

CONTENTS

PREFACE	xv
CHAPTER 1 • INTRODUCTION	1
1.1 An ink blot.....	1
1.2 Welcome to the digital age	2
1.3 Research design	5
1.4 Themes of this book.....	6
1.5 Outline of this book.....	9
What to read next.....	11
CHAPTER 2 • OBSERVING BEHAVIOR	13
2.1 Introduction.....	13
2.2 Big data	14
2.3 Ten common characteristics of big data	17
2.3.1 Big	17
2.3.2 Always-on.....	21
2.3.3 Nonreactive	23
2.3.4 Incomplete.....	24
2.3.5 Inaccessible	27
2.3.6 Nonrepresentative	29
2.3.7 Drifting.....	33
2.3.8 Algorithmically confounded.....	35
2.3.9 Dirty	37
2.3.10 Sensitive	39
2.4 Research strategies.....	41
2.4.1 Counting things	41
2.4.2 Forecasting and nowcasting.....	46
2.4.3 Approximating experiments	50

2.5	Conclusion	61
	Mathematical notes	62
	What to read next	70
	Activities	77
CHAPTER 3 • ASKING QUESTIONS		85
3.1	Introduction.....	85
3.2	Asking versus observing	87
3.3	The total survey error framework	89
	3.3.1 Representation	91
	3.3.2 Measurement	94
	3.3.3 Cost	98
3.4	Who to ask	99
3.5	New ways of asking questions	107
	3.5.1 Ecological momentary assessments	108
	3.5.2 Wiki surveys	111
	3.5.3 Gamification	115
3.6	Surveys linked to big data sources	117
	3.6.1 Enriched asking	118
	3.6.2 Amplified asking	122
3.7	Conclusion	130
	Mathematical notes	130
	What to read next	136
	Activities	141
CHAPTER 4 • RUNNING EXPERIMENTS		147
4.1	Introduction.....	147
4.2	What are experiments?.....	149
4.3	Two dimensions of experiments: lab–field and analog–digital	151
4.4	Moving beyond simple experiments	158
	4.4.1 Validity	161
	4.4.2 Heterogeneity of treatment effects	167
	4.4.3 Mechanisms	169
4.5	Making it happen	174
	4.5.1 Use existing environments	175
	4.5.2 Build your own experiment	178

4.5.3	Build your own product	182
4.5.4	Partner with the powerful	183
4.6	Advice	188
4.6.1	Create zero variable cost data	190
4.6.2	Build ethics into your design: replace, refine, and reduce	196
4.7	Conclusion	202
	Mathematical notes	203
	What to read next	209
	Activities	220
 CHAPTER 5 • CREATING MASS COLLABORATION		 231
5.1	Introduction.....	231
5.2	Human computation	233
5.2.1	Galaxy Zoo	234
5.2.2	Crowd-coding of political manifestos	241
5.2.3	Conclusion	244
5.3	Open calls	246
5.3.1	Netflix Prize	246
5.3.2	Foldit	249
5.3.3	Peer-to-Patent	252
5.3.4	Conclusion	254
5.4	Distributed data collection	256
5.4.1	eBird	257
5.4.2	PhotoCity.....	259
5.4.3	Conclusion	262
5.5	Designing your own	265
5.5.1	Motivate participants	265
5.5.2	Leverage heterogeneity.....	266
5.5.3	Focus attention	267
5.5.4	Enable surprise	267
5.5.5	Be ethical	268
5.5.6	Final design advice	269
5.6	Conclusion	271
	What to read next	272
	Activities	277

CHAPTER 6 • ETHICS	281
6.1 Introduction.....	281
6.2 Three examples	283
6.2.1 Emotional Contagion	284
6.2.2 Tastes, Ties, and Time.....	285
6.2.3 Encore	286
6.3 Digital is different	288
6.4 Four principles	294
6.4.1 Respect for Persons	295
6.4.2 Beneficence	296
6.4.3 Justice	298
6.4.4 Respect for Law and Public Interest	299
6.5 Two ethical frameworks	301
6.6 Areas of difficulty	303
6.6.1 Informed consent	303
6.6.2 Understanding and managing informational risk	307
6.6.3 Privacy	314
6.6.4 Making decisions in the face of uncertainty	317
6.7 Practical tips	321
6.7.1 The IRB is a floor, not a ceiling	321
6.7.2 Put yourself in everyone else's shoes	322
6.7.3 Think of research ethics as continuous, not discrete.....	324
6.8 Conclusion	324
Historical appendix.....	325
What to read next	331
Activities	338
CHAPTER 7 • THE FUTURE	355
7.1 Looking forward	355
7.2 Themes of the future.....	355
7.2.1 The blending of readymades and custommades.....	355
7.2.2 Participant-centered data collection	356
7.2.3 Ethics in research design	357
7.3 Back to the beginning.....	358

