Lecture 7: Differential Privacy

Oct 28

Prof. Nicolas Papernot

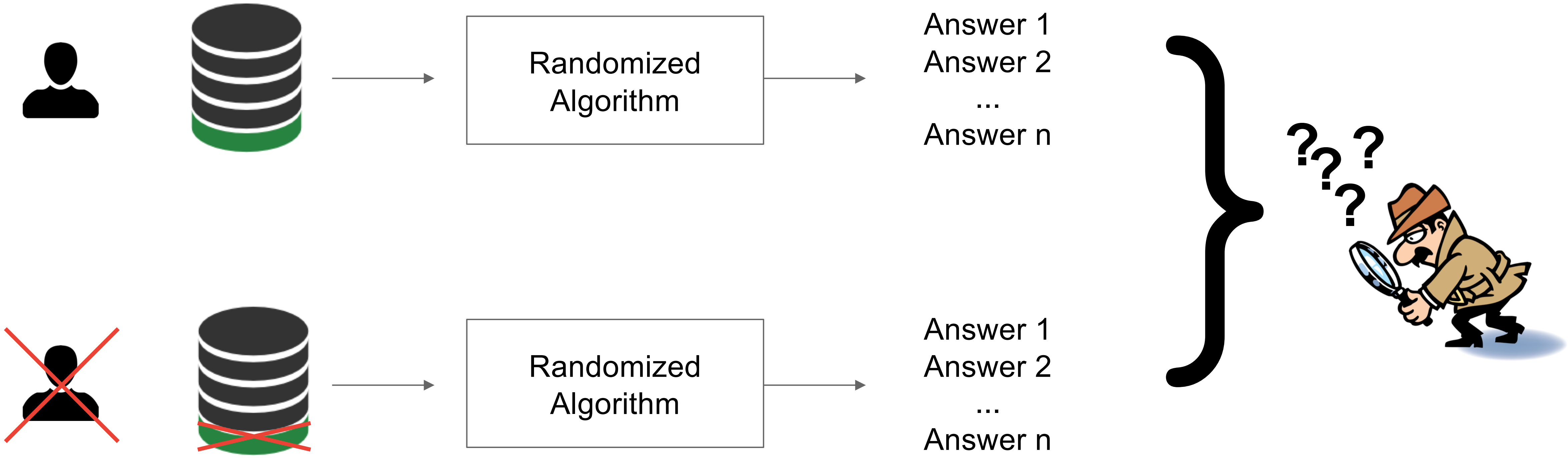
	Adversarial goal	Attack / defense example	Data owners	Model host	Consumers
Training	Data integrity	Data poisoning *(Koh and Liang, 2017)*	●		
	Model integrity	Backdoor *(Gu et al., 2017)*	●		
	Data confidentiality	Federated learning *(McMahan, 2017)*		●	
	Data privacy	RAPPOR *(Erlingsson, 2014)*		●	●
Inference	Model integrity	Adversarial examples *(Szegedy et al., 2013)*			●
	Data confidentiality	CryptoNets *(Dowlin et al., 2016)*		●	
	Model confidentiality	Model extraction *(Tramer et al., 2016)*			●
	Data Privacy	Membership inference *(Shokri et al., 2017)*			●

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)

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

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)

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

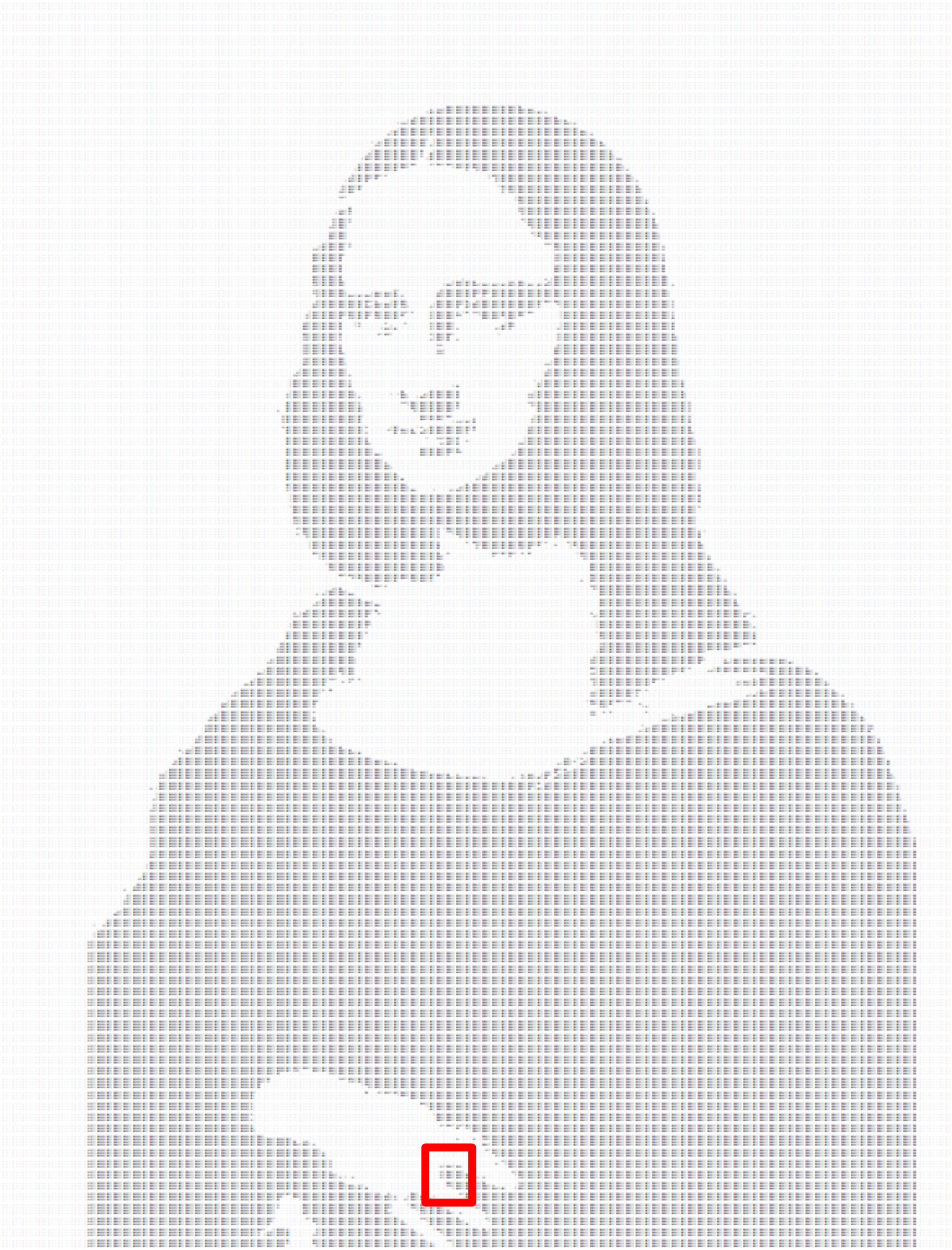
- Robust to composition
  - Algorithm M1 has eps DP
  - Algorithm M2 has eps DP
  - Algorithms M1 and M2 have 2eps DP
- Group guarantees

What does that mean for a user?

