THE RHETORICS OF DATA: INSIGHT AND KNOWLEDGE-MAKING AT A NATIONAL SCIENCE LABORATORY

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Trinity C. Overmyer

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF DISSERTATION APPROVAL

Dr. Patricia Sullivan, Chair

Department of English

Dr. Jennifer Bay

Department of English

Dr. Michael Salvo

Department of English

Dr. Benjamin Sims

Los Alamos National Laboratory

Approved by:

Dr. Manushag Powell

Head of the English Graduate Program

For Pearl and Vivi

ACKNOWLEDGMENTS

What seemed impossible up until the day of the defense became possible only with the support, guidance and enduring patience of an entire community of people. I am most grateful to all of my participants, who let me into another world of thought and research practice. I want to thank my advisor, Pat Sullivan, for helping to tease all the weird ideas out of my brain and into coherent thoughts, not just through the dissertation, but since the beginning of my graduate work. Thank you for supporting me, making me laugh, and for every word of insight and encouragement. Special gratitude goes to Jenny Bay, who is a model of mentorship, even going back to my UG days. Thank you for being fierce, smart and supportive beyond measure, for making me suck it up and get to work when I needed to, and for teaching me almost everything I know about engagement work, which is above everything else in my work, the thing that makes my heart soar. I am grateful for Michael Salvo, who has listened to me laugh and cry more than any other faculty member over the last seven years. You never cease to get excited with me when I have a new idea, and your sarcastic wit is contagious. Ben Sims, who went above and beyond as a committee member, mentor and advocate, you have my unending gratitude. I would have never imagined I could find someone at the lab who was so insightful about my interests and so selflessly supportive of the work. Getting to work with you was by far one of my favorite things about this process. I think it is rare to not only have such an impressive group of scholars on one committee, but also have it be filled with people I can call friends. I also want to thank the rest of the faculty as well, each of whom

have been tough, demanding, understanding, funny, kind and encouraging over the years.

To Sherri Craig, thank you for showing me being my tribe, for calling me out, for taking me at my worst and for the hundreds of hour-long (+) conversations that made me feel at home in the world. I could spend pages thanking all the friends who made this process unforgettable, especially Shane Kim, Don Unger, Erin Brock Carlson, Carrie Grant, Talisha Morrison, Isaac Wang, and Tony Bushner.

Stephen Harrell, thank you for being a bear, and for always reminding me I was good enough to make it this far. Thank you for musing me at a moment's notice, for every Atilla and snowshoe and window washing. You have been so supportive through all of this. You changed my life from the jump.

Finally, I'm ever grateful to my family: Diane, Kari, Travis, Jen, Case, Libby, Mariah, Ty, Pearl and Vivi Overmyer. Kari, you are my rock and my cherished one. I'll never be able to fully tell you all how much I love you, how grateful I am that you stood by me every day, and how you've all taught me more than a degree ever will.

TABLE OF CONTENTS

Page

LIST OF FIGURES viii

ABSTRACT ix

1 INTRODUCTION 1
1.1 Data Rhetorics 1
1.2 Site Selection 3
1.3 Chapter Overviews 4

2 DATA AS RHETORIC: A REVIEW OF THE LITERATURE 7
2.1 From Condemnation to Praxis 7
2.2 The Boundaries of Big Data: Definitions Beyond Size 8
2.2.1 History 9
2.2.2 Sensemaking 11
2.3 Corporate Vs. Scientific Big Data 14
2.4 The Expertise of Technical Writers 18
2.4.1 How Technical Writing Scholarship Approaches Data Rhetorics 19
2.5 Conclusion 24

3 METHODS, SITE SELECTION AND CODING MECHANISMS 29
3.1 Introduction 29
3.2 Research Questions 30
3.3 Dueling Institutional Review Boards (IRBs) 31
3.4 Researcher Positioning: Inside-Outsider 32
3.5 Research Methods 37
3.5.1 Interviews 37
3.5.2 Sampling and Recruitment 38
3.5.3 Informants 39
3.5.4 Privacy, Security and Technology 40
3.6 Data Analysis 41
3.6.1 Grounded Theory 42
3.6.2 Research Memos 43
3.6.3 Coding Structures 45
3.7 Challenges and Limitations 46

4 CHAPTER 4. DATA SETTINGS OVER DATA SETS 51
4.1 The Cases 51
4.2 Case 1: Data Workflows 51

Page

4.3 Case 2: The Analysts . . . 57
4.4 Case 3: The Novices . . . 62
4.4.1 Hackathon . . . 64

5 "WHAT CAN I MAKE OF THIS?": ANALYSIS AND DISCUSSION . . . 71
5.1 Data Settings Over Data Sets . . . 71
5.2 Data Dictionaries and Semantics . . . 72
5.3 Data Proxies . . . 76
5.4 Developing a Common Language . . . 82
5.5 Data-Driven Versus Situated Analysis . . . 85
5.5.1 Narrative and Problem Setting . . . 85
5.5.2 Expert Vs. Novice Analysis . . . 87
5.5.3 Composing Models . . . 88
5.5.4 Problem Setting . . . 90
5.5.5 Issues with Forgoing Data-Driven Approaches . . . 92
5.5.6 A Mix of Problem and Data Driven . . . 93
5.6 The Roles of Subject Matter Expertise in Data Analysis . . . 95
5.7 Structured Learning in Interdisciplinary Data Work . . . 98

6 DATA RHETORICS IN CURRICULA, INDUSTRY AND ENGAGEMENT 101
6.1 The Intellectual Work of Technical Communication Experts . . . 101
6.2 Data Rhetorics in Professional and Technical Writing . . . 103
6.2.1 Why Data is a Rhetorical Topic Needed in Technical Writing Programs . . . 103
6.3 Teaching Rhetorics of Data . . . 105
6.4 Professional and Technical Communicators in Data-Centric Industries 108
6.5 Is Doing Good with Data Enough? . . . 111
6.6 Future Research Directions . . . 113

REFERENCES . . . 116

LIST OF FIGURES

Figure Page

2.1 Comparison among definitions of big data, as outlined by (Kitchin & McArdle, 2016), (Kitchin, 2014), (Lupton, 2015), and (Boyd & Crawford, 2012). 12

3.1 Open Codes: Initial Coding 47

3.2 Axial Codes: Final Categories 48

3.3 Axial Code Density Across All Interviews 49

ABSTRACT

Overmyer, Trinity C. Ph.D., Purdue University, August 2020. The Rhetorics of Data: Insight and Knowledge-making at a National Science Laboratory. Major Professor: Patricia Sullivan.

This dissertation details one of the first lines of inquiry into the rhetorical strategies used in scientific data analysis. The study primarily concerns the relationships between data work and knowledge making in the analysis of so-called "big data," and how rhetoric and technical communication theories might inform those relationships. Hinging on five months embedded at a national science laboratory, this study uses ethnographic methods to detail the ways in which data analysis is neither purely data-driven and objective, nor purely situated in a local context or problem. Rather, data work requires both analytical processes and artful *techne* embedded in ongoing reflective praxis. As purely analytic, data work focuses on mathematical treatments, step by step procedures and rote formulas. As *techne*, data work requires interpretation. Rhetorical data analysis is not the opposite of data-driven work. Instead, rhetorical *techne* stands as the midpoint between the extremes of purely data-driven and purely context-driven analysis. Based on three cases that compare the practices of data novices, seasoned experts, and interdisciplinary teams, I argue that the ways in which scientists go about their data cleaning, collaboration, and analysis change based on their levels of expertise and the problem at hand. A number of principles that outline how data analysis is a form of rhetorical inscription are also defined, including the ways data dictionaries, model building and the construction of proxies intimately link scientific insights with language. The set of principles detailed in this dissertation are key areas that should be considered in both data science education

and professional and technical writing curricula. Therefore, the project should be of particular interest to instructors and administrators in both Technical Writing and Data Science programs, as well as well as critical data studies scholars.

1. INTRODUCTION

1.1 Data Rhetorics

The early decades of this century have been ruled by data. Decisions are made (and will continue to be made, possibly ad infinitum) based on the insights we glean from data analysis. There is promise and peril here. Big data, as both a concept and field of practice, wrestles with a "transparency paradox" (Richards & King, 2013, pp.42) while the myth of big data professes that access to more information equals better knowledge and a transparent understanding of the world, in practice, analytic data methods still largely exist inside a black box. Although data analysis frameworks and standards are emerging, ad hoc sense-making across sites, disciplines, and industries generates a lot of variability in data methods (Council, 2015). Even the technological tools of data analysis, like machine learning algorithms and their outputs, are not well-studied.

Although a generation has come to pass since Haraway (Haraway, 1988), Suchman (Suchman, 2002), and other feminist scholars argued against rendering scientific knowledge as objective and all-encompassing, the myth of big data and seeming transparency of data arguments has yet to be fully excavated through such lenses. In this myth, data science is rendered as its own ontic beast that is absolved of needing anything as trivial as interpretation. Data is transparent an cannot be denied or argued against. Data merely *is*. The study of data rhetorics is a fertile point of departure for technical communication scholars to examine the *techne* of big data as a mode of inquiry, and engage critically with data work as a medium of inscription. Studying data is a methodological and epistemological project where technical writers can pro-

ductively intercede. Data sets are not raw—they are productions (Gitelman, 2013). At each point in its lifecycle, data is constructed through interpretation. Big data sets are uniform but multiple; vast, but flat. And, at the same time, human. They are entangled with human processes of collection, sensemaking, and the transmission of meaning. Data are not a-rhetorical. Data are constructed through human-material-discursive ecologies and have influence in those ecologies, too. The era of big data brings with it a civic and professional imperative—to understand how knowledge is produced and deployed through big data. Renouncing big data because of its correlations with positivism, surveillance and power hierarchies closes off any prospect of using big data as a tool for insight and empowerment. Instead of running from big data, technical writing scholars might ask what "counter-data'‘ measures can be taken (Iliadis & Russo, 2016, pp.4). One such measure might be for critical scholars to work from the inside and become data-driven researchers (Leurs, 2017); (McNely, Spinuzzi, & Teston, 2015). Another way rhetoricians and technical writers might intercede is to interrogate the methodological and ontological assumptions made during data collection, cleaning, and analysis. For that, we must "follow the data scientists'‘ (Thatcher et al., 2016) (see also (Latour, 1987), which is what this project has done.

This study of rhetoric as a driving force in data analysis focuses on the epistemological and sociotechnical means through which data analysis and data arguments are constructed. In this research study, I attempt to reimagine data as a technological process along with the interfaces and computational processes that accompany data's deployment in the sciences. Similar to feminist technoscientific scholars working during the onset of the computational age (see the work of Sylvia Wynter, Lucy Suchman, Isabelle Stengers, Susan Leigh Star, Judith Butler, Donna Haraway, Sandra Harding), I aim to pursue an understanding of data work as a dynamic process. Instead of taking data as closed and unbiased, this project considers how even scientific argu-

ments are, "multiple, located, partial perspectives that find their objective character through ongoing processes of debate'' (Suchman, 2002, pp.92). Likewise, scientific data work itself is a continuously evolving set of skills that have to work across many settings and problem spaces, embedded in interwoven knowledge domains.

My research is not a proponent for a global set of standards when it comes to data work, and my goal is not to chastise the individuals, institutions or domains that rely heavily on data for their research practices. Instead, I aim to acknowledge and define the actual practices that have been developed through praxis, social networks and situated expertise over years and generations. I aim to look for "wiser interventions into [the] ecology'' of technologies, humans and action (Bennett, 2010, pp.4). Likewise, my goal is not to delineate a universal standard of practice, but to hone in on a starting point with partial models of how knowledges are made (Haraway, 1988, pp.187). Tracing the "relations of technology production and use'' (Suchman, 2002, pp.93) in terms of scientific data work is no small task. This project is only a small expedition into an entire world of similar relations that stretch from laboratory assistants all the way to the future of exascale computing and beyond.

1.2 Site Selection

Though normally sites of study are identified in a later phase after preliminary data collection, I purposefully selected LANL as the core site from the beginning of the project for several reasons. First, LANL has historically been at the forefront of big data research, computing and data visualization. Even more than most national labs, LANL attracts domain experts from across the globe, and emphasizes collaboration across disciplines in their project teams. Second, LANL staff have spearheaded research focused on how to combine quantitative and qualitative data to perform "bottom-up'' visual data analysis which is situated in the specific characteristics of

the data set, rather than born from a decontextualized list of best practices (Ahrens et al., 2006). In this work, Ahrens also acknowledges the role that "intuition'‘ plays in analyses and discovery of big data analysis and visualization, which aligns well with my lines of inquiry. Additionally, the ISTI program (Information Science and Technology Institute) offered me the chance to work full time at the lab each year while I conducted my research, which allowed me to integrate into the community culture I studied. I was given a workspace, as well as access to archives, project meetings, visiting scholars and staff scientists. At LANL, I was able to be a coworker with much more access to the community than I would have had researcher a group of scientists from the outside.

1.3 Chapter Overviews

Chapter 2 reviews literature concerning critical data studies; the differences between conceptions of data, evidence and fact; sociotechnical systems; sensemaking; and a discussion of the differences between scientific and corporate big data methods. Included here is a delineation of the ways the term *big data* is defined, both by its ontic and its rhetorical qualities. I then go on to discuss how Technical Communication approaches concepts of expertise and rhetorical notions of data work. I then discuss how Technical Communication as field approaches issues of expertise and rhetorical notions of data work. Through the lenses of the scholars introduced, I outline the ways in which data and its attending practices and technologies is a rhetorical construction rather than transparent fact.

In Chapter 3, I review the methods, methodologies and techniques that undergird the empirical study. I describe why Los Alamos National Laboratory was chosen as the primary focus of this study because of its cutting-edge data science and visualization division. I explore the tensions between researcher and practitioner that emerged

over the course of my work, and argue for positioning myself in the belly of the beast, so to speak. Specific methods such as interviews and recruitment are discussed, along with Grounded Theory and the ways in which I created coding structures throughout the study to analyze the data. Limitations and ethical concerns are highlighted in this chapter as well.

Chapter 4 is a short description of each of the three illustrative cases I constructed to compare different processes of data work. The first case focuses on a project concerning data workflows with a large collaboration of interdisciplinary scientists. Case 1 represents structured, well-planned work of a group of expert scientific data analysts. Case 2 highlights a data sprint that occurred over the course of four days. This case describes how the sprint structure became a site of contention as analysts were forced out of their typical workflows. The final case describes a data workshop where data analysis novices were taken through very preliminary steps of working with big data on a supercomputing cluster. This narrative allows me to discuss in Chapter 5 how attendees struggled through some of the more rhetorical aspects of data work, such as the construction of data proxies.

Chapter 5 presents results from my field work. I argue that the ways participants went about their data cleaning, collaboration, and analysis changed based on their levels of expertise and the problem at hand. Developing a common language around the data work among collaborators across disciplines is a key aspect of knowledge making. Data dictionaries and the construction of proxies are of particular importance and illustrate how meaning in data analysis is linked intimately with language. I go on to discuss the differences between concepts of data-driven analysis and situated analysis. Across the three cases, data workers illustrate that analysis is neither purely data-driven or purely situated in a local context, but is an intersection of both when it is successful. Rather than thinking of rhetorical data analysis as the opposite of data-

driven work, it is helpful to think about rhetorical *techne* as the midpoint between the extremes of purely data-driven and purely situated analysis. The construction of models is another activity that requires a lot of human decision-making. In the everyday work of scientific data analysis, models are built from scratch, taken from published literature, or most often, models are remixed from other scientific problems. Finally, I examine where problem setting and subject matter experts fall in situated analysis.

In the final chapter, I explore why data rhetorics should be included in professional and technical writing curricula. The need for publics to have the ability to access and participate in data analytics means that professional and technical writing programs should promote skills that allow for data citizenship. Data work requires analytical processes and is also artful *techne* embedded in ongoing reflective praxis. As purely analytic, data work focuses on mathematical treatments and step by step formulas and procedures. As *techne*, data work requires interpretation. I connect issues of education with areas where technical communication practitioners might find space in industries and fields that deal in big data, and particularly big scientific data. I then briefly discuss the differences between participating in critical data work and merely doing good with data, which can lead to new forms of epistemological colonialism. Finally, I briefly bring up future research avenues that are generative topics for technical writing scholars to consider.

2. DATA AS RHETORIC: A REVIEW OF THE LITERATURE

2.1 From Condemnation to Praxis

The concept of data rhetorics does not fit into customary boundaries of writing studies. Data work is composition, but it is not writing in the traditional sense. It is rife with space for critical technical writing scholars to intercede, but it currently does not have an abundance of professionals specifically educated in writing, ethics, social influence or the host of other areas writing scholars and practitioners excel. Some humanities scholars fear and condemn big data epistemologies, others are working to stake a claim on the trend.

My goal here is to outline some key issues in critical data discourse by touching on the extremes—from cautious to condemnation and from data worship to praxis. I neither denounce or laud big data here. Afterall, history is crowded with fears about how new communication technologies will ruin society or lead to information overload. Just after the Gutenberg's printing press was first wet with ink, people were already complaining about the dangers of having more books than an individual could read (Hobart & Schiffman, 2000). Telephones in the home were touted as immoral because a pair of sweethearts could talk to each other in the intimacy of their own bedrooms. This text is cautiously optimistic of how data can be put to work for the good, and it is also critical of the ways fears about big data confuse the practices and pitfalls of corporatized personal data and scientific or academics' data practice. The persuasive and situated practice of data analysis and visualization is essential

to the research process. Knowledge making is facilitated by data, not *made* by data. There is some academic fear surrounding big data, as if data science tries to remove human meaning from the processes of research or insight. However, in contrast to data fear, complex knowledge work practiced through big data still requires subject matter experts (SMEs). They are crucial to the process, in fact. As discussed later in this dissertation, those who work scientifically with data appreciate the need for SMEs and are critical of false correlation. In fact, early publications from the National Visualization and Analytics Center highlight that data insight is not self-evident and needs careful attention: "This analysis process requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information in the face of rapidly changing situations'' (Thomas & Cook, 2005, pp.22).

In this chapter, I discuss big data as a discursive construct and argue that data can only be defined as *big* in terms of the users' capabilities. After a brief note on the differences between data, fact and evidence, I go on to review literature concerning how knowledge making is a joint venture among sociotechnical systems and is situated, rather than objectively oriented. To counter some of the overt skepticism around data science, I argue that there is a stark difference between big data used by researchers and scientists, and big data collected and deployed for corporate gain, which tends to be the kinds of data power we should specifically turn our critical eye towards. Finally, I close by discussing how writing scholars have approached data studies, and the ways that professional and technical writing scholars and practitioners are poised to add immense value to data-centric work.

2.2 The Boundaries of Big Data: Definitions Beyond Size

Concepts are like mini-units of theory and are especially useful when looking at a term, like big data, which seemingly has no "canonical or historical status'' (Bal

& Marx-MacDonald, 2002, pp.23). When concepts travel across fields of knowledge, they have to be scrutinized, not just put into practice (Bal & Marx-MacDonald, 2002). The ways concepts are used in other disciplines and by other scholars have to be considered before each "trip"‘—that is, each time concepts are deployed. Such is the case with big data. Therefore, I spend considerable space in this chapter mapping out some definitional boundaries of the concept. First, big data is aggregative and granular, existing as a mass of little bits. Second, it is both material and ephemeral. Data only becomes data after it is collected or curated, so the material processes of collection are always engrained in its definition. However, big data is stored digitally by way of magnetic currents on a hard drive, which is translated into ones and zeros, stored in the cloud and also filtered through physical data centers. It is *effervescently solid*. Large data sets are always partial, fluid, situated, strategic and contradictory, even though popular discourse might conceive of them as impartial, complete and unaltered/unalterable. Big data sets often have no recognized author and no singular provider of data. They are at the same time quantifiable and elastic; granular but belonging to an undefined whole. Big data rests in epistemological tension with itself. Like many sites of contention, by exploring how it is understood by a range of actors, how it circulates discursively and how it is wielded methodologically, we may illuminate its innerworkings that belie the covert mythos of big data.

2.2.1 History

It is easy to claim that big data ushered in an entirely new world and had a discrete starting point, but like any paradigm shift, traces of its emergence can be identified long before the term hit the public imaginary or became widespread in professional practice. The term *big data* first appeared in scholarly publications in 2003 (Diebold, 2003) but was widely discussed in certain graphics sectors (Lohr, 2013)

and by NASA Ames researchers in the 1990s (Cox & Ellsworth, 1997). However, in "Big Data, Little History,'' (Barnes, 2013) Barnes locates the emergence of big data in the quantified geography movement of the 1950s (Barnes, 2013), directly after the second World War, when geographers began to use descriptive statistics to theorize the movement of people between urban and suburban areas. Others, such as Tung-Hui Hu in his book, *The Prehistory of the Cloud* (Hu, 2015), trace big data and its sister, cloud computing, through the early ages of community television, even comparing big data to flocks of carrier pigeons (Hu, 2015, pp.xix) along with a host of other digital and analogue starting points. In attending very briefly to its historical traces, I merely aim to illustrate that big data does in fact *have* a history and a discursive life. Though many professionals today consider the term to be more about marketing than substance, *big data* was at one time used to denote an intractable problem—not *big data* as opposed to *little*, but *big* as opposed to what computation methods at the time when the term surfaced could handle. The methodologies of data science come from a mix of statistics and computer science. It is unclear if data science will become a field unto itself or just a job title. Currently, no department or school within the academy is obviously responsible for education in data work; many programs in Data Science that the NSF considers "the most successful'' are interdisciplinary coalitions rather than departments (Berman et al., 2016a). Purdue's own Data Mine is an example of such a university-wide project that aims to educate students from all majors in data science and attending computational methods (Purdue.edu, 2020). Rather than solely focusing on a major, data science is grounded in statistical methods and uses interdisciplinary undergraduate learning communities to connect methods with departmental research thrusts. Most early definitions of big data focused on the size of the data set and computational resource constraints (Symons & Alvarado, 2016). Basic definitions today center around the

three Vs (*volume*, *velocity* and emphvariety), referring only to the data formatting and computation processes. Right now, big data usually entails terascale or petascale (think terabytes and petabytes), but currently supercomputing centers are working on exascale computing, which amounts to over a billion gigabytes or one quintillion calculations per second. Others have expanded to 6 Vs (adding rhetorical elements of emphvalue, emphveracity and emphvisualization), or the 13 Ps, which also focus primarily on rhetorical and social aspects of data and data work (see figure below). Kitchin's bibliographic analysis of scholarly articles that define big data (Kitchin & McArdle, 2016) reveals that of all the characteristics noted, velocity was the least present, even though my own fieldwork discussed in the remainder of this text leads me to believe that researchers in the areas associated with data science note its velocity more than any other characteristic. Kitchin's work uncovers that most of the characteristics of big data also exist in so-called small data, and that many data sets that would be defined as big do not even contain the 3V's. Big data is a contested and widely used term that is defined variously by fields, researchers, and marketing executives but does not have a unified meaning. Big data is jargon. It is a metaphor for a set of practices (Andrejevic, 2014) that should be examined critically.