ACKNOWLEDGMENTS	361
REFERENCES	367
INDEX	413

CHAPTER 1

INTRODUCTION

1.1 An ink blot

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

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

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

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

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

1.2 Welcome to the digital age

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

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

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

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

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

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

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

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

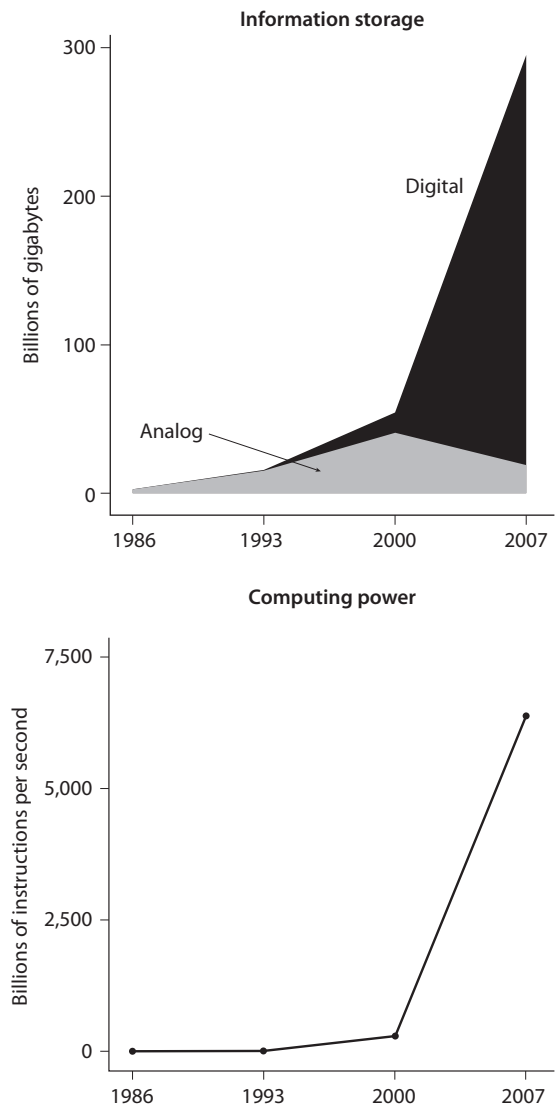


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

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

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

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

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

1.3 Research design

Research design is about connecting questions and answers.

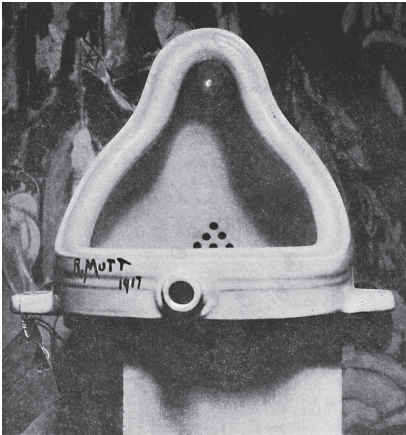
This book is written for two audiences that have a lot to learn from each other. On the one hand, it is for social scientists who have training and experience studying social behavior, but who are less familiar with the opportunities created by the digital age. On the other hand, it is for another group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

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

1.4 Themes of this book

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

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



Readymade



Custommade

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

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

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

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

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

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

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

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

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

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

1.5 Outline of this book

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

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

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

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

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

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

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

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

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

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

What to read next

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

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

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

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

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

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

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

- **Research design (section 1.3)**

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

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

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

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

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

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

INDEX

- Abelson, Hal, 11
- A/B tests, 185
- administrative records, 82–83; data linkage for, 140–41
- advertising, 186–87; return on investment tied to, 226–27
- Affluent Worker Project, 82
- African Americans: in Tuskegee Syphilis Study, 294, 325–28; *See also* race
- Agarwal, Sameer, 259
- AIDS and HIV, 262–63, 269
- Ai Weiwei, 44
- algorithmic confounding, 24, 34–36, 74; in Google Flu Trends, 49
- Allcott, Hunt, 164–68, 170–71, 212
- always-on data, 21–22, 71; in natural experiments, 53; pre-treatment information in, 157
- Amazon Mechanical Turk. *See* MTurk
- American Association of Public Opinion Research (AAPOR), 136
- amplified asking, 122–30, 141
- analog data, 2–3
- analog experiments, 152, 154–57, 210–11; costs of, 190
- ANCOVA (Analysis of Covariance), 209
- Anderson, Ashton, 74
- Anderson, Margo, 290
- Angrist, Joshua, 51, 53, 62–63
- animals, ethical issues in experiments using, 196–97
- anonymization of data, 8, 12, 28; to manage informational risks, 307–12, 336–37; of Netflix movie ratings, 39–40, 269
- Ansolabehere, Stephen, 119–21, 140, 356
- Antoun, Christopher, 144
- AOL (firm), 28, 72
- armada strategy, 189–90, 216
- asking questions, observing versus, 87–89, 137
- astronomy. *See* Galaxy Zoo
- auction behavior, 55–60
- Aurisset, Juliette, 201
- autonomy of people, 295
- auxiliary information: in non-probability sampling, 136; for nonresponse problem, 135; in stratified sampling, 132–33
- Back, Mitja D., 37, 38
- Bafumi, Joseph, 139
- Baker, David, 250, 270
- balance checks, 212
- Bamford, James, 27
- Banerji, Manda, 238, 239
- Banksy, 343–45
- Baracas, Solon, 11
- barnstars (Wikipedia awards), 150
- bathrooms, privacy in, 340–41
- Beauchamp, Tom L., 333
- behavior: ethical issues in observation of, 288; observational data on, 13; reported, 137
- behavioral drift, 34
- Belmont Report, 294–95, 301, 331; on assessment of risks and benefits, 318; on Beneficence, 296; Common Rule compared with, 329; history of, 328; on Justice, 298–99; on Respect for Persons, 295, 333
- Benard, Stephen, 153–54
- Beneficence, 296–98, 302, 317, 334
- Bengtsson, Linus, 83
- Benoit, Kenneth, 241–44, 278, 356
- Bentham, Jeremy, 288, 302
- Berges, Aida, 270
- Berinsky, Adam J., 77, 214
- Bernedo, María, 223
- Bethlehem, Jelke, 135, 136
- between-subjects designs, 160–61
- biases, 89; coverage bias, 93; in human computation mass collaborations, 237; nonresponse bias, 93, 135; social desirability bias, 108; trade-off between variance and, 138
- big data, 3, 60–62, 71; algorithmic confounded data in, 34–36; always-on characteristic of, 21–22; characteristics of, 17; definitions of, 13,

