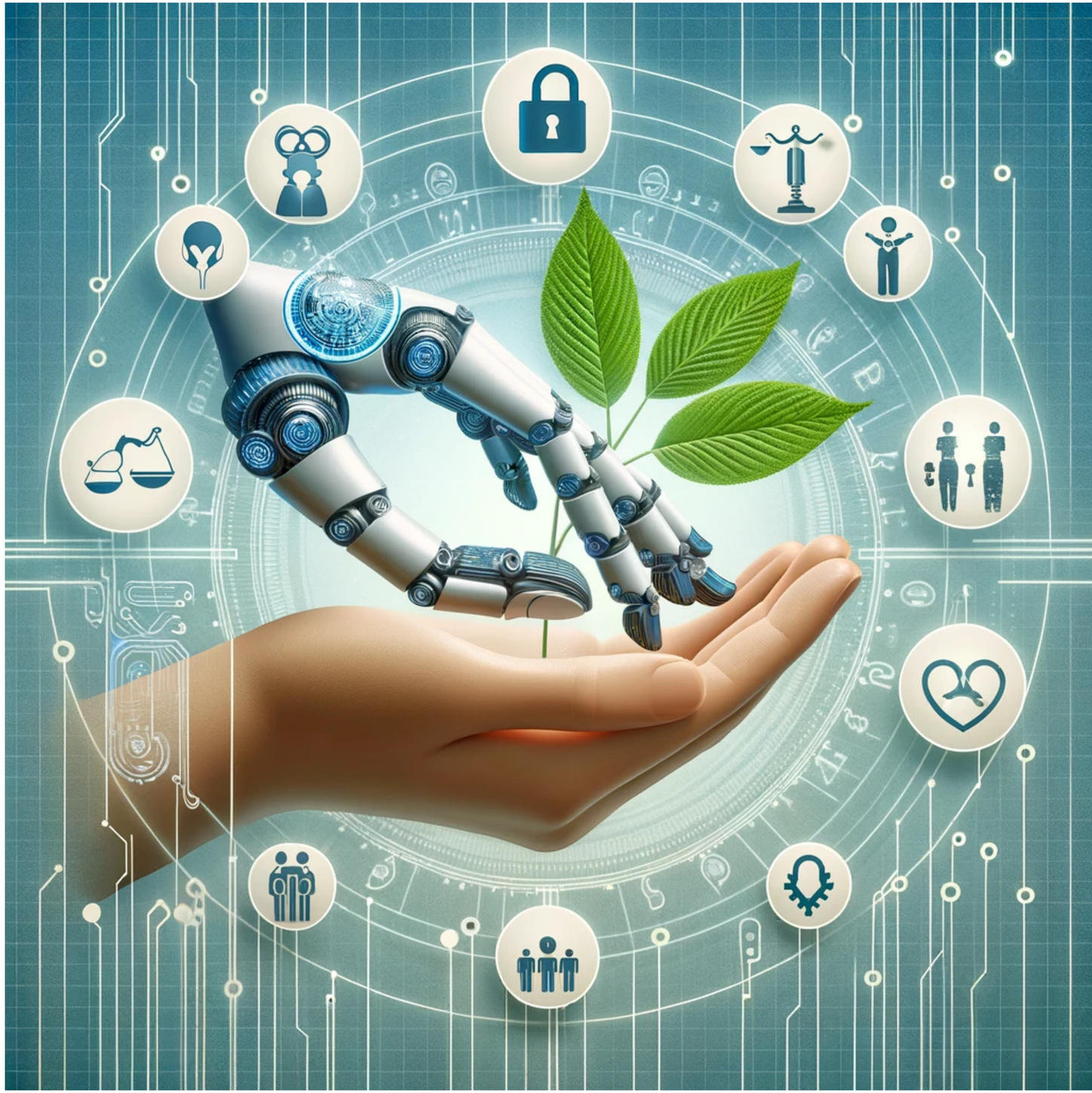
Responsible AI:

Seminar on Fairness, Safety, Privacy and more



https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2024

Discussion

- **Ethical Implications**: How does improving adversarial robustness in AI models intersect with ethical considerations? For instance, does making a model more robust also make it more or less likely to propagate biases or misinformation?
- **Trade-offs**: There are inherent trade-offs between adversarial robustness, model performance, and computational efficiency. How do these trade-offs impact the ethical deployment of these models?
- **Transparency and Explainability**: How can transparency and explainability in AI models help in understanding and mitigating adversarial attacks? Is there a tension between the complexity required for robustness and the need for understandable models?
- **Fairness and Equity in Robust AI Systems**: In what ways might efforts to increase adversarial robustness impact the fairness and equity of AI systems? How can we ensure that these efforts do not exacerbate existing inequalities?
- **LLMs**: How do adversarial attacks on large language models differ from those on other types of machine learning models, and what unique challenges do they present?
- **Responsibility and Accountability**: Who should be held accountable for failures in AI systems due to adversarial attacks – the developers, the users, or the AI itself?
- **Global Perspectives on AI Robustness**: How do perspectives on AI robustness and ethics vary across different cultures and countries? What can be learned from these diverse viewpoints?

Differential Privacy

Foundations and Applications

Part I: Foundation

- Privacy and Attack models
- Differential Privacy
- Common algorithms

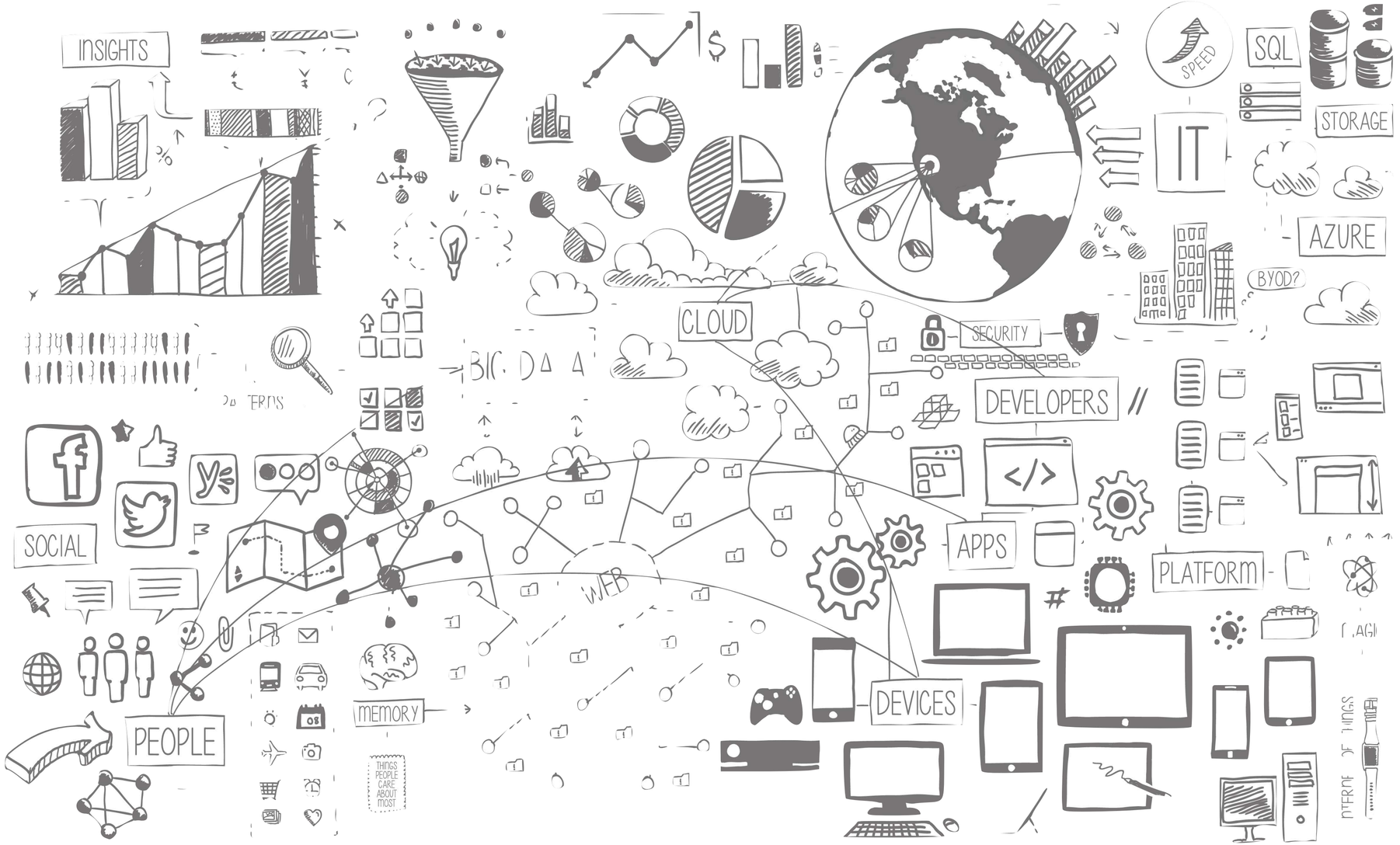
Part II: DP issues

- Consistency

Ferdinando Fioretto | University of Virginia

Part I: Foundation

We live in a data-driven world



Aggregated Personal Data is Invaluable

Online advertising

Medical applications

Mobility analysis

Releasing Personal Data can be Harmful Beyond Your Consent

Would your relatives consent?

Major DNA testing company is sharing genetic data with the FBI

By BLOOMBERG FEB 01, 2019 | 2:55 PM

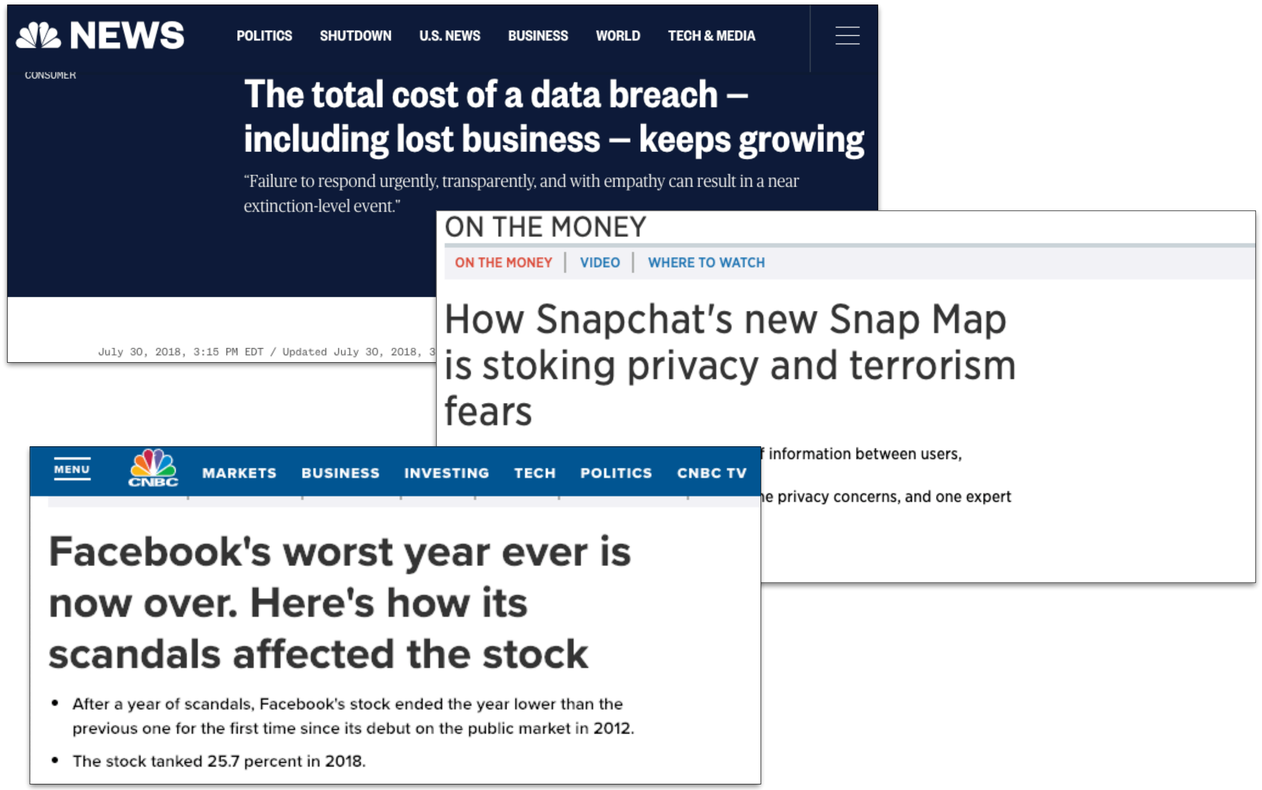
Releasing Personal Data can be Harmful Beyond Your Consent

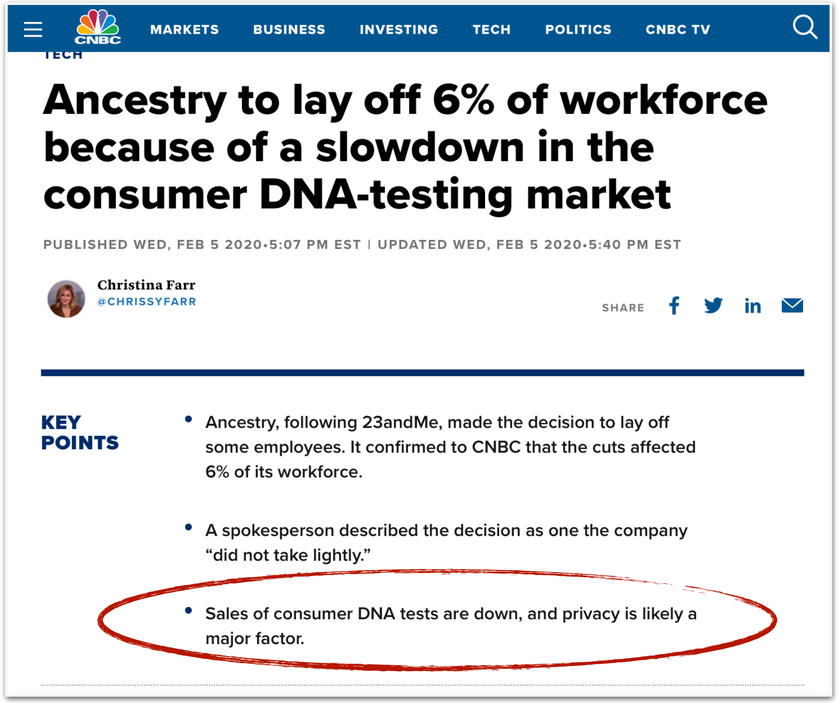
Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

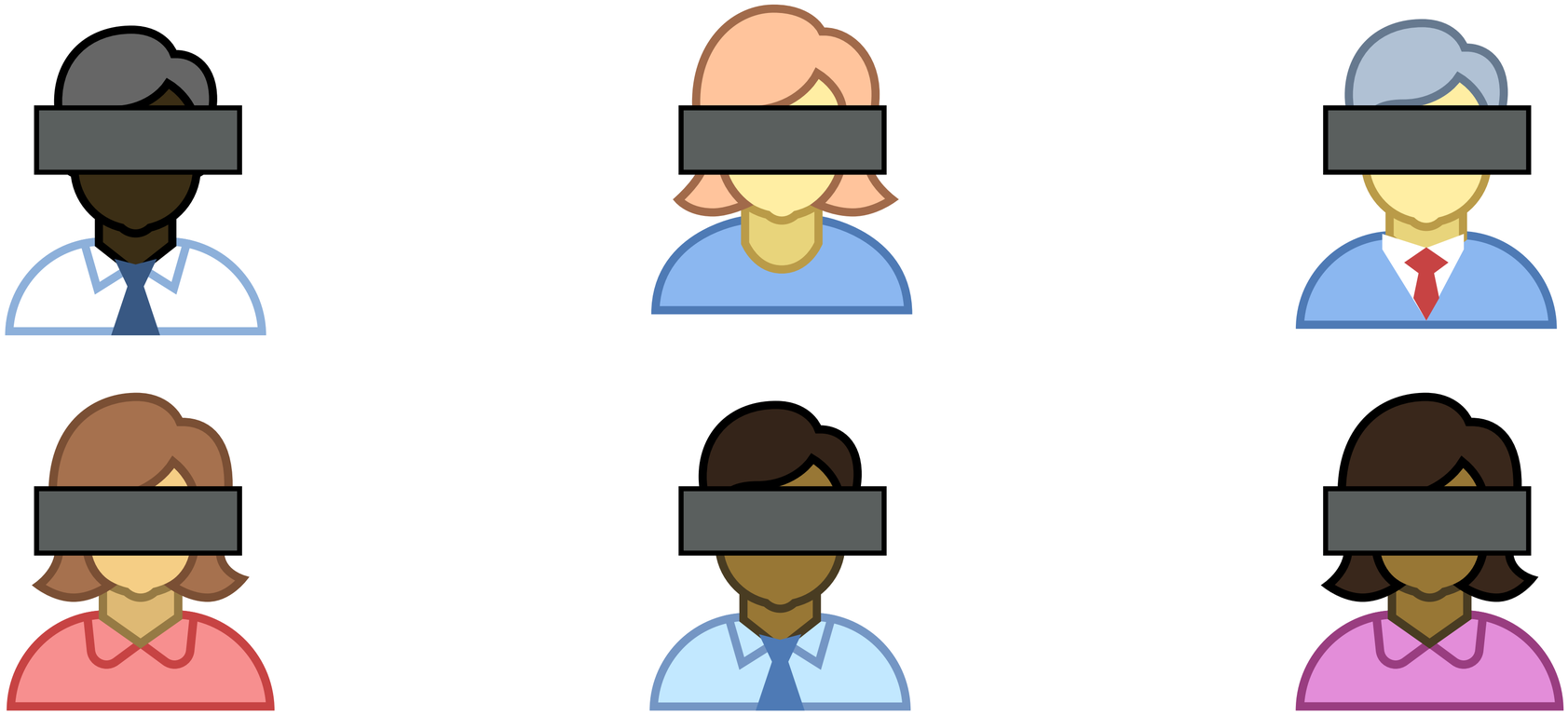
Data Breaches and Cost of Privacy

The average global cost of data breaches is $3.86M (2019) [IBM study]





Data Anonymization



How do you keep survey data private?

- The census bureau is supposed to keep the collected information confidential.
- How can you retain privacy while publishing results about the survey?

user	age	gender
Margaret	31	F
Luis	49	M
Maria	26	F
Carl	19	M
Isabelle	27	F

How do you keep survey data private?