2.2.2 Sensemaking

Technological processes like big data work give us room for new kinds of research encounters, which can open up to new insights (P. Sullivan, 2017). An encounter requires bumping up against something new and 'other.' An encounter implies attuned *activity*—a site of meaning making through acute attention to the landscape of objects, activities and humans that inhabit a space. Because the meaning making processes that occur in tandem with scientific data are largely unstudied, I spend some time reviewing concepts of knowledge work and sensemaking, which offers another

Ontic Characteristics	Processing & Relational Characteristics	Social & Rhetorical Characteristics
Volume/ Exhaustivity	Scalability	Veracity (Partial, noisy, the understanding that error and uncertainty exist)
Indexicality (Can be organized)	Indexical/Variable (Meaning shifts with contexts)	Indexical/Variable (Meaning shifts with contexts)
Variety	Value/Extensionality/ Polymorphous (Data can be repurposed for many forms of inquiry)	Value(able), Polyvalent
	Visualizable	Visualizable
Granular/Fine-grained	Practice-based	Practice-based

Figure 2.1. Comparison among definitions of big data, as outlined by (Kitchin & McArdle, 2016), (Kitchin, 2014), (Lupton, 2015), and (Boyd & Crawford, 2012).

layer of insight concerning humans as part of data analytic workflows. Knowledge work is not a passive operation, nor is it an abstraction only existing in people's minds. Knowledge is part of a shared, enduring sociotechnical system made up of people, processes, artifacts and culture, which maintains and propagates understanding about the world. In other words, knowledge is infrastructural (Edwards, 2010). Because knowledge is at the same time grounded in human and material existence, and also part of a larger complex infrastructure, it is useful to operate from overlapping epistemologies across differences and disciplines to explore knowledge work (Ingersoll, 2016);(Edwards, 2010). Sensemaking is a form of knowledge work that involves interpreting a situation. Sensemaking also goes beyond interpretation to include active creation of frameworks for understanding, where people impact and change the very events they are attempting to understand (Weick, 1995); (Weick,

Sutcliffe, & Obstfeld, 2005). In the process of sense work, people "establish labels in order to understand what is going on, but then the labels become part of what is going on" (Weick, 1995, pp.538). Within the process of interpretation, sense is both made and changed. Data work, which quite literally requires labels in its existence is an interesting process to consider socially-constructed forms of knowledge. Much sensemaking scholarship focuses on crises or unexpected problems that arise (Weick, 2010). Yet, *sensemaking* and *articulation* can be fruitful concepts when considering non-acute but highly complex labor that happens on granular levels of everyday work (Suchman, 1995); (Suchman & Suchman, 2007); (Star & Strauss, 1999); (A. Strauss, 1985). Sensemaking begins with information acquired through active, intentional exploration. It is stimulus-driven (Weick, 2010). In order for personal perceptions to translate to others, information gets bound into categories, schemas and narratives that try to explain the information—what Weick distinguishes as *enacted* sensemaking (Weick, 2010). Articulation work is active, intentional and often invisible. It is the work that gets things back on track (Star & Strauss, 1999), but it is often discussed as the kinds of work-behind-the-*real*-work that occurs through managers, machines or knowledge infrastructures (Star & Griesemer, 1989); (Bowker & Star, 1991). Articulation work is the making of incremental adjustments and the process of inhabiting a problem space—attuning to articulation means being so finely aware that one is able to see where small adjustments need made. Articulation and sensemaking both require an "education of attention" (Ingold, 2001, pp.167) and attunement to the knowledge environment (Coyne, 2010). Knowledge is forged from a lot of small, incremental adjustments. Like a navigator who continuously adjusts her course in tiny ways in the face of the vast amounts of information around her, sensemaking is a process of wading through and adjusting—of finding the signal in the noise. As her skill develops, the navigator can more easily interact with the envi-

ronment, constantly incorporating new knowledge (Ingold, 2000, pp.55-6). Her education of attention lets her continually respond and adjust her bearings. Knowledge is travel. The practice of knowledge is intimately focused in the problem at hand, being environmentally-oriented—meaning that it is forged from the complex space where disciplinary constructs, ideologies, and personal landscapes intersect. Knowledge is both the intersection and the process of traveling through that intersection (Ingersoll, 2016). See also (Bal & Marx-MacDonald, 2002).

2.3 Corporate Vs. Scientific Big Data

In a paper on *First Monday* in 2013, Tom Boellstorff contended that big data needed a corresponding "big theory'' in order for this new methodological shift to hold meaning rather than relying on spurious correlations (Boellstorff, 2013). He notes that while computers can translate from any language to any other, the computer does not have to "know" what is said. Boellstoroff's fears can be summed up in his statement: "a paradigm of semantics, of understanding, is becoming a paradigm of pragmatics, of search" (Boellstorff, 2013). Such fears have been taken up readily by other social and humanities scholars. Kitchin argues that big data has invoked a new epistemological paradigm, where instead of deductive reasoning—finding answers to questions through experimentation and collecting data—big data promotes inductive reasoning—finding insight by looking primarily and directly in the data (Kitchin, 2014, pp.4). Our ability to collect, store, and run data through algorithms that parse it for us has increased the amount of data available to researchers. The availability of big data also entices people with access to make use of it. So increasingly, scientists are exploiting large data set in their work and according to Kitchin, the logic of big data is seeping into the ways scientific research is done. The idea that correlation and prediction are 'good enough' ends contests the foundations of science itself, which

historically has positioned the real goal as understanding cause-effect relationships. The greatest challenge of data, Kitchin says, is "coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity" (pp. 2). While the first half of this statement holds true for many large data sets, the latter statement that data work occurs without research questions doesn't hold up in scientific settings. He goes on to note that as a new paradigm, big data research can go one of two ways: "Empiricism, wherein the data can speak for themselves free of theory, and data-driven science that radically modifies the existing scientific method by blending aspects of abduction, induction and deduction (pp. 10). In synch with Kitchin, Dalton, Taylor and Thatcher's article on the need for critical data studies (C. M. Dalton, Taylor, & Thatcher, 2016) notes that the bigness of data, the mythology of its infinitude, leads some to believe that "larger is always more clear, more comprehensive, and generally better, furthering a seductive pseudo-positivist research orientation in which "raw'' data becomes epistemic reality" (pp. 5). Approximately 2.5 exabytes of data are produced globally each day. And yes, massive amounts of both 'raw' and organized data are out there awaiting study. Repositories of data do admittedly encourage "data dredging," which means combing the recesses of a data set looking for relationships. Many scholars and data scientists alike criticize this process, seeing it as a-theoretical. Corollaries can always be found, even if they are meaningless in the end. But again, the a-theoretical dredging and aimless data exploration occurs primarily in the public and corporate sectors that rely on big data, not in scientific laboratories, as the rest of this dissertation will discuss. Even Boellstorff, who feared the impending "pragmatics of search" notes that those who actually work with big data know that data on its own, without context or theoretical frameworks "can be illuminating, [but] it is not unprob-

lematic. Any data set offers a limited representation of the world. . . " (Boellstorff, 2013). In disciplines where massive data sets have long existed, researchers are keenly aware of the issues of meaningless pattern identification and the biases that attend data dredging. For most scientific and academic researchers, data is still useless without a conceptual framework to make it meaningful. Still, Kitchen argues big data methods could be a *generative* disruption to science that "reveals relationships and patterns that [researchers] didn't even know to look for" (Kitchin, 2014, pp.4). What he is essentially praising here is the ability for big data sets to function as heuristics that value inquiry over deduction. Large data sets invite the serendipity of searching that undermines positivistic deduction, which is a topic that comes up in archival research. In the edited collection, *Working in the Archives* (Ramsey, Sharer, L'Eplattenier, & Mastrangelo, 2009), the authors spend considerable time heralding the joy and intellectual benefits of serendipity. Ostergaard even contributed an entire chapter dedicated to serendipitous findings. While the editors' introduction, as well as many other archival scholars, warn that "serendipity occurs because of preparation, awareness and hard work'' (Ramsey et al., 2009, pp.5); (See also (Gaillet, 2010). They also agree that accidental discovery through the act of browsing can lead to fruitful work. It is difficult to tease out how methods of exploration in an archive and exploration of a data set could render such opposite lengths of praise and condemnation. However, consider how much data is collected ambiently from consumers through consent that at best is only *technical* consent, and at worst bears no consent at all. The data rich also tend to be the corporations whose data is us: our conversations, information habits, even access to our homes. With such great power in the hands of such companies, it is no wonder there is a lot of trepidation when it comes to data dredging. But again, there is a stark difference between how trained scientists or other scholars approach data and how the average open data set user or

corporate analyst wields it. Most big data analytics is done in the corporate sector by analysts with no scholarly familiarity with the topics at hand, the social or scientific impact, or the specific concern for the data providers. Often business analytics do employ data dredging. Jobs that call for data analysts, which is currently the fastest growing position in the country, rarely require any kind of specific subject matter expertise beyond basic computational and statistical work. Critics of big data stress the need for interdisciplinary work and collaboration with SMEs to avoid at least some of these pitfalls (C. Dalton & Thatcher, 2014). Beyond the question of if such corporate analytics accurately reflect the world, there are huge ethical implications. As Bowker and Gitelman note, when looking for patterns, especially when it comes to human conditions, individual analysts are likely to succumb to bias, including racial and discriminatory conclusions (Bowker & Gitelman, 2013). Bowker and Gitelman recall that sarcoma was considered a Jewish disease and AIDS a gay disease because of the patient corollaries, which led to a systemic lack of research and treatment for these diseases, as well as violence and discrimination for those affected. Identity ceases to matter in the same way when numbers become policy without a deeper questioning of the context and material conditions of data. On the other hand, some scholars have also criticized open data as ushering in a "crisis in empirical sociology'' (Boyd & Crawford, 2012, pp.664). Sociological and user-generated data sets that were once only accessible to scholars and experts are now a few clicks away from anyone, regardless of their expertise or training. On one hand, more access means more opportunities for knowledge to be circulated and citizens to make arguments using this data. On the other, some scholars fear that access plus a lack of theoretical or methodological training means that such data sets can be used irresponsibly. Researchers are increasingly getting pressured by funding bodies to collaborate—not with other disciplines, but with corporations and enterprise (Taylor, 2015). Neither

of these scenarios lends itself to critical methodological work. Those who either work in the corporate sector or whose funding and collaborations depend on it are less likely to craft questions or put forth analyses that would make the company look bad and therefore jeopardize funding, jobs, or access to proprietary data. Those without access cannot evaluate the claims from research that uses proprietary data, nor can they reproduce findings. While scientists and other researchers use big data, they do also employ theoretical frameworks, disciplinary knowledge, and outside subject matter experts when needed. (Sadly, the exception to this rule might be when social science and humanities expertise is required. See Chapter 5 for further discussion.). However, in the hands of mere number crunchers, critics are correct: big data can be a powerful and detrimental tool.

2.4 The Expertise of Technical Writers

Understanding the language of the lab and the ways big data researchers work is crucial in order for trained technical communicators to enter these professional spaces. Interrogating any epistemology means we must first gather and understand the discourse through which the epistemology is articulated (Ingersoll, 2016). Language is a central to sensemaking (Weick, 2010); literacy practices and ways of knowing work in tandem, and cannot be separated from each other (Ingersoll, 2016); (Ong, 2002). Formal and informal, everyday scientific discourses are active co-creators of meaning in the complex workings of the scientific process, and feed directly into how knowledge is constructed. While sociologists of science, like Latour and Woolgar, were interested in how social processes affect scientific knowledge, they did not specifically consider the composing practices of scientific work (Latour & Woolgar, 1986). In order to understand big data epistemology, we need to first understand its language and the rhythms of data composition. In doing so, technical and professional

writers can position themselves as sociotechnical researchers who are integral to the meaning making processes of data work . [1] [2] While technical communicators offer diverse sets of expertise to scientific work, well beyond writing up articles and process documentation, domain scientists can have difficulty understanding technical writing expertise as knowledge work. However, with so much emphasis on interdisciplinarity, collaborative science and big data (See (Council, 2015), which requires finding and articulating patterns in largely unstructured information, trained technical writers have open spaces for research and insight to add. As sociotechnical work becomes more collaborative and more complex, articulation work becomes more invisible but more necessary (Star & Strauss, 1999). Therefore researching scientific workplaces in order to highlight and understand articulation and discursive work as integral to knowledge making is crucial for technical communication scholars looking to forge professional paths in high-tech and scientific environments.

2.4.1 How Technical Writing Scholarship Approaches Data Rhetorics

Several scholars have discussed big data and data-driven arguments as they relate to programmatic assessment in writing (M. Scott, 2017); (Anson, 2008); large-scale trends in student writing (Holcomb & Buell, 2018); and student writing assessment (Comer & White, 2016); (Dixon & Moxley, 2013); (Moxley, 2013);. Recently, *The Journal of Writing Analytics*, which was first published in 2017, began specializing in quantitative assessment of the writing process. While writing assessment and

[1]This statement comes with the awareness that professional and technical writers often cannot enter collaborative research teams on our own markers of traditional expertise, like credentials—instead, we have to continually make the case for our value (Hannah & Arreguin, 2017).

[2]The National Research Council recently issued a report, declaring anyone dealing with digital curation—be they researchers, computer scientists or archivists—must understand the "problems to be addressed, the goals to be pursued, as well as the customary methods, nomenclature, and practices of the fields in which the digital information assets are used" (Council, 2015, pp.63). The report's explicit attention to understanding the literacy and discursive practices of this work is an interesting way to understand how technical writers can position their work within big data research.

big data will likely be inextricably linked in the wake of higher education's move toward efficiency over efficacy, and though the pairing of writing assessment and big data shows promise, particularly in collaboration with corpus linguistics and natural language processing, these topics are quite separate from a discussion of composing, analyzing and communicating data as a rhetorical process. Scholarship in technical writing and rhetoric has called for added quantitative research and analysis education in graduate programs (Colombini & Hum, 2017). In the same year, Albers called for graduate programs to offer a rhetorically grounded education in quantitative inquiry, one that moved beyond simple statistical formulas:

> Teaching quantitative research analysis is teaching to analyze data; it is not teaching how to perform a t test and an analysis of variance (ANOVA). Rather, it is teaching the overall process and critical thinking required to develop an understanding of a data set to ask informed and meaningful questions and to connect the statistics with the context of the study. Quantitative data analysis is not about determining a p value, but it is about understanding relationships within the data and connecting those relationships to the research context. (Albers, 2017, pp.217)

Meloncon argued that the lack of empirical, and especially quantitative research in technical writing has hindered the acceptance of technical writers as expertise in science and high-tech industries (See (Albers, 2017). Her point tracks with various scholars who, for the past 20 years have made cases that writing scholarship in all its forms needs more quantitative research. Fulkerson's essay, "Composition at the Turn of the Twenty-First Century,'' notes that compositionists "have rejected quantification and any attempts to reach Truth about our business by scientific means, just as we long ago rejected 'truth' as derivable by deduction from unquestioned first principles" (Fulkerson, 2005, pp.662). This rejection of quantification and lack of large-scale, comparable research, Fulkerson argued, contributed to the field's lack of focus and coherency. Haswell's "NCTE/CCCC's Recent War on Scholarship" notably

tracks how the flagship journal rarely published empirical work that was replicable, aggregable and data supported (RAD) (Haswell, 2005); See also (Lang & Baehr, 2012). Like Fulkerson, Haswell implicated lack of work that can be extended past a single case study in the lack of cohesion and an understanding of major issues in the field. Other technical writing scholars have also focused on the need for pedagogy that emphasizes rhetorical considerations of data visualization. Wolfe (Wolfe, 2015) argued that although instructors and programs often include instruction on data visualization from the standpoint of visual rhetorics, the interpretive decisions made when choosing how to summarize and aggregate data during the visualization process are often not emphasized. Students in technical writing courses need more instruction and experience presenting data to public audiences and nonexperts (Wolfe, 2009). This is especially true for graduate students aiming to work professionally as science writers (Druschke et al., 2018). Beveridge (Beveridge, 2017) discusses how as data and visualization become more commonplace across public media, writing students need hands-on experience working with data to produce arguments. Beveridge notes here that it is especially important for students to confront issues of privacy, surveillance and other political aspects of data analysis in their writing classes. Dragga (Dragga, 1996) and Allen (Allen, 1996) each published articles examining the ethics of visual rhetoric and the need for technical communicators to understand ethical concerns and deception in data visualization. Later, (Dragga & Voss, 2001) argued that the abundance of technical illustrations and scientific visualization required for complex communication requires technical communicators to adopt a human-centered approach to visualization, whereby death statistics, information on violence and the like should be designed considering the humanity of those affected and the emotions of the audience over the ambivalence of statistics. Likewise, Hepworth's 2017 piece (Hepworth, 2017) advocated for practitioners to actively exercise empathy when cre-

ating data visualizations and add such work to our ethics of data communication. Some scholars have also conducted studies apart from pedagogical concerns on how best to visualize complex data for nonexpert audiences (Sorapure, 2019); (Meloncon & Warner, 2017). By analyzing studies from across a range of disciplines, Meloncon found little cohesion on effective strategies. She argues that technical writing scholars, as experts in visual rhetoric, should take up studies of information visualization. Based on the visualization and usability discourse she encountered in the study, she identified opportunities where technical writing scholars and practitioners might add value to visualization-heavy research and industries, especially in healthcare. Meloncon suggests that technical and professional communicators can contribute:

- An understanding of how context matters by designing and conducting user studies with actual users rather than random control trials
- More dynamic research questions that acknowledge the diversity of patients and publics and attend to the situated ways people access information
- More nuanced analysis of multimodal literacy, how literacy is assessed and taught, and user deficit models.

Owens' work on big data is one of the first pieces in the field of professional and technical communication that specifically addresses how technical writing researchers might dig deep into the ways human actors influence and interpret data (Owens, 2011). As "a species of human-made artifact," akin to other kinds of texts, it becomes important to encourage an understanding of how visualization developers attend to their choices while crafting interfaces or arguments. Salvo's piece, which combined issues of visual rhetoric and big data, argues that big data brings on "emergent genres [that knit] together invention, arrangement, style, memory, and delivery in ways that challenge conceptions of print based literacy and textuality'' and offers a space

that supports multiple narratives, which is rife for studies and practices of rhetorical invention (Salvo, 2012, pp39). Following Owens and Salvo, Firth argues that humans are "often rendered invisible" (Frith, 2017, pp169) in analytics, even though choices about data collection and communicating findings are decisions rather than rote procedures. Secondly, she highlights the importance of technical communication in data analysis and deployment. Data work usually involves a multitude of stakeholders and communication to them positions technical writers as key actors. Additionally, the rise of open data sets means that public access to data arguments is increasing. The goal of making data public is for more people to be informed and make evidence-backed decisions. However, merely granting access to data sets is not enough to make data accessible (Goldsmith & Crawford, 2014). Data has to be formatted, annotated, and often interfaces need to be created that allow potential stakeholders to make sense of the data. This opens a fruitful place to technical writing expertise to intercede. While storytelling has long been under the purview of rhetoricians and writers, new opportunities arise when combining data with narrative. With the understanding that data do not speak for themselves, scholars and practitioners in data visualization have taken interest in data narrative and data storytelling for the last decade (Segel & Heer, 2010); (Hullman & Diakopoulos, 2011); (Hullman et al., 2013); (Kosara & Mackinlay, 2013); (Argenta et al., 2014); (Bach, Wang, Farinella, Murray-Rust, & Henry Riche, 2018), though few if any draw on foundational or contemporary narrative scholarship from English. This year, Danner's piece in *Technical Communication Quarterly* (Danner, 2020) outlined a case of how data professionals approach storytelling. He concluded that narrative construction through data is not merely a static skill, but a rhetorical praxis that has to consider stakeholders and the "rich organizational and situational dynamics'' (pp. 184) of a given situation. Additionally, Danner highlights that data sets often carry numerous statistically sup-

portable insights, which denotes that narratives built through data depend on the craft of choosing the appropriate data depending on the author, the institutional culture, and the audiences. The themes emerging from this work note need for education on how decisions and human influences render data meaningful, an awareness of how to craft data arguments and stories for various stakeholders and publics, and also the need for technical writers to gain a basic understanding of data work in order to enter high tech and other industries that work with and are data-centric.

2.5 Conclusion

Scholars are only beginning to scratch the surface of how knowledge work operates through large scale data. Understanding epistemic data practices requires technical writers to embed in research groups and workplaces so they can attune to the minute interactions and discursive practices that other scholars and text-only studies might miss (Hinrichs, Seager, Tracy, & Hannah, 2017). Focusing on the discursive and knowledge making practices across interdisciplinary teams in big data environments can help professional and technical communicators (PTCs) understand these communities of practice in a way that is absent from the current literature in technical communication. Such a gap warrants a concerted effort on the part of technical writers to carve out space for ourselves and position writing experts as part of those larger knowledge infrastructures. In big data discourse and data-centric labs, there are already scholars discussing rhetoric, though they are not rhetoricians. With that in mind, this dissertation will enter a growing conversation about the place of rhetoric and technical communication in scientific data work. Data work is *techne*, situated in the local context of each encounter with a data set. *Techne* is a fluid knowledge practice and a complex system of small articulations, which allow for navigation through vast data. If *techne* is understood as an always-evolving practice that emerges from

communities, technologies, and complex sets of knowledges, to study *techne* means to study workers in action. Knowledges are grounded in the discourses from which they emerge, and they gain meaning from the human and social practices that produce them. Data work is a rhetorical process, defined by interactions with data and the community of practice through which it is interpreted. The observer and their community, therefore, are active participants in discovering patterns and making meaning. There is a long history of workplace studies that explore rhetoric and PTC as epistemological work. Early studies focused on genre research and investigated areas such as law (Perelman & Olbrechts-Tyteca, 1969), social and civic services (Odell & Goswami, 1982) and scientific research (Bazerman, 1988); (Bazerman, 1985). In the process of understanding the kinds of discourse and arguments used in various workplaces, early scholarship was able to demonstrate writing as an epistemic activity (Winsor, 1996) and a practice which precedes community and collaborative knowledge (Dias, Freedman, Medway, & Par, 2013).Technical writers were situated as key negotiators of meaning (P. A. Sullivan, 1996), fundamental to techno-scientific collaborative work. However, without professionals coming from PTC disciplinary expertise, technical practice in data-intense fields is limited. Those with the authority and access are those who are able to ask the questions. This determines which questions get asked and which ones do not. Currently, data work has to be done by those with high statistical and computational skills, which at present are still mostly male, white, and firmly grounded in STEM disciplinary identities, Such disciplines do not always foster critical engagement with ethics, identity and personhood. As Derrida noted, "Effective democratization can always be measured by this essential criterion... the participation in and access to the archive, its constitution, and its interpretation'' (Derrida, 1996, pp.11). As social and scientific are amassed daily, big data is nothing if not a messy archive of human affairs. In order to uncover a multi-

tude of perspectives, more than just computer scientists and statisticians need to be involved. As big data becomes normalized in discourse and practice, it recedes into the background, becoming more and more ubiquitous, and therefore, invisible (see (Suchman, 1995); (Dourish & Bell, 2011); (Szymanski & Whalen, 2011). Dalton and Thatcher suggest critical researchers use counter-data—i.e., using big data in their own research in ethical, reflective and socially situated ways (C. Dalton & Thatcher, 2014). The other option they offer is to look at the generative possibilities of big data, while keeping a critical eye on the processes that create insight. What new ways might we use data at scale ethically? Are there methodological practices that could negate some of the concern with big data raised earlier in this chapter? Is there a way to use big data to shed light on social problems and diverse knowledges and perspectives rather than slighting them? Critical data scholars must ask questions like "Whose view of the world does the visualization represent?'' (D'Ignazio & Klein, 2016, pp.3), but also whose view of the world does the analysis impact? What is the context of the data? how is it scrubbed and organized? Where does data come from, who is it representing, and for what ends? Strong argues that qualitative researchers have many opportunities with data work (Strong, 2014). For one, qualitative and critical researchers are poised to understand complex interactions and contexts linked to big data sets. Strong suggests qualitative researchers would be particularly valuable in discounting meaningless associations found in data and that such researchers have "the ability to challenge and influence established ways of seeing the world" (Strong, 2014, pp.339). Conversations are occurring in computational and domain sciences about the need for more qualitative (Ahrens et al., 2006) and human-focused big data work (D'Ignazio & Klein, 2016); (D'Ignazio & Bhargava, 2015), though in many respects, the work is still an uphill battle. It is clear that tools, even mathematics, are not neutral (Kwan, 2016); (Dourish, 2016); (Dixon-Roman, 2016); (Mittelstadt, Allo,

Taddeo, Wachter, & Floridi, 2016). Data are *wielded*. The capacity to wield them and the potential to act arises out of the shared relationship between human actors and our tools. One way that the big data phenomenon can have revolutionary impact is to use it as a tool to intercede. It is already a disruptive technology, and rather than using its logic to flatten perspectives, it has potential to open new pathways for inquiry. As with any consideration of seemingly neutral tools, the underlying processes that create and propagate power-knowledge relationships must be interrogated. As Foucault writes, "Knowledge linked to power, not only assumes the authority of 'the truth' but has the power to make itself true. All knowledge, once applied in the real world, has effects, and in that sense at least, 'becomes true'" (Foucault, 1977, pp.27). However, questions of power are not usual conversations in data science or other STEM fields. This begs critical scholars to ask how data operates as power-knowledge, both allowing and constraining certain operations within and outside of STEM. Scholars need to look at who speaks for data, who wields it and for what ends, because data "shapes and is shaped by a contested cultural landscape in both creation and interpretation," with technology as an actor in the process (C. Dalton & Thatcher, 2014). Dalton and Thatcher suggest a two-pronged approach to countering the (mis)use of big data. First, as noted above, researchers need to be critical of data and use thick, qualitative methods that delve into the same research problems as those making claims with big data. The second approach is to fight fire with fire. The authors note that "eschewing 'big data' entirely for its ties to surveillance, capital, and other exploitative power geometries forecloses the possibility of liberatory, revolutionary purposes" (C. M. Dalton et al., 2016). Instead of condemning data, researchers might ask if there are "counter-data" measures to be taken. One such measure might be for critical humanists to work from the inside and become data-driven researchers (Leurs, 2017). Another way to counter data is to "follow

the data scientists'‘ and interrogate the methodological and ontological assumptions made during data collection, scrubbing, and analysis. Just as Simmons and Grabill argued that "writing at and through complex computer interfaces is a required literacy for citizenship in the twenty-first century," (Simmons & Grabill, 2007, pp.441). I argue that a certain level of understanding when it comes to how data is processed, analyzed and wielded is crucial for students of rhetoric right now. Simmons and Grabill note that writing instructors need to put more focus on teaching such critical skills, not so students understand statistics, but so they know "how to make sense of public information from our own subject position as citizens and be able to write using multiple forms of evidence'‘ (pp. 441). Ridolfo and Hart-Davidson echo such concerns about the ways data and computing is a part of good citizenry: "Software is a medium of inscription... Ubiquitous computing now means that understanding software is a crucial part of civic life" (Ridolfo & Hart-Davidson, 2015, pp.22). They argue that as writing scholars and practitioners, we have to understand how the internal processes of computation map onto the outputs in order to engage critically with software and computation. Long before the NSF published its report on the need for data science curricula, Ball, Graban and Sidler wrote about the possibilities of opening up data to a wider range of publics, thus creating *data publics* (Ball, Graban, & Sidler, n.d.). Their goal in this piece was to advocate for a "collaboratory" over a "repository.": "Unlike a repository that might be guided by an ethic of preservation (and in turn, organized according to assumptions about what should make entities stable), a collaboratory promotes an environment in which researchers can leverage the mutability of shared data and communication tools, facilitating networks of research teams and promoting cross-pollination of inquiry, data, and projects" (pp. 2). Their project exemplifies how professional and technical communicators can intercede by participating in the knowledge work of data, rather than negating it.