- 14, 70; dirty data in, 36–39; drifting of, 33–34;
- inaccessibility of, 27–29; incompleteness of, 24–27; nonreactivity of, 23–24;
- nonrepresentativeness of, 29–33; nowcasting using, 46–50; purposes and scale of, 17–21;
- repurposing of, 14–17; sensitive data in, 39–41;
- surveys linked to, 117–30, 140–41
- biology, protein folding in, 249–51
- bird data, eBird for, 257–59
- Bitcoin, 306
- Bjerrekaer, Julius, 350–51
- blacks: in Tuskegee Syphilis Study, 294, 325–28;
 See also race
- blocked experimental designs (stratified
 experimental designs), 218–19
- Blumenstock, Joshua E.: amplified asking used by, 122–30; machine learning model used by, 145;
- mixed research design used by, 5, 8, 332;
- Rwanda telephone study by, 1–2, 358–59
- Boellstorff, Tom, 70
- bogus data, in mass collaborations, 236–37
- Bond, Robert M., 208, 216, 217
- boomerang effects, 159, 160, 168
- Bradburn, Norman, 94
- Brandeis, Louis D., 337
- Bravo-Lillo, Cristian, 320
- Brexit, 144–45
- Brick, J. Michael, 137
- British Doctors Study, 30
- British Election Study (BES), 144–45
- broad consent, 336
- Bruine de Bruin, Wändi, 223
- Buckley, Sue, 339
- Budak, Ceren, 21
- Burch, Rex Leonard, 196–97, 217
- Burke, Moira, 88, 118–19
- Burnett, Sam, 286
- business. *See* corporations
- Buxtun, Robert, 327
- Cadamuro, Gabriel, 125, 358–59
- call detail records (CDRs), 146
- Camerer, Colin, 74
- Campbell, Donald T., 211
- cancer: ethical issues in research into, 345–47;
- smoking and, 30–31
- Canfield, Casey, 223
- Cardamone, Carolin, 268
- Carter, Stephen L., 322
- casinos, 36, 224
- Castronova, Edward, 215, 270
- causality, 209; fundamental problem of causal
 inference, 64, 204–5; making causal inferences
 from non-experimental approaches, 62
- cause-and-effect questions, 147–48
- cell phones. *See* mobile phones
- censorship of Internet, 43–46, 74–75; Encore
 study of, 286, 297–98; *See also* Encore
- Census Bureau, US, 70
- census data, 122–23; secondary uses of, 290
- Centers for Disease Control and Prevention, US
 (CDC), 46–49
- Centola, Damon, 181–82
- Chatfield, Sara, 77
- Chetty, Raj, 18–20, 71, 72
- Chicago Jury Project, 334
- Childress, James F., 333
- China, censorship of Internet in, 43–46, 74–75
- cholera, 29–30
- Chowdhury, Abdur, 28
- Cialdini, Robert, 220
- Cinematch, 246–47
- citizen science, 10, 272
- Clark, Eric M., 78
- closed questions, 111–13
- collective intelligence, 272
- Common Rule, 328–29; attempts to modernize,
 293; based on Belmont Report, 294; Encore
 research not covered by, 286; updating of, 332
- compensation for participants, 217, 334; Belmont
 Report on, 299; in mass collaborations, 265–66,
 268–69
- compliance with laws, 300
- complier average causal effect (CACE), 68–69, 76
- computer-administered surveys, 108
- computer-assisted human computation system,
 239, 241, 274
- Computer Fraud and Abuse Act (US), 300
- computers, 3; humans compared with, 273
- computer science, ethical issues in, 329–30
- conditional average treatment effect (CATE), 206
- confidential data, 39–41; in study of
 ex-offenders, 111
- confounders, 147–48
- Conrad, Frederick G., 139
- consensus, in mass collaborations, 236–37
- consent. *See* informed consent
- consequentialism, 302–3
- CONSORT (Consolidated Standard Reporting of
 Trials) guidelines, 216
- construct validity, 25–26, 163, 199, 212
- consumer studies, 55–60

- context-relative informational norms, 315–17
- contextual integrity, 315–16
- control groups, 150, 197–98, 224
- Cooperative Congressional Election Study (CCES), 99
- Coppock, Alexander, 188
- corporations: administrative records of, 82–83; ethical issues in experiments with data from, 290–93; experimentation in, 222; inaccessibility of data held by, 27–29; nowcasting useful for, 46; partnering with, 184–88; sensitive data of, 39–41
- Correll, Shelley J., 153–54
- Costa, Dora L., 168
- costs: in building your own experiments, 182; of experiments, 190–96, 216–17; of human interviewers, 108; of survey research, 98–99, 138, 141; of working with powerful partners, 184
- counting, 41–46, 74–75
- coverage bias, 93
- coverage errors, 92–93, 100
- Coviello, Lorenzo, 200, 218
- credit for work, in mass collaborations, 269, 277
- crowdsourcing, 10, 272
- Cukier, Kenneth, 11
- cultural products, 192–96
- cumulative advantage, 176
- custommades, 7–8, 355–56

- data augmentation, 274
- database of ruin, 27, 289, 290
- data collection: distributed, 256–64; participant-centered, 356–57
- data protection plans, 312
- data scientists, 6, 355; ethical concerns of, 281–82; on repurposing data, 16
- Deaton, Angus, 76
- debriefing participants, 306, 335–36, 353–54
- Defense Advanced Research Projects Agency (DARPA) Network Challenge, 272
- deferred consent, 336
- demand effects, 210
- Demographic and Health Surveys, 2, 128–29, 358
- demographic forecasting, 75
- Deng, Alex, 201
- deontology, 302–3
- Desposato, Scott, 334
- difference-in-differences estimators, 201, 203, 208, 219; difference-in-mean estimators compared with, 224
- difference-of-means estimators, 205, 208–9; difference-in-differences estimators compared with, 224
- differential privacy, 358
- digital experiments. *See* experiments, digital
- digital fingerprints, 71
- digital footprints, 71
- digital traces, 13, 71
- dirty data, 36–39, 74
- distributed data collection mass collaborations, 232, 256–57, 262–64, 271, 274; eBird, 257–59; ethical issues in, 269; PhotoCity, 259–62
- Dodds, Peter, 192, 357
- Doleac, Jennifer, 175
- Doll, Richard, 30, 72
- Donoho, David, 12
- draft lottery (conscription), 51, 53–55, 63, 66–69, 76–77
- drift, 33–34, 49, 73
- Duchamp, Marcel, 6–7, 355
- Duncan, Otis Dudley, 33
- Dunn, Halbert L., 26
- Dunning, Thad, 76
- Dwyer, Patrick C., 221–22

