Lecture 8: Data Privacy

Nov 2

Prof. Nicolas Papernot

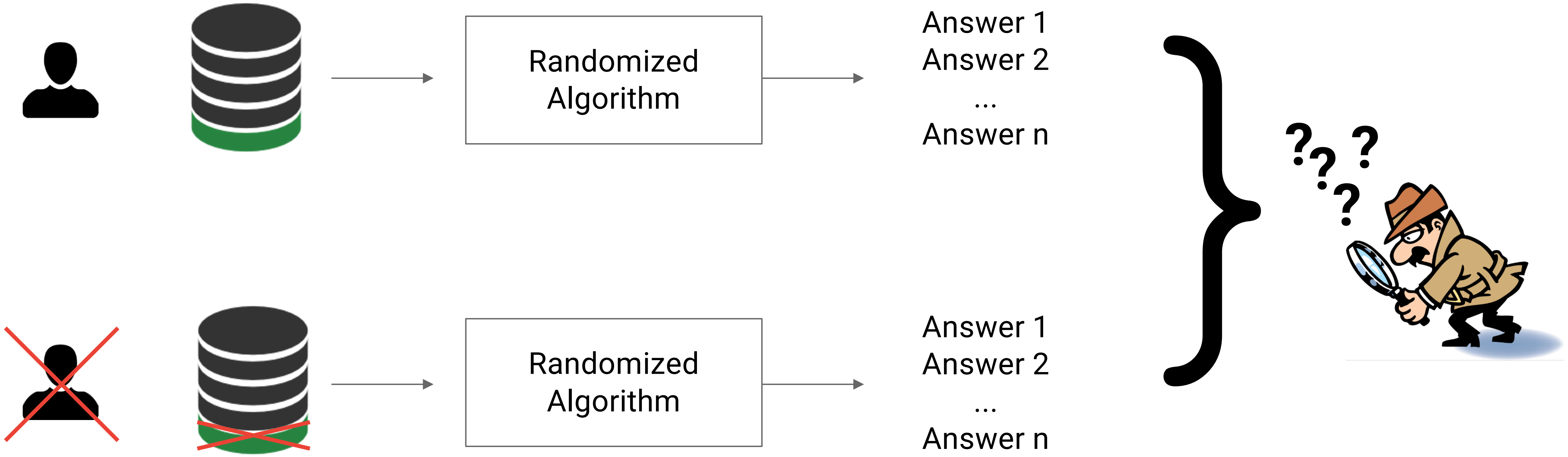
@NicolasPapernot

Limitations of previous definitions: the case of k-anonymity

- Each record must be indistinguishable from k-1 other records
  - Suppression -> replace features by wildcards
  - Generalization -> change age from number of years to bins

- Attacks:
  - Often use background knowledge
  - E.g., link attributes in private database and attributes from another database

What is a differentially private algorithm?



Differential privacy

S: any output

$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S]$$

d(d, d') = 1

Probability (algorithm M is randomized)

$Pr[M(d') \in S]$ $\quad$ $Pr[M(d) \in S]$ $\quad$ $Pr[M(d') \in S]$

Why DP improves upon previous definitions

- Made assumptions about adversaries:
  - Value of k in k-anonymity depends on capabilities of adversary
  - Instead DP guarantee does not depend on:
    - What adversary knows (capability)
    - What adversary wants (goal)
- Precise metric for privacy leakage (bound on epsilon)

$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S]$$

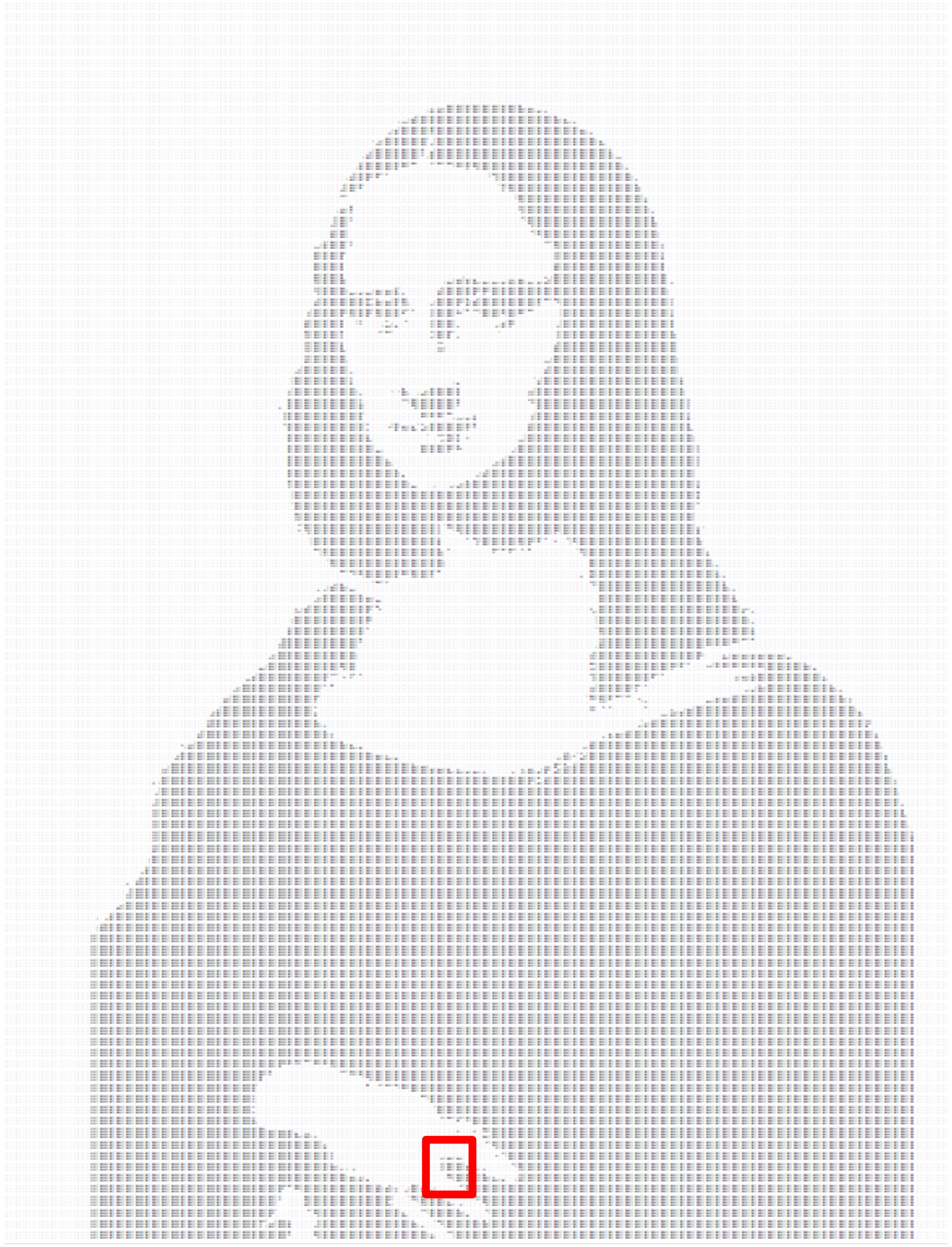
- Robust to composition
  - Algorithm M1 has eps DP
  - Algorithm M2 has eps DP
  - Algorithms M1 and M2 have 2eps DP
- Group guarantees

What does that mean for a user?

- Pessimistic perspective: privacy is already lost
- DP moves forward by estimating cost of participating in a dataset

-> *differential* privacy

A Metaphor For Private Learning



Slides adapted from Ulfar Erlingsson

An Individual's Training Data

```
...................................
..................................N
............MMMMMMMMM..............
............MMMMMMMMMM.............
............MMMMMMMMMMMMM..........
............MMMMMMMMMMMMMMMM..M....
..............MMMMMMMMMMMMMMMMM....
..............M.MMMMMMMMMMMMMMMMM..
.................MMMMMMMMMMMMMMMMMN
MMM.................MMMMMMMMMMMMMMN
MMMMMMM..M...........MMMMMMMMMMMMMN
```

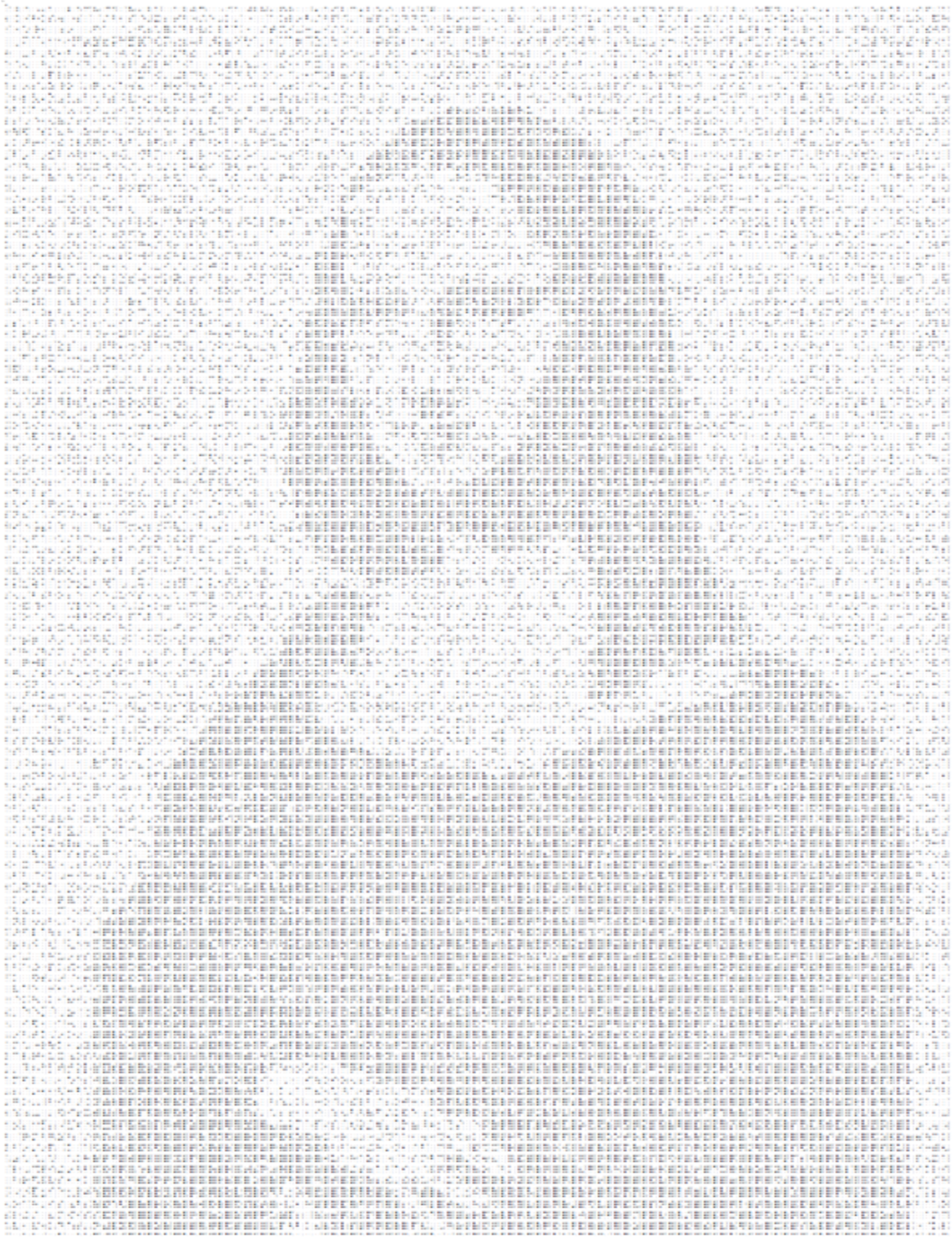
An Individual's Training Data

Each bit is flipped with probability 50%

```
.......M........MM.M.......MMM.M..
......................MM...MMMM...
....M..MM.MM..MMM.M.MM.M...M..MM..
.MM......MMM....MMMMMMMMM...M...MM
..M....M........MM..MMMMMMM...M...
M.......M..MM.MMMMMMMMMMMMMMM....M
.....M.....M.M.M.MMMMMM...MMMMM...
...M.....M.MM.M.MM..M..M..MM.MMMMM
M...M.M.....M.M..M..MMM.MMMMM.MMMM
.MMM.M....M.M.M........MMMMMMMMM.M
```


Slides adapted from Ulfar Erlingsson

Big Picture Remains!



