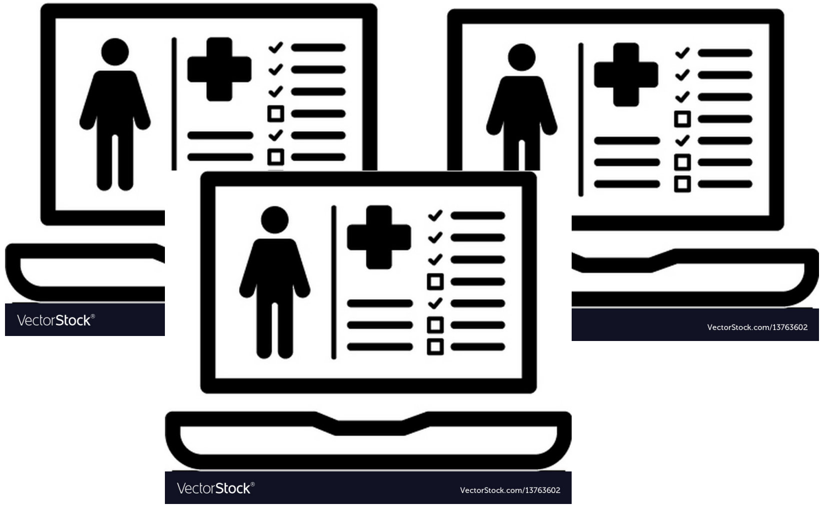
Three Flavors of Private Machine Learning

Nicolas Papernot*

University of Toronto & Vector Institute

***presenting the work of many others**

Why is the trustworthiness of ML important? *privacy*

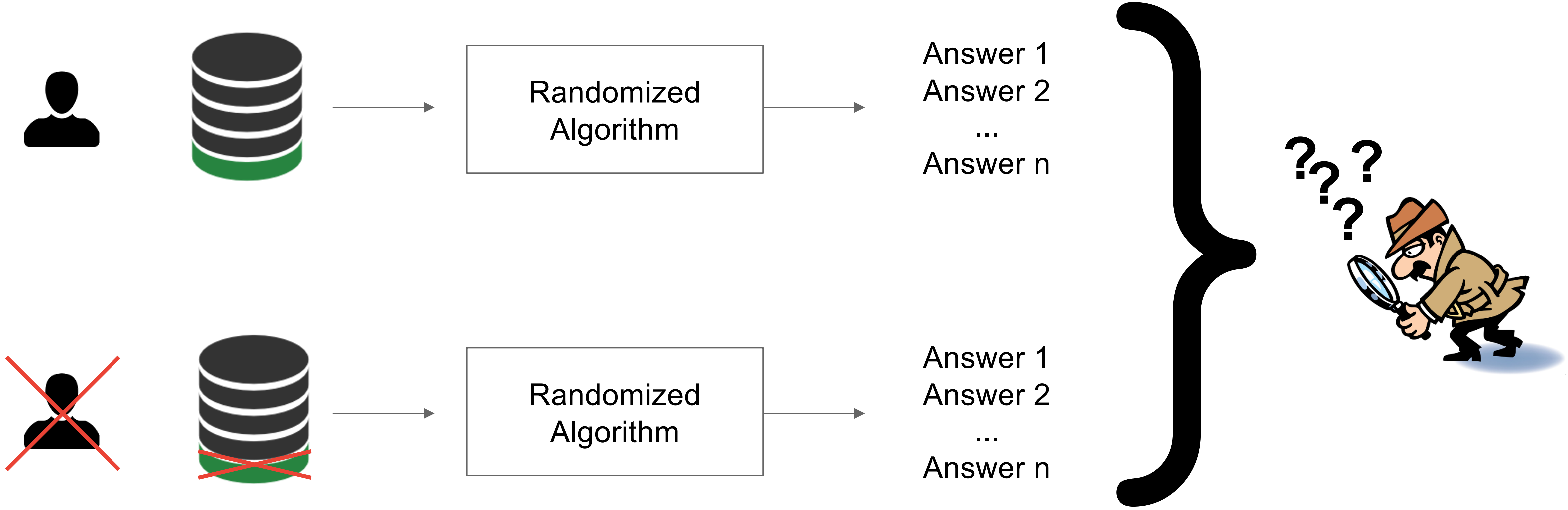


Membership inference attack

Adversary figures out whether data was in the training set from model predictions

Choquette-Choo et al. Label-Only Membership Inference Attacks

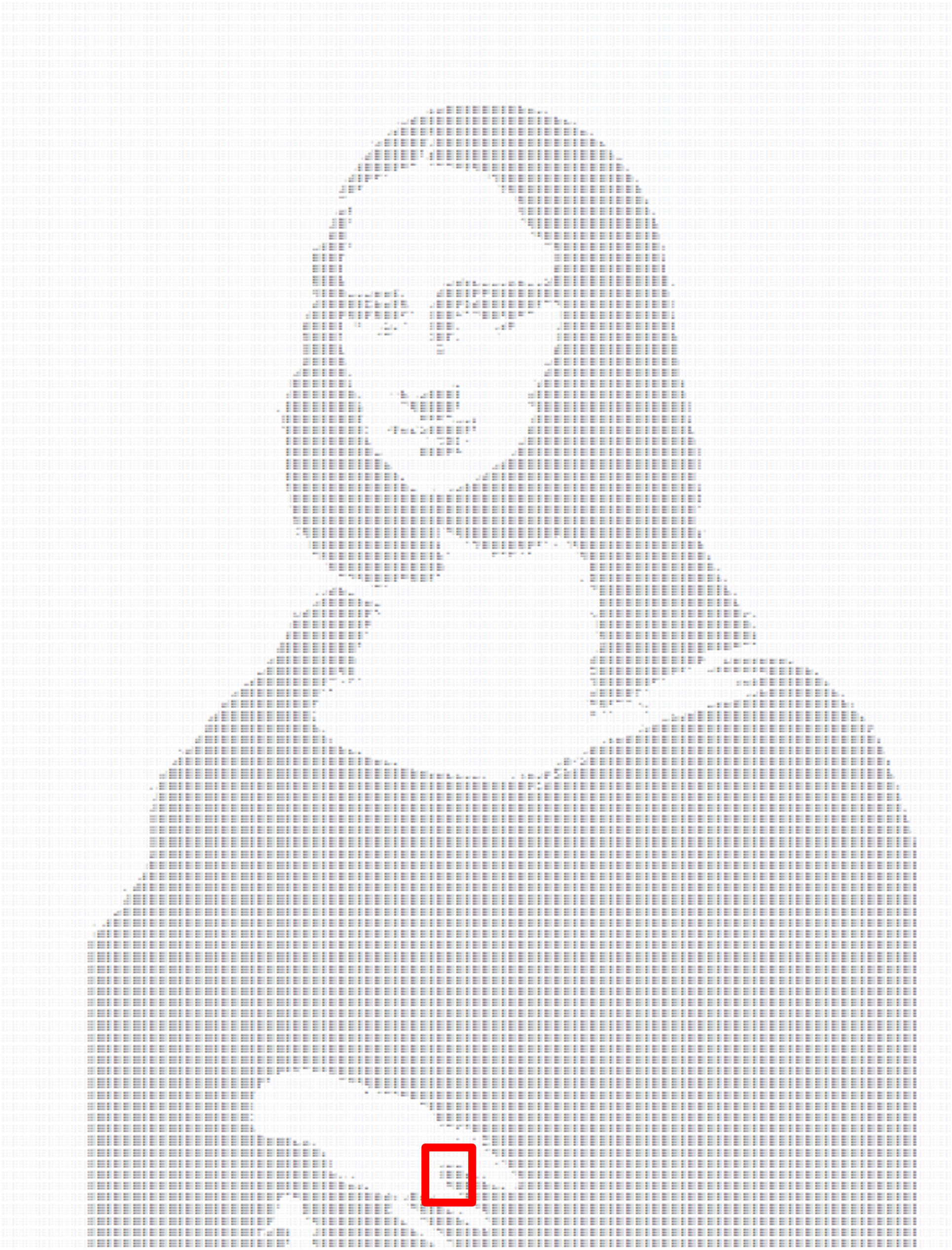
How to define trustworthiness? A successful attempt with privacy



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

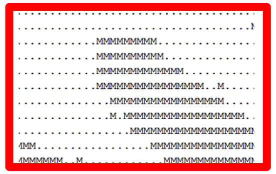
Dwork et al. *Calibrating noise to sensitivity in private data analysis.*

