Security and Privacy of ML

Differential Privacy

Shang-Tse Chen

Department of Computer Science
& Information Engineering
National Taiwan University

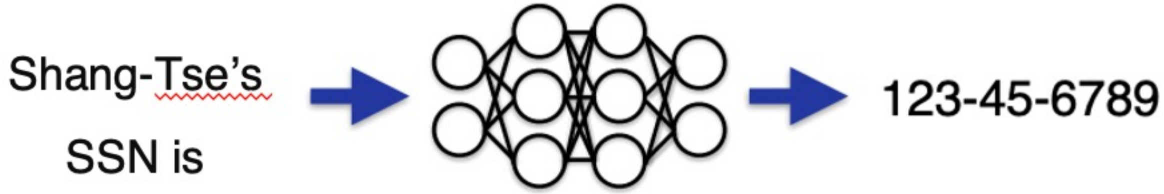
Review: Potential Data Leakage

- Model inversion attack

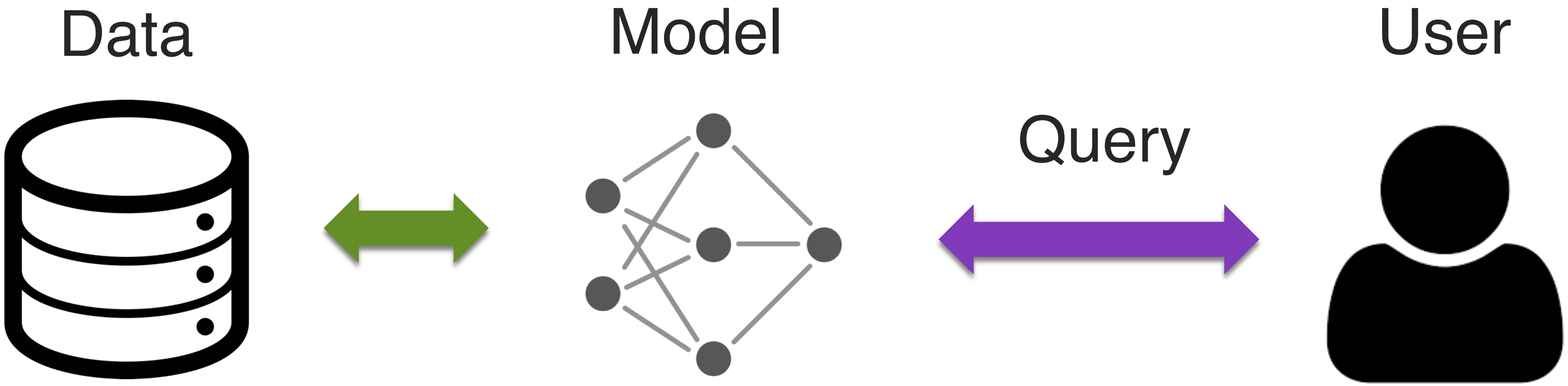
[Fredrikson et al. ‘15]

- Extract unintended memorization

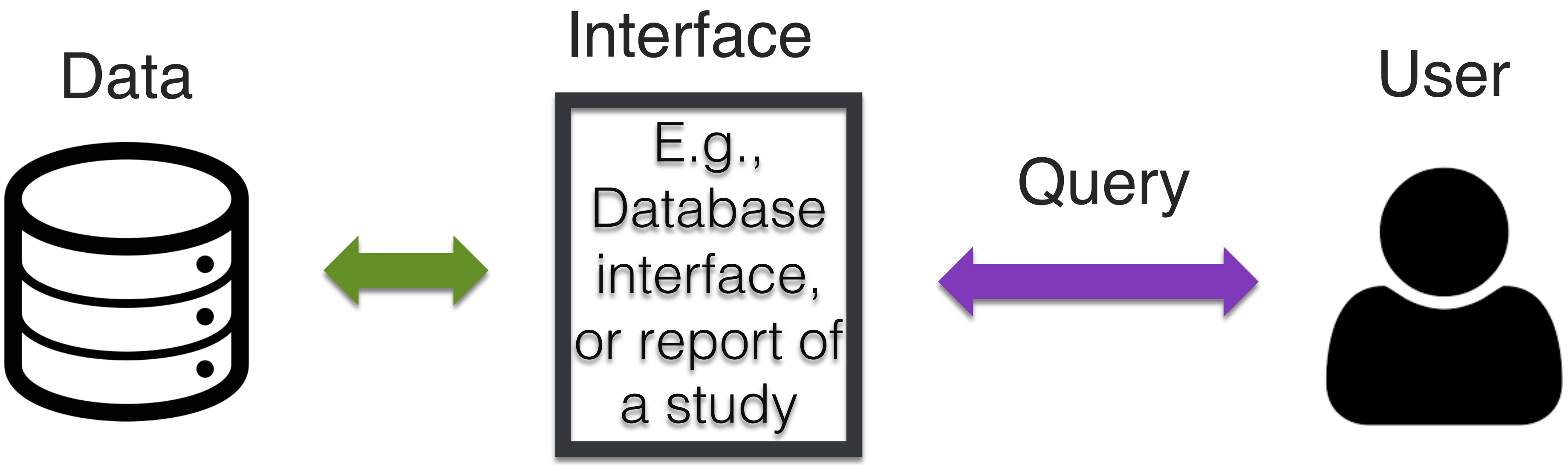
[Carlini et al. Usenix Security Symposium 2019]



