Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, Kunal Talwar

ICLR 2017

Contents

Machine Learning & Privacy

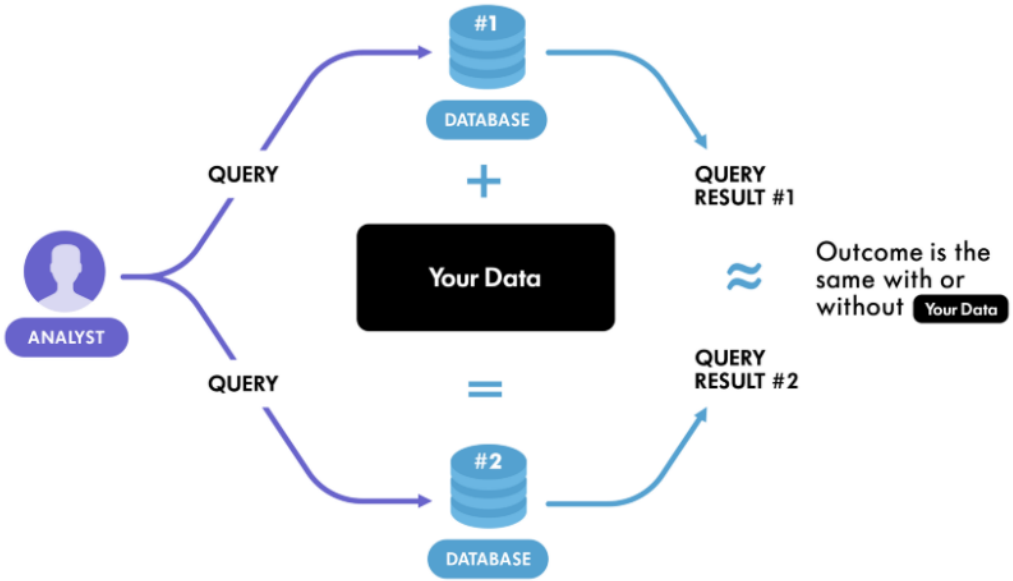
ML models can retain information (overlearning) and some memorise training data.

- e.g Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures(Fredrikson, Matt et al., 2015)

Threats:

1. Model querying (blackbox)
2. Model inspection (white box)

Differential Privacy

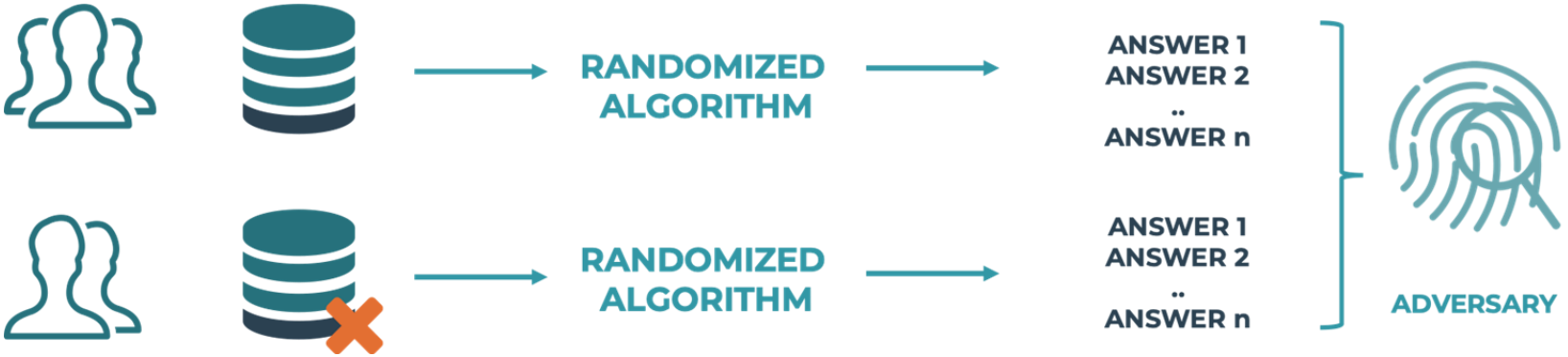


https://www.winton.com/research/using-differential-privacy-to-protect-personal-data

Differential Privacy

➔ same inference from query, whether an individual was included in the input data or not.

Differential Privacy in Machine Learning



https://2021.ai/machine-learning-differential-privacy-overview/

Differential Privacy

Definition 1: A randomized mechanism M with domain D and range R satisfies(ε,δ)-differential privacy if for any two adjacent inputs d, d’∈ D and for any subset of outputs S ⊆ R it holds that:

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

Smaller ε, stronger privacy guarantee and more noise.

δ is the tolerance bias.

