CSCI 699: Privacy Preserving Machine Learning - Week 2

Differential Privacy

Sai Praneeth Karimireddy, Sep 6 2024

Quantifying Privacy Leakage

Last week recap

- We saw many definitions of privacy
    - De-identification / suppression
    - K-anonymity
    - L-diversity
- We saw none of them really protected privacy and were easily broken
- Hinted at a more widely accepted definition.

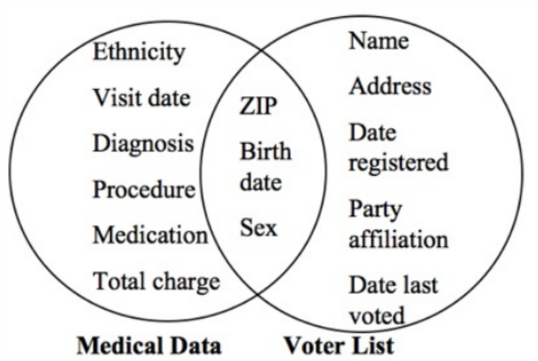
Last week takeaways

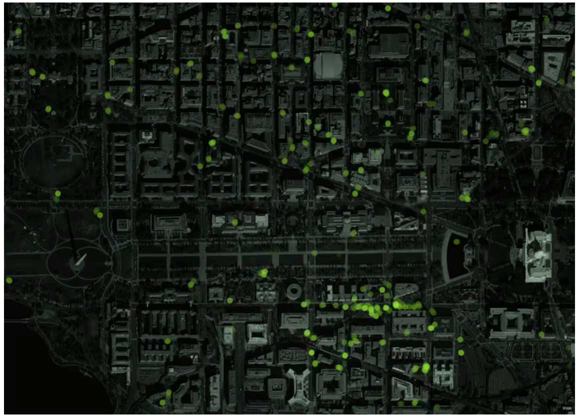
Requirements for privacy definition

- Unaffected by auxiliary information: we should not be able to combine extra data to undo privacy.

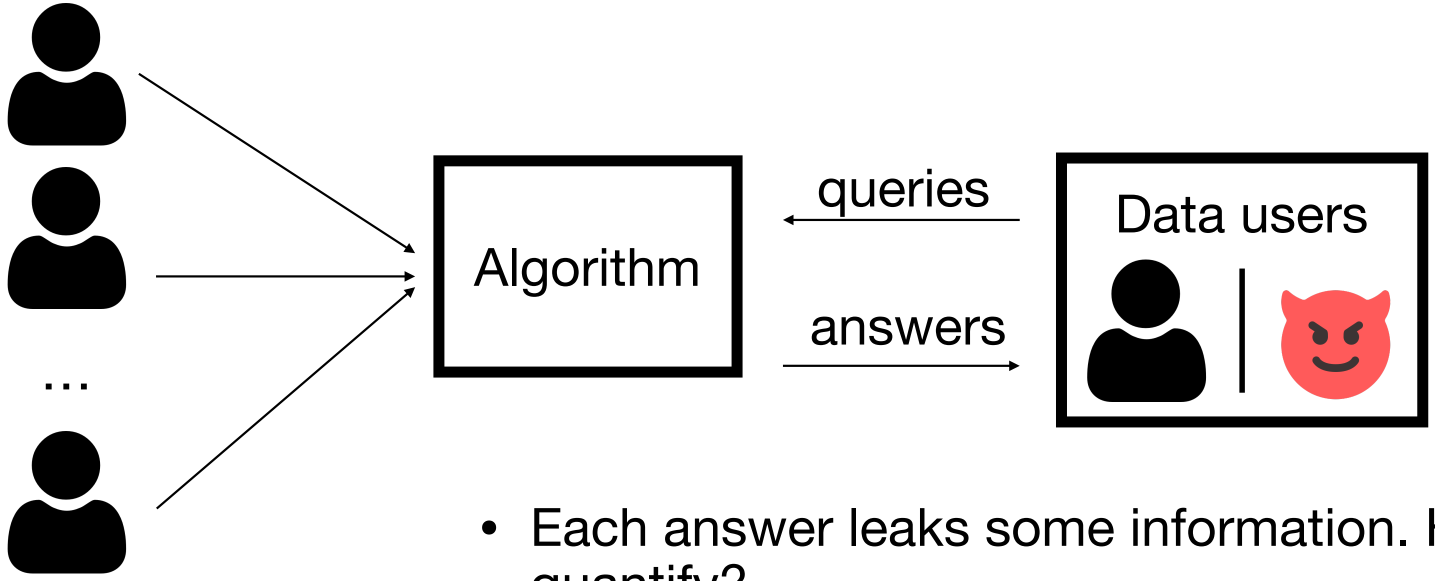
- Composition: We should understand what happens when data is continuously released.

- Today we will come with such a privacy definition.





Goals of PPML



- Each answer leaks some information. How to quantify?
- How to balance usefulness of answers vs. privacy being leaked?

Quantifying Privacy Leakage

Attempt 1

Absolute Privacy: quantify total information leaked

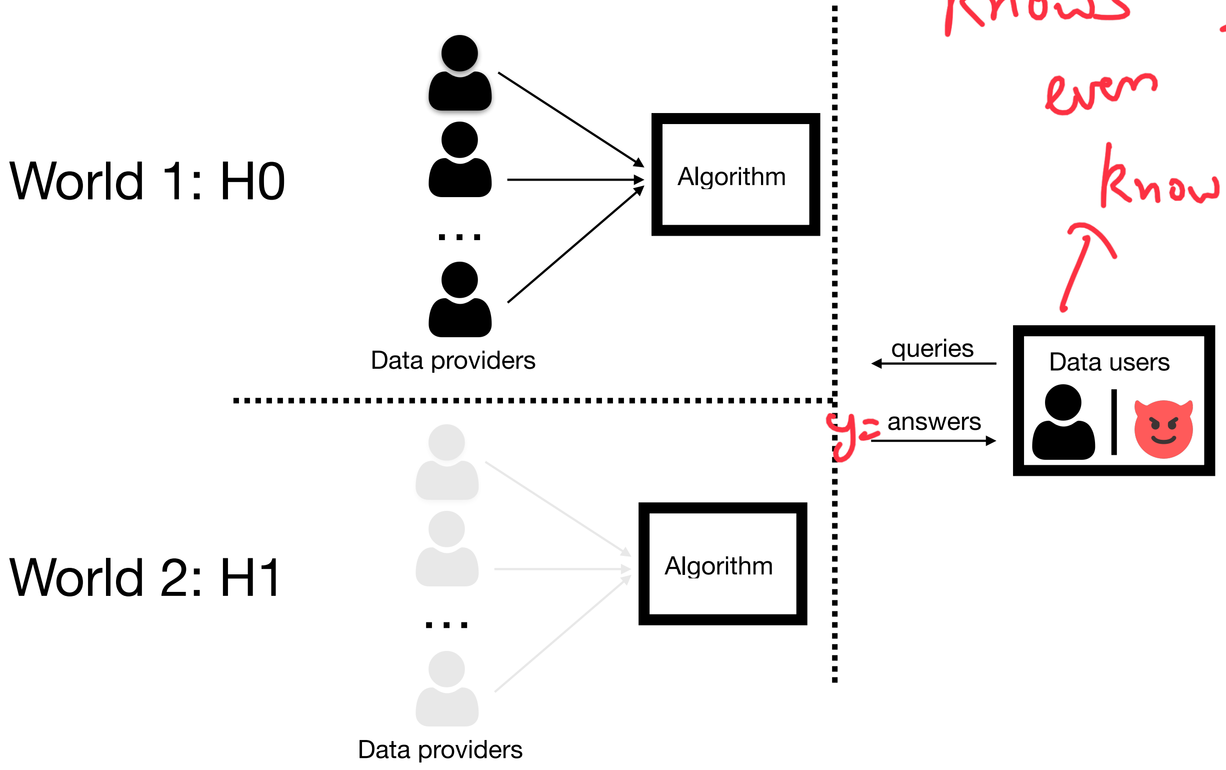
"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- Bayesian version: the posterior and prior are identical

Quantifying Privacy Leakage

Attempt 1

Knows everything even D! Only doesn't know if y from H0 or H1



We are either in world 1 or world 2. The adv cannot tell which world we are in.

Quantifying Privacy Leakage

Attempt 1

Absolute Privacy: quantify total information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 1**: Impossible to reveal anything useful about data since any useful answer will provide some previously unknown information.

