Dataset Inference: Ownership Resolution in Machine Learning

Pratyush Maini [1,2,3] Mohammad Yaghini [2,3] Nicolas Papernot [2,3]

1 IIT Delhi
Indian Institute of Technology Delhi

2 UNIVERSITY OF TORONTO

3 VECTOR INSTITUTE

Overview

- Why is model privacy important?
  - Primer on Model Extraction and Membership Inference
- Model Stealing – Threat Models
- Dataset Inference
  - Train-Test Prediction Margin
  - Blind Walk
  - Confidence Regressor
  - Ownership Resolution
  - Results

Developing High-performing ML models is expensive

Computational Cost

Private Data

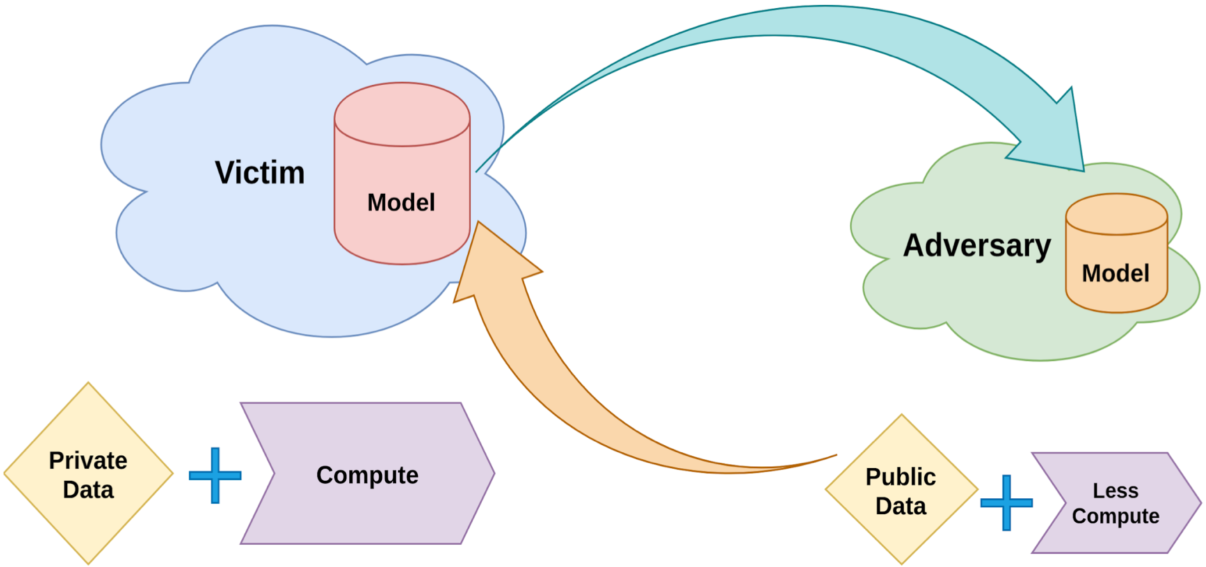
Intellectual Contribution

Model Stealing Attacks are a realistic threat

Copying a model's predictions with significantly lesser cost at the adversary's end.

Model Extraction

- Using predictions from an ML API (victim) to train a surrogate model using some publicly available dataset.



Membership Inference

- Inferring the membership of a data-point in a model's training set.

