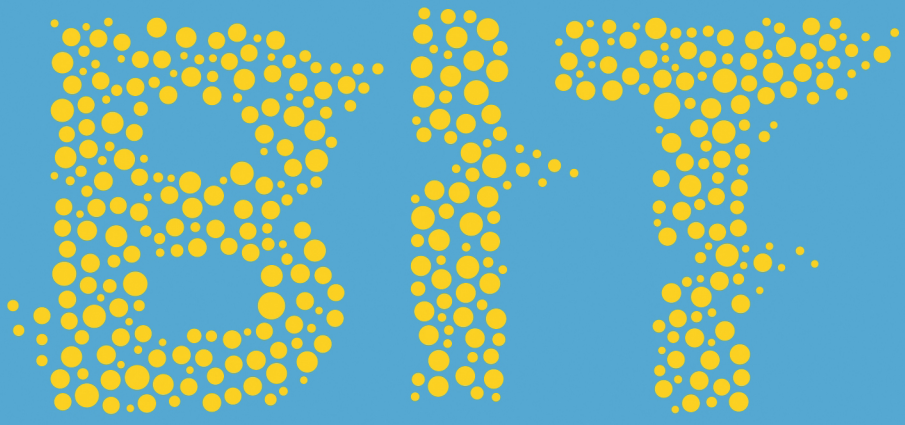
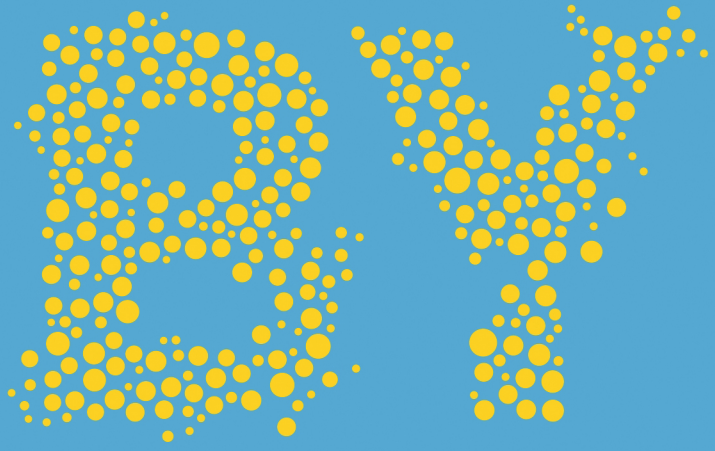


