Training/Testing Data Privacy

Neil Gong

ML confidentiality/privacy

- Model parameter/hyperparameter
- Training data
- Testing data
- Algorithms

Privacy vs. confidentiality

- Confidentiality: broader concept

- Privacy: related to sensitive information of human

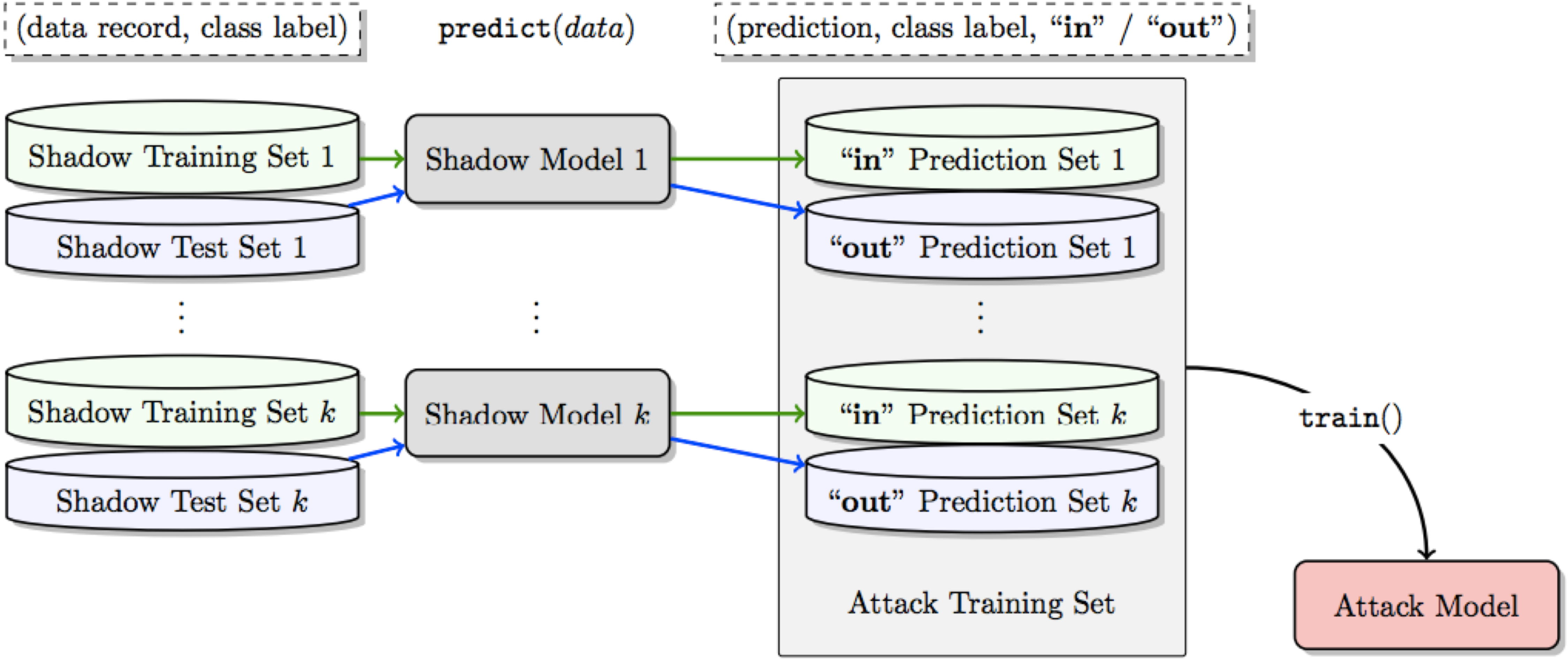
Training data privacy

- Attacker’s goal
  - Membership inference
  - Property inference
  - Training data distribution
  - Training data reconstruction
  - Attribute inference

