Trusting Machine Learning: Privacy, Robustness, and Interpretability Challenges

Reza Shokri

Joint work with: Amir Houmansadr, Prateek Mittal, Milad Nasr, Vitaly Shmatikov, Liwei Song, Congzheng Song, Martin Strobel, Marco Stronati, Yair Zick
Internet

Search engines, recommender systems, social networks, personalized services, …

Web trackers

Data aggregators
Internet

Search engines, recommender systems, social networks, personalized services, …

Data aggregators
Machine Learning

• Minimize the learning loss
• Maximize the predictive power
Machine Learning

- Massive amount of data
- Large models
Beyond Prediction Accuracy

- Privacy
- Robustness
- Transparency
  - Explainability
- Fairness
Components - Data
Components - Algorithms
Components - Computing Platforms

MLaaS
Machine Learning as a Service

Data
- Blobs and Tables
- Hadoop (HDInsight)
- Relational DB (Azure SQL DB)

Clients
- Model is now a web service that is callable
- Monetize the API through our marketplace

Integrated development environment for Machine Learning
Components - Users

Machine Intelligence LANDSCAPE

Core Technologies

Rethinking Enterprise

Sales
- PreEd
- Avizo
- NGDATA
- Intel

Security / Authentication
- Crossmatch
- Si2
- ThreatMetrix

Fraud Detection
- Sift Science
- Ver subtraction

HR / Recruiting
- TalentBin
- RecruitIQ
- Groundspeak

Marketing
- BrightMae
- CommandIQ
- RADIUS

Personal Assistant
- Telzor
- Movidio

Intelligence Tools
- Siri
- Cortana
- Xiao Murray

Rethinking Industries

ADTech
- Metanews
- Distillery
- YieldMo

Agriculture
- Blue River
- The Climate Corporation

Education
- Clarion
- CourseSource
- Kidaptive

Finance
- Bloomberg
- alphapense
- Datamatics

Legal
- Legal Machines
- CounselorList
- JUDICATA

Manufacturing
- Microscan
- Tevisys

Medical
- Parzival
- Zephyr

Oil and Gas
- Kaggle
- Ayasdi

Media / Content
- Outbrain
- sailthru

Consumer Finance
- EarlyPay
- LendIt

Philanthropies
- DataKind
- thorn

Automotive
- Google
- Cruise

Diagnostics
- Enlitic
- Scan

Retail
- BAY SENSORS
- Gen ease

Rethinking Humans / HCI

Augmented Reality
- APX
- Meta

Gestural Computing
- Leap
- myTouch

Robotics
- DreamWorks
- Intellectual Property

Emotional Recognition
- Mindset
- Gomint

Supporting Technologies

Hardware
- NVIDIA
- LG

Data Prep
- Talend
- Alteryx

Data Collection
- Diffbot
- KDDI

www.shionzilis.com/machineintelligence
Privacy Risks

Direct access to training data set

Access to sensitive query inputs
Privacy Risks

Input inference attack:
Given $f(x; W)$, infer input $x$

Fredrikson, et al. "Model inversion attacks that exploit confidence information …" CCS 2015
Privacy Risks

Model extraction attack: Given $f(x; W)$, infer model $W$

Privacy Risks

Reconstruction and Membership Inference attacks:
Given $W$ or $f(x; W)$, infer about the training set $X$

Privacy Risks

Data Inference in Distributed Training

Privacy in Machine Learning

Data Privacy

- Training could be outsourced, thus the training data is visible to (untrusted) entities

- Given the parameters or predictions, an attacker can infer the training data
Machine Learning as a Service
Machine Learning as a Service
Machine Learning as a Service
Machine Learning as a Service
Machine Learning as a Service

Prediction API

Training API

DATA
Machine Learning as a Service
Machine Learning as a Service

Model

Prediction API

Input data

Classification

Training API

DATA

airplane
automobile
...
ship
truck
Information Leakage

Do model’s predictions leak information about training data?
“Leakage”

being able to learn information about the training data, which cannot be learned from other models/data (from the same distribution)

Do model’s predictions leak information about training data?
Membership Inference Attack

Was this specific data record part of the training set?

