5

Privacy and Consent

A global virus pandemic is starting to abate, and different organizations are scrambling to put together 'back-to-work' plans to allow employees to return to their workplace after several months in lockdown at home. Toward this end, organizations are evaluating a (fictional) machine learning-based mobile app named TraceBridge. It supports the return to the office by collecting and modeling location traces, health-related measurements, other social data (e.g. internal social media and calendar invitations among employees), and administrative data (e.g. space planning information and org charts), to facilitate digital contact tracing: the process of figuring out disease-spreading interactions between an infected person and others. Is TraceBridge the right solution? Will organizations be able to re-open safely or will the employees be homebound for even more seemingly unending months?

The data that TraceBridge collects, even if free from many biases investigated in Chapter 5, is not free from concern. Does TraceBridge store the data from all employees in a centralized database? Who has access to the data? What would be revealed if there were a data breach? Have the employees been informed about possible uses of the data and agreed to them? Does the organization have permission to share their data with other organizations? Can employees opt out of the app or would that jeopardize their livelihood? Who gets to know that an employee has tested positive for the disease? Who gets to know their identity and their contacts?

The guidance to data scientists in Chapter 4 was to be wary of biases that creep into data and problem formulations because of the harms they can cause. In this chapter, the thing to be wary about is whether it is even right to use certain data for reasons of consent, power, and privacy.[1] Employers are now evaluating the app. However, when the problem owners, developers, and data scientists of TraceBridge were creating the app, they had to:

- weigh the need for consent, diffusion of power, and privacy,

[1] Eun Seo Jo and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, Jan. 2020, pp. 306–316.

- differentiate different kinds of anonymity for privacy, and
- question whether anonymity is the only way to achieve privacy.

Let’s critique their choices.

5.1 *Consent, Power, and Privacy*

The design of TraceBridge utilizes purposefully collected data, social data, and administrative data because using all of these data sources as features increases the performance of the underlying machine learning models. The app does not inform employees that it accesses internal social media, calendar invitations, and administrative data. In the app design, the employer’s organizational leadership has full control over the data.

As the designers, the TraceBridge team thought they were taking a simple and effective approach, but they did not understand that they were introducing problems of consent and power. The employer holds all the power in the deployment of the app because it can require the usage of the app as a condition of employment without any opportunity for the employee to give consent. Employees also have no opportunity to provide informed consent to the use of specific parts of their data. The employer holds the power to use the data not only for contact tracing of the viral infection, but also to track worker movements and interactions for other reasons like noting too many breaks and unwanted gatherings. Nothing prevents them from selling the data to other interested parties, leaving the employees powerless over their data. Overall, the design favors the powerful employer and fails the vulnerable employees.

Furthermore, the TraceBridge system design stores all personally-identifiable data it uses centrally without encryption or other safeguards for security, and makes it available without obfuscation to the organization’s management as the default behavior. When an infection is detected, an alert goes out to all people in the organization. Details of the identity of the infected person are transmitted to management and all inferred contacts.

The TraceBridge team may think they are providing a turnkey solution that does not overcomplicate things on the backend, but their design choices sacrifice *privacy*, the ability of individuals to withhold information about themselves. Privacy is considered an essential human right in many value systems and legal frameworks. The central repository of personally-identifiable information provides no protections to maintain anonymity in the employee’s data. The health status and movement of employees throughout the day is completely obvious by name. Furthermore, by revealing identifying information through alerts, there is no maintenance of anonymity. The TraceBridge team has been quite negligent of privacy considerations and any organization using the app will likely be on the wrong side of the law.

In a broad sense, data is a valuable commodity. It reveals a lot about human behavior at a gross level, but also about the behavior of individual people. Just like other natural resources, it can be extracted from the vulnerable without their consent and furthermore be exploited for their subjugation.[2] Some

[2] Shakir Mohamed, Marie-Therese Png, and William Isaac. “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence.” In: *Philosophy and Technology* 33 (Jul. 2020), pp. 659–684.

even argue that people should be compensated for their personal data because they are selling their privacy.[3] In short, *data is power*.

