

Scalable PATE

The Secret Sharer

work by the Brain Privacy and Security team and collaborators at UC Berkeley
presented by Ian Goodfellow

PATE / PATE-G

- Private / Papernot
- Aggregation / Abadi
- Teacher / Talwar
- Ensembles / Erlingsson
- Generative / Goodfellow

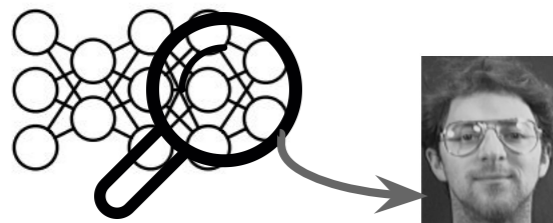
Threat Model

Types of adversaries and our threat model



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

Shokri et al. (2016) *Membership Inference Attacks against ML Models*
Fredrikson et al. (2015) *Model Inversion Attacks*



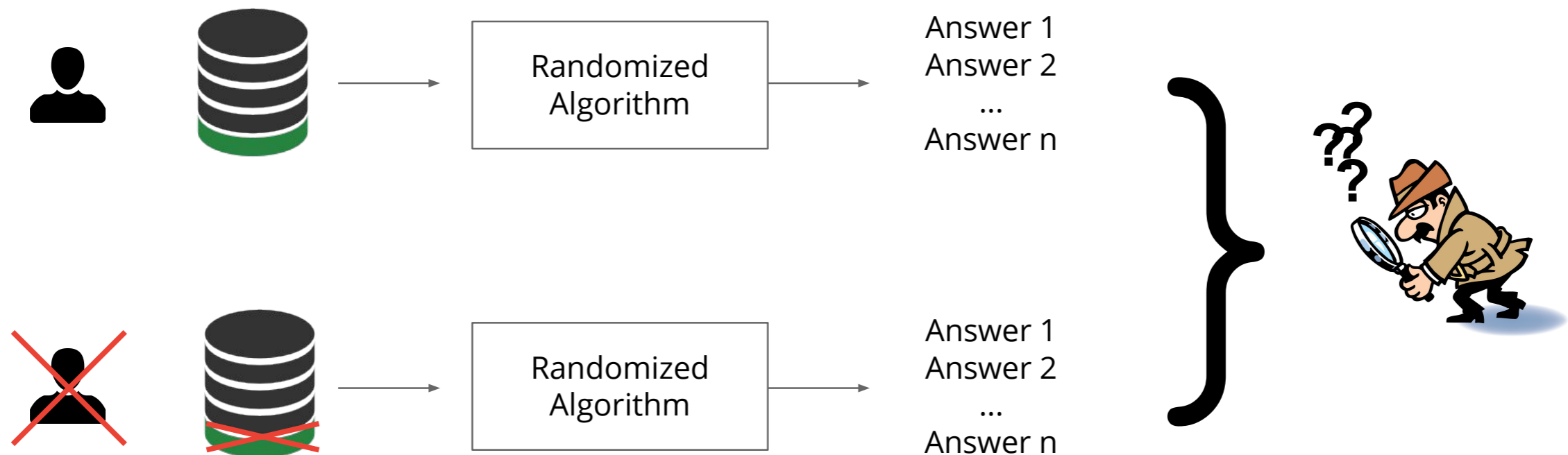
Model inspection (**white-box adversary**)

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

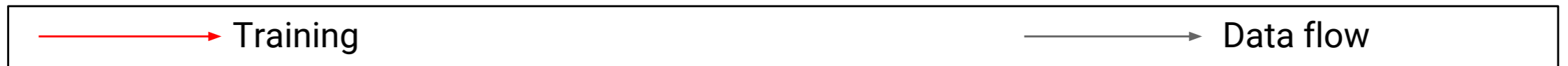
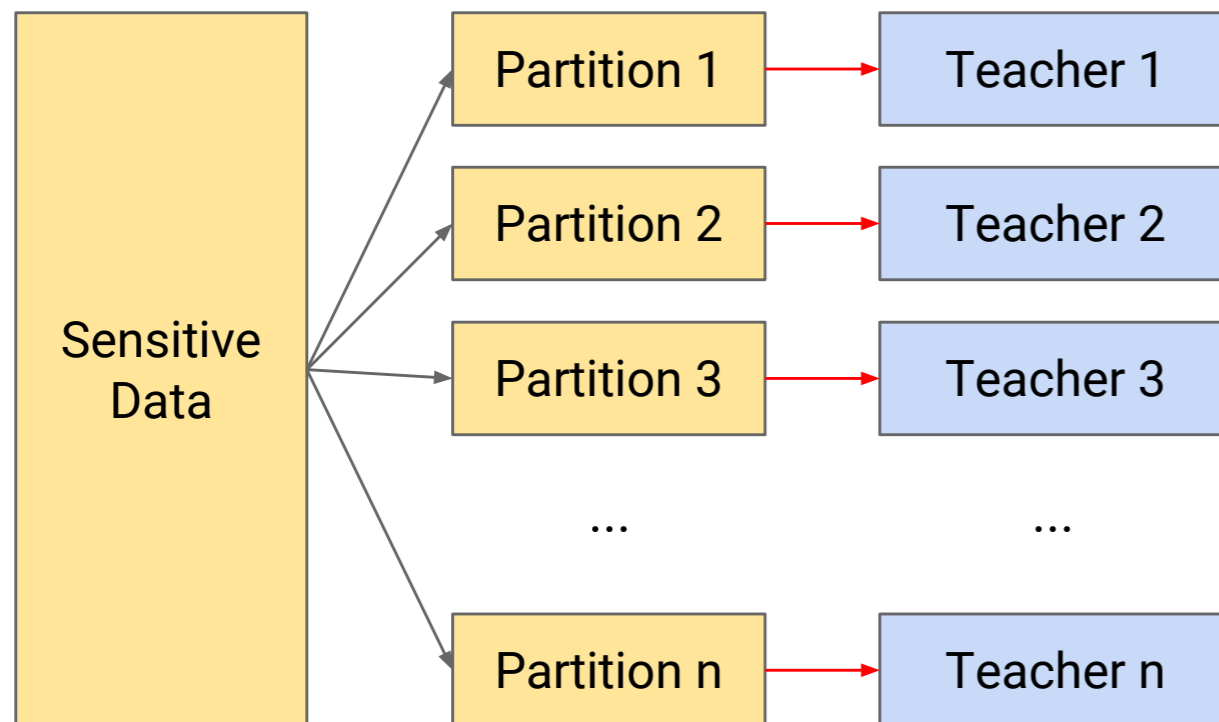
A definition of privacy: *differential privacy*



