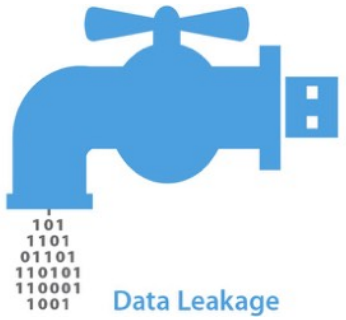
LOGAN: Membership Inference Attacks Against Generative Models

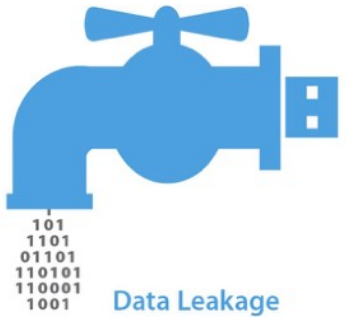
Jamie Hayes*, Luca Melis*, George Danezis,
and Emiliano De Cristofaro

Privacy in ML is 🔥🔥🔥



Privacy in ML is 🔥🔥🔥

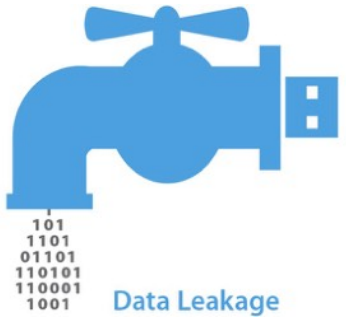
Most papers on privacy in ML focus on inferring:



Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

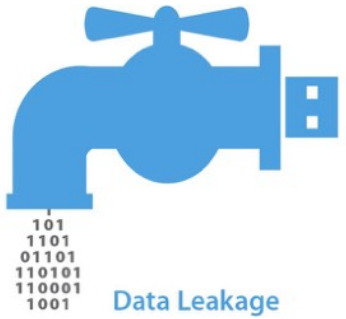


Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

2. What class representatives look like



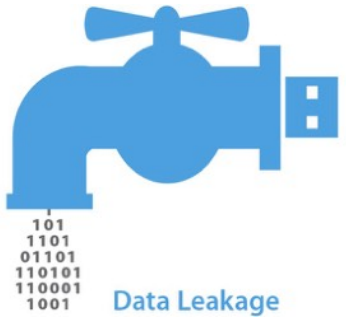
Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

2. What class representatives look like

3. Properties of training data



Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

Membership Inference

2. What class representatives look like

3. Properties of training data

Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

2. What class representatives look like

3. Properties of training data

Membership Inference

Model Inversion

Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set
2. What class representatives look like
3. Properties of training data

Membership Inference

Model Inversion

Property Inference

Privacy in ML is 🔥🔥🔥

Most papers on privacy in ML focus on inferring:

1. Inclusion of a data point in the training set

Membership Inference

This talk!

Model Inversion —> Fredrikson et al., Model inversion attacks that exploit confidence information and basic countermeasures. ACM CCS'15.

Property Inference —> Melis et al., Exploiting Unintended Feature Leakage in Collaborative Learning. IEEE S&P'19

Membership Inference

Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model

Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model

- Serious problem if inclusion in training set is privacy-sensitive

Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model

- Serious problem if inclusion in training set is privacy-sensitive
- E.g., main task is: predict whether a smoker gets cancer

Membership Inference

Adversary wants to test whether data of a target victim has been used to train a model

- Serious problem if inclusion in training set is privacy-sensitive
- E.g., main task is: predict whether a smoker gets cancer

Membership inference is a very active research area, not only in machine learning…

Membership Inference (cnt'd)

Membership inference is a very active research area, not only in machine learning…

Membership Inference (cnt'd)

Membership inference is a very active research area, not only in machine learning…

Given f(data), infer if x ∈ data (e.g., f is aggregation)

Membership Inference (cnt'd)

Membership inference is a very active research area, not only in machine learning…

Given f(data), infer if x ∈ data (e.g., f is aggregation)

[HSR+08, WLW+09] for genomic data

[Pyrgelis et al., NDSS'18] for mobility data

Membership Inference (cnt'd)

Membership inference is a very active research area, not only in machine learning…

Given f(data), infer if x ∈ data (e.g., f is aggregation)

[HSR+08, WLW+09] for genomic data

[Pyrgelis et al., NDSS'18] for mobility data

Well-understood problem, besides the more obvious leakage

Membership Inference (cnt'd)

Membership inference is a very active research area, not only in machine learning…

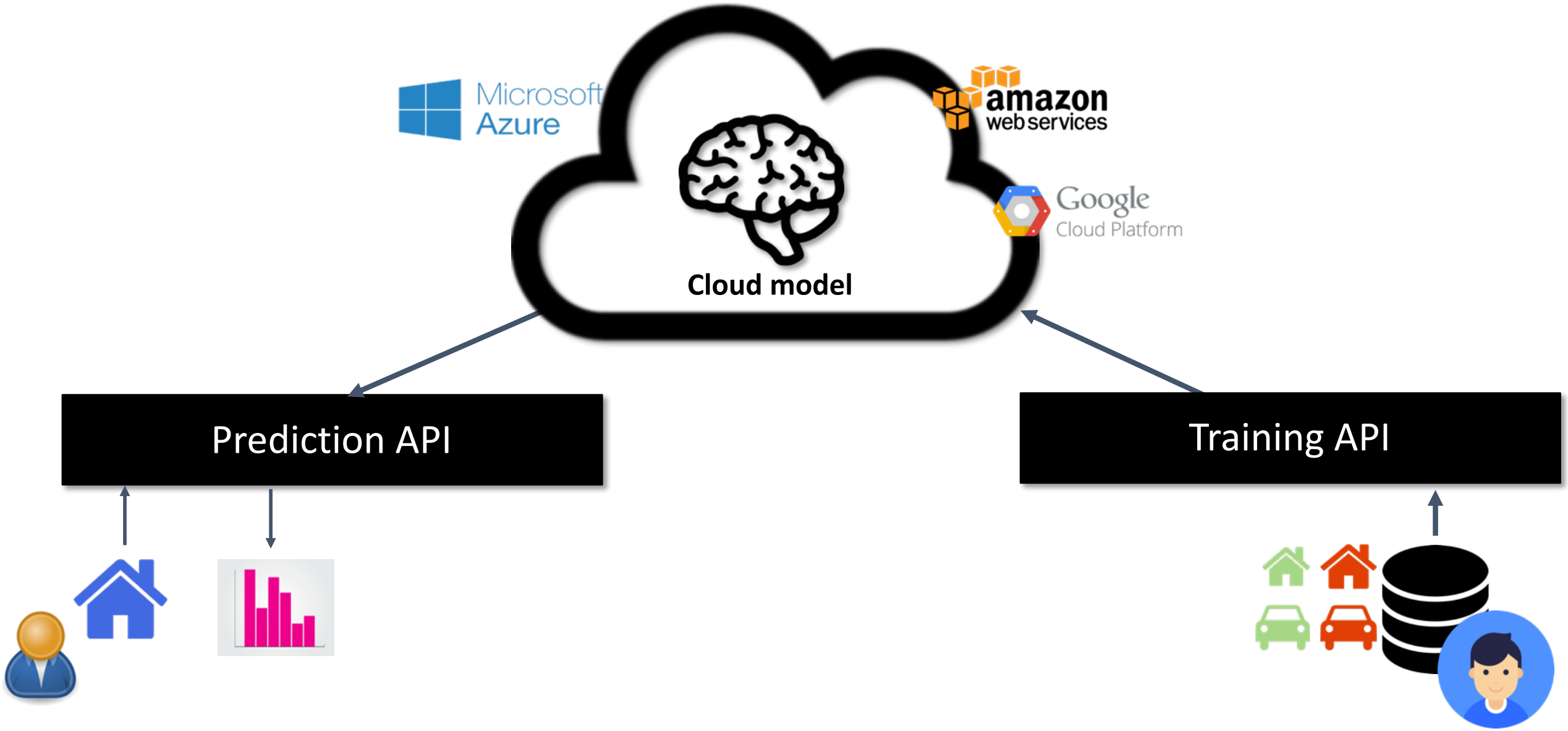
- Given f(data), infer if x ∈ data (e.g., f is aggregation)
- [HSR+08, WLW+09] for genomic data
- [Pyrgelis et al., NDSS'18] for mobility data

