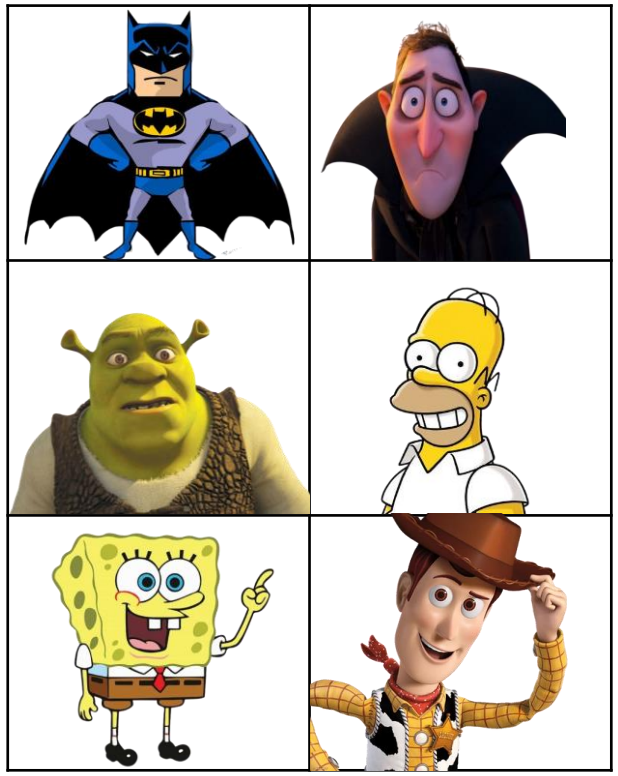
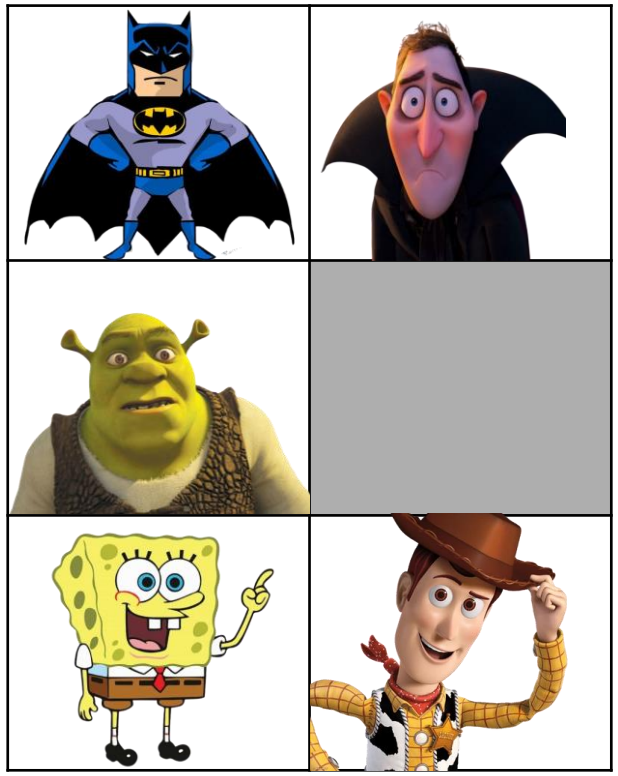
DIFFERENTIAL PRIVACY