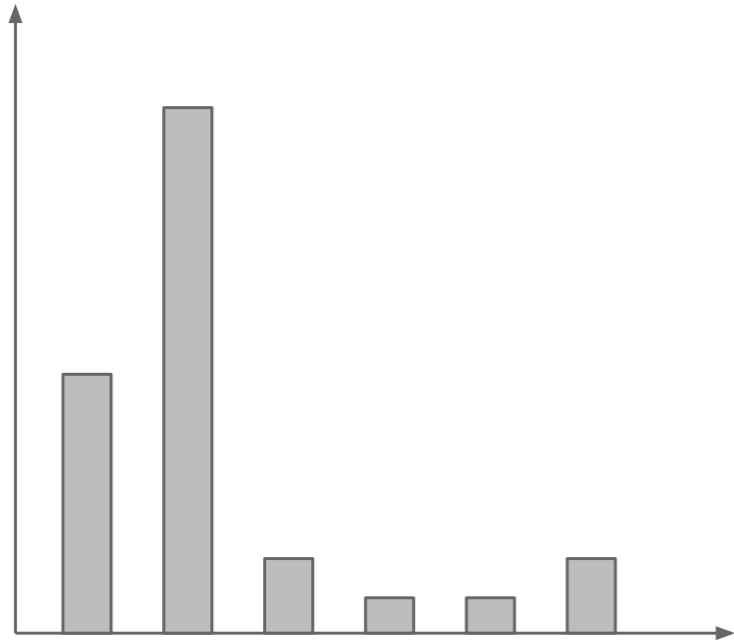
A tangent

- Which other fields need their “differential privacy moment”?
- Adversarial robustness needs a provable mechanism
- Interpretability needs measurable / actionable definitions
- Differential privacy is maybe the brightest spot in ML theory, especially in adversarial settings. Real guarantees that hold in practice

Different teachers learn from different subsets

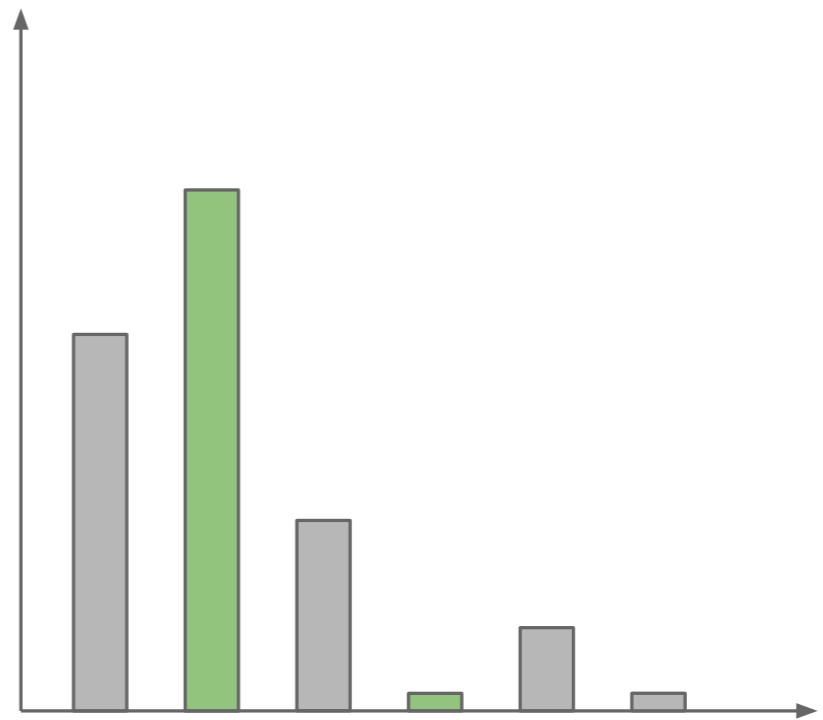


Aggregation



Count votes

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

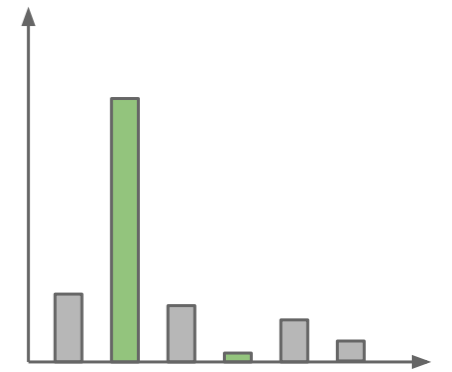


Take maximum

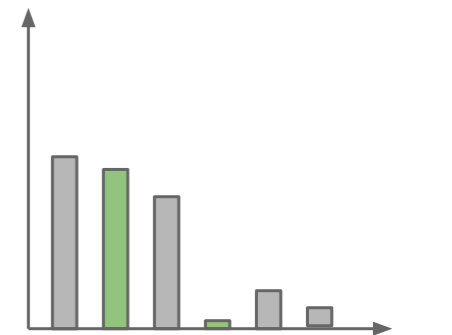
$$f(x) = \arg \max_j \left\{ n_j(\vec{x}) \right\}$$