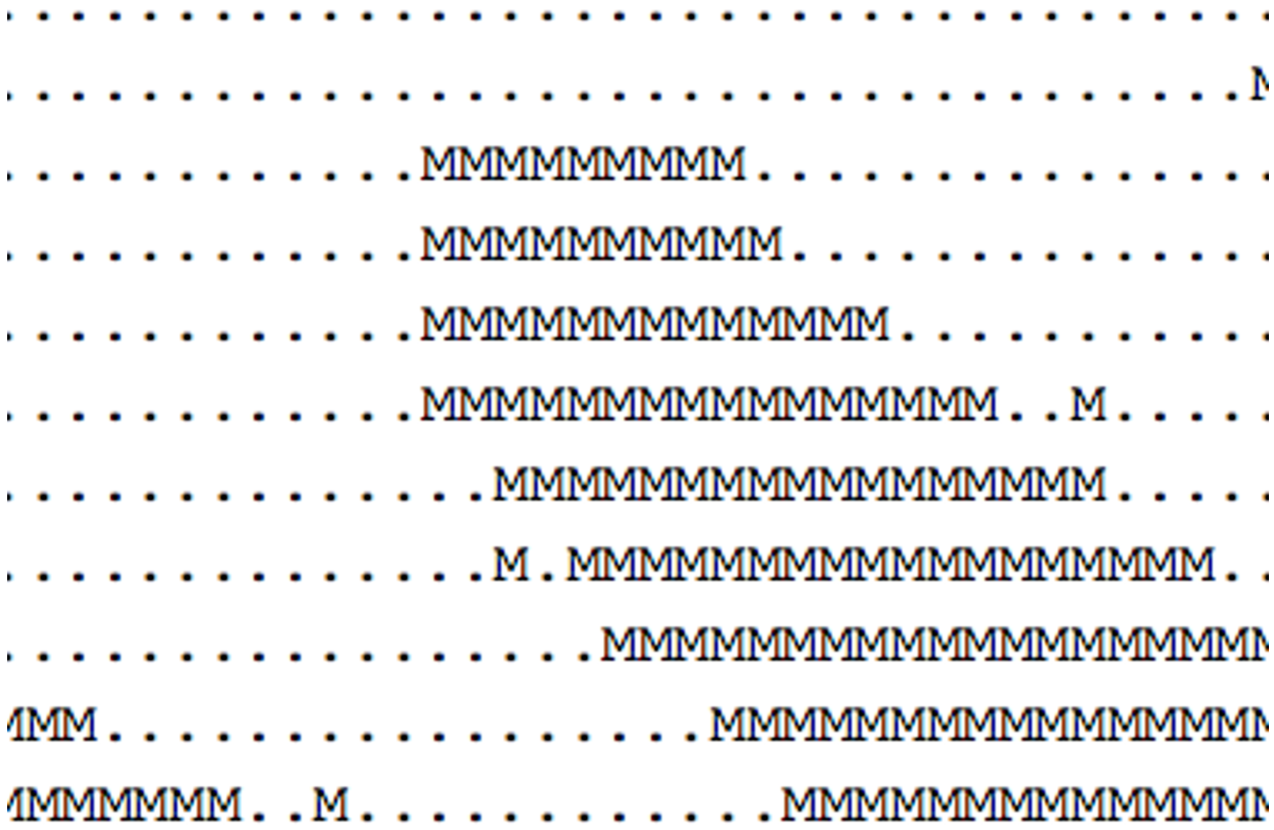
A Metaphor For Private Learning



Slides adapted from Ulfar Erlingsson

An Individual’s Training Data





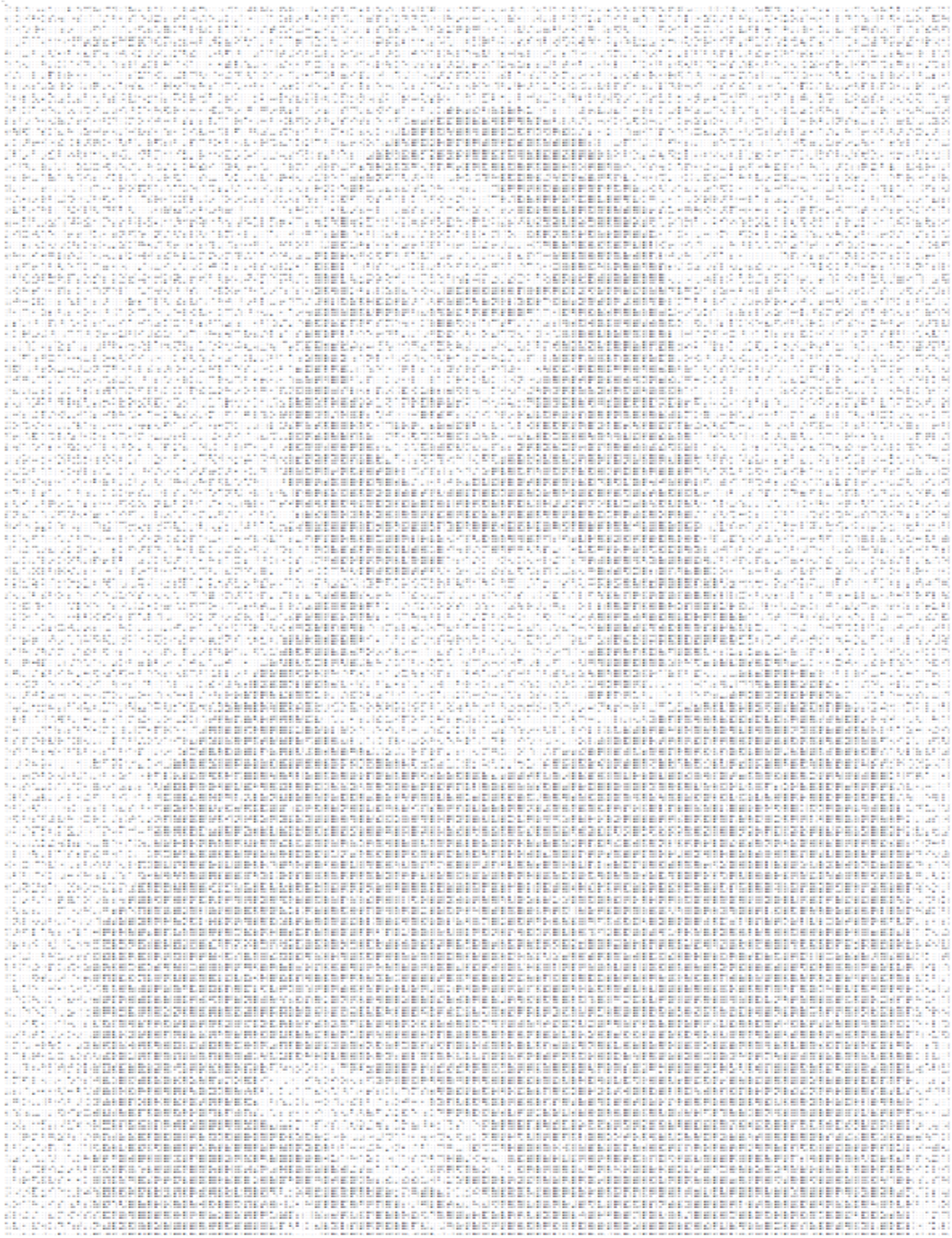
Slides adapted from Ulfar Erlingsson

An Individual's Training Data

Each bit is flipped with probability 25%

```
.......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
```

Big Picture Remains!



Slides adapted from Ulfar Erlingsson

How to train a model?

```
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
```