Data Processing



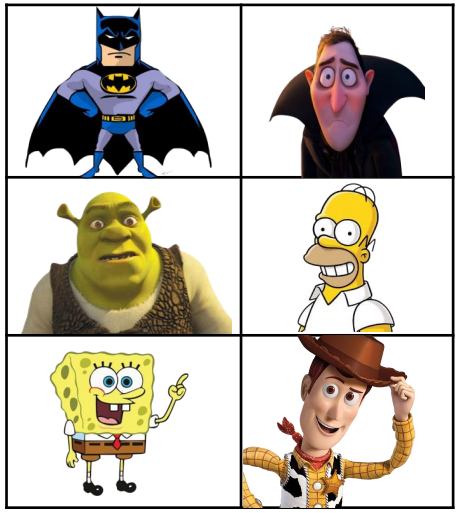
Process

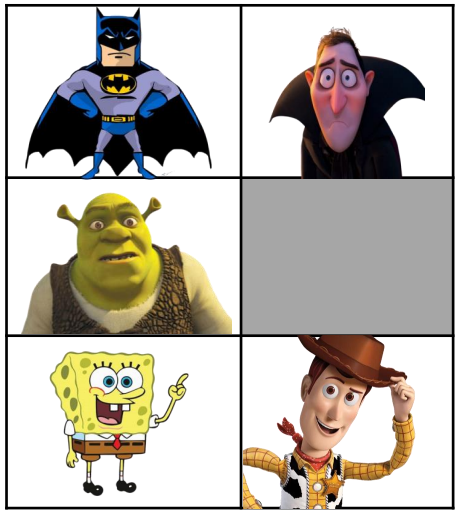
Data Processing



Process

Differential Privacy





Random Result

Probability of having the same result is close

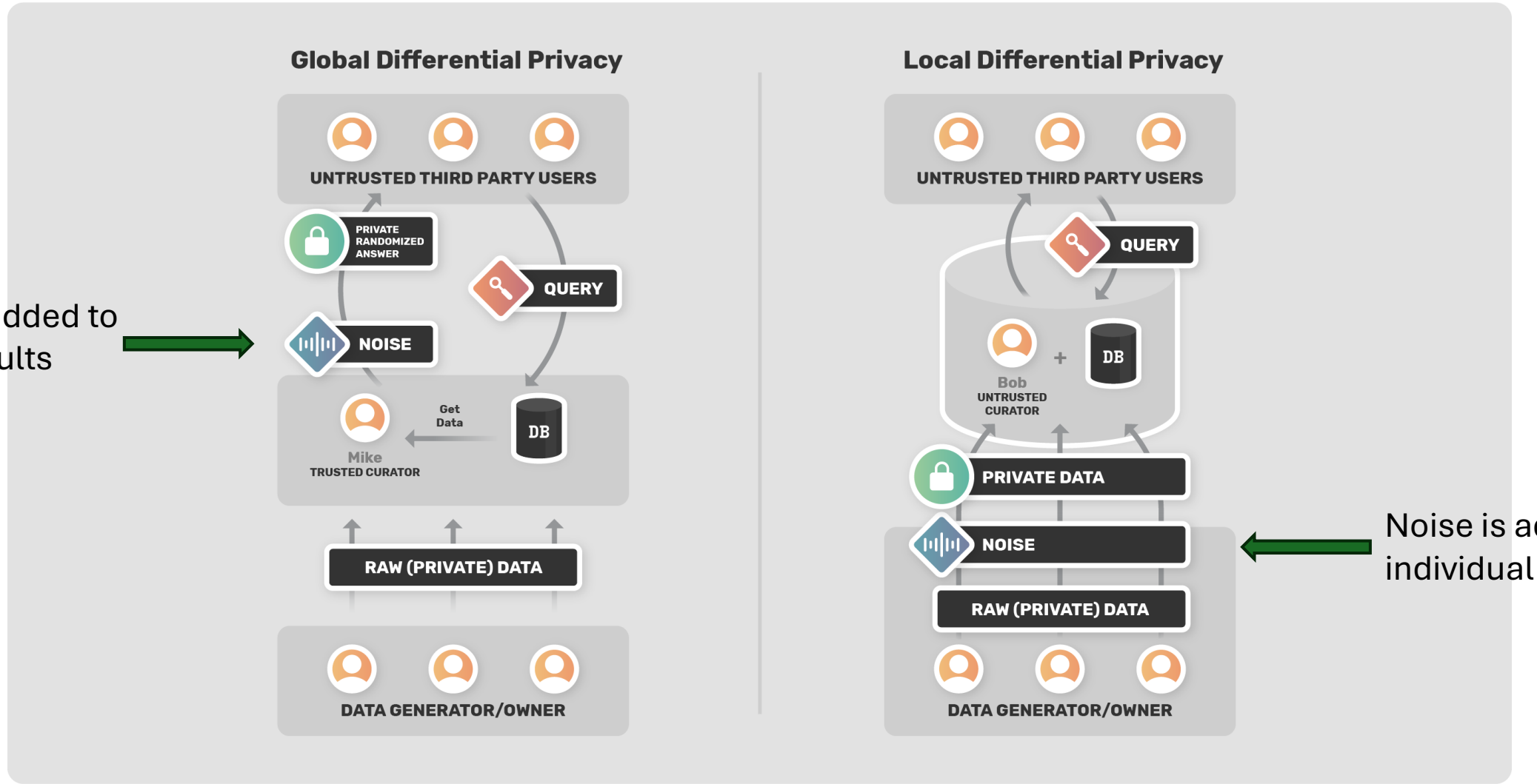
Random Result

Result = [.....A, B, C, D, E.....]

Desfontaines, D. (2021). A friendly, non-technical introduction to differential privacy. https://desfontain.es/blog/friendly-intro-to-differential-privacy.html

Differential Privacy

Noise is added to query results



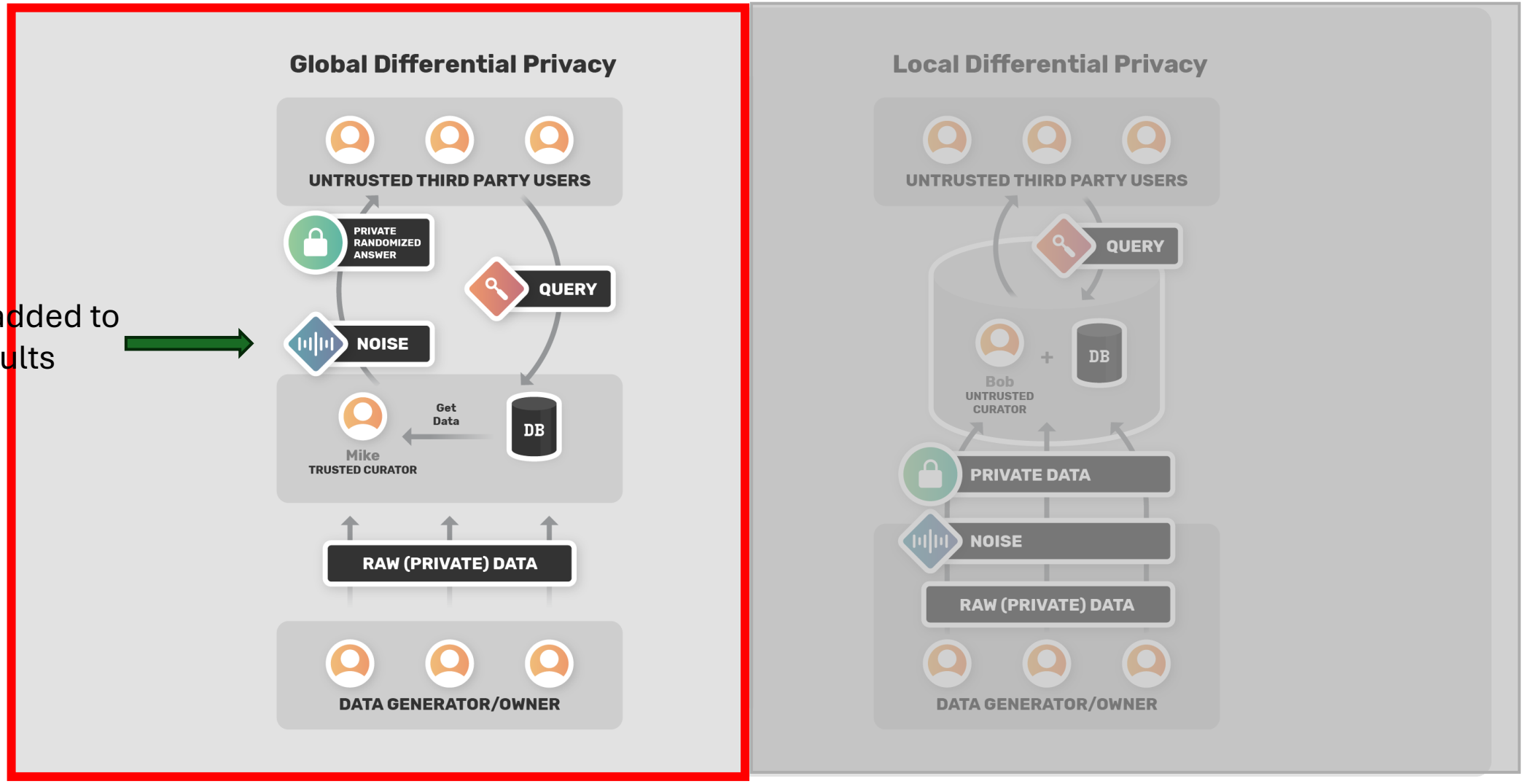
Noise is added to individual records

LOCAL VS GLOBAL DIFFERENTIAL PRIVACY (OpenMined 2021)

Differential Privacy

We Will Learn

Noise is added to query results



LOCAL VS GLOBAL DIFFERENTIAL PRIVACY (OpenMined 2021)

Indistinguishable yet Useful

When Adding Noise

What is the **color** of the T-shirt?

Original

Noisy#1

Noisy#2

Noisy#3

Noisy#4

Scenarios
Can we add noise?

- The government wants to find out the number of COVID patients in each Zip code

  I'm suspicious that Ann is in the dataset

Query "How many COVID patients are there in Zip = 1010"

Real Answer	20

Noisy Answer	20.1

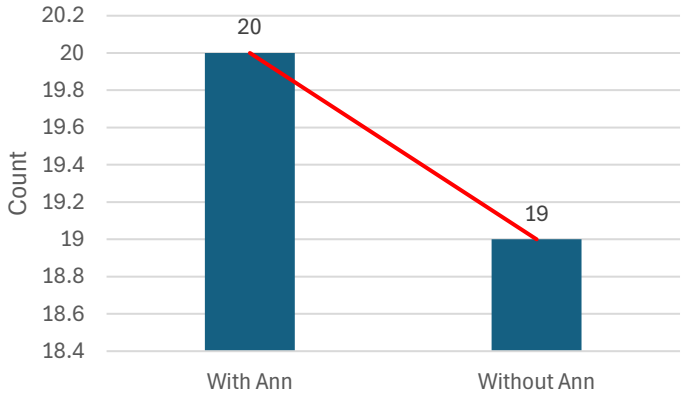
Error = 0.1

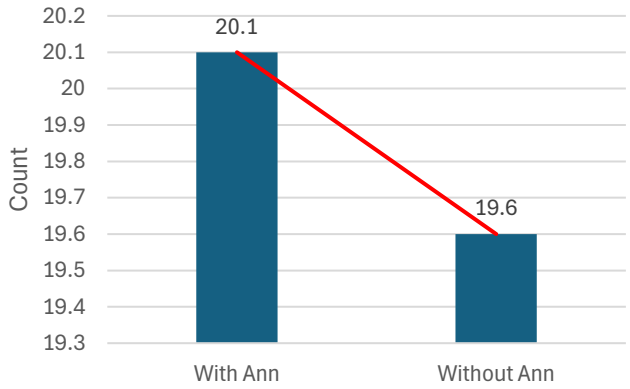
Query "How many COVID patients are there in Zip = 1010, where name ≠ Ann"

Real Answer	19

Noisy Answer	19.6

Error = 0.6





Scenarios

Can we add noise?