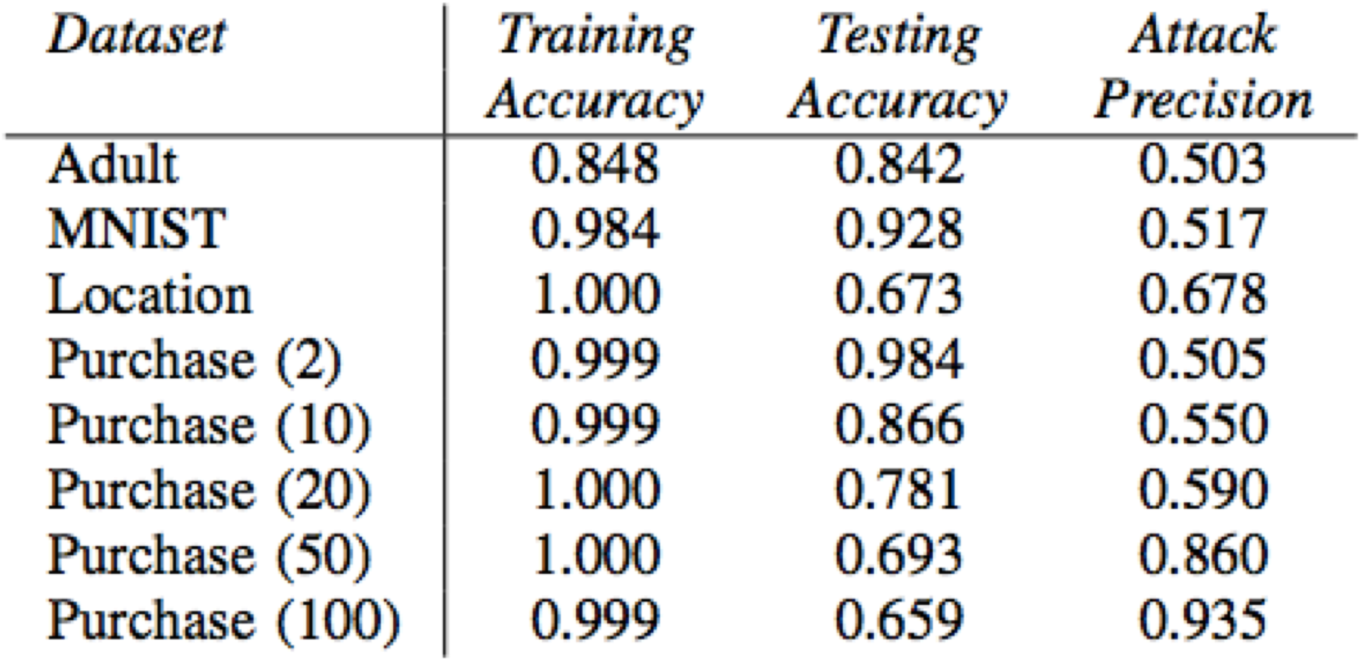
- Attacker’s background knowledge
  - Model parameters
  - Prediction API

- Attacker’s capabilities
  - Analyze model parameters
  - Querying the prediction API

Membership inference attacks

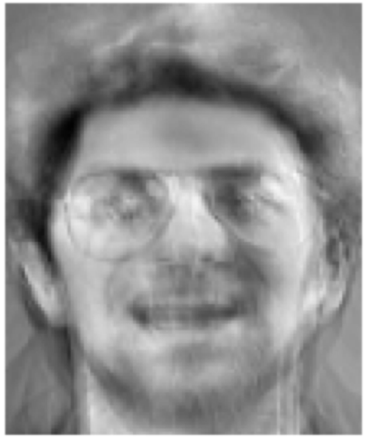


Results

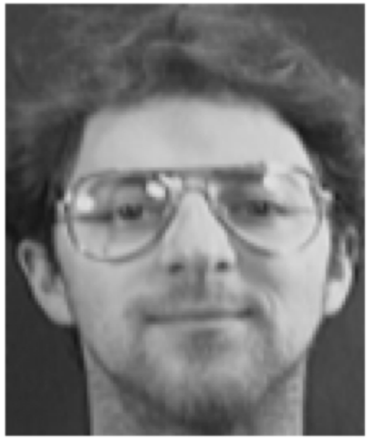


Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935

Training data distribution inference



Inferred image



Real image

Testing data privacy

- Machine learning as a cloud service
  - Query
  - Predicted label

- Attribute/Feature inference
  - Input
    - Prediction results
    - Partial features
  - Output
    - Missing features