Trusting Machine Learning: **Privacy**, Robustness, and Interpretability Challenges

Reza Shokri

NUS | Computing
National University of Singapore

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

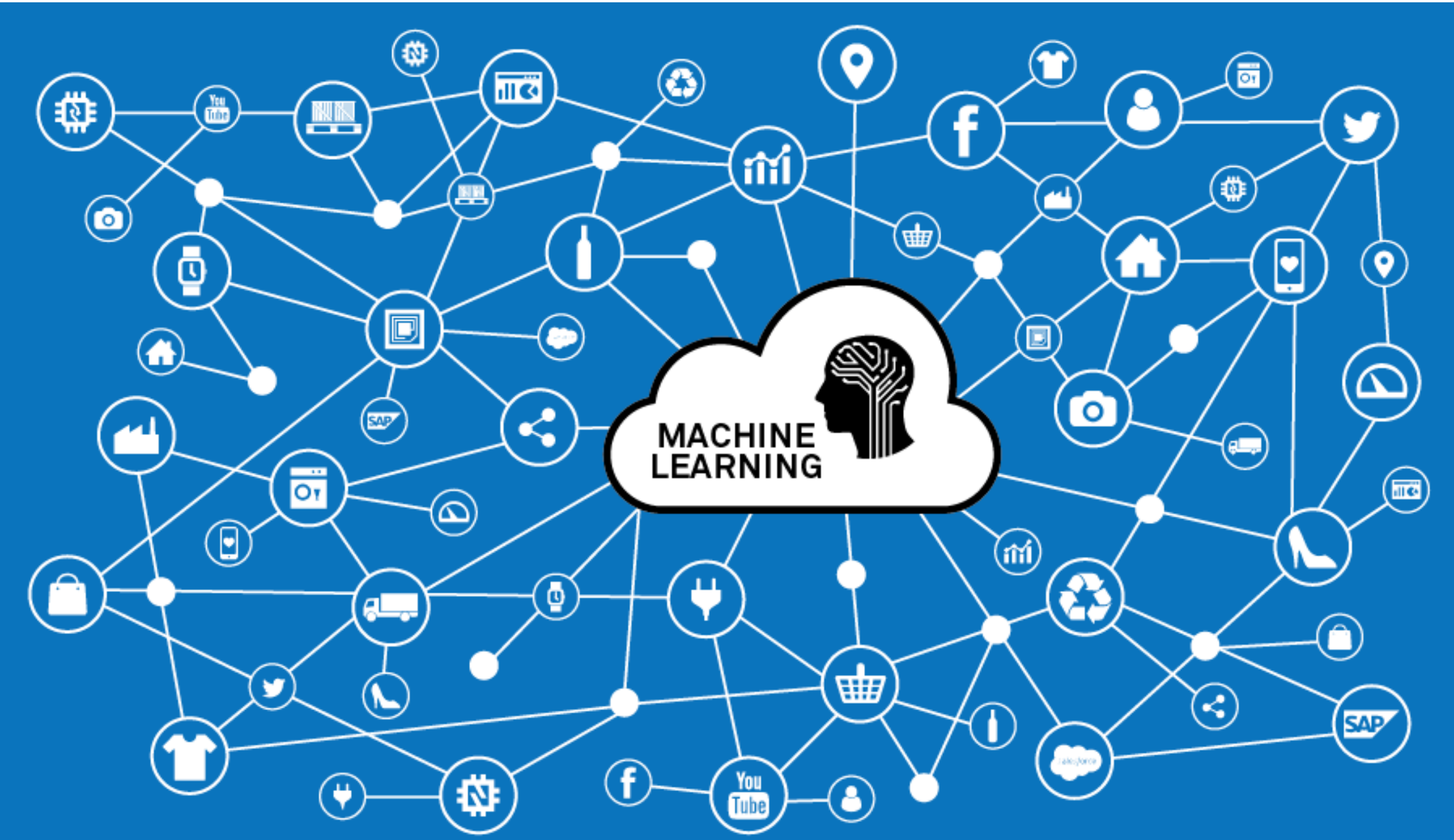
Internet

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

Web trackers

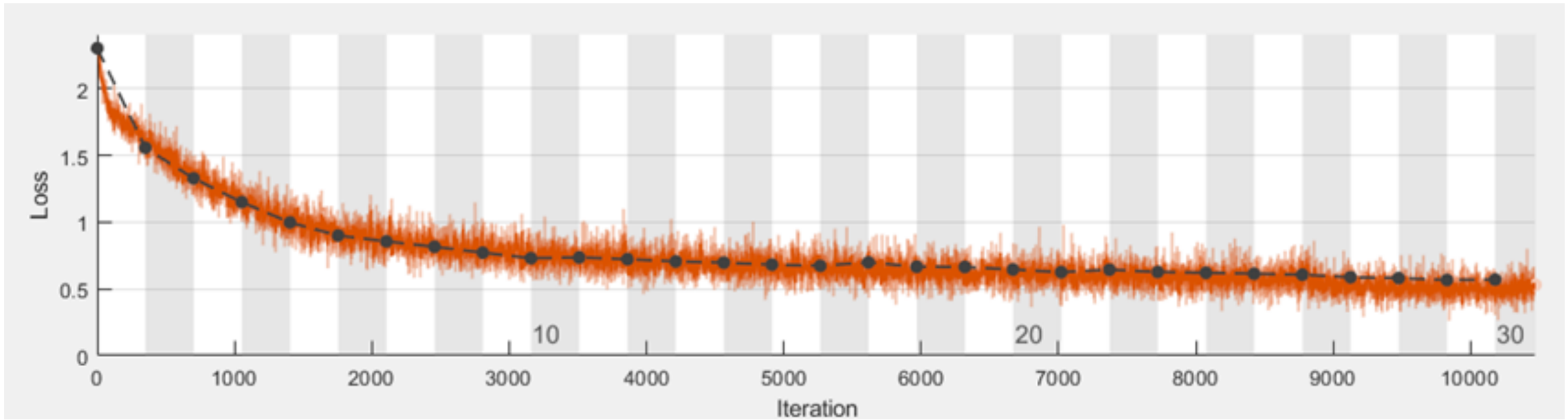
Data aggregators

Internet



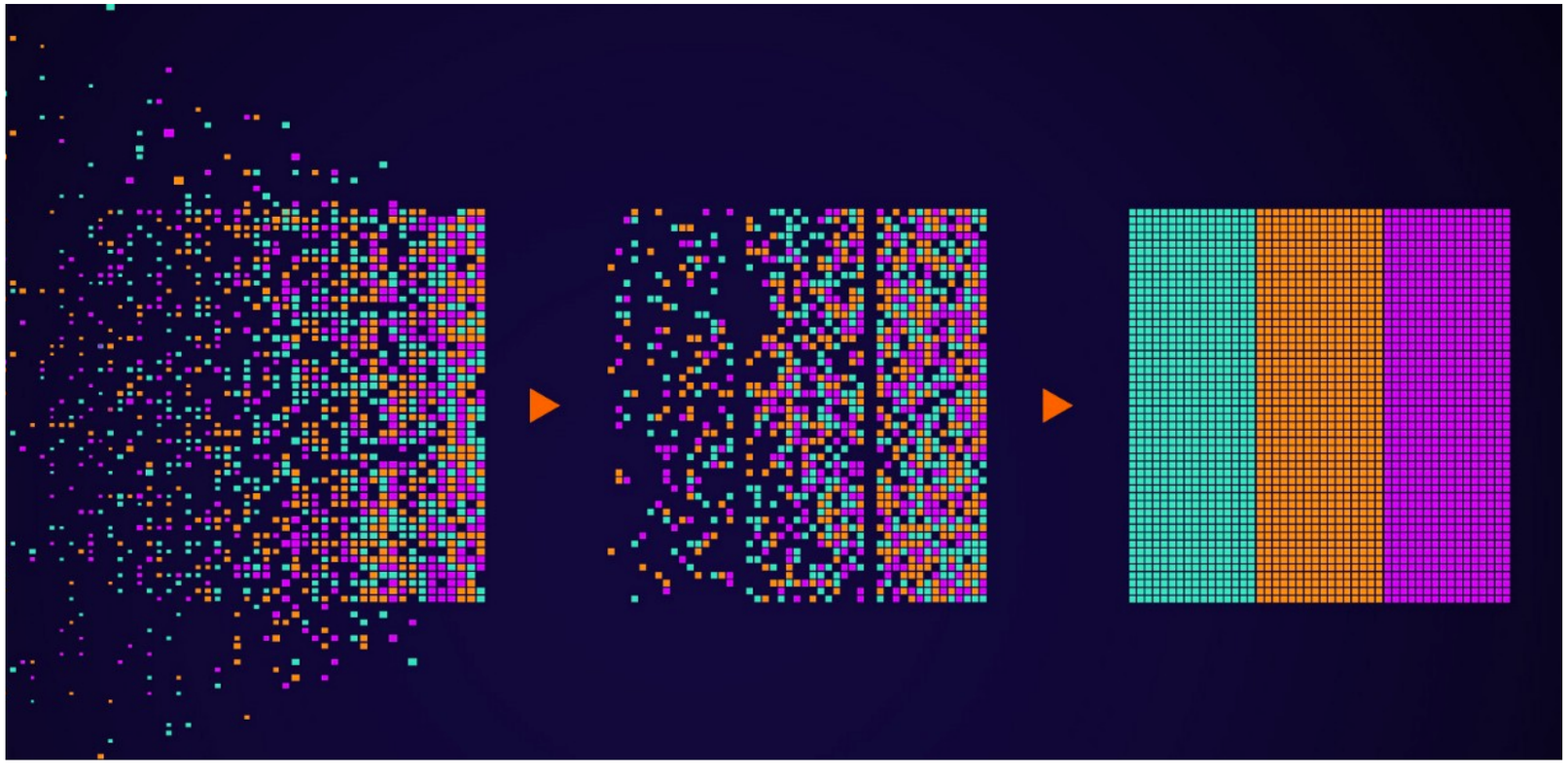
Machine Learning

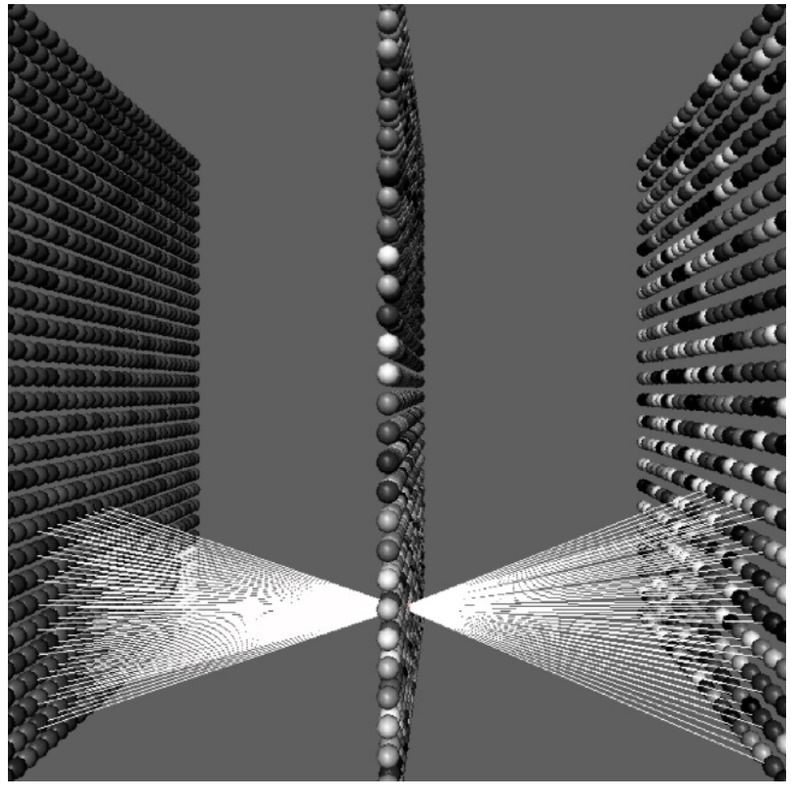
- Minimize the learning loss
- Maximize the predictive power



Machine Learning

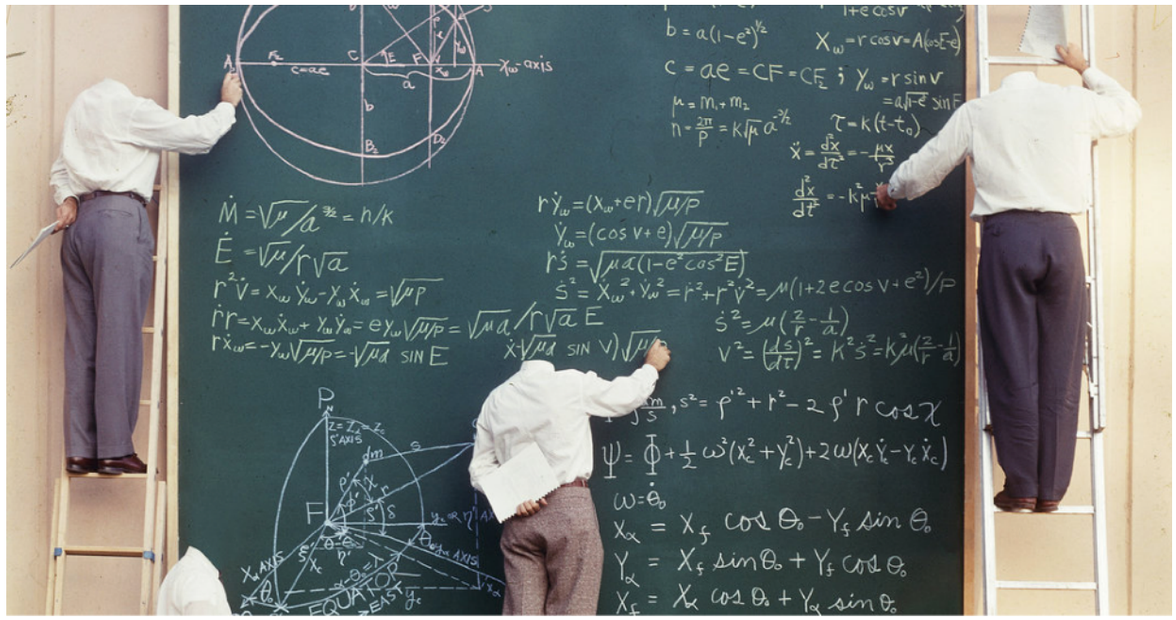
- Massive amount of data
- Large models

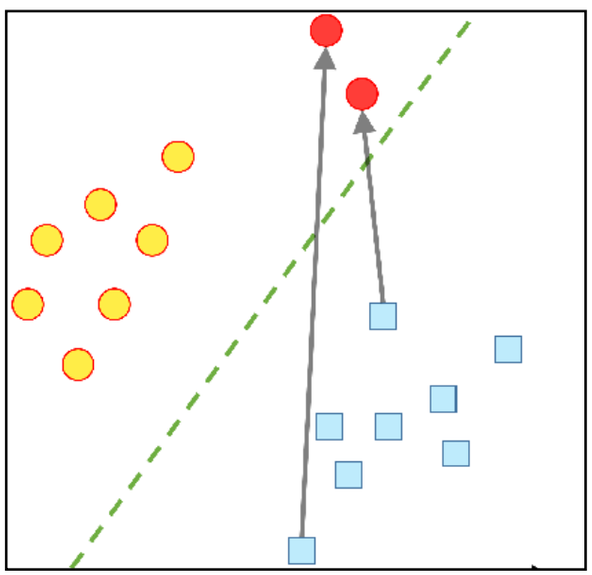




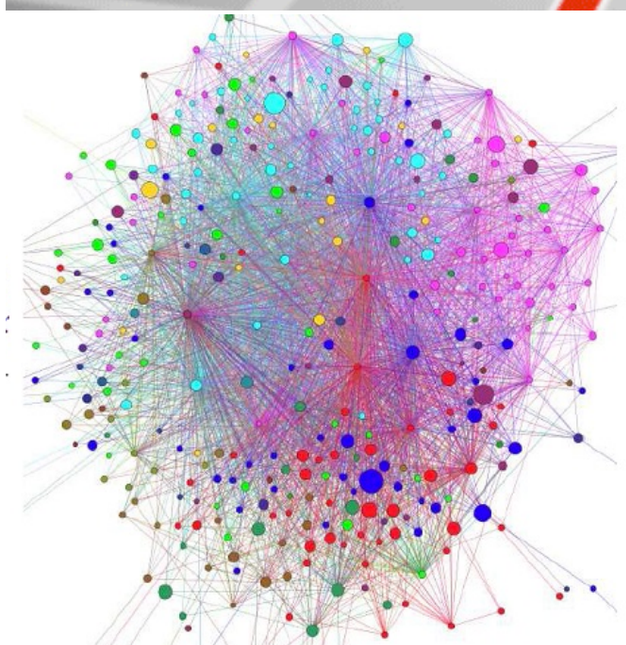
Beyond Prediction Accuracy

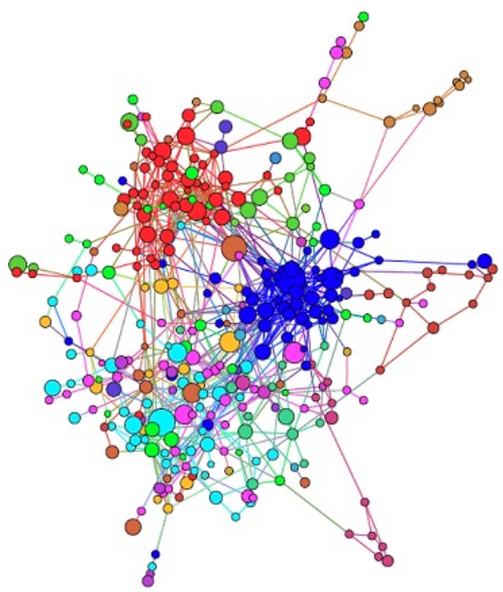
- Privacy
- Robustness
- Transparency
- Explainability
- Fairness

