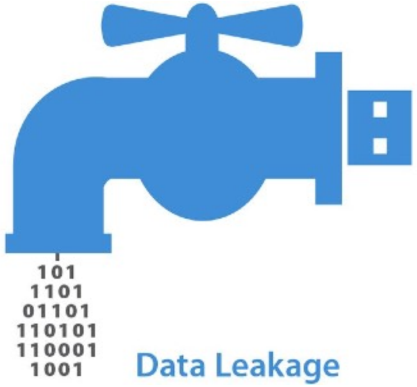
Membership and Property Inference Attacks Against Machine Learning

Emiliano De Cristofaro
https://emilianodc.com

Reasoning about "privacy" in ML



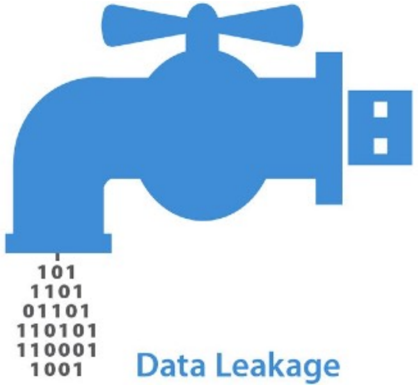
Reasoning about “privacy” in ML

Most papers on privacy attacks in ML focus on inferring:

Reasoning about "privacy" in ML

Most papers on privacy attacks in ML focus on inferring:

1. Inclusion of a data point in the training set
   (aka "membership inference")



Reasoning about "privacy" in ML

Most papers on privacy attacks in ML focus on inferring:

1. Inclusion of a data point in the training set (aka "membership inference")
2. What class representatives look like

1. Membership Inference

1. Membership Inference

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

1. 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

1. 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

1. 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
- [Shokri et al., S&P'17] show it for discriminative models

1. 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
- [Shokri et al., S&P'17] show it for discriminative models
- [Hayes et al. PETS'19] for generative models (later in the talk)

1. 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
- [Shokri et al., S&P'17] show it for discriminative models
- [Hayes et al. PETS'19] for generative models (later in the talk)

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 $x \in$ 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 $x \in$ data (e.g., f is aggregation)
- [Homer et al., Science'13] 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 $x \in$ data (e.g., f is aggregation)

[Homer et al., Science'13] 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 $x \in$ data (e.g., f is aggregation)
- [Homer et al., Science'13] 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

2. Inferring Class Representatives

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

Model Inversion [Fredrikson et al. CCS'15]

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

- Model Inversion [Fredrikson et al. CCS'15]
- GAN attacks [Hitaji et al. CCS'17]

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

- Model Inversion [Fredrikson et al. CCS'15]
- GAN attacks [Hitaji et al. CCS'17]

E.g.: given a gender classifier, infer what a male looks like

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

Model Inversion [Fredrikson et al. CCS'15]

GAN attacks [Hitaji et al. CCS'17]

E.g.: given a gender classifier, infer what a male looks like

But…any useful machine learning model does reveal something about the population from which the training data was sampled

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

- Model Inversion [Fredrikson et al. CCS'15]
- GAN attacks [Hitaji et al. CCS'17]

E.g.: given a gender classifier, infer what a male looks like

But…any useful machine learning model does reveal something about the population from which the training data was sampled

Privacy leakage != Adv learns something about training data

2. Inferring Class Representatives

Prior work shows how infer properties of an entire class, e.g.:

Model Inversion [Fredrikson et al. CCS'15]

GAN attacks [Hitaji et al. CCS'17]

E.g.: given a gender classifier, infer what a male looks like

But…any useful machine learning model does reveal something about the population from which the training data was sampled

Privacy leakage != Adv learns something about training data

Intuition

Intuition

How about if we inferred properties of a subset of the training inputs…

Intuition

How about if we inferred properties of a subset of the training inputs...

...but not of the whole class?

Intuition

How about if we inferred properties of a subset of the training inputs…

…but not of the whole class?

Intuition

How about if we inferred properties of a subset of the training inputs...

...but not of the whole class?

In a nutshell: given a gender classifier, infer race of people in Bob's photos

Agenda

1. Property Inference in Collaborative/Federated ML

2. Membership Inference against Generative Models

Agenda

1. Property Inference in Collaborative/Federated ML

2. Membership Inference against Generative Models

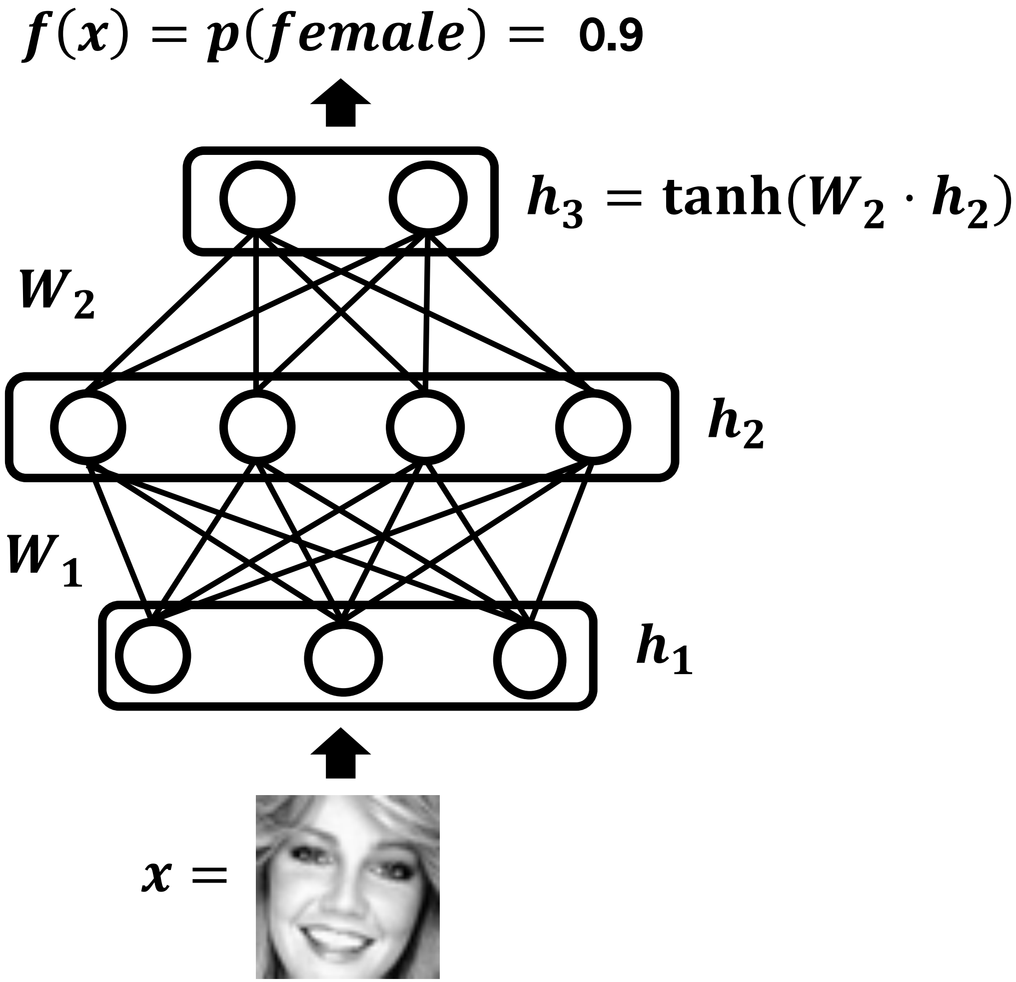
Agenda

1. Property Inference in Collaborative/Federated ML

2. Membership Inference against Generative Models

Luca Melis, Congzheng Song, Emiliano De Cristofaro, Vitaly Shmatikov. Exploiting Unintended Feature Leakage in Collaborative Learning. IEEE Symposium on Security & Privacy (S&P'19)

Deep Learning

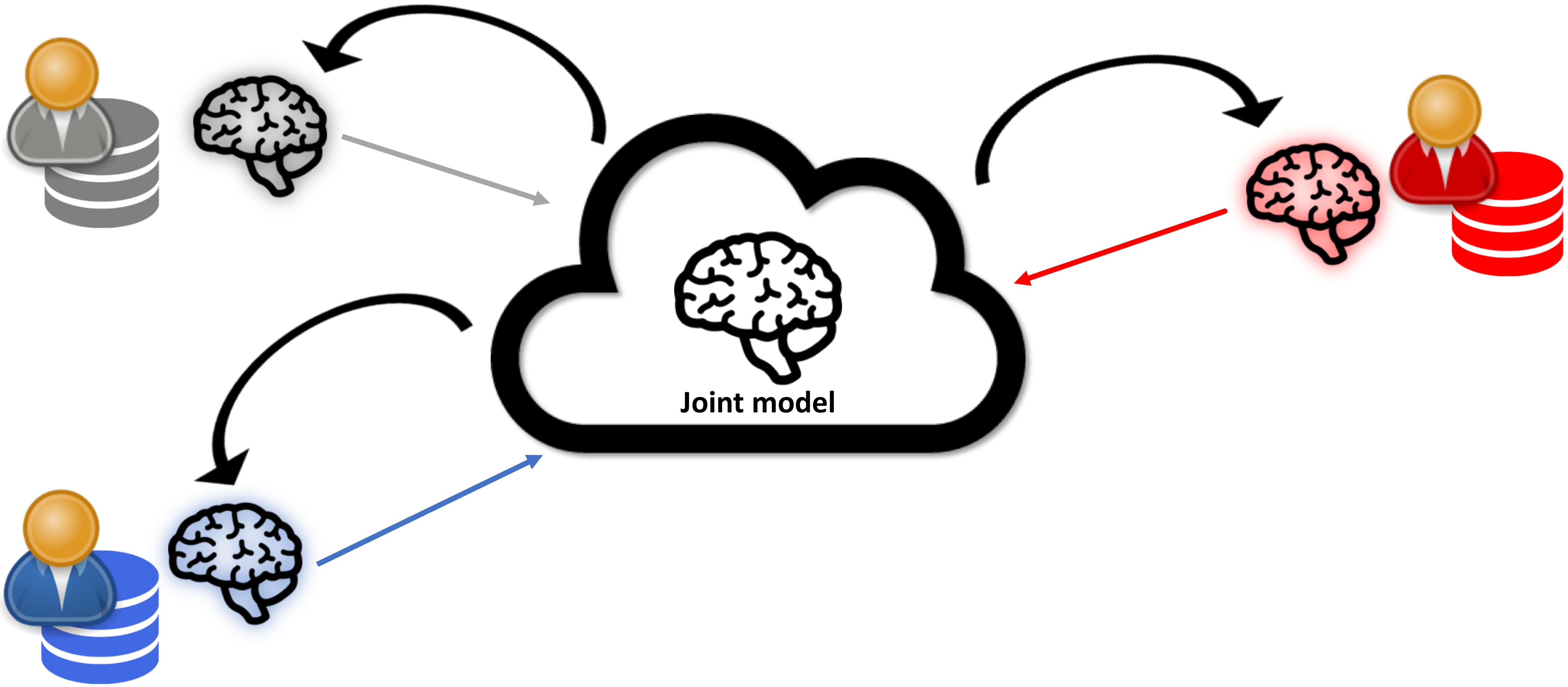


- Map input x to layers of hidden representations h, then to output y
- $h_{l+1} = a(W_l \cdot h_l)$ with parameter W_l
- Train model to minimizes loss:

$$W = \text{argmin}_W L(f(x), y)$$

- Gradient descent on parameters:
  - Each iteration train on a batch
  - Update W based on gradient of L

Collaborative/Federated Learning



Collaborative

Federated

Algorithm 1 Parameter server with synchronized SGD

Server executes:

Initialize θ_0
for $t = 1$ to T **do**
  for each client k **do**
    $g_t^k \leftarrow$**ClientUpdate**(θ_{t-1})
  end for
  $\theta_t \leftarrow \theta_{t-1} - \eta \sum_k g_t^k$
end for

ClientUpdate(θ):

Select batch b from client's data
return local gradients $\nabla L(b;\theta)$

Algorithm 2 Federated learning with model averaging

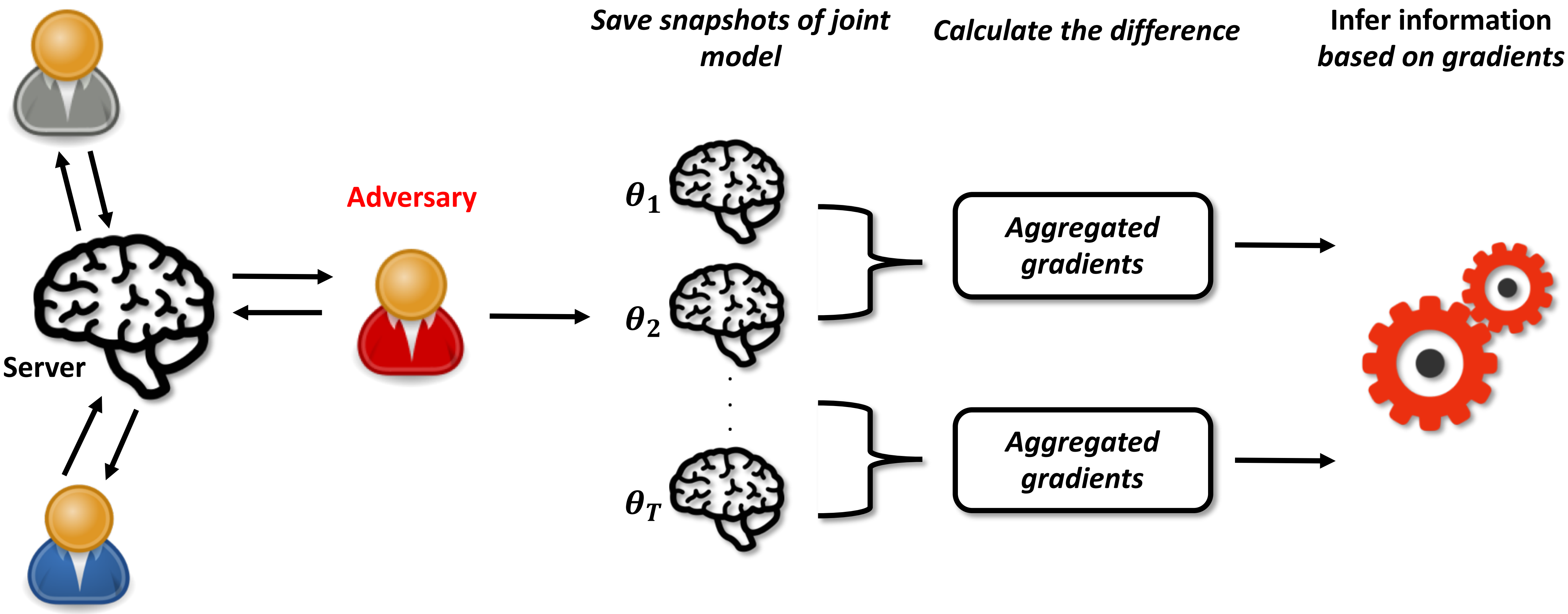
Server executes:

Initialize θ_0
$m \leftarrow max(C \cdot K, 1)$
for $t = 1$ to T **do**
  $S_t \leftarrow$ (random set of m clients)
  for each client $k \in S_t$ **do**
    $\theta_t^k \leftarrow$**ClientUpdate**(θ_{t-1})
  end for
  $\theta_t \leftarrow \sum_k \frac{n^k}{n} \theta_t^k$
end for

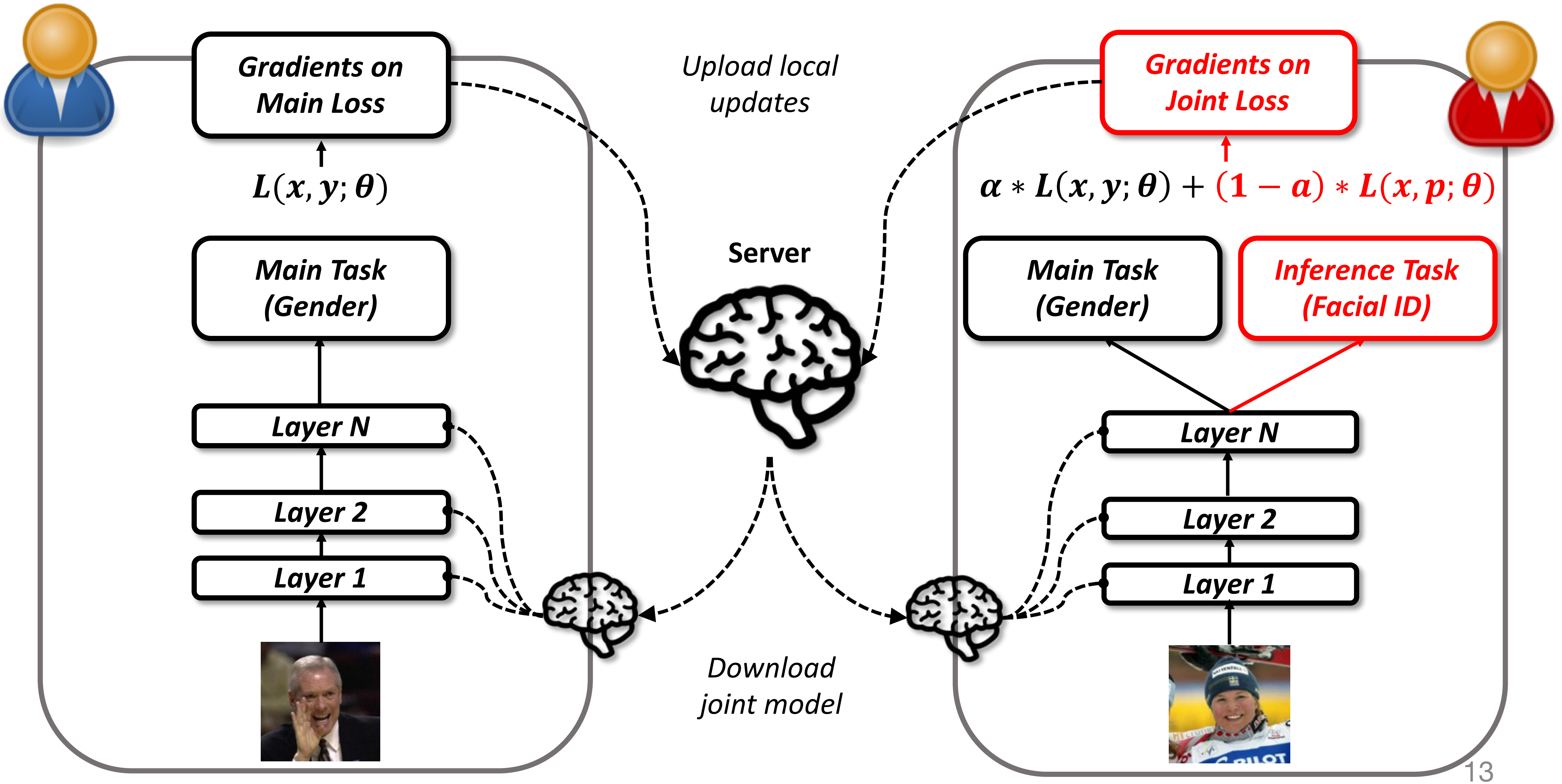
ClientUpdate(θ):

for each local iteration **do**
  for each batch b in client's split **do**
    $\theta \leftarrow \theta - \eta \nabla L(b;\theta)$
  end for
end for
return local model θ

Passive Property Inference Attack



Active Property Inference Attack

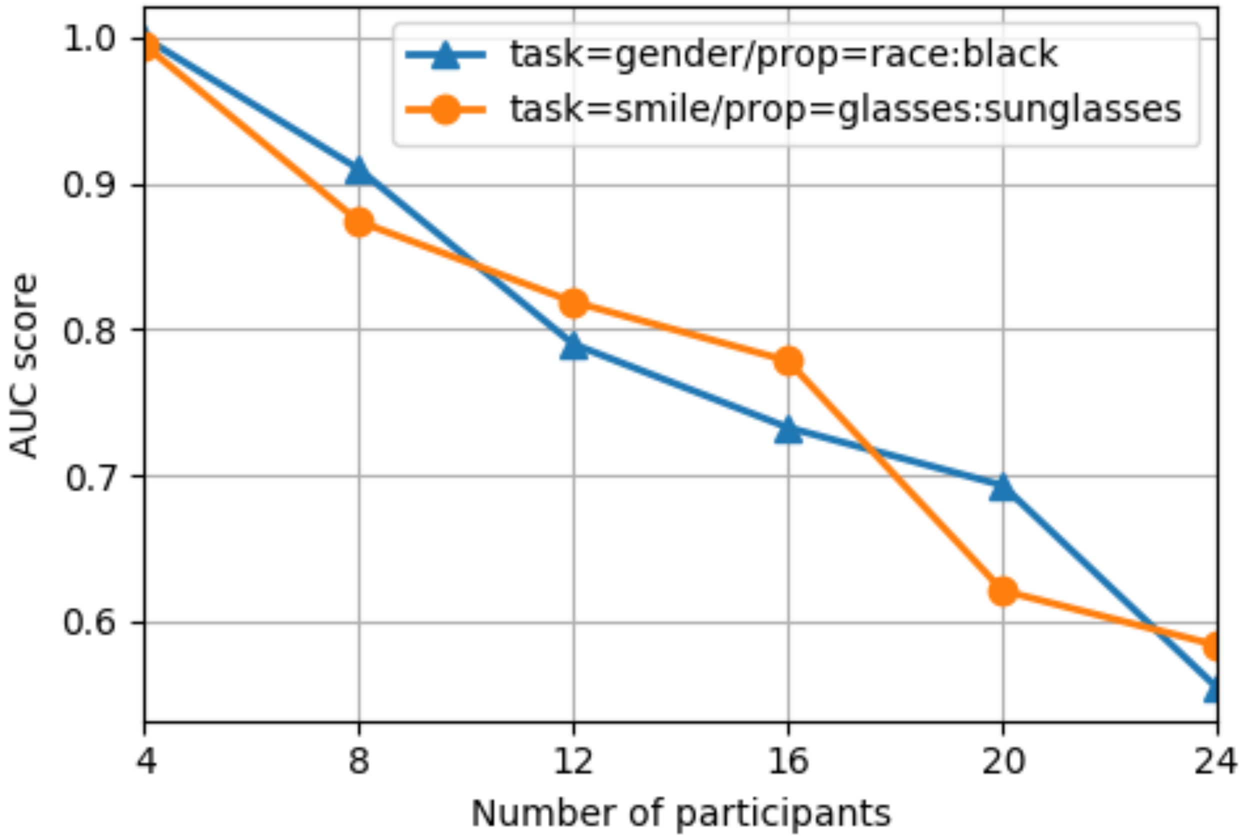


Dataset	Type	Main Task	Inference Task
LFW	Images	Gender/Smile/Age Eyewear/Race/Hair	Race/Eyewear
FaceScrub	Images	Gender	Identity
PIPA	Images	Age	Gender
FourSquare	Locations	Gender	Membership
Yelp-health	Text	Review Score	Membership Doctor specialty
Yelp-author	Text	Review Score	Author
CSI	Text	Sentiment	Membership Region/Gender/Veracity

