Security and Privacy of ML

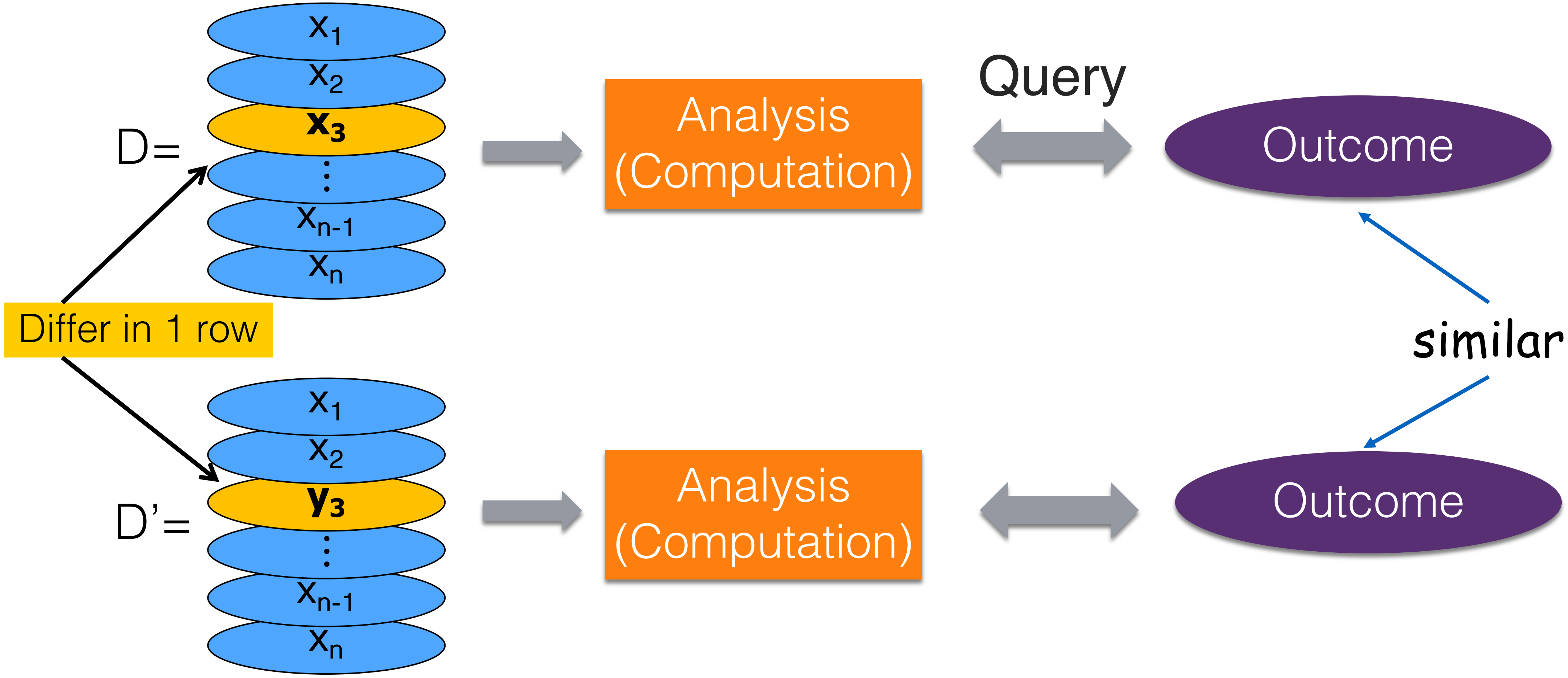
Differential Privacy (cont.)

Shang-Tse Chen

Department of Computer Science
& Information Engineering
National Taiwan University

Differential Privacy

[Dwork et al. '06]



(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,

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

Review: 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 $\varepsilon=\varepsilon_1+\ldots+\varepsilon_k$

Review: 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 $\varepsilon = \max\{\varepsilon_1,\ldots,\varepsilon_k\}$

Review: 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?

Naïve Algorithm

Return:

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

Error Analysis

Error:

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

Total Error:

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

Review: Sensitivity

- Let $f: \mathcal{D} \to \mathbb{R}^d$ 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 $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_i |x_i - y_i|$

Review: 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

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

Improving Utility of Algorithm 2

Compute:

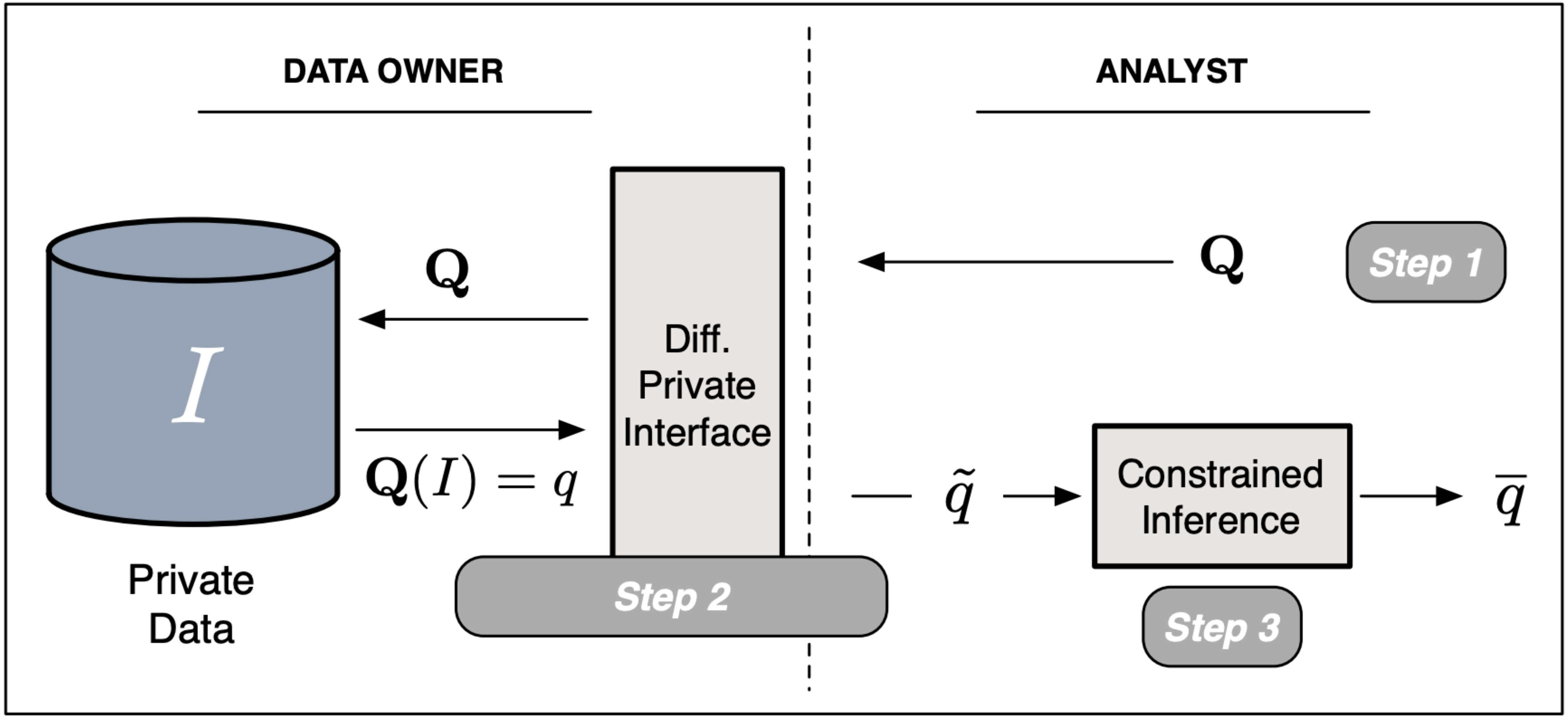
- $\widetilde{q_1}$ = # Males with BMI < 25 + Lap(1/ε)
- $\widetilde{q_2}$ = # Males with BMI > 25 + Lap(1/ε)