Review: Generic Framework



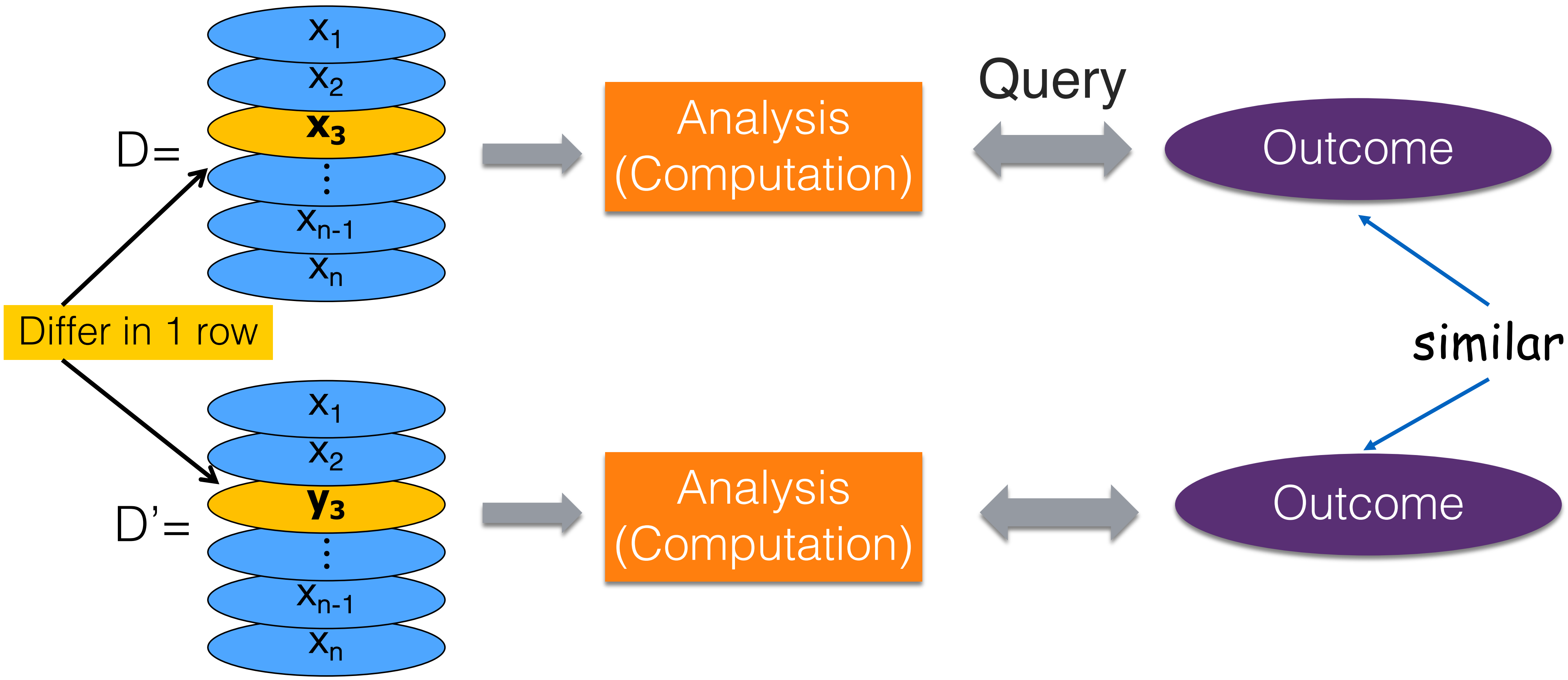
Review: Generic Framework



How do we provide useful information to user, while preserving privacy of individuals in the data?

Differential Privacy

[Dwork et al. '06]



Example Query: Counting Query

$$q' : \mathcal{X} \to \{0,1\}$$

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} q'(x_i)$$

E.g., Fraction of people having disease: 1/2

Disease (Y/N)
Y
Y
N
Y
N
N

Definition of Differential Privacy

A (randomized) algorithm $M: X^n \times Q \to T$ is ϵ-differential private if
for all datasets $x, x' \in X^n$ that differ on one entry and every query $\mathrm{q} \in Q$,
for all subsets S of the outcome space T,

$$\frac{\Pr_{\mathrm{M}}[M(x,q) \in S]}{\Pr_{\mathrm{M}}[M(x',q) \in S]} \leq A$$

- A should be close to 1
- If A >> 1, little privacy is guaranteed
- If A = 1, individuals have no effect on the results and there is zero utility

Definition of Differential Privacy

A (randomized) algorithm $M: X^n \times Q \to T$ is ϵ-differential private if
for all datasets $x, x' \in X^n$ that differ on one entry and every query $\mathrm{q} \in Q$,
for all subsets S of the outcome space T,

$$\frac{\Pr_{\mathrm{M}}[M(x,q) \in S]}{\Pr_{\mathrm{M}}[M(x',q) \in S]} \leq (1+\epsilon)$$

- A should be close to 1
- If A >> 1, little privacy is guaranteed
- If A = 1, individuals have not effect on the results and there is zero utility

Definition of Differential Privacy

A (randomized) algorithm $M: X^n \times Q \rightarrow T$ is ϵ-differential private if
for all datasets $x, x' \in X^n$ that differ on one entry and every query $\mathrm{q} \in Q$,
for all subsets S of the outcome space T,

$$\Pr_{\mathrm{M}}[M(x, q) \in S] \leq (1 + \epsilon) \Pr_{\mathrm{M}}[M(x', q) \in S]$$

Definition of Differential Privacy

A (randomized) algorithm $M: X^n \times Q \to T$ is ϵ-differential private if
for all datasets $x, x' \in X^n$ that differ on one entry and every query $\mathrm{q} \in Q$,
for all subsets S of the outcome space T,

$$\Pr_{\mathrm{M}}[M(x,q) \in S] \leq e^{\epsilon} \Pr_{\mathrm{M}}[M(x',q) \in S]$$

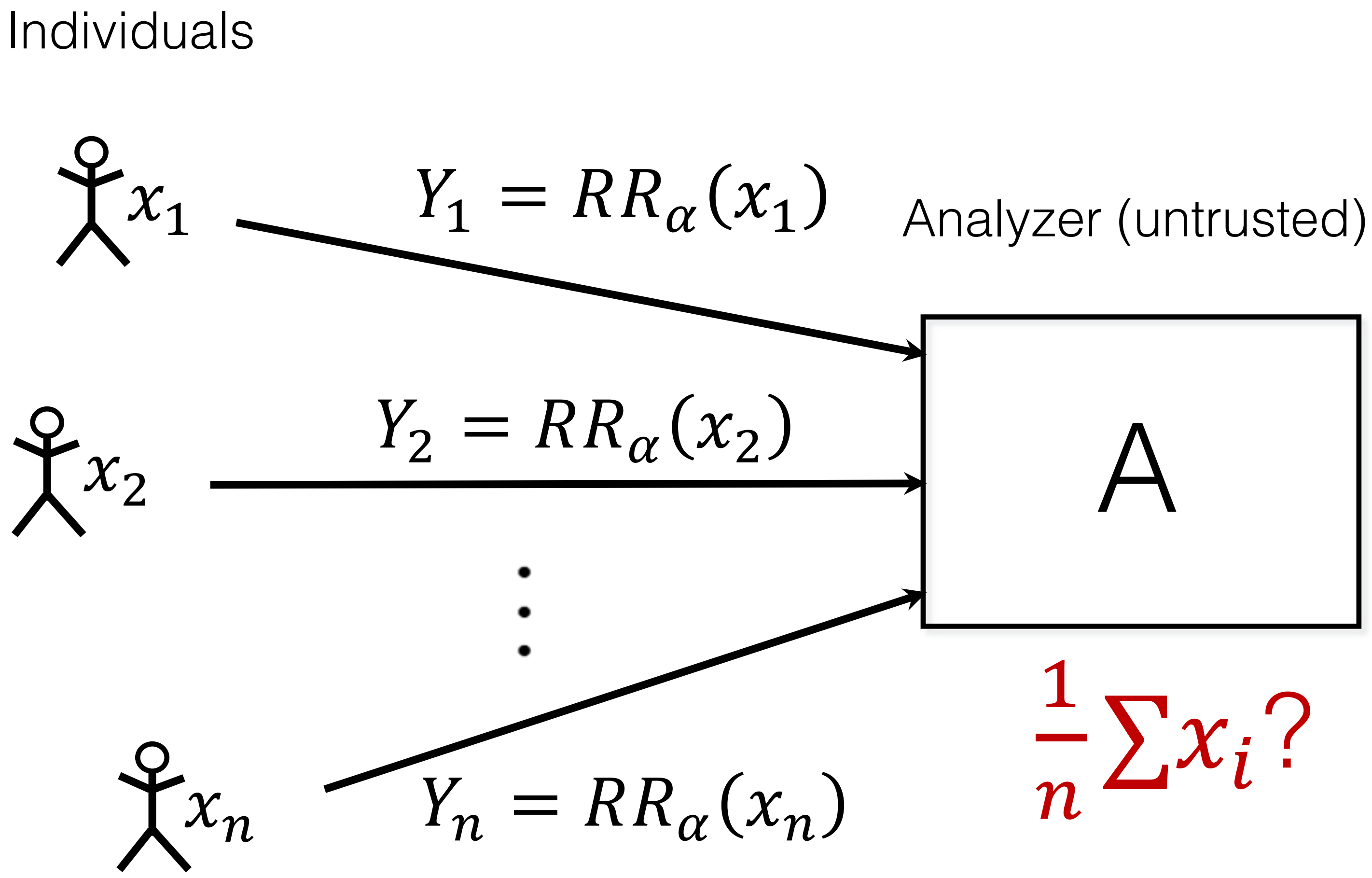
- For small ϵ: $e^{\epsilon} \approx 1 + \epsilon$, but is mathematically more convenient
- ϵ not small in cryptographical sense. Think $\epsilon \approx \frac{1}{100}$ or $\epsilon \approx \frac{1}{10}$
- This is called (pure) differential privacy