Privacy Loss

Definition 2: . Let M: D → R be a randomized mechanism and d, d' a pair of adjacent databases. Let aux denote an auxiliary input. For an outcome o ∈ R, the privacy loss at o is defined as:

$$c(o; \mathcal{M}, \mathit{aux}, d, d') \triangleq \log \frac{\Pr[\mathcal{M}(\mathit{aux}, d) = o]}{\Pr[\mathcal{M}(\mathit{aux}, d') = o]}.$$

Moments Accountant

Procedure that computes the privacy cost at each access to the training data, and accumulates this cost as the training progress

- At each step, we use the aggregation mechanism with noise $Lap(\frac{1}{\gamma})$ which is $(2\gamma, 0)$-DP.
- Thus over T steps, we get T steps, we get $(4T\gamma^2 + 2\gamma\sqrt{2T\ln\frac{1}{\delta}}, \delta)$-differential privacy.

PATE

Paper in a Nutshell

Private Aggregation of Teacher Ensembles - PATE

- ➔ Disjoint training sets and classifiers
- ➔ Noisy aggregation
- ➔ Semi supervised learning with PATE-G: teacher-student training of modified GAN model

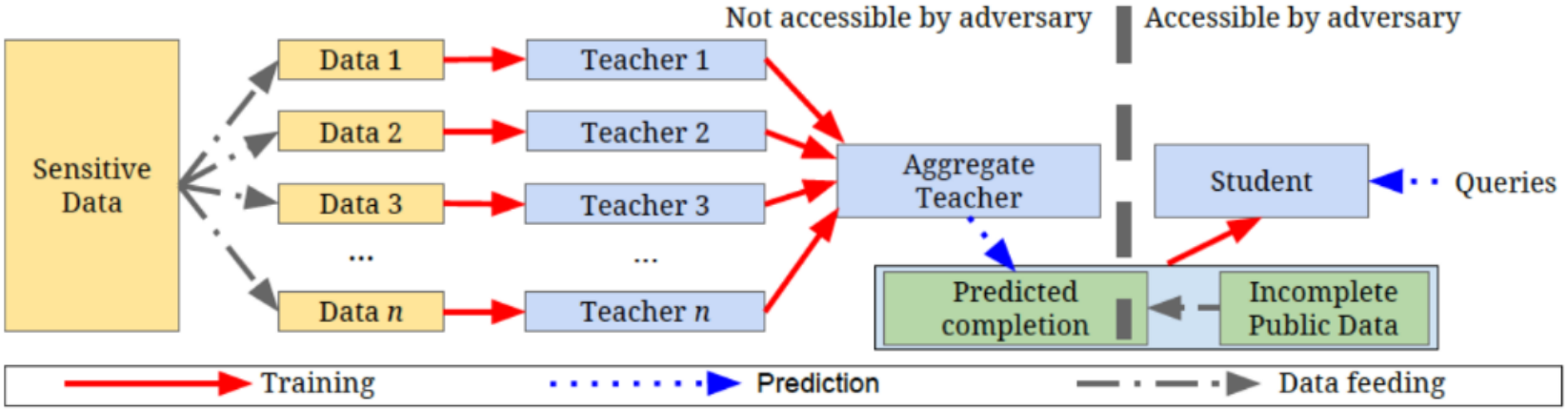
Threat model

- Trust ML service
- Do no trust user

Adversaries:

- Has access to model parameters
- Unlimited number of queries

PATE

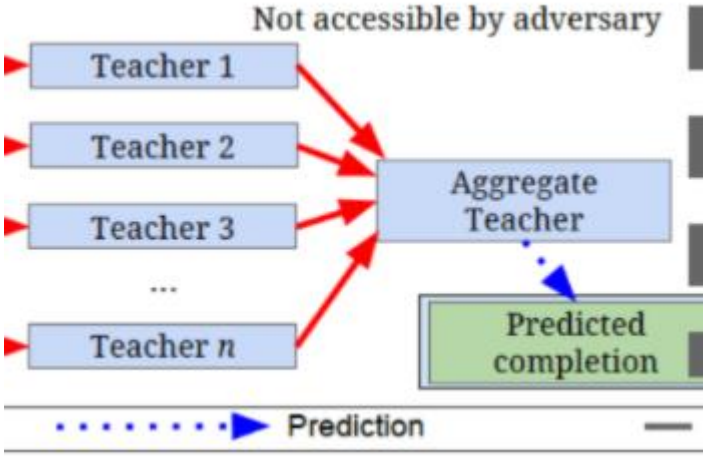

Not accessible by adversary
Accessible by adversary
Sensitive Data
Data 1
Data 2
Data 3
...
Data n
Teacher 1
Teacher 2
Teacher 3
...
Teacher n
Aggregate Teacher
Student
Queries
Predicted completion
Incomplete Public Data
Training
Prediction
Data feeding