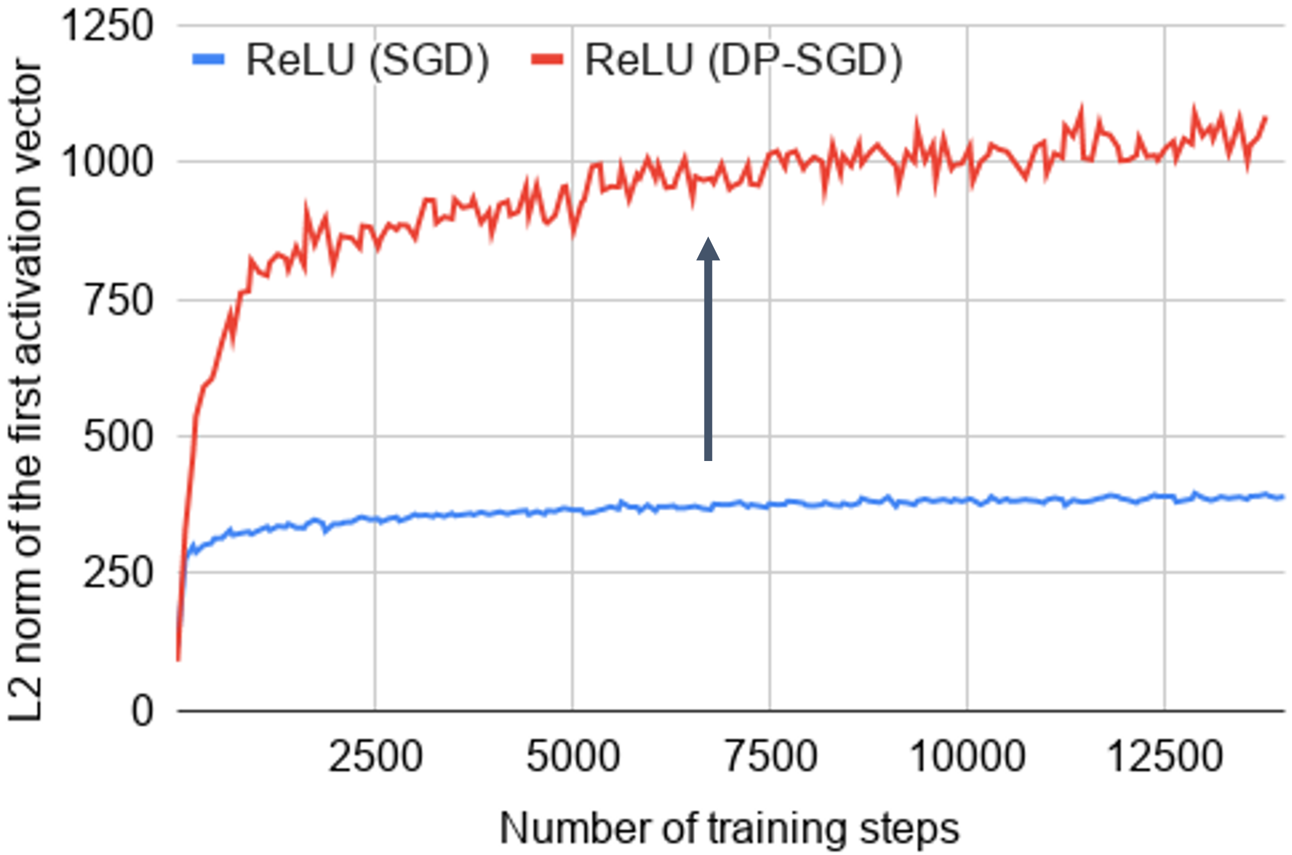
A first flavor: How to train a model with differential privacy?

```
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



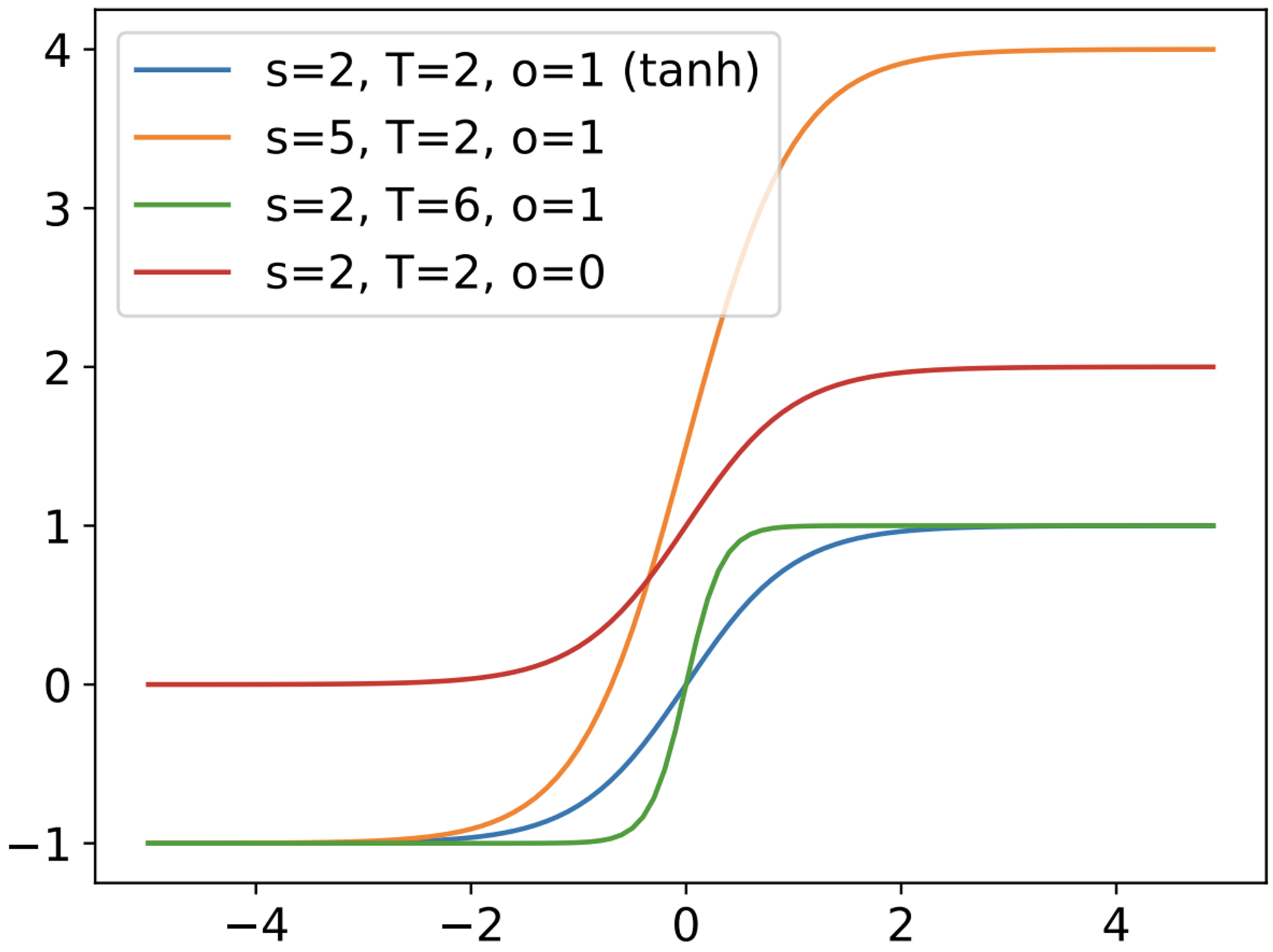
Tempered sigmoids: a family of bounded activation functions

scale

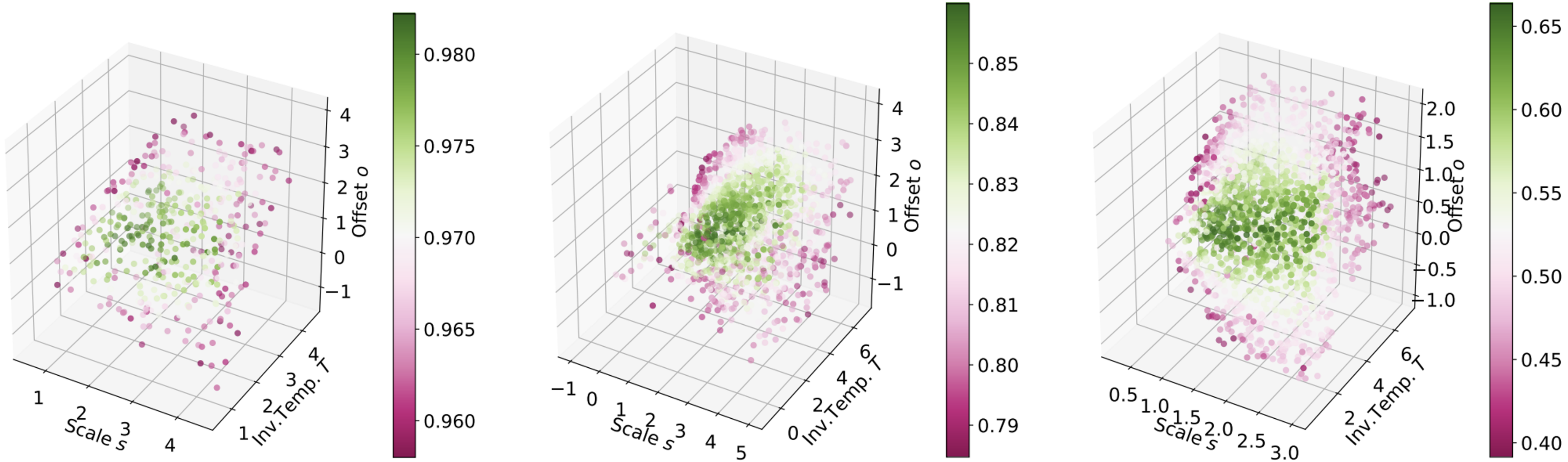
$$\phi_{s,T,o} : x \mapsto \frac{s}{1+e^{-T\cdot x}} - o$$

temperature

offset



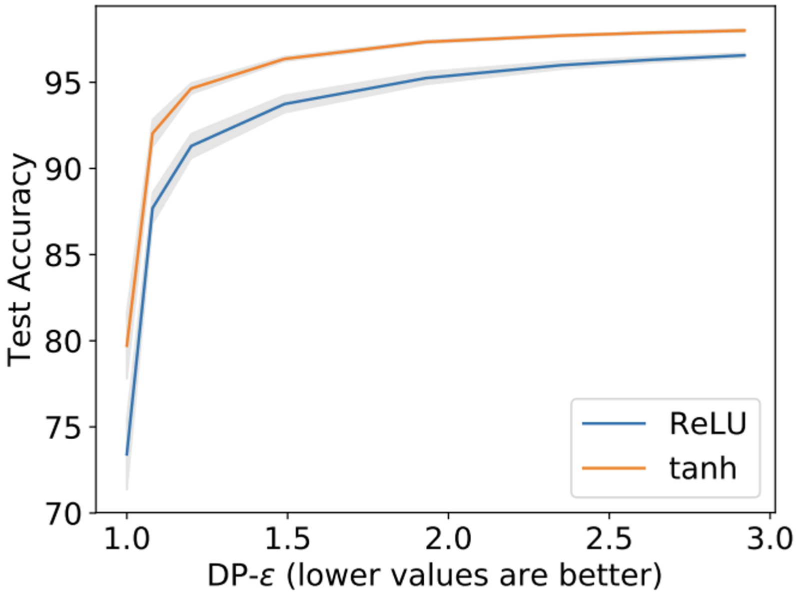
Improved privacy-utility tradeoffs with tempered sigmoids



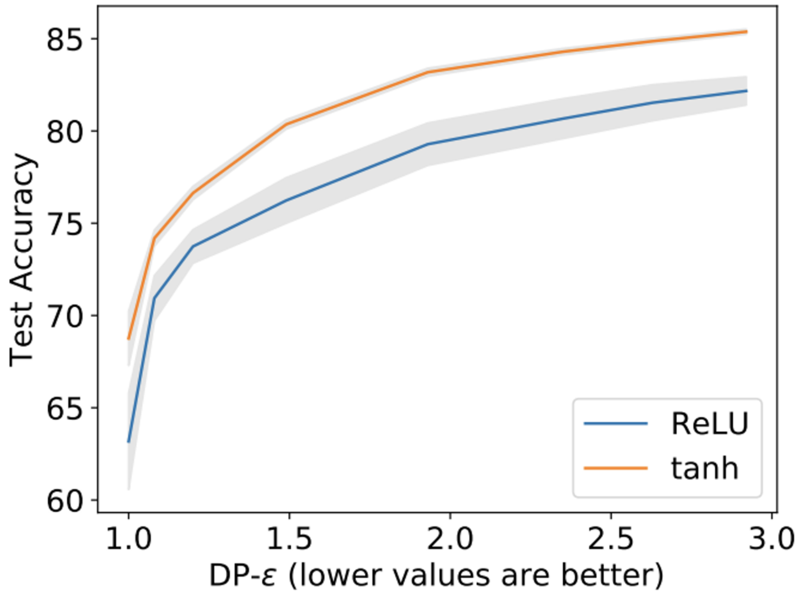
MNIST FashionMNIST CIFAR10

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

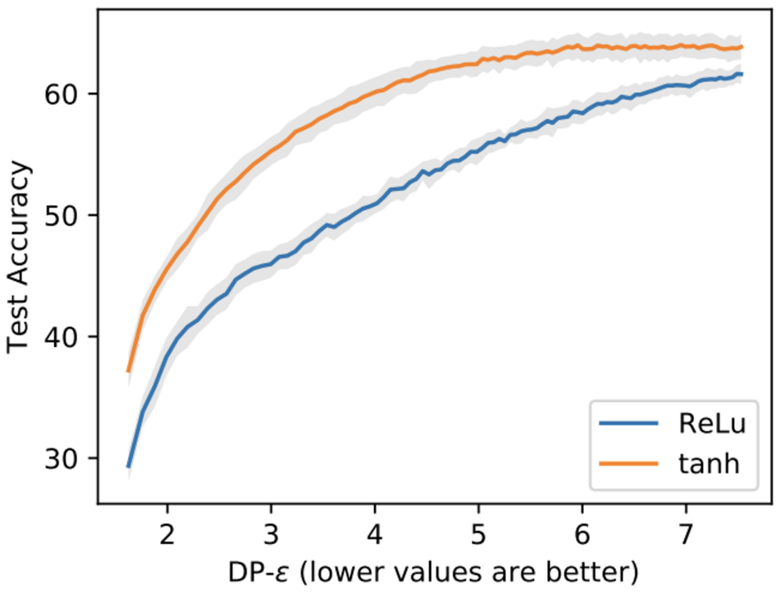
A particular case: `tanh`



MNIST

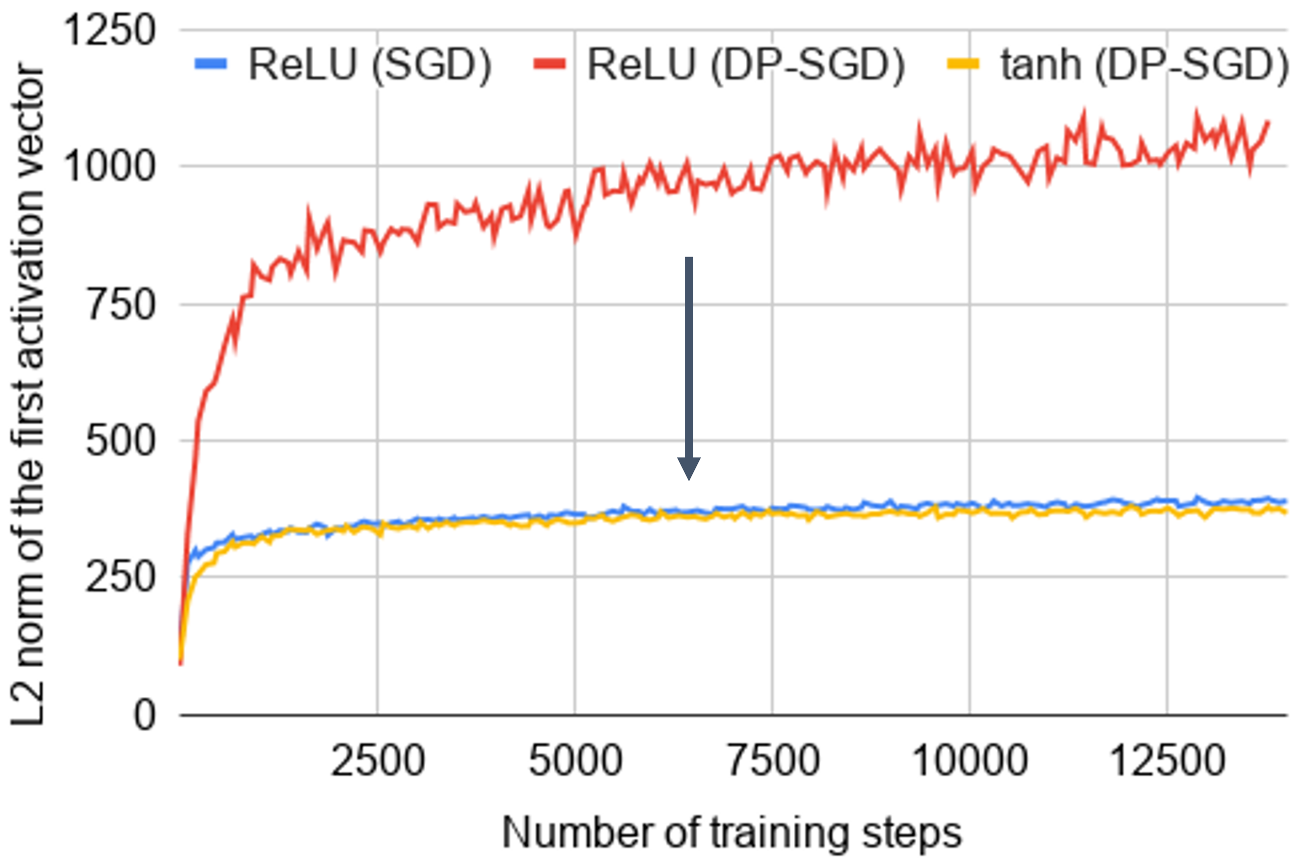


FashionMNIST



CIFAR10

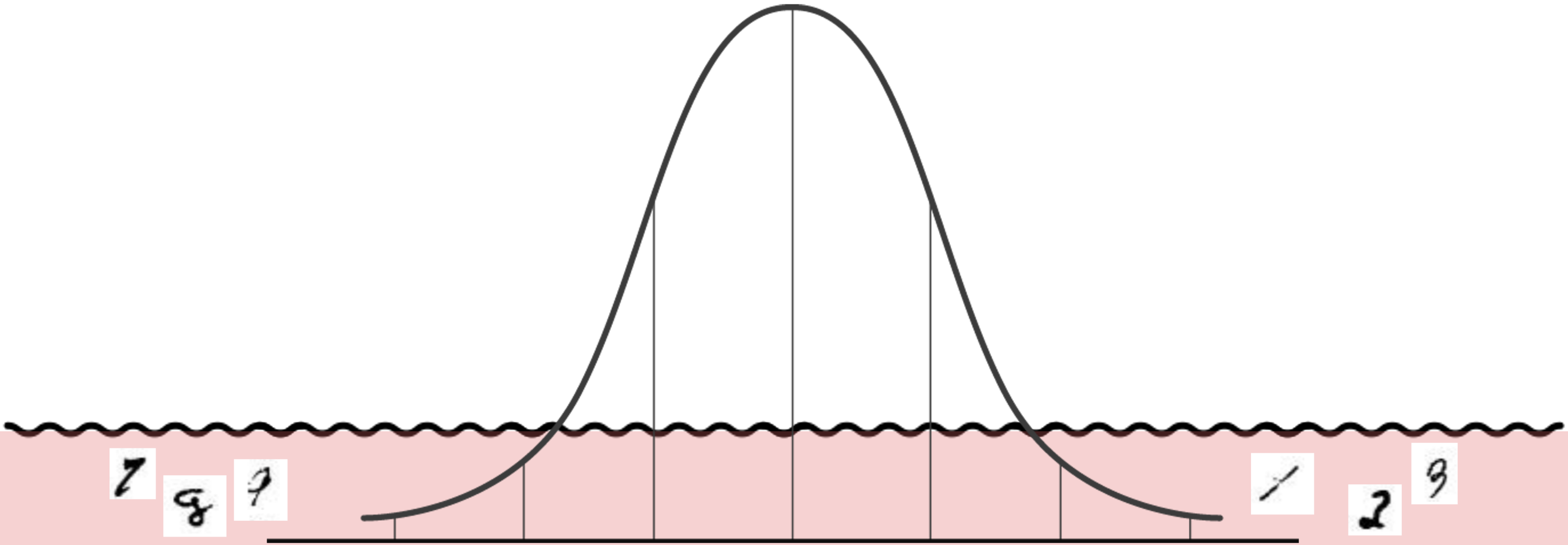
DP-SGD with tanh does **not** lead to exploding activations

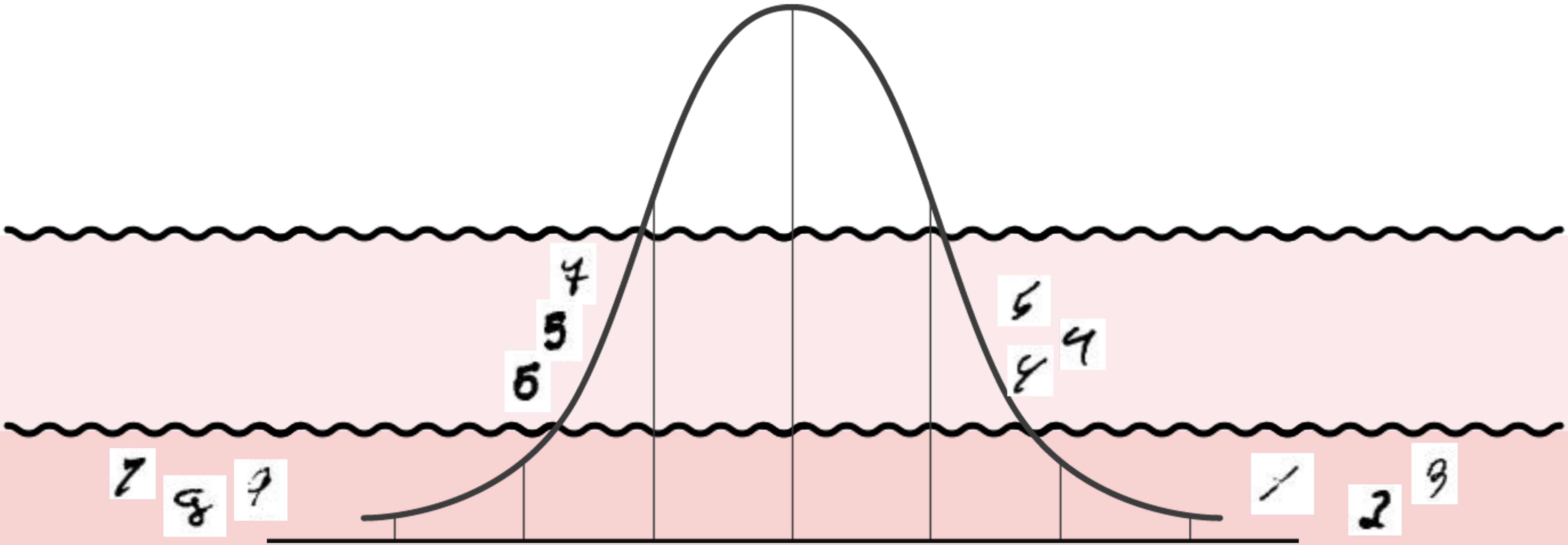


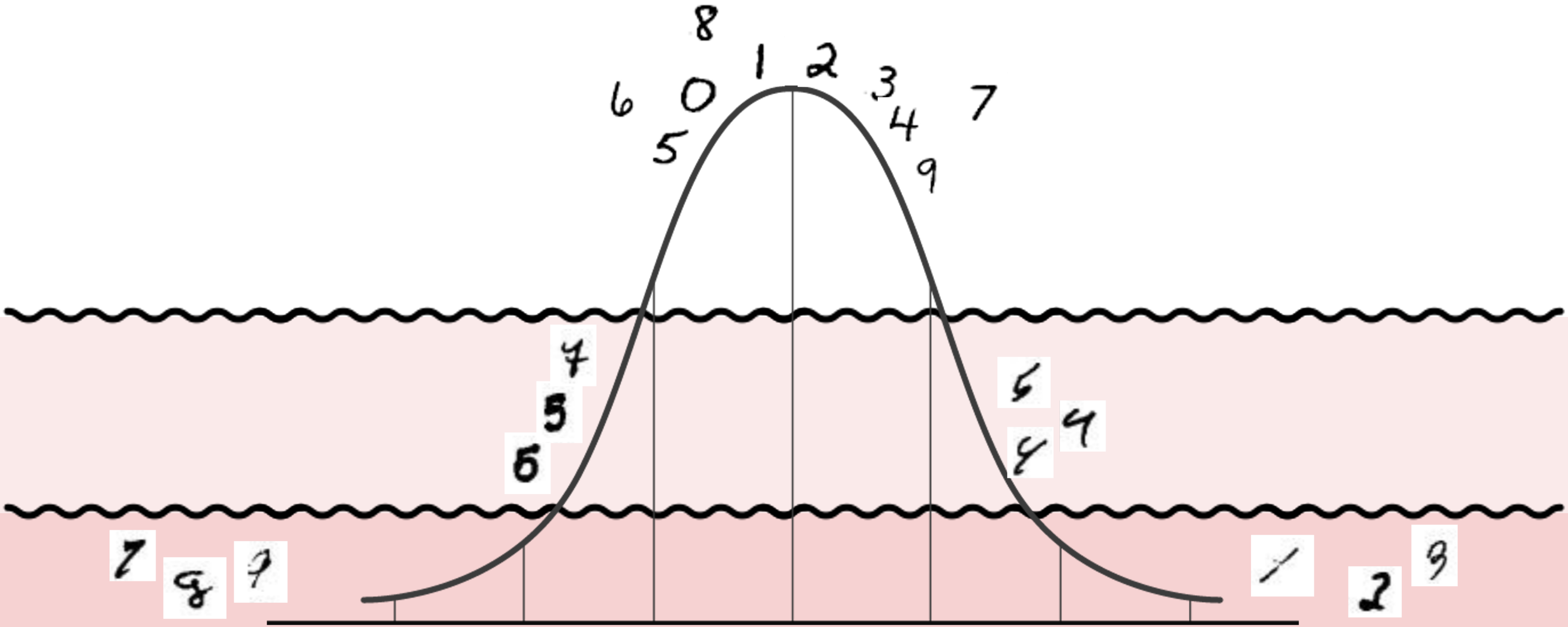
Improving the DP-SGD state-of-the-art with tanh

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







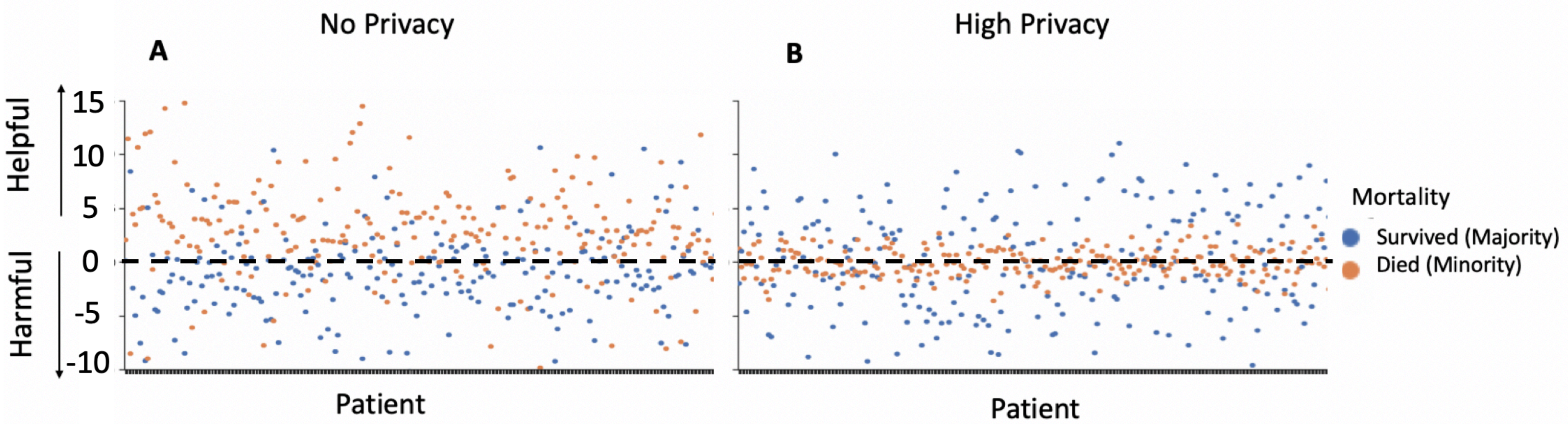
8
6 0 1 2 3 4 7
