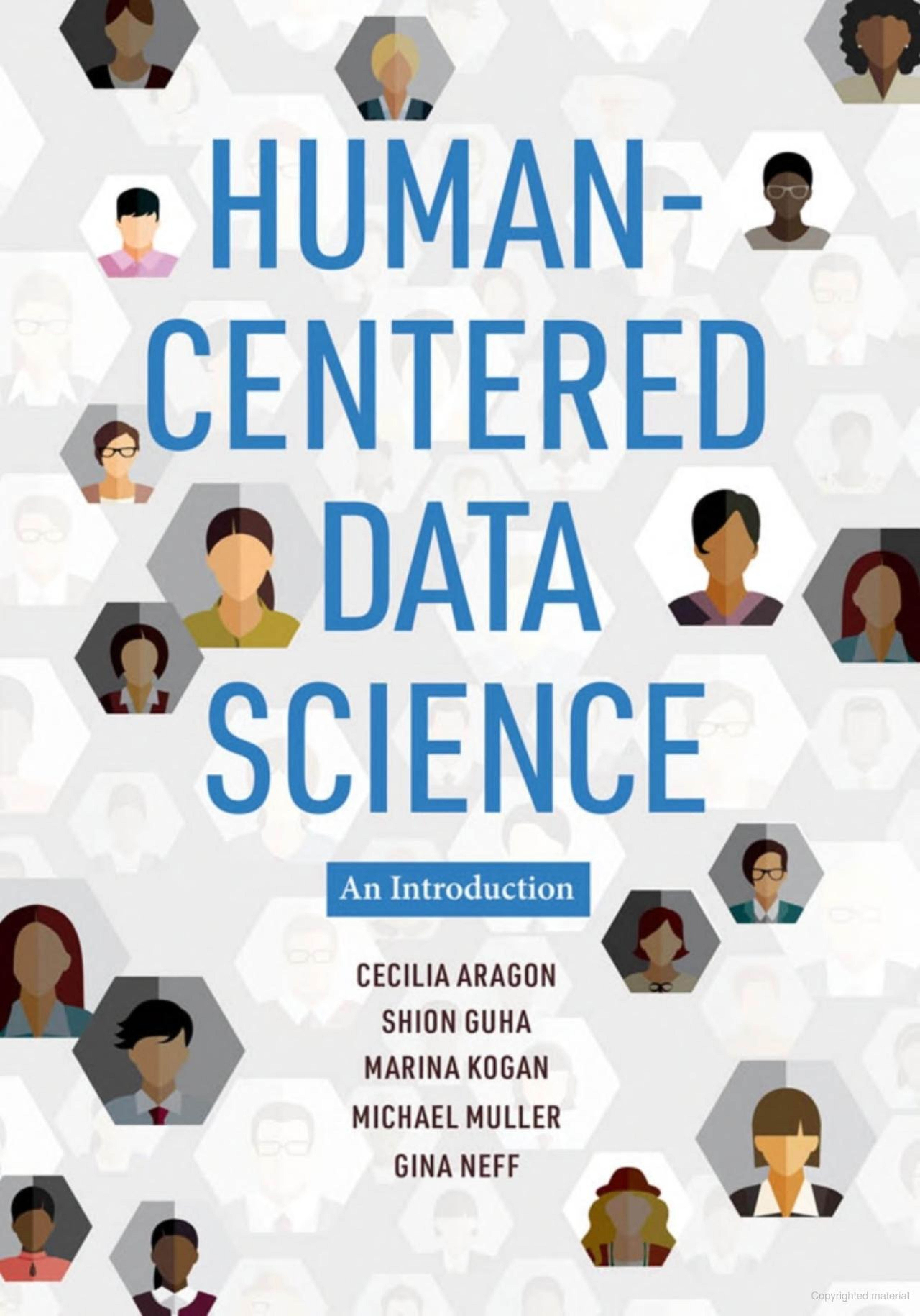


HUMAN-CENTERED DATA SCIENCE

An Introduction

CECILIA ARAGON
SHION GUHA
MARINA KOGAN
MICHAEL MULLER
GINA NEFF

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Times New Roman by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Aragon, Cecilia Rodriguez, author. | Guha, Shion, author. | Kogan, Marina, author. | Muller, Michael (Data scientist), author. | Neff, Gina, 1971– author.
Title: Human-centered data science : an introduction / Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021019747 | ISBN 9780262543217 (paperback)
Subjects: LCSH: Big data—Moral and ethical aspects. | Social indicators—Data processing. | Quantitative research—Moral and ethical aspects. | Information science.
Classification: LCC QA76.9.B45 A73 2022 | DDC 005.7—dc23
LC record available at https://lccn.loc.gov/2021019747

Contents

Acknowledgments

The authors wish to thank the many people who contributed to this book, beginning with our patient and highly responsive editor at the MIT Press, Gita Devi Manaktala.

Significant thanks are due to the attendees and co-organizers of the Computer-Supported Cooperative Work and Social Computing (CSCW) 2016 workshop on Human-Centered Data Science, where it first became clear we needed to document this new field. Next, we thank participants in the CHI 2019 workshop on Human-Centered Study of Data Science Work Practices; the GROUP 2020 workshop on Mapping out Human-Centered Data Science: Methods, Approaches, and Best Practices; and the CSCW 2020 workshop on Interrogating Data Science.

We particularly wish to thank Faith Bosworth from Book Sprints for keeping us on track with a gentle but relentless hand throughout our writing workshop in Seattle in February 2020. We also owe Jane Skau and Adeline Swires a great deal of gratitude for providing excellent logistical support during our writing workshop, keeping us well supplied and fed during a period of intense collaborative work.

We wish to thank the Gordon and Betty Moore and Alfred P. Sloan Foundations for their generous support of the eScience Institute at the University of Washington and for funding to help support this work. Support of a British Academy Mid-Career Fellowship is gratefully acknowledged.

We are grateful to Morgan Vigil-Hayes for her recommendations about Indigenous topics, to Theresa Jean Tanenbaum for consultations on gender identity, and to Andrea Simenstad for a careful reading of the final draft.

We thank the anonymous MIT Press reviewers whose detailed and thoughtful comments greatly improved the book and a brilliant team of editors and copyeditors who helped us get it in shape for publication.

Finally, we thank the authors of the case studies for sharing their excellent practical examples showcasing the wide variety in the field of human-centered data science.

