SAND2019-9275PE

Musgrave Ritual: Machine Learning Privacy Attacks and Defenses

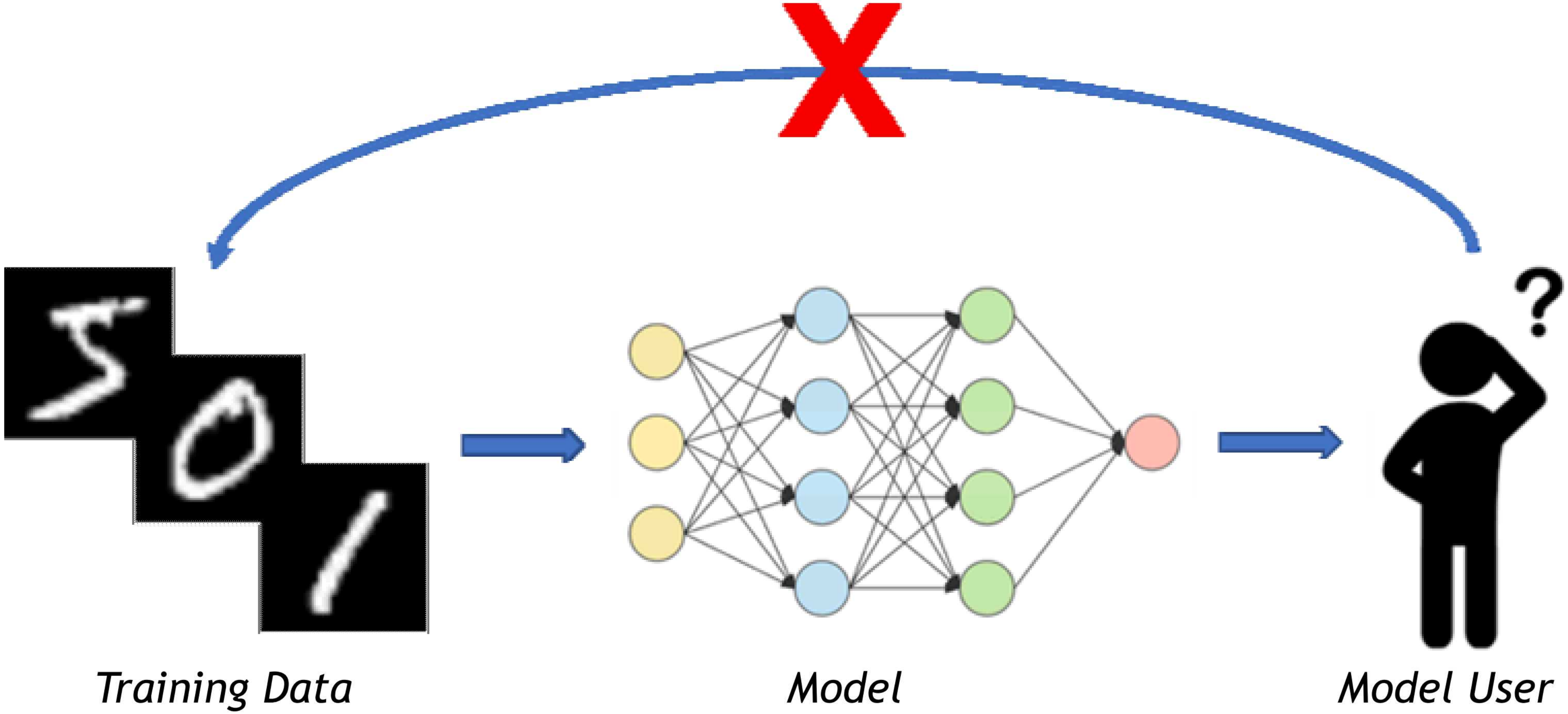
Gary Saavedra, Jeremy Wendt, Philip Kegelmeyer, Joe Bertino, Cosmin Safta