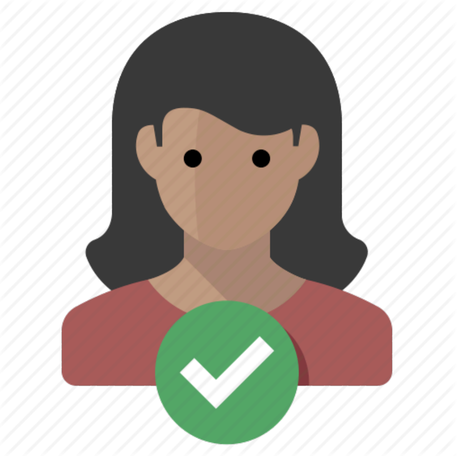
- How can you retain privacy while publishing results about the survey?
- Database Reconstruction Theorem: Every information released contributes to violate the privacy of an individual. The more information it is publicly released the more privacy is violated.

user	age	gender
~~Margaret~~	31	F
~~Luis~~	49	M
~~Maria~~	26	F
~~Carl~~	19	M
~~Isabelle~~	27	F

Avg.: 30.4

Avg.: 34 Male

Avg.: 28 Female



Irit et al. 2003

Anonymization is not Enough

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS

NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

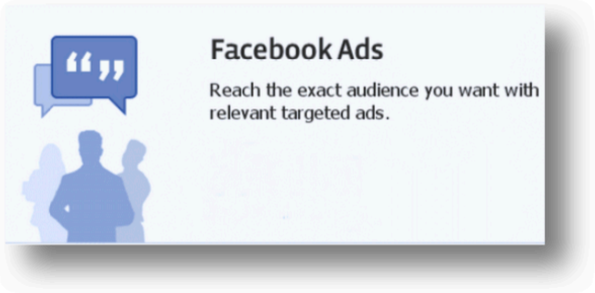
The Scientist » The Nutshell

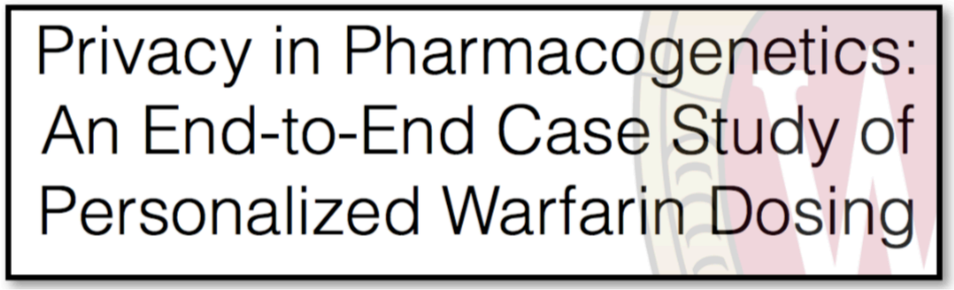
"Anonymous" Genomes Identified

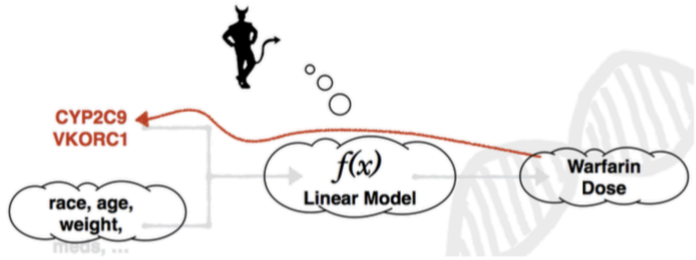
The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013

Anonymization is not Enough



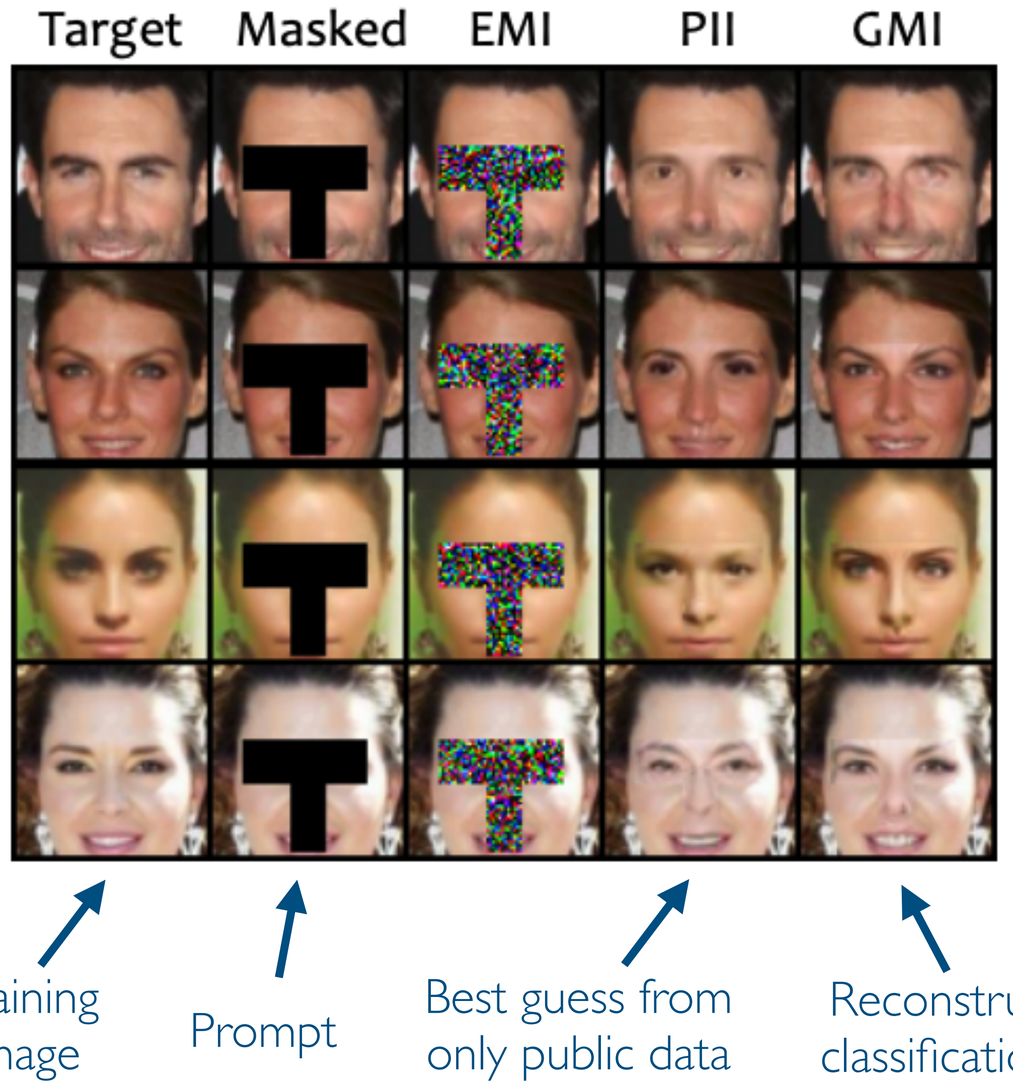




Why Anonymization is Hard?

Model inversion attacks

- Even if you don't release the raw data, the weights of a trained network might reveal sensitive information.
- Model inversion: recover information about the training data from the trained model.
- Example from a face recognition dataset, given a classifier trained on this dataset and a generative model trained on an unrelated dataset of publicly available images.

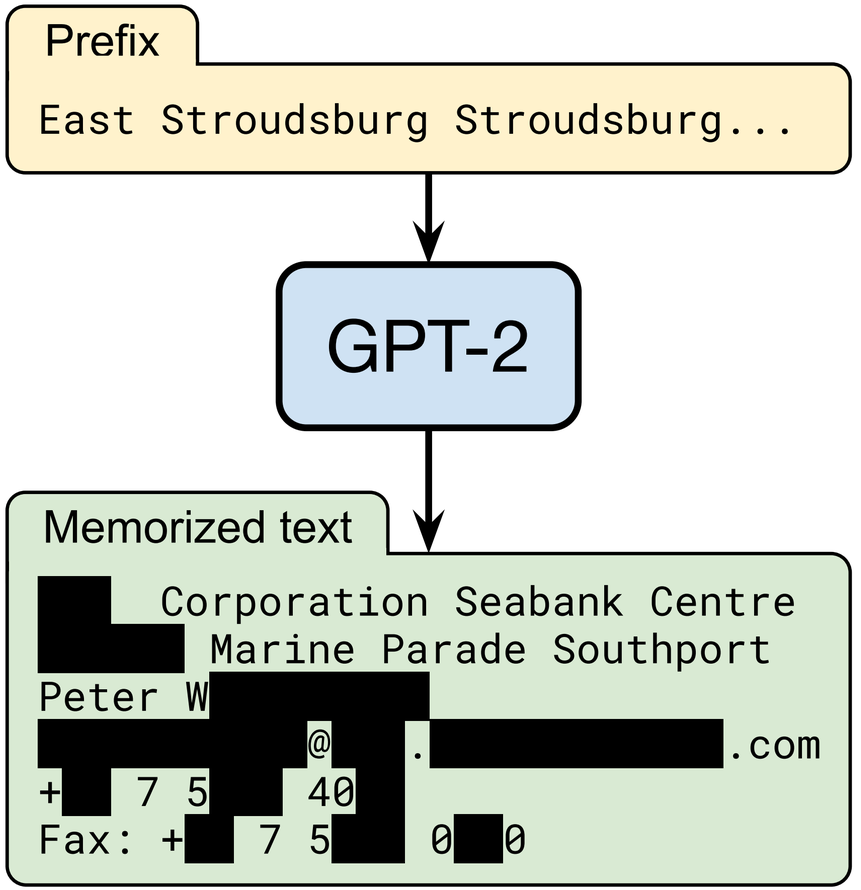


Source: Zhang et al., "The secret revealer: Generative model-inversion attacks against deep neural networks." https://arxiv.org/abs/1911.07135

Why Anonymization is Hard?

Extraction attacks

- Language models trained on scrapes of the public Internet.
- Extraction attack: extracts verbatim text sequences from the model's training data.
- Example from a GPT-2 model. Given query access, it extracts an individual person's name, email address, phone number, fax number, and physical address.



Carlini et al. 2021

Why Anonymization is Hard?

Reconstruction attacks

United States™ Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

308,745,548 people in 2010 release which implements some "protection"

Commercial databases

Linkage Attacks — Results from UC Census:

- Census blocks correctly reconstructed in all 6,207,027, inhabited blocks.
- Block, sex, age, race, ethnicity reconstructed:
  - Exactly: 46% of population (142M).
  - Allowing age +/- 1 year: 71% of population (219M).
- Name, block sex, age, race, ethnicity:
  - Confirmed re-identification: 38% of population.

Ramachandran et al. 2012

McKenna et al. 2018

Why Anonymization is Hard?

Needs for guarantees

- It's hard to guess what capabilities attackers will have, especially decades into the future.
  - **Analogy with crypto**: Cryptosystems today are designed based on what quantum computers might be able to do in 30 years.
  - To defend against unknown capabilities, we need mathematical guarantees.
- **Want to guarantee**: no individual is directly harmed (e.g. through release of sensitive information) by being part of the database, even if the attacker has tons of data and computation.

Part I: Foundation

Privacy Definitions

Ferdinando Fioretto | University of Virginia

Privacy is NOT

Encryption

- **Encryption:**
- Alice sends a message to Bob s.t. Trudy (the attacker) does not learn the message.
- Bob should get the correct message

- **Statistical Dataset Privacy :**
- Trudy can access the dataset
  - It must learn aggregate statistics, but
  - Must not learn new info about the individuals in the dataset

Privacy is NOT

Secure multiparty computation

- **Secure Multiparty Computation:**
- A set of agents, each with private input x_i
- Want to compute a function $f(x_1, \ldots, x_k)$
- Each agent can learn the true answer, but must not learn any other info than what can be inferred from their private input and the answer.

- **Statistical Dataset Privacy :**
- The function output must not disclose individual inputs.

What is privacy?

Privacy breach

A private mechanism M(D) that allows an unauthorized party P to learn sensitive information about any individual in D that P could not have learnt with access to M(D)

What is privacy?

The smoking causes cancer case

Alice

Alice has Cancer

Is this a privacy breach?

What is privacy?

Privacy Breach (revisited)

A private mechanism M(D) that allows an unauthorized party P to learn sensitive information about any individual **x in D** that P could not have learnt with access to M(D), **if x was not in D**

Brief Summary

- Statistical dataset privacy is the problem of releasing aggregate statistics while not disclosing individual records
- The problem is distinct from encryption and secure computation
- Defining privacy is non-trivial:
  - Requirements include resilience to background knowledge, as well as composition and closure under post-processing.

Randomized Response

Early privacy

Randomized response is a survey technique that ensures some level of privacy.

Example: Have you ever dodged your taxes?

1. Flip a coin.
2. If the coin lands Heads, then answer truthfully.
3. If it lands Tails, then flip it again.
    1. If it lands Heads, then answer Yes.
    2. If it lands Tails, then answer No.

Probability of responses

	Yes	No
Dodge	3/4	1/4
No Dodge	1/4	3/4

Randomized Response

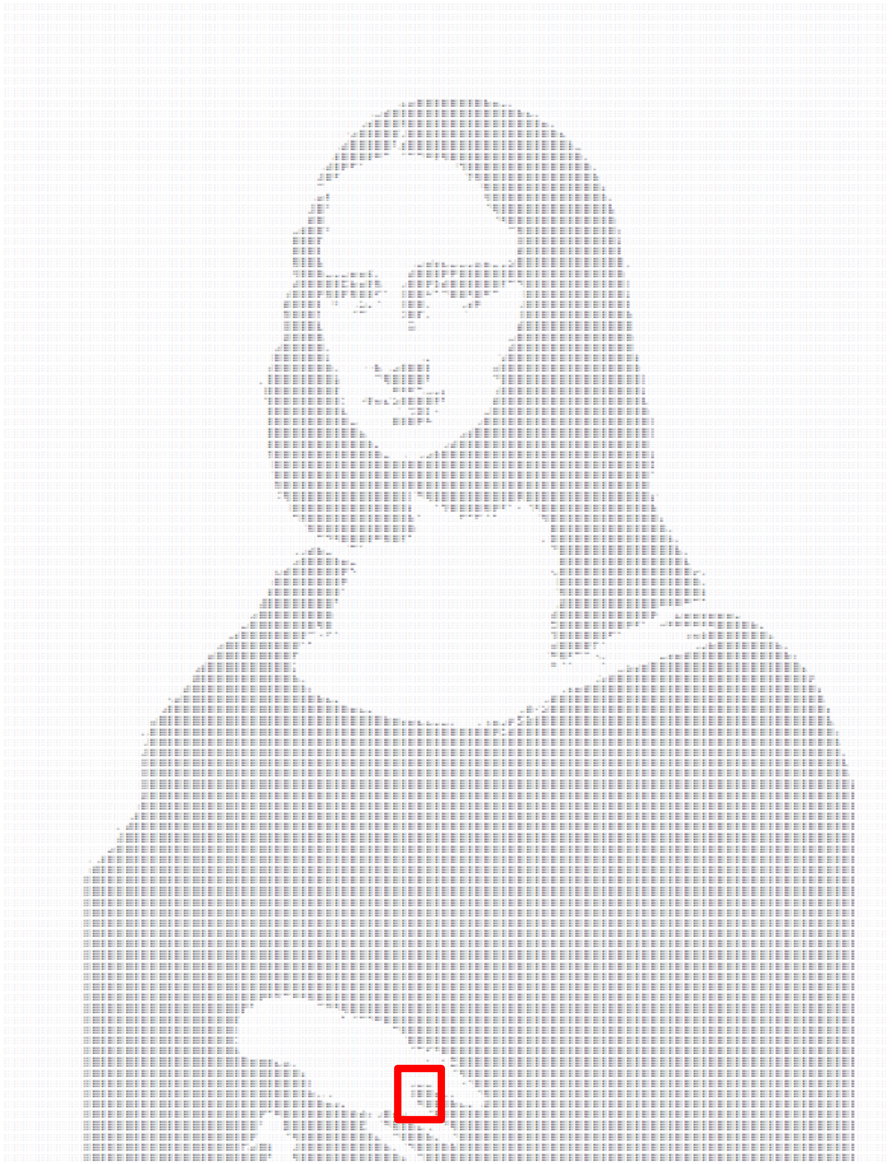
Early privacy

- Tammy, the tax investigator assigns a prior probability of 0.02 to Bob having dodged his taxes. Then she noticed he answered **Yes** to the survey.
- What is her posterior probability?

$$\begin{aligned}\Pr(\text{Dodge} \mid \text{Yes}) &= \frac{\Pr(\text{Dodge})\Pr(\text{Yes} \mid \text{Dodge})}{\Pr(\text{Dodge})\Pr(\text{Yes} \mid \text{Dodge}) + \Pr(\text{NoDodge})\Pr(\text{Yes} \mid \text{NoDodge})} \\ &= \frac{0.02 \cdot \frac{3}{4}}{0.02 \cdot \frac{3}{4} + 0.98 \cdot \frac{1}{4}} \\ &\approx 0.058\end{aligned}$$

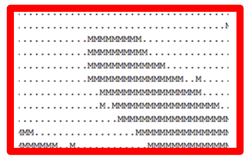
- Tammy's beliefs haven't shifted too much.
- Randomness turns out to be a useful technique for preventing information leakage.

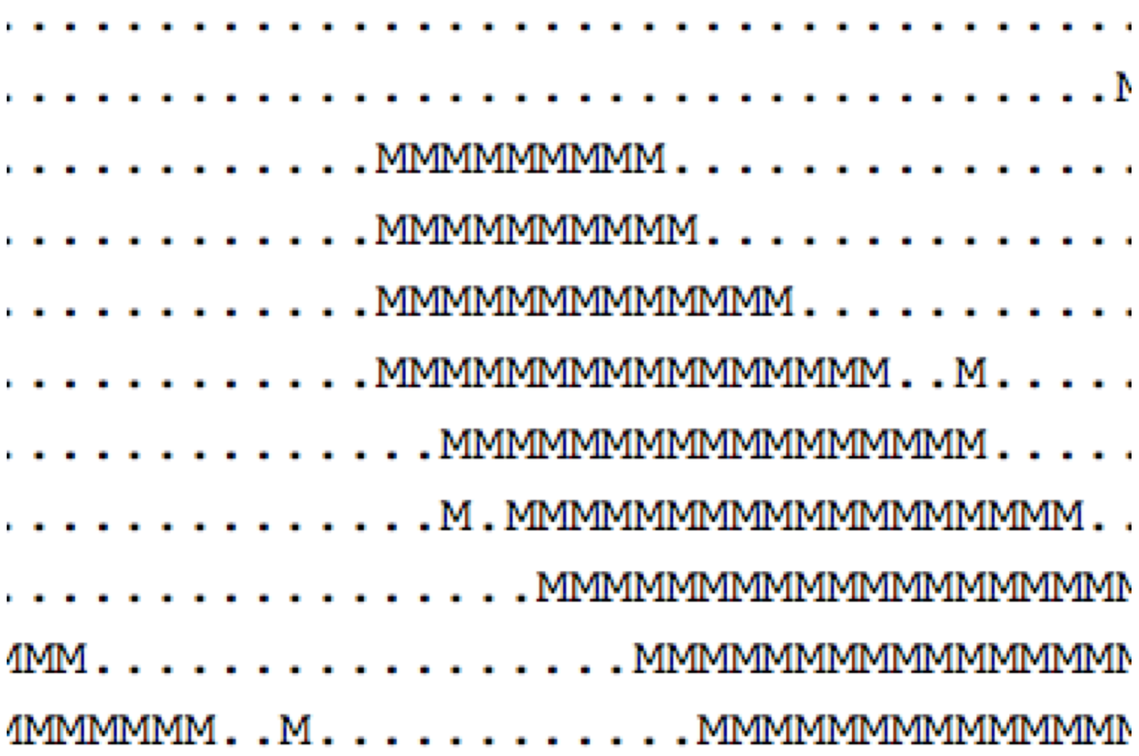
A Metaphor for Private Learning



A Metaphor for Private Learning

An individual's training data

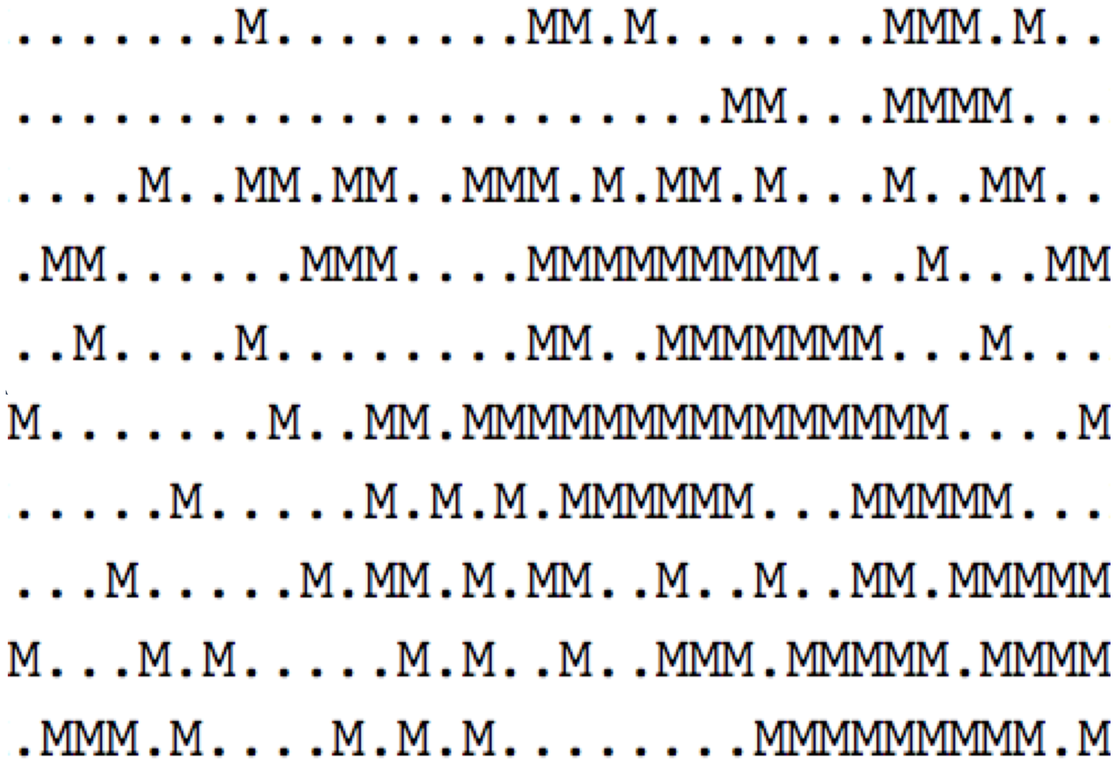




A Metaphor for Private Learning

An individual's training data

Each bit is flipped with probability 51%



A Metaphor for Private Learning

Big picture remains!