3. METHODS, SITE SELECTION AND CODING MECHANISMS

> "Ogni blocco di pietra ha una statua dentro di sé ed è compito dello scultore scoprirla."
>
> *Every block of stone has a statue inside it and it is the task of the sculptor to discover it.*
>
> —Michelangelo

3.1 Introduction

Ethnographic methods are used for complex landscapes and non-linear processes that require consideration of interlocking relationships, culture, histories and in this case, scientific goals and habits. Part of the ethnographic practice is to *create* the field of study by building professional and personal relationships with participants and in the workplaces studied. Ethnographers also create the object of their research by setting the problem and drawing boundaries around the questions they pursue. Generally, the boundaries of data collection are not found, and not even purposefully made in many cases, but drawn in the sand in order to work with manageable bits of information. The field that is created through ethnographic methods and texts, however, is sculpted, much like the way Michelangelo sculpted marble, knowing something is there, and uncovering it. The quote above illustrates the common push and pull between practitioner and material. Neither fully produces anything on their own. They work together through possibilities and constraints between the material data and the situated expertise and worldview of the practitioner. It is with this point

of view that I lay out my methods below and highlight when possible how my own positionality is a part of the research practice and the insights that emerged from it.

3.2 Research Questions

My overarching research interest in this study deals with the relationships between data work and knowledge making in the analysis of big data, and how rhetoric and technical communication might inform those relationships. Several questions emerged from these two main thrusts, which I have pursued in this study:

1. How do knowledge workers make sense of large-scale data sets?
    - How does insight emerge in the practice of working with data?
    - What factors influence the epistemological process when working with big data?
2. As a community of practice, how do scientific research groups enact technical communication discourse?
    - What are the discursive moves and terminology of this community?
    - What knowledge infrastructures, technologies and sociotechnical systems are at work?
    - How can rhetoricians and technical communicators inform work with large scale data?

Susan Leigh Star, a feminist sociologist and STS researcher, used to tell her students to look for two things in their fieldwork: first, the specific phrases, metaphors and language used by the communities studied, and second, anything strange, out of place, or any (Star & Strauss, 1999). Though they didn't emerge from Star, these research questions attend to both the language and the strange workings of scientists

whose practice and expertise relies heavily on *big data,* which is both a contested bit of jargon and an emerging area of technical practice that has and will continue to change the way technical communicators create and circulate knowledge.

3.3 Dueling Institutional Review Boards (IRBs)

The IRBs at both the University and laboratory played a significant role in this research. Because the laboratory is a secure site working towards national security and weapons research, the regulations for conducting qualitative work there were very important to understand and adhere to. Though I have been involved in several IRB protocols prior to this point, the review for this project posed several challenges above and beyond the already stringent requirements that Purdue has in place. Initially, I submitted a protocol through my home institution, Purdue University, soon after my position in LANL's ISTI program was confirmed. The IRB representatives stated that I would need approval from the lab's IRB, since I would be collecting data on site. In turn, the lab needed confirmation of an approved IRB from the University before they would approve an IRB for me there. Early in the research process, a lot of time was spent navigating the push and pull between the two IRBs, since neither would approve without first having the other's approval. Finally, Purdue agreed to review the protocol with an attached letter from LANL's IRB that confirmed their expectations and consent for the project. It was nearly a two month negotiation process. However, apart from the limitations on note taking and other forms of documentation, which were considerable (see the *Limitations* section of this chapter), there were very few constraints or special considerations for the handling of methods or data according to the final approved IRB protocol. The most influential constraint was that I would not be able to interview any graduate or undergraduate students,

which hindered my ability to compare different levels of expertise, but only mildly impacted the overall study.

3.4 Researcher Positioning: Inside-Outsider

Sullivan explores how a researcher's "narrative presence'' is "inscribed in the stories we tell''; qualitative work and especially ethnographic methods "[take] on the shading and hues of our own palette'' (P. A. Sullivan, 1996, pp.97). It is important for me to describe my position in the research in this text. Often feminist work does so in the form of identity characteristics—I am a first-generation college graduate and a white woman in her 30s from a low socioeconomic status. I have chosen to go beyond this identity statement to discuss the tensions between:

- Being a qualitative researcher and technical writing scholar in a quantitatively driven site
- Being a scholar and concurrently a paid practitioner
- Being an employee of the lab and also critiquing it
- Noticing power imbalances as a feminist researcher and even as researcher, being more part of a vulnerable population than my participants.

I positioned myself as an embedded researcher over the course of this study. Much like a participant-observer, I was collecting data and critiquing practices, while at the same time working in an official capacity as a staff member at the lab. The distinction between *participant-observer*,"'inside' outside observer'' (Jonas Salk, in his introduction to (Latour & Woolgar, 1986)), and *embedded researcher* is an important one for my purposes. Situating myself as embedded researcher or co-researcher in terms of my work at the site allowed me to act simultaneously as "rhetorician-critic'' and

practitioner, helping me to attune to the community, and at the same time, operate from a critical stance (J. B. Scott, 2003). As participant-observer, on the other hand, the focus of my status at the lab would be one of graduate student intern, rather than a researcher in my own right with my own expertise. As an intern, my job was to produce and be of use. As an intern without computational skills in a highly technical group, I was seen as an outsider, possibly even witless at times, because my skills did not match up within the normal workflow of staff and graduate student labor at the lab. The tension between positioning myself as an expert in rhetoric and technical communication and also a novice in data science and computation allowed me to approach the case study from a novel perspective, having a foot in both worlds. And yet, I wrestled with an internal tension between feminist research methodologies and the need to impress my particular brand of expertise upon the community of researchers I studied and with whom I worked. In my previous qualitative work, I would strive for participatory action research, where I worked *with participants to instigate change*, rather than *objectifying subjects*, which emphasizes the researcher as expert, above and outside the communities and phenomena she studies. Yet, at the lab as an intern among mostly male Ph.Ds. in high-level scientific work, I was always the lay person, inhabiting the lower rung of the power imbalance, even in my own research practice. In line with my feminist epistemology, I was almost subconsciously compelled to subvert my own authority as researcher, but at the same time, as a feminist scholar, I was also compelled to establish my authority and expertise in that space, both as a woman and a qualitative researcher. Certain interview techniques and my research and writing processes also amplified this cognitive tension. First, I have spent a lot of time developing my interview skills and have found that starting from the point of ignorance during an interview—even to the point of playing dumb—usually ended up with the most interesting and in-depth data. This technique also

helped me attempt to peel away some of my preconceived notions at the beginning of a study and instead, work from the grounded data. However, using this technique at the lab meant reifying the idea that I was underqualified to be there—and perhaps most detrimental to my data collection—not worth participants' time. And although in the end, the gendered culture of the lab seemed much less detrimental to women than many other workplaces and social situations I've experienced, I was constantly aware of being a woman among mostly men, in a field and a site that was historically white, male and socioeconomically privileged. In addition, using Grounded Theory and conducting a qualitative inquiry also put my research logic at odds with many of my participants' own practices. I was difficult for me to frame the goals of my research beyond the questions listed above. My goals were not hypothesis driven, and that became a sticking point as I interacted with lab employees. I often encountered questioning attitudes: How could mine be "research'' without a clearer understanding of what the outputs would look like? The deliverables I worked on as an employee were also sometimes nebulous, in that my writing process, while purpose-driven, was exploratory, especially when I would work on a text that was driven by interviews. The insights at the core of this study shifted during the writing process as my text matured, and this was at odds with the common practice of at the lab of writing a conference paper as the final research step, after all of the "real'' knowledge work was complete. The epistemological tensions I discuss above were expected in many ways, but it is important for me to note that I was also continually surprised by how often my research goals were accepted and even encouraged by participants and other staff scientists. Because of my informant/mentor, I was introduced to several researchers who were intrigued by qualitative and rhetorical concepts and saw potential for how such concepts could be implemented more concretely at the lab. From physicists to computer scientists to the more traditional humanities sectors such as archeology and

laboratory archivists, I was regularly offered resources and opportunities to implement my observations and methodologies in scientific practices around the lab. One of the most important insights I gained during my fieldwork related to how technical communicators have to navigate their entrance into high tech industries and disciplines. Here I am speaking particularly about practitioners who primarily identify as technical communicators as opposed to those who would consider themselves technical experts who write and communicate about their work. Some of the core skills that technical communicators practice are the ability to work with various subject matter experts (SMEs), a capacity to learn community discourse patterns and requirements, and to frame and reframe their skills and expertise in ways that make them intelligible and valuable in a given workplace or project space. A large part of earning respect, and therefore gaining access to certain insights and more participants, was understanding lab culture and learning the discourses of data, visualization, computation and physics. Though my interest and inquiry into these topics as they related to my research began long before I entered the lab, LANL as a site was very opaque before I stepped foot on the grounds. My first and most labor-intensive strategy was to read as much as I could in the disciplines (mainly in data visualization) to gain an understanding of the 1) terminology, 2) trends, 3) major issues and 4) connections to rhetoric and technical communication theory and practice. Reading wasn't just a preliminary step, but an ongoing practice during my work. I initiated many informal conversations to ask questions and help me situate the scholarly work into my own understanding. In doing both the discursive research and seeking out informal conversations, I was able to learn a lot about laboratory culture and institutional values along the way. This step also helped me form professional networks across the lab and render my research goals and technical writing expertise in terms scientists from various fields would recognize. On top of that, by building a bibliographic knowledge

of the field of data visualization, I was able to make strong contributions to my working groups by sharing and drafting literature reviews for publications, which in turn bolstered my value and expertise as seen through the eyes of my coworkers. Ethnographic research has a history of fears about "going native,‘ meaning worries that a researcher who enculturates into a group can no longer remain impartial in their analysis. Such fears even show up in Latour and Wolkgar, as they embark on the road to understanding the "alien‘ world of a biology lab (Latour & Woolgar, 1986, pp.44). As outsiders become insiders, they presumably become less able to articulate to other outsiders what is happening in the observations. The subjects' point of view becomes the researcher's, and the culturally internal logic manifests as part of the researcher's newly acquired habitus. And yet, as a technical communication practitioner and scholar, it is not only difficult to keep from "going native,‘ it is ineffectual. We have to both practice and critically reflect on those practices and bring those reflections to our peers in the field and in the workspaces we inhabit. I argue that technical communicators benefit from going native. Because of its rhetorical nature, writing and other forms of technical communication require an understanding of workplace culture, genre convention, stakeholders and the overarching epistemologies of both their collaborators and audiences. Technical writing is, as my participants would say, an applied field, like computer science or statistics. Our goal is to improve and understand writing practice and to usher our research into real world, practical solutions. So rather than fears about going native, the technical writing scholar should concern themselves with the problems that arise when remaining exterior to the work—a fear of staying outside of the discursive and material production of texts, decisions, and enacted knowledge in the workplaces we study.

3.5 Research Methods

This study is based on qualitative work performed over the course of two summers at Los Alamos National Laboratory in 2018 and 2019. The five months at LANL yielded a large amount of data, only a small portion of which was cited directly in this dissertation. The primary data included one on one interviews, group interviews, direct observations of meetings and scientific work, field notes, notes from lectures and research talks, my own research memos and reflections, archival materials from LANL, and a range of texts produced at the lab, including proposals, reports and slides. The body of collected data was supplemented by my own reflections and many informal questions and conversations with staff—most notably, sociologist Dr. Benjamin Sims, who met with me weekly as both a mentor and an informant of sorts, helping me make identify potential participants and make sense of laboratory culture.

3.5.1 Interviews

The first half of data collection was made up of 28 semi-structured interviews and countless hours of observations. The second half consisted of seven formal interviews directed specifically at those whose primary function at the lab is data work in some capacity, from analytics, to visualization, to computation, and domain science. The final data set consists of 35 interviews, which were approximately one hour each.

During the initial interview phase in year one, the interview questions were crafted in six areas:

1. Background on discipline, types of projects and scientific problems they work on, and characterization of the types of data they usually work with.

2. Working in groups and across disciplinary expertise.

3. Workflows and challenges of data work.

4. Data exploration processes and insight generation

5. Working with visualizations

6. Intuition in the scientific process.

In year two, based on my preliminary data analysis and developing understanding of data practice, I revised the interview questions to elicit more pointed feedback on how participants identify challenges in the process of data work, characterize insight, and the specifics of procedures compared to literature on data practice. Most notably, in the year two interviews, I gave the participants an ordered list of steps that were referenced in various ways throughout data visualization literature and asked them if that reflected their own practice, or if there were processes they enacted that didn't conform to the list. Each interview was immediately followed up by writing a research memo that outlined possible themes, statements of interest, and insights I had during the interview. These memos are an established technique in Grounded Theory methods, and I discuss them more in both the *Data Analysis* and *Challenges and Limitations* section of this chapter.

3.5.2 Sampling and Recruitment

I began interviews with a select few participants who were suggested by my various informants because 1) they worked in some aspect of the data workflow, and 2) because they were likely to be willing to take the time to discuss their work with me. I approached them via my LANL email to explain the study, request participation and schedule the interview. After each interview, I followed up with a thank you email and asked them to let me know if there was anyone else they knew who might be willing to contribute to this study. Often, participants would suggest a person or two during the interview with whom I should speak, even without me asking

them. Much of the official work I was asked to do as a LANL employee was closely related to this study, so I was able to make connections with potential participants that way. For instance, my main project as an intern was to investigate workflows and collaborative practices of a particular research team, in order to document ways that future teams might improve. Another deliverable I was tasked with as a staff member included interviewing various researchers inside my own working group and other affiliated sectors, which gave me valuable access to several more participants. Participants came from a variety of disciplines, genders, races, ages and expertise levels, but excluded any student interns on site, per the rules outlined in the IRB protocols.

3.5.3 Informants

Informants are a valuable part of ethnographic methods, because they act as a bridge between a researcher and the community. In my case, my informants and interview participants did not overlap. As noted earlier, my main informant was a seasoned researcher at the lab who was also tasked with officially serving as my internship mentor. In addition to answering questions about culture and data-specific procedures, he connected me with interviews and various social networks, helped me get on several research teams, and communicated my research to others, which allowed me to access a set of people and a pocket of knowledge that greatly aided my understanding. This informant worked with me before and after my fieldwork, and also helped me make sense of specific data cases. My secondary informant primarily connected me socially and professionally across the lab.

3.5.4 Privacy, Security and Technology

Participant privacy and anonymity is an important ethical concern, especially in qualitative discourse-based work with individuals. For this reason, several steps were taken to ensure privacy. First, all participants were voluntary and were given a full description of the research goals and means before consent. Because status, attitudes and other sensitive information that might be revealed during an interview could have very real material impact on a participant in the workplace, and because the lab is a place where work is conducted about national security at differing levels of clearance, it was particularly important for me to offer assurances to my participants beyond what may occur during a other studies. I notified each participant that they could speak off the record at any point, and I would remove my hands from the keyboard and keep all information out of my notes until they explicitly stated we were back on the record. I also allowed participants to read the transcripts at the end of the interview, before I saved them, so that they could strike or elaborate on any quotes or information I noted. This occurred very rarely, but the offer was important for the safety and comfort of the participants. I notified each participant that they had until a certain date, near the end of my time at the lab each summer, to re-read, redact, or change anything in their response by requesting my transcripts. After that date, all data was de-identified by removing all names, job titles, research project descriptions or any other reasonably direct or indirect identifying factors. The lab requires that all data and texts which will be open for public release must first go through a full review. In accordance with this policy, at the end of each summer all of my deidentified notes, including interview transcripts, were officially reviewed for any sensitive information before they were released back to me for analysis or publication. While I was always aware that sensitive information could potentially come forth in my interviews, the risks in reality were modest, because all lab employees are very

staunchly indoctrinated into the privacy and security culture of the lab. The line between what staff scientists can and cannot discuss and with whom is very well demarcated before they even begin a new project. All of the researchers I interviewed worked on projects considered *open science*, that is, research which is meant to be shared with publics beyond the lab. However, like most scientists there, many of my participants also worked on projects "behind the fence,'' which required a higher security clearance and could literally only be accessed on the other side of a 12 foot metal fence that carved an island of high security research out of the middle of the grounds.

At the beginning of my sitework, before I was even able to report to my workstation, I was required to attend trainings and pass tests on information and computer security, physical security, and health or operations-related procedures, including how to interact with the bears that occasionally appeared on campus. Part of the security measures I was trained in had a very large impact on the study, which was that I was not allowed to perform any kind of recording via audio or video on lab property. This meant that all of the "transcripts'' were typed in real time during the interview by me without the aid of recordings to refer to. I discuss the limitations of this in the *Challenges and Limitations* section below.

3.6 Data Analysis

There is a fine balance between method *planned*, method *followed* in a procedural manner—and method *enacted*, where embodied reason, wayfinding, invention, and even intellectual play compose a large part of the sensemaking process in qualitative analysis. During the process, researchers do not *find*, but *construct* the field of study, and this construction is crucial to the kinds of data we encounter and eventually the insights that are derived from it (Bourdieu, 1990). Researcher reflexivity then in

crucial to tracing how knowledge and research perspectives come to be. Powell and Takayoshi (Powell & Takayoshi, 2012) call for researchers in technical communication to be more diligent in tracking their processes, decisions, and techniques, and to include narratives of how researchers work, not as the complete research output, but as part of the work of positioning the methodology and the researcher in relation to the eventual insights.

Writing in its many forms, be it through memos, coding or fieldnotes is part of research and the knowledge construction process. It is "a contemplative act revealing further coherence and fresh patterns. . . a heuristic that guide[s] creativity and intellectual complexity'' Lauer, 2004, pp.86). The following section is an illustration of how I used Grounded Theory techniques to analyze the diverse and unwieldly amounts of qualitative data collected over the course of this study and how the field of study was constructed through my analysis. I outline my coding and memo-ing practices to shine a light on the ways the act of writing structures the final concepts derived from the data.

3.6.1 Grounded Theory

The research methodology begins with Grounded Theory (GT) (Glaser, Strauss, & Strutzel, 1968); (A. L. Strauss, 1987), which is a research practice especially for "generating and testing theory'' using qualitative methods (A. L. Strauss, 1987, pp.xi). Using a Grounded Theory approach means that data analysis is always present during the study, from pre-planning through the final write-up (Glaser et al., 1968), therefore this is not a static phase. GT allowed me to stay close to the data and follow the threads that emerge from my experiences during the fieldwork. It is particularly useful when combined with ethnographic methods in circumstances where the researcher is an outsider, entering a space that is not their own, because it offers

enough flexibility to follow a variety of threads and research questions as they arise during data collection and analysis. Glaser and Strauss outlined nine reasons for writing a practice-based text on Grounded Theory, but for this project, it is their three following reasons that make this method generative for my study:

1. GT works well when the data is very diverse in genre. Since I did not know from the onset which kinds of materials I might have access to over the course of the study beyond my direct interviews and observations, GT was a useful way to bring all materials together for analysis as the need arose.

2. Qualitative, social phenomena are complex and cannot be whittled down into simplicity. GT approaches to analysis and explication allow for dense theory and variation in the phenomena being studied, which is useful when working with complex data sets and complex sets of relations among experts, data and technology.

3. Theory developed without some sort of grounded practical component, whether that be experiential practice or other kinds of direct data, is ineffective (A. L. Strauss, 1987, pp.1-2). This is especially true in the realm of technical communication, which holds rhetorical efficacy and writing as action as core values.

3.6.2 Research Memos

A Grounded Theory approach requires the development of several concepts and their relationships in order to capture variation (A. L. Strauss, 1987). Phenomena are complex, and they require complex renderings of inquiry. Memos become a key analytical technique that allows complexity to develop while also identifying themes and concepts in the data. Alongside data collection and coding, memo-ing completes "the triad of analytic operation" (A. L. Strauss, 1987, pp.18). Based on my prelimi-

nary theoretical sampling, memos were often focused on how teams reify knowledge through discourse and technical work; data exploration and the differences between data-driven processes and expertise-driven analysis; and the ways scientists deal with outliers and anomalies in their data sets. Some memos took the form of literature reviews and notes about concepts I had encountered during the extensive reading I completed in the discipline (see *Researcher Positioning* in this chapter). However, many memos were also exploratory and open; they were used to highlight normal, everyday practices that reflected lab or scientific culture; key terms and jargon in the field that were used in discussion; early questions, assumptions, and hypotheses; possible codes; and any instances that were puzzling. General memos were written approximately five times a week, normally one for each day I was on site. However, separate memos were often written for specific meetings, lectures or other informal discussions that I took part in which stood out against my everyday practices in the group. Interviews also inspired their own separate memos written directly after the session was over or as soon as it was possible. Near the end of each summer, as programs began to wind down, I regularly reread and sifted through existing memos. I also wrote new memos that reflected on early notes, augmented them with new cases or information, or made connections across memos. All in all, the memos helped me address early questions and later contextualize those questions as my understanding and observations evolved. As Strauss notes, both the memo writing itself and research texts that evolved from memos felt "both analytic and creative'' to me, because they offered grounded data from the site and also and infrastructure from which I could work on making connections across a group of files (A. L. Strauss, 1987). Making connections in the early phases of memo writing felt like creative work. Grounded Theory memos are part data, part analysis, part exploration. Because they are both bound and separate from observational and interview notes, they became a liminal

space for me between *what is* and *what if*, where I was able to play. Having access to early memos and revisiting them also helped me re-enter that *de nova* positioning, when my understanding of the site and its attending phenomena were alien and fresh. While I have argued that it is not always useful for technical communication scholars to stay in the *de nova* phase, it was productive for my analysis to combine early questions with later insights; it helped me reflect on my own practice of entering a field and a discourse, which is part of the goals of this study—to find out what technical communication practitioners need to know to enter data-heavy industries, and to discover the ways such fields and industries would benefit from TC practitioners.