Model

Prediction

Input data

Classification

Training

Data

airplane
automobile
...
ship
truck
Exploit Model’s Predictions

Model

Prediction API

Training API

DATA
Exploit Model’s Predictions

Main insight: ML models overfit to their training data
Exploit Model’s Predictions

Main insight: ML models overfit to their training data

Model

Prediction API

Training API

Input from the training set

Classification

DATA
Exploit Model’s Predictions

Main **insight**: ML models overfit to their training data

**Model**

- **Prediction API**
  - Input from the training set
  - Classification
- **Training API**
  - Input **NOT** from the training set
  - Classification

**DATA**
Exploit Model’s Predictions

Model

Prediction API

Input from the training set

Classification

Classification

Input NOT from the training set

Training API

DATA

Recognize the difference
ML against ML

Train a ML model to recognize the difference
Train the attack model to predict if an input was a member of the training set (in) or a non-member (out)
Construct the Attack Model
Construct the Attack Model

Using the Attack Model

one single data record

Model

Prediction API

classification

Attack Model

Model

Prediction API

membership probability
Purchase Dataset — Classify Customers (100 classes)

Cumulative Fraction of Classes

Membership inference precision

Overall accuracy:
- Shadows trained on real data: 0.93
- Shadows trained on synthetic data: 0.89

Real Data
Marginal-Based Synthetic
Model-Based Synthetic
Without knowing the
- training algorithm
- model type and architecture
- model parameters
- data distribution

Cumulative Fraction of Classes

Membership inference precision

Overall accuracy:
- Shadows trained on real data: 0.93
- Shadows trained on synthetic data: 0.89

Purchase Dataset — Classify Customers (100 classes)
Privacy                   Learning

Note: information leakage wrt the model’s output (predictions)
Does the model leak information about data in the training set?

Note: information leakage wrt the model’s output (predictions)
Privacy

Does the model leak information about data in the training set?

Learning

Does the model generalize to data outside the training set?

Note: information leakage wrt the model’s output (predictions)
Privacy

Does the model leak information about data in the training set?

Learning

Does the model generalize to data outside the training set?

Overfitting is the common enemy!

Note: information leakage wrt the model’s output (predictions)
Privacy-Utility: Not in a Direct Conflict!

Utility
(prediction accuracy)

Privacy-preserving machine learning
Learning

Empirical loss over D:

\[ L_D(f) = \frac{1}{|D|} \sum_{(x,y)\in D} l(f(x), y) \]

Learning optimization:

\[ \min_f L_D(f) + \lambda R(f) \]

maximize prediction accuracy
Privacy as a Learning Objective

minimize inference accuracy

\[ \Pr((x, y) \in D) = h(x, y, f(x)) \]

maximize prediction accuracy
**Minmax Membership Privacy Game**

Adversarial Regularization

\[
\min_f \left( L_D(f) + \lambda \max_h G_{f,D,D'}(h) \right)
\]

inference model with maximum gain (distinguishing members from non-members)

classification model with minimum loss (predicting correct class for any input)

optimal inference

optimal privacy-preserving classification
Train using SGD

\[ h(x, y, f(x)) \quad h(x', y', f(x')) \]

\[ \text{inference model } h \]

\[ f(x) \quad f(x') \]

\[ \text{classifier } f \]

\[ x \quad x' \]

\[ D \quad D' \]

Gain: \( \frac{1}{2} \log(h(x, y, f(x))) + \frac{1}{2} \log(1 - h(x', y', f(x'))) \)

Loss: \( l(f(x), y) + \lambda \log(h(x, y, f(x))) \)

\( \min_f \left( L_D(f) + \lambda \max_h G_{f, D, D'}(h) \right) \)

optimal privacy-preserving classification
Membership Privacy Mechanism as a Regularizer

![Graph showing classifier loss over epochs with and without defense. The graph demonstrates a significant reduction in classifier loss when defense is applied.]
# Privacy

## Attack accuracy:

near random guess

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Without defense</th>
<th>With defense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training accuracy</td>
<td>Testing accuracy</td>
</tr>
<tr>
<td>Purchase100</td>
<td>100%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Texas100</td>
<td>81.6%</td>
<td>51.9%</td>
</tr>
<tr>
<td>CIFAR100- Alexnet</td>
<td>99%</td>
<td>44.7%</td>
</tr>
<tr>
<td>CIFAR100- DenseNET</td>
<td>100%</td>
<td>70.6%</td>
</tr>
</tbody>
</table>
### Generalization

