1

Data Science to Human-Centered Data Science

On October 9, 2017, Frank Lantz released a simple game based on the following premise: What if we created an AI (artificial intelligence) with an apparently innocent goal: make as many paperclips as possible, as efficiently as possible?

It sounds nonthreatening, even boring. But the clicker game, Universal Paperclips, promptly went viral. And as this story illustrates, the unintended consequences of a relatively simple algorithm can lead to the destruction of the universe. The problem is that an algorithm will execute exactly what its designer told it to do. If the designer forgot to program in bounds or stopping conditions (or common sense, ethics, or human values), the program will continue beyond the designer's original intent for an unbounded amount of time.

In this example, the lack of a stopping condition meant turning all the matter in the universe into 3×10^{55} paperclips.

At a moment of great optimism and enthusiasm for data science, coupled with a rising awareness of systemic societal injustices, this story resonates. Fields as diverse as computer science, artificial intelligence, social science, and even science fiction have been wrestling for decades with similar types of questions about ethics and boundaries and designers' responsibilities. Now we have access to much more data than humans can reliably make sense of, and these kinds of doomsday scenarios, along with algorithmic biases on smaller scales, have become greater risks.

It is common to hear people saying that we can guard against bias or racism in, say, mortgage lending with software that makes a less "biased" evaluation of applicants. That way, you remove "human error" and human prejudice from the equation, right? Those biased bankers or racist lenders will no longer be able to discriminate against people because of the color of their skin or what they were wearing on the day they walked into the office to apply.

In reality, of course, as numerous examples show, algorithms reflect the choices made by their human developers, including conscious and unconscious biases. What's worse, algorithms may amplify these biases, make them less transparent to other people, or make it harder to mitigate them. This is how we get Google Images returning pictures of white men when a person types in "doctor" or highly sexualized results for the phrase "Black girls" (Noble 2018; Wible 2016; see also Bradley et al. 2015). The CEO of a facial recognition software company said in 2018 that the software shouldn't be used by law enforcement to detect criminals because of the inherent racial bias (Brackeen 2018). IBM's CEO went further and declared that IBM was no longer in the facial recognition software business (Allyn 2020; Denham 2020). In June 2020, Microsoft and Amazon followed suit, announcing that they were putting in place a moratorium on selling their facial recognition software to police departments.

Recognition of the problem of *algorithmic bias* is becoming more widespread. The biases of individual designers, which may include the unconscious beliefs of a society's majority population, are inevitably reflected in the design of algorithms. Rather than the "wisdom of the crowd," perhaps we should be talking about the "bias of the crowd."

Beyond reflecting designer bias, the decisions that data scientists encode in algorithms may have unintended individual and societal consequences. Data science has developed significant power over human lives. As data scientists, we must think carefully about that power, its potential consequences, and our responsibilities.

Part of the work in human-centered data science lies in understanding and making transparent the mental models that govern the design of algorithms running over very large datasets. Research in visual analytics and human-centered data science shows that one of the most important elements for maximizing the effectiveness of an algorithm that is designed for humans to use is transparency or understandability (Baldassarre 2016; Brooks et al. 2013; Ye 2013).

And yet, consider the example of deep learning, which is so effective in solving many big data problems but which also creates models so complex, with many variables, that even the system designer cannot always know how a particular output is generated. All we know is that it works. How can we provide an understanding to the user when we don't even know why a model does what it does?

This is a good question, and one that machine learning developers have been wrestling with for many years. In the early 1980s, one of our colleagues attended a machine learning conference and listened to a speaker, a software developer at a bank, discussing the application of a backpropagation network, an early precursor of today's deep learning algorithms. The speaker explained improvements that the machine learning algorithm had made in the bank's mortgage lending process.

The colleague raised his hand and asked, "What do you do when an applicant wants to know why they've been denied a loan?"