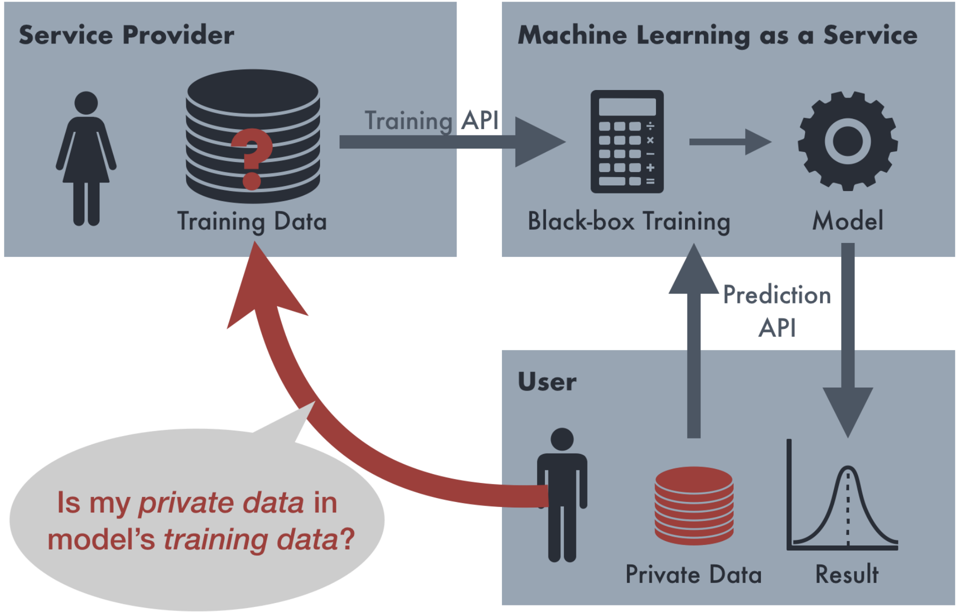
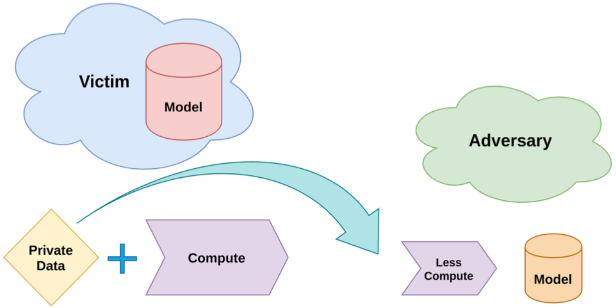


Image credits [Hisamoto et. al. 2020]

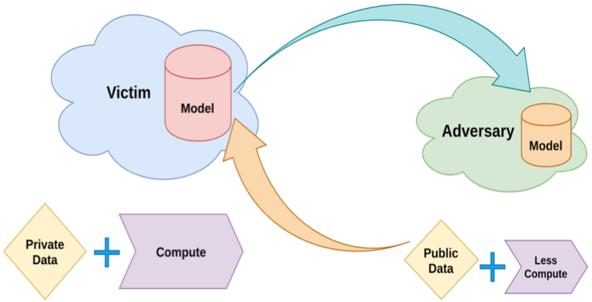
Model Stealing Attacks: How?

An adversary may gain varying degrees of access to your '*Knowledge*'



