Using Data Legally and Ethically

Legal Use of Datasets: Can I Use This Dataset?

In the age of big data, we are spoiled with large amounts of information that is accessible online and can be used in a number of ways. However, just because information is accessible does not mean it can be legally and ethically used. In this section, we'll review the laws that govern how, when, and by whom datasets can be used.

What Is Copyright Law?

When determining how a dataset may be used, the first legal protection we should look into is **copyright law**. In the United States, copyright law grants authors of "original works" rights over how their works can and cannot be used by others. Under copyright law, authors have exclusive rights to make and sell copies of their works, to perform or display their works, and to create other work derived from that original work. These rights are automatic; an author does not need to fill out any paperwork or go through a legal process in order to claim these rights.

So, what exactly constitutes an "original work?" U.S. copyright law provides some guidance on this topic by listing examples of works of authorship, such as the following:

- Literary works, such as books
- Musical compositions
- Dramatic works, like plays
- Choreography
- Visual art, including photography
- Audio and video recordings
- Architectural designs

However, the law does not limit copyright protection to the list of original work it provides. To make things more complicated, it specifies that the copyright protection of a work does NOT apply to the “ideas, concepts, or principles” that are introduced in that work.

One of the main goals of copyright law is to motivate innovation. It does this by:

- Giving a financial incentive to authors by allowing them to have an automatic legal monopoly over their original works; and

- By specifying that the ideas in the work can be referenced and built upon by others, as long as they don't infringe on the original work itself.

Distinguishing Facts and Frameworks from "Original Work"

Now let’s examine how the distinction between an original work and the ideas within the original work might play out in practice.

Let’s say that you like to unwind from a long day of cleaning messy datasets by experimenting with new ways of making sourdough breads. One day, inspired by a gluten-intolerant friend, you invent a genius new technique for making gluten-free sourdough bread that has the same airy-yet-substantial, chewy-yet-soft

texture of traditional sourdough bread. It's a complex process with some difficult-to-find ingredients, but it's worth it for the end result. You take studious notes, write up a recipe with detailed instructions, and post it on your blog. A few weeks later, you stumble across a recipe that uses your same method on the Bob's Red Mill baking blog with no attribution to you. Enraged, you get in touch with a lawyer friend of yours to find out if this blatant poaching of your idea is legal. Take a moment to think: What will your lawyer friend say?

After reviewing both blog posts, he shares the unfortunate news that Bob's Red Mill did not infringe on your copyright protection. Why is this the case? Although their recipe clearly drew on the methods that you pioneered, the author's written introduction to the recipe was original material, and the instructions for making the bread were phrased differently. There were even some slight changes to the ingredient list.

Your friend tells you that, while the text of your blog post is protected by copyright, the methods and ideas that you shared in the post are not. Even though he thinks that Bob's Red Mill *should* have attributed that process to you, copyright alone does not legally require them to. It's possible that in a court of law, a judge could interpret the law differently, but it's not likely.

Can a Dataset be Protected by Copyright?

So, what does this mean when the work in question is a dataset or database rather than a recipe? You will notice that the word “data” is conspicuously absent from the list of protected works in the copyright law. But this has not stopped companies from claiming that their databases are protected in a court of law.

The results of these court cases form **legal precedents**, which are decisions that are used as an authority for interpreting the law’s legal challenges. Judges are likely, but not guaranteed, to follow these precedents when it comes to interpreting whether or not a dataset can be considered an original work. If deemed an original work, it would be protected by copyright law. If not, the dataset will be identified as a collection of facts that can be freely used by anyone.

Feist v. Rural

One of the cases that set a legal precedent relevant to dataset copyright protections is **Feist v. Rural**, decided in 1991 by the U.S. Supreme Court. In this case, Rural Telephone Service Company, Inc., sued Feist Publications, Inc., for copyright infringement after Feist published a phone book that used Rural's subscriber information without its consent. The court ruled in favor of Feist, saying that Rural's subscriber list was a collection of facts and was not copyrightable.

According to the ruling, facts “are not original and therefore may not be copyrighted." A set of facts can only be protected if it features “an original selection or arrangement of facts.” The facts themselves may never be protected by copyright.

