CONTENTS

CHAPTER 1
INTRODUCTION

1.1 An ink blot

In the summer of 2009, mobile phones were ringing all across Rwanda. In addition to the millions of calls from family, friends, and business associates, about 1,000 Rwandans received a call from Joshua Blumenstock and his colleagues. These researchers were studying wealth and poverty by conducting a survey of a random sample of people from a database of 1.5 million customers of Rwanda's largest mobile phone provider. Blumenstock and colleagues asked the randomly selected people if they wanted to participate in a survey, explained the nature of the research to them, and then asked a series of questions about their demographic, social, and economic characteristics.

Everything I have said so far makes this sound like a traditional social science survey. But what comes next is not traditional—at least not yet. In addition to the survey data, Blumenstock and colleagues also had the complete call records for all 1.5 million people. Combining these two sources of data, they used the survey data to train a machine learning model to predict a person's wealth based on their call records. Next, they used this model to estimate the wealth of all 1.5 million customers in the database. They also estimated the places of residence of all 1.5 million customers using the geographic information embedded in the call records. Putting all of this together—the estimated wealth and the estimated place of residence—they were able to produce high-resolution maps of the geographic distribution of wealth in Rwanda. In particular, they could produce an estimated wealth for each of Rwanda's 2,148 cells, the smallest administrative unit in the country.

It was impossible to validate these estimates because nobody had ever produced estimates for such small geographic areas in Rwanda. But when Blumenstock and colleagues aggregated their estimates to Rwanda's thirty districts, they found that these estimates were very similar to those from the

Demographic and Health Survey, which is widely considered to be the gold standard of surveys in developing countries. Although these two approaches produced similar estimates in this case, the approach of Blumenstock and colleagues was about ten times faster and fifty times cheaper than the traditional Demographic and Health Surveys. These dramatically faster and cheaper estimates create new possibilities for researchers, governments, and companies (Blumenstock, Cadamuro, and On 2015).

This study is kind of like a Rorschach inkblot test: what people see depends on their background. Many *social scientists* see a new measurement tool that can be used to test theories about economic development. Many *data scientists* see a cool new machine learning problem. Many *business people* see a powerful approach for unlocking value in the big data that they have already collected. Many *privacy advocates* see a scary reminder that we live in a time of mass surveillance. And finally, many *policy makers* see a way that new technology can help create a better world. In fact, this study is all of those things, and because it has this mix of characteristics, I see it as a window into the future of social research.

1.2 Welcome to the digital age

The digital age is everywhere, it's growing, and it changes what is possible for researchers.

The central premise of this book is that the digital age creates new opportunities for social research. Researchers can now observe behavior, ask questions, run experiments, and collaborate in ways that were simply impossible in the recent past. Along with these new opportunities come new risks: researchers can now harm people in ways that were impossible in the recent past. The source of these opportunities and risks is the transition from the analog age to the digital age. This transition has not happened all at once—like a light switch turning on—and, in fact, it is not yet complete. However, we've seen enough by now to know that something big is going on.

One way to notice this transition is to look for changes in your daily life. Many things in your life that used to be analog are now digital. Maybe you used to use a camera with film, but now you use a digital camera (which is probably part of your smart phone). Maybe you used to read a physical

newspaper, but now you read an online newspaper. Maybe you used to pay for things with cash, but now you pay with a credit card. In each case, the change from analog to digital means that more data about you are being captured and stored digitally.

In fact, when looked at in aggregate, the effects of the transition are astonishing. The amount of information in the world is rapidly increasing, and more of that information is stored digitally, which facilitates analysis, transmission, and merging (figure 1.1). All of this digital information has come to be called "big data." In addition to this explosion of digital data, there is a parallel growth in our access to computing power (figure 1.1). These trends—increasing amounts of digital data and increasing use of computing—are likely to continue for the foreseeable future.

