4

Data Sources and Biases

The mission of the (fictional) non-profit organization Unconditionally is charitable giving. It collects donations and distributes unconditional cash transfers—funds with no strings attached—to poor households in East Africa. The recipients are free to do whatever they like with the money. Unconditionally is undertaking a new machine learning project to identify the poorest of the poor households to select for the cash donations. The faster they can complete the project, the faster and more efficiently they can move much-needed money to the recipients, some of whom need to replace their thatched roofs before the rainy season begins.

The team is in the data understanding phase of the machine learning lifecycle. Imagine that you are a data scientist on the team pondering which data sources to use as features and labels to estimate the wealth of households. You examine all sorts of data including daytime satellite imagery, nighttime illumination satellite imagery, national census data, household survey data, call detail records from mobile phones, mobile money transactions, social media posts, and many others. What will you choose and why? Will your choices lead to unintended consequences or to a trustworthy system?

The data understanding phase is a really exciting time in the lifecycle. The problem goals have been defined; working with the data engineers and other data scientists, you cannot wait to start acquiring data and conducting exploratory analyses. Having data is a prerequisite for doing machine learning, but not any data will do. It is important for you and the team to be careful and intentional at this point. Don't take shortcuts. Otherwise, before you know it, you will have a glorious edifice built upon a rocky foundation.

> "Garbage in, garbage out."
>
> —Wilf Hey, computer scientist at IBM

This chapter begins Part 2 of the book focused on all things data (remember the organization of the book shown in Figure 4.1).

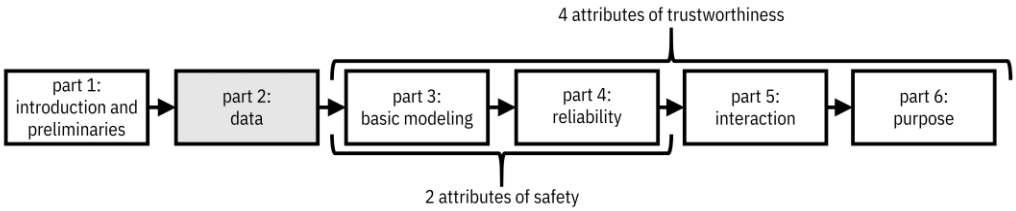

Figure 4.1. *Organization of the book. This second part focuses on different considerations of trustworthiness when working with data.* Accessible caption. A flow diagram from left to right with six boxes: part 1: introduction and preliminaries; part 2: data; part 3: basic modeling; part 4: reliability; part 5: interaction; part 6: purpose. Part 2 is highlighted. Parts 3–4 are labeled as attributes of safety. Parts 3–6 are labeled as attributes of trustworthiness.

The chapter digs into how you and Unconditionally's data engineers and other data scientists should:

- use knowledge of characteristics of different data modalities to evaluate datasets,
- select among different sources of data, and
- appraise datasets for biases and validity.

Appraising data sets for biases is critical for trustworthiness and is the primary focus of the chapter. The better job done at this stage, the less correction and mitigation of harms needs to be done in later stages of the lifecycle. Bias evaluation should include input from affected individuals of the planned machine learning system. If all possible relevant data is deemed too biased, a conversation with the problem owner and other stakeholders on whether to even proceed with the project is a must. (Data privacy and consent are investigated in Chapter 5.)

4.1 Modalities

Traditionally, when most people imagine data, they imagine tables of numbers in an accounting spreadsheet coming out of some system of record. However, data for machine learning systems can include digital family photographs, surveillance videos, tweets, legislative documents, DNA strings, event logs from computer systems, sensor readings over time, structures of molecules, and any other information in digital form. In the machine learning context, data is assumed to be a finite number of samples drawn from any underlying probability distribution.

The examples of data given above come from different *modalities* (images, text, time series, etc.). A modality is a category of data defined by how it is received, represented, and understood. Figure 4.2 presents a mental model of different modalities. There are of course others that are missing from the figure.

