MAKING DATA WORK
THE HUMAN AND ORGANIZATIONAL LIFEWORLDS OF DATA SCIENCE PRACTICES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Information Science

by

Samir Passi

December 2021

ABSTRACT

MAKING DATA WORK:
THE HUMAN AND ORGANIZATIONAL
LIFEWORLDS OF DATA SCIENCE PRACTICES

Samir Passi, Ph.D.
Cornell University 2021

It takes a lot of human work to do data science, and this thesis explains what that work is, how it is done, why it is needed, and what its implications are.

Data science is the computational practice of analyzing data using methods from fields such as machine learning and artificial intelligence. Data science in practice is a difficult undertaking, requiring a deep understanding of algorithmic techniques, computational tools, and statistical methods. Such technical knowledge is often seen as the core, if not the sole, ingredient for doing data science. Unsurprisingly, technical work persists at the center of current data science discourse, practice, and imagination.

But doing data science is not the simple task of creating models by applying algorithms to data. Data science is a craft—it has procedures, methods, and tools, but also requires creativity, discretion, and improvisation. Data scientists are often aware of this fact but not all forms of their everyday work are equally visible in scholarly and public discussions and representations of data science. Official narratives tell stories about data attributes, algorithmic workings, model findings, and quantitative metrics. The disproportionate emphasis on technical work sidelines other work, particularly ongoing and messy forms of human and organizational work involving collaboration, negotiation, and experimentation.

In this thesis, I focus on the less understood and often unaccounted forms of human and organizational work involved in the everyday practice of data science through ethnographic and qualitative research. Moving away from an individualistic and homogeneous understanding of data science work, I highlight the collaborative and heterogeneous nature of real-world data science work, showing how human and organizational work shape not only the design and working of data science systems but also their wider social implications. This thesis reveals the everyday origins of the social, ethical, and normative issues of data science, and the many ways in which practitioners struggle to define and deal with them in practice, reminding us that not all data science problems are technical in nature—some are deeply human, while others innately organizational.

BIOGRAPHICAL SKETCH

Samir Passi's interdisciplinary research program engages with the unaccounted and less visible forms of human and organizational work involved in AI and data science practices. He uses ethnographic and qualitative research to map the deep connections between the ethical and social implications of AI and data science systems and the everyday decisions, struggles, and work involved in building them. Samir currently works as a Researcher at Microsoft, where he analyzes the sociotechnical aspects within the design and use of Responsible AI systems.

Samir's research draws on and contributes to scholarship in several fields such as critical data studies; computer-supported cooperative work and social computing (CSCW); fairness, accountability, and transparency (FAccT); human-computer interaction (HCI); organizational studies; and science and technology studies (STS). His research has been published in top-tier conferences such as CSCW and FAccT, and in reputed journals such as Big Data & Society.

In 2021, Samir completed his PhD in Information Science at Cornell University. At Cornell, he was an affiliate of the Department of Science & Technology Studies, and a member of the Culturally Embedded Computing (CEMCOM) group, the Artificial Intelligence, Policy, and Practice (AIPP) initiative, and the Social Computing lab. He received his research master's degree in Science and Technology Studies from Maastricht University in The Netherlands and his bachelor's degree in Information and Communication Technology from Dhirubhai Ambani Institute of Information and Communication Technology in India.

To glance,
is to run.
One needs to walk
like the gaze.

ACKNOWLEDGMENTS

Not all heroes wear capes—some wear scrubs. I want to thank the doctors and nurses at Cedars-Sinai's Marina del Rey emergency for swiftly diagnosing my life-threatening condition on March 2nd, 2021. Without your due diligence, I would not have received the correct treatment in time at Cedars-Sinai's Smidt Heart Institute. Dr. Wen Cheng and Dr. Mohamed Hassanein—thank you for keeping me alive and for giving me another chance at life. I am forever in your debt. Thanks to the nurses, physicians, and hospital staff who helped me recover. I still get emotional when I recall the care, love, and patience that all of you gave me during my hardest time.

This dissertation, and my growth as a researcher and as a person, could not have happened without the help of several amazing people. Four of which were on my committee: Phoebe Sengers, Steve Jackson, Solon Barocas, and David Mimno.

Phoebe and Steve—I consider myself incredibly lucky to have met you. Thank you for being my advisors, counsellors, friends, teachers, and mentors. Words cannot do justice to the impact you both have had on my life. I hope I make you proud.

Phoebe, I will dearly miss our weekly meetings—a ritual that lasted for eight years! You taught me how to ask thoughtful questions, how to get comfortable with the messy nature of research, how to speak across perspectives, and how to nurture the values of grace and kindness. Thank you for showing me how to approach and make sense of the world and its problems. You are an amazing scholar. Thank you for making me a capable researcher, an effective mentor, and a better person—you inspire me to be the best version of myself.

Steve, there is so much I must thank you for—you helped me find my voice as a researcher, taught me the craft of writing, introduced me to Pragmatism, showed me how to look at the larger picture while keeping a close eye, and—most notably—taught me the values of humility and perseverance. It was you who showed me how to gather and unite diverse ideas—to find symbiosis among worldviews that appear unrelated at best and inconsistent at worst. Thank you for all this and much more—I would not be who I am today if not for you.

Solon, thank you for enabling me to find and nurture my normative self and for showing me how to think about my work in new ways. Our conversations played a key role in helping me cross the schism between relativism and practicality. Most importantly, thank you for always looking out for me—you are a great mentor.

David, thank you for your constant support and for always indulging my ideas. As a PhD student, I often found myself floating in the sea of doubt. But you helped me to see the value of my work, realize its potential, and better understand its impact. I wish I could become half as good of a *reflective practitioner* as you are.

Special thanks to my field informants for their knowledge, patience, and time; they helped me see and experience the world of data science in new ways.

At Cornell, I had the privilege of learning from several remarkable faculty. Special thanks to Michael Lynch and Malte Ziewitz. Michael, my research would look vastly different if I had not met you. You showed me new ways with which to look at the world and taught me the ins and outs of ethnomethodology. Our conversations and your zest for knowledge always reminded me of the joy of research. Malte, you pushed me out of my comfort zone and made me think about things differently. If it were not for you, it would have taken me much longer

to figure out how to make the familiar strange. Thanks also to Dan Cosley, Jon Kleinberg, Karen Levy, David Robinson, Susan Fussell, Trevor Pinch, and Ronald Klein for their compassion and wisdom over the years.

Thanks to the fellow members of the Culturally Embedded Computing (CEMCOM) research group, the Artificial Intelligence, Policy, and Practice (AIPP) initiative, the Infrastructure Studies group, and the Social Computing lab for the engaging discussions and splendid company. Thanks to Eric Baumer for several generative conversations. Thanks to Saleema Amershi for believing in me and for helping me navigate the world of industry research. Thanks to Jofish Kaye and Shay David for helping me sharpen my approach to gaining corporate access.

A PhD is a long journey, made pleasant by the friends we make along the way. At Cornell, I was lucky to cross paths with some of the most kind-hearted and talented individuals. Jessica Beth Polk, Stephanie Steinhardt, Brian McInnis, Vincent Tseng, Palashi Vaghela, Vera Khovanskaya, Upol Ehsan, Leo Kang, Saeed Abdullah, Jean Costa, Elizabeth Murnane, Stephanie Santoso, Ishtiaque Ahmed, Shion Guha, Dongwook Yoon, and Huaishu Peng—thank you for being a part of my journey. From dinners at Center Street and drinks at ICC to discussions in labs and hallways, our time together was special.

Thanks to the superb administrative staff of the IS department who ensured that I never needed to worry about the bureaucracy and logistics of graduate school. Eileen Grabosky, Barbara Woske, Terry Horgan, Ani Mercincavage, Julie Pagliaro, Lou DiPietro, Janeen Orr, Christine Stenglein, and Saba Alemayehu—thanks for taking care of everything.

Thanks to my funding providers for their generous support: National Science Foundation (CHS 1526155), Harvard-MIT Ethics and Governance of AI Initiative, Intel Science and Technology Center for Social Computing, and Cornell Information Science.

My journey into research started in 2008 when, as an undergraduate, I stumbled into the world of philosophy and social sciences. Shiv Visvanathan taught me how to reason like a social scientist, initiated me into STS, and got me to Maastricht University. Madhumita Mazumdar motivated me to think deeply about my analytic choices and lenses, helped me understand the value of carefully assessing scholarly ideas, and taught me how to craft constructive arguments and graceful critiques. Harmony Siganporia taught me how to put ideas into action and showed me that it is possible to foster compassion through research.

Outside formal institutions and spaces, I owe much to the first 'research lab' (if you can call it that) that I was a member of during my undergraduate days: the *Galla*. At face value, the Galla consisted of simply a bunch of street-side chai, soda, and food stalls outside college. But if you looked deeper, you would realize that the Galla was where several of us came together to give flight to ideas, space to curiosity, and shape to dreams. It was where I discovered, fostered, and felt comfortable with the contrarian in me. Utkarsh *'nUTs'* Srivastava, Paras *'ViRuS'* Mani, Sandeep *'Sandy'* Avula, Rohin *'Dark'* Reddy, Ishpal *'Paale'* Tuteja, Shitij *'Raivibo'* Kumar, Vaibhav *'Nasa'* Temani, Anurag *'Sahara'* Gupta, and others—thank you for creating a space where we all could find and be ourselves.

I left India eleven years ago as an engineer dabbling in concepts he did not yet fully understand. Two years later, I graduated from Maastricht University as an STS researcher—a transformation made possible by the incredible people at the Faculty of Arts & Social Sciences.

Wiebe Bijker inspired me to deeply engage with technical practices, to do interdisciplinary research, and to learn how to translate between different forms of knowledge. Jan de Roder showed me how to approach large projects and how to think freely. Anique Hommels helped me appreciate the value of research methods. Navin Goel, Ines Hülsmann, Chris Hesselbein, Charanya Chidambaram, and Rishabh Chawla made Maastricht even more special—I miss our conversations.

Thanks to Wiebe, I got to work with Harry Collins and Robert Evans—or, as I like to call them, the dynamic duo. Thank you both for helping me find the engineer in me, for showing me a whole new side of STS, and for teaching me how to marry my STS and engineering personalities. To this day, when confronted with a tricky problem, I often find myself saying 'what would Harry and Rob do?'

Special thanks to Sally Wyatt who, among several other things, taught me how to write (and get!) grants, gave me my first research job, showed me how to do STS on the ground, and has been a source of inspiration and knowledge. Sally, I will never forget that it was you who gave me my first big break—you believed in me before anyone else did, and for that I shall always remain eternally grateful. Thank you for everything. I hope I make you proud.

Thanks to my parents—Manisha Passi and Subhash Passi—for their love and patience, for standing by me, and for letting me chase my dreams. Thanks to my sister and brother-in-law—Shweta Rai and Dinesh Rai—for supporting me through thick and thin. I do not know what I would do without you all.

I am grateful to Ritwick Ghosh, Ann Bybee Finley, and Liz Clark for their friendship, love, and support. Your kindness towards everyone, passion for life, and dedication towards making the world a better place continue to inspire me. You all are nothing less than family.

None of this would be possible without Ranjit Singh, who is an extraordinary person with a heart of gold and a solid laugh. Ranjit calls me the *Titu*—the devil on his shoulder. The allegation is true. I often slip ideologies into his drinks and pour hypotheticals over his beliefs—but, in my defense, I act in good faith. Jokes aside, I consider myself extremely lucky to have met and spent considerable time with Ranjit. I have much to thank you for, Ranjit—and not enough time to ever repay you for your friendship. I would have left STS if not for you. I would not be at Cornell if not for you. I might have taken much longer to find my voice as a scholar if not for you. You are nothing less than a brother to me. Thank you for being at my side. If you ever need anything, all you must do is ask.

Midway through my PhD, I met the love of my life—Priya Gupta—at Cornell. Priya, I cannot thank you enough for your love and companionship. From patiently waiting for me to do things at my own pace to cheering me on during dark times, you have been my pillar of strength. You help me to keep the child in me alive—thank you for letting me be a part of your life. Now, if only you would spend a bit more time discussing my research…

Last but not the least, thanks to my cat Noodle for his company, headbutts, and love. He comforted me when no one else could. I will never forget that.

TABLE OF CONTENTS

I
Introduction

It takes a lot of human work to do data science, and this thesis explains what that work is, how it is done, why it is needed, and what its implications are.

Data science is the computational practice of analyzing large-scale data using techniques from fields such as artificial intelligence (AI), machine learning (ML), and natural language processing (NLP). Data science in practice is a difficult undertaking and requires a thorough understanding of algorithmic techniques, computational tools, and statistical methods. Such technical knowledge is often seen as the core, if not the sole, ingredient for doing data science. Technical work continues to persist at the center of data science discourse, practice, and imagination, making a data scientist's main work come across as applying algorithms to data to build models.

The emphasis on algorithms, data, models, and numbers in scholarly accounts of data science further amplifies such an understanding. Dominant narratives of data science work in research papers and presentations, for instance, usually go like this: *here is a question, this is the data, here is the algorithm, this is the model, here is the test strategy, these are the numbers, and here is the answer.* Such narratives represent the practice of data science as a linear journey from start (research question; high-level goal) to finish (statistical finding; software), painting a rather neat picture of data science work.

However, doing data science is not the simple task of creating models by applying algorithms to data. For instance, real-world problems rarely appear ready-formed as data science problems; they must be translated into data-driven problems. The world seldom fits neatly into rows and columns; it takes work to clean, complete, and turn the world's messy realities into sanitized datasets. Multiple algorithms can often address a given problem, and data scientists frequently try different algorithms, building many models before settling on one. Even model testing involves making choices between different, often divergent, performance criteria (e.g., *'do we want a model that identifies some of the correct answers and gives almost-no wrong answers or a model that identifies most of the correct answers but also gives a few wrong answers?')* and thresholds (e.g., *'do we want 75%, 80%, or 100% accuracy?'*).

None of this is surprising, at least not to data scientists. The ability to navigate uncertain and uncharted dataworlds is, after all, what separates a novice from an expert data scientist. Indeed, doing data science requires both theoretical and practical knowledge: know-*what*, but also know-*how*. Yet much of this know-how—discretionary choices and situated decisions made by data scientists—is left out in official narratives that generally contain accounts written from the point of view of technical artifacts alone: stories about data attributes, algorithmic workings, model findings, and quantitative metrics. The work accomplished *using* data, algorithms, models, and numbers come across as the work performed *by* data, algorithms, models, and numbers—a sentiment evident in the everyday use of phrases such as 'the data speaks for itself,' 'numbers do not lie,' and, my favorite, 'the model figured it out.'

Besides facilitating a technology-centered understanding of data science, such narratives also shape public data science discussions and debates. News reports tell heroic, and

at times horrifying, tales of the (wrong)doings of data science applications. Policy discussions, for instance those dealing with the responsible governance of data science applications, frequently center on technological progress as the main way forward. Governments and corporations continue to pour resources into collecting more data, building robust models, and developing better tests. The overvaluation of technical knowledge, work, and progress in and around the data science field makes it seem like nothing else matters.

This, however, is just one half of the story—a technical fable of computational possibility and machine work with faceless algorithms, unbiased models, and objective numbers as protagonists. In this thesis, I narrate the other half of the story—a sociotechnical tale of the human and organizational work involved in data science's everyday practice to highlight the choices and decisions that are crucial to the technical work of data science but often remain unaccounted for in data science learning and research and invisible in public debates and discussions.

I dive deeper into the messy world of data science, paying close attention to the everyday work of data science practitioners. Yes, practitioner-*s*. Plural. Data science, often from academic or research perspectives, is imagined as the sole work of data scientists interacting around reliable, standardized, and widely shared tools and conventions—a 'clean room' imagination of data and its relationship to the world. Although central to data science, data scientists are but one group involved in its practice, especially in *applied corporate data science practices*.

Much of this thesis focuses on applied data science practices in corporate organizations in which the goal is to use data science to solve business problems. The world of corporate data

science has data scientists, but also managers, analysts, designers, engineers, and executives. A relative newcomer to the business world, data science must rely on and support prevailing corporate practices such as business analytics, product design, software engineering, and project management that have longer histories and greater clout in organizational decision making.

As you will learn in this thesis, data scientists often neither have the only nor the final word in corporate projects. Moving away from an individualistic and homogeneous understanding of data science work (scientists and engineers on desks with computers), I highlight the collaborative and heterogeneous nature of corporate data science work (multiple practitioners working together on business problems with different assumptions, goals, and expectations). I explain *how* and *why* corporate data science work relies as much on individual discretion, business goals, constant negotiations, and organizational culture as on algorithms, data, models, and numbers.

1.1 The importance of human work

It is crucial to tell human-centered stories of data science work for several reasons—some of which I only slowly grasped over the years, all of which will hopefully become obvious as you read this thesis, and one of which is central to my research and thus deserves an up-front mention:

> Not accounting for human work paints a partial and misleading picture of the everyday practice of data science, making algorithmic results appear as unproblematic, objective facts about our world. A human-centered account sets the record straight, clarifying how and why data science results are not waiting-to-be-discovered neutral facts but normative *arti*-facts of a necessarily discretionary meaning-making process.

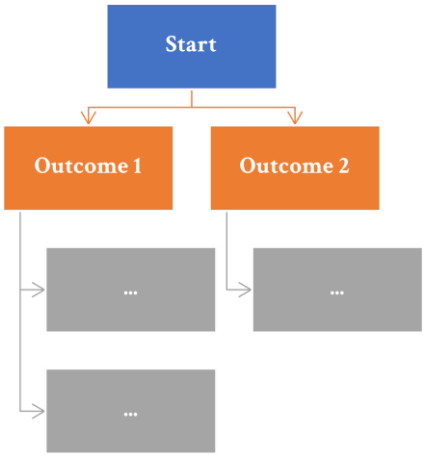

Figure 1: A sample flowchart

I stumbled upon a simpler version of this reason at the start of my doctoral research. Back then, data science seemed like an alien science to me—a way of knowing that at once implicated and remained elusive to most. Data scientists analyzed data far too large for manual analysis using methods that were often opaque even to experts. Owing to my CS background, I had a general idea about data science but little knowledge of what it entailed in practice. My research started with, what seemed at the time like, a simple question: *what is the work of a data scientist?*

I started reading a few highly-cited AI, ML, and NLP papers searching for an answer—pages upon pages of data science practitioners describing what work they had done and how. I also skimmed a bunch of well-known ML textbooks to see how they described data science work, including how you should or should not do it. Lastly, I also checked out the syllabi of a few online data science courses on Coursera and watched a handful of online data science tutorials on YouTube.

My efforts bore a futile but instructive payoff. Futile because, at the time, I could barely grasp the complex algorithms, dense math, and (literally) convoluted models described in textbooks, papers, and tutorials. Instructive because I learned that most of these depictions of data science work focused only on algorithms, data, models, and numbers—as if nothing else mattered.

Or that is what it seemed until I asked a few fellow doctoral students who did data science about their daily work; my conversations with them were remarkably helpful. I learned that doing data science requires making several non-obvious choices, and people who do data science are aware of the existence and import of such discretionary work. Here is a small excerpt from my conversation with a colleague (she later took a CS faculty position in an American university):

> "It is phenomenal how many times we act like there is not even a choice when [doing data science]. You are spending 80% of your time doing 'x' where 'x' represents a huge decision tree, and you just maybe state a couple of parts of that tree. *State!* You do not even justify. You just say it." (Interview, June 23, 2015, emphasis added)

A kind of flowchart, a decision tree denotes a top-to-down pathway from a starting point to different outcomes (Fig. 1). At the top of the tree is a single node that splits into other nodes. Each node indicates a different outcome, branching into more nodes. Each level contains many nodes; to move between levels you must choose among nodes to pick a single outcome. Every choice matters because selecting one node over another makes other choices and outcomes—i.e., other parts of the tree—inaccessible.

My colleague emphasized the importance of situated judgment and reasoning in data science practice by making an analogy between doing data science and traversing a decision tree. I must admit that her words, while instructive, left me confused. Data science in practice entails making several discretionary choices. People who do data science are aware of this; yet, when they document, describe, or present their work, they rarely mention such forms of quintessentially human work. Why?

I reckoned that perhaps an apparent reason for this was practicality. Variations between doing and describing work—routine occurrences in professional practices—are partly because of convenience.[1] Researchers, including myself, often present their work in a straightforward manner—dry and linear tales of what they did, how, and why. This thesis does some of that. Simplifying formal accounts of professional practice is not problematic. But how practitioners simplify—how they render specific work (in)visible—has deeper consequences, revealing how practitioners unevenly ascribe values to different kinds of work. Data science was no different. Data scientists place enormous value on some work (e.g., model building, statistical analysis), rendering them highly visible. In contrast, other kinds of work are not viewed as core parts of data science practices and rendered invisible (e.g., problem formulation, contingent fixes). The constant overappraisal of some (often technical) work bolsters the perception of those work as the main—if not the only—work of data science.

The implications of such differential (in)visibilities of data science work are, in fact, much more substantial. When data science is seen mainly as a domain of technical practice, the issues associated with it also appear technical in nature, making the work of fixing them come across as the 'technical' work of collecting more data, creating better models, or inventing robust evaluation strategies. This is evident in that present-day researchers attempt to address data science issues of accountability, bias, explainability, fairness, and trust chiefly through technical fixes. Technical solutions are essential, but data science's normative roots—as I argue

[1] There are other reasons why professionals may portray their work in certain ways—ways driven less by a desire to stay close to reality but more by, for instance, the need to demarcate certain, often expert-driven, forms of work as exclusive parts of their professional jurisdiction. For further details, see: Gieryn (1983) and Abbott (2014).

in this thesis—run deeper into its human foundation. Social implications of data science, as I show, are as much the artifacts of human discretion as of the technical properties of its systems. Making visible the consequential impact of human and organizational work on data science practice, I argue that technical work alone cannot—indeed, *must not*—exhaust the problem and solution space when we envision ways to address data science issues. Ethics of data science have *human* and *organizational*, and not just technical, roots.

What thus began as a mere curiosity of a first-year doctoral student over time turned into the research question underlying this thesis and my broader research program: *what is the nature, role, and implications of the human and organizational work involved in the everyday practice of data science?*

This thesis thus has three goals: (1) to describe the everyday practice of data science through stories from the oft-invisible world of corporate organizations where aspirations, data, goals, models, people, resources, and numbers continuously move between projects, quarters, rooms, and teams; (2) to explain how and why the successful practice of data science requires human and organizational, and not just technical, work; and (3) to highlight how human and organizational work shape the wider normative and social implications of data science systems.

The rest of the Introduction has four sections. **Section two**—*Sociotechnical scholarship on data science work*—focuses on the theoretical foundation of the thesis, key tropes in current research on data science work, and the scholarly relevance of studying the human and

organizational work of data science.[2] **Section three**—*A word on human, organizational, and technical work*—provides working descriptions of the three main types of data science work that I identify and examine in the thesis. **Section four**—*Research sites, data, and method*—describes my empirical fieldsites, qualitative data, and research methodology. **Section five**—*Chapters: A preview*—outlines the thesis's structure with short sketches of each chapter.

1.2 Sociotechnical scholarship on data science work

There has been a steady surge in sociotechnical research on data science practices and implications over the last few years in emerging fields such as *critical data studies* (boyd & Crawford, 2012; Iliadis & Russo, 2016), *fairness, accountability, transparency, and ethics* (Barocas & Selbst, 2016; Selbst et al., 2019), *human-centered data science* (Aragon et al., 2016; Baumer, 2017), and *human-AI interaction and collaboration* (Amershi, Weld, et al., 2019; Riedl, 2019).[3] Researchers have demonstrated, for instance, how and why data science systems implicate thorny issues such as those of bias and fairness (Buolamwini & Gebru, 2018; Selbst et al., 2019; Veale et al., 2018), disparate impacts (Barocas & Selbst, 2016; Rudin, 2018), epistemology (Kitchin, 2014a, 2014b; Leonelli, 2014), environmental costs (Hogan, 2015; Strubell et al., 2019), filter bubbles and misinformation (Bozdag, 2013; Gillespie, 2011; Seaver, 2012; Suwajanakorn, 2018), privacy (Barocas & Levy, 2020), stereotyping (Gillespie, 2016), and surveillance (Brayne, 2017). The list goes on.

[2] In-depth engagement with literature relevant to specific forms of data science work will happen in individual chapters.

[3] Sociotechnical scholarship on living and working with algorithms, data, and numbers more broadly has a much longer history in fields such as science and technology studies and organization studies. Works in these areas relevant to specific themes concerning the everyday work of data science will be introduced in individual chapters.

1.2.1 The many humans of data science

Scattered across these different strands of research are the countless humans of data science: those who use data science systems (Geiger, 2017), are affected by them (Rudin, 2019), champion the use of data science (Domingos, 2015), fund data science projects (DARPA, 2019), prepare datasets (Miceli et al., 2020), govern data science systems (Ziewitz, 2016), study data scientists[4] (Muller, Feinberg, et al., 2019), and those involved in the practices of designing and developing data science systems. I focus on this critical last group of *data science practitioners* who *do* different forms of data science work as part of their everyday job.

Data science work: What do we already know?

Over the past few years, researchers have worked to unpack the sociotechnical dimensions of data science practices. Such strands of research focus on aspects such as: how data scientists make sense of algorithmic results (Paine & Lee, 2017; Passi & Sengers, 2016; Ren et al., 2017), necessity of human discretion for algorithmic analysis (Diesner, 2015; Kleinberg et al., 2017; Zhang et al., 2019), similarities and differences between data science and other forms of discretionary professional work (Bolin & Schwarz, 2015; Muller et al., 2016; Thoreau, 2016), impacts of certain features of large-scale datasets on the knowledge produced using them (Busch, 2014; Domingos, 2012; Leonelli, 2014), broader consequences of seemingly technical data science problems (Dhar, 2013; McFarland & McFarland, 2015), transparency and intelligibility issues in data science practices (Ananny & Crawford, 2016; Burrell, 2016;

[4] See Forsythe (2001) for an early ethnographic study of the culture of AI research.

Neyland, 2016b; Selbst et al., 2019), and the importance of humanistic explanations of models and results (Bates et al., 2016; Riedl, 2019; Vaughan & Wallach, 2020).

Much of this research, however, has focused on the analysis of data science work in academic and research settings.[5] A small but growing body of research, however, has recently begun analyzing corporate data science work. Calling attention to the greater scale and complexity of applied data science work in corporate settings (relative to research contexts), these researchers focus on individual practices and steps in corporate data science workflows with the goal of finding ways to address the specific needs of industry practitioners (mainly data scientists).[6] Examples of such work include, but are not limited to, helping data scientists debug and build better models (Amershi et al., 2015), perform faster and more accurate model assessments (Ren et al., 2017), easily interpret complex models and results (Kahng et al., 2018), compare data features in large-scale datasets (Hohman et al., 2020), and effectively perform error discovery using better data exploration strategies (Suh et al., 2019).

Scale and complexity, however, as I also described earlier, are not the only distinguishing features of corporate data science work (when compared to similar work in academic and research settings). A very recent thread of qualitative research has thus started analyzing other defining features of corporate data science practices (though mostly with the same eventual goal of building tools and mechanisms to support such features as they intersect

[5] I describe the consequences of this limited focus in the next subsection.

[6] A common goal across such strands of research is the development of interactive tools to provide detailed insights on specific forms of data science work, opening new "space[s] for richer interactions and more context-aware support tailored to a practitioner's specific workflow" (Kery et al., 2020: 149). Some of these tools in fact provide holistic views of data science work itself by, for instance, helping data scientists to inspect the history—i.e., previous versions—of their datasets and models (Kery et al., 2019).

with the daily work of data scientists). Collaboration, for instance, is central to all kinds of organizational work. Corporate data science is no different. Recognizing this fact, Koesten et al. (2019) analyze twenty tools used for data work, making visible how industry practitioners collaborate around data and how different tools support such collaborative work. Finding the current tools lacking, they highlight the need for developing better tool support for data access, versioning, and documentation. But tools, as this thesis argues, can only go so far in addressing the collaborative foundations of corporate data science practices. As you will learn in this thesis, collaboration work is not always tool-based, often frustrated by, for instance, the inability to translate between data science and business knowledge or the very organizational culture and structure of an organization.

Another key facet of corporate data science work is how such work intersects with other existing corporate practices. Amershi, Begel, et al. (2019) zoom in on this area to study how AI development differs from traditional software development. Analyzing the workflow of Microsoft software teams building AI-based systems, they show how AI models are entangled in more complex technical ways compared to the relations between conventional software components. Best practices for software development do not necessarily port to AI development—we need to better understand the unique challenges within corporate AI practices, including how even known issues might arise in different and unexpected ways. In fact, the roots of such issues, as I show in the thesis, are not always found in technical differences between distinct industry practices. As you will learn, tight project timelines, organizational structure, business goals, and even other existing non-AI technologies also shape

how industry practitioners envision, design, and build AI systems (including how and to what extent they expect their systems to work in specific ways).

Data science work: What do we still not know?

Although the abovementioned strands of research have successfully highlighted specific attributes of data science work in general, several crucial aspects of real-world data science practices remain understudied. Rather than being results of simple oversights, such gaps in knowledge, I argue, exist for two key reasons.

First, much of existing qualitative research on data science work focuses on the study of such work in academic and research settings (e.g., Baumer, 2017; Burrell, 2016; Coletta & Kitchin, 2017; Hansen, 2020; Kay et al., 2015; Kocielnik et al., 2019; Neff et al., 2017; Paine & Lee, 2017; Paine & Ramakrishnan, 2019; Pink et al., 2017; Riberi et al., 2020). Faced by limits of access, confidentiality, and non-disclosure, the work of the heterogeneous set of practitioners—scientists, but also analysts, managers, designers, and executives—involved in corporate projects stays mostly invisible and unexplored.

This has created a "blind spot" (Crawford & Calo, 2016) in our research on the implications and work of real-world data science (including that of the many data science systems in use internally in companies). Researchers thus remain limited in effectively engaging with data science design, research, and practice without an in-depth understanding of corporate data science practices. This thesis is one of the first to address this limitation, joining a small but growing body of qualitative research on corporate data science work.

As described earlier, however, much of the existing qualitative research on corporate data science practice focuses on what industry practitioners *need* instead of how organizational and business aspects *shape* data science's applied practice in corporate settings. As a result, our understanding of the nature of corporate data science practices—including the situated and ongoing forms of work involved in them— is partial at best (and wrong at worst). In this thesis, I address this limitation by explicitly focusing on the nature and implications of the organizational dimensions of corporate data science work.

Second, existing qualitative research (on corporate data science specifically) is shaped as much by its methodology as by its focus. Most of the existing research on corporate data science uses interviews and user studies (often only with data scientists) as primary research methods. Such methods are useful for analyzing how data scientists approach their work but cannot address the unspoken and taken-for-granted aspects of their everyday work (Christin, 2020; Collins, 2010; Elish & boyd, 2018; Forsythe, 1993a; Seaver, 2017), let alone of the work done by other practitioners in corporate projects.

This is where ethnographic research—the primary research method underlying this thesis—is particularly useful. In addition to highlighting the nature of everyday work involved in corporate projects, ethnographic research—particularly participant observation that entails becoming an active member of a group to study the practices, work, and culture of its members—can help to capture the impact of aspects such as company culture, team dynamics, and group interactions on corporate data science work. Indeed, the need for and import of ethnographies are now gradually becoming evident to qualitative and quantitative researchers alike—evident in their calls for more ethnographic research on data science practices (e.g.,

Christin, 2020; Dourish & Cruz, 2018; Elish & boyd, 2018; Kitchin, 2016; Neyland, 2016b; Seaver, 2017; Ziewitz, 2017).

> "For the technological advancements to endure, it is imperative to ground both the practice and rhetoric of AI and Big Data. Doing so requires developing the methodological frameworks to reflexively account for the strengths and weaknesses of both the technical practices and the claims that can be produced through […data science] systems." (Elish & boyd, 2018, p. 74)

Ethnography is an effective means for such grounding work (Christin, 2020; Forsythe, 1993b, 2001). It makes it possible to identify and investigate the ongoing forms of everyday work done by data science practitioners and makes it possible to capture the development trajectories of data science systems. The latter affordance is crucial. A core challenge hindering research on data science harms is the difficulty of determining what went wrong, when, how, and why without access to a system's development history (Friedman & Nissenbaum, 1996)—*why was the system built this way?* Without detailed accounts of the nature of everyday work involved in system development, it is hard to detect and tackle issues arising from, for instance, prior assumptions, practical constraints, business imperatives, and emergent contexts of use (Hildebrandt, 2011).

Data science is a craft (L. A. Suchman & Trigg, 1993)—it has procedures, methods, and tools, but also requires creativity, discretion, and improvisation. While data scientists may be, and often are, aware of the human work at play in their practice (Barocas & boyd, 2017; Domingos, 2012; Luca et al., 2016), not all forms of such work are equally visible. The disproportionate emphasis on some kinds of technical work tends to slowly sideline other work, particularly ongoing and messy forms of human and organizational work involving collaboration, negotiation, and experimentation.

In this thesis, I focus on the less understood, seldom articulated, and often unaccounted forms of everyday work involved in corporate data science practices through ethnographic and qualitative research. In doing so, I show how human and organizational work can and do impact practitioners' understanding of technical artifacts such as data, models, and numbers, consequentially shaping the eventual design and working of data science systems.

1.2.2 (In)visible work

My focus on the *differential visibilities* of data science work draws from and builds on Star and Strauss' (1999) notion of "invisible work." They argue:

> "No work is inherently either visible or invisible. We always 'see' work through a selection of indicators: straining muscles, finished artifact, [and] a changed state of affairs. The indicators change with context, and that context becomes a negotiation about the relationship between visible and invisible work." (p. 9)

Although within data science we may not encounter strained muscles, everyday data science work—e.g., formulating data-driven problems, processing data, and interpreting model results—provide meaningful contexts for engaging with the range and implications of the various indicators through which different kinds of work are rendered less- and more-visible.

This thesis thus follows a long tradition of scholarship in Computer-Supported Cooperative Work (CSCW) in using the ethnographic analysis of work to understand and improve the fit between information technology and on-the-ground work practices (see Schmidt, 2011 for an overview). Key insights from this research area include: (1) even the most mundane work in practice is deeply context-dependent and rarely hews precisely to formal understandings and rules; (2) attempts to enforce seemingly logical and rational approaches to work (e.g., through structured technology support) often end disastrously; (3) this happens

because structured approaches to work often tend to make difficult or impossible the delicate, situated, and discretionary decision making and adaptation that make abstract procedures and rules work in practice; and, (4) the fit between people and information technology systems can still be improved by better understanding and addressing the partially articulable nature of work (Ackerman, 2000; Bannon, 1995; Dourish, 2001; L. A. Suchman, 2007). This interest in CSCW to acknowledge and support the less understood, poorly articulated, often invisible, and possibly non-formalizable forms of work is a suitable anchor for my focus on the human underpinnings of real-world data science.

1.2.3 Critical technical practice

My interest in using social science research to inform professional work also draws from Agre's (1997b, 1997a) work on *critical technical practice*. Inspired by qualitative social science research (e.g., L. A. Suchman, 1987), Agre explored what it would mean to construct artificially intelligent agents that support reactive, intelligent behavior instead of formal, rational thought. His subsequent work on behavior-based agent architecture transformed AI research (Agre & Chapman, 1990, e.g., 1987). In reflecting on this work, Agre argued that technical work can get caught in conceptual dead-ends caused by how we frame the problem we try to solve. The underlying cause of these dead ends is unclear because the way we approach the problem seems so natural that we do not realize any alternatives.

Agre argues for using social science research to effect change by identifying and modifying the conceptual assumptions underlying technical work. Sengers has further developed these concepts in AI (Sengers, 1999, 1998) and Human-Computer Interaction (HCI) (Boehner et al., 2007, 2005; Sengers et al., 2005) to change how practitioners conceive the

problems technologies solve, in turn facilitating new relationships between people and machines through the design of alternative technologies inspired by such a reconceptualization.

In this thesis, I build on this line of research to better integrate critical social science research and data science practices in two specific ways. *First,* I describe the human and organizational work involved in everyday data science practices to show how and why the dominant understanding of data science work as mainly, if not exclusively, technical is not just incorrect but also harmful. *Second,* by doing so, I show how the normative issues arising from data science systems have human and organizational, and not just technical, roots. Current research efforts to address data science's normative implications focus mainly on technical fixes. Broadening our understanding of the nature of everyday data science work, this thesis enables and fosters alternate understandings of data science work and implications, opening new ways to imagine solutions to data science problems.

1.2.4 From who to where: academic and corporate contexts

Where data science is done is as important as who does it since the context defines the nature of its work practices and the normative values at stake (Almklov, 2008; Crawford, 2015; Haraway, 1988; Helgesson, 2010; Leonelli, 2009; Neyland, 2016a; Saltz & Shamshurin, 2015; Selbst et al., 2019; Taylor et al., 2015), including why certain programs of action are devised and how they are executed in practice (L. A. Suchman, 2007). This thesis focuses on data science work in two empirical settings: academic learning environments and corporate organizations.

Academic learning environments

When we think of where data science happens, classrooms and workshops do not generally come to mind. But these are the places where individuals—would-be, future data science practitioners—often do data science for the first time. Learning is doing, with the goal of attaining expertise.

Learning a practice is, of course, a lifelong affair, not limited to classrooms. Yet while distinct from other contexts of professional practice (industry settings or research centers), academic learning environments do provide partial but meaningful sites to understand some of the crucial ways in which practitioners are immersed and encultured into the professional discourse (Goodwin, 1994; Lave & Wenger, 1991). In courses and workshops, instructors—as established practitioners—describe proven data science methods, theories, and techniques to newcomers in their field, helping us see how different pedagogic demonstrations and examples enable certain in-practice norms and heuristics (Burrell, 2016).

More importantly, a study of classrooms and workshops draws attention to the social aspects of learning: processes of participation and membership in a discourse, instead of just a set of individual experiences. Learning and professionalization in academic environments are accomplished via guided interactions between instructors, students, teaching assistants, course materials, quizzes, assignments, and exams. Learning environments thus function as important sites in which would-be data science practitioners learn integral aspects of what it means to *do* data science—a crucial rite of passage on their way to becoming members of their "communities of practice" (Lave & Wenger, 1991).

Data science courses and workshops mostly focus on technical training: a lot of know-what garnished with some amount of know-how. Take algorithms, for instance. In classrooms, an algorithm is often explained as a set of formal rules, a plan of action to organize data in actionable and predictable ways. Indeed, in universities, students learn all about various kinds of algorithms. However, when they begin to do data science, students realize that each dataset poses unique challenges (and opportunities), requiring them to figure out ways to accommodate variations in their routine acts of *applying* algorithms. Accommodations require discretion. To learn to do data science is to learn to see similarities and differences in how to put data and algorithms together. To be a data scientist is to see the unknown, the distinct, and the singular within the mundane and predictable.

