$$\frac{dy}{dx}$$

Differential Privacy

Dr. Balázs Pejó

www.crysys.hu

© 2022 CrySyS Lab, BME

Agenda

$$\frac{dy}{dx}$$

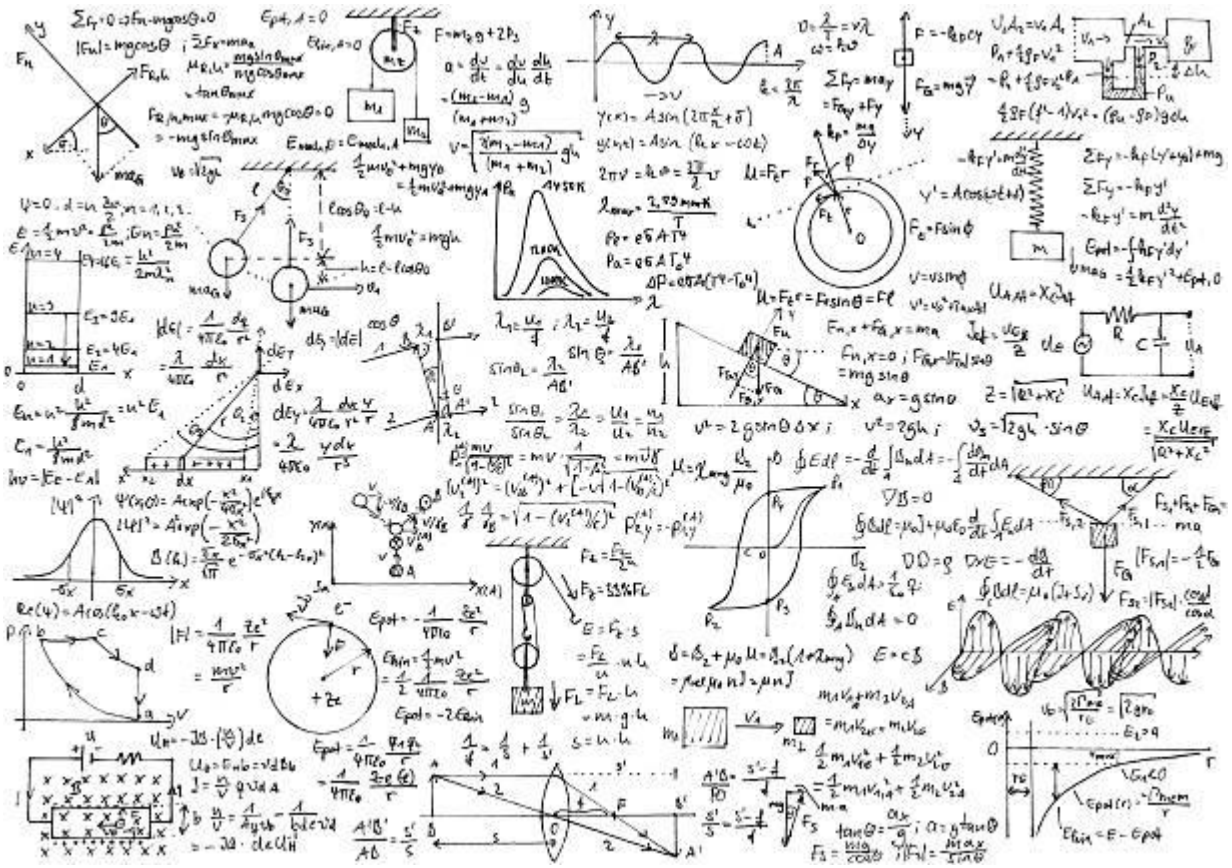
- Dark Patterns
- Tracking
- GDPR
- Deidentification
- Machine Learning
- Anonymization
- Cryptography

- Basics
  - Definition
- Mechanisms
  - Laplace / Gaussian
  - Sensitivity
- Properties
  - Composition
  - Axioms
- Settings
  - Central / Local
- Dimensions
- Deployments

Recap

$$\frac{dy}{dx}$$

- Using anonymity primitives in an AdHoc manner is not sufficient.
  - Determining the privacy model (e.g., the capabilities of the attacker) is crucial.
- K-Anonymity is wide-spread, but not perfect.
  - L-Diversity & T-Closeness are merely naïve fixes.
  - K-Map & D-Presence are weaker, and hard to use.
- Synthetic data generation is not a panacea.



$$\frac{dy}{dx}$$

Differential Privacy

Absolute vs Relative Privacy

$$\frac{dy}{dx}$$

- Absolute privacy: access to the published data should not enable the adversary to learn anything extra about any individual compared to no access to the data.
  - Not feasible in practice.

- One can learn from a dataset that the average height in Hungary is 175 cm.
- One knowns that Peter (not in the dataset) is taller by 10 cm than the average.

- Is this a privacy breach?
  - In an absolute sense, yes.

Dalenius Privacy Goal

$$\frac{dy}{dx}$$

- Background knowledge allows absolute privacy breach, even if the victim is not in the dataset!

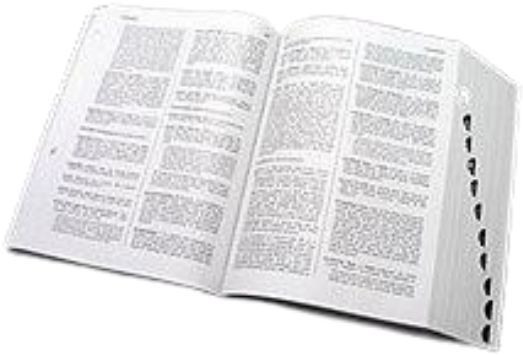
- Dalenius : nothing about an individual should be learnable from the database that cannot be learned without access to the database.
  - Impossible to achieve with non-trivial utility.

- Differentially privacy does not protect against absolute breach.
  - These breaches happen no matter if the target is in the data or not.
- Differential privacy aims to hide only those information that is specific to Peter (or any single individual in the dataset).

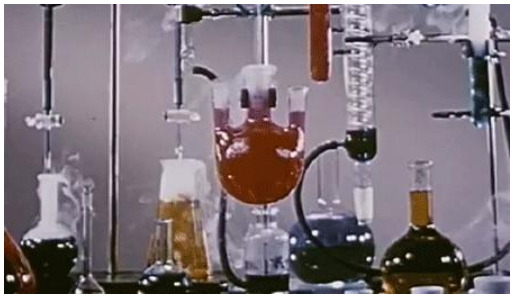
Other Privacy Notions

$$\frac{dy}{dx}$$

- All previous privacy notions needed some assumptions about the attacker.
  - How much prior knowledge do they have?
  - What auxiliary data are they allowed to use?
  - What kind of information do they want to learn?



- All previous privacy notions were shown to be broken in some circumstances.
- All previous notions were a property of the output data.
- Instead, Differential Privacy is a property of a process.
  - You can't look at the output data and determine whether it satisfies DP.
  - Rather, you have to know how the data was generated to determine that.
- The process can be anything.
  - Calculating some statistics.
  - A machine learning training process.
  - …

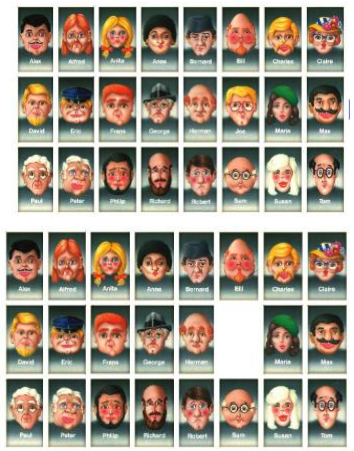


Basic Idea

$$\frac{dy}{dx}$$

- A creepy person trying to figure out whether the target is in the original data.

- By looking at the output, he cannot be 100% certain of anything.
  - The result could have come from a database with the target in it.
  - It could also have come from the exact same database, without the target.



DP Mechanism

Data

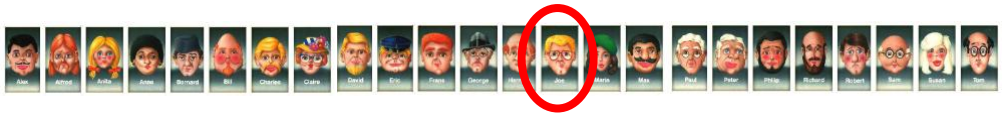
Data

Basically the same

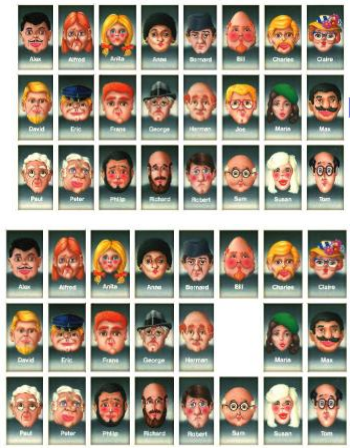
(Un)deterministic Process

$$\frac{dy}{dx}$$

- Attacker knows everything …
  - All data (even the target)



- … except whether a particular person is in the data or not.
- Deterministic process cannot satisfy DP.
- Randomness must be present in the process.
- Basically the same → the probabilities are close
- Uncertainty in the process means uncertainty for the attacker, which means better privacy.



