Anthropology by Data Science:

The EPIC Project with Indicia Consulting as an Exploratory Case Study

Stephen Paff
Fall 2018

This report is submitted in partial satisfaction of the requirements for the Master of Arts in Anthropology.

Table of Contents

Executive Summary

Overview:

- This report summarizes my work with Indicia Consulting on the technical Task 6 for the project, "Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency Program Design."
- Task 6's goal was to create ethnographic and machine learning decision tree models to determine an individual's cyber status.

Contribution within Anthropology:

- My practicum is an exploratory case study into anthropology *by* data science, integrating data science techniques into ethnographic and other anthropological work.
- The shifts within machine learning algorithm development gives impetus for incorporating quantitative techniques that are abductive, local, and interpretive into ethnographic practice
- Nick Seaver's concept of 'bastard methodology' is a helpful framework for how to integrate data science techniques into ethnography.

Project Results:

- My machine learning decision tree model had three parts: pre-development, development, and pruning.
- I then utilized insights from this process to develop random forests.
- My random forests had an 100% accuracy on the testing data.

Section 1: Introduction

This report summarizes my work with Indicia Consulting on Task 6 of the project, "Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency Program Design" (hereafter, EPIC Project). Indicia Consulting "is a mission-driven social enterprise" created by and made up of anthropologists that seeks to "increase in sustainability and subsequent improvement in the natural environment through engaging behavior through proven social science insights and methods" (EPIC 2018).

Indicia began work on this project in 2015; I initially joined during Task 4 in 2017 to provide statistical and data science analysis of their ethnographic findings, a role I have continued into future project tasks. We[1] conducted Task 6 from March to August 2018. The core team consisted of the Principal Investigator, Dr. Susan Mazur-Stommen, the Project Manager, Amy Bayersdorfer, and two Senior Research Analysts, Haley Gilbert and me. In addition, over the course of the project, the team had included three additional applied anthropologists, two professors of anthropology from the California State University (CSU) system, and 19 students from across the CSU system, as well as several junior analysts and interns. We also received guidance and assistance from members of a Technical Advisory Committee (TAC), in particular Dr Dan Fredman (University of Vermont/VEIC). Finally, we worked with Achla Marathe and Samarth Swarup of the Network Dynamics Simulation Science Laboratory at Virginia Tech

[1] Throughout this report, every time I use "we" I refer to the Indicia team working on this Epic project.

University who shared a synthetic population they constructed from American Census Survey and American Time Use Survey data (see Section 4.2).

In Task 6, we combined ethnographic decision tree modeling (EDTM) with machine learning decision tree modeling (CART) to understand the relationship between consumer technology usage and energy consumption. EDTM is an ethnographic technique anthropologists have developed to model collective decision making (C. Gladwin 1997, 2001; H. Gladwin 1984; Mukhopadyay 1987), and the CART is a machine learning computational methodology within data science (Mitchell 1997; Neville 1999; Rokach 2015). I will use Nick Seaver's concept of a "bastard methodology" to ground my discussion. For Seaver, a bastard methodology, which both ethnography and data science techniques are, refers to a mishmash approach combining several disciplinary strategies/techniques based on what works to address a specific research question(s) (2015: 43-45). Bastard methodology provides the theoretical groundwork for anthropology *by* data science, that is integrating data science techniques into ethnographic and other forms of anthropological research, and our work in Task 6 is an exploratory case study into how to potentially accomplish it.

The California Energy Commission called for research to understand how Californians understand and engage with energy usage, with the goal of developing informed, sustainable energy policies. In June 2015 the California Energy Commission accepted Indicia's proposal to analyze how relationships with technology influence electricity usage patterns using ethnography and other qualitative methods (e.g., surveys, coding, ethnographic software). The goal of the EPIC project initiative overall is to lower societal energy consumption by creating effective energy-efficiency programs.

Indicia designed the project to take advantage of the combined power of mixed-methods research. The project consists eight tasks (a phase of the overall research project), six of which we have completed at the time of this report. Tasks 2 and 3 detail how Indicia fielded a survey and received 415 responses, then used the survey participants as a pool for recruiting 48 households for in-depth interviews. Next, they used Atlas.ti to code interviews and conduct thematic analysis. In Task 4, we revisited the survey and interview data using statistical analysis, which I performed for them. In Task 5, we further tested our findings from the observations, interviews, and surveys against a larger data set of generic electricity consumption, drawn from zip codes roughly matching those of our participating households in Marin County, and the City of Long Beach, California (see Figure 2 in Section 3.1 for a summary of these tasks).

Task 6's specific goal was to create ethnographic and machine learning decision tree models to determine an individual's *cyber status*, a psychosocial measure Indicia developed in this project to understand an individual's relationship with and affective response to technology (Indicia Consulting 2018a:5).[2]

My primary motivation personally in both this project and this report is to explore how to broaden the scope of what constitutes anthropological work by cross-fertilizing with data science and data science techniques. Many anthropologists have advocated for cross-disciplinary approaches between anthropology and data science by conducting separate but complementary roles (Seaver 2015:36, T. Wang 2013, Madsen 2018, Norvaisas 2014, and Slobin 2010), but I advocate for an integration of anthropology and data science methods beyond simply applying them to the same project as separate specialists. The cross-fertilization I advocate for removes

[2] See Section 3.1 will provide a detailed chronology of all Epic Project Tasks and how the research project developed into the most recent manifestation of Task 6.

the boundaries between the two, allowing the disciplines to influence each other. Within anthropology, this manifests as anthropology by data science work: using data science techniques in ethnographies when applicable.

I consider research to be *ethnographic* when it seeks to holistically understand and express the lived experiences of actors in a sociocultural context(s) (Curran 2013:62; Edirisingha 2016). Because ethnographers seek to reformulate their conceptualizations throughout both fieldwork and analysis, ethnography is an abductive approach, which refers to an approach that iteratively refines or reformulates its conceptualizations as the actor(s) encounters data from new phenomena (Timmermans 2012:167). Abductive reasoning generally involves starting with some preconceptions yet iteratively altering them during the research, distinct from both inductive reasoning, which seeks to start research as a "blank slate" and build all conceptualizations from the ground up, and deductive reasoning, which starts research with already-decided conceptualizations used to interpret observations and data.

Abductive and interpretative techniques have developed significantly within data science, potential quantitative counterparts of anthropological qualitative research techniques (Curran 2013:63), and anthropologists, applied anthropologists especially, can and should incorporate these techniques into relevant ethnographic research. A bastard view of ethnography reinforces this strategic incorporation of other disciplines' techniques when they benefit the research project. My practicum project is an initial exploration into how to incorporate the two fields in the specific contexts of research.

In Section 2, I will discuss how abductive machine learning techniques have shifted the quantitative and qualitative debate creating a unique opportunity to integrate data science quantitative tools into ethnographic research, and how Nick Seaver's "bastard" methodological

and disciplinary approach (2015:44-45) facilitates this integrative work. In Section 3, I will summarize our project and how we attempted to combine the two as an exploratory example of integrative work, and in Section 4, I will analyze the strengths and weaknesses of our approach, what I would maintain and change, and how I intend for future research in our project to expand the strengths and address the weaknesses. Through this reflection, I intend to provide the methodological considerations necessary to replicate and refine this type of research in future.

Section 2: Anthropology and Data Science

2.1 Machine Learning Algorithms in the Quantitative/Qualitative Debate

Anthropologists have taken part in a long discussion and debate between quantitative and qualitative methodologies in social research within the Western modernity (Nafus 2018:4-5, 11). At its core, this debate has most often centered on the difference between universal, objective, top-down approaches to social research and flexible, local or particularistic, bottom-up approaches (12). Quantitative analysis has come to represent the former, and qualitative techniques the latter: Within data science, machine learning techniques are a partial shift towards situational and iterative quantitative analysis (Dawn 2018:15), which can and should modify how anthropologists relate to quantitative work: providing quantitative data techniques that are potentially local, abductive, and bottom-up to utilize. This strategic incorporation of data science techniques into anthropological work I term anthropology *by* data science.

Data sciences refers to an interdisciplinary discipline – or a bastard discipline (Seaver 2015:43) – that analyzes and extracts data through the development of computational algorithms. At its most practical level, *data science* refers to coding algorithms to analyze data, and to do so, it combines strands and aspects from computer science, mathematics (particularly statistics and linear algebra), engineering, the social sciences, and much more. It is a loose, decentralized discipline amalgam of various "people, epistemologies, and methods" (Seaver 2015:43), making a precise description of their methodology difficult.

Data science has a complicated relationship with statistics – which's most basic definition is the mathematical study of collecting, organizing, and interpreting data. Data science could technically be a field within statistics, given that it seeks to analyze patterns in data. But, data science is often both distinct from traditional statistics for three reasons:

1. Computer scientists and engineers, not primarily mathematicians, have prominently formed the disciple.
2. Data science emphasizes computation and algorithmic procedures.
3. Data scientists in the field cultivate unique skillsets based on its intersection between (traditional) statistics, computer science, and non-academic data problems (Nafus 2018:16; Lowrie 2018:43-44).

Here is an in-depth explanation of the difference between data science and statistics from a previous research paper of mine:

> A friend of mine and fellow data scientists provided the most basic yet most compelling definition of data science I have hear: "Data science is applied, computational statistics." Situating itself in terms of the analyzing data means that data science technically falls

> within or at least close to statistics,[3]… yet data science puts the emphasis on computational statistics: developing/programming algorithms to analyze data. At the same time, unlike academic or "pure" computational statistics (or academic statistics in general), data science has evolved in the context of analyzing specific data and solving "real world" problems…. As the sociologist/anthropologist Shreeharsh Kelkar (2014) demonstrates… data science both reflects and catalyzes a shift within the hard sciences towards computational analytical work, requiring "a certain kind of scientist who is skilled at both statistics and software-building." (Paff 2018:6-7)

Machine learning algorithms form the primary set of techniques or methods in the discipline to analyze data. Machine learning describes a classification of computer algorithms, in which the computer learns from a set of data. Although there are several different definitions of *learning* for a computer, I will use Herbert Simon's: "learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time" (Simon in Kononenko 2007:37). In summary, machine learning algorithms iterate through data and during each iteration, modify, and hopefully enhance how they model that data.

Machine learning techniques articulate a shift towards situational and iterative quantitative analysis (Nafus 2018:15). This might potentially revamp "the established opposition between quantitative and qualitative methods" (19); in particular, machine learning techniques and ethnography may have more in common. For John Curran, these machine learning

[3] My own experience has taught me that when anthropologists and many social researchers think of statistics, they conceive of very specific sets of practices (like certain hypothesis testing methods), which data science can involve but does not have to.

techniques are ethnographic (2013: 63): their aim "is not the epistemological search for fact but rather the epistemological search for meaning (more interpretative) [which] moves into an ethnographic space" (2013:70). For him, machine learning and ethnographic techniques (at least in the business world) have five characteristics in common (70):

1. Both are interested in the everyday culture.
2. Both explore patterns, movement, and networks.
3. Both are interested in the physical – how the body interacts with products and space.
4. Both can attempt to understand in relation to consumption and life.
5. Both can offer holistic and synchronic approaches to analysis.

Even though not all five of these points hold for every research project involving machine learning techniques or ethnography, he is articulating a key shift within machine learning towards abductive approaches.

Dawn Nafus illustrates this wider trend of machine learning based on a discussion of Bayesian statistics, popular within machine learning:

> Bayesian approaches to assembling data always start with a notion of partiality and contingency that we more readily take to be ethnography's bread and butter. It is a view of probability as 'orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information (Kotz and Johnson 2012). Bayesians then do not attempt to 'devise protocols that will purify the data in advance of analysis', as Strathern put it (2004:5-6), but fold in 'impure' propositions on an ongoing basis…. These incremental buildups of beliefs do not require reference to

> infinite sampling as the standard of evidence. Nor do they point towards a goal of universal truth. Bayesians do not tame chance, they proliferate it. (2018:16).

Both scholars are noticing a defining trend within machine learning – even though it is not the only trend – towards abductive techniques, which parallel qualitative ethnographic approaches.

"Quantitative and qualitative knowledge are not inherently separate, but the distinction between the two has been a longstanding Western cultural cleavage that has had the effect of separating them out" (Nafus 2018:11). In response to this, she concludes that we anthropologists need to reflect this relationship better:

> The longstanding distinction between qualitative and quantitative knowledge production lies in the background to this rejoinder of digital data and ethnography. If methods have a social life, this distinction becomes not an obstacle to be overcome but a particular social arrangement that needs to be better understood. (2018:11)

Nick Seaver echoes a similar solution: "Instead of taking the [ethnographic] encounter with big data as an occasion to rehearse this commonsense relationship between methods, we might take it as an opportunity to examine how methods relate" (2015:43). I agree that reflecting anew given this shift is necessary and helpful to understand the potentially shifting relationships and that ethnography provides an invaluable toolset to do this. But, I also would add that we need to enact this difference into our own practice as anthropologists. The shifts within machine learning algorithm development gives impetus for incorporating quantitative techniques that are abductive, local, and interpretive. The debate between universal and proscriptive vs. local and bottom-up knowledge production may not – or at least no longer – need to divide quantitative and qualitative techniques. Machine learning algorithms "leave open the possibility of situated

knowledge production, entangled with narrative," a clear parallel to qualitative ethnographic techniques (Nafus 2018:17).

At the same time, this shift towards abductive and flexible machine learning techniques is not total within data science: aspects of proscriptive frameworks remain, in terms of personnel, objectives, habits, strategies, and evaluation criteria. But, seeds of this new thinking exist, with the potential to significantly reshape data science and possibly quantitative analysis in general. As a discipline, data science is in its adolescence, making significant fluctuations as the data scientist community defines its standard practices. This is not only a key time for anthropologists to immerse themselves into data science and encourage and cultivate localized, abductive techniques but also a strategic time to incorporate quantitative approaches within its localized, abductive ethnographic practice.

The view that anthropology and data science do not possess fundamental theoretical or philosophical differences forms part of my justification for incorporating data science techniques into anthropology,[4] but a reader does not need to agree on this to accept the incorporation of data science techniques that I argue for in this paper. Anthropologists seeking such incorporation have typically based their justification by arguing that both anthropological and data science techniques operate on the same plane. For example, for Curran, both are ethnographic, although differ in scope (2013: 63). Other anthropologists have criticized this, articulating an invigorated sense of the differences between the two (c.f. Madsen 2018:185). To show that this is not the

[4] As both an anthropologist and data scientist, I see most discussions on supposed theoretical differences between the two disciplines to reflect disciplinary sectionalism rather than actual theoretical or philosophical differences. For example, in undergraduate, when I double-majored in anthropology and mathematics, going between anthropology and mathematics classes every day, I noticed a clear social division but never encountered strong philosophical differences between each, despite the fact that many on both sides justified this social division through supposed differences in philosophical dispositions.

only justification for incorporating data science techniques, here is an extended argument by the anthropologist Hannah Knox, that anthropologists should incorporate such techniques precisely because of a theoretical divide between the two disciplines:

> The problem with much of the discussion about the relationship between data and ethnography to date is that ethnography has been held up as a 'richer', or 'deeper' form of description that does not suffer from the reductions and abstractions of numerical data (Ingold 2007). In this framing, ethnography and data science are treated as if they operate on the same ontological plane, with the ethnographic knowledge being deeper and better (from the perspective of qualitative social scientists) than descriptions based on numbers. Ethnographic descriptions are implied to be more 'real' than numerical abstractions. However if we take the relational understanding of ethnography that has been put forward here as our starting point then I suggest that we might find ways in which ethnography might be understood not as more or less real than numerical descriptions, but as operating with a different relational understanding of reality and validity. What differentiates ethnographic description from computer models of ecosystems is not the proximity of the story that is told to an actually existing reality but the relational assumptions upon which claims to truth are made. This framing, I suggest, allows us to move beyond an opposition between those disciplines that deal in quantification or number and those that deal in qualitative data or text. It opens the way instead to considering whether numbers such as those derived from sensors and measurement devices that inform data science and climate science might actually be able to be incorporated into an ethnographic style of truth-making. Ethnography is a form of description that entails different epistemological parameters from data science, but that

> does not, *a priori*, have to reject the incorporation of numerical data into this analysis. Because of its epistemological commitments, ethnography will never be a method that uses data to establish general conditions before the fact. But there are, I suggest, ways in which might be able to better incorporate data into our analysis of sociotechnical entanglements, and this is by treating numerical data as we treat ethnographic facts – as points on a web that we respin through description. (Knox 2018:144-145)

Data science is multiplicitous, full of diverse and at times conflicting orientations. Both the bottom-up, unique and localized, learn-as-you-go and top-down, universal modeling exist within it. It is in a uniquely formative period, developing into a traditional practice, the strategic time to influence its workings. Both Curran's argument that data science is interpretive and thus also ethnographic and Knox's argument that the two have fundamentally different relations to truth are accurate, despite their seeming contradictory nature since both articulate a true thread within a multiplicitous discipline.

Whether anthropologists incorporate their techniques, despite data scientists or because of their relationship with data scientists, abductive machine learning techniques in data science could be both compatible with and useful to anthropology. The developing methodological techniques in and of themselves may be reshaping quantitative practice, deconstructing the divide between quantitative and qualitative data and allowing the incorporation of quantitative techniques into ground-up and learn-as-you-go ethnographic practice.

The biggest advantage lies in how the incorporation of abductive and interpretative quantitative techniques could potentially bring vitality to the field of anthropology. For Dawn Nafus, data science methods of themselves have a potential significance, despite the specific

individuals utilizing them. Speaking specifically of Bayesian statistical techniques, she argues that:

> There are more likely positivists than post-positivists currently using those techniques. Bayesian approaches to computation nevertheless do something epistemologically significant for ethnographers. They leave open the possibility of situated knowledge production, entangled with narrative (2018:17).