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

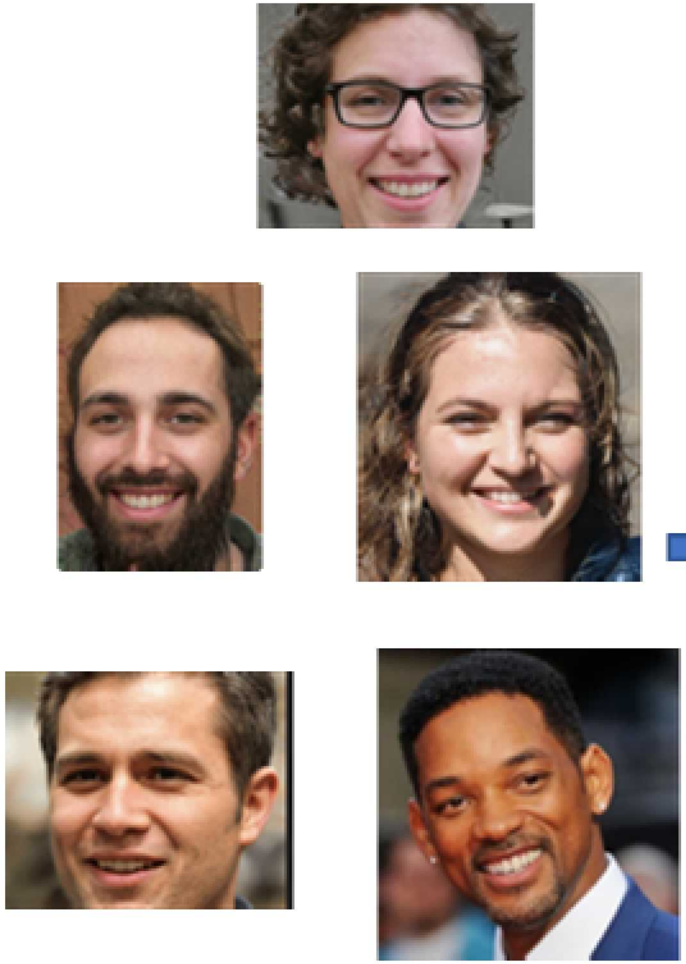
- What is privacy and why does it matter
- The membership inference attack and defending against it
- Experimental Results
  - The difference between defense and no defense
  - Effect of layers and regularization
  - The effects of noise

What does privacy mean in a machine learning context?

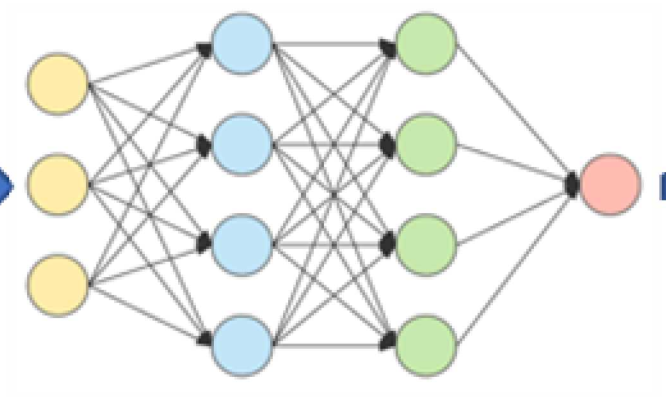


Data used to train a model will not be leaked by the model.