Randomized Response [Warner '65]

- $q(x) \in \{0,1\}$
- $RR_\alpha(x) = \begin{cases} q(x) & w.p.\ \frac{1}{2} + \alpha \\ \neg q(x) & w.p.\ \frac{1}{2} - \alpha \end{cases}$
- Claim: setting $\alpha = \frac{1}{2}\frac{e^\epsilon - 1}{e^\epsilon + 1}$, $RR_\alpha(x)$ is ϵ −differentially private
- Proof:
  - Neighboring databases: $\mathrm{q}(x_i) = 0;\ \ q(x_i') = 1$
  - $\frac{\Pr[RR(0)=0]}{\Pr[RR(1)=0]} = \frac{\frac{1}{2}(1+\frac{e^\epsilon-1}{e^\epsilon+1})}{\frac{1}{2}(1-\frac{e^\epsilon-1}{e^\epsilon+1})} = e^\epsilon$

small ϵ: $\mathrm{e}^\epsilon \approx 1 + \epsilon$
Get $\alpha \approx \frac{\epsilon}{4}$

Is Randomized Response Accurate?



Is Randomized Response Accurate?

- $E[Y_i] = x_i\left(\frac{1}{2}+\alpha\right) + (1-x_i)\left(\frac{1}{2}-\alpha\right) = \frac{1}{2} + \alpha(2x_i - 1)$
- Put $\widehat{x_i} = \frac{Y_i - \frac{1}{2} + \alpha}{2\alpha}$ then $E[\widehat{x_i}] = x_i$
- But $Var[\widehat{x_i}] = \frac{\frac{1}{4} - \alpha^2}{4\alpha^2} \approx \frac{1}{\epsilon^2}$ high!
- $E[\frac{1}{n}\sum\widehat{x_i}] = \frac{1}{n}\sum x_i$ and $Var[\frac{1}{n}\sum\widehat{x_i}] = \frac{1}{n}\frac{\frac{1}{4}-\alpha^2}{4\alpha^2} \approx \frac{1}{n\epsilon^2}$; stdev $\approx \frac{1}{\sqrt{n}\epsilon}$
- Useful when $n \gg \frac{1}{\epsilon^2}$

Laplace Mechanism

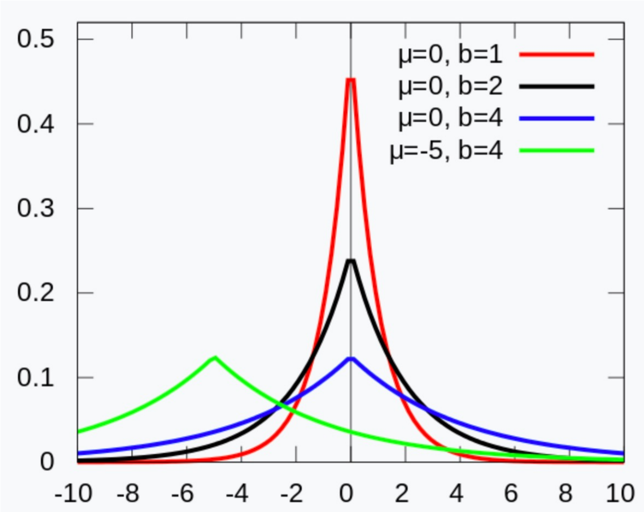
- Let q be a counting query
- Idea: $\mathrm{M}(x) = q(x) + z$, where z is some random noise
- How much noise is enough?
- If $x \sim x' \rightarrow |q(x) - q(x')| \leq \frac{1}{n}$
- $\Pr[\mathrm{M}(x) = y] = \Pr[q(x) + z = y] = \Pr[z = y - q(x)]$
- $\Pr[\mathrm{M}(x') = y] = \Pr[q(x') + z' = y] = \Pr[z' = y - q(x')]$
- $|z - z'| \leq \frac{1}{n}$
- Find a distribution that change by a factor of at most e^{ϵ} over intervals of length 1/n