5 9

Tension between differential privacy and fairness

Utility on Long Tailed Datasets

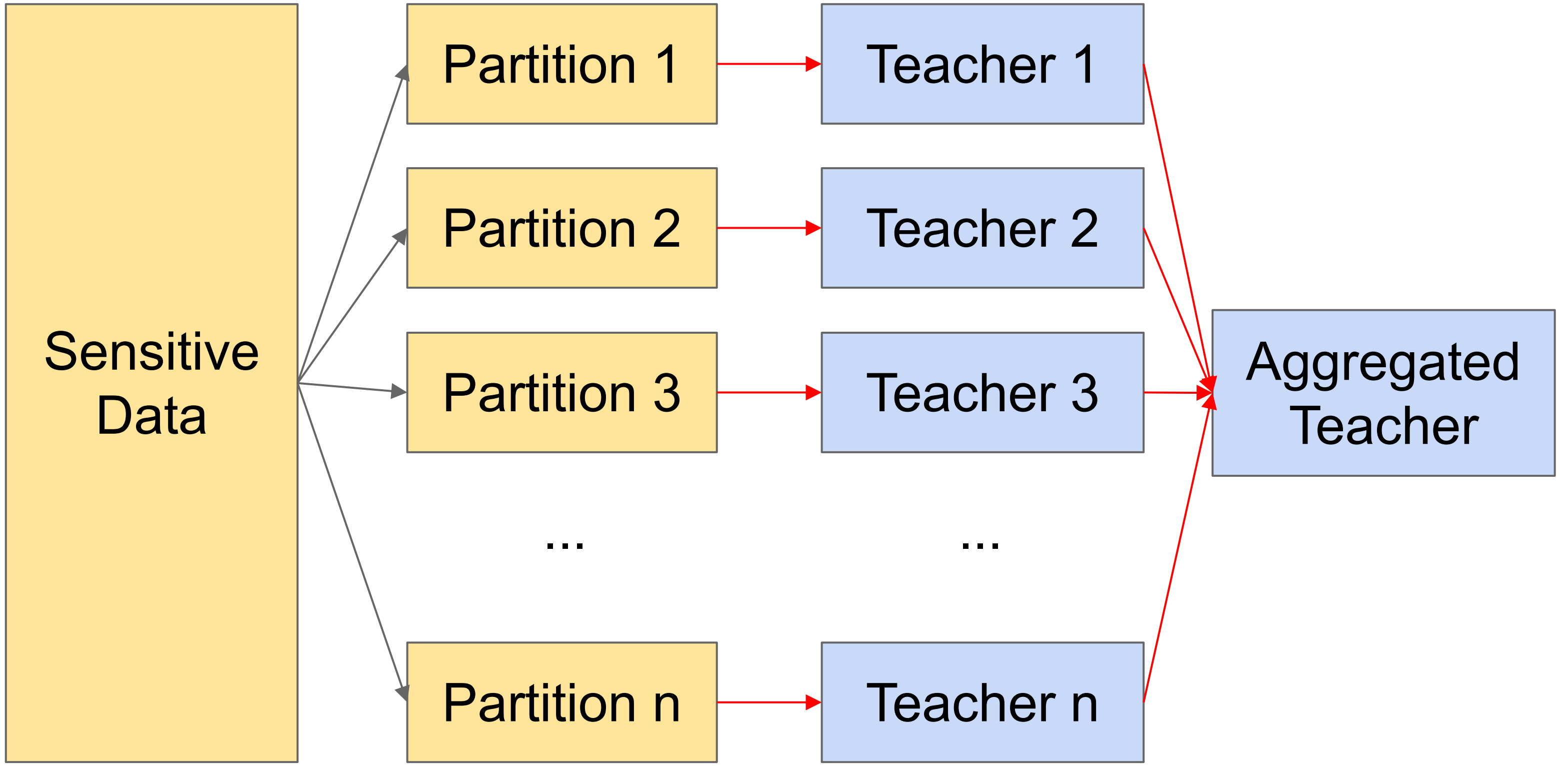
Task	Model	No Privacy	High Privacy (ε, δ)
MIMIC-III Mortality	Logistic Regression	**0.82 ± 0.03**	**0.60 ± 0.04 (3.54, 10^{-5})**
	GRU-D	**0.79 ± 0.03**	**0.53 ± 0.03 (2.65, 10^{-5})**
NIH Chest X-Ray Disease Prediction	Finetuned DenseNet-121	**0.84 ± 0.00**	**0.49 ± 0.00 (0.84, 10^{-6})**

Unfairness Due to Overinfluence of Majority Subgroups

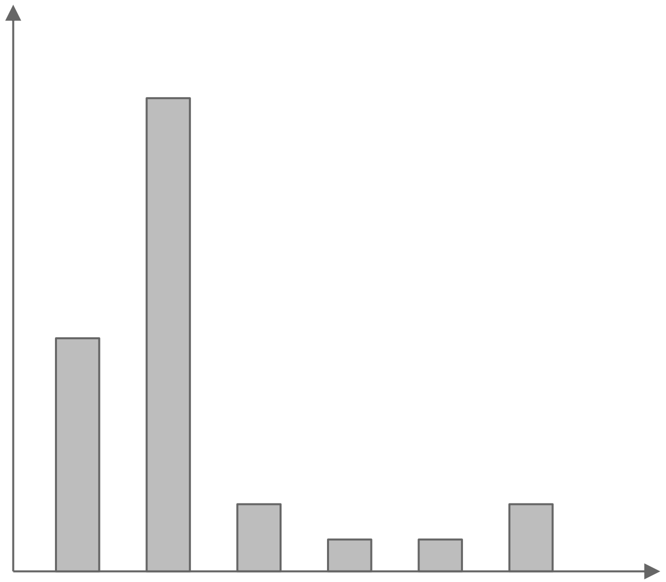


Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. (FAccT 2021)
Vinith Suriyakumar, Nicolas Papernot, Anna Goldenberg, Marzyeh Ghassemi.

A 2nd flavor: PATE aka Private Aggregation of Teacher Ensembles

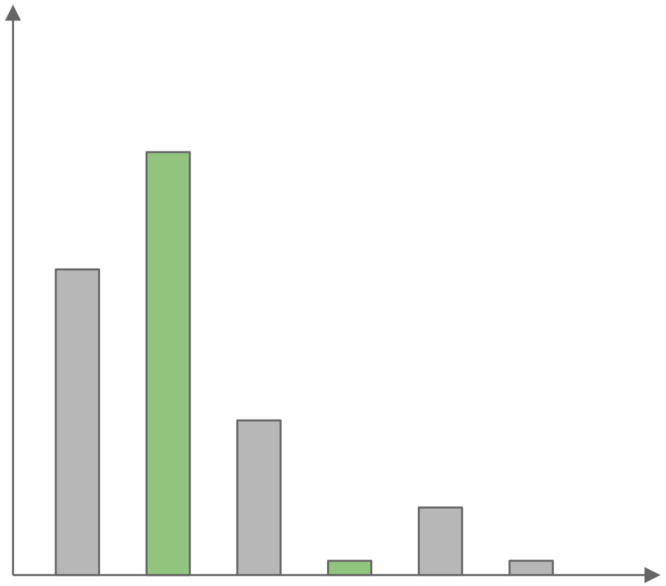


PATE: Private Aggregation of Teacher Ensembles



Count votes

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

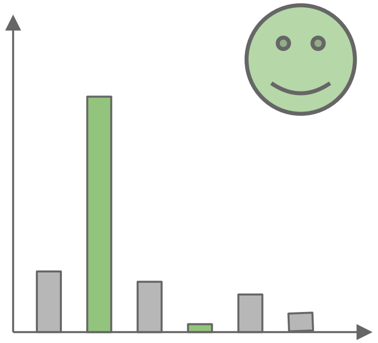


Take maximum

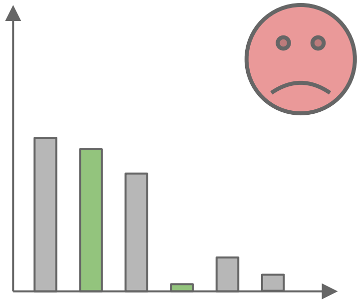
$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) \right\}$$

PATE: Private Aggregation of Teacher Ensembles

