The Culture of Data Science: Meaning and Authority in an Epistemological Landscape

Claire D'Elia Maiers
Colorado Springs, Colorado

B.A. Music, Colorado College, 2005
M.A. Music, Tufts University, 2009
M.A. Sociology, University of Virginia, 2012

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Sociology
May, 2017

TABLE OF CONTENTS

ACKNOWLEDGMENTS

As becomes evident to many of us in the pursuit of knowledge, the submission of a dissertation is not the end of inquiry, but the beginning. There are many who have supported me as I worked to locate and articulate and establish the initial footholds of my intellectual path. I would like to express thanks to my dissertation committee: Sarah Corse, Isaac Reed, Simone Polillo, and Siva Vaidhyanathan. These individuals offered continued encouragement and engaged in many productive conversations with me during the development of this project. I would like to especially thank Sarah Corse for her commitment as a mentor, for the many hours she spent reading preliminary material, for providing guidance through the research process, and for encouraging me to pursue my intellectual interests.

I have also benefited from a curious and critical group of colleagues at the University of Virginia. The incredible group of thinkers at the University of Virginia Scholars' Lab (Bethany Nowviskie, Wayne Graham, Jeremey Boggs, and Eric Johnson) gave me my first experience with coding and the digital humanities and provided a community in which to workshop my ideas, giving me the confidence to dive head first into the unknown investigation of big data, algorithms, and data science. Julia Ticona gave me continued feedback, work shopped ideas, and shared in a venture into the intersection of sociology and technology studies. Sarah Mosseri, Francesca Tripodi, and Hexuan Zhang also offered support and shared in the experience of the Ph.D. process. In addition, with Professor Caitlin Wiley and Professor Deborah Johnson I engaged in conversations that helped me to situate my work alongside the science and technology studies perspective.

Finally, I would like to thank my family for their support over the past seven years. Grandmothers, my godmother, my brother, and others frequently offered words of encouragement. Both of my parents continually expressed their support for my choice to commit to obtaining a doctorate. In more ways than one, they made it possible for me to dedicate the better part of a decade to this goal. I also benefitted from sharing the dissertation experience with husband, Pete. He helped to facilitate my ability to devote my time fully to writing and provided valuable copy editing of drafts. Most importantly, Pete has shared my passion for scholarly exploration and curiosity and provided a constant source of challenging and energizing conversations. Thanks to all of you.

PREFACE

As readers and writers of dissertations know, the dissertation project usually represents a narrow slice of the broader intellectual interests of the researcher. This project is no different. In the broadest sense, the pages before you represent my persistent curiosity in the role of meaning in shaping our lives. This is an interest that that been with me long before my discovery of sociology. In research for my master's thesis in music, I attempted to deal with the ways in which differing values and the perceived role of music's place in society shaped the business practices of music organizations in the Boston area. It was this inquiry into how groups found, assigned, and justified value in music that first brought me to the sociological literature.

After embracing the sociological discipline as my own and beginning graduate work at the University of Virginia, my focus narrowed to the ways in which groups make and justify claims. I first explored this topic in my master's thesis which dealt with ways in which various groups made claims to ancestry. In thinking about how groups establish claims, I have been interested not so much in the maneuvers or language by which groups are able to establish and convince others of their claims, but in the underlying assumptions or ways of seeing the world that can be discerned by examining such claims in context.

The work from scholars like Isaac Reed and his concept of "landscapes of meaning" and Gabrielle Abend's idea of the "moral background" helped me to find a vocabulary for talking about this interest and showed me that others working in the cultural sociology were similarly trying to deal with this aspect of meaning-making.

With a growing interest in quantification and datafication as means to claim making, I began to look for similar work in the sociology of knowledge. However, I found that despite the attention given to worldviews in the early sociology of knowledge, especially the work of Karl Mannheim, and despite a notable cultural turn in the sociology of knowledge during the past few decades, this perspective was missing. The sociology of knowledge instead focused on practices as the stuff of culture. Attention to the subjective aspects of belief was by and large lacking from the conversation on how science worked to generate claims.

As I began collecting data for the dissertation, observing professionals using and making data science and interviewing data scientists, I found that documentation of practices alone—especially in the case of data-driven technology used in the neonatal intensive care unit—did not fully capture the complex ways in which data analytics factored into decisions and knowledge claims. To make sense of my observations, I found myself needing again to think about worldviews and background assumptions. In the following pages, you will find the resulting examination of those ways of seeing the world that encourage and unfold alongside the use of data science.

In generating this investigation into the culture of data science, I make two overarching contributions. I argue that the current scholarship on big data, algorithms, and data science needs

to include on-the-ground ethnographic accounts. There are two reasons for this. 1) As I show through the empirical chapters, there are variations in the symbolic orders or what I call *epistemological landscapes* among the various settings where data science is practiced and invoked. This kind of variation goes missed when data science is viewed primarily from the distanced critiques or discourse analysis that has dominated the critical data studies literature. 2) In addition, this kind of ethnographic work is a key component in ascertaining the ways in which data science will impart structural or material consequences. This comes across most clearly in the chapters on the NICU and the chapter on the pragmatic data scientist where I show that local contexts play a large role in tempering and filtering the practices of data science.

The second contribution speaks more directly to the sociology of knowledge and how we try to address the culture of the knowledge society. I argue that despite a focus on culture, the recent work in the sociology of knowledge has not given enough attention to the subjective experiences that constitute worldviews and which provide a foundation upon which some knowledge claims or courses of action become possible. Especially in the chapter on the neonatal intensive care unit, I show how such ways of seeing, the epistemological landscape, and attitudes towards epistemological authority shape treatment decisions and understandings of particular patients as either sick or well. As I address in the conclusion, this has implications for depictions and approaches to the knowledge society as well. The moniker of the knowledge society is used to indicate that we have entered a period in which knowledge production is a driving force both economically and culturally in our society. As data science becomes more prevalent in society and as knowledge settings permeate the social landscape, it is not only the practices or claims generated by knowledge settings that may spread out to the rest of society, but subjective aspects of knowledge settings as well. These may become the subjective worldviews that underpin social life in general. However, as Mannheim's sociology of knowledge suggests, and as the empirical work in this dissertation shows, that does not necessarily mean there is a single blanket manifestation of the subjective aspects of the knowledge society. Though there are certainly some shared attributes, local and contextual factors may push or alter manifestations of culture. This becomes clear when comparing the epistemological landscape of data science alongside the institutional context of the neonatal intensive care unit. People in both settings are more comfortable with knowledge generated by numbers and worry over human subjectivity, but the ways in which they deal with domain expertise and experience is markedly different. This shows that continuing to attend to these subjective aspects, worldviews, and epistemological landscapes will be an important part of the sociological effort to understand what it is to live and to know in a knowledge society, as it will be upon these landscapes that we build meaning, construct experiences, and choose actions.

CHAPTER 1

Introduction

In early June of 2016, I wandered across the medical campus of Augustine University Hospital for the weekly meeting of the medical analytics team I had been observing for more than a year. Each week, the team would meet in one of the hospital's conference rooms over lunch to discuss their on-going progress, prepare research papers or conference presentation, and to strategize moving their work forward in collaboration with other labs and organizations. The conference room had a stately feel to it: large oil paintings of serious-looking, aged men in academic garb lined the wood-paneled walls. Several months into my field work, the members of the medical team had become accustomed to my presence. If the meeting was crowded, I would usually take up a seat on the periphery of the room, trying to be more of a fly on the wall than an active participant in the meetings. Given that Dr. Ibez, the founder of the team, started most of the meetings with a round of introductions, this became increasingly difficult over the months. I usually introduced myself as sociologists interested in understanding the ways in which data analytics influenced medical knowledge. Over time, Dr. Ibez and the others began asking for my opinions on the topics at hand, especially once I started shadowing the clinicians using their Horizon monitor in the neonatal intensive care unit (NICU). Wary of influencing the very processes I was there to study, I usually tried to avoid these moments. On this particular afternoon, the meeting was less crowded and I was compelled to take a seat at the conference table across from Dr. Ibez.

As the meeting opened, Dr. Ibez turned to me and said, "I don't want to put Claire on the spot, but—" This was a frequent and jovial tactic by which Dr. Ibez would invite members of the team to speak about recent developments or speculate on possible directions the team might take.

He continued to ask for my opinion on surveying clinicians prior to the introduction of some new algorithms and monitors that the team had developed for adult populations in the hospital. Dr. Ibez's group had become increasingly concerned with being able to demonstrate that their algorithms influenced the clinician's experience and being able to describe the mechanisms by which this occurred. They had missed the opportunity to document these aspects of medical care when they implemented Horizon in the NICU in the early 2000s and were eager not to miss the opportunity to do so as they launch the trial phase of their new technology in the adult intensive care unit. Dr. Ibez and his team were planning to administer a brief survey to clinicians both before and after its implementation. Earlier in the week, he had suggested to me that they use surveymonkey.com to administer a series of questions that, for the most part, prompted respondents to answer either "yes" or "no".

As we began to discuss this approach in the meeting, I suggested that the team might get better responses if they administered questions during a short focus group instead. Clinicians already would need to attend a training session before the new technology launched, and it seemed like a perfect time to engage in a more nuanced conversation about their care practices and means of detecting illness. The medical analytics team was utterly skeptical of my suggestion of a focus group, rather than a survey. One of the developers asked me, "how do you know change is good? How do you quantify it?" For him, any change that I might find through a series of focus groups could not be trusted without a clear way to capture it with numbers.

To my mind, the benefit of focus groups seemed obvious. They wanted to understand the mechanisms and whether or not the technology changed medical practice. Does it cause nurses and doctors to act differently? Do they communicate differently? Why might the technology prompt them to make certain choices? These are questions that can be answered by observing

and talking to clinicians. But the team was not satisfied. They needed something that could be quantified in a straightforward way, without a human observer. Quickly, the conversation turned away from me. My sociological training had clearly not provided them with an answer that they saw as a legitimate means of capturing reality. Instead, they began to propose methods that circumvented any human account—whether it be from the researcher or the clinicians themselves—and that relied heavily on quantitative analysis.

In place of the short, easily executable focus groups I recommended, they suggested technologically driven (and comparatively expensive) solutions. First, they considered installing eye-tracking software within the monitor so they could quantify how often clinicians viewed the screen and for how long. This solution wouldn't tell them how the monitor's algorithms influence the clinician's actions. So, a member of the team suggested having each clinician in the unit where a badge capable of tracking their movement during their shifts. They would collect this information for a time before the new technology was installed, and then again after installment to mathematically compare the patterns. Finally, they suggested calculating a time series from the electronic medical records for events both before and after launching the technology. Perhaps, they speculated, this would capture a difference in the timing of orders for lab tests and medicines, thus allowing them to show how their technology altered the timing of care decisions.

Any of the solutions that they suggested are feasible, though they would be considerably costly, requiring the purchase and integration of additional technology and presumably the hiring of specialists to install and manage this technology. Not to mention, the time required to identify, purchase, and install such technology would stall their progress a great deal. Short focus groups, conducted for free by a sociologist eager to secure access to her research site and

genuinely interested in being useful to her informants, would not incur any financial costs and be considerably faster to execute. The technologically-saturated and quantitative approaches suggested by the team may have also been an effective method for documenting some aspects of clinicians' behavior and patient care. It would tell them if clinicians paid attention to the monitor, and it would allow for some mathematical models that capture differences in movement and lab orders before and after installing the technology. However, it would not capture all of the ways in which this new predictive monitoring technology would change the clinical environment. The ways in which clinicians processed the information, whether or not they valued it, if they discussed it with their colleagues, and if clinicians derived the intended meanings from the algorithm would all remain outside the scope of this analysis. In short, these quantitative methods would tell the team some aspects of *what* changed after installation of the new technology, but it could do little to tell them *why*.

How do we make sense of this story in which a group of highly intelligent, highly-educated, well-intended, accomplished developers and physicians with a track record of improving patient care would so easily dismiss the accounts of clinicians and the work of a human observer as a way to know how and why new technology changed patient care? To be clear, for the doctors and nurses who participated in the medical analytics team, this meant that they dismissed the accounts of their own colleagues and members of their profession. They did not trust people just like themselves to recognize the role of technology in their own work and decision-making. The answer to this puzzle lies in understanding the epistemological landscape and institutions in which these professionals operate. To be fair, with some convincing Dr. Ibez and his team did allow me to carry out the very methods I suggested to them as part of my own research on their technology in the neonatal intensive care unit. Though I sensed constant

skepticism with regard to my research questions and approach, they were continually supportive, encouraging my presence at the meetings and sponsoring my access to sensitive patient spaces in the hospital such as the intensive care units. However, the conversation I have recounted here reflects the ways in which the epistemological landscape of data science shapes the ways in which people think about knowledge, evidence, and legitimate ways to access truth. In the following pages, I explore this epistemological landscape and how it unfolds among data scientists, in public presentations of data science, and in environments where data-driven tools influence decision-making. I argue that understanding this landscape and the subjective experiences of those who create and employ data science allows for a more nuanced appreciation of the ways in which data science is shaping our lives.

Data Science: An Institution

Data is an institution. It is a symbolic order for making sense of the world and a set of practices and processes that are increasingly ubiquitous across fields and organizational settings; it is a taken-for-granted way of making sense of the world (Douglas 1986). Though not always put in such blunt terms, others have recognized the institutional features of data as well. Statisticians and data scientist, themselves, have recognized the emergence of a new "algorithmic modeling culture" associated with data science (Breiman 2001). Crawford et al (2014) also come close to claiming that data is an institution when they assert that big data is not just a technological phenomenon, but a political, cultural, and economic one as well. Gillespie argues that the processes associated with data such as data mining and algorithmic assessment constitute a "*knowledge logic*, one built on specific presumptions about what knowledge is and how one

should identify its most relevant components" (2015). Bell recognizes that big data is embedded in a "socio-technical imagination," one that requires social scientific investigation (2015:9). As I will detail below, data science and its accompanying worldview is manifest through the software, processors, and databases needed to support its use, as well as a growing number of institutions, organizations, governments programs, and academic degrees dedicated to training a new generation of data scientists and employing its methods. This combination of worldviews, practices, material structures, and social organizations suggest that data science is an emerging institution, one that is likely to stick around and exert great influence over knowledge production and decision-making.

Below I outline material and cultural developments that point to the intuitional status of data. I begin by discussing the technological changes that practitioners often associate with the advent of big data and data science. I then elaborate on a few of these features. This is followed by a discussion of the ways in which data science and big data have captured public attention and inspired the foundation of new organizations.

The Material and Practical Aspects of Data Science

The Data

No matter the industry or sector, the use of data to generate knowledge relies on similar processes and capabilities: the means to collect a wide variety of data, large data warehouses capable of storing mass amounts of data, and the means of making sense of this data. Contrary to the claims of some observers, the dominance of data science is more than an extension of statistical methods. Data scientists may still draw upon traditional statistical models and research

practices, and they may work with "smaller" data sets that might not justify the title of "big data."[1] Among practitioners, there is no single, agreed upon definition of data science (Mayer-Schönberger and Cukier 2013). Those who use and produce data science usually refer to a bundle of technological conditions, practices, and perspectives when they use the term data science or related concepts such as data analytics, big data, and algorithms. For this reason—because data science is a cultural concept that signals a loose set of approaches to problem solving and knowledge production—I refrain from providing a specific definition. Instead, I outline some of the key associations and phenomena that frequent conversations about data science. I begin by outlining the three "V"s of big data. Though big data is not synonymous with data science (data science techniques can be executed on data sets that some would consider too small to count as big data), it is a strongly associated concept, and most data scientists recognize that changes in the "volume," "variety," and "velocity" of available data are related to their ability to do their work. I then discuss some of the new analytical techniques are often used by data scientists and provide brief description of algorithms. I conclude this section by describing the kinds of problems which data science is often called upon to solve.

Big data has become widely associated with three "V"s that describe the new data reality in which we live. These are volume, variety, and velocity (Laney 2012; Press 2013; Podesta et al. 2014). Though other organizations and authors have added additional "V"s such as veracity and value, the original three those most prominent in the discourse surrounding data.[2] These three features of big data focus on the attributes of the data itself, rather than what can be done

[1] There is little agreement on how big a data set needs to be to constitute "big data" (Bell 2015). A data set may be large in terms of the number of cases or in terms of the amount of data points associated with each case—both instances can create challenges for computing power. Further, a data set that may seem "small" in one industry or sector might be revolutionarily large in another.

[2] An IBM blog posts adds a fourth "V" to the characteristics of big data: they include "veracity," pointing to issues related to trusting the data and data quality (IBM 2017). Ishwarappa and Anuradha (2015) add a "value" as a fifth feature.

with it. Though the three "V"s have become such a taken-for-granted aspect of the concept of big data, that they are often not attributed to any given author, it seems likely that the source of these concepts comes from a 2001 Gartner Inc. report in which Doug Laney (2012) discussed the emerging data challenges her was seeing in the world of e-commerce. Although he does not use the term "big data" in the original article, his work was an early articulation of velocity, variety, and volume as defining characteristics of changing technological landscape. I begin my discussion of the technological changes that have encouraged the adoption of big data and data science with these three features because they simultaneously describe the conditions under which data science is used and begin to point toward how data scientists understand their work.

Velocity: Data velocity may refer to one of two features. First, it may refer to the "real-time" effect of data analytics to alter the experience of people as they shop, navigate a city with global positions systems (gps), or interact with social media (Podesta et al. 2014). Data is created and can then be analyzed and deployed faster than ever before. For example, when the human genome project was finished in 2003, scientist had spent 10 years to complete that project. The same amount of genetic data can now be sequenced in a single day (Mayer-Schönberger and Cukier 2013). Second, velocity may refer to the varying rates at which data is created and captured. The rate of data creation, or velocity, may vary between sources—Twitter data may stream in at faster rates than that created by sensors on weather balloons. At the same time, Twitter data may vary a great deal depending on the time of day, on day of the week, and on the occurrence of outside events such as the Super Bowl or the presidential election (Sicular 2013).

Variety: Data variety refers to vast number of sources from which data may be collected and stored. These include sources such as social media activity, credit card purchases, weather

sensors, personal tracking devices such as the popular Fitbit, video feed from security cameras or traffic cameras, vitals monitors in hospitals, arrest records, gps data from smartphones, data from parking meters, clickstream data from websites. In the age of the "internet of things," where an increasing number of devices, objects, and even buildings are embedded with network connectivity, the possibility for new data points is continually expanding.

Volume: Data volume is perhaps the feature that is most associated with big data. The increased variety of data sources, increase in digital activity, and advances in data storage have led to an unprecedented amount of data creation and capture of that data. Hilbert and Lopez (2011) estimate beginning in 1986, the world's storage capacity grew by 25% each year. By 2007, there were over 300 exabytes of stored data (ibid). Given that one exabyte is equivalent to one billion gigabytes, this is clearly an overwhelming amount of data. In addition, most of the data created and stored is in a digital format, making it more easily available for access, transfer, and analysis. According to their figures, 94% of all data storage capacity was digital in 2007 (ibid). As already mentioned, in addition to storage capacity, human activities increasingly take place through devices and locations that allow for digital capture of data. This trend is due in large part to the increased number of people participating in online and mobile activities. In 2010, there were 5 billion mobile phones in use, each producing its own stream of data (Manyika et al. 2011). According to the CTIA, within the United States alone, there were 377.9 million devices subscribed to wireless networks at the end of 2015 (2016). Facebook claimed to have 1.26 billion daily users as of December 2016 (http://newsroom.fb.com/Company-Info/ 2017). According to some estimates, we are currently generating 2.5 quintillion bytes of data every day (IBM 2014). This explosion in the amount of available data, has also led some to associate the "volume" of big data with a shift in method and analytical approach. While many traditional

statistical practices rely on relatively small samples of data, big data allows data scientists to analyze entire populations of data, including all available data in the analysis (e.g., Mayer-Schönberger and Cukier 2013).

The Processes

In addition to the 3 "V"s, there have been changes in what can be done with the data. An increase in processing power and the ease with which data can be stored are central. In 2010 (Manyika et al. 2011) it cost just 600 dollars to purchase enough storage space to store all of the world's music. Cloud-based services have also made storage and processing more accessible. Products like Amazon Web Services, launched in 2006, and Dropbox.com, founded in 2007, allow organizations and individuals to store vast amounts of data without purchasing and maintaining the physical hardware to do so themselves. Under these conditions, all of the data generated through remote sensory devices, social media, or other online activity can be captured, becoming the potential material for analysis through data science.

In addition to storage, services like Amazon Web Services can be used to rent processing power. In essence, anyone with a credit card and a basic computer can access a super computer through Amazon. Simply prepare the command you want to execute, navigate to Amazon Web Services, select the amount of power you need (more power will mean the analysis is executed more quickly), and let Amazon's computers do the work. Customers pay by the amount of time spent renting the computing space. The ability to store more and process faster has had significant implications for the ways in which people use computers to produce knowledge. Computational tasks that once took months can now be accomplished in a matter of hours. This does more than simply make analyses faster to execute. It changes how data scientists approach

their work. When tasks are so costly and time-consuming to execute, it becomes essential that they are done right the first time. The right model, correct algorithm, proper organization of the data must all occur in advance. Now, these aspects can be tweaked and altered, and the analysis can be run again. In short, many models, algorithms, and techniques can be experimented with and data scientists can pick the one that gets the best results for the task at hand.

Central to the processes of data science is the expansion of machine learning. In machine learning techniques, a computer program determines which algorithms, models, features, or even categories allow for the best assessment of the dataset. This contrasts with statistical methods where a human must decide which model and which features are most salient for the analysis. Machine learning is not new. Its origins date back at least to the 1950s when Arthur Samuel designed a computer program that learned to play checkers (Samuel 1959). However, the expansion of data collection into to new areas in tandem with increased computing power have made it easier to apply machine learning to almost any topic. As already suggested above, the use of machine learning represents a shift in how knowledge is created, allowing the computer rather than researchers to determine the model by which predictions and conclusions are made.

Along with machine learning, references to algorithms abound in discussions of data science and big data. An algorithm, in the simplest terms, is a set of instructions. A common algorithm that many of us encounter daily is Facebook's Edgerank algorithm. This algorithm determines the content that shows up in a user's Top Stories News Feed. The Edgerank algorithm relies on two kinds of inputs: objects and edges. Objects are any kind of content such as photos, status updates, videos, or links. Edges are actions that are taken by users in interacting with these objects, such as sharing content, clicking the like button, or leaving a comment.

Facebook creates a score for each object by calculating three scores related to edges: affinity, weight, and age. These three scores are then added together to generate the edge rank and ultimately which content gets displayed (Bucher 2012). It is through this set of directions that particular content is included, excluded, and ordered in a user's Facebook feed.

In data science, algorithms are both a means of production and a product. Data scientists may rely upon established algorithms to execute an analysis. Even methods associated with traditional statistics, such as a logistic regression, constitute an algorithm for producing analytics. In machine learning, algorithms are also important. For example, latent dirichlet allocation is an algorithm used in natural language processing. Through an unsupervised process, it identifies themes or topics contained within a corpus of texts (Blei 2012). This algorithm is readily available for data scientists to use as they produce an analysis. However, algorithms are also the product of data scientists. For example, the Rabin-Karp (Karp and Rabin 1987) is a set of instructions that tells a computer how to find strings (or a series of words) in texts that match a set of source texts. It is a product in the sense that it had to be developed by someone, but it also is a tool used to produce stand-alone analyses or to data-driven platforms, such as those used for plagiarism detection like Grammarly (see https://www.grammarly.com/).

The Applications

Although advocates for data science may praise its ability to solve any number of problems, I found that the data scientists with whom I spent time tended to solve a small set of problem types. These categories are not discrete—they may overlap quite a bit in practice. In addition, they are not meant to reflect differences in technique or methods. Instead, they point to the kinds of problems that data scientists are employed to solve.

1) Making Data Available: In some instances, data scientists are tasked with establishing or modifying a data infrastructure so that data may be collected and subsequently used for analysis. This involves dealing with both hardware and software. For example, data scientists may help to secure the right hardware that will allow weather sensors to detect aspects like temperature and wind speed, a server to store this data, and a process to transfer it from the sensors to the server. In addition, disparate data sources often need to be made commensurable. If an organization uses sensors from different manufacturers, they may find that this data is recorded in different formats, standards or measurements, or at different time intervals. In order to make this data useful for analysis, the data scientist may be tasked with coming up with a method to convert these disparate measurement to meet one standard. In addition, the data scientist might choose or design the format for the database, essentially determining the organizing principle for the data.

2) Data Marts: Sometimes data scientists will build upon the work above by designing a platform by which the data now collected and stored in an organization's database can be easily accessed and queried. This work often involves determining which data might be relevant for the organization and the criteria by which analysts may want to filter the data. For example, a data mart platform of sales data might allow marketers to filter the total amount of sales by date, by location, or by product. Most of this data is descriptive in nature, but data marts may also include some basic analytical tools, such as the ability to calculate growth of sales by quarter.

3) Relevance Algorithms and Platforms: Relevance algorithms are used to determine which information or cases to bring to a user's attention. For example, a relevance algorithm may determine which products to show to consumers visiting a webpage. When Amazon.com suggests additional products based on your search history, the algorithm is trying to determine

what products might be relevant to you. In other examples, a relevance algorithm might monitor thousands of insurance claims and suggest a small subset to an investigator as possible instances of fraud or analyze out-patient data, presenting a list of patients who might require interventions to keep their health on track. These algorithms are often deployed through platforms that monitor databases and then provide users with a list of possible cases of interest. In these cases, the purpose of the relevance algorithm is to make work more efficient and effective for some kind of human user. In the example of the insurance company, the idea is that either fewer human agents will be needed to investigate the same number of cases or that success rates and speed of these agents will improve when the algorithm assists them in identifying relevant cases.

4) Risk Algorithms and Platforms: Rather than survey a database for relevant cases, risk algorithms focus on a set of cases in which the human user already has an interest. The algorithm then produces a risk score that informs the user of the chances that a particular outcome might occur in that case. For instance, the Horizon monitor that I will discuss in chapter 5 uses a risk algorithm. It monitors the patients under the clinicians' care and produces a risk score for each one meant to indicate that patient's risk of getting sick. Similar platforms might predict the chances that a particular cargo ship will sink or get delayed in transit or the chances that improvised explosive devices (IEDs) have been buried along particular roads. These kinds of algorithms are designed to help humans make decisions: which route should the convoy take?; should this patient receive antibiotics?

This typology does not cover all of the possible applications of data science. However, these were the most common problems that the data scientists interviewed in this study addressed in their work and the kinds of solutions that they offered for their clients and collaborators.

Having outlined the technological changes, techniques, and some of the applications of data science, I now want to turn to the cultural side of the data science phenomenon.

The Cultural Phenomenon

The emergence of data science is as much as cultural phenomenon as it is a computational or material one. Advocates of data science have argued that it represents a new paradigm in producing knowledge and uncovering truths. In the popular imagination, data science, big data, and algorithms have been depicted as a great advance in the ability of science to make good on its claim to holding the keys to a better world.

Data Science in Popular Discourse

Media coverage of "big data" and "data science" has increased significantly in the last decade. In searching the Factiva database, a collection of 6,000 periodicals and newspapers, the term "big data" received just 211 hits in 2008, but that number increased dramatically by 2016 in which the same search rendered 62, 156 hits. Similarly, the term "data science" increased from just 38 hits in 2008 to 9,735 in 2016. If nothing else, this demonstrates that big data is getting significant press attention, increasing the chances that it enters the public's imagination. Part of the enthusiasm for data science, big data, and data analytics comes from its promise to get at the "real" facts and to avoid the pitfalls of human subjectivity. Newspaper and magazine articles extoll the potential of data analysis to generate business solutions (e.g., Chamorro-Premuzic 2014), to create a fair job market and happier workers (e.g., Peck 2013), to reduce disease (Rosenberg 2015), to better predict student success (e.g., Ungerleider 2013), and to design better cities (e.g., Gupta 2014).

In 2016 PBS aired a special called *The Human Face of Big Data*. In this program experts in the field convey a sense of hope contained within the promise of big data (Smolan 2016). We learn that "almost everything is measurable and quantifiable." "Almost everything we do today leaves a trail of digital exhaust." This is generally portrayed as a good thing because, "the more information we get, the larger the problems will be that we solve." Further, the narrative of the film teaches us that, when harnessed, we can think of data as "a microscope." With this new tool, "we are able to examine something that is around us" that has "a structure and patterns and beauty that are invisible without the right instruments, and all of this data is opening up our ability to perceive things around us." This orientation to the world—the inability of humans to properly perceive and the new technological ability of data to perceive on our behalf—structures the work of data scientists.

A New Paradigm?

Data scientist, its critics, and advocates are divided when it comes to determining whether or not or to what degree data science and big data represent a new paradigm. In the simplest terms, the scientific method has long focused on generating theories or covering laws that depict the ways in which the world works. Scientists generate a hypothesis, a statement of what might be true about the world, and then produce experiments to test this hypothesis. This work might lead to a theory or law that is then employed in future scientific work. When it comes to the new practices and capabilities of data, the question becomes whether or not theories, those explanations produced by experts that claim to depict the way in which the world works, are still necessary. In the following passage, I describe two examples of the argument against theory.

In a much cited and debated 2008 article from Wired Magazine, Chris Anderson begins:

> "All models are wrong, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

This quote from George Box will make a return in the following chapters on data scientists. For now, I want to draw your attention to the way in which Anderson draws a line in the sand, distinguishing the present from a somewhat inadequate past, one is which we had little choice but to settle to an epistemologically flawed way of knowing the world. Anderson continues to tell the reader how the present is different: there has been an exponential increase in the availability of data and increased computing power has equipped us to move forward without the make-shift tools of the past:

> The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years. Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. […] There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Anderson's stance on data science gives the impression that up until this point, scientists have been forced to suffer with blunt and inadequate tools, that of hypothesis, models, and theories. Thankfully, computational advancements will finally let us move away from these theories that "imperfectly" depicted reality. One gets the impression that holding on to the old ways is a dire mistake driven by nostalgia and little else.

Peter Norvig, currently the Director of Research at Google, has made similar claims about the outdated nature of theory or model-driven knowledge production. In a much cited 2011 blog post, Norvig offered his objections to comments made by Noam Chomsky during the 2011 MIT Brains, Minds, and Machines Symposium. Chomsky expressed his opposition to using statistical models as the primary means of generating knowledge, especially in his own area of linguistics. Instead, he argued for the continued role of theory to produce "why" explanations in science. Norvig understands Chomsky's remarks as an objection to the "algorithmic modelling culture," (Breiman 2001) or an approach in which "complex algorithmic approaches […] are used to estimate the function that maps from input to output variables, but we have no expectation that the *form* of the function that emerges from this complex algorithm reflects the true underlying nature" (Norvig 2011). This is in contrast to a "data modelling approach," in which "It is the job of the statistician to wisely choose an underlying model that reflects the reality of nature, and then use statistical data to estimate the parameters of the model" (ibid). In response, Norvig argues that the gathering of facts—not the generation of theory—is actually the dominant activity of science. Further, he disagrees with how Chomsky measures success in science. For Chomsky, explanations are paramount. Norvig argues that this is not necessarily the case. Instead, he tries to show how modeling the world, or successfully predicting future states, does suggest the success of science. He does so in an unexpected way.

Rather than making an epistemological argument at this point, he instead shows the popularity of this approach. Norvig is sure to say that common usage and financial success should not be metrics of scientific success, and yet he uses these justifications nonetheless. He proceeds to list the applications to which probabilistic modeling of language has been put (such as speech recognition or search engines). Of its prominence, he says, "clearly, it is inaccurate to say that statistical models (and probabilistic models) have achieved *limited* success; rather they have achieved a *dominant* (although not exclusive) position." He continues to say that, "another measure of success is the degree to which an idea captures a community of researchers," and then proceeds to describe how almost all of his colleagues have adopted these methods. Finally, he turns to financial justification, arguing that, "it is worth noting that the [probabilistic models of communication] create several trillion dollars of revenue each year, while the offspring of Chomsky's theories generate well under a billion."

Norvig closes his objections with a striking rhetorical move. He describes an incident from 2011 in which TV personality Bill O'Reilly claimed that the movements of the tides could not be explained without looking to God as the ultimate mover of things, the ultimate explanation. O'Reilly was chastised for his apparent lack of familiarity with the scientific explanations of tides. Norvig notes that, for O'Reilly's followers these scientific explanations do not matter because he has "gotten exactly to the key issue: *why.*" For some people, Norvig argues, "O'Reilly is correct that these questions can only be addressed by mythmaking, religion or philosophy, not by science." Then Norvig goes in for the kill when he states, "Chomsky is in complete agreement with O'Reilly." He continues, "Chomsky believes a theory of language should be simple and understandable, like a linear regression model where we know the underlying process is a straight line, and all we have to do is estimate the slope and intercept.

[…] The problem is that reality is messier than this theory." Through this move, Norvig equates the desire for explanations of why and the desire for theories to the "mythmaking," and to the misguided efforts of "religion or philosophy."

Although many data scientists (including some in this study) object to the perspectives of Anderson and Norvig, these kinds of statements contribute to the perception that data science will truly revolutionize our ability to know the world. The subtext suggests that this shift fits within a grand narrative of science, one in which we progress toward better, truer, and more complete understandings of the world. Just as science overcame the misplaced explanations of religious myths, so too will data science overcome the silly and inadequate explanations of theories and models. This perspective on the shifting epistemological paradigm will serve as a theme that cuts through much of this project. Needless to say, critics of data science have objected to these claims, noting that although it certainly changes the practices and epistemology of knowledge productions, data science is neither truly free of theory nor necessarily gets us closer to truth than other methods. I will return to these objections in the section on critical data studies and again in chapter 2.

The Reach of Data Science

As I have depicted above, the reach of data science goes beyond the offices and research labs in which data science is used. Instead, it has become a part of public discourse and popular imagination as people increasingly understand big data and data science as powerful means for producing knowledge and solving problems. For this reason, the degree to which the techniques of data science have become integrated with the rest of the social fabric cannot be fully captured

through quantified metrics. In his study of auditing practices in the United Kingdom, Michael Power (1997) makes a similar claim. While he offers up several possibilities for measuring the prevalence of auditing practices, he notes that,

> such an exercise would only conceptualize the rise of auditing in quantitative terms. [..] A quantitative approach like this would not capture the sense in which the growth of auditing is an explosion of an idea, an idea that become central to a certain style of controlling individuals and which has permeated organizational life (4).

So too, quantitative attempts to demonstrate the reach of data science fall short of conveying the ways in which this bundle of practices have permeated the imagination of the public, non-specialists, and data scientists alike as saving grace, capable of many good deeds. Nevertheless, there are material indicators of data science's increasing scope. Professionals in a wide variety of sectors have enthusiastically turned to data science. Data-driven tools and analysis have been employed in financial investment, marketing, academic research, health care, education, security, and many other areas. For example, the consulting firm Sociometric Solutions claims to maximize the potential of employees through interventions derived from data mining information collected from devices worn by employees to capture social signals such as face-to-face interactions, body movement, and employee location in space. The data firm Palantir uses large data sets to generate solutions to problems that range from terrorism to natural disasters. In healthcare, the American Society of Clinical Oncology has developed a database and analytical tool called CancerLinQ which culls information from electronic medical records and returns results to medical practitioners in real time, allowing them to alter treatment plans based on this data.

The U.S. government is also investing heavily in big data and data science. In 2012, the White House announced the "Big Data Research and Development Initiative" in which 200

million dollars were pledged to data science efforts that would advance solutions in healthcare, economic growth, and education. As part of this effort, the government funded the establishment of network of Big Data Regional Innovation Hubs. These hubs are designed to bring together data science professionals from across government, industry, and educational sectors. Individual hubs specialize in topics particularly relevant to their region. For example, the "West Hub" is affiliated with the University of Washington, University of California, San Diego and University of California, Berkeley and specializes in "big data technologies and data-intensive discovery, managing natural resources, and hazards and precision medicine." (Kalil et al. 2015).