3.6.3 Coding Structures

In Edwards' *Vast Machine*, he notes in his discussion of how data becomes knowledge that "an established fact is one that is supported by an infrastructure'' (Edwards, 2010, pp.22). In Grounded Theory and other kinds of qualitative analyses, part of the infrastructure that becomes insight or theory is the coding process. Often the details of coding processes are relegated to appendices or left out of final research texts altogether. However, I find it necessary here to briefly address the codes I began with and how those codes translated into concepts, in order to make the infrastructures of my research visible. During my first two phases of open coding, I created 48 distinct codes. Then, I combined and revised that initial set of codes during an axial coding session, which narrowed the scheme down to 19 codes, two of which were based on cases I wanted to sort rather than phenomena, such as "in situ'' and "data sprint.'' This is also called the Integration Phase, which helps codes evolve from "mere collections of incidents into theoretical constructs'' and helps ground any abstractions in data itself (Lindlof & Taylor, 2002, pp.222). With this in mind, a coding paradigm (A. L. Strauss, 1987) was developed using NVivo software, which allowed me to link

each code to the others in its vicinity. NVivo then creates a visual map, allowing the researcher to see the connections among codes as an in situ, discursive geography, rather than only a group of topics connected by definition or the researcher's whim. The final coding structure consists of 17 concepts from which I began to derive the remainder of my analysis.

3.7 Challenges and Limitations

Being the only rhetorician in a community of scientists posed challenges beyond only methodological and epistemological difference. Though differences in disciplinarity and methodology were tricky to navigate. I had to first come to a basic understanding of what any given person (coworker or potential participant) worked on, then find many different ways to frame the field of rhetoric and technical communication in terms that made sense. By *terms*, I mean both the actual terms and jargon of their worlds and the framework I used to situate my research and expertise into their disciplinary understanding. I also had to develop several different working descriptions of my study in order to help scientists see the potential of the research. Otherwise, it was difficult to get them commit to the time and access I needed for interviews and observations. While such continual reframing was a useful rhetorical practice, and certainly helped my research evolve, it was a constant exercise that I did not fully foresee when I began my fieldwork. This study was particularly complex because of the high-security environment of the lab. I regularly encountered unforeseen barriers to accessing people, research, and information that I hoped to attain. Some information which may have aided my analysis was simply inaccessible. Most notably, the security culture meant that I could not use my own laptop and could not use video or audio recordings, even for interviews. Having transcripts or other forms of detailed recordings is necessary in qualitative work because they allow

Codes	No. of Interviews Appeared in	No. of References Total	Codes	No. of Interviews Appeared in	No. of Referenc Total
Argument	11	14	Insight	15	36
Articulation Of Epistemology	14	23	Intuition	6	20
Articulation Work	10	13	Invention	11	18
Bias	9	18	Knowledge Is Always Carrying Over	2	4
Communicating Data To Others	9	18	LDRD	6	14
Communication Gap	17	28	Machine Learning, Deep Learning & AI	3	15
Community And Personal Networks	12	27	Mentor	1	1
Context Of Data	13	32	Missing Or Unused Data	6	12
Creativity	4	9	Modeling	9	24
Data Driven (Or Not)	10	22	Narrative	7	8
Data Friction	7	13	Noise & Signal	4	7
Data Scrubbing	5	10	Outliers	8	18
Data Sprint	1	4	Parataxis	7	18
Definitions	11	26	Proxies	1	2
Difference Between Soc Science And Science	2	7	Romanticism	4	4
Expertise	23	64	Situated Knowledge	10	22
Exploration	16	49	Sociotechnical Systems	12	23
Familiarity With Data	3	3	Teamwork	18	39
Human	10	22	\emph{techne}	7	11
Improvisation	2	3	Tuning Or Attuning	6	9
In Situ	2	3	Usability	16	29
Innovation	6	7	Visualization	15	33
			Wayfinding	8	13
			Workflow	24	60

Figure 3.1. Open Codes: Initial Coding

Epistemology	**Situated Knowledge**	**Context Of Data**
Articulation Of Epistemology	Situated Knowledge	Context
Workflow	Knowledge Is Always Carrying Over	Data Friction
Data Driven (Or Not)		Data Scrubbing
Bias		Missing Or Unused Data
Proxies		Outliers
		Noise & Signal
Insight	**\emph{techne}**	**Expertise**
Insight	Parataxis	Expertise
Exploration	Wayfinding	Familiarity With Data
Creativity	Tuning Or Attuning	
Improvisation	\emph{techne}	
Innovation	Articulation Work	
Intuition		
Invention		
Human	**Sociotechnical Systems**	**Teamwork**
		Teamwork
		Mentoring
Community & Personal Networks	**Modeling**	**Argument**
	Modeling	Argument
	Machine Learning, Deep Learning & AI	Communicating Data To Others
Communication Gaps	**Usability**	**Visualization**
Narrative	**Romanticism**	**In Situ**
	Data Sprint	

Figure 3.2. Axial Codes: Final Categories

a researcher to be present during the interview and refer back to recordings as the analysis develops. A primary limitation was that I had to be as present as possible during interviews while also typing as fast as possible to capture as much detail and language as I could. This definitely changed my process. In other instances, part of my interviewing *praxis* was to take note of topics of interest that might lead to further

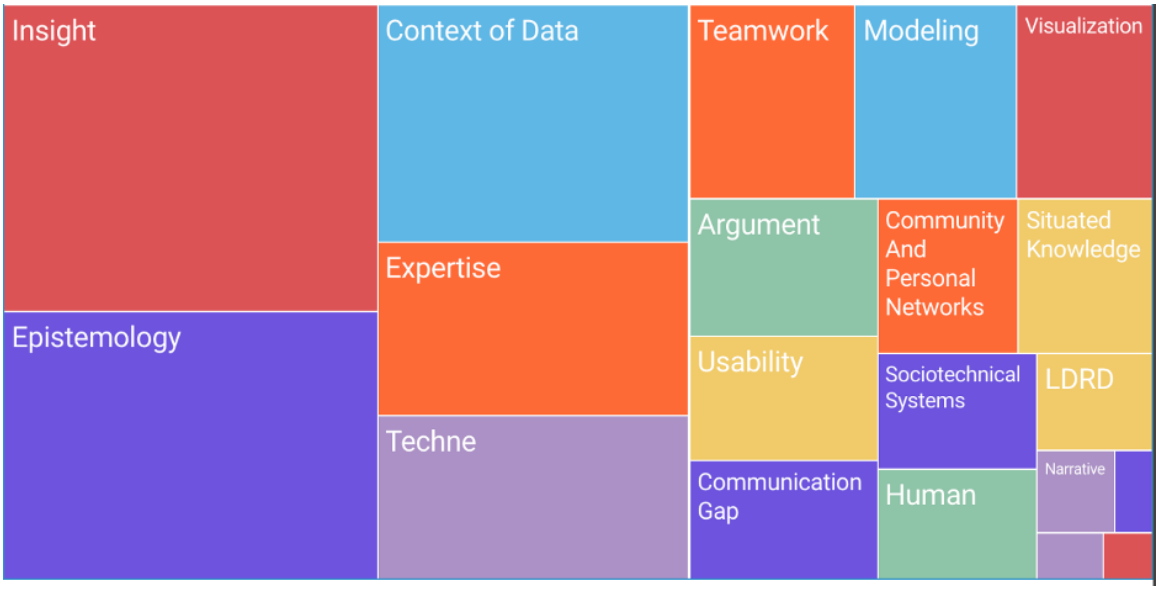

Figure 3.3. Axial Code Density Across All Interviews

inquiry. Without a transcript to fall back on, I noticed that my insights during and immediately after the interview were fewer than in other studies I have performed. Additionally, typing the transcripts live led to regular typos which only occasionally clouded meaning, but it also truncated the amount of data I could quote directly. Due to the schedules and workload of potential participants, I was unable to interview many people that would have likely contributed a great deal to the study, including students, which was a restriction laid out in the IRB protocol. Additionally, follow-up interviews were difficult to obtain for the same reason. To mitigate this limitation, I relied heavily on my informant to clarify a range of questions throughout the field work, rather than going back directly to the source. It was often a methodological challenge to read the research spaces at the lab during regular observations. By read, I mean gain a grasp of the micro-relations and tacit knowledge circulating in a space. Physical places and constructed workspaces "frequently carry social messages or require social behaviors and learning what these are is a central part of ethnographic

work. As such conventions are often tacit interviewing does not necessarily produce coherent accounts of them'‘ (Atkinson, Delamont, & Housley, 2008, pp.153). The issues of intellectual property, lab security, and the constraints on recording posed a challenge in reading research spaces, but so did the complexity of the epistemological processes I attempted to explore. Participants' intellectual, creative and disciplinary data habits were difficult to access in situ. I eventually discovered that the kinds of on-the-ground data work I wanted to explore was often not completed by the scientists, but by their graduate student interns. Because the IRB prevented me from working with students, it was not possible to intimately observe and track such practice. Another major challenge arose when I lost access to a large collection of memos, reading notes and a draft after my final phase of fieldwork. Because of information security procedures, any files moved from lab property, such as my laptop, to outside equipment has to pass through a portal and downloaded securely on the other side. I was unaware that files had a short window to be downloaded and therefore was unable to access much of the files from my LANL computer after I returned home. While some colleagues attempted to recover these documents on my behalf, the strict security protocols meant that all data was already wiped from my equipment and the other information systems that transferred these files. For some of these limitations, I created procedures and honed research techniques and social skills to mitigate them. I found valuable and reliable informants, worked ceaselessly to gain a baseline understanding of lab culture, learned to type very fast, and diligently used memos to think through data directly after each interview or meeting. However, the problem of lost and inaccessible data cannot fully be mitigated in the end.

4. CHAPTER 4. DATA SETTINGS OVER DATA SETS

4.1 The Cases

Though big data, especially in the corporate and public sectors, is thought of as a comprehensive tool that is able to uncover facts without any attending contextual information, "ways of inscribing data are always constrained by the local'' (Loukissas, 2019, pp.67). In The title of this chapter comes from Yanni Loukissas' book, *All Data are Local*, which argues that the local conditions of data collection, scrubbing and analysis are just as complex and important for how insights emerge as the data sets themselves. While "local'' can refer to geography or proximity, I want to define and address "local'' here as the ways knowledge infrastructures, levels of expertise and interdisciplinary backgrounds impact how meaning is inscribed with data. In this chapter, I outline three cases of data work that focus on differing levels of expertise in order highlight a few ways data is rhetorically constructed and situated in local practices rather than rote procedures.

4.2 Case 1: Data Workflows

J. Robert Oppenheimer was one of the world's leading theoretical physicists during the second World War, and in 1942, he was appointed the science director for Project Y, which evolved into what we now know as the Manhattan Project. Researchers of Project Y advocated for a top-secret facility where they could work in isolation, away from suspicion and security leaks. This lab was to function externally as a highly restricted space, but internally, the goal was to create an intellectual safe

haven where scientists could freely work together across disciplinary lines to produce the first nuclear weapons. A small, isolated mountain town composed almost solely of lab employees grew up around the laboratory and became known as "The Secret City.'' Los Alamos National Labs (LANL) in Los Alamos, New Mexico is still the home of a United States' nuclear testing facility, but today, tests are conducted through computational simulations instead of detonations. The lab is funded primarily by the National Nuclear Security Administration (NNSA), an agency within the Department of Energy (DOE) and is one of only three NNSA laboratories in the whole country.[1] LANL is a mission-driven laboratory, meaning all research conducted has to benefit issues of national security. While the weapons division is at the core of the research output, LANL researchers study a sizable range of phenomenon, including climate change and epidemiology. The lab houses many scientists who do not directly work on weapons research, though it's understood that any and all funded work will eventually be deployed in service of national security at some stage in the pipeline.

My interest in this site was piqued because of the lab's emphasis on scientific big data and data at scale. Since 1993, a group of computer scientists have regularly published what is known as The Top 500 List (*The Top 500 List*, 2020), which is an industry standard ranking of the world's 500 most powerful supercomputers. As of November 2019, LANL's supercomputer, *Trinity*[2] was ranked 7th in the world, being composed of almost one million cores. With this much compute power, LANL scientists are able to run larger data sets than almost anywhere in the world. Because of the diverse origins of big data, the potential material and social impact of its

[1]The other two labs in the "tri-lab system'' are Sandia National Laboratory in Albuquerque, NM and Lawrence Livermore National Lab in Livermore, CA. Research groups often form across these three sites, and at the same time, a healthy rivalry and competitive streak exists among researchers in the three labs.

[2]The other two labs in the "tri-lab system'' are Sandia National Laboratory in Albuquerque, NM and Lawrence Livermore National Lab in Livermore, CA. Research groups often form across these three sites, and at the same time, a healthy rivalry and competitive streak exists among researchers in the three labs

meanings, and the necessity of complex research questions across domains, scientific big data inquiry requires a diverse interdisciplinary approach. Most research teams working with large scale data include a wide range of domain scientists and computational experts, along with other technological staff. Employing more than 1,200 staff from 45 states and several countries internationally, LANL is no exception. I was hired in to LANL through their Information Science and Technology Institute (ISTI) as a graduate student intern. This program offers graduate students the chance to work on state-of-the-art projects with staff scientists for the summer. Because of the nature of the summer institute and its emphasis on student mentorship and development, I was able to negotiate time for my own doctoral research while also working full time in a team of scientists who design applications for new visualization methods using big data. During my time at LANL, I conducted interviews with a number of scientists connected to this division, and many more outside of it. Most of my participants were not connected to each other or to the visualization team by a specific project or working group. However, I did get an opportunity to delve quite deeply into a couple projects, where I was able to interview several people on the same team to get a sense of how big data-centric research functions as a collaborative exercise. The lab is split into divisions, and these divisions are usually bound by a certain field, such as molecular biology or data visualization. In any given division, staff tend to work closely in the same building or area of the lab. At the same time, all staff scientists are involved in several different projects, and they divide their time among them. At any given time, a division of 10 people could be dividing their work among 50 different problems, projects or working groups, which are often interdisciplinary. Because LANL is a mission-driven and primarily government-funded lab, most of the work is very goal-driven and project based. Research for the sake of research is not common. Funding also rarely allows staff to work in isolation. Big problems require

collaboration. Therefore, any discussion of how data is wrangled or how its meaning is constructed at the lab must include an awareness of how scientists collaborate. The following is a description of one such working group that illustrates some of the knowledge infrastructures that undergird the collaborative design of novel data workflows. When the average person imagines big data, they likely think about huge amounts of numeric information in tables and databases that is as a whole, incomprehensible for a human brain, at least without the help of basic data visualizations. As discussed in Chapter 2, incomprehensibility is often how data becomes defined as big. But what if the data you are working with is not a set of numbers and calculations, but instead, a set of images? Even with access to the most powerful supercomputers in the world, massive scientific data sets can take months and hundreds of iterative runs between data input and visual output. This is all the more true for image-based data, because images are far more computationally taxing to process. For many experimental scientists, the most problematic issue is not waiting months for their data to run, as much as waiting for their data, only to find that there were malfunctions with the equipment or methods that rendered the data unusable (Vogel et al., 2018). When breakdowns occur at this scale, especially for researchers who collect data in remote environments or for those whose experiments require considerable preparation time, unusable data outputs can set research back months or even years. The interdisciplinary team at hand vacillated between 10 and 25 active members all working on a very well-defined problem from the onset of the project: Can a new set of computational workflows augment the precision and speed up the timelines of data analysis? When the initial problem was set, the steps of the experiment and subsequent data analysis proceeded like this:

- Researchers would schedule a specialized lab for the experiment. Being a specialized laboratory, the equipment might only be free two days out of the year.

- The scientist planned the experiment. Because time in the facility was competitive and due to the massive cost of each experiment (in the hundreds of thousands of dollars), research plans were designed and revised meticulously for several months or a year prior to each run.

- When the time came, experiments lasted 36 hours nonstop, and the attending research team spent all 36 of those hours on the lab floor. Though this is a gross oversimplification of the experimental process, researchers would essentially shoot a beam of light at a material from different angles. The researchers took a single shot, collect measurements, recalibrate, and shoot again, in a constant loop. Because the data created were images, and because of the sheer amount of shots taken, the researchers did not have a feedback loop to make sure the shots they took were on target, which is why so much time and focus went into pre-planning.

- Once the experiment was over, the data took about two months to process, running continuously on one of the largest supercomputers in the world.

- At this point, hundreds of individual images were rendered for analysis. Because there was as yet no interface designed to work with such data, researchers arranged all the images in a single PowerPoint file. A slide deck was the primary way scientists analyzed their data.

- If after all that, the equipment wasn't calibrated correctly or the shots were off in any way, the researchers would spend months hundreds of thousands of dollars to repeat this process.

Based on the requirement of time, money and compute power, the goal of the team was to completely change the way data was handled by developing an in situ visualization workflow and attending interface, whereby scientists could see their data visualized

in real time during experimentation. This meant that for the first time, researchers would have the ability to intervene in the experiment. If the data was unclear, or instruments needed reset, researchers would be able to monitor the environment and make changes while they were still conducting the experiment, rather repeating the work 6-12 months down the road. Such a project would stretch over two to three years and required a very interdisciplinary team that included experimental scientists, visualization experts, computational physicists, statisticians and additional staff scientists with a range of other expertise. Generally each area of expertise was represented by only one to two people on the team, and there was not a lot of overlap in expertise as far as the project was concerned. In order to 1) get everyone up to speed on the project's goals, and 2) expose everyone to each team member and their area of specialization, the group spent six months at the onset of the project giving presentations to each other. These presentations covered specific cruxes of the problem, overviews of a certain field, discussions of possible solutions or previous work, as well as a lot of time discussing terms and definitions. Team members took turns educating the others on each of their own expertise and disciplinary perspectives as it related to the task at hand. This ramp-up phase was followed by a long span of weekly all-hands meetings and small working groups. New members were added as needs and new talent emerged, ushered in by team members as often as they were brought in by the team's leadership. Often, a small group of team members would create offshoot projects. Sometimes these were unrelated to the original data altogether. Though the meetings were large, they tended to be the place where the "real work'' of the project was accomplished, discursively. As opposed to the following two cases, this group was very structured in the sense that their plan of attack was detailed and problem-driven. In fact, the proposal and attending project management document that was originally written to apply for funding was a key piece of the group's knowledge infrastructure.

The proposal a primary way newcomers familiarized themselves with the project. The guiding document listed team milestones, but interestingly, it did not drill milestones down further. This was done purposefully in order to block out time for individuals to investigate possible interests and curiosities that emerged during the project. It was brought up at least once in each meeting I attended by several team members, not only by the leadership team. The proposal was so engrained in the team's work that by a certain point, they no longer had to refer to the text to know where the project was supposed to be on any given week.

4.3 Case 2: The Analysts

The data sprint, as its name suggests, was a short, intensive collaboration meant to turn data into insight over the course of four days. This gathering was made up of people with a wide range of skills and expertise, and collaborators were rapidly guided through structured invention and discovery phases. The sprint included rounds of ideation, heuristic development, discussions of how data parameters are defined, and the creation of personas. The data sprint structure is based on fast-paced ideation processes such as Agile software development and design thinking methods. Design thinking is a process of ideation and creation with the goal of bringing a human-centered focus to developing innovative designs and solving wicked problems.[3]. There are five iterative stages in the design thinking model: empathize, define, ideate, prototype, and test. While this method of ideation and production is often attributed to IDEO and the d.school at Stanford University (Pope-Ruark, Tham, Moses, & Conner, 2019), its foundations have been traced all the way back to John Dewey (Buchanan, 1992). Design sprint methods are commonly used during collaborative work, but

[3] *Wicked problems* are issues that are so ill-defined, complex and pervasive that actors need an alternative to linear, procedural thinking to tackle them (Buchanan, 1992)

data sprints are a relatively new concept whose strategies and methods are still being honed (Larson & Chang, 2016).

In this particular illustrative case, the data sprint structure became a site of contention as analysts were forced out of their typical workflows. The differences between sprint procedures and everyday data workflows that were highlighted by interviews allowed me to take a closer look at the epistemological and invention practices of data analysis. The sprint began with an overview of the research study and attending data that would serve as the focus of the next four days. The study aimed to identify and categorize group formation in the workplace and explore how emotions such as stress may play a role in professional practices. The original research study took place over several sites with thousands of total participants. Data included qualitative surveys, interviews and quantitative data from sensors that monitored a range of habits, not limited to phone use, geolocation, work schedules, interactions with others, and exercise and biorhythmic data, such as sleep patterns. In addition, a subset of sensors placed in participants' homes, public spaces and on participants' person collected voice and ambient sound. [4] Rather than analyzing the discourse in these recordings, researchers were interested in variables such as who participants spoke with, intonation, volume, the "energy'‘ or "intensity'‘ of speech, and the length of speaking time. All in all, there were millions of hours of sensor data, amounting to possibly billions of interlocking data points.

The goal of data sprint was exploratory in the sense that researchers began with a set of questions and points of interest as they collected the data, but did not have a clear hypothesis they wanted to test. Research questions of their study centered around group identification in the workplace:

[4]I am compelled to note here that all of their participants were aware of these study mechanisms, and while there may be ethical issues with ambient recording, such data collection is only as invasive as the average person's smartphone.

- What groups do participants belong to, and are participants aware they are part of these groups? Are the groups that reveal themselves through the data explicit or informal?
- Do these groups or physical and social networks influence behavior?
- Can groups be tracked over time as they evolve, emerge, disintegrate? What methods might we use to identify that in the data?

From these questions, the researchers hoped to identify trends and clusters of data points that might lend insight into these questions. With the mix of computational, statistical and social scientists involved in the sprint, the P.I.s hoped collaborators would isolate data clusters, propose methods to verify any model that would be created, and identify additional data that could be collected as the study moved forward. Data sprints take many different forms. According to a 2019 report on data sprint methods, sprints include intensive research and coding workshops where participants collaborate across disciplines and industries to work with a set of research questions through a data set (Laursen, 2017). Data sprints generally require an initial time period where organizers or subject matter experts orient the other participants to the problem at hand (Munk, Meunier, & Venturini, 2019). Laursen's informal study interviewed researchers across various design labs in Copenhagen who employ data sprint techniques. Their definitions focus on interdisciplinary expertise and process:

- "An intense period of working with data where you clean, explore, analyse, and visualise data. You could also add that you tell stories with data and that it is an iterative process.'‘
- "A data sprint is people with the right competencies to do a data project who meet for a week to do things first and think about them later.'‘

- "It is that you can work interdisciplinary and that you can learn something new; both technical but also about a subject that is not necessarily technical.''
- "It is about creating an open space and a partial structure for a process of development, where you put some people within a framework an unleash the process and then see what happens.''

According to Laursen, data sprints often follow a set of phases from sprint design to publication.

1. Pre-phase: Sprint methods are designed and a focus and/or data set is decided upon. Sometimes specific participants are chosen at this point.
2. Introductory Presentation Phase: Participants listen to presentations that illuminate the subject matter and the data in detail, based on expert knowledge.
3. The Sprint: This is the ideation phase, where facilitators walk participants through heuristic exercises, creating research questions or hypotheses, and other inventive group work that aims to break participants out of their normal workflows and habits.
4. Final Presentation Phase: Participants present their methods, findings or future research possibilities to stakeholders and each other.
5. Post-Phase: Write ups and publications are drafted and plans for future work are made that extend the modest outcomes of the sprint. (Laursen, 2017)

The Introductory Presentation Phase might look like one long lecture or several short presentations made by stakeholders or other experts. For example, presenters might outline the overall goals and history of the project. Others may discuss sampling and data collection techniques. Depending on the maturity of the research, presenters might also discuss any insights derived from preliminary analysis. The Sprint

Phase often consists of forming small interdisciplinary groups to brainstorm interests, explore the data, and create visualizations or data narratives (Laursen, 2017). The interests they focus on might concern:

- What collaborators want to know.
- How fields from the data set match their curiosities (or don't).
- If any additional context or data needs to be understood as they work with data.