The speaker responded, "Because that's a regulatory issue, we are required by law to let the applicant know the reason for any denials. So, what we do is gradually change one of the inputs until we run into the decision boundary of the algorithm. For example, we might raise their income, or increase the number of years they've spent in their job. As soon as the output changes, we can inform the applicant, 'If your salary was X dollars higher, or if your credit rating was so-and-so many points better, you would have received the loan.'"

This story is telling on many levels. First, it illustrates the importance of legal and policy guidelines to govern data and to govern what people are allowed to do with the results of the analysis of data. Second, it demonstrates how legal guidelines may inadvertently encourage developers to reverse-engineer obscure code to satisfy human needs. Even if the user or designer of this algorithm didn't really understand why it produced the results it did, the type of exploration described by this speaker can help produce a useful mental model for the end user. Finally, it speaks to the need to design machine learning algorithms whose parameters are more transparent, or to figure out ways to help people understand what's going on inside the "black box" of the algorithm—which is the goal of the field that has become known as *interpretable machine learning* (Molnar 2020).

Emergence of Human-Centeredness in Data Science

Large-scale data analysis presents opportunities in a wide variety of social, scientific, and technological areas and has led to a variety of innovations (Gluesing, Riopelle, and Danowski 2014; Luczak-Roesch et al. 2015). Yet, increasingly focusing on purely statistical or computational approaches may fail to capture social nuances, affective relationships, or ethical, value-driven community commitments and other human-centered concerns. Human-centered data science emerged to address precisely these issues.

Human-centered data science is a new interdisciplinary field of study that draws from human-computer interaction, social science, statistics, and computational techniques. It draws on the well-established traditions of human-centered design to inform better data science practice. Human-centered data science pushes computational approaches to large-scale data to include the kind of rich detail, contextual knowledge, and deep understanding that qualitative research and mixed methods can bring to the understanding of data and society. The authors of this book have helped to originate this area through writing about these practices, hosting workshops and tutorials to develop them with other researchers, putting them into practice in our own projects, and teaching them in our classrooms. We argue that a deep understanding of the human and social contexts of data helps data scientists to respond to the ethical thinking and choices that they face in their everyday practice. Human-centered approaches also help data scientists develop empathy for the subjects of their data, which can lead to greater acceptance and use of data science systems. Human-centered data science seeks to develop awareness of the complex nature of the interaction between society, technology, and human-generated data. Collectively, we have spent decades working in computational data science and in studying data science. We have come to the inescapable conclusion that without a human-centered approach, data science may not only cause irreparable harm to society but will also be ultimately unsuccessful in its lofty goals. These goals include addressing many widely accepted scientific and societal grand challenges, such as developing new medicines and vaccines trusted by the public, ameliorating hunger and poverty, and creating defenses against climate change.

Human-centered data science is necessarily broad. Our use of the term does not imply that every project should use every element that we define as human-centered. For example, a human-centered approach to crowdsourced scientific data might focus primarily on the dynamics of how people produce the data.

However, human impacts are more pervasive within data science than commonly believed. Often, large datasets for machine learning training are not thought of as being "human-centered." Our approach to human-centered data science, however, suggests making visible the human work that happens, for example, on platforms like Amazon Mechanical Turk so that we can label these training sets. The tools and techniques that we teach in this book help people learning data science to raise these questions, reflect on their work, do their work responsibly, carefully, and ethically, and above all keep the people affected by their projects at the forefront of the planning and execution of their projects.

Our approach goes beyond the "usability" engineering of data science models and pipelines. Our approach is pragmatic and directed at people who do data science projects. We take on board the concerns of social science and humanistic critiques, such as critical data studies. We aim, though, to translate them to an audience of soon-to-be practitioners. In this sense, what you won't find in this book are debates about the theories behind why we do data science the way we do. We are less interested in what we see as a false debate in data science between human-centered approaches and rigor. We show how you as a data science practitioner can build rigorous *and* ethical algorithms, design projects that use cutting-edge computational tools *and* address social concerns, and present results that could speak to statisticians *and* journalists. For us, the human-centered approaches that we present in this book, taken together, create the possibility for people working in data science to have the capacity to make positive changes in the world.