- eBay auctions, 55–60
- eBird (ornithology data), 257–59, 274
- Ebola virus, 281, 316
- e-cigarettes, 78
- ecological momentary assessments (EMA), 108–11
- The Economist* (magazine), 277–78
- Efrati, Amir, 81
- Egloff, Boris, 38
- Einav, Liran, 28, 55–60
- elections, 82; on Brexit, 144–45; Cooperative Congressional Election Study of, 99; errors in predicting results of, 100; experiment to simulate, 180–81; Facebook study of, 185–88; German parliamentary elections, 31–32; *Literary Digest* poll on, 91–94; non-probability samples for, 102–6; political manifestos crowd coded during, 241–44; problems in studying, 179–80; representation errors in samples predicting, 91–94; Twitter for data on, 33, 73; *See also* voting
- Emotional Contagion experiment, 197–201, 218, 221; ethical issues in, 284–85, 290–93, 319, 324, 331–32, 339; principle of Beneficence applied to, 297; principle of Justice applied to, 299; principle of Respect for Law and Public Interest

- applied to, 300, 301; worst-case scenarios in, 323
- empirically driven theorizing, 61–62
- employment discrimination, 304–5, 352
- Encore, 286–87; ethical issues in, 297–99, 318, 319;
 - informed consent issue for, 305; not under Common Rule, 330; principle of Respect for Law and Public Interest applied to, 301
- encouragement designs, 66, 228
- energy usage, 158–68, 170–71
- enriched asking, 118–22, 140
- ensemble solutions, 275
- environments for experiments: building new, 179–82; using existing, 174–79
- epidemiology, 46–49
- errors: in measurement, 94–98; in representation, 91–94; systematic, 20–21; total survey error framework, 89–91
- ESP Game, 273–74
- ethical issues, 8–11, 281–83, 324–25, 331–32, 357–58; in amplified asking, 141; in design of experiments, 196–202; in digital research, 288–93; ethical frameworks for, 301–3, 335; ethics as continuous, 324; in experiments in existing environments, 178–79; in Facebook study of Emotional Contagion, 284–85; of field experiments, 211; history of, 325–30; in inaccessible data, 27–29; informed consent, 303–7; institutional review boards and, 321–22; management of informational risk as, 307–14, 336–37; in mass collaborations, 268–69, 273, 277–78; nonreactivity as, 24; principles of, 294–301, 333; privacy as, 314–17, 337–38; in studies of ex-offenders, 110–11; in Tastes, Ties, and Time research project, 285; uncertainty and decisions on, 317–21, 338; in use of sensitive data, 40; worst-case scenarios in, 323–24
- ethical-response surveys, 320
- ethnography, 136
- exclusion restrictions, 54–55, 68
- existing environments, experiments within, 174–79
- ex-offenders, 109–11
- experiments, 147–49, 210; advice on design of, 188–90, 216; approximating, 50–60, 76–77; within businesses, 222; digital, 152, 154–57, 173–88, 210–11; digital, building, 179–82, 214–15; digital, building your own products for, 182–83, 215; digital, costs of, 190–96; digital, ethical issues in, 196–202, 288–93; digital, with powerful partners, 184–88, 215–16; digital, using existing environments, 174–79; elements of, 149–51; ethical issues in design of, 196–202; explanatory mechanisms (intervening variables) in, 169–73, 213–14; heterogeneity of treatment effects in, 167–69; lab versus field and analog versus digital, 151–57, 210–11; non-experimental data distinguished from, 209; potential outcomes framework for, 202–7; precision in, 207–9; survey experiments, 97; validity of, 161–67, 211–12
- external validity, 163–64
- Eyal, Nir, 305
- Facebook, 34; algorithmic confounding of data from, 35–36; data from merged with student data, 285; decline in sharing on, 81; in election study, 185–88; Emotional Contagion experiment using, 197–201, 218; Emotional Contagion experiment using, ethical issues in, 284–85, 290–93, 297; Friendsense survey on, 115–17; impact on friendships of, 88; predictions made by, 145; Tastes, Ties, and Time research project used data from, 285, 290, 297; on voting behavior, 72; voting behavior linked to, 140; *See also* Emotional Contagion experiment
- Farber, Henry, 42–43, 74, 311
- Feamster, Nick, 286
- feature engineering, 238–39
- Ferraro, Paul J., 171–73, 223
- field experiments, 151–54, 210–11; digital, 156–57; informed consent in, 304–5; simulating, 181–82
- Fisher, Dana R., 11
- “fishing,” 213, 219
- Fitzpatrick, J. W., 257
- fixed costs, 141, 190; in MusicLab, 194, 196
- Foldit, 249–51, 273, 275; credit for participants in, 269, 277; participants in, 270
- forecasting, 46–50, 75–76
- found data, 82–83
- Fragile Families and Child Wellbeing Study, 255–56
- frame populations, 92–93
- Friendsense (survey), 115–17, 357
- fundamental problem of causal inference, 64, 204–5
- Funk, Simon, 248–49
- Gaddis, S. Michael, 274
- Galaxy Zoo, 232–41, 245, 274, 357; credit for participants in, 269; new discoveries from,

- 267–68; participants in, 270
 Gallacher, John E. J., 72–73
 Gallup, George, 138
 games, 273–74
 gamification, 115–17
 Gardner, Howard, 25
 Gelman, Andrew, 102–4, 139, 354
 gender: employment discrimination based on, 352; post-stratification techniques for, 104
 generalizing from samples to populations, 31
 General Social Survey (GSS), 70, 116, 142; data collection for, 256–57; Twitter compared with, 14–15
 Gerber, Alan S., 207, 342, 343
 Gill, Michael, 351
 Ginsberg, Jeremy, 47
 Glaeser, Edward, 255
 Glaser, Barney, 61
 Gleick, James, 11
 Goel, Sharad, 102–6, 115, 117, 142, 357
 Goldman, William, 192
 Goldthorpe, John H., 82
 Google: Flu Trends reports of, 36, 47–49, 75, 77, 356; Google Books, 18; Google NGrams dataset, 79–80
 government administrative records, 15–16; data linkage to, 140
 governments: administrative records of, 82–83; ethical issues in experiments with data from, 290–93; inaccessibility of data held by, 27–29; nowcasting useful for, 46; sensitive data of, 39–41; as source of big data, 15–16, 70; voting records kept by, 119
 Graepel, Thore, 12, 145
 Grebner, Mark, 342, 343
 Green, Donald P., 207, 341–42
 Grimmelmann, James, 300
 grounded theory, 61–62
 Group Insurance Commission (GIC; Massachusetts), 308–9
 Groves, Robert M., 12, 137, 138
 Guess, Andrew, 188
 Guillory, Jamie, 197–201, 221