Property Inference on LFW

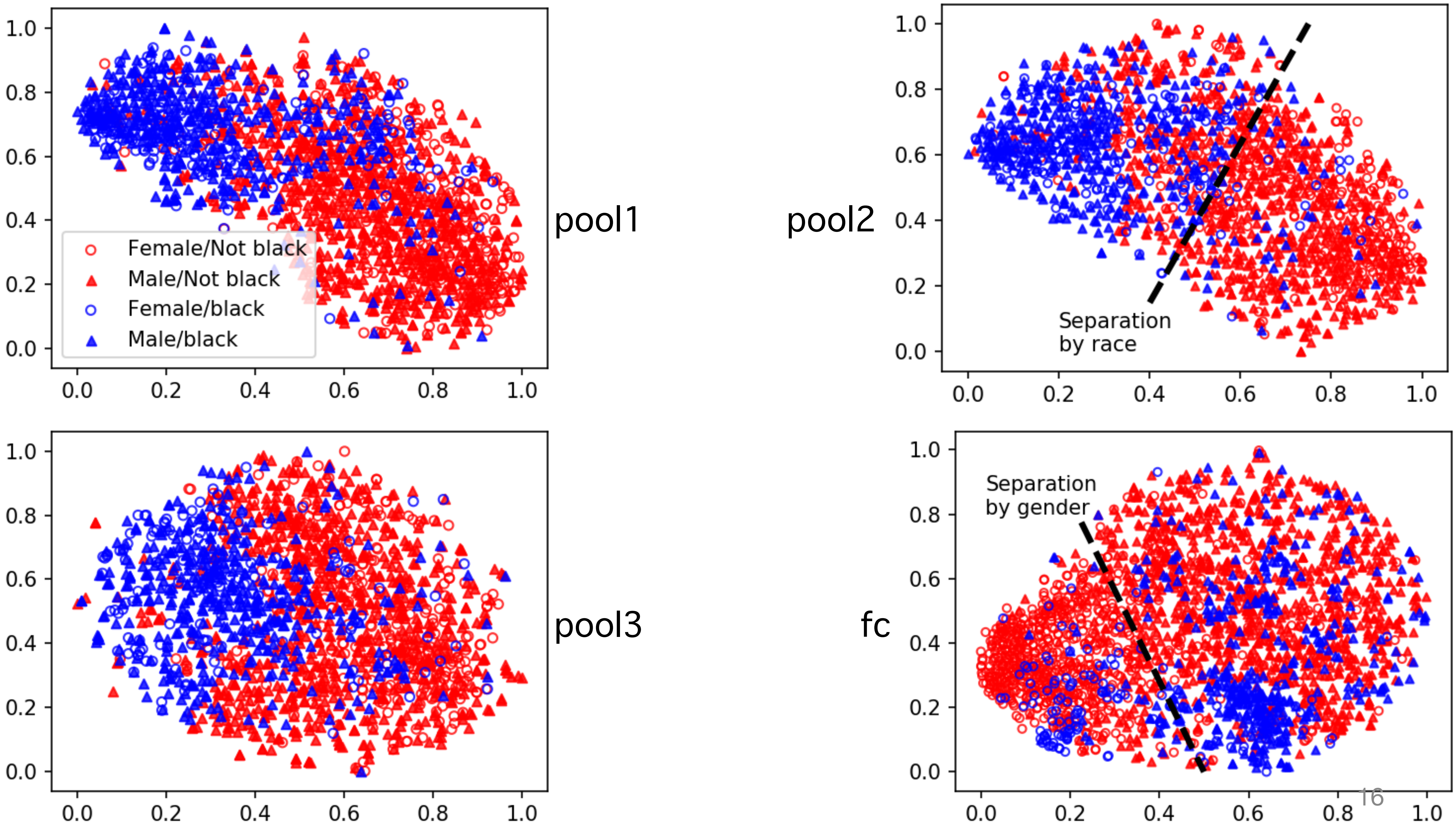
Main Task	Inference Task	Correlation	AUC score
Gender	Sunglasses	-0.025	1.0
Smile	Asian	0.047	0.93
Age	Black	-0.084	1.0
Race	Sunglasses	0.026	1.0
Eyewear	Asian	-0.119	0.91
Hair	Sunglasses	-0.013	1.0

Two-Party



Multi-Party

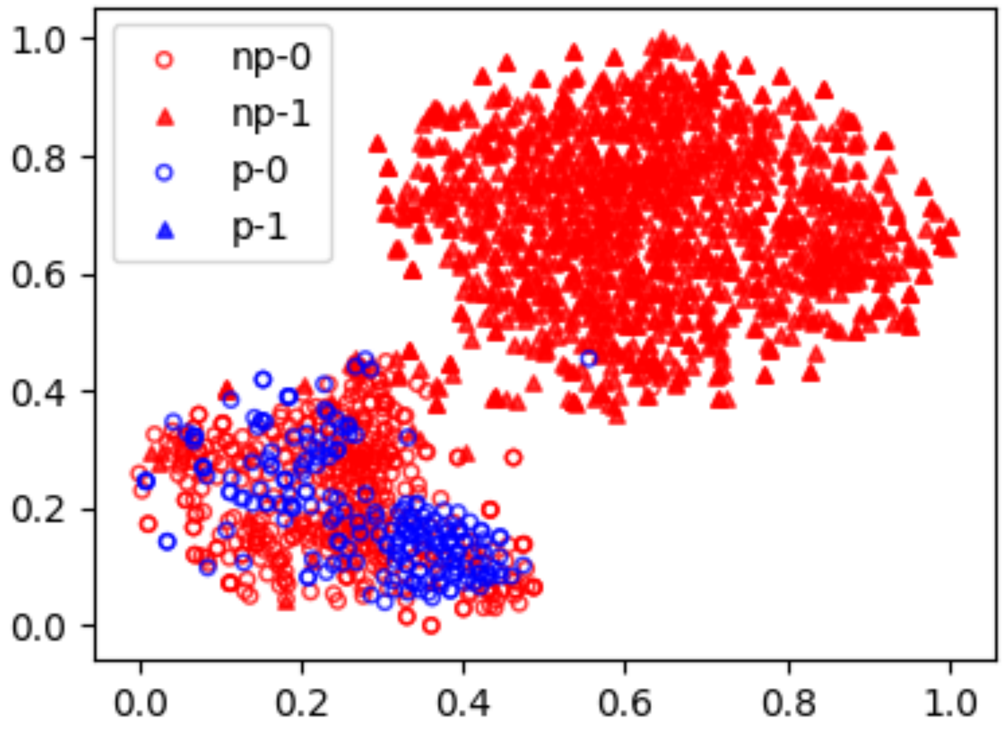
Feature t-SNE projection


pool1
pool2
pool3
fc
Female/Not black
Male/Not black
Female/black
Male/black
Separation
by race
Separation
by gender
0.0
0.2
0.4
0.6
0.8
1.0

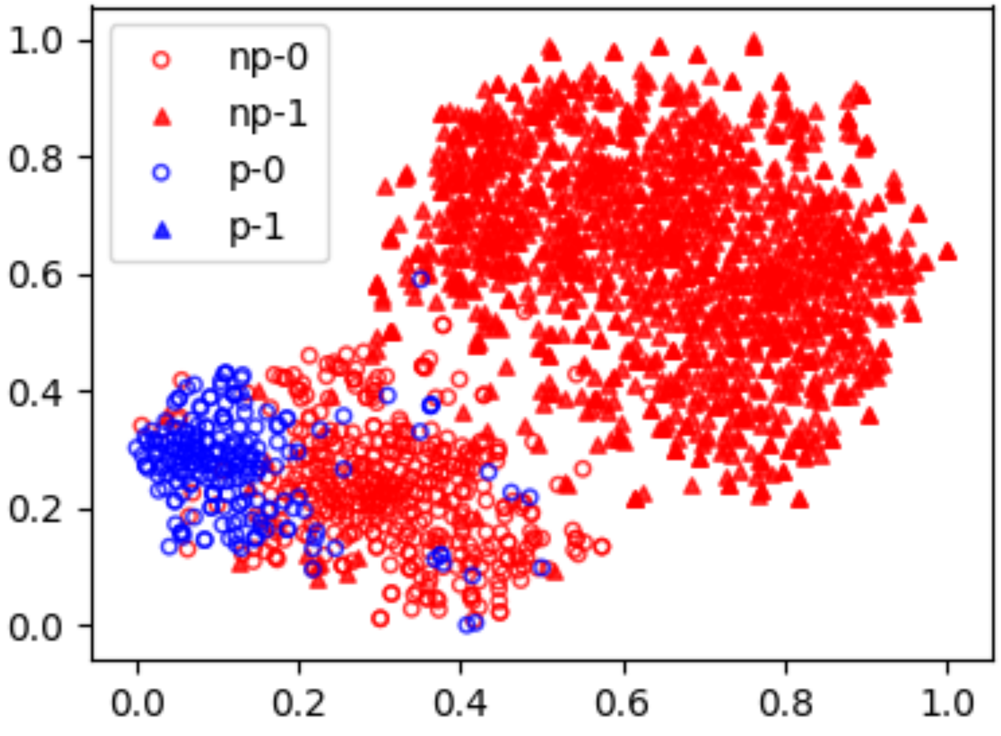
Passive vs Active Attack on FaceScrub

Main Task: ▲/●= female/male

Inference Task: Blue points with the property (identity)