Well-understood problem, besides the more obvious leakage

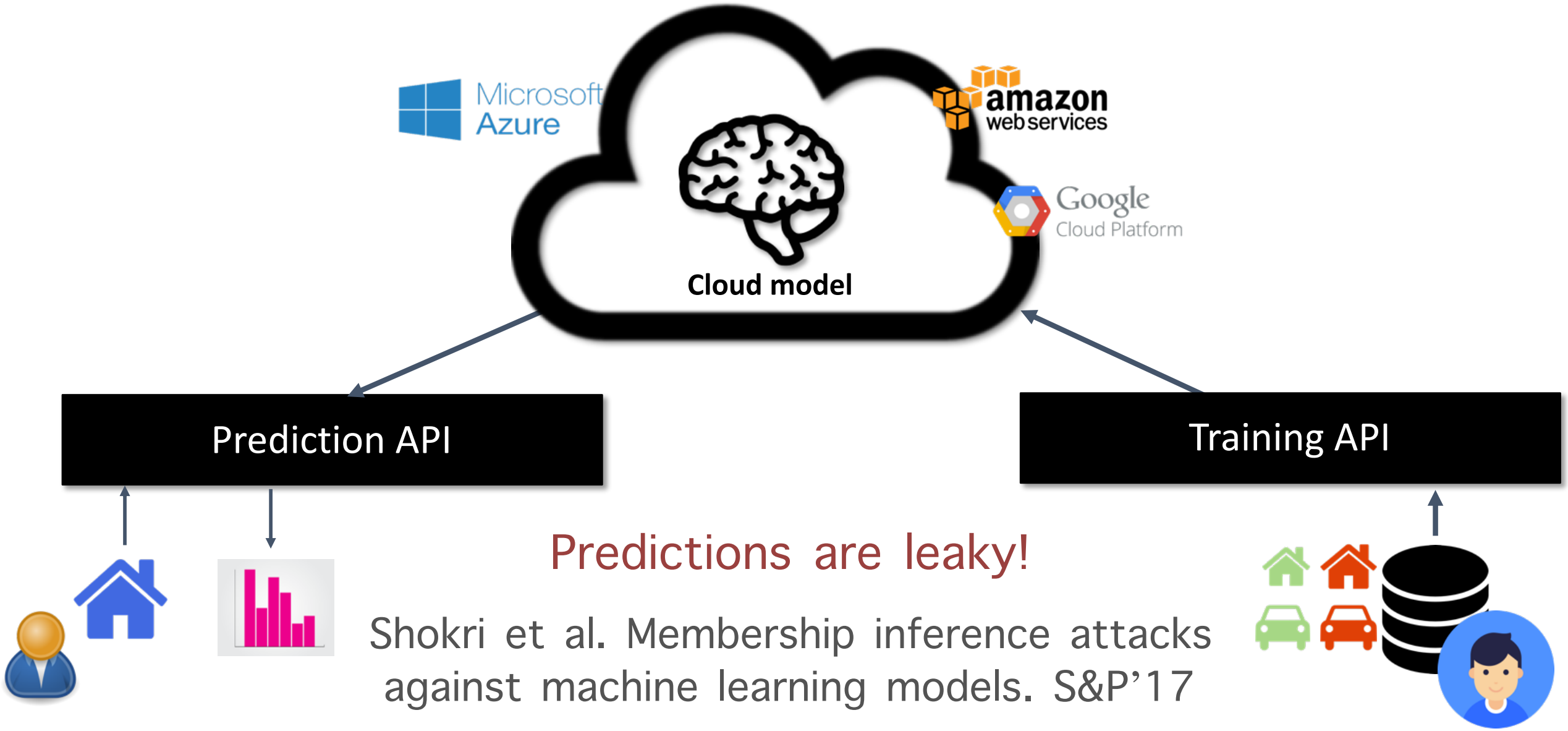
- Establish wrongdoing
- Assess protection, e.g., from differentially private defenses

