CHAPTER 2

OBSERVING BEHAVIOR

2.1 Introduction

In the analog age, collecting data about behavior—who does what, and when—was expensive and therefore relatively rare. Now, in the digital age, the behaviors of billions of people are recorded, stored, and analyzable. For example, every time you click on a website, make a call on your mobile phone, or pay for something with your credit card, a digital record of your behavior is created and stored by a business. Because these types of data are a by-product of people's everyday actions, they are often called *digital traces*. In addition to these traces held by businesses, there are also large amounts of incredibly rich data held by governments. Together, these business and government records are often called *big data*.

The ever-rising flood of big data means that we have moved from a world where behavioral data was scarce to one where it is plentiful. A first step to learning from big data is realizing that it is part of a broader category of data that has been used for social research for many years: *observational data*. Roughly, observational data is any data that results from observing a social system without intervening in some way. A crude way to think about it is that observational data is everything that does not involve talking with people (e.g., surveys, the topic of chapter 3) or changing people's environments (e.g., experiments, the topic of chapter 4). Thus, in addition to business and government records, observational data also includes things like the text of newspaper articles and satellite photos.

This chapter has three parts. First, in section 2.2, I describe big data sources in more detail and clarify a fundamental difference between them and the data that have typically been used for social research in the past. Then, in section 2.3, I describe 10 common characteristics of big data sources. Understanding these characteristics enables you to quickly recognize the

strengths and weaknesses of existing sources and will help you harness the new sources that will be available in the future. Finally, in section 2.4, I describe three main research strategies that you can use to learn from observational data: counting things, forecasting things, and approximating an experiment.

2.2 Big data

> Big data are created and collected by companies and governments for purposes other than research. Using this data for research therefore requires repurposing.

The first way that many people encounter social research in the digital age is through what is often called *big data*. Despite the widespread use of this term, there is no consensus about what big data even is. However, one of the most common definitions of big data focuses on the "3 Vs": Volume, Variety, and Velocity. Roughly, there is a lot of data, in a variety of formats, and it is being created constantly. Some fans of big data also add other "Vs," such as Veracity and Value, whereas some critics add "Vs" such as Vague and Vacuous. Rather than the "3 Vs" (or the "5 Vs" or the "7 Vs"), for the purposes of social research, I think a better place to start is the "5 Ws": Who, What, Where, When, and Why. In fact, I think that many of the challenges and opportunities created by big data sources follow from just one "W": Why.

In the analog age, most of the data that were used for social research were created for the purpose of doing research. In the digital age, however, huge amounts of data are being created by companies and governments for purposes other than research, such as providing services, generating profit, and administering laws. Creative people, however, have realized that you can *repurpose* this corporate and government data for research. Thinking back to the art analogy in chapter 1, just as Duchamp repurposed a found object to create art, scientists can now repurpose found data to create research.

While there are undoubtedly huge opportunities for repurposing, using data that were not created for the purposes of research also presents new challenges. Compare, for example, a social media service, such as Twitter, with a traditional public opinion survey, such as the General Social Survey.

Twitter's main goals are to provide a service to its users and to make a profit. The General Social Survey, on the other hand, is focused on creating general-purpose data for social research, particularly for public opinion research. This difference in goals means that the data created by Twitter and that created by the General Social Survey have different properties, even though both can be used for studying public opinion. Twitter operates at a scale and speed that the General Social Survey cannot match, but, unlike the General Social Survey, Twitter does not carefully sample users and does not work hard to maintain comparability over time. Because these two data sources are so different, it does not make sense to say that the General Social Survey is better than Twitter, or vice versa. If you want hourly measures of global mood (e.g., Golder and Macy (2011)), Twitter is the best choice. On the other hand, if you want to understand long-term changes in the polarization of attitudes in the United States (e.g., DiMaggio, Evans, and Bryson (1996)), then the General Social Survey is best. More generally, rather than trying to argue that big data sources are better or worse than other types of data, this chapter will try to clarify for which kinds of research questions big data sources have attractive properties and for which kinds of questions they might not be ideal.

When thinking about big data sources, many researchers immediately focus on online data created and collected by companies, such as search engine logs and social media posts. However, this narrow focus leaves out two other important sources of big data. First, increasingly, corporate big data sources come from digital devices in the physical world. For example, in this chapter, I'll tell you about a study that repurposed supermarket check-out data to study how a worker's productivity is impacted by the productivity of her peers (Mas and Moretti 2009). Then, in later chapters, I'll tell you about researchers who used call records from mobile phones (Blumenstock, Cadamuro, and On 2015) and billing data created by electric utilities (Allcott 2015). As these examples illustrate, corporate big data sources are about more than just online behavior.

The second important source of big data missed by a narrow focus on online behavior is data created by governments. These government data, which researchers call *government administrative records*, include things such as tax records, school records, and vital statistics records (e.g., registries of births and deaths). Governments have been creating these kinds of data for, in some cases, hundreds of years, and social scientists have been

exploiting them for nearly as long as there have been social scientists. What has changed, however, is digitization, which has made it dramatically easier for governments to collect, transmit, store, and analyze data. For example, in this chapter, I'll tell you about a study that repurposed data from New York City government's digital taxi meters in order to address a fundamental debate in labor economics (Farber 2015). Then, in later chapters, I'll tell you about how government-collected voting records were used in a survey (Ansolabehere and Hersh 2012) and an experiment (Bond et al. 2012).

I think the idea of repurposing is fundamental to learning from big data sources, and so, before talking more specifically about the properties of big data sources (section 2.3) and how these can be used in research (section 2.4), I'd like to offer two pieces of general advice about repurposing. First, it can be tempting to think about the contrast that I've set up as being between "found" data and "designed" data. That's close, but it's not quite right. Even though, from the perspective of researchers, big data sources are "found," they don't just fall from the sky. Instead, data sources that are "found" by researchers are designed by someone for some purpose. Because "found" data are designed by someone, I always recommend that you try to understand as much as possible about the people and processes that created your data. Second, when you are repurposing data, it is often extremely helpful to imagine the ideal dataset for your problem and then compare that ideal dataset with the one that you are using. If you didn't collect your data yourself, there are likely to be important differences between what you want and what you have. Noticing these differences will help clarify what you can and cannot learn from the data you have, and it might suggest new data that you should collect.

In my experience, social scientists and data scientists tend to approach repurposing very differently. Social scientists, who are accustomed to working with data designed for research, are typically quick to point out the problems with repurposed data, while ignoring its strengths. On the other hand, data scientists are typically quick to point out the benefits of repurposed data, while ignoring its weaknesses. Naturally, the best approach is a hybrid. That is, researchers need to understand the characteristics of big data sources—both good and bad—and then figure out how to learn from them. And, that is the plan for the remainder of this chapter. In the next section, I will describe 10 common characteristics of big data sources. Then, in the

following section, I will describe three research approaches that can work well with such data.