Are you a communist?

Algorithm:

1. Flip a first coin
2. If:
   a. First coin was heads -> return correct answer
   b. First coin was tails, flip second coin:
      a. report true if heads
      b. report false if tails

Plausible deniability

Is it still useful? What did you learn?

Result of survey

- If person is communist:
  - With probability __ they will respond correctly True
  - With probability __ they will respond with the second coin flip
    - With probability __ the second coin flip will return True
    - With probability __ the second coin flip will return False
- Probability to say True __
- Probability to say False __

- Repeat exercise for a non-communist

How private is our survey?

- Eps is such that 0.75 = e^eps * 0.25
- Eps = ln(3) ≈ 1.1

- If we changed probability of first coin flip to 75% saying truth:
  - Eps is now such that 0.75 + 0.25*0.5 = 0.875 = e^eps * 0.125
  - Eps = ln(7) ≈ 1.95

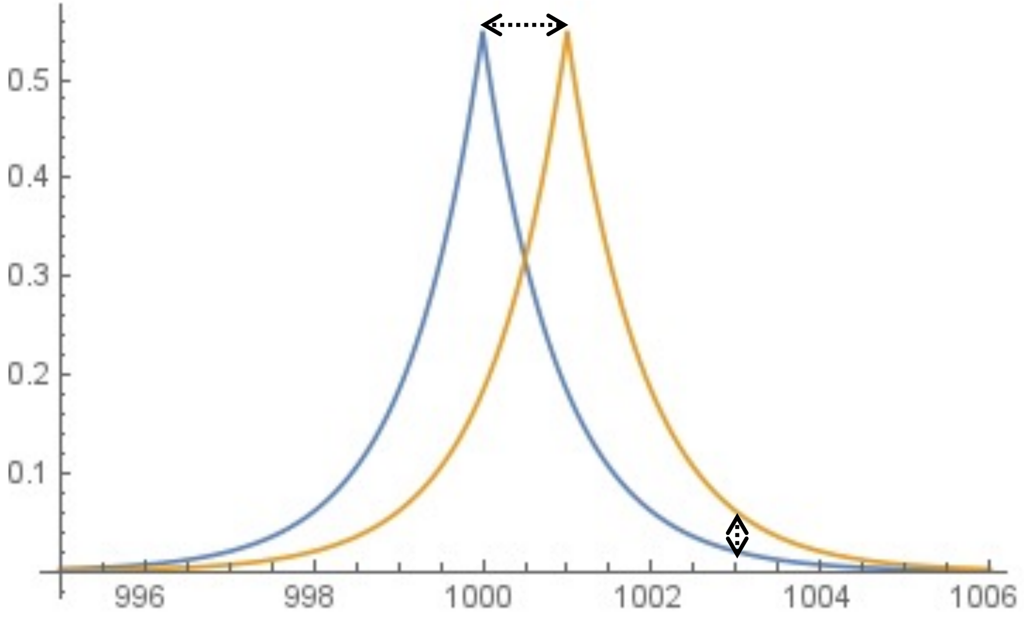
How to implement the survey in practice?

- Assume 10,000 participants
  - 3,000 say they are communist
  - 7,000 say they are not communist
- 50% answers are random so we remove 5,000/2 from each answer pool:
  - 500 are communist
  - 4,500 are not communist

Another example: a privacy-preserving count query

Query: how many users have green eyes?
Adversarial knowledge: all eye colors besides one person's

Real answer K=1000	Real answer K=1001
Respond 1000+Laplace(1/eps)	Respond 1001+Laplace(1/eps)
Output 1003	



Probability of K=1001 is e^eps more likely than K=1000

Example + illustration from desfontain.es

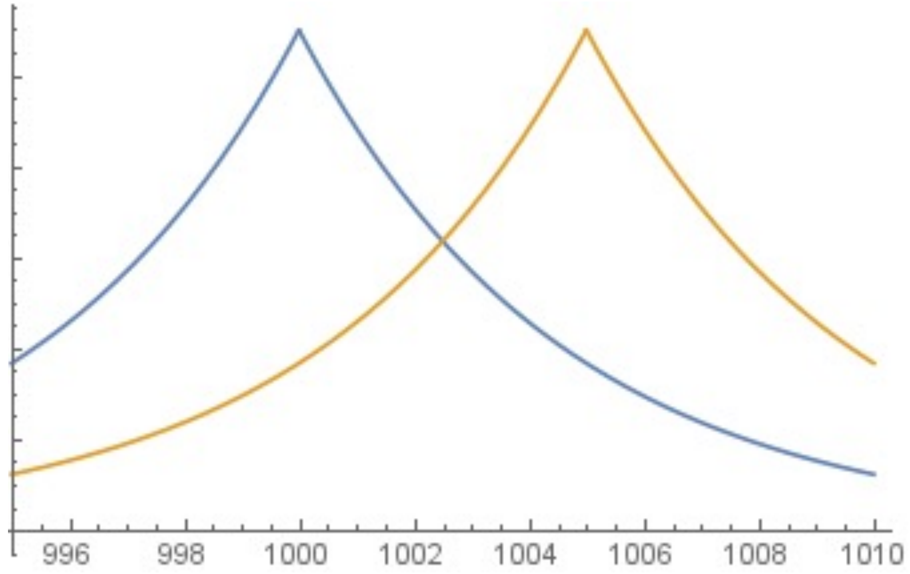
Another example: a privacy-preserving count query

Query: average rating (between 0 and 5) submitted by users

Average is same than sum / number of users

Adversarial knowledge: all ratings besides one person's sum up to 1000

Real answer K=1000 (user votes 0)	Real answer K=1005 (user votes 5)
Respond 1000+Laplace(5/eps)	Respond 1005+Laplace(5/eps)



One final consideration

- What if a user can contribute an outlier value?
  - Compute average of salaries where one individual has a very large salary
- Can pre-process data to remove outliers:
  - Good for privacy + accuracy when computing an average
  - Omission of data points creates new privacy issues
- Can relax definition of differential privacy:

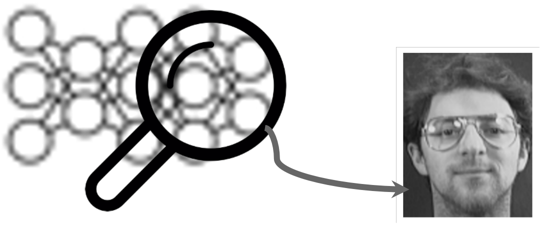
$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$$

For more details: https://www.youtube.com/watch?v=oQzaA5KG3pM (watch first 5 minutes)

Types of adversaries and our threat model

Model querying (**black-box adversary**)

Shokri et al. (2016) *Membership Inference Attacks*



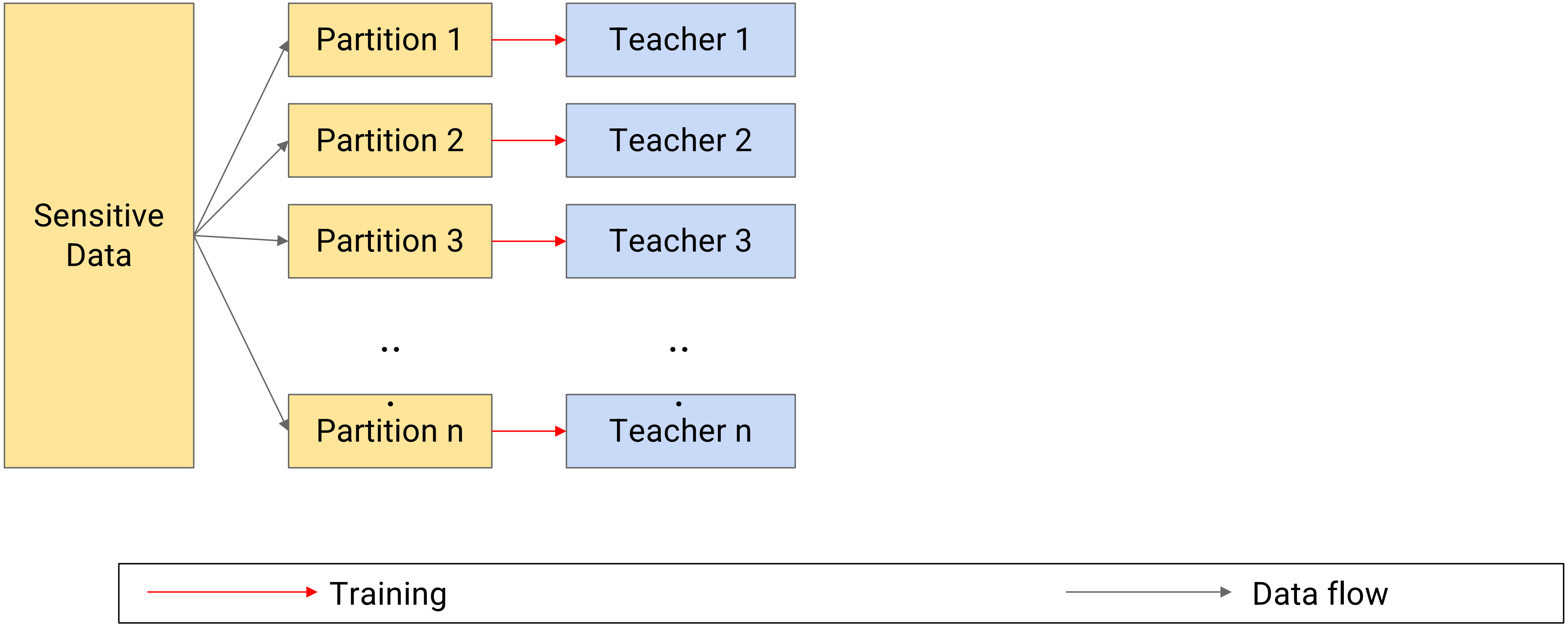
Model inspection (**white-box adversary**)

TBD

In our work, the threat model assumes:

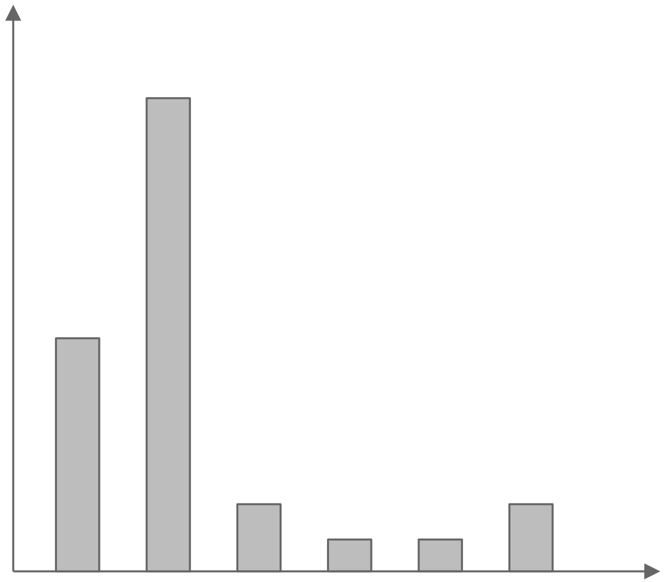
- Adversary can make a potentially unbounded number of queries
- Adversary has access to model internals