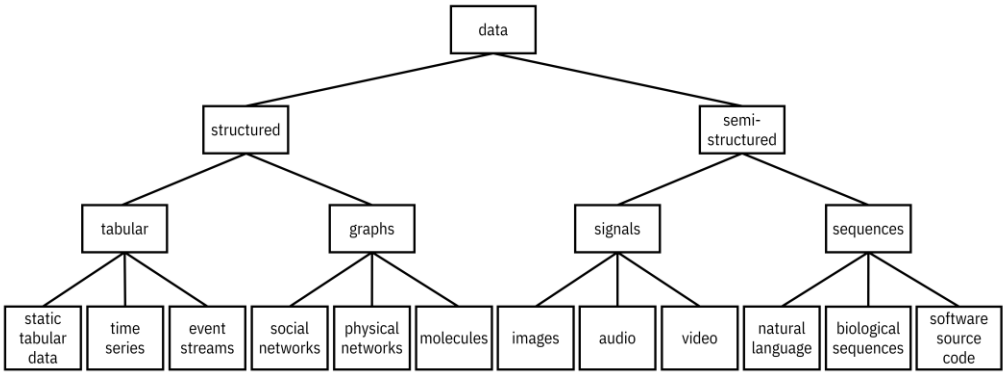

Figure 4.2. *A mental model of different modalities of data.* Accessible caption. A hierarchy diagram with data at its root. Data has children structured and semi-structured. Structured has children tabular and graphs. Tabular has children static tabular data, time series, and event streams. Graphs has children social networks, physical networks, and molecules. Semi-structured has children signals and sequences. Signals has children images, audio, and video. Sequences has children natural language, biological sequences, and software source code.

One of Unconditionally's possible datasets is from a household survey. It is an example of the *static tabular data* modality and part of the *structured* data category of modalities.[1] It is static because it is not following some time-varying phenomenon. The *columns* are different attributes that can be used as features and labels, and the *rows* are different records or sample points, i.e. different people and households. The columns contain numeric values, ordinal values, categorical values, strings of text, and special values such as dates. Although tabular data might look official, pristine, and flawless at first glance due to its nice structure, it can hide all sorts of false assumptions, errors, omissions, and biases.

Time series constitute another modality that can be stored in tabular form. As measurements at regular intervals in time (usually of numeric values), such data can be used to model trends and forecast quantities in time. *Longitudinal* or *panel* data, repeated measurements of the same individuals over time, are often time series. Household surveys are rarely longitudinal however, because they are logistically difficult to conduct. *Cross-sectional* surveys, simply several tabular datasets taken across time but without any linking, are logistically much easier to collect because the same individuals do not have to be tracked down.

Another of Unconditionally's possible datasets is mobile money transactions. Time stamps are a critical part of transactions data, but are not time series because they do not occur at regular intervals. Every mobile money customer asynchronously generates an event whenever they receive or disburse funds, not mediated by any common clock across customers. Transaction data is an example of the *event stream* modality. In addition to a time stamp, event streams contain additional values that are measured such as monetary amount, recipient, and items purchased. Other event streams include clinical tests conducted in a hospital and social services received by clients.

[1]There are modalities with even richer structure than tabular data, such as graphs that can represent social networks and the structure of chemical molecules.

Unconditionally can estimate poverty using satellite imagery. Digital *images* are the modality that spurred a lot of the glory showered upon machine learning in the past several years. They are part of the *semi-structured* branch of modalities. In general, images can be regular optical images or ones measured in other ranges of the electromagnetic spectrum. They are composed of numeric pixel values across various channels in their raw form and tend to contain a lot of spatial structure. *Video*, an image sequence over time, has a lot of spatiotemporal structure. Modern machine learning techniques learn these spatial and spatiotemporal representations by being trained on vast quantities of data, which may themselves contain unwanted biases and unsuitable content. (The model for the specified problem is a fine-tuned version of the model pre-trained on the large-scale, more generic dataset. These large pre-trained models are referred to as *foundation models*.) Videos may also contain *audio* signals.

One of your colleagues at Unconditionally imagines that although less likely, the content of text messages and social media posts might predict a person's poverty level. This modality is *natural language* or *text*. Longer documents, including formal documents and publications, are a part of the same modality. The syntax, semantics, and pragmatics of human language is complicated. One way of dealing with text includes parsing the language and creating a syntax tree. Another way is representing text as sparse structured data by counting the existence of individual words, pairs of words, triplets of words, and so on in a document. These *bag-of-words* or *n-gram* representations are currently being superseded by a third way: sophisticated *large language models*, a type of foundation model, trained on vast corpora of documents. Just like in learning spatial representations of images, the learning of language models can be fraught with many different biases, especially when the norms of the language in the training corpus do not match the norms of the application. A language model trained on a humongous pile of newspaper articles from the United States will typically not be a good foundation for a representation for short, informal, code-mixed text messages in East Africa.[2]