- The government wants to find out the number of COVID patients in each Zip code
- Apple wants to find the emojis that are used frequently in iMessages
- Uber wants to know peak booking times in Auckland region
- The Warehouse wants to analyse which products has higher online orders during the Christmas season
- The university health board wants to analyse abortion rates among students over the last five years

Differential Privacy

Definition

A mathematical guarantee which assures the indistinguishability of an individual within neighboring datasets

Dwork, C. (2006). Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. https://doi-org.ezproxy.auckland.ac.nz/10.1007/11787006_1

Differential Privacy

Definition

A mathematical guarantee which assures the indistinguishability of an individual within neighboring datasets

In simple terms……

DP makes it hard to deduce whether an individual is included in a dataset or not

Differential Privacy

"Differential Privacy is not a property of **data**, but a property of the **process**"

- By looking at the K-anonymous dataset can you say someone have applied K anonymity?
- You can't look at the output data and determine whether it satisfies differential privacy
- You have to know how the data was generated to determine whether it's differentially private

Differential Privacy

Neighbouring Datasets

A mathematical guarantee which assures the indistinguishability of an individual within neighboring datasets

If two datasets are different from **only one record**, they are neighbours

- Remove a record
- Add a record
- Edit a value/s in an existing record
- Queries trying to single out a person

  Average age of people in dataset X

  Average age of people in dataset X where name ≠ “Kate”

Differential Privacy

Neighbouring Datasets

If two datasets are different from only one record, they are neighbours

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	35	1

Name	Age	PTSD
Ann	21	1
David	16	0

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	64	1

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	35	1
Gibson	22	1

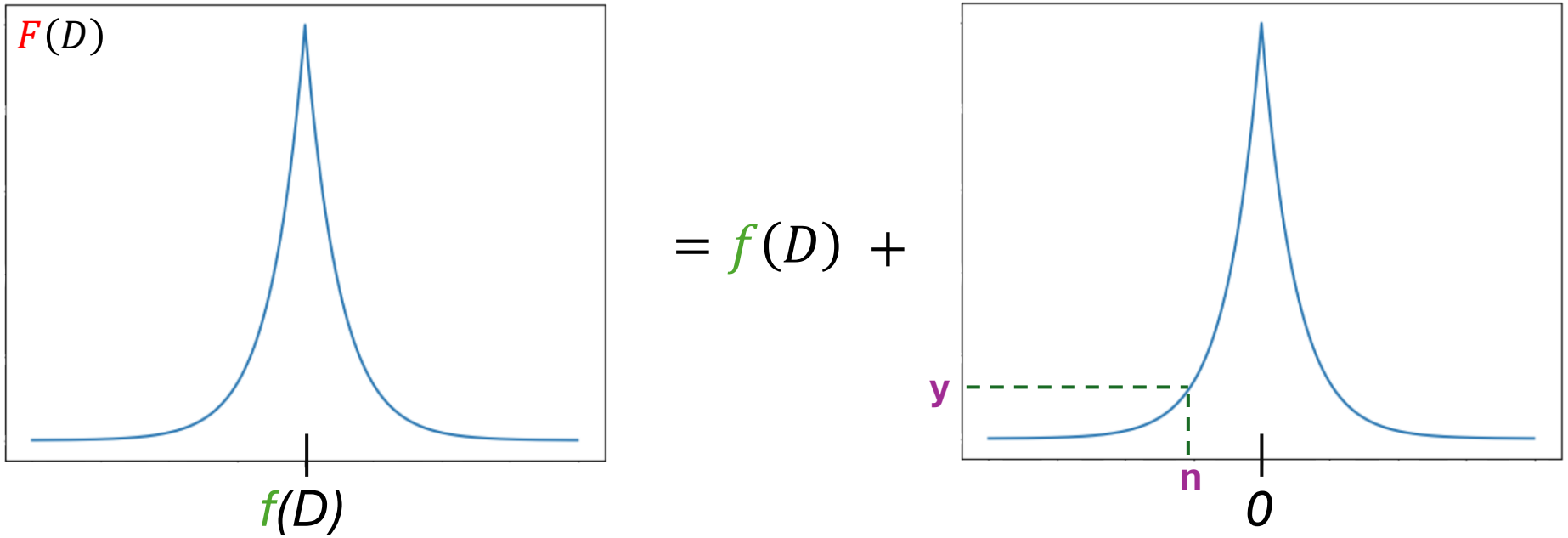
Differential Privacy

Adding Random Noise

$$F(D) = f(D) + randomised\_noise$$

$f(D)$: $query\ f\ applied\ on\ dataset\ D$ (ex : Count rows if PTSD = 1)

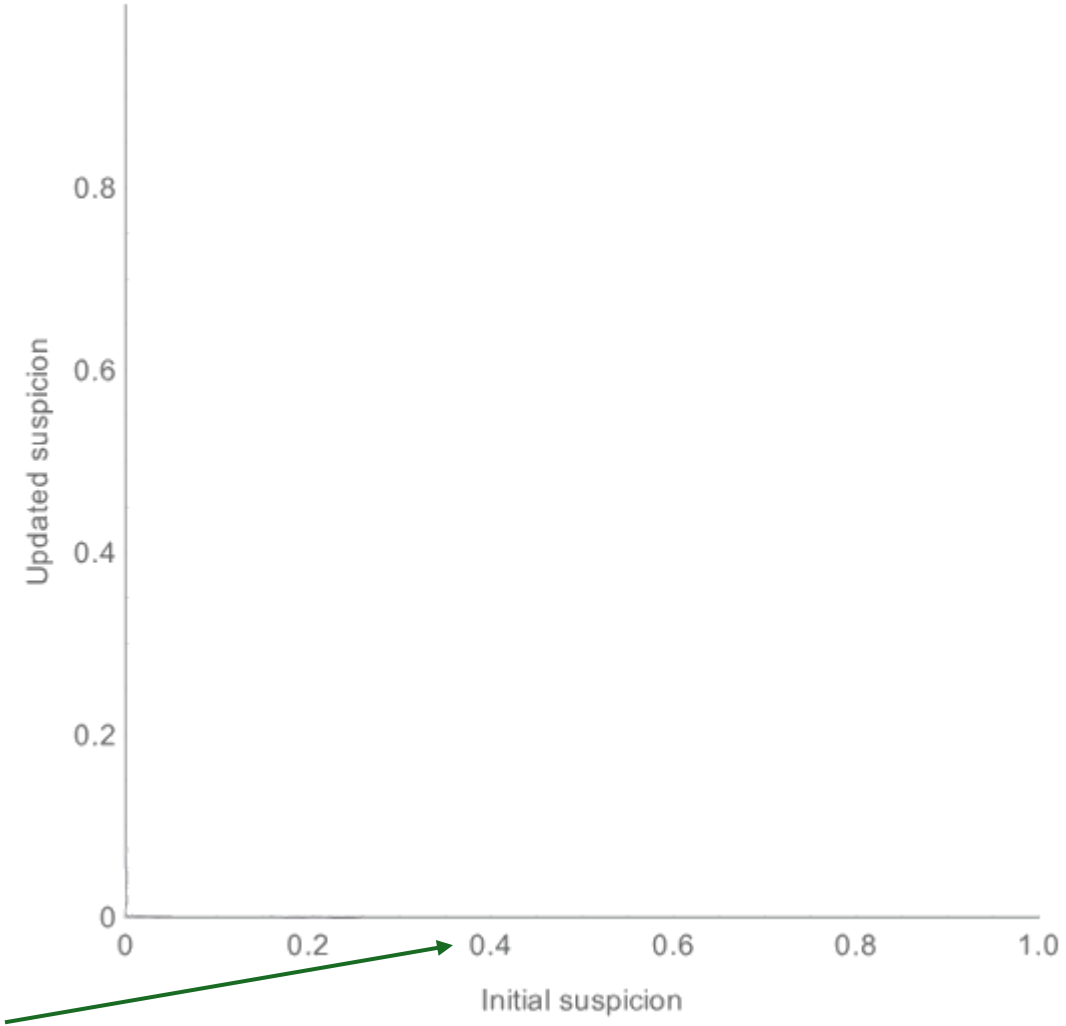
$F(D)$: $Differentially\ private\ mechanism\ F\ applied\ on\ dataset\ D$