For the purposes of social research, I think the most important feature of the digital age is *computers everywhere*. Beginning as room-sized machines that were available only to governments and big companies, computers have been shrinking in size and increasing in ubiquity. Each decade since the 1980s has seen a new kind of computing emerge: personal computers, laptops, smart phones, and now embedded processors in the "Internet of Things" (i.e., computers inside of devices such as cars, watches, and thermostats) (Waldrop 2016). Increasingly, these ubiquitous computers do more than just calculate: they also sense, store, and transmit information.

For researchers, the implications of the presence of computers everywhere are easiest to see online, an environment that is fully measured and amenable to experimentation. For example, an online store can easily collect incredibly precise data about the shopping patterns of millions of customers. Further, it can easily randomize groups of customers to receive different shopping experiences. This ability to randomize on top of tracking means that online stores can constantly run randomized controlled experiments. In fact, if you've ever bought anything from an online store, your behavior has been tracked and you've almost certainly been a participant in an experiment, whether you knew it or not.

This fully measured, fully randomizable world is not just happening online; it is increasingly happening everywhere. Physical stores already collect extremely detailed purchase data, and they are developing infrastructure to monitor customers' shopping behavior and mix experimentation into routine business practice. The "Internet of Things" means that behavior in the physical world will increasingly be captured by digital sensors. In other

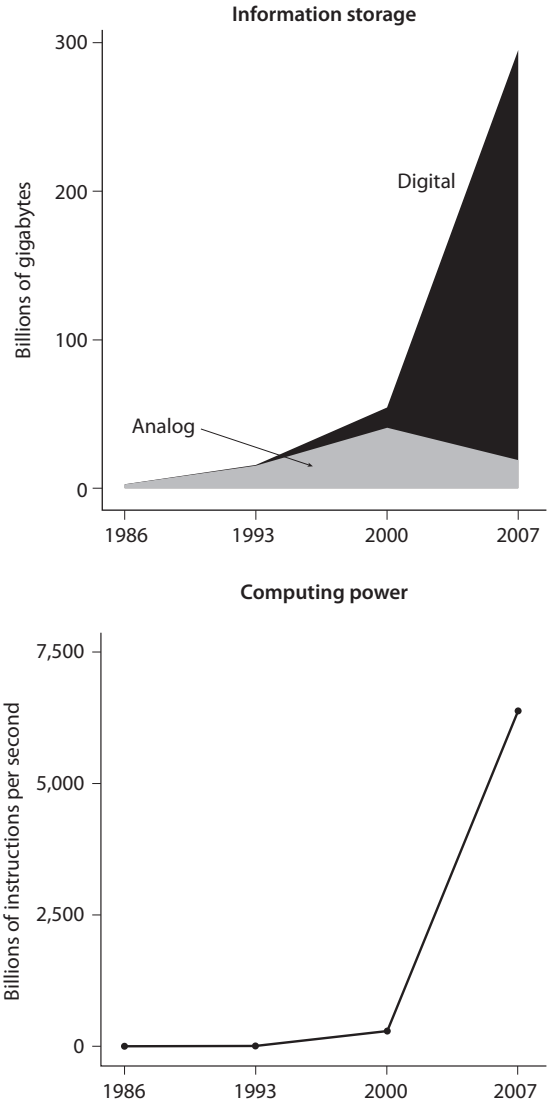


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

words, when you think about social research in the digital age, you should not just think *online*, you should think *everywhere*.

In addition to enabling the measurement of behavior and randomization of treatments, the digital age has also created new ways for people to communicate. These new forms of communication allow researchers to run innovative surveys and to create mass collaboration with their colleagues and the general public.

A skeptic might point out that none of these capabilities are really new. That is, in the past, there have been other major advances in people's abilities to communicate (e.g., the telegraph (Gleick 2011)), and computers have been getting faster at roughly the same rate since the 1960s (Waldrop 2016). But what this skeptic is missing is that at a certain point more of the same becomes something different (Halevy, Norvig, and Pereira 2009). Here's an analogy that I like. If you can capture an image of a horse, then you have a photograph. And if you can capture 24 images of a horse per second, then you have a movie. Of course, a movie is just a bunch of photos, but only a die-hard skeptic would claim that photos and movies are the same.

