Chapter 1

Overview and Fundamental Techniques

By Ferdinando Fioretto, Pascal Van Hentenryck and Juba Ziani

DOI: 10.1561/9781638284772.ch1

Published in *Differential Privacy in Artificial Intelligence: From Theory to Practice* by Ferdinando Fioretto and Pascal Van Hentenryck (eds.). 2025. ISBN 978-1-63828-476-5. E-ISBN 978-1-63828-477-2.

Suggested citation: Ferdinando Fioretto *et al.*. 2025. "Overview and Fundamental Techniques" in *Differential Privacy in Artificial Intelligence: From Theory to Practice*. Edited by Ferdinando Fioretto and Pascal Van Hentenryck. pp. 2–41. Now Publishers. DOI: 10.1561/9781638284772.ch1.

now
the essence of knowledge

1.1 Introduction

Data is continuously harvested from nearly every facet of our lives by corporations, service providers, and public institutions. Whether through smartphones, social media interactions, internet use, healthcare visits, or financial transactions, information is continuously gathered, shaping the foundation of the modern economy. Private companies leverage this vast pool of data to evaluate loan candidates, optimize transportation networks, improve supply chains, personalize services, and predict market demands, all to enhance decision making. Similarly, public policies and government initiatives rely heavily on this data, guiding resource distribution, monitoring public health crises, and driving urban development and sustainability efforts.

However, these datasets also contain a large array of sensitive information, including health, financial, or location data. Major privacy violations and breaches are commonplace, and can have severe negative impacts, not only on consumers and online users, but also on entire organizations and governments. For instance, the 2017 Equifax data breach [Wik24a] exposed the personal information of 147 million individuals, including social security numbers, birth dates, and addresses, leaving millions vulnerable to identity theft, fraud, and long-term financial harm. Similarly, the 2016 Facebook-Cambridge Analytica scandal [Wik24b], in which

the personal data of up to 87 million Facebook users was harvested without consent for political advertising purposes, raised concerns about its possible influence on the outcome of the 2016 presidential election.

Privacy concerns have become central in today's society, driving significant changes in government policy. Various regulatory frameworks have been established, with the United States and Europe leading efforts toward stronger privacy practices. In Europe, the General Data Protection Regulation (GDPR) sets strict standards for data management, focusing on consent and data minimization [PE16], while in the U.S., regulations like Title 13 [Uni98] govern the handling of census data and laws like the Health Insurance Portability and Accountability Act (HIPAA) and the California Consumer Privacy Act (CCPA) offer protections for health and consumer data [Cen96; Leg18]. This movement was further emphasized in October 2023 when the Biden administration issued an Executive Order on AI, ensuring the enforcement of consumer protection laws and introducing safeguards against privacy violations in AI systems. Government actions, such as the release of the AI Bill of Rights Blueprint [Hou23] in the US, underscore the increasing focus on privacy in both policy and technology.

Public policy has also devoted extensive research into technical solutions for privacy. Over the past three decades, this research has explored a wide range of privacy definitions and techniques, but one has emerged as a pivotal framework: *Differential Privacy* (DP) [Dwo+06]. DP has gained widespread recognition and adoption, not only by leading technology companies like Apple, Meta, Google, and LinkedIn, but also by the U.S. government, most notably in its landmark 2020 Census data release.

Differential Privacy is now widely regarded as the gold standard for privacy protection in statistical analyses and dataset releases. Its strength lies in providing a *formal* and *mathematical* definition of privacy, offering *precise* and *provable* guarantees. This is in stark contrast to historically ad-hoc and loosely defined privacy methods, which have repeatedly failed under attacks aimed at reconstructing part of the original dataset or identifying individuals in said datasets. As privacy challenges evolve, so too does Differential Privacy, expanding across diverse fields to meet new demands. This book aims at providing a comprehensive introduction to DP, particularly within the novel challenges brought by AI applications. It explores its foundational theories, applications in machine learning, and practical implementations, equipping readers with the knowledge to leverage this critical technology effectively.

Overview of the Chapter

This chapter is structured to provide an introduction to Differential Privacy. It begins by illustrating various attempts to protect data privacy, emphasizing where

and why they failed, and providing the key desiderata of a robust privacy definition (Section 1.2). It then defines the key actors, tasks, and scopes that make up the domain of privacy-preserving data analysis (Section 1.3). Following that, Section 1.4, formalizes the definition of DP and its inherent properties, including composition, post-processing immunity, and group privacy. The chapter also reviews the basic techniques and mechanisms commonly used to implement Differential Privacy in Sections 1.4 to 1.6. Finally, Section 1.8 concludes with an overview of Differential Privacy applications and some future directions in this field.

1.2 A Historical Perspective on Privacy

This section begins by posing a fundamental question: *What are the key desiderata and properties that a robust privacy definition must guarantee?* To address this, it examines historical failures of previous and current privacy definitions, highlighting the necessity for well-defined and formal guarantees. This section first outlines the main properties satisfied by Differential Privacy—these properties will be formally detailed later in this chapter. It then delves into specific examples of major privacy breaches over the past 30 years, identifying for each how adherence to certain privacy desiderata could have prevented the failure.

A central argument of this book is the importance of *well-defined and formal privacy guarantees*. A major weakness in many privacy techniques arises when the protections themselves are poorly specified, particularly when they fail to clearly define the classes of attacks they are designed to resist. Over the past three decades, numerous privacy attacks have exploited such ambiguities, often by applying privacy notions beyond their intended use cases. To address these challenges, this chapter focuses on four main *desiderata* that a strong privacy definition should satisfy:

1. **Desiderata 1: Compositionality.** A good privacy definition should ensure that its protections gracefully degrade when applied multiple times, whether across several datasets or through repeated private data analyses. In a data-driven world, where datasets are frequently analyzed multiple times and may contain overlapping information about individuals, composition is crucial. Without it, repeated analyses can cumulatively erode privacy safeguards and ultimately compromise individual privacy.
2. **Desiderata 2: Post-processing immunity.** Once data has been privatized using a privacy-preserving mechanism, any further data analyses should not degrade its privacy guarantees, provided that the original, non-privatized data remains inaccessible. This property assures that subsequent steps or transformations applied to the privatized output cannot compromise privacy.

Post-processing immunity offers a strong guarantee that allows data analysts to abstract away potential attack models, effectively providing *future-proof protection* against privacy violations.

3. **Desiderata 3: Group privacy.** Group privacy aims at controlling how privacy guarantees degrade when considering groups of individuals rather than single individuals. It ensures that a privacy mechanism does not arbitrarily fail to protect privacy beyond the individual level when data from multiple users is combined. While it is inevitable that privacy guarantees weaken as group sizes increase, since more information is encoded about them, the degradation should be controlled and quantifiable.
4. **Desiderata 4: Quantifiable privacy-accuracy trade-offs.** There is no free lunch in privacy: releasing accurate information about a group of people must necessarily and statistically encode some information about individuals. As privacy protection increases the accuracy of insights derived from the data may decrease. A good privacy definition should provide quantifiable trade-offs, allowing data analysts, decision-makers, and model builders to measure how much accuracy is sacrificed for a given level of privacy. This enables them to balance privacy and utility according to specific needs.

The following sections provide historical examples illustrating why privacy is complex, where traditional methods have failed, and how the above desiderata are essential for guaranteeing robust privacy.

1.2.1 Data Anonymization

A standard technique for privacy protection in various domains is *anonymization*. It involves the removal or masking of any identifying details to prevent the recovery of personal identities. Anonymization has been employed in areas such as the release of medical datasets under the Health Insurance Portability and Accountability Act (HIPAA) standards. In the mid-1990s, the Massachusetts Group Insurance Commission (GIC), a government agency responsible for purchasing health insurance for state employees, sought to promote medical research by releasing anonymized health data. The GIC approach involved removing what they considered "explicit" identifiers such as names, addresses, and social security numbers, while retaining hundreds of other attributes deemed non-identifiable. Supported by then-Governor William Weld, this initiative aimed at balancing data utility with privacy protection. However, in 1997, Dr. Latanya Sweeney, then a graduate student at MIT, set out to challenge the effectiveness of this anonymization. Using publicly available information, she re-identified Governor Weld's medical records within the dataset and sent them to his office, starkly demonstrating the vulnerability of supposedly anonymized data.

How was Dr. Sweeney able to uncover Governor Weld's personal medical information from the GIC's released data? One might assume that such an attack required sophisticated techniques and significant resources. In reality, her de-anonymization attack cost only $20 and limited time. The GIC's dataset included three crucial attributes for each individual: sex, zip code, and date of birth. Dr. Sweeney purchased voter registration records from Cambridge, Massachusetts, which contained names, addresses, zip codes, and dates of birth. By cross-referencing these two datasets, she found that only six people in Cambridge shared Governor Weld's birth date. Of those, only three were male, and just one resided in his zip code—uniquely identifying his medical records. This type of attack, known as a *linkage attack*, re-identifies individuals by linking anonymized data with external public records. In a subsequent report [Swe00], Dr. Sweeney demonstrated that her attack extended far beyond a single high-profile individual. She found that "87% of the population in the United States had reported characteristics that likely made them unique based only on 5-digit ZIP, gender, date of birth." Even at broader geographic levels, significant portions of the population could be uniquely identified with minimal information. "About half of the U.S. population are likely to be uniquely identified by only place, gender, date of birth, where place indicates the city, town, or municipality in which the individual resides."

Why Did Anonymization Fail?

Anonymization failed because it lacked formal privacy guarantees. Dr. Sweeney's attack was remarkably simple, yet unanticipated due to the absence of a precise attack model. The lack of *post-processing immunity* meant that, once the anonymized data was released, combining it with other publicly available datasets could reveal more information than intended. If the privacy mechanism had been robust to post-processing, additional analyses or data combinations would not have compromised individual privacy beyond what was already publicly accessible. This example motivates the need for formal privacy definitions that account for all potential avenues of data exploitation.

1.2.2 K-Anonymity

At this point, one might argue that the previous example does not represent a fundamental failure of anonymization as a privacy technique, but rather a misapplication in that specific instance. Is it possible to thwart de-anonymization attacks by simply withholding more attributes? For instance, would not releasing someone's zip code, date of birth, or gender resolve the issue? However, a significant challenge emerges in determining which combinations of publicly available attributes could uniquely

identify an individual. As the number of features in a dataset grows, it becomes practically impossible for modern computing to predict and guard against all potential attack vectors. Moreover, sensitive attributes often correlate statistically with non-sensitive ones, rendering anonymization susceptible to statistical attacks that can probabilistically reconstruct sensitive information through these correlations—for a particularly sensitive example involving genomic data, refer to [Hom+08].

Despite decades of deployment, anonymization has consistently failed to provide robust privacy protection. Other high-profile failures include the AOL search data release [BZH06], the Netflix Prize dataset [NS06], and studies demonstrating that individuals can be uniquely identified using just a few mobile phone location points [DHVB13]. So, what is the next step? Can the concept of anonymization be refined to address its shortcomings? A promising strategy might be to release only partial information about each attribute. For example, in demographic or medical analyses, knowing that an individual falls within a certain age *range*, such as "between 18 and 35," might suffice. By revealing less precise information, can re-identification attacks be made more difficult?