Example – Snapchat has a public model but private data



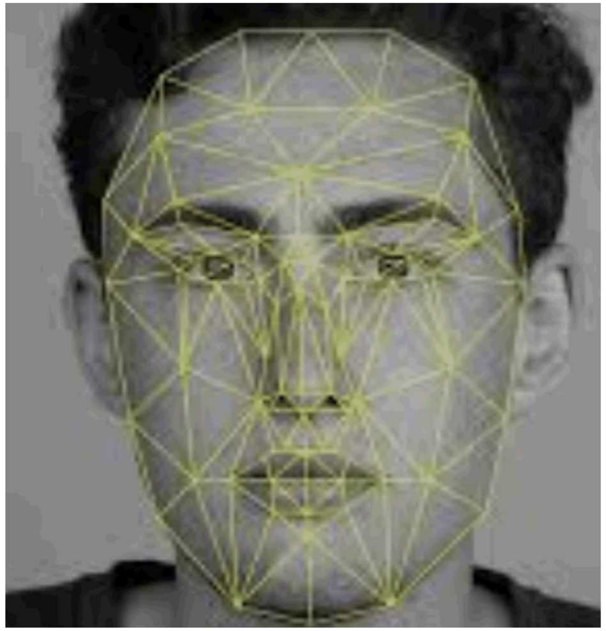
Private – user faces as training data



Public – face detection model

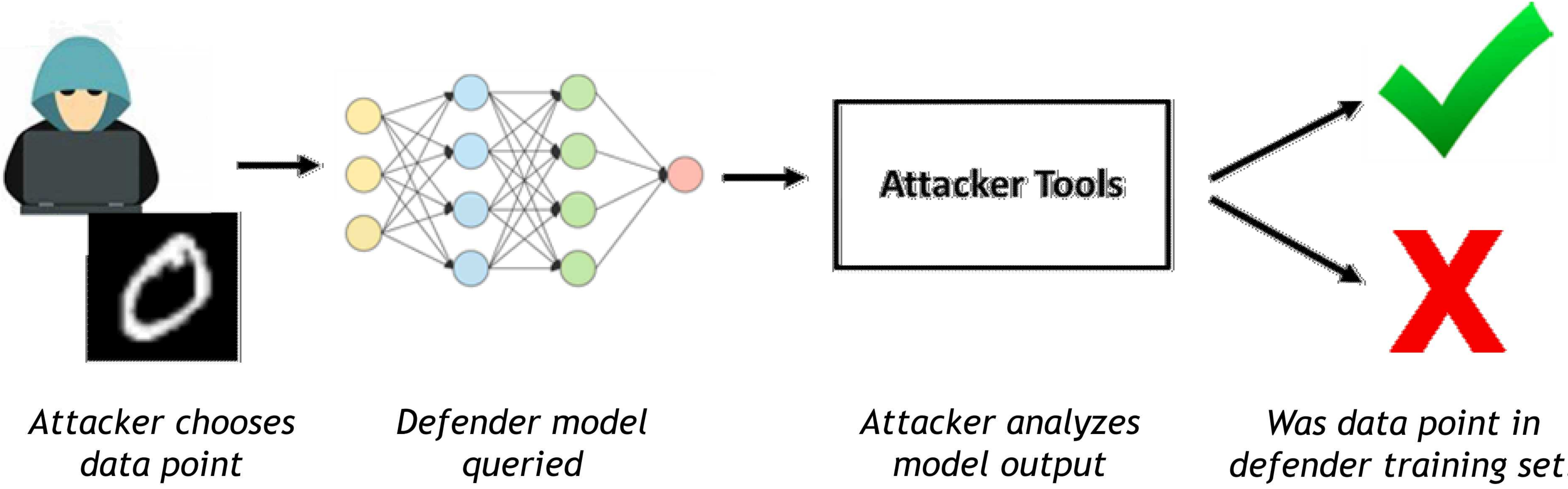
User is free to interact with model

Why does privacy matter?



- Legal risk to leaking information
- Competitive advantage to holding certain data
- Hinders applications of machine learning

