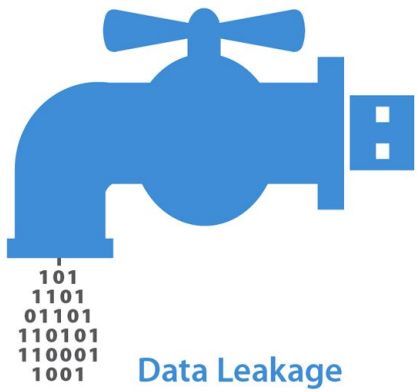
Privacy in ML... It's Complicated

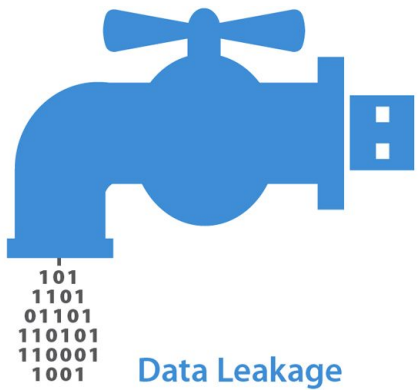
Emiliano De Cristofaro
https://emilianodc.com

Reasoning about "privacy" in ML



Reasoning about "privacy" in ML

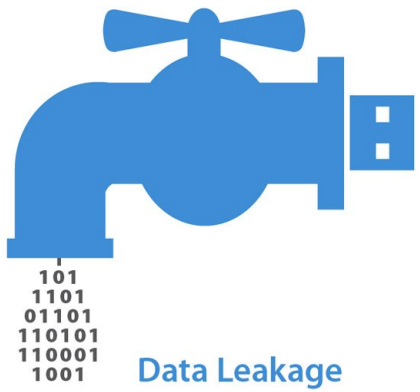
Most privacy attacks in ML focus on inferring either:



Reasoning about "privacy" in ML

Most privacy attacks in ML focus on inferring either:

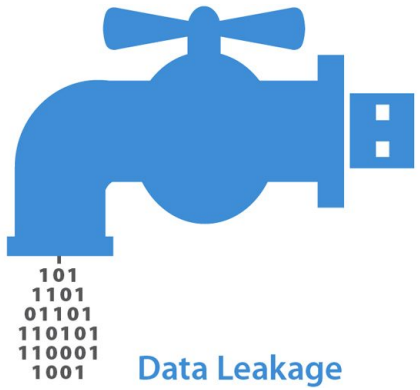
1. Inclusion of a data point in the training set
   (aka "membership inference")



Reasoning about "privacy" in ML

Most privacy attacks in ML focus on inferring either:

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



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 ∈ 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 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 leakage)

- Use it to establish wrongdoing
- Or to assess protection, e.g., with differentially private noise

2. Inferring Class Representatives

2. Inferring Class Representatives

Prior work focused on properties of an entire class, e.g.:

2. Inferring Class Representatives

Prior work focused on properties of an entire class, e.g.:

- Model Inversion [Fredrikson et al. CCS'15]

2. Inferring Class Representatives

Prior work focused on 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 focused on 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 focused on 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…shouldn't useful machine learning models reveal something about population from which training data was sampled

2. Inferring Class Representatives

Prior work focused on 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…shouldn't useful machine learning models reveal something about population from which training data was sampled

Privacy leakage != Adv learns something about training data

2. Inferring Class Representatives

Prior work focused on 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…shouldn't useful machine learning models reveal something about population from which 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

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

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

...but not of the whole class?

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

...but not of the whole class?