Private Aggregation of Teacher Ensembles (PATE)



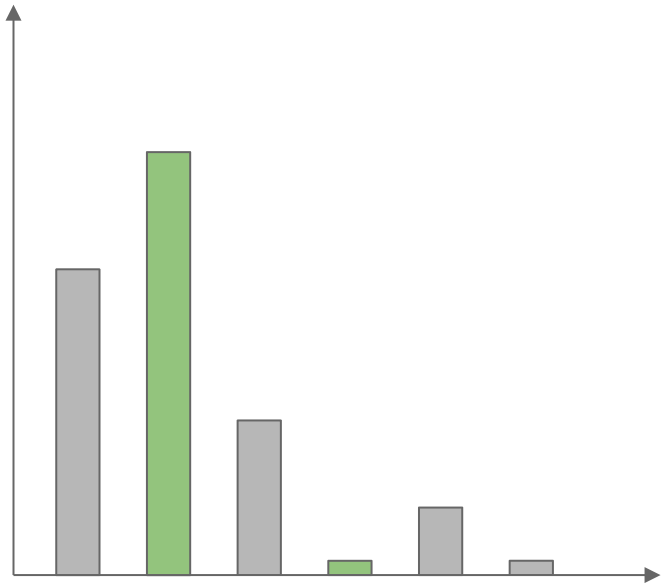
Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data [ICLR 2017 best paper]
Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar

Aggregation



Count votes

$$n_j(\vec{x}) = |\{i : i \in 1..n, f_i(\vec{x}) = j\}|$$

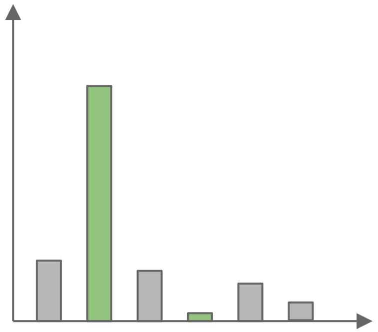


Take maximum

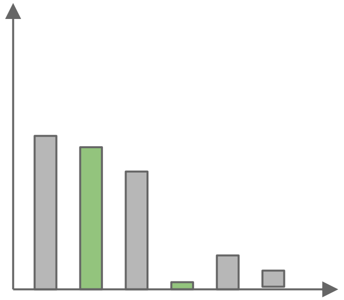
$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) \right\}$$

Intuitive privacy analysis

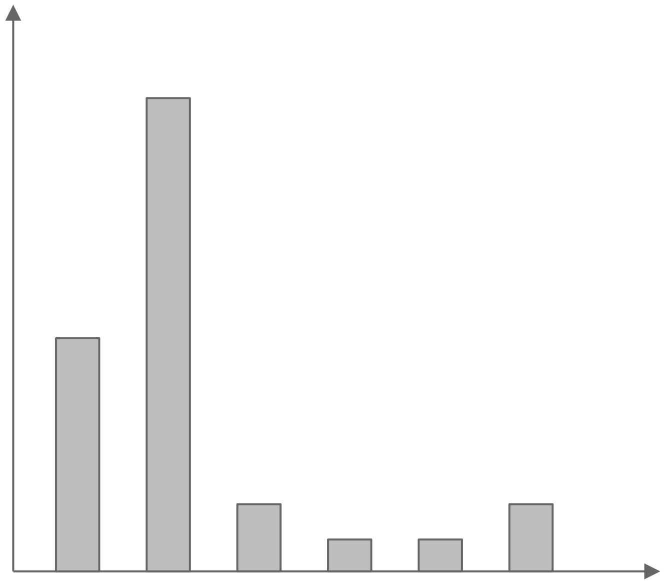
If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.



If two classes have close vote counts, the disagreement may reveal private information.

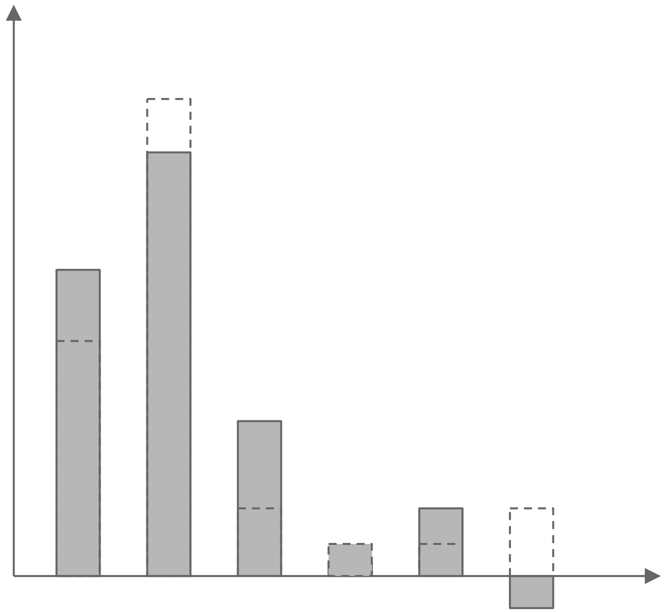


Noisy aggregation



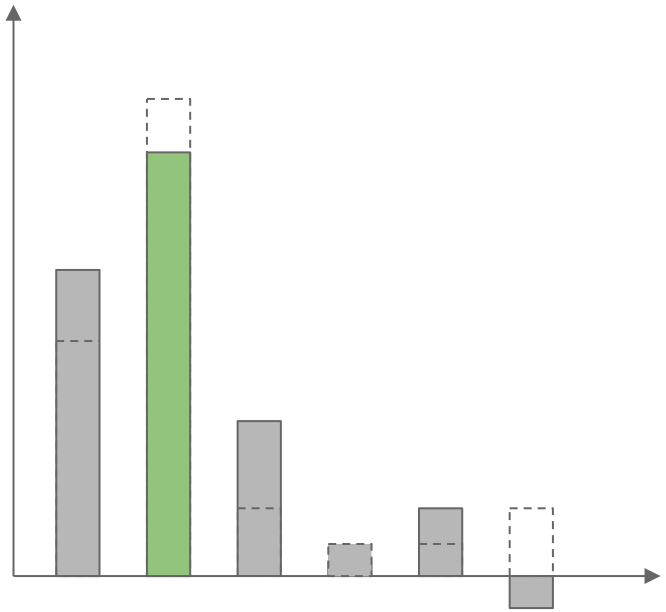
Count
votes

$$n_j(\vec{x}) = |\{i : i \in 1..n, f_i(\vec{x}) = j\}|$$



Add Laplacian

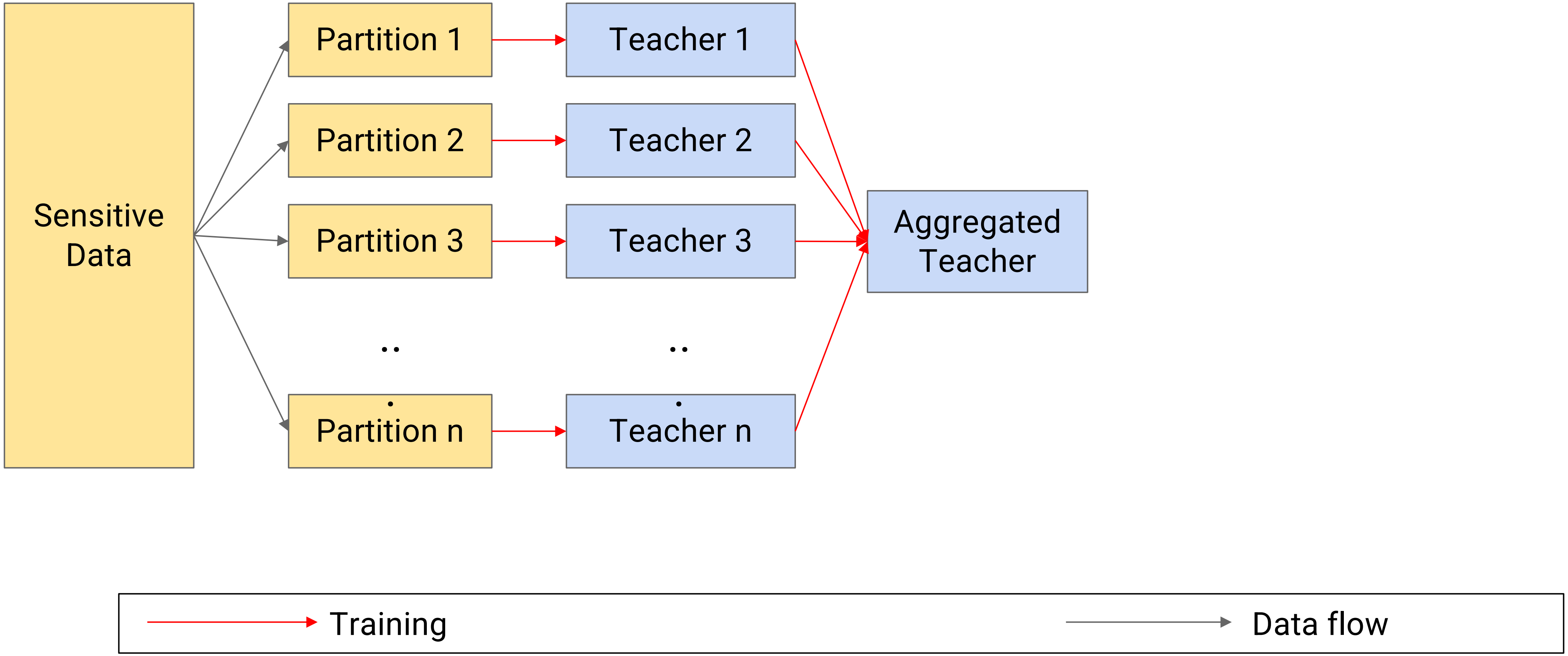
$$Lap\left(\frac{1}{\varepsilon}\right)$$