Passive attack

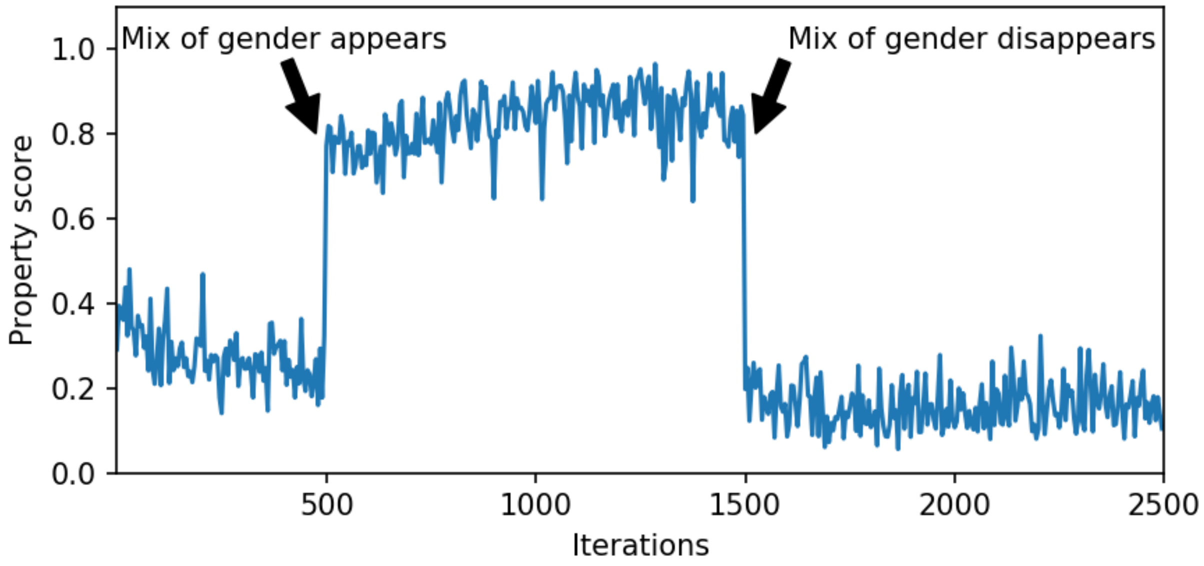


Active attack

Inferring when a property occurs

Inferring when a property occurs

Batches with the property appear

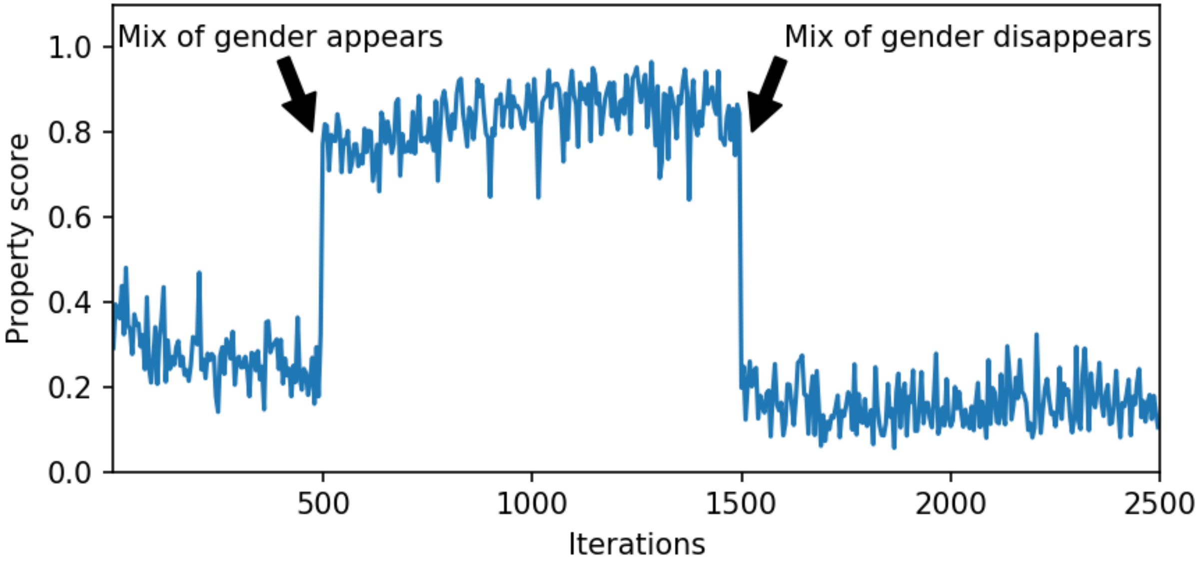


Main task: Age / Two-party

Inference task: people in the image are of the same gender (PIPA)

Inferring when a property occurs

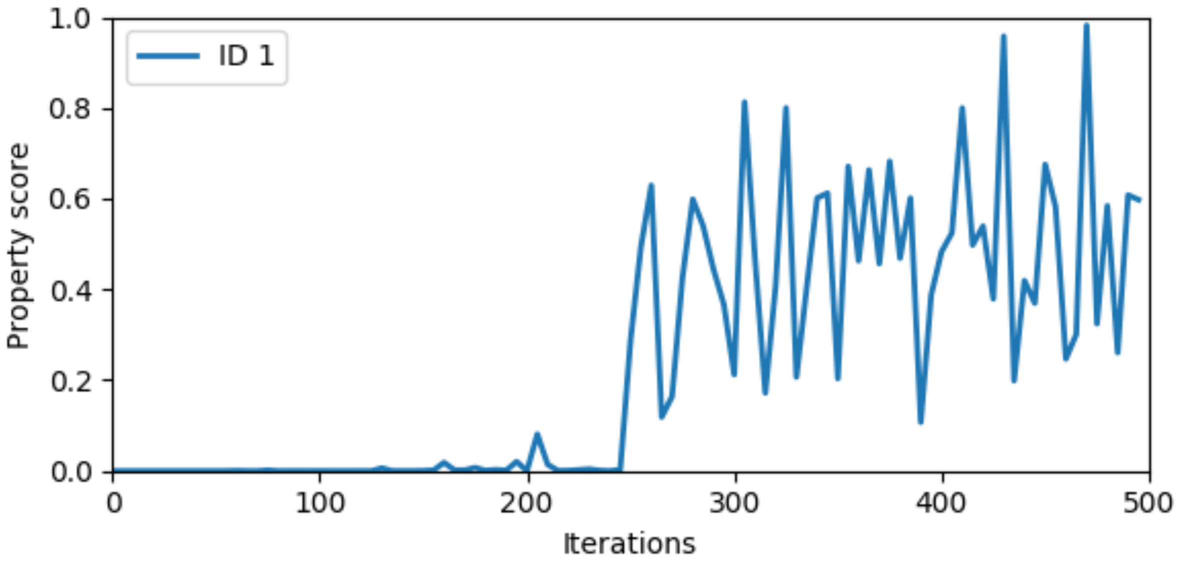
Batches with the property appear



Main task: Age / Two-party

Inference task: people in the image are of the same gender (PIPA)

Participant with ID 1 joins training



Main task: Gender / Multi-Party

Inference task: author identification

Defenses?

Defenses?

Selective gradient sharing

Dataset: Text reviews

Main Task: Sentiment classifier

Doesn’t really work…

Property / % parameters shared	10%	50%	100%
Top region	0.84	0.86	0.93
Gender	0.90	0.91	0.93
Veracity	0.94	0.99	0.99

Defenses?

Selective gradient sharing

Dataset: Text reviews

Main Task: Sentiment classifier

Doesn't really work…

Property / % parameters shared	10%	50%	100%
Top region	0.84	0.86	0.93
Gender	0.90	0.91	0.93
Veracity	0.94	0.99	0.99

Participant-level differential privacy

Hide participant's contributions

Only two mechanisms in the literature

Fail to converge for "few" participants

Agenda

1. Property Inference in Collaborative/Federated ML

2. Membership Inference against Generative Models

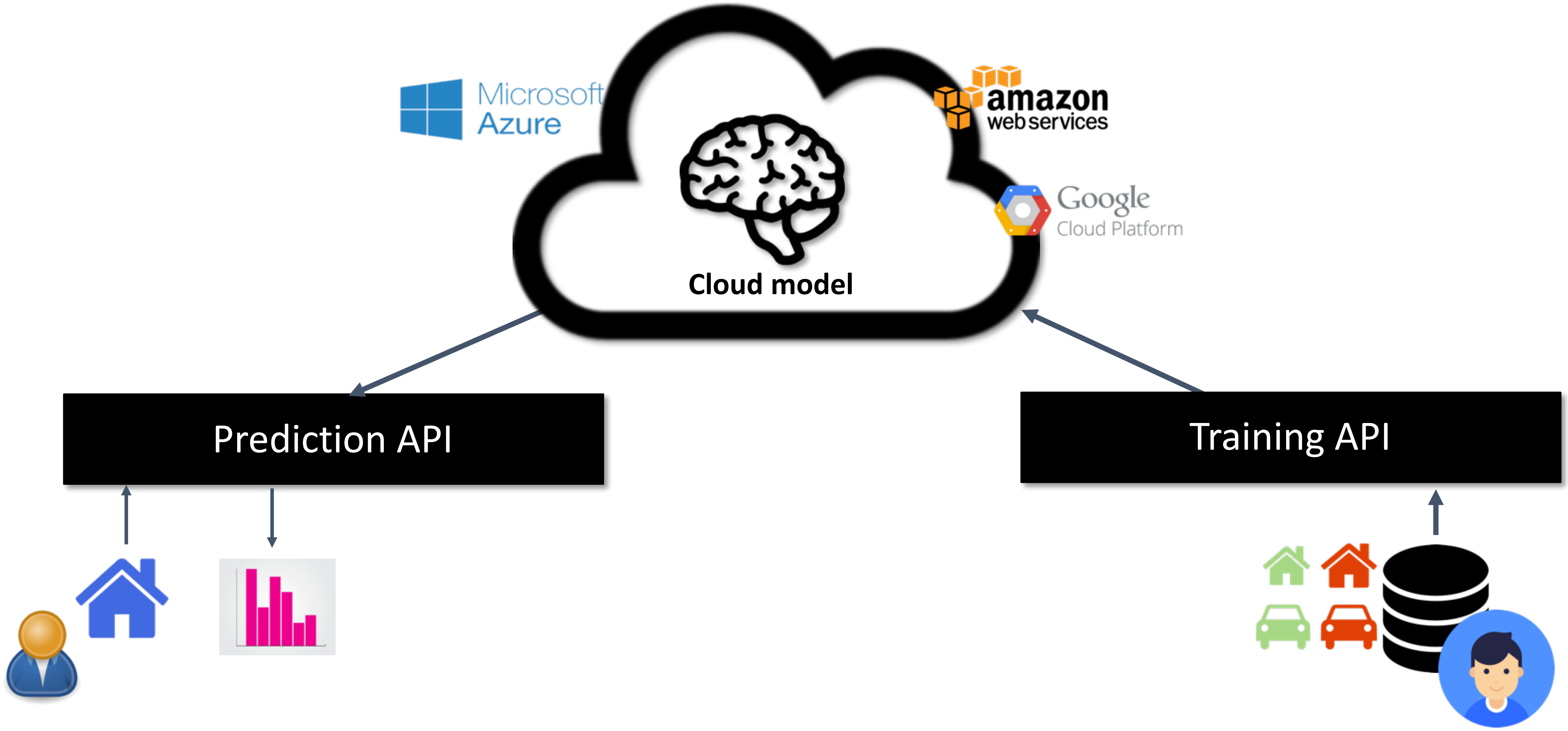
Agenda

1. Property Inference in Collaborative/Federated ML

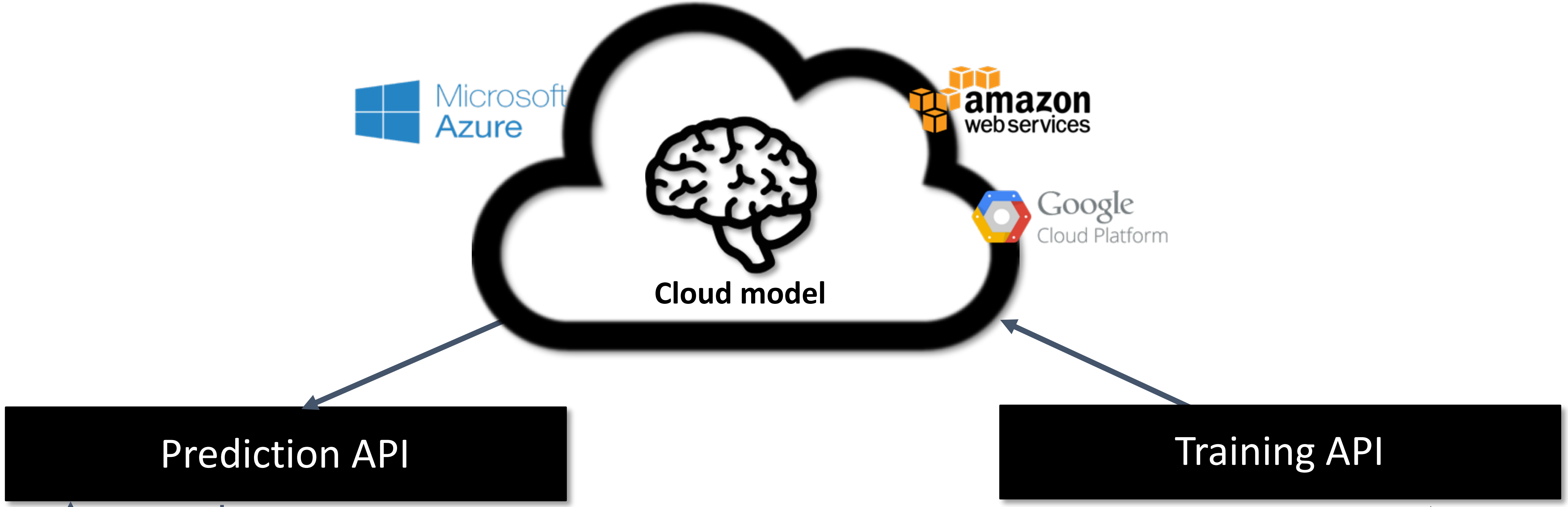
2. Membership Inference against Generative Models