"How many man are in the data?"

probabilities are close

21

21

Tri[illegible]y
dif[illegible]t

Attacker PoV

$$\frac{dy}{dx}$$

- The attacker cannot gain a lot of information about their target.
- The attacker knows that the actual database D can be either D_{in} or D_{out} depending on the presence or absence of their target.

- Attacker might have an initial suspicion about the database, i.e., $\Pr[D = D_{in}]$ $(= 1 - \Pr[D = D_{out}])$.

- The mechanism returns O, how much information did the attacker gain?
- Can be measured by how much the suspicion changed after seeing the output.
- DP ensures that the updated probability is never too far from the initial suspicion.

Noise

$$\frac{dy}{dx}$$

- The expected answer with the noise should be equal with the true answer, otherwise the utility would be reducing due to this bias.
  - The mean of the added noise is zero.
- The variance determines the level of privacy.
  - Bigger variance implies better protection but lower utility.
  - Smaller variance implies worst protection but higher utility.

Input data: D
Mechanism: M
E(M(D)) =
True Answer

ε-DP

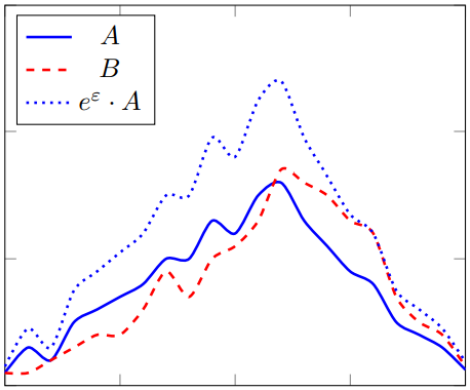
$$\frac{dy}{dx}$$

- The probabilities are close for any outcome.
  - Pr[M(D) = O] should be close to Pr[M(D') = O]
- ∀ O: Pr[M(D) = O] ≤ e^{ε} · Pr[M(D') = O]
  - ε > 0 is the privacy budget (parameter).
  - The degree of closeness depends on ε.

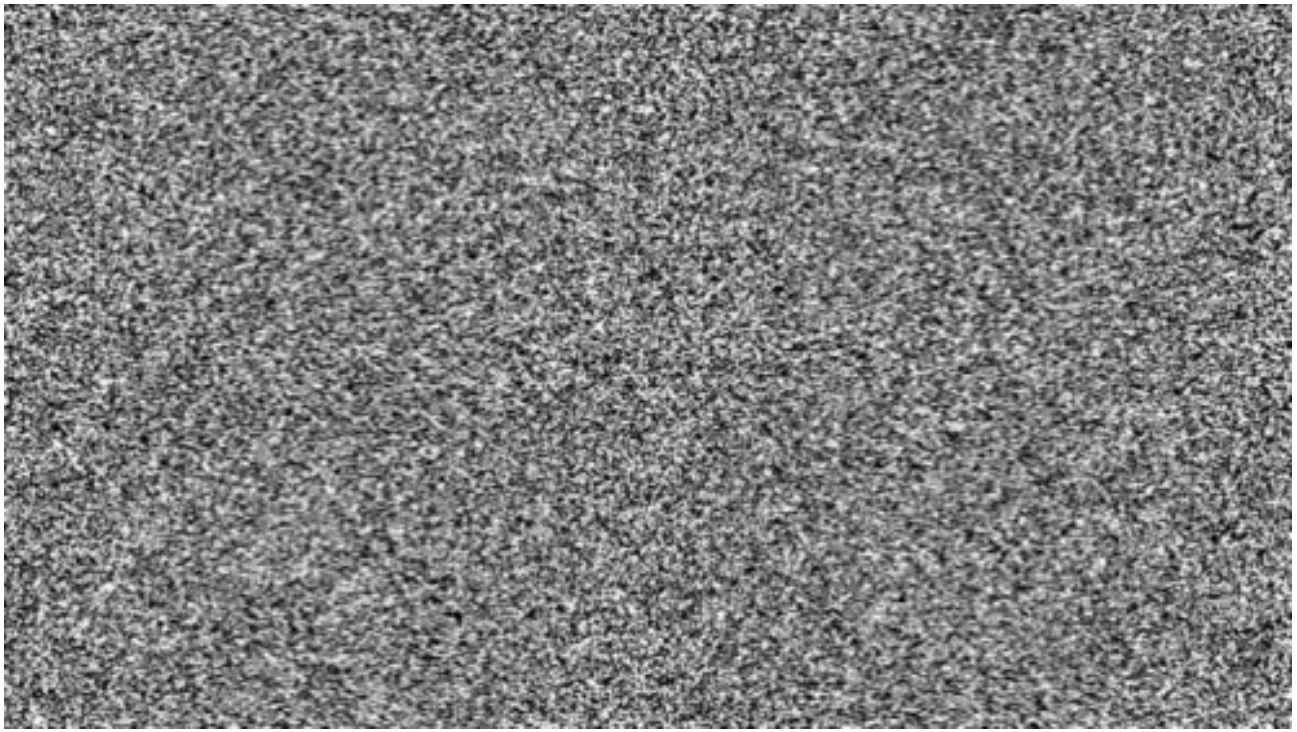
Input data: D, D'
(Differing in one record)
Privacy Mechanism: M
Output: O ∈ Ω

$$e^{-\varepsilon} \leq \frac{\Pr[M(D) = O]}{\Pr[M(D') = O]} \leq e^{\varepsilon}$$

- This is symmetric as D and D' are interchangeable.
  - $e^{-\varepsilon}$ · Pr[M(D') = O] ≤ Pr[M(D) = O] ≤ e^{ε} · Pr[M(D') = O]
- Originates from the cryptographic notion indistinguishability.
- Due to continuous event space (i.e., when infinite outcome is possible) the definition uses sets of outputs instead of individual outputs.
- ∀ S ⊆ Ω: Pr[M(D) ∈ S] ≤ e^{ε} · Pr[M(D') ∈ S]



$$\frac{dy}{dx}$$

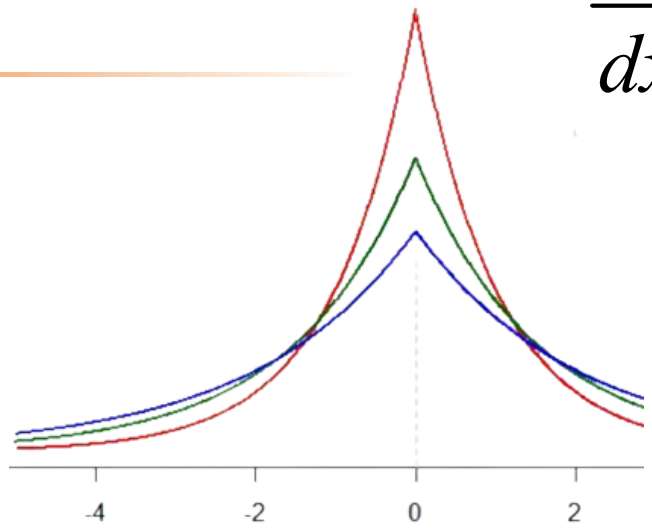


Noise

Laplacian Mechanism

$$\frac{dy}{dx}$$

- Lap(0, b) = Lap(b)
- ∀ O: Pr[M(D) = O] ≤ e^{ε} · Pr[M(D') = O]
  - Pr[M(D) = 21] = Pr[Noise = 2]
  - Pr[M(D') = 21] = Pr[Noise = 3]
  - Pr[Lap() → 2] ≤ e^{ε} · Pr[Lap() → 3]



$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

- Pr[Lap() → x] ≤ e^{ε} · Pr[Lap() → x ± 1]
  - Always true if variance (b) is 1/ε.