With this book we want to teach you how to reflect on the enormous responsibility that comes with the power of data science. This is because human-centered data science is about the people we study, the people who are involved at practically every stage of the data science cycle, and the people who base decisions on our projects. We show throughout this textbook that there are points where you get to make choices, and this has implications for the power and exclusion of people from this practice. Human-centered data science confounds the notion that automated or computational approaches can be "free" of bias and shows how the power to define categories and ask questions has huge implications for the work that we do.

Doing Human-Centered Data Science

There are multiple ways to "do" human-centered data science. Software developers and data scientists, such as some of the authors of this book, may apply human-centered techniques such as qualitative research to develop better algorithms for data science projects.

A human-centered approach to data science can produce more effective algorithms. One of us (Cecilia) worked as a data scientist and software developer with a group of astrophysicists who had, in their words, "too much" data. They were suddenly facing an onslaught of more than 100 times the amount of data they were used to. The problem was how to tell the good data from the bad. They were analyzing pictures of supernovae, and wanted to automate one of the steps in classifying the images. The team created a requirements document that suggested complete algorithm automation of the process. However, multiple attempts to accomplish this failed. It turned out that the requirements document did not capture the team's underlying needs. By spending time with the team, following human-centered methods that involved careful listening and understanding what they were trying to accomplish rather than merely what they thought they needed, Cecilia proposed new algorithms that were not mentioned in the requirements document but could help the team achieve their goals. For example, solely based on close observation of the team's process, she developed a Fourier contour descriptor algorithm to parse the specific curvature of supernovae from other images in the dataset. This approach turned out to be more efficient and ended up producing 40 percent fewer false positives, significantly improving the team's supernova search process.

Case Study 1.1

Different Ways of Seeing in Data Science

Steven Jackson, Cornell University

I work in *critical* and *interpretive* traditions that study order, value, and meaning as defining attributes of human activity in the world. To do this work, I mostly use ethnographic methods, but I inform them with readings from sociology, anthropology, law, policy, design, and science and technology studies (STS). Now that data science has become so important, I use these tools to study the work of people in data science. In this case study, I am particularly interested in the ethics of data science work.

The "Fairness, Accountability, and Transparency" (FAccT) field has made many important strides in opening the field of data science and algorithmic technique to being studied in terms of ethical assumptions, values, and practices. In addition to these formal studies, there are other virtues and practices essential to the real-world work of data science that are no less important and, taken collectively, constitute the *everyday practical ethics* of the field.

Like all forms of knowledge, data science (whether human-centered or otherwise) provides a *way of seeing*—that is, imagining or picturing the world that inevitably focuses attention on some things and ignores others. Sociologists might describe this as a kind of "standpoint epistemology"—from epistemology (how you know things) and from standpoint (you "see from where you stand"). We learn our standpoints over time: to be a data scientist is to learn to "see" the world in particular ways through the numbers and algorithms of data science (Passi and Jackson 2017). This is a powerful and, in its place, positive development.

Someone who acquires this knowledge can lose sight of the fact that it is one among many ways of seeing. At its worst, this tendency can make data science solutions appear to be inevitable and "objective," and can make data science practitioners seem to have exaggerated authority. A part of this effect may be traced to the stories we (and the world) tell about the nature of work in our field and the cleaned-up stories of practice that ignore the mundane realities of data science work. Despite elevated claims about data science, much of the work is in fact custodial or even janitorial in nature—the incessant effort to gather, clean, and repair data and to make datasets work well enough for the purpose at hand. Rather than a smooth or neutral mirror of reality, data is therefore best viewed as a messy and very human *accomplishment*—the end (or middle) point of a whole world of very ordinary human work.