Teacher Ensemble

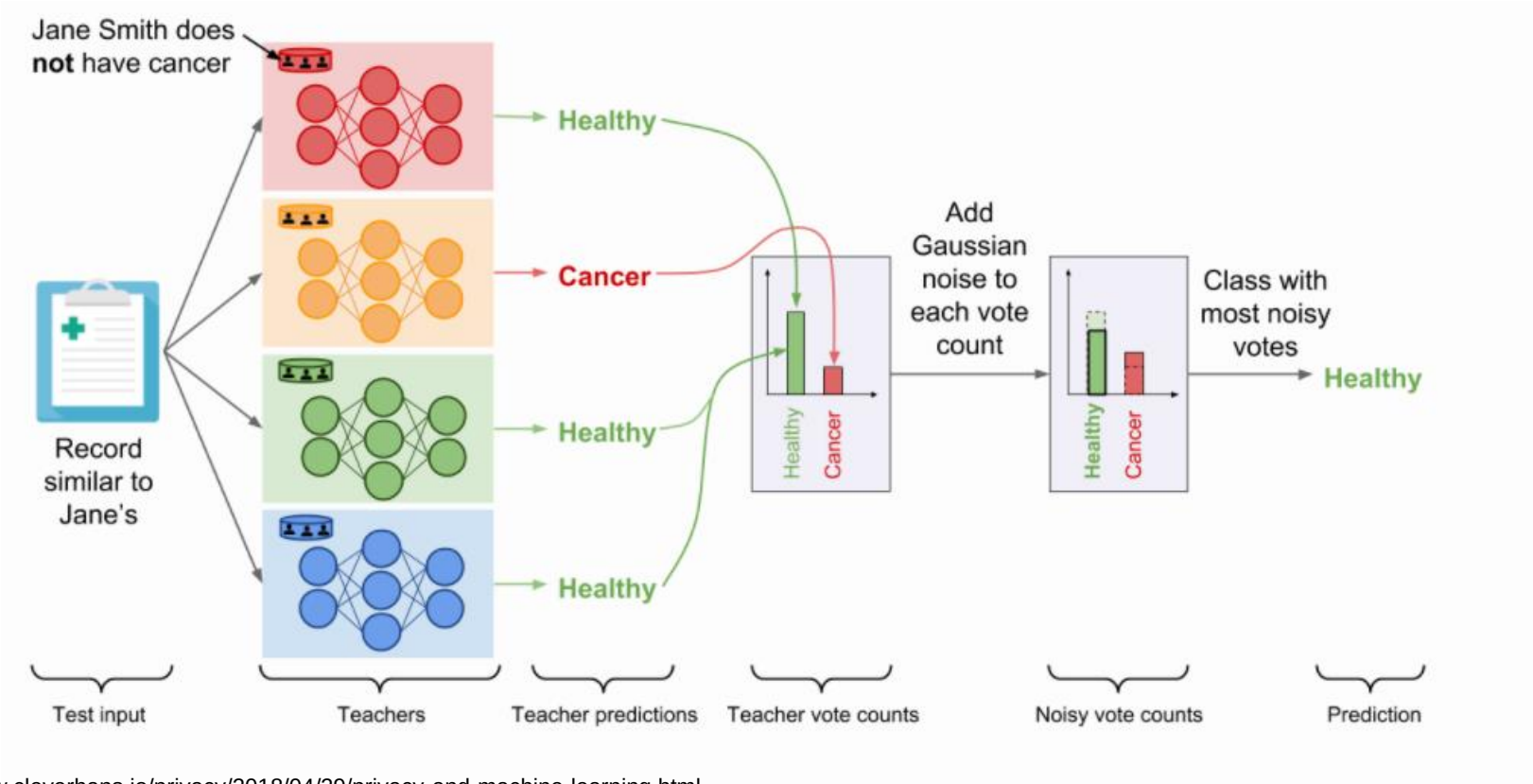
1. Divide dataset into n subsets
2. For each subsets, train a separate instance of a model (teacher)
3. At inference time, each teacher outputs its prediction
4. Count votes for each label, the one with the most counts is the predicted class
5. Users have access to predictions only (blackbox teachers models)

Aggregation Algorithm

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



- Count prediction for class j from each teacher given input $\vec{x}$
- Add Laplacian noise to each vote count n_j, γ being privacy parameter
- Return the prediction with the most votes