Such creative and improvisational forms of human work, while acknowledged, are often rendered secondary to seemingly 'real' forms of work such as that of building models in courses and workshops. Even outside of coursework, this is particularly evident in how instructors demonstrate and require students to document and present data science work—the same old story of questions, data collection, model building, testing, and numerical answers. Open conversations about the discretionary work essential to data science practices are rarely encouraged. Unsurprisingly, students are often unable to see beyond their tools and rules towards the broader impact of subjective judgment on data science and its necessity in everyday practice. Students learn about every hammer in their toolkit but are left to their own devices to understand how to make hammers, nails, and worlds work together.

This thesis focuses on data science education to analyze how students learn and are taught data science. I describe the different (in)visibilities of certain types of data science work

in formal instruction, showing how and why their effective practice requires students to have both technical and discretionary abilities. How are students trained to 'see' the world through data, models, and numbers? How do students learn to improvise around formal knowledge, methods, and tools in the face of emergent challenges? What can a study of data science training tell us about the challenges concerning real-world data science work? My analysis of the empirical context of data science learning—presented only in chapter 2—aims to answer such questions.

Corporate organizations

The second empirical context that I focus on in this thesis is corporate data science practice. Rapidly becoming data-driven, businesses now increasingly make decisions based on the quantitative analysis of data. Often called the "new oil" (Mayer-Schönberger & Cukier, 2013; Rotella, 2012) or a "force" to be controlled and reckoned with (Puschmann & Burgess, 2014), big data are now often considered a valuable form of capital (Yousif, 2015).

The advent of data science has only further exacerbated such tendencies. Turning towards an *AI first* sensibility, corporations now generally consider 'more data=more knowledge' as a truism. This is evident in the fact that in the socioeconomic imagination of most of today's businesses, the process of big data mining and analysis is sometimes called the process of "reality mining" (Pentland, 2009, p. 79). The situation is intense; some say that big data will soon signify the end of theory: given enough data, numbers speak for themselves, providing necessary explanations for all observations (Anderson, 2008). Similar arguments encompass, as evident in business publications and as argued by researchers, current ideology around corporate data science (Cukier & Mayer-Schönberger, 2013; Davenport, 2014;

Domingos, 2015; Franks, 2012; Hildebrandt, 2011; IBM, 2012; Kitchin & McArdle, 2016; Simon, 2013). Corporate data science practice is thus an obvious and crucial area of research.

What is the nature of the everyday work involved in corporate data science projects? How do industry practitioners design and build systems under complex and messy business and organizational conditions? What are the impacts of organizational culture, business goals, and managerial structures on the everyday work of data science in corporate settings? My main goal with analyzing the applied practice of data science in corporate contexts is to answer such questions.

In answering such questions, I continue to focus on the differential (in)visibilities of human, technical, and organizational work in corporate data science practices. In doing so, I draw on and contribute to scholarship in the field of organization studies on how organizational actors design, develop, and use technologies in the workplace (e.g., Feldman & Orlikowski, 2011; Leonardi, 2009; Neyland, 2016b; Orlikowski, 1992, 2007; Orlikowski & Gash, 1994; Reynaud, 2005; Strauss, 1988; L. Suchman, 2000). I also pay special attention to how corporate actors make sense of the nature of work involved in data science practices (Orlikowski & Gash, 1994; Weick, 1995) and how they make and justify their assumptions, choices, and decisions when doing data science (Boltanski & Thévenot, 2006; M. D. Cohen, 2007; Dewey, 1939; Stark, 2009).

The impact of organizational dimensions on data science work remains understudied in existing scholarship on corporate data science. As described earlier, even when their focus is on corporate data science work, researchers mainly analyze *what* industry practitioners need

instead of *how* the corporate context shapes data science work.[7] This is especially limiting as human and organizational work have huge repercussions for corporate data science practices. Central to such implications is the "responsibility gap" (Matthias, 2004)—the difficulty of determining who is really at fault when data science systems do harm. Is it the human, the company, the model? Traditional models of responsibility fail in such scenarios since, as is argued, "nobody has enough control over the machine's actions to be able to assume the responsibility for them" (ibid.: 177). The decision-making structures required to manage the design, working, and implications of data science systems often exceed the human and organizational resources available for oversight (Kitchin & McArdle, 2016). This leads researchers to try to solve the problems generated by data science by using and building more data science tools:

> "We cannot in principle avoid epistemic and ethical residues. Increasingly better algorithmic tools can normally be expected to rule out many obvious epistemic deficiencies, and even help us to detect well-understood ethical problem (e.g., discrimination). However, the full conceptual space of ethical challenges posed by […] algorithms cannot be reduced to the problems related to easily identified epistemic and ethical shortcomings" (Mittelstadt et al., 2016)

This is evident in that researchers mostly treat the ethical and normative issues arising from the use of data science systems as indicative of technical issues with data, models, or procedures (Busch, 2014; Domingos, 2012; Hajian & Domingo-Ferrer, 2013; Jacobs et al., 2020; Wattenberg et al., 2016). Seeing data science harms as outcomes of data science practitioners'

[7] For recent notable exceptions, see: Metcalf et al. (2019), Veale (2017), and Veale et al. (2018).

inability to recognize some form of bias or fairness, defined technically, researchers work hard to attempt to fix the issues through robust measures, fairness toolkits, or procedural checklists.

Such efforts are not in vain. Technicalities are key to data science, and technical solutions can and do bring positive changes. The question is not whether technical fixes are an effective way forward but if they are the *only* way forward. The answer is no. As you will learn in this thesis, normative implications of data science systems are as much a consequence of the human and organizational work involved in building them as of data attributes, algorithmic limitations, and model properties. Data science ethics have human and organizational, not just technical, roots. Seemingly mundane upstream choices and imperatives have severe downstream impacts. I reveal the everyday origins of ethical and normative issues, and the many ways in which practitioners struggle to define and deal with them in practice, describing how not all data science problems are technical in nature—some are deeply human, while others innately organizational. Researchers care deeply about the implications of data science. However, they often do not realize that the causes of issues they work on might differ from what they think and will thus need different solutions. We must change how we think about the origins of data science issues and implications. This thesis takes a step in this direction.

Organizing principle: Data science project workflow

There are several ways to narrate a story of corporate data science work. It can be broken down by projects or told from the perspective of practitioners. This thesis uses a data science project's typical workflow as the organizing principle for the chapters dealing with corporate data science. Each chapter focuses on a form of everyday work in the data science lifecycle—translating high-level goals into data-driven problems (chapter 3), building data science systems

(chapter 4), and assessing their credibility (chapter 5). I describe my reasons for choosing these three specific forms of work in each chapter—reasons based mainly on the scholarly relevance and practical importance of the human, technical, and organizational work involved at each step in the data science lifecycle.

Corporate data science projects are complex beasts, involving different aspirations, experts, goals, practices, teams, and forms of work. Focusing on specific stages in the data science lifecycle also makes it easier to describe the messy entanglement between the human, technical, and organizational aspects of one kind of everyday data science work.

1.3 A word on technical, human, and organizational work

As you may have noticed, I use three key terms—technical, human, and organizational work—to call out distinct kinds of data science work. I have thus far used these terms intuitively, relying on their everyday usage rather than trying to define them. Below I provide working descriptions for each.

My attempts to differentiate between the three forms of work is an analytic choice. In practice, human and technical work are intertwined in the everyday practice of data science—a mix further muddled by organizational work in corporate settings. My categorization of specific data science work as either of the three is thus not an exercise in classification. Instead, I use the terms to call out the salient aspects of specific kinds of data science work. In doing so, I borrow from Latour (1990), adapting one of his methodological maxims to my research—my goal is not to judge whether a form of data science work is human or technical, but to understand how in practice human, technical, or organizational aspects characterize specific kinds of data science work.

The descriptions below are thus clues, not definitions. Indeed, the identification and analysis of the human and organizational work implicated in data science practices are one of the primary contributions of my research. You have confronted these terms from page one of this thesis, and you will keep puzzling over them until its last page. The explanations below will help you on this journey.

1.3.1 Technical work

Every profession involves technical work—systematic applications of expert knowledge. Doing such work requires know-*what* (knowing the professional repertoire of facts, methods, and tools) but also know-*how* (knowing how to apply facts, methods, and tools to problems). The term *technical* finds its roots in the Greek *technē*—"the name […] for the activities and skills of the craftsman, but also for the arts of the mind and the fine arts" (Heidegger, 1954, p. 13). Theoretical and practical knowledge remain entangled in all forms of technical work. Data science is no different, and I use the term *technical work* to denote both theoretical and practical forms of data science work—knowing and doing.

From this perspective, every form of data science work may appear to fall under the umbrella of technical work—from organizing the world as data and crafting data-driven problems to making systems work in desired ways and assessing trust in models. But this is not what happens in practice. As described earlier, the overemphasis on some work drives other work towards, and at times outside, the technical margin. I use the term technical work to capture such movements. In labeling a form of work as technical, I call attention to its high visibility and overvaluation within data science practices. Examples of such work include, but are not limited to, data collection, feature engineering, and model selection.

1.3.2 Human work

I use the term *human work* to denote the discretionary work involved in data science practices—aspects of praxis that rely on creativity, improvisation, and subjectivity. Human work consists less of know-what and more of know-*how* and know-*why*—situated choices and practical decisions that drive real-world data science but are often unaccounted for in scholarly and public discourse.[8] An effective way to grasp how I use the term human work is to compare it to how I use the term technical work. If you bracket off all highly visible and overemphasized forms of technical work in data science practice, the forms of work that get left behind are, what I call, human work. Examples include translating high-level goals into tractable data-driven problems, resolving emergent challenges, and establishing trust in data.

While reading this thesis, keep in mind the difference between the *process* of human work and the *product* of such work. For instance, as I show in this thesis, the human work of interpreting model results consist of iterative and messy forms of "sensemaking" (Weick, 1995) on the part of practitioners. And while practitioners will often describe their interpretation of certain model results in papers and presentations, such descriptions only point to the eventual meaning practitioners ascribe to a finding and not to the way by which they arrived at the interpretation. Focusing on both the process and product of human work, in this thesis I report on both kinds of human work—those that are not recorded and those that are sometimes recorded but only in terms of their product.

[8] There are other ways to demarcate between technical and human work. For instance, technical work as the work done *by* algorithms and models and human work as the work of practitioners. I do not employ this view in this thesis. My goal is not to define what work are done by machines and what by people. Situating data science as a sociotechnical practice, I see practitioners generating knowledge *with* and *through*, and not merely *using*, machines.

1.3.3 Organizational work

I use the term *organizational work* to call out a specific form of human work done to align practice of data science with business goals, organizational culture, and other industry practices. Examples of such work include managing projects, people, and expectations; collaborating with individuals and teams; enabling and facilitating communication between people, teams, and leadership; aligning project outcomes with business objectives and key results (OKRs); and balancing between computational ideals and product requirements. I use the term organizational work to also cover "articulation work" (Strauss, 1988)—forms of meta project work to outline what must be done, by whom, when, how, and to what extent. Note that organizational work often, though not always, consists of the collective work of several practitioners.

1.4 Research sites, data, and method

1.4.1 Empirical fieldsites

The research in this thesis is based on multiple sets of ethnographic fieldwork and qualitative research conducted at a well-known American university on the East Coast (academic fieldsite) and at a large-scale American technology corporation based on the West Coast (corporate fieldsite).

Academic Fieldsite

I did two separate ethnographic participant-observations (four month long each) in two courses: a graduate-level ML course (~160 students, Fall 2014) and a senior-level NLP course (~40 students, Spring 2015). I enrolled as a student in each course and regularly attended classes. The ML course had two sections; I joined one of them (~80 students). Both courses had a mix of undergraduate, master, and doctoral students. I got approval from the instructors in each

course before doing research. I did not do audio/video recordings and did not collect any personal data on participants.

My approach to participant-observation in the two courses, however, differed on one aspect. In the NLP course, I told the students about my research on the first day of class. But in the ML course, I informed the students of my research project only after the last day of class. While emergent, this decision had an underlying logic. As a researcher, I could easily blend in as a student in the ML course because it had a large class size and was held in a big auditorium. However, the NLP course had fewer students and an intimate setting—a small classroom with students sitting together on benches. I figured this would make it difficult for me to blend in as I would spend most of my time taking fieldwork notes.

In both courses, students were informed—by the instructor and me—that their participation in research was voluntary and had no impact on their grades. Students could opt out of the research. If a student opted out, I would not take notes on what they say or do in class. This, however, posed a challenge for me in the ML course since it required me to delete a student's information after the fact. I did not collect any personal identifiers, making it challenging to identify specific students in research data. Thankfully, none of the students in the two courses opted out of the project.

Beyond courses, I did a participant-observation study of a series of digital humanities workshops organized at the same university (Spring 2015). The three-part workshop introduced students to NLP techniques for text analysis. Most participants were doctoral students in the English department. Each workshop lasted two hours, and the number of participants ranged from nine to thirteen. I did not do audio/video recording and did not collect any personal data

on participants. At the start of the first workshop, the instructor informed the group that I was there to conduct research. I informed the participants of my research project and provided consent forms to each participant for approval. Participants could opt out of research, but all of them willingly participated in the project.

Corporate Fieldsite

I conducted six months of ethnographic fieldwork with DeepNetwork[9], a multi-billion-dollar US-based ecommerce and new media technology corporation. To gain more immersive and participatory access to ongoing and ordinary work practices and experiences, I worked as a data scientist at the organization between June and November 2017, serving as lead data scientists on two business projects (not reported in this thesis) and participating in several others.

About the company. Founded in the nineties, DeepNetwork owns 100+ companies in domains such as health, legal, and automotive. Many of DeepNetwork's subsidiaries are multi-million-dollar businesses with several thousand clients each. The corporation has a primary data science team based at their US West Coast headquarters that works with multiple businesses across different domains and states. There are multiple teams of analysts, designers, managers, developers, and engineers both at DeepNetwork and its subsidiaries. Although not a research-driven corporation (in the fashion, for example, of Google, Amazon, or Microsoft), DeepNetwork is an extremely data-driven company with copious amounts of data across many businesses. However, its focus is not data science research (though it holds ML and AI patents) but business applications of data science.

[9] All organization, project, software, and personnel name are replaced with pseudonyms.

During my time at the company, DeepNetwork's primary data science team consisted of eight to eleven members (including myself). The team's supervisor was Martin—DeepNetwork's Director of Data Science with 30+ years of industry experience managing technical projects at several technology firms. Martin and the data science team reported to Justin—DeepNetwork's Chief Technology Officer (CTO) with 20+ years of experience in the technology industry.[10]

Research access, negotiation, and consent. I applied for an interning data scientist job position at DeepNetwork, going through a series of technical and behavioral interviews (phone and on-site). I informed all interviewees that my primary goal was to research corporate data science practices. I clearly stated that I would work at the company for a limited time and would need explicit permission to do research on company premises, gathering data such as fieldwork notes and audio-recorded interviews. DeepNetwork's final decision to hire me was based primarily on my data science ability and knowledge.

As part of the negotiation process, DeepNetwork and I settled on three crucial aspects of my research design. *First,* participation was optional, and each participant needed to sign a consent form (I provided a copy of the form for approval). Participants could opt out of research at any point, and all personnel and project names would be replaced with pseudonyms to preserve anonymity and privacy. *Second,* participants could consent to selective participation. For example, a participant could consent to the fieldwork (i.e., have their data recorded in fieldnotes) but not to the interview (I provided a copy of my interview topic guide for approval

[10] For context, I provide additional background information on each DeepNetwork employee as and when they are introduced in each chapter. Certain details have been omitted to preserve participant anonymity and privacy.

and explained that it would change over time as the research progressed). Participants could choose not to have interviews recorded. *Third*, DeepNetwork would define what data I can and cannot collect. However, the analysis of research data, including for reasons of inter-participant privacy, would be done solely by me.

I also provided a copy of Cornell IRB's research approval certificate. All submitted documents were vetted by DeepNetwork's Human Resources and Legal departments. DeepNetwork agreed to research participation. As part of the non-disclosure agreement, I agreed not to make copies of company data and proprietary code. In practice, all company employees, including company leaders, participated willingly and openly in research. Only two people asked not to record their interviews.

Affordances and limitations of my dual role. My dual role as both ethnographer and data scientist, communicated explicitly in my first meeting with each participant, presented both research opportunities and challenges.

On the one hand, my data scientist work helped me to build rapport with company personnel. For instance, after about four weeks, my daily meetings with other data science team members almost exclusively consisted of detailed and practical conversations around technical challenges, algorithmic issues, and ongoing projects. Even non-data-scientists began to see me primarily *as* a data scientist. For instance, even in research interviews, business analysts and project managers would ask me questions about project requirements and updates. I became the lead data scientist on two projects after about two months at the job. This further solidified my main identity as that of a data scientist—to the extent that other data scientists, for instance,

often explicitly articulated surprise (e.g., *'but, you knew that, right?'*) when I would ask them to explain something for the ethnographic record.

On the other hand, my data-scientist identity posed difficulties in personal communication with non-data-scientists. Business team members, for instance, hesitated to point out faults and issues with the data science team, sometimes resorting to conciliatory descriptions such as 'your team did its best,' 'maybe we are the problem,' and 'we love the work your team is doing.' In these moments, I had to work to make visible my dual role, reminding the interviewees that this was a research interview and assuring them that their critique would never be directly communicated to CTO Justin, director of data science Martin, data scientists, or other company personnel.

1.4.2 Qualitative data

Academic fieldwork

I collected the following data in courses and workshops: fieldwork notes and learning materials. The research in the ML and NLP courses produced a total of 384 pages of fieldwork notes (notebooks). The research in the workshops produced 30 pages of fieldwork notes (software[11]).

Corporate Fieldwork

During the six-month period of ethnographic research, I did 52 interviews with data scientists, project managers, business analysts, product managers, and company executives. The research also produced 426 pages of fieldwork notes (software) and 104 photographs. After completing fieldwork, I went back to DeepNetwork several times to give research talks. During those visits,

[11] For academic fieldwork, I jotted down research notes in my personal notebooks. For corporate fieldwork, I initially recorded my research notes in the Evernote software and later converted them to Microsoft Word files.

I conducted audio-recorded interviews with six data science team members (two of them were new hires). Additional interviews and informal conversations with company personnel during these visits are not an official part of this thesis's research data but have been valuable in sharpening my research analysis and focus.

1.4.3 Research Methodology

I categorized academic fieldwork data by courses (two separate sets of data for the ML and NLP course) and workshops (one set for workshops). For corporate fieldwork, I organized interview and fieldwork data in two ways: categorized by projects (e.g., all data on one project in one set) and categorized by actor groups (e.g., all data on a specific actor group, say business analysts, in one set). The former enabled me to analyze themes within and across projects; the latter allowed me to examine specific actor perspectives.

Categorizing data by projects also helped to identify and analyze common patterns and practices occurring at specific timeframes in the data science project lifecycle. As described earlier, how I structure the chapters on corporate data science in this thesis draws from this categorization, situating my analysis of the different kinds of everyday data science work within a corporate project's usual temporality.

Qualitative analysis

All interviews and fieldnotes were transcribed and coded according to the principles of grounded theory analysis (Charmaz, 2014; Strauss & Corbin, 1990), as articulated by Anselm Strauss and students and as previously applied in CSCW research by scholars such as Susan Leigh Star (Star & Ruhleder, 1994) and Ellen Balka (Balka & Wagner, 2006; Schuurman & Balka, 2009).

Grounded theory is a data-driven method for qualitative analysis in which working theories about a phenomenon under investigation are generated from the collected data through iterative and inductive analysis. I performed two rounds of coding—labeling and defining—for most of my research data using both in-vivo and open codes. [12] In-vivo coding involves generating codes from the data (e.g., using actor phrases as codes), while open coding involves codes manually assigned by a researcher based on ongoing data analysis. A core aspect of my analytic approach was the principle of "constant comparison" (Glaser & Strauss, 1967)—frequently juxtaposing and analyzing individual findings and themes that emerged from the ongoing analysis.

Not all analysis happened entirely after finishing data collection. A key part of the grounded theory approach is 'theoretical sampling'—altering research approach based on ongoing analysis of data during data collection. Chapter 5, in which I focus on how corporate practitioners assess the credibility of data, models, and numbers, provides an excellent example of theoretical sampling. The seeds of the analysis present in this chapter were sown during my time at DeepNetwork. During fieldwork, I encountered discrepancies between how different actors (such as data scientists, project managers, or business analysts) articulated problems with and confidence in algorithms, data, models, and numbers. I saw how data science projects consisted of diverse negotiations to tackle varying forms of uncertainties. I followed up on this theme during fieldwork by focusing my attention on the points of friction and collaboration between different actors and teams (in fieldwork notes and interviews). Doing so helped me identify a series of situated tensions that impacted actors' ability to manage uncertainties in

[12] Qualitative coding was done using the ATLAS.ti 8 (Windows) software.

corporate data science work (the four most salient topics I describe in the chapter). I chose the theme of trust as the chapter's organizing principle after multiple iterations of subsequent analysis.

A couple of practical points and a word on representativeness

Specific quotations in the empirical stories told in each chapter reflect a mix of transcript data from situations occurring, for instance, during course instruction or project meetings (thus embedded in the ordinary flow of everyday work in academic learning and corporate practices) and from separate interviews done with corporate practitioners. I label audio-recorded communications as *interviews* and non-audio-recorded conversations as *fieldwork notes*.

Finally, I chose specific situations and projects to report in this thesis because of their salience to the type of data science work I analyze in each chapter. Nonetheless, they are broadly representative of patterns that I found across the much larger number of cases not reported in the thesis and are consistent with my experiences as a student in courses/workshops and as project lead and collaborator on projects.

1.5 Chapters: A preview

The rest of the thesis consists of five chapters. In chapter 2, **Data Vision**, we dive into the world of data science classrooms and workshops.[13] I explain why doing data science requires the simultaneous mastery of two abilities: the *technical ability* to know and apply complex methods and tools and the *discretionary ability* to know how, why, and when to improvise the situated applications of methods when faced with on-the-ground challenges. Before embarking on data

[13] As mentioned earlier, this is the only chapter that uses data from my academic fieldwork.

science projects, practitioners must first grasp the intimate connections between formal rules, empirical contingency, and situated discretion in their work. In many ways, this chapter is a kind of proto chapter for the rest of the thesis. I focus on what must be in place for data science to even happen, analyzing how practitioners learn the technical and human work of data science and how certain forms of work remain differently (in)visible in formal instruction. To do data science—indeed, to become a data scientist—practitioners must first embody a distinct and powerful way of seeing the world as data, organizable and manipulatable by algorithms, models, and numbers. I create the concept of *data vision* to theorize such artful, not mechanical, practices of seeing the world.

From next chapter onwards, we move away from academia and into the corporate world to focus on data science work done in service of business goals. In chapter 3, **Problem Formulation and Fairness**, I focus on a crucial first step in data science projects: problem formulation—the work of translating high-level goals into data-driven problems. An uncertain and complicated process, problem formulation requires substantial discretionary work. Making visible the actors and activities involved in problem formulation, I describe how the specification and operationalization of data science problems in corporate projects are negotiated and elastic. In fact, I show that how practitioners choose to solve problems has design but also normative implications, even though practitioners rarely work out these formulations with explicit ethical concerns in mind. Normative assessments of data science often take problem formulations for granted even though they are seldom obvious and can raise distinct ethical issues. Linking the normative concerns that data science provokes to careful accounts of its everyday work, I explain how and why practitioners pose problems in specific

ways and why some formulations prevail in practice, even in the face of what may seem like normatively preferable alternatives.

In chapter 4, **Making Data Science Systems Work**, we take a step further into the corporate data science lifecycle, focusing on a crucial next step after problem formulation: building working data science systems. It may appear that whether a system works is merely a function of its technical design, but, as I describe in this chapter, it is also accomplished through ongoing forms of discretionary work. I show how and why a data science system's working is neither stable nor given but an improvised and resourceful artifact that remains *in the making* throughout a project. Practitioners constantly negotiate what work systems should do, how they should work, and how to assess their working. In this chapter, I analyze how such negotiations lay the foundation for how, why, and to what extent a system ultimately works relative to business goals, emergent design challenges, and other existing technologies. Through a detailed account of the on-the-ground realities of system development, I make visible the consequential relations between the working of data science systems and the everyday work of building them.

In chapter 5, **Trust in Data Science**, we analyze how practitioners trust the data science systems they build. Credibility checks occur throughout corporate data science projects, culminating in a crucial decision towards the end of a project: *is the system ready for deployment?* Trust is central to the effective development, adoption, and use of data science systems. Businesses spend considerable amounts of time, money, and effort to ensure the credibility of their data, models, and results. Established trust mechanisms (e.g., validation methods or performance metrics) engender, what I call, *calculated trust* through quantitative assessment—a highly visible form of technical data science work. I show, however, that

establishing trust in data science requires both calculative and collaborative work. Evaluation of a system and its parts is rarely the simple application of a set of performance criteria, often requiring situated work to negotiate between different, often divergent, criteria. In this chapter I explain how different practitioners describe issues with and confidence in data, models, and numbers in distinct ways. Highlighting the collaborative and heterogeneous nature of data science work, I illustrate how certain common tensions raise pertinent problems of trust in business projects and how practitioners manage, resolve, and even leverage these tensions via ongoing forms of practical work. Central to recent discussions and research on trust in data science, forms of calculated trust alone, as I argue in this chapter, fail to capture the plurality of expertise and modes of justification that typify problems and solutions of trust in real-world data science.

In chapter 6, **Conclusion**, we reach the end of our journey into the human and organizational lifeworlds of data science practices. I will first do quick recap of the previous four chapters, describing key learnings from each chapter (and relating them back to the concerns that I raised in this Introduction). Finally, I provide a set of high-level takeaways, outlining how the research presented in this thesis changes our conception of the everyday work involved in data science projects, the gaps between academic training and professional practice of data science, and the opportunities afforded by an engaged form of research, like the one in this thesis, for shaping our data science futures.

II
Data Vision

We begin our analysis of the different kinds of (in)visible work involved in everyday data science practices by focusing on how students learn and are taught data science. In this chapter, I explain how and why doing data science requires the cultivation of data vision—a way of 'seeing' the world as data, organizable and manipulatable via models and numbers. I highlight the different forms of human and technical work necessary to realize data vision: work bound by formalism, math, and rules but also work based around creativity, discretion, and improvisation. Both forms of work are foundational to doing data science, yet only one is visible and the other sidelined in representations of data science work. Such (in)visibilities—occurring as early as in learning—have consequences. Against the backdrop of a rule-bound verses a rule-based understanding of data science, I explore the real-world implications of data vision for addressing and accommodating the many (in)visibilities of human and technical data science work.

2.1 Introduction

Algorithmic data analysis has come to enable new ways of producing and validating knowledge (Gitelman, 2006; Leonelli, 2014). Algorithms are integral to many contemporary knowledge

[†] A slightly edited version of this chapter has been published as Passi & Jackson (2017). I proposed the initial research question and refined it with feedback from Steve Jackson. I designed and conducted the participant-observation study (with feedback and guidance from Phoebe Sengers, Steve Jackson, and Malte Ziewitz). I analyzed the data on my own; Steve Jackson gave regular feedback that shaped the direction of analysis. I wrote the first draft and later revisions with feedback from Steve Jackson.

practices, especially ones that rely on the analysis of large-scale datasets (Gillespie, 2014; Gitelman, 2006; Kitchin, 2014a, 2014b). At the same time, we know that algorithms can be selective (Gillespie, 2014), subjective (boyd & Crawford, 2012), and biased (Barocas, 2014); that they work on multiple assumptions about the world and how it functions (Bowker, 2013; Gillespie, 2014; Gitelman, 2006; Taylor et al., 2015); and that they simultaneously enable and constrain possibilities of human action and knowledge (Bowker, 2013, 2014). Algorithmic knowledge production is a deeply social and collaborative practice with sociocultural, economic, and political groundings and consequences.

In all these ways, data science embodies a distinct and powerful way of *seeing* the world. Data scientists learn to represent and organize the world via computational forms such as graphs, matrices, and a host of standardized formats, empowering them to make knowledge claims based on algorithmic analyses. The world, however, does not always neatly fit into spreadsheets, matrices, and tables. While data science is often understood as the work of faceless numbers and unbiased algorithms, a large amount of situated discretionary work is required to organize and manipulate the world algorithmically. Effective algorithmic analysis also demands proficiency in the ways that worlds and tools are put together, and *which* worlds and tools are so combined (across the wide range of methods, tools, and objects amenable to representation). Taken together, these two seemingly contradictory features constitute what I call *data vision:* the ability to organize and manipulate the world with data and algorithms, while simultaneously mastering forms of discretion around *why*, *how*, and *when* to apply and improvise around established methods and tools in the wake of empirical diversity.

Integrated, often seamlessly, in the practice of expert practitioners, these contradictory demands stand out with clarity in the moments of learning and professionalization through which novices learn to master and balance the intricacies of data vision. How do students learn to "see" the world through data and algorithms? How do they learn to maneuver and improvise around forms and formalizations in the face of empirical contingency? This chapter addresses such questions within the context of data science learning environments such as classrooms and workshops.

I describe two separate sequences of events—one from a ML classroom and another from a set of DH workshops—to showcase how learning to see through data requires students to balance between viewing the world through abstract constructs, while simultaneously adapting to empirical contingency. I advance a rule-*based* (as opposed to a rule-*bound*) understanding of data science practice, highlighting the situated interplay between formal abstraction and mechanical routinization on the one hand, and discretionary action and empirical contingency on the other hand, showing how it is the mastery of this interplay—and not just the practice of data science techniques in their formal dimension—that is central to the growing skill and efficacy of would-be data scientists. I argue that a better understanding of data vision in its more comprehensive and discretionary form can help instructors and researchers better engage and leverage the human dimensions and limits of data science learning and practice.

In the sections that follow, I begin by reviewing relevant scholarly literature on professional vision and situated knowledge before moving on to describe the two empirical cases. I conclude by discussing the implications of the notion and practice of *data vision*, and

the distinction between a *rule-bound* and *rule-based* understanding of data science, for data science learning, research, and practice.

2.2 Professional vision, situated knowledge, and discretionary practice

My understanding of data vision is based on a classic and growing body of work in the social sciences that has explored forms of identity, practice, and vision underpinning and constituting forms of professional knowledge. In his work on *professional vision,* Goodwin (1994) analyzes two professional activities (archaeological field excavation and legal argumentation) to show how professionals learn to "see" relevant objects of professional knowledge with and through *practice:* the exposure to and exercise of theories, methods, and tools to produce artifacts and knowledge in line with professional goals. Learning professional practice, he argues, help professionals make salient specific aspects of phenomena, transforming them into objects of knowledge amenable to professional analysis. Learning to see the world professionally, however, is not achieved through mere learning of generic rules and formal techniques. Instead, professional vision is slowly and carefully built through training, socialization, and immersion into professional discourse (Goodwin, 1994; Lave & Wenger, 1991; Pine & Liboiron, 2015; L. A. Suchman & Trigg, 1993). Professional vision, thus, is a substantively collaborative sociocultural accomplishment—a way of seeing the world constructed and shaped by a "community of practice" (Lave & Wenger, 1991).

A key aspect of professional vision, Abbott (2014) argues, is the way in which practitioners situate given problems within existing repertoires of professional knowledge, methods, and expertise. According to Abbott, the process of situating given problems—of "seeing" professionally—must be clear enough for professionals to create relations between a

problem and existing knowledge *(e.g., what can I say about a specific dataset?)*, yet abstract, even ambiguous, enough to enable professionals to create such relations for a wide variety of problems *(e.g., what are the different kinds of datasets about which I can say something?)*.

A similar interplay between abstraction, clarity, and discretion exists in data science practices. Algorithms, developed in computational domains such as AI, ML, and NLP, provide means of analyzing data. It is often argued that a specific algorithm can work on multiple datasets so long as the datasets are modeled in particular ways. However, data science requires much more work than simply applying an algorithm to a dataset. As Mackenzie (2015) argues, certain data science practices such as vectorization, approximation, and modeling often mask the inherent subjectivity of dataset and algorithms, imbuing them with a sense of inherent "generalization." A large amount of situated and discretionary work—e.g., data collection, data cleaning, data modeling, and other forms of pre- and post-processing—is required to make datasets work with chosen algorithms: from choice of analytic method and choices concerning data formatting to decisions about how best to communicate data analytic results to 'outside' audiences. Data scientists do not just learn to see and organize the world through data and algorithms, but also learn and discern meaningful and effective combinations of data and algorithms. As Gitelman et al. (2013) argue: "raw data"—at least as a workable entity—is an oxymoron. It takes *work* to make data work.

Abbott's (2014) example of chess is instructive in evoking the situated and discretionary work characteristic of all forms of practice. The opening and closing moves in a game of chess, Abbott contends, often appear methodical and rigorous. However, in between the two moves, he argues, is the game itself in which knowledge, expertise, and experience intermingle as the

game progresses. On the one hand, we can summarize and teach chess as a collection of formal *rules* (e.g., how a pawn or rook moves, ways to minimize safe moves for your opponent, etc.). On the other hand, however, we must acknowledge that all *applications* of such rules are situated, contingent to the specific layout of the game at hand. In this way, chess (and professional vision) is rule-*based* but not rule-*bound*—a distinction I return to in the discussion.

These insights are backed in turn by a long line of pragmatist social science work dealing with the nature of 'routines' and 'routinizable tasks' in organizational and other contexts. Building on Dewey's (1922) foundational work, Cohen (2007) argues against the common understanding of routinized tasks as collections of rigid and mundane actions, guided by "mindless" rules and mechanized actions; instead, he describes how every performance of a routine is unique:

> "For an established routine, the natural fluctuation of its surrounding environment guarantees that each performance is different, and yet, it is the 'same.' Somehow there is a pattern in the action, sufficient to allow us to say the pattern is recurring, even though there is substantial variety to the action." (Ibid.: 782)

This variety lets us recognize each 'application' of a routinized task as a unique performance, allowing us to draw out similarities and differences across multiple iterations of the 'same' routine.

Klemp et al. (2008) also draw on Deweyan roots to address these "similar, yet different" applications of routines through the vocabulary of plans, takes, and mis-*takes*. There might be a *plan* (a method, an algorithm, a script), and there might be known mistakes (incompatibility, inefficiency, misfit), but every application of the plan is a *take* ripe for mis-*takes* that occur when professionals are faced with something unexpected during the execution of formal and

established routines. Using the example of a Thelonious Monk jazz performance, the authors explore the complex discretionary processes by which musicians deal with mis-*takes:*

> "When we listen to music, we hear neither plans nor mistakes, but takes in which expectations and difficulties get worked on in the medium of notes, tones and rhythms. Notes live in connection with each other. They make demands on each other, and, if one note sticks out, the logic of their connections demands that they be reset and realigned." (Ibid.: 22)

Mis-*takes*, then, mark elements of "contingency, surprise, and repair [found] in all human activities" (ibid.: 4). Signifying the lived differences between theoretical reality and empirical richness, mis-*takes* necessitate situated, often creative, improvisations on the part of professionals and indeed other social actors involved in the practice.

Like Abbott's description of chess, Klemp et al.'s analysis draws out the discretionary and situated nature of professional knowledge and practice even in straightforward and routinized procedures. This point is further elaborated by Feldman & Pentland (2003) who show how routines are both ostensive (the structural rule-like elements of a routine) and performative (the situated and contingent execution of a routine). It is the interplay between these two aspects that allows for the seemingly discernable but always shifting reality of routinized work and professional practice. Along similar lines, Wylie's (2014) study of paleontology laboratories shows how adapting situated routines and practices to deal with new problems-at-hand is considered an integral aspect of learning by doing. "Problem-solving in ways acceptable to a field [...] can be an indicator of skill, knowledge, and membership in that particular field" (ibid.: 43).

However, the situatedness of a practice is not always visible. Ingold (2010, p. 98), using the example of a carpenter sawing planks, describes how to an observer, "it may look as though […a] carpenter is merely reproducing the same gesture, over and over again." Such a description, he reminds us, is incomplete:

> "For the carpenter, [...] who is obliged to follow the material and respond to its singularities, sawing is a matter of engaging 'in a continuous variation of variables…" (Ibid.)

To improvise routine tasks then is to "follow the ways of the world, as they open up, rather than to recover a chain of connections, from an end-point to a starting-point, on a route already travelled" (ibid.: 97).

Such social science insights on professional vision and discretionary practice have translated into important CSCW and HCI research programs. For instance, Suchman and Trigg (1993) demonstrate the role and significance of representational devices for ways in which AI researchers see and produce professional objects and knowledge. Mentis, Chellali, & Schwaitzberg (2014) show how laparoscopic surgeons demonstrate ways of "seeing" the body through imaging techniques to students: "seeing" the body in a medical image is not a given, but a process requiring discussion and interpretation. Mentis & Taylor (2013) similarly argue that "work required to see medical images is highly constructed and embodied with the action of manipulating the body." Situating objects or phenomena in representations, they argue, is a situated act: representations do not just *reveal* things, but also *produce* them, turning the "blooming, buzzing confusion" of the world (Joyce, 1922) into stable and tractable "objects" amenable to analytic and other forms of action.

Performing analytical and other forms of actions on the world, however, requires people to deal directly with empirical contingency. Suchman (2007) argues that "plans" of action are theoretical, often formulaic, representations of human actions and practices. "Situated action," however, requires people to work with continuous variation and uncertainty in the world. Human action, she argues, is a form of iterative problem solving to accomplish a task. Creativity often emerges within such discretionary and situated forms of problem solving. As Jackson & Kang's (2014) study of interactive artists shows, dealing with material mess and contingency (in this case, attached to the breakdown of technological systems and objects) may necessitate and drive forms of improvisation and creativity at the *margins* of formal order. Creativity—understood not as an abstract and free-standing act of cognition but instead as a *situated sociomaterial accomplishment*—emerges through the interplay between routines and applications, between plans, takes, and mis-*takes*, and between empirical mess and theoretical clarity.

Such situated and discretionary acts are no less central to forms of data science and algorithmic knowledge studied and practiced by CSCW and HCI researchers. Clarke (2015), for instance, analyzes the human collaborative work in data analytics that is often overlooked in the face of growing "popularity of automation and statistics." He analyzes the processes used by online advertising professionals to create targeted user models, bringing to light the mundane, assumptive, and interpretive deliberation work that goes into producing such "social-culturally constituted" models. He uses this insight to describe ways in which we could better design analytical as well as relationship management software. Pine & Liboiron (2015) study the use of public health data, showing how data collection practices are deeply social in nature.

One does not simply collect "raw data." Data collection practices are shaped by values and judgments about "what is counted and what is not, what is considered the best unit of measurement, and how different things are grouped together and 'made' into a measurable entity" (ibid.: 3147). Along similar lines, Vertesi & Dourish (2011) show how the use and sharing of data in scientific collaboration depends on the contexts of production and acquisition from which such data arise. Taylor et al. (2015) show how data materializes differently in different places by and for different actors. Indeed, it is precisely the erasure of these kinds of work that produces the troubling effects of neutrality, "opacity", and self-efficacy that all too often clouds public understanding of "big data," and makes algorithms appear 'magical' in learning, but also 'real-world' environments (Burrell, 2016).