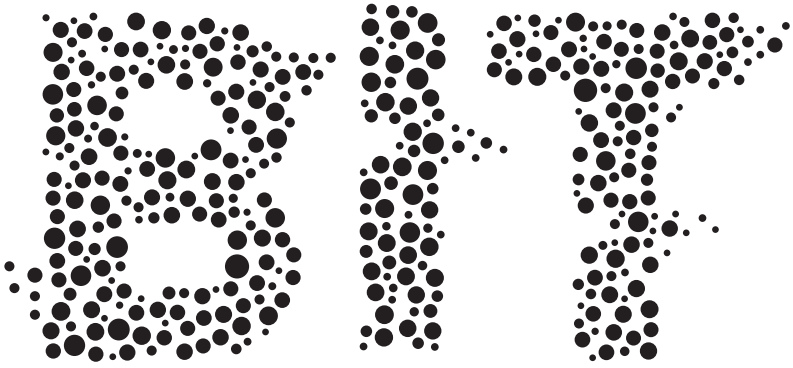
SOCIAL RESEARCH



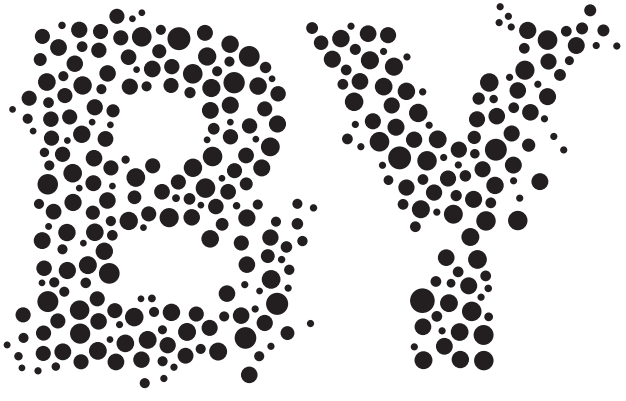
in the DIGITAL AGE

• MATTHEW J. SALGANIK •

BIT BY BIT



SOCIAL RESEARCH



in the DIGITAL AGE

• MATTHEW J. SALGANIK •

PRINCETON UNIVERSITY PRESS
PRINCETON & OXFORD

Published by Princeton University Press,
41 William Street, Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
6 Oxford Street, Woodstock, Oxfordshire OX20 1TR

press.princeton.edu

Cover design by Amanda Weiss

First paperback printing, 2019
Paperback ISBN 978-0-691-19610-7
Cloth ISBN 978-0-691-15864-8

Library of Congress Control Number: 2017935851

British Library Cataloging-in-Publication Data is available

This book has been composed in Minion Pro and DIN Pro

Printed on acid-free paper. ∞

Typeset by Nova Techset Pvt Ltd, Bangalore, India
Printed in the United States of America

To Amanda

SUMMARY OF CONTENTS

CONTENTS

PREFACE

This book began in 2005 in a basement at Columbia University. At the time, I was a graduate student, and I was running an online experiment that would eventually become my dissertation. I'll tell you all about the scientific parts of that experiment in chapter 4, but now I'm going to tell you about something that's not in my dissertation or in any of my papers. And it's something that fundamentally changed how I think about research. One morning, when I came into my basement office, I discovered that overnight about 100 people from Brazil had participated in my experiment. This simple experience had a profound effect on me. At that time, I had friends who were running traditional lab experiments, and I knew how hard they had to work to recruit, supervise, and pay people to participate in these experiments; if they could run 10 people in a single day, that was good progress. However, with my online experiment, 100 people participated *while I was sleeping*. Doing your research while you are sleeping might sound too good to be true, but it isn't. Changes in technology—specifically the transition from the analog age to the digital age—mean that we can now collect and analyze social data in new ways. This book is about doing social research in these new ways.

This book is for social scientists who want to do more data science, data scientists who want to do more social science, and anyone interested in the hybrid of these two fields. Given who this book is for, it should go without saying that it is not just for students and professors. Although I currently work at a university (Princeton), I've also worked in government (at the US Census Bureau) and in the tech industry (at Microsoft Research), so I know that there is a lot of exciting research happening outside of universities. So if you think of what you are doing as social research, then this book is for you, no matter where you work or what kind of techniques you currently use.

As you might have noticed already, the tone of this book is a bit different from that of many other academic books. That's intentional. This book emerged from a graduate seminar on computational social science that I have taught at Princeton in the Department of Sociology since 2007, and

I'd like it to capture some of the energy and excitement from that seminar. In particular, I want this book to have three characteristics: I want it to be helpful, future-oriented, and optimistic.

Helpful: My goal is to write a book that is helpful for you. Therefore, I'm going to write in an open, informal, and example-driven style. That's because the most important thing that I want to convey is a certain way of thinking about social research. And my experience suggests that the best way to convey this way of thinking is informally and with lots of examples. Also, at the end of each chapter, I have a section called "What to read next" that will help you transition into more detailed and technical readings on many of the topics that I introduce. In the end, I hope this book will help you both do research and evaluate the research of others.

Future-oriented: I hope that this book will help you to do social research using the digital systems that exist today *and* those that will be created in the future. I started doing this kind of research in 2004, and since then I've seen many changes, and I'm sure that over the course of your career you will see many changes too. The trick to staying relevant in the face of change is *abstraction*. For example, this is not going to be a book that teaches you exactly how to use the Twitter API as it exists today; instead, it is going to teach you how to learn from big data sources (chapter 2). This is not going to be a book that gives you step-by-step instructions for running experiments on Amazon Mechanical Turk; instead, it is going to teach you how to design and interpret experiments that rely on digital age infrastructure (chapter 4). Through the use of abstraction, I hope this will be a timeless book on a timely topic.