Finally, new educational and professional organizations have been founded to meet the demand for data analysts. The Institute for Advanced Analytics at North Carolina State University tracks statistics on masters programs in data science, business analytics, and analytics. For 2007, they list just a single program. By the year 2016, just 9 years later, they indicate that 125 master degrees are available within the United States. These include full-time, part-time, and online degree programs. Tuition for these programs ranges from $9,500 (The University of Alabama) to $75,000 (Sloan School of Management) (Institute for Advanced Analytics 2017). Among 21 of the top public and private universities in the United States, 16 now offer undergraduate or graduate degrees in data science. All of these programs were founded between 2013 and the present.[3] The techniques associated with data science have also captured the attention (and material support) of existing disciplines and institutions. A prime example of this is the digital humanities. Though its origins can be traced primarily to literary studies, the digital humanities, as a discipline, now focus on the creation of digital tools and application of digital

[3] This data comes from searching for data science degree programs and initiatives among the top ten universities in the U.S. News and World Report Rankings for "National University Rankings" and "Top Public Schools" lists from 2017.

methods to the reading and interpretation of texts ranging from Shakespeare to symphonies (Schreibman, Siemens, and Unsworth 2004). In place of nuanced and deep interpretations of a single or several texts, one variant of digital humanities advocates for "distant reading" (Moretti 2005). Much like those that advocate for data science due to its computational ability to consider entire populations of data, the methodological approach of distant reading suggests that large-scale data analysis of a multitude of texts is the best means to understanding the nature of literature.

It is clear that in the uncertain economic environment that followed the 2008 recession, individuals, universities, and industries are placing heavy bets on data science. In addition to the material resources being invested, data science is enjoying an increasing cultural authority. As media coverage makes clear, such authority is derived partially from the hope embedded in data science. Given the increasing ubiquity of this institution and the potential epistemological and cultural consequences suggested by its spread, social researchers need to attend to the ways of seeing and worldviews that inhere in practices surrounding data.

Critical Data Studies

The Social Construction and Social Consequences of Data

Not all the press surrounding big data and data science has been positive. For instance, a 2016 ProPublica article brought attention to racial bias in algorithmic risk scores designed to predict the chance that a criminal will commit additional, future offenses (Angwin et al. 2016). As data science, big data, and algorithms have grown in the application and entered the popular imagination, scholars too have begun to offer a critical perspective on these developments (e.g.,

Cheney-Lippold 2011; Andrejevic 2014; Gillespie 2014; Beer 2015; Crawford et al. 2014; Striphas 2015; Seyfert and Roberge 2016). Due to their powerful position in decision-making processes, many scholars have drawn our attention to the social construction of algorithms (Anderson 2012; Gitelman 2013) and emphasized the importance of opening up the "black boxes" through which algorithms turn big data into knowledge and decisions (e.g., Pasquale 2015). These studies usually attempt to examine the detailed ways in which algorithms work and to identify potential consequences. In outlining their makeup, researchers aim to better grasp how algorithms make decisions and shape the world in which they operate (Beer 2016). For example, Bucher (2012) unpacks the criteria by which Facebook's Edgerank algorithm determines what to show users in their feeds. She demonstrates that the assumptions and values embedded in the algorithm may lead to undesirable consequences, namely that some users are rendered invisible to others. In another example, Beer (2015) shows the ways in which the use of data analytics in football is altering the criteria by which players are recruited and even the skills which they choose to develop during training. Now, when players enter the field, they are not just playing a game, but "playing the stats" (ibid: 6). The scope of algorithmic and data-driven consequences is incredibly broad. They shape public discourse (Couldry and Turow, 2014; Gillespie, 2014), formations of the self and identity (Cheney-Lippold 2011), organizational activities (Ribes and Jackson, 2013), and structures (Andrejevic, 2014).

Cultural Accounts of Data

In addition to accounting for their makeup and effects, critical data scholars have also articulated an interest in unpacking the appeal of algorithms and data and for developing an account of what kinds of knowledge, world views, and selves unfold in the use of data (e.g., boyd & Crawford 2012; Beer 2016; Dalton, Taylor, & Thatcher 2016). With regard to

understanding the appeal of data science and big data, scholars have recognized that data and algorithms are attached to cultural conceptions of objectivity and truth and "evoked as a part of broader rationalities and ways of seeing the world" (Beer 2016:7). Data science fits neatly into a culture in which numbers and metrics are taken-for-granted representations of reality, thus endowing them with objectivity (Poovey 1998; Espeland and Stevens 2009). It is in this context, that advocates can claim that data brings an end to theory.

When considering the kinds of knowledge, worldviews, and selves unfold in the use of data, critical data studies has made the most advances in addressing the meanings and implications reflected in technical, systemic, and discourse aspects of data science. By looking at the techniques involved in data science, critics have pushed back against the end-of-theory argument (Bowker 2014; boyd and Crawford 2012). Of this claim, Bowker (2014) asks, "do we need theories, and do theories need categories?" (ibid:1796). He argues that categories are a type of theory—they are not natural phenomenon, but a social one. Whether we invoke the categories of gender or of more academic ideas such as socioeconomic status, we are using a model and theory of the world. Given that data cannot be constructed and organized without such categories, theories, on some level, are central to data science as well. He concludes that, "just because we have big data does not mean that the world acts as if there are no categories. And just because we have big (or very big, or massive) data does not mean that our databases are not theoretically structured in ways that enable certain perspectives and disable others" (ibid:1797). This same analytical tactic of considering the techniques of data practices have also allowed scholars to point out that data is never, "raw" but is shaped by human actors from its very inception (Gitelman 2013), that the reality portrayed through data is always partial (Gregg 2015),

and that despite claims to prediction, algorithms stabilize knowledge around *past* events and behavior (Bell 2014).

A Call to Study Subjectivities and Data

Fewer projects have tackled the worldviews, meanings, and subjective experiences that circulate throughout the contexts in which data analytics are constructed and employed. To fill this gap, many have called for ethnographic accounts of data (Seaver 2015; Beer 2016; Pink et al. 2016) that will allow us to better understand how people make sense of data, the degree to which it is endowed with the authority to make knowledge claims, or how it enters into decision-making processes. The importance of including such accounts in the critical study of data goes beyond satisfying scholarly curiosity. Instead, the empirical conditions by which data interacts with social outcomes make this kind of inquiry central to the effort to understand the role of data in society. While a great deal of attention has been given to the algorithms that, once constructed, deliver decisions and consequences—think for example of credit scoring (Fourcade and Healy 2013) or the Edgerank algorithm already discussed (Bucher 2012)—many algorithms do not impart consequences through such automated means. Instead, other contexts such as the professional settings in healthcare, criminal justice, government, or marketing act as conduits through which the results of data science and data analytics are filtered as they shape decisions and claims. In these settings, subjective human experiences, including social interaction and interpretation are important components of this filtering process.

Accounts of practitioners are important for a second reason. As indicated above, much of the research that attempts to address cultural aspects of data does so through an analysis of public discourse and rhetoric (e.g., Puschmann and Burgess 2014) or the analysis of quantified or

algorithmic objects (e.g., Bucher 2014) or systems (Cheney-Lippold 2011). If, as critical data scholars suggest, data and algorithms are becoming part of the process by which actors actively construct the realities in which they live, actual accounts of those actors are a crucial part of analyzes such processes and consequences. Making such claims without the subjective experience of actors risks a misrepresentation or misunderstanding of the cultural consequences of data. Despite the repeated call for research that addresses the subjective experience of data, only a handful of studies have ventured down this road (Christin 2014; Petre working paper).

This project takes up this call to unpack the meanings and worldviews that people draw upon as they create and utilize the techniques and products of data science. In doing so, this research helps to fill in the gap in critical data studies by providing an ethnographic and interview-based account of data science. In addition, by including the subjective perspective of those who encounter and practice data science, this project offers a check on the critiques of data science which have been most often made from a more distanced encounter with the data phenomenon. Fortunately, the analysis of the subjective aspect of data science also holds the potential to make a contribution to the sociology of knowledge. It is to developments in that area of scholarship that I now turn.

An Interpretive Approach to the Sociology of Knowledge

The Problem of Relativism

In addition to contributing to a better understanding of the ways in which data science works upon our society, this dissertation makes a contribution to the study of knowledge. This aspect of the project reflects my attempt to deal with how we, as sociologists, approach meaning

and action. In doing so, I take a cultural approach to data science, asking how data science both creates and unfolds within a particular way of seeing the world. I ask, how do data scientists claim "to know" something and in what kind of cultural contexts is data science seen as capable of answering questions and solving problems. Durkheim (1903) connected variations in knowledge to social organization, Mannheim suggested that knowledge varied by social groups ([1936]1968), and others connected knowledge to institutions (Foucault 1980, Douglas 1986). Most agree that what it means to know something varies across time and space as different kinds of thought and knowledge become possible during different historical periods and under different institutions.

When studying knowledge sociologists have proceeded along two tracks, one that addressed everyday knowledge and another focused on formal knowledge. The work of Karl Mannheim ([1936] 1968) is most often associated with the idea that different perceptions of the world are "differently formed in different social and historical settings," thus linking social forces to knowledge (238). Although he exempted mathematical and scientific knowledge from his analysis, Mannheim took a hermeneutical approach in which he analyzed the "*Weltanshauung*," the global outlook or worldview, of an entire era (1993). Through a consideration of the cultural products of a period, Mannheim argued one could begin to depict the spirit of an age. Particular manifestations of the *Weltanshauung* might vary by social group such as class or generation ([1936] 1968). The job of the sociologist of knowledge is to reconstruct the "subject's whole mode of conceiving things as determined by his historical and social setting" ([1936] 1985: 265). In addition to Mannheim's *Weltanshauung* (1993), Foucault's epistemes (1980), and the phenomenological approach of Berger and Luckmann (1967) also address the symbolic orders of actors and groups and how these orders are connected

to other social factors. Though they may take on scientific disciplines (Foucault 1980), they illustrate the "reality of everyday life" (Berger and Luckmann 1967:23) that structures the experiences and actions of entire groups and cultures.

Despite some shared origins in the work of Mannheim, a second track focused more on scientific settings. Merton (1973) was concerned with the how science worked to produce claims. However, his analysis did not extend to the very claims of science itself. Instead his work only addressed issues such as the selection processes of what to study or the mechanics of the production process. Kuhn (1964) also addressed knowledge in scientific communities. However, unlike Merton, Kuhn's concept of the paradigm and its susceptibility to social factors began to challenge the notion that scientific claims were exempt from social explanation. Kuhn argued that most science operates as *normal science*, meaning that it is not concerned with breaking or challenging the taken for granted assumptions of the field. Instead, knowledge does accumulate during this phase and operates under a given paradigm, or shared set of problems, solutions, and rules.[4] When a problem cannot be solved under the current paradigm, a revolution may occur, leading to a new paradigm (1964). In short, even formal scientific knowledge is subject to social factors.

The sociology of scientific knowledge (Bloor 1976; Barnes 1974) built upon this claim, arguing that sociologists must take a "symmetrical" approach, one in which the same mechanisms could explain how science arrived at *all* claims—both false and accurate ones. Eventually, the rejection of the idea that scientific claims were rooted in an underlying truth led to a crisis for the sociology of knowledge. It suggested that all knowledge claims are relative. If

[4] As Andersen (2001) notes, many have faulted Kuhn for a lack of clarity in his definition of the paradigm concept. In addition, Kuhn uses the concept somewhat inconsistently throughout his various publications.

this were so, sociologists could no longer claim any authority to make knowledge claims themselves (Collins and Yearly 1992; Zammito 2007; Shapin 1995).

Under this threat of relativism, the sociology of knowledge turned to a focus on practices. Instead of trying to place ideas within a causal explanation of action, the practice concept allowed sociologist to group a variety of human activity that may include skills, habituated behaviors, justifications, and tacit knowledge under a single category (Schatzki 2001). The work of the sociologists then became to identify the fields of practice which constitute social order (ibid). The practice concept also encouraged a methodological shift toward ethnography and laboratory studies. Rather than making grand claims about how knowledge is constructed, researchers turned toward a micro-level approach to understanding how specific scientific objects and claims are constructed in local and specific settings (Zammito 2007; Schatzki 2001).

Subject-Free Knowledge Cultures

Such studies remain dominant in the sociology of knowledge today. This perspective focuses on the processes for knowledge production that occur in local and specific settings (e.g., Camic et al 2011). The focus may be on articulating the way in which particular objects are constructed (e.g., Latour) or to articulations of the epistemic culture of each setting (Knorr Cetina 1999). Knorr Cetina defines epistemic cultures as "those amalgams of arrangements and mechanisms—bonded through affinity, necessity, and historical coincidence—which, in a given field, make up how we know what we know" (1999:1). She argues the practice concept takes us away from mental objects; An exploration of epistemic cultures is necessary because it will allow the analyst to incorporate the "orientations and preferences that inform a whole sequence

of actions" back into the sociological study of knowledge (9). As such, her work is concerned with bringing attention to symbols and meaning that underlie scientific work.

This kind of work in the sociology of knowledge has been incredibly important. In focusing on the activities that unfold within the lab, it has allowed scholars like Knorr Cetina and Latour (1988) to show that things are real *because* they are constructed. In trying to include symbols and meaning in her analysis, Knorr Cetina draws on the tradition of ethnomethodology (Garfinkel 1967) as she attends to the "symbolic structuring" of the lab (Knorr Cetina 1999: 11).

I follow Knorr Cetina by interrogating "how we know what we know," and through an explicit focus on the meaning systems that make such processes possible. However, I differ from Knorr Cetina in methodological approach and find her accounts of meaning to be partial. To a degree Knorr Cetina does accomplish the task of laying out the orientations that underlie action. For example, she analyzes discourse and vocabulary, showing the way in which scientific equipment used in high energy physics is endowed with human-like qualities and are thought of as either trustworthy or untrustworthy. However, her accounts lack the inclusion of human motives and fail to convey the scientists in her study as agentic actors with motivations and interior lives. In contrast, I draw more explicitly on hermeneutical tradition in an effort to see the world as others see it. As I will argue below, this interpretive approach is necessary in order to appreciate the emerging epistemic authority of data science and its capacity to exert effects on social processes and social structure. In an effort to demonstrate the lack of subjective accounts in the current sociology of knowledge, I provide brief assessment of Knorr Cetina's analysis below before moving on to describe my approach.

In her seminal work, *Epistemic Cultures* (2011), Knorr Cetina utilizes a comparative ethnography between experimental high energy physics and molecular biology. Though she situates herself in the tradition of laboratory studies (e.g., Latour and Woolgar 1986), she claims to be looking at something different. Instead of trying figure out how particular knowledge is constructed, she is concerned articulating the "epistemic machinery" through which science is made (3). This is, as she stresses, a cultural project. For Knorr Cetina, culture is the "aggregate patterns and dynamics that are on display in expert practice and that vary in different settings of expertise" (8).

She accesses this culture through a focus on the organizations and systems that produce science with less attention given to the actors that make up these scientific communities. This is an intentional aspect of her analytical approach to studying the systems and machineries of knowledge settings. Consistent with knowledge studies from Actor Network Theory (Latour and Woolgar 1986), Knorr Cetina rejects the automatic application of native categories in her analysis. This leads her to ask who are the "epistemic subjects in the laboratory?," rather than assume scientists as the actors (127). She notes that placing agency with the humans in the room is not necessarily wrong, but is "too limited when it comes to determining the cultural parts human entities play in the reconfigurations of self-other things with which I have associated laboratories" (127). As a result, the scientists in her study are rarely depicted as agentic actors with an inner life. This choice is reflected in her style of writing as much, or if not more so, than in the actual claims of the text. In her sentences, non-human phenomena are often depicted as the subjects of sentences and events are described in the passive voice. This is not to say that sentences in which people are the subjects are absent altogether, but they are much less common. Below I provide a few examples:

- "For example, the experienced body of the scientist, when it operates, naturally brings its experience to bear on the variations it concocts for selection by success" (109).
- "In this book, symbolic structuring will come into view through systems of classification, through the ways in which epistemic strategy, empirical procedure, and social collaboration are understood in the two fields investigated" (11).
- "Talk that recalls these factors, as we have seen, fills in the question marks in the test tube reactions" (109).
- "the detector is construed not as a mechanical or electronic device, but as a physiological organism" (136).

We can imagine the same analytical insights being expressed in slightly different terms. Below, I have created a table that points to the way in which the above sentences draw attention away from the human meaning-makers from which this analysis emanates. In the second column I identify why these quotes have such an effect. In the third column, I show that similar analyses could be expressed in ways that bring the attention's reader to the meaning-makers involved.

Original Sentence	**How Agency or Actors are Hidden**	**Reformulation**
"For example, the experienced body of the scientist, when it operates, naturally brings its experience to bear on the variations it concocts for selection by success" (109).	"The body" and "the scientist" are treated as separate entities. "The body" is the subject of the sentence that does the acting.	For example, when *scientists* draw on their bodily experience, *they* naturally bring *their* experience to bear on the variations *they* concoct for selection by success.
"In this book, symbolic structuring will come into view through systems of classification, through the ways in which epistemic strategy, empirical procedure, and social collaboration are understood in the two fields investigated" (11).	Passive voice obscures who or what does the understanding. With who/what the symbolic structuring lies is unclear.	In this book, symbolic structuring will come into view through systems of classification, through the ways in which epistemic strategy, empirical procedure, and social collaboration are understood in *by the professionals* in the two fields investigated.
"Talk that recalls these factors, as we have seen, fills in the question marks in the test tube reactions" (109).	"Talk" is the subject of the sentence and does the acting.	When *scientists* recall these factors, as we have seen, *they* fill in the question marks in the test tube reactions.
"the detector is construed not as a mechanical or electronic device, but as a physiological organism" (136).	Passive voice obscures who/what construes the detector. Where this cultural symbolism lies is unclear.	*Scientists construe* the detector not as a mechanical or electronic device, but as a physiological organism.

As I mentioned above, sentences in which the scientists are the actors or are the ones who believe in these symbolic aspects of epistemic cultures are not altogether absent from Knorr Cetina's writing. However, in reading her work, the reader does get a sense that culture, or at least epistemic cultures, somehow operate outside of actors who embody, imagine, call forth, and believe in this culture. Given that anyone who encounters her research is herself a meaning-making being who works to create meaning from a text, the presentation of Knorr's Cetina's analysis is not simply an inconsequential stylistic choice. We do not see the world as the actors in these contexts see it. While this approach may intentionally follow from her treatment of culture as a system and from a tradition in which networks or facts—rather than people—are the object of study (Latour and Woolgar 1986), the language with which the reader encounters the concept of epistemic cultures works to flatten the presence of human meaning-makers into the mire of a social world free of subjectivities.

I believe that this aspect of Knorr Cetina's treatment of culture stems from an intentional effort to treat all elements of the laboratory—the human and the material—from an even footing. If one approaches a human with the same observational techniques as one approaches microbes (Latour 1988) or lab equipment (Knorr Cetina 1999), there is no way to access deeper meaning structures. The researcher is left with the observation of discourse and behavior. Though they are important and useful aspects of sociological analysis, documentation of these phenomena alone cannot fully account for the meaning which structures human claims, understanding, and action.

Thus, although this style of sociological analysis claims to be dealing with meaning and culture and to take ethnomethodological approach, I find that a key element is missing: the subjectivities of actors. Though Garfinkel's (1967) ethnomethodology could appear to share this

feature in that he focuses on "practical activities" and on "observable-and-reportable" aspects of human interaction, he is also interested in the assumptions that are communicated but not said. In addition, in his empirical work, such as the study of Agnes, he gives a great deal of attention to her biography and the ways in which she understands the world—her worldview—and how this helps to explain her feelings and decisions. When developing sociological explanations, the meaning-making of humans matter. I find the current approach which tries to move away from the "native" categories of a case (Latour) and casts doubt on the status of human actors to be somewhat wrong-headed. When treating laboratory equipment by the same terms as human subjects, the sociologists neglects the fact that, though constructed, human-generated categories are *real* for the sites we study (Bowker 2014). In other words, to some degree, the fact that people think of themselves *as* people will influence how social patterns unfold in a given location. For this reason, I move away from the approach of Knorr Cetina and others that assumes humans (and the meaning-making capacities that they possess) are not central to sociological accounts. Given that these deeper structures make certain lines of thought and action possible, they are a key part of understanding how culture underlies the construction of knowledge (Garfinkel 1967; Abend 2014).

An Interpretive Approach to Knowledge Cultures

Instead of focusing on practices, my approach to the sociological analysis of knowledge relies upon the interpretive mode (Reed 2011). This means that the analyst endeavors to reconstruct the meanings "upon which social life proceeds" and which serve as the basis of an actor's "subjectivities and strategies" (ibid:110). Drawing upon Reed's concept of reconstructing "landscapes of meaning," my analysis relies upon the depiction of *epistemological landscapes* (ibid:109). Epistemological landscapes are the meanings and worldviews through

which actors see data science as the appropriate tool and method for producing legitimate knowledge claims and solving problems. The epistemological landscape includes beliefs about the purpose of science, about the nature of reality, about the human capacity to recognize that reality, and about valid ways to know the truth. As Reed argues (ibid), this approach to sociological explanation—of reconstructing meanings—requires a multitude of theoretical insights and the application of various theoretical tools as one works to bring a particular "landscape of meaning" into light (ibid:109).

When it comes to trying to understand the cultural frameworks that shape motivations, beliefs, and actions, the sociologist has a number of tools at her disposal. Many have noted the role of institutions in shaping our thinking and actions (Durkheim and Mauss [1903] 1963; Douglas 1986; Foucault 1980). I follow Foucault (1980) in treating the patterns of thought associated with institutions, not as ideological misunderstandings of the world, but as the productive discourses through which truths, meaning, and sense-making come about. Much as Foucault is interested in the kind of subjects that get produced through certain discourses, I ask what definitions of truth and reality inhere in the discourse and practices of data science.

In dealing with settings in which institutions are in flux or in which multiple intuitional frameworks may be present, I find Boltanski and Thevenot's (2006) recent theoretical framework particularly useful. In their empirical work, they identify not a single logic that guides thought and action, but a set of what they call *orders of worth*. Within each order, effect justifications must conform to acceptable forms of logic. Each order of worth possesses corresponding criteria used to assess justification and legitimacy. For instance, in an industrial order of worth, actions are evaluated based on the efficiency and productivity (mode of evaluation). Actors may draw upon multiple *orders of worth*, which overlap in social space,

despite being institutionally distinct. Different kinds of reasons are acceptable in different institutional realms. Importantly, a single individual may use many different orders of worth without conceiving of them as being in conflict, and individuals may strengthen their criticism or justification of an act by drawing together the evaluative criteria from more than one order. I use these insights to sensitize my analysis to the fact that, although I aim to locate the epistemological landscape that unfolds alongside data science, other ways of ordering the world may be present in the locations I study, they may influence both how actors feel about and approach their work in data science, and they may structure how data science is employed in practice.

With regard to the institution of science specifically, Daston and Galison (2010) have shown the ways in which the criteria by which scientific claims are produced have varied over time. Starting in the 18th century, they identify three "epistemic virtues," a set of values that are "preached and practiced in order to know the world. As I will discuss in chapter 3, the virtue they call "mechanical objectivity" most closely resembles the perspective of the data scientist interviewed and observed for this study. In bringing the insights from Daston and Galison into this study, it is useful to note their observation that epistemic virtues are never fully replaced by the subsequent virtues. Instead, "epistemic virtues do not replace one another like a succession of kings. Rather, they accumulate into a repertoire of possible forms of knowing" (113). This idea of successive orientations, virtues, and worldviews associated with producing knowledge should alert the researcher to the possible presence of multiple "virtues" in the settings that rely upon data science to produce claims.

The Epistemological Landscapes of the Knowledge Society

It is on these terms that I deal with the question of the knowledge society. In doing so, this project also builds toward a larger research agenda aimed at comprehending how knowledge and culture are constituted in modernity. Many have noted that knowledge is increasingly becoming the chief organizing principle in modern society (Bell 1973; Böhme and Stehr 1986; Castells 2000; Thrift 2005; Sennett 2006; Knorr Cetina 2007). Though I push back against their methodological approach, sociologists of knowledge have begun to make valuable contribution toward articulating the contours of such a society. In this knowledge society, knowledge becomes the "productive force that replaces capital, labour and natural resources as central value- and wealth-creating factors" (Knorr Cetina 2007:361). As Knorr Cetina argues, to claim that we are operating in a knowledge society does not simply mean that society is organized around knowledge. It also indicates that ours is a society, "permeated with knowledge settings, the whole sets of arrangements, processes, and principles that serve knowledge and unfold with its articulation" (Knorr Cetina 2007:361-2). This suggests that not only the economic relations are altered, but that the prevalence of knowledge and knowledge settings will have vast cultural consequences as well. Sociologists need to turn attention toward the many settings, both outside and within the traditional knowledge communities of academia, where knowledge production occurs.

The study of data science allows for the exploration of epistemological landscapes that cut across a variety of knowledge settings. As such, one contribution of this project is that I consider knowledge in non-laboratory, non-formal settings. These include both for-profit and consulting settings in which many of the data scientists interviewed for this study work and the medical setting of the neonatal intensive care unit in which I study the application of data

science. In addition, as data and the structures and techniques that accompany it become the increasingly ubiquitous means by which such knowledge settings operate, the epistemological landscape of the knowledge society may be ever more informed by the landscape that unfolds with data science. Therefore, in looking at the texture and interplay of epistemological landscapes within these contexts, I make an effort to unpack what it means to live in a knowledge society. This is a theme which I will return to more fully in the conclusion.

Data and Methods

My analytical strategy is focused on the collection and analysis of data science's epistemological landscape from a variety of angles. I do not attempt to emulate the model of the natural sciences by producing representative samples or mimicking the structure of experiments. Instead, my analysis is based upon data that allows me to depict the epistemological landscape of data science practitioners, of the users of data science, and of the public discourse that surrounds data.

To address practitioners, I interviewed 28 data scientists and conducted ethnographic observations of a medical analytics team. Interviews were conducted in 2015 and 2016 in person, by phone, and by Skype.[5] Respondents were identified primarily through snowball sampling. They range in age from 23 to 65 and work primarily in cities located in the Northeast, Mid-Atlantic, South, South-West, and the West Coast. All are college educated, 26 have advanced

[5] Not all the respondents included in the interview had "data scientist" in their official job title. This is due in part to the early stages of this emerging profession. When recruiting for this study, I focused on the kind of work conducted by data scientists, looking for professionals whose work involves 1) developing predictive analytics, 2) developing software or platforms that allow others to generate knowledge or predictions, or 3) analyzing large data sets to develop models for how systems, organizations, environments, or people behave.

graduate degrees, 20 are male, 8 are female, and 26 are white. In addition to the interview study, for 12 months between 2015 and 2016, I spent time in the offices of data science professionals and observed weekly meetings of a data science team working to build medical analytics and algorithms. These meetings varied from 9 to 25 individuals and included data scientists, mathematicians, computer scientists, and medical clinicians. Meeting discussions included brainstorming for new projects, sharing problems and progress on current projects, and strategizing ways in which to convince others of the value of medical analytics.

To explore the ways in which non-data scientists employ the products of data science in decision-making and knowledge production, I conducted an ethnography of a neonatal intensive care unit (NICU) that uses data-driven predictive algorithms to determine when an infant is likely to develop an infection. In total, I spent 60 hours conducting observations in the NICU. Approximately 35 attending physicians, fellows, medical students, nurse practitioners, and nurses, were present during observations. Observations consisted of shadowing physicians as they conducted rounds, shadowing nurses throughout shift work, and observing the unit as a whole from the nurses' station. In addition to many conversations in the field, I also conducted in-depth interviews with 11 clinicians. After completing an initial analysis of my field notes nad interviews, I was granted access to an independent data set of interviews with Horizon users (Robert H. Tai Research Group [RHTRG] 2012). I used this data to confirm the patterns indicated by my own data collection. Following observations of the medical analytics team and the NICU, I wrote detailed field notes. Field notes and transcripts were analyzed using an open-coding process (Charmaz 2006).

To access the public discourse surrounding data, I conducted a content analysis of 33 business-to-business white papers that address data analytics. Business-to-business white papers

often function as a type of advertising for data science products and tend to avoid overly technical language. They capture the ways in which data science is often presented to non-specialists and potential clients. I collected these papers through a variety of methods designed to mimic the ways in which potential consumers of data analytics might go about learning about data science and potential products available to them. I began by identifying leading and prominent companies that produce data analytics platforms. I identified these companies by using the Gartner[6] list of top analytics platforms for 2014 (Herschel, Linden, & Kart) and 2016 (Kart, Herschel, Linden, & Hare). I also asked my interviewees to identify companies and products that they track. I then visited the websites of these companies and downloaded available whitepapers that both address data analytics and were written primarily in non-technical language. In addition, I searched LinkedIn for white papers that fit similar criteria. This resulted in a corpus of 33 white papers from 24 different companies and organizations.[7]

I use the data describe above to reconstruct the *epistemological landscapes* of data science. Where Garfinkel (1967) relies on breeches of shared understandings to locate deeper meanings, my approach to constructing these epistemological landscapes is a hermeneutic one. I treat the conversations that I witnessed, interview responses, and statements from data science organizations as texts that both arise out of and make possible specific ways of understanding the world (Geertz 1973).

In order to produce a close reading of the texts, I follow Pugh's (2013) suggestion that such texts contain various kinds of information. As she describes, "a fundamental characteristic of in-depth interviews: they can access different levels of information about people's motivation,

[6] Gartner Inc. is a research and consulting company that specializes in information technology. Their reports are well-recognized within the information technology industry for assessing the industry as a whole.
[7] A full list of white papers and their source organizations is included in Appendix B.

beliefs, meanings, feelings and practices – in other words, the culture they use – often in the same sitting." (Pugh 2013:50). In analyzing interviews, field notes, and even written documents in this way, the analyst does not necessarily treat statements as evidence to back up particular claims. Pugh suggests that in-depth interviews contain at least four kinds of information: *the honorable*, *the schematic*, *the visceral*, and *the meta-feelings*. Both honorable and the schematic information are particularly useful for constructing the epistemological landscape. Honorable information includes belief statements and explanations intended to paint the respondent in a favorable light. This allows the researcher to get at the culture codes that suggest appropriate and justifiable beliefs and actions. In the case of data science, this assists in understanding the role of the data scientists within the epistemological landscape. Schematic information is communicated through "metaphors, jokes, turns of phrase and discursive innovations" (ibid). These features of the text allow the researcher to ascertain how the respondent sees the world. For this study, the analysis of these features and how people employ them allows for the reconstruction of what constitutes truth and legitimate claims within the landscape.

Outline of the Chapters

Chapters 2 and 3 deal with the epistemological landscape of data science from the perspective of data scientists. In chapter 2, I focus on the source of epistemic authority within the epistemological landscape. In other words, I ask what processes or conditions do data scientists see as legitimate means to making truth claims. Considering epistemic authority helps us to make sense of the ways in which data scientists approach "domain expertise," or knowledge based in a particular subject or area, rather than the techniques and processes

associated with data science. Given that domain experts are often the producers of theory, the placement of domain experts relates directly to the end-of-theory debate. Rather than engage with this debate directly, I offer an empirical and cultural account of these tensions, showing how such questions are answered when placed upon the epistemological landscape of data scientists. I find that, although data scientists still proclaim that domain experts are central to the process of knowledge production, they give little epistemic authority to these kinds of experts. Instead, the source of authority has shifted from the experience and accumulated knowledge of domain experts to the technical skills of data science. I suggest that the continued valued presence of domain experts in the epistemological landscape results from the overlapping institutions in which data scientists locate themselves—namely, an emerging identity as "data scientists" who draw on a particular narrative in which more "advanced" scientific techniques lead to better knowledge and the tradition of various disciplines which rely upon accumulated knowledge, such as physics or genetics.

In chapter 3, I continue to reconstruct the epistemological landscape of data science, focusing on the ways in which data scientists understand their role as scientists and their work as part of a larger project to better the world. When evaluating their work, I show that data scientists are caught between two worlds—one ideal and one pragmatic. As part of the ideal image, data scientists rely upon scientific techniques, methods, and criteria when evaluating their work. It is through an adherence to these trained skills that they ensure that their resulting knowledge claims are "right." However, data scientists also espouse an alternative set of criteria, a pragmatic approach in which quality is evaluated based upon the usefulness of an analysis. Under this pragmatic approach, data scientists sometimes accept work that is scientifically flawed because it provides some sort of improved outcomes for their task at hand, casting aside

concerns about epistemic authority. I argue that the pragmatic approach to data science raises questions about the legitimacy of knowledge claims, even when evaluated by the tenets of data science.

In chapter 4, I shift focus to the discourse of data, or what I call "data talk." Drawing on literary theories of metaphor, I argue that the words of data scientists and of data-driven organizations set the possible terms by which non-specialists and the public come to construct their own epistemological landscape of data science. I show that the metaphors contained in data talk depict an epistemological landscape in which truth and knowledge lie in the details of the data itself. This differs somewhat from the more expansive approach to epistemic authority taken by data scientists in which they focus on techniques and proper execution (which includes processes of data collection and creation of databases). Nevertheless, it creates a strong association between the individual data points contained in databases and truth. In this way, data science is still seen as containing the power to overturn and correct the knowledge claims of domain experts. In addition to dealing with epistemic authority, this chapter points out additional aspects of "data talk," specifically the ways in which this language associates data with objectivity and makes certain notions of data ownership seem natural and expected.

In chapter 5, the analysis moves to a setting in which the products of data scientists, predictive algorithms, contribute to the construction of knowledge and decision-making. As I have argued, many data science products are implemented in professional settings with their own cultures, worldviews, and values. Understanding how data science integrates with these settings is an important part of assessing the impact of data on society. Drawing upon ethnography and interviews conducted in a neonatal intensive care unit (NICU), I ask how clinicians use a data-driven predictive monitoring system called Horizon to determine if a patient is sick and to decide

whether or not to treat that patient. I find that clinicians draw on two attitudes toward epistemic authority when integrating Horizon into their assessment of a patient. First, clinicians draw most explicitly upon the tenets of evidence-based medicine. Second, though they are less likely to articulate it explicitly, clinicians espouse a great deal of value for the role of experience in determining a patient's condition. It is in a negotiation between experience and evidence-based medicine that clinicians make sense of Horizon. Contrary to fears that data science will undermine alternative forms of knowledge, I find that in this negotiation, Horizon's quantitative assessment may sometimes work to bolster the conclusions suggested by experiential knowledge, ensuring that they remain part of the knowledge production process in the NICU. However, I caution that this use of data analytics is only possible under particular organizational contexts, specifically those that allow knowledge workers to assess the outcomes of data analytics for themselves (rather than requiring automated responses) and those that encourage the development of experience with actual outcomes, and not just those predicted by data science.

In the conclusion, I draw together the themes that have cut through the empirical material of this project. First, I argue that in unpacking the epistemological landscape of data science, we can better understand the grounds on which action unfolds. This is especially apparent in chapters 4 and 5. The epistemological landscapes depicted in data talk and in the context of the NICU show how such symbolic orders clear the way for certain lines of action, making them seem more straightforward or less controversial. This work addresses both the dearth of ethnographic accounts of data science, algorithms, and big data within critical data studies and broadens the scope of the sociology of knowledge to once again include subjective experiences. This perspective, then, is especially helpful for developing a critique of data science and for suggesting ways in which a cultural assessment of science can facilitate policy recommendations

or regulations that allow us to benefit from the capabilities of data science without suffering from some of its deleterious effects.

In addition, I address to the question of the knowledge society. In particular, I suggest that returning the hermeneutic sociology of knowledge suggested by Mannheim may be a way forward in constructing the meaning landscape of the knowledge society. Under this theoretical framework, data science is one of the manifestations of the culture of the knowledge society. Accounting for the meanings that unfold within and encourage the use of data therefore provide a point of access into the symbolic order of the knowledge society.

CHAPTER 2

Epistemic Authority in the Land of Data Science

In this chapter and the following, I explore the epistemological landscape of data scientists. Central to depicting this landscape is the location of authority. From where or from what can data scientists derive the authority to justify their knowledge claims? I show that, for most, the authority to make a claim is derived from the processes associated with data science. To highlight this authority, I focus on the tension between *domain expertise*—expertise in a particular area that is usually accumulated by an individual or group over many years of experience and research—and the technical processes of data-driven analysis.

Rather than jump immediately into this debate, I begin by depicting some of the aspects of the epistemological landscape through which data scientists make sense of domain expertise and their own work. To do this, I draw on interviews, observations, and public discourse to show the many ways in which data science culture is consistent with broader cultural trends of western modernity—namely the suspicion of subjectivity and a trust in numbers and process to overcome this human shortcoming. As I argue, the methods of data science, including processes of both data collection and analysis, seem to promise a renewed ability to bypass the shortcomings of human subjectivity. I then return to the issue of domain expertise. In the views of data scientists, domain experts are both valued and associated with dangers of subjectivity. However, when faced with a conflict between domain expertise and their own findings, data scientist often reference the processes and techniques of data science as a means to settle the discrepancy and assign authority. In short, although data scientists espouse a reverence for domain experts and their theories, authority to make a knowledge claim is ultimately derived

from the data and techniques of data science. This perspective is somewhat problematic given the opportunities inherent within data science to introduce its own sources of bias and errors and, as I will show in the following chapter, the tendency of data scientists to accept flawed analysis in applied settings. Understanding where data scientists place epistemic authority is an important aspect of beginning to trace out how the expansion and application of data science will influence the role of expertise in knowledge production and problem solving (i.e., are the theories of domain experts built into the solutions and knowledge produced by data science) and in our cultural understanding of epistemic authority (i.e., what symbolic systems render some actors capable of making claims that are taken to be legitimate).