In 1998, Prof. Pierangela Samarati and Dr. Latanya Sweeney (the same Dr.Sweeney who highlighted the failures of basic anonymization) introduced a generalization called *k-anonymity* [SS98; Swe02]. A dataset satisfies k-anonymity if, for every record, there are at least $k - 1$ other records with identical values in a set of quasi-identifiers—attributes that could potentially be linked to external data to re-identify individuals. In this framework, it should be impossible to distinguish between any of the k individuals sharing the same quasi-identifiers. In particular, the larger the value of k, the stronger the privacy guarantee. Consider, for example, the dataset in Table 1.1 (left), containing information about state employees. One approach is to release this dataset by replacing sensitive names with random identifiers (see Table 1.1 (middle)). This technique provides only

Table 1.1. Three levels of anonymization on a demographic dataset. **Left:** original dataset. **Center:** masking the sensitive names for 1-anonymity. **Right:** generalizing Zip codes and age attributes for 4-anonymity.

Original Data

Name	Zip Code	Age
Rick	19456	67
Nathan	30309	33
Yani	19445	64
Xiao	30457	35
Luciana	19456	67
Anastasia	30271	38
Marcia	19456	31
Yuki	19456	62

Anonymized Data

ID	Zip Code	Age
1	19456	67
2	30309	33
3	19445	64
4	30457	35
5	19456	67
6	30271	38
7	19456	31
8	19456	62

4-anonymized Data

ID	Zip Code	Age
1	19***	60-70
2	30***	30-40
3	19***	60-70
4	30***	30-40
5	19***	60-70
6	30***	30-40
7	30***	30-40
8	19***	60-70

1-anonymity, which is essentially standard anonymization. However, each individual still has a unique combination of zip code and age, making them vulnerable to singling-out attacks [Swe00; Swe02]. In contrast, the table on the right demonstrates 4-anonymity: entries #1, 3, 5, and 8 are indistinguishable from each other, as are entries #2, 4, 6, and 7. Individuals are grouped into clusters of four, where each group shares the same generalized (zip code, age) attributes, significantly enhancing privacy.

Remark 1.1 (Privacy vs. Utility)**.** *In k-anonymization, increasing the value of k enhances* privacy *by making it more difficult to distinguish between individuals, as they are grouped into larger clusters with identical quasi-identifiers. However, this comes at the expense of* utility*. As k increases, the information becomes less precise, reducing the dataset's usefulness for analysis. For instance, in the* 4*-anonymized version of our dataset, the details about individuals' zip codes and ages are less specific compared to the* 1*-anonymized version. This trade-off between privacy and utility is a central theme in privacy research and will be addressed in the context of Differential Privacy in subsequent chapters.*

Where Does *k*-anonymization Fail? Reason #1: Lack of Group Privacy

At first glance, *k*-anonymity appears to address the shortcomings of basic anonymization by preventing the singling out of any specific individual within a dataset. In fact, for years, it was considered the state-of-the-art solution for preventing re-identification attacks. However, *k*-anonymity suffers from a significant limitation concerning the leakage of sensitive information, even when individuals are not directly identified. The core issue is not merely the potential to link a data subject to a specific record. Instead, the real problem lies in the exposure of sensitive information associated with individuals without explicit re-identification, as highlighted in [Des17]. In essence, one does not need to pinpoint a specific person to infer personal, sensitive details about them. Consider the previous example, but now suppose the dataset includes a sensitive attribute, such as credit scores (see Table 1.2, Left). Even without directly identifying anyone, an adversary could learn sensitive information about individuals based on the available data. Using the same linkage approach that Dr. Sweeney employed in her de-anonymization of the GIC medical records, one could cross-reference publicly available data to deduce that entries #2, 4, 6, and 7 correspond to Nathan, Xiao, Anastasia, and Marcia. Although it is impossible to match each person to their exact record, one can still infer that all four individuals have a credit score in the "Fair" category. This represents a significant privacy breach, as sensitive information is disclosed without explicit identification. Here, the property of *group privacy* is violated—the privacy guarantee collapses when aggregating data from as few as four individuals—leading to both group-level and *individual-level* harms.

Table 1.2. Two k-anonymized datasets augmented with credit score information. **Left:** State Employee Dataset. **Right:** The Dataset of Company Z.

ID	Zip Code	Age	Credit Score
1	19***	60-70	797
2	30***	30-40	650
3	19***	60-70	755
4	30***	30-40	590
5	19***	60-70	767
6	30***	30-40	597
7	30***	30-40	613
8	19***	60-70	775

ID	Zip Code	Age	Credit Score
A	30***	30-40	815
B	30***	30-40	613
C	30***	30-40	376
D	30***	30-40	727

Where Does k-anonymization Fail? Reason #2: Lack of Composition

A more subtle issue with k-anonymity arises from the concept of *composition*, described earlier in this chapter. Unfortunately, k-anonymity lacks fundamental composition guarantees and fails when multiple datasets are released. In fact, even releasing just two k-anonymized datasets can be sufficient to break its privacy protections in the worst-case scenario. To illustrate this, imagine a situation where it is known that Marcia is a state employee included in a dataset that has been 4-anonymized. Additionally, suppose that Marcia is a client of Company Z , which aims at helping individuals improve their credit scores. Company Z sells a separate 4-anonymized dataset about its customers (see Table 1.2, Right). Knowing that Marcia is present in both datasets, an adversary can cross-reference the state records with Company Z's records to find a *unique* match: the only individual appearing in both datasets is someone in the 30-40 age range, residing in zip code 30***, and having a credit score of 613. This individual must be Marcia, thereby uniquely identifying her. This scenario demonstrates a failure of *composition*: the privacy guarantees of k-anonymity break down when datasets are combined. While this example is simplified for clarity, extensive practical evidence has shown that the issues with k-anonymity are real and pervasive [NS08]. These limitations underscore the need for more robust privacy definitions that can withstand linkage attacks, data aggregation, and the release of multiple datasets.

1.2.3 Any Perfectly Accurate and Deterministic Privacy Notion Must Fail

Various strategies have been proposed to address the shortcomings k-anonymity without significantly reducing the utility of data for demographic and population-level analyses. One such method is *data swapping*, which involves exchanging parts of dataset entries among individuals to ensure that no single row corresponds

directly to one person, while still preserving overall demographic counts like the "number of people in dataset X that have property Y." This technique was employed in the release of U.S. Census data products prior 2020. Another approach is *data minimization*, which focuses on collecting as little data as necessary and discarding it after it has served its purpose. Despite these efforts, the challenge of ensuring that privacy guarantees degrade gracefully and predictably under repeated queries remains unresolved. To address this, it is important to highlight a *fundamental* property that must be satisfied by any robust privacy definitions, helping us narrow down the search for effective solutions. Specifically, the claim is that *no perfectly accurate and deterministic privacy technique can satisfy our requirements*, and that *randomization is essential for privacy*.

This crucial point can be illustrated with a simple example where the lack of randomness leads to a failure in *composition*. Imagine a hypothetical company named Gluble, which has 25 employees. Gluble publicly announces that the average salary of its employees is \$500,000, perhaps to attract top talent with its competitive compensation. After hiring a 26th employee named Rick, the company updates its public average salary to \$505,000. From these two pieces of information, one can deduce Rick's salary. Using basic arithmetic, Rick's salary x is obtained by solving $\frac{(x+25\times 500{,}000)}{26} = 505,000$, i.e., $x = \$630,000$. This amount is significantly higher than his colleagues' salaries. This scenario shows a *failure of composition*: while each individual data release seems innocuous, combining them allows an adversary to infer sensitive information about an individual. Even with access to just two queries, a differential attack reconstructed private data. Although this example is simplified, Dinur and Nissim [DN03] have shown that such differential attacks can be executed in far more complex settings, even when the query language is restricted. Importantly, the attack used no information about how the data was privatized. This vulnerability arises because the average salary at Gluble was released deterministically and exactly. What would happen if noise was added to Gluble's salary reports? Suppose that the average salary before hiring Rick was reported as approximately \$500,000, and after hiring, it was approximately \$505,000. It is no longer clear question whether the change is due to Rick's salary or simply a result of the added randomness. After all, the introduction of noise creates uncertainty, preventing exact inference of individual salaries. This concept of adding randomness to data releases is a cornerstone of Differential Privacy.

Observe that providing Differential Privacy is more complex than "just" adding noise. At a high level, the more noise is added, the better our privacy guarantees are going to be; however, adding too much noise is undesirable, as it destroys the utility of privately-released datasets and statistics. Therefore, noise must be carefully calibrated to balance privacy protection with data utility, enabling us to provide formal and provable privacy guarantees alongside precise *privacy-utility trade-offs*.

In fact, three years before Differential Privacy was formally introduced by Dwork et al. [DMNS06], Dinur and Nissim [DN03] laid the groundwork for understanding these trade-offs when incorporating privacy noise. Readers can consult their work, as well as subsequent studies [CNSU20b; CNSU20a], for a deeper exploration of the challenges in calibrating noise to protect against reconstruction attacks. The following sections delve deeper into Differential Privacy, what it protects against, its formal definition, guarantees, and the basic mechanisms to achieve it, providing a comprehensive understanding of the crucial role played by randomization in safeguarding privacy.

1.2.4 A Side Note: Other Types of Privacy Breaches

The discussion above highlights the importance of our privacy desiderata and illustrates how previous techniques that failed to meet these criteria have led to significant privacy failures. So far, the presentation relied on simple examples involving variants of anonymization techniques and the challenges associated with privatizing and releasing datasets. However, with the advent of increasingly complex models and large-scale machine learning applications, privacy failures have begun to emerge in more intricate and subtle ways—even when privatized datasets are never directly released. In particular, recent research has demonstrated that privacy can be compromised not only through released statistics but also via the models themselves. A notable example is *Federated Learning (FL)* [LSTS20], discussed in Chapter 8. The goal of FL frameworks is to protect privacy through decentralization: each user retains their data on their local device, performs computations locally, and only transmits aggregated updates (such as gradient information) to a central server. The intent is that no central entity ever accesses individual user data, thereby preserving privacy. Yet, recent work has shown that this is insufficient: the gradient updates themselves often encode sufficient information to be able to guess the original user data with high accuracy [ZLH19]. This is the topic of Chapter 8.

This issue is not confined to the training of machine learning models. Even after a model is trained and the original data is ostensibly deleted, the released models can still encapsulate information about the training data. This can lead to privacy breaches where models inadvertently memorize and reproduce parts of their training datasets, as discussed in Chapter 5. For instance, large language models trained on extensive text corpora have been found to occasionally output verbatim snippets from the training data when prompted in specific ways [Car+21]. A real-world example involves a South Korean AI company Scatter Lab [Dob21]. Scatter Lab used text and messaging data from users on South Korea's biggest test messaging company, KakaoTalk, to train a chatbot service. Despite efforts to remove

personally identifiable information, the chatbot reproduced memorized conversations from the training data when users interacted with it, inadvertently disclosing private and sensitive information about KakaoTalk users. These examples illustrate that privacy breaches can occur even without direct access to the underlying datasets. Thus there is a need for privacy-preserving techniques that extend beyond data anonymization and address the inherent risks in modern machine learning practices. Differential Privacy offers a framework to mitigate these risks by providing formal guarantees that limit the potential for information leakage, even when models are trained on sensitive data and released publicly. Part II of this book explores how Differential Privacy can be applied to machine learning and optimization tasks to safeguard individual privacy for increasingly complex data analysis.

1.3 What Protections Does Differential Privacy Provide?

1.3.1 What Does Differential Privacy Promise?

This section examines and defines what Differential Privacy does and does not protect against. It considers the scenario of an analyst or data curator who aims at collecting and aggregate personal and sensitive data for release in a privacy-preserving manner. This release can take various forms, such as a synthetic version of the dataset that masks private information, a set of sensitive population-level statistics about individuals in the dataset, or a model trained on the sensitive data. The common objective in all these cases is to release data that carefully conceal sensitive attributes at the *individual* and *group* level while retaining sufficient information to provide useful statistics or models at the *population* level. For example, an analyst might wish to determine the fraction of a population with a particular disease or calculate the average salary of employees in a company. In these instances, the data pertaining to each individual is private and sensitive, and individuals may prefer to keep it confidential.