2.3 Ten common characteristics of big data

> Big data sources tend to have a number of characteristics in common; some are generally good for social research and some are generally bad.

Even though each big data source is distinct, it is helpful to notice that there are certain characteristics that tend to occur over and over again. Therefore, rather than taking a platform-by-platform approach (e.g., here's what you need to know about Twitter, here's what you need to know about Google search data, etc.), I'm going to describe 10 general characteristics of big data sources. Stepping back from the details of each particular system and looking at these general characteristics enables researchers to quickly learn about existing data sources and have a firm set of ideas to apply to the data sources that will be created in the future.

Even though the desired characteristics of a data source depend on the research goal, I find it helpful to crudely group the 10 characteristics into two broad categories:

- generally helpful for research: big, always-on, and nonreactive
- generally problematic for research: incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, dirty, and sensitive

As I'm describing these characteristics, you'll notice that they often arise because big data sources were not created for the purpose of research.

2.3.1 Big

> Large datasets are a means to an end; they are not an end in themselves.

The most widely discussed feature of big data sources is that they are BIG. Many papers, for example, start by discussing—and sometimes bragging—about how much data they analyzed. For example, a paper published in

Science studying word-use trends in the Google Books corpus included the following (Michel et al. 2011):

> "[Our] corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion. The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over."

The scale of this data is undoubtedly impressive, and we are all fortunate that the Google Books team has released these data to the public (in fact, some of the activities at the end of this chapter make use of this data). However, whenever you see something like this, you should ask: Is that all that data really doing anything? Could they have done the same research if the data could reach to the Moon and back only once? What if the data could only reach to the top of Mount Everest or the top of the Eiffel Tower?

In this case, their research does, in fact, have some findings that require a huge corpus of words over a long time period. For example, one thing they explore is the evolution of grammar, particularly changes in the rate of irregular verb conjugation. Since some irregular verbs are quite rare, a large amount of data is needed to detect changes over time. Too often, however, researchers seem to treat the size of big data source as an end—"look how much data I can crunch"—rather than a means to some more important scientific objective.

In my experience, the study of rare events is one of the three specific scientific ends that large datasets tend to enable. The second is the study of heterogeneity, as can be illustrated by a study by Raj Chetty and colleagues (2014) on social mobility in the United States. In the past, many researchers have studied social mobility by comparing the life outcomes of parents and

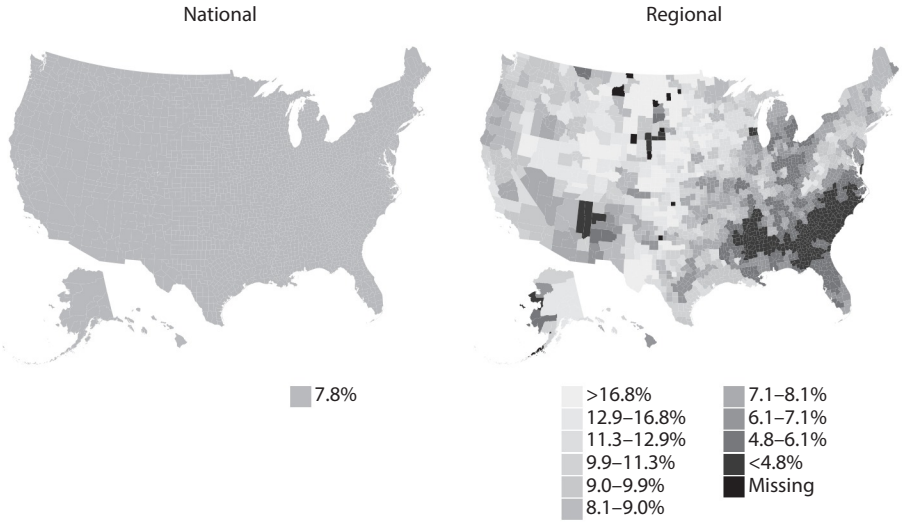

Figure 2.1: Estimates of a child's chances of reaching the top 20% of income distribution given parents in the bottom 20% (Chetty et al. 2014). The regional-level estimates, which show heterogeneity, naturally lead to interesting and important questions that do not arise from a single national-level estimate. These regional-level estimates were made possible in part because the researchers were using a large big data source: the tax records of 40 million people. Created from data available at http://www.equality-of-opportunity.org/.

children. A consistent finding from this literature is that advantaged parents tend to have advantaged children, but the strength of this relationship varies over time and across countries (Hout and DiPrete 2006). More recently, however, Chetty and colleagues were able to use the tax records from 40 million people to estimate the heterogeneity in intergenerational mobility across regions in the United States (figure 2.1). They found, for example, that the probability that a child reaches the top quintile of the national income distribution starting from a family in the bottom quintile is about 13% in San Jose, California, but only about 4% in Charlotte, North Carolina. If you look at figure 2.1 for a moment, you might begin to wonder why intergenerational mobility is higher in some places than others. Chetty and colleagues had exactly the same question, and they found that that high-mobility areas have less residential segregation, less income inequality, better primary schools, greater social capital, and greater family stability. Of course, these correlations alone do not show that these factors cause higher mobility, but they do suggest possible mechanisms that can be explored in further work,

which is exactly what Chetty and colleagues have done in subsequent work. Notice how the size of the data was really important in this project. If Chetty and colleagues had used the tax records of 40 thousand people rather than 40 million, they would not have been able to estimate regional heterogeneity, and they never would have been able to do subsequent research to try to identify the mechanisms that create this variation.

Finally, in addition to studying rare events and studying heterogeneity, large datasets also enable researchers to detect small differences. In fact, much of the focus on big data in industry is about these small differences: reliably detecting the difference between 1% and 1.1% click-through rates on an ad can translate into millions of dollars in extra revenue. In some scientific settings, however, such small differences might not be particular important, even if they are statistically significant (Prentice and Miller 1992). But, in some policy settings, they can become important when viewed in aggregate. For example, if there are two public health interventions and one is slightly more effective than the other, then picking the more effective intervention could end up saving thousands of additional lives.