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

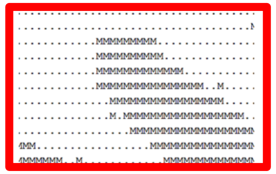
-> *differential* privacy

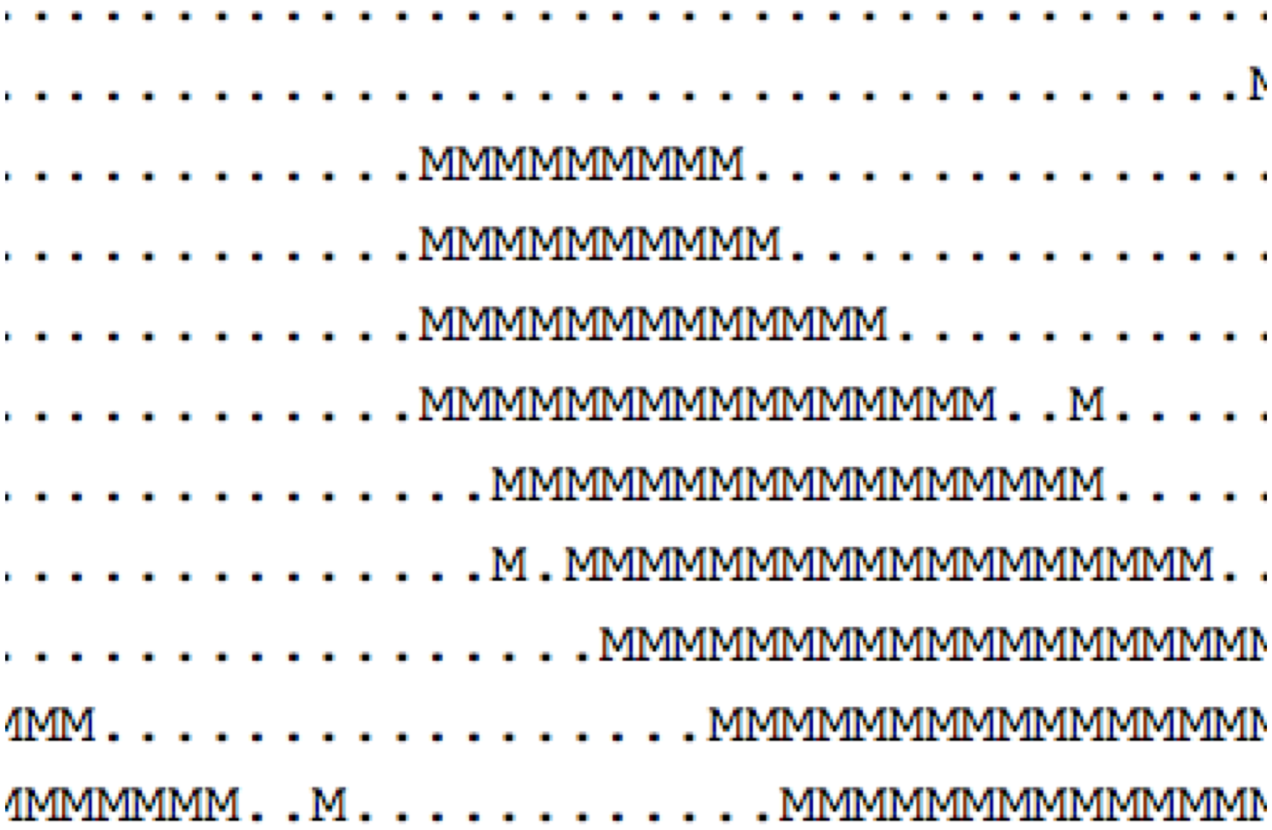
A Metaphor For Private Learning



Slides adapted from Ulfar Erlingsson

An Individual's Training Data





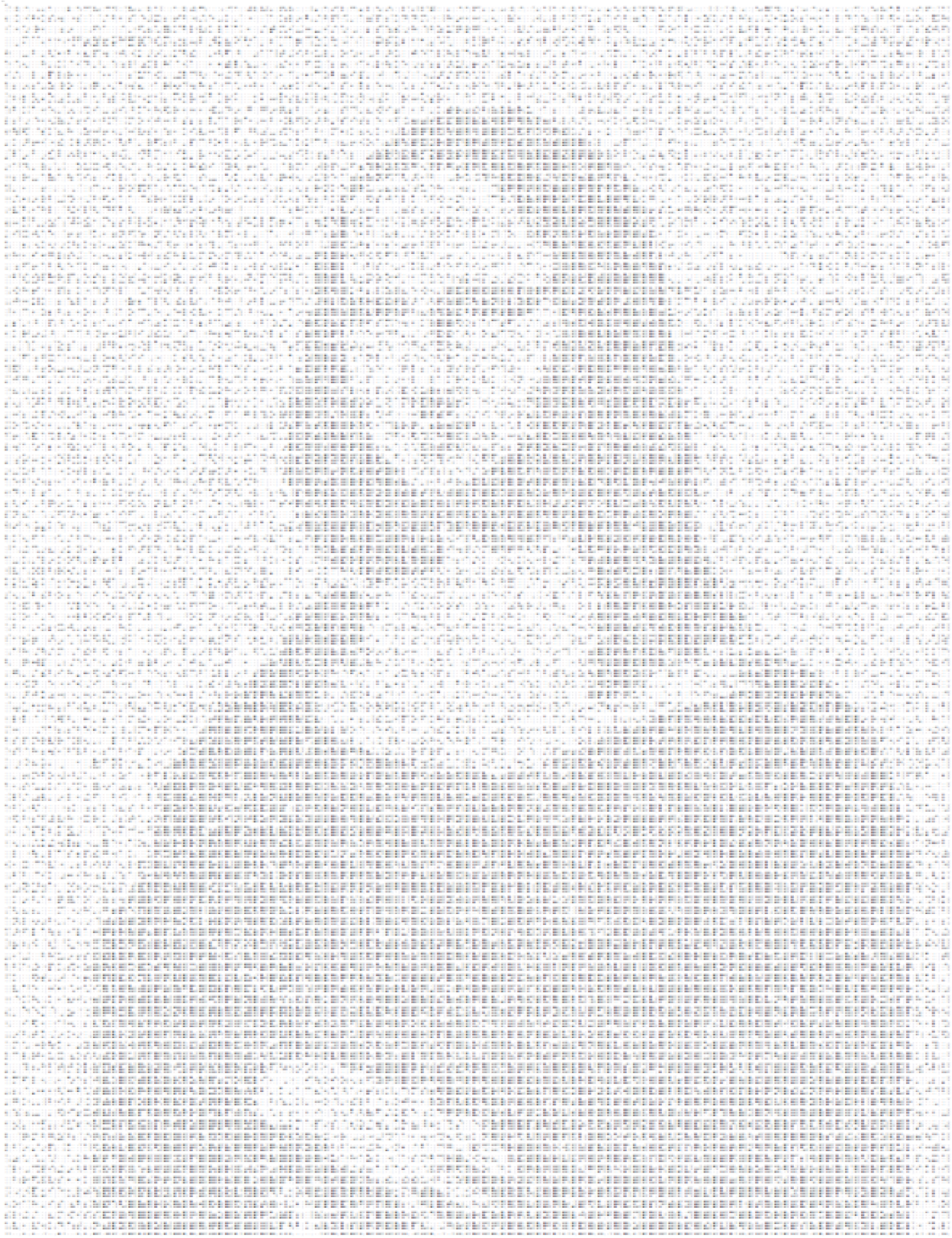
Slides adapted from Ulfar Erlingsson

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
```

Big Picture Remains!



Slides adapted from Ulfar Erlingsson

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 true:
  - 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?