Machine Learning as a Service

Machine Learning as a Service



Machine Learning as a Service

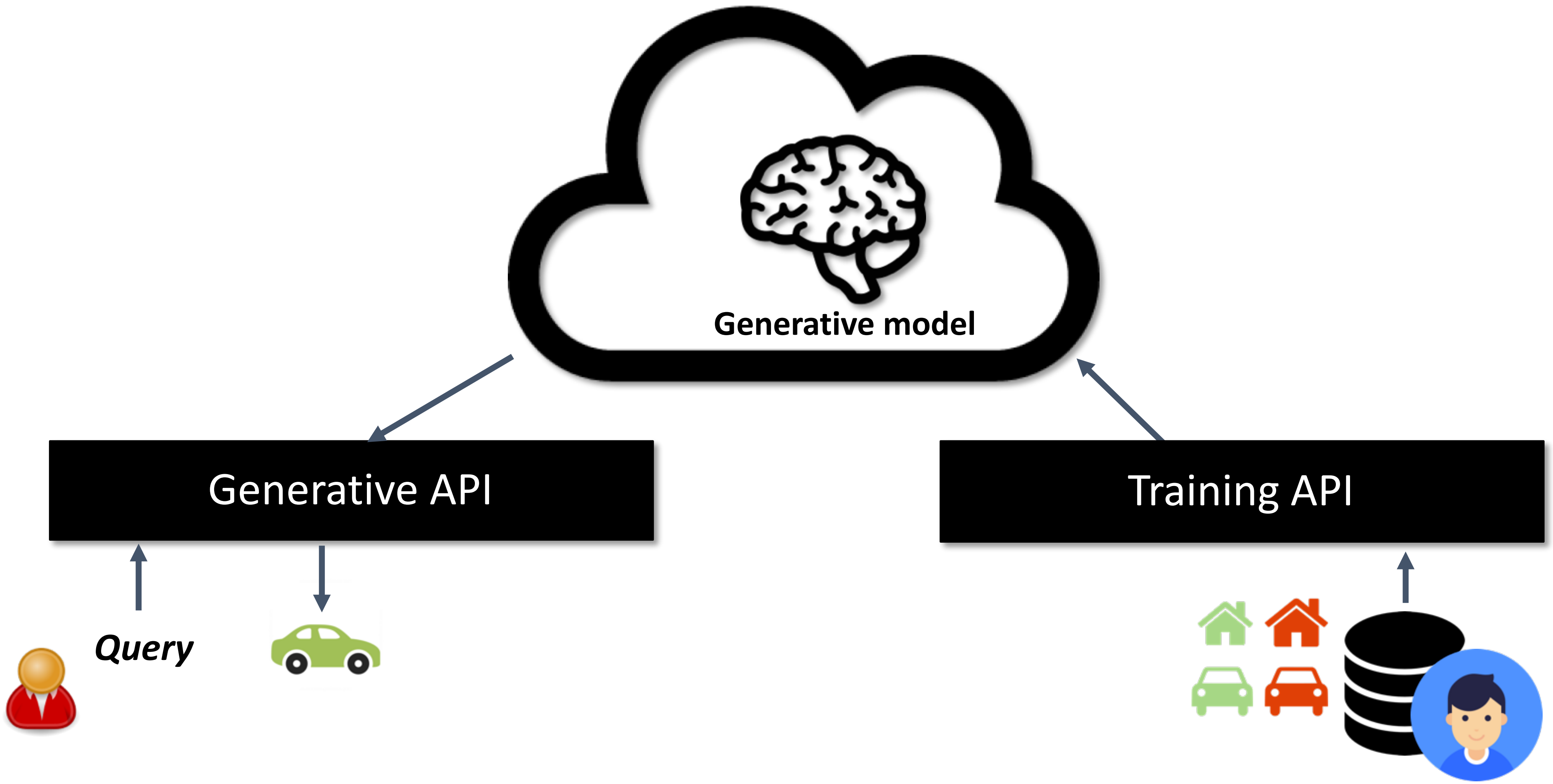


Predictions are leaky!

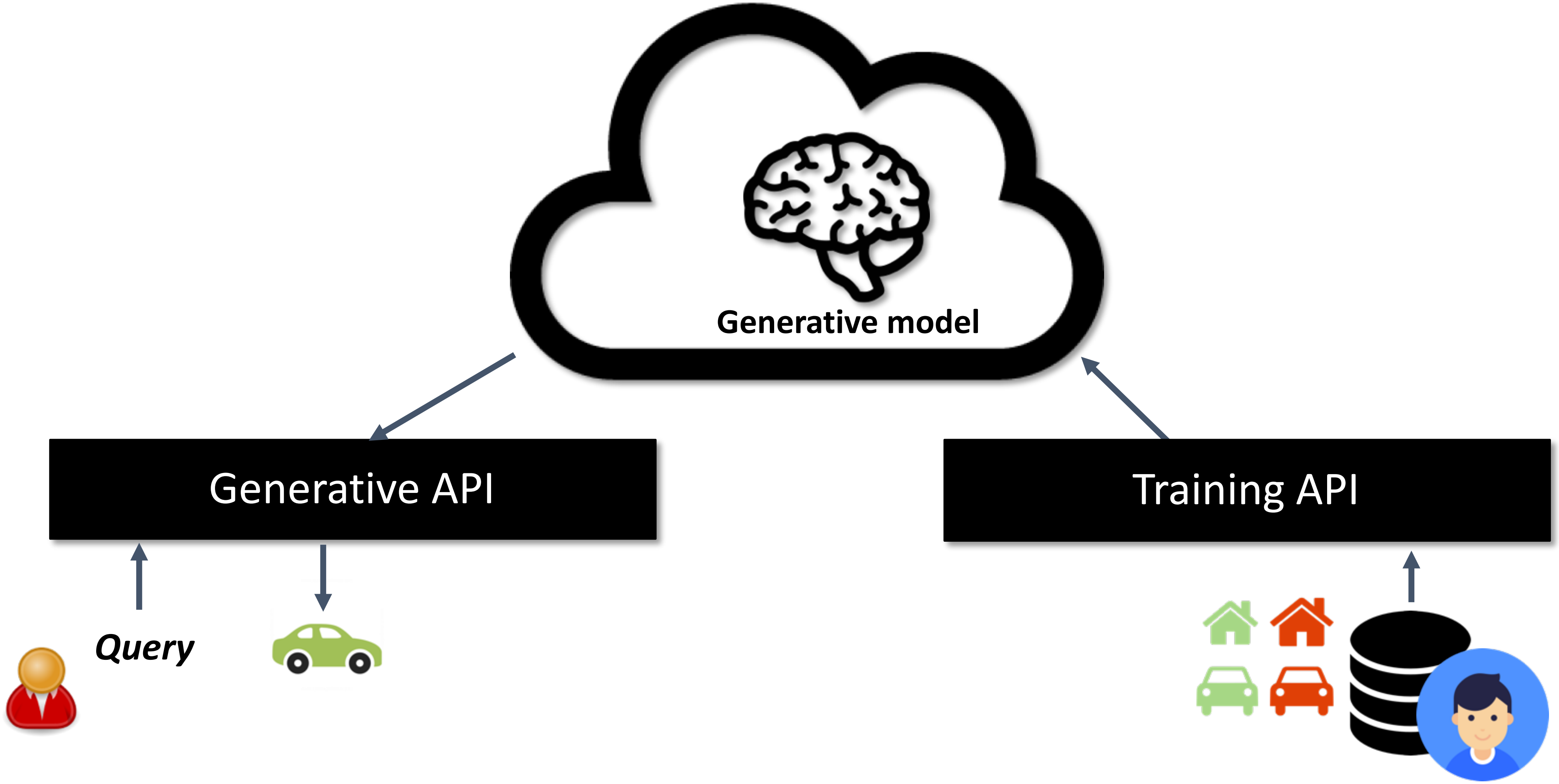
Shokri et al. Membership inference attacks against machine learning models. S&P'17

Membership Inference in Generative Models

Membership Inference in Generative Models



Membership Inference in Generative Models



Jamie Hayes, Luca Melis, George Danezis, Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. PETS 2019.

Inference without predictions?

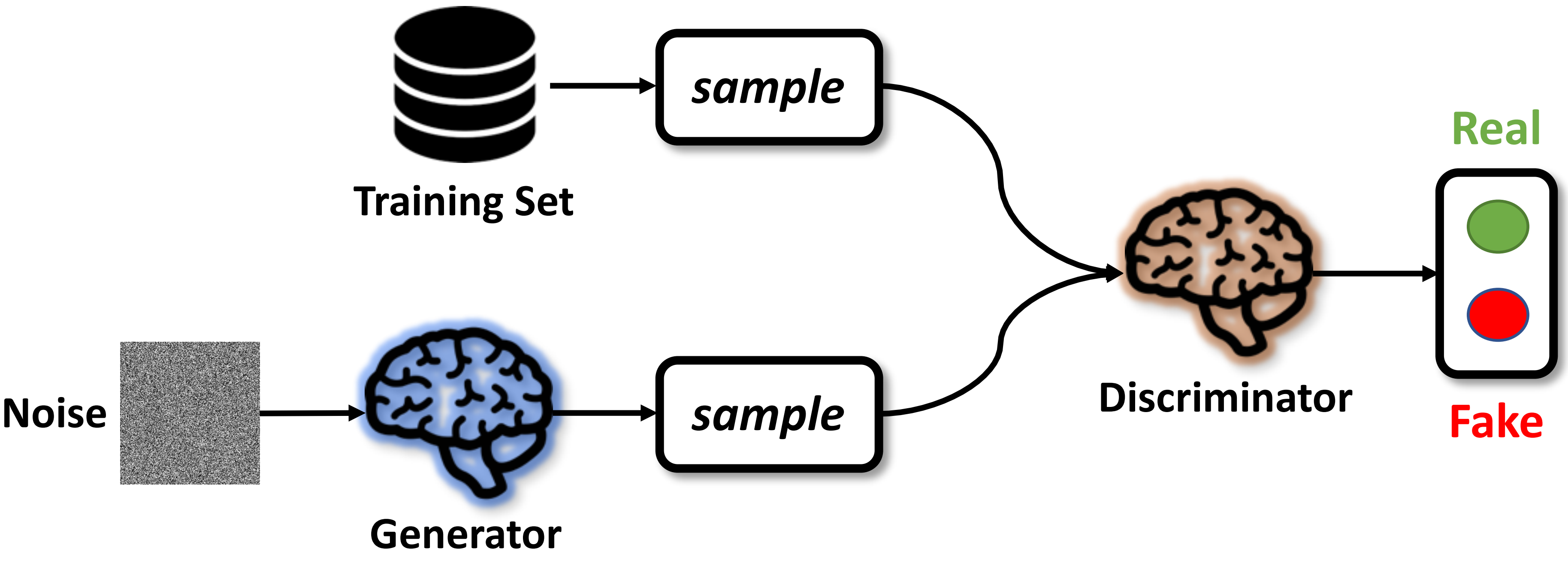
Use generative models!