Optimistic: The two communities that this book engages—social scientists and data scientists—have very different backgrounds and interests. In addition to these science-related differences, which I talk about in the book, I've also noticed that these two communities have different styles. Data scientists are generally excited; they tend to see the glass as half full. Social scientists, on the other hand, are generally more critical; they tend to see the glass as half empty. In this book, I'm going to adopt the optimistic tone of a data scientist. So, when I present examples, I'm going to tell you what I love about these examples. And when I do point out problems with the examples—and I will do that because no research is perfect—I'm going to try to point out these problems in a way that is positive and optimistic. I'm not going to be critical for the sake of being critical—I'm going to be critical so that I can help you create better research.

We are still in the early days of social research in the digital age, but I've seen some misunderstandings that are so common that it makes sense for me to address them here, in the preface. From data scientists, I've seen two common misunderstandings. The first is thinking that more data automatically solves problems. However, for social research, that has not been my experience. In fact, for social research, better data—as opposed to more data—seems to be more helpful. The second misunderstanding that I've seen from data scientists is thinking that social science is just a bunch of fancy talk wrapped around common sense. Of course, as a social scientist—more specifically as a sociologist—I don't agree with that. Smart people have been working hard to understand human behavior for a long time, and it seems unwise to ignore the wisdom that has accumulated from this effort. My hope is that this book will offer you some of that wisdom in a way that is easy to understand.

From social scientists, I've also seen two common misunderstandings. First, I've seen some people write off the entire idea of social research using the tools of the digital age because of a few bad papers. If you're reading this book, you've probably already read a bunch of papers that use social media data in ways that are banal or wrong (or both). I have too. However, it would be a serious mistake to conclude from these examples that all digital-age social research is bad. In fact, you've probably also read a bunch of papers that use survey data in ways that are banal or wrong, but you don't write off all research using surveys. That's because you know that there is great research done with survey data, and in this book I'm going to show you that there is also great research done with the tools of the digital age.

The second common misunderstanding that I've seen from social scientists is to confuse the present with the future. When we assess social research in the digital age—the research that I'm going to describe—it's important that we ask two distinct questions: "How well does this style of research work right now?" and "How well will this style of research work in the future?" Researchers are trained to answer the first question, but for this book I think the second question is more important. That is, even though social research in the digital age has not yet produced massive, paradigm-changing intellectual contributions, the rate of improvement in digital-age research is incredibly rapid. It is this rate of change—more than the current level—that makes digital-age research so exciting to me.

Even though that last paragraph might seem to offer you potential riches at some unspecified time in the future, my goal is not to sell you on any particular type of research. I don't personally own shares in Twitter, Facebook, Google, Microsoft, Apple, or any other tech company (although, for the sake of full disclosure, I should mention that I have worked at, or received research funding from, Microsoft, Google, and Facebook). Throughout the book, therefore, my goal is to remain a credible narrator, telling you about all the exciting new stuff that is possible, while guiding you away from a few traps that I've seen others fall into (and occasionally fallen into myself).

The intersection of social science and data science is sometimes called computational social science. Some consider this to be a technical field, but this will not be a technical book in the traditional sense. For example, there are no equations in the main text. I chose to write the book this way because I wanted to provide a comprehensive view of social research in the digital age, including big data sources, surveys, experiments, mass collaboration, and ethics. It turned out to be impossible to cover all these topics and provide technical details about each one. Instead, pointers to more technical material are given in the "What to read next" section at the end of each chapter. In other words, this book is not designed to teach you how to do any specific calculation; rather, it is designed to change the way that you think about social research.

How to use this book in a course

As I said earlier, this book emerged in part from a graduate seminar on computational social science that I've been teaching since 2007 at Princeton. Since you might be thinking about using this book to teach a course, I thought that it might be helpful for me to explain how it grew out of my course and how I imagine it being used in other courses.

