference and supported by projective techniques. Projective techniques are tools and formats represented in images, heat maps, journey maps, among others, that allow the visualization of unconscious elements based on perceptions and experiences that are not usually expressed during the conversation. These methods work as effective method as they draw out perceptions, desires and emotions in a creative way (Kalter 2016).

If the project had used only in-depth interviews, the participant could have only answered what he or she considered convenient or appropriate for the researcher or also obtain biased opinions based on factors such as the affinity between the participants and the EPS for being in the same geographic region or the testimonies based on negative trending topics on EPS. However, with these techniques, participants revealed perceptions that they had not mentioned through their statements.

The first part of the interview with the patients delved into general perceptions of the health system, health providers and specifically with the organization. Next, we inquired about their personal experience with different services to which they had accessed at some point. This last exercise was carried out based on mapping exercises on paper to oversee the experience route in each service. For this, a card sorting exercise with different health services, touchpoints and emotions was implemented. This allowed the team recognize pain points, common barriers, leverage points and elements with greater impact on the experience route of each service. It was necessary not only to know their perceptions but understand each detail throughout their experiences.

Shadowing

For shadowing, participants were also selected from the EPS´s information system that maintained data related to types of service and patients who were going to use them in upcoming days.

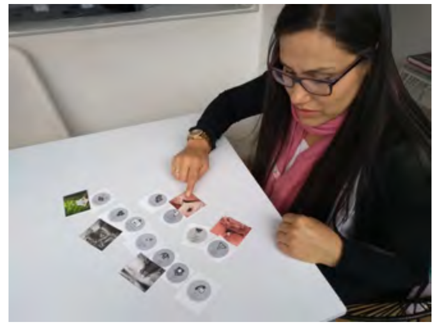


Figure 1 and 2. Example of informants narrating their stories through projective techniques. Photograph © A Piece of Pie, used with permission.

This accompaniment began at the patient´s home or place of work and permitted the researchers to accompany the participant in a real-life experience with the Emergency Rooms and other health services. For these observations, consent forms were essential. This made it possible to observe times and distances between these and the places where they received their services. With this technique, researchers were able to live the stories and be part of the experience, observing the step-by-step journey for different services: general practitioners, specialists, laboratories, diagnostic aids, claim of medicines, etc. The stories were captured firsthand.

Fly on the wall

Unlike shadowing, the "fly on the wall" observation methods do not concentrate on the experience of a particular patient but focus on different elements of the environment and the experiences of people present in a specific place. This exercise is where the sensitivity of the anthropologist flourishes because every detail is important: from the faces of the patients and their families that have been there waiting for too long to the expression of the nurses each time a sick baby enters the room. The interactions with different actors, space conditions, service protocols, times, among other aspects were observed.

These observations were made at IPS accessed by a large number of patients to request a variety of services. These IPS were selected by POP based on location, number of patients and health services offered where its General Manager was previously informed, and authorizations were carried out with the help of POP.

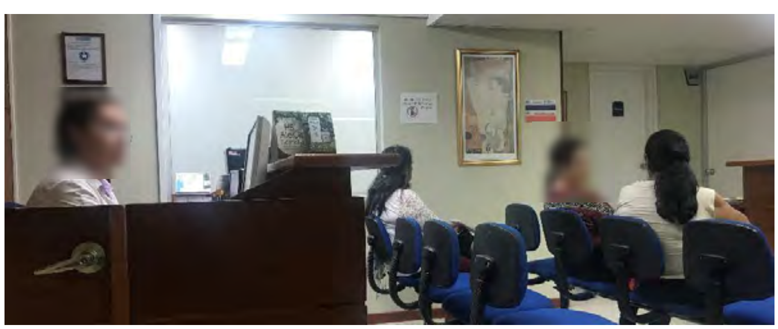


Figure 3. Example of Observation at IPS. Photograph © A Piece of Pie, used with permission.

ENTER STORIES: A USEFUL TOOL FOR COMMUNICATION

Once the research methods described above were approved and the informants recruited, the team headed out into the field. The chosen techniques were successful because they allowed the team to build empathy with the informants, which meant the informants went from being shy to sharing their experiences with great detail and enthusiasm, recalling not only what had happened but how they had felt. It did not take long for the team to perceive great potential in these experiences, and to understand them not just as experiences but as stories. Classic storytelling elements were easy to identify in their conversations with informants: there was a main character trying to achieve a goal, an antagonist and a setting in which a series of events took place. The narrator was also an important element, as he or she told the story with great emotion and authority. These elements will be easy to spot in the stories that will appear in further sections.

These powerful stories convinced the team that the project was not just about what the informants were saying, but how they were saying it. Their role both as narrators and characters showed they were not passive actors trapped in the country's healthcare system; instead, they were always making conscious decisions to improve their current situation and they even wanted to share these decisions. This way the team understood this project was about communicating these stories, guaranteeing they reached a wider audience and suggesting their power to lead a transformation from a process-centered organization to a human-centered one as will be discussed below.

Since patient experiences told through their stories had become such an important part of the project, the team decided to take a step back and review the use of stories in healthcare contexts. They found out that The British Medical Journal *The Lancet* published a report wrote by Helen King (2014) which suggested that healthcare systems were still unable to truly comprehend the contexts in which patients lived. Because of this, the journal called for refined methods to better understand these contexts, such as qualitative methods and, more specifically, stories. The importance of medical histories and individual accounts of illnesses and symptoms show that stories are already present in healthcare and medical contexts.

There are several reasons why stories are relevant evidence in healthcare and medical contexts. First, stories create empathy. They allow us to understand that other people will experience health services differently, or in other words, that there is not just one way in which these services are experienced. Empathy is also important because it is easier to relate to an individual than to an abstract group (think "Maria" or "Juan" rather than "patients" or "users"), and because content that is emotionally touching can be more enticing than plain data. Second, humans are accustomed to thinking in story terms, so they are more likely to retain information that was structured as a story. This means the use of stories need not be limited to discussions between researchers; researchers can use them to express the results of formal investigations in a way that appeals to different actors, including stakeholders, doctors, patients, policymakers, companies and even politicians. Finally, stories as relevant evidence in healthcare have encouraged organizations to design experiences rather than products or services. This has not been an easy transition, since it requires deep organizational changes as the creation of a new area or restructuring internal processes impacting operational and behavioral levels, or in other words, a true change of mind and heart.

Indeed, the benefits of narrative research are being identified and praised across this field. Stories have become a frequent topic in healthcare innovation, so medical actors as well as consultants and designers can easily access it via conferences, TEDx events, web tutorials, and case studies such as the present one. Additionally, several institutions such as hospitals, insurance companies, laboratories, governments and universities are actively encouraging the use of human-centered design in healthcare. In terms of hospitals and clinics, the San Joan de Deu Hospital in Barcelona is worth mentioning, for the creation of its Innovation Department almost ten years ago. In the United States, Mayo Clinic and Cleveland Clinic are just as pertinent, the first for being a pioneer in this whole process and the latter for organizing the annual Patient Experience Summit, an event that brings together patient experience leaders, healthcare CEOs, innovators, policy makers, stakeholders, industry experts and patients who are committed not only to the patient experience, but to the experience of other actors that may be found in health systems, from nurses to pharmaceutical companies. The reason for this is that listening to and ensuring the wellbeing of these actors will inevitably have a positive impact on the patient experience.

All this previous work was considered by the research team when it decided stories were going to be the main vehicle for transformation in this project. Just as Sir Isaac Newton humbly admitting his success as scientist was because he was standing on the shoulders of giants, so the team was thankful for all that had already been done in terms of narrative research in healthcare. However, convincing the client of the importance of qualitative methods and ethnography was one thing, but using stories and narrative research to communicate results was another. While on fieldwork, the team decided that decision makers needed to be invited into the field and attend the conversations with informants. The logic here was that it would be easier for the client to understand the power of stories if it witnessed them first-hand, as opposed to receiving them on a sheet of paper or a presentation. It was about showing rather than telling what the users were experiencing and feeling when interacting with their services. However, the team suggested the client should be prepared before heading out, so a few instructions were given here regarding proper behavior and dress code.

RESEARCH FINDINGS

User perceptions may contradict user experience

Perceptions are an important dimension of the user experience. These can be understood as the user's appreciation or envisioning of a product or service, and may determine the expectations that a person has before the interaction with said product or service occurs. Considering health and healthcare are important and recurrent topics in people's lives, it is safe to assume an individual that is about to have a doctor's appointment or walk into a hospital already has a perception and a series of expectations of this interaction.

This was exactly the case of the present investigation. Soon after fieldwork began, the team noticed that the informants' perceptions were quite different from what they actually experienced with POP. There were several reasons why this could be happening: first, POP is based on Medellín, a Colombian city located in the department of Antioquia. Unlike cities in other regions of the country, cities in Antioquia were not founded by Spanish colonists, but by Colombian families looking to settle down and become self-sustainable. *Paisas*, what people from the Antioquia region are called, have appropriated this past and still identify themselves as hardworking and community-oriented, which is why it is expected for them to praise and respect companies founded and based in Antioquia, such as POP. This positive perception of POP is so strong it might hinder negative experiences.

For instance, Miguel and Sofía have been together for more than 30 years, and Sofía has a severe medical condition for which she has to take more than one medicine. Her EPS, POP, always gives her the amount she has to take for a month, which means the couple had to visit the hospital three days every month in order for her to collect all the different medicines she had to take. Although the couple admitted this was inconvenient, when asked about POP they admitted they were very grateful and appreciated their EPS very much.

Figure 4. This painting by nineteenth century artist Francisco Antonio Cano has been appropriated by *paisas*, as it depicts the vision and determination of those Colombian families that first settled in Antioquia. These feelings or pride are so strong and engraved they determine what people in this region of Colombia think about themselves—including their companies.

Second, just as positive perceptions hindered negative experiences, so could negative perceptions hinder positive experiences. In daily conversations or small talk, it is not uncommon for Colombians to complain about the country's healthcare system, regardless of the EPS they are affiliated to. This is similar to what happens with traffic: since Colombians are accustomed to deficient traffic conditions, they are likely to complain about mobility, even on days when their experience was not bad at all. This generalized disapproval of the country's healthcare system is reinforced by the media, as news broadcasts emphasize corruption and other scandals within the system.

Because of this, the team figured addressing the challenges and pains contained in actual experiences would generate more value and have a greater impact than addressing those contained in perceptions. Thus, they decided to put the informants' perceptions aside and focus on singular stories, by asking informants to narrate a specific interaction with POP. This meant asking them what they did that day, how long it took them and how they felt, rather than inquiring about their feelings towards POP in general. When asked about a specific event, people are likely to recall their actual experience, which can be far better or far worse than their perceptions. While other research methods such as surveys would not have allowed the team to distinguish between perception and experience, singular stories did and even suggested these two could perfectly contradict.

To remain relevant and impact users, healthcare providers must go beyond healthcare

The team noticed there were elements in the informants' contexts that had a tremendous impact on their healthcare experience, even when these elements were not directly related to health, such as the education and socioeconomic status or the mobility in each city. Thus, POP had to go beyond healthcare and become more context-sensitive in order to positively impact its users. The client's first reaction was to reject this suggestion, since it could not understand why the research team was proposing something as idealistic as providing users adequate healthcare, while addressing their socioeconomics needs a well. However, the goal here was POP and the team to sit down and negotiate and redefine what was under the scope of the POP. It turned out there was a lot to be done once the informants' contexts were taken into account:

Like many *bogotanos*, Gloria wished traffic in the city would be better. This particularly bothered her whenever she had to visit her EPS POP, which at least was every three months, just to receive a printed medical authorization to get a medical exam and thus keep track of her chronic illness. She could not understand why she could not receive this authorization by email or why a single printed one could not apply for several exams. She told the research team it took her almost two hours to get to the nearest POP medical center, and that the transportation cost represented a huge impact in her household's economy. She confessed they were even thinking about getting a car, which until now had never been one of their priorities.

Pablo's stories also show the importance of taking context into account. Due to a difficult childhood, Pablo is illiterate. Although he was able to get a job as a teenager and paid for reading classes, he admits he still has trouble understanding what he reads. When Pablo was diagnosed with diabetes, his doctor gave him many brochures and other information, so he could understand more about this condition and what he could eat. After he gave him the information, the doctor gestured he could leave and Pablo felt too uncomfortable telling him he was unable to fully read.

These examples show that thinking beyond healthcare did not imply POP had to address all of its users' needs. They do however show that POP should consider contexts to make its healthcare services more compatible with users' socioeconomic conditions. This meant the company had to take an active and reactive stance rather than conceiving itself as a passive actor trapped in an unalterable context.

Additionally, by becoming context-sensitive, POP understood the importance of intervening different contexts differently. For example, user needs in Bogotá, Colombia's capital, were quite different to those of users in Medellín. Since Bogotá is a larger city, users were more affected by distance and traffic, implying POP had to make its services more accessible so users would not have to cover long distances. Thus, a higher degree of fragmentation was suggested for this city. Meanwhile, users in Medellín did not seem to have this problem, but most of the informants did complain about the waiting time before they were able to see a doctor. So rather than fragmentation, the issue here was increasing agility.

The users' contexts turned out to be a significant area of opportunity that would have been left out were it not for the informants' stories. When informants were asked to narrate an experience with POP, they did not limit themselves to the moment of interaction. Instead they mentioned other details that at first seemed irrelevant but really were not, such as the time they had woken up that day, or how long it took them to get there, or the money they were (or were not) prepared to spend that day. By listening and considering their contexts as valuable evidence, the team ensured POP was on the right track to becoming a user-centered company.

User stories question how quality of service is defined and measured

When it comes to the evidence that decision makers use to improve care, statistics and quantitative data tend to be more relevant—or located higher in the hierarchy of evidence—than stories and anecdotes. However this is starting to change. Throughout the past decade, researchers, policy makers, patient programs and other actors involved in healthcare have felt what some have called an overreliance on numbers. They have also challenged the "doctor knows best" paradigm, by giving similar or equal importance to the patient's perspective. All this has resulted in an increase in narrative research in healthcare, not as replacement but rather as complement to the scientific medical research that has historically been carried out. Regarding this case study, the research team could tell POP still had this overreliance on numbers, specifically when it came to understanding and measuring the quality of their services.

Ever since Colombia's healthcare system was reformed in the 1990s, the quality of the service provided by EPS, and whether it has improved or worsen, has been at the center of the discussion. While some experts suggest it is quite clear quality has decreased, and that this is due to private companies always attempting to optimize their profit and minimize costs, others argue quality is a highly subjective concept and very difficult to measure. In this context the research team saw an opportunity to position stories as powerful evidence when it came to measuring quality—and even more so when realizing the way POP was measuring it was insufficient:

POP understood there was a lot to be done to improve the service, but it was very proud with some of its indicators. For example, the EPS had data to show they were very efficient solving their users' questions and needs over their phone. However, after listening to stories such as Martina's, who tried to get in touch with POP during months until finally deciding to go to a medical center, the research team did exactly the opposite. When informing POP about these cases, the company was confused and argued their data showed they had an excellent service via phone. It turned out that the way they were measuring how good they were over the phone was very limited since they only took into account the duration of the calls. Average call duration was definitely not a good indicator, as they did not consider whether the user's issue had been resolved—or, like Martina, how many times the user had called before they actually answered.

With numbers and averages, negative experiences were being hidden by positive experiences, and little was being done to improve (or even listen to!) the negative ones.

These were just being regarded as data that was affecting the overall average, but not so much so as to inquire and intervene.

By listening to singular stories, POP was reminded of the importance of going beyond averages in a context as emotional and critical as healthcare. It was not easy for the client when the research team suggested singular stories revealed a much more pessimistic account of quality of service than the data they were accustomed too. The point here was not to completely replace how quality was being measured, but to use stories and qualitative evidence as valid input for quantitative measurement systems. It was about suggesting qualitative and quantitative information could and had to be integrated to understand the user experience.

IMPACT

A client transformed

The main impact of this project was the transformation of the client's mindset. Since the POP team had the opportunity to venture into the field with the research team, they listened first-hand to the user's stories—and proceeded to tell their own version of these stories.

The research team arrived one afternoon at Maria's house. She was a mother, so the team was quite excited to see how parents experienced healthcare for themselves, but also through their children. Seeing her daughter Valentina was a surprise; while an average 2-year old girl was expected, they met a fragile, tiny being, who could barely walk and looked more like an 8-month baby. Surprisingly, however, was not due to her development, but to the fact that this did not overshadow her capacity for happiness and excitement. She approached the team with the curiosity of a child and never stopped playing with the tape recorder and other interview materials. The team learned that Valentina suffers from a disease that affects her normal development and makes her bones extremely fragile: Congenital Osteogenesis Imperfecta, colloquially known as the "Brittle Bone Disease".

Throughout the interview, the team could not help but notice Maria and her mother's (Valentina's grandmother) tired faces. They told the team of the long journeys they had to make every day with Valentina, and how they tried their best to use the services they were offered, which were not adapted for a girl with this rare condition. Some of the more difficult situations they faced were physicians who generally treat people with more prevalent conditions and "normal" emergency situations, experts are scarce and do not usually provide services in government funded hospitals, long commutes, among other situations.

This story inevitably raised awareness of the importance of thinking about singularities, and this changed the mindset of doctors, engineers, administrators, and all the stakeholders that are accustomed to seeing members and patients from a numerical perspective. Numbers could reveal pains within the system, but these stories revealed the pains of existing individuals, whose expectation the system was unable to fulfill.

Since 8 people from POP were taken out of their daily duties and dedicated themselves 100% to this project, they were a fan of stories and qualitative research once they resumed these duties. So much they were even willing to confront colleagues (and even leaders!) that were still relying solely on numbers. Seeing the client have these discussions showed the research team the project had not been in vain.

A team transformed

One aspect that remained constant throughout the project was that the team's adaptability and resilience were constantly tested. This started out during the recruiting phase, when contacting users and patients proved extremely difficult. Recruitment is never easy in Colombia because individuals and communities have a hard time trusting strangers, but in this case users and patients were extremely resistant to talk about their health and their experiences in healthcare. So, the team decided to look for people with a similar accent to that of the users to gain trust and start fieldwork, since this cemented trust.

It is true recruiting and fieldwork were not completely smooth, but the real trouble began when the team sat down to analyze the collected information. After just a few sessions listening and re-listening to the collected stories, the researchers started to notice mapping a single user experience was not going to be possible. The project's objective, which at first had seemed achievable and even alluring, now seemed very far away, and this was due to the fact that users and patients were dealing with a system of services, not a single service. At this point the team considered fragmenting the user experience into many smaller experiences, each belonging to a specific service. However, there were far too many experiences for a qualitative study, and by doing this the team would be sacrificing the ability to understand the health system holistically. So, the team finally decided to map an experience based on the collected stories and on aspects that different services shared.

Once the experience was constructed and the project was completed, the client decided to work with a big data company, in order to map all of the experiences that the qualitative study had suggested. This is important because it shows how stories inspired the client to act on its own, and to use information that was qualitatively obtained as starting point to conduct further studies.

The project success, despite its complexity, was based on the team's adaptability, allowing experimentation to transcend methodologies, on its resilience so as not to yield to frustration in trials and errors though this experimentation and to learn at every step of the way and with the client's teamwork, achieving a synergy as a single team. Next, we will analyze the achievements and learnings that the team achieved thanks to this.

CONCLUSION

As ethnographers and designers working with business organizations, a challenge that we are constantly facing is proving the relevance of our research methods and results and explaining how this relevance is different to that of other types of evidence. In the service sector, decision makers want evidence that allows them to evaluate their organization internally and externally, which is why the key performance indicators (KPIs), satisfaction metrics and other quantitative customer experience indicators are permanently being implemented. But what happens when these organizations are operating in very humane and highly emotional contexts, such as healthcare? How can evidence obtained by qualitative means complement this need for constant evaluation, while generating additional value?

Throughout this case study it has been shown how a research team argued for the possibility of stories as valuable evidence in contexts such as healthcare, where organizations that are trying to provide broader and more efficient services should consider user stories as input for decision-making.

REFERENCES CITED

Barrera, Oscar David.
2014 Effects of the Law 100 of 1993 reform on coverage and health in Colombia. *Master's Thesis on Public Policy and Development, Paris School of Economics.*

Datos Abiertos. Gobierno digital Colombia
2017 "Número de afiliados por departamento, regimen y administradora para el año 2017" Accessed August 30, 2018. https://www.datos.gov.co/Salud-y-Protecci-n-Social/N-mero-de-afiliados-por-departamento-r-gimen-y-adm/7sbd-mgjf

Fiorencis, Alessandra.
2016 "The WHO Document on Narrative Research for Health Sector: Interview to Maria Giulia (Reviewer from WHO)." Medicina Narrativa.eu website, Sept 22. Accessed [July 20, 2018]. http://www.medicinanarrativa.eu/the-who-document-narrative-research-for-health-sector-interview-to-maria-giulia-marini-reviewer-for-who.

Forman, David & Sierra, Monica S.
2016 *Cancer in Central and South America: Introduction. International Journal of Cancer Epidemiology: S3-S10.*

Kalter, Steve.
2016 Using projectives to uncover "Aha moments" in qualitative research.In P. M. W. Hackett (Ed.), Researching social psychology. Qualitative research methods in consumer psychology: Ethnography and culture (pp. 131-146). New York, NY, US: Routledge/Taylor & Francis Group.

King, Helen
2014 "The art of medicine. Narratives of healing: a new approach to the past." *The Lancet* 393: 946-947.

Humanizing Quant and Scaling Qual to Drive Decision-Making

LAUREN MORRIS
Amazon Prime Video

REBECCA GATI
Amazon Prime Video

The Amazon Prime Video User Experience (UX) Research team endeavored to balance qualitative and quantitative insights and translate them into the currency that drives the business, specifically customer engagement, to improve decision-making. Researchers conducted foundational qualitative research to uncover what matters most to Prime Video customers, translated resulting insights into a set of durable, measurable customer outcomes, and developed a global, longitudinal online survey program that validated the importance and perception of these outcomes at scale. Researchers then systematically linked customers' attitudinal survey results to their usage patterns and overall satisfaction with the service. The resulting data showed how investing in improving a customer outcome is likely to increase service engagement, thus closing the loop between insights and business metrics for the first time. Prime Video executive leadership has not only embraced this integrated qualitative-quantitative system, but now uses it to prioritize projects on behalf of the customer. UX Research has shifted Prime Video culture from relying on analytics data alone and toward seeking comprehensive evidence to drive strategic decision-making.

INTRODUCTION

In the 2016 letter to Amazon shareholders, Jeff Bezos stated, "Good inventors and designers *deeply* understand their customers. … They study and understand many anecdotes rather than only the averages you'll find on surveys. … A remarkable customer experience starts with heart, intuition, curiosity, play, guts, taste." The concept of "anecdotal" evidence, or qualitative data, is embedded in Amazon culture—it is mentioned in everything from the company's guiding principles to training documents. However, when the company's initial streaming service launched in 2008, the team primarily used quantitative behavioral analytics (i.e. usage data) in decision-making and often used A/B testing for risk management before launch. After features shipped, the team used customer service contacts and app store reviews as a barometer for customer satisfaction.

As Prime Video grew, the team recognized the importance of expanding their understanding of the customer and subsequently invested more deeply in qualitative usability research and quantitative attitudinal research, primarily surveys, within the organization. Ultimately, UX Design hoped to evaluate design work more often and provide a method of measuring customer satisfaction in a way that could inform prioritization. However, the business did not have the expectation that research should provide strategic recommendations at the front end of the product development process. The greater organization recognized qualitative data included customer service contacts, app store reviews, and survey responses, and were intrigued by incoming usability test findings; however, they did not fully leverage results from these sources and questioned the validity of small sample size data, particularly when the findings did not match commonly held organizational beliefs. There was still a comfort and familiarity with extremely large sample sizes—e.g., A/B testing with millions of customers—that prevented teams from acting on the UX Research team's qualitative insights:

> "Amazon overall—not just [Prime] Video—has been really good at the intersection of behavioral and quantitative [methods]. This is something we actually pioneered in the industry—the use of extensive A/B tests to make determinations. ... It has been established as a scientific method with rigor and we have relied on that approach. That has been pretty well established for a really long time." – A Director of Engineering, Amazon Prime Video

Still, UX leadership saw the potential in investing in research to provide evidence that UX design quality and customers satisfaction are equally as important to address as moving key business metrics. At the time, product decisions *were* made on behalf of the customer, but from the perspective of increasing the number of customers and broadening content selection.

> "If you would have asked the folks at the time, 'How are you applying customer obsession?' They likely would have answered 'Well, for a customer that does not have access to our service on the device that they prefer, the best thing we can do for that customer is give them access to our service. For a customer who is in a market which we do not service today, the best thing we can do for that customer is exist in that market and give them access to our service'… So, those were the levers that seemed to have the largest customer impact and largest business impact at the time." – Principal Product Manager, Amazon Prime Video

UX leadership was hoping to inform product recommendations in a more holistic manner; the team believed that they could grow the impact of research at Prime Video and influence product stakeholders to focus on customer experience (CX) improvements with a diversely-skilled research group. The team hired two additional senior researchers—experts in qualitative user research and survey science—and they began gathering focused insights in their respective disciplines.

As the team grew, UX Researchers continued to collect and add to the existing wealth of independent insights and would occasionally come together to employ more traditional triangulation methods. They would, for example, examine the findings of a qualitative study and cross-reference with a similar quantitative study to validate results with the hopes of increasing confidence in results found in both studies. Researchers included inputs from other teams such as business analytics and market research to tell a more complete story. Converging inputs helped provide some recommendations to the business, but these attempts ultimately fell short to impact strategic business decisions. Studies were not always designed to have analogous objectives and operated on different roadmap timelines with different stakeholders; thus, the piecemeal nature made it difficult to compare results to create a comprehensive recommendation. For example, when examining how often customers clicked on certain elements on the website, analytics data showed heavy usage, leading stakeholders to believe this element was effective. There were also app store reviews that described a difficult UX and pointed to certain negative experiences but lacked detailed explanations. Additionally, a later qualitative study provided evidence that there was customer confusion around these elements, contributing to this heavy usage. Because these studies occurred on different timelines and were owned by different teams, the organization could not take clear action.

These triangulation attempts helped develop incrementally better solutions that statistically increased customers' usage of the service, but it was unknown if these solutions had a broader impact on the holistic experience or customers' overall perception of the service. The UX Research team was armed with qualitative insights about customer needs and pain points with the potential to spur exciting new ideas and concepts, but often asked themselves: "how could we quantify, more comprehensively measure, and prioritize what matters most to customers?" The research team recognized that data triangulation was not enough and sought out partnership with the analytics team. Together they aimed to more directly link all the inputs available, not only to gain a better understanding of Prime Video customers, but to help the business understand how to measure and act on insights and to influence strategic business investments that would drive innovation.

To develop a systematic approach for measuring and scaling insights, the team realized they needed to be more integrated in their research approach, enabling direct mixed-method research across qualitative and quantitative methods. The team sought to learn about each other's disciplines, design studies collaboratively, conduct fieldwork jointly, review data simultaneously, and leverage partnerships that each discipline unlocks. Direct participation in fieldwork provided crucial nuance that helped research scientists interpret quantitative data. Similarly, qualitative researchers leveraged quantitative insights to understand which patterns generalize to customers and scaled the impact of their findings, mitigating uncertainty of this method. By integrating qualitative and quantitative research disciplines and dedicating the roadmap to mixed-method research, the team was set up to build an integrated system that capitalized on each other's expertise. With a new team approach and new partners on board, they began the journey to answer what matters most to customers and which investments to prioritize. They quickly realized they needed a new, directly integrated method to do it.

METHODOLOGY

Determining What Matters Most to Customers

To determine what matters most to customers, the researchers aggregated existing knowledge from three years of generative and evaluative qualitative research, which included data from in-home interviews, longitudinal diary studies, usability studies, and customer service contacts. They mined for pain points, explicit and latent needs related to video services generally, and brand attitudes and perceptions. There were clear, consistent patterns across these various sources—enough for the team to begin confidently hypothesizing what matters most to current and potential Prime Video customers.

Initially, the team developed a framework that illustrated the hierarchy of customer needs for a video service, inspired by Maslow's Hierarchy of Needs theory (1954). Maslow's theory states that humans are governed by a hierarchy of needs; once their essential needs are met, they are motivated to pursue more advanced needs, but are unable to value those advanced needs until essential needs are met. The team found through their existing research that a similar paradigm existed for customers and had a strong hypothesis for which needs were essential and which were advanced. Maslow's theory is illustrated as a pyramid, which the team also leveraged for their framework.

In this initial framework, the base of the pyramid was named "Connection" which housed the customer need of "Awareness"; in other words, customers need to know the service exists. Once a customer is aware of the service, they can move through the next levels of the pyramid, as long as their needs at every level are adequately met. The top of the pyramid was named "Affinity," where the video service solves for needs that have not yet surfaced for customers—customers feel taken care of. At this phase, customers become loyal to the service and share their love for it with others. The goal is for the customer to reach "Affinity." For example, when customers become loyal and enthusiastically recommend Prime Video to others, the business can redirect resources toward further delighting customers.