These bodies of CSCW and HCI research call attention to aspects of formalism, contingency, and discretion at the heart of data science work. Advancing an understanding of data science as a rule-*based* (as opposed to a rule-*bound*) practice, in this chapter I argue that data vision is not merely a collection of formal and mechanical rules but a discretionary and situated process, requiring data scientists to continuously straddle the competing demands of formal abstraction and empirical contingency.

2.3 Empirical Case Studies

2.3.1 Case One: ML classroom

My first case follows an instance of data science learning revealed during a ML class. At the point I pick up the story, the instructor is about to introduce a type of algorithm that classifies things into groups (called *clusters*) such that things within a cluster are 'similar' to each other and things across clusters are 'different' from each other. The instructor starts by showing an image to the students (fig. 2) and asks: *how many clusters do you see?*

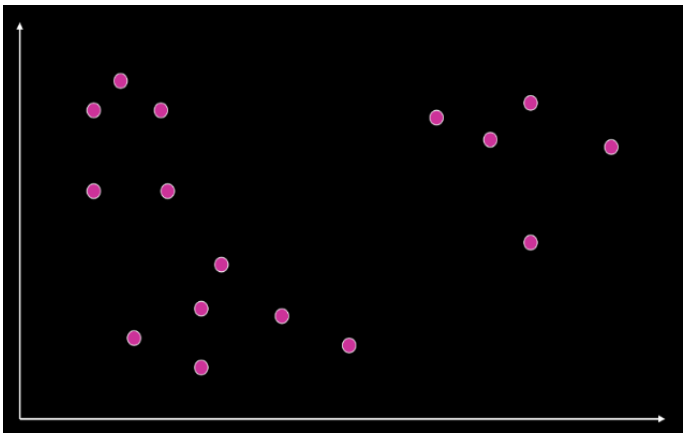

Figure 2: Class exercise to introduce the notion of 'clusters.'

Most students give the same answer: "three clusters." Anticipating the response, the instructor shows another image with three clearly marked clusters (fig. 3) and informs the students that the number of clusters in the image is unclear:

> How many clusters? I do not know. I have not even told you what the similarity measure is [i.e., how do you even know which two dots are similar to each other in this graph.] But you all somehow assumed Euclidean Distance [i.e., the closer two dots are, the more similar they are.]

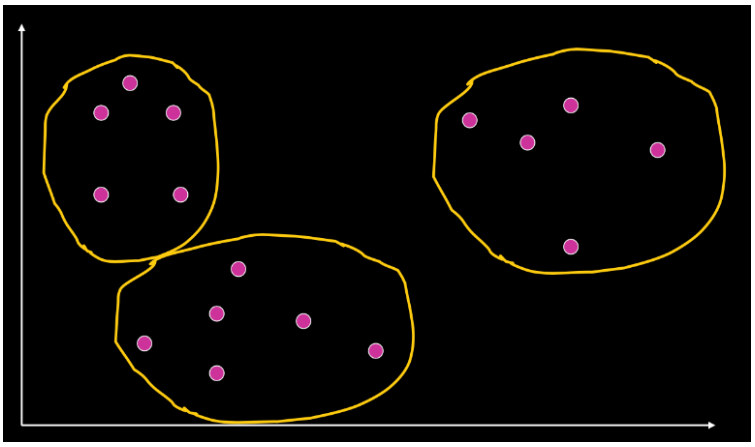

Figure 3: The three clusters that students 'saw.'

He now shows other types of clusters that could have been "seen" (fig. 4). As is clear from these images, there could have been two or three clusters. There could also have been distinct kinds of two (fig. 4a/4b) and three (fig. 4c/4d) clusters. After the students have had a chance to digest this lesson, the instructor goes on to introduce the concept of a clustering algorithm:

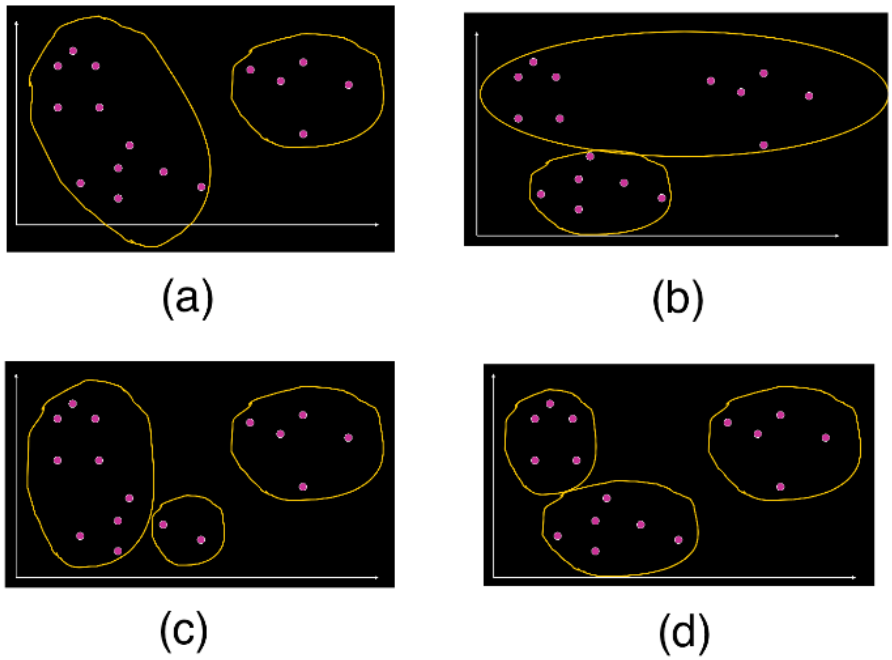

Figure 4: The different clusters that could have been 'seen.'

> A clustering algorithm does partitioning. Closer points are similar, and further away points are dissimilar. We have not yet defined what we mean exactly by similarity, but it is intuitive, right?

Having made this point, the instructor moves on to a more specific algorithm. The instructor explains that this algorithm works on a simple principle: *the similarity of two clusters is equal to the similarity of the most similar members of the two clusters*. The idea is to take a cluster (say, X), find the cluster that is most similar to it (say, Y), and then merge X and Y to make a new cluster. It is important to note that knowing the premise on which this algorithm functions is different from knowing how to apply it to data. How do we find a cluster most similar to a given cluster? What does it mean when we say, "most similar members of the two clusters"? Such questions, as we will see, are key to this algorithm's application.

The instructor now demonstrates the application of this algorithm by drawing a 2-dimensional graph marked with eight dots (figure 5a). The closer the two dots are, he explains,

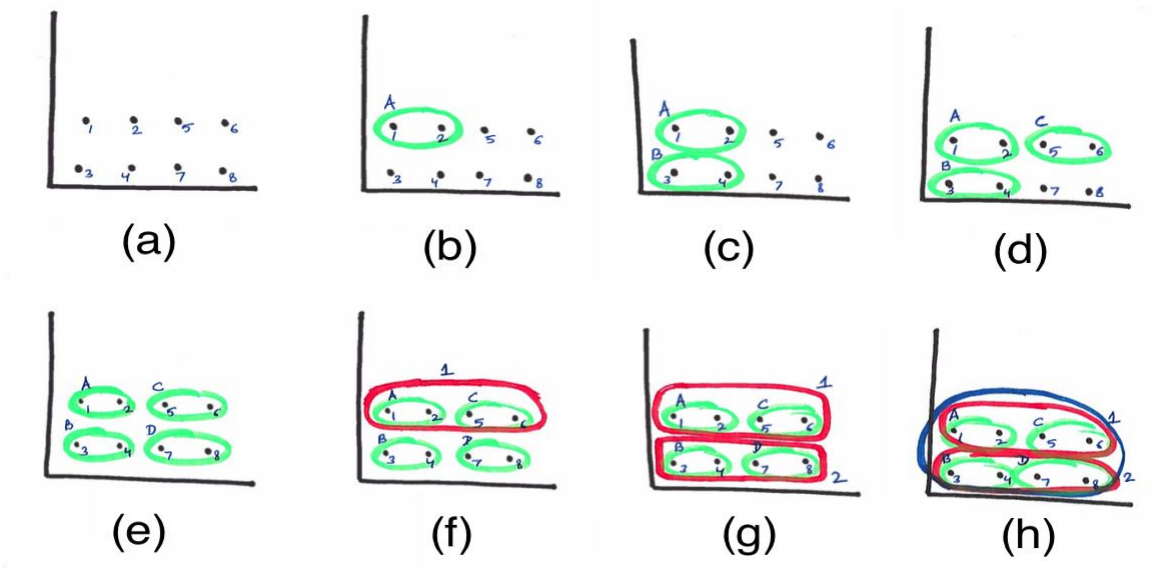

Figure 5: In-class exercise to learn a particular clustering algorithm.

the more similar they are for the purpose of this algorithm. At the start (figure 5a), there are no clusters but only a set of eight dots. The instructor tells the students that each dot will be treated initially as a cluster. He then starts to apply the algorithm beginning with dot-1. On visual inspection, the instructor and students infer that dot-1 is closer to dot-2, dot-3, and dot-4, than it is to the other dots. The instructor and the students then look again, and determine that of the three remaining points, dot-2 is the one closest to dot-1. Thus, based on the chosen similarity metric of physical distance, dot-1 and dot-2 are merged to form cluster-A (figure 5b).

The instructor now moves on to dot-3. Following the same logic, the instructor and students infer that dot-3 is closer to cluster-A and dot-4 than it is to the other dots. The instructor reminds the students that for this algorithm, two clusters are compared based on their most similar members (i.e., two dots – one in each cluster – that are closest to each other). Thus, comparing dot-3 and cluster-A, he says, means comparing dot-3 and dot-1 (as dot-1 is the dot in cluster-A that is closest to dot-3). Looking at dot-3, dot-1, and dot-4, the instructor and students infer that dot-4 is the one closest to dot-3; dot-3 and dot-4 are then merged to form cluster-B (figure 5c). In the next two steps, the instructor and students go on to dot-3 and dot-4, forming cluster-C (figure 5d) and cluster-D (figure 5e) respectively.

At this point, eight dots have been lost, and four clusters (with two dots each) gained (figure 4e). After reminding the students that comparing two clusters requires finding two dots – one in each cluster – that are closest to each other, the instructor moves on to cluster-A. A few students point out that the similarity between cluster-A and cluster-B is equivalent to the similarity between dot-1 and dot-3. Other students argue that it is equivalent to the distance between dot-2 and dot-4, as the distances between them look the same. The instructor agrees

with the students, and informs them that these distances represent the similarity between cluster-A and cluster-B. The students go on to perform the same analysis to compare cluster-A, -C, and –D.

Regarding cluster-A, the comparison is now down to three sets of distances: between (a) dot-2 and dot-4, (b) dot-2 and dot-5, and (c) dot-2 and dot-7. On visual inspection, the students observe that dot-2 is closest to dot-5. Cluster-A and cluster-C are therefore merged to form cluster-1 (figure 5f). A similar operation merges cluster-B and cluster-D to form cluster-2 (figure 5g). In the last step, cluster-1 and -2 are merged to form a single cluster containing all eight dots (figure 5h). With this, the instructor tells the students, they have reached the end of the exercise, having successfully "applied" the clustering algorithm.

There are three striking features about the in-class exercises described above. *First,* the step-by-step mechanical nature of the instructor's demonstration of the algorithm. Explicit in the algorithm's demonstration is a collection of formal rules specifying how to treat individual dots, how to compare two dots, how to compare a dot and a cluster, etc. Aspects of data vision, as we see here, are built sequentially with students learning an algorithm's application as a set of mechanical and routine steps through which data—represented as dots—are manipulated, enabling the formation of similarity clusters.

Second, the abstract nature of the represented and analyzed data. These exercises do not have a specific 'real-world' context supplementing them. The students were never told, and never inquired, what the dots and the graph represented. The dots were presented and analyzed simply as label-less dots on a nameless graph—as generic representations of all kinds of data that this algorithm can work on.

Third, the reliance on visuals to demonstrate the operation of the algorithm. We see how visual forms such as dots, circles, and graphs helped students learn to 'see' data in ways amenable to formal organization and representation. This allows students to learn to manipulate the world as a set of data points arrayed in two-dimensional space. The algorithm, it appears, 'works' so long as data is in the form of dots in n-dimensions.

While seeing and organizing the world through abstract representations and mechanical rules is key to data vision, students also need to learn to see the application of an abstract, generic method as a situated and discretionary activity. An instance of this appears in the case below.

2.3.2 Case Two: DH workshops

The second case follows the construction of data vision as revealed during a set of DH workshops. Broadly put, DH is a research area in which humanists and information scientists use algorithmic and interpretive methods to analyze data in domains such as history and literature. The vignette that follows describes how workshop conveners and students decide what dataset to work on and what happens when they begin to analyze the chosen dataset.

It has not been straightforward for the workshop conveners to decide what texts (i.e., data) the students should work on as a group not only because students have different research interests but also because not all texts are digitally available. In the first workshop session, there is a long discussion on how to get digitized version of texts (e.g., from Project Gutenberg, HathiTrust, etc.), what format to use (e.g., XML, HTML, or plain-text files), how to work with specific elements of a file (e.g., headers, tags, etc.), and how to clean the files (e.g., fixing formatting issues, removing stop-words, etc.). The students can, of course, simply download a

novel, and start reading it right away, but the point of the discussion is to find ways in which the students can make algorithms do the work of "reading".

While describing ways to convert files from one format to another, something catches the convener's eyes as he shows the students an online novel's source code. There is a vertical bar (|) in certain words such as 'over|whelming' and 'dis|tance.' At first, students suspect the digitized version has not been properly proofread. However, after noticing more words with the vertical bar symbol, the convener returns to the non-source-code version of the novel to discover that these are in fact words that cut across lines with a hyphen (-). The computer has been joining the two parts of these words with a vertical bar. At this point, a student asks about ways in which she can recognize such errors, separating "good" from "bad" data. A discussion ensues about ready-to-use scripts and packages. Several students observe that manual reading can help spot such errors, but the whole point of using algorithms is to allow work with much more text than can be read and checked in this way. The discussion ends with no clear answers in sight.

A second question concerns the dataset to be used for purposes of the common class exercises. This decision is reached only by the end of the second session: *English Gothic novels.* This choice is arrived at based on convenience rather than common interest—only one student has a research interest in Gothic literature. But a complete set of English Gothic novels in digital form is perceived to be easier to obtain than other candidates suggested by the group. "The allure of the available," as the convener remarks, "is a powerful thing." But this raises another issue: *what 'actually' qualifies as a Gothic novel?* Are gothic novels those that have the word Gothic in their title? Or those that are tagged as Gothic by the library? Or those that have been

acknowledged as Gothic by the literary community? After some discussion, the conveners and students agree to ask one of the library's digital curators to select a set of Gothic novels, and at the start of the third workshop session students are presented with plain-text files of 131 English Gothic novels.

While discussing ways in which this dataset can be used, a student inquires whether it is possible to create a separate file for each novel containing only direct quotes from characters in the novel. The workshop convener and students decide to try this out for themselves and immediately encounter a question: *how can an algorithm know what is and isn't a character quote?* After some discussion, the students decide to write a script that parses the text, inserting a section break each time a quotation mark is encountered. They surmise that this procedure will thereby capture all quotes as the text falling between sequential pairs of quotes. The total of such pairs will also indicate the number of quotes in each novel. Based on this understanding, the students create the below algorithm (in Python) to perform this work:

```
import sys
text = “”


for line in open(sys.argv[1]):
text += line.rstrip() + “ ”


quote_segments = text.split(“\””)
is_quote = False


for segment in quote_segments:
print “{0}\t{1}\{2}\n”.format(“Q” if is_quote else “N”, len(segment), segment)
## every other segment is a quote
is_quote = not is_quote
```

When tested against one of the novels in the set however the results are surprising: the script has produced just one section break. Most students feel that this result is "wrong." "Oh wow! That's it?" "I think it didn't even go through the file." "Just one quotation mark?" To see what went wrong, students scroll through the chosen novel, glancing through the first twenty paragraphs or so. Upon inspection, they conclude that there is nothing wrong with their script. It is just that the chosen novel does not have any quotes in it. (The single quotation mark that the script encountered was the result of an optical character recognition error.) This leads to a discussion of differences in writing styles between authors. A couple of students mention how some authors do not use quotation marks but a series of hyphens (-) to mark the beginning and end of character quotes. This raises a new problem. Is it safe to use quotation marks as proxies for character quotes, or should the script also look for hyphens? Are there still other variations that students will need to account for?

Out of curiosity, the students randomly open a few files to manually search for hyphens. Some authors are indeed using them in place of quotation marks:

> ------Except dimity, ------ replied my father.

Others, however, are using them to mark incomplete sentences:

> But 'tis impossible, ----

In some cases, hyphens have resulted because em-dashes (—) or en-dashes (–) were converted to hyphens by the optical character recognition system:

> Postscript--I did not tell you that Blandly…

It is now clear to the students that if hyphens sometimes mark speech, they are less robust than quotation marks as proxies for character quotes. They decide to use only quotation marks for the rest of the exercise to keep things "relatively simple."

It is now time to choose another novel to test the script. This time, the choice is not so random, as students want a novel that has many character quotes as a "good" sample or test case. The script is changed such that it now parses the text of all the novels, returning a list of novels along with the number of sections produced in each novel. These range from 0 to ~600. Since there are no pre-defined expectations for number of quotes in a novel, there is no way to just look at these numbers and know if they are accurate. However, some students still feel that something has gone "wrong." They argue that because every quote needs two quotation marks, the total number of "correct" quotation marks in a novel should be an even number. By the same logic, the number of sections produced on this basis should also be even. But the result returned shows odd numbers for almost half the novels. Students open some of these "wrong" novels to manually search for quotation marks. After trying this out on five different novels, they are puzzled. The novels do have an even number of quotation marks in them. *Why then is the script returning odd numbers?*

It does not take long to identify the problem. The students are right in saying that the number of quotation marks in a novel must be even. However, they seem to have misconstrued how the script creates sections in a novel. A student explains this by reference to one of the novels, Ann Radcliffe's *The Mysteries of Udolpho.* In the passage below, the python script will go through the text inserting four section breaks:

> She discovered in her early years a taste for works of genius; and it was St. Aubert's principle, as well as his inclination, to promote every innocent means of happiness. <> "A well-informed mind, <>" he would say, <> "is the best security against the contagion of folly and of vice." <> The vacant mind is ever on the watch for relief, and ready to plunge into error, to escape from the languor of idleness.

This example shows the students that they had been confusing *sections* with *section-breaks*. Although the script creates four section-breaks in the novel, the number of sections created by the script is, in fact, five. The students realize that the number of sections will thus be one more than the count of quotation marks. Since these will always be even, the number of sections created by the script must always be odd.

The problem has now reversed itself. Whereas earlier the participants believed that an odd number of sections was "wrong", they now agree that having an odd number of sections is actually "right". Why then, they puzzle, do some novels have an even number of sections? The participants manually check out a few "even" novels to search for quotation marks. They discover other optical character recognition errors, formatting issues, and variance in authors' writing styles that are producing "wrong," unexpected results. At the conclusion of the workshop session shortly thereafter, the students still do not have a script that can reliably extract all character quotes in an automated way.

There are several ways to explain what has happened here. One is to say that the novels were not in the "right" format—they had formatting issues, exhibited style inconsistencies, and had typographical errors. This, however, is true for most, if not all, kinds of data that analysts must deal with daily. Clean, complete, and consistent datasets—as every data scientist knows—are a theoretical fantasy. Outside of theory, data are often inconsistent

and incomplete. The requirement of prim and proper datasets, I argue, does not do justice either to the reality of the data world or to the explanation of this workshop exercise.

Another explanation is that the students simply lacked skill and experience, making what some would call "rookie mistakes." After all, these students were here to learn these methods, and were not expected to know them beforehand. However, the ability to identify and avoid "rookie mistakes" is itself an important artifact of training and professionalization. In large part, what makes a rookie a rookie is their inability to recognize and avoid these kinds of errors. As sites for learning and training, classrooms and workshops thus provide avenues for understanding how would-be professionals learn to "see" and avoid "rookie mistakes." Similar if less stark examples of such mistakes appeared in the ML class (using part of training data as a test case, confusing correlation for causation, etc.).

The workshop case brings together prior knowledge, discretionary choices, and empirical contingency. The choice of the dataset is not a given, but a compromise between thematic alignment and practical accessibility. Moreover, as seen in the case of vertical bars, hyphens, and quotation marks, data is often idiosyncratic in its own ways, necessitating situated and discretionary forms of pre-processing. Even clearly articulated computational routines (e.g., search for quotation marks, label text between marks as a section, count sections, put sections in a separate file) often require a host of situated decisions (e.g., what novels to look at, what stylistic elements to account for, how to alleviate formatting errors, how to infer and manage empirical contingency, etc.). In all these ways, algorithmically identifying and extracting character quotes is a situated activity that requires practitioners to find their way around specificities of the data at hand.

2.4 Discussion

The cases above provide important insights into the professionalization and practice of would-be data scientists. In case one, we saw how students not only learn to see data in forms amenable to algorithmic manipulation but also learn to see an algorithm's application as a collection of formal rule-like steps. The rules to be followed in applying the algorithm appear mechanical and methodical. The algorithm is demonstrated using an abstract representational form: label-less dots on name-less graphs. Whether it is discerning the similarity between two dots or knowing ways to compare and merge clusters of dots, students learn to organize and analyze the world using a fixed set of rules. Such demonstrations privilege an abstract understanding of data science, allowing students to learn to manipulate the world in actionable and predictable ways. This, I argue, is a major source of algorithmic strength: if the hallmark of real-world empirics is their richness and unpredictability, the hallmark of data science is its ability to organize and engage the world via abstract categorization and computationally actionable manipulation.

In case two, by contrast, we saw how processes of learning and practicing data science are also contingent and discretionary—in ways that abstract representations and mechanical demonstrations significantly understate. Multiple decisions were required to effectively combine the script with the given dataset ranging from identifying how to isolate character quotes and discerning how quotes appear in data to figuring out how to test the script. Unique datasets necessitate different fixes and workarounds, requiring a constant adjustment between prior knowledge, empirical contingencies, and formal methods. Making prior knowledge and abstract methods work with data is indeed hard work. Sometimes, data is not easy to find. Often,

needed data is unavailable or incomplete. Under such situations, practitioners must make do with what they can get, in ways that go against the abstracted application story usually shared in data science research papers and presentations.

Recognizing the incomplete nature of the abstracted data story helps situate an algorithm's application as a site not only for abstract categorization and formal manipulation but also for *discretion* and *creativity*. Learning to apply an algorithm, as we saw, involves a series of situated decisions to iteratively, often creatively, adapt prior knowledge, data science routines, and empirical data to each other. Elements of creativity manifest themselves as professional acts of discretion in the use of abstract, seemingly mechanical methods. While certain datasets may share similarities that support mechanical applications of rules across contexts, proficiency in operations in their mechanical form constitutes only one part of the professionalization of data analysts. Each dataset is incomplete and inconsistent in its own way, requiring situated strategies, workarounds, and fixes to make it ready and usable for data analysis. Data scientists are much like Suchman's (2007) problem solvers, Klemp et al.'s (2008) musicians, and Ingold's (2010) carpenters: constantly negotiating with and working around established routines in the face of emergent empirical diversity.

Viewing data science as an ongoing negotiation between rules and empirics helps mark a clear distinction between two ways of describing the professionalization and practice of data science that are relevant for CSCW and HCI researchers. One of these approaches data science as a rule-*bound* practice, in which data is organized and analyzed through the application of abstract and mechanical methods.

Casting data science as a rule-bound practice makes visible specific aspects of data science learning and practice. *First*, it allows data scientists to better understand the abstract nature of data analytic theories, facilitating novel ways of computationally organizing and manipulating the world. *Second*, it enables researchers to focus on constraints and limits of algorithmic analyses, providing a detailed look at some of the critical assumptions underlying data science work. *Finally*, it allows students to learn not only how to work with basic, yet foundational, data science ideas, but also how to organize and manipulate the world in predictable and actionable ways.

However, the same properties that make these aspects visible tend to render *in*-visible the empirical challenges confronting efforts to make algorithms work with data, making it difficult to account for the situated, often creative, decisions made by data analysts to conform empirical contingency to effective (and often innovative) abstraction. What is left is a stripped-down notion of data science—analytics as rules and tools—that only tells half the data analytic story, largely understating the breadth and depth of human work required to make data speak to algorithms. Significantly underappreciating the craftsmanship of data scientists, the rule-bound perspective paints a dry picture of data science—a process that often comprises of artful and innovative ways to produce novel forms of knowledge.

A more fruitful way to understand data science, I argue, is to see it not as a rule-bound but as a rule-*based* practice: structured but not fully determined by mechanical implementations of formal methods. In a rule-bound understanding, an algorithm's application requires organization and manipulation of the world through abstract constructs and mechanical rules. In a rule-based understanding, however, emergent empirical contingencies and practical

issues come to the fore, reminding us that the world requires a large amount of work for it to conform to high-level data analytic learning, expectations, and analyses. Following Feldman & Pentland's (2003) view of routines, a rule-based understanding of data science casts algorithms as ostensive as well as performative objects, highlighting how the performances of algorithms draw on and feed into their ostensive nature, and vice versa.

Seeing data science as a rule-based practice focuses our attention on the situated, discretionary, and improvisational nature of data analytics. It helps make salient not only the partial and contingent nature of the data world (i.e., data is often incomplete and inconsistent), but also the role of human decisions in aligning the world with formal assumptions and abstract representations of order as stipulated under abstract algorithmic methods and theories. Data science is a craft, and like every other form of craft it is never *bound* by rules, but only *based* on them. A rule-based understanding of data science acknowledges, and indeed celebrates, the lived differences between theoretical reality, empirical richness, and situated improvisations on the part of data scientists.

It is in and through these lived differences that data scientists gain data vision. As with Dewey's (1922), Cohen's (2007), and Feldman & Pentland's (2003) descriptions of routines and routinized tasks, we see in data vision the always-ongoing negotiation between abstract algorithmic "routines" and the situated and reflexive "applications" of such "routines." Data vision is much like an array of plans, takes, and mis-*takes* (Klemp et al., 2008)—a constant reminder of the situated and discretionary nature of the professionalization and practice of data analysis.

2.5 Implications for data science learning, research, and practice

An understanding of data vision informs data science learning, research, and practice in three basic ways. First, it helps focus attention on the role of human work in the professionalization and practice of data science as opposed to only focusing on models, algorithms, and statistics. Technicalities are important; however, focusing on the situated and collaborative nature of discretionary acts helps contextualize algorithmic knowledge, facilitating a better understanding of the mechanics, exactness, and limits of such knowledge. Algorithms and data do not produce knowledge by themselves. We produce knowledge *with* and *through* them. The notion of data vision puts humans back in the algorithm.

Second, it helps to better attend to human work in the ways in which data scientists collaboratively document, present, and write up algorithmic results. The choice of models, algorithms, and statistical tests is already an integral part of learning and presenting data science results. However, providing an explicit description of key decisions that data scientists take can not only help communicate a nuanced understanding of technical choices and algorithmic results but also enable students *and* practitioners to think through aspects of their work that may seem "non-technical" but greatly impact their knowledge claims. This helps to not only reduce the "opacity" (Burrell, 2016) of data science practices, but also better teach and communicate what some call the "black art" or "folk knowledge" (Domingos, 2012) of data science, contributing to the development of a complete and "reflective practitioner" (Schön, 1983).

Third, it helps inform professional training of as well as community conversations around data science. As described in the Introduction, in data science, and in many other forms of research (including my own!), we often present research setup, process, and results in a dry

and straightforward manner. Open and effective conversations about the messy and contingent aspects of research work—data science or otherwise—tend to escape the formal descriptions of methods sections and grant applications, reserved instead for water cooler and hallway conversations by which workarounds, 'tricks of the trade', and 'good enough' solutions are shared. The result is an excessively "neat" picture of data science that fails to communicate the real practices and contingencies by which data work proceeds.

This becomes even more difficult outside the classroom—as we will see in the rest of the chapters. In industry, research centers, and other contexts of algorithmic knowledge production, data scientists often work with huge volumes of data in multiple teams, simultaneously interfacing with a host of other actors such as off-site developers, marketers, managers, and clients. Where the results of data science meet other kinds of public choices and decisions (think contemporary debates over online tracking and surveillance, or the charismatic power of New York Times infographics) these complications—and their importance—only multiply. Data science results often travel far beyond their immediate contexts of production, taking on forms of certainty and objectivity (even magic!) that may or may not be warranted, considering the real-world conditions and operations from which they spring. Here as in other worlds of expert knowledge, "distance lends enchantment" (Collins, 1985).

More broadly, an understanding of data vision helps support the diverse forms of oft-invisible collaborative human data science work. Data science not only warrants algorithmic techniques and computational forms, but also comprises answers to crucial questions such as relations between data and question, what is answerable via available data, what are some of the underlying assumptions concerning data, methods, etc. By bringing such questions—and,

indeed, other forms of human work—to the fore, data vision directs our attention to forms of situated discretionary work enabling and facilitating data science practices. A large amount of work goes into making data speak for themselves. Data vision orients us towards identifying and building better acknowledgment and support mechanisms for sharing such folk knowledge that, though immensely useful, is often lost.

Data vision is not merely a way of perceiving the world, but a highly consequential way of seeing that turns perception into action. Data often speak specific forms of knowledge to power. Like all forms of explanation, data science has its own set of biases (Barocas, 2014; Gitelman & Jackson, 2013), assumptions (Boellstorff, 2013; boyd & Crawford, 2012; Gillespie, 2014), and consequences (Boellstorff & Maurer, 2015; Bowker, 2013, 2014). Understanding data vision allows us to better delineate and communicate the strengths as well as the limitation of such collaborative knowledge—indeed, of *seeing* the world with and through data.

2.5 Conclusion

Given our growing use of and reliance on algorithmic data analysis, an understanding of data vision is now integral to contemporary knowledge production practices, in CSCW and indeed many other fields. In this chapter I presented two distinct yet complementary aspects of learning and doing data science. I argued in favor of a rule-based, as opposed to a rule-bound, understanding of data science to introduce the concept of data vision—a notion I find fundamental to the professionalization and practice of data scientists. I described how a better understanding of data vision helps us to better grasp and value the intimate connection between methodological abstraction, empirical contingency, and situated discretion in data science

practice. Shedding light on the diverse forms of data science work, data vision produces a more open and accountable understanding of data science learning, research, and practice.

As I described in the Introduction, however, studying learning environments has its limitations. Classrooms are but one step in the professionalization of data scientists. Data science, like all practices, is a constant learning endeavor. To better understand the everyday practice of data science, we need to also study other contexts of algorithmic knowledge production such as those in the corporate sector.

As we will see in the rest of this thesis, in the corporate context data science is shaped by a much diverse set of professional expectations and business imperatives (forms of organizational work in addition to the human and technical work introduced in this chapter), further reminding us that the practice of data science remains a deeply social and collaborative accomplishment. What I have in this chapter are the first steps towards highlighting and theorizing the different types and (in)visibilities of data science work. In the next chapters I dig deeper into this realm, describing the uneven visibilities of forms of human, technical, and—in fact—organizational work in corporate data science practices.

III
Problem Formulation and Fairness

Moving out of the academic context, in the next three chapters we step into the realm of corporate data science—the world of the corporate organization DeepNetwork. As I said in the Introduction, the rest of the chapters are structured around the timeline of corporate data science projects—small windows into key moments involved in data science projects.

In this chapter, we focus on a crucial first step in such projects: problem formulation—the work of translating high-level goals into data-driven problems. Formulating data science problems is an uncertain and complicated process. It requires various forms of situated and discretionary work to translate strategic goals into tractable problems, necessitating, among other things, the identification of appropriate target variables and proxies (more on this below). These choices are rarely self-evident and often shaped by, as I show in this chapter, not just technical attributes but also human and organizational work. In this chapter I describe the complex set of actors and activities involved in problem formulation. I show that the specification and operationalization of the problem are always negotiated and elastic.

[†] A slightly edited version of this chapter has been published as Passi & Barocas (2019). I proposed the research question, refining it with feedback from Solon Barocas. I designed and conducted the ethnographic fieldwork (with guidance from Phoebe Sengers and Steve Jackson). I transcribed the interview data, analyzing the data and fieldwork notes on my own; Solon Barocas gave feedback that shaped the analysis of the empirical case presented in this chapter. I wrote the first draft and later revisions with feedback and guidance from Solon Barocas. Solon Barocas contributed significantly to section *3.1 Introduction* (subsections *3.1.1* and *3.1.2)* and subsection *3.2.1*.

In fact, I make visible how different problem formulations can raise profoundly different ethical concerns (even though normative assessments of data science projects often take problem formulations for granted). I argue that careful accounts of everyday data science work can help us to better understand *how* and *why* data science problems are posed in certain ways (rarely worked out with explicit normative considerations in mind)—and why specific formulations prevail in practice even in the face of what might seem like normatively preferable alternatives.

3.1 Introduction

Undertaking a data science project involves a series of difficult translations. As Provost and Fawcett (2013, p. 293) point out, "[b]usiness problems rarely are classification problems, or regression problems or clustering problems." They must be *made* into questions that data science can answer. Practitioners are often charged with turning amorphous goals into well-specified problems—that is, problems faithful to the original business objectives, but also problems that can be addressed by predicting the value of a variable. Often, the outcome or quality that practitioners want to predict—the 'target variable'—has not been well observed or measured in the past. In such cases, practitioners look to other variables that can act as suitable stand-ins: 'proxies.'

This process is challenging and far from linear. As Hand (1994, p. 317) argues, "establishing the mapping from the client's domain to a statistical question is one of the most difficult parts of statistical analysis," and data scientists frequently devise ways of providing answers to problems that differ from those that seemed to motivate the analysis. In most normative assessments of data science, this work of translation drops out entirely, treating the

critical task as one of interrogating properties of the resulting model. However, ethical concerns can extend to the formulation of the problem that models aim to address, not merely to whether models exhibit discriminatory effects.

To aid in hiring decisions, for example, machine learning needs to predict a specific outcome or quality of interest. One might want to use machine learning to find "good" employees to hire, but the meaning of "good" is not self-evident. Machine learning requires specific and explicit definitions, demanding that those definitions refer to something measurable. An employer might want to find personable applicants to join its sales staff, but such a quality can be difficult to specify or measure. What counts as personable? And how would employers measure it? Given the challenge of answering these questions, employers might favor a definition focused on sales figures, which they may find easier to monitor. In other words, they might define a "good" employee as the person with the highest predicted sales figures. In so doing, the problem of hiring is formulated as one of predicting applicants' sales figures, not simply identifying "good" employees.

As Barocas and Selbst (2016) demonstrate, choosing among competing target variables can affect whether a model used in hiring decisions ultimately exhibits a disparate impact. There are three reasons why this might happen. First, the target variable might be correlated with protected characteristics. In other words, an employer might focus on a quality that is distributed unevenly across the population. This alone would not constitute illegal discrimination, as the quality upon which the employer hinges its hiring decisions could be rational and defensible. But the employer could just as well choose a target variable that is a purposeful proxy for race, gender, or other protected characteristics. This would amount to a form of disparate treatment,

but one that might be difficult to establish if the decision rests on a seemingly neutral target variable. The employer could also choose a target variable that seems to serve its rational business interests but happens to generate an avoidable disparate impact—for instance, the employer could choose a different target variable that serves its business objective at least as well as the original choice while also reducing the disparate impact.

Second, the chosen target variable might be measured less accurately for certain groups. For example, arrests are often used as a proxy for crime in applications of machine learning to policing and criminal justice, even though arrests are a racially biased representation of the true incidence of crime (Lum & Isaac, 2016). In treating arrests as a reliable proxy for crime, the model learns to replicate the biased labels in its predictions. This is a particularly pernicious problem because the labeled examples in the training data serve as ground truth for the model. Specifically, the model will learn to assign labels to cases similar to those that received the label in the training data, irrespective of whether labels in the training data are accurate. Worse, evaluations of the model will likely rely on test data that were labeled using the same process, resulting in misleading reports about the model's real-world performance: these metrics would reflect the model's ability to predict the label, not the true outcome. Indeed, when the training and test data have been mislabeled in the same way, there is simply no way to know when the model is making mistakes. Choosing a target variable is therefore often a choice between outcomes of interest that are labeled more or less accurately. When these outcomes are systematically mismeasured by race, gender, or some other protected characteristic, models designed to predict them will invariably show discriminatory biases that do not show up in performance metrics.

Finally, different target variables might be more difficult to predict than others depending on the available training data and features. If the ability to predict the target variable varies by population, then the model might subject certain groups to greater errors than others.

Across all three cases whether a model ultimately violates a specific notion of fairness is often contingent on what the model is designed to predict, suggesting that we should pay far greater attention to the choice of the target variable—both because it can be a source of unfairness and a mechanism to avoid unfairness.

3.1.1 The non-obvious origins of obvious problems

This might not be surprising because some problem formulations may strike us as obviously unfair. Consider the case of 'financial-aid leveraging' in college admissions—the process by which universities calculate the best possible return for financial aid packages: the brightest students for the least amount of financial aid. To achieve this bargain, the university must predict how much each student is willing to pay to attend the university and how much of a discount would sway an applicant from competitors. In economic terms, 'financial-aid leveraging' calculates each applicant's responsiveness to price, which enables the university to make tailored offers that maximize the likely impact of financial aid on individual enrollment decisions. As Quirk (2005) explains:

> "Take a $20,000 scholarship—the full tuition for a needy student at some schools. Break it into four scholarships each for wealthier students who would probably go elsewhere without the discount but will pay the outstanding tuition if they can be lured to your school. Over four years the school will reap an extra $240,000, which can be used to buy more rich students—or gifted students who will improve the school's profile and thus its desirability and revenue."

Such strategies are in effect in schools throughout the United States, and the impact has been an increase in support for wealthier applicants at the expense of their equally qualified, but poorer peers (Wang, 2013, 2014).

One might, therefore, conclude, as Danielson (2009, p. 44) does, that "data mining technology increasingly structures recruiting to many U.S. colleges and universities," and that the technology poses a threat to such important values as equality and meritocracy. Alternatively, one could find, like Cook (2009) in a similar thought experiment, that "[t]he results would have been different if the goal were to find the most diverse student population that achieved a certain graduation rate after five years. In this case, the process was flawed fundamentally and ethically from the beginning." For Cook, agency and ethics are front loaded: a poorly formed question returns undesirable, if correct, answers. Data science might be the enabling device, but the ethical issue precedes the analysis and implementation. The goal was suspect from the start. For Danielson, however, certain ethics seem to flow from data mining itself. Data science is not merely the enabling device, but the impetus for posing certain questions. Its introduction affords new, and perhaps objectionable, ways of devising admissions strategies.

Though they are quite different, these positions are not necessarily incompatible: data science might invite certain kinds of questions, and 'financial-aid leveraging' could be one such example. One might say that data science promotes the formulation of questions that would be better left unasked. This, however, is a strangely unhelpful synthesis: it accords agency to the person or people who might formulate the problem but simultaneously imparts overwhelming influence to the affordances of data science. The effort of getting the question to work as a data

science problem drops out entirely, even though this process is where the question actually and ultimately takes shape. The issues of genuine concern—how universities arrive at a workable notion of student quality, how they decide on optimizing for competing variables (student quality, financial burden, diversity, etc.), how the results are put to work in one of many possible ways—are left largely out of view. The indeterminacy of the process, where many of the ethical issues are actually resolved, disappears.