For several years, I taught my course without a book; I'd just assign a collection of articles. While students were able to learn from these articles, the articles alone were not leading to the conceptual changes that I was hoping to create. So I would spend most of the time in class providing perspective, context, and advice in order to help the students see the big picture. This book is my attempt to write down all that perspective, context, and advice in a way that has no prerequisites—in terms of either social science or data science.

In a semester-long course, I would recommend pairing this book with a variety of additional readings. For example, such a course might spend two weeks on experiments, and you could pair chapter 4 with readings on topics such as the role of pre-treatment information in the design and analysis of experiments; statistical and computational issues raised by large-scale A/B tests at companies; design of experiments specifically focused on mechanisms; and practical, scientific, and ethical issues related to using participants from online labor markets, such as Amazon Mechanical Turk. It could also be paired with readings and activities related to programming. The appropriate choice between these many pairings depends on the students in your course (e.g., undergraduate, master's, or PhD), their backgrounds, and their goals.

A semester-length course could also include weekly problem sets. Each chapter has a variety of activities that are labeled by degree of difficulty: easy (), medium (), hard (), and very hard (). Also, I've labeled each problem by the skills that it requires: math (), coding (), and data collection (). Finally, I've labeled a few of the activities that are my personal favorites (). I hope that within this diverse collection of activities, you'll find some that are appropriate for your students.

In order to help people using this book in courses, I've started a collection of teaching materials such as syllabuses, slides, recommended pairings for each chapter, and solutions to some activities. You can find these materials—and contribute to them—at http://www.bitbybitbook.com.

BIT BY BIT

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

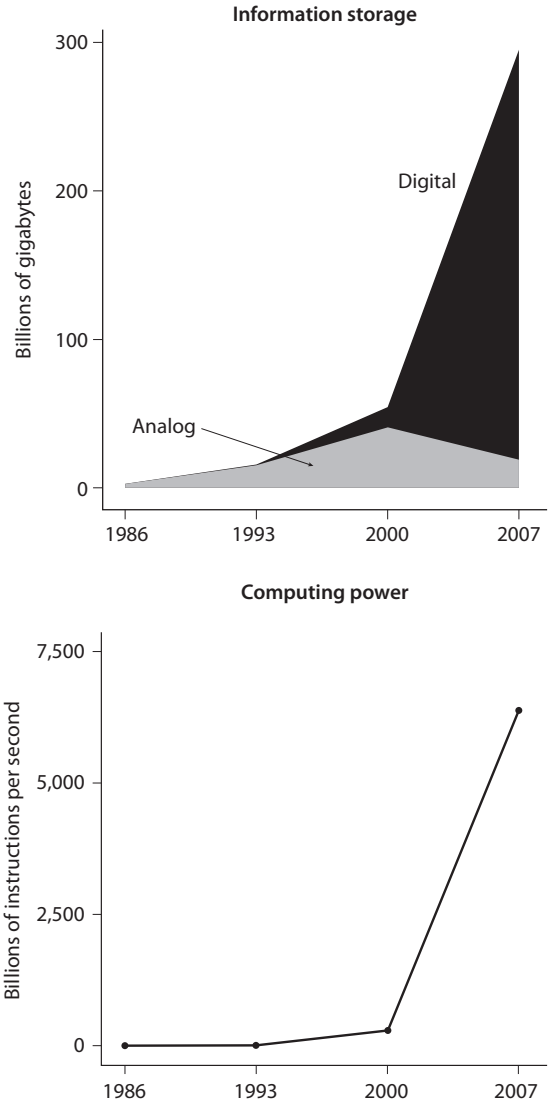


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

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.