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

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

Let's call this a
Property Inference Attack

Attacks

Attacks

1. Membership Inference

Attacks

1. Membership Inference

2. Property Inference in Collaborative/Federated ML

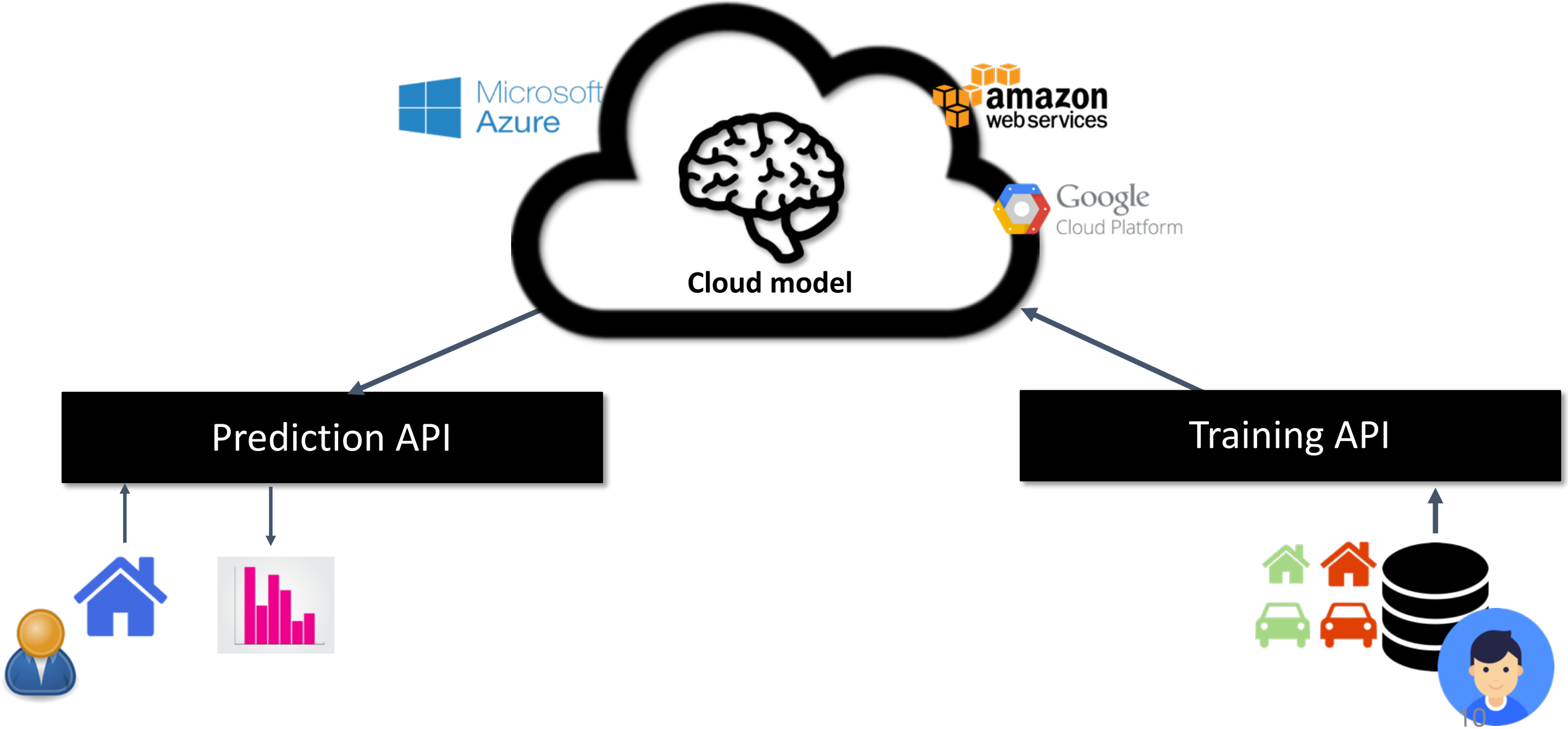
Attacks

1. Membership Inference

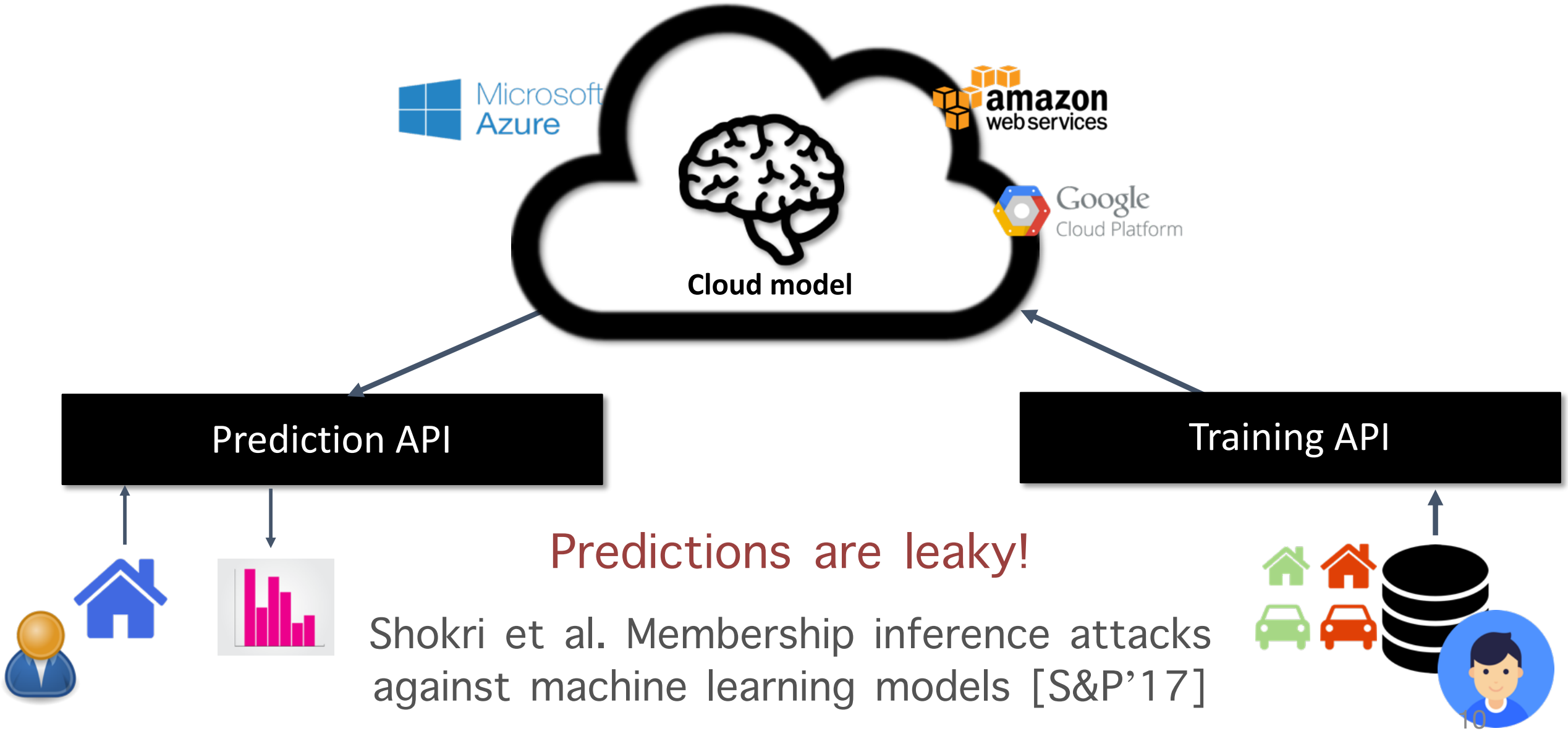
2. Property Inference in Collaborative/Federated ML

