Chapter 6

Professional Ethics

6.1 Introduction

Work in data analytics involves expert knowledge, understanding, and skill. In much of your work, you will be relying on the trust and confidence that your clients place in you. The term *professional ethics* describes the special responsibilities not to take unfair advantage of that trust. This involves more than being thoughtful and using common sense; there are specific professional standards that should guide your actions.

The best known professional standards are those in the Hippocratic Oath for physicians, which were originally written in the 5th century B.C. Three of the eight principles in the modern version of the oath [237] are presented here because of similarity to standards for data analytics.

- "I will not be ashamed to say 'I know not,' nor will I fail to call in my colleagues when the skills of another are needed for a patient's recovery."
- "I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know."
- "I will remember that I remain a member of society, with special obligations to all my fellow human beings, those sound of mind and body as well as the infirm."

Depending on the jurisdiction, these principles are extended and qualified by law. For instance, notwithstanding the need to "respect the privacy of my patients," health-care providers in the United States are required by law to report to appropriate government authorities evidence of child abuse or infectious diseases such as botulism, chicken pox, and cholera.

This chapter introduces principles of professional ethics for data analytics and gives examples of legal obligations as well as guidelines issued by professional societies. There is no data analyst's oath—only guidelines. Reasonable people can disagree about what actions are best, but the existing guidelines provide a description of the ethical expectations on which your clients can reasonably rely. As a consensus statement of professional ethics, the guidelines also establish standards of accountability.

6.2 Truthful falsehoods

The single best-selling book with "statistics" in the title is *How to Lie with Statistics* by Darrell Huff [114]. Written in the 1950s, the book shows graphical ploys to fool people

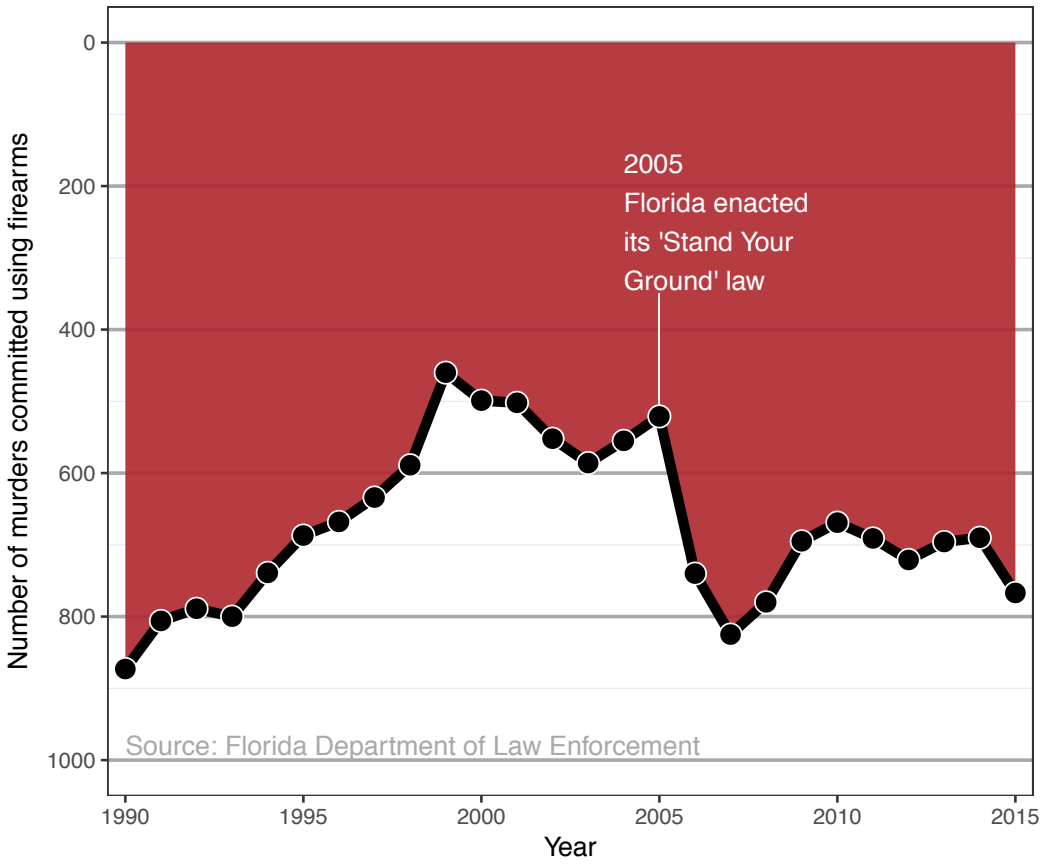

Figure 6.1: Reproduction of a data graphic reporting the number of gun deaths in Florida over time. The original image was published by Reuters.

even with accurate data. A general method is to violate conventions and tacit expectations that readers rely on when interpreting graphs. One way to think of *How to Lie* is a text to show the general public what these tacit expectations are and give tips for detecting when the trick is being played on them. The book's title, while compelling, has wrongly tarred the field of statistics. The "statistics" of the title are really just "numbers." The misleading graphical techniques are employed by politicians, journalists, and businessmen: not statisticians. More accurate titles would be "How to Lie with Numbers," or "Don't be misled by graphics."

Some of the graphical tricks in "How to Lie ..." are still in use. Consider these two recent examples.

