

# Privacy-Preserving Natural Language Processing

RUHR  
UNIVERSITÄT  
BOCHUM

RUB

## Lecture 3 – Introduction to Differential Privacy

---

Prof. Dr. Ivan Habernal

April 24, 2025

[www.trusthlt.org](http://www.trusthlt.org)

Chair of Trustworthy Human Language Technologies (TrustHLT)

Ruhr University Bochum & Research Center Trustworthy Data Science and Security



CENTER FOR TRUSTWORTHY  
DATA SCIENCE AND SECURITY

# Intro

---

- 1 Intro
- 2 Differential privacy basics
- 3 Dataset
- 4 Statistical queries

# Overview

First lecture: privacy definitions, arbitrary hard choices

Second lectures: still arbitrary choice in anonymization, but  
GDPR notion of possibility of deidentification

Today: towards probabilistic approach to privacy

# Differential privacy basics

---

- 1 Intro
- 2 Differential privacy basics**
- 3 Dataset
- 4 Statistical queries

# Precursors of differential privacy

"Let us begin with a short story. Envision a database of a hospital containing the medical history of some population. On one hand, the hospital would like to advance medical research which is based (among other things) on statistics of the information in the database. On the other hand, the hospital is obliged to keep the privacy of its patients, i.e. leak no medical information that could be related to a specific patient. The hospital needs an access mechanism to the database that allows certain 'statistical' queries to be answered, as long as they do not violate the privacy of any single patient."

I. Dinur and K. Nissim (2003). **"Revealing Information While Preserving Privacy"**. In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. SIGMOD/PODS03: International Conference on Management of Data and Symposium on Principles Database and Systems. San Diego California: ACM, pp. 202–210

# Precursors of differential privacy

"We focus on binary databases, where the content is of  $n$  binary (0-1) entries"

I. Dinur and K. Nissim (2003). **"Revealing Information While Preserving Privacy"**. In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. SIGMOD/PODS03: International Conference on Management of Data and Symposium on Principles Database and Systems. San Diego California: ACM, pp. 202–210

# First DP papers

"The database consists of some number  $n$  of rows. We are completely agnostic about the type of rows – they may be tuples of attributes, strings, or even pictures."

A. Blum, C. Dwork, F. McSherry, and K. Nissim (2005). **"Practical Privacy: The SuLQ Framework"**. In: *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. SIGMOD/PODS05: International Conference on Management of Data and Symposium on Principles Database and Systems. Baltimore Maryland: ACM, pp. 128–138

# Explicit notion of independence of rows

## Dependency Between Database Records

"We explicitly assume that the database records are chosen independently from each other, according to some underlying distribution  $D$ . We are not aware of any work that does not make this assumption (implicitly or explicitly)."

C. Dwork and K. Nissim (2004).

**"Privacy-Preserving Datamining on Vertically Partitioned Databases"**. In:

*Proceedings of the 24th Annual International Cryptology Conference - CRYPTO 2004*. Ed. by M. Franklin. Vol. 3152. Santa Barbara, CA, USA: Springer Berlin Heidelberg, pp. 528–544



# Explicit notion of independence of rows

"The intent of the independence assumption is to characterize what information is under the control of a given individual. Specifically, **if there is information about a row that can be learned from other rows, this information is not truly under the control of that row.** Even if the row in question were to sequester itself away in a high mountaintop cave, information about the row that can be gained from the analysis of other rows is still available to an adversary. It is for this reason that **we focus our attention on those inferences that can be made about rows without the help of others.**"

A. Blum, C. Dwork, F. McSherry, and K. Nissim (2005). **"Practical Privacy: The SuLQ Framework"**. In: *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. SIGMOD/PODS05: International Conference on Management of Data and Symposium on Principles Database and Systems. Baltimore Maryland: ACM, pp. 128–138

# Dataset

---

- 1 Intro
- 2 Differential privacy basics
- 3 Dataset**
- 4 Statistical queries

# Dataset (or Database)

## The essential assumption (in words)

Each person corresponds to a single row in a table

The semantics of columns is arbitrary:

- It could be an actual "table" with some sensitive attributes (e.g., age, income)
- It could be just an arbitrary unstructured object (e.g., personal photo, text)

# Dataset

## Another implicit assumption

Each row contains data (values or attributes) that "belong" to this row's person

## Implication of the above assumption

Example: If person A is removed from the dataset, you cannot certainly tell her private attributes from some other row B, C, D.