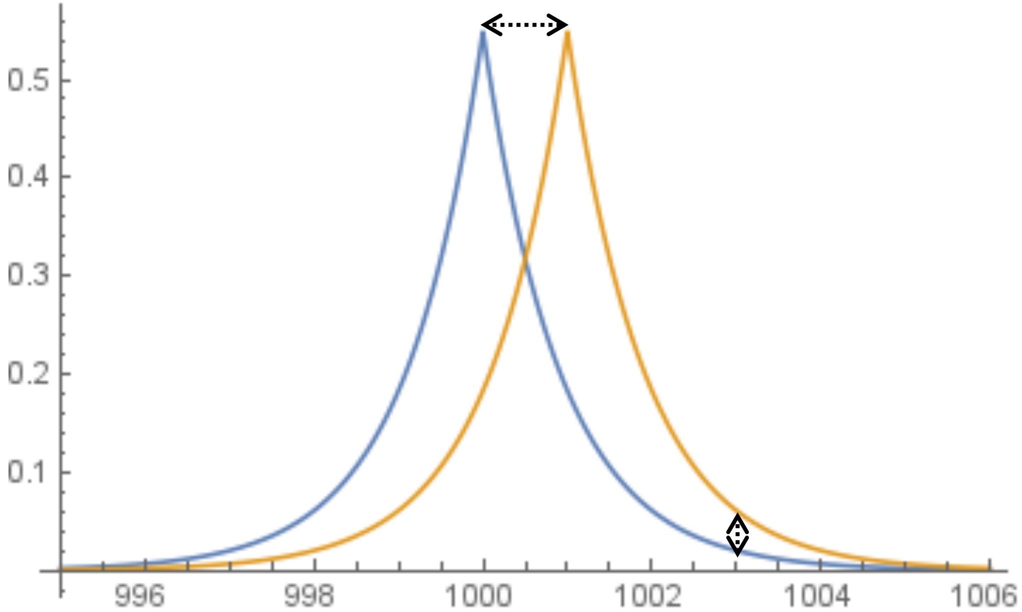
- Assume 10,000 participants
  - 3,000 say they are communist
  - 7,000 say they are not communist

- 50% answers are random so we remove 5,000/2 from each answer pool:
  - 500 are communist
  - 5,500 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

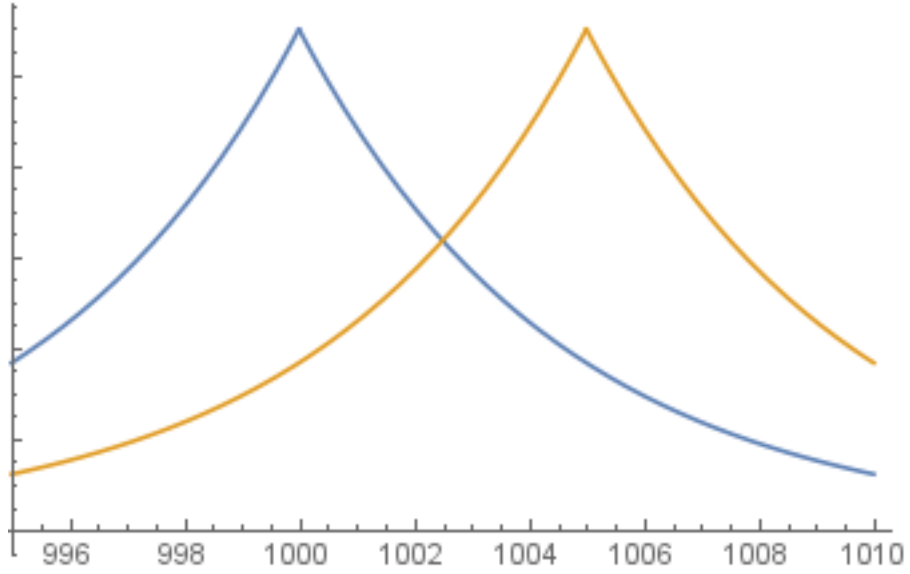
Another example: a privacy-preserving count query

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:

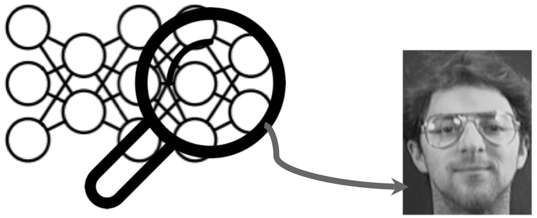
$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$$

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*



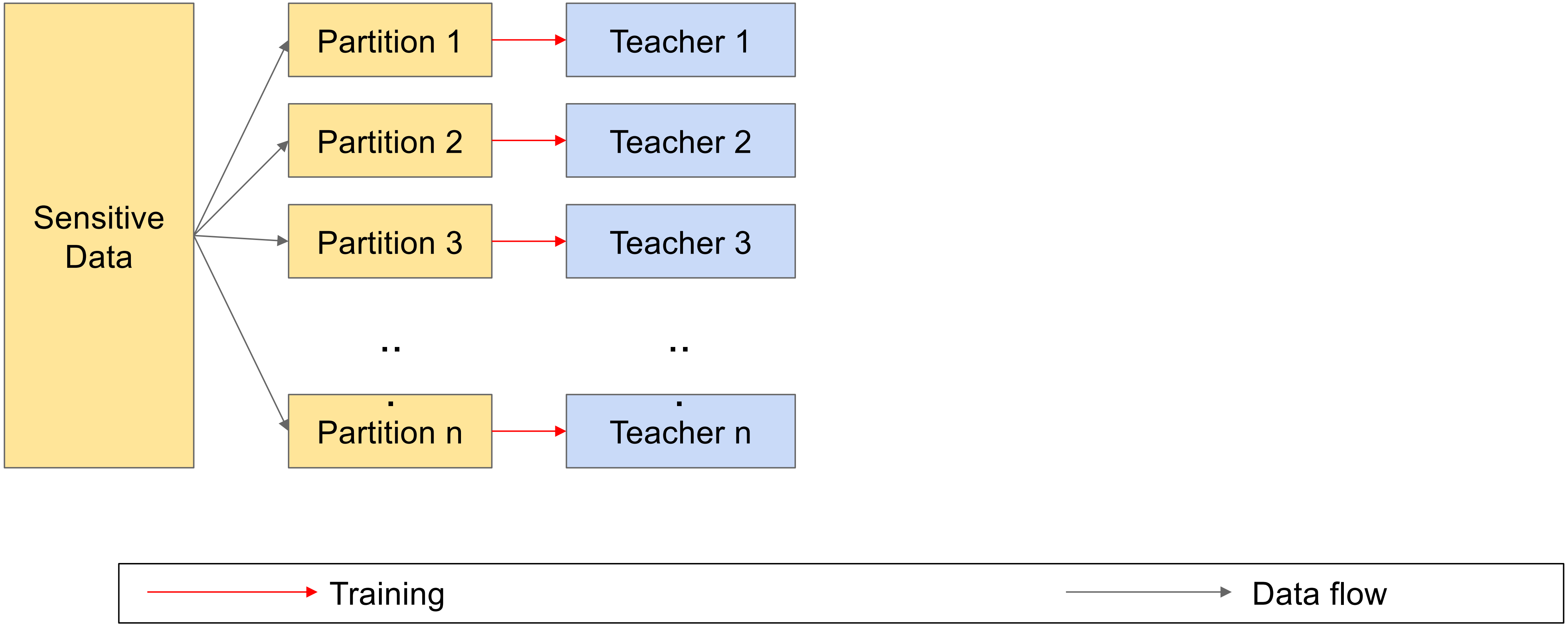