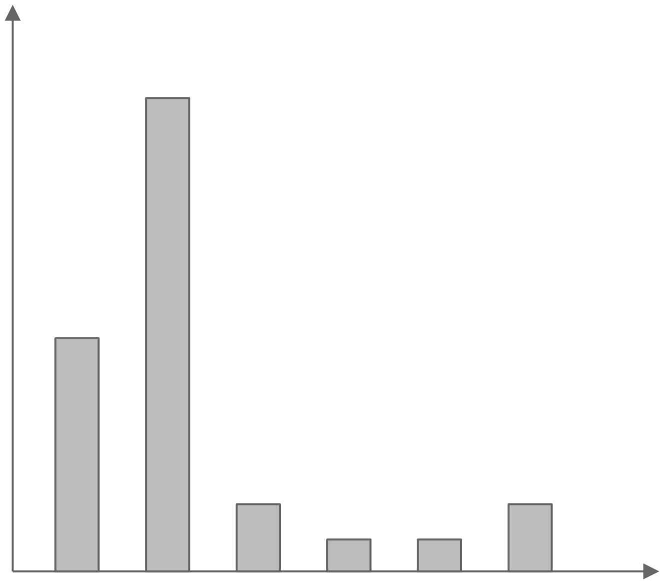
If most teachers agree on the label,
it does not depend on specific partitions,
so the privacy cost is small.



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

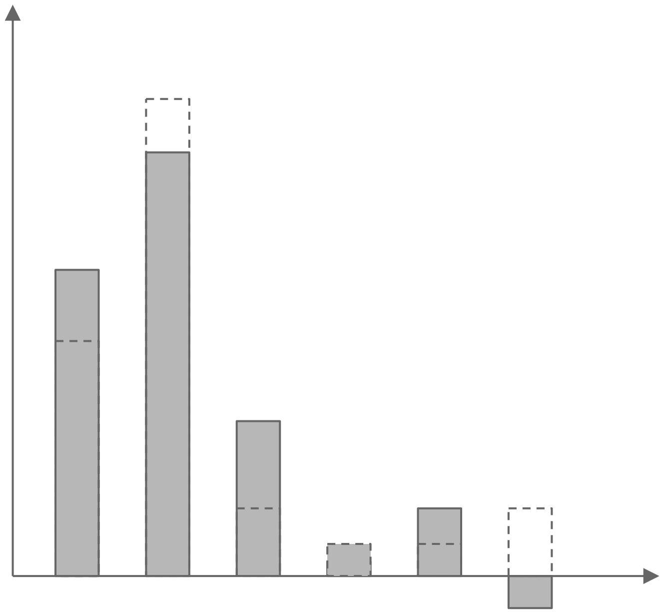


PATE: Private Aggregation of Teacher Ensembles



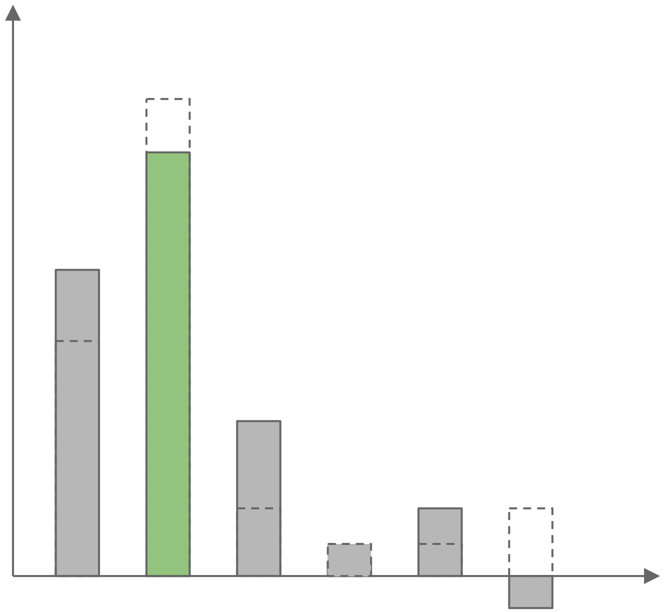
Count votes

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



Add Laplacian

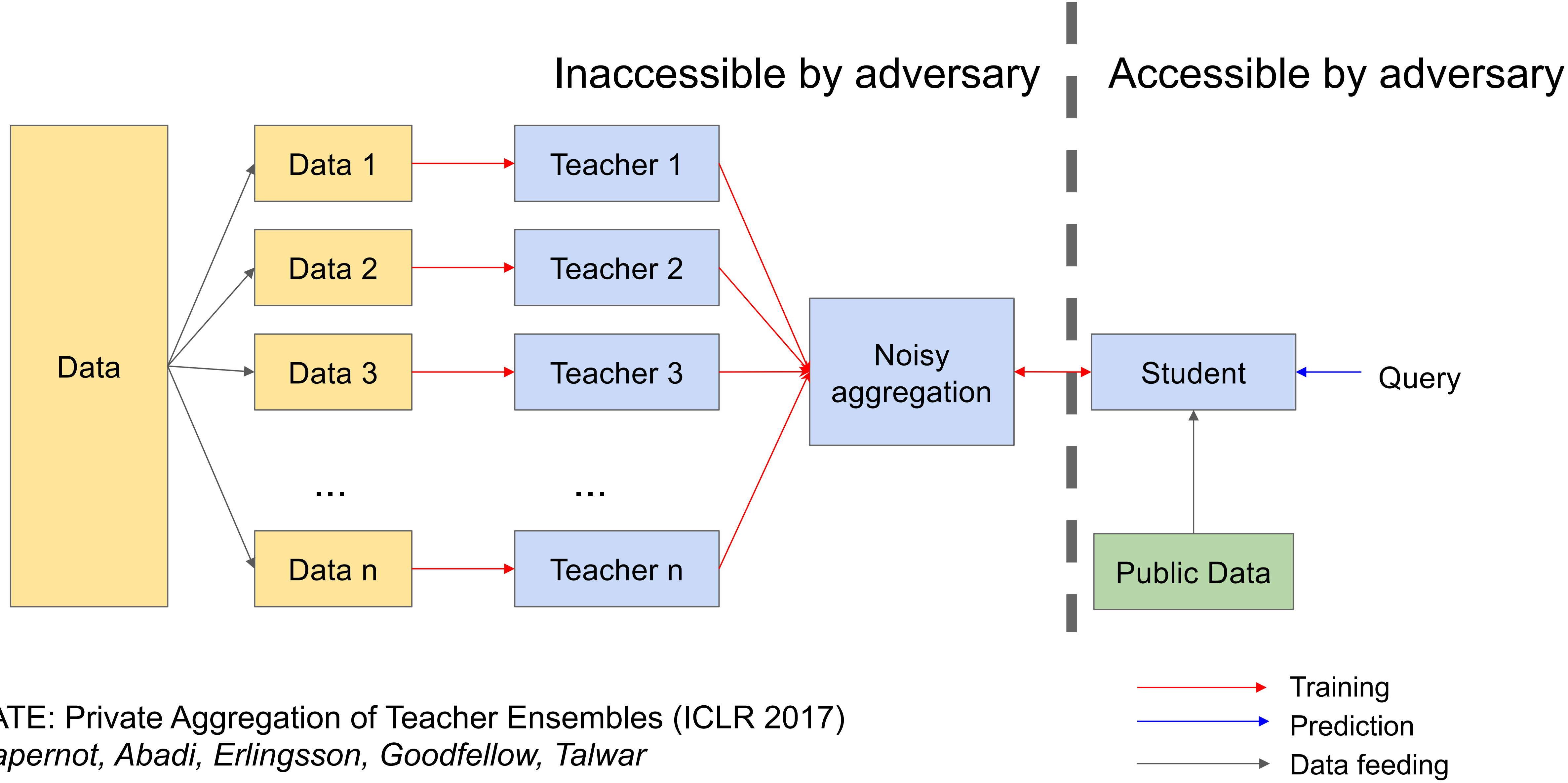
$$Lap\left(\frac{1}{\varepsilon}\right)$$



Take maximum

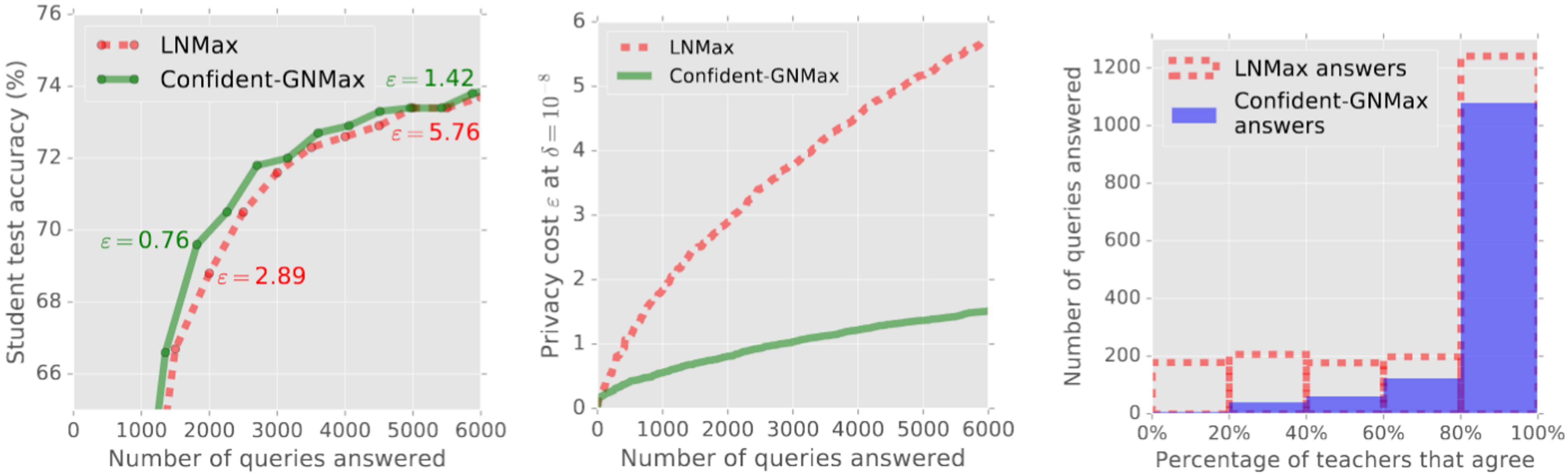
$$f(x) = \arg\max_j \left\{ n_j(\vec{x}) + Lap\left(\frac{1}{\varepsilon}\right) \right\}$$

PATE: Private Aggregation of Teacher Ensembles



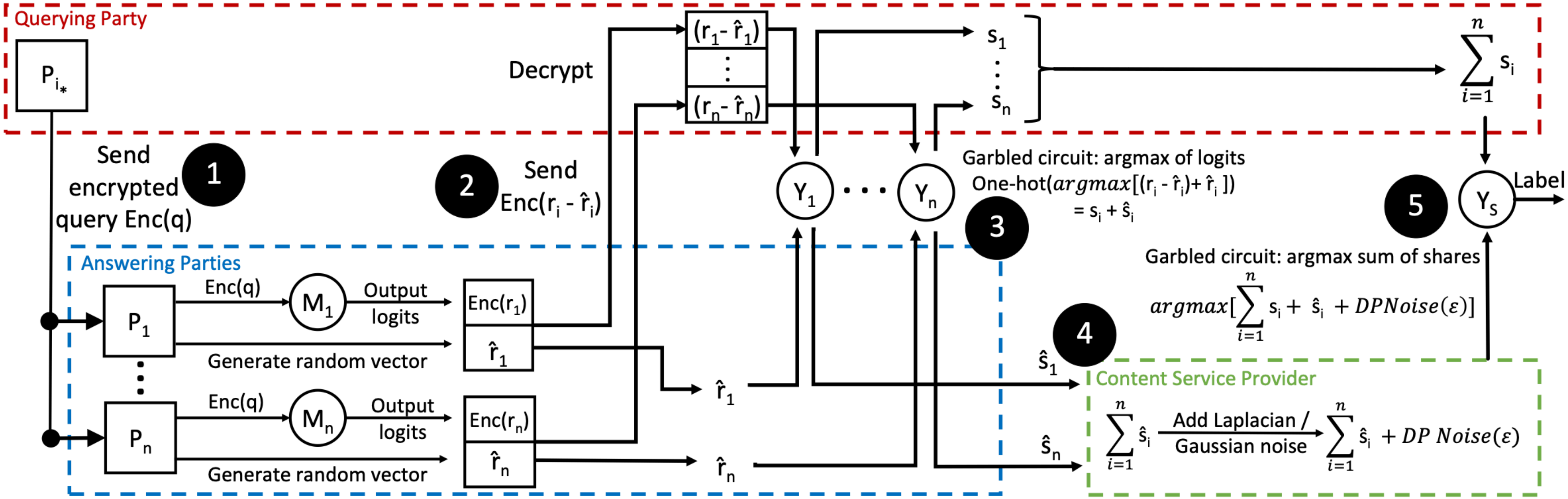
PATE: Private Aggregation of Teacher Ensembles (ICLR 2017)
Papernot, Abadi, Erlingsson, Goodfellow, Talwar

Aligning privacy with generalization



Scalable Private Learning with PATE (Papernot*, Song* et al., ICLR 2018)

A third flavor: Confidential and Private Collaborative Learning



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

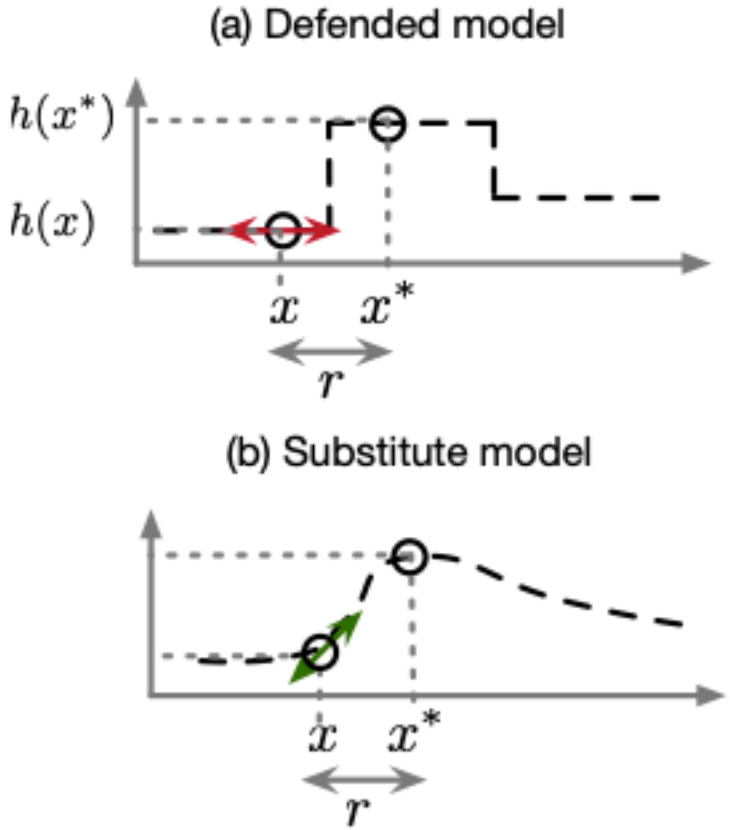
