

The background of the slide is a photograph of the Iowa State University campus, featuring the Old Capitol building with its prominent dome on the left and other university buildings in the distance. The entire image is covered with a semi-transparent red overlay. Two thin, horizontal gold lines are positioned above and below the text.

# IOWA STATE UNIVERSITY

**Department of Computer Science**

# Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning

Milad Nasr, Reza Shokri, Amir Houmansadr

*Yu-Pin Liang (Jasmine)*

# INTRODUCTION – Risk in Machine Learning

---

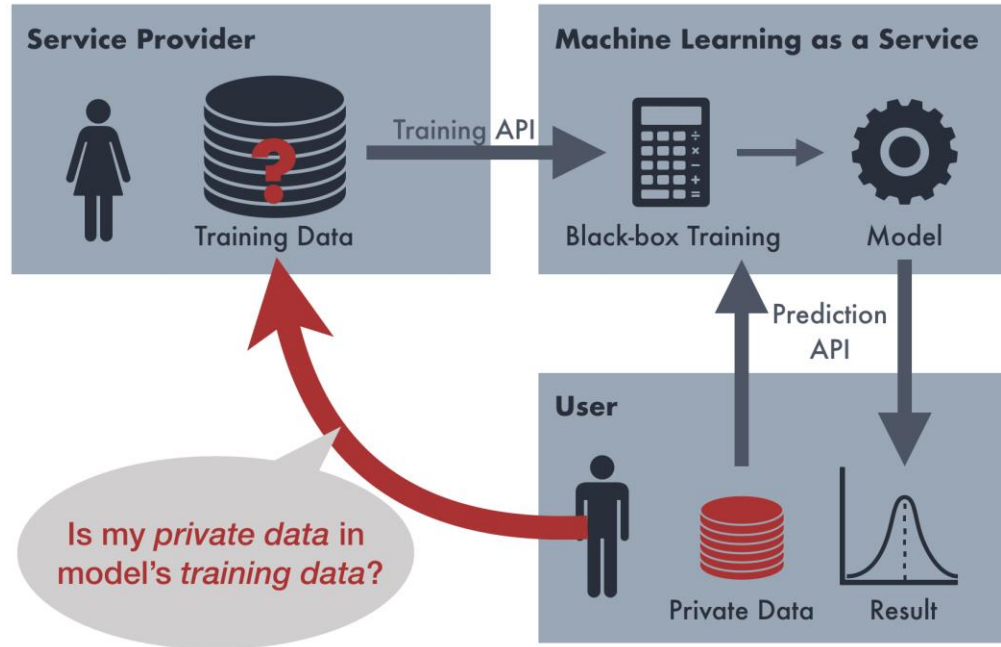
- Deep neural networks have shown unprecedented generalization for various learning tasks, from image and speech recognition to generating realistic-looking data.
- Wide range of **sensitive** data may leak due to the training process.
- What is the privacy risk of deep learning algorithms to individuals whose data is used for training deep neural networks? In other words, how much is the information leakage of deep learning algorithms about their individual training data samples?

# INTRODUCTION – Membership Inference Attack

---

- An adversary algorithm falls in to two types
  - -> tracing (a.k.a. membership inference) attacks
    - The attacker's objective is to infer if a particular individual data record was included in the training dataset. This is a **decisional** problem, and its accuracy directly demonstrates the leakage of the model.
  - -> reconstruction attacks
    - The attacker's objective is to infer attributes of the records in the training set.
- Membership inference attacks pose a threat to the privacy of individual data points in a dataset used for machine learning.

# INTRODUCTION – Membership Inference Attack





# Purpose

In this paper, authors present a comprehensive framework for the privacy analysis of deep neural networks, using white-box membership inference attacks.

