Machine Learning - Security and Privacy

Part I

Asst. Prof. Sinem Sav

Slides adapted from Carmela Troncoso, Reza Shokri, Bryan Ford, Jamie Hayes, Bogdan Kulynych, Nicolas Papernot, Nicholas Carlini, Rebekah Overdorf, Florian Tramer, Jean-Pierre Hubaux…

Machine Learning – Security and Privacy (I)

- Basics
- Model stealing
- Privacy issues
- *Altering the output*
- *Biases and fallacies*
- *Federated learning*

Machine Learning – Security and Privacy (I)

- **Basics**
- Model stealing
- Privacy issues
- *Altering the output*
- *Biases and fallacies*
- *Federated learning*

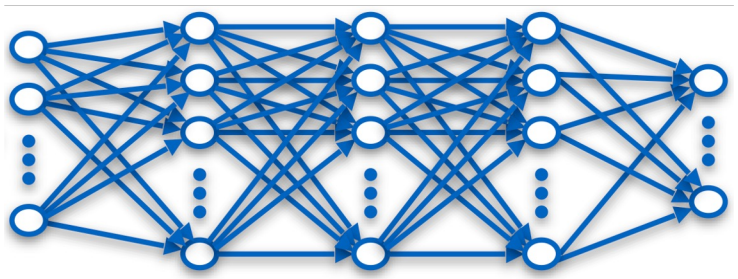
Machine Learning (ML)

Definition (Wikipedia)

Machine learning *[...] gives "**computers the ability to learn without being explicitly programmed**" [and] [...] explores the study and construction of **algorithms that can learn from and make predictions on data** – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs.*

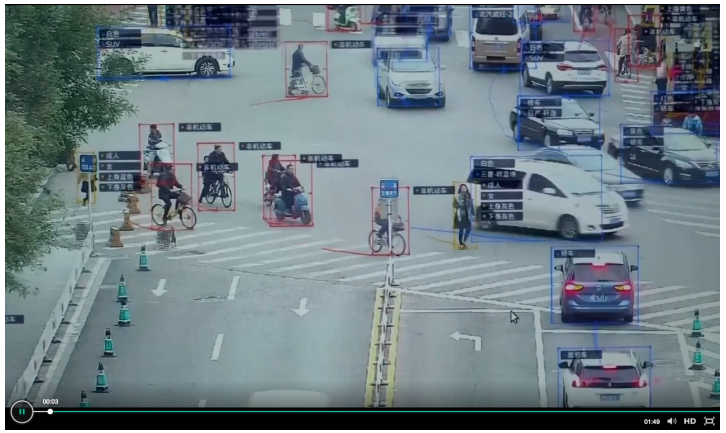
User data

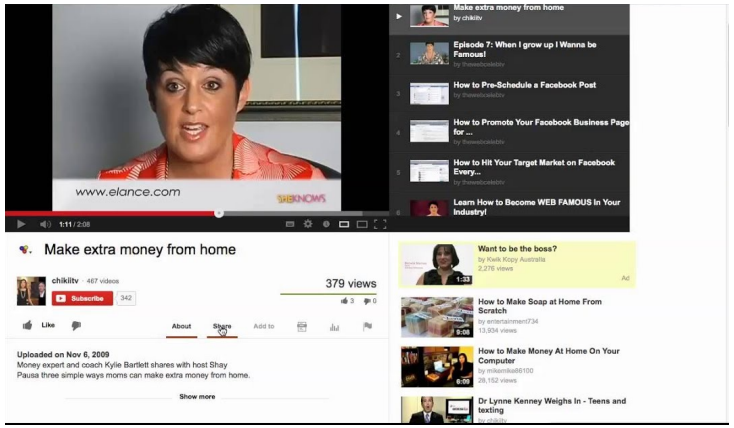
Machine learning

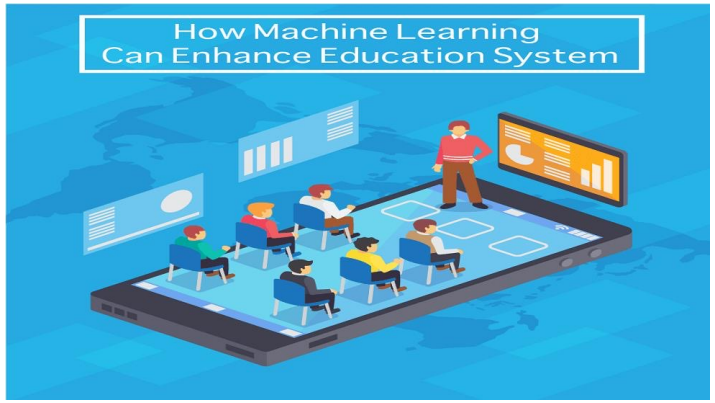


Services

Machine Learning Is Becoming Ubiquitous



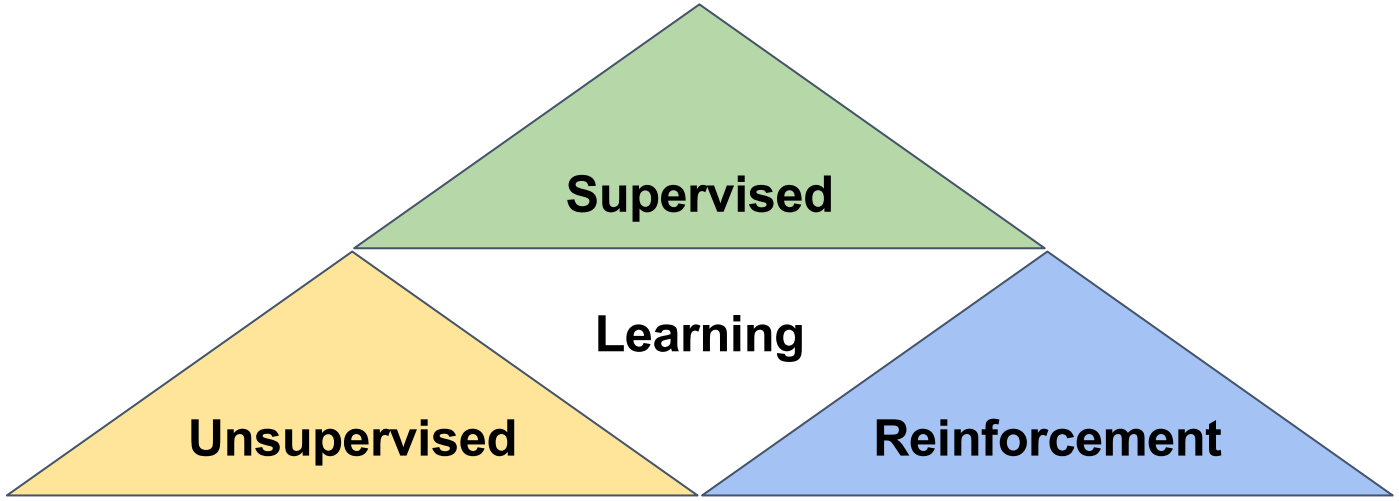




Machine Learning Taxonomy

Machine learning can be separated into 3 main categories

- Labeled data
- Direct feedback
- Predict outcome/future
- Example: estimate a house price



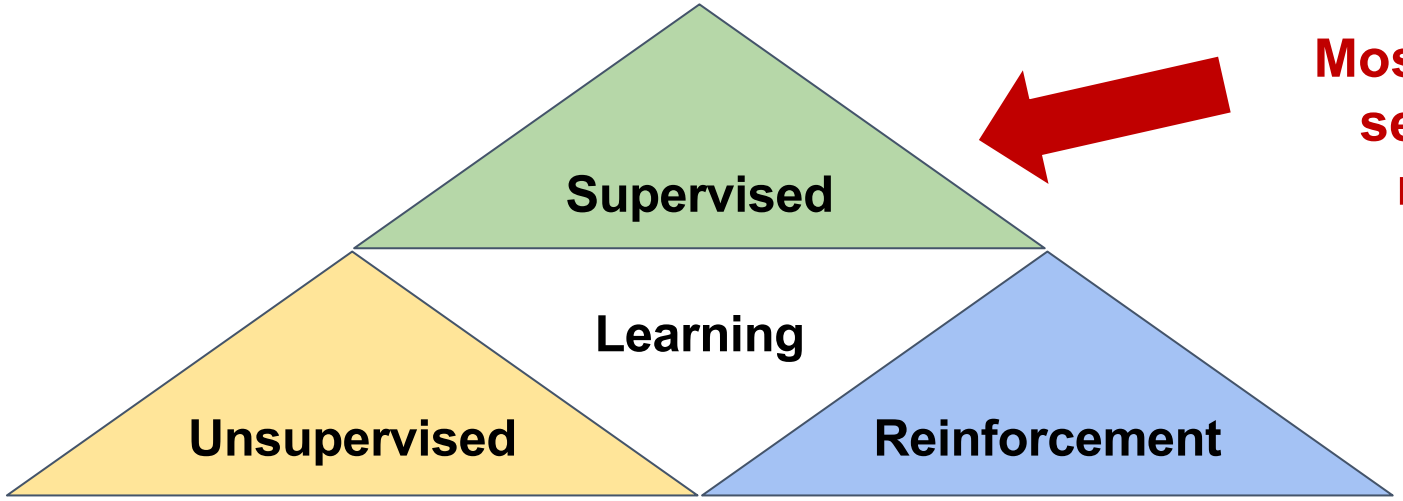
- No labels
- No feedback
- “Find hidden structures”
- Example: customer segmentation

- Decision process
- Reward system
- Learn series of actions
- Example: learning to play a board game (e.g., chess)

Machine Learning Taxonomy

Machine learning can be separated into 3 main categories

- Labeled data
- Direct feedback
- Predict outcome/future
- Example: estimate a house price