**Defeat overfitting**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Without defense</th>
<th>With defense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training accuracy</td>
<td>Testing accuracy</td>
</tr>
<tr>
<td>Purchase100</td>
<td>100%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Texas100</td>
<td>81.6%</td>
<td>51.9%</td>
</tr>
<tr>
<td>CIFAR100- Alexnet</td>
<td>99%</td>
<td>44.7%</td>
</tr>
<tr>
<td>CIFAR100- DenseNET</td>
<td>100%</td>
<td>70.6%</td>
</tr>
</tbody>
</table>

- Low Training accuracy,
- Smaller gap between the training & testing accuracies
## Generalization

**Defeat overfitting**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Without defense</th>
<th>With defense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training accuracy</td>
<td>Testing accuracy</td>
</tr>
<tr>
<td>Purchase100</td>
<td>100%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Texas100</td>
<td>81.6%</td>
<td>51.9%</td>
</tr>
<tr>
<td>CIFAR100- Alexnet</td>
<td>99%</td>
<td>44.7%</td>
</tr>
<tr>
<td>CIFAR100- DenseNET</td>
<td>100%</td>
<td>70.6%</td>
</tr>
</tbody>
</table>

**L2-regularization factor**

<table>
<thead>
<tr>
<th>L2-factor</th>
<th>Training accuracy</th>
<th>Testing accuracy</th>
<th>Attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (no regularization)</td>
<td>100%</td>
<td>80.1%</td>
<td>67.6%</td>
</tr>
<tr>
<td>0.001</td>
<td>86%</td>
<td>81.3%</td>
<td>60%</td>
</tr>
<tr>
<td>0.005</td>
<td>74%</td>
<td>70.2%</td>
<td>56%</td>
</tr>
<tr>
<td>0.01</td>
<td>34%</td>
<td>32.1%</td>
<td>50.6%</td>
</tr>
</tbody>
</table>

**Low Training accuracy, Smaller gap between the training & testing accuracies**

**High cost for same privacy**
White-box Privacy Analysis

What if the adversary observes the model parameters?
Extending Black-box Inference Attack to Activation Functions?
Stochastic Gradient Descent
Stochastic Gradient Descent
Gradient of Loss on Members vs. Non-members

![Graph showing the gradient of loss over training epochs for members and non-members. The graph indicates a sharper decline and stabilization for members compared to non-members.]
Attack Features

\[ h_1(x), h_2(x), \ldots, f(x) \]

\[ \frac{\partial L}{\partial W_1}, \frac{\partial L}{\partial W_2}, \ldots, \frac{\partial L}{\partial W_i} \]

\[ (f(x; W)), y \]

\[ L \]

\[ x, W_1, W_2, \ldots, W_i \]

\[ \text{target model} \]
Generalizability and (?) Privacy in the white-box setting

<table>
<thead>
<tr>
<th>Pre-trained</th>
<th>Target Model</th>
<th>Attack Accuracy</th>
<th>Attack Accuracy</th>
<th>Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dataset</td>
<td>Architecture</td>
<td>Train Accuracy</td>
<td>Test Accuracy</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>74.2%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td>62.2%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td>67.7%</td>
</tr>
<tr>
<td>Texas100</td>
<td>Fully Connected</td>
<td>81.6%</td>
<td>52%</td>
<td>63.0%</td>
</tr>
<tr>
<td>Purchase100</td>
<td>Fully Connected</td>
<td>100%</td>
<td>80%</td>
<td>67.6%</td>
</tr>
</tbody>
</table>

**High** generalizability (Best available models)

**Low** privacy (Significant leakage through parameters)
**Generalizability and (?!?) Privacy in the white-box setting**

<table>
<thead>
<tr>
<th>Pre-trained</th>
<th>Target Model</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Architecture</td>
<td></td>
<td></td>
<td>Black-box</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>74.2%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td>62.2%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td>67.7%</td>
</tr>
<tr>
<td>Texas100</td>
<td>Fully Connected</td>
<td>81.6%</td>
<td>52%</td>
<td>63.0%</td>
</tr>
<tr>
<td>Purchase100</td>
<td>Fully Connected</td>
<td>100%</td>
<td>80%</td>
<td>67.6%</td>
</tr>
</tbody>
</table>