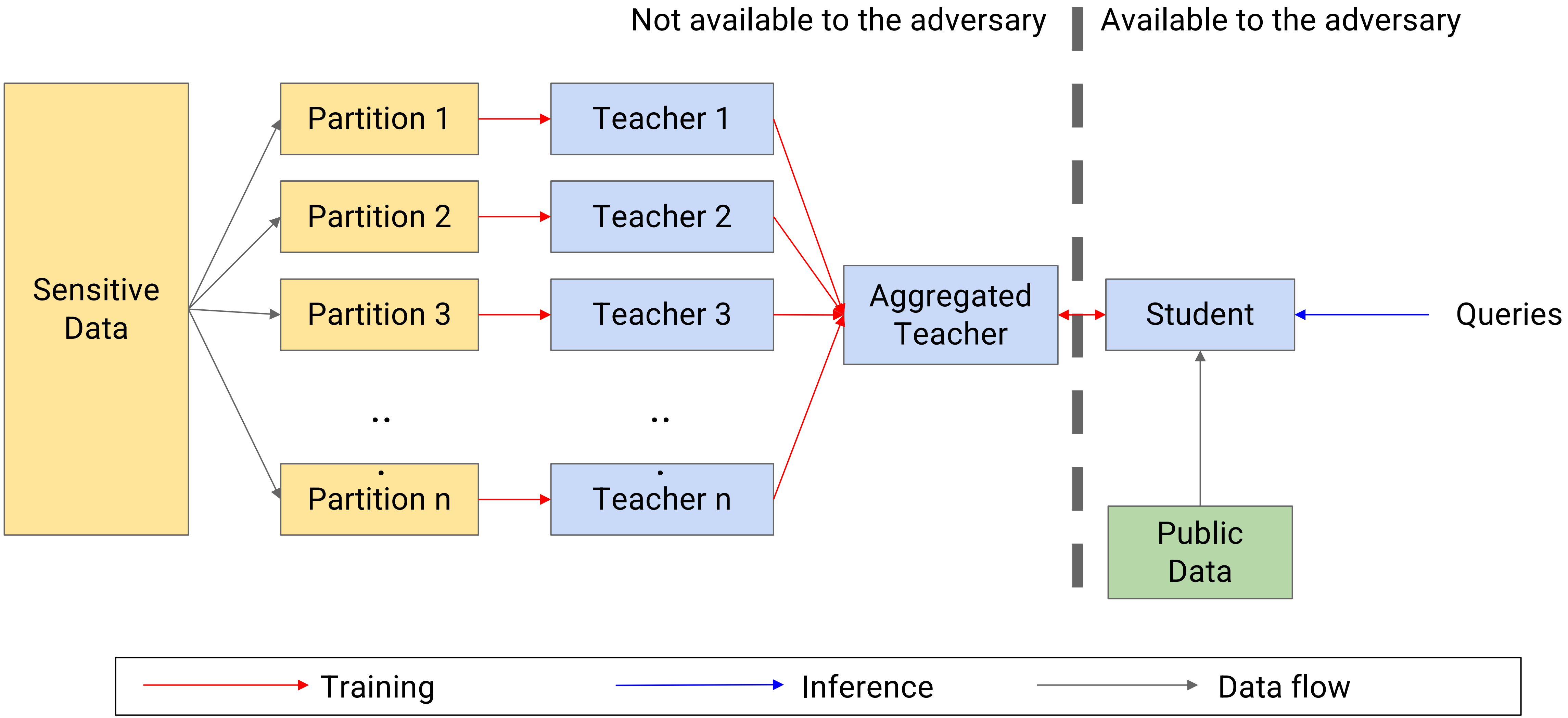
Take maximum

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

Teacher ensemble



Student training



Why train an additional “student” model?

The aggregated teacher violates our threat model:

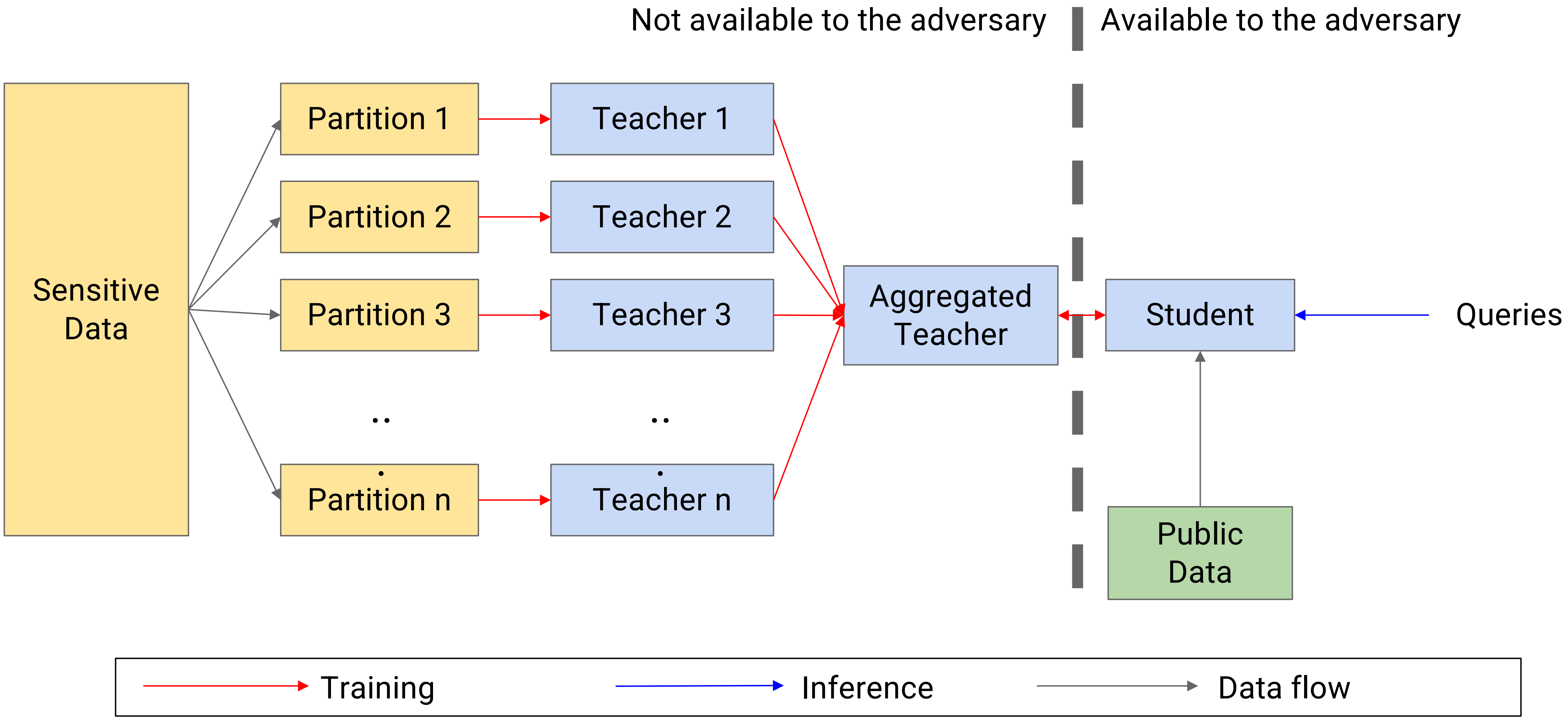
1. **Each prediction increases total privacy loss.**

   Privacy budgets create a tension between the accuracy and number of predictions.

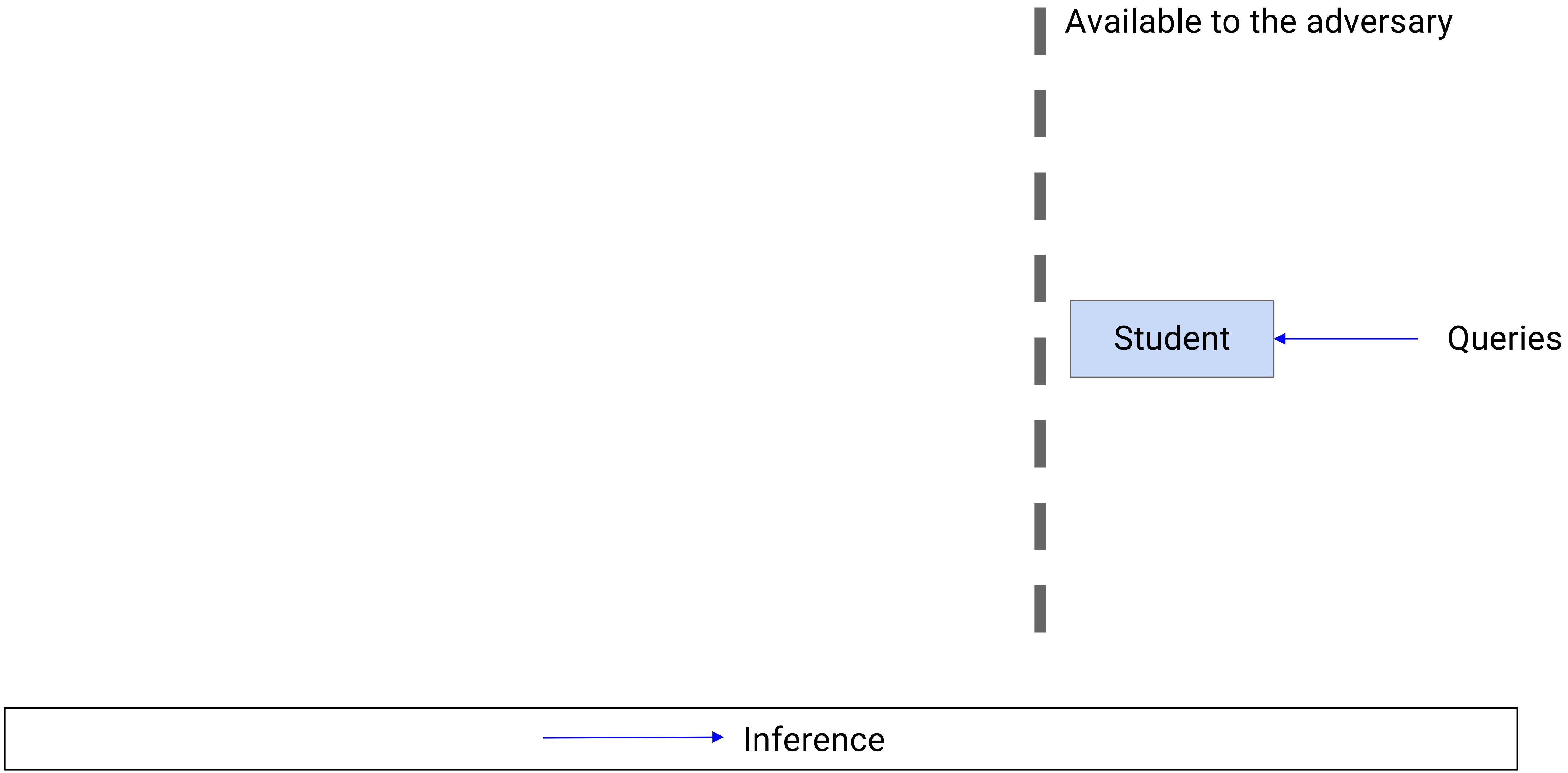
2. **Inspection of internals may reveal private data.**

   Privacy guarantees should hold in the face of white-box adversaries.

Student training



Deployment



Differential privacy analysis

Differential privacy:
A randomized algorithm *M* satisfies (ε,δ) differential privacy if for all pairs of neighbouring datasets (*d*,*d'*), for all subsets *S* of outputs:

$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$$

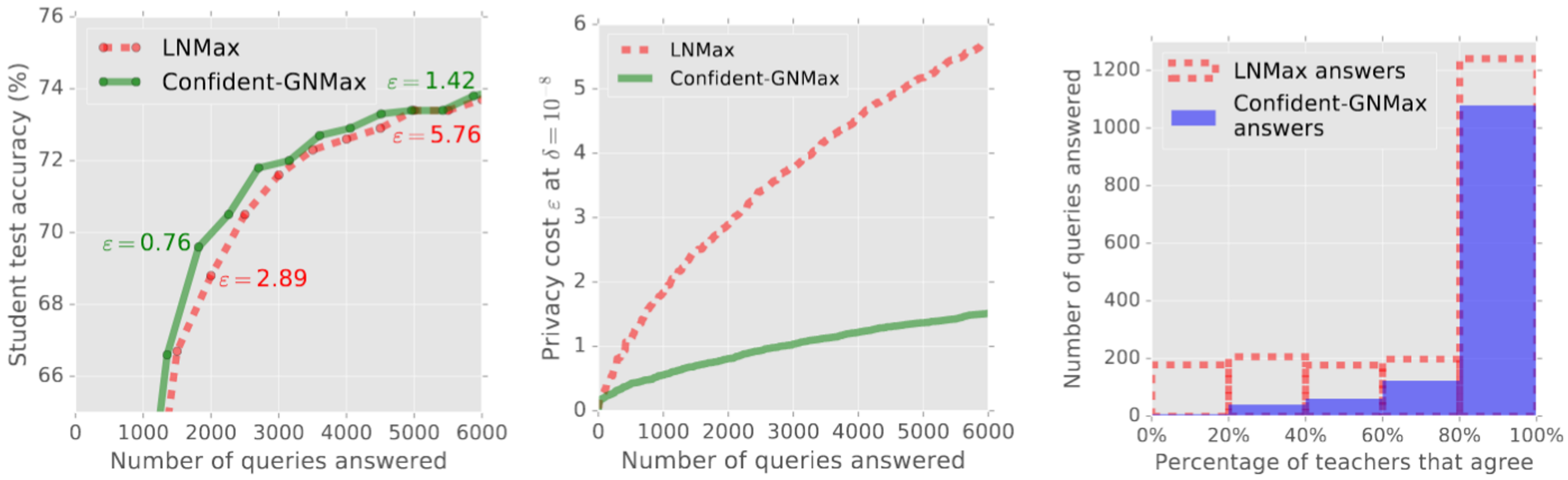
Application of the **Moments Accountant** technique (Abadi et al, 2016)