Quantifying Privacy Leakage

Attempt 1: Problems

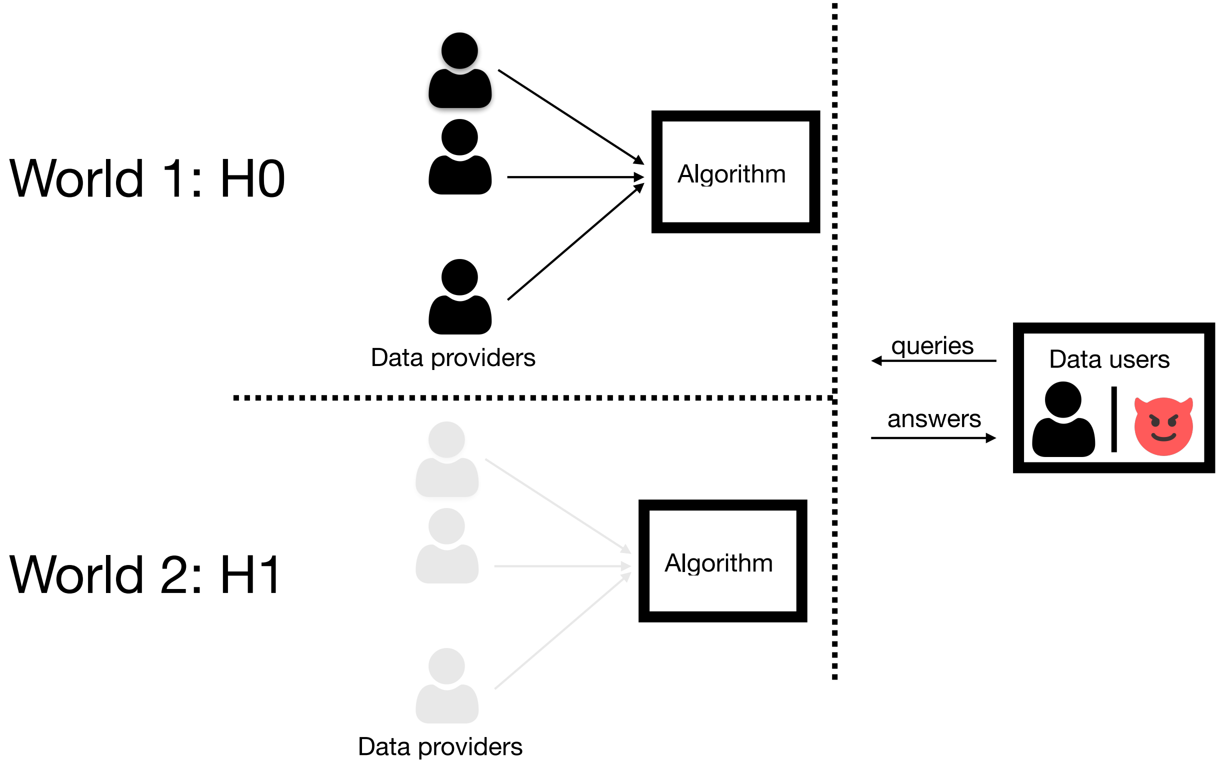
Absolute Privacy: quantify total information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 2**: What I know before changes with auxiliary information.
- Did the model leak information about Bob?
    - Bob is a smoker, but his data was not used to train the model.
    - The model said smokers have higher risk of disease.
    - Bob's insurance premiums were raised.

Quantifying Privacy Leakage

Attempt 1: Problems



Any information about the distribution reveals which world we are in.

Quantifying Privacy Leakage

Attempt 1: Problems

Absolute Privacy: quantify **total** information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 2**: What I know before changes with auxiliary information.
- We want to safeguard individual information (privacy) while revealing distributional/aggregate information (utility)

Quantifying Privacy Leakage

Attempt 2

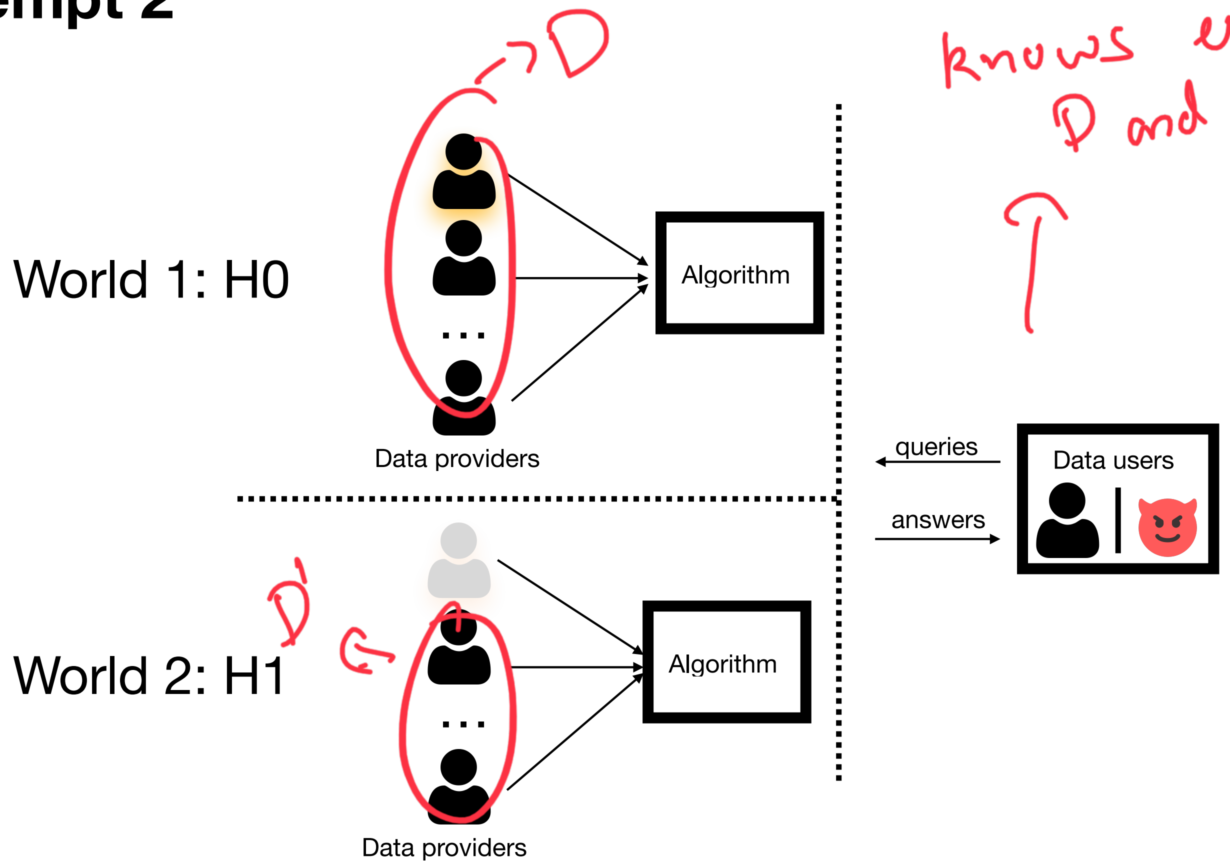
Relative Privacy: quantify **new** information leaked

"An analysis of a dataset is private if what can be learned about an individual in the dataset is not much more than what would be learned if the same analysis was conducted without them in the dataset"

Quantifying Privacy Leakage

Attempt 2

knows everything including D and D'. Only doesn't know H0 or H1.



- In world 2 only Bob is removed/ replaced.
- Now from the answer, how easily can guess the correct world?

Quantifying Privacy Leakage

Attempt 2

Relative Privacy: quantify new information leaked

"An analysis of a dataset is private if what can be learned about an individual in the dataset is not much more than what would be learned if the same analysis was conducted without them in the dataset"

- **Intuition**: Whether Bob is present in the data or not, the answer should not change much.
- Then, from looking at the answer, we will not learn whether Bob was present in the data or not.
- Gives Bob plausible deniability.

Aside: how is Putin's popularity calculated?

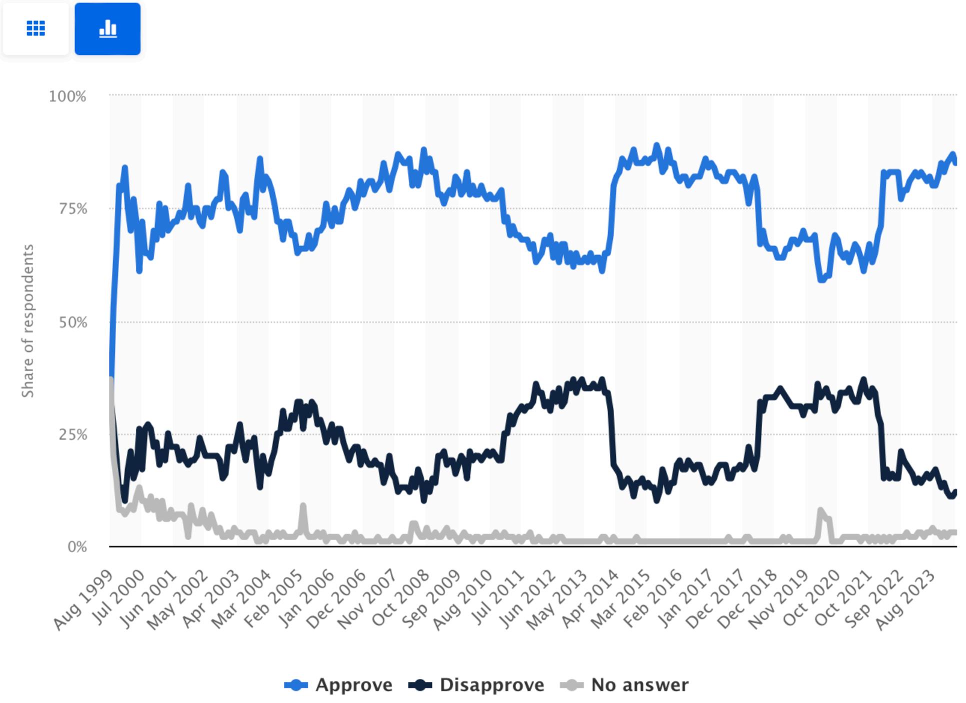
Plausible deniability as privacy

Poll: Russians Still Like Putin and Back the Ukraine War – but Are Anxious at Home

Most Russian survey respondents see the war in Ukraine as a broader conflict with the West and support it amid concerns about their own country's economy.