She concludes that machine learning techniques "are not just 'bigger' than surveys. They are ontologically quite different, much more like the objects of material culture research or archival work," with "the potential to elaborate forms of knowledge that conceptualise social regularities quite differently from" universal top-down social theorizations (18-19). I hesitate to dismiss data scientists in mass because they are supposedly positivists or postpositivist (whether that is true), but an important argument underlies her point: these techniques have an importance at least partially distinct from those employing it, and anthropologists should take these methods seriously, despite coming from an opposing "methodological clan" (Seaver 2015:39). I argue that the best way to take the methods seriously is to incorporate the techniques into ethnographic and other forms anthropological research, and Nick Seaver's bastard methodological provides a helpful framework to do this.

2.2 Anthropology *of*, *over*, *with*, vs. *by* Data Science

Given that data science may be shifting the historic relationship between qualitative and quantitative approaches (Nafus 2018:19; Seaver 2015:36-37), the relationships various

anthropological scholars and their theories have formed with data science matters. Within the anthropological and related social science literature about data science, four common, not mutually-exclusive relationships emerge: anthropology *of* data science, anthropology *over* data science, anthropology *with* data science, and anthropology *by* data science (see Figure 1). These prepositional relationships refer to theories, research, and practical work that each explicitly or implicitly reinforce a particular union between anthropology and data science: anthropology *of* data science to anthropological analysis of data science and/or data scientists; anthropology *over* data science to perspectives that anthropology should oversee, regulate, or monitor data scientists or data science work; anthropology *with* data science to collaborative work between the two; and anthropology *by* data science to using data science tools to conduct anthropological work. They are distinct in that each reflects separate emphases for how anthropologists should approach data science but do not logically contradict each other: on a theoretical and practical level, many anthropologists and other social scholars utilize and/or advocate for a combination in their work.

Theoretical Relationship	Definition
Anthropology *of* Data Science	Anthropological, most often ethnographic, analysis of data science, data scientists, and/or data science work
Anthropology *over* Data Science	Perspectives that anthropology should oversee, regulate, or monitor data science work
Anthropology *with* Data Science	Collaborative work between the anthropology and data science
Anthropology *by* Data Science	Using data science tools to conduct anthropological work

Figure 1 Potential Relationships for Anthropology towards Data Science

In my experience, most anthropologists interested in data science have fostered an anthropology *of* data science in their theories and conceptualizations, most often leading to an anthropology *over* data science and anthropology *with* data science. Theoretical foundations that foster anthropology *by* data science relationship are exceedingly rare. This *by* relationship, however, would help us navigate the potentially shifting relationship between quantitative and qualitative research and to incorporate abductive, quantitative machine learning techniques, and we at Indicia utilized an anthropology *by* data science framework in our project. In this subsection, I will show how anthropology *by* data science is distinct from anthropology *of*, *over*, and *with*.

Anthropology *of* data science – that is, using anthropological theories and methods to investigate and understand and unpack data science, its historical roots and how it influences the world – is useful and commendable. This applies not only to formally published ethnographies of various data science communities and practices but also to informal usages of ethnographic practices in the context of working with or within data science teams – such as the common but not always published practices among business anthropologists to use their ethnographic interview skills to ask their teammates (including but not limited to data scientists) deep questions about a prospective project to understand the full context. The *of* relationship becomes problematic, however, when such work foster a distant relationship with data science and data scientists. In my experience, this most often occurs by encouraging a distant, neutral gaze as analyzer, or by transitioning into anthropology *over* and/or *with* data science.

Anthropology *of* data science can transition to anthropology *over* data science when anthropologists conceive of their social analysis as a corrective for, and thus, as having a superior place over the data scientists' conceptualizations and/or practices. In my experience,

academic and professional territoriality at least partially explains the tendency by anthropologists to advocate for their superior place over data scientists. For example, according to Nick Seaver, "critical algorithm studies is, essentially, founded in a disciplinary transgression" of what data scientists and other programmers can claim from their algorithms about sociocultural phenomena (2017:2). Critical studies scholars perceive data scientists as crossing the professional "boundaries of expert communities" within academia through their social and societal research (2).[5] When the focal point of the analysis is to understand how and why the social researcher "knows best," then I believe that developed relationship is problematic.

Phil Agre, the critical algorithms scholar, in his framing of an ethnographic analysis of general artificial intelligence systems (AI), provides an exemplary way to situate one's research. He views his work as collaborative criticism intended to refine and improve: for him, "critical analysis quickly becomes lost unless it is organized and guided by an affirmative moral purpose," and in his work, this is confronting "certain prestigious technical methodologies [in AI] that falsify and distort human experiences" (1997:xii), which he sees Heidegger's concepts of self as a corrective for (xiii, 4-5). Yet, instead of just positing his criticism against AI work, he strives "to expand technical practice in such a way that the relevance of philosophical critique becomes evident *as a technical matter*" through the synthesis of critical and technical work into a "critical technical consciousness" (xiii). In other words, he seeks to incarnate his theoretical framework into the discipline of artificial intelligence by presenting algorithmic alternatives for the development of computerized intelligence that incorporate his proposed theory. I have noticed a tendency for critical algorithms scholars to utilize their chosen theoretical

[5] Despite quoting Seaver 2017, I am taking his idea in a direction that he does not explicitly do.

framework(s) – whether Heideggerian, Foucauldian, Marxist, etc. – to demonstrate how what the computer scientists or engineers are doing is wrong (morally problematic and/or ineffective, depending on the scholar and context), which articulates a relationship of critical scholar as a distant, naysaying observer. Agre, at the very least, seeks to provide an alternative that addresses his criticisms. Thus, his critical scholarship starts with a relationship over artificial intelligence – as based on a moral criticism of the field as a whole – but ends as an exploration into how to do artificial intelligence with and through ethnography: the two mutually learn from each other through the incarnation of this moral impetus into an alternative technical practice (4, 24).

He provides the best example I have encountered for how to conduct critical work that also seeks to pragmatically and strategically incorporate one's disciplinary methodological techniques into his artificial intelligence discipline of study. In this case, this would be artificial intelligence *by* ethnography,[6] a potential parallel for data science *by* ethnography or anthropology. Within anthropology, this incorporative approach would manifest as anthropology *by* data science.

Collaborative anthropology *with* data science frameworks still maintain and reinforce the distinction between the two disciplines; whereas, anthropology *by* data science seeks to traverse the established divide between them. For example, many scholars and business practitioners have advocated for anthropology *with* data science based on complementary roles between the two, particularly based on the view that ethnographic is "thick" and data science is "thin" (c.f. Seaver 2015:36; Madsen 2018; T. Wang 2013). In these accounts, ethnography and data science

[6] This would exemplify artificial intelligence *by* anthropology, but Agre not being an anthropologist, it is not a technically an example of it; hence I use the phrase "artificial intelligence *by* ethnography."

techniques possess a separate but complementary role, needing each other to address the other's weakness (Seaver 2015:36).

While incorporating the two is certainly strategically intelligent to provide a full understanding of the individuals and sociocultural phenomena studied, anthropology *with* data science frameworks do not always go far enough to change the modular nature of such work and thus reinforces the existence of both anthropology and data science as separate disciplinary "kingdoms." In reifying the boundaries, this collaborative approach still (re)creates the conditions of the turf wars, which have partially led to the sense of disciplinary transgressions discussed above.

Describing the two as separate but coming together still maintains the historic relationship between quantitative and qualitative data and techniques (Seaver 2015:37), instead of using abductive machine learning techniques as an impetus to rethink this distinction and develop quantitative techniques that are ethnographic. In occupational settings, whether academic, corporate, or other settings, the conceptualization of the two as complementary still maintains the juxtaposition between the quantitative and qualitative approaches, which has facilitated their establishment as opposing and competing disciplinary and organizational structures. In so far as anthropology *with* data science could be an intellectual halfway-house towards the radical breakdown of their barriers through their incorporation – by say, providing a space to reform positive interactions and cross-learning to do the latter – then anthropology *with* data science is beneficial, but in my personal experience, the advocacy and institutionalization of anthropology *with* data science has become an end to itself, actually stifling their incorporation by reifying their boundaries.

2.3 Bastard Ethnography

Nick Seaver (2015), through his depiction of data science as bastard algebra, develops the epistemological concept of a bastard discipline, which provides a framework for anthropology *by* data science in the EPIC project. It helps anthropologists creatively reimagine of the relationship between and across ethnography and data science techniques, encouraging ways to strategically incorporate the latter into the former.

He presents two potential examples of bastard disciplinary practices: data science as a "bastard algebra" and anthropologists conducting "bastard ethnography." "Data scientists work as professional bastard-makers, combining data sets, algorithms, and epistemologies in unauthorized ways to produce illicit offspring," forming data science into a mishmash of various "people, epistemologies, and methods" (including mathematics, computer science, engineering, economics, sociology, etc.) brought together to solve specific problems (43). Likewise, anthropologists created ethnography as a bastard practice especially in the first half of the twentieth century, "breeding descriptions from illicit encounters, mixing conceptual schemes, and stirring the blood of experience with the ink of theory," whatever worked to help understand cultures and societies (44).

Bronislaw Malinowski originally coined "bastard algebra" to describe mathematically-inspired kinship ethnographies popular in the 1930s, which he considered illegitimate offspring of mathematics and anthropology (Seaver 2015:38).

> Mathematicians would not claim this algebra as their own, and "the average anthropologist," as Malinowski writes, would not either. To this day, many of us [anthropologists] remain uneasy about this mixture, and to borrow another bit of kinship

> terminology, we tend to establish avoidance relationships with mathematics, when we are not dismissing it outright. Like anyone else, anthropologists are concerned with regulating our kinsmen. (38-39)[7]

Out of this, a "pure" concept of ethnography arose as a qualitative technique, and like a royal lineage, anthropologists revised their bastard ethnographic practices into an undiluted, noble disciplinary lineage.[8] Even though some anthropologists have utilized some quantitative techniques (Nafus 2018:4-5), anthropologists have historically defined the strengths of their ethnographic practice in contrast to and/or as a complement of various popular quantitative disciplines, such as cognitive psychology, quantitative sociology, and behavioral economics (Seaver 2015:36). Thus, "it should not be surprising that ethnography and big data appear as neatly opposed methodological moieties" that intersect in the rehearsed script of collaborative complementariness (36), instead of investigating the distinctiveness of abductive machine learning techniques among other quantitative approaches. "Anthropologists encounter data science as something utterly foreign to the practice of ethnography – a venture-funded epistemology that encroaches on our disciplinary territory, operationalizing concepts like 'culture' in troublesome ways," yet especially in corporate settings, they complement each other: ethnography "putting flesh on big data's bones" (36). In this disciplinary moiety system, the two *must* be distinct familial lines to form a complementary union: a yin to the other's yang.

---

[7] As someone who would consider himself both an anthropologist and a mathematician, I have seen these avoidance techniques (and reciprocal avoidance techniques among mathematicians) very frequently.

[8] I use *disciplinary purification* as the implied alternative to Seaver's bastard disciplinary approach: the attempt to develop, refine, or synthesize a discipline into a wholesome and well-defined practice or set of practices, typically presented as separate from and/or superior to those of other disciplines. The term is my logical extension of his concept of bastard disciplines and not his lexicon.

Such a complementary union is the most common model I have seen for anthropology *with* data science (see Seaver 2015:36 for his specific examples of this model in the anthropological literature, and see also T. Wang 2013, Madsen 2018, Norvaisas 2014, and Slobin 2010 for more examples). This simplistic, collaborative *with* model lacks any depiction of anthropology *by* data science, since conducting data science work within anthropological research would contradict the anthropology's familial purity and its complementary relationship. These anthropologists have many important insights about cross-collaborative work both with data scientists and those from related disciplinary backgrounds, but this overall model fails to merge each discipline's practices. Anthropologists would best do that by viewing anthropology as a non-pure, flexible bastard discipline and incorporating data science tools to understand complex sociocultural and socioenvironmental phenomena – that is, anthropology *by* data science. Instead, this complementary model regurgitates the "commonsense relationship" between anthropology and any quantitative discipline (Seaver 2015:43), which fails to consider how machine learning and other data science techniques are changing the nature of quantitative research and thus the historic relationship between qualitative and quantitative methods (Nafus 2018:19).

The bastard approach's rigorous flexibility would help anthropologists as they navigate this potentially shifting terrain. Not only do anthropologists need to reimagine the potential relationships during this shift, but this is also a strategic time to incorporate and redefine abductive quantitative techniques into our fold of ethnographic research. A bastard approach helps with both. As a secondary benefit, a bastard approach harmonizes with the values, frameworks, and techniques within data science, given that data science itself, as a bastard discipline, has developed based on what works to solve specific problems. Our project, for

example, became an epistemological bastard: a crossbreed between several disciplinary clans based on what would work best to address the specific problem at hand.

Section 3: Project Summary

Task 6 of the EPIC project is an exploration into how to utilize data science tools – machine learning decision tree modeling and random forests – alongside a conventional anthropological tool – ethnographic decision tree modeling – to understand a sociocultural phenomenon.

3.1 Project Overview

The table below (Figure 2) summarizes each task. At the most basic level, we analyzed how the relationship between people's relationship to technology connects with and potentially influences energy consumption behavior across diverse Californian populations and geographies, with the goal of developing energy saving campaigns that better reach various individuals and foster consumer energy saving in the overall Californian population.

Our project involved a slow shift from small-scale (that is, with small sample size) qualitative approaches to larger-scale statistical and quantitative tools: transitioning from in-home participant-observation and to interviews (Tasks 2-3), to quantitative survey analysis (in Task 4-5), to statewide population analysis through data science. Task 6 functions as a transition point between the survey and population analysis: in it we develop the analytical infrastructure –

decision tree models – to enable us to research at the latter scale. Starting with small-scale qualitative techniques and incrementally enlarging the scale through more complex quantitative techniques is a powerful strategy for conducting anthropology by data science.

Task #	Timeline	Task Name	Central Question(s)	Description
1	June 2015-Sept 2018	General Project Tasks	• How do we break down our project, and what is its timeline?	Developed project scope and timeline
2	July 2015 – July 2016	Documenting and analyzing emerging attitudes, emotions, experiences, habits, and practices around technology adoption	• How do we understand these cyber status categories? • What do they look like ethnographically?	Conducted ethnographic research to observe patterns of attitudes and behaviors among cybersensitives and cyberawares.
3	Sept 2016 – Dec 2016	Identifying the attributes and characteristics and psychological drivers of cybersensitives	• Do the observed behavioral and attitudinal differences in Task 2 relate to relate to detectable behavioral and attitudinal characteristics?	Conducted in-depth interviews coding for psych factor, energy consumption attitudes and behaviors, and technological device purchasing/usage.
4	Sept 2016 – July 2017	Assessing cybersensitives' valence with technology	• Can we also verify these differences statistically? • Do they relate to other known characteristics of these individuals (such as income, race, gender, etc.)?	Tested for statistically significant differences in demographics, behaviors, and beliefs/attitudes between cyber status groups
5	Aug 2017 – Dec 2018	Developing critical insights for supporting residential engagement in energy efficient behaviors	• Do the cyber groups consume energy differently? • Do cybersensitives and cyberawares. consume less energy than the general population, on average?	Analyzed utility data patterns of study participants, comparing them with the general population.
6	March 2018 – Aug 2018	Recommending an alternative energy efficiency potential model	• How can we determine someone's cyber status? • How can we model cyber status visually and computationally? • What benefit would segmenting cybersensitives and cyberawares. possess?	Constructed decision tree models to classify an individual's cyber status
7	TBD	Evaluation of Project Benefits	• How can we analyze cyber status and energy consumption in a wider population?	TBD

Task #	Timeline	Task Name	Central Question(s)	Description
8	TBD	Technology/Knowledge Transfer Activities	TBD	TBD

Figure 2 EPIC Project Task Summary (c.f. Indicia Consulting 2018c)

After initially scoping out the project contract, overall trajectory, and timeline (Task 1), the project's research started as Task 2: an initial ethnography of 45 households in Southern and Northern California. In Task 3, two team members conducted extended interviews about their technology usage and energy consumption and observations of these same features in their homes, coding the occurrences and/or intensity of the following three categories: "psychological factors (Psych), energy consumption attitudes and behaviors (Energy), and device purchase and usage (Device)" (Indicia Consulting 2014).

From the ethnographic observations and in-depth interviews, the team noticed patterns in how people related with technology and energy consumption behaviors, leading to the creation of cyber status categories. *Cyber status* are psychosocial characteristics relating to one's perceived emotional connection with technology and resulting behaviors. We tabulated the number of each of the codes, Psych, Energy, and Device, and used Atlas.ti to cluster these codes into the following five cybersensitive groups (ranked from least to greatest in terms of overall frequency of codes): cybersensitive, cyberaware, mainstream, low mainstream, and null (Indicia Consulting 2018a:12).

Cyber status is a psychosocial categorization. Psychosocial categorizations are ways of organizing a customer base along characteristics like lifestyle propensities and the purchase and usage patterns of products -- in this case personal technology, such as smartphones, tablets, laptops, or wearables. We hypothesized the X Factor to be a cybernetic behavioral/personality trait we labeled 'cybersensitivity' where the cybernetic aspect referred to the emotional

relationship between a person and their personal technology, and a heightened propensity for acting on information delivered via that device, aka feedback, where feedback is a cybernetic construct (Indicia Consulting 2018a:2).

In Task 4, we surveyed 400 Californian residents to test the behavioral and attitudinal patterns among cyber status groups against the population. The survey increased the number of research participants and the introduction of quantitative methods. Indicia initially hired me in March 2017 to analyze the survey data. The survey's goal was to determine:

1. Whether these cybersensitives and cyberawares (and to a lesser extent nulls) exhibited statistically significant differences in behavior and attitudes towards technology than the overall population.
2. Whether cyber status or these observed behaviors and patterns correlated with demographic characteristics (such as gender, income, race, age, occupation, etc.).