Researchers are in the process of making a change akin to the transition from photography to cinematography. This change, however, does not mean that everything we have learned in the past should be ignored. Just as the principles of photography inform those of cinematography, the principles of social research that have been developed over the past 100 years will inform the social research taking place over the next 100 years. But the change also means that we should not just keep doing the same thing. Rather, we must combine the approaches of the past with the capabilities of the present and future. For example, the research of Joshua Blumenstock and colleagues was a mixture of traditional survey research with what some might call data science. Both of these ingredients were necessary: neither the survey responses nor the call records by themselves were enough to produce high-resolution estimates of poverty. More generally, social researchers will need to combine ideas from social science and data science in order to take advantage of the opportunities of the digital age: neither approach alone will be enough.

1.3 Research design

Research design is about connecting questions and answers.

This book is written for two audiences that have a lot to learn from each other. On the one hand, it is for social scientists who have training and experience studying social behavior, but who are less familiar with the opportunities created by the digital age. On the other hand, it is for another group of researchers who are very comfortable using the tools of the digital age, but who are new to studying social behavior. This second group resists an easy name, but I will call them data scientists. These data scientists—who often have training in fields such as computer science, statistics, information science, engineering, and physics—have been some of the earliest adopters of digital-age social research, in part because they have access to the necessary data and computational skills. This book attempts to bring these two communities together to produce something richer and more interesting than either community could produce individually.

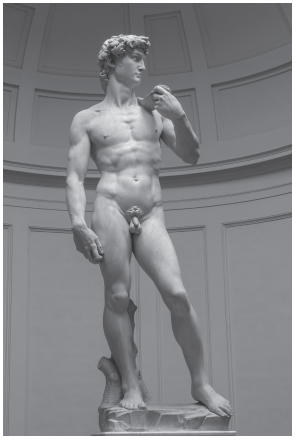
The best way to create this powerful hybrid is not to focus on abstract social theory or fancy machine learning. The best place to start is *research design*. If you think of social research as the process of asking and answering questions about human behavior, then research design is the connective tissue; research design links questions and answers. Getting this connection right is the key to producing convincing research. This book will focus on four approaches that you have seen—and maybe used—in the past: observing behavior, asking questions, running experiments, and collaborating with others. What is new, however, is that the digital age provides us with different opportunities for collecting and analyzing data. These new opportunities require us to modernize—but not replace—these classic approaches.

1.4 Themes of this book

Two themes in the book are (1) mixing readymades and custommades and (2) ethics.

Two themes run throughout this book, and I'd like to highlight them now so that you notice them as they come up over and over again. The first can be illustrated by an analogy that compares two greats: Marcel Duchamp and Michelangelo. Duchamp is mostly known for his readymades, such as *Fountain*, where he took ordinary objects and repurposed them as art. Michelangelo, on the other hand, didn't repurpose. When he wanted to

Readymade



Custommade

Figure 1.2: *Fountain* by Marcel Duchamp and *David* by Michelangelo. *Fountain* is an example of a readymade, where an artist sees something that already exists in the world and then creatively repurposes it for art. *David* is an example of art that was intentionally created; it is a custommade. Social research in the digital age will involve both readymades and custommades. Photograph of *Fountain* by Alfred Stieglitz, 1917 (Source: *The Blind Man*, no. 2/Wikimedia Commons). Photograph of *David* by Jörg Bittner Unna, 2008 (Source: Galleria dell'Accademia, Florence/Wikimedia Commons).

create a statue of David, he didn't look for a piece of marble that kind of looked like David: he spent three years laboring to create his masterpiece. *David* is not a readymade; it is a custommade (figure 1.2).

These two styles—readymades and custommades—roughly map onto styles that can be employed for social research in the digital age. As you will see, some of the examples in this book involve clever repurposing of big data sources that were originally created by companies and governments. In other examples, however, a researcher started with a specific question and then used the tools of the digital age to create the data needed to answer that question. When done well, both of these styles can be incredibly powerful. Therefore, social research in the digital age will involve both readymades and custommades; it will involve both Duchamps and Michelangelos.