Machine Learning as a Service

Machine Learning as a Service



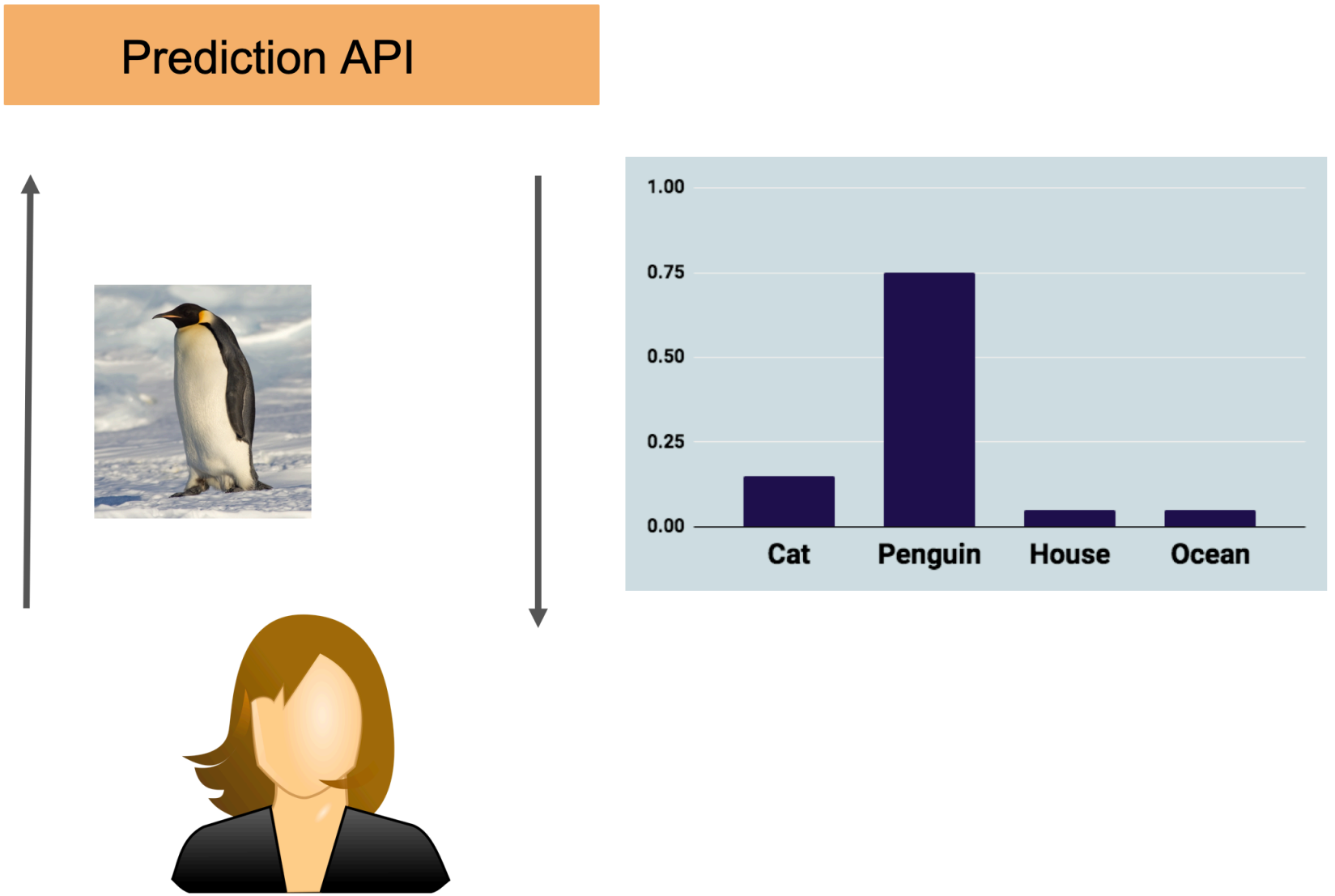
Machine Learning as a Service

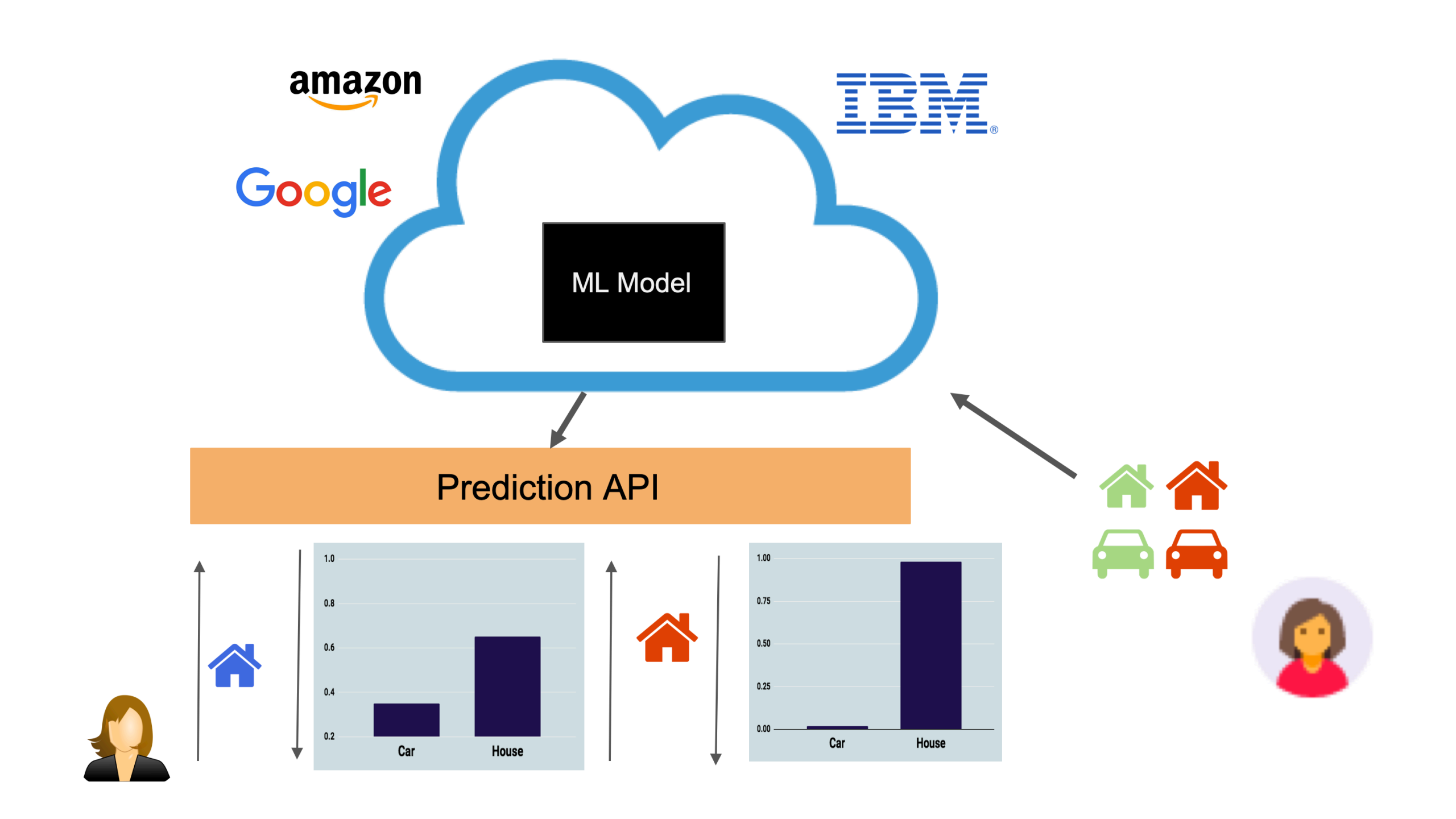


Predictions are leaky!

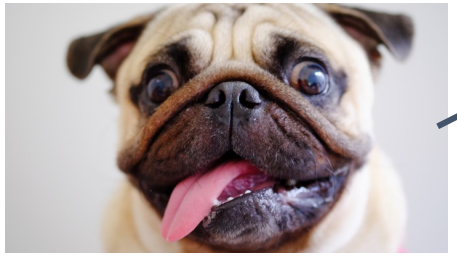
Shokri et al. Membership inference attacks against machine learning models [S&P'17]

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

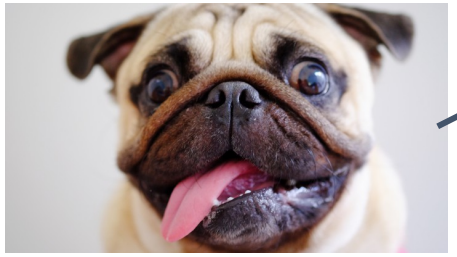
What About Generative Models?



Discriminative Model

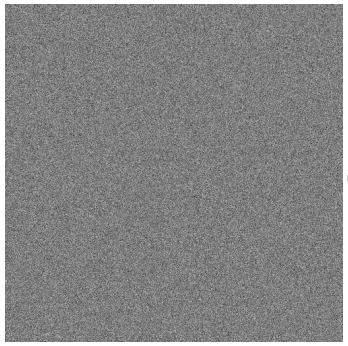
cat | dog

What About Generative Models?