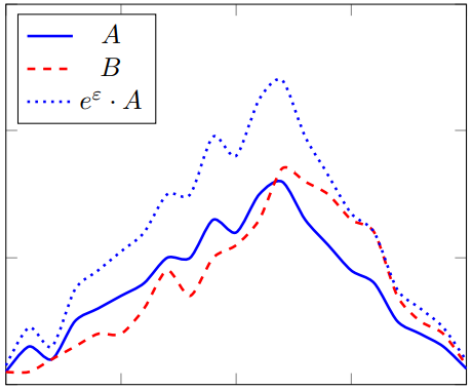
Approximated DP ((ε, δ)-DP)

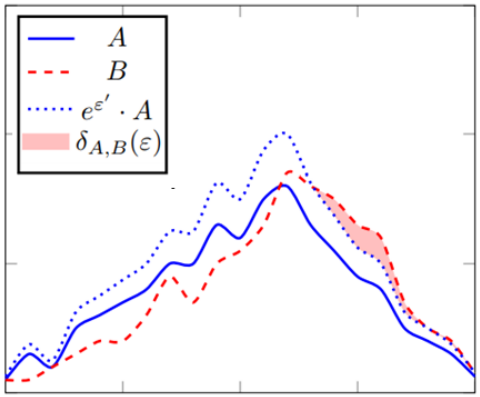
$$\frac{dy}{dx}$$

- In many settings achieving ε-DP is impossible or comes at a very high utility cost.
- $\forall S \subseteq \Omega$: $\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S] + \delta$

- Catastrophic failures of privacy:
  - Database D with n records.
  - Share randomly δ·n records.
  - Such mechanism is (0, δ)-DP.

- To avoid this in practice one must set δ small:
  - $\delta < n^{-1}$ implies the records are safe.



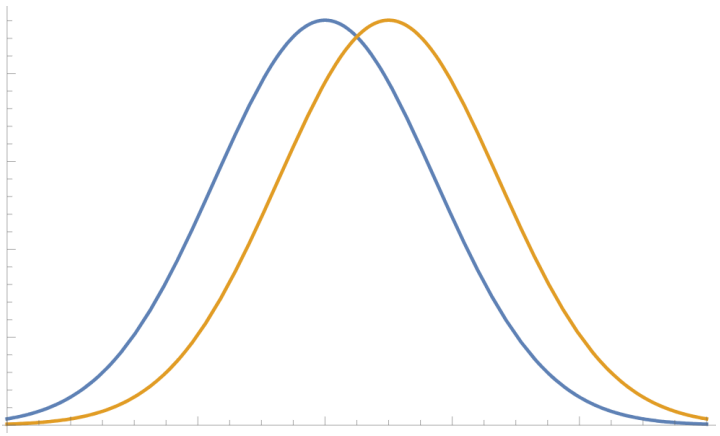


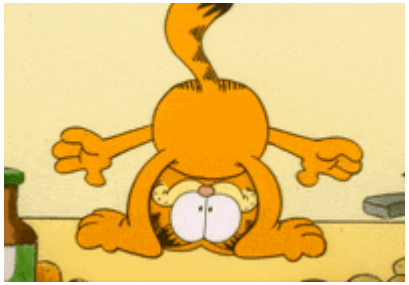
Gaussian Mechanism

$$\frac{dy}{dx}$$

- Laplace noise is exotic: it never shows up in the real world.
- Normal distributions are more familiar and friendly, many natural data distributions can be modeled with them.
- Via Gaussian noise one cannot obtain pure/vanilla ε-DP.
  - However, it is possible to obtain (ε, δ)-DP.
- If the privacy guarantee is weaker, why bother?
  - Tail behaves nicely.
- Draw 1.000.000 elements.
  - From Gauss approx. 0
  - From Laplace approx. 850

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$





will be outside
[- 5·σ, 5·σ].

Laplace vs Gauss

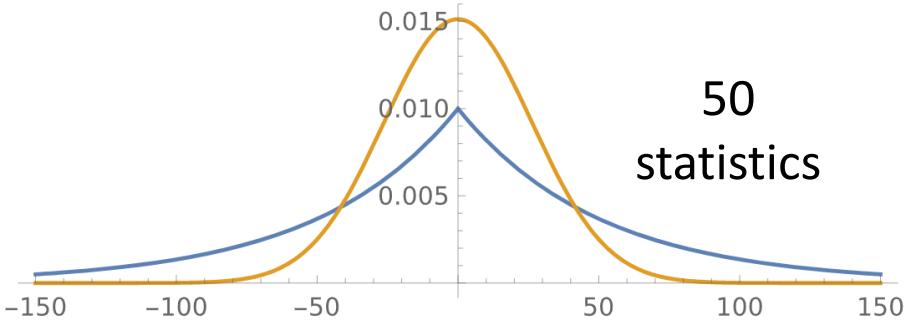
$$\frac{dy}{dx}$$

- Laplace is better if we are releasing only a few statistics.
- Gaussian is better if we are releasing multiple statistics.
- The noise is scaled with the sensitivity.
- The sensitivity of Laplacian is measured in L_1 (or Manhattan dist.).
  - Answering k query requires k times more noise.

$$\Delta_1(f) = \max \sum |f_i(D_1) - f_i(D_2)|$$

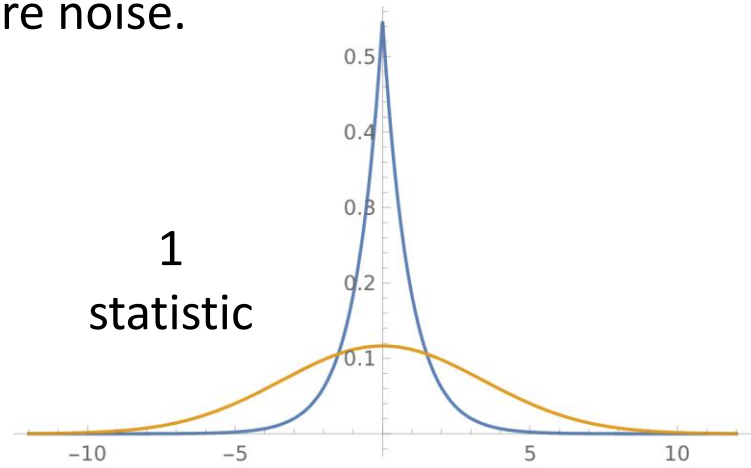
- The sensitivity of Gaussian is measured in L_2 (or Euclidian dist.).
  - Answering k queries requires √k times more noise.

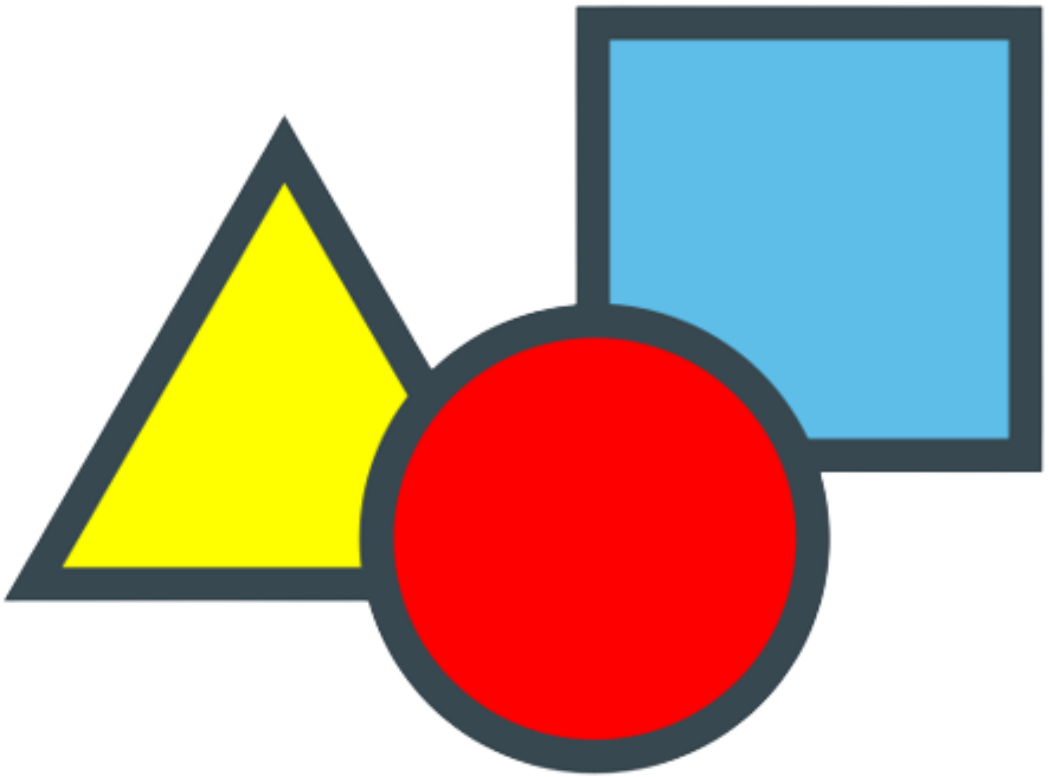
$$\Delta_2(f) = \max \sqrt{\sum |f_i(D_1) - f_i(D_2)|^2}$$