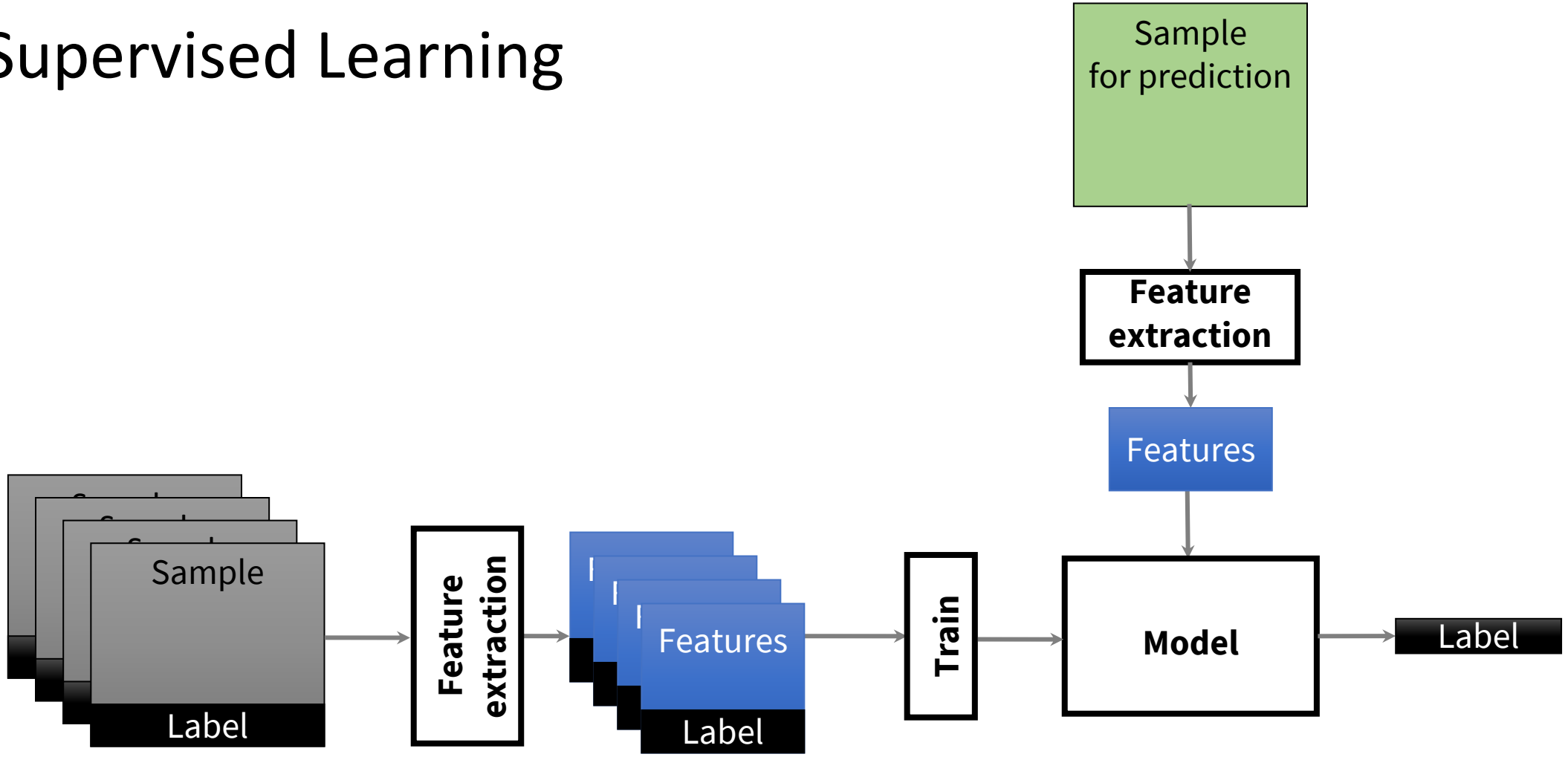


Most used and where security has been mostly studied

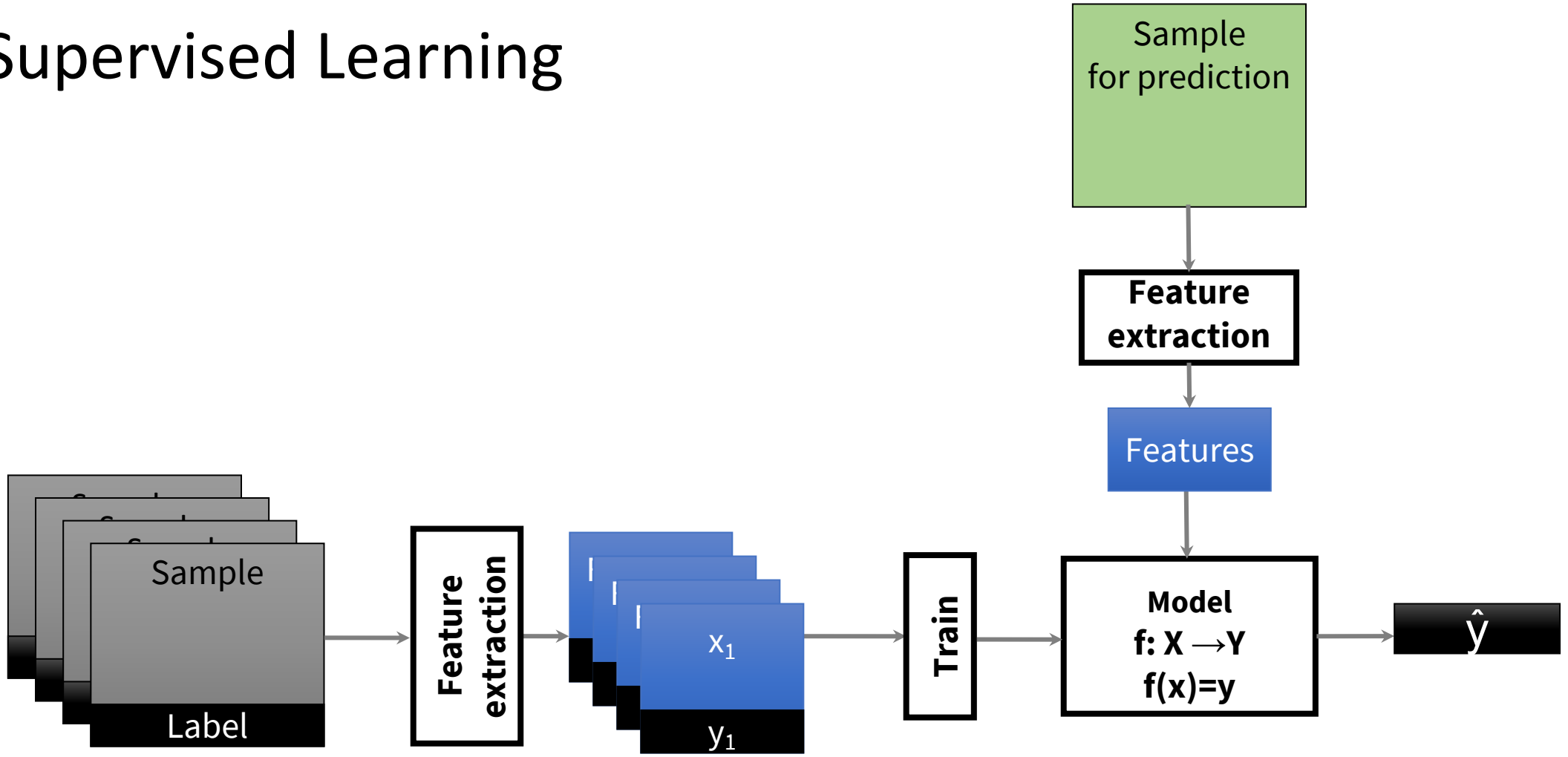
- No labels
- No feedback
- "Find hidden structures"
- Example: customer segmentation

- Decision process
- Reward system
- Learn series of actions
- Example: learning to play a board game (e.g., chess)

Supervised Learning



Supervised Learning



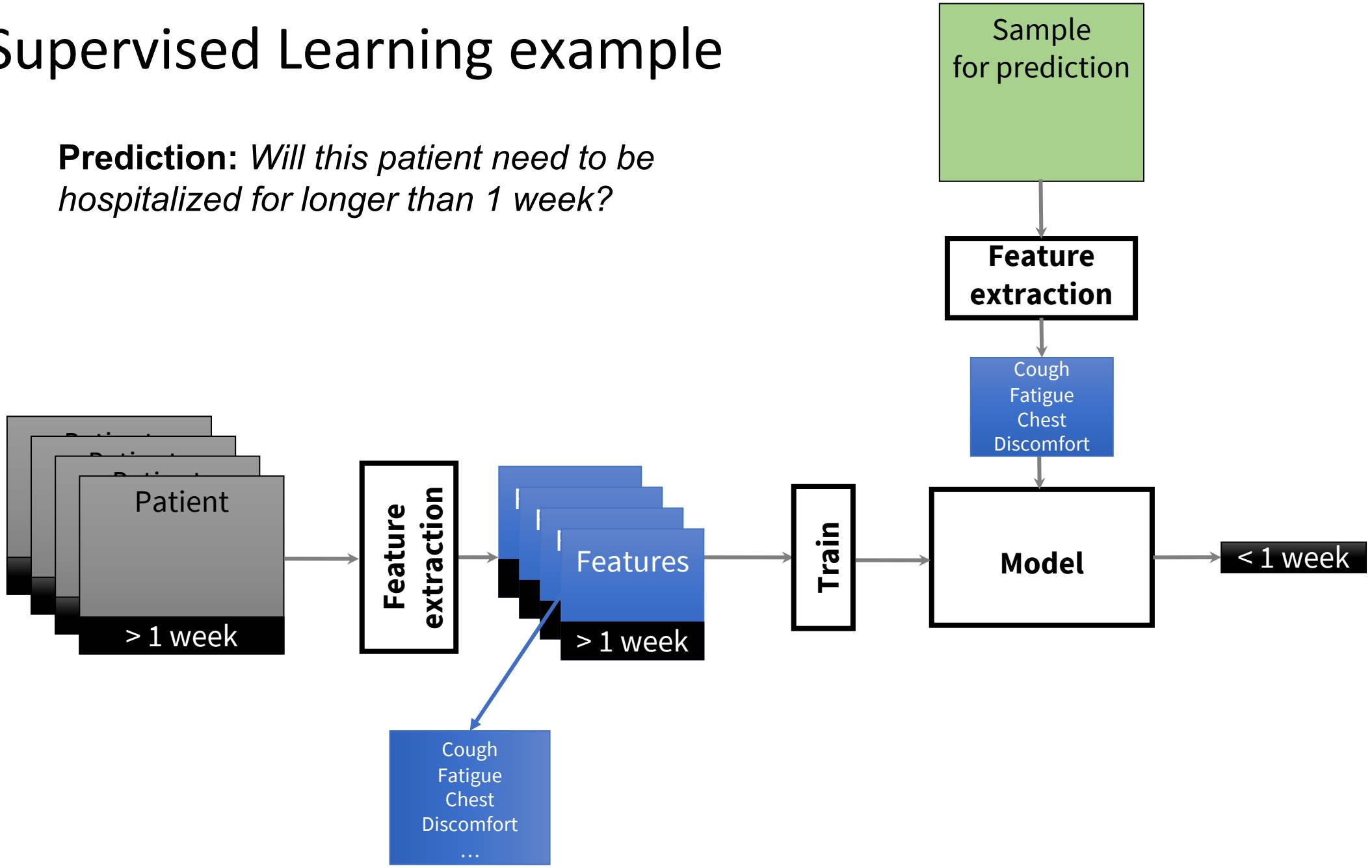
$\{(x_1,y_1),\ldots,(x_n,y_n)\} \in X \times Y$

X: Input space
Y: Output space
x_i: Feature vector
y_i: Output label (= class)

Classification: Y categorical
Regression: $Y = \mathbb{R}$

Supervised Learning example

Prediction: *Will this patient need to be hospitalized for longer than 1 week?*



Adversary Goals

Machine learning under adversarial conditions
Adversarial Goals

Confidentiality and Privacy

- Confidentiality of the model itself (e.g., intellectual property)
- Privacy of the training or test data (e.g., medical records)

Integrity and Availability

- Integrity of the predictions (wrt expected outcome)
- Availability of the system deploying machine learning

Machine learning under adversarial conditions
Adversarial Capabilities

Black-box attacks

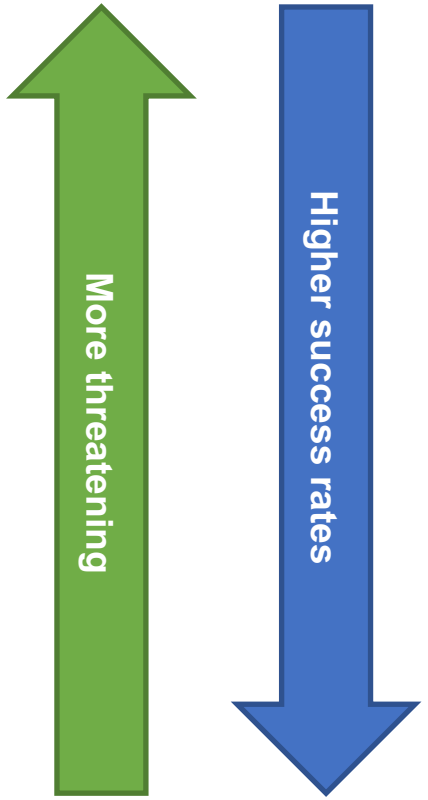
Model architecture and parameters unknown
Can only interact *blindly* with the model

Grey-box attacks

Model architecture known, parameters unknown
Can only interact with the model, but has information about the type of model

White-box attacks

Known architecture and parameters
Can replicate the model and use the model's internal parameters in the attack