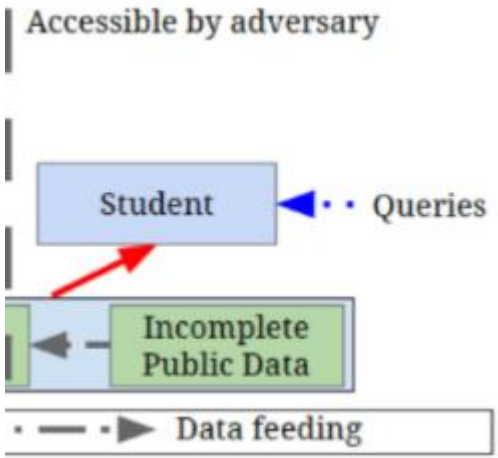


http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html

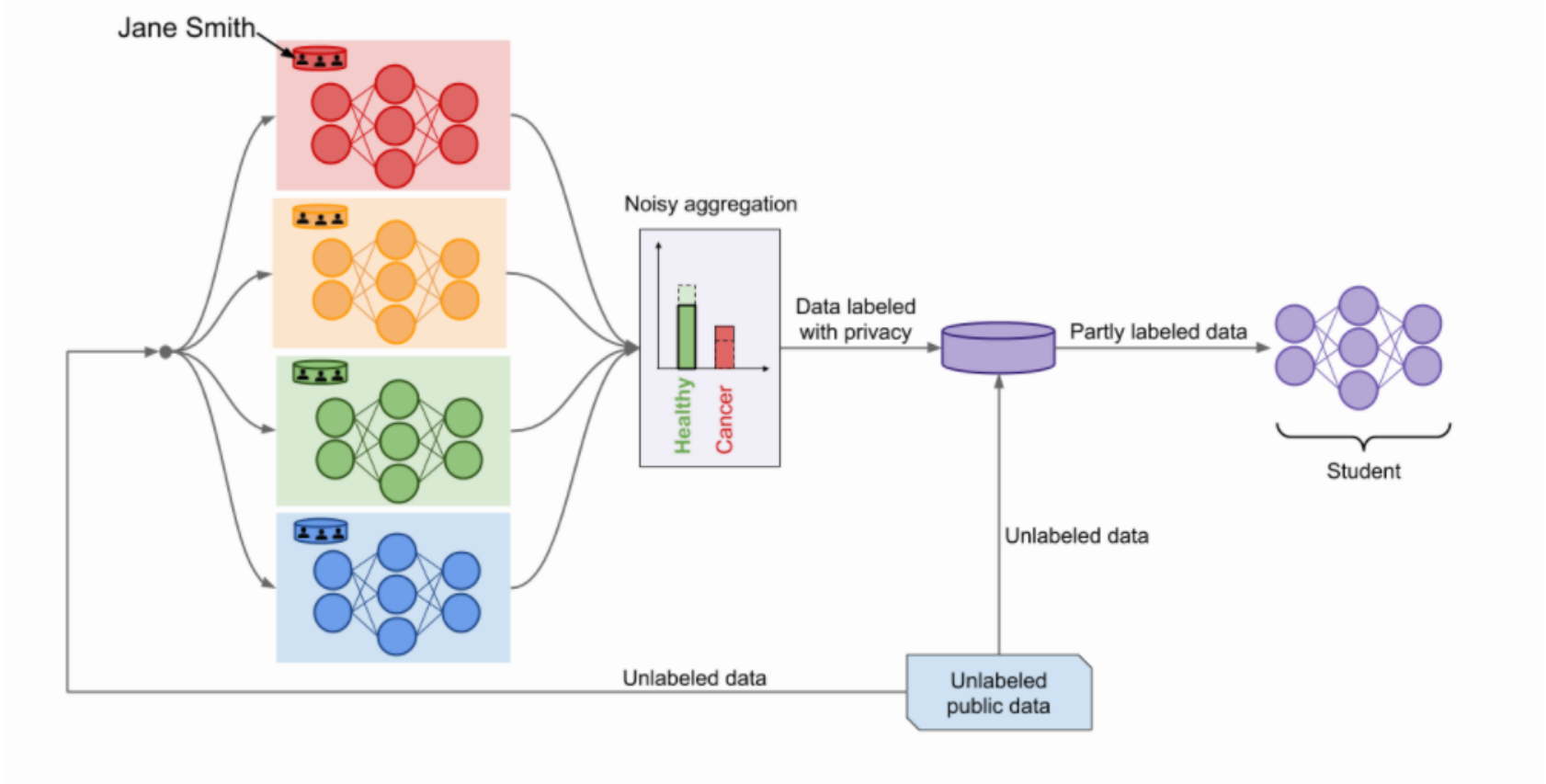
Privacy guarantee?

- Each query from the adversary increase the privacy costs
- Teachers' parameters could be discovered and consequently reveal the data they were trained

Student Model



- Student model trained on data partly labeled by teacher ensemble
- Available to users
- Data must be non sensitive
- Privacy loss is fixed to the number of queries allowed to the student model
- Privacy loss does not grow with number of queries to the student model