 Hancock, Jeffrey, 197–201, 221
 Hargittai, Eszter, 11, 12, 83
 Harper, F. Maxwell, 215
 Harrison, John, 274–75
 Hart, Nicky, 82
 hashtags, 34
 Hatch, Elizabeth E., 72–73

 Hauge, Michelle, 344
 Hausel, Peter, 196
 Hawthorne effects, 210
 Heckman, James J., 76
 Heller, Jean, 327
 Hersh, Eitan, 119–21, 140, 356
 heterogeneity: in mass collaborations, 266–67; of treatment effects, 167–69, 212–13
 Hill, A. Bradford, 30, 72
 Hill, Seth, 179–81
 history, sociology compared with, 82
 Holland, Paul W., 64, 204
 Home Energy Reports, 164–68, 170–71
 Homeland Security, US Department of: on ethical guidelines for information and communications technology, 330; Menlo Report of, 294; tracking and monitoring of social media by, 80
 homogeneous-response-propensities-within-groups assumption, 104–5
 Horvitz–Thompson estimator, 131–33
 Huber, Gregory A., 179–81, 214, 215
 Huberty, Mark, 73
 human computation mass collaborations, 231–34, 244–46, 271, 273–75; crowd-coding of political manifestos, 241–44; Galaxy Zoo project, 234–41; open call projects compared with, 254–55
 human interviewers, 108
 human rights abuses, 290
 Humphreys, Macartan, 278, 336, 339–40, 347

 Imbens, Guido W., 76
 imputation (user-attribute inference), 26; *See also* amplified asking
 inaccessibility of data, 27–29, 72
 inattentive participants, 215
 incarceration rates, 109
 incompleteness, 24–27, 72
 in-depth interviews, 85
 inferred consent, 336
 influenza (flu), 46–49, 75
 InfluenzaNet project, 277
 informational risk, management of, 307–14, 336–37
 informed consent, 295, 303–7, 333–36, 343; Belmont Report and Common Rule on, 329; for data linkage, 140; ethical frameworks on, 302; lacking in Encore research, 286–87; in mass collaborations, 269; in study of ex-offenders,

- 111; in Tuskegee Syphilis Study, 327
- injunctive norms, 159–60
- institutional review boards (IRBs), 281–82, 321–22; attempts to avoid, 332; Common Rule on, 329; on informed consent, 304; Montana voting study and, 350; nonresearchers on, 297
- instrumental variables, 66, 76
- intelligence, 25
- Intelligence Community Comprehensive National Cybersecurity Initiative Data Center (Utah Data Center), 27
- intergenerational social mobility, 19–20
- internal states, 88
- internal validity, 163, 212
- Internet: advertising on, 186–87, 226–27; Chinese censorship of, 43–46, 74–75; Encore study of censorship of, 286–87, 297–98; mass collaborations using, 231, 270; online courses on, 225–26; online dating sites on, 350–51; of things, 3; *See also* Encore; social media
- interrupted time series design, 80
- Inter-university Consortium for Political and Social Research, 314
- intervening variables (mechanisms), 169–73, 213–14
- interviewer effects, 108
- interviews, 85; by humans versus computer-administered, 108; new techniques versus, 107
- Issenberg, Sasha, 209, 343
- item nonresponse, 133

- Jensen, Robert, 83
- Jewish Chronic Disease Hospital, 345–47
- Johnson, Jeremy, 348
- Jones, Jason J., 216
- Judson, D. H., 140
- jury deliberations, 334
- Justice, 298–99, 334

- Kahn, Matthew E., 168
- Kahn, Robert Louis, 137
- Kant, Immanuel, 302
- Kauffman, Jason, 343
- Keeter, Scott, 98
- Keohane, Robert O., 11
- Khan, Shamus, 11
- King, Gary, 11, 43–46, 74
- Kirkegaard, Emil, 350–51
- Kleinberg, Jon, 74, 76
- Kleinsman, John, 339

- Kohavi, Ron, 212
- Konstan, Joseph A., 215
- Kosinski, Michal, 12, 145, 332
- Koun S. C., 49
- Kramer, Adam D. I., 197–201, 218, 221
- Kraut, Robert E., 88, 118–19
- Küfner, Albrecht C.P., 38

- lab experiments, 151–54, 210–11; building, 179–82, 214–15; costs of, 190; digital, 155
- Labour Party (United Kingdom), 241
- Lakhani, Karim, 278–79
- Landon, Alf, 92
- language, Google Books data on, 18
- Lapin, Lisa, 348
- Larimer, Christopher, 342
- Lazer, David, 49
- League of Conservation Voters, 188
- Le Comber, Steven C., 344
- Ledeen, Ken, 11
- Lenz, Gabriel S., 179–81, 214
- Levy, Karen E. C., 11, 113
- Lewis, Harry, 11
- Lewis, Randall A., 227
- Lintott, Chris, 235, 237, 245, 270, 357
- Literary Digest*, 91–94, 100, 138
- local average treatment effect (LATE), 68, 76
- Longford, Nicholas T., 210
- Longitude Prize, 274–75
- longitudinal studies, big data for, 21
- lotteries: draft lottery, 51, 53–55, 63, 66–69, 76–77; for research subjects, 217
- Loveman, Gary, 224
- Lowrance, William W., 314–15
- lung cancer, 30–31