By Elliott Davis Jr. | Jan. 9, 2024

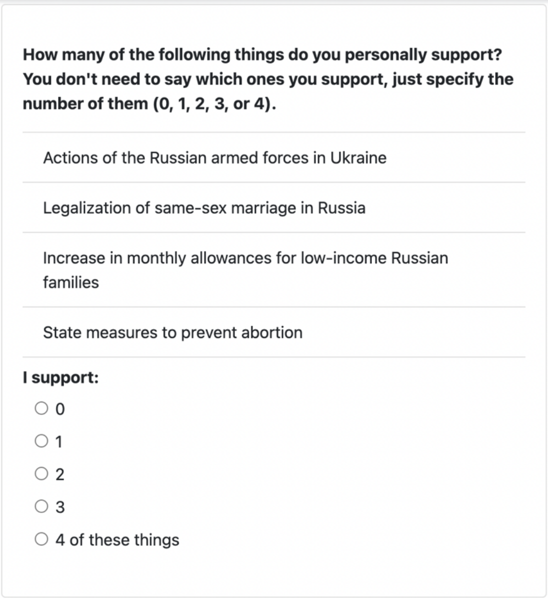
Do you approve of the activities of Vladimir Putin as the president (prime minister) of Russia?



Aside: how is Putin’s popularity calculated?

List Experiment

- Split users randomly into two groups
- Design a set of options very similar to the one you actually care about
- To control only ask about the rest. To the treatment include your option.
- Does this confer plausible deniability?



How many of the following things do you personally support? You don't need to say which ones you support, just specify the number of them (0, 1, 2, 3, or 4).

Actions of the Russian armed forces in Ukraine

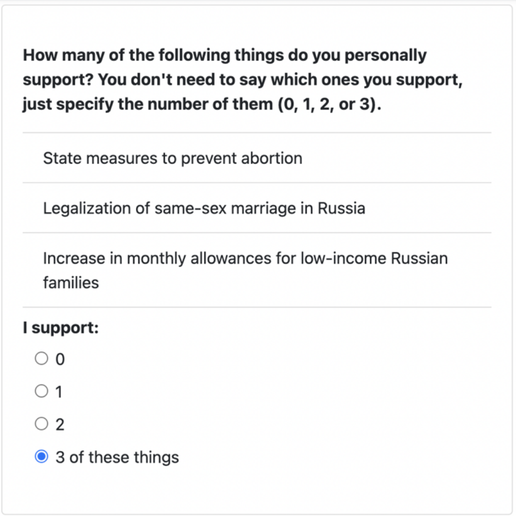
Legalization of same-sex marriage in Russia

Increase in monthly allowances for low-income Russian families

State measures to prevent abortion

I support:

0
1
2
3
4 of these things



How many of the following things do you personally support? You don't need to say which ones you support, just specify the number of them (0, 1, 2, or 3).

State measures to prevent abortion

Legalization of same-sex marriage in Russia

Increase in monthly allowances for low-income Russian families

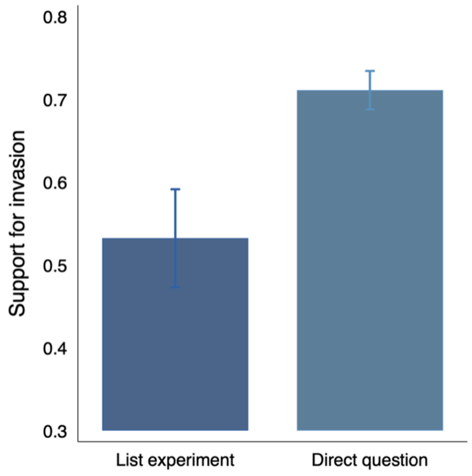
I support:

0
1
2
3 of these things

Chapkovski and Schaub 2022. “Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment” LSE Blog

Aside: how is Putin's popularity calculated?
List Experiment

Figure 2: Support for the Russian invasion of Ukraine



Note: Bars show averages, vertical lines show 95% confidence intervals.

Chapkovski and Schaub 2022. "Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment" LSE Blog

Quantifying Privacy Leakage

Attempt 2

Relative Privacy: quantify new information leaked

"An analysis of a dataset is private if what can be learned about an individual in the dataset is not much more than what would be learned if the same analysis was conducted without them in the dataset"

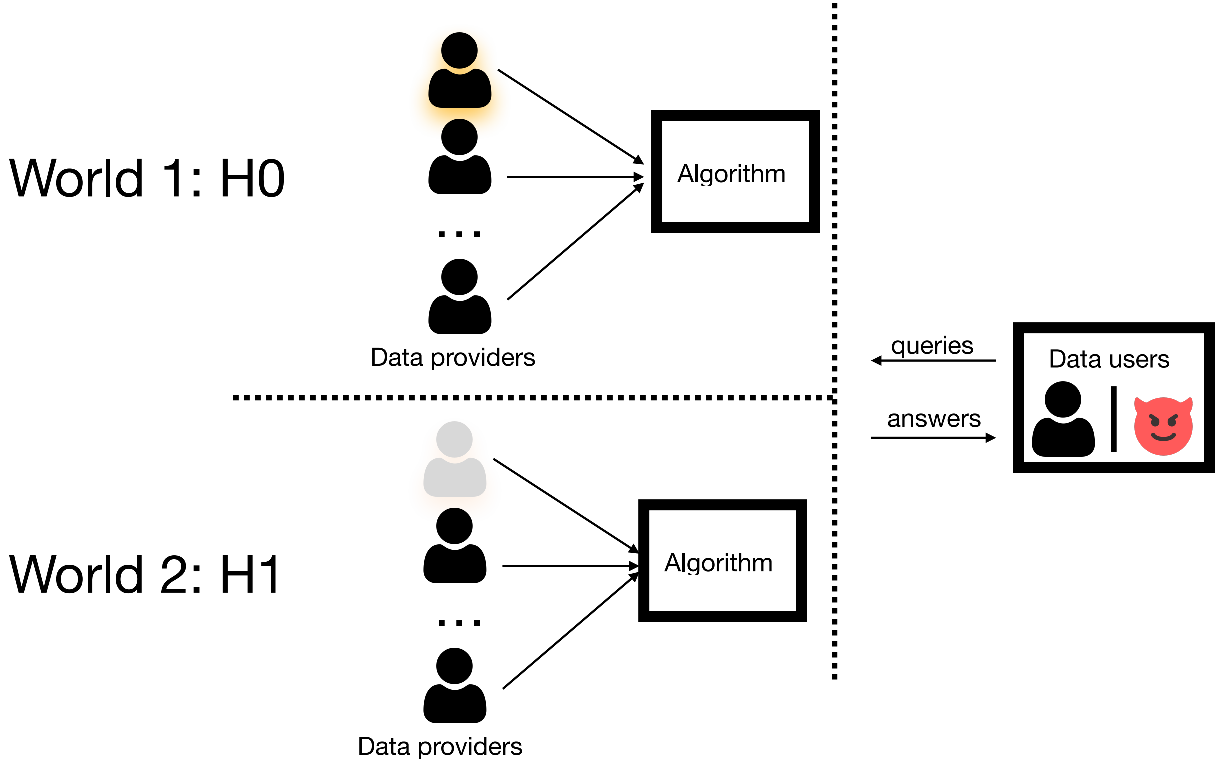
- **Question**: Can a deterministic algorithm be private?
- What if Bob is the only data point? Then can easily reverse-engineer Bob's data.

$$\min_x \ell(f(x), y)$$

- Only randomized algorithms can be private.

Membership Inference

As a definition of privacy

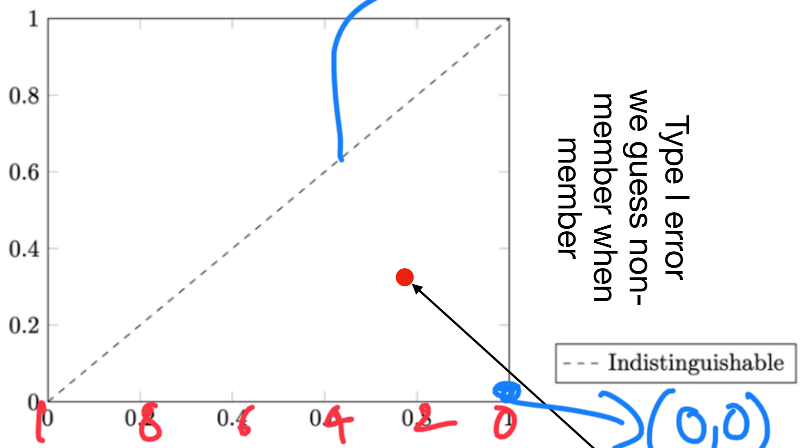


- In world 2 only Bob is removed/ replaced.
- Now from the answer, how easily can guess the correct world?
- Can have false positives, false negatives

Membership Inference

Quantifying attack success

random guess ⇒ perfect privacy



Type II error
we guess member when non-member

We get a point.

