CS334 Machine Learning

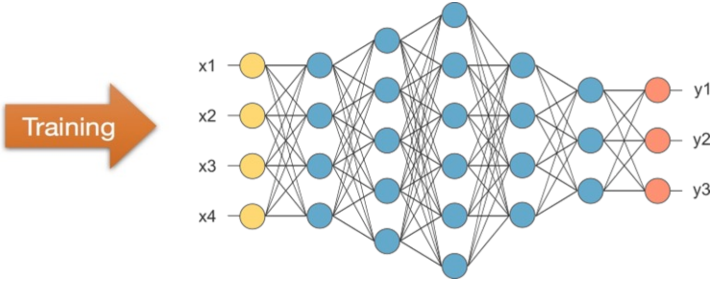
Emerging Topics: Privacy-Enhanced and Robust Machine Learning

Li Xiong
Emory University

EMORY
UNIVERSITY

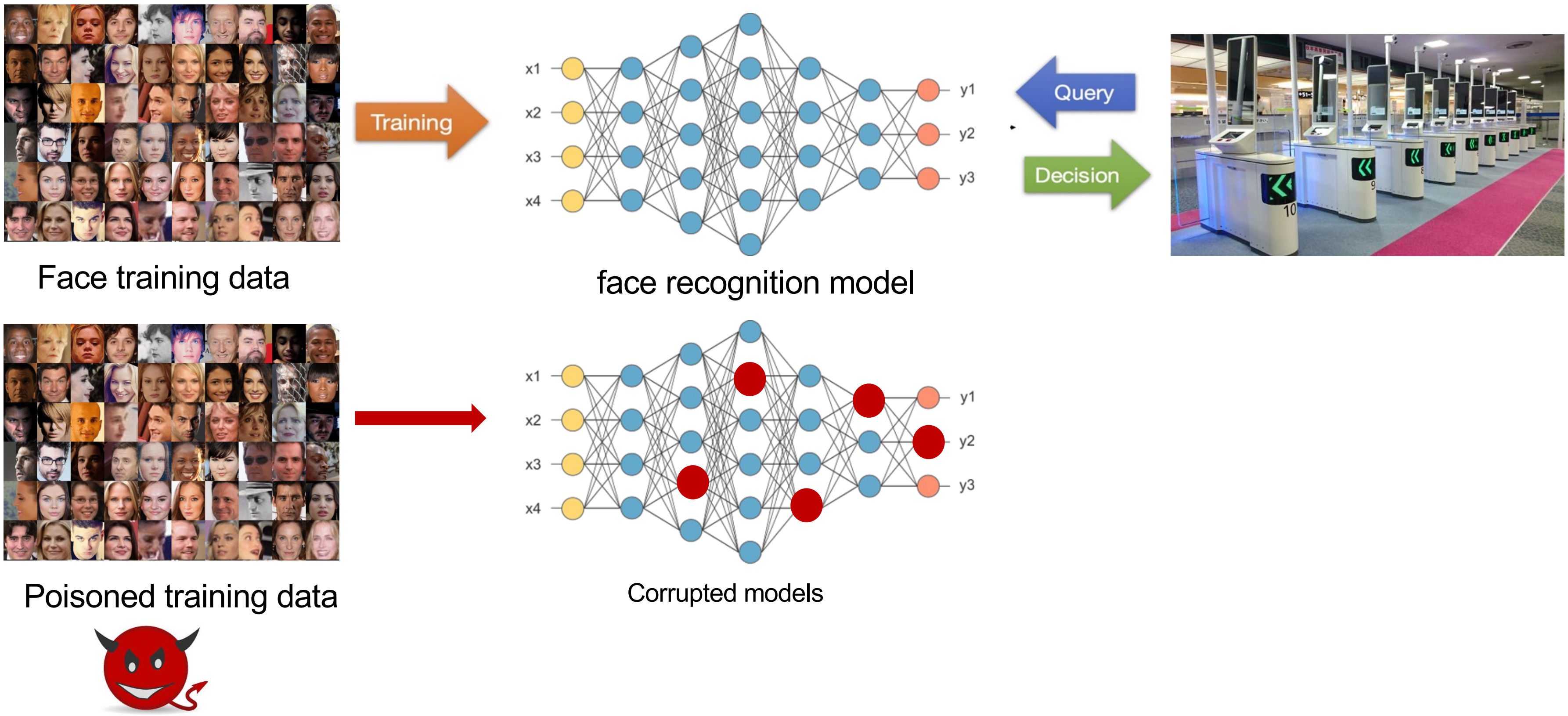
Machine Learning Pipeline

Face training data

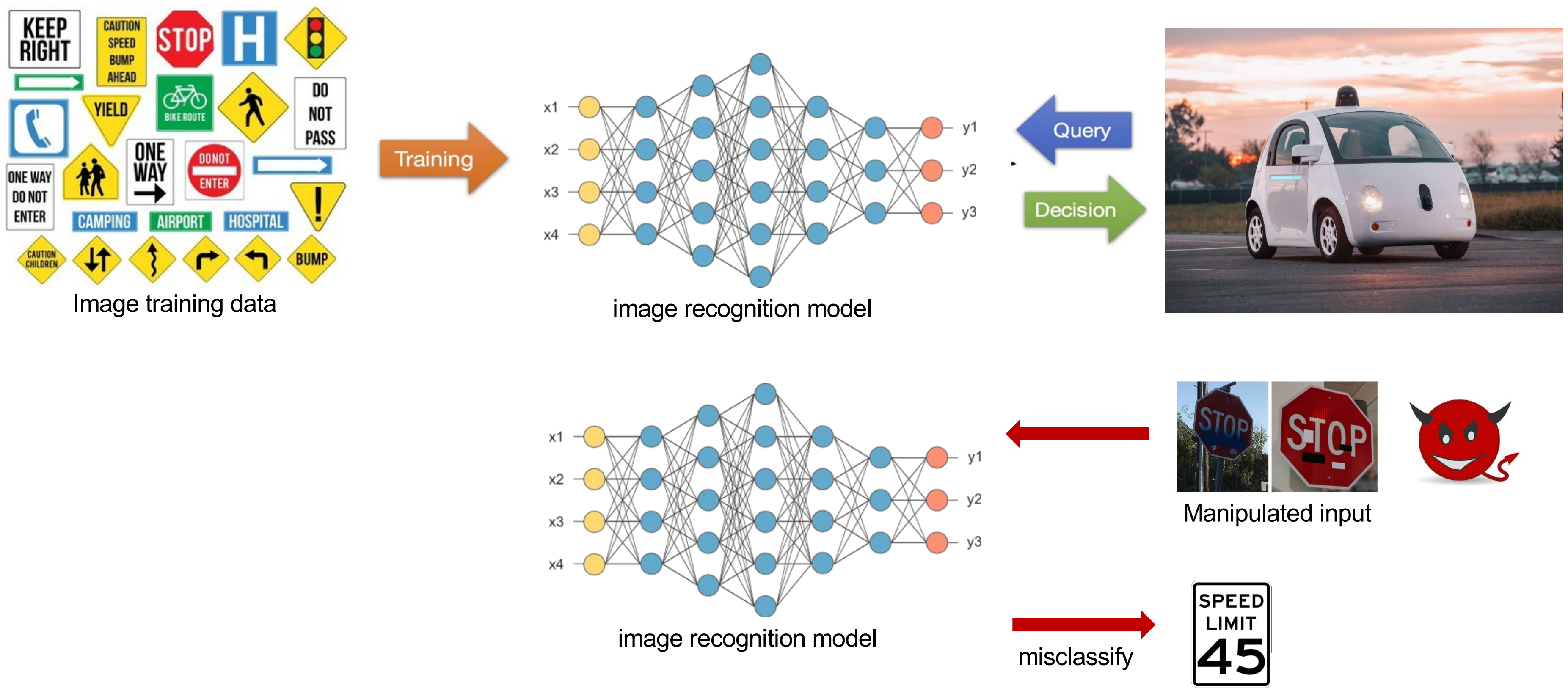


face recognition model

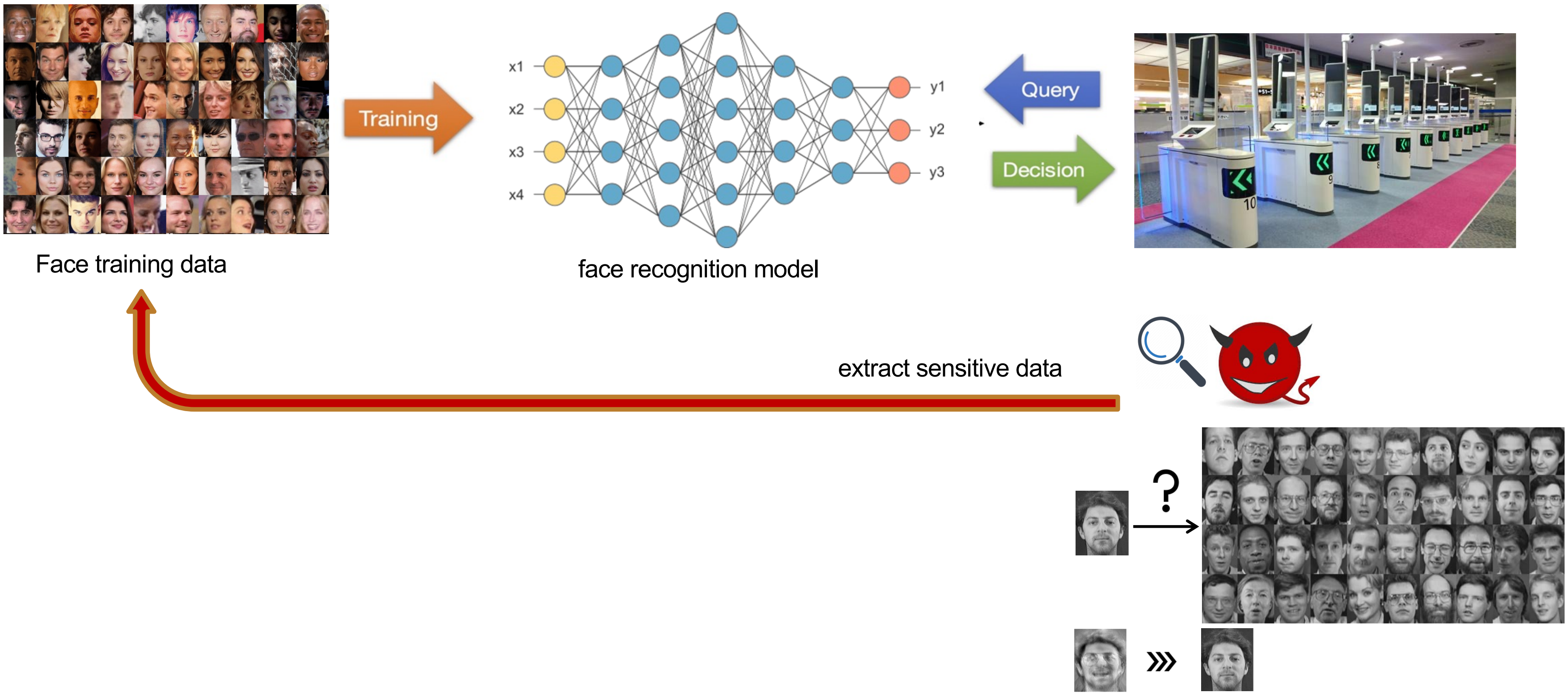
Data Poisoning Attacks (Training Stage)



