CSCI 699: Privacy Preserving Machine Learning - Week 2

Differential Privacy

Sai Praneeth Karimireddy, Sep 6 2024

Recap

- why privacy

- 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.

side information
"linkage attacks"

"negation"

Differential Privacy

Takeaways

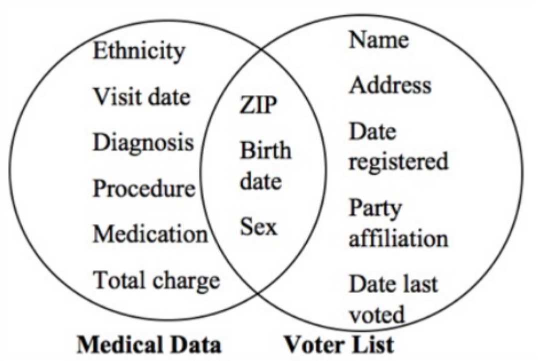
Requirements for privacy definition

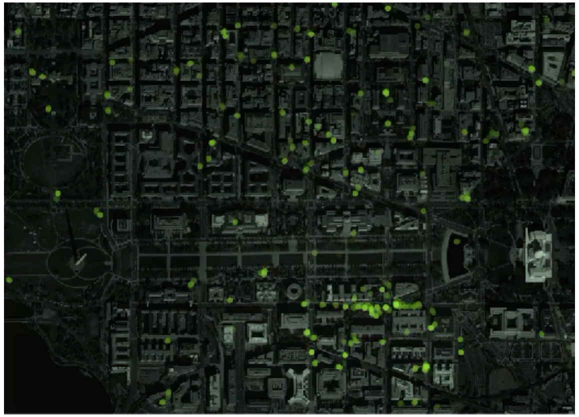
Post-processing

- 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.





Quantifying Privacy Leakage

Attempt 2

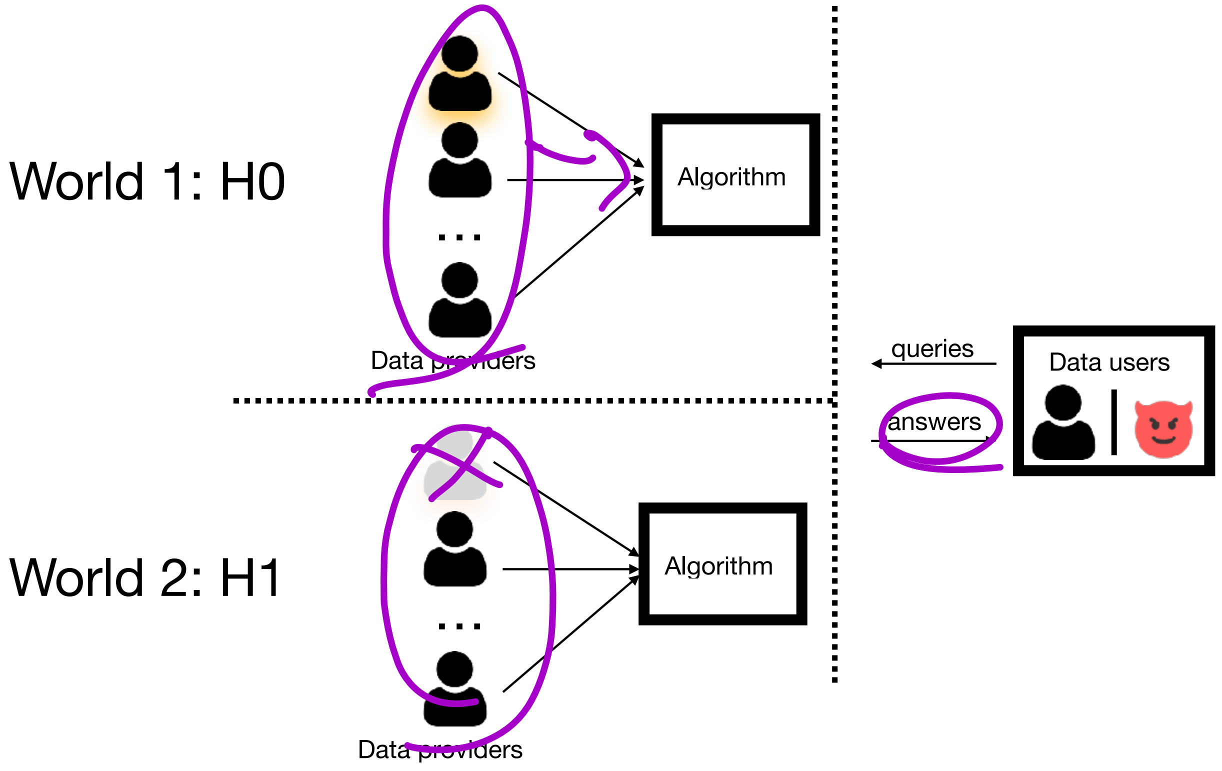
Relative Privacy: quantify **new** information leaked

"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.

