5

Human-Centered Approaches to Data Science Problems

We have looked at the data science cycle (chapter 2), the different steps where human-centered concerns can and should be addressed (chapter 3), and the different types of data science tools and models (chapter 4). Now we zoom out to look at the data science process as a whole. In this chapter we aim to introduce you to what makes human-centered data science different at every step of the process from other approaches to data science. We show you four key practices:

(1)Anchor your project with good questions and start by first figuring out what you want to know rather than what data you have.
(2)Develop ethical practices as one of the most important aspects of human-centered data science.
(3)Think about how your projects may get used by others.
(4)Reflect on the data science process to improve your practices and to become better at data science.

The approach to data science that we show here concerns both the *process* of doing data science and the *context* in which one does data science. We emphasize that our ethical reasoning about our practice is inseparable from our practice. To paraphrase internet researcher Annette Markham (2006), ethics is method—method is ethics. Ethical data science is good data science, and good data science relies on practices that support ethical design, choices, and actions. The real power of addressing ethics in human-centered data science is that it shows how ethics is practice that is informed by principles.

Asking Good Questions

Good questions, not convenient datasets, provide the anchor for good data science projects. Having a question that guides a project helps you think about what data you need, how much data is enough, and how the data connects to the question you want to answer.

Different disciplines may emphasize this to different degrees. When Shion, whose academic background was in statistics, was a graduate student in information science, he began collaborating with a senior social scientist studying media and health. He wanted to predict social media nonuse in a large dataset. He could work within the categories of data to show statistical correlations. He crunched a lot of numbers and built what he thought at the time was an excellent, rigorous predictive model based on the dataset at hand. His faculty collaborator took one look at his initial results and asked him, “But what is your question here? What is it that you are trying to find an answer to?”

“I don’t understand. How many people don’t use social media?”

She replied, “That’s a start. But how are you going to define and predict non-use [when people don’t use social media]? … After all, there’s no column label in the data that says non-use.” The better way would have been to start with the question first, not the qualities and categories of the data. The end of the story was that this question led Shion and his collaborator to conduct a large survey to help understand how people perceive their own continuum of use and non-use. It also led them to several research publications on the topic. Getting more specific on the right questions to ask can lead to better and richer results.

This story is about learning how to *operationalize* what we are measuring in our projects and how we do it. Knowing that two categories are related, or “exploring” a dataset to see what relationships emerge, may be one place to start. Take a step back from the dataset to figure out what it is you want to ask and what it is you want to answer. Sometimes building toward answers requires other data, different data, or looking at the data in a different way.

Correlations are easy enough when you know how to do them. Good questions take work. Without a good question, those correlations cannot explain relationships or dynamics in human behavior. With good questions, our results leap from describing the relationships in our datasets to telling the story about the relationships and dynamics in the world—and explaining human behavior.

Good questions are meaningful. At the core, asking good questions is about figuring out what you want to know. This seems intuitive, and yet we find from our experience of teaching data science that asking good questions is one of the hardest things to learn how to do. Good questions ask something that we can find the answer to. Good questions ask something novel, something we do not yet know. Good questions are clear enough to be understood, specific enough to be doable with a data science project, and yet broad enough to capture something meaningful. Good questions are:

- *Empirical*: They ask something about the world.
- *Falsifiable*: We can answer them with evidence.
- *Focused*: We can answer them within our project.
- *Important*: We will be able to answer something significant with them.

We know interesting questions when we hear them: Do people in different countries share misinformation on social media at different rates? Does the structure of databases on climate information support more scientific collaboration on some datasets than others? Which matters more for helping sick people get out of intensive care—close collaboration among nurses or close collaboration between nurses and specialists? These are questions that Gina's colleagues and students recently asked and answered. Asking good questions helps us design better data science projects.

The most important quality for good questions is that they are important enough for someone else to want to listen to their answers. That means that the questions are "interesting." In 1971, sociologist Murray Davis shared a simple observation about interesting questions: they often follow a format that begins, "What seems to be X in reality is really non-X" (Davis 1971, 313). A more nuanced form of the question might be, "What seems to be X in reality is really non-X, but only for people with Y." We know women often face discrimination, but what happens when coders are anonymous on GitHub: Do women still face discrimination? (Vedres and Vasarhelyi 2019). We know Chinese social media faces governmental controls: Are there differences in what and how social media posts are removed (King, Pan, and Roberts 2013)? We think we know the answer, but with data and insights good questions may show us something novel or different.

Coming up with these kinds of counterintuitive questions is not straightforward. It requires knowing what people expect or assume. Counterintuitive questions "work" because they show how data science adds value—by helping a company better understand their customers, a newspaper to understand their readers, or a hospital to understand their patients. Without showing something novel, important, or interesting, it is difficult to justify the time and resources of a data science project. Good questions depend on knowing who thinks a question is a good one. It may be important for a newspaper to know that more people read the sports section on Tuesdays than Wednesdays, but that will not help a data scientist explain the question of *who* reads sports and why. Good questions help us think about the *why* of a problem as much as the *what* of a problem.

This practice of asking questions first is different from starting with a dataset and asking, "What might it be able to tell me?" Good questions guide data scientists through a process of defining and scoping their projects. Focusing on questions first is connecting to concepts and discoveries that are known and designing ways to find out the "why" of relationships. In psychology and other statistically driven social sciences, there is now a practice of "preregistration" of hypotheses to prevent what is called "p-hacking," chasing the small effects that inevitably appear in noisy data (Nuzzo 2014). It is easy to find something statistically significant in a large enough dataset, but we should be looking for effects that are *meaningful* within the context of everything else we know. And that usually relates to focusing on good questions first.

Asking questions first also helps us think through the ethics of formulating our project. It helps us in exploring what we *should* look for. Questions help us think about the problem or puzzle to explore in the project. This practice keeps us from only looking under the streetlight of where it is easiest or most convenient to look.