Discriminative Model

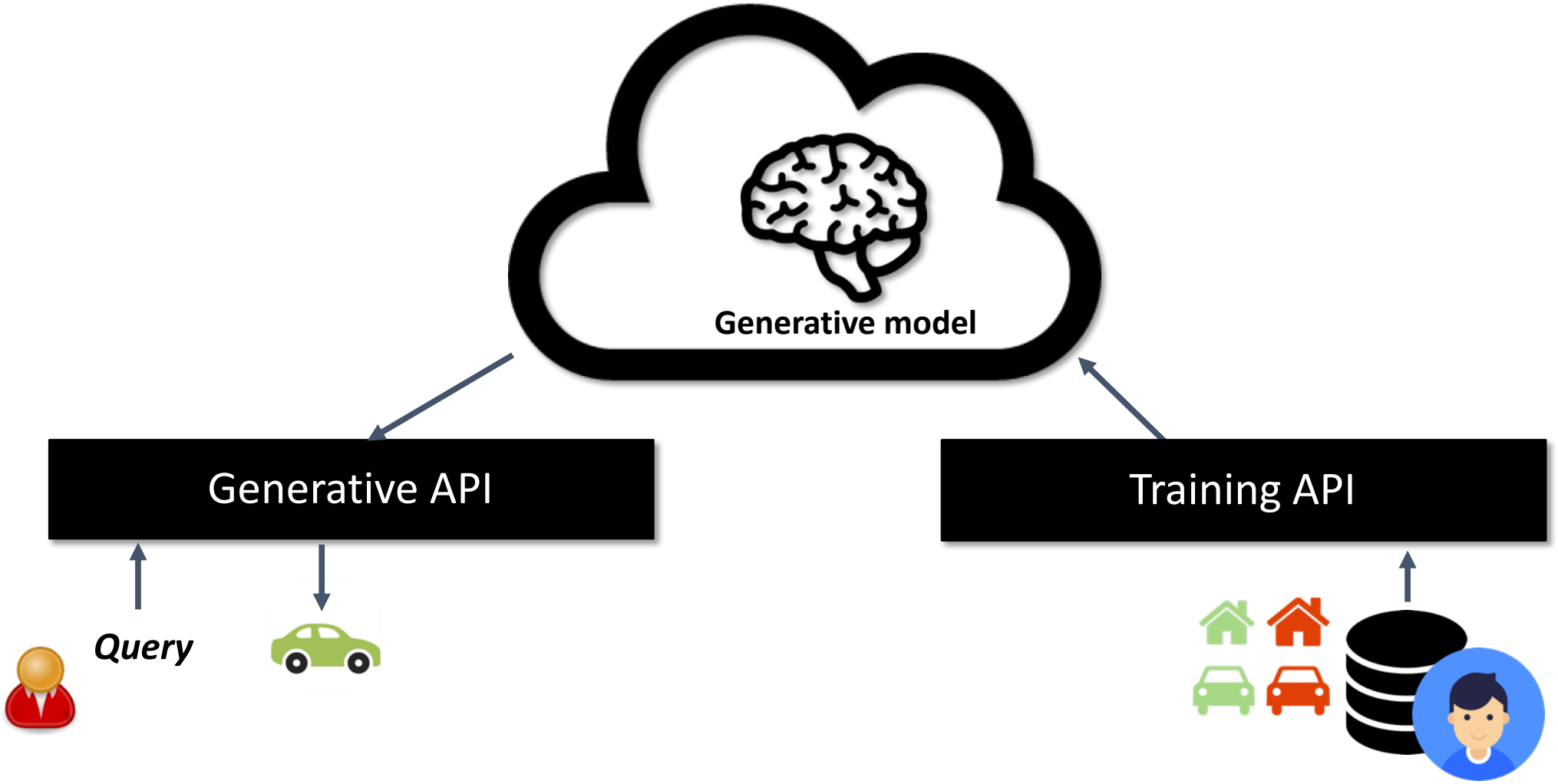
cat | dog



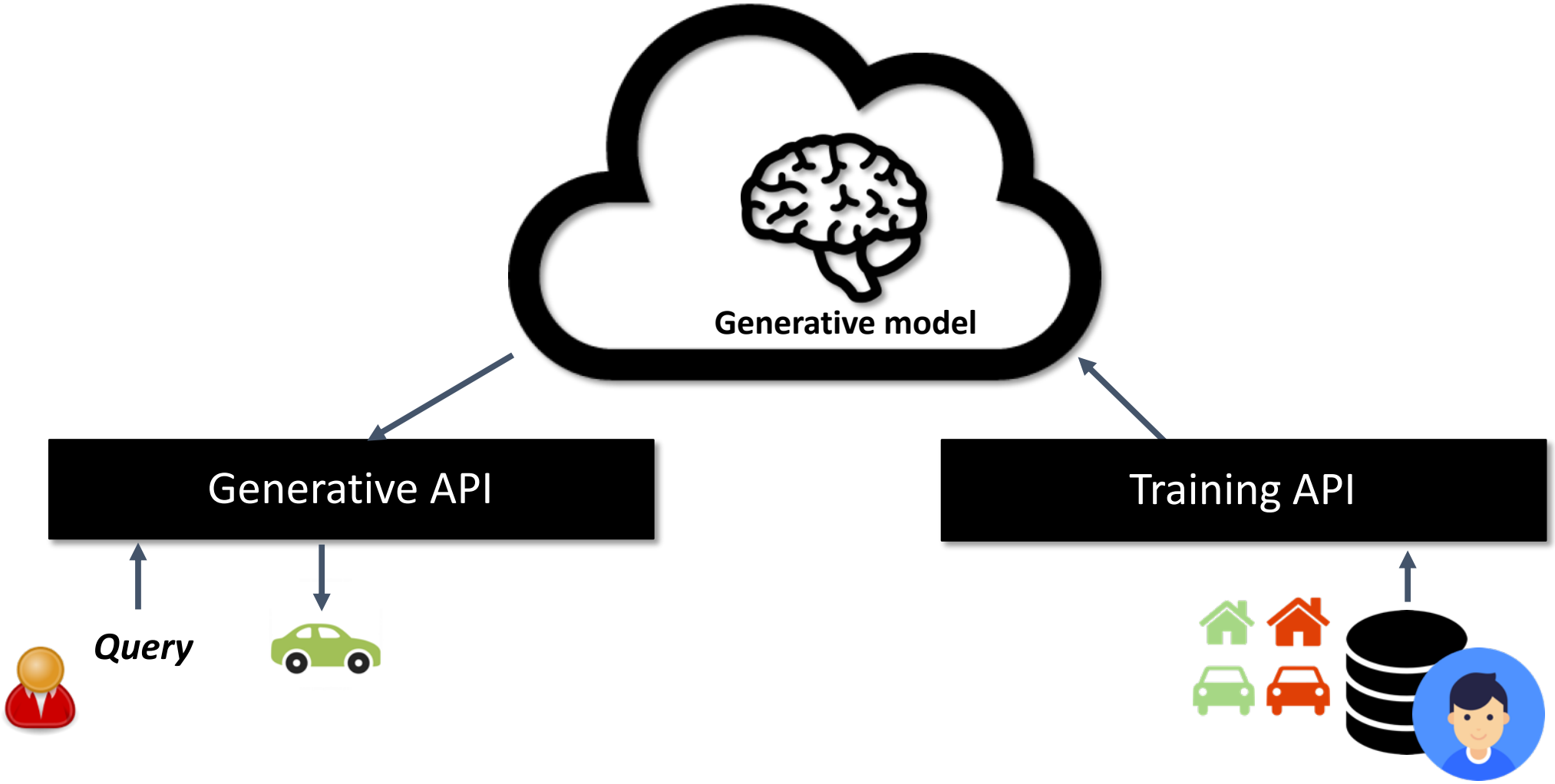
Generative Model

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?

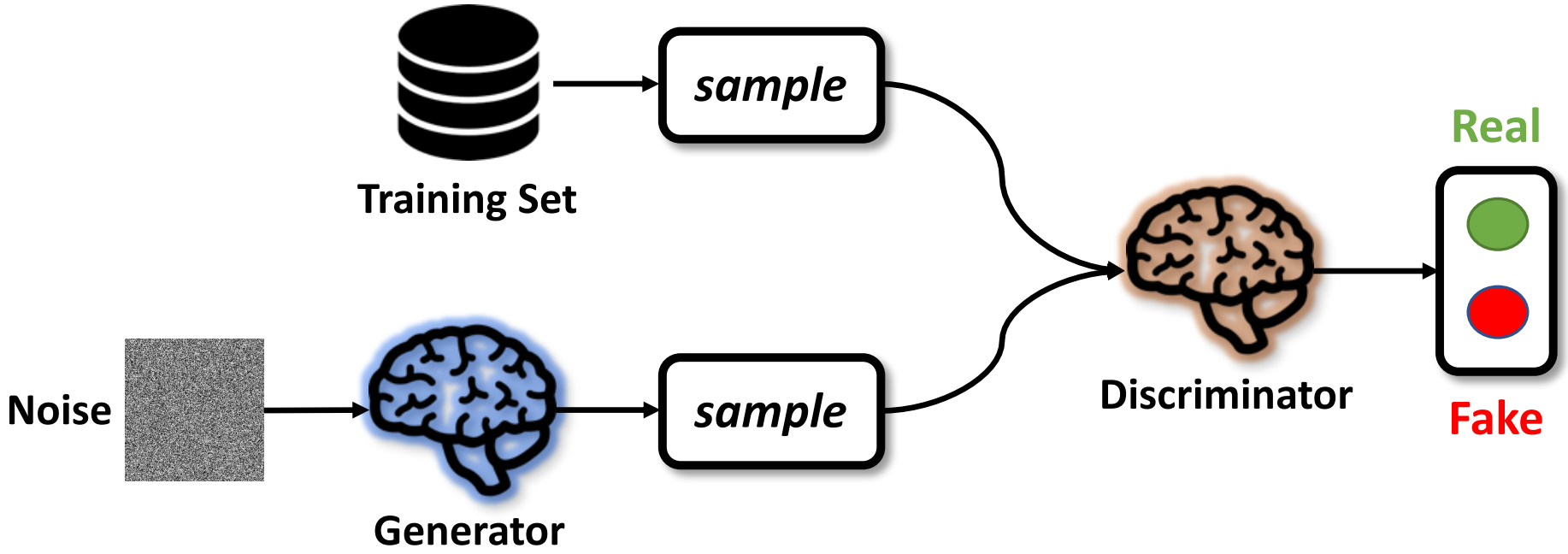
Use generative models!