Quantifying Privacy Leakage

Attempt 2



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

Quantifying Privacy Leakage

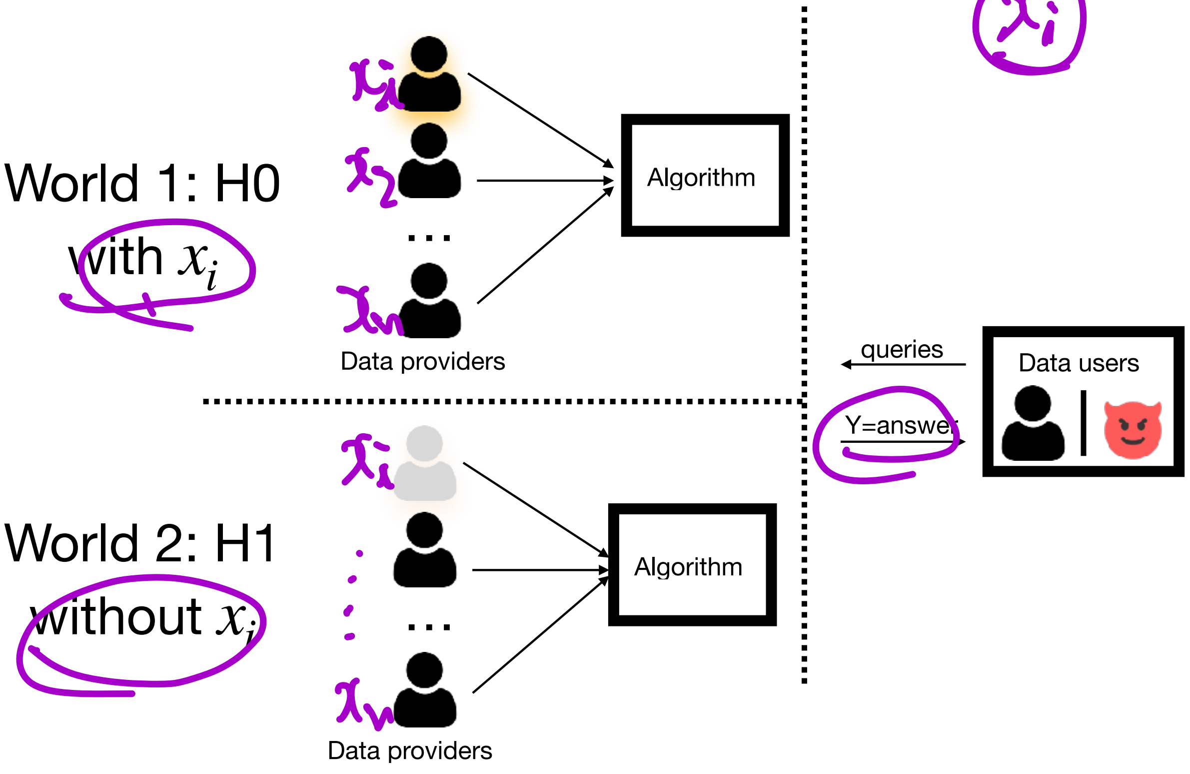
Membership Inference

attack

As a definition of privacy

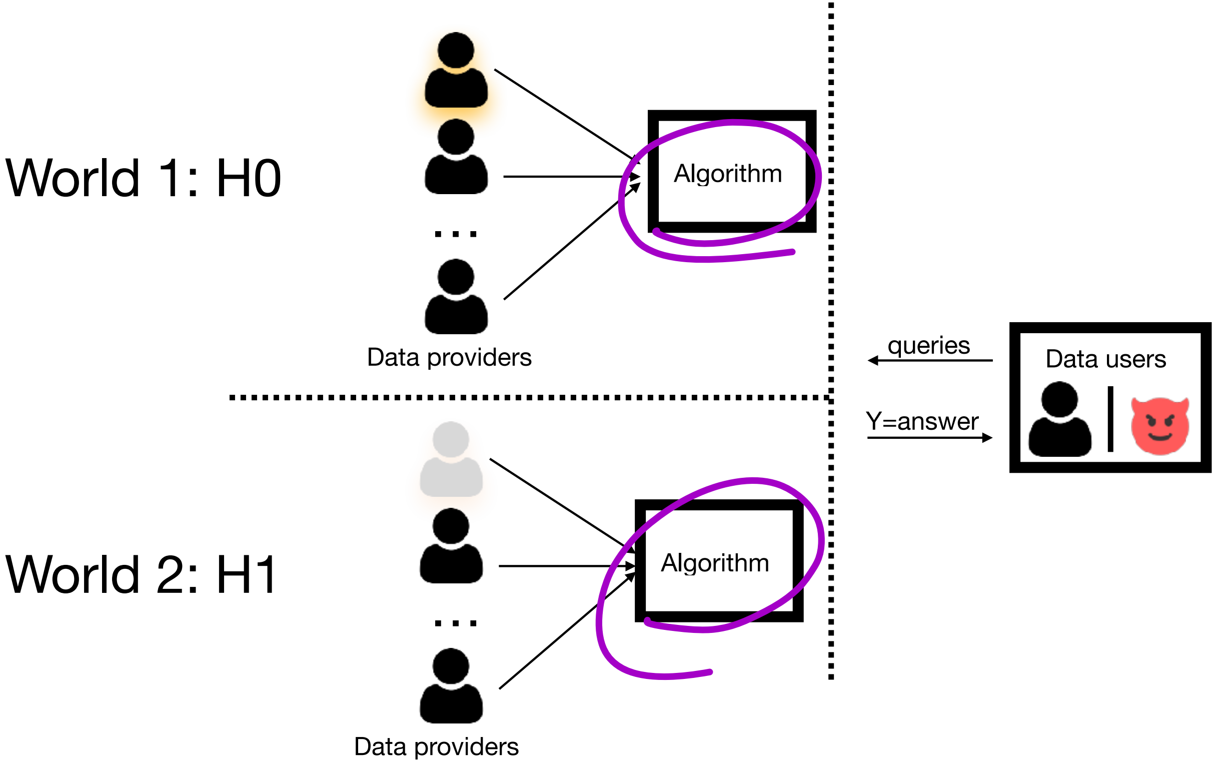
$D = \{x_1, \ldots, x_n\}$

x_i



- We know everything about the algorithm and even D, x_i,
- Only 1 bit unknown - H_0 or H_1?
- We observe an output Y
- Need to guess if it came from H0 or H1

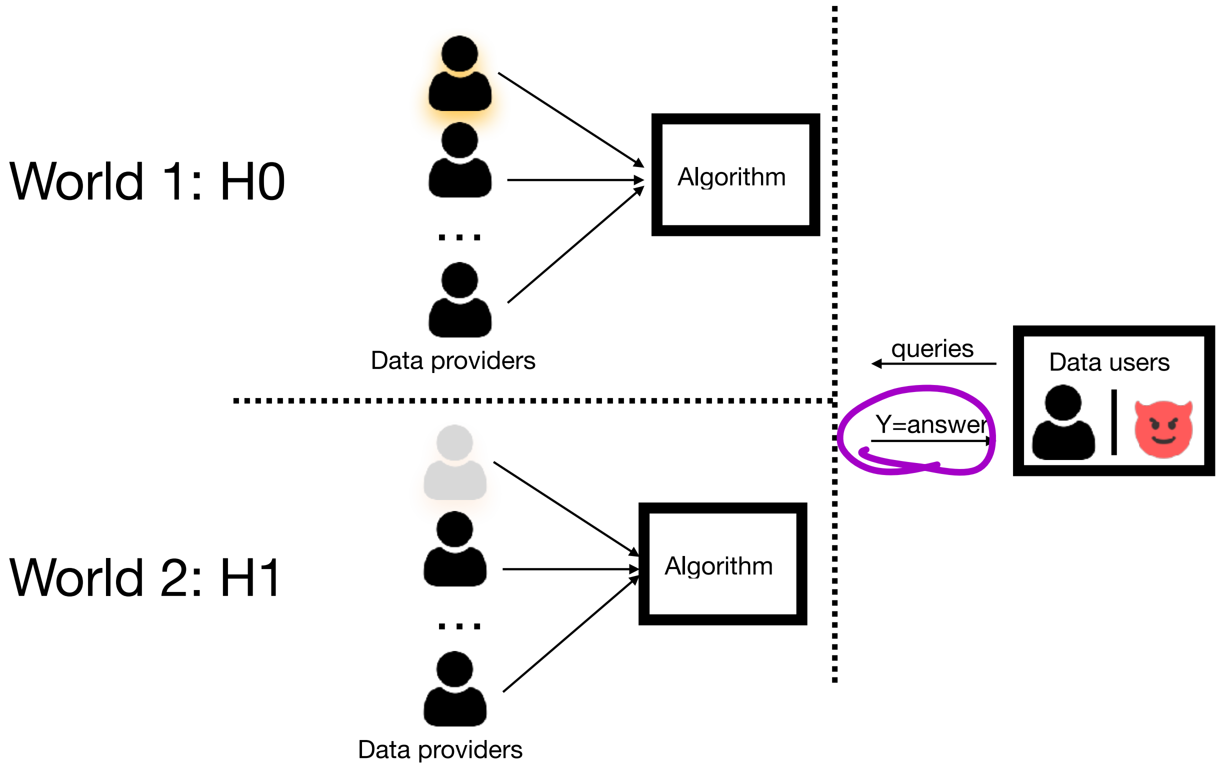
Membership Inference



- Can a deterministic algorithm be private?

Membership Inference

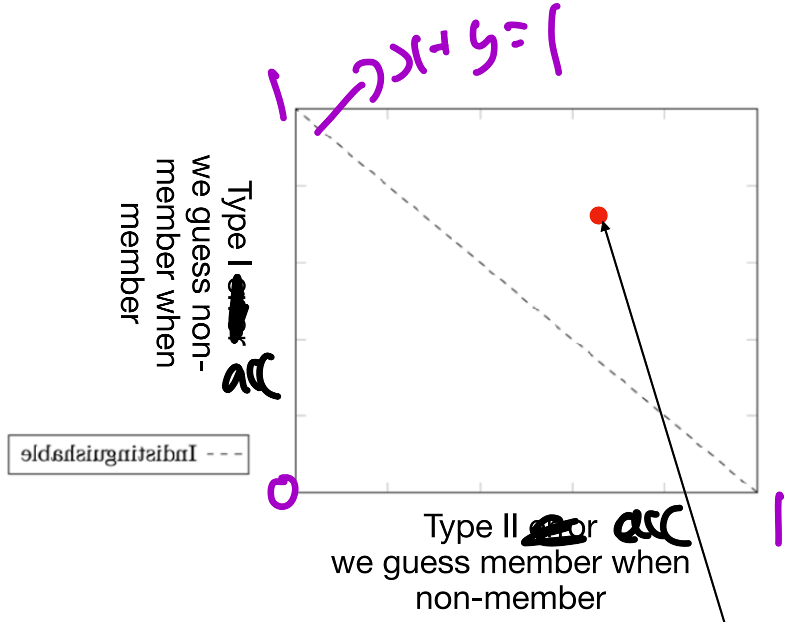
→ Y is a random variable
→ acc is also a random