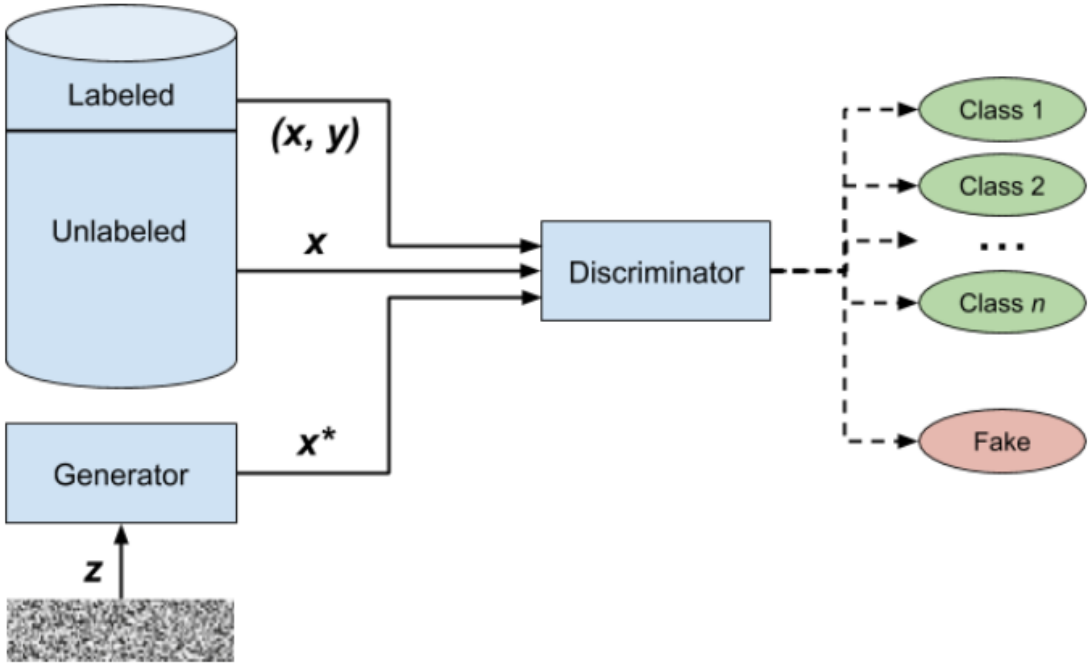

Jane Smith
Noisy aggregation
Healthy
Cancer
Data labeled
with privacy
Partly labeled data
Student
Unlabeled data
Unlabeled data
Unlabeled
public data

Experiments

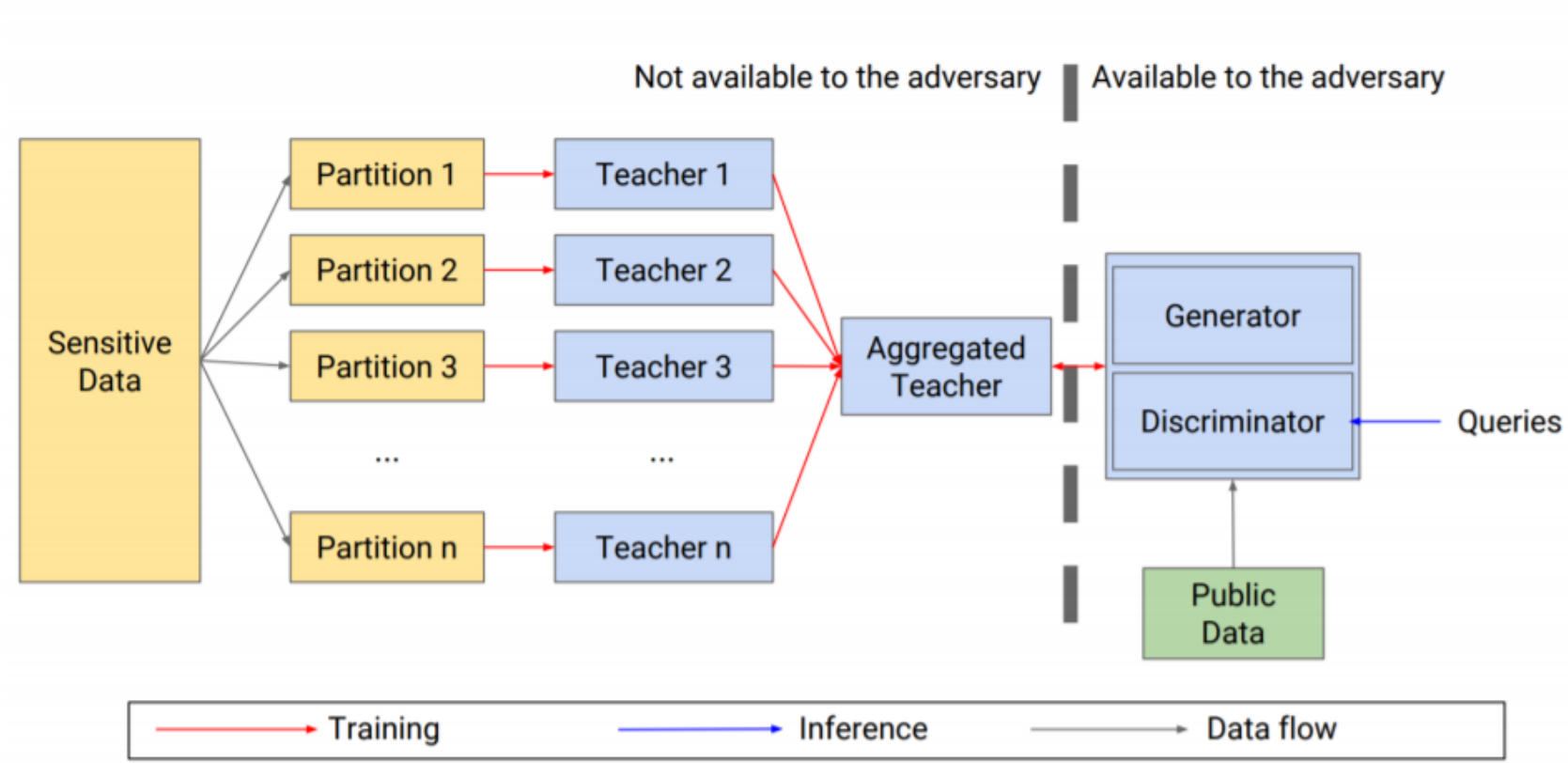
- ➔ Student model: PATE-G
- ➔ Number of teachers
- ➔ Privacy cost
- ➔ Utility (Accuracy)
- ➔ Experiments on MNIST and SVHN