⇒ (0,0) perfect accuracy

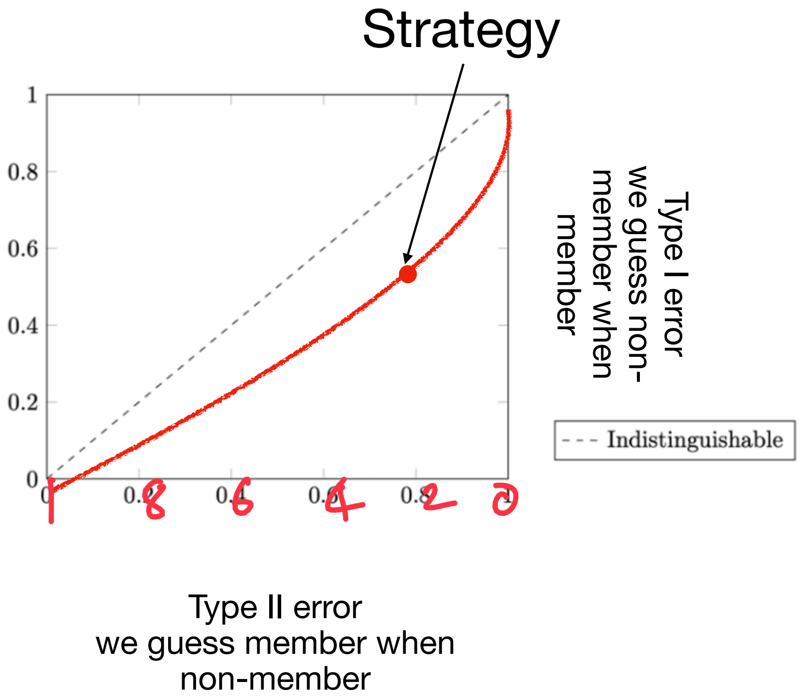
⇓

full leakage

- Suppose we run multiple runs
- Count the number of times the adv guesses H0 vs H1 correctly
- We can compute Type I and Type II errors.

Membership Inference

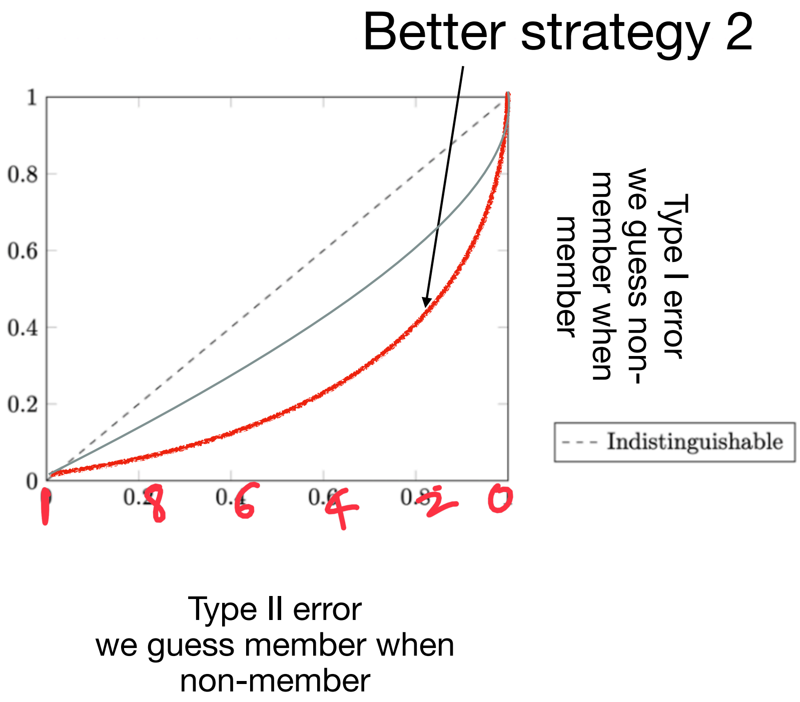
Tradeoff curve



- But sometimes we care asymmetrically
- E.g. its important not to miss anyone e.g. sending cat ads to pet owners
- Not ok if we are accusing them of a crime

Membership Inference

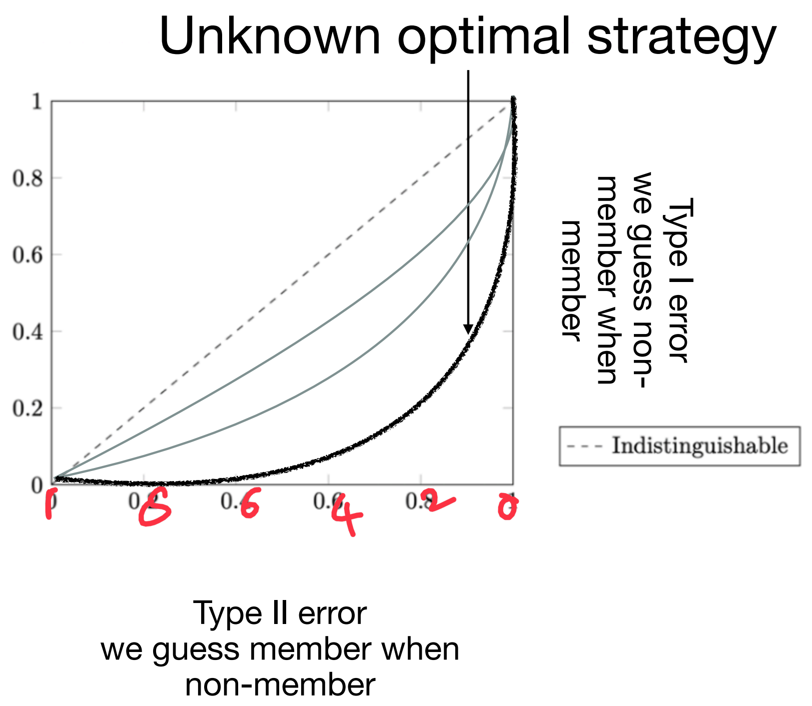
Comparing tradeoff curves



- Strategy 1 is better than Strategy 2 if the curve is uniformly above.
- Lower curve means we've found more privacy leakage

Membership Inference

Optimal tradeoff curve

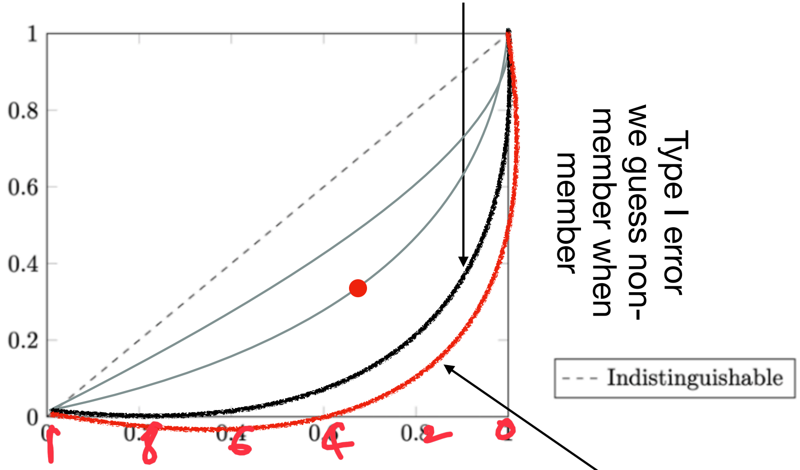


- There is an optimal strategy
- use this to quantify privacy leakage
- What if no single strategy is best?
- **Neyman–Pearson lemma** guarantees existence of uniformly most powerful test.

Membership Inference

Privacy from tradeoff curve

Unknown optimal strategy



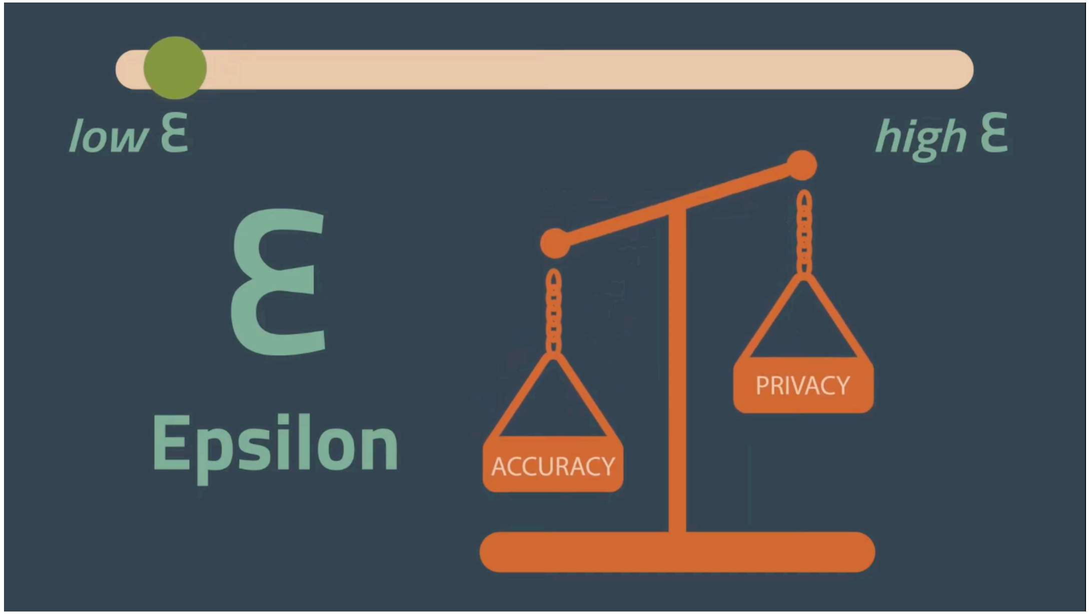
Type I error
we guess non-member when member

lower bound from theory

Type II error
we guess member when non-member

- Use optimal strategy to quantify privacy
- But empirical tests only give an upper-bound
- Need theory to give lower-bound

Differential Privacy



Differential Privacy

Calibrating Noise to Sensitivity in Private Data Analysis

2006

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3*]

2017 Gödel Prize

Differential privacy is a powerful theoretical model for dealing with the privacy of statistical data. The intellectual impact of differential privacy has been broad, influencing thinking about privacy across many disciplines. The work of Cynthia Dwork (Harvard University), Frank McSherry (independent researcher), Kobbi Nissim (Harvard University), and Adam Smith (Harvard University) launched a new line of theoretical research aimed at understanding the possibilities and limitations of differentially private algorithms. Deep connections have been exposed in other areas of theory (including learning, cryptography, discrepancy, and geometry) and have created new insights affecting multiple communities.