We concluded that the categorical clustering of these attitudinal and behavioral characteristics were psychosocial qualities within any Californian population, distributed independently of demographic characteristics.

In Task 5, we used interviewees' utility bills to analyze monthly energy consumption, furthering the degree of quantitative analysis through in-depth statistical analysis of consumption data. From this, we determined whether cyber status groups exhibited statistically significant differences in monthly energy consumption. Cybers (which refers to those who cybersensitive or who are cyberaware) within our ethnographic study consumed less energy per month on average than both the other three groups within our study and the regional populations. This result reinforced the need for and effectiveness of segmenting cybers for energy saving campaigns: targeting cybersensitives and cyberawares with energy saving programs would reduce overall

energy consumption in California. But, its conclusions are limited to just cybers in our ethnographic study.

In Task 6, we sought to determine whether this finding that cybers consume less energy on average also holds with the entire Californian population. In Task 5, we were only able to analyze the data from members of our ethnographic study. This produced enough data to be statistically rigorous, but we were unable to generalize our results across the whole regional and statewide populations, a significant limitation.[9] We developed the cyber status categorization system through several hour-long interviews and in-home observations. Such a time- and resource-intensive process is only doable for a small group of people, but to test energy consumption patterns across an entire statewide population, we would need to determine the cyber status of thousands of individuals. We needed a less intensive, scalable strategy for determining consumers' cyber status.

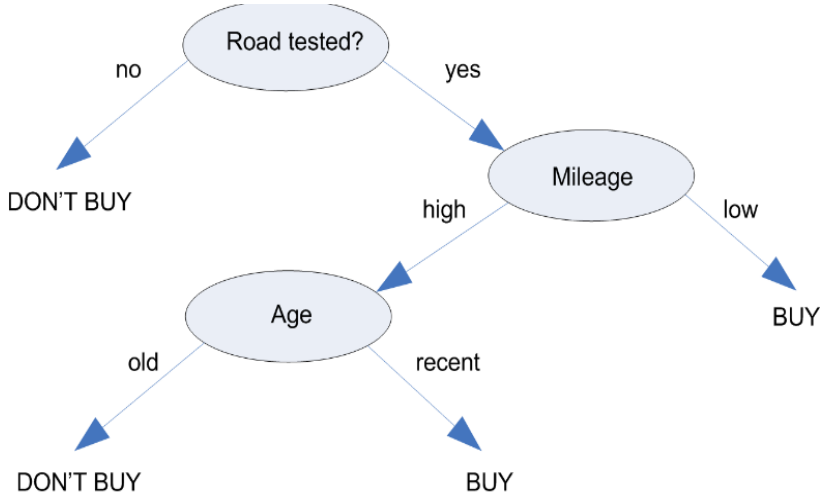
The goal of the decision tree model was to represent the criteria used to classify the cyber status of individuals, which we could then use to determine the cyber status of statewide (or nationwide) population and test whether cybers consume less energy on average. As shown, Task 4 and 5 involved strategic introduction of statistical techniques, with an increasing level of abstraction and participant scale. Task 6 involves the first introduction of data science techniques as our scale enlarges to the entire Californian population.

---

[9] Difficulties obtaining permission for the utility records further shrank the group for Task 5. In addition to the normal permissions required for consent for social research, the utility companies tried to frustrate the attempt to provide the data by requiring additional confirmations. Thus, we had to go back to the participants multiple times to receive consent. This naturally reduced the number of people who followed through with each successive bureaucratic hoop. Of the two major utility companies who served the regions where our participants lived, one company was more forthcoming than another, meaning the participants we did get utility data from tended to skew towards that company and towards the areas that company served.

3.2 Decision Tree Modeling

Figure 3 Example of a Decision Tree Model (IBM 2012; Indicia Consulting 2018a:6)



Decision trees model a decision-making flow through a hierarchy of generally binary Boolean decisions or criteria, such as true/false or yes/no conditions (Indicia Consulting 2018a:6). There are two broad types of decision tree modeling:

1. Ethnographic decision tree modeling on qualitative data
2. Machine learning decision tree modeling on quantitative data (Indicia Consulting 2018a:6)

Both techniques are abductive: building a unique model as they iterate through data. "In brief, the object of the game is to frame criteria, order or arrange them into a tree-like structure, and then test and revise the tree model" (Indicia Consulting 2018a:7, quoting C. Gladwin 2001).

Ethnographic decision tree models (EDTMs) use ethnographic interviews and observation to develop the decision-making process for a set of individuals on an activity, subject, or issue. EDTMs are “qualitative causal analyses that predict real, episodic behaviors, rather than…the intent to behave in a certain way" (Indicia Consulting 2018:7, quoting Ryan 2006). Christina Gladwin (2001:28-31) developed a detailed procedure for how to construct a decision tree model from interviews and observations, which we used as a template for our project:

> Researchers analyze and consolidate multiple informants' explanations of their decision-making processes into an overall decision tree model. "There are direct and indirect eliciting methods, but both require the ethnographic model builder to look for contrasts in decision behavior, ask the informant to explain the contrast (e.g., 'Why did you decide to evacuate with Hurricane Andrew but not with Hurricane Erin?') and then test that explanation on another informant" (C. Gladwin, 2001). (Indicia Consulting 2018a:7)

Classification and Regression Trees (CART) is a machine learning process for developing a predictive classificatory model through a set of machine-learning algorithmic strategies, most commonly built with quantitative data.[10] CART modeling centers around branch splitting: the "process to decide when to split into a new branch and which variable to choose to do so [with]" (Indicia Consulting 2018a:8). The process recursively loops through each subgroup of the data until all groups are homogeneous (or in some cases, the programmer provides an arbitrary stopping point).

> CART models typically involve three basic steps: pre-development, model development, and pruning through testing and/or [using] ensemble methods. Data cleaning refers to the process of organizing the data for the model, including developing independent and dependent variables, splitting the data into training and testing sets, resampling the data, etc. The next step is to develop the decision tree model. The most important consideration in this process is the method used to determine branching: that is, the equation and algorithm used to determine when to split the data into smaller branches and the variable

[10] Classification decision tree algorithms develop trees use categorical and ordinal data, and regression tree algorithms use continuous data. Because our data is primarily categorical or ordinal, we built a classification tree.

> to use to split it. After creating the initial model, developers test the accuracy model with cross-validation techniques and based on those results prune (or strategically adjust) the tree further. Various ensemble methods like random forests, bootstrapping, boosting, and bagging, are particularly effective [supplement to] pruning. This refines the nodes on the branches to ensure better accuracy and reliability (on both the data set and on potentially new potential data sets respectively) (Indicia Consulting 2018a:8).

For our project, decision tree modeling was the logical means to integrate qualitative and machine learning techniques, since decision tree modeling itself is an adaptive visual model – a crossbreed – made up of complexly integrated parts, each useful to several different parties with distinct disciplinary methodology (C, Gladwin 1997:8). It models the bastard disciplinary approach Nick Seaver described and exemplifies how including data science techniques as part of ethnographies can remain true to ethnography's cross-disciplinary dynamism.

Decision tree modeling provides not only an excellent means to classify a large population according to a category like cyber status but also an easy-to-read visualization tool to see the criteria that determine cyber status. This makes replicating and further testing through a variety of methods (including ethnographic, statistical, and data science-oriented) easy. Decision tree models are understandable because they tell the story of how to decide the group, not just the result – which most other machine learning techniques fail to do. Thus, they are a useful tool to help develop policy initiatives around segmentation.

Individual branches, which make up a decision tree model, are also themselves decision trees. This recursive layering enables intermixing of the best sections of several branches. Thus, one can combine qualitative and machine learning decision tree models into a new tree, garnering the advantages of both. Adaptively combining the best elements from traditional

ethnographic research (such as the ability to garner decision-making from individuals' explanations and observed actions) and data science techniques (such as the ability to prune and refine branches of the tree to improve accuracy) into a unique new crossbreed illustrates my vision for how to integrate the two disciplines. Through an anthropology *by* data science relationship, ethnography absorbs the advantages of several data science techniques through strategic absorption into ethnographic practice.

3.3 Methodology and Results

EDTM and CART modeling each have distinct methodologies, which I will unpack in this subsection to show how we combined them to conduct anthropology by data science.

3.3.1 EDTM Methodology and Results

Our EDTM model had three phases:

The first phase of EDTM development always consists of conducting a series of ethnographic interviews with the members of the group under scrutiny. The interviews are designed in such a way as to elicit the process of decision-making in the words of the group themselves. In the second phase of EDTM development, the ethnographer(s) reviews the verbatim responses and organizes the steps in the decision-process as captured in the ethnographic data. The third phase of EDTM development is to run the now diagrammed set of choices past another, similar, set of group members, to see where the ethnographer may have misunderstood or missed pertinent information. If the EDTM

> model is generally predictive (the literature suggests that a minimum 80% of cases should be properly accounted for by a well-done model), then the modeler can stop, although ideally the process of refinement can continue *ad infinitum*. (Indicia Consulting 2018a:11-12).

Another team member took the lead in developing the EDTM, and I only had an advisory role on this process.

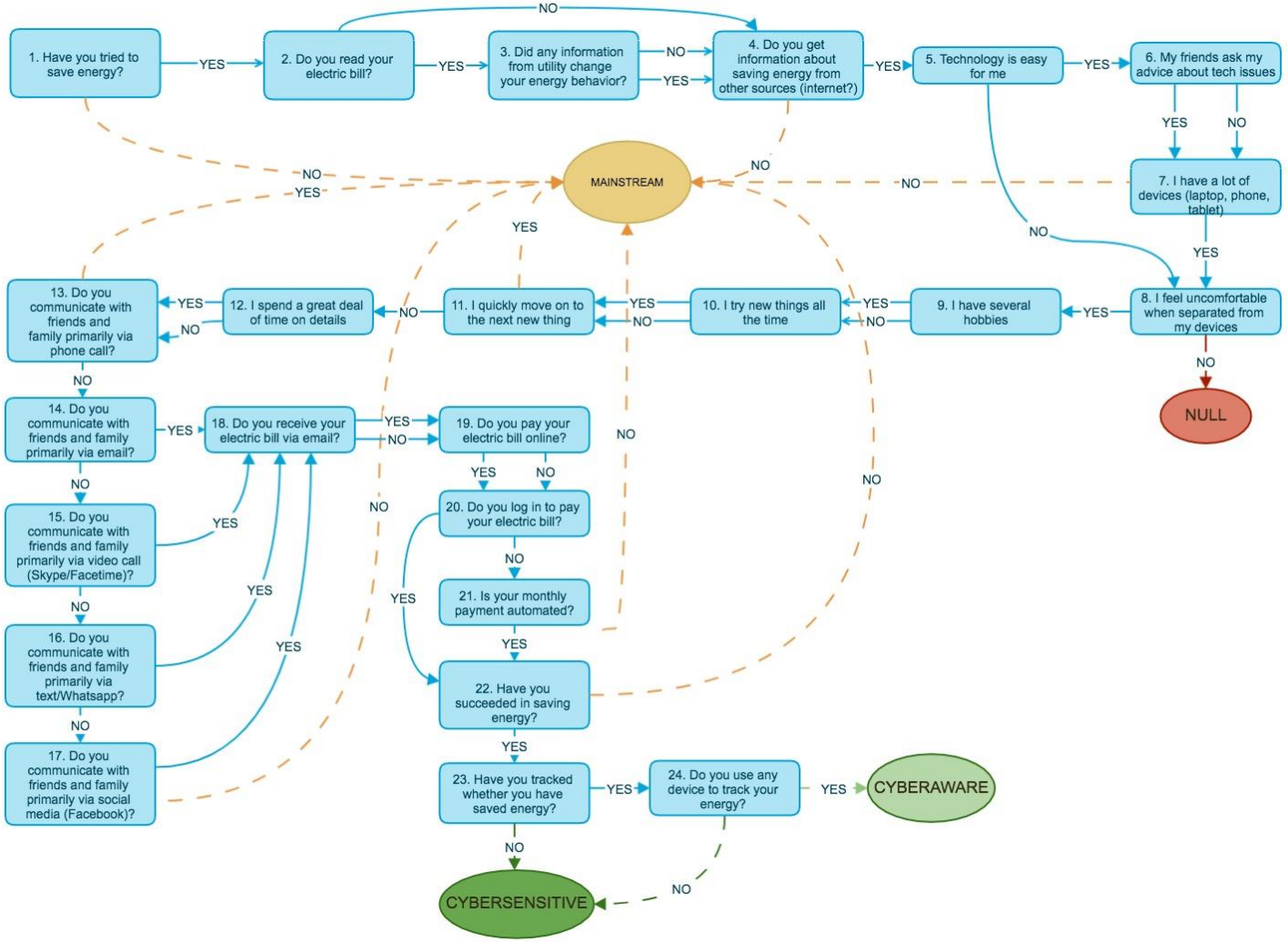


Figure 4 EDTM Model (Indicia Consulting 2018a:18)

3.3.2 CART Methodology and Results

I was the primary point-person for developing the CART model. I developed a CART model to classify participants according to cyber status and further used random forest modeling to improve the accuracy. My approach had three basic phases – Pre-Development, Development, and Pruning. For pre-development, I conducted bootstrap resampling to create a properly-sized dataset for CART analysis.

During the development state, I used Gini index function and a maximum branch of 8, testing the trees accuracy through Leave-One-Out-Cross-Validation (LOOCV). After testing between several different branching options and maximum depths, this produced the most accurate results. LOOCV goes through each value in the data, builds the model without that value, and tests to see whether that built model is accurately able to predict that value (Indicia Consulting 2018a:15). Our model had an accuracy of 76% (Indicia Consulting 2018a:16).

Pruning refers to ways to refine the built model to improve accuracy. Pruning conventionally refers to refining the tree by adjusting parameters and/or hyperparameters, but I also informally include use of ensemble methods as part of pruning, since data scientists often employ ensemble methods concurrently with pruning and since they also have the intent of improving model accuracy. “We employed an ensemble method called random forests to address overfitting. Ensemble methods combine several models together to improve the results, based on a collaborative approach” (Indicia Consulting 2018a:16).

> A random forest algorithm produces several CART decision trees (called forests, because they [are made of]many trees) based on randomly selected subsets of data points within the sample. For each data point, the random forest model then classifies based on the

> mode classification among all the trees, that is as the attribute [with the] most frequent classification. For example, if most trees constructs classify person X as a cybersensitive, it classifies him/her as a cybersensitive.
>
> Because random forests compare several different trees, they allow for significantly improved accuracy in their classifications, known to possess exceedingly accurate results. By creating and testing between several CART decision trees, random forests address potential tendency towards variation and overfitting inherent in decision trees, since they pick up the wide patterns between many specific decision tree models, filtering out both any irregular variation or narrow, over focus of a particular model by looking at the patterns of the models as a whole. (16-17)

Our random forest had an accuracy of 100%, meaning that it could correctly classify all research participants in our study. "Data scientists are most interested in which variables have been the most significant in helping to filter individuals, in other words, which have been most central variables when trying to classify all individual data points among all the trees" (quoting Bell 2018 in Indicia Consulting 2018a:20). The following graph shows the ten most significant variables for the random forest model when determining cyber status of an individual:

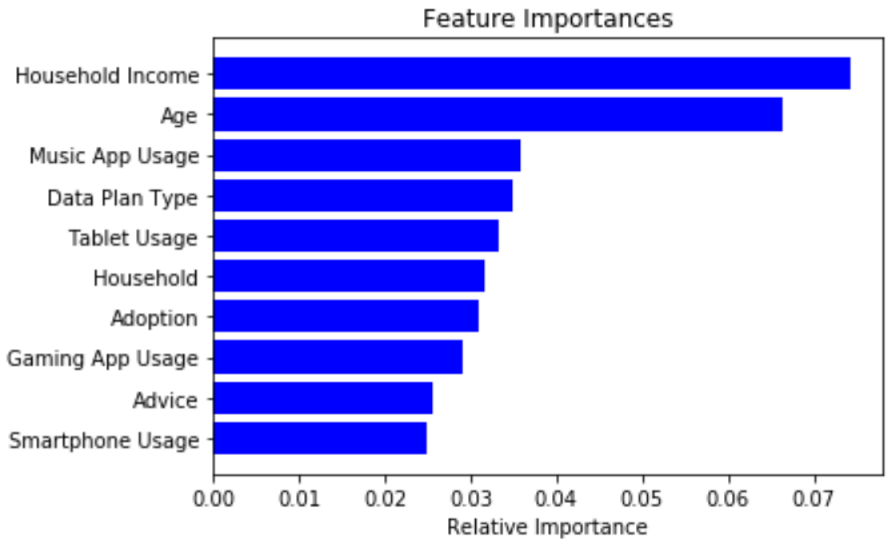


Figure 5 Random Forest Top Ten Most Important Variables (Indicia Consulting 2018a:21)

Among the trees generated in the random forest, the following tree was the most accurate, with an accuracy of 100%. As discussed, the random forest still aggregated the results from one thousand trees, of which this tree is just one tree it uses, but as the most accurate machine learning generated tree model developed so far, a graphical version of it is included below (21):