Disease (Y/N)
Y
Y
N
Y
N
N

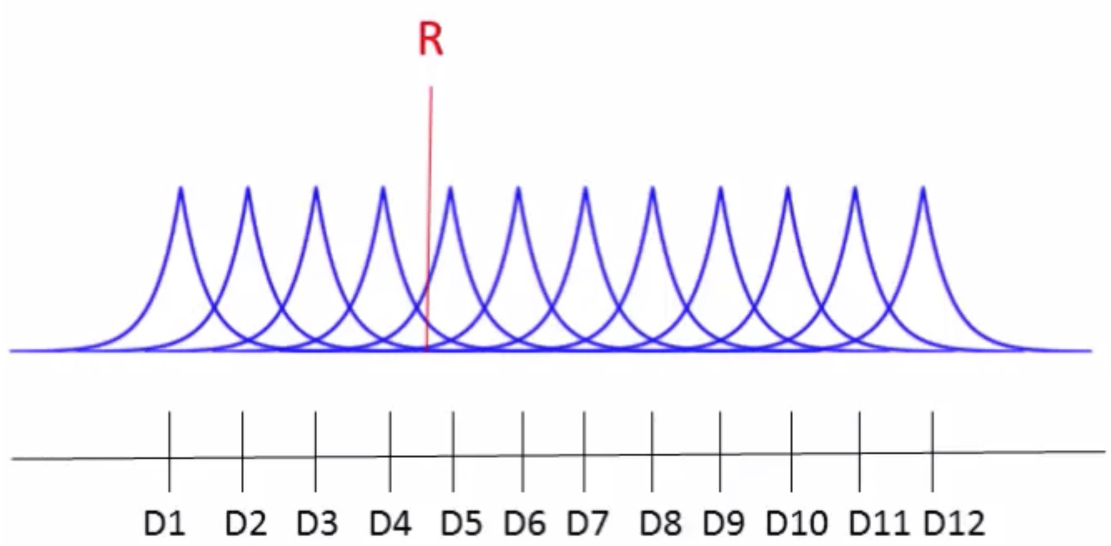
Laplace Mechanism

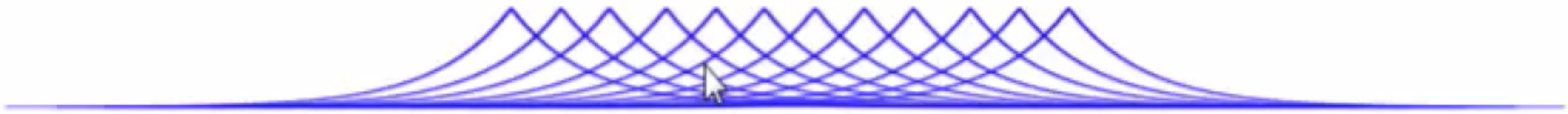
- Laplace distribution $\mathrm{Lap}(b)$
- Density of $\mathrm{Lap}(b)$ at z: $\frac{1}{2b}e^{-|z|/b}$
- If we set $\mathrm{b} = \frac{1}{\epsilon n}$:

$$\frac{\Pr\left[\mathrm{Lap}\left(\frac{1}{\epsilon n}\right) = z + \frac{1}{n}\right]}{\Pr\left[\mathrm{Lap}\left(\frac{1}{\epsilon n}\right) = z\right]} \leq e^{\epsilon}$$



Laplace Mechanism: Intuition





Accuracy of Laplace Mechanism

- Mean is accurate, because we add a zero-mean noise
- Std of $\mathrm{Lap}\left(\frac{1}{\epsilon n}\right)$ is $O\left(\frac{1}{\epsilon n}\right)$
- Significantly better than $\frac{1}{\sqrt{n}\epsilon}$ by randomized response

Global Sensitivity

- The analysis works for other types of queries
- Use $\mathrm{Lap}(\frac{\Delta f}{\epsilon})$ instead of $\mathrm{Lap}(\frac{1}{\epsilon n})$
- Global sensitivity $\Delta f = \max_{x \sim x'} |q(x) - q(x')|$

$$\frac{Pr(f(x)+Lap(\Delta f/\epsilon)=y)}{Pr(f(x')+Lap(\Delta f/\epsilon)=y)} = \frac{\exp\left(-\frac{|y-f(x)|\,\epsilon}{\Delta f}\right)}{\exp\left(-\frac{|y-f(x')|\epsilon}{\Delta f}\right)}$$

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

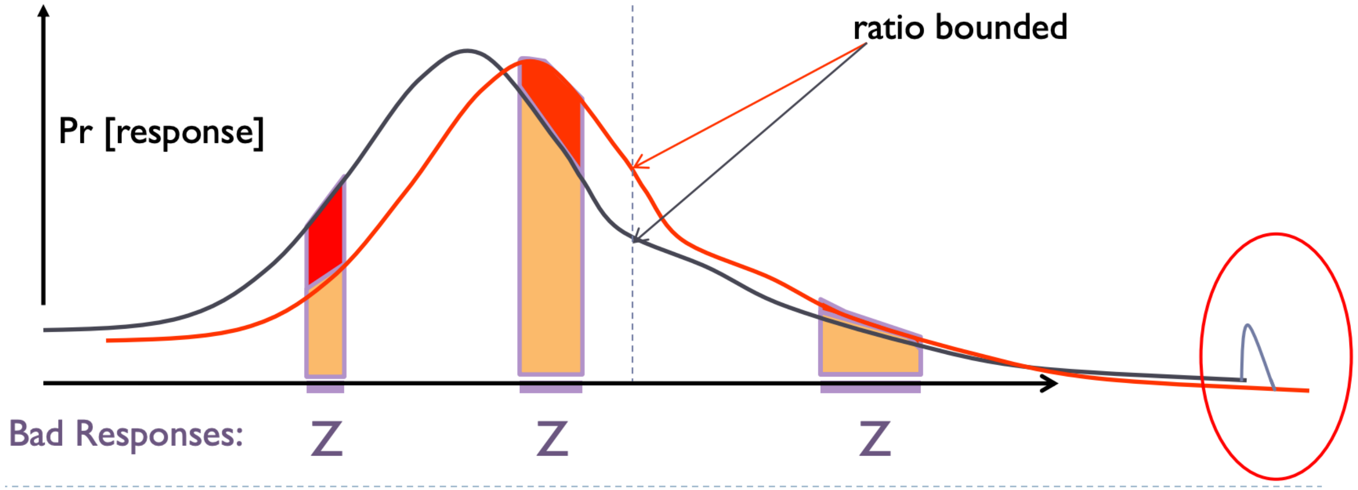
$$\leq \exp\left(\frac{\epsilon}{\Delta f}\left(|f(x) - f(x')|\right)\right) \leq e^{\epsilon}$$

$$\mathrm{P}[\mathrm{Lap}(b) = z] = \frac{1}{2b} e^{-|z|/b}$$

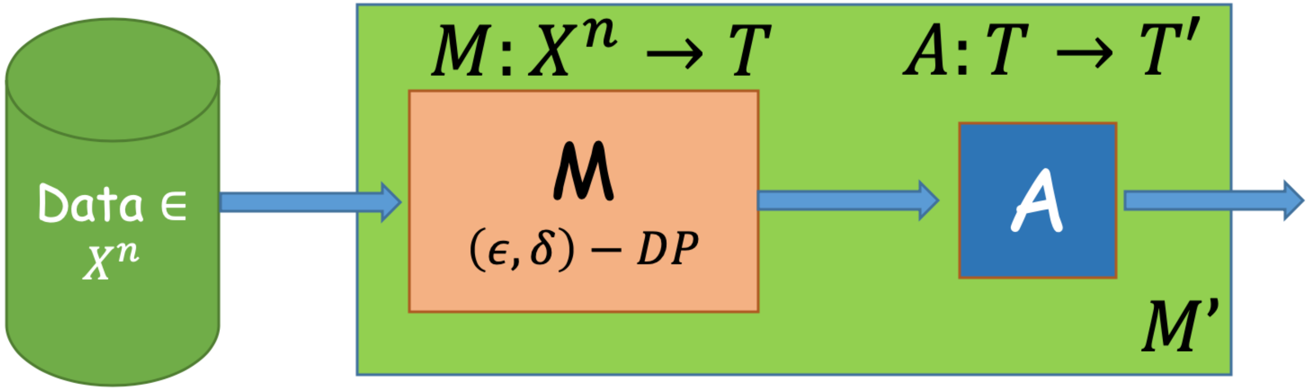
Approximate Differential Privacy

A (randomized) algorithm $M: X^n \times Q \to T$ is (ϵ, δ)-differential private if for all datasets $x, x' \in X^n$ that differ on one entry and every query $\mathrm{q} \in Q$, for all subsets S of the outcome space T,

$$\Pr_{\mathrm{M}}[M(x,q) \in S] \leq e^{\epsilon} \Pr_{\mathrm{M}}[M(x',q) \in S] + \delta$$