**Large capacity**

**High generalizability** (Best available models)

**Low privacy** (Significant leakage through parameters)
Federated Learning

Aggregator (global parameters $W$)

$f(x; W_1^{\{t\}})$  \hspace{1cm} $\downarrow$

$D_1$

$f(x; W_2^{\{t\}})$  \hspace{1cm} $\uparrow$

$D_2$

$f(x; W_N^{\{t\}})$  \hspace{1cm} $\downarrow$

$D_N$

$W_i^{\{t\}}$

$W^{\{t\}}$

$\ldots$
Federated Learning

• Multiple updates (every epoch)

<table>
<thead>
<tr>
<th>Observed Epochs</th>
<th>Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 10, 15, 20, 25</td>
<td>57.4%</td>
</tr>
<tr>
<td>10, 20, 30, 40, 50</td>
<td>76.5%</td>
</tr>
<tr>
<td>50, 100, 150, 200, 250</td>
<td>79.5%</td>
</tr>
<tr>
<td>100, 150, 200, 250, 300</td>
<td>85.1%</td>
</tr>
</tbody>
</table>

CIFAR100-Alexnet, Global Attack
Federated Learning

• Parameters from all parties are aggregated and sent back to all parties

• One can influence others’ models
Gradient Descent
Active Attack: Gradient “Ascent” on a Target Data Point
Active Attack: Gradient “Ascent” on a Target Data Point

It will be pushed back if the target point is in the training set
## Active Attacks

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Architecture</strong></td>
<td><strong>Global Attacker (the parameter aggregator)</strong></td>
</tr>
<tr>
<td></td>
<td>Passive</td>
</tr>
<tr>
<td></td>
<td>Gradient Ascent</td>
</tr>
<tr>
<td>Alexnet</td>
<td>85.1%</td>
</tr>
<tr>
<td>DenseNet</td>
<td>79.2%</td>
</tr>
</tbody>
</table>
Robustness in Machine Learning

Song, Shokri, Mittal, “Privacy risks of securing machine learning models against adversarial examples”
Robustness Threats

Data Poisoning attacks:
Add adversarially crafted data to X, to alter the model f
Robustness Threats

Parameter Poisoning attacks:
Share adversarially crafted parameter updates to alter the global model $f$

Federated Learning:
Distributed aggregation of local models
Robustness Threats

Adversarial Input (evasion) attacks:
Imperceptible perturbations of input to cause misclassification
Adversarial Inputs: Attack

© Aleksander Madry
Adversarial Inputs: Defense

Minimize the loss on adversarial perturbations
- empirical method (train on adv points)
- theoretical (compute upper bound on adv loss)
Robust Training

Baseline (standard SGD)

CIFAR10
Consequences

Robust Training

Baseline (standard SGD)

Training set members are more distinguishable from non-members

CIFAR10
## Inference on Perturbed Points

<table>
<thead>
<tr>
<th>Training method</th>
<th>train acc</th>
<th>test acc</th>
<th>adv-train acc</th>
<th>adv-test acc</th>
<th>inference acc ($I_B$)</th>
<th>inference acc ($I_A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>100%</td>
<td>95.01%</td>
<td>0%</td>
<td>0%</td>
<td>57.43%</td>
<td>50.85%</td>
</tr>
</tbody>
</table>

Not robust
Inference on Perturbed Points

- Lower loss on adversarial perturbations of training data

<table>
<thead>
<tr>
<th>Training method</th>
<th>train acc</th>
<th>test acc</th>
<th>adv-train acc</th>
<th>adv-test acc</th>
<th>inference acc ($I_B$)</th>
<th>inference acc ($I_A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>100%</td>
<td>95.01%</td>
<td>0%</td>
<td>0%</td>
<td>57.43%</td>
<td>50.85%</td>
</tr>
<tr>
<td>PGD-Based Adv-Train [30]</td>
<td>99.99%</td>
<td>87.25%</td>
<td>96.07%</td>
<td>46.59%</td>
<td>74.89%</td>
<td>75.65%</td>
</tr>
</tbody>
</table>
Fine-grained Fingerprinting: Targeted Adversarial Points
Fine-grained Fingerprinting: Targeted Adversarial Points