A different kind of distortion in seeing confronts the attributions of certainty and authority that are sometimes pressed on data science from the outside world. This is a tension that is long familiar to sociologists of science, and this tension is part of the inevitable loss of information during translation from one field to another. A sociologist might call this problem the "reification" of knowledge claims as they move from the researchers who initially generated them (for whom limitations, doubts, and uncertainties are hard to ignore), to more distant users and audiences. For this latter group (especially those with little working knowledge of the real-world practices of data science), the claim or finding may begin to take on a certainty and solidity that will begin to appear magical. In the words of sociologist Harry Collins (1985), “distance lends enchantment”—giving the output of data science work an authority that it may or may not deserve. This may also be seductive and therefore dangerous to data scientists themselves. Who doesn’t like to be believed? The sense of authority may also work against the sense of fallibility or “this-might-be-wrongness” that any human-centered data science must carry in its fundamental set of assumptions and work practices.

Finally, data scientists may struggle to recognize the essentially collaborative nature of their work—including with actors who may be closer to the everyday work of the domains or problem areas they seek to address and therefore have powerful standpoints of their own to contribute. Making these different perspectives play well together, whether in academic research or commercial firms, is essential to adding to our knowledge through the application of data science and is often negotiated in practice by the careful development and management of trust (Passi and Jackson 2018). Misplaced or unearned authority, or a data science that is uninterested in the knowledge and practices of real-world actors (“just give me the data”), is the enemy of this process. So are instances in which data science is brought in (for example, by management) to overrule local knowledge. One example is the essential knowledge and experience of experts in the domain. To paraphrase a point sometimes attributed to Winston Churchill, what is needed is a data science “on tap, not on top.”

Ordinary, humble, fallible, and collaborative: this would be a human-centered data science worthy of the name.

A completely different approach to human-centered data science has emerged from the social science fields, such as science and technology studies, involving studying the people who produce data science. What are the sociological factors at play in data science teams? What constitutes a successful data science collaboration? Reflecting on these types of questions can lead to changes in the human process of data science work and ultimately to more fruitful collaborations and better results.

Another way to do human-centered data science is to combine approaches. Small-scale qualitative approaches to data collection and analysis offer researchers the opportunity to obtain very rich, deep insights about specific phenomena—often in a very bounded or limited context (Zheng et al. 2015). Such studies often face challenges related to generalization, extension, verification, and validation. They also face problems of scale. It is possible to interview 100 people but very hard to interview one million. On the other hand, large-scale quantitative approaches to data collection and analysis give the opportunity to look at broad datasets, but the insights gleaned are often much shallower, lacking the rich detail associated with deep study (Green, Arias-Hernandez, and Fisher 2014). “Big data needs thick data,” as one anthropologist termed it (Wang 2016). There are now many people, including the authors of this book, who advocate for combining the power of data science tools to understand humans at scale with ways of understanding human behavior at depth through qualitative approaches that can provide powerful insights (Muller et al. 2016; Baumer et al. 2017).

Human-centered data science draws on work from both qualitative and quantitative traditions, involving practitioners with training in computer science, statistics, or social science. Examples include work that has integrated quantitative research methods into qualitative research workflows (Brooks et al. 2013; Goel and Helms 2014; Xing et al. 2015). Online, digital, or virtual ethnography has gained widespread adoption as qualitative researchers adapt traditional ethnographic methods to online spaces (Daniels, Gregory, and Cottom 2016; Markham and Baym 2009; Murthy 2011). Computational social scientists—that is, researchers primarily in social science areas who develop and use computational methodologies to ask and answer social science questions—have found significant recent success in developing computational methodologies for large-scale social data that account for a degree of contextual reasoning within analysis.

Human-centered data science also encompasses the process of creating data science tools that integrate seamlessly into the sociotechnical ecosystem of the domain they are designed for. Such tools have often demonstrated the greatest success. One well-known example is iPython (later Jupyter), first developed by Fernando Pérez in a human-centered fashion specifically to ease scientists’ workload (Pérez and Granger 2007). Human-centered design is particularly effective in the development of software for analyzing large datasets (Aragon and Poon 2007; Aragon, Poon, and Silva 2009; Faiola and Newlon 2011; Poon et al. 2008).