Ethics

Work in data science has the potential to harm people. *Ethics* refers to a set of principles that guide the behavior of a group or an individual to try to avoid harm and do good. In the United States, research on people—the "human subjects" of research—is governed by regulations that emerged from reports about unethical and harmful medical studies. *The Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) holds three general principles that must guide research with people, and these principles shape how research ethics are talked about today in university and medical settings: respect for persons, beneficence, and justice.

Case Study 5.1

Ethical Ethnography in Data Ethics

Katie Shilton, University of Maryland

I did not set out to work in data ethics. I started grad school to be an archivist and work with paper. But I got rerouted thanks to research happening in my program, and a bit of luck: finding the UCLA Center for Embedded Networked Sensing (CENS). CENS leaders were some of the first researchers to use mobile phones to collect data about people. This was before the era of smartphones—at the beginning, we were working with Nokia feature phones with an external GPS device stuck to the back and a huge external battery pack needed to power the thing through a full day.

CENS became the setting for my dissertation research. I conducted a participant-observation study of how CENS engineers grappled with the ethics of the data collections they were performing. Importantly, I was learning to be an ethnographer alongside the engineers who were learning how to collect pervasive data (personal information generated through digital interaction, like social media data, search histories, geolocation data, and wearable device data). I remember telling a CENS leader that I wanted to observe CENS researchers for my dissertation, and he replied, "Like monkeys?"

People were both welcoming of, and weirded out by, my presence as an observer, and I had to do a *lot* to mitigate that weirdness and build trust. I regularly presented on my work and engaged CENS members in discussions of what I thought I was seeing. I did formal interviews that allowed participants to tell me their version of the story. I checked quotes with participants before using them in publications. I sometimes chose not to publish vignettes that I thought might be identifiable, or even potentially hurtful. These are lessons I took with me: that people often don't like the idea of being studied, and that it takes work to build the trust to do so. And that building that trust means both *being* trustworthy and demonstrating that trustworthiness to participants.

And as studying social media, smart cities and IoT [internet of things] devices became increasingly common in research, my dual roles merged. First, I was lucky to have been thinking about the ethics of this kind of data collection early. And second, I was lucky to be thinking about it as an ethnographer—someone who had to negotiate trust to collect lots of observational data. I was just collecting it differently (with my eyes and my pen, rather than with sensors or online traces).

Over the years, I found academic friends and colleagues who were also interested in big data and privacy, ethics, and risk. Together, we started the PERVADE (Pervasive Data Ethics for Computational Research) project to study empirical questions in data ethics. PERVADE began in the context of a growing public backlash against data science research. People were angry about a set of big data projects focused on people and their habits, such as the Facebook emotional contagion study, releases of analyses by dating sites such as OkCupid, and notoriously, the selling of research data to Cambridge Analytica to attempt to influence elections. This backlash didn't surprise me; it felt like a throwback to my ethnographic work. People often object to being studied without their knowledge.

Ethnographers have spent more than three decades questioning the ethics of their practice, the meaning of their data, and the position of their methods within larger movements (such as the ethnographers who participated in first European and then American colonization of other areas of the world). Data science now needs a similar reckoning, and I think that data scientists can learn a lot from ethnographers, those original embedded observers. Like ethnographers, data scientists need to reflect on at least two principles: awareness and representation. How aware are the people we are studying of data creation, and of our ability to mine it? Studying intentionally created Tweets is different from studying the traces left by our phone GPS. We need to consider mechanisms for public awareness (ranging from traditional forms of giving "informed consent" to public scholarship to participatory methods) for particularly invisible or unobtrusive data collection. Second, data scientists should explicitly grapple with representation (i.e., who is included in the data, and who decides how their data is measured, structured, and recorded), and also power. We need to recognize when our data have historic biases or re-create injustices. We should use data science to increase the representation of those who have traditionally been left out of (or actively harmed by) research, and to hold our public institutions and powerful actors accountable. And we should reflect on the ways data science is embedded within larger social structures such as surveillance capitalism and historical inequities. We plan to use the PERVADE project to further each of those conversations.

- *Respect for persons* means that research should respect people's rights to exercise autonomy.
- *Beneficence* means research should not harm people and should, if possible, benefit them.
- *Justice* means that the benefits of research should be distributed fairly.

Society expects researchers to uphold these principles, and academic research is required to meet these standards. These principles are found in research ethics guidelines and regulations in many countries, even if the specific local regulations and laws differ. While there are different rules in different countries and cultures, we can build on these principles and uphold them in a way that expresses ethical pluralism—the notion that ethical concerns may vary around the world—and cross-cultural awareness of these differences.

Within these cultural differences there are different emphases on what constitutes *ethical*. In some countries, concern for individual autonomy guides how research should be conducted. In others, like the United States, some tolerance for "reasonable" risks to autonomy is granted for projects that might otherwise be expected to benefit the greater good (franzke et al. 2020). The Association of Internet Researchers created a document for conducting research online with people's social media and other digital data and for doing ethical research with artificial intelligence/machine learning (AI/ML) (franzke et al. 2020), and it guides our thinking on ethical practice in data science. The ethical guidelines they produced provide a way to think about projects, taking into consideration a wide range of potential scenarios and contexts rather than an overly simplified checklist or a rule-based approach for ethics, especially as it fits to one place.

Ethical action is guided by thinking through *how* research is done and what *context* it is done in. Human-centered data science relies on this approach to ethics as a practice: ethics is based on not just what people formally are supposed to do or the definitions of those actions, but what people do in practice. Ethics is not a checklist, because it is impossible to devise a set of rules to fit all situations. Ethics is not about a goal for people to optimize or to "game." In our view, ethics means being reflexive at every step—thinking of what could go wrong, who could be harmed, and who could benefit. Ethics is not just reflection, though, and simply being reflexive will not necessarily prevent harm. You may need to consult with others—especially with people whose data is being used in your analysis. As we show later in this chapter, each stage of the data science cycle can raise new questions and challenges for ethical concern.