Approaches to Expertise and Authority

The problem of expertise and the authority to make knowledge claims has been tackled from several different angles in the sociology and science and technology literature. Each approach is concerned, in some manner, with the relationship between expertise and power. In the first approach, experts are part of the rationalization process of modernity and come to specialize in particular fields (Weber [1952] 1991). In this view,

> An expert's knowledge includes specific, technical skill based on some wider appreciation of the field of knowledge in question. In academic areas, we say that someone "knows the literature," that is, knows the debates and the questions relevant to the use of the specialized knowledge at hand. The expert's knowledge is rooted in a body of knowledge well enough codified to be passed on through formal training. Expertise grows as well, over time, from clinical experience (Schudson 2006: 499).

As this definition suggests, the kinds of experts envisioned here are "domain experts" in that their expertise constitutes accumulated knowledge and experience of a particular topic. As

Schudson (ibid) notes, this expertise is not necessarily accurate or direct knowledge of a topic, but is socially constructed. Experts wield the power to make knowledge claims because society accepts claims from particular professional groups as legitimate.

Even though this perspective on experts recognizes that experts approach their work with a set of presuppositions or assumptions (Weber [1952] 1991) and that their knowledge does not necessarily reflect progress toward more and more accurate truths (Kuhn 1964), experts are important part of modern society. Both Weber and Schudson are concerned that without experts, the state will exert too much power over the public. Experts do not by default operate in service of the public. Much in the way that Medvetz (2012) argues politicians can now "shop" for expertise in the services and portfolios of think tanks, experts may be enlisted in the service of the state. However, with the proper conditions, experts also have the potential to "speak truth to power," clarify debates, and identify injustices and opportunities. For them to function in this way, experts need to occupy positions in which they neither fear the establishment nor defer to it (Schudson 2006:500). This provides experts with the potential to challenge and temper the power of the state.

A second view, manifested primarily in science and technology studies, is also concerned with power and expertise. Expertise in this research area is seen as both social and performative: "an expert has the tacit, social, and cultural knowledge needed for the performance of expertise" (Evans and Collins 2008:610). In addition, such expertise may not reflect direct access to truth; through the process of inclusion, or the socialization to a particular way of seeing the world that are associated with various professions or groups (Bijker 1995), experts may become blind to certain aspects of reality. Thus, they may be less capable of innovation. Like the first approach, empirical work in this area suggests that expertise is associated with particular topics and subject

areas; these experts are medical researchers (Epstein 1996), chemists, (Bijker 1995), or physicists (Wynne 2004).

Contrary to Weber and Schudson, work from science and technology studies approaches expertise primarily as a threat to democracy that stems from the very nature of expertise (Turner 2003, Evans and Collins 2008, Sismondo 2008). On the one hand, the power associated with expertise may be derived from preserving and defining a boundary between groups or between science and non-science (Gieryn & Figert 1986, Gieryn 1994).[8] This boundary work is "driven by a social interest in claiming, expanding, protecting, monopolizing, usurping, denying, or restricting the cognitive authority of science" (Gieryn 1994: 13). On the other hand, work in science and technology studies tends to treat expertise as something real that can be acquired (Epstein 1996). If this is the case, the power of expertise results from the inequalities between those who have access to the machinery of science and those who do not. Either way, this is problematic for a governing structure, such as democracy, predicated on the participation of all voices in society. As Turner (2001) points out, expertise is either a form of ideology or it treats the opinions of a particular group as privileged.

To deal with these problems, a number of solutions have been proposed. Collins and Evans (2002) take a normative approach and attempt to identify who should be involved in technical decision-making, given that experts are both useful and have a privileged position. As part of this effort, they identify different kinds of expertise. This move allows them to make the political claim that non-scientist citizens, too, bring valuable insights to debates and decision-making. In addition, the ways in which non-scientists have cultivated expertise has also been

[8] This approach to cognitive authority parallels developments within the sociology of professions, where legitimacy and authority are explained through the interested moves of groups as the work to professionalize themselves by acquiring jurisdictional rights or establishing a monopoly over particular activities or (Abbott 1988).

explored (Epstein 1996). Several (Turner 2003, Collins and Pinch 1996) have argued that by including the public to some degree—either through education or direct participation—democratic societies can avoid the potential threat of technocracy. In teaching the public to see science as a skill, not as knowledge, people will see that it can be fallible and therefore be more likely to challenge the authority of science (Turner 2001).

Despite their shared interest in the role of experts in society, both of these approaches have failed to deal with the deep symbolic orders by which experts and those that believe in their expertise derive the legitimacy of their knowledge claims. Instead, current approaches treat experts as interested actors who may either abuse the power derived from their expertise or use it for the good of the public. This neglect of the symbolic order may have led studies of experts to miss an important empirical shift. Though they differ in their exact treatment of the expert, both approaches recognize that experts are associated with knowledge in differing domains or realms. They approach the expert primarily as a domain expert. However, in a world increasingly infused with knowledge settings that integrate or rely upon data science, algorithms, and big data, this may no longer be a sufficient way to think about the expert. As data science is increasingly applied to solve problems and generate knowledge areas as diverse as medicine, policing, urban planning, or national defense, the kind of expertise involved in making claims and directing action changes. Rather than domain experts, the technical expertise of data scientists begins to claim the *epistemic authority* to direct action. I use the term epistemic authority to point to the differing people, practices, or objects that are viewed as possessing legitimate access to truth. In order to understand or advocate for particular relationships between society and experts, this new kind of expertise and its symbolic sources of power and legitimacy must be addressed. This requires inquiry into the underlying worldview that allows experts to

claim that their techniques and training lead to truths and to the power that inheres in these practices.[9]

Bringing Experts into Critical Data Studies

In the introduction, I reviewed the emerging literature from critical data studies. Although this body of work does not deal explicitly with expertise, one senses a concern that some kinds of expertise—those which are centered in a particular area or topic—are facing an eclipse by quantitative, technical expertise (boyd and Crawford 2012, Seaver 2015). In writing about the digital humanities Berry points to the way in which data science may circumvent domain expertise, arguing that "technology enables access to the databanks of human knowledge from anywhere, disregarding and bypassing the traditional gatekeepers of knowledge in the state, the universities and the market" (2011: 8). boyd and Crawford argue that the faith in numbers exhibited by the advocates for big data not only indicates a "dismissal of all other theories and disciplines," it also "reveals an arrogant undercurrent in many Big Data debates where other forms of analysis are too easily sidelined" (2012:666). This is more than strategic maneuvering by professional groups. Instead, it points to the shifting and conflicting symbolic locations of truth that are contained in the epistemological landscape of data science. Critical data studies, greatly influenced by Foucault (1980), approach data science with a recognition of the way in which power and knowledge are intertwined. Particular ways of seeing the world suggest differing claims to expertise.

[9] Even when Turner (2001) argues that experts' authority rests in different sources, he attributes this to differing relationships between an expert and her audience, rather than to symbolic orders. Turner also discusses Foucault's treatment of ideology and links this to expert authority, but he focuses on the way in which audiences are misled by such claims rather than stressing the symbolic order to which the experts come to see their own work and claims as authoritative.

To deal with this aspect of expertise, the symbolic ordering that makes certain claims to epistemic authority sensible, we need to attend to the makeup of the epistemological landscape that supports experts. Although I return briefly to the issue of experts and democracy at the end of this chapter, my main objective is to shift the conversation on expertise from a focus primarily on their social status and role in society to a consideration of the symbolic ordering that make claims to expertise possible. It is clear that different cultures and groups tend to locate epistemic authority differently. In her work on civil epistemologies, Jasanoff (2005) uses a cross-national comparison of the United States, Britain, and Germany to show how, at the level of the state, expertise operates differently in the national debates surrounding biotechnology. While she observes that expertise in the United States is associated with formal methods and professional skills, in Britain it is associated with experience. Essentially then, the epistemic authority of these experts comes from different sources.

Drawing on interviews with data scientists and observations of a medical analytics team, I begin to trace out the contours of the epistemological landscape of data science in the following pages. In doing so, I describe the symbolic ordering that allows data scientists the authority to make truth claims.

Contours of an Epistemological Landscape

Big Data Tools Surpass Human Capabilities

Much like the public discourse discussed in the introduction, the ability for data science to fill in gaps in human capabilities, outpace human capabilities, or counteract human bias and blind spots are frequent viewpoints in the epistemological landscape of data scientists. One of

the most common assertions is simply that there is too much data out there, and it is either impossible for humans to process entirely or would take too long to do so. Chris, a lead data scientist for the marketing team at a consulting firm commented that, "The main motivation I think for why someone like [our clients are] interested in it is because historically they have more data than they know how to handle, just as humans looking at it. If you have a dozen different data sets it's hard to make any holistic conclusions about what's going on in the marketplace, or what they should do, or what actions they need to take." Similarly, Jim, a data scientist who consults in the defense industry, told me that:

> The problem is that people have—at least this is what I see as being like *the* fundamental thing about data science, is that like you've got crap tons of data. And it's all a mess, and you really don't know what the heck's going on. Maybe you got it from multiple data sources, and so like there's very little overlap. But you wanna try to take all of it and actually extract information that you can use from it.

In contrasting the techniques of machine learning with human analysis, Adam, who is an analyst and the president of a company specializing in biomedical data, said that, "So the machine learning engine- I mean, it's just- the complexity is way too high for a human brain to put together the patterns on that." Even in areas where humans have traditionally tackled large scale or synthetic analysis, there is a sense that data analytics can accomplish the tasks more easily. As Kieran, a senior data scientist at a financial consulting firm, explained, there are:

> teams of human analysts who are experienced in a certain sector or whatever, and they'll just be doing their expert analyst forecasting. Oftentimes, we find that we end up being able to do those things better and more easily with machine learning. Sometimes we'll create something that's not new for the company. Sometimes we will replace an existing process with a data science or machine learning based one that we feel is better suited to our customers' needs. […] The fact that we have a machine learning model that is now able to produce these basic forecasts frees up

> our human resources to go do potentially more interesting work or more valuable work.

This use of data science to "free" human capacities for other problems is a common application. Taylor, who works as the lead engineer developer to design data-driven platforms for the defense and healthcare industries, described his work in the following way:

> "I think what we do that's different is we make---we really try to make knowledge accessible rather than just trying to gather data and mine data and leave it up to the human to make all the inference. We try to augment the human decision maker most effectively. So we want the user to be able to free up cycles, free up brain cells for thinking about harder problems, for providing more deeper inference than what they can give by actually having to go through all the steps themselves manually.

Even though both Kieran and Taylor continue to express value in human capabilities, their statements still reveal a view that human approaches are inefficient at certain tasks. Adam's comment also hints that human may not even be capable of producing some of the insights that can be achieved through data science. The data scientists with whom I spoke espoused a belief that the techniques associated with data science were able to enhance the human ability to process data by increasing the amount of data that could be considered and improve the processes for making sense of that data by making them either more efficient or more accurate.

Seeking Objectivity: Human Experience Need Not Apply

In addition, there is a concern that human bias, subjectivity, and blind spots can be roadblocks to new knowledge or hinder the research process. These human tendencies may make humans unreliable sources of information, lead people to discount insights from the data, or cause errors to occur in data collection Assumptions about the value of objectivity in knowledge production are so deeply seeded in our culture—not just the culture of data science—

that they can be hard to see. Daston and Galison (2010) trace the modern concept of objectivity to the middle of the 19th century. It was at this point that scientists began to see themselves as potential dangers to scientific knowledge and discovery. As they describe, "their fear was that the subjective self was prone to prettify, idealize, and in the worst case, regularize observations to fit theoretical explanations: to see what it hoped to see" (ibid: 34). Since that time, objectivity has become so tightly coupled to our notion of science, that we often see them as going hand in hand, despite the presence of alternative epistemic virtues that have characterized science at different times. In addition, the value of objectivity has become widespread across research and professional areas. Social scientific research has continued to cast doubt on people's ability to properly assess the world around them or even to accurately describe their own experiences (Dean and Whyte 1958, Golden 1992) and motivations (Vaisey 2009). Such concerns about subjectivity are embedded in the popular imagination as well. Although a 2013 article in *The Atlantic* ominously titled, "They're Watching You at Work," expresses some concerns about allowing machines to make human resources and hiring decisions, the author acquiesces to a preference for these risks over the risk of subjectivity. He writes:

> Should job candidates be ranked by what their Web habits say about them? Should the "data signature" of natural leaders play a role in promotion? These are all live questions today, and they prompt heavy concerns: that we will cede one of the most subtle and human of skills, the evaluation of the gifts and promise of other people, to machines; that the models will get it wrong; that some people will never get a shot in the new workforce. It's natural to worry about such things. But consider the alternative. A mountain of scholarly literature has shown that the intuitive way we now judge professional potential is rife with snap judgments and hidden biases, rooted in our upbringing or in deep neurological connections that doubtless served us well on the savanna but would seem to have less bearing on the world of work (Peck 2013).

Here, the author not only suggests that numbers may be safer than any undesirable consequences that result from the employment of data science, but he indicates that the danger of subjectivity is an inherent aspect of human biology and evolution, suggesting that this human flaw is unlikely to be overcome by training or effort.

The widespread assumptions about this aspect of knowledge and human obstacles to ascertaining it means that interviewees rarely articulated this view overtly. It was assumed that the need for objectivity was a taken for granted value that I shared with them, therefore needing little mention or discussion. Nevertheless, a close reading of their accounts shows this discomfort with subjective knowledge.

Adam recounted his frequent experience with doctors and genetic experts who would focus on genes that they study and with which they are already familiar: "When we first started working out at the National Cancer Institute, we would crunch this data, and we would go to some of the scientists working there that were, you know, that had been there for thirty years. They didn't want to see of all of [the data]. They just wanted to see their favorite gene and what it looked like, right, and that's what they want to work on." In this example, Adam expressed frustration that in expressing a subjective preference for some data over others, researchers were potentially missing out on important genetic data and insights.

While Adam's experience has made him wary of intentional bias, some data scientists expressed a concern at having humans overly involved in the production of knowledge in general. At the time of our interview, Isaiah was a recent graduate of a masters in data science program now working in cybersecurity. In his discussion of his current projects in reveal a hesitancy to trust human-reported data:

> So yeah, they've tried to impose a sort of structure to this, which is I think a step in the right direction. Uh, but at the same time, all of these things are what an analyst has already thought of. So, it's weird data, it's not like a nice, uh, rows and columns sort of data set. It's not *a* data set, it's data. It's raw, and-and most cases it's like the second tier of data. Which is--sort of—it's like—so an analyst has already looked at IP addresses or like some event, and then he reports this event using this system. So, this data is actually, uh, just capturing the information that this analyst captured. Uh, so it's like a weird second layer almost, and again, it's almost-it's almost meta-data at this point.

In our interview, Isaiah continued to express discomfort with conducting analysis on any data that has already been touched by human interaction. After telling me that "I would prefer more quantitative data basically," I asked Isaiah if he thinks that quantitative data is higher quality. He responded that,

> I think so, yeah. Cause then it's just math at that point. And it's- you're not introducing humans anymore. I-if it's just pure numbers then, I mean, assuming the humans that collected those numbers were not terrible, but, um... so if I'm doing the analysis of all this text, it's up to my discretion on what are- a-and it always is, with any analysis, but, I think, the more quantitative you get, the less subjective it becomes, I think.

Isaiah later continued,

> When I look at this XML data, I have to go through and say, oh this, uh...what could this mean, and I have to sit there, and myself, I have to look at it, and-and run through this analysis in my mind of all this textual data and go down certain paths of, like, what I think might be important and. Like I said, again you do that in sort of every analysis, but I think with quantitative variables, you can at least throw some math at it and get an answer back immediately. Uh, and there's- you- I mean, you can just measure the correlation of two variables and your response or something, and-and get an answer immediately, and, um. Again, analysis is always subjective, but I just- I think quantitative data cuts down on that.

Here, Isaiah reveals several things about the way he sees the world. First, he communicates a dislike for involving humans too much in the knowledge production process. Humans are contrasted with "just math," implying that math is simple, straightforward, or trustworthy,

whereas humans are not. Second, although he does acknowledge that there is some subjective aspect to any attempt to produce knowledge, he sees quantification as a more trustworthy process that reduces the threat of bias. As I will discuss in the next section, this intersection with quantification is a significant part of why data science is understood as holding the key to better knowledge.

The medical analytics team I observed also struggled to accept the idea that talking to clinicians could legitimately reveal the value of introducing predictive metrics to the intensive care unit. They trusted the comparatively thin accounts of eye tracking software and gps tracking systems over the words of their colleagues and even themselves. Much like the scientists chronicled in the work of Daston and Gallison, objectivity is something to be guarded against, and even the scientist herself cannot be trusted.

Data Science Circumvents Subjectivity:

Those Trusty Numbers

In some sense, data science provides the answer to this danger. Data science forces the representation of the world through numbers, it can avoid humans at multiple points in the process of collecting and analyzing data, and some of its methods even eschew human sense-making.

As already indicated in the quotes from Isaiah above, quantification is often seen as a route around human subjectivity and a means to accessing the truth. Numbers were not always associated with this kind of accurate, point-to-point, correlation to reality (Poovey 1998). In

addition to philosophical shifts, the associated between numbers and facts can be traced back to the emergence of double entry book keeping among merchants in the 15th century (Carruthers and Espeland 1991, Poovey 1998). The formal system and rhetoric of double entry book keeping was used by merchants to establish their virtue and trustworthiness, helping to facilitate a perception that numbers are transparent and neutral (ibid). Prior to this development, numbers were not used as accurate representations, but as didactic descriptions or imagery (Carruthers 2008, Espeland and Stevens 2008). Espeland and Stevens (2008) highlight this shift with an example from 1347. Louis Heylign Beeringen's account of the Plague describes the death toll by indicating there were 11,000 bodies. However, prior to the modern era, this figure, "communicated an almost unimaginable number, a multitude, rather than a precise number" (ibid: 406). Over time, numbers have come to be seen as trustworthy representations of reality (Espeland and Stevens 2008) that are strongly associated with objectivity (Daston 1992, Jasanoff 2005), especially in contexts where the involved parties cannot be trusted (Porter 1995). As such, numbers are powerful in any instance where subjectivity is feared to be interfering with knowledge production or decision-making. These practices have become so taken for granted, that it can be difficult to imagine alternative forms of operation. This trust in the veracity and necessity of numbers-based knowledge production and decision-making is both a driver of data science and its legitimacy and a perpetuated product of data science practices.

This authority of numbers over human insight persists among data scientists. Brent revealed this orientation toward quantification when he told me about his uneasiness with a previous employer:

> I could see how a lot of decisions were being made based on unfounded premises with very little data to inform them or really evidence supporting the argument. That worried me. To be in a company that was over a billion dollars a year in

> revenue and just seeing sort of "Well, we need to create these high priority accounts." I'm like, "How do you even define what a high priority account is? You don't have any metrics for assessing that. How are you going to do that?

Here, Brent emphasizes that "metrics," or quantified assessment, is the only way to really know what something *is*. He doesn't see high priority accounts as definable or even distinguishable from other kinds of accounts without a quantified means for identifying them.

For some, this connection between quantification, objectivity, and reality pushes beyond a mere method of representation. During one the medical analytics meetings that I attended, the group entered into a conversation where they were expressing their frustration with metrics that they felt did not accurately represent the improved success of care at Augustine University Hospital. They linked this problem to technical issues with measuring mortality rates. At this point, Dr. Osina exclaimed rhetorically, "What's the *true model* of why people die!" Later in the conversation other doctors and developers suggested that it may not be possible to measure the true mortality rate. They linked this challenge to the constantly changing patient population in the hospital. Dr. Osina responded forcefully that "You can! I don't understand the math enough, but there are probably only 30 variables." Although not all of the medical analytics team agreed with Dr. Osina that the *true model* could be identified, no one suggested that such a model does not exist. In addition, no one suggested that a non-quantitative approach might better capture the success of the hospital. Conversations like this one, as well as the group's efforts to produce algorithms that could translate measured vitals information into quantified representations of patient health, suggest a certain amount of faith in a world that operates according to mathematical models.

In another example, during an aside in my interview with Taylor, I expressed my frustration with teachers who grade on the bell curve despite having a class size that is too small

to meet the basic statistical assumptions of such a curve. Taylor asked me to elaborate on my frustration. During my explanation, I also indicated a frustration with an assumption that the world would actually unfold along the assumptions of such a model. Taylor nodded along, agreeing with me as I explained why I found such a practice inappropriate, even calling the use of the bell curve "stupid." Then, to my surprise, he followed up this agreement by saying, "I tend to believe much more in a lambda distribution." We then had a conversation in which Taylor explained and illustrated lambda distributions and when he thought they were appropriate to use. Both of these examples indicate that data scientists operate under the assumption that there is a "true model" of how the world works. There may be challenges in ascertaining that model such as difficulties in measurement or the selection of the appropriate statistical tool, but such models exist nonetheless. While this feature of the data landscape points both to, a sometimes unrecognized, use of theory and to the powerful way in which numbers constitute the world.

No Humans Required

Although data science and big data may use data that is not quantified in its initial form, the processes of data analysis inherently rely upon quantification at some level. As Berry describes, "a computer requires that everything is transformed from the continuous flow of our everyday reality into a grid of numbers that can be stored as a representation of reality which can then be manipulated using algorithms." (2011:2). Take textual analysis as an example. A basic data mining project might search through Twitter feeds or the text of digitized novels looking for particular words. These words are then *counted* perhaps allowing the analyst to assess the frequency of word occurrences, how these words are distributed across pages, novels, or the geographic locations of tweets. In short, the meaning rich context of word usage is rendered into

numerical representation. There are a number of more complex ways of dealing with text, such as topic modeling or decision trees. While these processes may not directly convert text into numbers, anytime time analysis or algorithms rely upon statistical concepts such as probabilities or correlations, aspects of the phenomena or their relation to other phenomena are quantified.

In addition to the inherent quantification involved in data science and big data, these practices often appear to avoid subjectivities through automated data collection processes. Automated sensors and data collection software found in smartphones, web applications, weather sensors, or wifi-enabled household items allow for data to be generated seemingly without human interference. This preference for automatically-collected data points is what Isaiah expresses above when he refers to his data as "second layer" data because it was recorded by people manually entering information into a system instead of being collected through an automated process. It is also evident the medical analytics team preference for automatically-generated time series data or eye-tracking software over clinicians' reports of how decisions are made and patient care unfolds.

Again, in our cultural context, this preference for automated data may seem to be the sensible and rational approach to knowledge production. After all, humans make mistakes and have been shown by some to be unreliable reporters of their own actions (Golden 1992). However, this preference for automation overlooks the many problems and opportunities for inaccuracy that arise with automated data collection as well. During my time observing the medical analytics team, I was surprised by how often they were unsure of the meaning or source of the data they were analyzing. They were sometimes unable to tell which equipment or monitors were producing the data subsequently stored on their servers or uncertain if data represented real measurement of a single patient or was due to equipment malfunction or the

transfer of equipment between patients. Stories of misidentified data were common among the data scientist. I asked Sienna to tell me about a time when a data field turned out not to mean what she thought it meant. She responded, "You should have asked this, is there a time that the data was exactly what I thought it meant. It would be once. I don't even know if that there'd be that many times." She went on to tell me about some data sets she struggles with because, although they are automatically generated, it is not clear what information they contain:

> Part of a bigger challenge at [our company] is that we have these different functional silos. Our finance department is figuring stuff out one way. Our procurement team is doing it another way. Marketing is looking, is importing information a completely different way and nobody's on the same page of what that is. You get a lot, lots of translation and there's not one owner so, as I was trying to sort through it, it was hard because I had to come and say, "This is what I think, sort of, is in the data," you know, given there's these missing pieces, there's these variations in the data that I don't know how to definitively say one way or the other what it does or doesn't include, because nobody else seems to know.

Examples like these may seem to show only errors in data analysis that are avoidable. And indeed, analysts are taught to be on the lookout for these problems. But the fact remains that humans must recognize the problems with the data collection for them to be resolved. Analysts are still making assumptions about the meaning of data, even when it is collected automatically. In addition to the problems I describe here, there may be faulty sensors, equipment, or flaws in software that lead to data errors.

No Theories Required: Unsupervised Learning

Finally, the methods of analysis associated with data science provide a path around human sense-making and theories. Often invoked in the same breathe as the term "big data," the process of "machine learning" promises to let us perceive truth without the bias of human

assumptions. Though data scientists may use a variety of methods to produce knowledge claims and generate predictive algorithms, I am going to focus here on what is often called "unsupervised machine learning." In the simplest terms, machine learning is a process that allows computers to develop methods for making predictions and interferences without providing pre-determined instructions and rules to the computer for which features or variable to use in making those inferences. It works by giving the computer a data set with labeled or classified phenomena, such as a collection of photos of pets that have been labeled as either "cat" or "dog" and telling the computer develop a method for telling those phenomena apart. As Alicia described it,

> So basically give [the computer] some training set, so things that you know are—so a classification, for example, like a binary classification. It's either "a" or "b." So you give it a bunch of things and you tell it, well these things are red and these things are blue, for example, right. And then you ask it, here's a big map, tell me what these things look like, and it'll label them for you, red or blue. And then the next step, if you want it to iterate, would be, okay you were right there, but you were wrong there. And you feed it all back in.

This approach already has an appeal to a culture worried about the limits of human subjectivity. It does not assume that humans are the best equipped to identify (and therefore program) the most salient features of a category. Unsupervised machine learning appears to take this omission of human sense-making a step further. In unsupervised learning, the computer is not provided with any labels or categories. Instead, when it receives a data set full of pictures of pets it determines the relevant categories by which they will be sorted and identified. While this could result in human-recognizable categories such as small pets, big pets, brown pets, or spotted pets, it could also result in features that are unrecognizable to interpretable to humans. This may have an appeal because in its avoidance of use of human-assigned categories, the resulting model also avoids human-generated theories that rely upon these categories.

In addition, the very use of the term "unsupervised learning" implies a certain level of distance from human subjectivity. First, by calling this a "learning" process, the impression of agency is imparted on the computer, drawing attention away from the human agency involved in these processes. The term "unsupervised" has a similar effect, obscuring the human work involved in producing models and algorithms through unsupervised machine learning processes. It implies that the computer generates rules, patterns, and categories on its own without human guidance. With regard to categories, this may be somewhat true (although in some forms of unsupervised learning humans do determine the numbers of categories). However, a great deal of human decisions go into generating these models. Leaving aside the choice of algorithms and analytical techniques, people must decide what information the computer has access to. Data scientists often refer to "feeding" data to the computer. This choice of data set has been shown to have significant consequences on the resulting model. In a rather infamous example from 2015, a Google Photos algorithm identified and labeled photos of black people as "gorillas." This offensive and problematic misrecognition by the algorithm may have stemmed, in part, from a data set trained on photos with an insufficient amount of diversity. Nevertheless, the practice of machine learning alongside the implications of the term "unsupervised learning" for the imaginary of the epistemological landscape present data science as providing powerful techniques for bypassing the dangers of human subjectivities.

Domain Experts

Up until this point, I have argued that the epistemological landscape of data science is one in which subjectivity and the limits of human perception threaten our ability to perceive the truth

and make sound decisions. The remedy to these dangers is to leverage the tools and techniques of data science to account for greater amounts of data and circumvent human interaction. Further, mathematical models are not representations of truth, but may be understood as the very laws or mechanics by which reality unfolds. In such a landscape, what is the role of the historian, the seasoned medical practitioner, the counter-terrorism expert, or the specialist in organizational design? These positions represent what many in data science refer to as "domain experts." They have specialized knowledge in a particular subject or area. This expertise is gained through a combination of training experience, such as immersion in other cultures or languages, the synthesis of a wide variety of texts or historical accounts, past mistakes and successes, or studies of financial or market trends. In short, experiences are an important aspect of their expertise and quantitative methods may or may not be part of what constitutes their claim to epistemic authority. This might lead to the conclusion that the knowledge of domain experts and the theories they generate would hold little sway with those that produce knowledge through data. While this is sometimes the case, I find that most data scientists still express value for theories and domain knowledge.

When Domain Expertise Fails

Both popular accounts and personal experiences have taught data scientists that data can overturn or outperform the knowledge of domain experts. In a 2014 article from *Harvard Magazine*, Gary King, Harvard professor and director of the Institute for Quantitative Social Science, recounts a story of data analytics out performing domain experts. Drawing on King's account, the article instructs the reader that, "the story follows a similar pattern in every field. The leaders are qualitative experts in their field. Then a statistical researcher who doesn't know the details of the field comes in and, using modern data analysis, adds tremendous insight and

value" (Shaw 2014). King then provides the following example. In 2002, a team of scholars developed a model to predict the outcomes of Supreme Court decisions (Ruger, Kim, Martin, & Quinn 2004). The model successfully predicted 75% of the outcomes of the cases from 2002. According to King, the research team, "collected six crude variables on a whole lot of previous cases and did an analysis" (ibid). The legal experts accurately predicted 59.1% of the outcomes. The research team also assembled a group of 87 law experts and asked them to predict the outcomes of the same cases. Quoting King the article states that, "the law professors knew the jurisprudence and what each of the justices had decided in previous cases, they knew the case law and all the arguments" (ibid). King uses this example to suggest that data is a more powerful way to know about the world. In doing the rhetorical work of promoting data science and big data, either King or the editors of *Harvard Magazine* neglected to mention an important aspect of this study. Although the model accurately predicted a greater percentage of outcomes than the experts in general, the authors of the study note that the experts exceeded the abilities of the models in some kinds of cases. Most notably, that the experts outperformed the model in certain types of cases, such as those involving judicial power (Ruger, Kim, Martin, & Quinn 2004). Instead of this nuanced understanding of which instances might allowed for more automated predictions and which require the expertise of legal scholars, the example given here simply communicates that pure numbers, when approached objectively, are better situated to explain the world that the subjective experiences of experts.

Some of the data scientists I spoke with had similar stories of data overturning or outperforming domain experts that they used to convey the usefulness of their craft. Milton, a senior data scientist with 30 years of experience and now working in healthcare, told me this story from one of his previous positions:

The company I was with was a startup. It was starting by, this was, they started, I think, around 2000. I joined in there about 2003. The system is widely used today. It's like, the people here [referring to the coffee shop in which we are seated], that work at Starbucks. You want to get a job at Starbucks, they send you a web page. You fill out a bunch of stuff, including a 50 question [form]. […] Twenty-five percent of all hourly workers that work for large groups, go through this system from this one particular company. You know, IO [Industrial Organizational] psychology, so it was founded by some IO psychologists and they built a whole bunch of these different texts. Their special advantage was that they made it electronic. It wasn't all paper based stuff, so it would make it easier to do. Because you probably know, in any of these kinds of tests, there is essentially a—it's based on a behavioral theory, etc., which I don't really report to understand at all, but there is, essentially, an answer key. That they know, if you answer this way, to this, then it means that based on their theory--No empirical, maybe some empirical valuation, but it was a great idea. It was all electronic but we also had the data on, for instance, for how long people stayed. In the hourly space, less today, but back in the early 2000s, turnover was huge. […] Yes, and it was just enormous. One of the things they did was to build an assessment that specifically looked at keeping people longer. It had questions that were based on that. We were able to actually track that. It didn't do very well. It didn't make it worse, but it was hard to say, it was statistically better, that they were staying longer. I joined, was recruited to join, along with a couple of other people--This was back, we didn't call it data science, but had lots and lots of data. Essentially, [we] had all the answers that everybody had answered and we knew how long they stayed etc., so we built a predictor model that didn't care what the questions were. There was no face validity. It just said, all right, let's run this through and come up with, empirically, the best answer keys.

It worked quite well. Much better than the, essentially, the theoretically-based one. But one of the things that was, but that ended up being a huge challenge in that the IO psychology folks. […] They were like, how, if you don't even know what the questions mean, how could you possible, how could this mean anything? In fact, what was, on each of the assessments, there was things you had to fill out at the beginning. Your name, address, all that kind of stuff. Those are required. It was required that you answer the 50 questions, but there was also some optional questions that had been put in, basically for some research that they were doing. They were free form questions and one of them was, the name of your 6th grade teacher. You remembered, just that you were supposed to put the name. It was free form. It didn't check. It didn't even, as long as you put something in. That turned out to be the most predictive, single thing, they had to pick one thing. Of course, they weren't—the IO psychologist went, they went crazy. This can't be

> possible and it's not. Eventually, they become comfortable with it by, essentially, reverse engineering theory, let's say.
>
> Claire: What's the explanation?
>
> Yes. This means that they're conscientious. If they're conscientious, because it was optional. It said right there, 'If you don't fill these out, we don't care.' The conscientious people stay longer. They came up with a story that made them comfortable with it and then they were happy.

Experiences like this one fit neatly with the views espoused by Anderson, Norvig, and King when they argue publically that theories and models—the products of domain experts—are no longer necessary in a world so saturated with data. Notice that although Milton is supportive of the idea that the theories of IO psychology might lead to solutions for retaining workers, he does not express any real concern over the eventual explanation that they generate for the patterns found in the data. He does not suggest that this data allowed his company to generate new theories or elaborate upon existing ones. Instead, he notes that the psychologists contented themselves with a "story," one that would suggest that an underlying theory was still at play. Milton himself seems little concerned with the veracity of such a theory and is instead content to note the pattern itself.

Yet, Theory Matters

Despite stories like these, most of the data scientists I spoke with and observed were troubled by the idea that theories were expendable in an age of big data and data science. This stemmed from both the perceived usefulness of theories and from the data scientists' commitment to more traditional forms of knowledge production. In interviews, I read an extended quote from Anderson's End of Theory[10] article and asked data scientists to reflect on it and tell me how it made them feel. Brent responded that it made him feel, "wildly

[10] For a summary of this article and the debate it fueled, see the introduction .

uncomfortable." When I asked Carter if he agreed with Anderson, he responded, "God no. […] I think that's dead wrong on so many levels." Almost all of the data scientists thought that theories were still needed to guide the generation of hypotheses, to design system changes and interventions, to move analysis beyond correlation to causal claims, or to produce general understanding of the world. Stephen, a data scientist who works on defense contracts in a consulting firm, expressed some of these perspectives in his response to Anderson:

> I don't know. I think there is definitely some truth to that. […] There's a difference between the correlations you can find with this kind of analysis that he's talking about and making deeper, more fundamental understanding of the underlying processes. I'm not sure if the scientific method gets you closer to that, but I think it's good—that's another thing—good to have this self-reflection on it. So I like that there are people working from both ends. I think that having people like philosophers to some degree thinking about holistically and just trying to understand it, conceptualize it from high up holistic, not down in the weeds. I think that there's a lot of benefit to having some people have this really, this really theoretical perspective on these sorts of things.
>
> I think that maybe from an applicability standpoint, the empirical modeling that he's talking about is potentially more directly applicable, but I think that there's a lot of, maybe it's just like a pursuit of knowledge thing of advancing or the harvesting of our knowledge, trying to uncover these fundamental kind of processes in this theoretical space and then proving them out in the empirical space. I think that that's really cool when that can happen. […] I think the give and take between the two paradigms is really beneficial for our pursuit of knowledge. I would be sad if it was just everybody doing more of what we do which is what he describes in that article of doing the number crunching, looking for these correlations. Like I said, I think that's more applicable, but it's not necessarily the best way or the only way to pursue the knowledge. We both will get a lot farther.

Brent continued his response along similar lines:

> I'll start more particularly on that and I'll just move into the more abstract. In particular, I think he's wrong. I think he's wrong because it's like saying we've got the saw so we don't need the ax anymore. Just because you have a better tool doesn't mean the tools you had before are somehow completely irrelevant. I would say he's also out of step with prevailing at least academic perceptions of

> data science as good practice. The guy's up at John Hopkins, who influenced a lot through their Coursera work, one of the fundamental things they say is, "What is the question you're approaching your data with?" If you don't ask a good question, you're not going to get a good result. That sounds a lot like theory to me. It sounds a lot like hypotheses to me. The point there is what we are doing is still science. The only way we can be assured of results as we were just discussing is that we have good scientific practice in what we do. I think moving into my more abstract, albeit, soap box there is a real, I at least feel, and this is personal so throw it out if you'd like, but I at least feel that there is a tremendous deficiency of philosophy in the quantitative world. There's a lack of appreciation for logic, surprisingly. Traditional ethics in logic, not is or is not or true or false. Not computer logic. Real logic. The fact that what we're doing is we're making an assertion about what is true in the world. We're making an assertion and we have to support that with good evidence. Just because the computer says so is not good support. There has to be some logic behind that. I think that's where that makes me really uncomfortable because it seems to have a lot of that sacrificing of the mind in it. Whereas it's just, "No. This doesn't have to be a well-considered issue. The computer will tell us." No. The computer won't.