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

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

(continued)

Case Study 1.1 (continued)

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

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).

A typical process starts with question or problem formulation, then goes through data collection, wrangling, cleaning, modeling, and finally representation, evaluation, and interpretation of results. Realistically, self-evaluation and feedback may exist at each step. We expand on each of these elements in the sections that follow, then conclude the chapter by distinguishing between models and pipelines and how they fit into a data science cycle.

Figure 2.1
The data science cycle.

Elements of the Data Science Cycle

One "standard" data science cycle lists nine elements (figure 2.1). The first step is often defined as *formulating a question*, although this step may not come first or the question may be revised multiple times. *Collecting data* can be accomplished in many ways. *Data wrangling and cleaning* refers to the unexpectedly lengthy and difficult process of finagling the data into a state that makes it easier to process, analyze, and visualize. *Feature engineering* involves selecting and extracting the attributes or features of the data that will be included in the model or algorithm, and sometimes making new combinations of features (e.g., ratios) or reassigning the labels of features ("classes") based on data outside the dataset. Next comes *data analysis and training* of the model, often on a subset of the full dataset. The first time through this step you will focus on data analysis for later training. Selecting labels or "ground truth values" for this training comes with its own set of decisions and potential pitfalls. Next, the iterative nature of the cycle becomes clearer as you *select and evaluate the model* and perhaps iterate on model selection. You will likely also return to the previous step and continue training the model. After that comes the decision of how to *represent the results of the data analysis* (i.e., communicate the results or provide for further exploration) to your audience—customers, decision makers, or anyone affected by the outcome of the analysis. *Distributing* the results of your analysis to others through publication is an important component of the cycle. Finally, *interpretation and communication* of your results to other people, through discussion or presentation, is a fundamental data science task. Throughout this cycle, the *iteration and feedback* process is critical. Iteration is not a setback but rather a process that deepens and strengthens the quality of your end results.