Adversarial Example Attacks (Inference Stage)



Privacy Attacks (Inference Stage)



Outline

- Privacy attacks
  - Membership inference attacks, model inversion attacks, secret sharer
- Privacy-preserving deep learning
  - Differential privacy, gradient perturbation, noisy ensemble, federated learning
- Security attacks
  - Adversarial example attacks, poisoning attacks, backdoor attacks
- Robust deep learning
  - Detection and reform, adversarial training, certified robustness

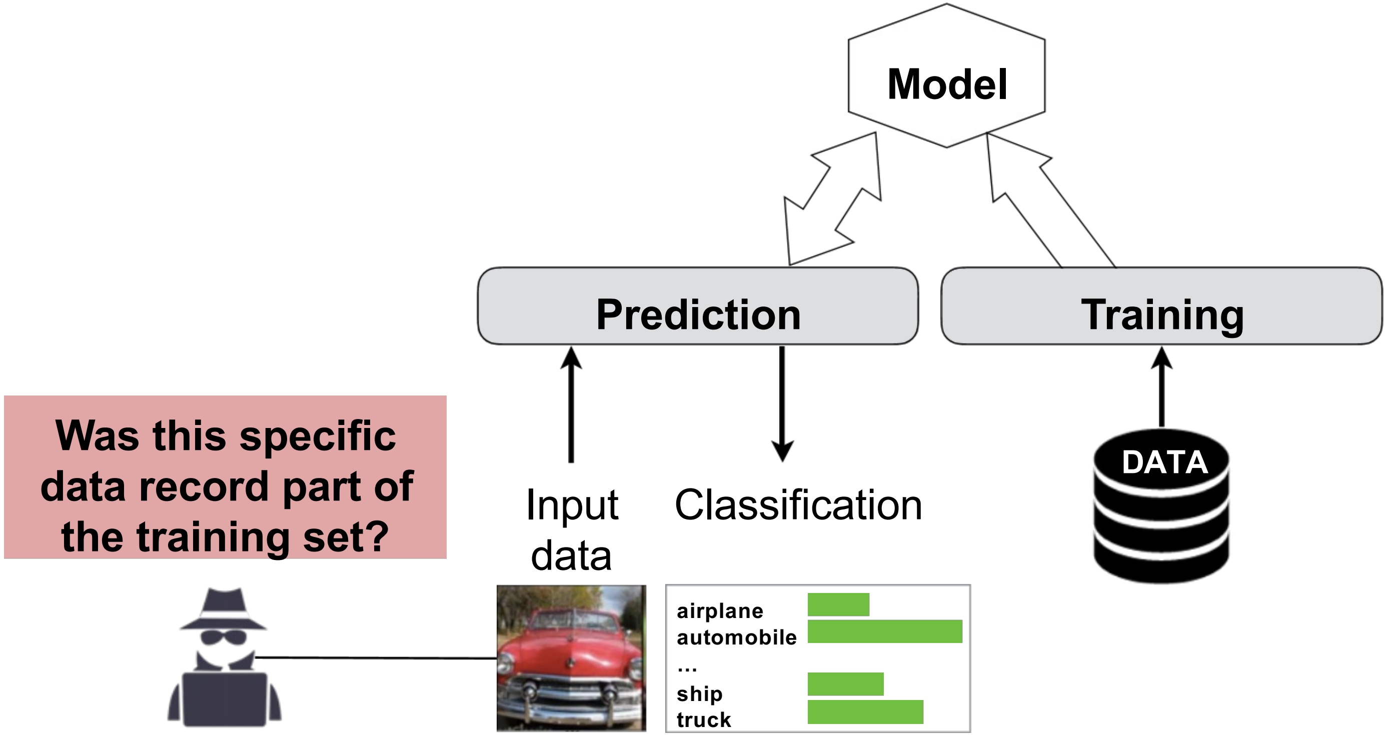
Membership Inference Attacks against Machine Learning Models