Basic Property of DP: Post Processing



- Claim: M' is (ϵ, δ)-differentially private
- Proof:
  - Let x, x' be neighboring databases and S' a subset of T'
  - Let $S = \{z \in T: A(z) \in S'\}$ be the preimage of S' under A

$$\Pr[M'(x) \in S'] = \Pr[M(x) \in S]$$
$$\leq e^{\epsilon} \Pr[M(x') \in S] + \delta \;\; = e^{\epsilon} \Pr[M'(x') \in S'] + \delta$$

Property of DP: Sequential Composition

- If M_1, M_2, ..., M_k are algorithms that access a private database D such that each M_i satisfies ε_i -differential privacy,

  then the combination of their outputs satisfies ε-differential privacy with $\varepsilon=\varepsilon_1+...+\varepsilon_k$

Property of DP: Parallel Composition

If M_1, M_2, ..., M_k are algorithms that access disjoint databases D_1, D_2, ..., D_k such that each M_i satisfies ε_i -differential privacy,

then the combination of their outputs satisfies ε-differential privacy with $\varepsilon = \max\{\varepsilon_1,...,\varepsilon_k\}$

Example Problem

Sex	Height	Weight
M	6′2″	210
F	5′3″	190
F	5′9″	160
M	5′3″	180
M	6′7″	250

Queries:

- # Males with BMI < 25
- # Males
- # Females with BMI < 25
- # Females

- ϵ-differentially private algorithm to answer all the questions?
- What is the total error?

Algorithm 1

Return:

- (# Males with BMI < 25) + Lap(4/ε)
- (# Males) + Lap(4/ε)
- (# Females with BMI) < 25 + Lap(4/ε)
- (# Females) + Lap(4/ε)

Privacy Analysis of Algorithm 1

- Sensitivity of each query is 1
- Each query is answered using a ε/4-DP algorithm
- By sequential composition, we get ε-DP

Utility Analysis of Algorithm 1

Error:

$$\sum E\left(\left(\tilde{q}(D) - q(D)\right)^2\right)$$

Total Error:

$$2\left(\frac{4}{\varepsilon}\right)^2 \times 4 = \frac{128}{\varepsilon^2}$$

Algorithm 2

Compute:

- $\widetilde{q_1}$ = (# Males with BMI < 25) + Lap(1/ε)
- $\widetilde{q_2}$ = (# Males with BMI > 25) + Lap(1/ε)
- $\widetilde{q_3}$ = (# Females with BMI < 25) + Lap(1/ε)
- $\widetilde{q_4}$ = (# Females with BMI > 25) + Lap(1/ε)

Return

- $\widetilde{q_1}$, $\widetilde{q_1}+\widetilde{q_2}$, $\widetilde{q_3}$, $\widetilde{q_3}+\widetilde{q_4}$

Privacy Analysis of Algorithm 2

- Sensitivity of count = 1. So each query is answered using a ε-DP algorithm.

- q_1, q_2, q_3, q_4 are counts on disjoint portions of the database. Thus by *parallel composition* releasing $\widetilde{q_1}, \widetilde{q_2}, \widetilde{q_3}, \widetilde{q_4}$ satisfies ε-DP.

- By the *postprocessing theorem*, releasing $\widetilde{q_1}$, $\widetilde{q_1}+\widetilde{q_2}$, $\widetilde{q_3}$, $\widetilde{q_3}+\widetilde{q_4}$ also satisfies ε-DP.

Utility Analysis of Algorithm 2

Error:

$$\sum E\left(\left(\tilde{q}(D) - q(D)\right)^2\right)$$

Total Error:

$$2\left(\frac{1}{\varepsilon}\right)^2 + 2 \cdot 2\left(\frac{1}{\varepsilon}\right)^2 + 2\left(\frac{1}{\varepsilon}\right)^2 + 2 \cdot 2\left(\frac{1}{\varepsilon}\right)^2 = \frac{12}{\varepsilon^2}$$

$\widetilde{q_1}$ $\widetilde{q_1} + \widetilde{q_2}$ $\widetilde{q_3}$ $\widetilde{q_3} + \widetilde{q_4}$

Generalized Sensitivity

- Let $f : \mathcal{D} \rightarrow \mathbb{R}^d$ be a function that outputs a vector of d real numbers. The sensitivity of f is given by:

$$S(f) = \max_{D,D' : |D \Delta D'| = 1} \|f(D) - f(D')\|_1$$

where $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_i |x_i - y_i|$

Generalized Sensitivity

- q_1 = # Males with BMI < 25
- q_2 = # Males with BMI > 25
- q = # Males with BMI

- Let f_1 be a function that answers both q_1, q_2
- Let f_2 be a function that answers both q_1, q

- Sensitivity of $f_1 = 1$
- Sensitivity of $f_2 = 2$

- An alternate privacy proof for Alg 2 is to show that the generalized sensitivity of $\widetilde{q_1}$, $\widetilde{q_2}$, $\widetilde{q_3}$, $\widetilde{q_4}$ is 1.

Improving Utility of Algorithm 2

Compute:

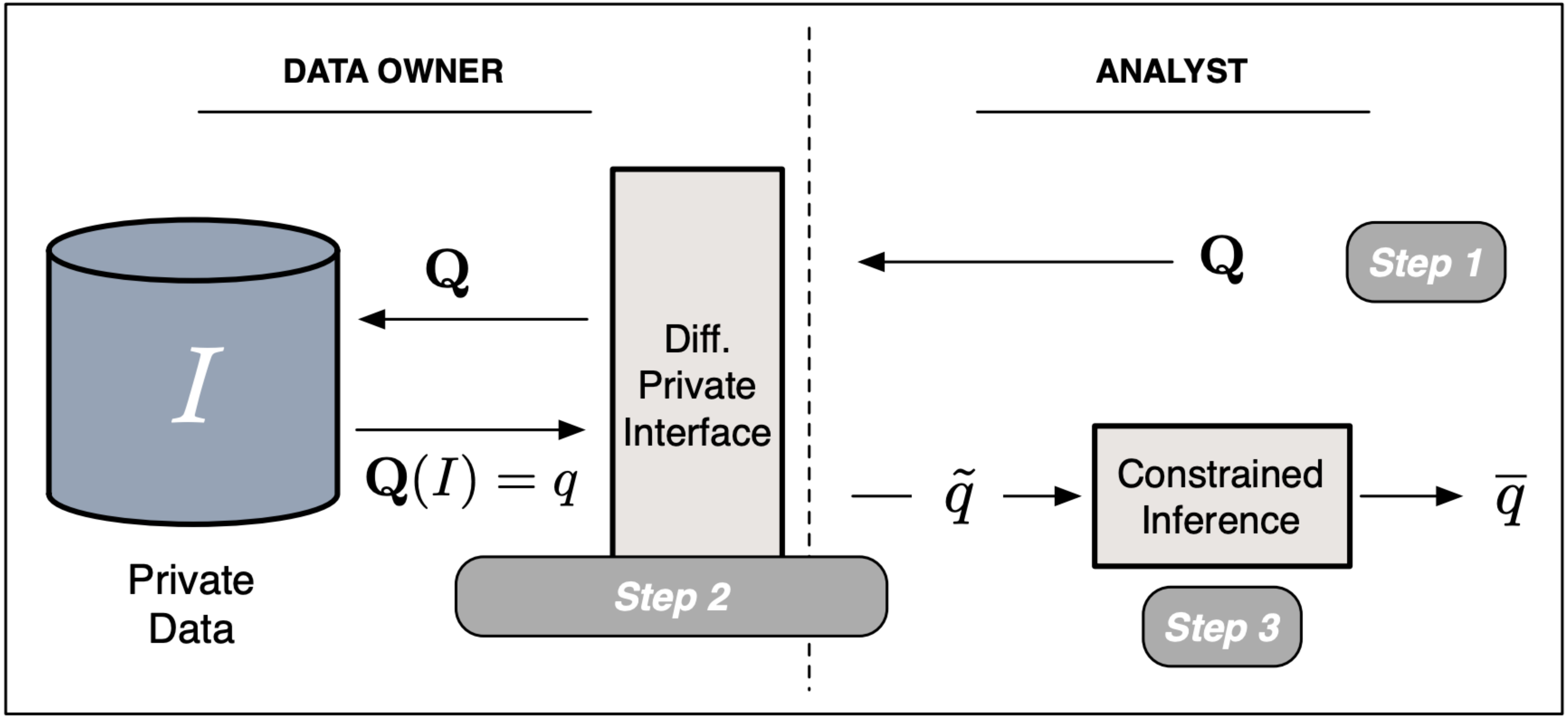
- $\widetilde{q_1}$ = # Males with BMI < 25 + Lap(1/ε)
- $\widetilde{q_2}$ = # Males with BMI > 25 + Lap(1/ε)

Return

- $\widetilde{q_1}$, $\widetilde{q_1}+\widetilde{q_2}$

We know $q_1 \leq q_1 + q_2$,
but $P[\widetilde{q_1} > \widetilde{q_1}+\widetilde{q_2}] > 0$

Constrained Inference



Least Squares Optimization

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

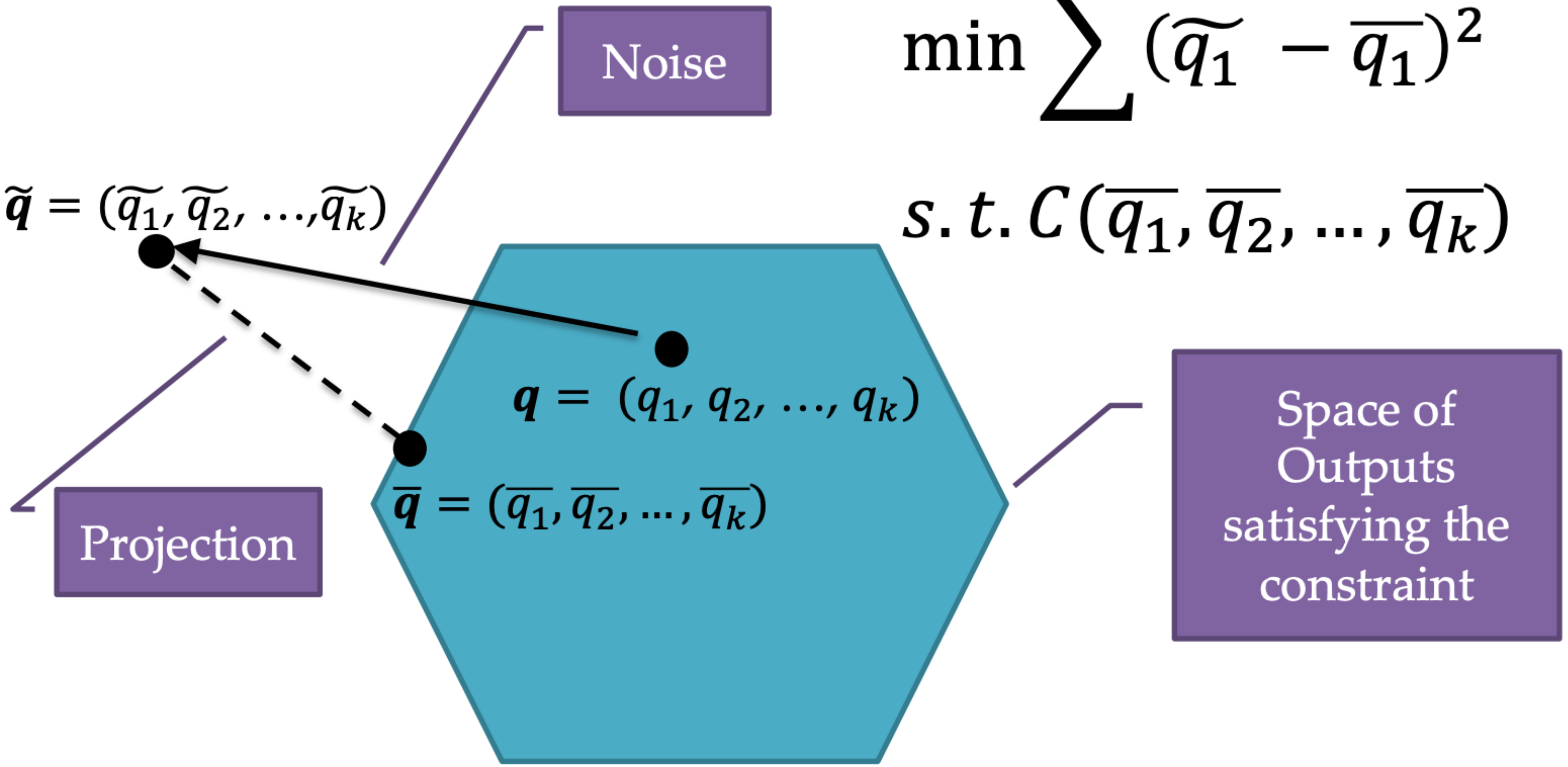
such that

$$\text{Constraint}(\overline{q_1}, \overline{q_2}, \ldots \overline{q_k}) = \text{True}$$