First Attempt: No Information Leakage

Ideally, no information about any specific individual should be leaked through the data release, i.e., "nobody can learn *any* information about a specific individual from the privatized computation." Achieving this level of privacy is theoretically straightforward: simply do not collect or use any data at all. However, this is impractical, as it precludes any meaningful data analysis. *Herein lies a fundamental tension highlighted earlier in this chapter: using more data enhances the accuracy and usefulness of the models and statistics but potentially compromises individual privacy.*

Second Attempt: *Almost* No Information Leakage

Rather than requiring that one learns *nothing* about any individual when conducting useful statistical analyses, perhaps one can accept learning *as little as possible* or *almost nothing* about them. Recall the example from the introduction concerning Rick's salary and the addition of noise for privacy. If the company Rick works for is large enough, adding a small amount of noise to the average salary can allow for releasing an approximate estimate of the average salary, while making it difficult to deduce Rick's specific salary. However, the problem here is subtle. It may still be possible to learn significant information about Rick, for instance, that he likely has a high salary because he works at a company where the average salary is close to $500,000. This may seem innocuous if Rick is expected to hold a high-paying position. But consider a more sensitive scenario: imagine that, in the early 1950s, Rick is a smoker participating in a novel medical study investigating the link between smoking and cancer. The study concludes that smoking does cause lung cancer. As a result, anyone who knows that Rick smokes now knows he is at a higher risk of developing lung cancer. His insurance company might increase his premiums or refuse coverage for cancer treatment, citing a pre-existing condition due to his smoking. Clearly, Rick has been *harmed* by the outcome of the study.

Refining the Definition of Privacy

The perspective adopted in this book and by Differential Privacy is that the above scenario does not constitute a privacy violation. Consider a counterfactual world where Rick did not participate in the study. The medical study would still have concluded that smoking causes cancer, and Rick would have faced the same potential harms. Rick's decision to share his data had *(almost) no impact* on the released statistical inference that smoking causes cancer. This outcome is unavoidable: *any* accurate statistical analysis revealing that smoking causes cancer would have had the same effect on Rick. This book takes the point of view that it is important to distinguish between harms arising from the ethical implications of certain statistical inferences and *privacy harms* that result specifically from the collection and use of an individual's data. This redefines what good statistical privacy guarantees should ensure and the refined desiderata: the goal is to ensure that *one can learn almost nothing new about an individual that could not have been inferred had they not shared their data*. It is important to emphasize that Differential Privacy is not an algorithm; it is a *definition* or *requirement* for privacy. The remainder of this chapter aims at accomplishing two goals: *(i)* to carefully formalize the definition and guarantees provided by Differential Privacy, and *(ii)* to cover basic algorithmic techniques and building blocks for achieving Differential Privacy.

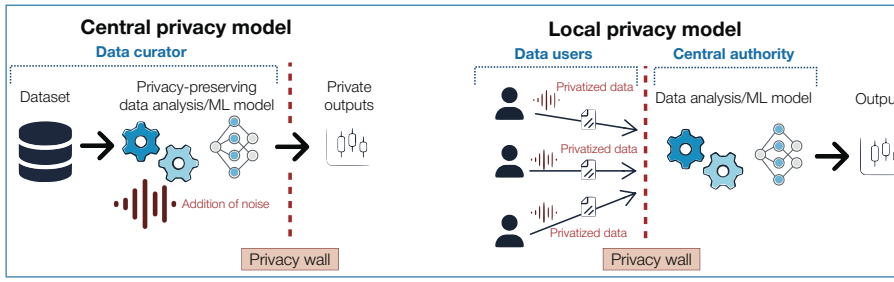

Figure 1.1. Actors and models in the Privacy Preserving data processing pipeline. Central privacy model (left) and Local privacy model (right).

1.3.2 Where to Guarantee Differential Privacy? Local vs Central Models

Implementing Differential Privacy requires careful consideration of the context in which privacy guarantees are applied, particularly regarding the underlying trust model. The degree of trust placed in data curators or aggregators significantly influences the design and effectiveness of privacy-preserving mechanisms. This book examines the two primary frameworks within privacy-preserving ecosystems: the *centralized* model and the *distributed* (or local) model, each with distinct characteristics and implications for privacy management.

In a *centralized* framework, all data collection, storage, and processing occur at a single, central location managed by a trusted data curator. This central entity has direct access to the raw data and is responsible for implementing and monitoring all privacy-preserving mechanisms. The assumption here is that the data curator will faithfully protect individual privacy and handle data responsibly. This setup represents the *central model* in Differential Privacy, as illustrated in Figure 1.1 (left). Conversely, a *distributed* framework keeps data decentralized, residing at its point of origin—such as personal devices or local databases. Privacy-preserving algorithms are executed locally by the data contributors themselves, and only essential, processed information is communicated to a central authority. For instance, in a typical federated learning setup, raw user data remains on their devices, and only privatized versions of the data—such as noisy data points or gradient updates—are sent to the central aggregator. This approach embodies the *local model* in Differential Privacy, depicted in Figure 1.1 (right). Both centralized and distributed frameworks offer distinct advantages and challenges concerning privacy.

Centralized systems concentrate data in one location, creating a single point of failure. If the central entity fails to protect the data—due to a breach or misuse—it can lead to widespread privacy violations affecting all users. Moreover, centralized frameworks require users to *trust* that the platform will implement privacy measures correctly and not exploit the data for unintended purposes. However, the

centralized setting offers significant advantages in terms of data utility and algorithmic flexibility. Because the data curator has access to the raw data, they can inject carefully calibrated noise at the aggregate level, often requiring much less noise to achieve the same privacy guarantees compared to the local setting. This means that analyses and models derived in the centralized setting can be of higher quality and accuracy. Additionally, the centralized model allows for the development of more complex algorithms that require inspecting the data and estimating joint statistics before adding noise—a process that is often challenging or infeasible in the local model.

In contrast, *the distributed model reduces the need for trust in a central authority since privacy is enforced locally by each user*. Even if the central aggregator is compromised, the attacker gains access only to the noisy, privacy-protected data that users have shared. This mitigates the risk associated with a central point of failure and enhances individual control over personal data. However, while the distributed model enhances privacy by minimizing trust requirements, it also introduces additional complexity in implementing privacy-preserving protocols. Each user must correctly execute the algorithms, which may involve sophisticated computations. Additionally, because each user adds noise to their data independently, the aggregated results may suffer from reduced accuracy due to the accumulation of noise. The centralized model, on the other hand, allows for more efficient privacy-utility trade-offs. Since the data curator has access to the raw data, they can add carefully calibrated noise at the aggregate level, achieving the desired privacy guarantees with potentially less impact on data utility. This centralized addition of noise can result in higher-quality data analyses compared to the distributed approach. An in-depth discussion of the local model of DP is provided in Chapter 2.

Distinguishing Data Privacy From Data Security

It is important to differentiate between *data privacy* and *data security* within the landscape of privacy-preserving technologies. Data *security* focuses on preventing unauthorized access to data, implementing measures such as encryption, authentication protocols, and intrusion detection systems to safeguard against breaches and cyber threats. These measures are designed to protect data from external attackers and unauthorized insiders. However, security alone is insufficient to prevent the *inference* of individual-level sensitive information from released data. In contrast, data *privacy*, as addressed in this book, aims at preventing *inference* of individual information when data, statistics, or machine learning models are released. Even when cryptographic security is fully implemented, computing a statistic or training a machine learning model can still allow an attacker to infer individual-level information from the computed statistics or the released model alone, without ever

breaching the system or accessing the original data. How this can occur was illustrated through our earlier example of a differential attack recovering Rick's salary or health status. Differential Privacy thus provides an *orthogonal and complementary* layer of protection to traditional data security techniques. *While data security aims at preventing unauthorized data access, Differential Privacy limits the potential harm from running inference or reconstruction attacks on released databases, statistics, and models.*

1.4 Differential Privacy: Formal Definition, Techniques, and Properties

Differential Privacy is a mathematical framework for measuring and bounding the individuals' privacy risks in a computation. The concept, first introduced in 2006 by Dwork, McSherry, Nissim, and Smith in [DMNS06], informally states that the presence or absence of any individual record in a dataset should not significantly affect the outcome of a mechanism. In this book, a *mechanism* is defined as any computation that can be performed on the data. Differential Privacy deals with randomized mechanisms, and a mechanism is considered *differentially private* if the probability of any outcome occurring is nearly the same for any two datasets that differ in only one record.

In this context, an adversary is any entity attempting to infer sensitive information about individuals from the output of a data analysis. Remarkably, the privacy guarantee of Differential Privacy holds even if the adversary possesses unlimited computing power and complete knowledge of the algorithm and system used to collect and analyze the data. Thus, even if the adversary were to develop new and sophisticated methods, including the attack methods discussed earlier, as well as new attacks that do not yet exist today, or even if new additional external information becomes available, Differential Privacy provides the exact same level of protection. In this sense, Differential Privacy is considered *future-proof*.

The section, next, reviews *Randomized Response*, a classic method adopted in surveys for ensuring the privacy of respondents. Originally developed as a survey technique to encourage honest responses to sensitive questions, Randomized Response leverages randomness to protect individual privacy while still allowing researchers to estimate population characteristics accurately. This method serves as a foundational example of how randomness can be systematically used to achieve Differential Privacy, illustrating the principles that guide more complex privacy-preserving mechanisms discussed later in this section, and throughout the book.

Randomized Response

Randomized response [War65] was proposed by Warner in 1965 to privately survey respondents for a potentially sensitive property. The setup is as follows: one wishes to test for how many individuals in a set of respondents have a certain property, $\mathcal{P}$, which might be a controversial one to possess, and this might ordinarily lead to a subset of respondents becoming what Warner described as a "non-cooperative" group, who might refuse to be surveyed or provide a dishonest answer, introducing unwanted bias in the survey results. To simultaneously ensure that respondents answer honestly (and, as a result, avoid bias due to the aforementioned non-cooperation) and that their privacy is not violated, randomized response provides respondents the ability to deny their response while also preserving the quality of the summary statistics inferred. This is ensured by introducing randomness into the process of surveying as follows.

1. The respondent takes a fair coin and flips it;
2. If *tails* is obtained, then the respondent answers truthfully, and if *heads* is obtained, the respondent flips the coin again and
   (a) Responds affirmatively if the outcome is heads;
   (b) Responds negatively if the outcome is tails.

Note that here the outcomes and numbers of coin flips are only known to the respondent. The property of *plausible deniability* allows respondents to be able to deny their responses, and this provides them with privacy guarantees (as it will be elaborated later). While the responses are partly perturbed due to this process, an analyst can recover the expected number of "Yes" responses accurately as follows,

$$\mathbb{E}[\text{Yes}] = \frac{3}{4}n(\text{has } \mathcal{P}) + \frac{1}{4}n(\text{does not have } \mathcal{P}),$$

where $\mathbb{E}[\text{Yes}]$ is the expected number of affirmative responses, and $n(X)$ is the number of respondents who claimed to satisfy property X.

As will become clear later in this section, the plausible deniability property of randomized response has a strong connection with Differential Privacy.

1.4.1 Differential Privacy, Formally

Prior to defining Differential Privacy formally, this section formalizes what this privacy notion aims at protecting (dataset) and the means by which an analyst interacts with data (queries).