Differential Privacy

Threat model

- Let χ be a the domain of training data
- A dataset $D \in \chi^n$ is a multiset of n records/rows of χ
- D (sensitive data) $\longrightarrow$ algorithm $\longrightarrow$ Y (answers)
- Attacker wants to infer some information about $D \in \chi^n$
    - observes Y
    - knows algorithm, domain χ, and potentially more prior information
    - cannot control what attacker knows

Differential Privacy

Threat model

- Attacker wants to infer some information about $D \in \chi^n$
  - observes Y, knows algorithm, domain χ, and prior information.
  - can compute likelihood of dataset:

$$Pr[D|Y] = \frac{\overset{\text{algorithm}}{Pr[Y|D]} \cdot \overset{\text{prior knowledge}}{Pr[D]}}{Pr[Y]}$$

Differential Privacy

Performing membership inference

- Attacker wants to infer presence of $x \in X$?
  - observes Y, knows algorithm, domain χ, and even $D \backslash x \in \chi^{n-1}$
  - can compute likelihood of x in dataset

$$Pr[x' \mid Y] = \frac{\overbrace{Pr[Y \mid x']}^{\text{algorithm}} \cdot \overbrace{Pr[x']}^{\text{prior knowledge}}}{Pr[Y]}$$

Differential Privacy

Performing membership inference

- Attacker wants to infer presence of $x \in X$?
  - can compute likelihood of x in dataset

algorithm prior knowledge

$$Pr[x' | Y] = \frac{Pr[Y | x'] \cdot Pr[x']}{Pr[Y]}$$

  - Can even recover x using max-likelihood

$$\hat{x} = \text{arg} \max_{x'} Pr[Y | x'] Pr[x']$$

Differential Privacy

Goal

- Attacker wants to infer some information about $D \in \chi^n$
    - can compute likelihood of seeing some dataset

$$Pr[D \mid Y] = \frac{\underbrace{Pr[Y \mid D]}_{\text{algorithm}} \cdot \underbrace{Pr[D]}_{\text{prior knowledge}}}{Pr[Y]}$$

- We design a private algorithm by controlling $Pr[Y \mid D]$

Differential Privacy

Strict definition

- Perfect relative indistinguishability: For all inputs, the output probability is the same.

$$\forall D, D', y : \quad \Pr[Y = y \,|\, \mathscr{D} = D] = \Pr[Y = y \,|\, \mathscr{D} = D']$$

- The mechanism does not leak any information about D

- However, achieving it is very hard, does not allow any information about D.

Differential Privacy

A better definition

- Some indistinguishability: For all similar inputs, the output probabilities are bounded.

$$\forall y, \forall \text{ similar } D, D' : \quad \frac{\Pr[Y = y \,|\, \mathscr{D} = D]}{\Pr[Y = y \,|\, \mathscr{D} = D']} \leq \text{ constant}$$

- It means by observing any Y, adversary is NOT able to distinguish between inputs x and x' beyond a bounded certainty.
- What does similar inputs mean?
  - Depends on use case
  - location positions that are within some range
  - datasets that differ in one individual row

Differential Privacy

Formal definition

ε-Differential Privacy:

An algorithm A satisfies ε-DP if for any similar datasets $D, D' \in \chi^n$ and $y \in \mathcal{Y}$

$$\frac{Pr[Y = y \,|\, D]}{Pr[Y = y \,|\, D']} \leq \exp(\varepsilon)$$

- Recall that D (sensitive data) $\longrightarrow$ algorithm $\longrightarrow$ Y (answers)
- So we have, $Pr[Y \,|\, D] = Pr[A(D) = Y]$

Differential Privacy

Formal definition

ε-Differential Privacy:

An algorithm A satisfies ε-DP if for any similar datasets $D, D' \in \chi^n$ and $y \in \mathcal{Y}$

$$\frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \exp(\varepsilon)$$

- $\varepsilon = 0$ means perfect privacy
- $\varepsilon \gg 0$ means not private

Differential Privacy

Source of randomness

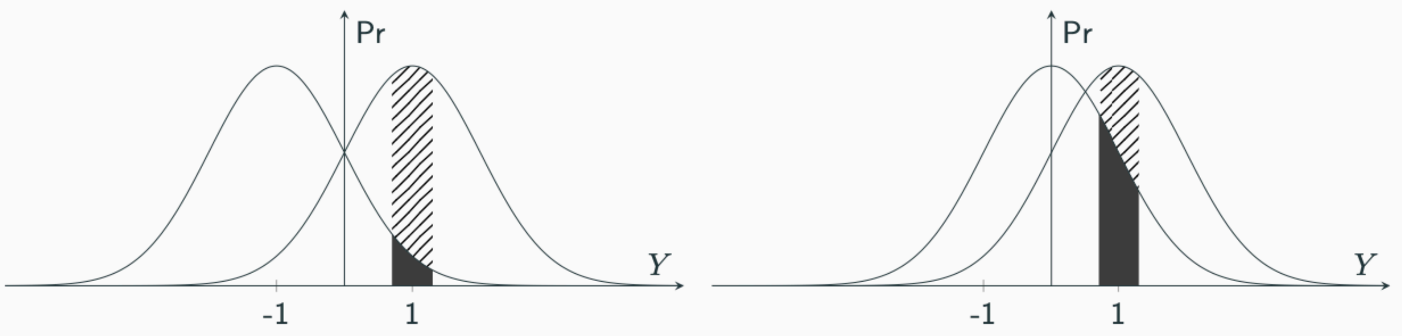
$$\forall y, \forall \text{ similar } D, D' : \quad \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \exp(\varepsilon)$$

- In $\Pr[A(D) = y]$, over what randomness is the probability defined?
    - The randomness of the algorithm?
        - Yes
    - Randomness of the data $D \in \chi^n$?
        - No.
        - We look at all possible values of D, D' i.e. worst case

Differential Privacy

Visual representation

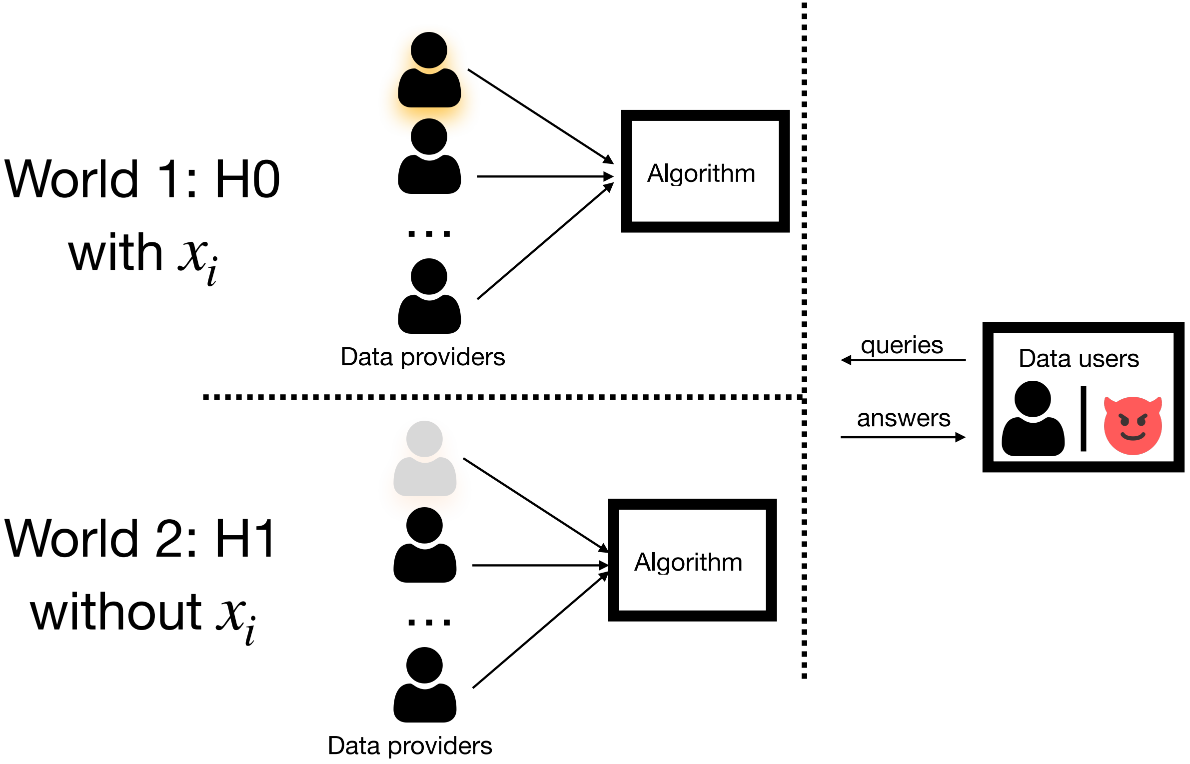
- Consider $D = \langle x_1, \cdots, x_i, \cdots x_n \rangle$, and a similar dataset $D' = \langle x_1, \cdots, \cancel{x_i}, \cdots x_n \rangle$
- ε-DP means $\dfrac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \exp(\varepsilon)$