Finally, Deana also expressed the need for theories:

> I think it's the kind of thing that is also spoken by one camp in the analytics world. I think it's incorrect because data doesn't give us knowledge or sense making or wisdom, it just gives us information. In my realm, in the field of learning analytics, the conversation is continuously about how we can make better connections between the data and the theory, because we have a long history of educational research that—coming from many different methodological schools—that has given us models and theories about how people learn, about how the brain works, about human cognition, about learning strategies, about learning design, effective course design and we're not going to just throw all of that away. The most valuable thing for us to be doing with our data is figuring out whether the data that we can get, which is partial and incomplete, and is never going to fully represent all of the nuances of what an individual learner is doing while they're trying to learn. How that data can help us confirm or deny if some of the theories and models that are in use or have been developed in other ways. I feel no drive, at least in this realm, to throw away theory at all. There is much more interest in making better connections between data and the theory.

As an additional indication of their faith in domain expertise, data scientists sometimes used a correlation between the results of data science and the conclusions of domain expertise

and theory to assert the veracity and authority of their work. In telling me about his colleague's work to predict which genes may be linked to cancer, Adam said:

> And if, um, so he ran lots and lots through our system, and some of the genes that popped out, these first couple ATM bracket to CD, um, KN2A [referring to various genes] and this and that—these are well known. Everybody knows that these are cancer-associated. So, that makes you feel good because you're seeing the ones you expect to see, but these bottom four were not known.

Adam continued, "some of the first predictions are well known, so it makes you feel good about the prediction." Rather than overturning domain knowledge, Adam used the fact that his company's analytics had identified genes already recognized as linked to cancer to bolster his faith that the new genes indicated by the analysis were indeed cancer-drivers too. In instances like this, data analytics and algorithms that mimic domain expertise—regardless of whether or not that expertise is accurate—are more likely to be valued, used, and trusted.

I suggest that the assertion that domain expertise and theories matter may stem from their education in a pre-data science era. Almost 80% of the data scientists I interviewed were 28 years old or older during my interview study, meaning that they had completed at least their undergraduate degree, if not also advanced degrees, before the onset of the big data phenomenon in business and consulting began to gain momentum and public awareness around 2008. In addition, only 2 of my interviewees were educated in an explicit data science program (due in part to the non-existence of most of these programs prior to 2013). Instead, 23 of the data scientists interviewed had a graduate degree. in various disciplines ranging from physics and genetics to clinical psychology or economics. These are fields where the goal is to produce generalizable knowledge, even when data-driven techniques are employed.

Their education and participation in institutions that value expertise and the production of theories may partially account for their placing value in domain knowledge. As more and more universities offer undergraduate programs and advanced degrees in data science, this perspective on domain knowledge may change. Data science programs offer training in particular methods that can be applied across disciplines rather than specialized knowledge and familiarity with particular domains. It is possible that the next generation of data scientists, trained in programs tailored to teach primarily the techniques of data science, will lose the value of domain knowledge.

Theories Still Lack Authority

Even in the current culture of data science that makes room for domain expertise, there may be little actual authority placed in domain expertise either in their minds or in practice. Given that most data scientists see value in theories which are produced by domain experts, I asked several data scientists to tell me about a time that domain knowledge conflicted with the data and how they decided if the data had been misinterpreted or if the experts got it wrong.

Veena, an education researcher and former econometrician for a large financial corporation, stressed this importance of domain knowledge in my conversation with her. She told me that "For us, we always start with the theory first," and she continued to tell me about an instance where her data appeared to be flawed until her team read more educational theory that actually predicted the outcomes contained in her dataset. Despite her defense of theory, when I asked her how she reconciles conflicts between data and theory, she said, "the data must be right because that's, I think, fundamentally the way we've come ... that's my worldview, I think, to some extent." Brent, who was "wildly uncomfortable" with the idea that theory might no longer

have a place in the world, also turned to the techniques of data science to settled such conflicts. When I asked him how he distinguishes between flaws in the data or analytics and a finding that challenges domain knowledge, he said:

> There is not a single expert. We are getting multiple opinions. Maybe sometimes that distinction can't be drawn. As consultants, we might not know that but chances are that with enough people in the room, I would be skeptical that the wrong view would prevail under those circumstances. The other thing too is that in the practice of what we do, we are trying to be as rigorous as possible. It's not the presentation of one interesting result. It's replication and repeatability, so that factors into it. We have quantitatively, [my boss] again, very big on this idea of target shuffling where we try and say, "How different is this result from random?" We put a number to that. We visualize that so that we're not finding signal in the noise. If it is just noise, we see that and we can show that. There isn't any storytelling that takes place there.

Notice that Brent's criteria for ensuring that the data is right and that the prevailing theories are wrong is to turn to the data and established processes of correctly analyzing it. Stephen, another advocate for theories and domain knowledge, did suggest that his team might reexamine the data, but he focused primarily on the usefulness of data for challenging experts:

> I think if you have the bandwidth, you always re-examine it and you say, "Okay, is this some kind of weird anomaly or is there something strange going on here?" A lot of times you don't discuss it. You're just running with what the software has basically fit. I think that even if [experts] do write it off, there's still maybe some benefit to that because then if they see that again and it happens, if that comes up in another model, then maybe they think about it a little bit more and maybe their frame of reference has shifted from the first time they saw it. They could say, "Okay, it doesn't make any sense this first time, but now I have more information. My frame of reference has changed a little bit and this thing, this result, maybe I should take another look at it." Even if it's doesn't jive with their sense—if we've done the modeling correctly, which is somewhat of and if, then there's some at least correlation there. I think it's valuable to have an unbiased mathematical thing making you rethink your thoughts. It's very easy to get confirmation bias when you're just staring at the same results every day or the same incidents every day, but if you have a tool that's fairly unbiased that can make you shift your frame of

> reference a little bit, then I think that can challenge the users and the subject matter experts to think a little bit differently. You're right, maybe a lot of times they'll just write it off, but even if they do nine times then the tenth time someone thinks about it a little bit more. Even if that all wasn't right, I think it's still good to potentially have an introspection on their thought process on their work that they're doing.

Stephen does acknowledge that there is the possibility that the data analysis process was not executed correctly. However, assuming that the analysis meets the criteria of data science, Stephen operates under the assumption that the data is right. He focuses instead on the issue of getting domain experts to accept this new information. It does not occur to him that the data analytics could be done correctly and still fail to correlate to the truth. He even describes the tools of data science as "unbiased" checks on the expert knowledge.

To some readers, these observations may come as no surprise—of course researchers change their theories based on the new results of analysis. However, there are few things to keep in mind. First, most of the data scientists are not also subject area or domain experts in the areas in which they work. Stephen is not trained as a counterterrorism or defense expert; Veena's background is not education, but economics and financial analysis; Brent is an engineer by training and now works on a variety of projects for a consulting firm. This lack of context may make it easier for faulty conclusions to enter the analysis, even when the mechanics of data science are executed correctly. Nevertheless, the data is almost always trusted over the expert. Second, their return to data science techniques to settle discrepancies between data and theory is surprising in light of their commitment to theory. We can imagine that if this belief in the power of theory and the authority of experts were central to their epistemological landscape that we would get a different kind of response from the data scientists. They might reply that in situations where domain expertise and data conflict that, "our data indicated a conflicting

conclusion. However, the experts in this area have such a good track record of generating testable and confirmed hypotheses/ or designing successful solutions and interventions, that it outweighed the conclusions of the analysis." Not a single data scientist with whom I spoke provided an answer anywhere close to this imagined, counterfactual response.

The Epistemological Landscape

The epistemological landscape of data science is to a great extent consistent with broader ways of seeing the world and notions of how to know that world: Humans face limits in their ability to perceive the world and may even introduce faulty claims due to our inherent tendency toward subjectivity. Quantification offers a potential route around this subjectivity. Data science, in particular, may offer a renewed path for avoiding subjectivity by providing automated methods of data collection and new methods that are perceived as reducing contact with human hands. Although some data scientists recognize that this process may itself introduce inaccuracies into knowledge claims, this danger is associated with potentially identifiable and resolvable technical problems, rather than dangers that are inherent to data science itself.

Surprisingly, data scientists still profess to value theory and domain expertise. However, these values are not an integral part of the landscape. Instead, they are scattered about as sometimes useful objects of a bygone era. To use the vocabulary of Boltanski and Thévenot (2006), we might think of the place of domain knowledge and expertise in the epistemological landscape of data science as the result of overlapping orders of worth. Some orders of worth—though still present—lose their ability to ground justification claims as new orders emerge. The epistemological landscape contains traces of an older order, one in which domain expertise and theories were paramount. Though data scientists may still use theories to generate hypothesis or

to confirm the results of their own analysis, the epistemic authority lies in the techniques of data science. Following the correct process is the means to ascertaining the truth and mechanics by which reality operates. As will become clear in the next chapter, data scientists try to protect these techniques so that they may successfully complete their scientific calling to make the world a better place.

Finally, what are the implications of this epistemological landscape for expertise and democracy? As data science becomes the means of generating knowledge and solving problems across a variety of domains, we are at risk of reducing expertise to a single type. Thus, although expertise may still appear to be diverse, providing the plurality of voices that have the potential to temper the power of the state, the source of epistemic authority circulating in these voices will remain singular. Because this single voice privileges claims built upon the epistemic authority of data science techniques, it may be ever more difficult for alternative voices to enter into debate. Although advances in computer processing and storage have made some aspects of data science more accessible, there is still a great deal of inequality. In the "big data divide," it tends to be only the powerful companies or government organizations that have access to the databases upon which the knowledge claims of data science are built (Andrejevic 2014). Under these conditions, we must question whether or not the potential solutions proposed by science and technology scholars—especially the notions of creating citizen-experts or teaching citizens to see science as a skill—will offer sufficient checks on data science experts.

CHAPTER 3

In Applied Contexts: The Pragmatic Data Scientist

Taylor is a lead developer at a small, but growing, contracting firm that designs data-driven tools for the defense and health sectors. In his late 30s, Taylor is trained as a physicist and is finishing his PhD in the subject as a part-time student. In the summer of 2015, I spent a day with him at his company, learning about a product he's been designing to identify gaps in organizational knowledge. When I asked him what makes a good model, our conversation unfolded as follows, moving from his preference for particular kinds of algorithms to thoughts on how he evaluates his own work:

> Taylor: I will trade extensibility and usability for accuracy. Because accuracy can be adjusted easier than extensibility or usability.
>
> Claire: Okay. You're gonna have to elaborate a little bit there for me.
>
> Taylor: So, let's say I pick up, um—let me find you an example that's not code related. Okay. So, I've got a pick-up truck. Okay? I can use it for just about everything, okay. So the extensibility's there, I can use it to carry cargo, I can use it to carry people, I can use it to haul stuff. Um, so the extensibility's there. The *usability,* eh, it's a little less. It's kind of bulky, it sucks to park occasionally, uh, gas mileage kind of stinks. Um, but I'm able to do a lot of other stuff with it. Um... *accuracy,* so this is where the metaphor is a bit screwy.
>
> Claire: Yeah, I was just gonna say, what's accuracy for a truck?
>
> Taylor: Um...you know, really, if we're talking about—optimal performance as an accurate measure, right? Um, meaning maybe gas mileage or gas consumption. Um, it's kind of all over the place depending on what I'm doing with it. Um, but there are ways that I can adjust that. Right, so I can't take a car, can't take a Honda, that gets fantastic gas mileage, fantastic performance, and make it haul a camper or trailer. I can however haul the trailer and the gear and all the things with my truck, and do certain things, change my driving mechanisms to increase the accuracy, the reliability of the performance of the car: by putting a tonneau

cover on it, by making sure my tires are always inflated properly, by adjusting how I drive it, using cruise control more or less. So I can adjust kind of the performance attributes of it. So maybe accuracy is the- is a wrong monitor- maybe it's *performance* instead. So extensibility and usability and performance are the three majors, right? Cause performance we can say precision, accuracy, etcetera, etcetera. So I can modify some things to increase my performance, but I can't necessarily make changes on the other end; on usability or extensibility. Um, as easily. Um, it's all made- I'm glad we made that change. So, for an algorithm, if I've got an algorithm, um, like the algorithm that I finally settled on for the modeling component, is broader than Bayesian's, um, it's a little...harder to use, but not much. Um, but it covers a much broader domain. Because I can use discreet probabilities, continuous probabilities, discontinuous probabilities, sets. Um, which I can't do in Bayesian. And my performance is milliseconds less. So I can take that hit because I've got the extensibility. Um, so that's kind of the reason I say that I would trade off a little bit of performance for the extensibility. Cause then I can use it across every problem that I come across rather than just one, or one really well-bounded problem.

Claire: Have you always had that mindset about it?

Taylor: No. I used to be a perfectionist. I wanted the best performance, the best extensibility, the best usability. Um-

Claire: So what changed?

Taylor: Oh [sighs] too much pain.

Claire: In trying to…?

Taylor: In trying to *find* that.

Claire: Achieve that, yeah.

Taylor: Yeah, and achieve it. And it's just- it got to be- it's exhausting, trying to find perfection when good enough is really good enough. You know, um, but then it comes down to managing expectations and understanding requirements, and things like that.

Claire: When did you let go of the...perfectionist?

Taylor: Oh, I'm not sure I fully let it go.

Claire: Or when did you start trying?

Taylor: Probably about five years ago.

Claire: Okay. So here in this context [meaning his work at his current company].

Taylor: Yeah. Because we were doing so much R&D [research and development] and so much cutting edge, right, you want to make it all perfect, but you've got to get enough done to *show* what you're doing. So you start going, okay, well yeah, that really doesn't matter. Or it doesn't matter as much now. Um, it may matter in the future, but I'm not gonna do anything to preclude it. I'll do things to include later. Um, so yeah, it's been much more of that dynamic R&D, have to create something.

In this conversation, Taylor links his preference for particular choices related to data analytics to the criteria by which he evaluates his own work. Through the detailed metaphor of the pick-up truck, Taylor tries to explain why he prefers algorithms that can function across a variety of contexts, even if they result in less accurate results. The tension he experiences between what he calls a "perfectionist" orientation to a "good enough" orientation is connected to his movement between contexts, specifically his transition to a data-driven consulting firm. In referring to performance aspects such as precision and accuracy, Taylor indicates his orientation toward scientific standards as the criteria by which models are assessed. However, in wanting something that is extensible, Taylor points to the demands in his work setting to produce something that will get the job done quickly, even if it sacrifices adherence to scientific standards to some degree. As I will show in this chapter, the tension that Taylor expresses is one shared by many of the data scientists with whom I spoke. Data scientists—especially those who work in more applied contexts—are caught between two sets of criteria for assessing value. Drawing on their training in academic settings, many of the data scientists rely upon an ideal-scientific

method for evaluating themselves and their work. However, in work contexts that are driven by the need to produce results, win contracts, and satisfy clients, data scientists become oriented to a pragmatic method for evaluating work, one that emphasizes the work data analytics can do, rather than their ability to produce knowledge. This is potentially problematic as it may lead to knowledge claims that are misleading or inaccurate, even according to the tenets of data science.

People make assessments and judgments according to shared cultural codes. Where some lines of institutional research address macro-cultural patterns or symbolic orders by which entire groups or societies discipline their behavior or make judgments (Durkheim and Mauss [1903] 1963, Douglas 1986, Foucault 1980), more recent approaches to institutions connect symbolic orders to particular settings and contexts. This line of work has established that modern societies exhibit a heterarchy or plurality from which people make judgments, evaluations, or justifications (Friedland and Alford 1991, Boltanski and Thévenot 2006). As such, a single individual may use many different "orders of worth" without conceiving of them as being in conflict (Boltanski and Thévenot 2006: 215). This stems, in part, from the association of particular justifications with particular contexts and situations. For instance, in an industrial order of worth, actions are evaluated based on efficiency and productivity, while in a market order, price is used as the primary mode for evaluation.

In considering the ways in which data scientists evaluate their work, it becomes clear that there are two overlapping and partially conflicting sets of virtues by which this assessment occurs. I call these two orientations the idealist and the pragmatic. These orientations come into contact when professionals with scientific training, generally focused on generating knowledge, enter applied settings. By applied settings, I mean that the goals of the organization are to provide information and analysis that allows for interventions in decision-making and outcomes.

I show that in applied settings, data scientists are often willing to accept results that violate the idealist criteria so long as they fulfill the pragmatic criteria for evaluation. Thus, despite the presence of the epistemological landscape described in chapter 2 and the first part of this chapter, in producing data science, data scientists sometimes shift to an alternative framework or logic. This alternative moves away from concerns about epistemological authority and toward a concern for legitimacy within a logic that more closely resembles Boltanski and Thévenot's (2006) industrial and market orders. In making this shift, however, data scientists are still able to adhere to important aspects of their epistemological landscape, namely the belief that data science has the potential to change the world. As I discuss, this shift recasts issues such as for whom data science betters the world.

In what follows, I begin by reconstructing a few more contours of the epistemological landscape of data science. Consistent with the imaginary of data science portrayed through media, data scientists conceptualize their work as having the potential to drastically improve the world. In addition, they situate themselves as "scientists" within this landscape, capable of directing the technical processes of data science to meaningful ends and charged with ensuring proper execution of these techniques. I argue that this responsibility to and reliance upon scientific standards, what I call the idealist approach, sometimes conflicts with a pragmatic approach in which the usefulness of their work matters more than its relationship to truth.

A Few More Contours of the Epistemological Landscape

During my time interviewing and observing data scientists, I heard many expressions of values that we might associate with the tradition of academic and scientific knowledge

production as data scientists positioned themselves as part of a scientific mission to better the world. Despite many of them being professionally located in for-profit industries, data scientists prided themselves on doing more than executing mindless analysis. They saw themselves as producing knowledge that would facilitate positive change in the world, as possessing the ability to identify useful areas of research, and as stewards of responsible research practices.

Data Science Can Change the World

Many of the data scientists with whom I spoke imagined that data held the power to greatly transform the world for the better. This view sometimes came from the actual projects on which they were working and sometimes from imagined potential projects. Alicia told me about how her work would allow for better responses to disasters. Matthew imagined that applying data science to medicine would allow for faster, more efficient, and cheaper discovery and advancement of much needed pharmaceuticals. Regardless of the use case, data scientists often focused on the ability of their work to provide more information or to synthesize information in ways that they thought could improve the world. For example, the medical analytics team that I observed saw their work as truly life-saving. As I will discuss in more detail in chapter 5, they believed that there was a great deal of medical data and information that was not currently being leveraged to produce better health outcomes and they worked to resolve that. A central mission of the group was to ensure that the information from vitals monitors, lab results, and electronic medical records were stored and available for data mining and analysis.

Ben, a graduate student working in a data science lab on algorithms to predict political unrest, also saw his work as holding the potential to greatly alter the world for the better. When I asked him how he imagines his models and algorithms might be used, he responded that there

were three ways in which predictive models are used. First, he briefly mentioned using predictive models to either prepare for inevitable outcomes or to improve particular outcomes. These are two aspects of data science that I will discuss in more detail as part of the pragmatic approach to data science. Ben spent much more time describing a third use for prediction to me: in the very act of predicting an outcome, Ben believed that data scientists could prevent an event from ever unfolding. He described this function as follows:

> The third one, which is predicting so that our predictions will prevent those things from ever occurring, is I think, the most interesting one. Just because it has—hypothetically—it has so much potential for preventing conflict. Obviously that's very hypothetical, and it's not like anything I do or anything anyone will do in our generation will achieve that, but--

I then asked Ben why he thought predictive models might someday be capable of preventing events from ever occurring just by predicting them. He said:

> There are kind of different game theoretic ways of explaining it, but I think the more intuitive way of explaining it is, like, imagine a poker game in which everyone could see everyone else's cards. The game really wouldn't make sense anymore because everyone who should fold would fold and everyone who would win would win. There wouldn't really be any competitive element to it. Now, with that said, people play games of chance. They still continue to play that, so who knows if humans are so irrational that we would continue to engage in conflict even when we knew it was costly, that's obviously a possibility.

Sentiments like these reveal several more contours of the epistemological landscape of data science. First, data scientists often associate negative aspects of life experience with either a lack of information or a lack of human ability to process it: poor responses to natural disasters, deaths in a hospital, and even war are at least partially attributable to insufficient access to or use of information by humans.

The Job of the Data Scientist: Asking the Right Questions

In positioning themselves among those who would use their work to better the world, data scientists often emphasized their role as *scientists*. This tendency was evident in how they distinguished themselves from other kinds of data analysts or technical workers and in their sense of responsibility to maintain proper research practices. Part of this scientific identity is marked by their ability to determine the research agenda and approach to solving problems. In our conversation, Angie emphasized that her job was not simply about exploring data to find unexpected or unknown relationships between phenomena. Instead, she told me:

> Part of what we do as data scientists is find out how to ask the right question and what is the data that we need to answer that question. To think about that, if I just throw all my data into a computer, and get a bunch of stuff out, how did that help me frame the question? I think it's still helpful to think about different theories.

When distinguishing between other analysts and data scientists, Brent also associated the job of the data scientist with asking questions:

> With a data scientist, there's more of a scientific approach. There is the question of posing questions of the data and asking something from it. I guess it's moving from that very low level, not mechanistic, but still really low level is the right way to describe it.

In addition to stressing the ability to ask the proper questions, Brent also contrasts this job with a more mechanical approach—implying that the data scientist must do more than appropriately execute processes and techniques associated with data science. Matthew also stressed the importance of these abilities:

> Data science is no longer about which algorithm fits the problem. We can try them all just at the push of a button. It's almost like spending all of our time on the upholstery in the vehicle. You've solved it. Okay. You can have one grade of leather or another grade. What really makes a slight difference whether it's a

> neural network or a random forest, or something else? We can do all that. What matters is how we think about the whole problem. That's where the creative process occurs in data science.

Similarly, in telling me about her experience transitioning from working as an astronomer at an observatory to a data scientist in the health care industry, Denise told me, "I really like the technical side of what I do, but there's the problem-solving and critical thinking is really important in data science. You're not just given programming tasks and you do them, you really have to think through the problem."

In their minds, data science is not just about the technical skills so often associated with this emerging profession. It is also associated with an ability to recognize what kinds of information are needed to solve problems, what kinds of answers the data is capable of providing, and the ability to think critically about the tasks before them.[11] Through these skills, data scientists stress that they have a role not just in executing data science techniques, but in guiding data science toward the appropriate ends.

The Idealist: Data Integrity and Responsible Research Practices

The emphasis on critical thinking, asking the right questions, and creative application does not mean that the technical processes are unimportant. Instead, the idealist approach to data science stresses the proper application of techniques, honest communication about results of analysis to others, and the evaluation of analysis and models according to established scientific criteria of fields like statistics and computer science. Borrowing the words of one of my

[11] Note also that this emphasis on critical thinking and knowing which questions to ask points to the continued use of theory in the work of data scientists. Angie even states this explicitly, indicating that theories can guide question-asking. For more on the tension between theory and data empiricism, see the previous chapter.

interviewees, I call this aspect of the idealist perspective “data integrity.” Each of these responsibilities can be seen in a conversation that I had with Sienna after asking her if it is possible to manipulate data analysis to get specific desired results:

> Sienna: Yeah. I have a person I work with who is decent enough with data but doesn't have the best, I guess, data integrity, and so definitely begins a lot of analyses with an idea in mind of what he wants the result to be. Then it's not like, he doesn't have any personal gain from it, it's not like he's adjusting his numbers or something, but he has an idea of what the story should be and is unusually successful in making the data match that in ways that don't often make sense to me. I think data is incredibly powerful and I think people just assume that if you know how to do some sort of analysis with it that it must be true. It's like the "lies, damned lies, and statistics." I think you have to be very, very careful about information that people give to you.
>
> Claire: I guess I have a question that's kind of two sides of the same coin. What do you do to guard against that happening or how do you evaluate for yourself whether someone has done that?
>
> Sienna: That's a good question. I guess on the evaluation piece, I really like to go back asking about the process because I think that can really eliminate a lot of how much the person thought about information I think. So many times people, like the people I work with, will just pick apart the actual numbers. They're like, "This 3%, shouldn't it be 5%, da-da-da?" That's not very good for you, but I really like to ask someone to walk through the process of like, "Okay, tell me how you put this together." That, I think, brings more light into what they're including, what they're not, and especially if they are able to disclose anything they feel a little uncertain about or that maybe seeming the little off about the data. Any time that comes up, I think that's a pretty good indicator that the person has thought about it pretty carefully. I think preventing against it is just having an open enough environment. People feel comfortable sharing those uncertainties, because I think there's so many times we're nervous about—and I've been guilty of it too—I’m nervous about sharing something that I know that I don't quite understand about maybe this piece of data looked weird or I'm not sure I can trust how clean this information is. We're nervous about sharing that because that's when somebody's going to think the entire thing is invalid. Really, it's just being honest about, "Here's where I see potential problems but I don't think it will impact the information core. If it does, I think it impacts it in this way."

In using the term "data integrity," Sienna signals a responsibility to let the data speak for itself. There is a sense among data scientists that if a data scientist follows the proper techniques, the truth of the data can speak through. Chris even indicated that if mathematical procedures were followed correctly, then "if two data scientists look at the same data, they will come to roughly the same conclusions from it, or at least the same mathematical output of it at least, objectively." This indicates that there is a message in the data that can be determined if procedures are followed correctly.

However, consistent with the concerns for subjectivity discussed in the previous chapter, data scientists can pollute this process. Alicia shared this view and told me, "the data is there and the numbers don't lie, it's the people who are really good at lying with them." I asked her if she thought that people did this intentionally, and she replied:

> Unintentionally, I think, are the people who don't know what they're doing, and intentionally, the people who really do have something to gain, usually. Like that-that example I was saying, a really simple one, right? Of where you have a very smooth line of some sort, and along it are tiny bumps. [Alicia refers here to a graphic visualization of data.] What you can do is you can stretch that colored gradient out and make something really green and make something really red, where really they're very, very similar. And so, yeah the data is telling you that there are differences, but how big are they compared to baseline—[…] So of course you can light them up and you can say, oh look how different these things are, where really they're all really pretty similar.

Consistent with Sienna's assertion that data scientists should be honest about potential shortcomings or limits of data analysis, Alicia communicates a sense of responsibility to accurately interpret the results of data analysis when presenting it to others. They both recognize the ability to use their skills to mislead others—whether with good or bad intentions—in either the execution or presentation of their research.

Data integrity also signals proper treatment of the data and execution of analysis. Due to the advanced technical, mathematical, and statistical skills needed to execute their work, data scientists are often the stewards and critics of their own work. In evaluating the quality of their work, one approach is to draw upon the evaluative criteria and methods of statistics and computer science. When I asked Veena what makes something a good model, she replied, "whatever explains the most variation in the data is what I consider a good model." She continued on by discussing the levels of acceptable p-values that her research team uses in their analysis. Similarly, Angie said, "a good statistical model is one that is repeatable. It stands up over a rigorous cross-validation." Along the same lines, Brent responded to this question by saying:

> I guess one thing would be robust. It stands up to new data. It doesn't do anything unexpected when there's new data. Responsive, so it does change when there's new data but it's not that it changes in some unexpected way and that it's heavily validated. We're not just excited about one really good result that we're seeing. That's kind of on robustness, but still I think it comes more to the scientific aspects of replications. It's tested, it's tested, it's tested again**.**

In each of these responses, data scientists invoked standards of evaluation that they most likely learned through their academic training. When Veena indicates that a good model explains a great deal of the variation in data or when Brent and Angie indicate that a good model is one that can be replicated on new data they are pointing to traditional ways in which statisticians and scientists have tried to ensure that their models are not random results, but are likely representations of true relationships that exist in the world. In this way, data scientists use the criteria of science to evaluate their work. And there is a certain logic to this idealist view: if the problems in the world result at least in part from imperfect information, then any attempt to

improve that world would include a commitment to making sure the results of science reflect the truth of the world.

These values are consistent with the ethics and values that are taught in research classes of scientific disciplines and presented in ethics and scientific methods handbooks. Through these experiences and materials, scientists learn that, "researchers have an obligation to honor the trust that their colleagues place in them," and that they "have an obligation to act in ways that serve the public" (National Academy of Sciences 2009:2). Clearly, these kinds of values continue to inform the ways in which data scientists view their work, encouraging them to focus on how their work will better the world and on the integrity required to produce and communicate their work.

These methods of evaluation could be considered a type of "epistemic virtue," or a set of values that are "preached and practiced in order to know the world" (Daston and Galison 2010: 37). It is clear that data scientists place epistemological authority in the techniques of data science. This can be seen both in the material from the previous chapter and through comments like the one Alicia made that assert that "numbers don't lie." In this world, the job of the data scientist is both to direct inquiry and to ensure that it unfolds according to processes that are trusted to reveal truths.

The Pragmatist: **"All Models are Wrong, but Some are Useful."**

Despite the presence of the idealist perspective in the minds of data scientists, a second orientation also guides data scientists in their approach and evaluation of their work. However, almost of my interviewees moved on to say that these criteria mattered less than the model's

usefulness. Like many others who invoked traditional evaluative criteria, Brent went on to tell me that there are additional criteria that matter:

> I guess over and above all of that, that's valuable. It's useful. It could academically be the purest model and algorithm ever but if it's not deployed, it's totally useless. It has to do something that makes someone's job easier, increases the bottom line, but still that it addresses the question that's there and that creates value.

This view is a more pragmatic one that stresses not ideal conditions, but practical applications and outcomes. These conflicting orientations are sometimes at odds with each other. While the pragmatic view does not necessarily contradict the desire to better the world or the emphasis on asking the right questions, it does recast these values. In addition, it challenges the criteria for evaluating work that stem from the idealist view.

When I asked Kieran what makes something a good model, he replied, "That's pretty easy because I work at a company and so the answer to that is that the model delivers good business results." Although few expressed this view so succinctly, this assertion that good models are useful models was part of the evaluation criteria for most of the data scientists whom I interviewed. Matthew expressed this view when he told me, 'When you start judging how well did this model do, you need to do how well did it do compared to what it did before, because the famous saying, all models are wrong, but some are useful, is absolutely the truth."[12][13] Angie also felt that the criteria of scientific investigation that she had been taught as a physicist were less applicable:

---

[12] Here, Matthew is using a quote usually attributed to statistician George Box. My interviewees used it again and again to assert that, while their models may not correlate perfectly with reality, they could "do" things with these models that offered value. See George E.P. Box. 1979. "Robustness in the strategy of scientific model building." *Robustness in Statistics*. R.L. Launer and G.N. Wilkinson, Editors. Pg. 202.

[13] Note the way in which the statement "all models are wrong," reinforces a notion that theories are less important. This fits with the "algorithmic modeling culture" (Breiman 2001) discussed in the introduction and with the placement of epistemological authority in techniques over domain knowledge and experience discussed in chapter 3.

> If it's not useful, then it's not a good model. The model we've been validating for our client, they actually, the sampling technique they used, statistically is not correct. They're assuming that they have distinct data points that are in fact not distinct. Me, this year, is not independent from me last year. There's a correlation. The work I did in Toronto. If you were to scan the same mouse over various points in its life, those data points are not independent. They didn't account for that at all in their model but for what they need the model to do it is useful. It's not wrong. It predicts things. The fact that it's not statistically robust is irrelevant. The model is explainable, it's reproducible, it holds up under cross-validation. All of these things.

In this case, the fact that the model Angie's client has been using is statistically inaccurate does not matter because, despite these statistical flaws, it allows the client to produce information that improves their business performance. These evaluation criteria might mean that data scientists decide to use a particular model, even when it falls short of the ideal scientific standards. Deana, an education researcher trained as a PhD in genetics, hit on exactly that tension when she was telling me about how she evaluates models:

> We put that [model] together, but what's interesting is the discussion about how good is good enough. People will look at that model or similar models and let's say 75% and 80% or-- The best predictive power I've seen for failure rate models of this kind isn't around by 80%, maybe further for more data mining approaches and this quibbling about well, in use, if that's the best accuracy we can get, how good is it? I keep saying "yeah, but before we had no information, now you have 80% accurate information, so isn't that better than no information?" I think there is a discussion to be had about how good is good enough for this purpose.

Here, Deana invokes many of the evaluation criteria typically used in scientific and research settings to evaluate models. She refers to 80% accuracy and predictive power, performance metrics that might be considered low in traditional academic research settings. But rather than focusing on these metrics as standards, she expresses frustration that these models get ignored by some in her research community. She feels that if a model adds information—even if that

information is flawed or not perfect—then it should be considered a valuable contribution. In Brent's comments on good models, he even pointed to the tension between scientific ideals and business-driven settings:

> It struck me in particular too because there's an element of perfection that exists when you're coming from an academic background where it's "Well, the best answer is always the most rigorous and complete one" as opposed to "Well, this is a good improvement over your base-line." That's really what you should be aiming for. Not "Is this the best answer, this is a good one."

It is notable that the same individuals who stress the importance of evaluation criteria derived from the sciences would be so quick to accept models that are flawed by these standards. As these quotes indicate, the applied setting in which most data scientists work required that models not only state knowledge claims, but that these claims be put to use. There are several features that data scientists often associate with the usefulness of a model: interpretability and actionability.

Interpretability

First, in applied settings, data scientists often stressed the benefit of models that are interpretable. For most, this means that the relationships indicated by the model can be described in terms of human-constructed concepts. In his list of desirable concepts for a good model, Carter said, "Second of all, is it interpretable? […] Can human beings figure out what it means? If we can't, there's a lot less we can do about it." As expressed by Carter, for many, this preference for interpretability is directly related to the desire to make interventions, especially those that change outcomes, based on the resulting models.

For others, their location in a business-driven setting also drove them toward a preference for interpretable results. Chris explained that his evaluation of models had two components:

> Probably I would say there's maybe two main things. One is, is it predictive? So does it accurately make predictions about what's going to happen based on the data that you have, and then at the same time, is it interpretable. Particularly if you're doing work for a customer like [our client] who has to make decisions and justify those decisions based on data, or models, or what have you, you can't just give them a black box that spits out the answer because they need to be able to say why is that the answer and justify it to whoever they answer to. Some types of models are like that, where it's basically a black box and it'll spit out answers and you don't really know, like a neural network, you don't really know why that's the answer. […]For our purposes that's not a great type of model to use because we can say what the answer is but we can't say why, whereas other types of models, like a linear regression, or even some other more exotic types of models, you can say this is the answer and these are the variables that are most important in driving that answer. This is their relative importance. This is the level of uncertainty around the answer, and you can caveat it and say what's important, why is it important, how much uncertainty is associated with the answer. It gives them a lot more context that they can use to justify the decision they're making or go to their bosses and make a case for something with a lot more rigor behind them. For us, having a model that is interpretable in that sense is just as important as being accurate and right.

Chris's response shows that in certain use cases interpretable results are desirable for reasons that have little to do with the accuracy of their results. In the instance that Chris describes, producing a knowledge claim, such as a prediction about markets, is not appealing to their clients unless they can also explain why they think this particular prediction will come true.[14] The need for interpretable results is driven more by the need to legitimate their claims according to the cultural frame of their clients. When those clients are not data scientists, this means that the explanation of the claims will need to rely upon human-constructed concepts such as seasonal trends, unemployment rate, or market share.

[14] Note that Chris's preference for results that can explain "why" also points to the importance of explanation and theory in these kinds of use cases in data science. For more on this issue, see the previous chapter.

Actionability:

Time and time again, data scientists told me that part of calling something a "good model" included an evaluation of its ability to alter decisions in applied settings. Stephen, a data scientist who worked with soldiers to develop predictive risk algorithms for the location of improvised explosive devises (IEDs) in Afghanistan, told me it's a good model, "if it's useful. I guess that you could argue that it's inherently good and the more useful you can make it, the better it is." As an example, he said, "You might say, 'Every one occurs on the road.' Yeah, that's a really good, great score from that." In this statement, Stephen implied that his hypothetical model would do very well against scientific evaluation, but as he stressed, "You'll get a perfect result, but that's not useful. They know that." Though such a model would be scientifically accurate, it would not facilitate better decision-making for the soldiers trying to avoid IEDs. In a similar sentiment, after asserting that performance of the model matters, Milton went on to stress the ability to transform the model's information into action:

> There's two different things. There's the model itself and then what you're going to do with it, so there's accuracy of the model and then whether the thing is intervenable or not. You could have, just like you could have a model that would be extremely accurate at predicting the weather, but you can't change the weather, but it would impact the way, whether you decide to carry an umbrella that day or not. I think it's the accuracy and then the intervenability.

