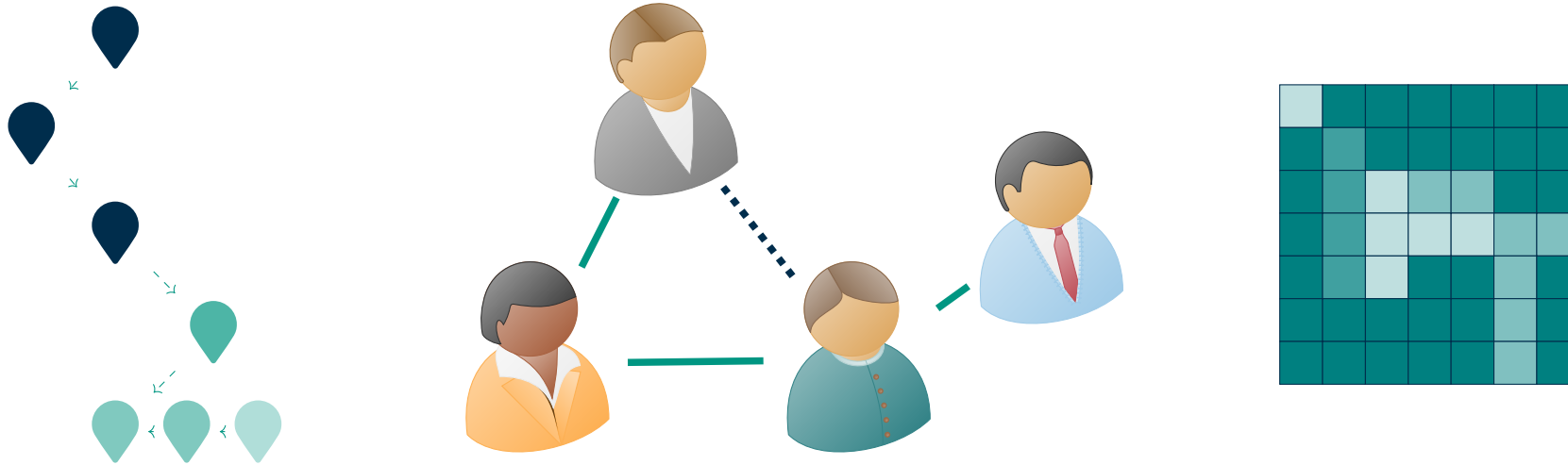




Connecting Differential Privacy to Real-World Attack Mitigation

MPI-SP

Patricia Guerra Balboa | 26th May 2026

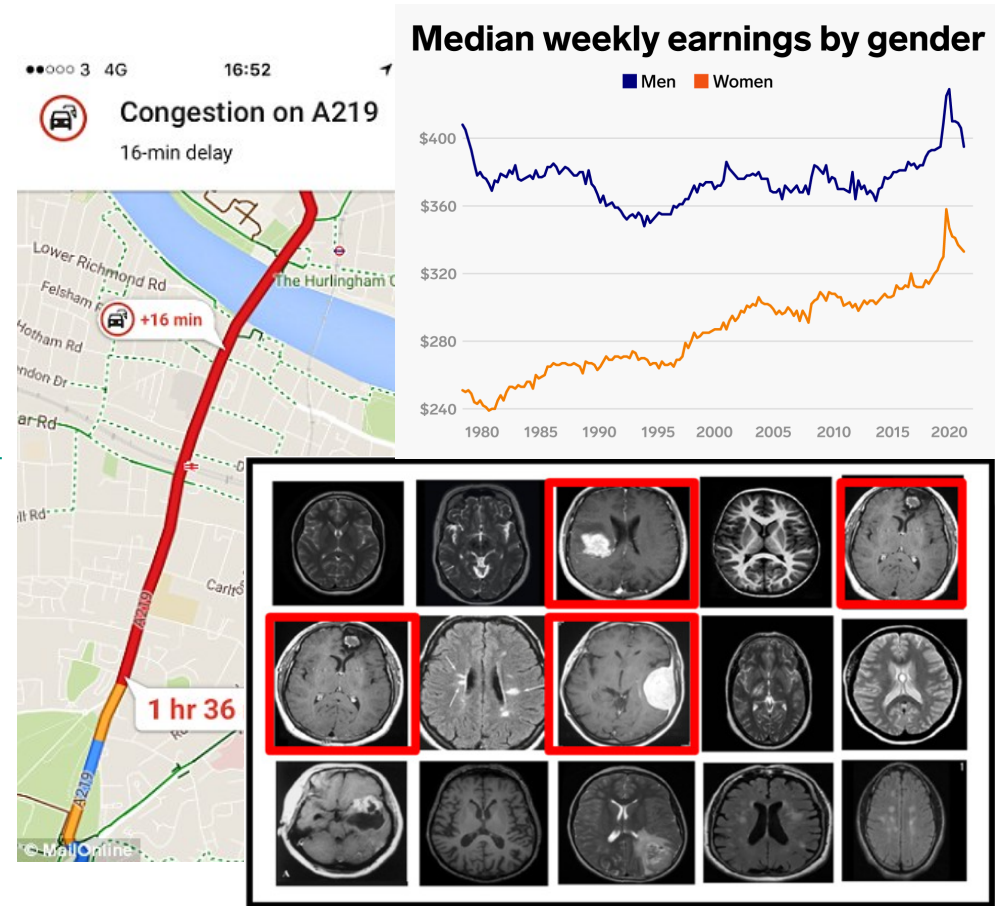
PRIVACY
AND SECURITY

 KASTEL

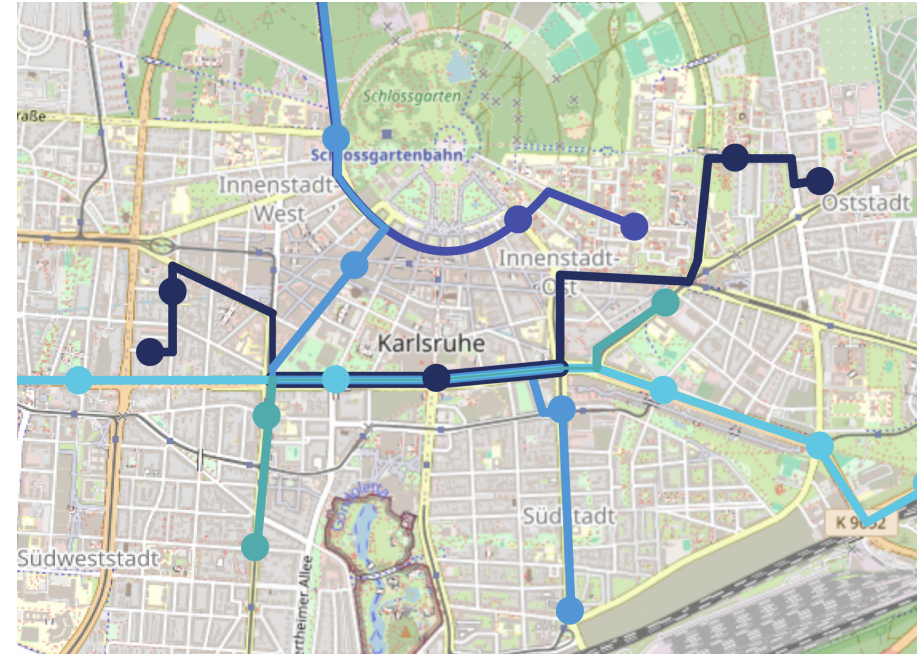
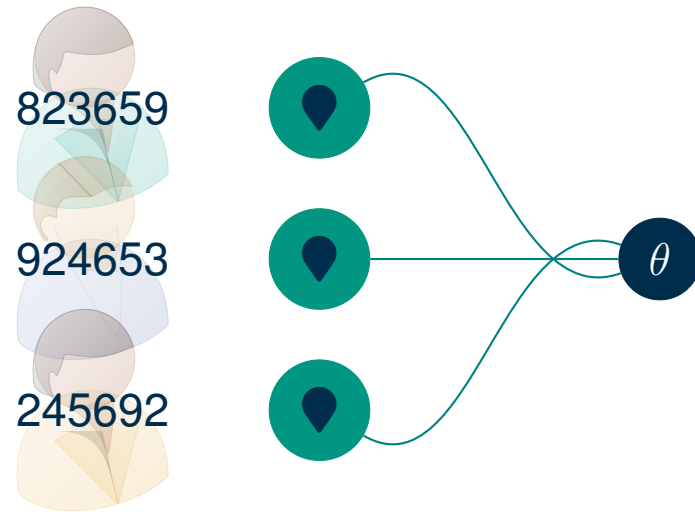
Privacy Risk in Data Analysis



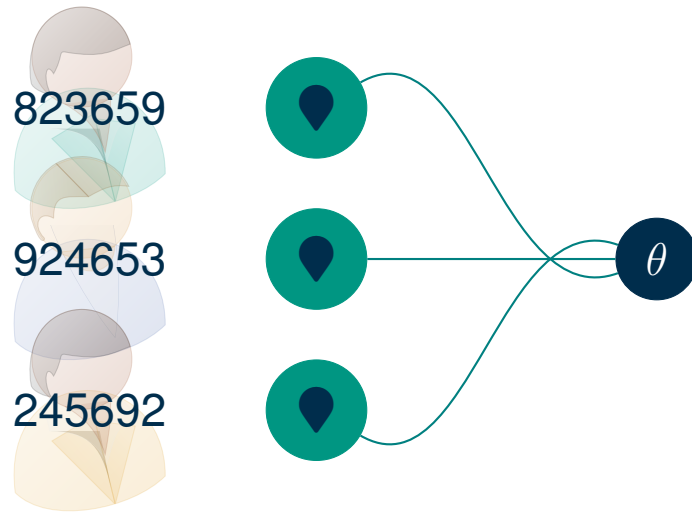
Data analyses & learning processes



Privacy Risk in Data Analysis

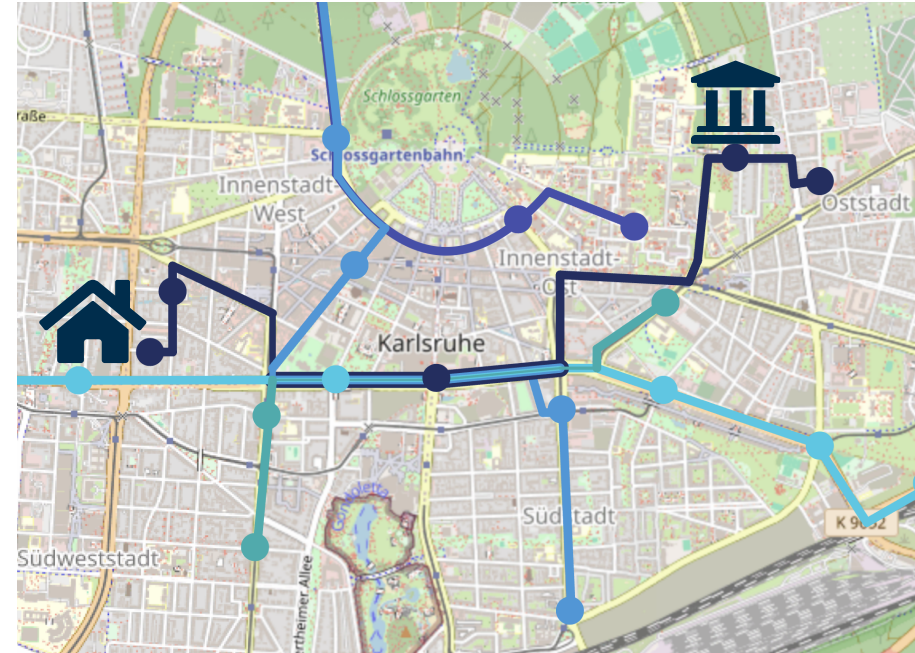


Privacy Risk in Data Analysis

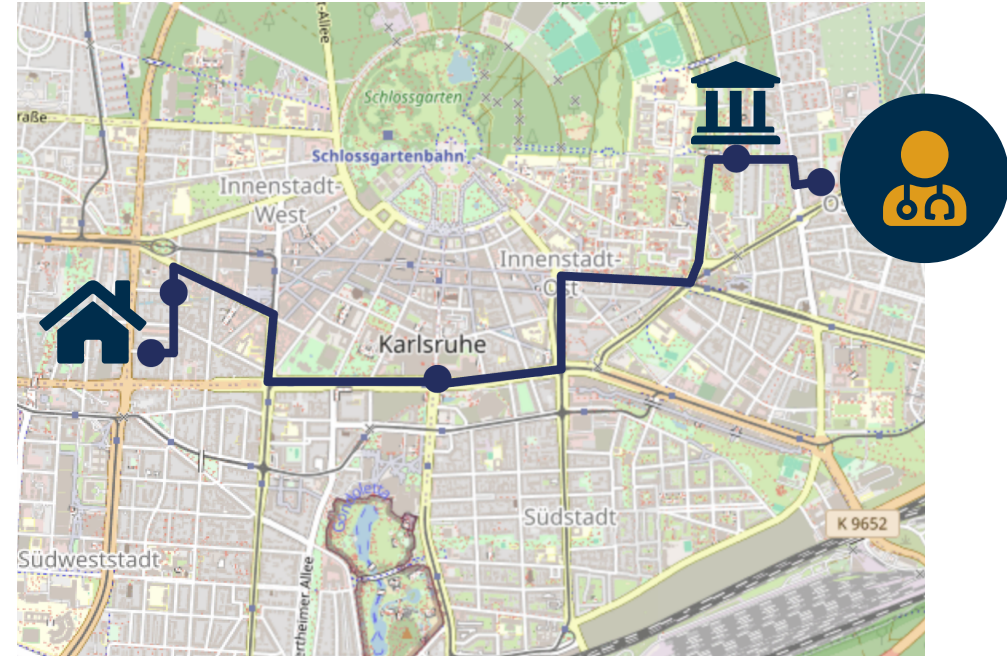
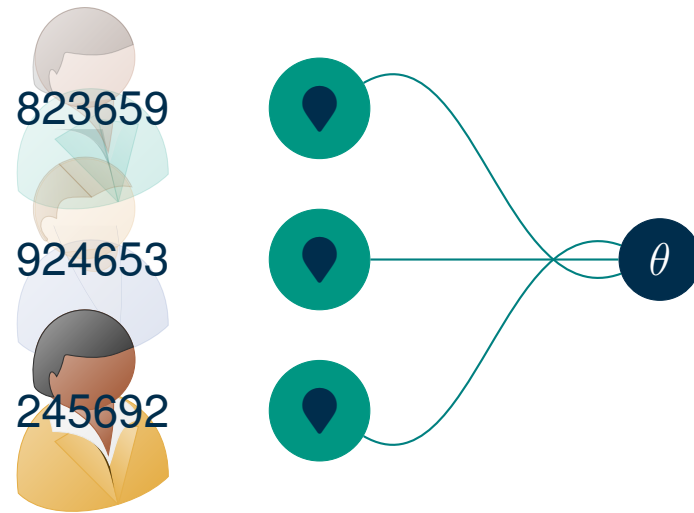


Auxiliary knowledge

-  Alice's home
-  Alice's job



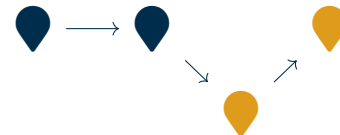
Privacy Risk in Data Analysis



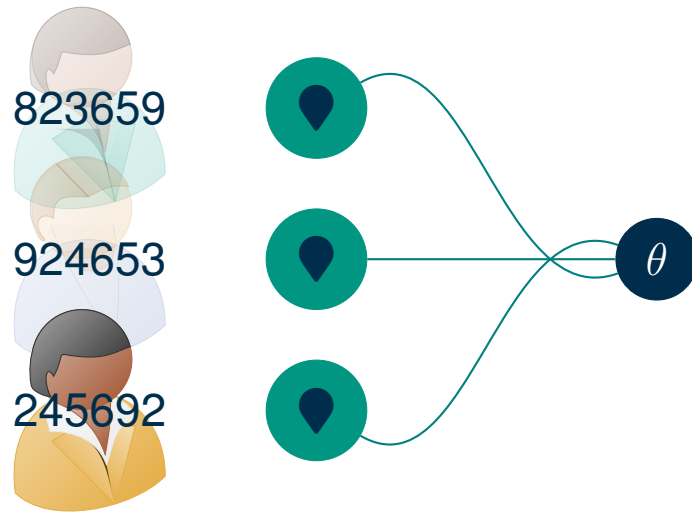
Auxiliary knowledge

-  Alice's home
-  Alice's job

Attack's output



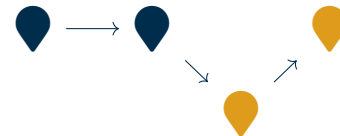
Privacy Risk in Data Analysis



Auxiliary knowledge

- Alice's home
- Alice's job

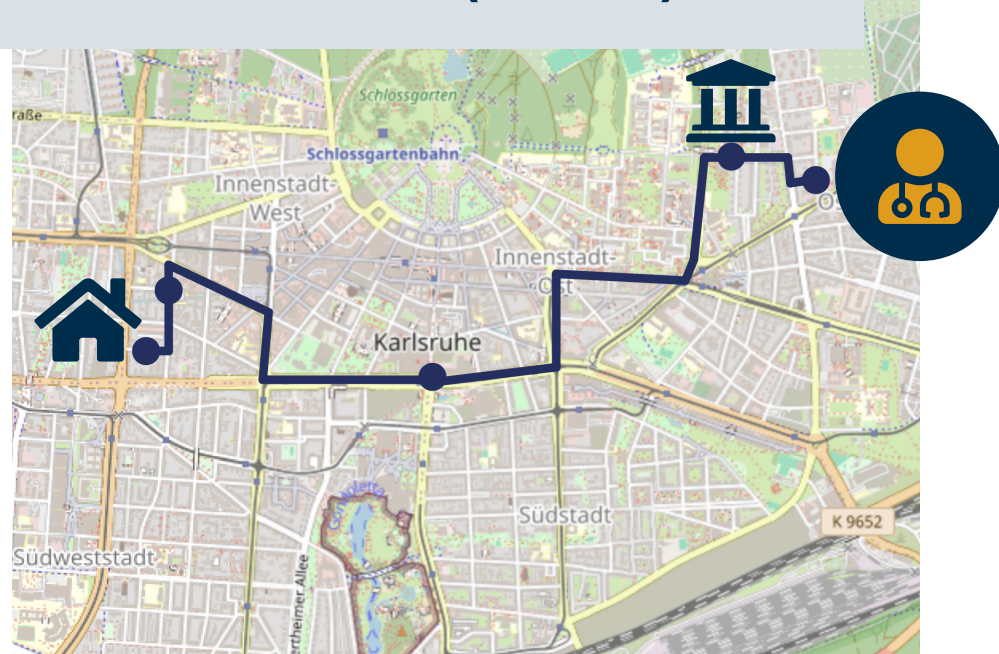
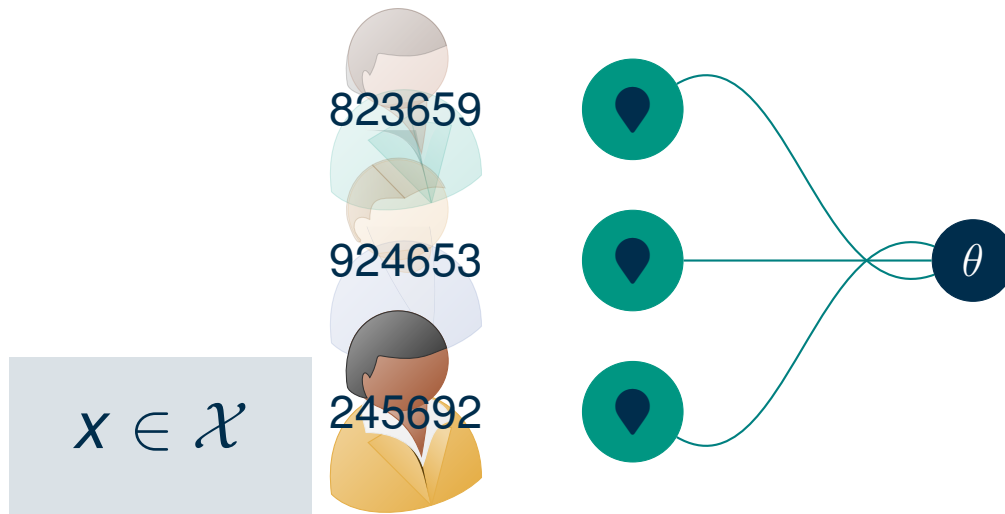
Attack's output



95% of users are uniquely identified with only 4 areas!
(de Montjoye et al.)

Privacy Risk in Data Analysis

Data Reconstruction Attacks (DRAs)

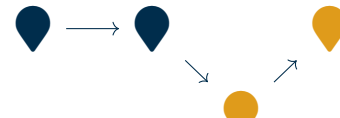


Auxiliary knowledge

-  Alice's home
-  Alice's job

$$a(x) \in aux$$

Attack's output



$$A(\theta, a(x)) = \tilde{x}$$

The attack is successful if the reconstruction error is less than η :

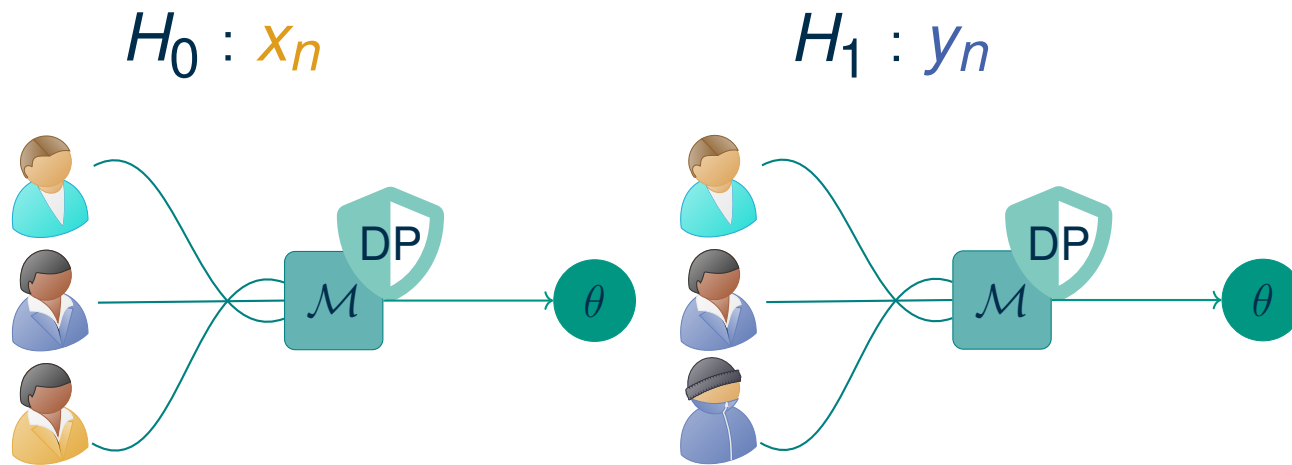
$$\ell(x, \tilde{x}) \leq \eta$$

Important Instances of Reconstruction Attacks

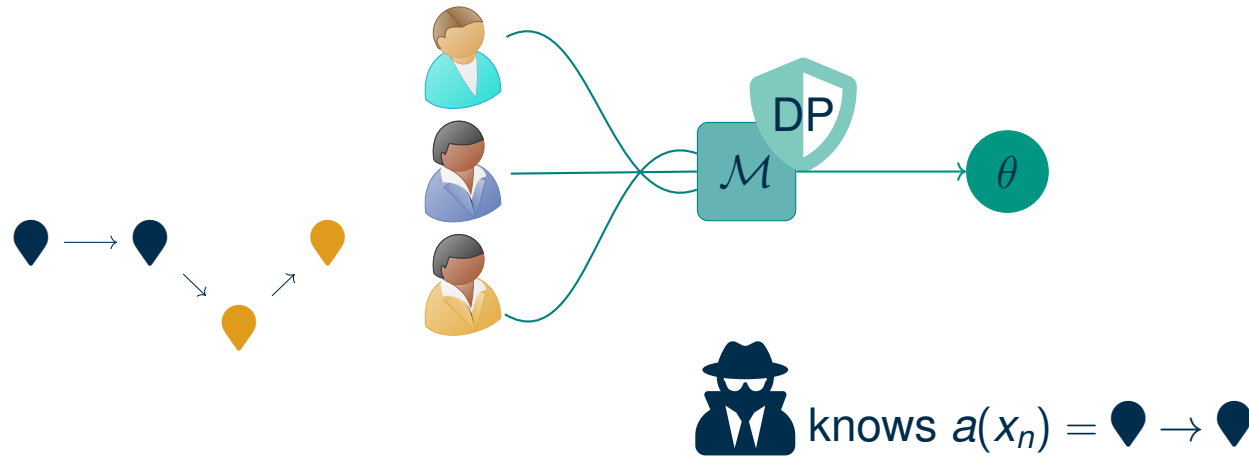
■ Membership Inference Attack:

$$a(x) = x, x_n \in D?$$

$$\ell(x, \tilde{x}) = \mathbf{1}_{\{x \neq \tilde{x}\}}$$



Important Instances of Reconstruction Attacks



■ Membership Inference Attack:

$$a(x) = x, x_n \in D?$$

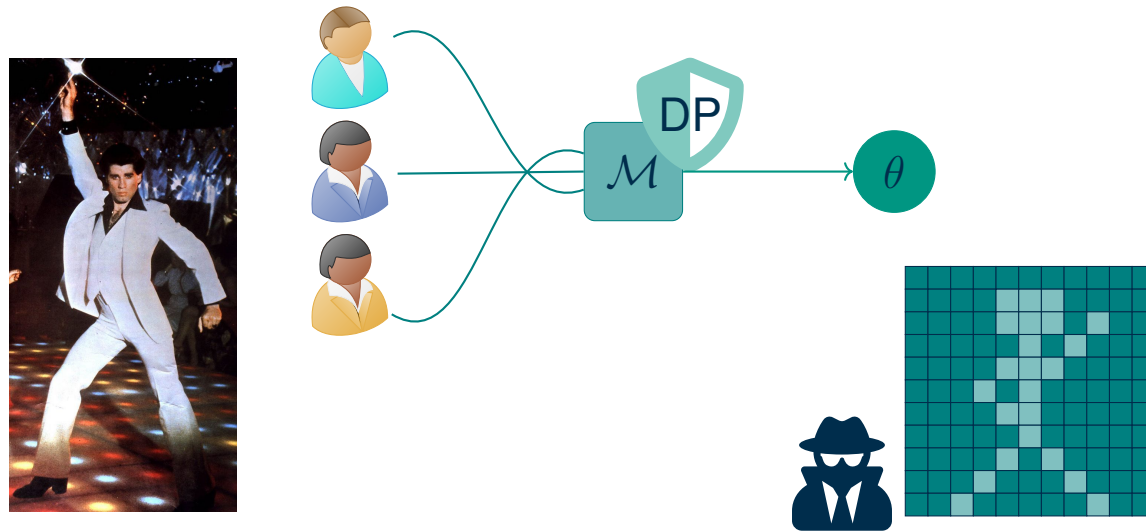
$$\ell(x, \tilde{x}) = \mathbf{1}_{\{x \neq \tilde{x}\}}$$

■ Attribute Inference Attack:

$$a(x) = x \setminus \varphi(x_n), \varphi(x_n)?$$

$$\ell(x, \tilde{x}) = \mathbf{1}_{\{\varphi(x) \neq \varphi(\tilde{x})\}}$$

Important Instances of Reconstruction Attacks



■ Membership Inference Attack:

$$a(x) = x, \mathbf{x}_n \in D?$$

$$\ell(x, \tilde{x}) = \mathbf{1}_{\{x \neq \tilde{x}\}}$$

■ Attribute Inference Attack:

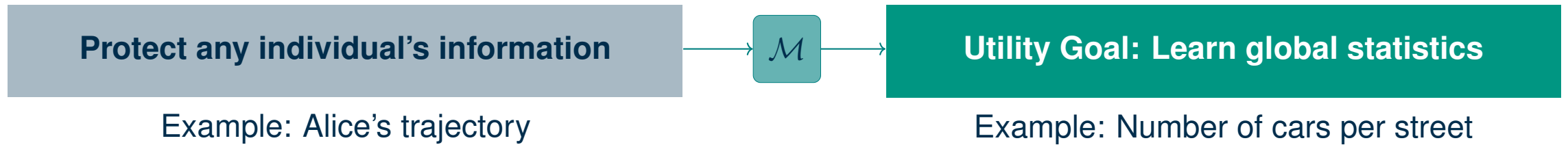
$$a(x) = x \setminus \varphi(\mathbf{x}_n), \varphi(\mathbf{x}_n)?$$

$$\ell(x, \tilde{x}) = \mathbf{1}_{\{\varphi(x) \neq \varphi(\tilde{x})\}}$$

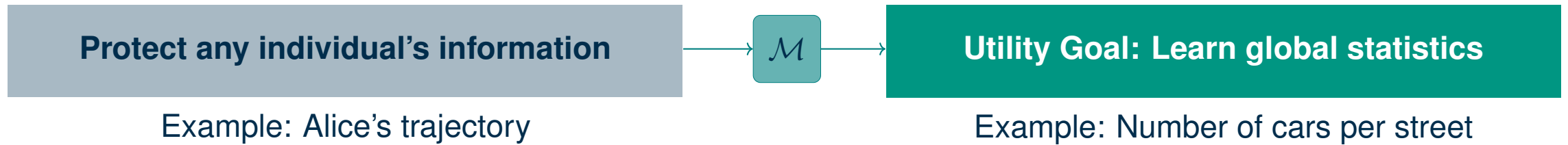
■ Data Reconstruction Attack:

$$a(x) = z, \ell(\mathbf{x}_n, \tilde{\mathbf{x}}_n) \leq \eta?$$

Privacy-preserving Data Analysis



Privacy-preserving Data Analysis



Differential Privacy

Privacy-preserving Data Analysis

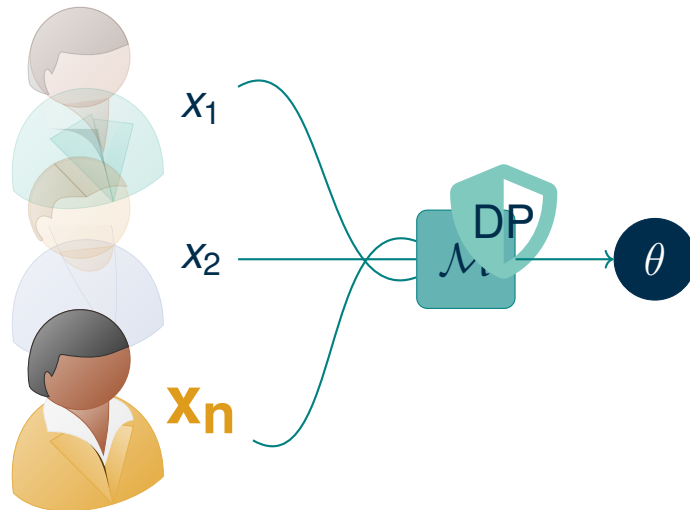
Protect any individual's information

Example: Alice's trajectory

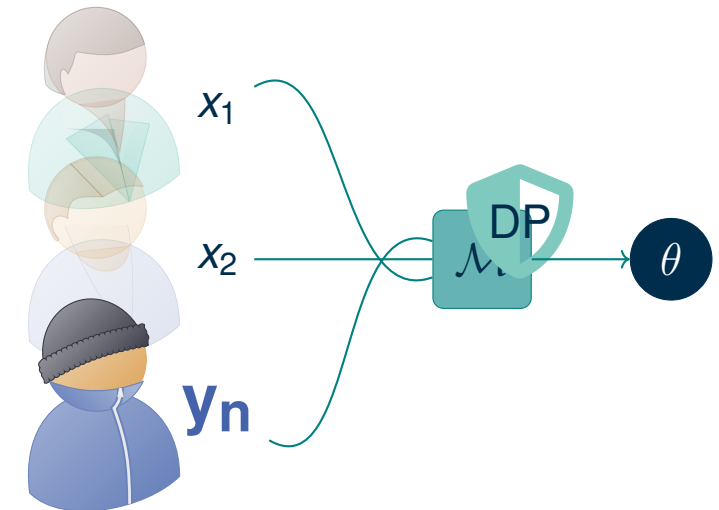


Utility Goal: Learn global statistics

Example: Number of cars per street



Differential Privacy



Privacy-preserving Data Analysis

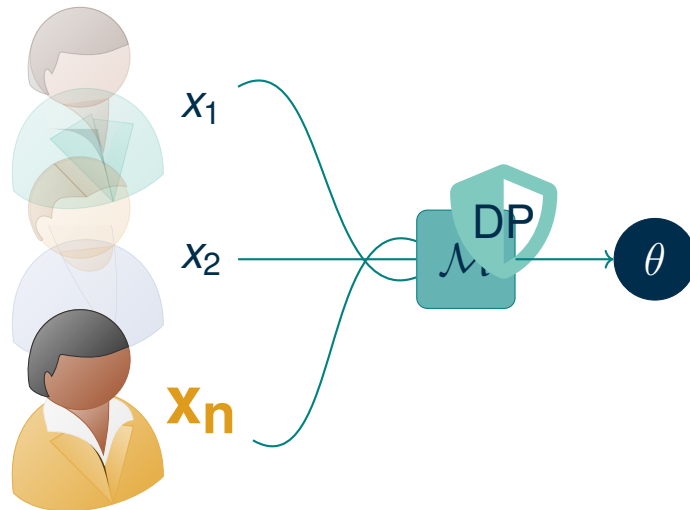
Protect any individual's information

Example: Alice's trajectory

$\mathcal{M}$

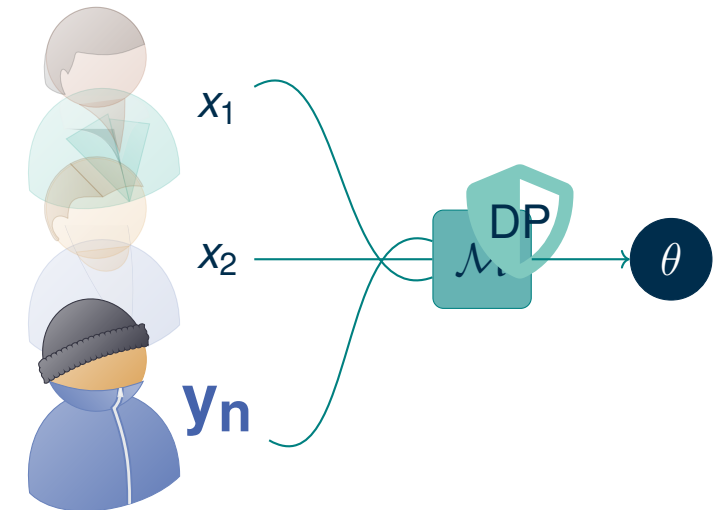
Utility Goal: Learn global statistics

Example: Number of cars per street



Differential Privacy

$$\frac{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{x}_n)}{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{y}_n)} \leq e^\epsilon$$



Privacy-preserving Data Analysis

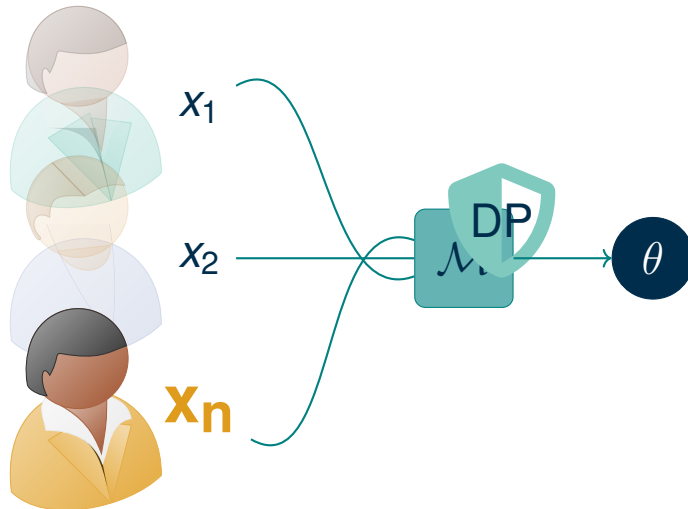
Protect any individual's information

$\mathcal{M}$

Utility Goal: Learn global statistics

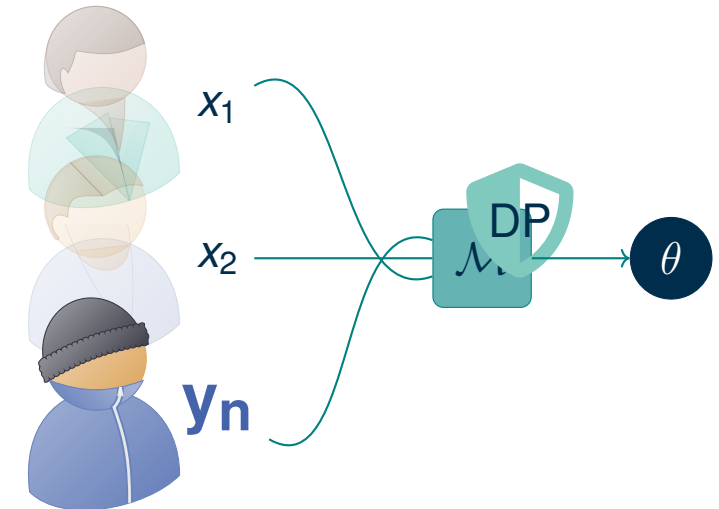
Example: Alice's trajectory

Example: Number of cars per street



Differential Privacy

$$\frac{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{x}_n)}{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{y}_n)} \leq e^\epsilon$$



Privacy-preserving Data Analysis

Protect any individual's information

$\mathcal{M}$

Utility Goal: Learn global statistics

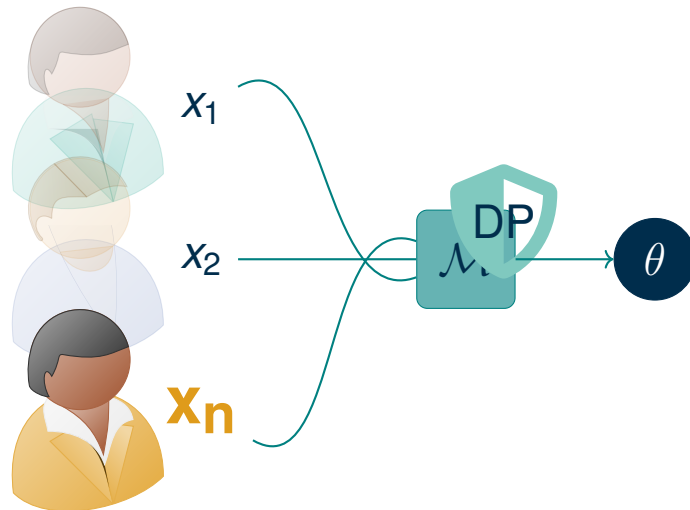
Example: Alice's trajectory

Example: Number of cars per street

0

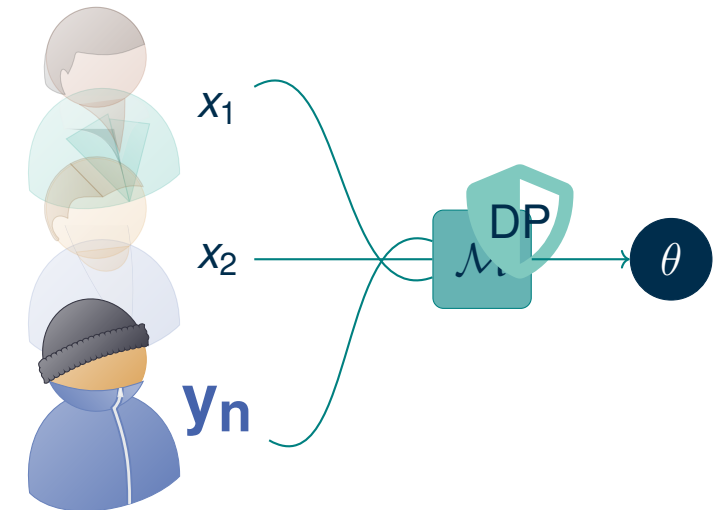
$\epsilon = 70$

$\rightarrow \infty$



Differential Privacy

$$\frac{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{x}_n)}{p_{\mathcal{M}}(\theta \mid x_1, \dots, x_{n-1}, \mathbf{y}_n)} \leq e^\epsilon$$



Differential Privacy and Attack Resilience

- ✓ Robust formal definition.
- ✓ We understand the impact on membership inference attacks.
- ✓ Great privacy–utility trade-offs:

Laplace mechanism


$$\text{Avg} = x \xrightarrow{+ \text{Noise}} \theta$$

$$\text{Error} \xrightarrow{N \rightarrow \infty} 0$$

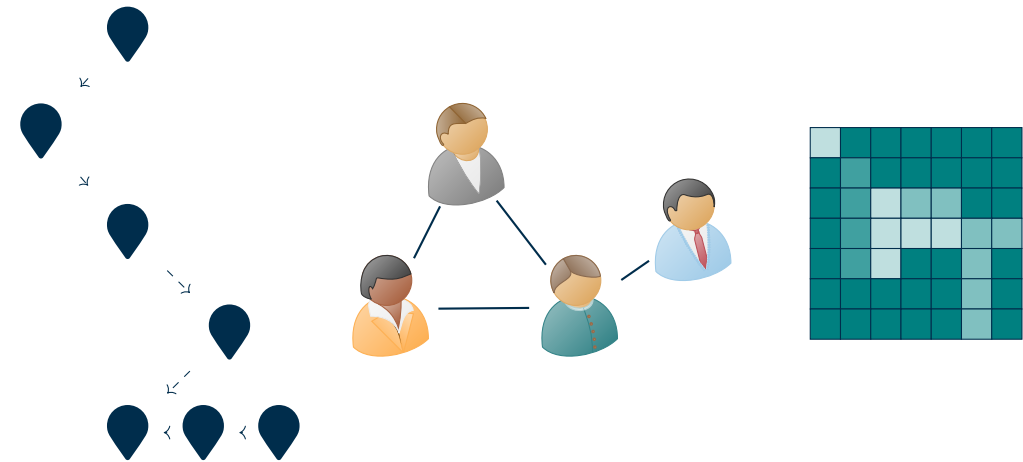
Differential Privacy and Attack Resilience

- ✓ Robust formal definition.
 - ✓ We understand the impact on membership inference attacks.
 - ✓ Great privacy–utility trade-offs:
- ✗ Yet, DP is not applicable in complex data domains.
 - ✗ There are other threats and attacks that DP literature does not cover.

Laplace mechanism


$$\text{Avg} = x \xrightarrow{+ \text{Noise}} \theta$$

$$\text{Error} \xrightarrow{N \rightarrow \infty} 0$$



Differential Privacy and Attack Resilience

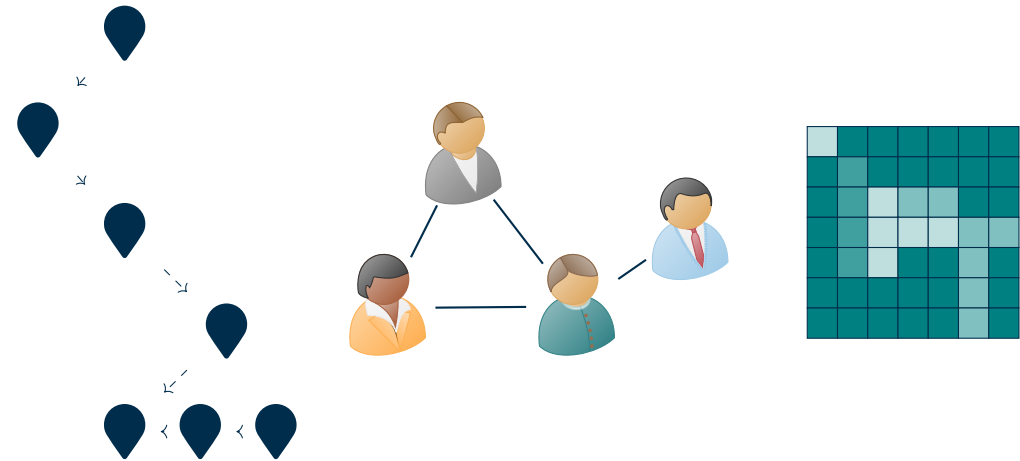
- ✓ Robust formal definition.
- ✓ We understand the impact on membership inference attacks.
- ✓ Great privacy–utility trade-offs:

Laplace mechanism


$$\text{Avg} = x \xrightarrow{+ \text{Noise}} \theta$$

$$\text{Error} \xrightarrow{N \rightarrow \infty} 0$$

- ✗ Yet, DP is not applicable in complex data domains.
- ✗ There are other threats and attacks that DP literature does not cover.



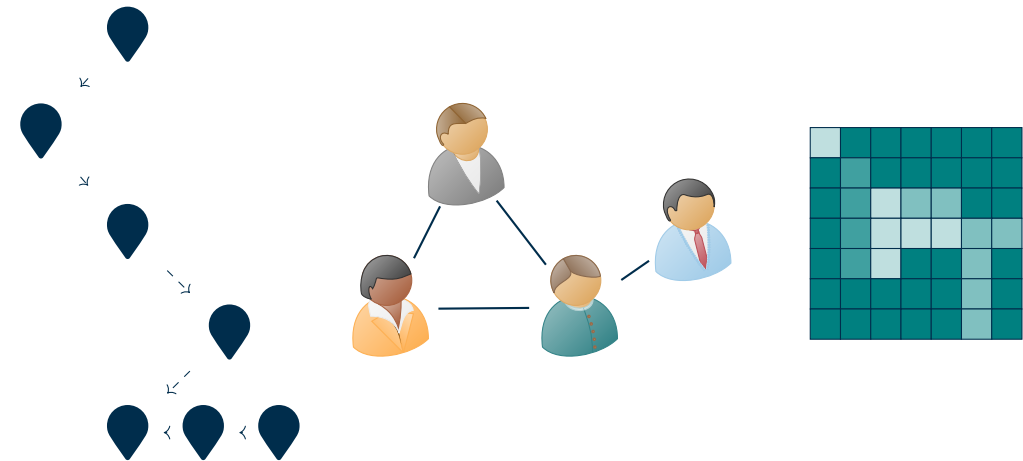
Differential Privacy and Attack Resilience

- ✓ Robust formal definition.
 - ✓ We understand the impact on membership inference attacks.
 - ✓ Great privacy–utility trade-offs:
- ✗ Yet, DP is not applicable in complex data domains.
 - ✗ There are other threats and attacks that DP literature does not cover.

Laplace mechanism


$$\text{Avg} = x \xrightarrow{+ \text{Noise}} \theta$$

$$\text{Error} \xrightarrow{N \rightarrow \infty} 0$$



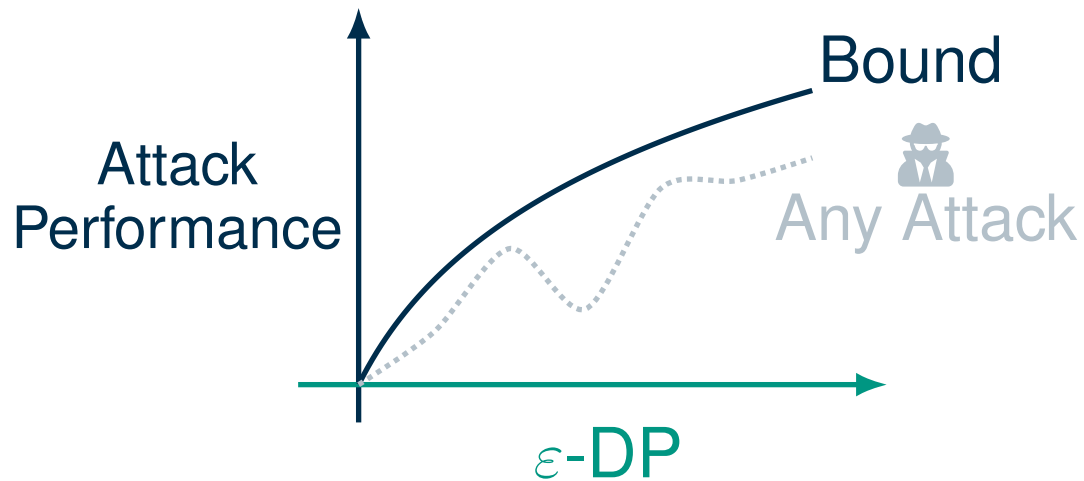
What is the actual risk when using a DP mechanism?

Understanding Disclosure Risk in Differential Privacy

WPES24, VLDB26

Tight and accurate bounds connecting **noise injection** to **actual risk** are crucial for:

 Interpretation  Utility  Auditing

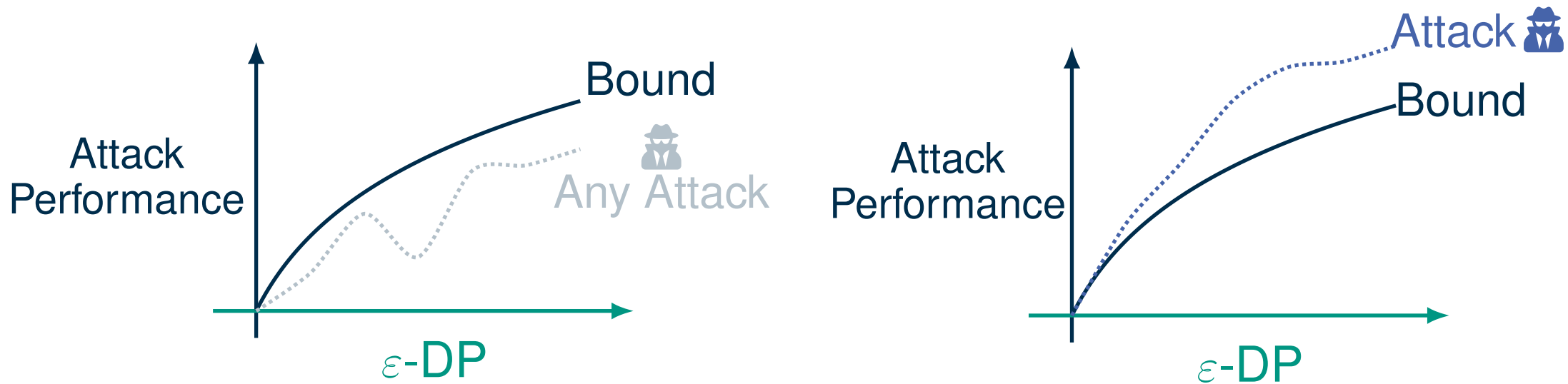


Understanding Disclosure Risk in Differential Privacy

WPES24, VLDB26

Tight and accurate bounds connecting **noise injection** to **actual risk** are crucial for:

 Interpretation  Utility  Auditing

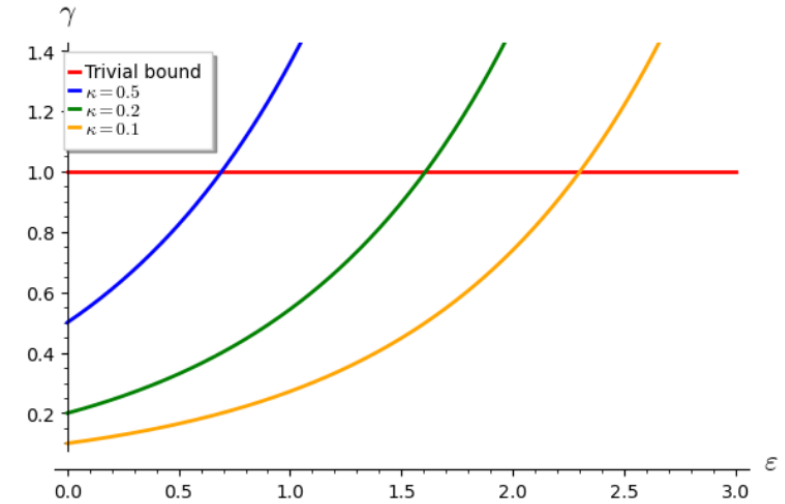


Connecting DP & Reconstruction Attacks

WPES24

η -ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$

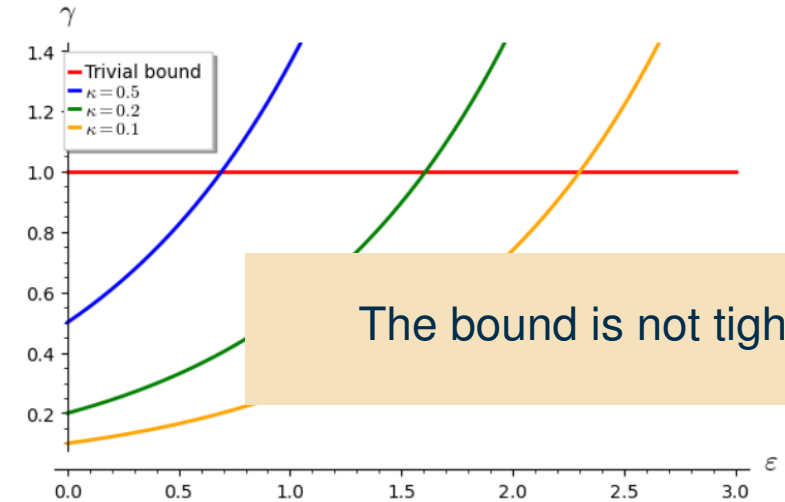


Connecting DP & Reconstruction Attacks

WPES24

η -ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$

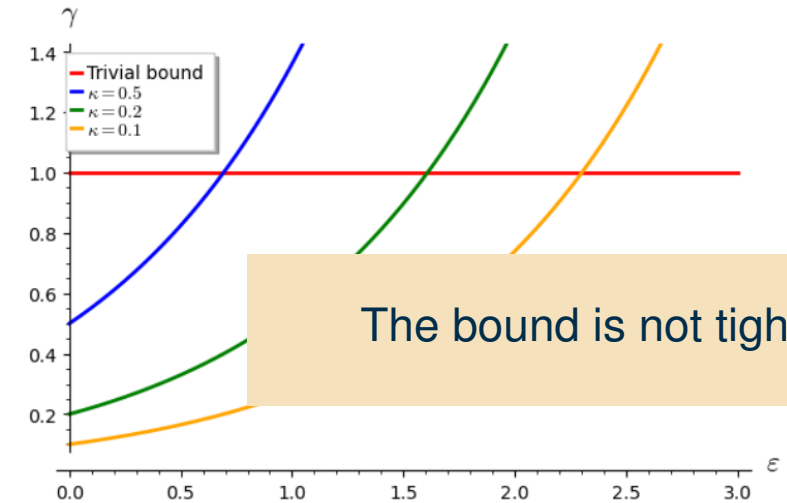


Connecting DP & Reconstruction Attacks

WPES24

η -ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$



The bound is not tight!

ReRo does not consider auxiliary knowledge!