In 2005, the Florida legislature passed the controversial "Stand Your Ground" law that broadened the situations in which citizens can use lethal force to protect themselves against perceived threats. Advocates believed that the new law would ultimately reduce crime; opponents feared an increase in the use of lethal force. What was the actual outcome?

The graphic in Figure 6.1 is a reproduction of one published by the news service Reuters showing the number of firearm murders in Florida over the years (see Exercise 4.18). Upon first glance, the graphic gives the visual impression that right after the passage of the 2005 law, the number of murders decreased substantially. However, the numbers tell a different story.

The convention in data graphics is that up corresponds to increasing values. This is not an obscure convention—rather, it's a standard part of the secondary school curriculum. Close inspection reveals that the y-axis in Figure 6.1 has been flipped upside down—the number of gun deaths increased sharply after 2005.

Figure 6.2 shows another example of misleading graphics: a tweet by the news magazine *National Review* on the subject of climate change. The dominant visual impression of the graphic is that global temperature has hardly changed at all.

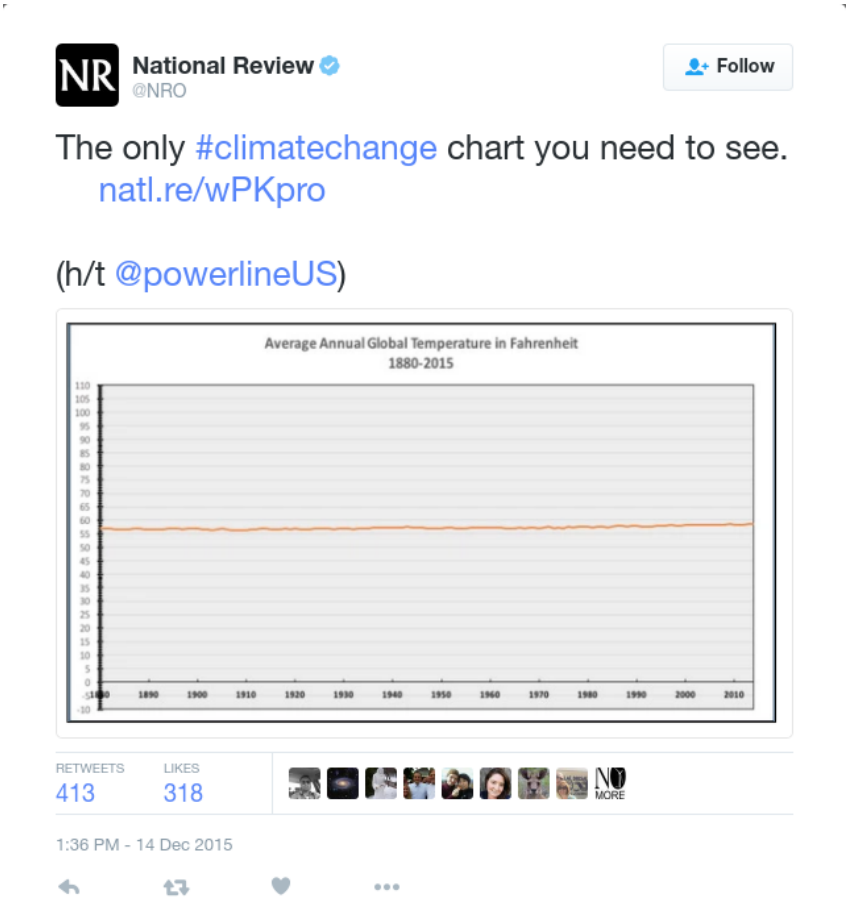

Figure 6.2: A tweet by *National Review* on December 14, 2015 showing the change in global temperature over time.

There is a tacit graphical convention that the coordinate scales on which the data are plotted are relevant to an informed interpretation of the data. The x-axis follows the convention—1880 to 2015 is a reasonable choice when considering the relationship between human industrial activity and climate. The y-axis, however, is utterly misleading. The scale goes from -10 to 110 degrees Fahrenheit. While this is a relevant scale for showing *season-to-season* variation in temperature, that is not the salient issue with respect to climate change. The concern with climate change is about rising ocean levels, intensification of storms, ecological and agricultural disruption, etc. These are the anticipated results of a change in global *average* temperature on the order of 5 degrees Fahrenheit. The *National Review* graphic has obscured the data by showing them on an irrelevant scale where the

actual changes in temperature are practically invisible. By graying out the numbers on the y-axis, the *National Review* makes it even harder to see the trick that's being played.

The examples in Figures 6.1 and 6.2 are not about lying with statistics. Statistical methodology doesn't enter into them. It's the professional ethics of journalism that the graphics violate, aided and abetted by an irresponsible ignorance of statistical methodology. Insofar as both graphics concern matters of political controversy, they can be seen as part of the blustering and bloviating of politics. While politics may be a profession, it's a profession without any comprehensive standard of professional ethics.

6.3 Some settings for professional ethics

Common sense is a good starting point for evaluating the ethics of a situation. Tell the truth. Don't steal. Don't harm innocent people. But professional ethics also require a neutral, unemotional, and informed assessment. A dramatic illustration of this comes from legal ethics: a situation where the lawyers for an accused murderer found the bodies of two victims whose deaths were unknown to authorities and to the victims' families. The responsibility to confidentiality for their client precluded the lawyers from following their hearts and reporting the discovery. The lawyers' careers were destroyed by the public and political recriminations that followed, yet courts and legal scholars have confirmed that the lawyers were right to do what they did, and have even held them up as heroes for their ethical behavior.