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



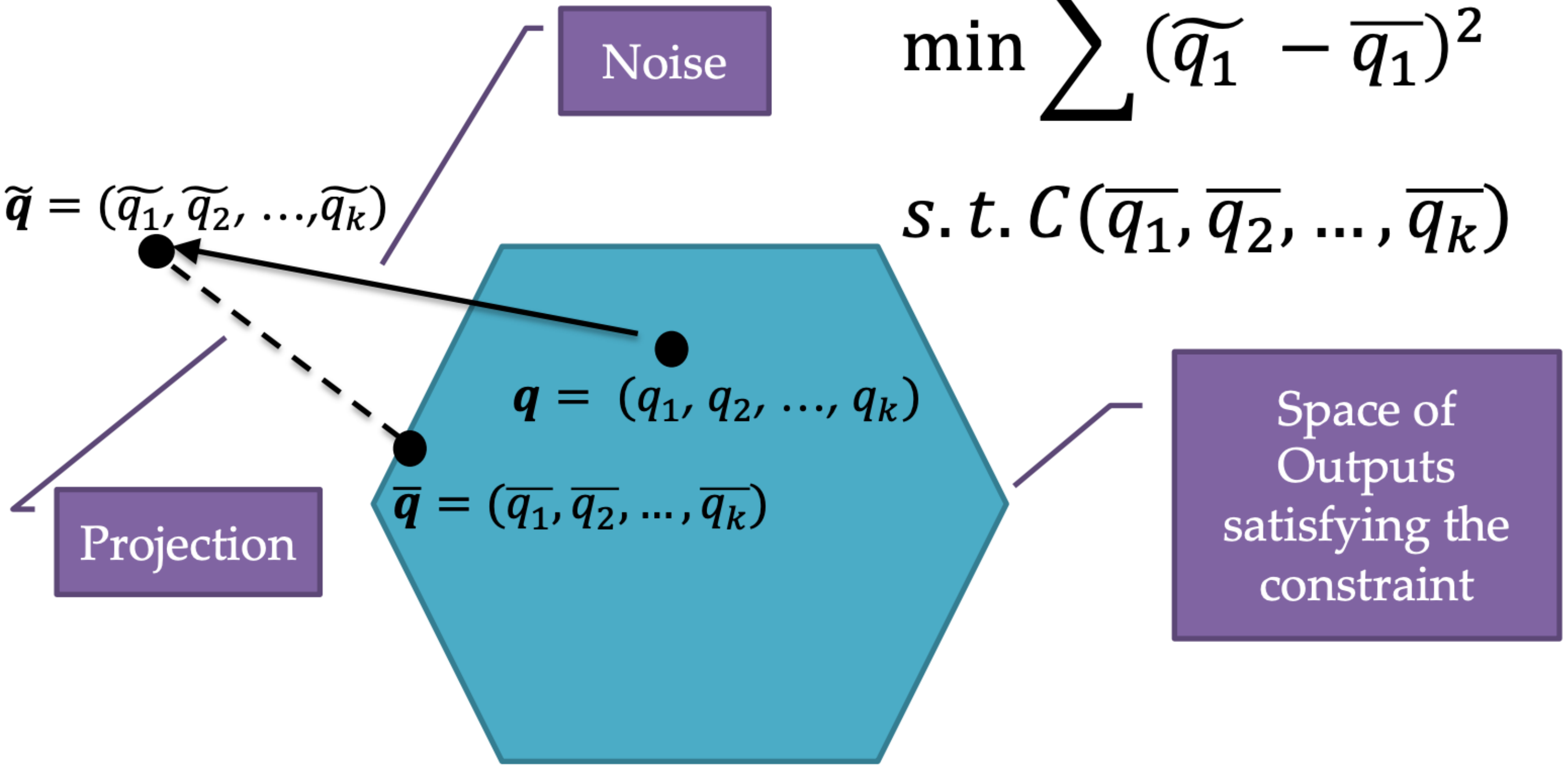
Least Squares Optimization

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

such that

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

Geometric Interpretation



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

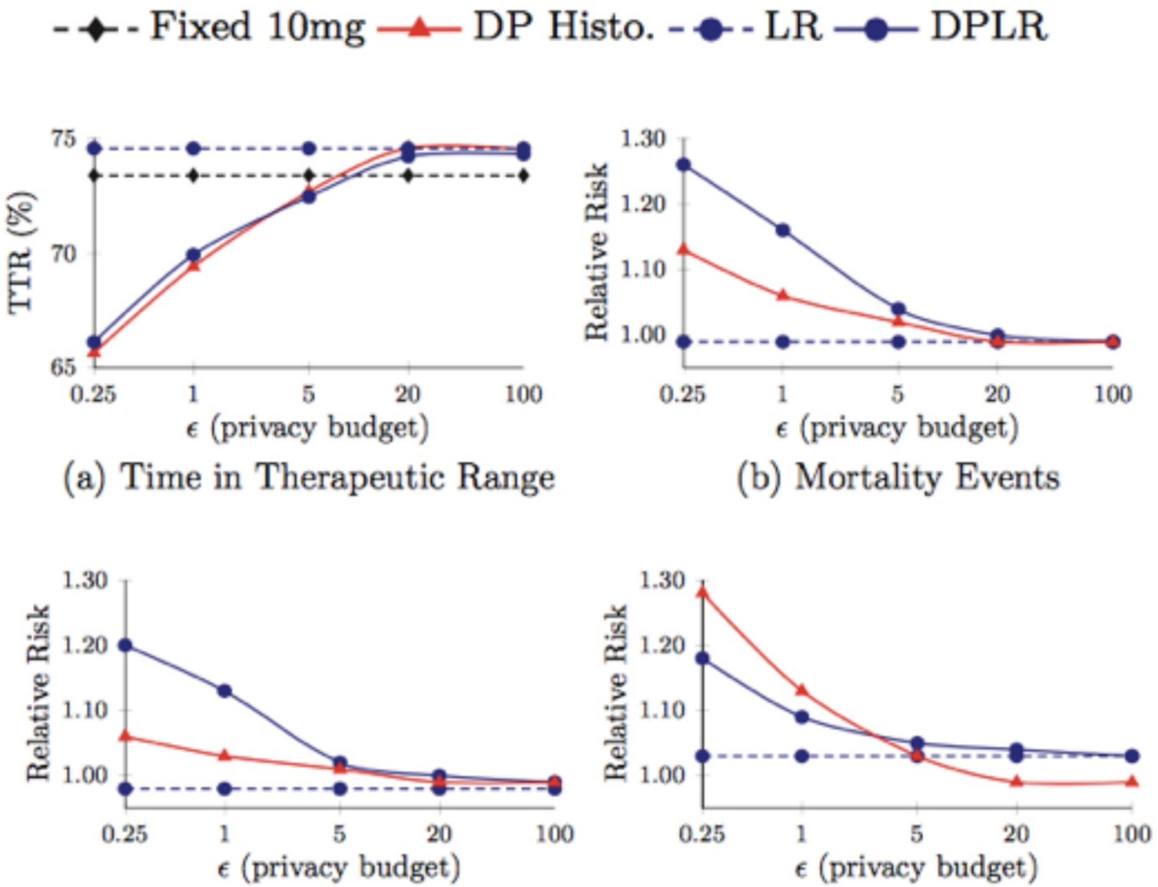
$$s.t.\, C(\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k})$$

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

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

Another Example: Range Queries

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

Queries:

- # people with height in [5′1″, 6′2″]
- # people with height in [2′0″, 4′0″]
- # people with height in [3′3″, 7′0″]
- …

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

Another Example: Range Queries

- Let $\{v_1, \ldots, v_k\}$ be the domain of an attribute
- Let $\{x_1, \ldots, x_k\}$ be the number of rows with values $v_1, \ldots, v_k$