Although bigness is generally a good property when used correctly, I've noticed that it can sometimes lead to a conceptual error. For some reason, bigness seems to lead researchers to ignore how their data was generated. While bigness does reduce the need to worry about random error, it actually *increases* the need to worry about systematic errors, the kinds of errors that I'll describe below that arise from biases in how data are created. For example, in a project I'll describe later in this chapter, researchers used messages generated on September 11, 2001 to produce a high-resolution emotional timeline of the reaction to the terrorist attack (Back, Küfner, and Egloff 2010). Because the researchers had a large number of messages, they didn't really need to worry about whether the patterns they observed—increasing anger over the course of the day—could be explained by random variation. There was so much data and the pattern was so clear that all the statistical statistical tests suggested that this was a real pattern. But these statistical tests were ignorant of how the data was created. In fact, it turned out that many of the patterns were attributable to a single bot that generated more and more meaningless messages throughout the day. Removing this one bot completely destroyed some of the key findings in the paper (Pury 2011; Back, Küfner, and Egloff 2011). Quite simply, researchers who don't think about systematic error face the risk of using their large datasets to get a precise estimate

of an unimportant quantity, such as the emotional content of meaningless messages produced by an automated bot.

In conclusion, big datasets are not an end in themselves, but they can enable certain kinds of research, including the study of rare events, the estimation of heterogeneity, and the detection of small differences. Big datasets also seem to lead some researchers to ignore how their data was created, which can lead them to get a precise estimate of an unimportant quantity.

2.3.2 Always-on

> Always-on big data enables the study of unexpected events and real-time measurement.

Many big data systems are *always-on*; they are constantly collecting data. This always-on characteristic provides researchers with longitudinal data (i.e., data over time). Being always-on has two important implications for research.

First, always-on data collection enables researchers to study unexpected events in ways that would not otherwise be possible. For example, researchers interested in studying the Occupy Gezi protests in Turkey in the summer of 2013 would typically focus on the behavior of protesters during the event. Ceren Budak and Duncan Watts (2015) were able to do more by using the always-on nature of Twitter to study protesters who used Twitter before, during, and after the event. And they were able to create a comparison group of nonparticipants before, during, and after the event (figure 2.2). In total, their *ex-post panel* included the tweets of 30,000 people over two years. By augmenting the commonly used data from the protests with this other information, Budak and Watts were able to learn much more: they were able to estimate what kinds of people were more likely to participate in the Gezi protests and to estimate the changes in attitudes of participants and nonparticipants, both in the short term (comparing pre-Gezi with during Gezi) and in the long term (comparing pre-Gezi with post-Gezi).

A skeptic might point out that some of these estimates could have been made without always-on data collection sources (e.g., long-term estimates of attitude change), and that is correct, although such a data collection for 30,000 people would have been quite expensive. Even given an unlimited

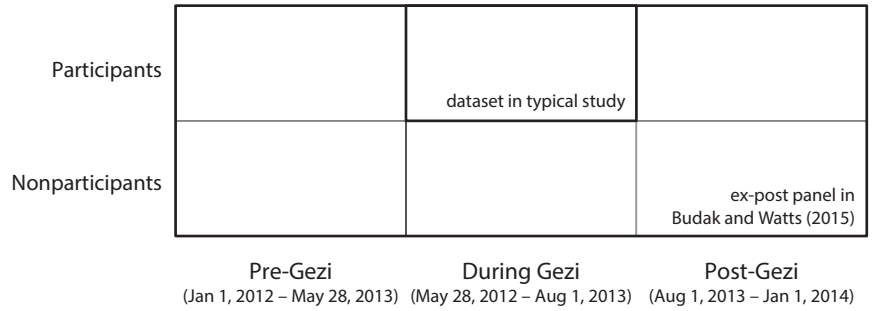

Figure 2.2: Design used by Budak and Watts (2015) to study the Occupy Gezi protests in Turkey in the summer of 2013. By using the always-on nature of Twitter, the researchers created what they called an *ex-post panel* that included about 30,000 people over two years. In contrast to a typical study that focused on participants during the protests, the ex-post panel adds (1) data from participants before and after the event and (2) data from nonparticipants before, during, and after the event. This enriched data structure enabled Budak and Watts to estimate what kinds of people were more likely to participate in the Gezi protests and to estimate the changes in attitudes of participants and nonparticipants, both in the short term (comparing pre-Gezi with during Gezi) and in the long term (comparing pre-Gezi with post-Gezi).

budget, however, I can't think of any other method that essentially allows researchers to *travel back in time* and directly observe participants' behavior in the past. The closest alternative would be to collect retrospective reports of behavior, but these would be of limited granularity and questionable accuracy. Table 2.1 provides other examples of studies that use an always-on data source to study an unexpected event.

In addition to studying unexpected events, always-on big data systems also enable researchers to produce real-time estimates, which can be important in settings where policy makers—in government or industry—want to respond based on situational awareness. For example, social media data can be used to guide emergency response to natural disasters (Castillo 2016), and a variety of different big data sources can be used produce real-time estimates of economic activity (Choi and Varian 2012).

In conclusion, always-on data systems enable researchers to study unexpected events and provide real-time information to policy makers. I do not, however, think that always-on data systems are well suited for tracking changes over very long periods of time. That is because many big data systems are constantly changing—a process that I'll call *drift* later in the chapter (section 2.3.7).

Table 2.1: Studies of Unexpected Events Using Always-On Big Data Sources

Unexpected event	Always-on data source	References
Occupy Gezi movement in Turkey	Twitter	Budak and Watts (2015)
Umbrella protests in Hong Kong	Weibo	Zhang (2016)
Shootings of police in New York City	Stop-and-frisk reports	Legewie (2016)
Person joining ISIS	Twitter	Magdy, Darwish, and Weber (2016)
September 11, 2001 attack	livejournal.com	Cohn, Mehl, and Pennebaker (2004)
September 11, 2001 attack	Pager messages	Back, Küfner, and Egloff (2010), Pury (2011), Back, Küfner, and Egloff (2011)

2.3.3 Nonreactive

> Measurement in big data sources is much less likely to change behavior.

One challenge of social research is that people can change their behavior when they know that they are being observed by researchers. Social scientists generally call this *reactivity* (Webb et al. 1966). For example, people can be more generous in laboratory studies than field studies because in the former they are very aware that they are being observed (Levitt and List 2007a). One aspect of big data that many researchers find promising is that participants are generally not aware that their data are being captured or they have become so accustomed to this data collection that it no longer changes their behavior. Because participants are *nonreactive*, therefore, many sources of big data can be used to study behavior that has not been amenable to accurate measurement previously. For example, Stephens-Davidowitz (2014) used the prevalence of racist terms in search engine queries to measure racial animus in different regions of the United States. The nonreactive and big (see section 2.3.1) nature of the search data enabled measurements that would be difficult using other methods, such as surveys.

