Differential Privacy using PATE

"Semi-supervised knowledge transfer for deep learning from private training data", by Papernot, et al.
https://arxiv.org/pdf/1610.05755.pdf

Dan Parshall

08/27/2021

Question

"How can you build a predictive model which can be used by any of your clients, while ensuring that private data from one client isn't leaked to another?"

Question

"How can you build a predictive model which can be used by any of your clients, while ensuring that private data from one client isn't leaked to another?"

- Hospitals
- Schools
- Banks

Differential Privacy

- Why we need differential privacy
- What *is* differential privacy?
- PATE model for differential privacy

Why we need differential privacy

- We want our users to feel comfortable sharing their data with us
- Even with limited query power, a database can be reconstructed
  - 40% of USA Census records can be uniquely identified
- ML models embed data in their weights
  - For SVM, the support vectors *are* boundary data points

The secret ingredient is noise

Flip a coin and tell me "Whom did you vote for in 2020?"

- If the coin comes up heads, tell me honestly...

The secret ingredient is noise

Flip a coin and tell me "Whom did you vote for in 2020?"

- If the coin comes up heads, tell me honestly...
- But if it comes up tails, then flip it again, and follow this rule:
  - Heads, tell me "Biden"
  - Tails, tell me "Trump"

The secret ingredient is noise

Flip a coin and tell me "Whom did you vote for in 2020?"

- If the coin comes up heads, tell me honestly...
- But if it comes up tails, then flip it again, and follow this rule:
  - Heads, tell me "Biden"
  - Tails, tell me "Trump"

Because we know the distribution of the added noise, we can infer the true population distribution, while being ignorant of any individual response

Differential privacy: definition & properties

- An algorithm is differentially private if the *difference* between a query on a dataset with any particular record is "very small" compared to one without; we will add however much noise is needed to achieve that

- Quantifiable
  - Can calculate noise needed as function of "privacy budget", ϵ (smaller ϵ means more noise)
- Cumulative
  - Two successive DP algorithms have a privacy loss of at most $\epsilon = \epsilon_1 + \epsilon_2$
- Future proof
  - Even if other datasets become available

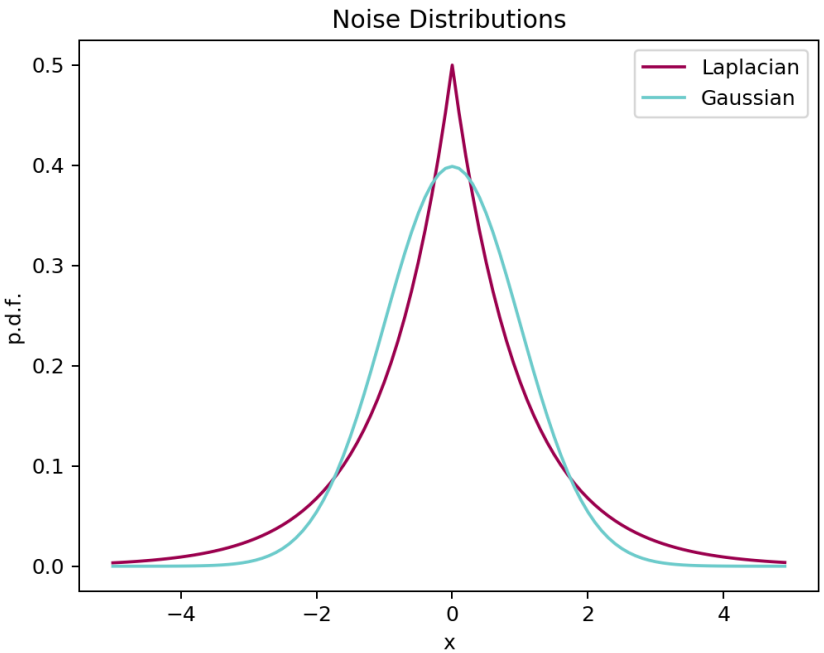
Calculating noise requires 2 parameters:

- "sensitivity" of the query

$$\Delta f = \max \|f(x) - f(y)\|$$

Calculated by finding the maximum change in the query function when each data point is removed.