Geometric Interpretation

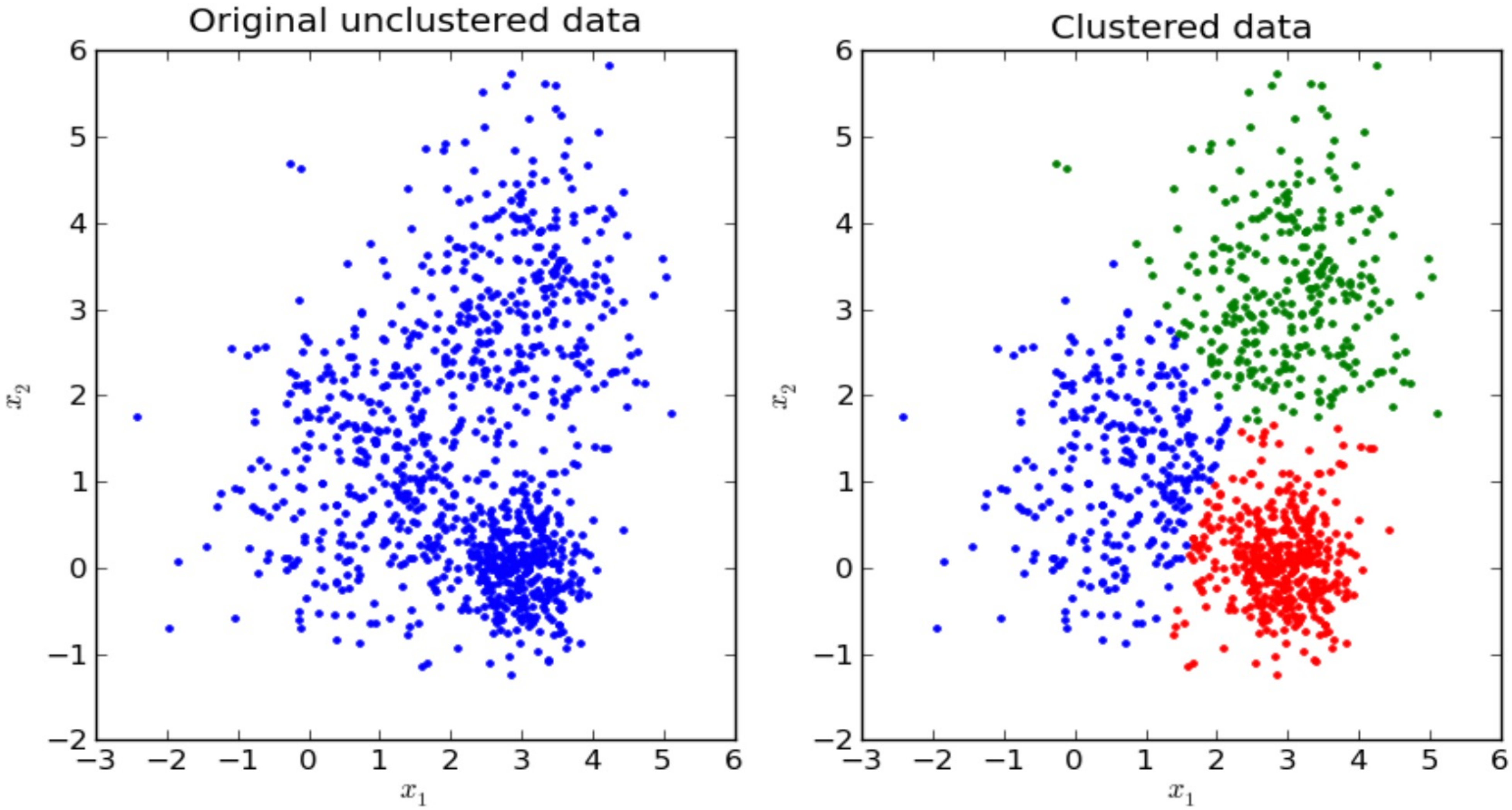
$$\min \sum (\widetilde{q_1} - \overline{q_1})^2$$

$$s.t.\, C(\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k})$$



Theorem: $\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_2 \leq \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|_2$ when the constraints form a convex space

Case Study: K-means Clustering



K-means: Problem

Partition a set of points $x_1, \ldots, x_n$ into k clusters $S_1, \ldots, S_k$ such that the following is minimized:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|_2^2$$

where μ_i is the mean of S_i

K-means: Algorithm

- Initialize a set of k centers
- Repeat until convergence:
  - Assign each point to its nearest center
  - Recompute the set of centers

- Output final set of k centers

Differentially Private K-means

[Blum et al. PODS '05]

- Suppose we fix the number of iterations to T
  - Each iteration uses ϵ/T privacy budget, total privacy loss is ϵ
- In each iteration (given a set of centers):
  - Assign the points to the new center to form clusters
  - Noisily compute the size of each cluster
  - Compute noisy sums of points in each cluster

Differentially Private K-means

[Blum et al. PODS ‘05]

Which of these steps expends privacy budget?

In each iteration (given a set of centers):

No
- Assign the points to the new center to form clusters

Yes
- Noisily compute the size of each cluster

Yes
- Compute noisy sums of points in each cluster

Differentially Private K-means

[Blum et al. PODS '05]

What is the sensitivity?

In each iteration (given a set of centers):

- Assign the points to the new center to form clusters
- Noisily compute the size of each cluster — 1
- Compute noisy sums of points in each cluster — data dependent e.g., if $x \in [0,1]^d$, then sensitivity $= d$

Differentially Private K-means

[Blum et al. PODS '05]

What noise do we add?

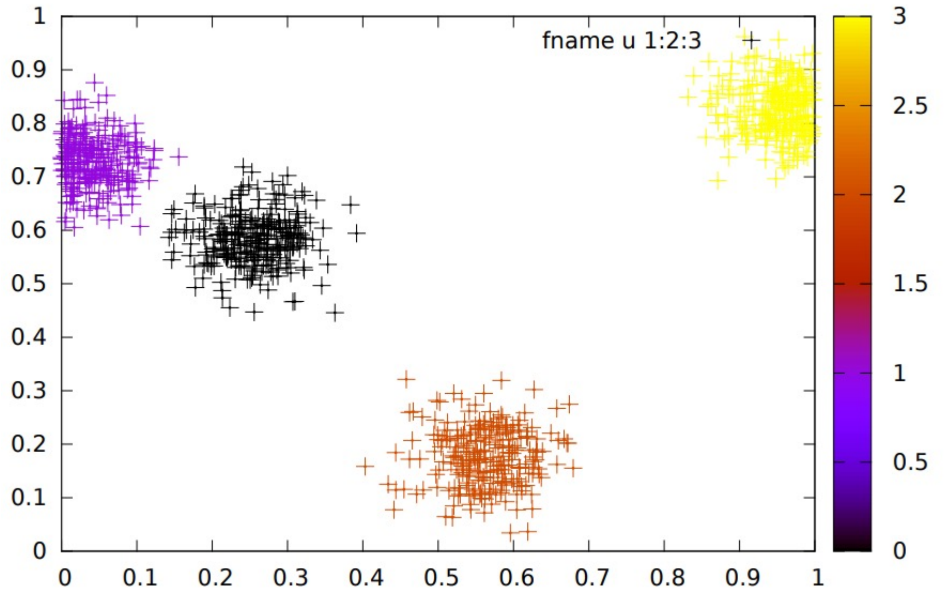
In each iteration (given a set of centers):