- Range Query: $q_{ij} = x_i + x_{i+1} + \ldots + x_j$

- Goal: Answer all range queries

Strategy 1

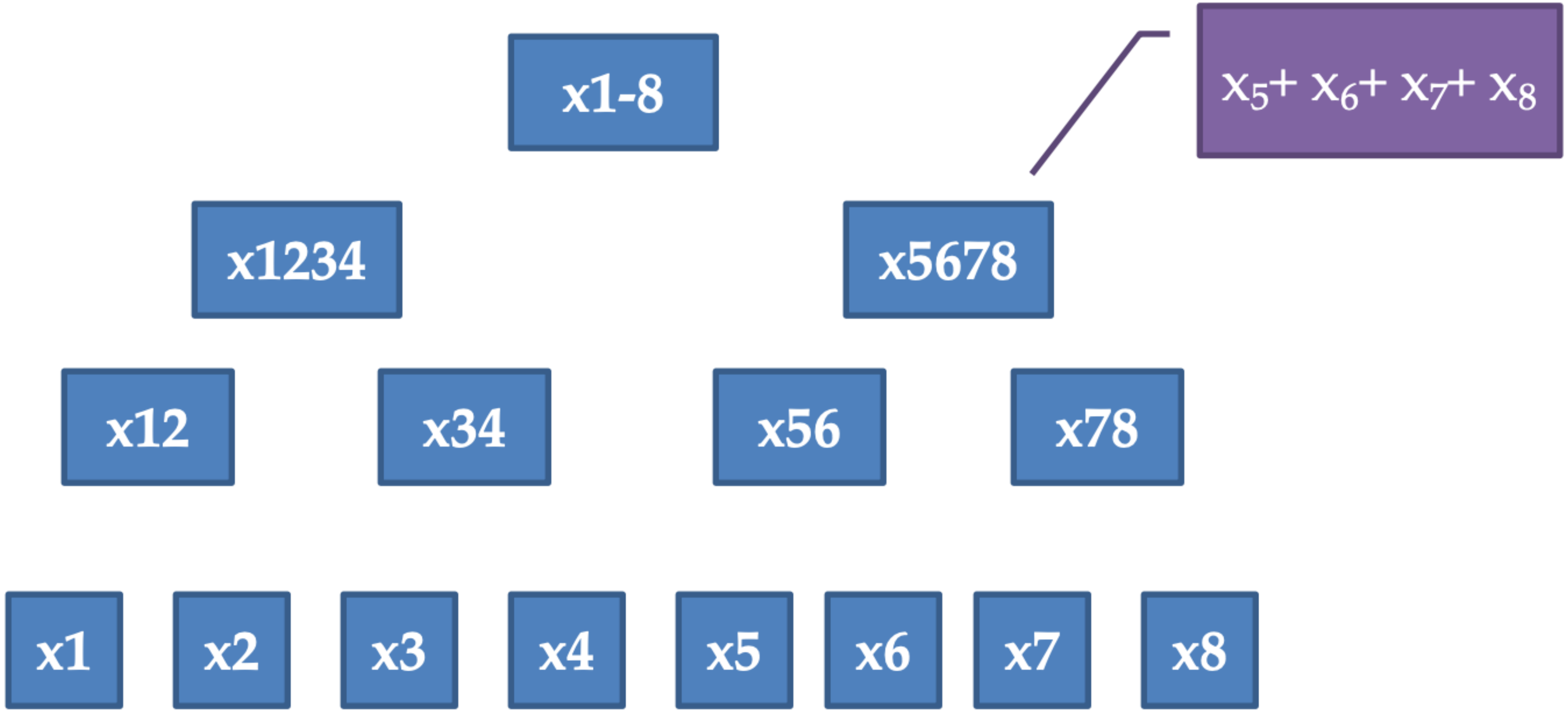
- Answer all range queries using Laplace mechanism
- Sensitivity: $O(k^2)$
- Total error: $O\left(\left(\frac{k^2}{\epsilon}\right)^2\right) = O(k^4/\epsilon^2)$

Strategy 2

- Estimate each individual x_i using Laplace mechanism
- Answer $q_{ij} = \widetilde{x_i} + \widetilde{x_{i+1}} + \cdots + \widetilde{x_j}$
- Error in each $\widetilde{x_i}$: $O(1/\epsilon^2)$
- Error in q_{1k}: $O(k/\epsilon^2)$
- Total Error: $O(k^3/\epsilon^2)$

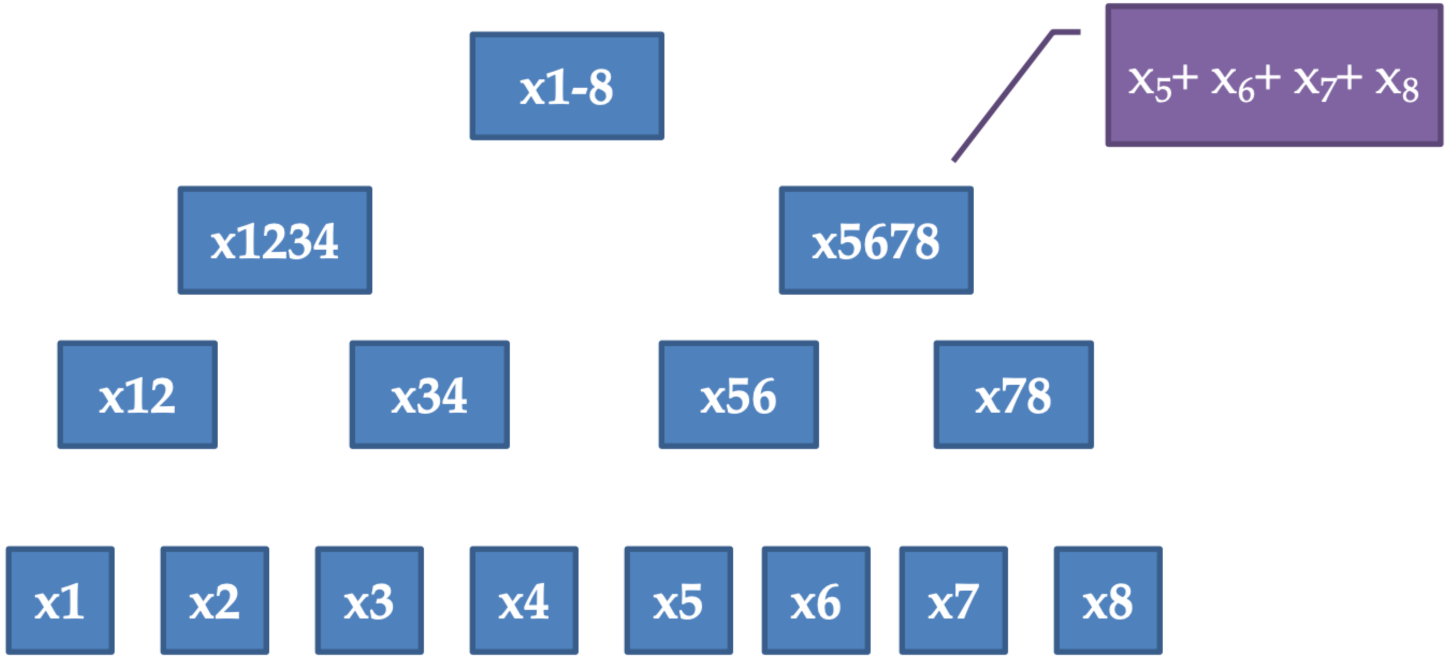
Strategy 3: Hierarchy

Estimate all the counts in the tree using Laplace mechanism



Strategy 3: Hierarchy

- Sensitivity: $O(\log k)$
- Every range query can be answered by summing up at most $O(\log k)$ nodes in the tree.



Strategy 3: Hierarchy

- Error in each node: $O\left((\log k)^2/\epsilon^2\right)$
- Max error on a range query: $O\left((\log k)^3/\epsilon^2\right)$
- Total Error: $O(k^2(\log k)^3/\epsilon^2)$
- Error can be further reduced by constrained inference
  - parent counts should not be smaller than child counts

General Strategy

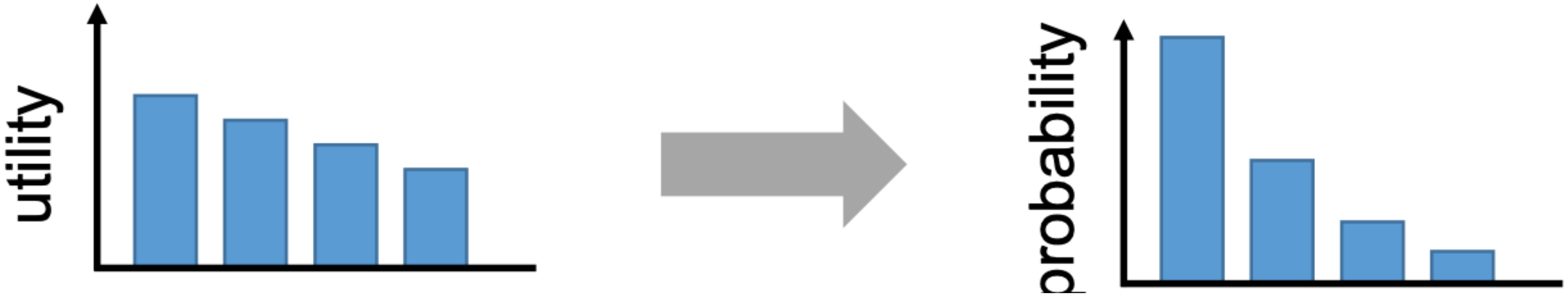
- Can think of nodes in the tree as coefficients
- Other algorithms use other transformations
  - Wavelets, Fourier coefficients