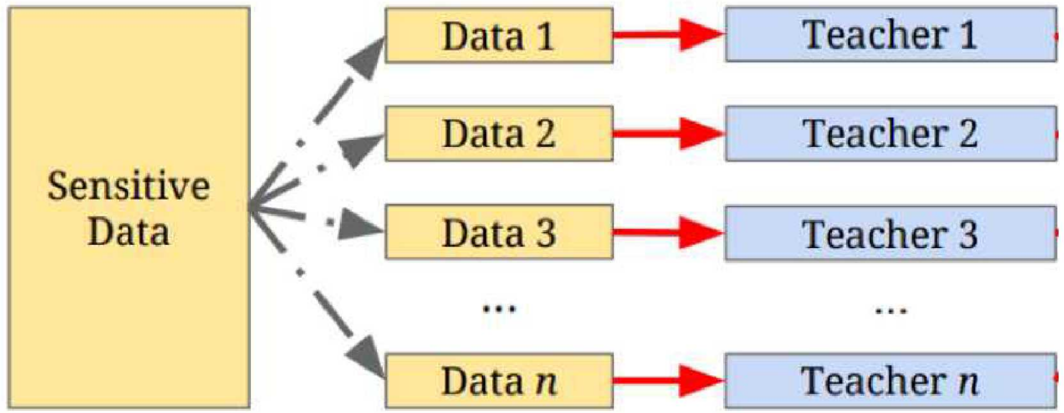
Membership inference attack



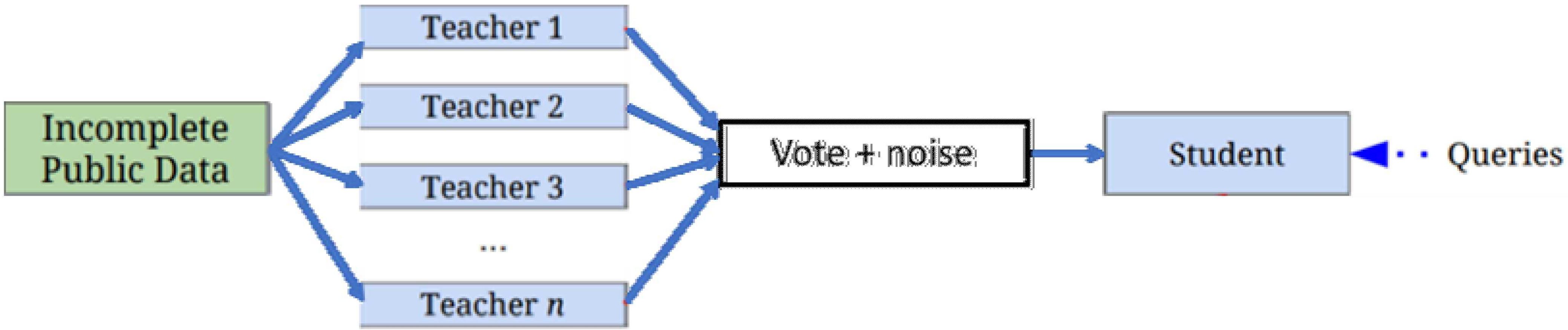
Attacker tests if a specific data point was part of the training set.

Defense - Private Aggregation of Teacher Ensembles (PATE)

Step 1:



Step 2:



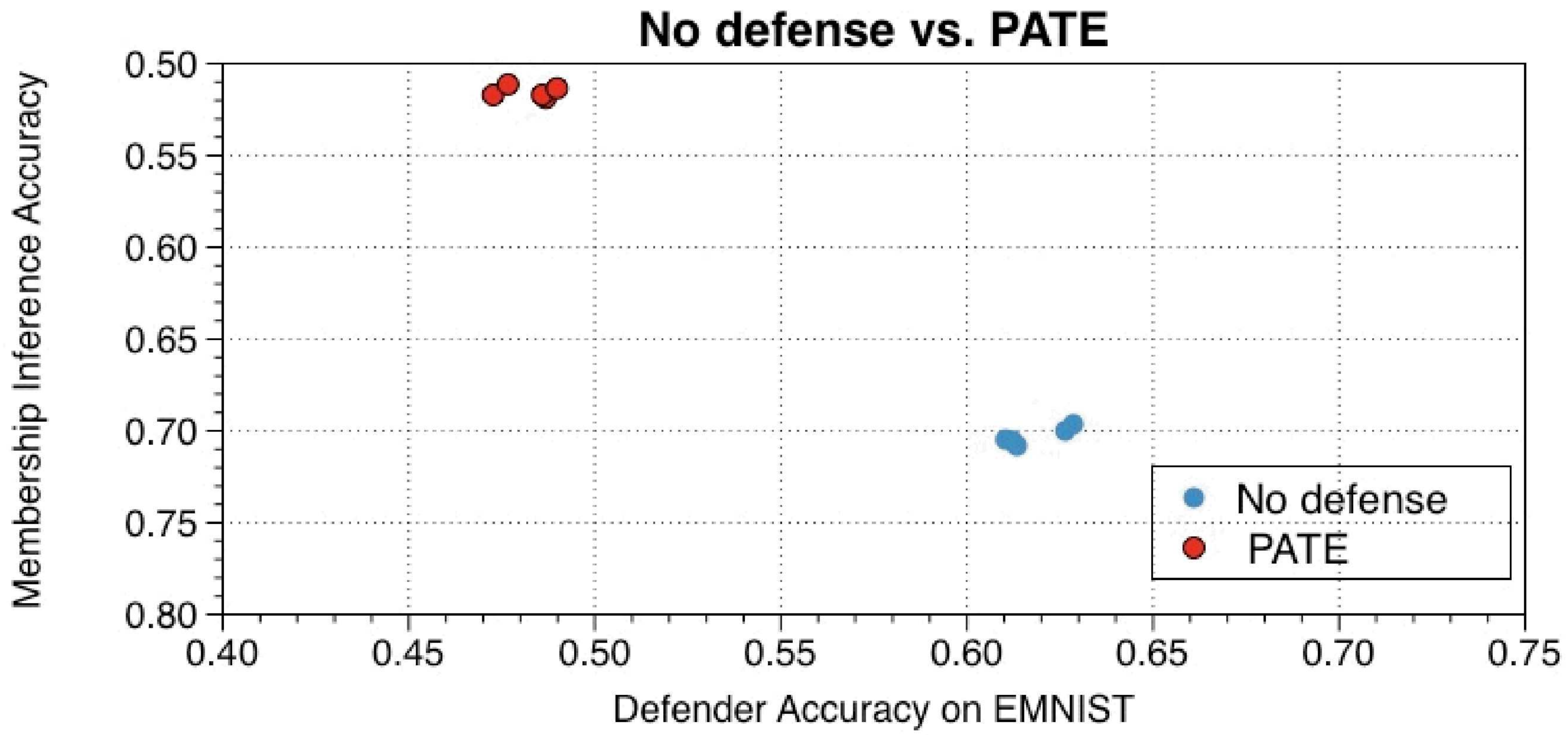
Defend sensitive data by using noise and data partitioning.