50 statistics

1 statistic





$$\frac{dy}{dx}$$

Properties

DP 1

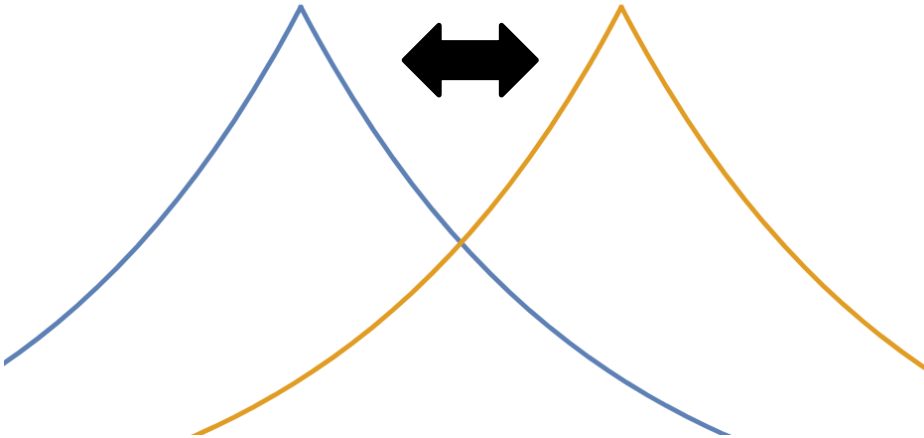
Libraries
(*Opacus*)

Sensitivity

$$\frac{dy}{dx}$$

- Counting is easy, one data influences the result at most by one.
- How about other statistics such as AVG or SUM?
- Sensitivity (Δ): the maximum influence (worst case) of a single datapoint on the desired statistics.
  - For instance, ratings from 10 to -10.
  - The possible biggest difference is 10 - (-10) = 20.

- Added noise must be scaled with the maximum possible change.
  - E.g., 20 · Lap(1/ε) = Lap(20/ε).



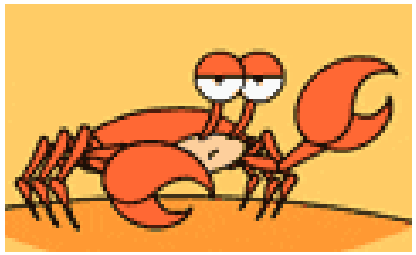
- Salary of Amazon employee:
  - John Smith – 65K
  - Sarah Parker – 58K
  - Michael Keen – 41K
  - …
  - Jeff Bezos – 676839272615385969607093827162638 4K

Clipping

$$\frac{dy}{dx}$$

- Clipping the values are necessary to bound the sensitivity.
  - Besides the amount of added noise, the size of clipping also influences the final accuracy.
- Estimating the scale (i.e., the minimum and maximum values) is difficult.
- DP-SGD
  - Clipping is heavily used in Machine Learning, were the sensitivity of the training process is not feasible to compute.
  - Bounding a single training step has negligible effect on the entire training process.
  - The gradients for each training step is clipped.

Initialize θ_0 randomly
for $t \in [T]$ **do**
Take a random sample L_t with sampling probability L/N
Compute gradient
For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$
Clip gradient
$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$
Add noise
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
Descent
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$



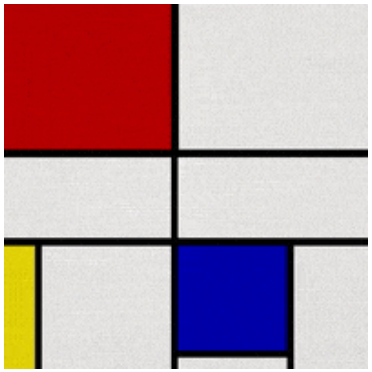
Composition

$$\frac{dy}{dx}$$

- The greatest risk to privacy is that an attacker will combine multiple pieces of information from the same or different sources and that the combination of these will reveal sensitive details.
  - Such as the case for K-Anonymity.

- One cannot study privacy leakage in a vacuum; it is important to reason about the accumulated privacy leakage over multiple independent analyses.

- The differential privacy guarantee of the overall system will depend on the number of analyses and the privacy parameters that they each satisfy.



Composition Types

$$\frac{dy}{dx}$$

- Sequential
  - M_1 is ε_1-DP, M_2 is ε_2-DP → $M(D) = (M_1(M_2(D))$ is $(\varepsilon_1+\varepsilon_2)$-DP.
- Parallel
  - M_1 is ε_1-DP, M_2 is ε_2-DP, $D = D_1 \cup D_2$, $D_1 \cap D_2 = \emptyset$ → $M(D) = (M_1(D_1), M_2(D_2))$ is $\max(\varepsilon_1, \varepsilon_2)$-DP.

ε_1-DP

ε_1-DP

ε_2-DP

DP 2

Example

BONUS POINTS

$\max(\varepsilon_1, \varepsilon_2)$-DP

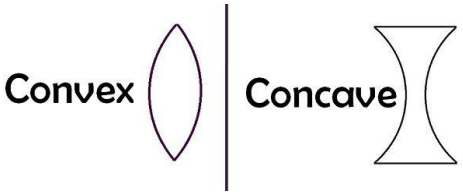
$(\varepsilon_1+\varepsilon_2)$-DP

ε_2-DP

Privacy Axioms

$$\frac{dy}{dx}$$

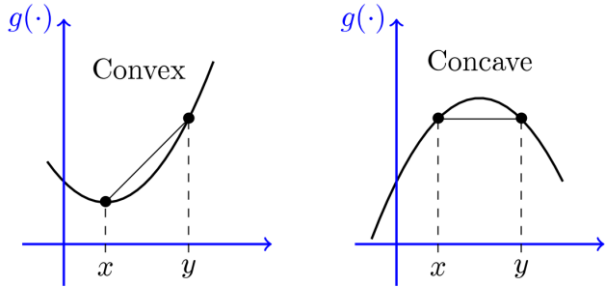
- These are not axioms in a sense that they assumed to be true; rather, they are consistency checks: properties that, if not satisfied by a data privacy definition, indicate a flaw in the definition.



- Convexity or privacy axiom of choice
  - For any two mechanisms M_1 and M_2 satisfying a privacy definition, the mechanism M defined by $M(D) = M_1(D)$ with probability p and $M(D) = M_2(D)$ with probability $1 - p$ also satisfies the same definition.

- Post-processing or transformation invariance

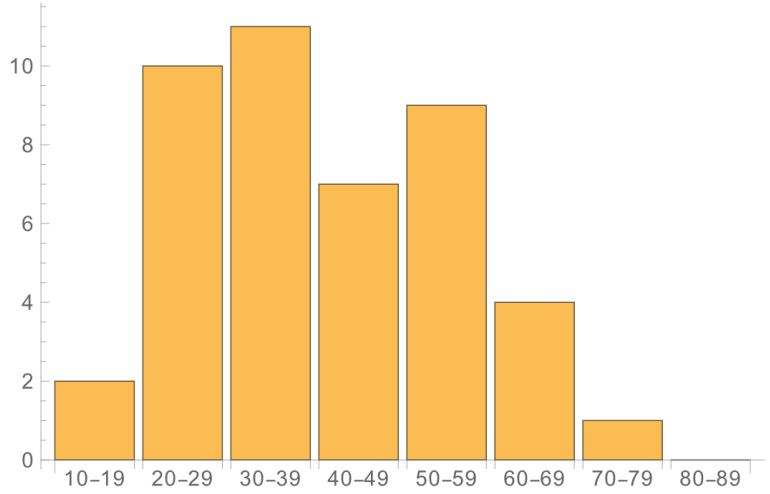
- Differential Privacy satisfy both axiom.



Post Processing

$$\frac{dy}{dx}$$

- Release the number of users depending on their age ranges.
- Each user will have an influence in at most one category.
  - Sensitivity is one for each bucket.
- To provide ε-DP, one must add Lap(1/ε) noise to each statistics.
- It looks a bit weird: the counts are not integers, and they could be negative too.
  - Round all counts and replace all negative numbers with 0.
  - Result is still ε-DP.



- Post Processing: if a mechanism M() is ε-DP, then f(M()) is also ε-DP for any f.