For someone working with data, this ruling has benefits and drawbacks. It facilitates the use of datasets and databases created by others without fear of copyright infringement, which can have significant legal consequences. However, it also means that the hours of work you might spend compiling a dataset from disparate sources, cleaning that data, and organizing and documenting it to make it usable, will likely not be protected by copyright.

You may attempt to protect your work by adding something original to your dataset, such as a ranking of your entries or a similarity index, which many companies do in an effort to prevent unauthorized use. However, originality is subject to interpretation and debate, and it's not always clear how a court will interpret attempts at originality in a collection of data.

Because of the legal precedent set in Feist v. Rural, many companies have turned to using contract law rather than copyright law as a way to protect the investment they have made in compiling and curating data.

What Are Contracts and Licenses?

A **contract** is an agreement that can be enforced by the law. Although you may think of contracts in familiar situations like buying a house or an arrangement for freelance work, you likely are interacting with contracts on a daily basis as you use different software products, websites, and datasets. **End-user license agreements,** which may also be called **terms of use** or **terms of service**, are contracts that most companies require you to agree with in order to use their products.

Sometimes, these contracts require explicit consent, such as the multi-page user agreement that requires approval when you update your phone. In other cases, the license is provided as a disclaimer, such as certain cookie notices or the licenses that you often see on GitHub repositories.

People have legally challenged the idea that a dataset can be protected by a contract that is not easily accessible for the average person to read, understand, or even, in some cases, locate. However, the courts have tended to support the legal enforceability of terms-of-use agreements that do not require explicit consent.

ProCD, Inc. v. Zeidenberg

An important legal precedent was set in the 1996 U.S. Court of Appeals Seventh Circuit case **ProCD v. Zeidenberg**. ProCD was a company that compiled the information from phone books across the country into a comprehensive computer database, which they sold as a physical CD-ROM. This CD-ROM included in its packaging an end-user license that restricted the use of the dataset to non-commercial purposes, among other limitations. Matthew Zeidenberg purchased this database, published its information online, and sold access to the online version of the database.

Zeidenberg argued that the end-user license was not a valid contract because it was only made available to him after he purchased the database, as it was within the shrink-wrapped CD case. The court did not agree with this argument, and ProCD won the lawsuit. This decision reinforced that there are many ways that a contract can be agreed upon, which is clearly expressed within the legal opinion:

> "A vendor, as master of the offer, may invite acceptance by conduct, and may propose limitations on the kind of conduct that constitutes acceptance. A buyer may accept by performing the acts the vendor proposes to treat as acceptance."

Because of this interpretation of contract law, it's crucial to search for and thoroughly read the terms of service or license of any dataset before using it. There may be restrictions on the context in which the dataset can be used that are important to consider before investing time and effort in analyzing it or using it to train an algorithm.

This can add some frustrating red tape to your work, but this interpretation of contract law can also be to your benefit. Before publishing any of your work as part of a GitHub portfolio or package, be sure to add a license that specifies how your work may be used by others.

Now let's switch gears from vetting the protections on an existing dataset to considering the legality of creating your own dataset via web scraping.

The Legality of Web Scraping

Another important consideration for determining if a dataset may be legally used is how it was created. Many publicly available datasets were created by **web scraping**, or the process of extracting information from websites. Web scraping is a powerful way to gather information. However, like many things in the world of data science, it has existed in a legal gray area as regulations and the courts have been catching up to advances in technology that allow for automated web scraping.

In many cases, datasets created by using web scraping are legal to use. However, there have been cases in which courts have ruled unauthorized web scraping to be a criminal act. The **Computer Fraud and Abuse Act (CFAA) of 1986** is a U.S. federal law created to prevent hacking. The CFAA is a criminal law, which means that the government can use this law to charge people with crimes. The law also includes civil provisions, which means that non-government entities (such as companies and individual people) can use the CFAA to charge people with hacking in civil courts.

In other words, this is a law that you don't want to break! And because of the many provisions in the CFAA and its history of being broadly interpreted, there are quite a few situations when scraping a website could be a violation of the CFAA.