This phase is meant to result in a range of interests and concerns. Like a design sprint, the invention phase of a data sprint is meant to allow loose brainstorming. Outputs of the invention part of the Sprint Phase can be purely data-driven, where analysts ask what is in the data and run statistical processes, or analysts' interests during this stage can be situated concerns about the context of the data, such as how it was collected, who it represents, or other questions about how the analysis impacts various stakeholders. Analysts may become frustrated with sprint methodologies, because they are designed specifically to break professionals out of their rote habits and workflows. Data sprints are meant to be exploratory whereby the insights one ends up with are not what was expected at the onset (Laursen, 2017). The tension between the structured sprint frameworks and the data habits developed through practice is meant to be a productive tension. In this case, collaborators' normal data habits had been subverted by the sprint methods, and some collaborators reverted back to their habitual and procedural methods of their normal workflow to analyze data. Practices included basic statistical methods, clustering and 2D visual plots. On the final day, the Final Presentation Phase begins, where collaborators present the insights they derived from the sprint, mainly focusing on the hands-on data work that occurred. Such deliverables are often humble in comparison to what analysts hope

for, but are intended to be preliminary starting points, not full, thick understandings of the entire issue at hand.

4.4 Case 3: The Novices

Tucked back in a quiet wing of an expansive hall, I sat in a room with 30 other people, all of us learning how to work with big data and computing clusters [5] for the first time, save the handful of instructors who designed this weeklong workshop. Its goal was twofold: 1) to expose underrepresented, early career professionals to analytic and cluster computing methods; and 2) to introduce computational analytics through pressing social issues that help attendees understand how data science and computing can be used to impact social inequalities. The group of attendees was purposefully selected to include people from both a diverse range of genders, races and ethnicities, as well a wide range of disciplinary backgrounds outside of computer science. Most computing fields have not been a welcoming place for women and domestic people of color. The diverse makeup of this workshop's attendees aimed to address the longstanding issue of homogeneity in the field of computer science, which is profoundly white and male in the United States. [6] The workshop asked attendees to focus on issues of maternal and infant mortality using a specific, curated big data set. Mortality is a complex issue, which is deeply intertwined with structural racism, classism, and sexism. Long historical traces of such inequalities continue to impact current policies and medical practices. Additionally, maternal and infant mortality

[5] *Computing clusters* or *cluster computing* refers to a computing architecture whereby several individual machines are linked together. They function logically as a single unit with much higher power and speed capabilities than a single computer.

[6] Only about 20%of computer programmers are women, 67% are white, % are Hispanic or Latinx, and another 8% are black. And in data science programs of study, demographics hover around 46% white, 35% women, 8% Hispanic or Latinx, 4% black in data science programs. Noting that these fields often pay well, and that the market for computer and information research experts is predicted to grow faster than average—16% in the next 10 years—it is particularly important to note the gender and racial makeup of a high-paid, fast-growing profession (of Statistics, 2015).

rates are cloaked in a long history of racism in the United States. Well known cases such as the Tuskegee Study and Henrietta Lacks' HeLa cells have come to public attention in the last 20-30 years. Many lesser known cases exists as well, such as Lucy (no family name documented), who was unwillingly put through gynecological surgeries without anesthesia and stories of white "night doctors'' in the 1800s, who would kidnap black people and corpses to use in experiments in Boston hospitals (Gamble, 1997). Though Tuskegee is often pointed to as the canonical case of medical racism, it is merely a single instance among many that demonstrates the ways in which people of color have systematically been terrorized by medical professionals and the healthcare industry. As such, people of color are less likely to seek medical care (Alsan & Wanamaker, 2018), less likely to trust medical professionals (Armstrong, Ravenell, McMurphy, & Putt, 2007), and more likely to receive lower quality of care than their white counterparts, even today (Nelson, 2002). Because of the ways race, gender, class and historical traces of medical terrorism are deeply intertwined in the United States, data about maternal mortality can only be insightful if it is intersectional and also put into conversation with the material realities and social injustices perpetrated against women—and especially women of color—which cannot always be quantified. The workshop purposefully sought to make issues of prejudice and injustice visible. Facilitators made a point to discuss how computational data analysis can be a tool to expose and address structural inequalities. By having to work with an intersectional and socio-materially complex topic such as maternal mortality, attendees were learning to see that issues of injustice could not be disentangled from data analysis. They were asked to confront their own inherent biases and the complex structural inequalities with each decision they made during data analysis. As opposed to the previous cases in this chapter, this workshop was not meant to operationalize the data or produce a specific deliverable. Quite the opposite—the main goal of

the workshop was to walk attendees through the steps of working with big data by guiding through exploratory exercises. As noted above, almost none of the attendees were familiar with computational analysis, and workshop facilitators encouraged them to "fail early and often.'' Attendees were expected to productively stumble as they learned to work with the data and tools introduced at the onset of the workshop. In fact, the facilitators opened the workshop with the following Theodore Roosevelt quote from "The Man in the Arena,'' a speech he gave in Paris railing against those who criticize and critique those who act and work to make a better world.

> It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly.
>
> — Theodore Roosevelt, "The Man in the Arena,'' April 23, 1910 in Paris France

The quote highlights some of the social and affective differences between this and the previous two cases. This workshop was designed to teach rather than produce, and attendees were invited specifically because they were data novices rather than experts. As novices, both the technologies and methods introduced during the workshop were foreign.

4.4.1 Hackathon

This workshop was marketed as a community hackathon. The term "hackathon'' showed up about 20 years ago, and is usually used to describe software or application development events with intensive periods of collaboration and a working prototype

at the end, much like a design sprint (Benham, 2018). However, a hackathon can be any event organized around a specific project and goal, where multiple people work together for a finite amount of time, often 24 hours stints or days in a row. This particular workshop lasted five days. Attendees were unaware of the week's focal topic prior to the start of the workshop. The facilitator announced on Day 1 by explaining the goals and research questions of the study that produced the data. Researchers were collecting past data in hopes that they could make predictions about an individual woman's health future health risks. In a sense, the researchers were using the data set to determine probabilities of maternal complications so they could put patients in "bins'' of probability based on their identity characteristics, demographics, geographic location and medical practices, such as how early they began seeing a doctor. The exigence of this research stems from the increase in maternal death compared to previous decades, while birth rates are going down. Although the United States as a whole ranks 46th in maternal health, certain states such as Louisiana have higher maternal death rates than half the countries in the world (Agency, 2020).

The introduction to the workshop continued like this on the first afternoon. Facilitators not only listed general morbidity statistics, they brought up specific insights gleaned from the data and put them into conversation with more data to craft a story or argument. For example, if 52% of birthing costs in a particular state are paid by Medicaid, such information may lead us to consider more research questions, but on its own doesn't create a story that helps us take action on maternal mortality. If we add in that 67% of maternal deaths in that same state occur 30 weeks after the birth of a child, and that Medicaid stops providing benefits at 26 weeks, then we can begin to put together a more nuanced question with which to begin an inquiry into mortality. Along with an overview of the study and composition of the data available,

facilitators spent considerable time defining terms. At first these definitions seemed to be merely more background, but I soon realized that most of the definitions discussed were important for understanding which data should be chosen over other similarly labeled datapoints. Having a clear understanding of the boundaries of each parameter is crucial for attendees' analysis. For examples of these definitional issues, see the discussion in Chapter 5. Along with definitions came with an introduction to the data dictionary, which is a very important component of any data set. A data dictionary is a file that lists several key factors that are key for sensemaking:

- A complete list of the data set's parameters,
- The type of data that is valid in each field (integers, text, percentages etc.),
- The possible valid ranges of each field, and
- A detailed descriptions of how labels and terms are defined.

In this case, the data dictionary noted that maternal mortality includes deaths for up to a year after birth. In the previous case of the data sprint, the dictionary defined "loud conversation'' as any speech recorded by sensors that lasted more than five seconds and occurred at 75 decibels or above. You can see then why the data dictionary is key to detailed and accurate analysis. Problems can arise if data workers do not thoroughly understand data fields or the significance of definitions. Additionally, data dictionaries can often be incomplete or haphazardly put together by those who compose the data set, for the same reason that documentation is often nonexistent in code. It takes a lot of extra time to document parameters, and if a professional works in a single data set over long periods of time, the terms and fields become tacit information that isn't made explicit for future data set users. The next topic discussed in the workshop would turn out to be possibly the most fascinating aspect of analysis: proxies. A proxy was described as a substitute for ground truth

data or other data that are not present in the data set at hand. If, for instance, one needed to understand the levels of access to emergency obstetric care in a geographic location, but did not have such data, they could find one or often a group of data they do have which can reasonably stand in for access to care. Caesarean section rates have been sed as a proxy, because credible studies show that 5-15% of births need them as an emergency procedure. Therefore if Cesarean sections in an location are below the standard range, one can assume that access to emergency obstetrics is lower than average (Stanton, Abderrahim, & Hill, 1997). Finally, at the end of the first day, the facilitators emphasized that there would be a long process of forming research questions, choosing data, and data scrubbing, before attendees would even begin to run computational analytics. Facilitators noted that medical data is known to be particularly messy because of the nature of documentation and the disparate ways data is formatted across the sites that produce medical data. In this particular instance, data scrubbing might also entail accounting for null data or reformatting various fields similarly so they could be compared. According to data science literature and the interviews I have conducted throughout this study, data scrubbing is the most time-consuming part of the analysis process. However, it is also the most invisible. At the beginning of the second day, the walls of the room were plastered with 14 posters, one for each of the steps attendees would take on the way to presenting their work at the end of the week. The steps were:

1. Identify your question.
2. Is the answer informed by the data we have?
3. What data do you need to tackle the question?
4. Does that data exist? If not, what proxies will you use?
5. Does your data need cleaning?

6. Ask your questions by analyzing your data.
7. Are you using the right tools to do your analysis?
8. Have you looked at your data? If no, why not?
9. Are you surprised by your answers?
10. Do you need more data? If yes, return to Step 3.
11. What are your findings? What is your story?
12. Put your story together visually.
13. Do you have any proposed solutions?
14. Tell your story: Presentations.

After a quick explanation of the process, attendees spent the rest of the day learning to use python and access the computing clusters they would use to process their data later in the week. Attendees then began forming research questions and were asked to skim the data set. However, it quickly became clear that as novice analysts and newcomers to the topic, attendees did not spend time with the data before forming their questions of interest. Instead, attendees' curiosity from the introduction to the issue of maternal mortality seemed to spawn their inquiry. More than half of the questions attendees first announced to the group were closely tied to the socio-material conditions of parent and child, and focused on thick, entangled problems dealing with mortality... but were not grounded in the data set. Although these research questions demonstrated dedication to human-centered inquiry, they were not questions that could be answered with the data set at hand. Attendees then spent several hours workshopping new questions and reworking their original questions based on what could be explored in the data set. On the following day,

several complications began to arise in the workshop. Almost all of the attendees were far behind in the process. While the workshop normally would be guiding attendees through analysis (steps 6-9) at this point, most of them were still focused on identifying lines of inquiry and understanding what the data set contained (steps 1-4). Many attendees began running into problems because they had not yet looked at the data dictionary or did not understand how to use it. Some have not looked through the data set enough to have a clear sense of what could be explored with this data. Even the individuals who had spent time with the data regularly got caught up in the meaning of terms and field labels. Although this particular data dictionary was complete and well organized, the fields were heavy with medical jargon that was not always defined in lay terms. Several attendees started relying on the facilitators to answer questions about definitions, and others chose to keep moving forward in the process without clarification. Noticing the bottlenecks in the process, the facilitators spent the afternoon working with attendees on composing useful proxies and thinking through complex interactions among data fields. One attendee chose to look at infant mortality, but was cautioned that mothers' conditions would of course play a factor—therefore, any full inquiry into infant mortality would need to consider mothers as well. Another attendee wanted to create a proxy for HIV/AIDS transmission rates. The data set did not have this information, so they used what they did have—Hepatitis B and C transmission rates—as a proxy. The logic behind this decision was that all of these diseases are viral, transmitted through bodily fluids and incurable. Though Hepatitis was not a perfect proxy, it did help the attendee move forward in their workflow. Facilitators stressed that many of the benefits of working with big data is that through some quick basic statistics, an analyst can see so much more of a big picture in a short time. This allows analysts to get a sense of the whole picture before digging into the data or even forming research questions with huge biases and

assumptions. If, however, only one data field or parameter is considered, analysts lose the power of having massive data sets. Facilitators explained that with big data, it's more useful to think in a continuum rather than working with the data and assuming you are going to pinpoint a specific insight with a few calculations. Their discussion aimed to specifically highlight a key lesson: Working with big data is not just a composite of smaller statistical processes. Rather, it allows analysts to integrate more complexity and nuance in data analysis, which with the right models, can illuminate how material conditions of peoples' lives are intertwined with intersectional concerns and healthcare. On the final day, attendees spent time creating 2D visuals to communicate their findings. For five minutes, each attendee presented their methods, their interest in the research questions, and their analysis. In the end, facilitators encouraged attendees to bring the information and technical knowhow back to their own communities, hoping that computational analysis and data practices might be useful tools in diverse areas beyond just data science industries.

5. "WHAT CAN I MAKE OF THIS?'": ANALYSIS AND DISCUSSION

5.1 Data Settings Over Data Sets

My overarching research interest in this study deals with the relationships between data work and knowledge making in the analysis of big data, and how rhetoric and technical communication might inform those relationships. In Loukissas' book, *All Data are Local*, he argues that those who study data as a cultural and epistemological object must learn to "engage data infrastructures not as large, homogenous sources of information but rather as sites of controversy where varied conceptions of data come into conflict'' (Loukissas, 2019, pp.59). Along these lines, Loukissas aims to work with "data settings instead of data sets" (pp. 29). With Loukissas' words in mind, I sought to consider how data practices bump up against one another in situated work. During my observations, I noticed data work was not a rote procedure that was followed in the same way across individuals. The ways participants went about their data cleaning, collaboration, and analysis changed based on levels of expertise and the problem at hand. In order to understand how analysis is situated as opposed to objective, it is necessary to look at how individuals work together to produce knowledge, that is, to look at communities' epistemological practices. Using the preceding cases, this chapter discusses some of the methods for how data workers set and solve problems, generate questions, and how they arrive at insights through the data with which they engage. The novices discussed in Case 3 were very process-driven. They focused on learning the tools of the trade, which included statistical processes and running their

data on supercomputing clusters. They were also driven by their own relationships to the data, as demonstrated by the ways research questions were set prior to and apart from data investigation. Experienced professionals in Cases 1 and 2 tended to be more data- and concept-driven. Their work relied heavily on collaborating with subject matter experts (SMEs) in iterative ways. In the remainder of this chapter, I compare and contrast the three cases, specifically considering the differences between novice and expert practice.

5.2 Data Dictionaries and Semantics

Working definitions of parameters or data fields in a data set are crucial to successful analysis. For example, infant mortality was defined in one data set as any death that occurs in the first year of the child's life. If instead, infant mortality was defined as death within the first month, you can see how inferences from the data—and therefore potential causes and solutions—would be quite different. Maternal death in Case 3 was categorized as death related to birthing complications that occur up to a year after the baby is born. This definition is specific to the United States—many countries define a death as *maternal mortality* only if it occurs within 45 days of delivery, which again, would lead to differing understandings of the same data. The ways binary terms are defined, like *prenatal* versus *perinatal* can also shift the insights gleaned from a data set. Prenatal refers to the period before birth for both mother and child. Perinatal refers to the time immediately surrounding birth. Even *immediately* can be dependent on medical definitions, policy making or context, each of which differs from country to country. If a data set holds both perinatal and prenatal information on healthcare access, nutrition and the like, then in order to create accurate analyses, an analyst has to consider if there are crossovers or duplicates within the data. The analyst should be cognizant of why one data field was chosen

over another other for any given analysis. There is also the question of how each of these terms is defined in any given data set, as noted above. The World Health Organization defines the perinatal period from 22 weeks' gestation to seven days after birth (*Maternal and perinatal health*, 2013). Additionally, timelines that define pre- and perinatal periods often vary among medical contexts worldwide. Though the timelines are based on scientific and medical knowledge, the length of time where one term is favored over the other is a site-specific construction. Depending on how a data set defines the length of time associated with each term—and how steadfast an analyst is when making analytic decisions—data insights can easily be over- or underrepresented as analytic decisions are made throughout the process. To be precise, of course the data set does not define its terms. A person defines them when they collect data and curate it into a set. The data set is an information technology, and as a technology, social influences and rhetorical decisions are baked into its original design. There are of course traces of power-knowledge apparati and institutional values woven into each one, but even looking at a data set that is the product of a single person's work, the set is designed for a purpose and decisions are made over and over again about which kinds of data to include, exclude and how to define terms in the data dictionary. A critical theory of technology might highlight that users and creators have to approach the technical artifact differently based on their subject positions:

> In the first, primary instrumentalization, a technology is materially designed to realize the social priorities of the technology's creators, such as profit or military purposes. In the secondary instrumentalization, that technology is adopted and actually used in the world. In practice, the uses of a technology for the secondary, subordinate user subjects might not match the intents of the designers and could even run counter to them. Nevertheless, secondary user subjects' actions are limited to a "margin of maneuver" that is defined by the material design of the primary instrumentalization, delimiting what could be done with the technology. (C. M. Dalton, 2018) (See also (Feenberg, 2012, pp.113).

While users (who Dalton dubs "secondary users'' here) are indeed constrained in their operations by how the database and data collection were designed, they also have opportunities to remix the data, augment it and use if for purposes unrealized by the data set's creators. Before any mathematical analysis even occurs, a "raw" data set is a field of rhetorical decisions made and asking to be made. Let's consider another data set, which is propagated with audio recordings of conversation in public places, rather than numbers. If a user wanted to look at differences between loud and quiet places, the user has to construct the boundaries whereby quiet becomes loud. The user must also consider where the threshold between the two is porous. If basic statistics are run on the data and a normal distribution is found, then extremes of loud and quiet can be put on a continuum in line with averages and norms. Perhaps the first and third quartiles are determined as the points where quiet and loud are defined respectively. Researchers would need to consider environmental and social factors that influence the data. If a site is ambiently loud, then conversations might regularly be louder to be heard over the ambience. However, such things can be statistically controlled for. But qualities like volume and intonation are raced and gendered characteristics. In taking a purely data-driven approach, a user could still identify norms and quartiles, but would then also have to consider who the populations are in these spaces and how social characteristics and gender norms might impact the data. For example, women and men's voices are perceived and judged very differently. Women received more professional backlash for talking more than their colleagues in the workplace (Brescoll, 2011), while men who speak up are often seen as innovative, helpful leaders (Grant, 2013). Volume levels and the amount of speaking time are impacted by a history of socially constructed gender and racial dynamics. When technical definitions are only driven by the data, rather than understood as situated in a field of place, context and culture, inferences can be skewed to reinscribe existing bias and social norms.

When those who wield data fail to include datapoints that represent a wide swath of identities and social groups, the consequences can be dangerous and particularly harmful for those under- or non-represented groups of people. Such under-represented groups are often women, minorities, people of color, and impoverished people. As a prime example of how data has material impact, consider the Ugandan Bureau of Labour Statistics, which I turn to because I have seen first-hand how data arguments are intertwined with national and international aid in the country. [1] The Ugandan government conducts regular labor force household surveys to collect information "high quality and timely data on population and socio-economic characteristics of households for monitoring development performance'' (of Statistics, 2015). According to the *BBC*, in the 1990s, the Ugandan survey asked for individuals' "primary activity or job,'' but transitioned in a later year to asking individuals to also mark their secondary activities (McDonald, 2015). With that small switch in the language of the survey, the number of working Ugandans went from 6.5 to 7.2 million. The steep rise from one year to the next means that well over half a million people were unrepresented in the first version of the survey.

The gap could be due to the gendered split in the workforce and social gender roles in Uganda, where women's "primary activity'' is usually understood as wife and/or mother, even if women perform other activities that bring in the bulk of a household's income (Boonabaana, 2014). These labor statistics specifically gathered "for the purposes of monitoring development,'' impact how government money and programs

[1]Prior to this study, I completed an internship with a nongovernmental organization (NGO) in Kampala, Uganda. During that position, I was allowed to attend a meeting where city NGOs gathered to listen to and watch the national budget allocations announced for the coming fiscal year. Here in Kampala was the first time I understood how one data set could be skewed, redefined, and represented differently based on the argument that an actor wanted to put forth. As the meeting progressed, NGO representatives argued against allocations, and budgetary officials came back with statistics to reinforce their budgetary cuts. Many of the conversations I was privy to that day centered around where the government was collecting their data from and how collection methods were constructed by the government to defend the government's goals, rather than their constituents' goals.

are allocated, how foreign aid is dispersed, and how other social programs are put into practice. Discounting the value of women's work, the gendered division of labor, and sociolinguistic implications of the survey has major material consequences for the role of women as well as the amount of aid that individuals receive. Consider that this report was written for and by Ugandans themselves; how much more might have been overlooked in the "exhaustive'' data set had it been collected and interpreted by someone from outside the culture, in the same way that many large data sets are collected and analyzed?

5.3 Data Proxies

Proxies are key rhetorical tools in data analysis. Data proxies are used heavily in all kinds of data work, from social data to physics. A *proxy* is a data field, or more often a set of fields, that can stand in for missing data. Proxies are approximations of the data you want, using the data you have. For instance, sea coral grows at different rates based on ocean temperatures. When researchers do not have historic ocean temperature data, they can measure the growth and density of coral to approximate ocean temperature fluctuations from the past. Using coral as a proxy for ocean temperature is relatively straightforward and scientific, but proxies are not always so clearly identified, especially when it comes to more complex workings of social data. One scientist I interviewed was working on epidemiology, specifically the transmission of Dengue Fever. There are many ways to make predictions on how a disease will spread, and now during the early phases of the COVID-19 outbreak, hosts of new methods are likely being developed. But still, predictions are difficult to construct concerning how, when and where a disease will spread in the future. Because Dengue is transmitted via mosquitos, the scientist was attempting to create a map of mosquito densities. Previous research had used physical sensors placed in the field that capture

and crush the insects to read their genetic material and look for disease. However, due to a lack of ground truth data from the sensors, the scientist needed to construct a proxy that would tell them that same information. The scientist had to ask what other data exists that might represent mosquito density, and therefore Dengue Fever density:

> *Mosquitos have a life cycle of several weeks. They need water—rain. Rain tends to lead to healthy vegetation. So we can detect vegetation with satellite data in order to find mosquito density.*

The scientist was able to gain access to satellite data, so pictures of the earth were collected and ranges of the color green per pixel and kilometer were measured to determine high vegetation. Ranges of blue and green were measured to determine water. High vegetation plus standing water leads to mosquito population density. The vegetation, images, green and blue scales are proxies for mosquito density, which is itself a proxy for the transmission of Dengue. Scientists who work with data understand the limits and perils of overusing proxies, and they also keep in mind that proxies are used as approximations. Over-trusting outcomes based on proxies is a danger that established scientists are well aware of. Proxies are constructed by looking at context and systems surrounding the missing data. Consider how Wikipedia has been integral to the study of epidemiology:

> *You have to consider how the data set relates to the real world. Some diseases come and go very fast. Foodborne illness go really fast, and so the hour by hour timeline is really important. Context plays into it.*

> *One study used Wikipedia access logs*[2] *to inform the spread of disease. You might end up on a Wiki page about the flu. So per hour we know how many people access every article on Wikipedia. The best we can do is assume language is a proxy for geography. I know that's not really true,*

[2]Access logs refer to the number of people who open each Wikipedia page in a given timeframe.

but we didn't have a better way at the time. English works pretty well as a proxy if you assume it's the U.S.