- Can a deterministic algorithm be private?
- No - adversary can simply compute $Y = f(D)$ or $f(D \backslash x_i)$?
- Need randomness - adversary will have type I and type II errors

Membership Inference

Quantifying attack success

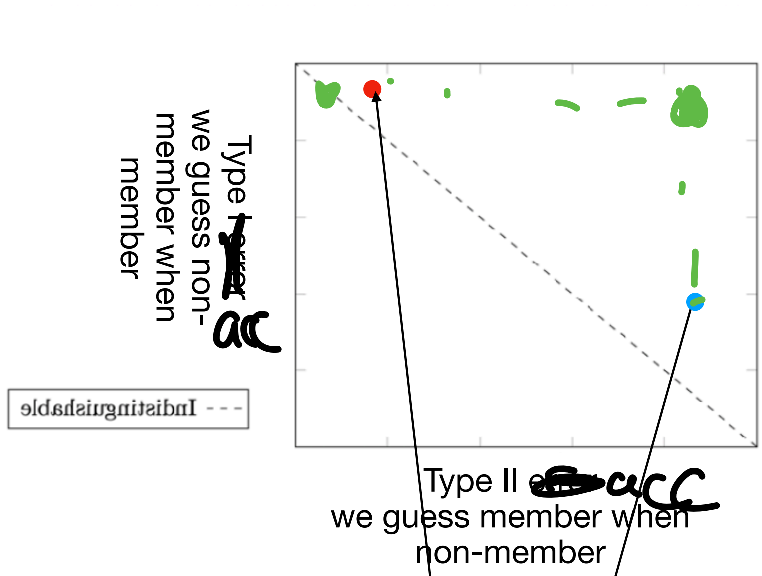


We get a point.

- 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

Quantifying attack success

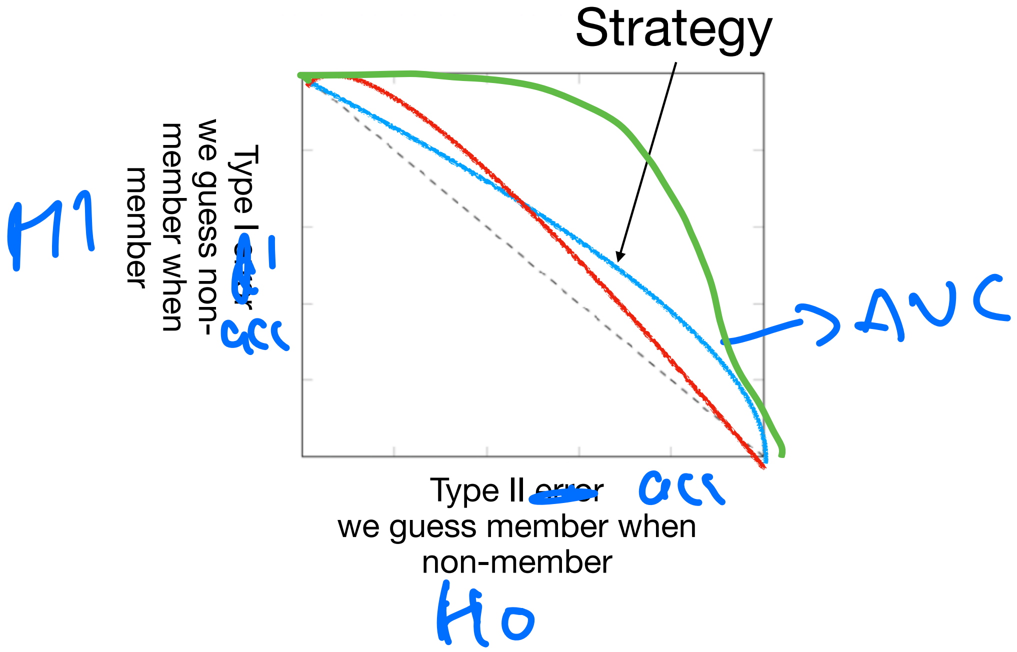


We have two points.

- Suppose we have two algorithms, each with different type I and type II errors.
- Which one has more privacy leakage?

Membership Inference

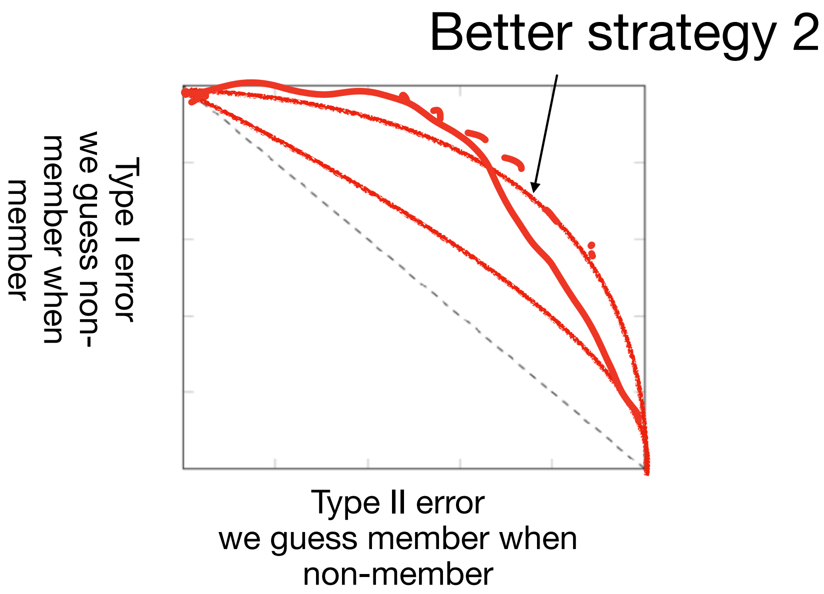
Tradeoff curve



- Depends on what we care
- E.g. its important not to miss anyone e.g. sending cat ads to pet owners - coverage
- Not ok if we are accusing them of a crime - precision much more important
- Impossible to compare individual points - need to compare entire trade off curves.

Membership Inference

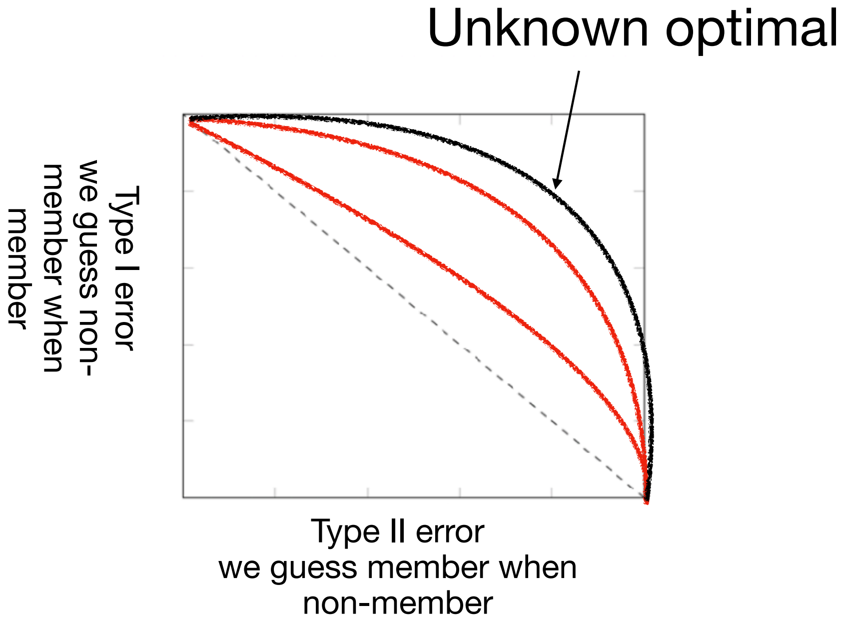
Comparing tradeoff curves



- Tradeoff curve depends on testing strategy adversary uses.
- Strategy 2is better than Strategy 1 if the curve is uniformly above.
- Higher curve means we've found more privacy leakage