The CFAA includes seven main offenses with associated sentences. In plain language, all of these offenses involve intentionally accessing a computer without authorization and obtaining information from any protected computer.

Is Web Scraping Legal?

To determine if the CFAA applies to work you are doing or a dataset you want to use, ask yourself the following questions:

- Are you accessing a computer?

  The law defines a computer as "an electronic magnetic, optical, electrochemical, or other high speed data processing device performing logical, arithmetic, or storage functions, and includes any data storage facility or communications facility directly related to or operating in conjunction with such device." The definition explicitly excludes devices such as typewriters and handheld calculators, but it is broad enough to include many devices that the average person would not think of as a computer, such as smart appliances and some children's toys. Essentially, anything with a microchip might be considered a computer under this definition.

- Are you authorized to access the computer in the way that you are accessing it?

  Accessing a computer is defined differently depending on whether the user is accessing the computer in person or virtually. Physical access could be something as seemingly innocent as turning the computer on and viewing the login screen. Virtual access usually requires entering a username and password, or otherwise circumventing the login process. Authorization is subject to interpretation depending on the context. Unauthorized access could refer to someone who has broken into a system using technical measures, such as guessing passwords or circumventing security measures. Access could also be unauthorized if it exceeds authorized access by violating a contract, such as an agreement between an employer and employee, or a social norm related to typical computer use.

- Are you obtaining information from a protected computer?

  The standard for obtaining information is very minimal; simply viewing the login screen could be considered obtaining information. A protected computer is defined in the law as any computer used by the government or a financial institution, any computer that is part of a voting

> system, or, most importantly, any computer that affects interstate or foreign commerce or communication. This means that any computer connected to the internet is a protected computer.

To summarize, web scraping may be illegal according to the CFAA when it involves unauthorized access to a computer, such as a web server, that is connected to the internet. However, whether or not an instance of web scraping is illegal is ultimately determined by a court of law.

Web Scraping Criminalized Under the CFAA

Like copyright and contract law, the CFAA is subject to interpretation by the justice system when it comes to determining whether or not scraping a site is in violation of the law. In this section, you'll learn about an important court case that has set legal precedents about what kinds of access may be considered unauthorized. Cases like this one should be considered alongside the law in judging whether a web scraping project is legal, or if a dataset was created legally.

Facebook, Inc. v. Power Ventures, Inc.

The 2008 case of Facebook, Inc. v. Power Ventures, Inc. was an early and prominent example of web scraping being labeled as illegal under the CFAA. Power Ventures was a social networking company that aggregated information from multiple social media sites on one website and allowed users to post to all of them at once. In the lawsuit, Facebook alleged that Power Ventures obtained information from Facebook servers when they were not authorized to do so. Power Ventures had collected usernames and passwords from Facebook users who opted into the social media aggregation services that Power Ventures offered. Initially, they accessed user information by using Facebook Connect to scrape information that was on Facebook and display it on their website: www.power.com.

Facebook ultimately won this lawsuit and was awarded damages for losses incurred during the time that Power Ventures was accessing Facebook's servers without authorization. In the lawsuit, a debate over when Facebook could begin incurring damages led to an interesting legal precedent.

According to the final ruling Power Ventures "was actively evading Facebook's attempts to block IP addresses associated with Power from accessing Facebook's servers," even after receiving a cease and desist letter from Facebook. This act of evasion was at the center of Facebook's case against Power Ventures.

The legal precedent set here is that attempts to evade security measures can in themselves be seen as evidence of unauthorized access. For this reason, if a website has security measures in place that are intended to evade web scraping, scraping the website is very likely illegal.

Determining what is legal and what is illegal can be challenging, because new legal precedents can be set that may change how the law is likely to be interpreted. For this reason, it's wise to be conservative. The case of Facebook, Inc v. Power Ventures, Inc. is one of many that have penalized certain instances of web scraping. However, there are other cases and legal precedents that limit how web scraping can be criminalized under the CFAA. The decision in the case of LinkedIn Corp. v. hiQ Labs, Inc. has been interpreted by many as a victory for the legality of web scraping. However, this ruling may not be final. In July 2021, the Supreme Court asked the Ninth Circuit to revisit this decision. This article has a more detailed summary of the ruling and what's next for this court case.