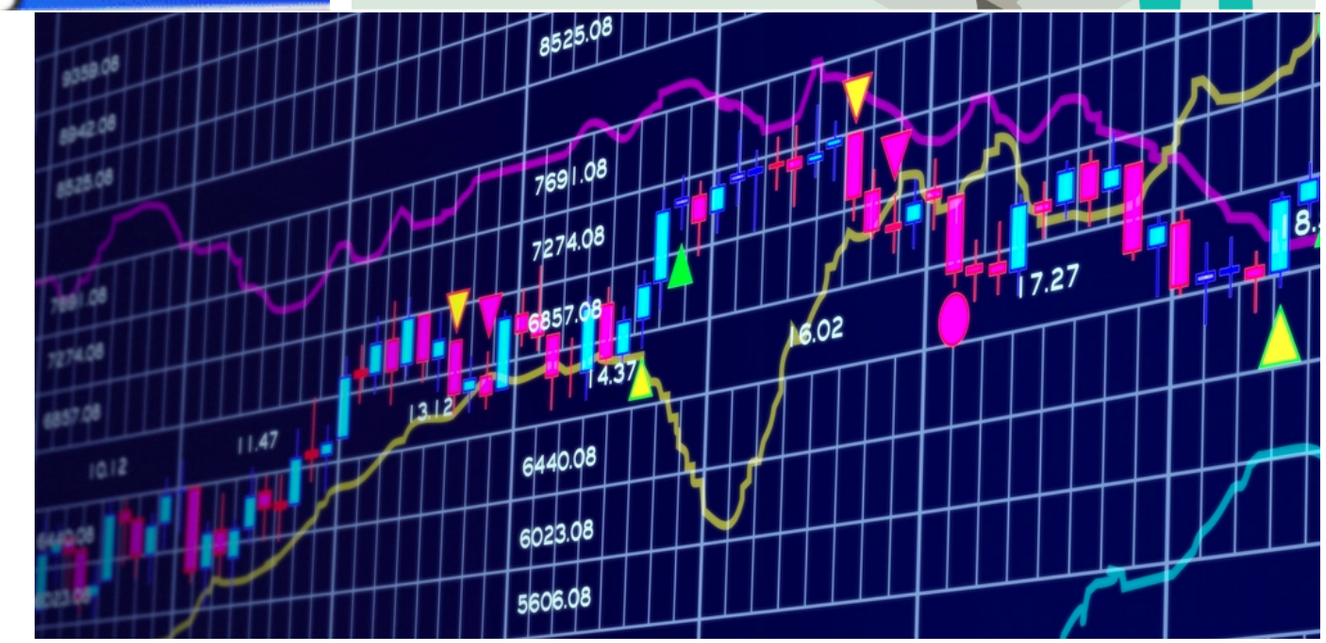




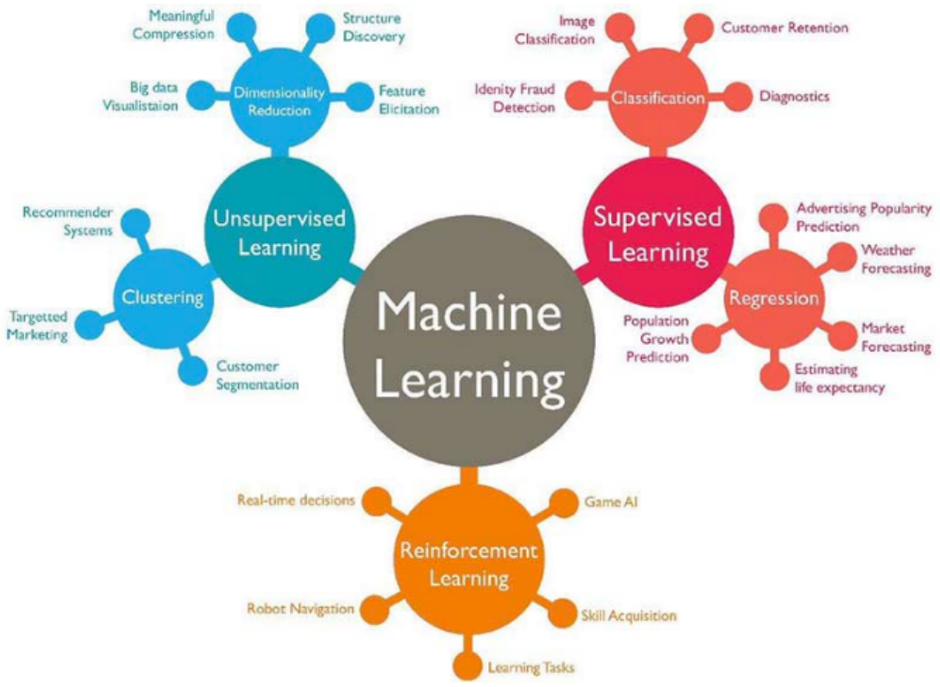
Components - Data

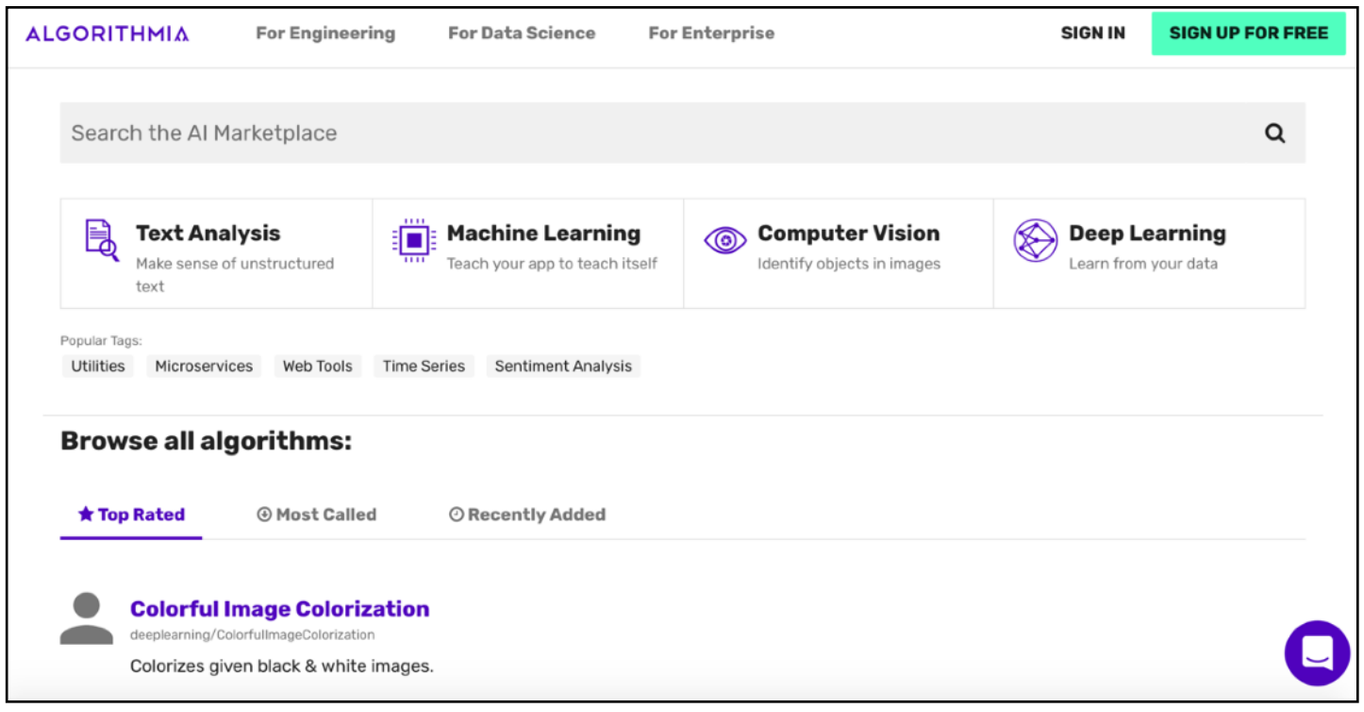


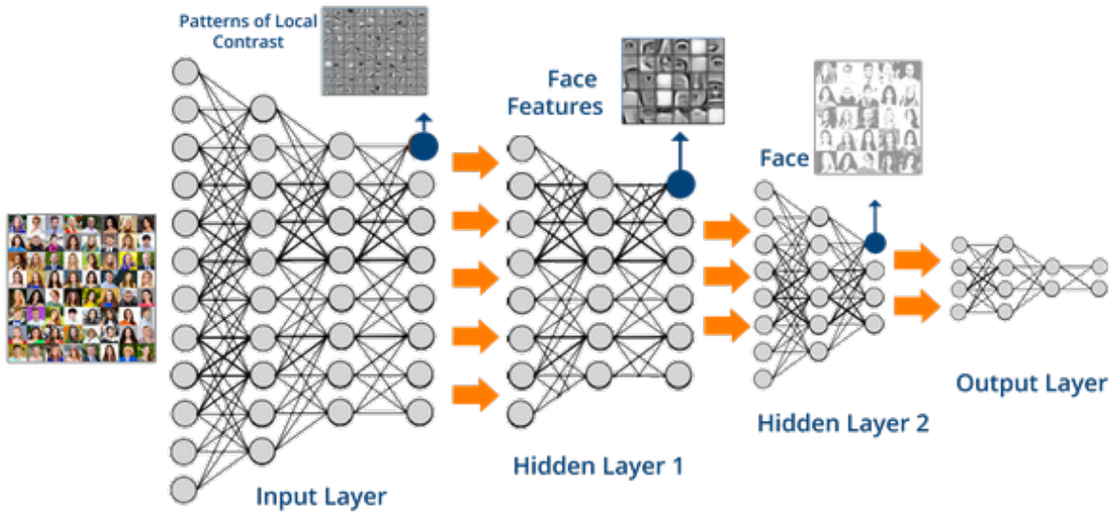


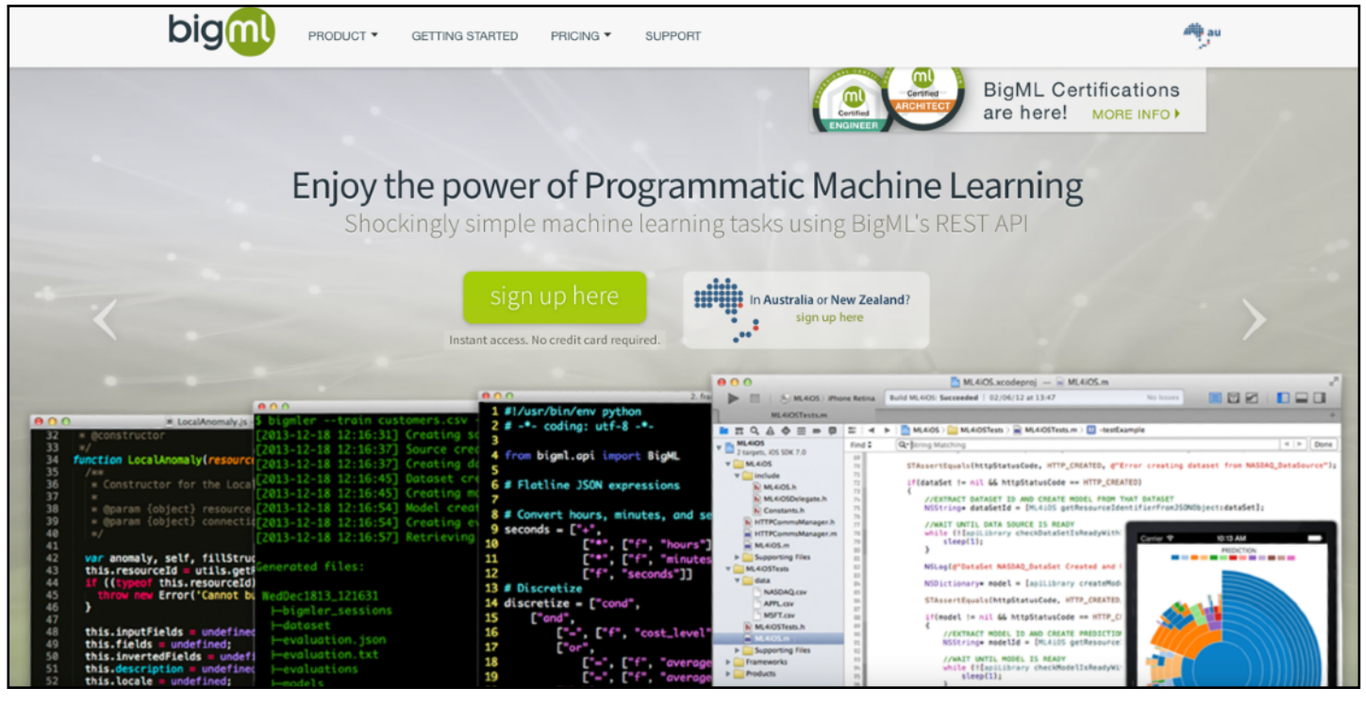


Components - Algorithms

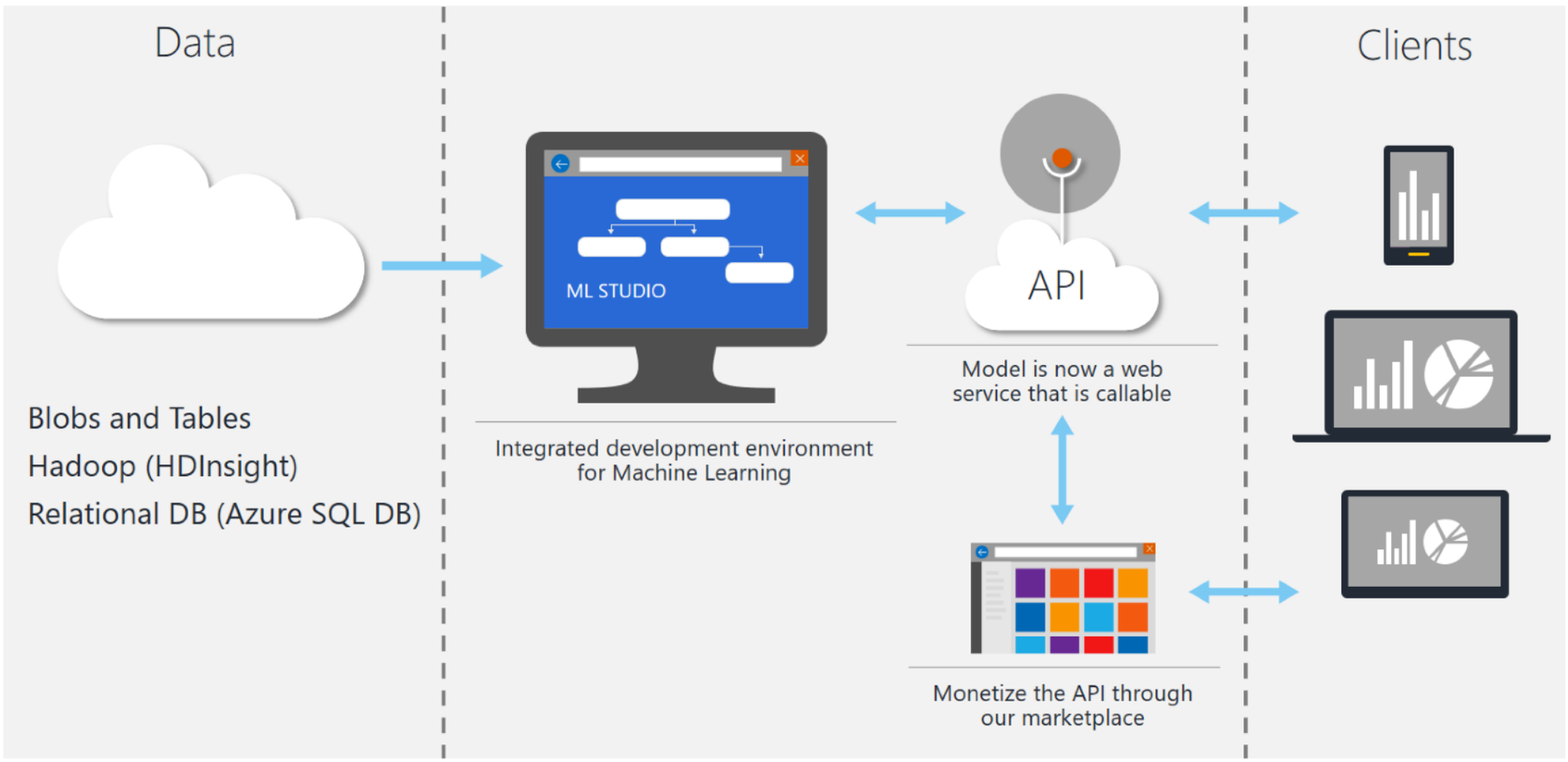




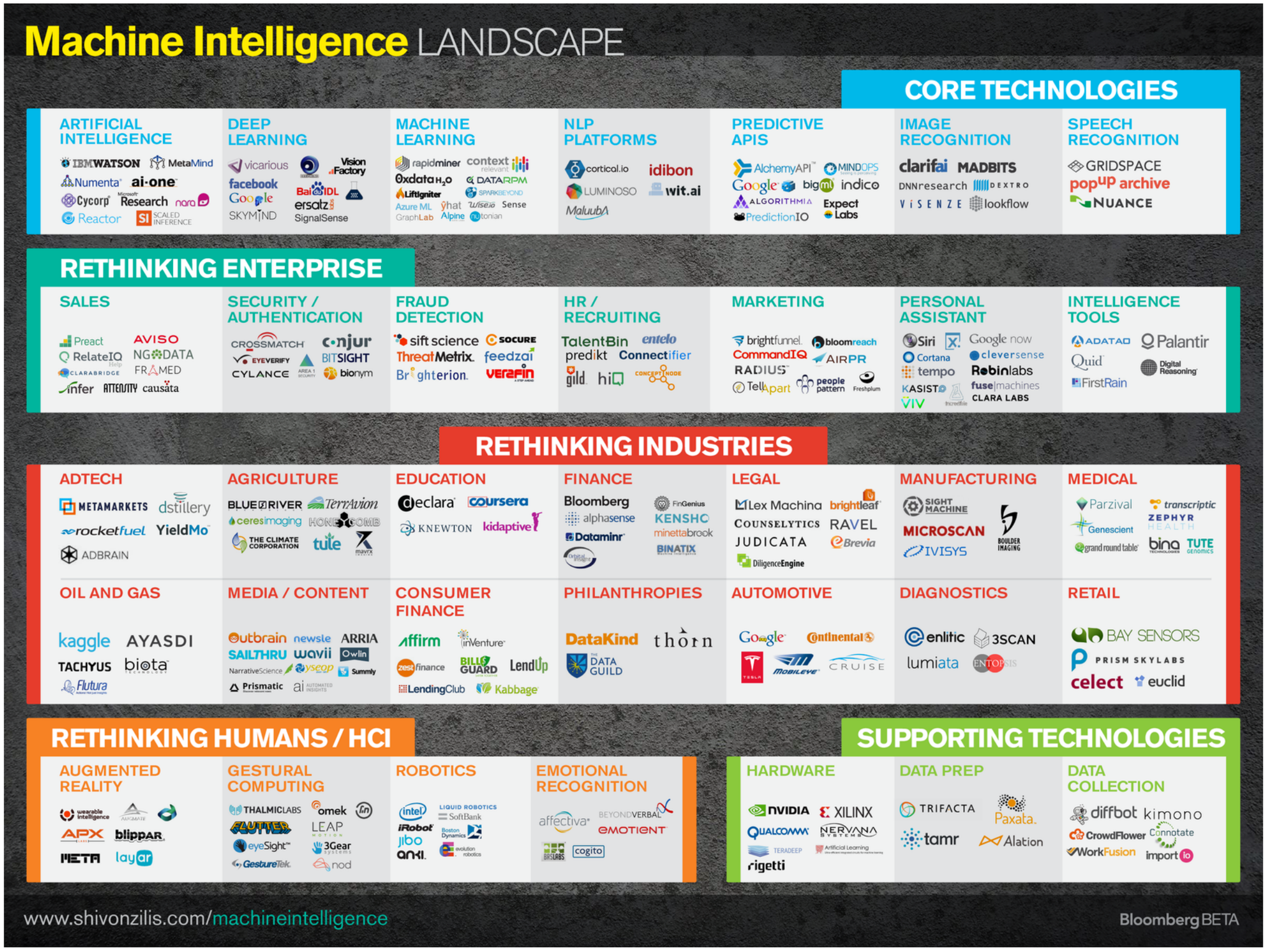




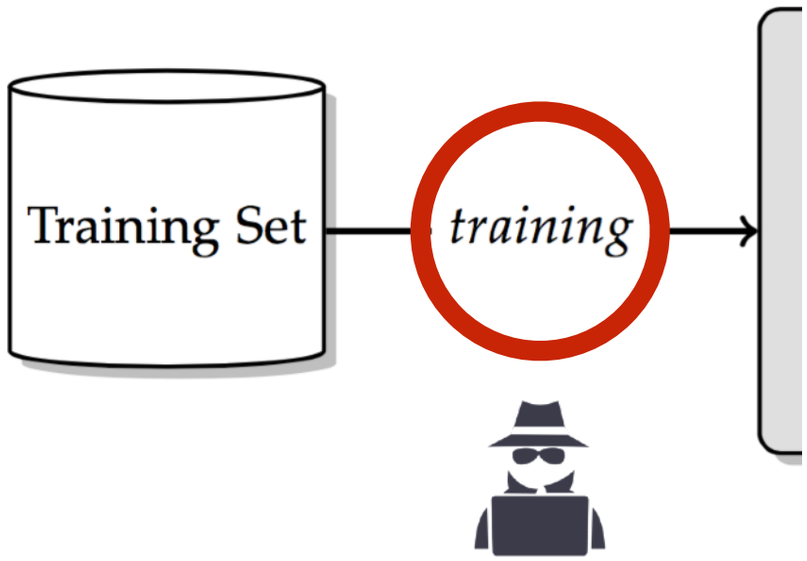
Components - Computing Platforms



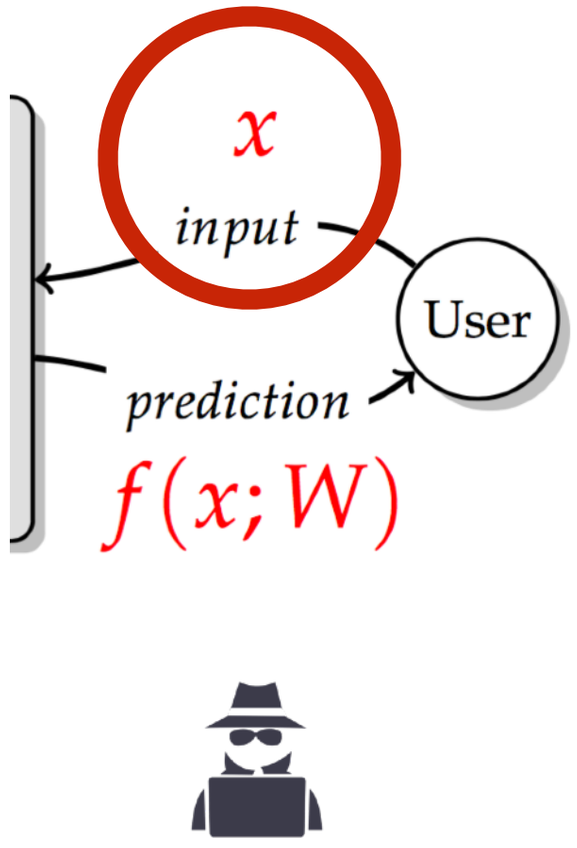
Components - Users



Privacy Risks

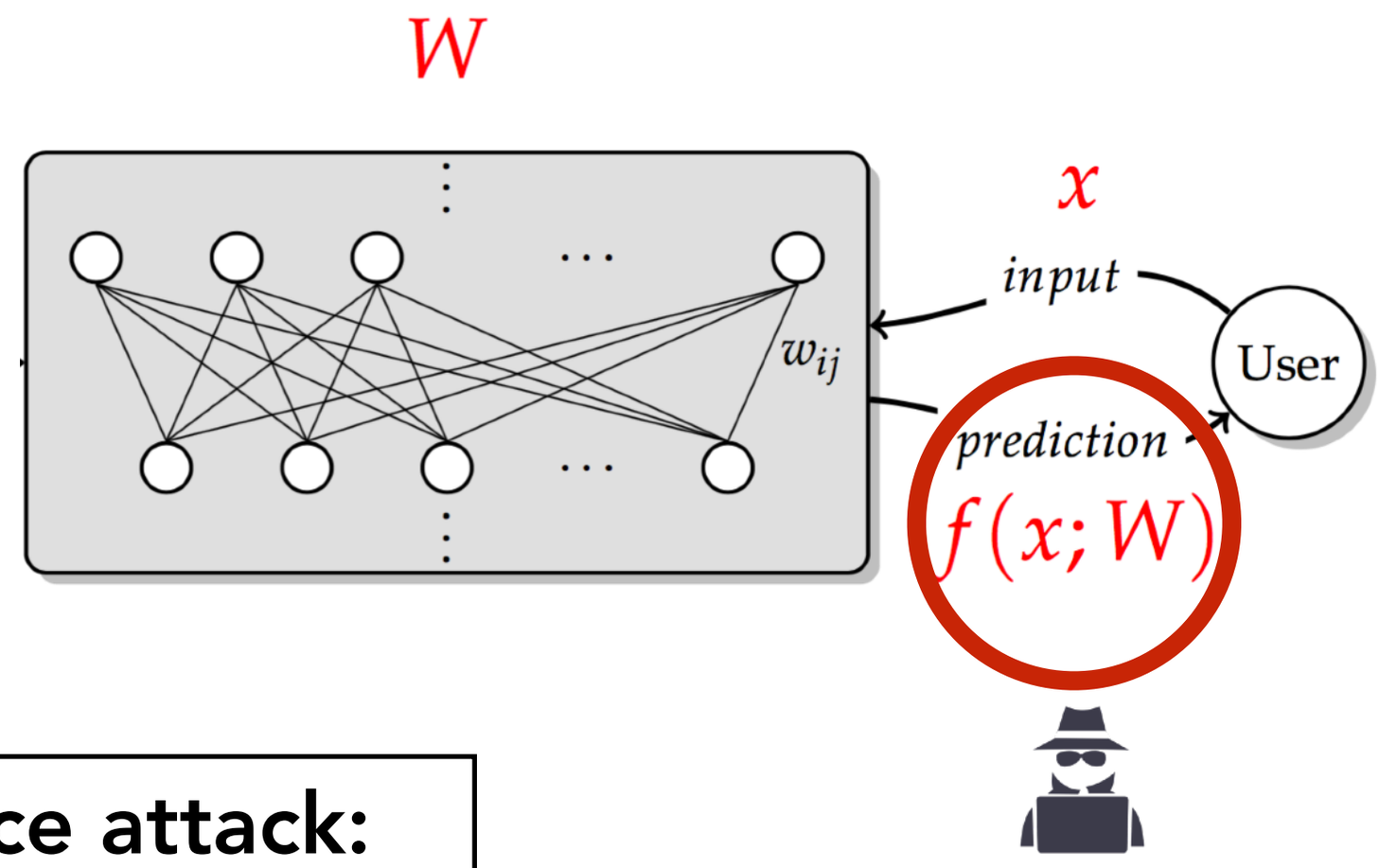


Direct access to training data set



Access to sensitive query inputs

Privacy Risks

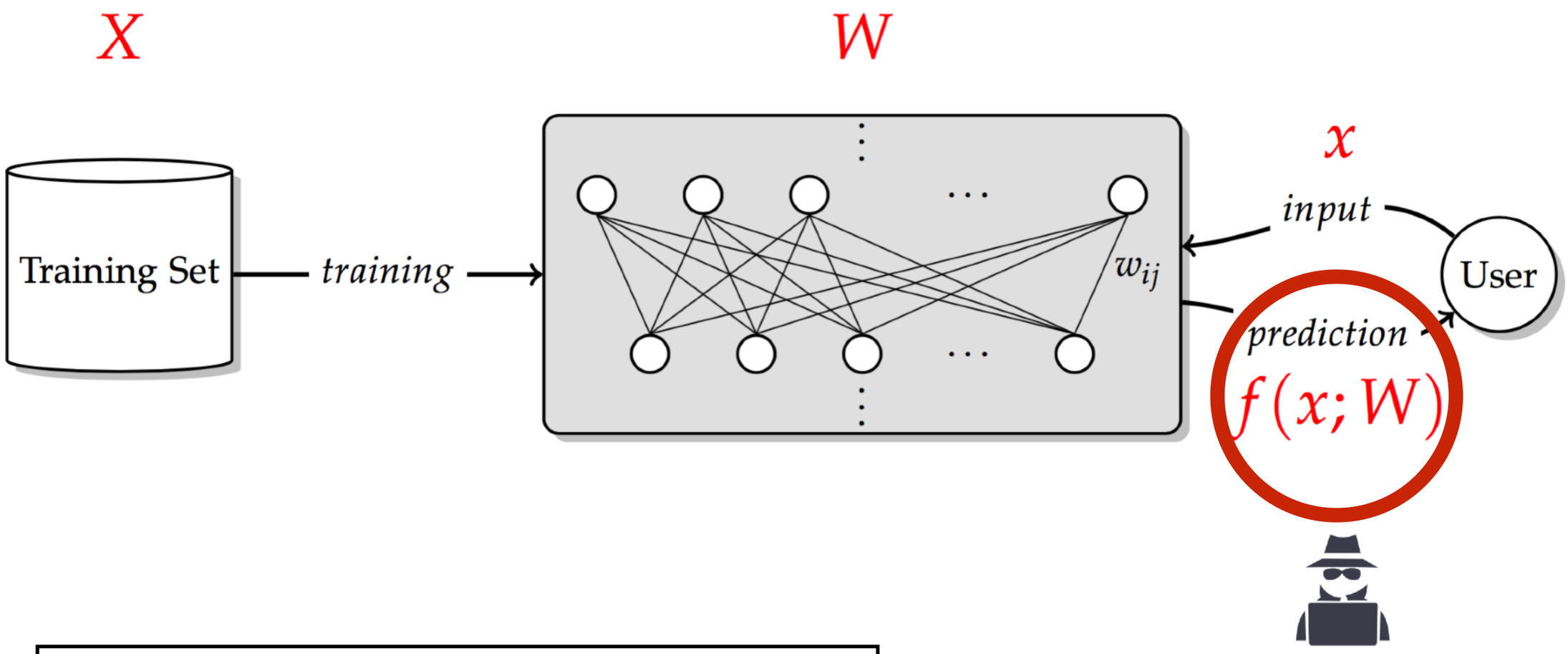


Input inference attack:
Given f(x;W), infer input x

Input/Data/Model Inference

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

Privacy Risks

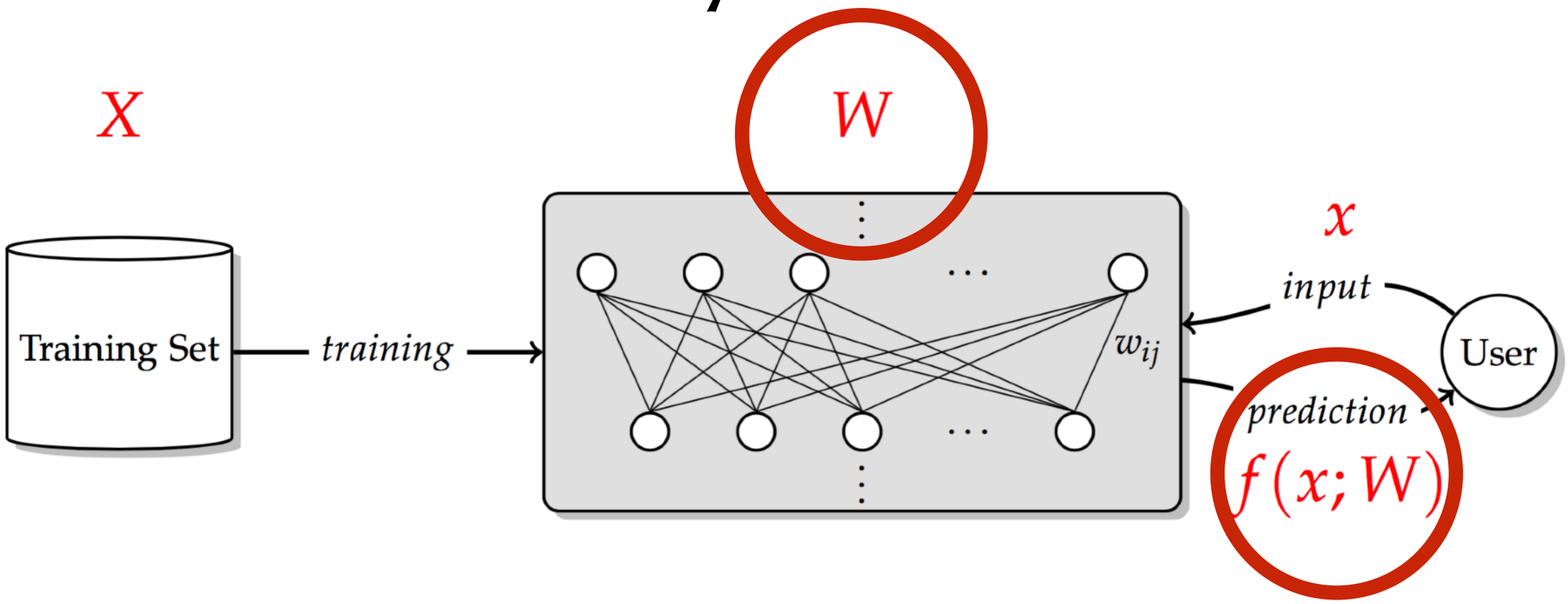


Model extraction attack:
Given f(x;W), infer model W

Input/Data/Model Inference

Tramèr, et al. "Stealing Machine Learning Models via Prediction APIs." USENIX Security 2016.

Privacy Risks

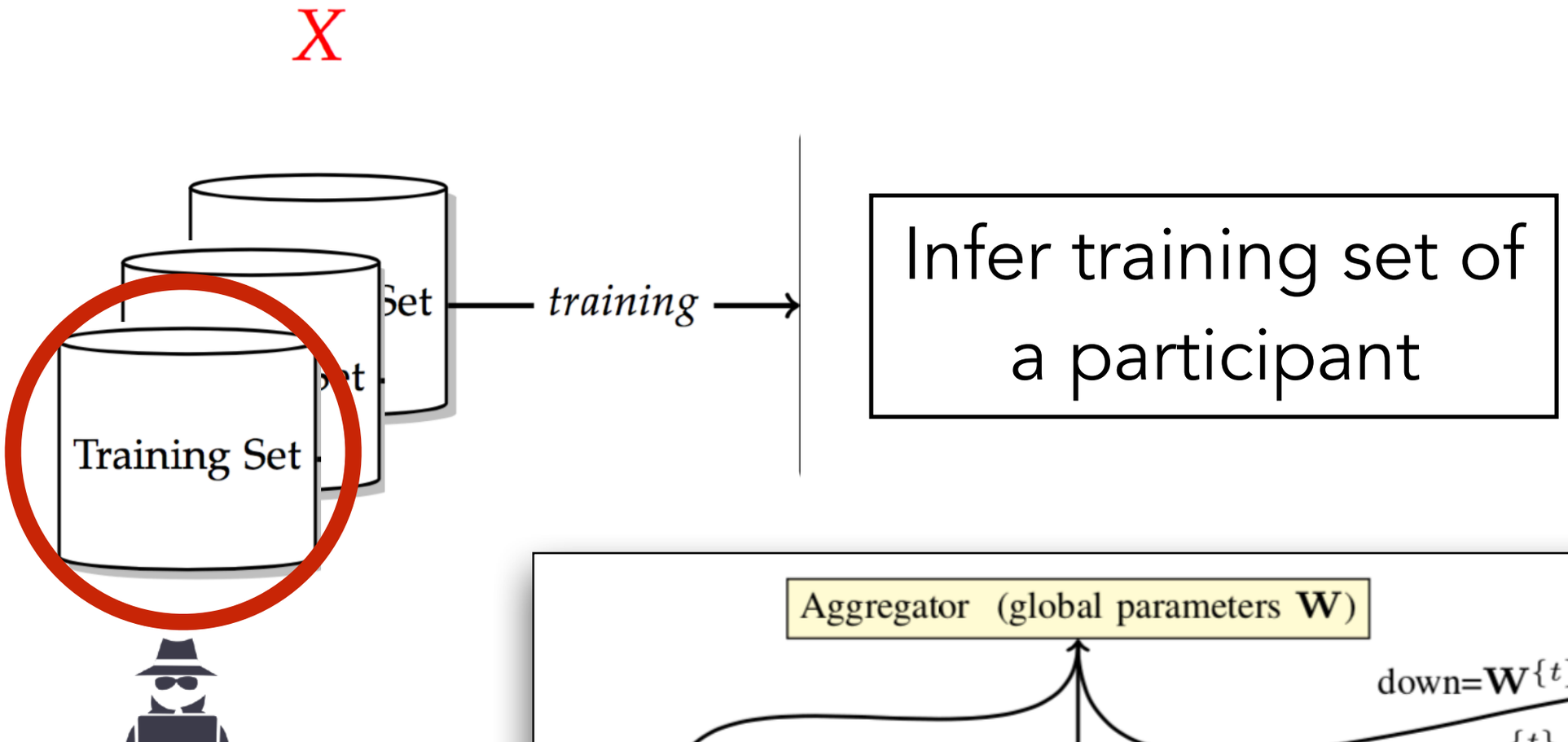


Reconstruction and Membership Inference attacks:
Given W or f(x;W), infer about the training set X

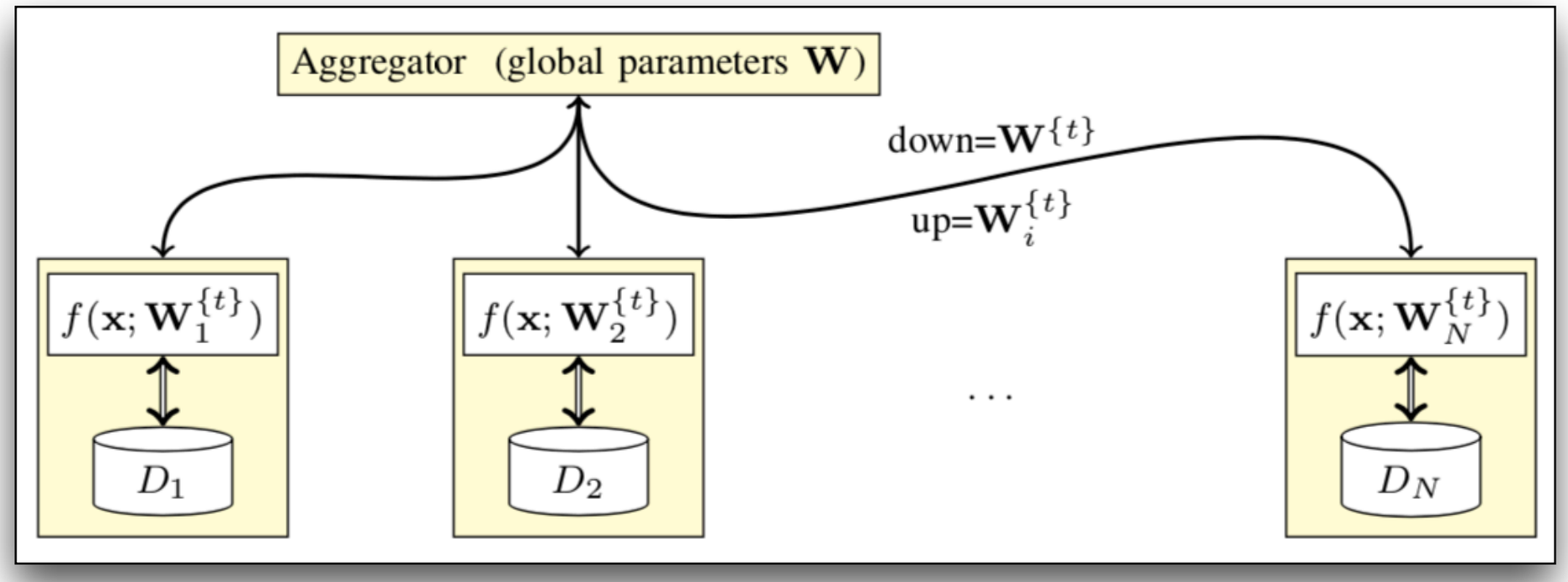
Input/Data/Model Inference

Shokri, et al. "Membership inference attacks against machine learning models." IEEE S&P 2017
Nasr, Shokri, Houmansadr, “Comprehensive Analysis of Deep Learning” IEEE S&P 2019

Privacy Risks



Data Inference in Distributed Training



Shokri, Shmatikov. “Privacy preserving deep earning." ACM CCS 2015
Nasr, Shokri, Houmansadr, “Comprehensive Analysis of Deep Learning” IEEE S&P 2019

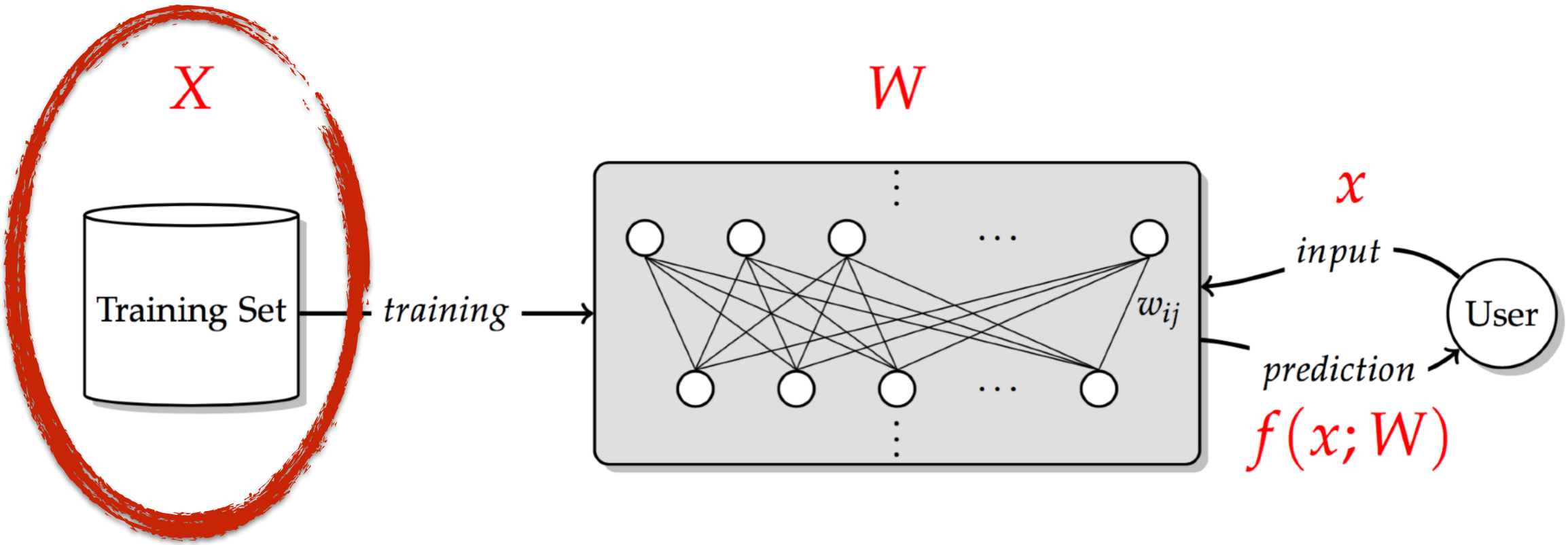
Privacy in Machine Learning

Shokri, et al. "Membership inference attacks against machine learning models." IEEE S&P 2017

Nasr, Shokri, Houmansadr, “Comprehensive Analysis of Deep Learning” IEEE S&P 2019

Nasr Milad, et al. "Machine Learning with Membership Privacy using Adversarial Regularization." CCS 2018

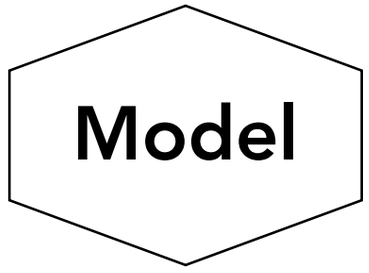
Data Privacy



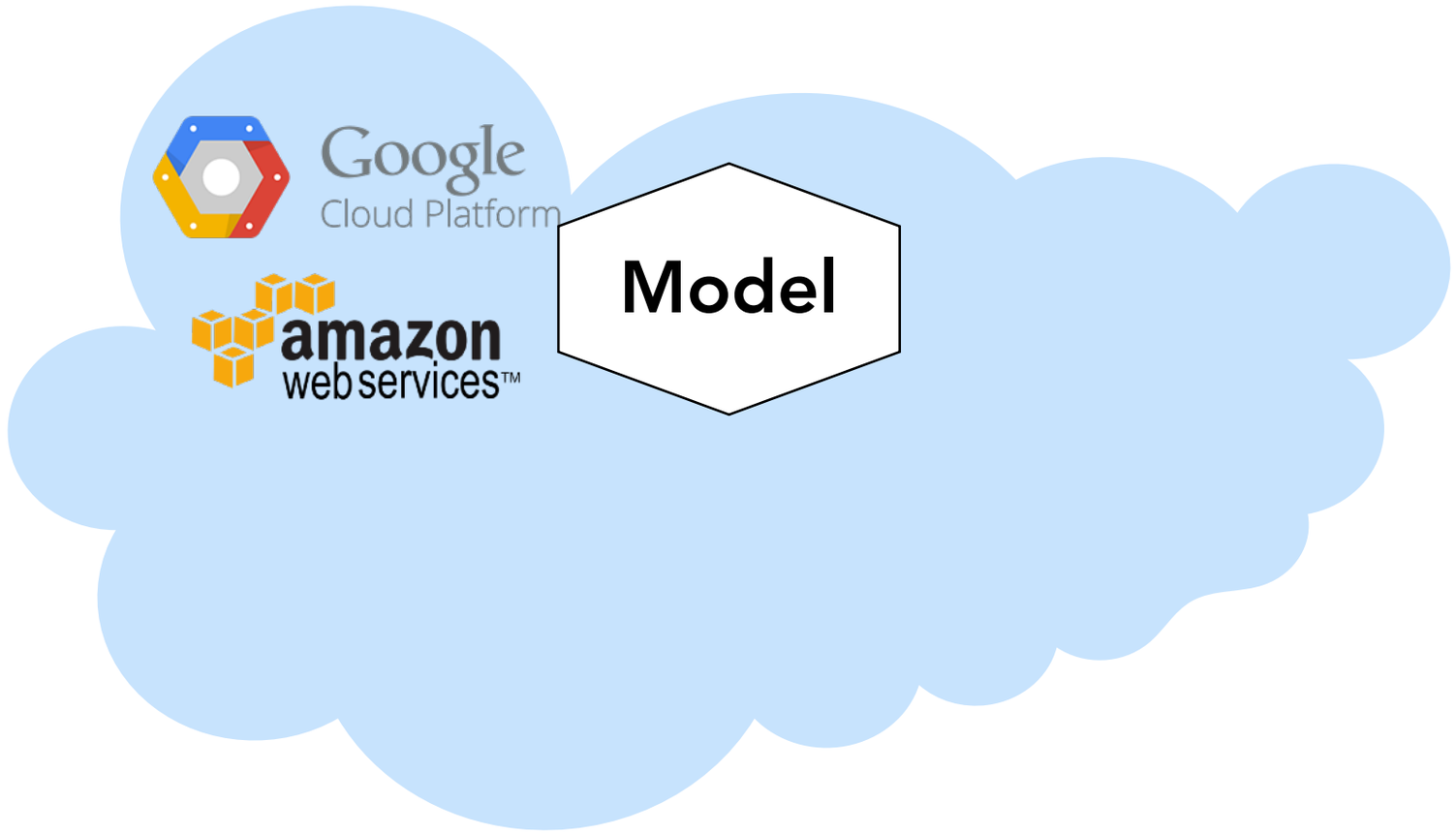
- Training could be outsourced, thus the training data is visible to (untrusted) entities

- **Given the parameters or predictions, an attacker can infer the training data**