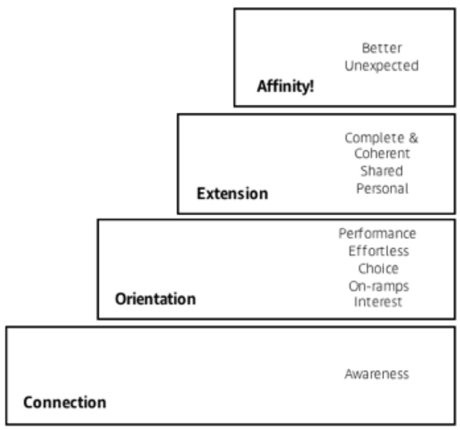


Figure 1. Prime Video's Hierarchy of Customer Needs Framework

The goal of this framework was to comprehensively and systematically capture and surface existing customer pain points, leveraging years of qualitative insights. The team translated these pain points into themes of customer needs, enabling teams to understand the breadth of customer needs that exist and consider this breadth when developing solutions in their domain. The team asserted the hierarchy of these themes to help the business prioritize investments, focusing on foundational customer needs first.

> "It was obvious that you would start at the bottom of the hierarchy. You wouldn't invest in solving some affinity problem if you couldn't even use the service, which was down in the basic level." — Director of UX Design, Amazon Prime Video

After collecting feedback internally, the research team uncovered three challenges with the framework: 1) utility—stakeholders, including members of the UX design team, did not understand how to leverage the framework because it did not provide specific recommendations, 2) theoretical basis—there was no quantitative validation of this hierarchy and Prime Video had not historically incorporated psychological theory-based frameworks into business decisions, and 3) ambiguity—the themes in the framework were scoped at a high level and thus open to interpretation, therefore stakeholders did not share a collective understanding of the themes. This limited the ability to measure whether or not customer needs were being met at each level.

"While the [pyramid framework] was interesting for creating a way of thinking about and conceptually prioritizing customer problems, it wasn't necessarily useful in terms of changing behavior – actually getting things done. We had the big band of things that were foundational, but which one of those was the most important thing to go focus on? We didn't have that yet." — Director of UX Design, Amazon Prime Video

Hoping to address these limitations, the team explored other effective, actionable frameworks that capture customer needs. A senior team member had applied and translated the "jobs-to-be-done" theory in multiple domains and recognized this could be a good fit for Prime Video culture. This theory was first articulated by Clayton Christensen (2003, 2016), and built upon by Anthony Ulwick (2005) and others. This theory asserts that people, in effect, "hire" products to do important jobs, and that they will "fire" a product if an alternative solution does these jobs better. The key to delivering successful products is identifying the jobs that are most important to customers and developing innovative solutions that do these jobs better than alternatives. The team set out to develop a more systematic approach that would allow them to not only build a framework of what matters most to customers with their insights, but to also quantify the magnitude and generalizability of these insights.

The research team began by engaging in multi-day working sessions over several weeks to develop a new approach. They built off of their existing Hierarchy of Customer Needs framework and applied the "jobs-to-be-done" theory lens, while also recognizing there were certain attributes of a framework that would be most effective (e.g., being able to appropriately scope customer needs, measure validity and generalizability of customer needs at scale, and ensure the future capability to provide recommendations). The result of these working sessions was the development of a set of "Customer Experience Outcomes" (CXOs) to articulate customer needs in an actionable, measurable way. A CXO captures a discrete customer need, derived from research insights, as 1) a "job" a product must do to address the need and 2) the criteria customers use to judge how well a product does the job. The anatomy of these CXOs, adapted from Ulwick (2005), included a *job statement* (e.g., "Protect my family from unwanted content") and a *success criterion* (e.g., "…with more confidence"), which enabled the team to measure how well the service does the job for the customer—both quantitatively and qualitatively.

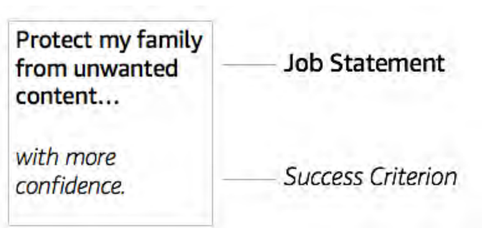


Figure 2. Anatomy of a Customer Experience Outcome (CXO)

By the end of this exercise, the research team had identified about 40 unique CXOs, loosely related to each section of the user journey that the organization was familiar with. The CXOs were all scoped at a granular level; the job statements contained only one verb that represented the smallest meaningful, but actionable, unit of customer value. Each one was measurable; they provided the basis for measuring the experience holistically. They intentionally had no mention of implementation or a solution, as they were meant to describe durable customer problems.

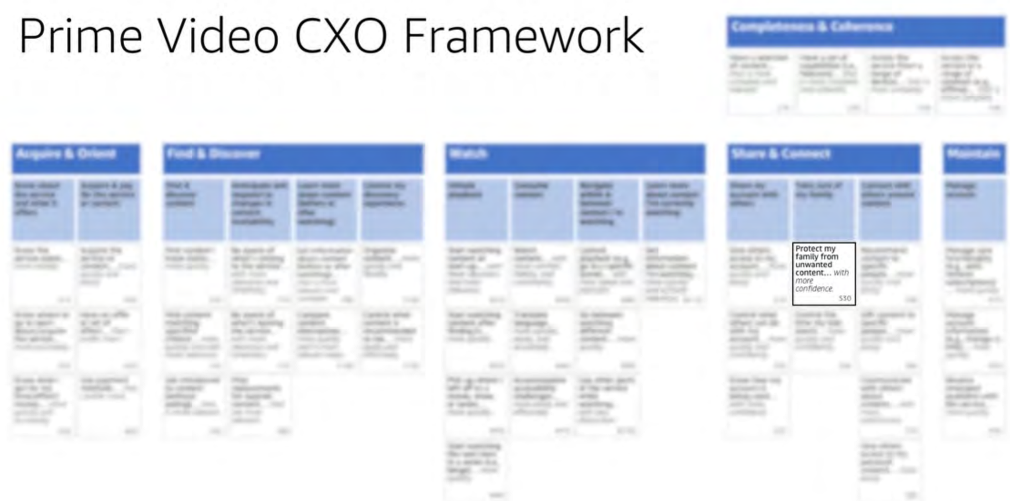


Figure 3. Prime Video's Customer Experience Outcome Framework

With the evolution of this framework, stakeholders were able to understand how to leverage it to make more customer-informed decisions and began to assess if their product features or designs did or did not map to these customer needs.

> "What started to change, in my mind, was when we started turning anecdotes, insights, survey results, whatever, into frameworks…it certainly resonated with me because I'm a strong believer in mechanisms. Mechanisms have an enduring power. They don't depend solely on one person's ability to evangelize them. If they make sense, they enable people to pick up those mechanisms and run with them, and essentially scale the impact of that mechanism." — Principal Product Manager, Amazon Prime Video

Validating What Matters Most to Customers

Although confident in the aggregated qualitative insights, the team believed it was necessary to validate these CXOs with customers and identify any missing pieces. The qualitative researchers designed a field study to validate the CXOs, using a semi-structured interview approach. Participants in two demographically-disparate cities were recruited for 2.5 hour-long in-home interviews. CXOs were written on cards and altered slightly to reflect customer-facing language rather than more technical job statements, although the team intentionally made sure the internal CXO framework language did not veer too far from an actual customer's voice. For example, the CXO "Protect my family from unwanted content…with more confidence" was changed to "Assure my kids have a safe and positive experience" in the interview. Much of the discussion focused on deeply understanding why certain CXOs were more important to participants than others.

Following a brief, unstructured interview consisting of a walk-through of the participant's home and their devices to ground and contextualize the discussion, participants were presented with the CXO cards. They were encouraged to read the CXOs, confirm whether or not they were relevant, and even suggested ways to clarify the CXO phrasing. They were also provided with blank cards to add any CXOs they judged to be missing. This helped ensure that the CXO framework comprehensively represented what mattered most to customers.

Participants were then asked to sort the CXOs into three roughly-equal groups: "most important," "next-most important," and "least important." It was key for the groups to be roughly equal because it forced the participant to make trade-offs and think critically about the most important aspects of a video service. From their "most important" group, the participant was asked to choose her or his top five CXOs. Again, the researchers observed the trade-offs that the participant made. While the goal of this exercise was to understand *why* certain CXOs were more important than others, the team's objective was to identify patterns across all sixteen participants' top five CXOs during analysis.

After the sorting exercises, participants were asked to comment on their satisfaction with all of their video services for their top five CXOs. A hand-drawn spectrum was provided to the participants for this exercise. The ends of the spectrum were defined by a question that participants had answered at the beginning of the interview—"Which brand/service/company have you had an amazing experience with, and which brand/service/company have you had an awful experience with?" Examples of amazing experiences included Taylor Guitars, Coach purses, and Nike running products. A notably high number of participants mentioned companies with poor customer service as their most awful experiences. The team did not limit participants' thinking to just video streaming companies, as the goal was to encourage participants to reflect on their emotional associations with products and services.

Using this spectrum, the participant set their own bar for an exceptional experience and assessed both Prime Video and other video services according to this bar. For example, if one of their top five CXOs was "Protect my family from unwanted content," they were asked "How good of a job does [service] do at assuring you that your kids are having a safe and positive experience?" The participant would then place each service somewhere on the spectrum. This process often elicited a visceral reaction from participants and helped the research team better understand how Prime Video compares to competitors on the things that matter most to customers.

After the field study, the entire team engaged in data synthesis. Every member of the team—including research scientists—watched recordings of multiple sessions and coded the interviews with the CXOs in mind. During synthesis, one member of the team was responsible for creating video clips that aligned to early themes and patterns as the team debriefed and discussed. Ultimately, synthesis of this research led the team to formalize a set of key themes, each of those tying back to specific CXOs. Now the team could illustrate the linkage between a qualitative insight and its corresponding measurable, actionable CXO.

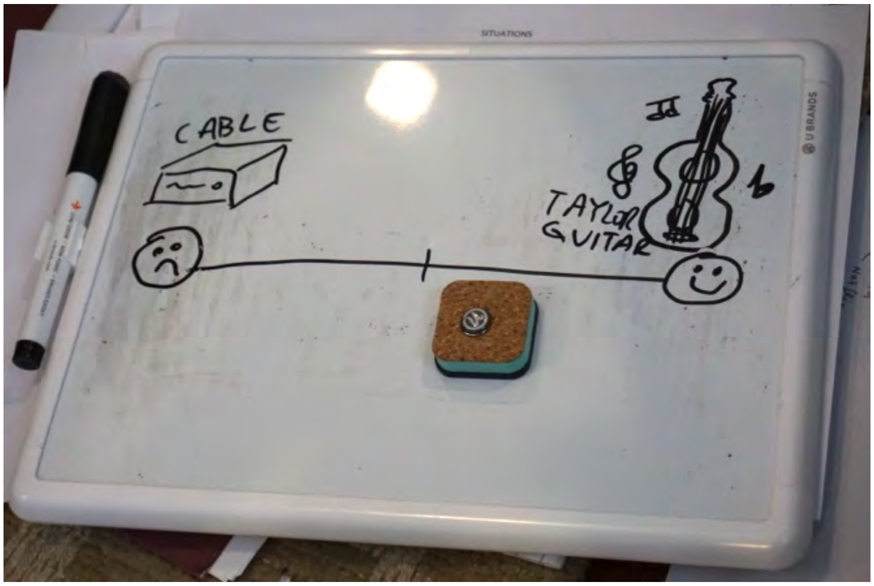


Figure 4. Example of a participant-defined experience spectrum. In this example, the participant drew their "worst experience," illustrated on the left, and their "best experience," illustrated on the right.

Influencing the Organization

The field study presented the first opportunity to fully integrate the qualitative and quantitative arms of the UX research team. The research scientists joined qualitative researchers in field, operating camera equipment and taking notes. Through their participation, they gained perspective that later directly influenced both quantitative study design and data interpretation. Additionally, the in-home interviews were a great opportunity to evangelize the value of applying ethnographic methods. The team invited senior leadership from a variety of disciplines into the field as notetakers, helping them build empathy with customers and gain a better understanding of field study as a valid research method. One senior leader later reflected on his appreciation of the experience, as he was reminded that customers live a completely different life than him. By participating, senior leaders were also reminded that customers' familiarity with technology and awareness of features are generally lower than the average Prime Video employee and are ultimately rooted in the context of their lives. At the conclusion of this study, leaders strongly encouraged their staff to become more involved in qualitative research opportunities.

Figure 5. A Director of Engineering taking notes in a participant's home

The team believed it was extremely important to curate and share high-production videos from this fieldwork to build empathy for customers and help stakeholders to challenge their assumptions about the customer. With Amazon being a heavy written-document culture, there was concern that sharing videos to stakeholders would not be effective or understandable. The result was a combination of the two approaches: 1) an immersive insights report, separated into sections for each of the key themes, with the integration of supplemental quantitative data to validate the insight from the research, and 2) five-minute customer videos correlating to the key themes. The team scheduled recurring customer insights immersion sessions where stakeholders from various teams were invited to watch these customer videos, with the UX Researchers in attendance to provide any necessary additional context. The researchers believed that these immersion sessions, specifically hearing directly from customers, would be a refreshing change for stakeholder attendees. It was also an opportunity to convey customer sentiment and emotion, which was a more effective tactic than having the researchers relay anecdotes from the field themselves. Session attendees were able to ask the research team questions about ethnographic methods, the study design, or the insights. The team followed up by sending the insights report to all session attendees, adhering to the Amazon written-document standard. The videos were embedded in the insights report, giving viewers a multi-faceted way to better understand their customers' needs and opportunities for innovation.

The outcome of this approach was the beginning of a culture shift. Leadership was intrigued and began requesting to attend future field studies. Key stakeholders left these workshops inspired by the opportunities to innovate. Teams that had never engaged with the research team in the past were now reaching out about how to work together. One of the researchers was even invited to speak at a Prime Video "all-hands" meeting, exposing the discipline of UX Research to thousands of employees. While some uncertainty of how to use qualitative data still existed, the team felt that ethnographic methods were gaining momentum.

> "It had the additional benefit for people like me. To put me face-to-face with a customer has a certain impact value that you don't really figure out looking at any number of metrics in a conference room. You just don't get that same feel for things. … It's a cool area that we work on, in terms of how directly it impacts consumers, and this is a very in-your-face reminder of that in a helpful way." — A Director of Engineering, Amazon Prime Video

The team acknowledged that in order to influence business strategy and organizational priorities, qualitative data must be validated at scale and linked to the currency that drives the business—customer engagement.

Scaling Qualitative Insights

After completing the qualitative study and working sessions, the research team needed to determine if findings generalized to the larger customer base. Historically, this is where the quantitative and qualitative arms of the research team would diverge and later triangulate data. To keep the research tightly aligned, they used a common currency—the CXOs. Using field study insights and collaborating closely with the qualitative researchers, the research scientists designed a cross-sectional, longitudinal survey program and disseminated the survey to major markets around the world. Its aim was to measure how well Prime Video does on a subset of CXOs, selected based on the qualitative ranking exercises, aggregated insights, and the project priorities for the year.

Developing this survey program was challenging, but necessary to scale qualitative findings. Survey research was still considered qualitative, low-sample size research by some stakeholders, as they were most familiar with reviewing usage data and multivariate analyses using several million "respondents" versus several thousand used in survey research. In response, research scientists conducted multiple educational seminars providing details around the scientific process of survey design, sampling, and analysis, which helped quell apprehension and laid the foundation to use this method for decision-making. It also served as a mechanism to contextualize, and ultimately humanize, usage data the business relied heavily on by providing a window into the attitudes and perceptions of those using Prime Video.

Some stakeholders were also convinced there was "more art than science" in designing surveys, which prompted the team to qualitatively test the survey with customers to ensure questions were interpreted as intended. This went beyond standard quality assurance measures the quantitative team typically used. They dug into both how questions were interpreted but also why participants selected their responses. This allowed stakeholders to hear directly from customers about how they were interpreting the questions and showed stakeholders the survey instrument was reliable, which ultimately increased their confidence. This was another instance of the qualitative and quantitative research experts working in tandem to help quell concerns about methodological rigor from stakeholders as well as generate empathy by hearing directly from customers.

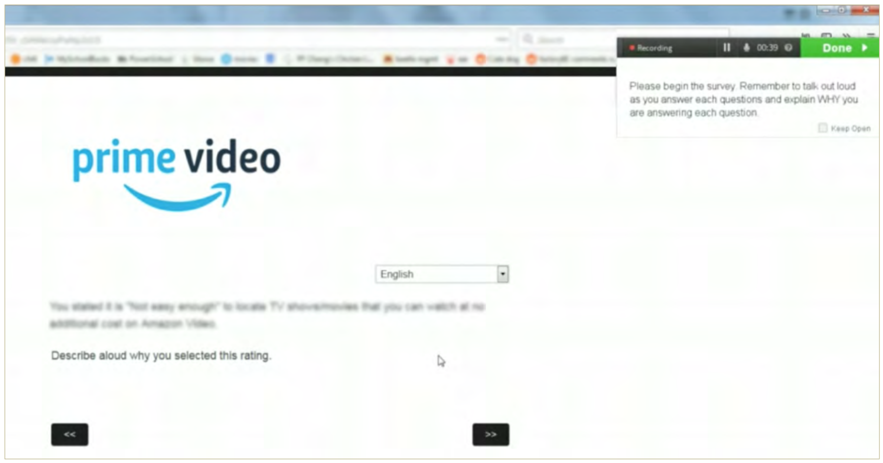


Figure 6. Screenshot of a participant using usertesting.com, the platform the team used to collect audio and video feedback from customers taking the survey. Participants were asked to talk aloud as they completed the survey. This helped the team understand how customers interpreted the survey questions and why they chose their responses.

There was also debate about which CXOs to prioritize in the survey instrument. There are always competing priorities on the product roadmap and product stakeholders wanted to ensure that their planned feature shipments would be covered. However, to maintain survey length best practices, informed by Dillman (2007), the research team had to assert that only around twenty-five CXOs would be covered. Although this left some stakeholders resistant to adoption, it forced teams to recognize that researchers would prioritize what to measure based on evidence of what mattered to customers instead of proposed product solutions.

The team also experimented with two waves of data before publishing to the organization, the first using a standard bipolar 5-point satisfaction scale (very satisfied to very dissatisfied) (Singleton & Straits, 2010) to assess CXO satisfaction and the second using a less traditional, categorical 3-point scale to assess performance of Prime Video on the same CXOs (very good, somewhat good, not good enough). Because the team piloted the program with both scales, research scientists were able to illustrate the similarity in results, ensure the robustness of the scale, and showcase the primary benefit of using the 3-point scale: persuasion. By requiring respondents to rate more extreme categories, it was much easier to interpret and determine whether the business should invest in an aspect of Prime Video when, for example, 20% customers stated it was "not good enough" vs. 4 on a satisfaction scale. The team wanted to reduce debate during results interpretation and encourage the business to act. Having polarizing categorical responses helped achieve this goal.

Through navigating these challenges, the research team showcased their expertise, helped stakeholders understand their processes, and ultimately elevated the opportunity to use survey data in decision-making. By bringing stakeholders in early and allowing them to challenge the teams' methods, adoption and evangelism was easier and not the sole responsibility of the research team.

After the data was available, the team was set up to integrate the survey dataset with behavioral usage data to help determine the relationship between satisfaction and engagement. This had been an ongoing topic of conversation amongst the research team and analytics partners. Both teams believed that this would be the most direct way to measure how customers were feeling about the service and if that had an impact on what they were doing on the service. Because the business focused so heavily on driving engagement, understanding the relationship of satisfaction to engagement would also be the most effective way to illustrate the importance of measuring and driving satisfaction, and that the two combined could determine which investments the business should make. However, complex new data infrastructure had to be built in Prime Video to enable this investigation and create a semi-automated process for ongoing measurement.

Creating New Systems to Support Integrated Data

In parallel to the survey program development, the research scientists knew they would need to work with partners outside the research team to develop a new system that could support both survey data and behavioral usage data. They would need to secure several months of technical resources which would overlap with the development and launch of major product releases. The research team had already garnered interest from the analytics team and engaged with them to discuss the details of leveraging each other's expertise to develop an integrated data system; they realized they would still needed engineering resources to develop the infrastructure. In conversation with engineering stakeholders, the initial response was that while the project was interesting, it would have to compete against other established priorities for the year. They continued to pitch the project, laying out their vision and the potential for the result. An engineering executive agreed to resource the project. With this support, the research scientists, analysts, and engineers were set to tackle this infrastructure challenge.

Since customer usage data was already being tracked for other reporting use cases, it made sense to try to ingest survey data into existing data warehouse structures. However, survey data is inherently flexible, semantic, and variable over time, while database structures are inherently inflexible, requiring rigid rules and consistency over time. The project team brainstormed for weeks and were ultimately able to create an adaptable system. The data engineers were able to extract and manipulate the new data, ultimately surfacing it in dashboards for the team to more easily work with. Connecting these data types was a huge victory, as this connection had never been done before in Prime Video. It enabled the team, for the first time, to begin investigating how the perception of Prime Video affected customers' usage of the service.

> "Before working with UX Research, our team would analyze data, hypothesize about root causes, and use our best judgment to describe results … but we never really knew if our judgment was correct. When we added data about what customers said, we no longer had to guess … we knew why there were differences in our datasets. This fundamentally changed the way we analyze and value data" — Senior Analytics Manager, Amazon Prime Video

Bringing Data Together

Once the data was connected, the team created a glanceable scorecard that illuminated areas that the business was doing well in and other areas that needed additional investment. There was so much new data that the team wanted to ensure stakeholders could quickly understand what mattered most to customers. To start, they displayed "live" satisfaction scores (versus a traditional survey report) housed within the CXO framework. This reinforced the CXO framework as a mechanism to highlight comprehensive customer needs, now backed by quantitative data. This dashboard was also built to be interactive, including features like survey question and response distribution details on hover, clickable CXO boxes that revealed the underlying data, and calculations that automatically coded the satisfaction red to green based on the responses from the latest survey. This was a departure from the standard published dashboards that primarily included rows of data, filters, and occasional annotations.

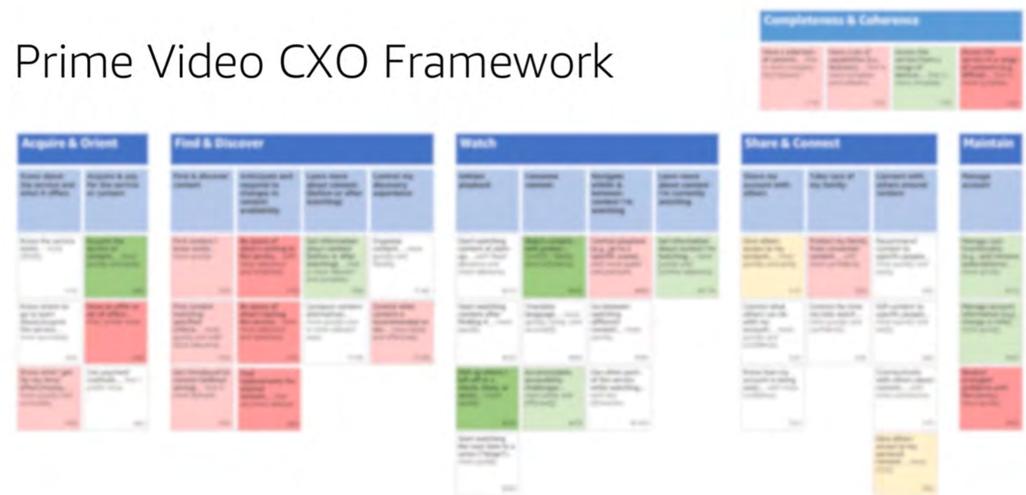


Figure 7. Prime Video Customer Experience Outcome Framework dashboard, including sample satisfaction ratings.

Because survey responses were linked to usage data, the team began investigating responses alongside corresponding engagement data, ultimately finding that customers who were unhappy with the service generally used the service less. They also pinpointed which aspect of the service, or which CXOs, correlated mostly strongly with engagement. It was the first time Prime Video had quantitatively-backed evidence that satisfaction was positively correlated with engagement, and the first time patterns of satisfaction and engagement could be examined together. The team could start to prioritize customer needs based on these correlations. It set the foundational understanding that satisfaction is equally important to measure and use in decision-making, and that integrating data could provide a framework for prioritization.

In order to make best use of the connected dataset, the team built several calculations to determine the "opportunity to influence overall service satisfaction" and the "opportunity to influence engagement." These calculations quantified the gap between satisfied customers and unsatisfied customers, giving each CXO a score. The result was an ordered list of CXOs based on 1) satisfaction with that CXO, 2) overall satisfaction with the service, and 3) overall engagement with the service, enabling ranking of CXOs across any of these dimensions.

To reduce complexity, the team decided to categorize CXOs and assign recommended actions based on uncovered patterns so that partners could more easily consume and act on this data. For example, an area that had low CXO satisfaction, a high correlation with overall service satisfaction, and a high correlation with service usage would be an area to *prioritize*. This pattern indicates that customers are generally unhappy with Prime Video performance on this CXO, have a relatively negative perception of the service, and those same customers use the service less. Alternatively, some CXOs had high satisfaction scores and high correlations with service satisfaction and usage. This pattern indicates that customers are happy with this aspect of the service, but if that CXO satisfaction score drops it will likely result in a decrease in service usage and overall satisfaction. This is an area to *fortify*.

Perhaps one of the more interesting patterns that could not have been uncovered examining satisfaction and engagement data independently was when a CXO had low satisfaction ratings and low correlations with service usage and overall satisfaction. This indicates that customers are unhappy with an area of the service, but it did not appear to influence their usage of the service. Puzzled, the research scientists consulted the qualitative researchers who had a bank of knowledge from the CXO validation fieldwork that included customers commenting on these particular aspects of Prime Video, and how no other service did a better job on this CXO. Customers had nowhere else to turn for a better alternative, so they were going to continue to use Prime Video (until something better came along). Keeping the two arms of the UX Research team in lockstep during development enabled the research scientists to interpret their data with fewer assumptions and allowed the qualitative researchers to use the aligned quantitative evidence to boost the impact of their findings. CXOs with this data pattern became known as *prospects*, or areas of opportunity for Prime Video to tackle to better compete and better serve unmet customer needs. These three categories became the primary recommended actions the team coined and supplied to the business as areas to invest in. With these recommendations automatically categorized and available in interactive dashboards, executives' prioritization questions were answered, truly, at a glance.

Challenges Integrating Data

Several challenges arose while developing this tool, many of which are still being iterated on. It was, and continues to be, difficult to distill multiple complex calculations into a glanceable scorecard that is useful to a range of stakeholders, including executive leadership, product managers, designers, marketers, and researchers. Most organizations request this type of scorecard to determine the health of their businesses; however, few are able to achieve a deliverable that is both comprehensive and usable. The quantitative team, guided by the qualitative team, usability tested this dashboard system with several stakeholders. They iterated on the tool's design multiple times, balancing methodological transparency with glanceability. Another challenge the team faced was training the organization on this tool. It required a dedicated communication strategy, with workshops and insights share-outs for more than 25 teams as well as the entire organization. After months of training, the organization now regularly uses the tool and its insights across a range of projects and job functions. It has also garnered interest from other Amazon businesses. Research scientists regularly consult with researchers, marketers, and product managers on how to build a similar system relevant to their respective businesses.

IMPACT

Since this integrated qualitative-quantitative program has come online, it has become a key input into the Prime Video decision-making process at all levels of the organization. The Prime Video roadmap is closely examined to understand which CXOs a given project influences. The quantity and priority of CXOs influenced determines how much support a project receives. Strategic goals have been set to improve satisfaction with particular CXOs, which are regularly reviewed with leadership. The CXO dashboard has the some of the highest views of all dashboards across the organization. Having CXO goals for each project helps all team members—from design, to product, to engineering—understand and stay aligned on the customer needs being addressed.

In addition to providing consistent updates to executive leadership, the UX Research team has concentrated on helping stakeholders think more holistically and more long-term about the customer experience—focusing on both satisfaction and engagement with the service. The team closely partners with product stakeholders to help them determine how their area of focus or specific feature compares to all other aspects of the service that customers care about. Engagement reporting meetings now include dedicated time to highlight satisfaction metrics because of the recognized relationship between business and attitudinal metrics. Additionally, they help product teams realize how introducing a new feature or changing one aspect of the service can come at the expense of negatively impacting a CXO.