There are different approaches to ethics. *Care ethics* involves thinking through ethics as care for others, not simply as trying to minimize harm to others or to maximize benefit. Cultivating the ability to put ourselves in the position of others helps us to design data science projects as if the people involved were our families and loved ones. Would that change our approach to the decisions that we make?

An Example of Ethics Practices and Principles to Develop and Extend

With these considerations in mind, we present one of many different sets of ethical guidelines for people in data science to develop and extend. These guidelines follow the Association of Internet Researchers 2020 Ethics 3.0 guidelines and principles We encourage our readers to read the whole document (franzke et al. 2020).

Respecting autonomy means that data science projects should respect people's rights to privacy and to making decisions that impact their lives, because datasets provide proxies for human subjects and their behavioral patterns.

Informed consent means working to ensure that people have the power to choose whether to participate in your data science project, as either the intended or unintended subjects of the research.

Legal frameworks and regulations govern laws about people's privacy. People's rights to data about them in large datasets varies by country (e.g., the European Union's General Data Protection Regulation, or GDPR, for privacy protection), by context (e.g., the US regulation on healthcare data and privacy, known as HIPAA), and within the United States by state (e.g., the California Consumer Privacy Act or CCPA). Doing ethical data science means being aware of and following the applicable laws. However, reflection is also necessary because what is minimally "legal" may not always be ethical.

Third-party data is important for many data science projects. When data scientists use existing datasets that were collected by parties other than the data scientists themselves or the people represented by the data, it is important to consider the relationship of power and responsibility with regards to how this third-party data was collected and what responsibilities data scientists have to the people represented in the data we work with or who are affected by the decisions that will be made with our results. We must also consider the potential impact of third-party data processing and the ethical risks that potentially poses. Europe's GDPR rules are particularly strong and specific about who processes the data.

Freely available data is not free. The rise of enormous amounts of data gathered about people in their everyday life exposes us to potential harms of reidentification and of products and services being designed that may present different levels of benefits or potential harm.

Data governance means how we take care of the data we are entrusted with and who governs the rights and responsibilities for handling data. Several well-publicized cases, such as the Cambridge Analytica example above or Latanya Sweeney's research on Massachusetts health care data we discussed in chapter 3, show how easy it is to expose sensitive private information through data science projects.

Case Study 5.2

Data Is People: The Ethics of Scraping Data

Casey Fiesler, University of Colorado

When you ask people whether it is ethical to scrape data from a website, there are two common responses: (1) Does scraping data violate Terms of Service? and (2) Is the data "public"?

Whether it is unethical for researchers to violate Terms of Service (TOS)—particularly data collection/scraping provisions—has been a topic of debate for years. Both the law and ethics on this topic are fuzzy. However, the problem with a TOS-based ethical decision on collecting data is that it assumes that (a) violating TOS is inherently unethical and (b) violating TOS is the *only* thing that could make collecting data unethical. I suggest that neither of these is true.

There may be situations in which violating TOS against the wishes of a company could be an ethical act. For example, in a recent court case, researchers claimed that TOS violations are necessary for algorithmic audits that can help uncover discriminatory practices—and the court agreed with the researchers. Another concern is that if research on a platform can only be conducted by researchers explicitly given access to that platform, this might distort scientific discovery, particularly if researchers may be constrained by the company that employs them. There are also many sites that *don't* prohibit data collection, but where the research might have ethical problems for any number of reasons. Even well-intentioned research or application might be harmful, regardless of legality.

In a recent research project, my collaborators and I analyzed the data scraping provisions from over 100 social networking sites and found that these provisions are vague, highly inconsistent across sites, and most importantly, almost entirely lacking in context. With the exception of a few vague mentions of "personal" data, most data collection provisions are entirely context-agnostic when it comes to, for example, what the specific data being collected is, who is collecting it, what it is being used for, and what the expectations and potential harms might be for the people who created that data.

However, if you are making decisions about collecting and using data that was created by people, context is critical. This brings us to the second common answer to the question, "Is it ethical to scrape data?" Unfortunately, for many people, the only context that matters is: "Is it *public*?"

"Public" is the magic word when it comes to research ethics. "But the data is already public" was the response from Harvard researchers in 2008, when they released a dataset of college students' Facebook profiles, and from Danish researchers in 2016, when they released a dataset scraped from OkCupid.

However, as with data scraping provisions, the "publicness" of data does not tell us what really matters. The idea that a tweet about what someone had for breakfast is exactly the same as a tweet revealing someone's sensitive health condition is absurd. After all, one of the potential harms of using public data is *amplification*—that is, spreading content beyond its intended audience. It probably will not do much harm if more people know about your breakfast cereal. Negative consequences might follow a tweet about someone having cancer, a sexually transmitted disease, or depression.

What you plan to *do* with the data also matters. Consider the example of Cambridge Analytica, where data collected for research was used for more than subjects were told—not for science, but for manipulating elections. Data scraped from public social media sites and dating profiles has also been used to create predictive machine learning algorithms designed to classify sexual orientation (to use the language of the project) or inappropriately label gender.

As researchers, we have a responsibility to acknowledge that factors like the type of data, the creator of that data, and our intended use for the data are important when it comes to collecting and using it. However, these factors are harder to judge and understand than "Does this violate TOS?" or "Is this public?"

It can be frustrating not to have explicit rules to follow. However, ethics is not necessarily so much about following rules as it is about being thoughtful and carefully examining each situation contextually. This requires care and often extra work, but it is important to remember that data is often not just *data*; it also represents people.