- Maddock, Jim, 339
- Maki, Alexander, 221–22
- Malawi Journals Project, 262–63, 269, 274
- Malchow, Hal, 341–42
- Manifesto Project, 241–44, 278
- Manzi, Jim, 11, 209, 222, 224
- Maryland, 300–301
- Mas, Alexandre, 52–53, 55
- Mason, Robert, 339
- Mason, Winter, 115, 357
- Massachusetts, 308–9
- mass collaborations, 10, 231–33, 271–77; crowd-coding of political manifestos, 241–44; designing your own, 265–71; distributed data collection, 256–57, 262–64; eBird, 257–59;

- Foldit as, 249–51; Galaxy Zoo project, 234–41; human computation for, 233–34, 244–46; Netflix Prize, 246–49; open calls, 246, 254–56; Peer-to-Patent, 252–54; PhotoCity, 259–62
- matching, 55, 59–60, 77
- Matorny, Tim, 269
- Mayer, Jonathan, 8
- Mayer-Schönberger, Viktor, 11
- McCulloch, Linda, 347
- measurement, 94–98, 138; errors in, 90
- mechanisms, in experiments (intervening variables), 169–73, 213–14
- mediating variables (mechanisms; intervening variables), 169–73, 213–14
- mediators, 211
- medical research: informed consent in, 305; staged trials in, 320–21; Tuskegee Syphilis Study, 325–28
- Menlo Report, 294–95, 330; on Respect for Law and Public Interest, 299–301
- metadata, from mobile phones, 123, 126, 337
- Metcalf, Jacob, 345
- Michel, Jean-Baptiste, 18, 79
- Michelangelo, 6–7, 355
- microsurveys, 107
- microtask labor market, 274; ethical issues involving, 269; used in mass collaborations, 242, 245, 265, 276
- Mill, John Stuart, 302
- Miller, Franklin G., 335
- minimal risk standard, 318–19
- Miranda, Juan Jose, 171–73
- Mitofsky, Warren, 87
- mobile phones (cell phones), 107, 143; call detail records from, 146, 347; data on mobility of people on, 281, 316; for distributed data collection, 264; in Rwanda, 123, 126
- moderators, 211
- modes of surveys, 107
- monotonicity assumption, 68
- Montana, 347–50
- Moore, David W., 144
- Moretti, Enrico, 52–53, 55
- motherhood penalty, 153–54
- motivating by absence, 41–42
- Motl, Jonathan, 347–48
- MovieLens project, 73, 182–83, 215
- movie ratings, 39–40; Netflix Prize for, 246–49
- MTurk (Amazon Mechanical Turk), 81, 155–56, 220, 279; cost of, 191; debate over use of, 222; recruiting participants using, 179–81, 214–15
- multilevel regression, 105; with post-stratification, 105–6, 138–39
- multiple comparison problems, 213
- MusicLab (website), 192–96, 216
- Narayanan, Arvind, 39–40, 269, 310
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 328
- National Security Agency, US (NSA), 27
- natural experiments, 29–30, 51–60, 76; mathematics of, 65–70; to replace larger experiments, 199–200, 218
- Netflix Prize, 39–40, 232, 246–49; as ensemble solution, 275; ethical issues in, 269, 310–11, 336–37
- Neumark, David, 352
- Newman, Mark W., 24
- Newton, Isaac, 274–75
- New York City, 113–14; restaurants in, 351–52
- NGOs (nongovernmental organizations), 188
- Nielsen, Michael, 267, 268
- Nissenbaum, Helen, 315–17
- non-experimental approaches, 50–51; experiments distinguished from, 209; making causal inferences from, 62; matching in, 55; potential outcomes framework for, 62–65
- non-probability sampling methods, 99–107, 136, 142; probability samples versus, 139
- nonreactive data, 23–24, 71–72
- nonrepresentative data, 29–33, 72–73
- nonresponse problem, 138; in probability sampling, 100; probability sampling with, 133–35; as source of bias, 93; survey costs and, 98
- Noveck, Beth, 252–53
- nowcasting, 46–50, 75–76
- Obama, Barack, 103–5
- observing: asking questions versus, 87–89, 137; observational data from, 13
- Occupy Gezi (Turkey), 21
- Ohm, Paul, 74, 289
- On, Robert, 125, 358–59
- one-sided noncompliance, 228
- online courses, 225–26
- online dating sites, 350–51
- online panels, 102
- open call mass collaborations, 232, 246, 254–56, 271, 274–76; ethical issues in, 269; Foldit as, 249–51; Netflix Prize, 246–49;

- Peer-to-Patent, 252–54
- open-ended questions, 111–13
- Opower (firm), 164, 212, 356
- organizations, 188
- out-of-sample generalizations, 30