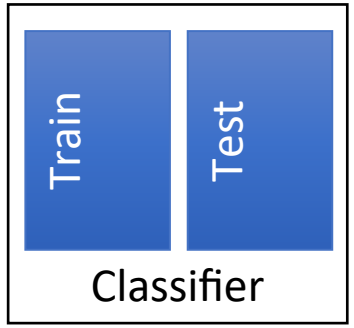


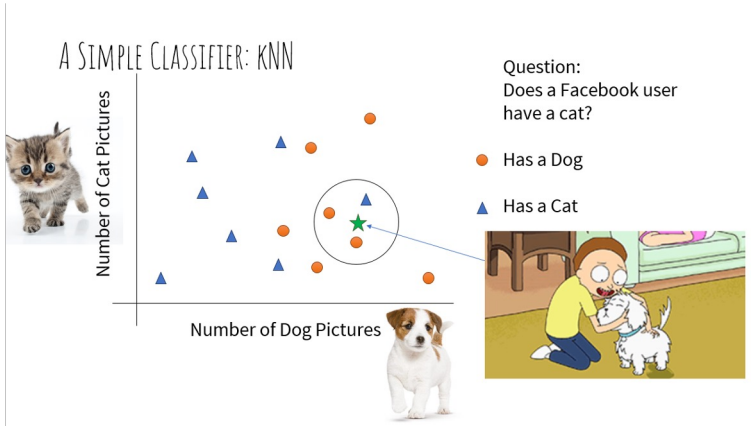
Machine Learning – Security and Privacy (I)

- Basics
- **Model stealing**
- Privacy issues
- *Altering the output*
- *Biases and fallacies*
- *Federated learning*

Machine Learning as a Service

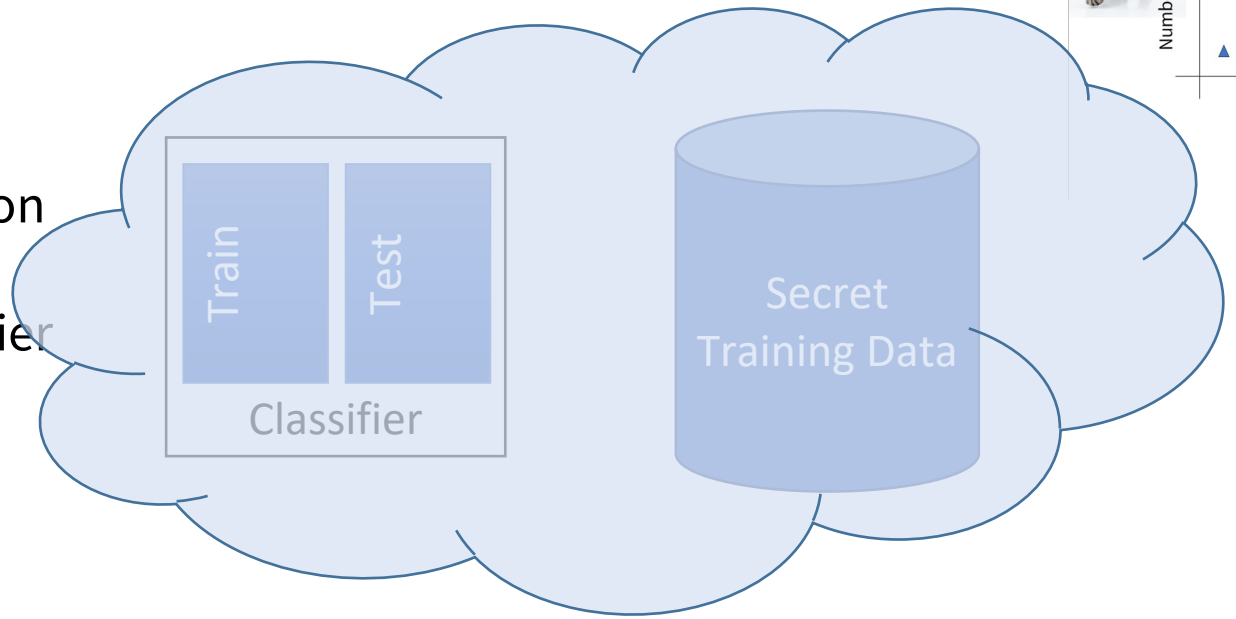
1. The Cloud, e.g., Amazon ML or Google Prediction API, (pre-)trains a classifier using their own data

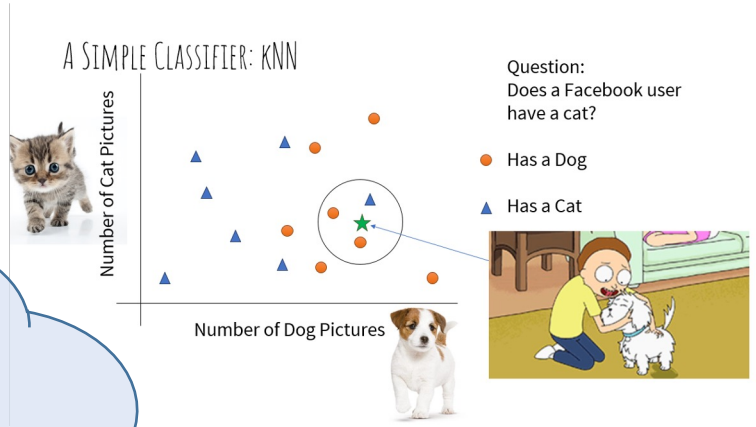




Machine Learning as a Service

1. The Cloud, e.g., Amazon ML or Google Prediction API, (pre-)trains a classifier using their own data

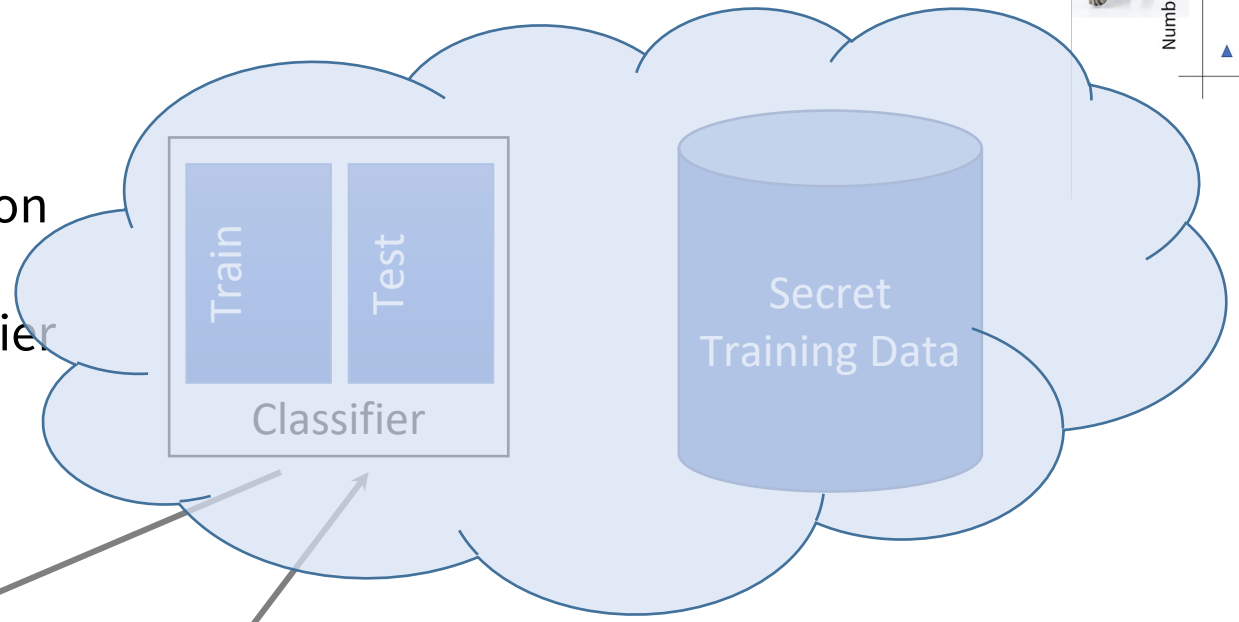


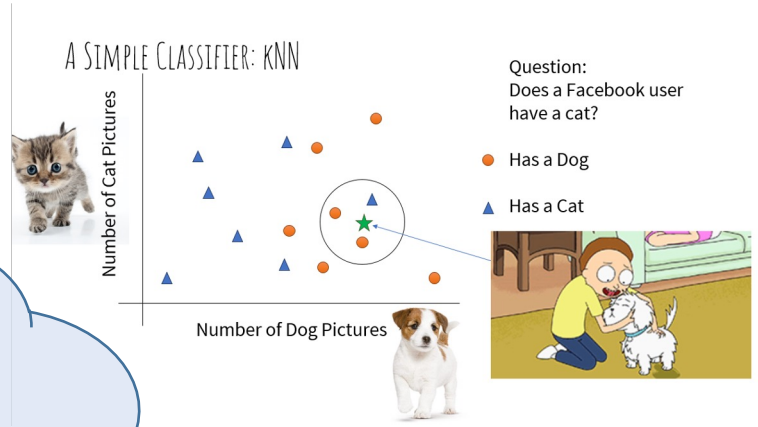


2. Make this classifier available as a service for users to query

Machine Learning as a Service

1. The Cloud, e.g., Amazon ML or Google Prediction API, (pre-)trains a classifier using their own data





2. Make this classifier available as a service for users to query

Cat

Bob's Profile

Features

3. The user makes a query: "Given his profile (photos, posts, metadata), what pet has this Facebook user?"

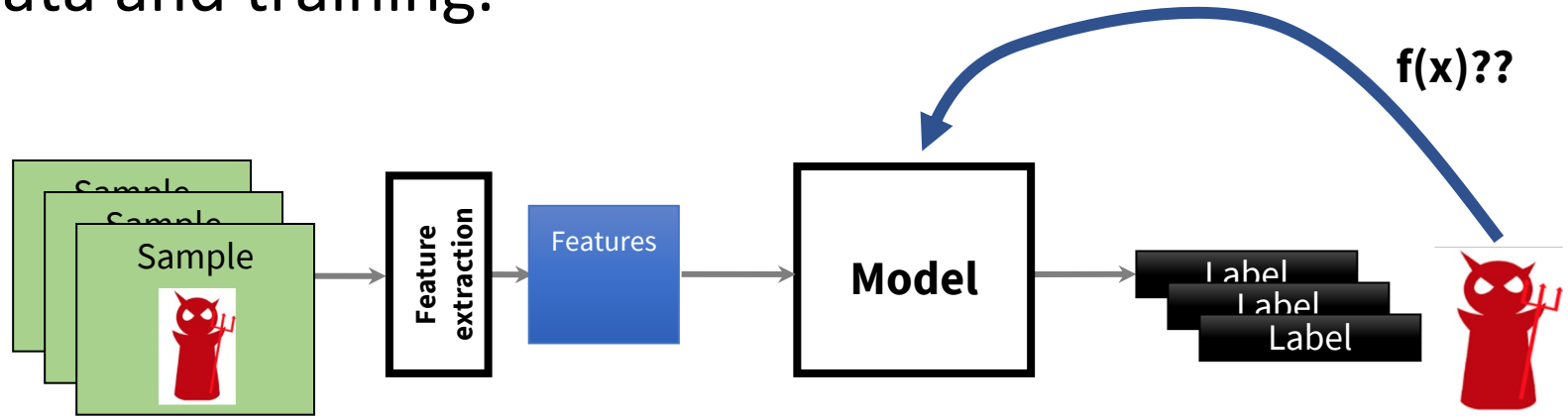
Model stealing (model extraction)

Confidentiality of the model itself (e.g., intellectual property)

Good ML models require considerable investment:

- Collecting data takes time and money
- Training infrastructure is expensive

Goal: "steal" the expensive model by observing its outputs with less cost than obtaining the data and training.