PATE-G

Student Model: Semi-supervised GAN



https://livebook.manning.com/book/gans-in-action/chapter-7/v-6/1

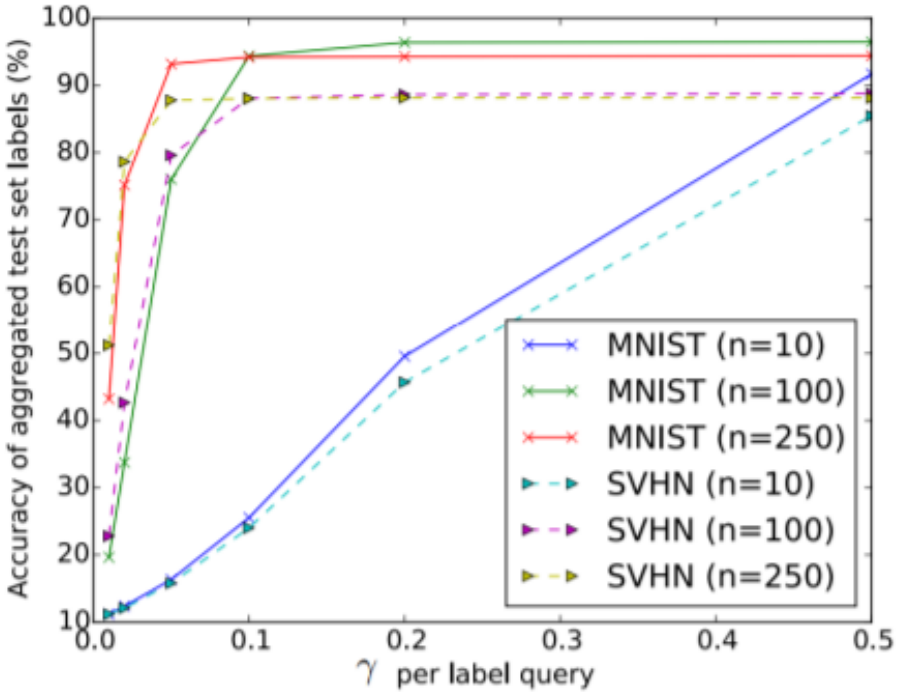

Not available to the adversary
Available to the adversary
Sensitive Data
Partition 1
Partition 2
Partition 3
...
Partition n
Teacher 1
Teacher 2
Teacher 3
...
Teacher n
Aggregated Teacher
Generator
Discriminator
Queries
Public Data
Training
Inference
Data flow