Expectations mean that people may have different expectations about how their data is used that may not be reflected in the terms and conditions of their agreements or regulated in the law. For example, when people post on social media, they may expect that their privacy is otherwise respected, even though they may have consented to particular uses of that data when they registered for their social media accounts. A survey of Twitter users found that the majority of respondents felt that "researchers should not be able to use tweets without consent" (Fiesler and Proferes 2018).

Research aims and risks of harm mean that the principles of respect, beneficence, and justice can be applied when we have clearly scoped data science projects with good questions. Clarifying the aims of the research means that they can be weighed against the potential harms.

Data science is a responsibility, to us and others, the communities we study, and society at large. People working in data science should take the time to be thoughtful about this responsibility.

Ethics is about practices guided by principles. Doing good human-centered data science is about thinking through the specific challenges of any project, not simply having a standard format for every project. There is no recipe or checklist that can solve every ethical dilemma.

Ethics means asking the right questions. Without a clearly scoped project and focused questions, data science projects can generate unnecessary risks to people through potential reidentification, exposure to risks of loss of privacy, or harm through classifying people into categories that have real-life consequences.

These ethical guidelines were based on careful thought and consideration, and there are other ethical guidelines from different perspectives, including from Indigenous peoples' or First Nations' points of view. These guidelines provide another way of *framing* the situation of racial and cultural minorities in their own languages and in their own ethical contexts (Smith 2013). The act of framing is important because a conceptual frame (the way you think about a person or a group) can powerfully influence your decisions about how to include or exclude a particular group's data and how you combine or distinguish their data, in relation to other groups. The Government of Nunavut published a set of principles for education that includes an extensive discussion of Nunavut normative ethics, organized into six major legal areas and detailed ethical principles within each area (Nunavut Department of Education 2007). Suvradip Maitra (2020) lists five Indigenous principles for the governance of AI that depend on relationships and relational ways of knowing. Reporting for the Canadian Commission for UNESCO's IdeaLab, Dick Bourgeois-Doyle (2019) summarizes a longer tradition of *two-eyed seeing*—that is, "seeing through the eyes" of one's Indigenous culture and simultaneously "seeing through the eyes" of a majority culture—to bring contemporary AI and Indigenous ways of knowing toward a common framework. These projects provide alternative ethical guidelines that emphasize more traditional Indigenous values of relationship, connectedness, and community.

Ethics of Representation of Data and Results

Ethical challenges can arise in seemingly simple aspects of data science. We will outline two of those aspects here. Please use these concerns *not* as a complete statement-of-concern but rather as a starting point into your own self-inquiry about the ethics of your projects and your protocols. As in our other examples, thinking these problems through will help you to do higher-quality data science work.

Data and results can be represented in many different ways. Some of these may suggest misleading results or obfuscate or detract from important elements of the findings. Different communities may draw different interpretations or implications from the same results. The more complicated or nuanced the results are, the harder it is to represent them appropriately to an audience. Data scientists face a challenge in making judgments about how to represent this information in a way that responsibly reflects the findings of the project.

This leads to storytelling with data, a topic we cover in more depth in chapter 8. The representation of data and results tells a story about the data science problem, the dataset, and the model. Data scientists should try to present this story as transparently and accountably as possible. Data representation is, at its core, a human-centered problem. We should consider the audiences for our work and what they need to understand it. Few students of data science are trained on how to think about the representation of results and data in this way. Instead, much of data science training focuses on the technical development of the model. But a human-centered approach to data science is different. We urge you to be aware of the critical points in the data science cycle where you make choices that shape your data, models, pipelines, results, and representations. Because data science is a cycle and not a linear path, data representation is an iterative process and may lead you to new questions.

Ethics of Training, Validating, and Testing Models

How we construct and test our models is also important for ethical practice and principles.

There are multiple lenses through which we can view the train-validate-test framework that is prevalent in data science courses. Yet the choices that underlie this framework are almost never unpacked. For example:

- Under what kinds of circumstances should one do a 50–50 train-validate split as opposed to a 70–30 train-validate split?
- When is it acceptable to decide that the trained model is validated well on a “held-out” validation dataset but should also be tested on a “never-seen-before” testing dataset?

For instance, the choice of *how* to split training-validation datasets, the choice of *when* to determine that validation is successful, and *how* and *when* the decision is made to test the final model against never-seen-before testing data—all of these choices are crucial. Thinking about the implications of training, validating, and testing is a human-centered data science problem. The best way to address this is to have appropriate background knowledge of the framework and decision points and maintain a freely available written narrative for the decisions that are made surrounding this framework.

Ethics alone is not a solution to all the problems facing people working in data science. Ethics is central to making a good data science project, but it can only provide guidance for your technical decisions; it cannot solve them. For example, ethics alone will not fix a poorly specified model. Ethics may lead you *toward* an interesting and good data science question before you then do the rest of the work, constructing a solid data science method and protocol to address the ethical question that you have chosen to investigate. Ethics is necessary but not sufficient for human-centered data science.

You may face many different ethical dilemmas, and you need to approach each one differently based on the situation. You may need to consult with different parties or organizations, depending on local groups and specific cultural differences. What should be your guiding principles? We suggest that one of your considerations should be the issue of fairness.

Fairness

Is your algorithm fair? This is a difficult question. There are many definitions of fairness, and often people don’t agree on the answer (Narayanan 2018). Despite the difficulty of this conversation, we should think it through as much as we can: Are certain groups benefiting more than others? Are certain groups being disadvantaged? Are people with specific demographics more vulnerable to harm by our system? Will our system amplify or perpetuate systemic inequities?