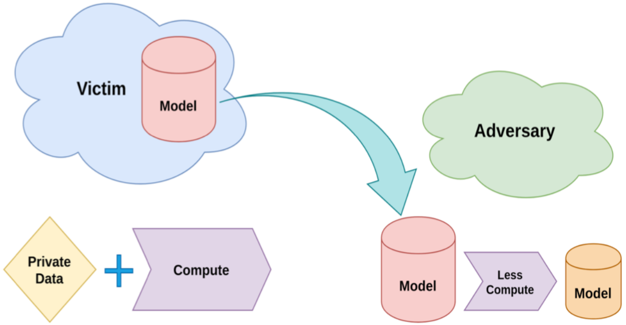
Data Access: A_D

-- Distillation
-- Train on different architectures
or hyperparameters



Query Access: A_Q

-- Label Access
-- Logit Access

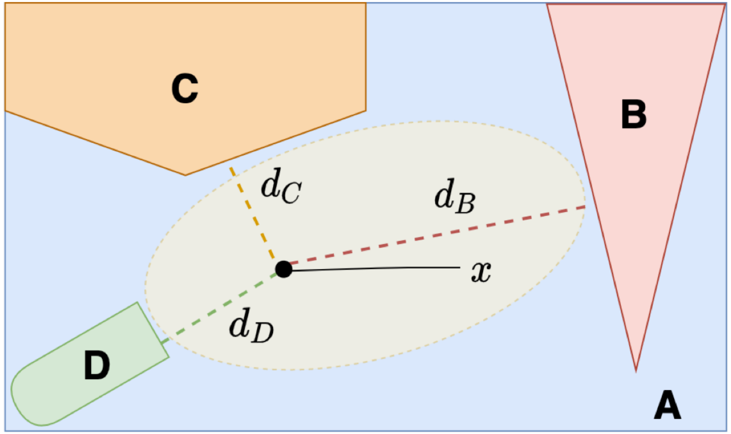


Model Access: A_M

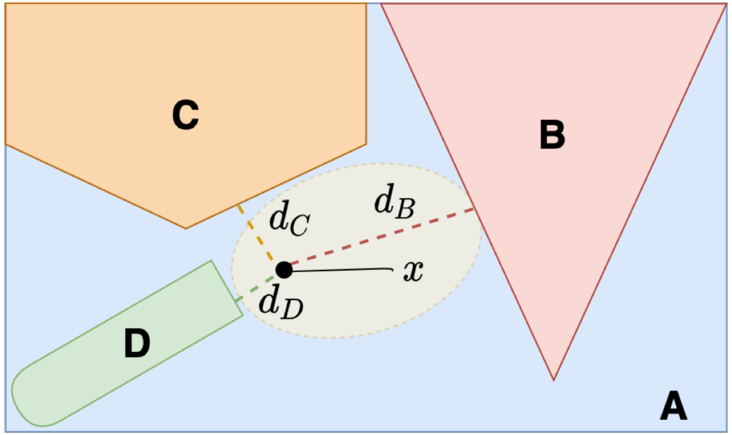
-- Zero shot learning
-- Fine tuning

Dataset Inference Exploits Train-Test Prediction Certainty

Prediction Margin if x was in Train set



Prediction Margin if x was in Test set



Analysis on a Linear Model

Binary Classification

$$y \sim \{-1, +1\};$$

Linearly Separable

$$\mathbf{x_1} = y \cdot \mathbf{u} \in \mathbb{R}^K,$$

Gaussian Noise

$$\mathbf{x_2} \sim \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^D$$

$$h(\mathbf{X}) = \mathbf{w_1} \cdot \mathbf{x_1} + \mathbf{w_2} \cdot \mathbf{x_2}$$

Linear Classifier

Theorem 1 (Train-Test Margin) *Given a linear classifier $h(.)$ trained to classify inputs $(\mathbf{X}, y) \in \mathcal{S} \subset \mathcal{D} \subset \mathbb{R}^{K+D}$, the difference in the expected prediction margin for $\mathbf{X}$ in $\mathcal{S}$ and $\mathcal{D} \setminus \mathcal{S}$, given by* $\mathbb{E}_{X \sim \mathcal{S}}[y \cdot h(\mathbf{X})] - \mathbb{E}_{X \sim \mathcal{D} \setminus \mathcal{S}}[y \cdot h(\mathbf{X})] = D\sigma^2$.

$$\mathbf{w_1} \leftarrow \mathbf{w_1} + \alpha y^{(i)} \mathbf{x_1}^{(i)}$$

$$\mathbf{w_2} \leftarrow \mathbf{w_2} + \alpha y^{(i)} \mathbf{x_2}^{(i)}$$

Dataset Inference Succeeds when Membership Inference Fails

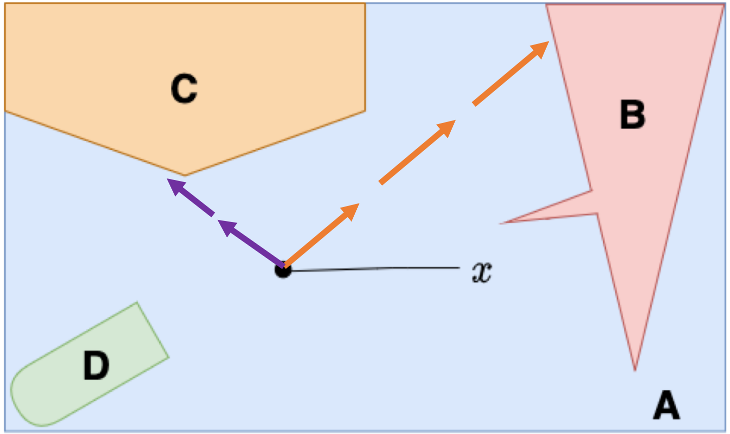
Theorem 2 (Failure of MI) Given a linear classifier $h(.)$ trained on $\mathcal{S} \subset \mathcal{D} \subset \mathbb{R}^{D+K}$, the probability that an adversary $\mathcal{M}$ correctly predicts the membership of inputs randomly belonging to the training or test set, $\mathbb{P}_{X\sim\mathcal{R}}\left[\mathcal{M}(\mathbf{X}, h(.)) = b\right] = 1 - \Phi\left(-\sqrt{\frac{D}{2m}}\right)$, and decreases with $|\mathcal{S}| = m$. Moreover, $\lim_{m\to\infty} \mathbb{P}_{X\sim\mathcal{R}}\left[\mathcal{M}(\mathbf{X}, h(.)) = b\right] = 0.5$.