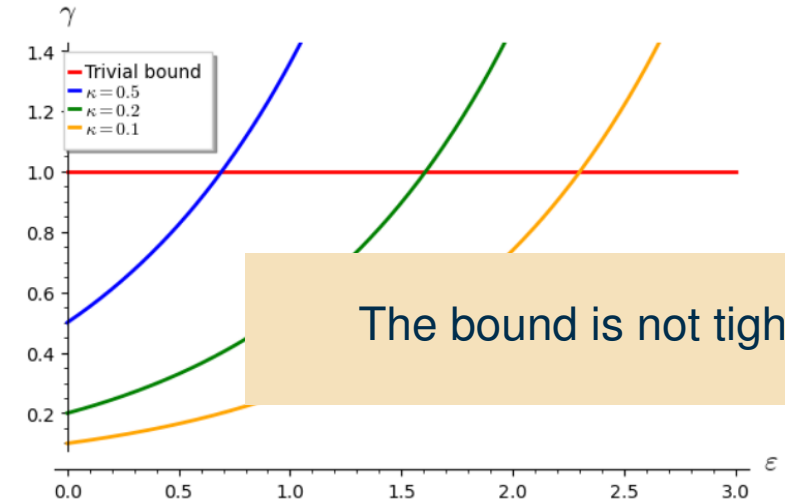


Connecting DP & Reconstruction Attacks

WPES24

η -ReRo
Balle et al.

$$\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta)) \leq \eta]$$

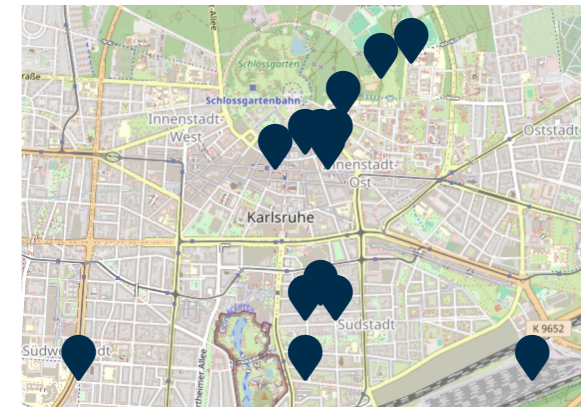


The bound is not tight!

ReRo does not consider auxiliary knowledge!

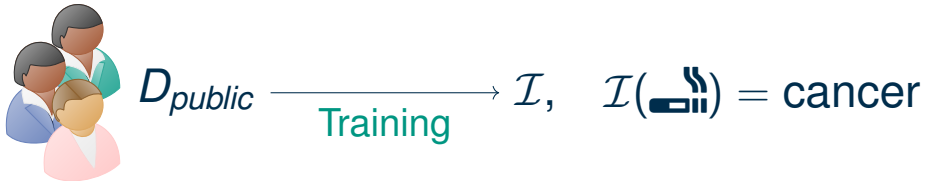


Risk overestimation due to imputation & prior knowledge!



Imputation is Not a Privacy Leakage

A public dataset D_{public} , trains an imputation model $\mathcal{I}: \mathcal{X} \rightarrow \mathcal{S}$ to predict the sensitive attribute from the non-sensitive ones.



DifferentialPrivacy.org

[Home](#) [About](#) [Posts](#) [Resources](#)

Statistical Inference is Not a Privacy Violation

On April 28, 2021, the US Census Bureau [released](#) a new demonstration of its differentially private Disclosure Avoidance System (DAS) for the 2020 US Census. The public were given a month to submit feedback before the system is finalized. This demonstration data and the feedback has generated a lot of discussion, including media coverage on [National Public Radio](#), in [the Washington Post](#), and via [the Associated Press](#). The DAS is also the subject of an [ongoing lawsuit](#).

photography to
Disclosure

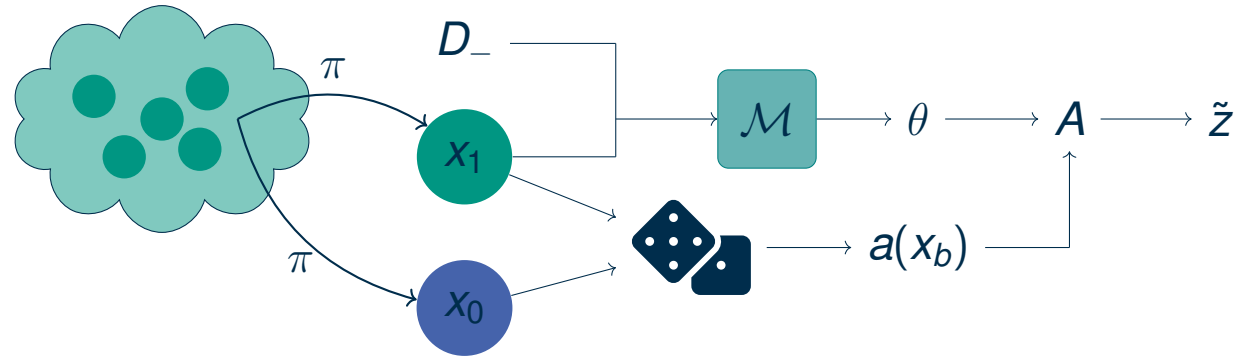
**Learning about the population
implies learning *something*
about the average individual!**

Imputation Attack without data access

Dataset	ReRo
Census	0.81
Texas	0.73

Reconstruction Advantage (RAD)

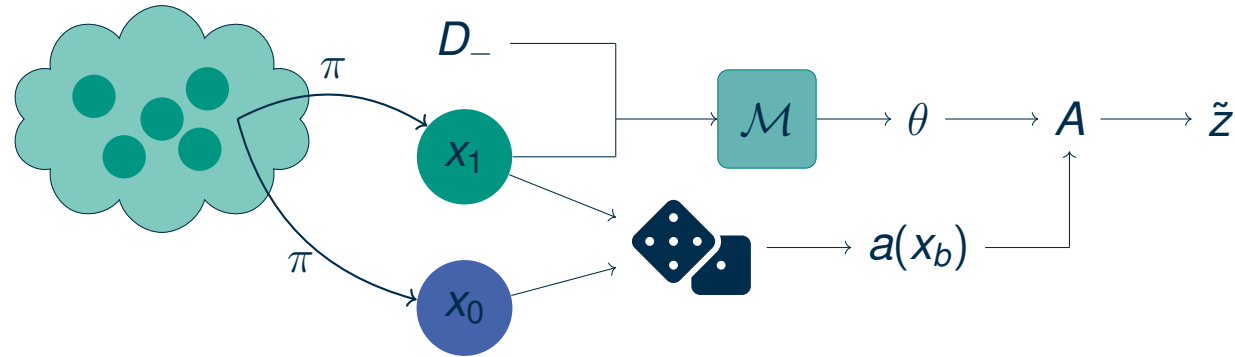
VLDB26



$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta] - \Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]$$

Reconstruction Advantage (RAD)

VLDB26



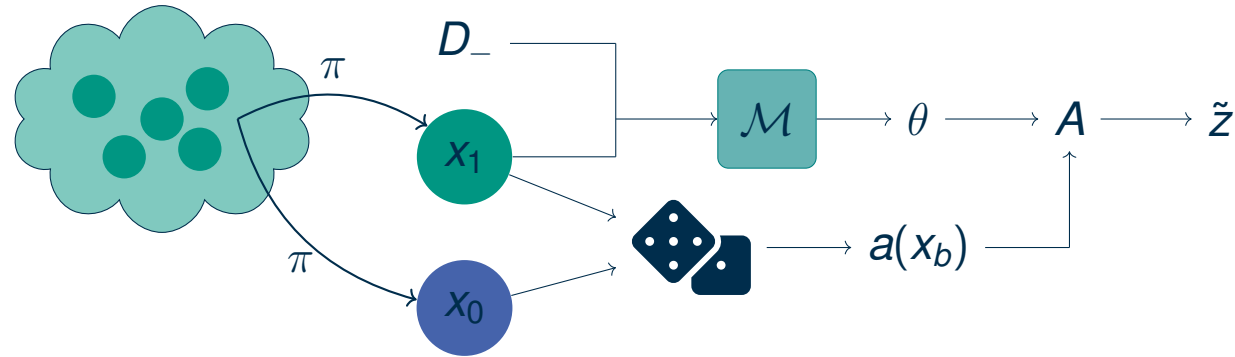
$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, \underbrace{a(X_1)})) \leq \eta] - \Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]$$

An arrow points from the underlined $a(X_1)$ in the first term to the text below.

- ✓ Takes the **auxiliary knowledge** into account!

Reconstruction Advantage (RAD)

VLDB26



$$\eta\text{-RAD} = \Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_1})}} [\ell(X_1, A(\theta, \underbrace{a(X_1)}_{\text{auxiliary knowledge}}))] \leq \eta] - \underbrace{\Pr_{\substack{X_0, X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta]}$$

✓ Takes the **auxiliary knowledge** into account!

✓ Deducts **imputation** and prior knowledge success!

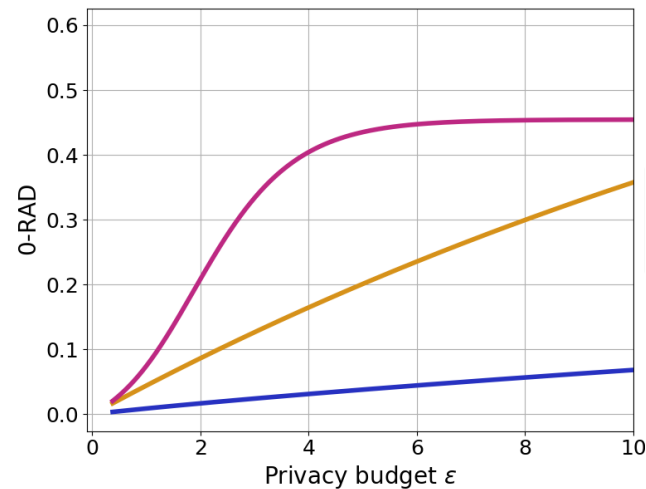
Connecting DP and Reconstruction Advantage

Theorem 5.3.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

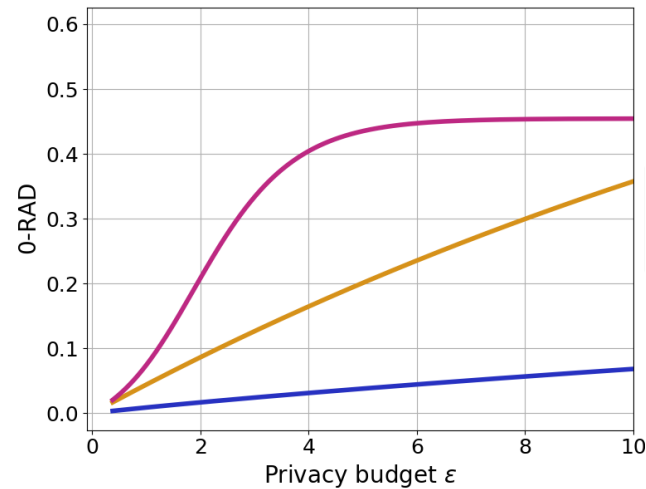
Connecting DP and Reconstruction Advantage

Theorem 5.3.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Different protection under
same ϵ !

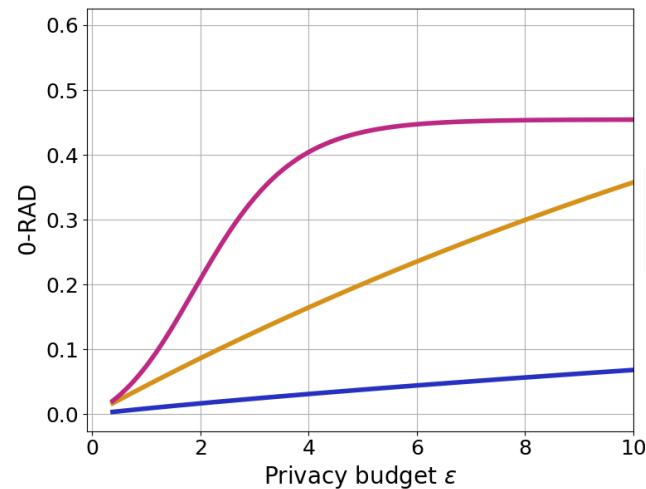
Connecting DP and Reconstruction Advantage

Theorem 5.3.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Algorithm 1: Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Different protection under
same ϵ !

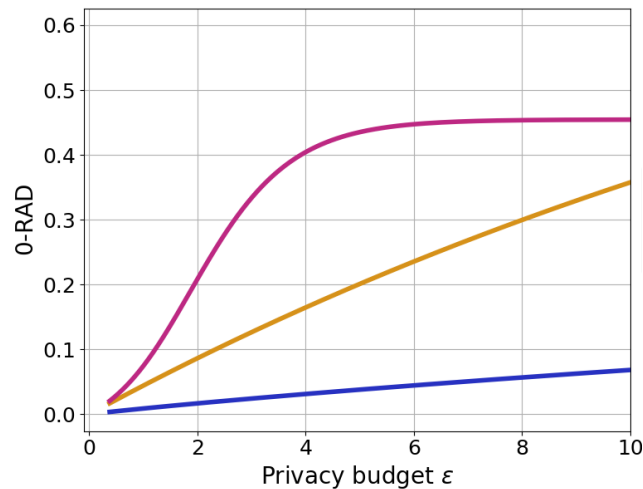
Connecting DP and Reconstruction Advantage

Theorem 5.3.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Algorithm 1: Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

Different protection under
same ϵ !

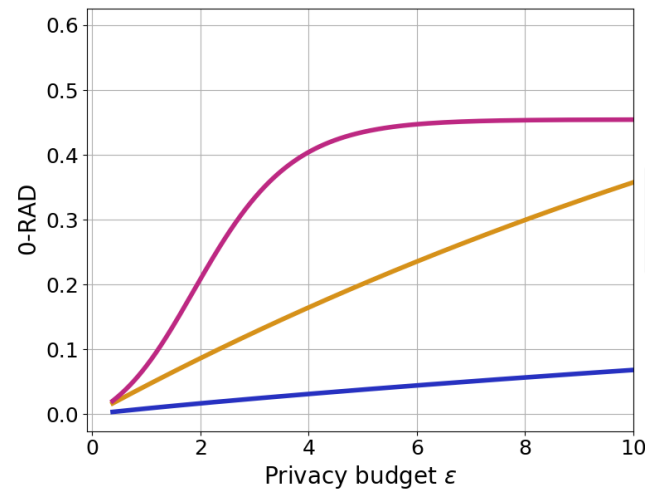
Connecting DP and Reconstruction Advantage

Theorem 5.3.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Algorithm 1: Attack 🕵️

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

Different protection under
same ϵ !

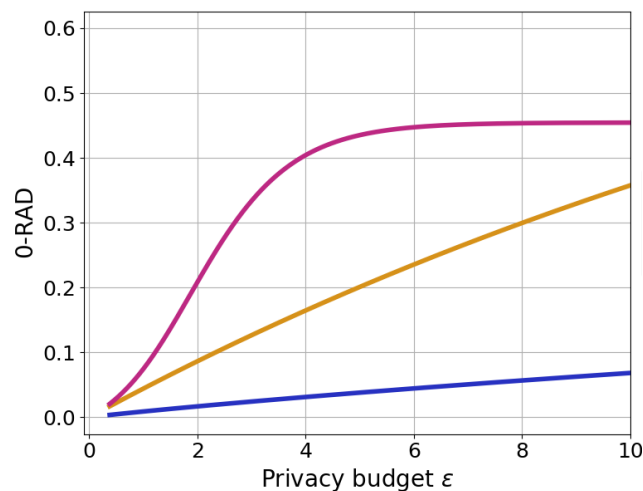
Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

Algorithm 1: Attack 🕵️

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

The bound is tight

Different protection under same ϵ !

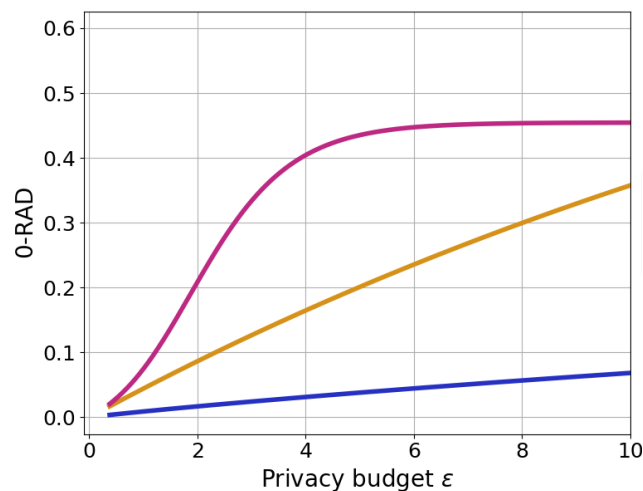
Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

Different protection under
same ϵ !

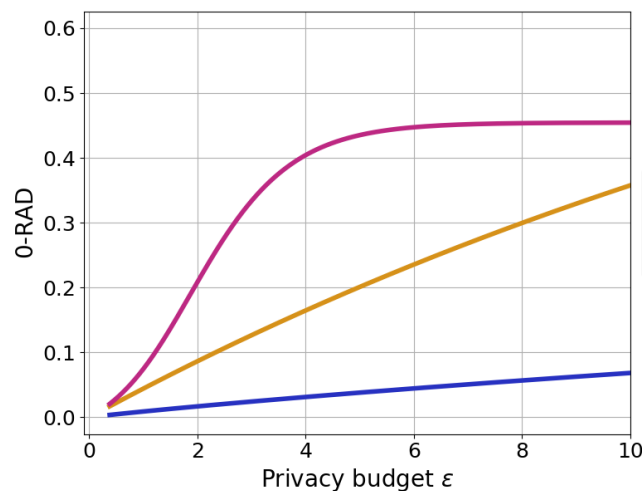
Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

— Th.5.3., $\text{aux} = \{\emptyset\}$, $\eta = 0$.



Gaussian, OUE, and Laplace.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

Different protection under
same ϵ !

Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_{\theta} \in \mathcal{X}} \sum_{\substack{\ell(x, x_{\theta}) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

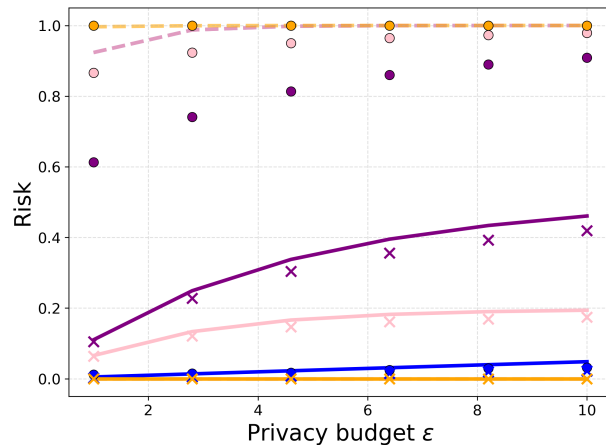
Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_{\eta}^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_{\eta}^z(x')$ (at random)

●=ReRo, ×=RAD, $\eta = 0, 40, 80, 100$ hours.



Laplace Mechanism

Different protection under same ϵ !



Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

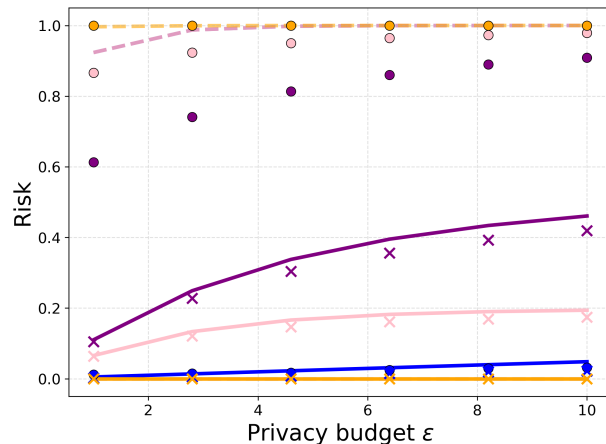
Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

●=ReRo, ×=RAD, $\eta = 0, 40, 80, 100$ hours.



Laplace Mechanism

Different protection under same ϵ !



✓ ReRo bounds are not tight vs. RAD bound is **universally tight**.

Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

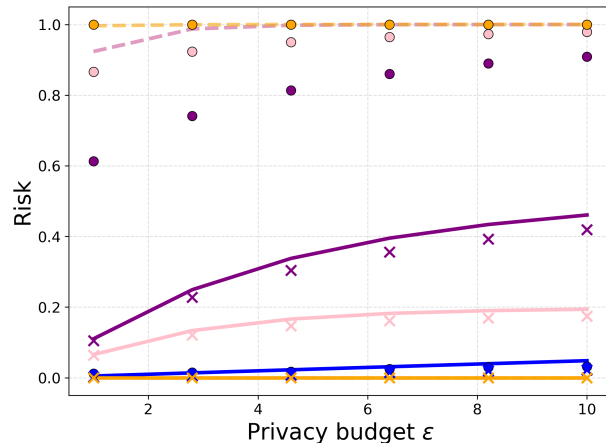
Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

●=ReRo, ×=RAD, $\eta = 0, 40, 80, 100$ hours.



Laplace Mechanism

Different protection under same ϵ !



- ✓ ReRo bounds are not tight vs. RAD bound is **universally tight**.
- ✓ RAD corrects **ReRo overestimation**.

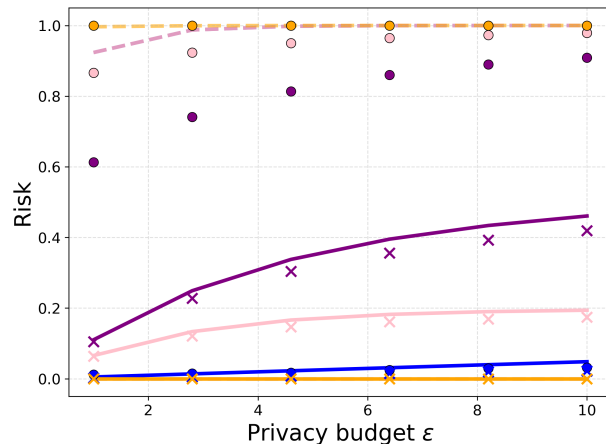
Connecting DP and Reconstruction Advantage

Theorem 5.3. (Universally Tight Bound)

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_{\theta} \in \mathcal{X}} \sum_{\substack{\ell(x, x_{\theta}) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x,$$

where $w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$.

●=ReRo, ×=RAD, $\eta = 0, 40, 80, 100$ hours.



Laplace Mechanism

- ✓ ReRo bounds are not tight vs. RAD bound is **universally tight**.
- ✓ RAD corrects **ReRo overestimation**.

The attack is optimal

The bound is tight

Algorithm 1: Optimal Attack 🛡️

Input : θ, η and $a(x) = z$

Output : $\tilde{x}$

Compute $a^{-1}(z) = \{x : a(x) = z\}$

for $x' \in \mathcal{X}$ **do**

$\mathcal{W}_{\eta}^z(x') = \sum_{x \in a^{-1}(z) : \ell(x, x') \leq \eta} w(\theta, x) \pi_x;$

Select $\tilde{x} \in \arg \max_{x'} \mathcal{W}_{\eta}^z(x')$ (at random)

Different protection under same ϵ !

Limitations:

- ✗ We need to know the **auxiliary knowledge**.
- ✗ The bound does not always have a **closed form**.

Upper Bounds on Reconstruction Advantage

- ✓ We provide **close-form upper-bounds** to apply when our optimal bound is not possible.

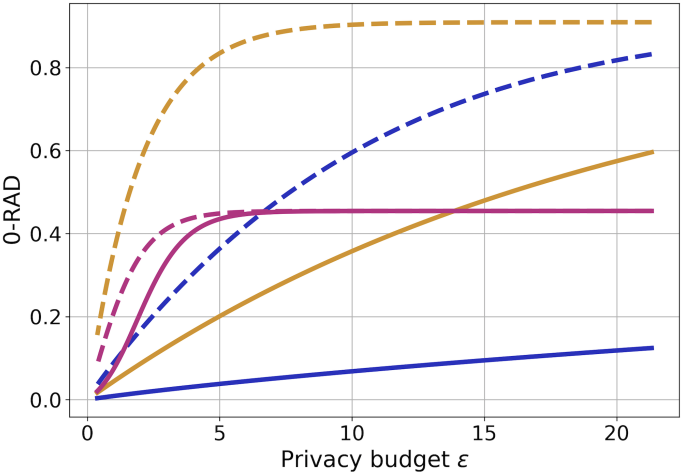
Notion	Assumptions	RAD	Composition	ReRo
TV	—	Th.5.2.	✓	⊥
f -DP	$aux = \{\emptyset\}$	Th.5.11.	✓	Hayes et al.
$\mathcal{M}$	aux known	Th.5.3.	×	⊥

Upper Bounds on Reconstruction Advantage

- ✓ We provide **close-form upper-bounds** to apply when our optimal bound is not possible.

Notion	Assumptions	RAD	Composition	ReRo
TV	—	Th.5.2.	✓	⚡
f -DP	$aux = \{\emptyset\}$	Th.5.11.	✓	Hayes et al.
$\mathcal{M}$	aux known	Th.5.3.	×	⚡

---Th.5.2. vs. —Th.5.3. ($aux = \{\emptyset\}$)



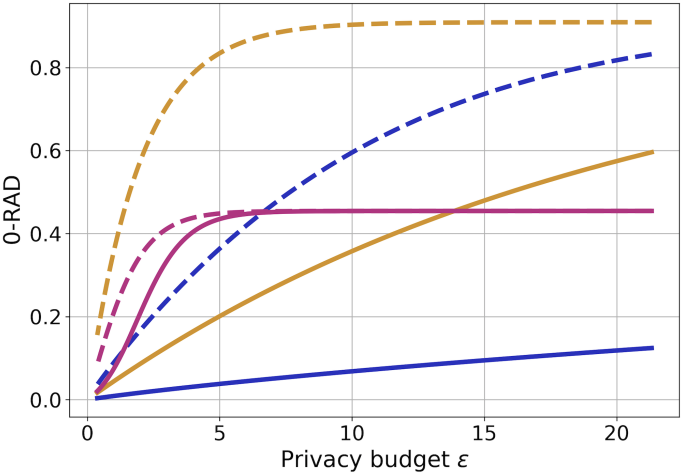
Gaussian, OUE, and Laplace.

Upper Bounds on Reconstruction Advantage

- ✓ We provide **close-form upper-bounds** to apply when our optimal bound is not possible.

Notion	Assumptions	RAD	Composition	ReRo
TV	—	Th.5.2.	✓	⚡
f -DP	$aux = \{\emptyset\}$	Th.5.11.	✓	Hayes et al.
$\mathcal{M}$	aux known	Th.5.3.	×	⚡
(ϵ, δ) -DP	—	Th.5.2.	✓	⚡
(ϵ, δ) -DP	$aux = \{\emptyset\}$	Prop.5.13.	✓	Balle et al.
(ϵ, δ) -DP	$aux = \{\emptyset\}, \eta = 0$	Th.5.15.	✓	Balle et al.

---Th.5.2. vs. —Th.5.3. ($aux = \{\emptyset\}$)



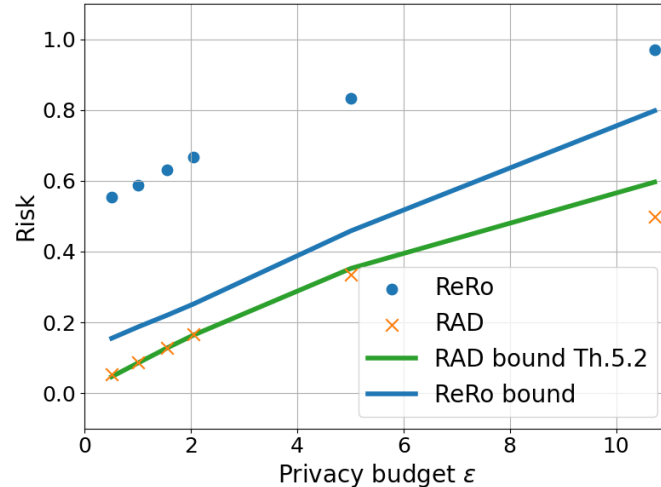
Gaussian, OUE, and Laplace.