In chapter 7 you will see that there are many systems that attempt to answer complicated societal questions through algorithmic means. For example, many places in the United States use algorithms to decide who goes home on bail and who stays in jail after an arrest. These algorithms were often developed because some people believe that an algorithm will be “fairer” than a human—after all, it is based on statistics and logic and therefore is supposed to be impartial. This is especially important because we know that human decision makers (judges, prosecutors, other people in the position of power) can have difficulty separating their personal feelings from their decision-making and so can (consciously or unconsciously) make decisions that are based on prejudice. People turn to algorithms to help solve these issues, but there are problems with that optimistic view.

First, as we mentioned previously, while many people believe that they know what fairness is, they do not necessarily agree with one another. One researcher listed twenty-one different definitions of fairness (Narayanan 2018). Shira Mitchell and colleagues detailed the derivations of multiple approaches to fairness and documented the tensions among them (Mitchell et al. 2019). Without critically considering these diverse definitions, it is possible to fall into “traps” when creating data science systems (Selbst et al. 2019). In view of these multiple definitions and risks of making mistakes, the problem of formal fairness may be “(im)possible” to achieve (Friedler, Scheidegger, and Venkatasubramanian 2016). These questions are challenging. As one set of researchers observed, “We cannot expect machines to reconcile these differences when society has not” (Rovatsos, Mittelstadt, and Koene 2019, 2). Clearly, there is no “universal” definition of fairness, and therefore it seems premature to trust that an algorithm could compute the “right” type of fairness.

Second, some of these data science systems are proprietary. Their algorithms are not available for inspection (Girasa 2020). Even if we agreed on a “universal” fairness approach, we could not determine whether that approach was implemented in these systems. Because many of the multiple approaches to fairness seem obvious to us, there is reason to worry that the implementers may be writing a version of fairness that “seems fair” or that “makes sense.” To whom? On whose authority? And who can evaluate whether their concept of fairness was fairly implemented?

Third, there may be systematic and structural problems with the data in some of these data science systems. A sentencing algorithm that imposes long prison sentences on a group of people will have the effect of removing those people from the free (out of prison) population. The predictive purpose of the sentencing algorithm is to prevent people from reoffending, but we will never know if this group will reoffend because they are in prison and are thereby *removed from the sample* of people whose behavior we can analyze. Our model *prevents them from reoffending* by locking them in prison, and thus makes it impossible to disconfirm the model’s prediction. Similarly, a bank loan system is designed to avoid granting loans to people who will probably not repay those loans. However, it systematically denies loans to the people who are predicted to be high-risk borrowers, and therefore there is no way to test the accuracy of the prediction because the high-risk people have been *removed from the sample* of people who received loans. Of course, we cannot know if they would repay a loan that they never received. Thus, these types of systems impose severe limitations on the collection of data that could validate or invalidate their own predictions.

Data scientists need to be accountable for their work and pay attention to the fairness of their algorithms. Accountability means that we cannot wash our hands of the project after we finish the pipeline. We should remain accountable to the people whose data the pipeline is using or whose lives might be changed by its effects. Also, accountability requires that we think carefully and critically about how our pipelines could be used for unexpected purposes. For example, research groups that developed neural-network-based video imitations of famous people did not initially see the potential for harmful misuse of their elegant generative models to produce “deep fakes” (Houde et al. 2020). Yet it is easy to imagine how deep fakes could be used for questionable political purposes or misinformation (Howard 2020). We should be considering what other applications could be found for our work—especially if it can have wide-ranging effects like this case.

Designing Projects for Others to Build On

Reproducibility means designing your projects so that others can test whether they would reach the same results. It is especially a concern for human-centered data science because it takes an approach that considers the people who use or reuse your data/model/pipeline, including possibly yourself at a future time.

- Think about the people who might reuse your data, your model, or your pipeline. Do they have everything they need to reuse them well? What might they need to make your data/model/pipeline better? Who can use your data, and how will they use it? Who can’t because they are excluded in one way or another?

Accessibility is another thing to consider. Are people with disabilities unintentionally excluded from building on your work? Does your project work for people who are color-blind or have other vision impairments? We often make visualizations with shades of red and green as the primary colors in diagrams. But those are the hardest colors to distinguish for people with red-green color vision deficiency, which affects about 9 percent of the population.

Openness is another consideration. Can you share your results and how you arrived at them? This is especially important when we consider public-sector projects and working with governments. But this is also important more broadly. If you want your results or pipelines to be useful, people need to be able to access them. For example, for those of us who work within the domain of natural disaster, it is important that people who work in disaster response and relief have access to our results. On the other hand, there are many cases where openness is neither feasible nor desirable—for example, health data or employee data. Thus, there are complex trade-offs between openness and privacy that you need to think through for each project. Consider the benefits when other people can build on the work you have done, and balance that with the potential for harm or loss of privacy.

Readability and legibility mean that it is important to write down your decisions and make them clear, so people don’t have to decipher or guess. Logs are a human-centered practice in part because they imagine a future user who will read those logs to understand what the data or the project “is about.”

- Have you ever come back to your old code after six months or more? Did you have enough comments to make sense of it? How much time did it take to become “fluent” again in what you did in that code?

Thinking about Your Process and Practice

One key thing that distinguishes human-centered data science is *reflexivity*—that is, actively thinking through your decisions and practices throughout the process. We discussed Donald Schön's work on reflective practice in chapter 3 (Schön 2002) and mentioned in previous chapters that many human decisions go into the data science life cycle. We add here some ways to be thoughtful about those decisions in each step.

We do many things in data science based not on a particular conscious decision but out of habit or convenience. These choices and decisions have an effect on the outcomes of the pipeline. For example, you might not have consciously decided to avoid commenting on your code, but a series of difficult decisions about alternative models may have taken all of your attention. And yet this lack of a conscious decision to prepare documentation will still affect the outcome: other people will have a much harder time using your pipeline and, as a result, may choose an ill-fitting dataset, misuse your models, or misinterpret their results.