- privacy budget ϵ gets "used up" with each query
  Computer scientists think ϵ should be ≈ 1
  Social scientists think ϵ should be ≈ 10
  FAANG use 2-4 for predictive text, emoji usage, etc.



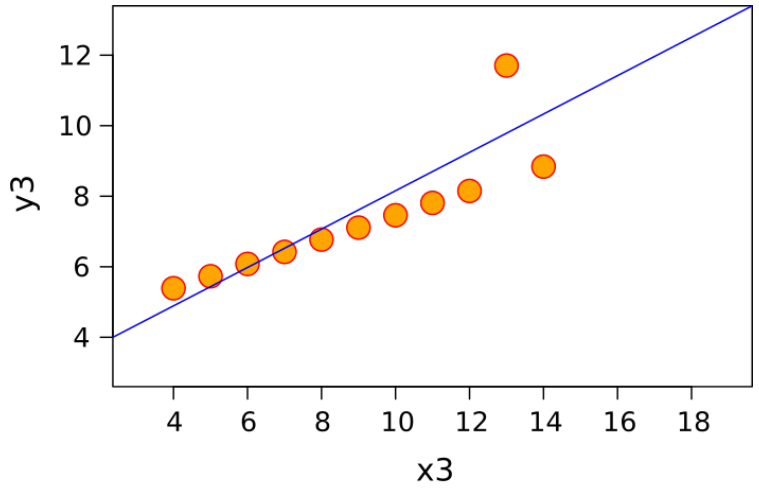
$$\frac{\varepsilon}{2 \cdot \Delta f} exp\left(\frac{-|x|\varepsilon}{\Delta f}\right)$$

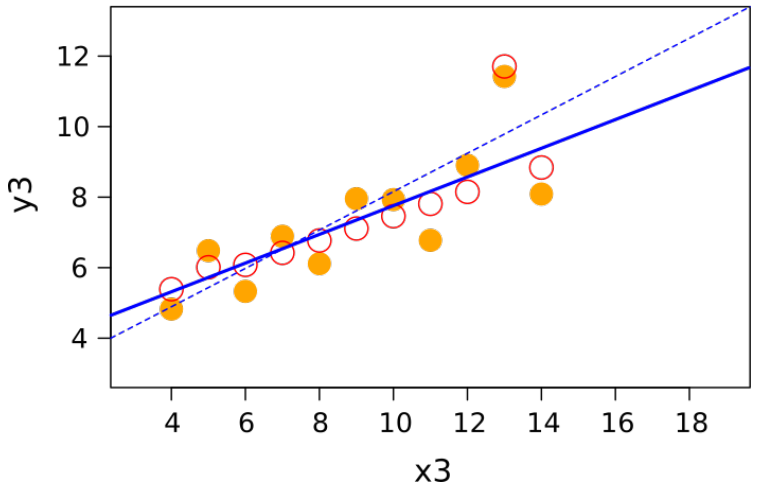
Types of differential privacy

- Local
  - Noise added to individual responses
  - at the time data is generated
  - without regard for other results
- Central
  - Noise added by trusted authority
  - after data have been collected
  - selectively, to preserve important correlations

DP & ML : like chocolate and peanut butter

- The noise added by DP must be on the same order as the outliers
  - Everyone receives "equal protection" (can be important legally)
  - This reduces overfitting to the outliers
- The data are noisier, so larger sample sizes are required
- Ideal for situations when data wouldn't have been available at all





PATE Overview

Private Aggregation of Teacher Ensembles

- Form disjoint partitions of private data
- Train a model on each partition ("teachers")
- Use ensemble prediction to label public data
- Add noise to the ensemble to provide privacy
- Train a "student" model using labeled public data