Model inspection (**white-box adversary**)

Zhang et al. (2017) *Understanding DL requires rethinking generalization*

In our work, the threat model assumes:

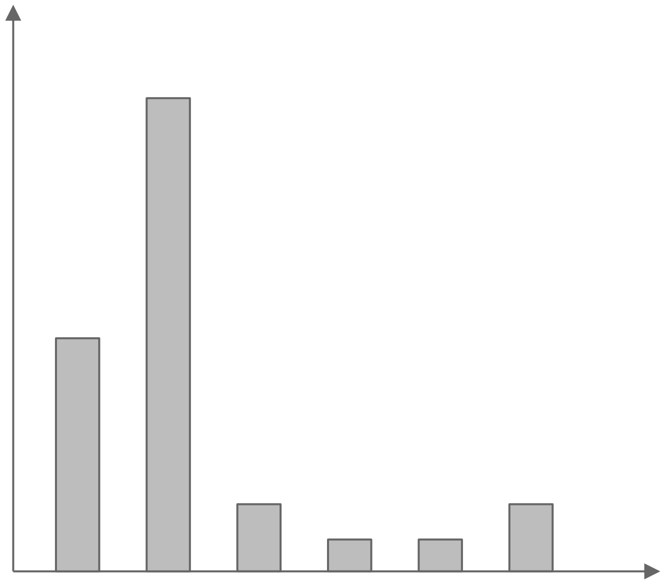
- Adversary can make a potentially unbounded number of queries
- Adversary has access to model internals

Private Aggregation of Teacher Ensembles (PATE)



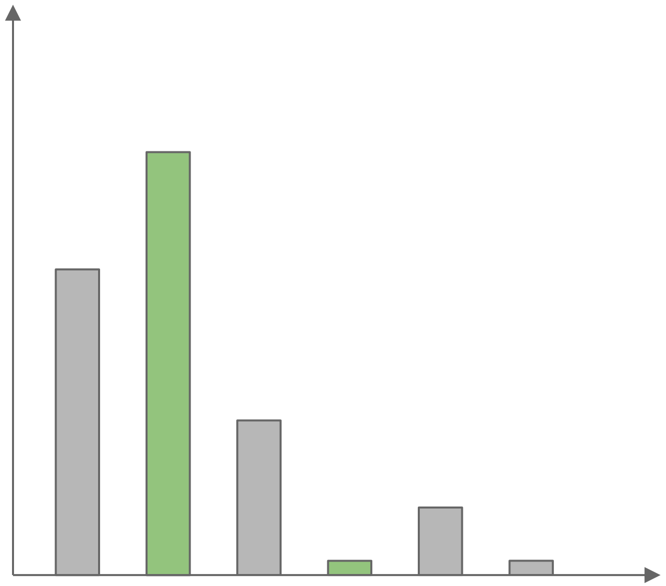
Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data [ICLR 2017 best paper]
Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar

Aggregation



Count votes

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