> "Data is the new oil."
>
> —Clive Humby, data science entrepreneur at dunnhumby

Data used in machine learning is often fraught with power and consent issues because it is often repurposed from other uses or is so-called *data exhaust*: byproducts from people's digital activities. For example, many large-scale image datasets used for training computer vision models are scraped from the internet without explicit consent from the people who posted the images.[4] Although there may be implicit consent through vehicles such as Creative Commons licenses, a lack of explicit consent can nevertheless be problematic. Sometimes copyright laws are violated in scraped and repurposed data.

Why does this happen? It is almost always due to system designers taking shortcuts to gather large datasets and show value quickly without giving thought to power and consent. And it is precisely the most powerful who tend to be least cognizant of issues of power. People from marginalized, minoritized, and otherwise less powerful backgrounds tend to have more knowledge of the perspectives of both the powerful and the powerless.[5] This concept, known as the *epistemic advantage* of people with lived experience of marginalization, is covered in greater detail in Chapter 16. Similarly, except in regulated application domains such as health care, privacy issues have usually been an afterthought due to convenience. Things have started to change due to comprehensive laws such as the General Data Protection Regulation enacted in the European Economic Area in 2018.

In summary, problem owners and data scientists should not have any calculus to weigh issues of power, consent and privacy against conveniences in data collection. For the fourth attribute of trust (aligned purpose), trustworthy machine learning systems require that data be used consensually, especially from those who could be subject to exploitation. No ifs, ands, or buts!

5.2 Achieving Privacy through Anonymization

After receiving unfavorable feedback from organizations that they risk breaking privacy laws, the TraceBridge development team is back to the drawing board. They must figure out what the heck privacy is all about, pick among competing frameworks, and then incorporate them into their system.

In preserving privacy, there are two main use cases: (1) data publishing and (2) data mining. *Privacy-preserving data publishing* is anonymizing data to fully disclose it without violating privacy. *Privacy-preserving data mining* is querying data while controlling the disclosure of information at the individual level. Privacy-preserving data publishing is also known as *non-interactive anonymization* and privacy-

[3] Nicholas Vincent, Yichun Li, Renee Zha, and Brent Hecht. "Mapping the Potential and Pitfalls of 'Data Dividends' as a Means of Sharing the Profits of Artificial Intelligence." arXiv:1912.00757, 2019.

[4] Abeba Birhane and Vinay Uday Prabhu. "Large Image Datasets: A Pyrrhic Win for Computer Vision?" In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. Jan. 2021, pp. 1536–1546.

[5] Miliann Kang, Donovan Lessard, and Laura Heston. *Introduction to Women, Gender, Sexuality Studies*. Amherst, Massachusetts, USA: University of Massachusetts Amherst Libraries, 2017.

preserving data mining is also known as *interactive anonymization*. TraceBridge may want to do either or both: publishing datasets for examination by organizational leaders or state regulators, and issuing contact tracing alerts without exposing individually-identifiable data. They have to go down both paths and learn about the appropriate technical approaches in each case: *syntactic anonymity* for data publishing and *differential privacy* for data mining.[6]

There are three main categories of variables when dealing with privacy: (1) identifiers, (2) quasi-identifiers, and (3) sensitive attributes. *Identifiers* directly reveal the identity of a person. Examples include the name of the person, national identification numbers such as the social security number, or employee serial numbers. Identifiers should be dropped from a dataset to achieve privacy, but such dropping is not the entire solution. In contrast, *quasi-identifiers* do not uniquely identify people on their own, but can reveal identity when linked together through a process known as *re-identification*. Examples are gender, birth date, postal code, and group membership. *Sensitive attributes* are features that people do not want revealed. Examples are health status, voting record, salary, and movement information. Briefly, syntactic anonymity works by modifying quasi-identifiers to reduce their information content, including suppressing them, generalizing them, and shuffling them. Differential privacy works by adding noise to sensitive attributes. A mental model for the two modes of privacy is given in Figure 5.1.