The disease we're looking at is active only in the summers. The U.S. is in the northern hemisphere, just like most other English-speaking countries, so the disease cycles are usually the same even if the English Wikipedia accessors aren't in the U.S..

Spanish is harder [to use as a proxy] because it's in the north and south hemisphere, so the seasons and mosquito breeding cycles are different among Spanish-speaking countries.

For context, you also have to understand the type of people that are contributing to the nontraditional data in the first place: they have internet, are literate... you have to understand how that sample is biased. In the Wiki study, we analyzed a bunch of different diseases to get a set. Using Wiki for Ebola didn't track well. The signal is buried in the noise. Only tens of people get Ebola in a given outbreak, so to find the signal [how many people are accessing Ebola page] with all the people that search Wiki is hard. People who do get Ebola often don't have internet access in villages in Africa. So searches for Ebola are usually from non-affected people. Access to the Ebola Wiki page during the 4-person Ebola outbreak in Dallas skyrocketed though. You have to understand the broader context to understand the ins and outs of people's practices.

Notice here that the scientist checked how their proxies tracked compared to infection rates, and did not take them at face value. For seasoned scientists and researchers, proxies can be useful tools, but novice analysts need more experience and instruction on how and why to construct them. Otherwise, analyses can easily be heavily biased due to assumptions.

In Case 2, some analysts attempted to decide how to determine social group formation using the data. Some suggested using personality traits of individuals to group them. One participant noted that it would be a big jump to go from individual traits to deriving characteristics of groups. Though the suggested method would indeed highlight groupings of data, it would not offer any useful information on the

actual social groups that formed, because social groups are often formed separate from personality traits. Social groups can congeal based on a number of reasons, such as geography, demographics, workplaces, habits and interests. Groups and interactions might also be formed around expertise, schedules, or goals. Here, it was as if because the proxy for social grouping could not be constructed with the data, the analysts were changing the research questions—from how social groups form to how groups can be clustered in the data set. The data-driven method not only eschewed research questions, but it also would have decontextualized the data to the point of not being useful. As discussed later in this chapter, purely data-driven analysis can remove bias, but it can also quickly nullify the contextual data environment, rendering the data less meaningful.

Workshop attendees in Case 3 struggled with constructing appropriate proxies. Attendees, like the scientist above, created and honed in on a research question before digging into a the data set. Of course the attendees only had days rather than months or years to work with the data and did not have the means to collect their own. They were dealing only with the provided data set. Therefore, they had to construct less nuanced proxies to answer their questions. One attendee wanted to understand more about HIV/AIDS and maternal mortality. The data clearly did not provide information on such cases. Individuals represented by the data may have had AIDS or HIV, but the data didn't note these rates. In order to continue pursuing the original question, the attendee decided to construct a proxy for AIDS transmission. Because the set included data on Hepatitis C and B, they used this data as a proxy. While the set included many different diseases, Hepatitis was chosen by the attendee because both Hepatitis and AIDS are viral. The workshop leaders (L) worked through the proxy decision making process with the attendee (A), suggesting that the virality was not enough to use Hepatitis as a proxy:

L: But the common cold is also viral, so why else?

A: Yeah but AIDS and Hepatitis aren't curable.

L: Neither is the cold. Do Hepatitis and AIDS have co-occurrence rates? Do the demographics of people who get AIDS also often get Hepatitis? Do demographics for each disease tend to be the same?

A: I don't know. Let me check. The CDC says that the demographics are not similar. Even Hep B and C don't affect similar demographics. But the amount of new cases each year is approximately the same as AIDS.

The attendee didn't initially think to consider other similarities and differences between the two diseases and was not critical about the value of their chosen proxy. Instead, they jumped to the easiest similarity, based on their common knowledge of both diseases. In the end, the attendee did make use of Hepatitis as a proxy for AIDS. However, it is doubtful that such a construct would hold up in peer-reviewed research. Proxies are key components of data analysis. While seasoned researchers take time and care to construct them and remain critical of proxies as estimates, it is unclear if other data workers outside of academia and research-focused areas do the same. What is clear is that decision making about proxies can be heavy-handed, and analysts need to be critical and transparent about how they are constructed from other data. Another attendee asked if they could augment the data set with data about the transmission of AIDS. But just like we cannot consider men's health or men's experiences as the standard for all genders, or expect that such data also relates to women, it is important that we consider how insights about one group cannot be expanded to another group. This was an issue that emerged in the workshop but is also a more global problem in big data analytics. Taylor writes:

> It is easy to confuse globally available data with globally representative data. People in lower-income places tend to produce sparser and less granular data because they have access to previous-generation devices... This means that data about lower-income places (i.e. most of the world) cannot

> tell us as much as data about higher-income places, yet sweeping claims are being made for it in terms of transforming human life and opportunities. (Taylor, 2015)

The problems that Taylor brings up here are composition fallacies being played out on a grand scale. Since big data already has so much impact on policy and the choices that constrain us, generalizing one population's attributes for other's is problematic. It can end in discriminatory practices, bad science and bad information. Such overstepping of proxy creation, especially when it comes to identity characteristics, is a form of data colonialism. We know that issues with data proxies have already played out historically, particularly when it comes to medical research. Since antiquity, considerable medical research was based on the male form, and in Western medicine, often the white, male form. Using a young or middle-aged white man as the standard medicalized body, anything that deviated from it would be abnormal. Besides the cultural problems with that assumption, it often played out such that women and non-white bodies were not studied. Therefore such individuals were not understood as well and not treated as comprehensively as white men (Epstein, 2009, pp.40-42). It wasn't until the middle of the 19th century that the United States graduated its first black doctor, and not until the turn of the next century that we saw our first woman doctor. Without women and minorities practicing as physicians, people in "nonstandard'' bodies were relegated as afterthoughts in medical research and practice. In fact, it wasn't until the 1990s, one hundred years after the United States had its first woman doctor, that the FDA began requiring that clinical trials of drugs include women Schiebinger, 1987, pp.114-17). Many argue that the effects of treating women and minorities as nonstandard bodies are still dangerously at play in the medical industry today. When data workers craft proxies, they are constructing an argument as to why certain kinds of data represent other kinds. Their decisions on how to craft a proxy are situated in their research questions and the data they have

access to. If the data set or the questions change, then so do the elements that construct a proxy. By definition, proxies are items that are missing from a data set. They are objects that coalesce from a combination of access, bias, and the situated perspectives a user has on what data means in the context of related data and in the context of the material and social world.

5.4 Developing a Common Language

As a unit, the team of scientists in Case 1 constructed their understanding of the project primarily during the weekly meetings, via a back and forth dialogue of questions, answers, and posing arguments of possible paths forward. According to the interviews, this dialogue-centered time is the core structure of how work gets accomplished in the group. While most meetings began by checking in with everyone or asking for updates on progress that occurred outside of the group meetings, it was not the phase where participants felt like "work is getting done.'' Instead, "real'' work was being done when 1) brainstorming and discussion of an issue occurred; 2) constraints or issues were uncovered; or 3) new ''to-dos" were added to the group's tasks from issues that arose during meeting discussions. The fact that people identified group meetings as central to the ''real work," rather than individual or solitary effort, may be a sign that their discourse served a primary role in productivity and scientific innovation. This team flipped classical Platonic dialectic on its head. For Plato and many of the ancient Greek philosophers, the goal of discourse was philosophical—to focus on abstract ideas and ideals, rather than situated, specific context and action (Yunis, 2011). However, dialectic, as a strategy to find insight, does not apply to real cases. It is only concerned with the ideal, the ephemeral, and purely theoretical. Instead, when principles have to be applied to actual cases, and solutions are pursued that are situated in context, knowledge-making must be dia-

logic, based on back and forth exchange of ideas that does not often follow linearly to an all-purpose solution. Applied knowledge is messy. It goes down many avenues and is continuously informed by various perspectives in order to arrive at working, concrete actions (Leff, 2000); (Gegeo & Watson-Gegeo, 2001); (Cissna & Anderson, 2008). Because of this dialogic model, regular meetings held a central position in the actual production of the project. For some members, this group is the only place they collaborate outside their disciplines, rather than working on their own piece of the puzzle. Much of the disciplinary work is being done individually or in pairs outside the main meeting, while the interdisciplinary work tends to (though not exclusively) happen during the meetings. This also makes the regular, paced meetings of prime importance to the group. Over half of these scientists mentioned what I call ''the structured learning phase" and the development of a common language as part of the group's collaborative apparatus. For experts from different fields, learning to speak the same language in terms of vocabulary is crucial for interdisciplinary work to succeed. Developing a common language though is more than differentiating terms. The language of problem setting has to operate on the same scale. One scientist noted the difference between the approximation through which physicists operate and the precision that mathematicians require. Their experience helps illustrate the scale of problem setting as it relates to a shared discourse:

> *For the first day of three days after flying out for this meeting, we are trying to find ways to talk to each other. [We got] very technical meanings from mathematicians. Over-technical. And we had to find a common language through the math, because physics and math take different approaches. Math doesn't buy it if you speak the physicist language until you speak the math language. The life of a physicist is to approximate, because you know you cannot actually solve anything. So we are ok deriving and doing approximations, where math has to prove things... The second day we start to speak a common language [through the precision of math], and then we can start brainstorming things. Intuition only brings you so*

> *far, eventually you have to convince someone else. That's where domain knowledge comes in, and that common language.*

Another statistician noted that it is their job as interdisciplinary data experts to learn the language of their collaborators:

> *Of course, communicating between different disciplines [is a primary part of the job]. Pretty typical for a data analysis person in a place like this. It's our job to go past half way to understand their language.*

Scientific and computational jargon can mean different things across disciplines, so one key way that language had to specifically be addressed was to delineate one term from the same term in another field. Team members often asked questions about the boundaries of terms or how one phrase relates to another term that was already understood by the group. Developing a set of common definitions was an ongoing part of their interdisciplinary collaboration. One analyst in Case 2 also noted that doing data work—especially with interdisciplinary experts and withdata that was provided to you rather than generated by you—is "really about creating a common vocabulary.'' In the sprint, it might have meant defining what a *group* means or how an *interaction* is defined. What are the limits of *colocation*? Do we define it in terms of time? Geography? Is three feet distance collocated? Or 20? In order to draw boundaries about what counts as colocation, analysts can rely on social context or the data set. Contextual arguments for defining *colocation* might consider real people and what it means to walk together through a park. A consistent span of one to four feet between them might mean they are walking together. Being collocated in an office might mean two people in the same conference room—ten to twenty feet would suffice in that case. Most data has mathematic breaks where the data separates, and each data set is different. Therefore for a data-driven approach to definitions of proximity in this case, you could look at the obvious breaks between close, medium range, and distant to define the boundaries of co-location. With complex data, both approaches

are needed. Understanding data as a text about a place or setting leads to a socio-contextual definition. Bouncing this definition off of the data-driven mathematical clustering could render the definition relevant for both the quantitative data and the person as data. However, in the end, the breaks or buckets of colocation (as in high colocation, low, no colocation, for example) are semi-arbitrary distinctions made by the data analysts. All in all, there are several specific ways that discourse impacts data analysis, both in social and technical ways. Discursive tasks emerged in each of the three cases. Beyond the development of common terms, data analysis inherently requires an attention to definitions and audiences, though this aspect of quantitative work is rarely foregrounded in data science scholarship.

5.5 Data-Driven Versus Situated Analysis

Across these three cases, data workers illustrate that analysis is neither purely data-driven nor purely situated in a local context, but is an intersection of both when it is successful. Rather than thinking of rhetorical data analysis as the opposite of data-driven work, I argue that it is helpful to think about rhetorical techne as the midpoint between the extremes of purely data-driven and purely situated data work.

5.5.1 Narrative and Problem Setting

One of my participants noted early in the study that the difference between data science and statistics is storytelling. As writing scholars, we know the power of narrative well. Information visualization is a common topic in our field, and we often point to visual arguments in the news, during elections and other infographics that communicate data in narrative and evocative ways. However, during this study, I found that scientists were also very aware of how necessary narrative visualization is

for them in their publications and presentations to stakeholders. As noted in Chapter 2, data visualization experts have been discussing narrative visualizations at length for five to ten years. But narrative also works on a simpler, older and more in-grained register in data work and visualization design. Narrative and problem-setting work hand in hand, especially when it comes to the metanarratives about the value of scientific work. During the formative problem characterization, metanarratives are useful tools. Munzner (Munzner, 2009)identifies several levels of tasks implicit in initially setting the data problem, which germinates with high-level science problems. These are essentially overarching, often romantic goals, such as aiding scientific discovery or taking action on issues of social justice related to maternal death, for example. Munzner defines "high level domain characterization" as the first task in a string of visualization tasks, but the narrative is often a pre-existing idea, rather than an actionable item. The metanarratives drawn on during scientific data work concern the value of the project. Scientists attach to such metanarratives in order to form meaning in their own and the team's work (Goffman, 2002).While metanarratives—the high-level science problems—are not detailed enough to directly inform workflows or applications, such narratives do frame the collaboration and individual motivation that underscores the success and longevity of a project or goal. The narrative has other practical purposes, like demonstrating the project's value and connection to the mission of the lab and other funding bodies. Though LANL is a mission-driven weapons lab, and that mission was explicitly brought up by the majority of my participants, the underlying narrative that scientists personally attach to tends to be a more romantic one. The scientist's narrative is a heroic story—as all the most powerful ones are—about innovation, experimentation and triumph of collaborative insight over technological barriers. It's a story about exploring, conquering and making the world better through scientific practice. The stories we tell ourselves are powerful

motivators. They color how we operate in the world, so awareness of narrative's place in scientific data work is useful for understanding what motivates individuals and how they situate their own values and purpose within their everyday workings with data.

5.5.2 Expert Vs. Novice Analysis

Based on these cases, experienced data workers first go to the data to set the problem by understanding the constraints and affordances of the data set. Their goal is not to consider what the data says as if it speaks for itself, like some scholars contend (see Chapter 2). Instead, getting a full view of what the data set includes helps analysts know what they can ask and how they might process the data based on their questions. Good data scientists are "informed sceptics who balance judgement with analysis and incorporate the key qualities of a sense of wonder, a quantitative knack, persistence and technical skills'' (Strong, 2014, pp.338). Novices, on the other hand, tended to approach data analysis in much the same way that inexperienced student writers approach textual research. They began with a question that interested them and a pre-determined expectation of the arguments they wanted to find in the data. Like a student in a first-year writing class who forms an opinion first before cherry picking scholarly evidence to support it, novice data workers were not as critical of how their own pre-determined biases and positionalities will influence analysis. Novices tended to be focused on solutions, as if they saw data work similar to a mathematical equation of inputs and outputs. Novices operated as if data analysis was a teleological event. Experienced data workers, on the other hand, gave the sense that analysis was more porous, interpretive, and ongoing. This could be due to the nature of the workshop where the novices interacted, which had a short timeline and a cutoff date. It could equally be due to the ways that research cultures value incremental work

that adds to a larger history of research. However, it is worth reiterating that I have not found that experienced scientists see data at face value. Instead, they use it the way I would use research and evidence: to get a more ecological understanding of a problem space before whittling down to an argument.

5.5.3 Composing Models

Models are key tools in scientific data analysis that require human decision-making and a thorough understanding of a data set. Models are mathematical frameworks that describe a system. In the everyday work of scientific data analysis, models are built from scratch, taken from published literature, or most often, models are built using bits and pieces of pre-existing, published models used in other scientific problems. Models are modular, in that can be used and recombined in a number of ways. Models can be an input of a research project, where scientists build the model to run their data. But models can also be research outputs. When new models are developed as tools for novel data problems, they are published and peer reviewed, just like other kinds of scientific results. I learned that the work of composing models involves a lot of creativity and decision making, which cannot be completed by computers. This is one phase of data analysis that many of my interviewees noted as being very important but outside of the rote steps that data visualization scholarship often discusses. As a key component of scientific data analysis, models are highly situated in the creator's experience, intuition, and the problem at hand.

> *There's a modeling step that's highly cognitive and intuitive. It's depending on the person and the experience of the person. If you have a background in theoretical understanding, you try to work out some mechanism or underlying process that might explain the data. Like, suppose you have an epidemic model. There are math models for epidemics. There are also heuristic models that map onto the theoretical models and the practices of epidemiologists. The notion of a contact is something that needs to be*

understood first. You don't just go into the data and pull stuff out. You need the contact rate, and you use that to make a theoretical model and then hypotheses and find trends.

Modeling goes back to your own human way of relating to the world. You can't know the world. You can only relate to it from that cognitive model. We use them to parse our experience as humans. Now we can do that with math. Any relationship between a description of the world and the world itself is a model...[If a problem is difficult, it concerns] what kind of model are you going to use. There's a lot of art to it, and you need to have experience. You can look at a task after a while and think that's going to be hard, because I've seen the problem before and I know it's hard. You develop a gut feeling over time.

In deep learning, if you have a lot of data, many millions of records, you can build strong predictive models that generalize. But when you don't have as much data, you turn to a model and hypothesize from that. And [the model] makes up for the data you don't have. And you never know if you are choosing the right data set to train your model. That's where the intuition lies in big data is—if you've collected the right and right amount of data for your model.

My whole life I've been modeling things, but recently the idea of modeling has become a thing: what makes a good model, how you represent it. It's a thing in recent decades. That's something that's hiding from the process [that is often discussed in data science literature]. The computer is doing some of it, and you are doing some of it. Take the idea of modeling as surrogate for the real word, and you can figure things out.

A model is a ***thing*** *[as in, a component in the research, just like the data itself]. And how do I describe my model from a catalogue of models to get you to choose it?*

...The modeling phase where you build the model. But I hesitate to call it [a phase] because it implies you just choose and stick with a model. You ***build*** *the model and show it to the domain scientist. And then I go back to the data phase or another phase. It's a loop. And my goal is the deployment phase when you have production ready software for users and it's out running somewhere.*

5.5.4 Problem Setting

The analysts in Case 2 had difficulty working with their data because they did not have a clear sense of their stakeholders' goals (i.e., the research team) during the sprint. Though analysts wanted to "get a feel for the data'' from the onset, the constraints of the sprint detoured them into more inventive practices first. The sprint's framework was over-exploratory. Problem setting was unlikely without the constraints of stakeholders, goals, and available resources (the data set). Analysts wanted to orient to the problem by understanding the data—what was there, what was missing, and how terms and fields were being defined in the data dictionary. The available means of the analysis was locked up in the data set. Though the room was filled with the highest levels of interdisciplinary experts, expertise was not enough to make sense of the problem if the problem was not directly put into conversation with the data or tools of analysis.

Problem setting was not a step prior to analysis, but a cooccurring task that grew and changed as the analysts worked with the data. Problem setting could *not* be done using only 1) the researcher's goals, 2) invention techniques, or 3) the range of a priori expertise that filled the room. Setting the problem space could only occur when analysts were able to find and create arguments with the data set and add it to the previous three items. As Cushman notes:

> In problem setting, processes are not a given because a process, in general, presupposes a means–end analysis in which both the starting position and the final goal are well defined. Processes require given problems. The practice of problem setting, however, means attuning to instability and indeterminate situations, acknowledging that processes and problems are mutually constituted. (Cushman, 2014, pp.330)

There is a complication I would add to Cushman's argument that processes require given problems. To some extent that is always true. However, experienced

analysts did have a set of processes in their back pockets that they wanted to use as soon as they first got access to the data set. Such practices of running basic statistics and data scrubbing existed outside of this particular problem space. Those practices are based on years of professional experience and training. Both basic statistics and data scrubbing, however, are practices that structure the rest of the knowledge making process. Because "well-designed work materials become integrated into the way people think, see, and control activities'' (Hollan, Hutchins, & Kirsh, 2000, pp.178), data are tools that are an invaluable part of distributed cognition.

Foundational work of statistics and scrubbing might be the data science equivalent of taking reading notes. As a writer, you may not yet know how the notes or your reading will influence any final research you produce, but it is necessary for you to understand new arguments as they are published in the field. Such work is both foundational and inventive in the rhetorical sense, but still often prior to specific problem being set. The process of running statistics and preliminary understanding of the components of a data set is where a data analyst's domain expertise weighs heavily in the larger analytic process. Though later steps, after the data has been "read,'' the subject matter expert's insights might weigh more heavily in the analysis that the data analyst's. The ebb and flow of various expertise in collaborative scientific data work is part of the techne of data analysis, which has to include an evolving response to novel situations as they develop, based on iterative problem setting and on a knowledge of the tools at hand. Sometimes these tools are data sets, and other times the tools are subject matter expertise. The overall point is that data analytic work is situated rather than objective. When knowledge work is situated, it influences and is influenced by specific happenings and environments from which it was produced. While Haraway's idea of situated knowledge hones in on how individual identity characteristics such as gender play a role in knowledge production, this case

highlights that 1) stakeholders, 2) the objects of study, and 3) access to interdisciplinary perspectives are also key in data analysis. Access to data without subject matter expertise can be hollow. Expertise without data is not useful either. In both cases, working with data is a situated experience where the analyst makes choices that guide the work, and those choices are specific to the analyst and their identity. One scientist described data analysis like using a recipe:

> *It's kind of like cooking in a way. I can hand you groceries and say make something, and in a way it's procedural, but what you actually do is make decisions about the tools you use, and that's pretty cut and dry. But then your final product is judged by somebody that may or may not like it for different reasons. "Here's a bunch of stuff. Make something." What I will make is going to be different from someone else, even with the same background. Within a general recipe you have to make smaller decisions, like what it means to mix things together. You have constraints, and depending on background, domain, experience... those constraints are slightly different. Even within the same workflow... there's a bunch of decisions we make. There's a lot of freedom in each area of the framework. And that's where the creativity comes in.*

5.5.5 Issues with Forgoing Data-Driven Approaches

As I argued in Chapter 2, conducting analysis that is purely data-driven and devoid of either subject matter expertise or theoretical backing can lead to serious problems with the insights drawn from data devoid of context. However, here I discuss some of the issues that arise when data work is approached from the opposite end, without centering the data set in the problem space. For workshop attendees, their preliminary research questions were created after they had an abstract knowledge of topics that the database contained but before attendees actually began looking directly at the available data. Because of their focus on topics that were interesting to themselves, rather than how they might use data to work through such topics, their research aims were void of the means to pursue them. Even if they would

have had the opportunity to pursue other data sets outside of the one presented to them, many of their questions were issues too complex to be explored using a single data set and would have required a deep dive into politics, healthcare infrastructures, policymaking, and cultural milieus. Attendees had trouble narrowing and focusing on questions that were specific and small enough to be pursued through the data set. In a sense, attendees questions were rooted in a problem space—the problem of racial and socioeconomic inequity of the U.S. healthcare system. But they were not grounded in the specific data or epistemological ecology at hand. Like the workshop attendees, analysts at the sprint were asked to form questions before looking at the data. Using such a structure meant that analysts were being asked to begin forming their inquiry around their *a priori* knowledge and expertise. This is the opposite of a data-driven approach where data takes precedent. For data sprints, which are designed specifically for exploratory work, participants might get a short glimpse of the study's mechanics and begin invention work before ever looking at the types and characteristics of the data. Therefore, for the analysts, the invention/heuristic phase would be very disconnected from the data, and any insight that could be constructed would not useful until the data comes into the problem space. If sprint participants are tasked with forming curiosities based on their nebulous sense of what may be in the data, rather than using the ground truth data as a starting point, the work will lack grounded, situated hypotheses.