Such extreme drama is rare. This section describes in brief six situations that raise questions of the ethical course of action. Some are drawn from the authors' personal experience, others from court cases and other reports. The purpose of these short case reports is to raise questions. Principles for addressing those questions are the subject of the next section.

6.3.1 The chief executive officer

One of us once worked as a statistical consultant for a client who wanted a proprietary model to predict commercial outcomes. After reviewing the literature, an existing multiple linear regression model was found that matched the scenario well and available public data were used to fit the parameters of the model. The client's staff were pleased with the result, but the CEO wanted a model that would give a competitive advantage. After all, their competitors could easily follow the same process to the same model, so what advantage would the client's company have? The CEO asked the statistical consultant whether the coefficients in the model could be "tweaked" to reflect the specific values of his company. The consultant suggested that this would not be appropriate, that the fitted coefficients best match the data and to change them arbitrarily would be "playing God." In response, the CEO rose from his chair and asserted, "I want to play God."

How should the consultant respond?

6.3.2 Employment discrimination

One of us works with legal cases arising from audits of employers, conducted by the United States Office of Federal Contract Compliance Programs (OFCCP). In a typical case, the OFCCP asks for hiring and salary data from a company that has a contract with the United States government. The company usually complies, sometimes unaware that the OFCCP applies a method to identify "discrimination" through a two-standard-deviation test outlined in the Uniform Guidelines on Employee Selection Procedures (UGESP). A

company that does not discriminate has some risk of being labeled as discriminating by the OFCCP method [41]. By using a questionable statistical method, is the OFCCP acting unethically?

6.3.3 Data scraping

In May 2016, the online OpenPsych Forum published a paper titled "The OkCupid data set: A very large public data set of dating site users". The resulting data set contained 2,620 variables—including usernames, gender, and dating preferences—from 68,371 people scraped from the OkCupid dating website. The ostensible purpose of the data dump was to provide an interesting open public data set to fellow researchers. These data might be used to answer questions such as this one suggested in the abstract of the paper: whether the Zodiac sign of each user was associated with any of the other variables (spoiler alert: it wasn't).

The data scraping did not involve any illicit technology such as breaking passwords. Nonetheless, the author received many comments on the OpenPsych Forum challenging the work as an ethical breach in *doxing* people by releasing personal data. Does the work raise actual ethical issues?

6.3.4 Reproducible spreadsheet analysis

In 2010, Harvard economists Carmen Reinhart and Kenneth Rogoff published a report entitled *Growth in a Time of Debt* [177], which argued that countries which pursued austerity measures did not necessarily suffer from slow economic growth. These ideas influenced the thinking of policymakers—notably United States Congressman Paul Ryan—during the time of the European debt crisis.

Graduate student Thomas Herndon requested access to the data and analysis contained in the paper. After receiving the original spreadsheet from Reinhart, Herndon found several errors.

> "I clicked on cell L51, and saw that they had only averaged rows 30 through 44, instead of rows 30 through 49." —Thomas Herndon [179]

In a critique [100] of the paper, Herndon, Ash, and Pollin point out coding errors, selective inclusion of data, and odd weighting of summary statistics that shaped the conclusions of the Reinhart/Rogoff paper.

Does publishing a flawed analysis raise ethical questions?

6.3.5 Drug dangers

In September 2004, drug company Merck withdrew from the market a popular product Vioxx because of evidence that the drug increases the risk of myocardial infarction (MI), a major type of heart attack. Approximately 20 million Americans had taken Vioxx up to that point. The leading medical journal *Lancet* later reported an estimate that Vioxx use resulted in 88,000 Americans having heart attacks, of whom 38,000 died.

Vioxx had been approved in May 1999 by the United States Food and Drug Administration based on tests involving 5,400 subjects. Slightly more than a year after the FDA approval, a study [36] of 8,076 patients published in another leading medical journal, *The New England Journal of Medicine*, established that Vioxx reduced the incidence of severe gastro-intestinal events substantially compared to the standard treatment, naproxen. That's good for Vioxx. In addition, the abstract reports these findings regarding heart attacks:

> "The incidence of myocardial infarction was lower among patients in the naproxen group than among those in the [Vioxx] group (0.1 percent vs. 0.4 percent; relative risk, 0.2; 95% confidence interval, 0.1 to 0.7); the overall mortality rate and the rate of death from cardiovascular causes were similar in the two groups."

Read the abstract again carefully. The Vioxx group had a much *higher* rate of MI than the group taking the standard treatment. This influential report identified the high risk soon after the drug was approved for use. Yet Vioxx was not withdrawn for another three years. Something clearly went wrong here. Did it involve an ethical lapse?

6.3.6 Legal negotiations

Lawyers sometimes retain statistical experts to help plan negotiations. In a common scenario, the defense lawyer will be negotiating the amount of damages in a case with the plaintiff's attorney. Plaintiffs will ask the statistician to estimate the amount of damages, with a clear but implicit directive that the estimate should reflect the plaintiff's interests. Similarly, the defense will ask their own expert to construct a framework that produces an estimate at a lower level.

Is this a game statisticians should play?