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov

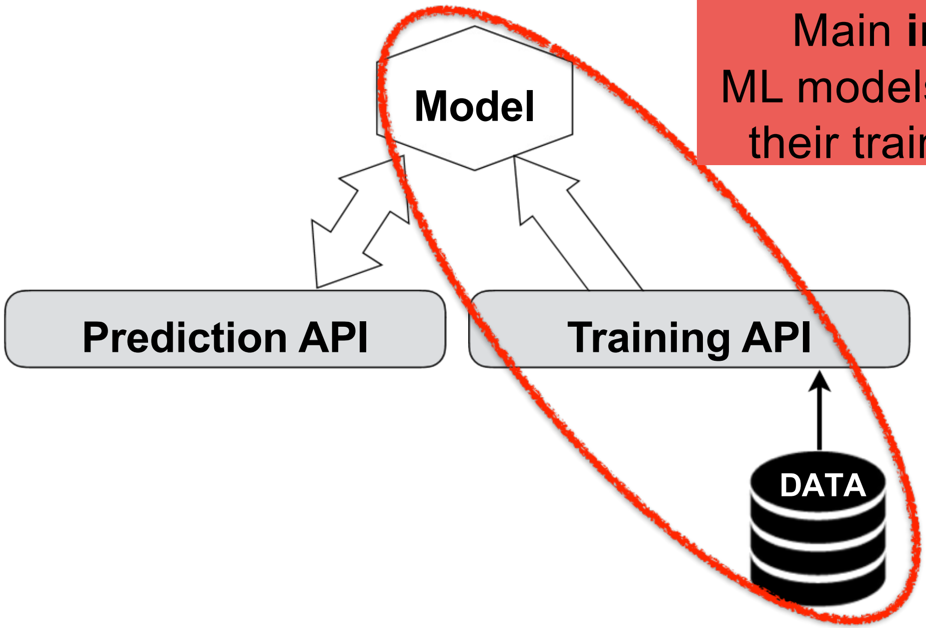
CORNELL UNIVERSITY FOUNDED A.D. 1865
CORNELL TECH

Membership Inference Attacks Against Machine Learning Models, S&P, 2017

Membership Inference Attack

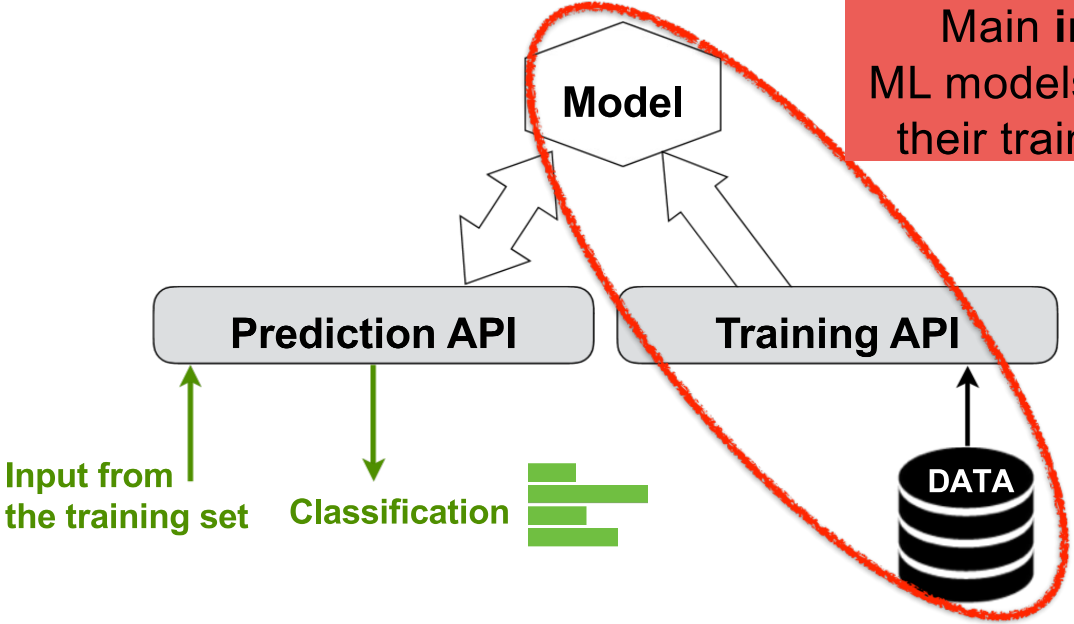


Exploit Model's Predictions



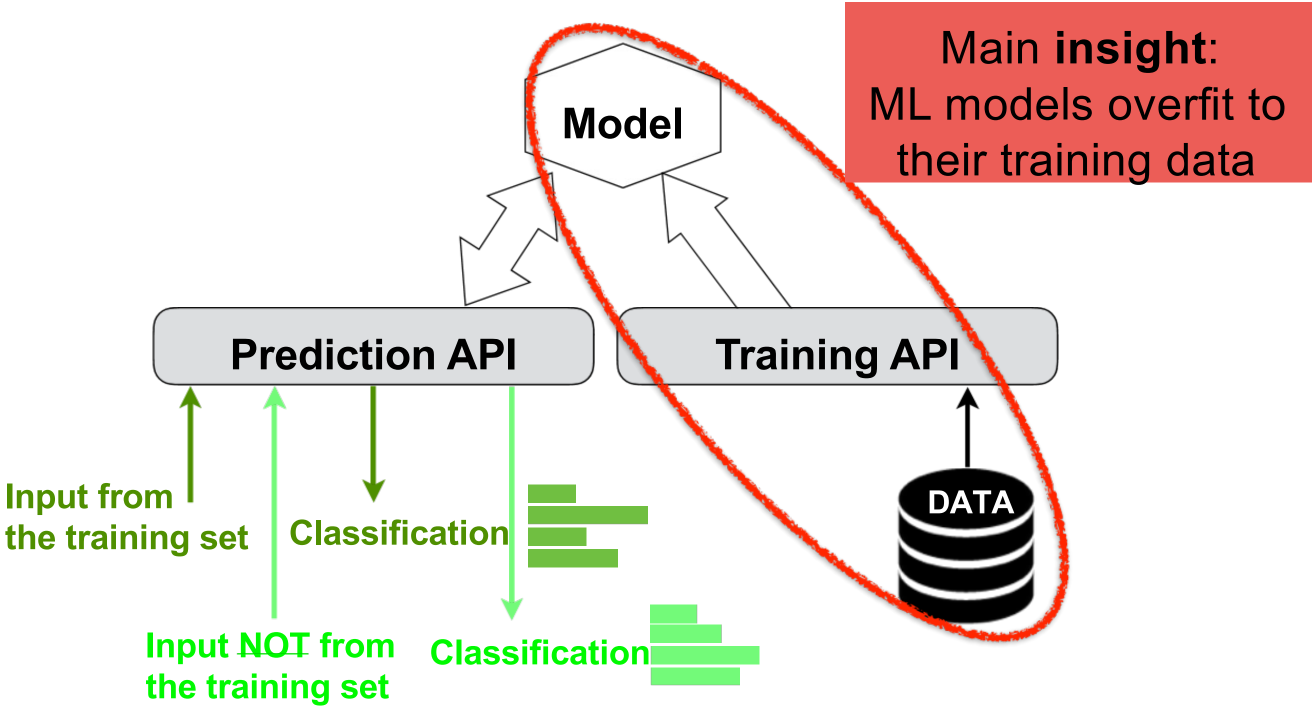
Main **insight**:
ML models overfit to their training data