As the case of Facebook Inc. vs. Power Ventures shows, just because you can scrape a website doesn't mean you should. Power Ventures was effectively shut down because of the way they scraped Facebook user data. Unfortunately, because the law on web scraping is complex and changing, it's not always clear when a web

scraping project is illegal. Before you scrape a website, do some research on recent legal findings around web scraping.

Next, we'll summarize what you should consider before using or creating a dataset, then check that you've understood the key points of this lesson.

A Checklist for Legally Using Datasets

To summarize what we've learned, here is a checklist that you should follow whenever you plan to use a dataset or create a new dataset based on other sources (such as existing datasets or information scraped from websites):

1. Check for copyright protections by asking if the dataset:

    - Includes information that could be considered an original creation rather than a fact; or
    - Is organized or structured in a creative and different way.

    If the answer to either of these questions is "yes", this dataset may be a copyrighted work. Review copyright law and make sure that the way you plan to use this dataset is within the bounds of fair use.

2. Document how you intend to use this dataset, now and in the future. Find any licenses or terms of use associated with the dataset and review them to confirm that your intended use is in compliance.

3. Investigate how the dataset was collected. Identify any indicators that the data was obtained from a source that the compilers were not authorized to access.

4. Research the current legal rulings on web scraping to determine if a scraped dataset is likely to be considered illegal. Because the law in this area is still in flux, be cautious not to violate the terms of use of any website that you scrape, and avoid scraping data that includes personally identifiable information.

Now that you've learned about the legal issues around using datasets, let's think about the ethical issues you should consider before using a dataset.

Should I Use This Dataset?

As a data scientist or analyst who works with a wide range of vast datasets every day, it's often easy to forget about the individuals whose lives are documented within that dataset. Datasets frequently include information about people or are used to inform decisions that impact people's lives. In this section, you'll consider how using different datasets can impact people both positively and negatively, and how to use data ethically.

Making Ethical Decisions

Data ethics is a growing field in which academics study the impact of big data on people and society. The term **ethics** refers to a system of principles that determine how people make decisions about what is good to do and what is bad to do. Data ethics is the application of ethical principles to data and technology. This field is rapidly expanding because there are many instances in which the growing applications of big data have caused various types of harm to people. The technical complexity of the systems that use big data can

make it challenging to understand how these problems were created and, thus, how to prevent future problems.

The good news is that, although the problems can be complex, being aware of the major ethical concerns around using data can help you use data to design technology that changes people's lives for the better.

In the following sections, we'll focus on some of the major ethical considerations when working with data. These considerations include privacy, consent, and issues of exclusion. We'll start by learning about privacy.

Privacy

One of the most important ethical considerations when working with data is privacy. Privacy is a complicated topic with many rules and regulations depending on the context and the industry. We'll examine some of those regulations later but, for now, we'll focus on what privacy means, why it matters, and how to protect it.

Before we define privacy in the context of big data, take a moment to consider the following questions:

- What are some circumstances in your life when you want privacy?
- When did you feel that your privacy was violated in person? How about online?
- When might you have violated someone else's privacy?
- Keeping your answers to the previous questions in mind, how would you explain the concept of privacy to someone?

In many cases, a violation of privacy is easier to identify than a formal definition of what privacy is. This is because privacy is defined differently depending on the situation, people involved, social norms, and other factors.

For example, someone peeking over your shoulder to see what you are reading might feel like a violation of privacy to you if what you're reading is a self-help book on overcoming divorce. On the other hand, it might feel like a simple expression of curiosity if you're reading a novel.

Generally speaking, you can think of privacy as a person's right to control how information about them is shared. But you should always be aware that the bounds of privacy will differ in different contexts, and that it is as much of a social issue as a legal one. Furthermore, your expectations of privacy may be different than your friend's expectations of privacy, as well as anyone who is from a different geographic, socioeconomic, or cultural background.

Almost any type of information can bring up privacy concerns. However, privacy specialists take special care to protect personally identifiable information, which we'll learn about in the next section.