Semi-supervised knowledge transfer for deep learning from private training data. Papernot et al.

Experimental Setup

- **Data** - Extended MNIST (EMNIST)
  - 47 classes
  - Digits and letters
- **Model** - Neural networks
  - Attacker only has access to confidence outputs
- Leveraged Kahuna to run parameter sweeps

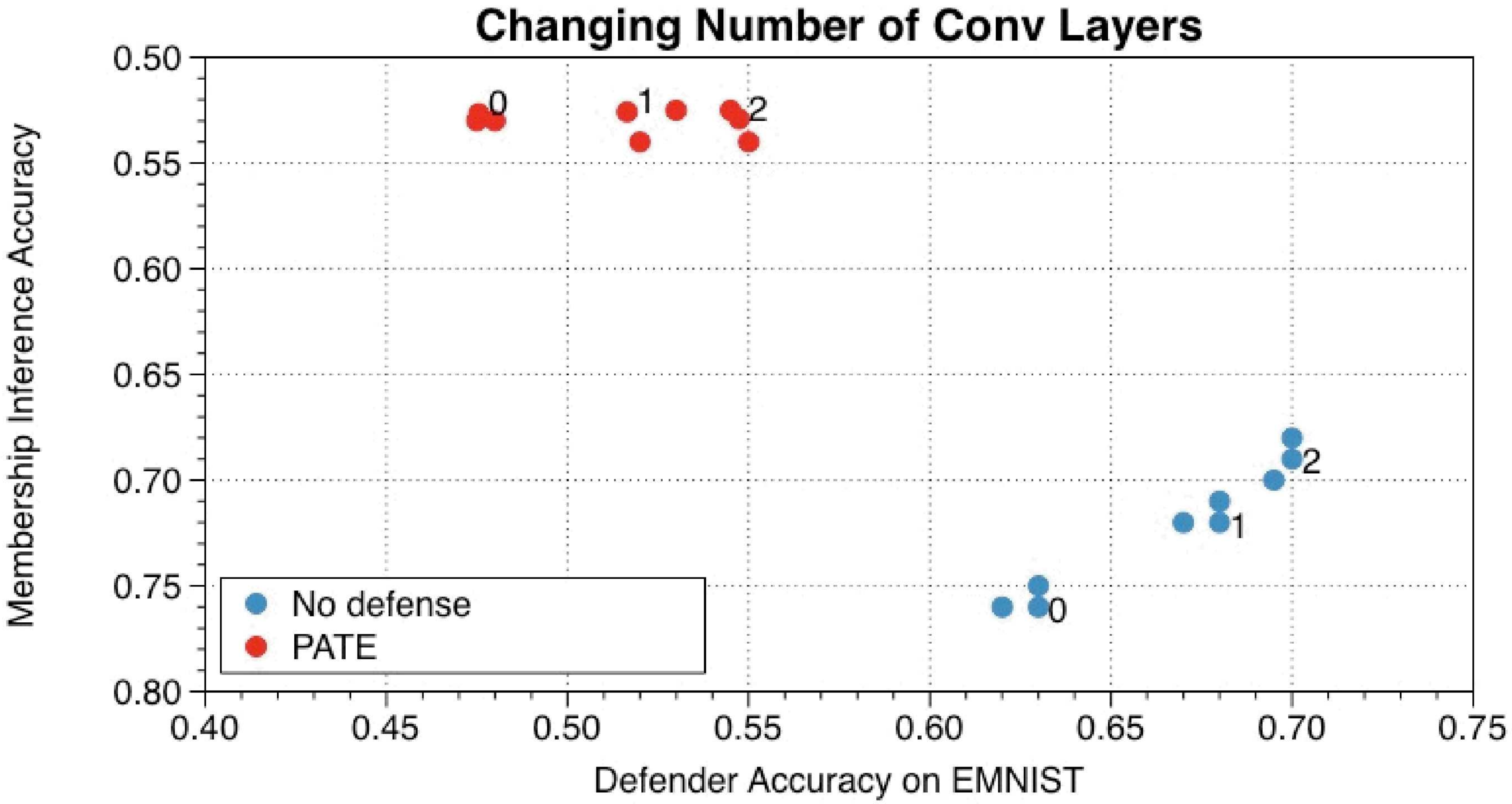
Results – No privacy vs. PATE protection