Nonreactivity, however, does not ensure that these data are somehow a direct reflection of people's behavior or attitudes. For example, as one respondent in an interview-based study said, "It's not that I don't have problems, I'm just not putting them on Facebook" (Newman et al. 2011). In other words, even though some big data sources are nonreactive, they are not always free of social desirability bias, the tendency for people to want to present themselves in the best possible way. Further, as I'll describe later in the chapter, the behavior captured in big data sources is sometimes impacted by the goals of platform owners, an issue I'll call *algorithmic confounding*. Finally, although nonreactivity is advantageous for research, tracking people's behavior without their consent and awareness raises ethical concerns that I'll describe in chapter 6.

The three properties that I just described—big, always-on, and nonreactive—are generally, but not always, advantageous for social research. Next, I'll turn to the seven properties of big data sources—incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, dirty, and sensitive—that generally, but not always, create problems for research.

2.3.4 Incomplete

> No matter how big your big data, it probably doesn't have the information you want.

Most big data sources are *incomplete*, in the sense that they don't have the information that you will want for your research. This is a common feature of data that were created for purposes other than research. Many social scientists have already had the experience of dealing with incompleteness, such as an existing survey that didn't ask the question that was needed. Unfortunately, the problems of incompleteness tend to be more extreme in big data. In my experience, big data tends to be missing three types of information useful for social research: demographic information about participants, behavior on other platforms, and data to operationalize theoretical constructs.

Of the three kinds of incompleteness, the problem of incomplete data to operationalize theoretical constructs is the hardest to solve. And in my experience, it is often accidentally overlooked. Roughly, *theoretical constructs* are abstract ideas that social scientists study, and *operationalizing* a

theoretical construct means proposing some way to capture that construct with observable data. Unfortunately, this simple-sounding process often turns out to be quite difficult. For example, let's imagine trying to empirically test the apparently simple claim that people who are more intelligent earn more money. In order to test this claim, you would need to measure "intelligence." But what is intelligence? Gardner (2011) argued that there are actually eight different forms of intelligence. And are there procedures that could accurately measure any of these forms of intelligence? Despite enormous amounts of work by psychologists, these questions still don't have unambiguous answers.

Thus, even a relatively simple claim—people who are more intelligent earn more money—can be hard to assess empirically because it can be hard to operationalize theoretical constructs in data. Other examples of theoretical constructs that are important but hard to operationalize include "norms," "social capital," and "democracy." Social scientists call the match between theoretical constructs and data *construct validity* (Cronbach and Meehl 1955). As this short list of constructs suggests, construct validity is a problem that social scientists have struggled with for a very long time. But in my experience, the problems of construct validity are even greater when working with data that were not created for the purposes of research (Lazer 2015).

When you are assessing a research result, one quick and useful way to assess construct validity is to take the result, which is usually expressed in terms of constructs, and re-express it in terms of the data used. For example, consider two hypothetical studies that claim to show that people who are more intelligent earn more money. In the first study, the researcher found that people who score well on the Raven Progressive Matrices Test—a well-studied test of analytic intelligence (Carpenter, Just, and Shell 1990)—have higher reported incomes on their tax returns. In the second study, the researcher found that people on Twitter who used longer words are more likely to mention luxury brands. In both cases, these researchers could claim that they have shown that people who are more intelligent earn more money. However, in the first study the theoretical constructs are well operationalized by the data, while in the second they are not. Further, as this example illustrates, more data does not automatically solve problems with construct validity. You should doubt the results of the second study whether it involved a million tweets, a billion tweets, or a trillion tweets. For researchers not

Table 2.2: Examples of Digital Traces That Were Used to Operationalize Theoretical Constructs

Data source	Theoretical construct	References
Email logs from a university (metadata only)	Social relationships	Kossinets and Watts (2006), Kossinets and Watts (2009), De Choudhury et al. (2010)
Social media posts on Weibo	Civic engagement	Zhang (2016)
Email logs from a firm (metadata and complete text)	Cultural fit in an organization	Srivastava et al. (2017)

familiar with the idea of construct validity, table 2.2 provides some examples of studies that have operationalized theoretical constructs using digital trace data.

Although the problem of incomplete data for capturing theoretical constructs is pretty hard to solve, there are common solutions to the other common types of incompleteness: incomplete demographic information and incomplete information on behavior on other platforms. The first solution is to actually collect the data you need; I'll tell you about that in chapter 3 when I talk about surveys. The second main solution is to do what data scientists call *user-attribute inference* and social scientists call *imputation*. In this approach, researchers use the information that they have on some people to infer attributes of other people. A third possible solution is to combine multiple data sources. This process is sometimes called *record linkage*. My favorite metaphor for this process was written by Dunn (1946) in the very first paragraph of the very first paper ever written on record linkage:

> "Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this book into a volume."

When Dunn wrote that passage, he was imagining that the Book of Life could include major life events like birth, marriage, divorce, and death. However, now that so much information about people is recorded, the Book of Life could be an incredibly detailed portrait, if those different pages

(i.e., our digital traces) can be bound together. This Book of Life could be a great resource for researchers. But it could also be called a *database of ruin* (Ohm 2010), which could be used for all kinds of unethical purposes, as I'll describe in chapter 6 (Ethics).

2.3.5 Inaccessible

> Data held by companies and governments are difficult for researchers to access.

In May 2014, the US National Security Agency opened a data center in rural Utah with an awkward name, the Intelligence Community Comprehensive National Cybersecurity Initiative Data Center. However, this data center, which has come to be known as the Utah Data Center, is reported to have astounding capabilities. One report alleges that it is able to store and process all forms of communication including "the complete contents of private emails, cell phone calls, and Google searches, as well as all sorts of personal data trails—parking receipts, travel itineraries, bookstore purchases, and other digital 'pocket litter'" (Bamford 2012). In addition to raising concerns about the sensitive nature of much of the information captured in big data, which will be described further below, the Utah Data Center is an extreme example of a rich data source that is inaccessible to researchers. More generally, many sources of big data that would be useful are controlled and restricted by governments (e.g., tax data and educational data) or companies (e.g., queries to search engines and phone call meta-data). Therefore, even though these data sources exist, they are useless for the purposes of social research because they are inaccessible.

In my experience, many researchers based at universities misunderstand the source of this inaccessibility. These data are inaccessible not because people at companies and governments are stupid, lazy, or uncaring. Rather, there are serious legal, business, and ethical barriers that prevent data access. For example, some terms-of-service agreements for websites only allow data to be used by employees or to improve the service. So certain forms of data sharing could expose companies to legitimate lawsuits from customers. There are also substantial business risks to companies involved in sharing data. Try to imagine how the public would respond if personal search data accidentally leaked out from Google as part of a university research project.

Such a data breach, if extreme, might even be an existential risk for the company. So Google—and most large companies—are very risk-averse about sharing data with researchers.