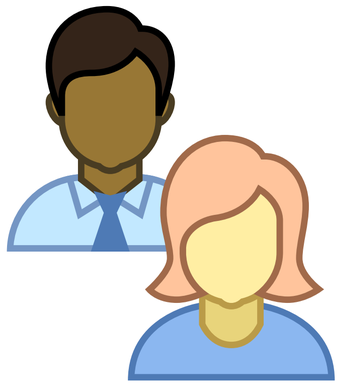
Part I: Foundation

Differential Privacy

Ferdinando Fioretto | University of Virginia

Differential Privacy

Intuition

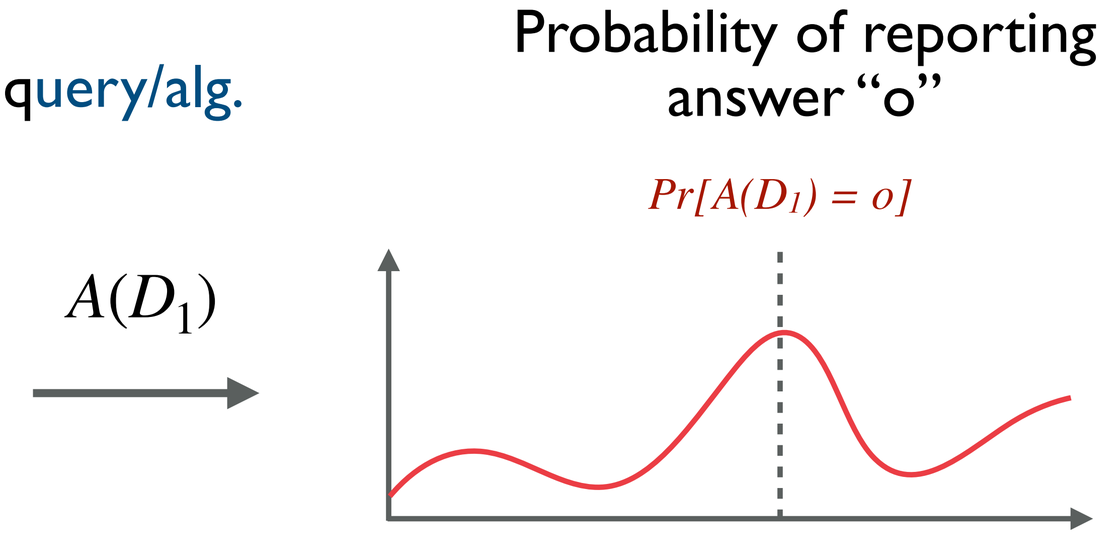


Dataset

D_1

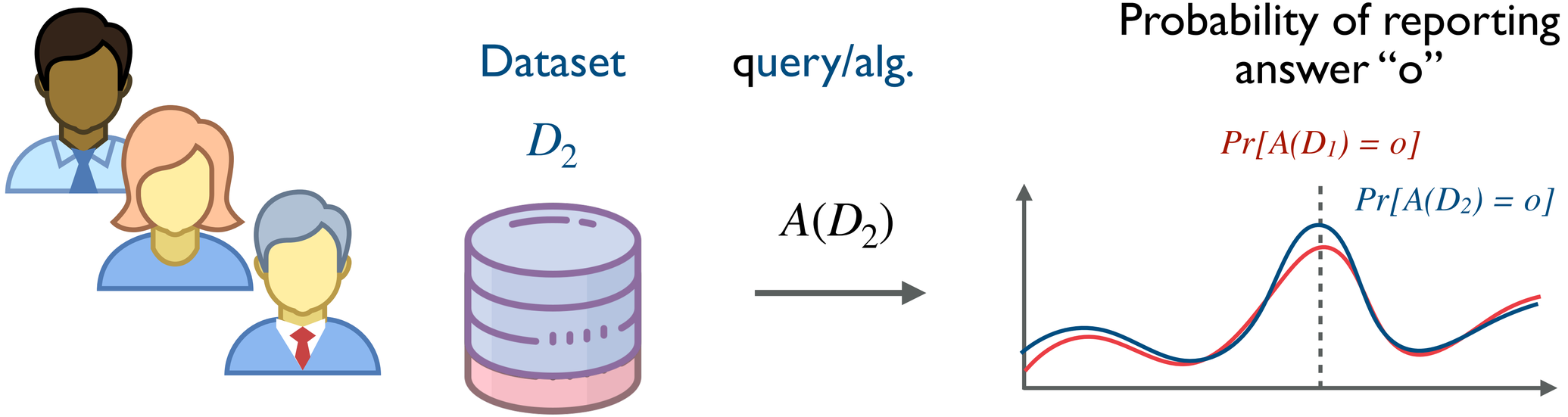
query/alg.

$A(D_1)$



Differential Privacy

Intuition



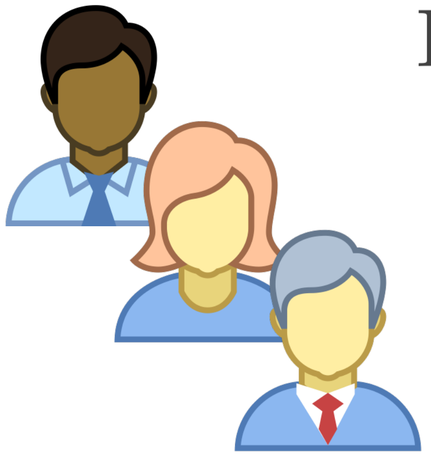
Should not be able do distinguish between weather the input was D1 or D2 no matter what the output is

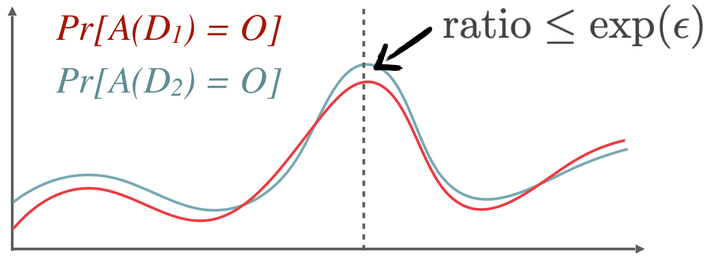
Differential Privacy

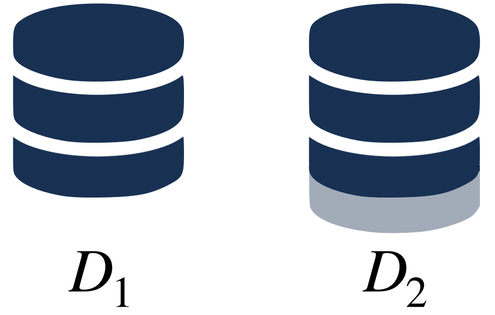
Definition

A randomized algorithm $\mathscr{A}$ is ε-differentially private if, for all pairs of inputs D_1, D_2, differing in one entry, and for any output O:

$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$



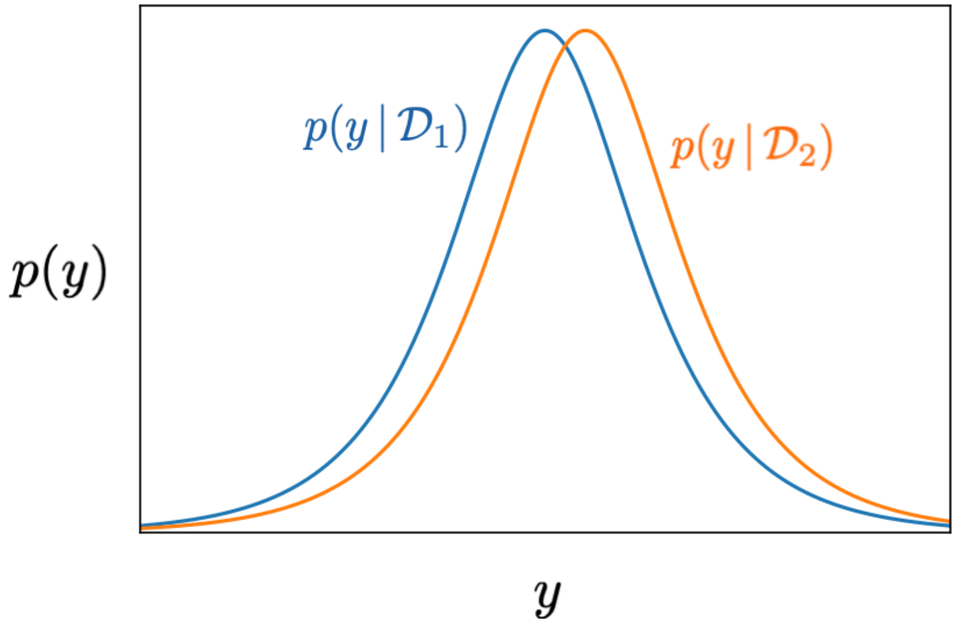


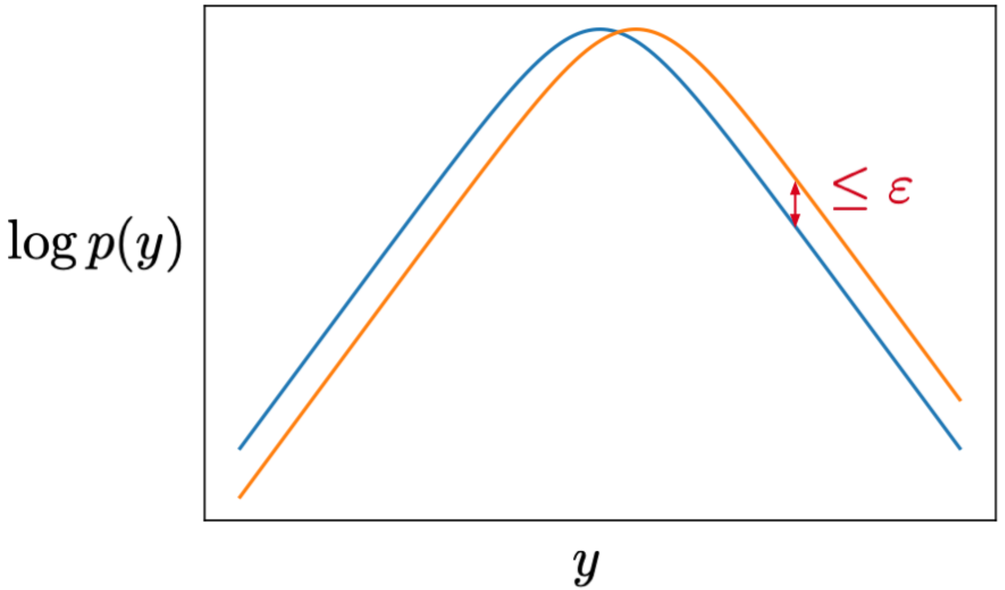


Intuition: An adversary should not be able to use output O to distinguish between any D_1 and D_2

Differential Privacy

Visually





- Notice that the tail behavior is important!

Differential Privacy

What can we infer?

- Alice is an attacker who wants to figure out if Bob(x) is in the cancer database D.
  Her prior for him being in the database is 0.4.
  D is ε-differentially private. She makes a query and gets back $y = M(D)$.

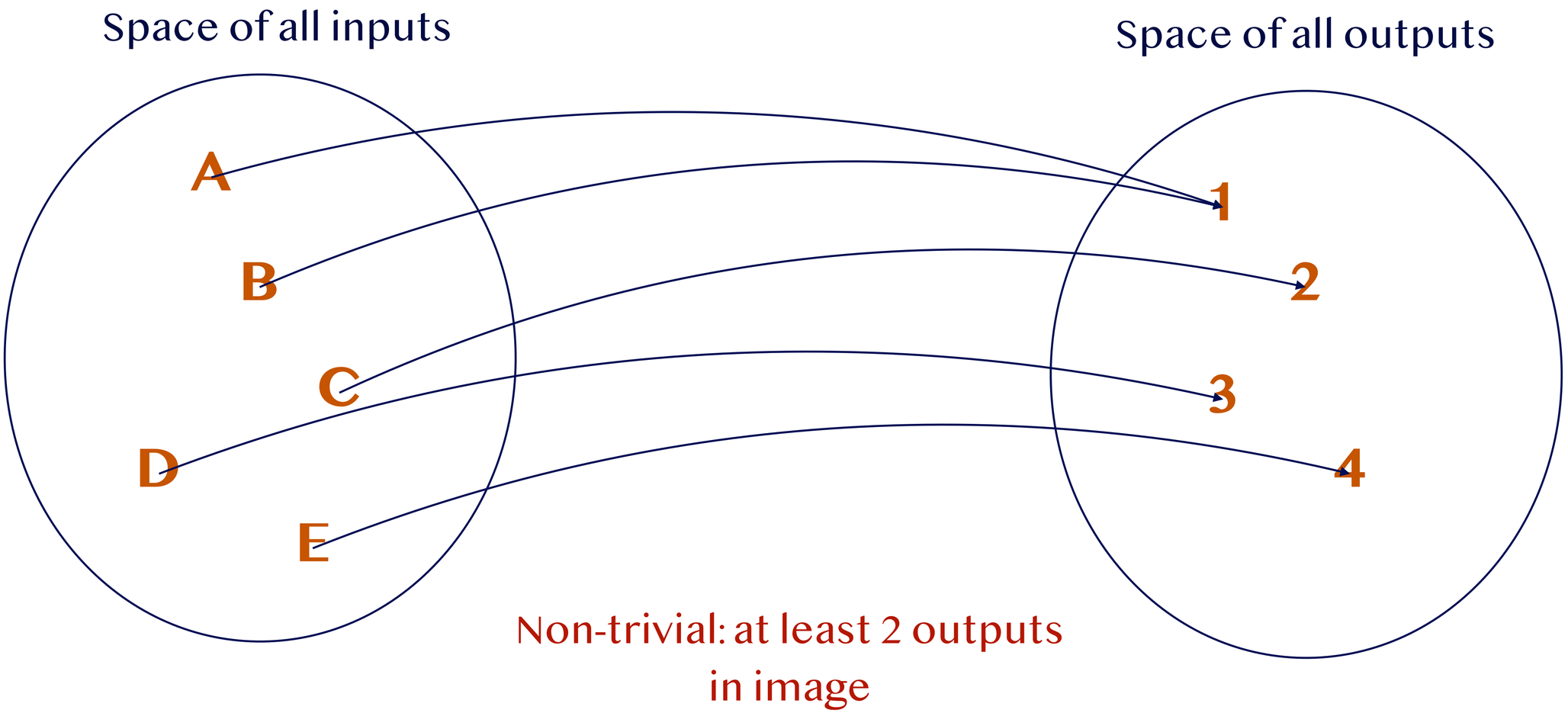
- After observing y, she computes the posterior using Bayes' rule:

$$\begin{aligned}
\Pr(x \in \mathcal{D} \mid y) &= \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \Pr(x \notin \mathcal{D}) \Pr(y \mid x \notin \mathcal{D})} \\
&\geq \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D}) \Pr(y \mid x \in \mathcal{D})} \\
&= \frac{\Pr(x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D})} \\
&\geq 0.4 \exp(-\varepsilon)
\end{aligned}$$

- Similarly $\Pr(x \in D \mid y) \leq 0.4 \exp(\epsilon)$ so Alice hasn't learn much about Bob.

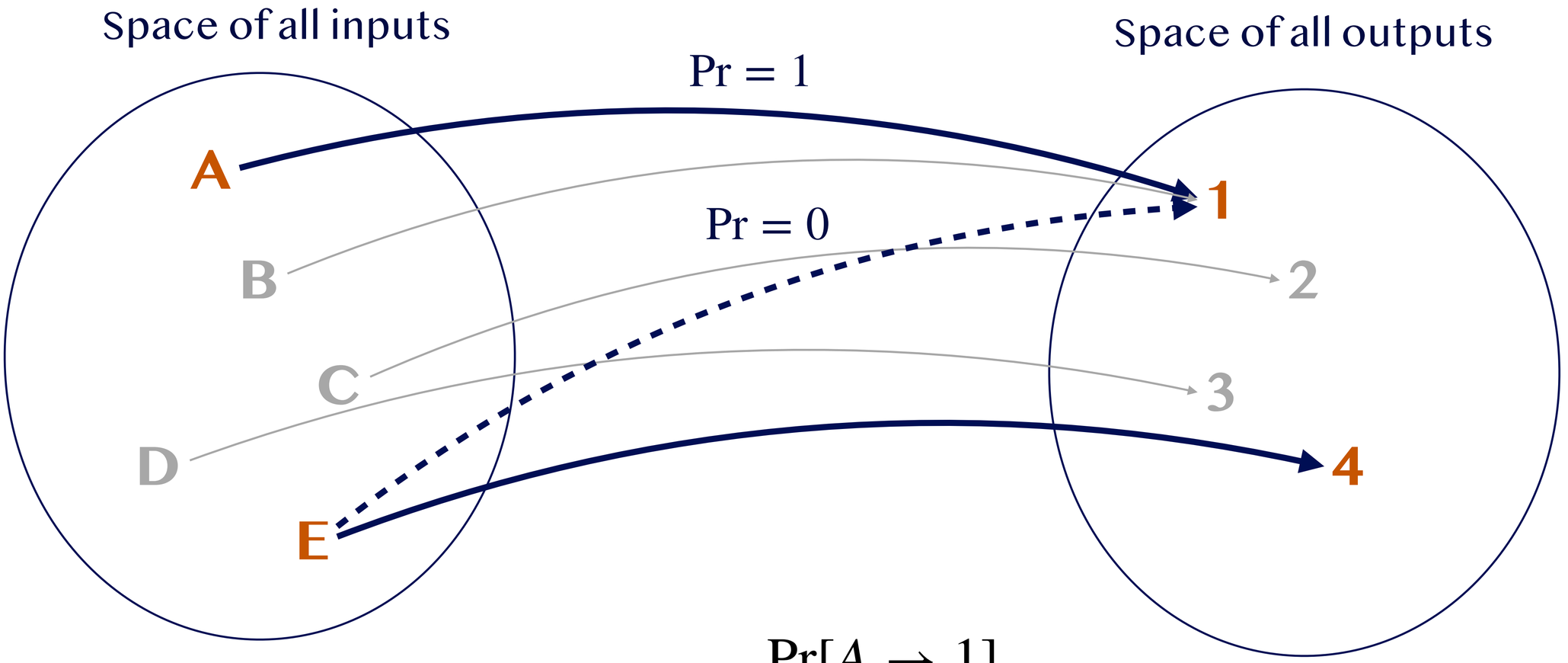
Differentially Private Algorithms

Can Non-trivial Deterministic Algorithms Satisfy DP?



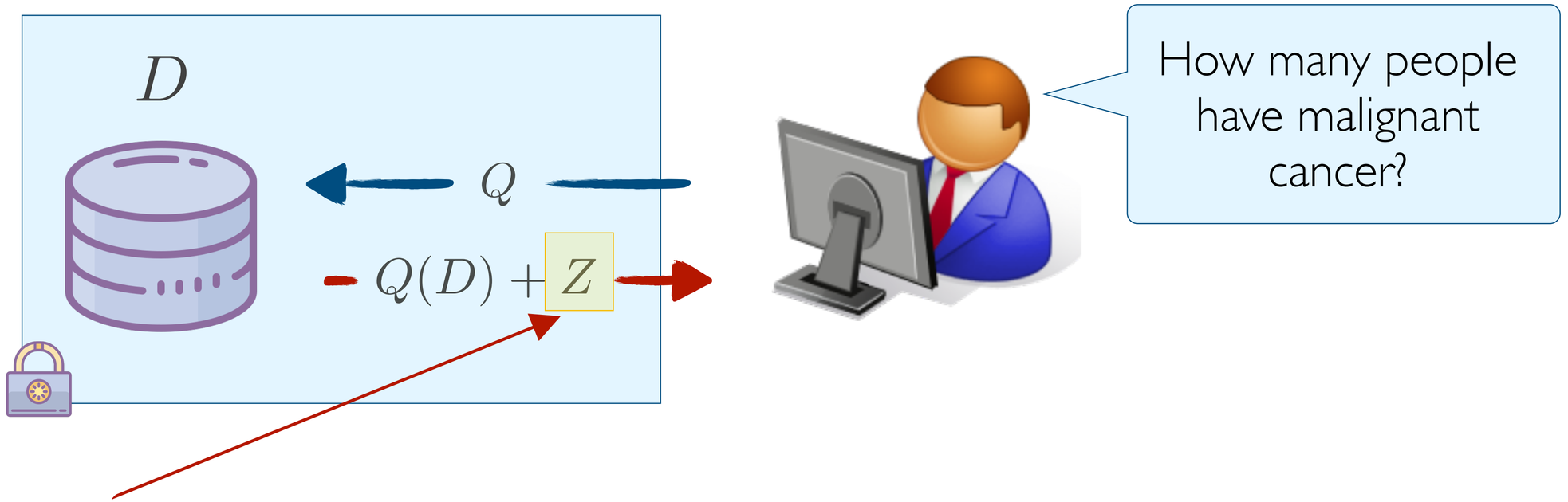
Differentially Private Algorithms

Can Non-trivial Deterministic Algorithms Satisfy DP?



$$\Pr[E \to 1] = 0 \quad \text{implies} \quad \frac{\Pr[A \to 1]}{\Pr[E \to 1]} = \infty$$

How do we design DP algorithms?