Experiments

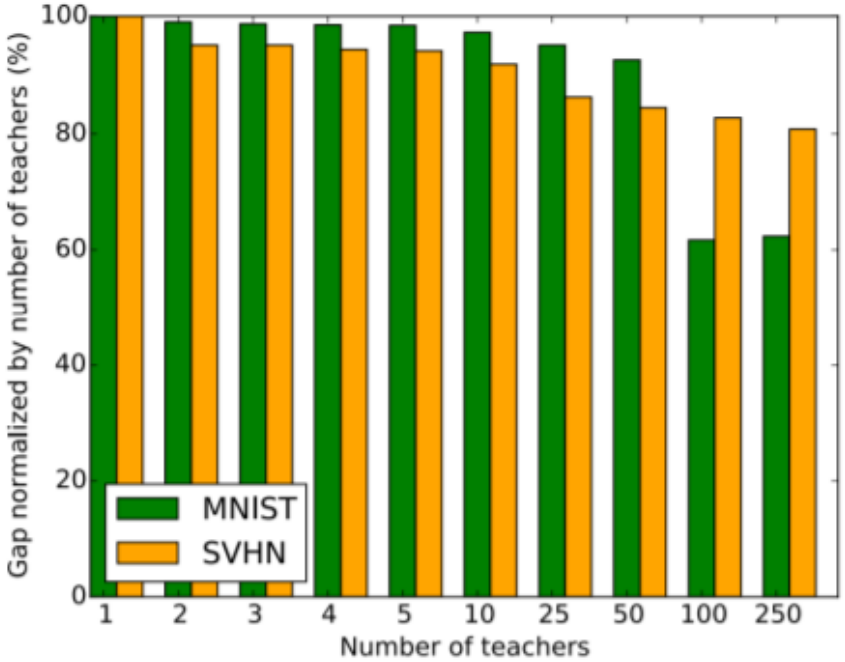
Experiment setup

Dataset	Teacher Model	Student Model	Student Public Data	Testing Data
MNIST	2 conv + 1 relu	GANs (6 fc layers)	test[:1000]	test[1000:]
SVHN	2 conv + 2 relu	GANs (7 conv + 2 NIN)	test[:1000]	test[1000:]

Accuracy per noise amplitude



Number of teachers and privacy



Results

Dataset	ε	δ	Queries	Non-Private Baseline	Student Accuracy
MNIST	2.04	10^{-5}	100	99.18%	98.00%
MNIST	8.03	10^{-5}	1000	99.18%	98.10%
SVHN	5.04	10^{-6}	500	92.80%	82.72%
SVHN	8.19	10^{-6}	1000	92.80%	90.66%

Discussion

1. Requires ~250 of teacher models and data subsets for SVHN and MNIST

1. Number of teachers are a parameter to optimise for each dataset

1. Impact on accuracy is relatively low

Limitations

1. Evaluated only on MNIST and SVHN tasks (relatively simple)

1. Require many teachers as number of output classes and complexity of models increases

1. Datasets must have private and public sets available

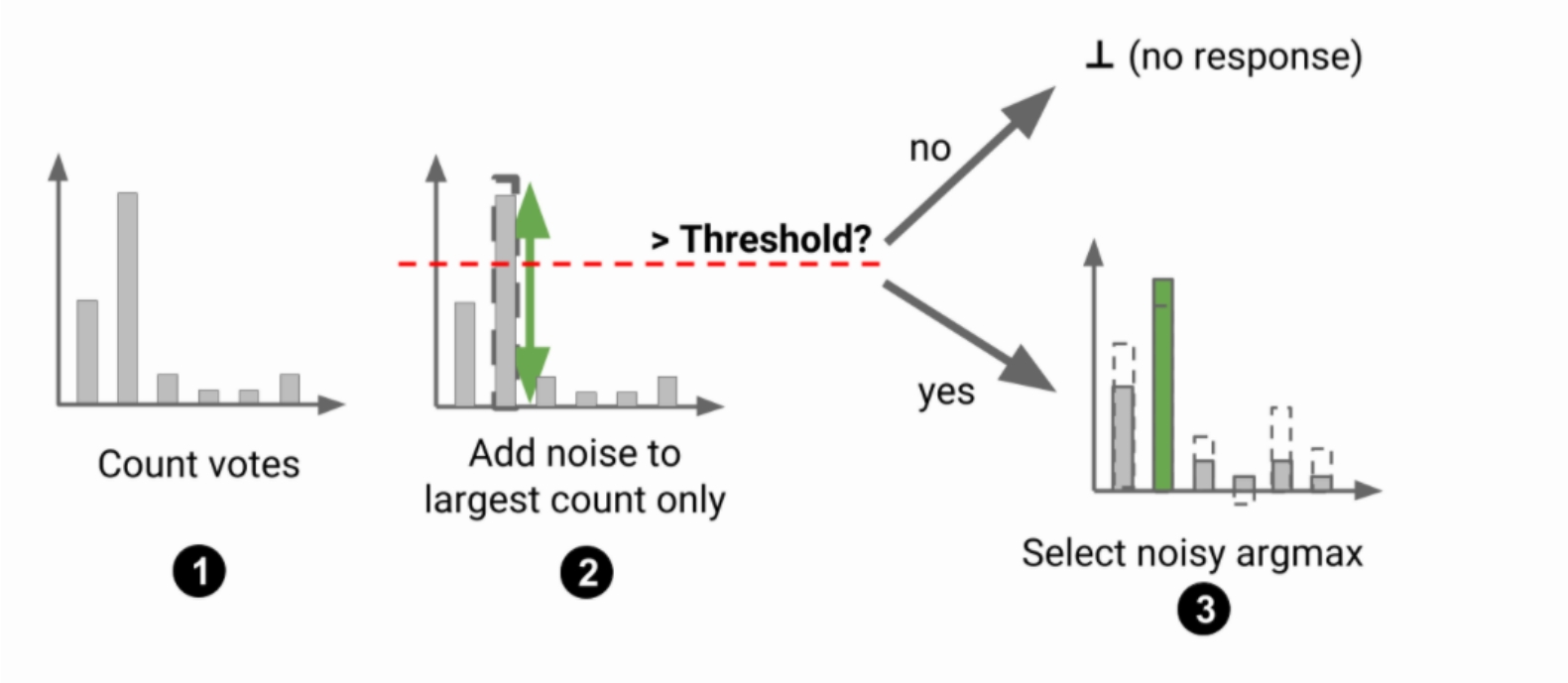
Scalable Private Learning with PATE (2018)