Strong **quorum** $\Longrightarrow$ Small privacy cost

Bound is **data-dependent**: computed using the empirical quorum

Synergy between utility and privacy. Win #1

1. Check privately for consensus
2. Run noisy argmax only when consensus is sufficient



Scalable Private Learning with PATE [ICLR 2018]
Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Ulfar Erlingsson

How to train a model with SGD?

```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute average loss L on batch B
  Compute average gradient of loss L wrt parameters θ



  Update parameters θ by a multiple of gradient average
```

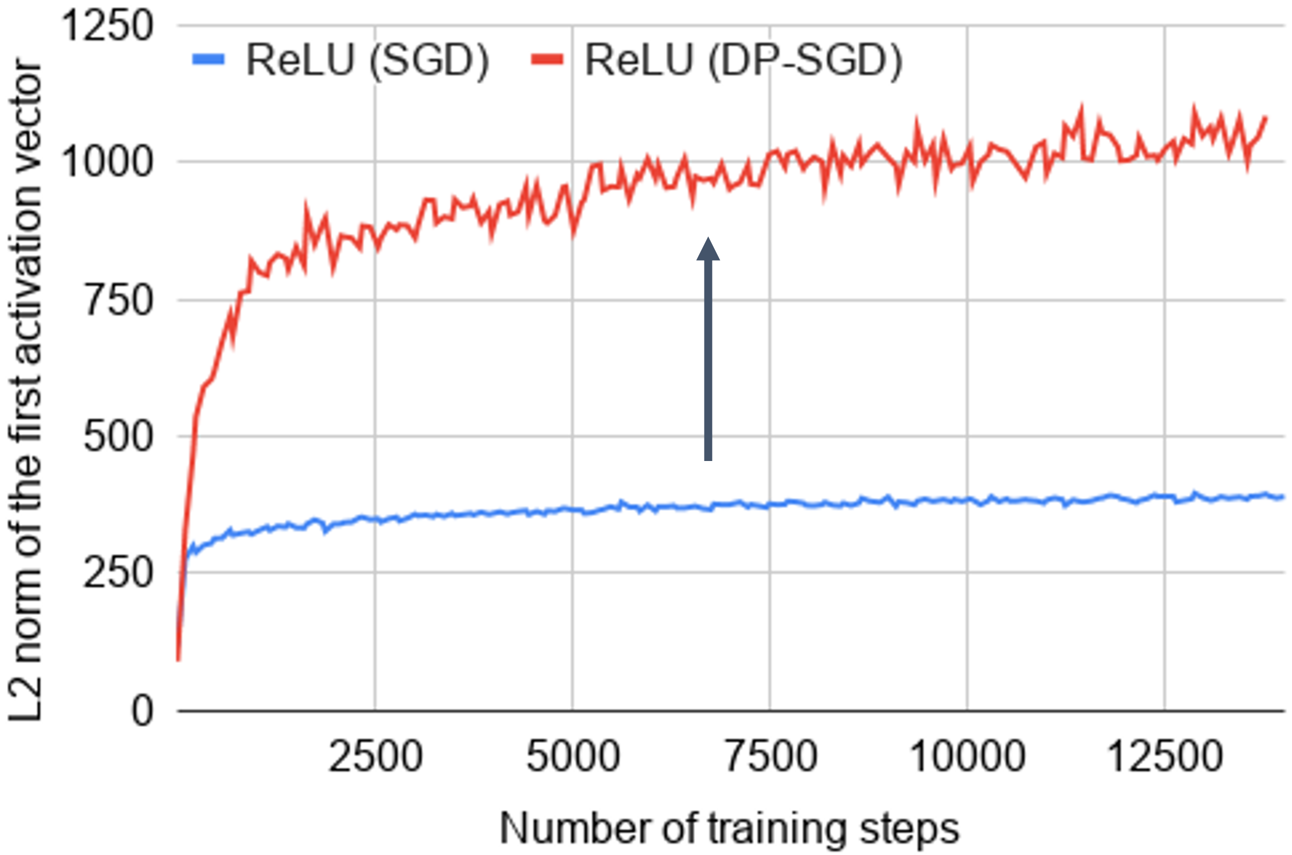
How to train a model with differentially private SGD?

```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute per-example loss L on batch B
  Compute per-example gradients of loss L wrt parameters θ
  Ensure L2 norm of gradients < C by clipping
  Add Gaussian noise to average gradients (as a function of C)
  Update parameters θ by a multiple of noisy gradient average
```

Deep Learning with Differential Privacy (CCS, 2016)
Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, Zhang

Our observation: DP-SGD leads to exploding activations



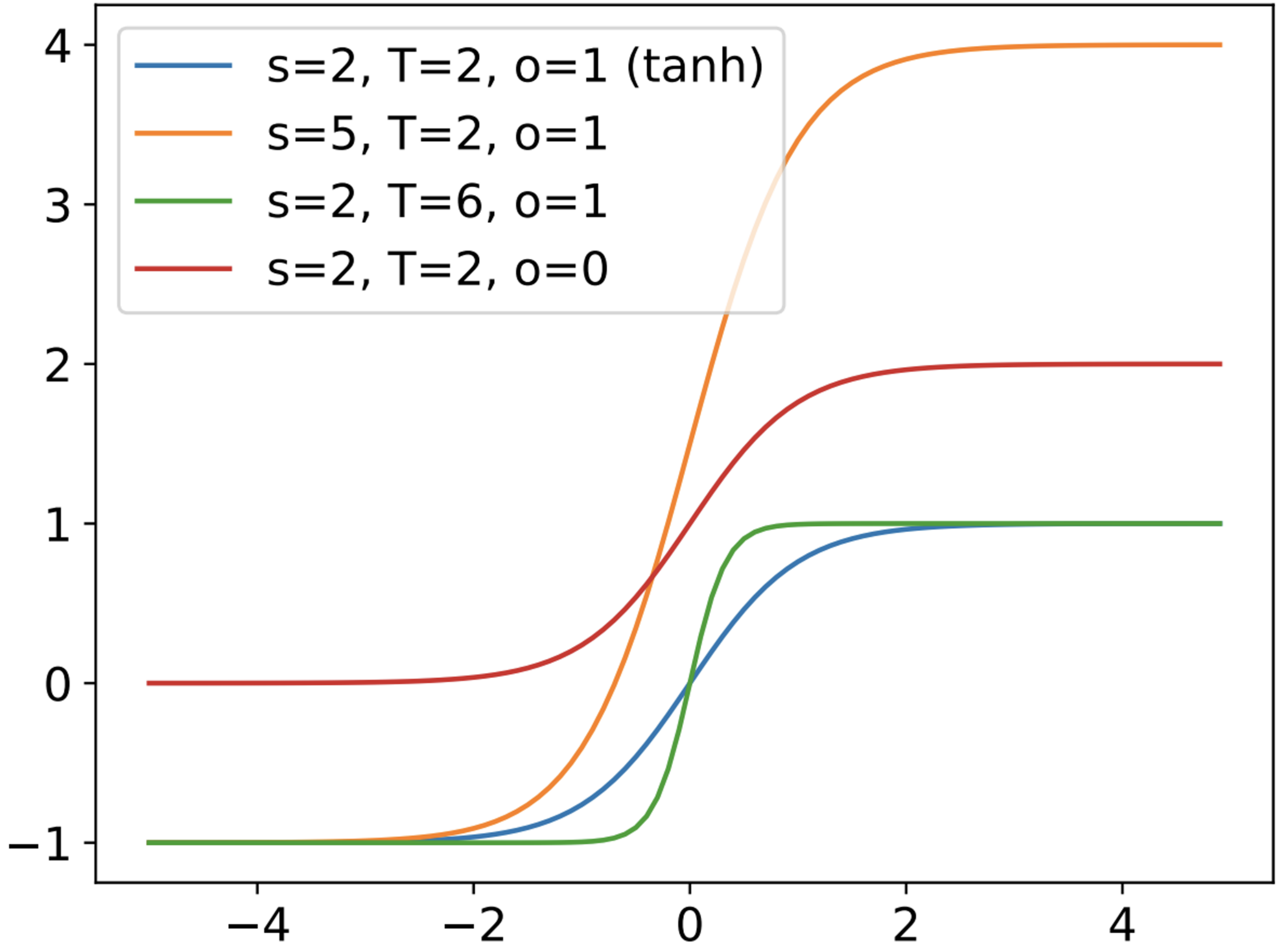
Tempered sigmoids: a family of bounded activation functions

scale

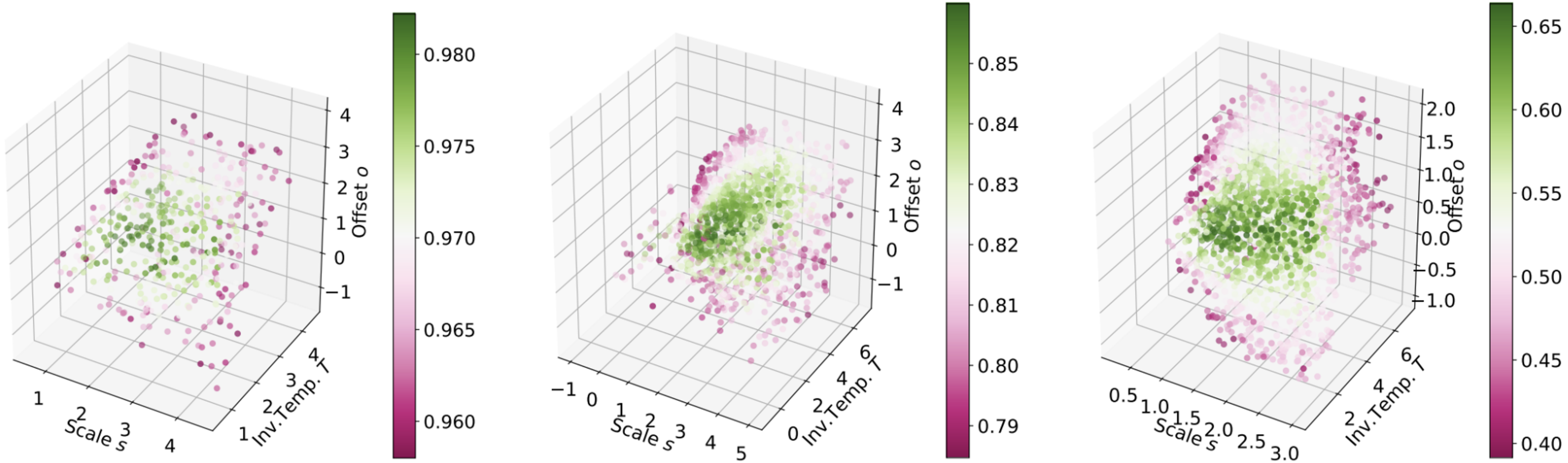
$$\phi_{s,T,o} : x \mapsto \frac{s}{1 + e^{-T \cdot x}} - o$$