6.4 Some principles to guide ethical action

As noted previously, lying, cheating, and stealing are common and longstanding unethical behaviors. To guide professional action, however, more nuance and understanding is needed. For instance, an essential aspect of the economy is that firms compete. As a natural part of such competition, firms hurt one another; they take away business that the competitor would otherwise have. We don't consider competition to be unethical, although there are certainly limits to ethical competition.

As a professional, you possess skills that are not widely available. A fundamental notion of professional ethics is to avoid using those skills in a way that is effectively lying—leading others to believe one thing when in fact something different is true. In every professional action you take, there is an implicit promise that you can be relied on—that you will use appropriate methods and draw appropriate conclusions. Non-professionals are not always in a position to make an informed judgment about whether your methods and conclusions are appropriate. Part of acting in a professionally ethical way is making sure that your methods and conclusions are indeed appropriate.

It is necessary to believe that your methods and conclusions are appropriate, but not sufficient. First, it's easy to mislead yourself, particularly in the heat and excitement of satisfying your client or your research team. Second, it's usually not a matter of absolutes: It's not always certain that a method is appropriate. Instead, there is almost always a risk that something is wrong.

An important way to deal with these issues is to draw on generally recognized professional standards. Some examples: Use software systems that have been vetted by the community. Check that your data are what you believe them to be. Don't use analytical methods that would not pass scrutiny by professional colleagues.

Note that the previous paragraph says "draw on" rather than "scrupulously follow." Inevitably there will be parts of your work that are not and cannot be vetted by the community. You write your own data wrangling statements: They aren't always vetted. In special circumstances you might reasonably choose to use software that is new or created just for the purpose at hand. You can look for internal consistency in your data, but it

would be unreasonable in most circumstances to insist on tracking everything back to the original point at which it was measured.

Another important approach is to be open and honest. Don't overstate your confidence in results. Point out to clients substantial risks of error or unexpected outcome. If you would squirm if some aspect or another of your work came under expert scrutiny, it's likely that you should draw attention to that aspect yourself.

Still, there are limits. You generally can't usefully inform your clients of *every* possible risk and methodological limitation. The information would overwhelm them. And you usually will not have the resources—time, money, data—that you would need to make every aspect of your work perfect. You have to use good professional judgment to identify the most salient risks and to ensure that your work is good enough even if it's not perfect.

You have a professional responsibility to particular stakeholders. It's important that you consider and recognize all the various stakeholders to whom you have this responsibility. These vary depending on the circumstances. Sometimes, your main responsibility is simply to your employer or your client. In other circumstances, you will have a responsibility to the general public or to subjects in your study or individuals represented in your data. You may have a special responsibility to the research community or to your profession itself. The legal system can also impose responsibilities; there are laws that are relevant to your work. Expert witnesses in court cases have a particular responsibility to the court itself.

Another concern is the potential for a conflict of interest. A *conflict of interest* is not itself unethical. We all have such conflicts: We want to do work that will advance us professionally, which instills a temptation to satisfy the expectations of our employers or colleagues or the marketplace. The conflict refers to the *potential* that our personal goals may cloud or bias or otherwise shape our professional judgment.

Many professional fields have rules that govern actions in the face of a conflict of interest. Judges recuse themselves when they have a prior involvement in a case. Lawyers and law firms should not represent different clients whose interests are at odds with each other. Clear protocols and standards for analysis regulated by the FDA help ensure that potential conflicts of interest for researchers working for drug companies do not distort results. There's always a basic professional obligation to disclose potential conflicts of interest to your clients, to journals, etc.

For concreteness, here is a list of professional ethical precepts. It's simplistic; it's not feasible to capture every nuance in a brief exposition.

1. Do your work well by your own standards and by the standards of your profession.
2. Recognize the parties to whom you have a special professional obligation.
3. Report results and methods honestly and respect your responsibility to identify and report flaws and shortcomings in your work.

6.4.1 Applying the precepts

Let's explore how these precepts play out in the several scenarios outlined in the previous section.

The CEO

You've been asked by a company CEO to modify model coefficients from the correct values, that is, from the values found by a generally accepted method. The stakeholder in this setting is the company. If your work will involve a method that's not generally accepted by the professional community, you're obliged to point this out to the company.

Remember that your client also has substantial knowledge of how their business works. Statistical purity is not the issue. Your work is a tool for your client to use; they can use it as they want. Going a little further, it's important to realize that your client's needs may not map well onto a particular statistical methodology. The consultant should work genuinely to understand the client's whole set of interests. Often the problem that clients identify is not really the problem that needs to be solved when seen from an expert statistical perspective.

Employment discrimination

The procedures adopted by the OFCCP are stated using statistical terms like "standard deviation" that themselves suggest that they are part of a legitimate statistical method. Yet the methods raise significant questions, since by construction they will sometimes label a company that is not discriminating as a discriminator. OFCCP and others might argue that they are not a statistical organization. They are enforcing a law, not participating in research. The OFCCP has a responsibility to the courts. The courts themselves, including the United States Supreme Court, have not developed or even called for a coherent approach to the use of statistics (although in 1977 the Supreme Court labeled differences greater than two or three standard deviations as too large to attribute solely to chance).

Data scraping

OkCupid provides public access to data. A researcher uses legitimate means to acquire those data. What could be wrong?

There is the matter of the stakeholders. The collection of data was intended to support psychological research. The ethics of research involving humans requires that the human not be exposed to any risk for which consent has not been explicitly given. The OkCupid members did not provide such consent. Since the data contain information that makes it possible to identify individual humans, there is a realistic risk of the release of potentially embarrassing information, or worse, information that jeopardizes the physical safety of certain users.