> "[Skepticism from others was mitigated by] a) the apparent level of rigor that seems to have gone into the articulation of that framework and b) it's undeniable that linking quantitative data to qualitative data was a big bridge to lead people over … it's likely that no product manager in Prime Video has ever had the luxury of having the time or the resources in place to say 'I'm going to look at the entirety of the experience and consider all the needs our customers have AND I'm going to make an attempt at getting a sense at how good or bad we're doing at these needs AND I'm going to try to see what the business opportunities are associated with each of these needs. That then becomes an overwhelming amount of evidence, and something that doesn't seem like it could be quickly replicated by anyone on the team … linking the attitudinal to behavioral data becomes more of an exercise in prioritization." — Principal Product Manager, Amazon Prime Video

From a research methods perspective, one of the most valuable aspects of this integrated data system is the ability to "close the loop" between behavioral usage data, survey data, and data from subsequent qualitative evaluative research. In usability testing, the qualitative team measures project-related CXOs using the same 3-point scale that the ongoing survey program uses, which enables direct comparison between data sources. Interestingly, when participants are asked to rate their satisfaction for key CXOs on this 3-point scale, the distribution of responses from usability testing are often similar to responses from 5,000-person surveys. This demonstrates the power of qualitative, smaller sample-sized data and increases the confidence stakeholders have in using this method.

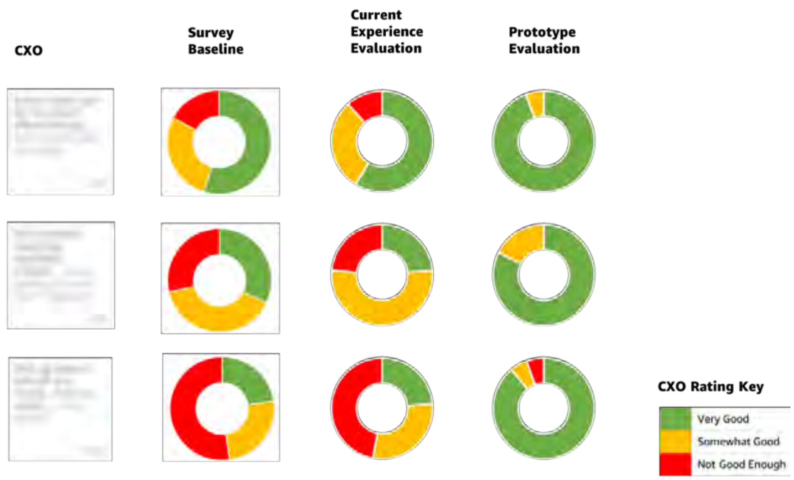


Figure 8. Sample comparison of survey, current experience evaluation, and prototype evaluation results. There are significant similarities between survey results and current experience evaluation, which increases confidence in qualitative findings.

As teams progress to using A/B and multivariate testing with greater sample sizes for their projects, results from usability testing have increased the confidence and decreased the cost of what the team includes in the tests. The team now provides recommendations for what should be included in multivariate or A/B testing based on confidence level and what is or is not well-suited for a usability lab setting. This has recently resulted in a multivariate test that had the highest increase in key business metrics in Prime Video history.

> "It's now clear that this has a level of rigor and thinking behind this that makes sense … this is a whole new way of approaching UX development and even product decisions … *That*, I feel, is a crucial inflection point. Now you have an approach that tells you what to prioritize in a data-driven way, which I think is going to be much less open to individual opinions than it had been before … Once you have an approach, doing user research to see if this actually helps that CXO improve … so you have high confidence before you launch that this is going to work because of the work that you've done up front, that is just completely new … I think that is a step function improvement in how we've gone about doing things." — A Director of Engineering, Amazon Prime Video

This is an ongoing program which directly informs what Prime Video prioritizes; the impact is continuously developing, and the team's knowledge is constantly evolving. However, there is clear evidence that with the integrated, closed-loop data system that the team has built, the organization is better equipped with the tools and insights they need to make confident, well-informed decisions on the customer's behalf. The customer's "voice" now has a seat at the table prioritizing roadmap investments. A senior leader summed up the research team's impact:

> "Seeing that we can build the case and hold our own using customer data, translated into business data … it's the first time I saw customer data at the business decision-making table and it breaking through. It basically bought its way onto the roadmap by being able to talk about it in the currency of the business." – Director of UX Design, Amazon Prime Video

Though this program has already been effective, in many ways it is still day one. The team is now embarking on extensions of this program and improving it as they learn more. Research scientists are working with other data scientists to explore how to capture customer satisfaction measures in multivariate tests. They are working with economists to determine the financial value of movement in satisfaction on key CXOs. Qualitative researchers are exploring ways to recruit participants who have responded to the survey in the past to test new concepts so they can compare their survey results to their evaluations of new concepts. The team is also continuing to qualitatively validate the CXO framework in new markets across the world—with recent studies completed in Germany and Japan—and spread the survey program into other key markets as well. These extensions help the UX Research team continue to expand their integrated, closed-loop system, and as more inputs are added to the system, the team will continue to expand how evidence is defined, balancing qualitative and quantitative inputs to drive decision-making for Prime Video.

NOTES

This work would not have been possible without the expertise of our colleagues Emily McGraw, Clemente Jones, Sarah Read, Kat Dellinger, and Bill Fulton; the critical and delightful partnership with analytical experts Vic Xiao and Steven Pyke and technical experts Sid Sriram and Jack Tomlinson; the impressive thought leadership and infinite guidance from Steve Herbst; and the continuous encouragement and enduring confidence from Alistair Hamilton.

Note that neither Amazon.com, Inc. nor any of its affiliates is a sponsor of this conference and this case study does not represent the official position of Amazon.com, Inc. or its affiliates.

All customer data collection methods follow standard best practices and is anonymized.

REFERENCES CITED

Christensen, Clayton M. and Michael E. Raynor
2003 The Innovator's Solution: Creating and Sustaining Successful Growth. Boston: Harvard Business School Publishing Corporation.

Ulwick, Anthony
2005 What customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services. New York: The McGraw-Hill Companies, Inc.

Christensen, Clayton M. and Taddy Hall, Karen Dillon, and David S. Duncan
2016 Competing Against Luck. New York: HarperCollins Publishers.

Dillman, Don A.
2007 Mail and Internet Surveys: The Tailored Design Method, 2007 Update with New Internet, Visual, and Mixed-Mode Guide. Hoboken: John Wiley & Sons, Inc.

Singleton, Royce A., Jr. and Bruce C. Straits
2010 Approaches to Social Research. New York: Oxford University Press.

Maslow, Abraham H.
1954 Motivation and Personality. New York: Harper & Row.

Below the Surface of the Data Lake: An Ethnographic Case Study on the Detrimental Effect of Big Data Path Dependency at a Theme Park

JACOB WACHMANN
ReD Associates

ANDREAS JUNI
ReD Associates

DAVE BAIOCCHI
ReD Associates

WILLIAM WELSER IV
ReD Associates

This case-study details how a team of anthropologists and a team of data scientists sought to help a Middle Eastern theme park make use of their big data platform to measure 'the good customer experience'. Ethnographic research within the theme park revealed that visitors yearned to bond with the other members of their group, as they rarely got the chance during their busy everyday lives back home. However, trying to build a measurement of how the theme park delivered on bonding – through the development of a 'bonding index' – turned out to be unfeasible, because the big data platform focused on capturing operational data. The decision to focus on operational data had unintentionally created a path dependency that made the big data setup unfit for answering some of the theme park's most fundamental questions. This is a problem ReD Associates has observed across clients and to solve it this paper suggests that companies start with an open-ended, ethnographic study of their big data needs before they build a big data platform. This will enable companies to be more strategic about their digitalization and thus maximize its impact.

THEME PARKS ARE AN ELDORADO FOR DATA SCIENTISTS

> "If you want to imagine how the world will look in just a few years (…) skip Silicon Valley and book a ticket to Orlando. Go to Disney World."
> (Wired, 2015)

Imagine a young girl called Liza going to Disney World for the first time in her life. She spots Pluto, her all-time favourite Disney character, and, as she walks toward him, he gets down on his knees, stretches out his arms readying them for a hug and calls out her name: "Liiiiza!". Pluto knows her name – but how could he?

The MagicBand Liza, and most other guests at Disney World Orland wear around their wrists enable the theme park to collect and make use of a wealth of data about their guests. Upon purchasing tickets, people will give up personal information like name, age, favourite character and credit card information. Inside the park this information can then be combined with geo-location data to provide someone like Pluto with the input he needs to create a special moment for kids like Liza. A truly magical Disney-experience – enabled by big data.

Pluto, and other characters calling out the actual names of kids (and possibly adult fans too) visiting Disney World is one use-case out of several imagined by the people behind the MagicBand. It is still in development, but families can already now be greeted by name before even opening their mouths when approaching a restaurant they've booked a table at. Another use-case, currently termed "The Story Engine", plans to combine geo-location data with the park's many video cameras (and possibly face-recognition software too) to create personalized videos for every single group visiting the park. Catching the candid moments and giving everyone a unique and shareable souvenir.

These moments of magic are enabled by a combination of a state-of-the-art big data setup, which reportedly cost Disney 1 billion USD to develop (Kuang: 2015), and a willingness to share personal data unparalleled by the outside world. This willingness to lay aside privacy concerns for a day of family-fun makes theme parks a unique fieldsite for both anthropologists and data scientists interested in what the future may hold at the intersection between big data and human experiences.

COMBINING THICK- & BIG DATA IN A THEME PARK

This case-study details the story of how another theme park, one placed in the Middle East, sought to utilize Big Data to improve its customer experience. The park, like Disney World, consisted of a closed space, where they owned all the restaurants (and their sales data), all the Wi-Fi routers (and their geo-location potential), all the rides (and their utilization data) and all the surveillance cameras – to name just a few data sources. To help improve the customer experience the park hired a team of anthropologists from the strategy consultancy ReD Associates in the summer of 2017. The team became a small part of a much larger ongoing project to build a Customer Data Platform (CDP). The theme parks executives hoped that the CDP could give them insights into how customers experienced their park and use it to guide strategic decisions going forward and to accurately measure the impact of new initiatives. The first roll-out of the CDP planned to collect data from 250 discrete data sources, which would then be stored in a Hadoop Data Lake. When the team of anthropologists joined the project 40 data scientists had already been on the ground at the theme park for three months, with a similar number working ad hoc remotely – primarily out of India. The anthropologists' involvement was set to last 6 weeks, whereas the first roll-out of the CDP was set to last a year followed by a support and adjustment phase.

At the onset of the project the division of labour for the collaboration between the anthropologists and data scientists was clear. The anthropologist would carry out an ethnographic study of the guests in order to identify what characterized a good user experience and the data scientists would then figure out how to measure that using big data analytics:

RESEARCH QUESTIONS

Anthropologists	*Data scientists*
What characterizes a good customer experience in the theme park?	How can the theme park measure and track whether it delivers a good customer experience?

The idea was thus, that after the anthropologists found out what mattered to visitors, the data scientists would measure how it was delivered on and track it over time as the park sought to improve it.

Conceptually, this this type of collaboration between the anthropologists' thick data (Geertz: 1973) and the data scientists' big data can be described as a 'Sounding Board Model'. The anthropologists job was to identify and "throw" insights at the CDP, which then, with the help of the data scientists, will return a quantified measure of the identified insight. One example, which the park executives and data scientists provided at the start of the project, illustrates how they imagined the collaboration:

- **Queuing time:** If the ethnographic research finds that queuing for rides is a major pain-point and something that is crucial to focus on in order to improve the customer experience, the CDP can be used to track queuing times and estimate the size of the problem. The effects of initiatives to alleviate the problem, such as e.g. guiding people towards less busy rides or planning shows during ride rush hours, can then be measured going forward.

This approach to integrating thick- and big data has the allure of seemingly being able to combine the two data types' biggest strength while simultaneously countering each other's biggest weakness, namely by adding scale to the thick data and depth to the big data (see Figure 1).

The Hadoop data lake, which was chosen as the type of big data setup for this theme park over the more traditional Data Warehouse setup, is well suited with the sounding board model of thick- and big data integration for two different reasons.

Firstly, a Hadoop data lake can store data at a fraction of the price compared to data warehouses. Whereas a terabyte of data stored in a data warehouse can cost \$250,000 a terabyte stored in a Hadoop data lake can cost \$2,500 – a reduction in price of 99% (PwC: 2014). This means that a larger 'sounding board' can be built with the same budget, thereby, in theory, making it possible to have enough data to quantify any insight the anthropologists might conjure up and throw at it. The drastic reduction in price can to a large extent be explained by the different data storage formats of data warehousing and data lakes. Whereas data warehousing requires a costly and time-consuming data integration and structuring up front, a data lake will store data in its native format – i.e. raw and unprocessed data, which can then be 'fished' out of the data lake when needed.

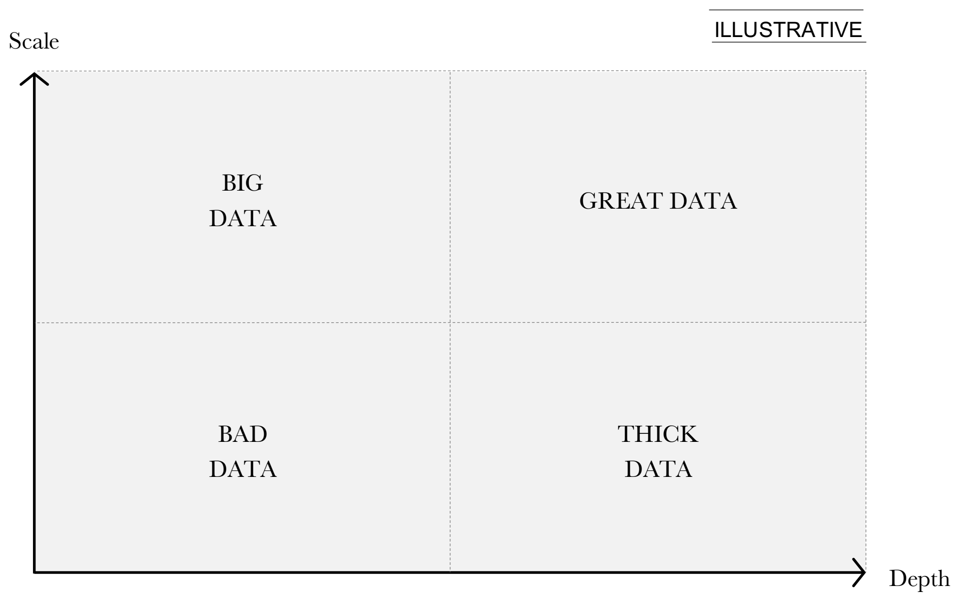


Figure 1: Illustrates how the combination of thick- and big data can theoretically produce a new type of data that contains both depth and scale.

Secondly, the fact that the data is stored in a native and unstructured format means it lends itself well to flexible and task-oriented structuring (PwC: 2014). The anthropologists' insights were thought to prompt this type of task-oriented structuring, which made the data lake setup well-suited for a sounding board type of collaboration. Or, to stay in the metaphor, the setup made it easy for the data scientists to go fish after the anthropologists have told them what to fish for.

ETHNOGRAPHIC RESEARCH APPROACH

The ethnographic fieldwork lasted just over two weeks during which the anthropologists would spend 1-2 days together with 12 different groups visiting the theme park. These groups were recruited to be approximately representative of the nationalities, age spread and group types – i.e. couples, families or groups of friends – that visit the theme park. A day of fieldwork would most often start with meeting the group visiting the park for breakfast and then stay with them until they arrived back at their hotel. The research would, when possible and appropriate, be documented using a dictaphone, a (waterproof) camera and a notebook. The anthropologist would go on rides with the groups or wait with part of the group as others went on rides, walk around the park with them, share meals with them, get lost in the park with them, and share motion sickness with them (especially the adults as they are more susceptible to this than kids and teenagers, see: CBS: 2015). The ethnographic research thus relied first and foremost on participant observation and on-the-go semi-structured interviews.

During fieldwork the team of anthropologists explored themes and research questions such as:

- What makes a good holiday?
- How does group dynamics impact the visit?
- What moods characterize peoples' visit to the theme park?
- How does a visit to a theme park fit into a larger holiday itinerary?
- What were the steps leading up to a visit to the theme park?
- What were people expectations and how did they match their actual experience?
- Which rides were peoples' favourite rides, and why?
- How do people share and recall their experience after their visit?

Following each day of fieldwork the researchers would write up their notes, upload them for the rest of the team to see and discuss.

THE SOCIAL STAKES ARE HIGH FOR HOLIDAYING

The analysis following fieldwork identified three key opportunity areas based on insights about what mattered most for improving the park's customer experience. The insights behind each of the three opportunity areas are sought summarized below in one sentence each:

- **GET CLOSER:** Guests want to bond within their social group when visiting the park.
- **GET BALANCE:** Guests want more variation in the mood-spaces offered within the park.
- **GET REAL:** Guests want deeper and more authentic experiences with the park's theme.

Of these three, the first opportunity area 'GET CLOSER' was found to be the most crucial and fundamental one to deliver on. For this reason, this was where the lion's share of the collaboration between the anthropologists and data scientists would focus. The remainder of this case-study will focus solely on this opportunity area.

During fieldwork, the ethnographers collected many stories of people who felt disconnected from the other people in their group during their everyday back home. Some examples include:

- Lorenzo (60, Italy): He lives in Italy and is divorced from the mother of his 7-year-old son, who moved to Canada with their son after the divorce. Lorenzo rarely gets to see his son and spend time with him, and as a consequence fears that he will lose contact with him as he grows older.
- Rodrigo (28, Philippines): Lives and works in a city two hours drive away from the city where his fiancé lives and works. About two hours of transportation separates them and they only get to see each other on weekends – unless they have a work shift during the weekend, which is relatively often the case.

- Molly (39, Denmark): Is the mother of two rollercoaster-loving tweens. She fears that this holiday might be the last where her kids really want to spend time with her before they grow up and become more independent. The last *real* family holiday, as she describes it.
- Robert (46, Belgium): Is a successful surgeon, divorcee and father of two teenage sons. He feels that work makes his everyday too busy to really bond with his sons, only leaving 5-10 minute intervals to interact on most days back home, which makes it hard for him to really feel connected to his sons and vice versa.

What these stories illustrate is that many people visiting the park has a need for bonding that is not satisfied in their everyday life, which they often described as fragmented. A description that resonates with Zygmunt Bauman's description of late modernity as liquid modernity (Bauman: 2000). Holidaying thus becomes a way for groups to seek out a social connection with their loved ones, which they struggle to feel in the everyday. And the stakes are very high. A study by two sociologists, associate professor Julie Brines and doctoral candidate Brian Serafini, found that divorce rates peak following winter and summer holiday seasons. The result of their research can be seen in the graph below (Brines & Serafini: 2016).

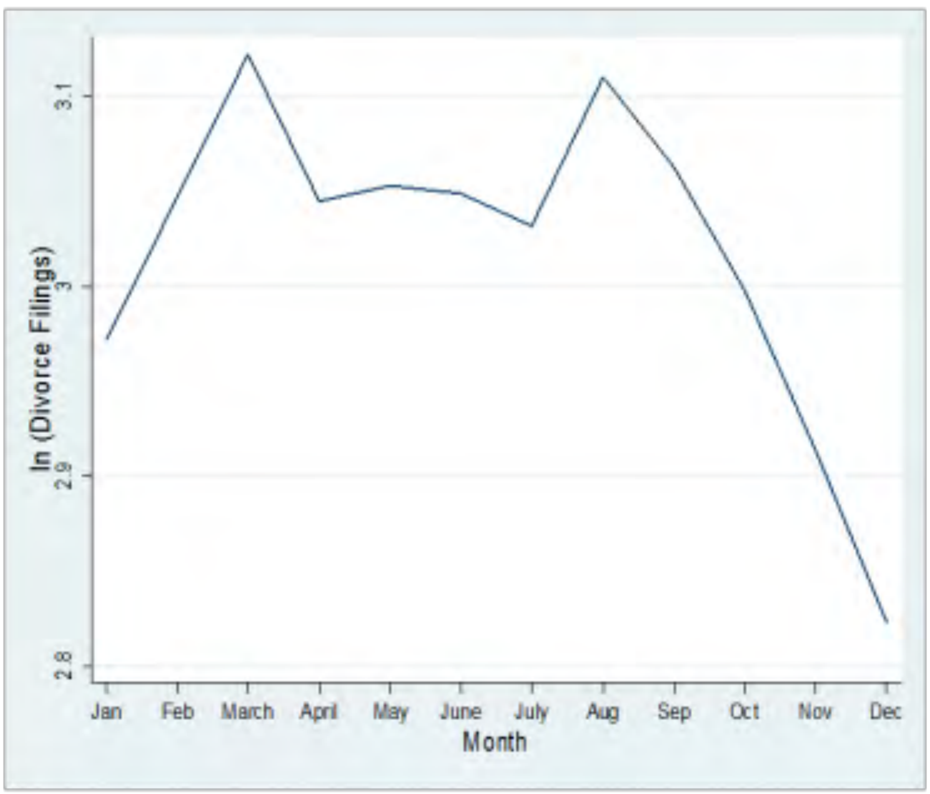


Figure 2: Illustrates divorce patterns in Washington state between 2001 and 2015 and shows that they spike after the winter and summer holiday seasons.

A visit to the theme park studied by the anthropologists should therefore not simply be understood as a destination for fun and thrills, but as a mean for facilitating the social bonding that groups don't feel they can achieve in their everyday. However, when analysing the theme park through the lens of bonding it became clear that there were a number of areas that currently had a negative impact on the customer experience.

ASSESSMENT OF THE CURRENT BONDING PERFORMANCE

When analysing the theme park's current performance through the lens of bonding, three distinct problems took shape.

Firstly, a lot of the rides and activities within the park create opportunities for competition between the members of the group. Upon finishing a ride, park visitors will get individual scores and a ranking so it's clear who did best and how much better they did. This often had the consequence that the older or more experienced members of a group would continuously triumph over the younger or more inexperienced members throughout the park visit, which tended to create a divide rather than a bonding experience. One mother said, for example, when her son and husband were driving go-karts on a track with 8 other park visitors: *"I hope he (her husband) doesn't humiliate him (her son) too much."* This quote illustrates the flipside of the competitive focus of many of the rides within the park. The problem was most prevalent with families and couples, whereas with groups of friends the competitive element often added to the bonding experience.

A second problem related to the often varying levels of *'theme park experience'* within groups. Theme parks are a rare or non-existent phenomenon in many places of the world, especially on the magnitude and degree of intensity of the theme park in question here, and many of the park's visitors are not from the Western world. Rather, a large proportion of the park's visitors were friends and families of immigrants who live and work in the country where the theme park is placed – a country were the native population only accounts for a minority of the overall population. While the local residents have ample choice of theme parks within the country, their guests have often never been to one before because it doesn't exist in their home country. During fieldwork, the anthropologists observed how this difference in theme park experience within groups meant that some members of the group were left feeling like their boundaries had been overstepped while others felt disappointed for not being able to share the positive experience they got from the most thrilling rides with the rest of group. Both these sets of feelings were not conductive to a bonding experience.

A third problem with the park's current performance on bonding related to group separation. Many groups would split up during the day. Often doing so intentionally to allow some members of the group go do something they enjoyed, but that the rest of the group did not – e.g. an intense rollercoaster ride. However, as they split up they often struggled to find each other afterwards and the narration of their separate experiences often feel flat. As a result the groups were often split up longer than intended and when re-united they struggled to share their separate experiences, which meant there was a lack of a shared narrative about the theme park experience.

RECOMMENDATIONS FOR IMPROVED IN-PARK BONDING

To address the three separate problems with bonding outlined above the team of anthropologists developed three different concepts.

To make competition more conductive to a bonding experience the team suggested a more social form of competition, where groups can compete together against other groups instead of with each other internally. This is aimed at creating an environment where the group tries to encourage and help out each other, which should support a bonding experience. A similar concept has been rolled out at a science museum in Copenhagen with great success (Experimentarium: 2018). Furthermore, the CDP should be able to always find some statistic that would make the group look good, as that is only a question of data granularity, e.g. "Best family from Jakarta (IDN) today!" or "Best mom (38) and daughter (12) combo this week!"

To better handle the differences in theme park experience, the team suggested to focus on helping the 'rollercoaster rookies' build up their 'rollercoaster confidence'. This could be done by inquiring about previous theme park experience at point of purchase and then recommend a route through the park that could slowly build confidence. Rollercoasters could also have a tag on them, which clearly stated the intensity level to be expected – similar to what has been done in the theme park Cedar Point in Ohio, which has the slogan "the roller coaster capital of the world" (see fig. 3 below). Today, it can be very difficult to guess the intensity of many of the rides in the theme park, as the ride itself is often hidden from view.

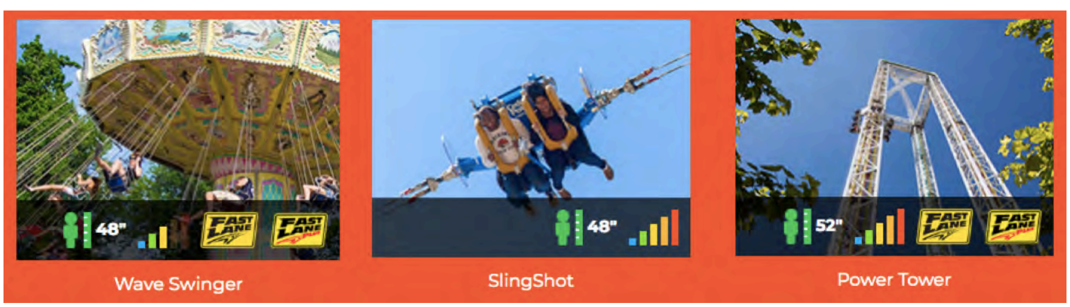


Figure 3: The theme park Cedar Point in Ohio clearly states the 'thrill level' of their rides, as illustrated with the 5 bars above (copied from Cedar Point's webpage, September 8th, 2018).

And thirdly, to address issues related to group separation the anthropologists suggested to develop a system that would let groups easily locate other members of the group by utilizing the geo-location data. Furthermore, the anthropologists also imagined a solution, where the park could use data to help guests enhance the narration of their separate experiences – e.g. by sending information to the guests' phones about other group-members experience such as speed, g-force, an action shot, heartrate or the vertical meters travelled on a roller coaster. This data could then be used when narrating one's experiences to the rest of the group, thereby narrowing the distance felt between group members after they had been separated.

Combined, these three concepts should help the park deliver on guests' social needs and change the focus away from the individual thrills, which had been the park's focus up until this point. The insights along with the concepts were shared with the team of data scientists working on the CDP. The goal for the data scientists was then to suggest ways to quantify how the park delivered on bonding today and how it could be tracked over time. This should be captured in a 'bonding index'.

The anthropologists' insights had now been thrown at the data scientists' sounding board, and the next meeting would reveal what the CDP would send back.

NO BONDING WITH THE BIG DATA

In a poorly lit, windowless room with a low ceiling the team of anthropologists sat down with a team of data scientists, to review the data scientists take on the bonding index. Following the anthropologists' presentation of their insights about what constitutes a good customer experience in the theme park, the data scientist had gone back to the CDP to identify the most relevant data sources in the data lake. The result was a suggestion to build the bonding index based on the following 4 data sources:

1. **Revenue from group tickets:** A group ticket being whenever 2 or more tickets are bought at the same time.
2. **Food & beverage sales to groups:** Defined as whenever 2 menus or more are bought in the same transaction.
3. **Number of guests on competitive rides vs. non-competitive rides:** With 6 rides within the park defined as competitive and all others as non-competitive.
4. **Number of guests on intense vs. relaxed rides:** With a distinction between high-intensity (approx. 50% of total rides), medium-intensity (30%) and low-intensity (20%) rides.

Lorenzo trying to connect with the son he rarely sees. Rodrigo wanting to make the most of the precious time he has with his fiancé. Molly fearing her current holiday will be the last real family holiday. And Robert's bi-annual (they also go skiing once a year) chance to connect with his teenage sons. Their unmet need for bonding in an age of liquid modernity had been boiled down into an index with the 4 KPIs described above.

The anthropologists' first reaction was an emotional one dominated by feelings of disappointment and of being underwhelmed. This was not what they had expected of a multi-million-dollar data setup, of which they constituted only a small part. The second reaction was a more rational analysis of the 4 KPIs. The first KPI was problematic, as it depended more on the bundles sold by the park or travel agents and the composition of the groups, rather than the theme park's ability to deliver a bonding experience. Furthermore, it turned out that almost 96% of current ticket purchases were already group tickets, which left insignificant room for measuring any meaningful changes. The second KPI was also problematic, as about 90% of meal purchases (which excludes sharable snacks such as popcorn) were already group purchases. Moreover, many groups of friends were likely to pay for their meals separately, unaffected by the park's ability to deliver a good bonding experience. The third KPI was based on a misunderstanding. When the anthropologists had argued that there was a problem with the way competition worked within the park presently (as it often creates unintended divides within groups), the data scientists had assumed that this would then be reflected in fewer people using those rides. There were, however, few visitors who consciously reflected on the negative effect of competition and next to none that would avoid the competitive rides for this reason – especially considering that these rides were among the park's biggest attractions. The unmet need leading the anthropologists to suggest social competition existed on a subconscious level and it was therefore not meaningful to assume that a significant amount of guests would actively avoid rides with a competitive element. Furthermore, the data scientists weren't able to measure *who* went on what ride, only *how many* went on it. As some visitors, especially group of friends, were likely to enjoy the competition they could skew this data further – had it been meaningful in the first place. The fourth KPI shared many of the problems of the third KPI. The data scientists had assumed that the existence of rollercoaster rookies would lead to a lower utilization of the most intense rides. However, the rollercoaster lovers, who would go on the most intense rides again and again, would statistically hide the experience of the rollercoaster rookies, as the CDP wasn't able to tell who went on what rides, only how many went on a given ride.