Among the many unanswered questions surrounding human-centered data science are issues of sampling, selection, and privacy. What are the ethical questions raised by processing vast datasets? How should we treat the workers who do necessary tasks on crowdwork platforms? Who owns personal medical data—the company whose machines and software collect it, the medical practitioner who interprets it, or the patient who generates it? Can design or other skills often considered to be the province of an individual human be effectively crowdsourced (Bean and Rosner 2014; Lasecki et al. 2015)? What policies do we need to develop to protect human rights in this new age of “big data” (Gray and Suri 2019)? Questions such as these are legion, and we are only beginning to explore the territory of potential answers.

About This Book: Themes

First, we would like to draw your attention to five recurring themes that are developed throughout the book. We ask you to reflect on each of these themes and consider how they are used as you read.

Human-Centered Data Science as Ethical Responsibility: “The Data Made Me Do It”

One of the difficulties in dealing with the “data deluge” is a facile assumption that the data can tell us everything—that it is unbiased, neutral, and somehow possessing wisdom far beyond the human. “If it’s big enough, it contains everything.”

Our approach puts human responsibility at the center of data science. People are involved at every stage of the cycle of collecting, cleaning, analyzing, and communicating data science results. Each stage presents a series of choices, and these choices matter for the responsible and ethical use of data.

Human-Centered Data Science as Looking in the Right Places: The Streetlight Effect

The streetlight effect is a kind of observational bias where people only search for something where it is easy to look. It refers to a joke that apparently dates to the 1920s. A police officer sees a person on hands and knees searching the ground around a streetlight at midnight and asks what they’re doing.

“I’m looking for my keys.”

The officer helps for a few minutes, doesn’t find anything, and eventually asks the person if they’re sure the keys were lost near the streetlight.

“No, I lost them across the street somewhere.”

“Then why look here?” asks the irritated officer.

“The light is much better here.”

The streetlight effect explains, perhaps, why many researchers (including, we have to admit, some of us) have turned to Twitter to study social phenomena. People take advantage of datasets that are easy to access or easy to convert into simple data structures for analysis. Research shows that Twitter data has serious limitations as a representation of public opinion. But because it is public, easily available, and vast in quantity, hundreds of research papers have been published using Twitter data. Certainly, data science can do better than look under the streetlight. A human-centered approach to data science asks where can we look first, before looking under the light of the easily available dataset.

Human-Centered Data Science as Collective Practice: We Are All Problem Seekers

Our approach to data science holds that many people can be empowered with data skills. You are reading this book and that is a start. We work with a range of communities, including self-trackers, child welfare agencies, community crisis response activists, astrophysicists, architects, journalists, nurses, pharmacists, and citizen or community scientists; clearly, data science in the human-centered approach is not a toolset reserved for the elite and the powerful. Our research shows us that working with people who have deep inside knowledge of the problems we are trying to solve helps improve our practice as data scientists. A human-centered approach figures out what needs to be known from the situation to create better models, more responsible data science pipelines, and more capacity for using data science tools—responsibly and ethically—to benefit people. We are encouraged by the Community Data Science Workshops and Urban Data Science, which host free and open meetings to train others (Hill et al. 2017; Rokem et al. 2015). We are inspired by people in the Data Science for Social Good (DSSG) movement who come together to learn experimentation and data science techniques from one another.

Human-Centered Data Science as Communication: We Are Communicators and Storytellers

Data science tools and methods are complex and multifaceted. We think of them as analytic lenses through which we look at the world or craft our version of the world. We also think of them as tools for reflection—on the data, on the tools, and on our own evolving understanding of our own ways of thinking about data. We use data science to tell stories about data and people who are affected by our choices of methods to analyze that data.