In fact, almost everyone who is in a position to provide access to large amounts of data knows the story of Abdur Chowdhury. In 2006, when he was the head of research at AOL, he intentionally released to the research community what he thought were anonymized search queries from 650,000 AOL users. As far as I can tell, Chowdhury and the researchers at AOL had good intentions, and they thought that they had anonymized the data. But they were wrong. It was quickly discovered that the data were not as anonymous as the researchers thought, and reporters from the *New York Times* were able to identify someone in the dataset with ease (Barbaro and Zeller 2006). Once these problems were discovered, Chowdhury removed the data from AOL's website, but it was too late. The data had been reposted on other websites, and it will probably still be available when you are reading this book. Ultimately, Chowdhury was fired, and AOL's chief technology officer resigned (Hafner 2006). As this example shows, the benefits for specific individuals inside of companies to facilitate data access are pretty small, and the worst-case scenario is terrible.

Researchers can, however, sometimes gain access to data that is inaccessible to the general public. Some governments have procedures that researchers can follow to apply for access, and, as the examples later in this chapter show, researchers can occasionally gain access to corporate data. For example, Einav et al. (2015) partnered with a researcher at eBay to study online auctions. I'll talk more about the research that came from this collaboration later in the chapter, but I mention it now because it had all four of the ingredients that I see in successful partnerships: researcher interest, researcher capability, company interest, and company capability. I've seen many potential collaborations fail because either the researcher or the partner—be it a company or government—lacked one of these ingredients.

Even if you are able to develop a partnership with a business or to gain access to restricted government data, however, there are some downsides for you. First, you will probably not be able to share your data with other researchers, which means that other researchers will not be able to verify and extend your results. Second, the questions that you can ask may be limited; companies are unlikely to allow research that could make them look bad. Finally, these partnerships can create at least the appearance of a conflict of

interest, where people might think that your results were influenced by your partnerships. All of these downsides can be addressed, but it is important to be clear that working with data that is not accessible to everyone has both upsides and downsides.

In summary, lots of big data are inaccessible to researchers. There are serious legal, business, and ethical barriers that prevent data access, and these barriers will not go away as technology improves, because they are not technical barriers. Some national governments have established procedures for enabling data access for some datasets, but the process is especially ad hoc at the state and local levels. Also, in some cases, researchers can partner with companies to obtain data access, but this can create a variety of problems for researchers and companies.

2.3.6 Nonrepresentative

> Nonrepresentative data are bad for out-of-sample generalizations, but can be quite useful for within-sample comparisons.

Some social scientists are accustomed to working with data that comes from a probabilistic random sample from a well-defined population, such as all adults in a particular country. This kind of data is called *representative* data because the sample "represents" the larger population. Many researchers prize representative data, and, to some, representative data is synonymous with rigorous science whereas nonrepresentative data is synonymous with sloppiness. At the most extreme, some skeptics seem to believe that nothing can be learned from nonrepresentative data. If true, this would seem to severely limit what can be learned from big data sources because many of them are nonrepresentative. Fortunately, these skeptics are only partially right. There are certain research goals for which nonrepresentative data is clearly not well suited, but there are others for which it might actually be quite useful.

To understand this distinction, let's consider a scientific classic: John Snow's study of the 1853–54 cholera outbreak in London. At the time, many doctors believed that cholera was caused by "bad air," but Snow believed that it was an infectious disease, perhaps spread by sewage-laced drinking water. To test this idea, Snow took advantage of what we might now call a natural experiment. He compared the cholera rates of households served by

two different water companies: Lambeth and Southwark & Vauxhall. These companies served similar households, but they differed in one important way: in 1849—a few years before the epidemic began—Lambeth moved its intake point upstream from the main sewage discharge in London, whereas Southwark & Vauxhall left their intake pipe downstream from the sewage discharge. When Snow compared the death rates from cholera in households served by the two companies, he found that customers of Southwark & Vauxhall—the company that was providing customers sewage-tainted water—were 10 times more likely to die from cholera. This result provides strong scientific evidence for Snow's argument about the cause of cholera, even though it is not based on a representative sample of people in London.

The data from these two companies, however, would not be ideal for answering a different question: what was the prevalence of cholera in London during the outbreak? For that second question, which is also important, it would be much better to have a representative sample of people from London.

As Snow's work illustrates, there are some scientific questions for which nonrepresentative data can be quite effective, and there are others for which it is not well suited. One crude way to distinguish these two kinds of questions is that some questions are about within-sample comparisons and some are about out-of-sample generalizations. This distinction can be further illustrated by another classic study in epidemiology: the British Doctors Study, which played an important role in demonstrating that smoking causes cancer. In this study, Richard Doll and A. Bradford Hill followed approximately 25,000 male doctors for several years and compared their death rates based on the amount that they smoked when the study began. Doll and Hill (1954) found a strong exposure–response relationship: the more heavily people smoked, the more likely they were to die from lung cancer. Of course, it would be unwise to estimate the prevalence of lung cancer among all British people based on this group of male doctors, but the within-sample comparison still provides evidence that smoking causes lung cancer.

Now that I've illustrated the difference between within-sample comparisons and out-of-sample generalizations, two caveats are in order. First, there are naturally questions about the extent to which a relationship that holds within a sample of male British doctors will also hold within a sample of

female British doctors or male British factory workers or female German factory workers or many other groups. These questions are interesting and important, but they are different from questions about the extent to which we can generalize from a sample to a population. Notice, for example, that you probably suspect that the relationship between smoking and cancer that was found in male British doctors will probably be similar in these other groups. Your ability to do this extrapolation does not come from the fact that male British doctors are a probabilistic random sample from any population; rather, it comes from an understanding of the mechanism that links smoking and cancer. Thus, the generalization from a sample to the population from which it is drawn is a largely a statistical issue, but questions about the *transportability* of pattern found in one group to another group is largely a nonstatistical issue (Pearl and Bareinboim 2014; Pearl 2015).

At this point, a skeptic might point out that most social patterns are probably less transportable across groups than the relationship between smoking and cancer. And I agree. The extent to which we should expect patterns to be transportable is ultimately a scientific question that has to be decided based on theory and evidence. It should not automatically be assumed that patterns will be transportable, but nor should be it assumed that they won't be transportable. These somewhat abstract questions about transportability will be familiar to you if you have followed the debates about how much researchers can learn about human behavior by studying undergraduate students (Sears 1986, Henrich, Heine, and Norenzayan (2010b)). Despite these debates, however, it would be unreasonable to say that researchers can't learn anything from studying undergraduate students.