F. Tramer et al. Stealing Machine Learning Models via Prediction APIs, USENIX Security 2016 https://arxiv.org/abs/1609.02943
Video: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

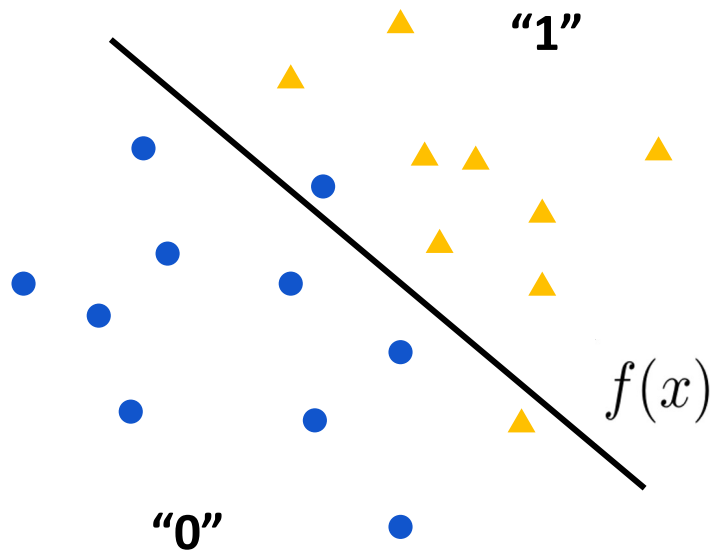
Machine Learning 101: Linear models

When used for classification, a linear model employs a linear function (separating hyperplane) to produce a decision

- E.g., logistic regression, SVM with linear kernel

$$f(x) = w \cdot x + b$$

If *f(x) > t,* the output class is “1”, otherwise is “0”.



Stealing a linear model

Assume adversary knows the model is a linear architecture (*Grey-box model*).

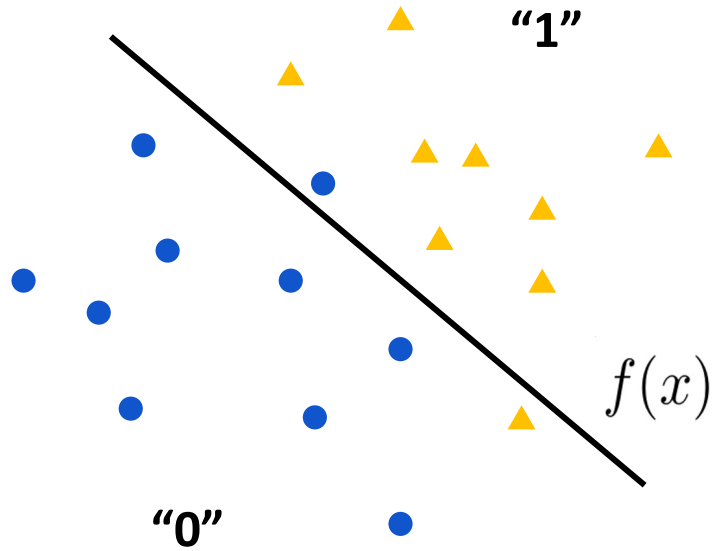
Assume **x** is **two**-dimensional:

$$f(x) = w_1 x_1 + w_2 x_2 + b$$

Adversary's goal: steal parameters w, b

How many input-output pairs *(x, f(x))*
the adversary needs to observe to steal the model?

What if the x was d-dimensional?

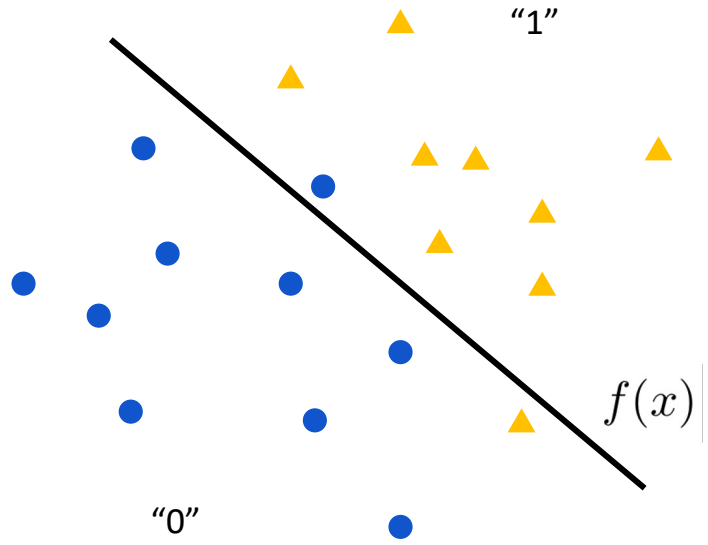


Equation-solving attack

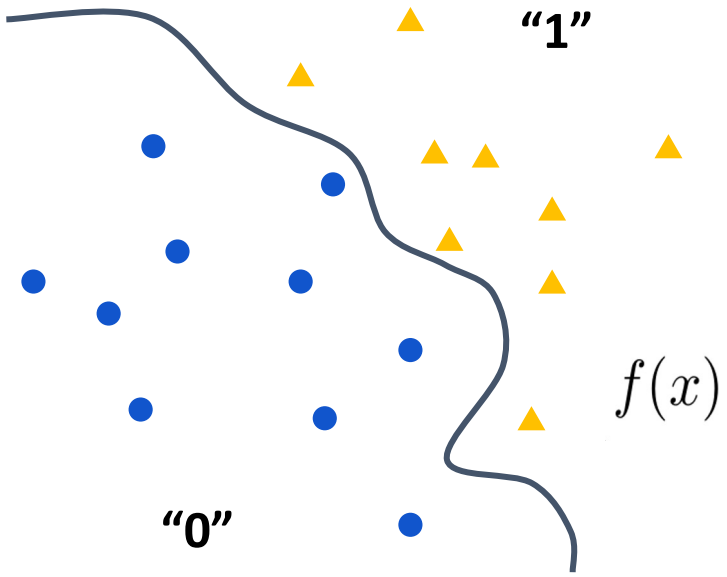
For **two-dimensional** x only need **3 queries!**

In general, if a linear model uses ***d features***, the adversary needs ***d*+1** different queries to steal by solving the linear system for w, b

$$w \cdot x^{(i)} + b = f(x^{(i)})$$



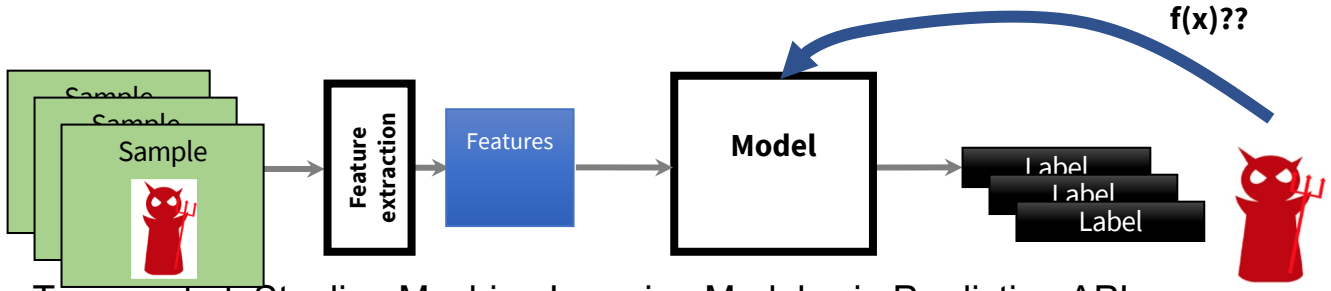
Stealing a non-linear model



Assume adversary knows the model's architecture

$$f_w(x)$$

Adversary's goal: steal parameters w

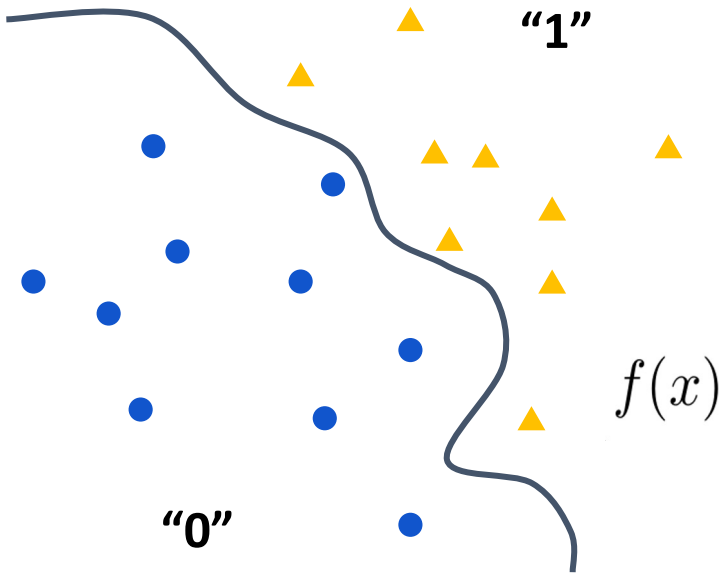


Tramer et al. Stealing Machine Learning Models via Prediction APIs

Retraining attack

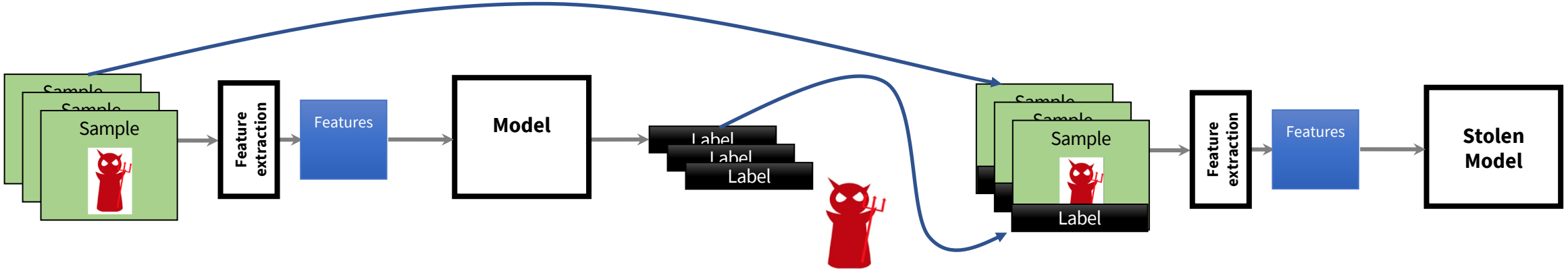
Observe many queries $X = (x, f(x))$, and fit the model on X like on any other training data!

$$f_w(x)$$



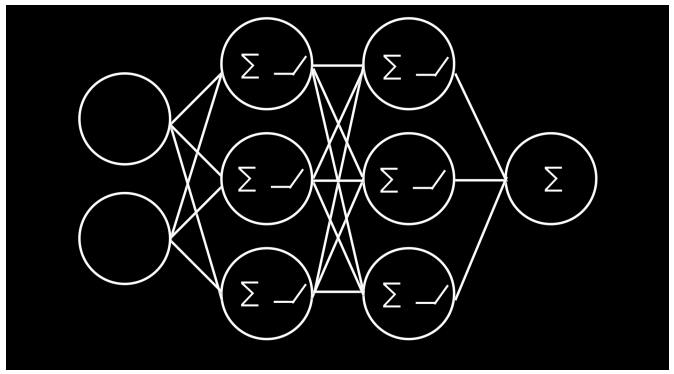
Takes many queries. For a neural network with 2K parameters, need 11K queries to get 99.9% similarity.

More recent work has reduced these numbers.