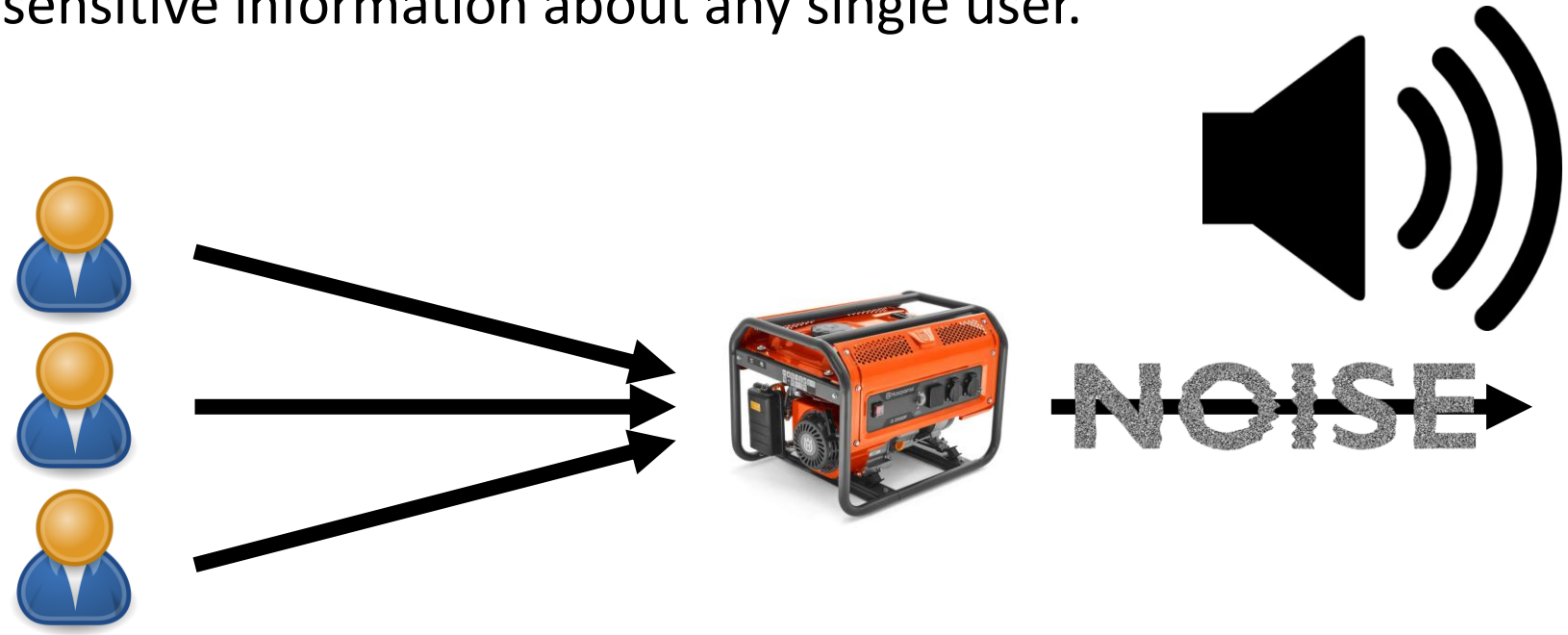
$$\frac{dy}{dx}$$

Settings

Central Model

$$\frac{dy}{dx}$$

- Trust the central server and nobody else.
- Users provides their data to the server, which runs the algorithm on the resulting dataset.
- The output of that algorithm, which is released to the (untrusted) world, needs to be private, and not reveal sensitive information about any single user.



Local Model

$$\frac{dy}{dx}$$

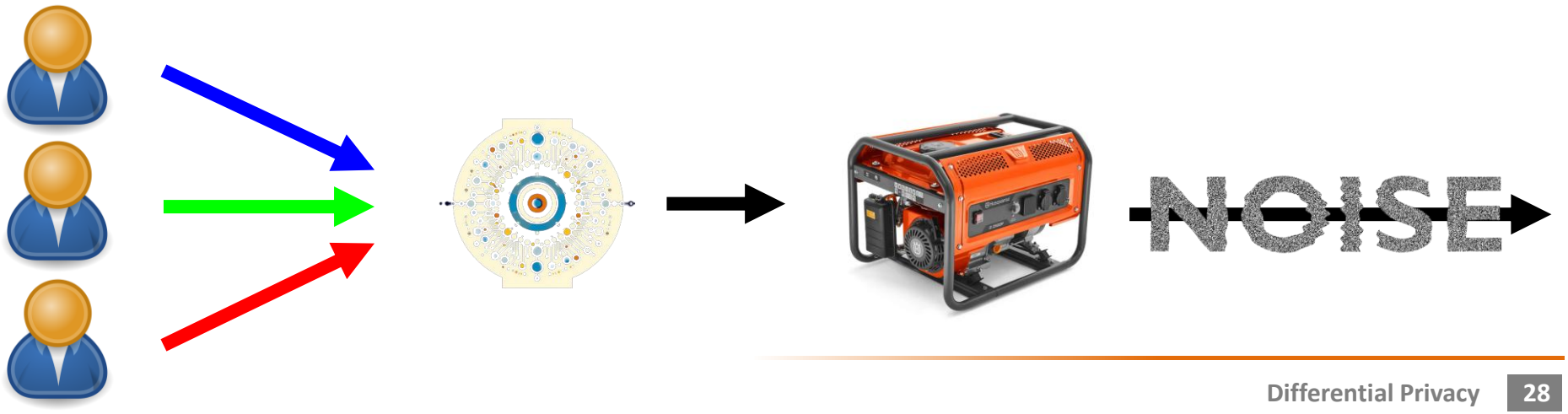
- The server itself is also untrusted.
- Any data communicated by the users must already be private, so even a prying server cannot learn much about any single user.

- A more stringent privacy model than the central DP.
- The utility one can obtain from the same amount of data is typically smaller than in the central model.

Shuffle Model

$$\frac{dy}{dx}$$

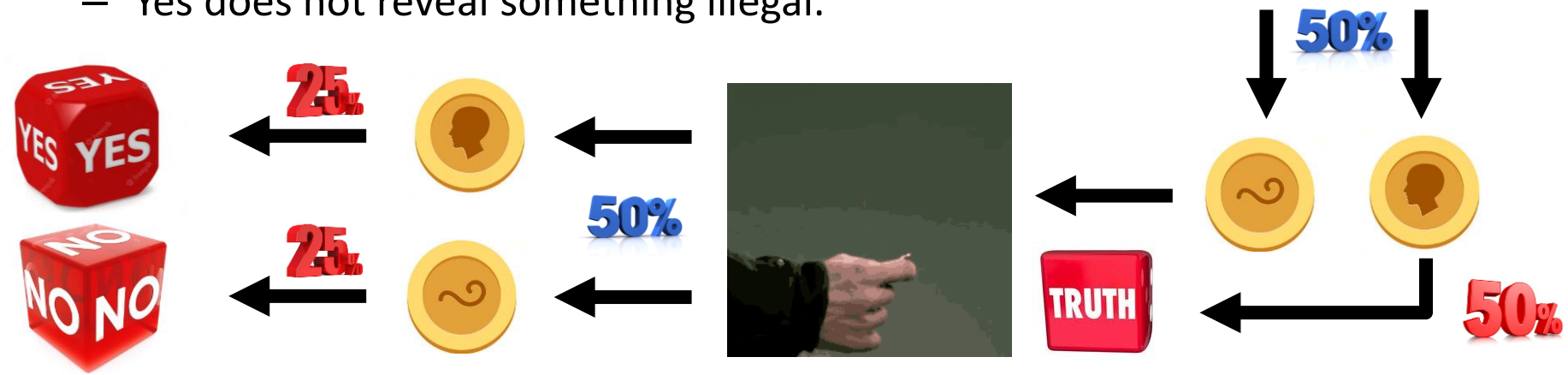
- Intermediate between the central and local models.
- Users do not trust the server but trust a black-box in the middle.
  - Mixer: when all users send their data to the untrusted server, this box-in-the-middle randomly permutes all the data points, so that the server had no idea who sent which part of the data.
  - Secure Aggregation (or in general SMPC) provides an even stronger solution by hiding the individual data.
- The goal is to try and provide stronger privacy than in the central model, while suffering a smaller utility loss than in the local model.



Random Response

$$\frac{dy}{dx}$$

- Survey to know how many people are illegal drug users.
  - Naively ask people, many will lie.
- Instead, each of them will flip a coin, without showing it to you.
  - On heads, the participant tells the truth (Yes or No).
  - On tails, they flip a second (secret) coin on which their answer is based on (heads – Yes, tail – No).
- Plausible Deniability: allows the subject to credibly say that they did not make a statement.
  - Yes does not reveal something illegal.