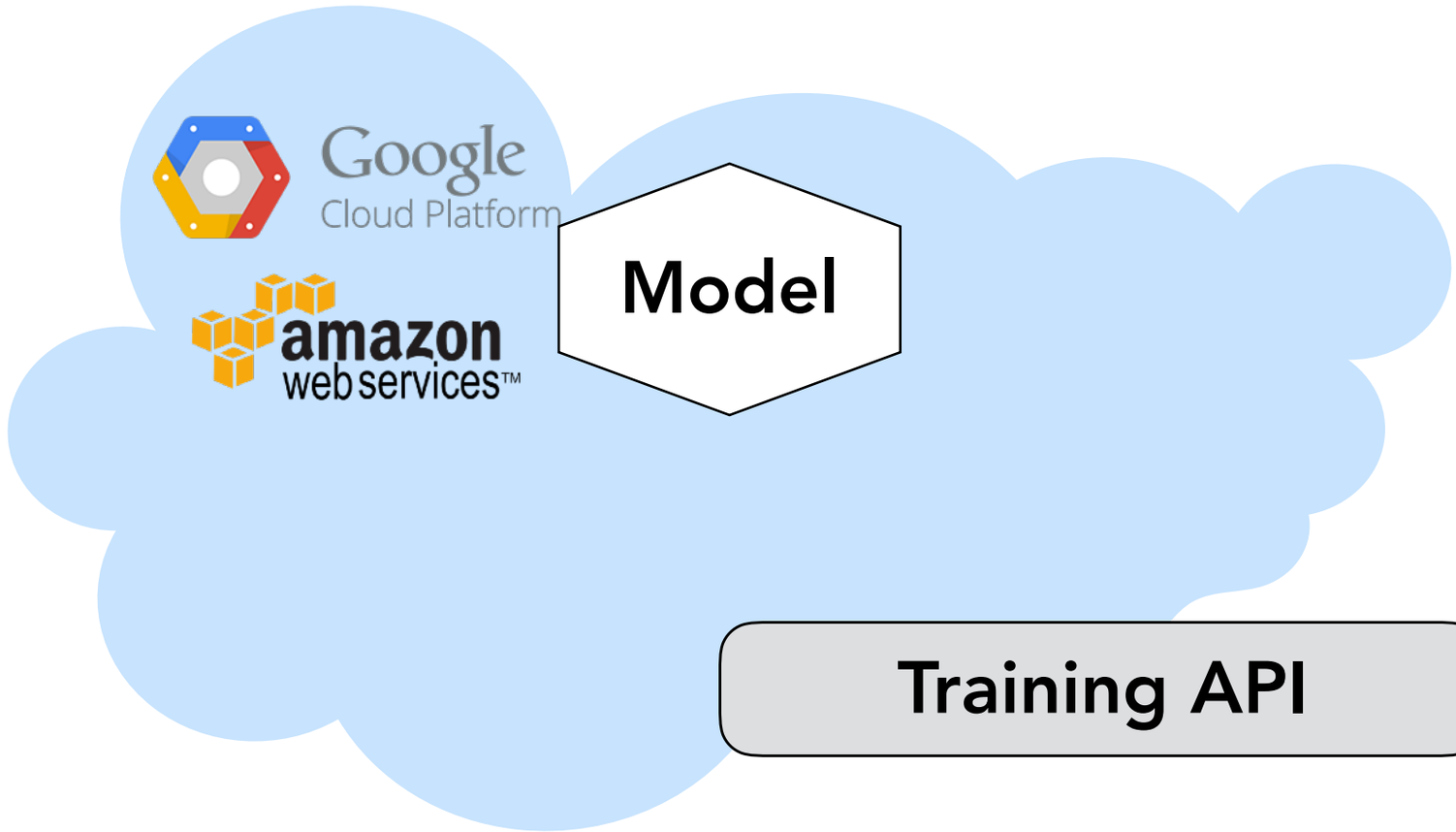
Machine Learning as a Service



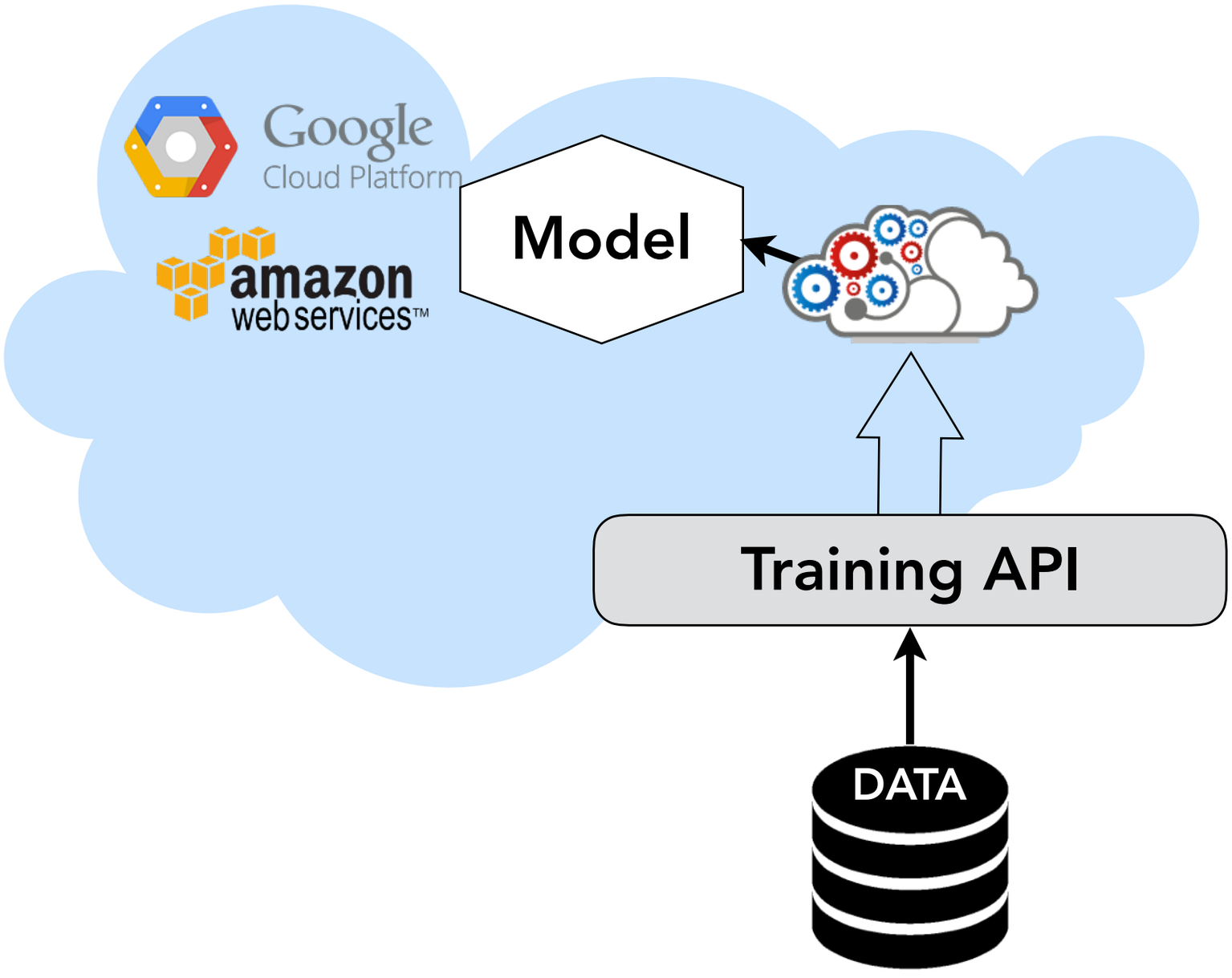
Machine Learning as a Service



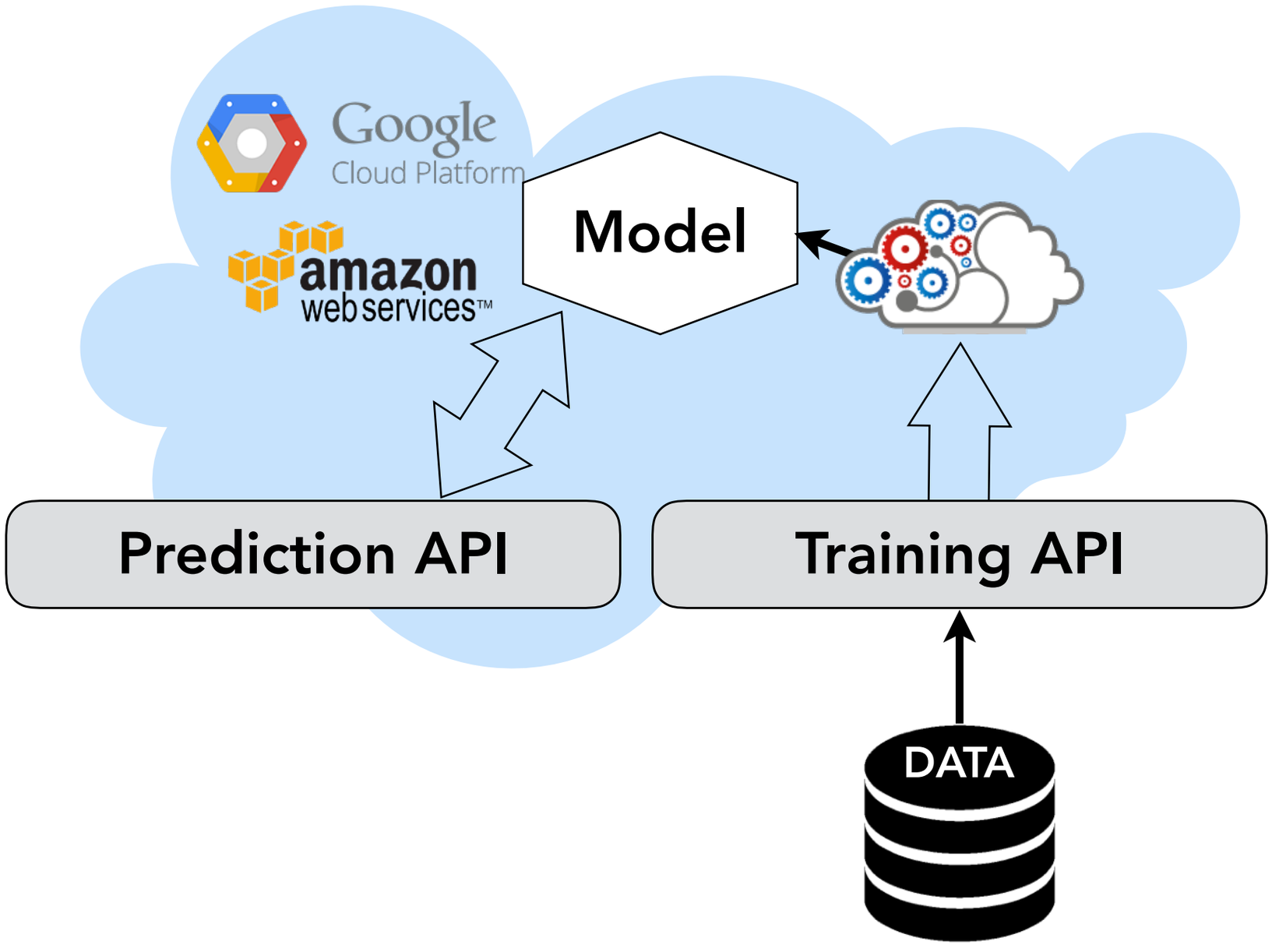
Machine Learning as a Service



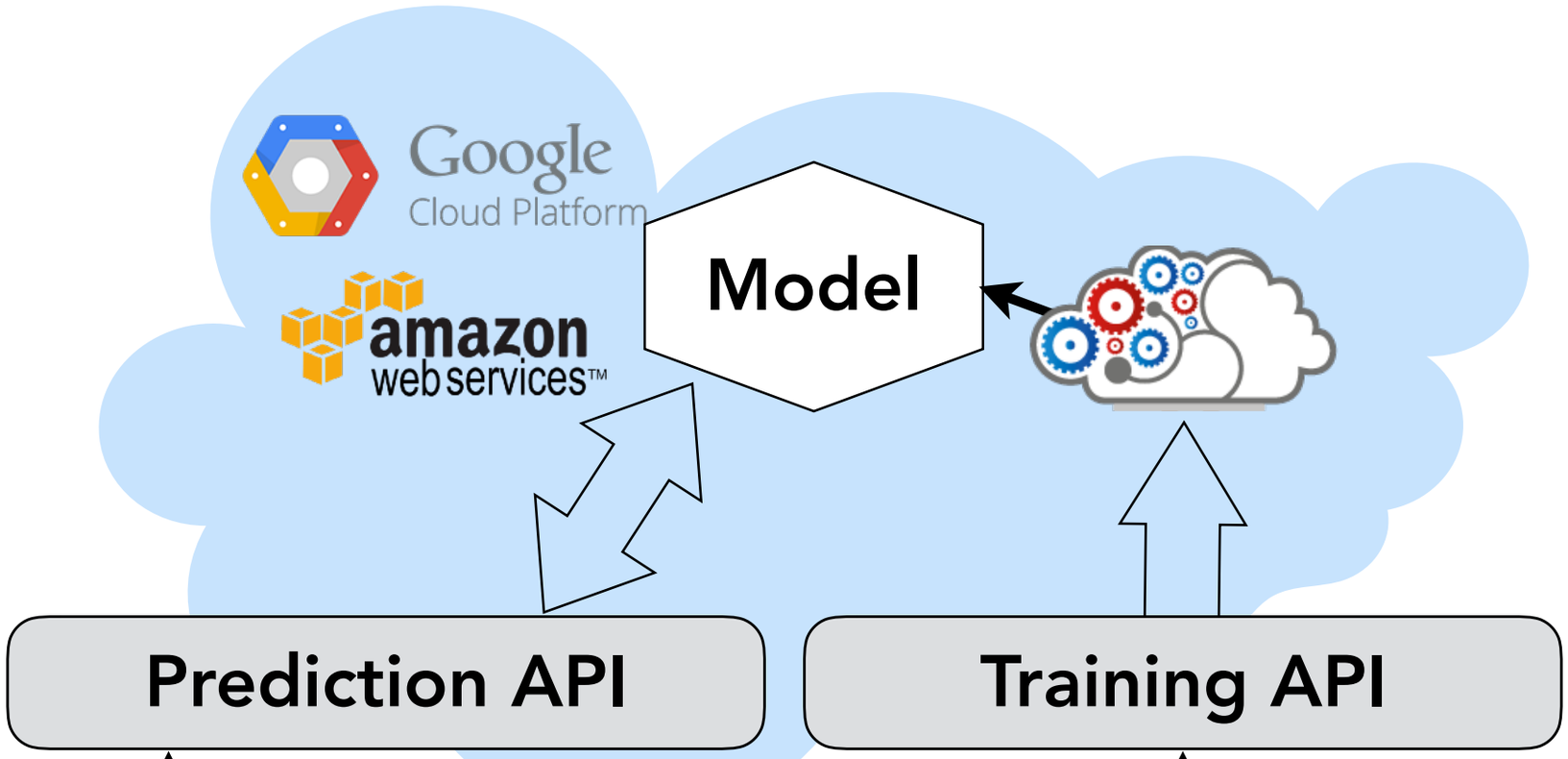
Machine Learning as a Service



Machine Learning as a Service

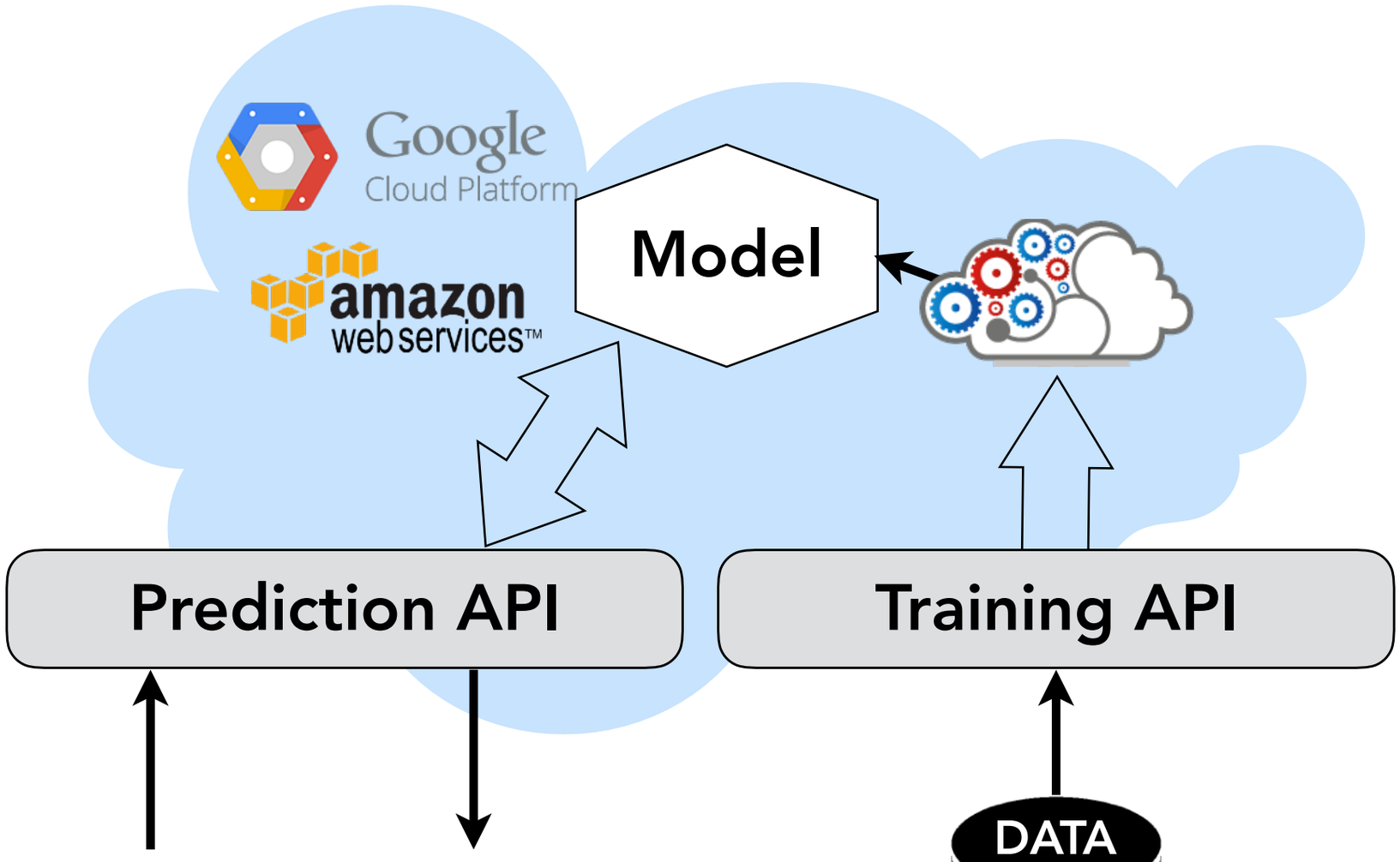


Machine Learning as a Service

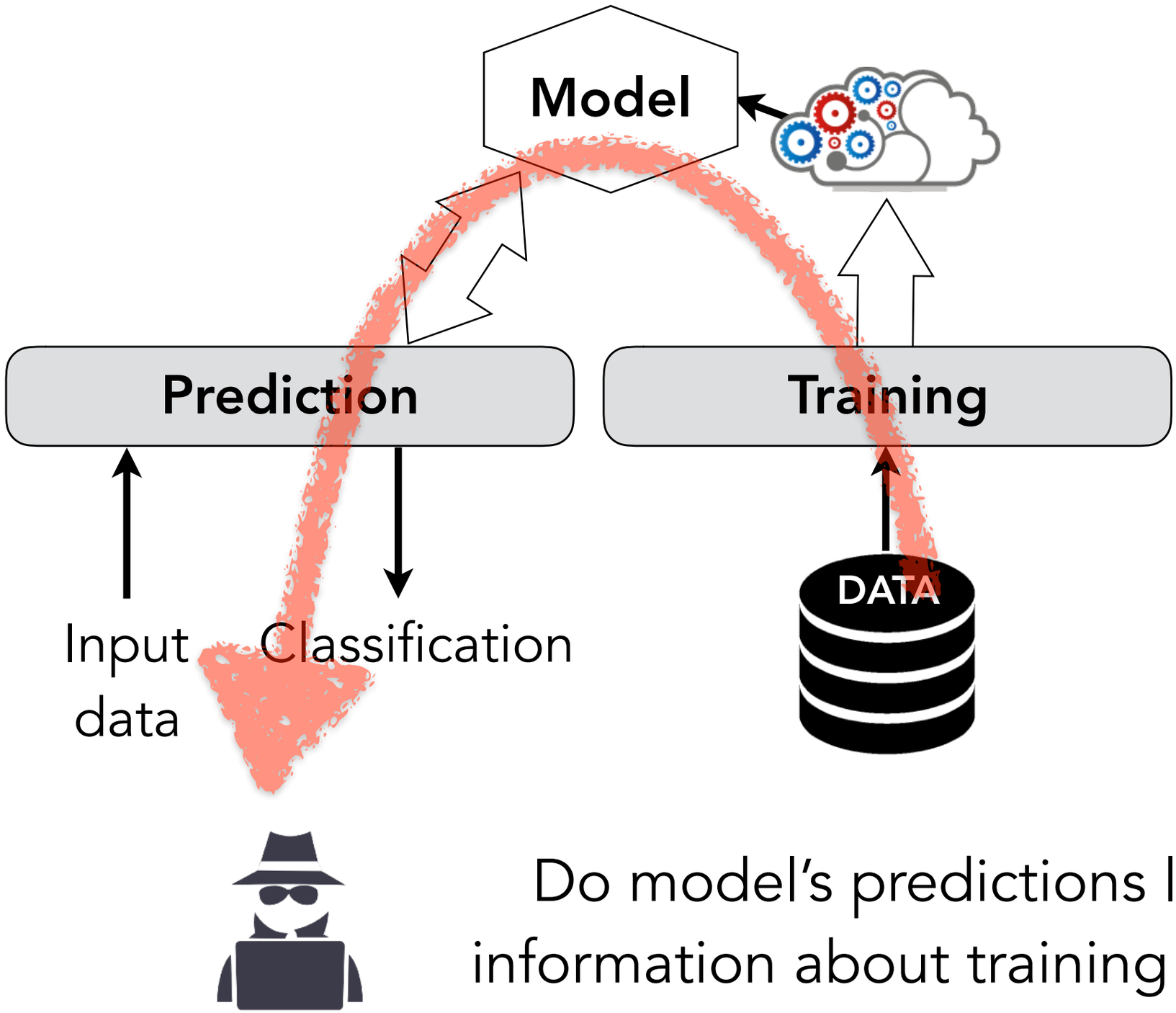


Input
data

Machine Learning as a Service



Information Leakage



Do model's predictions leak information about training data?

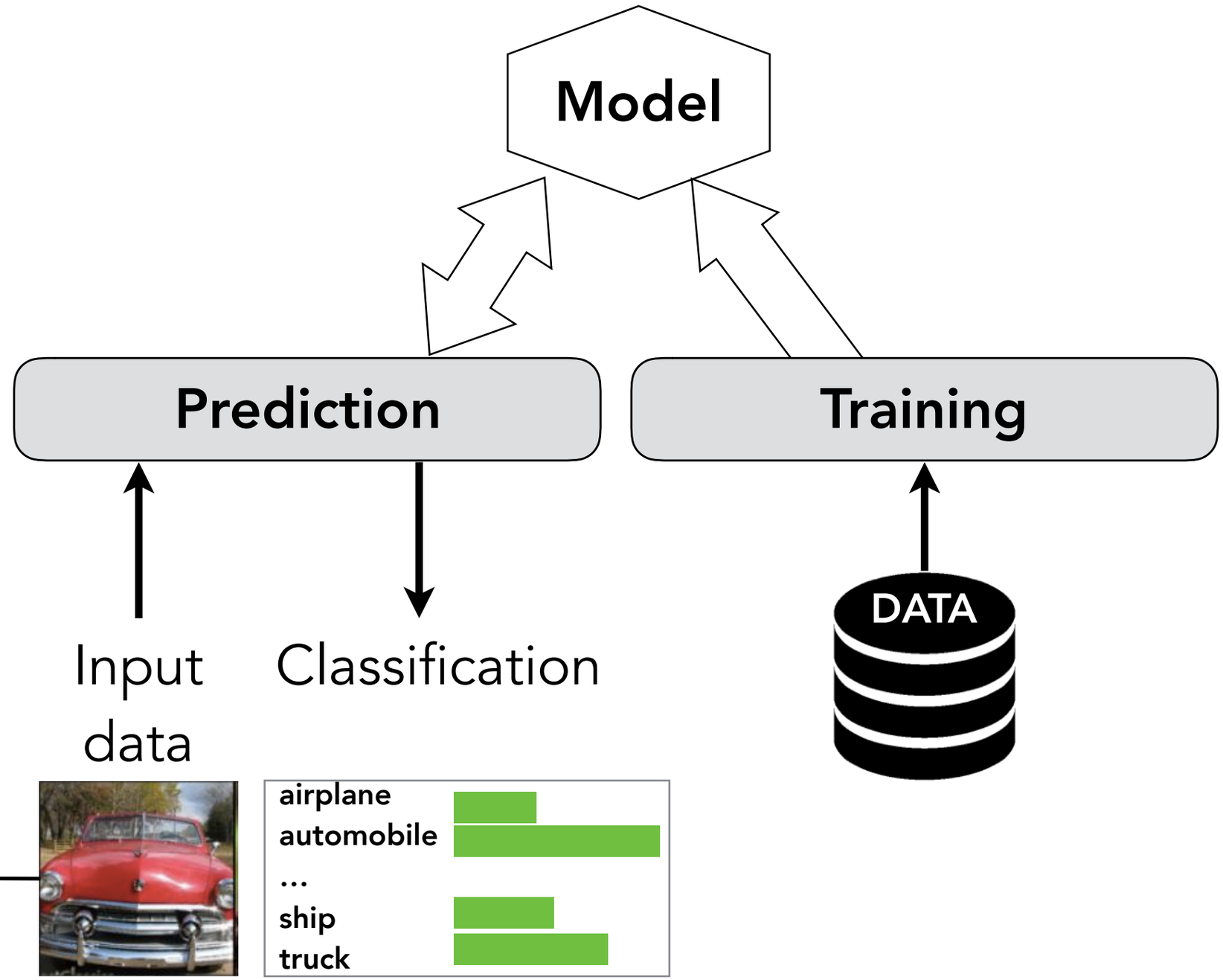
Machine Learning Privacy

"Leakage"

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

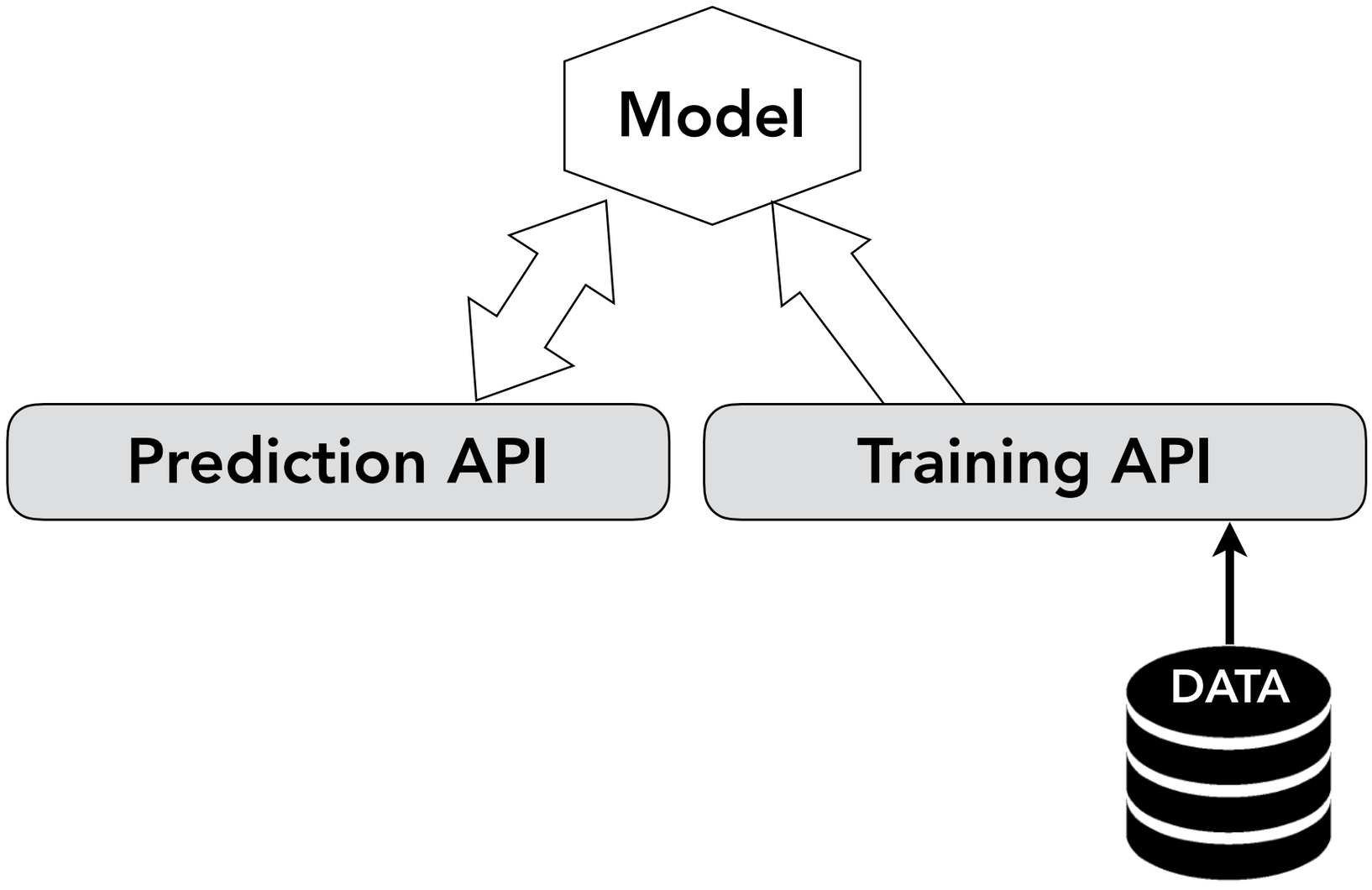
Do model's predictions leak information about training data?

Membership Inference Attack



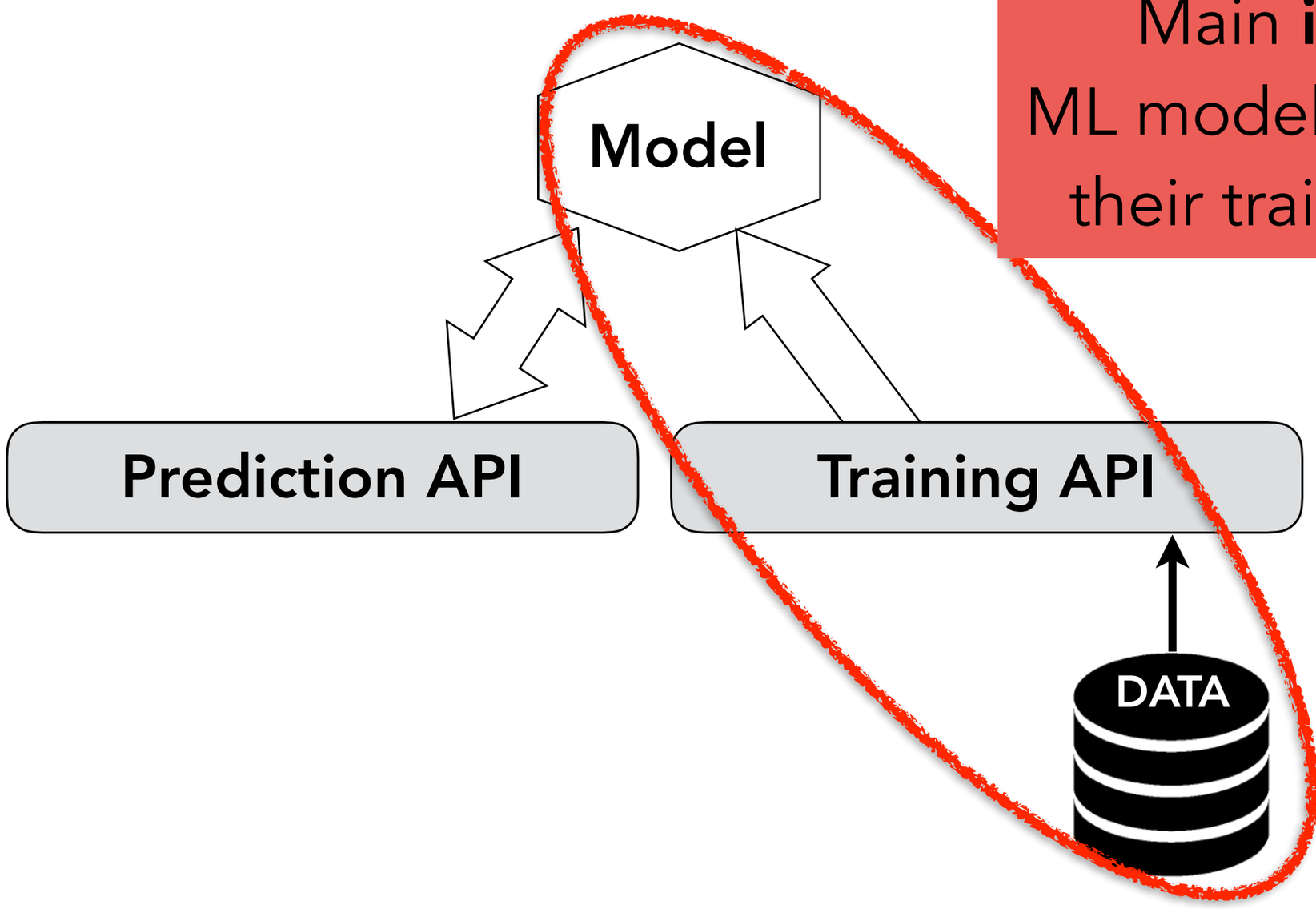
Was this specific data record part of the training set?

Exploit Model's Predictions

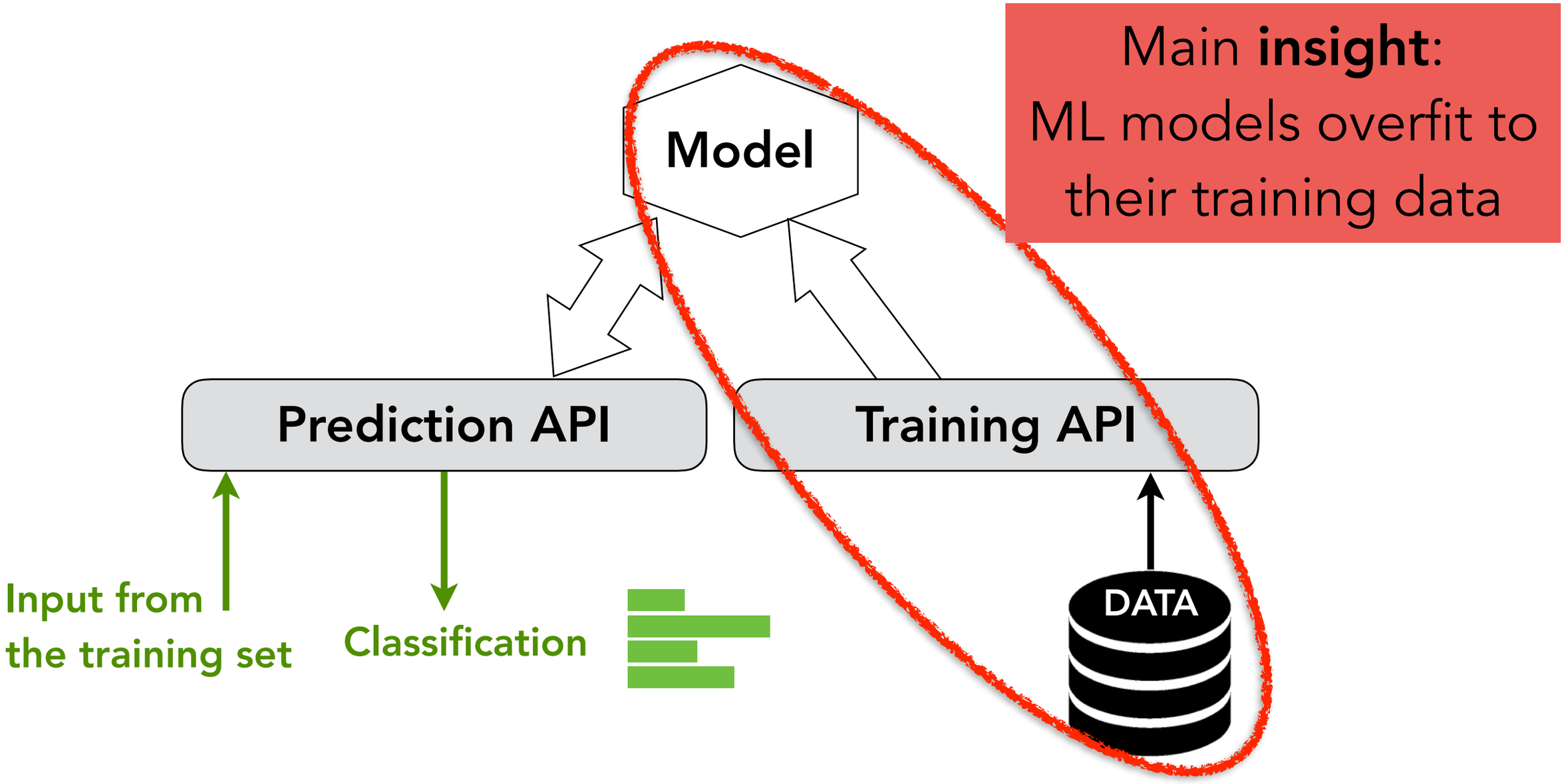


Exploit Model's Predictions

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



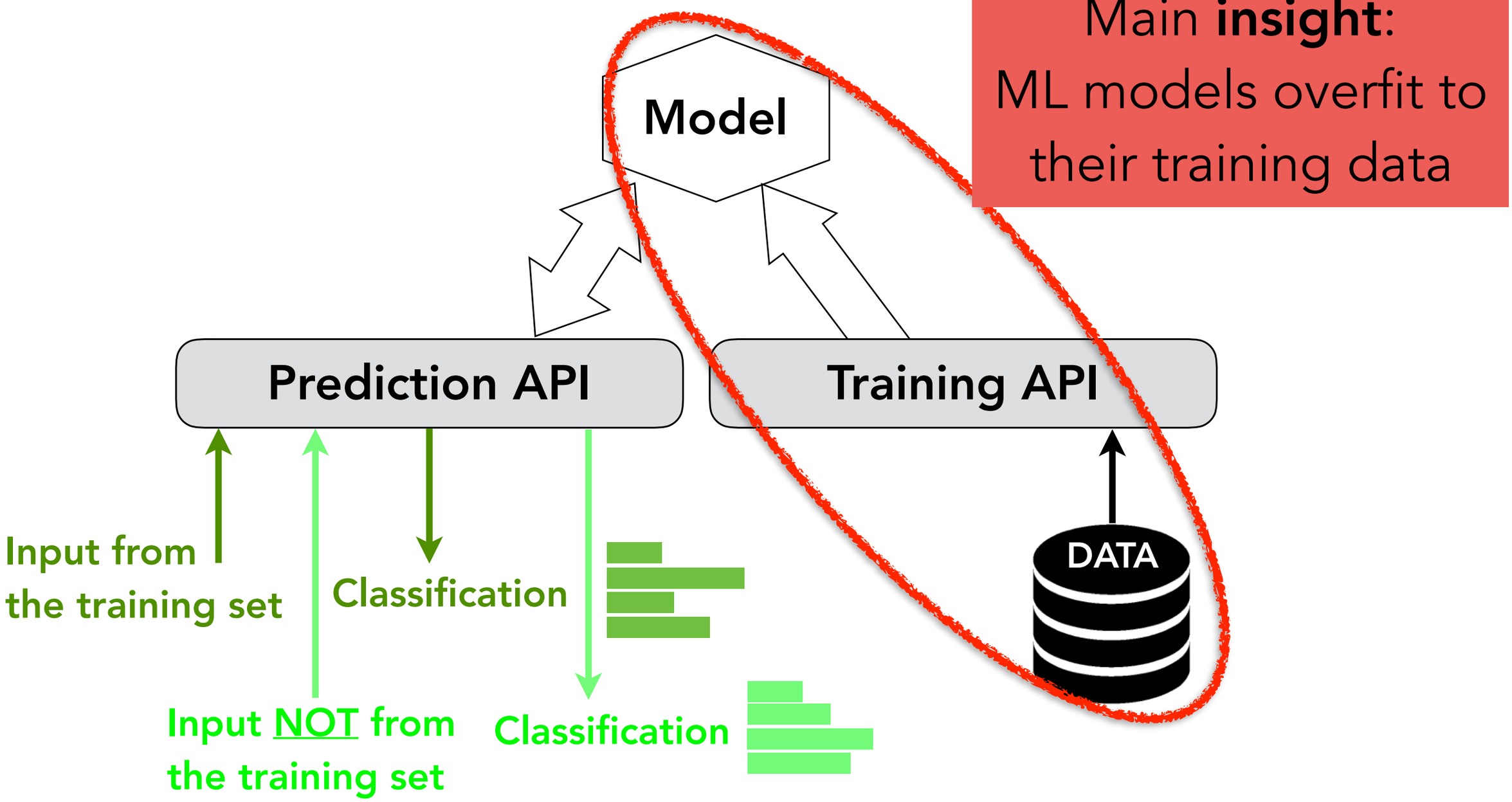
Exploit Model's Predictions



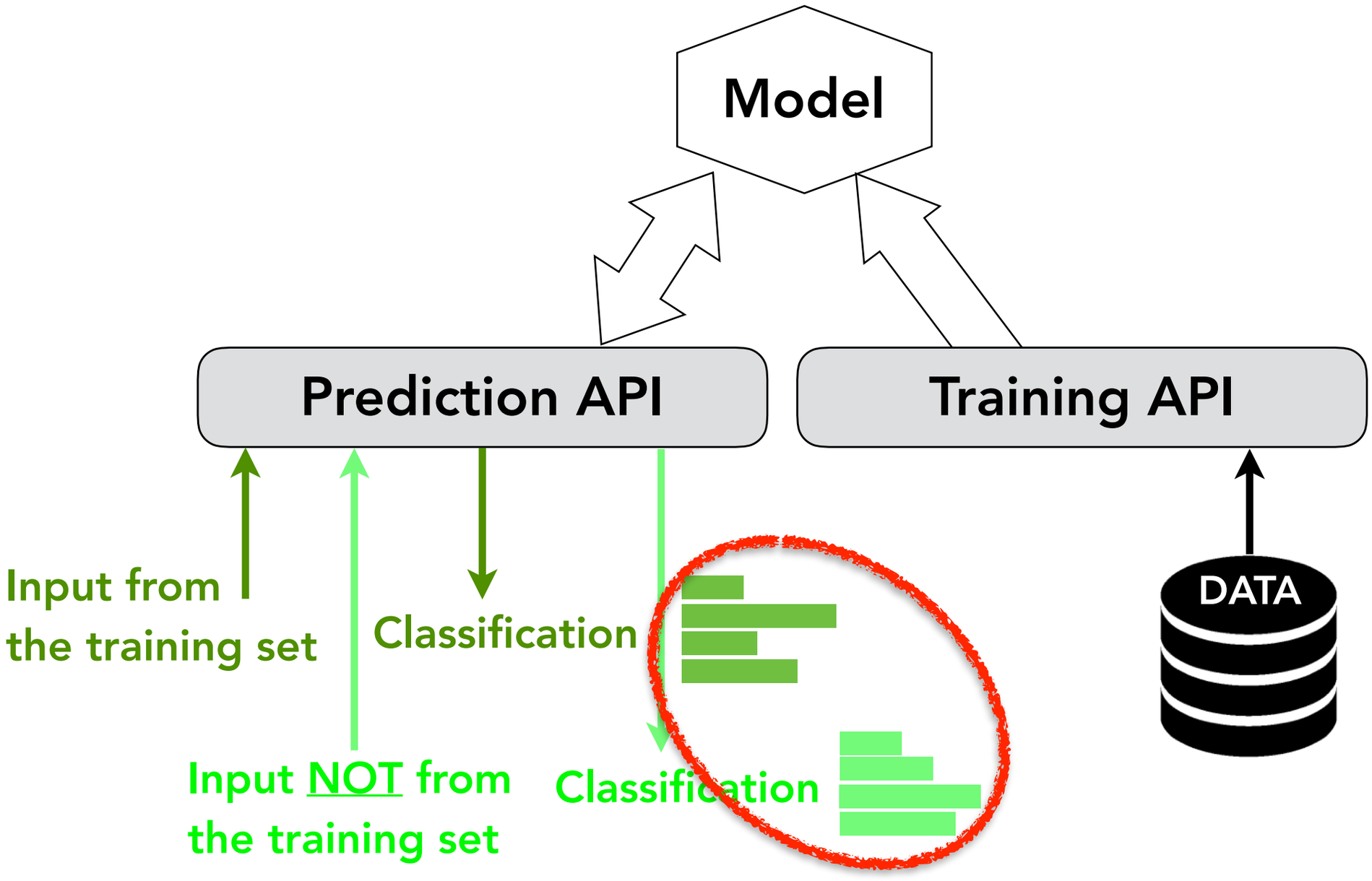
Main **insight**:
ML models overfit to their training data