- Assign the points to the new center to form clusters
- Noisily compute the size of each cluster Lap$(2T/\epsilon)$
- Compute noisy sums of points in each cluster Lap$(2dT/\epsilon)$

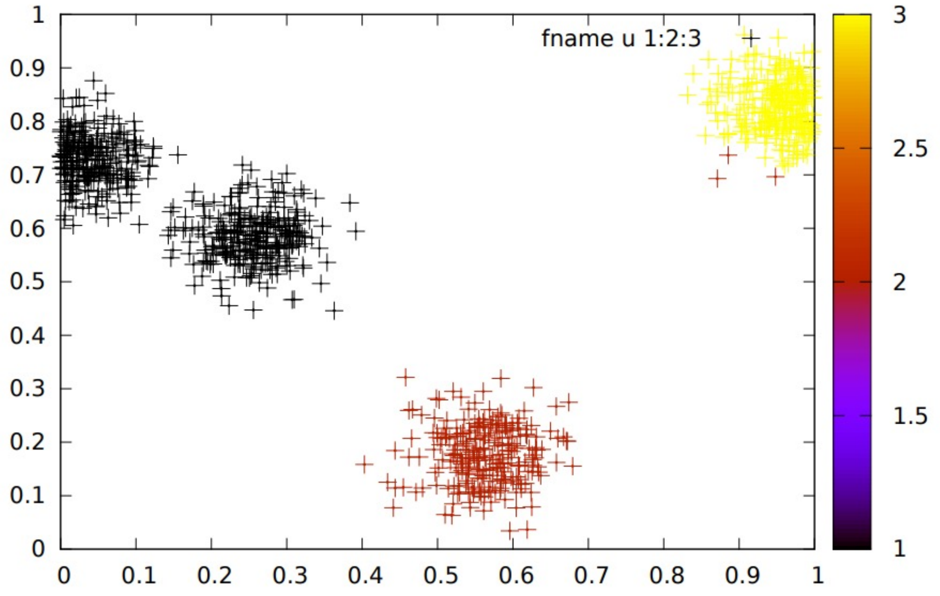
Results

- Can distinguish clusters that are far apart
- Can’t distinguish small clusters that are close by

Original Kmeans algorithm



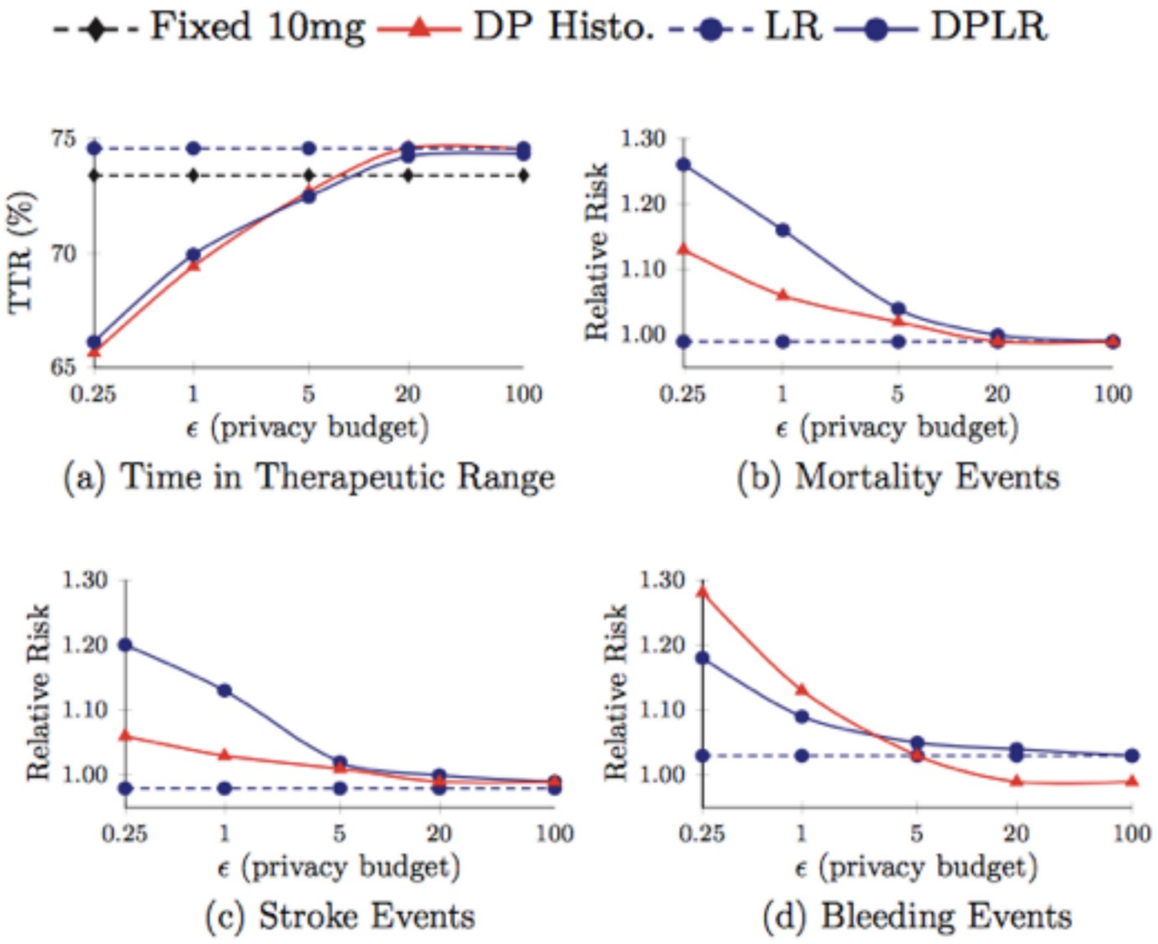
Laplace Kmeans algorithm



Application: Prevent Memorization

	Optimizer	ε	Test Loss	Estimated Exposure	Extraction Possible?
With DP	RMSProp	0.65	1.69	1.1	
	RMSProp	1.21	1.59	2.3	
	RMSProp	5.26	1.41	1.8	
	RMSProp	89	1.34	2.1	
	RMSProp	2×10^8	1.32	3.2	
	RMSProp	1×10^9	1.26	2.8	
	SGD	∞	2.11	3.6	
No DP	SGD	N/A	1.86	9.5	
	RMSProp	N/A	1.17	31.0	✓

Application: Pharmacogenetics



(a) Time in Therapeutic Range

(b) Mortality Events

(c) Stroke Events

(d) Bleeding Events

Goal: personalized dosing for warfarin

- see if genetic markers can be predicted from DP models
- small epsilon (< 1) does protect privacy but even moderate epsilon (< 5) leads to increased risk of fatality