Formulating a Question

Most versions of the data science cycle put formulating the question or stating the problem as the first step in the process. We agree and suggest that data scientists consider asking a specific question they would like to answer through the data science process, as opposed to only starting with a specific dataset. Starting with a dataset may limit the types of questions we are able to ask and answer, and it curtails imagination for what other types of data might be appropriate and may be even better suited to answering a particular question. Starting with the data puts us in the position of not proactively looking for an answer to a question or a solution to a problem; instead, we are reacting to the specifics of a particular dataset. Depending on how the dataset was collected, how the variables in it were measured, and by whom, the results of your analysis may be strongly skewed or even distorted. In addition, people often take the path of least resistance, choosing to work with datasets that are easy to get or provide easily quantifiable data. But these may not in fact be the best datasets for answering a particular question—just the most convenient ones. We want our dataset to be in service of our question, not the other way around.

- How have you approached your data science projects? Did you "start with the data" and try to "find a good question" from the data? Or did you "start with the question" and try to find the right data to answer that question? What strengths and weaknesses have you encountered with each of these approaches? What other approaches have you experienced?

Another concern with starting with the dataset is that it is harder to know whether you are really measuring what you intend to measure. Some disciplines (e.g., psychology) call this a *measurement plan*—laying out the steps to quantify and assess the concepts pertaining to your question. For example, in health-related machine learning we may want to predict mental health indicators to be able to help people in real time. But how do we know if specific behaviors we are measuring are indeed good indicators or proxies for, say, depression? Other disciplines call this *internal validity*: How valid is our measurement for capturing some concept? Starting with formulating a question will also allow you to think deeply about how you measure different concepts within it, instead of choosing variables that might be poor proxies but are part of an easily accessible dataset.

Collecting Data

The next step is usually data collection. There are many ways of collecting data: downloading existing datasets that have been curated by others (individuals and organizations), collecting data from surveys, capturing activity on various software systems through their logs (such as credit card transactions or mobile phone records), scraping data from the web, using application programming interfaces (APIs) to download structured data directly from websites or apps, collecting scientific data from sensors, and many others.

If you choose to work with an existing dataset that has been curated by others, first you need to learn how the data was collected and what additional information you may need to be able to make use of the data: what variables are in the dataset (usually as columns in a table), what does each variable measure and what does the metadata mean, who measured these things and how, and how reliable and trustworthy are these sources of information (the people who compiled the data). You also might need to request the data from the people who collected it. In this case, you would need to negotiate access.

If you scrape the data from the web, you need to consider the terms of service of the site you are interested in and how appropriate it is to harvest the data. In that case, you will want to consider the potential harm of disrupting the site with repeated requests, the dangers of combining the resulting data with other publicly available datasets that might lead to deanonymization, and other possibilities (Fiesler, Lampe, and Bruckman 2016). For a deeper discussion of legality and ethics of web scraping, consider reading Krotov and Silva (2018). You should be aware that there are diverse views on ethics versus legality of working with website data (Bruckman et al. 2017), and that there are different expectations and legal frameworks in different regions of the world (Voigt and Von dem Bussche 2017).

As you might have noticed, each type of data collection entails questions of values and ethics: what data is needed and from where should it be collected, whose worldviews are reflected in particular datasets, how are various concepts being measured and by whom.

- Laws and expectations for data access vary by country and culture. What have you had to do to make sure that your data access was ethical and legal in your location and institution? Did you experience the rules as helpful, obstructive, or protective? If they were protective, whom were they protecting?