x-axis: random chance = 0.02
y-axis: random chance = 0.5

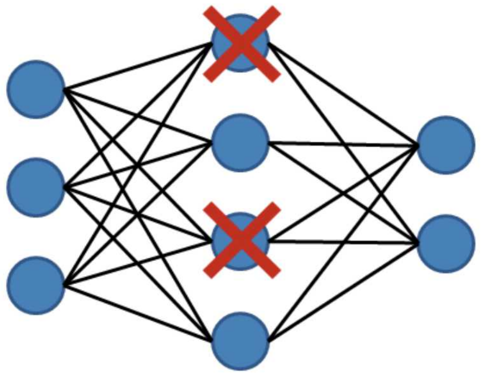
PATE drastically reduces vulnerability to membership inference.

Results – Effect of convolutional layers on privacy

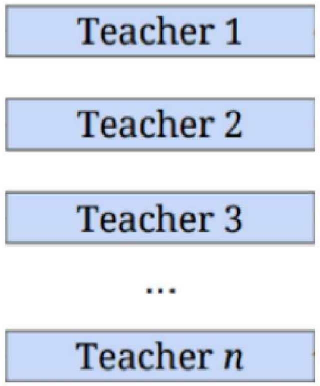


Convolution offers a way to improve accuracy and privacy simultaneously.

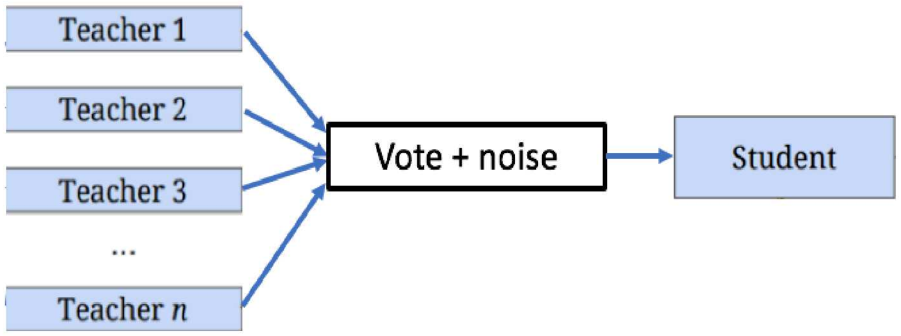
Results – Dropout and our variations of it