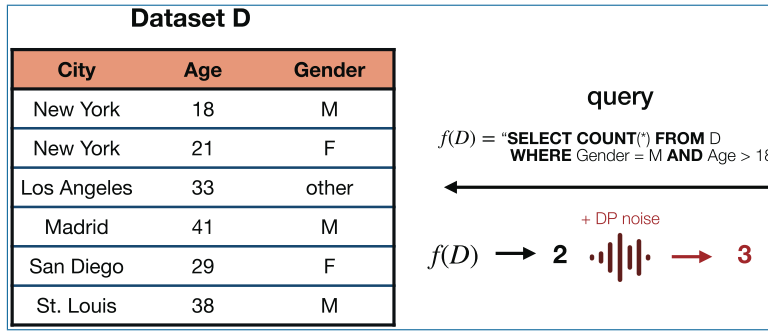

Figure 1.2. Example dataset and query.

Datasets and Queries

A *dataset* D is a multi-set of elements in the *data universe* $\mathcal{U}$. The set of every possible dataset is denoted $\mathcal{D}$. The data universe $\mathcal{U}$ is a cross product of multiple *attributes* $U_1, \ldots, U_n$ and has *dimension* n. For example, Figure 1.2, illustrates a dataset D with three attributes: city, age, and gender. If C is the set of all cities considered, the interval $A = [0, 100]$ the set of all ages considered, and $G = \{\text{M}, \text{F}, \text{other}\}$, then $\mathcal{U} = \mathcal{C} \times \mathcal{A} \times \mathcal{G}$. A *numeric query* is a function $f : \mathcal{D} \to \mathcal{R} \subseteq \mathbb{R}^r$ that maps a dataset in some real vector space. For instance, the query $f(D)$ could be an *SQL statement* that counts the number of male individuals over the age of 18 in dataset D, as illustrated in Figure 1.2.

The concept of *adjacency* is fundamental in DP. It frames the *unit of change* that Differential Privacy seeks to protect against, ensuring that the presence or absence of any single individual's data does not significantly alter the outcomes of data analysis. There are two common ways to define adjacency in the context of Differential Privacy, reviewed next.

Definition 1.2 (Add/remove adjacency)**.** *Two datasets D and D' are said adjacent under the add/remove notion, denoted as $D \sim D'$, if $|D \Delta D'| = 1$, where $\triangle$ is the symmetric difference of two sets.*

In other words, two datasets are defined as adjacent if one can be obtained from the other by either adding or removing the data of a single individual. This model is particularly relevant when considering the impact of an individual's participation or absence in the dataset.

Definition 1.3 (Exchange adjacency)**.** *Two datasets D and D' are said* adjacent *under the exchange notion, denoted as $D \sim_{\rightleftarrows} D'$, if D' is obtained from D by successively removing one record and then adding a (possibly different) record. That is, there exist elements $d \in D$ and $d' \in \mathcal{U}$ such that: $D' = (D \setminus \{d\}) \cup \{d'\}$. This implies that $|D| = |D'|$ and $|D \triangle D'| = 2$.*

In this notion, adjacent datasets differ in the data of exactly one individual but have the same size. This definition is suited to scenarios where the alteration of data within a constant-size dataset is the primary concern, and can be viewed as the removal followed by the addition of one individual.

The choice between add/remove or exchange adjacency has some implications for how Differential Privacy is applied as it directly affects the computation of global sensitivity, introduced next, which measures the maximum change in the output of a function for adjacent datasets. This chapter, and generally the book unless specified otherwise, adhere to the add/remove notion of adjacency.

Global Sensitivity

The impact of a single individual's data on the overall analysis is measured through the concept of *global sensitivity*. Formally, the global sensitivity of a function $f : \mathcal{D} \to \mathcal{R}$ is defined as the maximum difference in the output of f over all pairs of adjacent datasets $D \sim D' \in \mathcal{D}$, measured with respect to the ℓ_p norm:

$$\Delta_p f = \max_{D \sim D'} \|f(D) - f(D')\|_p. \tag{1.1}$$

In simpler terms, it measures how much the output of a function can change when an individual's data is added or removed from the dataset. This measurement provides a basis for determining the amount of noise that needs to be added to the function's output to achieve privacy. For example, the query considered in Figure 1.2 that counts the number of individuals satisfying a certain property in a dataset has global sensitivity 1, since adding or removing a single individual in the dataset can affect the final count by at most 1. Suppose instead that the task is to compute the average age of all individuals in the dataset. Then the global sensitivity of this average function would be

$$\Delta_p f = \frac{\max(A) - \min(A)}{|D|} = \frac{100}{|D|},$$

where A represents the range of possible ages (assuming ages range from 0 to 100). In this chapter, the ℓ_1-sensitivity $\Delta_1 f$ is denoted with Δf.

Differential Privacy

These examples illustrate how a single individual's data can influence the output of a function applied to a dataset. This influence is central to the concept of Differential Privacy. The impact of adding or removing an individual's data varies depending on the type of function in question—whether it's calculating sums, averages, or any other measure of the data. This sensitivity measurement tells us how much the output of the target function need to be adjusted in order to protect an individual's privacy. Differential Privacy achieves this by adding noise to the function's output,

by an amount calibrated to the function sensitivity. This approach ensures that the presence or absence of any single individual's data does not significantly alter the output, thereby masking their participation.

Definition 1.4 (Differential Privacy [DMNS06])**.** *A randomized mechanism* $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ *with domain* $\mathcal{D}$ *and range* $\mathcal{R}$ *is* (ε, δ)-differentially private *if, for any event* $S \subseteq \mathcal{R}$ *and any pair* $D, D' \in \mathcal{D}$ *of adjacent datasets:*

$$Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) Pr[\mathcal{M}(D') \in S] + \delta, \tag{1.2}$$

where the probability is calculated over the randomness of $\mathcal{M}$*.*

A differentially private mechanism maps a dataset to a distribution over the possible outputs because, e.g., it adds random noise or makes randomized choices. The released DP output is a single random sample drawn from this distribution. The level of privacy is controlled by the parameter $\varepsilon \geq 0$, called the *privacy loss*, with values close to 0 denoting strong privacy, and a secondary parameter δ which can be loosely interpreted as a margin of error.

First, observe that the inequality:

$$Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) Pr[\mathcal{M}(D') \in S],$$

(here with $\delta = 0$ for simplicity of exposition) holds for any D and D'. In particular, since it holds for any pair of neighboring databases, it also holds when *swapping* the roles of D and D' in the above definition. Hence, an $(\varepsilon, 0)$-differentially private algorithm must also satisfy

$$Pr[\mathcal{M}(D') \in S] \leq \exp(\varepsilon) Pr[\mathcal{M}(D) \in S].$$

This directly implies the "stronger" inequality below:

$$\exp(-\varepsilon) Pr[\mathcal{M}(D') \in S] \leq Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) Pr[\mathcal{M}(D') \in S], \tag{1.3}$$

which highlights that the probabilities of any event S under $\mathcal{M}$ applied to D and D' are close to each other, controlled by ε.

To intuitively understand these parameters, think of ε as a knob controlling the level of privacy. Lowering ε enhances privacy by making the outputs less sensitive to changes in any individual's data. As ε approaches zero (with $\delta = 0$), the inequality in Equation (1.3) forces the distributions $\mathcal{M}(D)$ and $\mathcal{M}(D')$ to become nearly identical. This means *more* privacy, as distinguishing between D and D', which is necessary to recover the data of the individual that differs across both databases, becomes harder. When $\varepsilon = 0$, $Pr[\mathcal{M}(D) \in S] = Pr[\mathcal{M}(D') \in S]$ for all S; i.e., the output is independent of and does not use the input dataset, providing *perfect* privacy, but *no utility*—mathematically, an easy implication of $\varepsilon = 0$ is that

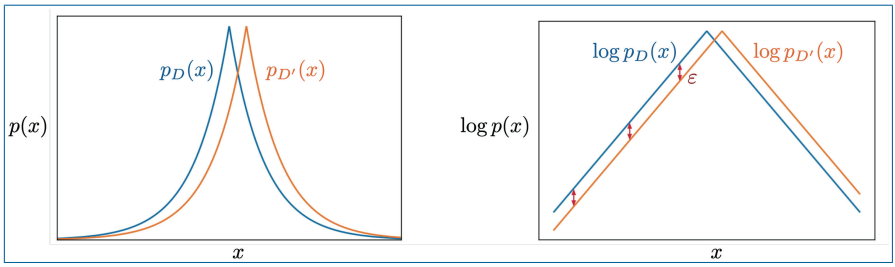

Figure 1.3. An illustration of the ε-DP guarantee (here, using the Laplace mechanism of Section 1.4.3). The log-probability of a value to be output by a mechanism given two neighboring datasets is bounded by ε.

our mechanism must have a (trivially) constant output across all datasets. When $\varepsilon \to +\infty$, the inequality is always satisfied by any mechanism and *no privacy* is guaranteed.

The parameter δ serves as a margin of error. It is typically a small number close to zero that defines a *failure threshold* allowing the DP guarantee not to hold with a probability of up to δ[1]. In practice, δ is chosen to be a negligible value, often much smaller than $\frac{1}{N}$, where N is the size of the dataset. This ensures that the likelihood of disclosing sensitive information about any individual remains extremely low. A mechanism satisfying $(\varepsilon, 0)$-differential privacy is said to satisfy *pure* Differential Privacy or ε-Differential Privacy.

The guarantees of Differential Privacy are illustrated in Figure 1.2, which shows the distribution of outputs from a differentially private mechanism applied to two adjacent datasets D and D'. The blue and red curves represent the probability distributions of the outputs for D and D', respectively. The left figure shows how the probability distributions over outputs must be close to each other for adjacent datasets. The right figure quantifies the difference between the probabilities, showing that the log-probabilities of any outcome x differ by at most ε. This means that the ratio of probabilities is bounded by e^{ε}, as required by the definition. One can see that requiring a smaller ε forces the distributions to be closer to each other across D and D', making it harder to distinguish between the two databases and hence providing stronger privacy protections.

1.4.2 Formal Properties of Differential Privacy

This section formalizes the properties guaranteed by Differential Privacy, and how they match the desirata described in Section 1.2. The composition, group privacy,

1. $(\varepsilon, 0)$-DP most of the time, except with probability δ, and (ε, δ)-DP are closely related but not exactly equivalent.

and post-processing properties are derived directly from the direction of Differential Privacy, and do not assume a specific mechanism like Randomized Response. As such, composition. group privacy, and post-processing hold for *any* differentially private mechanism, i.e. *any mechanism that satisfies requirement* (1.2).