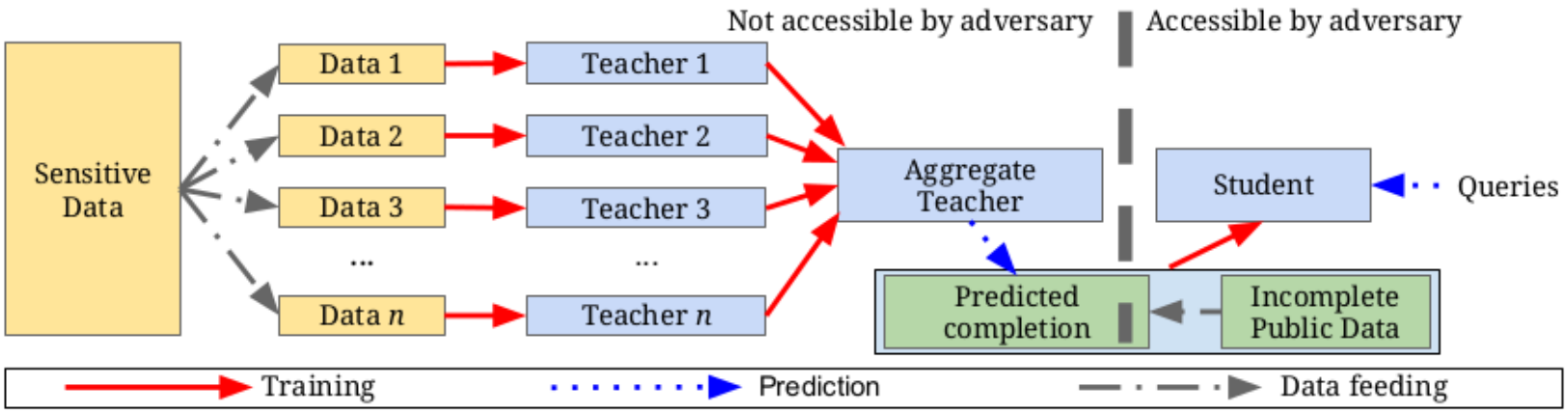


Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

Teacher Consensus

For a given dataset, privacy loss is lower with

- More consensus, therefore
- More teachers

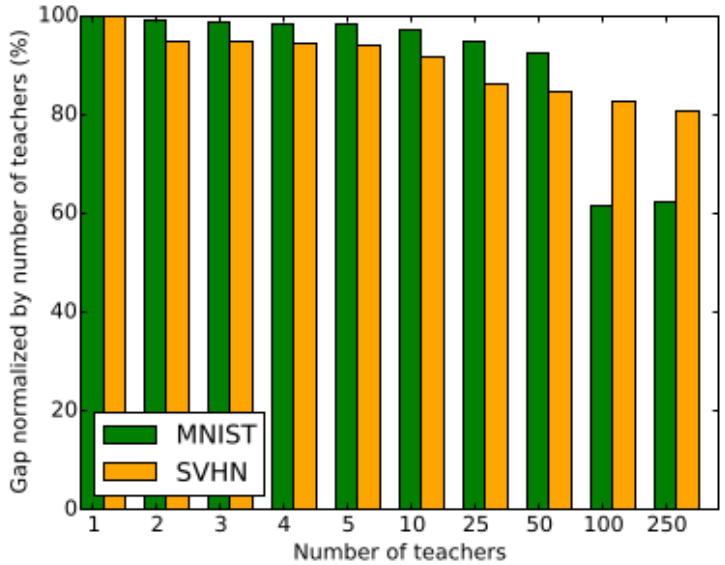


Figure 3: **How certain is the aggregation of teacher predictions?** Gap between the number of votes assigned to the most and second most frequent labels normalized by the number of teachers in an ensemble. Larger gaps indicate that the ensemble is confident in assigning the labels, and will be robust to more noise injection. Gaps were computed by averaging over the test data.