Let's start with being reflexive about formulating a question for a data science project. It starts with thinking about how you arrived at a particular question. Did you read what has been done before? Did you consider other studies, possibly from other domains? Have you thought through how answering your question will affect the communities represented in it, or the people whose data you are using? Do you need to work directly with those communities or persons to understand their needs, concerns, and vulnerabilities in relation to your work? Those are all important questions to keep in mind when formulating or reformulating a question for your data science project (Mao et al. 2019). This will help you to make choices based on careful consideration and social responsibility, as opposed to falling into habits or traps of convenience.

As you recall from chapter 2, you need to formulate and document a measurement plan that lays out the steps for how you plan to measure the variables of interest for your question. You should also think through your options for when things go wrong. One likely possibility is that you will not be able to obtain the exact dataset you want to answer your question. If you have to switch to a different dataset, how will you have to change the measurement plan? What kind of bias might this switch introduce into your analysis? Will you have to modify your question, and by how much (and then, is it even the same question)? Thinking through these issues will help you avoid drifting from one question to another to the point of losing track of your original motivation.

- Think of a time when you had to change to a new dataset. How did your way of measuring things change? Did the new dataset map well on your original ideas? If not, how did you need to adjust?

Being reflexive about your data is extremely important but can be difficult, especially if you did not collect the data yourself. You will need to thoroughly consider questions such as: How was the dataset created? Who entered the data? For example, was it the people that the data is about, professionals speaking for individuals, or an automated system? And what kind of organizational power and politics may have been involved? Do you know what the data contains? Do you fully understand what the metadata mean? It is impossible to design a meaningful data science project without a thorough understanding of what each column of data represents, how the variables were measured, and what constraints were placed on the measurement. You might need to contact the people (or organizations) who collected the dataset to get to the needed level of understanding. And then of course: How well does the data relate to your question?

The data practices that shape data help us to think through multiple issues. This brings us to another aspect of thinking through your dataset—namely, whether it was your first choice of data for answering your question. Hopefully, if this was your originally intended dataset, you have concluded that it is the best fit for your question. But if it wasn't your first choice, there are more things to consider: Is this dataset the best *available* fit for your question? Or was it just the most convenient? How are you now planning to answer your question with it?

Being reflexive about your methods is not just the right thing to do, it also leads to better models and better pipelines. Have you thought carefully about what methods you plan to use? What problems might you anticipate with applying those methods? How might they have to be changed or tweaked? For example, you might want to summarize and group social media posts using topic modeling. In data science, the most popular topic modeling technique is called Latent Dirichlet Allocation (LDA), which we first described in chapter 4. But if you think about this method carefully and read about its applications, you will quickly learn that LDA is not very good for working with short documents such as social media posts, and it may not be the right tool for the job. Fortunately, there are other topic modeling methods that work much better for short texts. You might consider all the alternatives and choose the method that is best suited for your question (like the biterm topic model). Are you done? Not quite. The preprocessing you do for the new model might be different, so you'll need to adjust your pipeline for that. The kinds of inputs might be different. And the results produced might deviate from your expectations based on your LDA plan. For example, with the biterm topic model (BTM), you will have to do some extra math to calculate the distribution of topics per document, while LDA automatically does that (however, bypassing this step is exactly what makes BTM more suitable for short text).

Case Study 5.3

Data Curation at the La Brea Tar Pits: Supporting Data Science by Understanding Data Practices

Andrea Thomer, University of Michigan School for Information

The La Brea Tar Pits are a cluster of incredibly rich fossil deposits located in the heart of Los Angeles—and the home to a unique kind of "big data." An estimated three to four million ice age fossils have been excavated from these deposits since 1901, ranging from microscopic pollen spores to dire wolf skulls to enormous mammoth tusks. These fossils need curation in the traditional, physical sense of "museum curation": that is, each must be cleaned, cataloged, and stored in order to be preserved as part of our cultural and natural heritage and to be used in future scientific investigations. Furthermore, the *data* associated with each specimen needs to be curated as well. These data range from field notes documenting each fossil's original location in the deposit, to databases indexing the site's massive collections, to detailed anatomical drawings and computed tomography (CT) scans used in evolutionary biology studies, to protein sequences and radioisotope measurements derived from the fossils. These digital objects need as much care as the physical fossils to be used in data-intensive science and to ensure that they are accessible by future generations.

Data curation (also called *digital curation*) is the work of making data usable, sharable, accessible, and preservation-ready over its lifecycle of interest to science and scholarship. This work—and the people who do it—are foundational to data science, but often overlooked. Data curation includes cleaning, reformatting, annotating, and standardizing data for analysis; describing data for later retrieval or reproducibility; and taking steps to ensure that data is stored in a stable format and trustworthy repository. Data curators go by many aliases: data librarians, research data managers, data stewards, data janitors, data engineers, and more. They also hold diverse roles in research teams: working directly with scientists as part of a lab, working in data repositories, or consulting from posts in an academic library, or anywhere in between. Regardless of their name or position, though, they share a common cause of making data *fit for use* and accessible to broad audiences now and in the future.

The work of data curation is rarely one-size-fits-all. Curation must suit the intended use of a dataset and the organizational context, and it must take into account existing *data practices*—the workflows, cultures, and moral economies (i.e., the ethical values governing the data "marketplace") of data use in a community (Strasser 2006). At La Brea, this entails supporting both very old and very new ways of collecting data. Scientists at La Brea have collected specimen data in the same way since 1969, and the data collected via this legacy workflow needs to be supported going forward. However, they must also manage data collected with novel methods, such as those from recent studies of "food webs" (the networks of predator-prey relationships in an ecosystem) at the site. While the curators at La Brea have recently invested in a new collections management system, there are limits to what kind of data it can store, and they have had to augment their database through *ad hoc* catalogs stored in spreadsheets. There will likely never be one database that will magically "solve" their data curation needs, but rather, a rich sociotechnical ecosystem of curatorial systems and workflows.