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

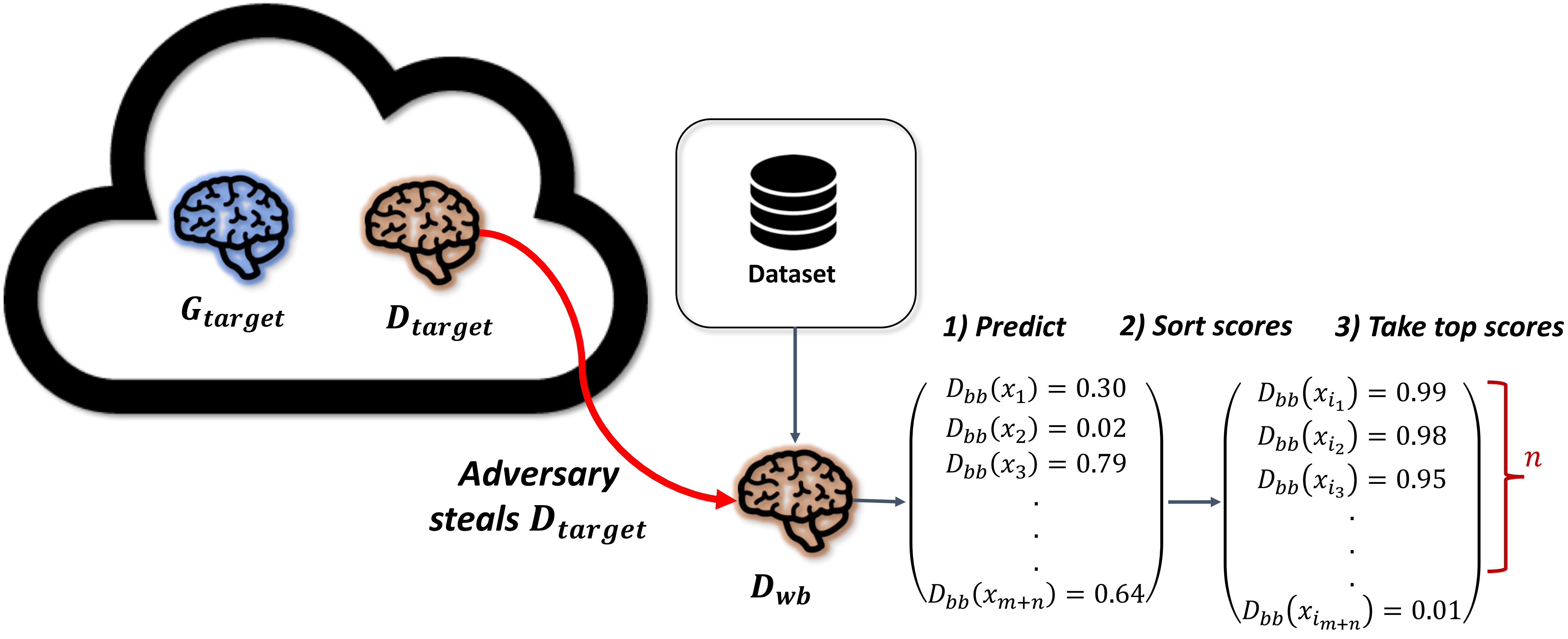
Inference without predictions?

Use generative models!

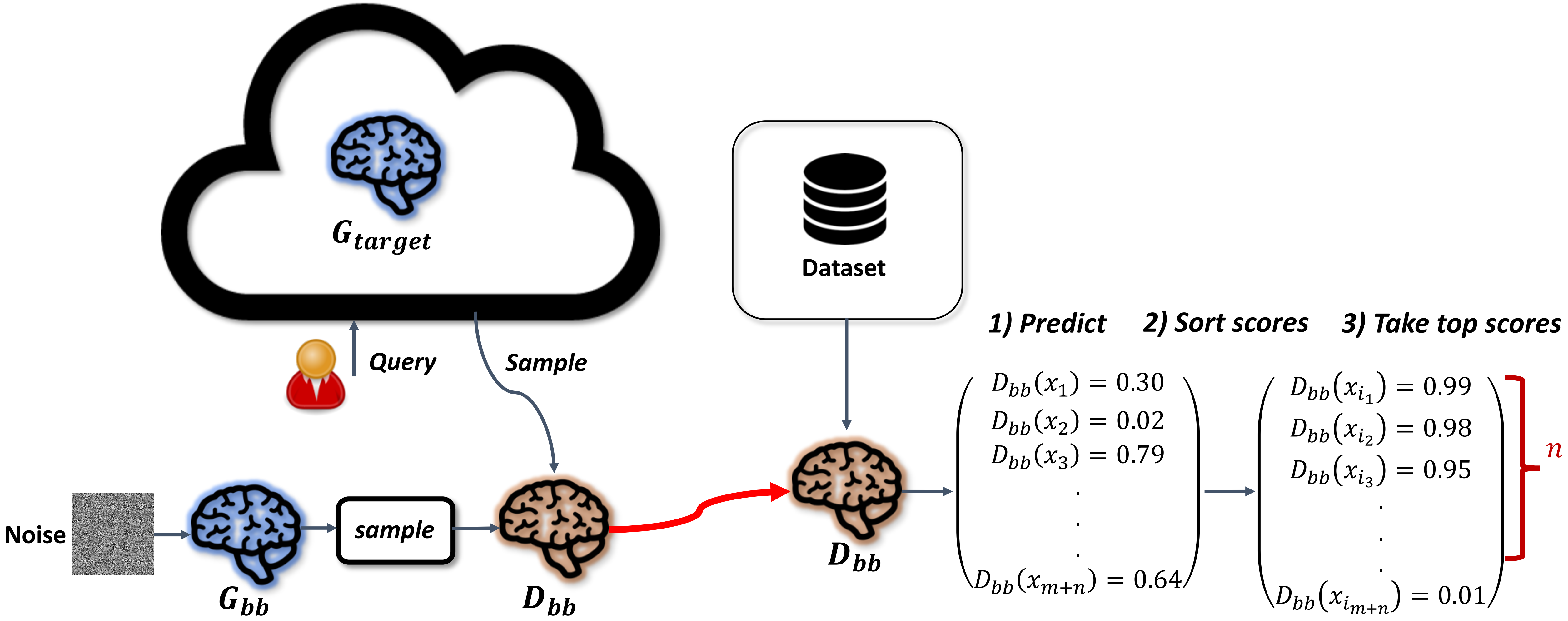
Train GANs to learn the distribution and a prediction model at the same time



White-Box Attack



Black-Box Attack

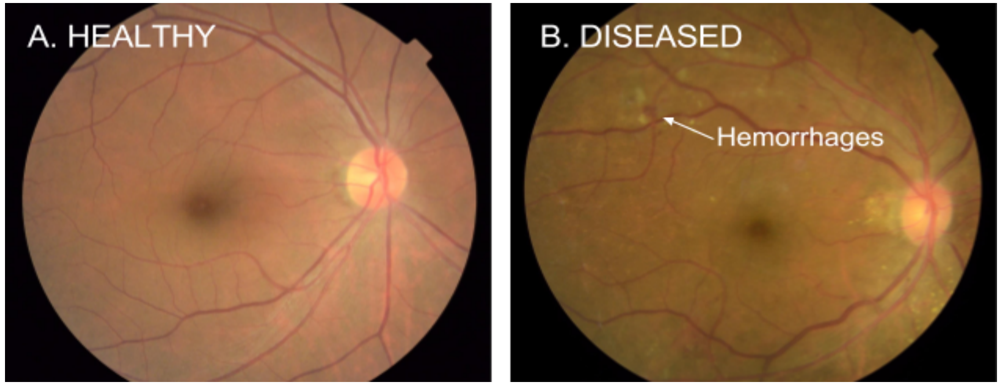


Datasets

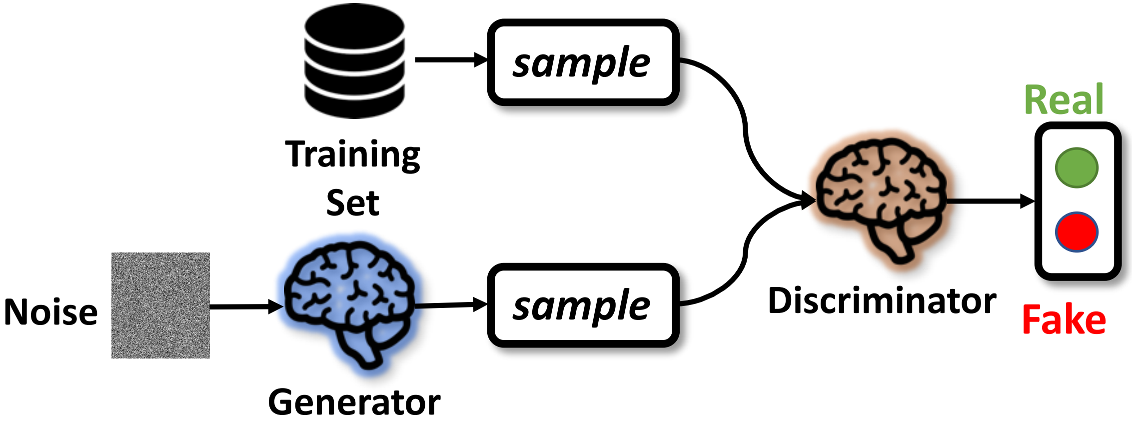
LFW

CIFAR-10

DR



Models



Attacker Model:

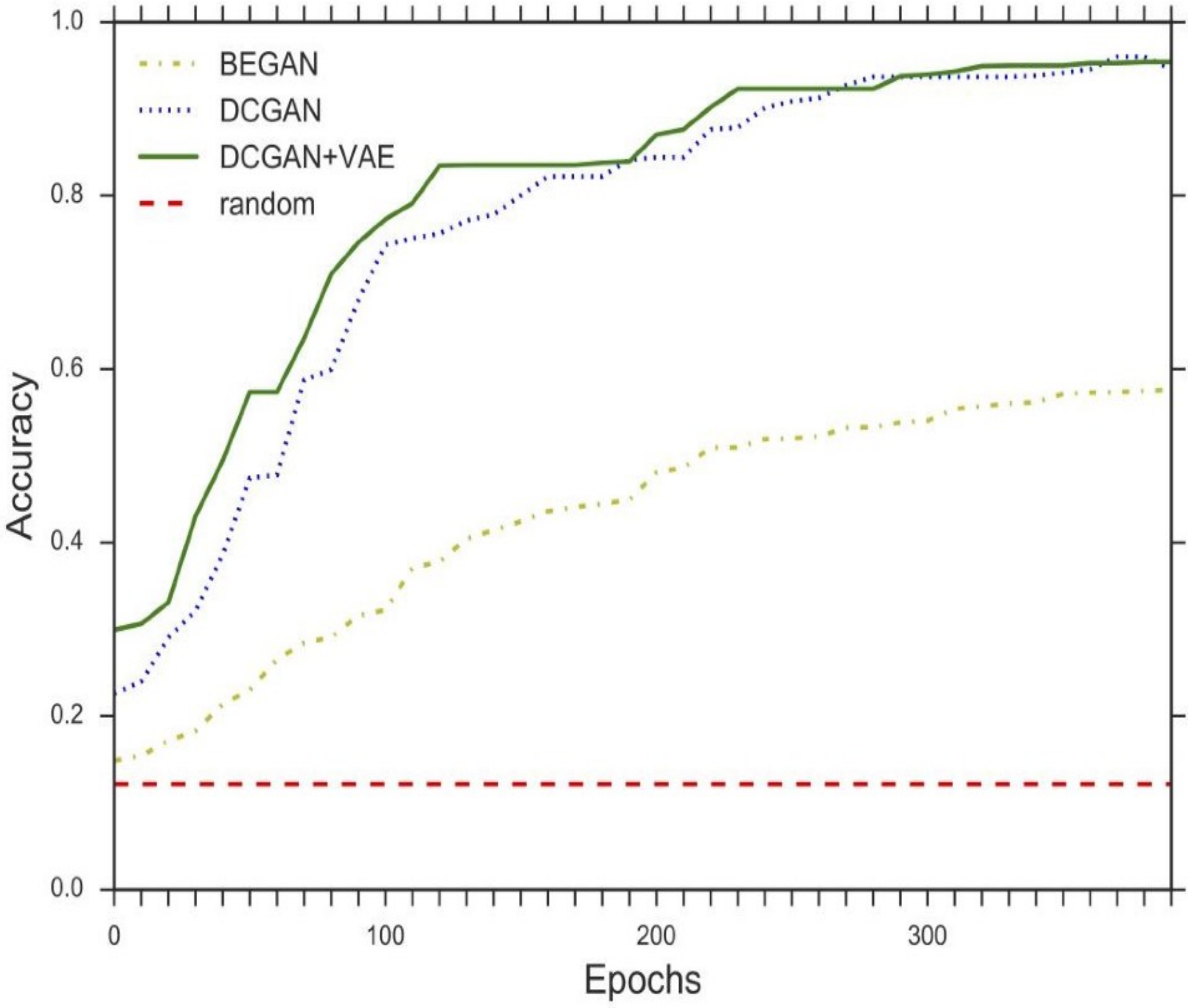
DCGAN

Target Model:

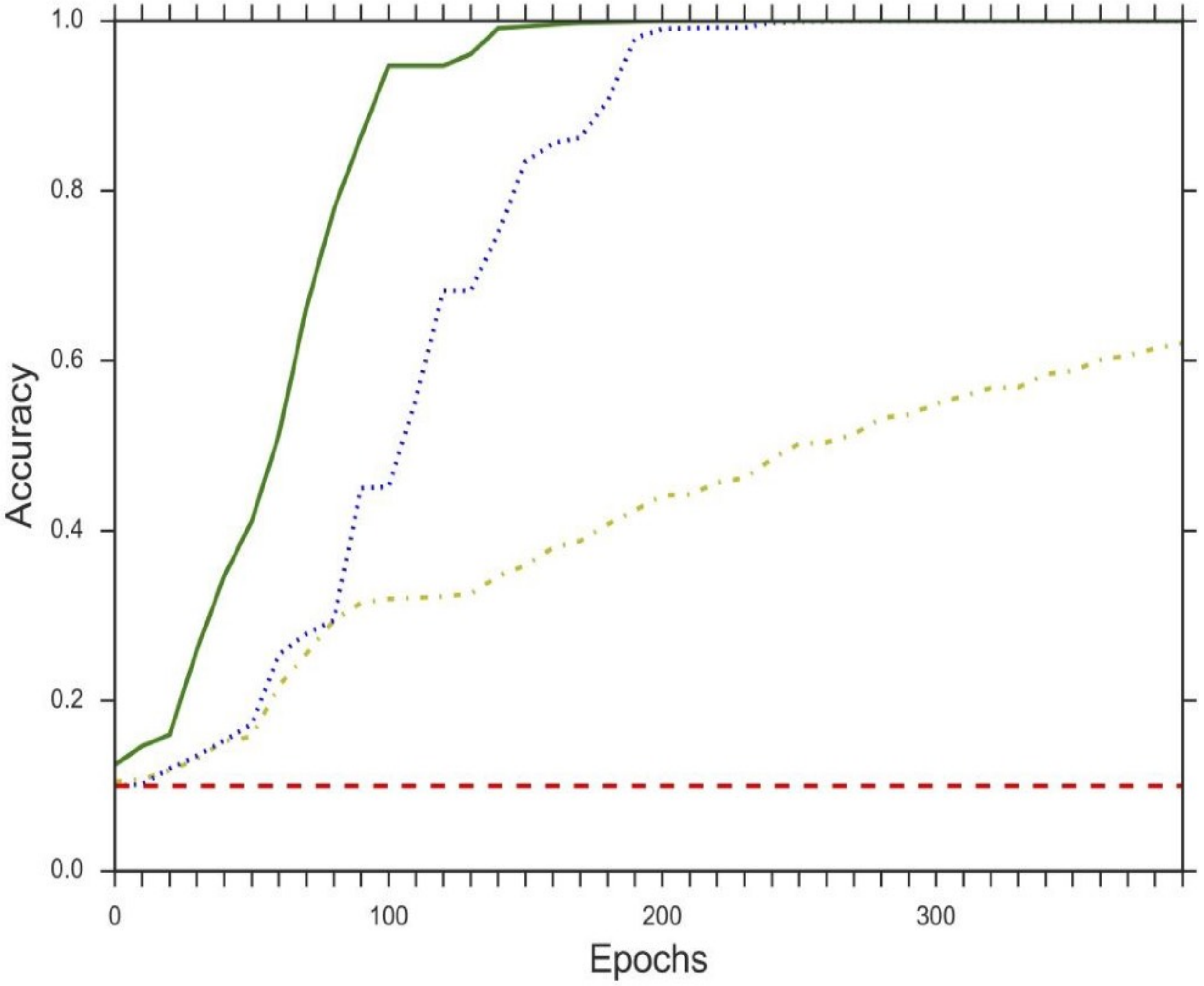
DCGAN, DCGAN+VAE, BEGAN

White-Box Results

LFW, top ten classes

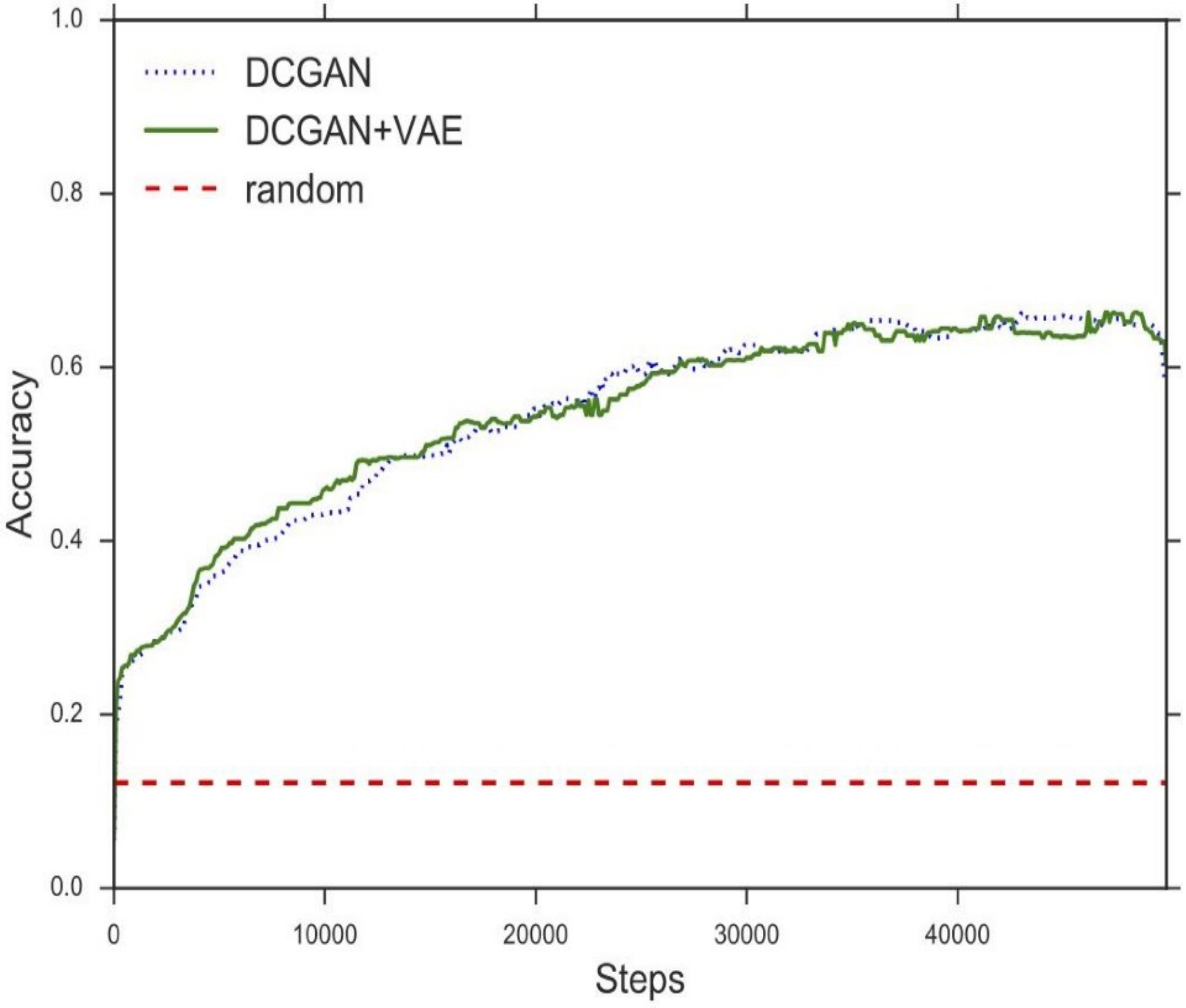


CIFAR-10, random 10% subset

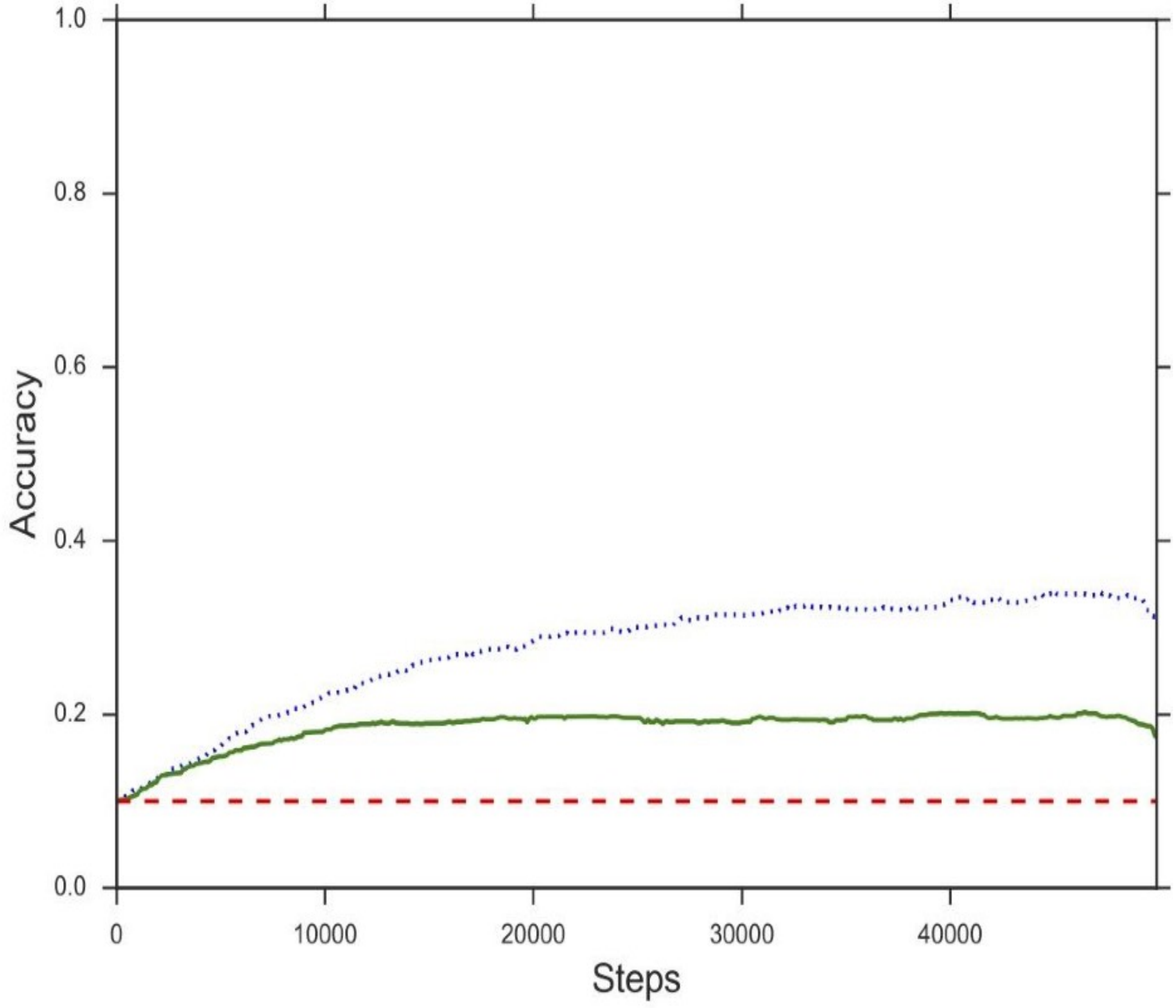


Black-Box Results

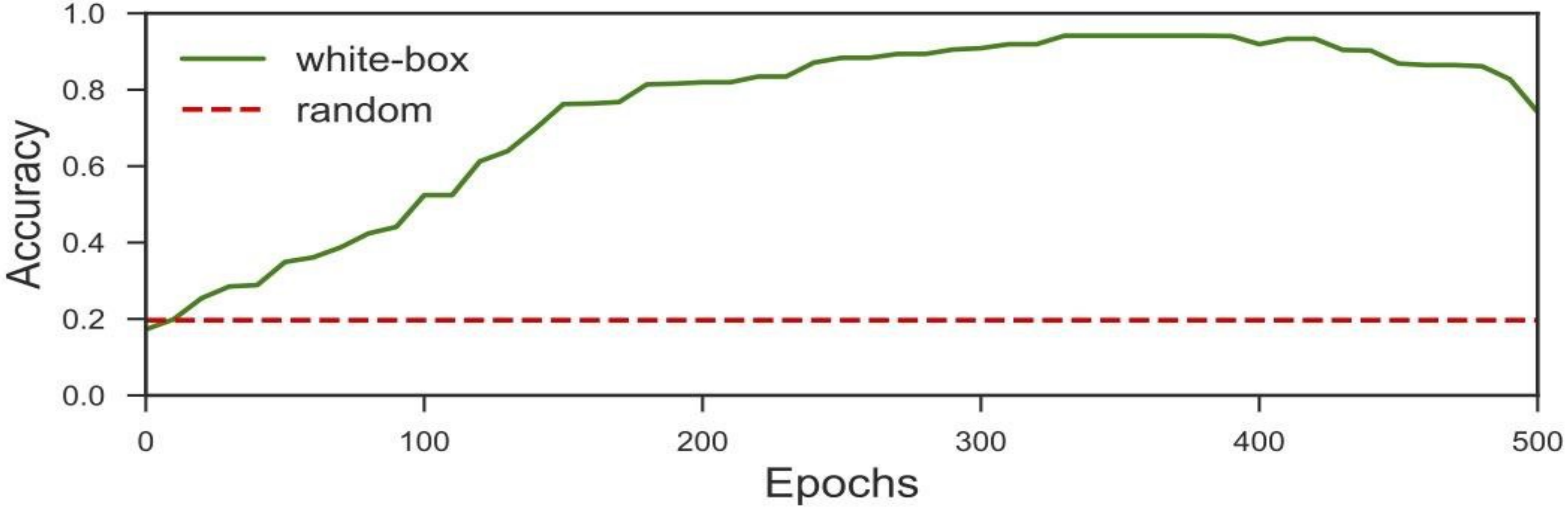
LFW, top ten classes



CIFAR-10, random 10% subset

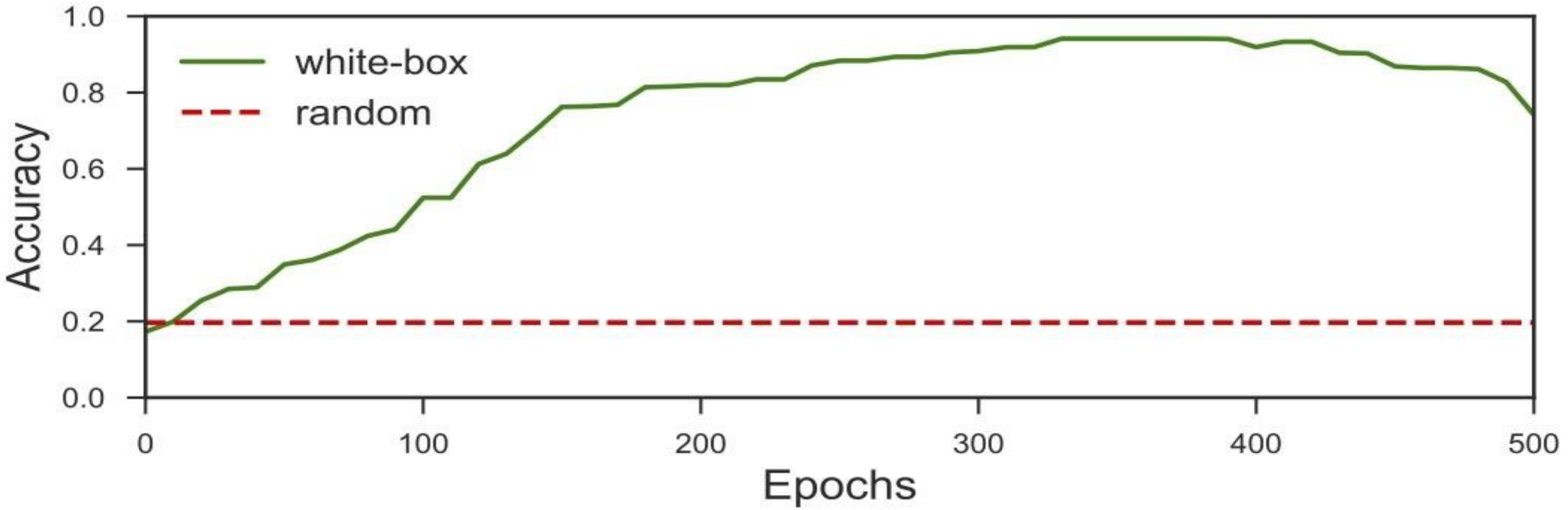


DR Dataset

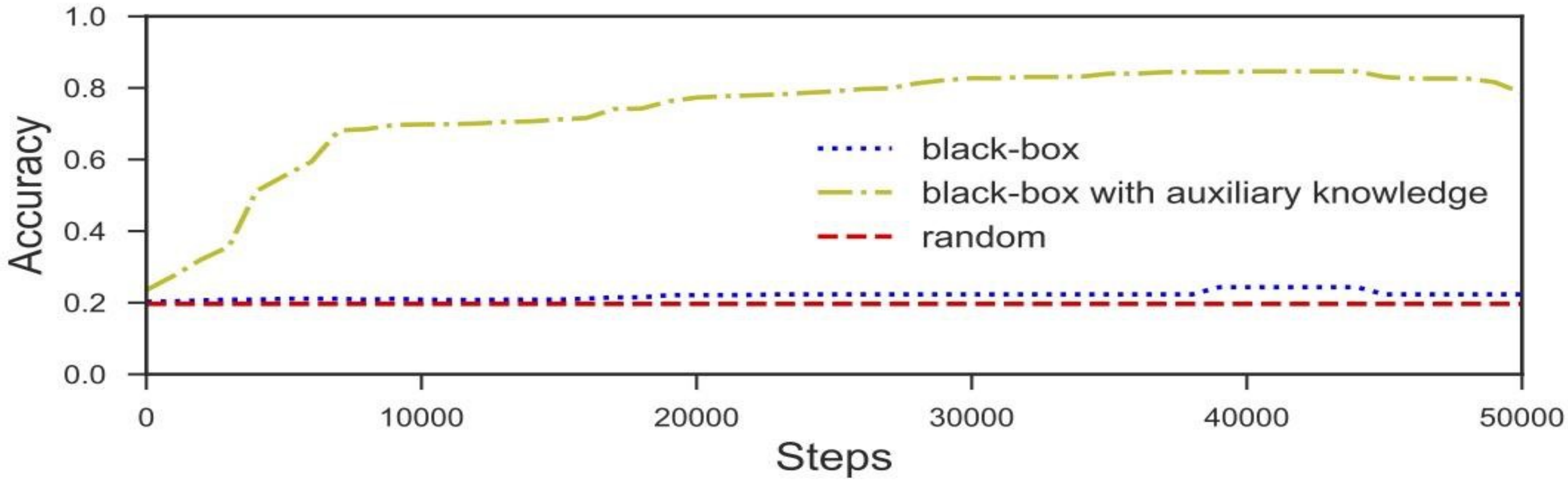