# Examples of implicit assumptions

| Name    | Income | Age |
|---------|--------|-----|
| Bob     | 700    | 21  |
| Alice   | 2,800  | 32  |
| Charlie | 3,500  | 45  |

**Table 1:** Looks legit, no clear violation of the assumptions

| Name    | Income | Age | Someone else's income |
|---------|--------|-----|-----------------------|
| Bob     | 700    | 21  | Alice: 2,800          |
| Alice   | 2,800  | 32  | Charlie: 3,500        |
| Charlie | 3,500  | 45  | Alice: 2,800          |

**Table 2:** Something looks wrong with this dataset

# Examples of implicit assumptions

| Name    | Income | Age |
|---------|--------|-----|
| Bob     | 700    | 21  |
| Alice   | 2,800  | 32  |
| Charlie | 3,500  | 45  |
| Alice   | 2,800  | 32  |

**Table 3:** Something looks wrong with this dataset

# Statistical queries

---

- 1 Intro
- 2 Differential privacy basics
- 3 Dataset
- 4 Statistical queries**

# Real-valued query

"Trusted server that holds a database of sensitive information. Given a query function  $f$  mapping databases to reals, the so-called true answer is the result of applying  $f$  to the database. [...]"

## Query

Let  $X$  be a database from a universe (a set) of all possible databases  $\mathcal{X}$

Real-valued query is

$$f : \mathcal{X} \mapsto \mathbb{R}$$

C. Dwork, F. McSherry, K. Nissim, and A. Smith (2006). **"Calibrating Noise to Sensitivity in Private Data Analysis"**. In: *Theory of Cryptography*. Ed. by S. Halevi and T. Rabin. Red. by D. Hutchison et al. Vol. 3876. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 265–284



## Query example – "counting" query

"A query consists of a pair  $(S, f)$  where  $S$  is a set of rows in the database and  $f$  is a function mapping database rows to  $\{0, 1\}$ . The true answer is  $\sum_{i \in S} f(d_i)$ "

- implicitly assuming  $d_i$  is the  $i$ -th entry from  $S$
- also note inconsistencies in notations, inevitable in every new field being just formalized!! Here  $f$  maps a "row" to zero or one, previous slide  $f$  would be the whole sum

C. Dwork, F. McSherry, K. Nissim, and A. Smith (2006). **"Calibrating Noise to Sensitivity in Private Data Analysis"**. In: *Theory of Cryptography*. Ed. by S. Halevi and T. Rabin. Red. by D. Hutchison et al. Vol. 3876. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 265–284

# How could a counting query leak private information?

"It's just a statistics (a sum), so I can hide in the crowd!"

## What is private information

The secret "bit"  $f(d_i) \mapsto \{0, 1\}$

which eventually corresponds to whether or not you are in the database

- If your are in the database, you cannot control others
- What if the database is very small?
- What if the database is queried repeatedly?
- What if you are an outlier?

# Statistical queries

---

How to protect privacy in counting queries

# Example

| Name    | Hospitalized in year | Age | Illegal drug use |
|---------|----------------------|-----|------------------|
| Alice   | 2024                 | 32  | yes              |
| Bob     | 2020                 | 21  | no               |
| Charlie | 2023                 | 45  | no               |
| ...     |                      |     |                  |
| Xander  | 2020                 | 31  | yes              |

(Important to clarify for context: If you are Alice, how did you end up in such a database?)

**Table 4:** A sensitive database example from a clinic

Query: How many persons in the database take illegal drugs?

## Example: How many persons in the database take illegal drugs?

| Name    | Hospitalized in year | Age | Illegal drug use |
|---------|----------------------|-----|------------------|
| Alice   | 2024                 | 32  | yes              |
| Bob     | 2020                 | 21  | no               |
| Charlie | 2023                 | 45  | no               |
| ...     |                      |     |                  |
| Xander  | 2020                 | 31  | yes              |

What might be public information: the size of the dataset  
(say  $n = 100$ )

The true answer to the query: 2

If Xander reveals his drug use, and we know Alice is in the database, her privacy is revealed!

# Statistical queries

---

How to protect privacy

# Privatizing the query result

How would you go about it?