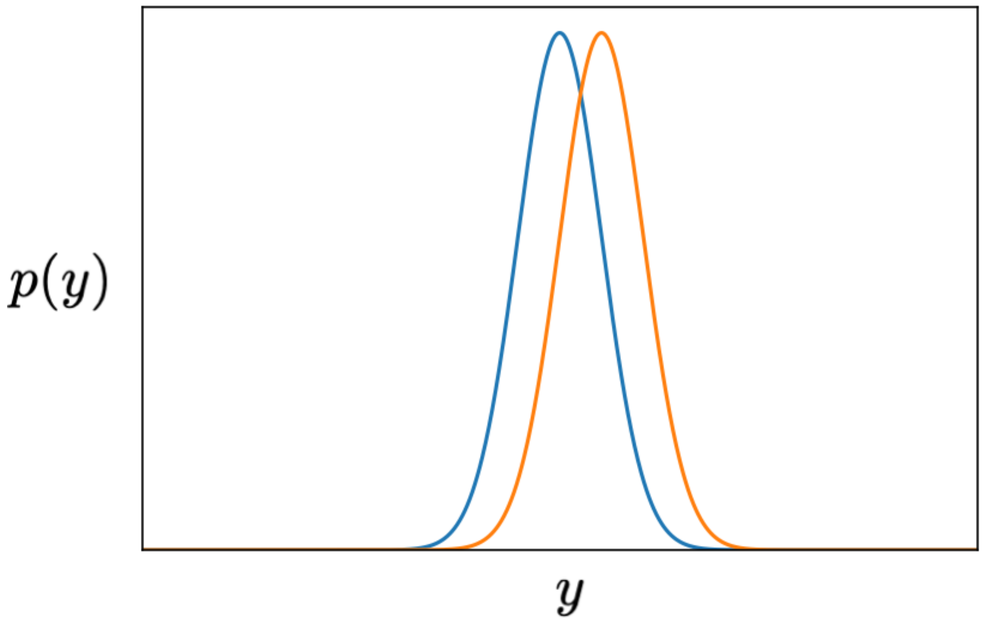


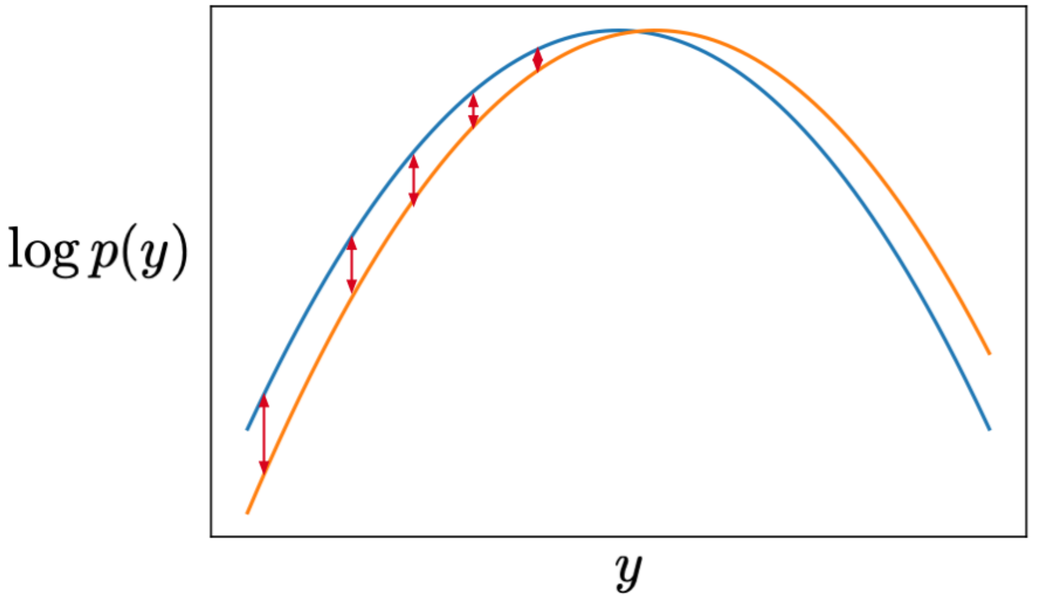
Add noise to the true answer such that:

1. Each answer does not leak too much information about the dataset
2. Noisy answers are close to the original ones

What kind of noise?

Gaussian Noise (first attempt)





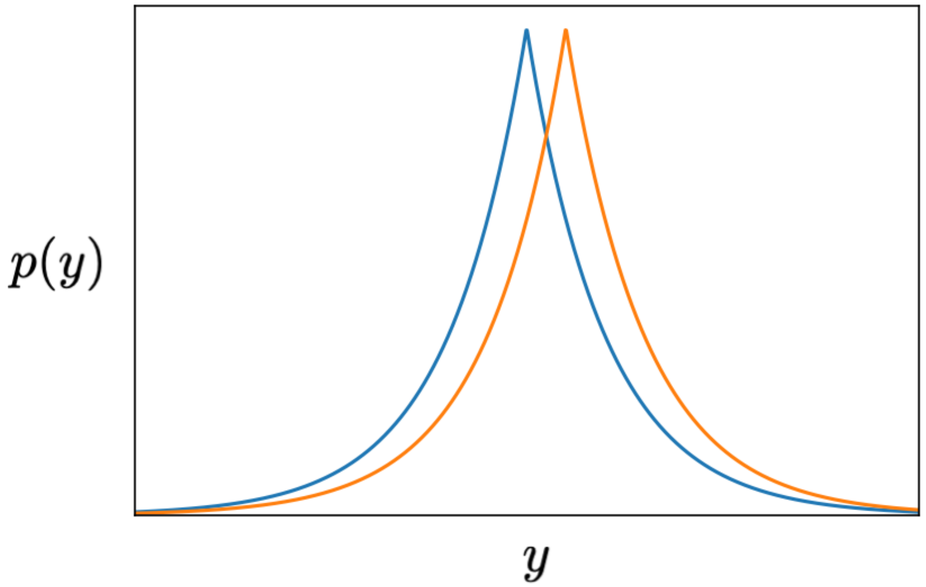
Gaussian noise violates our definition, but only because of the tails. It satisfies a different definition of differential privacy which allows violating the ε constraint with small probability, but that's beyond the scope of this slide.

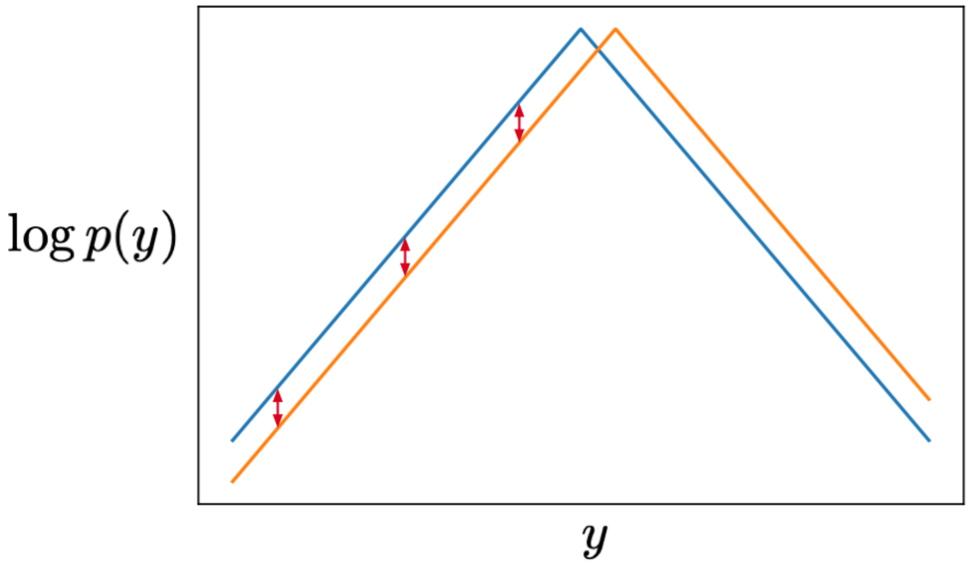
What kind of noise?

Laplace noise

The Laplace distribution $p(y;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|y-\mu|}{b}\right)$

Variance $= 2b^2$ (b is a parameter determining the scale of the distribution)





Is exactly what we need!

Laplace Mechanism

- **Global sensitivity:** Let f be a deterministic vector-valued function of a dataset. The L^1 ***sensitivity of f*** *is defined as:*

$$\Delta f = \max_{\substack{\mathcal{D}_1, \mathcal{D}_2 \\ \text{neighbours}}} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1.$$

Recall that $\|x\|_1 = \sum_i |x_i|$

- **Laplace mechanism** : returns a vector **y** whose entries are independently sampled from Laplace distributions

$$y_i \sim \text{Laplace}\left(f(\mathcal{D})_i, \frac{\Delta f}{\varepsilon}\right),$$

where $f(D)_i$ denotes the i-th entry of $f(D)$

Noise **calibrated** to the privacy requirement:
Higher sensitivity functions and tighter privacy constraints imply more noise

Sensitivity: Count query

D

Cancer
Y
N
N
N
Y
N
N
Y

3

D'

Cancer
Y
N
N
N
Y
N
N
Y
Y/N

3 or 4

How many people have malignant cancer?

$$\tilde{y} = f(D) + \text{Lap}\left(\frac{1}{\epsilon}\right)$$

What is the sensitivity of this query?

$$\Delta f = \max_{\substack{\mathcal{D}_1, \mathcal{D}_2 \\ \text{neighbours}}} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1.$$

$$\max(|3-3|, |3-4|) = 1$$

Sensitivity: SUM query

- Suppose all values x are in $[a, b]$, with $a, b \geq 0$
- What is the sensitivity of the SUM $= \sum_{x \in D}$ query?
- Sensitivity: **b**

E.g., a=3, b=5

value
3
4
5
3
5
4

24

value
3
4
5
3
5
4
[3, 5]

[27, 29]

Laplace Mechanism Privacy

- Consider neighboring datasets D1 and D2
- Consider some output O
-

$$f(x \mid \mu = 0, b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

$$\frac{\Pr[M(D_1) = O]}{\Pr[M(D_2) = O]} = \frac{\Pr[Q(D_1) + Z = O]}{\Pr[Q(D_2) + Z = O]}$$

$$= \frac{\exp(-|O - Q(D_1)|/b)}{\exp(-|O - Q(D_2)|/b)}$$

By triangle inequality

$$\leq \exp(|Q(D_1) - Q(D_2)|/b)$$

$$\leq \exp(\Delta_Q/\epsilon) = \exp(\epsilon)$$

$$\boxed{b = (\Delta_Q/\epsilon)}$$

Laplace Mechanism

Example

- Example: What fraction of Danish have blue eyes?
- Mechanism returns the counts (ξ_1, ξ_2) of Danish with and without blue eyes, plus Laplace noise.
- We'd like to satisfy a privacy constraint of $\varepsilon = 0.1$.
  How much Laplace noise should we add?
- Ans: $\Delta f / \varepsilon = 1/0.1 = 10$.
- The noise scale is **independent of the population size!**
- I.e., you can answer the query to within about ± 10 people, out of the population of Denmark. So you can obtain very accurate answers to queries over large populations.

.

Error of the Laplace Mechanism

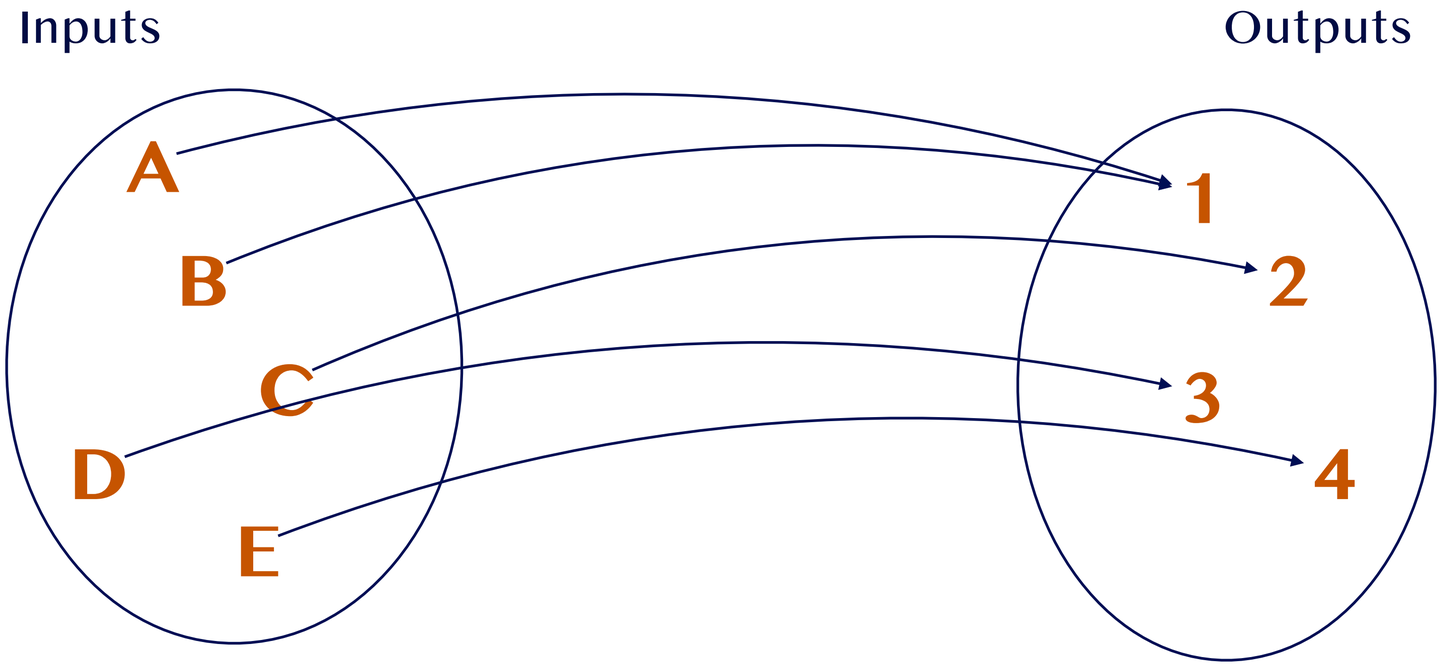
- Laplace mechanism works for any function that returns a real number
- Mean squared error:

$$\mathbb{E}\left[\tilde{x} - x\right]^2 = \mathrm{Var}\left(\mathrm{Lap}\left(\frac{\Delta_f}{\epsilon}\right)\right) = 2\left(\frac{\Delta_f}{\epsilon}\right)^2$$

Exponential Mechanism

Discrete mappings

- Consider some mapping *f (can be deterministic or probabilistic):*



How to construct a differentially private version of *f?*

Exponential Mechanism

- Suppose the goal is to make a decision f(D).
- We have a **scoring function** $\mathscr{L} : \mathscr{D} \times \mathscr{R} \rightarrow \mathbb{R}_+$

- **Example**:
  - D = dataset of nationality of a set of people
  - f(D): most frequent nationality in D
  - $\#_D(o)$: number of people in D with nationality "o"
  - $\mathscr{L}(D, o) = |\#_D(o) - \#_D(f(D))|$

Exponential Mechanism

- Given a function $f : \mathscr{D} \to \mathscr{R}$ and a loss function $\mathscr{L} : \mathscr{D} \times \mathscr{R} \to \mathbb{R}_+$
- Randomly sample an output o from $\mathscr{R}$ with probability:

$$\Pr(Y = y) \propto \exp\left(-\frac{\epsilon}{2\Delta\mathcal{L}}\mathcal{L}(D, y)\right)$$

- Where $\Delta_{\mathscr{L}} = \max_{o \in \mathscr{R}} \max_{D, D' \in \mathscr{D}} |\mathscr{L}(D, o) - \mathscr{L}(D', o)|$ is the sensitivity of the loss function
- The result is basically a softmax of $-\mathcal{L}$
- Note: for every output o, Pr[o is selected] > 0.

McSherry et al. 2007

Exponential Mechanism

- The exponential mechanism is ϵ-DP
- For two neighboring datasets D_1, D_2, and any value y

$$\frac{p(y \mid \mathcal{D}_1)}{p(y \mid \mathcal{D}_2)} = \frac{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}}$$

$$= \underbrace{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}}_{\leq \exp(\varepsilon/2)} \cdot \underbrace{\frac{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}_{\leq \exp(\varepsilon/2)}$$

- Hence $\frac{p(y \mid D_1)}{p(y \mid D_2)} \leq \exp(\epsilon)$

McSherry et al. 2007

Part I: Foundation

Properties of Differential Privacy

Ferdinando Fioretto | University of Virginia

Privacy Leakage

Dinur/Nissim Result

- To ensure some utility on the query answers, some information about each individual in the dataset must be leaked.
- We can only hope to bound the amount of disclosure.
- Therefore: There is a limit on the number of queries that can be answered.

Privacy Leakage

Dinur/Nissim Result

- To ensure some utility on the query answers, some information about each individual in the dataset must be leaked.
- We can only hope to bound the amount of disclosure.
- Therefore: There is a limit on the number of queries that can be answered.

- **Dinur/Nissim:** Given a dataset D of size n, a vast majority of records in D can be reconstructed when $n \log(n)^2$ queries are answered privately.
- True even if each query is altered with error up to $o(\sqrt{n})$

Dinur et al. 2003

Composition

Sequential Composition

- Given $A_1, \ldots, A_n$ algorithms that access a dataset D such that each satisfy ϵ_i-DP
- Then running all n algorithms sequentially satisfies:

$$\left(\sum_{i=1}^{n} \epsilon_i \right)\text{-differential privacy}$$

Dwork et al. 2006

Composition

Sequential Composition

- Given $A_1, \ldots, A_n$ algorithms that access a dataset D such that each satisfy ϵ_i-DP
- Then running all n algorithms sequentially satisfies:

$$\left(\sum_{i=1}^{n} \epsilon_i \right)\text{-differential privacy}$$

Parallel Composition

- If $A_1, \ldots, A_n$ are algorithms that access disjoint datasets $D_1, \ldots D_n$ such that each satisfy ϵ_i differential privacy
- Then running all k algorithms in parallel satisfies

$$\max\{\epsilon_1, \ldots, \epsilon_n\}\text{-differential privacy}$$

Dwork et al. 2006

Post-processing Immunity

- Post-processing immunity: If A enjoys ε-differential privacy and g is an arbitrary data-independent mapping, then $g \circ A$ is ε-differential private.

Dwork et al. 2006

Part I: Foundation
Summary

- Privacy breaches and no control on accessing side information
- Differential private algorithm ensure an attacker can't infer if an individual was in or not a dataset, based on any output
- Building blocks:
  - Laplace, exponential mechanism, (many others)
- Composition rules help using building blocks to create complex algorithms

Responsible AI:
Seminar on Fairness, Safety, Privacy and more

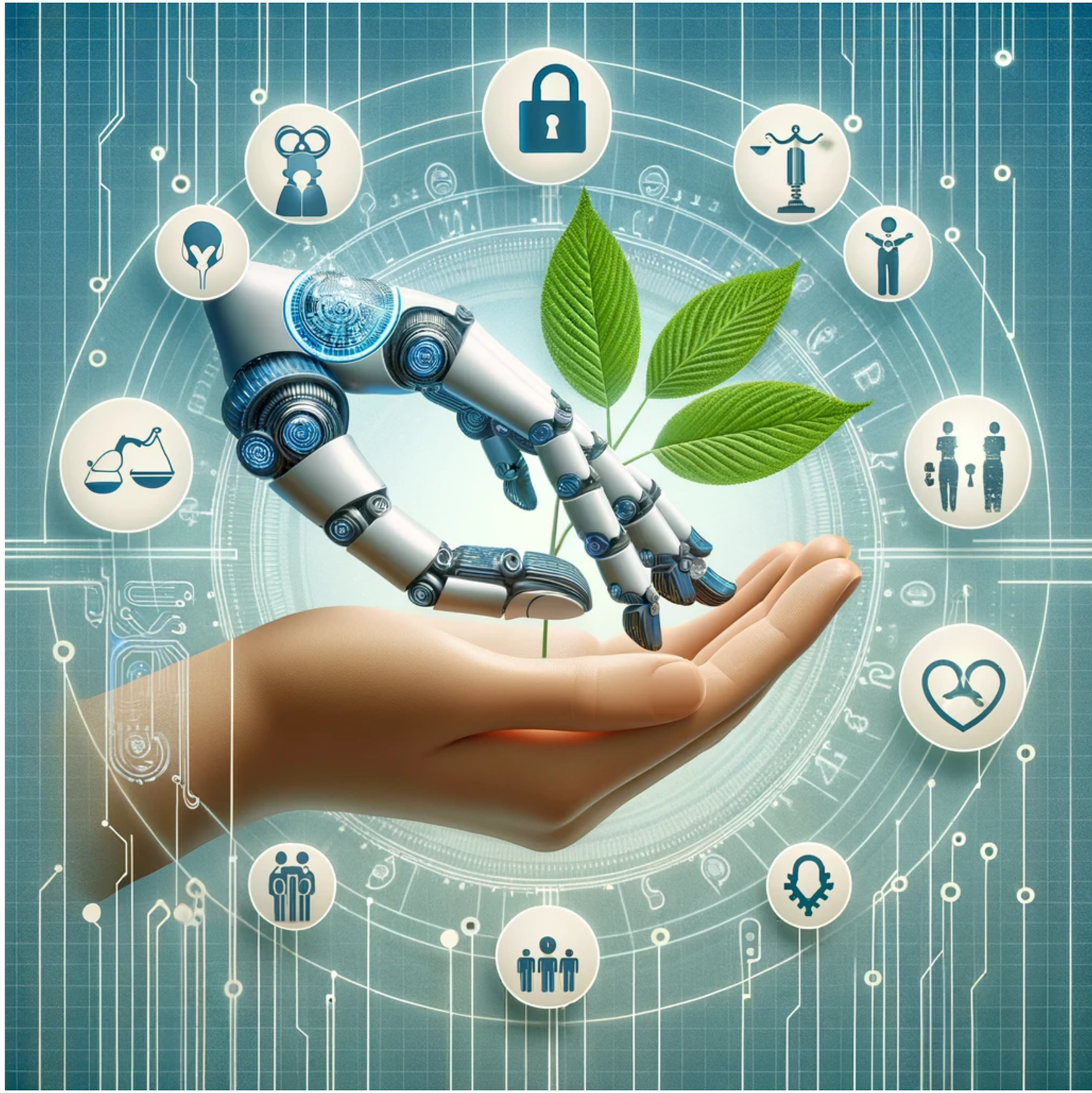
Thank you!