Exponential Mechanism

- Laplace/Gaussian mechanisms are for real-valued queries
- What if the queries output categorical values?
  - Choose the “best” item from a finite set of items

Exponential Mechanism

- Utility function $u(x, t)$ = "utility of t for dataset x"
- Goal: find $t \in T$ maximizing $u(x, t)$
- Sensitivity of u: $\Delta u = \max_{x, x', t} |u(x, t) - u(x', t)|$
- Output t with probability $\propto \exp\left(\frac{\epsilon}{2\Delta u} u(x, t)\right)$



Exponential Mechanism Preserves ϵ −DP

$$\frac{\Pr[\mathcal{M}_u(x) = t]}{\Pr[\mathcal{M}_u(x') = t]} = \frac{\dfrac{\exp\left(\frac{\epsilon}{2\Delta u} u(x,t)\right)}{\sum_{t' \in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x,t')\right)}}{\dfrac{\exp\left(\frac{\epsilon}{2\Delta u} u(x',t)\right)}{\sum_{t' \in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x',t')\right)}}$$

$$= \left(\frac{\exp\left(\frac{\epsilon}{2\Delta u} u(x,t)\right)}{\exp\left(\frac{\epsilon}{2\Delta u} u(x',t)\right)}\right) \cdot \frac{\sum_{t' \in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x',t')\right)}{\sum_{t' \in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x,t')\right)}$$

Exponential Mechanism Preserves ϵ −DP

$$= \exp\left(\frac{\epsilon\big(u(x,t) - u(x',t)\big)}{2\Delta u}\right) \cdot \frac{\sum_{t'\in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x',t')\right)}{\sum_{t'\in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x,t')\right)}$$

$$\leq \exp\left(\frac{\epsilon}{2}\right) \cdot \exp\left(\frac{\epsilon}{2}\right) \cdot \frac{\sum_{t'\in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x,t')\right)}{\sum_{t'\in T} \exp\left(\frac{\epsilon}{2\Delta u} u(x,t')\right)}$$

$$= \exp(\epsilon)$$

Accuracy of Exponential Mechanism

$$\mathrm{OPT}_u(x) = \max_{t \in T} u(x, t)$$

$$\Pr\left[u\big(\mathcal{M}_u(x)\big) \le \mathrm{OPT}_u(x) - \frac{2\Delta \mathrm{u}}{\epsilon}\left(\log\left(\frac{|T|}{|T_{\mathrm{OPT}}|}\right) + t\right)\right] \le e^{-t}$$

Pf:

$$\Pr\big[u\big(\mathcal{M}_u(x)\big) \le c\big] \le \frac{\Pr\big[u\big(\mathcal{M}_u(x)\big) \le c\big]}{\Pr\big[u\big(\mathcal{M}_u(x)\big) = \mathrm{OPT}_u(x)\big]}$$

$$\le \frac{|T| \exp\left(\frac{\epsilon c}{2\Delta u}\right)}{|T_{\mathrm{OPT}}| \exp\left(\frac{\epsilon \mathrm{OPT}_u(x)}{2\Delta u}\right)} = \frac{|T|}{|T_{\mathrm{OPT}}|} \exp\left(\frac{\epsilon\big(c - \mathrm{OPT}_u(x)\big)}{2\Delta u}\right)$$

Accuracy of Exponential Mechanism

rearrange

$$\Pr\left[\mathrm{OPT}_u(x) - u(\mathcal{M}_u(x)) \geq \frac{2\Delta \mathrm{u}}{\epsilon}\left(\log\left(\frac{|T|}{|T_{\mathrm{OPT}}|}\right) + t\right)\right] \leq e^{-t}$$

$t = \log\frac{1}{\beta}$

$$\Pr\left[\mathrm{OPT}_u(x) - u(\mathcal{M}_u(x)) \geq \frac{2\Delta \mathrm{u}}{\epsilon}\left(\log\left(\frac{|T|}{\beta|T_{\mathrm{OPT}}|}\right)\right)\right] \leq \beta$$

$|T_{\mathrm{OPT}}| \geq 1$

$$\Pr\left[\mathrm{OPT}_u(x) - u(\mathcal{M}_u(x)) \geq \frac{2\Delta \mathrm{u}}{\epsilon}\left(\log\left(\frac{|T|}{\beta}\right)\right)\right] \leq \beta$$

Accuracy of Exponential Mechanism

$$\Pr\left[\text{OPT}_u(x) - u(\mathcal{M}_u(x)) \geq \frac{2\Delta u}{\epsilon}\left(\log\left(\frac{|T|}{\beta}\right)\right)\right] \leq \beta$$

Compare with Laplace Mechanism

$$\Pr\left[|\mathcal{M}(x) - q(x)| \geq \frac{\Delta f}{\epsilon}\left(\log\left(\frac{1}{\beta}\right)\right)\right] \leq \beta$$

We have a dependency on the size of the output space

Exponential Mechanism

- Very general mechanism
- Unfortunately, when the output space is big:
  - Very costly to sample from it
  - Accuracy get worse

Private Data Release

Given a dataset $x \in \mathcal{X}^n$, a set of queries $Q = \{q_1, \dots, q_k\}$ and a target accuracy α, output a differentially private synthetic dataset $x' \in \mathcal{X}^m$ such that

$$\max_{q \in Q} |q(x) - q(x')| \leq \alpha$$

We focus on linear queries

$$q' : \mathcal{X} \to [0, 1] , \qquad q(x) = \frac{1}{n} \sum_{i=1}^{n} q'(x_i)$$

SmallDB Algorithm

1. Let $m = \frac{\log|Q|}{\alpha^2}$

2. Define utility function $u: \mathcal{X}^n \times \mathcal{X}^m \to \mathbb{R}$ as
$$u(x, y) = -\max_{q \in Q} |q(x) - q(y)|$$

3. Run exponential mechanism with u

Case Study: Linear Classifier

Empirical Risk Minimization (ERM):

$$\frac{1}{2}\lambda\|w\|^2 \quad + \quad \frac{1}{n}\sum_{i=1}^{n} L(y_i w^T x_i)$$

Regularizer (Model Complexity)

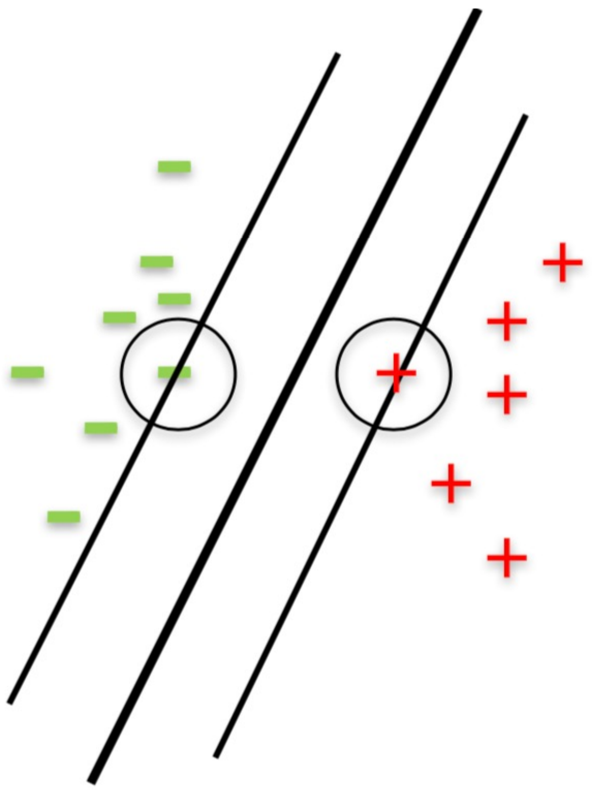
Risk (Training Error)

L = Logistic Loss ➡ **Logistic Regression**

L = Hinge Loss ➡ **SVM**

Why ERM Is Not Private For SVM?