# Solution: Alter the query result

Changes to the true query result must be **irreversible!**

"A natural approach, and one that has been explored by others in the 1980's, is to add random noise to the answer" (Blum, Dwork, McSherry, and Nissim, 2005)

"It is evident that without randomness there is no privacy: if everything is pre-determined, and all possible choices we make are predictable or pre-programmed by our adversaries, then there is nothing that we can build our privacy on." (Ekert and Renner, 2014)

A. Blum, C. Dwork, F. McSherry, and K. Nissim (2005). **"Practical Privacy: The SuLQ Framework"**. In: *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. SIGMOD/PODS05: International Conference on Management of Data and Symposium on Principles Database and Systems. Baltimore Maryland: ACM, pp. 128–138

A. Ekert and R. Renner (Mar. 2014). **"The Ultimate Physical Limits of Privacy"**. In: *Nature* 507.7493, pp. 443–447



# How to randomize the query result?

Recall: The counting query output is a natural number

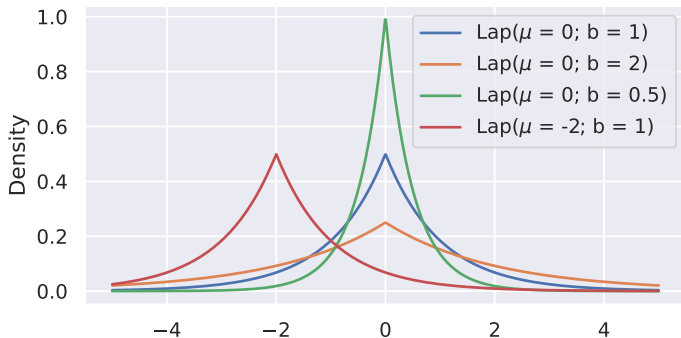
How should we randomize it?

- Draw a random value from a probability distribution
- Which one? How parametrized?

# What we want

The randomized output will be likely close to the true value,  
and unlikely far away

# The Laplace distribution



**Figure 1:** Laplace PDF (probability density function)

$$\text{Lap}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

# How to choose the parameters?

Location and scale

- Location: the true value
- Scale?

Let's pick  $b = 1$  just for lack of better ideas :)

# It's all nice and random, but what does it give us?

Query: How many persons in the database take illegal drugs?

| Name    | Hospitalized in year | Age | Illegal drug use |
|---------|----------------------|-----|------------------|
| Alice   | 2024                 | 32  | yes              |
| Bob     | 2020                 | 21  | no               |
| Charlie | 2023                 | 45  | no               |
| ...     |                      |     |                  |
| Xander  | 2020                 | 31  | yes              |

**Table 5:** Database  $D$ , including Alice

Assume that the true answer is 51

# It's all nice and random, but what does it give us?

What if Alice decided **not to be part of the data**?

– She has the right to decide about her privacy, GDPR, etc.

| Name    | Hospitalized in year | Age | Illegal drug use |
|---------|----------------------|-----|------------------|
| Bob     | 2020                 | 21  | no               |
| Charlie | 2023                 | 45  | no               |
| ...     |                      |     |                  |
| Xander  | 2020                 | 31  | yes              |

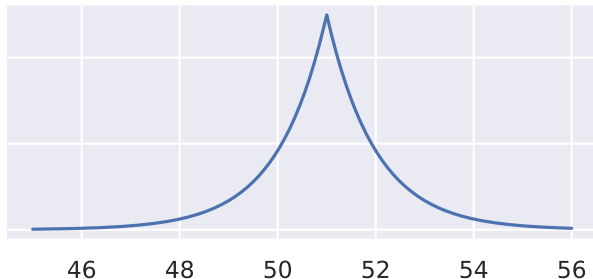
**Table 6:** Database  $D'$ , **excluding** Alice

The true answer to the query would be now 50

## We have two databases: With and without Alice

Privatize for  $D$  (with Alice), true answer is 51

$$Y \sim \frac{1}{2b} \exp\left(\frac{-|x - 51|}{b}\right) = \frac{1}{2} \exp(-|x - 51|)$$

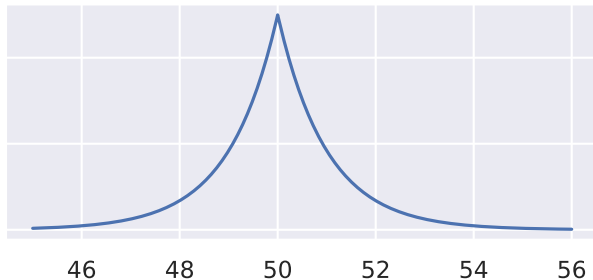


**Figure 2:** Laplace density,  $\mu = 51$ ,  $b = 1$

## Now without Alice

Privatize for  $D'$  (without Alice), true answer is 50

$$Y' \sim \frac{1}{2b} \exp\left(\frac{-|x - 50|}{b}\right) = \frac{1}{2} \exp(-|x - 50|)$$