Membership Inference

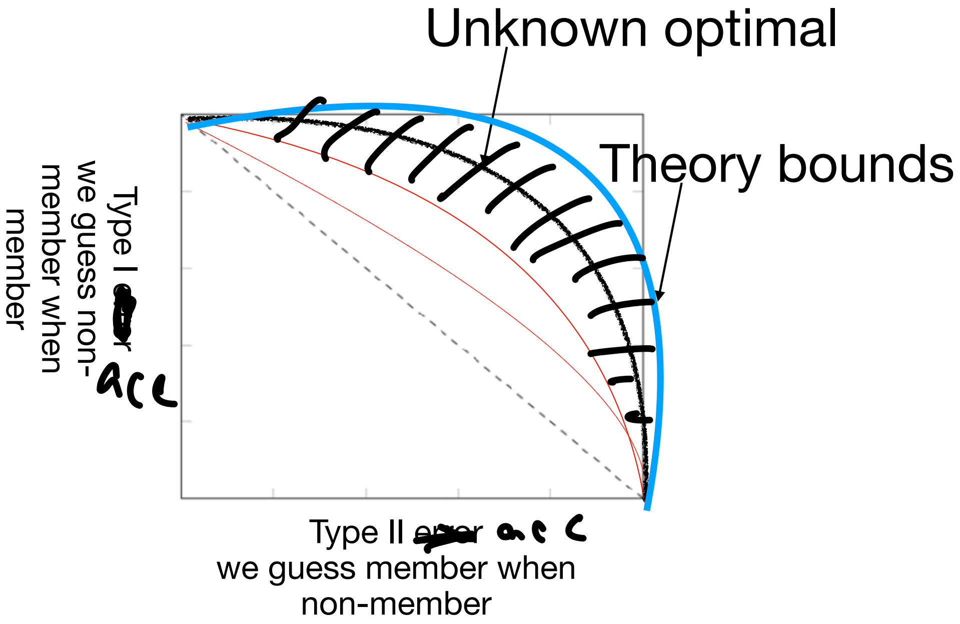
Optimal tradeoff curve



- 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.

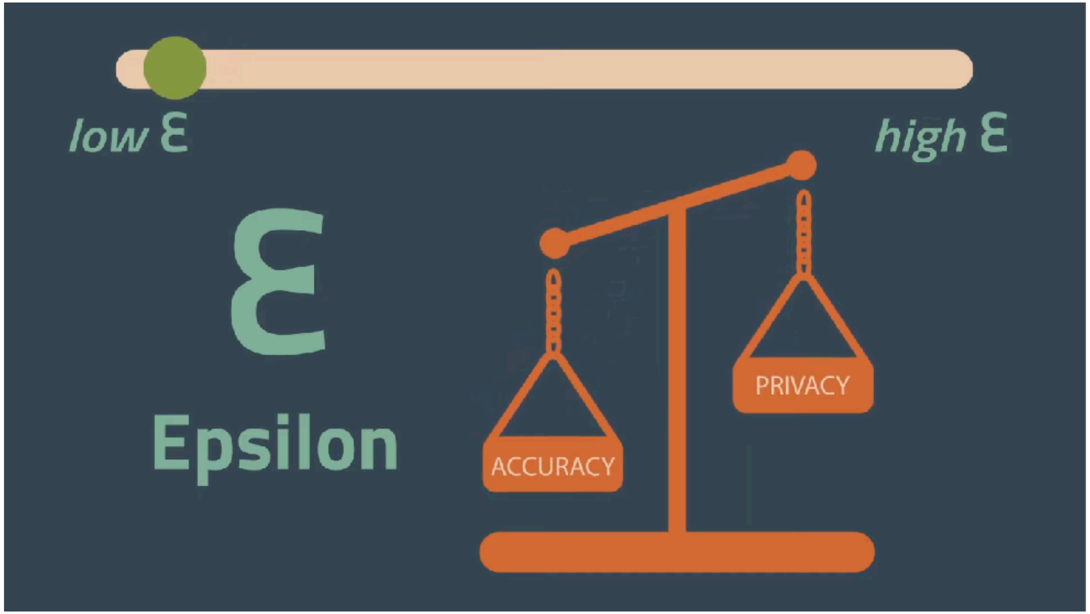
Membership Inference

Privacy from tradeoff curve



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

algorithm prior knowledge

$$Pr[D \,|\, Y] = \frac{Pr[Y \,|\, D] \cdot Pr[D]}{Pr[Y]}$$

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

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

  - Can even recover x using max-likelihood

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

Differential Privacy

Goal

$$Pr[D \mid Y = y]$$
$$= Pr[D \setminus \{x\} \mid Y = y]$$
$$\Uparrow$$
$$Pr[Y = y \mid D]$$
$$= Pr[Y = y \mid D \setminus \{x\}]$$

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

algorithm prior knowledge

$$Pr[D \mid Y] = \frac{Pr[Y \mid D] \cdot Pr[D]}{Pr[Y]}$$

- 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

ε-DP

- Some indistinguishability: For all neighboring datasets, the output probabilities are bounded.

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

- It means by observing any Y, adversary is NOT able to distinguish between inputs x and x' beyond a bounded certainty.

- What does neighboring datasets mean? Depends on use case
  - location positions that are within some range
  - datasets that differ in one individual row (our focus)
  - edit distance 1

Differential Privacy

Formal definition

ε-Differential Privacy:

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

Privacy

$$\log \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \varepsilon \text{ a.s.}$$

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

Differential Privacy

Formal definition

$0 \leq \varepsilon \leq 1$

sensitive

$\varepsilon = 10$

100

1K

ε-Differential Privacy:

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

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

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

Differential Privacy

Source of randomness

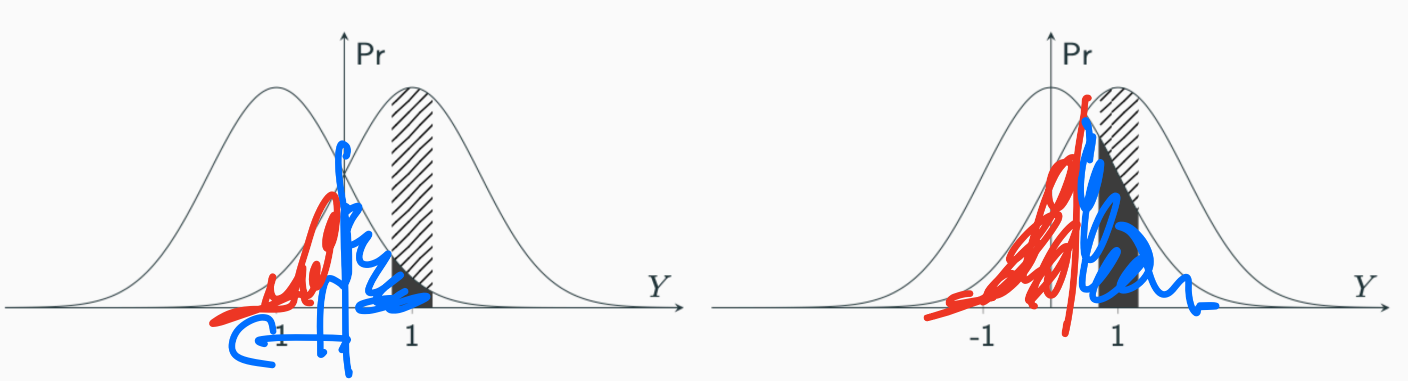
$$\log \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq \varepsilon \quad a.s.$$