Theorem 3 (Success of DI) Choose $b \leftarrow \{0, 1\}$ uniformly at random. Given an adversary's linear classifier $h(.)$ trained on $\mathcal{D}\backslash\mathcal{S} \subset \mathbb{R}^{K+D}$, if b = 0, and on $\mathcal{S} \subset \mathcal{D}$ otherwise. The probability $\mathcal{V}$ correctly decides if an adversary stole its knowledge $\mathbb{P}\left[\psi(\mathcal{D}, h(.)) = b\right] = 1 - \Phi\left(-\frac{\sqrt{D}}{2}\right)$. Moreover, $\lim_{D\to\infty} \mathbb{P}\left[\psi(\mathcal{D}, h(.)) = b\right] = 1$.

How do you calculate the prediction certainty?

Blind Walk : Black-box method to estimate the prediction certainty

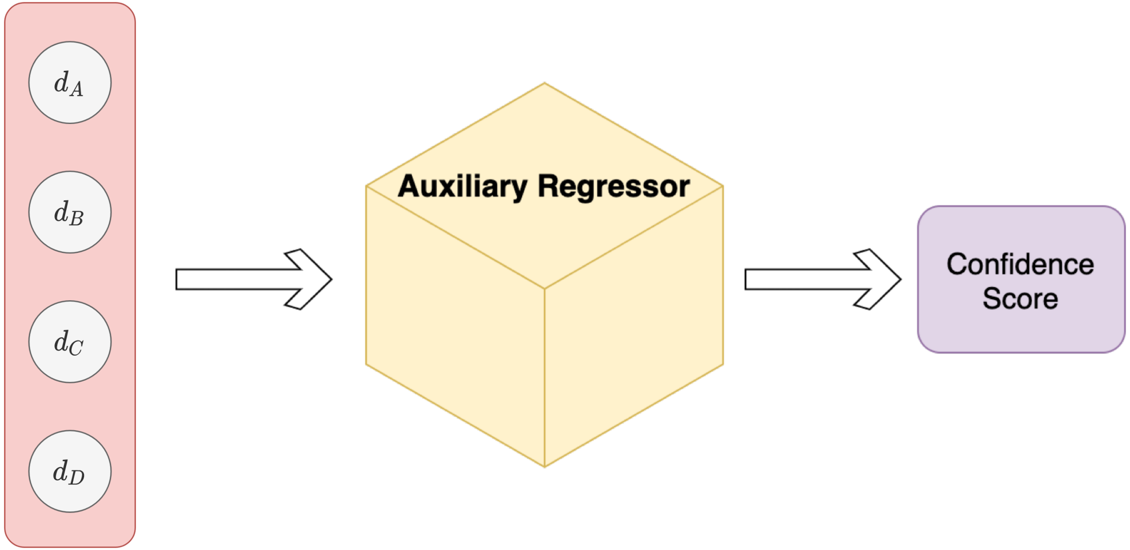
a. *Sample Random Noise*
b. *Take Small Steps in that direction till you reach class boundary*
c. *Aggregate the distance over many noise directions to create a feature embedding.*



$$\mathrm{emb}^{i}_{(\mathbf{x},y)}(f) = \min_{k \in \mathbb{N}} d(x, x + k\delta_i);$$