- pagers, data from, 37–38
- Paik, In, 153–54
- Pan, Jennifer, 43–44
- Panagopoulos, Costas, 315, 316
- panopticon, 288–89
- parallel programming, 71
- Park, David K., 139
- participants in experiments: Beneficence towards, 296; compensation for, 217; debriefing, 306, 335–36, 353–54; MTurk for recruiting, 214–15; paying, 191; recruiting, 179–81, 210; reducing number of, 197, 200–201; respect for, 295
- participants in mass collaborations, 265–66, 269–70, 276–77; data collection centered around, 356–57; ethical issues for, 268–69
- participatory sensing, 264
- Pasek, Josh, 140
- Pasteur, Louis, 184
- Pasteur's Quadrant, 184–85, 188, 215–16
- patents, 252–54
- Pearl, Judea, 62
- Pearson, Steve, 253
- Peer-to-Patent, 252–54, 275
- Pell, Jill P., 224
- Penney, Jonathon, 80, 81
- people-centric sensing, 264
- performativity, 35–36, 74
- personally identifying information (PII), 309, 337
- Persons, Respect for. *See* Respect for Persons
- perturb-and-observe experiments, 148–49
- Pirate Party (Germany), 32
- political manifestos, crowd-coding of, 241–44
- PolyMath project, 272
- populations: drifting of, 33–34, 73; generalizing from samples to, 31; sampling errors in, 92–94
- Porter, Nathaniel D., 274
- post-stratification techniques, 103–5, 138, 142; in non-probability sampling, 136; nonresponse bias in, 135; stratified sampling and, 132, 133
- potential outcomes framework, 62–65, 202–7
- Pothole Patrol, 264
- power analysis, 201, 319
- Precautionary Principle, 318
- precision, 207–9
- prediction, in social research, 276
- prediction markets, 272
- predictive models, 275
- pre-registration, 216
- presidential elections: of 1936, 91–94, 138; of 2012, 33, 102–6
- Presser, Stanley, 112, 143
- pre-testing questions, 98
- pre-treatment covariates, 218–19
- pre-treatment information, 157, 218
- Price, Michael K., 171–73, 223
- principles-based approach to ethical issues, 282
- prisons, 109
- privacy, 314–17, 337–38; in bathrooms, 340–41; differential privacy, 358; in mass collaborations, 269; in open calls, 275–76; in record linkage, 140; in sensitive data, 40; in use of Wikipedia, 80–81
- probability sampling methods, 99–100, 138; mathematics for, 131–33; non-probability sampling versus, 139; with nonresponse, 133–35
- Proceedings of the National Academy of Sciences*, 199, 284
- productivity studies, 52–53
- products, for experiments, 182–83, 215
- programmatic research, 216
- protein folding, 249–51
- pruning data, 55
- psychology, lab experiments in, 152; debriefing participants in, 336; deception in, 306
- psychometrics, 138
- Public Health Service, US (PHS), 325–28
- Purdam, Kingsley, 278
- Pury, Cynthia, 37–38

- question form effects, 95–96
- questionnaires, 96–98
- questions, 94–98; amplified asking of, 122–30; cause-and-effect, 147–48; enriched asking of, 118–22; how to ask, 139; new ways to ask, 107–17; observing versus, 87–89, 137; order of, 143–44; representation for, 99–107; who to ask, 138–39

- race: in economic transactions, 175–76; studies of discrimination by, 304–5, 352–54; in Tuskegee Syphilis Study, 294, 325–28
- random-digit dialing, 86, 87
- randomization, 77, 212; in experimental design, 151; of treatment and control groups, 205–6
- randomized controlled experiments, 148–49; elements of, 149–51

- Rao, Justin M., 227
rare events, 18
reactivity, 23–24
readymades, 7–8, 355–56
real-time estimates, 22
record linkage, 26–27, 119, 140–41
recruiting participants, 179–81, 210; MTurk for, 214–15
reduction in number of participants, 197, 217–19
redundancy, in mass collaborations, 236
refining the treatment, 197, 217, 218
re-identification of data. *See* anonymization of data
replacement, 197, 217–18
representation, 91–94, 99–107, 138; errors in, 90; representative data, 29–30
repurposing data, 219
research design, 5; counting in, 41–46; encouragement designs for, 66; ethical issues in, 357–58; experiments approximated in, 50–60; forecasting and nowcasting, 46–50
Respect for Law and Public Interest, 299–301, 335; Menlo Report on, 330
Respect for Persons, 295, 317, 333–34; based on deontology, 302; informed consent and, 306–7
respondents, 93; nonresponse problem and, 133–35; viral recruitment of, 116, 144
restaurants, 351–52
Restivo, Michael: ethical issues in work of, 296–97; participants conscripted by, 217; Wikipedia experiment by, 150–51, 155, 157, 190–91, 203–4, 206–9
retrospective data, 22, 37
return on investment (ROI), 226–27
retweets (Twitter), 78
risk/benefit analysis, 296–97
Roberts, Molly, 43–44
Rogers, Todd, 171
Rome (Italy), 259–62
Romney, Mitt, 103
Roosevelt, Franklin, 92
Rossmo, D. Kim, 344
Rothman, Alexander J., 221–22
Rothman, Kenneth J., 72–73
Rothschild, David, 102–6
Rubin, Donald, 62
Russell, William Moy Stratton, 196–97, 217
Rwanda, 1–2, 123–32, 332, 358–59
sample populations, 93
samples: errors in, 90–91; generalizing to populations from, 31
sampling, 99–107; errors in, 93; non-probability sampling methods, 99–107, 136; probability sampling, 131–33; probability sampling, with nonresponse, 133–35; representation in, 91–94; representativeness of, 29
sampling frames, 92
Sandvig, Christian, 11, 12, 83
Santillana, Mauricio, 49
Schawinski, Kevin, 234–35, 237, 245, 270, 357
Schechter, Stuart, 320
Schober, Michael F., 107, 139
Schultz, P. Wesley, 223; costs of research by, 190; energy conservation experiment by, 158–68, 171, 211, 219
Schuman, Howard, 112, 143
scurvy (disease), 170
self-report measures, 143
Seltzer, William, 290
sensitive data, 39–41, 74; Chinese censorship of social media on, 44; mobile phone metadata as, 123; re-identification of, 309–11; in study of ex-offenders, 111
September eleventh terrorist attacks, 20–21, 37–38
sharing data, 312–14
Shmatikov, Vitaly, 39–40, 269, 310
SIGCOMM (computer science conference), 287
simple experiments, 158
Singel, Ryan, 311
Singleton, Royce A., Jr., 11
smartphones, 107; for digital experiments, 155; ecological momentary assessments using, 109; *See also* mobile phones
Smith, Gordon C. S., 224
smoking, 30–31
Snapshot Serengeti, 274
Snow, John, 29–30
social bots, 74
social desirability bias, 108
social experiments, 211
social media: algorithmic confounding of data from, 35–36; in China, censorship of, 43–46; for elections data, 73; for real-time data, 22; as source of big data, 15; tracking and monitoring of, 80–81; viral recruitment of respondents using, 116; *See also* Facebook; Twitter
social mobility, 18–20
social networks, 181–82; Tastes, Ties, and Time research project on, 285
social scientists, 6, 355; ethical concerns of, 281–82; on repurposing data, 16