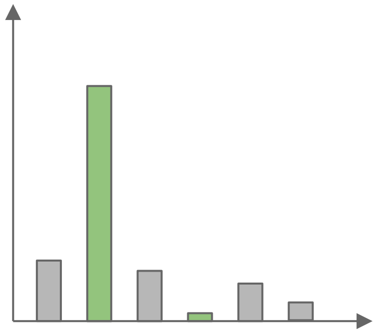


Take maximum

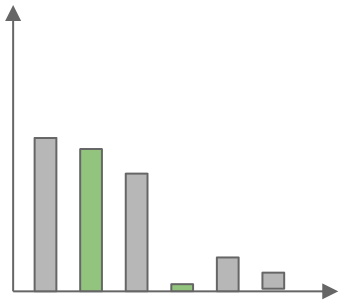
$$f(x) = \arg\max_j \left\{ n_j(\vec{x} \right\}$$

Intuitive privacy analysis

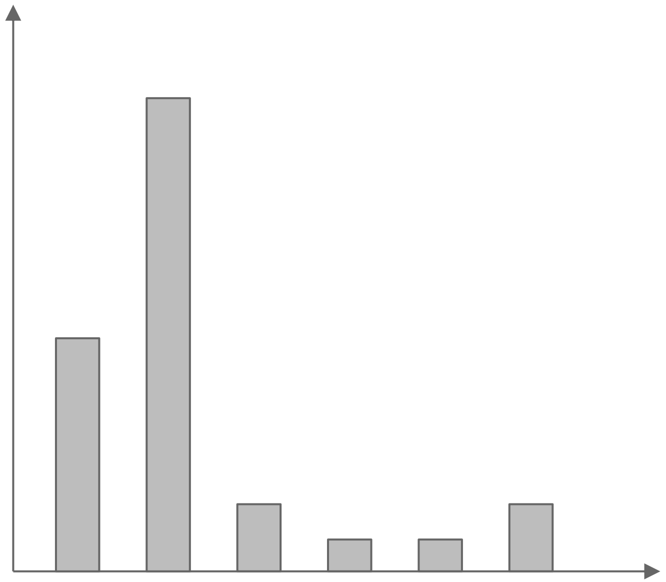
If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.



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

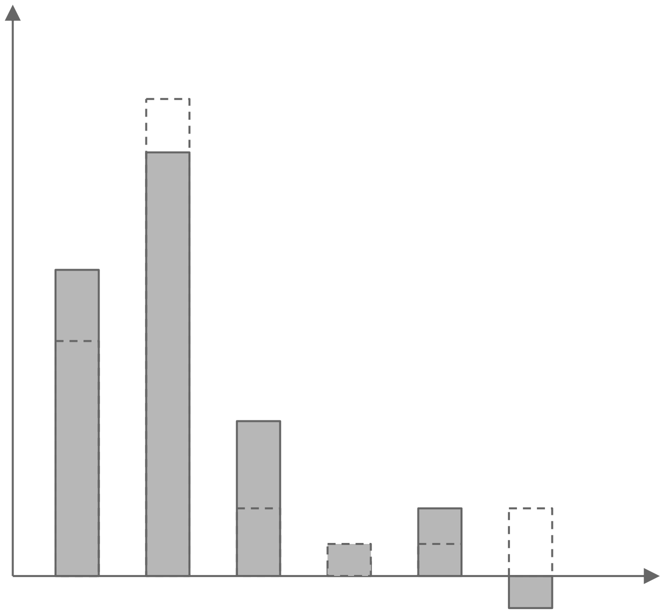


Noisy aggregation



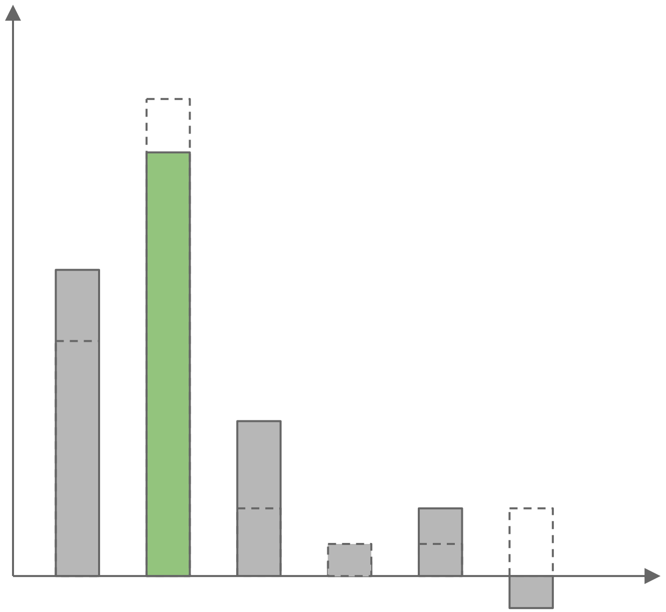
Count votes

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



Add Laplacian

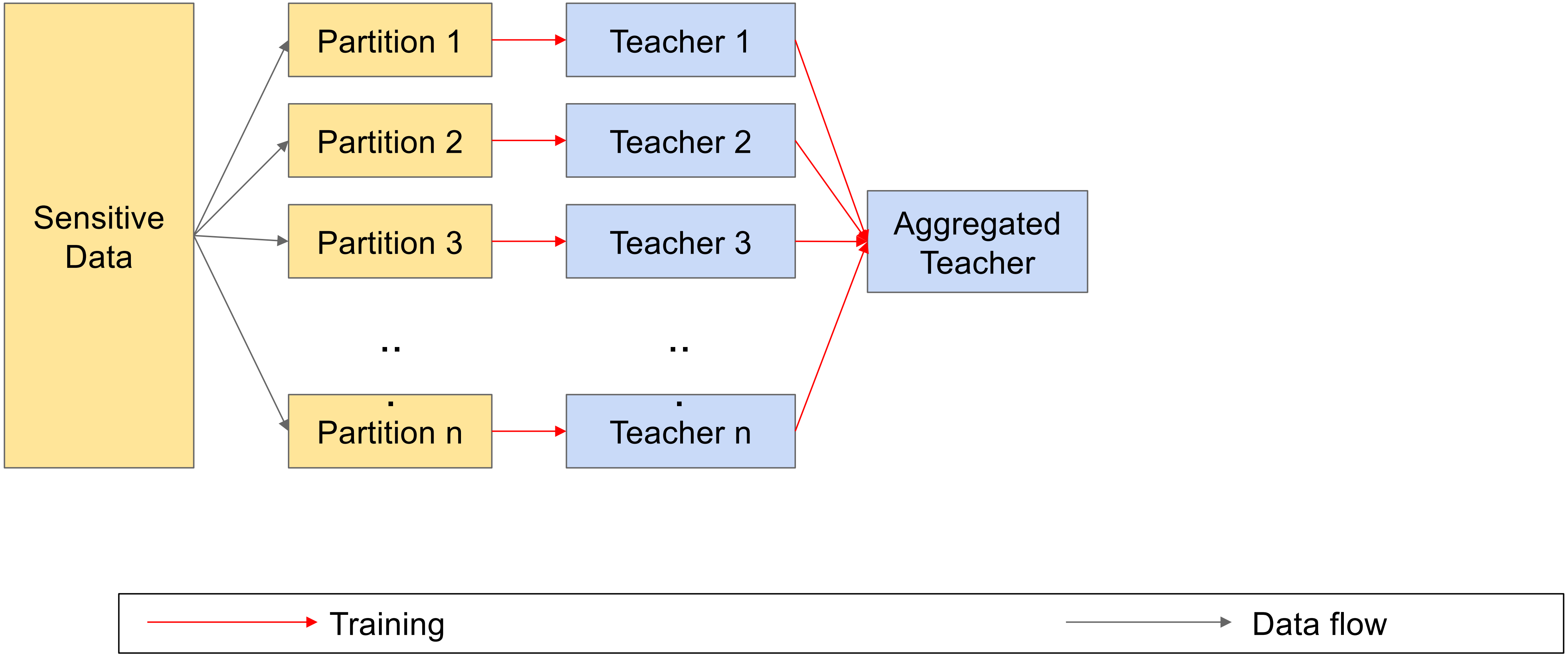
$$Lap\left(\frac{1}{\varepsilon}\right)$$