Human-Centered Data Science as Action: Make a World Where We Want to Live

Some of the authors of this book are software developers and data scientists with practical experience in industry. We look at the consequences of our technology, and we want to build technologies that create a world that we want to live in. For example, we would not want to build facial recognition technology that leads to false positives and wrongful convictions. We would not want to live in a world where digital surveillance is an everyday presence. As technology workers, we want to work toward a better future. Within human-computer interaction, this approach is sometimes referred to as *prefigurative* design and action, emphasizing both design practices and design outcomes that correspond to the future that we collectively envision (Asad 2019; Strohmayer, Clamen, and Laing 2019; Williams and Boyd 2019).

About This Book: Stories, Audience, and Our Purpose

We now shift gears from describing human-centered data science to helping you make the best use of this book.

Stories and Case Studies

We believe in the power of stories. In each chapter we use stories about data science practice to illustrate the main themes. Throughout the book, we present short case studies to illustrate some of the ramifications of human-centered data science. These case studies bring multiple authors into this book to present real-world examples of how to use human-centered data science, critique data science, and work with multiple communities.

Because we want this book to help people have a real-world impact, at the end of every chapter, we provide a set of recommendations and things to consider while doing a data science project. We also list recommended readings that go into more depth on the topics covered in each chapter.

Who This Book Is For

This book is addressed to people doing data science, learning data science, or managing data scientists. We imagine you, the reader, to be someone hoping to learn more about data science—either in a formal course or on your own, either as a student or a practitioner. We provide a brief overview and easily understandable explanation of many of the common statistical and algorithmic data science techniques to emphasize how a human-centered approach can enhance each one. You do not need any specialized knowledge in data science, computer science, or social science to learn from and benefit from this book, although we summarize and discuss many decades of research and experience from each of those fields. We don’t intend this book to teach you how to do the latest techniques in data science. However, we think, modestly, that you can’t do good data science without the practices that we cover here.

Why We Wrote This Book

Universities, businesses, and governments have rushed to train millions of people in the computational and statistical techniques necessary to process and extract insights from the vast amounts of structured and unstructured data. This computational turn toward so-called big data means the proliferation of more types of data generated and collected from a variety of sources. However, in the process, the social context and ethical considerations of data collection, analysis, use, and dissemination have often been overlooked. Many well-documented cases show how some approaches to data science can lead to severe ethical transgressions and significant harm, social bias, and inequality. And yet, from the purely computational perspective, many of these issues and complications may be hard to foresee, especially for aspiring data scientists who have no background in the ethics of data science from a human-centered perspective.

We have high hopes that this book can become a practical manual for data science practitioners who want to change the world. We do not say this lightly. We believe in the power of data science to help people discover new things, solve urgent challenges, create new services, and make things more efficient and better. We believe that people working in data science have a responsibility to the people who are affected by the results of their data, to the people whose data they use, to the people they work with and for, and to communities that may make use of their results. We believe that people want to—and can—work ethically and responsibly. Our approach to data science understands that people are often the source of data; that people do the work that it takes to create, label, source, and analyze data; that people are the audience for our results and rely on our ability to clearly communicate what we discover in our projects. We believe that data science can be done better if it is done in a way that is both technically rigorous and addresses concerns about bias, ethics, and inclusion. In that sense, we—a group of researchers and data scientists in universities and industry—are on a mission to help bring that urgency of human-centeredness to the next generation of leaders and practitioners in data science.

In this book, we explore how data-driven and qualitative research can be integrated to address complex questions in diverse areas, including but not limited to social computing; urban, health, or crisis informatics; and scientific, business, policy, technical, and other fields.

By training people in the human-centered data science methods described in this book, we hope to address concerns about the social impacts of large-scale data. This book can be used alongside a textbook for students of data science, both undergraduates and graduates. It can also be a useful handbook for professionals seeking to learn more about the responsible, effective, and human-centered use of data science to process the ever-growing quantity of human-generated data in the world today.