## Contribution

---

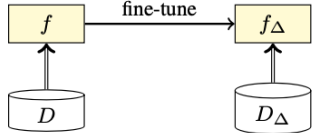
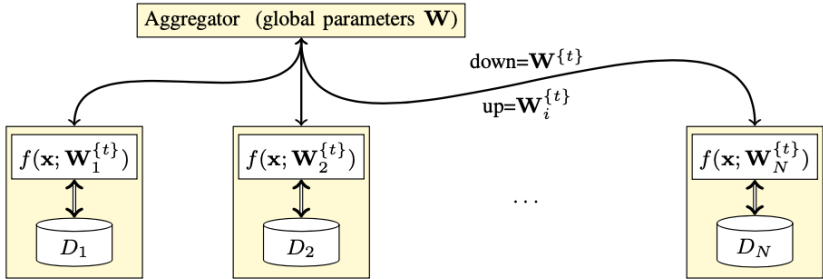
- Authors design **white-box inference attacks** that exploit the privacy vulnerabilities of the **stochastic gradient descent (SGD)** algorithm.
- Each data point in the training set influences many of the model parameters, through the SGD algorithm, to minimize its contribution to the model's training loss.

# Model - Black-box vs White-box

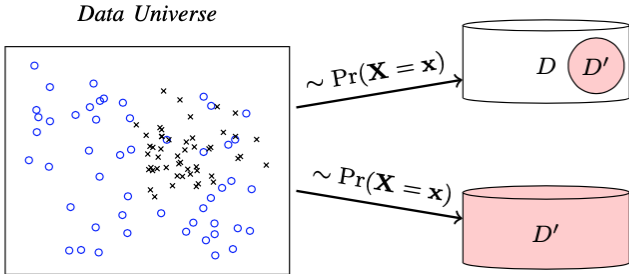
Observation	Black-box	<p>The attacker can obtain the prediction vector <math>f(\mathbf{x})</math> on arbitrary input <math>\mathbf{x}</math>, but cannot access the model parameters, nor the intermediate computations of <math>f(\mathbf{x})</math>.</p> <p><math>\mathbf{x} \longrightarrow \boxed{f} \longrightarrow f(\mathbf{x})</math></p>
	White-box	<p>The attacker has access to the full model <math>f(\mathbf{x}; \mathbf{W})</math>, notably its architecture and parameters <math>\mathbf{W}</math>, and any hyper-parameter that is needed to use the model for predictions. Thus, he can also observe the intermediate computations at hidden layers <math>h_i(\mathbf{x})</math>.</p> <p><math>\mathbf{x} \longrightarrow \boxed{\mathbf{W}_1} \quad h_1(\mathbf{x}) \rightarrow \boxed{\mathbf{W}_2} \quad h_2(\mathbf{x}) \rightarrow \cdots \rightarrow \boxed{\mathbf{W}_i} \longrightarrow f(\mathbf{x})</math></p>



# Model - Targets

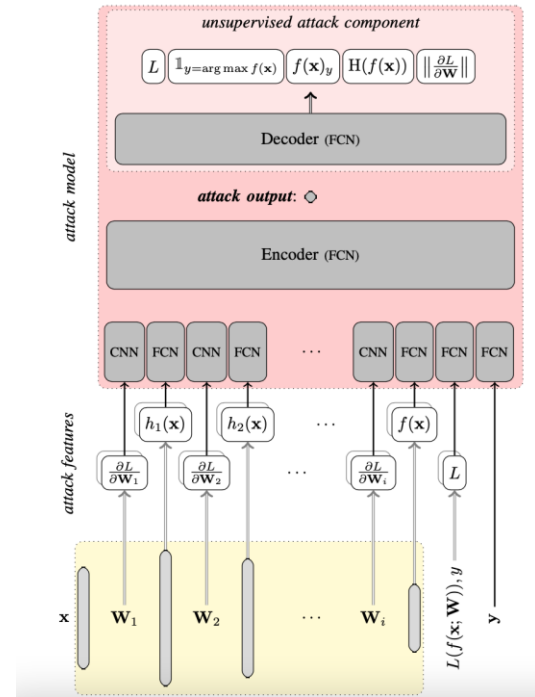
Target	Stand-alone	<p>The attacker observes the final target model <math>f</math>, after the training is done (e.g., in a centralized manner) using dataset <math>D</math>. He might also observe the updated model <math>f_{\Delta}</math> after it has been updated (fine-tuned) using a new dataset <math>D_{\Delta}</math>.</p> 
	Federated	<p>The attacker could be the central aggregator, who observes individual updates over time and can control the view of the participants on the global parameters. He could also be any of the participants who can observe the global parameter updates, and can control his parameter uploads.</p> 