Typically, structured modalities are their own representations for modeling and correspond to deliberative decision making by people, whereas semi-structured modalities require sophisticated transformations and correspond to instinctive perception by people. These days, the sophisticated transformations for semi-structured data tend to be learned using deep neural networks that are trained on unimaginably large datasets. This process is known as *representation learning*. Any biases present in the very large background datasets carry over to models fine-tuned on a problem-specific dataset because of the originally opaque and uncontrollable representation learning leading to the foundation model. As such, with semi-structured data, it is important that you not only evaluate the problem-specific dataset, but also the background dataset. With structured datasets, it is more critical that you analyze data preparation and feature engineering.[3]

4.2 Data Sources

Not only do the various kinds of data being considered by Unconditionally vary by modality, they also vary by how and where they come from, i.e., their *provenance*. As part of data understanding, your team

[2]Other strings with language-like characteristics such as DNA or amino acid sequences and software source code are currently being approached through techniques similar to natural language processing.

[3]There are new foundation models for structured modalities. Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. "Tabular Transformers for Modeling Multivariate Time Series." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Jun. 2021, pp. 3565–3569.

at Unconditionally must evaluate your *data sources* carefully to separate the wheat from the chaff: only including the good stuff. There are many different categories of data sources, which imply different considerations in assessing their quality.

4.2.1 Purposefully Collected Data

You may think that most data used in creating machine learning systems is expressly and carefully collected for the purpose of the problem, but you would be blissfully wrong. In fact, most data used in machine learning systems is repurposed. Some of the exceptions include data collected through surveys and censuses. These sources have the veneer of being well-designed and with minimal bias, but this might not always be the case. For example, if you rely on the recently completed national census in Kenya for either features or labels in Unconditionally's poverty prediction problem, you may suffer from non-responses and malicious data manipulation.

Another kind of purposefully collected data is generated from scientific experiments. Again, well-designed and well-conducted experiments should yield reliable data. However, there is a prevalent lack of trust in the scientific method due to practices such as misuse of data analysis to find patterns in data that can be selectively presented as statistically significant (p-hacking and the file drawer problem), lack of reproducibility, and outright fraud.

4.2.2 Administrative Data

Administrative data is the data collected by organizations about their routine operations for non-statistical reasons. Among Unconditionally's list of possible datasets, call detail records and mobile money transactions fit the bill. Data scientists and engineers frequently repurpose administrative data to train models because it is there and they can. Sometimes, it even makes sense to do so.

Since administrative data is a record of operations, which might have a direct bearing on an organization's bottom line or be subject to audit, it is usually quite correct. There are often difficulties in attempting to integrate different sources of administrative data within an organization due to their being siloed across teams. Such data can also contain traces of historical prejudices as well.

The most important thing for you to be aware of with administrative data is that it might not exactly match the predictive problem you are trying to solve. The machine learning problem specification may ask for a certain label, but the administrative data may contain columns that can only be proxies for that desired label. This mismatch can be devastating for certain individuals and groups, even if it is a decent proxy on average. For example, recall that in Chapter 2, we discussed how the number of doctor visits might not be a good proxy for how sick a patient is if there are external factors that prevent some groups from accessing health care. Also, the granularity of the records might be different than what is needed in the problem, e.g. individual phone numbers in call detail records instead of all activity by a household.

4.2.3 Social Data

Social data is data about people or created by people, and includes user-generated content, relationships between people, and traces of behavior.[4] Postings of text and images on social media platforms are a perfect example. Friendship networks and search histories are other examples. Similar to

[4]Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." In: *Frontiers in Big Data* 2.13 (Jul. 2019).

administrative data, these sources are not produced for the problem specification, but are repurposed for predictive or causal modeling. Many a time, just like administrative data, social data is only a proxy for what the problem specification requires and can be misleading or even outright wrong. The social media content of potential recipients of Unconditionally's cash transfer you analyze may be like this.

Since social data is created for purposes like communicating, seeking jobs, and maintaining friendships, the quality, accuracy, and reliability of this data source may be much less than administrative data. Text may include various slang, non-standard dialects, misspellings, and biases. Other modalities of social data are riddled with vagaries of their own. The information content of individual data points might not be very high. Also, there can be large amounts of sampling biases because not all populations participate in social platforms to the same extent. In particular, marginalized populations may be invisible in some types of social data.