Noise **n** is released with **y** probability

Differential Privacy

Probability of Attackers' Suspicion



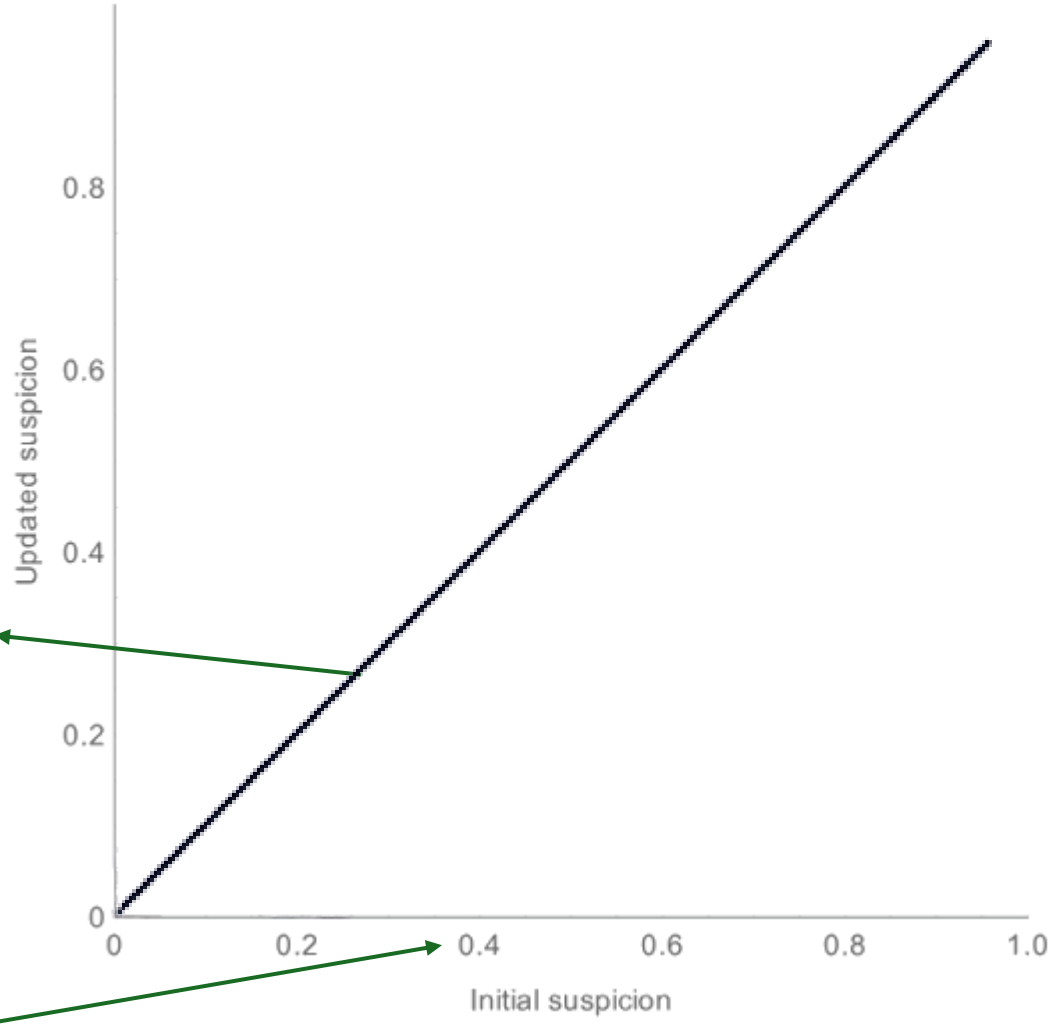
Example Suspicion :

P(Homer is in the dataset) = 0.4

Differential Privacy

Probability of Attackers' Suspicion

After looking at DP result
Suspicion is the same

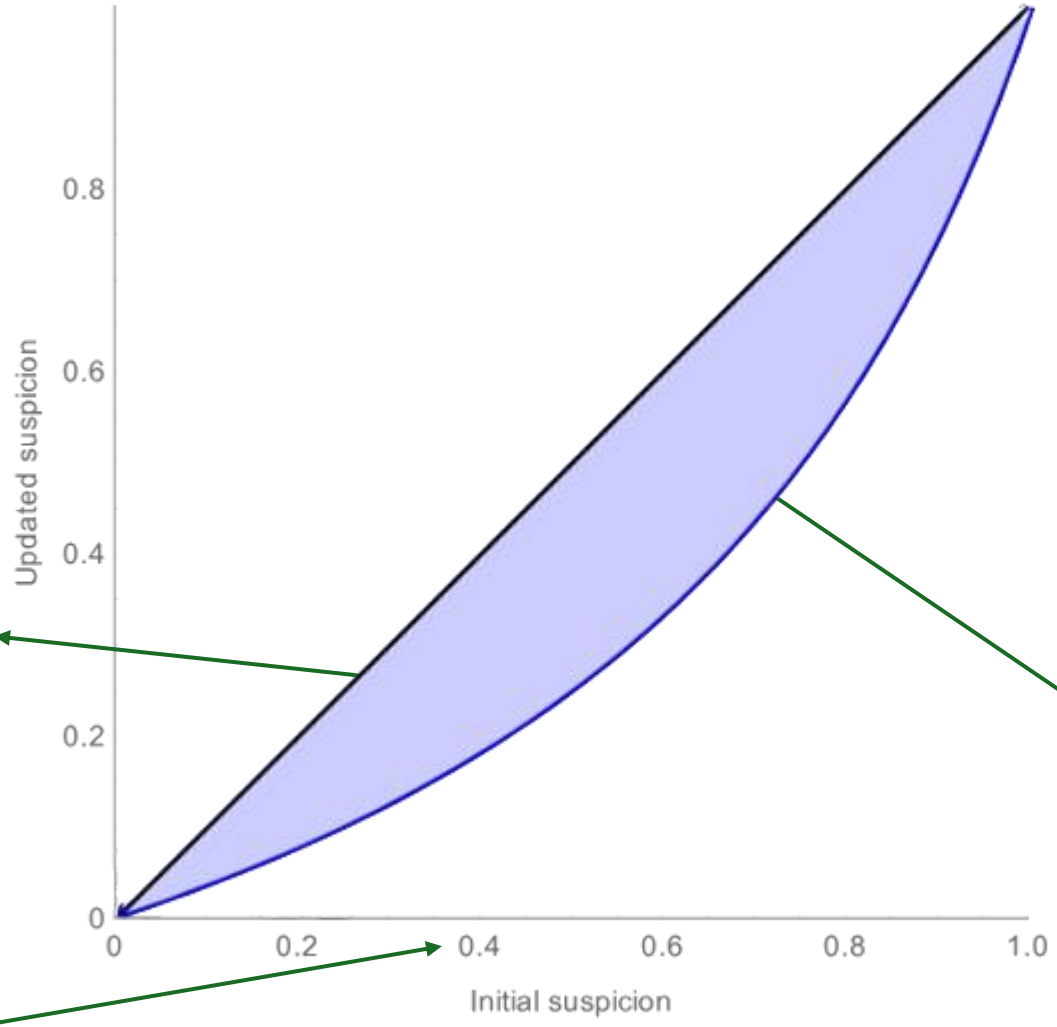


Example Suspicion :
P(Homer is in the dataset) = 0.4

Differential Privacy

Probability of Attackers' Suspicion

After looking at DP result
Suspicion is the same



After looking at DP result
Suspicion decreased

Example Suspicion :
P(Homer is in the dataset) = 0.4

Desfontaines, D. (2021). A friendly, non-technical introduction to differential privacy. https://desfontain.es/blog/friendly-intro-to-differential-privacy.html