Composition

Composition ensures that a combination of differentially private mechanisms (whether the mechanisms release privatized data, statistics on data, or learning models) preserves Differential Privacy. Composition is a key concept that enables the construction of complex algorithms by combining simpler primitives. It facilitates *privacy accounting*, the rigorous analysis of the overall privacy loss of a composite and potentially complex algorithm by aggregating the privacy guarantees of individual primitives. More formally, it can be stated as follows [DR14]:

Theorem 1.5 (Composition)**.** *Let $\mathcal{M}_i : \mathcal{D} \to \mathcal{R}_i$ be an ε_i-differentially private mechanism for $i \in \{1, 2\}$. Then, their composition, defined as $\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D))$, is $(\varepsilon_1 + \varepsilon_2)$-differentially private.*

Proof. For any $(R_1, R_2) \subseteq \mathcal{R}_1 \times \mathcal{R}_2$ and any two neighboring datasets $D \sim D'$,

$$\begin{aligned}
\frac{Pr[\mathcal{M}(D) \in (R_1, R_2)]}{Pr[\mathcal{M}(D') \in (R_1, R_2)]} &= \frac{Pr[\mathcal{M}_1(D) \in R_1]Pr[\mathcal{M}_2(D) \in R_2]}{Pr[\mathcal{M}_1(D') \in R_1]Pr[\mathcal{M}_2(D') \in R_2]} \\
&= \left(\frac{Pr[\mathcal{M}_1(D) \in R_1]}{Pr[\mathcal{M}_1(D') \in R_1]}\right)\left(\frac{Pr[\mathcal{M}_2(D) \in R_2]}{Pr[\mathcal{M}_2(D') \in R_2]}\right) \\
&\leq \exp(\varepsilon_1)\exp(\varepsilon_2) \\
&= \exp(\varepsilon_1 + \varepsilon_2).
\end{aligned}$$

□

This argument can be generalized to for k differentially private mechanisms by induction. More precisely, if $\mathcal{M}_i : \mathcal{D} \to \mathcal{R}_i$ is an ε_i-differentially private mechanism for $i = 1, \ldots, k$. Then, the composition $\mathcal{M}(D) = (\mathcal{M}_1(D), \ldots, \mathcal{M}_k(D))$ is $(\sum_{i=1}^{k} \varepsilon_i)$-differentially private. The result above is also called *simple composition*, as it deals with pure Differential Privacy mechanisms. An extensive treatment of composition in Differential Privacy is deferred to Chapter 3.

Group Privacy

The Differential Privacy notions discussed so far bound differences in output distributions of the mechanism for any pairs of adjacent datasets, i.e. for datasets D, D' such that $|D \Delta D'| = 1$. However, what is not immediately clear is the case when

two datasets differ in more than one individual's data. Fortunately, Differential Privacy yields group privacy guarantees that bound this difference for datasets that differ in k entries, for $k > 0$:

Theorem 1.6 (Group privacy). *Let $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ be an ε-differentially private algorithm. Suppose D and D' are two datasets that differ in exactly k entries. Then, for all $S \subseteq \mathcal{R}$:*

$$Pr[\mathcal{M}(D) \in S] \leq \exp(k\varepsilon) Pr[\mathcal{M}(D') \in S].$$

Proof. Let $D^{(0)} \triangleq D$ and $D^{(k)} \triangleq D'$, and let $D^{(0)}, D^{(1)}, \ldots, D^{(k-1)}, D^{(k)}$ be a sequence of datasets where $D^{(i)} \sim D^{(i+1)}$ for $i = 0, 1, \ldots, k-1$. The datasets in this sequence can be thought of as "intermediate" datasets when trying to obtain D' by starting with D and changing one entry at a time successively. Then by the DP guarantee of $\mathcal{M}$, for any $R \subseteq \mathcal{R}$ and $i \in [k-1]$,

$$Pr[\mathcal{M}(D^{(i)}) \in R] \leq \exp(\varepsilon) Pr[\mathcal{M}(D^{(i+1)}) \in R].$$

Then, for any $R \subseteq \mathcal{R}$,

$$\begin{aligned} Pr[\mathcal{M}(D) \in R] &= Pr[\mathcal{M}(D^{(0)}) \in R] \\ &\leq \exp(\varepsilon) Pr[\mathcal{M}(D^{(1)}) \in R] \\ &\leq \exp(2\varepsilon) Pr[\mathcal{M}(D^{(2)}) \in R] \\ &\vdots \\ &\leq \exp(k\varepsilon) Pr[\mathcal{M}(D^{(k)}) \in R] \\ &= \exp(k\varepsilon) Pr[\mathcal{M}(D') \in R]. \end{aligned}$$

□

Post-processing

Another key property of Differential Privacy is post-processing immunity. It ensures that privacy guarantees are preserved by arbitrary data-independent post-processing steps [DR14]:

Theorem 1.7 (Post-Processing Immunity). *Let $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ be a mechanism that is ε-differentially private and $g : \mathcal{R} \to \mathcal{R}'$ be a data-independent mapping. The mechanism $g \circ \mathcal{M}$ is ε-differentially private.*

Proof. The proof first considers a deterministic mapping $g : \mathcal{R} \to \mathcal{R}'$. Let $\tilde{S} \triangleq \{r \in \mathcal{R} : g(r) \in S\}, \forall S \subseteq \mathcal{R}'$. Then for any two neighboring datasets $D \sim D'$,

$$\begin{aligned} Pr[g \circ \mathcal{M}(D) \in S] &= Pr[\mathcal{M}(D) \in \tilde{S}] \\ &\leq \exp(\varepsilon) Pr[\mathcal{M}(D') \in \tilde{S}] \\ &= \exp(\varepsilon) Pr[g \circ \mathcal{M}(D') \in S]. \end{aligned}$$

This proves post-processing immunity for deterministic functions. To extend this guarantee to randomized functions, note that randomized functions can be viewed as a distribution over deterministic functions, and, in particular as a convex combination of deterministic functions. Given that a convex combination of differentially private mechanisms (here each mechanism is obtained by composing each deterministic function with the mechanism $\mathcal{M}$) is also differentially private, the result follows. □

This property ensures that, once Differential Privacy guarantees are applied, any further analysis or manipulation of the protected results will not compromise its privacy guarantees. Post-processing significantly expands the scope and applicability of Differential Privacy algorithms in real-world applications, as shown in Part III.

Quantifiable Privacy-accuracy Trade-offs

The last important property, mentioned in Section 1.2, the trade-off between privacy and accuracy can be quantified exactly. Privacy-accuracy trade-offs are *mechanism-level* properties: each mechanism has its own trade-off. The privacy-accuracy trade-offs of the main building blocks are described later in this section, including the privacy-accuracy trade-offs of *Randomized Response* in Section 1.7, of the *Laplace Mechanism* in Section 1.4.3, and of the *Gaussian Mechanism* in Section 1.5.1.

1.4.3 The Laplace Mechanism

The Laplace Distribution with 0 mean and scale b has a probability density function $\text{Lap}(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}$. The Laplace mechanism is a differentially private mechanism based on the Laplace distribution for answering numeric queries [DMNS06]. It is a fundamental building block for many DP algorithms described in this book, and itfunctions by simply computing the output of the query f and then perturbing each coordinate with noise drawn from the Laplace distribution. The scale b of the noise is calibrated to the query sensitivity Δf divided by ε:

Definition 1.8 (The Laplace Mechanism)**.** *Let $f : \mathcal{D} \to \mathcal{R} \subseteq \mathbb{R}^d$ be a numerical query, with d being a positive integer. The Laplace mechanism is defined as*

$\mathcal{M}_{Lap}(D; f, \varepsilon) = f(D) + Z$ *where* $Z \in \mathcal{R}$ *is a vector of i.i.d. samples drawn from* $Lap(\frac{\Delta f}{\varepsilon})$.

The Laplace mechanism adds random noise drawn from the Laplace distribution independently to each of the d dimensions of the query response.

Theorem 1.9 (Differential Privacy of The Laplace Mechanism)**.** *The Laplace mechanism,* $\mathcal{M}_{Lap}$*, achieves* $(\varepsilon, 0)$*-Differential Privacy.*

Proof. Let $D \sim D'$ be any two neighboring datasets in $\mathcal{D}$, and let p_D and $p_{D'}$ be the probability density functions of $\mathcal{M}_{\text{Lap}}(D; f, \varepsilon)$ and $\mathcal{M}_{\text{Lap}}(D'; f, \varepsilon)$, respectively. Then for any $r \in \mathcal{R}$,

$$\begin{aligned}
\frac{p_D(r)}{p_{D'}(r)} &= \prod_{i=1}^{d} \left(\frac{\exp\left(-\frac{\varepsilon |f(D)_i - r_i|}{\Delta f}\right)}{\exp\left(-\frac{\varepsilon |f(D')_i - r_i|}{\Delta f}\right)} \right) \\
&= \prod_{i=1}^{d} \exp\left(\frac{\varepsilon(|f(D')_i - r_i| - |f(D)_i - r_i|)}{\Delta f} \right) \\
&\leq \prod_{i=1}^{d} \exp\left(\frac{\varepsilon(|f(D')_i - f(D)_i|)}{\Delta f} \right) && \text{(By the triangle inequality.)} \\
&= \prod_{i=1}^{d} \exp\left(\frac{\varepsilon \cdot (\|f(D) - f(D')\|_1)}{\Delta f} \right) && \text{(By the definition of } \Delta f.) \\
&\leq \exp(\varepsilon).
\end{aligned}$$

The proof is similar for $\frac{p_{D'}(r)}{p_D(r)} \leq \exp(\varepsilon)$. □

A graphical representation of the densities and log density of two Laplace distributions associated with neighboring datasets D and D' are provided in Figure 1.3, respectively. Note how the difference between the log probabilities for x for each of the neighboring datasets $D \sim D'$ is bounded by ε.

Accuracy Guarantee of the Laplace Mechanism

The accuracy guarantees of the Laplace Mechanism is characterized by the following result.

Theorem 1.10. *For any numerical query* $f : \mathcal{D} \to \mathcal{R} \subseteq \mathbb{R}^d$*, and any database* $D \in \mathcal{D}$*,*

$$Pr\left[\left| f(D) - \mathcal{M}_{Lap}(D; f; \varepsilon) \right| \geq \ln\left(\frac{d}{\beta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right) \right] \leq \beta.$$

Proof. The proof is for $d = 1$ for simplicity, but it generalizes for $d > 1$. The proof follows from characterizations of the tails of the Laplace distribution. For a random variable $Z \sim Lap(b)$ and a real number $\alpha > 0$,

$$Pr\left[|Z| \geq \alpha\right] = \exp\left(-\alpha/b\right).$$

Therefore, given that $f(D) - \mathcal{M}_{Lap}(D; f; \varepsilon)$ is Laplace with parameter $b = \frac{\Delta f}{\varepsilon}$, it follows that

$$Pr\left[\left|f(x) - \mathcal{M}_{Lap}(D; f; \varepsilon)\right| \geq \alpha\right] = \exp\left(-\alpha \cdot \frac{\varepsilon}{\Delta f}\right) \triangleq \beta.$$

Solving for α in $\exp\left(-\alpha \cdot \frac{\varepsilon}{\Delta f}\right) = \beta$ leads to

$$\alpha \cdot \frac{\varepsilon}{\Delta f} = \ln\left(\frac{1}{\beta}\right),$$

hence

$$\alpha = \ln\left(\frac{1}{\beta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right).$$

This concludes the proof. □

The accuracy guarantee of the Laplace Mechanism provides a practical way to understand how the added noise affects the utility of the released data while ensuring differential privacy. Essentially, it quantifies the expected deviation between the true value of a numerical query and the noisy output produced by the mechanism.

1.4.4 Answering Private Queries in Practice

Next, we present two examples to illustrate how the Laplace Mechanism can be applied in practice.

Example 1: Computing the Average Age

Consider a dataset containing the ages of 10,000 individuals, with ages ranging from 0 to 100 years. The task is to compute the average age while ensuring differential privacy. A practical procedure follows the following steps:

1. *Determine the query function and its sensitivity.* In this task the query function is the average age,

$$f(\text{data}) = \frac{1}{n}\sum_{i=1}^{n} \text{age}_i,$$

where n is the number of individuals in the dataset. The global sensitivity Δf of the average function is the maximum change in the output when one individual is added or removed. Since the age can vary between 0 and 100, adding the data about a single individual can affect the sum by at most 100 units. Therefore, the sensitivity is:

$$\Delta f = \frac{\text{max age} - \text{min age}}{n} = \frac{100 - 0}{10{,}000} = 0.01.$$

2. *Apply the Laplace Mechanism.* The next step is to select the privacy parameter ε and add noise drawn from the Laplace distribution with scale parameter $\frac{\Delta f}{\varepsilon}$. Selecting $\varepsilon = 0.5$ to obtain a strong privacy guarantee adds the following noise:

$$\text{noise} \sim \text{Lap}\left(\frac{\Delta f}{\varepsilon}\right) = \text{Lap}\left(\frac{0.01}{0.5}\right) = \text{Lap}(0.02).$$

The private query thus reports $f(\text{data}) + \text{noise}$.