- SVM solution is a combination of support vectors. If one support vector moves, solution changes



First Attempt: Output Perturbation

$$\tilde{f}(D) = f(D) + \textcolor{red}{noise} =$$

$$\left[argmin_{\omega} \ \frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} l(\omega, (x_i, y_i)) \right] + \textcolor{red}{noise}$$

Theorem: [CMS11] If $\parallel x_i \parallel \leq 1$ and l is 1-Lipschitz, then for any D, D' with $\text{dist}(D, D') = 1$,

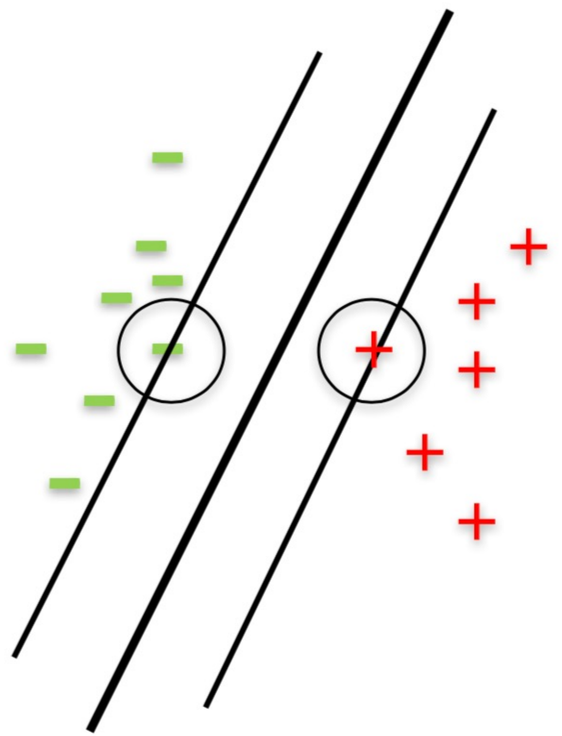
$$||f(D) - f(D')||_2 \leq \frac{2}{\lambda n} \quad (L_2\text{-sensitivity})$$

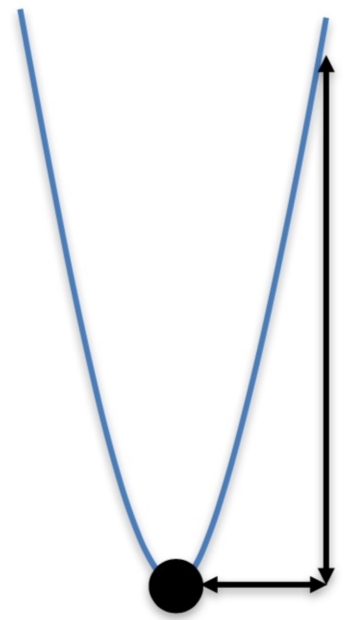
First Attempt: Output Perturbation

$$\tilde{f}(D) = f(D) + noise = \left[argmin_{\omega} \frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} l(\omega, (x_i, y_i)) \right] + noise$$

noise: $z \propto e^{-\frac{2}{\lambda n \epsilon}\|z\|_2}$

Property of Real Data





Optimization surface is very steep in some direction
→ High loss if perturbed in those directions

Better Solution: Objective Perturbation

[Chaudhuri et al. JMLR '11]

- **Insight:** Perturb optimization surface and then **optimize**

$$\tilde{f}(D) = argmin_{\omega} \left[\frac{1}{2} \lambda \parallel \omega \parallel^2 + \frac{1}{n} \sum_{i=1}^{n} l(\omega, (x_i, y_i)) + noise \right]$$

- **Main idea:** *add noise as part of the computation*:
  - Regularization already changes the objective to protects against overfitting.
  - Change the objective a little bit more to protect privacy.

Better Solution: Objective Perturbation

$$\underset{w}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i w^\top x_i) + \frac{1}{2} \lambda \|w\|^2 + \text{noise} \right\}$$

- Main idea: add noise as part of the computation
  - Regularization already changes the objective
  - Change the objective a little bit more to protect privacy

Better Solution: Objective Perturbation

$$\underset{w}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i w^\top x_i) + \frac{1}{2}\lambda \|w\|^2 + \text{noise} \right\}$$

- *noise* drawn from
  - Magnitude: drawn from $\Gamma(\mathrm{d}, \frac{1}{\epsilon})$
  - Direction: uniform at random

- **Theorem:** If l is convex and double-differentiable with $|l'(\mathrm{z})| \leq 1$, $|l''(\mathrm{z})| \leq c$ then Algorithm satisfy $\epsilon + 2\log\left(1 + \frac{c}{n\lambda}\right)$-DP. [CMS11]

Stochastic Gradient Descent (SGD)

- Initial ω_0
- Incremental gradient update for $t = 0 \dots T - 1$
  - Take a random example $(x_t, y_t) \in D$
  - Update $\omega_{t+1} = \omega_t - \eta_t(\nabla l(\omega_t, (x_t, y_t)))$
    - η_t is the step size

SGD with Differential Privacy

[Abadi et al. CCS'16]

- Initial ω_0
- Incremental gradient update for $t = 0 \dots T-1$
  - Take a random example $(x_t, y_t) \in D$
  - Update $\omega_{t+1} = \omega_t - \eta_t(\nabla l(\omega_t, (x_t, y_t)) + noise)$
    - η_t is the step size

Naïve Analysis

1. Choose $\sigma = \frac{\sqrt{2 \log 1/\delta}}{\varepsilon}$	$= 4$
2. Each step is (ε, δ)-DP	$(1.2, 10^{-5})$-DP
3. Number of steps T	10,000
4. Composition: $(T\varepsilon, T\delta)$-DP	$(12{,}000, .1)$-DP

Advanced Composition Theorem

Lemma 2.3 (basic composition). *If* $\mathcal{M}_1, \ldots, \mathcal{M}_k$ *are each* (ε, δ)*-differentially private, then* $\mathcal{M}$ *is* $(k\varepsilon, k\delta)$*-differentially private.*

However, if we are willing to tolerate an increase in the δ term, the privacy parameter ε only needs to degrade proportionally to $\sqrt{k}$:

Lemma 2.4 (advanced composition [42]). *If* $\mathcal{M}_1, \ldots, \mathcal{M}_k$ *are each* (ε, δ)*-differentially private and* $k < 1/\varepsilon^2$*, then for all* $\delta' > 0$*,* $\mathcal{M}$ *is* $\left(O(\sqrt{k \log(1/\delta')}) \cdot \varepsilon, k\delta + \delta'\right)$*-differentially private.*

Analysis With Advanced Composition

1. Choose $\sigma = \frac{\sqrt{2 \log 1/\delta}}{\varepsilon}$	$= 4$
2. Each step is (ε, δ)-DP	$(1.2, 10^{-5})$-DP
3. Number of steps T	10,000
4. Strong comp: $(\varepsilon\sqrt{T \log 1/\delta}, T\delta)$-DP	$(360, .1)$-DP

Amplification by Sampling

1. Choose $\sigma = \frac{\sqrt{2 \log 1/\delta}}{\varepsilon}$	$= 4$
2. Each batch is q fraction of data	1%
3. Each step is $(2q\varepsilon, q\delta)$-DP	$(.024, 10^{-7})$-DP
4. Number of steps T	10,000
5. Strong comp: $(2q\varepsilon\sqrt{T \log 1/\delta}, qT\delta)$-DP	(10, .001)-DP

Moments Accountant

1. Choose $\sigma = \frac{\sqrt{2 \log 1/\delta}}{\varepsilon}$ $= 4$
2. Each batch is q fraction of data 1%
3. Keeping track of privacy loss's **moments**
4. Number of steps T 10,000
5. Moments: $(2q\varepsilon\sqrt{T}, \delta)$-DP $(1.25, 10^{-5})$-DP

Tensorflow Integration

- https://github.com/tensorflow/privacy
- optimizer = tf.train.GradientDescentOptimizer()
- dp_optimizer_class = dp_optimizer.make_optimizer_class(tf.train.GradientDescentOptimizer)
- optimizer = dp_optimizer_class()

PATE: Private Aggregation of Teacher Ensemble

[Papernot et al. ICLR'17]