Differential Privacy

Probability of Attackers' Suspicion

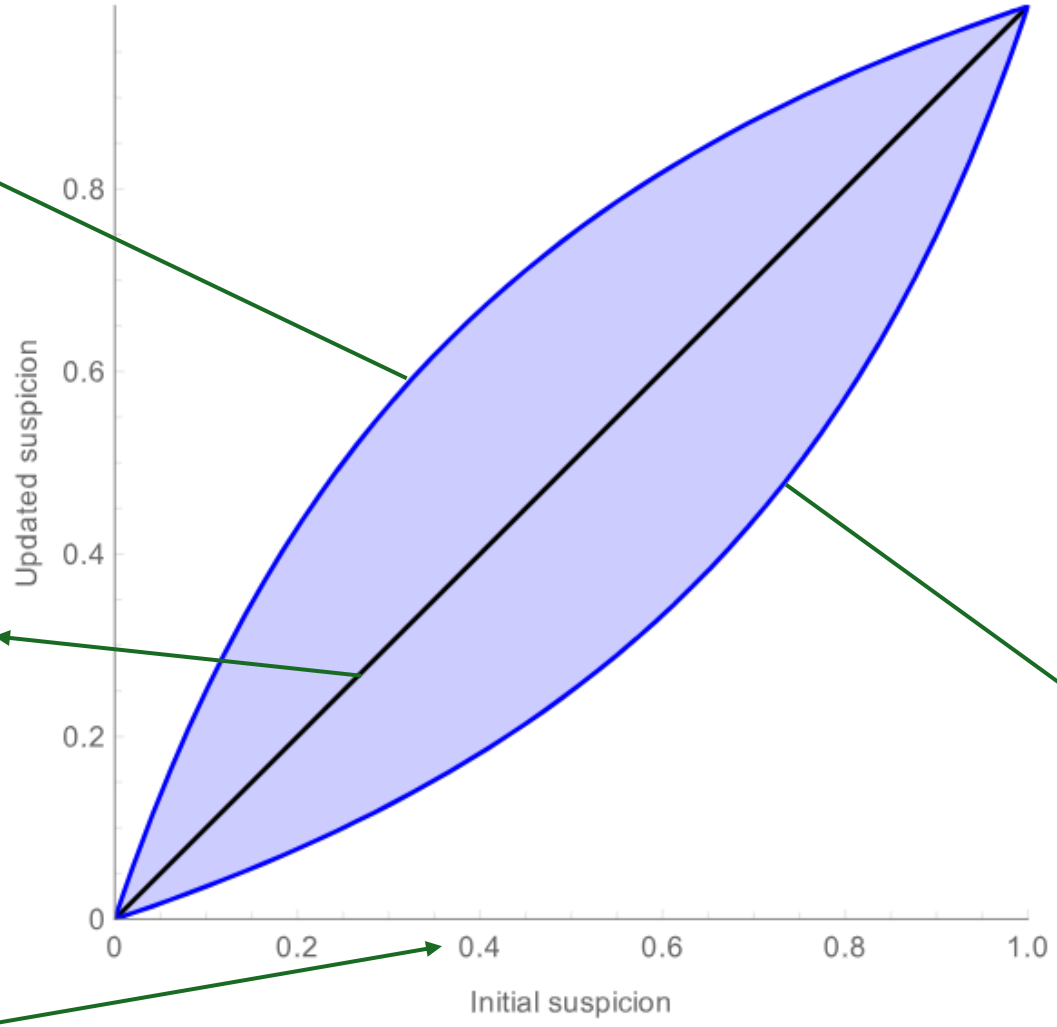
After looking at DP result
Suspicion increased

After looking at DP result
Suspicion is the same

After looking at DP result
Suspicion decreased

This area depends on the amount of noise

How much noise?
Initial Sus ≈ Updated Sus



Example Suspicion :
P(Homer is in the dataset) = 0.4

Desfontaines, D. (2021). A friendly, non-technical introduction to differential privacy. https://desfontain.es/blog/friendly-intro-to-differential-privacy.html

Differential Privacy

Why Magical?

- It works no matter what the attacker knows about the data or related background data
- We can **quantify** the privacy loss
- Multiple parties cannot conspire to reconstruct the original data

Privatize dataset D → give noisy results to Ann

Privatize dataset D → give noisy results to John

The union of the 2 results are still anonymous

Ɛ - Differential Privacy

$$\frac{Pr\ [F(D) \in \mathrm{O}]}{Pr\ [F(D') \in \mathrm{O}]} \leq\ e^{\varepsilon}$$

e – Euler's number ≈ 2.71828

Ɛ – privacy loss (>0)

D, D' – neighbouring datasets

O – possible set of outputs

If the above relationship holds for **all** neighbouring datasets:

"F satisfies Ɛ - Differential Privacy "

privacy loss/consumption (Ɛ) :

amount of privacy that an individual lose by participating in the analyses done by F

smaller Ɛ -> better privacy

Without Differential Privacy

f : Count rows if PTSD = 1

f(D) = 2

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	35	1

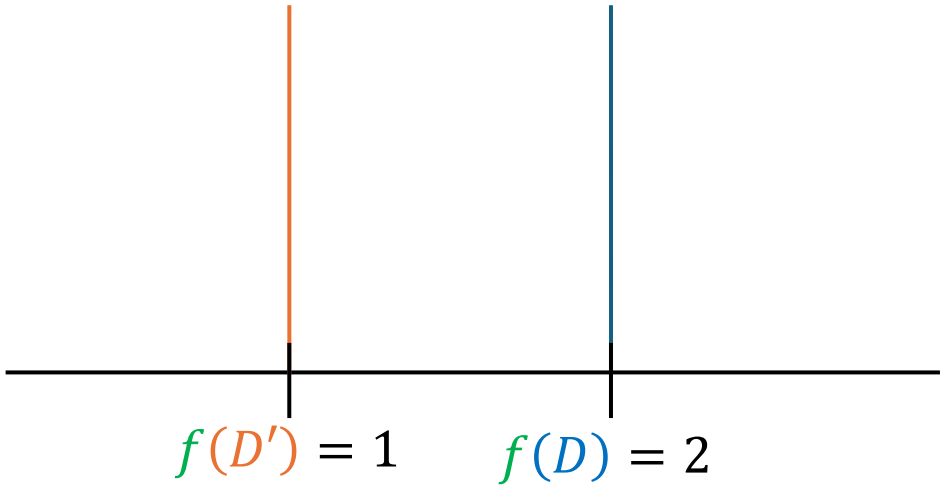
f(D') = 1

Name	Age	PTSD
Ann	21	1
David	16	0

Assume we run the query 10 times

Let's compare the probabilities of the result = 2

$$\frac{Pr\,[f(D)\ =2]}{Pr\,[f(D')\ =2]} = \frac{10/10}{0/10} = \frac{1}{0} = \infty$$



(∞) Higher Difference -> Easy to distinguish -> What about privacy?