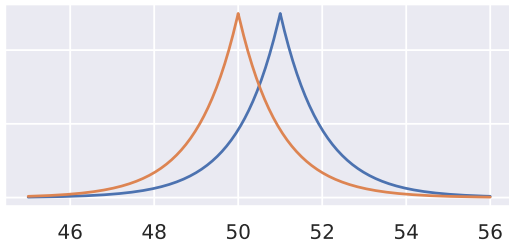


**Figure 3:** Laplace density,  $\mu = 50$ ,  $b = 1$



## Both variables in one plot

$$Y \sim \frac{1}{2} \exp(-|x - 51|) \quad Y' \sim \frac{1}{2} \exp(-|x - 50|)$$



**Figure 4:** Blue = with Alice, red = without Alice

Now observe a particular value  $y \in \mathbb{R}$ , for example 52

– Was it sampled from  $Y$  or from  $Y'$ ?

## If we observed 52, was it sampled from $Y$ or from $Y'$ ?

We cannot really tell!

But we can compare probabilities of this event<sup>1</sup>

$$\Pr[Y = 52] = \frac{1}{2} \exp(-|52 - 51|) = 0.5 \exp(-1)$$

$$\Pr[Y' = 52] = \frac{1}{2} \exp(-|52 - 50|) = 0.5 \exp(-2)$$

How much more likely from  $Y$ ?

$$\frac{\Pr[Y=52]}{\Pr[Y'=52]} = \frac{0.5 \exp(1)}{0.5 \exp(2)} = \exp(-1) = 0.367$$

How much more likely from  $Y'$ ?

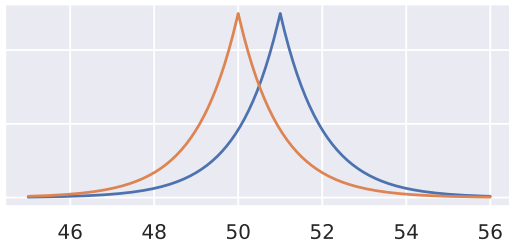
$$\frac{\Pr[Y'=52]}{\Pr[Y=52]} = \frac{0.5 \exp(2)}{0.5 \exp(1)} = \exp(1) = 2.718$$

---

<sup>1</sup>We must compare densities, so our math will not break

## Both variables in one plot

$$Y \sim \frac{1}{2} \exp(-|x - 51|) \quad Y' \sim \frac{1}{2} \exp(-|x - 50|)$$



**Figure 5:** Blue = with Alice, red = without Alice

Now observe a particular value  $y \in \mathbb{R}$ , for example 50

– Was it sampled from  $Y$  or from  $Y'$ ?

**If we observed 50, was it sampled from  $Y$  or from  $Y'$ ?**

$$\Pr[Y = 50] = \frac{1}{2} \exp(-|50 - 51|) = 0.5 \exp(-1)$$

$$\Pr[Y' = 50] = \frac{1}{2} \exp(-|50 - 50|) = 0.5 \exp(0)$$

How much more likely from  $Y$ ?

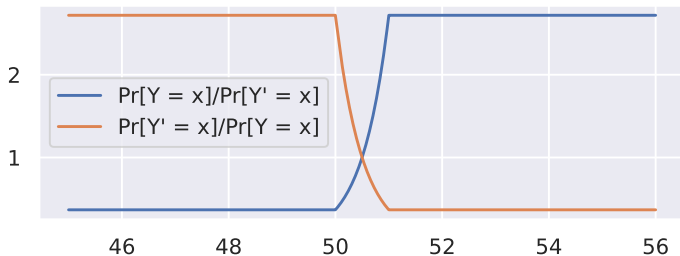
$$\frac{\Pr[Y=50]}{\Pr[Y'=50]} = \frac{0.5 \exp(-1)}{0.5 \exp(0)} = \exp(-1) = 0.367$$

How much more likely from  $Y'$ ?

$$\frac{\Pr[Y'=50]}{\Pr[Y=50]} = \frac{0.5 \exp(0)}{0.5 \exp(-1)} = \exp(1) = 2.718$$

It seems like the odds ratio is bounded from above...