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



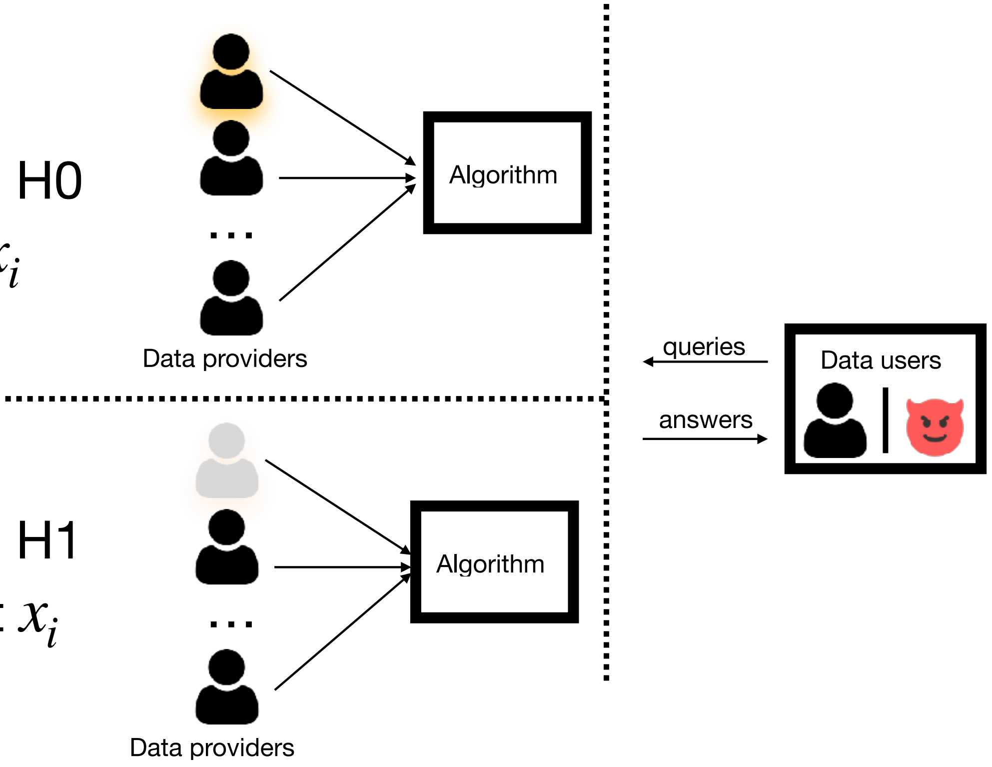
(a) Large ϵ (b) Small ε

Differential Privacy

Recall Membership Inference

World 1: H0
with x_i

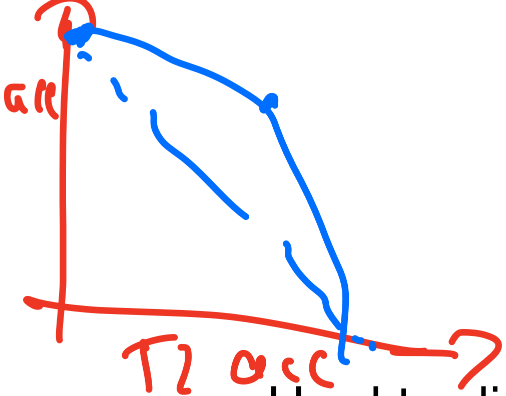
World 2: H1
without x_i



- 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

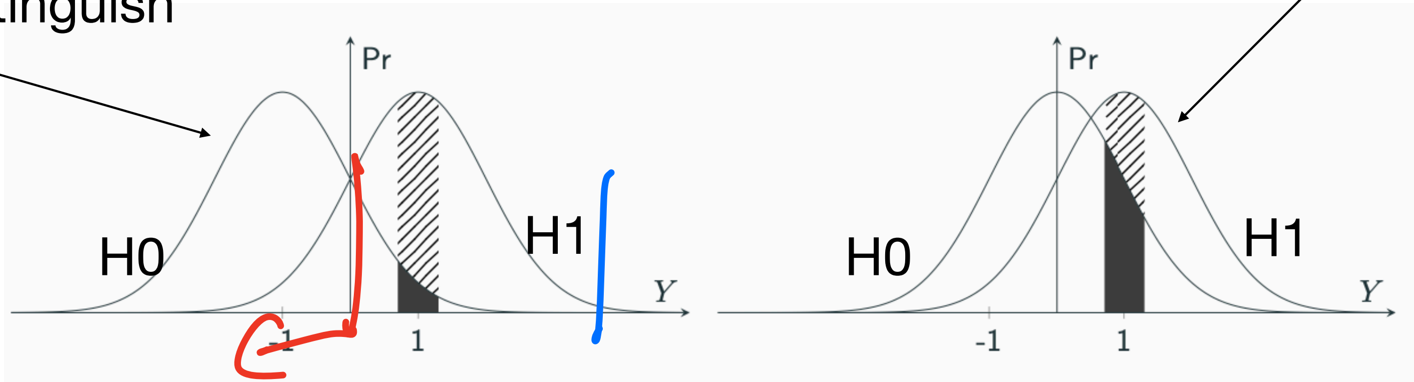
Differential Privacy

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?

Differential Privacy and membership inference

Quantifying connection

Theorem

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

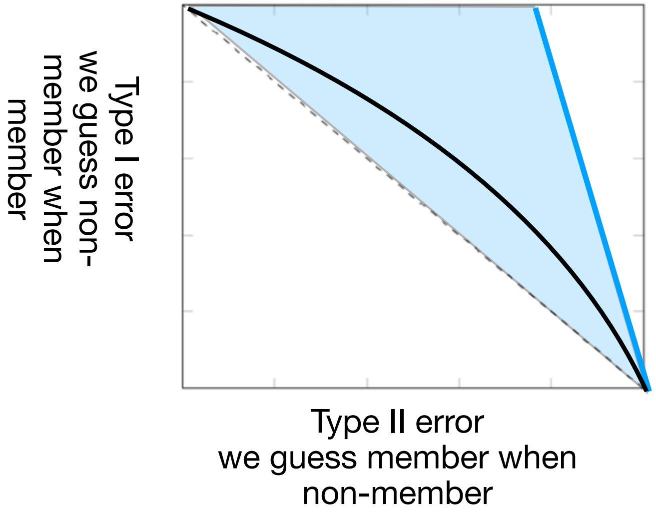
- $Pr[\text{guess H0} \mid H1] + e^{\varepsilon} Pr[\text{guess H1} \mid H0] \geq 1$
- $e^{\varepsilon} Pr[\text{guess H0} \mid H1] + Pr[\text{guess H1} \mid H0] \geq 1$