<table>
<thead>
<tr>
<th>label</th>
<th>(untargeted)</th>
<th>(targeted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>72.70%</td>
<td>74.42%</td>
</tr>
<tr>
<td>1</td>
<td>67.69%</td>
<td>68.88%</td>
</tr>
<tr>
<td>2</td>
<td>80.16%</td>
<td>83.58%</td>
</tr>
<tr>
<td>3</td>
<td>87.83%</td>
<td>90.57%</td>
</tr>
<tr>
<td>4</td>
<td>81.57%</td>
<td>84.47%</td>
</tr>
<tr>
<td>5</td>
<td>81.34%</td>
<td>83.02%</td>
</tr>
<tr>
<td>6</td>
<td>76.97%</td>
<td>79.94%</td>
</tr>
</tbody>
</table>

...
# Perturbation Budget

<table>
<thead>
<tr>
<th>Perturbation budget ($\epsilon$)</th>
<th>train acc</th>
<th>test acc</th>
<th>adv-train acc</th>
<th>adv-test acc</th>
<th>inference acc ($I_B$)</th>
<th>inference acc ($I_A$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/255</td>
<td>100%</td>
<td>93.74%</td>
<td>99.99%</td>
<td>82.20%</td>
<td>64.48%</td>
<td>66.54%</td>
</tr>
<tr>
<td>4/255</td>
<td>100%</td>
<td>91.19%</td>
<td>99.89%</td>
<td>70.03%</td>
<td>69.44%</td>
<td>72.43%</td>
</tr>
<tr>
<td>8/255</td>
<td>99.99%</td>
<td>87.25%</td>
<td>96.07%</td>
<td>46.59%</td>
<td>74.89%</td>
<td>75.65%</td>
</tr>
</tbody>
</table>

Higher perturbation budget → Lower Privacy
Model Capacity

- Resnet: Increase the number of output channels of the residual layers
Privacy vs Robustness vs Accuracy

• Limited capacity
  ✴ Low predictive power
  ✴ Low leakage

• Robustness
  ✴ Needs large capacity
Interpretability in Machine Learning
Interpretability

Explain the Output of f on x:

- Feature-based: highlight important features of x
- Record-based: identify influential data points in X
Feature-based Explanation

Compute the influence of each input feature on the output
Explain using Gradients

- A simple example: two data points \{(-1, 0), (+1,1)\}
- One single (sigmoid) activation function. \( f(x; w) = 1/(1+\exp(-wx)) \)
Explain using Gradients

• A simple example: two data points \{(-1, 0), (+1,1)\}

• One single (sigmoid) activation function. \(f(x; w) = 1/(1+\exp(-wx))\)

Gradient gets smaller on **training** data \(x=1\); and not on test data \(x=+/-1/2\)
Explain using Gradients

• A simple example: two data points \{(-1, 0), (+1,1)\}

• One single (sigmoid) activation function. \( f(x; w) = 1/(1+\exp(-wx)) \)

Gradient gets smaller on **training** data \((x=1)\); and not on test data \((x=+/-1/2)\)

| \(k\) | \(\theta_k\) | \(-\frac{\partial |1-f_{\theta}(1)|}{\partial \theta}(\theta_k)\) | \(f_{\theta_k}(1)\) | \(\frac{\partial f_{\theta_k}}{\partial x}(1)\) | \(\frac{\partial f_{\theta_k}}{\partial x}(\frac{1}{2})\) |
|------|--------------|------------------------------------------------|----------------|-----------------|----------------|
| 0    | 0.0000       | 0.2500                                           | 0.5000         | 0.0000          | 0.0000         |
| 1    | 0.2500       | 0.2461                                           | 0.5622         | 0.0615          | 0.0623         |
| 10   | 1.9069       | 0.1126                                           | 0.8707         | 0.2147          | 0.3829         |
| 100  | 4.5277       | 0.0106                                           | 0.9893         | 0.0479          | 0.3862         |
| 1000 | 6.8967       | 0.0010                                           | 0.9990         | 0.0070          | **0.2060**     |
Effect of Model Complexity