3.1.2 Problem formulation in practice

While a focus on the work of problem formulation in real-world applied settings has the potential to make visible the plethora of actors and activities involved in data science work, it has not been the focus of much empirical inquiry to date. Researchers still know very little about the everyday practice of problem formulation. This chapter attempts to fill this gap. How and why are specific questions posed? What challenges arise and how are they resolved in everyday practice? How do actors' choices and decisions shape data science problem formulations? Answers to these questions, I argue, can help us to better understand data science as a practice, but also the origin of the qualities of a data science project that raise normative concerns. As researchers work to unpack the normative values at stake in the uses of data science, I offer an ethnographic account of a special financing project for auto lending to make visible the work of problem formulation in applied contexts. In so doing, I show how to trace the ethical implications of these systems back to the everyday challenges and routine negotiations of data science.

In the following sections, I first broadly sketch out a longer history attending to the practical dimensions of data science, specifically the task of problem formulation, and then

move on to the empirical case-study. I conclude by discussing the implications of my findings, positioning the practical work of problem formulation as an important site for normative investigation and intervention.

3.2 Background

My understanding of the role of problem formulation in data science work draws from a long line of research within the history and sociology of science that describes how scientific methods are not just tools for answering questions, but in fact influence the kind of questions we ask and the ways in which we define and measure phenomena (Collins, 1985; Joerges & Shinn, 2001; Latour & Woolgar, 1985; Pinch & Bijker, 1984). Through different methods, scientists "mathematize" (Lynch, 1985, 1988) the world in specific ways, producing representations that are both contingent *(i.e., they change with a change in methods)* and real *(i.e., they provide actionable ways to analyze the world).* Our practical understanding of a given phenomenon is contingent on the data we choose to represent and measure it with.

Let us examine a concrete example to further illustrate this point. Levin (2014), in her ethnographic work on metabolomic science (i.e., the study of the molecules and processes that make up metabolism), describes the impact of metabolomic researchers' adoption of multivariate statistics as the preferred analytic method for studying metabolism. A key goal of metabolomic research is the identification of biomarkers: "measurable and quantifiable biological entities that can be statistically determined in relation to health and disease" (ibid.: 556). Biomarkers may occur, for instance, as physical (e.g., height) or molecular (e.g., biochemical compounds) attributes. The growth in the production of metabolomic research data through sophisticated high-tech experimentation instruments made it difficult, if not

impossible, for researchers to manually analyze research data. The researchers thus, not surprisingly, came to rely on computation. Computational methods, however, require data to be represented in specific ways. A variety of computational tools transformed discrete data points from research experiments (e.g., on urine and blood tests) into relational databases with thousands of variables, represented with enormous tables, matrices, and graphs. Such interconnected representations of the metabolic processes in data shaped the researchers' choice of multivariate statistics as the *preferred* data analytic method. In contrast with univariate statistics that focuses on the measurement of a single variable, multivariate statistics measures patterns across a combination of variables. Inspecting a small set of variables as a cause (i.e., a biomarker) for a biological process thus began to be considered an "incorrect" approach by researchers—in the world of multivariate statistics, a biomarker was a combination of several interdependent biochemical compounds. The data representations, and the subsequent use of multivariate statistics, influenced the researchers' perception of the scientific "reality" of metabolism—they saw metabolism as something "inherently and utterly complex" (ibid.: 557), multiple, and dynamic as opposed to being a linear or one-dimensional process. Multivariate statistics thus became not only the preferred way of doing metabolomic science, but also the "natural" and indeed the "correct" way of analyzing metabolism, allowing researchers to analyze metabolic "complexity" in specific practical ways—e.g., using principal component analysis to identify key constituents in the biomarker for a biological process. As we see through this example, the methods used by researchers were not merely tools for answering questions but allowed researchers to ask specific questions, in turn transforming how researchers even conceptualize the nature of the phenomena.

These STS works show how tools and methods are not just ways to solve established problems, but also shape the ways we formulate questions. Different methods not only help us to organize phenomena in specific ways *(e.g.: high-dimensional relational datasets)* but also enable specific forms of analysis *(e.g.: principal component analysis).*

The emerging field of critical data studies has brought similar insights to data science: data scientists do not just apply algorithms to data but work *with* algorithms and data, iteratively and often painstakingly aligning the two together in meaningful ways. Data science work, as I argued earlier in section 2.2, is not merely a collection of formal and mechanical rules, but a situated and discretionary process requiring data analysts to continuously straddle the competing demands of formal abstraction and empirical contingency. Algorithmic results embody specific forms of data vision: rule-*based,* not rule-*bound,* applications of algorithms, necessitating judgment-driven work to apply and improvise around established methods and tools in the wake of empirical diversity.

Data science requires "thoughtful measurement, […] careful research design, […and] creative deployment of statistical techniques" (Grimmer, 2015, p. 80) to identify units of measurement, clean and process data, construct working models, and interpret quantified results (boyd & Crawford, 2012; Busch, 2014; Gitelman, 2006; Muller et al., 2016; Pasquale, 2015). Subjective decision making is necessary throughout the process. Each of these *practical* choices can have profound *ethical* implications (Currie et al., 2016; Dourish & Cruz, 2018; Introna, 2016; McFarland & McFarland, 2015; Rieder & Simon, 2016), of which data scientists are sometimes well aware. Their everyday work is shot through with "careful thinking and critical reflection" (Barocas & boyd, 2017, p. 23). Neff et al. (2017), through ethnographic work on

academic data science research, show how data scientists often "acknowledge their interpretive contributions" and "use data to surface and negotiate social values." Data, the authors argue, are the starting, and not the end, points in data science.

In academic and research settings—the contexts that inform most of our current understanding of data science—the work of data science comes across mainly as the work of data scientists. Data science projects in applied corporate settings, however, as I will show later in chapter five, are inherently collaborative endeavors—a world as much of discretion, collaboration, and aspiration as of data, numbers, and models. In such projects, several actors work together to not only make sense of data and algorithmic results but also to negotiate and resolve practical problems. Project managers, product designers, and business analysts are as much a part of applied real-world corporate data science as are data scientists.

These strands of research call attention to the role of the work of problem formulation within data science. The relationship between formulated problems and the data we choose to address them is not a one-way street—data are not merely things to answer questions with. Instead, the very formulations of data-driven problems (i.e., the kind of questions we can and do ask) are determined by contingent aspects such as what data are available, what data we consider relevant to a phenomenon, and what method we choose to process them. Problem formulation is as much an outcome of our data and methods as of our goals and objectives. Indeed, defining the data science problem is not only about making the data science process fit specific and specifiable objectives but also making the objectives fit the data science process.

Data miners have long grappled with the role of human judgment and discretion in their practice. The field of Knowledge Discovery in Databases—an important predecessor to what

we now call data science—emerged to address how choices throughout the data mining process could be formalized in so-called *process* models.

3.2.1 Knowledge Discovery in Databases (KDD)

The iterative process of applied data mining

While KDD is commonly associated with data mining and machine learning, the history of the field has less to do with innovations around these techniques than with the process that surrounds their use. Dissatisfied with a lack of applied research in artificial intelligence, scholars and practitioners founded the new sub-field to draw together experts in computer science, statistics, and data management who were interested and proficient in the practical applications of machine learning [see: 18]. The significance of this move owed to a shift in the field's professional focus, not to a change in the substance of its computational techniques. When KDD established itself as an independent field[14], it also instituted a method for applying machine learning to real-world problems—the KDD process, consisting of a set of computational techniques and specific procedures through which questions are transformed into tractable data mining problems (Fayyad et al., 1996b; Frawley et al., 1992). Although the terms KDD and data mining are now used interchangeably—if they are used at all—the original difference between the two is telling. While data mining referred exclusively to the application of machine learning algorithms, KDD referred to the overall process of reworking questions into data-driven problems, collecting and preparing relevant data, subjecting data to analysis, and interpreting and implementing results. The canon of KDD devoted extensive attention not only to the range of problems that lend themselves to machine learning but also to the multi-

[14] Semi-annual workshops started in 1989. Annual conference began in 1995.

step process by which these problems can be made into *practicable* instances of machine learning. In their seminal paper, Fayyad, Piatetsky-Shapiro, and Smyth (1996a), for example, insist on the obvious applicability of data mining while paradoxically attempting to explain and advocate how to apply it in practice—that is, how to make it applicable. KDD covered more than just a set of computational techniques; it amounted to a method for innovating and executing new applications.

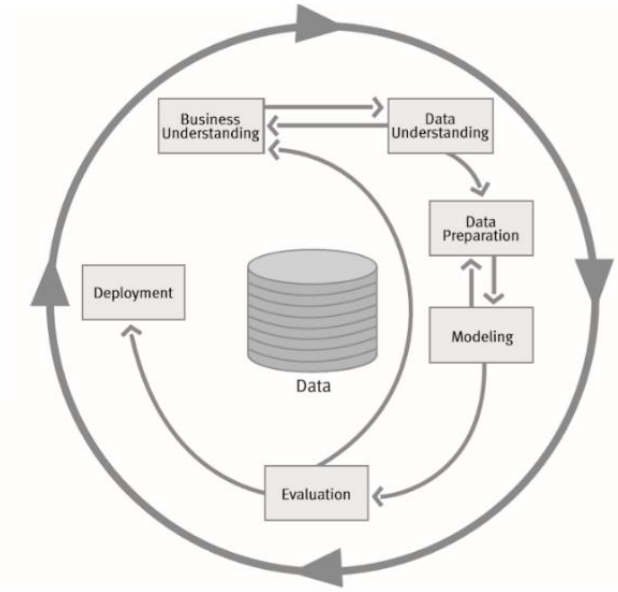

Figure 6: The Cyclic Nature of the Crisp-DM Process Model. Image taken from Chapman et al. (2000).

The focus on process led to the development of a series of *process models*—formal attempts to explicate how one progresses through a data mining project, breaking the process into discrete steps (Kurgan & Musilek, 2006). The Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000), the most widely adopted model, seems to simultaneously *describe* and *prescribe* the relevant steps in a project's lifecycle. Such an approach grows directly out of the earliest KDD writing. Fayyad, Piatetsky-Shapiro, and Smyth (1996a) make a point of saying that "data mining is a legitimate activity as long as one understands how to do it," suggesting that there is a particular way to go about mining data to ensure appropriate results. Indeed, the main impetus for developing process models were fears of mistakes, missteps, and misapplications, rather than a simple

desire to explicate what it is that data miners do. As Kurgan and Musilek (2006) explain, the push "to formally structure [data mining] as a process results from an observation of problems associated with a blind application of [data mining] methods to input data." Notably, CRISP-DM, like the earlier models that preceded it in the academic literature (Fayyad et al., 1996a, 1996b; Frawley et al., 1992), emphasized the iterative nature of the process and the need to move back and forth between steps. The attention to feedback loops and the overall dynamism of the process were made especially evident in the widely reproduced visual rendering of the process that adopted a circular form to stress cyclicality (Kurgan & Musilek, 2006).

Negotiated, not faithful, translations

Business understanding, the first step in the CRISP-DM model, is perhaps the most crucial in a data mining project because it involves the translation of an amorphous problem (a high-level objective or a business goal) into a question amenable to data mining. CRISP-DM describes this step as the process of "understanding the project objectives and requirements from a business perspective [and] then converting this knowledge into a data mining problem definition" (Chapman et al., 2000, p. 10). This process of 'conversion,' however, is underspecified in the extreme. Translating complex objectives into a data mining problem is not self-evident: "a large portion of the application effort can go into properly formulating the problem (asking the right question) rather than into optimizing the algorithmic details of a particular data-mining method" (Fayyad et al., 1996a, p. 46). Indeed, the open-endedness that characterizes such forms of translation work is often described as the 'art' of data mining (Domingos, 2012). Recourse to such terms reveals the degree to which the creativity of the translation process resists its own translation into still more specific parts and processes (i.e., it

is artistic only as far as it resists formalization). But it also highlights the importance of this first task in determining the very possibility of mining data for some purpose.

CRISP-DM and other practical guidance for data miners or data scientists [see: 6,26] tend to describe problem formulation mainly as part of a project's *first phase*—an initial occasion for frank conversations between the managers who set strategic business goals, the technologists that manage an organization's data, and the analysts that ultimately work on data. Predictably, those involved in data science work face the difficult challenge of *faithful translation*—finding the correct mapping between, say, corporate goals, organizational data, and computational problems. Practitioners themselves have long recognized that even when project members reach consensus in formulating the problem, it is a *negotiated* translation—contingent on the discretionary judgments of various actors and further impacted by the choice of methods, instruments, and data.

These insights speak to the conditions that motivate data science projects in a way that escapes the kind of technological determinism or data imperative that pervades the current discourses—as if the kinds of questions that data science can answer are always already evident. Getting the automation of machine learning to return the desired results paradoxically involves an enormous amount of manual work and subjective judgment (Barocas & Selbst, 2016). The work of problem formulation—of iteratively translating between strategic goals and tractable problems—is anything but self-evident, implicated with several practical and organizational aspects of data science work. As Hand points out, "[t]extbook descriptions of data mining tools […] and articles extolling the potential gains to be achieved by applying data mining techniques gloss over [these] difficulties" (Hand, 2006, p. 8).

In the following two sections, I look at the work of data science that is traditionally glossed over, showing how the initial problem formulation consists of a series of *elastic* translations—placeholder articulations susceptible to change as the project progresses.

3.3 Empirical Case Study: Special Financing

CarCorp, a DeepNetwork subsidiary, collects *special financing* data: information on people who need car financing but have either low/bad credit scores (between 300-600) or limited credit histories. The company's clientele mainly consists of auto dealers who pay to receive this data (called *lead data*) that include information such as name, address, mortgage, and employment details (sometimes even the make of the desired automobile). The company collects lead data primarily online: people who need special financing submit their data so that interested dealers can contact them. People requiring special financing face several challenges ranging from the lack of knowledge about available credit sources to difficulties in negotiating interest rates. As liaisons between borrowers and lenders, companies such as CarCorp and its affiliates act as important, sometimes necessary, intermediaries for people requiring special financing. CarCorp serves several dealers across the country.[15] Few dealers collect their own lead data as the money, effort, and technical skills required to do so is enormous. This is a key reason why dealers pay companies such as CarCorp to buy lead data.

[15] The exact number is omitted to preserve company anonymity.

CarCorp's technology development and project manager Bart[16] wanted to leverage data science to "improve the quality" of leads. Improving lead quality, Bart argued, will ensure that existing dealers do not churn (i.e., they continue to give their business to CarCorp).

> **Bart (project manager):** "The main goal [is] to improve the quality of our leads for our customers. We want to give actionable leads. […] That is what helps us make money, makes customers continue to use our services" *(Interview, November 1, 2017)*.

Initial discussions between the business and data science teams revolved around two themes: (a) defining lead "quality" and (b) finding ways to measure it. Defining lead quality was not straightforward. There were "many stakeholders with different opinions about leads" *(ibid.)*. Some described lead quality as a function of a lead's salary data, while some argued that a lead was good if the dealer had the lead's desired car in their inventory. Everyone on the business team, however, agreed on one thing—as CarCorp's business analyst Ray[17] put it: a "good" lead provided business to the dealer.

> **Ray (business analyst):** "The business team has been talking about [lead quality] for a long time. […] We have narrowed down the lead quality problem to how likely is someone to purchase or to be able to finance a car when you send them to that dealer?" (Interview, November 8, 2017).

Lead "quality" was equated with lead "financeability." It was, however, difficult to ascertain financeability. Different dealers had different special financing approval processes. A lead

[16] Bart has been with CarCorp for ~4 years. Before this, he had 14+ years of work experience in the technology industry in several roles such as business development director, technology and design manager, and data analyst.

[17] Ray has an undergraduate degree in economics and a graduate degree in applied statistics. He has been with DeepNetwork for 2+ years. Before this, he worked as a Research Assistant (Statistics) in an academic research center.

financeable for one dealer can be, for several reasons, unfinanceable for another. The goal thus was to determine *dealer-specific financeability* (i.e., predicting which dealer was most likely to finance a lead). The teams settled on the following definition of 'quality': *a good lead for a dealer was a lead financeable for that dealer*. This, in turn, framed the problem as one of *matching leads to dealers that were most likely to finance them.*

CarCorp had a large amount of historical lead data. In 2017 alone, the company had processed close to two million leads. CarCorp, however, had relatively less data on *which* leads had been approved for special financing (let alone data on *why* a lead was approved). The business team asked the data science team to contact data engineers to identify and assess the relevant data sources. The data science team, after investigating the data, however, declared that there was not enough data on dealer decisions—without sufficient data, they argued, it was impossible to match leads with dealers. Few dealers in CarCorp's network shared their approval data with the company. The scarcity of this data stemmed from the challenge of creating business incentives for dealers to provide up-to-date feedback. The incentives for dealers to share information about their approval process with CarCorp were too attenuated.

While the data science team instructed the business team to invest in the collection of up-to-date data on dealer decisions, further discussions ensued between the two teams to find alternate ways to predict dealer-specific financeability using the data that happened to be available already. In debates over the utility of the available data, business analysts and data scientists voiced several concerns ranging from inconsistency (e.g., discrepancies in data values) to unreliability (e.g., distrust of data sources). Business analyst Ray, for instance, was wary of the multiple ways in which the data was collected and generated:

> **Ray (business analyst):** "The entire complex lead ecosystem […] to outsiders does not make any sense. […Some data] came from an affiliate. […] They give a credit score range for a bunch of those leads. So, not exactly the score, but they could say 'this person is between 450-475, and 525-550,' different buckets. [...] Never realistic, but we pay money, and we have this data. [...] We [also] generate leads organically [online], then there are leads that we [buy from third parties]. […] Different pieces of information [are] appended to those leads depending on where it came from" (Interview, November 8, 2017).

Only a few CarCorp affiliates augmented lead data with additional information such as credit scores. Dealers run background checks on leads (with their consent) as part of the financing procedure and in the process get a lead's exact credit score from credit bureaus such as Equifax. The Fair Credit Reporting Act (FCRA) prohibits CarCorp from getting a lead's exact credit score from credit bureaus without a lead's explicit consent. Leads have no reason to authorize CarCorp to retrieve their credit data because the company does not make lending decisions; it only collects information about a lead's interest in special financing. CarCorp had to rely on either leads' self-reported credit scores that were collected by a few affiliates or on credit scores provided as part of lead data bought from third-party lending agencies.[18] Business affiliates and third-party agencies provide credit scores in the form of an approximate range (e.g., 476-525). CarCorp had hoped that this data would help them to, for example, differentiate between a subset of leads that appeared identical but exhibited different financeability.

> **Ray (business analyst):** "Two individuals [with] the same age, same income, same housing payment, same everything […] could have wildly different credit scores. [...] You have those two, and you send them to the same dealer. From our perspective, lead A and B are […] maybe not exactly [the] same, but close. [...] But, the dealer can finance person A, and they cannot finance person B [...] So, when they [dealers] are evaluating […] whether they would renew their product with us, if we had sent them a bunch from bucket B, and none from A, they are likely to churn. But we have no way

[18] Third-party lending agencies offer services to help put borrowers in touch with multiple lenders. These agencies get consent from people to perform a soft pull on their official credit reports from credit bureaus such as Equifax.

> of knowing who is in bucket A and who is in bucket B. […] You can have two individuals who measure the same on data points, and they have two different credit scores." (Interview, November 8, 2017).

It is not surprising that a lead with a credit score greater than another lead is more likely to secure special financing (even when the leads are otherwise identical). Credit score data is a significant factor in special financing approval processes. CarCorp had this data for about ~10% of their leads (~100,000). While business analysts found it challenging to predict credit score ranges from the other features using traditional statistical analyses, adding credit score ranges in as an additional feature did improve the company's ability to predict lead financeability. The business team wondered if it were possible to use data science to predict credit score ranges for the remaining 90% of the leads (i.e., perhaps machine learning could work where traditional analysis had failed). If successful, credit scores could help ascertain lead financeability—*a financeable lead was a lead with a high credit score.*

The data science team's attempt to assess if they could predict credit scores, however, faced a practical challenge. As mentioned above, credit score data received from affiliates took the form of ranges (e.g., 476-525), not discrete numbers. Different affiliates marked ranges differently. For example, one affiliate may categorize ranges as 476-525, 526-575, etc., while another as 451-500, 501-550, etc. It was not possible to directly use the ranges as class labels for training as the ranges overlapped. The data science team first needed to reconcile different ranges.

As data scientist Max[19] started working to make the ranges consistent, business analysts found a way to make this process easier. Pre-existing market analysis (and, to some extent, word-of-mouth business wisdom) indicated that having a credit score higher than 500 greatly increased a lead's likelihood of obtaining special financing approval. This piece of information had a significant impact on the project. With 500 as the crucial threshold, only two credit score ranges were now significant: *below*-500 and *above*-500. Max did not need to figure out ways to reconcile ranges such as 376-425 and 401-450 but could bundle them in the *below*-500 category. The *above*-500 credit score range could act as the measure of financeability—*a financeable lead was a lead with a credit score above-500.* The matching problem *(which leads are likely to get financed by a dealer)* was now a classification task *(which leads have a credit score of over 500).* Decreasing the number of classes to two helped attenuate the difficulty of reconciling different ranges but did not help to circumvent it.

> **Max (data scientist):** If the credit score is below 500, the dealer will simply kill the deal. […] The problem is there are too many records in the 476-525 range. […] This makes it difficult. (Fieldwork Notes, June 13, 2017).

Leads in the 476-525 range were an issue because this range contained not only *below*-500 leads, but also *above*-500 leads. Making mistakes close to the decision boundary is especially consequential for special financing where you want to find people just above the threshold. Max tried many ways to segregate the leads in this range but the models, according to him, did not work. Their accuracy, at best, was slightly better than a coin flip. Max attributed the model's

[19] Max has an undergraduate degree in information and computer science and a graduate degree in mathematics and statistics. He has been with DeepNetwork for a little less than two years. Before this, he worked as a Statistical Consultant.

bad performance not only to the presence of leads in the 476-525 range, but also to the limited number of available features (i.e., the data did not present sufficient information for the model to meaningfully differentiate between leads). While the in-house lead dataset was a good starting point, the data scientists knew that accurately classifying leads would require not only creative ways to work with available data, but also a wide variety of data. They had already been scouting for external datasets to augment the in-house lead dataset.

Director of data science Martin[20] had tasked data science project manager Daniel[21] with the work of finding third-party datasets that could help with classification. Their approach was first to use freely available data to see "how far we get" before buying paid data from companies such as Experian.[22] Free, yet reliable, datasets, however, were hard to come by. Daniel found only a few datasets—Internal Revenue Service (IRS) zip-code-level data on features such as income range, tax returns, and house affordability (i.e., how much an owner might be able to pay for a property). Data scientist Max tried each dataset but declared that none improved model performance. Members of the data science team wondered if it was worth investing in high-quality paid datasets.

[20] Martin has an undergraduate degree in electronics and electrical engineering, and graduate degrees in computer science and business administration. He has been with DeepNetwork for 5+ years. Before this, he had 25+ years of experience in the technology industry in several roles such as vice-president engineering, director engineering, and research engineer.

[21] Daniel has an undergraduate and a graduate degree in computer science. He worked at DeepNetwork for 2.5 years (he left in late 2017 to join as a process manager in an academic institute). Before this, he worked as a Technology Consultant. Outside the technological domain, he has eight years of work experience in industrial warehousing and public relations.

[22] For example, datasets such as Experian's Premier Aggregated Credit Statistics (https://www.experian.com/consumer-information/premier-aggregated-credit-statistics). Even such data only contained aggregated information on credit scores and ranges, and not lead-specific credit scores (which required consent).

> **Max:** I used all the data, but the model does not converge.
>
> **Daniel:** What about Experian data? We can get it if you think it will help.
>
> **Max:** We will have to buy it first for me to know if it helps with prediction.
>
> **Martin (director of data science):** Check out the online info on it, and then you and Daniel can figure it out (Fieldwork Notes, June 16, 2017).

Without access to the data, Max argued, it was not possible to clearly know its usefulness. If the data was not going to be helpful, however, it made no sense to buy it—the decision needed to be made based on the available description on Experian's website. They eventually ended up not investing in it. Even after analyzing the dataset's description, Max was doubtful that the data would increase model performance. Two months later, the project was halted in the absence of actionable progress.

Different actors justified the project's failure in separate ways. Data scientist Max blamed the data. Business analyst Ray felt that perhaps the business team unreasonably expected "magic" from data science. For him, the culprit was the nature of the problem itself:

> **Ray (business analyst):** "[It is a] selection-bias problem. We are not dealing with a random sample of the population. […] These individuals [...] why are they submitting a lead with us? Because it was not easy for them to get financed. By definition […] our population is people with at least not great credit, and usually bad credit. Why do you have bad credit? There is like one reason why you have good credit. There are a thousand reasons why you have […] bad credit. [...] If you show me someone with good credit, I will show you they pay bills on time, have a steady income, etc. If you show me someone with bad credit, they can have a gambling problem, they can be divorced, they could […] prove income now, but maybe it has been unstable in the past, and we have no way of knowing that. There are literally thousands of reasons that aren't that capture-able" (Interview, November 8, 2017).

Business analyst Ray described the failure not in terms of the initially articulated business goal but in terms of the project's current data science problem formulation—not the difficulty of defining the quality of a lead, but the challenge of classifying leads with scores in a specific part of the credit score spectrum. He believed it was possible to classify people with high/low credit scores on the full 300-850 credit score spectrum (e.g., differentiating between a person with a 750 score and a person with a 450 score). He argued, however, that CarCorp's focus on the special financing population meant that the goal was not to classify high/low scores on the full credit spectrum but to demarcate between *different kinds* of low scores on one side of the spectrum (roughly between 300 and 600). Note how Ray, a business analyst, describes the project's failure in relation to a *specific* definition of lead "quality"—it was difficult to know which leads were above or below the credit score threshold of 500. The project was halted when developing an accurate model based on this definition proved impossible. The business and data science team could not figure out any other way to formulate the problem at this stage with the data they had.

3.4 Discussion & Implications

The above description shows how the data science problem was formulated differently at different points in the project based on the two sets of targets variables and their possible proxies.

Proxy #1: Dealer Decisions. The business team initially described the project goal as the *improvement of lead quality*—a formulation of what the business team thought the dealers wanted. Note that this goal was in turn related to the broader objective of *minimizing churn rate*—a formulation of what CarCorp itself wanted. In this way, the problem specification was

just as much about keeping clients as it was about satisfying their business objectives. These high-level goals impacted the actors' initial understanding of the project's goal—the quality of leads was seen in relation to dealers and CarCorp's own success. CarCorp decided that if dealers could finance a lead, it was a good lead. The fact that different dealers had different special financing approval processes further impacted the contingent relationship between quality, dealers, and financeability: if a lead was financeable by a specific dealer, it was a good lead for *that* dealer. The data science problem, therefore, became the task of *matching* leads with dealers that were likely to finance them.

Proxy #2: Credit Score Ranges. Data available to support the use of dealer decisions as a proxy, however, were limited. While business analysts did not fully understand how dealers made decisions, they acknowledged, based on market research, the import of credit scores in the special financing approval process—leads with scores higher than 500 were highly likely to get special financing. Credit scores thus became a proxy for a dealer's decision, which was itself a proxy for a lead's financeability, which was, by extension, a proxy for a lead's quality—indeed, a *chain* of proxies. The data science problem thus became the task of *classifying* leads into below- and above-500 classes.

3.4.1 Problem formulation is a negotiated translation

At face value, the relationship between the project's high-level business goal (improving lead quality) and its two different problem formulations (the two sets of target variables and their proxies) may seem like a one-to-many relation—different translations of, in effect, the *same* goal. Such an understanding, however, fails to account not only for the amorphous nature of high-level goals (i.e., the difficulty of defining the quality of a lead), but also for the project's

iterative and evolving nature (i.e., problem formulations are negotiated, dependent on, for instance, actors' choice of proxy). In our case, actors equated (in order): lead quality with financeability, financeability with dealer decisions, and dealer decisions with credit score ranges. Each such maneuver produced different formulations of the objective, in turn shaping actors' articulation and understanding of the project's high-level goal (as seen, for instance, in the way business analyst Ray ultimately accounts for the project's failure).

This is not to argue that the high-level goal to improve lead quality, at some point, transformed into a *completely different* objective. Instead, it shows that the translation between high-level goals and tractable data science problems is not a given but a negotiated outcome—stable yet elastic. Throughout the project, the goal of improving lead quality remains recognizably similar but *practically* different, evident in the different descriptions, target variables, and proxies for lead quality. Each set of target variables and proxies represents a specific understanding of what a lead's quality is and what it means to improve on it. The quality of a lead is not a preexisting variable waiting to be measured, but an artifact of how our actors define and measure it.

3.4.2 The values at stake in problem formulation

Scholars concerned with bias in computer systems have long stressed the need to consider the original objectives or goals that motivate a project, apart from any form of bias that may creep into the system during its development and implementation (Friedman & Nissenbaum, 1996). On this account, the apparent problem to which data science is a solution determines whether it happens to serve morally defensible ends. Information systems can be no less biased than the objectives they serve.

These goals, however, rarely emerge ready-formed or precisely specified. Instead, navigating the vagaries of the data science process requires reconceiving the problem at hand and making it one that data and algorithms can help solve. In the empirical case, we do not observe a data science project working in the service of an established goal, about which there might be some normative debate. Instead, we find that the normative implications of the project evolve alongside changes in the different problem formulations of lead quality.

On the one hand, for the proxy of dealer-decisions, leads are categorized by their dealer-specific financeability—a lead is only sent to the dealer that is likely to finance them. In formulating the problem as a matching task, the company is essentially *catering* to dealer preferences. This approach will recommend leads to dealers that align with the preferences expressed in dealers' previous decisions. In this case, lead financeability operates on a spectrum. Financeability emerges as a more/less attribute: each lead is financeable, some more than others depending on the dealer. Effectively, each lead has at least a chance of being sent to a dealer (i.e., the dealer with the highest probability of financing a lead above some threshold).[23]

On the other hand, for the credit-score proxy, leads are categorized into two classes based on their credit score ranges and only leads with scores greater than 500 are considered financeable. In formulating the problem as the task of classifying leads above or below a score, the company reifies credit scores as the sole marker for financeability. Even if dealers had in the past financed leads with credit scores less than 500, this approach only recommends leads

[23] Of course, depending on the threshold, some leads will never be sent.

with scores higher than 500, shaping dealers' future financing practices. In this approach, financeability operates as a binary variable: a lead is financeable only if its credit score is higher than 500. Consequently, leads in the below-500 category may never see the light of day, discounted entirely because the company believes that these leads are not suitable for dealers.

3.4.3 Different principles; different normative concerns

Seen this way, the matching version of the problem may appear normatively *preferable* to the classification version. But is this always true? If we prioritize maximizing a person's lending opportunities, the matching formulation may seem better because it increases a lead's chances of securing special financing. If, however, we prioritize the goal of mitigating existing biases in lending practices (i.e., of alleviating existing dealer biases), the classification problem formulation may come across as the better alternative because it potentially encourages dealers to consider leads different from those they have financed in the past. Through the two scenarios, we see how proxies are not merely ways to equate goals with data but serve to frame the problem in subtly different ways—and raise different ethical concerns as a result.

It is far from obvious which of the two concerns is more serious and thus which choice is normatively preferable because shifting our normative lens alters our perception of fairness concerning the choice of target variables and proxies. In this chapter, I have demonstrated how approaching the work of problem formulation as an important site for investigation enables us to have a much more careful discussion about our own normative commitments. This, in turn, provides insights into how researchers and practitioners can ensure that projects align with those commitments.

3.4.4 Always imperfect; always partial

Translating strategic goals into tractable problems is a labored and challenging process. Such translations do necessary violence to the world that they attempt to model, but also provide actionable and novel ways to address complex problems. My intention to make visible the elasticity and multiplicity of such translations was thus not to criticize actors' inability to *find* the perfectly faithful translation. Quite the opposite: I recognize that translations are always imperfect and partial, and wanted to instead shift the attention to the consequences of different translations and the everyday judgments that drive them.

My actors, however, did not explicitly debate the ethical implications of their own systems—neither in the way we, as researchers, have come to recognize normative issues, nor in the way I, as an author, have analyzed the implications of their problem formulations. Practical and organizational aspects such as business requirements, the choice of proxies, the nature of the algorithmic task, and the availability of data impact problem formulations in much more significant and actionable ways than, for instance, practitioners' normative commitments and beliefs. Indeed, my analysis of the empirical case makes visible how aspects such as analytic uncertainty and financial cost impact problem formulations. For example, the prohibitive cost of datasets coupled with the challenge of assessing the data's efficacy without using it made it particularly challenging for actors to leverage additional sources of information.

Yet, as I show in this chapter, normative implications of data science systems do *in fact* find their roots in problem formulation work—the discretionary judgments and practical work involved in translations between high-level goals and tractable problems. Each translation galvanizes a distinct set of actors, aspirations, and practices, and, in doing so, creates

opportunities and challenges for normative intervention—upstream sites for downstream change. As Barocas et al. (2017) argue:

> "A robust understanding of the ethical use of data-driven systems needs substantial focus on the possible threats to civil rights that may result from the formulation of the problem. Such threats are insidious, because problem formulation is iterative. Many decisions are made early and quickly, before there is any notion that the effort will lead to a successful system, and only rarely are prior problem-formulation decisions revisited with a critical eye."

If we wish to take seriously the work of unpacking the normative implications of data science systems and of intervening in their development to ensure greater fairness, we need to find ways to identify, address, and accommodate the iterative and less visible work of formulating data science problems—*how* and *why* problems are formulated in specific ways.

3.5 Conclusion

In this chapter, I focused on the uncertain process by which certain questions come to be posed in real-world applied data science projects. I made visible how some of the most important normative implications of data science systems find their roots in the work of problem formulation. Attempts to make certain goals amenable to data science always involve subtle transformations of those objectives along the way—alterations that may have profound consequences for the very conception of the problem to which data science has been brought to bear and what consequently appear to be the most appropriate ways of handling those problems. The problems we solve with data science are thus never insulated from the larger process of getting data science to return actionable results. As I described, these ends are very much an artifact of a contingent process of arriving at a successful formulation of the problem, and they cannot be easily decoupled from the process at arriving at these ends. In linking the normative

concerns that data science has provoked to more nuanced accounts of the on-the-ground process of undertaking a data science project, I have suggested new objects for investigation and intervention: *which goals are posed and why; how goals are made into tractable questions and working problems;* and *how and why certain problem formulations succeed.*

In this chapter, we saw how during problem formulation, an essential first step, practitioners outline the system's intended working in the service of given goals. Even after problem formulation, however, practitioners' conceptions of how, whether, or in what ways their systems work remain in flux. In the next chapter, I unpack the ongoing forms of human and organizational work involved in corporate projects to show how much like a data science system's problem formulation, its working is also not stable but remains *in the making* throughout the project.

IV
Making Data Science Systems Work

Building data science systems is a laborious process, requiring extensive amounts of technical work.[24] Unsurprisingly, dominant narratives about the working of such systems—what work they do, how they work, and how to assess their working—remain technology centered, comprising formal accounts of algorithmic steps or performance metrics. Still, although models and numbers are vital to the process, building data science systems is a sociotechnical endeavor that requires not only technical but also human work (Baumer, 2017; Dourish & Cruz, 2018). Thus, understanding how these systems work requires also clarifying the human work this entails.

A good example, as we saw in the previous chapter, is the work of problem formulation (Hand, 1994)—translating high-level goals into data-driven problems. In the previous chapter, we learned that the expected working of a data science system is not given but negotiated during problem formulation through discretionary judgments of various actors and affected by choice of methods, instruments, and data. "Even the simplest piece of software has embedded within

† A slightly edited version of this chapter has been published as Passi & Sengers (2020). For this work, I proposed the initial research question, refining it with feedback from Phoebe Sengers. I designed and conducted the ethnographic study (with feedback and guidance from Phoebe Sengers and Steve Jackson). I transcribed the interview data, analyzed the data and fieldwork notes on my own; Phoebe Sengers provided regular feedback that shaped the analysis of the empirical case presented in this chapter. I wrote the first draft and subsequent revisions with feedback and guidance from Phoebe Sengers.

[24] My use of the term "system" refers to complex underlying sociotechnical assemblages, consisting of algorithms, code, hardware, models, software, and large amounts of "extra-algorithmic" (Cohn, 2019) human-organizational work (Paine & Lee, 2017; Vertesi, 2019).

it a series of architectural decisions about what 'works' regarding the purposes for which it was created" (Shaw, 2015, p. 2).

4.1 Introduction

In this chapter, I unpack the work involved in corporate projects, arguing how and why data science systems do not simply work on their own. Building a working data science system is much more than the seemingly simple task of technically implementing its problem formulation. Instead, as I highlight, practitioners *make* data science systems in some ways and not in others through ongoing and situated forms of human and organizational work. I situate 'working' as a system's ability to work *as intended* from the perspective of the practitioners who build the system. I use phrases such as 'the system now needed to work differently' to call out changes in practitioners' expectations of the system's working, highlighting changes done to align system working with shifting expectations. As exercises in "collective authorship" (Seaver, 2019, p. 418), building data science systems—and *making them work*—requires enormous subjective judgment.

In fact, I show that even determining what aspects of a system work or do not work is not always obvious or numerically determinable. One reason for this is that a system's working is multifaceted. The system works or does not work in distinct ways for different actors. Data scientists, for instance, often describe a system's working via performance metrics (Rieder & Simon, 2016). Numbers, however, remain tightly coupled with those aspects of working "that are most readily computationally quantifiable" (Baumer, 2017). Project managers, however, define working through the lens of business use cases, and product managers prioritize compatibility and feasibility as essential aspects of working—articulations embedded in wider

organizational imperatives. “Mathiness” (Lipton & Steinhardt, 2018) is but one feature of a data science system whose eventual working is a “practical accomplishment” (Garfinkel, 1967).

This chapter addresses the question: *how are data science systems made to work?* Unpacking this question is especially important with the growing impact of data science in virtually every sphere of modern life. When the (wrong)doings of systems become visible, system builders may argue that their systems are, in fact, not at fault—*they were not designed to work this way!* Such misalignments between systems’ intended and situated working are commonplace. One reason for this is that the perspective of people facing the most significant effects from a system’s output is often not included in the system’s design and development process (Binns et al., 2018). This absence creates a “de-coupling” between how system builders engineer, users use, and data subjects experience systems (Baumer, 2017). In such situations, without a clear understanding of the sociotechnical work involved in building data science systems, it is challenging to locate actionable sites of accountability and intervention. Critical data studies researchers strive to unpack the implications of data science systems, ranging from how they shape professional practices (Dudhwala & Larsen, 2019) and knowledge production (Bechmann & Bowker, 2019; Miles, 2019) to how they enable surveillance (Aradau & Blanke, 2015) and marginalization (Buolamwini & Gebru, 2018). Such research efforts aim to not only make visible the implications of data science systems but also shape their design and development practices.