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

Differential Privacy and membership inference

Visualizing connection

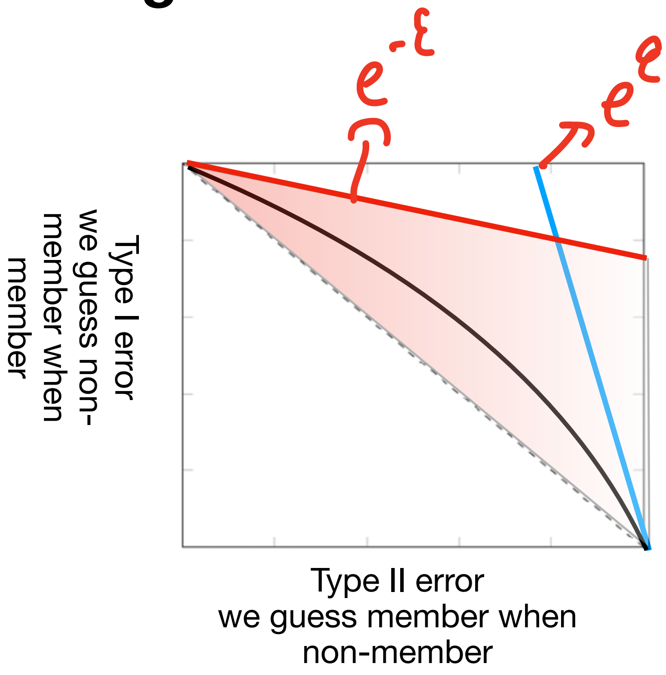
ε - DP



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

Differential Privacy and membership inference

Visualizing connection

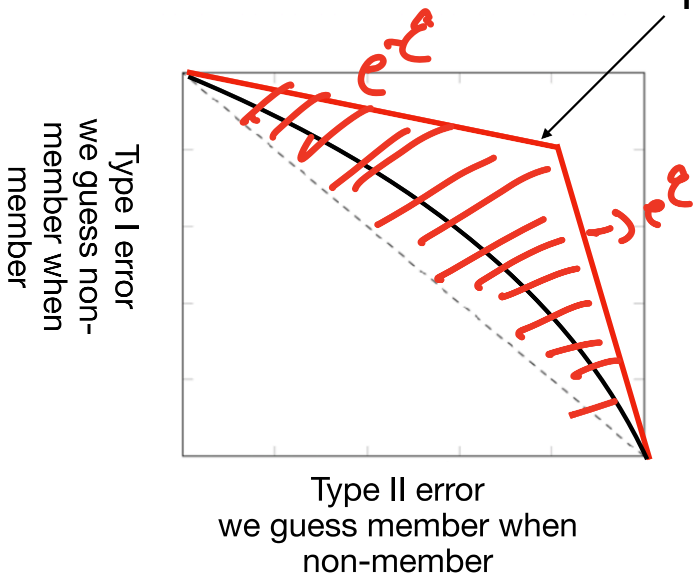


- $e^{\varepsilon} Pr[\text{guess H0} \mid H1] + Pr[\text{guess H1} \mid H0] \geq 1$
  - gives the red line with slope $e^{-\varepsilon}$

Differential Privacy and membership inference

Visualizing tradeoff curve of DP

Theoretical upper bound $\supseteq \varepsilon$-DP



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

what ε does this satisfy?

Aside: Is Putin's popularity calculation private?

List Experiment

n = 100

- 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 satisfy DP?

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

50 | 50

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

$D = ?$

$x_i = ?$

$Alg = ?$

$y = \sum_i y_i$ 3 4

$\to \{0, 1, 3, 4\}$

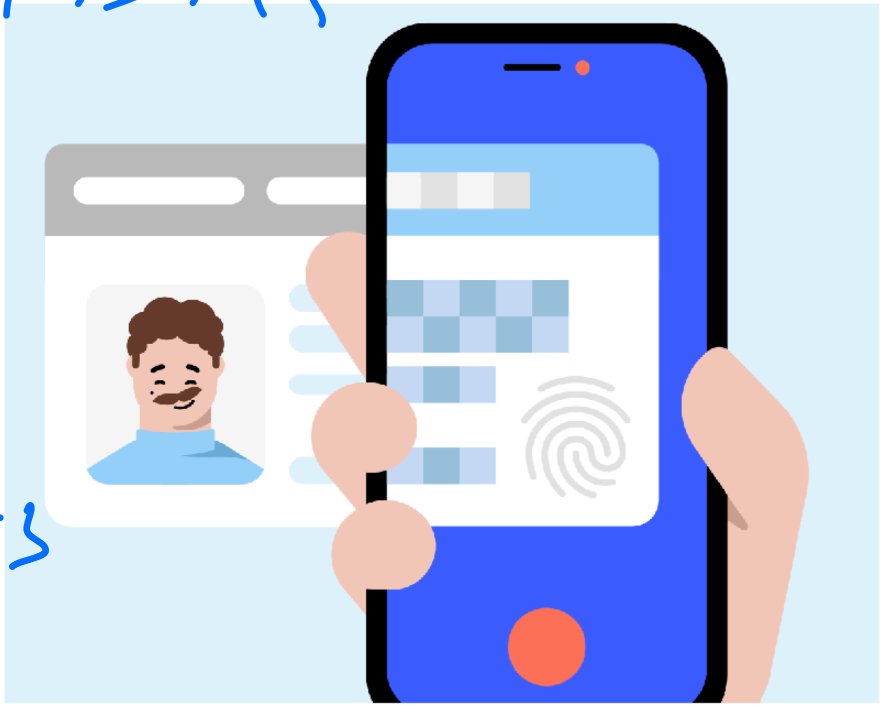
$\to i$

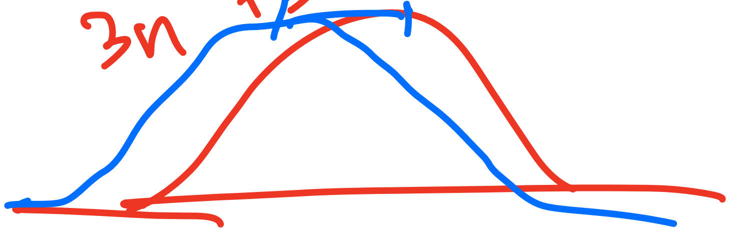
Algorithms for Differential Privacy

$H0$

$3n + Bin(n, 0.5)$

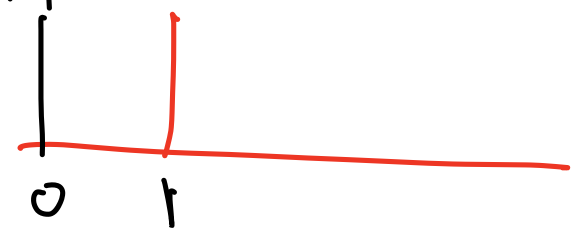
$H1 = H0 - 3$





$n = 1$ H

$H1$ $H0$



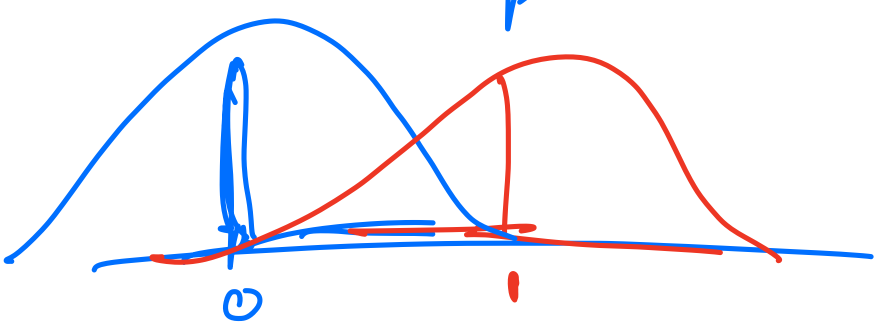
$Y = 0 \quad \leftarrow D$

$Y = 1 \quad 1 \quad \leftarrow D \setminus \{x\}$

output = Y + Noise

$\downarrow$

$\mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \propto e^{-t \cdot \varepsilon}$



$\leq e^{\varepsilon}$

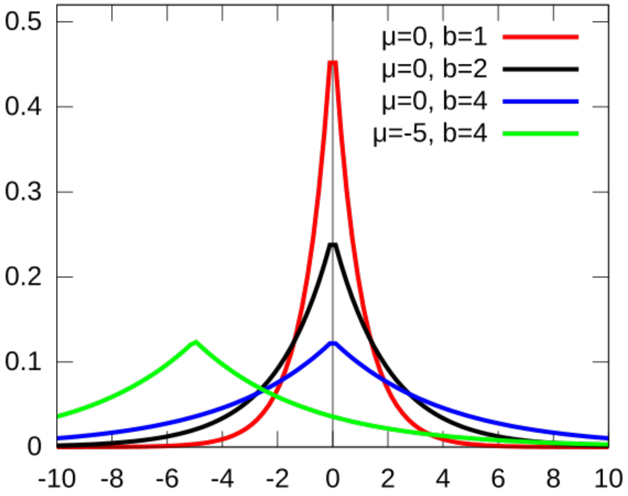
Laplace

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$(\mu, b) \Rightarrow p(z) = \frac{1}{2b} e^{\frac{-|z-\mu|}{b}}$