The second caveat is that most researchers with nonrepresentative data are not as careful as Snow or Doll and Hill. So, to illustrate what can go wrong when researchers try to make an out-of-sample generalization from nonrepresentative data, I'd like to tell you about a study of the 2009 German parliamentary election by Andranik Tumasjan and colleagues (2010). By analyzing more than 100,000 tweets, they found that the proportion of tweets mentioning a political party matched the proportion of votes that party received in the parliamentary election (figure 2.3). In other words, it appeared that Twitter data, which was essentially free, could replace traditional public opinion surveys, which are expensive because of their emphasis on representative data.

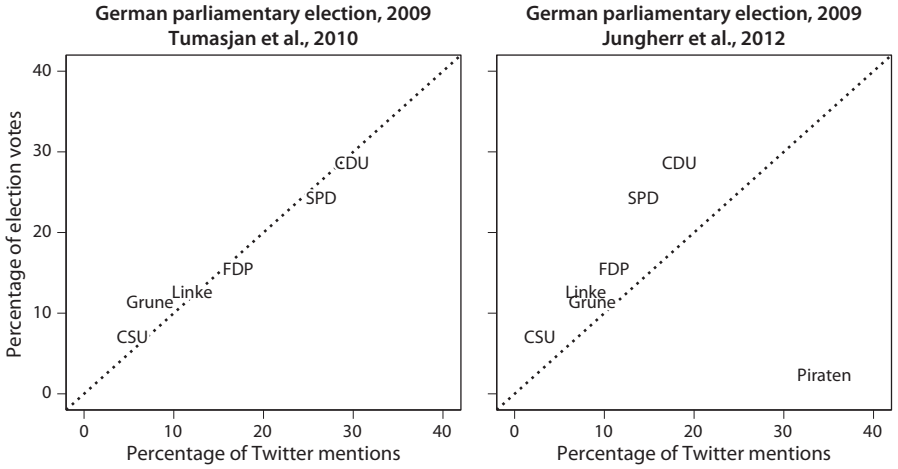

Figure 2.3: Twitter mentions appear to predict the results of the 2009 German election (Tumasjan et al. 2010), but this result depends on excluding the party with the most mentions: the Pirate Party (Jungherr, Jürgens, and Schoen 2012). See Tumasjan et al. (2012) for an argument in favor of excluding the Pirate Party. Adapted from Tumasjan et al. (2010), table 4, and Jungherr, Jürgens, and Schoen (2012), table 2.

Given what you probably already know about Twitter, you should immediately be skeptical of this result. Germans on Twitter in 2009 were not a probabilistic random sample of German voters, and supporters of some parties might tweet about politics much more often than supporters of other parties. Thus, it seems surprising that all of the possible biases that you could imagine would somehow cancel out so that this data would be directly reflective of German voters. In fact, the results in Tumasjan et al. (2010) turned out to be too good to be true. A follow-up paper by Andreas Jungherr, Pascal Jürgens, and Harald Schoen (2012) pointed out that the original analysis had excluded the political party that had received the most mentions on Twitter: the Pirate Party, a small party that fights government regulation of the Internet. When the Pirate Party was included in the analysis, Twitter mentions becomes a terrible predictor of election results (figure 2.3). As this example illustrates, using nonrepresentative big data sources to do out-of-sample generalizations can go very wrong. Also, you should notice that the fact that there were 100,000 tweets was basically irrelevant: lots of nonrepresentative data is still nonrepresentative, a theme that I'll return to in chapter 3 when I discuss surveys.

To conclude, many big data sources are not representative samples from some well-defined population. For questions that require generalizing results from the sample to the population from which it was drawn, this is a serious problem. But for questions about within-sample comparisons, nonrepresentative data can be powerful, so long as researchers are clear about the characteristics of their sample and support claims about transportability with theoretical or empirical evidence. In fact, my hope is that big data sources will enable researchers to make more within-sample comparisons in many nonrepresentative groups, and my guess is that estimates from many different groups will do more to advance social research than a single estimate from a probabilistic random sample.

2.3.7 Drifting

> Population drift, usage drift, and system drift make it hard to use big data sources to study long-term trends.

One of the great advantages of many big data sources is that they collect data over time. Social scientists call this kind of over-time data *longitudinal data*. And, naturally, longitudinal data are very important for studying change. In order to reliably measure change, however, the measurement system itself must be stable. In the words of sociologist Otis Dudley Duncan, "if you want to measure change, don't change the measure" (Fischer 2011).

Unfortunately, many big data systems—especially business systems—are changing all the time, a process that I'll call *drift*. In particular, these systems change in three main ways: *population drift* (change in who is using them), *behavioral drift* (change in how people are using them), and *system drift* (change in the system itself). The three sources of drift mean that any pattern in a big data source could be caused by an important change in the world, or it could be caused by some form of drift.

The first source of drift—population drift—is caused by changes in who is using the system, and these changes can happen on both short and long timescales. For example, during the US Presidential election of 2012 the proportion of tweets about politics that were written by women fluctuated from day to day (Diaz et al. 2016). Thus, what might appear to be a change in the mood of the Twitter-verse might actually just be a change in who is

talking at any moment. In addition to these short-term fluctuations, there has also been a long-term trend of certain demographic groups adopting and abandoning Twitter.

In addition to changes in who is using a system, there are also changes in how the system is used, which I call behavioral drift. For example, during the 2013 Occupy Gezi protests in Turkey, protesters changed their use of hashtags as the protest evolved. Here's how Zeynep Tufekci (2014) described the behavioral drift, which she was able to detect because she was observing behavior on Twitter and in person:

> "What had happened was that as soon as the protest became the dominant story, large numbers of people … stopped using the hashtags except to draw attention to a new phenomenon … While the protests continued, and even intensified, the hashtags died down. Interviews revealed two reasons for this. First, once everyone knew the topic, the hashtag was at once superfluous and wasteful on the character-limited Twitter platform. Second, hashtags were seen only as useful for attracting attention to a particular topic, not for talking about it."

Thus, researchers who were studying the protests by analyzing tweets with protest-related hashtags would have a distorted sense of what was happening because of this behavioral drift. For example, they might believe that the discussion of the protest decreased long before it actually decreased.

The third kind of drift is system drift. In this case, it is not the people changing or their behavior changing, but the system itself changing. For example, over time, Facebook has increased the limit on the length of status updates. Thus, any longitudinal study of status updates will be vulnerable to artifacts caused by this change. System drift is closely related to a problem called algorithmic confounding, which I'll cover in section 2.3.8.

To conclude, many big data sources are drifting because of changes in who is using them, in how they are being used, and in how the systems work. These sources of change are sometimes interesting research questions, but these changes complicate the ability of big data sources to track long-term changes over time.