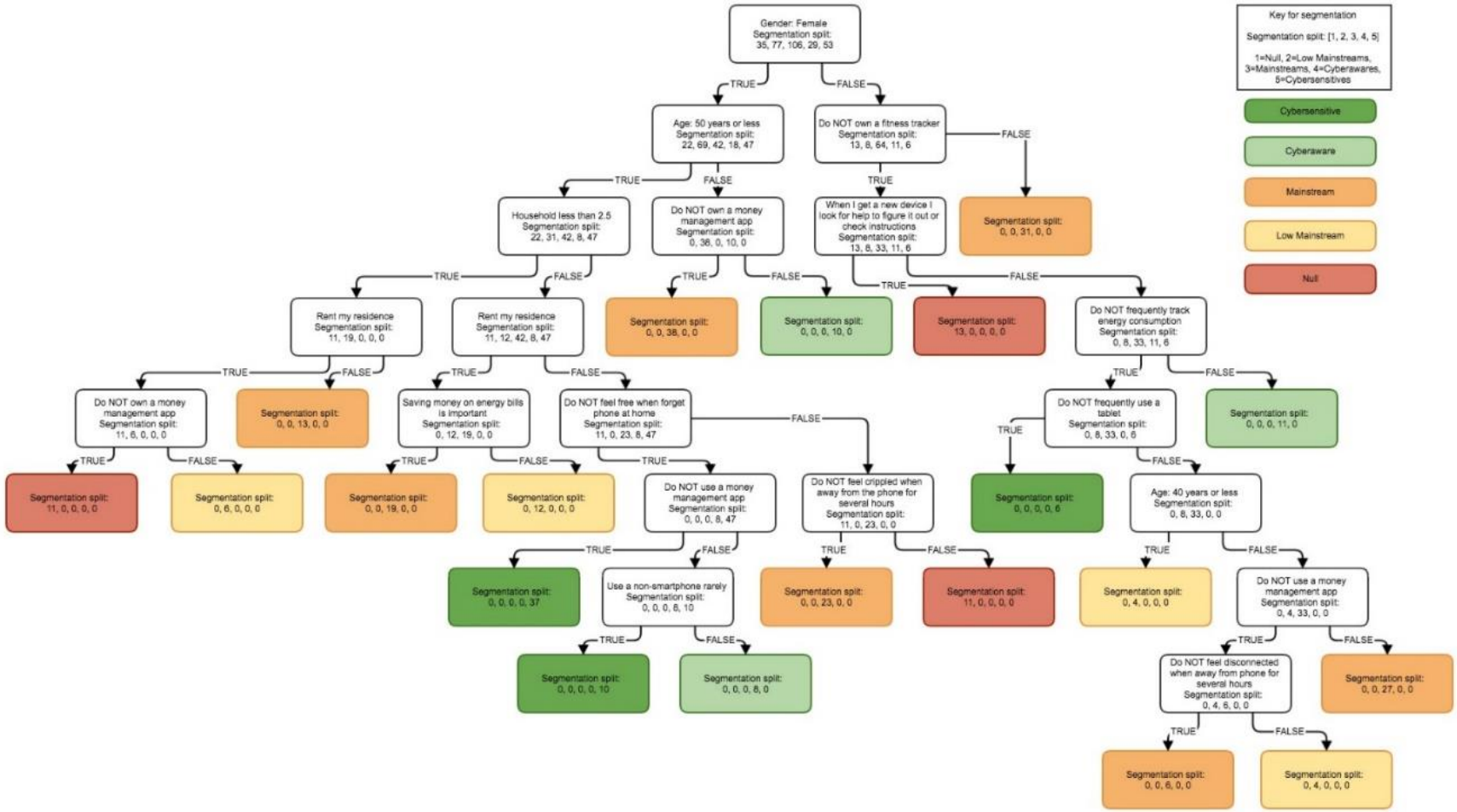


Figure 6 Most Accurate CART Model from Random Forest (Indicia Consulting 2018a:21

Section 4: Analysis of Both Methods

As an exploratory approach into conducting anthropology *by* data science, reflection on the project's strengths and weaknesses will not only refine this project but also drive ideas for future ethnographies incorporating data science techniques.

4.1 Methodological Advantages and Disadvantages

The limitations and strengths of EDTM and CART illustrate the specific fault points along which the two approaches could connect. EDTM offers a way to qualitatively synthesize the perspective of many in a group into an easily visualized decision framework. The models are hybrids, and their procedure is both replicable yet mutable to cover a wide variety of situations (C. Gladwin 1997:11-13). Their accuracy and explanatory power are also directly testable (13). The model is also encodable in parts: a researcher can easily convert the whole or any subsection into an algorithm for cross-comparative testing or assemblage with machine learning models.

"Human beings are complex, and because humans live in social groups like households, predicting energy consumption behavior at the household level from a few dozen questions is an imperfect art at best," its major disadvantage (Indicia Consulting 2018a:22). Its traditional application is to model collective decisions, and it fails when decision-making criteria do not fit easily into a stable, hierarchical structure understood and/or enacted by the individuals (Murtaugh 1984:243-244; Ryan 2006:103-104; Plattner 1984:252-253; Mathews 1987:54-56; Bauer 1996:190; Harris 1974:242; H. Gladwin 1984:217). The following situations do not fit

neatly into such a framework: 1) cases where no stable decision-making process exists between individuals and/or for one individual, 2) cases where a hierarchical decision-making inadequately represents decisions. Our EDTM model does not represent a decision-making process but a classification process, meaning that critiques of the practice as a decision-making modeling do not directly apply. But, our EDTM model still suffers the limitation of assuming a relatively stable transference of attitudes and behaviors into psychosocial characteristics. EDTMs "are qualitative, causal analyses that predict real, episodic behaviors, rather than—as does so much social research—the intent to behave in a certain way" (Ryan 2006:103).

CART decision tree modeling is also flexible method, whose results are easy to visualize and understand intuitively (unlike many other machine learning processes with obscure, unintelligible processes)" (Indicia Consulting 2018a:23). They "require little initial assumptions or requirements on the data" (23), being workable no matter its type: categorical, ordinal, interval, etc. It can also handle a mixture of types. CART decision models also "do not require normalization or extensive data cleaning to prepare the data" (23). The models "provide a flexible, easily manipulatable foundation for random forests and other methods that combine several different models/approaches" (23). Being easy to create and edit/prune, more complicated ensemble methods like random forests, boosting, and bagging can easily use them as a foundation (23).

CART's major disadvantage is overfitting and high variation. Overfitting, where "the model describes the specific data so accurately that it fails to generalize well to the overall population." These models "can become too large – that is, have too many branches – which can reflect considerations that are specific to the small group but do not occur for the population." (22) Small changes in data can also lead to a completely different tree. These slight variations

can lead to shifts in the variables decided on to branch with and because of decision trees' hierarchical structure, these shifts influence all future decisions down the tree. The best ways to address overfitting and variation is to increase the sample size (which as we intend to do when testing, as discussed in Section 4.2) and to use ensemble methods like random forests.

4.2 Project Modifications

Hindsight provides a clearer perspective for reflection than the moment, and hopefully, these reflections on the strengths of our approach and aspects to modify will benefit anyone conducting a project like this and provide helpful considerations for future anthropologists interested in combining ethnographic and machine learning decision tree modeling. My intent is that these reflections will stimulate conversation into how to use machine learning decision tree modeling and other data science techniques in ethnographic research in the future.

Indicia and the California Department of Energy initially developed the scope of work in 2015 when outlining all the tasks for this wider project. Their initial idea was to create a decision tree model to visualize cybersensitive and cyberaware energy consumption and from that represent and predict their energy consumption. Over the course of the project, however, the goal for the decision tree models shifted to classifying individuals according to cyber status. As discussed, Task 5 demonstrated the need to classify by cyber status for future analysis, which we addressed through these models. Ideally, we would have articulated the project this way from the start, but given that the specifics of a project task must develop as the project unfolds, we were unable to foresee that several years ago when developing the project scope.

If I could go back to the beginning, however, I would have consistently articulated the project's goal as what it became. As the individual who specialized in these data science approaches, ideally, I would have been involved in the project planning to ensure that the decision tree modeling plan addressed the correct set of questions for this technique and ensure that Indicia had planned the requisite actions to develop the model fully. I would have avoided the initial set of questions as beyond the scope of decision tree modeling, either advocating that we conduct decision tree modeling on classifying individuals by cyber status like we ended up doing, or arguing that we should employ another set of techniques altogether to address those initial set of questions, depending on the team's priorities.

For the latter, I would have directly incorporated testing into the decision tree modeling development. "Ideally, there would be two additional rounds of recruitment and testing, with two different samples: one to initially test assumptions and refine the model, and one to validate the findings. However, that lies beyond the scope of this project" (quoting myself in Indicia Consulting 2018a:10). Most studies that used either form of decision tree modeling further tested their initial model against a wider survey (Bell 2018:3; Ryan 2006:103; Mukhopadhyay 1984:227; Bauer 1996:187-190; Murtaugh 1984:246; Yu 2010:1639-1640). This was beyond the available resources for our project, however, but we may have been able to incorporate it if we had planned to do so from the beginning.

In EDTM, testing is frequently the next phase in the research process after decision tree development; whereas, for machine learning-based decision tree approaches – both CART and ensemble methods like random forests – testing should be concurrent with model development. "After creating the initial model, developers test the accuracy model with cross-validation techniques and based on those results prune (or strategically adjust) the tree further…. This

refines the nodes on the branches to ensure better accuracy and reliability" (Indicia Consulting 2018a:8). Data scientists typically do this testing both within their dataset by splitting their data into training and testing data and/or against another dataset(s). Testing against other data or datasets is particularly important for addressing overfitting, illuminating which features are over-specific to the original data. They typically juxtapose the internal dataset with an additional testing dataset to find a healthy balance between being accurate to the original data yet also accurately generalizing to the larger population (7-8).

A statewide or region-wide survey to test our data was beyond the project's scope and resources to develop. Indicia did not foresee this potential need when developing the project in 2015, but even so, the Californian Department of Energy would not likely have approved such an expensive endeavor in addition to all our participant-observation, interviews, and other activities. Instead, we decided in Task 7 to test our decision models against an already produced representative survey of Prince William County, Virginia, from a project with similar data by Virginia Tech University (Indicia Consulting 2018a:17).[11] Testing both models retroactively will help refine them, but if it was practical, I would have recommended that we incorporated testing into Task 6. In reality, however, I, the only data scientist on the team, joined years after the initial project development. For this project, that is water under the bridge: unable to go back in time, we must go forward where we are.

Finally, I would integrate the ethnographic and machine learning decision tree modeling during the development of each phase of research. The project exemplifies parallel research, where researchers enact qualitative and quantitative research approaches separately by in tandem

[11] I would like to thank the Network Dynamics Simulation Science Laboratory at Virginia Tech for sharing their data on the synthetic population, along with the detailed energy demand profiles.

(Johnson 2007:114, 119). We intend to integrate them during testing in Task 7: testing over a larger dataset enables us to prune or refine the trees, and both qualitative-based coding and quantitative-based machine learning pruning techniques will be helpful in this process. Each decision tree methodology requires a significant amount of legwork to develop – qualitative work for the EDTM, and quantitative and computational work for the CART and random forests. Because this intensive legwork lies solely on each's respective side of the qualitative-quantitative divide, integrating the two during the initial model development ultimately proved unhelpful.

Decision trees already invoke an implied horticultural imagery through horticultural terminology like *trees* and *pruning*, and I will elaborate on this analogy to make both the need for and proper space for the potential integration of the two approaches clearer. Like seeds, each methodology has a pre-packaged set of procedures employed to create or grow the initial tree. After creating them, a horticulturalist can then graft the trees – or specific branches of the trees – together as part of the pruning process. Analogously, researchers may employ the ready-made ethnographic and machine learning methods to develop the decision trees and then an integrative approach to prune them. Its goal would be to create a model that best matches the testing data by taking the most accurate parts of each decision tree model.

Section 5: Conclusion

Abductive and interpretative machine learning techniques have grown substantially within data science, which may shift the terms of the quantitative and qualitative debate (Nafus 2018:16-19). Anthropologists should incorporate these techniques into their ethnographic toolkit

to draw from when applicable in their work, pursuing what I call anthropology *by* data science. This would require a shift in how anthropologists conceptualize and relate to machine learning techniques: away those fostered in the "historic" quantitative and qualitative divide. Nick Seaver's bastard methodology allows for the conceptual flexibility to rethink "old-school" relationships and forge new methodological connections.

Indicia's ethnographic EPIC project is an exploration into how to conduct this type of anthropology *by* data science work: specifically, how to utilize machine learning decision tree modeling in an ethnographic research project. In this ethnography, our tasks slowly transitioned from small-scale qualitative techniques towards large-scale quantitative, statistical and data science techniques. In Task 6, we introduced CART and random forest analysis to help classify individuals according to cyber status, which in future tasks we will use to test statewide and potentially nationwide population samples. Our project had significant limitations, which we will seek to address in future tasks. It is only an initial exploration into how to conduct anthropology *by* data science, and future attempts to do this, by us and by other anthropologists, would help develop strategies to address these potential weaknesses and to further its strengths.

I do not base my argument that anthropologists should incorporate specific tools of data science on a sense that data science or specific machine learning techniques will transform the world - whether into a utopia or dystopia. These mythological narratives seem naïve and inaccurate (Paff 2018:13-16; also c.f. Ziewitz 2016:6-7). Rather, I base it on the view that this cross-pollinating disciplinary approach would help foster intellectual and methodological creativity within anthropology. Universalizing, top-down social research techniques are often problematic, and I think that anthropologists and other social researchers countering them is both necessary and beneficial. But, the division between qualitative and quantitative methodologies

has been an unfortunate biproduct – or causality – of these "fights." Many machine learning techniques offer alternative abductive approaches to conducting quantitative research, something which anthropologists should not only utilize but also help develop in ethnographies.

Appendix A: Python Code

The following records the Python code for this project, run in Jupyter notebook.

Decision Tree Modeling

The model uses decision trees to classify our study participants* according to their cyber status (cybersensitive, cyberaware, mainstream, low mainstream, and null) based on the survey and interview data. The goal of the model is to classify/predict what cyber status of an individual.

Data Cleaning

In [1]:
```
# Imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cross_validation import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
from sklearn import model_selection
from scipy import signal
```

In [2]:
```
# Loading the datasets from their respective files
filepath = 'C:\\Users\\swphi\\Documents\\Indicia\\Epic\\Task 4\\'
file1 = pd.ExcelFile(filepath + 'Master EPIC dataset - copy.xlsx')
file2 = pd.ExcelFile(filepath + 'all survey responses - copy.xlsx')

cyber_status = file1.parse('Cyber Status')
psych_codes = file1.parse('Psych')
energy_codes = file1.parse('Energy')
device_codes = file1.parse('Device')[
surveys = file2.parse('Interview Responses')
key = file2.parse('Question Key')
```