This relates to the problem of *justifying our data collection*. In data science, especially if we work in industry, we may have access to many different types of data that could potentially help us answer a question. Imagine a data scientist at Facebook aiming to answer a question about people's preferences for certain types of Facebook Groups (let's say around sports). It might be tempting for them to gather all the possible types of data that Facebook has on people who participate in these types of groups: data from groups themselves, their conversations and reactions within the groups, but also their conversations and reactions outside the groups, their private messages, and so on. There are a lot of data sources to consider. To justify the data collection, the data scientists (individually or within teams) should deeply interrogate what types of data they actually need for this analysis, as opposed to using everything that is available just because it is there. This interrogation—acting as your own adversary—will lead to more purposeful and intentional use of data.

In industry applications, when data scientists start with a particular question of interest to their clients, they may have to switch the dataset they rely on multiple times, based on the availability of the data, its granularity, and ease of access. This sometimes means that they repeatedly trade more precise and granular sources they had in mind for less accurate approximation of the behavior in the data that is more realistic to obtain. This means that the *measurement plan* they had at the beginning of their analysis also keeps changing. In this case, the data scientists need to be even more vigilant about to what degree they can answer their original question and with what level of confidence. These limitations need to be explicitly communicated along with the results of the analysis.

For example, in a study of crowding in a very large urban transit system, researchers first used surveys but needed to obtain data around the clock and in more transit stations categorical variables that only take on the value of zero or one, called *dummy variables*. There are also automatic ways of splitting multiclass variables into a weighted version of dummy variables; examples are effect coding in statistics and one-hot encoding in machine learning (Kugler, Dziak, and Trail 2018).

While features are intended to be a representation of things in the real world, some things are easier to represent as variables than others. Easily measured and quantified things are readily turned into variables or features. Human relationships, traditions, and local ways of making sense of the world may be harder to represent as quantifiable variables. And even when we work hard to find a way to measure these things and make them "legible," the resulting variables often reflect only a narrow view, a single aspect of these complex and multifaceted human endeavors (Scott 2020). For example, how do we represent how friendly and cordial a neighborhood is? It is not a well-structured, easily measurable aspect of a neighborhood, unlike the number of houses or streets. We may try to measure friendliness by the number of neighbors who say hello to each other on the street or welcome new neighbors with baked goods, or the number of children who have friends on their street. Clearly, all these ways of turning neighborhood friendliness into a quantifiable variable are only partial, one-sided representations of the complexity of neighborhood life, and we have to think carefully about which of these (or what combinations of them) are a better fit for our investigation.

Finally, we sometimes create features as a way of "controlling" or "normalizing" one variable with another. For example, if we want to include a predictive feature about how "characteristic" a certain word is in a document, then we might want to count the number of instances of that word (using natural language processing) and then divide by the number of words in the document to compute a percentage. Suppose our target word occurs ten times in a document. If the document is fifty words long, then the target word is probably *very* characteristic of that document. If the document is 10,000 words long, then the count of ten instances makes the target word less characteristic of the longer document. One of the tricky aspects of data science is deciding what is a "fair" way to normalize important numbers.

Since the new features are generated through various combinations and transformations of existing variables, feature engineering is an inherently human decision, even though there are often rules of thumb on what new combinations may be useful for different application domains (see chapter 3 for a discussion of feature engineering as the design of data). In addition, feature engineering generates and selects the features we consider using in the model. Here again, we need to think critically about these new features—whether they measure what we intend to measure, how they relate to our measurement plan, and whether they introduce any unintended bias. For example, using a geometric mean of three existing features might seem promising for some computational reasons, but what would it mean in terms of the concepts of interest, the question we are trying to answer? Would the simple arithmetic mean (average) be a better measure? Or the median?

- Have you had to construct new features from your data? Did you do this alone or with others? How did you decide which new features to engineer? How did you test them to see if they were useful and informative? Did you evaluate the possibility that the new features might have introduced bias?