If you generally use readymade data, I hope that this book will show you the value of custommade data. And likewise, if you generally use custommade data, I hope that this book will show you the value of readymade data. Finally, and most importantly, I hope that this book will show you

the value of combining these two styles. For example, Joshua Blumenstock and colleagues were part Duchamp and part Michelangelo: they repurposed the call records (a readymade), and they created their own survey data (a custommade). This blending of readymades and custommades is a pattern that you'll see throughout this book; it tends to require ideas from both social science and data science, and it often leads to the most exciting research.

A second theme that runs through this book is ethics. I'll show you how researchers can use the capabilities of the digital age to conduct exciting and important research. And I'll show you how researchers who take advantage of these opportunities will confront difficult ethical decisions. Chapter 6 will be entirely devoted to ethics, but I integrate ethics into the other chapters as well because, in the digital age, ethics will become an increasingly integral part of research design.

The work of Blumenstock and colleagues is again illustrative. Having access to the granular call records from 1.5 million people creates wonderful opportunities for research, but it also creates opportunities for harm. For example, Jonathan Mayer and colleagues (2016) have shown that even "anonymized" call records (i.e., data without names and addresses) can be combined with publicly available information in order to identify specific people in the data and to infer sensitive information about them, such as certain health information. To be clear, Blumenstock and colleagues did not attempt to identify specific people and infer sensitive information about them, but this possibility meant that it was difficult for them to acquire the call data, and it forced them to take extensive safeguards while conducting their research.

Beyond the details of the call records, there is a fundamental tension that runs through a lot of social research in the digital age. Researchers—often in collaboration with companies and governments—have increasing power over the lives of participants. By power, I mean the ability to do things to people without their consent or even awareness. For example, researchers can now observe the behavior of millions of people, and, as I'll describe later, researchers can also enroll millions of people in massive experiments. Further, all of this can happen without the consent or awareness of the people involved. As the power of researchers is increasing, there has not been an equivalent increase in clarity about how that power should be used. In fact, researchers must decide how to exercise their power based

on inconsistent and overlapping rules, laws, and norms. This combination of powerful capabilities and vague guidelines can force even well-meaning researchers to grapple with difficult decisions.

If you generally focus on how digital-age social research creates new opportunities, I hope that this book will show you that these opportunities also create new risks. And likewise, if you generally focus on these risks, I hope that this book will help you see the opportunities—opportunities that may require certain risks. Finally, and most importantly, I hope that this book will help everyone to responsibly balance the risks and opportunities created by digital-age social research. With an increase in power, there must also come an increase in responsibility.

1.5 Outline of this book

This book progresses through four broad research designs: observing behavior, asking questions, running experiments, and creating mass collaboration. Each of these approaches requires a different relationship between researchers and participants, and each enables us to learn different things. That is, if we ask people questions, we can learn things that we could not learn merely by observing behavior. Likewise, if we run experiments, we can learn things that we could not learn merely by observing behavior and asking questions. Finally, if we collaborate with participants, we can learn things that we could not learn by observing them, asking them questions, or enrolling them in experiments. These four approaches were all used in some form fifty years ago, and I'm confident that they will all still be used in some form fifty years from now. After devoting one chapter to each approach, including the ethical issues raised by that approach, I'll devote a full chapter to ethics. As mentioned in the preface, I'm going to keep the main text of the chapters as clean as possible, and each of them will conclude with a section called "What to read next" that includes important bibliographic information and pointers to more detailed material.

Looking ahead, in chapter 2 ("Observing behavior"), I'll describe what and how researchers can learn from observing people's behavior. In particular, I'll focus on big data sources created by companies and governments. Abstracting away from the details of any specific source, I'll describe 10 common features of the big data sources and how these impact researchers' ability to use these data sources for research. Then, I'll illustrate

three research strategies that can be used to successfully learn from big data sources.