4.2.4 Crowdsourcing

Supervised learning requires both features and labels. Unlabeled data is much easier to acquire than labeled data. *Crowdsourcing* is a way to fill the gap: crowd workers label the sentiment of sentences, determine whether a piece of text is hate speech, draw boxes around objects in images, and so on.[5] They evaluate explanations and the trustworthiness of machine learning systems. They help researchers better understand human behavior and human-computer interaction. Unconditionally contracted with crowd workers to label the type of roof of homes in satellite images.

In many crowdsourcing platforms, the workers are low-skill individuals whose incentive is monetary. They sometimes communicate with each other outside of the crowdsourcing platform and behave in ways that attempt to game the system to their benefit. The wages of crowd workers may be low, which raises ethical concerns. They may be unfamiliar with the task or the social context of the task, which may yield biases in labels. For example, crowd workers may not have the context to know what constitutes a household in rural East Africa and may thus introduce biases in roof labeling. (More details on this example later.) Gaming the system may also yield biases. Despite some platforms having quality control mechanisms, if you design the labeling task poorly, you will obtain poor quality data. In some cases, especially those involving applications with a positive social impact, the crowdworkers may have higher skill and be intrinsically motivated to do a conscientious job. Nevertheless, they may still be unfamiliar with the social context or have other biases.

4.2.5 Data Augmentation

Sometimes, especially in specialized problem domains, the amount of available data is not sufficient to learn high-performing models. *Data augmentation*—performing various transformations of the given dataset—may be used to increase data set size without actually collecting additional data. In image data, transformations for augmentation include rotations, flips, shifts, warps, additions of noise, and so on. In natural language data, transformations can include replacing words with synonyms. These sorts of heuristic transformations introduce some level of your subjectivity, which may yield certain biases.

---

[5]Jennifer Wortman Vaughan. "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research." In: *Journal of Machine Learning Research* 18.193 (May 2018).

Another way to perform data augmentation is through *generative machine learning*: using the given data to train a data generator that outputs many more samples to then be used for training a classifier. Ideally, these generated data points should be as diverse as the given dataset. However, a big problem known as *mode collapse*, which produces samples from only one part of the probability distribution of the given data, can yield severe biases in the resulting dataset.

4.2.6 Conclusion

Different data sources are useful in addressing various problem specifications, but all have biases of one kind or the other. Most data sources are repurposed. You must take care when selecting among data sources by paying attention to the more prevalent biases for any given data source. The next section describes biases from the perspective of their different kinds and where in the lifecycle they manifest.

4.3 Kinds of Biases

Your team is chugging along in the data understanding phase of the machine learning lifecycle. You know how different data modalities and data sources can go awry. These issues are your focus while appraising data for biases and lack of *validity* as it passes through various *spaces* in the machine learning lifecycle. A model of biases, validity, and spaces for you to keep in mind is given in Figure 4.3.

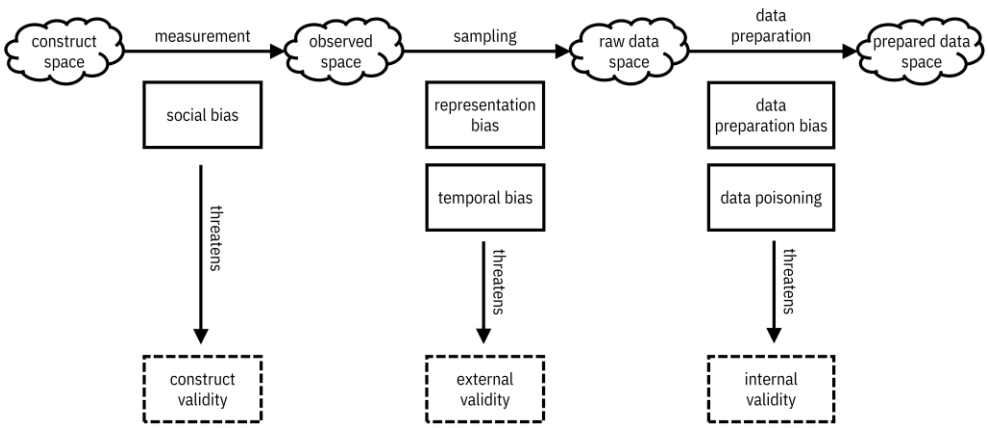

Figure 4.3. *A mental model of spaces, validities, and biases.* Accessible caption. A sequence of four spaces, each represented as a cloud. The construct space leads to the observed space via the measurement process. The observed space leads to the raw data space via the sampling process. The raw data space leads to the prepared data space via the data preparation process. The measurement process contains social bias, which threatens construct validity. The sampling process contains representation bias and temporal bias, which threatens external validity. The data preparation process contains data preparation bias and data poisoning, which threaten internal validity.