temperature

offset



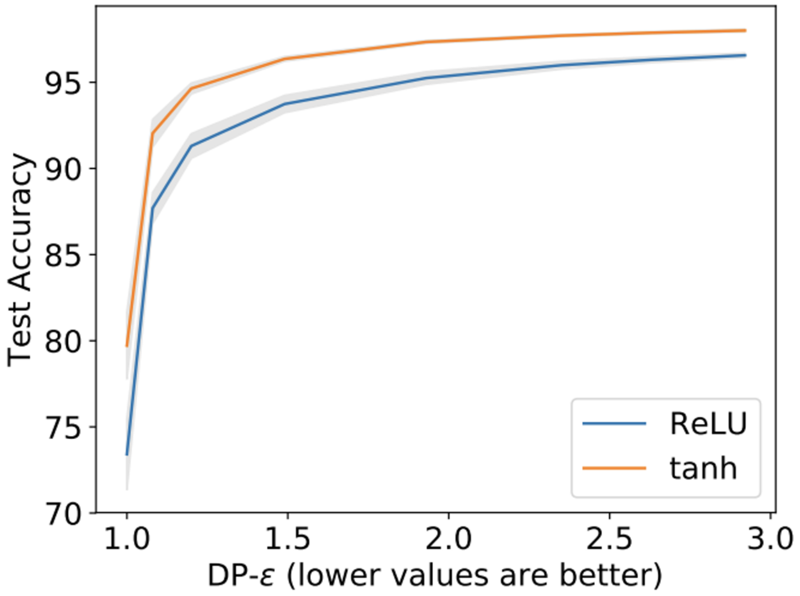
Improved privacy-utility tradeoffs with tempered sigmoids



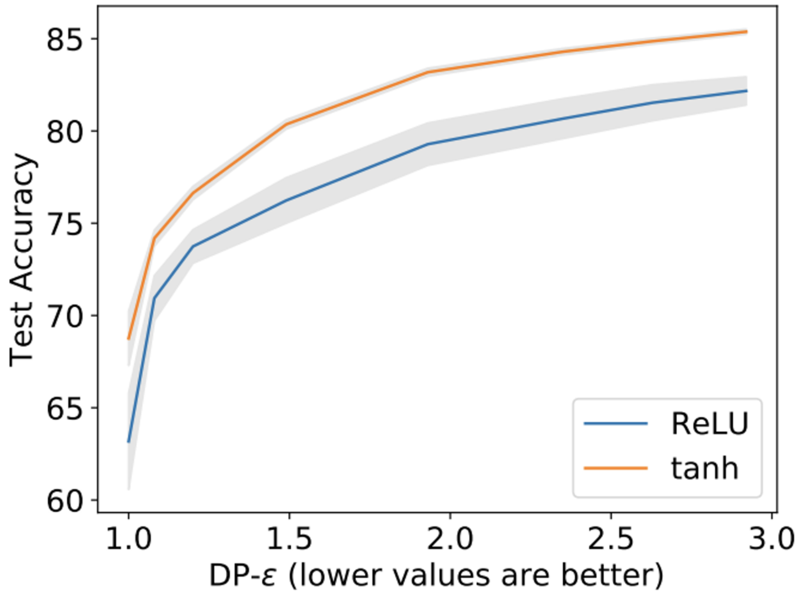
MNIST FashionMNIST CIFAR10

All 3D plots indicate accuracy using color (for a fixed privacy guarantee)

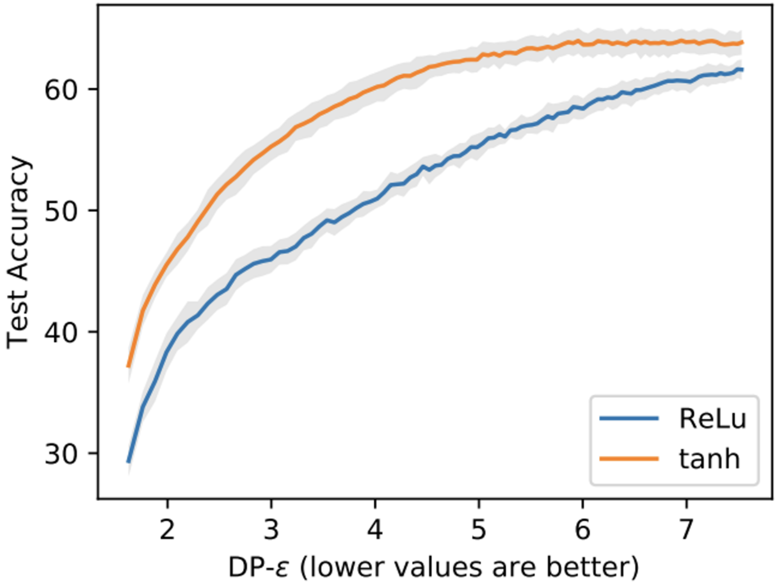
A particular case: `tanh`



MNIST

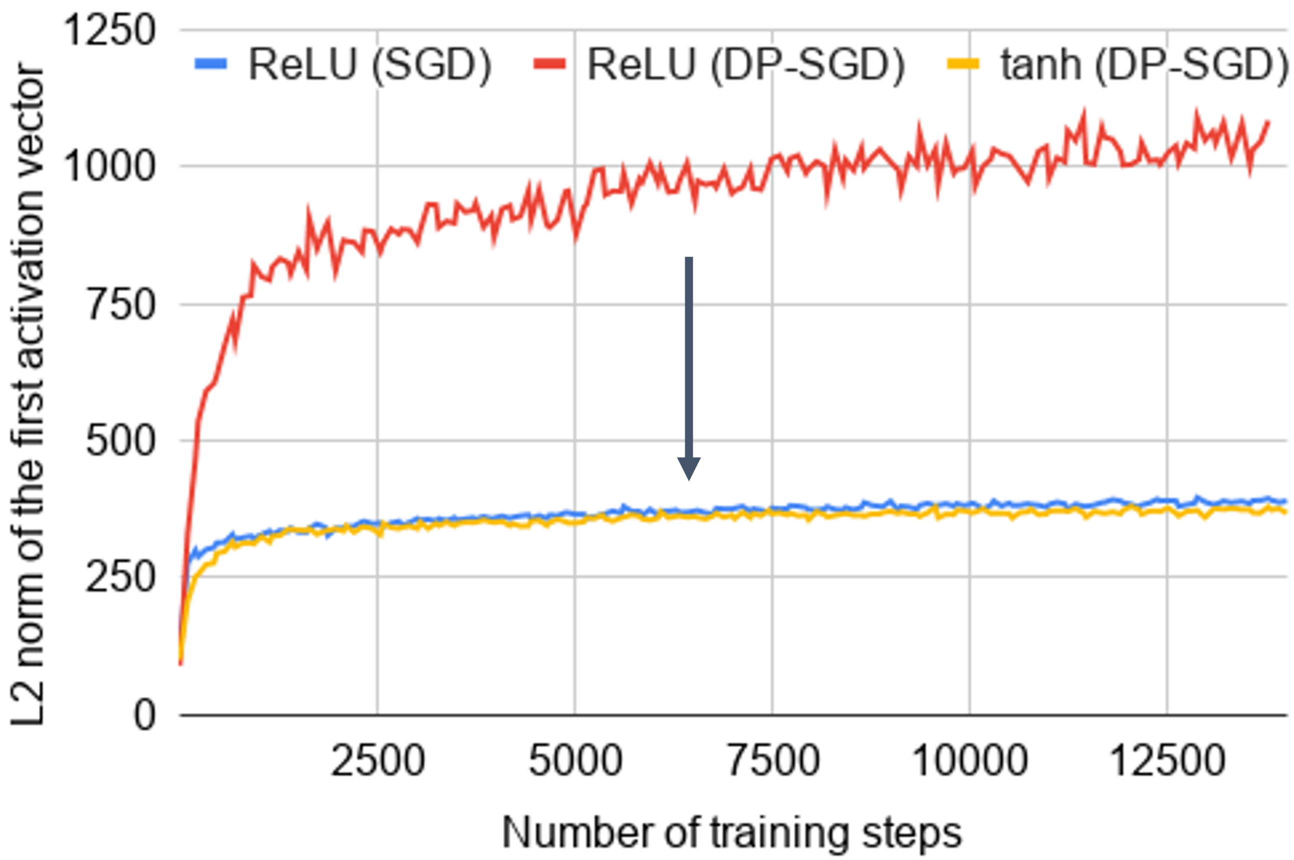


FashionMNIST



CIFAR10

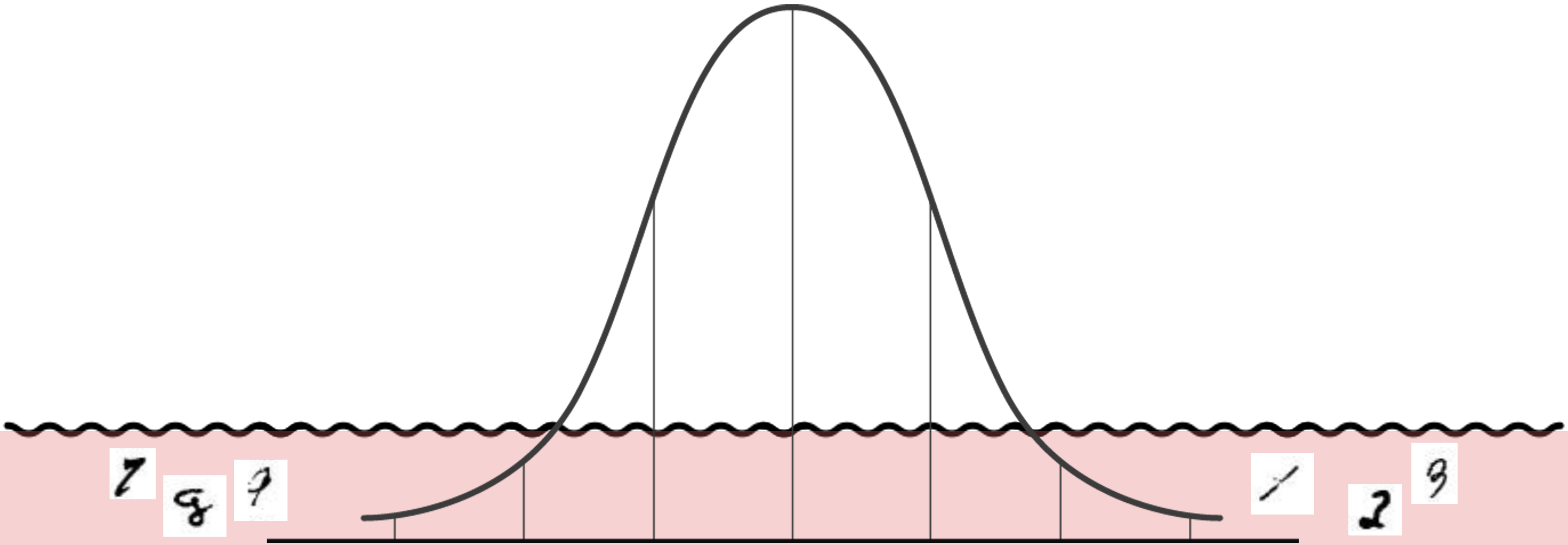
DP-SGD with tanh does **not** lead to exploding activations