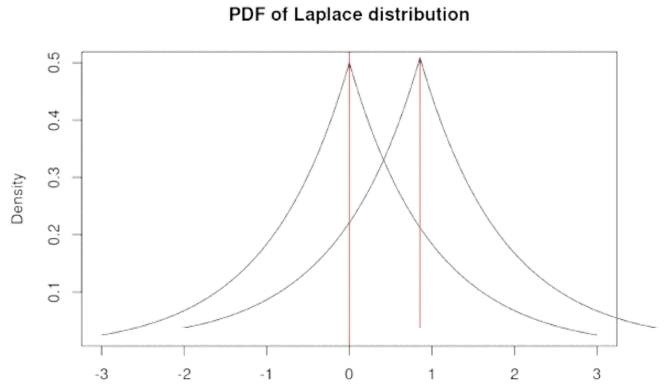
Differentially Private Algorithms

Just add Laplace noise

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

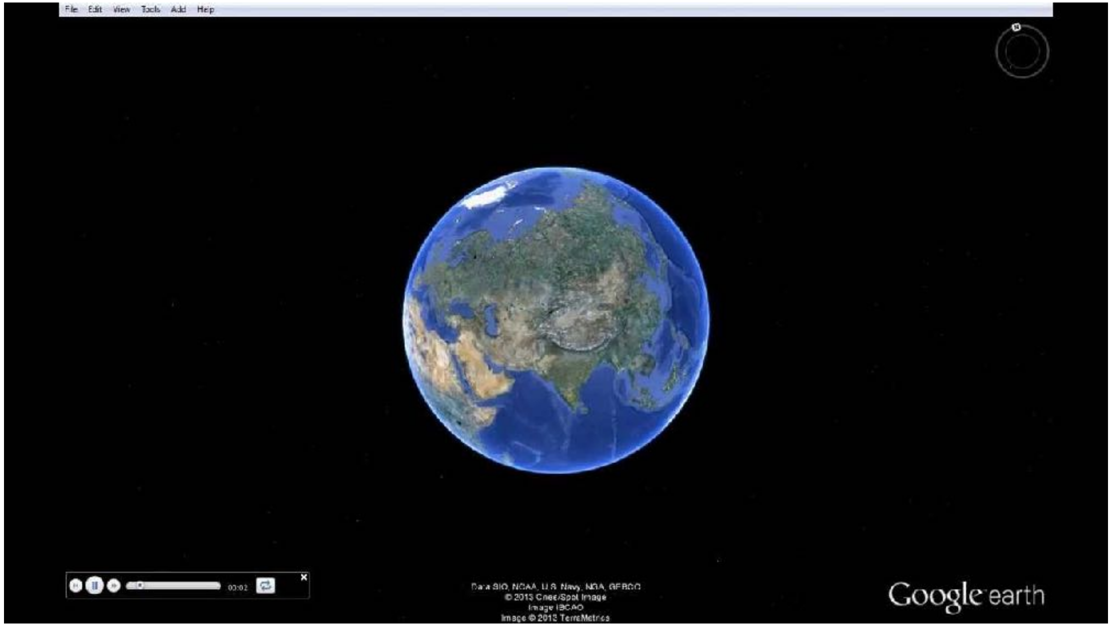
- Suppose A(D) = 0, A(D') = 1. Release $\hat{y} = y + \text{Laplace}(0, \varepsilon^{-1})$
- $Pr[\hat{y} \,|\, y = 0] = \text{Laplace}(0, \varepsilon^{-1})$ and $Pr[\hat{y} \,|\, y = 1] = \text{Laplace}(1, \varepsilon^{-1})$
- $$\frac{Pr[A(D) = y]}{Pr[A(D') = y]} = \frac{e^{-\varepsilon |y|}}{e^{-\varepsilon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

$$e^{\varepsilon \underbrace{(|y-1| - |y|)}_{\le 1}}$$



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

- **Definition:** Sensitivity of a function $f : (x_1, \cdots, x_n) \mapsto (y_1, \cdots, y_k)$ with respect to a norm $||\cdot||$ is
-
$$\Delta f = \max_{\text{similar datasets } D,D'} ||f(D) - f(D')|| \leq \Delta$$

$||\cdot||_2 \quad \Delta_2$

Theorem

Suppose f is Δ_1-sensitive with respect to $||\cdot||_1$. Then, the following satisfies ε-DP:

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

Δ_∞

$||f(D) - f(D')||_\infty \leq \Delta_\infty$

$$H0 = f(D) + Lap\left(0, \frac{\Delta}{\varepsilon}\right)$$

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

$$H1 = f(D') + Lap\left(0, \frac{\Delta}{\varepsilon}\right)$$

$$= Lap\left(f(D'), \frac{\Delta}{\varepsilon}\right)$$

$$\frac{P(Y=y \mid H0)}{P(Y=y \mid H1)} = \frac{\exp\left(-|\cancel{t}\, y - f(D)|\,\varepsilon/\Delta\right)}{\exp\left(-|\cancel{t}\, y - f(D')|\,\varepsilon/\Delta\right)}$$

$$= \exp\left(\frac{\varepsilon}{\Delta}\left(|y - f(D')| - |y - f(D)|\right)\right)$$

△'e ineq

$$\leq \exp\left(\frac{\varepsilon}{\Delta}\,\underbrace{|f(D) - f(D')|}_{\leq \Delta}\right)$$

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

$$Y_i \mid H0 \;\propto\; \mathrm{Lap}\left([f(D)]_i, \frac{\Delta}{\varepsilon}\right)$$

$$Y_i \mid H1 \;\propto\; \mathrm{Lap}\left([f(D')]_i, \frac{\Delta}{\varepsilon}\right) \quad d\Delta_\infty$$

$$\frac{\Pr[Y_i = y_i \mid H0]}{\Pr[Y_i = y_i \mid H1]} \leq \exp\left(\frac{\varepsilon}{\Delta} \left|[f(D)]_i - f(D')\right|\right)$$

$$\frac{\Pr[\vec{Y} = \vec{y} \mid H0]}{\Pr[\vec{Y} = \vec{y} \mid H1]} = \prod_{i=1}^{d} \frac{\Pr[Y_i = y_i \mid H0]}{\Pr[Y_i = y_i \mid H1]}$$

d-dim

$$\leq \exp\left(\frac{\varepsilon}{\Delta} \sum_i |f_i(D) - f_i(D')|\right)$$

$$\leq d\Delta_\infty$$

$$= \exp\left(\frac{\varepsilon}{\Delta} \underbrace{\|f(D) - f(D')\|_1}_{\leq \Delta}\right)$$

$$\leq \sqrt{d}\,\Delta_2$$

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

Differentially Private Algorithms

$\sum_i (f_i(D)-f_i(D'))^2 \le \Delta_2^2$

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 $||\cdot||_\infty$
- 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

$A(D) = 0$ $\qquad$ $A(D') = 1$

$$\hat{Y} = Y + N(0,\sigma^2)$$

$$\frac{\Pr[Y=y \mid H0]}{\Pr[Y=y \mid H1]} = \frac{N(0,\sigma^2)}{N(1,\sigma^2)}$$

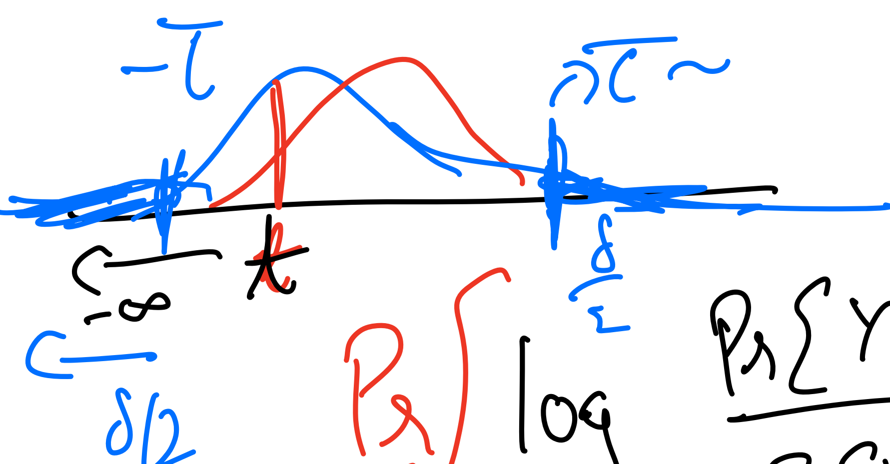
$$= \frac{\exp\left(-\frac{1}{2} y^2/\sigma^2\right)}{\exp\left(-\frac{1}{2}(y-1)^2/\sigma^2\right)}$$