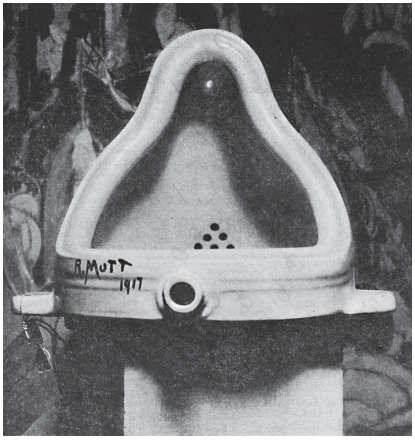
This book is written for two audiences that have a lot to learn from each other. On the one hand, it is for social scientists who have training and experience studying social behavior, but who are less familiar with the opportunities created by the digital age. On the other hand, it is for another group 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.

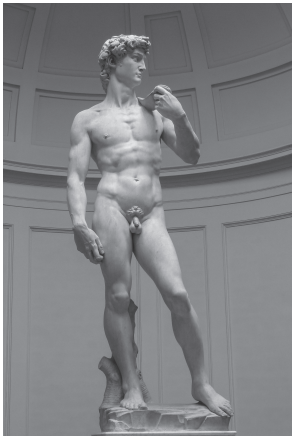
1.4 Themes of this book

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).

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 checkout 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

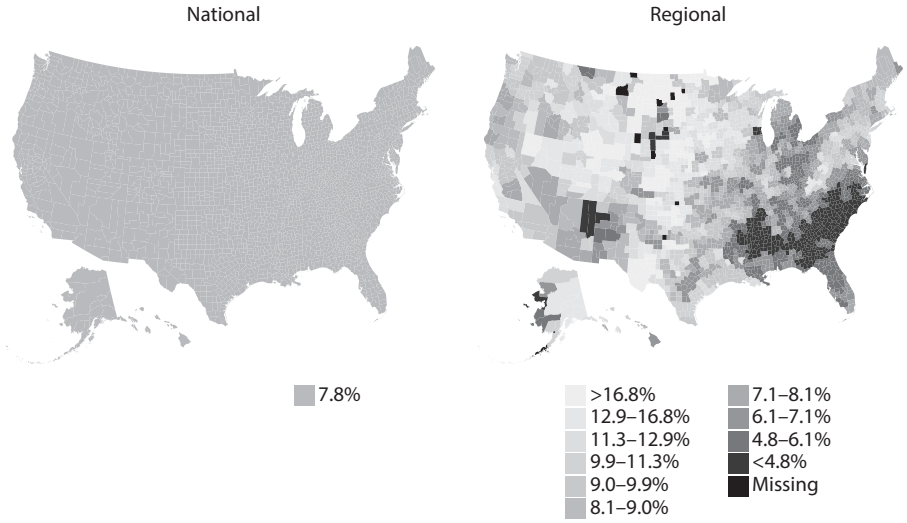


Figure 2.1: Estimates of a child's chances of reaching the top 20% of income distribution given parents in the bottom 20% (Chetty et al. 2014). The regional-level estimates, which show heterogeneity, naturally lead to interesting and important questions that do not arise from a single national-level 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

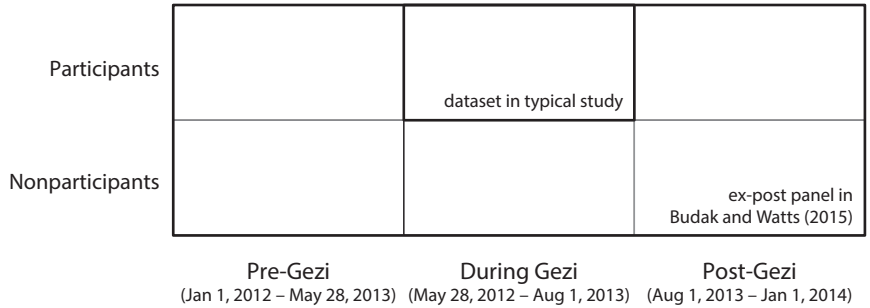


Figure 2.2: Design used by Budak and Watts (2015) to study the Occupy Gezi protests in Turkey in the summer of 2013. By using the always-on nature of Twitter, the researchers created what they called an *ex-post panel* that included about 30,000 people over two years. In contrast to a typical study that 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).