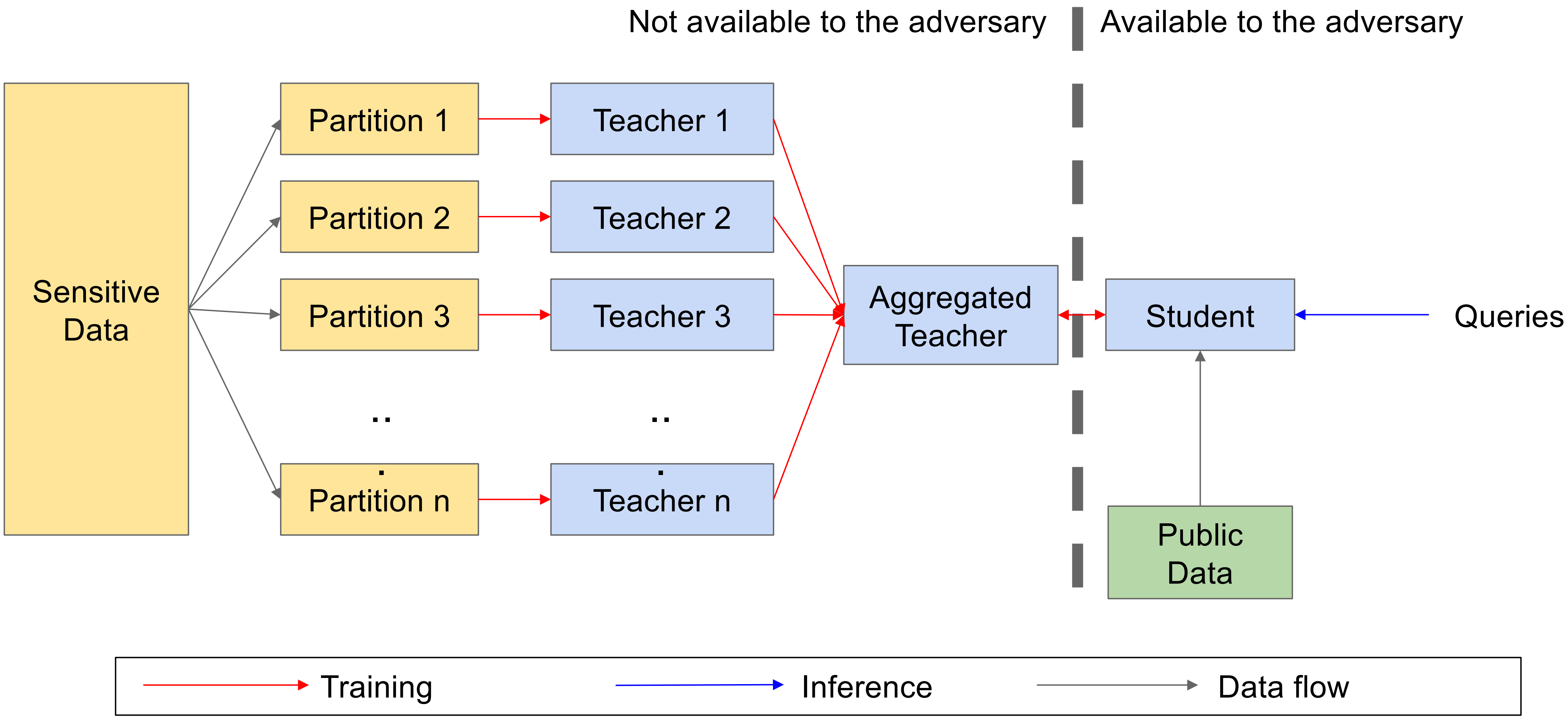
Take maximum

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

Teacher ensemble



Student training



Why train an additional “student” model?

The aggregated teacher violates our threat model:

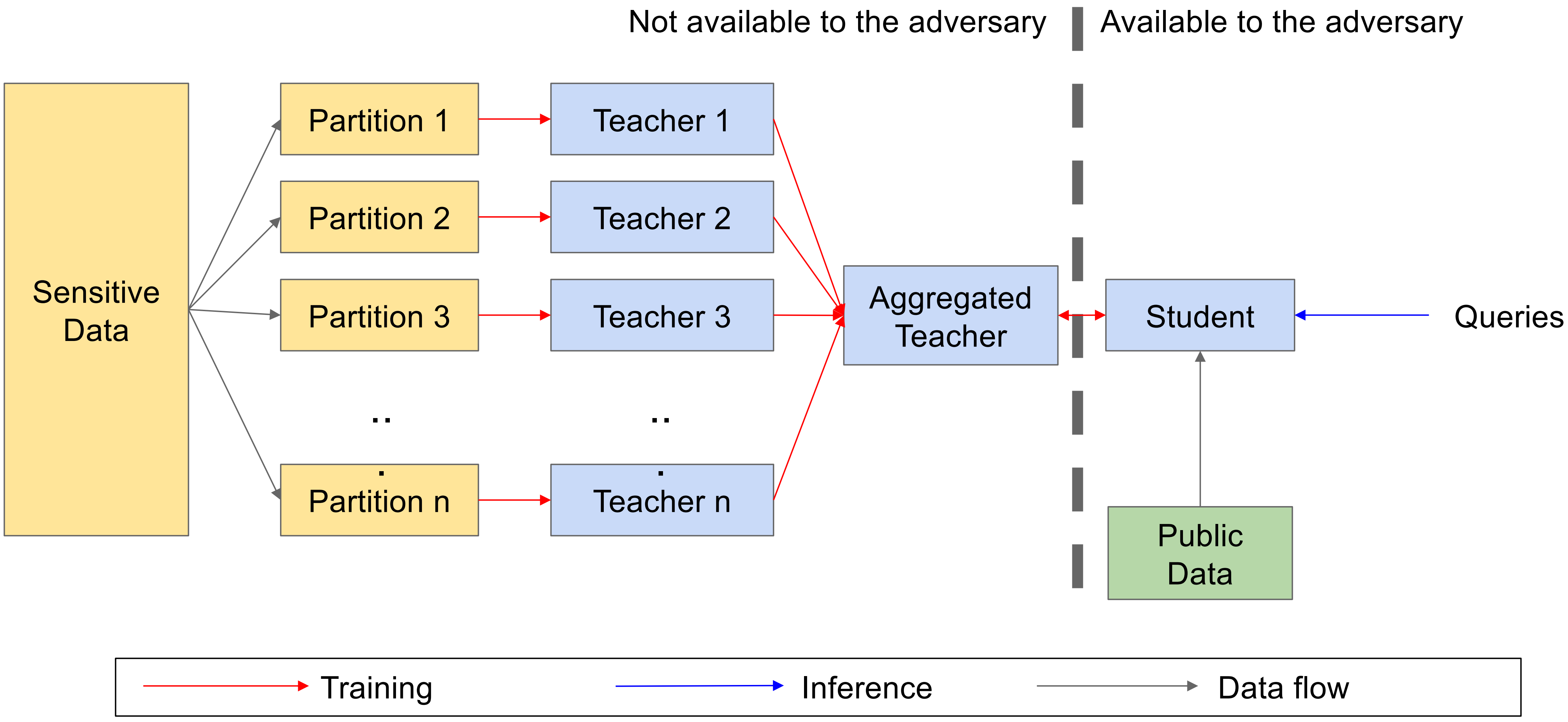
1. **Each prediction increases total privacy loss.**

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

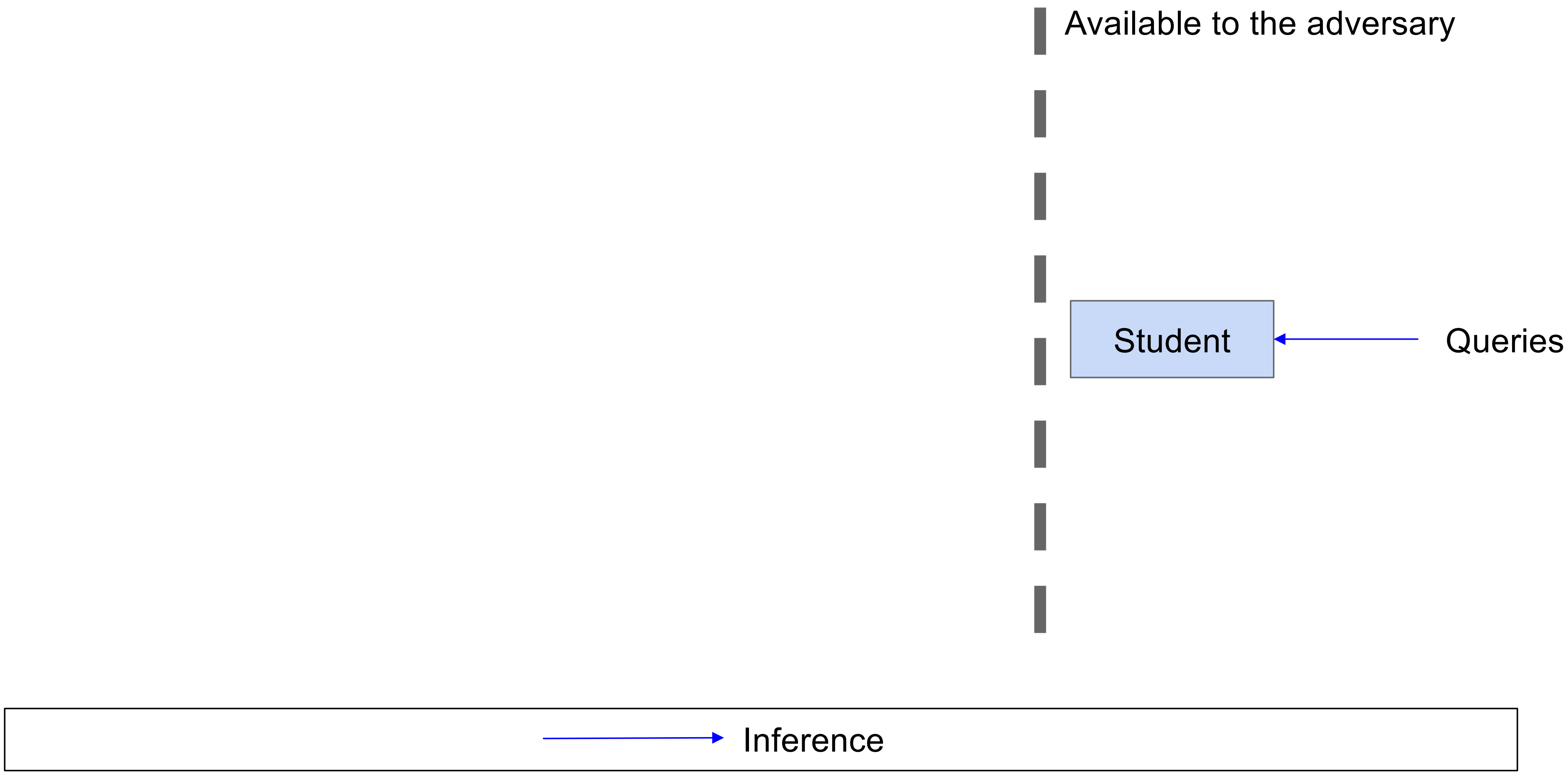
2. **Inspection of internals may reveal private data.**

   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:

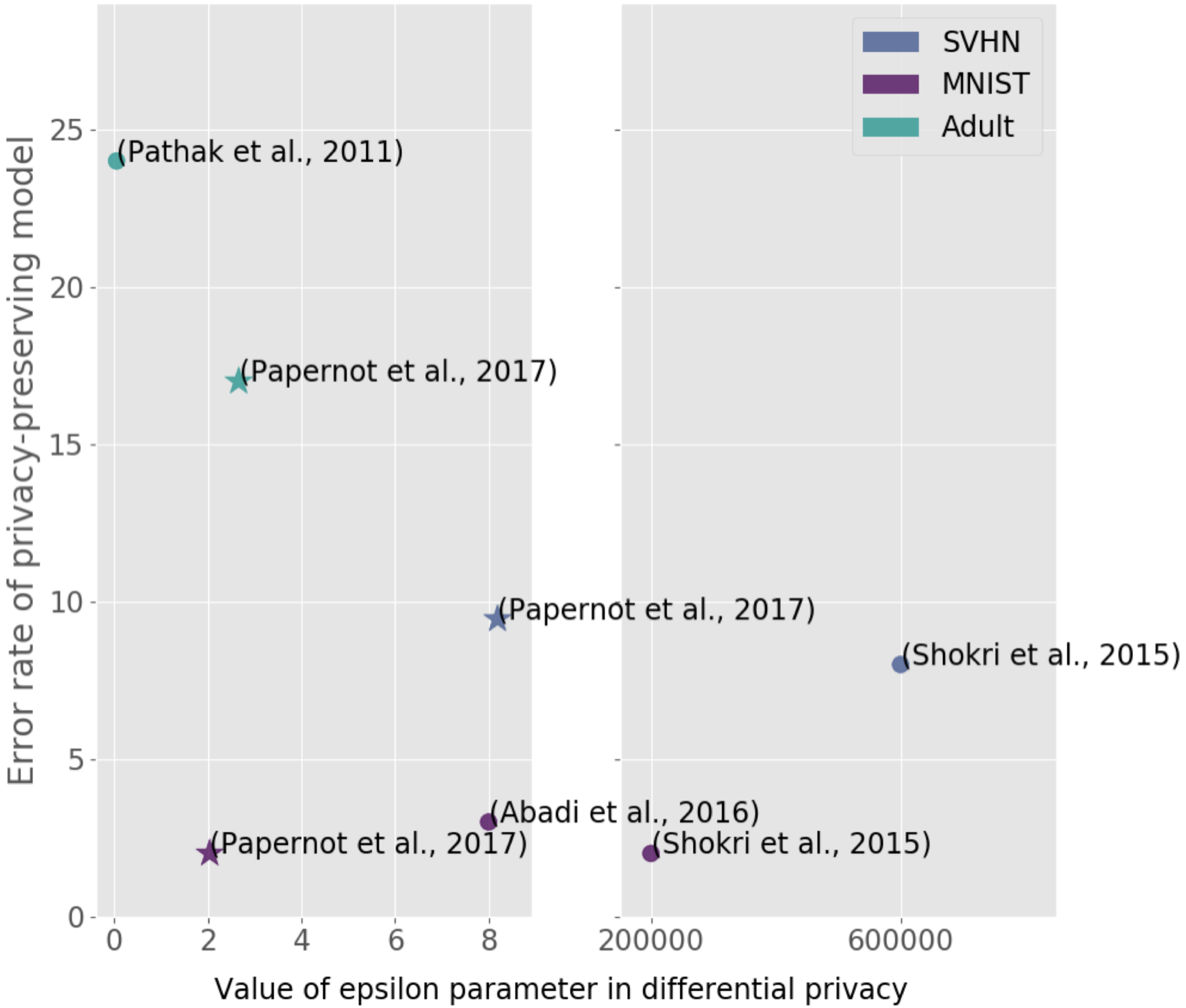