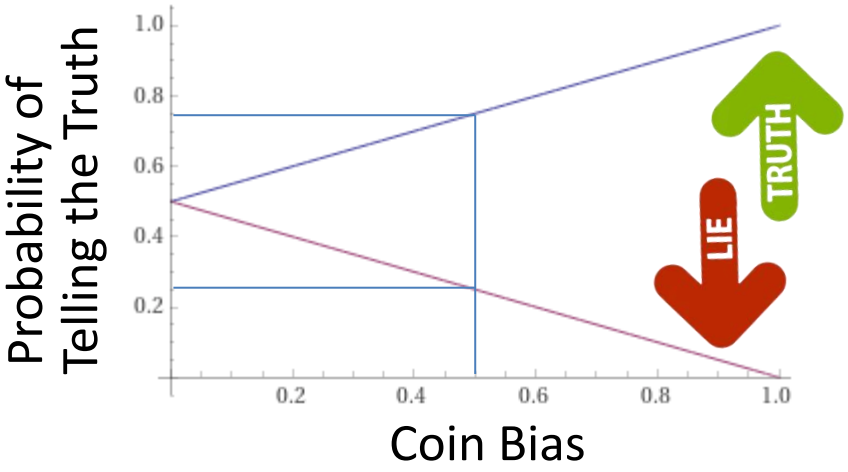
Math

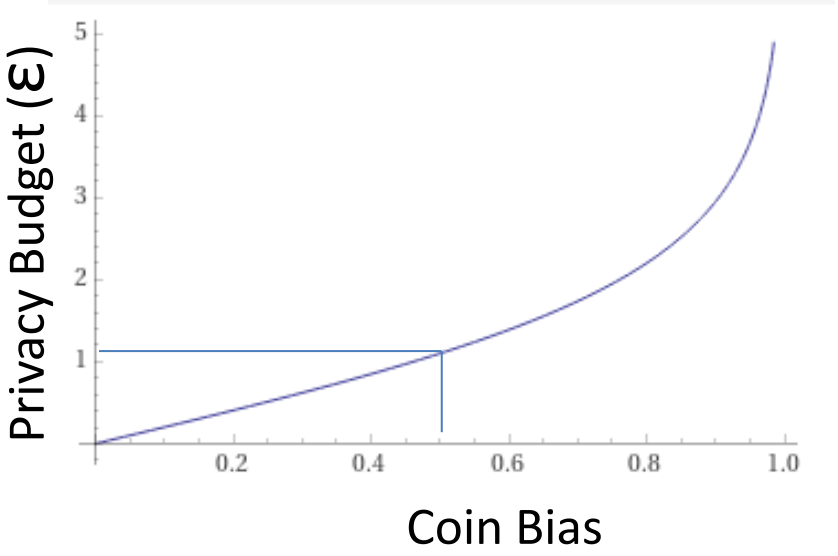
$$\frac{dy}{dx}$$

- $\forall$ O: $\Pr[M(D) = O] \leq e^{\varepsilon} \cdot \Pr[M(D') = O]$
  - 0.25 = $\Pr[M(\checkmark) = \oslash] \leq e^{\varepsilon} \cdot \Pr[M(\oslash) = \oslash] = 0.75$
    → Holds even with $\varepsilon = 0$.
  - 0.75 = $\Pr[M(\checkmark) = \checkmark] \leq e^{\varepsilon} \cdot \Pr[M(\oslash) = \checkmark] = 0.25$
    → Holds only if $\varepsilon \geq \ln(3) \approx 1.1$.

$\Pr[M(\oslash) = ?] \leq e^{\varepsilon} \cdot \Pr[M(\checkmark) = ?]$

- What if the first coin is biased?

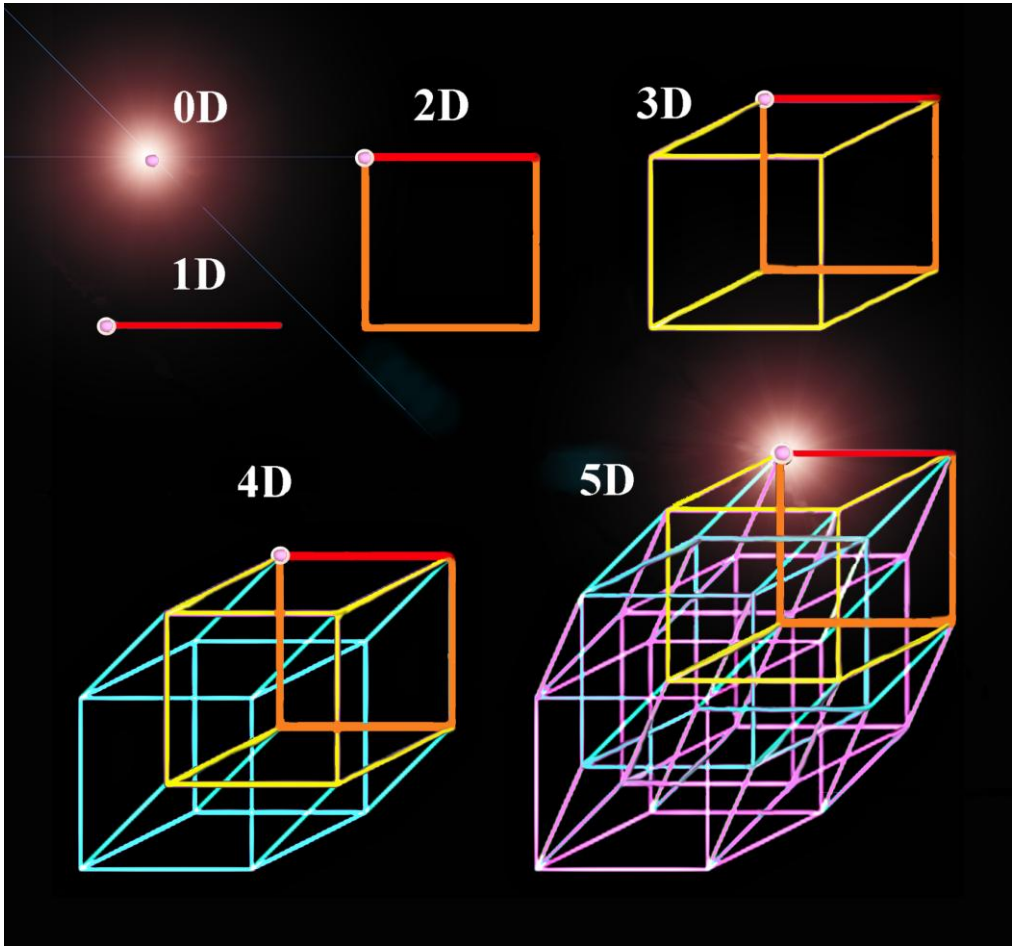




DeBiasing

$$\frac{dy}{dx}$$

- With a differentially private process like random response mechanism, some noise gets into the data.
- With enough answers, with high probability, the noise will cancel itself out.
- Suppose you have 1000 answers in total: 400 Yes and 600 No.
- About 50% of all 1000 answers are random, so you can remove 250 answers from each count.
- (400-250=) 150 Yes and (600-250=) 350 No out of 500 non-random answers means 30-70% ratio (instead of 40-60%).



$$\frac{dy}{dx}$$

Dimensions

Reformulate DP

$$\frac{dy}{dx}$$

- For all D and D' differing in a single element and for all output O:
  - $\Pr[M(D) = O] \leq e^{\varepsilon} \cdot \Pr[M(D') = O]$

An attacker with **perfect knowledge (B)** and **unbounded computation power (C)** is **unable (R)** to **distinguish (D)** whether someone is **in the data (N)**, **uniformly (V)** across users, even in the **worst-case scenario (Q)**.

(F) Change in Formalism

$$\frac{dy}{dx}$$

- Comparing distributions is not the only way to capture the idea that an attacker should not be able to gain too much information.

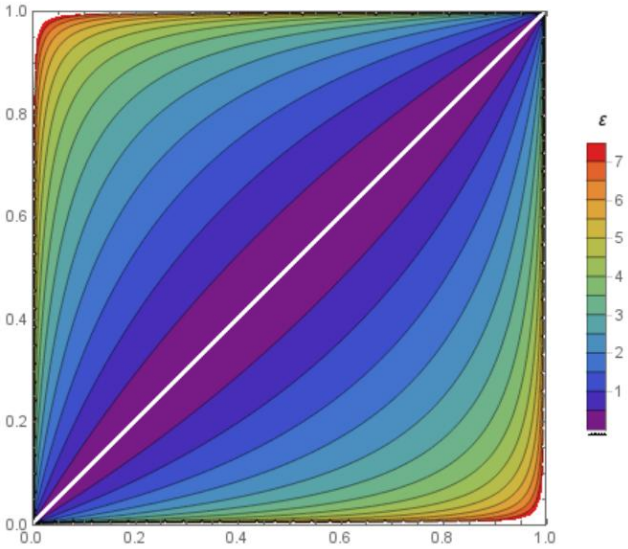
- Hypothesis Testing
  - An adversary wants to know whether the output O of a mechanism originates from D (the null hypothesis) or D' (the alternative hypothesis).
- The probability of false alarm (P_{FA}, type I error), when the null hypothesis is true but rejected.
- The probability of missed detection (P_{MD}, type II error), when the null hypothesis is false but retained.