The meeting concluded with both teams agreeing to go back to the drawing board to assess whether it was possible to come up with a better bonding index.

LOOKING BELOW THE SURFACE OF THE DATA LAKE

The first thing the team of anthropologists did after the initial meeting was to take a thorough look at the 250 different data sources contained in the data lake. This was the first time the team of anthropologists had done so, as it had been viewed as the data scientists' domain. The defined job of the anthropologists was to come up with a characterization of a good customer experience – not to decide what data sources should go into the CDP. However, the failure of the data scientists to come up with a meaningful bonding index prompted the anthropologists to take a first look below the surface of the data lake.

Here the anthropologists made two discoveries. Firstly, they realized that the vast majority of the data sources were operational in nature. The CDP would measure food and beverage sales, real-time inventory overview, number of ticket sales through different channels, how people clicked around the theme park's website, the effect campaigns had on ticket sales, nationality and country of residence for people visiting the park (if they bought the ticket online), guest acquisition cost (measured as AED – Advertising Elasticity of Demand), YoY (Year Over Year) key sales data, real-time sales numbers compared to budgeted numbers, customer visit duration, ride utilization rates and downtime, average length of stay of employees and social media followers growth per market – to name just a few data sources. The CDP was geared towards capturing the park's operational performance rather than data on the visitors' experience, which limited the number of KPIs directly relevant for the anthropologists.

The second discovery the anthropologists made was that the CDP was not able to capture data on groups. The CDP had been set up to capture some data on individuals and a lot of data aggregated on a park-wide level, but no data on groups. It was, for example, able capture data on what an individual bought throughout their day within the park, but not who the person buying it were together with and whether these people also bought anything themselves. And it can measure utilization rates of all rides throughout the day, but not who goes on those rides (not even on an individual level). Not knowing what guests formed a group was, obviously, a big obstacle for trying to measure their internal bonding experience.

Having gone through the data sources making up the data lake the team of anthropologists realized that there was no quick fix to the bonding index, as the existing data sources were not suitable for accurately measuring bonding. The sounding board model for collaboration between anthropologists and data scientists, between thick data and big data, had failed. This realization came at a point in time when the anthropologists' involvement in the project was nearing its end and there was no time to make amends within the scope of the current project.

During the last days of cooperation, the combined team of anthropologists and data scientists reviewed the data streams that the park executives were planning to include in future expansions of the CDP. Some of these data streams indicated that a meaningful measurement of bonding would likely be possible in the future, when the right data streams were in place. Some of these promising data streams included:

- **Geo-location:** The park was planning to introduce wristbands that, like the MagicBand at Disney World, would make it possible to track guests' movement throughout the park. By banding together wristbands when groups entered the park it would be possible to track how groups, and the individuals within them, moved around the park. This would allow the CDP to capture data on e.g. what rides a group go on together, how long time a group spends in the park and what the group members do when/if they split up during the day.
- **Point-of-purchase information:** When guests are buying tickets from the theme park's website, from a third party or at the gates, a future CDP planned to capture additional information, such as country of residence and group type. One use of this data could be to test whether the anthropologists' claim that there's a connection between country of residence and rollercoaster experience, as well as whether roller-coaster rookies will shorten their stay in the park if they try a too intense ride too early.
- **Social media listening:** A sentiment analysis of people's comments, tags and photos on e.g. Facebook, Twitter and Instagram was also in the pipeline. One use of this data could be to track the impact it would have, if the park started providing guests with social competition scores instead of the individual scores it provides today. Would guests post more, post more positively and use words related to bonding more often, as a result?
- **Survey review:** The theme park routinely conduct surveys among guests exiting the park and were planning to update their questionnaire. Adding questions related to bonding might help illustrate the size of the problem and track the impact of future initiatives. Furthermore, cross-referencing customer satisfaction scores with geo-location data and ride scores might give insights into how e.g. a younger sibling vs. an older sibling experienced the rides centred on individualized competition.

The above illustrates that it might be possible to build a meaningful bonding index in the future, when the right data streams are in place. Getting these in place was, however, outside of the mandate of this project. And as of this writing, a year after the final report was delivered, a bonding index has not materialized. The central task for the anthropologists and data scientists' collaboration thus fell flat. The remainder of this paper will deal with understanding the underlying reasons for why it failed in the hope that it can pave the for more fruitful collaborations in the future.

THE DETRIMENTAL EFFECT OF BIG DATA PATH DEPENDENCY

When evaluating the project, it became clear that the anthropologists spent the majority of their time on the project chasing a mirage. There were two reasons for this. Firstly, the current CDP included very little data on individuals and no data at all on groups. Capturing data on the customer *experience*, and not just the *effect* changes might have to the park's performance, was thus a futile endeavour from the get-go. Secondly, the project did not have the mandate to define and suggest new data sources to be included in the CDP. This meant that the team could only work with the data sources included in the current CDP.

Before the data scientists made a bid for the project the theme park executives had written a request for proposal (RfP) detailing what sources they wanted to include in the CDP. These were made up of a mix between what data the theme park was already collecting, but would like to automate, and best-practice within the industry. The advantage of defining what data sources is needed upfront is that it lends itself well for comparison between different bidders: What is the difference in price? What are the strengths and weaknesses of their different approaches? Etc. The disadvantage is that a decision about what data to include is, to some extent, taken *before* knowing what data is needed. And once a contract has been signed, it is difficult to alter course. Thus, when the theme park executives signed the contract with the bidder they found most appealing, they also made a decision that meant it would be possible to measure the park's operational performance but not the customer experience. This was not a deliberate decision, however, as they were unaware of the path dependency (Marquis & Tilcsik: 2013) their signatures created. And none of the proposals received by the park executives had challenged the wisdom of the inherent path dependencies created by the RfP.

Looking across other data integration projects that ReD Associates has been involved in, it is clear that there's a lack of recognition of the path dependencies created already when formulating a RfP. One explanation is, that the impetus for most data integration projects seem to stem from a desire to digitalize the company's current data practice rather than build a new digitally based strategy. To automate and consolidate current data collection, rather than rethink the new possibilities a digitalization of data could mean. The decision about what data to include thus largely becomes a common-sensical process, where companies formulate data integration RfPs based on what data they already collect, but haven't digitalized. Then, once data collection has been automated and consolidated, many companies hope to take advantage of this new entity they have created by asking: What else might they be able to learn from the vast amount of data? A question that anthropologists might be hired to help answer, by coming up with new questions to ask the new big data entity. However, due to the often-ignored path dependency the answer to that question is often: Not very much. Having lots of data doesn't necessarily translate into having the right data. As a result, many companies end up with big data platforms, but without the ability to answer some of the questions most fundamental to them. This was the result for the theme park described in this case-study, as the CDP left the theme park without any means to currently measure whether they deliver a good customer experience. In recognizing the detrimental effect that path dependency based on early data decisions can have for companies' big data projects, the authors of this case study suggest two new ways going forward.

Firstly, many companies would likely benefit from an open-ended analysis of their data needs, one that is not limited to what data is currently important for the company, before finalizing what data to include in a new data platform. This will enable companies to make an informed decision about what data sources to prioritize – i.e. a digitalization strategy. This suggested approach is illustrated in fig. 4 below:

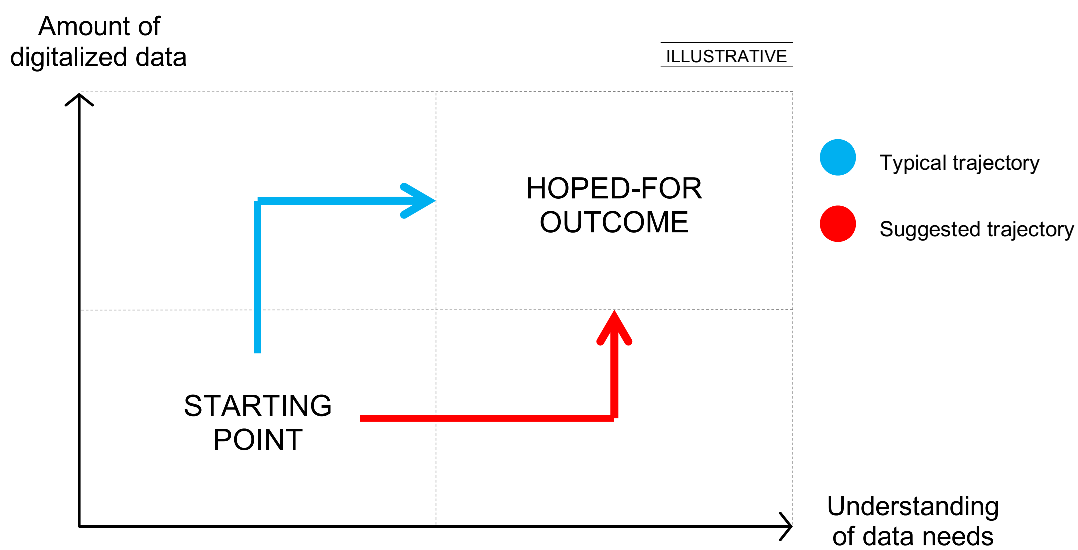


Figure 4: To build a digitalization strategy companies should first conduct an analysis of what they should use the data for (red arrow), rather than just digitalizing their current data sources (blue arrow), which is how most companies progress today.

Had the ethnographic research described in this case-study been carried out prior to the establishment of the CDP the theme park's executives might have prioritized e.g. geo-location data over digitalizing data on the average length of stay of employees. This example illustrates how anthropologists can help companies build a digitalization strategy, by identifying what is important to measure and what data implications it entails. Prompting companies to think through their data strategy, rather than just digitalizing their current setup.

Secondly, there's also a learning directed at anthropologists and data scientists staffed on projects with data sources clearly defined in a RfP. In hindsight, the anthropologists involved in the project described above were naïve in thinking that the CDP would be able to conjure up a measurement for whatever they found to be important for the customer experience and the data scientists were over-confident in thinking that they could. Had the anthropologists looked below the surface of the data lake early on in the project, or had the data scientists been more critical towards the possibilities with the current data sources, they would both have found that there was very little data likely to be relevant for measuring the customer experience. This could have prompted a conversation about how best to proceed, rather than chasing the mirage of trying to build a bonding index with the data available within the current CDP. A recommendation for future projects aimed at combining thick- and big data would thus be to start out with a critical assessment of the available data sources – understanding them as important actants (Latour: 1987, 1993, 2005) for the project. If these turn out to likely be incompatible with the hoped-for outcome, a re-scoping of the project will be necessary. Furthermore, even if the available data sources seem relevant it would be recommendable to try and build in the flexibility of including new ones. Had the CDP, for example, included a lot of data on individuals an early assessment of the data sources might not have raised any red flags. However, as the anthropologists' analysis found, what mattered most for the customer experience was the intra-group bonding. Thus, despite a lack of red flags in an early assessment of the data sources due to a lot of data on individuals, the CDP might still not be able to measure the thing that mattered most for a good customer experience if it wasn't able to capture data on groups. It would therefore be advisable for projects seeking to combine the strengths of both thick- and big data to avoid a strictly narrow scope, see fig. 5 below:

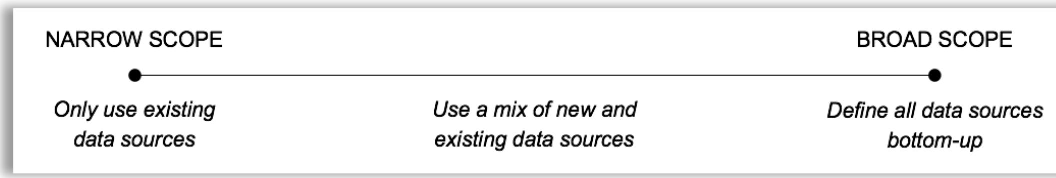


Figure 5: Projects aimed at combining the strengths of thick- and big data should avoid narrow scopes, in order to make room for the explorative power of the anthropological method.

The core strength of the anthropological method is its ability to understand people on their own terms (Malinowski: 1922), unbiased by pre-conceived ideas of what might and might not matter to them. This ability is what enables anthropologists to produce data that's deep, meaningful and original. A narrow scope risks short-circuiting this ability, if the anthropologists' understanding of the human experience has to take fit within a pre-confined set of metrics. Data scientists, on the other hand, often thrive in such conditions, where they can apply mathematical modelling to determine the most important factors in a defined multi-dimensional space – i.e. a dataset with a limited number of variables. However, as illustrated and argued in this case-study, the open-ended anthropological approach can have great value for data digitalization projects, by answering the question: What data should be included in a big data platform?

CONCLUSION

When companies consolidate their data in large, digital data platforms they increasingly start asking what else they might be able to do with this new asset. What other questions might it help them answer? What other truths about their company and their customers does it hold? This case-study has illustrated why many executives are likely in for a disappointment, when asking questions like these *after* a digital data platform has been created. Having lots of data doesn't necessarily enable companies to answer their most fundamental questions. Especially considering that the impetus for many data integration projects is to make existing data practices faster and smarter – not to answer new questions. This case-study argues for a new approach for companies, in suggesting that they start with identifying what questions are most important for them to answer *before* deciding on what data to collect. This will enable the companies to be strategic about their digitalization and make informed prioritizations about what data to collect when. Anthropologists are adept at identifying the fundamental questions that companies should ask at the beginning of large data integration projects. For these reasons, companies embarking on projects entailing the collection of large amounts of data would be wise to start out with an anthropological analysis.

REFERENCES CITED

Bauman, Zygmunt
2000. *Liquid Modernity.* Cambridge. Polity Press.

Brines, Julie & Serafini, Brian
2016 *Is divorce seasonal? UW research shows biannual spike in divorce filings.* In University of Washington News, August 21st, 2016.

CBS
2015 *Roller-coaster's 'weird sensations' perceived differently with age.* CBS News. Accessed July 2nd, 2018. https://www.cbc.ca/news/health/roller-coaster-s-weird-sensations-perceived-differently-with-age-1.3134223

Geertz, Clifford
1973 *The Interpretation of Cultures.* New York. Basic Books.

Kuang, Cliff
2015 *Disney's $1 Billion Bet on a Magical Wristband.* Wired Magazine. Accessed July 3rd, 2018. https://www.wired.com/2015/03/disney-magicband/

Latour, Bruno
1987 *Science in action: how to follow scientists and engineer through society.* Cambridge, Massachusetts: Harvard University Press.

Latour, Bruno
1993 *We have never been modern.* Translated by Catherine Porter. Cambridge, Massachusetts: Harvard University Press.

Latour, Bruno
2005 *Reassembling the social: an introduction to actor-network theory.* Oxford New York: Oxford University Press.

Malinowski, Bronislaw
1922 *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagos of Melanesian New Guinea.* Illinois. Waveland Press, Inc.

Marquis, Christopher & Tilcsik, András
2013 *Imprinting: Toward a Multilevel Theory.* Academy of Management Annals: 193-243.

Stein, Brian & Morrison, Alan
2014 *The enterprise data lake: Better integration and deeper analytics.* Accessed September 8th 2018. https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

Just Add Water: Lessons Learned from Mixing Data Science and Design Research Methods to Improve Customer Service

OVETTA SAMPSON
IDEO Chicago and DePaul University

This case study provides an inside look at what occurs when methods from the data science and ethnographic fields are mixed to solve perennial customer service problems within the call center and cruise industries. The paper details this particular blend of ethnographic practitioners with a data scientist resulted in changes to design approaches, debunking myths about qualitative and quantitative research methods being at odds and altering team member perspectives about the value of both. The project also led to the creation of innovative blended design research and data science methods to discover and leverage the right customer data to the benefit of both the customer and the call center agents who serve them. This paper offers insight into the untold value design teams can unlock when data scientists and ethnographers work together to solve a problem. The result was a design solution that gives a top-performing company an edge to grow even better by leveraging the millions of data records housed in its warehouse to the benefit of its customers.

Keywords: Design Research, Data Science, Ethnographic study, Artificial Intelligence, Call Center Industry, Customer Service

> It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
> (Conan Doyle)

BACKGROUND AND CONTEXT

Anyone who has ever called a call center more than once knows the process can be painful. Customers call with a quick question only to be forced to provide a slew of non-relevant information. Most often they're forced to give information to automated systems. Once their question goes unanswered; only then do they get to talk to a person. But by then they're forced to repeat the irrelevant information to a call center agent before they can actually ask a question.

Even getting a human agent doesn't guarantee the question gets answered if the information needed isn't at the agent's fingertips. Good luck if that call gets interrupted. The customer calls back and the frustrating process begins again. The process can be equally unpleasant for call center employees. Call center workers routinely complain of customer aggression and one study shows 1 out of 5 calls to contact employees are from angry customers, an average of 10 calls a day. The aggression from customers has risen over the years, so much so that the turnover rate within call centers has exploded from 19% in 2008 to 24% and rising today. (Dixon, Ponomareff & Turner 2017)

For every call center interaction between an agent and a customer, there's an equal and corresponding data creation and storage action. Each phone call a customer makes. Every email an agent sends to a customer. Each time a customer goes to a website and enters the search words of some product they want or exotic local they want to visit data collected and often stored. All of these actions create data points. These data points, or digital bits as they're known by the industry, make up what's called the digital universe – all the bits of data that are created, replicated, and consumed by people and businesses. (Gert & Reinsel 2013) But even though all this data available, call centers can't seem to recognize who is calling, why they're calling or how to more quickly solve their problem. This not only frustrates customers but annoys agents as well. It's like there's data everywhere but not a drop that's useful. It's a common problem across industries. A lot of data is being created, collected and stored but that data is incredibly difficult to track back to individuals. There's many reasons for this, but one of the central reasons is that realistically, 91 percent of companies globally report they have inaccurate data, according to a survey of more than 1,200 global companies on data quality commissioned by Experian. (SourceMedia Research 2018) The most common ingredients that lead to data inaccuracies include incomplete or missing data, outdated data and wrong data.

In addition to the data being notoriously inaccurate, data on individuals are collected through multiple channels that are rarely connected. A company may have data on one individual collected from at least three different channels, including email, a mobile application, a phone call or a text. A customer's information is split piece by piece into three, four, five, even 10 different databases, never giving companies a true "360-degree" view of the customer.

Cruise Company's Customer Service Woes

This challenge of having lots of information but not really knowing how to use it, is one many companies face, including, recently, a major international cruise line. Like many companies, the cruise line had collected millions of data records through multiple channels on customers and potential customers. Yet despite this, customers were still experiencing immense frustrations when booking a cruise. Though wildly successful, the cruise line's parent company's new CEO wasn't entirely happy. Rookie cruisers set on having fun in exotic locales, were finding booking cruises extremely difficult and not fun at all. The company's CEO decided to **conduct** some "mystery shopping," calls to his company's call center. He didn't like what he experienced. So, the CEO of a large cruise company issued a challenge to the largest of its operating cruise lines – improve the customer service. The design challenge was to improve the cruise line's customer service so that the booking process mirrored the fun cruisers had onboard the ships.

The company believed the solution was to be found in their call center. There was a list of metrics they wanted to improve including average call time, handle time, how many repeat calls etc. Fix that, they said and the problem goes away. There was also a cursory interest in using customer data to help improve the customer experience. The company had several work streams focused on familiar territory of trying to piece together a "360-degree view," of their customers. In an effort to improve its customer service efforts the cruise company hired the global design firm IDEO.

IDEO teams working on the project would soon discover that to truly solve the problem, to make sure customers felt heard, understood and confident when buying a cruise, the solution lie not just with data, but with people as well. During the first of several research phases, the team discovered that possessing data alone, it seems, couldn't solve the very human desire to be heard. A challenge most companies face is how to surface the right data, to the right person at the right time. That's why companies crave the consult of a data scientist. They seek someone to take largely unorganized and unruly data sets, wrangle them and come up with some salient insights that will give them a business edge. But the team discovered rather quickly that data wrangling wasn't the answer; at least not at first.

Some Level-Setting Definitions

Before explaining the methods used during this particular design project it's helpful to do some level-setting definitions. For this paper, let's define data science and data scientist as it is practiced at IDEO. Data scientists at IDEO are adept at shaping data as a resource for human-centered design. They sketch in pencil and in code, shape and reshape data, know what data can and can't tell us, help design feedback mechanisms, and prototype machine learning algorithms. They do all this to improve human experiences through the design of intelligent products, services, and systems. Data is the paint they use in the art of creating intelligent products.

The most easily understandable example of how data scientists and designers come together to create products is The Nest. This intelligent thermostat system developed by famed Apple designer Tony Fadell, and now owned by Google, is a smart device. The Nest processes data from a number of places including the Internet, the body heat of people inside a room, and adjusts a home's temperature accordingly. To create such smart devices, a company needs people who understand and can develop devices that use data to sense, listen (adjust) and then act. These people are data scientists. Generally, data scientists are seen as people who can wrangle large data sets. But data scientists can also be designers, people who use data as their art to create a different, more improved world. And yet, like most designers, data scientists often find it can be difficult to know where to start, because without context, millions of data bytes sit dormant unable to be rendered useful. That's where ethnographic research can illuminate pathways forward.

For this paper let's define ethnographic research, in the context of design research. In this context, ethnographic research is the study of how people live their lives in order to better understand their behavior, motivations, needs and aspirational wants to inspire new design. This approach is comprised of various methods including interviews, observations, role-playing games and journey mapping. This paper will show that using ethnographic research at the very beginning of the project, transformed the data scientist's initial assumptions about how to use data to solve the problem.

METHODOLOGY

To better prepare the team for upcoming field research, the design researcher on the project did a literature review about the cruising industry and cruiser. Using that information, she created an empathy exercise to help the team understand the reality of people who booked cruises over the phone.

Only two out of the team's six members had ever been on a cruise. And no team member had gone through the entire cruise booking process. Among the team, and in the broader media, the typical persona of someone who goes on cruises is a silver-haired, often retired, affluent couple looking to relax. A quick literature review of the cruise industry painted a very different picture what team members had envisioned. The average cruiser is a 49-year-old employed, married, college graduate with a six-figure income. (Cruise Lines International Association 2017) In addition, "Generation Xers," and "Millennials," were fast outpacing Baby Boomers as cruisers. In one study, Millennials were twice as likely to have taken a cruise than their Baby Boomer parents. (Sheivachman 2018)

Repeatedly, in articles about cruising the idea of seeing multiple locations without ever having to unpack a bag, made cruising appealing to all age groups. Exploring the data collected by the client really couldn't get to the central question, why was booking a cruise so darn difficult in the first place? Once that question was answered, the team felt confident they could design a satisfactory solution.

IDEO conducted two large-scale projects, Project Traveler* and Project Board* (not the projects real names), in multiple phases and with multiple design teams, to diagnose, prototype and design solutions to create a better cruise booking experience. There was 27 weeks of work though the projects spanned more than a year. Only two designers remained constant throughout all phases – the data scientist and the design researcher. The design teams used iterative research and design approaches that included a variety of methods. The project essentially had two user groups: customers who were booking cruises and call center agents who were helping them to book the cruise. The complexity and multi-variant challenge to the problem required an extensive array of design and data science research methods to illuminate a path toward design solutions.

First Phase: Diagnosing the Problem

In the first phase, which lasted five months, the team conducted in-depth interviews with first-time cruisers, cruise-curious people, repeat cruisers in their homes as well as talked to cruisers onboard a cruise to understand their pain points and challenges as well as aspirations when it came to booking a cruise. These interviews were conducted in three rounds before, during and after concept design. During the interviews, team members used card sorting, and role-playing games to help gaining a better understanding of customers' beliefs, motivations and aspirational needs before, during and after the booking process. The number of customers and experts interviewed during this phase totaled 20.

For its second user group, call center agents, the team not only conducted in-depth interviews but also did on-site observations and multiple co-design sessions with agents. For the on-site observations, teams split off into twos, and sat with agents at their desks while they took phone calls. Using headphones specially made for listening to agents while on calls, team members were able to hear both customers calling in and agents' responses. The team made multiple trips to the client's call center to listen in on phone calls. Interviews with agents had to be kept short, because every minute agents were taken from the floor to talk to the team was a minute they lost the opportunity to make money or serve a customer. Because the team didn't want to inconvenience agents or disrupt their job, the team leaned heavily on observation to learn what was needed.

In addition, the team conducted several co-design sessions one of which included creating a gallery walk of proposed design concepts that was installed on-site at the call center for a weeklong feedback session. Agents were able to view design concepts at their leisure and vote on the ones they liked the most. The team interviewed 40 agents in two rounds of on-site observations and interviews. To help client executives and agents better understand insights gathered through field research with consumers the team also conducted several role-playing and empathy exercises.

Second Phase: Testing Solution Prototypes

Once design direction became more concrete (read about how and why in the Findings section below) the team also conducted two live prototyping sessions. In the first phase, the team and the client designed a two-week prototype which altered the organizational structure of the call center to better meet customers' needs. In the second phase the team conducted a large-scale usability test to help solidify the desire for a propose software solution. The team tested a proposed new software application live with 11 agents using anonymized customer data. To create ensure that our usability test with agents were as real as possible the team used real customer data that had all personal identifying details expunged but still had real information that agents would recognized. The team created various behavior typologies that we reconstructed from real customer calls and hired four actors from the Chicago-based Second City Comedy Troupe to embodies those behavioral typologies and into the call center to book a cruise.

Agents knew they were testing a new software application but they had no idea the calls weren't real until after the "customers," hung up. The actors didn't receive scripts, rather the team used various behavioral data points to create behavioral frameworks for actors to embody when they made the phone call. Four behavioral frameworks were constructed from anonymized customer data the team collected and analyzed from the client. The behavioral frameworks used during the agent user testing phase stemmed from a new mixed-method approach developed by the data scientist and the ethnographers on the team.

Mix Method Creations: Data/Human Journey Mapping

It is a customary method for data scientists to map where data is housed and how it flows within throughout systems. (Loukides 2018) This is done for a variety of reasons but most acutely so data scientists can pinpoint how data is created and transformed as it is moves from one system to another. Mapping data flows helps to pinpoint flaws, biases and errors in data. It also can yield opportunities for creating new models that help to take what is known as unstructured data, data that is notoriously difficult to analyze and make sense of, and turn that data into something useful for an intelligent model. Plus, it's just basic to know where data comes from before a data scientists starts massaging and working on it.

For example, a data scientist may map a system and see that a company may not keep track of customer complaints using software but does tape all its phone calls. Locating the audio recordings within the company's systems is paramount for a data scientist who wants to detect patterns and anomalies in customer speech from those phone calls. While the data scientist may request all the audio recordings, who exactly stores and archives them may be a missing point of information. Often who enters and extracts data is left out of the data mapping process.

During this project, the team decided not only to map the data, but map who created, transferred, transformed and extracted the data. In addition, the team decided to map at which point in the data flow, data was created, transferred or transformed and for what purpose and, of course, by whom. This method was dubbed "A Data/Human Journey

Map." This detailed mapping not only told the team what data was being stored, but where it came from and how it was used by agents. It also allowed the design team to pinpoint the exact data that was most important to both customers and agents, the systems where that data was housed and became a key part of the resulting design of a new software application. Data mapping becomes an essential ethical exercise when you pair this flow mapping with people. A data/human journey map is simply a journey map that shows the input, transference and transforming of data throughout a company's information technology systems ***and*** who creates, touches, transfers or transforms that data as it flows through the system. Creating a data/human data flow journey map from the user's perspective is best when ethnographic methods such as observations and user interviews are combined with data exploration. While journey maps have always included people and systems, this map included people, systems and data, actual information points created, stored and retrieved by customers and agents as part of the journey.

For example, when a customer called and gave their name and where they were from, it was mapped to the database or system that information was transferred to or retrieved from by an agent. These systems roundtable interviews were also key as they helped the team to determine what data was stored where within the company. Those roundtable discussions gave the team insight into the current data and systems the client used. It also allowed the team to find ways to leverage the client's existing stored data and systems in a new and innovative way without significant expense or technology investment. This would prove key in getting buy in from the client on design recommendations.

> "Having a data scientist on the team, as a form of research with the understanding and the ability to leverage what data exists was huge," the project leader said. "As designers, we always think to start from scratch. But the [data scientists] pushed us to think about what data we could lever to make the experience better. She pushed us to think what are the systems that our client is using, how do they operate. Normally we wouldn't consider that until later on."