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2024

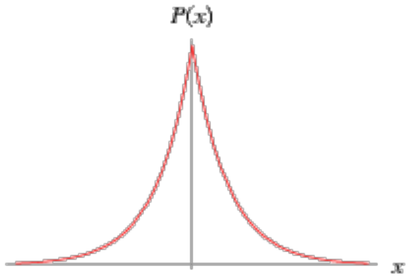


Part II: Algorithms and Consistency Issues

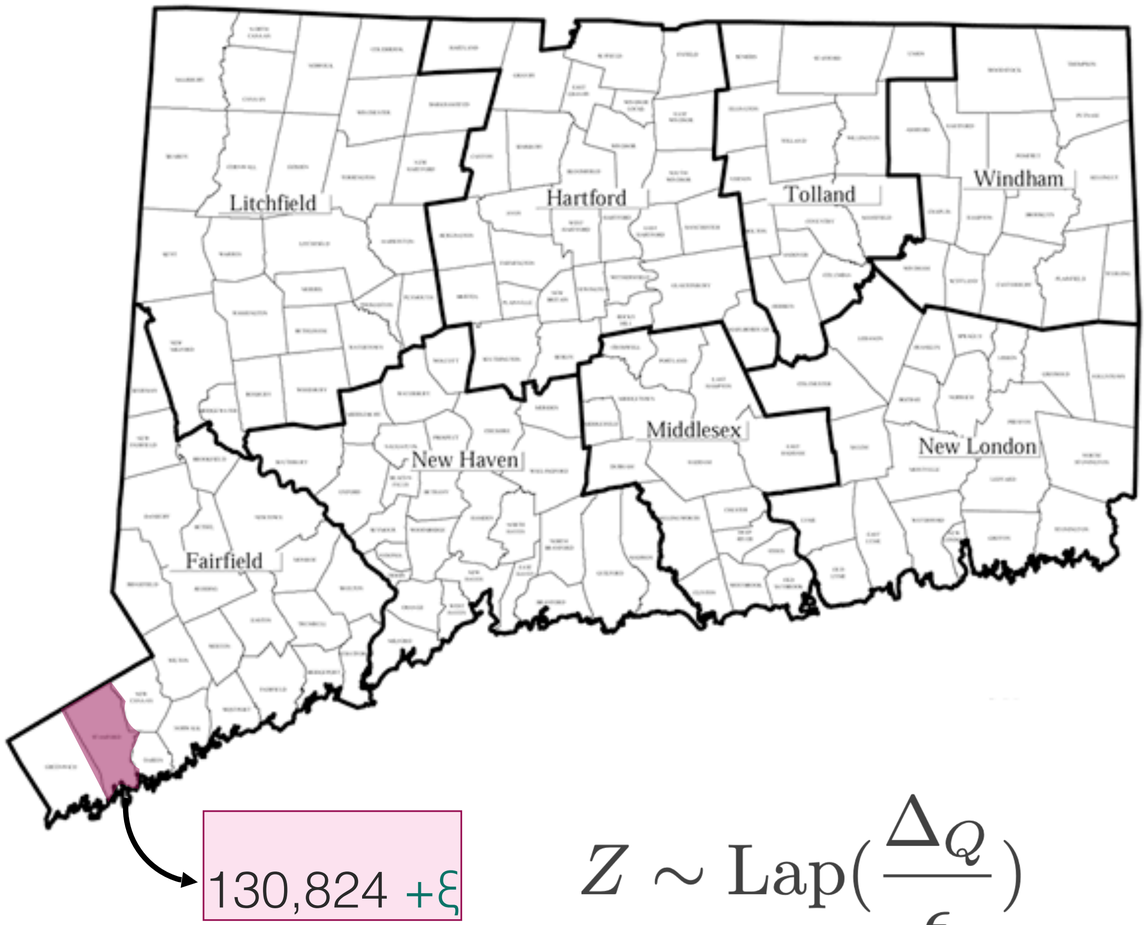
The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

1. Census blocks
2. Counties
3. States
4. National level



PDF $f(x \mid \mu = 0, b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$

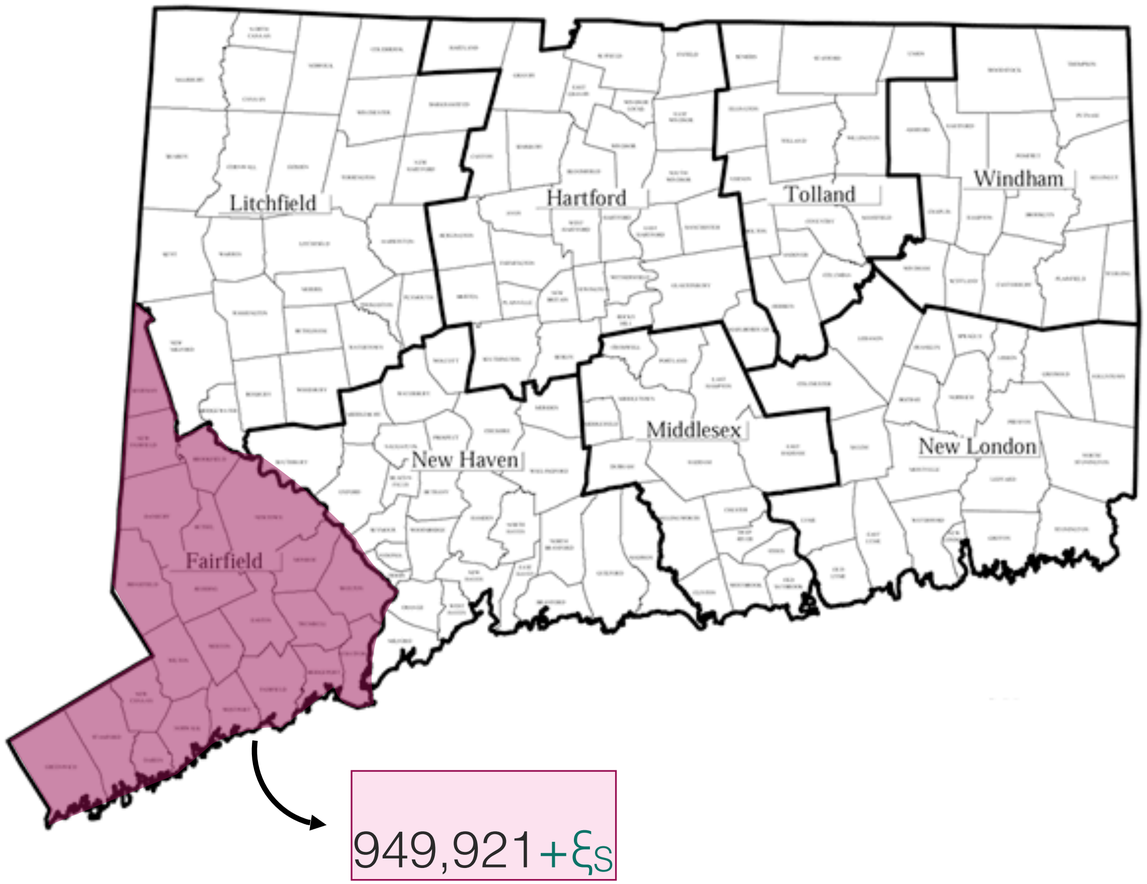


$$Z \sim \text{Lap}(\frac{\Delta_Q}{\epsilon})$$

The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

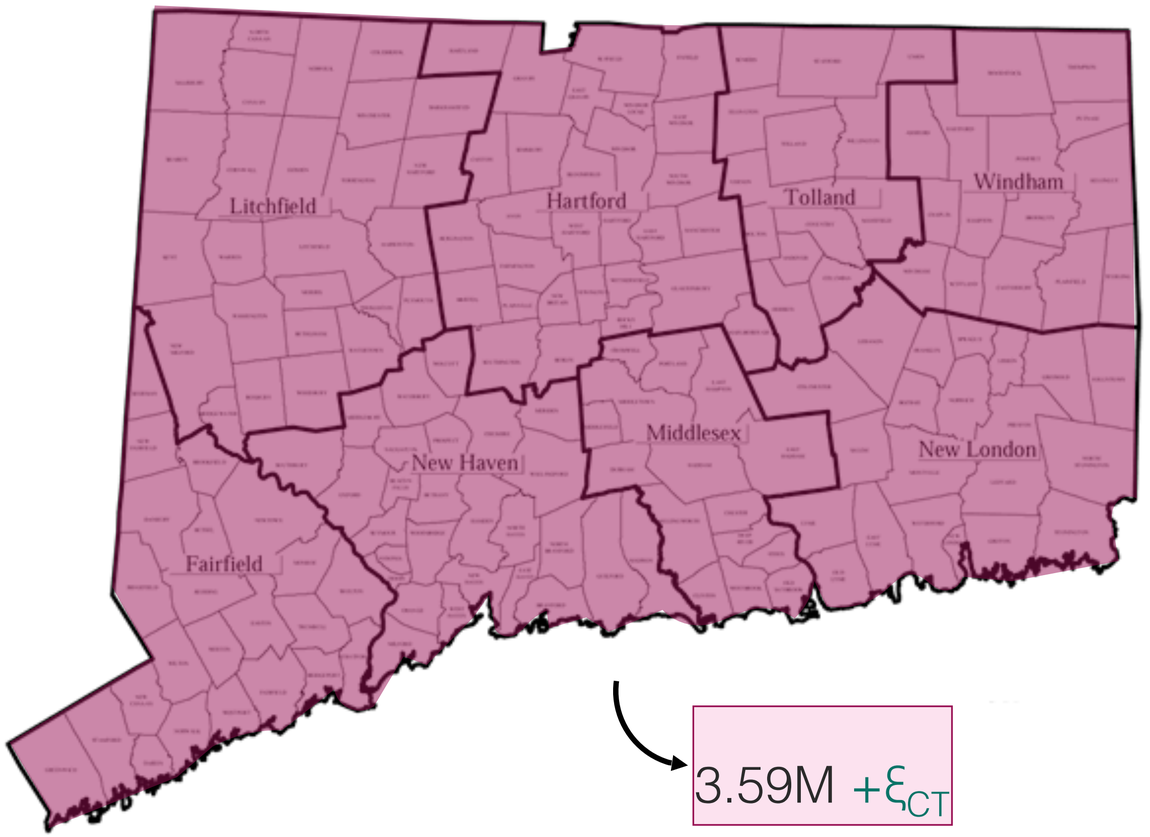
1. Census blocks
2. Counties
3. States
4. National level



The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

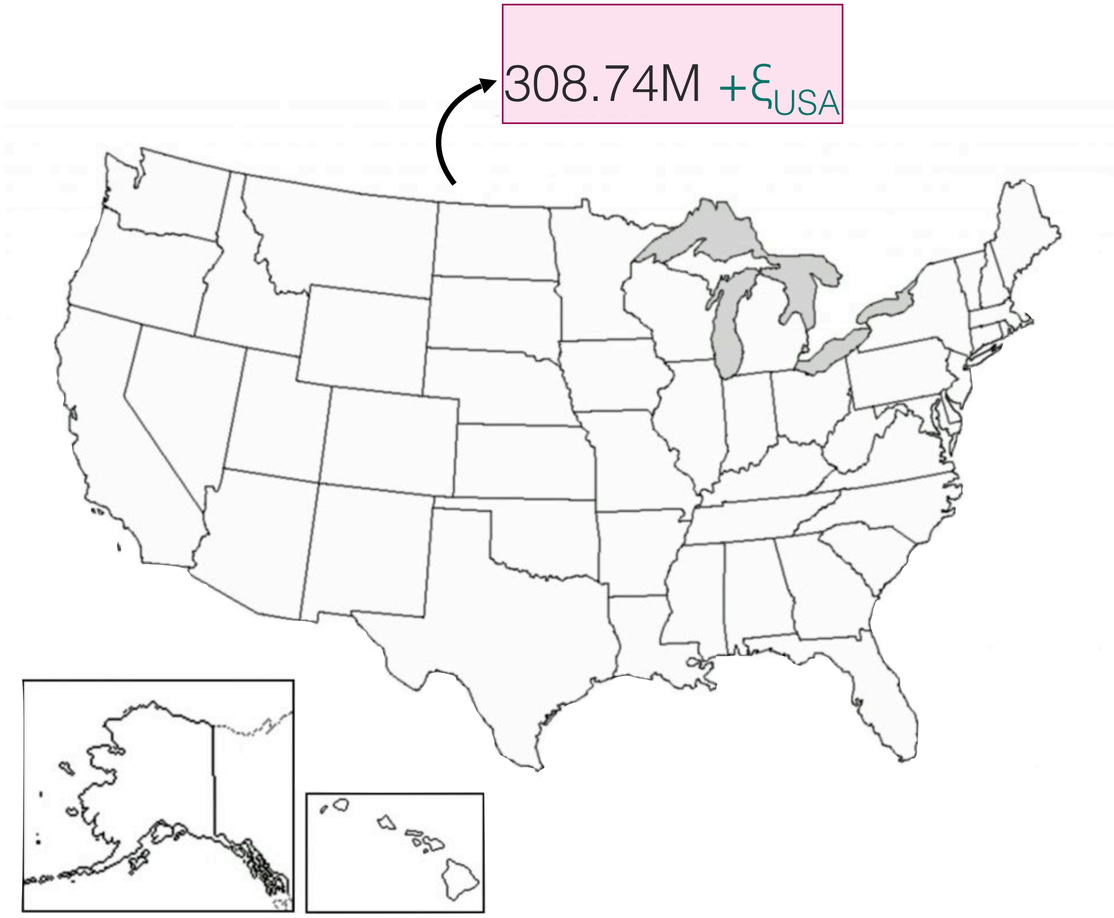
1. Census blocks
2. Counties
3. States
4. National level



The Census Data Release Problem

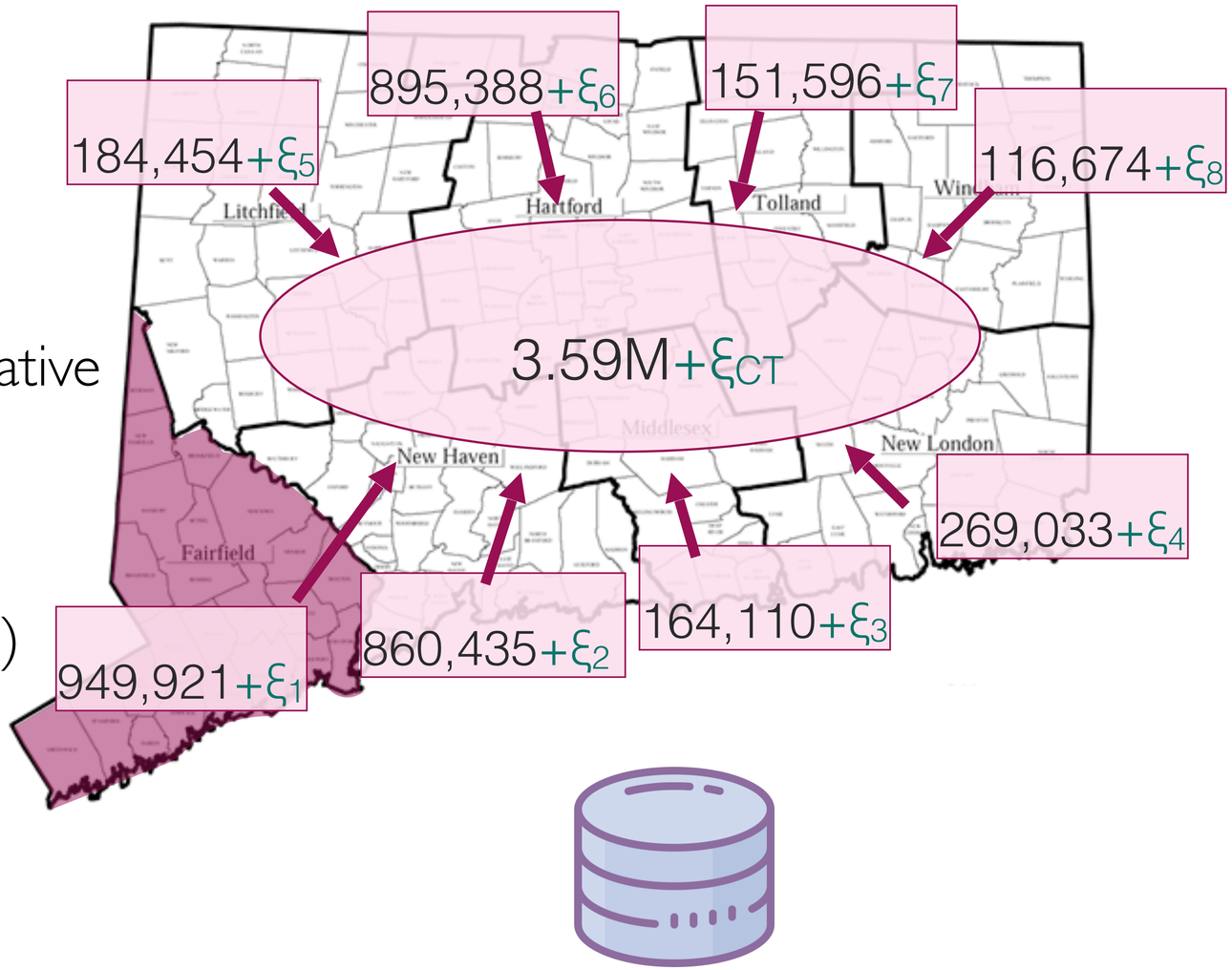
- GOAL: Release socio-demographic feature of the population grouped by:

1. Census blocks
2. Counties
3. States
4. National level

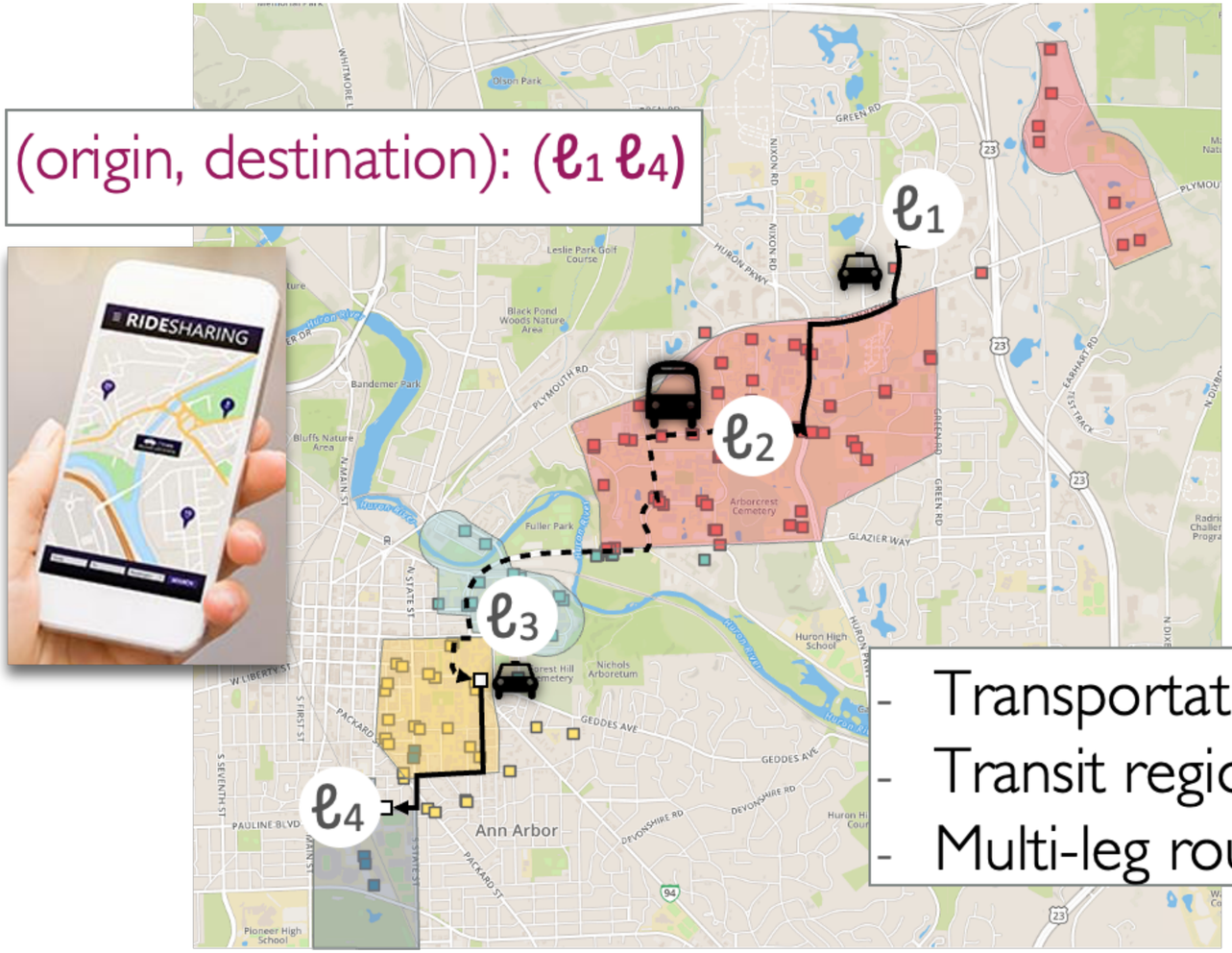


The consistency issue

- **Requirements:**
  1. **Privacy**
  2. Hierarchical **Consistency**
  3. **Validity**: The private values are non-negative
- Noise is applied independently to each estimate
- The noisy quantities at a "level" (e.g., state) are inconsistent with the sum of the noisy quantities at the "children levels" (e.g., counties of that state)



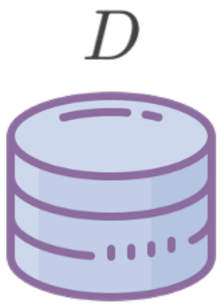
Transportation Application

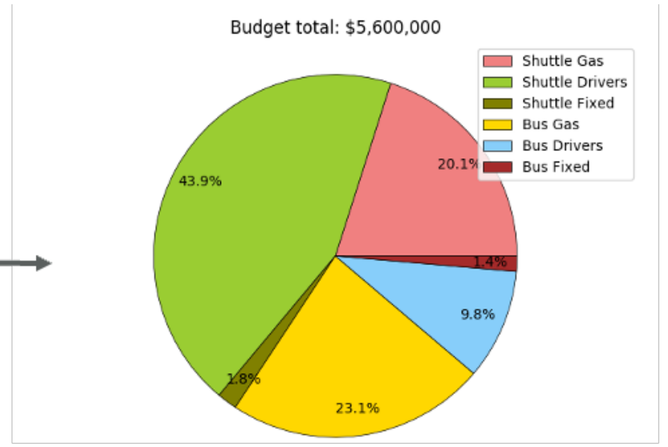


Data universe: ~30k OD-pairs

time	origin	destination	count
8:00-8:30	bbr	eecs	34
8:00-8:30	food_crt1	nw2	0
8:00-8:30	lsa	ort3	1
8:00-8:30	sprt	nw4	0
8:00-8:30	bbb	chi	12
...	...	...	...