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

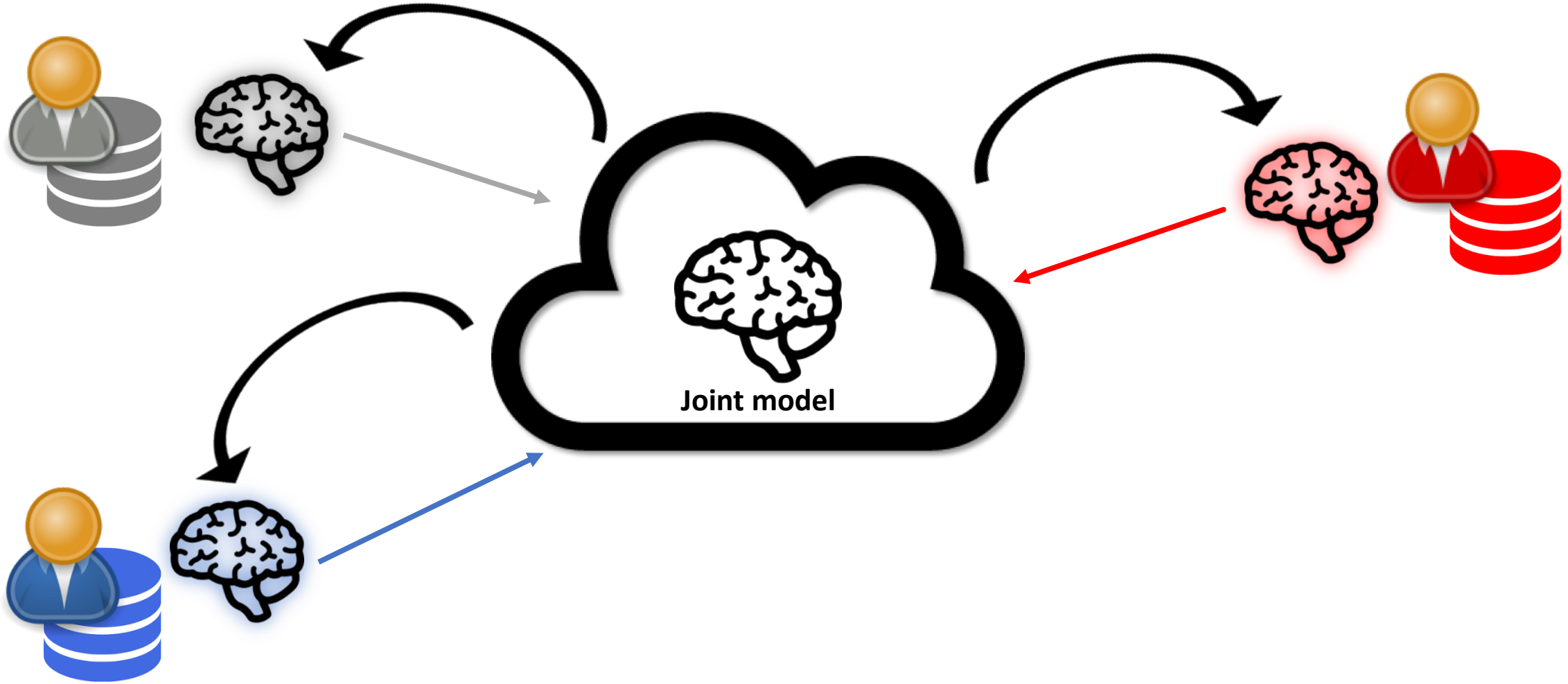


Attacks

1. Membership Inference

2. Property Inference in Collaborative/Federated ML

Collaborative/Federated Learning



Collaborative

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 $\nabla L(b; \theta)$