Noise vs Accuracy

For a given dataset, privacy loss is lower with

- More consensus
- More teachers

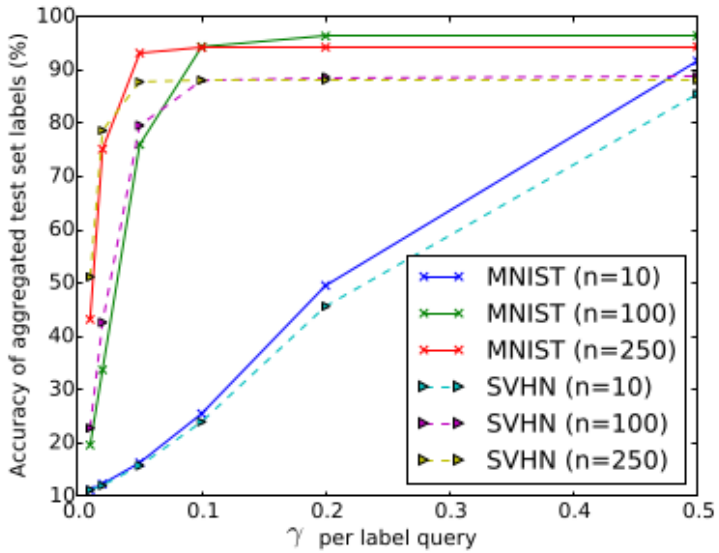


Figure 2: **How much noise can be injected to a query?** Accuracy of the noisy aggregation for three MNIST and SVHN teacher ensembles and varying γ value per query. The noise introduced to achieve a given γ scales inversely proportionally to the value of γ: small values of γ on the left of the axis correspond to large noise amplitudes and large γ values on the right to small noise.

Noise vs Accuracy

For a given dataset, privacy loss is lower with

- More consensus
- More teachers

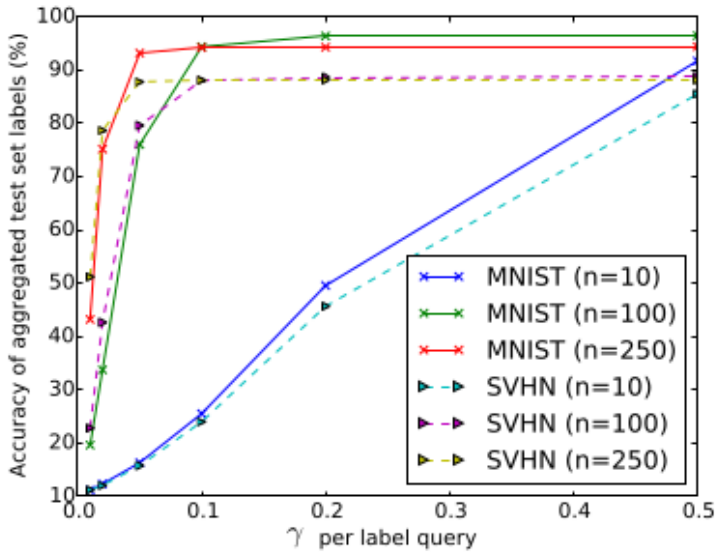


Figure 2: **How much noise can be injected to a query?** Accuracy of the noisy aggregation for three MNIST and SVHN teacher ensembles and varying γ value per query. The noise introduced to achieve a given γ scales inversely proportionally to the value of γ: small values of γ on the left of the axis correspond to large noise amplitudes and large γ values on the right to small noise.

Further reading

- PATE / Universal Model
- `https://arxiv.org/abs/1610.05755` - Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data (original paper)
- `https://arxiv.org/abs/1802.08908` - Scalable Private Learning with PATE (updated algorithm, includes Gaussian noise and careful choice of training examples)
- RAPPOR / text analysis
- `https://arxiv.org/abs/1407.6981` - RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response
- `https://arxiv.org/abs/1503.01214` - Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries

Thank you!

Attribution

- https://commons.wikimedia.org/wiki/File:Coin_Toss_(3635981474).jpg
- https://commons.wikimedia.org/wiki/File:Anscombe's_quartet_2.svg
- https://commons.wikimedia.org/wiki/File:Emoji_u1f3eb.svg
- https://commons.wikimedia.org/wiki/File:NeuralNetwork.png