Upper Bounds on Reconstruction Advantage

✓ We provide **close-form upper-bounds** to apply when our optimal bound is not possible.

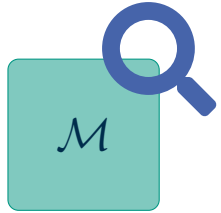
👤 We use our **optimal attack** to test their **tightness** empirically.

DP-SGD for image classification,
 $a(x)$ = image label, $\eta = 0$.



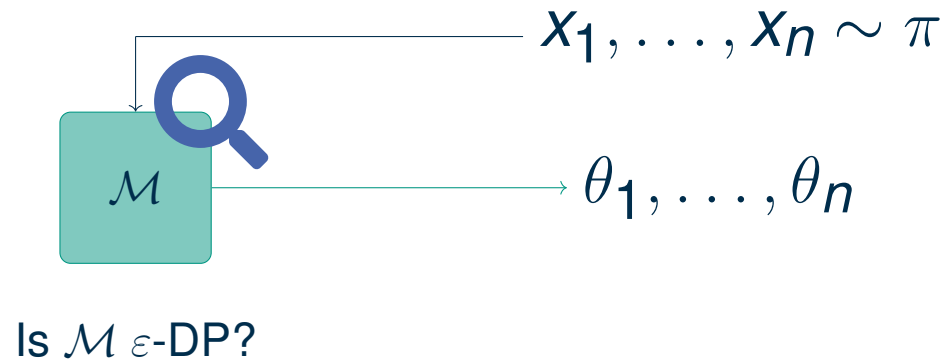
Notion	Assumptions	RAD	Composition	ReRo
TV	—	Th.5.2.	✓	⚡
f -DP	$aux = \{\emptyset\}$	Th.5.11.	✓	Hayes et al.
$\mathcal{M}$	aux known	Th.5.3.	×	⚡
(ϵ, δ) -DP	—	Th.5.2.	✓	⚡
(ϵ, δ) -DP	$aux = \{\emptyset\}$	Prop.5.13.	✓	Balle et al.
(ϵ, δ) -DP	$aux = \{\emptyset\}, \eta = 0$	Th.5.15.	✓	Balle et al.

RAD-based DP Auditing



Is $\mathcal{M}$ ϵ -DP?

RAD-based DP Auditing



RAD-based DP Auditing

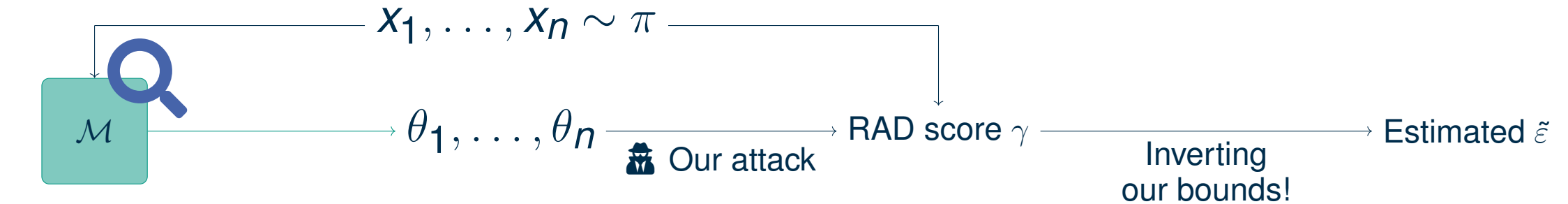


RAD-based DP Auditing



**We improve scalability
over previous auditing
methods!**

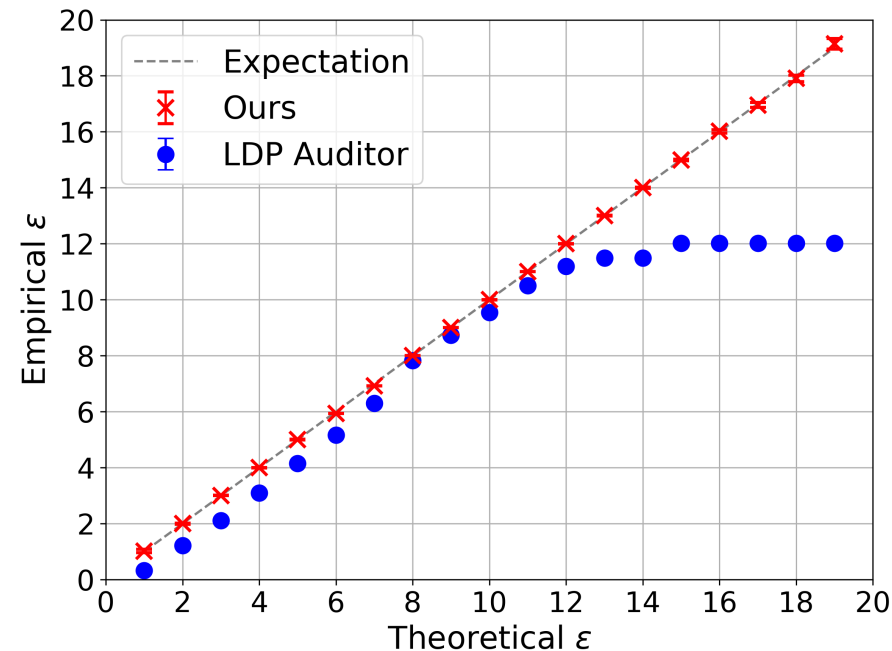
RAD-based DP Auditing



Is $\mathcal{M}$ ϵ -DP?

**We improve scalability
over previous auditing
methods!**

**We improve accuracy
over LDP Auditor!**



Takeaways

ϵ is not enough!

**RAD & our bounds help
understanding participation
risk**

Participation risk $\neq$ harm

- ✗ RAD does not help understanding harms
- ✗ RAD does not help understanding how sharing my data can influence others

Takeaways

ϵ is not enough!

RAD & our bounds help understanding participation risk

Participation risk $\neq$ harm

- ✗ RAD does not help understanding harms
- ✗ RAD does not help understanding how sharing my data can influence others

RAD & our bounds help verifying and auditing protection mechanisms

- ✓ We can detect errors and pitfalls better!
- ✓ We can detect scams and misconducts!

Takeaways

ϵ is not enough!

RAD & our bounds help understanding participation risk

Participation risk $\neq$ harm

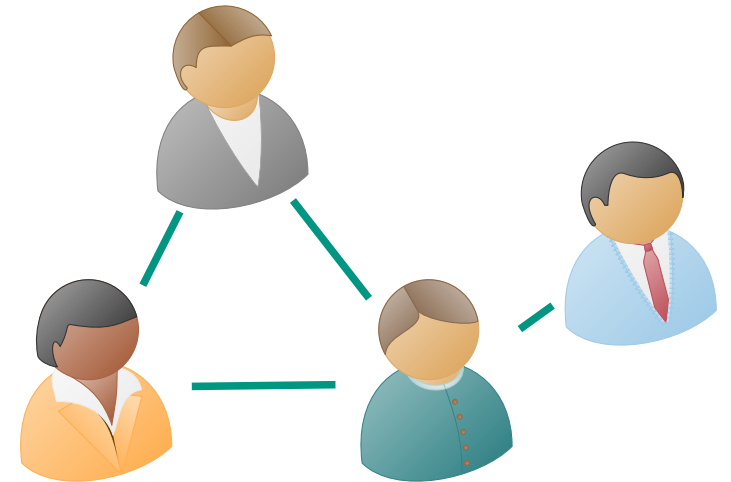
- ✗ RAD does not help understanding harms
- ✗ RAD does not help understanding how sharing my data can influence others

RAD & our bounds help verifying and auditing protection mechanisms

- ✓ We can detect errors and pitfalls better!
- ✓ We can detect scams and misconducts!

Limitation:

Neither DP or RAD take the effect of correlation into account!



Correlation Challenge in DP

Differential privacy fails to measure privacy leakage under correlation

  Theoretically exposed

  Empirically confirmed

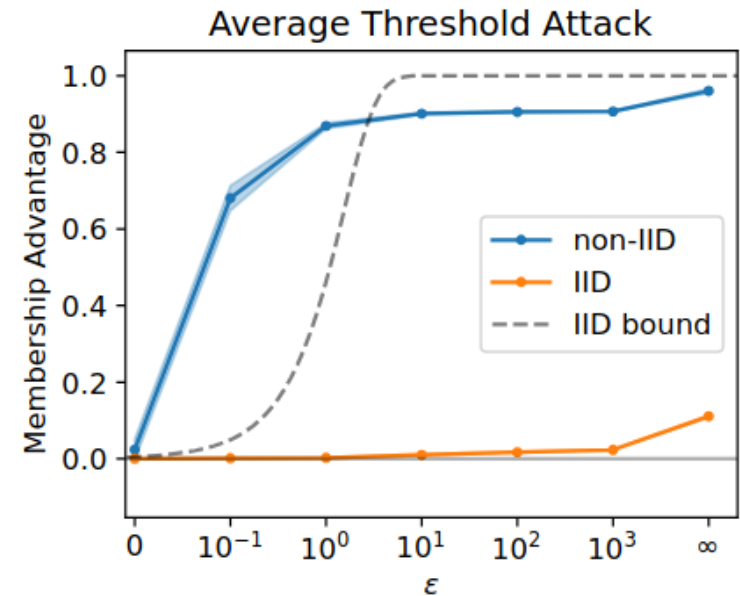


Figure: Humphries et al.'s attack breaks DP guarantees.

Correlation Challenge in DP

Differential privacy fails to measure privacy leakage under correlation

  Theoretically exposed

  Empirically confirmed

Does everybody agree?

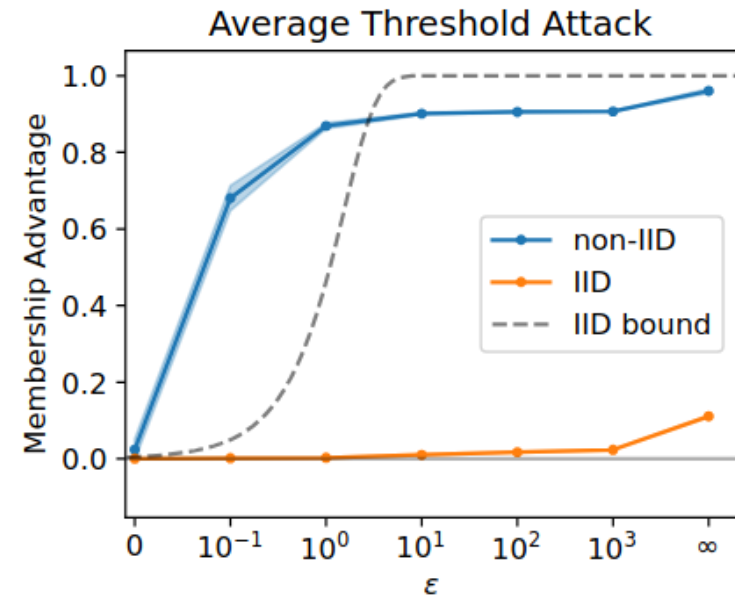


Figure: Humphries et al.'s attack breaks DP guarantees.

Pufferfish: A Framework for Mathematical Privacy Definitions

DANIEL KIFER, Penn State University
ASHWIN MACHANAVAJHALA, Duke University

In this article, we introduce a new and general privacy framework called Pufferfish. The Pufferfish framework can be used to create new privacy definitions that are customized to the needs of a given application. The goal of Pufferfish is to allow experts in an application domain, who frequently do not have expertise in privacy, to develop rigorous privacy definitions for their data sharing needs. In addition to this, the Pufferfish framework can also be used to study existing privacy definitions.

We illustrate the benefits with several applications of this privacy framework: we use it to analyze differ-

3

Yes, correlation is a big deal!

DifferentialPrivacy.org Home About Posts Resources

Statistical Inference is Not a Privacy Violation

On April 28, 2021, the US Census Bureau released a new demonstration of its differentially private Disclosure Avoidance System (DAS) for the 2020 US Census. The public were given a month to submit feedback before the system is finalized. This demonstration data and the feedback has generated a lot of discussion, including media coverage on National Public Radio, in the Washington Post, and via the Associated Press. The DAS is also the subject of an ongoing lawsuit.

NO! It is a statistical inference fallacy!

SoK: Differential Privacy as a Causal Property

Michael Carl Tschantz*, Shayak Sen*, and Anupam Datta*
*International Computer Science Institute / Carnegie Mellon University

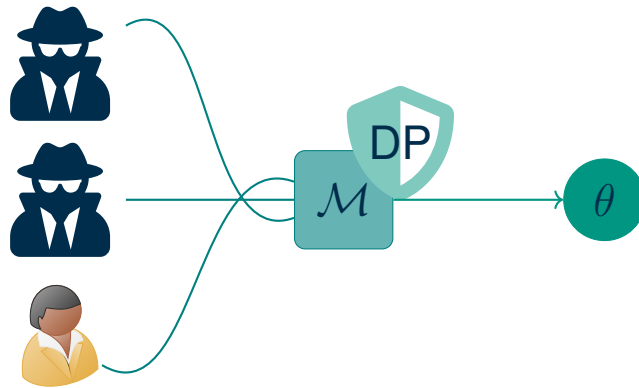
Abstract—We present formal models of the associative and causal views of differential privacy. Under the associative view, the possibility of dependencies between data points precludes a simple statement of differential privacy's guarantee as conditioning upon a single changed data point. However, we show that a simple characterization of differential privacy as limiting the effect of a single data point does exist under the causal view, without independence assumptions about data points. We believe this characterization resolves disagreements and confusions in prior work about the consequences of differential privacy. The associative view needing assumptions boils down to the contra-

Both views are relevant!

Membership Inference Attack Knowing D_-

The attacker receives θ and aims to distinguish between:

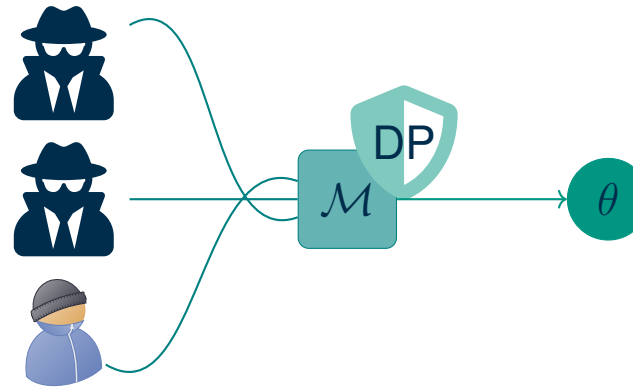
$$H_0 : x_n$$



Type I error:

$$\alpha = \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{x_n})$$

$$H_1 : y_n$$



Type II error:

$$\beta = 1 - \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{y_n})$$

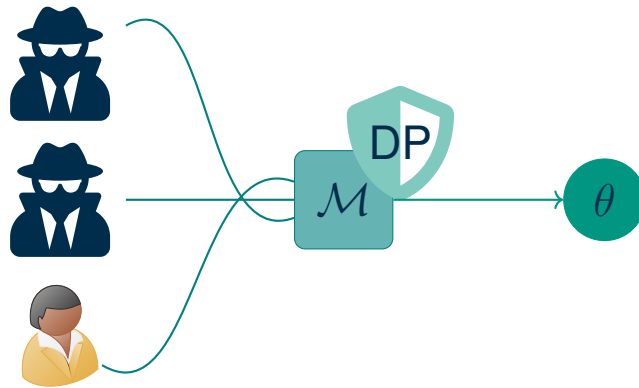
D_- is known:

$$H_0 = D_{x_n} \text{ Vs. } H_1 = D_{y_n}$$

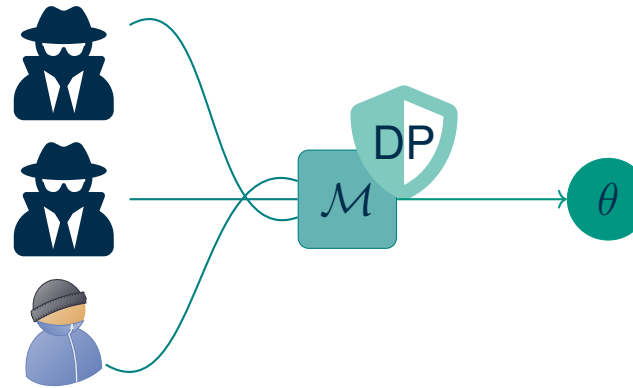
Membership Inference Attack Knowing D_-

The attacker receives θ and aims to distinguish between:

$H_0 : x_n$



$H_1 : y_n$



D_- is known:

$H_0 = D_{x_n}$ Vs. $H_1 = D_{y_n}$

Type I error:

$$\alpha = \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{x_n})$$

Type II error:

$$\beta = 1 - \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{y_n})$$

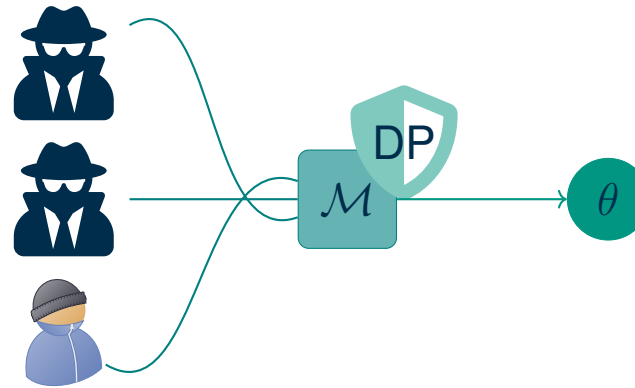
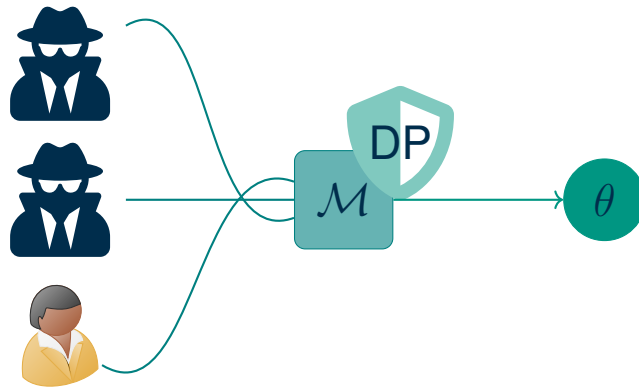
$$A \circ \mathcal{M} \text{ is } \varepsilon\text{-DP} \Rightarrow \begin{aligned} 1 - \beta &\leq e^\varepsilon \alpha \\ \alpha &\leq e^\varepsilon (1 - \beta) \end{aligned}$$

Membership Inference Attack Knowing D_-

The attacker receives θ and aims to distinguish between:

$H_0 : x_n$

$H_1 : y_n$



D_- is known:

$H_0 = D_{x_n}$ Vs. $H_1 = D_{y_n}$

Type I error:

$$\alpha = \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{x_n})$$

Type II error:

$$\beta = 1 - \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{y_n})$$

$$A \circ \mathcal{M} \text{ is } \varepsilon\text{-DP} \Rightarrow \begin{aligned} 1 - \beta &\leq e^\varepsilon \alpha \\ \alpha &\leq e^\varepsilon (1 - \beta) \end{aligned}$$

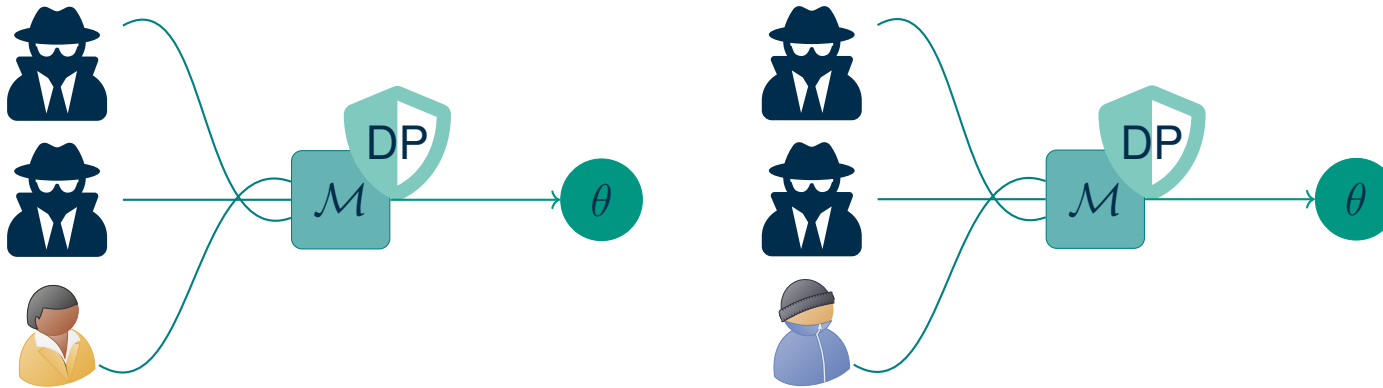
$$\beta \geq \max\{1 - e^\varepsilon \alpha, e^\varepsilon (1 - \alpha)\}$$

Membership Inference Attack Knowing D_-

The attacker receives θ and aims to distinguish between:

$H_0 : x_n$

$H_1 : y_n$



Type I error:

$$\alpha = \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{x_n})$$

Type II error:

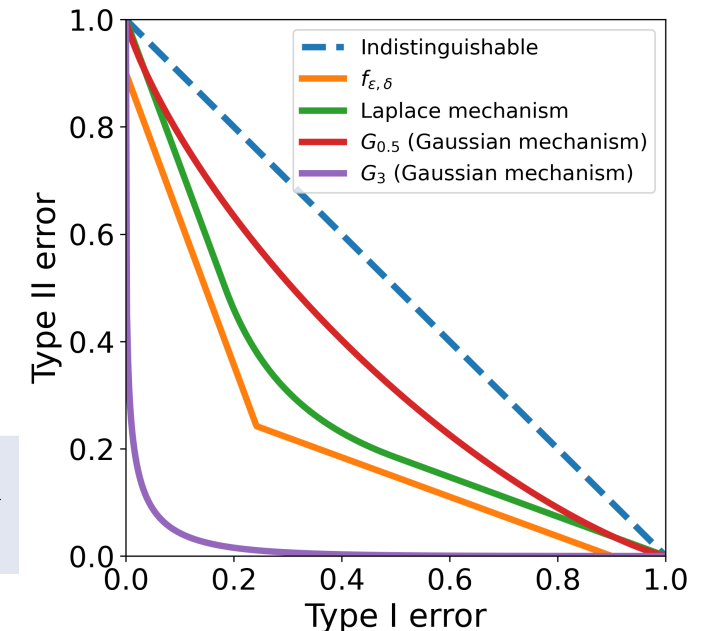
$$\beta = 1 - \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid D_{y_n})$$

$$A \circ \mathcal{M} \text{ is } \varepsilon\text{-DP} \Rightarrow \begin{aligned} 1 - \beta &\leq e^\varepsilon \alpha \\ \alpha &\leq e^\varepsilon (1 - \beta) \end{aligned}$$

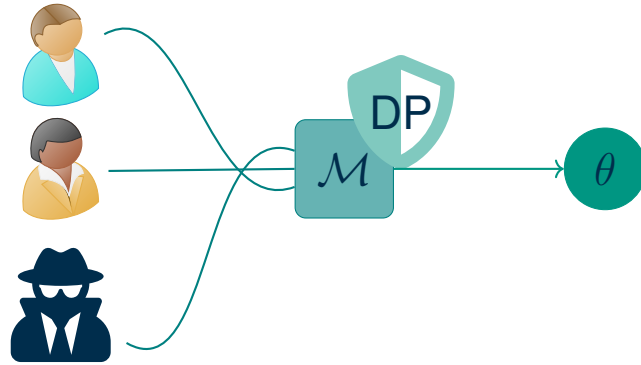
$$\beta \geq \max\{1 - e^\varepsilon \alpha, e^\varepsilon (1 - \alpha)\}$$

D_- is known:

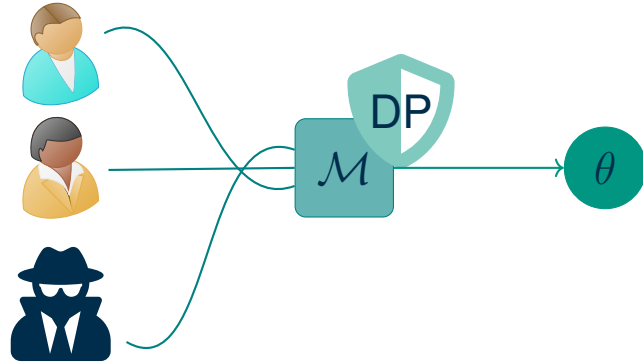
$H_0 = D_{x_n}$ Vs. $H_1 = D_{y_n}$



What about other attacker models?



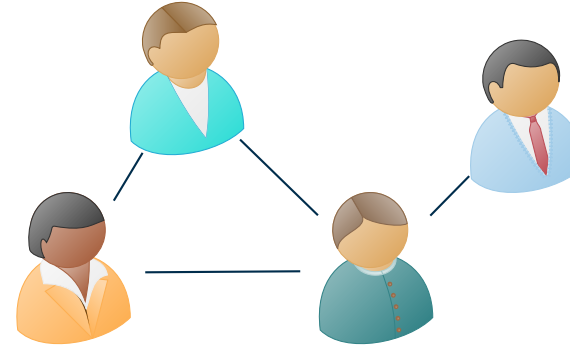
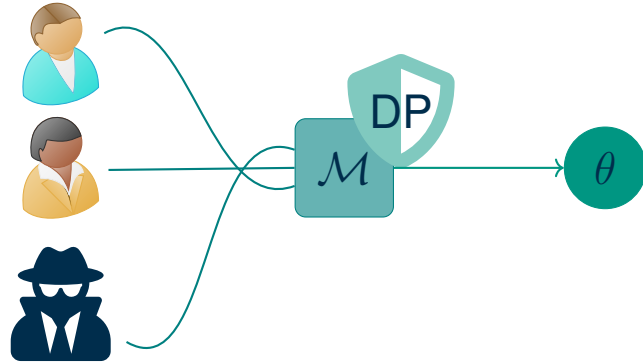
What about other attacker models?



Statistical Independence

The strongest attacker is the worst-case one, and we have at least the same protection.

What about other attacker models?



Statistical Independence

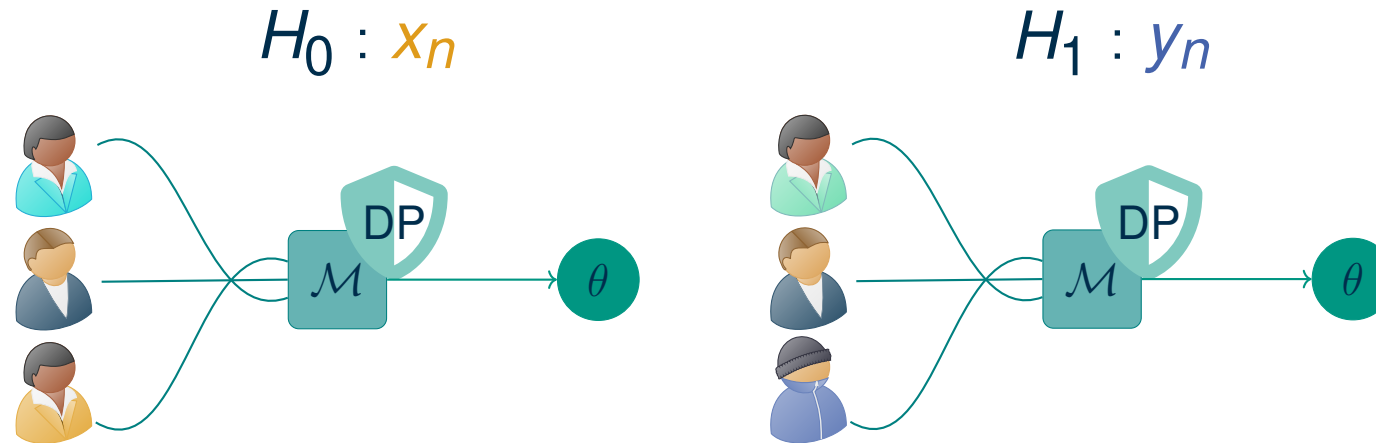
Dependencies between
Records

The strongest attacker is the worst-case one, and we have at least the same protection.