Federated

Algorithm 2 Federated learning with model averaging

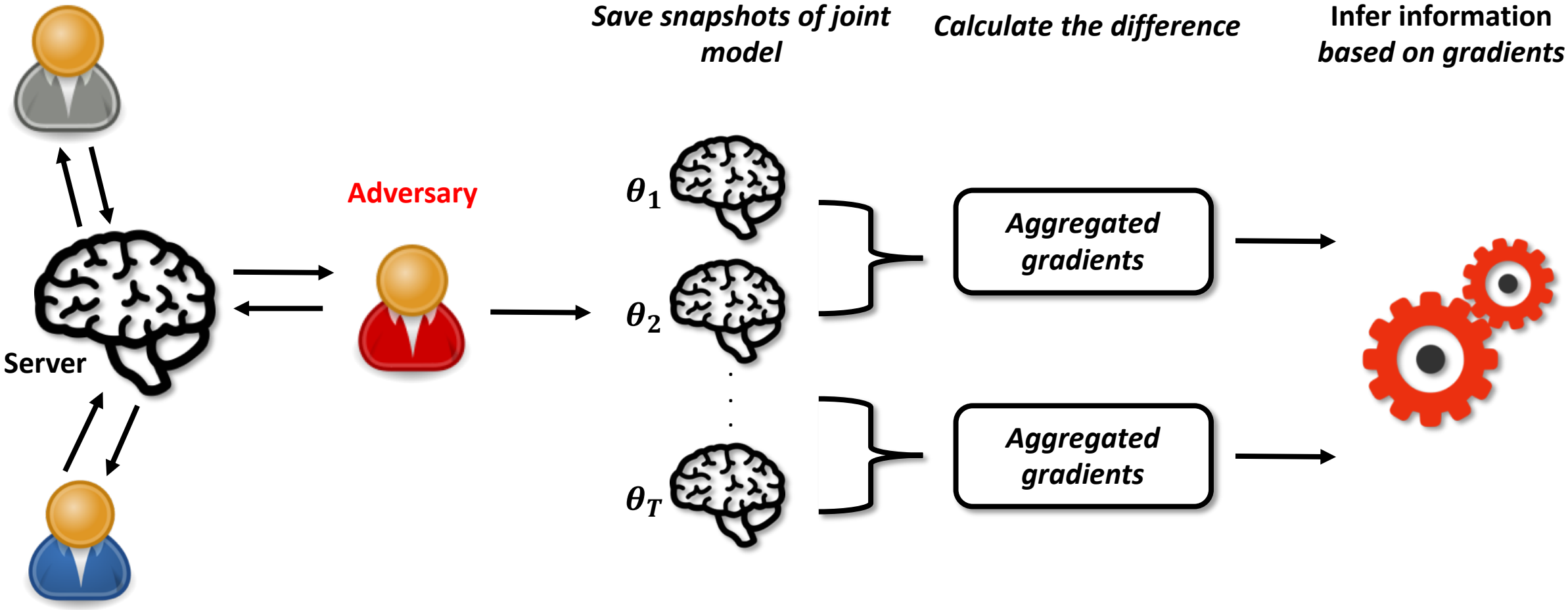
Server executes:

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

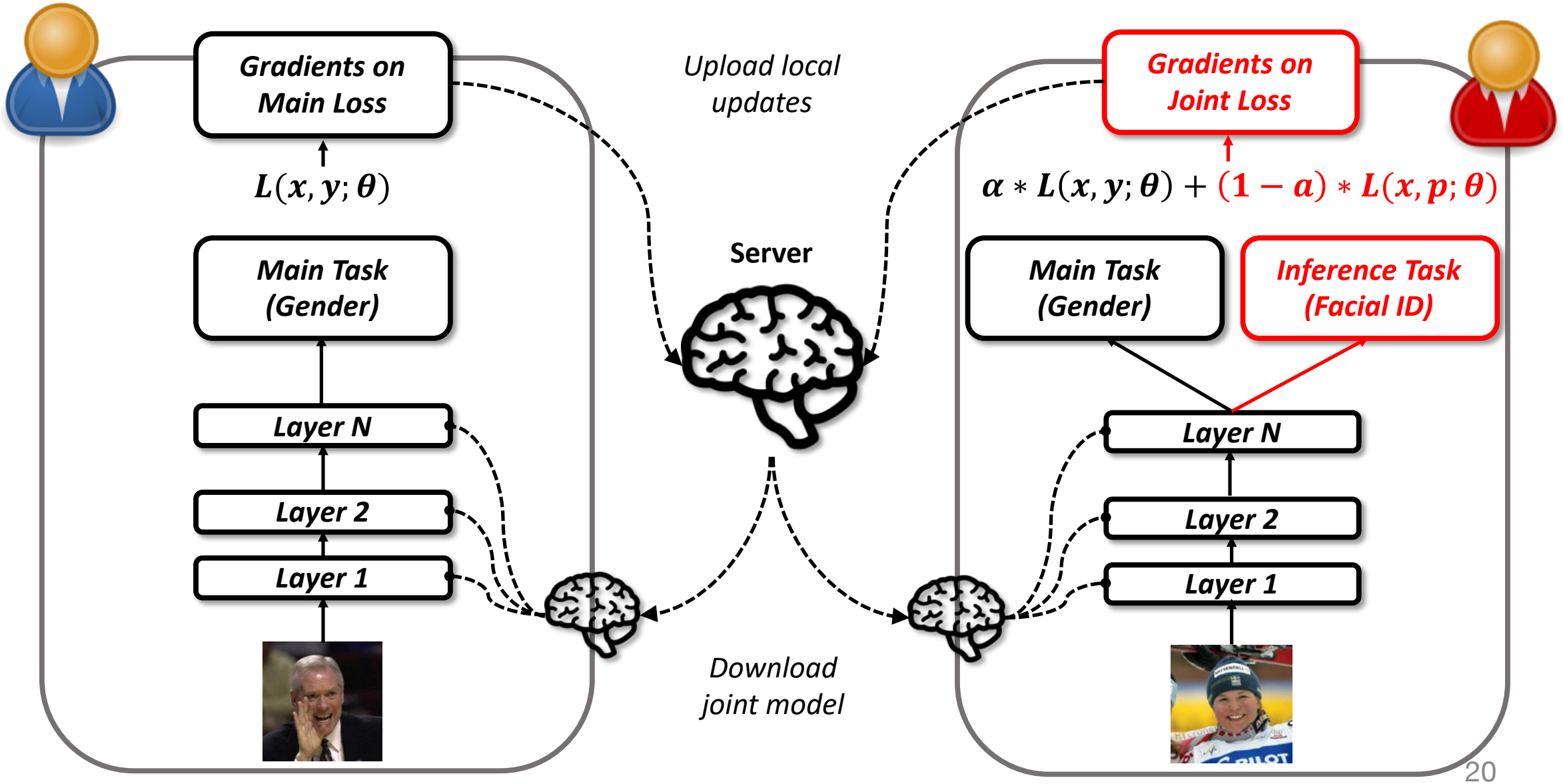
ClientUpdate(θ):