Ɛ - Differential Privacy

f : Count rows if PTSD = 1

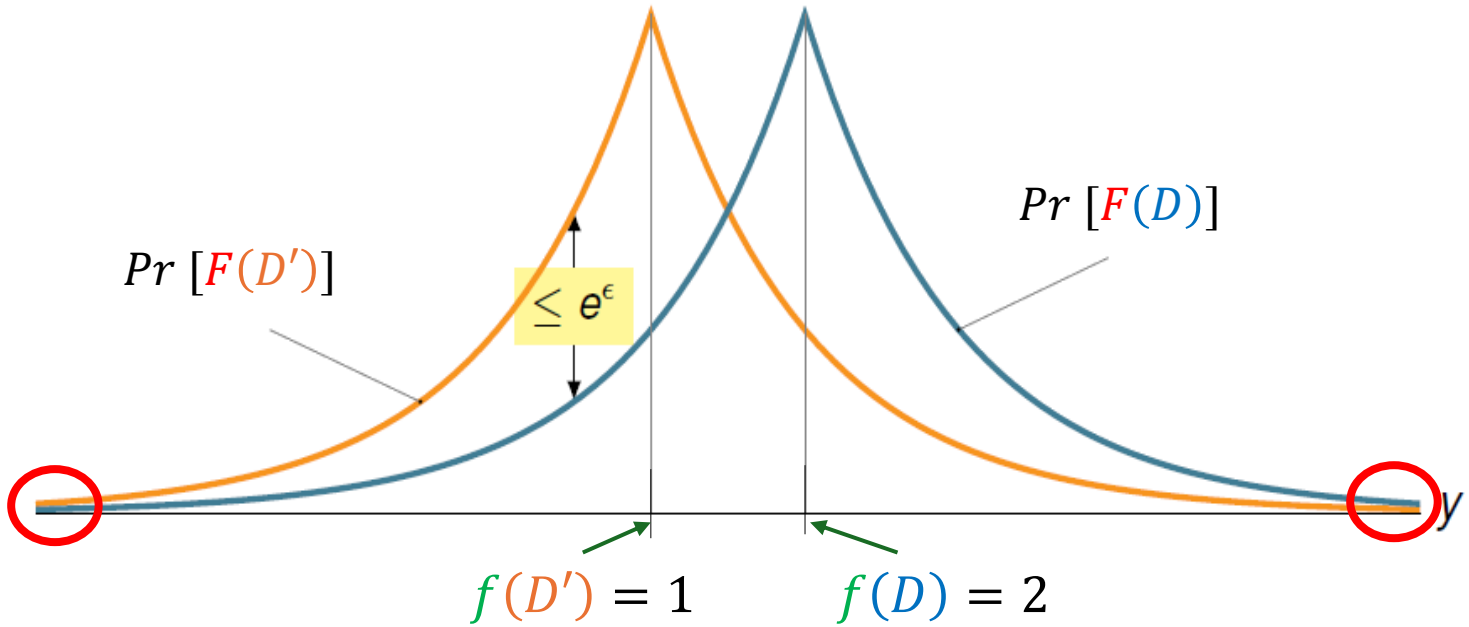
f(D) = 2

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	35	1

f(D') = 1

Name	Age	PTSD
Ann	21	1
David	16	0

$$\frac{Pr\,[F(D) \in O]}{Pr\,[F(D') \in O]} \leq e^{\varepsilon}$$



What happens when Ɛ increases?

What happens when Ɛ decreases?

Ɛ - Differential Privacy

What kind of noise

- To achieve DP for a query we need to add **randomised** noise to the query result
- To calculate the noise, there are several mechanisms in the DP field
- For Ɛ - DP we use the **Laplacian mechanism** to generate noise

$$F(D) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right)$$

$Lap\left(\frac{\Delta f}{\varepsilon}\right)$ Laplace distribution with mean 0 and **sensitivity** (Δf) of query $f(real\ query)$

Δf : the **maximum change** of $f(real\ query)$ *result when applied to all neighbouring datastes*

Ɛ - Differential Privacy

Sensitivity

Δf : the maximum change of $f's\ result\ when\ f\ is\ applied\ to\ all\ neighbouring\ datastes$