(a) Large ϵ **(b)** Small ε

Differential Privacy

Recall Membership Inference



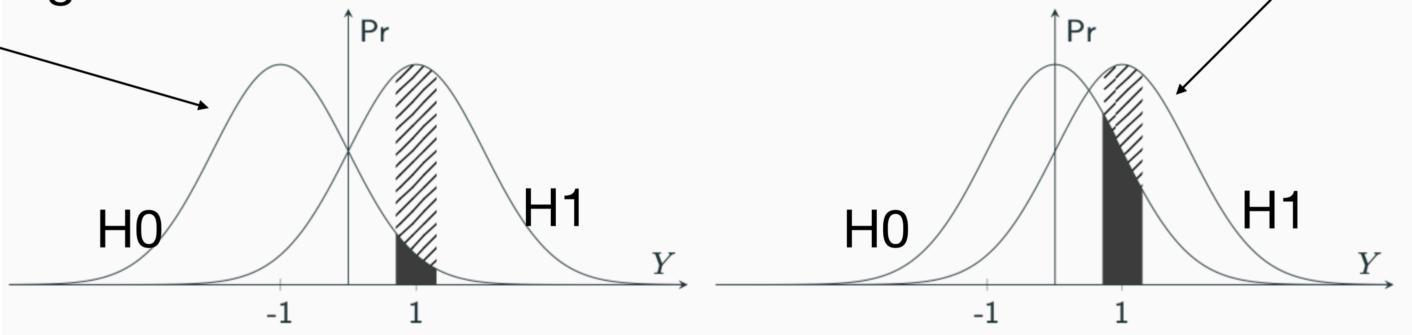
- We know everything about the algorithm and even $D \backslash x_i$
- We observe an output Y
- Need to guess if it came from H0 or H1

Differential Privacy

Connection to Membership Inference

Hard to distinguish

Easy to distinguish



(a) Large ϵ **(b)** Small ε

- We observe Y = 1.
- Can you guess H0 or H1?

Small $\varepsilon \Rightarrow$ higher error

Differential Privacy and membership inference

Quantifying connection

Can trade some Type I error for Type II, but sum of errors is lower-bounded.

Theorem

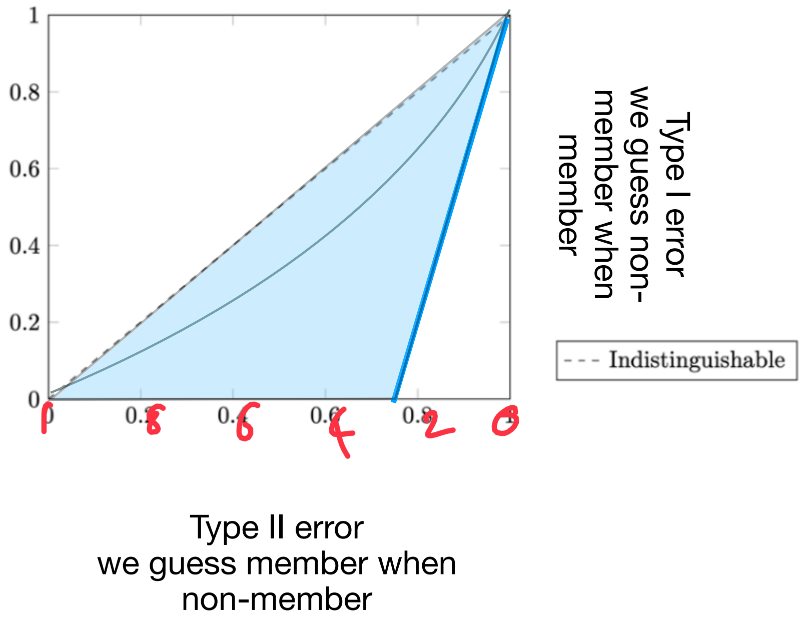
Suppose A satisfies ε-DP for datasets D, D' which differ by one datapoint. Then, we have

- $Pr[\text{guess H0} \,|\, H1] + e^{\varepsilon} Pr[\text{guess H1} \,|\, H0] \geq 1$
- $e^{\varepsilon} Pr[\text{guess H0} \,|\, H1] + Pr[\text{guess H1} \,|\, H0] \geq 1$

- Type I error $= Pr[\text{guess H0} \,|\, H1]$
- Type II error $= Pr[\text{guess H1} \,|\, H0]$

Differential Privacy and membership inference

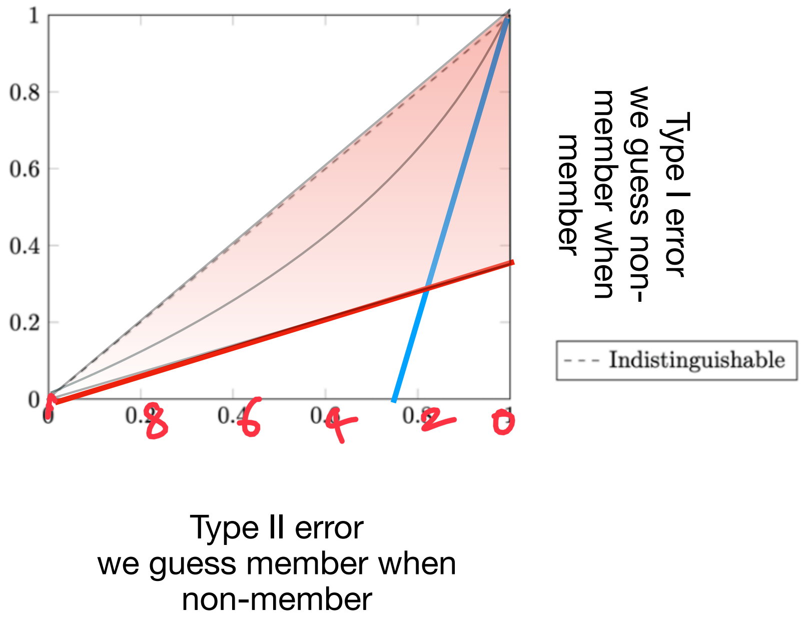
Visualizing connection



- $Pr[\text{guess H0} \,|\, H1] + e^{\varepsilon} Pr[\text{guess H1} \,|\, H0] \geq 1$
  - gives us blue line with slope e^{ε}

Differential Privacy and membership inference

Visualizing connection

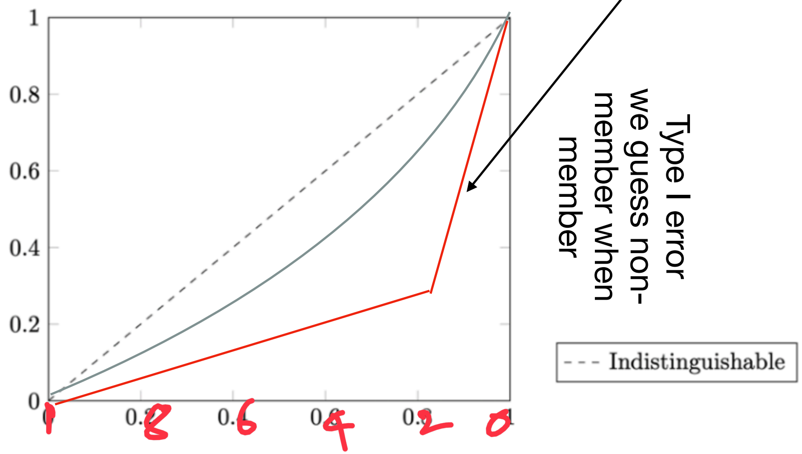


- $e^{\varepsilon} Pr[\text{guess H0} \,|\, H1] + Pr[\text{guess H1} \,|\, H0] \geq 1$
    - gives the red line with slope $e^{-\varepsilon}$

Differential Privacy and membership inference

Visualizing tradeoff curve of DP

Theoretical lower bound



Type II error
we guess member when non-member

- $Pr[\text{guess H0} \,|\, H1] + e^{\varepsilon} Pr[\text{guess H1} \,|\, H0] \geq 1$
    - gives us blue line
- $e^{\varepsilon} Pr[\text{guess H0} \,|\, H1] + Pr[\text{guess H1} \,|\, H0] \geq 1$
    - gives the red line