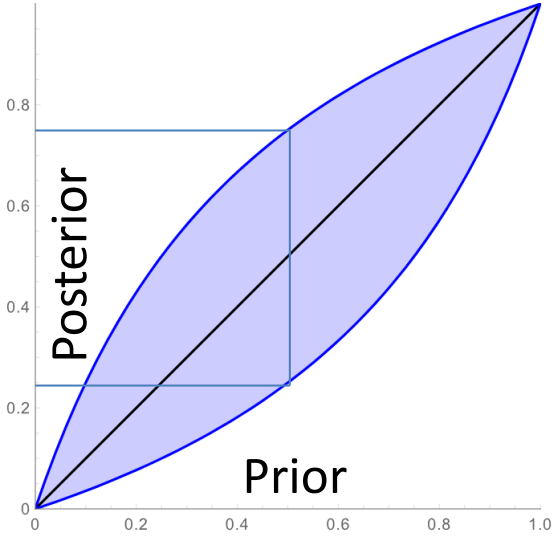
$$\begin{aligned}
\varepsilon\text{-DP} &\Leftrightarrow \begin{matrix}\mathbb{P}_{FA} + e^{\varepsilon}\mathbb{P}_{MD} \geq 1 \\ e^{\varepsilon}\mathbb{P}_{FA} + \mathbb{P}_{MD} \geq 1\end{matrix} \quad \text{for all } S \subseteq \mathcal{O} \\
\varepsilon\text{-DP} &\Leftrightarrow \mathbb{P}_{FA} + \mathbb{P}_{MD} \geq \frac{2}{1+e^{\varepsilon}} \\
(\varepsilon, \delta)\text{-DP} &\Leftrightarrow (\mathbb{P}_{FA}, \mathbb{P}_{MD}) = \begin{matrix}\{(\alpha, \beta) \in [0,1] \times [0,1] : \\ (1 - \alpha \leq e^{\varepsilon}\beta + \delta)\}\end{matrix}
\end{aligned}$$

(F) Bayesian Interpretation

$$\frac{dy}{dx}$$

- Prior: Pr [D = D_{in}]
- Posterior: Pr [D = D_{in} | M(D) = O]
- ε = Ln(3) ≈ 1.1
- 50 % Prior → ± ≤ 25 % → 25 – 75 % Posterior

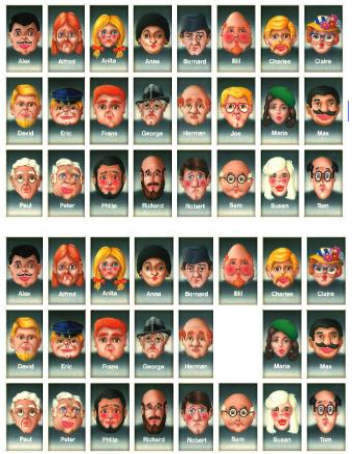




(N) Neighbourhood Definition

$$\frac{dy}{dx}$$

- For all D and D' differing in a single element and for all output O:
  - $\Pr[\, M(D) = O \,] \leq e^{\varepsilon} \cdot \Pr[\, M(D') = O \,]$



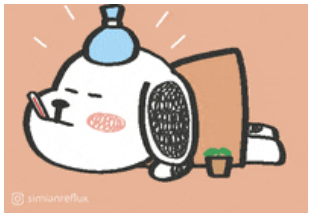
- DP considers datasets differing in one record.
  - The same size and differ only on one record.
  - One is a copy of the other with one extra record.
- These two options do not protect the same thing: the former protects the value (called Bounded DP), the latter the existence (called Unbounded DP).

- It is possible to protect any property which is deemed sensitive.
  - The difference is in the protected class or minority group.

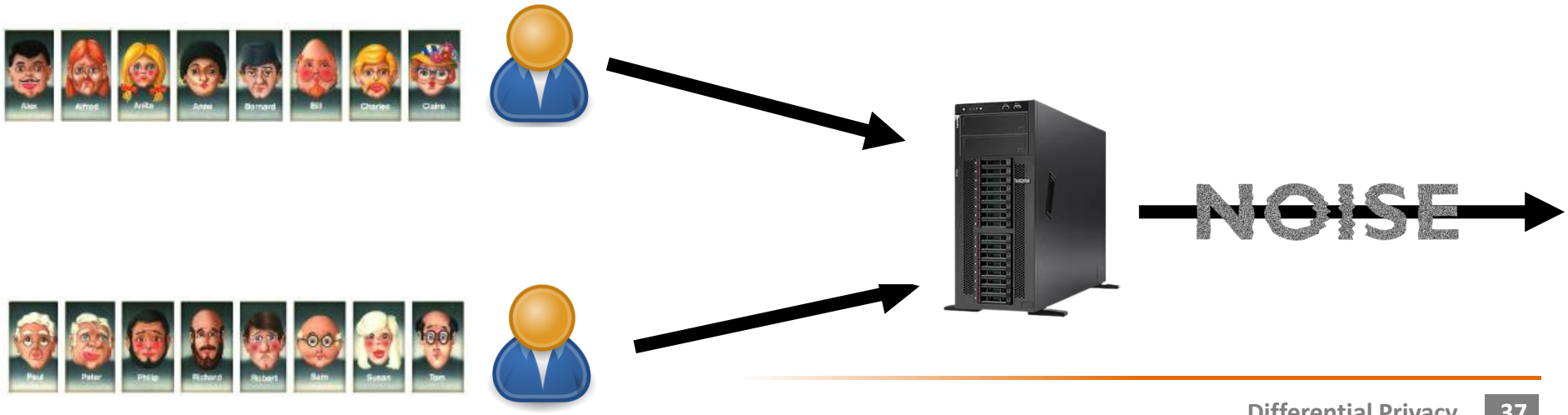
(N) Variants & Extensions

$$\frac{dy}{dx}$$

- For all D and D' differing in a single element, and for all output O:
  - $\Pr[M(D) = O] \leq e^{\varepsilon} \cdot \Pr[M(D') = O]$
- User-level / Group-level / Client-level
  - Consider datasets differing in more record.



- Individual DP: Only consider one (the actual) dataset.
- Asymmetric DP: Learns that the target does has cancer is worst than learning it does not.



(V) Variance of the Privacy Loss

$$\frac{dy}{dx}$$

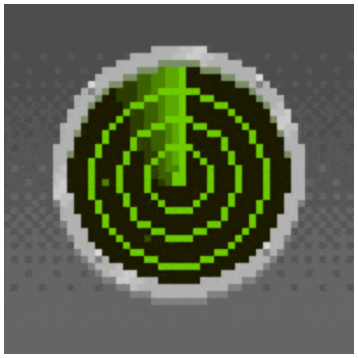
- For all D and D' differing in a single element, and for all output O:
  - $\Pr[M(D) = O] \leq e^{\varepsilon} \cdot \Pr[M(D') = O]$
- The privacy parameter ε is uniform: the level of protection is the same for all protected users or attributes.
- In practice, some users might require a higher level of protection.
- Personalized DP: varying the privacy level across inputs.
- Random DP: randomizing the variation of privacy levels.
  - Instead of requiring DP to hold for any possible datasets, it is natural to only consider realistic datasets, and allow unrealistic datasets to not be protected.

(N&V) Multi Dimension

$$\frac{dy}{dx}$$

- Dimensions are independent, instances from different dimensions could be combined.
- N + V: both personalization and neighborhood can be naturally captured together via distance functions.

- For all D and D', and for all output O:
  - $\Pr[M(D) = O] \leq e^{\varepsilon \cdot d(D,D')} \cdot \Pr[M(D') = O]$



- The closer D and D' is, the stronger the protection.
  - Applicable for instance for location data.
- The neighborhood relationship between two datasets is not binary but quantified with a distance metric.
- If d() is the Hamiltonian difference, D-Privacy is equivalent with ε-DP.

Comparison

$$\frac{dy}{dx}$$

- There is no ultimately privacy definition.
  - One cannot say that any of the notions is better than the others.
- They all aim at modeling different scenarios and provide mostly incomparable guarantees.
- Depending on your situation, pick the one that fits best.