DP interpretation does not hold anymore.

Membership Inference Attack With Dependencies

The attacker receives θ and aims to distinguish between:

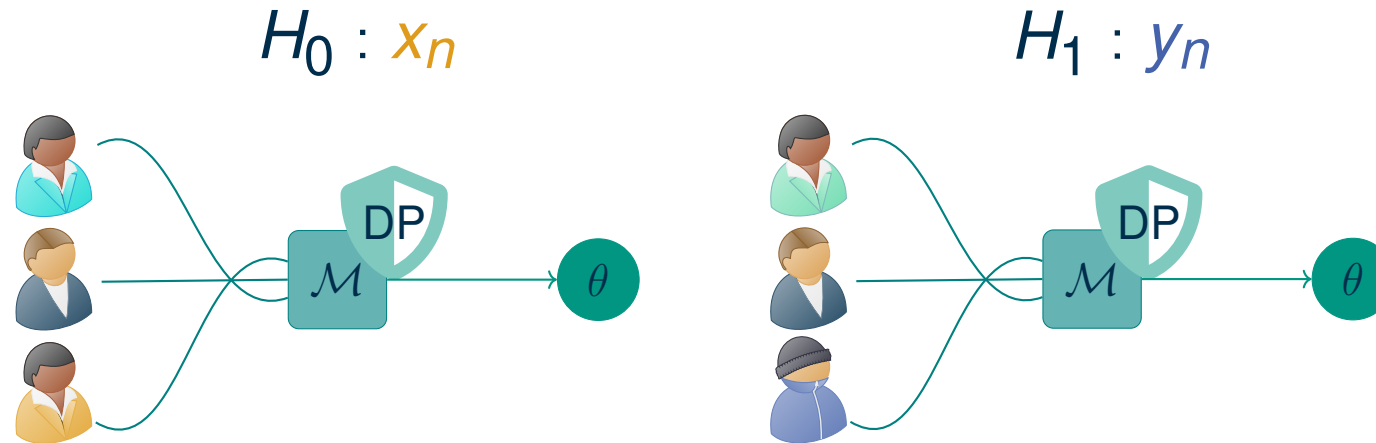


D_- is **unknown**:

$$H_0 = \{D: x_n \in D\} \text{ Vs. } H_1 = \{D: y_n \in D\}$$

Membership Inference Attack With Dependencies

The attacker receives θ and aims to distinguish between:



D_- is **unknown**:

$$H_0 = \{D: x_n \in D\} \text{ Vs. } H_1 = \{D: y_n \in D\}$$

Type I error:

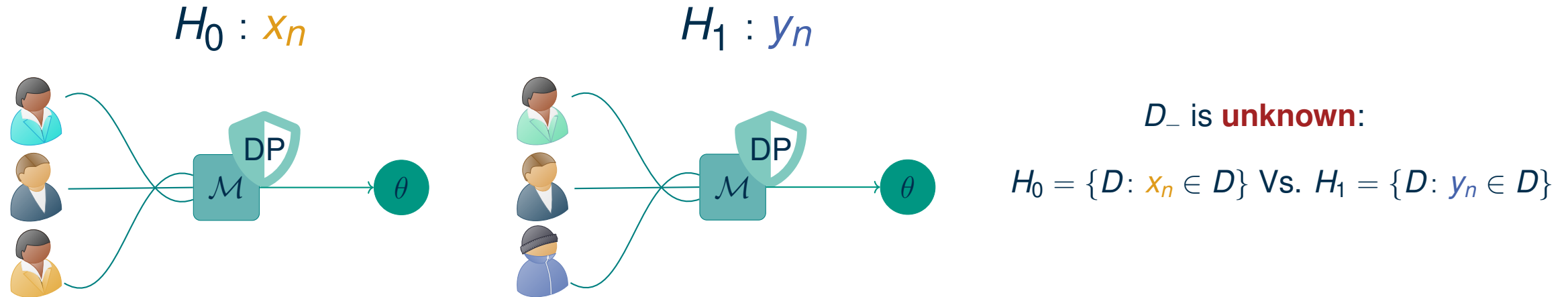
$$\begin{aligned} \alpha &= \Pr_{A \circ \mathcal{M}}(y_n | x_n) \\ &= \sum_{D_-} \Pr_{A \circ \mathcal{M}}(y_n | D_{x_n}) \pi(D_- | x_n) \end{aligned}$$

Type II error:

$$\begin{aligned} \beta &= 1 - \Pr_{A \circ \mathcal{M}}(y_n | y_n) \\ &= 1 - \sum_{D_-} \Pr_{A \circ \mathcal{M}}(y_n | D_{y_n}) \pi(D_- | y_n) \end{aligned}$$

Membership Inference Attack With Dependencies

The attacker receives θ and aims to distinguish between:



Type I error:

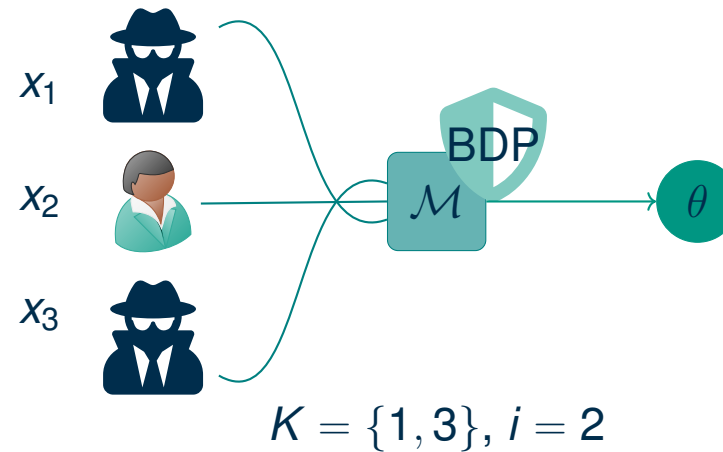
$$\begin{aligned} \alpha &= \Pr_{A \circ \mathcal{M}}(y_n \mid x_n) \\ &= \sum_{D_-} \Pr_{A \circ \mathcal{M}}(y_n \mid D_{x_n}) \pi(D_- \mid x_n) \end{aligned}$$

Type II error:

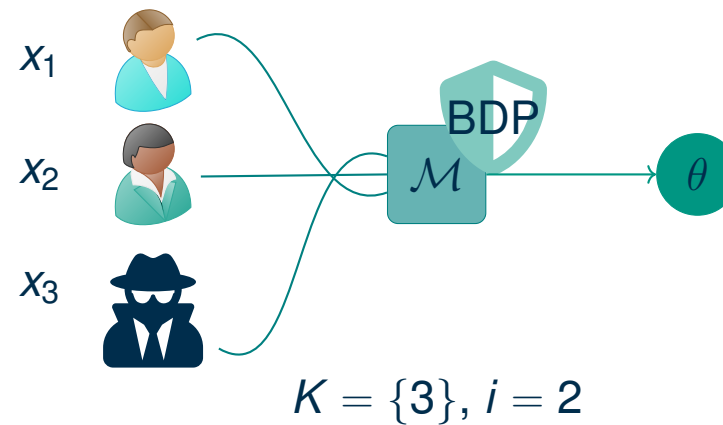
$$\begin{aligned} \beta &= 1 - \Pr_{A \circ \mathcal{M}}(y_n \mid y_n) \\ &= 1 - \sum_{D_-} \Pr_{A \circ \mathcal{M}}(y_n \mid D_{y_n}) \pi(D_- \mid y_n) \end{aligned}$$

$$A \circ \mathcal{M} \text{ is } \varepsilon\text{-DP} \Rightarrow 1 - \beta \leq \sum_{D_-} e^\varepsilon \Pr_{A \circ \mathcal{M}}(y_n \mid D_{x_n}) \pi(D_- \mid y_n) = e^\varepsilon \sum_{D_-} \Pr_{A \circ \mathcal{M}}(y_n \mid D_{x_n}) \pi(D_- \mid y_n) \neq e^\varepsilon \alpha$$

Proposed Solution: Bayesian Differential Privacy



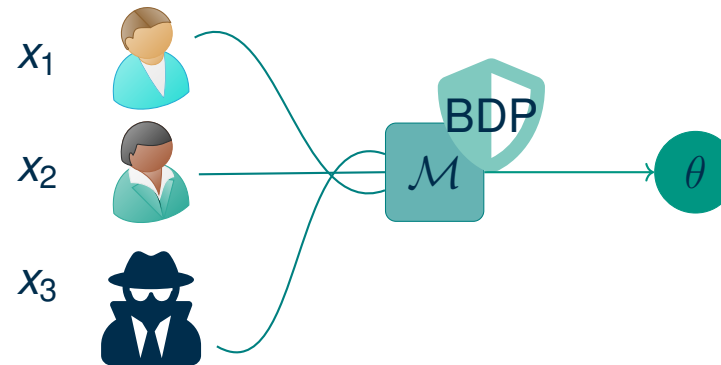
Proposed Solution: Bayesian Differential Privacy



Proposed Solution: Bayesian Differential Privacy

Bayesian DP leakage (Yang et al. 2017)

$$\text{BDPL}_{(K,i)} = \sup_{x_i, x'_i, \mathbf{x}_K, S} \ln \frac{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x'_i]}, \text{ then } \varepsilon = \sup_{K,i} \text{BDPL}_{(K,i)}.$$



Proposed Solution: Bayesian Differential Privacy

Bayesian DP leakage (Yang et al. 2017)

$$\text{BDPL}_{(K,i)} = \sup_{x_i, x'_i, \mathbf{x}_K, S} \ln \frac{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x'_i]}, \text{ then } \varepsilon = \sup_{K,i} \text{BDPL}_{(K,i)}.$$

We check our MIA again, with $H_0 : \mathbf{x}_n \in D$, $H_1 : \mathbf{y}_n \in D$ ($K = \emptyset$)

Type I error:

$$\alpha = \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid \mathbf{x}_n)$$

Type II error:

$$\beta = 1 - \Pr_{A \circ \mathcal{M}}(\mathbf{y}_n \mid \mathbf{y}_n)$$

$$A \circ \mathcal{M} \text{ is } \varepsilon\text{-BDP} \Rightarrow \begin{aligned} 1 - \beta &\leq e^\varepsilon \alpha \\ \alpha &\leq e^\varepsilon (1 - \beta) \end{aligned}$$

$$\beta \geq \max\{1 - e^\varepsilon \alpha, e^\varepsilon (1 - \alpha)\}$$

Proposed Solution: Bayesian Differential Privacy

Bayesian DP leakage (Yang et al. 2017)

$$\text{BDPL}_{(K,i)} = \sup_{x_i, x'_i, \mathbf{x}_K, S} \ln \frac{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x'_i]}, \text{ then } \varepsilon = \sup_{K,i} \text{BDPL}_{(K,i)}.$$

Privacy

- ✓ Effective measure and resistance to correlation-based attacks.

Utility

- ✗ Pessimistic impossibility results!

Against Arbitrary Correlations It Is Impossible

Kifer and Machanavajjhala 2014:

Pufferfish (including BDP)

&

$\Rightarrow$

Free-lunch Privacy

$\Rightarrow$

No utility.

arbitrary correlation

Against Arbitrary Correlations It Is Impossible

Kifer and Machanavajjhala 2014:

Pufferfish (including BDP)

&

arbitrary correlation

$\Rightarrow$

Free-lunch Privacy

$\Rightarrow$

No utility.

We express this in term of (α, β) -**accuracy** for any numerical target query f :

(α, β) -accuracy

$$\Pr(|f(D) - \mathcal{M}(D)| \geq \alpha) \leq \beta$$

$1 - \beta = \text{confidence}$

$\alpha = \text{error interval}$



Against Arbitrary Correlations It Is Impossible

Kifer and Machanavajjhala 2014:

Pufferfish (including BDP)

&

arbitrary correlation

$\Rightarrow$

Free-lunch Privacy

$\Rightarrow$

No utility.

We express this in term of (α, β) -**accuracy** for any numerical target query f :

(α, β) -accuracy

$$\Pr(|f(D) - \mathcal{M}(D)| \geq \alpha) \leq \beta$$

$1 - \beta = \text{confidence}$

$\alpha = \text{error interval}$

Our result (informal):

$$\beta < \frac{1}{e^\epsilon + 1} \Rightarrow \alpha > \frac{1}{2} \text{Range}(f).$$



Against Arbitrary Correlations It Is Impossible

Kifer and Machanavajjhala 2014:

Pufferfish (including BDP)

&

arbitrary correlation

$\Rightarrow$

Free-lunch Privacy

$\Rightarrow$

No utility.

We express this in term of (α, β) -**accuracy** for any numerical target query f :

(α, β) -accuracy

$$\Pr(|f(D) - \mathcal{M}(D)| \geq \alpha) \leq \beta$$

$1 - \beta = \text{confidence}$

$\alpha = \text{error interval}$

Our result (informal):

$$\beta < \frac{1}{e^\epsilon + 1} \Rightarrow \alpha > \frac{1}{2} \text{Range}(f).$$



Against Arbitrary Correlations It Is Impossible

Kifer and Machanavajjhala 2014:

Pufferfish (including BDP)

&

arbitrary correlation

$\Rightarrow$

Free-lunch Privacy

$\Rightarrow$

No utility.

We express this in term of (α, β) -**accuracy** for any numerical target query f :

(α, β) -accuracy

$$\Pr(|f(D) - \mathcal{M}(D)| \geq \alpha) \leq \beta$$

$1 - \beta = \text{confidence}$

$\alpha = \text{error interval}$

Our result (informal):

$$\beta < \frac{1}{e^\epsilon + 1} \Rightarrow \alpha > \frac{1}{2} \text{Range}(f).$$



Out of 100
how many
are infected?



Is it 0 or
100? 🤔



Few Correlated Records, Same Disaster

Our result (informal)

Privacy decreases linearly proportional to number of correlated records:

$$\varepsilon\text{-DP} \Rightarrow m\varepsilon\text{-BDP}$$

How does it impact utility?

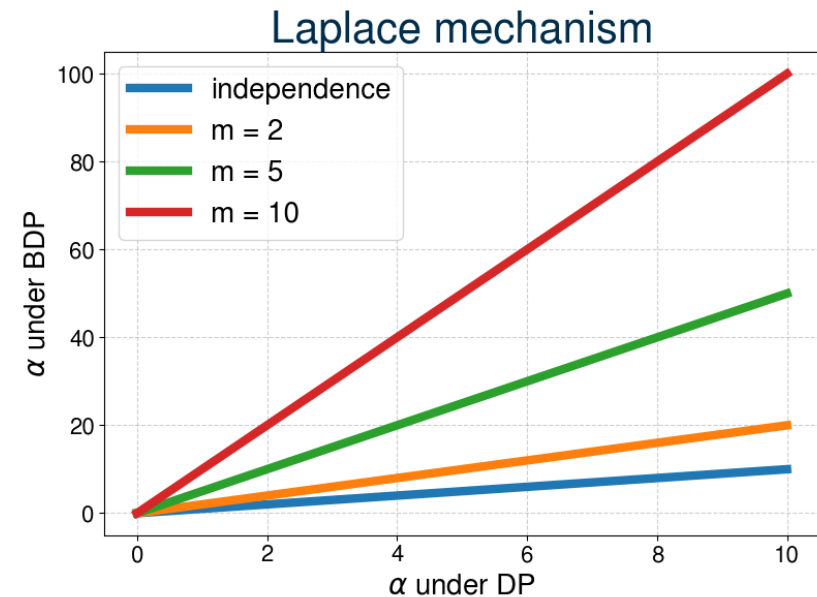


Figure: For the same confidence level, the upper bound on the query error α increases sharply.

Few Correlated Records, Same Disaster

Our result (informal)

Privacy decreases linearly proportional to number of correlated records:

$$\varepsilon\text{-DP} \Rightarrow m\varepsilon\text{-BDP}$$

This result is tight! Even excluding edge-cases.

How does it impact utility?

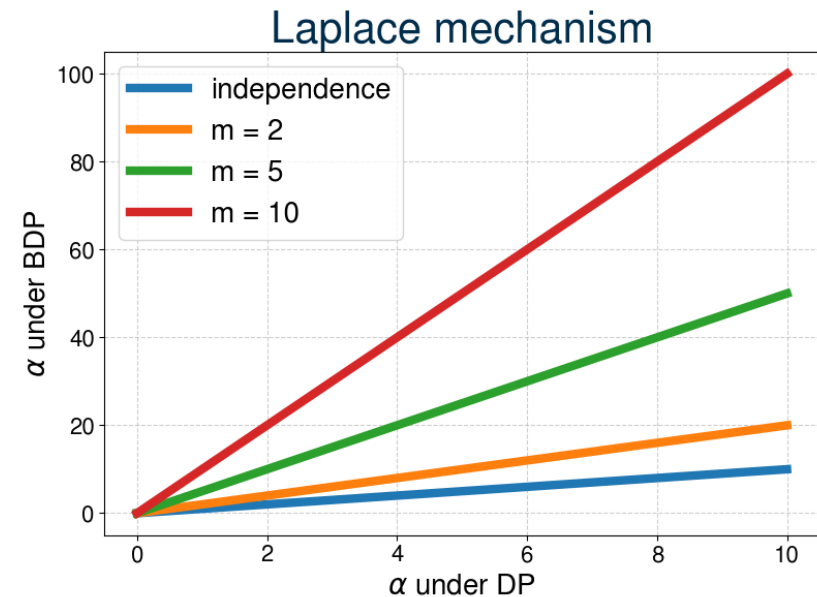


Figure: For the same confidence level, the upper bound on the query error α increases sharply.

Few Correlated Records, Same Disaster

Our result (informal)

Privacy decreases linearly proportional to number of correlated records:

$$\varepsilon\text{-DP} \Rightarrow m\varepsilon\text{-BDP}$$

This result is tight! Even excluding edge-cases.

Conclusion: We need to target specific correlation models π to obtain utility

How does it impact utility?

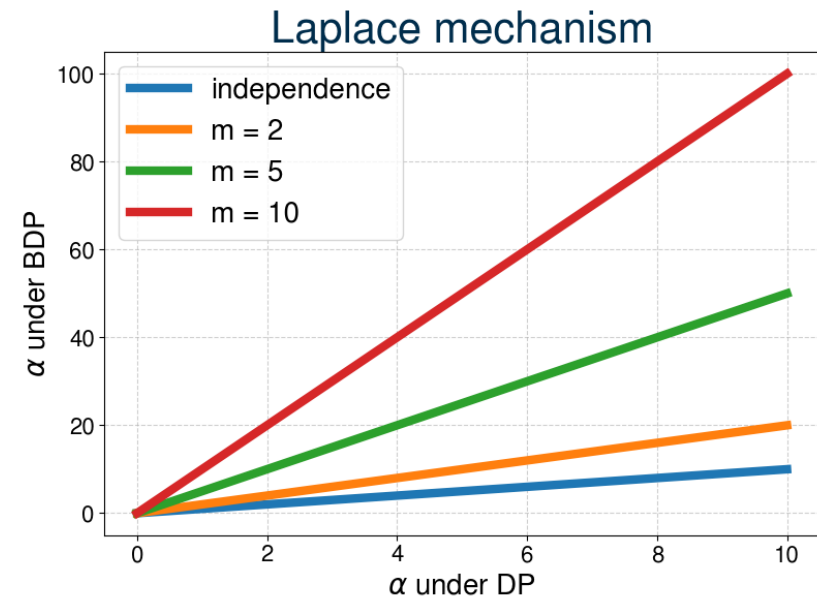


Figure: For the same confidence level, the upper bound on the query error α increases sharply.

Balancing Privacy and Utility in Correlated Data

VLDB25

Research Question:

Is it possible to perform accurate data analysis while protecting against correlation-based attacks?

Balancing Privacy and Utility in Correlated Data

VLDB25

Research Question:

Is it possible to perform accurate data analysis while protecting against correlation-based attacks?

Assumption:

 does not have more knowledge about π than the data curator.

Balancing Privacy and Utility in Correlated Data

VLDB25

Research Question:

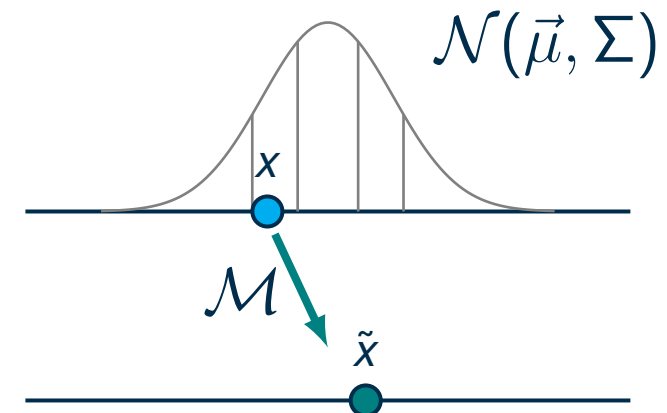
Is it possible to perform accurate data analysis while protecting against correlation-based attacks?

Assumption:

 does not have more knowledge about π than the data curator.

Our Strategy:

Given a specific prior π
 ϵ -DP $\implies$???-BDP



Connecting DP and Bayesian DP

Multivariate Gaussian Theorem

Assumptions: $\rho(m-2) \leq 1$ & ℓ_1 -privacy:

Markov Chain Theorem

Assumptions: $P_{xy} > 0$ & Stationary:

Connecting DP and Bayesian DP

Multivariate Gaussian Theorem

Assumptions: $\rho(m-2) \leq 1$ & ℓ_1 -privacy:

$$\varepsilon\text{-DP} \implies \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) \varepsilon\text{-BDP}$$

Markov Chain Theorem

Assumptions: $P_{xy} > 0$ & Stationary:

$$\varepsilon\text{-DP} \implies (\varepsilon + 4 \ln \gamma)\text{-BDP}$$

$$\text{where } \gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}.$$

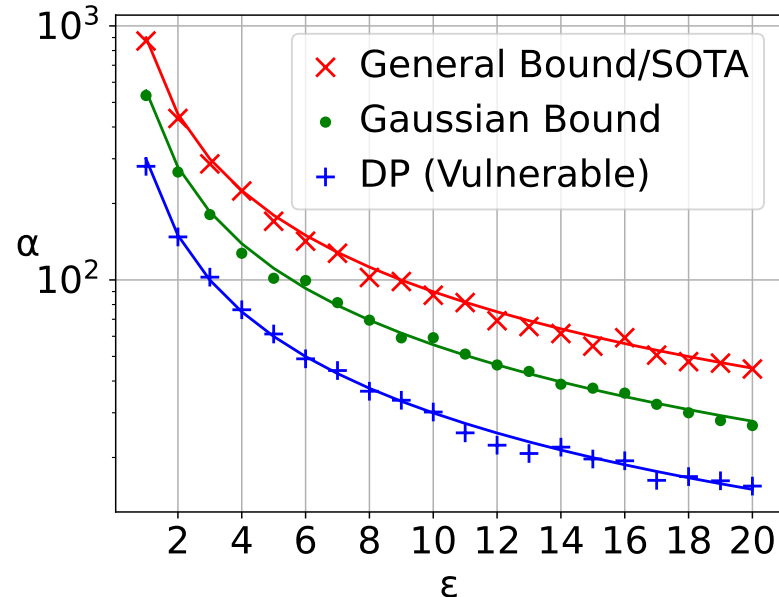
Connecting DP and Bayesian DP

Multivariate Gaussian Theorem

Assumptions: $\rho(m-2) \leq 1$ & ℓ_1 -privacy:

$$\varepsilon\text{-DP} \implies \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) \varepsilon\text{-BDP}$$

Galton Data



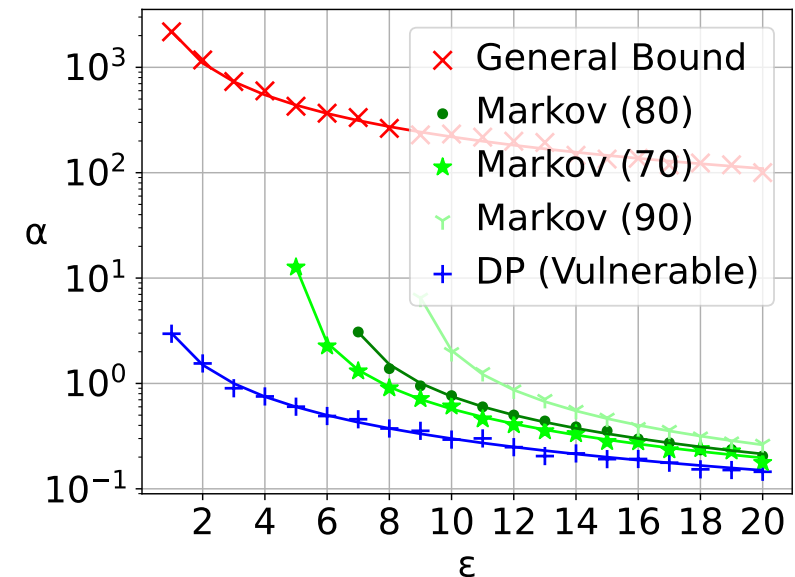
Markov Chain Theorem

Assumptions: $P_{xy} > 0$ & Stationary:

$$\varepsilon\text{-DP} \implies (\varepsilon + 4 \ln \gamma)\text{-BDP}$$

$$\text{where } \gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}.$$

Electricity



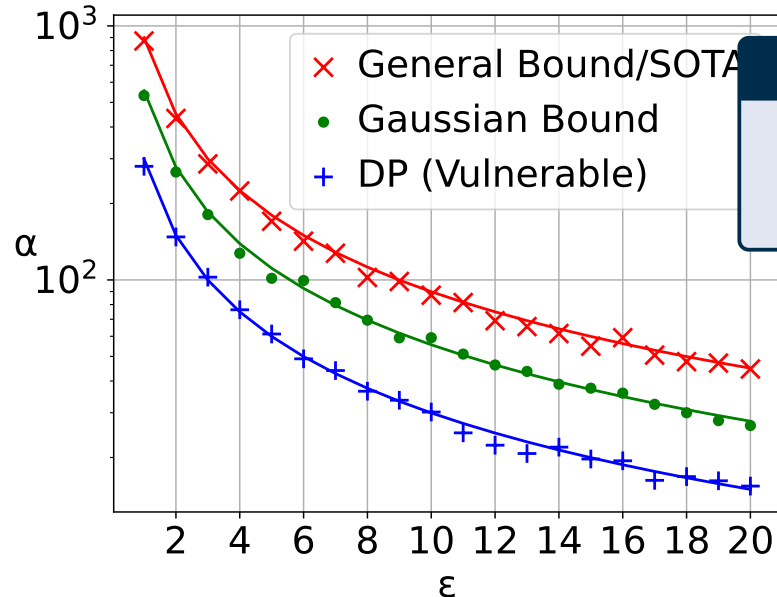
Connecting DP and Bayesian DP

Multivariate Gaussian Theorem

Assumptions: $\rho(m-2) \leq 1$ & ℓ_1 -privacy:

$$\varepsilon\text{-DP} \implies \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) \varepsilon\text{-BDP}$$

Galton Data



Key Takeaway:

We significantly improve the **accuracy** under BDP protection!