In [3]:
```
# Merging all the datasets into one data for the model
data = cyber_status.merge(psych_codes, left_on = 'Name', right_index =
True)
data = cyber_status.merge(energy_codes, left_on = 'Name', right_index =
True)
data = cyber_status.merge(device_codes, left_on = 'Name', right_index =
True)
data = cyber_status.merge(surveys, left_on = 'Name', right_on = 'Name')
```

```
In [4]:
# Converting non-numerical data into numbers so that it will run
through the decision tree packages

# Converts it to a quantitative variable based on the given key
def quantize(x, key):
    for k in key:
        if x == key[k]:
            return k
    return x

'''
Key for Variable Edits:
rent_or_own_quant: 0 if the user rents and 1 if the user owns
gender_quant: 0 if female and 1 if male
age_quant: Goes to smallest age in the range
household_income_quant: Goes to the smallest income and -1 if prefer
not to answer
community_quant: 0 if rural, 1 if suburban, 2 if urban
instructions_quant: 0 if wait for someone else, 1 if use instructions,
2 if figure out on your own
region_quant: 0 if southern california, 1 if northern california
smartphone_usage_quant: 0 if never, 1 if a few times a month, 2 if a
few times a week, 3 if once a day
nonsmartphone_usage_quant: 0 if never, 1 if a few times a month, 2 if a
few times a week, 3 if once a day
gaming_console_usage_quant: 0 if never, 1 if a few times a month, 2 if
a few times a week, 3 if once a day
laptop_usage_quant:0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
ipod_usage_quant: 0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
fitness_tracker_usage_quant: 0 if never, 1 if a few times a month, 2 if
a few times a week, 3 if once a day
health_fitness_app_usage_quant: 0 if never, 1 if a few times a month, 2
if a few times a week, 3 if once a day
home_automation_system_usage_quant: 0 if never, 1 if a few times a
month, 2 if a few times a week, 3 if once a day
home_security_usage_quant: 0 if never, 1 if a few times a month, 2 if a
few times a week, 3 if once a day
energyConsumption_app_usage_quant: 0 if never, 1 if a few times a
month, 2 if a few times a week, 3 if once a day
money_management_app_usage_quant: 0 if never, 1 if a few times a month,
2 if a few times a week, 3 if once a day
```

```
tablet_usage_quant: 0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
music_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
gaming_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a
few times a week, 3 if once a day
data_plan_type_quant: 0 if they do not have a plan, 1 if 1-5 GB plan, 2
if more than 5 GB, 3 if unlimited, -1 if unsure what plan they have
energy_awareness_quant: 0 if not aware, 1 if generally aware, 2 if
always aware
social_adoption_quant: 0 if one of the last to adopt any new
technology, 1 if wait until somewhat widely adopted, 2 if one of the
first ones to adopt
'''

quant_keys = {'Rent or Own': ['rent_or_own_quant', 'Do you currently
rent or own your primary residence?', {0: 'Rent', 1: 'Own'}],
              'Gender': ['gender_quant', 'Please indicate your gender
below:', {0: 'Female', 1: 'Male'}],
              'Age': ['age_quant', 'Please indicate your age below:',
{25: '25-34', 35: '35-44', 45: '45-54', 55: '55-64', 65: '65-74'}],
              'Household Income': ['household_income_quant', 'Which of
the following ranges best indicates your annual household income?',
{2000: '$20,000 to $49,999', 50000: '$50,000 to $99,999', 100000:
'$100,000 to $149,999', 150000: '$150,000 to $199,999', 200000:
'$200,000 or more', -1: 'Prefer not to answer'}],
              'Community': ['community_quant', 'How would you describe
the type of community you reside in?', {0: 'Rural community', 1:
'Suburban community', 2: 'City or Urban community'}],
              'Instructions': ['instructions_quant', 'Which of the
following statements best describes you?', {0: 'When I get a new
device, I usually wait for someone else to help me figure out how to
use it', 1: 'When I get a new device, I jump right into the
instructions and learn how to use it in detail', 2: 'When I get a new
device, I figure out how to use it on my own, and look at the
instructions only if I get stuck'}],
              'Region': ['region_quant', 'Region', {0: 'Southern
California', 1: 'Northern California'}],
              'Smartphone Usage': ['smartphone_usage_quant', 'Mobile
phone with internet capability:How often, if ever, do you use or access
the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A
few times a week', 3: 'At least once a day'}],
              'Non-Smartphone Usage': ['nonsmartphone_usage_quant',
'Mobile phone without internet capability:How often, if ever, do you
```

```
use or access the following?', {0: 'Never', 1: 'A few times a month or
less', 2: 'A few times a week', 3: 'At least once a day'}],
              'Gaming Console Usage': ['gaming_console_usage_quant',
'Gaming console (such as Playstation, X-box, etc.):How often, if ever,
do you use or access the following?', {0: 'Never', 1: 'A few times a
month or less', 2: 'A few times a week', 3: 'At least once a day'}],
              'Laptop Usage': ['laptop_usage_quant', 'Laptop:How often,
if ever, do you use or access the following?', {0: 'Never', 1: 'A few
times a month or less', 2: 'A few times a week', 3: 'At least once a
day'}],
              'ipod usage': ['ipod_usage_quant', 'Portable digital
music player (such as iPod, etc.):How often, if ever, do you use or
access the following?', {0: 'Never', 1: 'A few times a month or less',
2: 'A few times a week', 3: 'At least once a day'}],
              'Fitness Tracker Usage': ['fitness_tracker_usage_quant',
'Health / fitness tracker (such as Fitbit, Garmin watch, etc.):How
often, if ever, do you use or access the following?', {0: 'Never', 1:
'A few times a month or less', 2: 'A few times a week', 3: 'At least
once a day'}],
              'Health/Fitness App Usage':
['health_fitness_app_usage_quant', 'Health / fitness tracking apps on
your mobile phone (such as RunKeeper, Strava, etc.):How often, if ever,
do you use or access the following?', {0: 'Never', 1: 'A few times a
month or less', 2: 'A few times a week', 3: 'At least once a day'}],
              'Home Automation System Usage':
['home_automation_system_usage_quant', '', {0: 'Never', 1: 'A few times
a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
              'Home Security Usage': ['home_security_usage_quant',
'Home security or home monitoring system (such as ADT, web-enabled
surveillance devices, etc.):How often, if ever, do you use or access
the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A
few times a week', 3: 'At least once a day'}],
              'Energy Consumption App Usage':
['energy_consumption_app_usage_quant', 'MEnergy consumption tracking
devices, apps or services:How often, if ever, do you use or access the
following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few
times a week', 3: 'At least once a day'}],
              'Money Management App Usage':
['money_management_app_usage_quant', 'Money management applications
(such as Mint, Quicken, etc.):How often, if ever, do you use or access
the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A
few times a week', 3: 'At least once a day'}],
              'Tablet Usage': ['tablet_usage_quant', 'Tablet (such as
iPad, etc.):How often, if ever, do you use or access the following?',
```

```
{0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week',
3: 'At least once a day'}],
              'Music App Usage': ['music_app_usage_quant', 'Music apps
on my phone:How often, if ever, do you use or access the following?',
{0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week',
3: 'At least once a day'}],
              'Gaming App Usage': ['gaming_app_usage_quant', 'Gaming
apps on my phone:How often, if ever, do you use or access the
following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few
times a week', 3: 'At least once a day'}],
              'Data Plan Type': ['data_plan_type_quant', 'What type of
the data plan do you have on your mobile phone?', {-1: 'I have a data
plan, but not sure about the type', 0: 'I do not have a data plan', 1:
'1-5 GB per month', 2: 'More than 5 GB per month', 3: 'Unlimited data
plan'}],
              'Energy Awareness': ['energy_awareness_quant', 'Consider
the level of energy consumption in your household. Which of the
following statements would you agree with the most?', {0: 'I am not
aware of the level of energy consumption in my household. I generally
do not participate in making changes around the house to reduce our
energy consumption.', 1: 'I am generally aware of some aspects of
energy consumption in my household. I do not monitor all aspects of
energy usage in great detail, but participate in making changes to our
energy consumption whenever it is convenient.', 2: 'I am fully aware
of, and monitor the level of energy consumption in my household. I have
made and continue to make many changes wherever possible to our energy
usage, and lead the charge in this aspect in my household.'}],
              'Social Adoption': ['social_adoption_quant', 'Which of
the following statements best describes you?', {0: 'I am usually the
last one to adopt any new technology', 1: 'I wait for new technologies
to be somewhat widely adopted before adopting them myself', 2: 'I am
usually among the first ones to buy the latest electronic devices'}]
              }

for q in quant_keys:
    current = quant_keys[q]
    data[current[0]] = data[q].apply(quantize, args = (current[2],))
    data[current[0]] = data[current[0]].fillna(0)
    key = key.append(pd.Series(current, index = key.columns),
ignore_index = True)
```

In [5]:

```
# Develops the X and y matrices for model development
```

```
dep_vars = ['region_quant', 'Smartphone', 'Non-Smartphone', 'Gaming
Console ',
            'Laptop', 'iPod', 'Fitness Tracker',
           'Home Automation System', 'Health/Fitness App', 'Home
Securtiy',
           'Energy Consumption App', 'Money Management App', 'Tablet ',
           'Music Apps', 'Gaming Apps ', 'smartphone_usage_quant',
           'nonsmartphone_usage_quant', 'gaming_console_usage_quant',
'laptop_usage_quant',
           'ipod_usage_quant', 'fitness_tracker_usage_quant',
'home_automation_system_usage_quant',
           'health_fitness_app_usage_quant',
'home_security_usage_quant',
           'energy_consumption_app_usage_quant',
'money_management_app_usage_quant',
           'tablet_usage_quant', 'music_app_usage_quant',
'gaming_app_usage_quant', 'data_plan_type_quant',
           'Relieved', 'Happy', 'Anxious', 'Unaffected', 'Concerned',
           'Bored', 'Free', 'Excited', 'Crippled', 'Stressed',
'Unproductive', 'Frustrated',
           'Disconnected', 'Adoption', 'Breadth', 'Explore',
           'Advice', 'Bills', 'Energy Saving', 'Fun', 'Easy',
'Technology News',
           'social_adoption_quant', 'instructions_quant',
'gender_quant', 'age_quant',
           'household_income_quant', 'Household', 'Minors',
'community_quant', 'rent_or_own_quant']

X = data[dep_vars]
y_qual = data['Cyber Status']

cyber_key = {0: 'Null', 1: 'Low Mainstream', 2: 'Mainstream', 3:
'Cyberaware', 4: 'Cybersensitive'}
y = y_qual.apply(quantize, args = (cyber_key,))
y_quant = y
```

Resampling

In this section, we use bootstrap resampling to create a larger version of the sample data with a size of 300. Decision tree models vary impulsively on smaller datasets, and this will help address this. Resampling allows one to create a distribution, which is larger (or sometimes smaller) than the original, but which still represents the overall distribution.

In [6]:

```
X_resample = signal.resample(X, 300)
X_resample = pd.DataFrame(X_resample, columns = X.columns)
X_resample = X_resample.round()
y_resample = signal.resample(y_quant, 300)
y_resample = y_resample.round()
y_resample = pd.DataFrame(y_resample, columns = ['Cyber Status'])
```

```
In [7]:
# Develops a qualitative version the y_resample for models that require
a qualitative version
def qualitize(x, key, minimum, maximum):
    if x < minimum:
        return key[minimum]
    if x > maximum:
        return key[maximum]
    return key[x]

y_resample_qual = pd.DataFrame(columns = ['Cyber Status'])
y_resample_qual['Cyber Status'] = y_resample['Cyber
Status'].apply(qualitize, args = (cyber_key, 0, 4,))
```

I then tested the means and standard deviations of each column to ensure that the resample is still representative of the original data. They are pretty much the same with only slight variation several decimal places in (the one exception to this is income, which because it contains 5- to 6-digit values means that its slight variation is proportionally larger. It is still a slight variation). This indicates that the resampling still matches the sample's distribution.

```
In [8]:
test = pd.DataFrame(columns = ['Original Mean', 'Resampled Mean',
'Original Std', 'Resampled Std'])
test['Original Mean'] = X.mean()
test['Resampled Mean'] = X_resample.mean()
test['Original Std'] = X.std()
test['Resampled Std'] = X_resample.std()
test['Mean Difference'] = abs(test['Original Mean'] - test['Resampled
Mean'])
test['Std Difference'] = abs(test['Original Std'] - test['Resampled
Std'])
display(test)
```

Decision Tree Development

We will first determine the most accurate decision tree model to construct. In particular, we will determine the best type of splitting function - Gini index or entropy - and what are maximum depth is. Below are the accuracies when conducting Leave-One-Out Cross-Validation (LOOCV) for each type of function type for each depth.

Two conclusions from the table:

1) Gini index are consistently more accurate than entropy.

2) The accuracies start to level off at a maximum depth of 7, indicating that 7 is the best maximum depth

Thus, we will use the Gini index with a maximum depth of 7.

```
In [10]:
loocv = model_selection.LeaveOneOut()
accuracy = pd.DataFrame( columns = ['Maximum Depth', 'Gini Index
Accuracy', 'Entropy Accuracy'])

for depth in range(1, 34):
    score = [depth]

    clf_gini = tree.DecisionTreeClassifier(criterion = "gini",
random_state = 100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_gini, X_resample,
y_resample_qual, cv=loocv)
    score.append(results.mean())

    clf_entropy = tree.DecisionTreeClassifier(criterion = "entropy",
random_state = 100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_entropy, X_resample,
y_resample_qual, cv=loocv)
    score.append(results.mean())

    accuracy.loc[len(accuracy)] = score

accuracy
```

```
Out[14]:
```

	Maximum Depth	Gini Index Accuracy	Entropy Accuracy
0	1.0	0.253333	0.333333
1	2.0	0.313333	0.353333
2	3.0	0.336667	0.436667
3	4.0	0.453333	0.476667
4	5.0	0.626667	0.570000
5	6.0	0.686667	0.596667
6	7.0	0.696667	0.666667
7	8.0	0.760000	0.720000
8	9.0	0.753333	0.726667
9	10.0	0.743333	0.740000
10	11.0	0.763333	0.746667
11	12.0	0.766667	0.746667

```
12    13.0  0.763333    0.746667
13    14.0  0.763333    0.746667
14    15.0  0.763333    0.746667
15    16.0  0.763333    0.746667
16    17.0  0.763333    0.746667
17    18.0  0.763333    0.746667
18    19.0  0.763333    0.746667
19    20.0  0.763333    0.746667
20    21.0  0.763333    0.746667
21    22.0  0.763333    0.746667
22    23.0  0.763333    0.746667
23    24.0  0.763333    0.746667
24    25.0  0.763333    0.746667
25    26.0  0.763333    0.746667
26    27.0  0.763333    0.746667
27    28.0  0.763333    0.746667
28    29.0  0.763333    0.746667
29    30.0  0.763333    0.746667
30    31.0  0.763333    0.746667
31    32.0  0.763333    0.746667
32    33.0  0.763333    0.746667
In [11]:
accuracy.to_csv('Decision Tree LOOCV Results v2.csv')
```

Build initial decision tree.

Here is the primary decision tree. I use the gini index function and has a maximum depth of 8. It's LOOCV accuracy score is 76%.

```
In [12]:
max_depth = 8
dtree = tree.DecisionTreeClassifier(criterion = "gini", random_state
= 100, max_depth=max_depth, min_samples_leaf=1)
dtree.fit(X_resample, y_resample)
tree.export_graphviz(dtree, out_file='tree.dot',
feature_names=X.columns, filled=True, rounded=True )
In [13]:
results = model_selection.cross_val_score(dtree, X_resample,
y_resample_qual, cv=loocv)
print('LOOCV Accuracy Score: ' + str(100*results.mean()) + '%.')
LOOCV Accuracy Score: 76.0%.
```

Random Forests

Here I build a random forest with the resampled set. I then tested it on both the resampled set and original sample. For both, it has an accuracy of 100%. (Accuracy is measured by the percentage of individuals the model can accurately predict.)

In [14]:
```
rf = RandomForestClassifier(n_estimators = 1000)
rf.fit(X_resample, y_resample)
```

Out[14]:
```
RandomForestClassifier(bootstrap=True, class_weight=None,
criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

In [15]:
```
pred = rf.predict(X_resample)
correct = 0
for x in range(len(X_resample)):
    if pred[x] == y_resample['Cyber Status'][x]:
        correct += 1
score = correct/len(X_resample)
print('Accuracy on resampled set: ' + str(100*score) +'%')
```

```
Accuracy on resampled set: 100.0%
```

In [16]:
```
pred = rf.predict(X)
correct = 0
for x in range(len(X)):
    if pred[x] == y[x]:
        correct += 1
score = correct/len(X)
print('Accuracy on original sample: ' + str(100*score) +'%')
```

```
Accuracy on original sample: 100.0%
```

Plot the model.

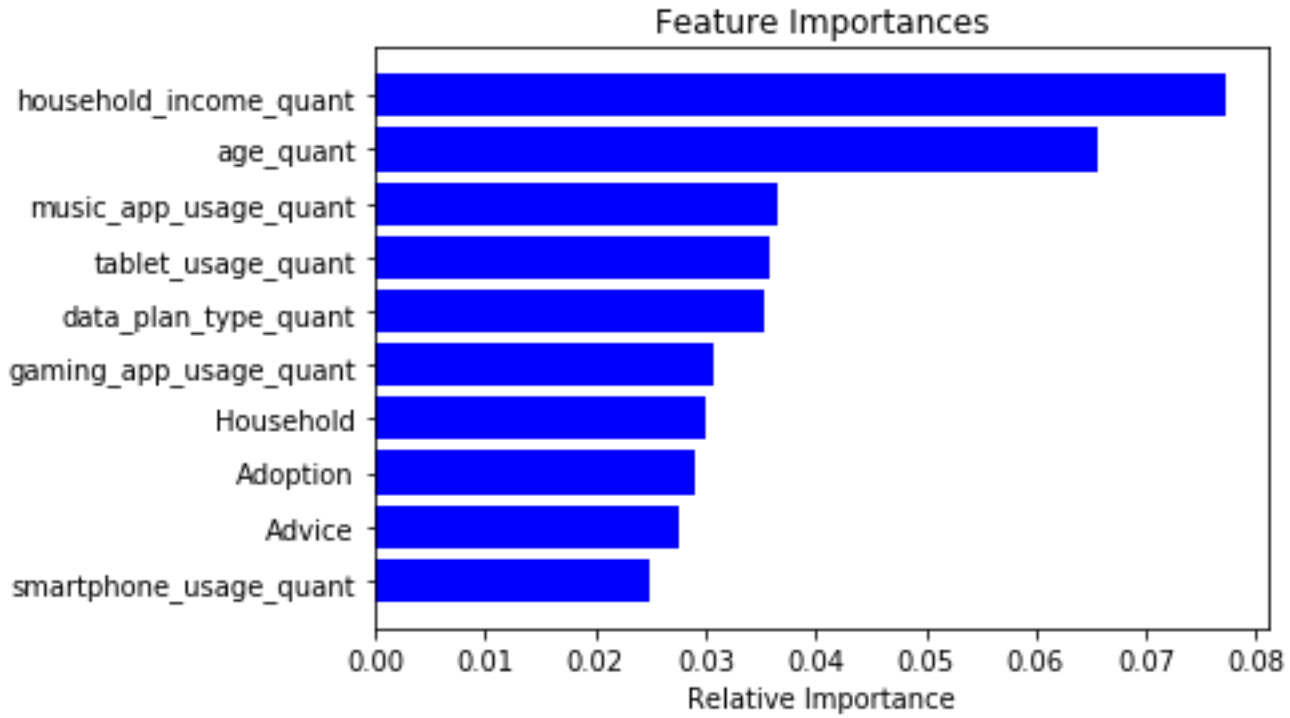
The graph below shows the 10 most important variables, which have been most significant variables when trying to classify each individual among all the trees. A larger, full table with the relative significance of all the variables is provided below that. (I made it latter table for your records; it is too much information for a final report, especially since most of it is extraneous. The graph provides the most important variables to discuss.)

In [17]:
```
# The code for this graphing approach is based on a similar model from
http://www.agcross.com/2015/02/random-forests-scikit-learn/
features = X.columns
```

```
importances = rf.feature_importances_
indices = np.argsort(importances)
l = len(indices)
top_indices = indices[l-10:l]
plt.figure(1)
plt.title('Feature Importances')
plt.barh(range(len(top_indices)), importances[top_indices], color='b',
align='center')
plt.yticks(range(len(top_indices)), features[top_indices])
plt.xlabel('Relative Importance')
```

Out[17]:

```
Text(0.5,0,'Relative Importance')
```



In [18]:

```
def column_name(column_number, df):
    return df.columns[column_number]

def get_importance(index, list_of_importances):
    return list_of_importances[index]

def print_full(x):
    pd.set_option('display.max_rows', len(x))
    print(x)
    pd.reset_option('display.max_rows')

importances_ranking = pd.DataFrame({'Column Number': indices})
importances_ranking['Column Name'] = importances_ranking['Column
Number'].apply(column_name, args = (X,))
```