Exploit Model's Predictions

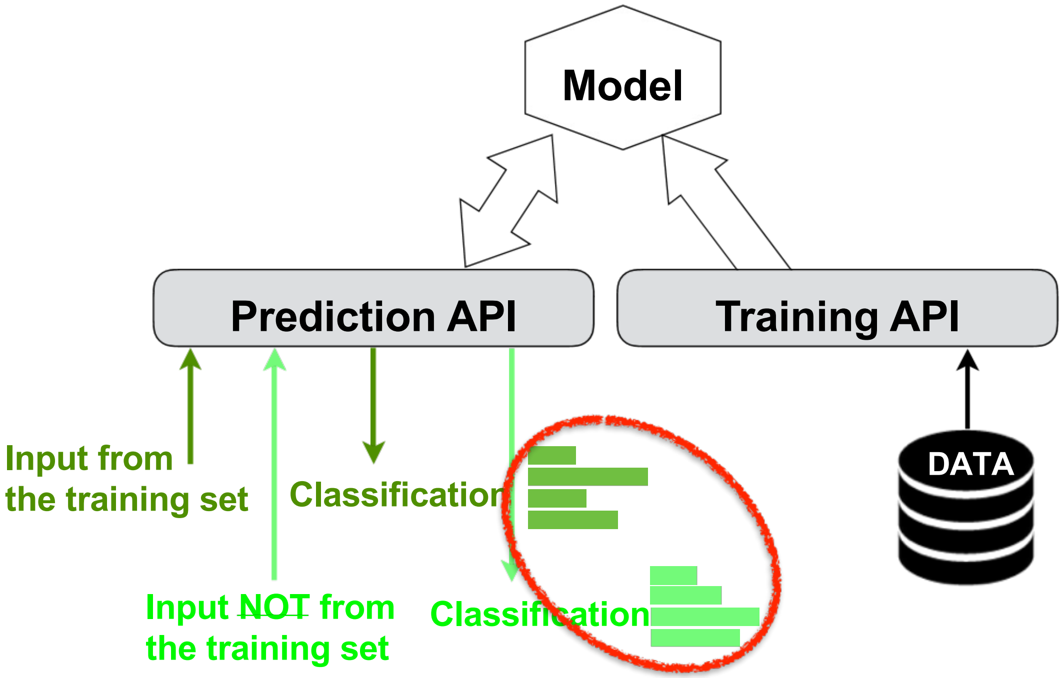


Main **insight**:
ML models overfit to their training data

Exploit Model's Predictions

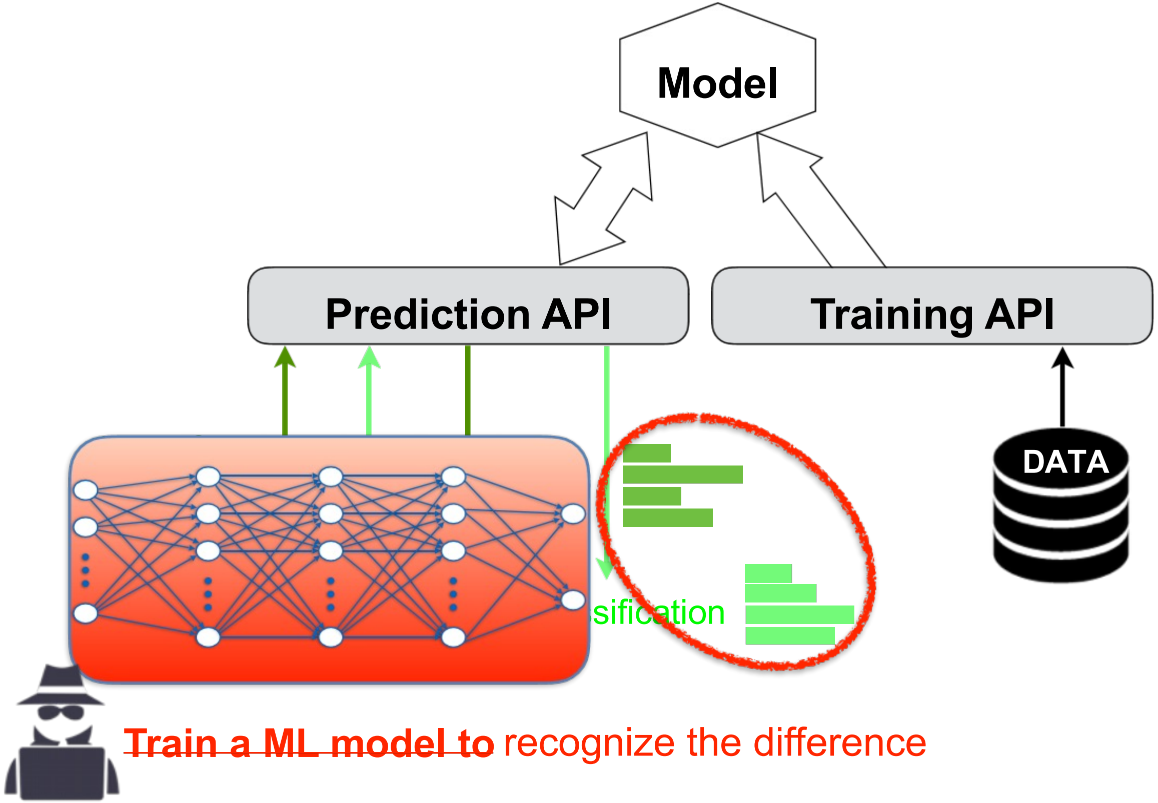


Exploit Model's Predictions



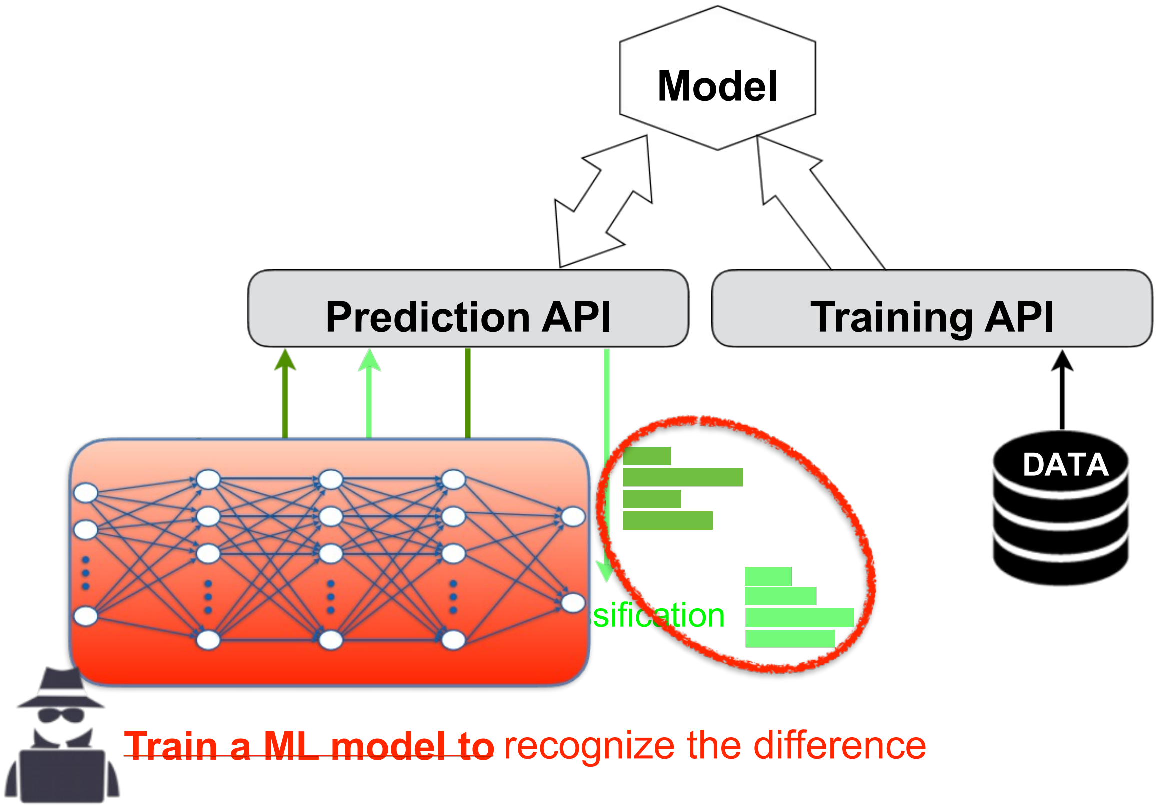
Recognize the difference

ML against ML



~~Train a ML model to~~ recognize the difference

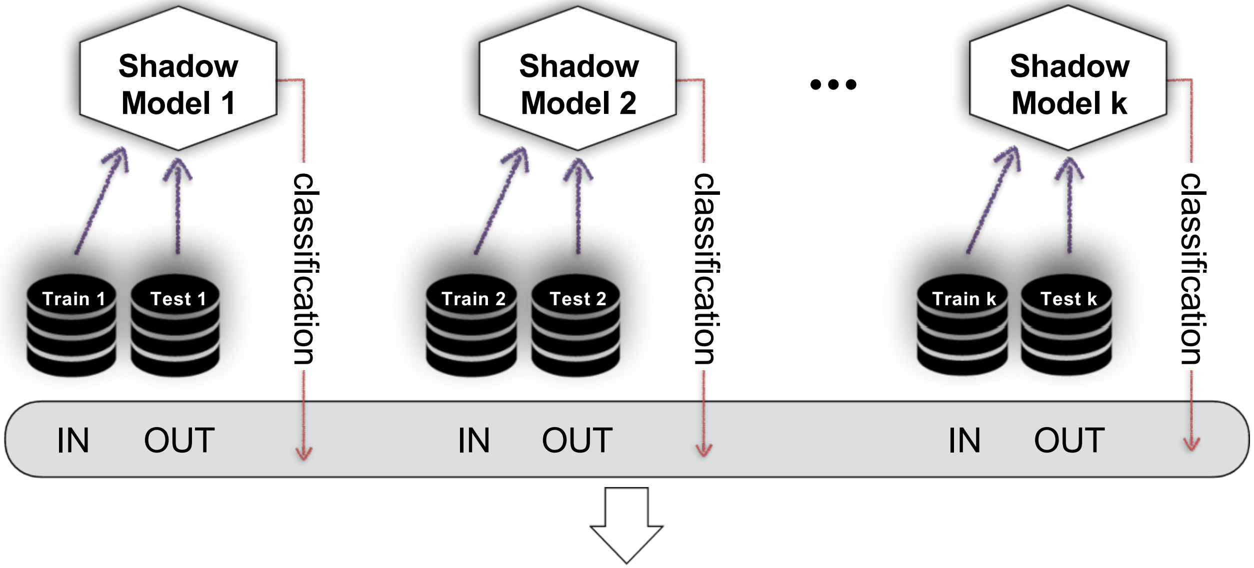
ML against ML



~~Train a ML model to~~ recognize the difference

What kind of training data is needed for training the attack model?

Train Attack Model using **Shadow Models**

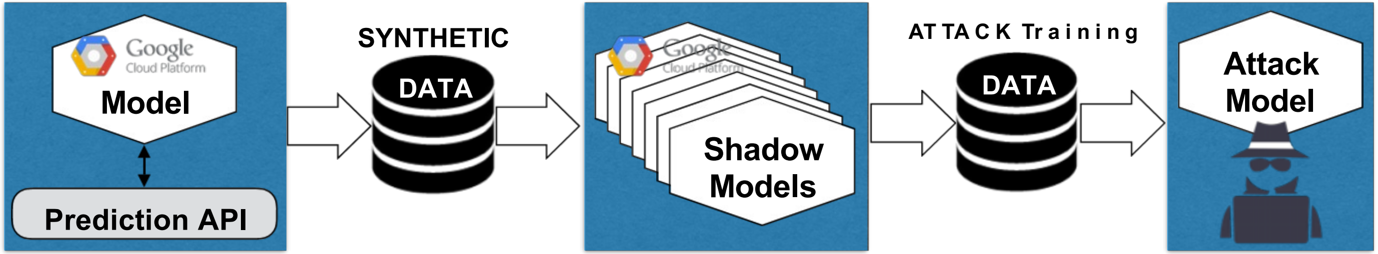


Train the attack model

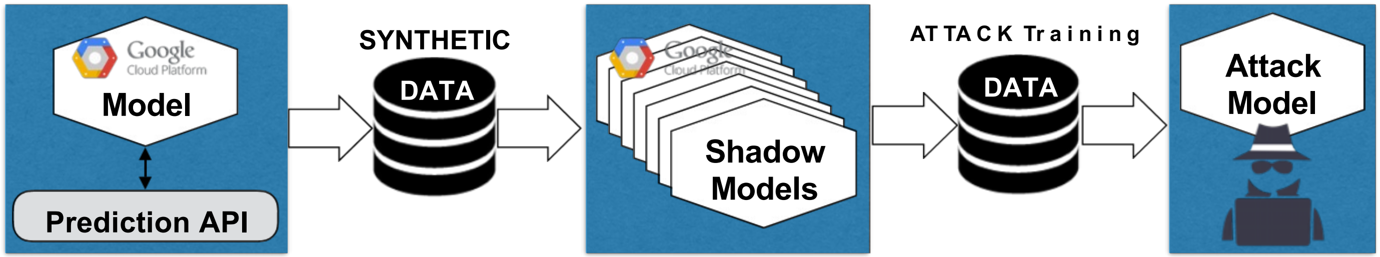
to predict if an input was a member of the training set (in) or a non-member (out)

How to get the training data?