CaPC Learning: Confidential and Private Collaborative Learning (ICLR 2021)
Christopher A. Choquette-Choo, Natalie Dullerud, Adam Dziedzic, Yunxiang Zhang, Somesh Jha, Nicolas Papernot, Xiao Wang

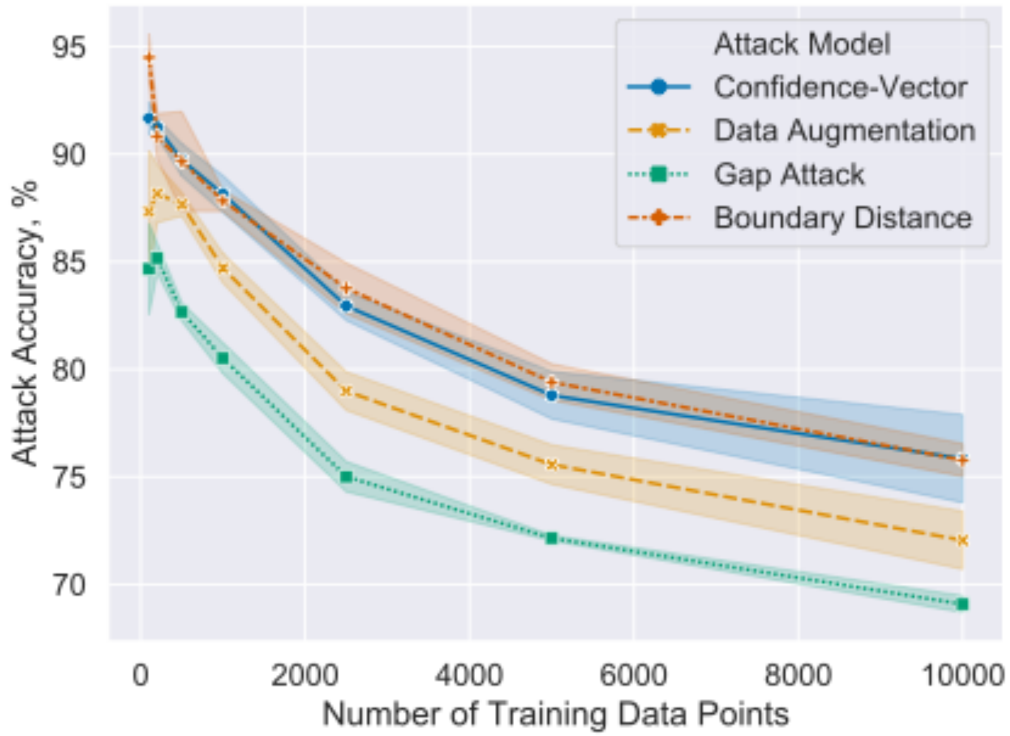
Is achieving trustworthy ML any different from ~~real-world~~ computer security?

*"Practical security balances the cost of protection and the risk of loss, which is the cost of recovering from a loss times its probability" (*Butler Lampson, 2004)

Is the ML paradigm fundamentally different in a way that enables systematic approaches to security and privacy?

Gradient masking vs. Confidence masking

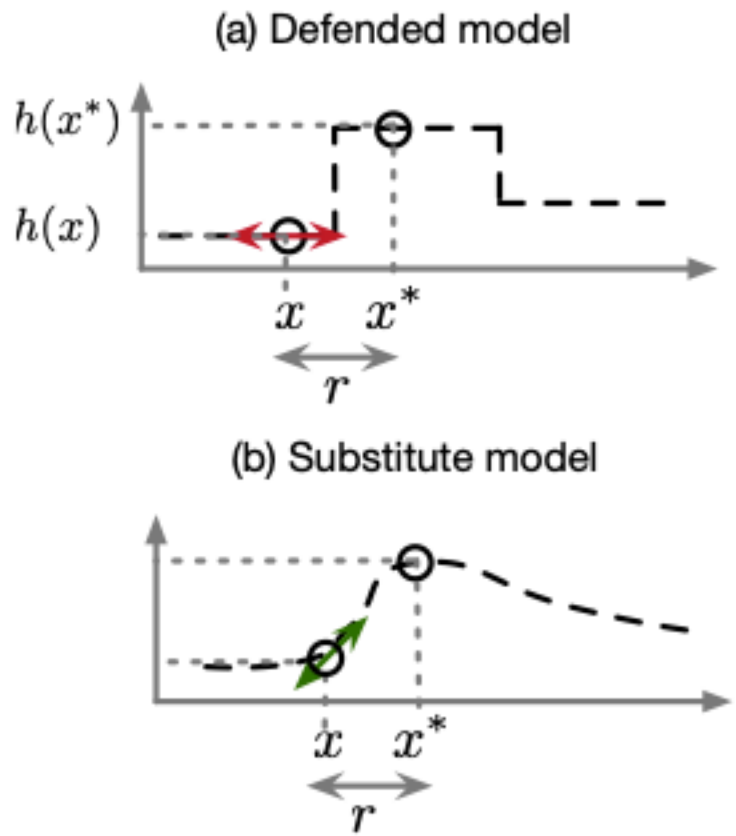




Practical Black-Box Attacks against Machine Learning. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami.

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

Gradient masking vs. Confidence masking



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

Why is differential privacy in ML successful?

- Definition of robustness to adversarial examples using simplistic distances like L_p norms directly conflicts with generalization