Personally Identifiable Information

To protect individual privacy, there are some universal standards for handling personal information.

Personally identifiable information (PII) is a term used to refer to information that can pinpoint the identity of a specific person. Someone's name is an obvious example of PII; less obvious is someone's birthday. Here are some examples of PII provided by the National Institute of Standards and Technology (NIST):

- Full name
- Personal identification numbers, such as Social Security number, driver's license number, or passport number
- Directory information, such as address, email address, and telephone number
- Technical identifiers such as IP address
- Personal characteristics, such as an image of a person, recording of a voice, or other biometric information
- Information that can be linked to any of the above, such as date of birth, race, employment information, and education information

There are rules and regulations for handling data that vary by industry and geographic area. We'll cover those regulations later, but, for now, know that you should always closely review any dataset you are working with to determine whether it includes PII. If it does, you will need to take precautions in order to protect the privacy of individuals included in your data. Depending on how you plan to use the data, this may mean anonymizing the dataset by removing any PII or putting limits on who can access the data and in what settings.

Anonymizing data by removing or concealing PII is an important strategy to protect individual identities. However, because information is interconnected and anonymized datasets can be linked with publicly available information, removing PII may not always be enough to protect privacy.

Anonymized Data and Identification

In NIST's list of personally identifiable information, the last item is not a specific category of information but rather a broad list of types of information that could be linked to an individual. This is important to highlight because privacy researchers have demonstrated that there are numerous ways to identify specific people in a dataset by using information that might seem anonymous.

Privacy scholar Dr. LaTanya Sweeney has repeatedly demonstrated how easy it is to identify specific people from an "anonymized" dataset, exposing significant systemic flaws and threats to privacy in large record systems. In the late 1990s, after the Massachusetts Group Insurance Commission publicly released "anonymized" data on hospital visits by state employees, Dr. Sweeney showed that she was easily able to identify a specific person within these records by finding the health records of the governor of Massachusetts at the time, Bill Weld. She did this by simply linking Cambridge voter-registrations with information in the hospital visits dataset.

Although Dr. Sweeney's work has led to reforms in privacy laws and regulations, her team at Harvard has continued to find instances where joining two publicly available datasets leads to identifying sensitive information about individuals.

Dr. Sweeney's Data Privacy Lab at Harvard created this tool to show how easy it is to identify a person by using basic information. If you feel comfortable doing so, try inputting your information into the tool to see how unique that combination of information is. What did you find? How does this result make you feel?

Now think about how many times you have shared details like your birthday and ZIP code. Think about all of the datasets you have seen or used that include this information. Do you feel like your data is protected? How do you feel about your privacy?

Now we'll switch gears to learn about a related, but distinct aspect of data ethics: consent. You may feel more comfortable revealing private information if you actively chose to do so by providing your consent. But what is consent, exactly? We'll learn about this in the next section.

Consent

Another important ethical consideration related to privacy is **consent**, or agreeing for something to happen. An important part of having privacy is being able to control what happens with our information. We can exert control over our information by:

- Providing consent, or permission, for our information to be used in a certain way; and
- Refusing to consent, or refusing to give permission, for our information to be used in a different way.

Like privacy, the concept of consent has a long history in many different contexts. **Informed consent,** which is defined as knowing possible consequences when you agree that something can happen, is a critical part of medicine, law, and research, and these fields have rigorous standards for what constitutes informed consent.

The **Belmont Report** was written in 1979 to provide ethical guidelines for research performed on human subjects. Although data analytics and data science may not always constitute research, the guidelines it presents are a helpful starting point for thinking through ethical concerns regarding the use of data.

The Belmont Report

The Belmont Report centers around three basic ethical principles: **respect for persons**, **beneficence**, and **justice**.

1. Respect for Persons

This principle refers to the following:

- Individuals are autonomous, which means they are capable of making decisions about their own lives and goals, and acting on those decisions and goals.
- Individuals who have less autonomy for some reason, such as their age, illness, mental disability, or other impairment, should be protected from harm

2. Beneficence

This principle refers to the following:

- Treating people ethically means not only respecting them as autonomous and/or protecting them from harm, but also putting in effort to secure the well-being of people; and
- Acting beneficently, not causing harm to people, maximizing possible benefits, and minimizing possible harms.