Exploit Model's Predictions

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

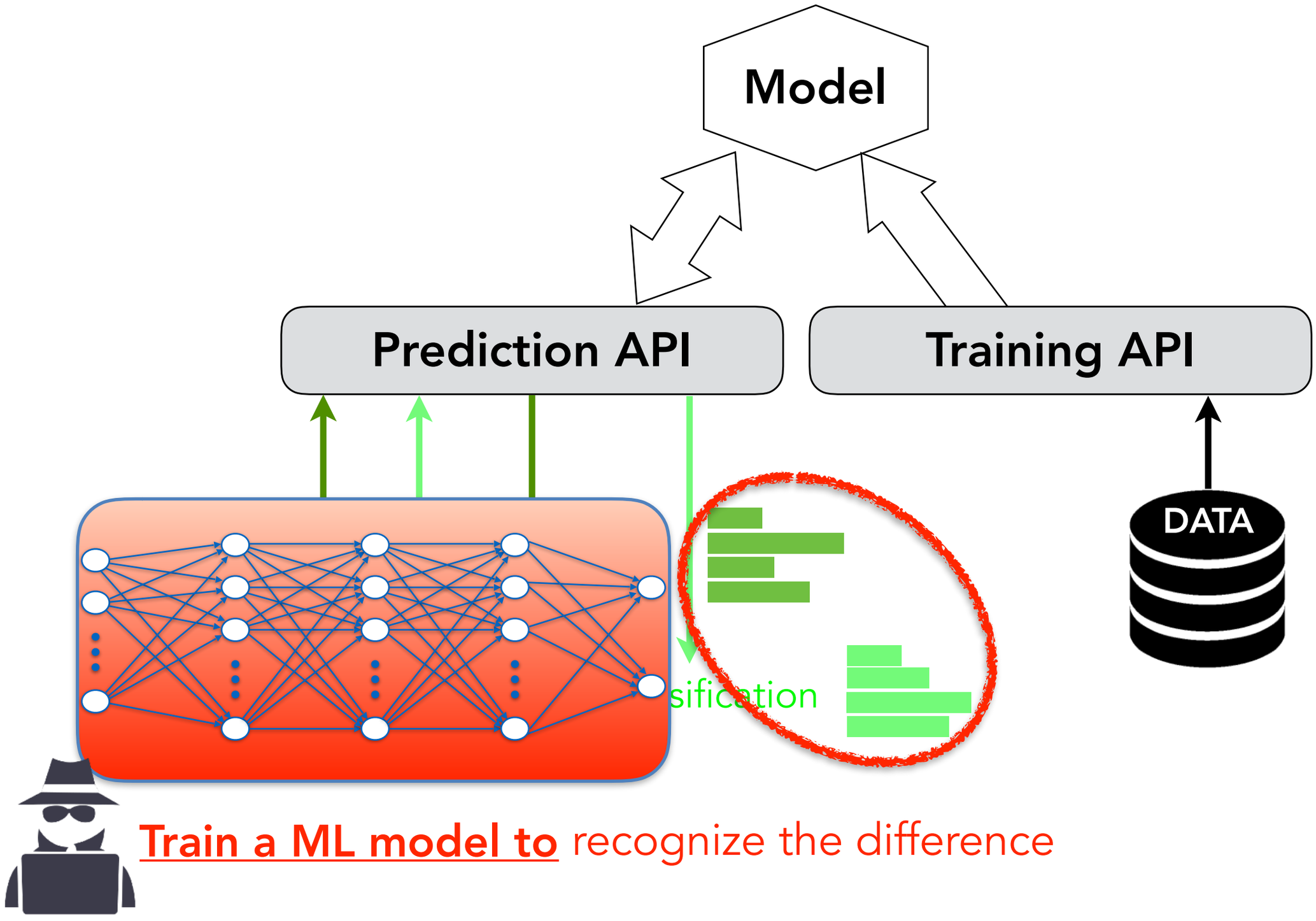


Exploit Model's Predictions



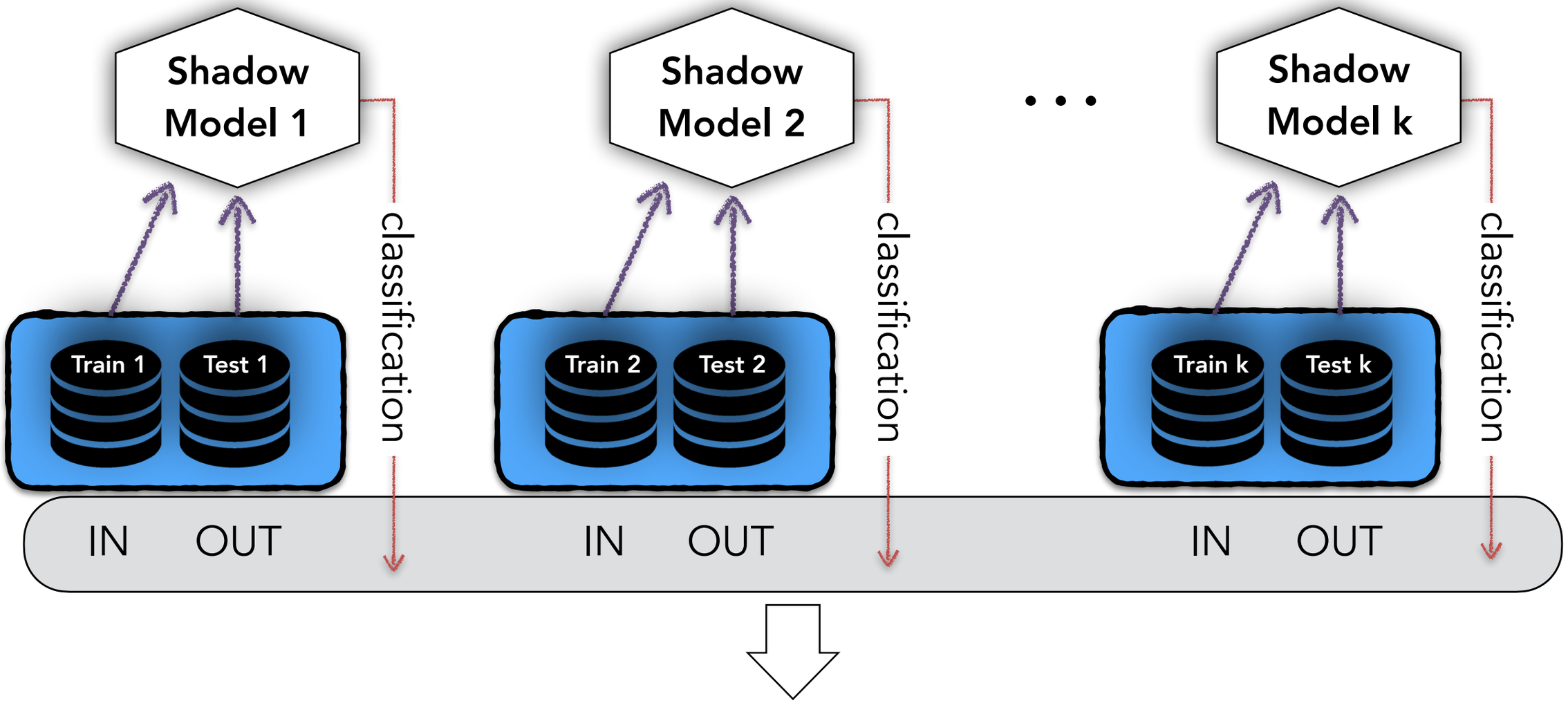
Recognize the difference

ML against ML



Train a ML model to recognize the difference

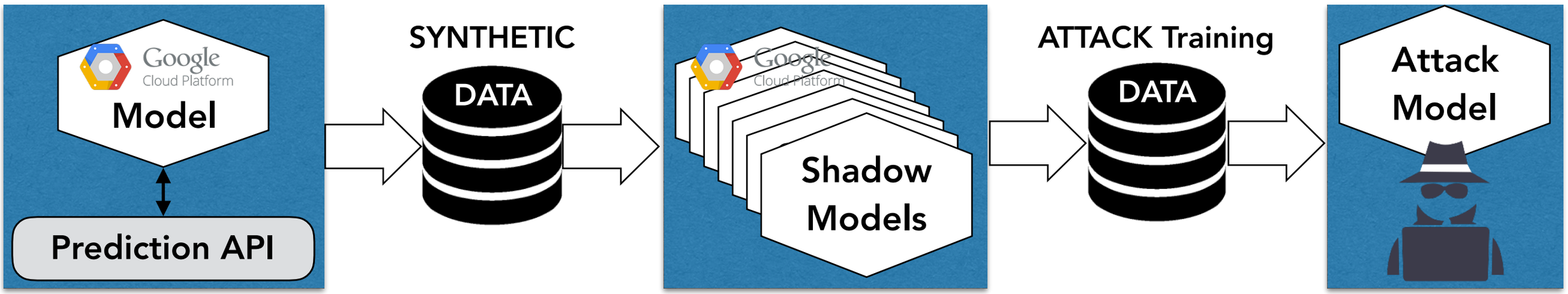
Train Attack Model using Shadow Models



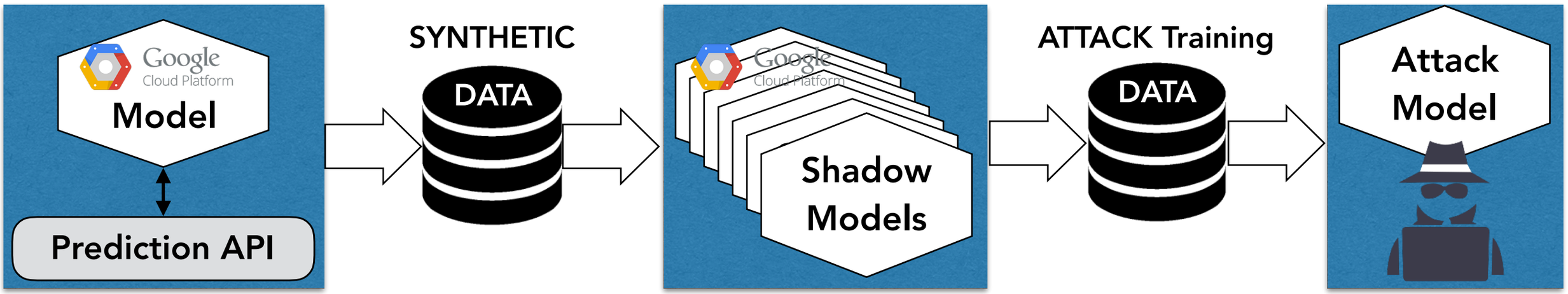
Train the attack model

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

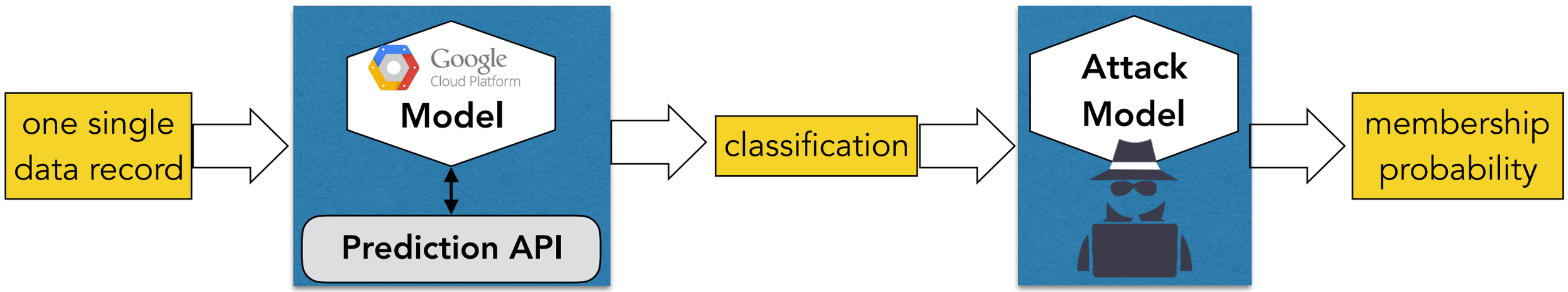
Construct the Attack Model

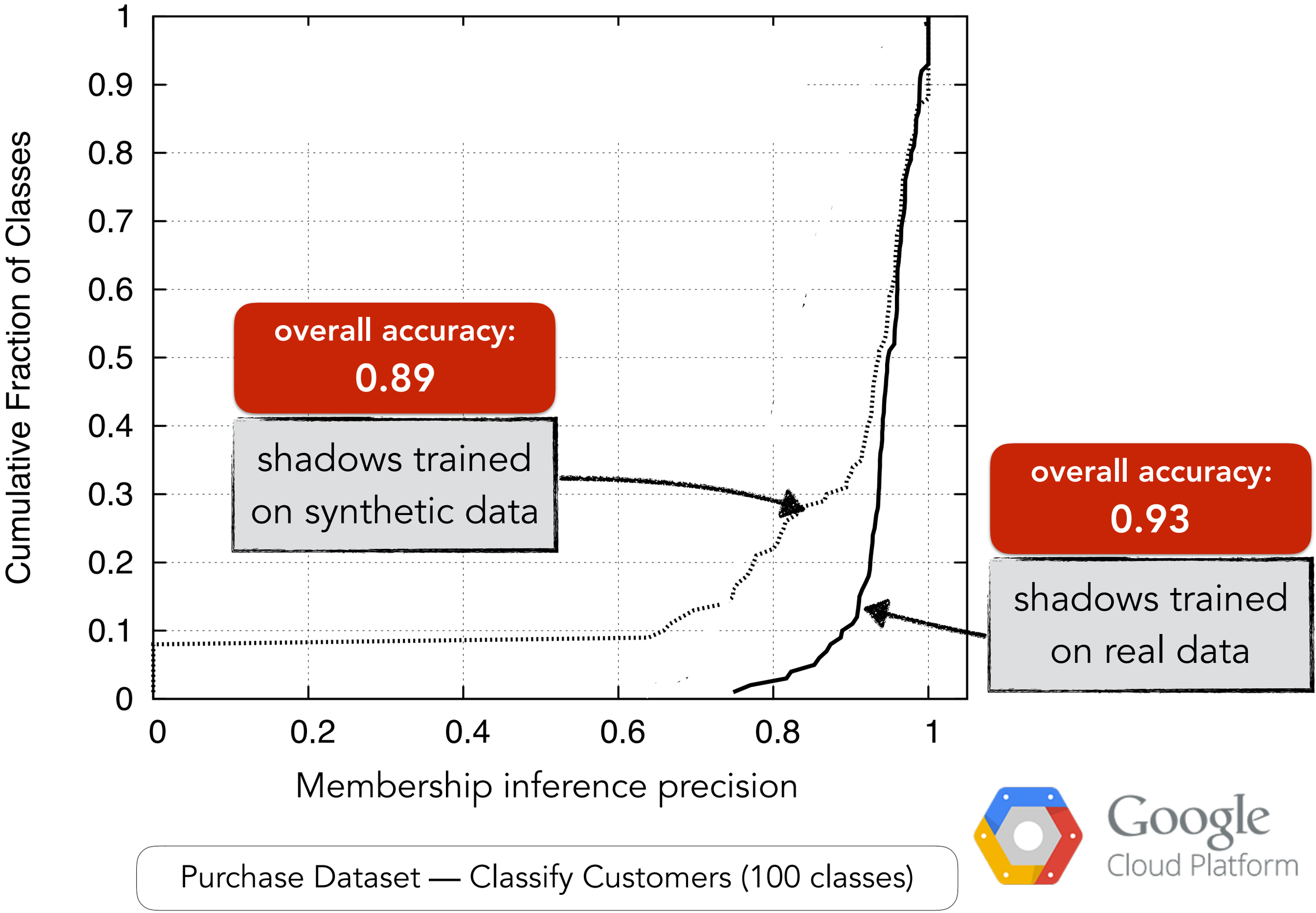


Construct the Attack Model



Using the Attack Model



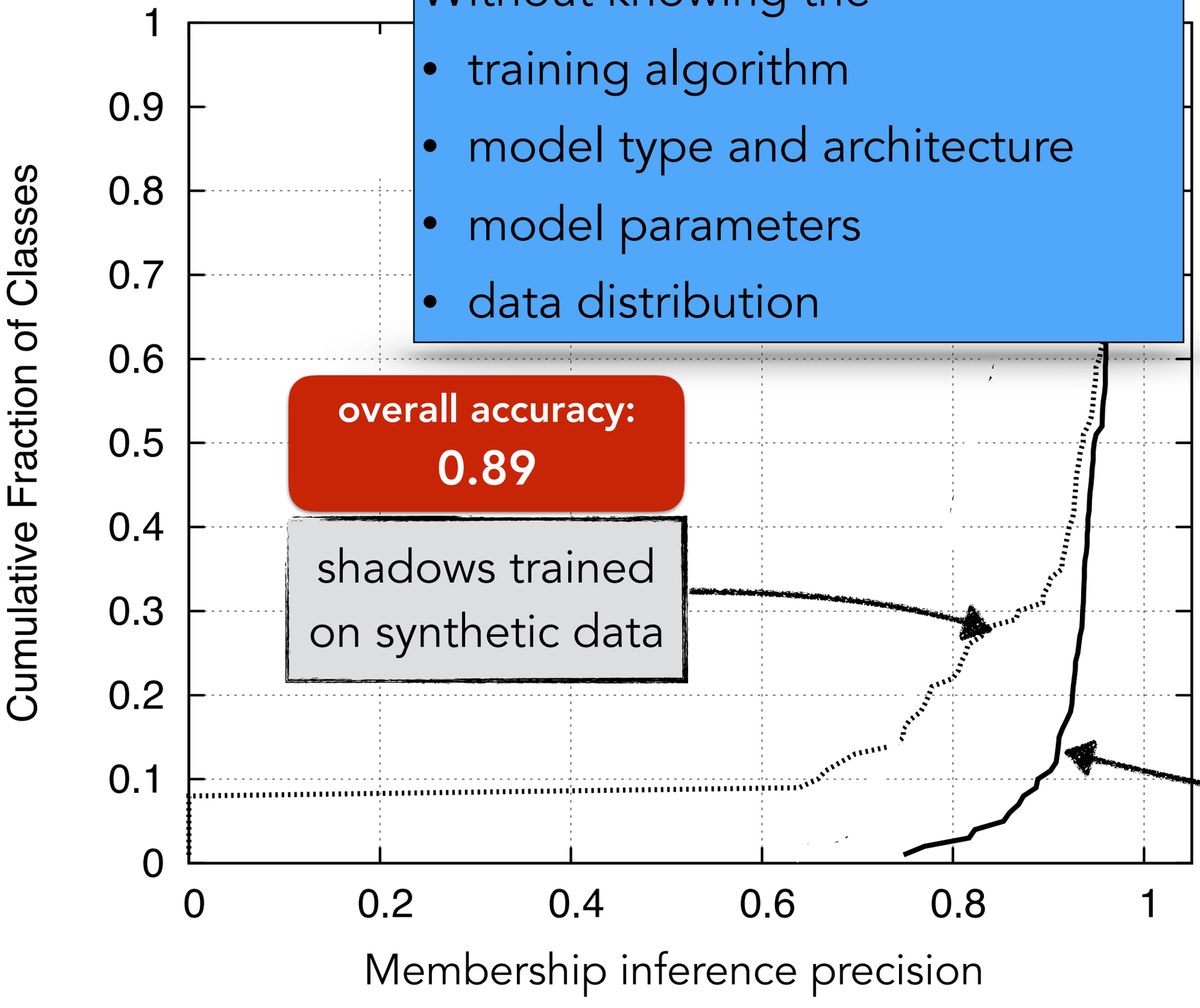


Purchase Dataset — Classify Customers (100 classes)

Google Cloud Platform

Without knowing the

- training algorithm
- model type and architecture
- model parameters
- data distribution

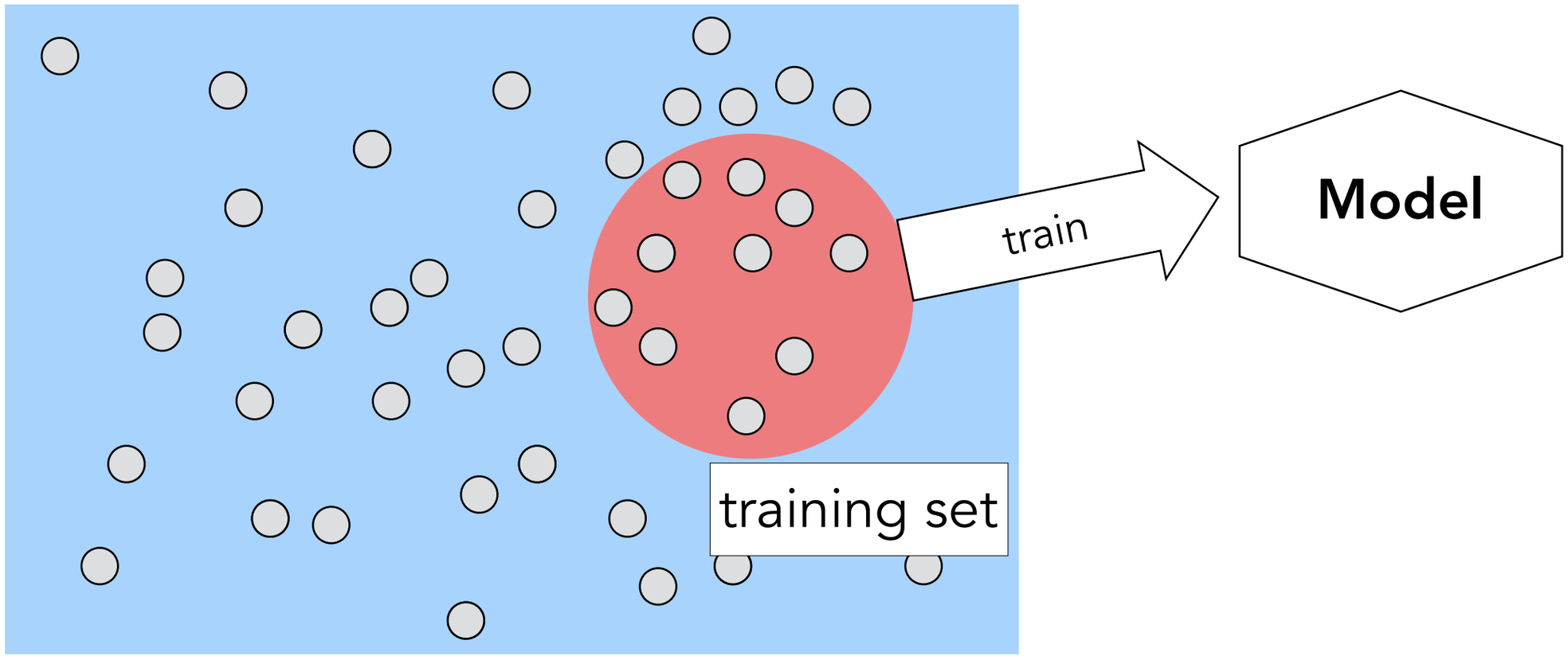


Purchase Dataset — Classify Customers (100 classes)

Google Cloud Platform

Privacy

Learning

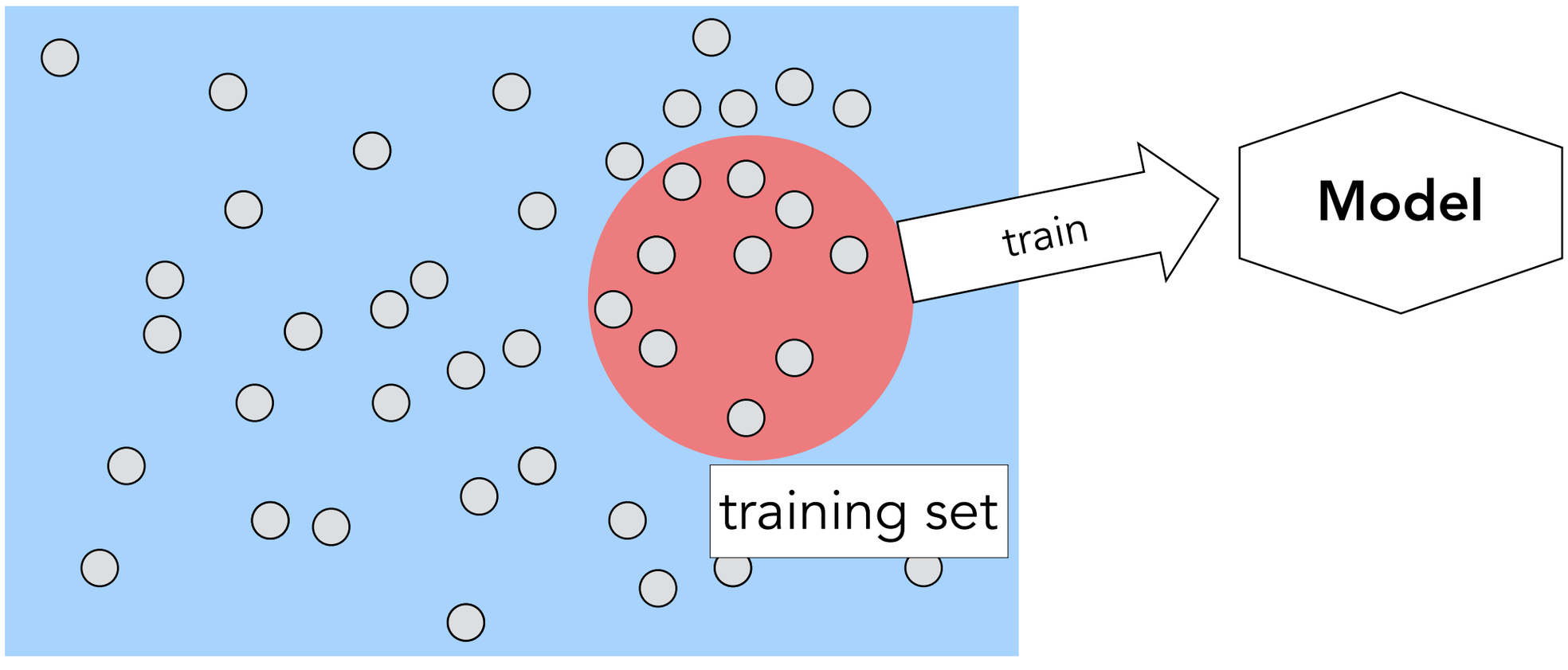


Note: information leakage wrt the model's **output** (predictions)

Privacy

Learning

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



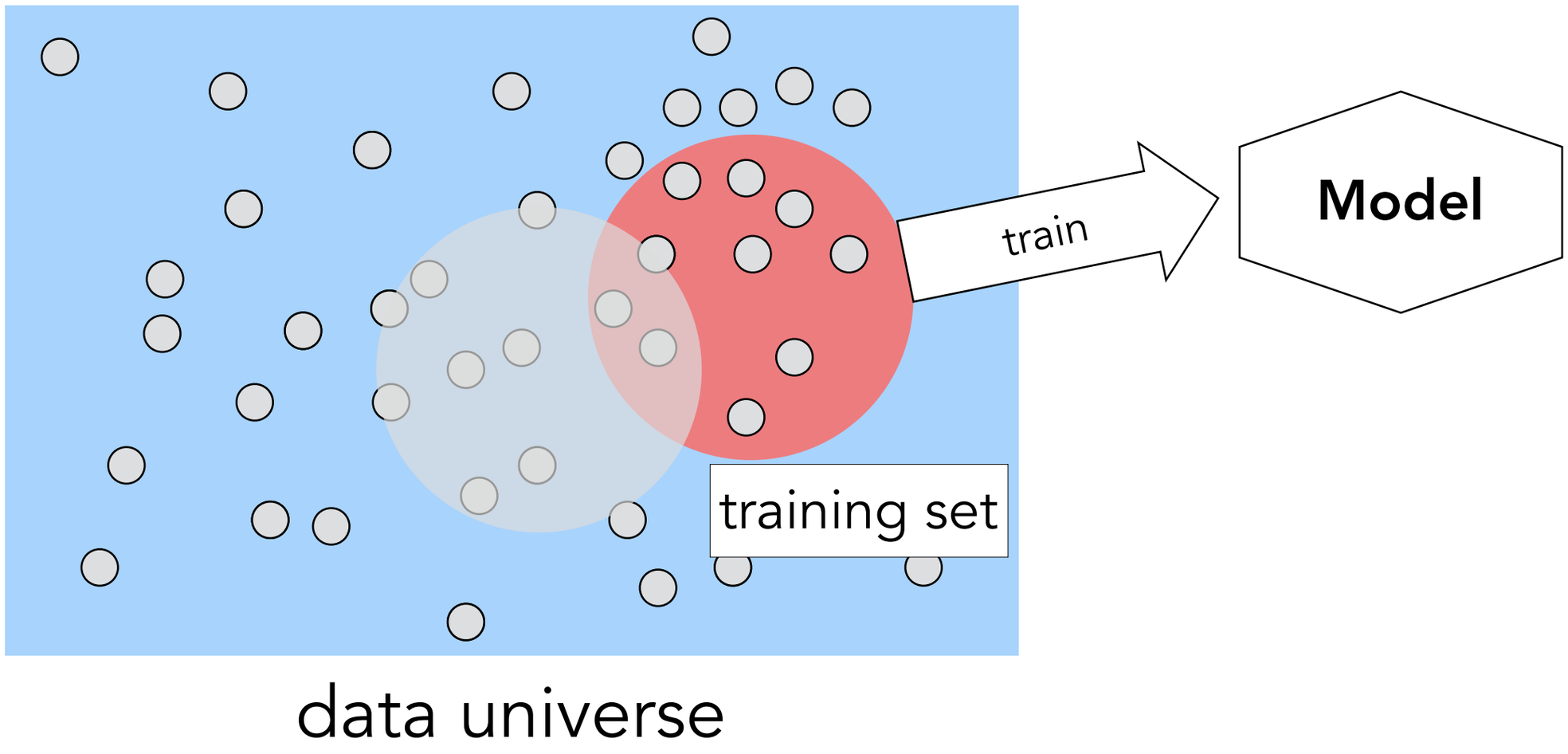
Note: information leakage wrt the model's **output** (predictions)

Privacy

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

Learning

Does the model generalize to data outside the training set?



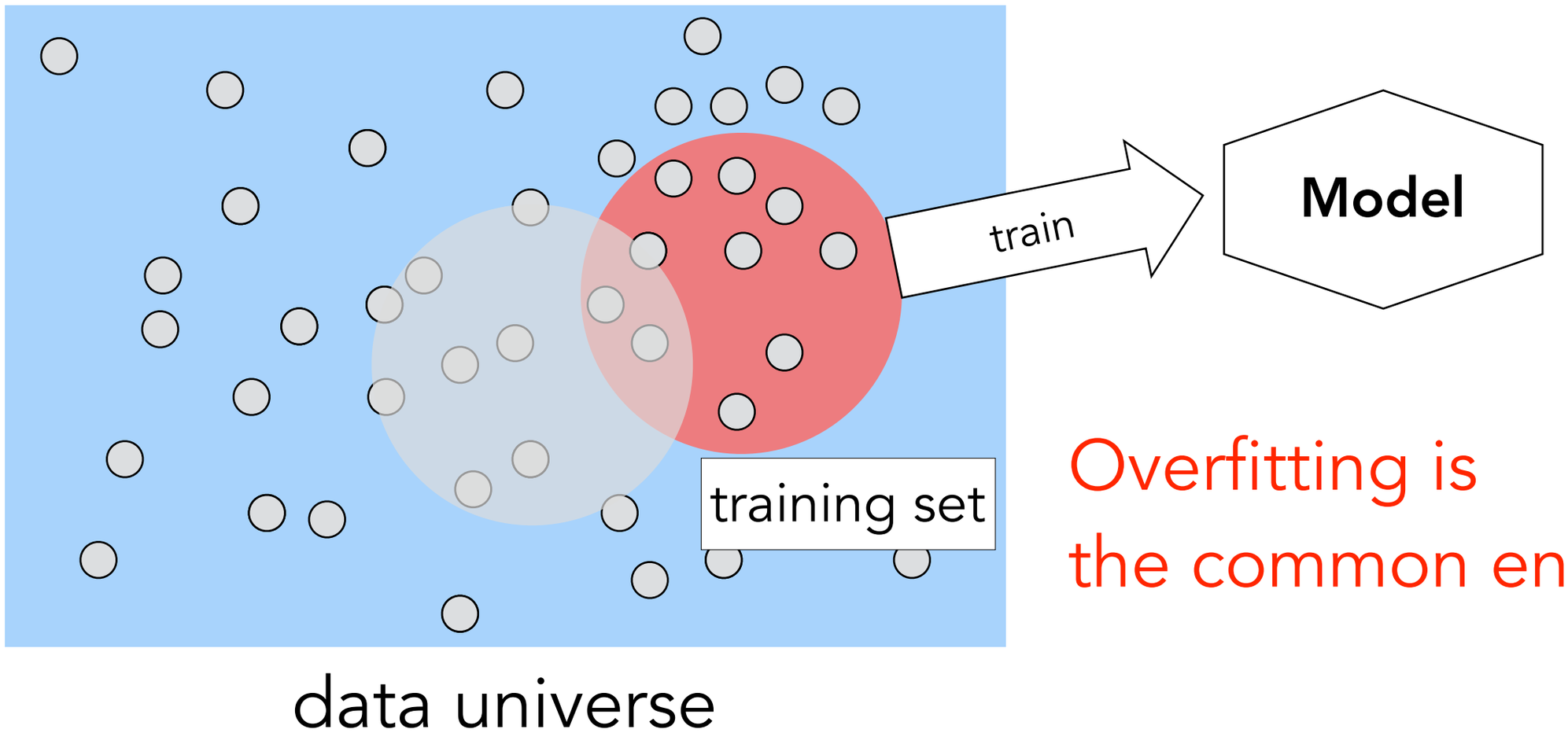
Note: information leakage wrt the model's **output** (predictions)

Privacy

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

Learning

Does the model generalize to data outside the training set?