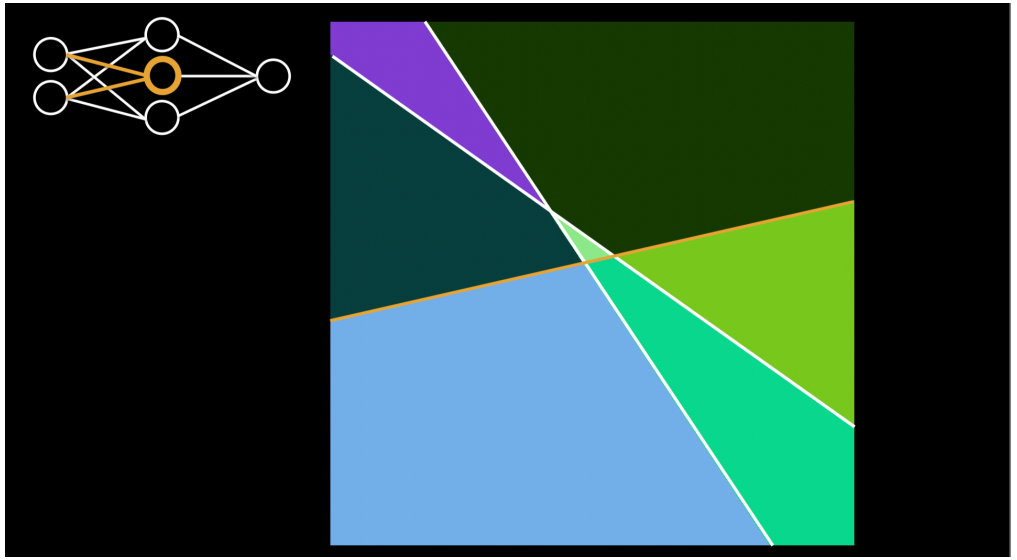


Tramer et al. Stealing Machine Learning Models via Prediction APIs

More on stealing neural networks



Given (oracle) query access to a neural network, learned through stochastic gradient descent, can we extract a *functionally equivalent* model?



Nicholas Carlini et al, Cryptanalytic Extraction of Neural Network Models, EuroCrypt 2020 https://arxiv.org/abs/2003.04884
Video available at: https://nicholas.carlini.com

Implication of this attack against neural networks

- Achievement: extracted a 100,000 parameter neural network trained on the MNIST digit recognition task with 2^21.5 queries in under an hour
  ➔ Implications for ML and cryptographic research
- Assumption of the field of secure inference: observing the output of a neural network does not reveal the weights
  ➔ This assumption is false, and therefore the field of secure inference will need to develop new techniques to protect the secrecy of trained models

Nicholas Carlini et al, Cryptanalytic Extraction of Neural Network Models. 2020

Preventing model stealing

First attack = 2016 First defenses ~ 2019

Output perturbation [1]

Add noise to the probabilities output by the model to hinder reconstruction, but not accuracy

Detect suspicious queries [2]

Identify deviations from expected on distribution of successive queries from a client

Very recent techniques! We don't yet know how robust they are

[1] Lee et al. Defending Against Neural Network Model Stealing Attacks Using Deceptive Perturbations. 2019
[2] Juuti et al. PRADA: Protecting Against DNN Model Stealing Attacks. 2019

Takeaways on model stealing

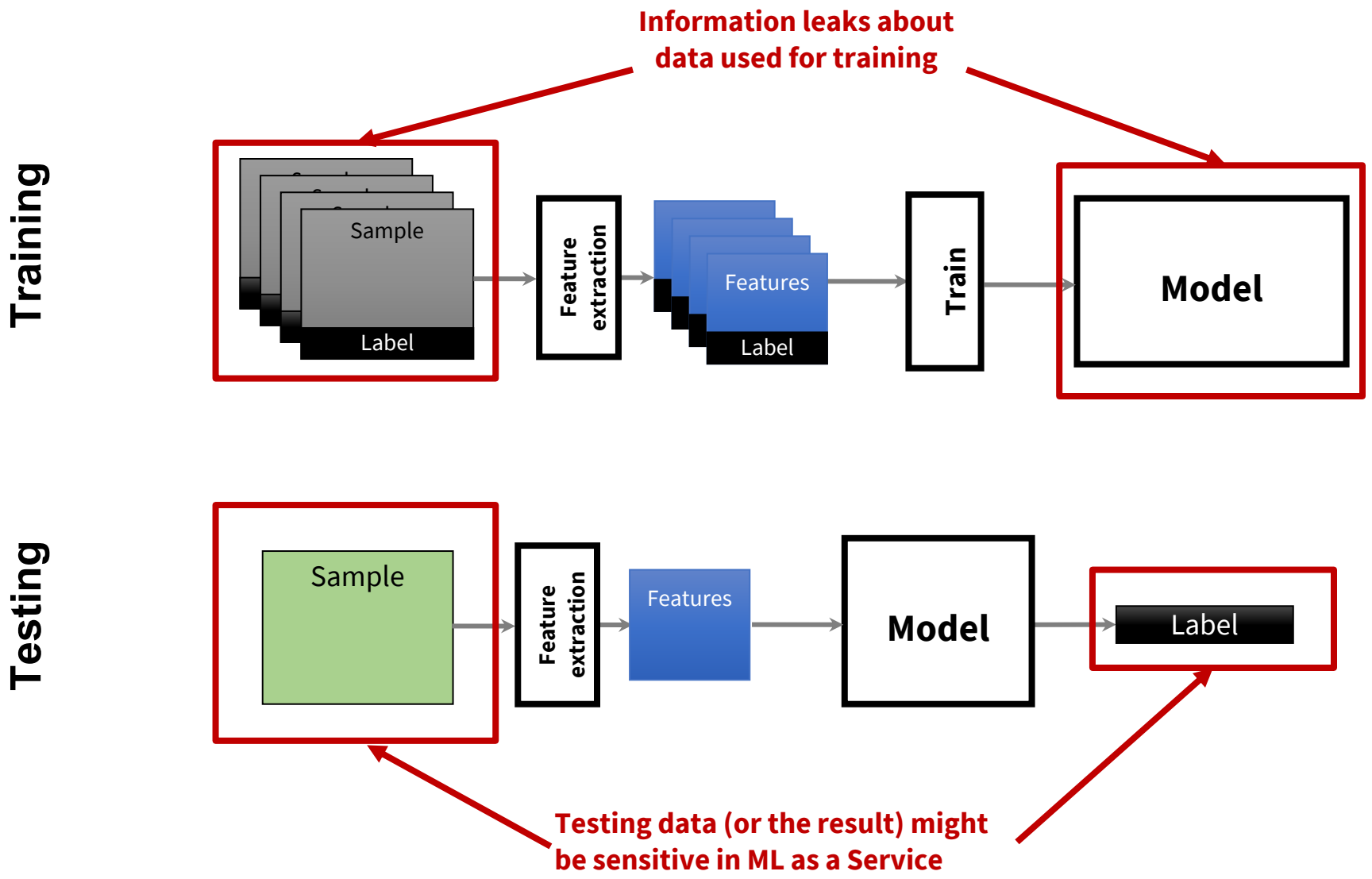
- Many models are susceptible to stealing
- Complex topic, lots of ongoing research
- Crucial issue for ML as a service
- Many problems, some solutions ☺️

- Very young field, the situation will (probably) improve!
  ➔ Are you willing to help?

Machine Learning – Security and Privacy (I)

- Basics
- Model stealing
- **Privacy issues**
- *Altering the output*
- *Biases and fallacies*
- *Federated learning*

Privacy risks in machine learning



Machine Learning is a threat to privacy even if one does not voluntarily participate in the model

Privacy risks before the ML era

Users' data

Services

Privacy Threats

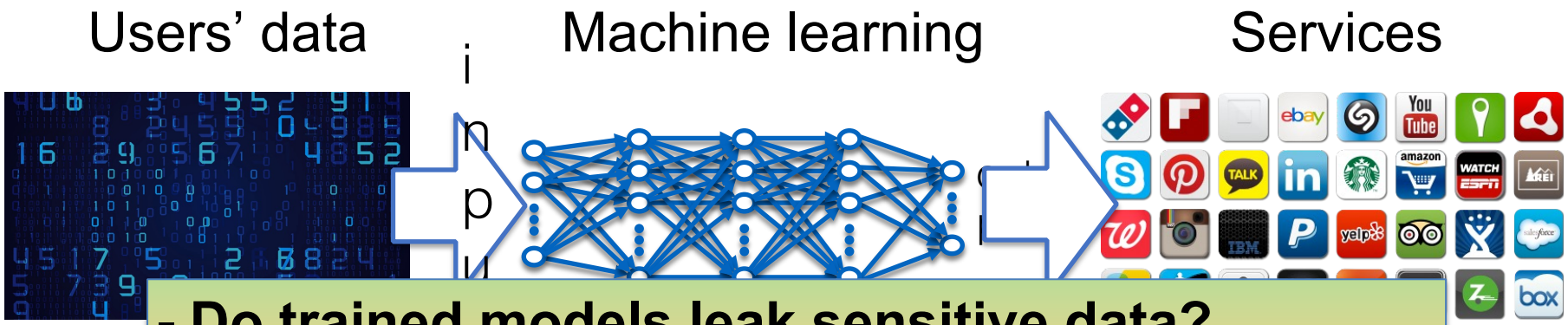
–Collection of sensitive personal data

–Re-identification/reconstruction

–Membership inference attacks

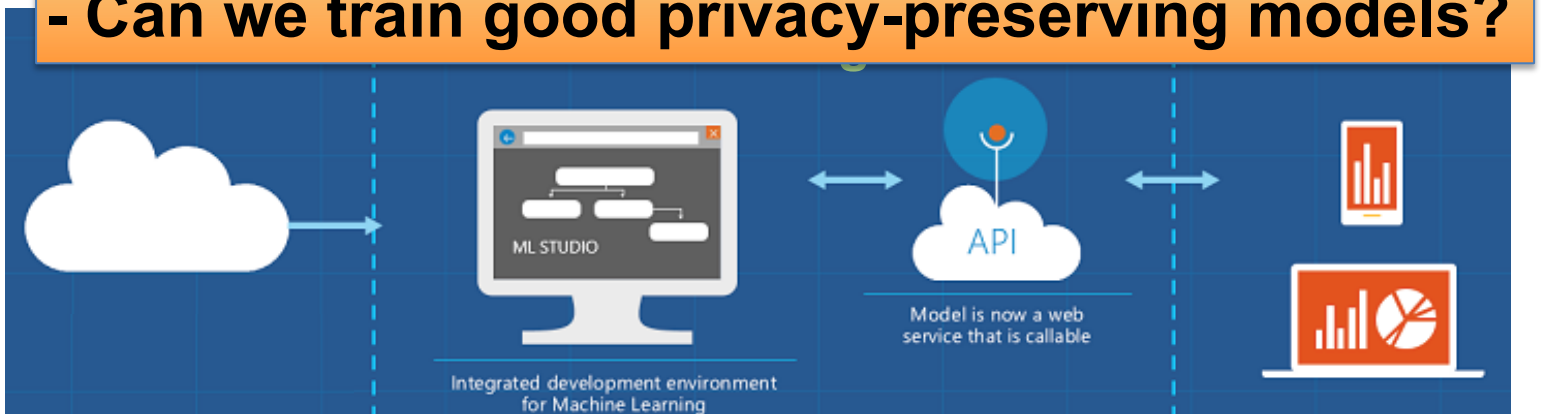
Shokri et al. 2017

Privacy risks in the ML era



- Do trained models leak sensitive data?