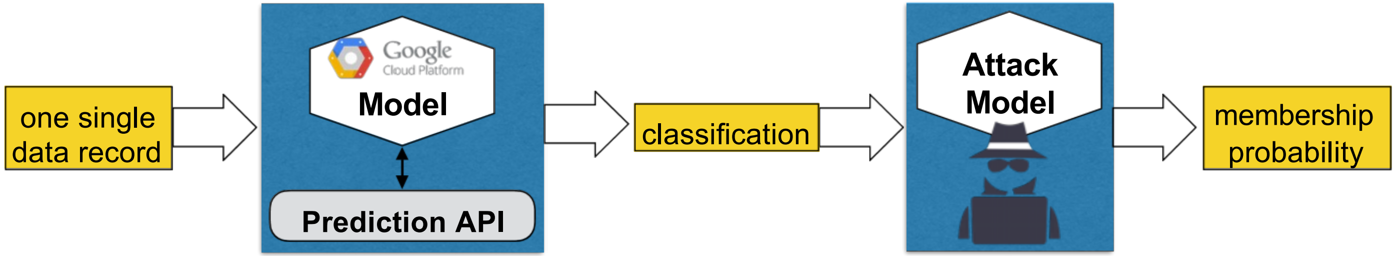
Constructing the Attack Model

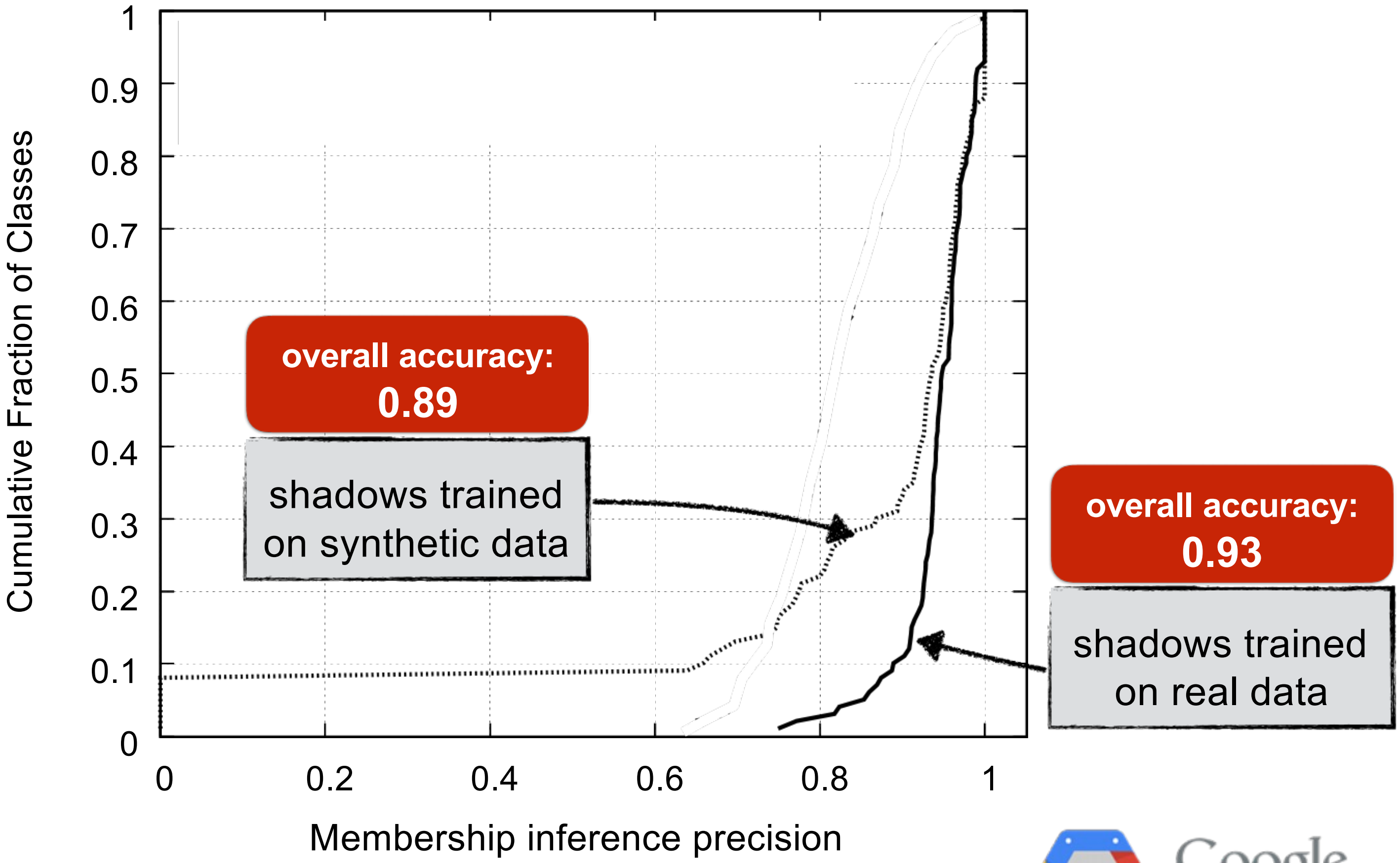


Constructing the Attack Model



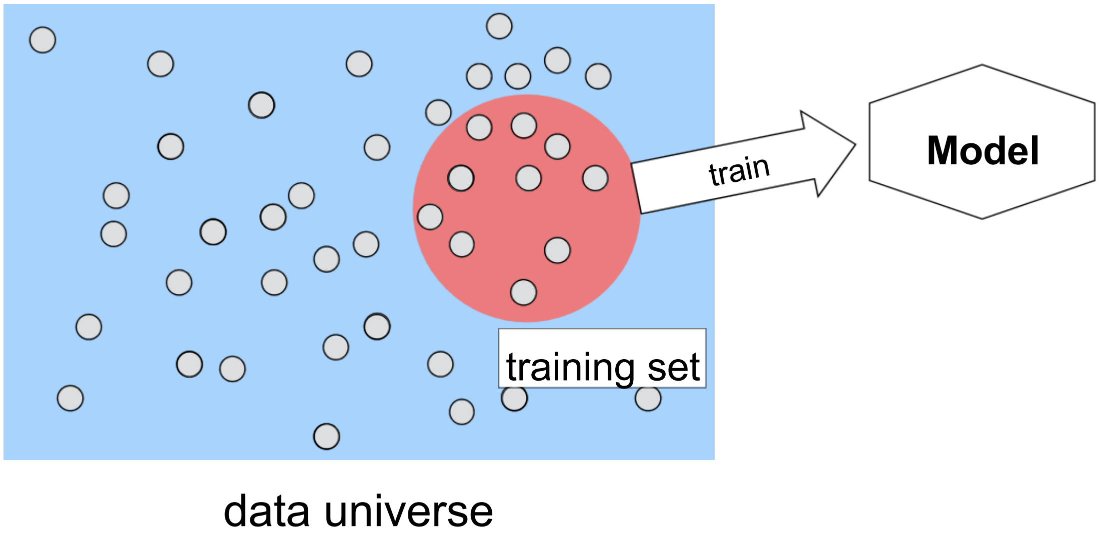
Using the Attack Model





Purchase Dataset — Classify Customers (100 classes)

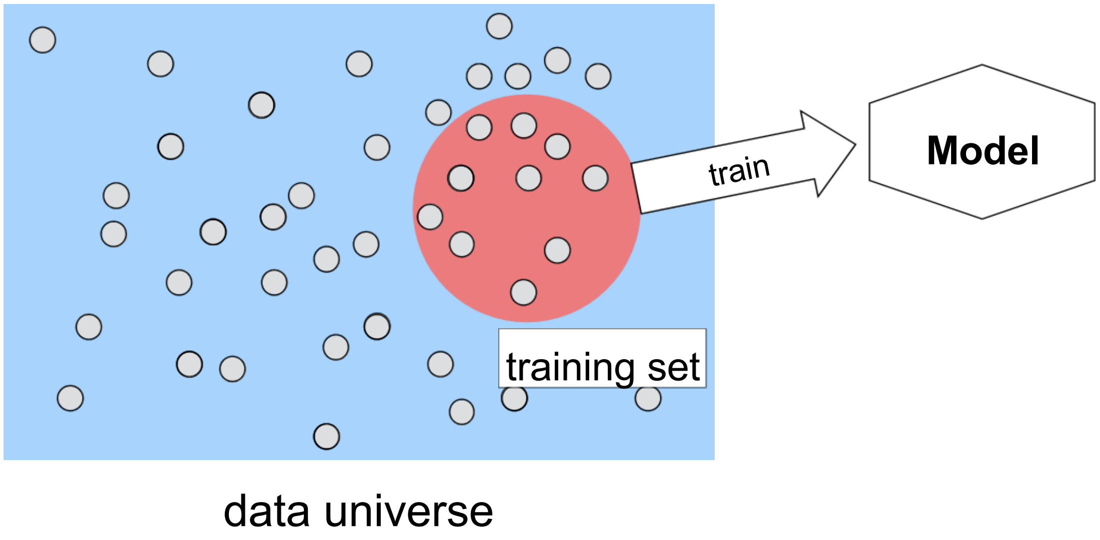
Privacy Learning



Privacy

Learning

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

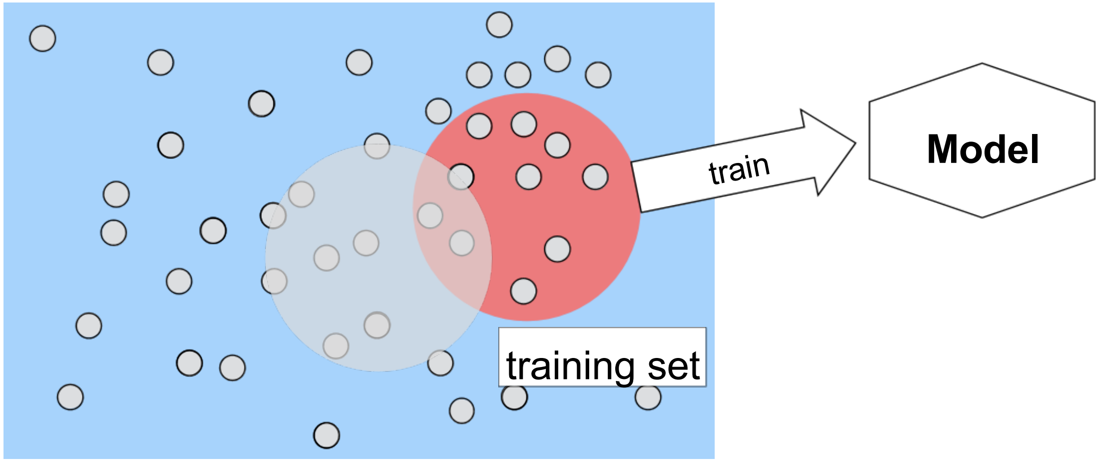


Privacy

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

Learning

Does the model generalize to data outside the training set?



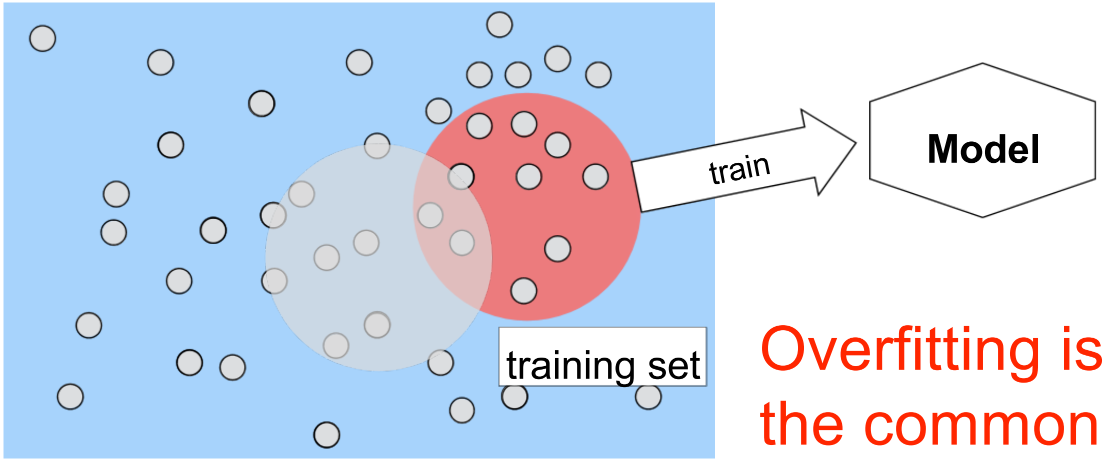
data universe

Privacy

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

Learning

Does the model generalize to data outside the training set?