Studying and understanding data practices isn't just important for the development of effective data curation protocols. Understanding data practices is also critical in surfacing the cultural norms, assumptions, biases, accidents, and perspectives that go into a dataset's creation. Because of the unique data collection method used by La Brea researchers, a dataset collected at La Brea includes variables and details that likely wouldn't be collected at another paleontological site. This doesn't mean one dataset is better or worse than the other—but rather, that the *people* who collected each dataset chose to record different data points and therefore emphasize (or obscure) different aspects of their study sites. Thus, studying data practices can help shed light on the ways that human choices shape data that may otherwise seem "objective" or "value neutral."

- Think of a time when you had to pivot from one method/model to another. How much work was entailed in preprocessing? What about lining up the model inputs with your pipeline? How different was the format of the results?

Our goal here is to make you aware that thinking critically about your process and how it affects your results, as well as the people whose behavior is represented in your dataset, is needed every step of the way. Thinking through your process may be even harder when some parts of your system are automated. In this case, you have less control over that part of your pipeline, as it may feel like an opaque box. We still encourage you to be as reflexive as possible about these parts of your pipeline by getting enough information about the automation to understand exactly what it does and how. Otherwise, you will be importing decisions wholesale, and you might be surprised by their outcomes in terms of whom they affect and how.

Platform Affordances and Data Schema

In thinking through your process, there is another source of potential trouble. Because so much of the data used in data science projects comes from various online platforms, we need to highlight the platform affordances and how they affect data schema.

Affordance is what the environment offers the individual. The term originally comes from psychology, but it was appropriated by people working in human-machine interaction. In that new context, affordances are the possible actions that technology makes available to someone based on how they perceive them in their environment (Nagy and Neff 2015). Here is an example from the realm of online platforms. Until a few years ago, the only way you could react to a friend's post on Facebook, outside of writing a comment, was to click the Like button. Clicking on the button was easy but not always very meaningful. If your friend shared some sad news and you wanted to show that you were paying attention and being sympathetic, a *Like* probably did not literally mean that you liked the bad news. But the affordances of the platform put you in a difficult position. How would your friend interpret your use of the Like button? Here the Like button is an affordance: it enables certain visible actions and constrains our space of possibilities for others (such as expressing more nuanced emotions). Understanding this, in 2016 Facebook introduced its Reactions feature, broadening the type of emotional reactions we could signify with a button click—and so broadening the affordances of its interface.

As you might imagine, the affordances of an interface—what you can easily do within online platforms—translate into the data we can collect from these platforms. Now, instead of one column documenting how many *Like*s each Facebook post got, we would have six different columns showing how many reactions of each type it received. Hence, the affordances of the platforms can be traced directly to the data schema of the data that we collect from them: how the data is organized, what columns are present, and what they measure. And, of course, this determines the kinds of questions you can answer and the kinds of analyses you can run with this data. It is important to think carefully about what the data schema from the platforms affords you—that is, what it allows you to do or restricts you from doing—as a data scientist.

How data is packaged, stored, and distributed makes certain analyses easier than others. For example, data you can get from various Twitter APIs is organized around individual tweets. At first glance, this is a reasonable organization. But as we will describe in more detail in chapter 6, people use the system conversationally. They refer to their own previous tweets (as you would in a conversation, expecting that others remember your previous statements) and to tweets by others (as you would build on what the other person said in a dialogue). From this perspective, isolated tweets are not very useful, and we need whole Twitter conversations to understand and model what people are saying. And yet very few data science projects are doing that, because reconstructing those conversations from isolated tweets is a lot of work. It's much easier to rely on individual tweets, which are directly available from the Twitter APIs.

What is directly available in the data schema greatly affects the types of analysis that are common. We all know how easy it is to pick the path of least resistance, to look under the streetlight. This means that we often choose the same easy proxies for the things we want to measure for our questions. And of course, these are not always the best or the right ways of measuring those concepts and answering our questions. So just like with other parts of the data science pipeline, we have to be very thoughtful about how the data schema affects our analyses.

- Think about a time you chose a concept to analyze because there was a column for it in the dataset. Were there other, better ways to answer your question? Were they more labor-intensive? How were your results affected by this choice?

Many people use social media data in data science projects. There are so many sources of information and so many loud voices there. Which ones do we attend to? Attention on social media is easiest to measure through the traces of what people do with the content: when they click to retweet, favorite, reply, or mention. Those metrics are directly available in the data schema, and they are easy to count. But they are also completely dependent on social media users taking action—pressing a button in the interface. What about measuring the attention that we pay to posts we just read, without a mouse click? If no action was taken, there is no digital trace, so this information is not recorded in the social media data. To measure that kind of attention, we would have to go beyond the easily accessible data from platform APIs to capture what people are doing through other means: eye-tracking studies, direct observation, or interviews. Of course, these other methods are more labor-intensive and time-consuming than working with existing datasets.

Carefully thinking through the constraints of the data schema, and how they affect your analysis choices, is an important part of being reflexive about your process. People in data science often use the following techniques for reflecting on their practice:

- *Ask others* in data science how they do their work and how they think about their work.
- *Observe others* as they do data science work and as they work with demos.
- *Ask* data science workers to lay out their work practices in detail, and to explain those work practices in detail as they lay them out.
- *Analyze code and documentation* that others write—or do not write—as they do their work.
- *Read* the online descriptions and instructions at popular sites that collect user-provided data.