There are three main kinds of validity: (1) *construct validity*, (2) *external validity*, and (3) *internal validity*.[6] Construct validity is whether the data really measures what it ought to measure. External validity is whether analyzing data from a given population generalizes to other populations. Internal validity is whether there are any errors in the data processing.

The various kinds of validity are threatened by various kinds of bias. There are many categorizations of types of bias, but for simplicity, let's focus on just five.[7] *Social bias* threatens construct validity, *representation bias* and *temporal bias* threaten external validity, and *data preparation bias* and *data poisoning* threaten internal validity. These biases are detailed throughout this section.

It is useful to also imagine different spaces in which various abstract and concrete versions of the data exist: a *construct space*, an *observed space*, a *raw data space*, and a *prepared data space*. The construct space is an abstract, unobserved, theoretical space in which there are no biases. *Hakuna matata*, the East African problem-free philosophy, reigns in this ideal world. The construct space is operationalized to the observed space through the measurement of features and labels.[8] Data samples collected from a specific population in the observed space live in the raw data space. The raw data is processed to obtain the final prepared data to train and test machine learning models.

4.3.1 Social Bias

Whether it is experts whose decision making is being automated or it is crowd workers, people's judgement is involved in going from labels in the construct space to labels in the observed space. These human judgements are subject to human cognitive biases which can lead to implicit social biases (associating stereotypes towards categories of people without conscious awareness) that yield systematic disadvantages to unprivileged individuals and groups.[9] If decision makers are prejudiced, they may also exert explicit social bias. These biases are pernicious and reinforce deep-seated structural inequalities. Human cognitive biases in labeling can yield other sorts of systematic errors as well.

There can also be structural inequalities in features too. If an aptitude test asks questions that rely on specific cultural knowledge that not all test-takers have, then the feature will not, in fact, be a good representation of the test-taker's underlying aptitude. And most of the time, this tacit knowledge will favor privileged groups. Historical underinvestment and lack of opportunity among marginalized social groups also yield similar bias in features.

Your team found an interesting case of social bias when appraising your crowdsourced labels of roofs seen in satellite images in East African villages. The crowd workers had marked and labeled not only the roof of the main house of a household compound, but also separate structures of the same household such as a free-standing kitchen and free-standing sleeping quarters for young men. They had no idea that this is how households are laid out in this part of the world. The bias, if not caught, would have led to incorrect inferences of poverty.

---

[6]Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." In: *Frontiers in Big Data* 2.13 (Jul. 2019).

[7]Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle." In: *Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Oct. 2021, p. 17.

[8]Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Mar. 2021, pp. 375–385.

[9]Lav R. Varshney and Kush R. Varshney. "Decision Making with Quantized Priors Leads to Discrimination." In: *Proceedings of the IEEE* 105.2 (Feb. 2017), pp. 241–255.

4.3.2 Representation Bias

Once operating in the observation space of features and labels, the data engineers on your team must actually acquire sample data points. Ideally, this sampling should be done in such a way that the acquired data set is representative of the underlying population. Often however, there is *selection bias* such that the probability distribution in the observed space does not match the distribution of the data points. External validity may be affected. A specific example of selection bias is unprivileged groups being either underrepresented or overrepresented in the dataset, which leads to machine learning models either ignoring their special characteristics to satisfy an average performance metric or focusing too much on them leading to systematic disadvantage. Upon appraisal of one of Unconditionally's mobile phone datasets, the data engineers found that senior citizens were underrepresented because mobile phone ownership was lower in that subpopulation. They also found that the national census may have been undercounting people in some counties because the statistics authority had not provisioned enough census takers there.

Representation bias need not only be selection bias. Even if present, the characteristics of the features and labels that come from one subpopulation may be different than those from another. Representativeness is not only a question of the presence and absence of data points, but is a broader concept that includes, among others, systematic differences in data quality.