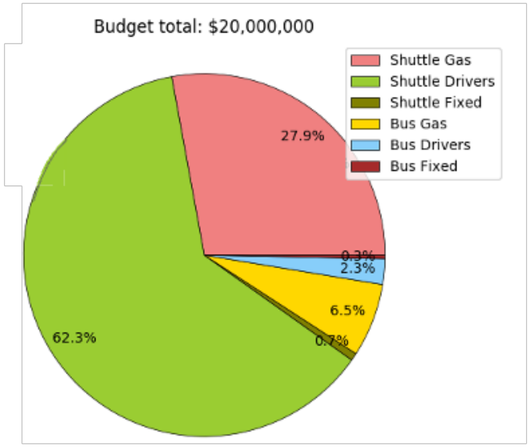
Differential Privacy Challenge for Mobility



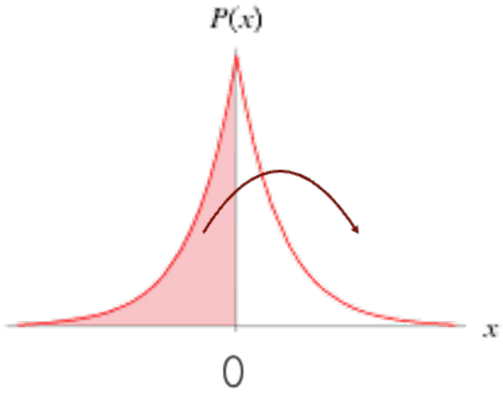


37,714 trips

Laplace Mechanism



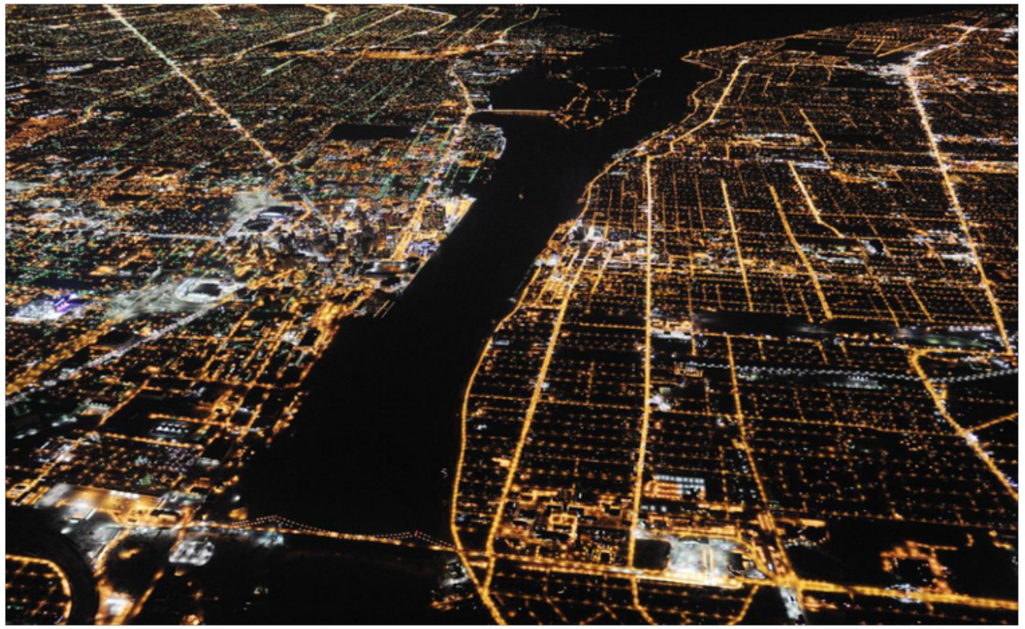
261,032 trips



$\varepsilon = 1.0$

Fioretto:AAMAS-18

Energy Optimization



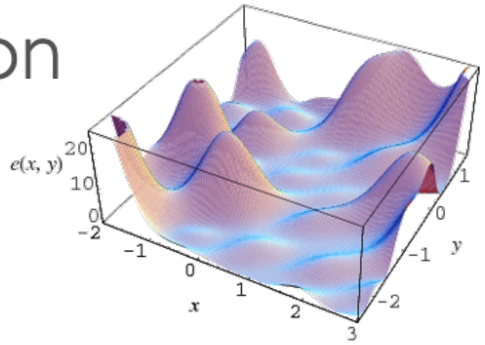
Le réseau de transport d'électricité

Customers

Loads

optimization

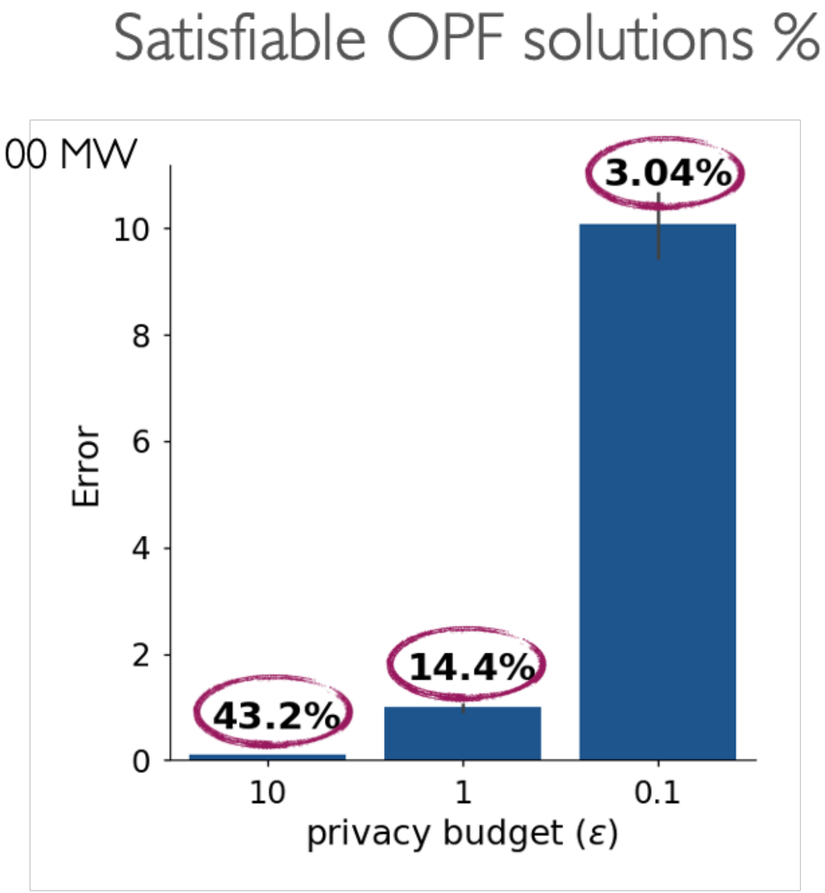
OPF



Fioretto:CPAIOR-18; IEEE-TPS:20, IEEE-TSG:20

Differential Privacy Challenge for Energy

Issues with the problems' constraints



Fioretto:CPAIOR-18; IEEE-TPS:20, IEEE-TSG:20

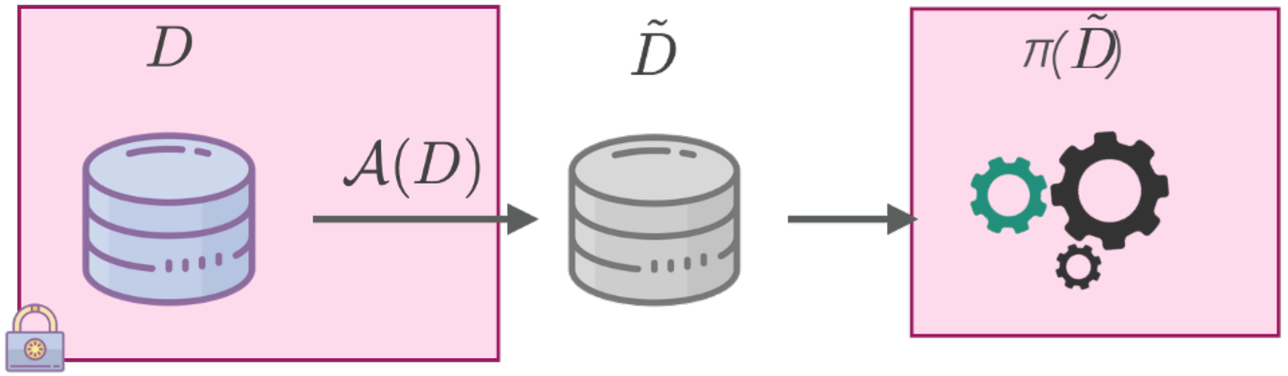
Differential Privacy is oblivious to the structure of the dataset and the constraints of the underlying problem

Constrained based DP

We are interested in solving:

$$\begin{aligned} \text{minimize}_{\mathbf{x}\in\mathbb{R}^n} \quad & f(D, \mathbf{x}) \\ \text{subject to} \quad & g_i(D, \mathbf{x}) \leqslant 0, \quad i = 1, \ldots, p \end{aligned}$$

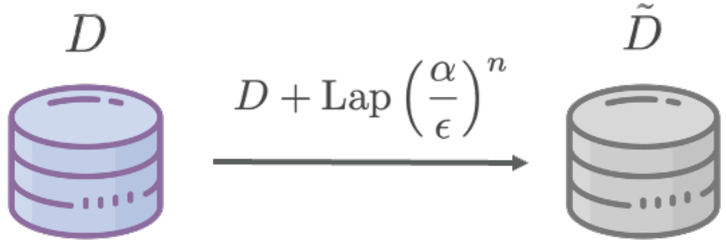
Sensitive Data



Constrained based DP

Consider a dataset $D \in \mathbb{R}^n$ of values

#1



(a)

#2

$$\begin{aligned} &\text{minimize}_{\hat{D},\mathbf{x}\in\mathbb{R}^n} \|\hat{D} - \tilde{D}\|_2^2 \\ &\text{subject to} \quad |f(\hat{D},\mathbf{x}) - f^*| \leqslant \beta \\ &\qquad\qquad g_i(\hat{D},\mathbf{x}) \leqslant 0, i = 1,\ldots,p \end{aligned}$$

(b)

(c)

#3

Release $\hat{D}$

Fioretto:CPAIOR-18; TSG-19; JAIR-19; AIJ-21

Constrained based DP

Properties

- : CBDP achieves ε-DP
  - By application of the privacy-preserving mechanism on D to calibrated noise, and composition
- **Efficiency**: When the constraint space is convex, CBDP runs in polynomial time in the size of the universe and number of constraints
- **Accuracy**: For convex problems, the optimal solution to the optimization model of CBDP satisfies:

- $\|x^* - c\|_{2,w} \leq \|\tilde{c} - c\|_{2,w}$
-

Constrained-based DP

Properties

- Privacy: CBDP achieves generalized ε-Differential Privacy
  - By the properties of the (Polar) Laplace mechanism.

- Accuracy: The optimal solution to the optimization model of CBDP satisfies:

$$\|\hat{S}^* - S\|_2 \leq 2\|\tilde{S} - S\|_2$$

that is, it is at away from optimality by at most a factor of 2.

Fioretto et al. IEEE TPWRS 2020

Constrained-based DP

Geometrical intuition

- Let $Q_1, Q_2, \ldots, Q_k$ be a collection of queries.
- Let $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_k$ be their private, noisy, answers.
- Constraint $\mathbb{C}(Q_1, Q_2, \ldots, Q_k)$ is satisfied (in all data sets) on the true query answers, but does not on noisy answers

Objective: Find $x_1, x_2, \ldots, x_k$ such that:

1. Close to $\tilde{Q}_1, \tilde{Q}_2, \ldots, \tilde{Q}_k$
2. Satisfy $\mathbb{C}(x_1, x_2, \ldots, x_k)$

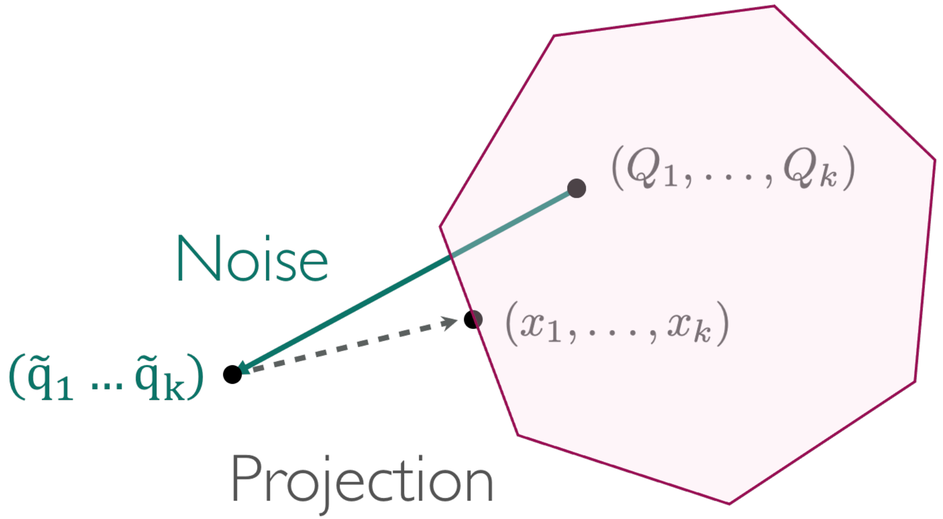
Constrained-based DP

Geometrical intuition

- Consider a set of queries $\{Q_1, \ldots, Q_k\}$ with private answers $\tilde{q}_1, \ldots, \tilde{q}_k$

$$\min \sum_{i=1}^{k} (x_i - \tilde{q}_i)^2$$

$$\text{s.t.: } \mathbb{C}(x_1, x_2, \ldots, x_k)$$



Space of outputs satisfying $\mathbb{C}$

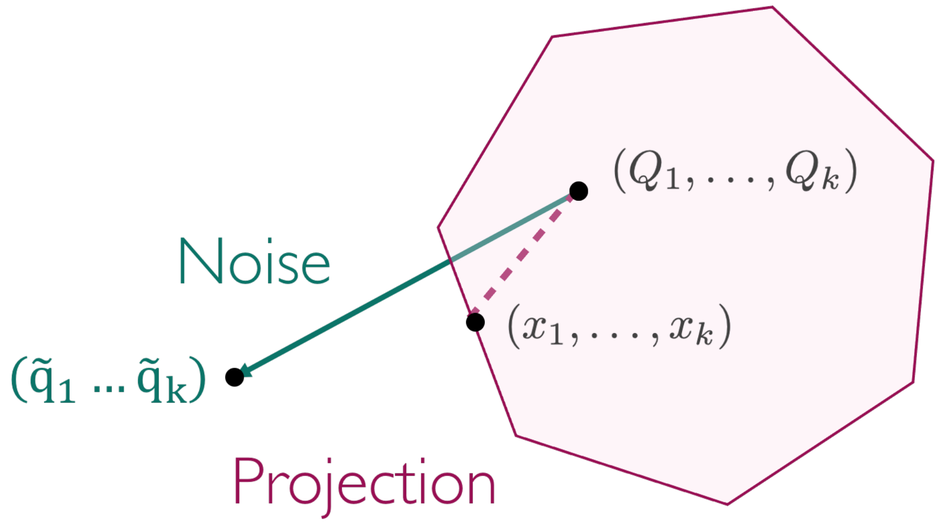
Constrained-based DP

Geometrical intuition

- Consider a set of queries $\{Q_1, \ldots, Q_k\}$ with private answers $\tilde{q}_1, \ldots, \tilde{q}_k$

$$\min \sum_{i=1}^{k} (x_i - \tilde{q}_i)^2$$

$$\text{s.t.: } \mathbb{C}(x_1, x_2, \ldots, x_k)$$



$$\|x - Q\|_2 \leq \|\tilde{Q} - Q\|_2$$

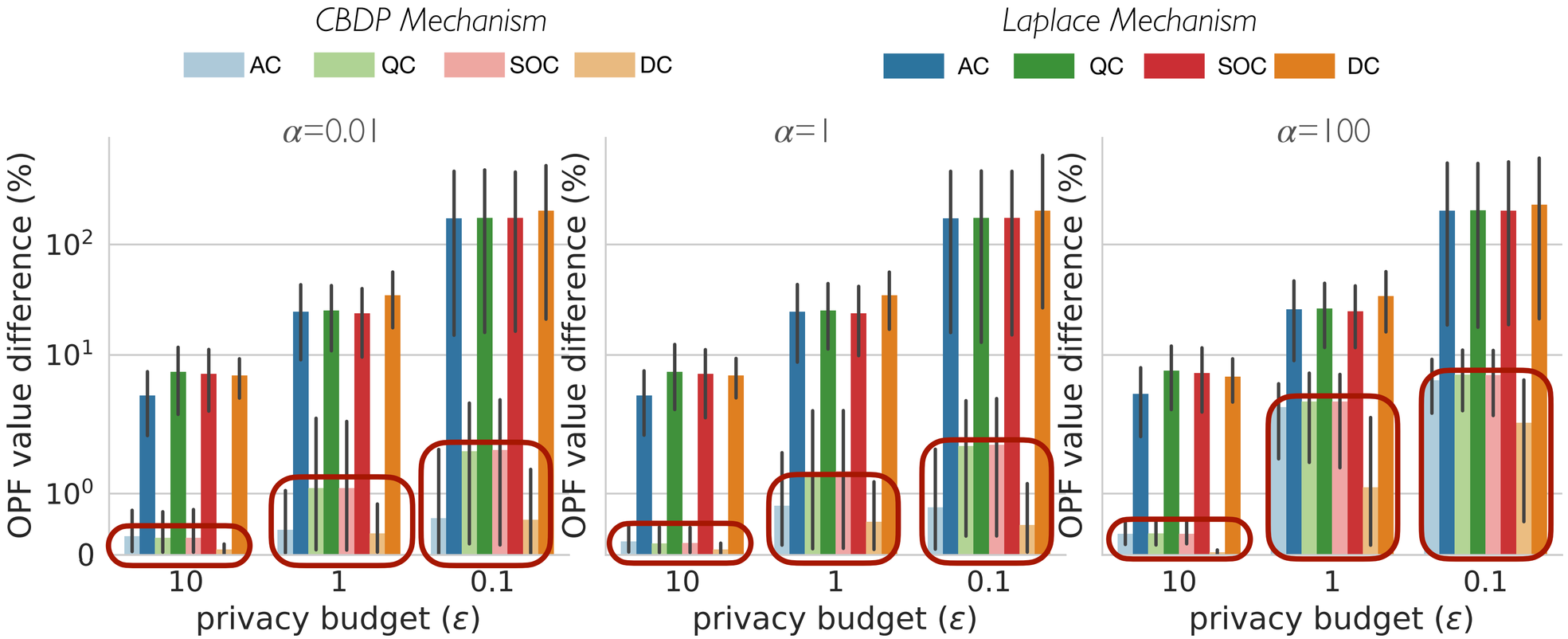
When $\mathbb{C}$ forms a convex space

Protecting Loads

Analysis of the OPF

Summary:

- 1-2 order of magnitude improvements, for all ε and α
- Difference of OPF values between BiLevel-DP and original data is <10%



Summary of Results

Fioretto et al. AAMAS 2018

Mobility

- Features: Total trips (2), Operating zones (3), Transit modes (4)
- Baselines: **Laplace** [Dwork:08], **Hierarchical** [Hay:10, Cormode:13]
- Key results: >= 1 order magnitude improvement

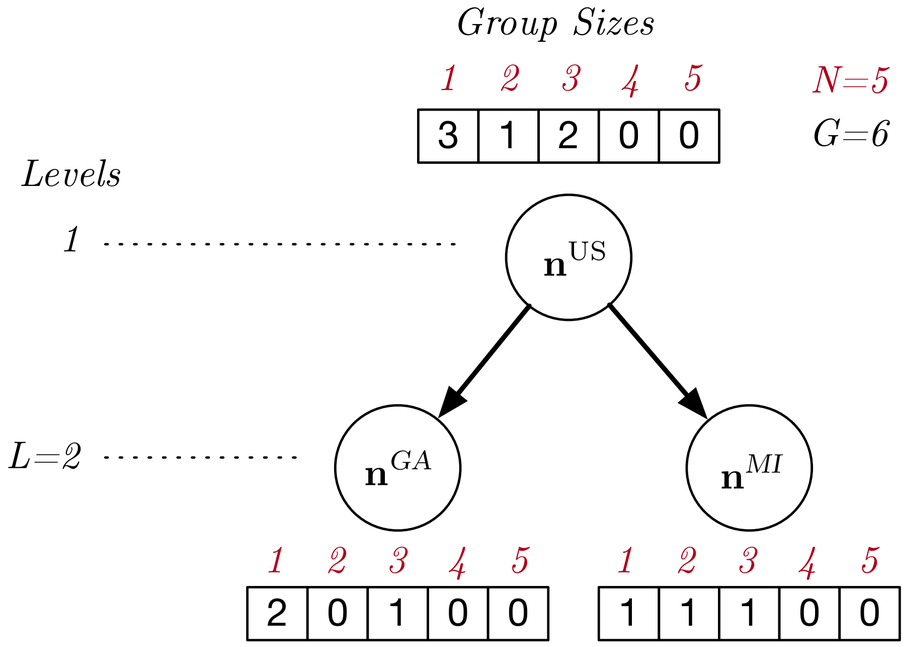
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\qquad \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\qquad\qquad \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