```
importances_ranking['Relative Importance'] =
importances_ranking['Column Number'].apply(get_importance, args =
(importances,))
print_full(importances_ranking.sort_values('Relative Importance',
ascending = False))
```

Out[18]:

	Column Number	Column Name	Relative Importance
60	56	household_income_quant	0.073520
59	55	age_quant	0.067818
58	27	music_app_usage_quant	0.036126
57	29	data_plan_type_quant	0.034547
56	26	tablet_usage_quant	0.033653
55	28	gaming_app_usage_quant	0.030865
54	57	Household	0.030570
53	43	Adoption	0.030422
52	46	Advice	0.027976
51	15	smartphone_usage_quant	0.025409
50	18	laptop_usage_quant	0.025143
49	44	Breadth	0.022487
48	19	ipod_usage_quant	0.022474
47	53	instructions_quant	0.021726
46	25	money_management_app_usage_quant	0.021551
45	58	Minors	0.021307
44	51	Technology News	0.021102
43	45	Explore	0.020535
42	60	rent_or_own_quant	0.019648
41	21	home_automation_system_usage_quant	0.019525
40	49	Fun	0.019162
39	22	health_fitness_app_usage_quant	0.018548
38	50	Easy	0.018303

37	47	Bills	0.017428
36	13	Music Apps	0.016551
35	48	Energy Saving	0.015454
34	14	Gaming Apps	0.014398
33	17	gaming_console_usage_quant	0.013776
32	59	community_quant	0.013560
31	11	Money Management App	0.013262
30	20	fitness_tracker_usage_quant	0.013014
29	5	iPod	0.012586
28	52	social_adoption_quant	0.010916
27	34	Concerned	0.010694
26	42	Disconnected	0.010602
25	36	Free	0.010070
24	3	Gaming Console	0.010000
23	33	Unaffected	0.009874
22	24	energy_consumption_app_usage_quant	0.009765
21	0	region_quant	0.009690
20	8	Health/Fitness App	0.009564
19	54	gender_quant	0.009377
18	23	home_security_usage_quant	0.009216
17	41	Frustrated	0.009004
16	32	Anxious	0.008769
15	30	Relieved	0.008455
14	12	Tablet	0.008248
13	7	Home Automation System	0.007824
12	40	Unproductive	0.006298
11	6	Fitness Tracker	0.005899

10	1	Smartphone	0.005579
9	10	Energy Consumption App	0.005430
8	31	Happy	0.005138
7	9	Home Securtiy	0.005135
6	38	Crippled	0.004793
5	4	Laptop	0.003623
4	39	Stressed	0.003595
3	16	nonsmartphone_usage_quant	0.003097
2	37	Excited	0.002992
1	2	Non-Smartphone	0.002963
0	35	Bored	0.000944

Appendix B: Work Cited

Agre, Philip. 1997. *Computatoin and Human Experience.* Cambridge: Cambridge University Press.

Bauer, Mark, and Anne Wright. 1996. "Integrating Qualitative and Quantitative Methods to Model Infant Feeding Behavior among Navajo Mothers." *Human Organization* 55 (2).

Bell, Andrew, Jennifer Zavaleta Cheek, and Frazer Mataya. 2018. "Do As They Did: Peer Effects Explain Adoption of Conservation Agriculture in Malawi." *Water* 10 (1): 51.

boyd, danah, and Kati Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* 662-679.

Bucher, Taina. 2016. "Neither Black Nor Box: Ways of Knowing Algorithms." In *Innovative Methods in Media and Communication Research,* by Sebastian Kubitschko and Anne Kaun, 81-98. Cham: Palgrave Macmillan.

Chandler, David. 2015. "A World without Causation: Big Data and the Coming of Age of Posthumanism." *Millennium: Journal of International Studies.*

Cheney-Lippold, John. 2017. *We Are Data: Algorithms and the Making of Our Digitalized Selves.* New York: New York University Press.

Countee, Astrid. 2016. "An Engineering Anthropologist: Why tech companies need to hire software developers with ethnographic skills." *Ethnography Matters.* https://anthrocode.com/2016/07/31/an-engineering-anthropologist-why-tech-companies-need-to-hire-software-developers-with-ethnographic-skills/.

Crawford, Kate. 2014. "Big Data Anxieties: From Squeaky Dolphin to Normcore." *Epic.*

—. 2017. *The Trouble with Bias.* 12 10. https://www.youtube.com/watch?v=fMym_BKWQzk.

Curran, John. 2013. "Big Data or 'Big Ethnographic Data'? Positioning Big Data within the Ethnographic Space." *EPIC.*

Edirisingha, Prabash. 2016. *Ethnography, lived experience and consumer research.* August 23. https://prabash78.wordpress.com/2016/08/23/ethnography-lived-experience-and-consumer-research/.

EPIC. 2018. *Indicia Consulting.* https://www.epicpeople.org/business-directory/4430/indicia-consulting/.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Gillespie, Tarleton, and Nick Seaver. 2016. *Critical Algorithm Studies: a Reading List.* 12 15. https://socialmediacollective.org/reading-lists/critical-algorithm-studies/.

Gladwin, Christina. 1997. *Ethnographic Decision Tree Modeling.* Newbury, CA: Sage.

Gladwin, Christina, Hugh Gladwin, and Walter Peacock. 2001. "Modeling Hurrican Evacuation Decisions with Ethnographic Methods." *International Journal of Mass Emergencies and Disasters* 19 (2): 117-143.

Gladwin, Hugh, and Michael Murtaugh. 1984. "Test of a Hierarchical Model of Auto Choice on Data from the National Transportation Survey." *Human Organization* 43 (2): 217-226.

Harris, Marvin. 1974. "Why a Perfect Knowledge of All the Rules One Must Know to Act like a Native Cannot Lead to the Knowledge of How Natives Act." *Journal of Anthropological Research* 30 (4): 242-251.

Indicia Consulting. 2018. *Engaging Cybersensitives and Cyberawares in Energy Efficiency Part 1.* Epic Project Task 6, EPIC PON 14-306.

Indicia Consulting. 2014. *Epic Project Fact Sheet.* Epic Project Report, Indicia Consulting.

Indicia Consulting. 2018. *Scope of Work.* Epic Project Report, Indicia Consulting.

Ingold, Tim. 2007. *Lines: A Bried History.* London: Routledge.

Johnson, Burke, Anthony Onwuegbuzie, and Lisa Turner. 2007. "Toward a Definition of Mixed Methods Research." *Journal of Mixed Methods Research* 112-133.

Kelkar, Shreeharsh. 2014. *On the Porous Boundaries of Computer Science.* June 18. http://blog.castac.org/2014/06/on-the-porous-boundaries-of-computer-science/.

Knox, Hannah. 2018. "Baseless Data? Modelling, ethnography and the challenge of the anthropocene." In *Ethnography for a Data-saturated World*, by Dawn Nafus and Hannah Knox, 128-150. Manchester: Manchester University Press.

Kotz, Samuel, and Norman Johnson. 2012. Breakthroughs in Statistics: Foundations and Basic Theory. New York: Springer Science & Business Media.

Lowrie, Ian. 2018. "Becoming a real data scientist: expertise, flexibility, and lifelong learning." In *Ethnography for a Data-Saturated World*, by Dawn Nafus and Hannah Knox, 62-81. Manchester: Manchester University Press.

Mackenzie, Adrian. 2017. *Machine Learners: Archaeology of a Data Practice.* Cambridge: The MIT Press.

Mackenzie, Adrian. 2012. "More parts than elements: how databases multiply." *Society and Space.*

Mackenzie, Adrian, and Ruth McNally. 2013. "Living Multiples: How Large-scale Scientific Data-mining Pursues Identity and Differences." *Theory, Culture & Society.*

Madsen, Matte My, Anders Blok, and Morten Axel Pedersen. 2018. "Transversal collaboration: an ethnographic in/of computational social science." In *Ethnography for a Data-saturated World*, by Dawn Nafus. Manchester: Manchester University Press.

Mathews, Holly. 1987. "Predicting Decision Outcomes: Have We Put the Cart before the Horse in Anthropological Studies of Decision Making." *Human Organization* 46 (1): 54-61.

Mitchell, Tom. 1997. *Machine Learning.* Redmond: McGraw-Hill.

Mukhopadhyay, Carol. 1984. "Testing a Decision Process Model of the Sexual Division of Labor in the Family." *Human Organization* 43 (3): 227-242.

Murtaugh, Michael. 1984. "A Model of Grocery Shopping Decision Process Based on Verbal Protocol Data." *Human Organization* 43 (3): 243-251.

Nafus, Dawn, and Hannah Knox. 2018. *Ethnography for a Data-Saturated World.* Manchester: Manchester Univeristy Press.

Neville, Padraic G. 1994. *Decision Trees for Predictive Modeling.* SAS Institute, 1-24.

Noble, Safiya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.

Norvaisas, Julie, and Jonathan Karpfen. 2014. "Little Data, Big Data and Design at LinkedIn." *EPIC*.

Paff, Stephen. 2018. *The Anthropology of Machine Learning.* Memphis: University of Memphis Anthropology Department.

Plattner, Stuart. 1984. "Economic Decisin Making of Marketplace Merchants: An Ethnographic Model." *Human Organization* 43 (3).

Roark, Kendall. 2018. "Participatory Big Data Ethics: Against AI Gaydar and Other Creepy Machines." *Society for Applied Anthropology*.

Roark, Lior, and Oded Z Maimon. 2015. *Data Mining with Decision Trees: Theory and Applications.* New Jersey: World Scientific.

Ryan, Gery W, and H Russell Bernard. 2006. "Testing an Ethnographic Decision Tree Model on a National Sample: Recycling Beverage Cans." *Human Organization* 65 (1): 103-114.

Seaver, Nick. 2017. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." *Big Data & Society* 1-12.

Seaver, Nick. 2015. "Bastard Algebra." In *Data, Now Bigger and Better*, by Tom Boellstorff and B Maurer, 27-46. Chicago: Prickly Paradigm Press.

Slobin, Adrian, and Todd Cherkasky. 2010. "Ethnography in the Age of Analytics." *EPIC.*

Strathern, Marlilyn. 2004. Commons and Borderlands: Working Papers on Interdisciplinarity, Accountability nd the Flow of Knowledge. Oxford: Sean Kingston Publishing.

Thomas, S, D Nafus, and J Sherman. 2018. "Algorithms as fetish: Faith and possibility in algorithmic work." *Big Data & Society* 1-11.

Timmermans, Stefan, and Iddo Tavory. 2012. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *American Sociological Association* 167-186.

Wang, Tricia. 2013. *Why Big Data Needs Thick Data.* 5 13. https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7.

Wheeler, Schaun. 2018. *Data is a stakeholder.* 2 14. https://towardsdatascience.com/data-is-a-stakeholder-31bfdb650af0.

Yu, Zhun, Benjamin C.M. Fung, and Hiroshi Yoshino. 2010. "A Decision Tree Method for Building Energy Demand Modeling." *Energy and Buildings* 42 (10): 1637-1646.

Ziewitz, Malte. 2016. "Governing Algorithms: Myth, Mess, and Methods." *Science, Technology, & Human Values* 3-16.

Appendix C: Curriculum Vitae

Stephen Paff

Education:

University of Memphis, *Memphis, TN* 2017-2019
Master's in Applied Anthropology

Springboard, *San Francisco CA* 2017
Program: Data Science Career Intensive
Data Science Certificate

Wheaton College, *Wheaton IL* 2009-2013
Majors: Anthropology and Math
Bachelor of Arts in Anthropology and Bachelor of Arts in Math
Human Needs and Global Resources Certificate in Development Studies
GPA: 3.88, Summa cum Laude
Awards/Honors:
Wheaton College Scholastic Honors Society
Lambda Alpha Honors Society
Pi Kappa Delta Honor Society
Dean's List all semesters

Publications:

- "Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency Program Design," co-authored as team member on EPIC Project with Indicia Consulting, Task 6, 2018
- "Cybersensitive Electricity Consumption Patterns," co-authored as team member on EPIC Project with Indicia Consulting, Task 5, 2018
- "Cybersensitive Response to Technology," co-authored as team member on EPIC Project with Indicia Consulting, Task 4, 2017
- "Market Segmentation," co-authored as team member on Avangrid Project with Indicia Consulting in consortium with Brand Cool, 2017
- "Religious Life among Students at Cornerstone Leadership Academy for Boys: A Multi-Faith Ugandan High School" in Wheaton College's Human Needs and Global Resources Journal, 2013

Conferences, Presentations and Presentation Awards:

- "The Anthropology of Machine Learning: Steps Moving Forward" at 2018 Epic Conference in Honolulu, Hawaii; University of Memphis SGA Travel Award recipient

- Awarded Heinz von Foerster Award by the American Society for Cybernetics for presentation "Machine Stories: Machine Learning as Computerized Narrative Design," Chicago 2018
- "Computerized Knowledge Production: Machine Learning Models as Social Actors" at the 2018 Society for Applied Anthropology Annual Conference, Philadelphia, Pennsylvania
- Awarded 1st place by Evangelical Missiological Society for presentation, "Missions and Money," Chicago 2012

Experiences:

Research Analyst, *Indicia Consulting,* Washington DC
2017 to Present[12]

- Develop data science and statistical systems to analyze results from ethnographic studies, developing descriptive and predictive models
- Conduct segmentation and cluster analysis of consumer electric usage and behavior
- Facilitate discussions between qualitative and quantitative social research to best understand consumer behavior
- Build statistical and machine learning models in Python/Jupyter

Data Analyst, *ServiceMaster,* Memphis, TN
2018 to Present

- Conduct long-term customer research
- Integrate statistical, ethnographic, and interview techniques to understand clients
- Help develop research techniques with other team members

UX Design Researcher, *ServiceMaster,* Memphis, TN
2017-2018

- Conducted ethnographic research to understand technology usage
- Ran usability testing on software interfaces to empathize with users
- Designed user-friendly software systems
- Developed user personas and journey maps

Math Teacher, *Ralph Ellison High School,* Auburn Gresham, Chicago, IL
2016

- Developed and taught statistics curriculum for both AP Statistics and Statistics to 90 high school seniors
- Taught common-core Algebra 2 content to about 100 high school juniors
- Created Calculus curriculum and integrated it into the school's math education program

[12] The *present* throughout the CV refers to the time of completion of my Masters in Anthropology at University of Memphis in May 2019.

Facilitator, *Museum of Science and Industry,* Chicago, IL
2015-2016

- Regularly spoke in front of 100+ guests on a variety of science topics
- Conducted hands-on tours and workshop with groups of 10-20 guests in areas of specific interest
- Facilitated guests' learning experiences by developing methods to stimulate and teach guests

Middle School Math Teacher, *Aspire Charter Academy,* Gary, IN
2014-2015

- Developed and taught curriculum for about one hundred sixth, seventh, and eighth grade accelerated math
- Developed data-driven policies through Excel for the Aspire Charter Math Department
- On average students exceeded expected mathematics growth by 150%, according to NWEA test scores
- Provided comprehensive, individualized support for academically at-risk students

Math Teacher and Ethnographer, *Cornerstone Leadership Academy* Nakasongola, Uganda
2012

- Taught advanced-level secondary mathematics and Java programming
- Conducted for-credit ethnographic research project: "Religious Life among Students at Cornerstone Leadership Academy for Boys: A Multi-Faith Ugandan High School"
- Integrated cross-culturally to rural Central Uganda through a seven-month service-learning internship

Research Assistant for Dr. Brian Howell, *Anthropology Department,* Wheaton College
2011-2013

- Mentored first-time students conducting ethnographic research projects
- Created promotional brochures to attract potential anthropology students
- Conducted secondary research and edited research articles and books for publication
- Created and managed data tables for anthropology classes in Excel

Lab Assistant, *Math Department,* Wheaton College, Wheaton, IL
2010-2013

- Conducted weekly Calculus study sessions to assist struggling students
- Proctored weekly quizzes and exams
- Graded Calculus 1 homework for Dr. Isihara

Instructor, *Computing Workshop,* Pittsburgh, PA
2005-2011

- Instructed autistic teenagers on various lessons, including mathematics and programming topics
- Led social skills and career-planning sessions
- Observed and wrote reports on learning capacities and best teaching methods for autistic individuals

Research Projects:
Ethnographic Religious Study of CLA, *Nakasongola, Uganda*
2012

- Conducted for-credit ethnographic research on religious differences at CLA
- Integrated participant-observation, dozens of interviews, and survey results into research analysis
- Published in Wheaton College's Human Needs and Global Resources journal: "Religious Life among Students at Cornerstone Leadership Academy for Boys: A Multi-Faith Ugandan High School"

Ekitangaala Ranch Grocery Shop Ethnography, *Nakasongola, Uganda*
2012

- Conducted a 1.5 month-long ethnographic for-credit field research project of a grocery store in Ekitangaala Ranch in Nakasongola, Uganda
- Developed proficiency in Luganda to communicate with non-English speaking grocery store customers

Theosophical Society Ethnographic Research Project, *Wheaton, IL*
2012

- Conducted a 2.5 month-long ethnographic field research project at Theosophical Society's National Headquarters in Wheaton, IL for Anthropology Field Research Methods course
- Observed daily work of Theosophical staff's daily work, interviewed 20 members, and participated with a religious discussion group
- Synthesized and presented findings for each project in a 20+ ethnographic paper and class presentation

Volunteer Experiences:

Chairman, *STEM Committee, Boy Scouts of America,* Chicago, IL
2015-2017

- Managed the development of STEM projects among Scouting youth
- Developed after-school robotics program for Chicagoan youth
- Organize four events per year on STEM topics for hundreds of youth in Chicago