Engaging with the design of data science systems, however, is challenging. Opening the data science black box is neither easy nor straightforward. Analyzing such systems requires multiple forms of knowledge. Access to data science practitioners, particularly those working

in corporations, is restricted. Critical data studies scholarship on the working of data science systems thus centers on their use, rather than their design—their impact in the way they work as opposed to *how* and *why* practitioners build them to work the way they do. As researchers, we know that building technological systems is a non-linear process, requiring extensive amounts of collaborative, discretionary, and situated work (L. Suchman et al., 2002; L. A. Suchman, 1987; L. A. Suchman et al., 1999). However, regarding data science, as we continue to focus on the implications of data science systems, we know little about *how* and *why* practitioners build such systems to work in some ways and not in others. If we, as researchers, have stakes in how data science systems *should* work, we must attempt to understand how these systems are *made* to work.

In this chapter I address this gap, contributing to a growing body of scholarship on data science system design practices (e.g., Amershi, Begel, et al., 2019; Muller, Lange, et al., 2019; Saltz & Grady, 2017; Saltz & Shamshurin, 2015). I describe how corporate data science practitioners negotiate central aspects of a system's working: what work the system should do, how the system should work, and how to assess whether the system works. These negotiations, as you will learn in this chapter, lay the foundation for how, why, and to what extent a system ultimately works the way it does. I focus on a specific corporate data science project—a self-help legal chatbot—to highlight these negotiations and the invisible human work that goes into determining whether and how systems work. Through a detailed recounting of corporate data science in action, I develop a more general account of how a system's working is not only affected by algorithmic and technical design but also continually reworked in the face of competing technologies, implementation challenges, and business goals.

In the following sections, I first describe the empirical case study and then explain my findings, finally concluding by describing ways for the field of critical data studies to move forward based on them.

4.2 Empirical Case Study: Self-help Legal Chatbot[25]

Law&You, a DeepNetwork subsidiary, offers online marketing services to attorneys, lawyers, and law firms. Clients pay to integrate Law&You's digital tools into their websites to convert website visitors into paying customers. One such tool is an online chat module that connects users with human chat operators, who guide users towards self-help legal resources or collect information on users who need professional legal services. If a user needs professional help, the operator collects data such as the user's name, address, and contact, and forwards it to the client.

In late 2016, Law&You started replacing human chat operators with automated guided chats. A guided chat is a scripted list of options presented to the user one at a time. Depending on the user's selection, guided chat moves to the next set of options. Guided chat generates its initial options based on, for instance, keyword analysis. If the user's request is, for example, "I want to file for bankruptcy," guided chat will identify the keyword 'bankruptcy' and provide three options: *Personal Bankruptcy, Corporate Bankruptcy,* and *Something Else*.

Law&You, however, faced two challenges with guided chats. First, *users described legal issues in multiple ways.* Although "I want to file for bankruptcy" is one way to describe

[25] In this chapter, I focus on one project for its salience to the chapter's theme, but I observed similar dynamics across other projects. The chatbot project was in its initial stage when I began fieldwork. I was only an observer in this project.

bankruptcy, “I do not have money” and “I have debt” are also valid bankruptcy descriptions. It was impossible to hard code all the ways in which people describe legal issues. Second, *users often went off script.* Instead of selecting an option, users often responded with open-ended text or asked questions. Follow-up questions and transferring users to human operators clarified such situations but Law&You felt that repeated questioning and transfers lowered user and client confidence in the product. Guided chat’s inability to handle such situations meant that users often left chats midway, providing no data at all. This was unacceptable since the collection of user data was a business priority. Law&You lost out on profitable data, spending further resources to clean the messy data captured by guided chat.

In May 2017, Law&You’s director of technology Paul[26] approached DeepNetwork’s director of data science Martin with the idea of a smart chatbot. Law&You’s move to chatbots was partly a response to the fact that several of their competitors now provided AI-driven marketing tools. The chatbot, from Paul’s perspective, was the future of digital marketing. The data science team had never built a chatbot before but had experience with natural language processing (NLP) tools. The team had previously developed NLP-based systems using in-house models and third-party services provided by companies such as Amazon, IBM, and Microsoft. At the beginning of the project, as part of requirement gathering, Martin asked Paul for a definition of the chatbot’s use case—*what was it supposed to do?* In a later meeting, Paul restated the chatbot’s business use case:

[26] Paul has an undergraduate degree in CS. He has been with DeepNetwork for ~1 year. Prior to DeepNetwork, Paul has ~20 years of work experience in the technology industry in positions ranging from a software engineer and technology consultant to tech advisor and CTO.

> **Paul (director of technology):** People in legal chat do not always follow a script. We want to find ways to anticipate conversational pathways. We are agnostic to technology as long as it provides value—user information. (Fieldwork Notes, July 31, 2017)

Guided chat's failure to perfectly determine *what* users say hindered the business goal of data collection. A chatbot could help overcome this challenge. Given this business use case, as Martin described later, the data science team saw its task as building a chatbot to guide open-ended conversations:

> **Martin (director of data science):** "[We wanted to go] beyond guided chat. Guided chats are easy because they are guided […] to the point where it is a request-and-response kind of scripted conversation. The challenge was—how to go from a guided chat to, what I would call, an open-ended chat. […] If [users] went off rails, can [we] somehow guide them back to the original conversation and let them not give up on the conversation? […] Try to get them back to a valuable conversation. An open-ended conversation can get out of control very quickly" (Interview, October 26, 2017).

The data science team broke down the chatbot's functionality into two tasks: (1) "knowing*"* what users say (identifying what users want to accomplish with the chatbot) and (2) *"*guiding*"* conversations (helping users by making the chatbot talk to them in a human-like manner).[27] Initially, data scientists focused on the first task—determining what users say. When they researched existing chatbots, they found that third-party AI platforms powered most chatbots.

[27] "Knowing" and "guiding" were terms used by participants. The term "guiding" was unanimously used by everyone, but the use of the term "knowing" deserves clarification. Data scientists initially described the "knowing" task as that of "identifying user intentions." This usage, however, diminished over time. The chatbot's working was anthropomorphized, especially by business members. The chatbot was at times "rude," "smart," and "childish," had "abilities," and could be "told" things. This, in part, led participants, even data scientists, to adopt the vocabulary of "knowing." This speaks to professional challenges in interdisciplinary collaborations in which certain groups adopt the vocabulary used by other, often senior, groups.

In earlier projects, the data science team had "tried" most third-party AI platforms; given their experiences, they favored VocabX[28] and considered it "smarter than most systems."

The data science team believed that building a VocabX-enabled chatbot required the use of several VocabX services working in tandem: (1) *Grasp*—a service to analyze unstructured text to extract topics and keywords, (2) *Erudition*—a service to train VocabX on a specific domain (it comes pre-trained on several topics), and (3) *Colloquy*—a service to enable VocabX to hold conversations. Given their previous experience with VocabX, data scientists stated that using VocabX to identify legal topics should be "relatively easy" using two services—parse the text through *Grasp* and pass the output to *Erudition*. Right out of the box, VocabX "already knew a lot" about several topics, further bolstering the team's confidence. Data science team members often shared examples of VocabX's "success" in team meetings:

> **Martin (director of data science):** We fed VocabX a line deliberately trying to confuse it. We wrote, 'I am thinking about chapter 13 in Boston divorce filing.' VocabX figured out the two topics:
>
> 1. business and industrial / company / bankruptcy
> 2. society / social institution / divorce (Fieldwork Notes, May 31, 2017).[29]

They considered the line "confusing" because it had keywords for both divorce and bankruptcy[30]. Such attempts to confuse the chatbot[31] were commonplace, considered an

[28] Product-specific names are replaced with pseudonyms. VocabX is an AI platform.

[29] Each item also has a relevancy score (not shown).

[30] Put broadly, Chapter 13—as part of United States Bankruptcy Code—allows individuals with a regular income to propose plans for handling their finances while under bankruptcy court's protection.

[31] From here on, the term 'chatbot' replaces 'VocabX-enabled chatbot.'

informal, yet reasonable, form of testing. The chatbot "knew" what users said if it identified the "correct" topics.

The more significant challenge was the second task—guiding conversations. When users went off track, the data science team wanted the chatbot to guide back the conversation in a "natural" manner. Data scientists tried but were unsuccessful in making the chatbot hold conversations. The chatbot could identify legal topics but data scientists could not successfully configure *Grasp, Erudition,* and *Colloquy* services to work together. Director of data science Martin mentioned that perhaps the team did not "fully understand" VocabX. He believed it was possible to configure VocabX for open-ended legal conversations and organized a meeting with VocabX representatives to resolve the problem.[32] Three data scientists, data science project manager Daniel, director of data science Martin, two VocabX representatives, director of technology Paul, and DeepNetwork's CTO Justin[33] attended the meeting. The meeting's goal was to figure out how to configure VocabX for open-ended conversations. In the meeting, however, another problem emerged when Martin asked for a specific demonstration.

> **Martin:** Can you give a demo or an example of VocabX in action, particularly when a user query goes outside the scope of the conversation?
>
> **VocabX representative #1:** You mean when the query hits outside the bounds of the algorithm?

[32] VocabX representatives were not data scientists or machine learning engineers. Vendor presentations are often done by consultants and sales representatives.

[33] Justin has been with DeepNetwork for 12+ years (10+ years as CTO and 2+ years as Vice President, Technology). Before this, he had 7+ years of work experience in technology industry in roles such as engineering manager and senior engineer.

> **Martin:** Yes. If the user is, let us say, talking about legal stuff and then suddenly types 'oh, what's the weather in Chicago?' That kind of thing. How to bring the conversation back on track? Guiding it in the direction we want.
>
> **VocabX representative #1:** What you are asking for is an in-practice thing. With VocabX you can configure different ways of saying the same thing. So, if you encounter something that is below a certain confidence interval threshold or outside the confidence boundary, that is maybe out of scope. You can also call *Equals* [another VocabX service] for misspelled words. You can tell the system that 'piza,' 'pizzza,' and 'pizzzza' are all just 'pizza.'
>
> **Martin:** Yes, for entities, but can we do it for intents as well?
>
> **Justin (CTO):** Remind me–what is the difference between entities and intents?
>
> **Martin:** Intent is the end goal, and entities are the small parts that make you reach that goal. For example, 'order a pizza' is an intent, and things such as 'large,' 'pepperoni,' etc. are entities of the intent. We can give a minimum of six instances of entities to tell the VocabX model about the entity range of an intent, but ideally it should not take that many if the system is capable of NLP, which I think it is.
>
> **Paul (director of technology):** If you had enough entities, figuring out intent should not be hard (Fieldwork Notes, June 5, 2017).

Martin asked how the chatbot would work when users went "outside the scope of the conversation." The data science team often discussed the two tasks (*knowing* what users say and *guiding* conversations) separately but recognized that they overlapped: to *guide* users, you must *know* what they said. As data scientist Alex[34] put it: 'the bot has to know why you are on the platform!' The task of 'knowing' required discerning not only what a user said but also what

[34] Alex has an undergraduate and master's degree in CS. He has been with DeepNetwork for a little over a year. Prior to joining here, Alex has ~2 years of work experience in the technology industry primarily working as a software engineer. Alex's focus was on model deployment—i.e., working with engineering teams to ensure that built models were successfully deployed and integrated into existing business infrastructure.

they left unsaid. As Martin had described a few days before the meeting: ‘it is almost like we want to read someone’s mind.’

The meeting surfaced a pertinent challenge. Guiding conversations required knowing what a user says *as well as* identifying whether what they say is on or off track within the context of the ongoing conversation. It was possible to configure the chatbot to identify themes related to a topic by manually inputting intent-entity combinations. Intent refers to users’ desired goals—the reason they are talking to the chatbot (e.g., to get legal advice on divorce). Entities refer to themes within users’ intended goals—unique aspects of what users want (e.g., to get advice on annulment). The chatbot could identify that entities such as ‘marriage,’ ‘annulment,’ and ‘custody’ are all linked to divorce. If users asked about ‘custody,’ the chatbot would use configured confidence thresholds to determine that users were on track in a legal conversation on divorce.

The situation grew complex when users said things that were not already specified as entities and intents. For example, when users describe love for their kids in divorce conversations. Pretrained on general topics, the chatbot can identify that the user is perhaps on the ‘family’ or ‘parenting’ topic. If the chatbot does not identify that these topics are also related to divorce, then it might incorrectly determine that the user is off track. Law&You wanted to avoid such situations that could make users feel that they were not understood or taken seriously, making them drop conversations midway. For the legal use case, the chatbot needed to identify *whether* a topic was within or outside legal discussions. Identifying on- and off-track topics required either manually inputting all intent-entity relations or the chatbot to learn intent-

entity relations on its own. The former was not even a possibility—it was after all the very problem with guided chats. CTO Justin asked if the latter was possible.

> **Justin:** I am happy to give it six entities to populate it if it is then able to figure out the next 50 entities on its own. Is that possible? We want to get intent. If the user says: 'I have no money, and I have to sell my car,' we want to figure out that the user is talking about bankruptcy. We can do that manually. We can have a button that says: 'click here to file for bankruptcy.' But, instead of having preprogrammed buckets, asking people to self-identify, we want to extract it directly out of the conversation.
>
> **VocabX representative #2:** You can use Wisdom [a VocabX product] to manually label your data and train the system to identify domain-specific things (Fieldwork Notes, June 5, 2017).

Director of technology Paul and CTO Justin turned down manual labeling, arguing that it incurred high financial and personnel costs. They could allocate resources to manually label a small set of entities, but not all of them.

The data science team organized the meeting to know how to "guide" conversations. However, much of the meeting centered on the work of identifying legal issues—i.e., "knowing" what users say—that the data science team believed it had already accomplished. The team realized that the full scope of the task of knowing what users say includes knowing the difference between "really off-track" users and on-track users "describing things differently."

My description of the meeting may seem incomplete, consisting only of discussions among business personnel—Martin, Paul, and Justin—and VocabX representatives. Although data scientists were at the meeting, they are absent in my description. During fieldwork, I observed similar dynamics between data scientists and business personnel. Data scientists were

vocal in data science team meetings (even in the presence of their director or project manager). However, they were far less vocal in meetings with business personnel, especially senior personnel. For example, in requirement gathering or status update meetings with business teams, Martin and Daniel often spoke on behalf of the team.

The same dynamic was at play in the VocabX meeting. One reason for this, as I learned through discussions with data scientists, lies in their understanding of the goal of such meetings. It may seem that the meeting's goal was technical—to learn how to configure VocabX to hold open-ended conversations. Data scientists, however, felt differently. For them, the meeting's goal was primarily business in nature. They recognized that if VocabX could power the chatbot, the company would need to invest a substantial amount of money into VocabX to get access to their cloud infrastructure to handle the thousands of customers that would use the chatbot daily. The meeting's technical and business goals were intertwined—a common occurrence in the world of corporate data science.

Going back to my story, the meeting surfaced additional problems but resolved some previous ones. Data scientists now better understood what they could accomplish with VocabX (e.g., identify preconfigured intents and entities) and what they could not (e.g., discern whether new entities were part of legal intents)—as discussed in the debrief session after the meeting. They decided that they needed to figure out a way to make the chatbot automatically learn legal intent-entity relations. The team discussed two ways to achieve this. First, using VocabX's *Erudition* service, which provided a way to learn new things, to train the chatbot on data containing legal books and articles. Second, developing in-house NLP models trained on the same data. The former prioritized compatibility with VocabX; the latter enabled greater control

over learning. Irrespective of the choice, the immediate task was to create a training dataset. Law books and articles, however, describe legal issues in legal vocabulary. It was possible to use them as training data to train models, but only as these topics appeared in legal terminology. This data, however, was inadequate to learn how users colloquially describe legal issues, a point raised in the meeting. The practical difficulty of curating data on people's everyday descriptions of legal issues posed challenges for the learning task. Data scientists attempted to resolve this issue by also including archived chat transcripts with human operators and forum discussions as part of their training data. Besides VocabX, the chatbot would now also use in-house models.

The data science team presumed that VocabX *could* easily learn new intent-entity relations (just like their earlier assumption that VocabX *could* hold open-ended conversations). The final verdict, however, was against VocabX, as described later by director of data science Martin:

> **Martin:** "My expectation of their [VocabX] service was a little bit higher than what they were actually delivering. The reason I started looking at VocabX is because I was hoping that going beyond guided chat… that their technology had progressed to the point where they can help me with the less guided or open-ended chat. That is when we started asking questions to VocabX team—does your technology actually help in that regard? What ended up happening was that they acknowledged that their Colloquy service really was not… <*pause*> I should not have expected their service to do that. Instead, they offered yet another service [Wisdom]. Even that [is] not that smart. You still have limitations." (Interview, October 26, 2017).

A few weeks later, the data science team provided an update on the chatbot's development. Martin, data science project manager Samuel,[35] data scientists Max and Alex, Law&You's director of technology Paul, and Law&You's software engineer Richard attended the meeting.

[35] Samuel was hired to replace Daniel, who left the company in July.

Martin told Paul that the chatbot was a work in progress. It performed better than before but was far from ready for deployment.[36] There were "too many edge cases" in which conversations did not go as planned. Paul, however, was not convinced that the chatbot's performance was as bad as Martin described. For him, the chatbot just had to be good enough.

> **Paul:** Maybe we need to think about it like an 80/20 rule. In some cases, it works well, but for some, it is harder. 80% of the time everything is fine, and in the remaining 20%, we try to do our best.
>
> **Martin:** The trouble is how to automatically recognize what is 80 and what is 20.
>
> **Paul:** I agree. Let us focus on that. We just want value. Tech is secondary.
>
> **Max:** It is harder than it sounds. *(Martin laughs, Paul asks Max: 'In what way?').* One of the models I have is a matching model trained on pairs of legal questions and answers. 60,000 of them. It seems large but is small for machine learning.
>
> **Paul:** That is a lot. Can it answer a question about say visa renewal?
>
> **Max:** If there exists a question like that in training data, then yes. But with just 60,000, the model can easily overfit, and then for anything outside, it would just fail.
>
> **Paul:** I see what you are saying. Edge cases are interesting from an academic perspective, but from a business perspective the first and foremost thing is value. You are trying to solve an interesting problem. I get it. But I feel that you may have already solved it enough to gain business value (Fieldwork Notes, July 31, 2017).

For the data science team, the chatbot was better than before but still far from perfect. Paul did not require perfection. The chatbot had business value, even if it worked 80% of the time. Paul differentiated between academic and business perspectives. Edge cases posed exciting data

[36] The team's confidence stemmed from their testing of the chatbot with, what they considered, hard test cases. This included texts with several legal topics and with off-track queries. Archived chat transcripts between users and human operators also provided quantitative assessment benchmarks.

science challenges. Solving them, however, was outside the project's scope. The business gained value from a good-enough chatbot even if it was not ideal from a computational perspective.

It is not surprising that imperfect, good-enough systems can provide business value. What was surprising was that Paul argued that the chatbot's failures were, in fact, not failures at all!

> **Paul:** Edge cases are important, but the end goal is user information, monetizing user data. We are building a legal self-help chatbot, but a major business use case is to tell people: 'here, talk to this lawyer.' We *do* want to connect them with a lawyer. Even for 20%, when our bot fails, we tell users that the problem cannot be done through self-help. Let us get you a lawyer, right? That is what we wanted in the first place (Fieldwork Notes, July 31, 2017).

For Paul, the primary goal was to collect user data and sell it to clients. If the chatbot did this most of the time, it worked. It was acceptable to fail, and failures did not mean that the chatbot did not work. In such cases, the chatbot can inform users that their legal problem (the problem it failed to identify) is not a self-help problem and requires professional help. The users should thus provide their information so that the company could put them in touch with lawyers. There were no failures, only data.

4.3 Findings

I began the chapter with the question: *how are data science systems made to work?* In this section, I answer this question by examining how actors negotiated central aspects of the chatbot's working. First, *what work should the chatbot do?* I show how existing technologies shape the chatbot's intended working. Second, *how should the chatbot work?* I show how the resolution to the challenge of identifying on- and off-track users influences the way actors expected the chatbot to work. Third, *how to assess whether the chatbot works?* I show how

actors evaluate the chatbot in distinct ways and finally agree to assess its working by skewing the balance between business and data science goals. Making visible the ongoing forms of discretionary work essential to building data science systems, I complement existing critical data studies research with a detailed account of the human work required to make data science systems work.

4.3.1 Existing technologies: The Old and the New

The working of data science systems is not just an artifact of their technical features but entangled with existing technologies. In this subsection, I analyze how two existing technologies—one *other than* the chatbot and one *making up* the chatbot—shaped the chatbot's intended working.

First, assessments regarding whether a technology other than the chatbot—guided chat—worked shaped actors' articulations of the chatbot intended working. Director of technology Paul initially described the chatbot as different from guided chat. The chatbot was a novel technology, signifying the company's move towards AI. The chatbot's initial problem formulation, however, was motivated by the perceived performance of guided chat. The guided chat *successfully* increased profit margins by reducing the cost of hiring human chat operators but *unsuccessfully* collected reliable data because of problems with scripted chats.

The company could revert to hiring chat operators, but this was undesirable. Hiring back human operators would incur a high financial cost. The company would also lose market share to competitors who already provided AI-driven digital marketing tools. Reverting to human operators would mean that to remedy guided chat's failure (unreliable data collection), Law&You would also have to give up on guided chat's success (higher profit margin). The

company instead went for the chatbot that promised reliable data collection *and* maintained, if not increased, profit margins and market share. What guided chat *could or could not* do shaped what the chatbot *should or should not* do. If users followed guided chats, the company would not even require a chatbot, at least for data collection. The *need* for the chatbot to anticipate conversational pathways, to know what users say, and to guide them were founded in the perceived poor performance of guided chat.

Second, an important technical feature of a technology making up the chatbot shaped actors' understanding of users' legal issues, affecting how actors expected the chatbot to work. For my actors, users had legal intents, which were combinations of distinct legal entities. The chatbot needed to determine intent by identifying entities. *How did actors settle on this understanding?* The answer lies within VocabX's technical design in which texts are analyzed *as* combinations of intents and entities. For VocabX, the meaning of a piece of text is equal to the text's intent-entity makeup. My actors' understanding of legal issues directly mirrored VocabX's technical design[37], enormously impacting the chatbot's intended working. The work of identifying what users say became the work of identifying intents and entities. In doing so, actors also configured a particular type of user (Woolgar, 1991) in their system—a user with apparent intentions, which they described using recognizable terms.

4.3.2 Emergent Challenges: Situated Resolutions and System Working

The working of data science systems is shaped not only by problem formulation and algorithmic design but also by situated forms of work to resolve emergent implementation challenges. In

[37] Most third-party AI platforms use this approach.

this subsection, I examine how the resolution to the problem of identifying *on-* and *off-track* users shaped how actors expected the chatbot to work.

The data science team initially equated the work of knowing what users say with identifying legal intents. In doing so, the team made two assumptions. First, users described legal issues—i.e., users were on track. This assumption was apparent in that the text used to test the chatbot contained valid accounts of legal issues (e.g., bankruptcy and divorce). Second, users used recognizable legal words in their descriptions. This assumption was evident in the use of specific keywords (e.g., chapter 13, divorce, and child custody) in test cases. The chatbot worked through the correct identification of legal intents of on-track users who described issues using recognizable legal words.

The VocabX meeting, however, surfaced additional challenges. The meeting's goal was to configure the chatbot to *guide* users (actors assumed that the work of *knowing* what users say was already done). Director of data science Martin asked how to guide a user asking about the weather. Everyone agreed that this query was outside legal discussions. Through this query, Martin invoked an exemplary instance of an off-track user who said things "completely unrelated" to law. Guiding, even identifying, such off-track users required the chatbot to perform additional work to identify what is or is not a valid part of legal discussions.

Actors resolved this challenge by proposing that the chatbot must identify *all* kinds of on-track users. The chatbot should learn the distinct ways in which users remain on track in legal discussions. If the chatbot did this successfully, it could identify off-track users since their queries would not correspond to the chatbot's model of on-track users. Recognizing the many kinds of on-track users, however, required the chatbot to work differently. In fact, the chatbot

needed to perform multiple kinds of work: identifying the *content* of legal topics (e.g., is annulment a part of the divorce topic?), mapping *relations* between legal topics (e.g., are bankruptcy and divorce connected?), and discerning the *scope* of legal discussions (e.g., is caring for kids a part of a divorce discussion?). One way the chatbot could learn legal content, relations, and scope was through training data prepared by manually labeling legal texts. Doing so incurred high financial and personnel costs, and this was not how the actors wanted the chatbot to work. It needed to learn on its own—a requirement made difficult by differences between formal and colloquial descriptions of legal issues. In the end, the chatbot needed the technical ability to identify the content of *and* relations among legal topics *besides* accounting for the scope of legal discussions *to differentiate* between on- and off-track users. At face value, this change to the chatbot's working might seem like a mere redefinition of its working rather than an actual change. However, it is crucial to note that this redefinition consequentially altered the chatbot's technical setup and working.

4.3.3 Negotiated Balance: Business and Data Science Considerations

Whether a data science system works is neither obvious nor given (for a more general account, see: Collins, 1985; Rooksby et al., 2009; L. Suchman et al., 2002); the perceived success and failure of its working depend as much on business goals as on data science imperatives. In this subsection, I unpack how actors evaluated the chatbot in distinct ways, agreeing to assess its working in a practical way founded in a negotiated yet skewed balance between business and data science goals.

Business and data science actors evaluated the chatbot in different, somewhat divergent, ways. The data science team focused on assessing the algorithmic efficacy of the chatbot's

working. From this perspective, the chatbot was far from perfect because of its inability to account for several edge cases. The data science team's fixation on scoping and resolving edge cases was not an arbitrary choice. An essential part of director of data science Martin's project goal was that the chatbot should know when users went off track and guide them back. It should not come as a surprise that most, if not all, edge cases were off-track user queries—queries at the heart of the data science team's assessment criteria.

For the business team, the chatbot's assessment depended on the practical efficacy of its working. Director of technology Paul argued that solving edge cases was an interesting academic challenge but not the project's business goal—the chatbot needed to work for most, *not all,* cases. This articulation of the difference between academic challenges and business goals is in line with recent work, e.g., Wolf (2019), on how industry practitioners perceive differences between applied and scholarly data science. For Paul, the chatbot's success did not just depend on its algorithmic prowess. The chatbot was also a competitive tool. A good-enough chatbot was already a huge success, signaling the company's uptake of AI-driven technologies to clients and competitors.

My finding that practitioners often need systems to work in good-enough, and not perfect, ways echo similar findings concerning other systems (Gabrys et al., 2016; Keller, 2000; Lewis et al., 2012). However, what is surprising is how the business team reframed the situations which the data science team considered as failures as potential sites of success. The chatbot's computational failures were, for the business team, a result of the complexity of users' legal issues and not of the chatbot's technical inadequacies. In such cases, the chatbot could inform users that their legal issue required professional help—*I cannot help you because your*

issue is not a self-help issue, thus give me your information, and I will connect you with lawyers. The chatbot worked 100% of the time from a business perspective. Paul's 80/20 rule became the new ground for assessment, establishing an accountable definition of a successful chatbot. The data science team could disagree with this new assessment criteria but not entirely reject it, especially given that the data science team described its central organizational role as that of "adding value" to businesses—an aspect I observed in this and many other projects.

4.4 Discussion

In this chapter, I described how actors make data science systems work relative to existing technologies, emergent challenges, and business goals. Enormous amounts of discretionary, often invisible work lay the foundation for how, why, and to what extent systems ultimately work.

One way to explain this is to see the chatbot's final working as a mundane consequence of data science practice. From this perspective, a plausible narrative of the chatbot project would be that a problem (anticipating conversational pathways) hinders a business goal (data collection). Data scientists break down the problem (identify topics, learn relations among topics) and build a chatbot to solve it (by knowing what users say and guiding them). The chatbot's working is thus a foregone conclusion—always stable, merely realized through development.

This framing, however, does not account for the choices and decisions that alter the working of systems throughout development—sometimes in ways that are invisible even to practitioners. Building systems—indeed, making systems work—requires "ordering the natural, social, and software worlds at the same time" (Vertesi, 2019, p. 388). Business goals

and existing technologies shape problem formulations. The design of existing technologies configures the work systems must do and assumptions about how users will interact with the systems. Considerations of financial cost, personnel time, and resource allocation lead actors to require systems to do specific work in automated ways. The working of data science systems is not just an account of their technical design but made *account-able* (Garfinkel, 1967) through ongoing work by many practitioners (Neyland, 2016a).

Researchers continue to analyze data science implications, recommending how systems should or should not work. Researcher's ability to effectively address responsible design requires understanding and addressing how and why practitioners choose to make systems work in specific ways. Through a detailed description of the negotiated nature of the working of data science systems, my empirical findings call attention to the consequential relationship between the working of data science systems and the everyday practices of building them, highlighting three takeaways for critical data studies.

First, practitioners have different, sometimes divergent, assumptions and expectations of system's intended working. We saw how data science and business team members differed in their approach to evaluating the chatbot's working, highlighting underlying differences in their understanding of how the chatbot was intended to work. Data science is as much the process of managing different expectations, goals, and imperatives as of working with data, algorithms, models, and numbers. Building data science systems requires work by many practitioners, but not all forms of work are considered equal. We saw how business team members had more power than data scientists in the chatbot project (Saltz and Shamshurin (2015) point to similar dynamics at another company). The organizational culture at Aurelion

reinforced the notion that the data science team's job was to "add value" to business products. Data science teams remain one of the most recent additions to corporate organizations. But their everyday work intersects with already-existing project, product, and business practices—with teams that often have more weight in organizational decisions.

In making visible the impact of organizational aspects, I suggest researchers orient toward how differences among practitioners and teams are managed, resolved, or sidelined within projects. Who gets to take part in negotiating a system's working? Who decides the nature of this participation? Who gets to arbitrate the outcome of negotiations? In studying the implications of data science systems, we also must engage with the culture and structure of organizations that design such systems to understand how specific viewpoints are (de)legitimized by practitioners (Haraway, 1988; Harding, 2001). This engagement can help us better understand the entangled relationship between the organizational context and product development practices of corporate organizations (Boltanski & Thévenot, 2006; Reynaud, 2005; Stark, 2009).

The second takeaway concerns itself with new empirical sites for analyzing the work of building data science systems. The challenging nature of gaining access to corporate practice continues to limit critical data studies scholarship. Data science systems, however, do not exist in isolation but are embedded in wider sociotechnical ecosystems. Justifications concerning the working of systems may lie, as I have shown, not within but outside them—in the technologies that systems replace or the technologies that make up systems. We must keep in mind that existing technologies that data science systems replace—even those that seem to have nothing to do with data science—also shape how practitioners envision what data science systems can

or cannot, and should or should not, do. What existing practices and technologies do data science systems augment or replace? In what ways do existing technologies seem deficient or superior to data science systems? Engaging with these questions is useful, helping us see how data science systems intersect with existing ways of doing and knowing.

As I have shown, third-party AI platforms make up and shape the working of data science systems in important ways. As we continue to work to gain access to corporate data science practices, analyzing these systems promises to provide meaningful insights into how, why, and to what extent systems work the way they do. How do third-party AI services define and solve common problems such as identifying user intent? What are the affordances and limits of different AI platforms? Describing how AI platforms work and what futures they promise to their clients falls within the purview of our efforts to unpack the working and implications of data science systems.

Third, a data science system's working is impacted as much by how practitioners anticipate the kinds of work the system can perform as by the actual work of practitioners to build the system. Initially, my actors believed that the chatbot could successfully solve guided chat problems. In two other instances, data science team members expected the chatbot to easily recognize legal intent, guide conversations, and learn new information. Beyond the project's immediate goals, business actors believed that the chatbot was the future of digital marketing. Such forms of 'anticipation work' (Steinhardt & Jackson, 2015) shaped how actors imagined viable approaches and solutions to problems at hand, further affecting how and why actors built the chatbot to work the way it did. In our case, however, anticipations often faltered, requiring

work by actors to adjust their programs of action. Situated in the present, articulations of plausible futures consequentially shape everyday data science work.

Proclaiming the efficacy of actions before performing them is not the mistake of putting the cart before the horse but, in fact, an essential attribute of professional work. Pinch et al. (1996) describe how professionals use forms of anticipation to decide what is and is not doable, imagining and selecting among viable actions when faced with uncertainty. Examples of such anticipatory skills range from the ability of mechanics to provide accurate time estimates for repairs to the ability of aerospace engineers to assess the safety of a rocket before flight. Similarly, data science practitioners must know not only *how* to build systems to work in specific ways but also *whether* certain forms of working are possible. Like all professionals, data science practitioners often think and act "in the future perfect tense" (Schuetz, 1943: 40):

> "We cannot find out which of the alternatives will lead to the desired end without imagining […an] act as already accomplished. […We] have to place ourselves mentally in a future state of affairs which we consider as already realized, though to realize it would be the end of our contemplated action."

For critical data studies researchers, analyzing and participating in the anticipatory work of practitioners is crucial because it is through such forms of work that practitioners imagine possible futures—worlds in which systems can, do, and must work in some ways and not in others. As researchers, our efforts to assist practitioners in designing better systems thus must include the work of fostering alternative, even somewhat radical, imaginations of futures to help practitioners envision new possibilities. We get the systems we imagine, but not necessarily the ones we need. The work of building better systems begins with working with practitioners to imagine better futures.

I realize that such engagements will sometimes frustrate both critical data studies researchers and data science practitioners; the two may, and often do, have different normative goals. Data science practitioners may cater to a different set of ethics, caring more about the clients they are in relationship with than about the concerns espoused by critical data studies. Sometimes critical data studies researchers may argue that it is better not to build a system. Data science practitioners may still go ahead with it because of the perceived business value or because they believe that if they do not build it, someone else will. In such situations, it may seem better to not engage with practitioners at all.

My aim is not to simplify the lives of practitioners and researchers and create a binary divide between them. *Both* practitioners and researchers are entangled in the current data science moment in complex ways. Still, I want to make visible what I believe is increasingly becoming a challenge within critical data studies scholarship: calls, such as mine, to engage with practitioners are often seen as futile (at best) or appalling (at worst)—to put it at its most extreme, *how could you work with, and not against, these evildoers?*

I strongly believe that we need more research on the capitalistic underpinnings and negative implications of data science—certainly, you *do not* need to always *work with* practitioners to do important and valuable research. But I am troubled when critical data studies researchers seem to treat ethical values and normative goals as always stable *a priori* frameworks that just need to be implemented in systems. It is almost as if practitioners and researchers are thought to have no ethics of their own. Not only do ethics exist in all practices, but they are also often worked out as part of everyday work. What is normatively better depends as much on people's normative stance as on the practical judgments that drive their everyday

work. Practitioners and researchers struggle with different sets of constraining forces, but it is important to be reflexive and remember that *both* groups perceive and act on the world through constraints that shape what they believe to be good and possible. The goal of engaging with practitioners is thus neither to school them nor to do their bidding. Instead, my call to work *with* data science practitioners is best understood as embarking on a difficult journey to learn more about the situated nature of data science practice and research, making visible the differences and similarities in our normative goals.

4.5 Conclusion

In this chapter I provided a process-centered account of the everyday work of building data science systems, showing how and why the working of a system is neither stable nor given, but a resourceful and improvised artifact that remains *in the making* throughout development. Through this work, I advance the sociotechnical scholarship in critical data studies on the everyday practices of doing data science, offering researchers new pathways into effectively engaging with the entangled relationship between the everyday work of building data science systems and their eventual working and social implications. I make a case for analyzing the human and organizational work by which practitioners decide *how* and *why* systems can and should work in some ways and not in others, including forms of anticipatory work that drive practitioners towards certain technological futures.

Once practitioners build a 'working' system, however, comes the task of deciding whether the system is ready for real-world use—can we trust that the system will do its job? In the next chapter, we will dig deeper into how practitioners make such decisions. How do data science practitioners ascertain trust in data science systems—in data, models, and numbers?

Much like the working of data science systems, the next chapter will show how even the perceived trustworthiness of data science systems is a situated accomplishment, made possible by different forms of human and organizational work on the part of different practitioners.

V
Trust in Data Science

Finally, in this last empirical chapter, we focus on yet another important stage in a corporate data science project—deciding whether to deploy the built data science system. Central to the practice of making such decisions is the notion of trust. Before data science practitioners decide to deploy their systems to their users, they must first trust their own system. It is thus not surprising that companies spend enormous effort and resources to ascertain the credibility of data, models, and metrics. But how is trust established in corporate data science projects? Is it simply done through quantitative analysis—producing and assessing metrics such as accuracy scores?

While a key credibility mechanism, quantitative analysis is but one half of data science's trust puzzle. In this chapter, I show how and why the everyday work of managing and negotiating trust in corporate data science projects is as much collaborative as calculative. To do this, we will take a deeper look at: (a) how certain common tensions problematize practitioners' trust in data science projects and (b) how practitioners perform different forms of human and organizational (and not just technical) work to address and resolve such problems. Assessing the trustworthiness of a corporate data science system is not a *one-shot* decision at

† A slightly edited version of this chapter has been published as Passi & Jackson (2018). For this work, I proposed the initial research question, refining it with feedback from Steve Jackson. I designed and conducted the ethnographic study (with feedback and guidance from Steve Jackson and Phoebe Sengers). I transcribed the interview data, analyzed the data and fieldwork notes on my own; Steve Jackson provided regular feedback that shaped the analysis of the empirical case presented in this chapter. I wrote the first draft and subsequent revisions with feedback and guidance from Steve Jackson.

all (though it may look like that in hindsight). Trust in corporate data science work is a process—one which involves more than one kind of work and more than one type of practitioner.

5.1 Introduction

Data science is a sociomaterial practice (Orlikowski, 2007) in which human and technical forms of work intertwine in specific, significant, and mutually shaping ways. The limits and tensions of everyday data science work in its collaborative dimensions, however, are not yet fully scoped. Researchers argue that "everything might be collected and connected, [but] that does not necessarily mean that everything can be known" (Raley, 2013). Data are never "raw" (Gitelman, 2006), often speaking specific forms of knowledge to power (Bowker, 2014; Pasquale, 2015; Willson, 2017). In the current push towards algorithmic analyses, what counts as valid and reliable knowledge remains contested (Kitchin, 2014a, 2014b; Kitchin & McArdle, 2016; Leonelli, 2015). Algorithmic results based on "sufficient" data, however, are often considered "credible" enough to serve as actionable evidence (Hildebrandt, 2011; Zarsky, 2016). "[The] idea of data-led objective discoveries entirely discounts the role of interpretive frameworks in making sense of data which [are] a necessary and inevitable part of interacting with the world, people and phenomena" (Mittelstadt & Floridi, 2016, p. 320).

If data science enables us to ask different questions and generate novel insights, it also requires varied and distributed forms of interpretation and discretionary judgment that ensure meaningfulness, confidence, and reliability of algorithmic results. The effective practice and governance of data science therefore remain top priorities for data science researchers and practitioners alike.