Intuitive Privacy Analysis

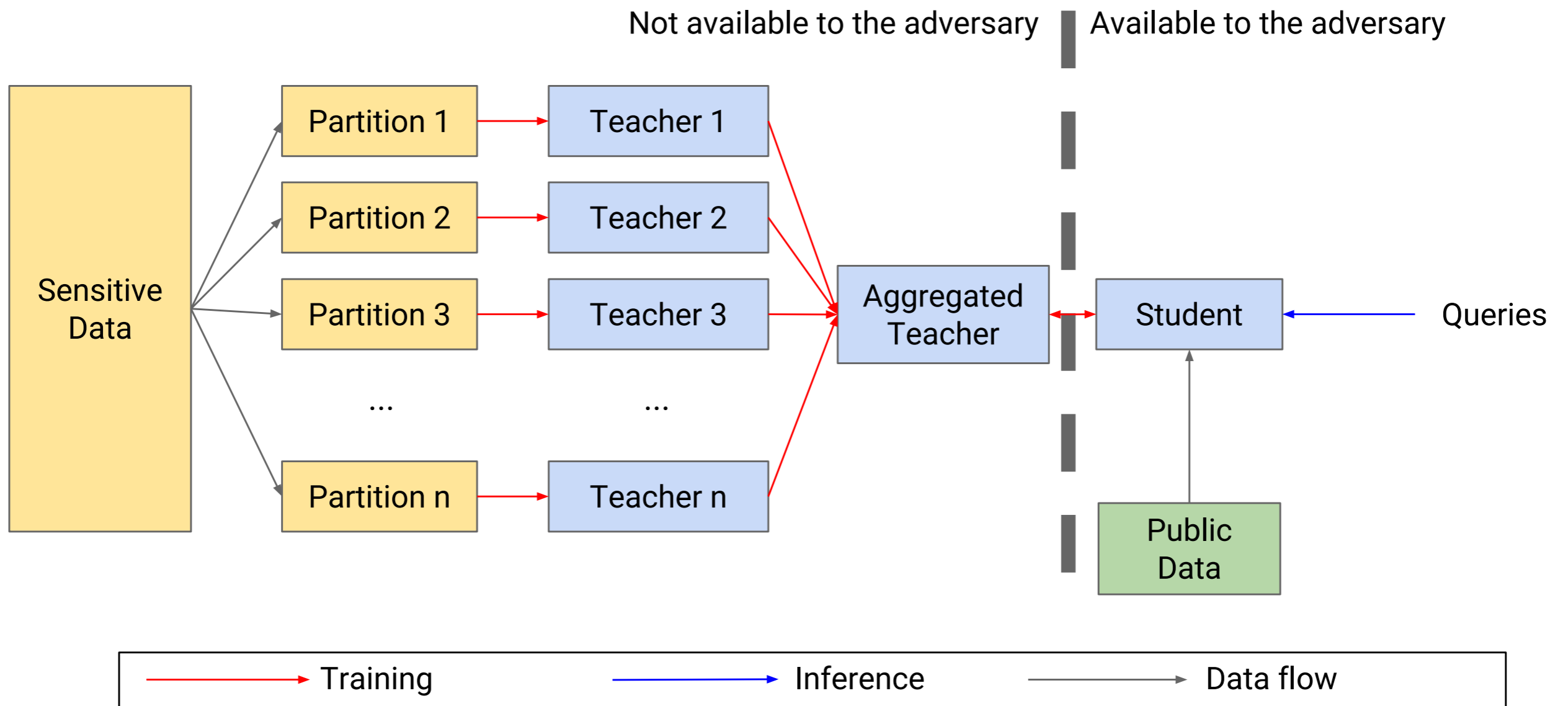
If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.



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



Student Training



Why train a student model?