To make this mental model more concrete, let's see how it applies to an actual sample dataset of employees and their results on a diagnostic test for the virus (specifically the cycle threshold (CT) value of a polymerase chain reaction test), which we treat as sensitive. The original dataset, the transformed dataset after k-anonymity with $k = 3$, and the transformed dataset after differential privacy are shown in Table 5.1, Table 5.2, and Table 5.3 (details on k-anonymity and differential privacy are forthcoming).

Table 5.1. *A sample original dataset.*

Name	Department	CT Value
Joseph Cipolla	Trustworthy AI	12
Kweku Yefi	Neurosymbolic AI	20
Anjali Singh	AI Applications	35
Celia Sontag	Compute Acceleration	31
Phaedra Paragios	Software-Defined Architecture	19
Chunhua Chen	Thermal Packaging	27

Table 5.2. *The sample original dataset under k-anonymity with $k = 3$.*

Organization	CT Value
AI	12
AI	20
AI	35
Hybrid Cloud	31
Hybrid Cloud	19
Hybrid Cloud	27

[6] John S. Davis II and Osonde A. Osoba. "Privacy Preservation in the Age of Big Data: A Survey." *RAND Justice, Infrastructure, and Environment* Working Paper WR-1161, 2016.

Table 5.3. *The values returned for queries under differential privacy with Laplace noise added to the sensitive attribute in the sample original dataset.*

Name	Department	CT Value
Joseph Cipolla	Trustworthy AI	13.5
Kweku Yefi	Neurosymbolic AI	12.8
Anjali Singh	AI Applications	32.7
Celia Sontag	Compute Acceleration	35.9
Phaedra Paragios	Software-Defined Architecture	22.1
Chunhua Chen	Thermal Packaging	13.4