Not only does Milton stress the importance of being able to do something with the model, but he points to two different ways in which this might occur. First, it might be possible to actually change the state of the world by altering phenomena or features included in the model. To give another example, data scientists might do some modeling to determine the features that contribute to successful student outcomes in college course and then alter the course curriculum to maximize outcomes. This is an example of direct actions designed to change outcomes.

However, models might also be actionable if they allow for an improved response to an outcome that cannot be changed, such as the weather forecast or likely location of IEDs along roads in Afghanistan.

It is worth noting that the preference for interpretability and actionability may not be necessary in all applications of data science. They matter most for the cases and demands that data scientists face in applied settings: instances in which data science is being used to construct interventions and change outcomes or to respond to predicted outcomes. Interpretability and actionability are especially important in cases where data scientists or their clients intend to alter outcomes. It is difficult, if not impossible, to alter classroom curriculum and course design to facilitate better student outcomes if you do not understand the features which the model uses to predict student outcomes. In the effort to change outcomes, actionability matters as well. As Carter, the education researcher put it:

> A lot of things are interpretable that aren't actionable, and I'll give you an example. We've known for a while that you're much less likely to go to college if your parents don't make much money, but you can't tell a kid, "Hey, you're not likely to go to college, so tell your dad to go get a better job." It just doesn't work. It's not actionable.

In this case, Carter points out that certain features in a model might refer to concepts that cannot be altered, and therefore they do little good for interventions. However, actionable models also refer to the ability to respond to the predicted outcomes of the model, such as expected weather, predicted location of IEDs, or risk scores for hospital patients getting an infection. In these cases, interpretability matters less. Instead, the focus is on the value provided by the resulting prediction and the ability to make decisions based on this information.

In contrast to the ideal perspective, the pragmatic data scientist evaluates her work based on what it can accomplish, rather than its adherence to scientific principles or the likelihood of producing knowledge claims that reflect the truth. One way that this conflict can be described is as a tension between an emphasis on the means versus the ends. Though there is a clear, desirable end in the idealist view—the desire to produce accurate knowledge and to better the world—the means are a very important aspect of meeting those goals. In the pragmatic view, the means matter much less so long as the desired end is achieved. To be sure, data scientists still care about scientific criteria, but work in applied settings also drives them toward these conflicting priorities.

Implications of the Pragmatic Approach

How is it possible that the very same people who stress the importance of data integrity and scientific methods of evaluation are so quick to accept flawed models as long as they are useful? Although some data scientists, such as Taylor, expressed a slight sense of remorse or conflict in describing the tension between the idealist and pragmatic orientation, most seemed quite comfortable with the mutual reliance upon these conflicting criteria. In the final section of this chapter, I offer a possible explanation of the co-existence of the ideal and pragmatic data scientist, explore the theoretical implications of data scientists' adoption of the pragmatic approach, and present some potential empirical concerns resulting from the pragmatic approach.

First, it is likely that these conflicting views stem from the overlapping institutional contexts in which data scientists operate. This is especially true for those who work in applied settings. As I discussed in the previous chapter, most of the data scientists in this study were

educated in a pre-big data era. Their formal training was centered on the mission of particular disciplines. These missions focus on asking the right questions to generate plausible knowledge claims about phenomena such as outer space, the human body, or social patterns. The particular technical skills associated with many disciplines are a means to these kinds of ends. When data scientists take the skills that they learned to achieve these ends into more applied settings, the values associated with their training get recast and the focus of knowledge production shifts.

How do we make sense of this transition to a pragmatic approach from a cultural or institutional perspective? Does this demonstrate that the epistemological landscape is not powerful in shaping the actions and production practices of data scientists in applied contexts? Does epistemological authority really lie in the techniques of data science? Recall that Boltanski and Thévenot (2006) suggest that people are able to hold conflicting worldviews at the same time. Based on my discussions with data scientists, they rarely experience the ideal and the pragmatic as in conflict. Instead, these views are themselves ordered. As Stephen's example of a perfect scientific model that tells soldiers only that all IEDs are located on a road indicates, the scientific criteria of models matter, but they are secondary to usefulness. Data scientists clearly are still influenced by the ideal approach; they express dismay toward colleagues who do not exhibit data integrity and they still structure their analyses toward scientific standards. When the ideal and pragmatic perspectives align, data scientists still place epistemological authority in techniques over the experiences of domain experts, for example. It is important to note that the focus on usefulness and pragmatic results may even align with aspects of the epistemological landscape, such as the desire to better the world. In focusing on useful models, data scientists create interventions in the world, rather than only constructing abstract knowledge claims.

However, the shift to the pragmatic approach and the lessening concern for epistemological authority does point to the usefulness of the practice approach to the sociology of knowledge (e.g., Latour and Woolgar 1986, Knorr Cetina 1999). So far as the construction of knowledge claims is concerned, we can see that it may not always unfold fully according to the espoused processes of scientific investigation. Continuing to investigate the practices of data science will be an important aspect of data studies moving forward, allowing researchers to ascertain the ways in which the diverse environments in which data science is deployed shape the relationship between data science and social outcomes. This is an effort which I take up more fully in chapter 5 as I explore the use of data-driven algorithms in the medical context.

Though data scientists can retain some aspects of the epistemological landscape during their shift to the pragmatic approach, we must ask what happens to these ideals when they are recast through the pragmatic lens. Data scientists still believe in the ability of their work to better the world. In the pragmatic frame, for whom does the world get better? The same is true of their scientific responsibility to ask the right questions and solve problems. Asking the right questions to do what? Solving what kinds of problems? As Kieran's observation that better business outcomes are the marker of a good model in a for-profit setting makes clear, these questions are now answered by the organizations for which data scientists work. Though some applied settings still work toward what we might consider universal goods, such as improving patient outcomes in a hospital, many of the settings in which data scientists work use the technical skills of data science to increase profits and maximize efficiency for businesses and clients. This should lead us to question the progressive narrative of hope and goodwill often contained in the epistemological landscape.

In shifting to a pragmatic approach, data scientists may produce analysis and products that, in addition to overlooking alternative forms of knowledge, no longer adhere to the epistemological authority of data science. While this may seem unproblematic when data science is used to solve problems, rather than generate knowledge claims, the distinction between these two applications is a fragile one. To give a greatly simplified example, if an algorithm relies upon a correlation between an individual's zip code (a common proxy for racial identity) and their likelihood of defaulting on a loan in order to determine whether they should receive a home loan, it is only a short cognitive jump to start believing that people living in that zip code (or of a certain race) do not repay loans. The problem is that more careful analysis might demonstrate that individuals in a particular zip code have more precarious work situations or access to fewer financial resources. These factors might better correlate with the likelihood of defaulting on a loan. While the practices of data science might already push away from creating a kind of theoretical explanation of these connections, data science in a pragmatic mode can be even more problematic. In an effort to meet deadlines and produce *something useful* for clients, data scientists might find the zip code to be an easier, more efficient way to reach effective predictions about loans. With time, data scientists who work on financial markets or loan officers "learn" that people from particular areas do not repay loans and therefore should not receive them. This suggests that while critical data studies have been concerned about the suggested dominance of technical expertise that is part of the epistemological landscape, the practice of data science in applied contexts may be even more troubling, reinforcing and generating knowledge claims that are not supported by the epistemological authority of any worldview, but are instead supported by an orientation that comes closer to a market or industrial logic.

Although the presence of and shift to a pragmatic approach in applied settings suggests that factors outside of the epistemological landscape also contribute to shaping knowledge production in some settings, the symbolic order of data science is still an important piece of understanding the societal consequences of data science. First, it is still factors into the construction of knowledge in applied settings although it may be secondary to pragmatic results. Second, as I will discuss in the following chapter, it is the contours of the epistemological landscape that are often presented to the public and to potential consumers of data science. It is through this presentation of the epistemological landscape that the broader public will learn about the capabilities, promises, and authority of data science. And it will be on these terms that they accept or reject the products of data science, no matter the processes by which they are produced.

CHAPTER 4

Data Metaphors: Going Deep for Objective Truths

"Without the application of data science, big data is no better than an inert pile of coal."
-Excerpt from ThreatTrack white paper.

"These things that are more advanced, they might do a better job [….] at predicting, because they go a little deeper."
-Denise, Data Scientist working in Health and Medicine

The ways in which we talk—our language, turns of phrase, and metaphors—tell us something about our culture and how we see and experience the world. Likewise, the ways in which people talk about data reveal and inform our cultural conceptions about the capabilities of data science and analytics. What does it communicate when data scientists equate data to coal? What do we hear when Denise says that better methods can go deeper? This discourse, or "data talk," as I call it, is as instructive as it is reflective. It provides the possible imaginaries through which people—especially novices, non-specialists, and the public—come to grasp data's abilities and role in generating knowledge. In other words, data talk reflects and shapes one version of the epistemological landscape of data.

In previous chapters, I have outlined some aspects of this landscape by focusing on the perspective of data scientists. Here, I use an exploration of the metaphors contained in data talk to reveal the possible implications of this talk when encountered by the public and those who are not trained in the methods of data science. Through an examination of the metaphors that infuse the discourse of data, I examine what data is and how it relates to truth claims according to the imagery provided in data talk.

Through an exploration of two different types of metaphor contained in data talk, as well as the language that surrounds these metaphors, I show that data talk reinforces ideas about the objectivity of data, obscures the existence of those who produced the data, and facilitates certain expectations about data ownership and use. In addition, I find that the metaphors of data depict a particular relationship between data and knowledge claims, one in which truth and knowledge lie in the details of the data itself. I conclude by arguing that, due to its potential to shape our broader culture and approach to knowledge, examining data talk is central to a sociological account of data science and a key to success in advancing protective data policies and the ethical orientation of data scientists.

Discourse and Metaphor as a Window into Culture and Epistemology

Sociologists frequently turn to the examination of discourse as a means of better understanding a group's culture and worldviews (Alexander and Smith 1993; Boltanski and Thevenot 2006; Olick 1999; Wagner-Pacifici 1994; Wuthnow 1987). In this chapter, I use discourse analysis with a special focus on metaphor to ascertain the epistemological landscape of data science. While recent work in the sociology of knowledge has focused on primarily on the analysis of practices to depict the epistemic cultures of various laboratories or professions (e.g., Knorr Cetina 1999), I argue that the epistemic landscape can be studied through the same methods by which sociologists study other beliefs and cultural constructs. Like other cultural sociologists, I begin by treating the discourse that surrounds data as analytically independent from other social phenomena (Alexander and Smith 1993). In other words, this chapter is not concerned with tracing the determinants of this discourse, historical variation, or spatial

variation. While these relationships are worth investigating, my intent in this chapter is to produce a thick description of the epistemic imaginary that is available in data talk.

I do this through an exploration of one aspect of discourse, that of metaphor. The philosopher Paul Ricoeur (1987) argues that metaphor and text contain possible ways of seeing the world. He argues that "texts speak of possible worlds and of possible ways of orientating oneself in those worlds" (ibid: 144). When people make meaning from text, they are "grasping of the world-propositions opened up by the nonostensive references of the text" (ibid:144). Although sociologists have been slow to take advantage of metaphor analysis, other disciplines have effectively used this approach to show how experiences and knowledge become structured and understood through familiar cultural constructs (e.g., Lakoff and Núñez 2000; Winter 2001). For example, Cohn (1987) uses metaphor analysis to show how the gendered and "technostrategic" language of defense professionals allows them to participate in plans for nuclear violence. Similarly, Martin (1991) shows how biological concepts are steeped in gendered metaphors. In examining biological texts on reproduction, she shows how sperm are described with masculine words that emphasize agency, while eggs are painted as passive participants in the reproductive process. Even when more recent medical research revealed the active role of the egg in fertilization, gendered metaphors persisted and the egg was described as dangerously aggressive, if any agency was given to it at all. In pointing out the reliance upon gender to structure other experiences such as defense strategy or biological concepts, works like these use metaphor analysis to bring our attention to unexamined assumptions or bias and show how language facilitates particular actions.

While I do use metaphor analysis to those ends, I also use metaphor to produce a thick description of the epistemological landscape that unfolds within data use and discourse. In doing

so, my work aligns with recent efforts in the sociology of culture. Abend's (2014) work on the moral background accomplishes a similar task. While Abend does not focus on metaphor exclusively, he shows that there is a "conceptual repertoire" available within a culture, and that more specific claims and beliefs are built upon these repertoires. They are the categories and concepts with which we think. While they are not determinant of particular beliefs or claims, they set the terms of what beliefs and claims become possible. Focusing on moral arguments and orientations of business ethics between 1850 and 1930, Abend argues that our actions are dependent upon these conceptual repertoires and shows how they can support varied—though not unlimited—moral orders. Similarly, Hochschild (2016) uses the metaphor of what she calls a "deep story" to show how some Americans conceive of their political, historical, and cultural context. Hochschild argues that it is this deep story that has facilitated recent political activity, including the election of Donald Trump. In examining the metaphor of data talk, I similarly attempt to bring to light the underlying world views and conceptions that structure and reflect our experience of data.

To accomplish this task, I draw on the literary and hermeneutic tradition of Ricoeur (1987). Ricoeur described metaphor as a modifier of a subject: it applies specific attributes to something else. When it comes to making meaning, Ricoeur argues that the meaning of metaphor, and text in general, is contained within the text itself and not the mind of the author. Instead, meaning unfolds in the interaction between the "the world that the work displays," and the world of the reader (ibid;144). The reader makes an interpretation by "grasping of the world-propositions opened up by the nonostensive references of the text (144)." In other words, the reader brings her life experience into the interpretive process of the making meaning from the text. I follow Ricoeur in this treatment of meaning. For the present case, this means that the

implications of data talk may differ between the data scientists and organizations that speak or write these words and those that hear or read them. In the analysis presented here, data scientists and the organizations that produce these white papers are the authors. It is important then to keep in mind the differing experiences and references that these authors may draw upon in creating meaning when compared to other groups. Non-specialists lack their own experience with data science practices and theories and therefore have little basis for countering or tempering the imaginaries provided by data talk. Similar patterns have been observed between accountants and statisticians. Though these groups use the same tools, they hold different world views and conceptions of reality—a factor that is related to their differing levels of familiarity with statistical processes (Desrosieres 2001). Once the words of data scientists and data-driven organizations are encountered the public or clients, the intentions or understandings of the data scientists matter little in how these non-specialists will construct the world form data talk. Given the unfamiliarity of non-specialists with the techniques of data science, the metaphors of data, which rely on familiar concepts that are already part of the reader's "mode of being" are especially powerful for shaping how non-specialists come to understand what data science is, what it can accomplish, how it relates to truth, and what constitutes just use of data science. It is this aspect of metaphor and meaning that provide data talk with the powerful epistemological implication that truth lies in the details of data points and the power to naturalize the ownership and use of data by those who have the resources to access and collect it.

While I use the theoretical lens of Ricouer's treatment of metaphor to ground my empirical argument, I also draw on a second area of metaphor studies that takes a more cognitive approach. While I am not engaged in the same theoretical and causal arguments as those scholars of metaphors within the cognitive tradition, I do find their typologies and analytical

concepts to be useful methodological tools.[15] Much like Ricouer, Lakoff and Johnson (1980/2003) treat metaphor as the practice of "understanding and experiencing one kind of thing in terms of another" (ibid: 5). They articulate this definition through the example of "argument is war." The language that we use to discuss argument reveals this metaphor: "His criticisms were *right on target*," "I've never *won* an argument with him," and "He *shot down* all of my arguments" (emphases original, ibid: 4). In this case, war serves as the "source domain," meaning that we understand argument according to the features associated with war. Therefore our experience of argument is one of conflict and adversaries. Metaphors have this effect, in part, because in framing an experience or phenomenon through a source domain, they are able to highlight some aspects of phenomena while hiding others. With regard to arguments, Lakoff and Johnson note that the war metaphor brings disagreement to the forefront, while obscuring the level of cooperation required for arguments to take place. In short, metaphors facilitate the experience of phenomena according to the dynamics, features, and expectations associated with something else with which we are familiar. This process tends to shape our experience as it hides some aspects of phenomena while stressing others.

In addition to the concept of the source domain, I borrow from Lakoff and Johnson's (1980/2003) typology of metaphors. To understand arguments as war is a *structural metaphor*, meaning that the language and practices of war structure the experience of arguing; the concepts of one phenomenon structure the other. *Ontological metaphors* structure phenomena as a type of

[15] Given their grounding in a cognitive perspective, scholars working in this tradition of metaphor analysis are engaged in a debate as to whether or not metaphors actually structure experience and action or simply are used in speech to reflect cultural constructs. In her study of American marriage, Quinn (1991) argues that the presence of a discrete set of metaphors used to describe marriage indicates that metaphor reflects underlying cultural models, rather than shaping those models. In contrast, Cohn (1987) argues explicitly that it is the language of the defense industry that produces a particular mind-set and therefore enables the activities of defense professionals. Like Abend (2014) and Hochschild (2016), I am less concerned with using this case to articulate the relationship between metaphor or language and the related cultural constructs. Instead, I use metaphor as a means to access aspects of culture that shape beliefs and action.

entity or substance. This may occur in a general sense, such as when *inflation* is discussed as an entity that may be quantified, compared, or the cause of various effects. However, ontological metaphors may also be more specific. Lakoff and Johnson (ibid) use the examples of the *mind is a machine* and *the mind is a brittle object.* These metaphors are expressed through phrases like, "I'm a little rusty today" and "her ego is very fragile," respectively (ibid: 27-28). Though these phrases reveal two different models of the mind that exist in our culture, they both equate the mind with another known entity. Lakoff and Johnson (ibid) also discuss *orientational metaphors*. *Orientational metaphors* organize systems of concepts in relation to each other. This organization is often spatial in nature. For example, we think of the future as lying *ahead* and the past as being *behind*. Phrases like, "I'm feeling *up* today, but she seems *down in the dumps*," suggest that we experience good as up and bad as down. These kinds of orientational metaphors do not just organize single concepts, but rather form a coherent system, meaning that there is consistency across expressions; for example, good is almost always associated with up and rarely associated with down.

Ontological Metaphors of Data

It is well documented that the language surrounding technology and science is suffused with metaphor (Ignatow 2003; Lombard 2005; Markham 2013; Ryall 2008). The discourse of data science is no different. As Puschmann and Burgess (2014) argue, the rhetoric and imaginary that surrounds data is still in a phase of interpretive flexibility (Pinch and Bjiker 1984). This means that metaphor may be especially important in analyzing the possible imaginaries of data at this point in time when both data science and its discourse are in a stage of

development. In addition, metaphors may be more powerful in structuring the experience of certain groups. As Gregg (2015) argues, leading companies in the technology industry do the work of selling data science to others and convincing potential clients of its value. In her words, these companies do the "rhetorical work" of "assembling the data spectacle" and establishing shared assumptions about the nature of data (Gregg 2015). These organizations occupy a powerful position in their ability to set the discourse by which data is discussed by both data scientists *and* non-specialists such as users of data-driven tools or the public.

With regard to big data, data analytics, and data science, scholars have focused primarily on the ontological metaphors and the aspects of data that they obscure or hide. For example, both Peters (2014) and Gregg (2015) note that the metaphor of cloud computing obscures the immense physical activities and on-the-ground infrastructures that are required to provide this kind of remote data storage and analysis. Peters (ibid) also makes the point that in achieving the widespread association between cloud-like imagery and remote computing and data storage, the IT industry has successfully drawn attention away from the risks associated with allowing these IT structures to control our data. Based on an analysis of the public discourse surrounding big data in popular newspapers, business journals, and several organizations' publications, Puschmann and Burgess (2014) argue that big data is often understood through two ontological metaphors that equate data with nature. The first, is the idea that big data is a natural force to be controlled, an aspect that is often depicted through terminology related to water.[16] For example, we see phrases like *there is a deluge of data* or *companies are drowning in data*. Puschmann and Burgess note that "the allusion to water supports the notion that data is all at once essential, valuable, difficult to control, and ubiquitous," but there may also be "the danger of 'torrents' of

[16] Metaphors equating water to data have also been discussed by Seaver (2015).

data in which one can ‘drown,’ ‘floods’ that overwhelm us, and ‘tsunamis’ that leave destruction in their wake,” (ibid: 1699). They argue that this data metaphor obscures the fact that data is not a natural discovery, but produced by people. Further, by suggesting that value is inherent in the data, this metaphor hides the processes by which data is massaged and interpreted into something that has value. The second metaphor that they discuss is that data is a natural resource to be consumed. This comes across most clearly in phrases such as *data drives* decisions or *data is fed* into the system. Puschmann and Burgess argue that this metaphor treats data as a commodity and frames the results of data analysis as resulting in inevitable, self-evident conclusions.

While the current work on data and metaphors has done a great deal to advance our understanding of the ways in which data are conceptualized and which aspects of data are obscured, the focus thus far has been primarily on *ontological metaphors*. Although orientational metaphors reveal a great deal about cultural conceptions of our world, the implications of these kinds of metaphors in data talk remain underexplored.

In what follows, I undertake two main tasks. The first is to expand upon the existing analysis of ontological metaphors. In doing so, I begin by examining additional terminology that supports the claim that data talk is saturated with nature metaphors and then move on to analyze metaphors of data talk that treat insights and findings as entities. While I complicate the findings of Puschmann and Burgess (2014) by pointing to the varying degrees by which different nature metaphors obscure or point to the role of human labor in obtaining value from data, I focus primarily on the shared aspects of these two metaphors. Specifically, I find that both nature metaphors and the insights-as-entities metaphor promote a notion of objectivity associated with data and set our expectations for data ownership and use. My second task is to examine the orientational metaphors contained in data talk. I argue that this discourse contains a spatial

mapping of data, one in which data is organized along vertical planes, with the details and individual data points located at lower levels than general knowledge. When combined with language that suggests truth also lies in the lower levels of this spatial plane, the resulting epistemological landscape is one in which truth lies in the details of the data. In the discussion and conclusion, I combine the metaphors to paint a picture of the epistemological landscape of data science. I then explore the degree to which this landscape represents the understanding and experience of data science for data scientists versus non-specialists before concluding with some possible consequences that result from comprehending data science through this lens.

Data and Methods:

The material for this analysis is drawn primarily from a content analysis of 33 business-to-business white papers that address data analytics.[17] I draw on interviews and ethnographic observations to provide additional instances of data talk that correspond to the data talk of the white papers.[18] These sources are particularly apt for assessing the epistemological landscape contained in data talk. While interviews and observations allow me to capture the ways in which data scientists talk and think about their work in practice, the white papers represent discourse through which data science is presented to outsiders and potential clients. This is a feature of business-to-business white papers which often function as a type of advertisement for potential clients and tend to avoid overly technical language. These leading companies in the technology industry do the work of selling data science to others and convincing potential clients of its

[17] A full list of white papers and their source organizations is included in Appendix B.
[18] For more information about the collection process of these white papers, see the data and methods section of the introductory chapter.

value. As such, they are in a position to define much of the discourse that surrounds data for both insiders and outsiders of the industry.

Although Lakoff and Johnson (1980/2003) provide a general theoretical model for metaphor, they do not offer an explicit analytical method for the social sciences. As of yet, there is no well-established method for integrating the analysis of metaphor into the sociology of culture. In an effort to develop the methods of the sociology of cognition, Ignatow offers a quantitative approach to assessing the metaphors of the high technology industry (2003) and shipyard workers (2004). His work on the technology industry relies upon counting metaphors and their features from a rather large corpus of lexicons over a 32 year period. In particular, Ignatow is interested in the presence of profane metaphors in the industry. From this analysis, Ignatow is able to show how the prevalence and nature of metaphors changed over time. This allows him to postulate some possible causes of these language shifts and the presence of profane metaphors. This method works quite well for Ignatow's purpose and research questions. However, this approach is less effective for a smaller, in-depth data set that contains a great deal more complexity than that found in lexicons. In addition, because he identifies his categories of metaphor in advance of the analysis (metaphors of the profane), Ignatow's method does not allow for the kind of analysis that interests me here: an unpacking of the epistemological landscape that constitutes the data imaginary.[19]

In order to use metaphor to this end, I relied upon a multi-step coding process. I read each interview, field note, and white paper, making note of metaphor use and expressions. It became clear that there were several metaphors that appeared repeatedly throughout the data. I

[19] In addition to this misfit between his method and my analytical goals, Ignatow (2014) explicitly distinguishes his cognitive approach to culture from the kind of cultural sociology or the "interpretive epistemic mode" (Reed 2011) upon which this analysis relies.

then returned to the documents a second time, coding for these specific metaphors. Although the metaphors discussed below are used throughout the data, I was not interested in their frequency as such. This is for several reasons. First, as Puschmann and Burgess (2014) argue, the rhetoric and imagery that surrounds data is still in a phase of interpretive flexibility (Pinch and Bjiker 1984). This means that data talk has not yet stabilized around one cohesive system, and therefore we would not expect any given metaphor to appear across all or most of the sample. Secondly, I am more interested in the contextual aspects of these metaphors and what their use can tell us about the way that data science is envisioned. This kind of analysis requires a reading of the text that surrounds each use of the metaphor rather than a more distanced view of the metaphor's use across documents. Therefore, once instances of particular metaphors were identified, I then examined the use of the metaphor in context in order to determine the implications of the metaphor. In addition, I followed Lakoff and Johnson (1980/2003) as well as others who have conducted a metaphor analysis (Martin 1991, Quinn 1991) by examining the source domain of each metaphor. In my analysis, I rely upon common cultural understandings of and associations with the sources domain to provide possible implications of the metaphors contained in data talk.

Deep Diving for Truth

Ontological Metaphors: Data as Objective and Free for the Taking

In this section, I focus on the ontological metaphors of data talk. This involves two parts. Puschmann and Burgess (2014) draw our attention the metaphor of data as a natural phenomenon. In this section, I reexamine these nature metaphors as well as metaphors that

frame the results of data analytics as entities and objects themselves. Puschmann and Burgess (ibid) find that data metaphors frame data as objective, imply that value is inherent in data, and obscure the human activity involved in producing data. My analysis complicates and expands upon these conclusions. The use of water-related terminology, especially when contrasted with the metaphors used to describe more traditional data processes, reinforces the claim that data is seen as a ubiquitous mass that takes form without human intervention. Metaphors that treat data analytics as entities similarly convey the objectivity found in water metaphors. However, another nature metaphor, that of data as oil, complicates the claim that data metaphors entirely obscure the human labor that goes into massaging and preparing data for analysis. Finally, all of the ontological metaphors that I analyze promote a particular understanding of data ownership and rights to usage.

Metaphors of Lakes and Silos

In recent years, the term *data lake* has begun to appear in the data discourse. Although there is no industry-wide standard for defining a *data lake* (Gartner, Inc 2014), the general idea is that data lakes are databases that store data in their initial form. In other words, data is not manipulated and translated into standardized formats ready for inquiry and analysis. This is thought to increase the amount of data to which analysts have access. This technique is often contrasted with *data silos*, a more traditional way of storing data in structured collections (ibid).

> Knowledgent White Paper: "This is where the *data lakes* come in. The massive environment will house data in its most raw form, giving analysts the option to format and standardize it when needed to make it machine readable and easy to use"

> Gov Loop White Paper: "That's why many are turning to a *data lake*, which is one big data *storage pool* that houses different forms of data"
>
> Booz Allen White Paper: "This high-speed analytic connection is done within the *Data Lake*, as opposed to older style sampling methods that could only make use of a narrow slice of the data. In order to understand what was in the lake, you had to bring the data out and study it. Now you can *dive into the lake*, bringing your analytics to the data"
>
> Gov Loop White Paper: "Agencies must break down legacy storage *infrastructure silos* while improving performance and increasing capacity"
>
> Booz Allen White Paper: "In the wake of the transformation, organizations face a stark choice: you can continue to build *data silos* and piece together disparate information or you can consolidate your data and *distill* answers."

Consistent with the findings of Puschmann and Burgess (2014), the notion of the data lake equates data with a natural resource and reinforces the idea that data in this form has not been manipulated or shaped by human hands. The quote from the Knowledgent white paper even refers to this data as *raw*, another indication that it is untouched by human intention, and suggests that data in this state is the most valuable. Further, this imagery is strengthened by comparing the new methods of data storage to the traditional methods to "siloing." To speak of data housed in silos brings forth images of farmed grains stored in literal silos. The matter contained in those silos, while natural in some sense, is also cultivated and stored through human labor. The metaphor of data silos forefronts the human activity necessary to produce the data. By moving to the metaphor of the data lake, data becomes a vast natural resource, one that exists regardless of human's efforts and stands ready for human domestication and use. However, as becomes evident from observing data scientists work with data and from research such as Gitelman (2013), the process of collection, recording, and storing data inherently involves human decisions, categorization, and manipulation. Despite this reality of data science, these

metaphors obscure this aspect of data analysis and facilitate an experience of data in which it is not produced by humans, but there for our consumption and use.

Data as Oil: Raw Data, Clean Data, and Drilling for Data.

In addition to describing data through water metaphors, data is sometimes referred to as another kind of natural resource, that of oil or coal. This metaphor has been occasionally made explicit, as in an article from *Forbes Magazine* that asked "Is Data the New Oil?" (Rotella 2012) or an editorial from Intel that claimed, "Just as oil has transformed our world over the last century, data is poised to transform our world for the next hundred years – and beyond," **(**Krzanich 2016**)**. In interviews and white papers, this metaphor was often more subtle. For example, just as oil must be refined before it is ready for use, data is often discussed as either *raw* or *clean.*

> Raph: "One of the first tickets that I got when I came onto Arpeggio was simplifying the way that the front-end took in data from the back-end. All the manipulation of the *raw data* was happening in the browser."
>
> Isaiah: "I think data science was sort of pursued as a field of study because it allows us to turn all of this *raw information* into a decision"
>
> Chris: "We usually just call it *clean, cleaned data* versus, we don't really call it dirty data, but *raw data* I guess."
>
> Mikey: "it starts out with like the data extract from pubmed, and then goes into the *data cleaning* and then, you know, into-into the actual like bottling, and then visualization."

The data as oil metaphor also becomes evident when data scientists refer to "drilling" into the data and "extracting" value or insights from it.

> Sienna: "You think about on the small package side, they've got scans of every single package that's been delivered. You can *drill down* into the most minute detail."
>
> Palentir Cyber White Paper: "Analysts can leverage a robust suite of analytical applications that enable organizations to triage alerts, *drill down* on the most critical ones, and quickly assess the extent of exposure."
>
> SAS White Paper: "Using systems that provide automated monitoring and alerting, predictive modeling, advanced analytics and reporting, and KPI dashboards with *drillable alerts*, they are containing maintenance costs and minimizing maintenance-related disruptions of their operations."
>
> IBM Analytics Paper: "Planning Analytics can *extract* data, metadata and security profiles for use in essential financial performance management processes"

This framing of data as oil overlaps with Puschmann and Burgess's assessment that data is often treated as a resource to be consumed. However, the source domain of oil contains additional implications for the data imaginary as well. While this language might still obscure the human effort involved in producing the original data and datasets, it also suggests that, much like a natural resource that must be purified and refined before it becomes something useful or valuable, data requires human intervention in order for it to have value. When described as a natural resource that must be laboriously removed from the earth and transformed by human processes into something valuable, the human activities of cleaning and massaging data enter the data imaginary and become part of the visible process of data analysis.

Together, the metaphor of data as oil compared to metaphors of data as water paint two slightly different pictures of data. To see data as oil indicates the need for human interaction before value can be derived from the data. In contrast, the image of a data lake obscures the human effort required to produce the data and suggests that data is more valuable in its supposedly raw, untouched form. Therefore, the degree to which value is perceived as inherent

in the data may fluctuate depending upon which natural metaphor is used. To suggest that data is akin to coal or oil does point to some of the human labor and efforts involved in transforming that data into something usable and useful.

Although the metaphors differ in the ways in which they hide or reveal human activity, they share other features that may lead to real implications for how this data is encountered, used, sold, and regulated. Both water and oil exist prior to the human effort to collect and use these materials. Even though the data as oil metaphor points to some of the human work required to get value from data, both metaphors contribute to the idea of data as objective by suggesting that it takes shape regardless of human subjectivities and interventions. There are also implications for ownership and data rights. Consider that with regard to minerals and oil, those that can access these resources usually have the right to use and sell them. In other words, the owner of the land has the right to whatever resources lie beneath and can be extracted from that access point, unless those rights have been sold or signed away to others. While modern water rights may have complicated the metaphor, water too, is generally available for use to those who can access it. In referring to data as a natural resource, we may reinforce the notion that any data an organization can access is fair game for the organization to use as it sees fit. Similar conceptions of data ownership and the objectivity of data science are expressed when data talk treats the results of data analytics as an entity. It is that analysis that I turn to next.

Findings and Insights of Data Analytics as Entities

While speaking of data as a natural resource equates it with a specific kind of phenomenon and the features of that source domain, data talk also treats the results of data collection and analysis as an entity in a more general sense. Like the nature metaphors, this

language reinforces a notion of objectivity in data science because it describes the insights and results of data science and data analysis as objects themselves. In addition, the language that surrounds and signals the use of this metaphor contributes to the particular ways in which data talk primes us to conceptualize the ownership, appropriate use, and nature of data.

The ways in which the results of data analysis are treated as an entity can be seen most clearly in phrases that refer to *hidden* insights and the work of *surfacing* or *discovering* these objects.

> Tera White Paper: "Big data analytics examines large data sets to *unearth hidden* patterns, trends, preferences, unanticipated correlations, and other useful actionable knowledge."
>
> Will: "Anything can repeat and can memorize a pattern and repeat it. What we try to get towards is some model that *uncovers hidden* relationships that aren't directly what we model on."
>
> SAS Predictive Analytics White Paper: "With the reality of big data, new techniques are being explored by companies to *leverage the value hidden* in new types of data. Being able to explore *all* of your data quickly and in an interactive manner is driving the need for data visualization techniques and interactive predictive modeling on very big amounts on data—fast."
>
> FICO White Paper: "Used properly, Big Data can help a business decide when to launch a new product, at what price and in which geographical regions. Or it can help *reveal previously hidden* risks associated with a loan or investment."
>
> IBM Analytics White Paper: "It then performs the statistical analysis to *uncover* the factors influencing or predictive of late payments. Once those factors are identified, further analysis may *reveal* process changes or improvements that could be made in those various influencing factors, such as bill type, whether paper based or electronic."
>
> Palentir Cyber White Paper: "Analysts can search across all data sources at once, visualize relationships, explore hypotheses, *discover unknown* connections, *surface previously hidden* patterns, and share insights with other teams."

This language suggests that the patterns, relationships, and risks articulated through data analysis exist prior to the process of analysis. They are not created or crafted. Rather, they sit out there in the world waiting for the data scientists to detect them. Further this language suggests that we lack the ability to recognize or make visible these findings through traditional methods.

The same notion of pre-existing findings and insights is communicated through a language of discovery.

> Sunera White Paper: "The interactivity of visual analytics opens the door to *discovering* trends, anomalies, opportunities, and causes and effects that may have been missed otherwise"
>
> IBM Analytics White Paper: "Users can *discover* new insights from their data automatically and apply these insights into plans, analyses and reports"

Much like referring to insights as *hidden* or *covered*, the language of discovery suggests that the results of data analysis exist independent of the work of data scientists. Data is not created or forged. Instead an existing object becomes visible and knowable. This aspect of data is further reinforced by language use present in my interviews. Interviewees described data that was not collected explicitly for scientific study as "found data," a turn of phrase that also indicates that the data scientist had little to do with its creation. This reinforces the idea of an objective truth that awaits detection.

To describe data science as a *discovering* process also has implications for how we conceptualize ownership of data. In stories of exploration, we often refer to the discovery of new lands. Much like the conception of truth that unfolds with data analytics, this land exists prior to the explorer's arrival. Importantly though, ownership of the discovered lands often goes to the

explorer or the group he represents (despite the fact that there may already be local inhabitants). One only need to think of the many examples of colonial expansion and images of Europeans planting flags on shores that were new to them. To discover a place meant to claim it. Again, the terminology of "found data" also suggests that there is not a current, identifiable owner, much in the way that we refer to the "lost and found" box. Much like metaphors of natural phenomena, the language of discovery may have significant implications for how we treat the results of data analytics. Despite the fact that in many instances such analysis is only possible due to the vast amounts of data produced by the activities of countless individuals, the language of discovery inclines us to consent to the discoverer's right to control the results of data analytics and use them as they see fit.