Machine Learning as a Service

Machine Learning as a Service



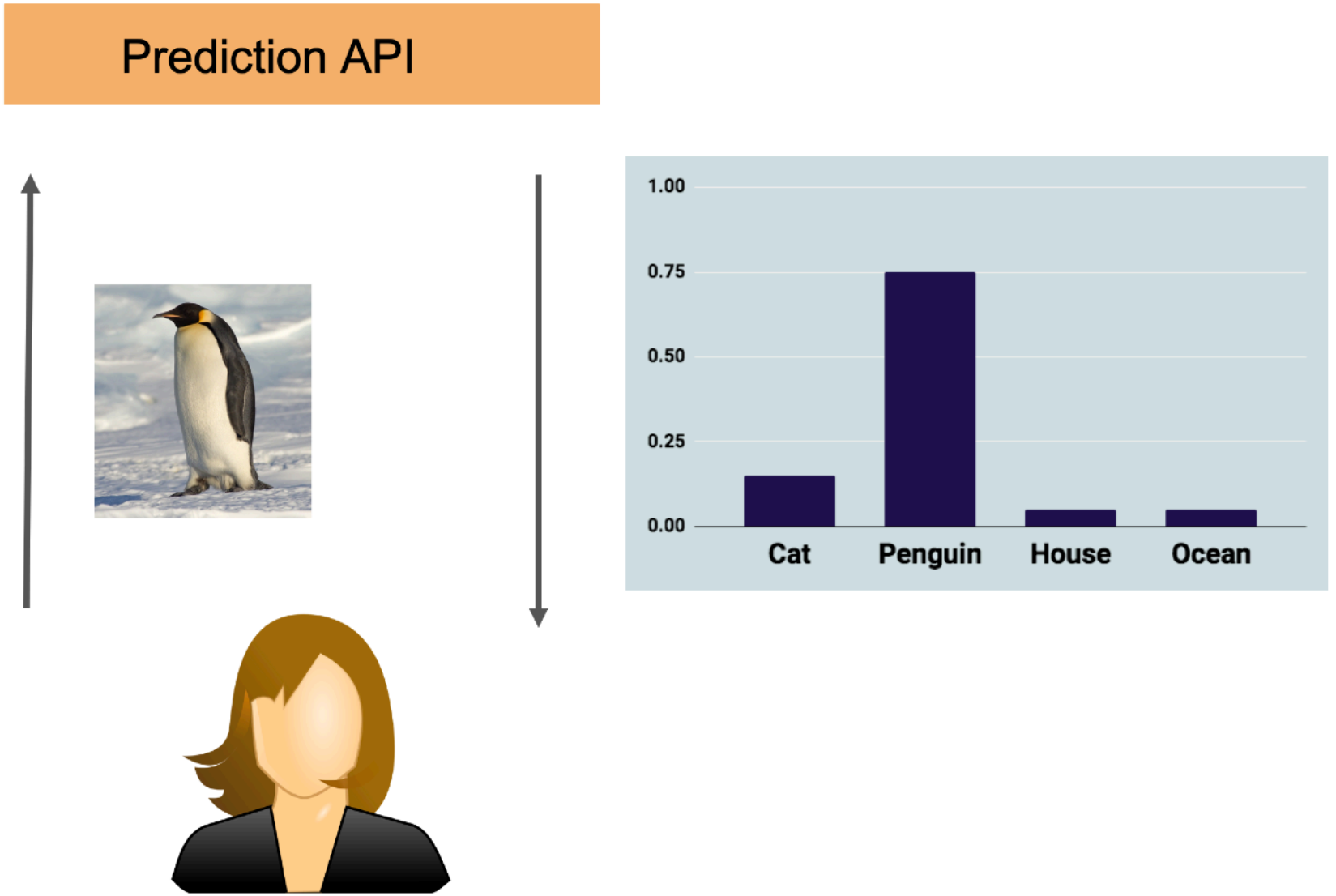
Machine Learning as a Service

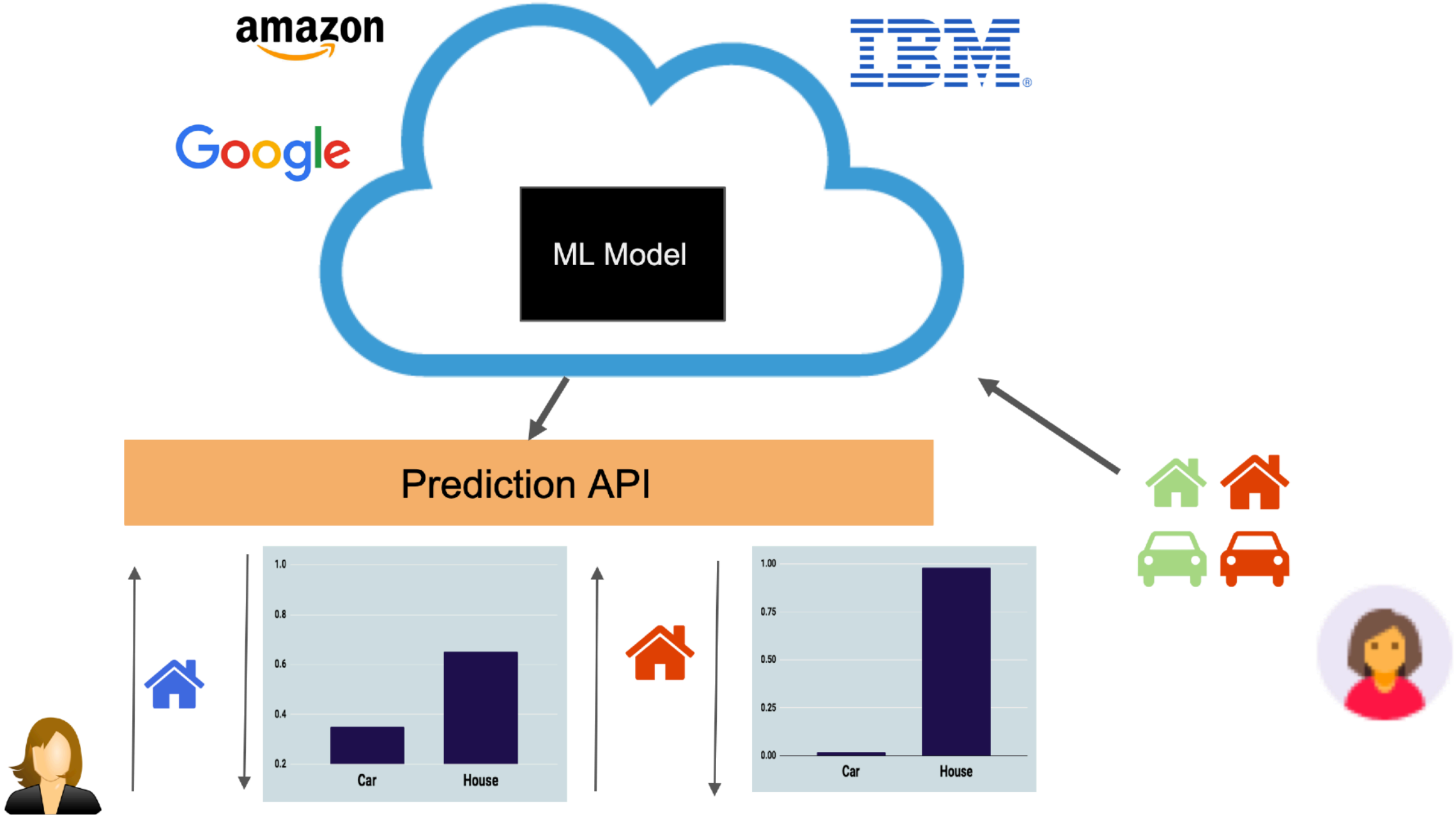


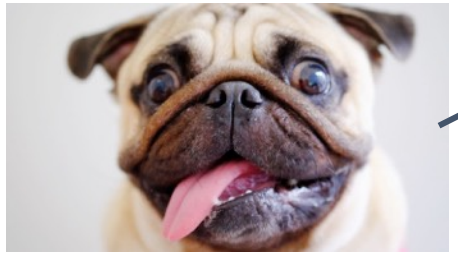
Predictions are leaky!

Shokri et al. Membership inference attacks against machine learning models. S&P'17

Membership Inference/Discriminative

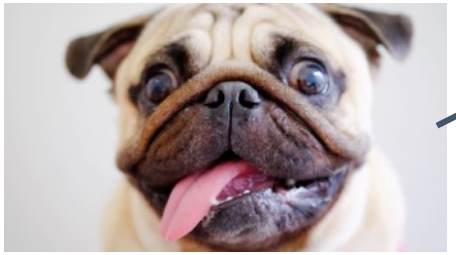



amazon
IBM
Google
ML Model
Prediction API
1.0
0.8
0.6
0.4
0.2
Car
House
1.00
0.75
0.50
0.25
0.00
Car
House



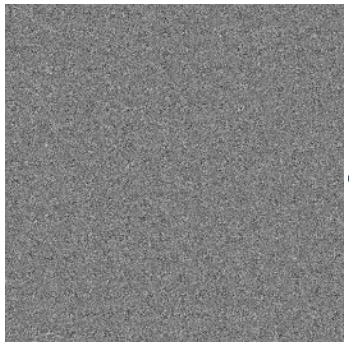
Discriminative Model

cat | dog



Discriminative Model

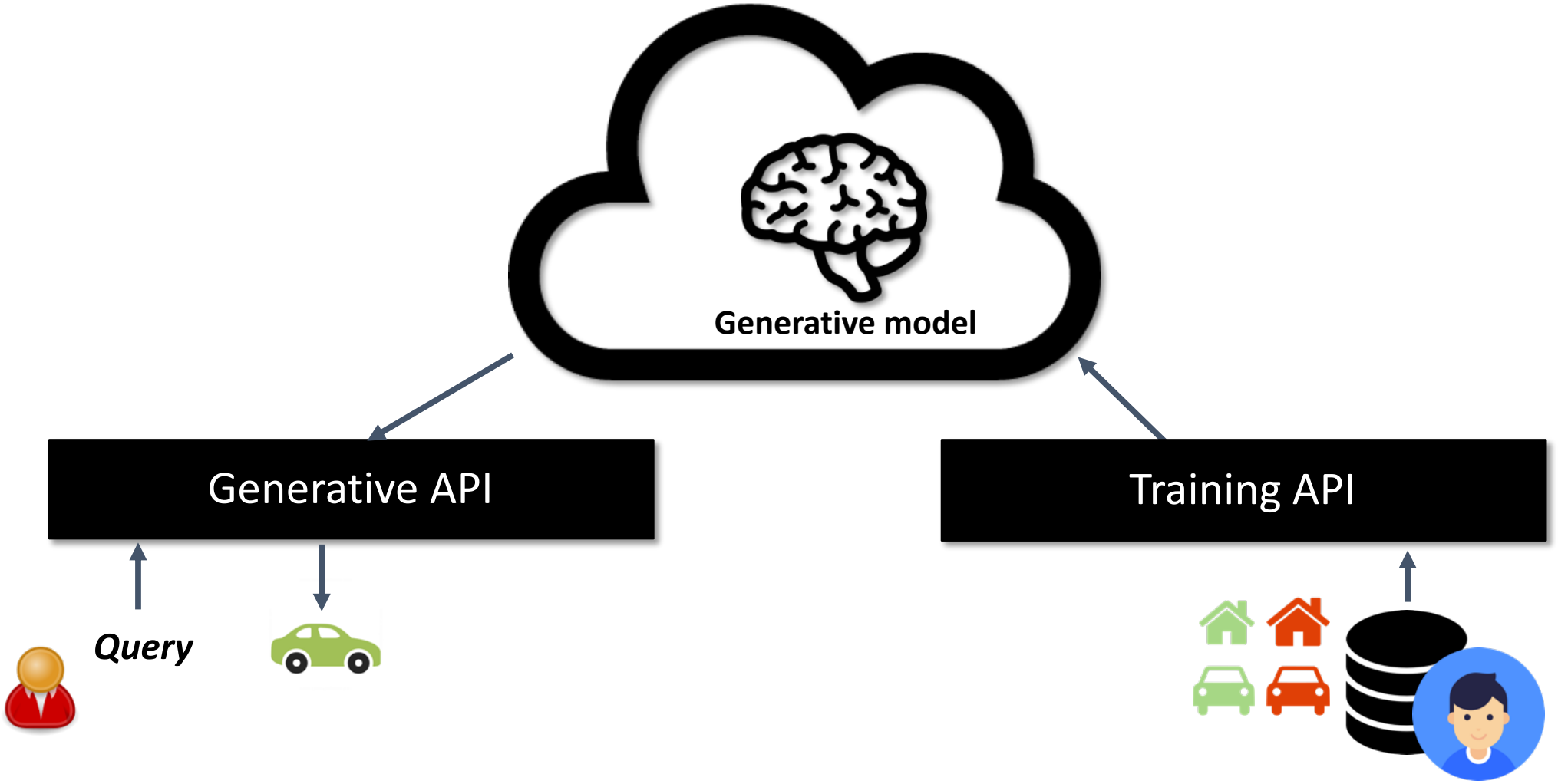
cat | dog



Generative Model

Membership Inference in Generative Models?

Membership Inference in Generative Models?



Inference without predictions?

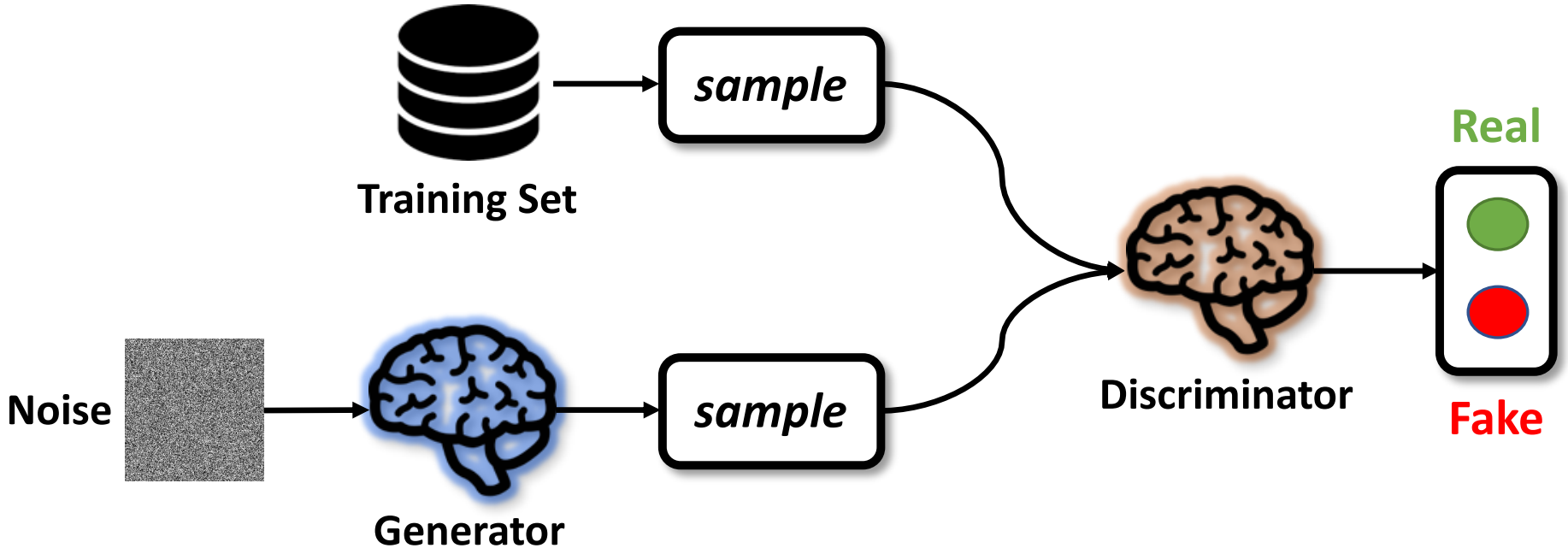
Use generative models!