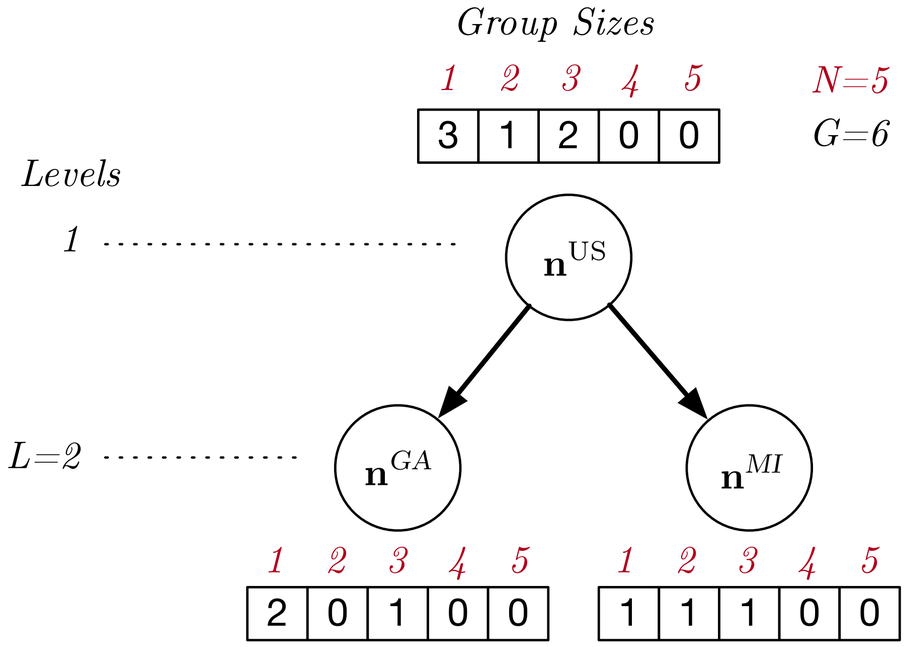
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
\text{s.t:} \quad & \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
& \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
& \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

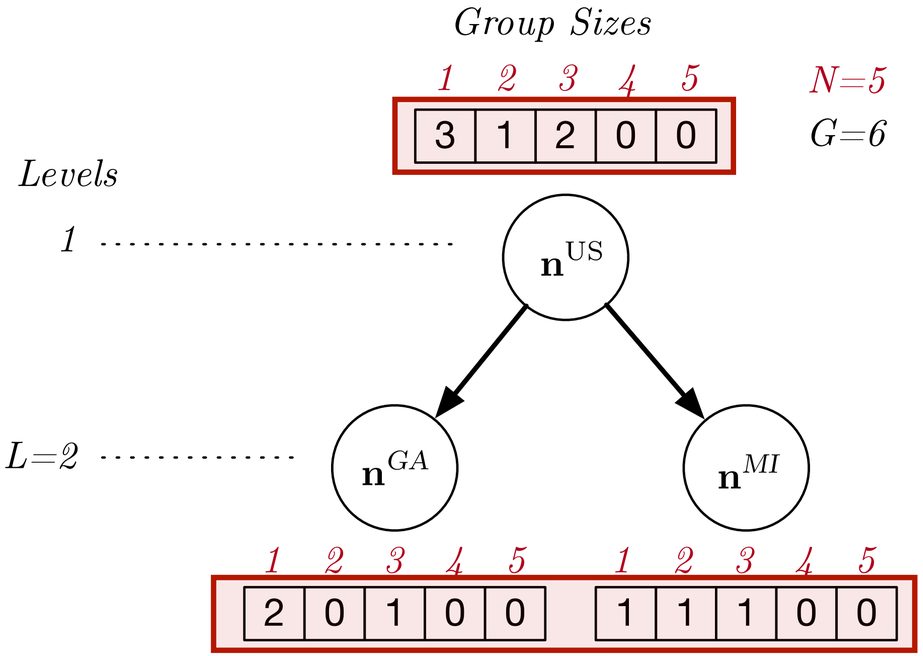
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\qquad \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\qquad\qquad \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

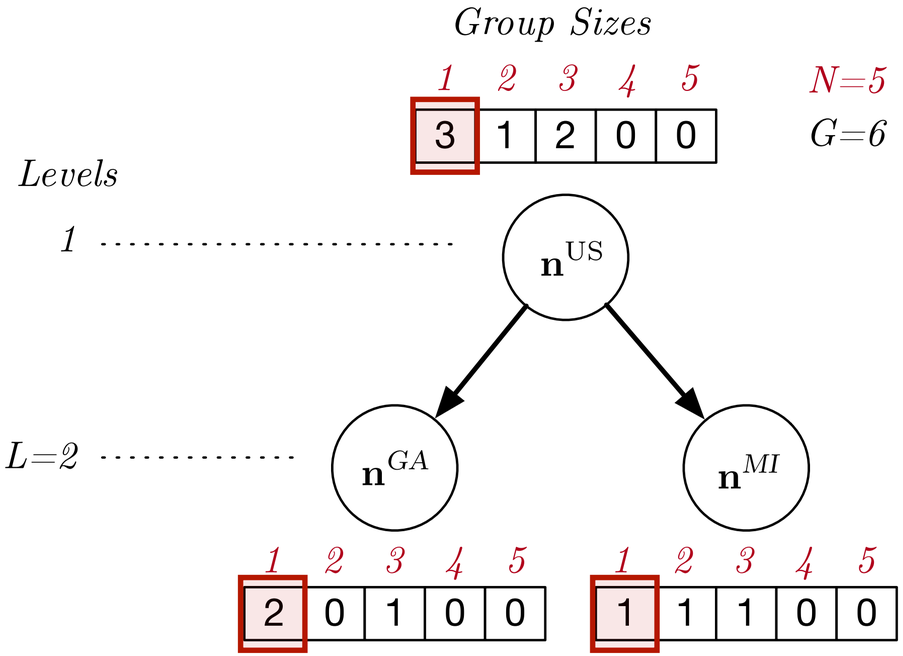
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

A Differentially Private Optimization Approach

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 \qquad \text{(H1)}$$

$$\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} \qquad \text{(H2)}$$

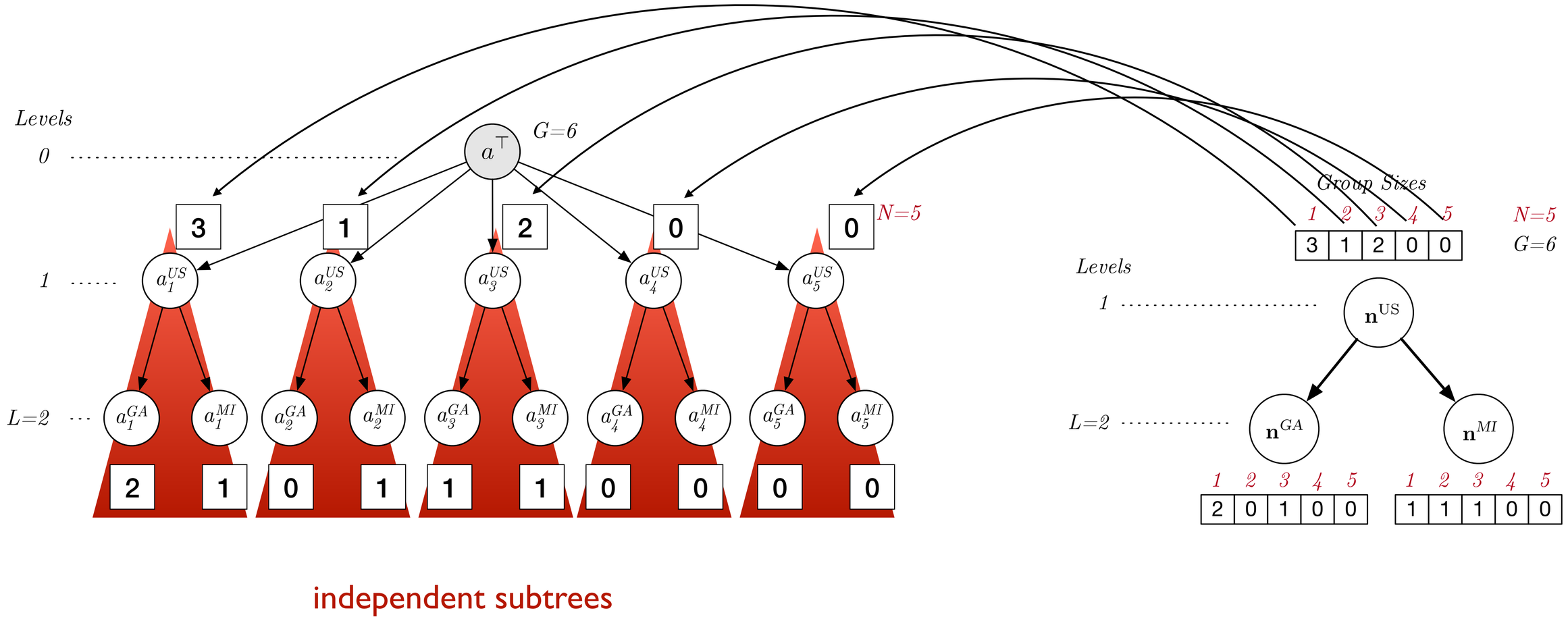
$$\sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] \qquad \text{(H3)}$$

$$\hat{n}_s^r \in D_s^r \quad \forall r \in \mathcal{R}, s \in [N] \qquad \text{(H4)}$$

- Solving this QIP is intractable for the datasets of interest to the census bureau.
- Relax the integrality constraint.
- The resulting optimization problem becomes convex but presents two limitations:
  1. Its final solution may violate the **consistency** and **faithfulness** conditions
  2. The mechanism is still too slow for very large problems!

Exploiting the Problem Structure

A Dynamic Programming Solution

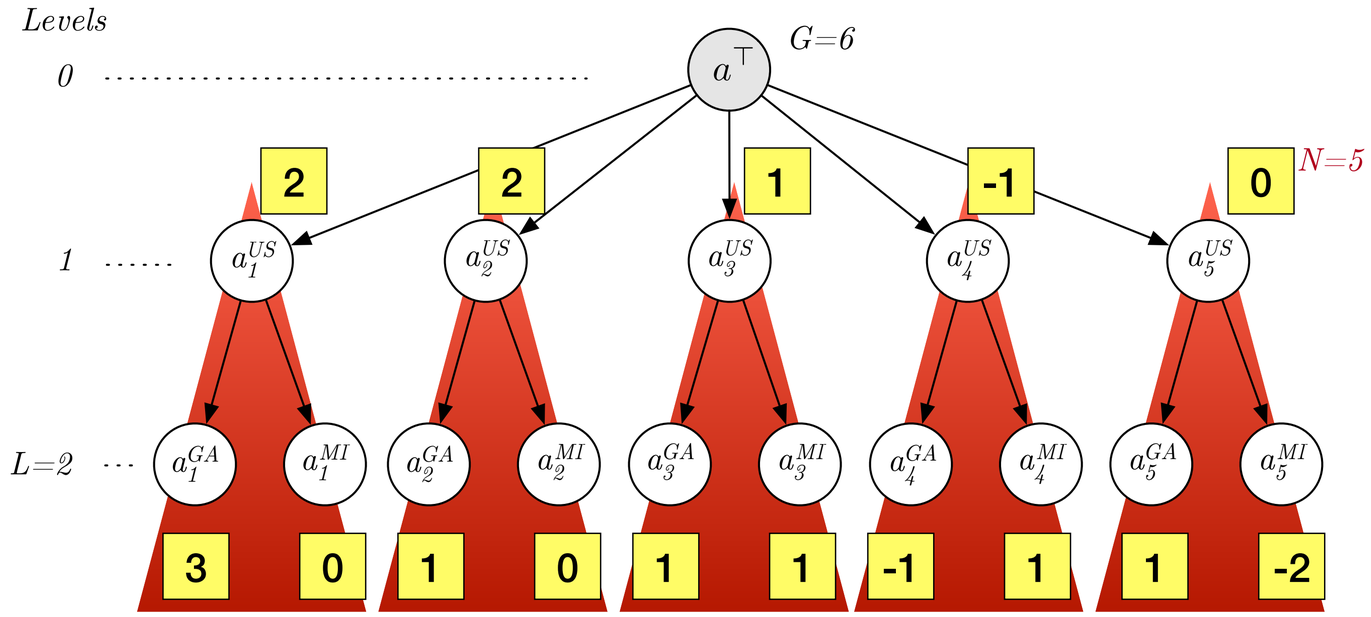


independent subtrees

Exploiting the Problem Structure

A Dynamic Programming Solution

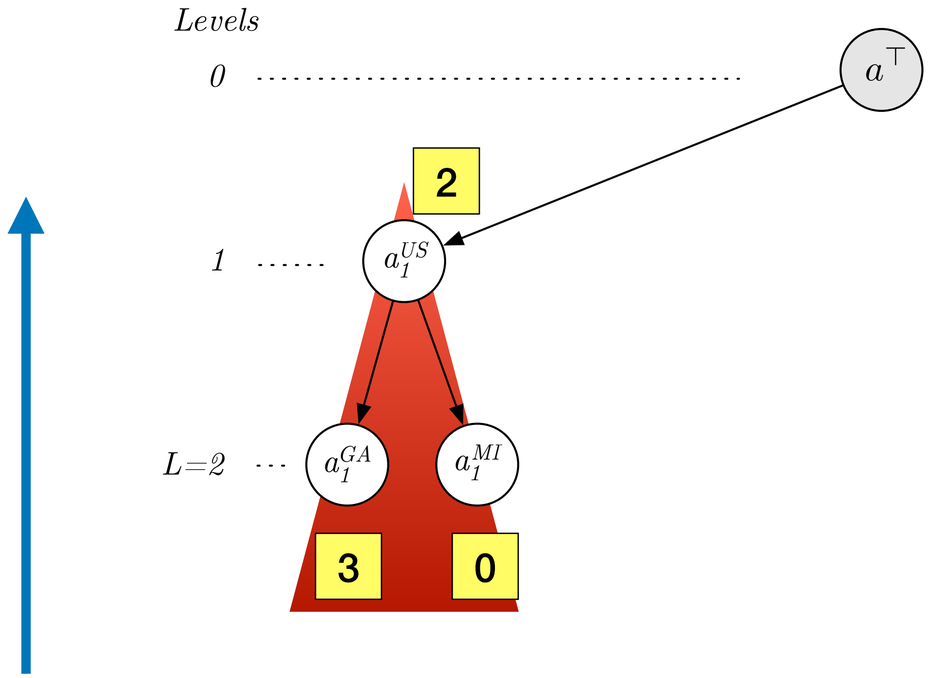
1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$



$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

independent subtrees

Exploiting the Problem Structure
A Dynamic Programming Solution



2. Bottom-up phase

- Find new, group sizes $\hat{n}^r$ that satisfy the consistency properties.
- Each node of the tree, computes a table $\tau^r : D^r \rightarrow \mathbb{R}_+$ mapping values (group sizes) to costs.
- $\tau^r(v)$ is the optimal cost for $\hat{n}^r$ in the subtree rooted at region **r** when $\hat{n}^r = v$
- The optimal cost for $\tau^r(v)$ can be computed from the cost table τ^c of region **r** children $c \in ch(r)$

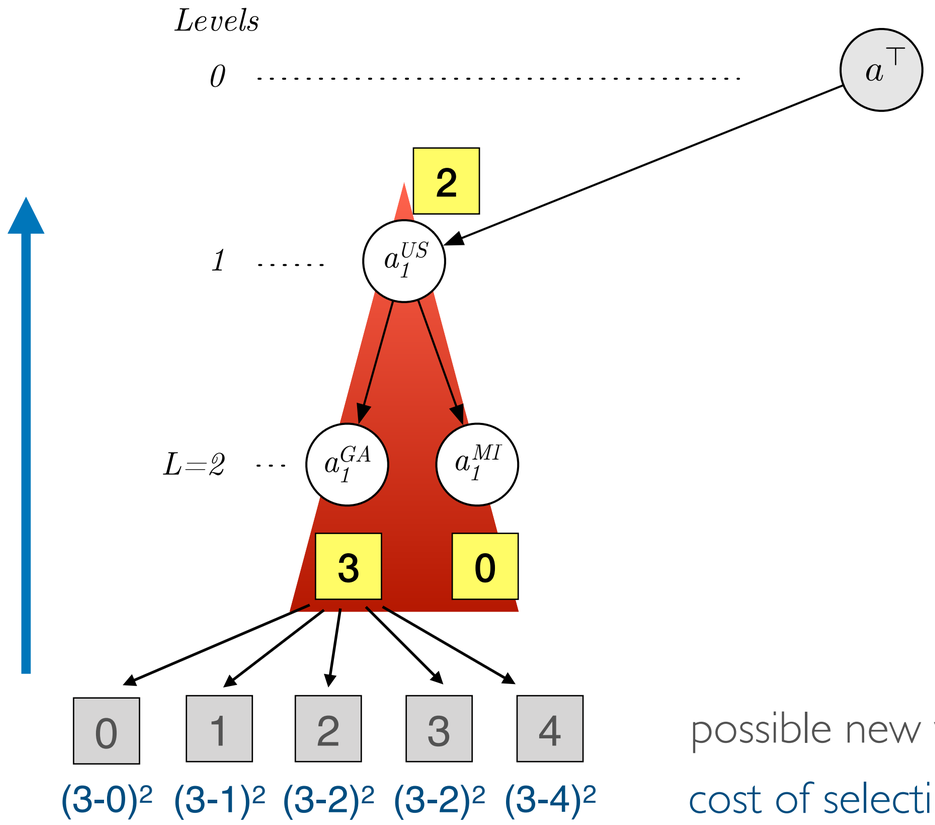
$$\boldsymbol{\tau}^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

$$\boldsymbol{\phi}^r(v) = \min_{\{x_c\}_{c \in ch(r)}} \sum_{c \in ch(r)} \boldsymbol{\tau}^c(x_c) \quad \text{(d2)}$$

$$\text{s.t.} \sum_{c \in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \quad \text{(d4)}$$

Exploiting the Problem Structure
A Dynamic Programming Solution



possible new values v for $\hat{n}_1^{GA}$

cost of selecting value v

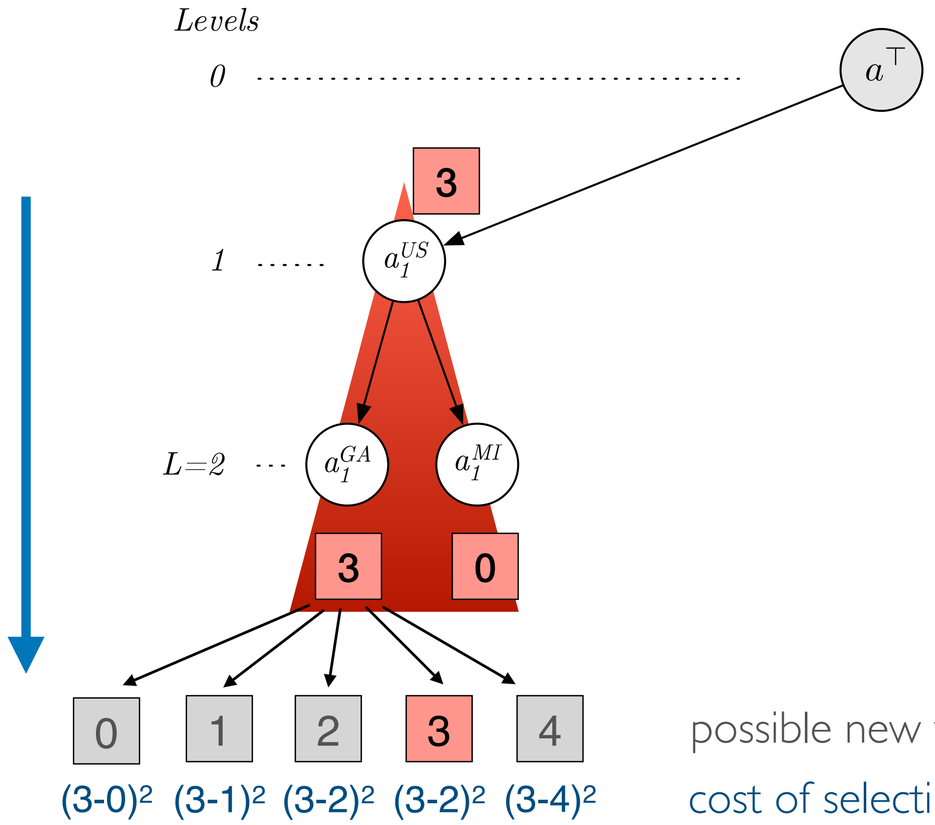
2. Bottom-up phase

- Find new, group sizes $\hat{n}^r$ that satisfy the consistency properties.
- Each node of the tree, computes a table $\tau^r : D^r \rightarrow \mathbb{R}_+$ mapping values (group sizes) to costs.
- $\tau^r(v)$ is the optimal cost for $\hat{n}^r$ in the subtree rooted at region **r** when $\hat{n}^r = v$
- The optimal cost for $\tau^r(v)$ can be computed from the cost table τ^c of region **r** children $c \in ch(r)$

cost tables

v	τ_1^{GA}	τ_1^{MI}	τ_1^{US}
0	9	**0**	$4+\min(9+0)$
1	4	1	$1+\min(9+1;\ 4+0)$
2	1	4	$0+\min(0+4;\ 4+1;\ 1+0)$
3	**0**	9	$\mathbf{1}+\min(0+0;\ 1+1;\ 4+4;\ 9+9)$
4	1	16	$4+\min(1+0;\ 0+1;\ 1+4;\ \ldots)$