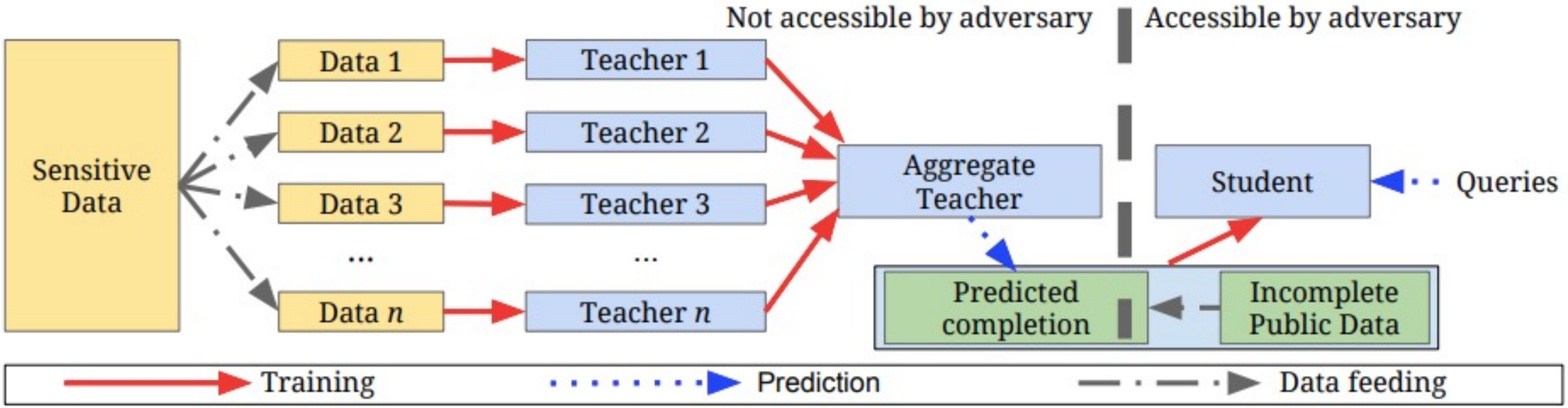
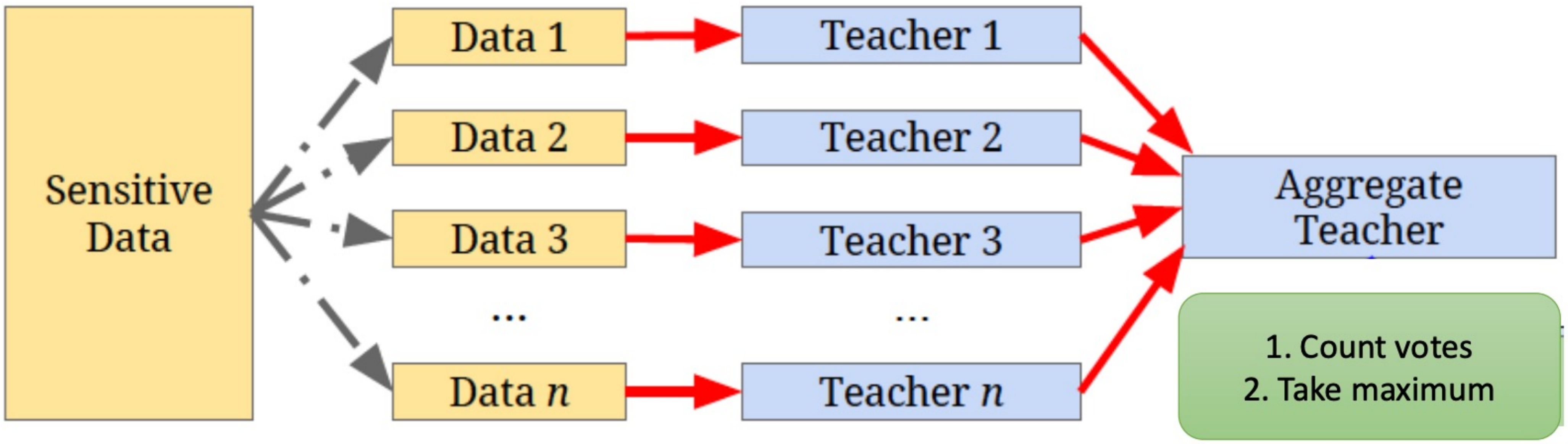


Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

PATE: Private Aggregation of Teacher Ensemble

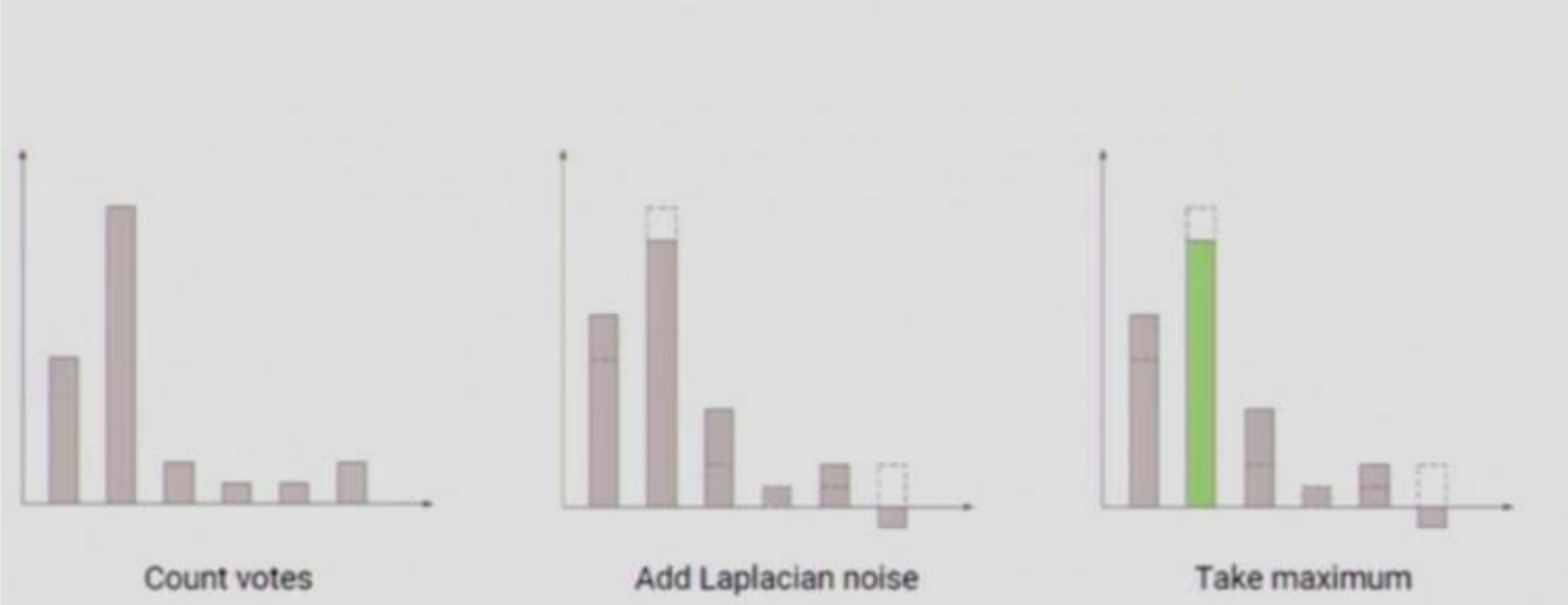
[Papernot et al. ICLR'17]



Intuitive privacy analysis:

- If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.
- If two classes have close vote counts, the disagreement may reveal private information

Noisy Aggregation

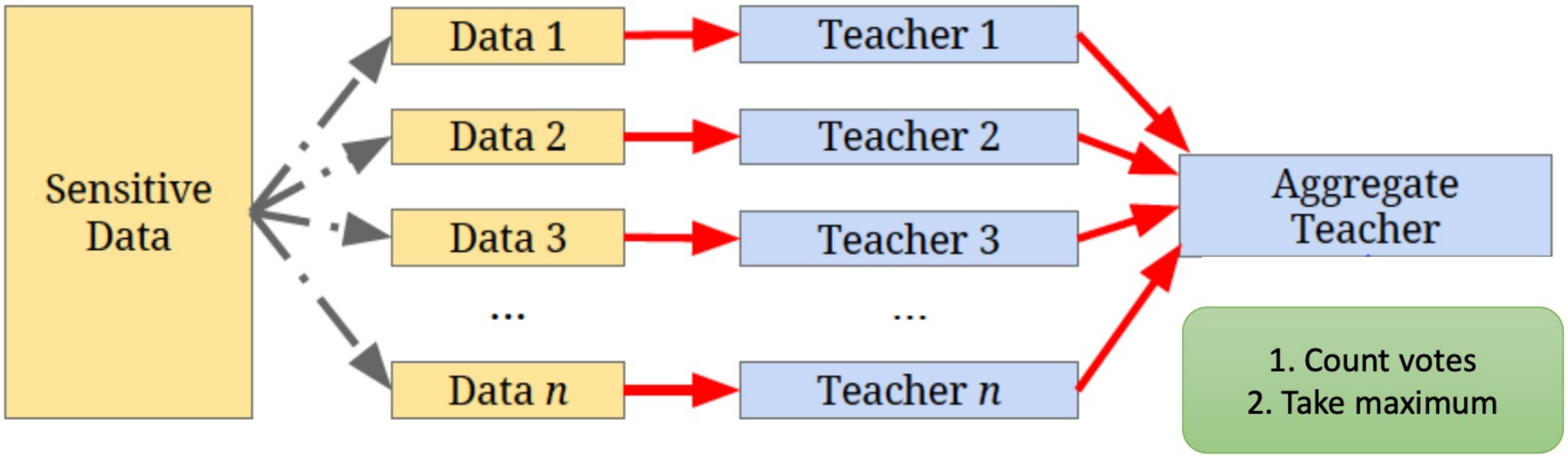


$$n_j(\vec{x}) = |\{i : i \in [n], f_i(\vec{x}) = j\}|$$

$$Lap\left(\frac{1}{\gamma}\right)$$

$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) + Lap\left(\frac{1}{\gamma}\right) \right\}$$