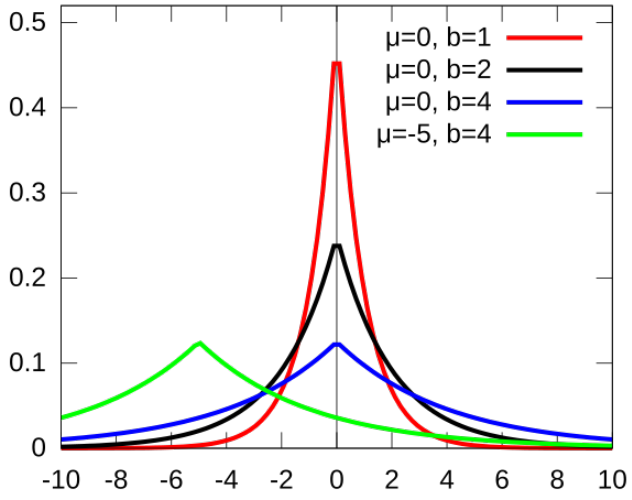
Algorithms for Differential Privacy

Differentially Private Algorithms

Just add Laplace noise

$$\forall y, \forall \text{ similar } D, D' : \quad \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \exp(\varepsilon)$$

- Suppose A(D) = 0, A(D') = 1.
- Release $\hat{y} = y +$ Laplace$(0, \varepsilon^{-1})$
- $z \sim$ Laplace$(\mu, b) \Rightarrow p(z) = \frac{1}{2b} e^{\frac{-|z-\mu|}{b}}$

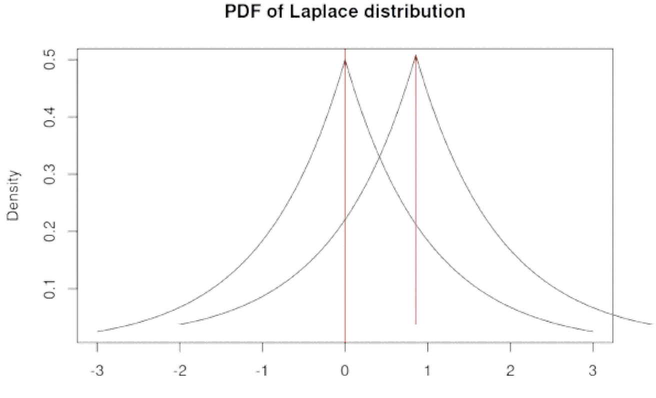


Differentially Private Algorithms

Just add Laplace noise

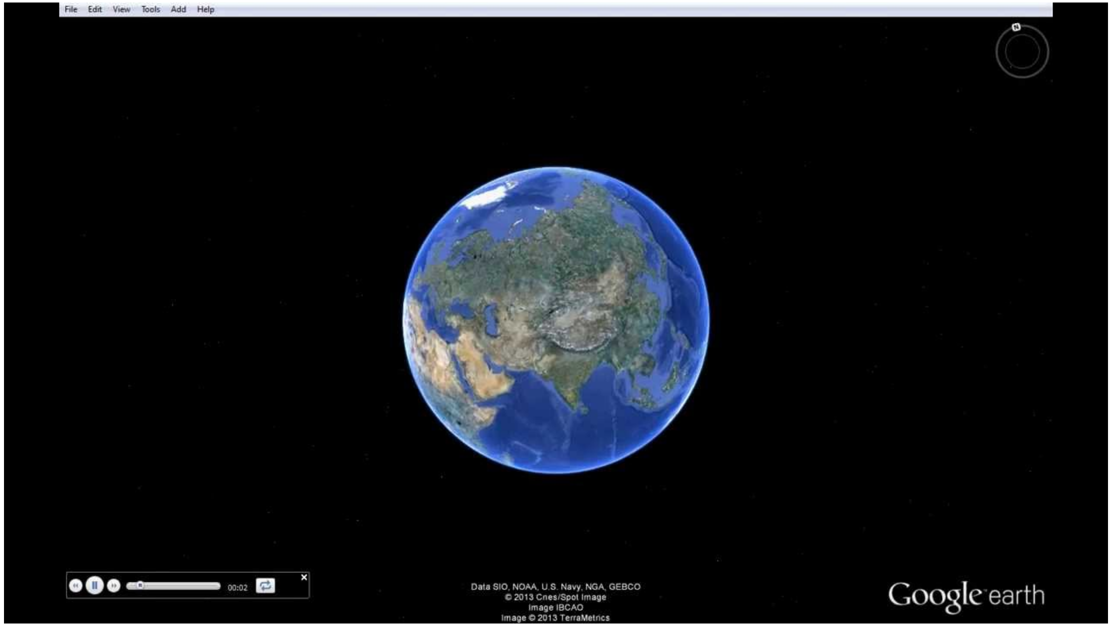
$$\forall y, \forall \text{ similar } D, D' : \quad \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \exp(\varepsilon)$$

- Suppose A(D) = 0, A(D') = 1. Release $\hat{y} = y + \text{Laplace}(0, \varepsilon^{-1})$
- $Pr[\hat{y} \,|\, y = 0] = \text{Laplace}(0, \varepsilon^{-1})$ and $Pr[\hat{y} \,|\, y = 1] = \text{Laplace}(1, \varepsilon^{-1})$
- $\dfrac{Pr[A(D) = y]}{Pr[A(D') = y]} = \dfrac{e^{-\varepsilon|y|}}{e^{-\varepsilon|y-1|}} = e^{\varepsilon}$



Differentially Private Algorithms

Sensitivity



- I release average income at different zoom levels. Added Lap(0,1).
- Do they all leak same amount of privacy?

Differentially Private Algorithms

Sensitivity and Laplace mechanism

$y \in \mathbb{R}^d$

- **Definition:** Sensitivity of a function $f : (x_1, \cdots, x_n) \mapsto (y_1, \cdots, y_d)$ with respect to a norm $||\cdot||$ is

$f(D) \in \mathbb{R}^d$

- $$\Delta f = \max_{\text{similar datasets } D,D'} ||f(D) - f(D')||$$

Theorem

Suppose f is Δ-sensitive with respect to $||\cdot||_1$. Then, the following satisfies ε-DP:

$$[A(D)]_i = [f(D)]_i + \text{Laplace}(0, \Delta\varepsilon^{-1}) \quad \forall i \in d$$

Proof: $\Pr[A(D) = y] = \Pr\left[[f(D)]_i + \mathrm{Lap}(0, \Delta\epsilon^{-1}) = y\right] \quad \forall i \in d$

$$= \frac{\epsilon}{2\Delta} \prod_{i=1}^{d} \exp\left(\frac{-\epsilon}{\Delta} \left| y_i - [f(D)]_i \right|\right)$$

$$\Rightarrow \frac{\Pr[A(D) = y]}{\Pr[A(D') = y]} = \frac{\prod_{i=1}^{d} \exp\left(\frac{-\epsilon}{\Delta} \left| y_i - [f(D)]_i \right|\right)}{\prod_{i=1}^{d} \exp\left(\frac{-\epsilon}{\Delta} \left| y_i - [f(D')]_i \right|\right)}$$

$$= \exp\left(\frac{\epsilon}{\Delta} \sum_{i=1}^{d} \left(\left| y_i - [f(D')]_i \right| - \left| y_i - [f(D)]_i \right| \right)\right)$$

$$\stackrel{(a)}{\leq} \exp\left(\frac{\varepsilon}{\Delta}\sum_{i=1}^{d}\left|[f(D')]_i - [f(D)]_i\right|\right)$$

$$\stackrel{(b)}{\leq} \exp\left(\frac{\varepsilon}{\Delta}\cdot\Delta\right)$$

$$\leq \exp(\varepsilon)$$

(a) is because $|a| - |b| \leq |a-b|$ (triangle ineq)

(b) is because f is Δ-sensitive wrt $\|\cdot\|_1$ norm

$\blacksquare$

Differentially Private Algorithms

Sensitivity and Laplace mechanism

- **Definition:** Sensitivity of a function $f : (x_1, \cdots, x_n) \mapsto (y_1, \cdots, y_k)$ with respect to a norm $||\cdot||$ is

$$\Delta f = \max_{\text{similar datasets } D, D'} ||f(D) - f(D')||$$

- How much noise should we add if we have Δ-sensitivity wrt $||\cdot||_\infty$

Can convert to $||.||_1$, then apply prev theorem.