Train GANs to learn the distribution and a prediction model at the same time

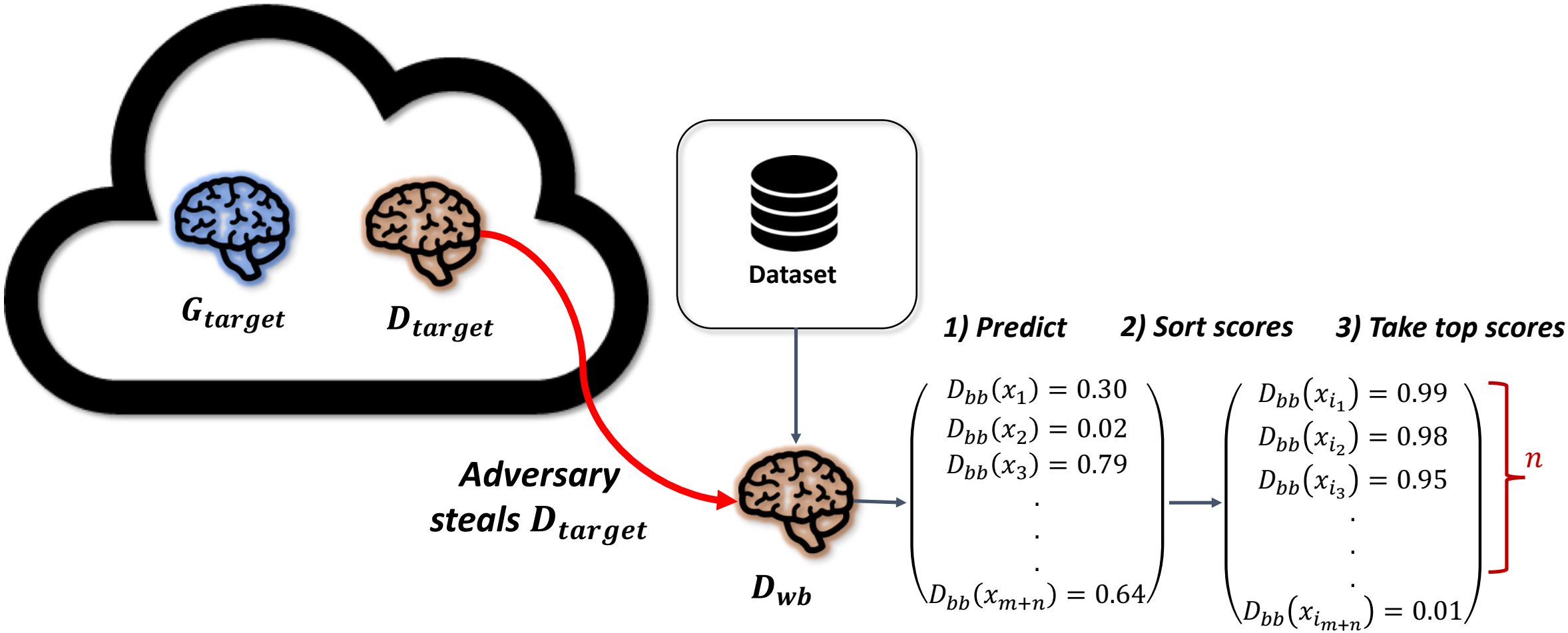
Inference without predictions?

Use generative models!

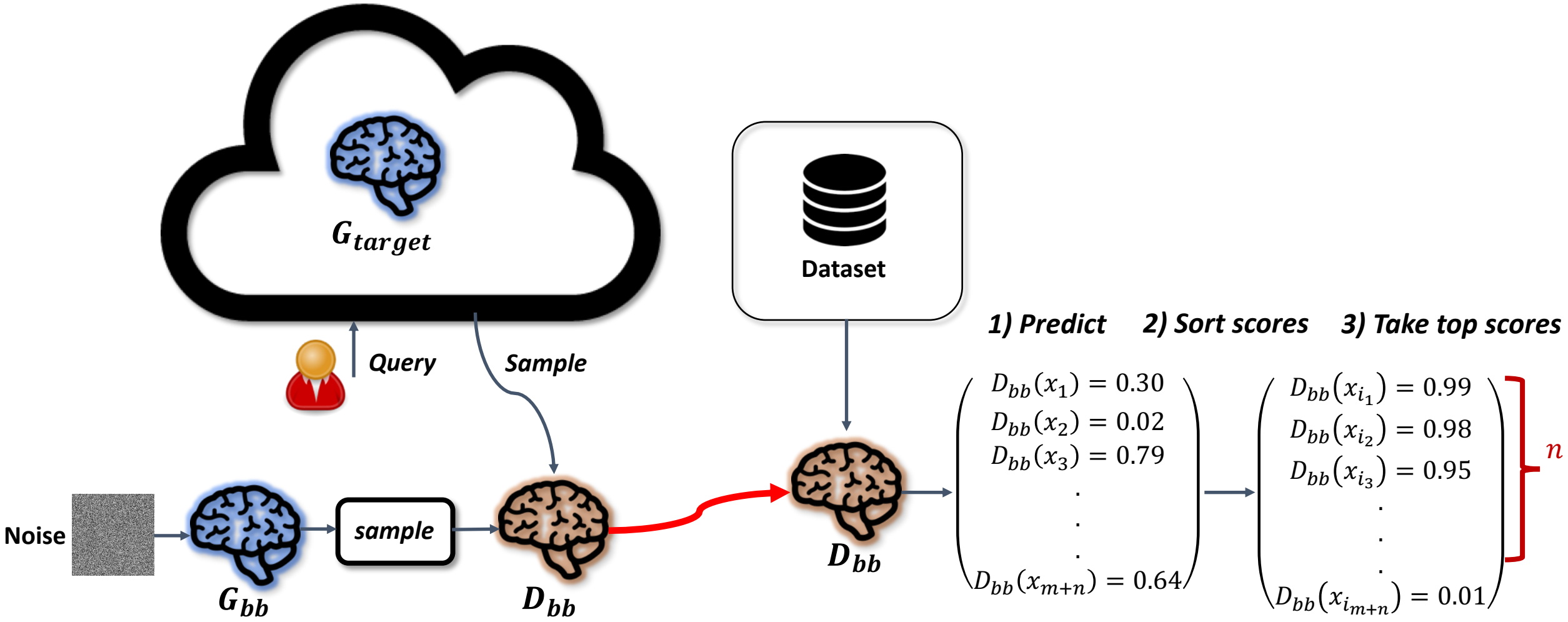
Train GANs to learn the distribution and a prediction model at the same time