Data Analysis and Training

In some data science techniques, a *training dataset* is used to help create and fine-tune the model. This requires *labeled* data: a subset of data points annotated with labels with categories or classes of interest. Labeling may also be called *annotating.* The labels or annotations are used as the *ground truth*—an accurate representation of the world from which the model is supposed to learn to generate labels when presented with new, unlabeled data (see case study 2.1 for the importance and implications of labeling the data). The choices made in how data science models are trained have implications for the results. How the training dataset is selected is an important choice. The size of the training data influences the results, but so does its representativeness.

We need to carefully consider who and what are included in the training set, as that will influence what the model finds in the data. Thus, types of social values and choices that are included in the training data will have implications for the model. A model trained to identify shoes but trained only on athletic shoes will miss other types of shoes. Similarly, a training dataset that relies on cases drawn from only one neighborhood may miss key factors from the rest of the city. The main consideration here is that training data should represent the phenomenon across all relevant diversity measures. This can become a difficult decision. If you are sampling across people, is race an important type of diversity for the phenomenon that you are working on? Is gender identity? What about age? You may need to look at other, similar projects to see which diversity variables they have used. You will also need to exercise your own good judgment and be attentive to your own biases, and you may want to consult with a diverse group of other people who have worked on these questions.

In addition, assigning labels can often be a difficult and costly process, and there are many decisions about strategies for data labeling (who labels it and how). These issues are discussed in more detail in chapter 3. Once we have a labeled set of training data, we train the model based on these labels. In practice, we often train multiple models to determine which type of model represents the data best, including a set of model parameters. One important concern is to avoid *overfitting*, or creating a model that fits the existing data so closely that it is not generalizable. How to do this is discussed in chapter 4.

Selecting and Evaluating the Model

There are different ways to select the best model for the job. In data science, we often use measures of accuracy. When making a prediction, the simplest measure we often use is called *classification accuracy.* It measures the proportion of the correct predictions out of all the predictions we made. It works well if there are equal numbers of cases in each class but can be misleading if classes are mismatched in size. For example, if 95 percent of our data had green labels and 5 percent had red, the model simply predicting all the labels to be green would be 95 percent accurate. But of course, such a model is not very useful, as it would not perform well on other datasets.

Overall accuracy is also not very useful, as we want to ensure that the model does well for every class. For example, if the model predicts men's behavior with very high accuracy but predicts the behavior of women somewhat poorly, the overall accuracy would still be rather high. Instead, we are looking for a model that has high accuracy for each class.

Case Study 2.1

Combining Knowledge in Data Science about Mental Health

Stevie Chancellor, Northwestern University

I build machine learning and data science tools to identify high-risk mental health behaviors in online communities. By high-risk, I mean behaviors like self-injury, suicidal ideation, and disordered eating and exercise behaviors. My aim is to build models on millions of public posts that can support better decisions about this data—like when to make an intervention. Online communities are an important source of support for those with mental illness, and I hope to support these people and those who care for them.

An important piece of my human-centered data science work is *construct validity*: that is, getting high-quality, accurate labels for my data. For me, a label might be a diagnosis of depression or quantified risk evaluation. Because I use labeled data to build reliable and precise models, it is essential that the labels are as close as possible to the corresponding clinical concept. For example, when we study depression, we should look at diagnostic criteria and clinical definitions. This is critical to the success of my work: if my research team's data labels are not valid, we could misclassify someone as suffering from a disorder, or we risk missing someone who needs help.

To support strong construct validity, I work with subject experts who are familiar with the clinical concepts my labels are trying to replicate. That often means working with doctors and medical researchers—but sometimes it also means moderators or community members themselves. I talk to many different "experts" with different experiences to help triangulate these concepts—triangulation is a research practice that uses information from multiple sources to converge on a conclusion. This makes my research very collaborative and fun and helps ensure the labeled data is a valid representation of mental illness.

In one project, we explored high-risk mental illness behaviors on Instagram. Severe mental illness is defined as major cognitive, judgment, and behavior impairments that make it hard to take part in day-to-day life. In our study, we focused on suicidal ideation, extreme weight control behaviors (like binging and purging, overexercising, and extreme food restrictions), and self-injury.