5.5.6 A Mix of Problem and Data Driven

One LANL scientist noted that during his normal process of data analysis, he begins with basic scrubbing and statistics, then after he "gets a feel for the data,'' the process shifts:

> *I have to stand back and think, what can I make out of this? Then stand back and look at it after it's made to figure out what else needs to be done or what else could potentially be done.*

When the participant asks themselves *What can I make out of this?*, they did not mean how can I understand it, but what can I construct or put together from this? This participant described their process as a physical, embodied practice, feeling out the data and constructing an object from which to find insight. He begins analysis with rote procedures, changing them slightly as he feels around the data. Then he works to decide on the curiosities and arguments that the data set might highlight, based on his own disciplinary knowledge and professional experience working with data. He then often enlisted the collaboration of SMEs to go further into what possibilities the data set might hold in terms of the scientific problem being addressed. Another scientist noted that some decisions are conscious and backed by expertise, while others are unconscious, but still affecting the analysis. However, they highlight that taking humans out of the process has its own perils. They described what happens when humans are pulled out of the data workflow in lieu of machine learning:

> *You think you know what you need, then you start worrying and pulling in more data So there can be unconscious decision processes in limiting the data at the collection or the feature level. Then there are techniques to see which of the features is most predictive. You can get to this point and still have no intuition about why it's important. So what you've done is deferred the modeling process beyond stats. And you can keep deferring it with machine learning until you have an answer, but you don't know why you have it or what it means.*

Many of the study participants who were well-versed in data analytic work employed a mix data-driven and problem-based, situated knowledge work. Both practices on the spectrum require distinct decisions to be made that are focused from a particular discipline disciplinary vantage point. Rhetorical *techne* that experts

perform tends to exist somewhere between the two extremes, ebbing and flowing variously depending on the stage of data analysis, the immediate environment and tools available to them, and the research questions being pursued.

5.6 The Roles of Subject Matter Expertise in Data Analysis

Silver, in his book, *The Signal and the Noise* (Silver, 2012), warned that the increase of available data has not increased the number of "meaningful relationships'' proportionately. The rise of big data instead risks simply generating a larger number of false positives. In Strong's piece, "The challenge of 'Big Data': What does it mean for the qualitative research industry?'' he echoes the same fears. Our brains are trained to see patterns. However, Strong argues that "when looking at a truly random sequence, we tend to think there are patterns in the data because they somehow look too ordered or 'lumpy'... which leads us to believe that there are real relationships in the data where, in fact, the linkage is trivial'' (Strong, 2014, pp.338). Strong and others have criticized big data as a whole for providing a field where the numbers speak for themselves but may turn out to be empty (see Chapter 2). However, when subject matter experts work closely with data analysts, trivial linkages can often be explored and explained rather than taken at face value.

One scientist-statistician explained it well:

> *Your classes tell you to do this do this do this, and you can program part of that, but what about borderline, you have to decide on arbitrary lines sometimes. And that's where you have to bring the subject matter experts. If I see something odd in data, I ask the SMEs to see if it is human error. We [statisticians] are not quite data analysts because we've had more research experience, and we've been trained in trivial / basic methods. But it's ours to decide what kinds of models don't fit because of very specific things in the data... You can say the outlier is never going to happen, but a machine will just make the model bigger rather than see as a mistake, so we have a lot of back and forth with SMEs. The creativity lies*

> *in changing the model. You have to be creative to vary the model. To be able to change it constantly because of the information you get back from the SME.*

Another scientist passionately highlighted how data work can easily become less meaningful without SMEs:

> *Don't drop the research question thing. Cause otherwise people would find trends, but not know what it means. [SMEs] are helpful to form questions and make sense of trends and give more information about connections and the kinds of people [represented in the data]. I know what the clusters are, but SMEs need to make sense of what the clusters mean. Even with exploratory work, [SMEs] are important.*

SMEs help create meaning by adding context needed to convert data into knowledge. Those who I would dub data scientists practice back and forth work with SMEs regularly in their daily work. According to one participant, normal data practice includes:

1. Running descriptive (statistical) analyses from a data set,
2. Working with SMEs to define terms, fields and how data might interrelate, and
3. Helping the SMEs identify the kinds and amounts of data that are missing while also helping them hedge their insights based on the data that is available at a given time.

While data scientists at the lab tended to have productive working relationships with other researchers, depending on the site, statisticians and other types of data scientists are often brought into a team at the end, rather than seen as collaborators and trusted coresearchers throughout a project (National Academies of Sciences, Engineering, and Medicine, 2017). One participant at the sprint noted differences between a social scientist, who might identify a theory or model based on their expertise, and the people we might call data scientists, who have to "sneak up on what

might be interesting in the data.'' Their meta goals are the same for this project, but methods are clearly different based on the tools (theory or statistics) that come from their discipline. Disciplinarity can be critiqued for being staunchly rooted in only one way of seeing the world. Subject matter experts understand problems in part through their technological tools, including theory, algorithms, statistics, and applications. The adage—*If you are a hammer, everything looks like a nail*—comes to mind. However, when interdisciplinary teams come together with each member orienting to the problem through the lens of their own expertise, teams can thrive. Particularly when disciplinary knowledge is purposefully shared with the group to form a baseline of understanding, each discipline's point of view can assemble and diverge productively as a project evolves. According to Schön (Schön, 1938), professionals often see theoretical knowledge as higher than knowledge of specific practice. This is true to some extent at the lab as well, where theoretical and other forms of high-level physics is privileged over other domains. However, as professional scientists, individuals also have to be productive members of teams that produce work, so being able to put abstract or theoretical knowledge into practice is still a valued trait. Productive contribution is held possibly higher than abstract knowledge in working teams. Teams at the lab often value interdisciplinarity, because the scale and scope of big scientific problems requires input from a multitude of perspectives. Subject matter experts can open or constrain what questions are asked or not asked and which solutions are pursued. The existence of SMEs in the data process is part of the situatedness of analysis. Disciplines are, as Manoff notes, "discursive formations or systematic conceptual frameworks that define their own truth criteria'' (Manoff, 2004); See also (Foucault, 1982). Having a range of disciplinary backgrounds on a project offers perspectives from a range of different truth criteria. It also offers an opportunity to meld various disciplinary truth criteria through the process of collaboration. Adding

subject matter experts to a big data project can move the insights from statistical to workable models that reflect disciplinary theories. Subject matter expertise helps bring big data from the brink of a-theoretical information to knowledge grounded in deep understandings of a particular problem space.

5.7 Structured Learning in Interdisciplinary Data Work

Each case included a kind of structured learning phase, though they occurred with varying degrees of success. Data sprints include an introductory time period where participants listen to various researchers about the purpose of the study and its methods. During the Data Sprint, analysts knew there were people from social science, statistics and computing, among other areas, but they did not have a clear sense of individuals' disciplinary identities. They had a sense that they were meant to work from various disciplinary vantage points, but the structure of the sprint did not allow them to purposefully form multidisciplinary groups. There was a desire for all parties to understand the various areas of expertise that existed in the room, especially as it related to the problem at hand. Yet in practice, there wasn't enough emphasis how to make that occur. One participant noted that the interdisciplinary teams need to be oriented to the larger landscape of expertise in the group. This orientation would help analysts figure out the boundaries of the study and what was possible and also how their own expertise fit into the group. The participant noted that the lack of orientation could be a reason why the sprint team struggled. In essence, while the structured learning phase existed, it wasn't enough to have a variety of expertise in the room if as a group they were either unclear as to what others' expertise meant or how their foci would work with other experts in the room to approach the analysis. The workshop also included an introduction to the data set and the larger issue of maternal mortality. However, being that there were few subject matter experts as

attendees, it wasn't necessary for each of them to let the others know how their previous knowledge or experience might come into play. However, workshop leaders did clearly introduce their own disciplinary areas such as computation, social science or mentorship. Leaders let the participants know who they could go to with certain kinds of problems. The workshop as a whole was a kind of structured learning phase combined with practice. Most of the teaching here was dedicated to working with the supercomputing clusters rather than the meaning of the data or how to create models from the data that held up to real life. The scientists took full advantage of the structured learning phase. For months, the team dedicated time for each member to give an in-depth presentation to the group on their background expertise and to explain the technical details of the problem from their disciplinary vantage point. The team's knowledge as a whole was forged from the incremental articulation work of learning others' fields and technical terms, then continually resituating their own ideas and expertise in the evolving landscape of the others. Multiple scientists noted that though some teams can be minefields of egos and arguments, this team worked together very smoothly, in part because of mutual trust. Their trust was not built from personal familiarity with each other. It was a result of each person having a clear understanding of how their work was valued in the group, how it fit into the larger landscape of knowledge, and that scientists with expertise outside their own were willing to learn from each other. The team was structured so that members willingly and enthusiastically shared knowledge and developed skills among each other. They were bound together by the scientific problem at hand, yes, but also by the understanding that working collaboratively helps individuals create a body of shared knowledge, practices and skills that they could not achieve alone. The structured learning phase in this team illustrated that to some degree they understood

that knowledge is a social act, a participatory one, rather than an object that exists in the minds of individuals.

6. DATA RHETORICS IN CURRICULA, INDUSTRY AND ENGAGEMENT

6.1 The Intellectual Work of Technical Communication Experts

Much of my time and intellectual energy during this study was spent developing an understanding scientists' task environments as they structure their work with big data. Task environments might include lab culture, the tools and technologies used, or domain-specific discourse. Basic concepts, vocabulary and field concerns were foreign to me at the onset of this research. However, by the second summer of my fieldwork, I realized how much clout came from being able to speak in the language of my participants and coworkers. Terms, jargon, and a cursory understanding of a small range of scientific issues were the tools I relied on to gain access to scientific experts. In large part, I intentionally studied three key areas in order to become conversationally proficient at the lab:

1. Computation: Jargon, machine architectures, HPC, exascale computing issues, artificial intelligence, neural networks and machine learning

2. Data analysis: Modeling, simulation, statistical terms and processes

3. Data visualization: Data provenance, task taxonomies, in situ visualization, mesh-based physics simulations, color mapping

4. Lab culture: Low-level physics discourse, lab hierarchies, confidentiality and security, lab history.

My process of learning basic discourse is perhaps an illustration of the intellectual work and expertise that technical writers perform when they enter a new workplace or tackle a novel project. Beyond just an understanding of audience or how to translate technical jargon for various publics, technical writers need to understand the basics of technical processes, tools and workplace discourse. I studied the technical, social and discursive elements in formal and informal ways. Of course, as a technical writing scholar, I was particularly attuned to how the people around me spoke about their work and the terms they used. Once I thought I had a grasp of a concept, I explained it to others and asked them what they would add to my understanding. This was perhaps my most useful tactic. Formally, I undertook an extensive bibliographic review of issues in data visualization (see Chapter 3). In this final chapter, I discuss how and why we should teach a rhetorical understanding of data to our technical writing students. In order to prepare students for professional practice, we as instructors need to consider the theoretical concerns as well as situated tactics that allow students—especially our professional and technical writing majors—to enter a workplace having some comfort with data processes and data as arguments. Data work is a composite "blend of human and computer meanings'' (Manovich, 2001, pp.46), and understanding meaning is a key skill that our students need to wield professionally. Additionally in this chapter, I put forth a discussion on data citizenship and some intersectional feminist interventions made possible through data practice. Finally, I consider future research and the tributaries of inquiry that are emerging from this study.

6.2 Data Rhetorics in Professional and Technical Writing

6.2.1 Why Data is a Rhetorical Topic Needed in Technical Writing Programs

Data does not speak for itself, nor does it emerge out of the ether. Data is *constructed* at each point in its life cycle. Data sets are rhetorical forms of inscription that are composed, which communicate and actively shape meaning. Data are not static bits of information, but ecologically oriented components of a larger infrastructure of expertise, *techne* and sensemaking, leveraging the affordances of technological interfaces, situated goals, and domain-specific epistemologies. The interconnection between data, evidence, fact and argument is tightly coupled with how insight—scientific or otherwise—is shared and put to work on the world's problems. Data work requires analytical processes and also artful *techne* that is situated in ongoing reflective praxis. As purely analytic, data work focuses on mathematical treatments and step by step formulas and procedures. As *techne*—defined by Hawk as "techniques for situating bodies in context'' (Hawk, 2004, pp.381)—data work requires interpretation, from translating insight into meaning to defining terms and acknowledging biases. Hawk argues that *techne* involves both "a rational, conscious capacity to produce and an intuitive, unconscious ability to make'' (Hawk, 2004, pp.372). Scientists derive their insights from kairotic moments, lived experience, and disciplinary expertise rather than from only a formal set of scientific procedures or explicit training in statistics. Understanding *techne* in data work as artful rather than simply analytical is a generative way to produce richer, more contextualized data practice and insights.

In Case 3, workshop attendees continually used the word *solutions* and focused on finding solutions with the data. Their penchant for the word was in part due to

the language used in the workshop by facilitators. Yet, facilitators also emphasized that attendees were unlikely to find solutions in the time period allotted, either to the wicked problem of maternal mortality or to their own specific lines of inquiry. *Solution* carries with it a teleological bearing that leads to closure, to Fact and Truth. I want to suggest here that *insight* is a more rhetorically accurate term when it comes to evidence constructed through data analysis, because where solutions are final, insights are only fragments of a larger knowledge infrastructure. Solutions denote finality; insights denote possibility. Solutions do not emerge from data, but from an ecology of human decision making, argumentation and actions in contexts of social life. Students need to understand the *why* behind analytic choices. This does not mean that technical writing and rhetoric courses need to take up the instruction of statistics. However, I am suggesting that there is space in our programs for more quantitative and data-centric rhetorical pedagogy that deals with data collection, visualization, ethics and the ways in which arguments are constructed through data in various publics and industries. Research skills that will be valued in the future include data analysis (Boyd & Crawford, 2012). Even the NSF's report on data science education emphasizes that education in analysis is not a recipe, but an attention to the mechanics of interpretation and decision making (Berman et al., 2016b). Students across the university should be taught how to understand data apart from someone's treatment or visualization of it. They need to understand how to interrogate data arguments and evaluate credible sources of data. It is imperative in the age of data that students are versed in the practices of data argumentation as well as the social and rhetorical theory that they need in order to make meaning from such arguments.

sub

6.3 Teaching Rhetorics of Data

Data visualization already falls under the purview of technical communication curricula. While visualization is useful in most data arguments, scientific data analysis that falls under the purview of big data *requires* it. Big data cannot be understood without the help of simple 2D visualizations and sometimes more complex mesh or 3D objects. When it comes to big data, visualization is epistemic. As a field, we discuss visual rhetoric often in terms of information visualization, where practitioners are creating visuals for others to interpret. We have yet to delve into the realm of visualization as a tool in data sensemaking, not for communication to others, but as a cognitive tool for the analysts. Our students should understand how to ethically and effectively construct a range of visualization types, but they should also understand how to variously create visualizations to help themselves understand data. Visualization is a type of inquiry when data-based arguments come into play in students' professional and civic work. Visualization is a bourgeoning field where students of technical communication can contribute (Salvo, 2012). Narrative is also an intrinsic part of data visualization, and data storytelling is currently of interest to data visualization scholars and technicians. Stories help audiences make sense of arguments, and stories also make those arguments more interesting, memorable, and sometimes more palatable. While other fields might do more work in the technical aspects of data analysis, no other field is positioned better to contribute to narrative development in visualization more than technical writing and rhetoric. As our students enter their professional fields, the ability to narrativize data and create data stories for research, marketing, grant writing and a number of other fields will help industries understand the value of students with rhetorical and composition educations. According to the NSF, ethics is a required part of any responsible data science curriculum (Berman et al., 2016b). However, in my informal search through data science curricula at uni-

versities across the country, there are very few required courses that focus on either ethics or visualization. It is easy to unintentionally create misleading visuals if one does not know where the common pitfalls are or is not versed in critical understandings of data visualization. It is also perhaps easy to forget the human and material concerns impacted by data analysis if people have not made a practice of considering the implications of the numbers they analyze. There are already "growing discontinuities between the research practices of data science and established tools of research ethics regulation" (Metcalf & Crawford, 2016, pp.1). The discontinuities are in part due to the speed at which tools and practices are being developed. The ethics of research, visualization, and human subjects are areas where technical communication curricula should lead. The area of data rhetorics could also cover socio-cultural education as it pertains to data, which would go hand in hand with students' technical education. Topics might include foundational knowledge in intersectionality, systemic and cultural oppression, situated knowledge production, the ways evidence and fact have been produced throughout history, as well as issues of visualization accessibility and the ethical issues of blanket, covert data collection as it relates to internet and smartphone users. We have to teach students to look at who collected the data and why. We should teach lessons on how data has been decontextualized from a specific problem space, and why that matters for data insights derived from decontextualized data. The challenges of big data sensemaking are also the strengths of our English and Professional Writing majors. In Kitchin's foundation work in critical data studies, he frames the challenges as: "coping with abundance, exhaustively and variety, timeliness and dynamism, messiness and uncertainty, high relationality. . . '" (Kitchin, 2014, pp.2). His statement brings to mind the work of dealing with textual research and composing arguments. Writing is an iterative process of coping with variety, abundance and messiness. Rhetoric's purview is timeliness, relationality and

how arguments relate to human activity. Qualitative and discursive research skills lend themselves well to such challenges as well. If we consider some of our often-used qualitative methodologies and apply them to how we teach students about data practice, we can help them see the rich connections and available arguments in a data set, rather than jumping to the most basic statistics and obvious data insights. Some additional data topics that technical communication instructors are poised to offer may include:

- The construction of data proxies, including experience arguing for the proxies they chose.
- The decoding and construction of data dictionaries, which can be either an obstacle or a key tool in analytics.
- Experience linking intersectional methodology with data work.
- Finding and critiquing data sets based on inquiry into who created them, for what end, and with which populations.
- Tracing how an open data set is deployed by various actors over time and media to see how arguments are countered and used for vastly different purposes.
- The differences between data inquiry from a specific data set versus looking for data with a topic in mind. These are different processes with different affordances.

In our rhetoric and professional writing courses, we should introduce students to the critical discourse on data. We can frame data work as having the potential to be both big and rich and the ability to spur decisions while still existing as contingent forms of arguments. Many industries right now have a certain reverence for data. Analysis and data work is a highly valued, highly sought-after set of skills. Our students

would benefit from being able to speak the language of data analysis and having the expertise to narrativize and communicate data to publics. Our students should learn the language and jargon of data analysis in order to enter these industries. Our field is armed to include such critical work in our classrooms in order to send students into their professional and civic spaces well-armed.

6.4 Professional and Technical Communicators in Data-Centric Industries

There is room for rhetorical scholars and technical writing practitioners to intercede in data work when it claims to be transparent or without interpretation. Multiple points of entry exist where scholars could and should intervene. These are not limited to studies of inquiry and hermeneutics. Ethnographic and archeological, thick descriptive work on data and its cultural and discursive impacts is needed. It is not enough to end with mapping big data. The goal is to change how data is wielded and to bring to light the deep and complex interactions between history, epistemology and the ways the influence of non-situated data arguments radiate through culture and material domains. If the split between qualitative and quantitative research continues to be glorified and widened, rather than using each to enrich the other, everyone loses. Qualitative work might continue to be obscured in the public age of big data, but the overlapping features I have touched upon are starting points that help us consider how qualitative and rhetorical scholarship can create critical and generative connections with data science. Technical writing practitioners have a part to play in data as it interacts with governance and social policy, as well as more theoretical issues of power, privacy, control and narrative. The NSF's report on data science education highlights that it is "critical to develop clear and useful policies with respect to how organizations, institutions, and projects deal with data (what data is

kept, who owns it and its byproducts, and who has access to the data or to parts of it)'‘ (Berman et al., 2016a, pp.17). Additionally, the House of Representatives recently passed the Evidence-Based Policymaking Commission Act of 2016, and the White House Office of Science and Technology Policy's National Science and Technology Council established a Data Science Interagency Working Group (Berman et al., 2016a). Discourse about the importance of data work is being institutionalized at the highest levels of American government, which illustrates that data arguments, ethics, and use are becoming a focus of policymakers. Technical communication practitioners already offer expertise to areas of policy making and advocate for citizens on this front. With a concerted effort to understand the intricacies of data practice, professional and technical communicators could insert themselves into the upcoming policy deluge that will deal with data. Currently, few laws and regulations exist that deal with the ethics around how corporate agents collect and deploy user data. As a result, there is a power disparity between those who provide data (often unknowingly) and the companies that buy, sell, and turn a profit on the virtual bodies of their users. In describing the system in which "data sets are centralized behind corporate proprietary or national security barriers and require specialized infrastructure, tools, and training,'‘ Boyd and Crawford note the creation of "a new kind of digital divide: the Big Data rich and the Big Data poor'‘ (Boyd & Crawford, 2012, pp.674). Producers of data, such as the average smartphone user, greatly outnumber those who have "access, expertise, or facilities to do the kind of data analysis done inside large institutional and corporate contexts'‘ (C. M. Dalton, 2018, pp.160). Promoting education in data literacy from a rhetorical standpoint fosters possibilities for data citizenship, where users are able to have domain over their own data and the power to leverage and gain insight from it. The data divide doesn't just relate to the data *haves* and *have nots*. There are also issues when data analysis is a skill only held by a few

privileged professionals. I bring up two specific instances here. The first is the data divide in grant writing and the second concerns who gets to ask the questions and make the arguments in research and development. In my experience working closely with local nonprofit and grassroots organizations, data arguments are key pieces to changing legislation and municipal policies, and to acquiring grants that fund such organizations. Professionals at locally based service organizations are often already low on staff and funding. If they do not have experience finding and working with large data sets that make a case for their programs, then staff have fewer opportunities than large organizations to win grants. The digital divide has real material consequences for nonprofits that do not always have the resources to hire others who do data work. The data divide in this case can leave a legacy of social inequity. When data analytics are positioned as the most valuable skill, especially when it comes to arguing for grant funding in the community, we have to take notice of the power differential in data literacy and ask "Who is advantaged and who is disadvantaged in such a context?" (Boyd & Crawford, 2012, pp.674). In research and development, the same holds true. Big data can be a powerful tool for understanding the world on a scale we have yet to fully understand. But when computational scientists are the ones with skills and access to data, then those without do not get to ask the same questions. In general, those from the Liberal Arts traditions are less likely to be data savvy, which means new hierarchies emerge. If you cannot get a seat at the table, then your perspectives are left out. Research that relies only on big data can easily leave out human-centered perspectives if these professionals do not have the skillsets needed to work with data; therefore, humanities perspectives are easily left out of research and development practices. STEM-educated researchers will ask different questions than rhetorical scholars. They will construct different conclusions. Therefore, the data divide has the potential to obscure many potential lines of work

and inquiry. Considering the lack of women and people of color statistically in computational and data sciences, the discussion goes beyond STEM versus humanities and ventures into the space where white men and their perspectives will continue to hold epistemological privilege. Because of the power imbalances present in big data work, D'Ignazio and Bhargava argue for a data literacy that empowers users from the ground up, rather than expecting power will be passed down from those at the top (Bhargava & D'Ignazio, 2015). They use Freirean frameworks of literacy and empowerment to consider how a data literacy can empower students, community members, and others who traditionally do not have access to it. They define data literacy as "the ability to read, work with, analyze and argue with data,'' though they note other popular definitions are built solely on statistical literacy or defined by putting data into action (pp. 1). D'Ignazio and Bhargava's definition is rhetorically-focused and takes into account the multivocal ways that data can be interpreted and put into action for social change.