- Instead differential privacy encourages generalization

1. No necessary trade-off between privacy and ML objective

2. Degrades smoothly to not learning when it cannot be done privately

A fourth flavor of privacy?

Concrete problem: let's say we

- noticed one of our training points was poisoned
- One of our users wants to delete their data

how do we patch a model once we've trained and deployed it?

-> machine "unlearning"

GDPR

Consumer Privacy Act

PIPEDA

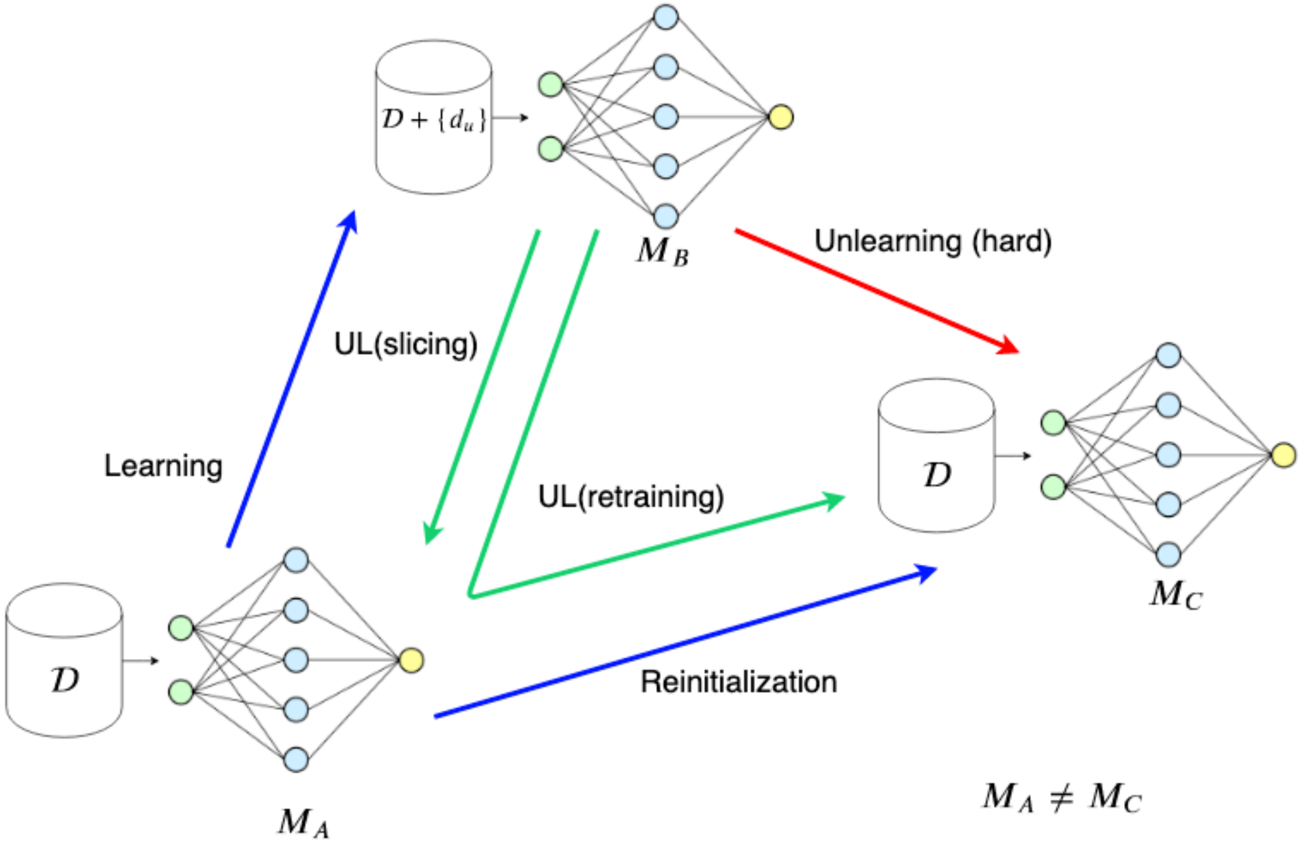
Is differentially private training enough?

- Not really: differentially private training only bounds how much we've learned from each training example
- If we wanted to use differential privacy, we would have to set epsilon to 0.

Why is machine unlearning difficult?

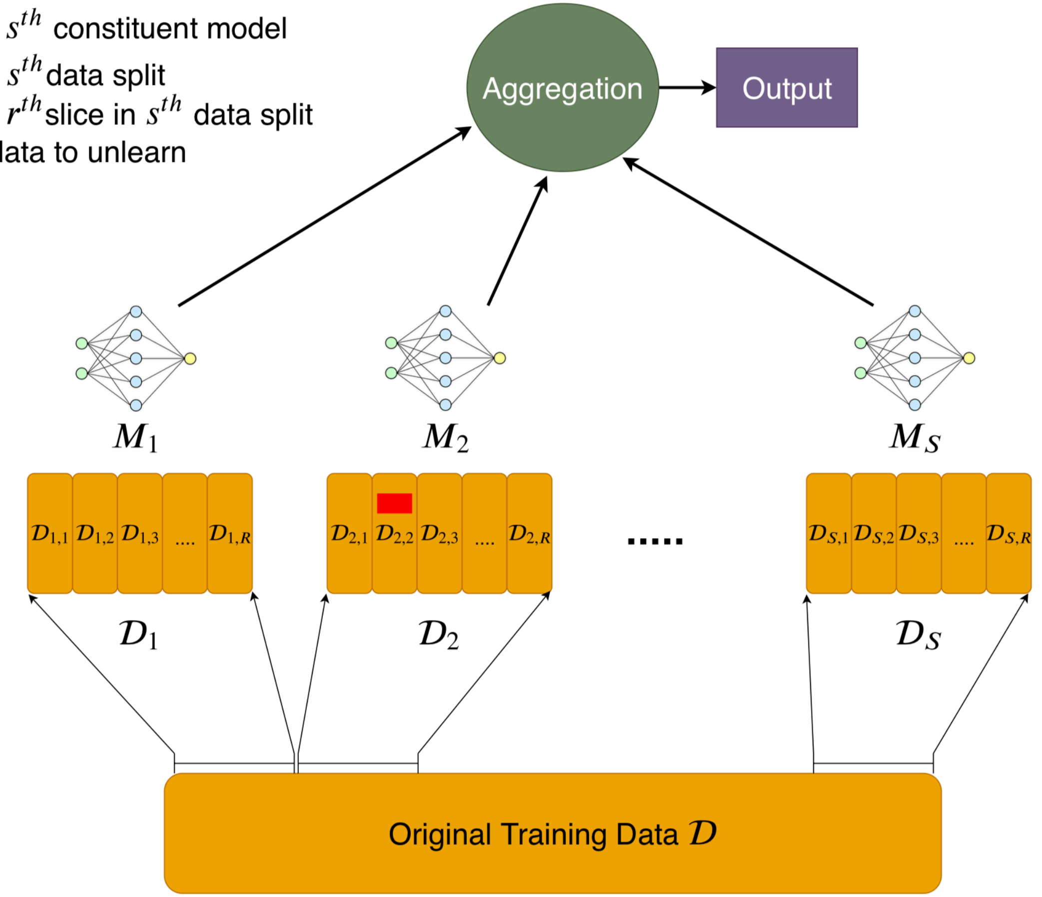
- Difficult to estimate influence of each training example on parameters and predictions
- Stochasticity in
  - training algorithms: batch sampling, …
  - learning itself: multiple minima
- Training is incremental

What is machine unlearning?



Sharded Isolated Sliced Aggregated Training