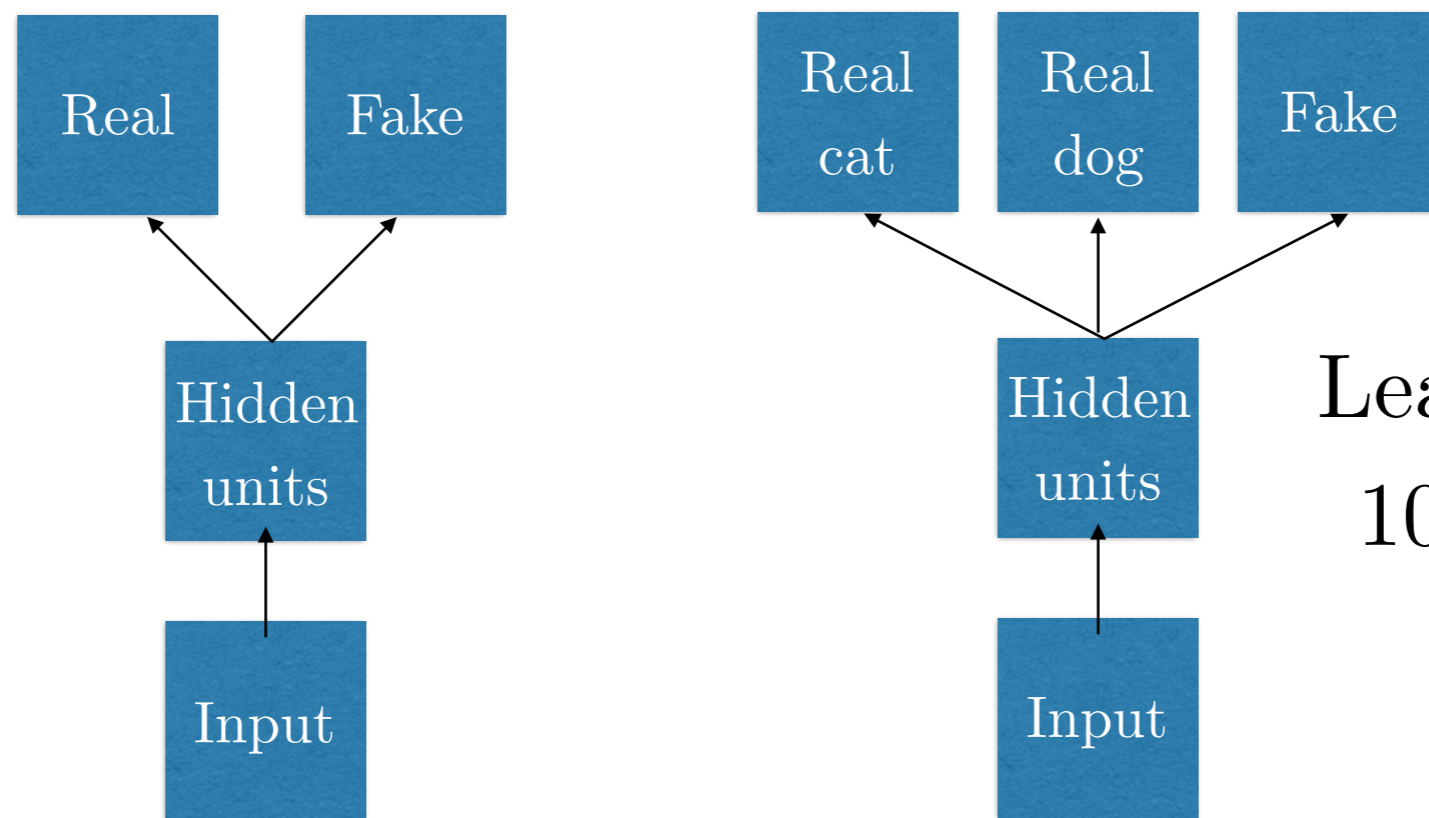
The aggregated teacher violates our threat model:

- 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.

Label-efficient learning

- More queries to teacher while training student = more privacy lost
- Use semi-supervised GAN (Salimans et al 2016) to achieve high accuracy with few labels

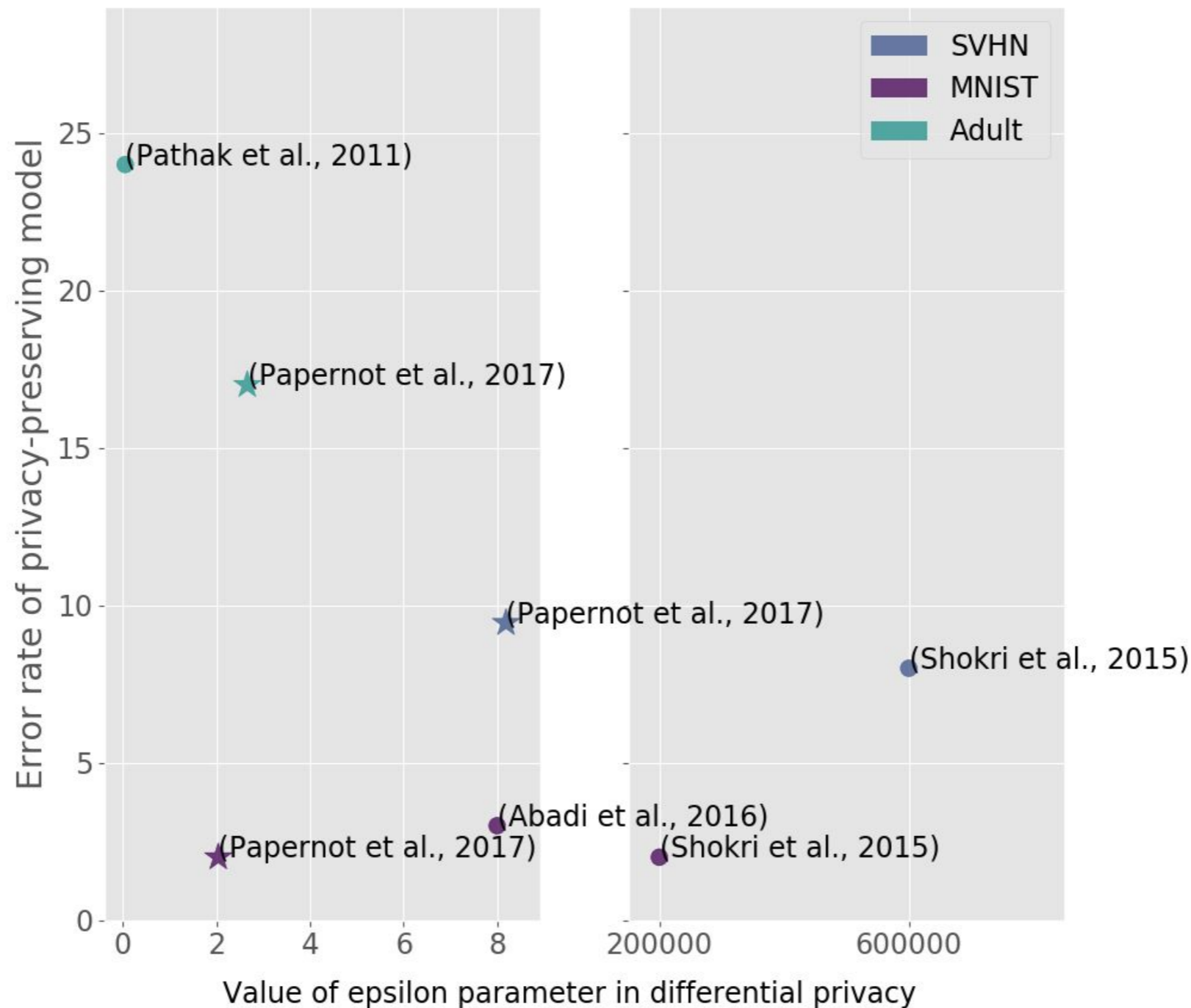
Supervised Discriminator for Semi-Supervised Learning