$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$$

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

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

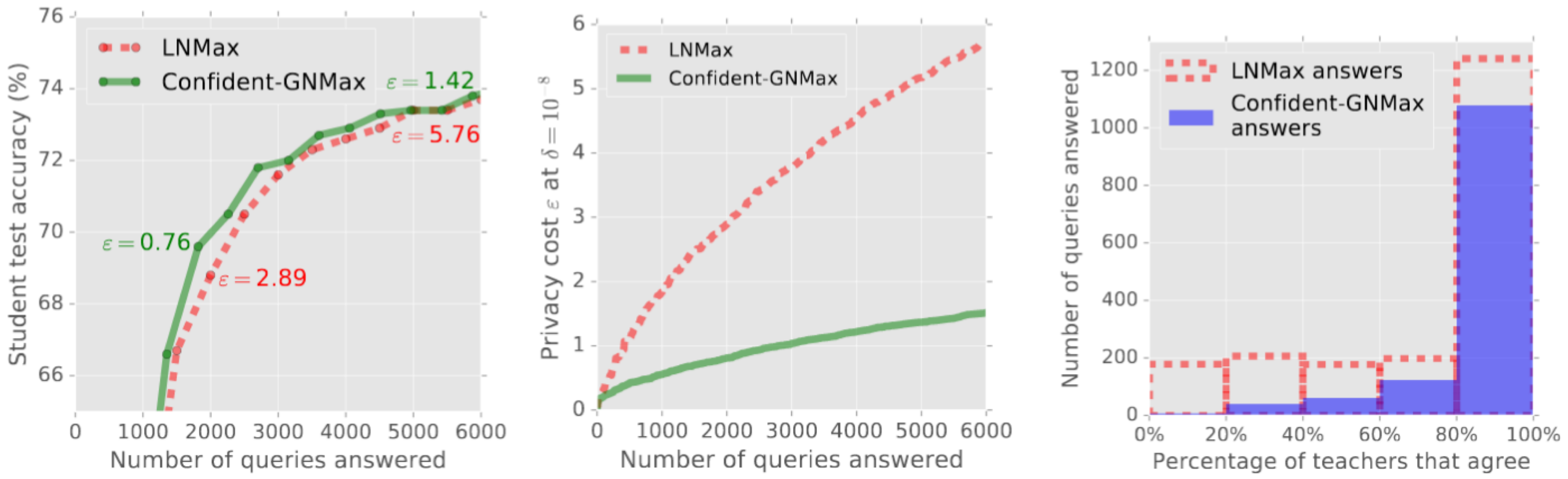
Bound is **data-dependent**: computed using the empirical quorum

Trade-off between student accuracy and privacy



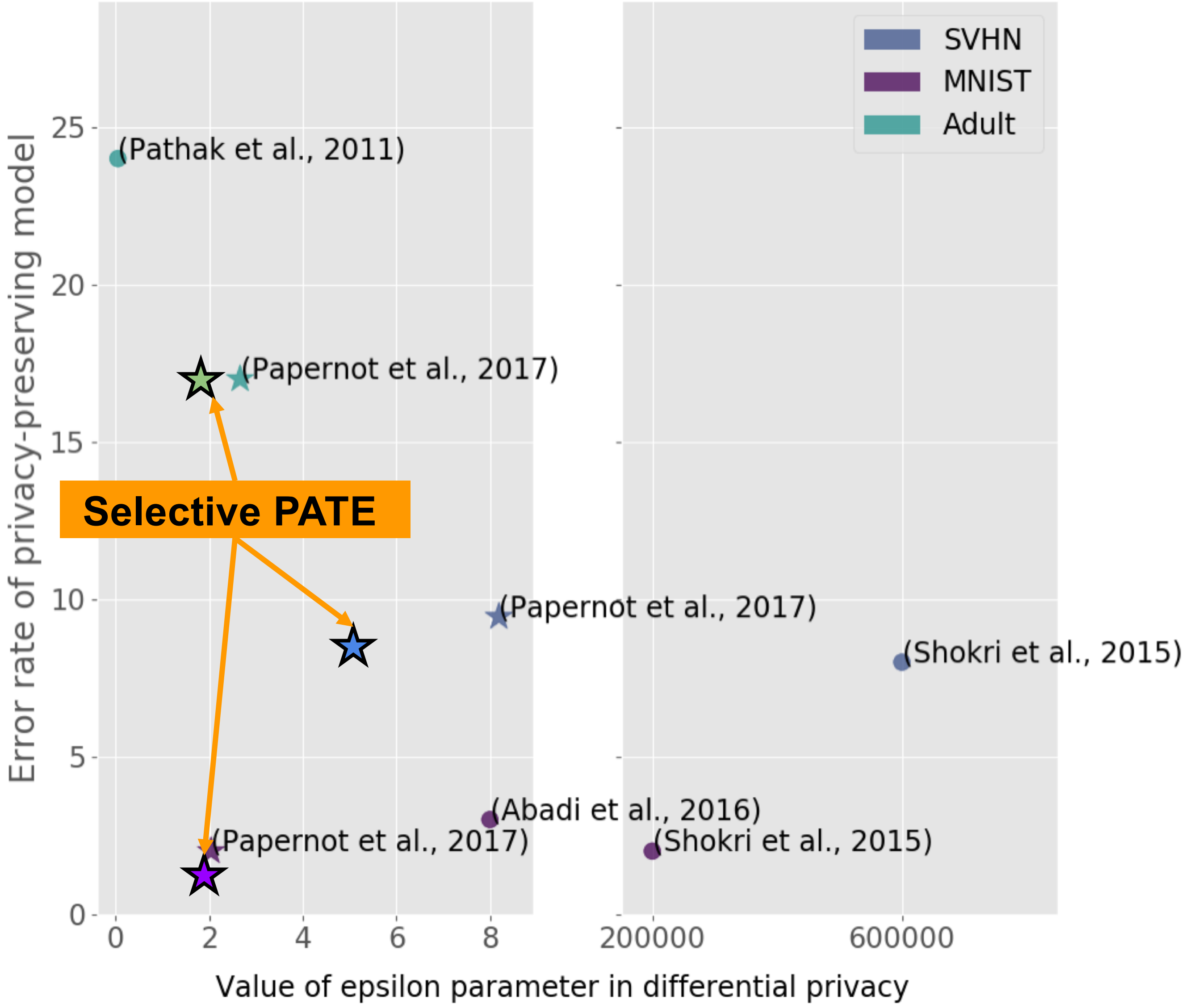
Synergy between utility and privacy

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

Trade-off between student accuracy and privacy



How to train a model with SGD?

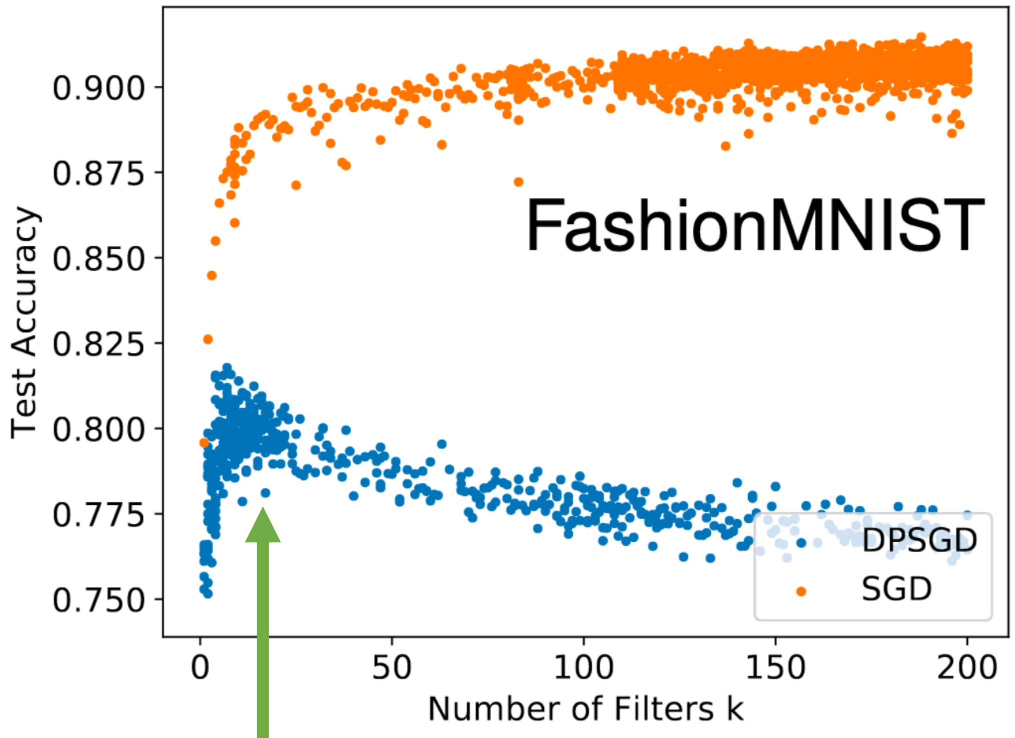
```
Initialize parameters θ