- What about Δ-sensitivity wrt $||\cdot||_2$

- Laplace mechanism is great for functions with small ℓ_1 sensitivity, not so much for small ℓ_2 sensitivity

Really bad if $d \gg 1$

Use $\|v\|_1 \le d \|v\|_\infty$

$\|v\|_2 \le \|v\|_1 \le \sqrt{d} \|v\|_2$

Differentially Private Algorithms

Gaussian mechanism

- Suppose A(D) = 0, A(D') = 1.
- Release $\hat{y} = y +$ Gaussian$(0,\varepsilon^{-1})$
- $z \sim$ Gaussian$(\mu, \sigma^2) \Rightarrow p(z) \propto \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$
- $Pr[\hat{y} \,|\, y = 0] =$ Gaussian$(0,\varepsilon^{-1})$ and $Pr[\hat{y} \,|\, y = 1] =$ Gaussian$(1,\varepsilon^{-1})$
- $\frac{Pr[A(D) = y]}{Pr[A(D') = y]} = ?$ What happens at the tails?

$$\frac{\Pr[A(D)=y]}{\Pr[A(D')=y]} = \exp\left(\frac{\varepsilon}{2}\left((y-1)^2 - y^2\right)\right)$$

$$= \exp\left(\frac{\varepsilon(1-2y)}{2}\right)$$

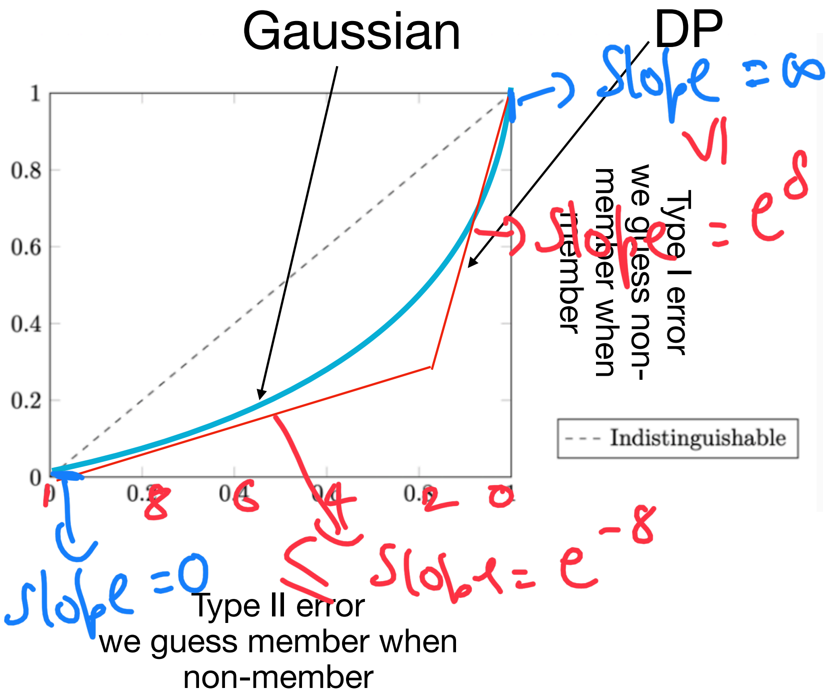
When $y \to -\infty$, this blows up to ∞

$\Rightarrow$ ε-DP impossible $\forall \varepsilon$!

$\Rightarrow$ No privacy?

Differentially Private Algorithms

Visualizing tradeoff curve of DP and Gaussian mechanism



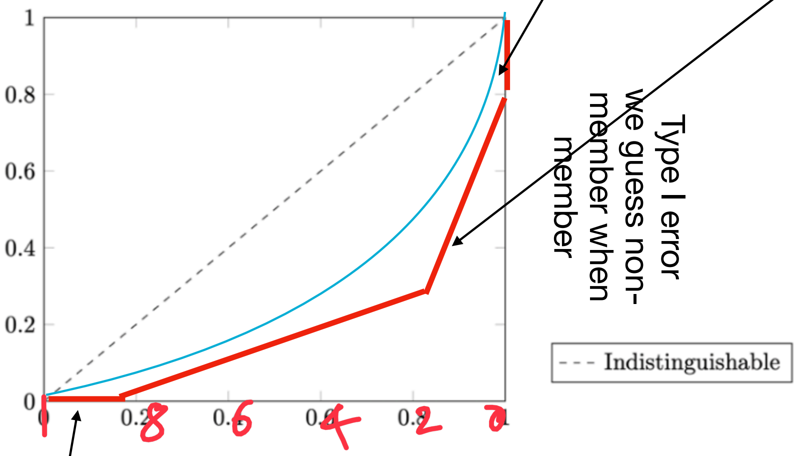
- At the edges, the slope of gaussian mechanism is vertical
- Impossible to get DP guarantee for any value of ε
- Does this mean Gaussian mechanism is not private?

Differentially Private Algorithms

Approximate DP

Vertical line of size δ

Approximate (ε, δ)-DP



Type II error
we guess member when non-member

Horizontal line of size δ

- Add flat lines of length δ at the edges to make some space for Gaussian mechanism
- Now chance for Gaussian mechanism to show privacy!

Differentially Private Algorithms

Aproximate Differential Privacy

(ε, δ)-Differential Privacy:

An algorithm A satisfies (ε, δ)-DP if for any similar datasets $D, D' \in \chi^n$ and $y \in \mathcal{Y}$

$$Pr[A(D) = y] \leq Pr[A(D') = y] \cdot \exp(\varepsilon) + \delta$$

- With δ probability anything can happen
- Typically δ is chosen very small $\delta \leq n^{-1}$

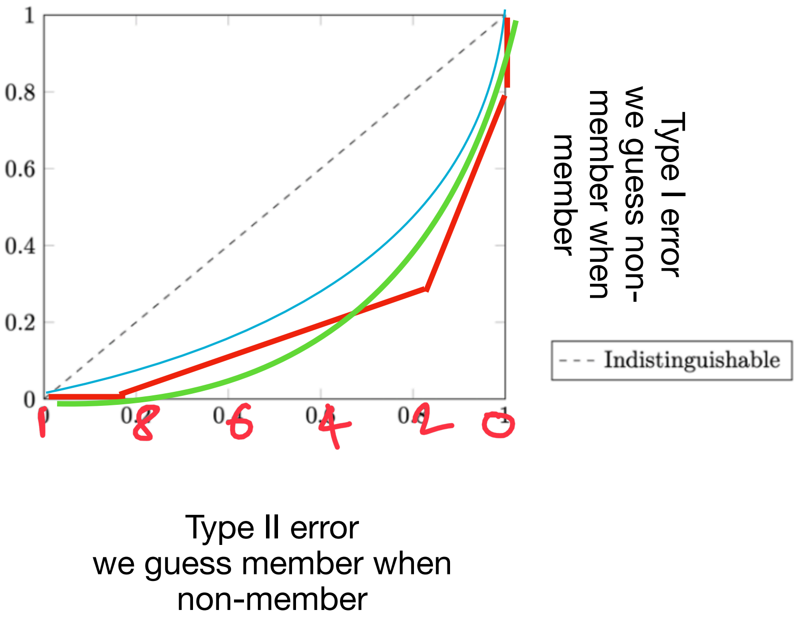
Differentially Private Algorithms

Gaussian mechanism

- Suppose A(D) = 0, A(D') = 1. Release $\hat{y} = y + \text{Gaussian}(0, \varepsilon^{-1})$
- $z \sim \text{Gaussian}(\mu, \sigma^2) \Rightarrow p(z) \propto \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$
- $Pr[\hat{y} \,|\, y = 0] = \text{Gaussian}(0, \varepsilon^{-1})$ and $Pr[\hat{y} \,|\, y = 1] = \text{Gaussian}(1, \varepsilon^{-1})$
- $\frac{Pr[A(D) = y]}{Pr[A(D') = y]} = ?$ what happens now?

Differentially Private Algorithms

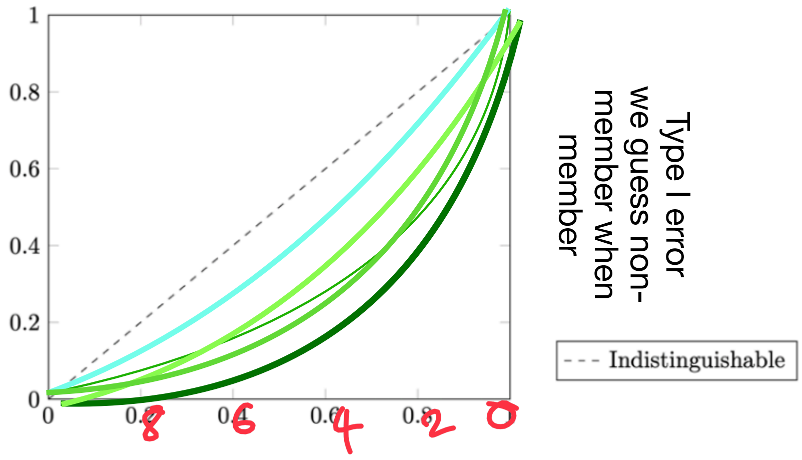
f-DP and Gaussian DP



- All this seems a bit ad-hoc. Is there a “canonical” definition of privacy?
- **Definition.** An algorithm A satisfies **f-DP** if the optimal tradeoff curve is below the function **f.**
- Generalizes all previous notions. What **f** should we pick? Both green and red curves satisfy.

Differentially Private Algorithms

f-DP and Gaussian DP



Type II error
we guess member when non-member

- There is a special family of curves: Gaussian tradeoff curve
- **Definition.** An algorithm A satisfies μ -**Gaussian Differential Privacy** if it is harder to distinguish between A(D) vs. A(D') than $\mathcal{N}(0,1)$ vs. $\mathcal{N}(\mu,1)$

Differentially Private Algorithms

Gaussian mechanism

- **Definition:** Sensitivity of a function $f : (x_1, \cdots, x_n) \mapsto (y_1, \cdots, y_k)$ with respect to a norm $||\cdot||$ is

$$\Delta f = \max_{\text{similar datasets } D,D'} ||f(D) - f(D')||$$

- What about Δ-sensitivity wrt $|| \cdot ||_2$
- Gaussian mechanism with GDP is great for ℓ_2 sensitivity!