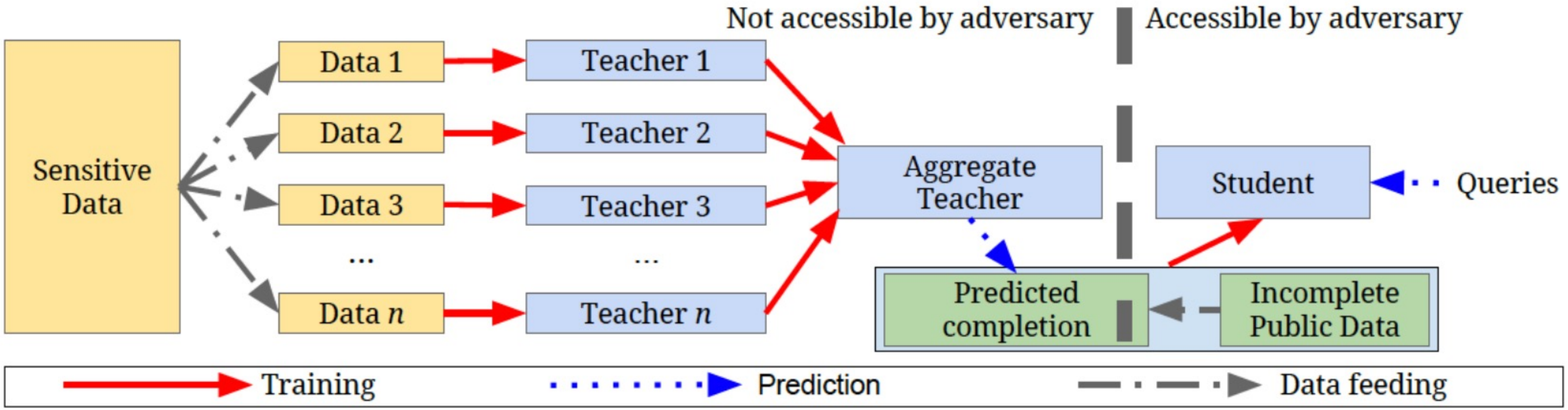
Why Not Just Use the Teacher Model?



The aggregated teacher violates the threat model:

- **Each prediction increases total privacy loss.**
  privacy budgets create a tension between the accuracy and number of predictions
- **Inspection of internals may reveal private data.**
  Privacy guarantees should hold in the face of white-box adversaries

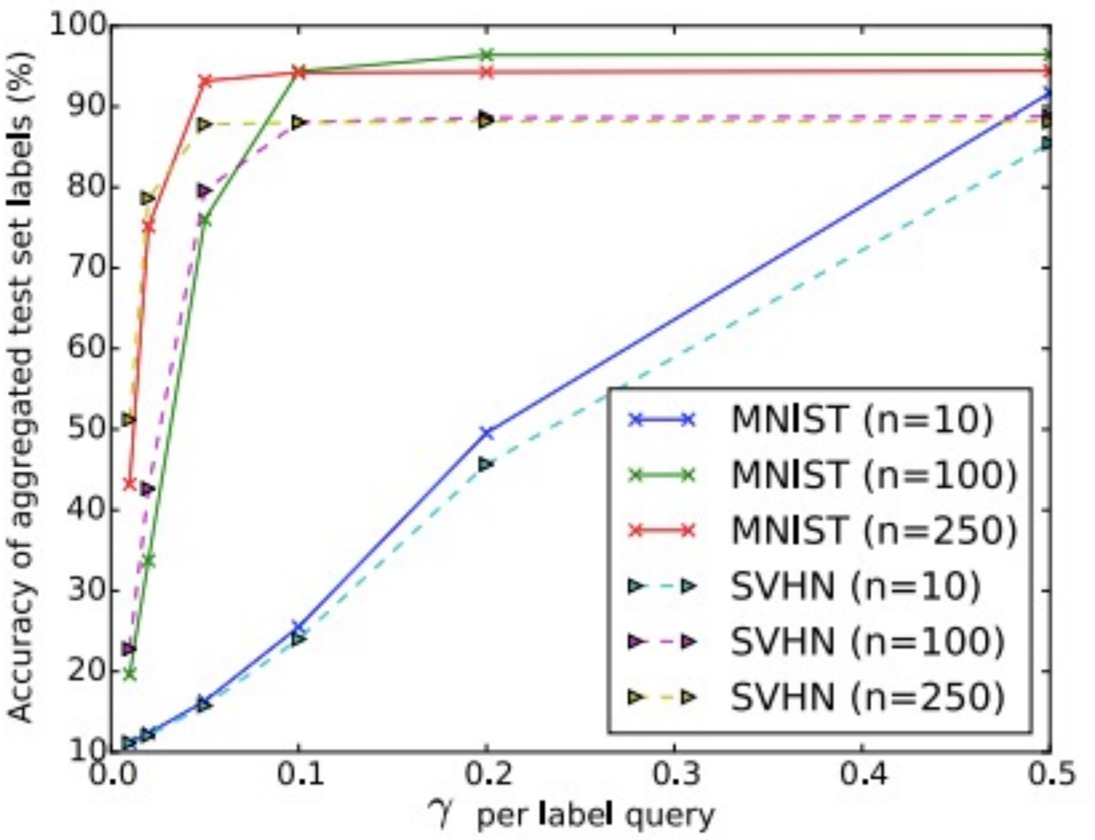
Benefits of Using the Student Model

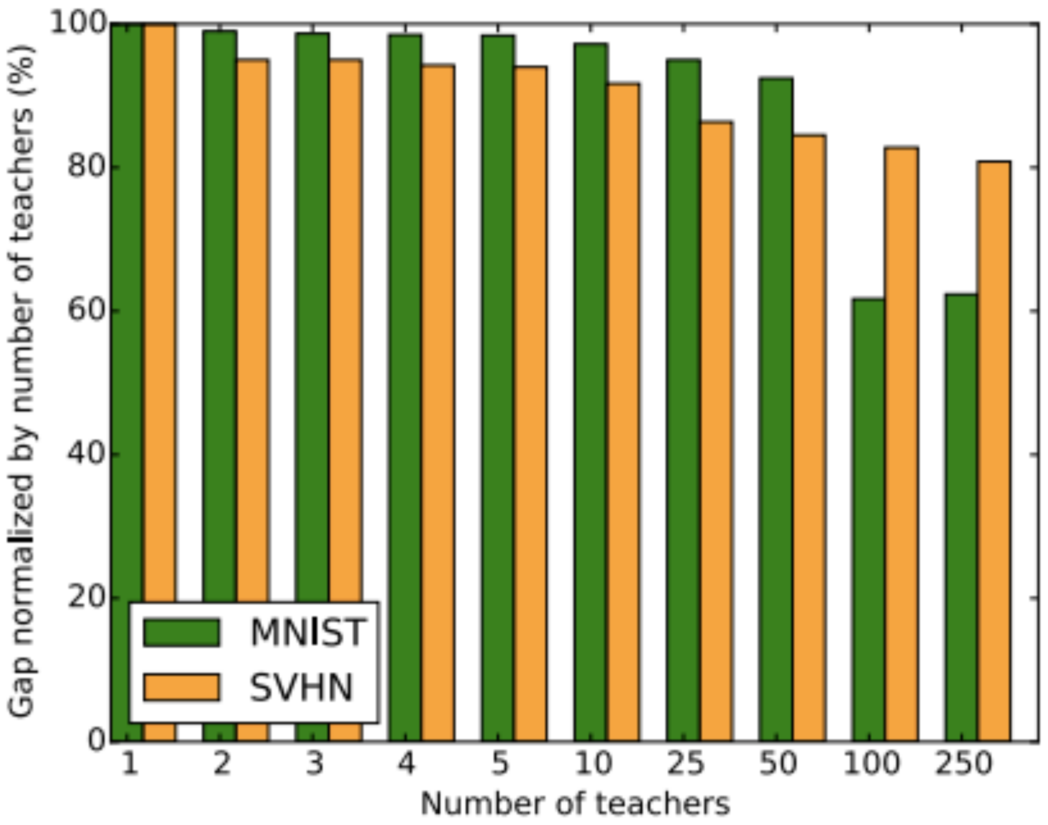


Privacy Analysis:

- Privacy loss is fixed after the student model is done training.
- Even if white-box adversary can inspect the model parameters, the information can be revealed from student model is unlabeled public data and labels from aggregate teacher which is protected with privacy

Experiment Results



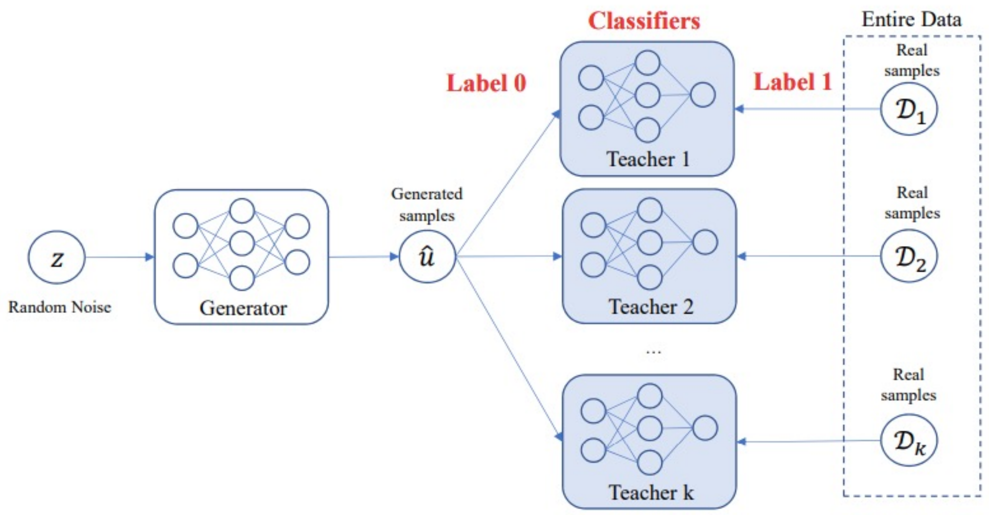


Gap increases as number of teachers increases -> Less Privacy Loss, but there will be acc. tradeoffs

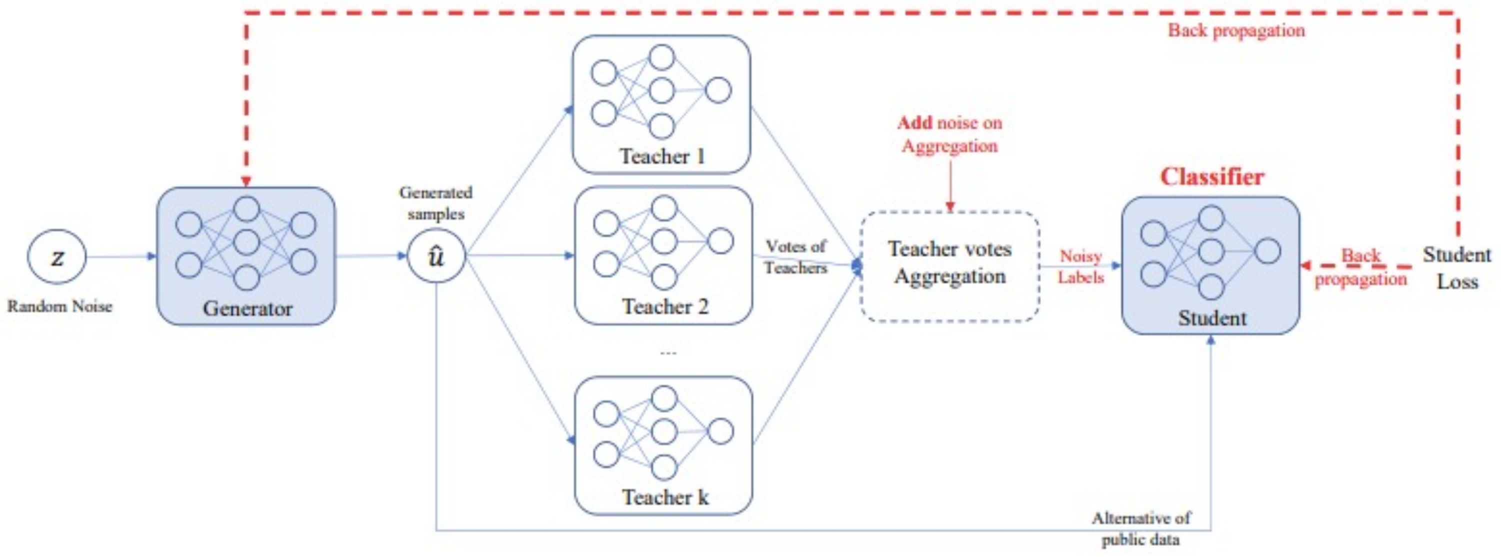
PATE-GAN

[Jordan et al. ICLR'19]

Training Procedure for Teacher Discriminators



Training Procedure for Generator and Student



Visual Results

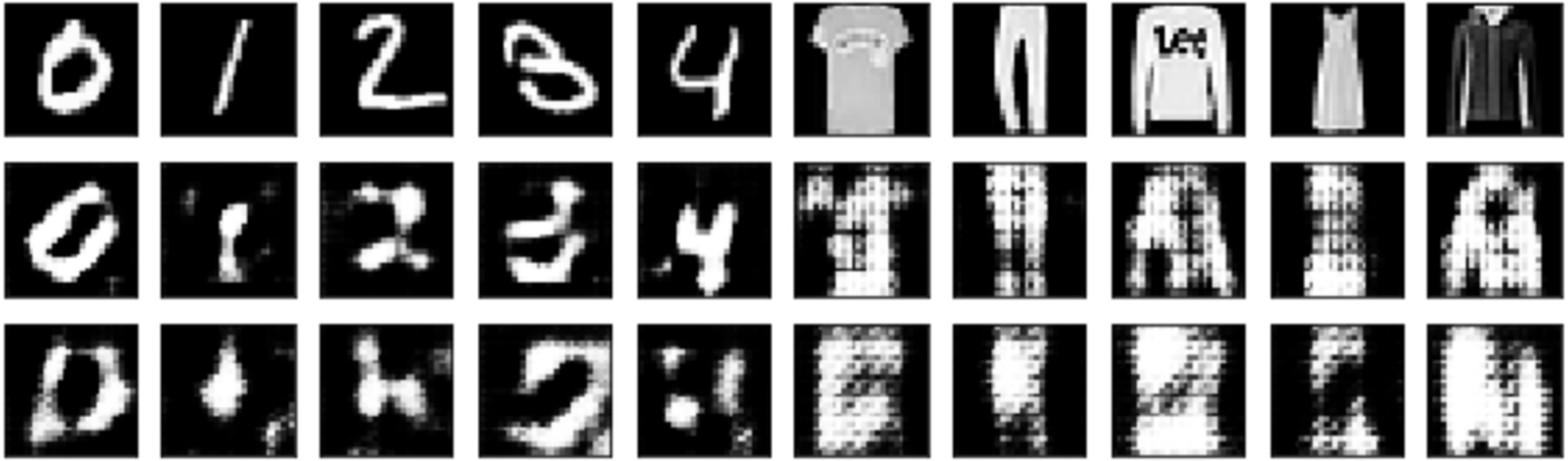


Figure 2. **Visualization of generated instances by G-PATE.** Row 1 (real image), row 2 ($\varepsilon = 10, \delta = 10^{-5}$) and row 3 ($\varepsilon = 1, \delta = 10^{-5}$) each presents one image from each class (the left 5 columns are MNIST images, and the right 5 columns are Fashion-MNIST images). When $\varepsilon = 1$, G-PATE does not generate high-quality images. However, it preserves partial features in the training images, so the synthetic images are useful to preserve data utility which can be seen from our quantitative results.

Summary

- Differential privacy: a systematic way to guarantee privacy
- Many useful tools for building strong algorithms
- Many opportunities in adapting traditional data-oriented tasks and algorithms to the privacy-preserving setting