3. Justice

This principle refers to the following:

- Burdens and benefits of a project should be distributed in a manner that is considered to be fair.

- There are different approaches to fair distribution. In general, take care to review who benefits from your work and who is harmed by it. Avoid situations where the benefits of a project primarily go to those in power, such as the wealthy or high-status, and the risks and harms rest on those without power, such as the poor or imprisoned.

The Belmont Report and Informed Consent

The Belmont Report considers informed consent to be a critical component of the "respect for persons" ethical principle. It stipulates that, in order for someone to give informed consent to something, they must:

- Have information about what they are consenting to;
- Understand that information as well as the consequences of their consent; and
- Not be pressured to consent in any way.

When you work with a dataset, think about the people whose information is included in that dataset. Did they consent to being in that dataset? If so, did they understand the different ways that their data may be used, now or in the future?

Let's review some examples of when information was used without the kind of informed consent that the Belmont Report recommends.

Examples of Consent Issues

Issues of consent are common in data science because as technology continues to advance, there are more and more interesting, innovative, and exciting ways to use data. However, this can result in people's information being used in ways that they didn't consent to, with unanticipated consequences.

In this section, we'll examine two situations in which information was used without appropriate consent and consider the consequences. Learning about examples like these can help you recognize situations where lack of consent could lead to harm.

Genetic Information

There are many cases that show how genetic information has been used without consent. One of the most famous cases is that of Henrietta Lacks, whose cancerous cells were taken without her consent at a doctor's office in 1951. They have subsequently been used for medical research for years after her death, continuing to this day.

Genetic information is sensitive and is shared within a family, not just an individual. Henrietta Lacks was never given a chance to consent to her cells being used for research, but even if she had, she never could have known that:

- Her cells could be replicated indefinitely.
- Her genome could be sequenced.
- All of her future relatives could easily be identified through her donation of cells.

This violation of Henrietta Lacks' right to consent has brought huge advancements to medicine and helped millions of people. Her cells have been used to develop vaccines, understand cancers, and manufacture drugs to treat a number of different illnesses.

Her cells have also enriched pharmaceutical industries and brought fame and adulation to scientists and research organizations, without acknowledgement of Henrietta Lacks herself or her family. Her children did not find out that her cells had been used until 1973, over two decades after her cells were taken.

Because Henrietta Lacks' family were never given an opportunity to consent to the removal and use of her cells, they were not able to share in the profits or to celebrate the impact their genetic line had on the world. Instead, they are part of a dark legacy in American medicine of using African Americans for medical experiments without their knowledge or consent.

Read more about Henrietta Lacks and her legacy at the Henrietta Lacks Foundation website, and in the 2010 book *The Immortal Life of Henrietta Lacks*.

Facial Recognition

With all the consent forms you've likely signed in medical offices over the years, the Henrietta Lacks story may seem like a relic of the past. But compare that case to this story about Clearview AI, which was fined $22.6 million in November 2021 for using images collected via web scraping without individuals' consent to build a facial recognition algorithm.

Facial recognition is in its infancy, just like genetic research was in the 1950s when Henrietta Lacks' cells were taken. The people behind the faces included in Clearview AI's training dataset never had a chance to consent to being used as part of a facial recognition system. They also have no way of knowing how this technology might evolve over time and the impact their inclusion may have on their lives.

Our inability to predict the future of technology should not prevent us from pursuing interesting projects that can change the world for the better. But in order to work ethically and respect people, we need to do our best to use data gathered from people who have consented to their information being used in the way that we intend to use it.

So far, the ethical concerns we've discussed have mostly related to the people represented in a dataset. That is, what if being included in this dataset reveals who someone is and that results in sensitive information being leaked? What if someone didn't consent to being included in this dataset? What if they did consent, but they didn't understand how that information might ultimately be used? Next, we'll switch gears by thinking about what can happen if a dataset excludes a group of people.