In all, the team spent the better part of 18 months researching, concepting, prototyping and user testing more than 21 different design concepts. Teams spent interviewed or observed more than 70 stakeholders, customers and agents in homes, offices and onboard a cruise.

GENERAL FINDINGS

The team's in-depth interviews quickly lead them to discover a universal truth–no one cruises alone. The complexity of the booking processes compounds this truth. Even if a person is going on the cruise solo, he or she usually has to consult someone else to make the decision. One person may be on the phone, but to make the final decision it could be more than 10 people involved. The team imagined the cruise booking process was simpler than it actually was. After conducting interviews with both first-time cruisers and those who had not been on a cruise yet but were actively looking, it became apparent that cruising ain't easy. Take Michelle *(not her real name).* Michelle is a 37-year-old mother of two who lives off the Florida coast. The team interviewed her at her home just as she was in the midst of booking her next cruise. Even though she has been cruising since she was a teen-ager, Michelle says she's still a novice when it comes to booking a cruise.

> "While I'm booking [a cruise], I'm stressed," she said softly as the team sat at her dining room table in rattan chairs. The stress of the process weighing on her face as she dictated all the decisions she had to make. "I need to figure out the date, the room type, do they have it, is it the right port that I want to go to, so it's kind of stressful. "I'm not an expert," she continued. "I've done a lot of cruising, but not so much recently. I know about the tipping procedures and what you can bring on board and what you can't. I mean, I think I do. Who knows? It might have changed since then."

Even though Michelle had been through the booking process repeatedly, she had been on nine cruises, she still felt like a novice. Her insight that the booking process is ever changing making it difficult for anyone rookie or not to navigate was particularly inspiring to the team. The team sat in Michelle's living room, outfitted with rattan chairs and a beautiful glass table, covered with a white sheet to protect its shine, as she made a phone call to a cruise line. She said she wanted to call two different cruise lines to get pricing and availability for her next cruise. On the day the team went to interview Michelle, she called the client's cruise line. (Incidentally, researchers knew about the call before and asked Michelle's permission to listen to it. It was totally coincidental that Michelle called the client's cruise line.*)* The call was painful. Cruise line stakeholders visibly cringed as they listened to the Michelle struggle to get her questions answered. She had to repeat her destination desires multiple times and never really did get the answers she was seeking. Shortly after that call, she called another cruise line, the client's competition. The call went better, Michelle thought, but she was still unsatisfied.

> "I feel okay. I don't feel fantastic with either of them. I guess I feel better from this call. The other guy seemed just kind of like, you know, wanting to talk about his own trip. … But both of them, I don't have all my questions answered, so…" and her voice trailed off in soft disappointment.

Cruise Booking: The Super Bowl of Decision-Making

Michelle's experience was hardly isolated. As the team sat in its project space and debriefed each interview with cruisers, the room seemed to get brighter from the collective light bulbs going off. Selecting the right cruise is the Super Bowl of decision-making. The complicated booking process turned the simple information processing model humans unconsciously do in milliseconds to say, decide whether to buy that new dress from Amazon, into a long drawn out string of second-guesses and "I don't knows."

There was the destination decision. The room decision. Add on to that the decision about the location of the room - front or back of the boat? Need two beds or four? Want a view or not? What about dinner? Eat at 5 p.m. or after 8 p.m.? Double those decisions when taking along a friend or a spouse. What about the kids? Don't live in Florida or one of the many coastline states where cruises embark? On average, people who cruise take between three and 18 months to plan and book a cruise. (Cruise Lines International Association 2017) That's an extremely long sales cycle. Add airline and departure information to a list of decisions and the process gets even more complicated. In all, a rookie cruiser would have to make about 10 decisions, including how much to pay just to get past the basics. And that's not even including what to do on the cruise. Imagine doing that on a phone call in between lunch breaks at work?

In addition to talking to rookie cruisers, the design researcher on the team also conducted an autoethnography experience, reflecting on the exceedingly long and difficult journey she took to book a cruise and rooms for all team members. (Four members did end up taking a cruise from this exercise.)

In the age of Kayak, Expedia, Priceline and Orbitz who have made traveling as easy as ordering a pizza, booking a cruise remains maddeningly difficult. Some quotes from our field research:

> "It's overwhelming." … "It was bad." The first time was bad 'cause I got frustrated…Because I thought you could just go online, Google, there you go. No, it's not like that."
> "My head's hurting. You know like you're reading something so long, or you're studying so long that you give yourself a headache? Just put everything to the side and I'll come back to that."
> "[With groups] …it's like herding cats…too much frustration. I decided not to do that anymore."

Processing what they had heard from the field and the reflection of a team member's own experience booking a cruise, it was at that moment It was at this moment, just weeks into a multi-month project, that the team's data scientist, would plant the seed that grew like a foundational vine unifying one of the team's eventual design solutions.

> "If there is an entire space of possible ideas and how to get there, the way that a data scientists would think, we would be to say 'What are the pie in the sky ideas? What does the future look like?" Ann* the team's data scientist quipped. "But this is where having a collaborative team with multiple disciplines including ethnographic researchers can make a difference and can come into play. The collaborative team, comes at it from a point of where are the human pain points followed by where are the business opportunities and data comes from it at another angle, how can we make systems more intelligent. We're looking through all those lenses together to serve a very specific human problem."

The field interviews yielded some interesting "truths" about customer exchanges with cruise call centers:

- Rookie cruisers didn't book their cruises in one call. They often made repeated calls because questions popped up as they learned about cruising.
- Rookie cruisers rarely talked to the same person when they call about their trips which forced them to have to repeat the same information multiple times
- Rookie cruisers were often calling for advice but got a sales pitch instead

After the field interviews, the team was left feeling as frustrated as the customers it interviewed. Why was this process so difficult? It didn't take long to decide that it wasn't the website. It wasn't the app. It wasn't the call center. It was all of the above. The team couldn't wait to reconcile what they heard in the field with what went on at the client's call center. And they were able to do just that when they visited the company's call center for some marathon observation sessions.

Ethnographic Methods Makes Intelligent World Understandable

Walking into the bright, spacious building of the cruise line's call center was akin to jumping on a merry-go-round going full tilt. Dozens, and dozens of call center agents sat row after row, some standing, others leaning on their chairs some bouncing even, as they cheerily but assertively, talked to customers about booking cruises. There were shouts, and bells ringing when cabins were booked and sold. Leaderboards adorned the entire back wall as names of top sales agents beckoned the competition to come and get them. Sales coaches and managers paced the floors walking up and down the cubicle aisles giving advice. It was a call center, filled with extroverts, deal-making and fun, LOTS of fun. Timing and scheduling forced the team to be on site at the call center for just three days. Yet, that was enough time for center's infectious energy to engulf the team. Being on-site at the call center allowed the team to see the company's culture, understand employee incentives and see where service played a part in the sales journey. Listening to frontline agents take and make calls to customers was, by far, one of the most lucrative research activity the team did. The phone calls told a truth that was buried, hidden and obscured by disparate systems, human nature and a lack of transparency. After a couple of hours of listening to phone calls; it was clear that calling into a cruise line's 1-800 number was an adventure. But not a fun one. It didn't take long listening to agent's phone calls with real customers to hear echoes of the frustrations expressed by cruisers during our earlier field research phase.

Team members noticed that the calls weren't great for agents either. When the customer called in often agents didn't know anything about the customer, beyond name, address and phone number. Agents would have no clue the customer called in yesterday asking questions about a Caribbean cruise. They wouldn't know that the customer was afraid of flying and booked a hotel two days in advance from their cruise date and needed directions from the nearest train station.

Ethnographic Methods Makes Data Science More Powerful

Adhering to the research plan which prompted team members to document information exchange and transfer between agents, customers and systems, members noticed agents would have to search through multiple software systems to retrieve crucial information about a customer's history with the company. This lack of understanding about who was calling into the call center and why was a major roadblock to the customer experience. It led to long wait times as agents sifted through a half a dozen software applications to find the exact information needed to help the customer. It was also obvious that enterprise systems were a barrier to agents having information readily accessible to help customers.

During these observations, the team also discovered something that never would have been found through data exploration alone–the sheer amount of relevant customer data that was being captured by agents but not stored by the cruise line. The team observed frontline agents repeatedly recording information from customers in a notebook or on their desktops in Notepad or Word documents, only to delete or throw away that information after a call was completed.

Sometimes the information was thought to be unnecessary. Other times the agent didn't really know what to do with the information. And sometimes the agent didn't see the information as valuable because it didn't lead to a sale. We also saw agents using these notebooks to "double check," their sales. Though all sales were marked in their system when a customer booked, some agents weren't sure they were getting credit for all their sales. So, they recorded every sale in their own notebooks. This let the team know that agents didn't trust their systems to record data accurately. Coincidentally, many customers didn't either. This observation was something the data science designer found invaluable to see before crafting an intelligent model using data for a solution. Such human data exclusion and manipulations are virtually undetectable during a data science usual methodology for dissecting and analyzing data. Data scientists usually only see the output of data. They see data without context. Seeing how data is created and transformed by humans is essential to devising useful data-driven solutions for businesses. Paring what the team saw in observation with the data mapping exercise allowed the team to see opportunities for design. This map pairing allowed the team to see the system barriers agents and customers had when creating or extracting information for use. Mixing the traditional data mapping exercise with the human data interaction exercise allowed the team to actually "see," how data began and ended in the client's technology systems. It was also great way to see how user data was actually used by agents. Visualizing this journey was key to finding the root cause of the pain points both agents and customers felt during calls:

- Agents spent a lot of time on calls manually pulling customer information from multiple applications
- Agents struggled to build rapport with customers because personal information about customers was missing or incomplete
- Agents struggled with better visibility into customers' past interactions with the company forcing customers to repeat information frequently
- Agents wanted to take ownership of their guest relationships but feel they had to depend upon systems they didn't wholly trust.

After the observations, the team developed a "behavioral data map," that allowed it to see the connective tissue the customer's information traveled through during a phone call. The team discovered that the customer's information was split between three main information systems and it took considerable amount of effort and system know-how to connect them to give a coherent and contextual understanding of a customer's need, if the agent could do that at all. The team could pinpoint design opportunities to leverage existing data to the benefit of the customer. Without the on-site observations, the data flow mapping would have given an incomplete picture to the team.

> "Mapping the people to the data and the journey both take with data makes it easier to identify data priorities, redundancies and how data is transformed and changed throughout the system," said Ann*, the data scientist on the project. "This gives data scientists the edge they need to focus on the information that really matters and find opportunities to augment human intelligence, instead of just taking a bunch of data without context and plugging it into some new technology model and seeing what comes out."

Frontline inbound sales agents, who were the agents that answered the cruise line's 1-800 number were often the ones who recorded the data. It was incredible the amount of personal details these agents were capturing on customers that never ended up in the cruise line's databases. Information gathered on calls where a customer did not actually book a cruise was just "lost," preventing other agents from "picking up where another agent left off," when a customer called back.

> "For a company that is very focused on conversion and selling quickly for them to have an agent to capture that information and take the time out to capture that information is huge," Cheryl*, the project lead said. "As designers, we need to show them the value, to show them how much that information is worth and why it's worth collecting."

Using Empathy for Paradigm Shifts

Because the client collected a lot of data about its current and prospective customers–basically anyone who contacts them through their call center or website–the team determined that leveraging that data that lay dormant and often inaccessible to agents could go a long way in shoring up some of their customer service issues. But just like there was a separation of sales and service agents, there was a separation of sales and service data. This left the data the company collected on consumers as fragmented as the service experience.

Still one truth remained–ethnographic research revealed that leveraging data alone wouldn't solve most of the customer service problems. Optimizing the company's ability to use data without changes to the culture and organizational structure wouldn't give agents the freedom and incentive to use that data to serve customer. Infusing agents who were focused just on sales and not service with customer data might benefit the agents, but it would do nothing to improve the customer experience. But how to tell a client expecting some tweaks to a script or a series of new training modules that they needed to change their entire employee organizational structure before they could give their information data systems a modern facelift to serve customers better? Not an easy task. But the team decided that building empathy for the customer's pain would give the client a sense of urgency that made them conducive to large-scale change.

After returning from the research trip in the field the team holed up in a project space to process what they learned. The team had just a few weeks to synthesize what we had learned. Plastering quotes and notes from our field debriefing sessions the team holed up on their project space to find themes and patterns. In addition, the team's researcher branched off to read transcripts and listen to audio from various observation sessions the team witnessed. This dissection of transcripts lead to the "Anatomy of the Call," and exercise that broke down customer calls to agents, section by section marking information and data used during those called. This allowed the team to see what data was most frequently requested and used during phone calls. Subsequent interviews with agents and data analysts and software engineers confirmed the team's hunches on information most used by agents and most requested by customers. In addition to synthesizing data capture, the team also worked through patterns found in interviews from customers and listening in on calls. Clustering quotes and deconstructing their meaning, the team centered around four actionable insights.

- People want support that acknowledges who they are and adapts to where they are in the journey.
- People seek confidence in their decisions, and rushing to commit before they're ready makes them feel insecure.

- People are delighted by a streamlined process and deflated by repetition.
- People naturally collaborate during the travel planning process, but they feel overwhelmed by managing this task

The team felt the insights lead to good design opportunities and was excited to share these with the client. But team members knew it would be difficult for the client to act upon the insights without winning the hearts and mind argument. The idea of sales and service being separate was entrenched in the client's culture. The team had to figure out a way to get the clients to see how disruptive this split was to the customer experience. The stakeholders needed to see their roles and the roles of the agents within their organizations differently, from the customer's view. Sharing what customers said was one way, but the design researcher and the team wanted their clients to actually feel the barriers that guest felt when accessing their call center. So in true design research form, the design team designed a series of empathy exercises that put the client core team through a one-hour timed game that simulated the way their customers felt. In this newly stylized version of television gameshow "The $10,000 Pyramid," players were positioned back-to-back and were asked to guess words based upon clues given by their partner.

One partner could see the word and had to give clues to the other partner, while not saying the actual word. On average, there were three to six words to guess. Like a phone call to a call center, partners couldn't see each other. Each round got harder as the rules changed. The "rules," mimicked many of the barriers customers encountered when calling the call center, the communication constraints (can only give three clues), the system constraints (players couldn't skip words) and, of course, time constraints. Players were rushed to make decisions. The giggles were equally matched by groans as player after player failed to win. Older players used clues that younger players didn't understand. "Who is Mick Jagger?" One player went down in flames because he couldn't think of the word "Limbo," though he kept slinking down in his chair as if sliding under a phantom pole. As the words got increasingly abstract the players sounded like they were speaking foreign languages to each other. Only one pair guessed all the words in the allotted time. It was a fun, but poignant way to have them walk in their customer's shoes. Through the exercise the executives felt much like their customers did on the phone, frustrated, misunderstood and ignored. It was a powerful illustration of just how frustrating the entire booking process could be. And it exhibited the true uniqueness of design research. The client had collected raw data about customer satisfaction. But having to go through an exercise where they were treated like their customers helped them to understand the feelings behind the data. From one executive after the meeting,

> "I feel like the shackle has fallen from my eyes. No matter what becomes of this project, I know that I will not think the way I did in the past again. This process has changed my whole outlook." – Cruise Executive

This meeting was a turning point for the project. It marked the point where the client core team fully came on board. The sincere desire to change exhibited by executive leadership at that meeting gave the team license to expand design opportunities in new and innovative ways. It allowed the team to take some bold design risks including live prototyping a new organizational structure that teamed up sales and service agents to take calls together. But by far the most exciting change the client glommed onto was an opportunity to leverage existing data to make its call center agents more knowledgeable and responsive to customers' need in a quick and efficient way. Though the one design recommendation the client wanted to implement right away was a data-driven solution, it was the tried-and-true ethnographic method of empathy-building that created the opportunity for this bold new design.

LESSONS LEARNED

Convinced in the new opportunities presented, the client embraced the expansive design solutions proposed. The team landed on four design recommendations to help the company improve their customer experience.

Three of them focused on organizational and communications changes. For example, the team's research suggested that the way inbound agents were incentivized was a huge barrier to the customer experience. So, a design recommendation devised a different way to organize the two sales divisions and one service department to better meet customers' needs when they call in. However, there was one design recommendation that sprang solely from collaboration with a data scientist. This design concept, created to envision a new way to organize and surface customer information to agents was designed to make customers feel heard, understood and served and is called "The Board.

Through design research, the team had mapped three parallel journeys - the customer's, data and systems. Using text from customer phone calls, the team matched all the information that customers talked about during a phone call to the software or enterprise system where that data was retrieved or stored. The team did this to see if there was a better and quicker way to surface this information during phone calls to help eliminate the frustrations customers felt during calls.

Real People, Real Data Usher in Ethics Realities

But now that the laser focus was on real customer data, another challenge surfaced-just what data was needed and exactly how would consumers feel about agents having a fuller picture of them? As human-centered designers, the team felt the idea of surfacing more data on potential and current customers to agents deserved more than a passing thought. Sure, the design idea was to serve customers better, but would having too much data on customers make customers feel vulnerable? How would customers feel about agents already knowing who they were and why they were calling? What data was off limits? What data was OK?

Designing this testing exercises with real data forced the team to have several discussions about the ethical use of data. As a matter of course, the design team decided as a team and company, not to take custody of personal identifiable information (PII) of cruise line customers. Personal identifiable information includes name, address, birthdate, social security numbers and any other information that could immediately identify a person. This decision stems from a universal truth handling data ethically, to only take as much data as one needs to solve a problem no more. The team, thanks from input from the data scientists and a software engineer guide on the project, rationalized that it didn't need the caller's real name, phone number, exact address etc., to analyze the calls. Instead the team focused on the content of what the caller was saying, where did they want to book a cruise and what systems and tools they used to learn about those destinations.

While ethnographic researchers are used to getting consent from everyone they engage. This is feasible. But when it comes to data, teams will also have to engage hundreds, if not thousands of people through their data. Getting consent for that engagement on an individual basis is nearly impossible. So, as a practical and ethical practice, it's good to just ask for the data needed to actually solve the problem and no more. Some tips on doing this include:

- Scrub the PII from the data and use the rest, all but the personal information, for data simulations, prototypes or data visualization models.
- Create new data that simulates the characteristics of real data without the PII to run data simulations, prototypes or data visualization models.
- In rare cases, become an approved vendor with all the data compliance requirements met and transfer raw data (not scrubbed) to company's servers to work simulations and visualization models. This is very rare and not recommended because it ups liability and often is unnecessary.

The team also took its data ethics discussion to cruisers through concepts that visualized the use of customer data in design. The team talked to current cruisers to get their sense of what data was too much information for their cruise company to have. Far from being invasive, customers said they would welcome their cruise company to use personal information to serve them. Specifically, customers felt knowing intimate details like their birthdays or whom they usually travel with weren't off limits to their cruise company. They felt having such data would shorten their process and get them closer to the fun quicker. Again, none of the data the team proposed to use was not already collected by the cruise company. The team just wanted to use it in service of the customer. As a cruising couple said to the team, "We've been cruising with them for 12 years, they already have all this information on us. Why not use it to our advantage?"

Data Science Methods Amplifies Design Solutions

The team's research revealed an important insight when it comes to using customer data. As long as the customer felt they were receiving a benefit for the use of their personal data, its use seemed more sanctioned. Even with customers' willingness to have cruise companies use their data to their benefit, the team still wanted to ensure that what they designed would truly help customers first, and not just benefit the cruise line. This was a point of view came from the team's dedication to human-centered design. An example of this dedication happened when it came to financial data on customers, specifically historical data on how much money cruisers spend on board the ship.

After concepting, a few versions of a new design that would surface various information data points including how much onboard spending a customer had done on previous cruises, the team decided not to include this financial data in the dashboard for sales agents. After much discussion with the client core team, it was decided that showing such financial data would put customers at a disadvantage of an aggressive salesperson. The sales agent could use the onboard spending average as a barometer for a person's wealth capacity, tempting them to oversell a package that did not align with what the customer actually wanted. They didn't want customers to feel agents had inside information that could be used against them. Data should be leveraged for the customer's, not the agent's benefit. This type of discussion is what the design firm later called "white hat/black hat."

It's looking at the data to be used or model to be created and trying to figure out the worst thing that could happen if a bad actor got ahold of that data or model being created. If the consequences harm people then the team needs to adjust the model, create safeguards, or even abandon it for a new model.

Data Science Methods Amplifies Design Solution

Using the data mapping and system mapping methods of the data scientist on the team gain the ability to not only define the data both customers and agents used most, but map that data to internal systems so it could be resurfaced and previsualized in a new way. In addition, the data scientists offered new ways to bring in external data, such as local weather data to make the calls feel more personal and less transactional. The Project Board team created a newly visualized software application which surfaced the most used information to agents during customer calls, but excluded sensitive information that could categorize people by personal spending habits.

The Board draws from three main cruise line systems to create a more complete picture of customers when they call. Results from the usability testing showed it was nearly unanimous among agents that the experience was better than their current system. Both agents and callers felt the new "Dashboard," made for a smoother and better customer conversation.

"The feedback I get just from looking at a page like that is awesome," one agent said. "It provided a lot of information that I usually have to go looking for so it makes it easier to respond to guest," another said. "This is everything," another agent exclaimed.

The design also included room for the cruise line to incorporate more sophisticated data modeling including machine learning to help improve a customer's experience over time.

DATA AND ETHNOGRAPHY SHARE EQUAL BILLING IN PROJECT OUTCOMES

It took about eight months and another project to bring the "Dashboard," into fruition, but the seed of this new software application was planted a year before in a small project room. That's when a design researcher, a data scientist, an environments designer, a communication designer, an organizational designer and a business designer all poured over transcripts of customer calls to the call center.

They all felt the pain of both agents and customers with the current systems and processes, and a data scientist saw that same pain and offered a way to leverage data to help make the process better. And an entire design team worked together to create an innovative software product that the client could implement into existing systems to get customer service value right away. There's no way to know of course if the results would have been as good without the mix of a multidisciplinary team. But it's a certainty that it took both ethnographic research and data science to pull off the right solution.

"That's where the design research comes in," the data scientist said. "We got an accurate deep learning of what people need today and what the pie in the sky was for tomorrow and we worked backward from there. We created the first version given the data limitations they have today but we created it in such a way they can get further and more sophisticated in their data use over time. We created the foundation for an intelligent dashboard. And it's not just a dashboard that surfaces basic customer information, it's collecting that information such that when the company is ready to invest in something that's more intelligent, that uses machine learning for example, they have the foundation to do so."

Project Board is an example of pairing the highly effective skill of call center agents with the processing power of software programs to get a better result than either alone. For this reason alone, not the data that is integral to the design, but the people who will make use of this surfaced information, Project Board is a good example of the future of human-centered AI. It's not about the data. It's about the people, and how surfacing the data made people better at their job, happier doing their job and customers happier when interacting with customer service agents.

Ovetta Sampson is a design research lead at IDEO in Chicago where she uses a variety of qualitative and quantitative methods to inspire innovation in various industries including IoT, transportation, and service. She works with clients to visualize ways to ethically and with compassion serve people through digital and intelligent products. She also an adjunct professor in the School of Design at DePaul University in Chicago. Contact her at ovetta@ovetta-sampson.com.

NOTES

The paper's author would like to acknowledge the many people who helped make this publication possible. This includes the leader of the project and data scientists who worked on both projects, both of whom are not named at the request of the author's employer. In addition, the author would like to acknowledge the many data scientist and design research colleagues at IDEO and at DePaul University who helped her highlight the most illuminating and innovative parts of this very complex and unique project. This includes Jess Freaner, Dan Perkel, Lisa Nash, and many others. Special thanks to EPIC Mentor Ryoko Imani for your selfless commitment to this first-time EPIC author and your endless patience answering all of my incessant questions.

The information included is in this paper is accurate but does not represent the official case study about the project from the author's employer IDEO.

REFERENCES CITED

Cruise Lines International Association
2017 "Cruise Travel Report." Annual Survey.

Dixon, Matthew, Lara Ponomareff, Scott Turner, and Rick DeLisi
2018 "Kickass Customer Service." 1 Jan–Feb 2017. *HBR.org* 25 September 2018. https://hbr.org/2017/01/kick-ass-customer-service.

Gert, John and David Reinsel
2012 *The Digital Universe IN 2020: Big Data, Bigger Digital Shadows and the Biggest Growth in the Far East – United States. 2012 IDC Country Brief*. Study. IDC. Framingham: EMC, 2012.

Loukides, Mike
2018 *The Ethics of Data Flow*. 11 September 2018. 18 September 2018. https://www.oreilly.com/ideas/the-ethics-of-data-flow.

Neff, Jack.
2018 *Ad Age*. 11 July 2018. https://adage.com/article/news/ai-models-real-consumers-reveal-research-answers/314137/

Sheivachman, Andrew.
2018 *Millennials Now Enjoy Cruising More than Boomers*. 8 May 2017. 8 July 2018. <https://skift.com/2017/05/08/millennials-now-enjoy-cruising-more-than-boomers/>.

SourceMedia Research.
2018 *The State of Data Quality in Enterprise 2018*. New York, 1 January 2018. Report.

Scale, Nuance, and New Expectations in Ethnographic Observation and Sensemaking

ALEXANDRA ZAFIROGLU
Intel Corporation

YEN-NING CHANG
Intel Corporation

We consider new expectations for ethnographic observation and sensemaking in the next 20-25 years, as technology industry ethnographers' work unfolds in the increasing presence of the type of analytical capabilities specially trained (and self-training) machines can do 'better' and 'cheaper' than humans as they can take in, analyze and model digital data at much higher volumes and with an attention to nuance not achievable through human cognition alone. We do so by re-imagining three of our existing ethnographic research projects with the addition of very specific applications of machine learning, computer vision, and Internet of Things sensing and connectivity technologies. We draw speculative conclusions about: (1) how data in-and-of-the world that drives tech innovation will be collected and analyzed, (2) how ethnographers will approach analysis and findings, and (3) how the evidence produced by ethnographers will be evaluated and validated. We argue that these technology capabilities do offer compelling new ways to model and understand the contexts in which ethnographic encounters take place. Yet because ethnography has never been solely about describing behavior, or about testing hypotheses to ultimately generate laws, these new tools will never get us on their own to the type of truths the ethnographer values above all else: the meanings given to experiences by humans.

INTRODUCTION: LIQUEFACTION OF EPISTEMOLOGICAL GROUNDS

Ethnographers employed in various silicon geographies — Valley, Forest, Ally, Wadi, Cape, Gorge, Roundabout[1] — have experienced destabilizing epistemological ground shifts with the emergence of technologies that create new types of data and evidence more valued and trusted by their co-workers to inform product innovation and company business strategies than qualitative, applied ethnography. This epistemological earthquake started with the shudders around big data in the mid aughts, followed quickly by the jostles and rolls of advanced analytics and has culminated in the liquefied ground brought on by the real — and richly imagined *potential* — force of artificial intelligence fueled by data produced by the blanketing of the physical world with the sensing and communication capabilities of the Internet of Things. In geological earthquakes, liquefaction destabilizes the support for building foundations and other objects on the ground; liquefaction of epistemological grounds similarly destabilizes what data count as evidence. In either case, liquefaction is not a permanent state though the ground/data and the objects on it/evidence never revert to their previous arrangements. Instead, attempts to salvage, rebuild, and stabilize on new landscapes ensue. These altered landscapes offer opportunities to refresh or build physical and intellectual structures anew. City planners prioritize structural engineering and update building codes. Technology companies valorize new knowledge systems and data (in the philosophical sense) for characterizing the world, and new expert workers who create and work with this data (in the computing sense).

In this paper, we explore what will irrecoverably change, what will be contested and what will be the net gains (and losses) for technology industry ethnographers in this new landscape. First, we outline fundamental changes to the narratives the industry tells itself about how innovation happens, and the new competition ethnographers face in observing and making sense of human experiences. We situate this new competition in the larger context of the advent of the fourth industrial revolution and the changes in how data act in the world. Then, through a series of re-imaginings of our current and past ethnography research situated in artificial intelligence and computer vision saturated environments, we will tease out what types of evidence become possible, which types of evidence become more and less relevant or trusted, and re-imagine the role and activities of future technology industry ethnographers.