- M_s : s^{th} constituent model
- $\mathcal{D}_s$: s^{th} data split
- $\mathcal{D}_{s,r}$: r^{th} slice in s^{th} data split
- : data to unlearn



Machine Unlearning. (IEEE SP 2021)

Lucas Bourtoule, Varun Chandrasekaran, Christopher Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot

Resources:
cleverhans.io
github.com/cleverhans-lab/cleverhans
github.com/tensorflow/privacy

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Private ML is an opportunity to make ML better

?

Contact information:
nicolas.papernot@utoronto.ca
@NicolasPapernot

I'm hiring at UofT & Vector:

- Students and postdocs
- Faculty positions at all ranks

Resources:
cleverhans.io
github.com/cleverhans-lab/cleverhans
github.com/tensorflow/privacy

UNIVERSITY OF TORONTO

VECTOR INSTITUTE


Adam Dziedzic


Adelin Travers


Ilia Shumailov


Lucas Bourtoule


Mohammad Yaghini


Natalie Dullerud


Nick Jia


Tejumade Afonja


Varun Chandrasekaran


Vinith Suriyakumar

Contact information:
nicolas.papernot@utoronto.ca
@NicolasPapernot

I’m hiring at UofT & Vector:
- Students and postdocs
- Faculty positions at all ranks