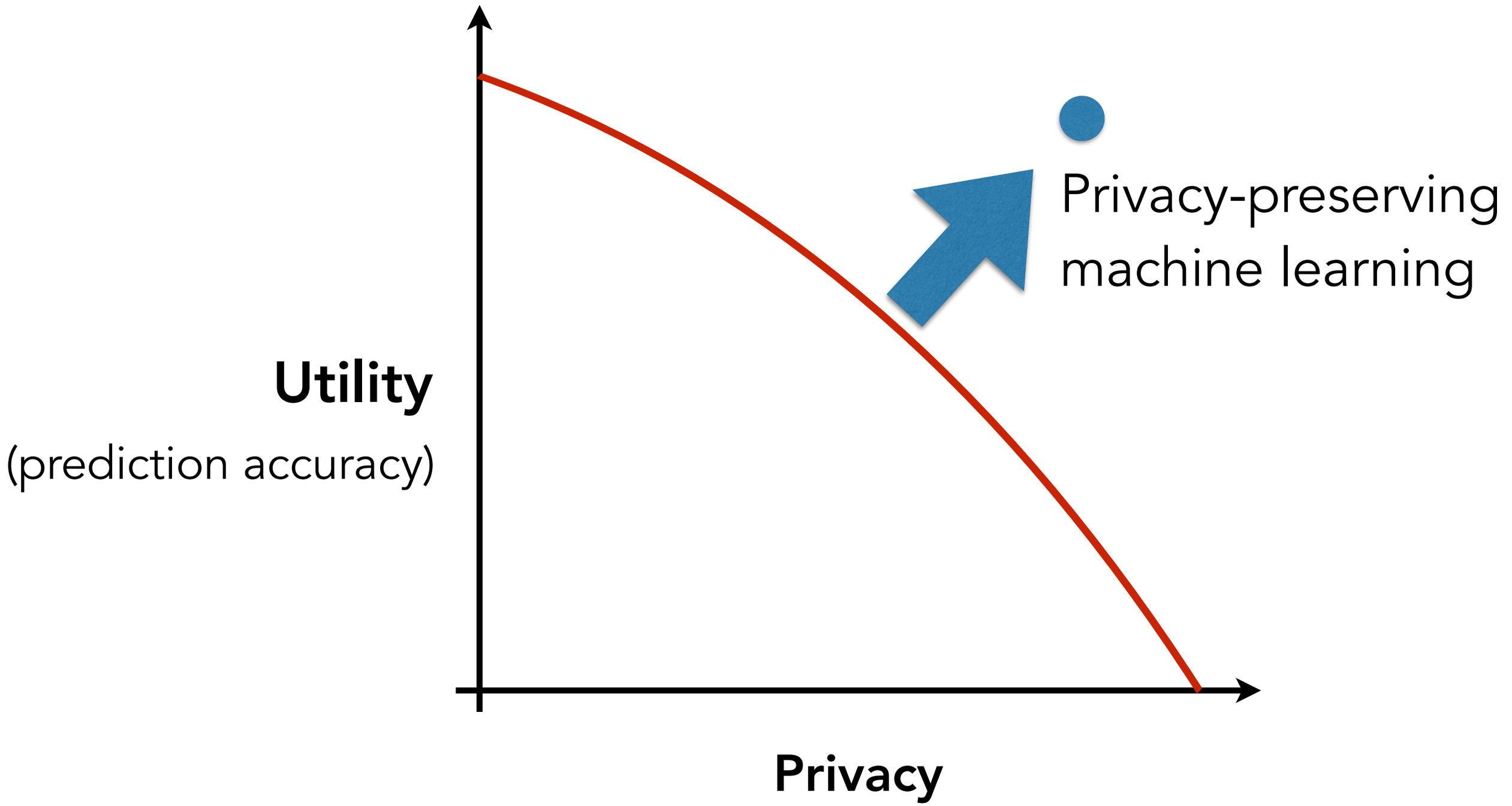


Overfitting is the common enemy!

Note: information leakage wrt the model's **output** (predictions)

Privacy-Utility:

Not in a Direct Conflict!



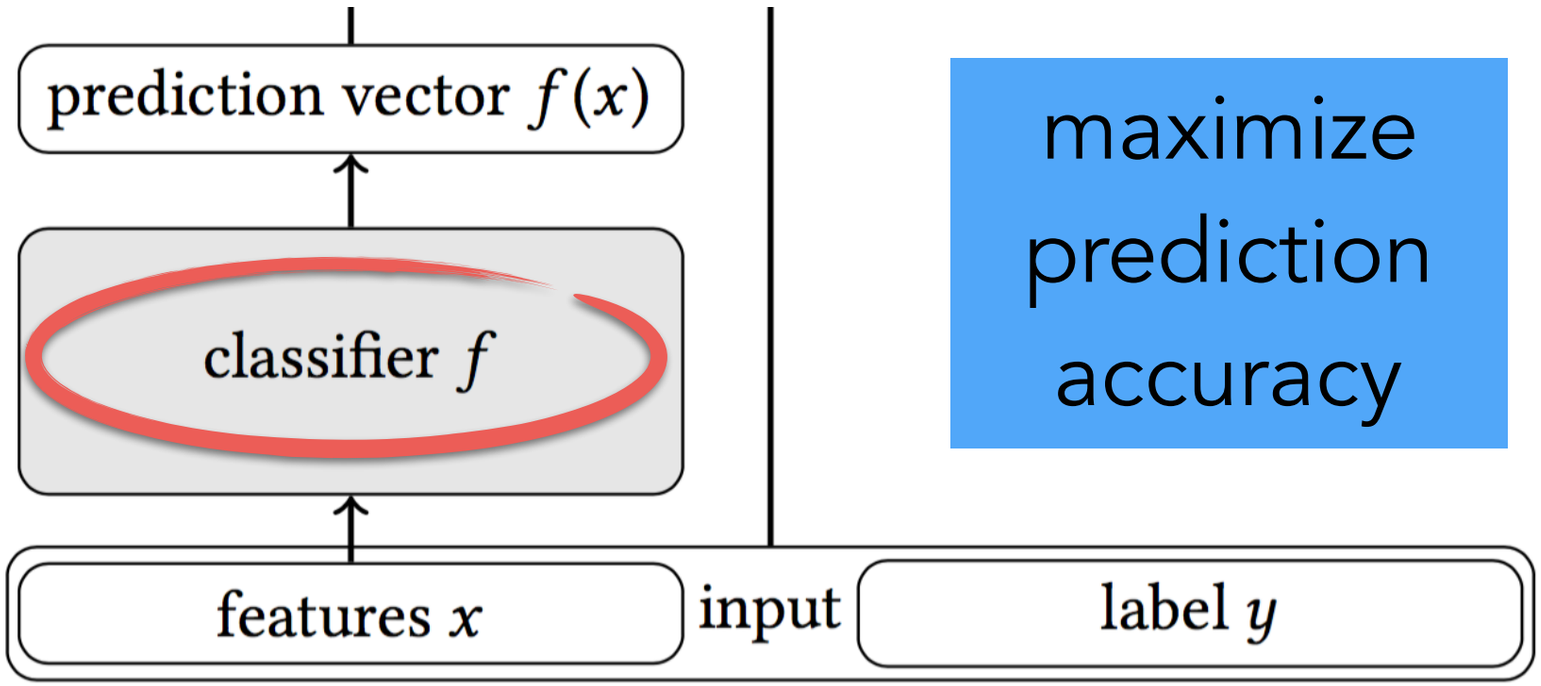
Learning

Empirical loss over D:

$$\mathrm{L}_D(f) = \frac{1}{|D|} \sum_{(x,y)\in D} l(f(x), y)$$

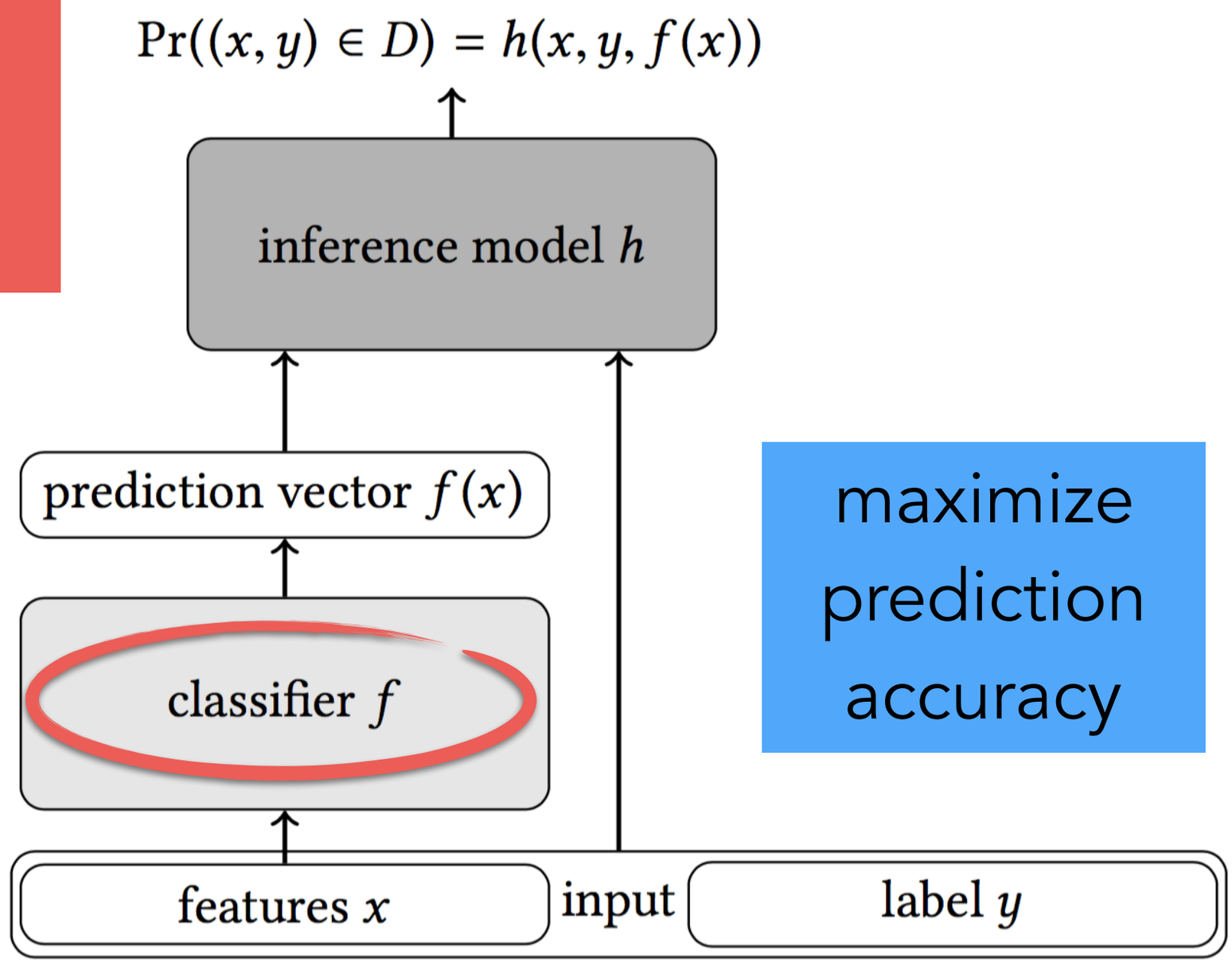
Learning optimization:

$$\min_f \mathrm{L}_D(f) + \lambda R(f)$$



Privacy as a Learning Objective

minimize inference accuracy



maximize prediction accuracy

Minmax Membership Privacy Game

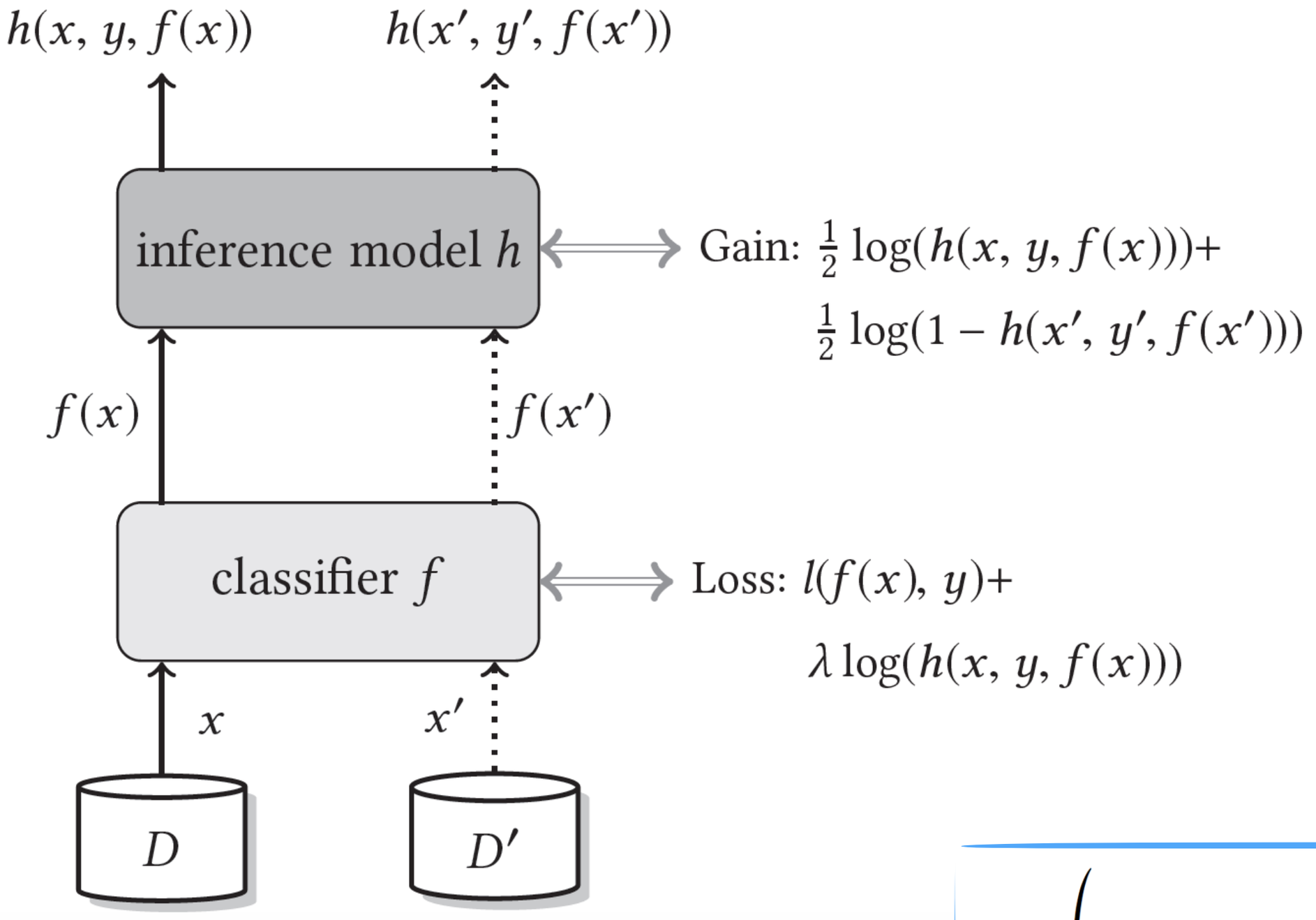
Adversarial Regularization

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

$$\min_{f} \Big(\mathrm{L}_D(f) + \lambda \underbrace{\max_{h} \mathrm{G}_{f,D,D'}(h)}_{\text{optimal inference}} \Big)$$

optimal privacy-preserving classification

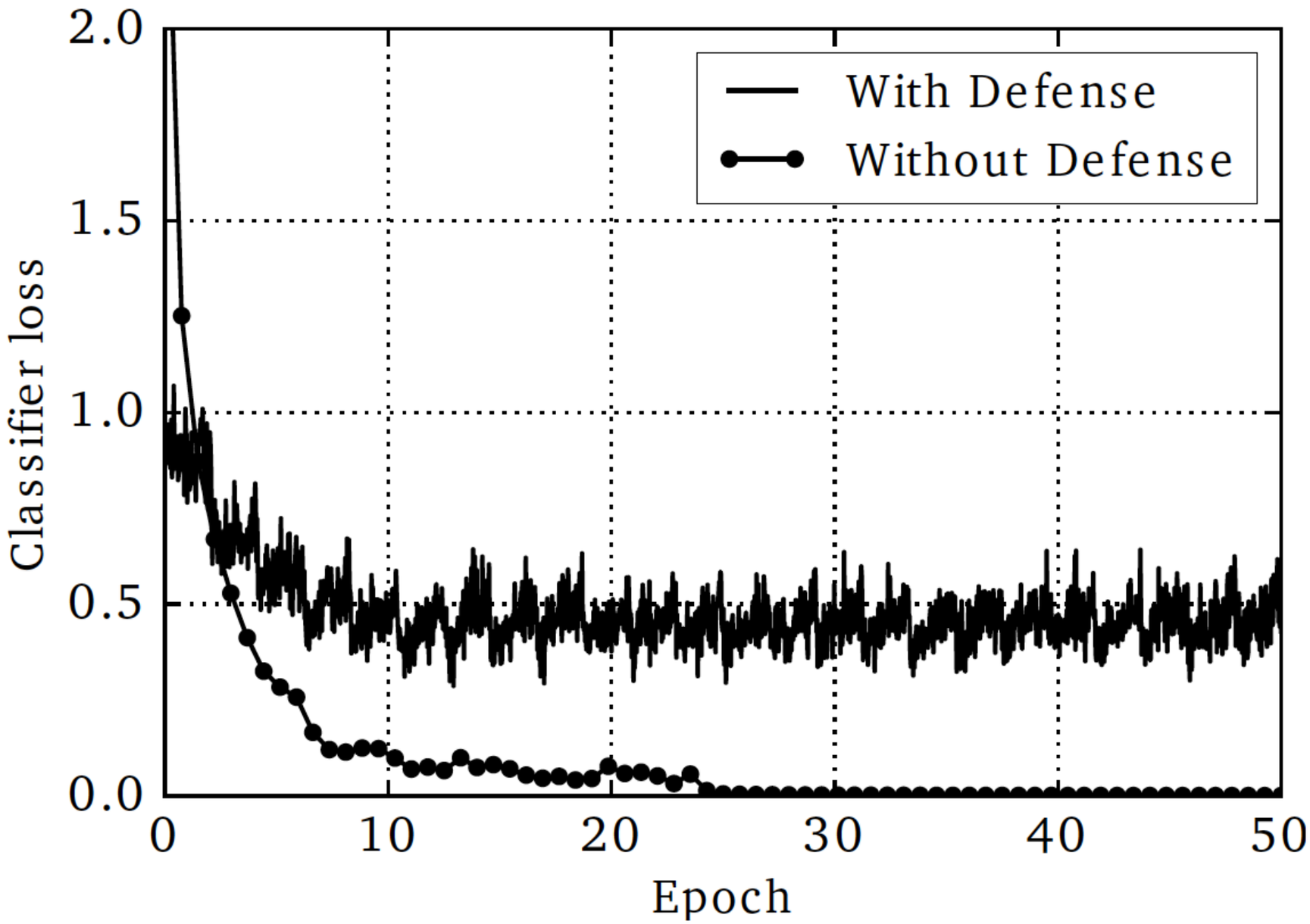
classification model with minimum loss (predicting correct class for any input)



Train using SGD

$$\underbrace{\min_{f}\left(\mathrm{L}_D(f) + \lambda \underbrace{\max_{h} \mathrm{G}_{f,D,D'}(h)}_{\text{optimal inference}}\right)}_{\text{optimal privacy-preserving classification}}$$

Membership Privacy Mechanism as a Regularizer



Privacy

	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

Attack accuracy:
near random guess

Generalization

Defeat overfitting

	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

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

Generalization

Defeat overfitting

	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

L2-regularization factor	Training accuracy	Testing accuracy	Attack accuracy
0 (no regularization)	100%	80.1%	67.6%
0.001	86%	81.3%	60%
0.005	74%	70.2%	56%
0.01	34%	32.1%	50.6%

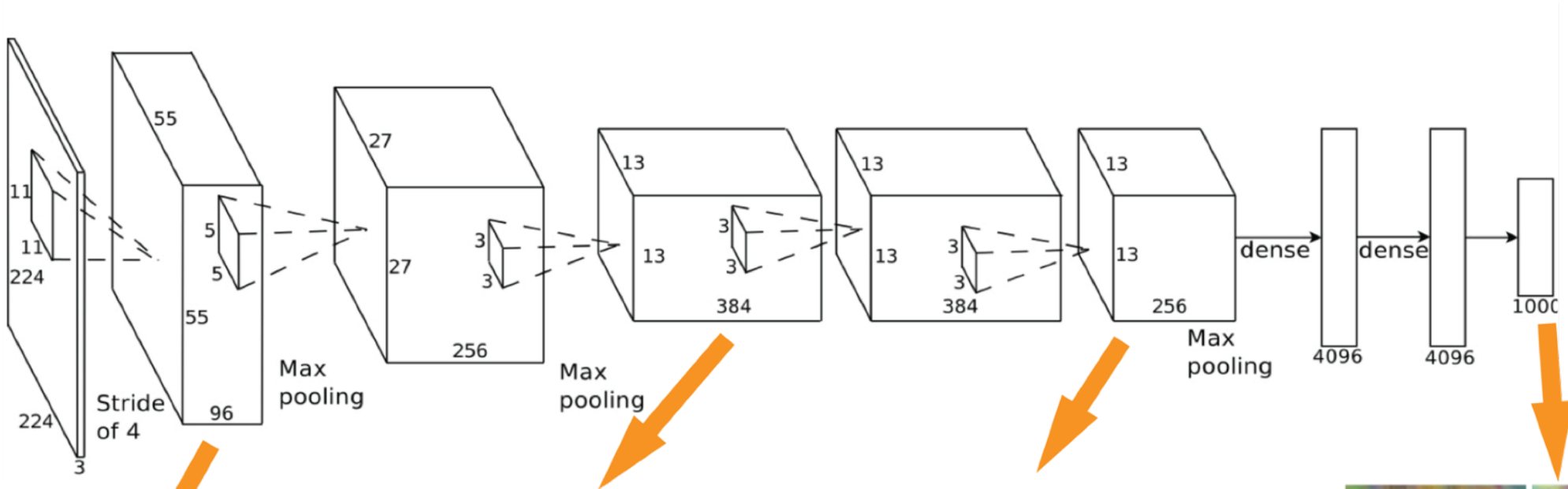
High cost for same privacy

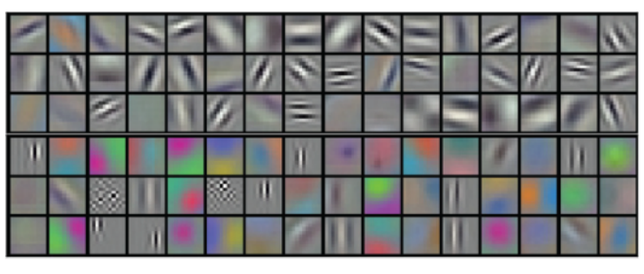
Low Training accuracy, Smaller gap between the training & testing accuracies

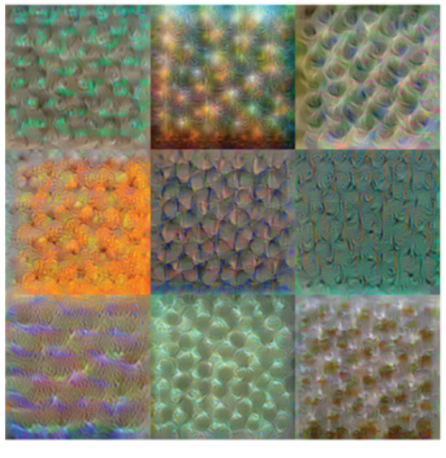
White-box Privacy Analysis

What if the adversary observes
the model parameters?

Extending Black-box Inference Attack to Activation Functions?

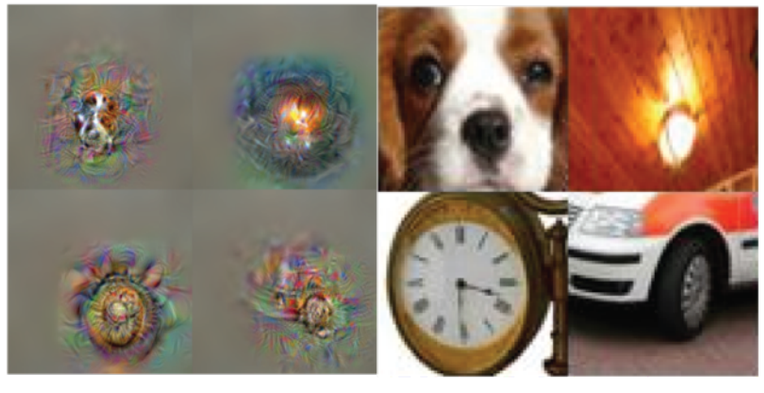




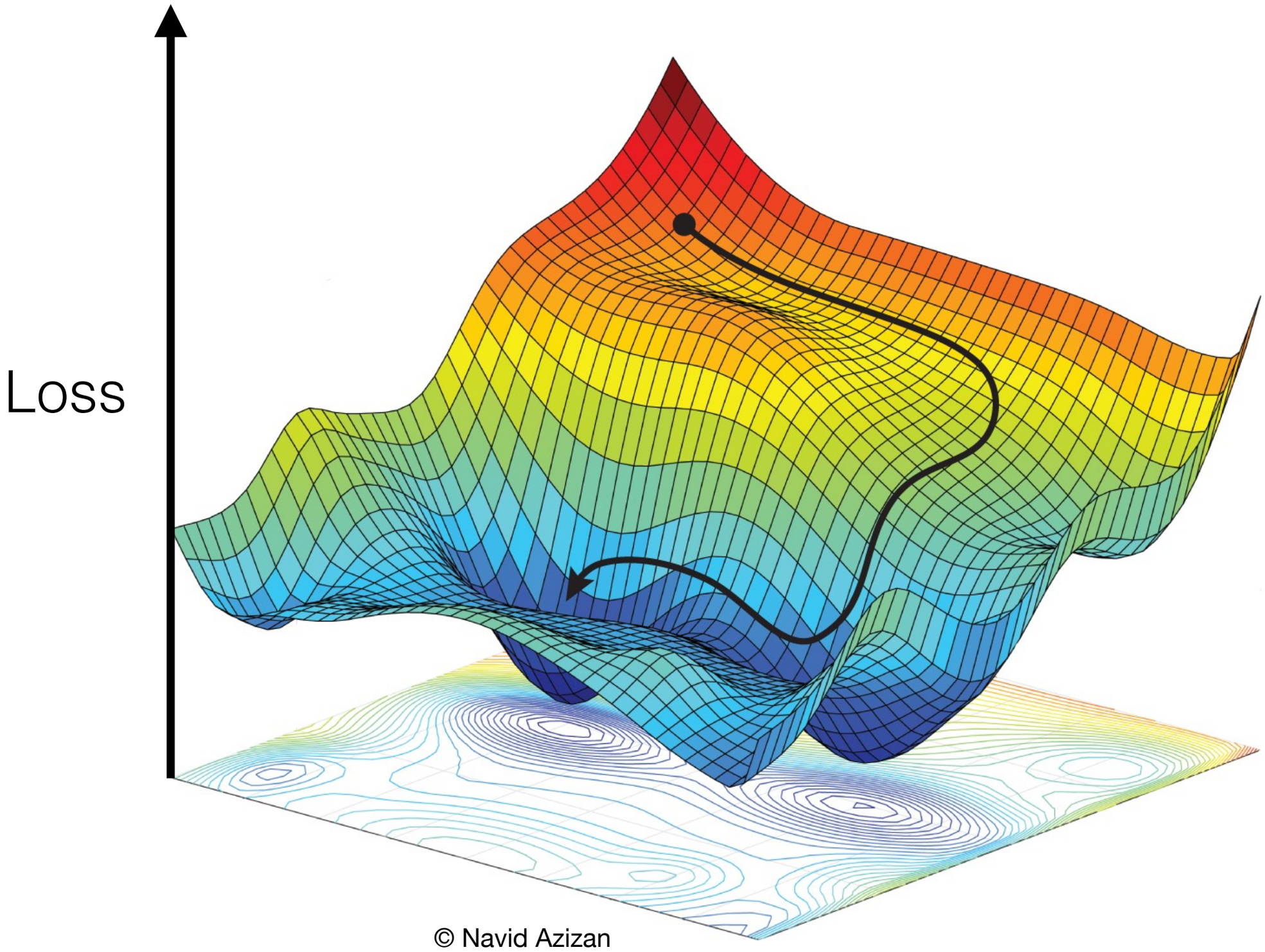


Numerical

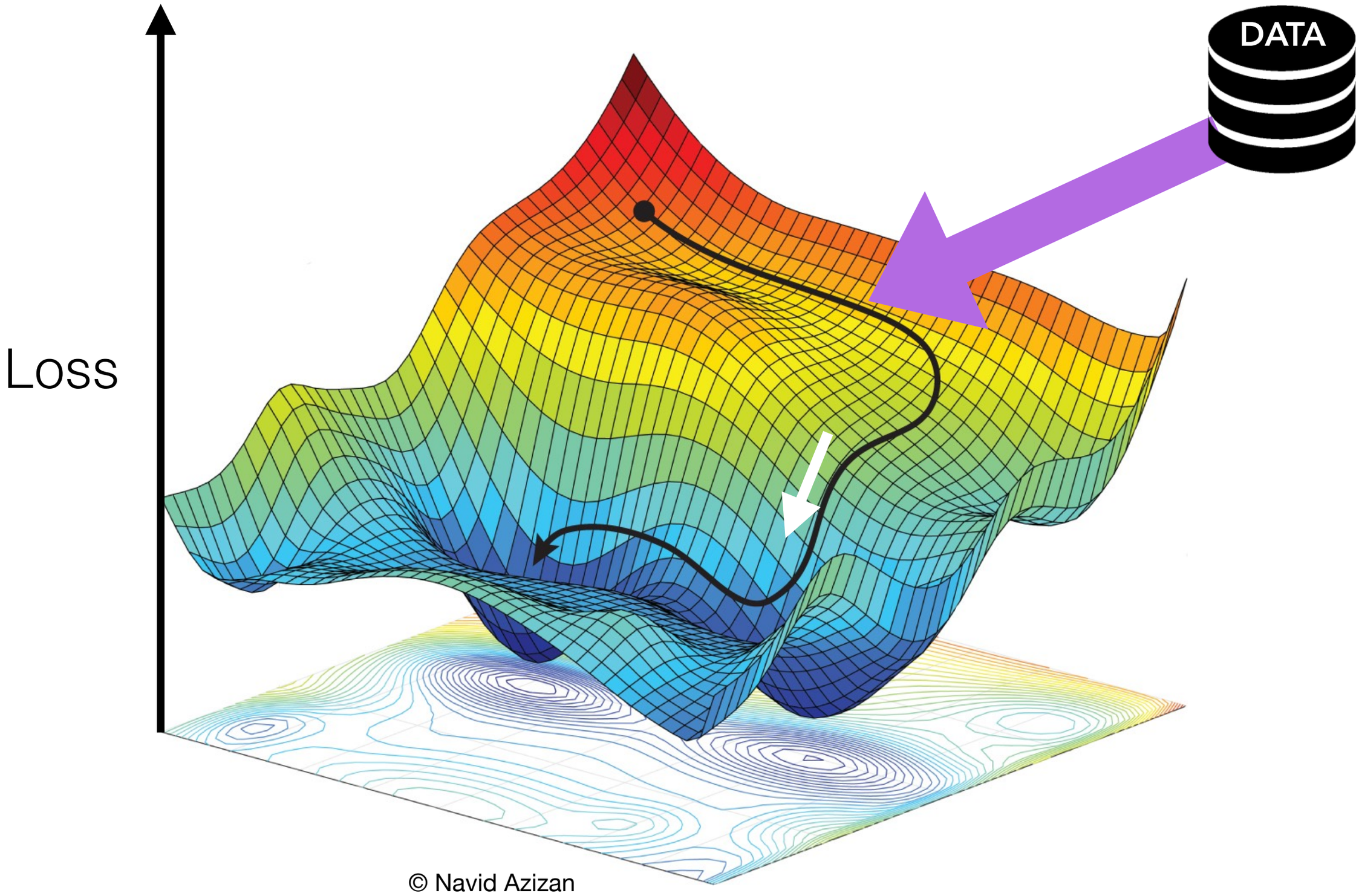
Data-driven



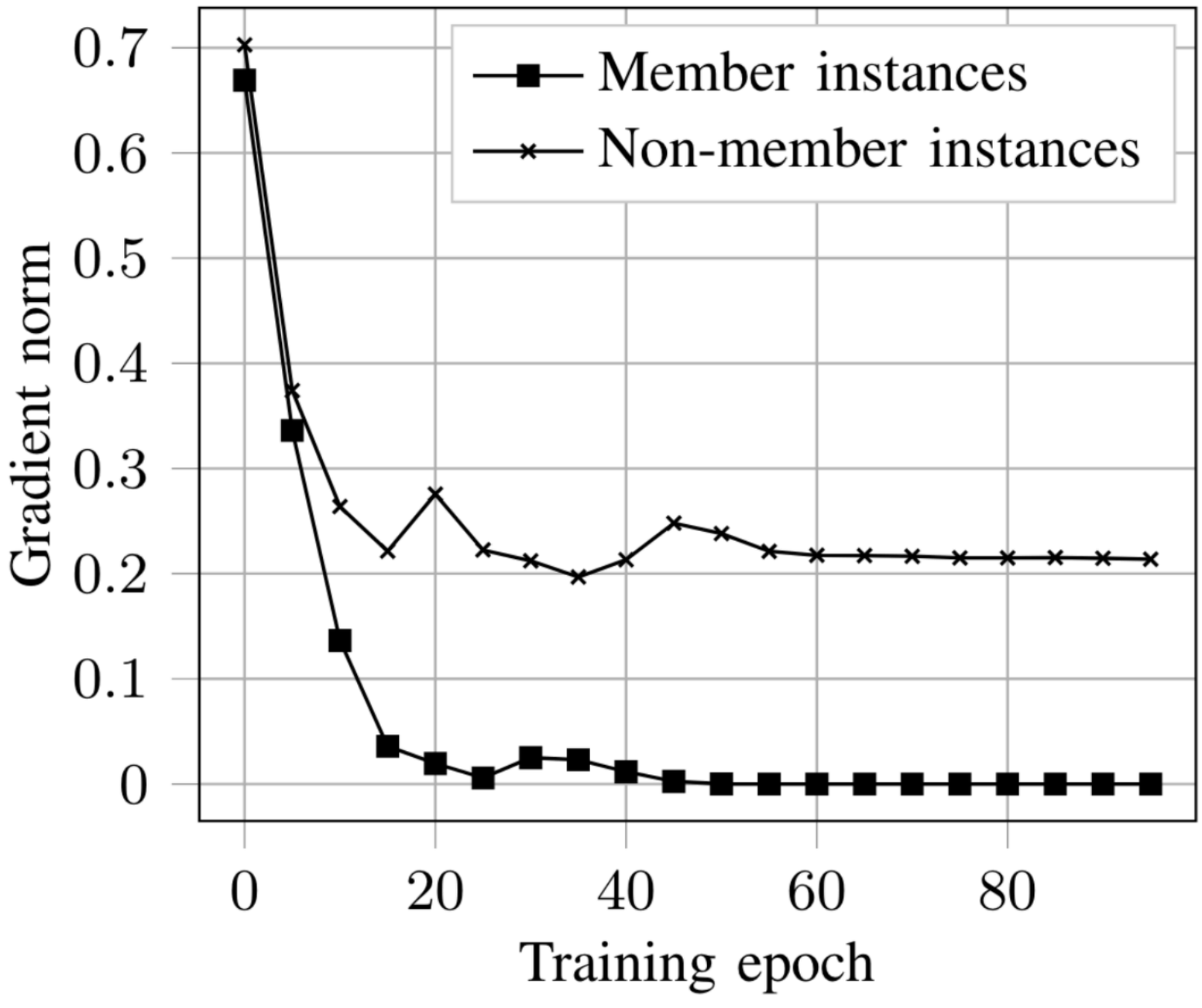
Stochastic Gradient Descent



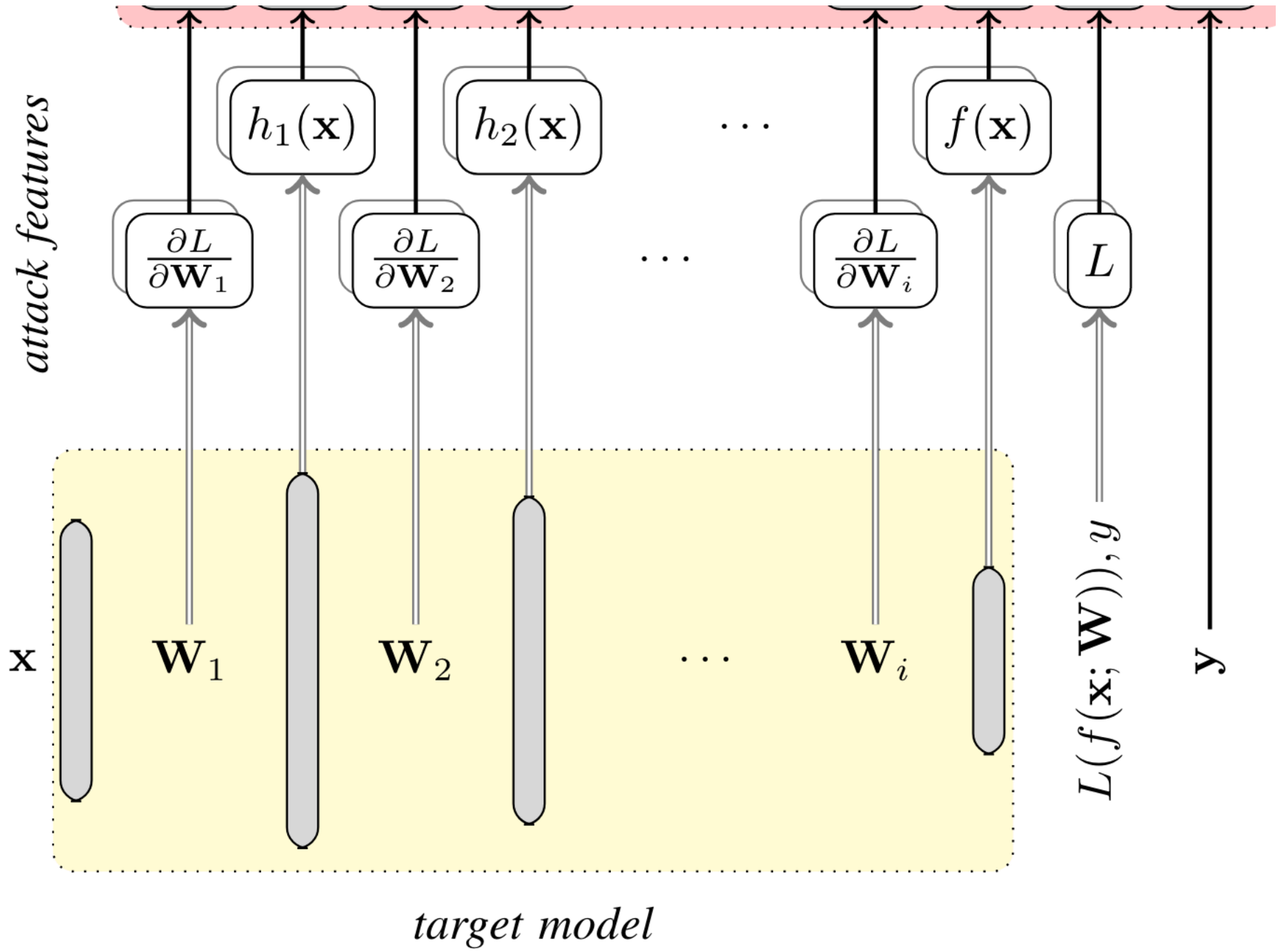
Stochastic Gradient Descent



Gradient of Loss on Members vs. Non-members



Attack Features



Generalizability and (?!) Privacy

in the white-box setting

Pre-trained 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%

High generalizability
(Best available models)