Central to all of this is trust. The credibility of algorithmic results is adjudged variously—at times through statistical probabilities and performance metrics, and at other times through prior knowledge and expert judgments. The reliability and valence of algorithmic results, as data scientists well know, is also impacted by factors such as the framing of questions, the choice of algorithms, and the calibration of models.[38] Established credibility mechanisms such as statistical significances, quantified metrics, and theoretical bounds comprise forms of *calculated trust*, but their effective use remains challenging. Data scientists may have the necessary expertise to test systems, but the manual intractability of large-scale data coupled with the complex, sometimes opaque, nature of state-of-the-art models makes it hard even for them to clearly articulate and ascribe trustworthiness to approaches and insights. This is even harder for lay users who lack specialized data science knowledge but are sometimes the users of these systems or those most impacted by them. CSCW and HCI researchers have begun to take on these challenges through efforts to make results more understandable (Ribeiro et al., 2016a, 2016b) to effectively manage data science systems (Kleinberg et al., 2016; Luca et al., 2016), to understand user perception of performance metrics (Kay et al., 2015), and to ascertain ways to foster trust through system design (Knowles et al., 2014, 2015).

Contemporary understanding of data science in research, as I described in the Introduction, is largely based on the analysis of data science work in academic and research sites; the large body of applied data science work, especially in corporate settings, has received much less attention. Crawford & Calo (2016) argue that the lack of research focus on already

[38] The term 'algorithm' refers to the underlying statistical/computational, approach (e.g., 'random forests' is an algorithm). The term 'model' refers to the specific application of an algorithm to a dataset (e.g., using 'random forests' on 'user data' produces a model).

in-use data science systems has created a "blind spot" in our understanding of data science work. Based on ethnographic fieldwork in a corporate data science team, this chapter attempts to bridge this gap by describing how problems of trust and credibility are negotiated and managed by actors in applied and corporate settings, with a focus on two separate projects: *churn prediction* and *special financing.* I make visible four common tensions in corporate data science work—(un)equivocal numbers, (counter)intuitive knowledge, (in)credible data, and (in)scrutable models. I describe how each of these tensions raises different problems of trust, in turn highlighting the practices of skepticism, assessment, and credibility through which organizational actors establish and re-negotiate trust under uncertain analytic conditions—work that is simultaneously calculative and collaborative. Highlighting the heterogeneous nature of real-world data science, I show how management and accountability of trust in applied data science work depends not only on pre-processing and quantification, but also on negotiation and translation—producing forms of what I identify as 'algorithmic witnessing' and 'deliberative accountability.' Trust in data science is therefore best understood as a deeply *collaborative* accomplishment, undertaken in the service of pragmatic ways of acting in an uncertain world.

In the following sections, I begin by reviewing history and sociology of science, social science, critical data science, and CSCW literature on trust and credibility in science and technology more broadly and data science and organizations more specifically. I then move on to two empirical examples illustrating the challenges and complexity of trust in applied data science work. I conclude by describing my findings concerning the negotiation of trust in real-

world data science work, highlighting the implications of my findings for the data science field, both within and beyond CSCW.

5.2 Trust, Objectivity, and Justification

My conceptualization of trust begins with an important vein of work in history and sociology of science on the relation between science, trust, and credibility. Influential work by Shapin (1994) and Shapin & Schaffer (1985) show how "working solutions" to problems of credibility in early experimental science were found in the social perception of gentlemen as "reliable truth-tellers." A gentleman was a person of noble birth, raised in the traditions of civility and therefore marked by a commitment to honor, virtue, and righteousness. The seventeenth-century cultural association of "gentility, integrity, and credibility" provided forms of *social scaffolding* to negotiations of scientific trust—gentlemen rejected "notions of truth, certainty, rigor, and precision which were judged suitable for scholarly inquiry, but out of place in civil conversations" (Shapin, 1994, p. xxx). The perception of gentlemen as oracles of truth, however, rendered other knowers such as laboratory technicians invisible (Shapin, 1989). Early experimental science was embedded in (and built on) a "moral economy of truth" led by gentlemanly scientists pursuing science with "epistemological decorum." At the same time, techniques of "*virtual witnessing"*—organized around mechanized experiments, standardized (in principle) reproducible methods, and conventionalized forms of reporting—helped discipline the senses and lent certainty to experimental knowledge. This combination of social order and mechanical apparatus working together—and not uncredited and single testimonies—instilled trust and ultimately power in experimental results. Early experimental science was thus

a collective practice simultaneously social and technical: facts emerged through specific forms of sociability embedded within the experimental discourse.

It was only towards the mid-nineteenth century that "objectivity," as we understand it today, gained prominence as a central scientific ideal (Daston & Galison, 1992, 2007). Daston & Galison (2007), for instance, describe a shift from "truth-to-nature," a mode of objectivity characterized by "reasoned images" of natural reality produced by scientists based on theoretical selection and aesthetic judgments of exemplary specimens to "mechanical objectivity," organized around the credibility of mechanized means such as photography to produce images untouched by (troubling) individual scientific judgments. As the work of interpretation shifted from the maker to the reader, scientific artifacts became open to interpretation, making consensus challenging. Agreements between multiple ways of seeing required differentiating between right and wrong interpretations: science required the correct "professional vision" (Goodwin, 1994). As mathematicians and physicists argued for "structural objectivity" characterized by forms of measurement and replicability, the role of "trained judgments" became salient. Scientists thus chased truth with "calibrated eyes" and standardized tools, simultaneously questioning the credibility of their findings, instruments, and knowledge.

Today, objectivity goes hand in hand with quantification: mechanically-produced numbers standing in for factual representations of reality (Desrosieres, 1998; Espeland & Stevens, 2008; Joerges & Shinn, 2001; O'Connell, 1993; Porter, 1995). Numbers do not just lend credibility but are themselves "couched in a rhetoric of factuality, imbued with an ethos of neutrality, and presented with an aura of certainty" (Rieder & Simon, 2016, p. 4). Numbers

ascribe but also command trust. Trust in numbers, however, as Porter (1995) argues, is but one form of "technologies of trust." Through studies of accounting and engineering practices, Porter shows how the "authority of numbers" was enacted as a pragmatic response to institutional problems of trust. The twentieth-century thrust for precision coupled with the rise of mechanical objectivity brought in a "regime of calculation" in which rule-bound numbers fostered a "cult of impersonality." Sanitizing the inherently discretionary nature of professional practices, quantification became a rigorous judgment criterion: "a way of making decisions without seeming to decide" (Porter, 1995, p. 8). Numbers, however, do not contain within themselves their own logic of interpretation. As a form of intervening in the world, quantification necessitates its own ecologies of usability and valuation. Trust in numbers is therefore best understood as a variegated "practical accomplishment" (Garfinkel, 1964, 1967), emanating from efforts at standardization and mechanization, along with forms of professional and institutional work (Hacking, 1990; Harper, 2000; Power, 1997; Shapin, 1995a, 1995b).

A second line of work central to problems of trust in complex organizational settings is found in pragmatist traditions of social and organizational science. Dewey (1939) argues that instead of existing as *a priori* criteria, values—as perceived or assigned worth of things—are continuously negotiated within decision-making. Processes of valuation are simultaneously evaluative (how to value?) and declarative (what is valuable?). Boltanski & Thévenot (2006) focus on the "plurality of justification logics" comprising valuation practices. They describe six orders of worth—c*ivic*, *market*, *industrial*, *domestic*, *fame*, and *inspiration*—within which "justifiable agreements" are achieved in social and professional practices. Each order of worth signifies its own reality test based on "the testimony of a worthy person, […] the greatest

number, […] the general will, […] the established price, […or] the competence of experts" (Boltanski & Thévenot, 2006, p. 32).

Applying these insights to organizational decision-making, Stark (2009)builds on Boltanski & Thévenot's (2006) work to build a theory of "heterarchies"—organizational forms in which professionals operate according to diverse and non-commensurate valuation principles. In his ethnographic studies of a factory workshop, a new media firm, and an arbitrage trading room, Stark analyzes how actors use diverse "evaluative and calculative practices" to accomplish practical actions in the face of uncertainty. In heterarchies, decision-making requires negotiations between different, often competing, performance metrics. Stark distinguishes his account of heterarchies from Boltanski & Thévenot's orders of worth in two significant ways. First, while Stark argues that decisions are embedded within multiple forms of valuation, he does not believe that decisions are confined within a pre-defined matrix of worth. Valuation criteria, for Stark, remain contextual, emerging differently in different situations. Second, while Boltanski & Thévenot position orders of worth as ways of *resolving* uncertainty, Stark sees the plurality of valuation as providing opportunities for action by *creating* uncertainty. Organizational conflicts are not roadblocks, but integral to organizational diversity in which the "productive friction" between multiple ecologies of valuation helps accomplish justification and trust—(dis)trusting specific things, actions, and worlds.

Data science is no different. Integral to several contemporary knowledge practices, algorithmic knowledge production is here and now. Big data pushes "the tenets of mechanical objectivity into ever more areas of applications"—data science is not just interested in calculating what *is*, but also "aspires to calculate what is yet to come" (Rieder & Simon, 2016,

p. 4). Given enough data, numbers appear to speak for themselves, supplying necessary explanations for all observations (Anderson, 2008; Hildebrandt, 2011; Leonelli, 2014; Rieder & Simon, 2016). The problem is exacerbated given that automation—as operationalized in and through data science—is often understood as an absence of bias (Bozdag, 2013; Dourish, 2016; Gillespie, 2014; Naik & Bhide, 2014). The mathematical foundation of algorithms, coupled with the manual intractability of large-scale data, has shaped the use of quantified metrics as key indicators to ascertain a data science system's workability.

The increased management of data science systems is a priority (Luca et al., 2016; Symons & Alvarado, 2016; Zook et al., 2017), but "the complex decision-making structure" needed to manage them often exceeds "the human and organizational resources available for oversight" (Mittelstadt et al., 2016). One limiting factor is the opacity of data science systems (Tabarrok, 2015; Veale, 2017) arising from a variety of reasons: (a) algorithms are often trade secrets, (b) data science requires specialized knowledge, and (c) novel analytic methods remain conceptually challenging (Burrell, 2016). Neural networks are often critiqued for their black-boxed nature,[39] but researchers argue that even simpler models are not necessarily more interpretable than their complex counterparts (Lipton, 2018). The fact that models do not always furnish explanations has sparked efforts to produce humanly understandable explanations of data science workings and results (Guidotti et al., 2018; Ribeiro et al., 2016b; Singh et al., 2016). Another limiting factor is the paywalled or proprietary nature of datasets. Several data correlations established in large datasets are neither reproducible nor falsifiable (Ioannidis,

[39] A neural network contains multiple layers with each layer containing several nodes. Broadly, each node acts as a simple classifier, activating upon encountering specific data attributes. Output from one layer forms the input for the next layer.

2005; Lazer et al., 2014). The combination of data inaccessibility and analytic fragility makes "virtual witnessing" (Shapin & Schaffer, 1985) challenging, if not at times impossible. While making data public may seem a good first step (though not in the best interests of corporations), transparency remains a difficult and problematic ideal (Ananny & Crawford, 2016). Data science comprises myriad forms of discretionary human work—visible only to those in its close vicinity. Algorithmic results comprise specific articulations of "data vision"—the situated rule-*based* traversal of messy data through clinical rule-*bound* vehicles of analysis (chapter two). The absence of development histories of systems further camouflages aspects arising not only from personal and professional values, but also from the emergent sites of algorithmic re-use as opposed to their contexts of development (Friedman & Nissenbaum, 1996; Hildebrandt, 2011; Kery et al., 2017). This combination of oversight challenges often causes "non-expert" data subjects to lose trust in the enterprise of data science more broadly (G. Cohen et al., 2014; Dourish, 2016; Rubel & Jones, 2014).

CSCW and critical data science researchers have thus put forward an agenda for "human-centered data science," arguing that focusing on computational or statistical approaches alone often fails to capture "social nuances, affective relationships, or ethical, value-driven concerns" in data science practices (Aragon et al., 2016, p. 2). Such failures cause rifts between developers' understanding of what the system does and the users' interpretation of the system in practice, causing user and societal confidence in these systems to plummet. This is particularly challenging because the systems are not usually designed "with the information needs of those individuals facing significant personal consequences of model outputs" (Binns et al., 2018, p. 10). While quantification provides calculative relief in the face of uncertainty,

questions of trust are "ultimately questions of interpretation," and within meaning-making processes "limitations of traditional algorithm design and evaluation methods" are clearly visible (Baumer, 2017, p. 10). Researchers thus focus not only on fostering trust in computational systems (Knowles et al., 2015) but also on understanding users' perception of quantified metrics (Kay et al., 2015).

These bodies of history and sociology of science, critical data studies, CSCW, and social science research highlight the import of trust and credibility justifications in data science work. Problems of trust in science are solved variously, characterized by a set of key approaches: *social scaffolding*, *virtual witnessing*, *mechanical objectivity*, *trained judgments*, and *quantification*. Beyond a taxonomy of trust mechanisms, this list also makes visible the sociocultural and political factors encompassing scientific negotiations on trust. As we—researchers and society alike—grapple with problems of societal trust in science and technology, it is important to understand how trust and credibility are established in scientific and technological practices at large. Data science charts new knowledge frontiers, but its maps remain opaque and distant to all but a few. In complex organization settings, data science is transected by multiple experts, interests, and goals, relying upon and feeding into a plethora of practices such as business analytics, product design, and project management. Applied data science needs not only scientists and engineers, but also managers and executives (Kleinberg et al., 2016; Luca et al., 2016).

In this chapter, I unpack the situated work of organizational actors to engender trust in data, algorithms, models, and results in corporate data science practices. Highlighting the plurality of valuation practices in organizational decision-making, I describe common tensions

in data science work and mechanisms by which actors resolve problems of trust and credibility. As forms of justification, these mechanisms exemplify pragmatic, not absolute, forms of trust, helping actors act and move forward in an uncertain world—valuing specific forms of knowing over others. My goal in this chapter is not to merely argue the partial, social, and messy nature of data, models, and numbers (a fact readily acknowledged by my actors), but to show *how* actors negotiate and justify the worth of data science solutions to identify opportunities for practical action when faced with such conditions. Evaluation of a data science system, as I show in this chapter, is rarely the simple application of a specific set of criteria to a technology, and often requires specific mechanisms to negotiate and translate between multiple situated, often divergent, performance criteria. In this chapter, I address two of these mechanisms: *algorithmic witnessing*, in which data scientists assess model performance by variously, mostly technically, reproducing models and results; and, *deliberative accountability*, in which multiple experts assess systems through collaborative negotiations between diverse forms of trained judgments and performance criteria.

5.3 Empirical Case Studies

5.3.1 Case One: Churn Prediction

For case one, I focus on a *churn prediction* project—i.e. the detection of currently active customers who are likely to cancel paid services in the future—associated with DeltaTribe, a multi-million-dollar marketing and management subsidiary owned by DeepNetwork, that provides online customer management services to thousands of clients across the United States in domains such as medical, dental, veterinary, and automotive. In business terminology, customers who cancel paid services are called ‘churned,’ and active customers who might cancel paid services are called ‘likely to churn.’ In what follows, I describe key moments from

the project to show how tensions emerge as actors negotiate the trustworthiness of numbers and intuition.

(Un)equivocal Numbers

When I began fieldwork, the project's data scientists—David[40] and Max—had already settled on two models after trying several approaches.[41] Their models differed not only in their algorithmic approaches, but also in their data. David's model used *subscription data* consisting of information such as the services enabled for a customer and their platform settings. Collected monthly, this data was available for the past two years—a total of 24 datasets. Max's model used *action data* containing data on customers' actions such as the modules they accessed and when. This data was collected every day but was not archived—each day's collection overwrote the previous day's data. David's subscription-based model therefore had more historical data to work with than Max's action-based model. David's model, however, had lower accuracy than Max's model.[42] Both believed that "more data was better," but the jury was out on whether more data led to better performance. In fact, Max's model had an accuracy of ~30%—a number higher only in comparison to the ~20% accuracy of David's model. Director of data science Martin was "disappointed" in both models.

[40] David has an undergraduate degree in electrical and electronics engineering and several online data science certifications. He has been with DeepNetwork for a little less than two years. Before this, he worked as a Product Engineer.

[41] Model one was based on gradient boosting algorithm (xgboost), and model two on random forests.

[42] In churn prediction, the accuracy score refers to the percentage of correctly identified likely-to-churn customers (who do cancel their services in the future) in the total number of likely-to-churn customers identified by a model.

Over time, accuracy scores did not increase, but the data science team's disappointment in their models decreased. This was because accuracy score—as a performance metric—was devalued by team members over the course of the project. For instance, data science team's project manager Daniel found accuracy scores problematic given his prior experiences of working with business teams:

> **Daniel (project manager):** When we say to business teams that we have 90% accuracy, they think: 'oh, you are saying with a 90% confidence that these people are going to churn.' NO, THAT'S NOT WHAT WE ARE SAYING! They think in black and white: likely to churn or not likely to churn. In total we have four answers: (a) people we say will churn, and who churn [true positives], (b) people we say will churn, but do not [false positives], (c) people we say will not churn, and do not [true negatives], and (d) people we say will not churn, but churn [false negatives]. Business folks don't know what the numbers *really* mean (Fieldwork Notes, June 8, 2017).

This point was borne out in a subsequent interview with Parth[43], DeltaTribe's business analyst, who described his perception of low accuracy scores:

> **Parth (business analyst):** "Expectation-wise […], this project was going to be the silver bullet. Then, expectations were, I guess, neutralized or you know, brought down to earth. […] I thought [it] was going to be 80% or 90% accuracy. […] In reality, it was more like 20-30%. That's where the expectations for me were shattered, I guess" (Interview, October 26, 2017).

For Parth, a low accuracy score signaled failure. His assessment was largely based on performance numbers that he believed provided definite information about the model. Both Parth and Charles[44]—another business analyst—were also skeptical of numbers produced by the models: the likely-to-churn customer probabilities. A likely-to-churn probability, generated

[43] Parth has been with DeltaTribe for 8+ years. Before this, he was an Assistant Manager in a corporation (4+ years).

[44] Charles has been with DeltaTribe for 5+ years. Before this, he was a consultant in a corporation (5+ years).

by the model, indicates the model's perception of a customer's likelihood to churn (the higher the number, the higher the likelihood). Parth and Charles found the probabilities helpful but incomplete on their own without further information on how they were generated.

> **Charles (business analyst):** "We were looking for [likely-to-churn probabilities]. But, keep in mind that, that number does not tell the full story. It is a good indicator, but it is not the absolute truth for us. […] It [only] helps us identify which customers to go for, [but not why]."

> **Parth (business analyst):** "The more we understand how models work […] allow[s] us to understand: 'ok, this score actually means this or when there is a change, [it] means that'." (ibid.)

These probabilities were not always a part of the results. Initially, the results only contained customer IDs and a label indicating whether the model perceived a customer as 'likely-to-churn' or 'not-likely-to-churn.' The move to probabilities, as I show below, was largely an outcome of the devaluation of accuracy scores. In addition to voicing concerns about the business interpretation of accuracy scores, and performance metrics in general, project manager Daniel and data scientists David and Max argued that low scores in the project were not a sign of "bad models," but of the "extremely hard" nature of churn prediction itself.

> **David (data scientist):** It is tough. Customers churn for any reason, and you can't know that. E.g., they want to save money, or their boss told them to cancel. Even with all the data, we cannot be perfect. Almost impossible (Fieldwork Notes, August 16, 2017).

> **Daniel (project manager):** "[Accuracy] is so low because you are trying to predict a *very* difficult concept, […] to explain human behavior. […] It is random to a degree. There is stuff that you can predict and model, but […] it just seems unreasonable to […] a data scientist that you can create a model that perfectly models human behavior. […] The challenge with non-technical people is they think that computers can do more than they really can" (Interview, June 30, 2017).

Most predictions made by the models were incorrect, but the data science team gradually saw value in the small handful of correct predictions: *30% was better than nothing.* As Daniel put it, "even bad results or less than ideal results can be good—it is more than what you know now." The perceived value of having *some* predictions over *none* further shaped the data science team's outlook towards model results and metrics. This prompted the move towards likely-to-churn probabilities. While binary labels—likely-to-churn/not-likely-to-churn—drew attention to which labels were right and which were wrong, churn probabilities instead brought to light customers at varying levels of churn risk. The results were now in the form of an ordered list of customers with decreasing likely-to-churn probabilities.

> **Martin (director of data science):** Think of it as a list that if you are on it, you are going to die. We provide the ones with the highest probability of dying and tell our business team that maybe you should reach out to these folks. By doing this we deemphasize the 30% and say: 'how can a low 30% score still be made useful for businesses?' (Fieldwork Notes, May 30, 2017).

(Counter)intuitive Knowledge

The results were presented to DeltaTribe's business analysts Parth and Charles. A feature importance list accompanied the results.[45] Director of data science Martin and project manager Daniel considered the list an important part of the results—an explanation mechanism to help the business team "see" what the models considered noteworthy. Within minutes of seeing the results, both Parth and Charles described the results as "useful." Their perception of the usefulness of results, however, was not grounded in model performance metrics (mentioned

[45] In machine learning, *features* refer to measurable attributes. E.g., 'whether email is enabled' is one feature of a customer. Algorithms analyze features to calculate results. A feature importance list, produced by certain algorithms, contains features and their weights. The weights signal the relative importance of each feature in the algorithmic model's analysis.

once in the hour-long meeting) or in the manual inspection of customer accounts (not done in the meeting at all). Instead, their perception was based solely on the feature importance list.

> **Daniel (project manager):** Here [points to a slide with feature importances] we see that whether customer enabled email and voice were important. So was use, how much the customer engages. It might mean engagement with the platform is good, and customers who engage remain active. Also [points to a different part] communication was important.
>
> **Martin (director of data science):** Do these results match your gut feeling, your intuition?
>
> **Charles (business analyst):** Definitely! These are things we already focus on (Fieldwork Notes, June 9, 2017).

Certain highly weighted features matched business intuitions, and everyone in the meeting considered this a good thing. Models that knew "nothing about the business" had correctly identified certain aspects integral to business practices. Such forms of intuitive results were important not only for business analysts, but also for data scientists. In an interview, director of data science Martin described how "seeing" intuitive patterns had helped him corroborate results:

> **Martin (director of data science):** "We saw patterns [showing that] there is this cohort of customers that are less than one year old, and they are clustered in their behavior. […] This other cluster of customers that have been with us more than X number of years and […] their behavior looks like this. […] The graph was SO important. […] To me, it is easier to look at a graph, especially one […] so *obvious* in terms of patterns" (Interview, August 23, 2017).

In the meeting with business analysts, however, not all feature importances mirrored business expectations. Parth asked Daniel why a specific feature that DeltaTribe considered important

did not show up in the top ten features identified by the models. Daniel argued that even if expected features were absent, it did not mean that the models were necessarily wrong:

> **Daniel (project manager):** If you see a feature in the middle of the pack but expect it to be higher, it might mean that in your business you already focus on it. Its importance perhaps went down over time. If you focus on these [points to top features] that are prioritized by models, we expect that these will also go down over time. You focus on customers that we say are at risk. They don't cancel. This is good. But, it means that features will change (Fieldwork Notes, June 9, 2017).

Regarding counter-intuitive feature importances, Daniel reminded Parth and Charles that machine-learning models do not approach data in the same way humans do. He pointed out that models use "a lot of complex math" to tell us things that we may not know or fully understand. While DeltaTribe may consider a feature important for their business, models may spot the over-time devaluation of the feature's importance for a specific business practice (here, churn). The feature importance list signaled the flux of business practices and priorities. If an "intuitive" feature had been sufficiently incorporated in business practices, Daniel and Martin argued, it would – from the model's perspective – cease to be "important." Counter-intuitive findings were *also* valuable. This tension between intuitive results and counter-intuitive discovery showed up again in my interview with Jasper[46]—DeltaTribe's CTO. As he explained, one way to assess the efficacy of a counter-intuitive result was to juxtapose it with the intuitiveness of the data science workflow.

> "When I get an end-result type of an answer, I like to understand […] the factors that went into it, and how they [data scientists] weighed it to make sure that it makes intuitive sense. […] I understand that a lot of the times, findings will be counterintuitive. I am not immediately pre-disposed to distrust anything that does not

[46] Jasper has been DeltaTribe's CTO for 2+ years. Before this, he had 25+ years of work experience in the technology industry in several roles such as chief information officer, chief operating officer, and senior systems analyst.

> make intuitive sense. I just like to understand the [significant] factors. […] Understanding the exact algorithm, probably *less* so. […] Understanding process is really important […] to trust that *(pause)* science was allowed to take its proper course" (Interview, November 14, 2017).

The tension between intuition and counter-intuition was not limited to algorithmic results. The business goals themselves, Jasper further argued, may often appear counter-intuitive. This was the case, for instance, with Jasper's requirement of sensitivity over precision. A model configured for sensitivity prioritizes recall, aiming to maximize the number of correct predictions in its results.[47] A model configured for precision, however, prioritizes positive predictive value, aiming to maximize the correctness of its predictions.[48]

> **Jasper (CTO, DeltaTribe):** "As a pragmatist, what I am looking [for] are things that are highly sensitive, and their sensitivity is more important to me than their accuracy. […] If you can ensure me that of [say] the 2700 [customers] we touch every month, all 500 of those potential churns are in that, that is gold for me. […] If you could tell me [to] only worry about touching 1000 customers, and all 500 are in it, that'd be even better. But […], let us start with [making] sure that all the people I need to touch are in that population, and make maximum value out of that. […] It is about what outcomes I am trying to optimize to begin with, and then what outcomes am I trying to solve for and optimize after. […] You want your model to be completely sensitive and completely accurate. Of course! […But,] you don't want to optimize too soon. […] I probably want to talk about dialing up accuracy a little bit [later]. [Our current approach] is so inherently inefficient that there is an enormous order of magnitude [of] optimization possible without being perfect. […] It is maybe a little bit counter-intuitive, but the goal I am trying to solve for is: can I spend a subset of my total resources on a population that is going to return well for me, but what I am not doing is avoiding spending resources on people that I should have?" (ibid.)

[47] Recall is the ratio between True Positives and Total Positives. True Positives are customers that churn and are correctly identified by the model as likely-to-churn. Total Positives are the total number of customers that churn. A sensitive model maximizes recall.

[48] Positive predictive value is the ratio between True Positives and the sum of True Positives and False Positives. False Positives are customers that do not churn but are identified by the model as likely-to-churn. A precision model maximizes the accuracy of its predictions.

Business requirements were thus two-fold. On the one hand, the goal was to minimize churn. This led to an initial preference for sensitivity. On the other hand, the aim was to optimize resource allocation. This led to a subsequent preference for precision. For Jasper, the goals came one after the other—first build a sensitive model, and later tune it for precision. The data science team, however, tackled the problem differently. Their models were not configured for sensitivity or precision but for specificity. A model configured for specificity does not focus on maximizing the number or accuracy of its correct predictions but on minimizing the number of its incorrect predictions. The aim was to ensure that healthy customers were not incorrectly identified as likely-to-churn. By minimizing incorrect predictions, the team hoped to ensure that customers classified as likely-to-churn were, in fact, problematic, guaranteeing that resources were not wasted on healthy customers. For Jasper, the two goals married different corporate values with different computational ideals at different points in time. The need to separate the two goals ("you don't want to optimize too soon"), Jasper argued, may seem counter-intuitive to data scientists. Indeed, for the data science team the two goals went hand in hand—the 'specific' solution found in incorrect answers instead of correct ones. At the end of the meeting, business analysts Parth and Charles agreed to conduct pilot tests with the results to see if they lowered churn rate.

There are two striking features evident in the vignette above. *First,* quantified metrics such as precision, recall, specificity, and accuracy remain integral to model assessment, but they are often valued differently by different actors. The data science team considered likely-to-churn probabilities a valuable resource, but the business team found them incomplete in the absence of knowledge about how they were generated. Conversely, while the data science team devalued the usefulness of the accuracy score, the business analysts nevertheless considered the score an

important signal in assessing model usefulness. The business team wanted to first focus on recall and then on precision, but the data science team saw in specificity a way to marry both concerns. Thus, while different organizational actors differently articulate and negotiate the efficacy of numbers, their use in complex collaborative settings engenders specific forms of practical action (e.g., focusing on varying levels of churn risk as opposed to the binary of churn or no-churn) and understanding (e.g., prioritizing specificity over precision or recall).

Second, while intuition acts as a form of a reality check in data science practice, it also runs the risk of naturalizing or rendering invisible biases and precluding alternate findings. Intuitive results and familiar or expected patterns engender trust in algorithmic results, in turn ascribing forms of obviousness to data science approaches. While the role of intuition in making sense of data science results and processes is revealing, this project surfaced another aspect of the role of intuition in data science practice. Organizational actors negotiate not only what is and is not intuitive, but also when counter-intuition is and is not warranted. They leverage both the intuitiveness *and* counter-intuitiveness of results and processes to negotiate trust and credibility in data science systems. The possibility of unexpected patterns and counter-intuitive knowledge—one of the great promises of data science work after all—stands in contrast and partial tension with intuitive assessment, requiring ongoing forms of work to balance between the two. This is especially apparent in projects in which business goals themselves may seem to follow a counter-intuitive trajectory to computational ideals.

5.3.2 Case Two: Special Financing

For the next case, I turn to the project on *special financing*—loan financing for people who have either low or bad credit scores (usually 300-600) or limited credit history—that we previously

saw in chapter three. To give a quick recap: one of DeepNetwork's subsidiaries—CarCorp—helps people to get special financing loans to buy new and used cars. CarCorp's clientele consists of money lenders and auto dealers who pay the company to receive "leads"—information on people interested in special financing. This information comprises several details including demographics, address, mortgage, and current salary and employment status. Knowing which leads will get loan approval by lenders/dealers is not simple. CarCorp wanted to use data science to predict which leads are likely to get financed *before* they are sent to clients. The company assigned Ray, its business analyst, to work with the data science team on this project.

(In)credible Data

Both data scientists and business personnel accepted that some leads were good and some bad. For Ray, a lead was good if a lender/dealer approved it for financing. Different lenders and dealers, however, had different requirements for loan approval. The data science team thus approached the question as a matching problem: *how do you match leads to lenders and dealers that are more likely to finance them?* As data scientist Max began work on the project, something troubled him. He described his concern in a data science meeting:

> **Max (data scientist):** Surprisingly, more than 50,000 people who applied for a [special financing] loan earn more than $5000 a month! Why do you *(throws his hands in the air)* need a [special financing] loan for a car if you earn that much money? Maybe there is noise in the data. I need to fix that first (Fieldwork Notes, May 31, 2017).

Max found it odd that people with $5000 monthly earnings applied for special financing. Underlying Max's inference was his assumption that people with bad credit scores do not earn such salaries. Director of data science Martin and DeepNetwork's CTO Justin, however, believed

otherwise. Arguing that the relationship between credit score and salary was tenuous at best, they told Max that his interpretation was incorrect.

> **Martin (director of data science):** That isn't true. Your current salary cannot tell you whether someone would or would not need a [special financing] loan. Do not fix anything, please!
>
> **Justin (CTO):** Yeah, you might have accrued debt or filed for bankruptcy (ibid.).

While Max's concern was dismissed as a false alarm, it did not mean that data was not suspect in other ways. Throughout the project, there were several discussions around credibility ranging from questions of data accuracy to the reliability of the underlying data sources themselves. Business analyst Ray, for example, raised concerns about how data used in the project was generated and collected:

> **Ray (business analyst):** "How this business works […] is very strange. […] The fact that [some] dataset [come] from affiliate[s].[49] We [also] generate our leads organically [online…], then there are leads that we get from syndication that we bid on. [...] Different pieces of information [are] appended to those leads depending on where […they come] from. […] There is one affiliate […], they give a credit score range for a bunch of those leads. So, not exactly the score, but they could say 'this person is between 450-475 and 525-550,' different buckets. [...] Never realistic but we pay money, and we have this data." (Interview, November 8, 2017).

Acquired through multiple means (e.g., web forms, bidding), lead data was, as we saw earlier, sometimes augmented with additional data such as credit score ranges by some of CarCorp's business affiliates. The role of credit score data was particularly revealing. Credit score ranges were a part of leads bought from a small number of third-party lending agencies. CarCorp's business analysts wondered whether credit score range data could help them in this project. Credit

[49] The affiliates consist of third-party lending agencies and business associates.

score ranges, for instance, were already used by business analysts as one way to distinguish between two leads that appeared identical but exhibited different financeability.

> **Ray (business data analyst):** "Two individuals that have the same age, same income, same housing payment, same everything […] could have wildly different credit scores. [...] You have those two individuals, and you send them to the same dealer. From our perspective lead A and B are […] maybe not exactly same but close. [...] But, the dealer can finance person A, and they cannot finance person B [...] So, when they [dealers] are evaluating from month to month whether they would renew their product with us, if we had sent them a bunch from bucket B, and none from A, they are likely to churn. But, we have no way of knowing who is in bucket A and who is in bucket B. […] Two individuals who measure the same on data points, [could] have two different credit scores." (ibid.)

The data science team's attempt to assess the usefulness of credit score range data, however, as we know, faced certain practical challenges. The data was incomplete, only available for a few leads (~100,000 out of ~1,000,000). The data was approximate, not in the form of a single number (e.g., 490) but as a range (e.g., 476-525). The data was inconsistent since different affiliates marked ranges differently. For example, a credit score of 490 might be put in the 476-525 range by one affiliate and in the 451-500 range by another. Credit score data, considered a key factor by the business team, was thus at best sparse, rough, and uneven. As data scientist Max attempted to make credit score ranges received from different affiliates consistent with each other, business analysts found a way to make this work easier. Pre-existing market analysis (and, to some extent, word-of-mouth business wisdom) showed that leads with credit scores greater than 500 were highly likely to get special financing approval. This piece of information simplified the work of achieving consistency across different credit score ranges. Max did not need to figure out ways to reconcile overlapping ranges such as 376-425 and 401-450 but could simply include them in the *below-500* category. Only two credit score ranges were now important: *below-500* and *above-*

500. The solution to the matching problem (*which leads are likely to get financed by a lender)* was now a classification task (*which leads have a credit score of over 500).*

> **Max (data scientist):** [CarCorp] really care about 500. If the credit score is below 500, the dealer will simply kill the deal. They now want me to tell them whether credit score is below or over 500. The problem is there are too many records in 476-525—a very tight group. This makes it difficult (Fieldwork Notes, June 13, 2017).

The presence of several leads in the 476-525 credit score range was a problem given that 500—a number now central to the project—fell right in the center of the range. This made it hard to figure out which leads in the 476-525 range were above or below 500. The threshold of 500 had helped attenuate the effect of inconsistency but not circumvent it. Max tried several ways to deal with this problem, but each attempt decreased the accuracy of his model. Later, he achieved a classification accuracy of 76%.[50] He did this by completely deleting all the leads from the dataset that were in the 476-525 range. He acknowledged that this was an imperfect solution but argued that it was an effective way forward. The model was now accurate for all but a quarter of the leads (better than chance, or 50/50), but at the cost of removing leads near the crucial threshold of 500.

(In)scrutable Models

When Max shared results with the business team, he received mixed reactions. The business team was unimpressed with the accuracy scores and uncertain about how the model produced results:

> **Bart (project manager):** 'We were given a PowerPoint with like, a short note saying here are the numbers and here is this technique. [...] We weren't told much other than that. I personally felt that we weren't evaluating the model, but it was like—do you

[50] The number of correctly identified above-500 leads in the total number of leads that the model identified as above-500.

> like these numbers? That wasn't helpful. We didn't like the numbers. [But] we got no explanation of how things worked, how insights were generated. We all just were not on the same page. [...] Max tried to explain to us, but it was explained in a manner [...] we maybe did not get? [...] It soon got overwhelming with all the formulas and models' (Interview, November 1, 2017).

Max provided detailed descriptions of the algorithm but the "manner" in which he explained did not resonate with the business team. Max had used a neural network. The exact working of neural networks on data is mostly black boxed even for data scientists. In an interview, Max mentioned that the large scale of the data coupled with the complex nature of neural networks made it difficult for him to explain to the business team how the model made decisions. Max, in fact, argued that understanding how the model made decisions was "not that important."

> **Max (data scientist):** The reason we use machine learning, we let the machine learn. Like a child learns a language. When we say 'hi,' the child says 'hi' back. We see that, but we don't know why. We don't ask why the child says 'hi.' I don't get it. We can use tools without understanding the tools. E.g., stock markets. There are charts, lines, and we make decisions based on them. We don't know how they work! (Fieldwork Notes, July 13, 2017).

The data science team was not opposed to explaining the principles underlying their models. It was the *in situ* working of his model that Max argued was blackboxed but also unimportant. In a later interview, director of data science Martin described what he thought was the best way to proceed in such projects—a combination of "implicit trust" and "explicit verification":

> **Martin (director of data science):** "[We need] an implicit trust that the models [produce] correct outputs. […] We can explain at some layman's level […] what algorithms the models are based on, […] what that black box was based on, but please do not ask me – because I cannot tell you anyway – how it got to the result that we are offering to you. I can tell you that […the] model happened to be based on […this algorithm] with 10-fold validation. […] I'll *even tell you* how […the models] are created, what their characteristics are, what [are] the pros and cons of the [model's] algorithm. But, for your case, why did that 0.9 come up versus the 0.84 for customer ID 'x' versus 'z'? Could not tell you. […] I am hoping for implicit trust with explicit verification [from a pilot test]. Because if it turns out during the pilot that the

> effectiveness wasn't there, I am also perfectly okay to call it a failure. It did not meet the business requirements, and we did not add value, and I am okay with that" (Interview, August 23, 2017).

For Martin, implicit trust corroborated by explicit verification in the form of "real-world" tests remedied the lack of explanations. He wanted to conduct a pilot test with model results to see if lender perception of lead efficacy improved. The business team felt otherwise. The absence of model explanations did not inspire enough business trust even for a pilot test. With this feedback, Max returned to work on his model.

Two key insights emerge from the description of this project. *First*, notions of trustworthiness in data science practices are entangled with the perceived credibility (or lack thereof) of data itself. Differentiating between signal and noise in data is challenging. While data scientist Max assumed a set of data values as noise, business analyst Ray doubted the data's reliability given how it was prepared. Incomplete data is a deterrent (e.g., credit score data only available for a handful of leads), but even data assumed complete is often inconsistent (e.g., different ranges provided by different affiliates) or misaligned (e.g., leads in the 475-525 range stood in opposition to treating 500 as a threshold). Working solutions to data-driven problems require creative mechanisms and situated discretion to work *with* messiness (e.g., finding ways to make credit score ranges consistent across affiliates) and *around* messiness (e.g., deleting problematic leads).

Second, stakeholders' trust in data science systems stems not only from model results and performance metrics, but also from some explanation or confidence in a model's inner working—an explanation or confidence which may prove challenging to port between members of the data science and business teams. In this project, we see the presence of at least two

possible ways to unwrap black-boxed models. One way is to explain *how* a model's algorithm works in principle. Although such explanations are possible and provided by data scientists, we saw in this project (and others) that such information was sometimes considered unhelpful by business teams who preferred a second kind of explanation: *why* the model makes a specific decision. Such explanations are neither straightforward nor always available. Data scientists describe the lack of these explanations not as an impediment but as a trade-off between in-depth understanding and predictive power. It is thus not surprising that data scientist Max considered such explanations unnecessary. The absence of these explanations necessitates additional work on the part of the data science team to help foster business trust in black-boxed models (e.g., combining implicit trust and explicit verification).