- Social Security Administration, US, 51, 140
- sociology, history compared with, 82
- Sommer, Robert, 340
- Sommers, Roseanna, 335
- Southam, Chester M., 346–47
- spam (dirty data), 36–39, 74
- Spirling, Arthur, 351
- split–apply–combine strategy, 273; used in Galaxy Zoo, 234, 237–38; used in Manifesto Project, 242
- Stable Unit Treatment Value Assumption (SUTVA), 206–7
- staged trials, 320–21
- Starbird, Kate, 339
- statistical conclusion validity, 161–63
- statistical data editing, 74
- Statistics Netherlands, 73
- Statistics Sweden, 140
- Stein, Luke, 175
- Stephens-Davidowitz, Seth, 23
- Stevenson, Mark, 344
- Stewart, Neil, 215
- Stillwell, David, 12, 145
- Straits, Bruce C., 11
- strata, 132
- stratified experimental designs (blocked experimental designs), 218–19
- stratified sampling, 132–33
- Strauss, Anselm, 61
- Subrahmanian, V. S., 74
- Sudman, Seymour, 94
- Sugie, Naomi, 109–11
- supervised learning, 44–46, 239
- surveillance, 80–81, 288–90, 332
- survey experiments, 97, 211
- survey research, 85–87; costs of, 98–99; enriched asking in, 118–22; gamification used in, 115–17; linked to big data sources, 117–30, 140–41; measurement in, 94–98; new ways to ask questions in, 107–17, 139; nonresponse problem in, 133–35; representation in, 91–94; total survey error in, 89–91; wiki surveys, 111–15
- Sweeney, Latanya, 308–9
- syphilis, 294, 325–28
- systematic error, 20–21
- system drift, 34, 36, 73
- target populations, 92; stratified sampling of, 132–33
- Tastes, Ties, and Time research project, 285, 290, 297, 299, 343
- taxi drivers, 42–43, 311
- telephones, 85–87; *Literary Digest* survey using, 91–94; metadata from, 337; nonresponse in surveys using, 100; *See also* mobile phones; smartphones
- Tenuto, Justin, 278
- Terentev, Ivan, 269
- terms-of-service agreements, 300, 335
- Ternovski, John, 188
- terrorism: September eleventh terrorist attacks, 20–21, 37–38; tracking and monitoring of social media for, 80–81
- text messages, 107
- theoretical constructs, 24–25; big data and, 61–62; construct validity with, 163; performativity and, 35
- time-dependent mobilizations, 272
- Tockar, Anthony, 311
- Toole, Jameson L., 146
- total survey error framework, 89–91, 98, 107, 137–38
- transitivity, in social networks, 35–36
- transparency-based accountability, 300
- transparency paradox, 333
- transportability of patterns, 31
- treatment effects: conditional average treatment effect, 206; explanatory mechanisms for, 169–73; heterogeneity of, 167–69, 212–13; precision in, 207–9
- triadic closure, 74
- Tucker, Clyde, 137
- Tufekci, Zeynep, 34
- Tuite, Kathleen, 260
- Tumasjan, Andranik, 31–32
- Turkey, 21, 34
- Tuskegee Syphilis Study, 294, 325–28
- Twitter: for election predictions, 31–33, 73, 75; repurposing of data from, 14–15; retweets, 78; spam on, 38–39; for stock market predictions, 77; to study protests in Turkey, 21, 34; triadic closure in, 74
- 2^k factorial design, 172–73
- two-sided noncompliance, 229
- Ugander, Johan, 35
- uncertainty, decisions in face of, 317–21, 338
- unexpected events, 21–22
- unit nonresponse, 133–35
- Urzúa, Sergio, 76
- user-attribute inference (imputation), 26

- Utah Data Center (Intelligence Community Comprehensive National Cybersecurity Initiative Data Center), 27
- validity, 161–67, 211–12
- van Arkel, Hanny, 268
- van de Rijt, Arnout: ethical issues in work of, 296–97; low costs of research by, 190–91; random rewards experiment by, 176–77, 220; Wikipedia experiment by, 150–51, 155, 157, 203–4, 206–9, 217
- van der Windt, Peter, 278, 347
- variable costs, 190
- variables: instrumental variables, 66, 76; intervening variables (mechanisms), 169–73
- variance, 89; trade-off between bias and, 138
- Verba, Sidney, 11
- Verdery, Ashton M., 274
- viral recruitment, 116
- vitamin C, 170
- von Ahn, Luis, 273
- voting: Affluent Worker Project study of, 82; on Brexit, 144–45; Facebook study of, 185–88; get-out-the-vote experiments, 216, 347–50; privacy and, 315, 341–43; research into, 119–21
- Waksberg, Joseph, 87
- walled garden approach, 313–14
- Wang, Wei, 102–6, 136, 138–39
- Wansink, Brian, 94
- Warren, Samuel D., 337
- Watkins, Susan, 262–63
- Watts, Duncan: Friendsense survey by, 115, 357; MusicLab website by, 192, 196; Twitter data used by, 21
- Webb, Eugene J., 71
- Weld, William, 308–9
- Wichita Jury Study, 334
- Wickham, Hadley, 273
- WikiLeaks, 351
- Wikipedia, 113; bot-created edits to, 39; ethical issues in study of, 296–97; as mass collaboration, 231; privacy concerns in using, 80–81; rewards in, 150–51, 191, 202–4, 206–9
- wiki surveys, 111–15
- winner-take-all markets, 192, 194, 217
- within-sample comparisons, 30, 33
- within-subjects designs, 160
- women: employment discrimination against, 352; motherhood penalty for, 153–54; study of voting by, 82
- Wong-Parodi, Gabrielle, 223
- wording effects, 95–96
- Xbox users, 102–6, 136, 138–39
- Xie, Huizhi, 201
- Yang, Shihao, 49
- YouGov (firm), 144–45
- Zaccanelli, Scott “Boots,” 270
- Zevenbergen, Ben, 286–87