But we are computer scientists, not doctors—how are we going to understand this phenomenon, stay true to its clinical definitions, and ensure construct validity? To accomplish this, we worked with two medical researchers to identify signs of mental illness severity in eating disorder communities. Stephanie Zerwas and Erica Goodman are two clinical researchers who actively see patients as well as conduct research on eating disorders online. They were true research partners on this project and led on key aspects of our approach. For instance, they evolved our labeling guidelines from simply recording whether the post in the sample was "severe" to a more nuanced three-tiered approach that reflected their clinical reasoning. We then developed an annotation system using both clinical judgments and computational linguistics that annotated 26 million posts on Instagram for mental illness severity.

Working with domain experts like Stephanie and Erica is a crucial part of my research for a few reasons. First, these researchers make my models more robust because the experts think through examples I would never know as a computer science person. Accurately labeled rare and surprising cases are key to developing machine learning models with high sensitivity. Working with domain experts gives me confidence that the data labels we have are closer to a real idea of mental illness severity that may be used in a clinic or by a doctor, since Stephanie and Erica see patients in addition to conducting research. The second benefit is true of all collaborative research: I enjoy working with smart people, and their approaches encourage me to think about my problems in new ways. Our work together helps them consider theirs in novel ways, too.

(continued)

Case Study 2.1 (continued)

> I've used this approach to great success in studies about other behaviors like eating disorders and opioid addiction. Recently, I have been working with domain experts about suicidal ideation in online communities with the US Centers for Disease Control and Prevention to help them develop monitoring for suicide risk. I aim to involve experts in my work from the very beginning of my projects and view this as a core component of human-centered data science. "Construct validity" was the term I used at the outset, but ultimately this means not losing the humans represented by labeled training data. Keeping humans, and the experts that know about the human phenomena we're studying, at the center of our work helps ensure that the results we generate are valid, useful, and grounded in appropriate domain expertise for the people we are trying to help.

In addition, we would like to know exactly in what way the model fails in its prediction and where it does well. A *confusion matrix* presents the full performance of the model in a matrix form. Assuming a binary classification problem, we have two real classes in the data: 0 and 1. It is common to use 1 for the presence of something (or a positive outcome) and 0 for its absence. We also have our classification predictions, which may or may not match the true class of each sample.