6.5 Is Doing Good with Data Enough?

What should be done about the infrastructures of power-knowledge that operate in favor of the few over the many ? How do we conceptualize a more egalitarian, intersectional agenda for data rhetoric? One option that garners a lot of positive attention is doing good with data, that is, using the power of big data to press upon the world's problems. Several large and international organizations exist that work to give individuals and nonprofits access to large data sets. Some focus on supplying education and training, while others, like Datakind, organize volunteer data scientist and pair them with social organizations to help maximize data that organizations already collect. Datakind operates under the attitude that "data is a force for good'' and should be leveraged to help organizations fulfill their goals for social change

(DataKind, 2020 (accessed February 3, 2015)). Data Scientists Without Borders was an international organization modeled on Doctors Without Borders, where professionals volunteer to travel around the world to work on data solutions for organizations. One of their key initiatives was training refugees to become data scientists so they can support themselves. other projects include work on cholera outbreaks, economic crises, and youth unemployment in Bosnia (DataKind, 2020 (accessed February 3, 2015)). Here, data science is being applied to real world, pressing problems and using skilled volunteers to do it. Providing temporary volunteers may not address structural inequalities that are proliferated, but it does create temporary solutions for projects that are doing the work to decrease other structural inequalities. However, even in these well-meaning organizations, the same methodological and structural problems exist as do in industry. Analysts come into organizations from the outside, trying to solve complex problems through only the power of data. Many of the projects in volunteer organizations are for complex issues that need expert teams of people to address. For instance, one Data Scientists Without Borders was to "resolve Africa's education crisis'"—the goal being to increase attendance in secondary school. Even if we do not discuss the problem of scale—that one volunteer will solve all of Africa's education problems—or the problem of speaking about Africa widely when the project is located only in rural Nigeria—this is a project that requires expertise in education, Nigerian culture, politics and social inequity. The project website explains that the current problem organizations want help addressing are issues with retention, such as: ... insufficient motivation, interest in their study, indiscipline, low retention, and association with wrong peers, bad teachers among students as well as the effect of unmotivated teachers/access to correct information, quality educational programs and counseling services (DataKind, 2020 (accessed February 3, 2015)). Such a complex mix of issues calls for much more than a single data scientist and more than just

quantitative data to address. The moral of the story is this: data without situated knowledge and context is not a full solution, and well-meaning stop gaps for data inequities that do not address more systemic issues of power and marginalization are partial or temporary solutions at best. Further, volunteer projects like those described above can easily fall into data colonialism. The problems are situated in specific contexts. Therefore, the methods to solve these problems—''the products of research communities, economic actors, and educational practices that span the globe''—need to also be produced in situ (Philip, Irani, & Dourish, 2012, pp.6).

6.6 Future Research Directions

In addition to studying socially situated knowledge work within data analytics, there are several other areas of data work that could yield interesting research. Specifically, research dedicated to ways humans are implicated in high performance computing workflows would benefit computational research and practitioners. As scientists push computing architectures to exascale, there is a lot of uncertainty in how this will change computing paradigms. More specifically, the work currently being done to rewrite foundational application code for exascale architectures would benefit from the study of collaborative coding as a writing practice. Working from the outside of a field of expertise can lead to dangerous assumptions in addition to covering old ground. I argue that true interdisciplinary work has to be collaborative, which means decidedly declining to come in from outside a problem to enter the conversation, scholarly or otherwise. Instead, for scholars who aim to do work in data or computationally-focused fields, I encourage collaboration with data scientists and others who might benefit from the rhetorical standpoint of our work. Based on my work at the lab, my fact-finding missions at tech conferences, and my extensive reading in data visualization literature, I believe that there are fruitful avenues for technical

writing scholars to consider data provenance, which refers to the tools and methods used to track data work and decision-making, often via sophisticated interfaces that record practices tacitly in the background, akin to the ways eye tracking software is used in usability research. Visualization fields are working heavily with narrative, not only in terms of how to turn a number into a story, but how to create scientific visualization interfaces that are designed with narrative sensemaking in mind (Turton, Banesh, Overmyer, Sims, & Rogers, 2020). So much of the visualization literature that emphasizes how narrative can be put to work for data science forgoes any deep dive into scholarship on narrative, specifically scholarship from composition, rhetoric, literature, creative writing or narrative inquiry. I believe that scholars in rhetoric and technical writing would have a lot to add to these scholarly conversations, which could have a great impact on application design and visual analytics. Because the rise of big data has changed the epistemological process, a theoretical inquiry into the implications of new modes of knowledge production would push our understanding of the paradigm shift forward. Ong argued that the shift from oral cultures to literate cultures fundamentally changed how we compose our thoughts, solve problems and essentially how we interact socially. Though I do not contend that the availability of data has made the same kind of shift, I do think it is worth considering how the rise of data and computational thinking may be changing our relationship to our own ideas. Ong argues that in oral cultures, with no way to write notes as memory aids, the only way to proceed was to "think memorable thoughts'' (Ong, 2002, pp.32). Memory has taken on additional meanings since Ong's work, and now our everyday lives are saved, shared and backed up digitally. We have historical facts, current events and our own personal histories pressed behind glass and always accessible. Information is granular and modular. If, when we moved to writing culture, we no longer had to think thoughts that were quite as memorable or grand, then where might that leave us now?

Some Luddites will jump to grief over the loss. I offer a different idea: because we no longer *have to* think grand thoughts, we can now put our mental dexterity toward making small connections and putting old thoughts into new contexts, that we might cultivate richer understandings of knowledge making.

REFERENCES

REFERENCES

Agency, U. S. C. I. (2020). *The world factbook 2020.* Author.

Ahrens, J., Heitmann, K., Habib, S., Ankeny, L., McCormick, P., Inman, J., ... Ma, K.-L. (2006). Quantitative and comparative visualization applied to cosmological simulations. In *Journal of physics: Conference series* (Vol. 46, pp. 526–534).

Albers, M. J. (2017). Quantitative data analysis—in the graduate curriculum. *Journal of Technical Writing and Communication*, *47*(2), 215–233.

Allen, N. (1996). Ethics and visual rhetorics: Seeing's not believing anymore. *Technical communication quarterly*, *5*(1), 87–105.

Alsan, M., & Wanamaker, M. (2018). Tuskegee and the health of black men. *The quarterly journal of economics*, *133*(1), 407–455.

Andrejevic, M. (2014). Big data, big questions— the big data divide. *International Journal of Communication*, *8*, 17.

Anson, C. M. (2008). The intelligent design of writing programs: Reliance on belief or a future of evidence. *Writing Program Administration*, *32*(1-2), 11–37.

Argenta, C., Benson, J., Bos, N., Paletz, S. B., Pike, W., & Wilson, A. (2014). Sensemaking in big data environments. In *Proceedings of the 2014 workshop on human centered big data research* (pp. 53–55).

Armstrong, K., Ravenell, K. L., McMurphy, S., & Putt, M. (2007). Racial/ethnic differences in physician distrust in the united states. *American journal of public health*, *97*(7), 1283–1289.

Atkinson, P., Delamont, S., & Housley, W. (2008). *Contours of culture: Complex ethnography and the ethnography of complexity.* Rowman Altamira.

Bach, B., Wang, Z., Farinella, M., Murray-Rust, D., & Henry Riche, N. (2018). Design patterns for data comics. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–12).

Bal, M., & Marx-MacDonald, S. (2002). *Travelling concepts in the humanities: A rough guide.* University of Toronto Press.

Ball, C. E., Graban, T. S., & Sidler, M. (n.d.). The boutique is open: Data for writing studies. *Networked Humanities: Within and Without the University*.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written communication*, *2*(1), 3–23.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science* (Vol. 356). University of Wisconsin Press Madison.

Benham, H. (2018, May). *The history of the hackathon.* One iota. Retrieved from `https://medium.com/oneiota/the-history-of-the-hackathon-52551433af31`

Bennett, J. (2010). *Vibrant matter: A political ecology of things.* Duke University Press.

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., ... Stodden, V. (2016a). *Realizing the potential of data science: Final report from the national science foundation computer and information science and engineering advisory committee data science working group.* National Science Foundation Computer and Information Science and Engineering ....

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., ... Stodden, V. (2016b). *Realizing the potential of data science: Final report from the national science foundation computer and information science and engineering advisory committee data science working group.* National Science Foundation Computer and Information Science and Engineering ....

Beveridge, A. (2017). Writing through big data: New challenges and possibilities for data-driven arguments. *Composition Forum*, *37*.

Bhargava, R., & D'Ignazio, C. (2015). Designing tools and activities for data literacy learners. In *Workshop on data literacy, webscience.*

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, *18*(10).

Boonabaana, B. (2014). Negotiating gender and tourism work: Women's lived experiences in uganda. *Tourism and Hospitality Research*, *14*(1-2), 27–36.

Bourdieu, P. (1990). *In other words, trans. matthew adamson.* Polity Press, Cambridge.

Bowker, G., & Gitelman, L. (2013). *"raw data" is an oxymoron.* MIT Press.

Bowker, G., & Star, S. L. (1991). Situations vs. standards in long-term, wide-scale decision-making: The case of the international classification of diseases. In *Proceedings of the twenty-fourth annual hawaii international conference on system sciences* (Vol. 4, pp. 73–81).

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Brescoll, V. L. (2011). Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, *56*(4), 622–641.

Buchanan, R. (1992). Wicked problems in design thinking. *Design issues*, *8*(2), 5–21.

Cissna, K. N., & Anderson, R. (2008). Dialogic rhetoric, coauthorship, and moments of meeting. *Dialogue and rhetoric*, *2*, 39.

Colombini, C. B., & Hum, S. (2017). Integrating quantitative literacy into technical writing instruction. *Technical Communication Quarterly*, *26*(4), 379–394.

Comer, D. K., & White, E. M. (2016). Adventuring into mooc writing assessment: Challenges, results, and possibilities. *College Composition and Communication*, 318–359.

Council, N. R. (2015). *Preparing the workforce for digital curation*. National Academies Press.

Cox, M., & Ellsworth, D. (1997). Managing big data for scientific visualization. In *Acm siggraph* (Vol. 97, pp. 21–38).

Coyne, R. (2010). *The tuning of place: sociable spaces and pervasive digital media*. MIT press.

Cushman, J. (2014). Our unstable artistry: Donald schön's counterprofessional practice of problem setting. *Journal of Business and Technical Communication*, *28*(3), 327–351.

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? seven points for a critical approach to 'big data'. *Society and Space*, *29*.

Dalton, C. M. (2018). Big data from the ground up: Mobile maps and geographic knowledge. *The Professional Geographer*, *70*(1), 157–164.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. *Big Data & Society*, *3*(1), 2053951716648346.

Danner, P. (2020). Story/telling with data as distributed activity. *Technical Communication Quarterly*, *29*(2), 174–187.

DataKind. (2020 (accessed February 3, 2015)). *About us.* Retrieved from `https://www.datakind.org/about`

Derrida, J. (1996). *Archive fever: A freudian impression*. University of Chicago Press.

Dias, P., Freedman, A., Medway, P., & Par, A. (2013). *Worlds apart: Acting and writing in academic and workplace contexts*. Routledge.

Diebold, F. X. (2003). "big data'': Dynamic factor models for macroeconomic measurement and forecasting: A discussion of the papers by lucrezia reichlin and by mark w. watson. In M. Dewatripont, L. P. Hansen, & S. J. Turnovsky (Eds.), *Advances in economics and econometrics: Theory and applications, eighth world congress* (Vol. 3, p. 115–122). Cambridge University Press. doi: 10.1017/CBO9780511610264.005

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*(4), 241–256.

Dixon-Roman, E. (2016). Algo-ritmo: More-than-human performative acts and the racializing assemblages of algorithmic architectures. *Cultural Studies? Critical Methodologies*, *16*(5), 482–490.

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 2053951716665128.

Dourish, P., & Bell, G. (2011). *Divining a digital future: Mess and mythology in ubiquitous computing*. Mit Press.

Dragga, S. (1996). " is this ethical?": a survey of opinion on principles and practices of document design. *Technical communication*, *43*(3), 255–265.

Dragga, S., & Voss, D. (2001). Cruel pies: The inhumanity of technical illustrations. *Technical communication*, *48*(3), 265–274.

Druschke, C. G., Reynolds, N., Morton-Aiken, J., Lofgren, I. E., Karraker, N. E., & McWilliams, S. R. (2018). Better science through rhetoric: A new model and pilot program for training graduate student science writers. *Technical Communication Quarterly*, *27*(2), 175–190.

D'Ignazio, C., & Bhargava, R. (2015). Approaches to building big data literacy. In *Proceedings of the bloomberg data for good exchange conference.*

D'Ignazio, C., & Klein, L. F. (2016). Feminist data visualization. In *Workshop on visualization for the digital humanities (vis4dh), baltimore. ieee.*

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Mit Press.

Epstein, S. (2009). Beyond the standard human. In M. Lampland & S. Star (Eds.), *Standards and their stories: How quantifying, classifying, and formalizing practices shape everyday life* (pp. 35–53). Cornell University Press.

Feenberg, A. (2012). *Questioning technology*. Routledge.

Foucault, M. (1977). Discipline and punish. *Pantheon New York*.

Foucault, M. (1982). The subject and power. *Critical inquiry*, *8*(4), 777–795.

Frith, J. (2017). Big data, technical communication, and the smart city. *Journal of Business and Technical Communication*, *31*(2), 168–187.

Fulkerson, R. (2005). Composition at the turn of the twenty-first century. *College Composition and Communication*, 654–687.

Gaillet, L. L. (2010). Archival survival: Navigating historical research. *Working in the archives: Practical research methods for rhetoric and composition*, 28–39.

Gamble, V. N. (1997). Under the shadow of tuskegee: African americans and health care. *American journal of public health*, *87*(11), 1773–1778.

Gegeo, D. W., & Watson-Gegeo, K. A. (2001). " how we know": Kwara'ae rural villagers doing indigenous epistemology. *The contemporary pacific*, 55–88.

Gitelman, L. (2013). *Raw data is an oxymoron*. MIT press.

Glaser, B. G., Strauss, A. L., & Strutzel, E. (1968). The discovery of grounded theory; strategies for qualitative research. *Nursing research*, *17*(4), 364.

Goffman, E. (2002). The presentation of self in everyday life. 1959. *Garden City, NY*, *259*.

Goldsmith, S., & Crawford, S. (2014). *The responsive city: Engaging communities through data-smart governance*. John Wiley & Sons.

Grant, A. M. (2013). Rocking the boat but keeping it steady: The role of emotion regulation in employee voice. *Academy of Management Journal*, *56*(6), 1703–1723.

Hannah, M. A., & Arreguin, A. (2017). Cultivating conditions for access: A case for "case-making" in graduate student preparation for interdisciplinary research. *Journal of Technical Writing and Communication*, *47*(2), 172–193.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, *14*(3), 575–599.

Haswell, R. H. (2005). Ncte/cccc's recent war on scholarship. *Written Communication*, *22*(2), 198–223.

Hawk, B. (2004). Toward a post-*techne*-or, inventing pedagogies for professional writing. *Technical Communication Quarterly*, *13*(4), 371–392.

Hepworth, K. (2017). *Big data visualization: promises & pitfalls*. ACM New York, NY, USA.

Hinrichs, M. M., Seager, T. P., Tracy, S. J., & Hannah, M. A. (2017). Innovation in the knowledge age: implications for collaborative science. *Environment Systems and Decisions*, *37*(2), 144–155.

Hobart, E. M., & Schiffman, S. Z. (2000). Orality and the problem of memory. *Information ages: literacy, numeracy, and the computer revolution*, 11–31.

Holcomb, C., & Buell, D. A. (2018). First-year composition as "big data'": Towards examining student revisions at scale. *Computers and Composition*, *48*, 49–66.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174–196.

Hu, T.-H. (2015). *A prehistory of the cloud*. MIT press.

Hullman, J., & Diakopoulos, N. (2011). Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics*, *17*(12), 2231–2240.

Hullman, J., Drucker, S., Riche, N. H., Lee, B., Fisher, D., & Adar, E. (2013). A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics*, *19*(12), 2406–2415.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 2053951716674238.

Ingersoll, K. A. (2016). *Waves of knowing: A seascape epistemology*. Duke University Press.

Ingold, T. (2000). *The perception of the environment: essays on livelihood, dwelling and skill*. Psychology Press.

Ingold, T. (2001). From the transmission of representations to the education of attention. *The debated mind: Evolutionary psychology versus ethnography*, 113–153.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & society*, *1*(1), 2053951714528481.

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 data sets. *Big Data & Society*, *3*(1), 2053951716631130.

Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. *Computer*, *46*(5), 44–50.

Kwan, M.-P. (2016). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), 274–282.

Lang, S., & Baehr, C. (2012). Data mining: A hybrid methodology for complex and dynamic research. *College Composition and Communication*, 172–194.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, *36*(5), 700–710.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard university press.

Latour, B., & Woolgar, S. (1986). *Laboratory life the construction of scientific facts*. Princeton Univ. Press.

Lauer, J. M. (2004). *Invention in rhetoric and composition*. Parlor Press LLC.

Laursen, C. (2017, Feb). *What is a data sprint? an inquiry into data sprints in practice in copenhagen.* Ethos Lab. Retrieved from https://ethos.itu.dk/2017/02/15/caecilie-laursen/

Leff, M. (2000). Rhetoric and dialectic in the twenty-first century. *Argumentation*, *14*(3), 241–254.

Leurs, K. (2017). feminist data studies: using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review*, *115*(1), 130–154.

Lindlof, T. R., & Taylor, B. C. (2002). *Qualitative communication research methods*. Sage publications.

Lohr, S. (2013, Feb). *The origins of 'big data': An etymological detective story ...* New York Times. Retrieved from https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.

Lupton, D. (2015). The thirteen ps of big data. *This Sociological Life*.

Manoff, M. (2004). Theories of the archive from across the disciplines. *portal: Libraries and the Academy*, *4*(1), 9–25.

Manovich, L. (2001). *The language of new media*. MIT press.

Maternal and perinatal health. (2013, Oct). World Health Organization. Retrieved from https://www.who.int/maternal_child_adolescent/topics/maternal/maternal_perinatal/e

McDonald, C. (2015, May). *Is there a sexist data crisis?* BBC. Retrieved from www.bbc.com/news/magazine-36314061

McNely, B., Spinuzzi, C., & Teston, C. (2015). Contemporary research methodologies in technical communication. *Technical Communication Quarterly*, *24*(1), 1–13.

Meloncon, L., & Warner, E. (2017). Data visualizations: A literature review and opportunities for technical and professional communication. In *2017 ieee international professional communication conference (procomm)* (pp. 1–9).

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Moxley, J. (2013). Big data, learning analytics, and social assessment. *The Journal of Writing Assessment*, *6*(1), 1–10.

Munk, A. K., Meunier, A., & Venturini, T. (2019). Data sprints: A collaborative format in digital controversy mapping. *digitalSTS: A Field Guide for Science & Technology Studies*, 472.

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, *15*(6), 921–928.

National Academies of Sciences, Engineering, and Medicine. (2017). *Refining the concept of scientific inference when working with big data: proceedings of a workshop*. National Academies Press.

Nelson, A. (2002). Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association*, *94*(8), 666.

Odell, L., & Goswami, D. (1982). Writing in non-academic settings. *Research in the Teaching of English*, *16.3*, 201–223.

of Statistics, U. B. (2015, September). *Mandate, vision and mission.* Uganda. Retrieved from https://www.ubos.org/about-us/vision-and-mission/

Ong, W. J. (2002). *Orality and literacy*. Routledge.

Owens, T. (2011). Defining data for humanists: Text, artifact, information or evidence. *Journal of Digital Humanities*, *1*(1), 6–8.

Perelman, C., & Olbrechts-Tyteca, L. (1969). The new rhetoric: A treatise on argumentation. 1958. *Trans. John Wilkinson and Purcell Weaver. Notre Dame: U of Notre Dame P*.

Philip, K., Irani, L., & Dourish, P. (2012). Postcolonial computing: A tactical survey. *Science, Technology, & Human Values*, *37*(1), 3–29.

Pope-Ruark, R., Tham, J., Moses, J., & Conner, T. (2019). Introduction to special issue: Design-thinking approaches in technical and professional communication. *Journal of Business and Technical Communication*, *33*(4), 370–375.

Powell, K. M., & Takayoshi, P. (2012). Revealing methodology. *Practicing research in writing studies: Reflexive and ethically responsible research*, 1–30.

Purdue.edu. (2020). *Introducing the data mine learning communities.* Retrieved 2020-06-29, from `https://datamine.purdue.edu/about/`

Ramsey, A. E., Sharer, W. B., L'Eplattenier, B., & Mastrangelo, L. (2009). *Working in the archives: Practical research methods for rhetoric and composition*. SIU Press.

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. *Stan. L. Rev. Online*, *66*, 41.

Ridolfo, J., & Hart-Davidson, W. (2015). *Rhetoric and the digital humanities*. University of Chicago Press.

Salvo, M. J. (2012). Visual rhetoric and big data: Design of future communication. *Communication Design Quarterly Review*, *1*(1), 37–40.

Schiebinger, L. (1987). The history and philosophy of women in science: A review essay. *Signs: Journal of Women in Culture and Society*, *12*(2), 305–332.

Schön, D. (1938). The reflective practitioner. *New York*, *1083*.

Scott, J. B. (2003). Extending rhetorical-cultural analysis: Transformations of home hiv testing. *College English*, *65*(4), 349–367.

Scott, M. (2017). Big data and writing program retention assessment. *Retention, Persistence, and Writing Programs*, 56.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, *16*(6), 1139–1148.

Silver, N. (2012). *The signal and the noise: why so many predictions fail–but some don't*. Penguin.

Simmons, W. M., & Grabill, J. T. (2007). Toward a civic rhetoric for technologically and scientifically complex places: Invention, performance, and participation. *College Composition and Communication*, 419–448.

Sorapure, M. (2019). Text, image, data, interaction: Understanding information visualization. *Computers and Composition*, *54*, 102519.

Stanton, C., Abderrahim, N., & Hill, K. (1997). *Dhs maternal mortality indicators: an assessment of data quality and implications for data use* (No. 4). Demographic and Health Surveys, Macro International, Inc.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology,translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science*, *19*(3), 387–420.

Star, S. L., & Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work (CSCW)*, *8*(1-2), 9–30.

Strauss, A. (1985). Work and the division of labor. *Sociological quarterly*, *26*(1), 1–19.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge university press.

Strong, C. (2014). The challenge of "big data'": What does it mean for the qualitative research industry? *Qualitative Market Research: An International Journal*.

Suchman, L. (1995). Making work visible. *Communications of the ACM*, *38*(9), 56–64.

Suchman, L. (2002). Located accountabilities in technology production. *Scandinavian journal of information systems*, *14*(2), 7.

Suchman, L., & Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Sullivan, P. (2017). Beckon, encounter, experience: The danger of control and the promise of encounters in the study of user experience. *Rhetoric and experience architecture*, 17–40.

Sullivan, P. A. (1996). Ethnography and the problem of the'other.'. *Ethics and representation in qualitative studies of literacy*, 97–114.

Symons, J., & Alvarado, R. (2016). Can we trust big data? applying philosophy of science to software. *Big Data & Society*, *3*(2), 2053951716664747.

Szymanski, M. H., & Whalen, J. (2011). *Making work visible: Ethnographically grounded case studies of work practice*. Cambridge University Press.

Taylor, L. (2015). Towards a contextual and inclusive data studies: A response to dalton and thatcher. *Society and Space blog*.

Thatcher, J., Bergmann, L., Ricker, B., Rose-Redwood, R., O'Sullivan, D., Barnes, T. J., ... Cinnamon, J. (2016). Revisiting critical gis. *Environment and Planning A*, *48*(5), 815–824.

Thomas, J. J., & Cook, K. (2005). *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society.

The top 500 list. (2020). The Top 500. Retrieved from `https://www.top500.org/`

Turton, T. L., Banesh, D., Overmyer, T., Sims, B. H., & Rogers, D. H. (2020). Enabling domain expertise in scientific visualization with cinemascience. *IEEE Computer Graphics and Applications*, *40*(1), 90–98.

Vogel, S. C., Biwer, C. M., Rogers, D. H., Ahrens, J. P., Hackenberg, R. E., Onken, D., & Zhang, J. (2018). Interactive visualization of multi-data-set rietveld analyses using cinema: Debye-scherrer. *Journal of applied crystallography*, *51*(3), 943–951.

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Weick, K. E. (2010). Reflections on enacted sensemaking in the bhopal disaster. *Journal of Management Studies*, *47*(3), 537–550.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science*, *16*(4), 409–421.

Winsor, D. (1996). Writing well as a form of social knowledge. *Nonacademic writing: Social theory and technology*, 157–172.

Wolfe, J. (2009). How technical communication textbooks fail engineering students. *Technical Communication Quarterly*, *18*(4), 351–375.

Wolfe, J. (2015). Teaching students to focus on the data in data visualization. *Journal of Business and Technical Communication*, *29*(3), 344–359.

Yunis, H. (2011). *Plato: Phaedrus*. Cambridge University Press.