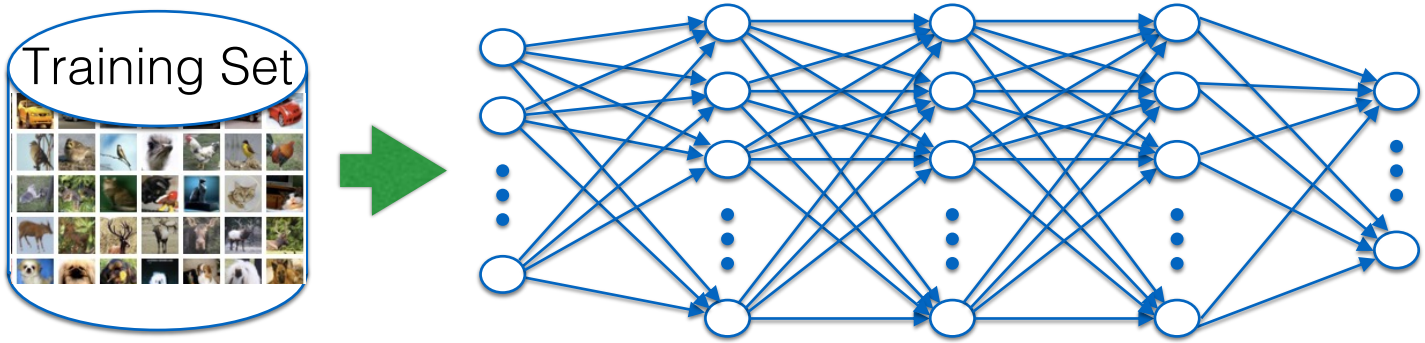
- Can we train good privacy-preserving models?



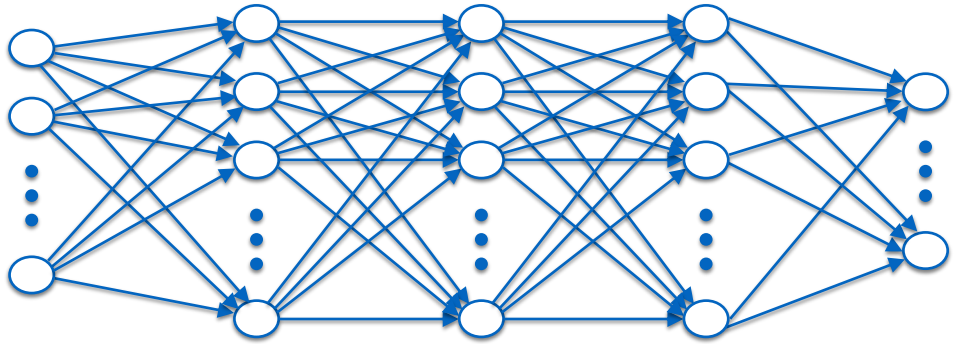
Google Cloud Platform amazon web services™ Microsoft Azure bigml®

Shokri et al. 2017

Typical Task: Classification



Query

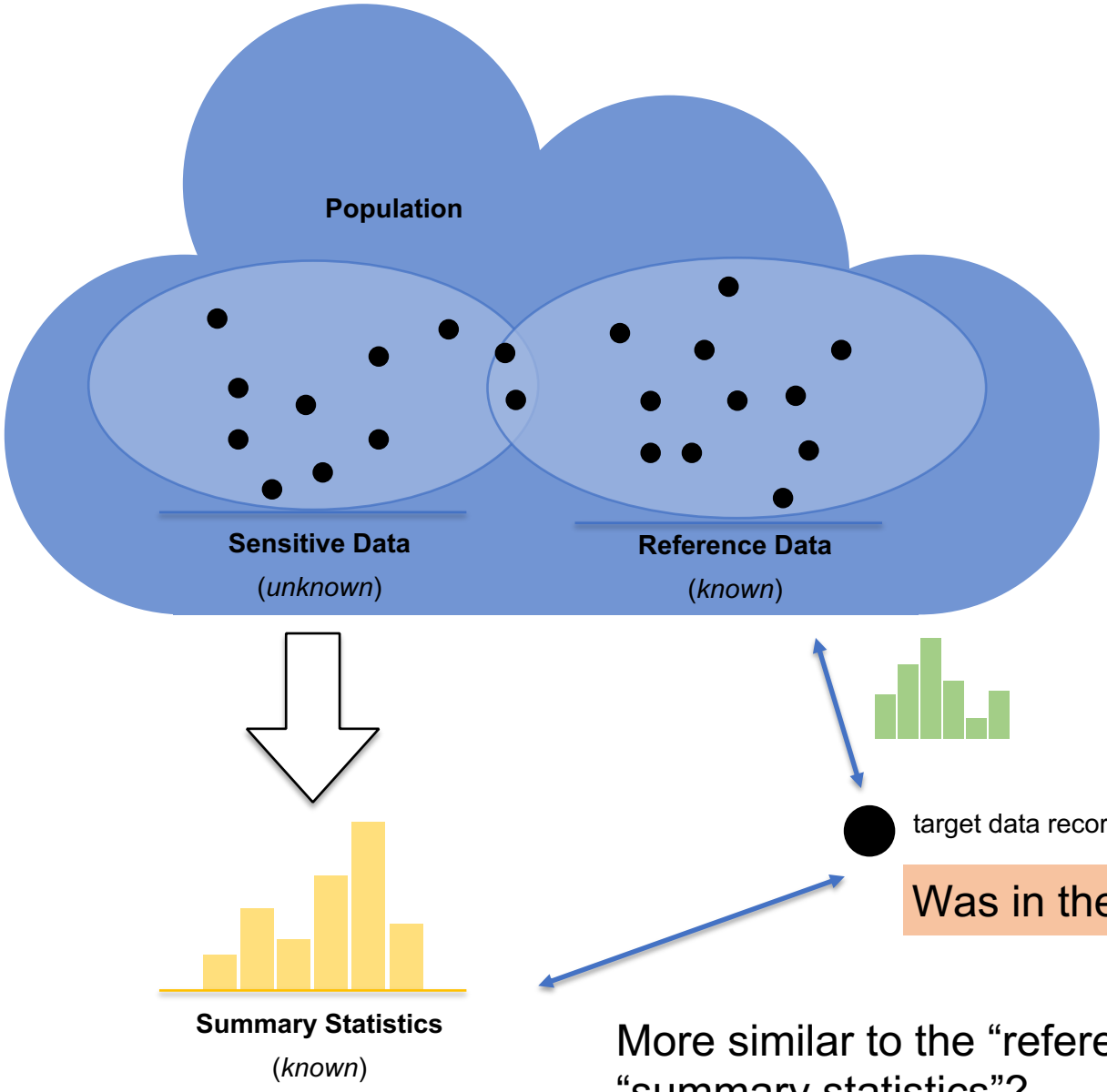


Prediction

airplane
automobile
...
ship
truck

Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

Membership Inference

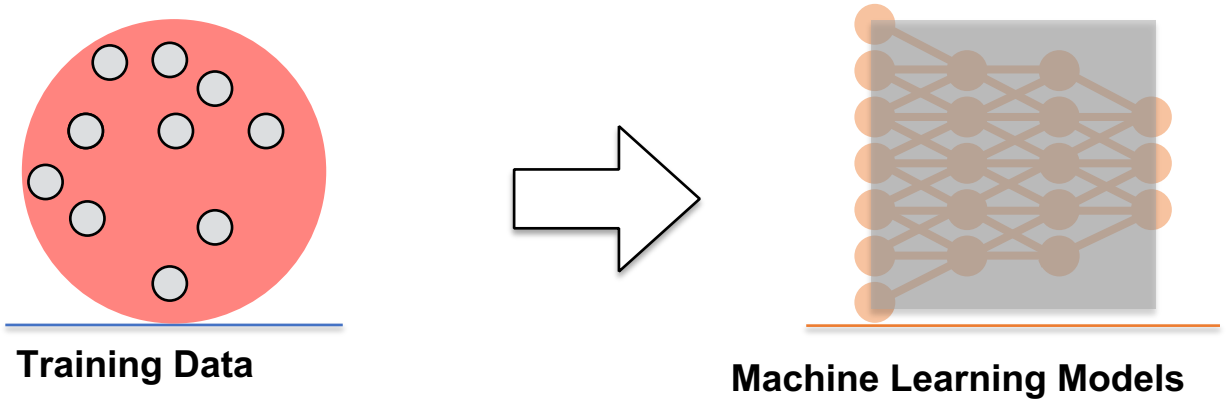


[Homer et al. (2008)]
[Dwork et al. (2015)]
[Backes et al. (2016)]

Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

Membership Inference (against ML)

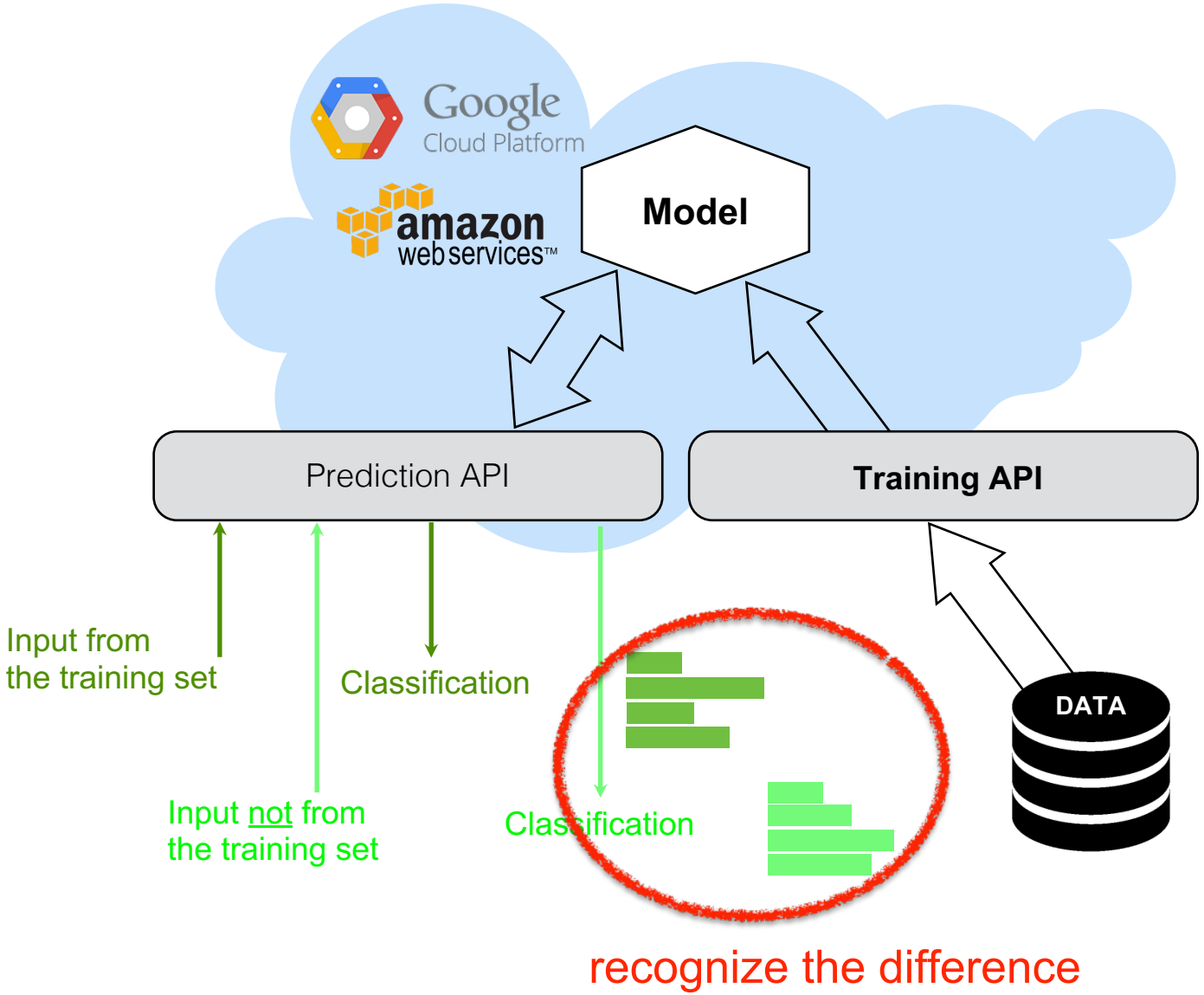
Assumptions



No knowledge of the model **parameters** (grey-box attack)