In chapter 3 ("Asking questions"), I'll begin by showing what researchers can learn by moving beyond preexisting big data. In particular, I'll show that by asking people questions, researchers can learn things that they can't easily learn by just observing behavior. In order to organize the opportunities created by the digital age, I'll review the traditional total survey error framework. Then, I'll show how the digital age enables new approaches to both sampling and interviewing. Finally, I'll describe two strategies for combining survey data and big data sources.

In chapter 4 ("Running experiments"), I'll begin by showing what researchers can learn when they move beyond observing behavior and asking survey questions. In particular, I'll show how randomized controlled experiments—where the researcher intervenes in the world in a very specific way—enable researchers to learn about causal relationships. I'll compare the kinds of experiments that we could do in the past with the kinds that we can do now. With that background, I'll describe the trade-offs involved in the two main strategies for conducting digital experiments. Finally, I'll conclude with some design advice about how you can take advantage of the real power of digital experiments, and I'll describe some of the responsibilities that come with that power.

In chapter 5 ("Creating mass collaboration"), I'll show how researchers can create mass collaborations—such as crowdsourcing and citizen science—in order to do social research. By describing successful mass collaboration projects and by providing a few key organizing principles, I hope to convince you of two things: first, that mass collaboration can be harnessed for social research, and, second, that researchers who use mass collaboration will be able to solve problems that had previously seemed impossible.

In chapter 6 ("Ethics"), I'll argue that researchers have rapidly increasing power over participants and that these capabilities are changing faster than our norms, rules, and laws. This combination of increasing power and lack of agreement about how that power should be used leaves well-meaning researchers in a difficult situation. To address this problem, I'll argue that researchers should adopt a *principles-based* approach. That is, researchers should evaluate their research through existing rules—which I'll take as given—and through more general ethical principles. I'll describe four established principles and two ethical frameworks that can help guide researchers'

decisions. Finally, I'll explain some specific ethical challenges that I expect will confront researchers in the future, and I'll offer practical tips for working in an area with unsettled ethics.

Finally, in chapter 7 ("The future"), I'll review the themes that run through the book, and then use them to speculate about themes that will be important in the future.

Social research in the digital age will combine what we have done in the past with the very different capabilities of the future. Thus, social research will be shaped by both social scientists and data scientists. Each group has something to contribute, and each has something to learn.

What to read next

- **An ink blot (section 1.1)**

  For a more detailed description of the project of Blumenstock and colleagues, see chapter 3 of this book.

- **Welcome to the digital age (section 1.2)**

  Gleick (2011) provides a historical overview of changes in humanity's ability to collect, store, transmit, and process information.

  For an introduction to the digital age that focuses on potential harms, such as privacy violations, see Abelson, Ledeen, and Lewis (2008) and Mayer-Schönberger (2009). For an introduction to the digital age that focuses on research opportunities, see Mayer-Schönberger and Cukier (2013).

  For more about firms mixing experimentation into routine practice, see Manzi (2012), and for more about firms tracking behavior in the physical world, see Levy and Baracas (2017).

  Digital-age systems can be both instruments and objects of study. For example, you might want to use social media to measure public opinion or you might want to understand the impact of social media on public opinion. In one case, the digital system serves as an instrument that helps you do new measurement. In the other case, the digital system is the object of study. For more on this distinction, see Sandvig and Hargittai (2015).

- **Research design (section 1.3)**

  For more on research design in the social sciences, see Singleton and Straits (2009), King, Keohane, and Verba (1994), and Khan and Fisher (2013).

  Donoho (2015) describes data science as the activities of people learning from data, and offers a history of data science, tracing the intellectual origins of the field to scholars such as Tukey, Cleveland, Chambers, and Breiman.

For a series of first-person reports about conducting social research in the digital age, see Hargittai and Sandvig (2015).

- **Themes of this book (section 1.4)**

For more about mixing readymade and custommade data, see Groves (2011).

For more about failure of "anonymization," see chapter 6 of this book. The same general technique that Blumenstock and colleagues used to infer people's wealth can also be used to infer potentially sensitive personal attributes, including sexual orientation, ethnicity, religious and political views, and use of addictive substances; see Kosinski, Stillwell, and Graepel (2013).

INDEX