Exploiting the Problem Structure
A Dynamic Programming Solution



possible new values v for $\hat{n}_1^{GA}$

cost of selecting value v

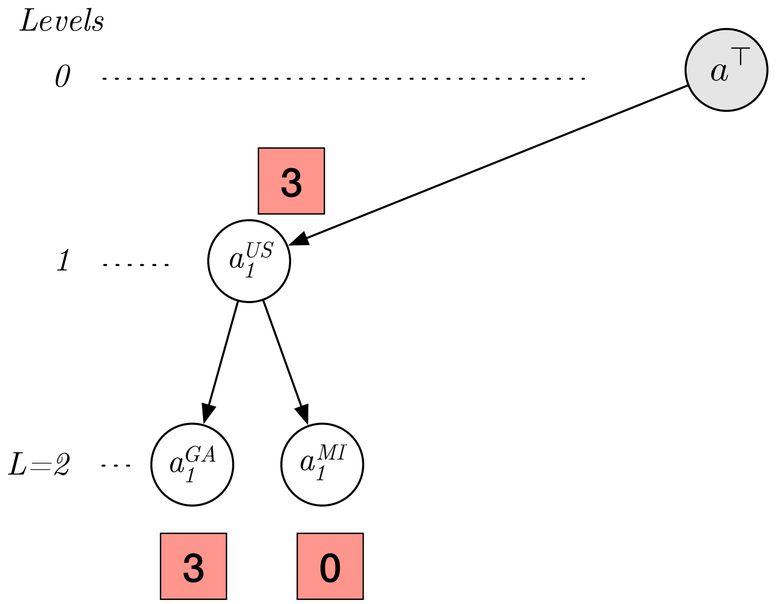
3. Top-down phase

- Each node select the value v that minimizes its cost table
- Using the parent value, choice each children repeats the process

cost tables

v	τ_1^{GA}	τ_1^{MI}	τ_1^{US}
0	9	**0**	$4+\min(9+0)$
1	4	1	$1+\min(9+1;\ 4+0)$
2	1	4	$0+\min(0+4;\ 4+1;\ 1+0)$
3	**0**	9	$\mathbf{1}+\min(0+0;\ 1+1;\ 4+4;\ 9+9)$
4	1	16	$4+\min(1+0;\ 0+1;\ 1+4;\ \ldots)$

Exploiting the Problem Structure
A Dynamic Programming Solution



The issue

- The construction of the data hierarchy requires solving $O(|R|N\bar{D})$ optimization problems, with $\bar{D} = \max_{s,r} |D_s^r|$

R = regions, s = groups sizes (1, 2, 3 , …)

$$\boldsymbol{\tau}^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

$$\boldsymbol{\phi}^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \boldsymbol{\tau}^c(x_c) \quad \text{(d2)}$$

$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \quad \text{(d4)}$$

Exploiting the Cost Functions Structure

A Polynomial Time Mechanism

Main Result: The function ϕ_s^r used to compute the values $\tau_s^c(v)$ convex piecewise linear (CPWL)

$$v_c^k = \begin{cases} v_c^{k-1} + 1 \text{ if } \; c = \text{argmin}_c \, \tau^c(v_c^{k-1}+1) - \tau^c(v_c^{k-1}) \\ v_c^{k-1} \qquad\qquad \text{otherwise.} \end{cases}$$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

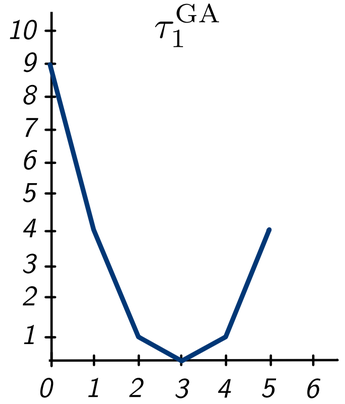
$$\phi^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \tau^c(x_c) \quad \text{(d2)}$$

$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \;\; \forall c \in ch(r) \quad \text{(d4)}$$

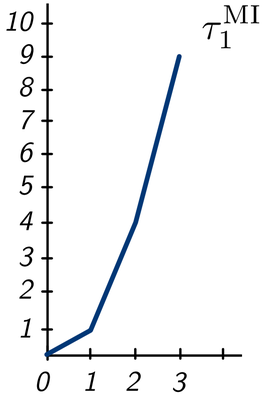
τ_1^{GA}

j	$cost$
0	9
1	4
2	1
3	0
4	1
5	4



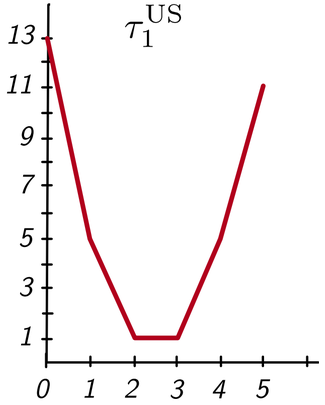
τ_1^{MI}

j	$cost$
0	0
1	1
2	4
3	9



τ_1^{US}

j	$cost$
0	4+(0+0)
1	1+(0+4)
2	0+(0+1)
3	1+(0+0)
4	4+(1+0)
5	9+(1+1)



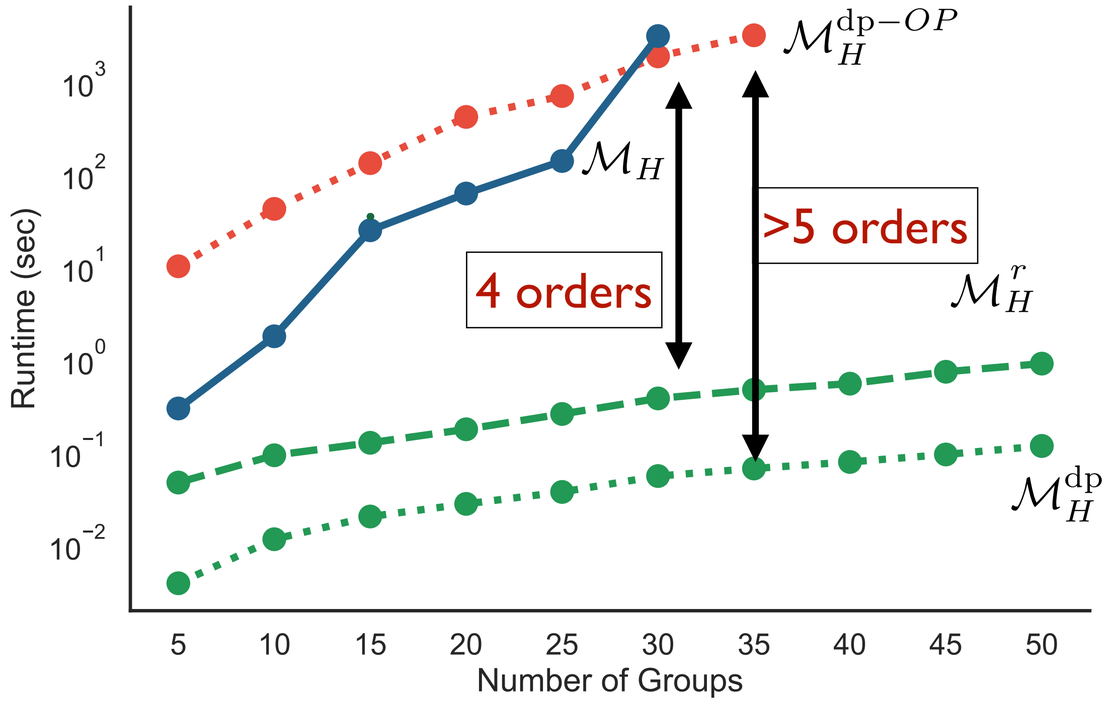
Corollary: The cost table function of each node of the tree is CPWL

Experimental Evaluation

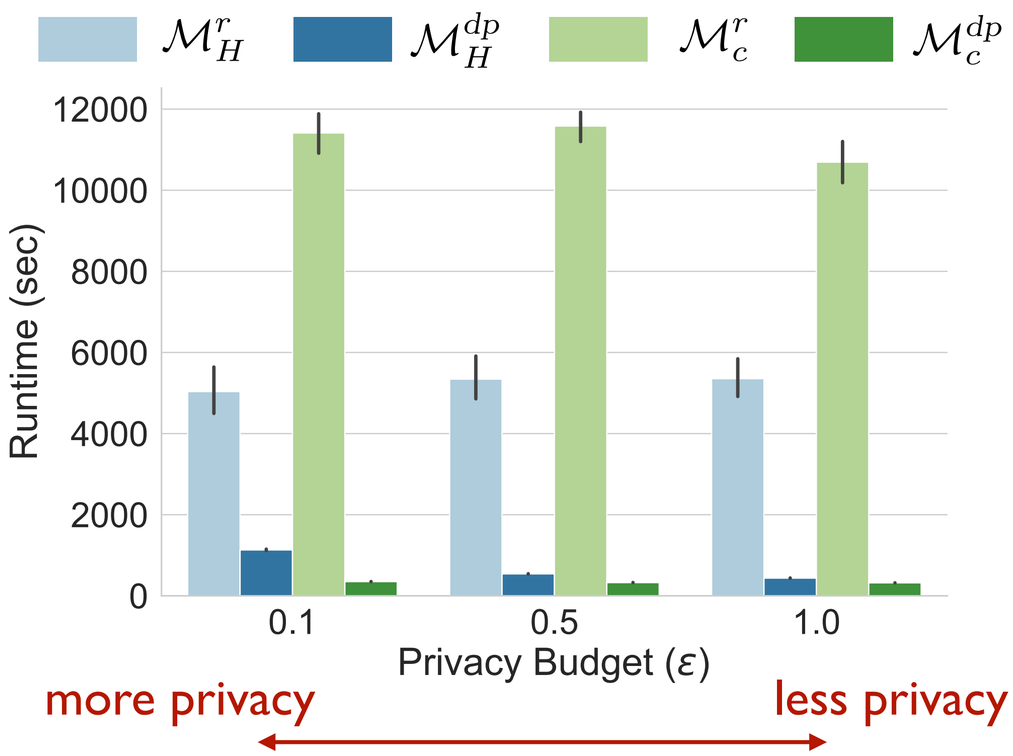
Runtime

Main Result: Exploiting convexity and structure of the problem provides up to **2 order magnitude improvements** w.r.t. the relaxed QIP method

Rule out the "naive" and QIP options



US Census Dataset



Discussion

- **A bit of quizzing**:
  - Can you explain the concept of 'ε' (epsilon) in Differential Privacy? How does changing its value affect privacy and data utility?
  - What are some common mechanisms (like Laplace or Gaussian mechanisms) used to achieve Differential Privacy? How do they work?
- **Practical thinking**:
  - Discuss a real-world scenario where DP could be effectively applied. What challenges might arise in this implementation?
  - How would you approach the trade-off between data utility and privacy when implementing DP in a large-scale public dataset?
- **Responsible use**:
  - What are the ethical implications of not using Differential Privacy in data-driven projects?
  - Can DP always guarantee the protection of individual's data? Are there any scenarios where it might fail?
- If you had to argue against the use of DP, what points would you raise?
- How might advances in technology (like quantum computing) impact the effectiveness of Differential Privacy?
-

Discussion

- **Balancing Privacy and Public Good:**
  - In situations where AI can benefit public health or safety, how should the balance between individual privacy and the collective good be managed?
  - Should there be exceptions to privacy rules for AI applications in critical areas like healthcare, criminal justice, or national security?
- **Consent and Awareness:**
  - How should informed consent be obtained or considered especially when data is used for AI training?
  - What level of understanding should individuals have about Differential Privacy before their data is used? How can this be realistically achieved?
- **Transparency:**
  - Should organizations be required to disclose the use of DP in their AI systems?
  - Who should be held accountable when an AI system, using differentially private data, makes an erroneous or harmful decision?
- **Long-term Effects and Society:**
  - What could be the long-term societal impacts of widespread adoption of Privacy in AI. Could it lead to a more privacy-conscious culture or create new challenges?
  - How might the evolution of AI technologies influence the future development and implementation of DP principles?

Responsible AI:
Seminar on Fairness, Safety, Privacy and more

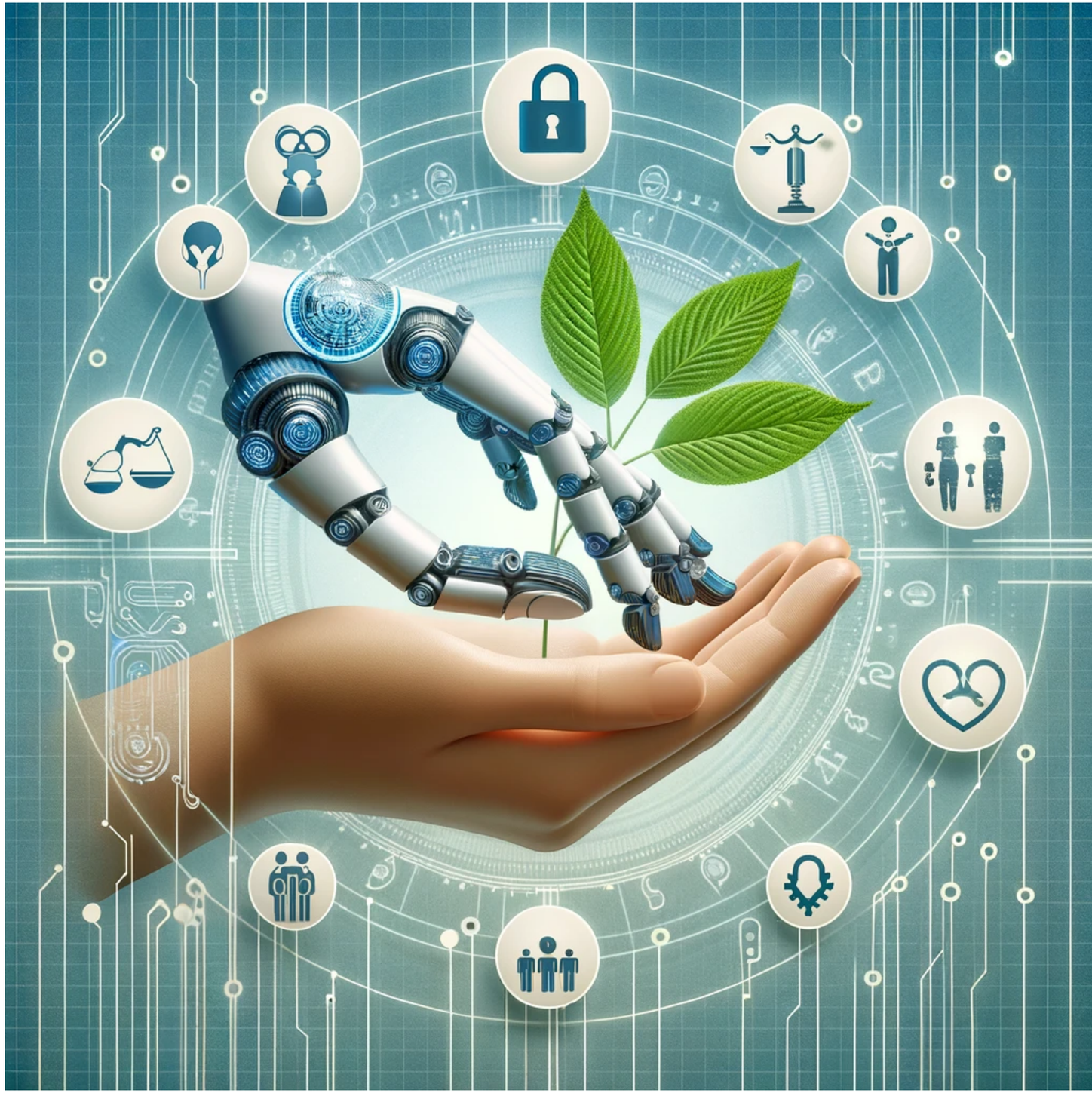
Thank you!

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

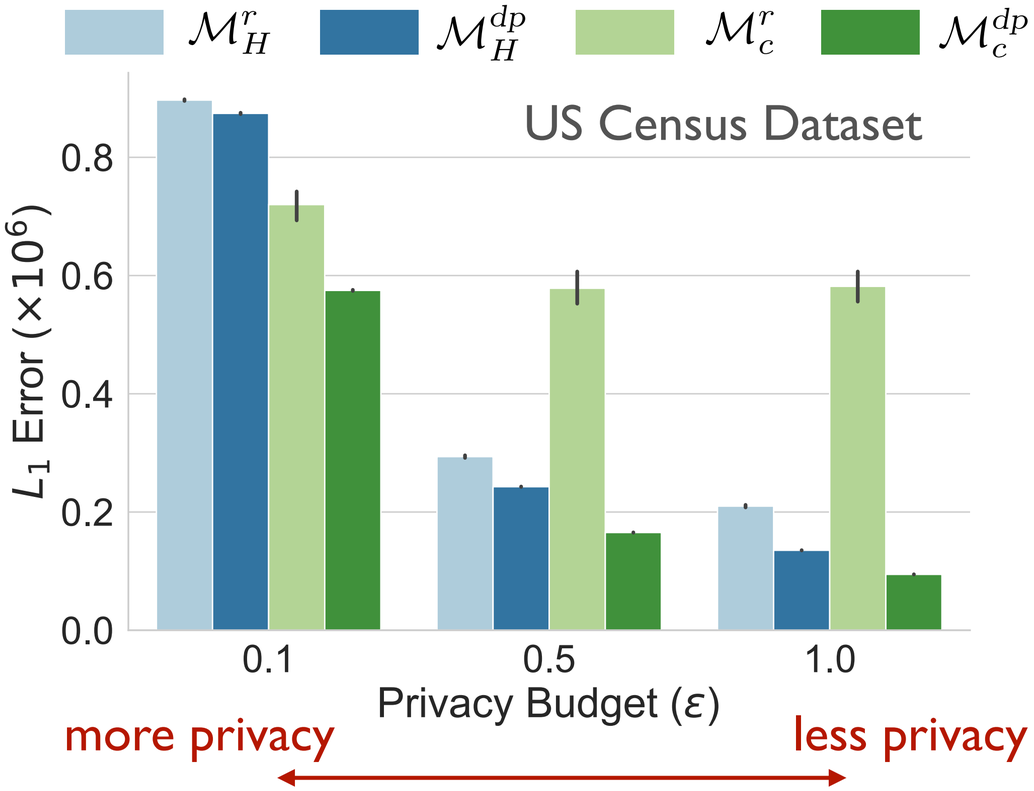
Ferdinando Fioretto @UVA Spring 2024



Experimental Evaluation

Accuracy

Main Result: The CPWL-dynamic programming versions produces **less errors** than the QP counterparts **violating no constraints**



		Taxi Data				Census Data			
		L_1 Errors ($\times 10^4$)			#CV	L_1 Errors ($\times 10^3$)			#CV
ϵ	**Alg**	Lev 1	Lev 2	Lev 3		Lev 1	Lev 2	Lev 3	
0.1	$\mathcal{M}_H^r$	25.4	158.7	904.4	18206	40.3	54.3	802.1	1966
	$\mathcal{M}_H^{dp}$	26.6	121.9	915.7	**0**	10.3	38.4	825.4	**0**
	$\mathcal{M}_c^r$	47.9	153.2	**551.6**	19460	23.1	64.5	632.2	1715
	$\mathcal{M}_c^{dp}$	**19.9**	**65.6**	644.3	**0**	**0.9**	**23.2**	**550.6**	**0**
0.5	$\mathcal{M}_H^r$	8.6	81.2	364.2	18591	39.4	37.9	216.3	1990
	$\mathcal{M}_H^{dp}$	5.5	31.0	408.9	**0**	2.4	9.4	230.8	**0**
	$\mathcal{M}_c^r$	46.7	153.5	450.7	19531	23.1	61.0	494.2	1718
	$\mathcal{M}_c^{dp}$	**4.0**	**16.4**	**352.9**	**0**	**0.2**	**5.8**	**159.1**	**0**
1.0	$\mathcal{M}_H^r$	7.7	77.2	**279.0**	18085	40.7	39.2	130.0	1989
	$\mathcal{M}_H^{dp}$	3.1	19.8	328.5	**0**	1.2	5.1	128.8	**0**
	$\mathcal{M}_c^r$	47.1	154.2	447.1	19706	24.1	63.0	494.5	1728
	$\mathcal{M}_c^{dp}$	**2.0**	**8.7**	307.8	**0**	**0.1**	**3.2**	**91.0**	**0**