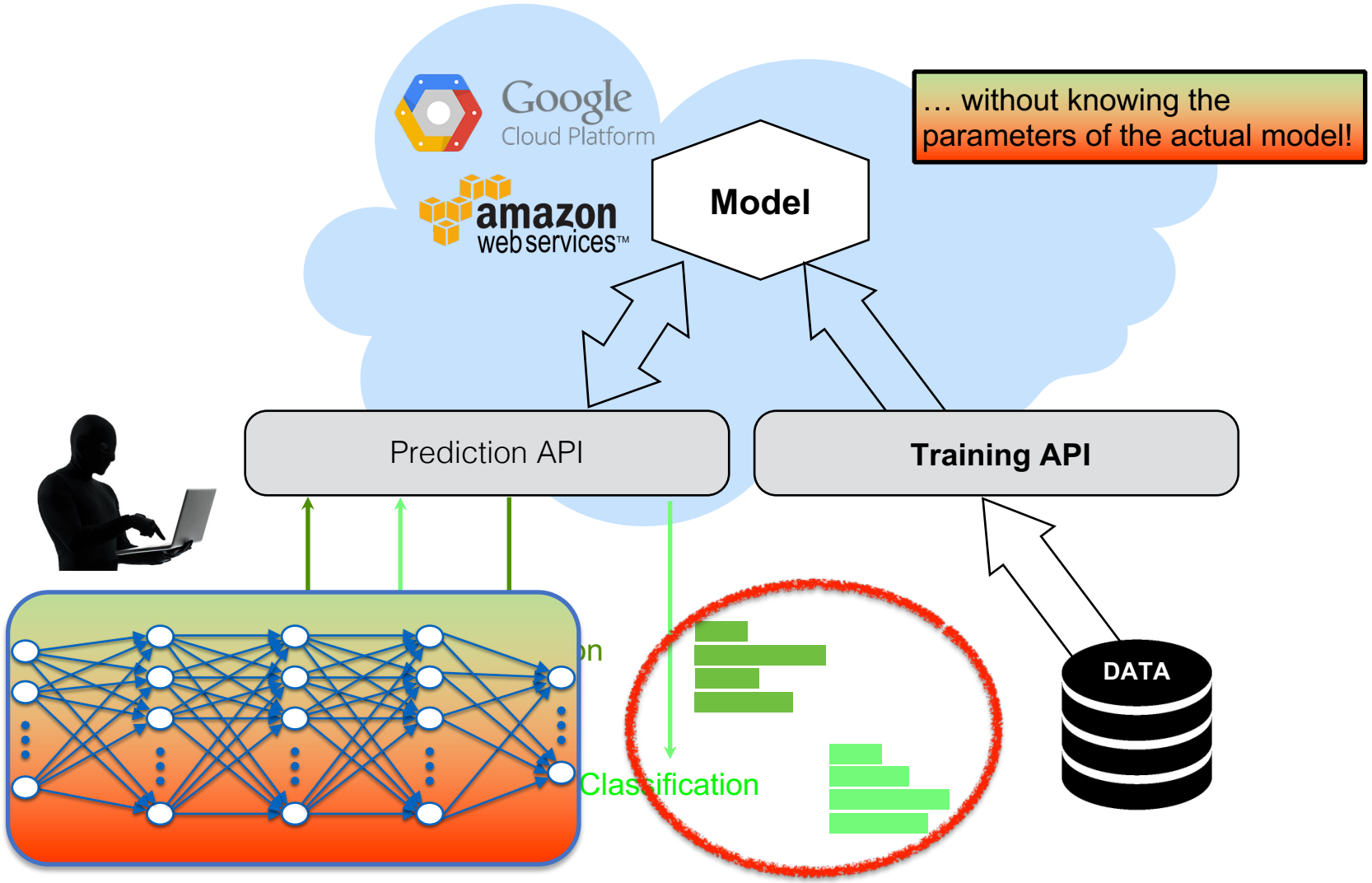
Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

Exploiting Trained Models



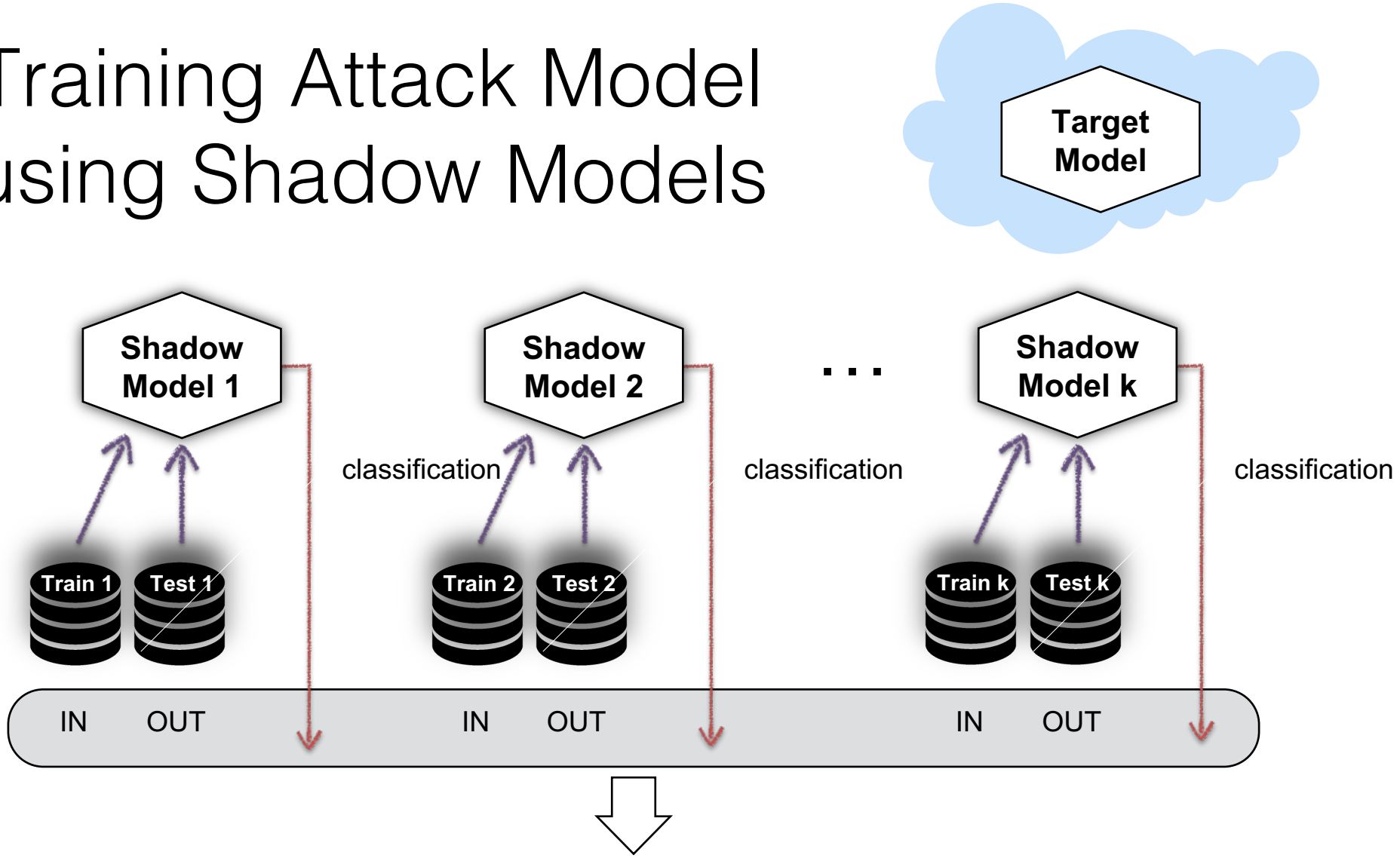
Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