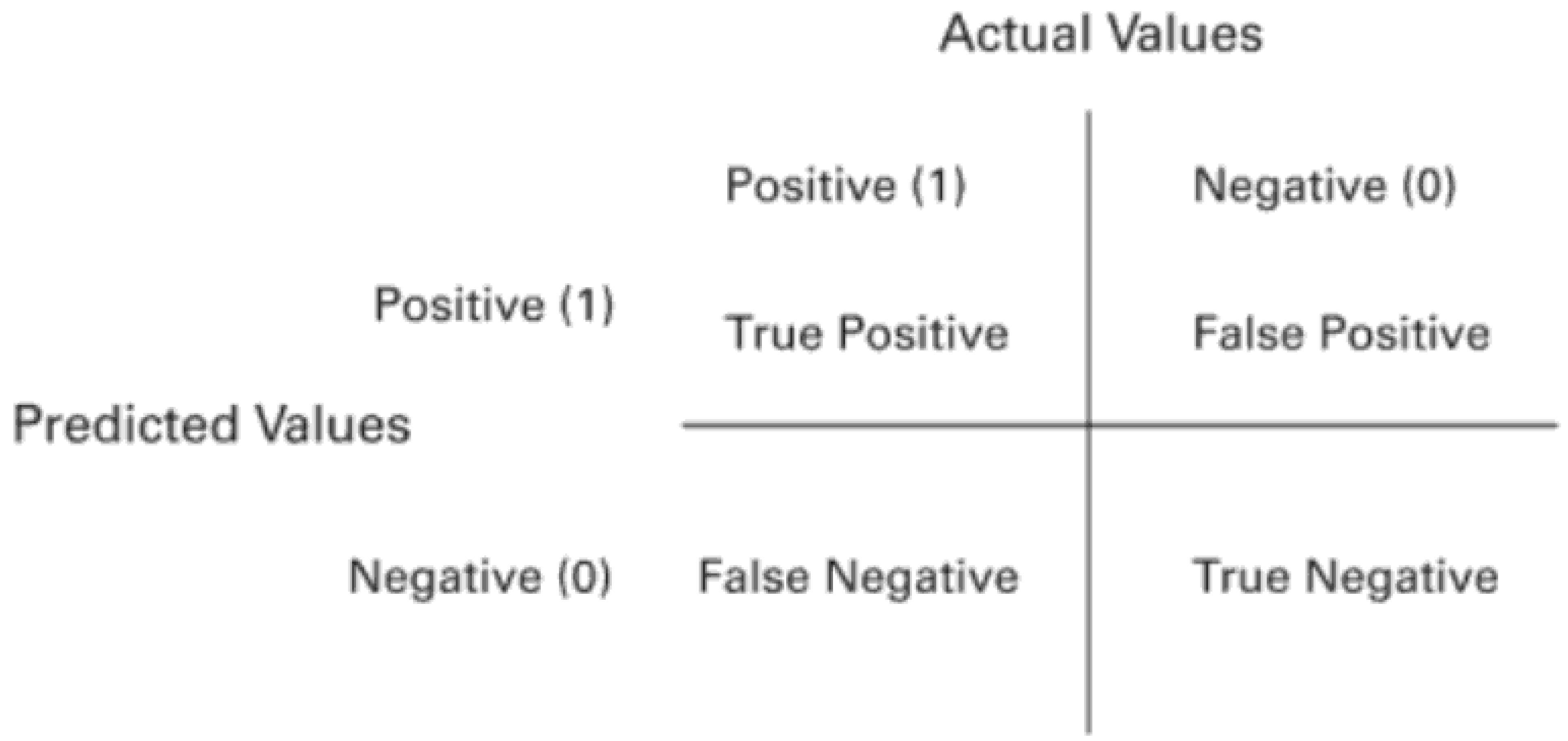
The confusion matrix (figure 2.2) cross-tabulates the actual class labels (rows) and the labels predicted by our model (columns). The resulting cells contain frequencies of four important measures. Suppose that we are trying to predict outcomes that can be represented numerically as 0 or 1, where 0 represents a negative outcome and 1 positive. Then we can define the following measures:

- True positives (TP): the cases in which we predicted 1 and the actual output was also 1.
- True negatives (TN): the cases in which we predicted 0 and the actual output was 0.
- False positives (FP): the cases in which we predicted 1 and the actual output was 0.
- False negatives (FN): the cases in which we predicted 0 and the actual output was 1.

False positives and negatives are important measures for diagnosing the weaknesses of our model and determining where it failed and how.

Another way to evaluate the success of a model is through the metrics of precision and recall. *Precision* is the number of correct positive results divided by the number of positive results predicted by the classifier. This is the portion of true positives out of all the positive results predicted by the model (true and false). It signifies how precise the model is—that is, how many positive instances it classifies correctly. *Recall* is the number of correct positive results divided by the total number of relevant samples. Essentially, this is the proportion of true positives out of all the predictions. It signifies how robust the model is (i.e., it doesn't miss a significant number of positive instances). High precision but lower recall means an accurate model, which at the same time misses a large number of instances that are difficult to classify.

One of Michael's participants described a project to determine whether the audiogram of a human cough was a symptom of a viral infection or a bacterial infection (Muller,



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive	False Positive
	Negative (0)	False Negative	True Negative

Figure 2.2
Confusion matrix.