## Can we generalize it for any observed $x$ ?



Seems like the maximum we can get is  $2.718 = e = \exp(1)$

# Let's try to prove it

We need triangle inequality for absolute values:<sup>2</sup>

$$|x| - |y| \leq |x - y|$$

$$\begin{aligned} \frac{\Pr[Y = x]}{\Pr[Y' = x]} &= \frac{\frac{1}{2} \exp(-|x - 51|)}{\frac{1}{2} \exp(-|x - 50|)} = \exp(|x - 50| - |x - 51|) \\ &\leq \exp(|(x - 50) - (x - 51)|) \\ &= \exp(|x - 50 - x + 51|) \\ &= \exp(|1|) = \exp(1) \end{aligned}$$

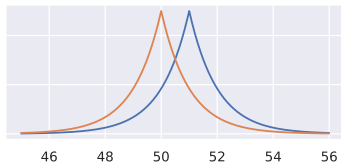
---

<sup>2</sup>Homework: Prove it!  $x, y \in \mathbb{R}$

## Let's try to prove it (part 2)

$$\begin{aligned}\frac{\Pr[Y' = x]}{\Pr[Y = x]} &= \frac{\frac{1}{2} \exp(-|x - 50|)}{\frac{1}{2} \exp(-|x - 51|)} = \exp(|x - 51| - |x - 50|) \\ &\leq \exp(|(x - 51) - (x - 50)|) \\ &= \exp(|x - 51 - x + 50|) \\ &= \exp(|-1|) = \exp(1)\end{aligned}$$

# What does that mean?



**Figure 6:** Blue = with Alice, red = without Alice,  
 $Y \sim \frac{1}{2} \exp(-|x - 51|)$        $Y' \sim \frac{1}{2} \exp(-|x - 50|)$

$$\frac{\Pr[Y=x]}{\Pr[Y'=x]} \leq \exp(1) \quad \frac{\Pr[Y'=x]}{\Pr[Y=x]} \leq \exp(1)$$

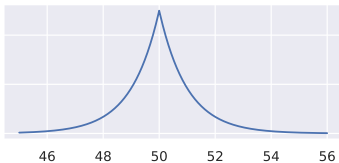
No matter what value we get after privatizing the counting query – we can only get some limited "information" about whether it came from  $Y$  or  $Y'$ .



## What if the true answer would be the same for $D$ and $D'$ ?

Alice did not take drugs

$$Y \sim \frac{1}{2} \exp(-|x - 50|) \quad Y' \sim \frac{1}{2} \exp(-|x - 50|)$$



**Figure 7:** Both have the same distribution

$$\frac{\Pr[Y=x]}{\Pr[Y'=x]} = 1 \leq \exp(1) \quad \frac{\Pr[Y'=x]}{\Pr[Y=x]} = 1 \leq \exp(1)$$

So the maximum "loss" only when the presence of Alice changes the query output

# What if the true answer would be 22?

Alice did not take drugs

$$Y \sim \frac{1}{2} \exp(-|x - 22|) \quad Y' \sim \frac{1}{2} \exp(-|x - 22|)$$

– bounded by 1 as in the previous slide

Alice did take drugs

$$Y \sim \frac{1}{2} \exp(-|x - 23|) \quad Y' \sim \frac{1}{2} \exp(-|x - 22|)$$

– our general proof still holds!

$$\frac{\Pr[Y=x]}{\Pr[Y'=x]} \leq \exp(1) \quad \frac{\Pr[Y'=x]}{\Pr[Y=x]} \leq \exp(1)$$

# Summary

Four counting queries, the maximum difference of the query result is 1 (when a particular person is not in the dataset)

We used scale  $b = 1$  for the Laplace distribution, which gives us upper bound on privacy loss

$$\frac{\Pr[Y=x]}{\Pr[Y'=x]} \leq \exp(1) \quad \frac{\Pr[Y'=x]}{\Pr[Y=x]} \leq \exp(1)$$

# License and credits

Licensed under Creative Commons  
Attribution-ShareAlike 4.0 International  
(CC BY-SA 4.0)



## Credits

Ivan Habernal

Content from ACL Anthology papers licensed under CC-BY

<https://www.aclweb.org/anthology>

Partly inspired by lectures from Antti Honkela, Aurélien Bellet, Gautam Kamath