Another stakeholder is OkCupid itself. Many information providers, like OkCupid, have *terms of use* that restrict how the data may be legitimately used. Such terms of use (see Section 6.5.3) form an explicit agreement between the service and the users of that service. They cannot ethically be disregarded.

Reproducible spreadsheet analysis

The scientific community as a whole is a stakeholder in public research. Insofar as the research is used to inform public policy, the public as a whole is a stakeholder. Researchers have an obligation to be truthful in their reporting of research. This is not just a matter of being honest, but also of participating in the process by which scientific work is challenged or confirmed. Reinhart and Rogoff honored this professional obligation by providing reasonable access to their software and data.

Note that it is not an ethical obligation to reach correct research results. The obligation is to do everything feasible to ensure that the conclusions faithfully reflect the data and the theoretical framework in which the data are analyzed. Scientific findings are often subject to dispute, reinterpretation, and refinement.

Since this book is specifically about data science, it can be helpful to examine the Reinhart and Rogoff findings with respect to the professional standards of data science. Note that these can be different from the professional standards of economics, which might reasonably be the ones that economists like Reinhart and Rogoff adopt. So the following is

not a criticism of them, *per se*, but an opportunity to delineate standards relevant to data scientists.

Seen from the perspective of data science, Microsoft Excel, the tool used by Reinhart and Rogoff, is an unfortunate choice. It mixes the data with the analysis. It works at a low level of abstraction, so it's difficult to program in a concise and readable way. Commands are customized to a particular size and organization of data, so it's hard to apply to a new or modified data set. One of the major strategies in debugging is to work on a data set where the answer is known; this is impractical in Excel. Programming and revision in Excel generally involves lots of click-and-drag copying, which is itself an error-prone operation.

Data science professionals have an ethical obligation to use tools that are reliable, verifiable, and conducive to reproducible data analysis (see Appendix D). This is a good reason for professionals to eschew Excel.

Drug dangers

When something goes wrong on a large scale, it's tempting to look for a breach of ethics. This may indeed identify an offender, but we must also beware of creating scapegoats. With Vioxx, there were many claims, counterclaims, and lawsuits. The researchers failed to incorporate some data that were available and provided a misleading summary of results. The journal editors also failed to highlight the very substantial problem of the increased rate of myocardial infarction with Vioxx.

To be sure, it's unethical not to include data that undermines the conclusion presented in a paper. The Vioxx researchers were acting according to their original research protocol—a solid professional practice.

What seems to have happened with Vioxx is that the researchers had a theory that the higher rate of infarction was not due to Vioxx, *per se*, but to an aspect of the study protocol that excluded subjects who were being treated with aspirin to reduce the risk of heart attacks. The researchers believed with some justification that the drug to which Vioxx was being compared, naproxen, was acting as a substitute for aspirin. They were wrong, as subsequent research showed.

Professional ethics dictate that professional standards be applied in work. Incidents like Vioxx should remind us to work with appropriate humility and to be vigilant to the possibility that our own explanations are misleading us.

Legal negotiations

In legal cases such as the one described earlier in the chapter, the data scientist has ethical obligations to their client. Depending on the circumstances, they may also have obligations to the court.

As always, you should be forthright with your client. Usually you will be using methods that you deem appropriate, but on occasion you will be directed to use a method that you think is inappropriate. For instance, we've seen occasions when the client requested that the time period of data included in the analysis be limited in some way to produce a "better" result. We've had clients ask us to subdivide the data (in employment discrimination cases, say, by job title) in order to change p-values. Although such subdivision may be entirely legitimate, the decision about subdividing—seen from a purely statistical point of view—ought to be based on the situation, not the desired outcome (see the discussion of the "garden of forking paths" in Section 7.7).

Your client is entitled to make such requests. Whether or not you think the method being asked for is the right one doesn't enter into it. Your professional obligation is to

inform the client what the flaws in the proposed method are and how and why you think another method would be better. (See the major exception that follows.)

The legal system in countries such as the U.S. is an *adversarial* system. Lawyers are allowed to frame legal arguments that may be dismissed: They are entitled to enter some facts and not others into evidence. Of course, the opposing legal team is entitled to create their own legal arguments and to cross-examine the evidence to show how it is incomplete and misleading. When you are working with a legal team as a data scientist, you are part of the team. The lawyers on the team are the experts about what negotiation strategies and legal theories to use, how to define the limits of the case (such as damages), and how to present their case or negotiate with the other party.

It is a different matter when you are presenting to the court. This might take the form of filing an expert report to the court, testifying as an expert witness, or being deposed. A deposition is when you are questioned, under oath, outside of the court room. You are obliged to answer all questions honestly. (Your lawyer may, however, direct you not to answer a question about privileged communications.)

If you are an expert witness or filing an expert report, the word "expert" is significant. A court will certify you as an expert in a case giving you permission to express your opinions. Now you have professional ethical obligations to apply your expertise honestly and openly in forming those opinions.

When working on a legal case, you should get advice from a legal authority, which might be your client. Remember that if you do shoddy work, or fail to reply honestly to the other side's criticisms of your work, your credibility as an expert will be imperiled.