Lange et al. 2019). Coughs were being measured in low-resourced field clinics, with very limited supplies. The goal was to make an audio recording of a cough using a smartphone, then use it to make a diagnosis. Precision tells us about the "positive" result—in which case the cough is due to bacterial infection and we should treat it with an antibiotic. Of course, we want high precision in order to save lives. Recall tells us about the "negative" result in which the cough is due to a virus; in this case, an antibiotic would not be an effective treatment. Because the field clinics had a limited supply of antibiotics, it was also important not to deplete that limited supply when the antibiotic would not help the patient. The clinic didn't want to run out of antibiotics and then not be able to treat someone with a bacterial infection. Because precision and recall are both important for most data science applications, the data scientists in this situation also computed *F1*, which is a harmonic mean between precision and recall. It provides a balanced metric that accounts for both how precise and robust the model is.

Representing Results

There are many potential ways to represent the results of a data science project. The two most common approaches are data visualization and modeling (usually using machine learning or statistical analysis). There are trade-offs between these two approaches. Most modeling approaches provide a quantitative summary of how well we predicted whatever it is that we formulated a problem for. Data visualization, on the other hand, offers a graphical (and often more intuitive, for most people) way of communicating the results. Of course, visualization can also go beyond being merely a representation of the results of modeling and may itself form a critical part of the data science cycle, one that works in parallel with modeling (see the section on visual analytics in chapter 4).

To represent the results of data science, that information needs to be conveyed to humans. The human visual system is the highest-bandwidth channel into the human brain (Ware 2020). We can process far more bits of information through sight (with our eyes) than we can through any of our other senses. Thus, visual representations of data are often the best ways to powerfully, effectively, and memorably convey the meaning of large amounts of data.

the time and effort in a data science project does *not* involve the model but takes place during data cleaning and feature engineering. These stages are essential because they provide the model with the clean and regularized datasets it needs for computation.

Thus, what is actually *delivered* from a data science project is the pipeline, including the methods and parameters for inputting data, cleaning data, and engineering features from the data—and *then* the model. Often, the client or "user" of a data science outcome wants to receive a "product" that they can put into use in their project, service, or product. They may not want to "look inside" the pipeline to understand each step and stage. It is common for data scientists to provide a single block of code (in the form of a pipeline or a notebook or of a compiled system) that contains the detailed steps that we have described. We may like to think that precision, recall, and F1 are properties of our model. However, when we deliver the outcome to the client, it is the solution (i.e., the pipeline) that has functional properties such as precision, recall, and F1. Inadequate data cleaning and feature engineering can negatively impact those accuracy metrics. Worse, these inadequacies can also produce incorrect outcomes that can harm people's fates in medical procedures, banking systems, or criminal justice systems. Thus, in this book, we focus on the pipeline as the client-oriented outcome of data science.

Pipelines are especially easy to share in the form of Jupyter notebooks or other integrated environments. Version control systems like Git and sites that implement them such as GitHub are excellent repositories for sharing pipelines, because they keep track of all the changes in the pipeline over time. They also easily allow others to create their own copies of the pipeline. This is especially important for *reproducibility*—the idea that we need to make all the resources available and transparent so that others who are interested in related questions could reproduce our work. While traditional data science is very much concerned with reproducibility, it is especially important for human-centered data science as it concerns the people who will potentially use our results or the entire pipeline. We will discuss the importance of reproducibility in more detail and how to cultivate it at every step of the data science cycle in chapter 5.

Interpreting and Communicating Results

A key component of the data science cycle is interpreting the results and communicating them to others. We want others to be able to reproduce our results, and that requires getting out the word on what we did and how. Thoroughly documenting and describing our pipeline is important for others to be able to follow along, and especially for making explicit our mental model of how to address the questions of interest by using the dataset we chose. We discuss documentation further in chapter 3 and chapter 7.

Another important aspect of others being able to use your pipelines is the *interpretability* of your model, meaning how well the model results can be explained. In data science, this is also sometimes called *explainability*. Explainability matters because if the model is very difficult to interpret, others might not understand how it works and how to use it properly. This means that people might apply it to questions or datasets for which it is well suited, but they might apply it incorrectly and reach erroneous conclusions or use it correctly but misinterpret the results. Worse, people may apply the model to questions or datasets for which it is not well suited, thus drawing conclusions based on an invalid premise. These are all problems leading to bad data science outcomes. Hence, data