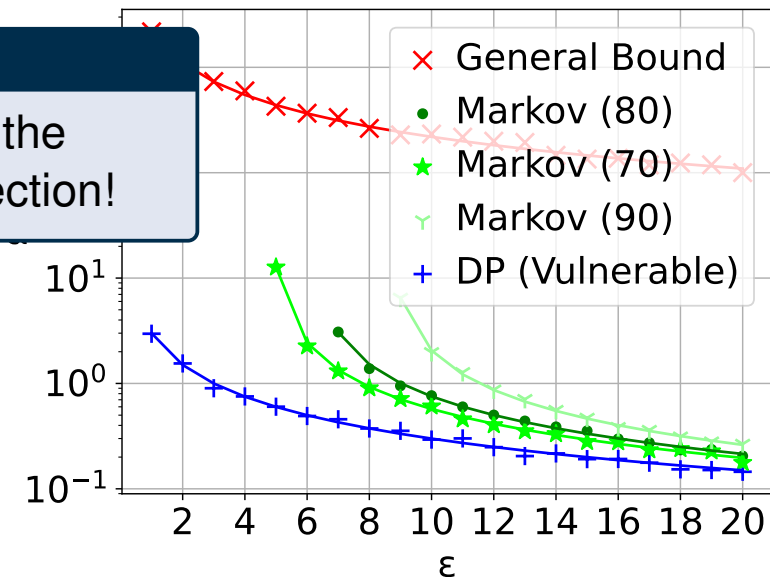
Markov Chain Theorem

Assumptions: $P_{xy} > 0$ & Stationary:

$$\varepsilon\text{-DP} \implies (\varepsilon + 4 \ln \gamma)\text{-BDP}$$

$$\text{where } \gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}.$$

Electricity



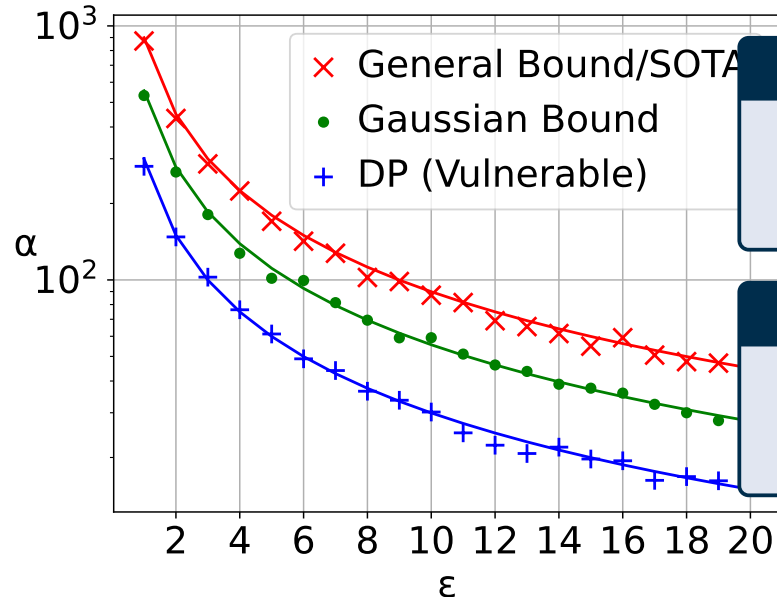
Connecting DP and Bayesian DP

Multivariate Gaussian Theorem

Assumptions: $\rho(m-2) \leq 1$ & ℓ_1 -privacy:

$$\varepsilon\text{-DP} \implies \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) \varepsilon\text{-BDP}$$

Galton Data



Key Takeaway:

We significantly improve the **accuracy** under BDP protection!

Key Limitation:

Assumptions limit applicability & privacy choices!

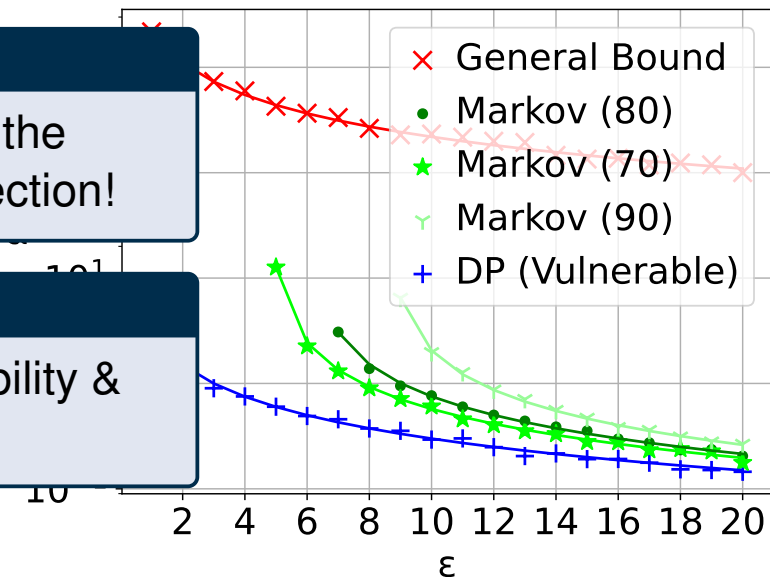
Markov Chain Theorem

Assumptions: $P_{xy} > 0$ & Stationary:

$$\varepsilon\text{-DP} \implies (\varepsilon + 4 \ln \gamma)\text{-BDP}$$

$$\text{where } \gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}.$$

Electricity



Conclusion

Our contributions:

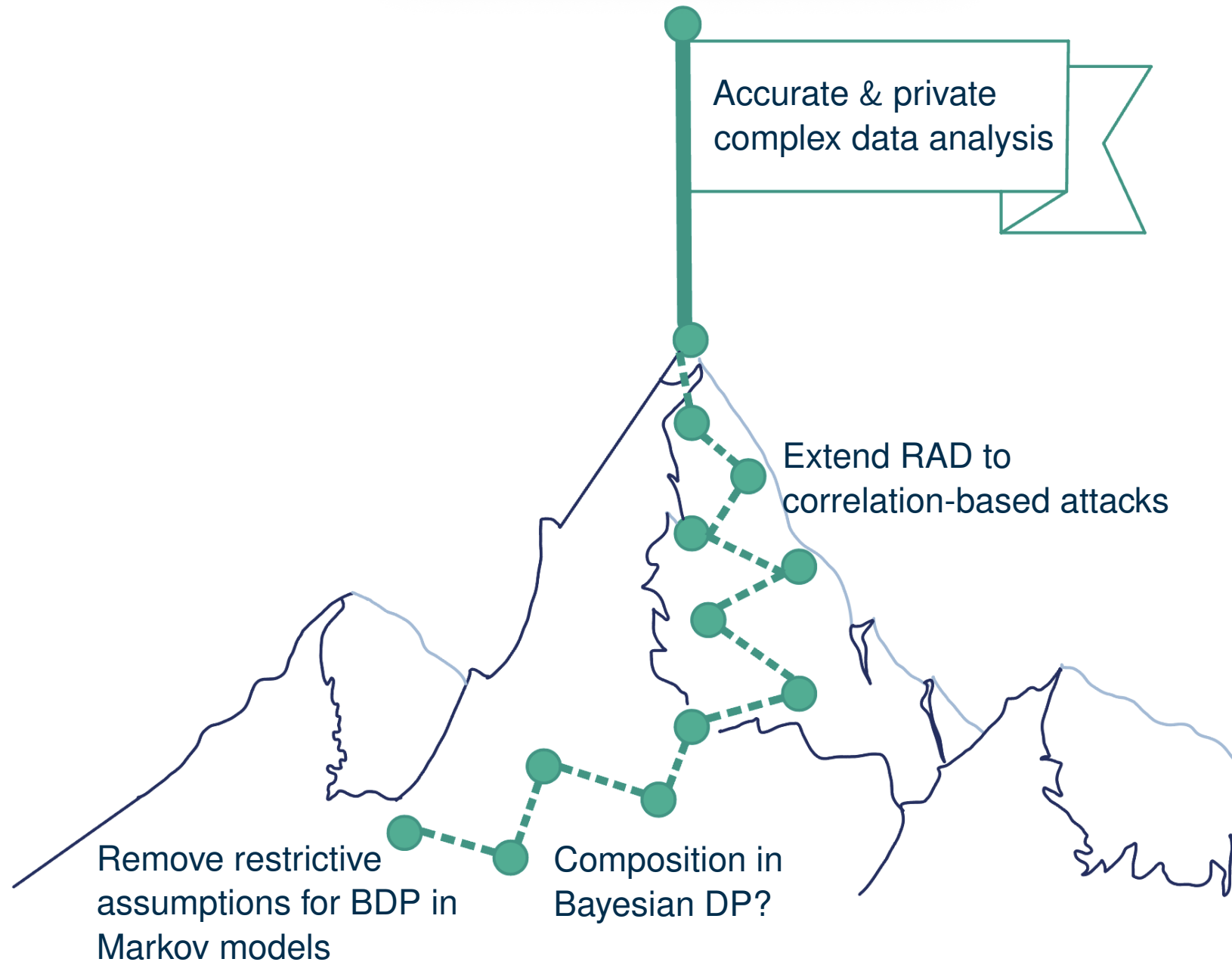
Improved attack bounds and optimal attacks

Novel results and algorithms improving utility in correlated data

Takeaways:

Attack resilience offers a more complete view of the real impact and protection of our mechanisms than ε .

It is possible to give accuracy and protection against correlation-based adversaries! (If we know the correlation model)



Conclusion

Our contributions:

Improved attack bounds and optimal attacks

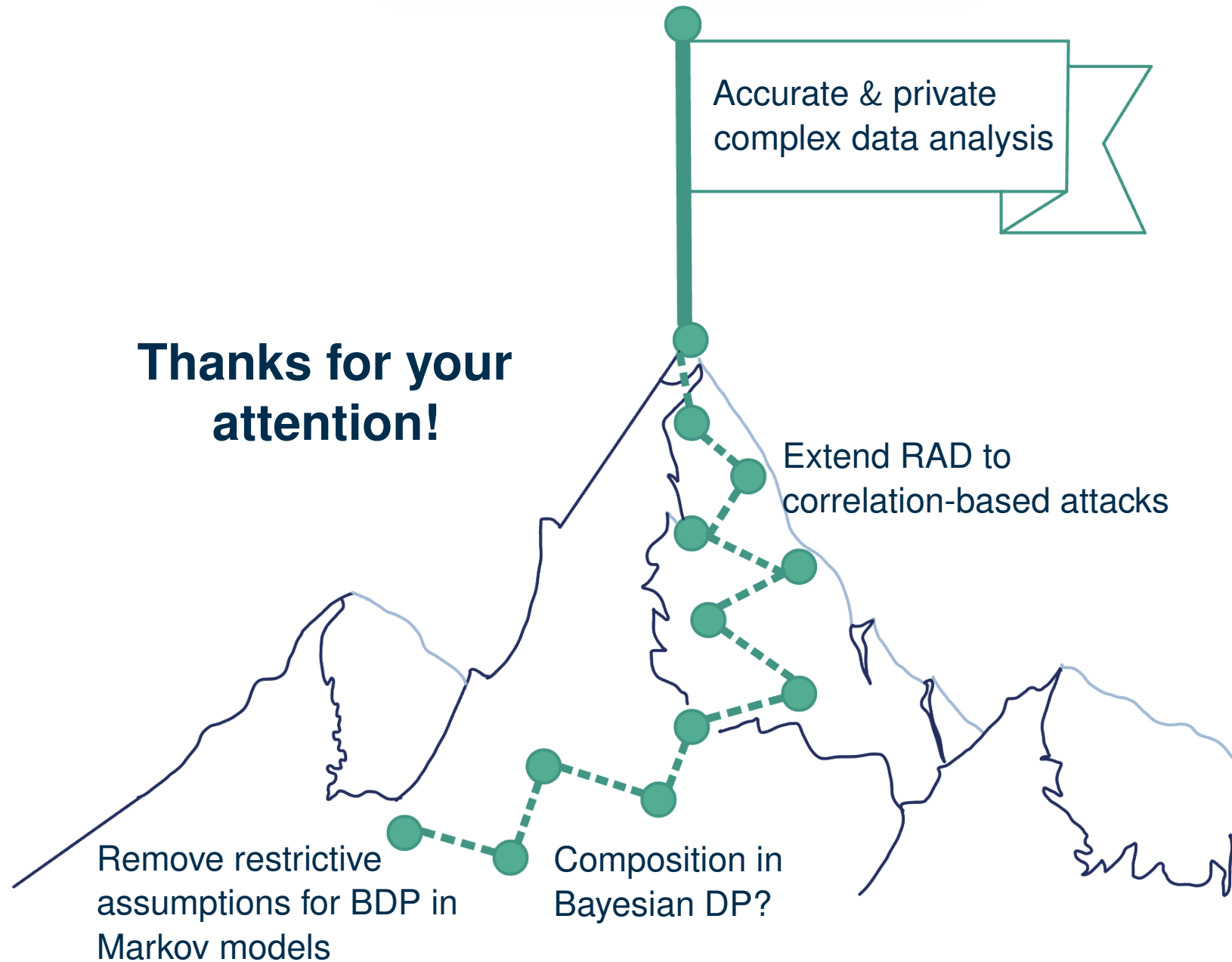
Novel results and algorithms improving utility in correlated data

Takeaways:

Attack resilience offers a more complete view of the real impact and protection of our mechanisms than ε .

It is possible to give accuracy and protection against correlation-based adversaries! (If we know the correlation model)

Thanks for your attention!



Additional Slides

Understanding Attack Resilience in DP

Theorem 5.2

Theorem $((\varepsilon, \delta)$ -DP implies η -RAD)

Let $\pi, \ell, \eta \geq 0$ as in Def. 5.1., and $\kappa_\pi = \Pr_{X, X' \sim \pi}[X = X']$. If a mechanism $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{D}(\Theta)$ satisfies (ε, δ) -DP, then for any attack $A: \Theta \times \text{aux} \rightarrow \mathcal{D}(\mathcal{X})$, and database D_- we have

$$\eta\text{-RAD} \leq \text{TV}(\mathcal{M})(1 - \kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1 - \kappa_\pi).$$

Theorem 5.3

Theorem

Given $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{D}(\Theta)$ and $a: \mathcal{X} \rightarrow \text{aux}$ measurable, then for any attack $A: \Theta \times \text{aux} \rightarrow \mathcal{D}(\mathcal{X})$, we have

$$\eta\text{-RAD} \leq \int_{\Theta} \int_{\text{aux}} \max_{x_{\theta} \in \mathcal{X}} \left(\int_{S_{\eta}^z(x_{\theta})} w(\theta, x) \pi_x \, d\mu_z(x) \right) d\nu(z) d\mu(\theta),$$

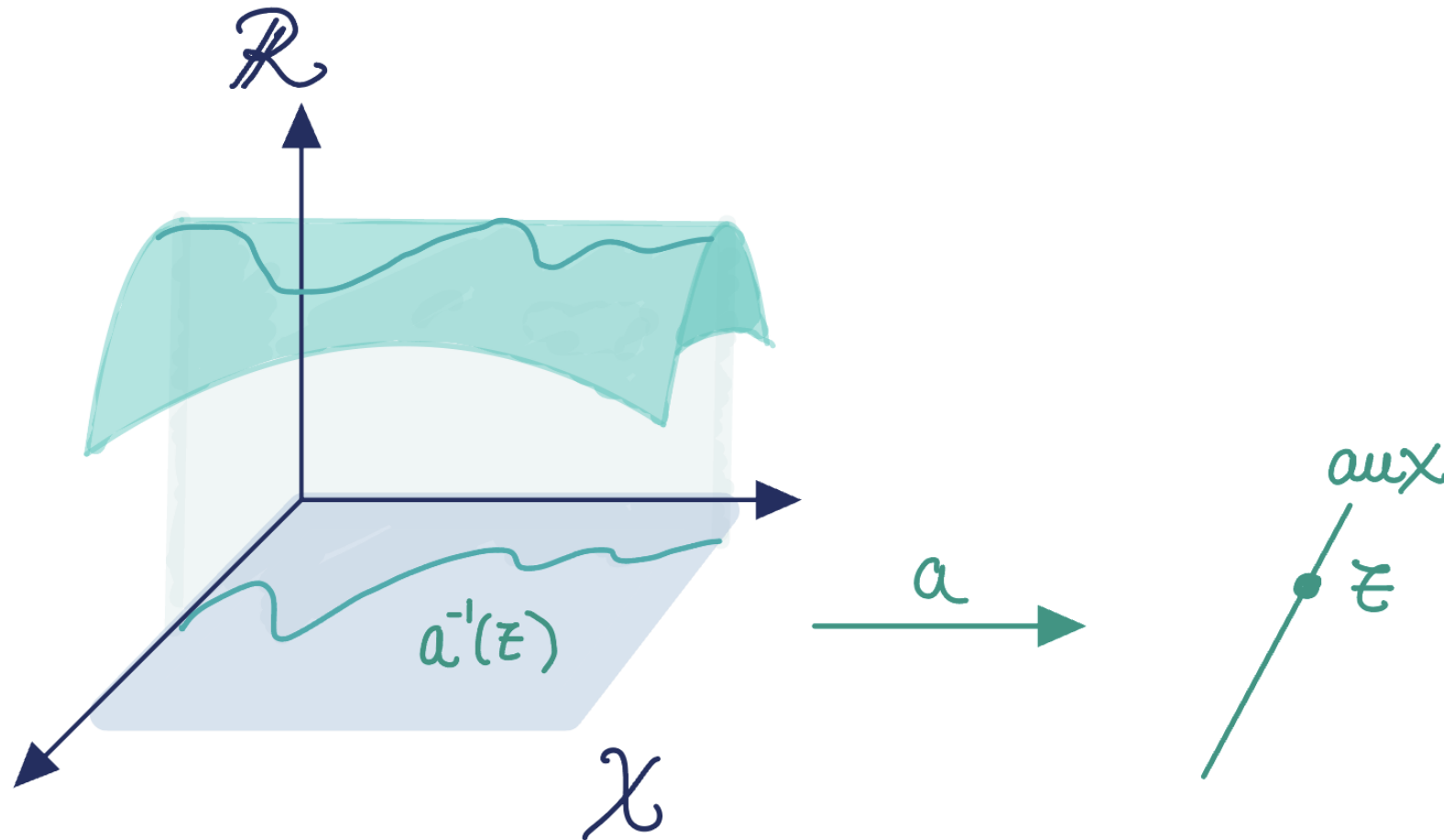
where μ the counting (or Lebesgue measure) and π_x mass (or density function) in the discrete (or continuous case). Additionally, $w(x, \theta) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$, $S_{\eta}^z(x_{\theta}) = \{x: a(x) = z \wedge \ell(x_{\theta}, x) \leq \eta\}$, $\nu(z) = \mu \circ a^{-1}(z)$ and μ_z the disintegration theorem measure.

The discrete case simplifies to

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{x_{\theta} \in \mathcal{X}} \sum_{\substack{\ell(x, x_{\theta}) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x$$

by direct application of Remark 2.29.

Disintegration Theorem



Optimal Bound for Particular Cases

Membership Inference Attack (MIA):

- $a(x) = x$
- $\ell(x, x') = 1$ if $x = x'$ and 0 otherwise.

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{\substack{x \in \mathcal{X} \\ w(\theta, x) > 0}} w(\theta, x) \pi_x$$

Generalized Randomized Response

$$\eta\text{-RAD} \leq \frac{e^\varepsilon - 1}{e^\varepsilon + m - 1} (1 - \kappa_\pi)$$

Optimal Unary Encoding

$$\eta\text{-RAD} \leq \frac{1}{2} \frac{e^\varepsilon - 1}{e^\varepsilon + 1} (1 - \kappa_\pi)$$

Perfect reconstruction:

- $a(x) = \emptyset$ for every $x \in \mathcal{X}$
- $\eta = 0$

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in \text{aux}} \max_{\substack{x \in \mathcal{X} \\ w(\theta, x) > 0}} w(\theta, x) \pi_x$$

Laplace mechanism, uniform prior

$$\eta\text{-RAD} \leq \left(1 - e^{-\frac{\varepsilon}{2(m-1)}}\right) \frac{m-1}{m}$$

Gaussian mechanism, uniform prior

$$\eta\text{-RAD} \leq \left(2\Phi\left(\frac{1}{2\sigma(m-1)}\right)\right) \frac{m-1}{m}$$

Upper-bounds for $aux = \{\emptyset\}$

Theorem

If a mechanism $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{D}(\Theta)$ satisfies f -DP, then for any attack with $aux = \{\emptyset\}$, $A: \Theta \rightarrow \mathcal{D}(\mathcal{X})$, it satisfies

$$\eta\text{-RAD} \leq \max_{\alpha \in [\kappa_{\pi, \ell}^-(\eta), \kappa_{\pi, \ell}^+(\eta)]} 1 - f(\alpha) - \alpha.$$

If $\mathcal{X}$ is discrete, then it also holds

$$\eta\text{-RAD} \leq (1 - \kappa_{\pi}) \max_{\alpha \in [0, \frac{\kappa_{\pi, \ell}^+(\eta)}{1 - \kappa_{\pi}}]} 1 - f(\alpha) - \alpha.$$

Theorem (0-RAD under (ε, δ) -DP)

Given $|\mathcal{X}| = m$ with prior $\pi_1(1 - \pi_1) \geq \dots \geq \pi_m(1 - \pi_m)$ and $\mathcal{M}: \mathcal{X}^n \rightarrow \mathcal{D}(\Theta)$ an (ε, δ) -DP mechanism, for any attack $A: \Theta \rightarrow \mathcal{D}(\mathcal{X})$

$$0\text{-RAD} \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} K_{\pi} + R \max_{i > K} \pi_i$$

where $K \in [m]$ is the largest index satisfying $R = (m - 1) \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1} - (K - \sum_{i=1}^K \pi_i) \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} \geq 0$ and $K_{\pi} = \sum_{i=1}^K (1 - \pi_i) \pi_i$.

Comparison of black-box bounds when $aux = \{\emptyset\}$

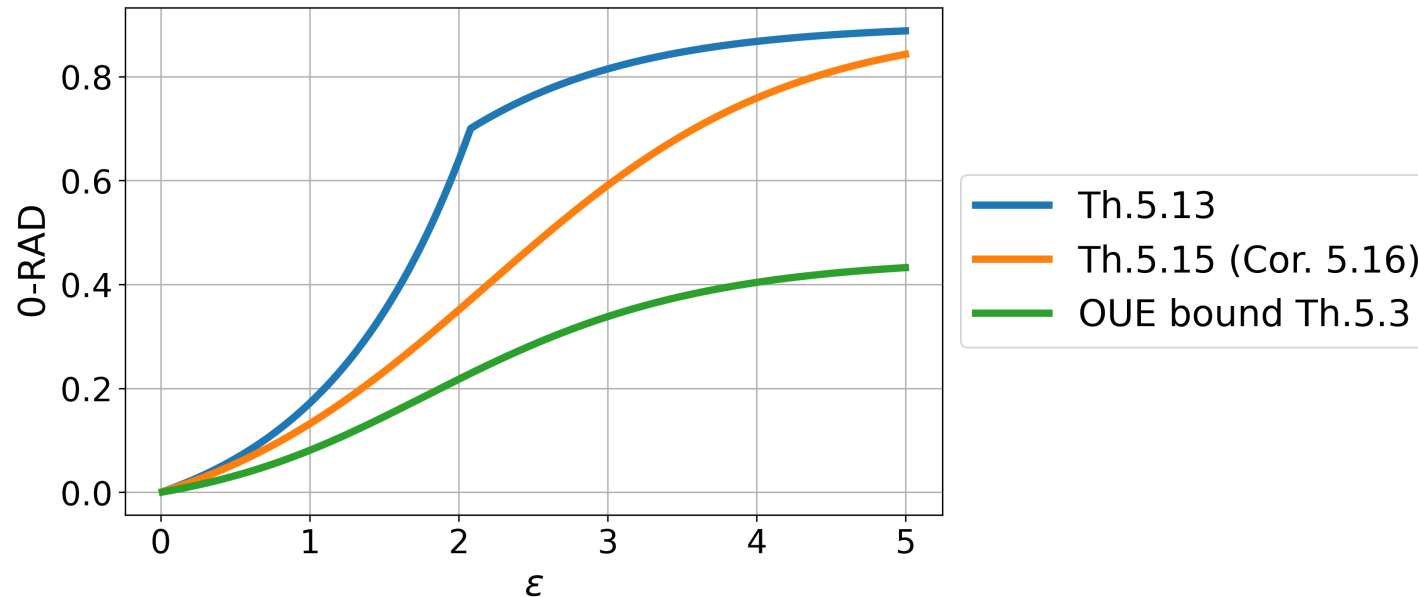


Figure: $\pi = U[10]$ and $\delta = 10^{-5}$.

- The bound given in Th.5.13 is more general and applies in any setting.
- Th.5.15 is specific to categorical data but provides a tighter risk estimate when applicable.
- If the mechanism is known—here, OUE—it is always preferable to use universally tight bound (Th.5.3).

Optimal Attack against DP-SGD

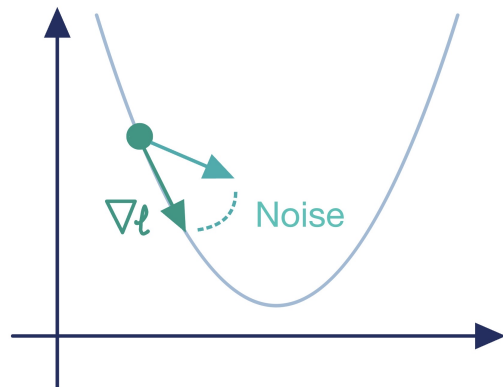
- $\mathcal{X} = \{x_1, \dots, x_m\}$, π uniform.
- Intermediate weights $\theta = (\theta_1, \dots, \theta_T)$ and
- privatized gradients $\{\bar{g}_1, \dots, \bar{g}_T\}$ released during training:

$$\bar{g}_t = \sum_x \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x)) + \mathcal{N}(0, c^2 \sigma^2 I),$$

where $\text{clip}_C(\vec{v}) = \vec{v} \min(1, \frac{C}{\|\vec{v}\|_2})$.

- Noisy contribution of the target's gradient t :

$$g_t = \bar{g}_t - \sum_{x \in D_-} \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x)).$$



Algorithm 2: Optimal Attack for DP-SGD

Input : $\theta = (\theta_1, \dots, \theta_T)$, $a(x) = z$ and $g = (g_1, \dots, g_T)$

Output : $\tilde{x}$

for $x: a(x) = z$ **do**

 | compute $\sum_t W(\bar{g}_t, \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x)))$;

Select $x^* = \arg \max_{x: a(x)=z} \sum_t W(\bar{g}_t, \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x))) \pi_x$;

if $W(\bar{g}_t, x^*) > 0$ **then**

 | $\tilde{x} = x^*$;

else

 | $\tilde{x} \leftarrow U[\mathcal{X} \setminus \{x: a(x) = z\}]$;

We can simplify w maximization to

$$\arg \max_{x: a(x)=z} w(\theta, x) = \arg \max_{x: a(x)=z} \sum_t W(g_t, \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x)))$$

where $W(u, v) = \langle u, v \rangle - \frac{1}{m} \sum_{x_i \in \mathcal{X}} \langle u, \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x_i)) \rangle$.