3. *Analyze the error bound.* Additionally, by setting a confidence level $\beta = 0.05$ (meaning that one is 95% confident in the error bound), the error bound can be computed as,

$$\text{Error Bound} = \frac{\Delta f}{\varepsilon} \ln\left(\frac{1}{\delta}\right) = \frac{0.01}{0.5} \ln\left(\frac{1}{0.05}\right) \approx 0.06 \text{ years}.$$

This means that, with 95% confidence, the noisy average age returned by the Laplace Mechanism will differ from the true average age by no more than approximately 0.06 years. If the privacy parameter is set to $\varepsilon = 1$, allowing for slightly less privacy in exchange for greater accuracy, the error bound decreases to about 0.03 years. Thus, selecting ε and β appropriately ensures that the released data remains both useful and privacy-preserving.

Example 2: Releasing a Histogram

Suppose a statistical agency wants to release a histogram showing the number of individuals in different age groups, segmented by gender and region, from a dataset containing a large number of respondents. The age groups could be categorized in intervals (e.g., 0–9, 10–19, …, 90+). The goal is to release this histogram while ensuring differential privacy. Note that this is different from the previous task where a single quantity wasreleased. The procedure again follows the the three same steps:

1. *Determine the query function and its sensitivity.* The query function is the count of individuals in each combination of age group, gender, and region. For count queries, the global sensitivity Δf is 1 because adding or removing one individual can change the count in one category by at most 1.

2. *Apply the Laplace Mechanism.* The next step consists in selecting a privacy parameter $\varepsilon = 0.5$ for each count in the histogram and adding independent Laplace noise to each cell (i.e., each combination of age group, gender, and region) in the histogram. Let $c_{i,j,k}$ be the true count for age group i, gender j, and region k, and $\tilde{c}_{i,j,k}$ is the private counterpart to be released. The counts are linked by the following formula:

$$\tilde{c}_{i,j,k} = c_{i,j,k} + \text{Noise}_{i,j,k}, \quad \text{where } \text{Noise}_{i,j,k} \sim \text{Laplace}\left(\frac{\Delta f}{\varepsilon}\right) = \text{Laplace}(2).$$

3. *Post-processing to ensure valid counts.* Notice that the application of real-valued noise to each count may render the resulting privacy-preserving counterpart negative or non-integers, thus producing invalid ouputs. These issues can be corrected by applying a post-processing step, that set any negative noisy counts to zero and round the noisy counts to the nearest integer. Such post-processing steps do not alter the privacy guarantees of the original release and are commonly applied in deployments [CDMS21].
4. *Analyze privacy and utility.* Each count is ε-differentially private with $\varepsilon = 0.5$. Since each individual's data affects only one count, and the counts are disjoint, the overall privacy guarantee remains $\varepsilon = 0.5$. The added Laplace noise has a mean of zero and a scale of 2 and thus the expected absolute error for each count is 2. For categories with large counts, this noise has a relatively small impact. However, for categories with small counts, especially in less populated age groups or regions, the noise can significantly affect the accuracy. A further analysis on disparate impacts of Differential Privacy on different subpopulations is discussed in Chapter 17.

Note that other mechanisms can produce integer counts directly without additional rounding, by using discrete noise mechanisms, such as the Geometric mechanism [GRS12] and the discrete Laplace mechanism [KS12].

1.5 Approximate Differential Privacy

The discussion in the previous section focused on pure Differential Privacy and the mechanisms and guarantees associated with it. The case where $\delta > 0$ for (ε, δ)-DP constitutes a variant of Differential Privacy known as *Approximate Differential Privacy*. Recall that $\delta \in (0, 1)$ is the failure probability of the privacy loss bound in the relaxed variant of pure DP, and is meant to be a cryptographically low quantity—that is, so small it is considered negligible for practical purposes, often much less than $\frac{1}{N}$ where N is the dataset size. This allows practitioners to apply other mechanisms which yields better utility than the Laplace mechanism in exchange for a

marginal failure probability. Importantly, approximate Differential Privacy retains the composition, group privacy, and post-processing immunity properties provided by pure Differential Privacy.

1.5.1 The Gaussian Mechanism

The canonical mechanism for (ε, δ)-DP is the Gaussian mechanism [DR14]. Where the Laplace mechanism adds noise proportionally to the ℓ_1 sensitivity of a query f, Δf, the Gaussian mechanism uses the ℓ_2 sensitivity, denoted by $\Delta_2 f$, and defined as in Equation (1.1) with $p = 2$. The ℓ_2 and ℓ_1 norms enjoy the following relationship: for a vector $x \in \mathbb{R}^d$, $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$. Thus, the ℓ_2 sensitivity can be up to a factor $\sqrt{d}$ less than the ℓ_1 sensitivity. The Gaussian distribution with 0 mean and standard deviation σ has the probability density function $\mathcal{N}(x|\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2)}{2\sigma^2}\right)$.

Definition 1.11 (Gaussian Mechanism)**.** *Let $f : \mathcal{D} \to \mathcal{R}$ be a numerical query. The Gaussian mechanism is defined as $\mathcal{M}_{Gauss}(D; f, \varepsilon) = f(D) + z$ where $z \in \mathcal{R}$ is a vector of i.i.d. samples drawn from $\mathcal{N}\left(0, \sigma^2 I\right)$ where $\sigma \geq \sqrt{2\ln(\frac{1.25}{\delta})}(\Delta_2 f/\varepsilon)$.*

As with the Laplace mechanism, the numerical query response is d-dimensional for some integer $d > 0$ as well. Gaussian noise is added to each dimension of the query response independently by the Gaussian mechanism. To highlight a key distinction between the Laplace and Gaussian mechanisms, consider the context of computing the mean of a multivariate dataset, revisited from [Kam20]. Consider a dataset $D \in \{0, 1\}^{n \times d}$ aiming to compute the mean in a privacy-preserving manner, denoted by $f(D) = \frac{1}{n}\sum_{i=1}^{n} D_i$. The maximum discrepancy in f across adjacent datasets is $\frac{1}{n}1$, yielding a vector with ℓ_1 norm of $\frac{d}{n}$ and ℓ_2 norm of $\sqrt{d/n}$ as the ℓ_1 and ℓ_2 sensitivities. The following theorem defines the (ε, δ)-DP guarantees for the Gaussian mechanism.

Theorem 1.12. *The Gaussian mechanism, $\mathcal{M}_{Gauss}$, achieves (ε, δ)-Differential Privacy, for $\varepsilon \in (0, 1]$ and $\delta \in [0, 1]$.*

For the proof of this theorem, see Appendix A of [DR14]. Notice that, in the original proposition, also reviewed in [DR14], the mechanism is restricted to use ε within $(0, 1]$. However, it is not uncommon to see values of $\varepsilon > 1$ in practice, including in various discussions in this book. This restriction was studied and overcome in [BW18], which provided a more general *analytical Gaussian mechanism* that holds for $\varepsilon > 1$ as well. While the details of the DP guarantee of the analytical Gaussian mechanism are beyond the scope of this text, the mechanism and the associated (ε, δ)-DP is defined as follows.

Theorem 1.13. *(Analytical Gaussian Mechanism [BW18]). Let $f : \mathcal{D} \to \mathcal{R}$ be a numerical query with global ℓ_2 sensitivity $\Delta_2 f$. $\forall \varepsilon > 0$ and $\delta \in [0, 1]$, the Gaussian mechanism $\mathcal{M}_{Gauss}(D; f, \varepsilon) = f(D) + z$ with $z \sim \mathcal{N}(0, \sigma^2 I)$ satisfies (ε, δ)-Differential Privacy if and only if*

$$\Phi\left(\frac{\Delta_2 f}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2 f}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta_2 f}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2 f}\right) \leq \delta.$$

Where $\Phi(t) = Pr[\mathcal{N}(0, 1) \leq t] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-y^2/2} dy$ is the CDF of the standard univariate Gaussian distribution.

The reader is referred to [BW18] for the details on the analytical Gaussian mechanism.

Discussion of Accuracy

The exact formal accuracy guarantees is left as an exercise to the reader. The proof is similar to that of the accuracy guarantee for the Laplace mechanism, simply quantifying tails on the Gaussian distribution. Note that, in high-dimensions, the Laplace mechanism introduces noise scaled by $\frac{d}{n\varepsilon}$ to each dimension, providing an ε-DP estimate of f with an ℓ_2 error scaling as $O(\frac{d^{3/2}}{n\varepsilon})$. In contrast, the Gaussian mechanism introduces noise with a scale of $O(\sqrt{\frac{d \log(1/\delta)}{n\varepsilon}})$ per dimension, resulting in an (ε, δ)-DP estimate of f with ℓ_2 error approximately $O(\frac{d}{n\varepsilon})$. Thus the Gaussian mechanism shaves of a factor of $O(\sqrt{d})$ from the noise, improving accuracy significantly for large d at a slight cost to the privacy guarantee, positing it as a potentially more effective approach for multi-variate estimations.

1.6 Beyond Statistical Queries: Differentially Private Selection

Numerical queries form an important class of computations over which privacy can be enforced. However, in many natural situations, *the goal may be to output an object selected according to certain criteria among other objects, rather than just a numerical value.* Consider the following example, adapted from [DR14]. Suppose that a retailer is selling an amount of items for which there are 3 potential buyers A, B, and C. Each buyer has a maximum price they are willing to pay for the item, known as their *valuation*. The buyers wish to keep their valuations private, to avoid disclosing sensitive information about their purchasing strategies or financial standing. *Hence the task of the retailer is to set a sale price to maximize their total revenue without revealing the valuations of the buyers in the process.*

Assume that the valuations of buyers A, B and C are, respectively \$1.00, \$1.01, and \$3.01. Consider the possible pricing options:

- **Price at \$1.00**: All three buyers are willing to purchase at this price, thus the total revenue is $\$1.00 \times 3$ buyers $=$ **\$3.00**.
- **Price at \$1.01**: Buyers B and C are willing to purchase, thus the total revenue is $\$1.01 \times 2$ buyers $=$ **\$2.02**.
- **Price at \$3.01**: Only buyer C is willing to purchase, thus the total revenue is $\$3.01 \times 1$ buyer $=$ **\$3.01**.

To maximize revenue, the retailer should set the price at \$3.01. However, since the buyers' valuations are private, the seller cannot directly know the optimal price. The seller needs to select a price in a privacy-preserving manner. One naive approach might be for the seller to add random noise to the buyers' valuations to preserve their privacy. Suppose the seller adds noise to buyer C's valuation, and it becomes, say, \$3.02. Based on this noisy valuation, the seller decides to set the price at \$3.02 apiece. However, this approach leads to a problem: at a price of \$3.02, none of the buyers are willing to purchase the item, since their true valuations are all below this price. Consequently, the total revenue would be **\$0**, which is worse than any of the previous pricing options. This illustrates that simply adding noise to the valuations is not suitable for such a setting. Adding noise to the valuations can lead to suboptimal pricing decisions. Small changes in the valuations (due to noise) can result in significant differences in the optimal price, which may drastically reduce the seller's revenue or eliminate it altogether. This is particularly problematic when the output is an object selection (the optimal price) rather than a simple numerical query.