Low privacy
(Significant leakage through parameters)

Generalizability and (?!) Privacy

in the white-box setting

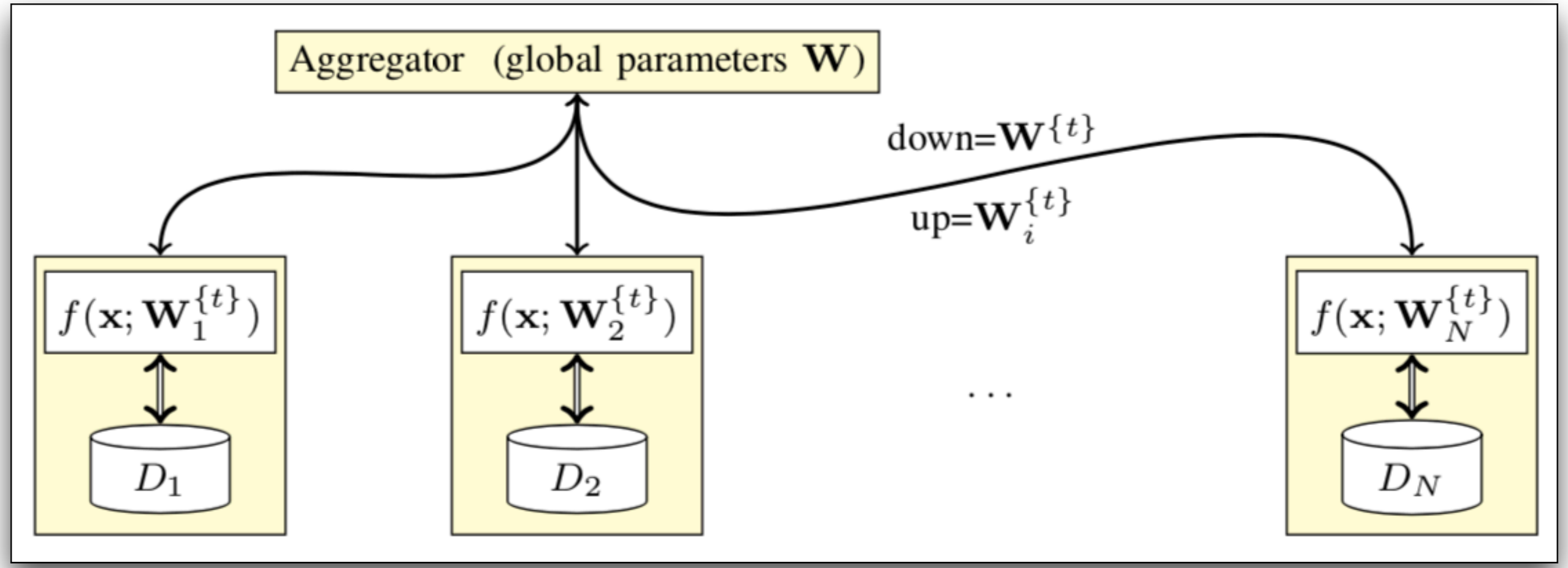
Pre-trained Target Model				Attack Accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%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

Large capacity

High generalizability (Best available models)

Low privacy (Significant leakage through parameters)

Federated Learning



Federated Learning

- Multiple updates (every epoch)

Observed Epochs	Attack Accuracy
5, 10, 15, 20, 25	57.4%
10, 20, 30, 40, 50	76.5%
50, 100, 150, 200, 250	79.5%
100, 150, 200, 250, 300	85.1%

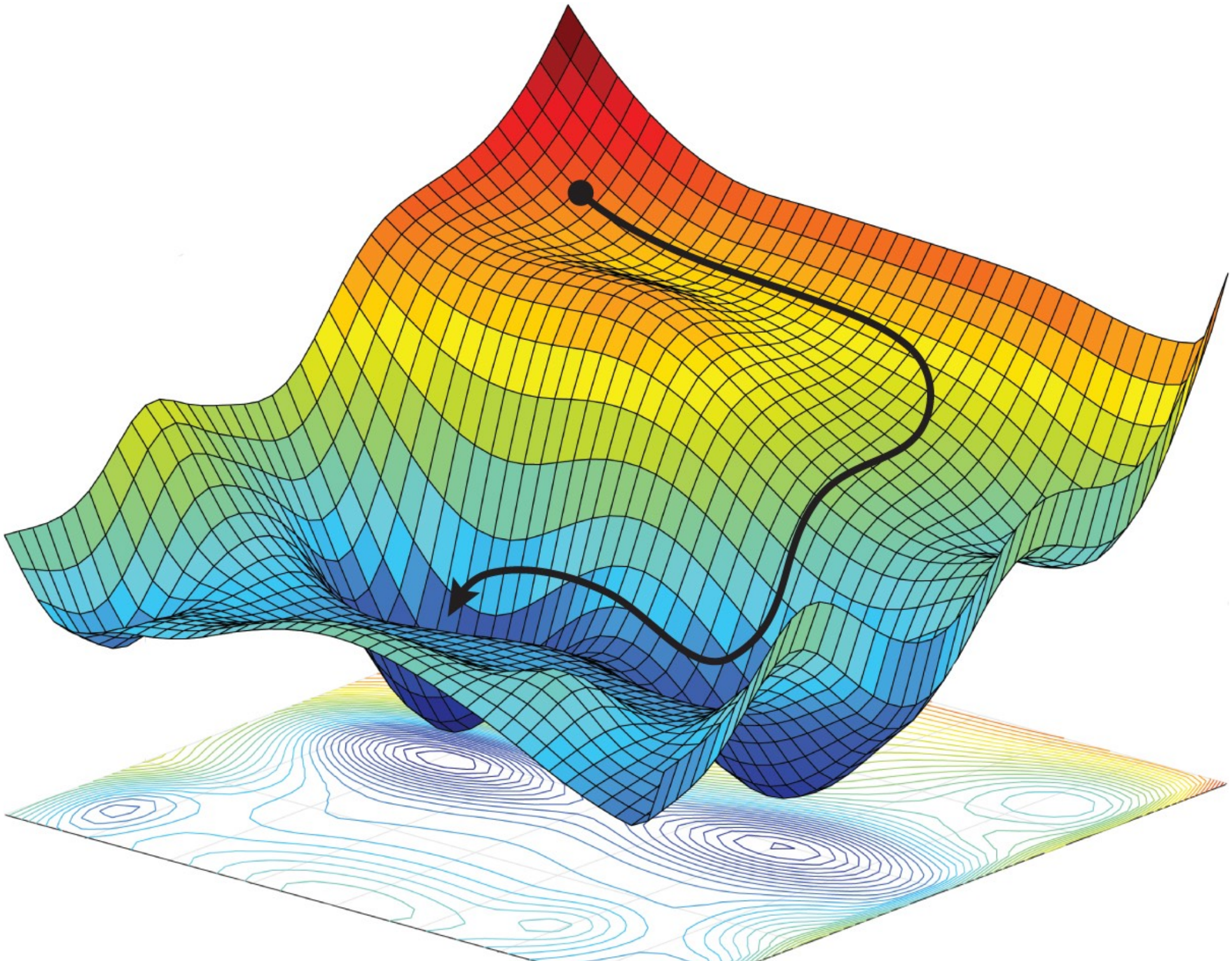
CIFAR100-Alexnet, Global Attack

Federated Learning

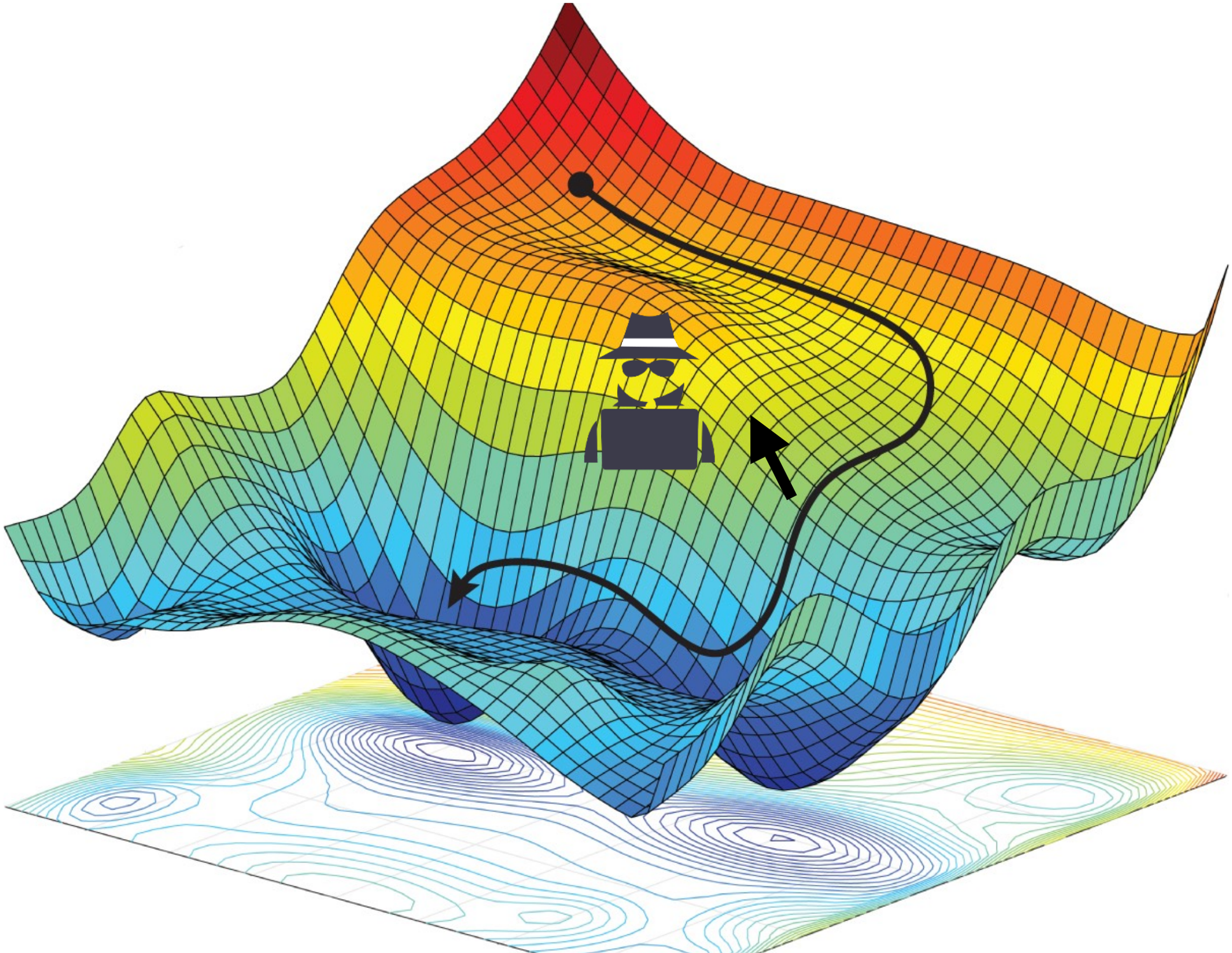
- Parameters from all parties are aggregated and sent back to all parties

- One can influence others' models

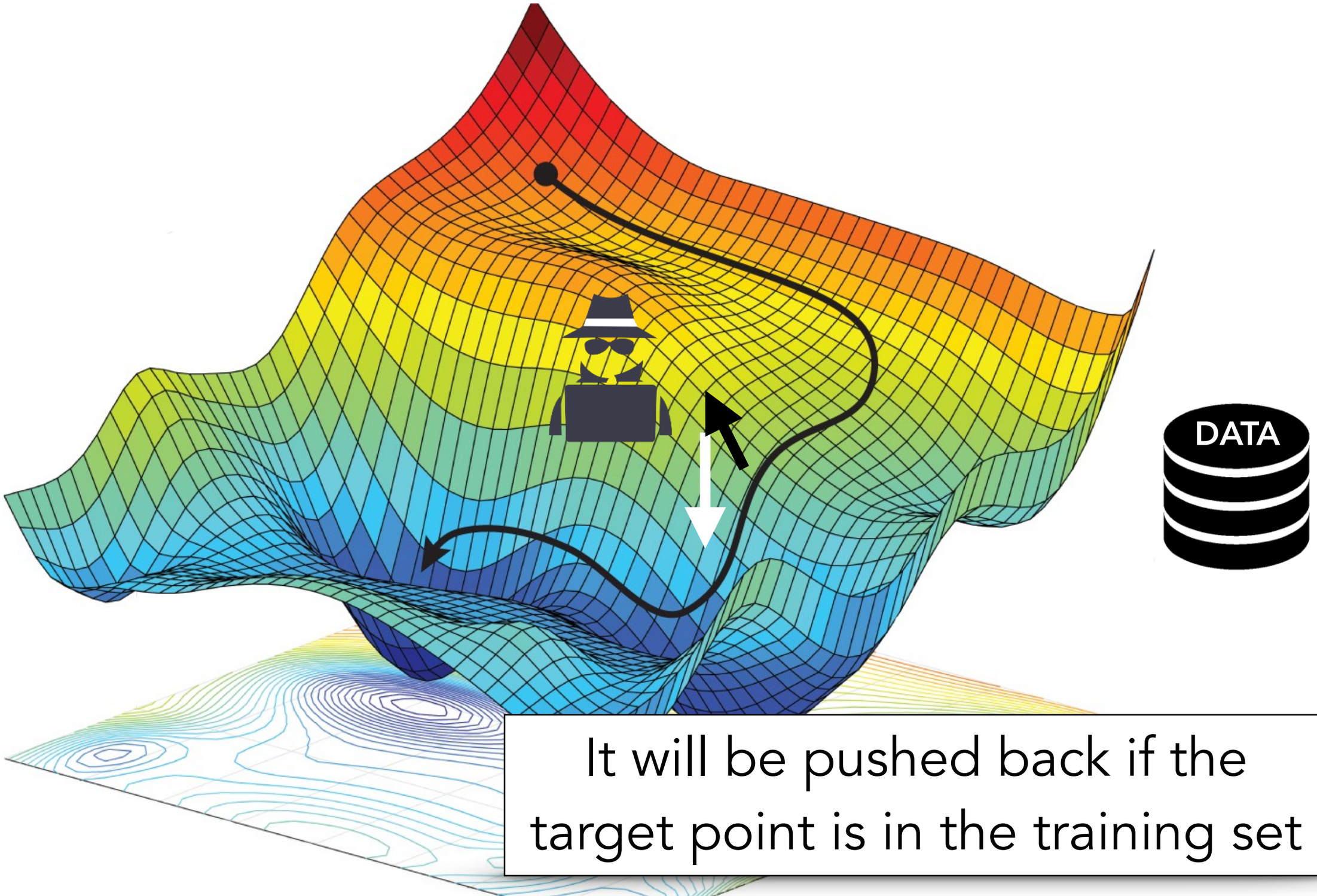
Gradient Descent



Active Attack: Gradient "Ascent" on a Target Data Point



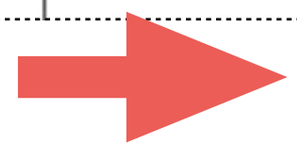
Active Attack: Gradient "Ascent" on a Target Data Point



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

Active Attacks

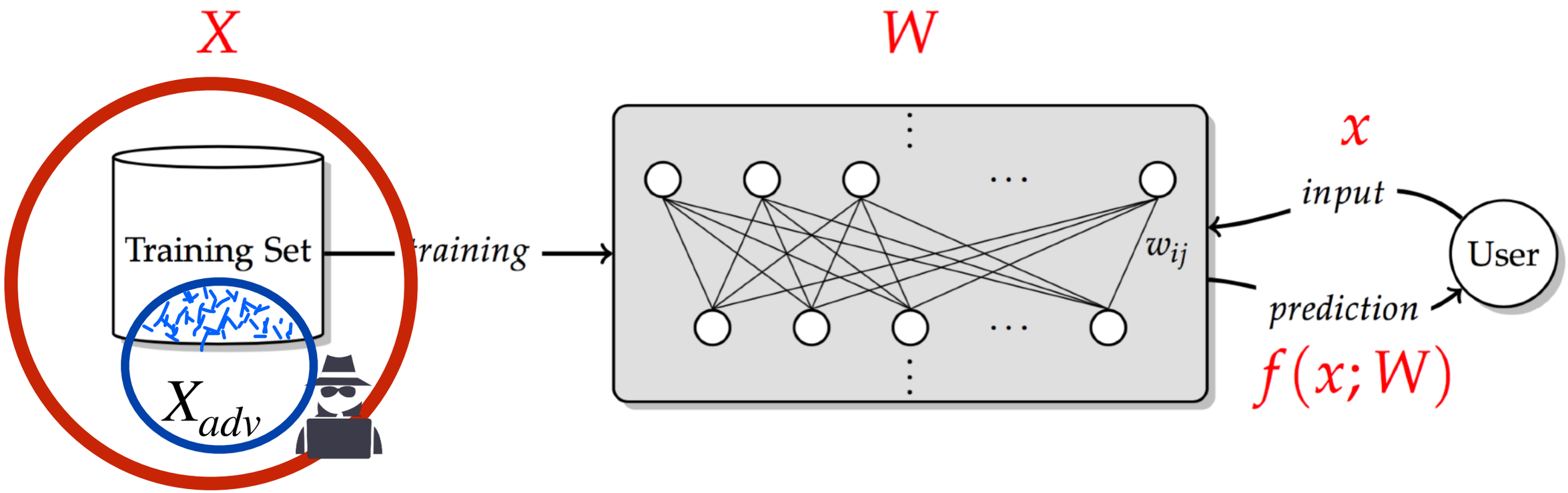
Model	Global Attacker (the parameter aggregator)				Local Attacker (a participant)	
CIFAR100	Passive	Active			Passive	Active
Architecture		Gradient Ascent	Isolating	Isolating Gradient Ascent		Gradient Ascent
Alexnet	85.1%	88.2%	89.0%	92.1%	73.1%	76.3%
DenseNet	79.2%	82.1%	84.3%	87.3%	72.2%	76.7%



Robustness in Machine Learning

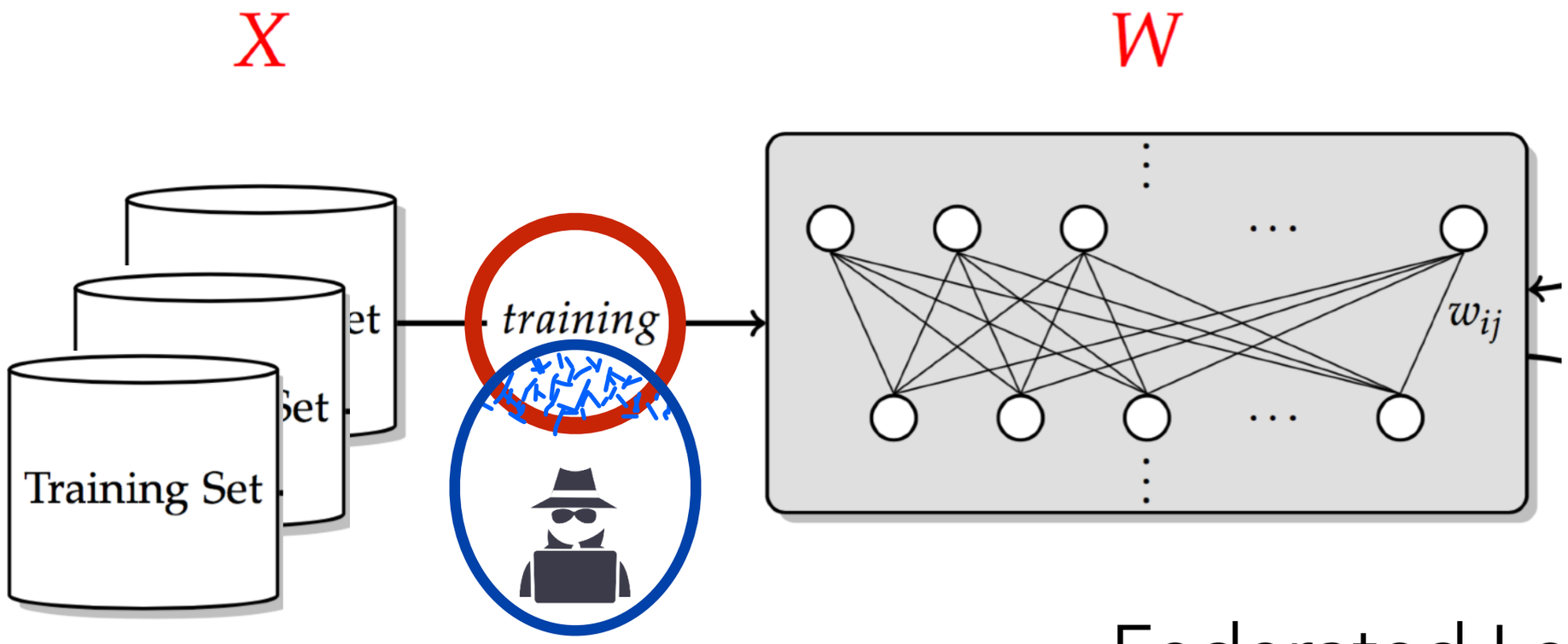
Song, Shokri, Mittal, “Privacy risks of securing machine learning models against adversarial examples”

Robustness Threats



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

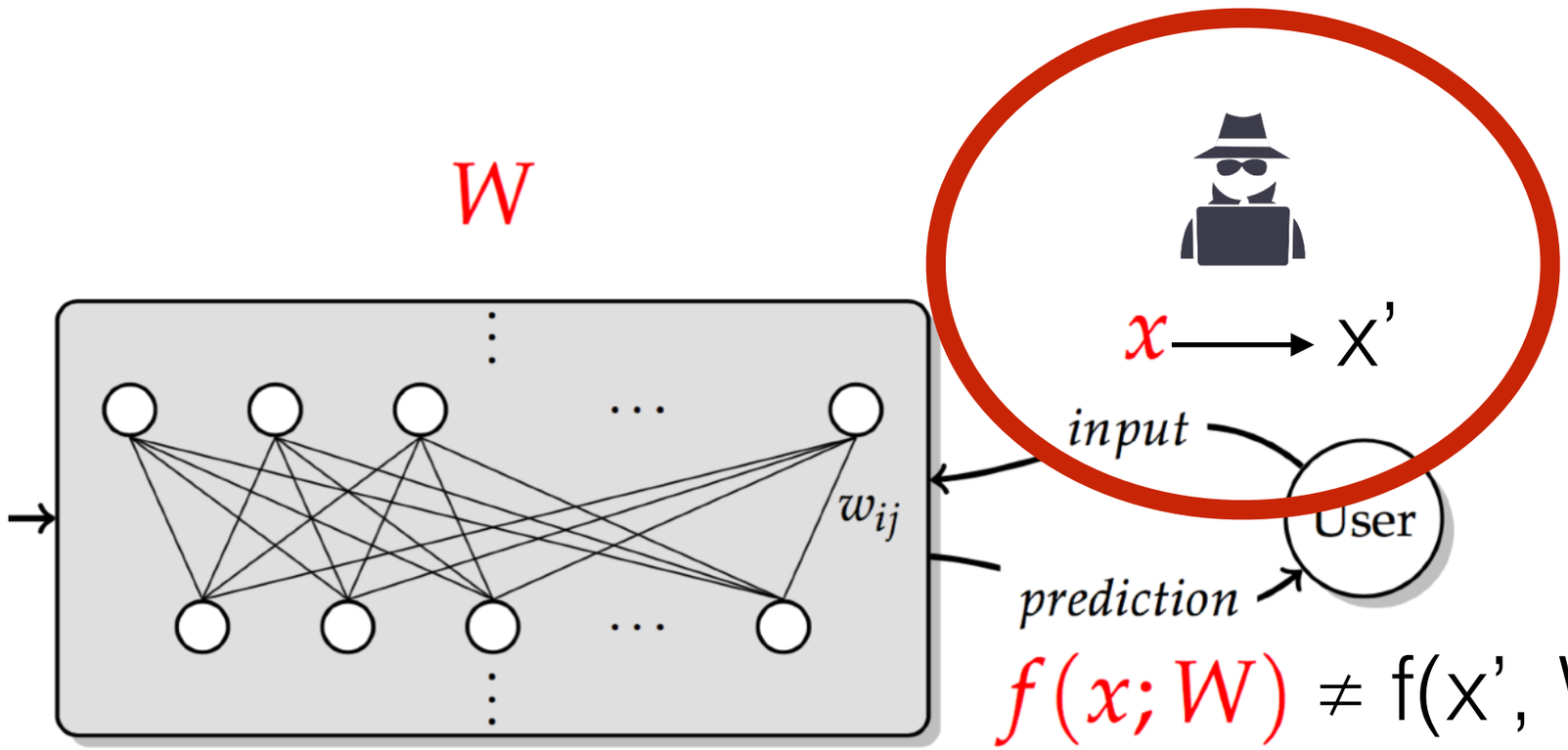
Robustness Threats



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

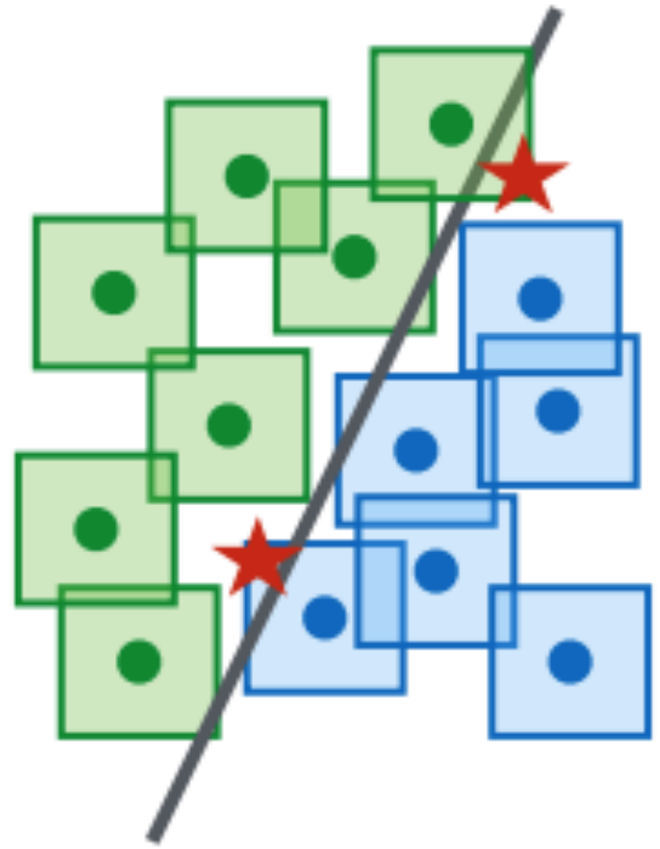
Federated Learning: Distributed aggregation of local models

Robustness Threats

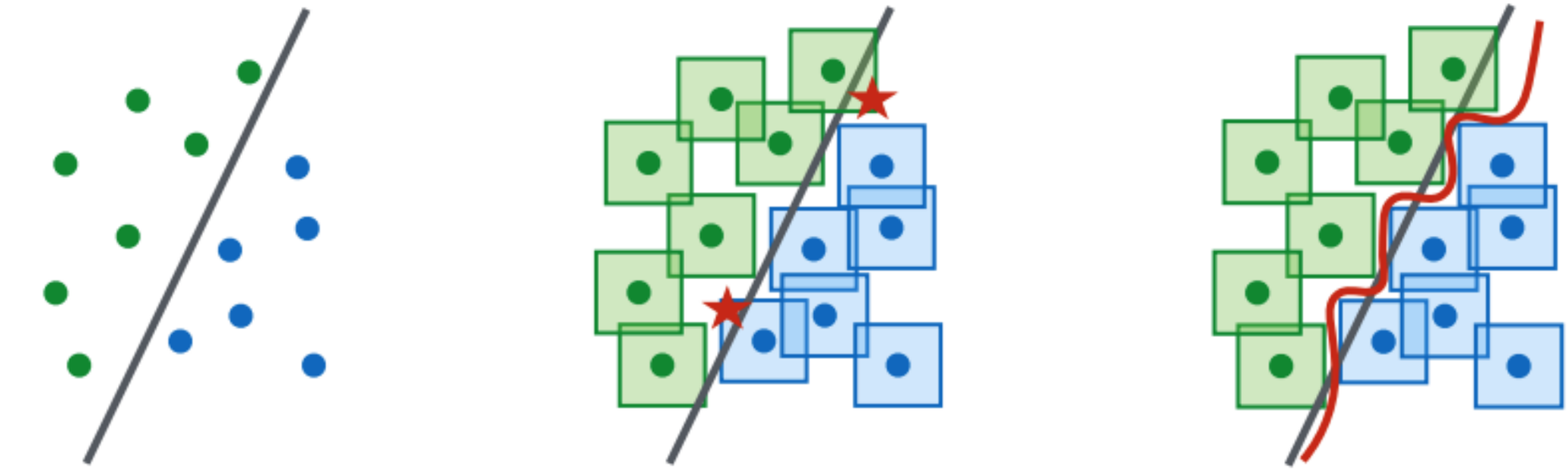


Adversarial Input (evasion) attacks:
Imperceptible perturbations of input to cause misclassification