White-Box Attack



Black-Box Attack

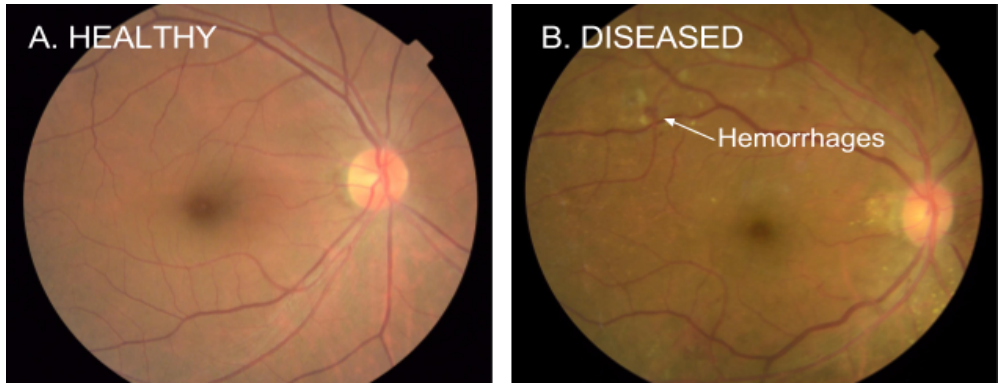


Datasets

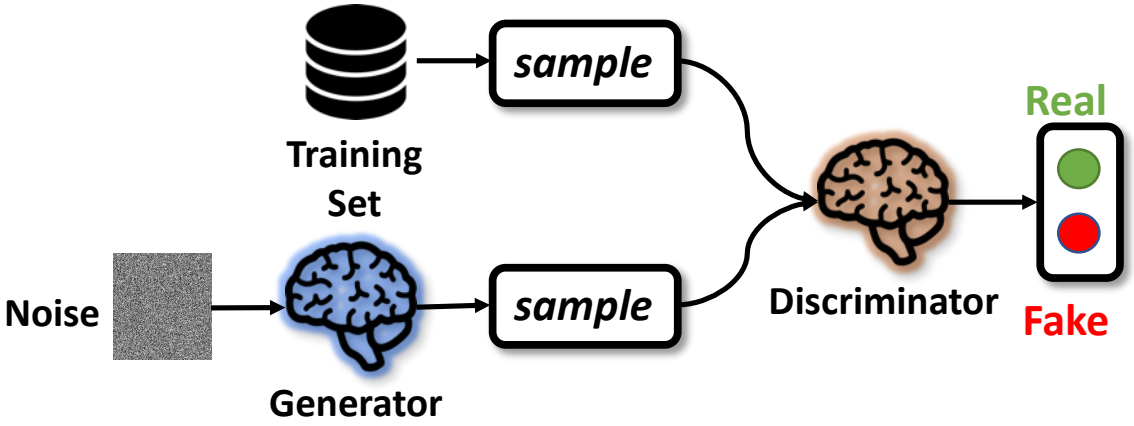
LFW

CIFAR-10

DR



Models



Attacker Model:

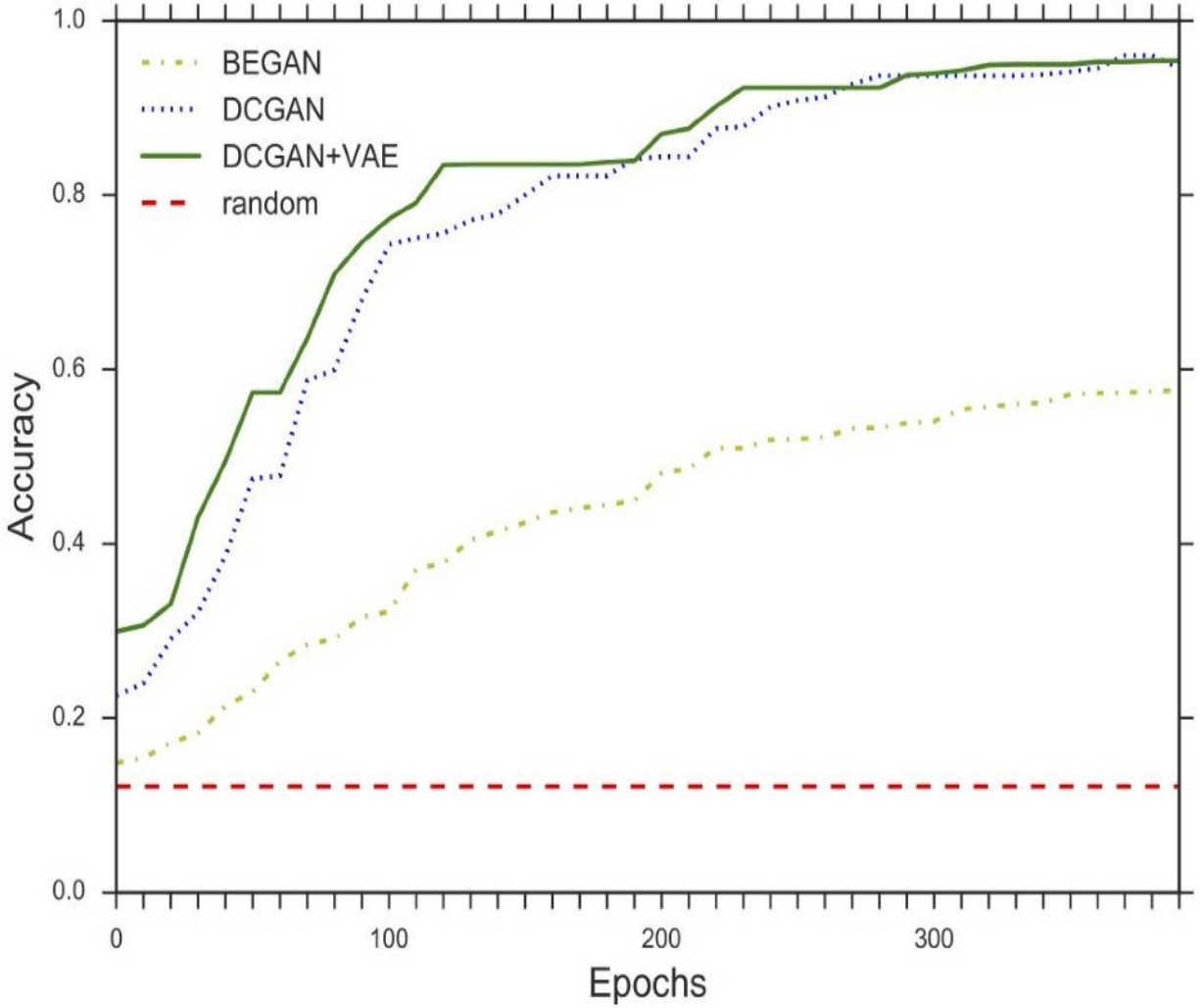
DCGAN

Target Model:

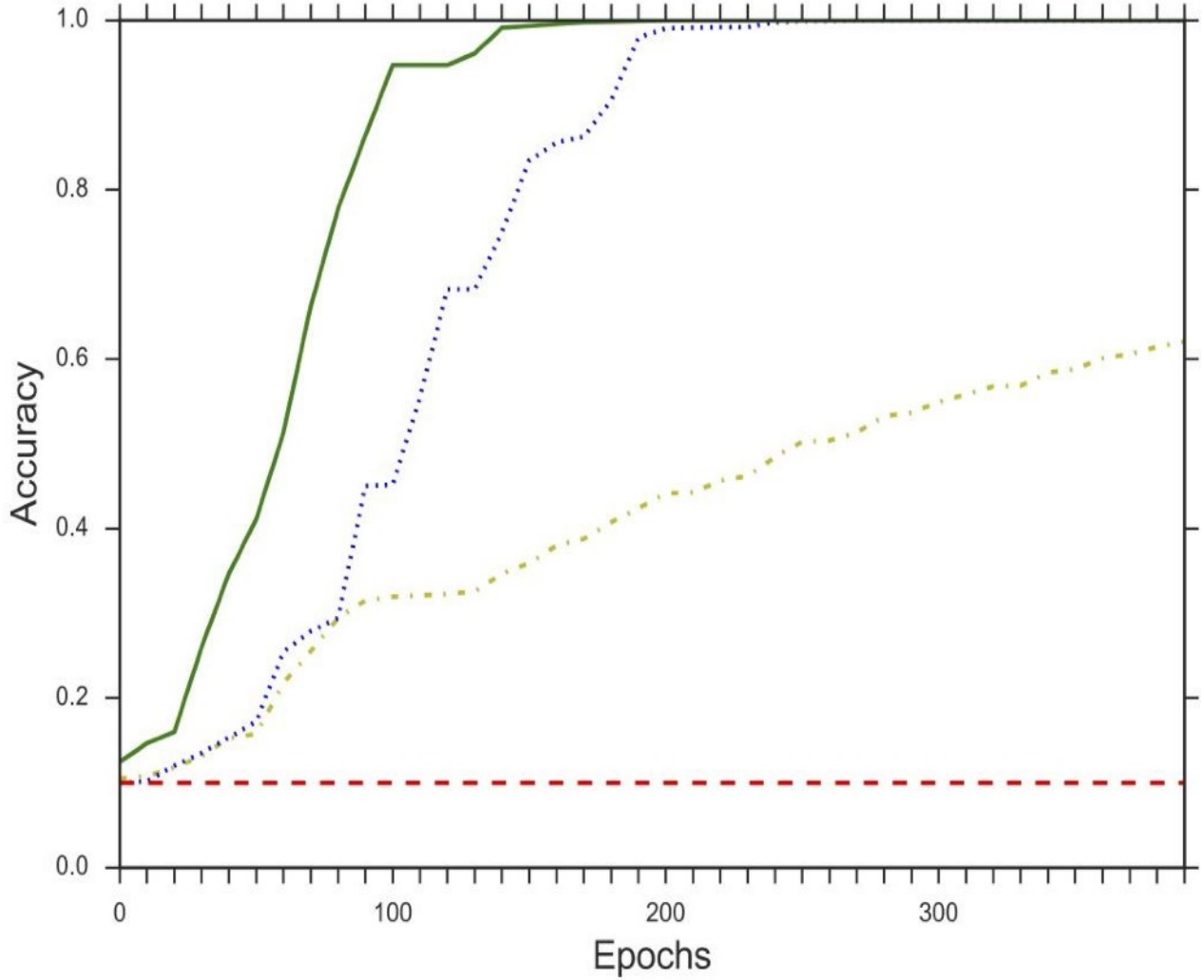
DCGAN, DCGAN+VAE, BEGAN

White-Box Results

LFW, top ten classes

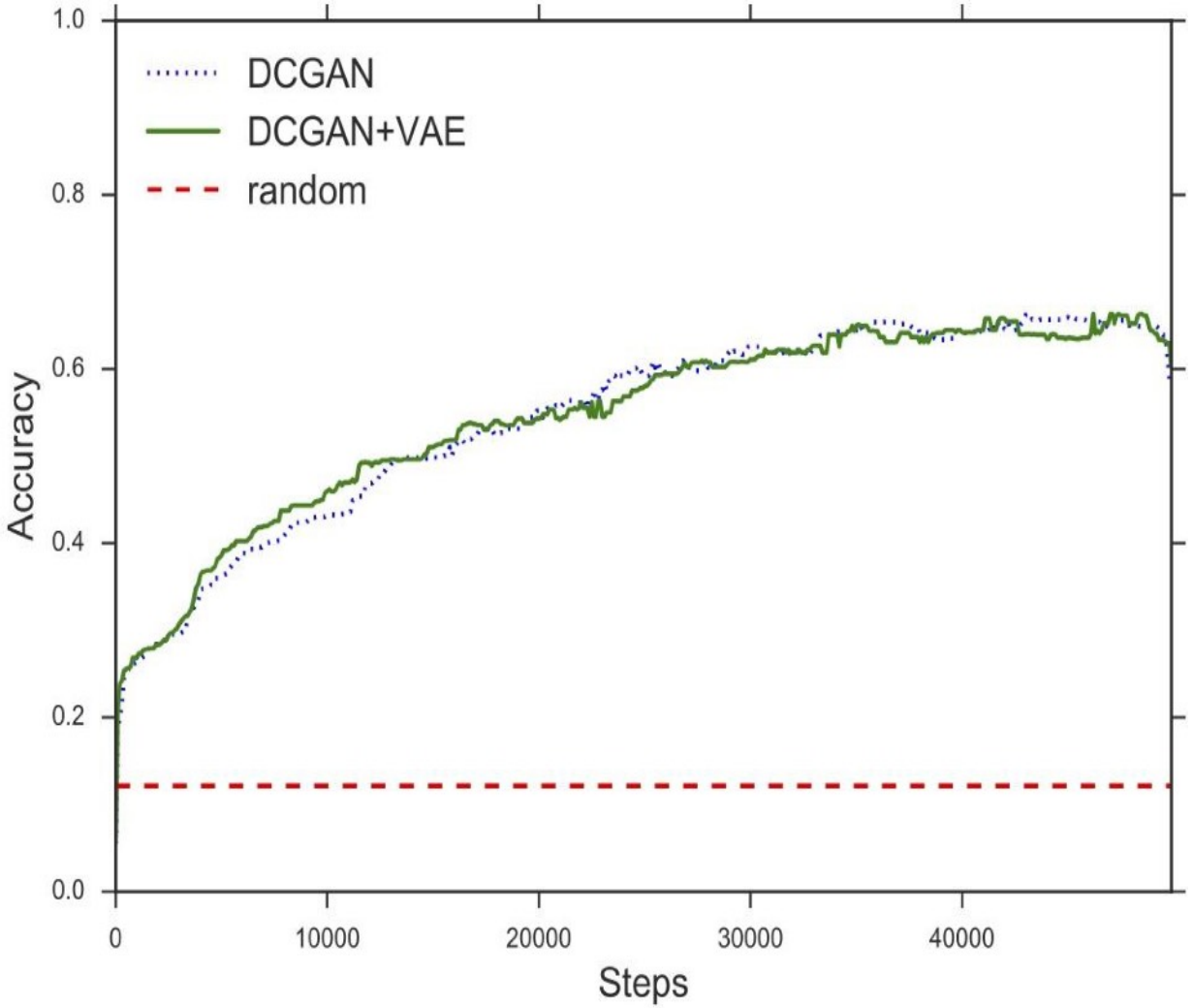


CIFAR-10, random 10% subset

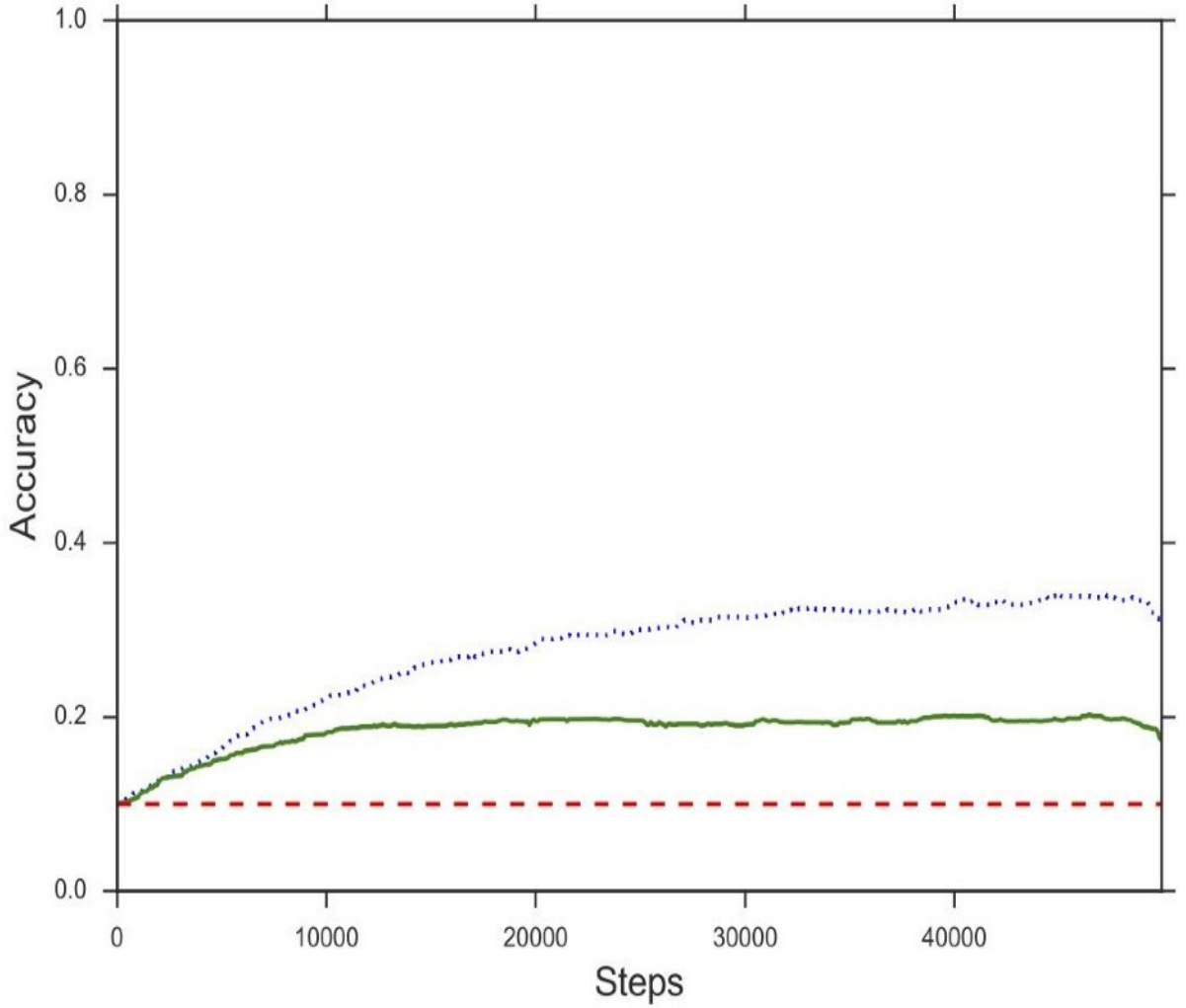


Black-Box Results

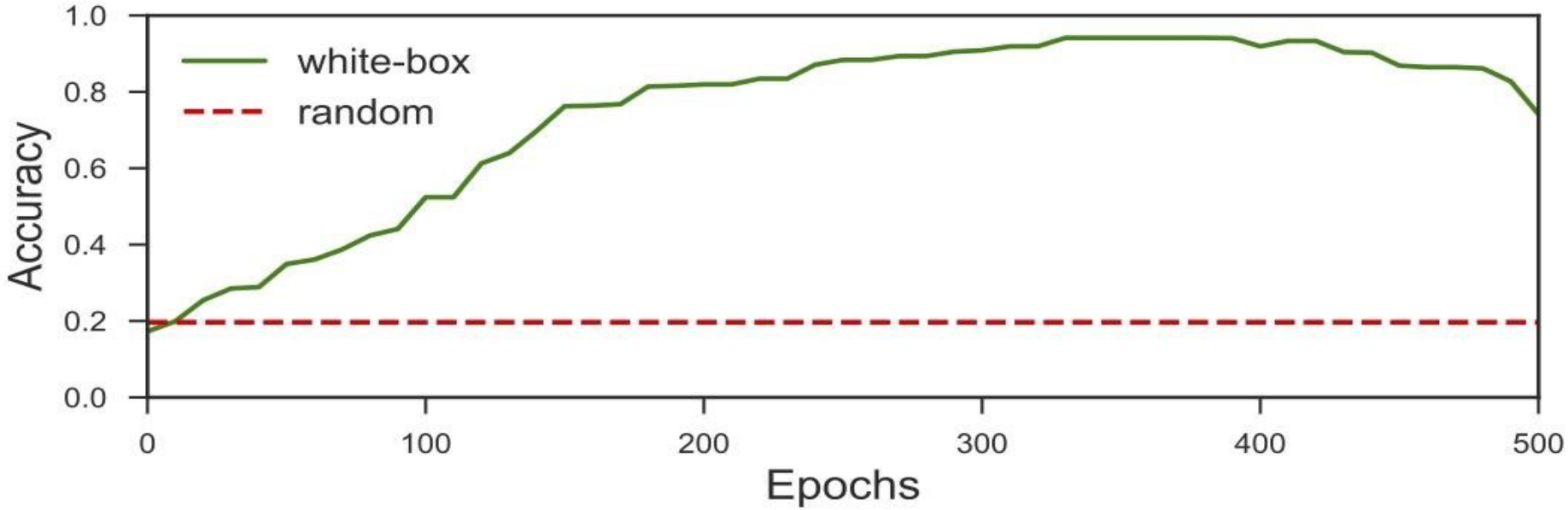