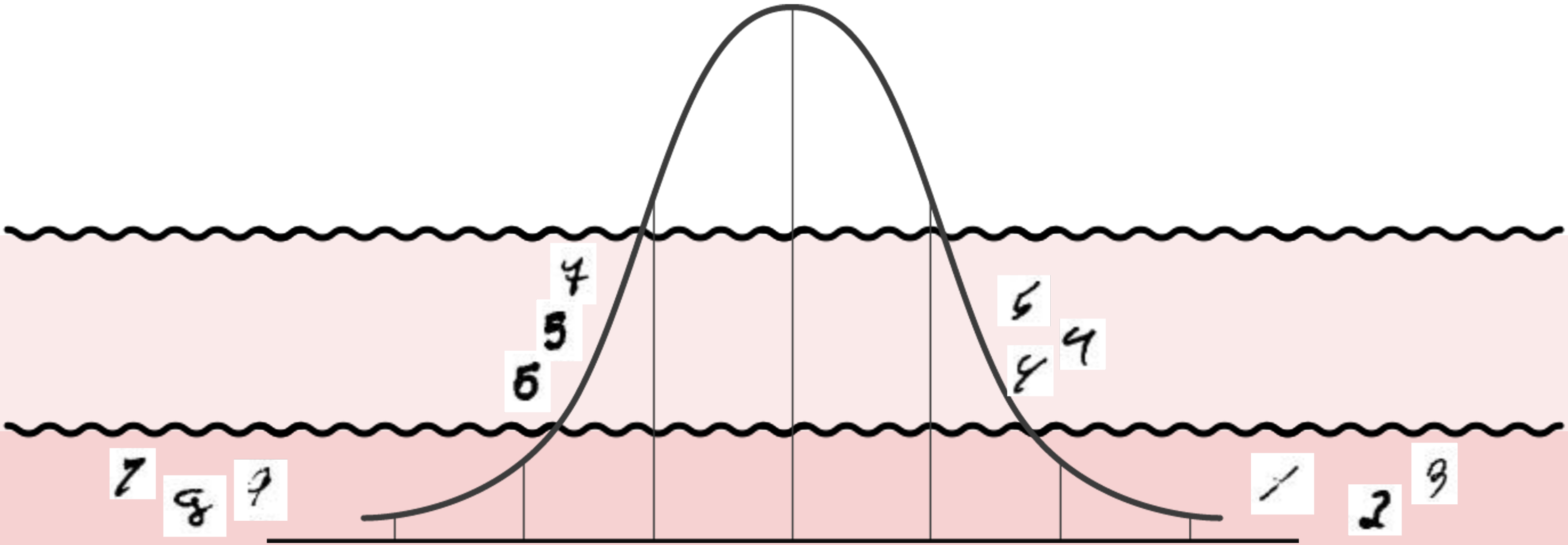


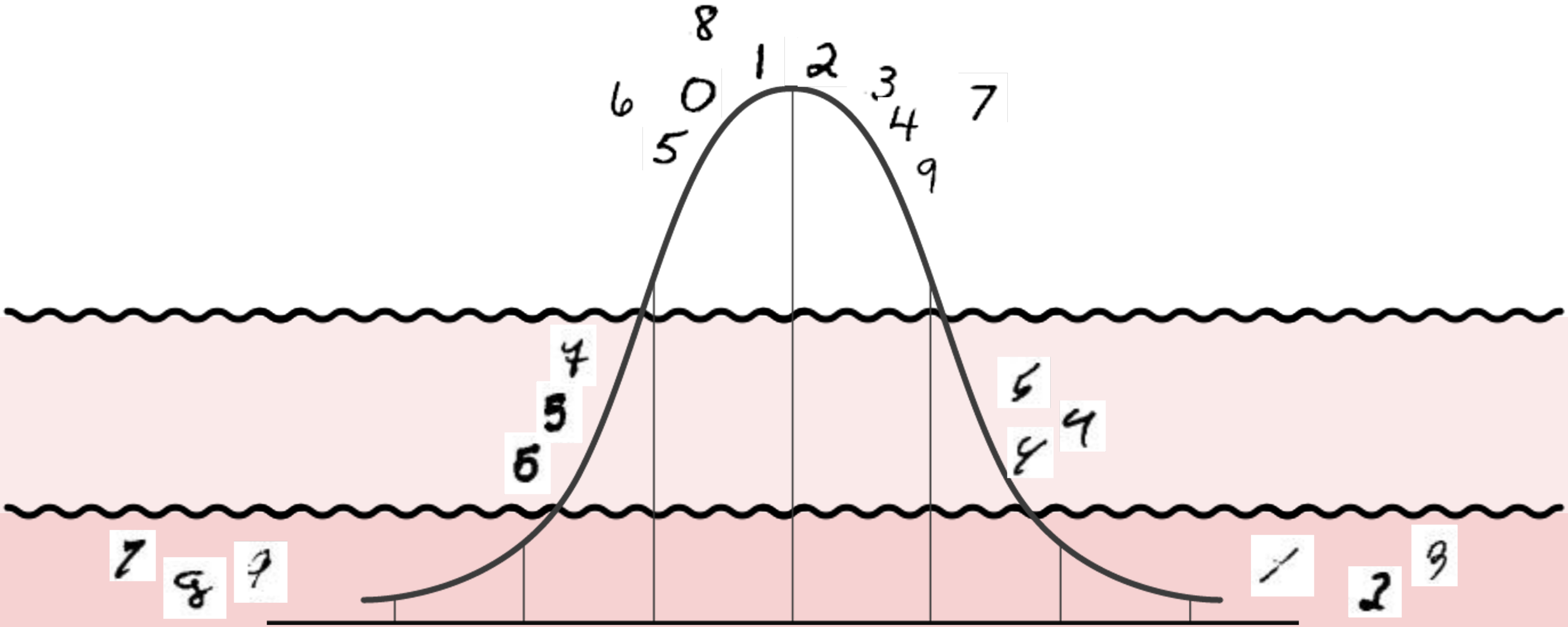
Improving the DP-SGD state-of-the-art with tanh

Dataset	Technique	Acc.	ε	δ
MNIST	SGD w/ ReLU (not private)	99.0%	∞	0
	DP-SGD w/ ReLU	96.6%	2.93	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**98.1%**	**2.93**	10^{-5}
FashionMNIST	SGD w/ ReLU (not private)	89.4%	∞	0
	DP-SGD w/ ReLU	81.9%	2.7	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**86.1%**	**2.7**	10^{-5}
CIFAR10	SGD w/ ReLU (not private)	76.6%	∞	0
	DP-SGD w/ ReLU	61.6%	7.53	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**66.2%**	**7.53**	10^{-5}

Tempered Sigmoid Activations for Deep Learning with Differential Privacy (AAAI 2021)
Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, Úlfar Erlingsson



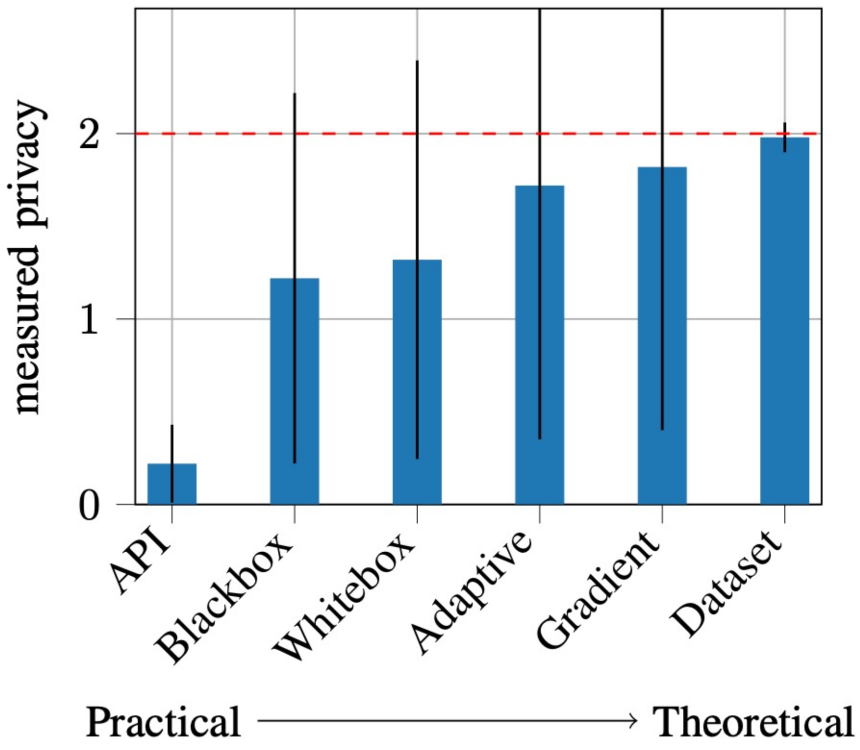




[arxiv:1910.13427]

DP is not a silver bullet, reason #1: privacy still comes at the price of average case performance on cu

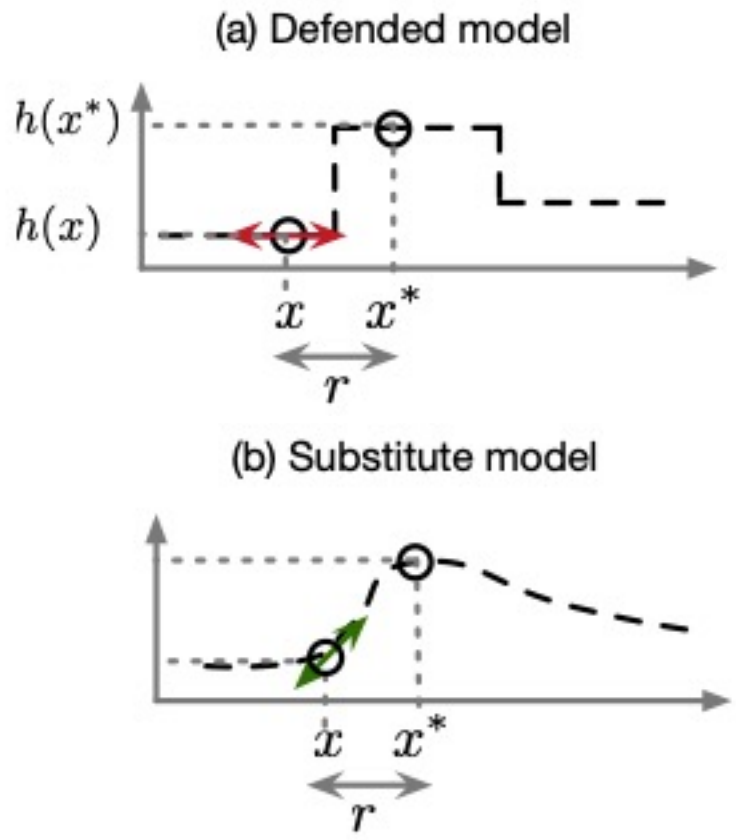
Dataset	Technique	Acc.	ε	δ
MNIST	SGD w/ ReLU (not private)	99.0%	∞	0
	DP-SGD w/ ReLU	96.6%	.93	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**98.1%**	**.93**	10^{-5}
FashionMNIST	SGD w/ ReLU (not private)	89.4%	∞	0
	DP-SGD w/ ReLU	81.9%	2.7	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**86.1%**	**2.7**	10^{-5}
CIFAR10	SGD w/ ReLU (not private)	76.6%	∞	0
	DP-SGD w/ ReLU	61.6%	.53	10^{-5}
	DP-SGD w/ tempered sigmoid (tanh) [ours]	**66.2%**	**.53**	10^{-5}