Learn to read with
100 labels rather
than 60,000

(Odena 2016, Salimans et al 2016)

Trade-off between accuracy and privacy



Scalable PATE

- Nicolas Papernot*, Shuang Song*, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Úlfar Erlingsson

Limitations of first PATE paper

- Only on MNIST / SVHN
 - Very clean
 - 10 classes (easier to get consensus)
- Scalable PATE
 - More classes
 - Unbalanced classes
 - Mislabeled training examples

Improvements

- Noisy votes use Gaussian rather than Laplace distribution
 - More likely to achieve consensus for large number of classes
- Selective teacher response

Selective Teacher Response

- Check for overwhelming consensus
 - Use high variance noise
 - Check if noisy votes for argmax exceed threshold T
- Consensus? Publish noisy votes with smaller variance
- No consensus? Don't publish anything, student skips
 - Note: running the noisy consensus check still spent some of our privacy budget

Background: adversarial training

Labeled as bird

Still has same label (bird)



Decrease
probability
of bird class



Virtual Adversarial Training

Unlabeled; model
guesses it's probably
a bird, maybe a plane



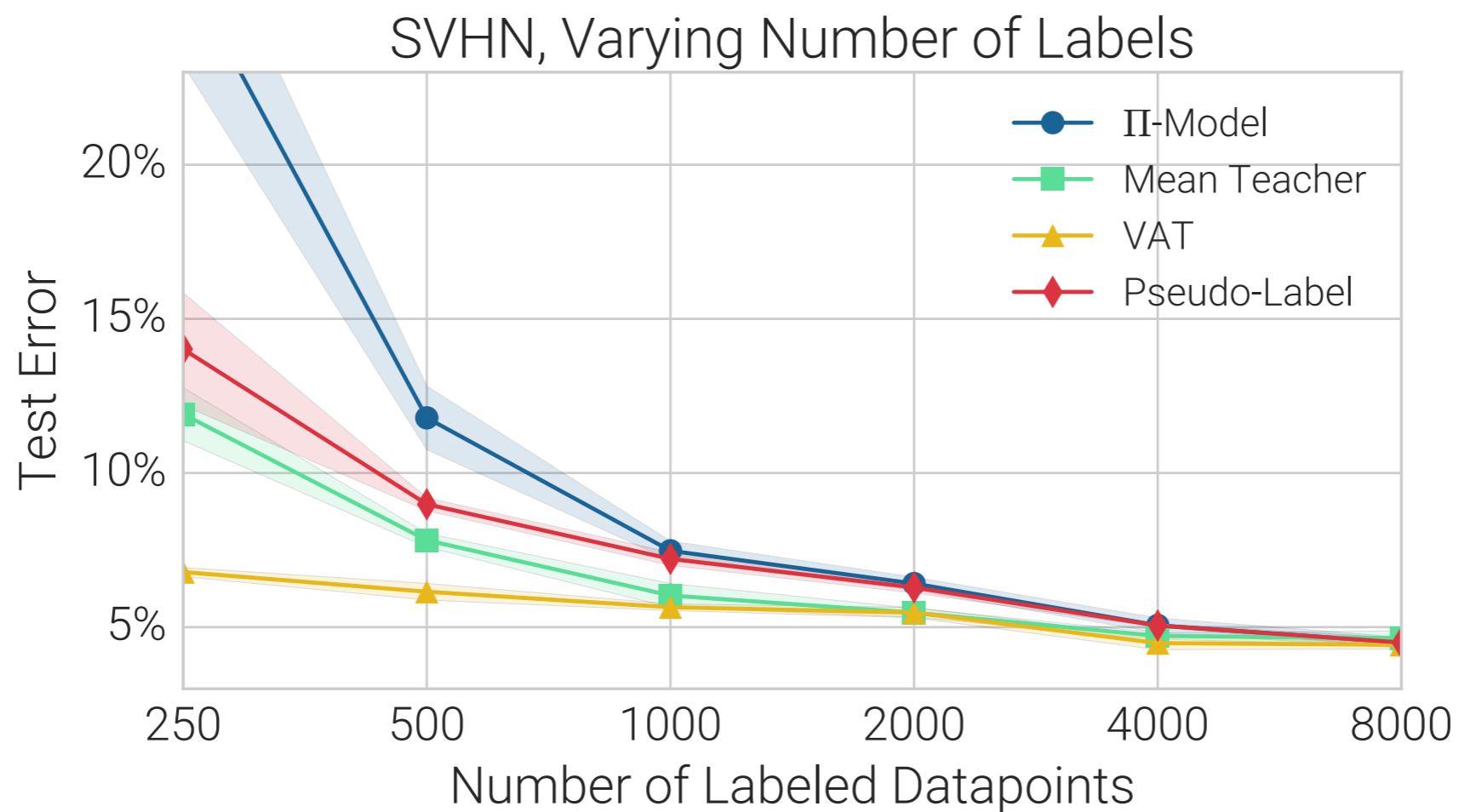
New guess should
match old guess
(probably bird, maybe plane)