Conclusion

In this chapter we introduced you to many things you should consider when designing and implementing your data science project. These things are what distinguish human-centered data science: formulating a meaningful question, considering issues of ethics and fairness, designing projects that others can easily build on, and incorporating reflexivity into your entire process. Many of these suggestions are concerned with people: the people represented in your dataset, the people whose lives may be affected by the results of your analysis, the people who might want to reuse your pipeline or just your results (such as classification labels), and even yourself at a later point in time. These strategies make your data science project more human-centered and sensitive to the people who are a part of practically every step of the cycle.

These strategies also make for a better data science project. Formulating an interesting question that matters to others and can be tested with well-suited data will make for a powerful finding. Carefully considering the constraints of the data schema lets you refine and answer more nuanced, more powerful questions. You will get stronger results because you are deliberate in what you are trying to measure when you have carefully planned your measurements. Thinking through your model, including considerations of fairness, will produce more precise and generalizable results and pipelines.

Ethics, like reflexivity, is not a one-time consideration. That is why recipes and checklists don't work in this domain: we never know when or what parts of our work will require careful and critical consideration. We emphasize reflexivity as a practice—something you do (and continue training yourself to do) at every step of the process. It is a skill that gets easier with time, as long as you are open to asking yourself difficult and critical questions.

Recommended Reading

Bourgeois-Doyle, Dick. 2019. “Two-Eyed AI: A Reflection on Artificial Intelligence.” Canadian Commission for UNESCO’s IdeaLab. https://en.ccunesco.ca/-/media/Files/Unesco/Resources/2019/03/TwoEyedArtificialIntelligence.pdf. This paper describes how members of an oppressed culture or group may need to maintain two ways of perceiving social realities: one in terms of the truths of their own culture, and the other in terms of the normative beliefs of the oppressing culture.

franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and the Association of Internet Researchers. 2020. “Internet Research: Ethical Guidelines 3.0.” https://aoir.org/reports/ethics3.pdf. This extensive ethics guide for responsibly using online and social media information was developed by an international organization of academic researchers.

Quinton, Sarah, and Nina Reynolds. 2018. *Understanding Research in the Digital Age*. London: SAGE Publications. This general guide to digital research is organized around the themes of ethics, expectations, and expertise. The book provides a way to think through a research project that focuses on digital data.

Smith, Linda Tuhiwai. 2013. *Decolonizing Methodologies: Research and Indigenous Peoples*. London: Zed Books. This work helps you to reconsider your own cultural frame as it influences your data science method. While not about data science as such, this book is very much about how we see other people—how we categorize them and how we make them visible or invisible in our analyses and our reports. Although it focuses on colonized peoples, its lessons apply to any situation that involves two or more groups with different degrees of social power.

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

UK Government Digital Service. 2020. “Data Ethics Framework.” https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923108/Data_Ethics_Framework_2020.pdf. This is an example of how a national-level strategy for the ethical use of data might be implemented. This framework document is also useful for its clarity and ability to explain data science ethics questions in practical and applied terms to those outside the data science profession.

6

Human-Centered Data Science Methods

This chapter outlines key traditions from social science, design studies, and critical theory to show other ways to study human behavior and how they could be incorporated into or combined with data science. The results of these combinations and experiments at the intersections of data science are quite exciting as they merge computational capacity with our abilities to interpret, build, and transform the social world we live in.

Social Science Methods for Rethinking Data Science

Emile Durkheim, a French philosopher, is credited with founding the field of sociology at the end of the nineteenth century. He worked at a time when society was undergoing enormous changes. In Europe, the rise of manufacturing and the use of more machines on farms meant many people were forced to move from close-knit villages to relatively more anonymous cities. Traditional village values were giving way to more connected, cosmopolitan, and global world views. Durkheim suggested that there were what he called "social facts," ways of acting, thinking, and feeling that are influenced by the communities we live in. These social facts, Durkheim hypothesized, have some influence over us (Durkheim [1895] 2014).

In 1897 Durkheim published *Suicide*, one of the first books that used data to explain the behavior of people in society. Durkheim wanted to test theories about social cohesion as a social fact. How much social cohesion was needed to hold society together? How much was too much? The problem was that social cohesion and social integration were difficult to measure. Durkheim's idea—that cohesion was an attribute of groups instead of being an attribute of individuals—was interesting. But he couldn't *see* social integration, so how could he measure it?

He did it by explaining patterns in variation, and building those patterns into larger explanations, not unlike the process of modern data science. However, Durkheim painstakingly hand-calculated death rates using death certificates from different cities and countries. In this data, Durkheim found sets of patterns: substantial differences between men and women, between Catholic and Protestant communities, between city and countryside. People who underwent major life changes and those who lived in rapidly changing places were more at risk for suicide. Men whose wives had recently died were more likely to commit suicide than women whose husbands had died, and they were more likely to commit suicide than men who had never married. Older people were more at risk for suicide than younger people. These patterns all suggested that a third variable—changes in life circumstances—had some significant influence on a person's tragic, seemingly deeply personal choice. If suicide were simply a function of individual psychology, these patterns would be difficult to explain. Durkheim used these differences to develop a hypothesis about how people are influenced by the rules of everyday life. Too many or too few rules, or too much or too little integration with others, put people at a greater risk for suicide. Durkheim used simple tables and charts to map these differences in variation and to create a theory about the role of social rules on people's lives (Durkheim [1897] 1979). His data science tools were rudimentary, but the spirit of his approach to demographic data is very much in keeping with the spirit that motivates data science today.

Social science has over a hundred years of experience with studies to explain human behavior. Our goal in this chapter is not to convert readers into social scientists. Rather, we want to show how social science and other methods can bring a human focus to data science, while pointing interested readers to the resources they need to dive deeper if they wish. We introduce a range of methods so that these approaches can be considered in designing, evaluating, and feeding back into the data science pipeline. They can be thought of as ways to inform projects with the contextual information that is often missing from large-scale data;