6.5 Data and disclosure

6.5.1 Reidentification and disclosure avoidance

The ability to link multiple data sets and to use public information to identify individuals is a growing problem. A glaring example of this occurred in 1996 when then-Governor of Massachusetts William Weld collapsed while attending a graduation ceremony at Bentley College. An MIT graduate student used information from a public data release by the Massachusetts Group Insurance Commission to identify Weld's subsequent hospitalization records. The disclosure of this information was highly publicized and led to many changes in data releases. This was a situation where the right balance was not struck between disclosure (to help improve health care and control costs) and nondisclosure (to help ensure private information is not made public). There are many challenges to ensure disclosure avoidance [244, 151]: This remains an active and important area of research.

The Health Insurance Portability and Accountability Act (HIPAA) was passed by the United States Congress in 1996—the same year as Weld's illness. The law augmented and clarified the role that researchers and medical care providers had in maintaining protected health information (PHI). The HIPAA regulations developed since then specify procedures to ensure that individually identifiable PHI is protected when it is transferred, received, handled, analyzed, or shared. As an example, detailed geographic information (e.g., home or office location) is not allowed to be shared unless there is an overriding need. For research purposes, geographic information might be limited to state or territory, though for certain rare diseases or characteristics even this level of detail may lead to disclosure. Those whose PHI is not protected can file a complaint with the Office of Civil Rights.

The HIPAA structure, while limited to medical information, provides a useful model for disclosure avoidance that is relevant to other data scientists. Parties accessing PHI need to have privacy policies and procedures. They must identify a privacy official and

undertake training of their employees. If there is a disclosure they must mitigate the effects to the extent practical. There must be reasonable data safeguards to prevent intentional or unintentional use. Covered entities may not retaliate against someone for assisting in investigations of disclosures. They must maintain records and documentation for six years after their last use of the data. Similar regulations protect information collected by the statistical agencies of the United States.

6.5.2 Safe data storage

Inadvertent disclosures of data can be even more damaging than planned disclosures. Stories abound of protected data being made available on the Internet with subsequent harm to those whose information is made accessible. Such releases may be due to misconfigured databases, malware, theft, or by posting on a public forum. Each individual and organization needs to practice safe computing, to regularly audit their systems, and to implement plans to address computer and data security. Such policies need to ensure that protections remain even when equipment is transferred or disposed of.

6.5.3 Data scraping and terms of use

A different issue arises relating to legal status of material on the Web. Consider Zillow.com, an online real-estate database company that combines data from a number of public and private sources to generate house price and rental information on more than 100 million homes across the United States. Zillow has made access to their database available through an API (see Section 5.5.2) under certain restrictions. The terms of use for Zillow are provided in a legal document. They require that users of the API consider the data on an "as is" basis, not replicate functionality of the Zillow website or mobile app, not retain any copies of the Zillow data, not separately extract data elements to enhance other data files, and not use the data for direct marketing.

Another common form for terms of use is a limit to the amount or frequency of access. Zillow's API is limited to 1,000 calls per day to the home valuations or property details. Another example: The Weather Underground maintains an API focused on weather information. They provide no-cost access limited to 500 calls per day and 10 calls per minute and with no access to historical information. They have a for-pay system with multiple tiers for accessing more extensive data.

Data points are not just content in tabular form. Text is also data. Many websites have restrictions on text mining. Slate.com, for example, states that users may not:

> "Engage in unauthorized spidering, scraping, or harvesting of content or information, or use any other unauthorized automated means to compile information."

Apparently, it violates the Slate.com terms of use to compile a compendium of Slate articles (even for personal use) without their authorization.

To get authorization, you need to ask for it. For instance, Albert Kim of Middlebury College published data with information for 59,946 San Francisco OkCupid users (a free online dating website) with the permission of the president of OkCupid [125]. To help minimize possible damage, he also removed certain variables (e.g., username) that would make it more straightforward to reidentify the profiles. Contrast the concern for privacy taken here to the careless doxing of OkCupid users mentioned above.

6.6 Reproducibility

Disappointingly often, even the original researchers are unable to reproduce their own results. This failure arises naturally enough when researchers use menu-driven software that does not keep an audit trail of each step in the process. For instance, in Excel, the process of sorting data is not recorded. You can't look at a spreadsheet and determine what range of data was sorted, so mistakes in selecting cases or variables for a sort are propagated untraceably through the subsequent analysis. Researchers commonly use tools like word processors that do not mandate an explicit tie between the result presented in a publication and the analysis that produced the result. These seemingly innocuous practices contribute to the loss of reproducibility: numbers may be copied by hand into a document and graphics are cut-and-pasted into the report. (Imagine that you have inserted a graphic into a report in this way. How could you, or anyone else, easily demonstrate that the correct graphic was selected for inclusion?)

Reproducible analysis is the practice of recording each and every step, no matter how trivial seeming, in a data analysis. The main elements of a reproducible analysis plan (as described by Project TIER (`https://www.haverford.edu/project-tier`) include:

Data: all original data files in the form in which they originated,

Metadata: codebooks and other information needed to understand the data,

Commands: the computer code needed to extract, transform, and load the data—then run analyses, fit models, generate graphical displays, and

Map: a file that maps between the output and the results in the report.

The American Statistical Association (ASA) notes the importance of reproducible analysis in its curricular guidelines. The development of new tools such as R Markdown and `knitr` have dramatically improved the usability of these methods in practice. See Appendix D for an introduction to these tools.