2.3.8 Algorithmically confounded

> Behavior in big data systems is not natural; it is driven by the engineering goals of the systems.

Although many big data sources are nonreactive because people are not aware their data are being recorded (section 2.3.3), researchers should not consider behavior in these online systems to be "naturally occurring." In reality, the digital systems that record behavior are highly engineered to induce specific behaviors such as clicking on ads or posting content. The ways that the goals of system designers can introduce patterns into data is called *algorithmic confounding*. Algorithmic confounding is relatively unknown to social scientists, but it is a major concern among careful data scientists. And, unlike some of the other problems with digital traces, algorithmic confounding is largely invisible.

A relatively simple example of algorithmic confounding is the fact that on Facebook there are an anomalously high number of users with approximately 20 friends, as was discovered by Johan Ugander and colleagues (2011). Scientists analyzing this data without any understanding of how Facebook works could doubtless generate many stories about how 20 is some kind of magical social number. Fortunately, Ugander and his colleagues had a substantial understanding of the process that generated the data, and they knew that Facebook encouraged people with few connections on Facebook to make more friends until they reached 20 friends. Although Ugander and colleagues don't say this in their paper, this policy was presumably created by Facebook in order to encourage new users to become more active. Without knowing about the existence of this policy, however, it is easy to draw the wrong conclusion from the data. In other words, the surprisingly high number of people with about 20 friends tells us more about Facebook than about human behavior.

In this previous example, algorithmic confounding produced a quirky result that a careful researcher might detect and investigate further. However, there is an even trickier version of algorithmic confounding that occurs when designers of online systems are aware of social theories and then bake these theories into the working of their systems. Social scientists call this *performativity*: when a theory changes the world in such a way that it bring the world more into line with the theory. In the case of

performative algorithmic confounding, the confounded nature of the data is likely invisible.

One example of a pattern created by performativity is transitivity in online social networks. In the 1970s and 1980s, researchers repeatedly found that if you are friends with both Alice and Bob, then Alice and Bob are more likely to be friends with each other than if they were two randomly chosen people. This very same pattern was found in the social graph on Facebook (Ugander et al. 2011). Thus, one might conclude that patterns of friendship on Facebook replicate patterns of offline friendships, at least in terms of transitivity. However, the magnitude of transitivity in the Facebook social graph is partially driven by algorithmic confounding. That is, data scientists at Facebook knew of the empirical and theoretical research about transitivity and then baked it into how Facebook works. Facebook has a "People You May Know" feature that suggests new friends, and one way that Facebook decides who to suggest to you is transitivity. That is, Facebook is more likely to suggest that you become friends with the friends of your friends. This feature thus has the effect of increasing transitivity in the Facebook social graph; in other words, the theory of transitivity brings the world into line with the predictions of the theory (Zignani et al. 2014; Healy 2015). Thus, when big data sources appear to reproduce predictions of social theory, we must be sure that the theory itself was not baked into how the system worked.

Rather than thinking of big data sources as observing people in a natural setting, a more apt metaphor is observing people in a casino. Casinos are highly engineered environments designed to induce certain behaviors, and a researcher would never expect behavior in a casino to provide an unfettered window into human behavior. Of course, you could learn something about human behavior by studying people in casinos, but if you ignored the fact that the data was being created in a casino, you might draw some bad conclusions.

Unfortunately, dealing with algorithmic confounding is particularly difficult because many features of online systems are proprietary, poorly documented, and constantly changing. For example, as I'll explain later in this chapter, algorithmic confounding was one possible explanation for the gradual breakdown of Google Flu Trends (section 2.4.2), but this claim was hard to assess because the inner workings of Google's search algorithm are proprietary. The dynamic nature of algorithmic confounding is one form of system drift. Algorithmic confounding means that we should be cautious

about any claim regarding human behavior that comes from a single digital system, no matter how big.

2.3.9 Dirty

Big data sources can be loaded with junk and spam.

Some researchers believe that big data sources, especially online sources, are pristine because they are collected automatically. In fact, people who have worked with big data sources know that they are frequently *dirty*. That is, they frequently include data that do not reflect real actions of interest to researchers. Most social scientists are already familiar with the process of cleaning large-scale social survey data, but cleaning big data sources seems to be more difficult. I think the ultimate source of this difficulty is that many of these big data sources were never intended to be used for research, and so they are not collected, stored, and documented in a way that facilitates data cleaning.

The dangers of dirty digital trace data are illustrated by Back and colleagues' (2010) study of the emotional response to the attacks of September 11, 2001, which I briefly mentioned earlier in the chapter. Researchers typically study the response to tragic events using retrospective data collected over months or even years. But, Back and colleagues found an always-on source of digital traces—the timestamped, automatically recorded messages from 85,000 American pagers—and this enabled them to study emotional response on a much finer timescale. They created a minute-by-minute emotional timeline of September 11 by coding the emotional content of the pager messages by the percentage of words related to (1) sadness (e.g., "crying" and "grief"), (2) anxiety (e.g., "worried" and "fearful"), and (3) anger (e.g., "hate" and "critical"). They found that sadness and anxiety fluctuated throughout the day without a strong pattern, but that there was a striking increase in anger throughout the day. This research seems to be a wonderful illustration of the power of always-on data sources: if traditional data sources had been used, it would have been impossible to obtain such a high-resolution timeline of the immediate response to an unexpected event.

Just one year later, however, Cynthia Pury (2011) looked at the data more carefully. She discovered that a large number of the supposedly angry messages were generated by a single pager and they were all identical. Here's