- Lead over a team of a few dozen to implement STEM initiatives in Chicago's five districts

Tutor, *World Relief,* Wheaton, IL
2010 - 2012

- Assisted with homework for high school refugees
- Tutored ESL for a Nepalese refugee couple

Programming Languages and Software Expertise:

Java, Python, Jupyter, SQL, Visual Basic, Matlab, SPSS, Mathematica, Qualtrics, and Hotjar

Appendix D: Selected Readings

The Selected Readings highlights provides an in-depth bibliographic reference of relevant literature for the reader. It includes sources analysis of data science, machine learning, data, big data, and other related topics from anthropologists and a host of other scholars (ranging from social scientists to data scientists).

Selected Readings

Abe, Akinori and Moeko Hayashi. "On communication assistance via bots —towards IMDJ." *ScienceDirect* (2016): 1657-1665.

Acheson, James and Ann Acheson. "Offshore wind power development in Maine: A rational choice perspective." *Maine Policy Review* (2016): 42-55.

Adomavicius, Gediminas and Alexander Tuzhilin. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions." IEEE Transactions on Knowledge and Data Engineering (2005): 734-749.

Agre, Philip. *Computatoin and Human Experience*. Cambridge: Cambridge University Press, 19bucher97.

Anderson, R.J. and W.W. Sharrock. "Ethical Algorithms: A brief comment on an extensive muddle." (2013).

Anderson, Susan and Michael Anderson. *How Machines Can Advance Ethics*. 2009. <https://philosophynow.org/issues/72/How_Machines_Can_Advance_Ethics>.

Bail, Christopher. "The cultural environment: measuring culture with big data." *Social Theory* (2014).

Bailey, Arlene and Ojelanki Ngwenyama. "Toward Entrepreneurial Behavior in Underserved Communities: An Ethnographic Decision Tree Model of Telecenter Usage." *Information Technology for Development* (2013): 230-248.

Barros, Rodrigo, Andre de Carvalho and Alex Freitas. "Automatic Design of Decision-Tree Induction Algorithms." *Springer* (2015).

Bateson, Gregory. *Steps to an Ecology of Mind*. New York: Ballantine Books, 1972.

Bauer, Mark, and Anne Wright. "Integrating Qualitative and Quantitative Methods to Model Infant Feeding Behavior among Navajo Mothers." *Human Organization* 55, no. 2 (1996): 183-92.

Baumer, Eric, et al. "Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?11." *Journal of the Association for INformation Science and Technology* (2017): 1387-1410.

Bell, Andrew, Jennifer Zavaleta Cheek, Frazer Mataya, and Patrick Ward. "Do As They Did: Peer Effects Explain Adoption of Conservation Agriculture in Malawi." *Water* 10, no. 1 (2018): 51.

Bharwani, Sukaina. "Understanding Complex Behavior and Decision Making Using Ethnographic Knowledge Elicitation Tools (KnETs)." *Social Science Computer Review* (2006): 78-105.

Bishop, Christopher. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

Bleecker, Julian. "Why Things Matter: A Manifesto for Networked Objects — Cohabiting with Pigeons, Arphids and Aibos in the Internet of Things." *Creative Commons* (2005).

Blue, Alexis. *Digital Archaeology Project to Use Big Data*. 28 8 2017. <https://uanews.arizona.edu/story/digital-archaeology-project-use-big-data>.

Boellstorff, Tom and Bill Maurer. *Data, Now Bigger and Better!* Chicago: Prickly Paradigm Press, 2015.

boyd, danah and Kati Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* (2012): 662-679.

Brightman, Marc and Jerome Lewis. *The Anthropology of Sustainability: Beyond Developmen and Progress*. London: Palgrave MacMillan, 2017.

Brondo, Keri and Linda Bennett. "Career Subjectivities in U.S. Anthropology: Gender, Practice, and Resistance." *American Anthropologist* (2012): 598–610.

Bucher, Taina. "Neither Black Nor Box: Ways of Knowing Algorithms." Kubitschko, Sebastian and Anne Kaun. *Innovative Methods in Media and Communication Research*. Cham: Palgrave Macmillan, 2016. 81-98.

Burrell, Jenn. "How the machine ‘thinks’: Understanding opacity in machine learning algorithms." *Big Data & Society* (2016).

Burri, Thomas. "Machine Learning and the Law: Five Theses." (2017): 1-4.

Burrsettles. *Machine Learning and Social Science: Taking The Best of Both Worlds*. 5 2 2012. <https://slackprop.wordpress.com/2013/02/05/machine-learning-and-social-science/>.

Button, Graham. *Technology in Working Order: Studies of Work, Ineraction, and Technology*. New York: Routledge, 1993.

Caplan, Robyn and danah boyd. "Isomorphism through algorithms: Institutional dependencies in the case of Facebook." *Big Data & Society* (2018).

Cardon, Dominique. "Deconstructing the algorithm: four types of digital information calculations." Seyfert, Robert and Jonathan Roberge. *Algorithmic Cultures: Essays on Meaning, Performance, and New Technologies*. London: Routledge, 2016. 95-110.

Cates, Caleb, Lane Bruner and Joseph Moss. "Recuperating the Real: New Materialism, Object-Oriented Ontology, and Neo-Lacanian Ontical Cartography." *Philosophy & Rhetoric* (2018): 151-175.

CGP Grey. *How Machines *Really* Learn. [Footnote]*. 18 12 2017. <https://www.youtube.com/watch?v=wvWpdrfoEv0>.

—. *How Machines Learn*. 18 12 2017. <https://www.youtube.com/watch?v=R9OHn5ZF4Uo>.

Chandler, David. "A World without Causation: Big Data and the Coming of Age of Posthumanism." *Millennium: Journal of International Studies* (2015).

Chatfield, Tim. *AI will simplify talent acquisition*. 20 9 2017. <https://venturebeat.com/2017/09/20/ai-will-simplify-talent-acquisition/>.

Chen, Kai-Ying and Yeh Chih-Feng. "Factors affecting adoption of smart meters in the post-Fukushima era Taiwan: an extended protection motivation theory perspective." *Behaviour & Information Technology* (2017): 955-969.

Cheney-Lippold, John. *We Are Data: Algorithms and the Making of Our Digitalized Selves*. New York: New York University Press, 2017.

Clement, Maxime and Matthieu Guitton. "Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia." *Computers in Human Behavior* (2015): 66-75.

Countee, Astrid. "An Engineering Anthropologist: Why tech companies need to hire software developers with ethnographic skills." *Ethnography Matters* (2016). <https://anthrocode.com/2016/07/31/an-engineering-anthropologist-why-tech-companies-need-to-hire-software-developers-with-ethnographic-skills/>.

Crawford, Kate. "Big Data Anxieties: From Squeaky Dolphin to Normcore." *EPIC* (2014).

—. *The Trouble with Bias*. 10 12 2017. <https://www.youtube.com/watch?v=fMym_BKWQzk>.

Croll, Alistair. *Data Everywhere: Data Anthropology, Quantified Self, Machine Data, Human Centered Design, and more*. 4 2 2014. <http://www.oreilly.com/pub/e/2997>.

Cunningham, Sally Jo. "Machine learning applications in anthropology: automated discovery over kinship structures." *Computers and the Humanities* (1996).

Curran, John. "Big Data or 'Big Ethnographic Data'? Positioning Big Data within the Ethnographic Space." *EPIC* (2013).

Cyborg Anthropology. *What is Cyborg Anthropology?* 2011. <http://cyborganthropology.com/What_is_Cyborg_Anthropology%3F>.

Dahan, Haim, Shahar Cohen, Lior Rokach, and Oded Maimon. *Proactive Data Mining with Decision Trees*. New York, NY: Springer New York, 2014.

Davidai, Shai, Thomas Gilovich, and Lee D. Ross. "The Meaning of Default Options for Potential Organ Donors." Proceedings of the National Academy of Sciences 109, no. 38 (2012): 15201- 5205.

De Meur, Gisele. *New Trends in Mathematical Anthropology*. London: Routledge & Kegan Paul, 1986.

Deleuze, Gilles. "Postscript on the Societies of Control." *The MIT Press* (1992): 3-7.

Diakopoulos, Nicholas. *Sex, Violence, and Autocomplete Algorithms*. 2 8 2013. <http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html>.

D'Oca, Simona, Stefano Corgnati and Tiziana Buso. "Smart meters and energy savings in Italy: Determining the effectiveness of persuasive communication in dwellings." *Energy Research & Social Science* (2014): 131-142.

Domingos, Pedro. "A Few Useful Things to Know about Machine Learning." *ACM* (2012).

—. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, 2015.

—. "The Five Tribes of Machine Learning And What You Can Take from Each." *ACM* (2015).

Dougherty, Anne, Courtney Henderson, Amanda Dwelley, Mallika Jayaraman, Edward Vine, and Susan Mazur-Stommen. Energy Efficiency Behavioral Programs: Literature Review, Benchmarking Analysis, and Evaluation Guidelines. Report. Conservation Applied Research & Development (CARD) Final Report. Madison, WI: Illume Advising, 2015. 1-114. Prepared for Minnesota Department of Commerce, Division of Energy Resources.

Drackle, Dorle and Werner Krauss. "Ethnographies of Wind and Power." *Reviews in Anthropology* (2011): 313-330.

Driscoll, Kevin and Shawn Walker. "Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data." *International Journal of Communication* (2014): 1745-1764.

Ducheneaut, Nicolas, Nicholas Yee and Victoria Bellotti. "The Best of Both (Virtual) Worlds: Using Ethnography and Computational Tools to Study Online Behavior." *EPIC* (2010): 136-148.

Duda, Richard, Peter Hart and David Stork. *Pattern Classification*. Wiley-Interscience, 2000.

Edirisingha, Prabash. *Ethnography, lived experience and consumer research*. 23 August 2016. <https://prabash78.wordpress.com/2016/08/23/ethnography-lived-experience-and-consumer-research/>.

Edwards, Chad, et al. "Differences in perceptions of communication quality between a Twitterbot and human agent for information seeking and learning." *Computers in Human Behavior* (2016): 666-671.

—. "Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter." *Computers in Human Behavior* (2014): 372-376.

Elish, M.C. and danah boyd. "Situating methods in the magic of Big Data and AI." *Communication Monographs* (2018): 57-80.

Ehrhardt-Martinez, Karen, Kat A. Donnelly, John A. Laitner, Dan York, Jacob Talbot, and Katherine Friedrich. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Report no. E105. Washington, D.C.: American Council for an Energy-Efficient Economy, 2010. 1-128.

Eslami, Motahhare, et al. "First I “like” it, then I hide it: Folk Theories of Social Feeds." *Curation and Algorithms* (2016).

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press, 2018.

Evans, Bob. "Paco—Applying Computational Methods to Scale Qualitative." *EPIC* (2016): 348-368.

Faßler, Manfred. "Human-Computer-Inter-Creativity: A Co-Evolutionary Approach." (2013).

Feldman, Joseph. "Big data and ethnology." *Anthropology Today* (2017).

Fischer, Michael. *Anthropological Futures*. Durham: Duke University Press, 2009.

Flach, Peter. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press, 2012.

Foucault, Michel. *Discipline & Punish*. New York: Random House, Inc., 1977.

—. *Madness and Civilization*. New York: Random House, Inc, 1965.

—. *Power/Knowledge*. New York: Random House, Inc, 1977.

—. *The Final Foucault*. Cambridge: MIT Press, 1988.

Friedman, Uri. "Big Data: Anthropology of an Idea." *Foreign Policy* (2012).

Gandomi, Amir and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management* (2015).

Geng, Hwaiyu. *Internet of Things and Data Analytics Handbook*. Hoboken: John Wiley & Sons, 2016.

Geiger, Stuart and David Ribes. "Trace ethnography: Following coordination through documentary practices." (2011).

Geiger, Stuart. "Computational Ethnography and the Ethnography of Computation." 14 9 2017.

—. "Does Habermas Understand the Internet? The Algorithmic Construction of the Blogo/Public Sphere." *Gnovis: A Journal of Communication, Culture, and Technology* (2009).

—. "The Lives of Bots." Lovink, Geert and Nathaniel Tkacz. *A Critical Point of View: A Wikipedia Reader*. Amsterdam: Network Cultures, 2011. 78-93.

George, Stephen, Eric Bell, Aimee Savage, Alexandra Dunn, and Benjamin Messer. California Statewide Opt-in-Time-of-Use Pricing Pilot: Interim Evaluation. Report. Nexant and Research Into Action, 2017. 1-451. Prepared for The TOU Working Group, under contract to Southern California Edison Company.

Giaccardi, Elisa, Chris Speed and Neil Rubens. "Things Making Things: An Ethnography of the Impossible." (2014).

Giaccardi, Elisa, et al. "Thing Ethnography: Doing Design Research with Non-Humans." *DIS* (2016).

Gillespie, Tarleton and Nick Seaver. *Critical Algorithm Studies: a Reading List*. 15 12 2016. <https://socialmediacollective.org/reading-lists/critical-algorithm-studies/>.

Gladwin, Christina H., Hugh Gladwin, and Walter Gillis Peacock. “Modeling Hurricane Evacuation Decisions with Ethnographic Methods” International Journal of Mass Emergencies and Disasters 19, no. 2 (2001): 117-143

Gladwin, Christina H. *Ethnographic Decision Tree Modeling*. Newbury, CA: Sage, 1997.

Gladwin, Hugh, and Michael Murtaugh. "Test of a Hierarchical Model of Auto Choice on Data from the National Transportation Survey." *Human Organization* 43, no. 3 (1984): 217-226.

Gram-Hanssen, Kirsten. "“Home is where the smart is”? Evaluating smart home research and approaches against the concept of home." *Energy Research & Social Science* (2017): 94-101.

Gray, Mary. "Big Data, Ethical Futures." *Anthropology News* (2013).

Gray, Mary, et al. "The Crowd is a Collaborative Network." *Social and Behavioral Sciences: Sociology* (2016).

Gustafon, Steven. *The human-side of artificial intelligence and machine learning*. 20 6 2016. <http://ethnographymatters.net/blog/2016/06/20/the-human-side-of-artificial-intelligence-and-machine-learning/>.

Haanpaa, Leena. "Consumers' Green Commitment: Indication of a Postmodern Lifestyle?" *International Journal of Consumer Studies* (2007): 478-486.

Haines, Julia. "Towards Multi-Dimensional Ethnography." *EPIC* (2017).

Hale, Scott. "What is Data Science?" *SAGE Publications* (2017).

Harris, Marvin. "Why a Perfect Knowledge of All the Rules One Must Know to Act like a Native Cannot Lead to the Knowledge of How Natives Act." *Journal of Anthropological Research* 30, no. 4 (1974): 242-251.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

Hendricks, Bracken, Bill Campbell, and Pen Goodale. Efficiency Works: Creating Good Jobs and New Markets Through Energy Efficiency. Publication. Center for American Progress and Energy Resource Management, 2010. 1-48.

Henning, Annette. "Climate change and energy use: The role for anthropological research." *Anthropology Today* (2005).

Hill, Kashmi and Mattu Surya. *The House That Spied on Me*. 7 2 2018. <https://gizmodo.com/the-house-that-spied-on-me-1822429852?rev=1518027891546>.

Huang, Hsiu-Li, and Mei Chang Yeh. "Introduction to Ethnographic Decision Tree Modeling." *Journal of Nursing* 53, no. 3 (2006): 60-68.

IBM Corporation. (2012eavs). *Decision Tree Models*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm

Indicia Consulting. "Engaging Cybersensitives and Cyberawares in Energy Efficiency Part 1." EPIC Project Task 6. 2018.

—. "Engaging Cybersensitives and Cyberawares in Energy Efficiency Part 2: Recommendations." EPIC Task 6 Report. 2018.

—. "EPIC Project Fact Sheet." EPIC Project Report. 2014.

—. "Scope of Work." EPIC Project Report. 2018.

—. *Task 5 Background*. Washington D.C., 2018.

Ingold, Tim. *Lines: A Bried History*. London: Routledge, 2007.

Introna, Lucas. "The algorithmic choreography of the impressionable subject." Seyfert, Robert and Jonathan Roberge. *Algorithmic Cultures: Essays on Meaning Performance and New Technologies*. London: Routledge, 2016. 25-51.

Irani, Lilly. *Justince for "Data Janitors"*. 15 1 2015. <http://www.publicbooks.org/justice-for-data-janitors/>.

Isenhour, Cindy, Gary McDonogh and Melissa Checker. *Sustainability in the Global City: Myth and Practice*. Cambridge: Cambridge University Press, 2015.

Jasper, Agnes. "'I am not a goth!': The Unspoken Morale of Authenticity within the Dutch Gothic Subculture." *Ethnofoor* (2004): 90-115.

Johnson, Burke, Anthony Onwuegbuzie and Lisa Turner. "Toward a Definition of Mixed Methods Research." *Journal of Mixed Methods Research* (2007): 112-133.

Johnson, Jim. "Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer*." *Social Problems* (1988).

Kaghan, William and Geoffrey Bowker. "Out of machine age?: complexity, sociotechnical systems and actor network theory." *Journal of Engineering and Technology Development* (n.d.).

Keane, Webb. *Christian Moderns*. Berkeley: University of California Press, 2007.

Kelion, Leo. *London police trial gang violence 'predicting' software*. 29 October 2014. <http://www.bbc.com/news/technology-29824854>.

Kelkar, Shreeharsh. *On the Porous Boundaries of Computer Science*. 18 June 2014. <http://blog.castac.org/2014/06/on-the-porous-boundaries-of-computer-science/>.

Kelty, Christopher. "Two Fables." *Pulse* (2016): 490-514.

Khawaja, M. Sami, and James Stewart. Long-run Savings and Cost-effectiveness of Home Energy Report Program. Report. Waltham, MA: Cadmus Group, 2017. 1-20.

Kippen, James and Bernard Bel. "Can a computer help resolve the problem of ethnographic description?" *Anthropological Quarterly* (1989): 131-144.

Kippen, James. "On the Uses of Computers in Anthropological Research." *Current Anthropology* (1988): 317-320.

Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Infrastructures & Their Consequences*. Los Angeles: SAGE, 2014.

Klein, B., & Schlagenhauf, T. (2018). *What are Decision Trees*. Retrieved from Python Machine Learning Tutorial: https://www.python-course.eu/Decision_Trees.php

Kockelman, Paul. "The anthropology of an equation." *HAU: Journal of Ethnographic Theory* (2013).

Kononenko, Igor and Matjaz Kukar. *Machine Learning and Data Mining*. Elsevier: Philadelphia, 2007.

Kotz, Samuel and Norman Johnson. Breakthroughs in Statistics: Foundations and Basic Theory. New York: Springer Science & Business Media, 2012.

Kubat, Miroslav. *An Introduction to Machine Learning*. Springer Publishing Company: Cham, 2015.

Kubitschko, Sebastian and Anne Kaun. *Innovative Methods in Media and Communication Research*. Cham: Palgrave Macmillan, 2016.

Ladner, Sam. 7 12 2017. <www.samladner.com/why-machine-learning-isnt-about-machines/>.

Lalwani, Mona. *The next wave of AI is rooted in human culture and history*. 16 8 2016. <https://www.engadget.com/2016/08/16/the-next-wave-of-ai-is-rooted-in-human-culture-and-history>.

Lane, Justine. "Big Data and Anthropology." (2016).

Latour, Bruno. *Politics of Nature: How to Bring the Sciences into Democracy*. Boston: Harvard University Press, 2004.

—. *Reassembling the Social*. New York: Oxford University Press, 2005.

—. "The New Climate." *Harper's Magazine* (2017).

—. *We Have Never Been Modern*. Cambridge: Harvard University Press, 1993.

Latour, Bruno and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton: Princeton University Press, 1986.

Lee, Saro, and Chang-Wook Lee. "Application of Decision-Tree Model to Groundwater Productivity-Potential Mapping." *Sustainability* 7, no. 10 (2015): 13416-3432.

Lovins, Amory. “The Negawatt Revolution.” *The Conference Board Magazine*, 27 no. 9 (1990): 18-23.

Lowrie, Ian. "Becoming a real data scientist: expertise, flexibility, and lifelong learning." Nafus, Dawn and Hannah Knox. *Ethnography for a Data-Saturated World*. Manchester: Manchester University Press, 2018. 62-81.

Mackenzie, Adrian and Ruth McNally. "Living Multiples: How Large-scale Scientific Data-mining Pursues Identity and Differences." *Theory, Culture & Society* (2013).

Mackenzie, Adrian. *Machine Learners: Archaeology of a Data Practice*. Cambridge: The MIT Press, 2017.

—. "More parts than elements: how databases multiply." *Society and Space* (2012).

—. "Multiplying numbers differently: an epidemiology of contagious convolution." *Distinktion: Journal of Social Theory* (2014).

—. "Programming subjects in the regime of anticipation: Software studies and subjectivity." *Subjectivity* (2013).

—. "Set." Lury, Celia and Nina Wakeford. *Inventive Methods: The Happening of the Social*. London: Routledge, 2012. 219-231.

—. "The production of prediction: What does machine learning want?" *European Journal of Cutlural Studies* (2015).

—. "The Strange Meshing of Impersonal and Personal Forces in Technological Action." *Culture, theory & critique* (2006): 197-212.

Mackenzie, Donald. *Mechanizing proof: computing, risk, and trust*. Cambridge: MIT Press, 2001.

—. *A Sociology of Algorithms High-Frequency Trading and the Shaping of Markets*. Edinburgh: University of Edinburgh, 2014.

Madadipouya, Kasra. "A New Decision Tree Method for Data Mining in Medicine." *Advanced Computational Intelligence: An International Journal* 2, no. 3 (2015): 31-37.

Madsen, Matte My, Anders Blok and Morten Axel Pedersen. "Transversal collaboration: an ethnography in/of computational social science." Nafus, Dawn. *Ethnography for a Data-saturated World*. Manchester: Manchester Univeristy Press, 2018.

Marchese, Francis. "Tables and Early Information Visualization." (2013).

Mathews, Holly. "Predicting Decision Outcomes: Have We Put the Cart before the Horse in Anthropological Studies of Decision Making?" *Human Organization* 46, no. 1 (1987): 54-61.

Maurer, Bill. "Transacting ontologies: Kockelman's sieves and a Bayesian." *HAU: Journal of Ethnographic Theory* (2013).

McCarthy, Matthew. *Enacting the Semantic Web: Ontological Orderings, Negotiated Standards, and Human-machine Translations*. Milwaukee: University of Wisconsin-Milwaukee, 2017.

McGee, Kyle. *Bruno Latour: The Normativity of Networks*. New York City: Routledge, 2014.

McNally, Ruth and Adrian Mackenzie. "Understanding the 'Intensive' in 'Data Intensive Research': DataFlows in Next Generation Sequencing and EnvironmentalNetworked Sensors." *The International Journal of Digital Curation* (2012).

Mingers, Josh. "An empirical comparison of pruning methods for decision induction." Machine Learning (1989): 227-243.

Miller, Adam and Levi Bryant. *Speculative Grace: Bruno Latour and Object-Oriented Theology*. New York City: Forham University Press, 2013.

Mingers, John. "An Empirical Comparison of Selection Measures for Decision-tree Induction." *Machine Learning* 3, no. 4 (1989): 319-42.

MIT Technology Review. *Computational Anthropology Reveals How the Most Important People in History Vary by Culture*. 23 2 2015. <https://www.technologyreview.com/s/535356/computational-anthropology-reveals-how-the-most-important-people-in-history-vary-by/>.

Mitchell, Tom. "Key Ideas in Machine Learning." (2017).

—. *Machine Learning*. Redmond: McGraw-Hill, 1997.

Miyazaki, Shintaro. *ALGORHYTHMICS: UNDERSTANDING MICRO-TEMPORALITY IN COMPUTATIONAL CULTURES*. 28 9 2012. <http://computationalculture.net/algorhythmics-understanding-micro-temporality-in-computational-cultures/>.

Morita, Atsuro. "The Ethnographic Machine: Experimenting with Context and Comparison in Strathernian Ethnography." *Science, Technology, & Human Values* (2014): 214-235.

Morton, Timothy. *The Ecological Thought*. Boston: Harvard University Press, 2012.

Moss, Manny. *Crowdwork for Machine Learning: An Autoethnography*. 26 9 2017. <http://blog.fastforwardlabs.com/2017/09/26/crowdwork-for-ml.html>.

Mukhopadhyay, Carol. "Testing a Decision Process Model of the Sexual Division of Labor in the Family." *Human Organization* 43, no. 3 (1984): 227-242.

Murphy, Kevin. *Machine Learning: A Probabilitist Perpsepctive*. Cambridge: The MIT Press, 2012.

Murtaugh, Michael. "A Model of Grocery Shopping Decision Process Based on Verbal Protocol Data." *Human Organization* 43, no. 3 (1984): 243-251.

Mutzel, Sophie. "Facing Big Data: Making sociology relevant." *Big Data & Society* (2015): 1-4.

Nafus, Dawn and Hannah Knox. *Ethnography for a Data-Saturated World*. Manchester: Manchester Univeristy Press, 2018.

Nafus, Dawn and Tye Rattenbury. *Data Science and Ethnography: What's Our Common Ground, and Why Does It Matter?* 7 3 2018. <https://www.epicpeople.org/data-science

Nafus, Dawn. *Ethnography for a Data Saturated World*. Manchester: Manchester University Press, 2018.

Navega, David, Catarina Coelho, Ricardo Vicente, Maria Teresa Ferreira, Sofia Wasterlain, and Eugénia Cunha. "AncesTrees: Ancestry Estimation with Randomized Decision Trees." *International Journal of Legal Medicine* 129, no. 5 (2014): 1145-153.

Neville, Padraic G. *Decision Trees for Predictive Modeling*. Report. SAS Institute Inc. 1994. 1-24.

Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

Norvaisas, Julie and Jonathan Karpfen. "Little Data, Big Data and Design at LinkedIn." *EPIC* (2014).

Nyberg, Roy. "Using 'smartness' to reorganise sectors: Energy infrastructure and information engagement." *International Journal of Information Management* (2018).

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing, 2016.

Paff, Stephen. *The Anthropology of Machine Learning*. Memphis: University of Memphis Anthropology Department, 2018.

Patel, Neal. "For a Ruthless Criticism of Everything Existing: Rebellion Against the Quantitative-Qualitative Divide." *EPIC* (2013): 43-60.

Plattner, Stuart. "Economic Decision Making of Marketplace Merchants: An Ethnographic Model." *Human Organization* 43, no. 3 (1984): 252-264.

Physics arXiv. *When A Machine Learning Algorithm Studied Fine Art Paintings, It Saw Things Art Historians Had Never Noticed*. 14 8 2014. <https://medium.com/the-physics-arxiv-blog/when-a-machine-learning-algorithm-studied-fine-art-paintings-it-saw-things-art-historians-had-never-b8e4e7bf7d3e>.

Plessis, Elsabe and Robert Lorway. "What really works: Understanding the role of “local knowledges” in the monitoring and evaluation of a maternal, newborn and child health project in Kenya. Monitoring and Evaluation." Bell, Stephen and Peter Aggleton. *Health and Social Development: Interpretive and Ethnographic Perspectives*. New York: Routledge, 2016. 47-62.

Potter, Jennifer, Stephen S. George, and Lupe R. Jimenez. SmartPricing Options Final Evaluation: The Final Report on Pilot Design, Implementation, and Evaluation of the Sacramento Municipal Utility District's Consumer Behavior Study. Report. Sacramento, CA: Sacramento Municipal Utility District, 2014. 4-192. Prepared for U.S. Department of Energy.

Powell, Allison. "Data Walks and the Production of Radical Bottom-up Data Knowledge." *International Communications Association* (2017).

Rattenbury, Tye, Dawn Nafus and Ken Anderson. "Plastic: a metaphor for integrated technologies." *UbiComp* (2008).

Ribes, David and Geoffrey Bowker. "Between meaning and machine: Learning to represent the knowledge of communities." *Information and Organization* (2009).

Rieder, Bernhard. "Scrutinizing an algorithmic technique: the Bayes classifier as interested reading of reality." *Information, Communication & Society* (2017): 100-117.

Roark, Kendall. "Participatory Big Data Ethics: Against AI Gaydar and Other Creepy Machines." *Society for Applied Anthropology* (2018).

Rokach, Lior, and Oded Z. Maimon. *Data Mining with Decision Trees: Theory and Applications*. New Jersey: World Scientific, 2015.

Ryan, Gery W., and H. Russell Bernard. "Testing an Ethnographic Decision Tree Model on a National Sample: Recycling Beverage Cans." *Human Organization* 65, no. 1 (2006): 103-114.

Seaver, Nick. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." Big Data & Society (2017): 1-12.

—. "Bastard Algebra." Boellstorff, Tom and Bill Maurer. *Data, Now Bigger and Better*. Chicago: Prickly Paradigm Press, 2015. 27-46.

—. "The nice thing about context is that everyone has it." *Media, Culture & Society* (2015).

Selman, Bill. *Why Do We Conduct Qualitative User Research?* 30 10 2014. <https://blog.mozilla.org/ux/2014/10/why-do-we-conduct-qualitative-user-research/>.

Seyfert, Robert and Jonathan Roberge. *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies*. London: Routledge, 2016.

Shove, Elizabeth and Gordon Walker. "What Is Energy For? Social Practice and Energy Demand." *Theory, Culture, & Society* (2014).

Siemens, George. "Learning Analytics: The Emergence of a Discipline." *American Behavioral Scientist* (2013).

Simonite, Tom. *Should Data Scientists Adhere to a Hippocratic Oath?* 8 2 2018. <https://www.wired.com/story/should-data-scientists-adhere-to-a-hippocratic-oath/#ampshare=https://www.wired.com/story/should-data-scientists-adhere-to-a-hippocratic-oath>.

Sinders, Caroline. *Caroline Sinders on Ethical Product Design for Machine Learning*. 23 3 2017. <https://design.blog/2017/03/23/caroline-sinders-on-ethical-product-design-for-machine-learning/>.

—. *The Most Crucial Design Job Of The Future: What is a data ethnographer, and why is it poised to become so important?* 24 7 2017. <https://www.fastcodesign.com/90134155/the-most-crucial-design-job-of-the-future>.

Slobin, Adrian and Todd Cherkasky. "Ethnography in the Age of Analytics." *EPIC* (2010).

Solanki, Aakash and Sarvesh Tewari. "#GoingEthno in the Indian Bureaucracy." *EPIC* (2016).

Solomoff, R.J. "An Inductive Inference Machine." (1956).

State Energy Efficiency Resource Standards (EERS). Policy Brief. American Council for an Energy- Efficient Economy. 2012. 1-6.

State Energy Efficiency Resource Standards (EERS). Policy Brief. American Council for an Energy-

Efficient Economy. 2017. 1-9.

Steinberger, Julia and Maryline Sahakian. "Energy Reduction Through a Deeper Understanding of Household Consumption." Journal of Industrial Ecology (2011).

Strathern, Marlilyn. Commons and Borderlands: Working Papers on Interdisciplinarity, Accountability nd the Flow of Knowledge. Oxford: Sean Kingston Publishing, 2004.

Subbiah, Rajesh; Animitra, Pal; Eric K. Nordberg; Achla Marathe; and Madhav V. Marathe. "Energy Demand Model for Residential Sector: A First Principals Approach." *IEEE Transactions on Sustainable Energy* 8, no. 3 (2017): 1215-224.

Suchman, Lucy; Weber, Jutta. "Human-Machine Autonomies." Button, Graham. *Technology in working order: studies of work, interaction, and technology*. Cambride: Cambridge University Press, 2016. 75-102.

—. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge University Press, 2007.

Sussman, Reuven, and Maxine Chikumbo. Behavior Change Programs: Status and Impact. Report no. B1601. Washington, DC: American Council for an Energy-Efficient Economyc2016. 1-94.

Szinai, Julia, Merrian Borgeson, and Emily Levin. Putting Your Money Where Your Meter Is: A Study of Pay-for-performance Energy Efficiency Programs in the United States. Report. 1-62. Prepared for the Natural Resources Defense Council and Vermont Energy Investment Corporation.

Taylor, Alex. "Machine Intelligence." (2009).

Thacker, Eugene. *The Global Genome*. Cambridge: The MIT Press, 2005.

"The social life of Learning Analytics: cluster analysis and the performance of algorithmic education." *Learning, Media and Technology* (2016): 3-16.

Thomas, Matthew and Djuke VELDHUIS. *Learning to Trust Machines That Learn*. 11 10 2017. <https://www.sapiens.org/column/machinations/game-theory-anthropology/>.

Thomas, Suzanne, Dawn Nafus and Jamie Sherman. "Algorithms as fetish: Faith and possibility in algorithmic work." *Big Data & Society* (2018): 1-11.

Thorve, Swapna, Samarth Swarup, Achla Marathe, Young Yun Chun Baek, Eric K. Nordberg, and Madhav V. Marathe. "Simulating Residential Energy Demand in Urban and Rural Areas." Report. Department of Computer Science, Department of Agricultural and Applied Economics Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute, Virginia Tech. Blacksburg, VA, 2018.

Timmermans, Stefan and Iddo Tavory. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *American Sociological Association* (2012): 167-186.

Tso, Geoffrey, and Kelvin Yau. "Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks." *Energy* 32, no. 9 (2007): 1761-768.

Veale, Michael and Reuben Binns. "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data." *Big Data & Society* (2017).

Vertesi, Janet. *What robots in space teach us about teamwork: A deep dive into NASA*. 7 7 2016. <http://ethnographymatters.net/blog/category/editions/co-designing-with-machines/>.

Wallach, Hanna. *Machine Learning for Social Science*. 15 7 2016. <https://www.youtube.com/watch?v=oqfKz-PP9FU>.

Wang, Tricia. *Why Big Data Needs Thick Data*. 13 5 2013. <https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7>.

Wang, Xizhao, et al. *Machine Learning and Cybernetics: 13th International Conference, Lanzhou, China, July 13-16, 2014. Proceedings*. New York: Springer, 2014.

Wang, Yilun and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." *PsyArXiv* (2017).

Werner, Oswald. "Converting decision tables to flow charts with DTREDUCER." *Cultural Anthropology Methods Journal* (1997): 17-20.

Wheeler, Schaun. *Data is a stakeholder*. 14 2 2018. <https://towardsdatascience.com/data-is-a-stakeholder-31bfdb650af0>.

Wilf, Eitan. "Toward an Anthropology of Computer-Mediated, Algorithmic Forms of Sociality." *Current Anthropology* (2013): 716-739.

—. *Case Study Research: Core Skills in Using 15 Genres*. Emerald Group Publishing, 2016.

Woodside, Arch G. *Case Study Research: Core Skill Sets in Using 15 Genres*. Bingley, UK: Emerald, 2017.

York, Dan, "Positive on Negawatts: The Increasing Role of Energy Efficiency as a Utility Resource." Presentation. American Council for an Energy-Efficient Economy, 2008. Barros, Rodrigo C., De Carvalho, André C. P. L. F, and Alex A. Freitas. *Automatic Design of Decision-Tree Induction Algorithms*. Cham: Springer International Publishing, 2015.

Yu, Zhun, Fariborz Haghighat, Benjamin Fung, and Hiroshi Yoshino. "A Decision Tree Method for Building Energy Demand Modeling." *Energy and Buildings* 42, no. 10 (2010): 1637-1646.

Yu, Zhun, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. "A Systematic Procedure to Study the Influence of Occupant Behavior on Building Energy Consumption." *Energy and Buildings* 43, no. 6 (2011): 1409-1417.

Ziewitz, Malte. "A not quite random walk: Experimenting with the ethnomethods of the algorithm." *Big Data & Society* (2017).

—. "Governing Algorithms: Myth, Mess, and Methods." *Science, Technology, & Human Values* (2016): 3-16.

—. "What does transparency conceal?" *Privacy Research Group* (2013).