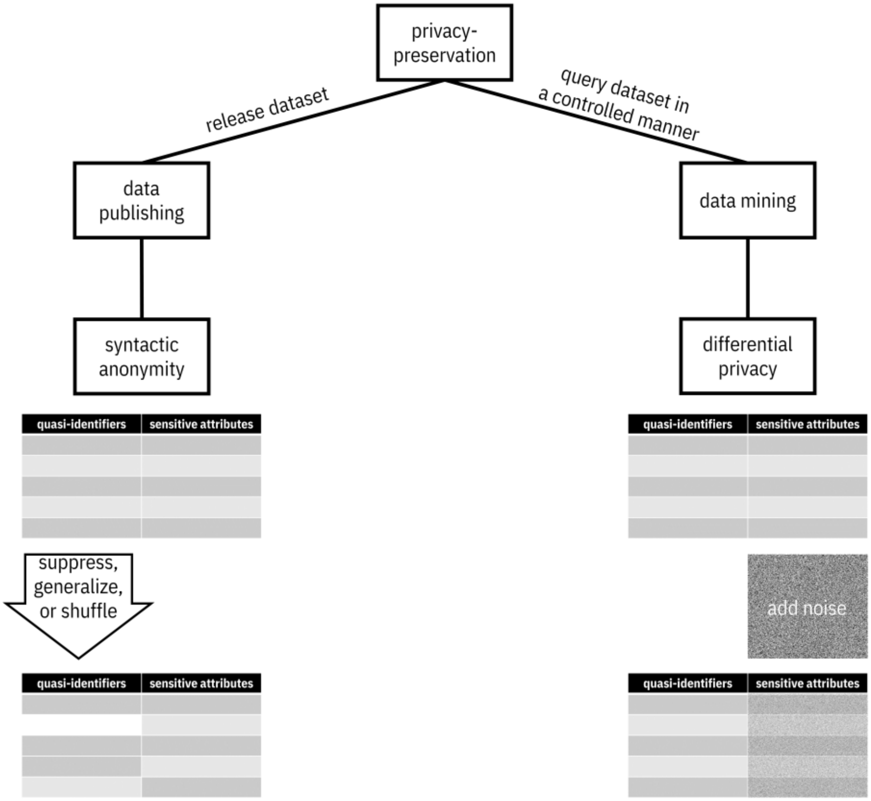

Figure 5.1. *A mental model of privacy-preservation broken down into two branches: data publishing with syntactic anonymity and data mining with differential privacy.* Accessible caption. A hierarchy diagram with privacy-preservation at its root. One child is data publishing, which is done when you release dataset. The only child of data publishing is syntactic anonymity. Syntactic anonymity is illustrated by a table with columns for quasi-identifiers and sensitive attributes. By suppressing, generalizing, or shuffling quasi-identifiers, some rows have been reordered and others have taken on a different value. The other child of privacy-preservation is data mining, which is done when you query dataset in a controlled manner. The only child of data mining is differential privacy. Differential privacy is also illustrated by a table with columns for quasi-identifiers and sensitive attributes. By adding noise to sensitive attributes, all the rows are noisy.

5.2.1 Data Publishing and Syntactic Anonymity

The simplest form of syntactic anonymity is *k-anonymity*.[7] By means of suppressing values of quasi-identifiers (replacing the value with a null value) or generalizing their values (for example replacing 5-digit zip codes with only their first three digits), the idea of k-anonymity is to create groups of records of cardinality at least k that have exactly the same modified quasi-identifier values. All the records within a group or cluster become equivalent and cannot be distinguished. Randomly shuffling identifiers within a quasi-identifier group achieves the same effect. If there are n data points in the original dataset, then there should be about n/k groups in the anonymized dataset, each of approximately the same cardinality.

Weaknesses of k-anonymity include susceptibility to the *homogeneity attack* and the *background knowledge attack*. The homogeneity attack takes advantage of many records within a k-member cluster having the same sensitive attributes, which means that even without precise re-identification, the sensitive information of individuals is still revealed. The background knowledge attack takes advantage of side information of subgroups having specific distributions of sensitive attributes to home in on likely sensitive attribute values of individuals. An extension of k-anonymity known as is *l-diversity* overcomes these vulnerabilities.[8] It further requires each k-member group to have at least l distinct values of sensitive attributes.

A further enhancement of k-anonymity and l-diversity is *t-closeness*.[9] Starting with the basic definition of k-anonymity, t-closeness further requires that the suitably-defined distance between the sensitive attribute probability distribution of each k-member group and the global sensitive attribute probability distribution of all records in the dataset is less than or equal to t. Simply put, all the groups should be similar in their distribution of sensitive attributes. Finding a t-closeness transformation of a given dataset is computationally difficult.

The re-identification risks of k-anonymity, l-diversity, and t-closeness have interpretations in terms of mutual information, which was introduced in Chapter 3. If X is the random variable for quasi-identifiers in the original dataset, $\tilde{X}$ is the random variable for quasi-identifiers in the anonymized dataset, and W is the random variable for sensitive attributes, then we have the following quantitative problem specifications:

- $I(X,\tilde{X}) \leq log\,\frac{n}{k}$ (k-anonymity),
- $I(W,\tilde{X}) \leq H(W) - log\,l$ (l-diversity), and
- $I(W,\tilde{X}) \leq t$ (t-closeness).[10]

Through k-anonymity, the reidentification risk is reduced down from that of the full dataset to the number of clusters. With l-diversity or t-closeness added on top of k-anonymity, the predictability of the

[7] Latanya Sweeney. “k-Anonymity: A Model for Protecting Privacy.” In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (Oct. 2002), pp. 557–570.

[8] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. “l-Diversity: Privacy Beyond k-Anonymity.” In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (Mar. 2007), p. 3.

[9] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.” In: *Proceedings of the IEEE International Conference on Data Engineering*. Istanbul, Turkey, Apr. 2007, pp. 106–115.

[10] Michele Bezzi. “An Information Theoretic Approach for Privacy Metrics.” In: *Transactions on Data Privacy* 3.3 (Dec. 2010), pp. 199–215.

sensitive attributes from the anonymized quasi-identifiers is constrained. These relationships are valuable ways of reasoning about what information is and is not revealed due to anonymization. By expressing them in the common statistical language of information theory, they can be examined and studied alongside other problem specifications and requirements of trustworthiness in broader contexts.