LFW, top ten classes



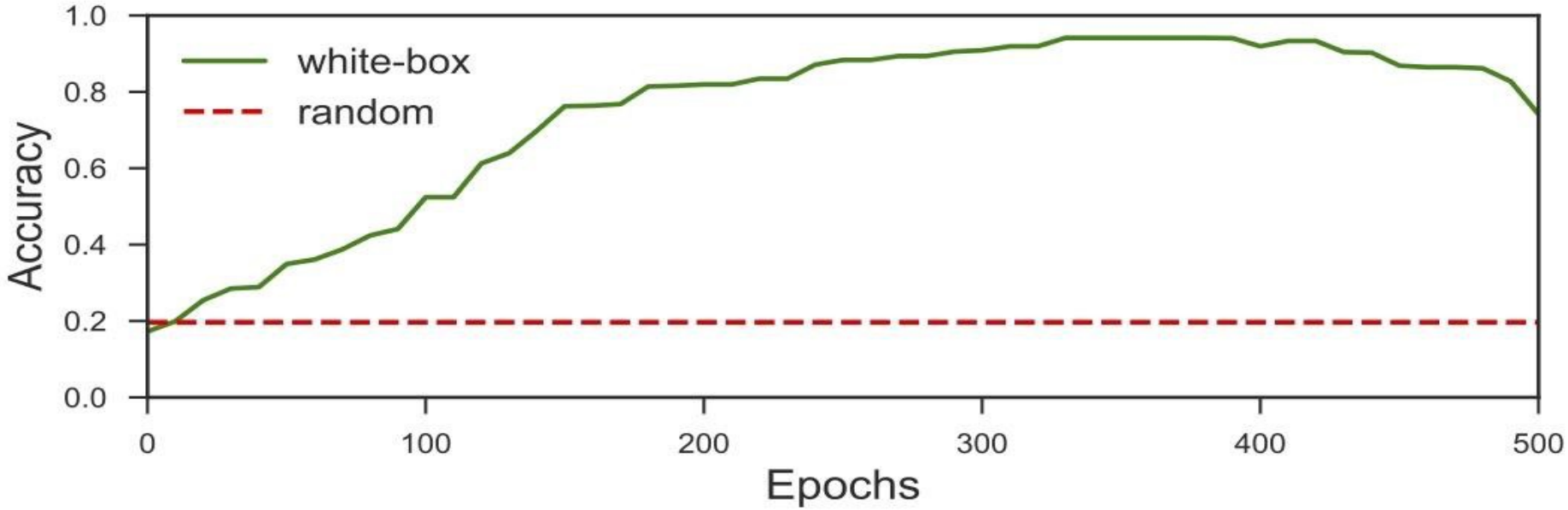
CIFAR-10, random 10% subset

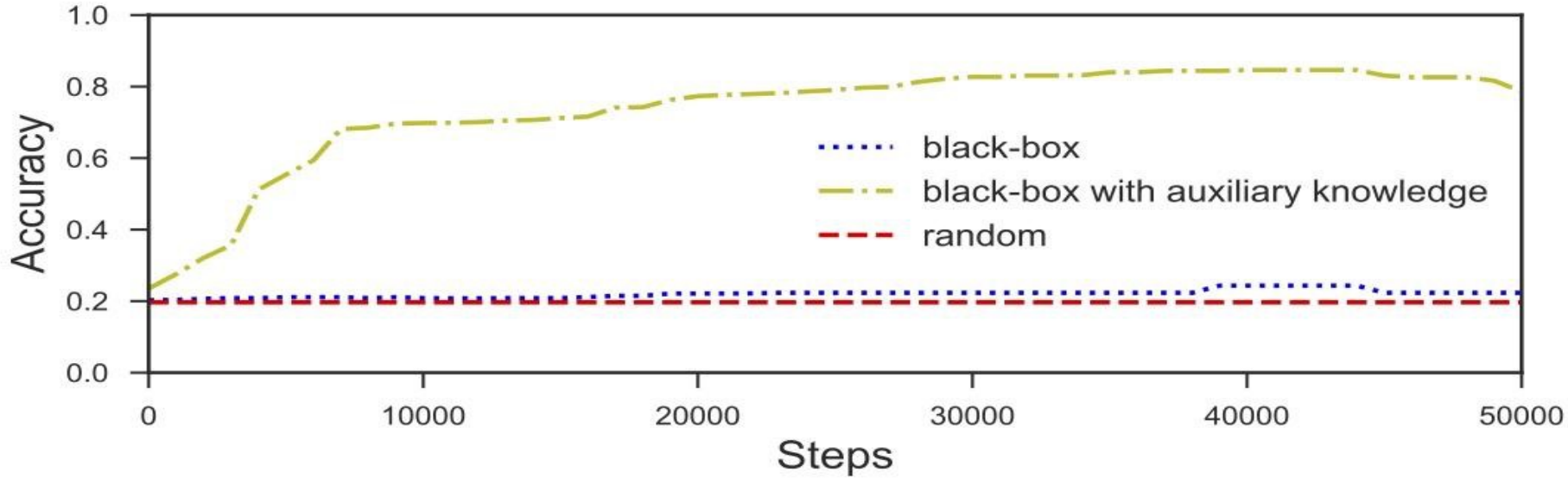


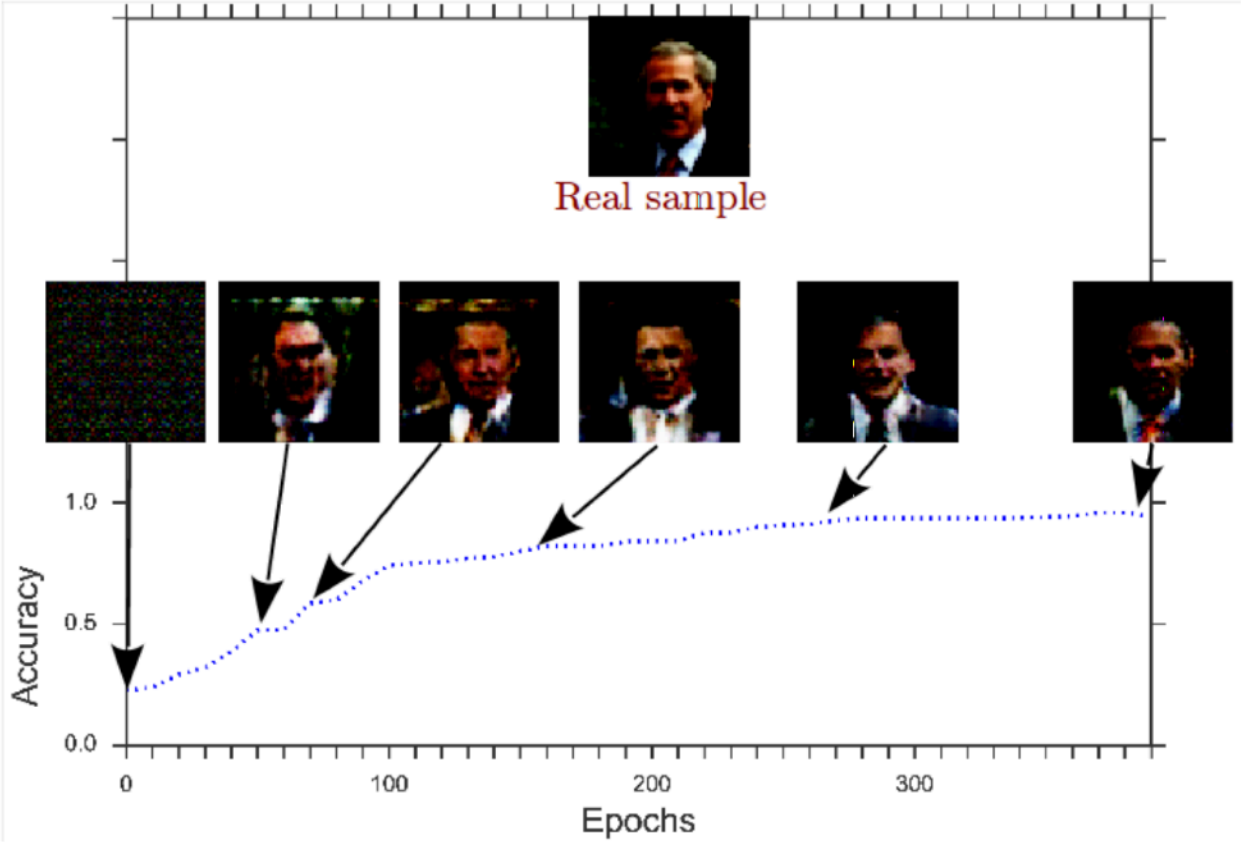
DR Dataset


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
white-box
random
0
100
200
300
400
500
Epochs