5.4 Discussion

The two cases above surface crucial tensions within real-world data science practices. In case one—*churn prediction*—we see negotiations concerning the (un)equivocality of numbers and the (counter)intuitiveness of results. Quantified metrics, integral to assessing the workings of data science solutions, exhibit a certain degree of plasticity. The perceived value of metrics shifts as numbers move between people and teams marked by different, often divergent, valuation practices. Recourse to intuition may engender confidence but at the risk of camouflaging or dismissing novel insights. Balancing between expected and unexpected results is therefore central to the validity of intuitive assessments. Through case two—*special financing*—we see how actors assess data (in)credibility and rationalize model (in)scrutability. Prior knowledge and established goals shape a dataset's perceived effectiveness, requiring discretion to work with and around challenges of inconsistency, unreliability, and

incompleteness. Explaining why models make certain decisions (i.e., their situated application) is often as important as describing how they work (i.e., their abstract working). The (in)scrutability of a model shapes its evaluation in significant ways. These tensions problematize the actors' ability to trust data, algorithms, models, and results. In the face of uncertainty, I see actors using specific mechanisms to resolve and circumvent such problems—solutions to problems of trust that help enable pragmatic action. I describe these mechanisms in the following four sub-sections.

5.4.1 Contextualizing Numbers

In case one, we see actors qualify the effective value of quantified metrics in specific ways. A first strategy involves *decomposing a problematic number into its constituent parts.* For the data science team, a single accuracy score made invisible, especially to business teams, the four scores constituting it (true positives, true negatives, false positives, false negatives). An overall accuracy score of, say, 75% demonstrates a model's success for three-quarters of the data but signals the model's failure for the remaining quarter. Decomposing the score pluralizes the notion of a mistake, recasting the missing 25% as a combination of four *different kinds* of mistakes embedded within statistical ideals of precision, recall, or specificity. The business team wanted to first build a sensitive model and then tune it for precision. The data science team, however, focused on specificity, trying to kill two birds with one stone—minimizing incorrectness to improve recall as well as sensitivity. Different approaches enable prioritization of certain mistakes and values over others, facilitating the re-negotiation of the model's perceived success or failure. During fieldwork, I saw several instances in which numbers

considered sub-optimal were broken down into their constituent parts, while numbers assumed adequate or sufficiently high were often communicated and interpreted at face value.

A second strategy involves *situating suspect numbers in a broader context.* The accuracy score provided information about model performance, but the score's interpretation extended beyond the model. The data science team juxtaposed low accuracy scores with the description of churn prediction as a "very difficult" problem. Arguing that customers often churn for reasons uncapturable in datasets, they found it unreasonable to assume that human behavior could be modeled perfectly. Low performing models shaped the data science team's understanding of the project's complexity. The value of "even bad results or less than ideal results" was found in their ability to provide information not already available to the business team. The accuracy score of 30 was not 70 less than 100 but, in fact, 30 more than zero. Sub-optimal models were still better than nothing. Throughout my fieldwork, I saw instances in which numbers, especially large numbers, acted as "immutable mobiles" (Latour, 1987)—as stable and effective forms of data science evidencing. But in many others, actors leveraged the inherent mutability of numbers in specific and significant ways. The plasticity of numbers in such contexts is therefore partial but strategic.

5.4.2 Balancing Intuition

Actors balance intuition and counterintuition in specific ways. A first strategy comprises *leveraging intuition as an informal yet significant means to ratify results and processes.* We saw this mechanism at play in case one when data science team members inquired whether computed feature importances matched existing business insights. The convergence between model results and prior knowledge engendered confidence in the models even when their inner

workings were not available for inspection. The model's capability to "discover" already-known business facts inspired trust in its analytic ability. In addition to endorsing results, intuition aids assessing project workflow. In case one, and multiple times during my fieldwork, I saw business actors enquiring into data science processes with an explicit intent to ensure that followed protocols "made intuitive sense"—to the extent that the intuitiveness of data science workflows was considered a way to assess the efficacy of counter-intuitive results. Different from the scrutiny data scientists already employ in their everyday work (e.g., preventing model overfitting), such examinations were considered a way to uncover erroneous decisions (e.g., data reorganization) or configurations (e.g., flawed assumptions).

Deploying intuition as a form of assessment, however, has its own problems. Intuitive results call further attention to the subset of counter-intuitive results. When certain feature importances matched business expectations, business analysts questioned the absence of other expected features. Upholding the validity of intuitive assessment in such situations required a way to explain counter-intuitive findings while not entirely relinquishing the ability for intuitive ratification. This was achieved through a second strategy: *demarcating between algorithmic and human analytical approaches to justify perceived differences.* Data science team argued that, unlike humans, algorithms statistically traverse the uneven contours of data, producing results that may sometimes appear unrecognizable or different. Counter-intuitive findings, they argued, can at times comprise novel forms of knowledge and not model mistakes. Balancing between intuition, counter-intuition, and trust requires work on the part of organizational actors to negotiate the relations between prior knowledge and novel discoveries. Often, intuitive results are made visible to inspire trust in models, while sometimes counter-intuitive results are

posited as moments of new discoveries. The excessive overlap between model results and prior expectations is also problematic at times since intuitive results can stem from overfitted or over-configured models. For example, in a different project, I saw that a model whose results *completely* mirrored existing insights was deemed a failure by business personnel who argued that the model "lacked business intelligence" because it furnished "no new information."

5.4.3 Rationalizing and Reorganizing Data

In case two, we see how actors negotiate and accomplish data credibility in at least two different ways. A first mechanism involves *rationalizing suspect data features.* The data scientist questioned the high salary figures for certain customers with low/bad credit scores. (Differentiating between high/low salaries is itself a matter of interpretation). Assuming that people with low/bad credit do not earn high salaries, he wanted to get rid of this data. Business personnel, however, invoked the fragile relationship between fiscal circumstances and monthly earnings, arguing that people with seemingly high salaries were not atypical but ordinary occurrences in the world of special financing. Such a form of rationalization involved the contextualization of data in prior knowledge to articulate felt, yet practically-oriented, experiences of data inconsistency, unreliability, and illegitimacy. Technical examination and statistical measures are highly visible forms of data credibility arbitration within data science. In this case, and many others, however, I saw that the lived experience of data is a significant yet largely under-studied form of credibility assessment.

A second mechanism comprises *reorganizing data in different ways to mitigate identified problems.* Throughout the project, several problems with the special financing dataset were identified such as issues of consistency (credit ranges varied across affiliates) and

interpolation (leads in the 476-525 range needed approximation around the 500 threshold). Nevertheless, such issues did not obstruct the project. Identified problems were tackled in specific, often creative, ways. The data scientist tried several ways to achieve compatibility between inconsistent credit score ranges. The effect of inconsistency was lessened by the fact that business analysts characterized the credit score of 500 as a significant cutoff (scores greater than 500 were considered highly likely to get special financing approval). There was no need to make all divergent ranges compatible. The leads could simply be restructured into two buckets: above-500 and below-500. Leads in the 476-525 range, the border between the two classes, were now a significant problem. Placing these leads in either bucket required work and discretion to interpolate scores in some manner. The problem was resolved by expunging all the leads in the 476-525 range—a solution considered imperfect but practical.

5.4.4 Managing Interpretability

In case two, two kinds of explanations were discussed: how a model works (i.e., the abstract working of the model's underlying algorithm) and why a model makes specific decisions (i.e., the situated application of the model on a data point). Explaining the in situ working of, for instance, neural networks is difficult and often impossible. The decision process of even relatively simpler models is hard to grasp for large-scale data. Data scientists focused instead on the abstract working of the model's underlying algorithm. Few business personnel (particularly those with technical backgrounds) found such explanations useful. The majority considered them impractical, wanting to understand the rationale behind model decisions more specifically or model complexity more generally. Business personnel's ability to trust results,

as they repeatedly told us, was severely affected when faced with such forms of "opaque intelligence" (Tabarrok, 2015).

At several instances during my fieldwork, and as described in case two, data science team members tried to alleviate this problem by *accentuating the perceived import of model results,* in turn *deemphasizing the need to understand algorithms and models.* Business personnel desired predictive prowess *and* analytic clarity. Data scientists argued for a trade-off between understandability and effectiveness—state-of-the-art models were not entirely inspectable. As one data scientist said, the complexity of models is not a problem but the very reason they work—a resource for model performance instead of a topic for analytic concern. Transparency remained a problematic ideal caught between multiple interpretations of inscrutability (Lipton, 2018). Opacity was often perceived as a function of models' black-boxed nature, necessitating detailed descriptions of algorithmic workings. Even when translucent, models remained recondite—their workings were complex; their results were hard to explain. Underscoring the import and value of results in these circumstances deemphasized complex descriptions and absent explanations. The question changed from how or why models worked to whether or how well they worked. "Implicit trust" took the place of complex descriptions. "Explicit verification" from real-world tests supplanted absent explanations. What remained unresolved, however, was the foreign nature of algorithmic approaches themselves. Models were opaque and abstruse, but also alien—their complexity was described and explained, but not justified (Selbst & Barocas, 2018).

5.5 Implications for Data Science Research and Practice

These findings hold important implications for the growing data science field, both within and beyond CSCW. Trustworthy data science systems are a priority for organizations and researchers alike—evident, for instance, in rubrics for assessing a data science system's production-readiness (Breck et al., 2016) or rules for conducting responsible big data research (Zook et al., 2017). Such forms of advice address a range of sociotechnical challenges, helping data scientists manage aspects ranging from performance evaluation and feature engineering to algorithmic harm and ethical data sharing. As CSCW and critical data studies researchers work to make data science approaches transparent, metrics humanistic, and methodologies diverse, their tools often travel far from their academic and research contexts of development, finding new homes in company servers, business meetings, and organizational work. The insights in this chapter add three further dimensions to the effective practice and management of data science work by explicating how specific tensions problematize trust and credibility in applied data science systems, and how these problems are variously negotiated and resolved.

First, rather than a natural or inevitable property of data or algorithms themselves, the perceived trustworthiness of applied data science systems, as I show in this chapter, is a collaborative accomplishment, emerging from the situated resolution of specific tensions through pragmatic and ongoing forms of work. As my actors repeatedly told us, data are messy and models only approximations (or as the classic line on models has it: "all models are wrong, but some are useful" (Box, 1979)). Perfect solutions were not required. Models just needed to be "good enough" for the purpose at hand (Keller, 2000). Actors' trust in data science did not therefore depend on the flawless nature of its execution, rather on the reasoned practicality of

its results. Much like actors in a "heterarchy" (Stark, 2009), we saw organizational actors treat everyday uncertainties less as impediments and more as sites for justifying the "worth" (Boltanski & Thévenot, 2006) of data, models, and results through actionable strategies. Organizational actors often acknowledged the always-already partial and social nature of data and numbers. Their attempts to negotiate and justify the worth of data science systems were thus aimed at identifying pragmatic ways to make the "best" out of inherently messy assemblages. Such uncertain moments comprise forms of focused skepticism—doubt in and negotiation of specific aspects of data science work (e.g., counter-intuitiveness) require trust in several other aspects (e.g., data sufficiency, model efficacy). This further speaks to the intimate relationship between trust, skepticism, and action: or, as Shapin (1994, p. 19) argues, "distrust is something which takes place on the *margins* of trusting systems."

Data science, particularly from an academic or research perspective, is often imagined from the outside as the work of data scientists interacting around reliable and widely shared tools, norms, and conventions—a "clean room" imagination of data and its relationship to the world. As my cases show, however, corporate data science practice is inherently heterogeneous, comprised by the collaboration of diverse actors and aspirations. Project managers, product designers, and business analysts are as much a part of applied real-world corporate data science as are data scientists—the operations and relations of trust and credibility between data science and business teams are not *outside* the purview of data science work, but *integral* to its very technical operation. A more inclusive approach to the real-world practice of corporate data science helps us understand that while quantified metrics and statistical reasoning remain visible and effective forms of *calculated trust*, the crystallization of trust in applied data science

work is both calculative *and* collaborative. Quantified metrics allow close inspection of data and models, yet numbers appear differently to different actors—sometimes stable, and at other times mutable. Numbers not only signify model performance or validity, but also embody specific technical ideals and business values. Understanding the pragmatic ways of working *with* the plastic and plural nature of quantified trust and credibility metrics can further nuance existing CSCW and HCI research on the design of trustworthy systems (Greis et al., 2017; Knowles et al., 2015; Ribeiro et al., 2016b) and reliable performance metrics (Amershi et al., 2015; Kay et al., 2015; Powers, 2011)—managing numbers is as important as engineering them.

Second, I show how the collaborative accomplishment of trust requires work on the part of diverse experts to translate between different forms of knowledge. For instance, data scientists work to explicate algorithmic approaches to business analysts, and business teams strive to explain business knowledge to data scientists. I see in such forms of translation work a common trait—the recourse to stories and narratives to not only explain algorithmic results (Gabrys et al., 2016; Taylor et al., 2014) but also describe suspect data and model attributes. Narrativization serves various purposes ranging from delineating the abnormal from the ordinary (e.g., what is and is not noisy data) to rendering opaque technicalities natural and commonplace (e.g., models are inscrutable, but so are human brains). As exercises in world-making (Haraway, 2007), narratives invoke specific lifeworlds (Habermas, 1992; Husserl, 1970) to explain what things mean or why they are a certain way. Algorithmically produced numbers may be black boxed, but tales about and around such numbers engender forms of intuitive and assumptive plausibility. While datasets comprise information on people and their practices, people remain largely invisible in data, dismembered into rows, columns, and

matrices of numbers. Forms of narration and storytelling, however, are often all about people, significantly shaping their identity, agency, and forms of life. Narrativization, as a form of doing, implicates data science between reality and possibility, between signal and noise—indeed, between life and data.

These insights on the narrativization of data and results add new dimensions to existing CSCW and HCI research on explainable machine learning systems (Amershi et al., 2015; Ehsan et al., 2021; Ribeiro et al., 2016b; Selbst & Barocas, 2018; Wattenberg et al., 2016) and human perception of data representations and algorithmic working (Dasgupta et al., 2017; Kay et al., 2015; L. M. Koesten et al., 2017; Yang et al., 2018), making visible not only the plurality of reasonings and modes of justification (Boltanski & Thévenot, 2006) that actually subtend applied data science work but also the multiple forms of expertise that constitute such work in complex real-world settings. Workable data, for instance, is computationally accomplished through multiple forms of pre-processing—each attempt at reorganization adds value, but also removes possibilities. Researchers strive to create better tools to identify and resolve issues with real-world data, but even data assumed or made workable by data scientists are sometimes distrusted by other organizational actors. The identification of mess is an exercise not just in statistics and computation, but also narrativization and interpretation. Understanding and articulating the relation between distinct forms of data curation and their interpretational and narrative affordances, for instance, can complement current technical work on data pre-processing and curation—artifacts that are simultaneously partial and practical. Imagined as an exclusively algorithmic venture, data science would appear as the stronghold of data scientists working with specialized and sophisticated computational tools. Acknowledging the domain-

specificity of data, however, surfaces the many other forms of necessary expertise supplied by diverse organizational actors. These different experts influence the development of a data science system in different ways, pulling the project in specific, sometimes contradictory, directions. Understanding the work of these experts can provide new pathways for CSCW, HCI, and critical data studies researchers into the study, design, and management of data science systems.

Third, I show how different experts hold accountable different parts of data science systems, highlighting the distributed and multifarious nature of data science trustworthiness. For instance, data scientists cross-validate results, while business analysts inquire about data scientists' business knowledge and assumptions. In some cases, trust is placed not in the analysis, but in the identity of the analyst. At a few points in my fieldwork, I saw that data and results were assumed correct by business stakeholders because of the trust they placed in specific individuals. Data scientists trusted datasets more if they came from business analysts as opposed to data engineers. Business teams trusted results coming from senior analysts and scientists. As forms of 'social scaffolding,' people's perceived reputation and knowledgeability at times provided working solutions to problems of credibility. Embedded within these different forms of trust valuations are distinct approaches to data science auditing. On the one hand, audits function as a form of *algorithmic witnessing*— backtracking technical procedures to ensure the reliability and validity of model results. As exercises in corroboration, such audits necessitate data science expertise on the part of the auditors. On the other hand, audits contribute to *deliberative accountability* (Jackson, 2006)—situating model contingencies, technical discretion, and algorithmic results in the broader social, cultural, and organizational milieu.

Acknowledging the role of other experts, such audits encompass multiple ways of 'seeing'—of witnessing data science results and processes. Between *algorithmic witnessing* and *deliberative accountability*, I see the everyday work of applied corporate data science betwixt and between algorithms, businesses, calculations, and aspirations—technically valid, but also practically sound.

Juxtaposing algorithmic witnessing with deliberative accountability provides new research pathways into the effective evaluation, governance, and management of data science systems. As CSCW and HCI researchers work to make data science models transparent and results explainable, their focus should include not only unpacking algorithmic black boxes, but also studying how data and models move between teams, products, and services. This enables us to better understand how the inability of organizational actors, sometimes even of data scientists, to understand models is an artifact not only of their black-boxed nature, but also, for instance, of their counter-intuitiveness. This is especially problematic given that a large part of data science's appeal is its ability to surprise us. Even when opened and made tractable, model innards and results remain complex and foreign in their movement between people, practices, and teams. Like other forms of alternate knowledge, data science's alien-ness complicates the attribution of trust and credibility to unaccountable and inscrutable truths—its foreignness sometimes mistaken by some for its incorrectness, and at other times celebrated as a novelty. As researchers make visible the rules comprising models and describe the application of these rules, they also need ways to explain and unpack the complex and alien nature of the rules themselves: "while there will always be a role for intuition, we will not always be able to use intuition to bypass the question of why the rules are the rules" (Selbst & Barocas, 2018).

My findings also suggest some more immediately practical takeaways for data science work in the contexts of academic education, organizational practices, and data science research more generally. In learning environments, would-be data scientists are encultured into data science's professional discourse, largely comprising of technical work such as data pre-processing, model selection, and feature engineering. As students go on to take data science jobs in corporations, their everyday work, however, comprises working not only with data and models, but also with project managers and business stakeholders. The collaborative and heterogeneous nature of real-world data science work remains as of now largely invisible in current data science curricula and training. Several of my actors argued that the data scientists they interacted with lacked, among other things, the "vocabulary" of working with business teams. As James—a senior business team member—put it: "data scientists […are] very eager to crunch numbers […], train the system, and see what kind of output they […can] get" (Interview, 6 November 2017). The incorporation of collaboration (e.g., interacting with non-data-scientists) and translation (e.g., effective communication of results) work into data science curricula and training is thus a good first step to ensure that would-be data scientists not only learn the skills to negotiate the trust in and credibility of their technical work, but also learn to see such forms of work as integral to the everyday work of data science. Or, to put it in terms of sociologists Harry Collins and Robert Evans, real-world applied data science projects require forms of both "contributory" *and* "interactional" expertise (Collins & Evans, 2007).

In corporate organizations, as I show in this chapter, the development of data science systems comprises a combination of algorithmic and deliberative accountability. Integral to both approaches are the need and role of documentation. At the organization, I initially

discovered that there was much emphasis put on code documentation, but the everyday discretionary work of data scientists in pre-processing data, selecting models, and engineering features remained less visible and documented (I attempted to address this at the organization by initiating detailed project and decision documentation). This remains a hindrance not only for forms of inter-organization accountability, but also for the compliance and management of such systems in the wake of laws and policies such as the General Data Protection Regulation (GDPR) and the Right to Explanation in Europe. With current calls for more open documentation, corporate organizations need to document not only algorithmic functions and data variables, but also data decisions, model choices, and interim results. Organizations need to allocate additional resources and efforts to make visible and archive the seemingly mundane, yet extremely significant, decisions and imperatives in everyday data science work.

Lastly, highlighting the existence and role of diverse experts in applied data science projects, my work helps to further unpack the distinction between the designers and users of data science systems in existing CSCW and critical data studies research. Studies of in-use public-facing data science products often work on a clear distinction between designers and users (e.g., Google data scientists made Google Search, which is now used by internet users). Unpacking the design and development work of such corporate systems, however, stands in contrast with rigid binaries—corporate organizations are not monolithically comprised of data scientists working with their algorithmic toolkits to produce computational artifacts. Project managers, product managers, and business stakeholders, as I show, are not merely the "users" of data science systems, but also in part their managers, stakeholders, and designers. Interpretability remains multiple (Lipton, 2018), but so do the people requiring explanations.

As they focus on studying and building interpretable models and trustworthy systems, researchers must also consider *who* attributes trust and requires explanations, *how*, and for *what* purposes. The decision to deploy a data science solution in a real-world product remains a negotiated outcome—one in which data scientists play an important, yet partial, role.

5.6 Conclusion

In this chapter, I showed how tensions of (un)equivocal numbers, (counter)intuitive knowledge, (in)credible data, and (in)scrutable models shape and challenge data science work. I described a range of strategies (e.g., rationalization and decomposition) that real-world actors employ to manage, resolve, and sometimes even leverage problems of trust emanating from these tensions, in the service of imperfect but ultimately pragmatic and workable forms of analysis. As we work to guarantee and foster trust in applied data science, understanding the situated work of trust and credibility negotiations generates new possibilities for research and practice—from the implications of narrative affordances and plurality of model opacity to the management of numbers and leveraging of expertise.

As data science grows, so do its problems of trust. This chapter is my attempt to show that trustworthiness in corporate data science work emerges through a mix of *algorithmic witnessing* and *deliberative accountability*—a calculated but also collaborative accomplishment. The work of data science consists of not just pre-processing and quantification but also negotiation and translation. These processes together, and not any taken singly, ultimately accounts for the trustworthiness and effectiveness of data science.

VI
Conclusion

This chapter marks the end of our journey. The more I thought about what I wanted to say at the end, the more I realized how similar the ending of a thesis is to a data science system. Both are useful but not perfect, reasonable but not neutral, subjective but not irrational, self-contained but not exhaustive, and—to be honest—foreseeable artifacts but not foregone conclusions. Much like our data science practitioners who wish to build working systems, my goal with this chapter was to construct a *working argument*—weaving, not finding, threads that bring things together.

6.1 A quick recap

Before we get to the moral of the story, let me do a quick recap. Data science is often erroneously represented and understood mainly as a technical practice—the work of data, algorithms, models, and numbers alone. Such a technology-centered vision, as I argued, has consequences. It paints a partial picture of data science's everyday practice—technical work gets high visibility while human and organizational work remain invisible. In fact, such a misleading understanding of data science work affects how people imagine and resolve the wider issues arising from data science applications.

In this thesis, however, I described how and why doing data science requires as much human as technical work. And in highlighting the sociotechnical nature of data science practices, I showed that the human and organizational work of data science shape not only the

design and working of data science systems but also the wider implications arising from their (ab)use.

My goal, as I said in the Introduction, was to tell human-centered tales of the everyday world of data science practitioners. In this thesis, I achieved this goal using a combination of two analytic lenses: (1) I focused on how practitioners do data science (instead of, as is the norm, on what they use to do it) and (2) in doing so I paid special attention to the varying (in)visibilities of data science work (and the immediate and wider consequences of those (in)visibilities).

Chapter two **opened a space for a sociotechnical articulation and analysis of data science work.** We analyzed how data scientists learn to see the world as data—as things that can be actionably organized and predictably manipulated with algorithms, models, and numbers. Students must learn both technical and discretionary abilities to do data science. On the one hand, algorithmic tools and rules provide data scientists with stable devices with which to unravel the world. On the other hand, there is a gulf between knowing and doing data science—its everyday practice is far more involved and messier than the simple application of tools and rules. Fusing the social and technical sides of data science, *data vision* enabled us to approach data science as an artful practice—a rule-based, as opposed to a rule-bound, journey into the dataworld, characterized by creativity and improvisation. The stage was thus set—there is an unexplored 'invisible' human side to data science, and it is crucial to its technical practice.

Wittgenstein (1958) said that rules do not contain within themselves the logic of their own application. The rest of the thesis chapters thus became a journey to explore how practitioners navigate the gulf between fully specified rules and the wildly messy world on

which they apply those rules through situated forms of collaboration, creativity, discretion, improvisation, and negotiation.

This expedition led us to the world of corporate data science. We analyzed the everyday work involved in corporate data science projects, examining how practitioners build data science systems while working under complex business and organizational conditions. We did so by looking at three important forms of work involved at different steps in data science projects: formulating data science problems, building working systems, and assessing the trustworthiness of data, models, and numbers. We had already begun to unravel the human face of data science in chapter two, but it was only from chapter three onwards that we started seeing the organizational face of data science.

Chapter three **showed how human and organizational work are crucial to corporate data science practices, starting from the very first step in a project—turning real-world goals into data science problems.** In current discussions of data science, the complex process of problem formulation—translating high-level goals into tractable problems—is often taken for granted. But, as we saw, such a process is far from linear, and success is never guaranteed. Practitioners frequently struggle to turn amorphous goals into well-specified problems; varying aspirations at times push projects in different, sometimes divergent, directions. Organizational imperatives and business goals often lead practitioners to practical, not perfect, problem formulations. Problem formulation work, in fact, has broader implications. Problem formulations can change throughout projects, and the normative implications of systems evolve alongside changes in the different problem formulations. The problems practitioners choose to solve—and *how* and *why* they choose to solve them in specific ways—configure consequential

forms of bias and fairness in system design (often before a line of code is even written!). Ethics of data science have human and organizational, and not just technical, roots.

Post problem formulation comes the work of building data science systems. Chapter four took us further into the system development lifecycle, highlighting how the working of data science systems are deeply affected by the everyday practices of building them. Data, algorithms, models, and numbers are vital to the process of building data science systems. But, as we learned, **the real-world practice of building such systems is remarkably heterogeneous—in no way limited to the work of data scientists or to technical work alone.** Many different practitioners—analysts, executives, managers, and scientists—work together to negotiate crucial design aspects such as what systems should do, how they should work, and how to assess their working. These judgments affect *how* and *why* systems finally work the way they do (or do not work the way others expect them to). Practitioners *choose to make* data science systems work in some ways and not in others. Building such systems is as much a practice of managing expectations, goals, and priorities as of working with data, algorithms, models, and numbers. A data science system's working is entangled with the context and culture of the organization building it and impacted by practitioners' imagination of how their systems can and should work. Practitioners make systems work through an effective combination of human, technical, and organizational work—and not either of these alone.

After building a system comes the task of deciding whether it is ready for deployment—is the rubber ready to hit the road? Central to this task (and to data science projects more generally), as we saw, is the notion of trust. Practitioners spend a lot of money and effort to ensure the credibility of data, models, and results. Established credibility mechanisms (e.g.,

performance metrics or validation methods) engender forms of calculated trust accomplished via quantification. But in chapter five we learned that **establishing trust in data science requires both calculative and collaborative work.** Different practitioners describe issues with and confidence in data, models, and numbers in different ways. Common tensions raise pertinent problems of trust that practitioners manage, resolve, and sometimes leverage in various ways. Calculated trust is central to contemporary discussions and research on responsible data science but taken alone fails to capture the plurality of expertise and modes of justification that typify problems and solutions of trust in data science work. Multiple experts work together to assess systems using diverse criteria and trained judgment in the service of imperfect but pragmatic applications. Organizational actors establish and re-negotiate trust under messy and uncertain conditions through collaborative practices of skepticism, translation, and accountability.

Human and organizational work—and ethical and normative discretion, more generally—do not just show up in certain places in the data science system development lifecycle. As all scientific endeavors, data science is a social practice—shot through with aspirations, compromises, creativity, discretion, and improvisation from start to finish.

6.2 Four high-level takeaways

Until now, I have offered a series of micro- and meso-level analyses of the everyday work of data science—from ethnographic vignettes and empirical findings to conceptual takeaways and practical implications. Taken together, I believe, these help us better understand the human and organizational work of data science, including how such forms of work intersect with more visible forms of technical work. The recap in the previous section brought us back full circle,

reminding us of the questions and issues with which this thesis began and the answers and resolutions with which we now end it.

In this section, I take one final step back to outline a set of four high-level takeaways. In these takeaways, I reflect on three themes. The first two takeaways focus on how this thesis alters our view of the expertise and nature of the work involved in data science. The third takeaway focuses on the gaps between the academic training and professional practice of data science—and how we realistically cannot hope to achieve effective change without also overhauling the way we train data scientists. The last takeaway focuses on the opportunities afforded by (and methods to foster) an engaged form of qualitative-empirical research, like the one in this thesis, for shaping the design and management of data science systems.

6.2.1 A team effort: From data scientists to data science practitioners

In contemporary discussions and research on responsible data science, one specific notion has quickly gained considerable popularity and momentum—that of the *stakeholder*. It is common today to say that data scientists build systems, some people use them, and everyone else has, in one way or another, a stake in the system. Data scientists, users, stakeholders—the holy trinity.

While not incorrect, such a narrow categorization of the humans involved in data science is harmful. It severely limits how we imagine concerns, engagements, and interventions within the space of responsible data science. The eventual goal comes across largely as that of ensuring that data scientists care about their users and think about their stakeholders. Unsurprisingly, researchers now call for more ‘democratization’ and ‘participation’ in data science—involving users and stakeholders in system design practices.

These are noble, much-needed goals. But we must recognize that not all stakeholders or users are alike. In this thesis I demonstrated how and why the everyday practice of building data science systems is a team effort. Data scientists do not build systems. Data science *practitioners* build systems—through the collective work of the many different experts involved in the everyday work of data science. Managers, designers, analysts, and executives—as stakeholders—of course have much to lose or gain when systems fail or work. But these 'stakeholders' also bring to the table different forms of expertise that are required to build working systems—sometimes to the same or greater extent than that of data scientists themselves. In fact, at times stakeholders *are* the users. The business team involved in the design of a data science solution, as we saw, may itself be the system's intended user. The terms 'users' and 'stakeholders' fail to capture the dual role certain practitioners (data scientists but also others) play in a corporate data science project, in turn undermining their agency to bring about change.

The change in vocabulary—moving from data scientists to data science practitioners—is key, marking a crucial shift in how we can and *must* approach the range of expertise and work involved in doing data science. Data scientists are not the only ones who do data science. The everyday practice of data science requires work on the part of multiple professional groups—data scientists and software engineers, but also product managers, business analysts, project managers, product designers, business heads, and corporate executives. There is more than one way to do, to practice data science. Each way has its own flaws and problems but also comes with its own affordances and opportunities for change and engagement.

We need more participation in data science. But we must recognize that the everyday practice of corporate data science is always-already participatory—just not how we presently imagine and desire participation in data science. I do not wish to argue for approaching or appreciating corporate projects as sites of participatory design. Instead, I wish for us to acknowledge the role different experts play in everyday corporate data science work—experts who are often labelled merely as stakeholders (or, at times, users) when, in fact, they significantly shape key design decisions and discussions concerning the nature and working of systems.

Unpacking the collaborative work in corporate projects as forms of participation raises several pertinent questions—questions that open new sites of engagement and intervention for responsible data science work. What happens when we consider business teams as data science users? What forms of participation should characterize the collaboration between data science and other corporate teams? Are there specific stages within the data science design process in which we may explicitly want or not want the participation of certain practitioner groups? Arising from a heightened sensitivity towards the diversity of work and plurality of expertise involved in the everyday practice of data science, such questions, I hope, will force us to think carefully about the forms of participation that do or do not exist in data science and that we do or do not want in data science.

6.2.2 Not just that: Multiple ways to imagine and do data science

Not all questions, however, remain unanswered. This thesis has helped to illustrate potential answers to a vital question: how can we, as researchers, leverage existing forms of human and organizational work to shape how corporate practitioners build data science systems?

We learned that not all data science problems are the same. Recognizing the diversity of work involved in everyday data science practices sensitizes us to the *different causes* of data science problems. Not all issues—such as those of accountability, bias, explainability, fairness, method, transparency, or trust—stem from technical work or the work of data scientists alone. Some are borne out of technical attributes, but others out of business mindsets, organizational culture and structure, prior assumptions and knowledge, practical imperatives, and resource constraints. It thus makes little sense to expend all our attention and resources in one place (the technical domain). We must strive for a holistic analysis and resolution of data science problems—problems that are as much business, design, engineering, management, product, or organizational issues as they are algorithmic, computational, or mathematical ones. There are multiple different kinds of roadblocks on the path towards responsible data science.

In fact, there are multiple paths to imagining and doing responsible data science itself. If data science problems are simultaneously human, technical, and organizational, why should their solutions be limited to one form of work? Multifaceted problems require multidimensional solutions. Technical work alone cannot solve all our problems. Better algorithms, robust measures, and detailed checklists are vital to addressing data science problems. But we cannot stop there. This thesis has shown us new ways to imagine and do responsible data science, highlighting, for instance, how addressing challenges with project timelines, organizational culture, or project reporting are also valid, and at times necessary, ways to solve data science problems.

Operationalizing human- and organizational-centered solutions, however, is not easy. Doing so requires us to first acknowledge the role and import of different practitioners in

corporate projects. We must leverage existing forms of organizational work (e.g., how can we use the already-central role that project managers play in corporate projects to advance responsible data science work?) and enable new collaborations (e.g., creating avenues for participation for practitioners that lack representation in decisions and discussions). The goal is not to have *every*one participate in project meetings, discussions, or decisions. But instead to first try to understand *how, why,* and *when* different practitioners currently participate in projects (including who gets to decide the nature of such participation, and how) and then work to reimagine the flow and representation of expertise in corporate data science projects.

Involving different practitioners in responsible data science endeavors is key to success. A big reason for this is that each practitioner, as an expert, has their own ways of knowing and doing. We saw how practitioners describe issues in different ways since different experts imagine and use distinct forms of assessment. But that is not all. The multiple modes of justification at play in corporate data science projects also represent the vast differences between the seldom-spoken normative goals and concerns of individual practitioners. Each practitioner involved in the everyday practice of data science (including us, the researchers) has their own normative understanding of right/wrong and good/bad. The on-the-ground reality of data science work is not characterized by one kind of right/wrong or good/bad. As we saw in chapter three, choosing one form of doing good may in fact hurt someone else!

As researchers, we wish to differentiate between (capital R) *Right* and (capital W) *Wrong* ways of doing data science. This is a noble goal. But before we attempt to outline how data science must be done, it is important to first analyze how different notions of right/wrong and good/bad manifest in everyday data science work, and why certain kinds of normativity

prevail in practice (especially when in hindsight we can easily envision better alternatives). Doing responsible data science requires first acknowledging the fact that responsibilities themselves are of different, often divergent, kinds. Being responsible towards shareholders is not the same as being responsible towards employees, users, data subjects, or society at large. We must think long and hard about how best to, and if we want to, balance these different kinds of responsibilities as we slowly work towards shaping the future of responsible data science work.

6.2.3 Alternative approaches to data science training: Human and organizational work

While most of this thesis focused on corporate data science practices (chapters three, four, and five), it dealt in part with academic practices of data science learning and professionalization (chapter two). The difference between the two analyses was stark, to say the least. For instance, chapter two never mentioned aspects such as organizational work, the multiplicity of data science practitioners, or the need to balance between different aspirations and goals.

The absence of such notions in my analysis of the academic practices of data science training was not an oversight but a stark reminder of existing gaps between the academic training and professional practice of data science. Data science courses largely focus on technical training. Human work, while acknowledged, is rendered secondary to the seemingly 'real' work of, say, building models. This is evident in how instructors demonstrate, and require students to document and present, data science work—question, data collection, model building, model testing, answer. Open conversations about the discretionary work essential to the effective practice of data science are rarely encouraged.

In making visible the human work essential to the effective technical practice of data science, this thesis calls for a sociotechnical approach to data science training in which students learn to reflect on their decisions to better grasp the strengths and limits of their tools and findings. Such an approach can enable students to see beyond their tool and rules towards the broader impacts of subjective judgment on data science and its necessity in everyday practice.

Recently, I have had fellow researchers push back on such remarks. They argue that it is impossible to teach everything in one course. A course on ML, they reckon, should largely focus on teaching different algorithmic methods and tools to students as opposed to exposing them to the functioning of corporate organizations. I agree. We cannot teach everything in one course. However, no matter what we choose to teach in a course, we must ensure that students learn data science concepts and ways of doing within the real-world contexts of their applications. If it is impossible to do the technical work of data science without also doing the human work of data science, then how can we expect that it is possible to teach students what to do with new and shiny algorithms without also teaching them the discretionary nature of their applications or the real-world contexts of their potential (ab)use?

Moving on, when human work itself is mostly overlooked in data science training, we can only imagine the plight of organizational work. In classrooms, students see data science as a homogeneous practice done *by* and *for* data scientists. As they go on to take data-science related jobs, however, students realize—much like what we learned in this thesis—that they need to work with not only scientists and engineers but also managers, analysts, and designers. Effectively collaborating with other experts requires knowing how to navigate different forms of professional knowledge and practices—a skill neglected in data science training. The work

of justifying assumptions and results is not *outside* the purview of data science but *integral* to its very technical operation. An organizational approach to data science training can help students learn that data science is not an isolated scientific effort, but deeply entangled with business and product practices.

We must make a conscious effort to make classrooms more like the outside. This is not to say that we should give up scientific ideals in the name of business goals. Instead, a sociotechnical-organizational approach to data science training exposes students to state-of-the-art methods, tools, frameworks, and platforms, while simultaneously ensuring that students understand how their work fits into and must speak to existing practices and ways of knowing.

6.2.4 Up close and personal: Engaging with data science practitioners

The research in this thesis is based on ethnographic research. My approach to ethnographic research involves not just observing the work of data science practitioners but *working with* them as part of their everyday practice through immersive participant observation. This is the reason why I decided to work as a data scientist at DeepNetwork while doing research. My commitment to this research style is part of my larger goal to advance an *engaged form of qualitative-empirical research* that addresses and contributes to the technical foundation, organizational dimensions, and human underpinnings of data science.

My engaged form of research has in turn helped me to develop generative collaborations with data science practitioners, enabling me to shape real-world data science practices. For instance, during corporate fieldwork, I worked as a data scientist at DeepNetwork, but my contributions were not just technical. During and after fieldwork, the company invited me to give talks to their scientists, managers, and designers, in which I discussed with them the

normative aspects of their work. For instance, I pinpointed how certain business decisions (e.g., project scope negotiation) impact the technical working of ML models (e.g., choice and measure of performance metrics). This led the company to start new archival practices of recording discretionary choices and team decisions in project documentation. The company's CTO Justin and director of data science Martin have expressed interest to continue working with me in my future research projects. In fact, quite recently the company has tried to make explicit the work of problem formulation in the service-level agreements (SLAs) between the data science and business teams. What this means is that at the start of a new data science project, the business and data science teams devote some amount of time on the mapping between business goals and data science problems.

Moving on, my academic fieldwork involved an ethnographic study of how students learn and are taught data science. During this time, I involved myself with the research practices of a faculty working at the intersection of ML, NLP, and humanities (the same faculty who taught one of the two courses that I look at in this thesis). Over a year, I worked with him to unpack the role of human judgment in practices of analyzing text and images as data and numbers. This collaboration helped steer his lab's focus towards accounting for the impact of human discretion on data curation and processing and of subjective reasoning on assessments of algorithmic success and failure.