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

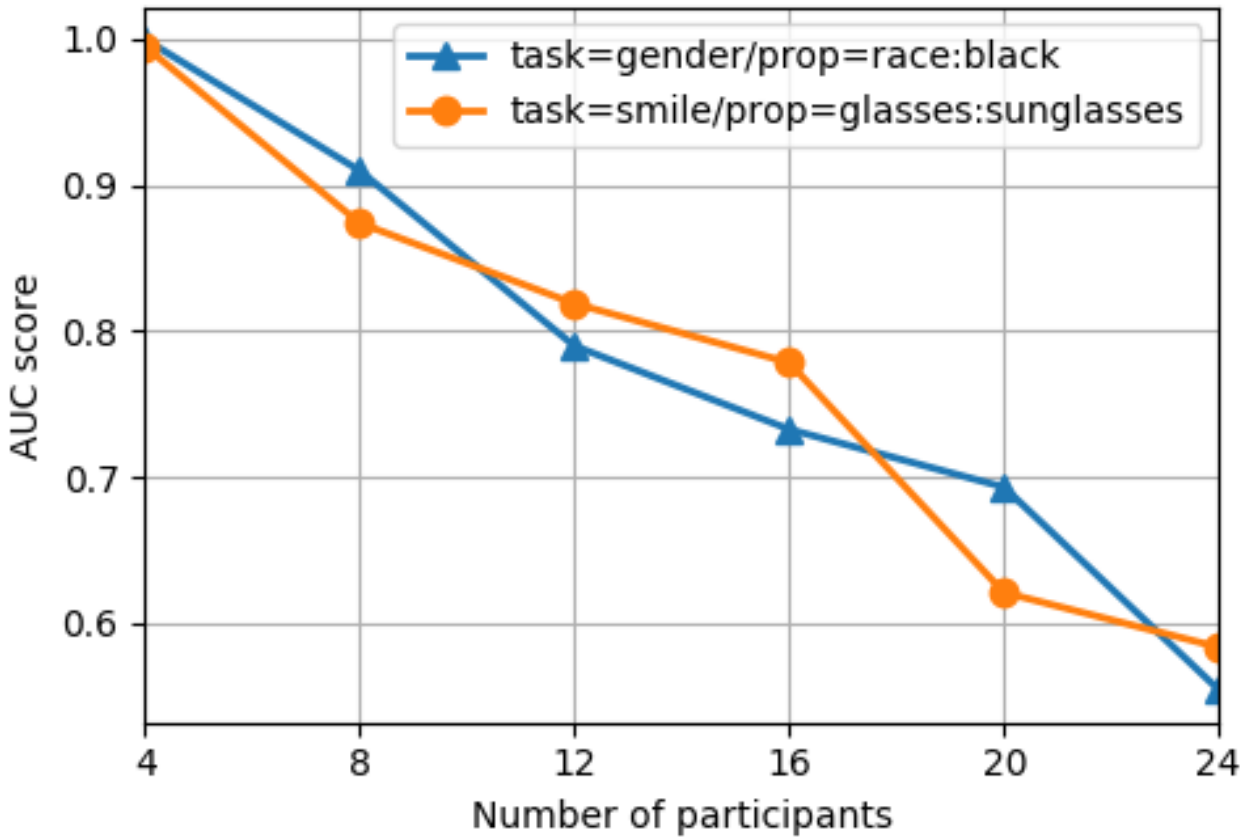


Dataset	Type	Main Task	Inference Task
LFW	Images	Gender/Smile/Age Eyewear/Race/Hair	Race/Eyewear
FaceScrub	Images	Gender	Identity
PIPA	Images	Age	Gender
FourSquare	Locations	Gender	Membership
Yelp-health	Text	Review Score	Membership Doctor specialty
Yelp-author	Text	Review Score	Author
CSI	Text	Sentiment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

Property Inference on LFW

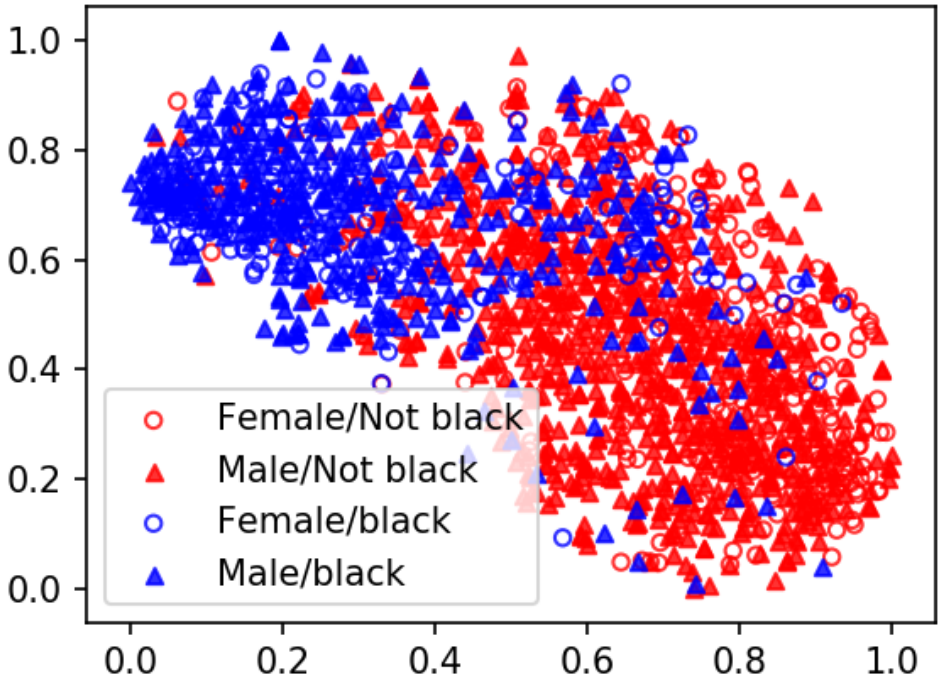
Main Task	Inference Task	Correlation	AUC score
Gender	Sunglasses	-0.025	1.0
Smile	Asian	0.047	0.93
Age	Black	-0.084	1.0
Race	Sunglasses	0.026	1.0
Eyewear	Asian	-0.119	0.91
Hair	Sunglasses	-0.013	1.0