1.6.1 The Exponential Mechanism

To be able to perform selection privately while also preserving the quality of the selection made, McSherry and Talwar defined the exponential mechanism [MT07]. Given a set of objects $\mathcal{H}$, a dataset $D \in \mathcal{D}$, and a score function $s : \mathcal{D} \times \mathcal{H} \to \mathbb{R}$, the exponential mechanism chooses an object $h \in \mathcal{H}$ that maximizes the score function in a differentially private manner.

Definition 1.14 (Exponential Mechanism)**.** *The exponential mechanism, denoted by $\mathcal{M}_{exp}$, takes as input a dataset $D \in \mathcal{D}$, a set of objects $\mathcal{H}$, and a score function $s : \mathcal{D} \times \mathcal{H} \to \mathbb{R}$ and outputs $h \in \mathcal{H}$ with probability proportional to* $\exp\left(\frac{\varepsilon s(D,h)}{2\Delta s}\right)$, *where* $\Delta s \triangleq \max_{h \in \mathcal{H}} \max_{D \sim D'} |s(D,h) - s(D',h)|$.

In this pricing example, the seller defines a utility function $u(D, p)$ that calculates the total revenue generated by setting a price p, given the buyers' valuations in

the dataset D. The exponential mechanism then selects a price p with probability *proportional* – the actual probability needs to be renormalized to sum to 1 – to:

$$\exp\left(\frac{\varepsilon \cdot u(D,p)}{2\Delta u}\right),$$

where ε is the privacy parameter controlling the level of privacy, and Δu is the global sensitivity of the utility function—that is, the maximum change in $u(D,p)$ when a single individual's valuation in D is modified. The seller thus probabilistically chooses a price that is likely to yield high revenue. The probability of selecting a particular price is influenced by the total revenue it generates, but is also smoothed to prevent any single buyer's data from having too much impact on the computation. This smoothing out is controlled by ε. When $\varepsilon \to 0$, all prices become equally likely independently of the buyers' valuations D and the revenue $u(D,p)$, leading to perfect privacy. As ε increases, the mechanism introduces less smoothing out and gives more importance to the revenue $u(D,p)$, providing more utility—by putting more mass on higher revenues—but less privacy. This mechanism thus allows the seller to achieve a balance between maximizing revenue and preserving the privacy of the buyers. The exponential mechanism provides Differential Privacy.

Theorem 1.15. *The exponential mechanism, $\mathcal{M}_{exp}$, achieves $(\varepsilon, 0)$-Differential Privacy.*

Proof. The proof assumes that $\mathcal{H}$ is a finite set. For any two neighbouring datasets $D \sim D'$ and some outcome $h \in \mathcal{H}$,

$$\begin{aligned}
\frac{Pr[\mathcal{M}_{\exp}(D) = h]}{Pr[\mathcal{M}_{\exp}(D') = h]} &= \frac{\left(\frac{\exp(\varepsilon s(D,h)/2\Delta s)}{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D,h')/2\Delta s\right)}\right)}{\left(\frac{\exp\left(\varepsilon s(D',h)/2\Delta s\right)}{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D',h')/2\Delta s\right)}\right)} \\
&= \exp\left(\frac{\varepsilon(s(D,h) - s(D',h))}{2\Delta s}\right) \frac{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D',h')/2\Delta s\right)}{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D,h')/2\Delta s\right)} \\
&\leq \exp\left(\frac{\varepsilon}{2}\right) \exp\left(\frac{\varepsilon}{2}\right) \frac{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D,h')/2\Delta s\right)}{\sum_{h' \in \mathcal{H}} \exp\left(\varepsilon s(D,h')/2\Delta s\right)} \\
&= \exp(\varepsilon).
\end{aligned}$$

The inequality follows due to the definition of Δs. □

Accuracy Guarantee

For the exponential mechanism, accuracy is not measured in terms of how close the mechanism is to the optimal hypothesis h. Rather, the objective is to guarantee that, with high probability, the output by the mechanism has a high score, as close as possible to optimality.

Theorem 1.16. *Let us fix a database D, and let $\mathcal{H}_{OPT} = \{h^* \in \mathcal{H} \text{ s.t. } s(D,h) = \max_h s(D,h)\}$ be the set of elements in $\mathcal{H}$ that achieve the maximum possible utility score. Then, the exponential mechanism guarantees*

$$Pr\left[s(D, \mathcal{M}_{exp}(D)) \geq OPT - \frac{2\Delta s}{\varepsilon}\left(\ln\left(|\mathcal{H}|/\beta\right)\right)\right] \geq 1 - \beta.$$

where $OPT = \max_h s(D,h)$.

Proof. Take any $c \in \mathbb{R}$. It follows that

$$\begin{aligned}
Pr\left[s(D, \mathcal{M}_{\text{exp}}(D) \leq c\right] &= \frac{\sum_{h:\ s(D,h)\leq c} \exp\left(\varepsilon s(D,h)/2\Delta s\right)}{\sum_{r\in\mathcal{H}} \exp(\varepsilon s(D,h)/2\Delta s)} \\
&\leq \frac{\sum_{r:\ s(D,h)\leq c} \exp\left(\varepsilon c/2\Delta s\right)}{\sum_{r\in\mathcal{H}_{\text{OPT}}} \exp(\varepsilon \text{OPT}/2\Delta s)} \\
&\leq \frac{|\mathcal{H}| \exp(\varepsilon c/2\Delta s)}{|\mathcal{H}_{\text{OPT}}| \exp(\varepsilon \text{OPT}/2\Delta s)} \\
&= \frac{|\mathcal{H}|}{|\mathcal{H}_{\text{OPT}}|} \exp\left(\frac{\varepsilon(c - \text{OPT})}{2\Delta s}\right) \\
&\leq |\mathcal{H}| \exp\left(\frac{\varepsilon(c - \text{OPT})}{2\Delta s}\right).
\end{aligned}$$

The result follows by plugging in

$$c \triangleq \text{OPT} - \frac{2\Delta s}{\varepsilon} \ln\left(|\mathcal{H}|/\delta\right).$$

□

Practically, this means that, although the mechanism introduces randomness to protect individual privacy (e.g., the buyers' valuations in our example), it still ensures that the selected output (the price) will yield a utility (the revenue) that is close to the best possible. E.g., in our example, the maximum possible revenue was \$3.01 at price \$3.01). Moreover, the utility loss due to privacy is limited and can be controlled by adjusting the privacy parameters.

1.7 Randomized Response, Revisited

Before concluding this chapter, it is useful to revisit the concept of randomized response. Consider Figure 1.4: its left side presents a pixelated version of the Mona Lisa, where each pixel is represented by either an 'M' or a '.' character. By implementing a random process that flips each pixel with a probability of 0.25,

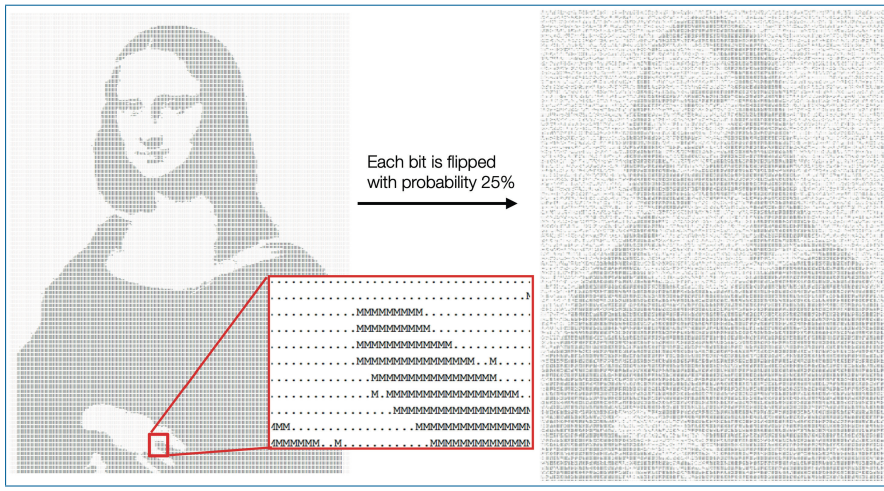

Figure 1.4. A metaphor for private data analysis: Perturbing each bit of the image on the left by flipping it with a random probability of 25% prevents inferring with high probability whether each single bit was originally an "M" or a ".", while still allowing to observe conclusions from the big picture. Figure adapted from slides presentation of Ulfar Erlingsson [Nam17].

the figure on the right emerges as locally perturbed yet retains the overall image, enabling recognition of the iconic Mona Lisa painting. This metaphor demonstrates that, although plausible deniability is afforded for the original value of each pixel, the outcomes of data analysis can still be preserved with considerable accuracy.

Revisiting Randomized Response

Figure 1.4 happens to be an instance of using randomized response to obscure individual responses while providing accurate summary statistics. Indeed, there is an equivalent formulation of randomized response that satisfies ε-DP in a stronger setting called *local Differential Privacy*, where instead of having a trusted curator that perturbs raw data to provide Differential Privacy, each data contributor perturbs its own data prior to its release. The topic of local DP is the subject of study of Chapter 2. Given $\varepsilon > 0$, for every private bit X, the mechanism is defined as follows:

$$\mathcal{M}(X) = \begin{cases} X, \text{ with probability } = \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}; \\ 1-X, \text{ with probability } = \frac{1}{1+\exp(\varepsilon)}. \end{cases}$$

Privacy Guarantees

Randomized response has the following Differential Privacy guarantees.

Theorem 1.17. *Randomized Response is* $(\varepsilon, 0)$*-differentially private.*

Proof. Let $p = \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$ for simplicity of exposition. The proof obligation is to upper bound the probability of ratios of probabilities for the two possible outcomes $\mathcal{M}(X) = X$ and $\mathcal{M}(X) = 1 - X$ for any $X \in \{0, 1\}$ and the neighbouring $X' = 1 - X$, i.e.,

$$\frac{Pr[\mathcal{M}(X) = X]}{Pr[\mathcal{M}(X') = X]} = \frac{Pr[\mathcal{M}(X) = X]}{Pr[\mathcal{M}(1 - X) = X]},$$

and

$$\frac{Pr[\mathcal{M}(X) = 1 - X]}{Pr[\mathcal{M}(X') = 1 - X]} = \frac{Pr[\mathcal{M}(X) = 1 - X]}{Pr[\mathcal{M}(1 - X) = 1 - X]}.$$

Note that the first quantity is equal to $\frac{p}{1-p} = \exp(\varepsilon)$, while the second quantity is equal to $\frac{1-p}{p} = \exp(-\varepsilon)$. This is enough to conclude the proof. □

Accuracy of Randomized Response

To provide the accuracy guarantee of Randomized Response, consider a collection of n data points $X_1, \ldots, X_n$. The goal is to compute the average of these data points, given by $\mu \triangleq \frac{1}{N}\sum_{i=1}^{n} X_i$. Consider the following simple linear estimator that corrects for the bias introduced by flipping X to the wrong answer, $1 - X$, with probability $p \triangleq \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$:

$$\hat{X} = \frac{1}{(2p-1)N}\left(\sum_{i=1}^{n} \mathcal{M}(X_i) + p - 1\right).$$

Lemma 1.18. *$\hat{X}$ is an unbiased estimator of $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$. Further, with probability at least $1 - \beta$,*

$$\left|\hat{X} - \mu\right| \leq \frac{\sqrt{1/\beta}}{2(2p-1)\sqrt{n}}.$$

Before providing the proof of this accuracy bound, consider what Differential Privacy promises. Remember that $p \triangleq \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$. Plugging this in the bound above,

$$\left|\hat{X} - \mu\right| = O\left(\frac{(1 + e^{\varepsilon})}{2\,(e^{\varepsilon} - 1)\,\sqrt{n}}\right).$$

As $\varepsilon \to 0$, the $1 + \exp(\varepsilon)$ term goes to 1; the $1 - \exp(\varepsilon)$ term can be approximated by ε given a first-order Taylor expansion. Hence, it follows that, as ε is small,

$$\left|\hat{X} - \mu\right| = O\left(\frac{1}{\varepsilon\sqrt{n}}\right).$$

In particular, given a small ε, to obtain an accuracy of α, requires that $n \sim \frac{1}{\varepsilon^2\alpha^2}$ samples.