- The same is true for the privacy model, e.g., local vs global.

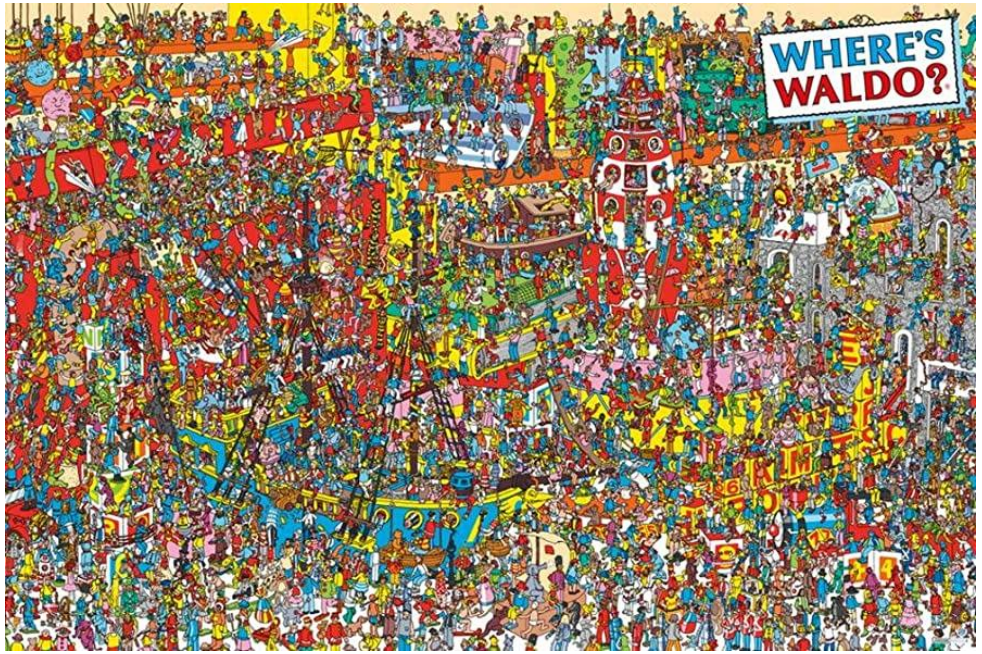
$$\frac{dy}{dx}$$

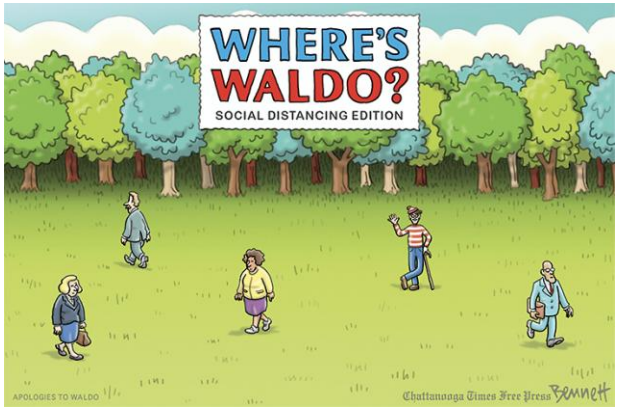
Real World Deployments

Robust vs Sensitive

$$\frac{dy}{dx}$$

- Robust if
  - … estimating large population sizes.
  - … understanding correlations between features in a large dataset.
  - … producing usage metrics for a service with many users.
  - … computing statistics over large groups.
- Sensitive
  - Small populations
  - Finding outlier individuals
  - Preserving linkability





Apple & Google

$$\frac{dy}{dx}$$

- QuickType suggestions learns previously-unknown words typed by sufficiently many users (ε=16).
- Emoji suggestions calculates which emojis are most popular among users (ε=4).
- Health Type Usage estimates which health types are most used in the HealthKit app (ε=2).

- Next-word prediction model on Gboard for the Spanish-language version ($\varepsilon=6.92, \delta=10^{-5}$).
- Google shared mobility data with researchers ($\varepsilon=0.66, \delta=2.1\cdot10^{-29}$).
- Community Mobility Reports quantify changes in mobility patterns during the COVID-19 pandemic (ε=2.64).

Facebook & LinkedIn

$$\frac{dy}{dx}$$

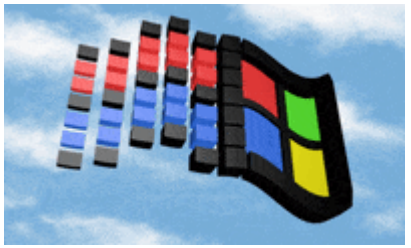
- The Full URLs Data Set provides data on user interactions with web pages shared on Facebok (99% of users are protected with $\varepsilon=0.45$, $\delta=10^{-5}$).
- The Movement Range Maps quantify the changes in mobility of Facebook users during the COVID-19 pandemic ($\varepsilon=2$).

- The Audience Engagements API is an interactive query system allowing marketers to get information about users engaging with their content ($\varepsilon= 34.9$, $\delta=10^{-9}$).
- The Labor Market Insights measure trends in people changing their occupation ($\varepsilon=28.8$, $\delta=2.4\cdot10^{-9}$ for hiring events, and $\varepsilon=0.3$, $\delta=3\cdot10^{-10}$ for skill information).

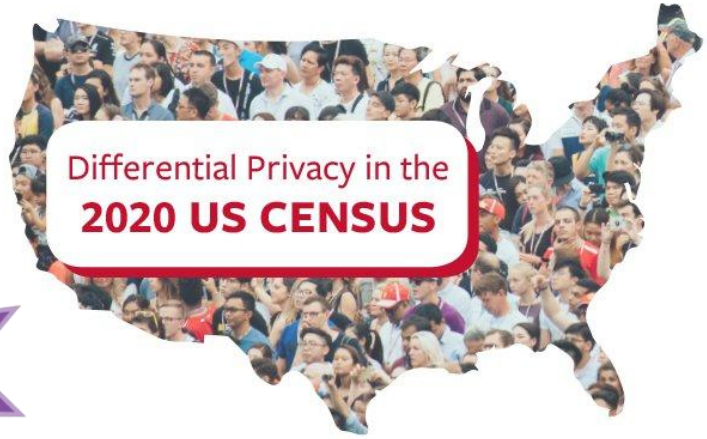
Microsoft & US Census Bureau

$$\frac{dy}{dx}$$

- The Global Victim-Perpetrator Synthetic Dataset provides information about victims and perpetrators of trafficking ($\varepsilon=12$, $\delta=5.8\cdot10^{-6}$).



- Microsoft collects telemetry data in Windows, e.g., how much time users spend using particular apps ($\varepsilon=1.672$).
- The U.S. Broadband Coverage Dataset quantifies the percentage of users having access to high-speed Internet across the US ($\varepsilon=0.2$).
- The 2020 Census Redistricting Data contain US population data and demographic information ($\varepsilon=13.64$, $\delta=10^{-5}$).



DP 3

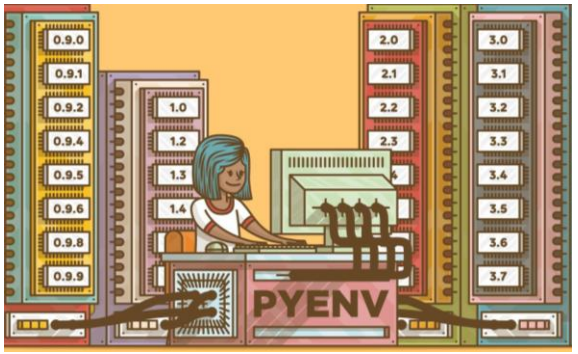
Example

Take Away

$$\frac{dy}{dx}$$

- Absolute privacy breaches are inevitable, DP focuses on the individual (e.g., to counter Membership Inference Attacks).
- Compositions (and the privacy axioms) are crucial properties of DP.
- To achieve DP, uncertainty must be added (e.g., in the form of Laplacian or Gaussian noise).
- Global (aka central) and local settings provide different kind of protection.

- There are many variants and extensions of DP, one must be careful in selecting the right one for a particular use-case.
- DP is not just theory anymore, many real-life deployment exists.



Control Questions

$$\frac{dy}{dx}$$

- Enlist three dimensions of how Differential Privacy can be changed, and give motivations and examples for them!
- Give a high-level explanation about the Gauss and Laplace mechanisms, detail their pros and cons, and explain what sensitivity means in this context!
- What are the local and the global models for Differential Privacy and how do they differ?

References

$$\frac{dy}{dx}$$

- Ted's Blog
- TED talk
- Guide to Differential Privacy
- DP: What is all the noise about?
- DP Blog
- Video1 & Video2
- Blog1 & Blog2 & Blog3
- Census