$$\Delta f = max|f(D) - f(D')|$$

It is easy to calculate the sensitivity of counting queries

Ɛ - Differential Privacy

Sensitivity

It is easy to calculate the sensitivity of counting queries

$$\Delta f = max|0,1,1| = 1$$

Sensitivity of counting queries is always 1

f : Count rows if PTSD = 1 is performed on dataset D

f(D) = 2

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	35	1

f(D'') = 1

Name	Age	PTSD
Ann	21	1
David	16	0

f(D') = 2

Name	Age	PTSD
Ann	21	1
David	16	0
Kate	36	1

f(D''') = 1

Name	Age	PTSD
David	16	0
Kate	35	1

Ɛ - Differential Privacy

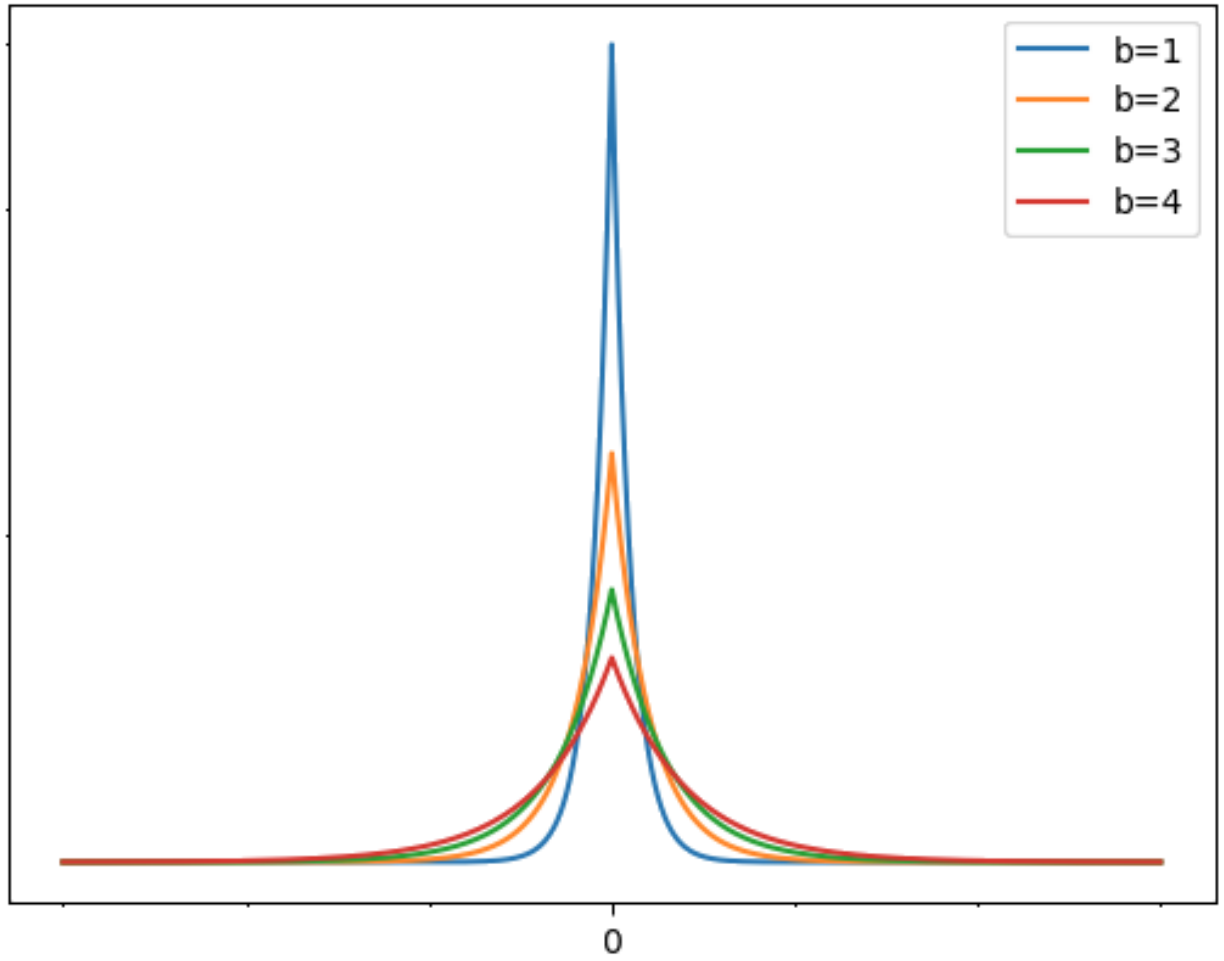
Laplacian Distribution

$$Lap(x) = \frac{1}{2b} e^{\frac{-|x-\mu|}{b}}$$ *Gives the noise* ***distribution***

x is a random real number

μ mean of the distribution (= 0 for DP)

b **scale** of the noise distribution ; $b = \frac{\Delta f}{Ɛ}$



What happens to the amount of noise?

For a given query increase Ɛ...

For a given query decrease Ɛ...

Ɛ - Differential Privacy

Laplacian Noise

In Python

```python
import numpy as np

sensitivity = 1;
epsilon = 0.1

laplace_noise =  np.random.laplace(loc=0, scale=sensitivity/epsilon)
```

loc = μ
sensitivity = Δf
epsilon = Ɛ

In Java

```java
import java.util.Random;

class DifferentialPrivacy {
    public static double laplaceNoise(double sensitivity, double epsilon){
        double randomNum = new Random().nextDouble() - 0.5; //random number range [-0.5, 0.5], you can give any
            range
        return -sensitivity * Math.signum(randomNum) * Math.log(1 - 2 * Math.abs(randomNum)) / epsilon;

}
```

Ɛ - Differential Privacy

Sensitivity

Give the sensitivity of each query

- How many people are in the dataset?
- How many people have an educational qualification level equal to or below 7?
- What is the sum of the ages of people with an educational qualification level above 7?
- What is the average age of people with an educational qualification level above 7?

Ɛ - Differential Privacy

Clipping

- It is difficult to calculate the sensitivity of some queries (e.g., summation and average)
- Need clear lower and upper bounds for the query answers
- Clip (limit) the data at reasonable lower and upper bounds
- How can we find the clipping bounds:
    1. Known knowledge about the data values
    (e.g., can be provided as metadata or we know it by general knowledge)
    2. Executing differentially private queries
- Aggressive clipping bounds can cause information loss

    If the difference between the bound is smaller => smaller sensitivity => low noise

Ɛ - Differential Privacy

Clipping

1. Known knowledge about the data values

   f : What is the sum of the ages of people with an educational qualification level above 7?

   An individual can likely to live up to 90 years

   Lower bound = 0

   Upper bound = 90 ; if there are ages greater than 90, they have to be set to 90 (clip)

   Within neighboring datasets,
   maximum change in the output of above summation query is 90

   Therefore, sensitivity (Δf) is 90

Ɛ - Differential Privacy

Clipping

2. Executing differentially private queries

f : What is the sum of the ages of people with an educational qualification level above 7?

*Select a value **i** for **age** in a range [0, 100]*

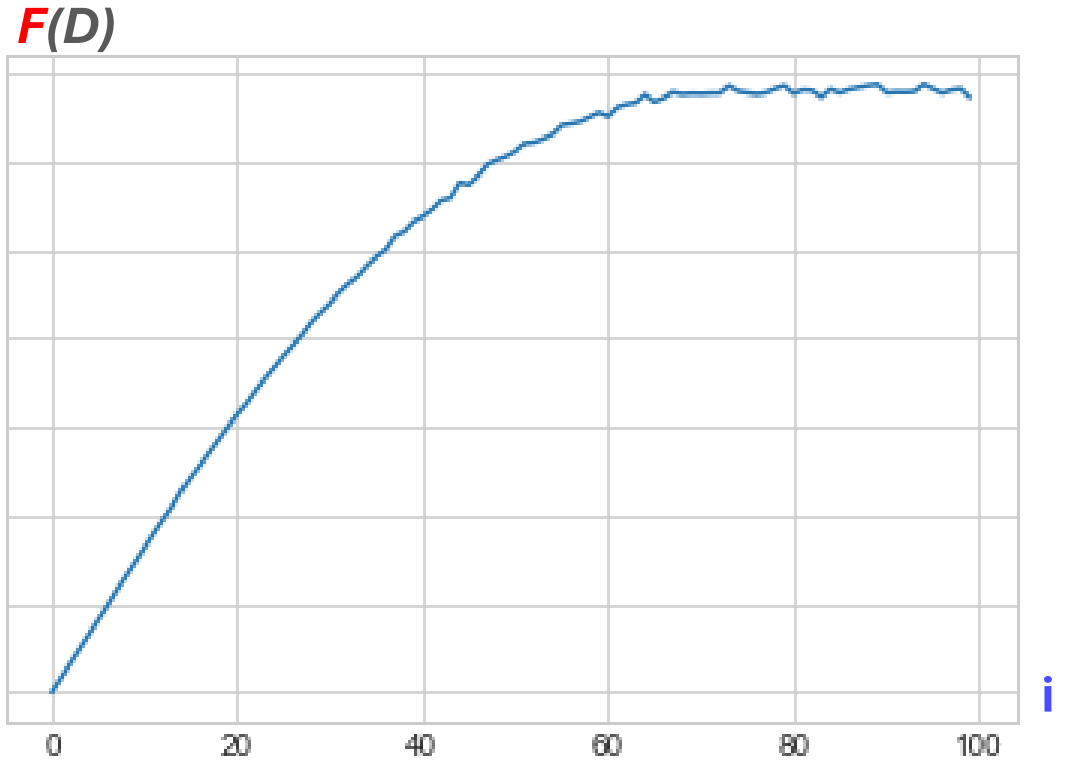
*For each **i***

*We clip all ages at **i***

Then generate noisy result

$$F(D) = f(D) + Lap\left(\frac{i}{\varepsilon}\right)$$

Real Ages	[10, 20, 65]	$f(D)$ = 95
Ages clipped **i** = 10	[10, 10, 10]	$f(D)$ = 30
Ages clipped **i** = 20	[10, 20, 20]	$f(D)$ = 50
Ages clipped **i** = 80	[10, 20, 65]	$f(D)$ = 95