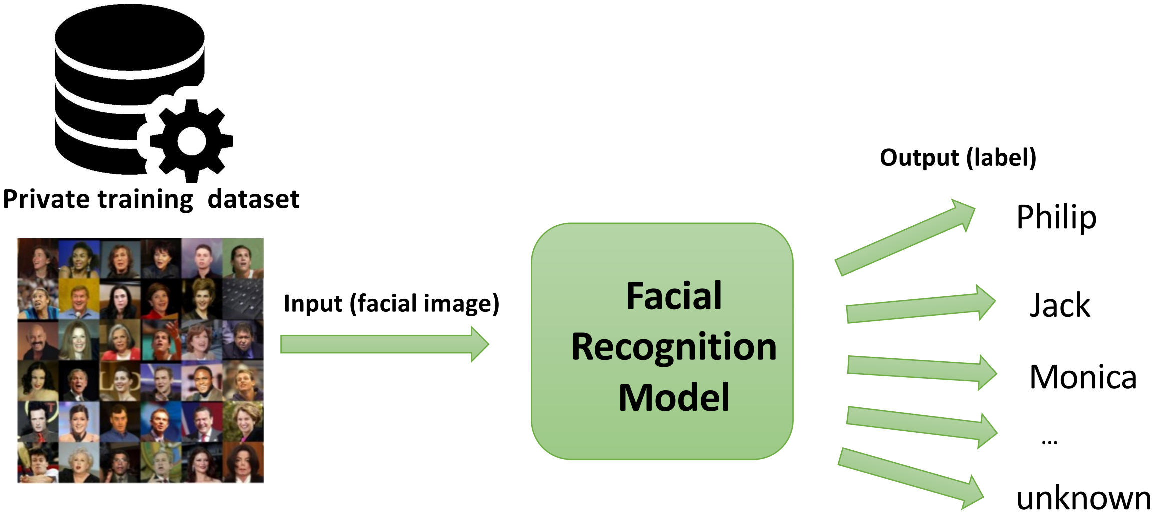


data universe

Overfitting is the common enemy!

Feature Inference Attacks

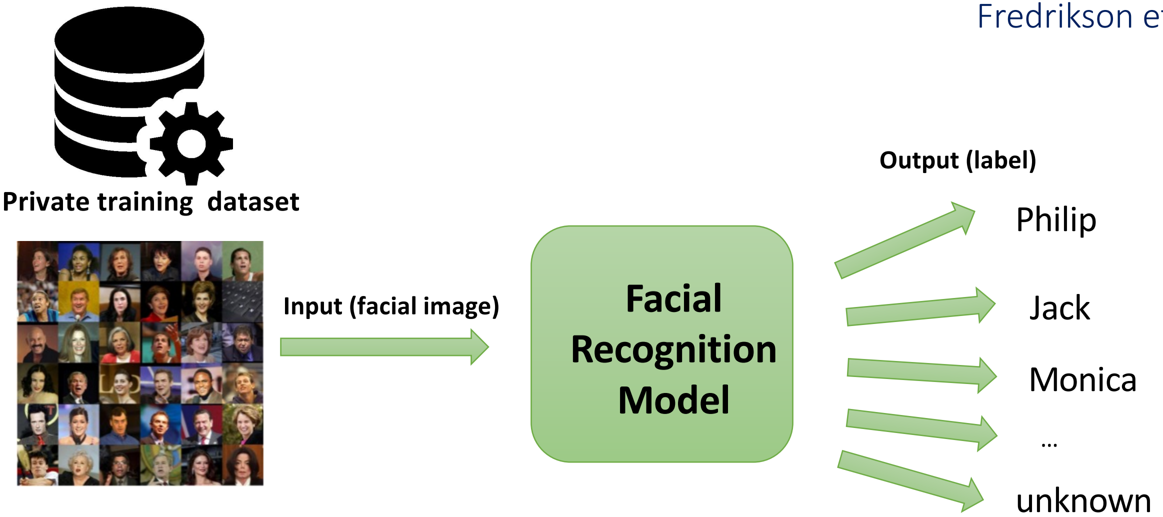
Fredrikson et al. 2015



Can I reconstruct the image of someone?

Feature Inference Attacks

Fredrikson et al. 2015



Can I reconstruct the image of someone?

Find **x** to minimize $c(\mathbf{x}) = 1 - f_{label}(\mathbf{x})$

Feature Inference Attacks

Fredrikson et al. 2015

Private training dataset

Input (facial image)

Facial Recognition Model

Output (label)

- Philip
- Jack
- Monica
- …
- unknown

Can I reconstruct the image of someone?

Find **x** to minimize $c(\mathbf{x}) = 1 - f_{label}(\mathbf{x})$

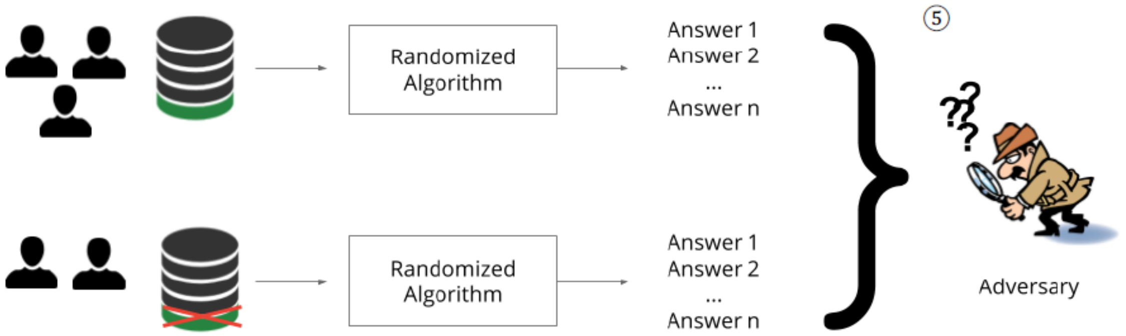
Use Gradient Descent (require white box access of the model)

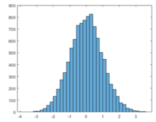
Outline

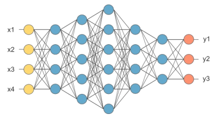
- Privacy attacks
  - Membership inference attacks, model inversion attacks, secret sharer
- Privacy-preserving deep learning
  - Differential privacy, gradient perturbation, noisy ensemble
- Security attacks
  - Adversarial example attacks, poisoning attacks, backdoor attacks
- Robust deep learning
  - Detection and reform, adversarial training, certified robustness

EMORY UNIVERSITY

Differential Privacy (DP) [Dwork 06]







Differential Privacy (DP) in practice

Differential Privacy

Definition 2.4 (Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private.

quantifies information leakage

allows for a small probability of failure

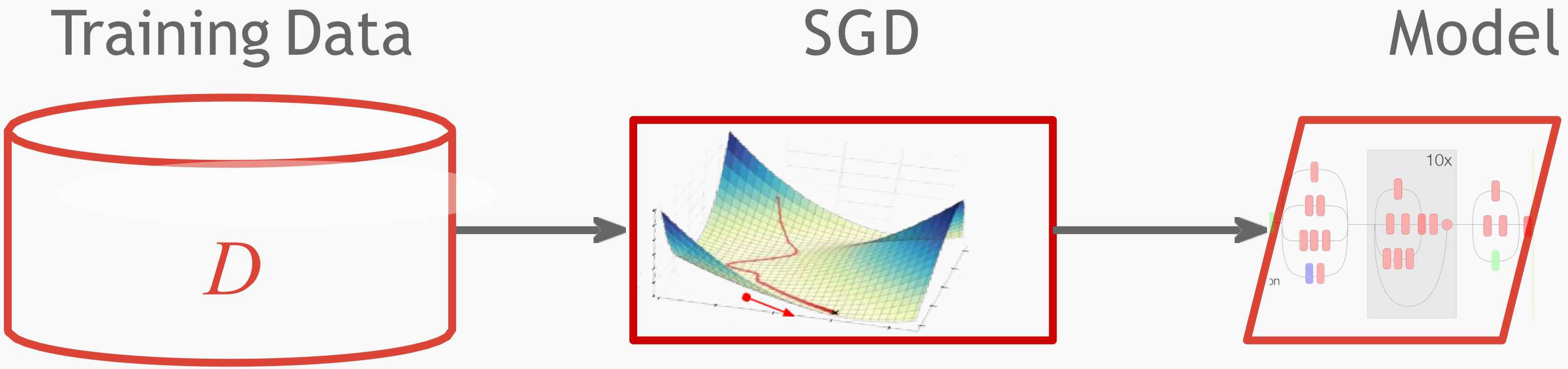
DEEP LEARNING WITH DIFFERENTIAL PRIVACY

Martin Abadi, Andy Chu, Ian Goodfellow*,
Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang
Google

* OpenAI

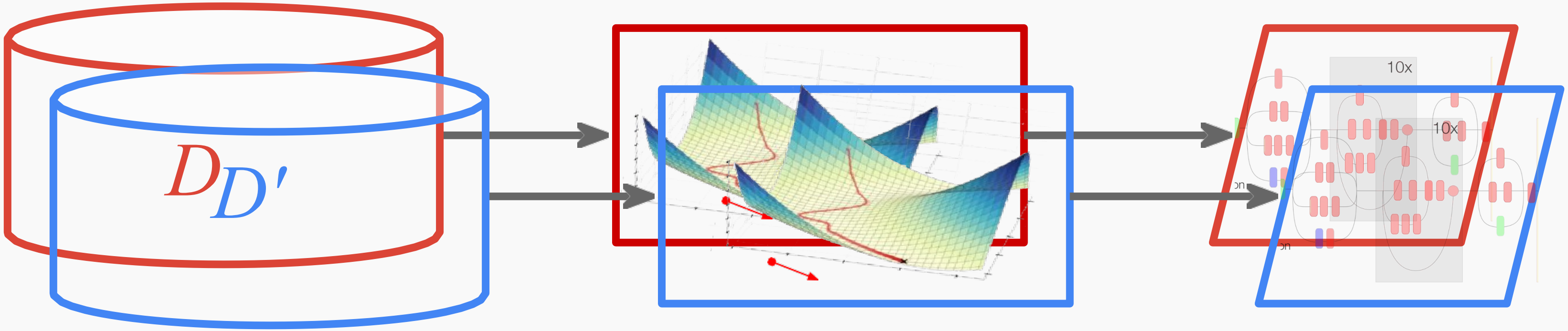
Deep Learning with Differential Privacy, CCS, 2016