→
Adversarial
perturbation
intended to
change the guess

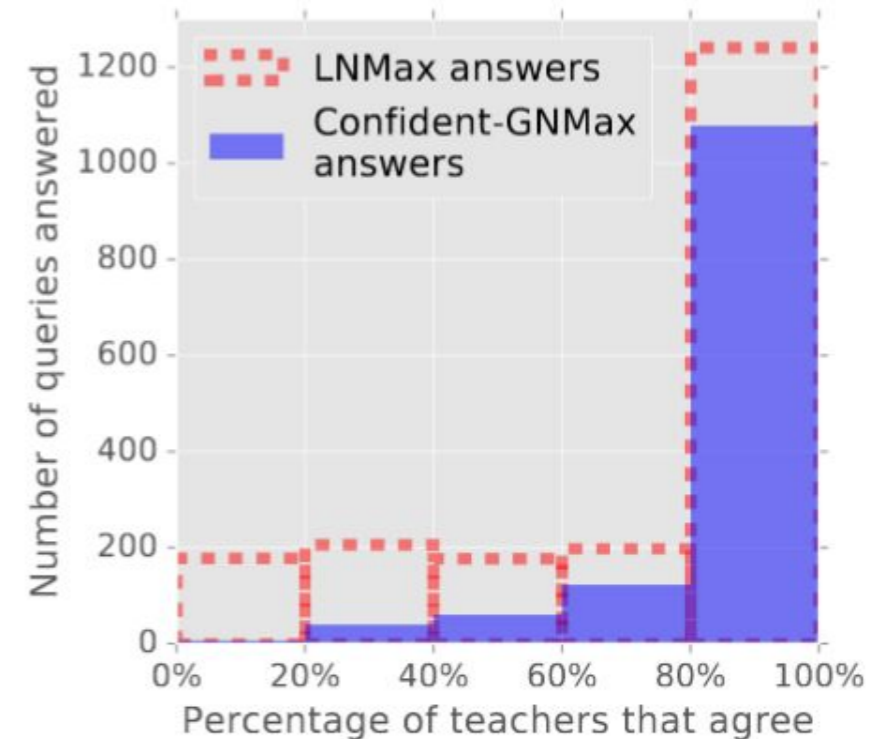
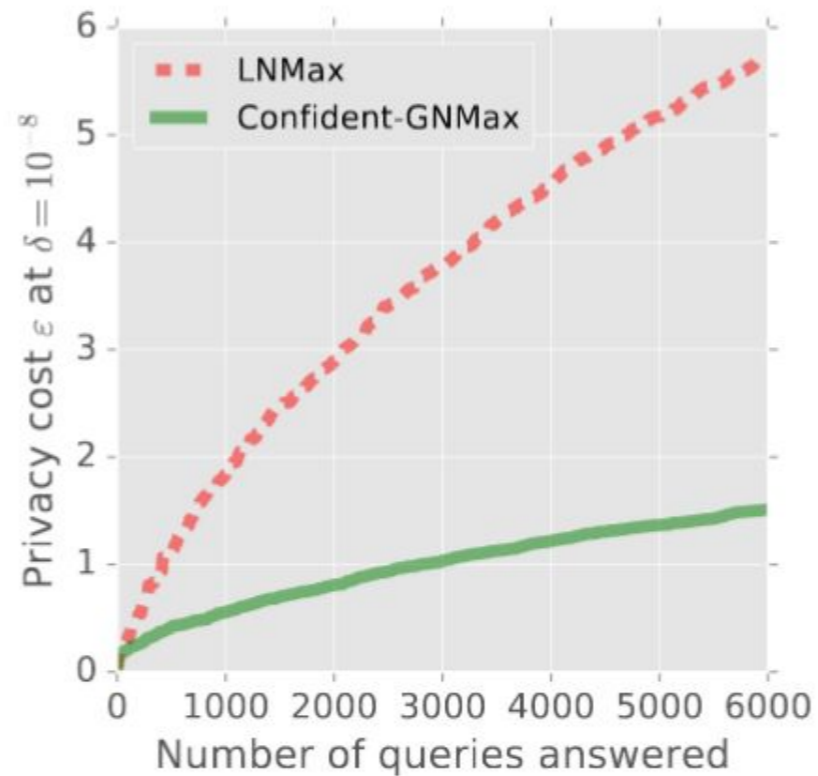
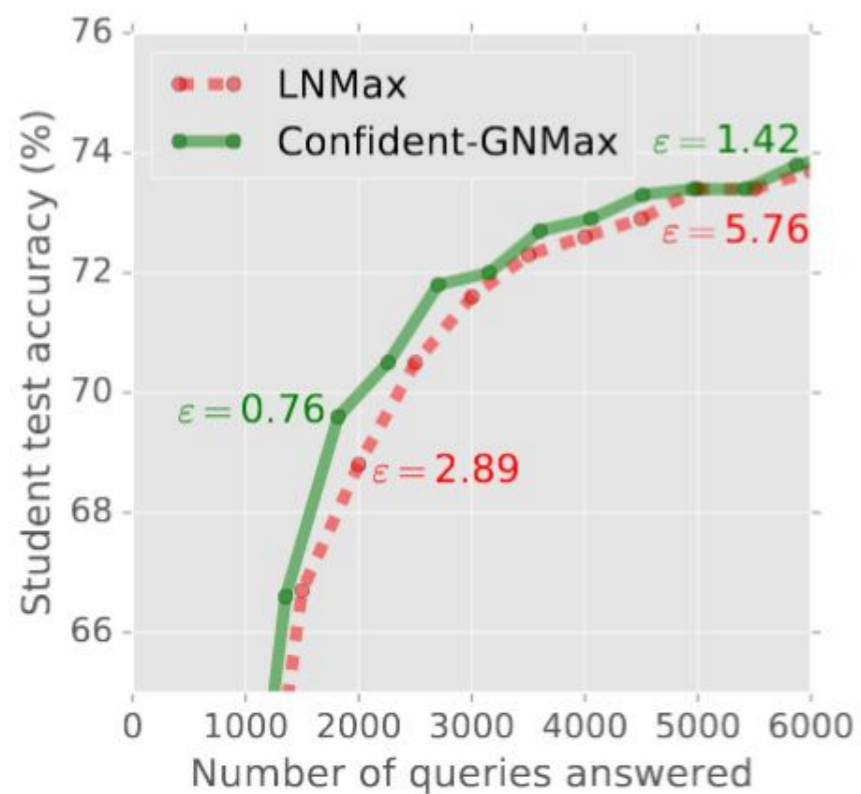
(Miyato et al, 2015)