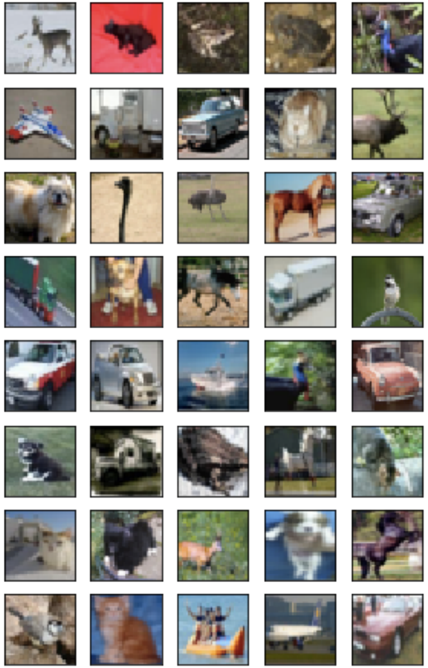
$$s.t. f(x + k\delta_i) = t; \;\; t \neq y$$

Training an Auxiliary Regressor



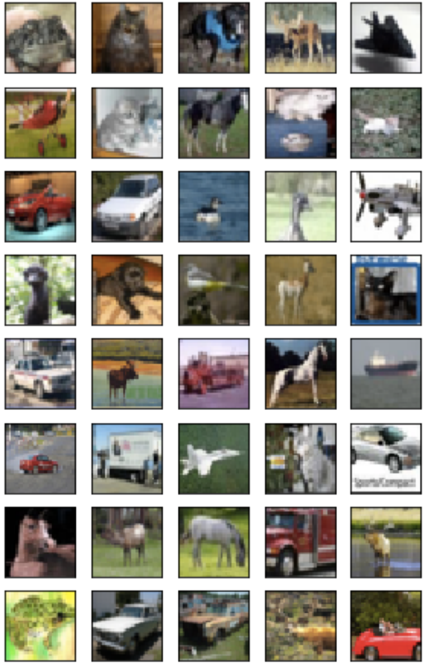
Ownership Resolution by aggregation of Confidence Scores