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

$$= \exp\left(\underbrace{\frac{1}{2\sigma^2}(1-2y)}_{\le \varepsilon}\right)$$

$\sum \left| f_i^2(D) - f_i^2(D') \right|$

$\le \sum \left(f_i(D) - f_i(D')\right)$



$$\Pr_{t \sim A(D)}\left[\log \frac{\Pr\{Y=t \mid H0\}}{\Pr\{Y=t \mid H1\}} \ge \varepsilon\right] \le \delta$$

$$\exp\left(\frac{1}{2\sigma^2}(1+2\tau_\delta)\right) \le e^\varepsilon$$

$$\sigma^2 = \frac{1+2\tau_\delta}{2\varepsilon}$$

$$\Rightarrow (\varepsilon, \delta)\text{-DP}$$

gaussian mechanism

$$\|f(D) - f(D')\|_2 \leq \Delta_2 \quad \forall D, D'$$

$$Y_i = f_i(D) + \mathrm{Lap}\left(\frac{\sqrt{d}\,\Delta_2}{\varepsilon}\right)$$

$$Y_i = f_i(D) + N\left(0, \frac{\Delta_2}{\varepsilon}\right)$$

independent of d

Differentially Private Algorithms

Gaussian mechanism

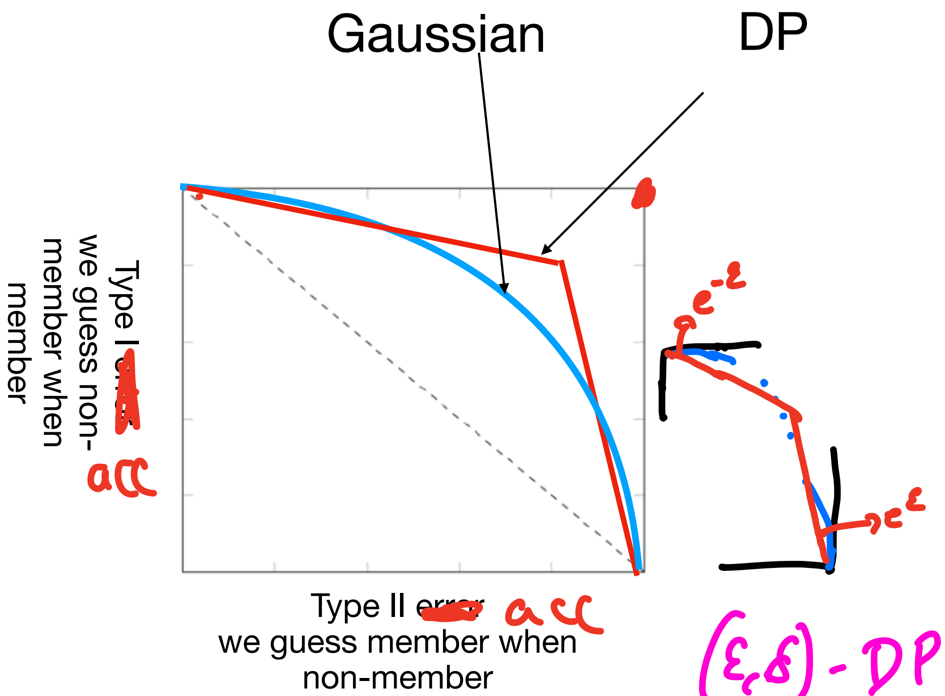
(ε, δ) - Approx. D.P

(ε, δ)-DP

- 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 $Pr[\hat{y} \,|\, y = 1] = \text{Gaussian}(1, \varepsilon^{-1})$
- $\dfrac{Pr[A(D) = y]}{Pr[A(D') = y]} = ?$ What happens at the tails?

Differentially Private Algorithms

Visualizing tradeoff curve of DP and Gaussian mechanism



(ε, δ)-DP

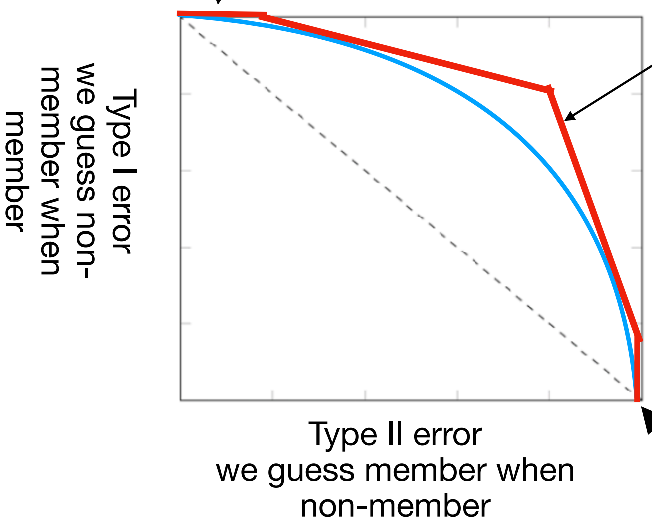
- At the edges, the slope of gaussian mechanism is vertical
- Impossible to get DP guarantee for any value of ε
- Does this mean Gaussian mechanism is not private?

Differentially Private Algorithms

Approximate DP

Horizontal line of size δ

Approximate (ε, δ)-DP



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

Vertical line of size δ

Differentially Private Algorithms

Approximate Differential Privacy

(ε, δ)-Differential Privacy:

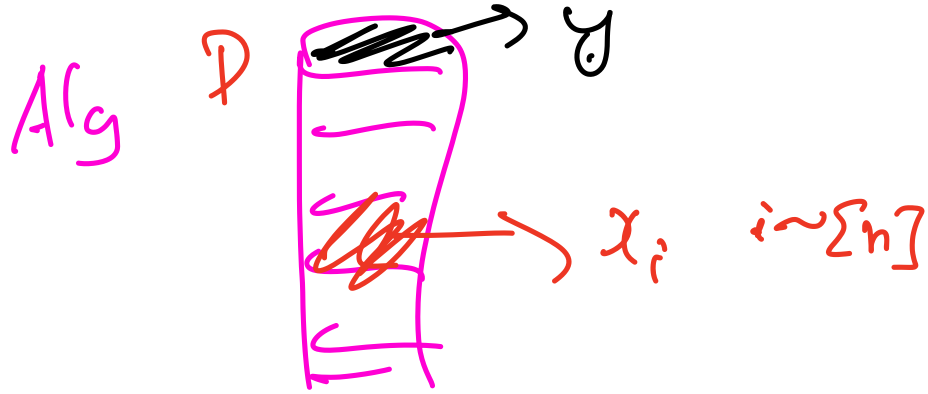
Let us draw a variable $t \sim A(D)$. Then the privacy loss random variable:

$$\mathscr{L}_{D,D'} := \ln\left(\frac{Pr[A(D) = t]}{Pr[A(D') = t]}\right)$$

A satisfies (ε, δ)-DP iff for any neighboring datasets $D, D' \in \chi^n$ we have

$$Pr\left[\mathscr{L}_{D,D'} \geq \varepsilon\right] \leq \delta$$

- With δ probability, arbitrarily bad things can happen.
- Ideally δ is chosen very small $\delta \leq n^{-1}$, or more common in fixed to 10^{-5}.



$$\Pr\left[\log \frac{\Pr[Y=y \mid D]}{\Pr[Y=y \mid D \setminus \{x_i\}]} > 0\right] \le \frac{1}{n}$$

$\varepsilon = 0, \ \delta = \frac{1}{n}$ $\quad$ $\left(0, \frac{1}{n}\right)$-DP

Differentially Private Algorithms

Gaussian mechanism

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