Issues of Exclusion

There are many ways that data can be used, but when you simplify most data projects, they are almost always using data, or information, to make a decision. Which movie should I watch on Netflix? Is this person likely to pay back a loan? Is this tumor cancerous? These are all questions that can be answered by using data to make a prediction.

When you are using data to inform decision-making, it's critical to ask yourself the following questions:

- Who will be impacted by the decisions that are made based on this data?
- Are all of those people represented in this dataset?

Consider the example of an algorithm that determines if a tumor is cancerous or benign. This algorithm would have a huge impact on medical decision-making, so its accuracy is extremely important. A false negative, meaning a cancerous tumor that the algorithm thought was benign, could cause a patient to die if a doctor does not pursue treatment.

Algorithms like this one usually make their decisions based on patterns they have found in a dataset. If a tumor looks a lot like tumors that had been confirmed to be cancerous, the algorithm will guess that this tumor is cancerous. If a tumor looks a lot like tumors that are known to be benign, the algorithm will say that the tumor is benign.

Imagine that your cancer-predicting algorithm made its decisions based on a dataset of labeled skin cancer images. Would you want a doctor to use this algorithm to determine if a tumor in the lung was cancerous? Hopefully you can answer that quickly: no!

The decisions that this algorithm makes are high stakes. You don't need to be an oncologist to suspect that skin cancer and lung cancer tumors are likely to be different from each other. Using an algorithm trained on skin tumors to predict cancer in lung tumors is very likely to harm people who have lung tumors.

This is an example of what can happen when data-driven decisions are made based on data that excludes or leaves out people impacted by those decisions. People with lung tumors were not included in the skin tumor dataset used to determine if tumors were cancerous. Consequently, they would be harmed by decisions that were made solely based on that dataset.

Examples like this are clear-cut but, in reality, issues of exclusion are not always obvious to people who are making data-driven decisions. Let's review an example of a well-intentioned project that still managed to cause harm to people who were excluded from the data used to make decisions.

Favoring the Young and Rich by Excluding the Old and Poor

In 2012, the city of Boston released an innovative new project called Project Street Bump, which was designed to help fix potholes in the city. Project Street Bump was a smart phone app that Bostonians could use to collect data about road conditions while driving. That data was then shared with the city government so that it could fix problems with roads. The city also used this data to inform planning of long-term projects.

Sounds cool, right? But think a little bit more about this project.

Question: In 2012, what types of people do you think owned smartphones? Of the people who had smartphones, who do you think would be most likely to use an app like this?

Answer: If you guessed younger, wealthier people, you'd be right. Initially, after the Street Bump app was launched, reports were primarily coming from higher-income neighborhoods. The makers of the Street Bump app were able to identify this problem early on and work to correct this bias. Had they not identified this problem, city resources would have quickly become distributed inequitably.

Drawing inferences from datasets can feel magical and powerful, and it is. But to use that power responsibly, we must remember that there are many different types of people in the world.

Before you make decisions or design technology based on data, consider who is included in that data and who is likely excluded from that data. Find creative ways to make sure that your work accounts for everyone it will impact, not just those who are represented in your work.

Checklist for Ethically Using Data

Follow this checklist when collecting and analyzing data:

1. **Review your dataset for any personally identifiable information (PII)**. If it contains PII, remove it or otherwise anonymize your dataset, and control who is able to access the data. Be aware that it's almost impossible to truly anonymize a dataset, as anonymized datasets can frequently be linked with other datasets to identify the people included within it.

2. **Investigate how your dataset was collected**. Ask yourself questions like the following: Did the people whose information is represented in the data consent to being included in the dataset? Did they consent to their information being used in this way? If you can't confirm consent, try finding a different data source or contacting the people included to obtain informed consent. As you make decisions about your work, be sure to consider how it may harm or benefit the people in the dataset or their relatives.

3. **Think about who the people represented in your dataset are, and how they relate to the people who your work is likely to impact**. Consider questions such as the following: Are the people in your dataset representative of the full population? Are they from different socioeconomic classes and racial backgrounds? Are there different cultures represented? Make sure that you consider what types of people may have been excluded from your dataset.