Optimal Attack Against DP-SGD

Why can we simplify w to W ?

Given θ, x , under DP-SGD the privatized gradient at step t is

$$g_t \sim \mathcal{N}(\mu_x, C^2 \sigma^2 I), \quad \mu_x = \text{clip}_C(\nabla_{\theta} \ell(\theta_t, x)),$$

where C is the clipping parameter and I the identity function of dimension d . Hence

$$p_{\mathcal{M}}(g_t | x) = \underbrace{\frac{1}{(2\pi C^2 \sigma^2)^{d/2}}}_A \exp\left(\underbrace{-\frac{1}{2C^2 \sigma^2}}_B \|g_t - \mu_x\|^2\right),$$

where both A, B are independent from x . Consequently,

$$w(g, x) = \prod_t p_{\mathcal{M}}(g_t | x) - \prod_t p_{\mathcal{M}}(g_t) > 0 \Leftrightarrow A^T \left(\prod_t e^{B\langle g_t, \mu_x \rangle} - \prod_t \frac{1}{m} \sum_i e^{B\langle g_t, \mu_{x_i} \rangle} \right) > 0 \Leftrightarrow \quad (1)$$

$$e^{B\sum_t \langle g_t, \mu_x \rangle} > \prod_t \frac{1}{m} \sum_i e^{B\langle g_t, \mu_{x_i} \rangle} \Leftrightarrow B \sum_t \langle g_t, \mu_x \rangle > \sum_t \ln \left(\frac{1}{m} \sum_i e^{B\langle g_t, \mu_{x_i} \rangle} \right) \Leftrightarrow \quad (2)$$

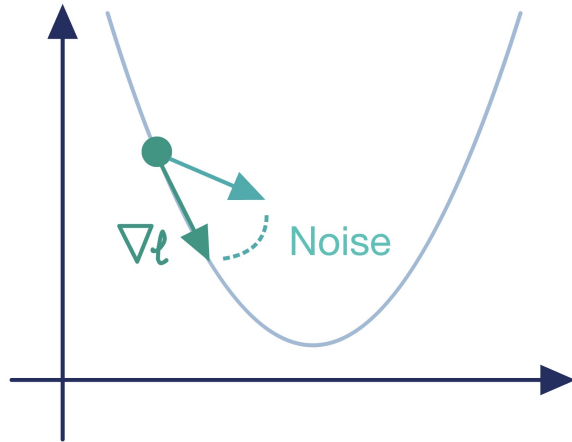
$$B \sum_t \langle g_t, \mu_x \rangle > \sum_t \frac{1}{m} \sum_z \ln(e^{B\langle g_t, \mu_{x_i} \rangle}) \Leftrightarrow B \sum_t \langle g_t, \mu_x \rangle > \sum_t \frac{B}{m} \sum_i \langle g_t, \mu_{x_i} \rangle \Leftrightarrow \quad (3)$$

$$\sum_t \langle g_t, \mu_x \rangle - \sum_t \frac{1}{m} \sum_{x_i} \langle g_t, \mu_{x_i} \rangle > 0 \Leftrightarrow \sum_t W(g_t, x) > 0. \quad (4)$$

Experimental Details: Mechanisms

ML Image Classification

DP-(S)GD



Architecture: Standard multilayer perceptron (MLP)

Hyperparameters:

- Batch sampling rate $q = 1$, Learning rate 9.
- Epochs $T = 100$, Clipping rate $C = 0.1$.

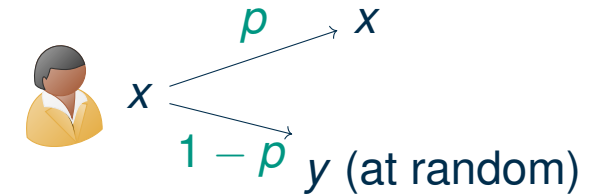
Aggregate Statistics

Laplace Mechanism

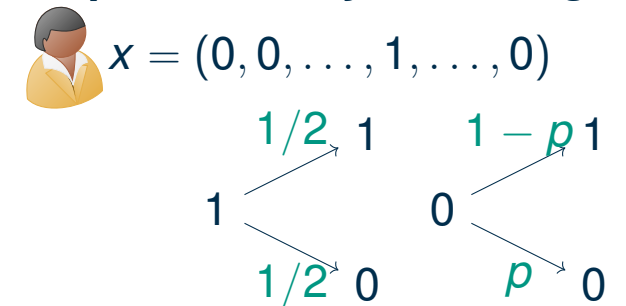


LDP for location data

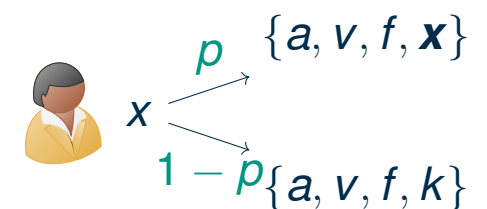
Generalized Randomized Response



Optimal Unary Encoding



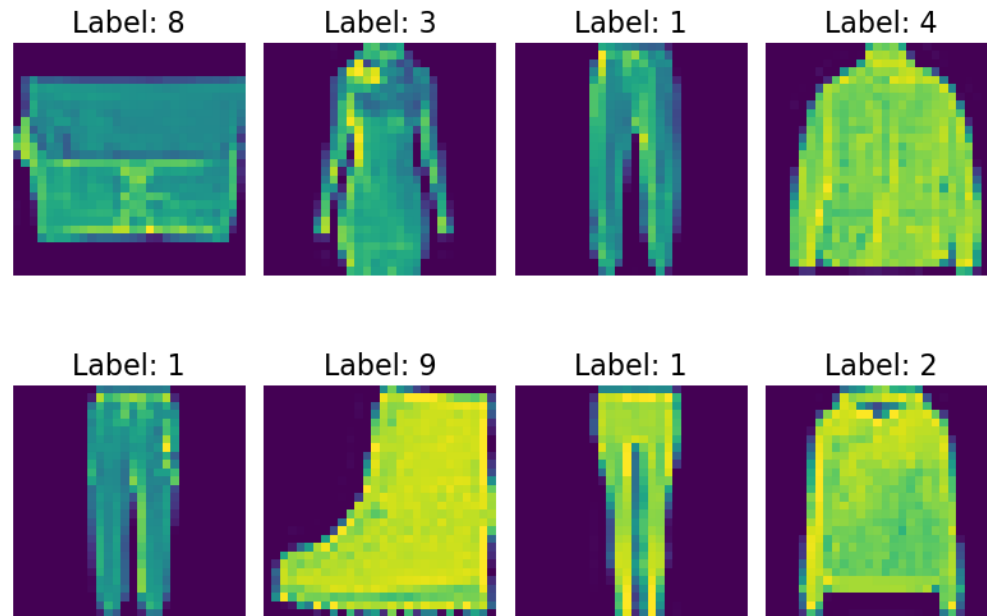
Subset Selection



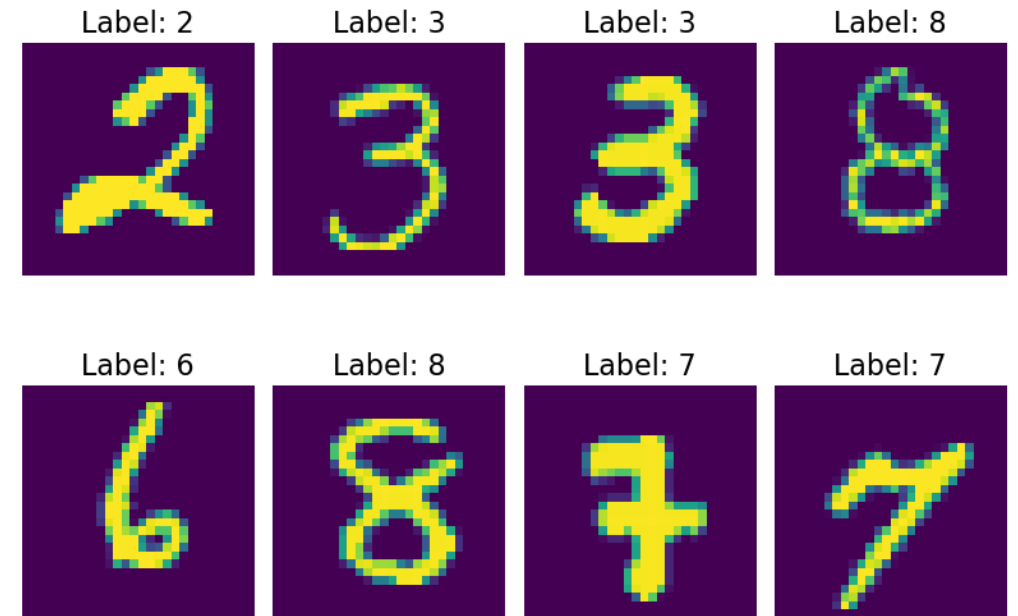
Experimental Details: Datasets

DP-SGD

Fashion-MNIST



MNIST

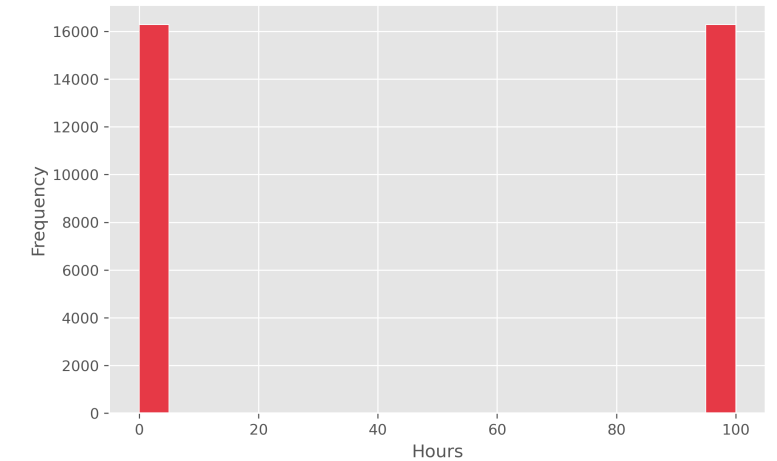
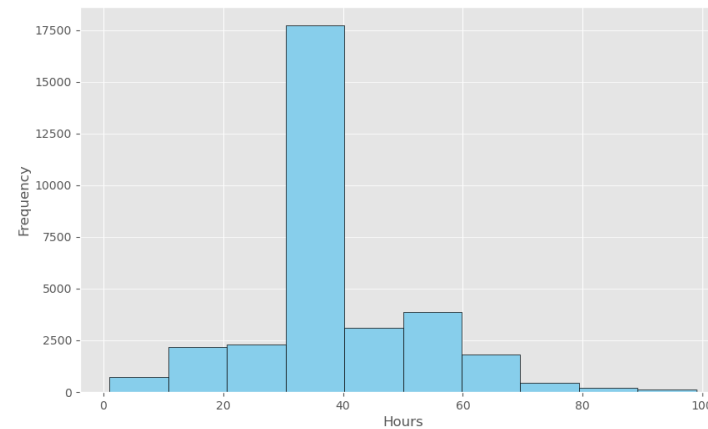
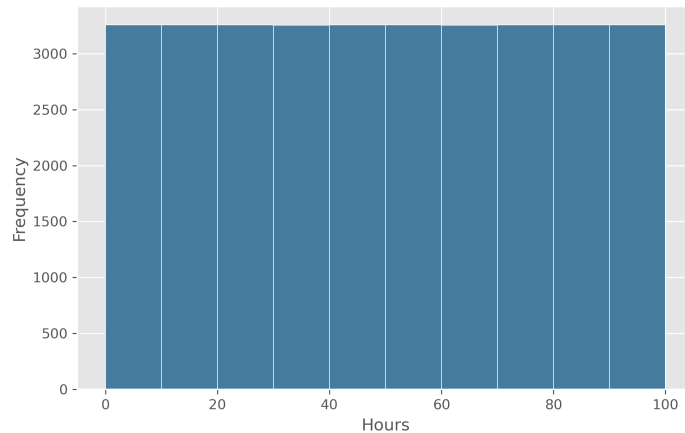


- We set the uniform prior with size $|\mathcal{X}| = 8$ (disjoint from D_-), meaning that $\kappa_{\pi,0}^+ = \kappa_{\pi} = 0.125$.
- We train on $D = 1\,000$ images.

Experimental Details: Datasets

Laplace Mechanism

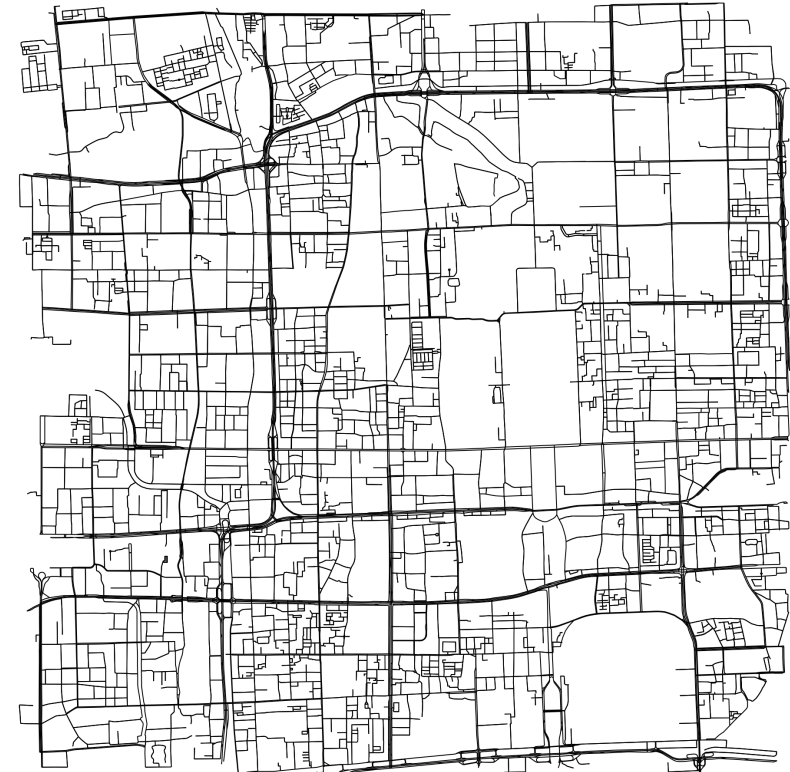
- Adult dataset (Census), $|D| = 32\,561$
- $\mathcal{X} = [0, \dots, 100]$ working hours per week.



Experimental Details: Datasets

LDP for Locations

- **Porto:** 83,409,386 location reports that we map to the OSM roadgraph at Porto's city center (41.1475° N, 8.5870° W) with a 2.7 km radius, capturing the urban core. $|\mathcal{X}| = 3\,052$.
- **Geolife:** 24 876 978 locations that we mapped to an OSM graph centered near Tiananmen Square (39.9130° N, 116.3703° E) with a 5 km radius covering major central districts, $|\mathcal{X}| = 5\,356$.



Experimental Details: Datasets

Imputation Attack in Census Data

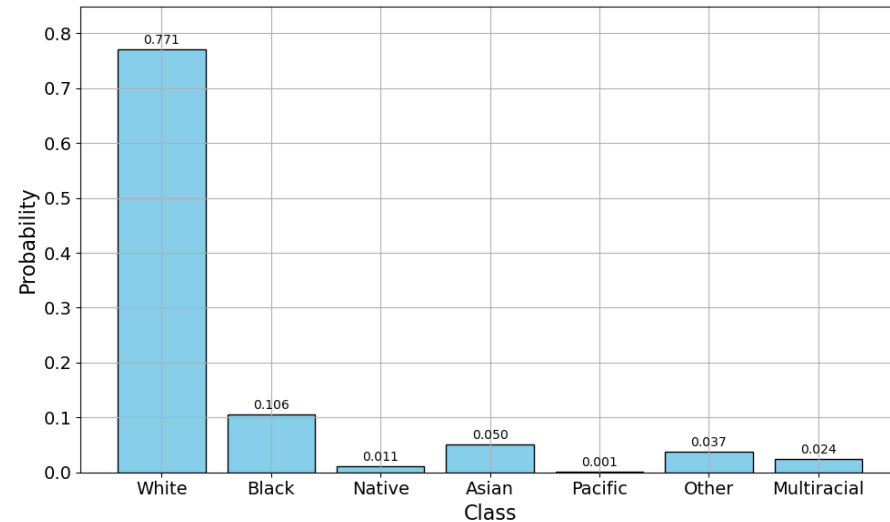


Figure: Ethnicity (Census).

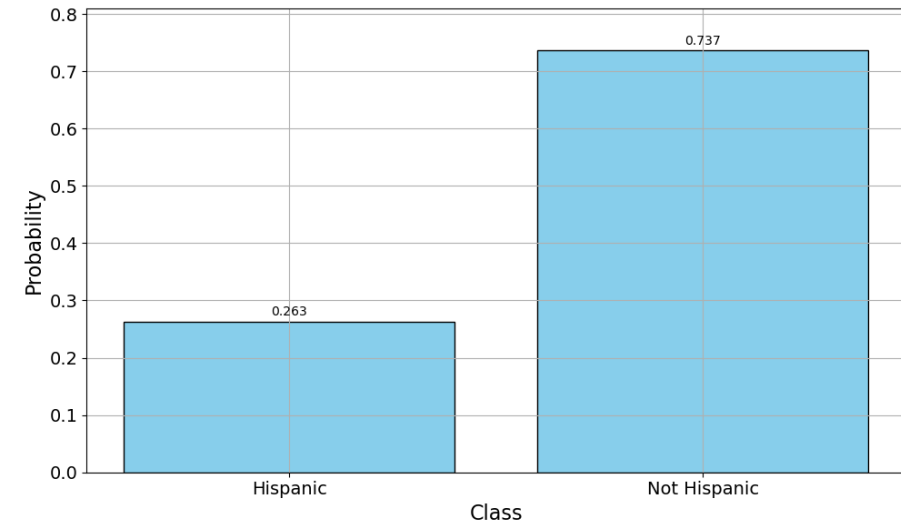
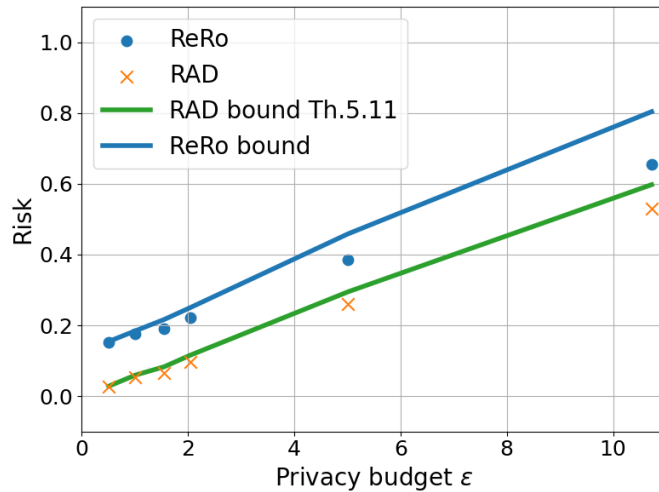
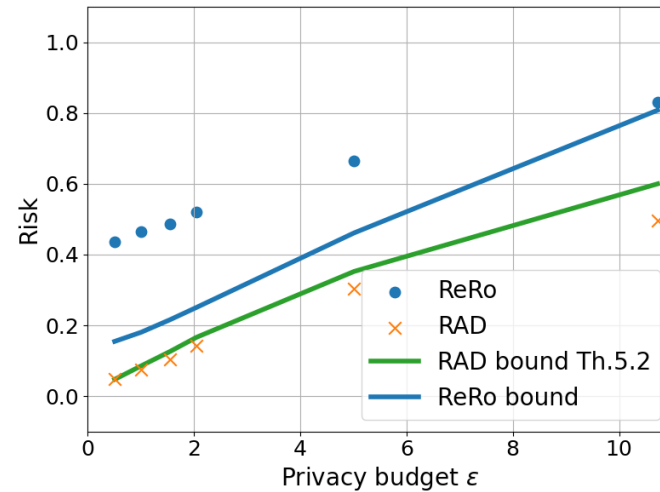


Figure: Ethnicity (Texas-100X).

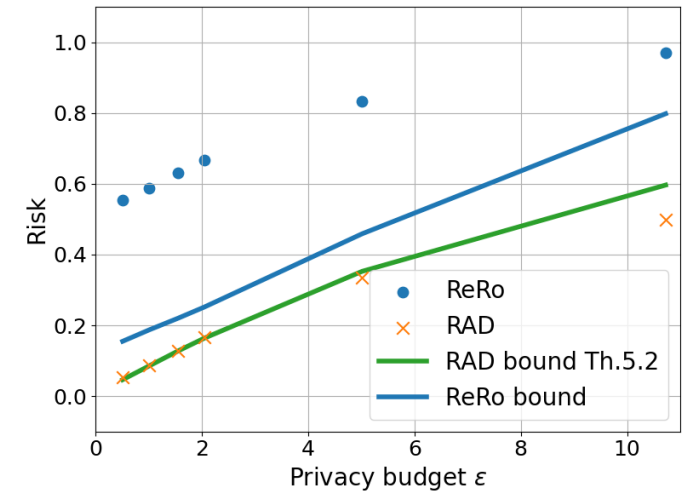
Results Optimal Attack against DP-SGD



DRA, $aux = \{\emptyset\}$.



DRA, $a(x) = \text{image label}$.



MIA, $a(x) = x$.

Experimental details:

- $\eta = 0$
- We set $|D_-| = 999$ and $|D_- \cup \{x\}| = 1\,000$
- Full-batch DP-SGD for $T = 100$ steps.
- We set the clipping rate, $C = 0.1$ and $\delta = 10^{-5}$

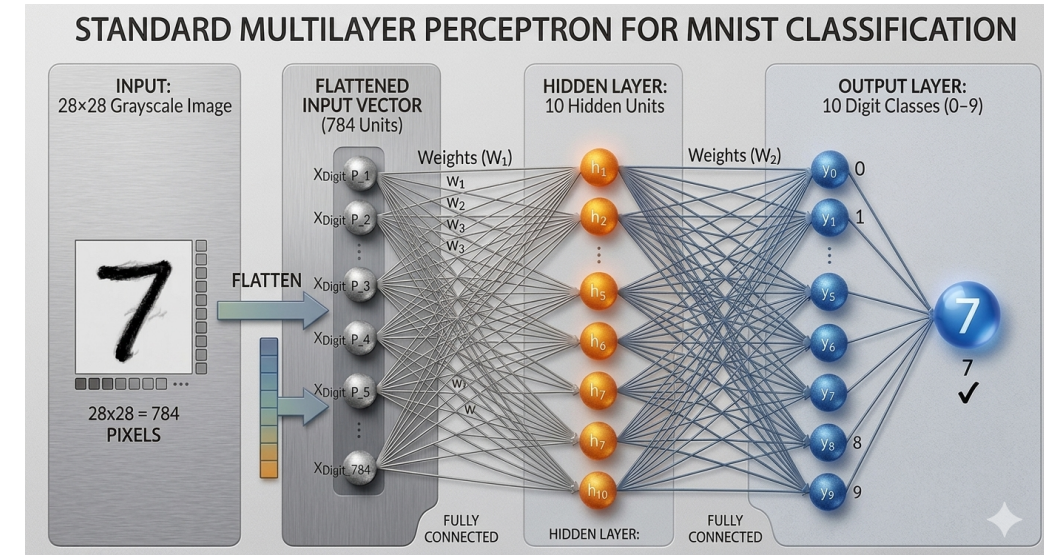
Details about DP-SGD Experiments

Detailed Architecture:

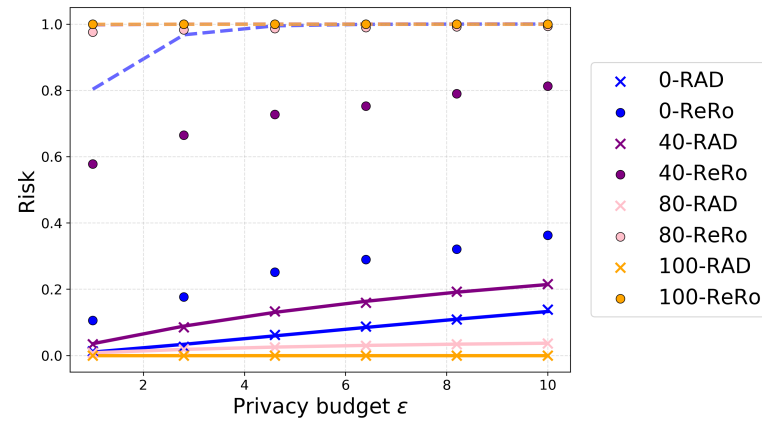
- Standard **multilayer perceptron** (MLP) with 2 layers.
- **Input:** 28×28 grayscale image (784-dimensional vector).
- The input passes through a **fully connected** layer with 10 hidden units, followed by a **ReLU** activation.
- The output layer consists of **10 units**, one for each digit class (0–9).

ϵ	MNIST	Fashion
1	0.46	0.46
10	0.75	0.75
20	0.75	0.77
50	0.75	0.80

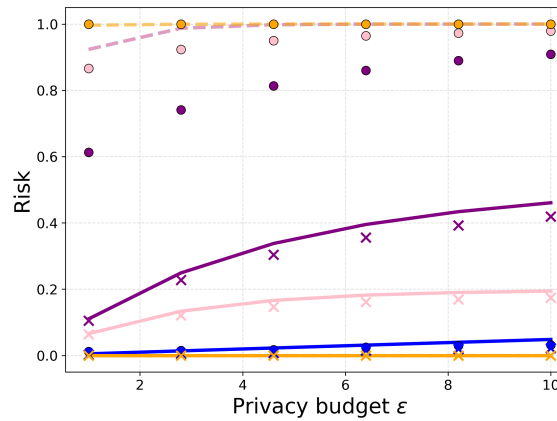
Table: Average test accuracy for different ϵ values under DP-SGD on MNIST and Fashion.



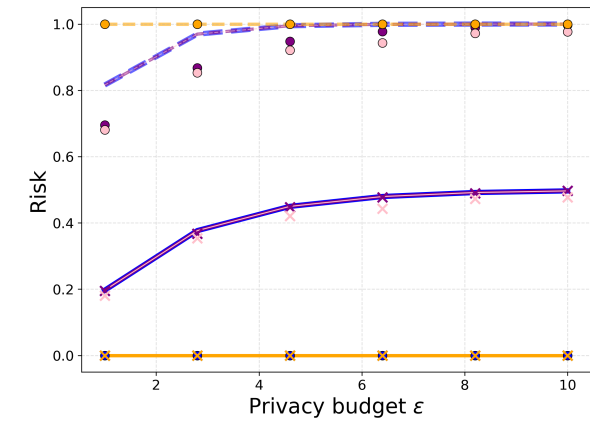
Results Against Laplace Mechanism



Adult distribution.

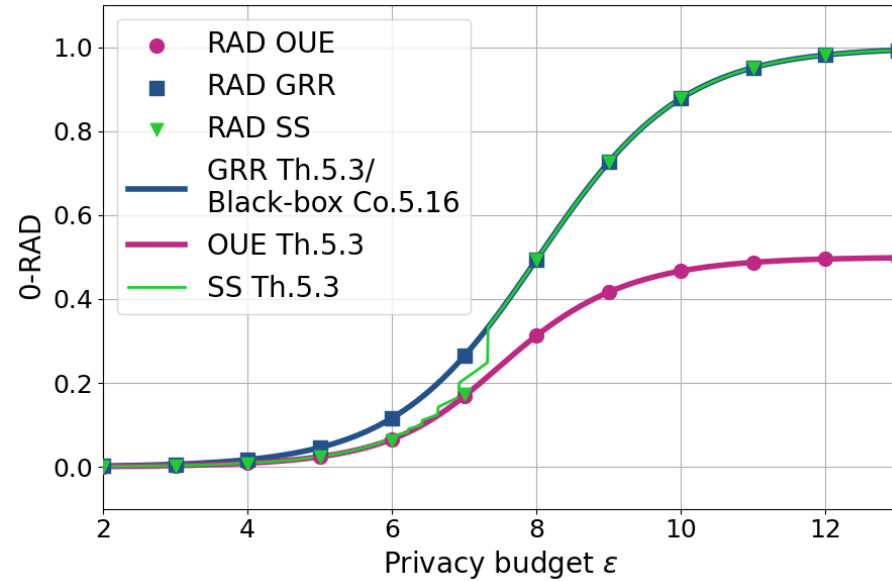


Uniform distribution.