When we look holistically at the ontological metaphors of data talk, we can see that several metaphors—data as water, data as oil, and data analytics as an entity—converge on similar implications. Data *and* the resulting claims made through data collection and analysis are experienced as objective truths that exist independent of human intention and intervention. In addition, these metaphors have shared and powerful implications for data ownership; data belongs to the one who finds it, claims it, and uses it. The producer and the origin of data are obscured by these metaphors and are therefore less prominent aspects of the data imaginary. In attempting to resist this language and the ways in which corporations are able to take advantage of the data that people regularly produce, Gregg (2015) advocates for using the metaphor of "data sweat" to draw our attention to the way in which we are related to and produce data; it is undeniably tied to human activity and bodies, it may show up when we do not want it, and it leaves a trace of our behavior and mental state. Her efforts demonstrate one avenue through which we might try to overcome the assumptions contained in data talk. In addition to educating

the public and consumers of data analytics on the limits, constraints, and conditions of producing knowledge through data, the introduction of new terminology that highlights, rather than obscures, the problematic aspects of data science may prove useful.

Orientational Metaphors: Details and Truth below the Surface

The Vertical Organization of Data

In additional to the ontological metaphors already discussed, data talk is saturated with orientational metaphors. These metaphors organize concepts along spatial planes. In describing orientational metaphors, Lakoff and Johnson (1980/2003) give examples such as "good is up; bad is down" and "high status is up; low status is down" (16). Data talk contains similar metaphors that map the epistemological practices of data science onto a spatial orientation. I find the spatial aspects of the epistemological landscape are sometimes expressed through the ontological metaphors described above, but they can also clearly be seen through additional language use in data talk.

For example, in the data as oil metaphor, data scientists use the phrase "drill down" to describe the process of examining data closely. This phrasing may be used when data scientists refer to their own work, or it may appear when data scientists are describing the ways in which users of data-driven platforms are able to make use of their technology. For example, the medical analytics team that I observed developed what they called a "drill down feature" which allowed doctors to click on patients with high risk scores in order to access more details about those patients. Consider these additional examples:

> Matthew: "if there are problems, and they [his clients] *drill down*, and they say okay, there's that area, oh that's in my region, and that's red, let me *drill down further* and as they *drill down*, they end up with individual cases that are the

> highest priority, and they organize their work for the week to go and schedule business and things like that"
>
> Gov Loop White Paper: "By looking at a lot of data and *drilling down to the appropriate level*, we have reasonable assurance that we've exhausted what we can do, and we can base any conclusions on the work that we've done"
>
> PepperSack Big Data White Paper: "Data science can be applied to allow you to *drill down* into the effects of technical issues on your business"
> Adam: "to a certain extent, I think yes, you do better science if you look at all the data first and see where it's pointing you, and then *drill down* into that"
>
> Alteryx White Paper: "Armed with a better understanding of their wireless network's dynamics, providers can use analytics to *drill down* to the individual subscriber level and discover trends that can help them introduce innovative new services."

Each of these examples suggest that data and knowledge are oriented along a vertical plane. Some information lies *below the surface* and requires that the data scientist move down to access it. In addition, note the frequent use of the word "level" to describe data, indicating that within the vertical space, data are organized and grouped in stackable planes. Data and the phenomena they represent are organized on levels. Further, these quotes indicate that an initial analysis of the surface can tell you where to look closer for more insight. This suggests that the makeup of the data that sits below the surface directly shapes the contours of the surface that data scientists can discern through their analytical techniques. Many of these quotes also indicate that in drilling down, analysts will discover more nuanced details that allow them to better assess the situation. You can see this specifically in the quotes from Matthew and the Alteryx white paper that associate this lower level of data with information about individual cases, clients, or data points. So in a very simplified sense, this metaphor suggests that the details and nuance are down and that a general overview, or surface view, is up. This is consistent with already familiar

language in research and academic efforts in which we conceptualize individual details as "closer to the ground" and general concepts or theories as "abstracted up."

A similar spatial orientation is present in other language of data talk. Data scientists and firms frequently refer to *deep* or *in-depth* analysis, and this further reinforces the vertical organization of data and the benefits of examining the lower levels. For example:

> Isaiah: Where you're using data analysis to detect an anomaly, so every single event you're watching, say, hey this is out of the ordinary. And then you sort of *dive deeper* into that and do some more analysis.
>
> Teradata White Paper: "Today, in the era of big data, private industry is catching on— leveraging data to channel its efforts and influence its customers, and signaling a crucial opportunity for government to step up its game. And one way government can achieve this is by starting *to look deeper* into its treasure trove of documents.
>
> Palentir Cyber: "Palantir Cyber leverages advanced detection and alert enrichment technologies, allows analysts to seamlessly pivot from detection to *deep-dive investigations* to reduce incident response time, and captures analyst insights to enable organizations to harden their defenses, providing a holistic, end-to-end cyber solution.

In these examples, the spatial metaphor is consistent with the idea of drilling down into data. They suggest that an initial analysis can tell you where to look more closely and suggest that the benefit lies in the ability to view the details of a dataset.

Truth is Located in the Depths of Data

However, the same terminology may also be employed in ways that suggest additional aspects of the epistemological landscape of data science. In the following examples, *deep* is

associated with what data scientists consider newer and better techniques. This language suggests that truer claims are often found deeper in the data:

> Oracle White Paper: "In considering all the components of a big data platform, it is important to remember that the end goal is to easily integrate your big data with your enterprise data to allow you to conduct *deep analytics* on the combined data set"
>
> Denise: "These things that are more advanced, they might do a better job [….] at predicting, because they *go a little deeper*. They solve a lot of problems that you can get from other models, from those more simple models, but you can't explain them"
>
> IBM Cognitive Systems White Paper: "When accuracy is needed over precision, we use *deep natural language processing*, or *deep NLP*, that analyzes context in evaluating a question. Watson is a deep NLP system that assesses as much context as possible that it derives from immediate information, from more broadly available information, from the knowledge base (called a *corpus*), and from source databases (emphasis original).

While these quotes share the use of the spatial metaphor with the notion of "drilling down," they communicate something a bit different. First, it should be noted that these quotes suggest that *deep* is associated with something good and presented as better than alternative types of analysis. Further, in these quotes, "deep analytics" indicates particular kinds of methods and techniques that promise to bring more insight or value. Denise's words perhaps lay this out most clearly. In discussing certain techniques, she claims that they are "better" because they go "deeper," allowing her to better solve the problems she is tasked with solving as a data scientists. This spatial metaphor also contains the suggestion that deeper analysis will get the data scientist closer to the truth or valuable information. While this is implicit in most cases, it is occasionally an explicit claim made by the advocates of data science. For example, the authors of the IBM paper quoted above compare the *deep natural language processes* that they are promoting with

alternative "*shallow natural language processing*, […], which can be precise within narrow confines, but it is often not accurate" (emphasis original). This is significant because, in the terminology of data science, accuracy indicates proximity to the true answer, while precision indicates the reproducibility or consistency of results. This statement is intended to communicate that the *deep* techniques will be more likely to discern the truth. This suggests that in the epistemological landscape of data science, the truth lies in the depths.

The Epistemological Landscape of Data Talk

What is the epistemological landscape that unfolds within data talk? What are the features of data and its relationship to truth that become apparent in light of both the ontological and orientational metaphors used to describe data science? This landscape is saturated with data. Data exists prior to human analysis and intervention. Therefore, it is objective and a strong representative of the truth. Though it is unclear if data is best in its untouched state or after human refinement processes, it is surely valuable in some form. Data belongs to no one in particular. If it can be found, accessed, or collected, those that obtain it may use it as they see fit.

Data is organized on vertical levels that have relationships to various aspects of knowledge. Details and individual observations or facts reside on lower levels that shape the contours of the surface. The surface level provides an overview of the world that the data scientist tries to know and therefore can point the data scientist to useful places to investigate individual level information.

In addition, the best methods of data science allow the analyst to produce claims and insights that are closer to the truth than the results of other, often older, methods. Curiously, this

is possible because these new techniques are able to open up levels of information that are hard to access because they reside on a lower level than previously accessible. Given that both the details and the objective truth lie in the depths of data, truth itself can be found in the details of data.

Placing Truth in the Details

This is the epistemological landscape contained in the talk of data scientists and data analytics firms. I have suggested that it contains a somewhat counterintuitive conclusion that objective truth lies in the details of the data. This assessment is a consequence of interpreting the vertical organization of data in combination with the metaphors that place truth in the depths. However, it presents a puzzle when considering the actual practices of data science. The advanced techniques that are associated with deeper insights and analysis are often the very techniques used to produce that surface, overview assessment of the data. It is through these techniques that an analyst is clued into the areas where she needs to *dive deeper* or *drill down.* How can both the details and the general overview be in the depths? In addition, we usually think of the patterns produced by machine learning and statistical analysis as abstracting *up and away* to truth. This is the method for producing general knowledge claims. How can it be that both the details and the general truth of large scale patterns lie beneath?

This becomes less of a puzzle when considering the current state of data science discourse and the conditions under which non-specialists encounter data talk. Given the state of interpretive flexibility in which the data science industry and discourse currently exists, it may be that the vertical organization and truth in the depths metaphors represent two competing models of data science, one of which may eventually win out. Or, consistent with Lakoff and Johnson's

(1980/2003) observation that there are multiple models of the mind in our culture, it may be that they represent two models of data science. In that case, practitioners of data science most likely move between these two models depending on their task at hand.

Regardless, I am more interested in the ways in which this aspect of data talk may shape the data imaginary of the non-specialists, users, the public, and future data scientists and less interested in the degree to which these combined metaphors represent the actual mindset of current data scientists. For non-specialists, users, and the public, it is primarily data talk that will shape their conceptions of data's capabilities and its relation to truth. Much like Desrosieres' (2001) accountants, these groups may not have exposure to the actual practices or theories of data science that allow them to distinguish between these two models or to cultivate an understanding of data science that is less defined by these metaphors. This lack of experience, along with the fact that many are new to data science and its capabilities, may make them more likely to combine or conflate the various metaphors as they conceptualize data science and its relation to truth. For these groups, data talk paints an epistemological landscape where objective data correlates to truth and it is the details of that data, more so than overarching theories derived from the data, that allow for the formation of truth claims.

This assessment is of particular consequence in light of recent epistemological debates on the role of theory that have unfolded among data scientists and their critics. Recall that in his argument for the end of theory, Chris Anderson (2008) proclaimed that big data would make theories and models—aspects of the epistemological landscape usually perched abstracted *up* and *away* from the details—obsolete for scientific inquiry. Instead of reliance on models, he claimed:

> There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

As I discussed in the introduction, scholars responded with objections to this view and made a case for the continued role of theory and models in knowledge production (e.g., Bowker 2014, boyd and Crawford 2012). And indeed, as I detail in chapter 2, it is clear that data scientists still respect domain experts and the theories of fields related to their projects at hand. Though they are endowed with lesser epistemic authority, the opinions and knowledge of domain experts ranging from nurses and doctors to marketing professionals to military specialists were factors in many of the projects I discussed with data scientists. Nevertheless, the language used in data talk lays the groundwork for a particular understanding of the world and how to know it. It obscures aspects of knowledge production—namely the role of theories and experts in producing claims through data science—and instead suggests that access to the truth lies solely in the details of the data. This is not simply a view that is promoted explicitly through the arguments of data science advocates. The possibility of experiencing the world in this way is built into the very language of data talk. Although additional research is needed to examine the degree to which non-specialists actually understand the world and data science in these terms, the point that I want to make is that the metaphors of data talk are primed to facilitate this particular kind of world view and epistemological landscape.

Life in this Landscape

Unpacking the contours of this landscape is an important part of formulating cultural theories about data science as well as shaping future discourse and policies of data science. With regard to the sociology of knowledge, we can see that language and imagery—not just the processes, practices, and resulting decisions of science—are important aspects of the ways in which it may shape us socially and culturally. The same is true of data science. The implicit assumptions of data science contained within the metaphors explored here are poised to shape our culture's approach to knowledge production, decisions, and the justification of various claims and actions. As I discuss in the introductory chapter, this is due in part to the growing influence of data science as it becomes institutionalized through government programs, its use in both the profit and non-profit sectors, and the establishment of educational and research organizations dedicated to data science. While data scientists themselves may have their education and experience to contribute to their understanding of data science, for most of the public, data talk is the primary conduit through which they will experience data and its capabilities. As this discourse becomes increasingly common, it may also become the primary way in which we understand knowledge production in general. As such, we need to take the possible readings of data talk seriously.

Data talk contains a particular depiction of truth and a prescription for accessing that truth. As already discussed, it is an objective truth protected from human subjectivities and interference. The key to knowing this truth lies in looking closely at the available data points—which are understood as direct correlates to this objective truth—with the tools and techniques of data science. While the notion of an objective truth may not be a new addition to our modern cultural fabric, we should be vigilant to the consequences of aligning that notion of truth so

closely to the bits of data captured and stored in the databases of IT companies, social media companies, government agencies, educational and health organizations and the like. Each of these organizations structure databases and collection processes to their needs and according to particular cultural categories and frames (Gitelman and Jackson 2013). Even if we were to accept the existence of an objective truth severed from human meaning-making and investigation, the building blocks of data science do not seamlessly correlate to such a truth. Further, we should be wary of the ways in which this language obscures important aspects of knowledge production. It is possible that in associating truth with detailed data and hiding the role of domain experts and specialized knowledge, the metaphors of data talk could contribute to broader cultural trends toward devaluing the role of expert knowledge (Collins 2007). As our cultural understandings of knowledge and truth are increasingly informed by the language of data science, this perceived void of domain specialists may make it difficult to challenge the conclusions and decisions that are derived from the seemingly authoritative—though imperfect—databases and analytic techniques.

In addition, the epistemological landscape of data talk primes us for particular uses and treatment of data. Data talk is infused with metaphors that obscure issues of data ownership and rights. Debates on data ownership and use have focused on issues of privacy (Crawford and Schultz 2014; Kerr and Earle 2013), informed consent (boyd 2014), and contextual integrity (Barocas and Nissenbaum 2014). While these are indeed relevant issues to this conversation, the language of data science may be working against those efforts. In framing the collection, storage, and analysis of data as an unproblematic practice akin to finding and keeping a leaf when walking along a nature trail, the language of data science may make it difficult to challenge the practices of data scientists and analytics firms. The public and data scientists alike may

simply not see the problem, making it difficult to gain traction for these issues. This may be especially important for efforts aimed at instilling an ethical and moral code in the minds of data scientists (e.g., Tijerina 2016). In pointing out the way in which their language obscures issues of data ownership or in actively working to reshape this language, cultural sociologists may be better poised to encourage the development and employment of ethical practices in the work of data scientists.

CHAPTER 5

Data Science in Action: Users and Predictive Data in the Neonatal Intensive Care Unit

As I discuss in the introduction, researchers have increasingly recognized the need to investigate algorithmic and data-driven technology, practices, and culture (Beer 2015; Gillespie 2014; Striphas 2015). Due to the powerful position of algorithms in decision-making processes, many scholars have drawn our attention to their social construction (Anderson 2012; Gitelman 2013) and emphasized the importance of unpacking how algorithms work by opening up the "black boxes" through which algorithms turn data into knowledge and decisions (Pasquale 2015). Here I expand the conversation and suggest that understanding the way in which users make sense of algorithmic output is as important as the affordances contained within the technology itself. Drawing on interviews and observations of developers and clinicians, I explore this aspect of data and algorithms through a case study of predictive data analytics employed in the medical environment.[20]

Clinicians at Augustine University Hospital (AUH) use Horizon, a data-driven monitor that predicts the chances that an individual patient will develop an infection, to make treatment decisions. The high stakes of life and death decision-making combined with the uncertainty that stems from the complexity of corporeal illness makes the life-saving potential of predictive algorithms quite appealing to these clinicians. These conditions, the practical setting of the hospital, and the limited familiarity of clinicians with algorithmic design, make AUH a rich setting for studying data analytics in knowledge production. I examine the practices surrounding Horizon and ask how clinicians make sense of this technology and use it to construct knowledge

[20] To protect the confidentiality of respondents, the names of new technology, individuals, and locations have been replaced by pseudonyms.

of patients. In the neonatal intensive care unit (NICU) that I observed, clinicians work to detect illness in the tiniest of patients. To ease this task, the developers designed Horizon to work as an early warning system, capable of detecting illness in advance of typical symptoms. My analysis shows that, far from providing stand-alone deterministic indications of health, Horizon is integrated into the established practices of the medical setting. This may lead it to be discounted altogether or to be considered in relation to other signs. This interpretation of Horizon unfolds amidst a negotiation between experience and intuition, on the one hand, and a doctrine of evidence-based medicine on the other. These factors lead to a set of interpretive processes that I call *conditioned reading* and *accumulative reading*.

Considering users of data and algorithms

Data analytics, big data, and algorithms have become increasingly ubiquitous parts of our lives (Schäfer & van Es 2017). As a data-driven and predictive tool, the study of Horizon offers insights into processes of quantification, datafication (van Dijck 2014), and the epistemologies that accompany these processes[21]. There is a growing effort in the medical community to increase the use of data and data analytics in clinical decision-making. Inspired by the plethora of health information collected by electronic medical records and concerns over rising health costs (e.g., Manyika et al 2011), engineers and data scientists have devised a number of health tracking tools for both personal and hospital use. For example, companies like Jvion and Castlight offer predictive analytics that claim to determine which patients are most critical within a hospital unit

[21] van Dijck uses the term datafication to refer to the collection and analysis of internet data. I use it here in more broadly to speak to the collection and analysis of data in general. As van Dijck indicates, "Datafication as a legitimate means to *access*, *understand* and *monitor* people's behavior is becoming a leading principle, not just amongst techno-adepts, but also amongst scholars who see datafication as a revolutionary research opportunity to investigate human conduct" (2014:198). I use there term here to signal this approach to knowledge.

or analyze personal communications data to predict health conditions of individuals and prompt them into corrective action. Enthusiasm for applying data analytics to health care has led to a number of efforts to increase the availability of health data. One such example is the National Institutes of Health's Big Data to Knowledge Initiative which works on advancing the data ecosystem, developing data-driven tools, and making data more available for research and development.[22] Programs such are these are likely to increase the presence of predictive algorithms in medical settings.

In attempting to grasp the implications of datafication, scholars have begun to outline aspects of their production, use, and effects. For example, despite their perceived association with objectivity, it is well-recognized that data and algorithms are socially constructed, sometimes containing aspects of the social imaginary held by their creators (e.g., Gitelman 2013). Once unleashed on the world, data and algorithms can impart significant consequences, shaping public discourse (Couldry & Turow 2014; Gillespie 2014), formations of the self and identity (Cheney-Lippold 2011), organizational activities (Ribes & Jackson 2013), and structures (Andrejevic 2014). Central to this research agenda is the methodological practice of opening up the "black boxes" of algorithms and data to peer at their inner workings (Pasquale 2015). In outlining their makeup, researchers aim to better grasp how algorithms make decisions and shape the world in which they operate (Beer 2016).

However, there is also an interest in unpacking the landscape of meaning that encourages and unfolds within the use of data (e.g., boyd & Crawford 2012; Beer 2016; Dalton, Taylor, & Thatcher 2016). How can we account for the appeal of algorithms and data? What kind of knowledge, world views, and selves unfold in their use? Will data practices reduce what it

[22] See https://datascience.nih.gov/bd2k/about

means "to know" (Bowker, 2014)? Beer has suggested that data and algorithms are attached to cultural conceptions of objectivity and truth and "evoked as a part of broader rationalities and ways of seeing the world" (2016:7). Like scholarship on quantification (e.g., Espeland & Stevens 2008), this suggestion rejects a narrative that links the usefulness of analytics purely to technical ends. The power and appeal of numbers reside in their symbolism and association with objectivity, their ability to standardize, and their increasing association with accountability (Espeland & Vannebo 2007; Porter 1995). Similar dynamics are likely at play with regard to data and algorithms.

Currently, much of the research that approaches these cultural aspects of datafication focuses on public discourse and rhetoric (e.g., Puschmann & Burgess 2014), organizational or institutional dynamics (e.g., Espeland & Sauder 2009), or analysis of quantified or algorithmic objects (e.g., Bucher 2014) without looking at "how people sense and make sense of data," in practice (Pink et al. 2016). There are at least two reasons for scholars to look closely at the practices of users. The first is related to the empirical conditions through which data analytics influence outcomes. While some algorithms have an automated effect without further human interaction—think for example of the automated way in which a search algorithm alters the results presented on your computer screen—this is not true in all applications of data analytics and algorithms. Many require human interaction and interpretation before they can be transformed into actions and consequences. Horizon is one such example. There is no predetermined action that occurs in response to Horizon's analytics. Clinicians must decide how to react to these predictions. In cases such as these, a consideration of users is central to an understanding of the ways by which algorithms and data impart broader consequences. Secondly, studying users is central to understanding the epistemological and cultural aspects of

datafication. In addition to influencing formal knowledge, data and algorithms are becoming part of the process by which actors actively construct the realities in which they live. This role of datafication goes missed when viewed only from an organizational or discourse perspective. Capturing these kinds of everyday epistemological dynamics requires an analysis of contexts and interactions through which knowledge about the world is formed (Garfinkel 1967). In the case of data and algorithms, this includes a consideration of the ways in which users integrate the products of data analytics into their conceptions of reality and truth.

Early research on the users of data analytics and algorithms is beginning to bear important insights. Consistent with the growing literature on users in science and technology studies (Oudshoorn & Pinch 2005), it is clear that the users of data and analytics do not always make sense of and use analytics in ways intended by their designers. For example, Nafus and Sherman (2014) show how members of the Quantified Self movement exercise "soft resistance" by altering and shifting the categories handed down to them by big data apps and technology. Similarly, a few studies on data, analytics, and professions suggest that, much like in formal research settings (Knorr-Cetina 1999), local epistemological contexts greatly influence the interpretation of data (Parasie 2015). Fiore-Gartland and Neff (2015) suggest that even within a shared industry, people approach data with a variety of "data valences," or expectations and values that surround data, that can lead people to understand data as "materially different things" (p.148).

This study contributes to this growing research area in two ways. First, this case provides an additional context, that of life and death decision-making in medical practice, through which to analyze the role of data and algorithms in the meaning-making process. Second, I make an effort to identify some generalizable patterns that may occur in other, nonmedical settings. How

do the users make sense of data analytics or use them to construct knowledge? How are these analytics interpreted in relation to and reconciled with other signs? What features of the environment, organization, or culture are salient for structuring the process of interpretation?

In what follows, I give analytical attention to the ways in which users interpret and, by extension, mediate the products of data-driven knowledge. I do so by illustrating the ways in which users integrate predictive data analytics with the local environmental, practical, and contexts of the neonatal intensive care unit (NICU) at Eastern University Hospital. I begin by describing the developers of Horizon, their conception of the role that analytics play in medicine, and their intended use for Horizon. I then provide the environmental, practical, and cultural context of the intensive care unit where Horizon is deployed. These two sections lay the ground work for viewing Horizon's use in practice. Horizon factors into clinical practice in more complex ways that the developers imagined. This is due, in part, to the contexts in which is embedded. I conclude by suggesting some possibilities for why this context matters for interpretation and outlining some additional research questions called forth by the case of predictive medical analytics.

***Horizon's Origins and Purpose*:**

I was first introduced to Horizon in 2014 at a seminar on data analytics and ethics. Pat Brine, Horizon's lead statistician, gave a presentation in which he discussed the life-saving potential of data analytics and the hurdles that researchers face due to policies like HIPAA and

other institutional barriers to collecting and sharing data.[23] In his presentation, we learned that the developers of Horizon believe it to be a powerful tool for saving the lives of premature and low-birth-weight infants because it functions as an advanced warning system, predicting the onset of illness *before* other symptoms become visible to clinicians. This belief is rooted in the success of a randomized control trial in which Horizon was deployed across several hospitals and several thousand patients. The patient group monitored by Horizon demonstrated a greater than 20% reduction in mortality—a startlingly successful outcome. The predictive nature of this data-driven technology, its translation into the practical setting of the intensive care unit, and its daily use by professionals who may have limited training in statistics, algorithmic design, and data analytics make it an intriguing lens through which to examine the interplay between data analytics and knowledge.

At Pat's invitation, in 2015 I began attending weekly meetings of the medical analytics team in as they worked to extend Horizon's capabilities to adult populations. This opportunity allowed me to learn about the ways that the developers of medical analytics think about their work, approach problems, and conceptualize the role of large data sets and data analytics in medical contexts. After navigating my way through the maze of a large medical complex on the AUH campus, I arrived for my first meeting to find a small group sitting in a stately conference room with plush carpet and large portraits of serious-looking men lining the walls. The session began with introductions. Dr. Ibez, a cardiologist by training and the leader of the group, started off. As we went around the stable, I was struck by the professional diversity in the room. In addition to Pat, there was also a professor of systems engineering, several nurses, medical

[23] Health Insurance Portability and Accountability Act of 1996 (HIPAA) includes regulations for the ways in which health care information can be used and transferred. Many of the data scientists who work in medicine and healthcare see it as an obstacle to collecting, storing, and using data that could advance medical analytics.

students, a medical fellow, a surgeon, consultants and a graphic designer from a local data analytics firm, and several data warehouse specialists. Each week over lunch, the team would explain their purpose and goals to any guests, discuss challenges they were experiencing, formulate plans for the coming week, and review publication and conference materials that they were preparing for submission. It is through these conversations that I came to learn about how the group views the role of data analytics in medicine.

Analytics as Lifesaving

The explicit goal of the medical analytics team is to improve patient care and reduce mortality. As Dr. Ibez put it, we have "a single objective: to save lives," and "we do that through predictive monitoring." While there may be other motivations, such as the prestige that comes from designing and implementing new medical interventions, this mission to save lives is a sincere one. The physicians in the medical analytics groups often told stories and lamented over what they saw as preventable deaths. In these instances, the doctors understand the patient's death as resulting from missed signs; the patient was sick, but no one noticed. A key example of this belief in insufficient data or failure to properly interpret data as a cause of death comes in the form of a story that Dr. Ibez and Dr. March would frequently tell to visitors as a way of introducing their work to apply predictive analytics to patient care. With frustration in his voice, Dr. March would tell the story as follows:

> Several years ago, a police officer was shot during a domestic dispute. He was in rough shape and not expected to live, but he was miraculously saved by a team of four surgeons. During his recovery, he spent one month in the intensive care unit.

> Once he improved, he was sent to the floor [a section of the hospital that houses healthier patients and where patients are not monitored as closely]. One night, the internist—who was not an expert or upper level doctor—was called to his bed side twice in one shift. His condition worsened, and in the morning he was sent back up to the intensive care unit where he died a few hours later.

Instances like this clearly disturbed the clinicians involved in Horizon's development. They believed that, with the proper training, protocols, and technology any patient who made it to the hospital should survive. Stories like this reflected the perceived shortcomings in the hospital, all of which relate the death to a lack of information or the inability to accurately assess information. First, patients on the floor are not monitored by technology as closely; they may not be attached to electronic monitoring systems that automatically alarm when a patient's vital signs are in distress. In addition, there are fewer nurses and physicians per patient on the floor. Where each nurse in the ICU is assigned to just two patients, he may have many more to keep track of on the floor. While the team worked to encourage that continuous monitoring of vitals be extended to all hospital beds (a move that would enable them to expand their tracking, collection, and analysis of patient data), they also saw their work in medical analytics as being able to provide knowledge about deteriorating patients to the divided attention of medical staff who care for multiple patients simultaneously. Finally, in telling the story, Dr. March would often imply that the response of the inexperienced internist was insufficient. He did not recognize how sick the patient was and failed to respond soon enough. This occurred despite being called to the bedside by the nurse. After learning of the police officer's death, Dr. March started to believe that analytics could compensate for this lack

of experience: “I thought to myself, I wonder if an expert system could monitor our patients. Could the mind of the expert be put in the mind of the novice? Could we construct some objective summary of physiology?” Dr. March was always hopeful that data analytics could allow “the novice to have the awareness that the patient may be sicker that they realize.” Believing that a lack of information or inability to make sense of that information was linked to unnecessary deaths, the analytics team tried to remedy these problems by developing algorithms and monitoring systems that would produce risk scores. These scores are intended to alert the medical staff to quickly assess which patients were likely to worsen and needed the most attention, an approach that Dr. March referred to as “electronic watchdogs”.

The efforts of the medical analytics team reveal a particular understanding of a patient’s condition, how to construct knowledge about their condition, and the obstacles to constructing that knowledge. First, the developers and their approach indicate that there is a single, accurate assessment of a patient’s condition and that this assessment can be captured and depicted through quantified measurements of physiology and algorithms. Another story helps to reveal this belief. During a meeting where the team was discussing hospital performance metrics and potential new projects, Dr. Osina exclaimed, “What’s the *true model* of why people die?” He was suggesting that developing such a model might be a project for the team to tackle. His emphasis of the word “true” suggests that there is one explanation for death. Secondly, Dr. Osina’s exclamation also reveals that we are capable of grasping this explanation through a “model,” an approach that requires translating information into a quantified form and then integrating it. If this is true, then the obstacles for knowing the *true* cause of death or health problems lies in a lack of information, inaccurate information, or failure to properly integrate that information. Comments like this one,

as well as the group's efforts to produce algorithms that could translate vitals into quantified representations of patient health, paint the clinician as a person tasked with identifying an objective cause of illness through proper attention to metrics and proper recognition of their *true* meaning.

From the perspective of the developers, the primary obstacles to accomplishing this task and eliminating preventable deaths are two fold; there may be insufficient information or the medical practitioner may fail to appropriately recognize the information. This failure of recognition may be caused by multiple demands on the attention of the clinician or due to lack of experience. For these reasons, the developers and clinicians who participate in the medical analytics team believe that lives can be saved by providing medical staff with more information. It is important to note that this is not the only response available to remedy preventable deaths. The analytics team rarely discussed solutions such as getting better equipment, hiring more nurses to work on the floor, reorganizing the hospitals shift schedule, or reconsidering which kinds of doctors or teams are assigned during various hours. Of course, these solutions might seem out of the realm of an *analytics* team. However, their same efforts toward data collection and analysis could very well be applied to other efforts such as identifying the combination of clinicians that leads to the lowest mortality rate. Instead, they focused their efforts on projects that they felt would help the doctors to know their patients and their condition better. This may involve making previously inaccessible information readily available or producing summative scores of physiology that clue the physician into developing problems. Horizon combines these approaches and adds a predictive component.

How Horizon Solves a Knowledge Problem

Perhaps more than any other areas of the hospital, the NICU is saturated with knowledge problems. Distinguishing between signs of illness and the more benign symptoms of premature infants is especially challenging. As Dr. Walker shared with me, symptoms that might be concerning in older patients, are often just "standard things for premature babies." Added to that uncertainty is the problem that, unlike most adult patients, the infants are not able to communicate how they feel or changes in their health status (at least not through explicit speech). In addition, research on neonates is relatively undeveloped compared to other areas of medical research. This means in contrast to other areas of medicine, relatively few studies or evidence exist upon which to base care decisions and practices. Each of these conditions adds a layer of obscurity in the clinicians' efforts to "know" if their patients are well or sick.

Sepsis presents an additional problem. Sepsis is an infection that enters the blood stream and causes inflammation throughout the body. It is quite serious. In 2009, sepsis was among the top 10 causes of death for both infants and adults in the United States (Kochanek et al. 2011), and according to the hospital's Quality Assurance Unit, sepsis accounts for 40% of all deaths at AUH. Once it is too late to save a patient, the presence of sepsis is obvious. However, detecting it earlier is often difficult.[24] As a syndrome, no gold standard for diagnosing sepsis exists. However, in practice, physicians hope to detect sepsis through a positive blood culture of the underlying infection. This functions as the gold standard for detecting sepsis in practice. This test requires that blood be drawn from the baby. Lab technicians then wait to see if bacteria

[24] Detecting sepsis has been a pernicious problem for the medical community in recent years. In addition to struggling to detect and treat sepsis, the medical community often disagrees on how to define the syndrome. This leads to discrepancies in how various studies of sepsis operationalize the syndrome. In other words, what "counts" as sepsis varies by studies. This has led to a series of conferences aimed at revising and standardizing the sepsis definition.

grows from the blood sample. A positive test indicates that bacteria has been detected. The problem is that this test may take several days to produce results. At that point, it may be too late to treat the patient. In addition, clinicians recognize what they call "culture negative sepsis." In this case, patients are believed to be septic even when the test says otherwise. This means that clinicians must make decisions about whether or not a patient may have sepsis without their preferred standard of proof.

Horizon purports to ease this problem. As the promotional material for the data-driven monitor make clear, Horizon's value is in its predictive capability. While typical monitoring methods indicate that a patient is "*currently* deteriorating," Horizon signals problems "*prior*" to symptoms and acts as an *"early warning* of patient deterioration" (emphasis original). The developers began working on Horizon in the early 2000s. For four years, the team carefully recorded the vital signs of each infant in two different hospitals' neonatal units. They matched this data with recorded cases of sepsis. While exploring the data, the team honed in on aspects of the heart rate as the most predictive sign of sepsis. They noticed that the infant's heart rate variability, or the amount of change in the time between each heartbeat, was linked to the onset of illness. Although patient heart rates are monitored in ICU settings, there is no monitor or display that measures and reports the degree of variability in in patient heart rates. Horizon works by translating the patterns in an infant's heart beat into predictions about sepsis. This can be thought of as a two-step process. First, it measures heartrate variability, a piece of information otherwise unavailable through typical monitoring methods and technology. Given that the doctor's and developers believe that insufficient information may be a cause of death, Horizon is extremely valuable precisely because it accesses information previously hidden to the clinician. Secondly, Horizon uses an algorithm to transform that data into a risk score for sepsis.

These predictions take the form of a *risk score* that presents the likelihood that a baby will develop sepsis within the next 24 hours on a scale from 0-7. Each increase in number represents a 100% increase in risk. For example, a score of three indicates that there is a three-fold increase in risk that a baby will develop sepsis. In addition to the current score, the monitor also displays a line graph that tracks the changes of the score for the previous 5 days. The score and five day trend are displayed on a monitor in each pod. The scores can also be accessed from both the bedside computer monitors and the computers in the nurses' stations. During randomized clinical trials of Horizon, the monitor was simply turned on to monitor some infants and turned off for others. The group of monitored infants experienced a reduction in mortality of over 20%. This was a startlingly successful outcome. In response, Horizon has been integrated in over 1500 NICU beds in seven different countries. In the following section, I outline the context in which I observed Horizon in practice, the NICU of EHU, a unit where Horizon has become a staple of premature infant care.

The NICU: Horizon's Environmental Context

Just before 7:00 each morning, nurses and doctors—most with coffee in hand—make a quiet commute through the lobby of AUH on their way to work in various units in the medical complex. Those who board the employee elevators and hit the button for the 9th floor are headed for their shift in the NICU. The NICU of AUH serves as the primary care unit for premature and sick infants for a large, mostly rural region. As such, the unit regularly has over 40 patients in residence. Entering and navigating the unit takes some acclimation. Changes in physical space, light, sound, and activity signal that you've entered a space unlike other areas of the hospital.

Rather than hallways with attached rooms, the unit is arranged into "pods," or segments of the unit that are somewhat separated from each other. The resulting layout resembles an asymmetrical section of honeycomb with openings to pass between each section. In the center of each pod there is a nurses' station, which contains several computers, chairs, and vitals monitors. Making a U-shape along three walls of each pod is conglomeration of technological devices, including computer stations, monitors for tracking vitals, pumps that administer medicine and food, and ventilator systems for breathing. Amidst all that equipment, there are seven to ten babies nestled against the walls. Some are in open bassinets and under phototherapy lamps that give off a purple light; others are in isolette incubators with quilted blankets drawn over them. The lights are kept off as much as possible, leaving each pod bathed in a combination of dim natural light and the violet glow of the bassinets. Within this space, there is constant movement: various specialists come in to check on babies who fall under their care; social workers pass through the pod checking on patients and families; giant carts capable of taking x-rays or ultrasounds of the babies are wheeled in and out; parents come to visit and hover over the incubators. From the walls of devices comes the constant and persistent call of alarms as they ding and blare for attention. The telephone in the nurses' station, cell phones from the nurses and physicians, and pagers chirp constantly. Beneath it all, there is the background hum of medical equipment and the cry of babies.