New Competition in Observing and Making Sense of Human Experiences

In this altered landscape, one certain irreversible change is that silicon geography ethnographers face new competition for their expertise in observing and making sense of human experiences. The established innovation narrative common in the technology sector that justifies the expertise and value of employing ethnographers faces a compelling new narrative in which complex and increasingly ubiquitous computing solutions and systems observe the world, create models of it, make inferences and act with ever-decreasing reliance on direct human intervention. The humans most critical in this new narrative are computer scientists specializing in machine learning and data scientists, experts who can "obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning" (Woods 2011[2]). In 2010, Hal Varian described statisticians as holders of "the sexy job in the next ten years" (McKinsey 2009). By 2012, *statistician* had morphed into *data scientist*, in recognition of the computer science and machine learning expertise entailed, and sexy morphed to sexiest with a much-extended expiration date in a widely cited Harvard Business Review article provocatively titled "The Sexiest Job of the 21st Century" (Davenport and Patil 2012). With this new innovation narrative sweeping the technology industry and beyond, LinkedIn reported machine learning engineer and data scientist lead the fastest growing job roles in the United States in 2017, followed closely by big data developer (#5) and director of data science (#8) (Bowley 2017). In early 2018, GlassDoor announced data scientist the "top job" in the US for the third year in a row, based on the number of job openings listed on their service, the salary levels reported by those with this job title, and these workers' overall satisfaction with their jobs (Glassdoor 2018). As the liquefied ground for what data count as evidence for innovation re-stabilizes, innovation-focused ethnographers working in/on/around/about the technology industry (and others) have reacted how we might expect from aggrieved qualitative social scientists; with denial (qualitative data matters!), with bargaining (we need both big data and thick data!); with testing (how do we combine quant and qual data in new ways?) and acceptance (it appears to be here to stay, so let's do what social scientists do best and cut it down to size by interrogating our assumptions, our language and what's actually possible now)[3]. Here, we add to these voices with a contribution to the far end of the acceptance literature, as we imagine the new concretized landscape and how silicon geography ethnographers will work there in the near and far futures.

We refer to the established technology innovation narrative as *the people-centric narrative* and the new one as *the data-centric narrative*. These are the **idealized stories** the industry tells itself and others about how it uses evidence in-and-of the world to create new products and services customers will value and pay for. As idealized stories, they are neither nuanced nor subtle and thus not true to how teams of diversely trained researchers often collaborate at technology companies, as we will illustrate later. While they grossly oversimplify how work actually happens, they do provide a useful framework to highlight differences in the underlying idealized *logic, methods, analysis, outputs* and *end goal* flow for innovation processes in the technology sector. The coexistence of these differences constitute the unstable ground silicon geography ethnographers currently find themselves occupying.

The people-centric narrative emphasizes the **logics** of opportunity areas for technology intervention and the existence of unmet needs, **methods** for their discovery through qualitative, ethnographic inquiry and **analysis** informed by social science theory with an **output** of implications for design then acted upon by product and strategy teams[4]. It is an unrelentingly human-centric model. The unmet needs may be of atomized individuals (consumers, workers) or of social and/or economic organizations (households, enterprise) but the **end goal** is always addressing the desires and needs of people. This narrative now co-exists, often in epistemological friction, with the new data-centric narrative of innovation. This narrative privileges the **logic** of powerful yet invisible patterns all around us, **methods** for their discovery through instrumenting the world to digitize the stubbornly analog and sucking up massive amounts of resulting data which are then combined with existing digital data. These data sets are then cleaned and readied for **analysis** by data scientists and computer scientists using new, increasingly complex and progressively more opaque algorithmic processes that then **output** models of the world that machines can use to directly understand and act in the world. The **end goal** in this narrative is to create compute systems that exhibit a specific behavior: artificial intelligence (AI). This innovation model is unrelentingly machine and data centric, with the people so important in the earlier narrative relegated to the edges; as (often unwitting) providers of behavioral inputs to feed the voracious data appetites of increasingly autonomous intelligent systems, and as subjects acted upon by such systems as they go about their daily activities — usually polarized between delighted (in corporate vision videos) or alarmed (in critiques of this narrative) (see O'Neil 2016).

This data-centric innovation narrative is part and parcel of the broader narrative about the fourth industrial revolution and confluence of big data, advanced analytics, machine learning and its subfield of computer vision, the Internet of Things (IoT) and other technologies that in combination enable machines to exhibit artificial intelligence, and that increasingly blur the lines between the physical, the digital and the biological (Schwab 2017). Various potential futures await earth's inhabitants, and we are challenged to decipher which voices in the popular debates around AI are Pollyannas, Chicken Littles, or Cassandras, and which are simply exploiting the current context in which "technologists, businesspeople, and journalists wield the idea like a magic wand that turns ordinary computer software and devices into world-saving (or world-ending) marvels" (Bogost 2017), simply to further their personal or company brand strategies.

The envisioned full expression of the fourth industrial revolution is the arrival of High Level Machine Intelligence (HLMI), "achieved when unaided machines can accomplish every task better and more cheaply than human workers" (Grace et al 2018), a clever definition as it is broad enough to be both meaningless (what's a task? how is cost measured?) and to genuinely alarm observers. A recent survey of machine learning experts that explored their beliefs about how quickly the world is progressing towards HLMI suggests (if the experts are right) that we have roughly 120 more years before AI can automate all human jobs. These surveyed experts expect AI systems to be able to tackle more and more ambitious analytical and creative tasks in the intervening century, such as folding laundry "as well as and as fast as the median human clothing store employee" (in 5.6 years), assembling any Lego (in 8.4 years), generating a top 40 Pop song (in 11.4 years) or writing a New York Times bestseller (in 33 years) (Grace et al 2018). Considering the work done by the average silicon geography ethnographer to be somewhere between writing a pop song and writing a best seller, we may have approximately another 20-25 years of work ahead of us before the machines takes over.

Of course, we are being facetious. We believe the future that technology industry ethnographers need to consider is not when and how sentient machines will take our jobs, enslave, or murder us. Rather, we need to consider new expectations for ethnographic observation and sensemaking in the next 20-25 years, as our work unfolds in the increasing presence of the type of analytical capabilities specially trained (and self-training) machines can do 'better' and 'cheaper' than humans as they can take in, analyze and model digital data at much higher volumes and with an attention to nuance not achievable through human cognition alone. As we examine these new expectations here, we will avoid clichéd tropes about the nature of work, Pollyanna, Chicken Little or Cassandra proclamations about artificial intelligence, as well as any mention of sentient robots. We will contemplate these new expectations by imagining and then dissecting very specific applications of machine learning, computer vision, and Internet of Things sensing and connectivity technologies to our current and past ethnographic projects. We will use these case studies to draw speculative conclusions about: (1) how data in-and-of-the world that drives tech innovation will be collected and analyzed, (2) how ethnographers will approach analysis and findings, and (3) how the evidence produced by ethnographers will be evaluated and validated. These imaginings and dissections will highlight the current frictions in approaches to making sense of the world between technology industry ethnographers and some of their workplace collaborators, stakeholders or audiences, namely: engineering, computer sciences and machine learning experts, as well as marketing teams and business leaders. Given their bases in long-standing disciplinary differences and the hype around big data and the 'magic wand' of AI, these epistemic frictions are unlikely to be resolved. However, as is the case with most frictions, we believe these epistemic ones will be productive. So before we turn to our imaginings, we will first consider three key differences in how we as technology industry ethnographers and our workplace collaborators approach observation and sensemaking of the world. These differences provide the context for how we imagine working in the world saturated with the types of sense and sensemaking technologies that our colleagues are creating.

EXISTING FRICTIONS AND THE RUMBLING INNOVATION LANDSCAPE

As technology industry ethnographers, we have endured repeated and sometimes arched-eyebrow questioning by our colleagues about how we observe and make sense of the world. We have had (our own!) sales people introduce us to customers as appetizers (tasty and insubstantial) before the main course (meaty and serious) of technology discussions. We've been informed that our work is fluffy (not important), and our data anecdotal (not evidence). We are familiar with the polite thanks for sharing our interesting (not useful) work at the conclusion of meetings. We've been told that what we do is simply descriptive observation (no theory needed), for which advanced training is not truly necessary and that we should/can be easily replaced with cheap interns. The common misconception behind these critiques is an assumption that we are pure empiricists dedicated solely to inductive reasoning — that we take a highly limited amount of data and reason directly from it to grand conclusions. The theory that informs our data collection and analysis strategies is unrecognized by our peers primarily for two reasons. First, because we eschew name dropping theorists and alienating stakeholders with inside baseball explanations of the theoretical foundations of our research decisions as a starting point for discussions of research findings[5]. Second, because the blend of abductive, deductive and inductive reasoning we favor differs substantially from established methods of observing and making sense of the world rooted in models of stating, testing, and rejecting or accepting hypotheses, and to the types of empirical data favored in the engineering and computer science disciplinary training of our colleagues. After recently presenting a thoughtfully considered multi-pronged research program incorporating very specific proposed research projects to identify near, mid and long term strategic and tactical opportunities for a business line, we were criticized for not understanding how research works (despite our Ph.Ds.). A primary stakeholder corrected our approach, offering that in order to do research we needed to identify hypotheses to test, and proceeded to give us several examples of how to correctly formulate a testable hypothesis.

In full disclosure, these occasionally disheartening — but mostly amusing — encounters are generally with internal stakeholders removed from day-to-day R & D work, and not the computer scientists and engineers we work with closely and for extended periods of time on innovation projects. Turning to these latter colleagues, we recognize some fundamental differences in the assumptions we bring to creating and using data in and of the world to inform innovation. The frictions our differing assumptions create has been overwhelmingly productive — getting us to new questions and new knowledge. We will use examples from one project, *Local Experiences of Automobility* (LEAM for short) that we will revisit later as one of our future case studies, to illustrate these frictions. The LEAM team, led by one of the authors, consisted of anthropologists, an electrical engineer, a psychologist, computer scientists, and a data visualization intern, and is an excellent example of close collaboration between colleagues with different disciplinary training and of the epistemological frictions that ensued and that will likely be exacerbated in coming years as the ground concretizes around data-centric innovation.

You Are Giving Us Too Many Problems

When we started LEAM in 2010, our idea of a suitable ethnographic research question for exploring the future of autonomous vehicles was simple, open-ended, and broad: *what is a car?* This left us open to a number of research methods and tools that would generate many types of data to identify what we called opportunity areas for the development and application of new technology solutions for road-based transportation. As social scientists we assumed our job was to generate data to identify *opportunities* to make change in the world. As computer scientists and engineers, our colleagues assumed their job was to generate and use data to *create* change in the world. One teammate in particular was fond of reminding us:

> Engineers work in the solution space. We like to solve problems. Anthropologists like to find new problems, even when we are still working to solve the ones you gave us last week. You are giving us too many problems. [6]

Her idea of an appropriate question was: *40% of the build of materials cost of a car is sensors and electronics. What if the sensors were so good that a car couldn't be hit? Then it could be an egg – no airbags, no crash frame, and no bumper.* This approach enabled her to focus specifically on increasing the value derived from ingredients currently going into a solution — solving the problem of justifying the BOM (build of materials) cost of a car devoted to sensors, by decreasing the amount devoted to bending metal. As a team, we overcame this friction by learning to announce at the beginning of activities which mode we were in: problem or solution, and by prioritizing the problems we identified so as not to overwhelm our team.

This guy can't be trusted! This data is no good. Throw it out!

As the lineup of researchers was still forming in the earliest days of LEAM, we had team members each review a single ethnographic interview transcript with a car owner as a means of building empathy and awareness of current automobility practices. Upon meeting to talk about what we had read, one of our teammates, a computer scientist highly skilled in machine learning, quickly pointed out that the participant in his transcript contradicted himself in describing how he uses his car. He then dramatically threw his hands in the air and declared: "This guy can't be trusted! This data is no good! Throw it out!" For him, the data was bad; how could he build a model on logically inconsistent data? In both ethnographic analysis and in building machine learning solutions, data needs to be assessed, cleaned and criteria established to determine what data to include and what not to include in the analysis. In this case, our computer scientist teammate applied the rules of his training to a data type that didn't follow the validity criteria needed to build a predictive model. Our inability to convince him that such inconsistencies did not invalidate the data, they merely suggested a different type of analysis than he was accustomed to, meant his stint as on the team ended there.

Excellent! We'll get to ground truth

When we started scoping the research methods for LEAM, we recognized that because we were after the mundane, taken-for-granted details of owning, using, caring for and being around private passenger cars, interviews alone were insufficient. We settled on a mix of methods that we have described in detail elsewhere (Zafiroglu, Healey, Plowman 2012), and that we will briefly summarize here as semi-structured interviews and 'carchaeology' inventories of objects in and on vehicles on the one hand (getting us to 'Remembered Drives') and 30-day collections of GPS data, mobile phone use during and around drives, and car-interior sensing of sound pressure, lighting, accelerometer, temperature (getting us to 'Recorded Drives').

Feeling confident that these methods would allow us to create thick data from two thin sources; the after-the-fact remembrances and explanations provided in interviews and during carchaeology sessions, and the machine-produced numbers and patterns from the sensors, GPS and phone monitoring app, we were taken by surprise by the interpretation of our methods by our electrical engineer team mate. Referring to the Recorded Drives data set, she proclaimed her excitement that we would finally get to 'ground truth' about how people were using their cars and their phones while in their cars. Concerned (like our erstwhile computer scientist team member) that people are inconsistent and lie when they talk to us, she argued the machine data would provide better evidence for how people *actually* used their phones. We could, in fact, catch participants when they lie! This was a goal the ethnographers hadn't even imagined. What ensued were months of animated discussions about 'truth' in research data and how we create and trust data created through our interactions with people vs. people's interactions with machines. Our colleague's confidence in the empirical objectivity of digital traces created by people through their interactions with technology to get us closer to what we were trying to measure — to ground truth — was never truly reconciled with the social science trained ethnographers keen interest in the multiple truths people produce and live in simultaneously. A data point from a participant in China drove home our different takes on evidence. Our GPS and phone monitoring data showed the participant using the camera function on his mobile phone during his commute to work while travelling at relatively high speeds. When we queried him about taking pictures while driving, he explained he was merely placing his phone in a holder attached to his dashboard and as the camera button was on the side of the phone, he sometimes inadvertently took a picture. For our computer science and engineering teammates, the data point was errant; while it did get us closer to the ground truth of when a phone was handled while driving, it was incorrectly labeled by our phone monitoring software and therefore as currently labeled would not be useful for training a machine learning model. For the ethnographers, the phone monitoring data point wasn't errant; it was simply data about a mundane detail of phone and car use the participant would not have thought to mention to us and that we would not have known to ask about without the machine data. Of these three frictions, this was the most difficult for us to reconcile as a team, as half us wanted 'ground truth' data useful for training a machine learning algorithm and the other half was just being exposed to the idea of machine learning and the data rules it requires.

Past the Shudders, Jostles and Rolls and Onward to Liquefaction

In addition to these long standing frictions between social science and computer science and engineering research methodologies and understandings of evidence, in recent years we've lived through the transformation of our company, like so many others, to being data-driven. 'Data-driven' attests not to the type of evidence produced by qualitative ethnographic research, but by big data, advanced analytics and IoT capabilities that are creating the massive amounts of data that drive the new data-centric innovation model. In a recent keynote for a "Data-Centric Innovation Summit" hosted by our company, an executive vice president presented a "vision for a new era of data-centric computing" and explained the opportunity before us as the biggest TAM (total available market) growth opportunity in our company's history.

> I think one of the most stunning statistics is that 90% of the world's data has been created in the last two years and even more stunning perhaps is that only about 1% of that data is being utilized to create any sort of meaningful business value. I see tons of room for our industry to grow. (Intel 2018a)

This VP's enthusiasm succinctly illustrates the ascendency of empiricist models for innovation fueled by the increasing generation and availability of big data in business. This is a compelling model, if as Kitchin argues, those in business believe "the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory" (2014, 3). With the adoption of the data-centric innovation narrative, the evidence needed to fuel innovation has shifted from being *about people and their lived experiences* to being *digital traces created by people*, often at very high scales and from multiple sources. The outcome, as Metcalf and Crawford argue, is "the familiar human subject is largely invisible or irrelevant to data science" (2016, 3). The vagaries and dissimulations of the familiar human subject no longer matter if the focus is on the (imagined) veracity of digital traces. If researchers focus on *digital traces* that are imaged to be objective and capable of speaking for themselves then, (for example) a contracting eyelid is empirically a contracting eyelid that can be analyzed using new data analytics methods; there is no need to engage with the whole of the person attached to said lid. In the age of computer vision and big data Geertz's argument for thick description becomes particularly troubling for the ethnographer.

> Consider . . . two boys rapidly contracting the eyelids of their right eyes. In one, this is an involuntary twitch; in the other, a conspiratorial signal to a friend. The two movements are, as movements, identical; from an I-am-a-camera, "phenomenalistic" observation of them alone, one could not tell which was a twitch and which was wink, or indeed whether both or either was twitch or wink. Yet the difference, however unphotographable, between a twitch and a wink is vast; as anyone unfortunate enough to have had the first taken for the second knows. (Geertz 1973, 6)

In the 45 years since Geertz wrote about "I-am-camera" observation, the world has profoundly changed; such differences may no longer be "unphotographable". We can now imagine a world in which a camera connected to computer vision capabilities can distinguish a wink from a blink, if trained with a sufficiently large data set of humans contracting the eyelids of their right eyes. While conceivable, we are still left wondering *why,* both in terms of the significance of the contraction (a sign of disease? a sign of playfulness? serious ill intent?) and in terms of what the return on investment might be for developing and training such a model (who would pay for it and why?). In other words, as Kitchin notes, "while data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is likely to be anemic or unhelpful as it lacks embedding in wider debate and knowledge."(Kitchin 2014, 5)

And yet, in our work environment as technology industry ethnographers, there exists tremendous confidence in AI systems built on computer vision and internet of things sensing that promise to model and interpret the cyber-physical world and to automate decision making and actions with limited human contextual and domain-specific expertise needed beyond the initial set up of the systems by data and computer scientists. Big data, advanced analytics, AI and IoT promise to simplify and speed up innovation, by creating a "new mode of science, one in which the modus operandi is purely inductive in nature" (Kitchin 2014: 4). Our colleagues believe this and our leaders espouse the primacy of data and such solutions to our future financial success.

Regardless of our clear understanding of why such a purely inductive science is preposterous[7], as ethnographers our work will be increasingly situated and carried out in contexts where AI, computer vision and machine learning algorithms are constantly sensing and building models of people, activities and objects, that feed services which will significantly reconfigure our social, political, and economic activities and our daily behaviors and interactions. In the not-so-distant future, we will work, move through, live, shop, recreate and more in environments where machines will create models independent of human interpretation to inference what's happened, what's happening, and what's *likely* about to happen, or to *create* what happens by arranging local conditions towards outcomes desired by those deploying such tools. As Paglen argues, those deploying computer vision and related tools will be able to "exercise power on dramatically larger and smaller scales than has ever been possible", seemingly objectively as the ideological foundations of the algorithms informing interpretations of images "function on an invisible plane and are not dependent on a human seeing-subject"(Paglen, 2016). 2017 news coverage of toilet paper dispenser kiosks using facial recognition to limit patrons to nine sheets of tissue each per 15 minutes at Beijing's Temple of Heaven is an excellent example of one such exercise of power.

As ethnographers employed in the technology industry, we face a challenge and responsibility as we may both be *acted upon* by such systems and our work (ideally) will *shape* how these systems and data exercise power by contributing to how they are created, managed, updated, and scrapped. Considering how we may be *acted upon*, the presence of these systems in field sites and in our work environments means we face existential and tactical questions about our practice. The growing literature on the existential questions addressed pressing concerns including: are our skills still needed and valued for innovation (Madsbjerg 2017)? Can our employers replace us outright with algorithms and AI systems and a few data scientists (until the machines become 'good enough'?) Does the technology industry need fewer ethnographers and more human factors engineers who study how experts will build such systems? Or who study how their business customers will interact with the outputs of these systems? What are the burning questions that will keep ethnographers relevant and employed (boyd and Crawford 2012)?

In the remainder of this paper we will focus on tactical questions we face as ethnographers being acted upon by the technologies of the fourth industrial revolution, including: How will AI and machine learning applications such as computer vision, text analytics, and speech understanding reshape how we collect and analyze data in-and-of-the world that drives tech innovation? What data will and will not count as evidence? How will the evidence produced by ethnographers be evaluated and validated by our colleagues and by our employers potentially using these self-same technologies and tools? In short, how will expectations for ethnographic practice in the technology industry change in the next 20-25 years?

ETHNOGRAPHY ON THE (NEW CONCRETIZED) GROUND

To answer these questions requires a bit of imagination, but not a full flight of fancy into Chicken Little proclamations about sentient job-killing robots that we promised to avoid. Our imagination of the future needs grounding in our current ethnographic practices and the technology capabilities at our disposal today and in the foreseeable future. Therefore, we will cut down the abstract and idealized fourth industrial revolution and AI into more realistic yet still ambitious applications of machine learning, computer vision, natural language and audio processing, and Internet of Things sensing and connectivity technologies that we then imagine included in three real (not idealized) ethnography projects we have undertaken in the tech industry.

We will avoid declarations about the exact timing of the blanketing of the earth in these new sense and sensemaking technologies, and instead imagine each project in a *low*, or a *medium* or a *high* presence of advanced sensing and sensemaking capabilities present in our research settings and in our work settings (i.e. where we analyze our data).

For each case study, we will describe the original project, our methods, the data we produced and the project outcomes. We will then re-imagine the project with new combinations of sensing and sensemaking capabilities in our research and data analysis locations and explore how adding these would have first order and second order consequences. Here, first order consequences refer to what will change in ethnographers' research practices and methods; how we will approach sense and sensemaking in our research and how data will and be created to achieve our research goals. In contrast, we define second order consequences here as how professional expectations for technology industry ethnographers will change; what may be new expectations for future ethnographers' responsibilities and scope of work, and new professional standards for training and for how our work is evaluated and validated. We will avoid long discussions on the mechanics of our access to such new data. We will assume that access to research participants' own data will be arranged with and consented to directly by them, and access to some subsets and/or versions of broader public or private data sets (such as security camera data from a housing complex; utility usage data for a neighborhood) will be arranged with the data set owners, and the uses to which we put these data sets will follow the ethical guidelines of the American Anthropological Association, and the privacy requirements and research approval process of our employer, which jointly define our current research practice.

Table 1. Imagined Levels of Advanced Sense and Sense-Making Capabilities

Presence of Advanced Sense and Sensemaking Capabilities	In Field Locations: Data Creation, Data Access and Sense Making Capabilities	At Ethnographers' Office Locations: Tools for Sensemaking	Interaction Between Computing Systems and Ethnographer
Low	Access to historic data such as utility usage, home access systems as provided by participants or by service providers	Access to computer vision, big data analytics, speech and audio detection and interpretation systems that we use to analyze the field data we have already created, our participants have already created, and stored data already generated by others (utility usage, for example)	Computing systems alert ethnographers to patterns or anomalies within field data, and in field data within context of larger data sets
Medium	Close to real-time access to existing data being collected and analyzed by service providers and other actors. Ethnographers' field equipment has machine learning and computer vision capabilities that can act on data as it is created and present analysis to researchers.	Same as low level	Computing systems proactively suggest topics and follow up questions to ethnographers based on patterns and anomalies in field data.
High	New sensors, devices, connectivity and computing solutions that create models of environment and can respond with real-world action to local conditions and to human or machine commands	Access to advanced data analysis and research management systems	Computing systems generate protocols and questions based on priorities and inputs from ethnographers, and present for approval; computing systems automate some interactions with research participants

Smart Home Economics (SHE) and 21st Century Homemaking

In 2014, our strategists and product development teams considered home security the entry point for the smart home market, from which we could add on home automation capabilities such as automating locks, lighting, and HVAC systems. After security and basic home automation, what other usages would be valued? As ethnographers, we set out to think beyond security and simple automation to other experiences that could be possible, appropriate, and valued in homes equipped with new IoT sensing and sensemaking through field visits with householders and with home services professionals. In Smart Home Economics (SHE) we explored possibilities around practices of caring for the physical structure of a home and around managing a household.

Our field methods reflected our goal of thinking beyond incremental improvements to existing home security products. We explored *homemaking* activities at two scales. At a large scale: how are homes purchased/leased/otherwise come to be occupied, finished, maintained and updated? How do people come to live where they do? What processes are involved? At a small scale: what are considered normal and necessary daily practices in homes and how are they achieved through people, devices, systems, expectations and conventions? We visited eight households each in greater Shanghai and greater Atlanta. We spent an initial three hours interviewing in homes, then had participants answer personalized follow up questions via a smart phone app with which they took videos and pictures and answered questions. We then revisited them a week later and reviewed their follow up answers. These interviews were in English or a combination of English and Mandarin (using a translator) and were video recorded. In parallel, we conducted in-person interviews with home services professionals about their work practices, skills, perspectives on what was changing in their markets, and their outlooks on how digital technologies were impacting their work. These experts included: real estate agents, property managers for luxury, non-luxury apartment, and other housing complexes, home inspectors, building and home automation solution designers and installers, home and building security specialists and installers, security guards for apartment complexes, construction managers, interior designers, and architects. These interviews were in English or Mandarin. Atlanta expert interviews were video recorded and Shanghai expert interviews audio-recorded.

SHE resulted in a data-informed critique of then current home automation and security products. We contended such products were chasing the tail end of the 20th century by seeking to further digitize homemaking practices that had already been almost fully automated in the past two centuries. Have a vacuum? Make it robotic! Have a dishwasher or washing machine? Connect it to the internet! Have electric lights? Control them from afar! We argued to truly think beyond relatively small tweaks to previously achieved dramatic improvements in homemaking practices, 21st century smart home solutions must position householders for domesticity in a world with very different environmental, political, social and economic contexts than the previous century. We offered experiential statements for domestic life in these new contexts and different priorities around security or automation usages responsive to new concerns and situated in broader networks of services, systems and actors within and beyond the home. We followed SHE with SHIFT (Smart Home Information Flow Technologies), an ethnographic project on householders' expectations and preferences for sharing information about daily homemaking activities with others[8]. SHE and SHIFT were central to the comprehensive Smart Home usage roadmap we authored that our business and technology teams relied on extensively when defining new product capabilities for the smart home market.

SHE in a Low Sense and Sensemaking World

Let's now reimagine SHE with a low presence of advanced sensing and sense making technologies in the homes and work locations we visited, and in our own work settings as we analyze the data. In this alternative reality, we imagine 'low' presence in *our field locations* simply means that we have some level of access to stored (not real time) data already being generated by other actors that we didn't have in 2014, such as:

- video and audio files from apartment complexes, residential neighborhoods and individual home security systems.

- event logs from security, home access and communication systems, such as when key passes were used to enter a gated community, or park a car in a garage, or when calls were made to an apartment from an entry call box.
- Current and historic logs of utility usage in homes; patterns of electricity, heating, gas, water consumption as captured by service providers for our participants and for the areas they live in (an apartment complex; a city; a region)

We imagine as we analyze the research data in *our own work settings*, 'low' presence means we have access to computer vision, big data analytics, speech and audio detection and interpretation systems that we could use to analyze the field data we have already created in person (our interview video and audio recordings, our still images, our audio field notes), our participants have already created (video and text data created through the smart phone app) and the historic utility use, security footage and event log data from home settings we now have access to. In other words as the level of sense and sensemaking technologies are low, we assume we will *not* have real-time access to any conditions, actions or events happening in our field locations, we will simply be able to analyze our existing data differently after it has been digitally collected and stored.

So, what might meaningfully change in terms of first order consequences, i.e. how we conduct research? Given that we will not have real time access to machine data, the main differences in how we sense and sensemake will happen as we review and analyze existing data types in office. Currently, the audio and visual recordings we create and our participants create on our behalf are primarily useful as a way to supplement and extend *our human powers of observation and memory* of events, activities, interactions, locations, objects and actors. We review images and videos and audio transcripts in order to recall details we may have failed to originally observe or simply forgotten between visits and analysis days or weeks later. How was that living room arranged? What exactly did that broken refrigerator ice maker work around look like? Were there informative turns of phrase we didn't notice during an interview? An off-hand comment that we didn't quite grasp or didn't realize was noteworthy (pun intended) at the time but on review significantly informs our analysis? If we were able to analyze this data using computing systems trained to detect patterns or anomalies, these digital artifacts would shift from simple observation and memory aids *for us* to potentially powerful tools to observe at nuances we can't detect and to sensemake using massive memory of other events we do not personally possess. With trained models, we could open our field data to new machine interrogation which might be able to detect patterns or significant events we captured but failed to notice. In other words, the trained machines may be able to see what we couldn't in our videos, or hear what we couldn't in our recordings. In the majority of current applications of computer vision analysis of video in law enforcement and commercial security services, such use is forensic; seeking after-the-fact evidence of a crime committed. Here we might use forensic in a different way to call attention to the 'crime' of leaving data collected by ethnographers unnoticed or unanalyzed. We imagine the following five ways we as ethnographers may forensically employ AI systems in office to interrogate our field data.

We See In Office What We Couldn't See While In Field – Using computer vision enabled computing systems in our offices, we could be alerted to objects, movements, behaviors, activities that we were not attuned to in field and still do not or cannot notice later when we review images and videos. We imagine our office systems producing reports on field work video footage and still images that provide descriptions or categorizations of people, objects, and conditions using criteria not available to the unaided human eye alone. Concerning people, this could include such things as patterns in facial expressions (perhaps glossed as 'emotions'), body posture, and the body language of participants and of researchers. For objects, this could include recognizing objects or arrangements of objects in the environment that are common or uncommon to a larger household demographic, or to create novel classifications of household types based on visual evidence of homemaking practices that we might not have reached on our own (or not reached as quickly). For environmental conditions we cannot visually analyze with precision on our own, this could include evaluations of indoor and outdoor air quality based on our video or still images (see Zhang et al 2016) or standardized evaluations of patterns in lighting practices and conditions in homes using metrics valued by our engineering colleagues. We could achieve the attention to the physical environment Reichenbach and Wesolkowska (2008) argue is often missing in ethnographic research.