Training a deep learning network

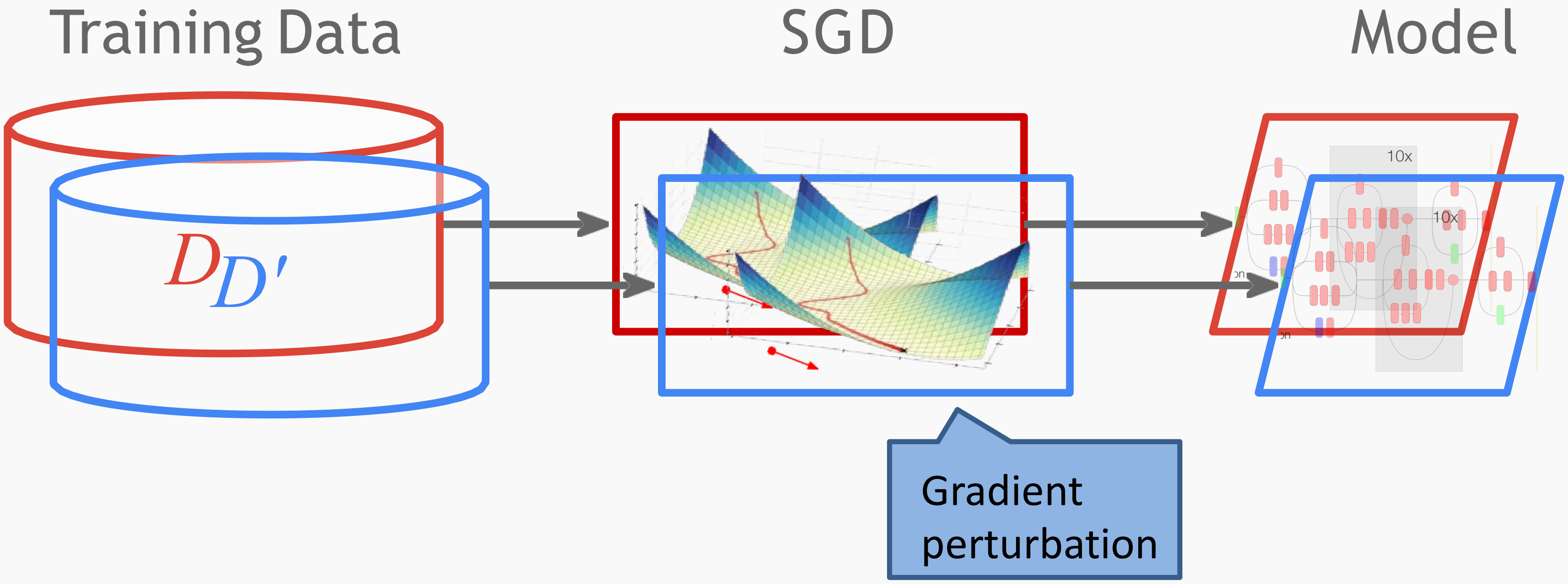


Interpreting Differential Privacy

Training Data SGD Model



Achieving Differential Privacy - DPSGD



Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$ Clipping with bound C

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$ Add noise

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method. Privacy composition

Our Datasets: "Fruit Flies of Machine Learning"

MNIST dataset:

70,000 images

28×28 pixels each

CIFAR-10 dataset:

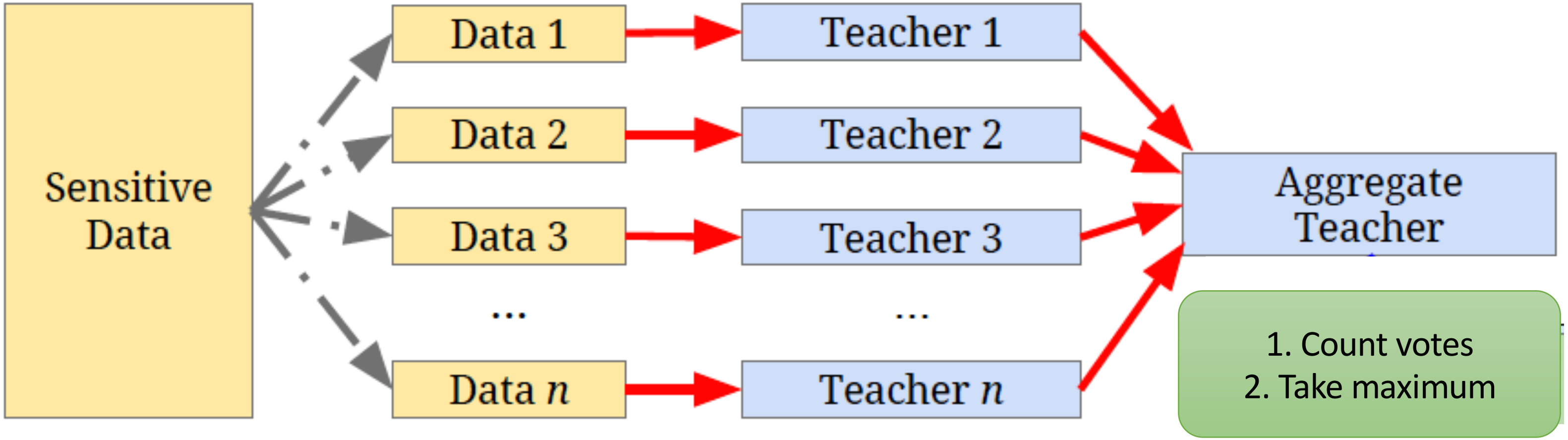
60,000 color images

32×32 pixels each

Summary of Results

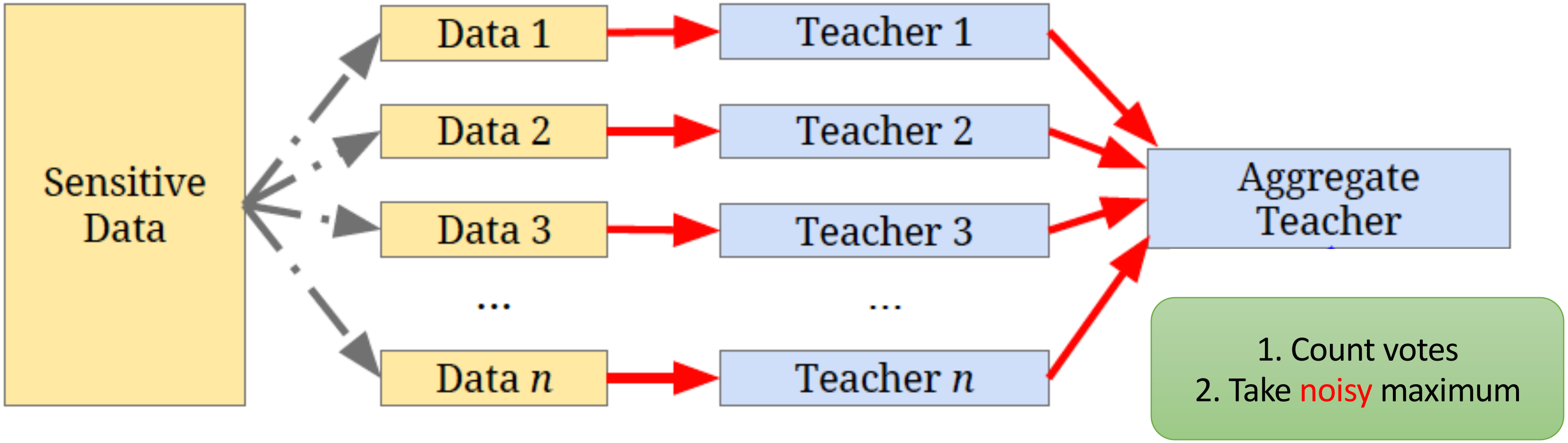
	Baseline	[SS15]	[WKC+16]	this work		
	no privacy	reports ε per parameter	$\varepsilon = 2$	$\varepsilon = 8$ $\delta = 10^{-5}$	$\varepsilon = 2$ $\delta = 10^{-5}$	$\varepsilon = 0.5$ $\delta = 10^{-5}$
MNIST	98.3%	98%	80%	97%	95%	90%
CIFAR-10	80%			73%	67%	

Private Aggregation of Teacher Ensembles (PATE)



How can we ensure DP for the teacher ensemble?

Private Aggregation of Teacher Ensembles (PATE)



The noisy aggregated teacher:

- **Each prediction increases total privacy loss.**

  privacy budgets create a tension between the accuracy and number of predictions

Evaluation

Dataset	ε	δ	Queries	Non-Private Baseline	Student Accuracy
MNIST	2.04	10^{-5}	100	99.18%	98.00%
MNIST	8.03	10^{-5}	1000	99.18%	98.10%
SVHN	5.04	10^{-6}	500	92.80%	82.72%
SVHN	8.19	10^{-6}	1000	92.80%	90.66%