Horizon's Practical Context

In this environment, clinicians are tasked with detecting and responding to illness in the tiniest of patients. The smallest babies in the NICU may weigh less than 2 pounds, their arms are

no thicker than an adult's pinky finger, many cannot open their eyes, most cannot eat without the help of a feeding tube, some cannot tolerate touch, many struggle to breathe without the help of ventilator machines, and those that can cry make more of a whispered screech than a robust wail. Nurses are assigned between one and three babies at a time while physicians and nurse practitioners, depending of their position, oversee anywhere from 5 to 25 infants at a time. Most of the patients in the NICU are long-term patients who are there for support while they grow and develop. Therefore the nurses and doctors often focus on detecting conditions of sub-optimal health, such as infection, that may impede growth and endanger the infant. In some sense then, the job of the clinician is to constantly categorize patients as either sick or well.

The practices through which this categorization is assigned can most easily been seen during rounds. Each morning the rounding team moves to each baby to discuss her progress and to craft a plan of care for that day. The rounding team consists of seven to ten physicians, residents, pharmacists, nutritionists, and respiratory therapists who push around a number of "WOWs" or computer "workstations on wheels." When the team arrives in front of each isolette, the process begins with a member of the team giving an update on the baby's history and condition. This involves listing 20-30 pieces of discrete quantified information. The information includes things like delivery date, gestational age, date of birth, weight, how often they are feeding and how much they ate, how many times they have urinated and number of stools they have had in 24 hours, amounts of medication and how frequently they receive it, ventilator settings, white blood cell count, blood gas numbers, measures of various nutrients, and the number of apneic events. In addition to this, they review the baby's physical appearance and results of physical exam, baby's mood and reactions to various stimuli, whether they are on light therapy, if they are eating breast milk or formula, patient history, mother's history, any events or

procedures from the last 24 hours, and social conditions or concerns. More seasoned practitioners may be more selective about which information they present, seemingly presenting only that information that they see as relevant to the plan of care they will recommend to the attending. Nevertheless, the litany of information that junior practitioners share with the team conveys a sense of the vast inventory of information which passes through the clinicians' minds as they care for patients.

When I asked the clinicians to tell me about how they detect signs that babies in their care might be ill, it became clear that medical practitioners not only assess large constellations of information each day, but that they track that information over time and often look for deviations. In describing her process for knowing if a baby is getting sick, Melissa, a registered nurse (RN) said:

> Symptoms can be very subtle or they can be very obvious, and oh! there are so many. But not all the time do they have all the symptoms. But that decrease in tone, that lethargy, that the patient is kind of limp. They really look sick. They can have subtle color change, being more pale. We say they almost look green or grey. Again, they just look sick. They can have more apnea bradycardia events, more significant ones, more frequent ones. There can be, if it's a feeding thing, the feeding intolerance. A respiratory thing; there can be an increase in the amount of secretions, a change in color in the secretions that they have. I've seen tachycardia be a sign that's not related to like just being angry or hot, but just a baby that's lying very still and weak and tachycardia can be a sign. Let me think. There are so many subtle signs. So like laboratory signs. That's getting more into the doctor territory, so a white count that's high, platelet level might be low. Of

> course, a differential on a CBC, looking for a shift in the CBC. Yeah, the blood counts are really—that's like a later sign after we've looked at the blood counts to see those things.

Notice how often Melissa used words that indicate that she was looking for changes in the patient's condition. Her sentences are filled with words like "decrease," "increase," "change," and "shift" as she describes the signs that point to illness. Cindy, another RN, responded similarly by saying she worries about illness:

> When they don't look good. When they require extra IV or fluid or medication, when you see their complete blood count profile change. […]They have a change in their color, […] extra fluid where it shouldn't be—puffy. They maybe are not as responsive, they are more like a rag doll, you know, just listless, not as interactive. We follow lots of different blood studies. […] There are a lot of things you put together to get a clinical picture.

The practical context of the NICU is one in which medical practitioners track a constellation of information in order to construct knowledge about their patients' health. These signs are rarely considered in isolation, but instead in relation to one another. In addition, constructing knowledge about a patient does not occur solely through the reporting of absolute numbers. Changes and trends are especially important in the attention of the medical team and may alter their decisions about care.

Horizon's Cultural Context: The Apparent Dominance of Evidence-based Medicine

Evidence-Based Medicine

Within the NICU there is a cultural paradigm that favors particular kinds of measurable markers as evidence of disease. Like the broader culture of American hospitals, the clinicians of the NICU are steeped in the culture of *evidence-based medicine* (EBM). This approach gained traction in the 1980s and has since become the dominant way to approach practice in mainstream medicine. EBM focuses on informing practice with "a clearly defined hierarchy of available evidence," the best of which relies on randomized controlled double-blind clinical trials (Timmermans 2010:309). The nature of randomized controlled trials is one that often relies upon metrics, numeric operationalization of phenomena, and statistical methods to provide evidence. Therefore the values of EBM promote a viewpoint in which quantifiable and measurable information is often seen as more powerful while other forms of information, such as narrative and qualitative descriptions, are seen as somewhat suspect.

These judgments about the validity of various kinds of information are not only held by researchers. Instead, they are integrated into the education of nurses and physicians (Timmermans and Angell 2001) and get passed into the very practice of medicine. For example, during a meeting with a lead physician and his ICU team, a group of residents presented some new research on ICU protocols. The group evaluated these studied by discussing their statistical significance and aspects of research design. The attending physician concluded by stating, "Is there adequate evidence to change our practice? Is there adequate evidence to uphold our practice? Is there not enough data to say?—these are *always* the questions you should ask yourself." Through this statement, he reinforced these evaluative practices and instructed the

residents to bring those practices to medical care and decision-making. Similar statements were also common during conversations in which the medical team was deciding how to care for particular infants during rounds and intake meetings. During rounds, doctors often referred to "recent papers" on various health conditions and regularly used concepts like statistical significance, odds ratio, and p-values to evaluate these studies. Invoking these standards and criteria for assessing knowledge during the conversations through which decisions about patient care are made reinforces the values of evidenced-based medicine in the practice of care within the NICU. Clinicians also explicitly articulated this preference for particular kinds of evidence. As Jenn, a nurse practitioner explained, "It's fair to say that I think they would want to have some sort of lab work to back it up. We're very numbers oriented and we want growth, we want a bacteria, you know. Something to say, 'Oh, yes, this baby actually is sick.'" In short, as clinicians work to construct knowledge of patients' health, they are more comfortable with quantified or measurable evidence as a form of reliable information.

Intuition, Experience, and the Corporeal Signs of Babies

Despite the fact that quantifiable and measurable criteria are valued as the most legitimate form of information in the NICU, there are other forms that frequently inform the decisions and practices of nurses and physicians. Intuitive feelings, experiential knowledge, and the corporeal conditions of the baby figure heavily into the knowledge practices of the NICU, even if these aspects of knowledge construction are less recognized by the clinicians themselves.

For example, doctors may integrate feedback from parents or nurses when they believe that something is wrong. As Dr. Peterson indicated, "The parents that are very involved will give

their point of view, and we weigh it pretty heavily if it's a parent that is there a lot and has a good feeling for what their baby's like" (RHTRG 2012). Crucially, this "feeling" of what their baby is like is not based on the kind of evidence preferred by the paradigm of EBM such as lab results or deviations from the acceptable heart rate. Rather, it may be based on changes in mood, behavior, or something less tangible.

In addition, knowledge of a particular baby and her idiosyncrasies factors into decisions about care. Nurses and nurse practitioners frequently emphasized their ability to recognize that something was wrong with "their" baby based on personalized factors. When I asked Terry, an RN, what causes her to suspect that a baby might have sepsis she responded by saying,

> um… they just don't—they don't act right. A lot of them don't act like *themselves*. Like if they're a baby who never drops their heart rate, they might start dropping their heart rate all of a sudden. Or if they're a baby who never has snot, all of a sudden they have like a ton of snot. You're like "oh that's weird. They don't normally have boogers."

Similarly, Anna, an RN, told me a story of how she diagnosed a patient with sepsis: "he started having apneic events and bradycardia where he would stop breathing, drop his heart rate more frequently than usual, *which wasn't himself*." While parents and nurses may not always be in decision-making roles with regard to patients, doctors usually take this information seriously because parents and nurses spend more extended time with individual patients. Thus, clinicians often filter the signs and signals from their patients through a set of individually cultivated expectations before interpreting those signs as either indicators of illness or normal conditions.

In addition to filtering signs through expectations set by intuition or past experience with a patient, the corporeal conditions and symptoms of the infants occasionally outweigh the prescribed responses to evidence that we might expect from a context completely dominated by an evidence-based paradigm. This comes across most clearly in instances in which the evidence from various tests or lab results do not match up with other signs. For example, as Jenn indicated, "So...you may do all these lab tests and they all come back normal, yet the baby is saying, "I'm sick." So you're gonna treat 'em. So. It's kind of interesting. With that technology, you still can't rely a hundred percent on the lab work." In the case of sepsis, clinicians may decide to treat a patient even when the blood culture test is negative, meaning that there is no measurable indication of infection. In these cases of what doctors refer to as "culture negative sepsis," doctors believe sepsis to be present *despite* a lack of conclusive evidence.

This reliance on alternative forms of knowledge reveals a tension in medical practice. Although quantifiable and measurable markers of a patient's physiology are preferred by the formal doctrine for making decisions, the nature of the human body, disease, and current medical practices push back. Not all of the important signs about a patient can be transformed into this kind of information. Although doctors are entrenched in EBM and do explicitly articulate their adherence to this epistemology, the existence of phenomena such as culture negative sepsis demonstrate that they rely heavily on other forms of knowledge—those that cannot be accounted for quantitatively or documented through laboratory tests—in practice. It is this tension that interacts with Horizon in practice. In the NICU, Horizon does more than prognosticate early warnings of which infants will contract sepsis. Through the practices of *conditioned reading* and *accumulative reading*, it may also be discounted or serve to buttress existing suspicions of infection.

Horizon and Two Ways of Reading

Most of the clinicians I spoke with expressed considerable enthusiasm for the Horizon monitor. However, it is clear that Horizon does not function solely as intended by the developers, by predicting the onset of sepsis *before* any clinical signs are present. Essentially, an elevating Horizon score is designed to prompt the clinician to examine a seemingly-healthy infant for signs of illness. Closer examination of the patient might lead to further tests. About two-thirds of my interviewees explicitly asserted that Horizon works in this way, suggesting that Horizon does sometimes work as envisioned. However, when I asked clinicians to tell me stories about using Horizon, almost all of these stories revealed alternative processes by which Horizon influences the categorization of infants as either sick or well. I describe these processes as *conditioned reading* and *accumulative reading*. These readings of Horizon do not occur in a predetermined order, and they may overlap. Nevertheless, these processes are the frames through which Horizon's output is integrated with other signs and used to construct knowledge of a patient's health.

Conditioned Reading

Almost all of the clinicians described the process of conditioned reading in their discussion of Horizon. The concept of conditioned reading points to processes through which users of data-driven technology temper, filter, discount, or place trust in its output. During my observations of the NICU, clinicians consistently told me about how they make judgements about when to react to a change in Horizon and when to ignore it. As Robin, a neonatal nurse practitioner, told me, she always asks herself, "okay, do I believe Horizon in this kid or not?"

Much like other indicators of illness, that judgment about the case-by-case validity of Horizon is based primarily on the clinician's existing experience with a particular infant or experiential information passed on from other clinicians. As Terry described:

> I think you just learn whether they have- like what their trend is. You learn whether they're a baby who's always steady on like, low, or you learn whether they're a baby that spikes every night and then comes down during the day. […] I feel like that's how it helps […] Like if I know a baby really well, and I get report from a nurse who has never had the baby, and they're like, 'Oh my gosh, their Horizon is 3.' And I'm like, 'Oh no, it's goes to 3 every night, don't worry, it'll come back down.' Like kind of like that. We kind of learn their trend, I think.

As is the practice with other signs of patient health, clinicians track Horizon's scores over time. This can lead them to discount its trustworthiness in particular patients. It can also lead Horizon to be discounted for entire patient populations. For example, when discussing an infant with several congenital conditions, the nurse told the rounding team that the baby was doing poorly. Her Horizon score had been sitting at 7, the highest risk score possible, since I arrived at the NICU that morning. In reference to Horizon, the nurse commented that, "I know it's been high, but we took no action." The nurse then implied that this was due to the assumption that Horizon is not trustworthy for infants with this condition. The rounding team concurred, concluding that the cause of this baby's troubles was not infection. In this case, past experience had convinced the team Horizon is untrustworthy for infants with congenital problems. Despite the high score, they did not test for infection or begin antibiotics.

The formation of these conditioned readings is a continuous, iterative process. On occasion, clinicians have an experience that causes them to recast their readings of Horizon. As Dr. Walker described to me:

> I know there are cases where we haven't started, you know, we haven't necessarily changed our management, and the patient has decompensated and, looking back the Horizon score was elevated, but maybe we were attributing it to something else, or maybe it's just been elevated for days, and it just wasn't a big change in the trend.

In this instance, the Horizon score was not initially conditioned to be interpreted as a trusted sign of sepsis. It was only through the appearance of infection via other signs that, in retrospect, Horizon and the other symptoms were made interpretable as indicators of infection.

The concept of conditioned reading shows one way in which the process by which clinicians interpret Horizon is made consistent with other knowledge practices in the NICU. Rather than focus on absolute numbers alone, clinicians track the score over time, relating it to the corporeal signs of the infant. This tracking of both the scores and the resulting change or lack of change in patient health allows them to establish a baby's version of "normal" and temper Horizon's predictions. This can lead them to either take the score seriously or to discount it. When discounted, Horizon drops out of the constellation of information through which clinicians construct knowledge about patients. When trusted, Horizon does not simply dictate care, but contributes to decisions about care and treatment through a process of accumulative reading.

Accumulative Reading:

When Horizon is believed to be a legitimate indicator of illness, it is included in the constellation of information used to construct knowledge about patients. Accumulative reading describes the process by which data-driven technology is layered upon other information and sometimes used to assess the meaning of other information. Respondents frequently described the process of diagnosing patients as putting together "pieces of a puzzle." This metaphor reveals much about the way that clinicians see illness. A puzzle can only be put together one way; it reveals a single picture. Similarly, clinicians approach troubling symptoms with the assumption that a true and single explanation (though it may involve several causes) can be revealed if all the available information is interpreted correctly. As Tina, a neonatal nurse practitioner, indicated, clinicians treat Horizon as "a piece of the puzzle in trying to diagnose sepsis."

Within that puzzle or constellation, Horizon has a distinct relationship to other signs. In contrast to its intended use as an early warning, in practice Horizon often helps to reinforce or dissuade *existing suspicions* of infection. In telling me stories about how they use Horizon to diagnose patients, clinicians rarely articulated specific experiences where an increase in the Horizon score was the first and primary indicator of sepsis. They were more likely to tell me stories about instances in which they noticed or sought out Horizon after the onset of other symptoms or simultaneously with other symptoms. Some clinicians were quite explicit about this process and their use of Horizon as a check on other symptoms and signs. Consider this story that Anthony, an RN, told me about a baby that developed sepsis:

> So there's one baby…I think he was on CPAP [a device that helps the baby breathe]. […] He was doing fine, but then near the end of our shift—his temperature was fine—but then near the end of our shift his residuals…say he was getting 10, his residual was 8! And it looked more bloody-ish rather than just undigested food, and I was like, "Amy [another nurse], I don't know about this. This is not looking right." And she was like, "Yeah you're right, you should page the docs." […] And I was like, "Also, his stomach's a little bit rounder. It feels firm. I'm kind of concerned. This is not how he was three hours ago." And she was like, "you're right. He wasn't like this three hours ago." They didn't come around until after we'd left, but when we came back the next morning, the nurse said they did a full sepsis workup on him last night because he had a Horizon spike right after you guys left and his temperature started going up. It's those small things where you're like, oh maybe it's just a feeding thing, but over the course of two or three hours, these other things started happening.

While Anthony knew that his patient was not well, he did not yet have a particular interpretation of what was causing the symptoms. Reading the constellation of signs, which included the rise in the Horizon score, the rest of the medical team suggested that sepsis was the cause and initiated a medical intervention.

However, the inverse can occur with accumulative readings; Horizon can also dissuade suspicions of infection, encouraging clinicians to dismiss signs of illness. For example, Dr. Manning said, "I guess the Horizon score I use sometimes in a confirmation that things are okay. Knowing—the nurse is telling me that the baby is a little more lethargic […], but the Horizon

score is down is a little more reassuring that, okay, well we can see how the baby does instead of jumping right then and doing something about it" (RHTRG 2012).

Finally, tracking Horizon over time may matter more than the absolute number for accumulative readings as well. On one of my first visits to the NICU, the rounding team was discussing an infant who had a fever. The medical team was uncertain what was causing it. Dr. Kapoor, the resident assigned to this patient, told the team that "everything is up, except the Horizon score," indicating that there were worrying changes in the infant's vitals and lab reports. Dr. Walker, who was overseeing Dr. Kapoor, looked at the monitor and said, "but the Horizon *is going up*. We should probably do a full work up," and asked the team to start antibiotics. Although the absolute number was under 2, a score that the team would usually accept as normal, it had been trending upward over the last few hours. The combination of the low, but rising, score in tandem with other symptoms led the medical team to treat the baby and look for infection. The rising score solidified the interpretation of other signs and reinforced a decision to treat for sepsis.

Regardless of whether it persuades or dissuades, in accumulative readings, Horizon can act as a flip switch on the meaning-making process. The Horizon score alters the way in which other signs within the constellation are read; a sign like lethargy becomes either a foreshadow of illness or a quirk of the baby's mood. Crucially, this depends on Horizon also having a conditioned reading in which it is a trusted marker of the baby's condition. In cases where Horizon is taken as a legitimate sign, it can be a powerful force in shaping the meaning of other signs and the categorization of infants as either sick or well.

Discussion:

Meaning is interactional. As meaning-making beings, people must be able to account for their actions and beliefs in ways that are recognizable to others (Berger & Luckmann 1967). Acceptable explanations and accounts vary according to social context; different institutions have different ways of reasoning that count as legitimate (Boltanski & Thévenot 2006; Mills 1940). Clinicians must make judgments about care within the appropriate framework attached to the institution of medicine. The acceptable framework within the neonatal intensive care unit (NICU) is undoubtedly that of evidence-based medicine (EBM). The tension between the tenets of EBM and the implicit value of alternative forms of knowledge and information are central to the meaning-making process and provide insight into the patterns by which Horizon intervenes in constructing knowledge.

Although Horizon is a quantified metric, a highly valued kind of information according to the tenets of EBM, its use and interpretation are intimately linked to the presence of alternative forms of knowledge in the NICU. Horizon would work quite differently in an environment that operated solely by the criteria espoused by EBM. Due to its success in the randomized control trial, each rise in the score would need to be taken seriously and discounting Horizon might rarely occur. Instead, experience and qualitative knowledge factor heavily into care decisions. Over time, clinicians spend countless hours caring for infants, building up experiential knowledge of infants in general and for the particular patients under their care. This experience attunes them to small changes and provides them with insights about a baby's condition. This sometimes creates tension for clinicians, who themselves are more comfortable with metrics and lab results as true indications of illness. Further, these hunches are often insufficient for making a diagnosis or treatment decisions. Diagnoses require evidence, usually in the form of

quantifiable vitals and lab results. Qualitative and narrative accounts cannot fully account for a legitimate interpretation of a baby's condition. Horizon is powerful because it provides a quantified metric through which to filter other signs the clinicians receive.

Crucially, Horizon's use in this way pushes beyond rhetorical strategies used to convince others. As Mills (1940) argues, accounts must be sensible to our selves as well. It is not simply that clinicians use Horizon to make their concern legitimate to others. It is that they may convince *themselves* that subtle signs like a round belly are either significant signs of illness or meaningless, depending upon their assessment of Horizon's signal. It is this location within an institution that formally sees the evidence-based paradigm as the only legitimate means of constructing knowledge that calls for and allows Horizon to function as a check on other suspicions of illness.

The resulting practices of conditioned and accumulative readings stand in stark contrast to Horizon's intended use as an early warning system. To function in this way, Horizon would need to stand in isolation, outside of the constellation of information involved in accumulative readings, and it would need to avoid being discounted as sometimes occurs in conditioned readings. This is not to say that Horizon is not useful. In an epistemological context disciplined by EBM, Horizon's role in the interpretive process may actually help to make other forms of knowledge accountable and encourage that they remain part of the constellation of information by which clinicians construct knowledge claims about infants.

However, as the medical community embraces approaches like precision medicine—the tailoring of treatment plans to particular patient populations based upon the collection and analysis of large data sets—and pushes for the increased use of algorithmic and computer-

supported decisions, the possibility of conditioned and accumulative readings may become less likely. Precision medicine associates better understandings of medical conditions with the increased ability to collect and analyze data harvested through electronic medical records and sensory devices (Leff and Yang 2015). As hospitals are pressured to become more efficient, to reduce their use of resources, and to legitimate their practices in accordance with precision medicine, we can imagine a world in which treatment decisions are made by doctors sitting at far-off computer screens, depriving clinicians of the opportunity to develop conditioned readings of algorithmic output. To some degree, there is already a tendency in this direction among younger physicians. Dr. March, the surgeon who participated in the medical analytics team, frequently mentioned that he was at pains to insist that the residents he supervised in the intensive care unit go and actually look at the patients for whom they were caring, rather than simply make treatment decisions from their computer stations. If such a trend away from patient experiences continues, reactions to predictive analytics may become more standardized, potentially removing idiosyncratic knowledge of particular patients from the constellation of information by which clinicians determine patient conditions. Given that the practices of NICU clinicians reveal that data analytics and metric-based assessment fail to fully capture the phenomenon of illness[25], medical professionals and scholars should think carefully about the conditions under which data analytics are employed and work to design organizational contexts that leverage the advantages of data analytics while preserving other forms of knowledge as well.

Outside of medicine, conditioned and accumulative readings may already be beyond the scope of possibility. With Horizon, organizational protocols permit the medical team to decide

[25] This observation parallels critiques of another type of reading associated with data, that of "distant reading" (Moretti 2005). Originating in literary studies, this methodological approach suggests that large-scale data analysis is the best means to understanding the nature of literature. Critics argue that this method neglects the production of meaning and thereby fails to account fully for a phenomenon (see Kitchin 2014).

how to respond to the algorithm, thereby encouraging interpretation in the first place. In addition, clinicians are confronted with the bodily reality of the object they are trying to know. The baby gets sick or well in spite of what the algorithm predicts, and clinicians experience the mismatch of data and the corporeal reality of the baby. This experience is a significant aspect of why conditioned readings are able to unfold. The same may not be true in other contexts. For example, with predictive policing, unreported crimes are unable to alter the user's experience of the algorithm. If the algorithm tells them not to patrol an area, they cannot know if an unreported crime happened there anyway. The crime analyst may remain ignorant of the mismatch between her numbers and the reality of criminal events, and this in turn may lead to greater trust or more deterministic reactions associated with algorithmic output.

Conclusion:

Through this case, I have argued that unpacking algorithms alone falls short of fully appreciating the way in which data analytics impart consequences on the world. The subtle processes by which predictive analytics are discounted, trusted, and interact with other signs are missed when critiques of data analytics are abstracted from practice or when staring solely into the black boxes of data warehouses and algorithms. Though the professionals in this study espouse a deep belief in the ability of metrics and analytics to reflect the truth, their sense-making practices draw heavily upon other forms of knowledge and complicate the processes by which analytics are transformed into action. The concepts of accumulative and conditioned readings point to knowledge and expertise that cannot be rendered as numbers or analytics and draw attention to the meaning-making process by which algorithms are interpreted and used.

These processes are central to a holistic understanding of the both the scope (Pink et al. 2016) and societal consequences (Schäfer & van Es 2017) of datafication. Examining other settings to identify the factors that facilitate conditioned and accumulative readings, their role in producing knowledge claims, and their connection to experiential knowledge and expertise will allow scholars to develop a general framework for the means by which users turn analytics into knowledge and action and allow practitioners to cultivate settings that encourage such processes when desired.

CHAPTER 6

Conclusion: The Work of an Epistemological Landscape

In the introduction, I argued that the institutionalization of data and the incredible amount of material resources and cultural authority attached to it called for investigation into the symbolic order or worldview that accompanies data science. Where work within critical data studies had begun to articulate the contours of meaning in this new paradigm, few studies have attended to data science in practice but have instead leveraged somewhat distanced critiques through analyses focused on discourse or technical and processual aspects of big data and data science, rather than engaging in ethnographic inquiry. In contrast, the sociology of knowledge has focused intensely on practices and ethnographic methods in recent years. I take up this approach to studying data-driven knowledge settings, but argue that the sociology of knowledge has moved away from an inclusion of subjective experiences and beliefs, even in the explicitly cultural work of Knorr Cetina (1999). I argue that this is a central part of understanding the ways in which data science will influence society and advocate for an investigation of the epistemological landscape, the meanings and worldviews through which actors see data science as the appropriate tool and method for producing legitimate knowledge claims and solving problems.

How, then, does attending to the epistemological landscape help us to make sense of occurrences such as the opening story of this project? Why would the data scientists and medical professionals of the medical analytics team be so quick to discount the assessments of their colleagues or of a trained observer? As I depicted in chapters 2 and 3, the epistemological landscape of data science is one in which, much like broader cultural shifts, the subjectivity of human experience is seen as a threat to true insights. Trained experts and scientists themselves

are not exempt from the potential blindfold of subjectivity, and in fact, their training may make them more susceptible to it. The techniques of data science such as automated data collection and machine learning seemingly provide routes to the truth that steer data scientists safely around these dangers. With the new tools that data science has provided, science can fulfill its promise to better the world. It is within this epistemological landscape that the medical analytics team saw quantified data collection and analysis techniques as central to their effort to explain how their interventions changed medical practice. Their beliefs about human nature, in addition to other features of the landscape, encouraged the preference for particular means in discerning how and why their algorithms changed medical practice.

Capturing Differing and Complex Social Consequences

However, when we begin to look at data science in a variety of contexts, it becomes clear that the cultures, organizational structures, and goals or demands of various settings filter the ways in which the epistemological landscape of data science shapes the production of knowledge. The effects are not always the same; data science and its products may push practices and culture in varying ways. On the one hand, chapter 3 demonstrated that when data scientists are pressured by deadlines, contracts, or clients to produce useable results, they move away from concerns about epistemological authority and focus instead on methods that allow them to improve outcomes. This shift may be even more troubling than the critiques leveraged by critical data studies. In this version of data science, not only does technical expertise outweigh experiential knowledge, but flawed techniques may be found acceptable. Due to the potential slippage between statements used to solve problems (i.e., including zip code in a model

improves our ability to predict defaulting on loans by 20%) and knowledge claims (i.e., people in this zip code do not pay their loans), this is an especially troubling manifestation of data science.

On the other hand, in some contexts the tools of data science may be integrated into existing practices and cultures. In the case of the NICU and Horizon, clinicians still attribute some epistemological authority to experiential knowledge. Experiential knowledge of infants, disease, or even of Horizon itself lead clinicians to condition the meaning of Horizon's predictive risk scores. In some instances, Horizon can actually buttress experiential knowledge, ensuring that it remains part of the practices by which clinicians construct knowledge about their patients. As I argue in the chapter, this results from the fact that, although clinicians reference and rely upon experience a great deal as they care for patients, they are more likely to explicitly recognize the epistemological authority of the quantified and measurable phenomena recognized as evidence by the institution of evidence-based medicine. As I note, however, this use of Horizon is contingent upon organizational structures that bring clinicians in physical contact with patients, allowing them to build experiential knowledge of both patients and Horizon's match or mismatch with patient outcomes and upon protocols that give clinicians the discretion to choose how to respond to Horizon's risk scores. These aspects of the Horizon case show that data science tools may be integrated into knowledge settings in productive ways that take advantage of the capabilities of data without pushing out other forms of knowledge. However, pressures toward increased quantification, datafication, and efficiency should caution us to the intentional effort required to preserve these aspects of knowledge settings.

The issues discussed in these chapters show that we cannot fully understand the way in which data science, big data, and algorithms will impart social and cultural consequences without attending to their manifestations in particular contexts. The ways in which data shapes

decisions and knowledge varies greatly between the work places of data scientists and that of the NICU. It is likely that other settings in which data science is increasingly used to make decisions, such as courtrooms, police departments, policy and government organizations, or media organizations will each integrate data and data-driven tools in disparate ways. In addition, understanding why data is used in particular ways requires attending to the symbolic orders by which people orient themselves and their work to data. As a result, any effort to construct organizational environments in which the benefits of data science may be leveraged without also introducing problems associated with data science needs more than the distanced criticism offered by most of the critical data studies scholarship. Instead, effective policies will depend upon the continuation of this kind of research.

Considering Data in the Knowledge Society

In addition to pointing to the diverse ways in which data science may lead to complex material or social outcomes, an investigation of the epistemological landscape of data science and its integration into varied contexts helps to shed light on the potential cultural consequences of data science as well. As I indicated in the introduction, a multitude of social theorists have claimed that knowledge is the defining characteristic of our time. The term *knowledge society* (Böhme and Stehr 1986, Knorr Cetina 2007) and similar monikers (Giddens 1990, Castells 2000) are used to signal the ways in which information and knowledge have become the productive forces that drive economic (Bell 1973), political (Böhme and Stehr 1986) and cultural activity (Knorr Cetina 2007). Most assessments of the knowledge society have focused on structural, material, or economic changes. As expert systems expand, they disembed social relations from

local interactions (Giddens 1990). Knowledge is now recognized as a sector of the economy and the use of knowledge production practices to do work other than generating scientific knowledge is expanding (Gross 2012). Under these empirical conditions, the study of non-laboratory, non-academic, and non-research settings becomes a key piece of assessing the knowledge society. The consulting firms in which the data scientists work as well as the medical analytics team and NICU have provided a few examples of these kinds of knowledge settings.

When it comes to cultural transformations, Knorr Cetina provides a framework for assessing the symbolic aspects of the knowledge society. First, she suggests that epistemic cultures constitute the knowledge society (Knorr Cetina 2007). When epistemic cultures become dispersed, Knorr Cetina argues that sociologists should study the "macro-epistemics" of various networks. She provides the example of the Global Financial Architecture and the ways in which this macro-epistemic focuses on "news" rather than "truth" when trying to ascertain financial activity. Studying these phenomena, Knorr Cetina suggests, are ways to access Knowledge Culture. By this, she signals the treatment of "general culture as a kind of knowledge culture" (ibid: 369-370). Knowledge does not simply inform other realms such as economic and political life. Instead, knowledge culture cuts through and constitutes activities in these areas. She suggests that the societies in which epistemic settings are contained may come to reflect aspects of local epistemic cultures. Through this connection, epistemic cultures may influence important features of the "lifeworld" or general knowledge culture (371). This means that the symbolic orders of knowledge settings may influence ways in which society constitutes objects such as selves or identity.

If we accept Knorr Cetina's argument, then studying the symbolic order and epistemological landscapes of data science is essential because it taps into broader cultural

constructs. There are at least two ways in which the epistemological landscape of particular knowledge settings are related to the knowledge society. First, as I have argued throughout this project, cultural products and the possible symbolic orders they provide constitute one of the conduits by which data science interacts with broader social and cultural patterns. While studies of the construction of objects and knowledge products (Latour 1988) or the technical and systemic aspects of data science (Bucher 2012, Beer 2015) do address certain aspects of the productive force of data science, they do not offer a complete picture. The symbolic orders that inhere in the language and presentation of data science products will shape how society responds to, assesses, and employs data science as a means of knowledge production and problem solving. For example, as I show in chapter 4, data talk suggests a certain location of epistemological authority—in the details of data points—and normalizes certain approaches to data ownership. In addition, media coverage of data science contains the same notions of human nature and the danger of subjectivity contained in the epistemological landscape of data science.

Second, rather than thinking of knowledge settings or macro epistemics as shaping the knowledge culture through their practices, machineries, or production of knowledge claims, we can treat these settings as manifestations of the broader cultural outlook (manifestations that may in turn filter out to the knowledge society through the symbolic products discussed above). This approach requires a consideration of the way in which the broader culture meshes or clashes with the experience of local contexts. As such it includes an assessment of the subjective experiences of actors in these settings and under particular conditions. In integrating the subjective, this kind of analysis resembles the sociology of knowledge of Karl Mannheim (([1936] 1985).

Though I am not suggesting that we pick up Mannheim's theories in their entirety, his concept of *Weltanshauung* closely resembles what Knorr Cetina describes as a knowledge

culture. Mannheim sees the *Weltanshauung* as a global outlook of an era ([1936] 1985). Importantly, he notes that this worldview may manifest itself in different ways by different groups depending on their experiences. In his empirical work, Mannheim outlines a variety of political views and what happens when they are put in contact with particular situations. He argues that a key feature of human existence is that our "characteristics emerge in the course of [our] concrete conduct and in confrontation with actual problems" (ibid: 169). Further, the various manifestations of a *Weltanshauung* may contain utopian leanings, future orientations that drive people to encourage changes to current social realities. To reconstruct these worldview and the visions that they contain requires and attending to both the symbolic order and the locations in which individuals are situated.

Mannheim gives two missions to the sociologists: One is to reconstruct the epistemology of worldviews. The other is to trace out social determinants of knowledge. Through an exploration and reconstruction of the epistemological landscape of data science (found in chapters 2 and 3) and a focus on epistemological authority, I have provided an analysis of one strand of thought that circulates in the Knowledge Society. In looking at the manifestations of this worldview in different settings, I have begun to identify some of the concrete situations that lead to variations in this worldview. We can see that when confronted with the task of solving problems and implementing solutions, at least some data scientists shift to a pragmatic mode in which legitimacy is derived from usefulness rather than correlations to the truth. The metaphors contained in data talk suggest that without the tempering experiences of conducting data science themselves, non-specialists and the public may receive a slightly different version of the epistemological landscape, one in which authority lies in the details of the data rather than the techniques used to analyze it. Finally, though the clinicians of the NICU are oriented toward

evidence-based medicine, a perspective which resonates with the epistemological landscape of data science, their context presents a partial mismatch with this perspective and tempers the authority of data-driven claims.

The epistemological landscape of data scientists, of data talk, and of the medical analytics team all contain utopian elements whereby advances in our ability to harness and analyze data in ways free of human subjectivity will finally enable us to solve some of the world's most difficult problems—whether that be the problem of disease, global conflict, or mitigating natural disasters. As these visions motivate individuals and organizations to increasingly invest in data collection and storage, to apply data science techniques in new settings, and to advocate for the advancement of data science as the determinant of truths and solutions, we must ask if the epistemological landscape of data science or one of the variants explored in the chapters here will become the primary worldview of the knowledge society. Will data itself or the techniques associated with data science become the ultimate source of authority in intellectual or political debate? Will epistemological authority fail to matter at all as the perceived success of data-driven interventions come to reinforce the assumptions contained in algorithmic models as truth? To be sure, the ways in which data science produces claims and the material and social consequences of specific data science endeavors will continue to be of important sociological investigation as well. But, if data science, big data, and algorithmic knowledge are able to continue to dominate public imagination and become the primary arbiters of truth, it will be through an expansion and persistence of an epistemological landscape resembling that which has been outlined in these pages.

WORKS CITED

Abend, Gabriel. 2014. *The Moral Background: An Inquiry into the History of Business Ethics*. Princeton University Press.

Abbott, Andrew. 1988. *The System of Professions*. Chicago: University Of Chicago Press.

Alexander, Jeffrey C. and Philip Smith. 1993. "The Discourse of American Civil Society: A New Proposal for Cultural Studies." *Theory and Society*. 22:151-207.

Ananny, Mike and Kate Crawford. 2014. *A Liminal Press: Situating News App Designers within a Field of Networked News Production*. Rochester, NY: Social Science Research Network. Retrieved January 23, 2017 (https://papers.ssrn.com/abstract=2448736).

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *WIRED*. Retrieved October 20, 2014 (https://www.wired.com/2008/06/pb-theory/).

Anderson, C. W. 2012. "Towards a Sociology of Computational and Algorithmic Journalism." *New Media & Society* 15(7):1005–21.

Andrejevic, Mark. 2014. "The Big Data Divide." *International Journal of Communication* 8:1673–89.

Angwin, Julia, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." *ProPublica*. Retrieved March 21, 2017 (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Barnes, Barry. 1974. *Scientific Knowledge and Sociological Theory*. London and Boston: Routledge and Kegan Paul Ltd.

Barocas, Solon and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." Pp. 44–75 in *Privacy, Big Data, and the Public Good Frameworks for Engagement*, edited by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge University Press.