We Hear In Office What We Were Not Attuned To While In Field – We imagine running our video and audio through machine learning audio software that analyzes non-speech sounds that we didn't notice, that we tuned out over the course of fieldwork, or that were inadvertently captured on video recordings produced for us by participants. These could include inside-the-home and outside-the-home soundscapes of domestic life, from appliances or consumer electronics running, to heating and cooling systems turning on and off, to neighbors moving about, to traffic, garbage pick up, deliveries, or grandmas blasting music while practicing Tai Chi in the condo complex courtyard. Such software could listen and classify these sounds for us, and are not so fanciful given current product offerings from companies such as Audio Analytic, SoundHound and several others. Reports on these sounds could turn unrecognized sounds into recognized ones, and could spark us to ask new questions about homemaking. While beyond the capabilities of current audio AI product offerings, we know university researchers have shown how computer vision analysis of small movements of inanimate objects (a potted plant, an empty potato chip bag) in a silent video can be independently used to recreate the audio (including speech) occurring when the video was created (Feltman 2014). We imagine running existing video footage from security cameras, which often don't include audio, through computer vision algorithms to identify audio events and patterns around home exteriors — from multi-tenant dwellings to single family homes — that could spur us to ask more informed questions about home security conditions and practices.

We Understand Speech That We Didn't Understand In Field – We imagine running our audio (or video) files through AI language translation and analysis software that can flag points in conversations where misunderstandings may have occurred, so we can decide to explore more with participants in the follow up interviews. Such language translation and analysis capabilities would be useful between completely different languages (Mandarin and English) and between dialects and slang within a language (Atlanta/SE United States English and West Coast English). We may be alerted to regional accents we don't recognize, or to idioms we don't recognize as having local significance. We can also imagine needing to train the software to recognize intentional miscommunications; when a topic is skirted to save face, or because it is too personal, or a follow up question is not needed as the people present already recognized and sorted the miscommunication non-verbally.

***We Contextualize Our Data In Ways Satisfactory To Our Colleagues* –** The memory used by machine learning systems (the sets of data on which a machine learning algorithm are trained) is much larger than we as humans can retain and recall. Given we will have access to home utility data, as well as audio/video we have captured, we imagine that our machine learning tools can help us understand if a research participant is representative of a larger group of householders. Is the amount and pattern of utility use of this household typical or unusual for their neighborhood or complex? Or an extreme in some way? Using computer vision capabilities, could we understand if there is an object present, a home decorating style, a pattern in the arrangement of objects, or the range of objects in this home that have social, cultural, economic, or political significance in the field locale? As researchers, we value both participants that represent larger groups as well as extremes in behavior, product ownership, income, etc. The machine learning software will help us better contextualize who we have included in our study.

***We Extend The Usefulness of Ethnographic Data Over Time* –** In addition to subjecting the data from SHE to analysis by machines, we may consider adding the SHE data to the broader data set that feeds the machine learning models we've used in the study. This of course raises a number of questions about how we do this, for which we will need to collaborate closely with data scientists. Beyond the obvious difficulties of cleaning, structuring and properly labeling the data, we face practical questions about data handling. How do we write a consent form for participants that encompasses use in training models? How do we explain who has access to the raw data in the future? Currently, we specify "only the immediate research team". Will we need to ask for consent for additional researchers? For machines? What might that look like? Furthermore, several years later, would we want to — and legally and ethically could we — re-run the data from SHE through updated machine learning tools to see if other insights are possible with retrained and updated AI tools?

These new in-office analysis capabilities will create novel data artifacts in the form of text-based reports that we will treat similarly to our existing field notes, field audio transcripts, and image data; as another input to be analyzed. They won't replace the type of social theory informed analysis that we current undertake, but they will change two aspects of our research practice. First, we will extend our ability to identify areas of interest based on nuance and scale of observation that we did not have before. Second, we will create richer, more scientifically precise descriptions of the locations for which we are designing new technologies that we can share with product development teams.

Given these outcomes, what might meaningfully change in terms of second order consequences, i.e. professional expectations for technology industry ethnographers? We see three potential outcomes.

***Ethnographers Face New Expectations For Proving Data Validity* –** We will now have better means to explain who our research participants are. With analysis of our participants demographics and behaviors in comparison to larger data sets we can get closer to ground truth with our internal engineering trained research audiences about how representative, or how unique, our participant sample is. We can now better explain and support our decisions for who we have included in our studies, and our data are less likely to be doubted as anecdotal by our colleagues.

Ethnographers Face New Expectations To Wring More Insight From The Same Type and Amount of Qualitative Data – Because we can now perform forensic data analysis on our standard digital research files that extends the scale at which we can analyze our data and the nuance in observation beyond our human sense and sense-making abilities, the type and breadth of deliverables from a single project will likely increase. As a simple example, the types of data amenable to creating 'ground truth' to train algorithms by our computer science colleagues will be expected as part of research findings. (how well-lighted are American vs. Chinese homes?)

New Professional Standards For Ethnographers' Skills And Fluency In Engaging and Contesting Machine Data – Ethnographers will be expected to be able to parse the types of reports produced by machine learning systems. Even if the front end UI available for accessing the analysis on our work machines resembles those used by consumer wearable or smart home services, ethnographers will need be able to look behind the UI and be able to understand enough of how the analysis was performed to contest or at least understand how the machines came to a conclusion. This will particularly be an important skill if the ethnographer independently reaches a different conclusion than the trained machine, and the ethnographer needs to give input on how the training models should be updated.

Local Experiences of Automobility (LEAM) and the Future of Transportation

Imagining a bit more intense presence of sense and sensemaking technologies in our field and work environments we return to Local Experience of Automobility (LEAM) from which we drew some of our earlier examples of disciplinary frictions. The ultimate goal of the research was to prepare Intel to design vehicle and transportation system solutions as we entered a decade of transformation of cars, road infrastructure and ecosystems through advanced sensing technologies, computational systems and services. While we did several rounds of research in seven countries, the richest methodology was used with car owners in two cities each in Brazil, China and Germany. With these participants, we completed car inventories, semi-structured interviews, video diaries and 30 days of car use data including GPS data, car-interior sensing of sound pressure, lighting accelerometer, air temperature and use of mobile phones before, during and after drives. Participants were visited once for an initial three hour interview, car inventory and ride-along, and to have sensors and GPS installed in their car, and tracking software installed on their phones for thirty days. They were revisited again approximately 6-8 week later to review the phone and GPS data the research team had visualized using Google Maps.

The outcomes of LEAM included the generation of over three hundred use cases for in-vehicle infotainment systems, advanced driver assistance systems and semi-autonomous driving solutions. The project was notable because it created a direct tie between foundational qualitative research and product definition, as well as generating over forty awarded patents, and multiple internal prototypes and projects with car manufacturers.

LEAM in Medium Sense and Sensemaking World

Let's now imagine LEAM with a medium presence of advanced sensing and sense making technologies in our research equipment, in the cars and the road systems and transportation infrastructure in the six cities we visited in China, Germany, and Brazil, and in our office settings as we analyzed the ethnographic data. In this alternative reality, medium presence in *our field locations* might mean that in addition to the 'low' presence capabilities of the last case study, we have close to real-time access to existing data that is being collected and analyzed by other actors such as:

- Current road conditions and traffic patterns
- City-level mobile phone location and use data, including phones being used in moving and still vehicles
- Histories of a research participants' cars presence on and use of roads during the past 6 months (based on phone data or transponders for road fees), presented on digital maps, and anonymized comparisons to other car owners and averages in the municipal area

And it might mean that the cameras and audio recording systems we bring with us to the field incorporate computer vision and machine learning capabilities that act on or analyze data as we are create it and present analysis to us by showing us patterns or alerting us to anomalies. We imagine as we analyze the research data in *our own work settings*, in addition to computer vision, big data analytics and speech detection and interpretation systems that we could use to analyze the field data, we have means to realistically (not 'effortlessly', but not so difficult as to be not worth the effort needed) combine our time-and-place-specific data with time-and-space congruent 'big data' sets such as traffic conditions and social media postings.

So, what might meaningfully change in terms of first order consequences, i.e. how we conduct research? Given that we will have real time access to machine data, the main differences in how we sense and sense-make shift from happening exclusively after the fact/in the office, to a mix of in real time/in the field and in office. We imagined two significant outcomes for our field research practices.

***We See and Hear in Field What We Couldn't Before* –** Observation and sensemaking by machines move closer together in time and space, with real-time, or close to real-time, pattern and anomaly detection happening in the field on our data capturing equipment rather than after the fact in our offices. We imagine this means our field equipment will alert us to movements, behaviors, activities, environmental conditions, and to the presence or arrangement of objects or to turns of phrase, language miscommunications and non-speech audio events that seem significant in the moment. In addition to simply noting and alerting us to patterns or anomalies, our equipment might also suggest actions to take such as a question to ask about an object as we unpack a car, a rearrangement of objects on our sorting sheet as we inventory the contents of a car, or an data analysis point and a suggested follow up question such as: *83.5% of the car contents belong to Fernando, but Ana Luisa is the primary user of this car: how did this come to be?* The frequency and intensity of these alerts and suggestions will likely increase as a given interview progresses, and over the course of a project as the number of interviews we complete increases, as the equipment could be adding to its knowledge set over time depending on how learning and inferencing are architected.

Interviews Are More Exhausting for Ethnographers – Currently during fieldwork we are intensely and actively listening to participants and noticing everything around us, even as we formulate our next question and continually recalibrate the overall flow of a conversation with participants. This requires intense focus and is, frankly, exhausting. Indeed, on first exposure to field work during LEAM, our electrical engineer teammate marveled that the ethnographers could keep an enthusiastic conversation going with a participant for as long as we normally did. With sense and sensemaking smart machines in field with us, such work will require we expertly handle and incorporate an extra stream of information coming at us, and fluidly incorporate it into our orchestration of conversation and observation.

In our office locations, as we analyze data between field visits and at the conclusion of the data collection, we see one significant change.

We Better Distinguish and Flag Possible Differences in Causality of Data Patterns – We imagine being able to combine our GPS/telemetry or other machine-created data with other data sets so that we can understand our individual participant actions in the context of larger time- and space-specific events to highlight possible connections. In our original fieldwork, we almost missed an important story in Brazil when we mistook the GPS and phone data from a participant to indicate she had trouble parking (she made a phone call and she drove very slowly in a meandering pattern through a neighborhood). We brought biases to our interpretation of the data from similar data patterns we had seen elsewhere in Brazil and in Germany, and commented to the participant that her data from one Friday night seemed to show a hoh-hum evening of looking for a parking spot. The participant corrected us and explained that she had been alerted that night to an *arrestão* in the neighborhood (a group of criminals moving through an area and robbing everyone they encounter) by a phone call from a friend, she heard gunshots in the distance, and spent a frantic twenty minutes moving slowly out of the area hoping not to be robbed. Had we been able to run our data against social media postings or police reports from that neighborhood at that time, we could have been alerted to follow up with the participant in the second interview. We would, as well, get closer to the 'ground truth' our colleague imagined machines could detect.

Combined, these new field and office based sensing and sensemaking capabilities will shape how we create data and how we quickly or frequently we iterate our interviews or other research protocols. Given these outcomes, what might meaningfully change in terms of second order consequences, i.e. professional expectations for technology industry ethnographers? We see three additional outcomes beyond those outlined in the low level environments.

New Professional Training and Expectations For Handling Smart Equipment, Information, and Analysis Coming at the Ethnographer Real-Time in the Field – We expect such training opportunities will likely show up first in professional organizations, such as EPIC, where professionals can update their already expert ethnography skills. Moreover, we don't imagine that the presence of sense and sensemaking equipment in field means companies can hire 'cheap interns' as the human work is reduced to description and guided by machines; rather such skilled human work will require more training and more experience to deftly combine human questions and analysis and suggestions from equipment.

***New Expectations for Fluency in Collaborating with Data Scientists and Computer Scientists* –** Underlying the examples of smart equipment we've given is an assumption that the software on the equipment is trained to analyze and to learn over time. We do not believe it likely that all ethnographers will be fluent in the broad range of intense work that goes into building AI solutions: collecting data, creating a model, tuning it, implementing it on equipment and maintaining the software and hardware over time. They will, however, be expected to know how to effectively partner with experts who can do such work and to work together to define the expected outputs from the smart equipment. Ethnographers will need some ability or sensitivity to 'think like a computer or data scientist' just as we currently have some sensitivity and ability to 'think like an engineer' (and not give them too many problems). We need to be able to conceptualize: how can we translate what we need into data requirements that our colleagues can use when building a solution to train a machine?

***New teams, New Research Protocols and New Standards of Data Analysis* –** Tying together the first two secondary outcomes, is the larger employment context in which technology industry ethnographers will work. In a world with mid-level presence of advanced sensing and sense making capabilities in our research and work settings, ethnographers will need to work with computer and data scientists *before* fieldwork, as they collaboratively scope projects, research goals and protocols. Much as ethnographers made a shift in the 1990s and early 2000s to conducting 'digital ethnography', in the 2020s technology industry ethnographers will shift to truly working in, as well as studying, cyber-physical worlds.

Home Instrumentation and Sensing Study (HISS)

Our last case study will be our most extreme, as we imagine adding a high level of machine sensing and sensemaking into our field and office locations. For our Home Instrumentation and Sensing Study (HISS) completed three months ago, our goal was to revamp our existing Smart Home usage roadmap to explicitly include householders' domestic lives with a new roommate: artificial intelligence. Quite a bit has changed in three years: Intel now describes the Smart Home on our external corporate website as "perceptive, responsive and autonomous", a three-adjective shorthand for enabled by artificial intelligence. A vision video on the same site titled "Smart Homes Are like Us: real-time collecting | analyzing | diagnosing" illustrates usages requiring advanced sense and sensemaking capabilities working near real time in the home. (Intel 2018b). In our offices, we repeatedly find ourselves in conversations with computer scientists, engineers and other product team members who make assumptions about 'always on' sensing in homes, often through cameras (coupled with computer vision algorithms and other capabilities) and microphones (coupled with automatic speech recognition, natural language processing, acoustic event detection and other capabilities). As social scientists, we find these conversations alarming and fascinating

— in other words, urgently in need of data in-and-of-the-world to validate or disprove the assumptions being made about the necessity and desirability of always-on sensing in homes.

With the HISS research protocol we honed in on exactly how, where, when and why householders might (and might not) want their homes to be perceptive, responsive and autonomous although we kept our research protocol light-hearted and free of such jargon or any mention of machine learning, computer vision, artificial intelligence or even the term smart home. We will not describe the entire protocol here, but rather limit ourselves to the parts that could be most radically altered if a high level of advanced sense and sensemaking systems were already widely present in American homes.

In HISS, we used a smart phone ethnography tool to have 101 US householders create a corpus of scenarios for living in imaginary versions of their own homes, specifically ones instrumented with sensing and inferencing capabilities that could make novel experiences possible. Participants started with two assignments that allowed us to understand what they valued in their homes now. With videos and text answers, participants shared details around the three areas of their homes in which they spend the most time awake. They then detailed two changes they would make to their homes if they won a "Complete Home Makeover With an Unlimited Budget", and explained why they chose each change and how each would affect their home lives. We then shifted participants to a series of five assignments that playfully teased out householders' overall expectations for the experience of living in an AI-enabled home, where and when they might want to partner with AI to a specific end, how an AI enabled home should know what's happening in and around it, and how they expect an AI home should function, including their expectations for the accuracy and consistency of the sensing and inferences made. How did we accomplish this tall order? We asked participants to imagine winning a "free upgrade" to their home makeover so that their homes could always-and-all-over be able to sense and make sense based on one input modality. First they imagined life in a "smell-o-matic" version of their homes, then, (alternatively, and in turn) "sight-o-matic", "touch-o-matic", "hear-o-matic" or "taste-o-matic" versions. In each assignment, householders shared how they use their human sense at home now ("I walk in the door and smell dinner cooking and I know we'll be eating kimchi stew for dinner"), acted out in video what it would be like to live in a sensing home, and then in text gave us an additional three "if this than that" style scenarios in which their home senses objects, actors or activities, then infers and acts to change their home lives. We also asked participants for a scenario in which sensing in their homes would not be welcome[9]. Moving participants through the staged protocol, checking the quality and completeness of their answers, asking clarifying questions, or for re-dos was a full time job (weekends, evening, and holidays) for us. We felt, at times, overwhelmed by the data coming in and the need to keep participants moving.

As we only recently finished the data analysis, the concrete outcomes of HISS are still developing. We are partnering with product teams to apply the research insights to the definition of smart home AI prototypes and products. Moreover, HISS was planned as the first of a series of projects to create AI experience roadmaps in different usage contexts and we are moving forward with other research[10].

HISS in a Sense and Sensemaking Saturated Environment

Let's now imagine HISS with a high presence of advanced sensing and sense making technologies in our research participants' homes and in our corporate offices. In this alternative reality, high presence in our field locations might mean that householders already:

- live in homes with sensors, devices, connectivity and computing solutions that makes it possible to:
  - detect odors and signature scents in the air (i.e. smell)
  - detect and decode vibrations that travel through the air (i.e. hear)
  - detect and decode meaning through tactile sensations like pressure, hot and cold, wet and dry, and vibration (i.e. touch)
  - create a two or three dimensional likeness of the environment (i.e. see)
  - detect chemicals and substances in liquids or solids (i.e. taste)
- live in homes that have connected infrastructure, appliances and devices that can respond to conditions and commands from humans or machines, and can make changes in the home (actuators and automation)
- are fluent in interacting with an in-home and on-phone interface that lets them direct, respond to, and experiment with new usages of this system. The popularity of smart speakers - recent research predicts almost 50% of US households will have a smart speaker by 2019 (Adobe 2018) – makes this at first far-fetched scenario not so unimaginable.

In our work locations, we imagine we have a means of securely accessing data and event logs from participants' homes during the course of the study, from sources including utilities, internet providers, and of course the sensors, devices and computing solutions within the homes.

So, what might meaningfully change in terms of first order consequences, i.e. how we conduct research? We imagine four significant outcomes to our practice.

Machines Co-Design Research with Ethnographers – We promised we would not venture to imagining sentient robots conducting fieldwork, and as we are re-envisioning a remote ethnography study there will be no need to do so[10]. Instead, inspired by Autodesk's much heralded 2017 Elbo chair created as a collaboration between a human designer and Dreamcatcher software, we imagine ethnographers similarly collaborating with AI software. The future ethnographer may input parameters for research goals, participant criteria, data types and data reliability, research cost and duration constraints, and have an AI system trained with details and successful/failures, impacts and outcomes of earlier projects produce multiple research plans for remote ethnography that can be further tweaked as the ethnographer provides more inputs. Can an AI system write more creative and engaging mobile ethnography questions than we did for HISS? Certainly we expect it could be faster, as we spent a *generous* amount of time designing our questions. Even if the system could provide variations on our existing questions, this could save us weeks of time.

Machines Co-Execute Research with Ethnographers – In a high presence of advanced sensing and sense making technology world, we imagine machines can simplify recruiting and managing research participants in at least three ways. First, gone would be traditional screeners we use now; in would be our machines matching potential participants to our research criteria based on the data from their already instrumented homes that householders have consented to provide for commercial purposes. Our machines could contact these participants through their in-home systems or smart phones and ask if they would like to participate, and gather further information from them so that the human ethnographers can review and choose the most suitable participants, saving us weeks of work. Second, as participants are accepted into the study, our machines could negotiate access to relevant and limited home systems and secure informed consent from participants about how and when the data will be collected, stored, analyzed and deleted. Third, as the research goes to field, the ethnographers could rely on our machine partners to check video, audio and text responses for quality based on criteria we provide. If responses are unclear or inadequate, the system could generate follow up questions or instructions, and only alert the ethnographers when a resolution to inadequate data cannot be negotiated. Overall this could easily save us 100 hours of work over the course of a project the size and scope of HISS.

Ethnographers More Often Experiment and Evaluate Using Data to Create Change In the World – Recall that as social scientists we assume our job is to generate data to identify opportunities to make change in the world. In a high presence world this changes. We likely more often than not end up guiding participants to generate and use data to create change in their own worlds, moving us as ethnographers closer in research practice to our computer science and engineering colleagues. While as ethnographers in technology industry we have certainly engaged in evaluative research from concept testing to proof of concept prototype evaluations to alpha and beta testing of product releases, what we propose here is novel type of evaluative intervention. Rather than just imagining and play acting scenarios in HISS prompted by our research questions, future participants could detail what they would want to be possible and their home systems would attempt to deliver the experience. In part, our research data would then consist of the experiences participants proposed to their systems and the data inputs they ask the systems to consider; the choice their home systems make to meet experience and data criteria; the success, failure and reactions of participants to the new home capabilities; and the negotiating and updating participants would do with their homes systems to achieve more satisfactory results. We would still have participants reflect with us on the value of the usage; combined with the above data, we would indeed be much closer to the ground truth data inputs that our colleagues need to build better AI solutions.

.

Mediate Better Transparency Between Established and Emerging Experts in Cyber-Physical Worlds – Currently we think of computer and data scientists, engineers, and developers as experts on machine learning, computer vision, natural language and audio processing, and Internet of Things sensing and connectivity technologies. In the future, we will also consider people who work, move through, live, shop, recreate and more in environments with high presence of sense and sensemaking technologies to be experts on these systems. These people will be knowledgeable and skilled at interacting with, training, working around, and occasionally subverting such systems. As ethnographers are adept at handling data *about people* and will be adept with data *traces by people* we will play an important role in tech companies deploying such systems. Through our research and advocacy, we will make the priorities, concerns, desires and actions of both types of experts more transparent to one another. Ethnographers will champion ethics, accountability, and data rights to guide strategies and policies their companies use in solution development. Given the ability of these systems to exercise real power in people's lives, having ethnographers who can facilitate understanding and empathy for the positions of all experts will be critical for adoption of solutions that do not create a total surveillance society. Returning to Paglen's (2016) critique of the deployment of computer vision systems as "an active, cunning, exercise of power, one ideally suited to molecular police and market operations–one designed to insert its tendrils into ever-smaller slices of everyday life" we will have an obligation to shape how that power is used. As ethnographers, our deliverables will need to speak to much broader and much more senior audiences at our companies; we will no longer be a nice to have, or anecdotal or fluffy; but a crucial part of mature AI solution development practices.

Given these outcomes, what additional changes to the professional expectations for technology industry ethnographers might follow? We see two additional outcomes beyond those outlines in the low and medium level environments.

***Future Ethnographers are Trained Differently* –**Ethnographers must be fluent in querying how digital traces are created, and in assessing what they can and cannot understand from them. They must also develop new methods for engaging research participants skilled in living in ever more algorithmically generated and mediated settings and fieldsites that can rapidly be reconfigured based on input from lay and professional experts. These skills will become standard part of training at universities as well as professional development courses. Australia National University's 3A institute, with a charter to deliver "a new applied science to enable the safe, ethical and effective design, integration, management and regulation of cyber-physical systems" is one example of where we already see happening in education. (ANU 2018)

***Ethnographers Become More Visible at and More Important for Accountable Technology Companies* –** Future ethnographers, as experts who sit comfortably at the intersection of the cyber and physical worlds, will share peer stature with the data scientists who have come to the fore in the new data-driven innovation narrative popular in technology companies. With complementary skills, they will be jointly responsible for evaluating and justifying the types of sense and sense making capabilities their companies enable, that in practice can quickly devolved into the surveillance and exercises of power Paglen notes. Ethnographer's stars will rise at *accountable technology companie*s; companies that expect to create products explainable to those whose lives are shaped by them and who expect the actions and decision these systems make to them. Ethnographers working for unaccountable technology companies can expect to battle to shift their companies to accountability, or be reduced to apologists for controversial exercises of power using invisible technologies.

CONCLUSION

Through re-imagining three of our existing ethnographic research projects, we have attempted to draw reasonable expectations for how ethnographic practice in the technology industry may change in the next 20-25 years. Our explorations have been unapologetically ethnography-centric. We did not choose to imagine a dystopian future in which the ethnographer's expertise and skills in engaging with, understanding, and thoughtfully giving voice to people outside tech will lose all currency with our employers. Rather, we provided specific and rather pedestrian examples of how we might integrate new technology capabilities into our existing practices without completely blowing up the concept of applied ethnography as a practice and an approach to sensemaking grounded in social science theory and methods. We realize that we have, at times, taken great liberties in our imagining; we have glossed over the complexities, for example, in building machine learning systems that could perform the types of anomaly and pattern recognition we have imagined. These complexities including sourcing suitable data for training the system, and the time and cost in developing such a solution, among many others. In this way, as well, we have been ethnography-centric, choosing to focus on how the impressive and complex work undertaken by our data scientist, computer scientist, engineering, software programming and other highly trained technology industry colleagues could serve *our* professional ends.

We have argued that the combination of technology capabilities that together are common glossed as AI, including machine learning, Internet of Things, computer vision and speech and audio processing technologies, do offer compelling new ways to model and understand the context in which ethnographic encounters take place. They will most likely become indispensable in allowing ethnographers to see and hear data we may not realize we have. They can help us interpret our specific data in the context of larger events or larger patterns. In both ways, they expand our capabilities as researchers and allow a space for us as ethnographers to engage in the new data-centric innovation narrative in ways that acknowledge our expertise in understanding people, and provide an avenue for us to influence how such technologies are developed over time.

In some fashion adding these capabilities to our research practice does get us closer to ground truth about some human behaviors; they do offer better representations of the things we set out to measure. But as we argue that ethnography has never been about testing hypotheses to ultimately generate laws, these new technology tools will never get us *on their own* to the type of truths the ethnographer values above all else; the meanings given to experiences by humans, meanings for which the ground beneath is always unstable. We look forward to the next twenty years, to see how our practices change (and what we got right and wrong) and to the excitement and creative frictions that future rumbles, jostles and liquefaction of our grounds for building evidence will produce.

Alexandra Zafiroglu is a principal engineer at Intel. As a cultural anthropologist, she immerses herself in others' daily routines and realities and formulates explanatory and predictive models for how people and organizations create and will respond to techno-economic transformations in a data-centric world. Contact her at: zafiroglu@gmail.com

Yen-ning Chang is a design researcher at Intel. With a background in Journalism and HCI, she is passionate about communicating people's stories. She brings concepts to life by creating empathy within multi-disciplinary teams to ensure the design work is underpinned by a strong human-centered foundation. Contact her at: yenning.c@gmail.com

NOTES

The research projects explored here were sponsored by Intel Corporation. The arguments presented in this paper do not represent the official position of Intel. We thank the range of research partners for these projects, including: Jennifer Healey, Tim Plowman, David Graumann, Georgios Theocharous, Philip Corriveau, Susan Faulkner, Laura Rumbel, Kathi Kitner, Heather Patterson and Faith McCreary.

1. https://en.wikipedia.org/wiki/List_of_technology_centers

2. Here Woods is paraphrasing Hilary Mason (Mason and Wiggins 2010)

3. Most writing on this topic is a mix of each of these arguments, rather than the playful 'stages of grief' dissection that we've done here. Examples include: Anderson et al 2009; Boyd and Crawford 2012; Burrell 2016; Elish & Boyd 2017; Madsbjerg 2017; Wang 2016.

4. We recognize a number of critiques of this model, including Dourish's (2006) oft-cited critique of the 'implications for design' model, and Amirebrahim's (2016) critique of the flattening of social science through user experience. While these critiques are important, they are not our focus here.

5. Rather than name drop and speak inside baseball, we tend to use key concepts backed up with specific examples. If stakeholders want to know more, we gladly share – but we never overtly start with theory. Elisabeth Shove's definition of 'convenience' in *Comfort, Cleanliness + Convenience: the social organization of normality* (2003) was critical in explaining the history and future of the smart home to stakeholders in our first case study in this paper. Jenna Burrell's 2016 paper on machine learning has been a useful piece for helping us explain the differences between human and machine cognition, and how we ask question 'about AI' to non-experts in our last case study in this paper.

6. Paraphrased from memory; presented at an invited talk at NWWiC Regional Conference 2013 (Zafiroglu & Healey 2013).

7. The work of Boyd & Crawford (2012), Kitchin (2014), Wang (2014) and Boyd & Crawford (2017) as well as the collective body of work published in *Big Data & Society* amply makes these points.

8. A perspective on the role of ethnographers' responsibilities for designing for privacy in smart homes based on SHIFT was presented at EPIC 2016 (Zafiroglu, Patterson, McCreary 2016).

9. The entire protocol is longer and more detailed than is relevant in this case study. We are currently working on a conference paper that details HISS and findings.