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?

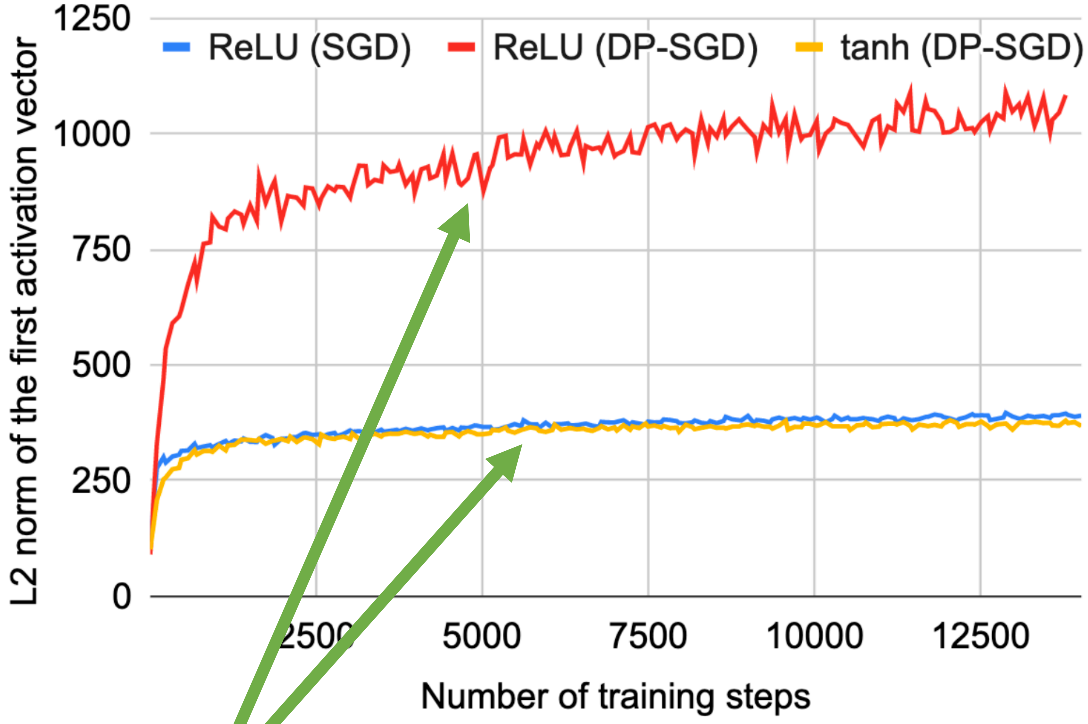
```
Initialize parameters θ

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

Architectures for DP-SGD learning



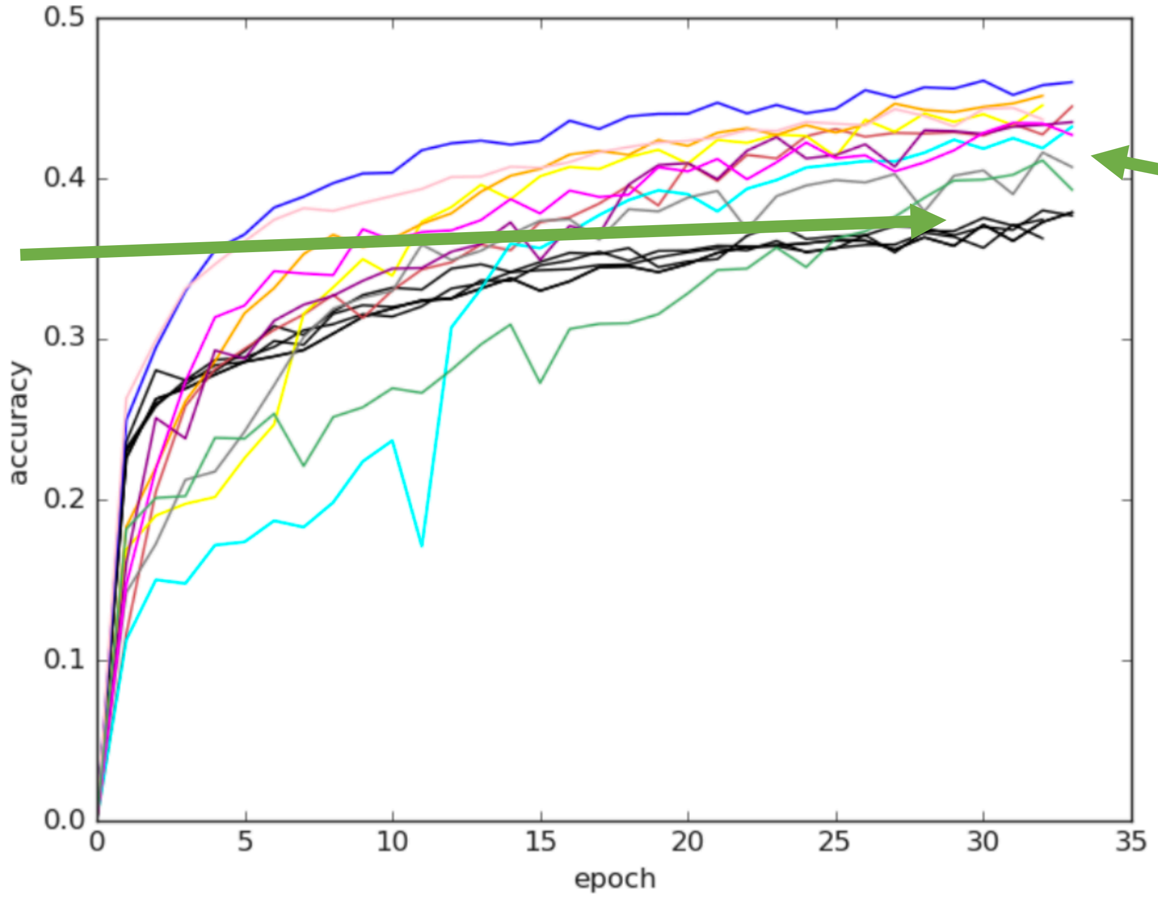
More capacity is not always helpful



Bounded activations help

Initializations for DP-SGD learning

Low variance between different initializations when learning without differential privacy



High variance between initializations when learning with DP-SGD and initializing with Raghu et al.

Hyperparameters for DP-SGD learning

			Non-private		Differentially-private	
Optimizer	Batch size	Epochs	Learning Rate	Test Acc.	Learning Rate	Test Acc.
SGD	256	40	$1.07 \cdot 10^{-1}$	90.3%	$3.32 \cdot 10^{-1}$	86.1%
	1024	7	$3.68 \cdot 10^{-1}$	86.3%	4.46	85.1%
Adam	256	40	$1.06 \cdot 10^{-3}$	90.5%	$1.32 \cdot 10^{-3}$	86.0%
	1024	7	$4.32 \cdot 10^{-3}$	88.7%	$7.08 \cdot 10^{-3}$	85.1%

Training with large batches for few epochs can be competitive in terms of wall-clock time

Best hyperparameter for non-private learning is not best hyperparameter for private learning

Adaptive optimizers are not necessarily helpful

Architectures, initializations, hyperparameters for DP-SGD learning

Dataset	Technique	Acc.	ε	δ	Assumptions
MNIST	SGD w/ tanh (not private)	99.0%	∞	0	-
MNIST	DP-SGD w/ ReLU	96.6%	2.93	10^{-5}	-
MNIST	**DP-SGD w/ tanh (ours)**	**98.1%**	**2.93**	10^{-5}	-
Fashion	SGD w/ ReLU (not private)	89.4%	∞	0	-
Fashion	DP-SGD w/ ReLU	81.9%	2.7	10^{-5}	-
Fashion	**DP-SGD w/ tanh (ours)**	**86.1%**	**2.7**	10^{-5}	-
CIFAR10	Transfer + SGD (not private)	75%	∞	0	-
CIFAR10	Transfer + DP-SGD (Abadi et al.)	67%	2	10^{-5}	Public Data
CIFAR10	**Transfer + DP-SGD (ours)**	**72%**	**2.1**	10^{-5}	Public Data

Making the Shoe Fit: Architectures, Initializations, and Tuning for Learning with Privacy

Papernot, Chien, Thakurta, Song, Erlingsson (in submission)

Useful resources

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