Finally, more recently, Microsoft Research AI (MSR AI) invited me to examine the challenges faced by Microsoft product teams in addressing the ethical aspects of the AI-based systems that they develop. I did an interview-based analysis of the product development practices of different Microsoft product teams, in turn highlighting how aspects of their

everyday work impacted their ability to scope and resolve FATE in AI issues (the results of this study are currently not public). I charted practical programs of action to address identified issues; several of which were incorporated into Microsoft's internal organizational and technical initiatives on responsible AI.

Because of the impact my research has had on real-world data science work, I am often asked to describe elements of my research style that have made such impact possible. Listing these elements out has not been easy because when I began this research, I was not thinking about developing collaborations or affecting changes along the way. I wanted to partake in the everyday practice of data science, yes, but mainly to gain immersive access to the everyday reality of data science work. In that sense, I developed my engaged research style along the way—mainly through trial and error.[51]

Over time, however, I have been able to pinpoint three distinct elements of my research style that have been immensely helpful for me when collaborating with other data science practitioners. I briefly describe these below. But before I do that, I wish to make it clear that I am not trying to argue that my way is the "best" way to research data science practices. You neither must *do* data science nor *work with* data scientists to do good, impactful research (most of the critical data studies, FATE in AI, and human-centered data science scholarship that I draw from in this thesis is proof of that). However, what I wish to argue is that there *are* ways to effectively collaborate, engage, and work with data science practitioners—to turn

[51] There are both pros and cons to such an engaged style of research. If you recall, in subsection 3.1.2 of the first chapter (Introduction), I described the limitations of playing dual roles of researcher and practitioner.

constructive critiques into in-practice elements. I highlight the three elements of my research philosophy below to help those who may wish to partake in similar efforts.

First, we must acknowledge that all practices are messy and can, at best, only partially represent the complexities of the world. This includes the practice of doing data science as well as the practice of researching data science. Each way of doing and knowing is biased, political, has limitations and blind spots, and does necessary violence to the rich world to enable certain actionabilities. Such humility and reflexivity go a long way in ensuring that when you study others, you are not just looking at their faults or fixating on their failures. *All* practices are deficient—we all make mistakes. That does not mean that, say, data scientists should get a free pass when they do harm. Of course not. But pointing fingers and shaming failures, in my opinion, are not going to lead us anywhere. The focus must not be on trying to find mistakes but instead, for instance, on trying to understand how and why mistakes happen in everyday practices of data science (indeed—what even counts as a mistake in the moment or in hindsight). An engaged style of research strives to understand and connect with the lived experiences of data science practitioners rather than seeking to examine, observe, and interrogate them.

Second, it is important to keep in mind that there are multiple ways to *do* data science and thus, by extension, to engage with it. As I have made clear, we must move away from seeing data science as a solely scientific practice done mainly by data scientists to seeing it as a sociotechnical-organizational practice involving different practitioners. Such a vision of data science opens new ways to affect change. You realize that you are not restricted to engaging just with data scientists. Business analysts, project managers, product designers, and corporate

executives, for instance, are all key roles in corporate data science projects—each of these roles is a good place to start. In fact, no matter what role you choose to start with, doing data science requires collaborating with other roles. Your options to affect changes are not limited to your interactions with those who are in the role that you choose to look at or in the same role as you. Every interaction is a possible engagement.

Lastly, we must not forget that effectively engaging with data science practitioners requires translating between different forms of knowledge. Whether it be between data science and social science or data science and project management, knowing how to bridge discourses is a powerful engagement device. But developing this ability is not easy. I was fortunate to have an interdisciplinary training in computer science, information science, and STS, but I still had to spend a lot of time learning the ropes of doing data science during academic ethnography and through online data science certification courses. Even now, I do not consider myself an expert in doing this. It is a continuous learning process—one in which you get better with time. Note that the ability to build bridges is different from the ability to outline design implications (though the latter is not possible without the former). A design implication is one kind of translation—taking a piece of information from one discourse and transforming it into another. Locating the source of a FATE issue within the everyday work of data scientists or project managers is another kind of translation—mapping between a high-level concern in a discourse and a situated action in another. Remember that the work we put in to hone the ability to speak to different data science practitioners does not just represent our investment in a research skill. It also highlights that we understand that the everyday practice of data science is a

sociotechnical mess—one in which solutions to pertinent problems often lie betwixt and between discourses and changes in one discourse can propagate to others.

6.3 Future Research

This thesis has helped to advance the sociotechnical understanding of data science work, impacting not only how we understand the discretionary and normative dimensions of the everyday practice of data science but also the wider organizational practices of doing responsible data science. Much of the thesis has focused on the work of building data science systems, and what different forms of human, technical, and organizational work tell us about wider normative issues with data science systems.

Below I describe two projects that point towards an important future research direction that complements the research presented in this thesis, extending my work on analyzing how teams build data science systems to understanding what happens when data science systems themselves become team members. This is, of course, not a far-fetched reality but an increasingly common occurrence in different professions and organizations that are now moving towards an *AI-first* sensibility to augment professional decision making through human-AI collaboration. Such collaborations happen broadly in two important contexts.

6.3.1 AI-as-a-service: The changing nature of data science work

The first context of human-AI collaboration is found within the data science field itself where AI and ML-driven application programming interfaces (APIs), platforms, tools, and libraries are increasingly becoming a routine part of the everyday work of building data science systems. Traditionally, the ability to build and deploy AI systems was limited to a select few. This is not

the case anymore. Platforms run by Amazon, Google, Microsoft, and others have made AI accessible through cost-effective and easy-to-use products. AI-as-a-service opens new opportunities for creative applications, but also raises unique challenges. For instance, the drag-and-drop style of AI possible on these platforms makes it possible even for non-data-scientists to build and deploy models, implicating thorny issues of expertise, accountability, and trust. Emerging tools such as *AutoAI* further streamline the human work of data science for data scientists—upload data, select labels, and click buttons to get already-trained, optimized models. Black boxing the artful nature of everyday data science work, AI platforms further complicate practitioners' ability to trust their data, models, and results.

6.3.2 Hybrid work: The future of human-AI collaboration

The second context of human-AI collaboration occurs in areas far removed from data science where new, shiny, and often black boxed AI tools are increasingly shaping how professionals make everyday decisions. Data science tools, specifically AI tools, are now finding new homes in domains such as business, healthcare, law enforcement, and agriculture. AI, however, does not just complement professional work, but transforms its very nature to that of *hybrid work*, characterized by human-AI collaboration. Doctors, traders, and police officers now use AI to make everyday decisions, raising pertinent FATE issues and governance challenges. In fact, today AI tools and systems tremendously shape out very access to information. The nature of human-AI interaction and collaboration—of *hybrid work*—varies considerably across domains given differences in work practices and normative values. General findings about human-AI interaction and collaboration must be supplemented with a deeper understanding of contextual practices, professional commitments, and normative goals across different domains. It is

imperative that we analyze the everyday work of AI-based decision making across different professional domains to better understand how to effectively address the affordances and limitations of hybrid work, particularly in professions far removed from data science.

Analytic approach. It is important to note that researchers have been working on examining the implications of such forms of hybrid work (both within and outside data science fields) for some years now, especially in domains such as healthcare, law enforcement, and the criminal justice system. The approach that I have in mind—and one that I have demonstrated in this thesis—however is different. In these future endeavors, we must analyze and engage with these forms of hybrid work in a deeply human- and organization-centered manner, as situated and ongoing forms of everyday work on the part of people and professionals. My goal with outlining these research projects (and this research approach) is to help us break away from the current regime of thinking in which high-level goals such as fairness, interpretability, explainability, and trust are seen as already-given, already-stable *a priori* categories that we just need to implement in data science systems to do responsible AI.

Quite the opposite. Using the lens of *everyday work,* we must approach these and other goals as things-in-the-making, as constantly evolving, shifting, and contested values that people negotiate and arrive at in their everyday interactions with AI-based systems. The larger question underlying this research direction is the following: *how do people work with, and sometimes against, AI-based tools in their everyday and professional practices?* Answering this question, I believe, can take us yet another step closer to my bigger goal of fostering responsible data science futures by engaging with and learning from people's everyday lived experiences of doing and living with data science.

BIBLIOGRAPHY

Abbott, A. (2014). *The System of Professions: An essay on the division of expert labor*. University of Chicago Press.

Ackerman, M. S. (2000). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Hum.-Comput. Interact.*, *15*(2), 179–203. https://doi.org/10.1207/S15327051HCI1523_5

Agre, P. (1997a). *Computation and Human Experience*. Cambridge University Press.

Agre, P. (1997b). Towards a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In G. C. Bowker, S. L. Star, W. Turner, & L. Gasser (Eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide* (pp. 131–158). Psychology Press.

Agre, P., & Chapman, D. (1990). What are plans for? *Robotics and Autonomous Systems*, *6*(1), 17–34.

Agre, P., & Chapman, D. (1987). PENGI: An Implementation of a Theory of Activity. *Proceedings of the Sixth National Conference on Artificial Intelligence*, 286–272.

Almklov, P. G. (2008). Standardized Data and Singular Situations. *Social Studies of Science*, *38*(6), 873–897. https://doi.org/10.1177/0306312708098606

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmerman, T. (2019). Software Engineering for Machine Learning: A Case Study. *International Conference on Software Engineering (ICSE 2019) - Software Engineering in Practice Track*, 291–300.

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, 337–346.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 1–13.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Wired. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Aradau, C., & Blanke, T. (2015). The (Big) Data-security assemblage: Knowledge and critique. *Big Data & Society*, *2*(2).

Aragon, C., Hutto, C., Echenique, A., Fiore-Gartland, B., Huang, Y., Kim, J., Neff, G., Xing, W., & Bayer, J. (2016). Developing a Research Agenda for Human-Centered Data Science. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 529–535.

Balka, E., & Wagner, I. (2006). Making Things Work: Dimensions of Configurability As Appropriation Work. *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 229–238.

Bannon, L. J. (1995). The Politics of Design: Representing Work. *CACM*, *38*(9), 66–68.

Barocas, S. (2014). Data Mining and the Discourse on Discrimination. *Proceedings of Data Ethics Workshop at the 2014 ACM Conference on Knowledge Discovery and Data-Mining (KDD), August 24*, 1–4.

Barocas, S., & boyd, danah. (2017). Engaging the Ethics of Data Science in Practice. *Communications of the ACM*, *60*(11), 23–25.

Barocas, S., Bradley, E., Honavar, V., & Provost, F. (2017). *Big Data, Data Science, and Civil Rights*. A White Paper Prepared for the Computing Community Consortium Committee of the Computing Research Association. https://cra.org/ccc/resources/ccc-led-whitepapers/

Barocas, S., & Levy, K. (2020). Privacy Dependencies. *Washington Law Review*, *95*(2), 555–616. https://digitalcommons.law.uw.edu/wlr/vol95/iss2/4

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review*, *671*.

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2).

Baumer, E. P. S. (2017). Toward human-centered algorithm design. *Big Data & Society*, *4*(2), 1–12.

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1).

Binns, R., Kleek, M. Van, Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, 1–14.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291. https://doi.org/10.1016/j.ijhcs.2006.11.016

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2005). Affect: From Information to Interaction. *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, 59–68. https://doi.org/10.1145/1094562.1094570

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

Boellstorff, T., & Maurer, B. (Eds.). (2015). *Data, Now Bigger and Better!* Prickly Paradigm Press.

Bolin, G., & Schwarz, J. A. (2015). Heuristics of the algorithm: Big Data, user interpretation and institutional translation. *Big Data & Society*, *2*(2).

Boltanski, L., & Thévenot, L. (2006). *On Justification: Economies of Worth*. Princeton University Press.

Bowker, G. C. (2013). Data Flakes: An Afterword to "Raw Data" Is an Oxymoron. In L. Gitelman (Ed.), *"Raw Data" Is an Oxymoron* (pp. 167–171). MIT Press.

Bowker, G. C. (2014). The Theory/Data Thing. *International Journal of Communication*, *8*, 1795–1799.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and Scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227.

Brayne, S. (2017). Big Data Surveillance: The Case of Policing. *American Sociological Review*, *82*(5), 977–1008. https://doi.org/10.1177/0003122417725865

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2016). What's your ML Test Score? A rubric for ML production systems. *Reliable Machine Learning in the Wild - NIPS 2016 Workshop, NIPS*.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12.

Busch, L. (2014). A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets. *International Journal of Communication*, *8*, 1727–1744.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step by step data mining guide*.

Charmaz, K. (2014). *Constructing Grounded Theory (Introducing Qualitative Methods series) 2nd Edition*. Sage.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, *49*(5), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Clarke, M. F. (2015). The Work of Mad Men that Makes the Methods of Math Men Work: Practically Occasioned Segment Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*, 3275–3284.

Cohen, G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*(7), 1139–1147.

Cohen, M. D. (2007). Reading Dewey: Reflections on the study of routineReading Dewey: Reflections on the study of routine. *Organization Studies*, *28*(5), 773–786.

Cohn, M. L. (2019). Keeping Software Present Software as a Timely Object for STS Studies of the Digital. In J. Vertesi & D. Ribes (Eds.), *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 423–446). Princeton University Press.

Coletta, C., & Kitchin, R. (2017). Algorhythmic governance: Regulating the 'heartbeat' of a city using the Internet of Things. *Big Data & Society*, *4*(2), 2053951717742418. https://doi.org/10.1177/2053951717742418

Collins, H. M. (1985). *Changing Order: Replication and Induction in Scientific Practice*. Sage.

Collins, H. M. (2010). *Tacit and Explicit Knowledge*. University of Chicago Press.

Collins, H. M., & Evans, R. (2007). *Rethinking Expertise*. University of Chicago Press.

Cook, J. (2009). Ethics of Data Mining. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 783–788). IGI Global.

Crawford, K. (2015). Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values*, *41*(1), 77–92.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*, 311–313.

Cukier, K., & Mayer-Schönberger, V. (2013). The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs*, *92*(3), 28–40.

Currie, M., Paris, B. S., Pasquetto, I., & Pierre, J. (2016). The conundrum of police officer-involved homicides: Counter-data in Los Angeles County. *Big Data & Society*, *3*(2), 1–14.

Danielson, P. (2009). Metaphors and Models for Data Mining Ethics. In E. Eyob (Ed.), *Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions* (pp. 33–47). IGI Global.

DARPA. (2019). *AI Next Campaign*. Defence Advanced Research Projects Agency (DARPA). https://www.darpa.mil/work-with-us/ai-next-campaign

Dasgupta, A., Burrows, S., Han, K., & Rasch, P. J. (2017). Empirical Analysis of the Subjective Impressions and Objective Measures of Domain Scientists' Visual Analytic Judgments. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1193–1204.

Daston, L., & Galison, P. (1992). The Image of Objectivity. *Representations*, *40*, 81–128.

Daston, L., & Galison, P. (2007). *Objectivity*. MIT Press.

Davenport, T. H. (2014). *big data @ work: Dispelling the Myths, Uncovering the Opportunities*. Harvard University Press.

Desrosieres, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press.

Dewey, J. (1922). *Human nature and conduct: An introduction to social psychology*. H. Holt & Company.

Dewey, J. (1939). *Theory of Valuation*. University of Chicago Press.

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*, *56*(12), 64–73.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, *2*(2), 1–6.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, *55*(10), 78–87.

Domingos, P. (2015). *The Master Algorithm*. Basic Books.

Dourish, P. (2001). *Where the Action is: The Foundations of Embodied Interaction*. MIT Press.

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 1–11.

Dourish, P., & Cruz, E. G. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society*, *5*(2), 1–10.

Dudhwala, F., & Larsen, L. B. (2019). Recalibration in counting and accounting practices: Dealing with algorithmic output in public and private. *Big Data & Society*, *6*(2).

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *ArXiv*, *2107.13509*, 1–24.

Elish, M. C., & boyd, danah. (2018). Situating methods in the magic of Big Data and AI. *Communication Monographs*, *85*(1), 57–80. https://doi.org/10.1080/03637751.2017.1375130

Espeland, W. N., & Stevens, M. L. (2008). A Sociology of Quantification. *European Journal of Sociology / Archives Européennes de Sociologie / Europäisches Archiv Für Soziologie*, *49*(3), 401–436.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, *39*(11), 27–34.

Feldman, M. S., & Orlikowski, W. J. (2011). Theorizing Practice and Practicing Theory. *Organization Science*, *22*(5), 1240–1253.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing Organizational Routines as a Source of Flexibility and Change. *Administrative Science Quarterly*, *48*(1), 94–118. https://doi.org/10.2307/3556620

Forsythe, D. E. (1993a). The Construction of Work in Artificial Intelligence. *Science, Technology & Human Values*, *18*(4), 460–479. https://doi.org/10.1177/016224399301800404

Forsythe, D. E. (1993b). Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science*, *23*(3), 445–477. https://doi.org/10.1177/0306312793023003002

Forsythe, D. E. (2001). *Studying Those who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press.

Franks, B. (2012). *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. John Wiley and Sons.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: an Overview. *AI Magazine*, *13*(3), 57–70.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gabrys, J., Pritchard, H., & Barratt, B. (2016). Just good enough data: Figuring data citizenships through air pollution sensing and data stories. *Big Data & Society*, *3*(2), 1–14.

Garfinkel, H. (1964). Studies of the Routine Grounds of Everyday Activities. *Social Problems*, *11*(3), 225–250.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Prentice Hall.

Geiger, S. R. (2017). Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717730735

Gieryn, T. F. (1983). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, *48*(6), 781–795. http://www.jstor.org/stable/2095325

Gillespie, T. (2011). *Can an algorithm be wrong? Twitter Trends, the specter of censorship, and our faith in the*

algorithms around us. Culture Digitally: Examining Contemporary Cultural Production [Blog].

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167–194). MIT Press.

Gillespie, T. (2016). #trendingistrending: when algorithms become culture. In R. Seyfert & J. Roberge (Eds.), *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies* (pp. 64–87). Routledge.

Gitelman, L. (2006). *Raw Data is an Oxymoron*. MIT Press.

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw Data" Is an Oxymoron* (pp. 1–14). MIT Press.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transactions.

Goodwin, C. (1994). Professional Vision. *American Anthropologist*, *96*(3), 606–633.

Greis, M., Avci, E., Schmidt, A., & Machulla, T. (2017). Increasing Users' Confidence in Uncertain Data by Aggregating Data from Multiple Sources. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, 828–840.

Grimmer, J. (2015). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, *48*(1), 80–83.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models. *ArXiv Preprint*, *1802.01933*.

Habermas, J. (1992). *Autonomy and Solidarity: Interviews*. Verso.

Hacking, I. (1990). *The Taming of Chance*. Cambridge University Press.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459.

Hand, D. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 317–356.

Hand, D. (2006). Protection or Privacy? Data Mining and Personal Data. In W.-K. Ng, M. Kitsuregawa, J. Li, & K. Chang (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1–10). Springer Berlin Heidelberg.

Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, *7*(1), 2053951720926558. https://doi.org/10.1177/2053951720926558

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599.

Haraway, D. (2007). Modest_Witness@Second_Millenium. In *Modest_Witness@Second_Millenium.FemaleMan©_Meets_Oncomouse™*. Routledge.

Harding, S. (2001). Feminist Standpoint Epistemology. In M. Lederman & I. Bartsch (Eds.), *The Gender and Science Reader* (pp. 145–168). Routledge.

Harper, R. (2000). The social organization of the IMF's mission work: An examination of international auditing. In M. Strathern (Ed.), *Audit cultures: Anthropological studies in accountability, ethics, and the academy* (pp. 21–53). Routledge.

Heidegger, M. (1954). The Question Concerning Technology. In W. Lovitt (Ed.), *The Question Concerning Technology and Other Essays* (pp. 3–35). Harper Torchbooks.

Helgesson, C.-F. (2010). From dirty data to credible scientific evidence: Some practices used to clean data in large randomised clinical trials. In C. Will & T. Moreira (Eds.), *Medical Proofs, Social Experiments: Clinical Trials in Shifting Contexts* (pp. 49–67). Ashgate.

Hildebrandt, M. (2011). Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology*, *24*(4), 371–390.

Hogan, M. (2015). Data flows and water woes: The Utah Data Center. *Big Data & Society*, *2*(2).

Hohman, F., Wongsuphasawat, K., Kery, M. B., & Patel, K. (2020). Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376177

Husserl, E. (1970). *The Crisis of European Sciences and Transcendental Philosophy* (D. Carr (Trans.)). Northwestern University Press. https://doi.org/10.1017/CBO9781107415324.004

IBM. (2012). *Bringing Big Data to the Enterprise*. http://www-01.ibm.com/software/in/data/bigdata

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

Ingold, T. (2010). The textility of making. *Cambridge Journal of Economics*, *34*(1), 91–102.

Introna, L. D. (2016). Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology, & Human Values*, *41*(1), 17–49.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jackson, S. J. (2006). Water Models and Water Politics: Design, Deliberation, and Virtual Accountability. *Proceedings of the 2006 International Conference on Digital Government Research*, 95–104.

Jackson, S. J., & Kang, L. (2014). Breakdown, obsolescence and reuse. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 449–458.

Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). The Meaning and Measurement of Bias: Lessons from Natural Language Processing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 706. https://doi.org/10.1145/3351095.3375671

Joerges, B., & Shinn, T. (2001). A Fresh Look at Instrumentation an Introduction. In B. Joerges & T. Shinn (Eds.), *Instrumentation Between Science, State and Industry* (pp. 1–13). Springer Netherlands.

Joyce, J. (1922). *Ulysses*. Project Gutenberg.

Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. (2018). ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 88–97. https://doi.org/10.1109/TVCG.2017.2744718

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, 347–356.

Keller, E. F. (2000). Models of and Models for: Theory and Practice in Contemporary Biology. *Philosophy of Science*, *67*(September 2000), 72–86.

Kery, M. B., Horvath, A., & Myers, B. (2017). Variolite: Supporting Exploratory Programming by Data Scientists. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1265–1276.

Kitchin, R. (2014a). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage Publications.

Kitchin, R. (2014b). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12.

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication, and Society*, *20*(1), 14–29.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1–10.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Ludwig, J., & Mullainathan, S. (2016, December). *A Guide to Solving Social Problems with Machine Learning*. Harvard Business Review. https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning

Klemp, N., McDermott, R., Raley, J., Thibeault, M., Powell, K., & Levitin, D. J. (2008). Plans, Takes, and Mis-takes. *Outlines: Critical Practical Studies*, *10*(1), 4–21.

Knowles, B., Harding, M., Blair, L., Davies, N., Hannon, J., Rouncefield, M., & Walden, J. (2014). Trustworthy by Design. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1060–1071.

Knowles, B., Rouncefield, M., Harding, M., Davies, N., Blair, L., Hannon, J., Walden, J., & Wang, D. (2015). Models and Patterns of Trust. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 328–338.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Koesten, L., Kacprzak, E., Tennison, J., & Simperl, E. (2019). Collaborative Practices with Structured Data: Do Tools Support What Users Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300330

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017). The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1277–1289.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The*

Knowledge Engineering Review, *21*(1), 1–24.

Latour, B. (1987). *Science in Action*. Harvard University Press.

Latour, B. (1990). Postmodern? No, Simply Amodern. Steps Towards an Anthropology of Science: An Essay Review. *Studies in History and Philosophy of Science Part A*, *21*(1), 145–171.

Latour, B., & Woolgar, S. (1985). *Laboratory Life: The Construction of Scientific Facts* (2nd ed.). Princeton University Press.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Leonardi, P. M. (2009). From Road to Lab to Math: The Co-evolution of Technological, Regulatory, and Organizational Innovations for Automotive Crash Testing. *Social Studies of Science*, *40*(2), 243–274. https://doi.org/10.1177/0306312709346079

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science*, *76*, 737–749.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 1–11.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821.

Levin, N. (2014). Multivariate statistics and the enactment of metabolic complexity. *Social Studies of Science*, *44*(4), 555–578.

Lewis, J., Atkinson, P., Harrington, J., & Featherstone, K. (2012). Representation and Practical Accomplishment in the Laboratory: When is an Animal Model Good-enough? *Sociology*, *47*(4), 776–792.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lipton, Z. C., & Steinhardt, J. (2018). Troubling Trends in Machine Learning Scholarship. *ArXiv*, *1807.03341*, 1–15.

Luca, M., Kleinberg, J., & Mullainathan, S. (2016, January). *Algorithms Need Managers, Too*. Harvard Business Review. https://hbr.org/2016/01/algorithms-need-managers-too

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Lynch, M. (1985). Discipline and the Material Form of Images: An Analysis of Scientific Visibility. *Social Studies of Science*, *15*(1), 37–66.

Lynch, M. (1988). The externalized retina: Selection and mathematization in the visual documentation of objects in the life sciences. *Human Studies*, *11*, 201–234.

Mackenzie, A. (2015). The production of prediction: What does machine learning want? *Journal of Cultural Studies*, *18*(4–5), 429–445.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. John Murray Publishers.

McFarland, D. A., & McFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society*, *2*(2).

Mentis, H. M., Chellali, A., & Schwaitzberg, S. (2014). Learning to see the body: supporting instructional practices in laparoscopic surgical procedures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, 2113–2122.

Mentis, H. M., & Taylor, A. S. (2013). Imaging the body: embodied vision in minimally invasive surgery. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, 1479–1488.

Metcalf, J., Moss, E., & Boyd, D. (2019). Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, *86*(2), 449–476.

Miceli, M., Schuessler, M., & Yang, T. (2020). Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW2). https://doi.org/10.1145/3415186

Miles, C. (2019). The combine will tell the truth: On precision agriculture and algorithmic rationality. *Big Data & Society*, *6*(1).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21.

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, *22*(2), 303–341.

Muller, M., Feinberg, M., George, T., Jackson, S. J., John, B. E., Kery, M. B., & Passi, S. (2019). Human-Centered Study of Data Science Work Practices. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, 1–8.

Muller, M., Guha, S., Baumer, E. P. S., Mimno, D., & Shami, N. S. (2016). Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. *Proceedings of the 19th International Conference on Supporting Group Work*, 3–8.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 126:1--126:15.

Naik, G., & Bhide, S. S. (2014). Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics*, *3*(3), 50–53.

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data*, *5*(2), 85–97.

Neyland, D. (2016a). Bearing accountable witness to the ethical algorithmic system. *Science, Technology, & Human Values*, *41*(1), 50–76.

Neyland, D. (2016b). Bearing Account-able Witness to the Ethical Algorithmic System. *Science Technology and*

Human Values, *41*(1), 50–76. https://doi.org/10.1177/0162243915598056

O'Connell, J. (1993). Metrology: The Creation of Universality by the Circulation of Particulars. *Social Studies of Science*, *23*(1), 129–173.

Orlikowski, W. J. (1992). The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Studies*, *3*(3), 398–427.

Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. *Organization Studies*, *28*(9), 1435–1448.

Orlikowski, W. J., & Gash, D. C. (1994). Technological frames: making sense of information technology in organizations. *ACM Transactions on Information Systems*, *12*(2), 174–207. https://doi.org/10.1145/196734.196745

Paine, D., & Lee, C. P. (2017). "Who Has Plots?": Contextualizing Scientific Software, Practice, and Visualizations. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW).

Paine, D., & Ramakrishnan, L. (2019). Surfacing Data Change in Scientific Work. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, & B. Nardi (Eds.), *Information in Contemporary Society* (pp. 15–26). Springer International Publishing.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press.

Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 39–48.

Passi, S., & Jackson, S. J. (2018). Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28.

Passi, S., & Jackson, S. J. (2017). Data Vision: Learning to See Through Algorithmic Abstraction. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, 2436–2447.

Passi, S., & Sengers, P. (2020). Making Data Science Systems Work. *Big Data & Society*, *7*(2), 1–13.

Passi, S., & Sengers, P. (2016). "From what I see, this makes sense:" Seeing meaning in algorithmic results. *Computer-Supported Cooperative Work (CSCW) 2016 Workshop "Algorithms at Work: Empirical Diversity, Analytic Vocabularies, Design Implications,"* 1–4.

Pentland, A. (2009). Reality Mining of Mobile Communications: Towards a New Deal on Data. In S. Dutta & I. Mia (Eds.), *The Global Information Technology Information Report 2008-2009: Mobility in a Networked World* (pp. 75–80). Palgrave Macmillan.

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441.

Pinch, T. J., Collins, H. M., & Carbone, L. (1996). Inside Knowledge: second order measures of skill. *The Sociological Review*, *44*(2), 163–186.

Pine, K. H., & Liboiron, M. (2015). The Politics of Measurement and Action. *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (CHI '15), 3147–3156.

Pink, S., Sumartojo, S., Lupton, D., & Bond, C. H. La. (2017). Mundane data: The routines, contingencies and accomplishments of digital living. *Big Data & Society*, *4*(1), 2053951717700924. https://doi.org/10.1177/2053951717700924

Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Scienceand Public Life*. Princeton University Press.

Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford University Press.

Powers, D. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol*, *2*(1), 37–63.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Puschmann, C., & Burgess, J. (2014). Metaphors of Big Data. *International Journal of Communication*, *8*, 1690–1709.

Quirk, M. (2005, November). The Best Class Money Can Buy. *The Atlantic Monthly*.

Raley, R. (2013). Dataveillance and Counterveillance. In L. Gitelman (Ed.), *"Raw Data" Is an Oxymoron* (pp. 121–145). MIT Press.

Ren, D., Amershi, S., Lee, B., Suh, J., & Williams, J. D. (2017). Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 61–70. https://doi.org/10.1109/TVCG.2016.2598828

Reynaud, B. (2005). The void at the heart of rules: routines in the context of rule-following. The case of the Paris Metro Workshop. *Industrial and Corporate Change*, *14*(5), 847–871.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-Agnostic Interpretability of Machine Learning. *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 91–95.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Riberi, V., González, E., & Rojas Lasch, C. (2020). An Ethnography of Vulnerability: A New Materialist Approach to the Apparatus of Measurement, the Algorithm. *Anthropology & Education Quarterly*, 1–24. https://doi.org/10.1111/aeq.12359

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 1–6.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33–36. https://doi.org/10.1002/hbe2.117

Rooksby, J., Rouncefield, M., & Sommerville, I. (2009). Testing in the Wild: The Social and Organisational Dimensions of Real World Practice. *Computer Supported Cooperative Work (CSCW)*, *18*(5), 559.

Rotella, P. (2012, April). Is Data The New Oil? *Forbes*.

Rubel, A., & Jones, K. M. L. (2014). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, *32*(2), 143–159.

Rudin, C. (2018). *Algorithms and Justice: Scrapping the 'Black Box.'* The Crime Report: The Criminal Justice Network. https://thecrimereport.org/2018/01/26/algorithms-and-justice-scrapping-the-black-box/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. *2017 IEEE International Conference on Big Data (Big Data)*, 2355–2361.

Saltz, J. S., & Shamshurin, I. (2015). Exploring the process of doing data science via an ethnographic study of a media advertising company. *2015 IEEE International Conference on Big Data (Big Data)*, 2098–2105.

Schmidt, K. (2011). The Concept of `Work' in CSCW. *Comput. Supported Coop. Work*, *20*(4–5), 341–401. https://doi.org/10.1007/s10606-011-9146-y

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think In Action*. Basic Books.

Schuetz, A. (1943). The Problem of Rationality in the Social World. *Economica*, *10*(38), 130–149.

Schuurman, N., & Balka, E. (2009). alt.metadata.health: Ontological Context for Data Use and Integration. *Computer Supported Cooperative Work (CSCW)*, *18*(1), 83–108.

Seaver, N. (2012). Algorithmic Recommendations and Synaptic Functions. *Limn*, *2*. https://limn.it/articles/algorithmic-recommendations-and-synaptic-functions/

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 2053951717738104. https://doi.org/10.1177/2053951717738104

Seaver, N. (2019). Knowing Algorithms. In J. Vertesi & D. Ribes (Eds.), *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 412–422). Princeton University Press.

Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, *Forthcomin*.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Sengers, P. (1999). Designing Comprehensible Agents. *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, 1227–1232. http://dl.acm.org/citation.cfm?id=1624312.1624393

Sengers, P. (1998). Do the thing right: an architecture for action-expression. *AGENTS '98: Proceedings of the Second International Conference on Autonomous Agents*, 24–31. https://doi.org/10.1145/280765.280770

Sengers, P., Boehner, K., David, S., & Kaye, J. "Jofish." (2005). Reflective Design. *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, 49–58. https://doi.org/10.1145/1094562.1094569

Shapin, S. (1989). The Invisible Technician. *American Scientist*, *77*(6), 554–563.

Shapin, S. (1994). *A Social History of Truth: Civility and Science in Seventeenth Century England*. University of Chicago Press.

Shapin, S. (1995a). Cordelia's Love: Credibility and the Social Studies of Science. *Perspectives on Science*, *3*(3), 255–275.

Shapin, S. (1995b). Trust, Honesty, and the Authority of Science. In R. E. Bulger, E. M. Bobby, & H. Fineberg (Eds.), *Society's Choices: Social and Ethical Decision Making in Biomedicine* (pp. 388–408). National Academies Press.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life*. Princeton University Press.

Shaw, R. (2015). Big Data and reality. *Big Data & Society*, *2*(2).

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data*. John Wiley and Sons.

Singh, S., Ribeiro, M. T., & Guestrin, C. (2016). Programs as Black-Box Explanations. *ArXiv E-Prints*, *1611.07579*.

Star, S. L., & Ruhleder, K. (1994). Steps Towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-scale Collaborative Systems. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 253–264.

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*, 9–30.

Stark, D. (2009). *The Sense of Dissonance: Accounts of Worth in Economic Life*. Princeton University Press.

Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation Work: Cultivating Vision in Collective Practice. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 443–453.

Strauss, A. (1988). The Articulation of Project Work: An Organizational Process. *The Sociological Quarterly*, *29*(2), 163–178. https://doi.org/https://doi.org/10.1111/j.1533-8525.1988.tb01249.x

Strauss, A., & Corbin, J. M. (1990). *Basics of Qualitative Research: Grounded Theory Techniques and Procedures*. Sage.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Suchman, L. (2000). Organizing Alignment: A Case of Bridge-Building. *Organization*, *7*(2), 311–327.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

Suchman, L. A. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge University Press.

Suchman, L. A., Blomberg, J., Orr, J. E., & Trigg, R. (1999). Reconstructing Technologies as Social Practice. *American Behavioral Scientist*, *43*(3), 392–408.

Suchman, L. A., & Trigg, R. H. (1993). Artificial Intelligence as Craftwork. In S. Chaiklin & J. Lave (Eds.), *Understanding Practice: Perspectives on Activity and Context* (pp. 144–178). Cambridge University Press.

Suchman, L., Trigg, R., & Blomberg, J. (2002). Working artefacts: ethnomethods of the prototype. *The British Journal of Sociology*, *53*(2), 163–179.

Suh, J., Ghorashi, S., Ramos, G., Chen, N.-C., Drucker, S., Verwey, J., & Simard, P. (2019). AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.*, *10*(1). https://doi.org/10.1145/3241379

Suwajanakorn, S. (2018). *Fake videos of real people - and how to spot them*. Ted Talk. https://www.ted.com/talks/supasorn_suwajanakorn_fake_videos_of_real_people_and_how_to_spot_them#t-379720

Symons, J., & Alvarado, R. (2016). Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society*, *3*(2), 1–17.

Tabarrok, A. (2015). *The Rise of Opaque Intelligence*. Marginal Revolution. http://marginalrevolution.com/marginalrevolution/2015/02/opaque-intelligence.html

Taylor, A. S., Lindley, S., Regan, T., & Sweeney, D. (2014). Data and life on the street. *Big Data & Society*, *1*(2), 1–7.

Taylor, A. S., Lindley, S., Regan, T., Sweeney, D., Vlachokyriakos, V., Grainger, L., & Lingel, J. (2015). Data-in-place: Thinking through the Relations Between Data and Community. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*, 2863–2872.

Thoreau, F. (2016). 'A mechanistic interpretation, if possible': How does predictive modelling causality affect the regulation of chemicals? *Big Data & Society*, *3*(2), 1–11.

Vaughan, J. W., & Wallach, H. M. (2020). A Human-Centered Agenda for Intelligible Machine Learning. In M. Pelillo & T. Scantamburlo (Eds.), *Machines We Trust: Getting Along with Artificial Intelligence*. http://www.jennwv.com/papers/intel-chapter.pdf

Veale, M. (2017). Logics and Practices of Transparency and Opacity in Real-world Applications of Public Sector Machine Learning. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017), Halifax, Canada.*

Veale, M., Kleek, M. Van, & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of ACM CHI 2018*.

Vertesi, J. (2019). From Affordances to Accomplishments PowerPoint and Excel at NAS. In J. Vertesi & D. Ribes (Eds.), *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 369–392). Princeton University Press.

Vertesi, J., & Dourish, P. (2011). The value of data: considering the context of production in data economies. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '11)*, 533–542.

Wang, M. (2013, September 11). Public Universities Ramp Up Aid for the Wealthy, Leaving the Poor Behind. *ProPublica*. https://www.propublica.org/article/how-state-schools-ramp-up-aid-for-the-wealthy-leaving-the-poor-behind

Wang, M. (2014, February 25). How Exactly Do Colleges Allocate Their Financial Aid? They Won't Say. *ProPublica*. http://www.propublica.org/article/how-exactly-do-colleges-allocate-their-financial-aid-they-

wont-say

Wattenberg, M., Viégas, F., & Hardt, M. (2016). Attacking discrimination with smarter machine learning. *Google Research*. https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Weick, K. E. (1995). *Sensemaking in Organizations*. SAGE Publications.

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150.

Wittgenstein, L. (1958). *Philosophical Investigations* (G. E. M. Anscombe (Trans.); 6th ed.). Blackwell.

Wolf, C. T. (2019). Conceptualizing Care in the Everyday Work Practices of Machine Learning Developers. *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, 331–335.

Woolgar, S. (1991). Configuring the User: The Case of Usability Trials. In J. Law (Ed.), *A Sociology of Monsters: Essays on Power, Technology and Domination* (pp. 58–100). Routledge.

Wylie, C. D. (2014). 'The artist's piece is already in the stone': Constructing creativity in paleontology laboratories. *Social Studies of Science*, *45*(1), 31–55.

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). Investigating How Experienced UX Designers Effectively Work with Machine Learning. *Proceedings of the 2018 Designing Interactive Systems Conference*, 585–596.

Yousif, M. (2015). The rise of data capital. *IEEE Cloud Computing*, *2*(2), 4.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science Technology and Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zhang, Z., Singh, J., Gadiraju, U., & Anand, A. (2019). Dissonance Between Human and Machine Understanding. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW). https://doi.org/10.1145/3359158

Ziewitz, M. (2016). Governing Algorithms: Myth, Mess, and Methods. *Science Technology and Human Values*, *41*(1), 3–16. https://doi.org/10.1177/0162243915608948

Ziewitz, M. (2017). A not quite random walk: Experimenting with the ethnomethods of the algorithm. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717738105

Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology*, *13*(3).