• **Synthetic** data; 2 classes; n_info features; n_cluster clusters in each class

\[ n_{\text{info}} = 3, \ n_{\text{cluster}} = 1 \]
Distinguishability of Members and Non-Members

\[
\frac{\text{median}\{\|\phi_{GRAD}(\vec{x})\|_1 | \vec{x} \in X_{\text{test}}\}}{\text{median}\{\|\phi_{GRAD}(\vec{x})\|_1 | \vec{x} \in X_{\text{train}}\}}
\]
Distinguishability of Members and Non-Members

\[
\frac{\text{median}\{\|\phi_{\text{GRAD}}(\tilde{x})\|_1 | \tilde{x} \in X_{\text{test}}\}}{\text{median}\{\|\phi_{\text{GRAD}}(\tilde{x})\|_1 | \tilde{x} \in X_{\text{train}}\}}
\]
Membership Inference Attack

- Complete output vector
  \( c(\vec{x}) \)
- Label + Explanation
  \( \phi_{GRAD}(\vec{x}) \)

![Graph showing accuracy vs. dropout with two lines: one for complete output vector and one for label + explanation.](Image)
Membership Inference Attack

Complete output vector

\[ c(\vec{x}) \]

Label + Explanation

Same leakage as observing the full precision output
Impact on Minorities

![Graph showing the relationship between attack accuracy on class and number of records in class. The graph includes a trend line and a red circle highlighting a specific data point.]

- **Y-axis**: Attack accuracy on class
- **X-axis**: Number of records in class
- **Trend**: A downward trend is observed as the number of records increases.
Record-based Explanation

- Release the most influential data points in the training set on the classification of a target data point (by approximating the change in loss on the target point if a training point is removed)
Record-based Explanation

- Release the most influential data points in the training set on the classification of a target data point (by approximating the change in loss on the target point if a training point is removed)
It Discloses Training Data!

Membership Inference: Query with a record and see if it appears as influential for itself

<table>
<thead>
<tr>
<th>% of data</th>
<th>$k = 1$</th>
<th>$k = 5$</th>
<th>$k = 10$</th>
</tr>
</thead>
</table>

Attack accuracy

Diabetics Hospital Dataset (medical test)
**It Discloses Training Data!**

Membership Inference: Query with a record and see if it appears as influential for itself

<table>
<thead>
<tr>
<th></th>
<th>% of data</th>
<th>$k = 1$</th>
<th>$k = 5$</th>
<th>$k = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Whole data set</strong></td>
<td>100%</td>
<td>34%</td>
<td>64%</td>
<td>77%</td>
</tr>
<tr>
<td><strong>Age 0 -10</strong></td>
<td>&lt;0.1%</td>
<td>67%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Age 0 -20</strong></td>
<td>&lt;1%</td>
<td>20%</td>
<td>58%</td>
<td>92%</td>
</tr>
<tr>
<td><strong>Caucasian</strong></td>
<td>74%</td>
<td>34%</td>
<td>64%</td>
<td>77%</td>
</tr>
<tr>
<td><strong>African American</strong></td>
<td>19%</td>
<td>38%</td>
<td>68%</td>
<td>81%</td>
</tr>
<tr>
<td><strong>Hispanics</strong></td>
<td>2%</td>
<td>39%</td>
<td>64%</td>
<td>76%</td>
</tr>
<tr>
<td><strong>Unknown race</strong></td>
<td>1%</td>
<td>35%</td>
<td>60%</td>
<td>77%</td>
</tr>
<tr>
<td><strong>Asian American</strong></td>
<td>&lt;1%</td>
<td>25%</td>
<td>64%</td>
<td>89%</td>
</tr>
</tbody>
</table>

Diabetics Hospital Dataset (medical test)
Reconstruction Attack
Reconstruction Attack

![Graph showing the comparison between Shadow model and Sampling over the number of queries. The graph indicates that the Shadow model generally performs better than Sampling as the number of queries increases.](image-url)
Interpretability vs. Privacy

• Complex decisions (on minorities)
  ✴ When we need to explain
  ✴ Explanation leaks information
Privacy Challenges for Generalizability, Robustness, Interpretability

- Model Capacity
- Data Distribution (minorities)
- Task Complexity

- ML Privacy by design (cannot be easily “added” to existing robustness/explaining algorithms, without significant overhead)
We are hiring (PhD, Postdoc)!