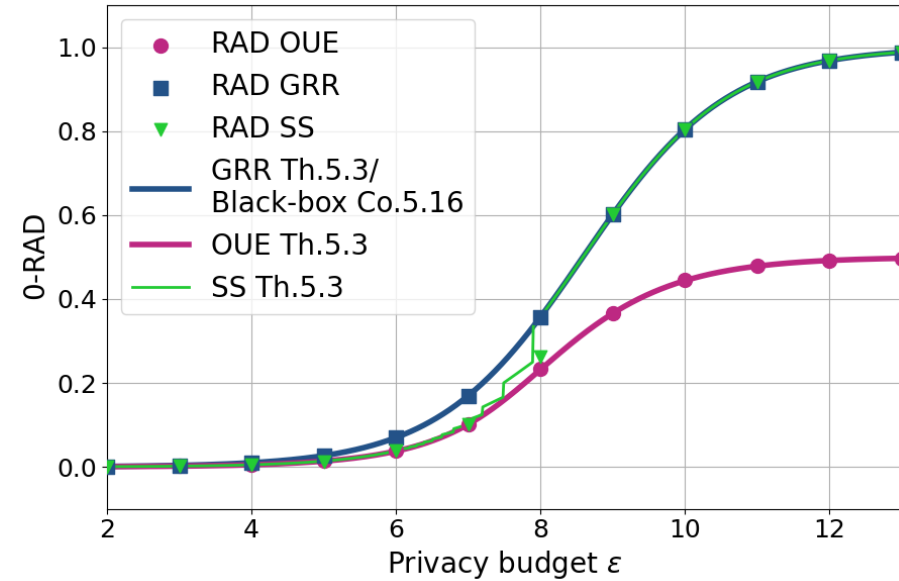


$\pi(0) = \pi(100) = \frac{1}{2}$.

Results Against LDP mechanisms

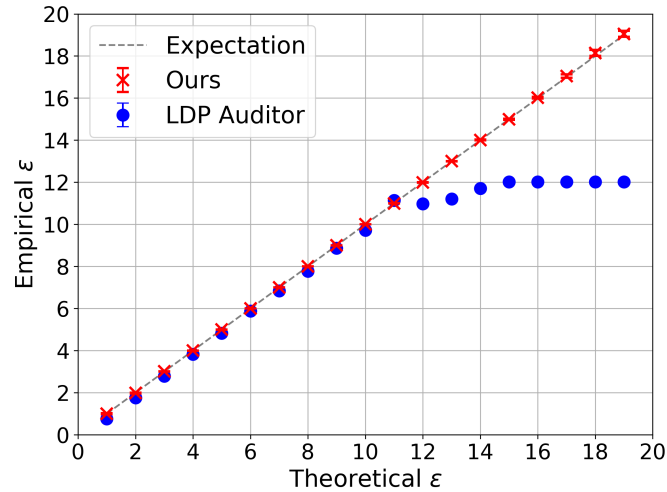


Porto

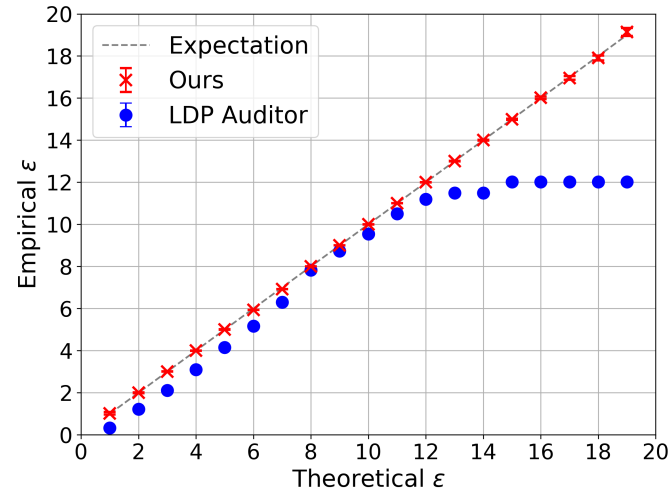


Geolife

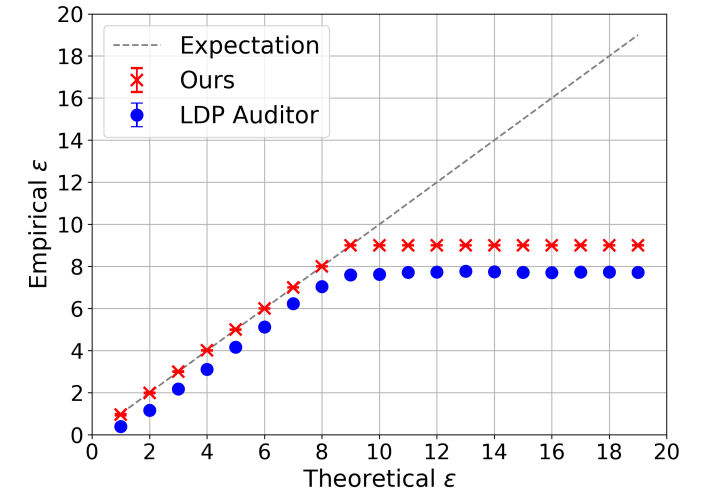
Results RAD-based Auditor



(GRR)



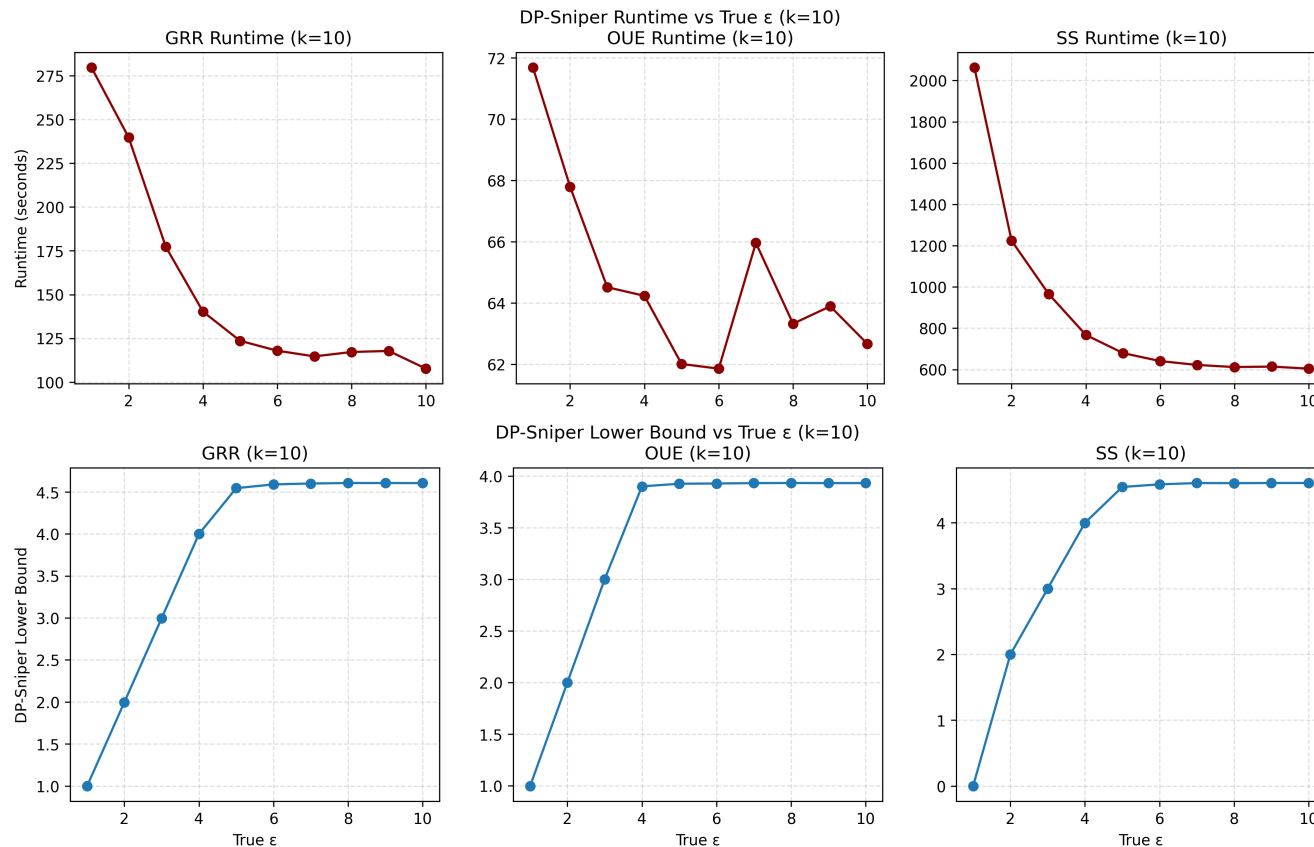
(SS)



(OUE)

Scalability Problem of ML-based Auditing

■ $k = |\mathcal{X}|$, Porto, $|\mathcal{X}| = 3\,052$, Geolife, $|\mathcal{X}| = 5\,356$.



DP-Sniper evaluation on a small categorical domain ($k = 10$) for GRR, OUE, and SS.

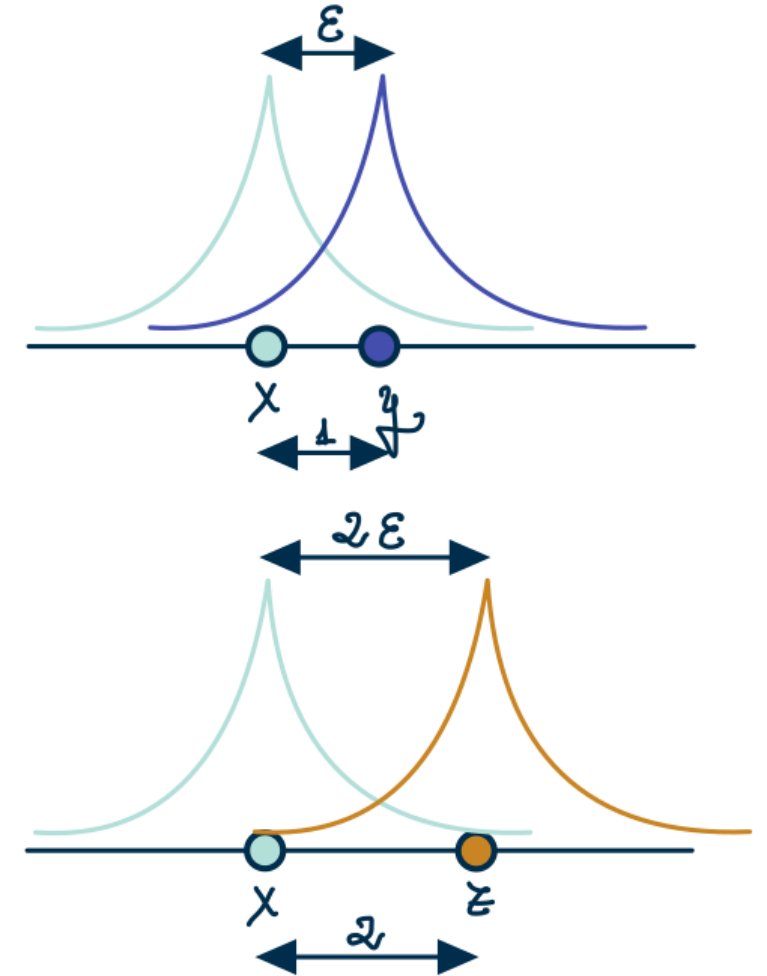
Top: inferred DP lower bounds.
Bottom: wall-clock runtime per ϵ value.

- DP-Sniper saturates near $\epsilon \approx 4.6$ due to its internal clipping mechanism,
- and requires several minutes per audit.
- Conclusion: computationally infeasible for large categorical LDP mechanisms.

Multivariate Gaussian Correlation

Main Result (Informal)

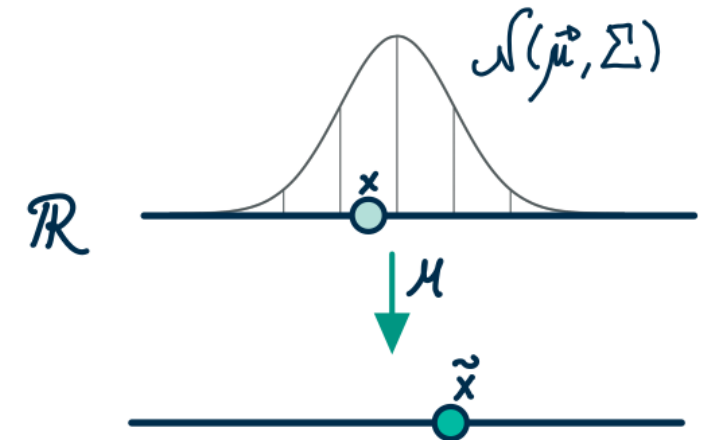
- Let $\mathcal{M}$ be an $\varepsilon\ell_1$ -private mechanism,



Multivariate Gaussian Correlation

Main Result (Informal)

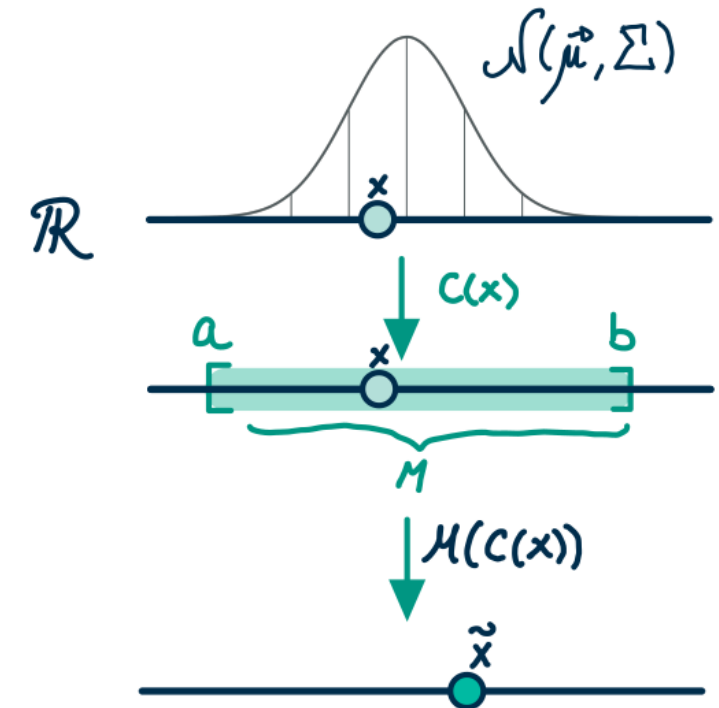
- Let $\mathcal{M}$ be an ε_{ℓ_1} -private mechanism,
- input data drawn from a multivariate Gaussian distribution



Multivariate Gaussian Correlation

Main Result (Informal)

- Let $\mathcal{M}$ be an $\varepsilon\ell_1$ -private mechanism,
- input data drawn from a multivariate Gaussian distribution
- $\rho(m-2) < 1$ is the maximum correlation coefficient.

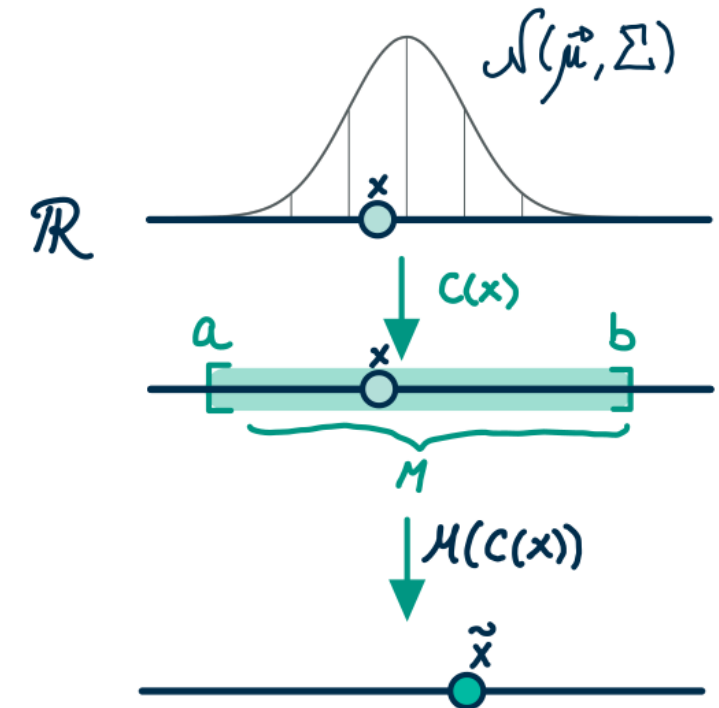


Multivariate Gaussian Correlation

Main Result (Informal)

- Let $\mathcal{M}$ be an $\varepsilon\ell_1$ -private mechanism,
- input data drawn from a multivariate Gaussian distribution
- $\rho(m-2) < 1$ is the maximum correlation coefficient.

Then, using clipping as preprocessing step, $c_l(D)_i = \max(a, \min(b, D_i))$, we obtain $\mathcal{M}_l$ satisfying



Multivariate Gaussian Correlation

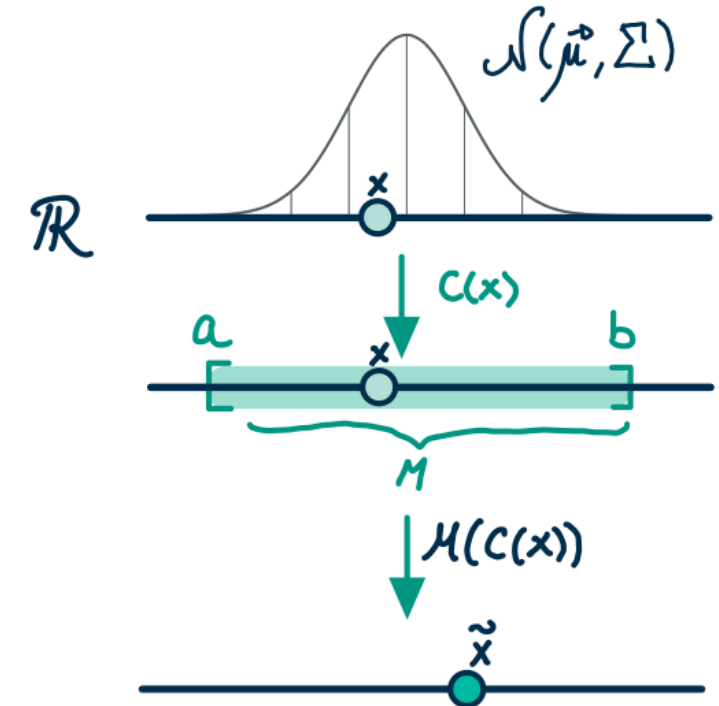
Main Result (Informal)

- Let $\mathcal{M}$ be an $\varepsilon\ell_1$ -private mechanism,
- input data drawn from a multivariate Gaussian distribution
- $\rho(m-2) < 1$ is the maximum correlation coefficient.

Then, using clipping as preprocessing step, $c_I(D)_i = \max(a, \min(b, D_i))$, we obtain $\mathcal{M}_I$ satisfying

$$\text{BDPL}(\mathcal{M}_I) \leq \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) M\varepsilon.$$

where M is the diameter of the interval $I = [a, b]$



Multivariate Gaussian Correlation

Proof Sketch

Trivial case: set of unknown records V , none is correlated with the target, $U = \emptyset$.

Otherwise, for any adversary H , we reorder $K = \{m - k, \dots, m - 1\}$ and $i = m$, denote $T = K \cup \{m\}$ and we show that if the principal **submatrix** Σ_T **is invertible**, then

$$\text{BDPL}_{(H,m)}(x_m, x'_m) \leq \varepsilon |x_m - x'_m| (\|\Sigma_{U;T} \Sigma_T^{-1} \mathbf{e}_{k+1}\|_1 + 1) \quad (5)$$

where $\mathbf{e}_{k+1} \equiv (0, \dots, 0, 1)^\top \in \mathbb{R}^{k+1}$ and the notation of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is reordered as

$$\Sigma_\rho = \begin{pmatrix} \Sigma_U & \Sigma_{U;T} & \mathbf{0} \\ \Sigma_{U;T}^\top & \Sigma_T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_S \end{pmatrix}.$$

If $m = 2$, given that $U \neq \emptyset$, then $k = 0$ and $\Sigma_{U;T} \Sigma_T^{-1} = \frac{\rho \sigma_1 \sigma_2}{\sigma_2^2} \leq \rho$. Using Eq. 5 we obtain:

$$\text{BDPL}_{(H,i)}(x_i, x'_i) \leq (\rho + 1) \varepsilon |x'_i - x_i|.$$

Markov Chain Correlation Model (Theoretical Results)

Main result (Informal)

- Let $\mathcal{M}$ be an ε -DP mechanism,
- input data sampled from Markov chain with transition matrix $P \in \mathbb{R}^{s \times s}$ and initial distribution $w \in \mathbb{R}^s$ with the following properties:

(H1) For all $x, y \in \mathcal{S}$ we have $P_{x,y} > 0$ and, (H2) $wP = w$.

Then, $\mathcal{M}$ is an $(\varepsilon + 4 \ln \gamma)$ -BDP mechanism where $\gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}$.

Markov Chain Correlation Model (Theoretical Results)

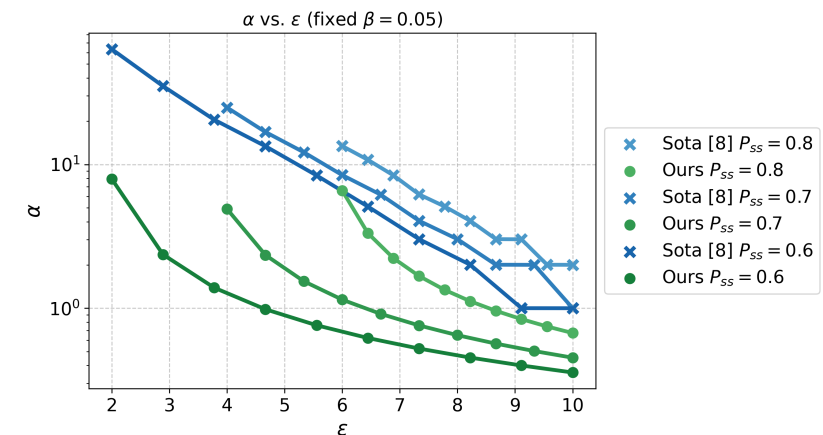
Main result (Informal)

- Let $\mathcal{M}$ be an ε -DP mechanism,
- input data sampled from Markov chain with transition matrix $P \in \mathbb{R}^{s \times s}$ and initial distribution $w \in \mathbb{R}^s$ with the following properties:

(H1) For all $x, y \in \mathcal{S}$ we have $P_{x,y} > 0$ and, (H2) $wP = w$.

Then, $\mathcal{M}$ is an $(\varepsilon + 4 \ln \gamma)$ -BDP mechanism where $\gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}$.

Previous mechanism	Ours
$P_{xy} > 0$	$P_{xy} > 0$
stationary	stationary
lazy	
binary	
symmetric	
$\varepsilon' > 0$	$\varepsilon' > 4 \ln(\gamma)$

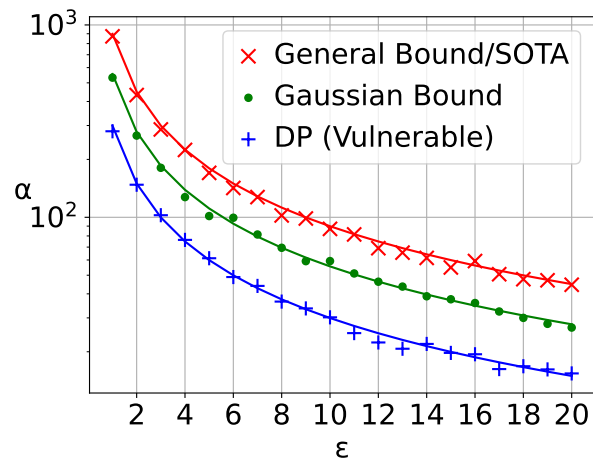


Experiment Details

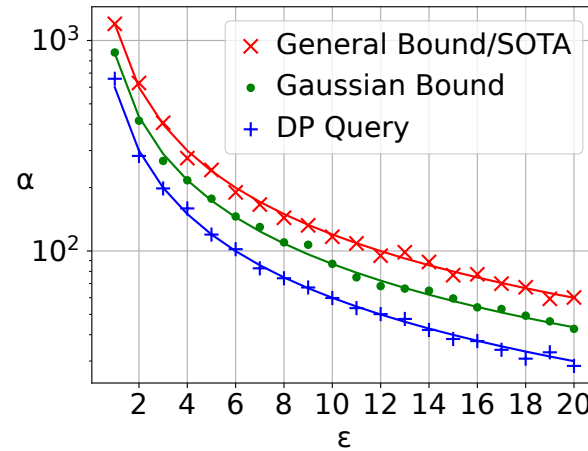
Database	n	m	Parameters	Sensitivity
Galton	897	3	$\rho = 0.275$	$\Delta q = 254cm$
FamilyIQ	868	2	$\rho = 0.4483$	$\Delta q = 120$
SyntheticIQ	20000	2	$\rho = 0.45$	$\Delta q = 120$
Activity	17568	n	$\gamma = 7.54$	$\Delta q = 1$
Activity Single Day	288	n	$\gamma = 7.54$	$\Delta q = 1$
Electricity	731	n	70 kWh, $\gamma = 3.29$ 80 kWh, $\gamma = 4.49$ 90 kWh, $\gamma = 8.43$	$\Delta q = 1$

Table: Data description. m is the max number of correlated records and n the total amount.

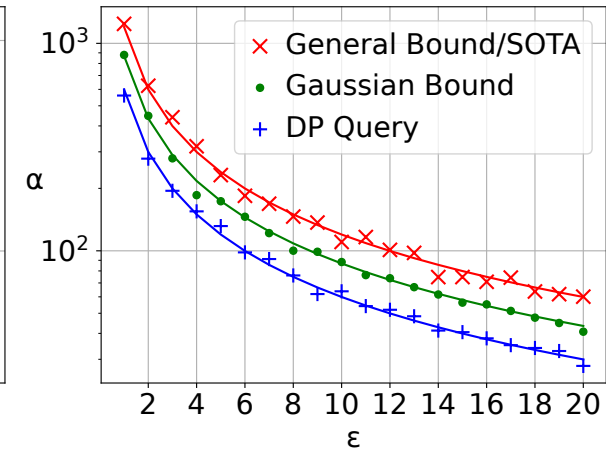
Multivariate Gaussian More Results



(a) Galton, $n = 897$, $m = 3$



(b) FamilyIQ, $n = 868$, $m = 2$.



(c) SyntheticIQ, $n = 20000$, $m = 2$.

Figure: Gaussian data results. Lines show theoretical error at $\beta = 5\%$ and markers indicate empirical 95% upper bounds.