1. Refined aggregation algorithm: Confident Aggregator

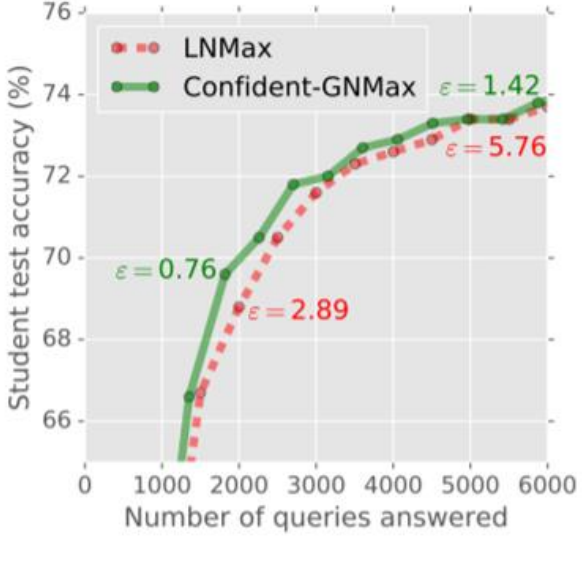
1. Smaller privacy loss

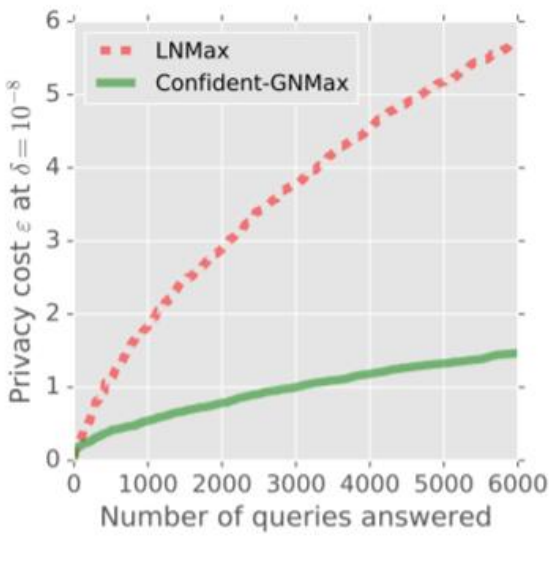
1. Better predictions performance

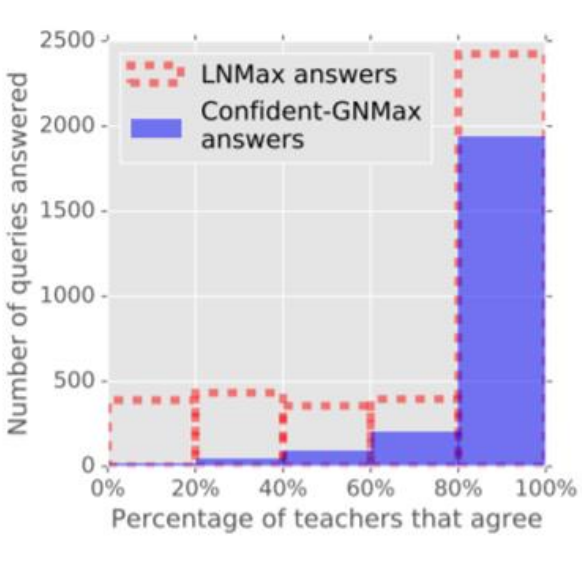
Confident Aggregator



Scalable Private Learning with PATE (2018)







Conclusion

1. PATE is a general framework for privacy preserving machine learning training using a black box training strategy on sensitive data and noisy aggregation that provides differential privacy guarantee

1. Compared to previous work, make no assumption about models, parameters or student model. Student model is free and can only access top k votes.

1. PATE acts as a regularizer and can improve generability

1. Requires large datasets with private and public data available

Related Work

1. Differentially private stochastic gradient descent
    a. Deep Learning with Differential Privacy (Abadi et al., 2016) [here]

1. Federated Learning
    a. Federated Machine Learning: Concept and Applications (Yang, Qiang, et al, 2019) [here]