- **Typical use** – randomly drop nodes during training process

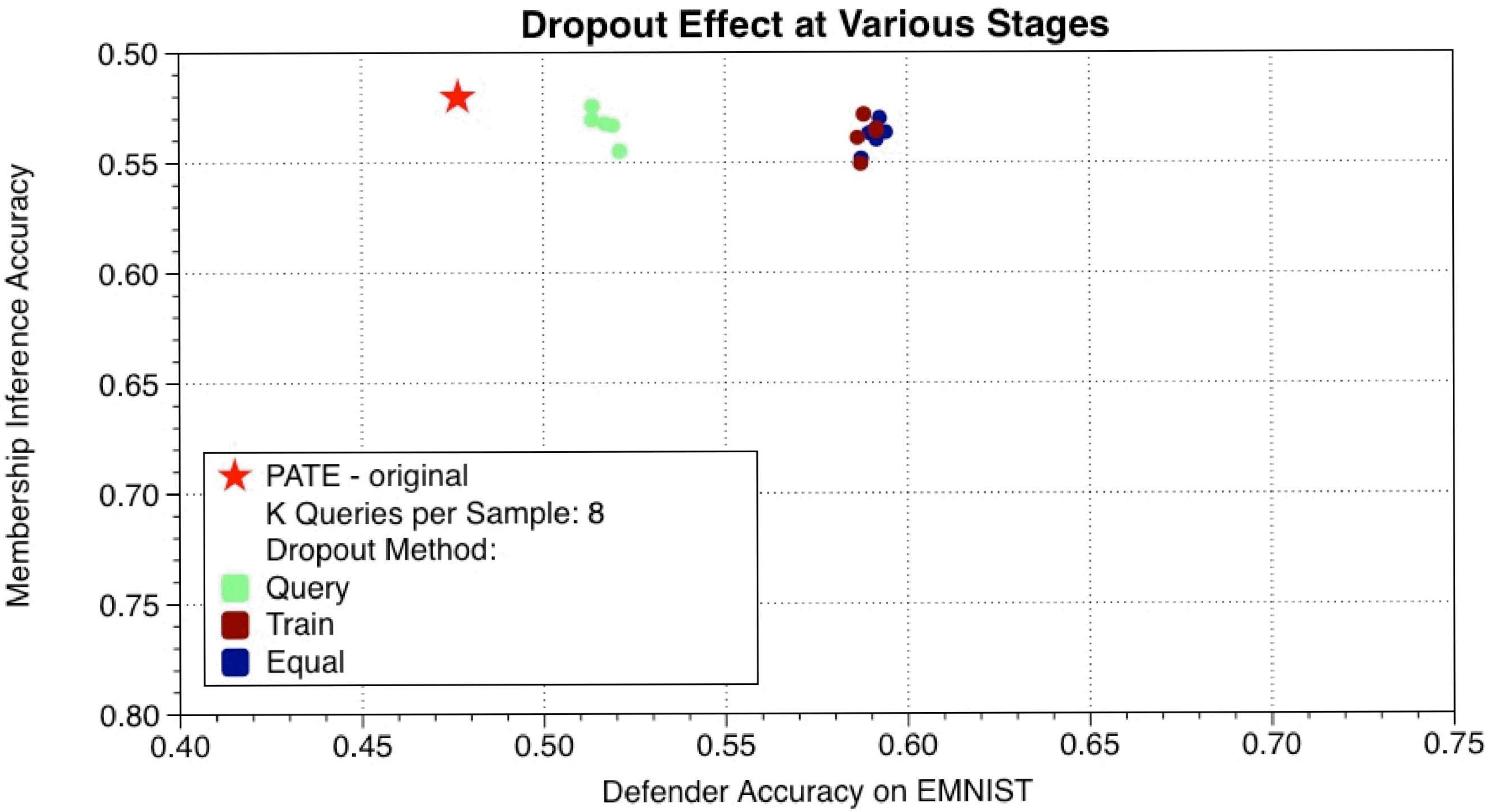


- **train** – dropout teacher nodes during teacher training



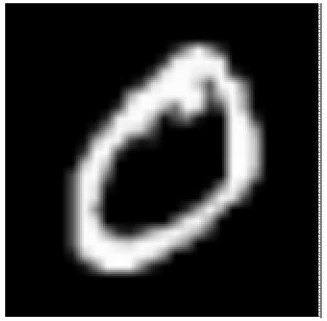
- **query** – dropout teacher nodes when student queries