Beer, David. 2015. "Productive Measures: Culture and Measurement in the Context of Everyday Neoliberalism." *Big Data & Society* 2(1):2053951715578951.

Beer, David. 2016. "The Social Power of Algorithms." *Information, Communication & Society* 0(0):1–13.

Bell, Daniel. 1973. *The Coming Of Post-Industrial Society*. Basic Books.

Bell, Genevieve. 2014. "Author Meets Critics: Pressed for Time: The Acceleration of Life in

Digital Capitalism." Retrieved March 20, 2017 (https://www.youtube.com/watch?v=uR1xDyh2oXY).

Bell, Genevieve. 2015. "The Secret Life of Big Data." in *Data: Now Bigger and Better!*, edited by T. Boellstorff and B. Maurer. Chicago, IL: Prickly Paradigm Press.

Berger, Peter L. and Thomas Luckmann. 1967. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor.

Berry, David M. 2011. "The Computational Turn: Thinking About the Digital Humanities." *Culture Machine* 12(0).

Bijker, Wiebe E. 1995. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. Cambridge, Mass.: The MIT Press.

Blei, David M. 2012. "Prababilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Bloor, David. 1976. *Knowledge and Social Imagery*. London: Routledge and Kegan Paul Ltd.

Böhme, Gernot and Nico Stehr, eds. 1986. *The Knowledge Society: The Growing Impact of Scientific Knowledge on Social Relations*. Dordrecht, Boston, Lancaster, Tokyo: D. Reidel Publishing Company.

Boltanski, Luc and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Princeton: Princeton University Press.

Bowker, Geoffrey C. 2014. "Big Data, Big Questions| The Theory/Data Thing." *International Journal of Communication* 8(0):1795–99.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15(5):662–79.

boyd, danah. 2014. "What Does the Facebook Experiment Teach Us?" *Social Media Collective*. Retrieved December 30, 2014 (https://socialmediacollective.org/2014/07/01/facebook-experiment/).

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16(3):199–231.

Bucher, Taina. 2012. "Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook." *New Media & Society* 14(7):1164–80.

Camic, Charles, Neil Gross, and Michèle Lamont, eds. 2011. *Social Knowledge in the Making*. 1 edition. Chicago : London: University Of Chicago Press.

Chamorro-Premuzic, Tomas. 2014. "Should big data analytics decide whether you get promoted?" *The Guardian* Feb 7. Retrieved 2/25/14. (http://www.theguardian.com/media-network/media-network-blog/2014/feb/07/people-analytics-office-politics-hr/print).

Charmaz, K. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* London: Sage Publications.

Carruthers, Bruce G. and Wendy Nelson Espeland. 1991. "Accounting for Rationality: Double-Entry Bookkeeping and the Rhetoric of Economic Rationality." *American Journal of Sociology* 97(1):31–69.

Carruthers, Mary. 2008. *The Book of Memory: A Study of Memory in Medieval Culture*. 2nd ed. Cambridge, UK ; New York: Cambridge University Press.

Castells, Manuel. 2000. *The Rise of The Network Society: The Information Age: Economy, Society and Culture*. 2nd ed. Oxford and Malden, MA: Blackwell.

Cheney-Lippold, John. 2011. "A New Algorithmic Identity Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28(6):164–81.

Christin, Angèle. 2014. "When It Comes to Chasing Clicks, Journalists Say One Thing but Feel Pressure to Do Another." *Nieman Lab*. Retrieved January 17, 2017 (http://www.niemanlab.org/2014/08/when-it-comes-to-chasing-clicks-journalists-say-one-thing-but-feel-pressure-to-do-another/).

Crawford, Kate and Jason Schultz. 2014. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55:93.

Crawford, Kate, Kate Miltner, and Mary L. Gray. 2014. "Special Section Introduction *:* Critiquing Big Data: Politics, Ethics, Epistemology" *International Journal of Communication* 8, 1663–1672.

Cohn, Carol. 1987. "Sex and Death in the Rational World of Defense Intellectuals." *Signs* 12(4):687–718.

Collins, H. M. and Steven Yearly. 1992. "Epistemological Chicken." in *Science as Practice and Culture*, edited by A. Pickering. Chicago and London: The University of Chicago Press.

Collins, Harry and Robert Evans. 2007. *Rethinking Expertise*. Chicago: University Of Chicago Press.

Couldry, Nick and Joseph Turow. 2014. "Advertising, Big Data and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy." *International Journal of Communication* 8:1710–26.

Dalton, Craig M., Linnet Taylor, and Jim Thatcher (alphabetical). 2016. "Critical Data Studies: A Dialog on Data and Space." *Big Data & Society* 3(1):2053951716648346.

Daston, Lorraine. 1992. “Objectivity and the Escape from Perspective.” *Social Studies of Science* 22(4):597–618.

Daston, Lorraine and Peter Galison. 2010. *Objectivity*. Zone Books.

Dean, John and William Whyte. 1958. “How Do You Know If the Informant Is Telling the Truth?” *Human Organization* 17(2):34–38.

Desrosières, Alain. 2001. “How Real Are Statistics? Four Possible Attitudes.” *Social Research* 68(2):339–55.

Douglas, Mary. 1986. *How Institutions Think*. Syracuse: Syracuse University Press.

Durkheim, Emile and Marcell Mauss. [1903] 1963. *Primitive Classification.* Translated and edited by Rodney Needham. University of Chicago Press.

Epstein, Steven. c1996. *Impure Science : AIDS, Activism, and the Politics of Knowledge*. Berkeley : University of California Press. Retrieved (http://search.lib.virginia.edu/catalog/u2633411).

Espeland, Wendy Nelson and Michael Sauder. 2007. “Rankings and Reactivity: How Public Measures Recreate Social Worlds.” *American Journal of Sociology* 113(1):1–40.

Espeland, Wendy Nelson and Mitchell L. Stevens. 2008. “A Sociology of Quantification.” *European Journal of Sociology* 49(3):401–36.

Espeland, Wendy Nelson and Berit Irene Vannebo. 2007. “Accountability, Quantification, and Law.” *Annual Review of Law and Social Science* 3(1):21–43.

Evans, Robert and Harry Collins. 2008. “Expertise: From Attribute to Attributions and Back Again?” Pp. 609–30 in *The Handbook of Science and Technology Studies*, edited by E. J. Hackett, O. Amsterdamska, M. E. Lynch, and J. Wajcman. Cambridge and London: The MIT Press.

Fiore-Gartland, Brittany and Gina Neff. 2015. “Communication, Mediation, and the Expectations of Data: Data Valences Across Health and Wellness Communities.” *International Journal of Communication* 9(0):1466–84.

Foucault, Michel. 1980. *Power/Knowledge: Selected Interviews and Other Writings 1972-1977.* New York: Pantheon Books.

Fourcade, Marion and Kieran Healy. 2013. “Classification Situations: Life-Chances in the Neoliberal Era.” *Accounting, Organizations and Society* 38(8):559–72.

Friedland, R. and R.R. Alford. 1991. "Bringing Society Back In: Symbols, Practices, and

Institutional Contradictions." Pp. 232-266 in *The New Institutionalism in Organizational Analysis*, W. W. Powell and P. DiMaggio (eds). Chicago: University of Chicago Press.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Cambridge, UK: Polity.

Gartner, Inc. 2014. "Gartner Says Beware of the Data Lake Fallacy." Retrieved October 31, 2016 (http://www.gartner.com/newsroom/id/2809117).

Geertz, Clifford. 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Gieryn, Thomas F. and Anne E. Figert. 1986. "Scientists Protect Their Cognitive Authority: The Status Degradation Ceremony of Sir Cyril Burt." Pp. 67–86 in *The Knowledge Society, Sociology of the Sciences*, edited by G. Böhme and N. Stehr. Dordrecht, Boston, Lancaster, Tokyo: D. Reidel Publishing Company.

Gieryn, Thomas F. 1994. "Boundaries of Science", in Sheila Jasanoff, Gerald E. Markle, James C. Petersen and Trevor Pinch (eds), Handbook of Science and Technology Studies (Newbury Park, CA: Sage/4S): 393-443.

Gillespie, T. 2014. "The relevance of algorithms." Pp. 167-194 in Gillespie, T., Boczkowski, P. J., & Foot, K. (Eds.), *Media Technologies : Essays on Communication, Materiality, and Society*. Cambridge, MA: The MIT Press.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, Massachusetts ; London, England: The MIT Press.

Gitelman, Lisa and Virginia Jackson. 2013. "Introduction." in *"Raw Data" Is an Oxymoron*, edited by Lisa Gitelman. Cambridge, Massachusetts ; London, England: The MIT Press.

Golden, Brian R. 1992. "Research Notes. The Past Is the Past—Or Is It? The Use of Retrospective Accounts as Indicators of Past Strategy." *Academy of Management Journal* 35(4):848–60.

Gregg, Melissa. 2015. "Inside the Data Spectacle." *Television & New Media* 16(1):37–51.

Gupta, Shalena. 2014. "Bright lights, big cities, bigger data" *Fortune.* Oct 30. Retrieved 12/3/14. (http://fortune.com/2014/10/30/bright-lights-big-cities-bigger-data/).

Hacking, Ian. 1986. "Making Up People." in *Reconstructing Individualism.* T. Heller et al (eds). Standford University Press.

Hacking, Ian. 1992. "The Self Vindication of Laboratory Sciences" in *Science as Practice and Culture.* Andrew Pickering (ed). Chicago and London: University of Chicago Press.

Herschel, Gareth, Alexander Linden, and Lisa Kart. 2014. "Magic Quadrant for Advanced Analytics Platforms." *Gartner.com*. Retrieved December 5, 2016 (https://www.gartner.com/doc/2667527/magic-quadrant-advanced-analytics-platforms).

Hilbert, Martin and Priscila López. 2011. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332(6025):60–65.

Hochschild, Arlie Russell. 2016. *Strangers in Their Own Land: Anger and Mourning on the American Right*. New York: The New Press.

Ian, Kerr and Earle Jessica. 2013. "Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy." *Stanford Law Review Online* (65). Retrieved December 5, 2014 (https://www.stanfordlawreview.org/online/privacy-and-big-data-prediction-preemption-presumption/).

IBM. 2014. "What is Big Data?" Retrieved 11/6/14. (http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html).

Ignatow, Gabriel. 2003. "'Idea Hamsters' on the 'bleeding Edge': Profane Metaphors in High Technology Jargon." *Poetics* 31(1):1–22.

Ignatow, Gabriel. 2004. "Speaking Together, Thinking Together? Exploring Metaphor and Cognition in a Shipyard Union Dispute." *Sociological Forum* 19(3):405–33.

Ignatow, Gabriel. 2014. "Ontology and Method in Cognitive Sociology." *Sociological Forum.* 29(4): 990-994.

Institute for Advanced Analytics. 2017. *Degree Programs in Analytics and Data Science – Master of Science in Analytics | Institute for Advanced Analytics*. Retrieved March 21, 2017 (http://analytics.ncsu.edu/?page_id=4184).

Jasanoff, Sheila. 2005. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton, NJ: Princeton University Press.

Kalil, Tom, Jim Kurose, and Fen Zhao. 2015. "Big Announcements in Big Data." *Whitehouse.gov*. Retrieved March 21, 2017 (https://obamawhitehouse.archives.gov/blog/2015/11/04/big-announcements-big-data).

Karp, R. M. and M. O. Rabin. 1987. "Efficient Randomized Pattern-Matching Algorithms." *IBM Journal of Research and Development* 31(2):249–60.

Kart, Lisa, Gareth Herschel, Alexander Linden, and Jim Hare. 2016. "Magic Quadrant for Advanced Analytics Platforms." *Gartner.com*. Retrieved December 5, 2016 (https://www.gartner.com/doc/reprints?id=1-2WQY2ZJ&ct=160121&st=sb).

Knorr Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge and London: Harvard University Press.

Knorr Cetina, Karin. 2007. "Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures." *Interdisciplinary Science Reviews* 32(4):361–75.

Kochanek, Kenneth D., Xu, Jianquan, Murphy, Sherry L., Minino, Arialdi M., and Kung, Hsiang-Ching. 2011. "Deaths: Final Data for 2009." *National Vital Statistics Reports* 60(3):1–117.

Krzanich, Brian. 2016. "Data Is the New Oil in the Future of Automated Driving | Intel Newsroom." *Intel.com*. Retrieved (https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/).

Kuhn, Thomas S. 1964. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakoff, George and Mark Johnson. 1980/2003. *Metaphors We Live By*. 1st edition. Chicago: University Of Chicago Press.

Lakoff, George and Rafael Nuñez. 2000. *Where Mathematics Come From: How The Embodied Mind Brings Mathematics Into Being*. unknown edition. New York, NY: Basic Books.

Laney, Doug. 2012. "Deja VVVu: Others Claiming Gartner's Construct for Big Data." *Doug Laney*. Retrieved February 27, 2017 (http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/).

Latour, Bruno. 1988. *The Pasteurization of France*. Cambridge, Mass.: Harvard University Press.

Latour, Bruno and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts, 2nd Edition*. 2nd edition. Princeton, N.J: Princeton University Press.

Lyman, Stanford M. and Marvin B. Scott. 1989. *A Sociology of the Absurd*. 2nd ed. Altamira Press.

Mannheim, Karl. [1936] 1968. *Ideology and Utopia: An Introduction to the Sociology of Knowledge.* New York: Harcourt, Brace, and World, Inc.

Mannheim, Karl. [1936] 1985. *Ideology and Utopia: An Introduction to the Sociology of Knowledge.* Louise Wirth and Edward Shils (trans). San Diego, New York, London: Harcourt Brace Jovanovich.

Mannheim, Karl. 1993. *From Karl Mannheim*. edited by K. Wolf. New Brunswick and London: Transaction Publishers.

Manyika, James et al. 2011. *Big Data: The next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. Retrieved (http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation).

Markham, Annette N. 2013. “Undermining ‘data’: A Critical Examination of a Core Term in Scientific Inquiry.” *First Monday* 18(10). Retrieved October 20, 2016 (http://firstmonday.org/ojs/index.php/fm/article/view/4868).

Martin, Emily. 1991. “The Egg and the Sperm: How Science Has Constructed a Romance Based on Stereotypical Male-Female Roles.” *Signs* 16(3):485–501.

Mayer-Schönberger, Viktor and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. edited by N. W. Storer. Chicago: University Of Chicago Press.

Mills, C. Wright. 1940. “Situated Actions and Vocabularies of Motive.” *American Sociological Review* 5(6):904–13.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London and New York: Verso.

Nafus, Dawn and Jamie Sherman. 2014. “Big Data, Big Questions| This One Does Not Go Up To 11: The Quantified Self Movement as an Alternative Big Data Practice.” *International Journal of Communication* 8(0):1784–94.

National Academy of Sciences, National Academy of Engineering, and and Institute of Medicine. 2009. *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition*.

Norvig, Peter. 2011. “On Chomsky and the Two Cultures of Statistical Learning.” Retrieved March 9, 2017 (http://norvig.com/chomsky.html).

Olick, Jeffrey K. 1999. “Genre Memories and Memory Genres: A Dialogical Analysis of May 8, 1945 Commemorations in the Federal Republic of Germany.” *American Sociological Review* 64(3):381–402.

Oudshoorn, Nelly and Trevor Pinch, eds. 2005. *How Users Matter: The Co-Construction of Users and Technology*. Cambridge, MA and London: The MIT Press.

Parasie, Sylvain. 2015. “Data-Driven Revelation?: Epistemological Tensions in Investigative Journalism in the Age of ‘Big Data.’” *Digital Journalism* 3(3):364–80.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. 1 edition. Cambridge: Harvard University Press.

Peck, Don. 2013. “They're Watching You at Work” *The Atlantic*. Retrieved 12/3/14 (http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/).

Peters, John Durham. 2014. “Cloud.” *Culture Digitally*. Retrieved December 5, 2016 (http://culturedigitally.org/2014/06/cloud-draft-digitalkeywords/).

Petre, Caitlin. Working Paper. “Becoming Data: Web Analytics, Journalism, and Emotional Dimensions of Rationalizing Technologies.”

Pinch, Trevor J. and Wiebe E. Bijker. 1984. “The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other.” *Social Studies of Science* 14(3):399–441.

Pink, Sarah et al. 2016. “Data Stories.” *Data Ethnographies Lab*. Retrieved January 17, 2017 (https://dataethnographies.com/paper-iv-data-stories/).

Podesta, John, Penny Pritzker, Ernest J. Moniz, John Holdren, and Jeffrey Zients. 2014. “Big Data - Seizing Opportunities, Preserving Values.” Retrieved November 1, 2015 (https://www.whitehouse.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf).

Pollner, Melvin. 1987. *Mundane Reason: Reality in Everyday and Sociological Discourse*. Reissue edition. Cambridge University Press.

Poovey, Mary. 1998. *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. 1 edition. Chicago: University Of Chicago Press.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Reprint edition. Princeton, N.J: Princeton University Press.

Power, Michael. 1997. *The Audit Society: Rituals of Verification by Michael Power*. Oxford and New York: Oxford University Press.

Press, Gil. 2013. “A Very Short History Of Big Data.” *Forbes.com*. Retrieved March 17, 2017 (https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#2fe80e3b65a1).

Pugh, Allison J. 2013. “What Good Are Interviews for Thinking about Culture? Demystifying Interpretive Analysis.” *American Journal of Cultural Sociology* 1(1):42–68.

Puschmann, Cornelius and Jean Burgess. 2014. “Big Data, Big Questions| Metaphors of Big Data.” *International Journal of Communication* 8(0):20.

Quinn, Naomi. 1991 "The Cultural Basis of Metaphor." In Beyond Metaphor: The Theory of Tropes in Anthropology, ed. J. W. Fernandez. Stanford, California: Stanford University Press. Pp. 56-93.

Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. 2004. "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking." *Columbia Law Review* 104(4):1150–1210.

Reed, Isaac Ariail. 2011. *Interpretation and Social Knowledge: On the Use of Theory in the Human Sciences*. Chicago ; London: University Of Chicago Press.

Ribes, D., & Jackson, S.J. 2013. "Data Bite Man: The Work of Sustaining a Long-Term Study." Pp. 147-166 in Gitelman, L. (Ed.), *"Raw Data" is an Ooxymoron*. Cambridge, Massachusetts; London, England: The MIT Press.

Ricoeur, Paul. 1978. *The Philosophy of Paul Ricoeur: An Anthology of His Work,* edited by Charles E. Reagan and David Stewart. Boston: Beacon Press.

Robert H. Tai Research Group. 2012. "Informing Medical Decision-Making Based on Heart Rate Characteristics Monitoring." [Transcribed Interviews]. Used with the permission of the principle investigator.

Rosenberg, Tina. 2015. "Turning to Big, Big Data to See What Ails the World." *The New York Times Opinionator*. Retrieved March 21, 2017 (https://opinionator.blogs.nytimes.com/2015/04/09/turning-to-big-big-data-to-see-what-ails-the-world/).

Rotella, Perry. 2012. "Is Data The New Oil?" *Forbes*. Retrieved October 26, 2016 (http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/).

Ryall, Emily. 2008. "The Language of Genetic Technology: Metaphor and Media Representation." *Continuum* 22(3):363–73.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3(3):210–29.

Schatzki, Theodore R. 2001. "Introduction" in *The Practice Turn in Contemporary Theory.* Schatzki, Theodore R., Knorr Cetina, Karin, and von Savigny, Eike (eds). London and New York: Routledge.

Schudson, Michael. 2006. "The Trouble with Experts – and Why Democracies Need Them." *Theory and Society* 35(5–6):491–506.

Schreibman, Susan, Ray Siemens, and John Unsworth. 2004. *Companion to Digital Humanities (Blackwell Companions to Literature and Culture)*. Hardcover. Oxford: Blackwell Publishing Professional. Retrieved (http://www.digitalhumanities.org/companion/).

Seaver, Nick. 2015. “The Nice Thing about Context Is That Everyone Has It.” *Media, Culture & Society* 37(7):1101–9.

Sennett, Richard. 2006. *The Culture of the New Capitalism*. Yale University Press.

Seyfert, Robert and Jonathan Roberge. 2016. *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies*. London ; New York: Routledge, Taylor & Francis Group.

Shapin, Steven. 1995. “Here and Everywhere: Sociology of Scientific Knowledge.” *Annual Review of Sociology* 21(1):289.

Shaw, Jonathan. 2014. “Why ‘Big Data’ Is a Big Deal.” *Harvard Magazine*, February 18. Retrieved January 25, 2017 (http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal).

Sicular, Svetlana. 2013. “Gartner’s Big Data Definition Consists of Three Parts, Not to Be Confused with Three ‘V’s.” *Forbes*. Retrieved February 27, 2017 (https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/print/).

Sismondo, Sergio. 2008. “Science and Technology Studies and an Engaged Program.” in *The Handbook of Science and Technology Studies*, edited by E. J. Hackett, O. Amsterdamska, M. E. Lynch, and J. Wajcman. Cambridge and London: The MIT Press.

Smolan, Sam. 2016. *The Human Face of Big Data*. Against All Odds Inc.

Striphas, Ted. 2015. “Algorithmic Culture.” *European Journal of Cultural Studies* 18(4–5):395–412.

Thrift, Nigel. 2005. *Knowing Capitalism*. SAGE.

Tijerina, Bonnie. 2016. **“**Campus Support Systems for Technical Researchers Navigating Big Data Ethics**.”** *Educause Review*. Retrieved November 30, 2016. (http://er.educause.edu/articles/2016/6/campus-support-systems-for-technical-researchers-navigating-big-data-ethics).

Timmermans, Stefan. 2010. “Evidence-based medicine: sociological explorations.” Pp. 309-323 in Bird, C. E., Conrad, P., Fremont, A. M., & Timmermans, S. (Eds.), *Handbook of medical sociology*, (6th ed.). Nashville: Vanderbilt University Press.

Turner, Stephen. 2001. “What Is the Problem with Experts?” *Social Studies of Science* 31(1):123–49.

Turner, Stephen P. 2003. *Liberal Democracy 3.0 : Civil Society in an Age of Experts*. London, Thousand Oaks, CA, New Dehli: SAGE Publications Ltd.

Ungerleider, Neil. 2013. "Colleges Are Using Big Data To Predict Which Students Will Do Well—Before They Accept Them" Oct 21, 2013 Retrieved 12/3/14. (http://www.fastcoexist.com/3019859/futurist-forum/colleges-are-using-big-data-to-predict-which-students-will-do-well-before-the).

Vaisey, Stephen. 2009. "Motivation and Justification: A Dual-Process Model of Culture in Action." *American Journal of Sociology* 114(6):1675–1715.

van Dijck, J. 2014. "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." *Surveillance & Society* 12(2): 197-208.

Wagner-Pacifici, Robin. 1994. "A Framework for Articulating Horror," pp. 1-10 in *Discourse and Destruction: The City of Philadelphia versus MOVE*. Chicago: University of Chicago Press.

Weber, Max. [1952] 1991. "Science as a Vocation," pp. 129-158 in *From Max Weber*. H.H. Gerth and C. Wright Mills (eds). Abingdon: Routledge.

Winter, Steven L. 2001. *A Clearing in the Forest: Law, Life, and Mind*. Chicago: University Of Chicago Press.

Wuthnow, Robert. 1987. *Meaning and Moral Order: Explorations in Cultural Analysis*. Berkeley and Los Angeles: University of California Press.

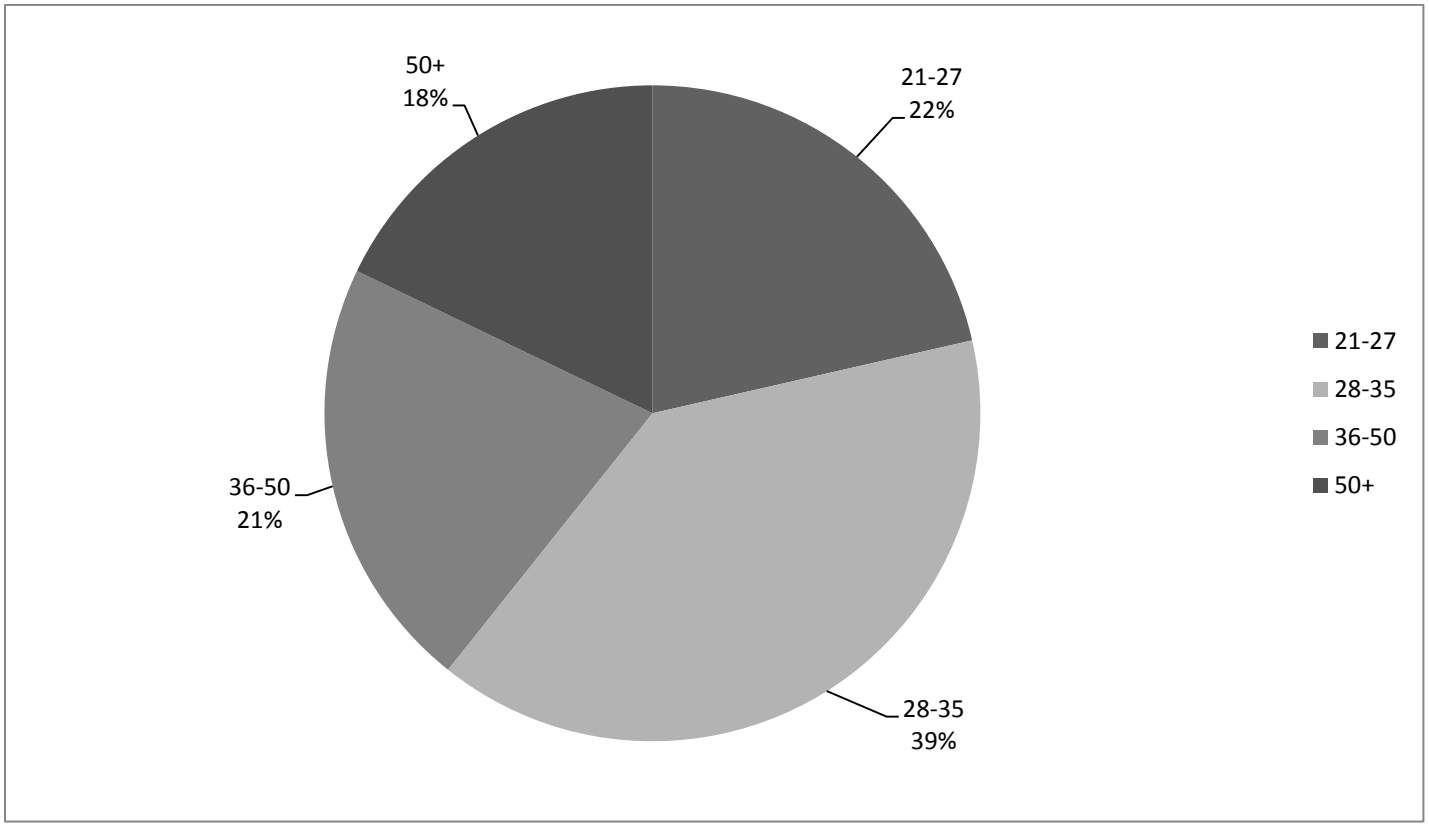
Wynne, Brian. 2004. "May the Sheep Safely Graze? A Reflexive View of the Expert-Lay Knowledge Divide." in *Risk, Environment and Modernity: Towards a New Ecology.* S. Lash, B. Szerszynski, and B. Wynne (eds). Sage.

Zammito, John. 2007. "What's 'New' in the Sociology of Knowledge." in *Philosophy of Anthropology and Sociology.* Stephen Turner and Mark Risjord (eds). Elsevier.

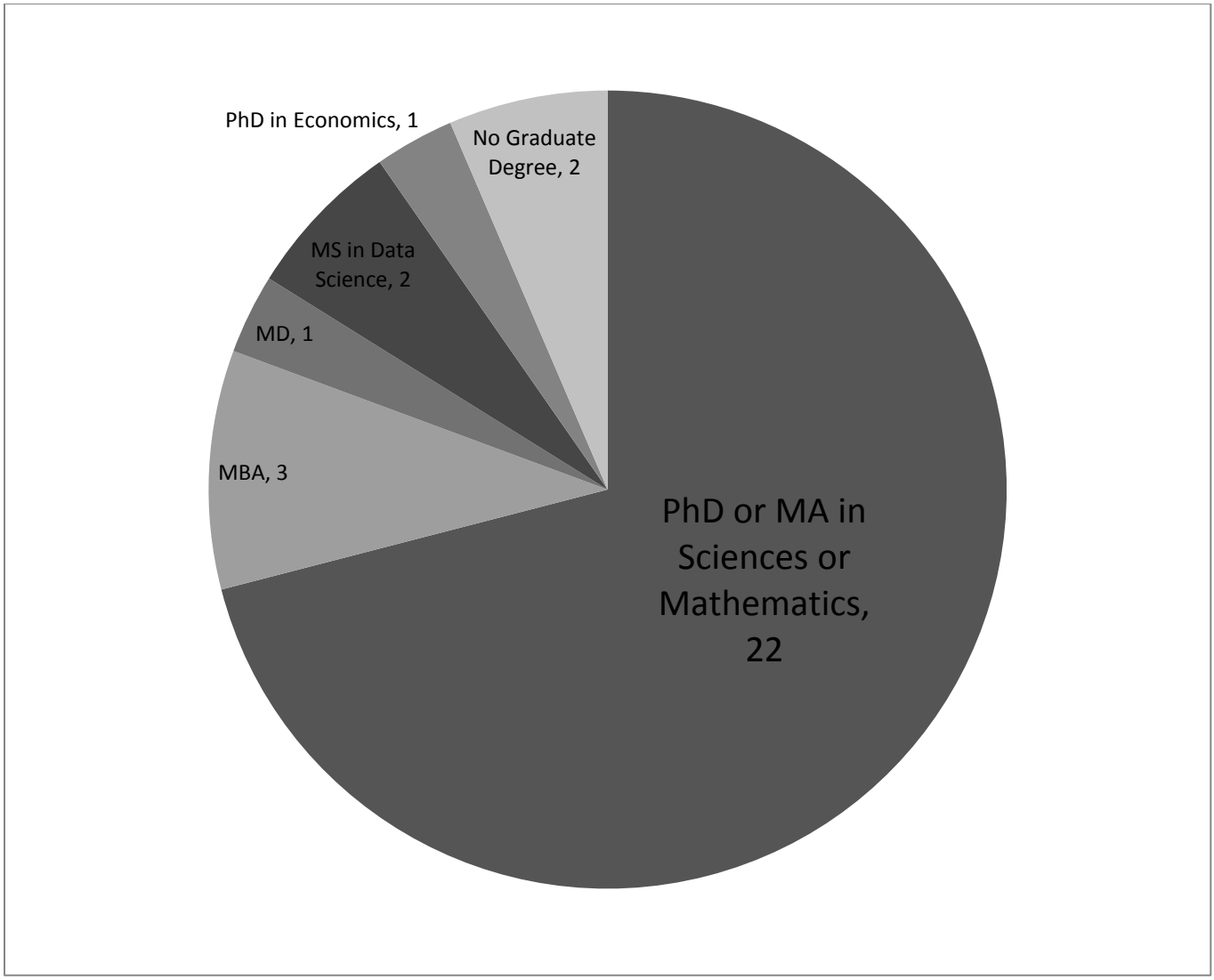
Appendix A:

Data Scientist Interviewee Sample

Percentage of Interviewees According to Age Ranges



Graduate Degrees among Data Scientists by Subject Area



Appendix B: White Paper Sample

White Paper Title	Organization(s)	# of pages	Year*	link
How to Improve Efficiency and Quality in Hospital Healthcare	Actuate Analytics	10	x	http://birt.actuate.com/resources/white-papers
Build Data-Driven Visualizations, Dashboards and Reports	Actuate Analytics	2	2015	http://birt.actuate.com/resources/white-papers
Four Pillars of Business Analytics	Actuate Analytics	6	2015	http://birt.actuate.com/resources/white-papers
You're Doing it Wrong: How to Get Data Science Right	Alpine Data	10	2015	http://alpinedata.com/white-paper-how-to-get-data-science-right/
Unlocking Big Data	Alteryx	8	2015	http://pages.alteryx.com/rs/alteryx/images/Unlocking%20Value%20from%20Big%20Data%20for%20CSPs.pdf
Finding truth	Alteryx	6	2014	http://pages.alteryx.com/rs/alteryx/images/alt-qlik-wp-spatial-analytics.pdf
Field Guide to Data Science	Booz Allen	110	2013	http://www.boozallen.com/insights/2015/12/data-science-field-guide-second-edition
5 Ways to Boost Your Business IQ	Dell	5	2014	http://i.dell.com/sites/doccontent/business/solutions/whitepapers/en/Documents/5ways_to_boost_your_business_IQ_White_Paper_Dec_2014.pdf
Predictive Supply Chain	DHL	12	2016	http://supplychain.dhl.com/LP=704?nu_ref=Vanity-URL
Data Mining with Qualitative and Quantitative Data	Elder Research	6	2010	http://resources.idgenterprise.com/original/AST-0036087_DataMining.pdf
The Future of Deciding	FICO	14	2016	http://www.fico.com/en/latest-thinking/white-papers/the-future-of-deciding
How You Can Use Data Analytics to Change Government	GovLoop	40	2015*	https://www.govloop.com/resources/how-you-can-use-data-analytics-to-change-government/
Better Decisions Through Data Analytics	HFMA	5	2015*	https://images.magnetmail.net/images/clients/HFMA/attach/HFMA2016Seminars_WhitePaper_DataAnalytics.pdf
Seven Capabilities in Performance Management	IBM	12	2015	http://www-01.ibm.com/software/analytics/learn-center/index.jsp?path=Business/Finance/Finanancial-Performance-Management/&featured=464

Transforming the Way Organizations Think with Cognitive Systems	IBM	5	2012	http://www.redbooks.ibm.com/abstracts/redp4961.html
The Era of Cognitive Systems: An Inside Look at how Watson Works	IBM	16	2012	http://www.redbooks.ibm.com/abstracts/redp4955.html
Starting the Workforce Analytics Journey	IBM	26	2015	http://cng.files.cms-plus.com/IBM%20Starting%20the%20Workforce%20Analytics%20Journey%20June%202015.pdf
Customer Segmentation Comfortably from a Web Browser: Combining Data Science and Business Analytics	KNIME	18	2016	http://www.knime.org/files/white-papers/customer_segmentation.pdf
Risk Scoring: Big Data and Andvanced Analytics Further Evolve the Healthcare Model	Knowledgent and Teradata	10	2015	https://knowledgent.com/whitepaper/risk-scoring-big-data-and-advanced-analytics-further-evolve-the-healthcare-model/
Datalake Opportunities in Health Care	Knowledgent	9	2015	https://knowledgent.com/whitepaper/data-lake-opportunities-in-healthcare/
Nara Logics' Synaptic Network	Nara Logics	8	2013*	https://brainstorm.naralogics.com/nara-logics-synaptic-network
Oracle: Big Data for Enterprise	Oracle	16	2013	http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf
Palentir Cyber	Palentir	16	2014	http://www.palantir.com/wp-assets/wp-content/uploads/2014/03/Solution-Overview-Palantir-Cyber.pdf
Palentir Health	Palentir	7	x	http://www.palantir.com/wp-assets/media/capabilities-perspectives/Palantir-Health.pdf
How Advanced Analytics is Providing Insight into How to Win PPC Customers	Peppersack	15	2016	https://www.peppersack.com/whitepaper
Data Science, Big Data, and SEO	Peppersack	14	2016	https://www.peppersack.com/whitepaper
Predixion: Improving Predictive Maintenance Time-to-Value with Real-Time Visual Edge Analytics	Wind River	7	2016	http://events.windriver.com/wrcd01/wrcm/2016/08/WP-HDC-Predixion.pdf
Big Data Meets Big Data Analytics	SAS	13	2012	https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf
The Early Warning Project on Predictive Maintenance	SAS	26	2008	https://www.sas.com/en_us/whitepapers/early-warning-project-on-predictive-maintenance-103617.html

Predictive Analytics: Revolutionizing Business Decision Making	TWDI and SAS	9	2014	http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/tdwi-predictive-analytics-107459.pdf
5 Common Pitfalls of Data Analytics	Sunera	19	2015	https://focal-point.com/article/data-analytics/five-common-pitfalls-data-analytics
Text Data Analytics: In Service of Smart Government	Teradata	8	2016	http://www.teradata.com/Resources/White-Papers/Text-Data-Analytics-In-Service-of-Smart-Government/
Data Science: The Engine to Power Next-Generation Cybersecurity	ThreatTrack Security	5	2015*	http://land.threattracksecurity.com/Data-Science-The-Engine-to-Power-Next-Generation-Cybersecurity.html