# Defenses against Poisoning Attacks

Neil Gong

# General strategy

- Prevention
- Detection
- Recovery

# Prevention – robust mean gradient estimation

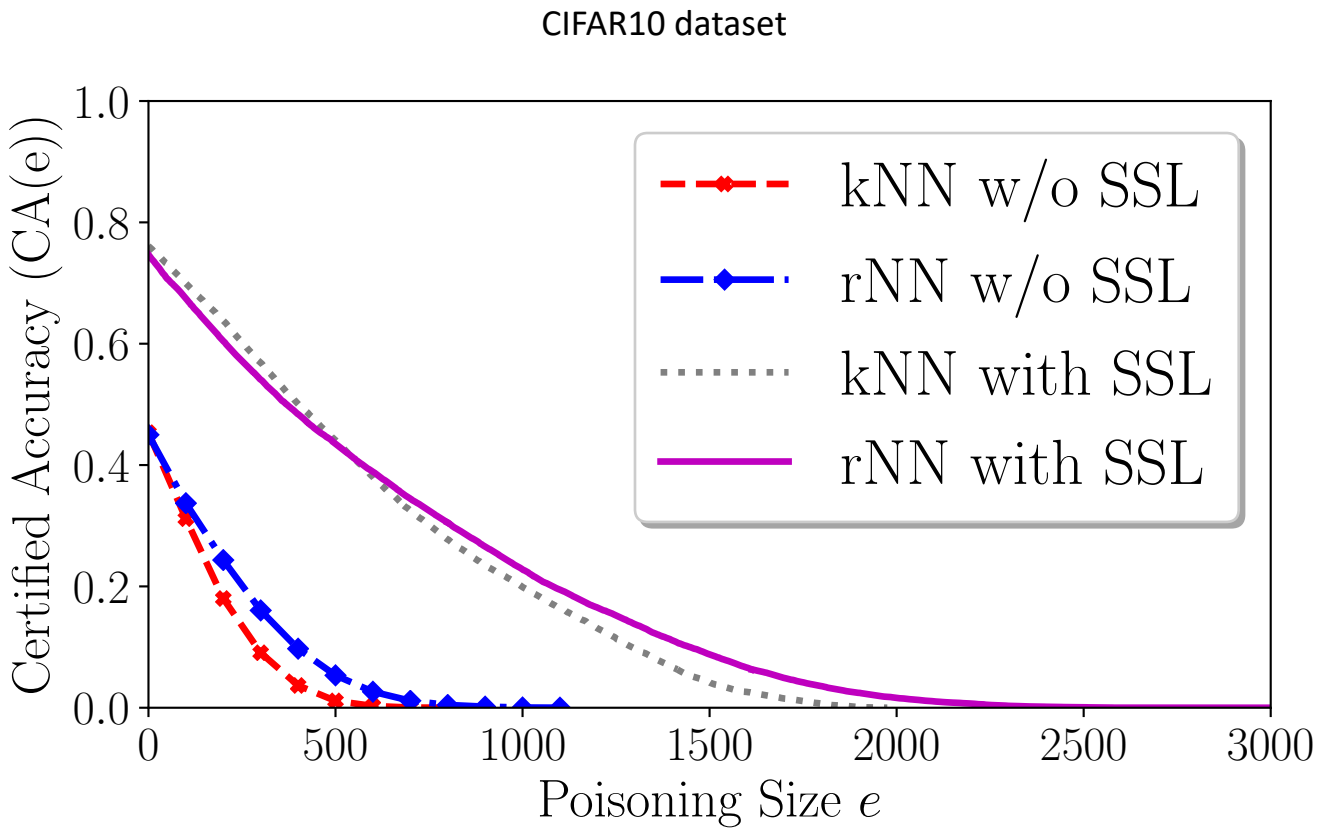
- Data poisoning
  - Median
  - Trimmed mean
- Model/Computation poisoning in federated learning
  - Median
  - Trimmed mean

# Prevention –provably robust defenses

- Data poisoning
  - Randomized smoothing
    - *Certified Robustness to Label-Flipping Attacks via Randomized Smoothing*
  - Differential privacy
    - *Data Poisoning against Differentially-Private Learners: Attacks and Defenses*
  - Majority vote
    - Bagging
    - Nearest neighbor
  - Generalization to graph-based methods
  - Generalization to recommender systems?

- Model/Computation poisoning in federated learning
  - Majority vote
    - Bagging

# Some results



# Detection

- Data poisoning
  - Remove “bad” training data points
  - Outlier detection
  - Graph anomaly detection
  - Fake user detection in recommender systems

- Model/Computation poisoning in federated learning
  - Detecting malicious clients

# Recovery

- Detecting poisoned training examples after training

- How to efficiently recover a model without re-training from scratch?
  - *Towards Making Systems Forget with Machine Unlearning*
  - *DeltaGrad: Rapid retraining of machine learning models*
  - *FedRecover: Recovering from Poisoning Attacks in Federated Learning using Historical Information*