10. The next planned project will be for AI in manufacturing, and will, naturally, be named MISS for 'Manufacturing Instrumentation and Sensing Study'.

11. For those of you craving a good robot-ethnography story, Alicia Dudek's delightful "Lou and Cee Cee prepare for fieldwork in the future: a world where robots conduct ethnography" (2016) will not disappoint!

REFERENCES CITED

Adobe Digital Insights
2018 State of Voice Assistants First published September 10 2018 https://www.cmo.com/features/articles/2018/9/7/adobe-2018-consumer-voice-survey.html#gs.CcIsBMw

Amirebrahim, Shaheen
2016 The Rise of the User and the Fall of People. *Ethnographic Praxis in Industry Conference Proceedings* 2016:71–103.

Anderson, Ken, Dawn Nafus, Tye Rattenbury and Ryan Aipperspach
2009 Numbers Have Qualities Too: Experiences with Ethno-Mining. New York: *Ethnographic Praxis in Industry Conference Proceedings* 2009:123–140.

ANU
2018 3AI Institute Home Page accessed September 20 2018 https://3ainstitute.cecs.anu.edu.au/#about

Bogost, Ian
2017 Why Zuckerberg and Musk are Fighting About the Robot Future *The Atlantic*, July 27 Accessed September 1, 2018. https://www.theatlantic.com/technology/archive/2017/07/musk-vs-zuck/535077/

Bowley, Rachel
2017 The Fastest-Growing Jobs in the U.S. Based on LinkedIn Data first published December 17, 2017 https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data

Boyd, Danah and Kate Crawford
2012 Critical Questions for Big Data *Information, Communication, & Society* 15(5):662-679

Burrell, Jenna
2016 How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* first published January 6, 2016 https://doi.org/10.1177/2053951715622512

Davenport, Thomas H. and D.J. Patil
2012 Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90(10):70-76, 128

Dourish, Paul
2006 Implications for Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* New York: ACM 541-550.

Dudek, Alicia
2016 Lou and Cee Cee prepare for fieldwork in the future: a world where robots conduct ethnography *Ethnography Matters* first published June 17, 2016 http://ethnographymatters.net/blog/2016/06/17/lou-and-cee-cee-prepare-for-fieldwork-in-the-future-a-world-where-robots-conduct-ethnography/

Elish, M.C. and danah Boyd
2017 Situating Methods in the Magic of Big Data and AI. *Communication Monographs* 85(1):57-80

Feltman, Rachel
2014 MIT Researchers Can Listen to Your Conversation by Watching Your Potato Chip Bag. *Washington Post* August 4, 2014. https://www.washingtonpost.com/news/speaking-of-science/wp/2014/08/04/mit-researchers-can-listen-to-your-conversation-by-watching-your-potato-chip-bag/?utm_term=.ba5473c31113

Geertz, Clifford
1973 The Interpretation of Culture: Selected Essays New York: Basic Books

Glassdoor
2018 Fifty Best Jobs in America Accessed September 1, 2018. https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang and Owain Evans

2018 When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research* 62: p. 729-754 https://arxiv.org/abs/1705.08807

Intel Corporation
2018a Media Alert: Data-Centric Innovation Summit – Data Center Platform and Products Fueling Intel's Growth https://newsroom.intel.com/news-releases/data-centric-innovation-summit-data-center-platform-products-fueling-intels-growth/ (August 7 2018)

Intel Corporation
2018b Smart Homes: Perceptive, Responsive, and Autonomous accessed September 1 2018 https://www.intel.com/content/www/us/en/smart-home/overview.html

Kitchin, Rob
2014 Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* first published April 1, 2014 https://doi.org/10.1177/2053951714528481

Madsbjerg, Christian
2017 *Sensemaking: The Power of the Humanities in the Age of the Algorithm* New York: Hachette.

Mason, Hilary and Chris Wiggins
2010 A Taxonomy of Data Science *Dataists.com* First published September 25, 2010 http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

McKinsey
2009 Hal Varian on How the Web Challenges Managers *McKinsey Quarterly* January 2009 https://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers

Metcalf, Jacob and Kate Crawford
2016 Where are the Human Subjects in Big Data Research? The Emerging Ethics Divide *Big Data & Society* first published June 1, 2016 https://doi.org/10.1177/2053951716650211

O'Neil, Cathy
2016 *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* New York: Crown Books

Paglen, Trevor
2016 Invisible Images (your pictures are looking at you) *The New Inquiry* December 6, 2016 https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/

Reichenbach, Lisa and Magda Wesolkowska
2008 All that is Seen and Unseen: The Physical Environment as Informant *Ethnographic Praxis in Industry Conference Proceedings* 2008:160-174.

Schwab, Klaus
2017 *The Fourth Industrial Revolution* New York: Crown Publishing

Shove, Elizabeth
2003 *Comfort, Cleanliness and Convenience: The Social Organization of Normality* Oxford: Berg

Wang, Tricia
2016 Co-Designing with Machines: Moving Beyond the Human/Machine Binary first published June 13, 2016 http://ethnographymatters.net/blog/2016/06/13/co-designing-with-machines-moving-beyond-the-humanmachine-binary/

Woods, Dan
2011 LinkedIn's Daniel Tunkelag on "What is a Data Scientist" *Forbes* October 11, 2011 https://www.forbes.com/sites/danwoods/2011/10/24/linkedins-daniel-tunkelang-on-what-is-a-data-scientist/#2ab65e3311cc

Zafiroglu, Alexandra, Jennifer Healey and Tim Plowman
2012 Navigation to Multiple Local Transportation Futures: Cross-Interrogating Remembered and Recorded Drives AutomotiveUI '12: Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications p. 139-146

Zafiroglu, Alexandra and Jennifer Healey
2013 "From Ground Truth to Common Ground: Negotiating Between Problem and Solution Spaces" invited talk at NWWiC Regional conference Portland, OR, 2013

Zafiroglu, Alexandra, Heather Pattterson and Faith McCreary
2016 Living Comfortably In Glass houses 2016 Ethnographic Praxis in Industry Conference Proceedings, p. 540, ISSN 1559-8918, https://www.epicpeople.org/living-comfortably-glass-houses/.

Zhang, Chao and Juchi Yan, Changsheng Li, Xiaoguang Rui, Liang Liu, and Rongfang Bie
2016 On Estimating Air Pollution from Photos Using Convolutional Neural Networks Proceedings of the ACM on Multimedia Conference Pages 297-301 doi>10.1145/2964284.2967230

Confessions across Digital Distances

JESS SHUTT
Salesforce

Digital environments can expand the distances between people. While this is often challenging, it can also be leveraged to do great things. This Pecha Kucha explores how we can take this negative divide between people and flip it on its head to discover more powerful insights. By looking at a range of studies focused on sensitive subjects, we explore how technology has the power to create a safer environment for vulnerable participants. While technology is an often underutilized research tool, these technology enabled environments can lead to richer data and insights. As a result, we, as researchers, can create the space needed to share some of our most intimate stories.

Unspoken Thoughts, by Jess Shutt

Jess Shutt has dedicated her career to studying & creating technology to help democratize complex processes & systems from finance to consumer robotics. Her current work as the Lead User Researcher of Einstein at Salesforce focuses on applying these concepts to artificial intelligence.

“Empathizing” with Machines

CHRIS BUTLER
Philosophie

When we study human systems and organizations we have a job that requires to empathize or at the very least be compassionate towards the experiences others are having. This allows to understand their goals, problems, and how we can best make their lives better. When machines start to do things that we can’t imagine how do we continue to work with them? What is necessary to create great combinations of humans and machines? What is a machine’s purpose? Very simply: it is to serve human purposes. As technology continues to build facades that hide the human element we need to pull back the curtain (like the one in the Wizard of Oz) and see that the tools we build are really us reflected back. We have the choice to make tools that are good or bad for us.

Empathy Mapping for the Machine for Patrick, Field Service Operations project spring, 2016, Chris Butler

Chris Butler, is the Director of AI at Philosophie and frequently speaks on the intersection of product, design, and AI. He has extensive experience from Microsoft, Waze, KAYAK, among others. Through his practice he has created Empathy Mapping for the Machine and Confusion Mapping to align teams when building intelligent systems. *chrisbutler@philosophie.is*

Play it Back: Research as Intervention

NATALIE NAPIER
InWithForward

The social welfare system was built to protect the vulnerable through the provision of basic needs. I left my social service job to join an organization with a mission to shift that system from safety nets to trampolines - from services designed to maximize safety, to those that develop agency and resilience. That's meant interrogating and renewing my principles for ethical engagement with people who are getting the poorest outcomes from services. Returning people's data to them, in the form of a story is now a practice at the heart of my relationships to the people with whom I do research. At the best of times this interaction is an intervention in and of itself, validating someone's experience and allowing them to open themselves up to new self-narratives. But the goal of story return in not a positive reception; rather, it's about following through on our ethical commitment to recognize people's ownership over their own data, and allowing them the opportunity to benefit from it. Ultimately, the story is never evidence for once policy direction or another, but, through the way in which it's gathered, it becomes evidence for how human connection is the foundation for our growth.

Photo by Natalie Napier

Natalie Napier is Lead Coach at InWithForward and divides her time between rooming houses, seniors' residences, shelters, newcomer communities, and working with the organizations that serve them, to build capacity for community-based social R&D. She is currently living in Montreal, Canada. natalie@inwithforward.com.

Life and Death of Evidence: The Role of Digital Interactions during Mexico's Earthquake

FRANCISCO JAVIER PULIDO RAMIREZ
INSITUM

Social media played a fundamental role on Mexico's earthquake, it bring us new solutions but created some other problematics that were unexpected. Millions of users shared their experiences faster than any other traditional media but the use and abuse of their evidence impacted the way we faced the crisis. Earthquakes are extreme case scenarios where social medias couldn't forecast the different consequences of their design decisions that impacts people's lifes. As producers of contents, all our evidence is storage on the digital sphere, always available, unchangeable, static, waiting to be rescue for interpretation. Most of the evidence that generate chaos after the earthquake happened because they were digitally alive, being shared over and over without control, for hours and days and when it finally reach you it was no longer useful. But on a scenario where temporality is crucial and minutes can define life or death, should we kill our evidences in pro for a better communication?

Photo by Dean Chahim

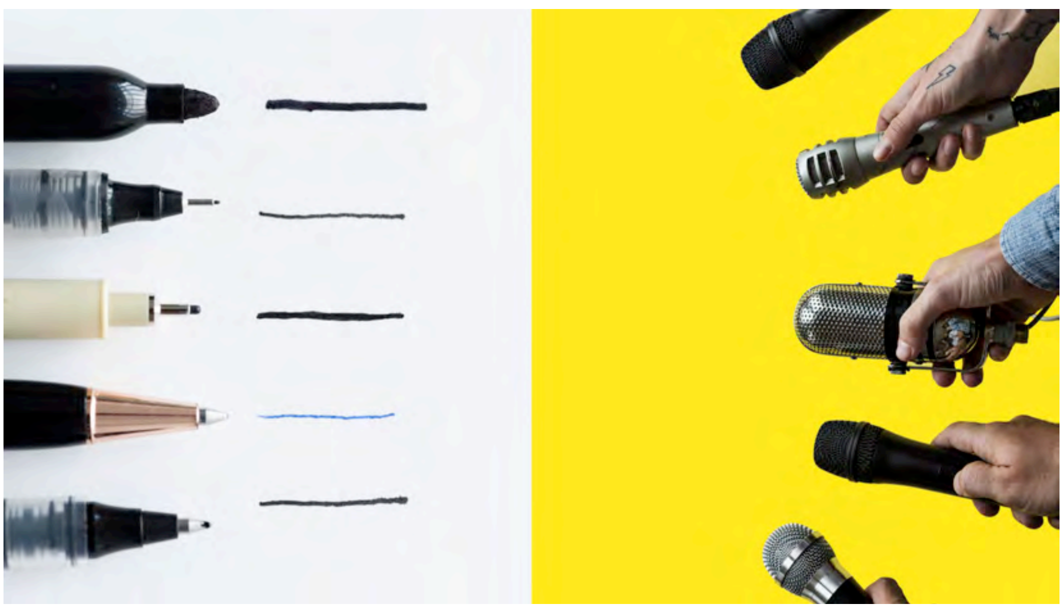
Francisco J. Pulido Ramírez, is an innovator consultant focused on LATAM market. He is passionate on how societies are modifying their behavior by using digital tools, his experience in strategic research and digital strategy feeds his desire to understand how we behave the way we do. franciscopulido@insitum.com

The Story as Evidence: It's Yours, It's Mine, It's Theirs

NIK JARVIE-WALDROM
Empathy

I've been reflecting on my role in the use and abuse of evidence — in the past as a radio producer and more recently as a writer in a design research company. Storytelling is held aloft as something businesses need to do more of — and be better at — but often the narratives do not belong to businesses. We are re-tellers. The work of a writer presenting design research isolates evidence from its source. There are limits to what we can do to make sure evidence is considered alongside the intention it was gathered with. I started working on this because I wanted to share my indignation at evidence I gathered being misrepresented. My editors have turned stories of triumph into stories of disaster to get more clicks. But I've noticed the similarity between my questioning of editors, and the anthropologists I work with questioning me.

Evidence exists in relation to questions. Defining the things we're curious about helps us focus, and decide which evidence to seek out. Ethnographers need to be so careful about the way we communicate throughout a project, lest the evidence we gather is carried away to serve a question we never wanted to ask.

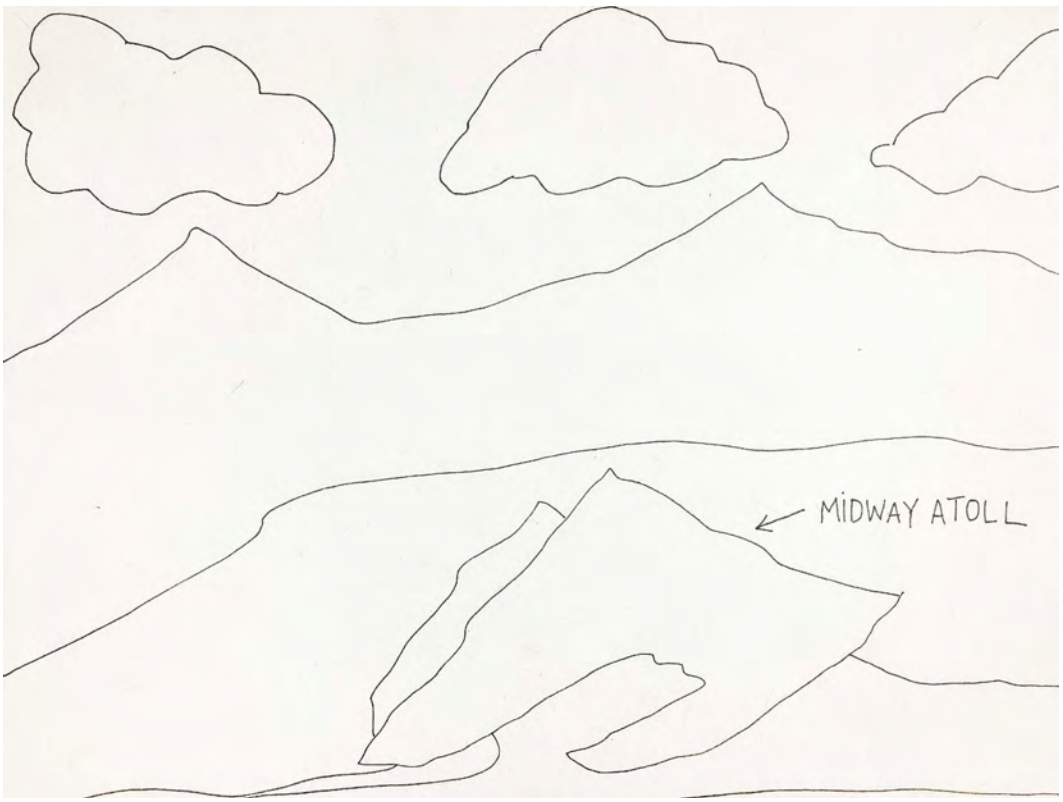


Nik Jarvie-Waldrom is a writer and content designer who finds efficient, powerful ways to help people understand each other. Her experience as a producer of human interest radio documentaries feeds into her creative approach to writing, and equips her to identify, structure and clarify relatable narratives. *nik@empathydesign.com*

Midway Atoll

SARAH BROOKS
IBM

We live our lives in contexts of overlapping systems. Developing the skill to connect dots of evidence between social, ecological and economic evidence offers the potential for more effective interventions in complex challenges.



Sarah Brooks, Sarah Brooks' teaching and design practice sits at the intersection of design research, service design, and social innovation. She currently serves as a Design Executive and Distinguished Designer at IBM.

Invisible Evidence: Our Disconnection with Broadband Connectivity

SUSAN FAULKNER
Intel Corporation

When our broadband connectivity at home stops working, it's a crisis. When it does work it is magic, an invisible miracle most of us don't understand. Our evidence for whether it is working great or barely working at all, is scant, murky and elusive. The telecommunications industry language used to measure and characterize connectivity is obtuse. Data transfer rates, data storage rates, and wireless frequency rates, all sound similar and make no sense to most people. So, what's wrong with that? Who cares how many megabits per second download speed I'm getting? As long as I can stream The Crown, what difference does it make? That's what I thought when I began doing research about people's relationships with connectivity, and then I met people who changed my point of view. Connectivity is fundamental to how we experience the world and to our sense of well-being. We need ways to connect with our connectivity.

Susan Faulkner is a Senior Researcher at Intel Corporation where her work has focused on the future of connectivity, media creation and consumption, the role technology plays in people's daily social transitions, and the role of women in the community of hackers and makers that are transforming the way physical objects are created. susan.a.faulkner@intel.com

PechaKucha 3 – Systemic Evidence: Seen and Not Seen

Rejected!: Design Research, Publics and the Purging of New Technologies

LAURA CESAFSKY
Nissan Research Center, Silicon Valley

A challenge for design research today lies in naming, knowing and accounting for people who are not direct users of our technologies, but who are nonetheless affected and compelled to interact with them in daily life. This Pecha Kucha takes us to the streets of Bogotá, Colombia, where a new bus system that was roundly rejected becomes a cautionary tale on the perils of ignoring the painpoints of 'non-direct users.' Drawing from pragmatist political science, I propose we can usefully understand this latter group as a 'technological public,' and I touch on key difficulties of designing for publics.

Laura Cesafsky is an urban geographer, transportation nerd, and Human-Centered Systems Design Researcher at Nissan Research Center in Silicon Valley. laura.cesafsky@nissan-usa.com

Working with Intuition

ANISH NANGIA
eBay Inc.

Evidence, today, can have very narrow definitions. For digital products, this type of evidence usually includes clicks and engagement metrics. I believe that in our effort to only listen to numbers and data, we have created a culture that looks down on intuition as something messy and to be rejected. As a result, we train ourselves to cast our intuition aside. When we do listen, we dare not speak up about it.

Intuition is a part of what makes us human. It is often the primary way we make sense of our work (and the world around us). Yet, we continue to focus only on rationality and logic. Ultimtately, the goal of this piece is to present intuition from a different lens – to understand it, learn to listen to it and create spaces where we talk about it.

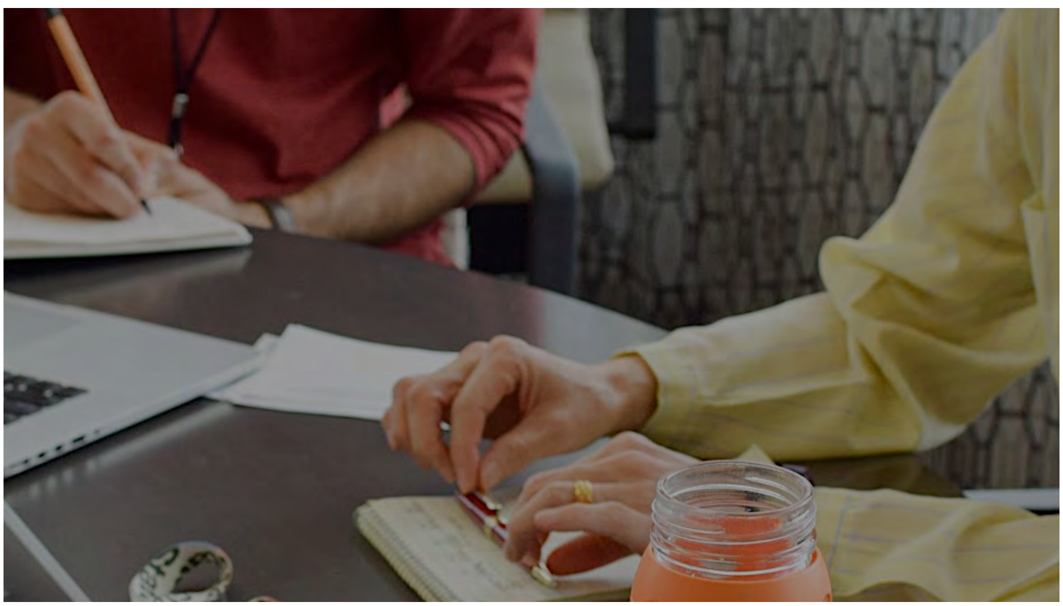


Photo by Amoli Mehta

Anish Nangia is a User Experience Researcher that uses human centered design to understand complex problem spaces and inspire compelling products. anishnangia@gmail.com

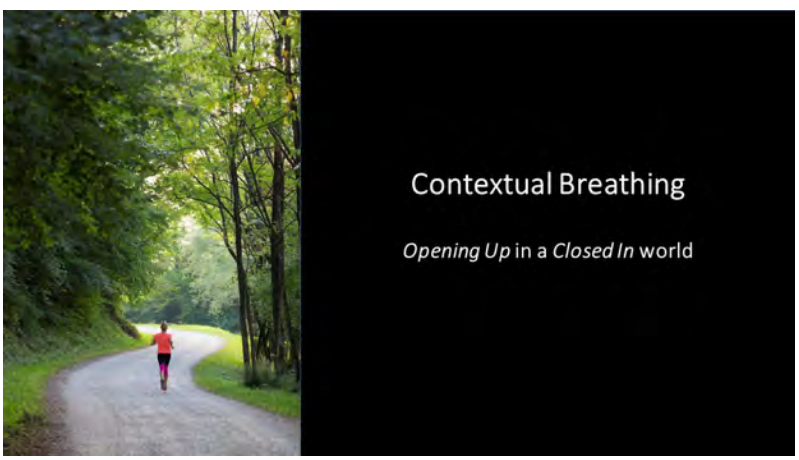
Contextual Breathing

APRIL JEFFRIES
Ipsos Understanding Unlimited

Title: Contextual Breathing – The importance of opening in an increasingly closed world.
Context cannot be ignored. The ability to pull back, observe and listen deeply balanced with internal analysis and reflection has significant impact on our individual and societal health. Myopic views that ignore or distort what is happening around us have resulted in a social, cultural and political bipolar effect that occurs within a narrow spectrum of isolation. Extreme swings from close-minded tribes to secluded self dialogue, wreak havoc on our broader needs for transcendence and compassion.

A study of middle-class moms in America, found a pull toward insular communities in unexpected places. Hostile or challenging political arguments were increasingly infiltrating conversations in venues ranging from Facebook to book club. Emotional eruptions in previously "safe spaces" caused retreats to like-minded groups. Women who may have otherwise enjoyed open curiosity or stimulating debate, in these situations, were ill-equipped to handle feelings of rejection and separation. Tribal political behavior was a result of stress and a deep sense of loss.

The rhythmic balance of convergence and divergence is as necessary as breathing, to fairly assess external realities and form rational individual opinions. Ethnography is a prime opportunity to build this context. Revealing non-judgmental stories of those unlike ourselves, we draw back from the myopic focus, which initiates a return to the center with deeper understanding and empathy.



April Jeffries, Global President of Observation at Ipsos, has held progressive leadership positions at Kraft Foods, Campbell's Soup, and Pinnacle Foods. She has an Engineering degree from MIT, an MBA from Wharton B-School and a Certificate in Multimedia Technology from NYU. She is an award-winning writer, singer and producer of original works in various mediums.
April.jeffries@ipsos.com

Diagnosing the World Pulse

CHRISTOPHER A. GOLIAS, PH.D.

This Pecha Kucha explores the ethics of interpreting data by employing an extended metaphor of data as the lifeblood of the connected world. It begins by exploring two distinct viewpoints on medical pulse diagnosis, starting from the perspective of the acupuncturist diagnosing a patient's pulse and continuing through differences between Eastern pulse diagnosis and biomedical pulse diagnosis. I explore data as lifeblood, and imagine more visceral ways to read data (e.g., auguring data) and the ethical implications of such a reading. I envision data as a flowing river filling a lake, in which diagnostic specialists observe society's reflection. In the process, I contrast utopian visions of a data driven world with dystopian ones before resolving tension by returning to the central comparison of data scientist and medical doctor. The presentation concludes by recalling medicine's Hippocratic Oath, an ethical charter binding practitioners to a code of conduct, and implying that data science consider a similar injunction.

The Agnew Clinic by Thomas Eakins. Public Domain.

Christopher Golias is a technology ethnographer who has conducted applied anthropological research across various areas including retail, healthcare, indigenous rights, substance use, mobile technology, retail, governance, and information technology. He holds a Ph.D. in Anthropology from the University of Pennsylvania.

Helping People Heal

REBEKAH PARK
ReD Associates

EMILY PRESTON
Idea Couture

A documentary-style short film exploring how people in the United States heal themselves once they leave the doctor's office. This film is based on in-depth field research and illustrates how healing depends on the alignment between what doctors expect from patients, and what people actually do in their everyday lives. Exploring and entering the lives of three individuals, Helping People Heal reveals the ways that people put their recommended treatments into practice—and why these personalized applications achieve better health outcomes beyond simply doing what the doctor ordered.

Emily Preston is a Toronto-based cinematographer and director whose cinematography work has been showcased at many recognized festivals, including the Atlanta Film Festival and the Atlantic Film Festival. Currently employed by Idea Couture, Emily has produced over fifty films for Fortune 500 companies. She is currently producing a documentary series, Women in Digital, for Cognizant Technical Solutions that focuses on women in leadership positions and the challenges they face.

Rebekah Park is an applied medical anthropologist and a manager at ReD Associates, where she specializes in health and healthcare. Rebekah holds a PhD in socio-cultural anthropology UCLA and is currently on the board of the Association for Political and Legal Anthropology. Her book The Reappeared: Argentine Former Political Prisoners was published by Rutgers University Press in 2014.

Seeing Double

KARL MENDONCA
Google

Dekho, Purani Dilli (Seeing Double) *is an essayistic film that illustrates the experiential and contextual nature of ethnography, while reflecting on cultural predispositions that often frame ways of seeing and knowing. The film was produced as part of a design conference in Delhi, where conversations centered on 'modernizing' traditional modes of production, distribution and the object itself. My goal was to question a peculiar narrative of modernity and development that positioned design and digital technology as an answer to longstanding social, cultural and political relations. Working in the tradition of structuralist film making, the film is composed as a series of encounters in Old Delhi, where object and people play the role of interlocutors that prompt a turn to history and a deeper engagement with place. As an evidentiary form, the film is an inversion of the typical 'deliverable' serving as a critique of accelerated design culture and a meditation on the critical lens ethnography can provide.*

Karl Mendonca is a Lead Design Researcher at Google and an artist, educator and filmmaker. His mixed media work has shown at a number of galleries and film festivals including the Lower Manhattan Cultural Council, The Queens Museum of Art, the Oxford Film Festival, Stuttgart Filmwinter, Jersey City Museum and Experimenta (India). Karl was an adjunct faculty at The New School where he taught hybrid courses on media theory and production. He is also finishing a PhD in Film & Digital Media at the University of California, Santa Cruz.

Tracey's Struggle

SHELLEY O'NEIL
Jump the Fence

CHARLIE COCHRANE
Jump the Fence

Despite the 'foodie' reputation perpetuated by reality TV, cooking shows, and food blogs, many Australians struggle with eating healthy food. MasterFoods brand has been a household name in Australian supermarket groceries for 70 years and have a 2020 brand ambition to help all Australians enjoy the rewards that good food can bring. So they asked Jump the Fence to conduct an ethnographic study to understand people with challenges around physical access, time, money, and skill with food. Tracey from Southwest Sydney is one of five people the study focused on. Tracey's welcoming generosity and honesty are profound and reveal the complex layers of psychological, cultural and situational barriers she has to eating well.

Tracey in her kitchen in Southwest Sydney

Shelley O'Neil and Charlie Cochrane are partners in *Jump The Fence*, a qualitative research agency. We help our clients make brands, services, communications and strategies that connect with people's experience, emotion and intention by taking a collaborative approach with our research participants. We use a diverse range of qualitative methods, but our combined backgrounds and specialties make video ethnography our 'super power'. Our ethnographic films paint pictures that reflect people's complexity and tell stories that everyone can relate to. They enable empathetic and contextual understanding that data alone cannot.