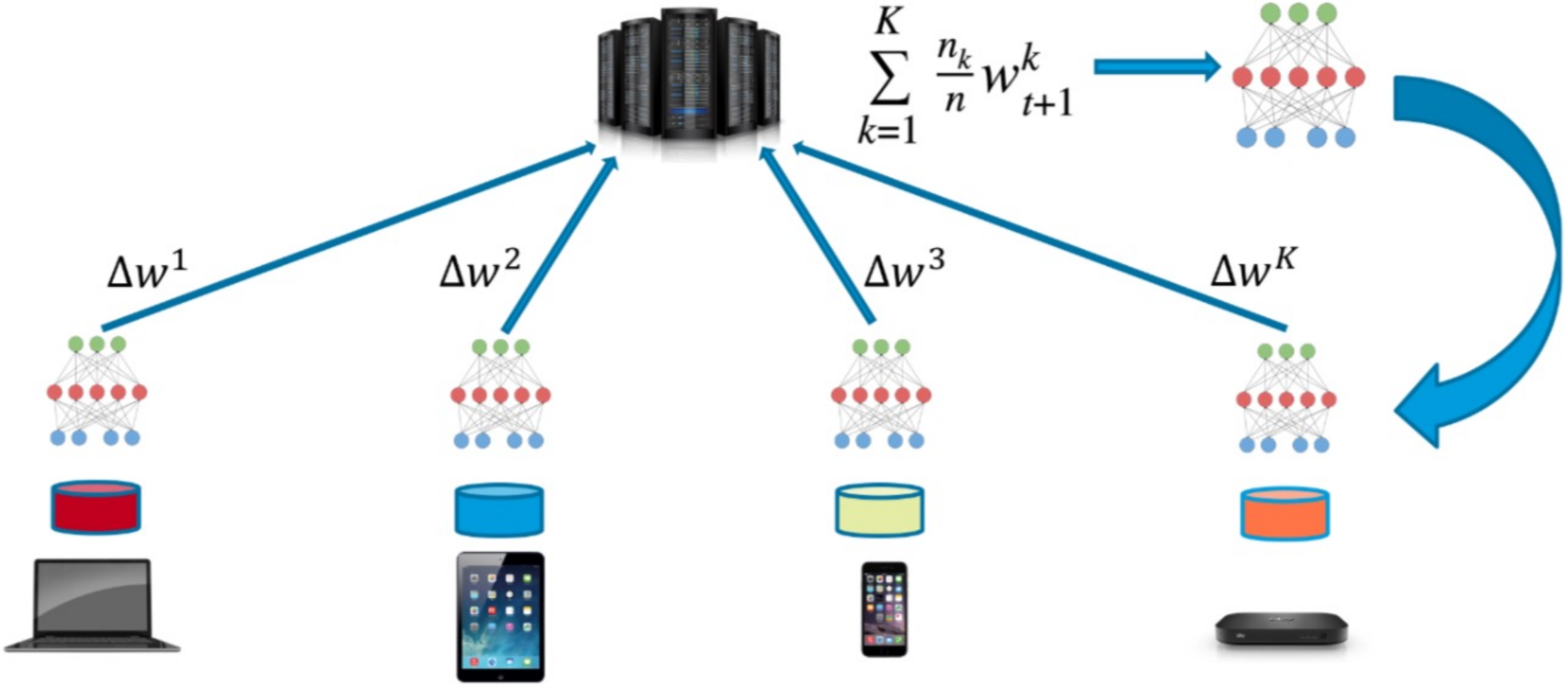
M Abadi et al. (2016) *Deep Learning with Differential Privacy*

$(0.5, 10^{-5})$ 90%

$(2, 10^{-5})$ 95%

$(8, 10^{-5})$ 97%

Federated Learning



Federated Averaging Algorithm

Server executes:

initialize x_0

for each round $t = 1, 2, \ldots,$ T **do**

$S_t \leftarrow$ (random set of M clients)

for each client $i \in S_t$ **in parallel do**

$x^i_{t+1} \leftarrow \text{ClientUpdate}(i, x_t)$

$x_{t+1} \leftarrow \sum_{k=1}^{M} \frac{1}{M} x^i_{t+1}$

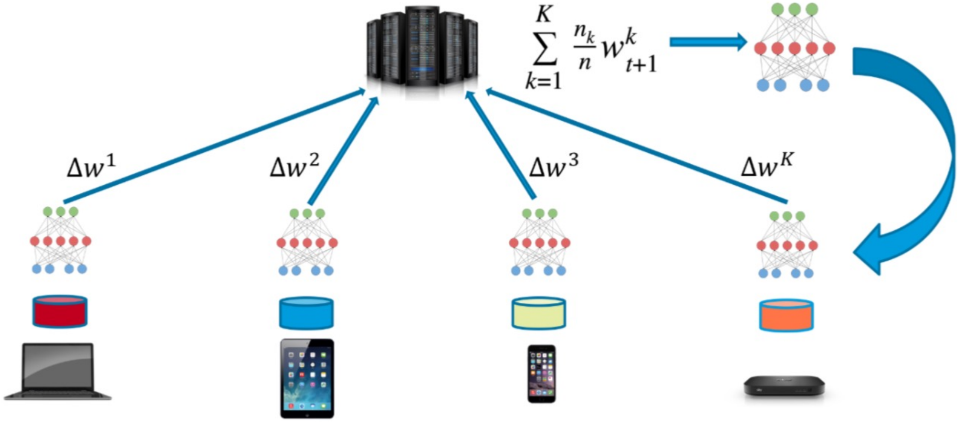
ClientUpdate(i, x)**:**

for local step $j = 1, \ldots, K$ **do**

$x \leftarrow x - \eta \nabla f(x; z)$ for $z \sim \mathcal{P}_i$

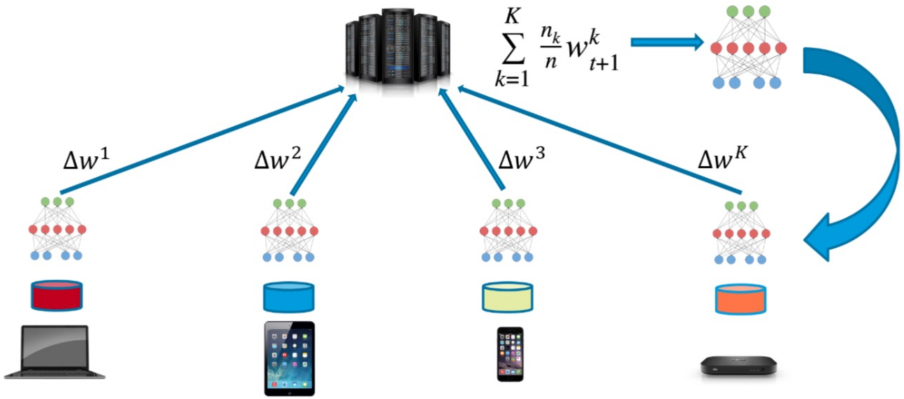
return x to server

Algorithm 1: Federated Averaging (local SGD), when all clients have the same amount of data.



Federated Learning with Differential Privacy

- Server is trusted – ensure DP for global model
  - DP at server
- Server is not trusted – ensure DP for gradients and resulting model
  - DP at client

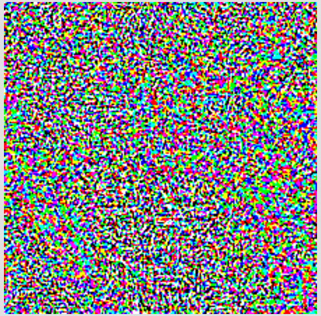


Outline

- Privacy attacks
  - Membership inference attacks, model inversion attacks, secret sharer
- Privacy-preserving deep learning
  - Differential privacy, gradient perturbation, noisy ensemble
- Security attacks
  - Adversarial example attacks, poisoning attacks, backdoor attacks
- Robust deep learning
  - Detection and reform, adversarial training, certified robustness

ADVERSARIAL EXAMPLES

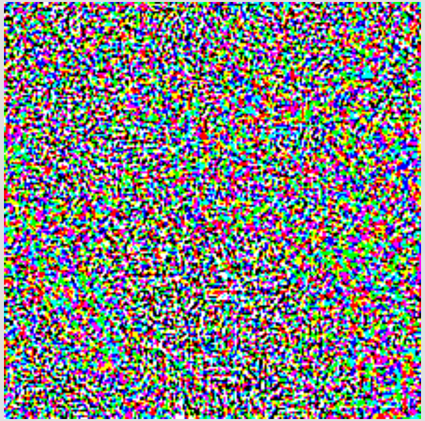
Classified as panda



Small adversarial noise

Classified as gibbon

ADVERSARIAL EXAMPLES



Small adversarial noise

ADVERSARIAL EXAMPLES

Constraint:
$\| x^{*} - x \|_{p} \leq \epsilon$
$p = 0, 1, 2, \infty$

1. Non-targeted attack:
   $y_{true} \neq y^{*}$

2. Targeted attack:
   y^{*} is the target label specified by the adversary

Figure by Qiuchen Zhang

Carlini and Wagner (C&W) (2017)

- Followed L-BFGS work
- Dealt with box constraints by change of variables: X^{adv} = 0.5(tanh(w) + 1)
- K: determine confidence level
- Used Adam optimizer

$$\|x^{adv} - x\|_p + c \underbrace{\max\left(\max_{i \neq Y} f(x^{adv})_i - f(x^{adv})_Y, -\kappa\right)}_{\text{Loss function}} \rightarrow \text{minimum}$$

0.5(tanh(w) + 1)

confidence

Adversarial Example Defenses

- Adversarial training (training stage)
- Detection and reformation (inference stage)
- Preprocessing (inference stage)
- Randomized smoothing for certified robustness (inference stage)

References
Privacy attacks and privacy-preserving deep learning

- Membership Inference Attacks Against Machine Learning Models, S&P, 2017
- Model inversion attacks that exploit confidence information and basic countermeasures, CCS, 2015
- The secret sharer: Evaluating and testing unintended memorization in neural networks, USENIX Security, 2019
- The Algorithmic Foundations of Differential Privacy, 2014 (book Ch 3)
- Deep Learning with Differential Privacy, CCS, 2016
- Private Stochastic Nonconvex Optimization with Better Utility Rates, IJCAI 2021
- Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data, ICLR, 2017
- Evaluating Differentially Private Machine Learning in Practice, USENIX Security 2019

EMORY UNIVERSITY

References
Adversarial example and poisoning attacks and robust deep learning

- Explaining and harnessing adversarial examples, ICLR 2015
- Towards Evaluating the Robustness of Neural Networks, S&P, 2017
- Learning to Attack: Adversarial Transformation Networks, AAAI 2018
- Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks, S&P 2016
- Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018
- MagNet: a Two-Pronged Defense against Adversarial Examples, CCS 2017
- Certified robustness to adversarial examples with differential privacy, S&P, 2019
- Certified adversarial robustness via randomized smoothing, ICML, 2019
- Integer-arithmetic-only Certified Robustness for Quantized Neural Networks, ICCV 2021
- Poisoning Attacks against Support Vector Machines, ICML 2012
- Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, S&P, 2018
- Data Poisoning against Differentially-Private Learners: Attacks and Defenses, IJCAI 2019