Proof. Note that

$$\begin{aligned}\mathbb{E}[\mathcal{M}(X)] &= Pr[\mathcal{M}(X) = X] \cdot X + Pr[\mathcal{M}(X) = 1 - X] \cdot (1 - X) \\ &= pX + (1-p)(1-X) \\ &= (2p-1)X + (1-p).\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathcal{M}(X)] = (2p-1)\mu + (1-p),$$

immediately implying unbiasedness of $\hat{X}$. Now note that the variance of estimator $\hat{X}$ is given by

$$\text{Var}\left[\hat{X}\right] = \frac{1}{(2p-1)^2N^2}\sum_{i=1}^{N}\text{Var}\left[\mathcal{M}(X_i)\right] \leq \sum_{i=1}^{N}\frac{1}{4(2p-1)^2N^2} = \frac{1}{4(2p-1)^2N},$$

where the first equality follows from the fact that $\text{Var}[cX] = c^2\text{Var}[X]$ and $\text{Var}[X + c] = \text{Var}[X]$ for a constant c, and the inequality follows from the fact that $\mathcal{M}(X)$ is a Bernoulli random variable and has variance at most $1/4$. Using Chebyshev's inequality with $k = \frac{1}{\sqrt{\beta}}$, it follows that

$$Pr\left[\left|\hat{X} - \mu\right| \geq \frac{\sqrt{1/\beta}}{2(1-2p)\sqrt{n}}\right] \leq \beta.$$

□

The above bound is an example of privacy-accuracy trade-off. To obtain an accuracy level of α (i.e., the estimator does not mis-estimate μ by more than α) with high probability $1 - \beta$, one needs to pick the value of p such that

$$\frac{\sqrt{1/\beta}}{2(1-2p)\sqrt{n}} \leq \alpha.$$

This immediately gives the desired value of ε, given us a trade-off between the accuracy level α and the privacy level ε. Here, decreasing ε towards 0 (or equivalently decreasing p towards $1/2$) yields a worse accuracy guarantee, as the denominator decreases and eventually goes to 0. This goes in the expected direction: the more privacy is required, the more the accuracy suffers.

1.8 Concluding Remarks

This chapter discussed foundational concepts and mechanisms that are the bedrock of Differential Privacy. Since its conceptual introduction, Differential Privacy has seen considerable evolution, both in theoretical development and practical applications. Researchers have refined the mathematical guarantees, offering tighter bounds on privacy leakage and more effective mechanisms for trading utility with privacy. Practically, Differential Privacy has been applied across diverse sectors, from healthcare to social science, to engineering systems, as reviewed in Part III. These applications demonstrate the flexibility and robustness of Differential Privacy in safeguarding personal information while maintaining data utility. The implications of adopting Differential Privacy extends beyond the technical realm, influencing regulatory policies around data privacy [Exe23], as also discussed in Part V of this book. As organizations increasingly rely on data-driven decision-making, the implementation of DP can help build trust with stakeholders by demonstrating a commitment to privacy-preserving practices. This trust is crucial for compliance with international data protection regulations and for fostering a more privacy-conscious data ecosystem. Furthermore, the principles of Differential Privacy can guide ethical considerations in data usage, promoting a balance between innovation and individual rights to privacy.

Acknowledgements

This work was partially supported by NSF grants SaTC-2345483, CAREER RI-2401285, CAREER HCC-2336236, and by a Google Scholar Research Award. Its view and conclusions are those of the authors only.

References

[BW18] B. Balle and Y.-X. Wang. “Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising”. In: International Conference on Machine Learning. 2018. URL: https://api.semanticscholar.org/CorpusID:21713075 (cit. on pp. 29, 30).

[BZH06] M. Barbaro, T. Zeller, and S. Hansell. “A face is exposed for AOL searcher no. 4417749”. In: New York Times 9.2008 (2006), p. 8 (cit. on p. 7).

[Car+21] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. "Extracting training data from large language models". In: 30th USENIX Security Symposium (USENIX Security 21). 2021, pp. 2633–2650 (cit. on p. 11).

[CDMS21] A. Cohen, M. Duchin, J. Matthews, and B. Suwal. "Census Top-Down: The Impacts of Differential Privacy on Redistricting". In: Proceedings of the 2nd Symposium on Foundations of Responsible Computing. Ed. by K. Ligett and S. Gupta. FORC '21. 2021, 5:1–22 (cit. on p. 28).

[Cen96] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/. 1996 (cit. on p. 3).

[CNSU20a] A. Cohen, A. Nikolov, Z. Schutzman, and J. Ullman. Reconstruction Attacks in Practice. DifferentialPrivacy.org. https://differentialprivacy.org/diffix-attack/. Oct. 2020 (cit. on p. 11).

[CNSU20b] A. Cohen, A. Nikolov, Z. Schutzman, and J. Ullman. The Theory of Reconstruction Attacks. DifferentialPrivacy.org. https://differentialprivacy.org/reconstruction-theory/. Oct. 2020 (cit. on p. 11).

[Des17] D. Desfontaines. k-anonymity, the parent of all privacy definitions. https://desfontain.es/blog/k-anonymity.html. Ted is writing things (personal blog). Aug. 2017 (cit. on p. 8).

[DHVB13] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. "Unique in the crowd: The privacy bounds of human mobility". In: Scientific reports 3.1 (2013), pp. 1–5 (cit. on p. 7).

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating noise to sensitivity in private data analysis". In: Theory of cryptography conference. Springer. 2006, pp. 265–284 (cit. on pp. 11, 16, 20, 24).

[DN03] I. Dinur and K. Nissim. "Revealing information while preserving privacy". In: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2003, pp. 202–210 (cit. on pp. 10, 11).

[Dob21] L. Dobberstein. Korean app-maker Scatter Lab fined for using private data to create homophobic and lewd chatbot. 2021. URL: https://www.theregister.com/2021/04/29/scatter_lab_fined_for_lewd_chatbot/ (cit. on p. 11).

[DR14] C. Dwork and A. Roth. "The Algorithmic Foundations of Differential Privacy". In: Found. Trends Theor. Comput. Sci. 9 (2014), pp. 211–407. URL: https://api.semanticscholar.org/CorpusID:207178262 (cit. on pp. 22, 23, 29, 30).

[Dwo+06] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. "Our Data, Ourselves: Privacy Via Distributed Noise Generation". In: Advances in Cryptology - EUROCRYPT. Vol. 4004. Lecture Notes in Computer Science. Springer, 2006, pp. 486–503. URL: https://doi.org/10.1007/11761679%5C_29 (cit. on p. 3).

[Exe23] Executive Office of the President. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register. Available online: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Nov. 2023 (cit. on p. 37).

[GRS12] A. Ghosh, T. Roughgarden, and M. Sundararajan. "Universally Utility-Maximizing Privacy Mechanisms". In: SIAM Journal on Computing 41.6 (2012), pp. 1673–1693 (cit. on p. 28).

[Hom+08] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". In: PLoS genetics 4.8 (2008), e1000167 (cit. on p. 7).

[Hou23] T. W. House. Blueprint for an AI Bill of Rights. Nov. 2023. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (cit. on p. 3).

[Kam20] G. Kamath. Approximate Differential Privacy. CS 860: Algorithms for Private Data Analysis - Fall 2020 Lecture Notes. Available online at: http://www.gautamkamath.com/courses/CS860-fa2022-files/lec5.pdf. 2020 (cit. on p. 29).

[KS12] V. Karwa and A. B. Slavković. "Differentially Private Graphical Degree Sequences and Synthetic Graphs". In: Privacy in Statistical Databases. Ed. by J. Domingo-Ferrer and I. Tinnirello. Vol. 7556. Lecture Notes in Computer Science. Springer, 2012, pp. 273–285 (cit. on p. 28).

[Leg18] C. S. Legislature. California Consumer Privacy Act (CCPA) — oag.ca.gov. Online at https://oag.ca.gov/privacy/ccpa. 2018 (cit. on p. 3).

[LSTS20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. "Federated learning: Challenges, methods, and future directions". In: IEEE signal processing magazine 37.3 (2020), pp. 50–60 (cit. on p. 11).

[MT07] F. McSherry and K. Talwar. "Mechanism Design via Differential Privacy". In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). 2007, pp. 94–103 (cit. on p. 31).

[Nam17] A. Name. RAPPOR Talk for DIMACS Workshop, April 2017. Slide presentation at the DIMACS Workshop on Big Data Integration. Available online at: http://archive.dimacs.rutgers.edu/Workshops/BigDataHub/Slides/RAPPOR-talk-for-DIMACS-workshop-April-2017.pdf. DIMACS, Apr. 2017. URL: http://archive.dimacs.rutgers.edu/Workshops/BigDataHub/Slides/RAPPOR-talk-for-DIMACS-workshop-April-2017.pdf (cit. on p. 34).

[NS06] A. Narayanan and V. Shmatikov. "How To Break Anonymity of the Netflix Prize Dataset". In: ArXiv abs/cs/0610105 (2006). URL: https://api.semanticscholar.org/CorpusID:1086763 (cit. on p. 7).

[NS08] A. Narayanan and V. Shmatikov. "Robust de-anonymization of large sparse datasets". In: 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE. 2008, pp. 111–125 (cit. on p. 9).

[PE16] E. Parliament and C. of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). May 2016 (cit. on p. 3).

[SS98] P. Samarati and L. Sweeney. "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression". In: Proceedings of the IEEE Symposium on Research in Security and Privacy. 1998 (cit. on p. 7).

[Swe00] L. Sweeney. "Simple Demographics Often Identify People Uniquely". Working paper. 2000. URL: http://dataprivacylab.org/projects/identifiability/ (cit. on pp. 6, 8).

[Swe02] L. Sweeney. "k-anonymity: a model for protecting privacy". In: Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. URL: https://doi.org/10.1142/S0218488502001648 (cit. on pp. 7, 8).

[Uni98] United States Census Bureau. Title 13 of the United States Code: Census. https://www.census.gov/about/history/bureau-history/agency-history-timeline/title-13.html. Accessed: 2024-04-27. 1998 (cit. on p. 3).

[War65] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: Journal of the American Statistical Association 60.309 (1965), pp. 63–69. ISSN: 01621459. URL: http://www.jstor.org/stable/2283137 (cit. on p. 17).

[Wik24a] Wikipedia. 2017 Equifax data breach — Wikipedia, The Free Encyclopedia. Online at http://en.wikipedia.org/w/index.php?title=2017-Equifax-data-breach&oldid=1241882235. 2024 (cit. on p. 2).

[Wik24b] Wikipedia. Facebook–Cambridge Analytica data scandal — Wikipedia, The Free Encyclopedia. Online at https://en.wikipedia.org/wiki/FacebookâĽ"Cambridge_Analytica_data_scandal#Governmental_actions. 2024 (cit. on p. 2).

[ZLH19] L. Zhu, Z. Liu, and S. Han. "Deep leakage from gradients". In: Advances in neural information processing systems 32 (2019) (cit. on p. 11).