Individuals and organizations have been working to develop protocols to facilitate making the data analysis process more transparent and to integrate this into the workflow of practitioners and students. One of us has worked as part of a research project team at the Channing Laboratory at Harvard University. As part of the vetting process for all manuscripts, an analyst outside of the research team is required to review all programs used to generate results. In addition, another individual is responsible for checking each number in the paper to ensure that it was correctly transcribed from the results. Similar practice is underway at The Odum Institute for Research in Social Science at the University of North Carolina. This organization performs third-party code and data verification for several political science journals.

6.6.1 Example: Erroneous data merging

In Chapter 4, we discuss how the *join* operation can be used to merge two data tables together. Incorrect merges can be very difficult to unravel unless the exact details of the merge have been recorded. The `dplyr inner_join()` function simplifies this process.

In a 2013 paper published in the journal *Brain, Behavior, and Immunity*, Kern et al. reported a link between immune response and depression. To their credit, the authors later noticed that the results were the artifact of a faulty data merge between the lab results and other survey data. A retraction [124], as well as a corrected paper reporting negative results [123], were published in the same journal.

In some ways this is science done well—ultimately the correct negative result was published, and the authors acted ethically by alerting the journal editor to their mistake. However, the error likely would have been caught earlier had the authors adhered to stricter standards of reproducibility (see Appendix D) in the first place.

6.7 Professional guidelines for ethical conduct

This chapter has outlined basic principles of professional ethics. Usefully, several organizations have developed detailed statements on topics such as professionalism, integrity of data and methods, responsibilities to stakeholders, conflicts of interest, and the response to allegations of misconduct. One good source is the framework for professional ethics endorsed by the American Statistical Association (ASA) [58].

The Committee on Science, Engineering, and Public Policy of the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine has published the third edition of *On Being a Scientist: A Guide to Responsible Conduct in Research*. The guide is structured into a number of chapters, many of which are highly relevant for data scientists (including "the Treatment of Data," "Mistakes and Negligence," "Sharing of Results," "Competing Interests, Commitment, and Values," and "The Researcher in Society").

The Association for Computing Machinery (ACM)—the world's largest computing society, with more than 100,000 members—adopted a code of ethics in 1992 (see `https://www.acm.org/about/code-of-ethics`). Other relevant statements and codes of conduct have been promulgated by the Data Science Association (`http://www.datascienceassn.org/code-of-conduct.html`), the International Statistical Institute (`http://www.isi-web.org/about-isi/professional-ethics`), and the United Nations Statistics Division (`http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx`). The Belmont Report outlines ethical principles and guidelines for the protection of human research subjects.

6.8 Ethics, collectively

Although science is carried out by individuals and teams, the scientific community as a whole is a stakeholder. Some of the ethical responsibilities faced by data scientists are created by the collective nature of the enterprise.

A team of Columbia University scientists discovered that a former post-doc in the group, unbeknownst to the others, had fabricated and falsified research reported in articles in the journals *Cell* and *Nature*. Needless to say, the post-doc had violated his ethical obligations both with respect to his colleagues and to the scientific enterprise as a whole. When the misconduct was discovered, the other members of the team incurred an ethical obligation to the scientific community. In fulfillment of this obligation, they notified the journals and retracted the papers, which had been highly cited. To be sure, such episodes can tarnish the reputation of even the innocent team members, but the ethical obligation outweighs the desire to protect one's reputation.

Perhaps surprisingly, there are situations where it is not ethical *not* to publish one's work. "Publication bias" (or the "file-drawer problem") refers to the situation where reports of statistically significant (i.e., $p < 0.05$) results are much more likely to be published than reports where the results are not statistically significant. In many settings, this bias is for the good; a lot of scientific work is in the pursuit of hypotheses that turn out to be wrong or ideas that turn out not to be productive.

But with many research teams investigating similar ideas, or even with a single research team that goes down many parallel paths, the meaning of "statistically significant" becomes

clouded and corrupt. Imagine 100 parallel research efforts to investigate the effect of a drug that in reality has no effect at all. Roughly five of those efforts are expected to culminate in a misleadingly "statistically significant" ($p < 0.05$) result. Combine this with publication bias and the scientific literature might consist of reports on just the five projects that happened to be significant. In isolation, five such reports would be considered substantial evidence about the (non-null) effect of the drug. It might seem unlikely that there would be 100 parallel research efforts on the same drug, but at any given time there are tens of thousands of research efforts, any one of which has a 5% chance of producing a significant result even if there were no genuine effect.

The American Statistical Association's ethical guidelines state, "Selecting the one 'significant' result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading." So, if you're examining the effect on five different measures of health by five different foods, and you find that broccoli consumption has a statistically significant relationship with the development of colon cancer, not only should you be skeptical but you should include in your report the null result for the other twenty-four tests or perform an appropriate statistical correction to account for the multiple tests. Often, there may be several different outcome measures, several different food types, and several potential covariates (age, sex, whether breastfed as an infant, smoking, the geographical area of residence or upbringing, etc.), so it's easy to be performing dozens or hundreds of different tests without realizing it.

For clinical health trials, there are efforts to address this problem through trial registries. In such registries (e.g., `https://clinicaltrials.gov`), researchers provide their study design and analysis protocol in advance and post results.