Book Outline

In chapter 2, we introduce the data science cycle, or the stages involved in a typical data science project. We address data collection, cleaning, feature engineering, analysis, representation, bias detection/mitigation, and distribution, with attention to the ethical and human-centered considerations that inform each step. We emphasize a cyclical, rather than linear, trajectory for data. We then describe several techniques commonly employed in data science and discuss the most popular tools used to carry out each technique. Our goal is not to teach these topics in-depth but to show how human-centered approaches might be useful for improving them.

Chapter 3 foregrounds the social aspects of human-centered data science work, exploring the relationships among various stakeholders in data science work: subjects, researchers, and their audiences. We show how people who work in data science usually have to intervene between “the data” and “the model” in data science pipelines. We discuss how data can represent both voluntary participants and unknowing subjects, and we address some of the ethical issues involved in data collection. We also examine populations that have historically been most vulnerable to exploitation by data science projects, and we think about how data scientists might interrogate both their methods and their data with regard to protection of privacy.

In chapter 4, we present a high-level overview of commonly used tools in the data science toolkit: machine learning, statistical analysis, automated tools for constructing data science solutions, visualization tools, and others, retaining a human-centered focus. These tools and techniques form the core of textbooks that our book aims to supplement. We spend no more than a few paragraphs on each of these techniques, with a focus on advice about when each might be most appropriate to a given research project. The concepts in chapter 4 may help anchor some readers who are new to data science in the tools and techniques of data science. Reviewing these concepts also gives us the opportunity to reflect on the human-centered challenges of these tools.

Chapter 5 focuses specifically on human-centered approaches to asking and answering data science problems. We start with the history of ethical research design to show how human-centered data science differs from other data science practice. Human-centered data science starts by formulating a meaningful question, considers issues of ethics and fairness, designs projects that others can easily build on, and incorporates reflexivity into the process. Many of the suggestions we make in chapter 5 urge you to think about your project's impact on people represented in your dataset, people whose lives may be affected by the results of your analysis, people who might want to reuse your pipeline or just view your results, people who may reuse your results, and even yourself at a later point in time.

In chapter 6, we extend the human-centered data science toolkit by examining how methods from other fields might be integrated and combined in innovative ways with data science. We draw on the rich tradition of more than a century of social science research into human behavior to explore recent research on new ways of merging computational tools with the methods that people have developed to interpret and transform our social world.

In chapter 7, we address the multiple types of collaborations that are needed for human-centered data science. Most data science projects are conducted in teams, and we consider how collaboration, especially with people from other disciplines, is an integral part of the process. We also discuss the different roles people take on in data science teams throughout the project lifetime. We examine various ways data scientists can collaborate with AI in their projects. While considering how collaboration works in organizations, we delve more deeply into the notion of a stakeholder, giving you tools for untangling complex webs of interaction and power in large organizations. Finally, this chapter emphasizes the importance and often complicated nature of working with the communities whose data you plan to use or whose problems you intend to solve with data science projects.

In chapter 8, we consider data storytelling and offer suggestions for how to translate human-centered data science principles into action. We draw on the rich literature of storytelling with special focus on why stories matter for data science practice. We then focus on how visualizations work as a great storytelling medium and provide recommendations for how data scientists can use visualizations (and storytelling, more broadly) to connect with various communities, including business leaders, experts, academics, researchers, and policymakers.

Chapter 9 addresses the future of human-centered data science as a discipline and discusses how researchers and practitioners can be advocates for human-centered data science practices among students, business leaders, policymakers, researchers in other domains, and the public. We review the five cross-cutting themes that we first introduced in chapter 1 and reflect on the examples and challenges of each. We conclude with a call to action posed by the poet Adrienne Rich (1986)—"With whom do you believe your lot is cast?"—and look ahead to how we all can shape the future of data science.