ML Against ML



Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

Training Attack Model using Shadow Models

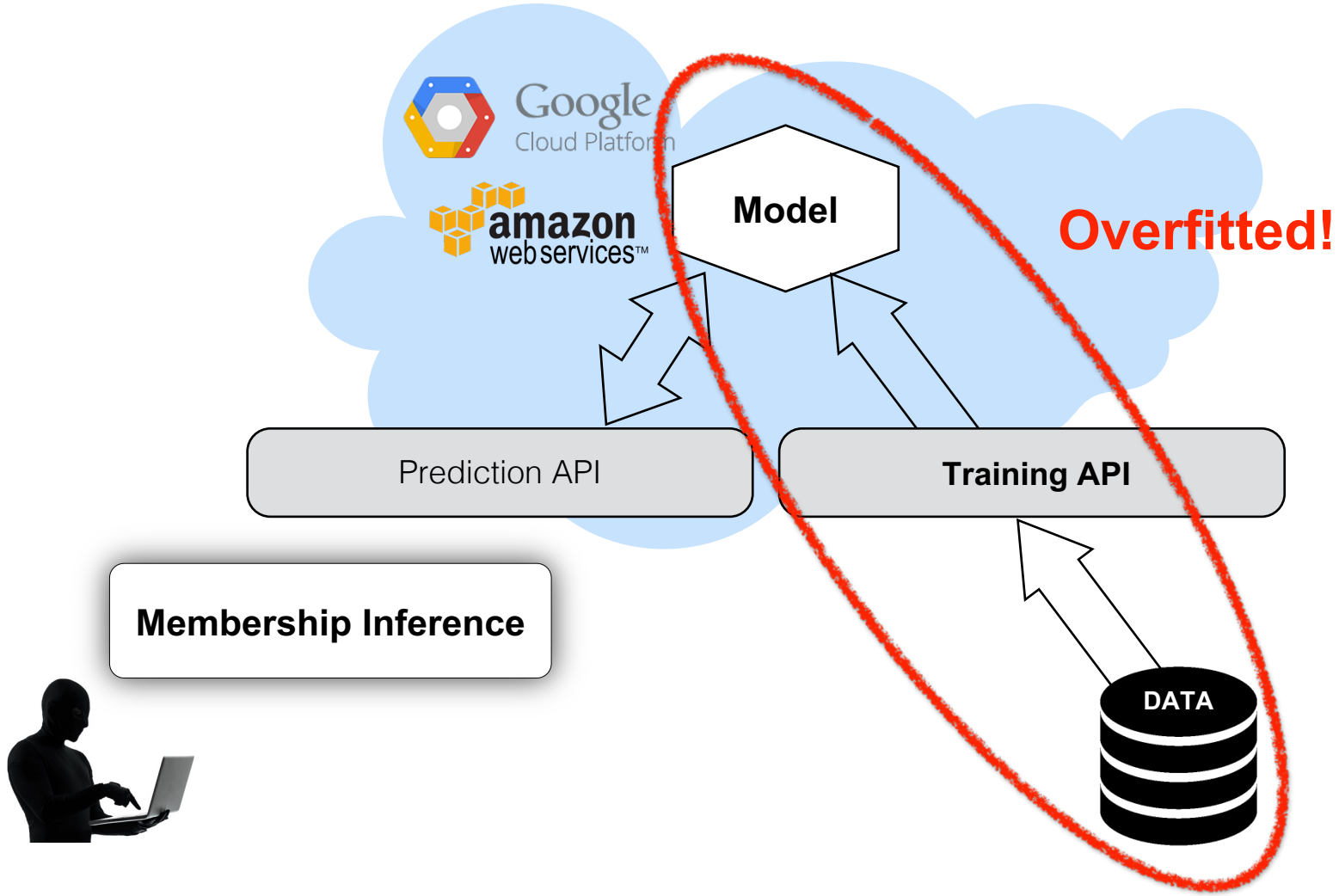


Train the attack model

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

Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

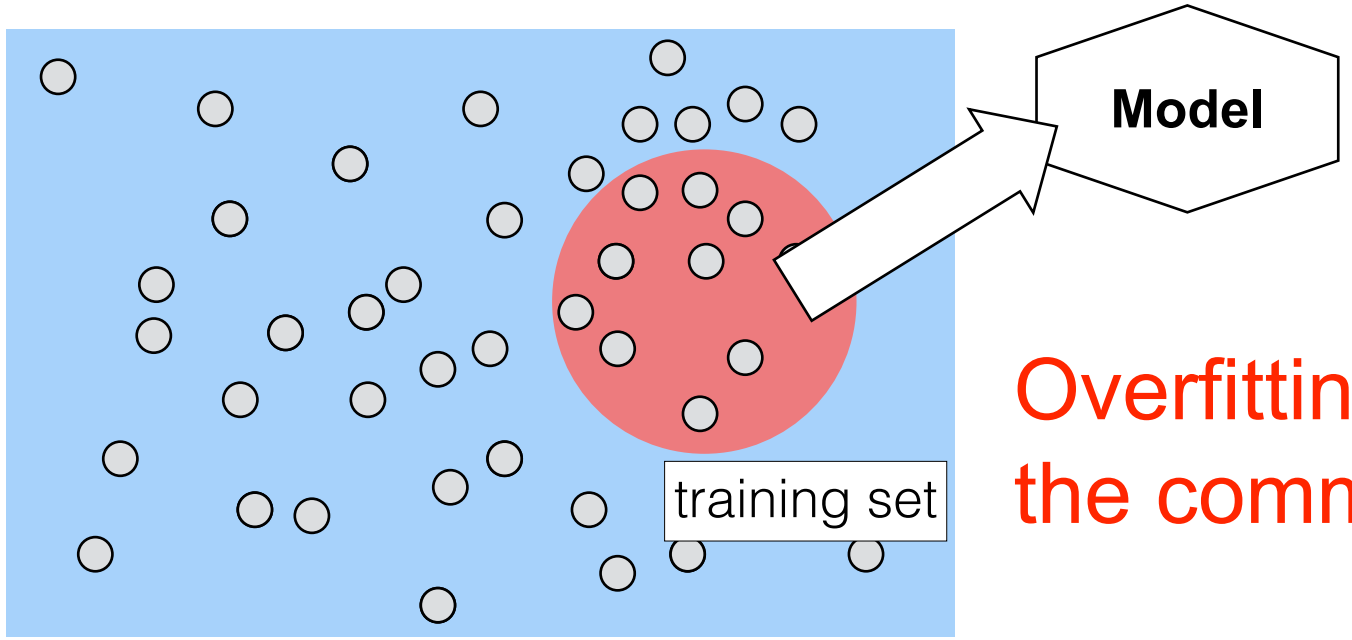
Why Do These Attacks Work?



Shokri et al., Membership Inference Attacks Against Machine Learning Models,2017

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

Learning:
Does the model generalize to data outside the training set?

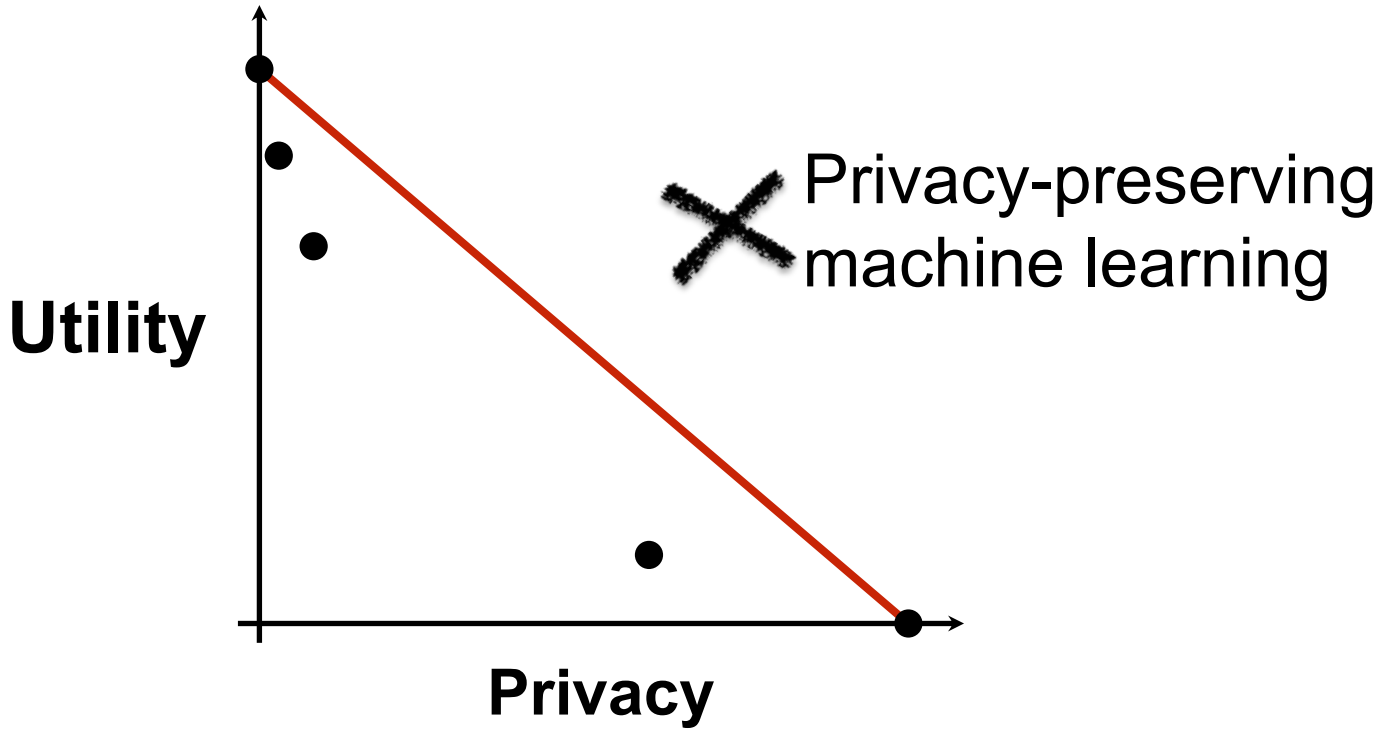


data universe

Overfitting is the common enemy

Shokri et al. 2017

- For once, privacy and utility are not in conflict: overfitting is the common enemy
  - Overfitted models leak training data
  - Overfitted models lack predictive power
- Need generalizability and accuracy



Shokri et al.
2017

Summary until now…

Privacy problem #1: ML needs data to learn! (training phase)

Machine learning is based on data to

Find features + Train the model

Data is highly unique! Allows many inferences

Anonymizing may not work…

Aggregation affects utility and requires careful evaluation

Hide data

Noise –> Differential privacy

Encryption –> Homomorphic encryption, secure multiparty computation,…

Privacy problem #2:
To obtain value, you must give data!
(testing phase)

To use the model you need to provide a sample...

If the model is remote (ML as a service) you are giving this sample out!

We can do **privacy-preserving model evaluation**:

predict on noisy data: utility hit

use advanced cryptography: performance hit

Privacy problem #3:
The output reveals information!
(testing phase)

Membership inference

Given the answer of the classifier, infer whether the queried example was used in training.

Attribute inference/reconstructions

Given the answer of the classifier, infer whether a training sample had a particular attribute

Privacy problem #4:
Machine learning is **VERY** good at inferring

Deploy the program!: Use the learnings to classify/predict on new data

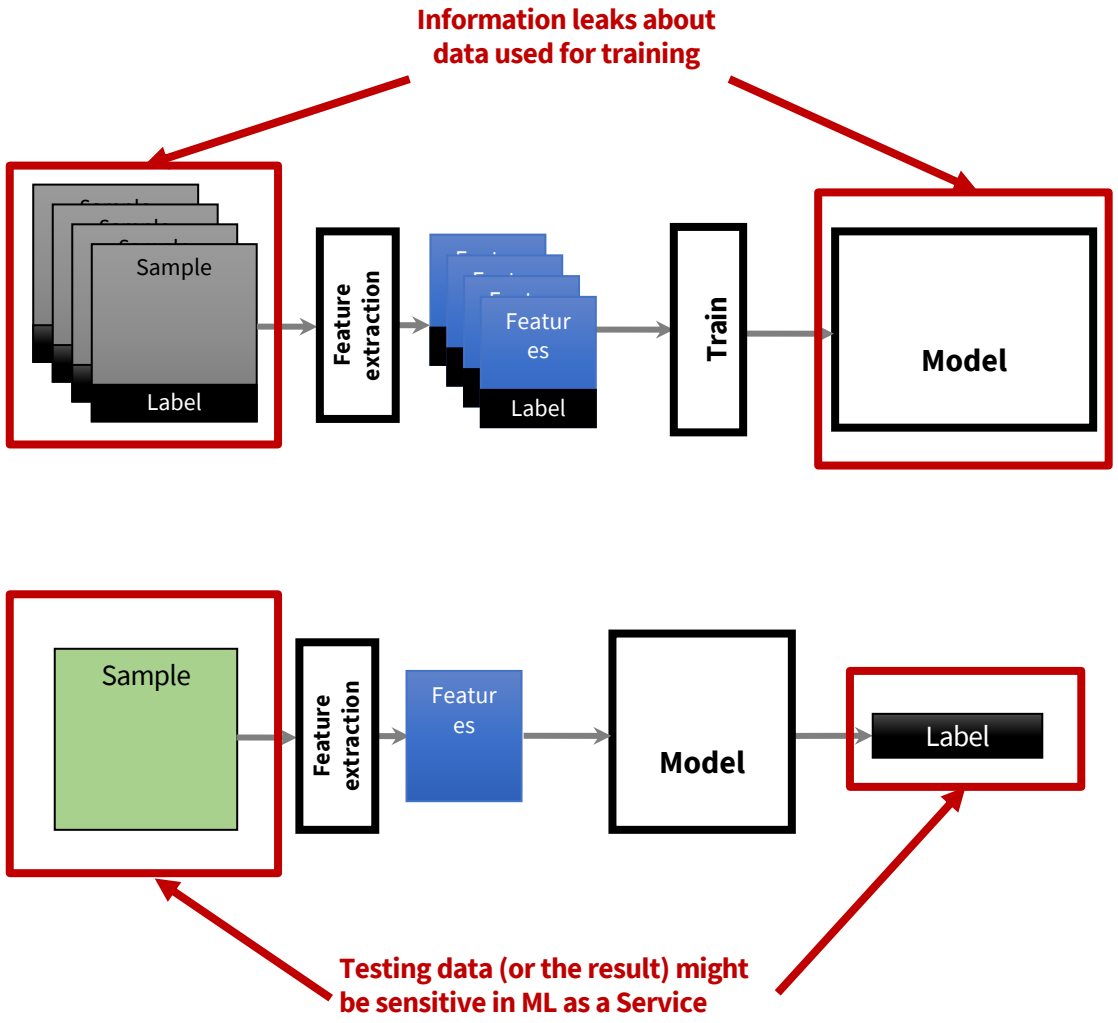
The ML model can be used to breach privacy of that new data (or associated entities)

Privacy problem #4:
Machine learning is **VERY** good at inferring

Deploy the program!: Use the learnings to classify/predict on new data

The ML model can be used to breach privacy of that new data (or associated entities)

Takeaways on ML Privacy issues



Many problems, some solutions ☺️

Very young field, the situation will improve!

Are you willing to help?