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
white-box
random
0
100
200
300
400
500
Epochs

DR Dataset

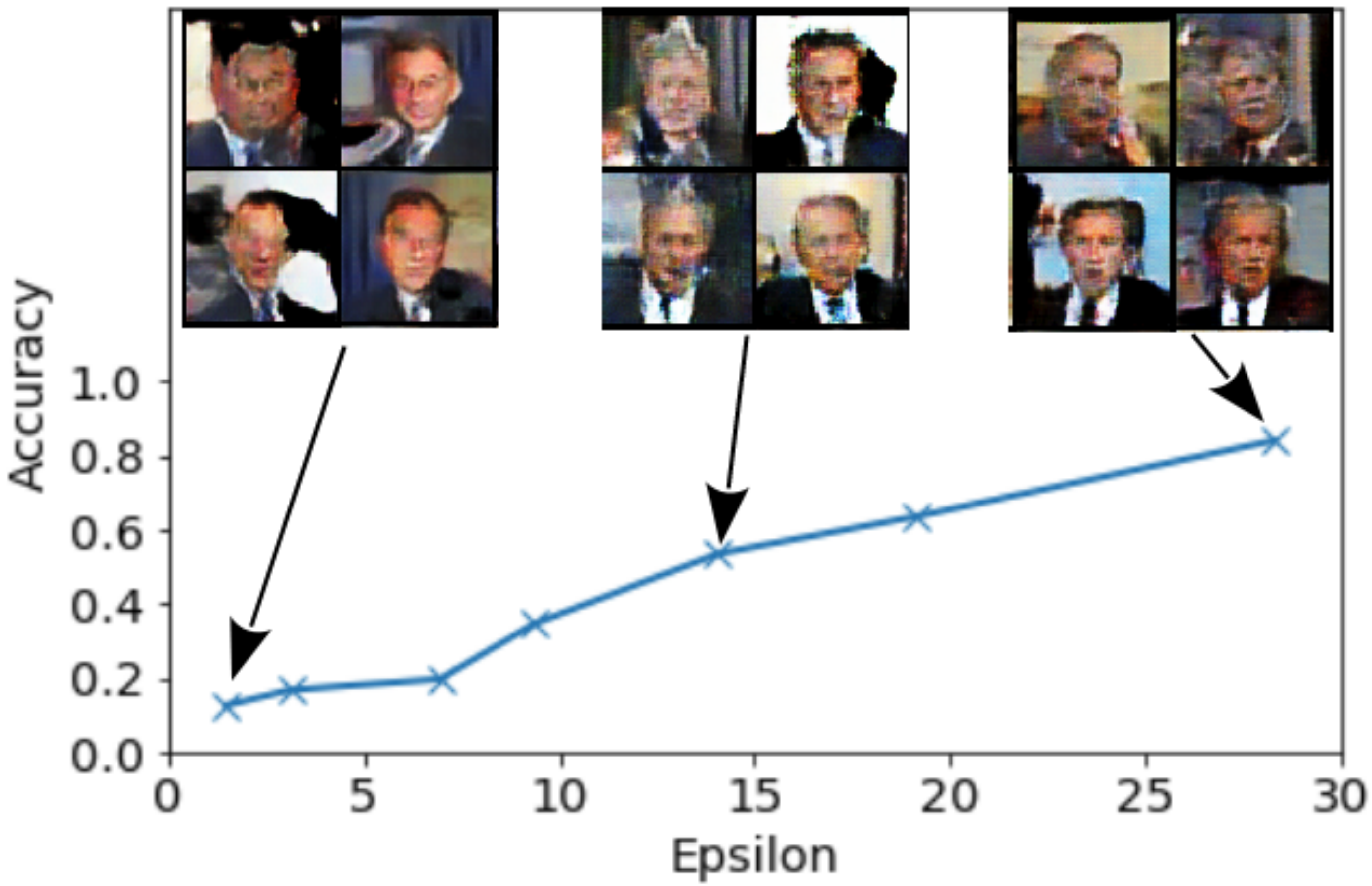


white-box
random
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Epochs



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

Defense? Differentially Private GAN*



White-box, LFW, top ten classes

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

Thank you!

Thank you!