VAT performance



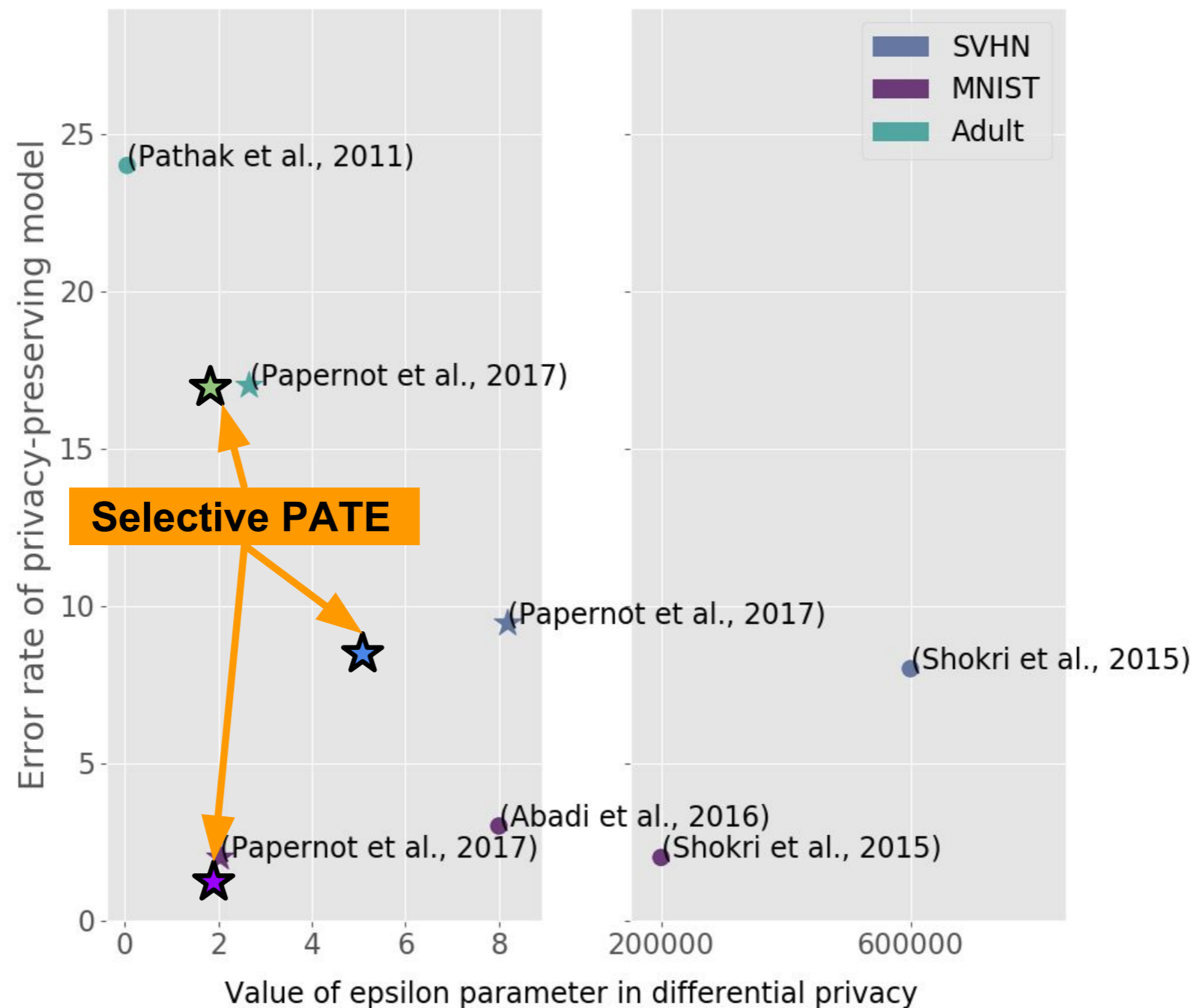
(Oliver+Odena+Raffel et al, 2018)