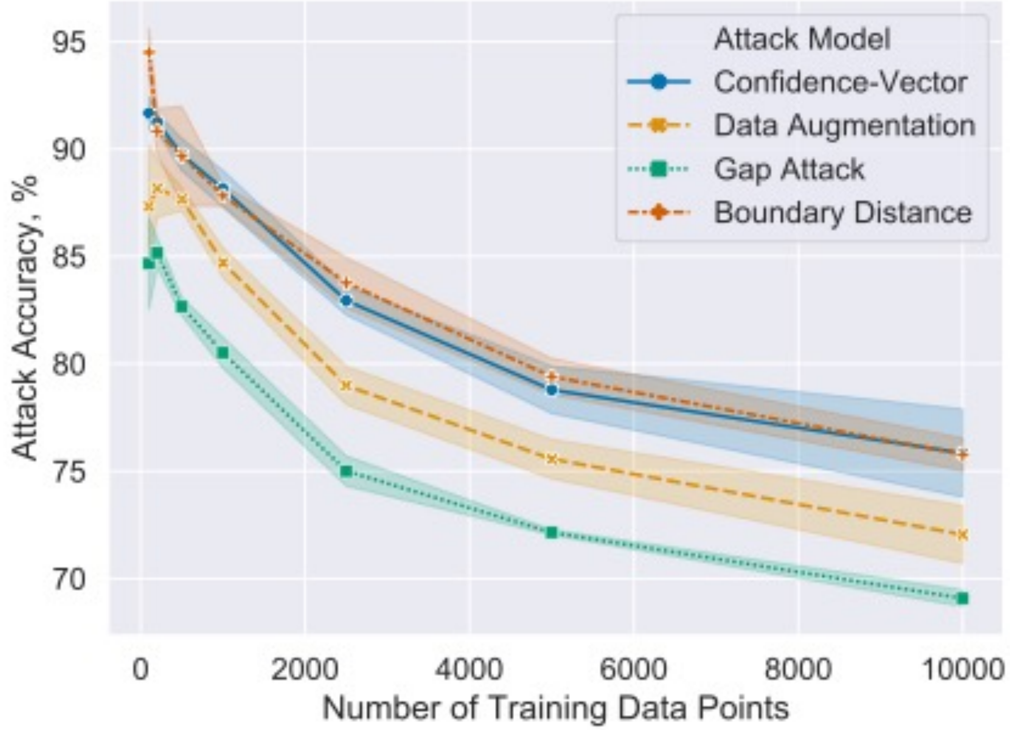


Tempered Sigmoid Activations for Deep Learning with Differential Privacy
Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, Úlfar Erlingsson

Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning (IEEE SP 2021)
Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, Nicholas Carlini

Gradient masking vs. Confidence masking

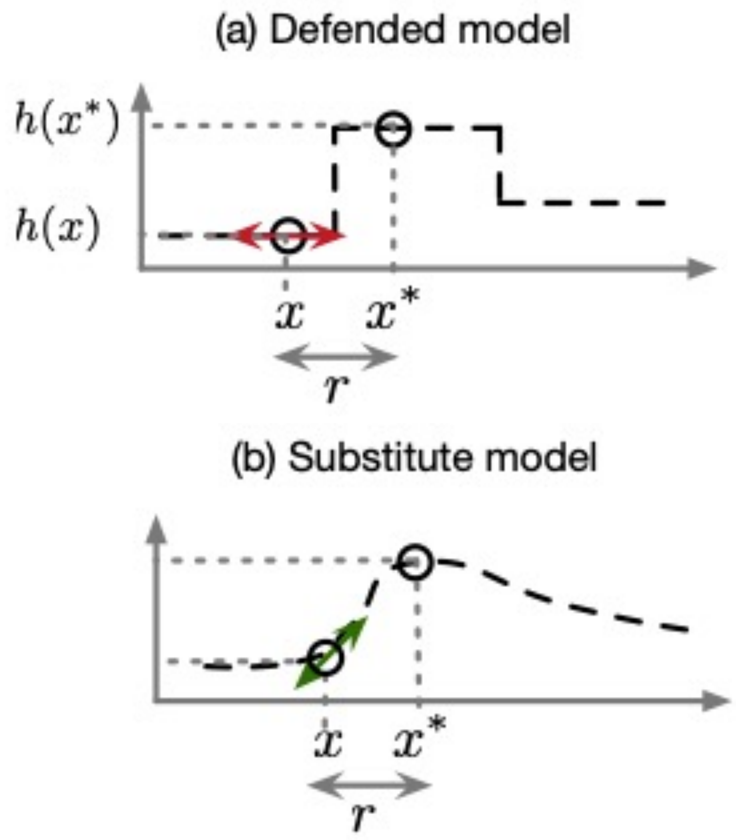




Practical Black-Box Attacks against Machine Learning. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami.

Label-Only Membership Inference Attacks Christopher A. Choquette Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot

Gradient masking vs. Confidence masking



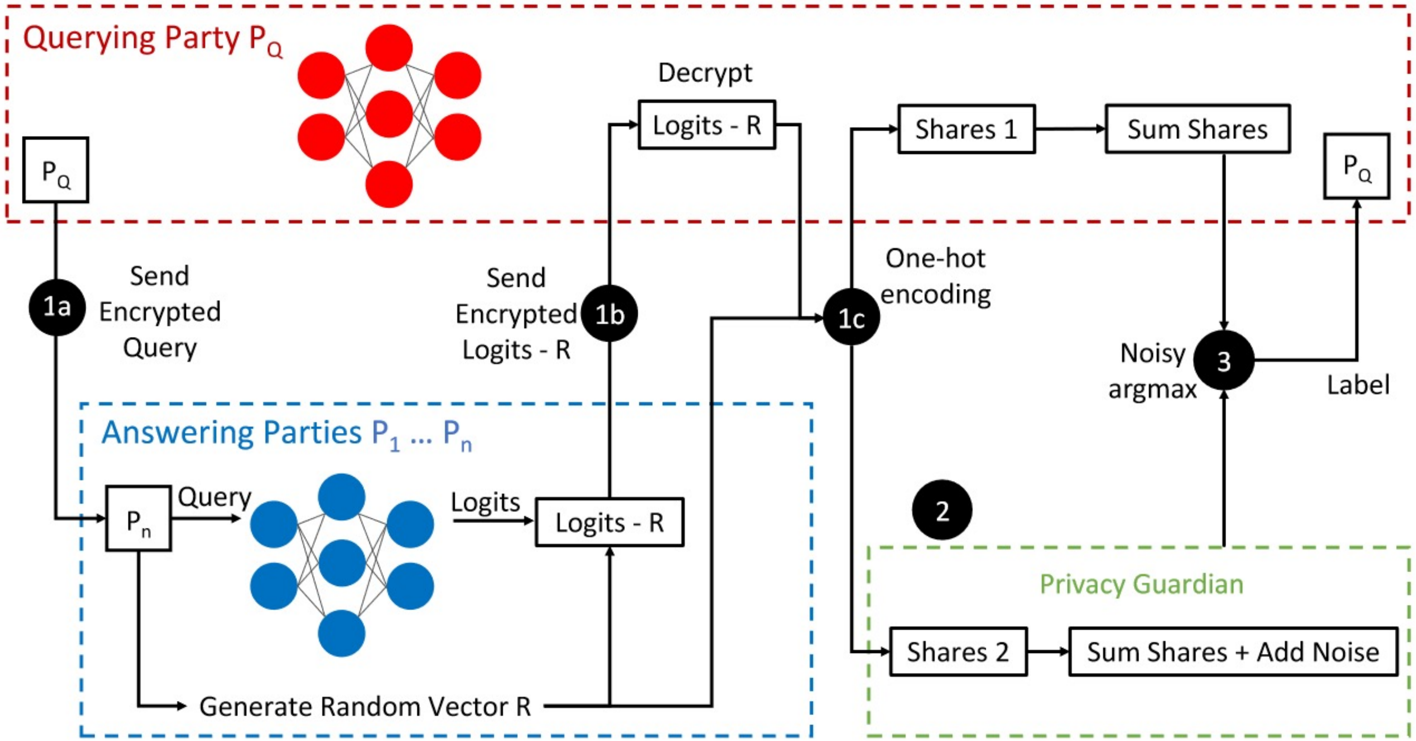
Win #2

Practical Black-Box Attacks against Machine Learning. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami.

Label-Only Membership Inference Attacks Christopher A. Choquette Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot

DP is not a silver bullet, reason #2: it does not provide confidentiality

Example solution



- Few distributed participants, can use heterogeneous architectures
- Evaluation shows improvements to accuracy and balanced accuracy (fairness)

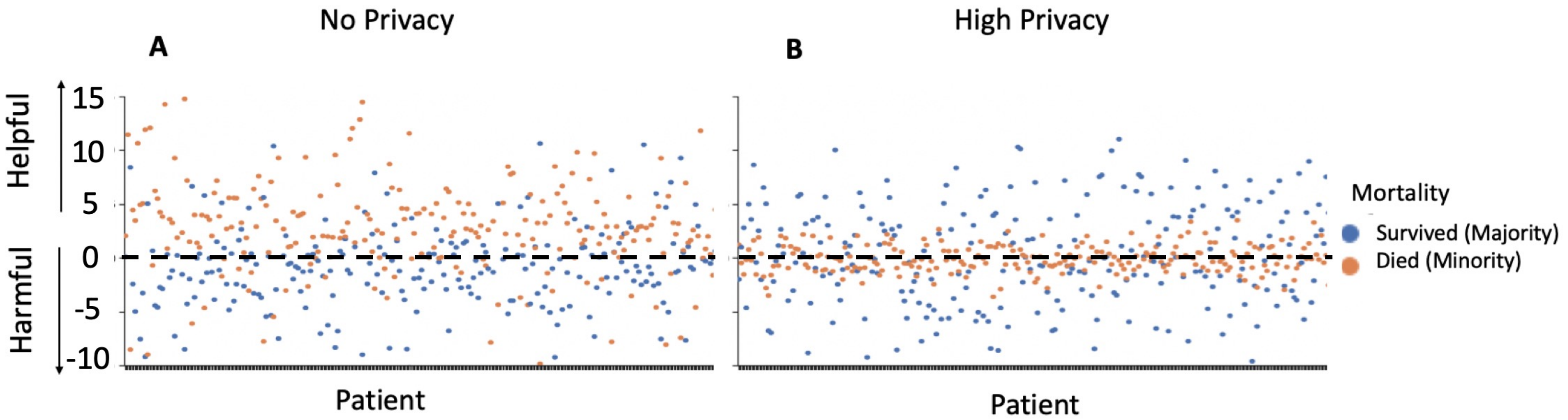
CaPC Learning: Confidential and Private Collaborative Learning (ICLR 2021)
Christopher A. Choquette-Choo, Natalie Dullerud, Adam Dziedzic, Yunxiang Zhang, Somesh Jha, Nicolas Papernot, Xiao Wang

DP is not a silver bullet, reason #3: fairness tradeoff

Utility on Long Tailed Datasets

Task	Model	No Privacy	High Privacy (ε, δ)
MIMIC-III Mortality	Logistic Regression	**0.82 ± 0.03**	**0.60 ± 0.04 (3.54, 10^{-5})**
	GRU-D	**0.79 ± 0.03**	**0.53 ± 0.03 (2.65, 10^{-5})**
NIH Chest X-Ray Disease Prediction	Finetuned DenseNet-121	**0.84 ± 0.00**	**0.49 ± 0.00 (0.84, 10^{-6})**

Unfairness Due to Overinfluence of Majority Subgroups



Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. (FAccT 2021)
Vinith Suriyakumar, Nicolas Papernot, Anna Goldenberg, Marzyeh Ghassemi.

Useful resources

- https://desfontain.es/privacy/differential-privacy-awesomeness.html
- https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
- https://github.com/tensorflow/privacy