# Model - Mode and Knowledge

Mode	Passive	The attacker can only observe the genuine computations by the training algorithm and the model.
	Active	The attacker could be one of the participants in the federated learning, who adversarially modifies his parameter uploads $\mathbf{W}_i^{\{t\}}$ , or could be the central aggregator who adversarially modifies the aggregate parameters $\mathbf{W}^{\{t\}}$ which he sends to the target participant(s).
Knowledge	Supervised	<p>The attacker has a data set <math>D'</math>, which contains a subset of the target set <math>D</math>, as well as some data points from the same underlying distribution as <math>D</math> that are not in <math>D</math>. The attacker trains an inference model <math>h</math> in a supervised manner, by minimizing the empirical loss function <math>\sum_{d \in D'} (1 - \mathbb{1}_{d \in D})h(d) + \mathbb{1}_{d \in D}(1 - h(d))</math>, where the inference model <math>h</math> computes the membership probability of any data point <math>d</math> in the training set of a given target model <math>f</math>, i.e., <math>h(d) = \Pr(d \in D; f)</math>.</p> 
	Unsupervised	The attacker has data points that are sampled from the same underlying distribution as $D$ . However, he does not have information about whether a data sample has been in the target set $D$ .

# Model - Unsupervised attack

- The architecture of white-box inference attack.
- Given target data  $(x, y)$ , the objective of the attack is to determine its membership in the training set  $D$  of target model  $f$ .



# Result

TABLE VI: Attack accuracy for various sizes of the attacker’s training dataset. The size of the target model’s training dataset is 50,000. (The CIFAR100 dataset with Alexnet, stand-alone scenario)

Members Sizes	Non-members Sizes	Attack Accuracy
10,000	2,000	73.2%
15,000	2,000	73.7%
15,000	5,000	74.8%
25,000	5,000	75.1%

TABLE VII: Accuracy of our unsupervised attack compared to the Shadow models approach [6] for the white-box scenario.

Dataset	Arch	(Unsupervised) Attack Accuracy	(Shadow Models) Attack Accuracy
CIFAR100	Alexnet	75.0%	70.5%
CIFAR100	DenseNet	71.2%	64.2%
CIFAR100	ResNet	63.1%	60.9%
Texas100	Fully Connected	66.3%	65.3%
Purchase100	Fully Connected	71.0%	68.2%

TABLE VIII: The attack accuracy for different datasets and different target architectures using layer outputs versus gradients. This is the result of analyzing the stand-alone scenario, where the CIFAR100 models are all obtained from pre-trained online repositories.

Target Model				Attack Accuracy		
Dataset	Architecture	Train Accuracy	Test Accuracy	Black-box	White-box (Outputs)	White-box (Gradients)
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%
Texas100	Fully Connected	81.6%	52%	63.0%	63.3%	68.3%
Purchase100	Fully Connected	100%	80%	67.6%	67.6%	73.4%

TABLE IX: Attack accuracy on fine-tuned models.  $D$  is the initial training set,  $D_{\Delta}$  is the new dataset used for fine-tuning, and  $\bar{D}$  is the set of non-members (which is disjoint with  $D$  and  $D_{\Delta}$ ).

Dataset	Architecture	Train Acc.	Test Acc.	Distinguish $D$ from $D_{\Delta}$	Distinguish $D$ from $\bar{D}$	Distinguish $D_{\Delta}$ from $\bar{D}$
CIFAR100	Alexnet	100.0%	39.8%	62.1%	75.4%	71.3%
CIFAR100	DenseNet	100.0%	64.3%	63.3%	74.6%	71.5%
Texas100	Fully Connected	95.2%	48.6%	58.4%	68.4%	67.2%
Purchase100	Fully Connected	100.0%	77.5%	64.4%	73.8%	71.2%

## Conclusion

---

- Author designed and evaluated novel white-box membership inference attacks against neural network models by exploiting the privacy vulnerabilities of the stochastic gradient descent algorithm.