4.3.3 Temporal Bias

Temporal bias is another bias that happens when the observation space is sampled to collect the raw data. It also puts external validity at stake. Once a dataset has been collected, it can get out of sync with the distribution in the observation space if the observation space drifts and shifts over time. *Covariate shift* refers to the distribution of the features, *prior probability shift* refers to the distribution of the labels, and *concept drift* refers to the conditional distribution of the labels given the features. These drifts and shifts can be gradual or sudden. (Distribution shift is covered in greater detail in Chapter 9.) An example of covariate shift in Unconditionally's satellite image dataset is that some locations were observed in the rainy season and some were observed in the dry season.

4.3.4 Data Preparation Bias

The data preparation phase follows the data understanding phase in the machine learning lifecycle. Many biases can be introduced in data preparation that limit internal validity. For example, the data engineers on your team must do something to rows containing missing values. If they follow the common practice of dropping these rows and the missingness is correlated with a sensitive feature, like a debt feature being missing more often for certain religious groups, they have introduced a new bias. Other biases can enter in data preparation through other data cleaning, data enrichment, and data aggregation steps, as well as in data augmentation (see Section 4.2.5).

A sometimes overlooked bias is the use of proxies in the labels. For example, arrests are a problematic proxy for committing crimes. Innocent people are sometimes arrested and more arrests happen where there is more police presence (and police are deployed unevenly). Health care utilization is a problematic proxy for an individual's health status because groups utilize health care systems unevenly. Data preparation biases are often subtle and involve some choices made by the data engineer and data scientist, who are influenced by their own personal and social biases. You can help mitigate

some of these social biases by taking input from a diverse panel of informants in the data understanding phase of the lifecycle. (The role of a diverse team in data understanding is covered in greater depth in Chapter 16.)

4.3.5 Data Poisoning

Finally, a malicious actor can introduce unwanted biases into a dataset, unbeknown to you. This kind of adversarial attack is known as *data poisoning*. Data poisoning is accomplished through different means, including *data injection* and *data manipulation*. Data injection is adding additional data points with characteristics desired by the adversary. Data manipulation is altering the data points already in the dataset. (Data poisoning is covered in more detail in Chapter 11.)

Most of the biases introduced in this chapter can be implemented deliberately to reduce the performance of a machine learning system or otherwise degrade it. Security analysts commonly focus on attacks on accuracy, but other considerations like fairness can also be attacked. In addition to degrading performance, data poisoning can introduce a so-called backdoor for the adversary to exploit later. For example, someone trying to swindle Unconditionally might introduce satellite images of households next to rivers always labeled as severe poverty to trick your model into giving more cash transfers to riverside communities.

4.3.6 Conclusion

The different categories of bias neutralize different types of validity. Appraising data and preparing data are difficult tasks that must be done comprehensively without taking shortcuts. More diverse teams may be able to brainstorm more threats to validity than less diverse teams. Assessing data requires a careful consideration not only of the modality and source, but also of the measurement, sampling, and preparation. The mental model of biases provides you with a checklist to go through before using a dataset to train a machine learning model. Have you evaluated social biases? Is your dataset representative? Could there be any temporal dataset shifts over time? Have any data preparation steps accidently introduced any subtle biases? Has someone snuck in, accessed the data, and changed it for their malicious purpose?

What should you do if any bias is found? Some biases can be overcome by collecting better data or redoing preparation steps better. Some biases will slip through and contribute to epistemic uncertainty in the modeling phase of the machine learning lifecycle. Some of the biases that have slipped through can be mitigated in the modeling step explicitly through defense algorithms or implicitly by being robust to them. You'll learn how in Part 4 of the book.

4.4 Summary

- Data is the prerequisite for modeling in machine learning systems. It comes in many forms from various sources and can pick up many different biases along the way.
- It is critical to ascertain which biases are present in a dataset because they jeopardize the validity of the system solving the specified problem.
- Evaluating structured datasets involves evaluating the dataset itself, including a focus on data preparation. Evaluating semi-structured datasets that are represented by foundation models and

learned representations additionally involves evaluating large-scale background datasets.

- The source of most data for machine learning is repurposed data created and collected for other reasons. Evaluating the original reason for data creation provides insight into a dataset's bias.
- No matter how careful one is, there is no completely unbiased dataset. Nevertheless, the more effort put in to catching and fixing biases before modeling, the better.
- Trustworthy machine learning systems should be designed to mitigate biases that slip through the data understanding and data preparation phases of the lifecycle.