5.2.2 Data Mining and Differential Privacy

The other branch of anonymization is differential privacy and applies to use cases involving querying a dataset, not simply releasing it. The setup is that an organization has a dataset and knows exactly which query it will be dealing with. Some example queries are to return the count of a certain value in the dataset or the average value of a column in the dataset. A query could even be as complicated as returning a machine learning classifier trained on the dataset. Some queries are easier and some queries are harder to anonymize. In the differential privacy setup, the organization has to maintain control over the dataset and all queries to it. This is in contrast to syntactic anonymity where once the dataset has been anonymized, it is footloose and free. The basic method of differential privacy is adding noise to sensitive attributes.

Getting down into a little more detail, let's say that TraceBridge has a dataset W_1 of all employees positive for the viral disease. Another employee is detected to be positive and is added to the dataset giving us a new dataset W_2 that only differs from W_1 by the addition of one row. Let's say that the query function $y(W)$ is the number of employees who have cancer, which is important for better understanding the relationship between cancer and the viral disease.[11] Cancer diagnosis status is considered sensitive. Instead of actually returning $y(W)$, a differentially-private system gives back a noisy version of $y(W)$ by adding a random value to it. The value it returns is $\tilde{Y}(W) = y(W) + \text{noise}$. $\tilde{Y}$ is a random function which we can think of as a random variable that takes sample value $\tilde{y}$. The goal of differential privacy is expressed by the following bound involving the probabilities of queries from the original and new datasets:

$$P\big(\tilde{Y}(W_1) = \tilde{y}\big) \leq e^{\epsilon} P\big(\tilde{Y}(W_2) = \tilde{y}\big), \text{for all } \tilde{y}.$$

Equation 5.1

The ϵ is a tiny positive parameter saying how much privacy we want. The value of e^{ϵ} becomes closer and closer to one as ϵ gets closer and closer to zero. When ϵ is zero, the two probabilities are required to be equal and thus the two datasets have to be indistinguishable, which is exactly the sense of anonymity that differential-privacy is aiming for.[12] You can't tell the difference in the query result when you add the new person in, so you can't figure out their sensitive attribute any more than what you could have figured out in general from the dataset.

[11] https://rebootrx.org/covid-cancer

[12] We should really write $P\big(\tilde{Y}(W_1) \in S\big) \leq e^{\epsilon} P\big(\tilde{Y}(W_2) \in S\big)$ for some interval or other set S because if $\tilde{Y}$ is a continuous random variable, then its probability of taking any specific value is always zero. It only has a specific probability when defined over a set.

The main trick in differential privacy is solving for the kind of noise and its strength to add to $y(W)$. For lots of query functions, the best kind of noise comes from the Laplace distribution.[13] As stated earlier, some queries are easier than others. This easiness is quantified using *global sensitivity*, which measures how much a single row of a dataset impacts the query value. Queries with smaller global sensitivity need lower strength noise to achieve ϵ-differential privacy.

Just like with syntactic privacy, it can be easier to think about differential privacy alongside other problem specifications in trustworthy machine learning like accuracy, fairness, robustness, and explainability when expressed in terms of information theory rather than the more specialized terminology used in defining it earlier. To do so, we also need to say that the dataset W is a random variable, so the probabilities that we want to be close to each other are the noised query results conditioned on the dataset realizations $P(\tilde{Y} \mid W = w_1)$ and $P(\tilde{Y} \mid W = w_2)$. Then we can pose our objective of differential privacy as wanting the mutual information between the dataset and noisy query result $I(W, \tilde{Y})$ to be minimized. With some more specifics added to minimizing the mutual information, we can get back a relationship in terms of the ϵ of ϵ-differential privacy.[14] The idea of examining the mutual information between the dataset and the query is as follows. Since mutual information measures the reduction in uncertainty about the dataset by the knowledge of the query, having zero (or small) mutual information indicates that we don't learn anything about the dataset's composition from the query result, which is exactly the idea of differential privacy.