Adversarial Inputs: Attack



Adversarial Inputs: Defense

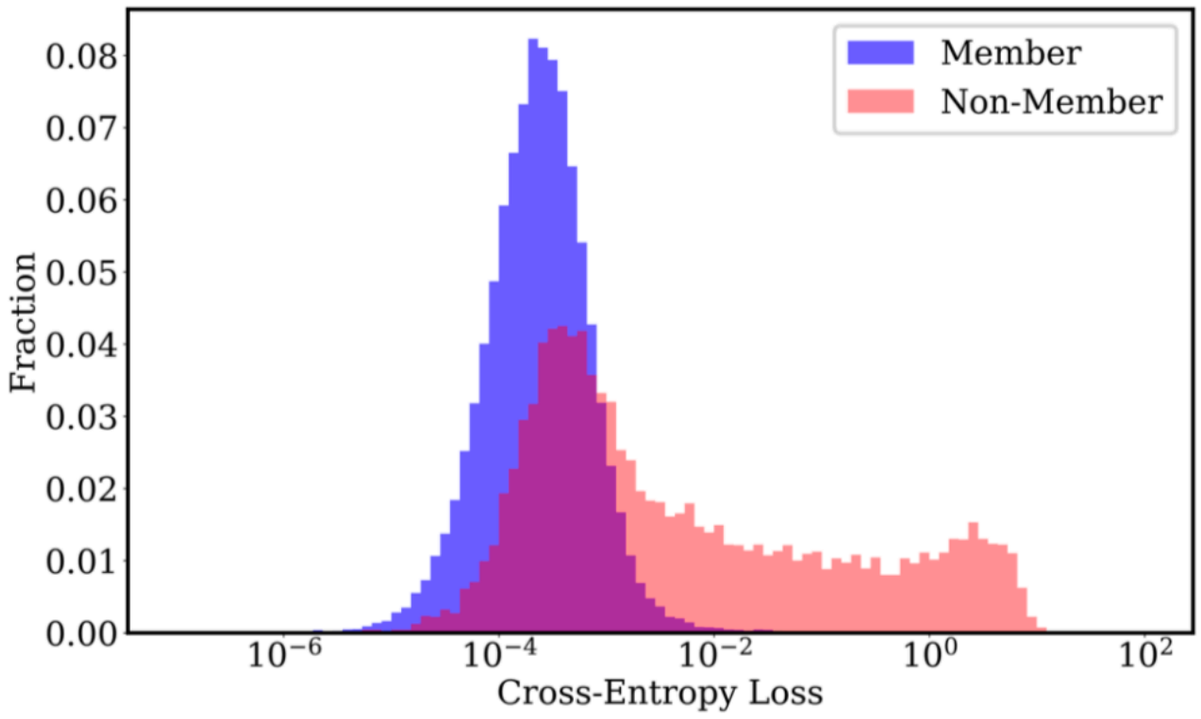


© Aleksander Madry

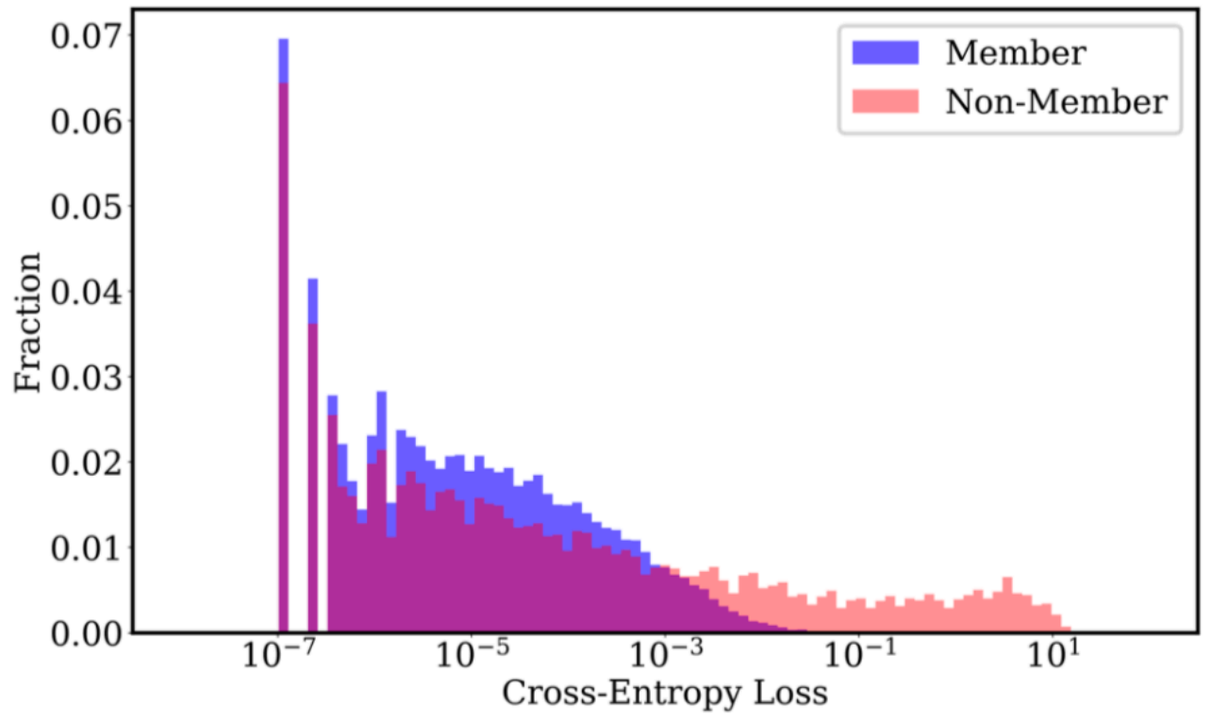
Minimize the loss on adversarial perturbations

- empirical method (train on adv points)
- theoretical (compute upper bound on adv loss)

Consequences



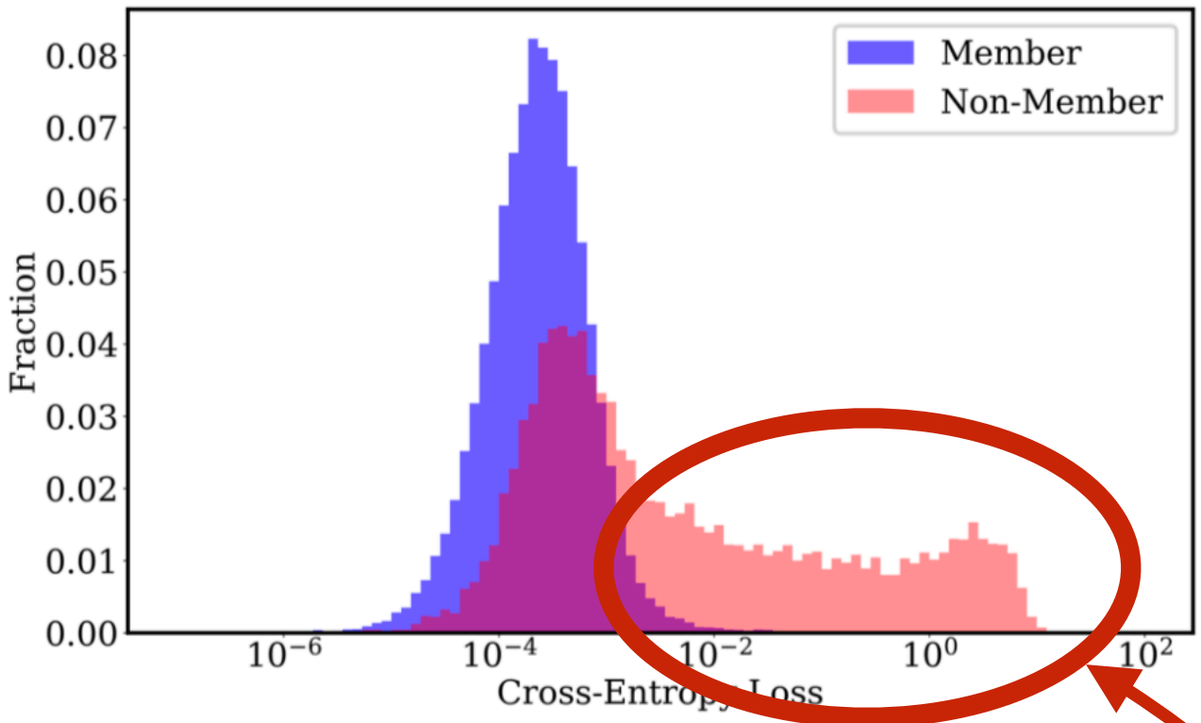
Robust Training



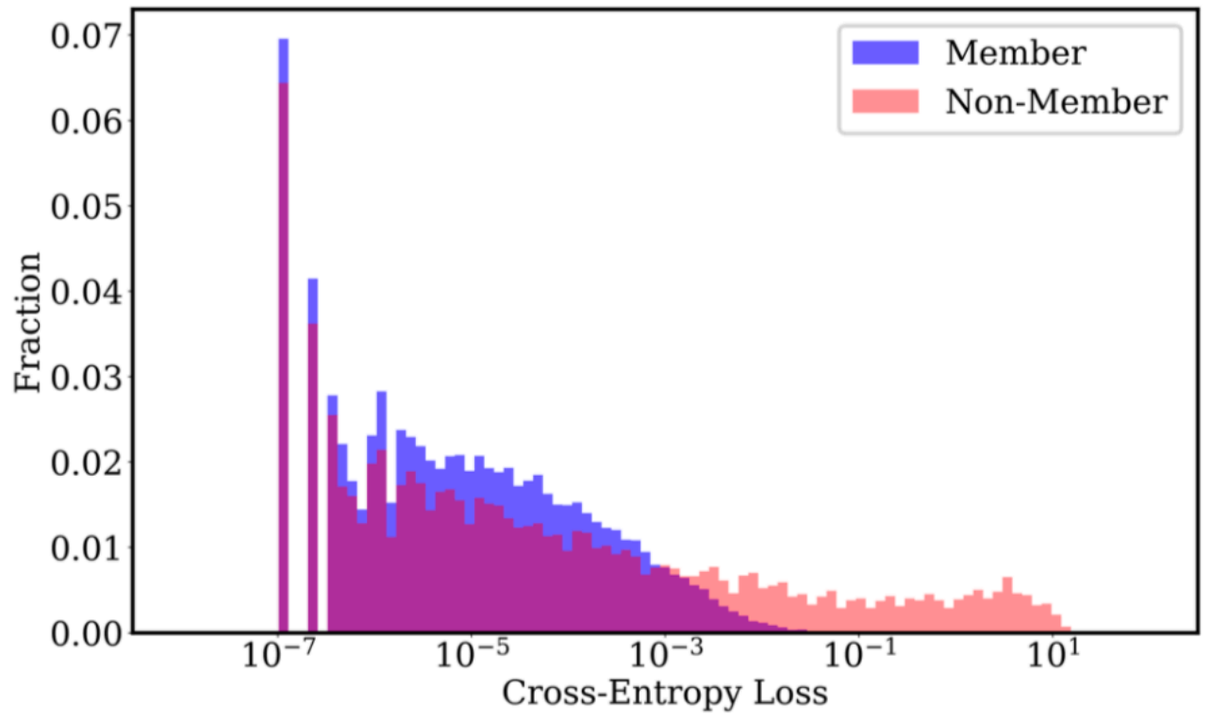
Baseline (standard SGD)

CIFAR10

Consequences



Robust Training



Baseline (standard SGD)

CIFAR10

Training set members are more **distinguishable** from non-members

Inference on Perturbed Points

Not robust

Training method	train acc	test acc	adv-train acc	adv-test acc	inference acc ($\mathcal{I}_{\mathrm{B}}$)	inference acc ($\mathcal{I}_{\mathrm{A}}$)
Natural	100%	95.01%	0%	0%	**57.43%**	50.85%

Inference on Perturbed Points

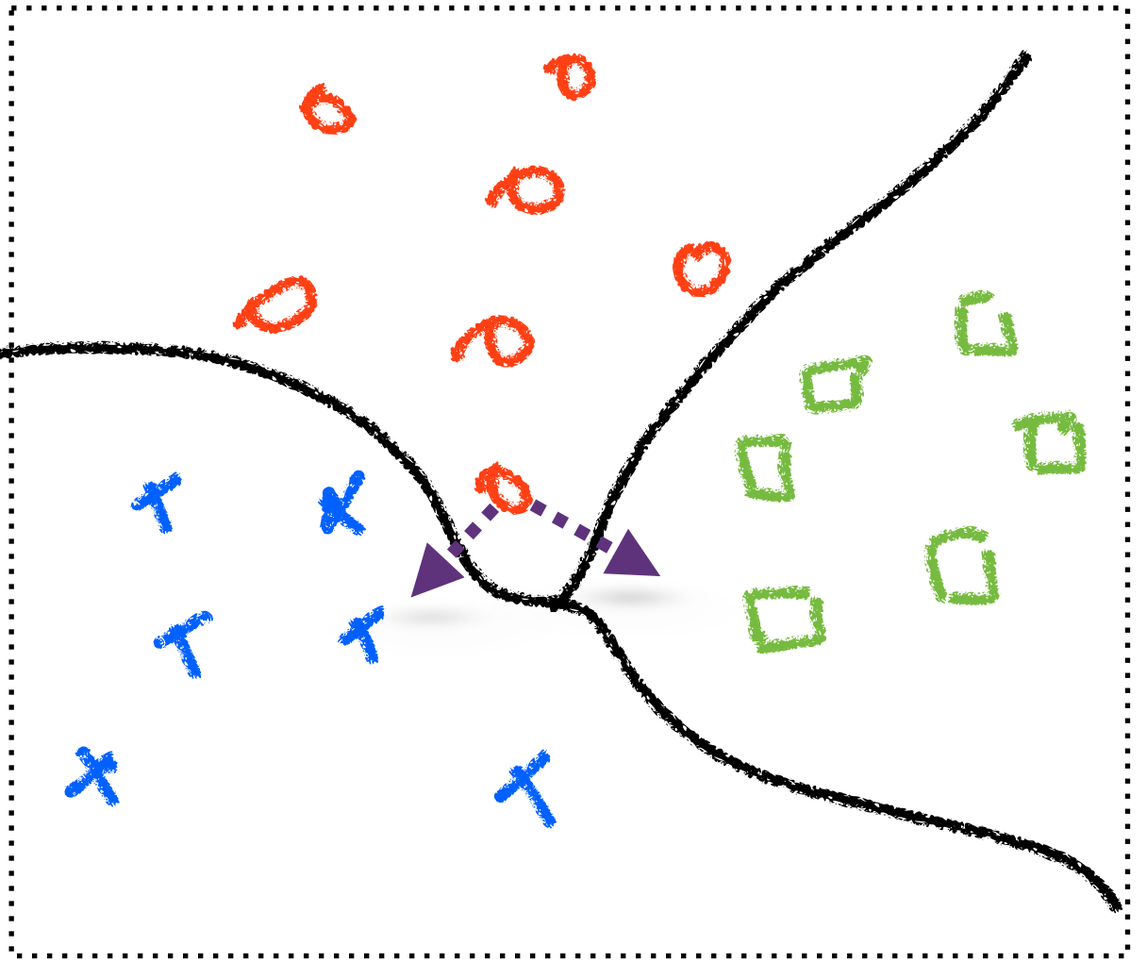
- Lower loss on adversarial perturbations of training data

Not robust

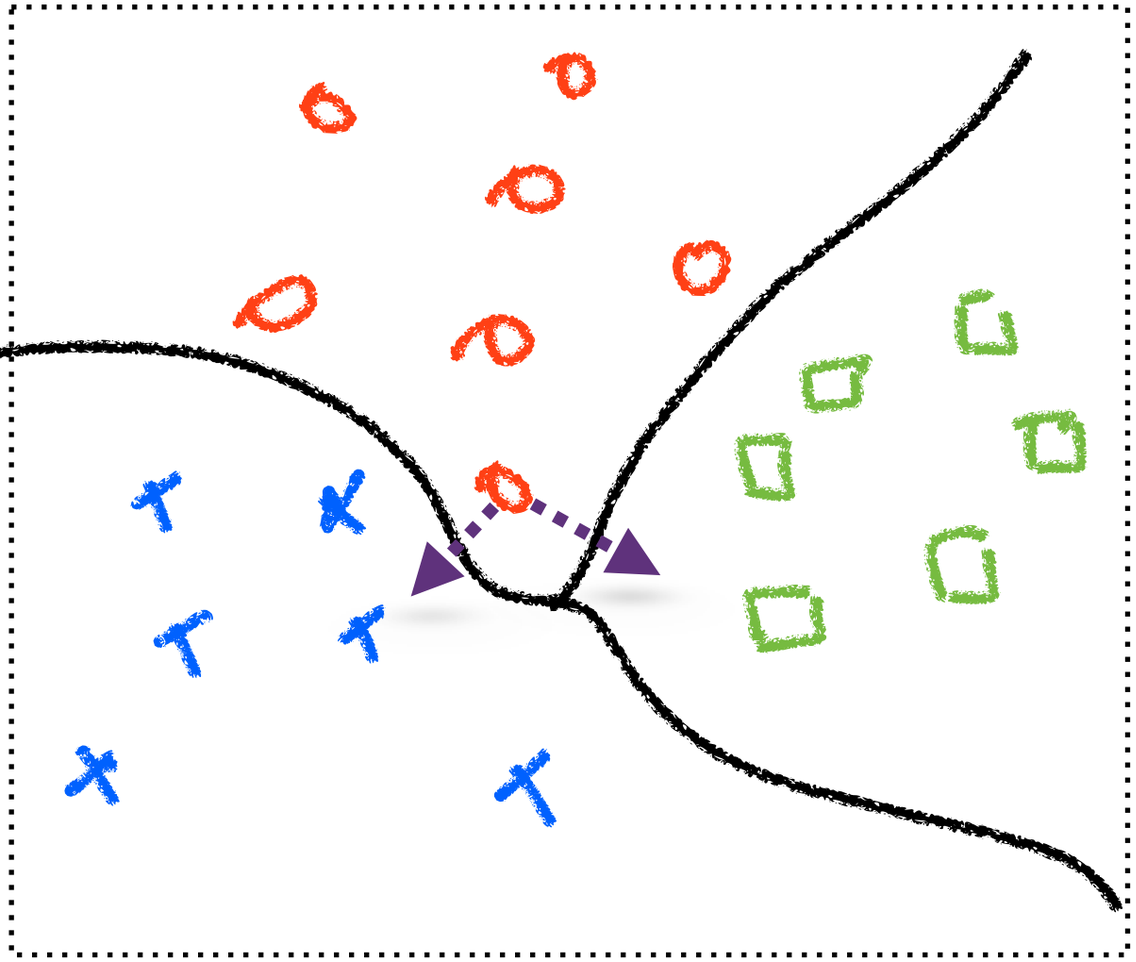
Training method	train acc	test acc	adv-train acc	adv-test acc	inference acc ($\mathcal{I}_B$)	inference acc ($\mathcal{I}_A$)
Natural	100%	95.01%	0%	0%	**57.43%**	50.85%
PGD-Based Adv-Train [30]	99.99%	87.25%	96.07%	46.59%	74.89%	**75.65%**

Not privacy-preserving

Fine-grained Fingerprinting: Targeted Adversarial Points



Fine-grained Fingerprinting: Targeted Adversarial Points



label	(untargeted)	(targeted)
0	72.70%	**74.42%**
1	67.69%	**68.88%**
2	80.16%	**83.58%**
3	87.83%	**90.57%**
4	81.57%	**84.47%**
5	81.34%	**83.02%**
6	76.97%	**79.94%**

...

Perturbation Budget

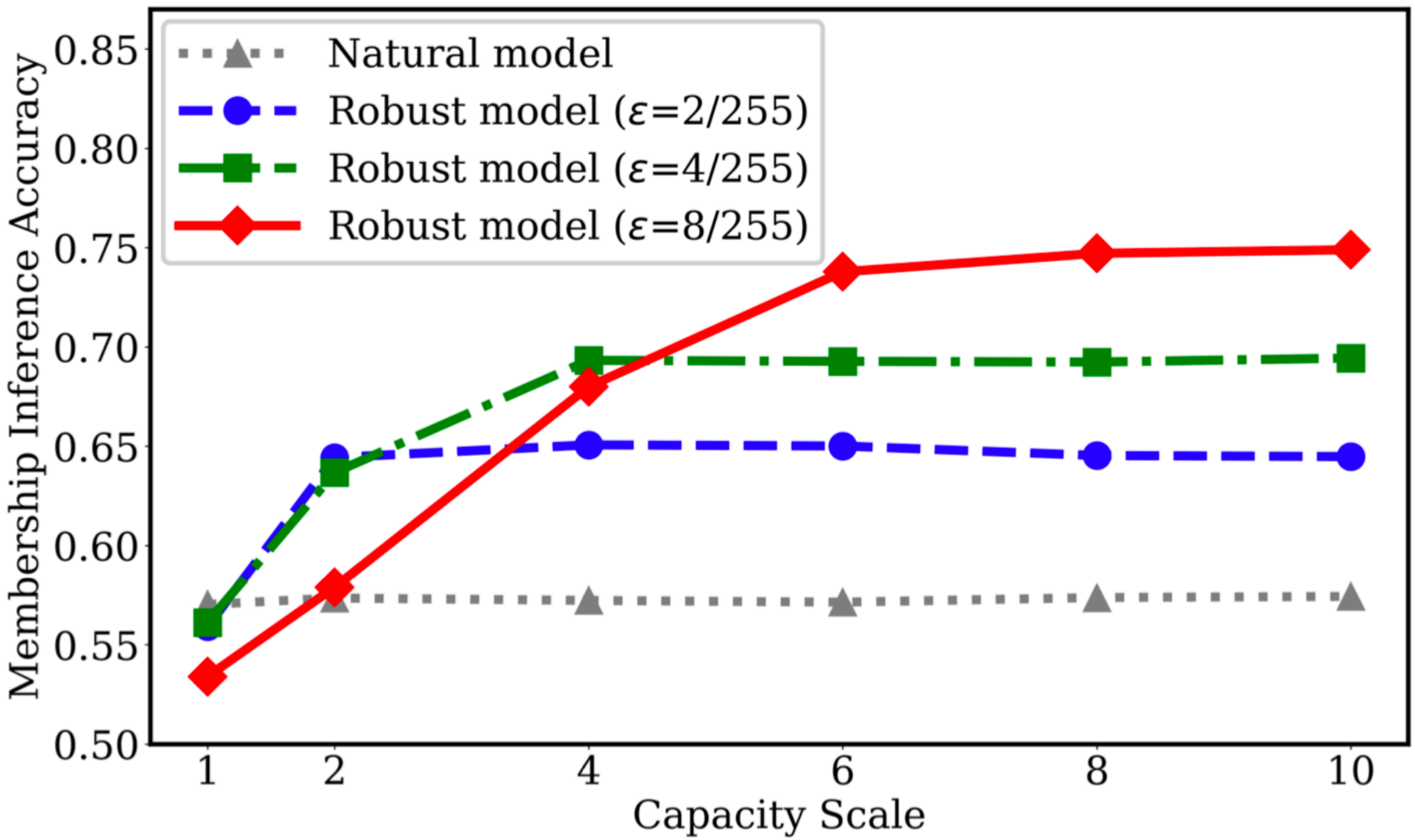
Perturbation budget (ϵ)	train acc	test acc	adv-train acc	adv-test acc	inference acc ($\mathcal{I}_\mathrm{B}$)	inference acc ($\mathcal{I}_\mathrm{A}$)
2/255	100%	93.74%	99.99%	82.20%	64.48%	**66.54%**
4/255	100%	91.19%	99.89%	70.03%	69.44%	**72.43%**
8/255	99.99%	87.25%	96.07%	46.59%	74.89%	**75.65%**

Higher perturbation budget

Lower Privacy

Model Capacity

- Resnet: Increase the number of output channels of the residual layers

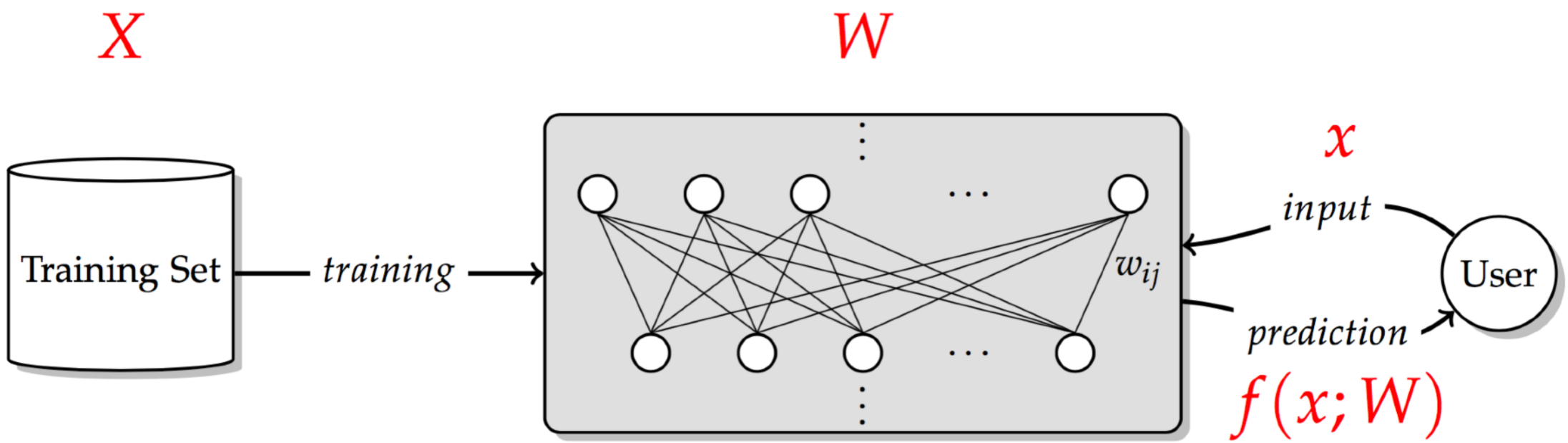


Privacy vs Robustness vs Accuracy

- Limited capacity
  - ✳ Low predictive power
  - ✳ Low leakage
- Robustness
  - ✳ Needs large capacity

Interpretability in Machine Learning

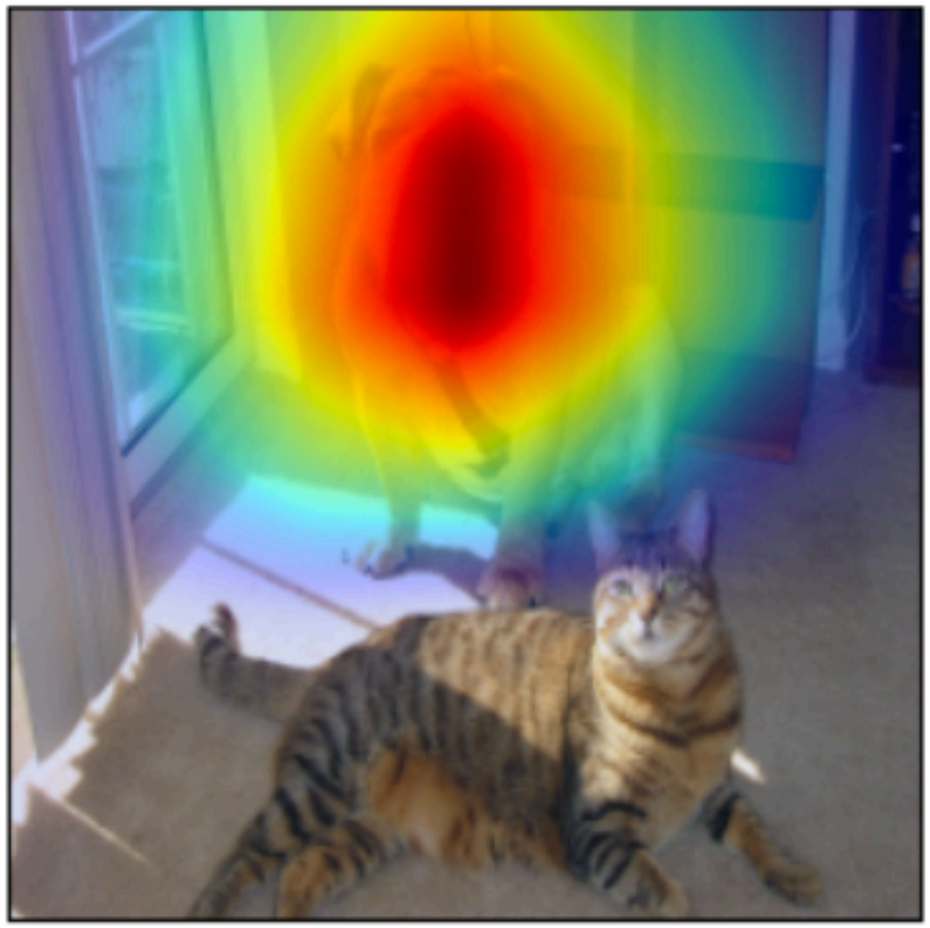
Interpretability



Explain the Output of f on x:

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

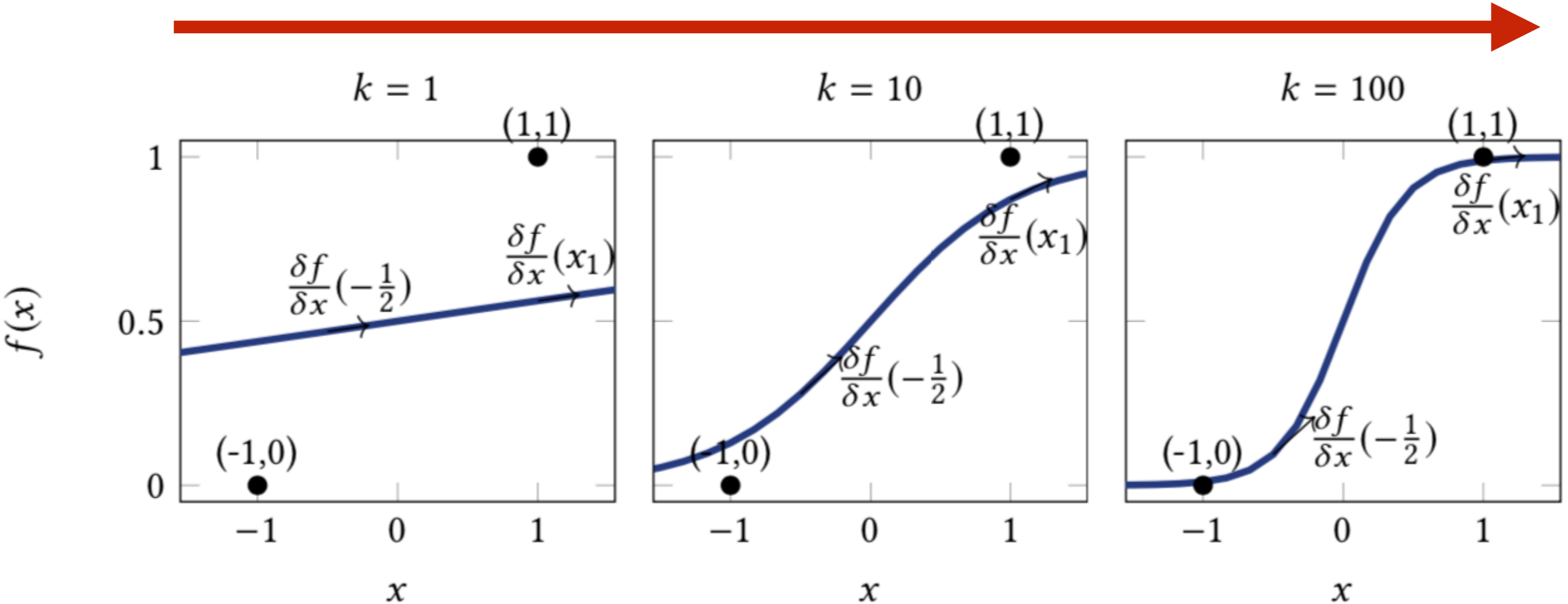
Feature-based Explanation



Compute the influence of each input feature on the output

Explain using Gradients

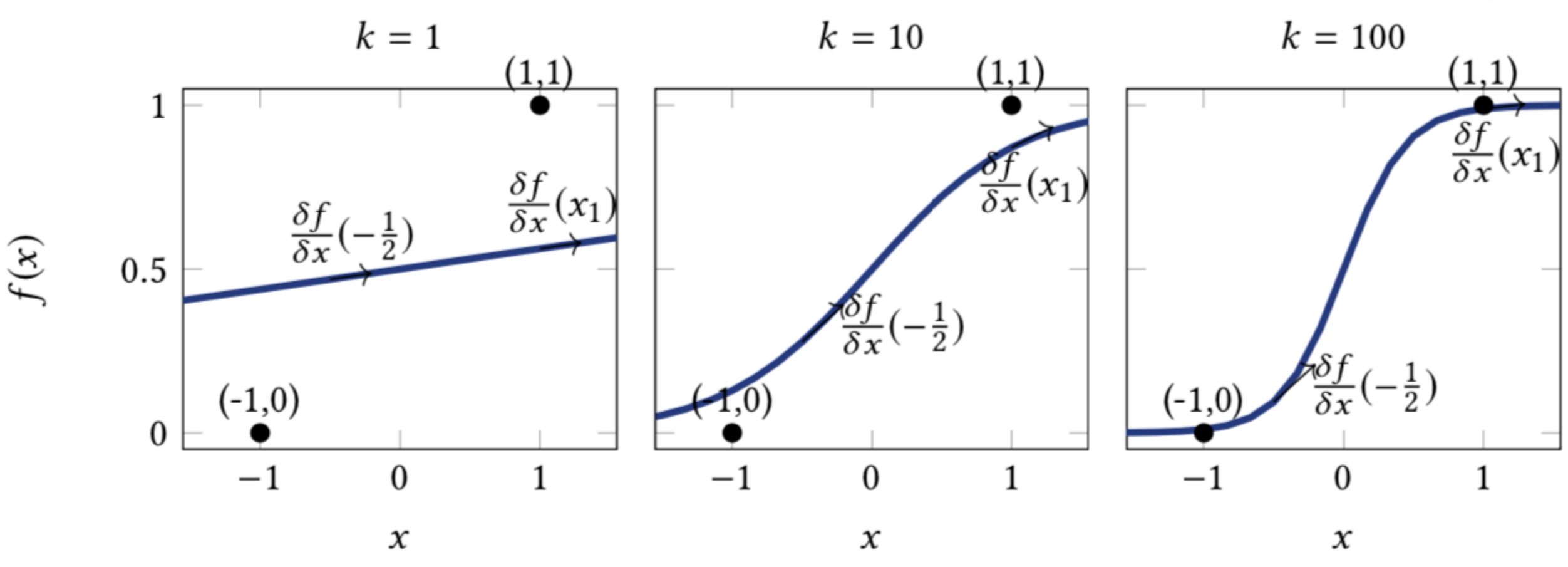
- A simple example: two data points {(-1, 0), (+1,1)}
- One single (sigmoid) activation function. f(x; w) = 1/(1+exp(-wx))



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

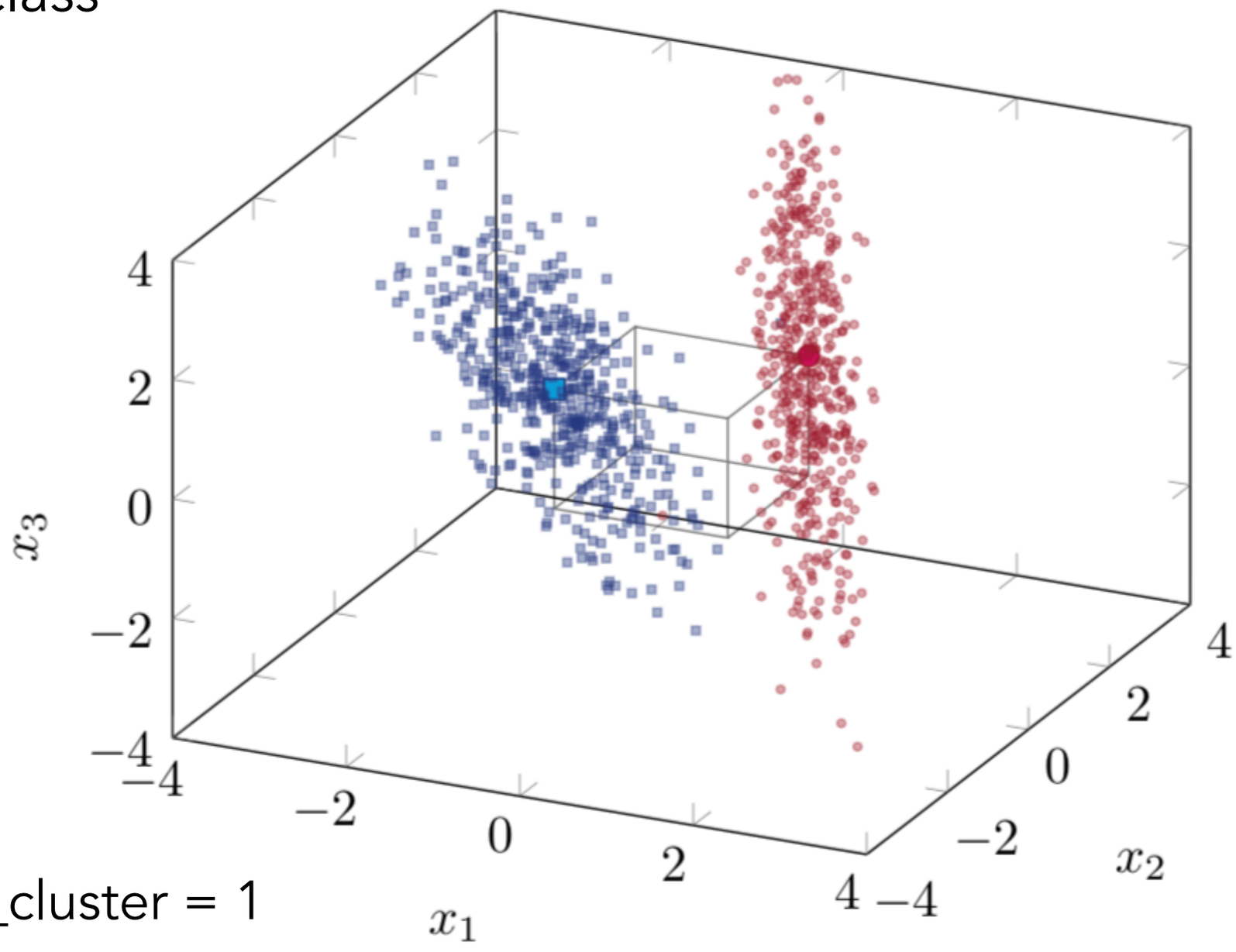
- A simple example: two data points {(-1, 0), (+1,1)}
- One single (sigmoid) activation function. f(x; w) = 1/(1+exp(-wx))