Results – Dropout as a privacy defense

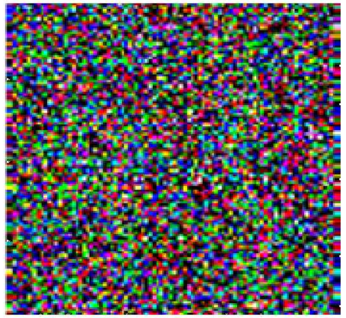


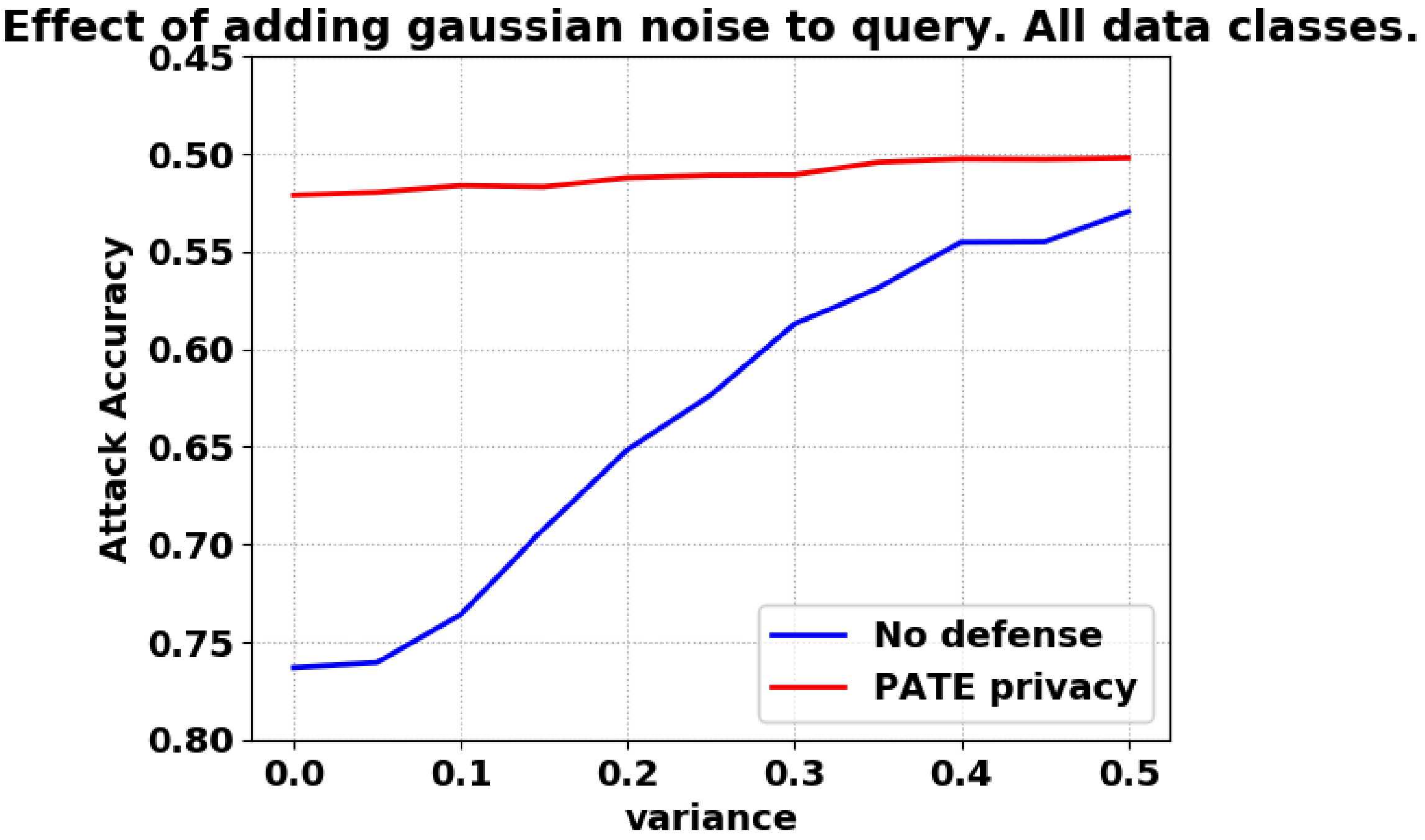
The effectiveness of dropout as a defense depends on where it is applied.

Results – Does the attacker need an exact copy of the data point?



+





Attack can still be successful even with noisy version of training points.

Conclusion and Future Work

- **Important takeaways**
  - Various hyperparameters and regularization schemes affect privacy
  - Even black box models are vulnerable to membership attacks
  - Privacy in machine learning is still a young field
- **Future work**
  - Understand extent to which dropout offers protection
  - Vary images in different ways – rotations, cropping, etc. and test the effect on membership inference
  - Develop new attacks and defenses
  - Try different datasets

Feel free to contact us with questions or comments.

- **Presenter:**
  - Gary Saavedra
  - gjsaave@sandia.gov

- **Project lead:**
  - Jeremy Wendt
  - jdwendt@sandia.gov