Differential privacy is intended to prevent attributing the change of the query's value to any one person's row of data. Nevertheless, one criticism of differential privacy as imagined in its information theory presentation is that it can allow the value of sensitive attributes to be inferred if there are correlations or associations among the rows. Stricter information-theoretic conceptions of differential privacy have been developed that require the rows of the dataset to be independent.

5.2.3 Conclusion

TraceBridge has a few different use cases as part of their app and contact tracing system. One is publishing sensitive data for the leaders of the organization, external regulators, or auditors to look at. A second is to interactively query statistics from the sensitive data without revealing things about individuals. Each use has different appropriate approaches: syntactic anonymity for the first and differential privacy for the second, along with different requirements on the system design and the infrastructure required. Existing legal protections in various jurisdictions and application domains are mostly for the first use case (data publishing), but the regulations themselves are usually unclear on their precise notion of privacy. TraceBridge may have to go with both approaches to privacy in developing a trusted system.

We've reached the end of this section and haven't talked about the tradeoff of privacy with utility. All measures and approaches of providing privacy should be evaluated in conjunction with how the data is going to be used. It is a balance. The tradeoff parameters are there for a reason. The usefulness of a dataset after k-anonymization is usually pretty good for a decent-sized k, but might not be so great after

[13] The pdf of the Laplace distribution is $p_X(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$, where μ is the mean and b is a scale parameter such that the variance is $2b^2$.

[14] Darakshan J. Mir. "Information-Theoretic Foundations of Differential Privacy." In: *Foundations and Practice of Security*. Montreal, Canada, Oct. 2012, pp. 374–381.

achieving t-closeness for a decent t. Similarly, a criticism of differential privacy for typical queries is that the usefulness of the query results is not that great for a small ϵ (adding a large magnitude of noise). However, there are no blanket statements to be made: these intuitions have to be appraised for specific scenarios and datasets, and privacy parameters have to be chosen carefully without taking shortcuts and incorporating input from multiple stakeholders.

5.3 Other Ways of Achieving Privacy

The two technical approaches that yield anonymized data (syntactic anonymity for data publishing and differential privacy for data mining) are not the only ways for TraceBridge to achieve privacy. Here's a really easy way for them to achieve privacy: lock up the data and throw away the key. If they don't publish the data and provide no ability to query it, they have perfect privacy. But they don't have any utility from the data either. So what else can they do?

One answer is to set up *institutional controls* and procedures so that only qualified individuals have access to data, and only for specific approved uses. Cleared data scientists may only be allowed to access the data on closed computing systems with strong enforcement to prevent data breaches. Keeping data in a decentralized manner rather than all in one centralized place can also help prevent breaches.

A second answer is to bring the algorithm to the data rather than the other way around. Working through a decentralized system where different pieces of data are kept in different places, *secure multi-party computation* allows a value to be computed using data from different sources without revealing the inputs sent by each data source to other data sources.

A third answer is encryption. TraceBridge can use *fully homomorphic encryption* to compute things on encrypted data and get the answer they would have gotten if the data hadn't been encrypted. This approach can be a computational beast, but is getting more and more computationally tractable every day.

With all three of these approaches: institutional controls, secure multi-party computation, and fully homomorphic encryption, the question of *what* is computed remains open. People and organizations can be using these techniques and still be outputting some value or summary statistic that discloses sensitive individual information. It may thus make sense to combine these methods with, for example, differential privacy.

5.4 Summary

- Data is a valuable resource that comes from people. The use of this data should be consensually obtained. If there is no consent, do not proceed.
- It is easy for data scientists to set up machine learning systems that exploit and subjugate vulnerable individuals and groups. Do not do it. Instead, be careful, thoughtful, and take input from powerless groups.
- By consenting to the use of their data, people give up their privacy. Various methods can be used to preserve their privacy.
- Syntactic anonymization methods group together individuals with similar quasi-identifiers and then obfuscate those quasi-identifiers. These methods are useful when publishing individual-

level data.

- Differential privacy methods add noise to queries about sensitive attributes when users can only interact with the data through known and fixed queries. These methods are useful when statistically analyzing the data or computing models from the data.
- Securing access to data provides an alternative to data anonymization.