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

k	θ_k	$-\frac{\partial \lvert 1-f_\theta(1)\rvert}{\partial \theta}(\theta_k)$	$f_{\theta_k}(1)$	$\frac{\partial f_{\theta_k}}{\partial x}(1)$	$\frac{\partial f_{\theta_k}}{\partial x}(\frac{1}{2})$
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

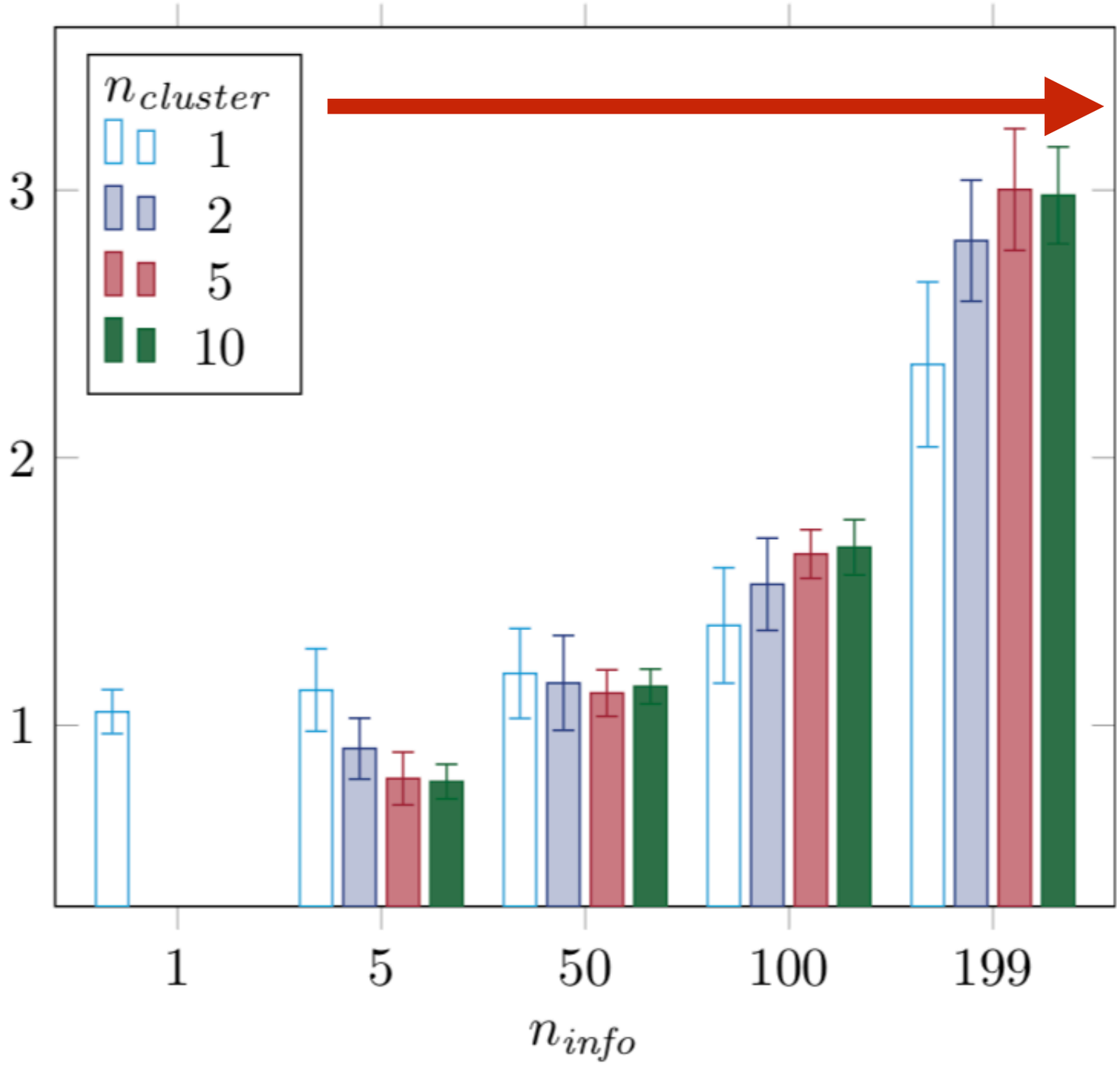
- **Synthetic** data; 2 classes; n_info features; n_cluster clusters in each class



n_info = 3, n_cluster = 1

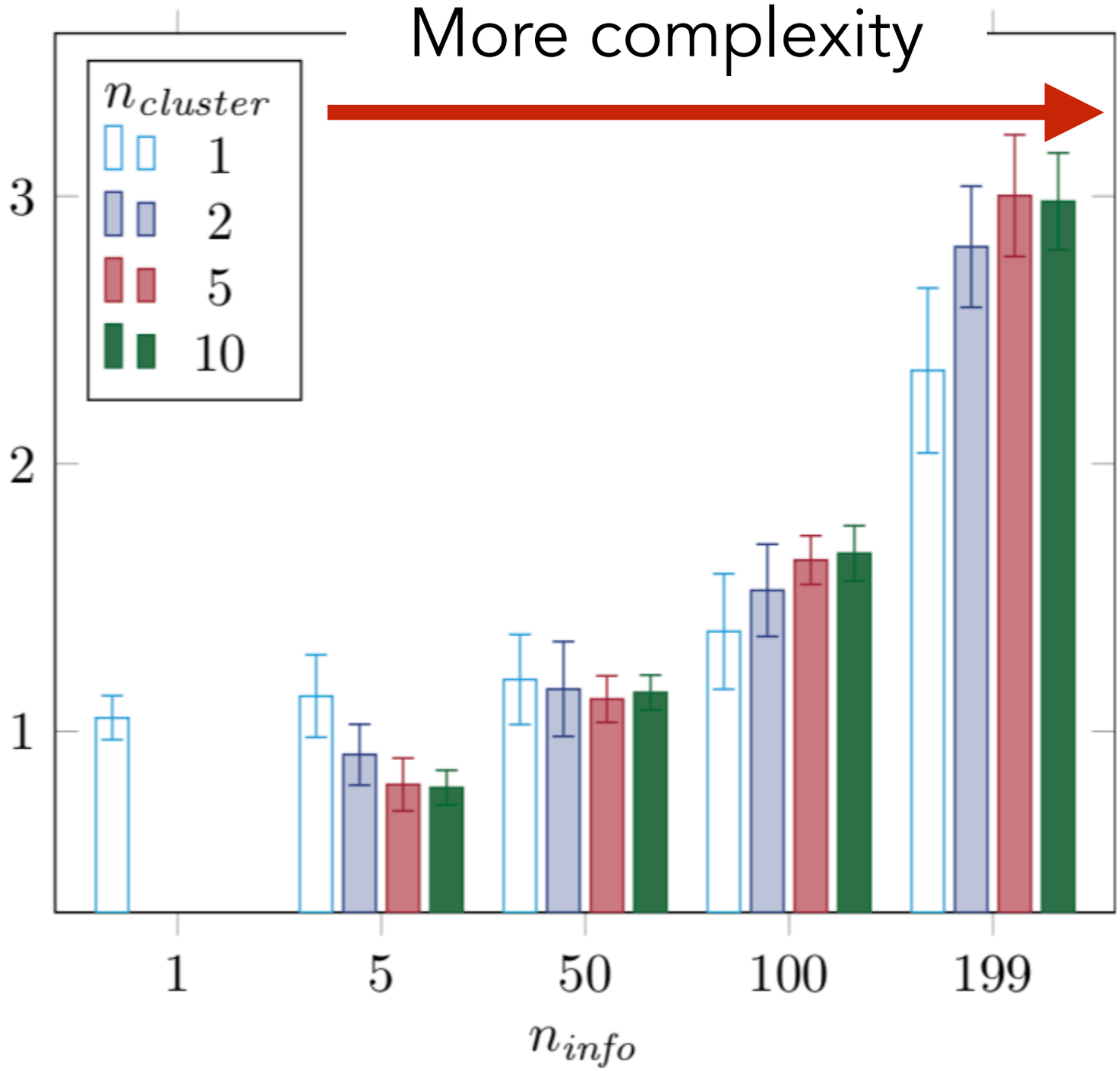
Distinguishability of Members and Non-Members

$$\frac{median\{||\phi_{GRAD}(\vec{x})||_1 | \vec{x} \in \mathcal{X}_{\text{test}}\}}{median\{||\phi_{GRAD}(\vec{x})||_1 | \vec{x} \in \mathcal{X}_{\text{train}}\}}$$



Distinguishability of Members and Non-Members

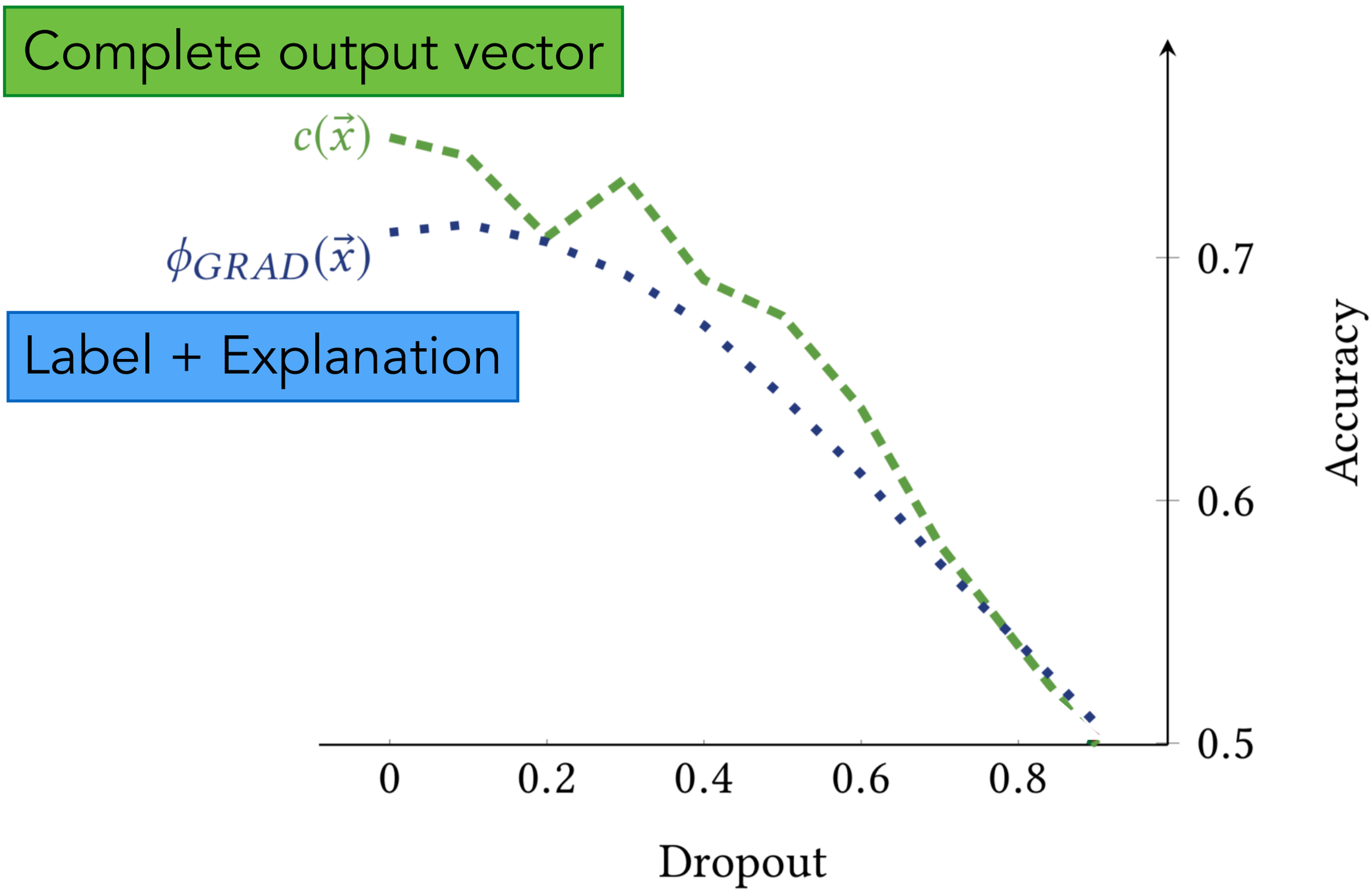
$$\frac{median\{||\phi_{GRAD}(\vec{x})||_1 | \vec{x} \in \mathcal{X}_{\text{test}}\}}{median\{||\phi_{GRAD}(\vec{x})||_1 | \vec{x} \in \mathcal{X}_{\text{train}}\}}$$



Membership Inference Attack

Complete output vector

Label + Explanation



Membership Inference Attack

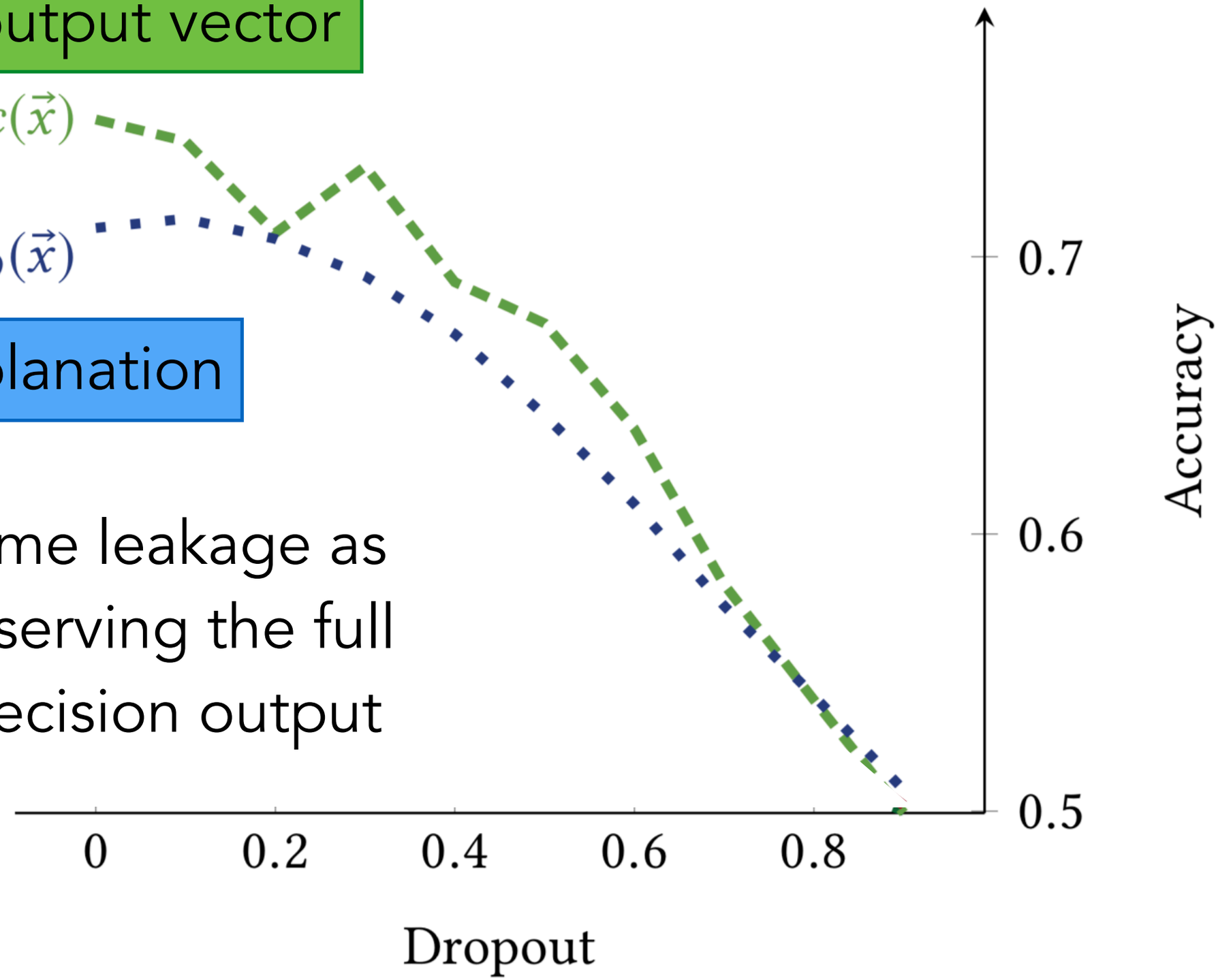
Complete output vector

$c(\vec{x})$

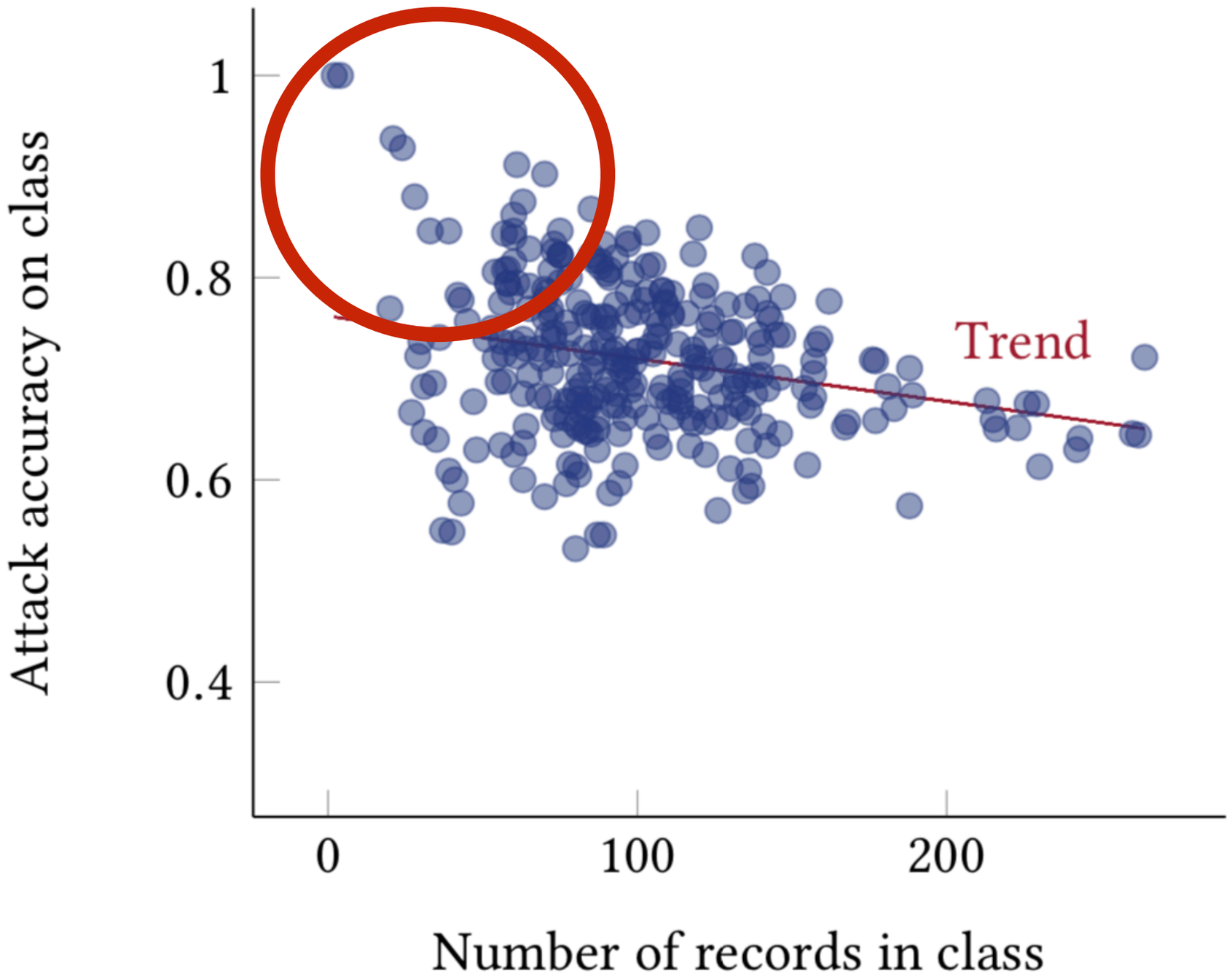
$\phi_{GRAD}(\vec{x})$

Label + Explanation

Same leakage as observing the full precision output

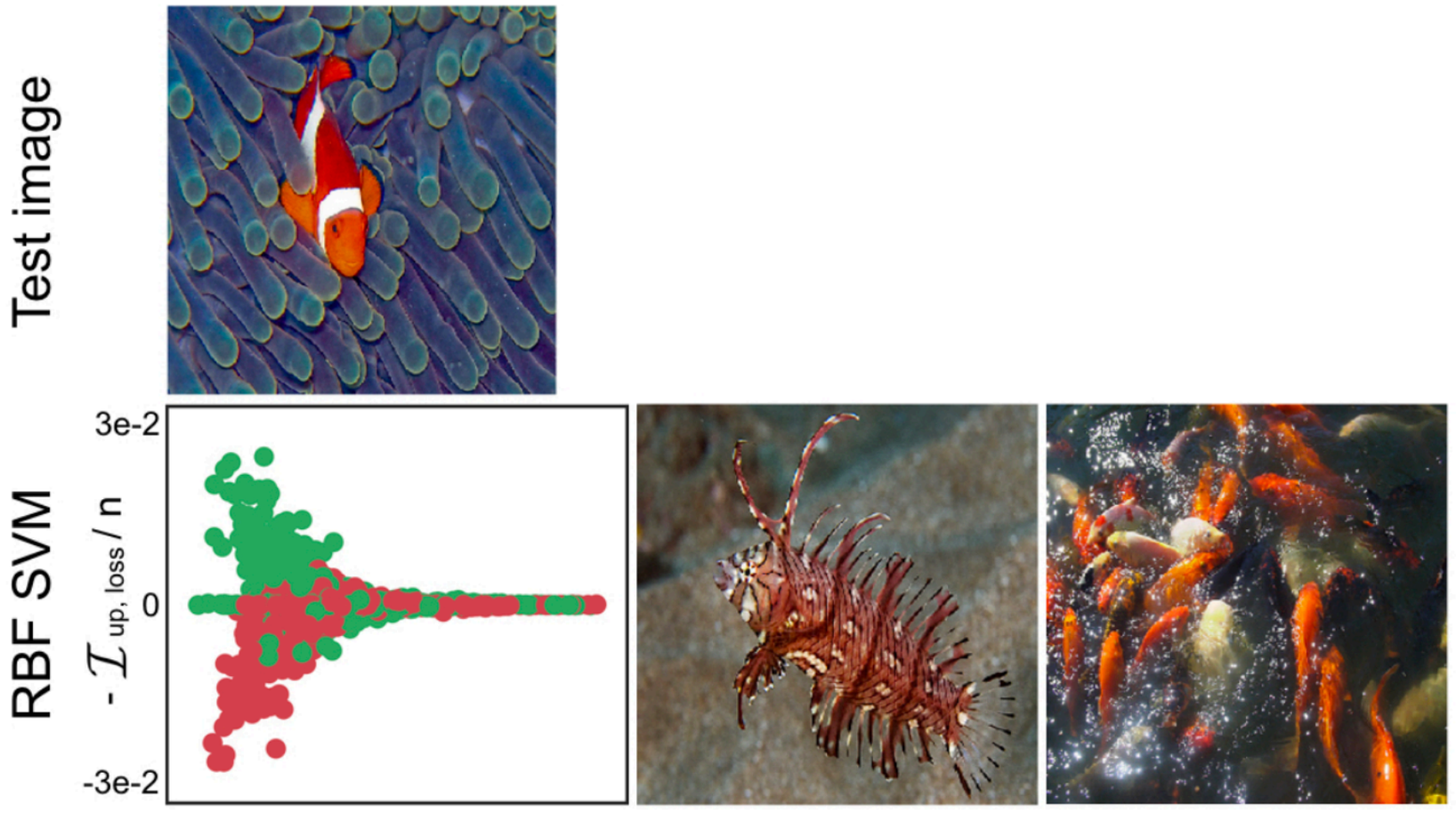


Impact on Minorities



Record-based Explanation

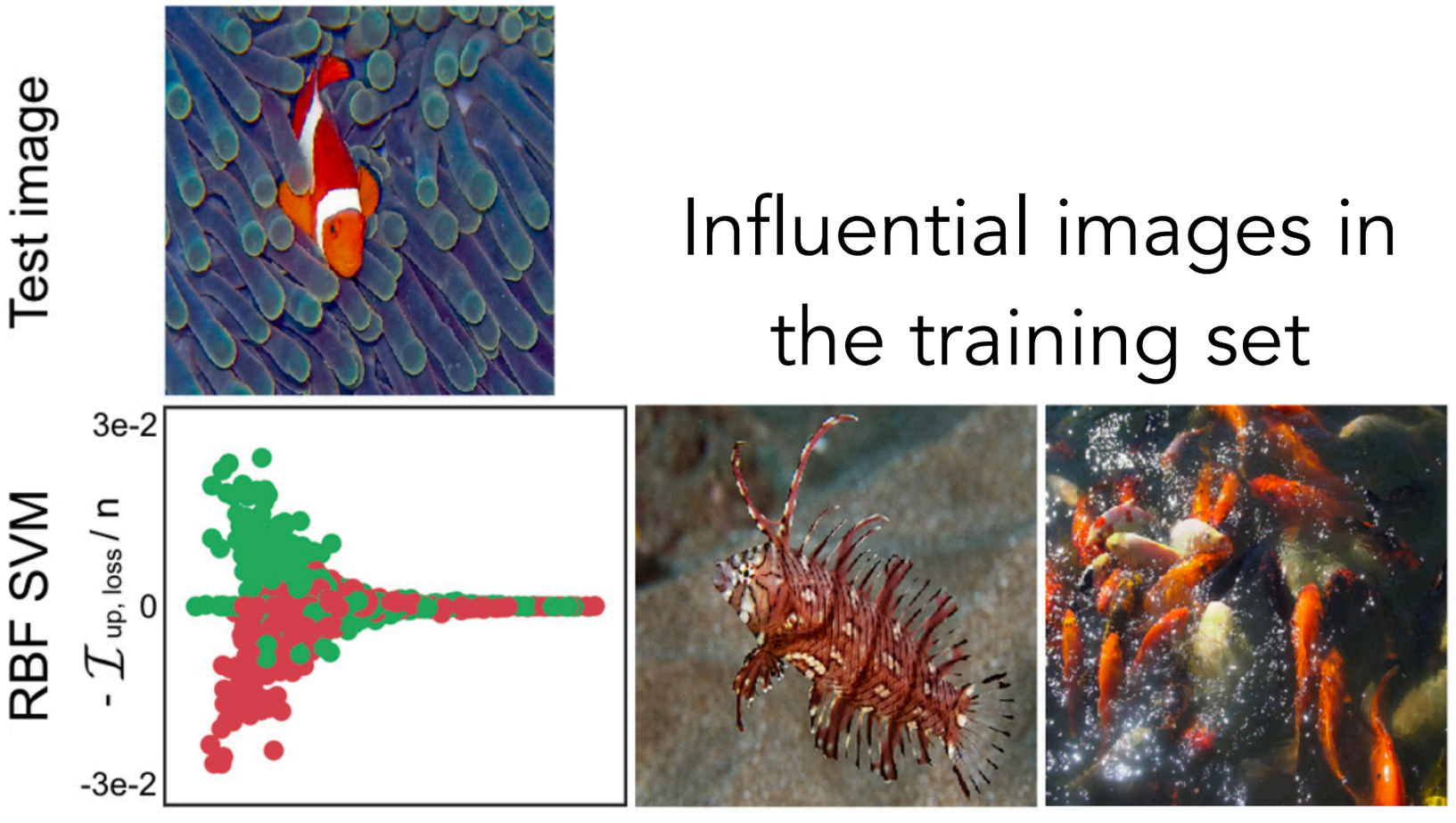
- Release the most influential data points in the training set on the classification of a target data point (by approximating the change in loss on the target point if a training point is removed)



Pang Wei Koh and Percy Liang

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)



Pang Wei Koh and Percy Liang

It Discloses Training Data !

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

% of data	$k = 1$	$k = 5$	$k = 10$

Attack accuracy

Diabetics Hospital Dataset (medical test)

It Discloses Training Data !

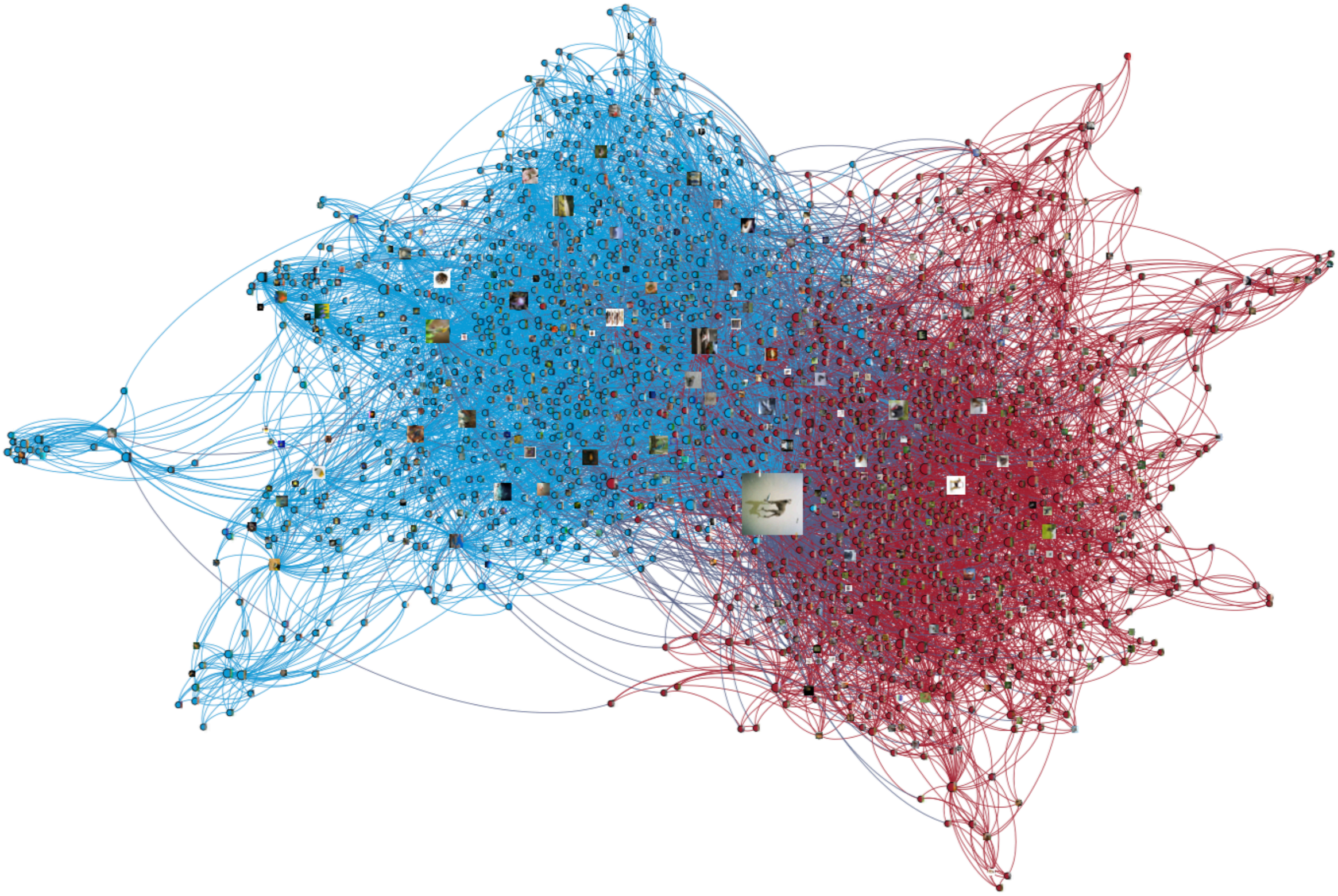
Membership Inference: Query with a record and see if it appears as influential for itself

	% of data	$k = 1$	$k = 5$	$k = 10$
Whole data set	100%	34%	64%	77%
Age 0 -10	<0.1%	67%	100%	100%
Age 0 -20	<1%	20%	58%	92%
Caucasian	74%	34%	64%	77%
African American	19%	38%	68%	81%
Hispanics	2%	39%	64%	76%
Unknown race	1%	35%	60%	77%
Asian American	<1%	25%	64%	89%

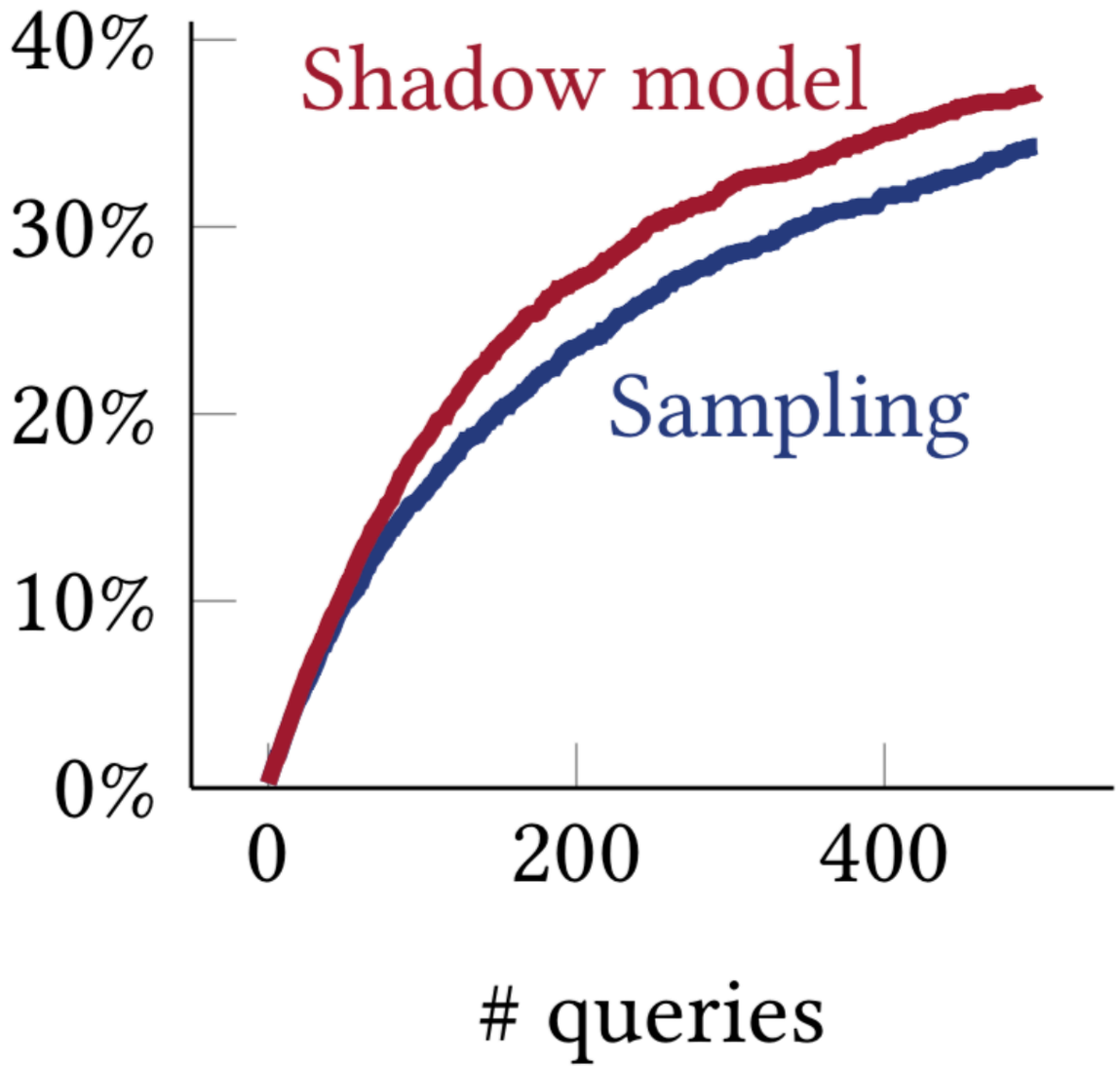
Minorities

Diabetics Hospital Dataset (medical test)

Reconstruction Attack



Reconstruction Attack



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)

NUS | Computing

National University of Singapore

We are hiring (PhD, Postdoc)!