Who We Are

The authors are a diverse group of researchers with long-term experience in human-centered data science. In February 2016, four of us came together at a workshop titled "Developing a Research Agenda for Human-Centered Data Science" at the Computer-Supported Cooperative Work and Social Computing (CSCW 2016) conference. Although we had all been working in this area for a few years, this workshop served as a catalyst for us to develop focused research to build the field.

Cecilia Aragon, originally trained as a mathematician and computer scientist, has been conducting qualitative and quantitative research in human-centered data science for over a decade in academia as a professor at the University of Washington (UW), after fifteen years of hands-on experience as a data scientist and software developer in industry. She coined the term "human-centered data science" and organized the first workshop on the topic at the 2016 CSCW conference. She is Professor and Director of the Human-Centered Data Science Lab at UW and a strong advocate for the use of human-centered techniques throughout data science. As founding faculty director of the interdisciplinary data science master's program at the University of Washington, she developed the original curriculum for its course in human-centered data science.

Shion Guha has formal academic training in economics, statistics, and information science and is a professor of human-centered data science in the Faculty of Information at the University of Toronto. He uses computational and qualitative methods to examine how data-driven algorithms are designed, deployed, and evaluated in public services, particularly in the child welfare and criminal justice systems. He is building the undergraduate and graduate programs as well as curriculum in human-centered data science where questions of ethics, inequalities, and social justice take precedence in academic discussions.

Marina Kogan is a professor in the School of Computing at the University of Utah. Her research focuses on how people self-organize and problem-solve on social media during disasters. Her methodological focus is on developing methods that attempt to both harness the power of computational techniques and account for the highly contextual nature of the social activity in crisis. She extends and develops human-centered versions of network science models, natural language processing (NLP) techniques, and probabilistic models.

Michael Muller is enthusiastic about working with users (including data scientists and *their* users) for increased mutual understanding and collective action to make better outcomes for everyone. He has researched data science work practices at IBM Research AI. His research methods span qualitative and quantitative approaches, including the grounded theory analysis in his 2019 paper on data science workers and quantitative survey analysis in his 2020 paper on collaboration patterns in data science teams. Michael's background includes extensive and sometimes passionate work in participatory design, organizational social media, and allyship for social justice.

Gina Neff leads qualitative research teams on data science studies, looking at how data science is made in practice in industry settings. She directs the Minderoo Centre for Technology and Democracy at Cambridge. Her research focuses on work and collaboration, and she uses these insights to advise universities, startups, and nonprofit research organizations, including Data and Society and AI Now.

2
The Data Science Cycle

> The invitation was simple enough: meet for a coffee in the coolest café in a neighborhood known for its density of tech company headquarters. One of the authors (Gina) had met someone at a health innovation conference who had experience in luxury consumer goods and left to work on a digital watch that would serve to gather data about users' daily exercise and movements. The question he asked of her was simple and yet hard to answer, "What are we going to do with all this data?"
>
> —Neff and Nafus 2016

It is the way data science often works: start with the dataset, then figure out something to do with it. Social scientists start with the question, then figure out the data needed to answer it. From wireless sensors to mobile phone geolocation to social media, society is awash in "all this data."

We anticipate that people learning data science are asking the same question: What are we going to do with all this data? This question has motivated an explosion of data science opportunities and jobs. And it is one of the ways the data science cycle begins in many industries. Our goal in this chapter is to introduce this standard cycle, which is covered in-depth in other textbooks. We present it here to show readers the typical structure for the data science process and at the same time use it as a springboard to discuss the human-centered approaches that we focus on in this book. In chapter 3 we will examine many assumptions inherent in the traditional data science cycle. This chapter lays the foundation by walking you through the standard cycle as is currently practiced by many data scientists.

The data science cycle is often imagined as a series of sequential, interconnected steps. It is a cycle because there is an element of self-evaluation and feedback in every step that circles back to the initial stage of asking questions (see figure 2.1).