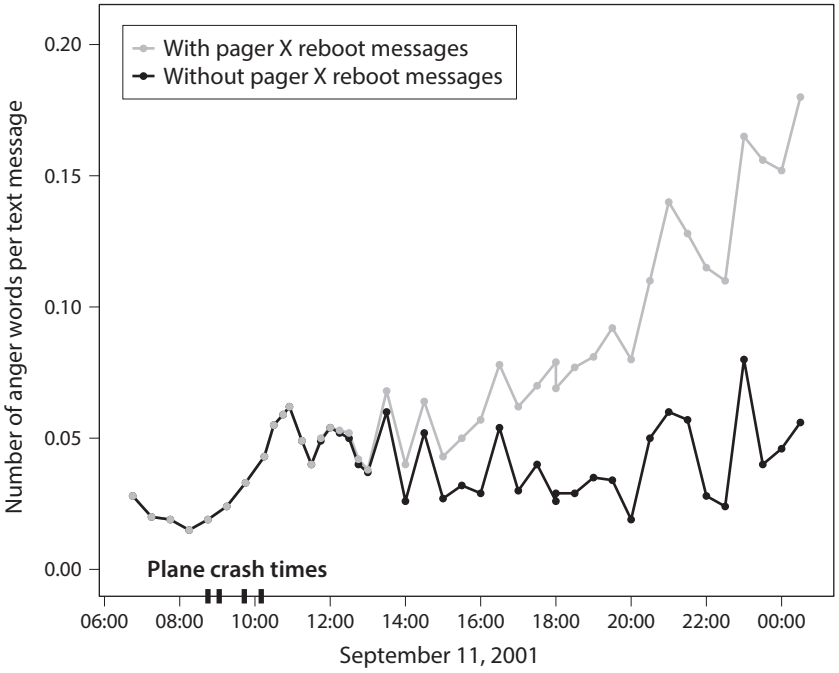

Figure 2.4: Estimated trends in anger over the course of September 11, 2001 based on 85,000 American pagers (Back, Küfner, and Egloff 2010; Pury 2011; Back, Küfner, and Egloff 2011). Originally, Back, Küfner, and Egloff (2010) reported a pattern of increasing anger throughout the day. However, most of these apparently angry messages were generated by a single pager that repeatedly sent out the following message: "Reboot NT machine [name] in cabinet [name] at [location]:CRITICAL:[date and time]". With this message removed, the apparent increase in anger disappears (Pury 2011; Back, Küfner, and Egloff 2011). Adapted from Pury (2011), figure 1b.

what those supposedly angry messages said:

> "Reboot NT machine [name] in cabinet [name] at [location]: CRITICAL:[date and time]"

These messages were labeled angry because they included the word "CRITICAL," which may generally indicate anger but in this case does not. Removing the messages generated by this single automated pager completely eliminates the apparent increase in anger over the course of the day (figure 2.4). In other words, the main result in Back, Küfner, and Egloff (2010) was an artifact of one pager. As this example illustrates, relatively

simple analysis of relatively complex and messy data has the potential to go seriously wrong.

While dirty data that is created unintentionally—such as that from one noisy pager—can be detected by a reasonably careful researcher, there are also some online systems that attract intentional spammers. These spammers actively generate fake data, and—often motivated by profit—work very hard to keep their spamming concealed. For example, political activity on Twitter seems to include at least some reasonably sophisticated spam, whereby some political causes are intentionally made to look more popular than they actually are (Ratkiewicz et al. 2011). Unfortunately, removing this intentional spam can be quite difficult.

Of course what is considered dirty data can depend, in part, on the research question. For example, many edits to Wikipedia are created by automated bots (Geiger 2014). If you are interested in the ecology of Wikipedia, then these bot-created edits are important. But if you are interested in how humans contribute to Wikipedia, then the bot-created edits should be excluded.

There is no single statistical technique or approach that can ensure that you have sufficiently cleaned your dirty data. In the end, I think the best way to avoid being fooled by dirty data is to understand as much as possible about how your data were created.

2.3.10 Sensitive

> Some of the information that companies and governments have is sensitive.

Health insurance companies have detailed information about the medical care received by their customers. This information could be used for important research about health, but if it became public, it could potentially lead to emotional harm (e.g., embarrassment) or economic harm (e.g., loss of employment). Many other big data sources also have information that is *sensitive*, which is part of the reason why they are often inaccessible.

Unfortunately, it turns out to be quite tricky to decide what information is actually sensitive (Ohm, 2015), as was illustrated by the Netflix Prize. As I will describe in chapter 5, in 2006, Netflix released 100 million movie ratings provided by almost 500,000 members and had an open call where

people from all over the world submitted algorithms that could improve Netflix's ability to recommend movies. Before releasing the data, Netflix removed any obvious personally identifying information, such as names. But, just two weeks after the data was released Arvind Narayanan and Vitaly Shmatikov (2008) showed that it was possible to learn about specific people's movie ratings using a trick that I'll show you in chapter 6. Even though an attacker could discover a person's movie ratings, there still doesn't seem to be anything sensitive here. While that might be true in general, for at least some of the 500,000 people in the dataset, movie ratings were sensitive. In fact, in response to the release and re-identification of the data, a closeted lesbian woman joined a class-action suit against Netflix. Here's how the problem was expressed in this lawsuit (Singel 2009):

> "[M]ovie and rating data contains information of a … highly personal and sensitive nature. The member's movie data exposes a Netflix member's personal interest and/or struggles with various highly personal issues, including sexuality, mental illness, recovery from alcoholism, and victimization from incest, physical abuse, domestic violence, adultery, and rape."

This example shows that there can be information that some people consider sensitive inside of what might appear to be a benign database. Further, it shows that a main defense that researchers employ to protect sensitive data—de-identification—can fail in surprising ways. These two ideas are developed in greater detail in chapter 6.

The final thing to keep in mind about sensitive data is that collecting it without people's consent raises ethical questions, even if no specific harm is caused. Much like watching someone taking a shower without their consent might be considered a violation of that person's privacy, collecting sensitive information—and remember how hard it can be to decide what is sensitive—without consent creates potential privacy concerns. I'll return to questions about privacy in chapter 6.

In conclusion, big data sources, such as government and business administrative records, are generally not created for the purpose of social research. The big data sources of today, and likely tomorrow, tend to have 10 characteristics. Many of the properties that are generally considered to be good for research—big, always-on, and nonreactive—come from the fact in the digital

age companies and governments are able to collect data at a scale that was not possible previously. And many of the properties that are generally considered to be bad for research—incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, inaccessible, dirty, and sensitive—come from the fact that these data were not collected by researchers for researchers. So far, I've talked about government and business data together, but there are some differences between the two. In my experience, government data tends to be less nonrepresentative, less algorithmically confounded, and less drifting. One the other hand, business administrative records tend to be more always-on. Understanding these 10 general characteristics is a helpful first step toward learning from big data sources. And now we turn to research strategies we can use with this data.

2.4 Research strategies

Given these 10 characteristics of big data sources and the inherent limitations of even perfectly observed data, I see three main strategies for learning from big data sources: counting things, forecasting things, and approximating experiments. I'll describe each of these approaches—which could be called "research strategies" or "research recipes"—and I'll illustrate them with examples. These strategies are neither mutually exclusive nor exhaustive.

2.4.1 Counting things

> Simple counting can be interesting if you combine a good question with good data.

Although it is couched in sophisticated-sounding language, lots of social research is really just counting things. In the age of big data, researchers can count more than ever, but that does not mean that they should just start counting haphazardly. Instead, researchers should ask: What things are worth counting? This may seem like an entirely subjective matter, but there are some general patterns.

Often students motivate their counting research by saying: I'm going to count something that no-one has ever counted before. For example, a student might say that many people have studied migrants and many people have studied twins, but nobody has studied migrant twins. In my experience, this