Training Set



Step 1:
Sample inputs from the train & test set

Test Set

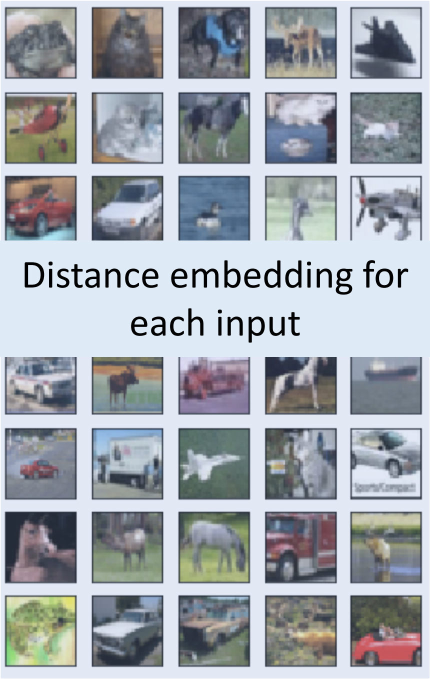


Ownership Resolution by aggregation of Confidence Scores

Training Set

Step 2:
Generate embeddings for prediction margin

Test Set



Ownership Resolution by aggregation of Confidence Scores

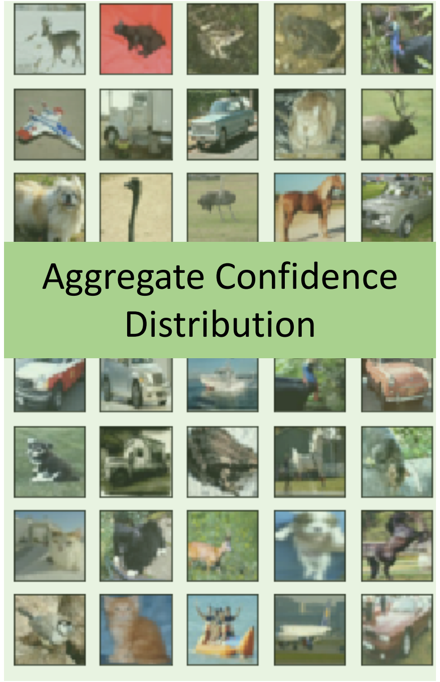
Training Set

Step 3:
Pass embeddings through auxiliary regressor

Test Set

Ownership Resolution by aggregation of Confidence Scores

Training Set

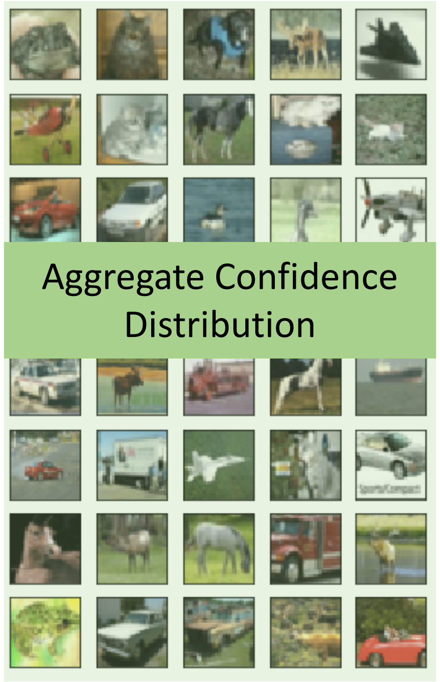


Step 4: One sided t-Test:

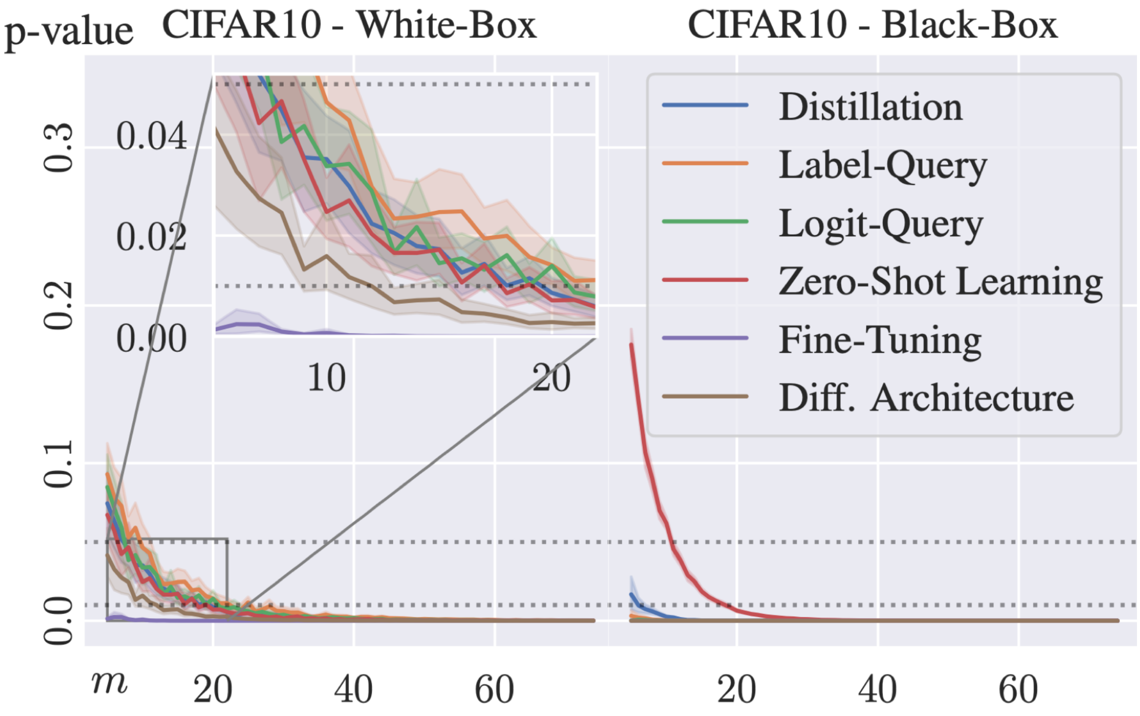
H_0: $\mu_{test} \geq \mu_{train}$

If stolen, H_0 would be rejected.

Test Set



DI is successful across CIFAR10, SVHN, CIFAR100 and ImageNet



p-value against number of revealed samples (m)

Dataset Inference resolves ownership by revealing fewer than 60 private samples, with FPR < 1%

Key Take-aways from Dataset Inference (DI)

1. Requires few private points to prove ownership.
2. Can be performed in less than 30,000 queries to the adversary.
3. White-box access is not essential to DI
4. Out-of-the-box solution that does not require overfitting or retraining.
5. Does not have a trade-off with task accuracy.