6.9 Further resources

For a book-length treatment of ethical issues in statistics, see [113]. A historical perspective on the ASA's Ethical Guidelines for Statistical Practice can be found in [70]. The University of Michigan provides an EdX course on "Data Science Ethics." Gelman has written a column on ethics in statistics in *CHANCE* for the past several years (see, for example [84, 86, 85]). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* describes a number of frightening uses of big data and algorithms [153].

The Center for Open Science—which develops the Open Science Framework (OSF)—is an organization that promotes openness, integrity, and reproducibility in scientific research. The OSF provides an online platform for researchers to publish their scientific projects. Emil Kirkegaard used OSF to publish his OkCupid data set.

The Institute for Quantitative Social Science at Harvard and the Berkeley Initiative for Transparency in the Social Sciences are two other organizations working to promote reproducibility in social science research. The American Political Association has incorporated the Data Access and Research Transparency (DA-RT) principles into its ethics guide. The Consolidated Standards of Reporting Trials (CONSORT) statement at `http://www.consort-statement.org` provides detailed guidance on the analysis and reporting of clinical trials.

Many more examples of how irreproducibility has led to scientific errors are available at `http://retractionwatch.com/`. For example, a study linking severe illness and divorce rates was retracted due to a coding mistake.

6.10 Exercises

Exercise 6.1

A researcher is interested in the relationship of weather to sentiment on Twitter. They want to scrape data from `www.wunderground.com` and join that to Tweets in that geographic area at a particular time. One complication is that Weather Underground limits the number of data points that can be downloaded for free using their API (application program interface). The researcher sets up six free accounts to allow them to collect the data they want in a shorter time-frame. What ethical guidelines are violated by this approach to data scraping?

Exercise 6.2

A data analyst received permission to post a data set that was scraped from a social media site. The full data set included name, screen name, email address, geographic location, IP (Internet protocol) address, demographic profiles, and preferences for relationships. Why might it be problematic to post a deidentified form of this data set where name and email address were removed?

Exercise 6.3

A company uses a machine learning algorithm to determine which job advertisement to display for users searching for technology jobs. Based on past results, the algorithm tends to display lower paying jobs for women than for men (after controlling for other characteristics than gender). What ethical considerations might be considered when reviewing this algorithm?

Exercise 6.4

A reporter carried out a clinical trial of chocolate where a small number of overweight subjects who had received medical clearance were randomized to either eat dark chocolate or not to eat dark chocolate. They were followed for a period and their change in weight was recorded from baseline until the end of the study. More than a dozen outcomes were recorded and one proved to be significantly different in the treatment group than the outcome. This study was publicized and received coverage from a number of magazines and television programs. Outline the ethical considerations that arise in this situation.

Exercise 6.5

A data scientist compiled data from several public sources (voter registration, political contributions, tax records) that were used to predict sexual orientation of individuals in a community. What ethical considerations arise that should guide use of such data sets?

Exercise 6.6

A *Slate* article (`http://tinyurl.com/slate-ethics`) discussed whether race/ethnicity should be included in a predictive model for how long a homeless family would stay in homeless services. Discuss the ethical considerations involved in whether race/ethnicity should be included as a predictor in the model.

Exercise 6.7

In the United States, most students apply for grants or subsidized loans to finance their college education. Part of this process involves filling in a federal government form called the Free Application for Federal Student Aid (FAFSA). The form asks for information about family income and assets. The form also includes a place for listing the universities to which the information is to be sent. The data collected by FAFSA includes confidential

financial information (listing the schools eligible to receive the information is effectively giving permission to share the data with them).

It turns out that the order in which the schools are listed carries important information. Students typically apply to several schools, but can attend only one of them. Until recently, admissions offices at some universities used the information as an important part of their models of whether an admitted student will accept admissions. The earlier in a list a school appears, the more likely the student is to attend that school.

Here's the catch from the student's point of view. Some institutions use statistical models to allocate grant aid (a scarce resource) where it is most likely to help ensure that a student enrolls. For these schools, the more likely a student is deemed to accept admissions, the lower the amount of grant aid they are likely to receive.

Is this ethical? Discuss.

Exercise 6.8

In 2006, AOL released a database of search terms that users had used in the prior month (see `http://www.nytimes.com/2006/08/09/technology/09aol.html`). Research this disclosure and the reaction that ensued. What ethical issues are involved? What potential impact has this disclosure had?

Exercise 6.9

In the United States, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) governs the confidentiality of data collected by agencies such as the Bureau of Labor Statistics and the Census Bureau. What are the penalties for willful and knowing disclosure of protected information to unauthorized persons?

Exercise 6.10

A statistical analyst carried out an investigation of the association of gender and teaching evaluations at a university. They undertook exploratory analysis of the data and carried out a number of bivariate comparisons. The multiple items on the teaching evaluation were consolidated to a single measure based on these exploratory analyses. They used this information to construct a multivariable regression model that found evidence for biases. What issues might arise based on such an analytic approach?

Exercise 6.11

An investigative team wants to winnow the set of variables to include in their final multiple regression model. They have 100 variables and one outcome measured for $n = 250$ observations). They use the following procedure:

1. Fit each of the 100 bivariate models for the outcome as a function of a single predictor, then

2. Include all of the significant predictors in the overall model.

What does the distribution of the p-value for the overall test look like, assuming that there are no associations between any of the predictors and the outcome (all are assumed to be multivariate normal and independent). Carry out a simulation to check your answer.