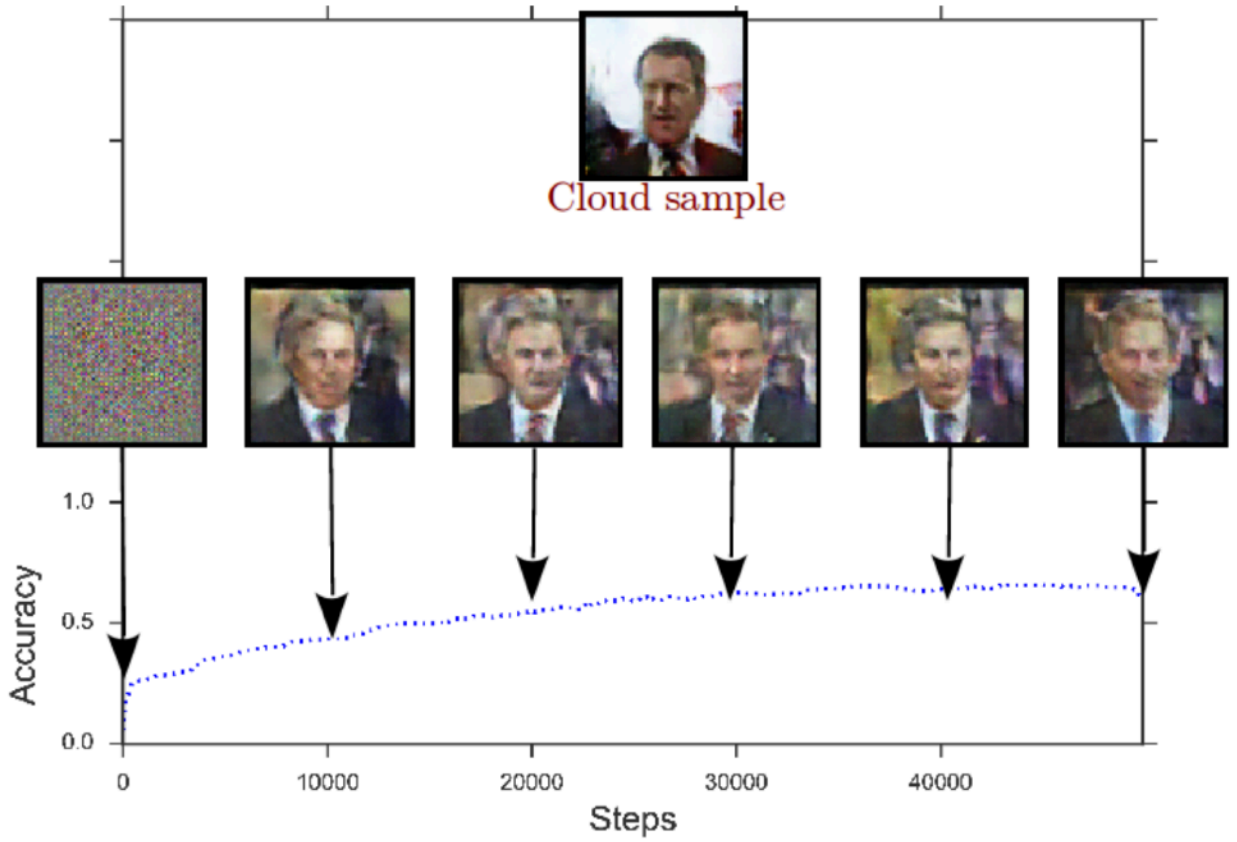
DR Dataset


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
white-box
random
0
100
200
300
400
500
Epochs


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
black-box
black-box with auxiliary knowledge
random
0
10000
20000
30000
40000
50000
Steps



(a) White-box attack

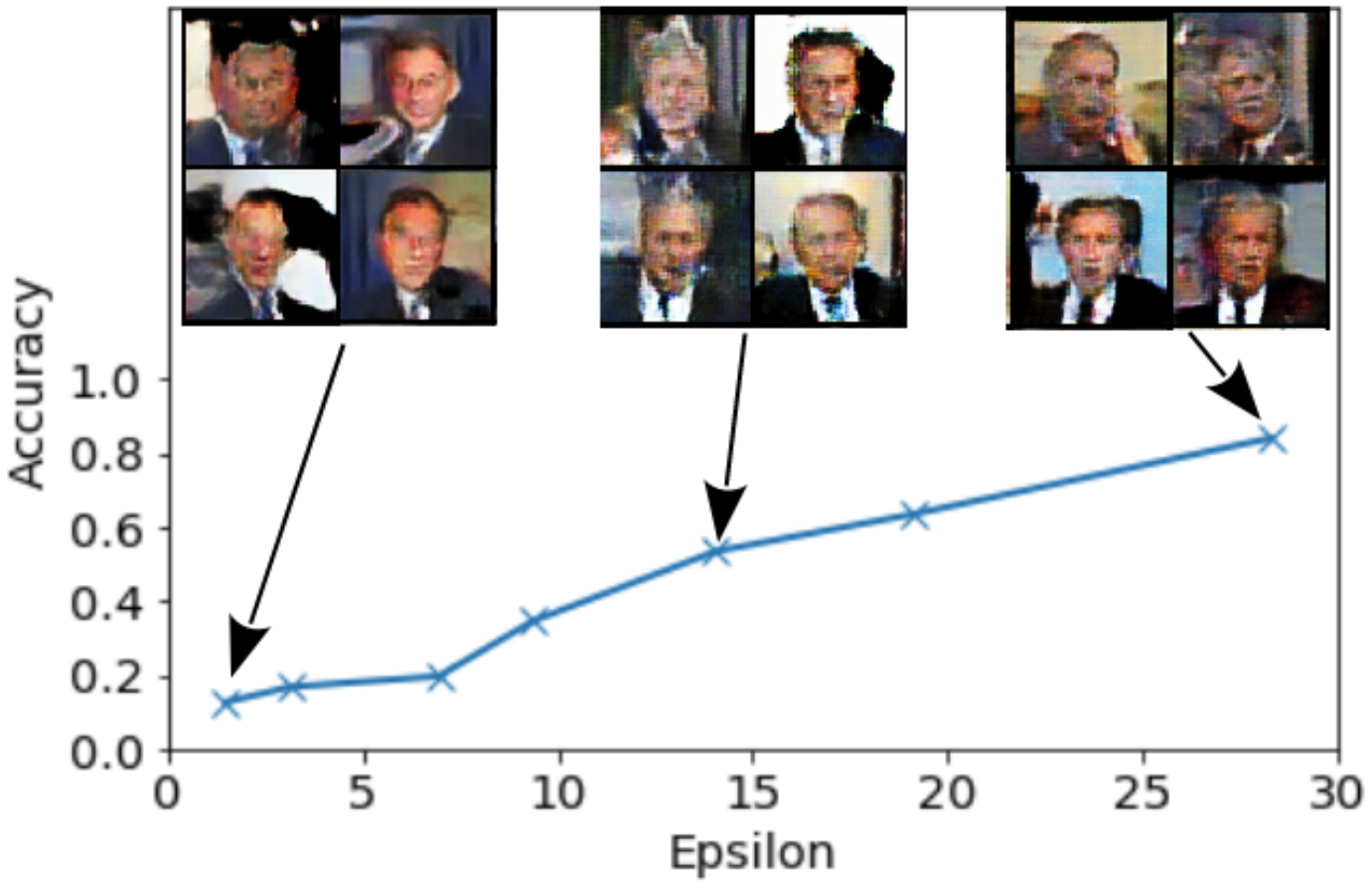


(b) Black-box attack

In a nutshell…

Attack	LFW	CIFAR-10	DR
White-box	100%	100%	95%
Black-box	40%	37%	22%
Black-box with aux knowledge	60%	58%	81%
Random guess	10%	10%	20%

Defense? Differentially Private GAN?



White-box, LFW, top ten classes

*Triastcyn et al. “Generating differentially private datasets using GANs.” arXiv 1803.03148

Thank you!