Scalable PATE: Improved Results



(LNMax=PATE, Confident-GNMax=Scalable PATE)


Scalable PATE: Improved tradeoff



The Secret Sharer

- Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson,
Dawn Song

Secret with format known to adversary

- “My social security number is _____ - _____ - _____”

Secret

- Measure memorization with *exposure*

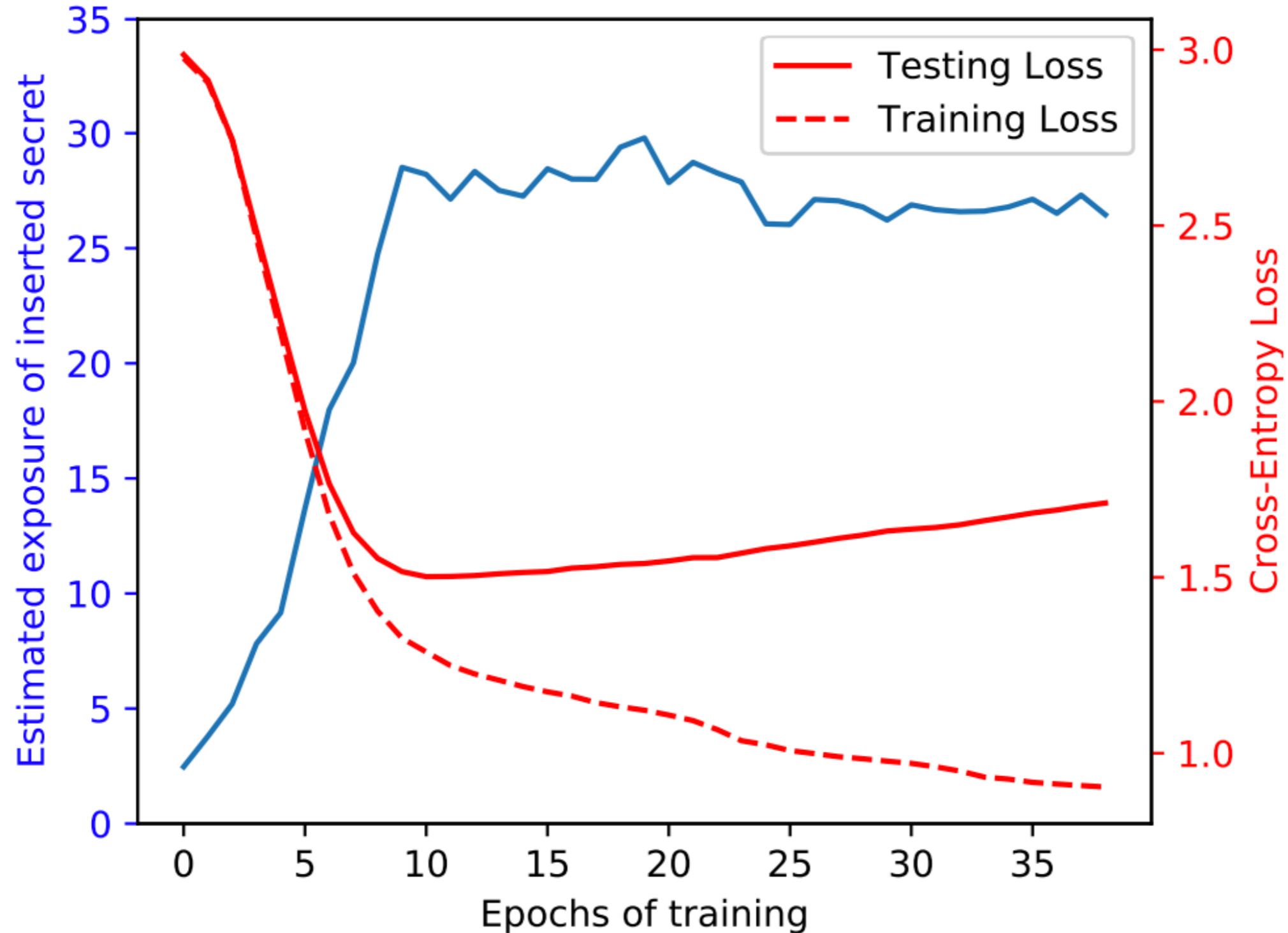
Definitions

- Suppose model assigns probability p to the actual secret
- The *rank* of the secret is the number of other strings given probability $\leq p$
 - Minimum value is 1
- *Exposure*: negative log prob of sampling a string with probability less than p
- equivalent: *Exposure*: $\log(\# \text{ possible strings}) - \log \text{rank}$

Practical Experiments

- Can estimate exposure via sampling
- Can approximately find most likely secret value with optimization (beam search)

Memorization during learning



Observations

- Exposure is high
- Exposure rises early during learning
- Exposure is not caused by overfitting
 - Peaks before overfitting occurs

Comparisons

- Across architectures:
 - More accuracy \rightarrow more exposure
 - LSTM / GRU: high accuracy, high exposure
 - CNN: lower accuracy, lower exposure
- Larger batch size \rightarrow more memorization
- Larger model \rightarrow more memorization
 - Secret memorization happens even when compressed model smaller than compressed dataset
- Choice of optimizer: no significant difference

Defenses

- Regularization does not work
 - Weight decay
 - Dropout
 - Weight quantization
- Differentially privacy works, as guaranteed
 - Even for very small epsilon, with little theoretical guarantee, the exposure measured in practice decreases significantly

Questions