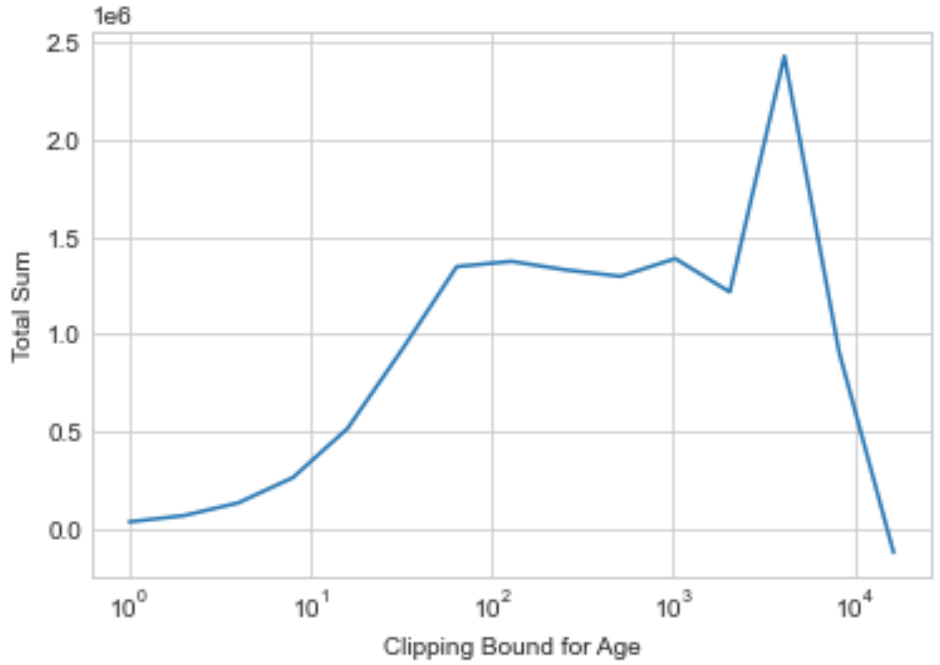


*Select the **i** value where F(D) starts to get **saturated***

Ɛ - Differential Privacy

Clipping

- But for some data fields it is hard to do the calculation between small range of ***i***
- What about data fields such as annual income
- Instead of setting the sensitivity to ***i***, use a logarithmic value such as 2^{i}



Differential Privacy

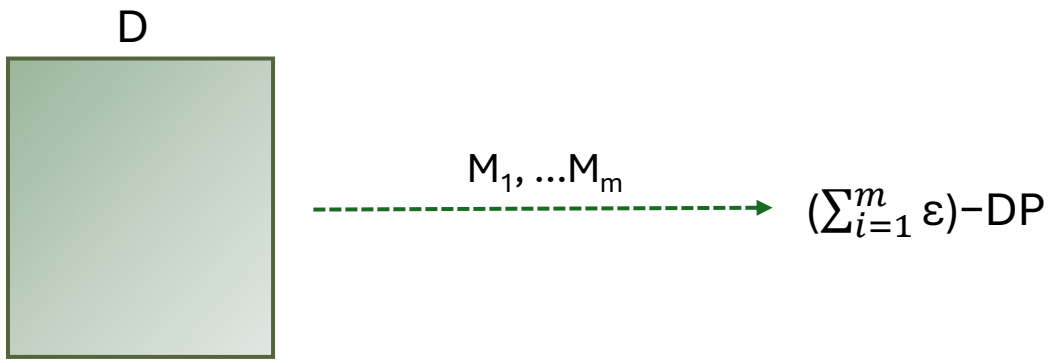
Properties

1. Sequential Composition

Suppose we have a set of privacy mechanisms $M = \{M_1, \ldots M_m\}$, each performed on the **same** dataset

if each M_i provides ε_i privacy guarantee on the entire dataset

then M will provide $(\sum_{i=1}^{m} \varepsilon)-DP$



Have you heard about the 20 Questions game?

Differential Privacy

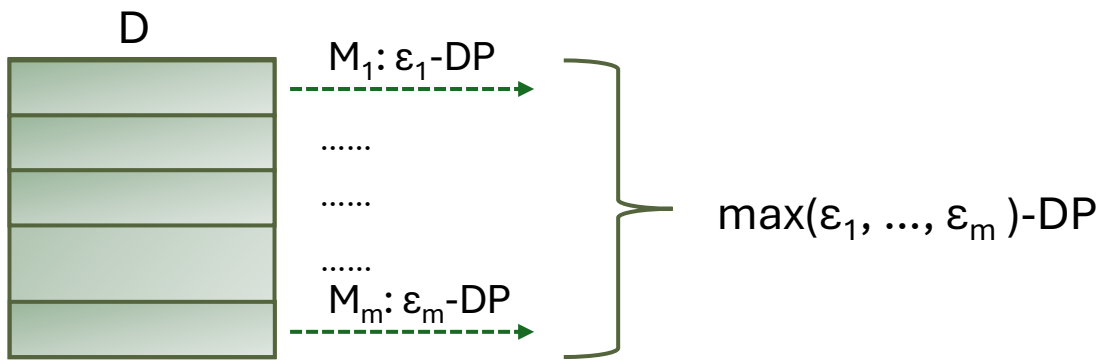
Properties

2. Parallel Composition

Suppose we have a set of privacy mechanisms $M = \{M_1, \ldots M_m\}$,

if each M_i provides ε_i privacy guarantee on a **disjointed subset** of the entire dataset

then M will provide $\max(\varepsilon_1, \ldots, \varepsilon_m)$-DP



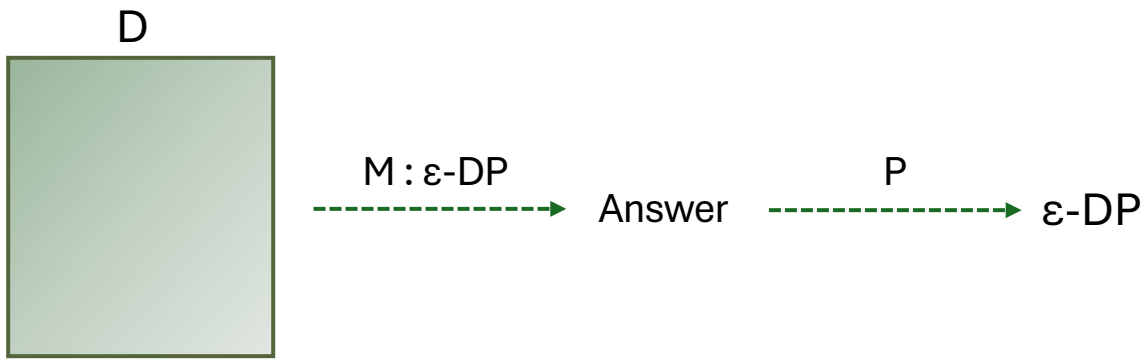
Differential Privacy

Properties

3. Post-processing

Suppose we have a privacy mechanisms M that provide ε differential privacy

Any function P that tries to further process the output of M also satisfies ε-differential privacy



Differential Privacy

Privacy Budget

- The total amount of privacy we are **willing** to lose
- The total ε must not exceed the budget
- Pre-defining a privacy budget will limit the number of queries that can be run on a dataset
- Parallel composition allows more efficient use of the privacy budget
- Even the queries we use to find the sensitivity bounds in clipping consume the privacy budget

Differential Privacy

How to decide DP is suitable

- Analysis handles personal data
- Analysis does not require to link back to the individuals
- Analysis on the data are not very sensitive to changes in the data
- Has enough resources to perform the analysis

Differential Privacy

Weaknesses

- Privacy guarantee become gradually weak for sequential composition
- The performance of the DP enabled applications can decrease
- Might require higher computational power
- The privacy budget limits the chances of gaining insights
- Not suitable for higher sensitivities

Differential Privacy

Other Types

- (Ɛ , δ) - Differential Privacy
- Rényi Differential Privacy
- Pufferfish Privacy

Summary

- Differential Privacy
- Laplacian mechanism
- Sensitivity
- Drawbacks of Differential Privacy