Two-Party



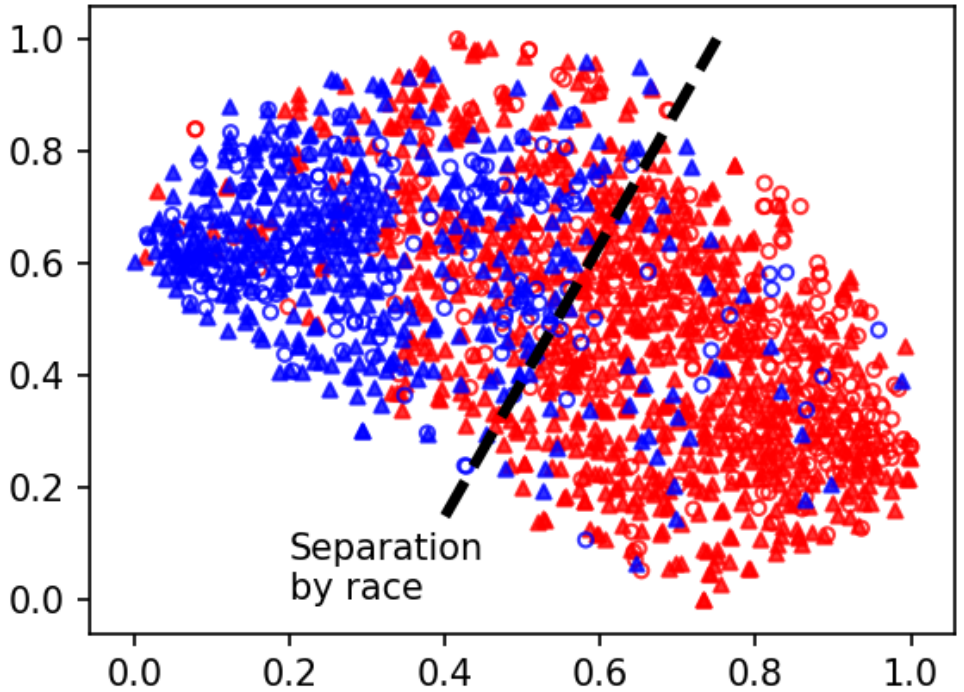
Multi-Party

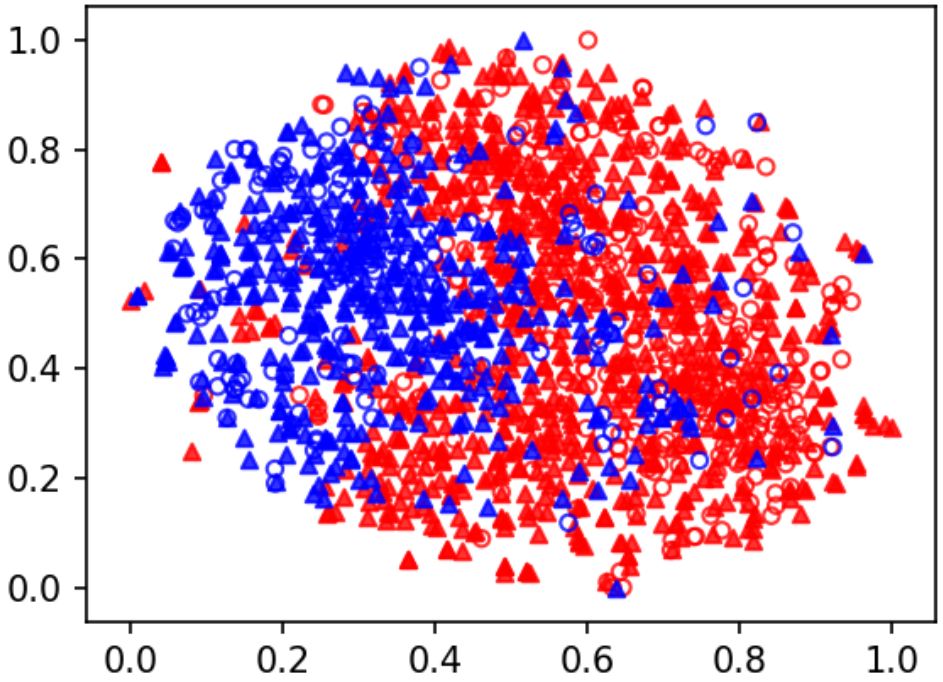
Feature t-SNE projection



pool1

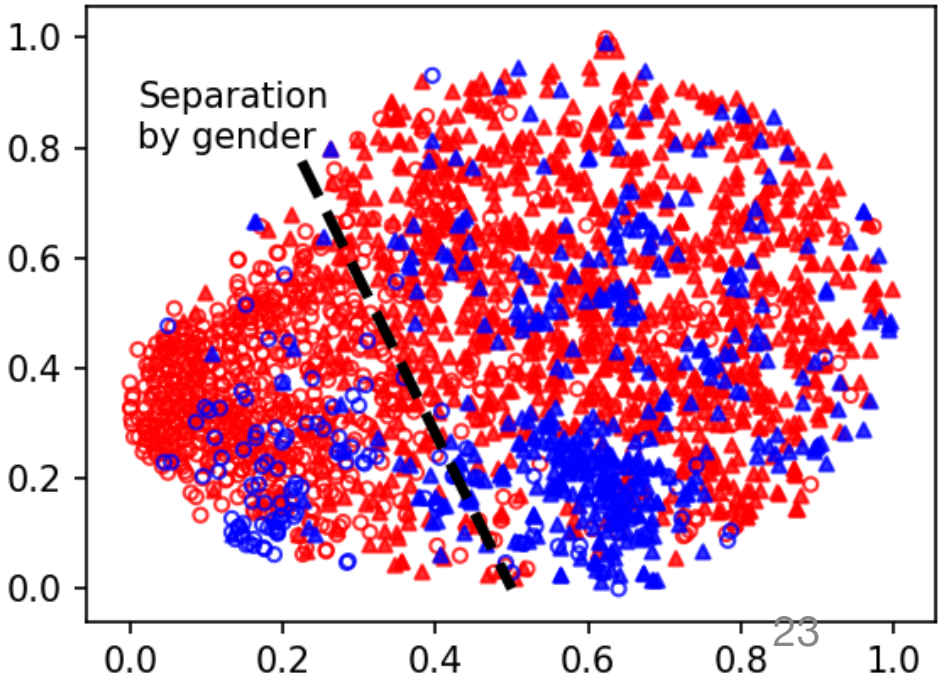
pool2



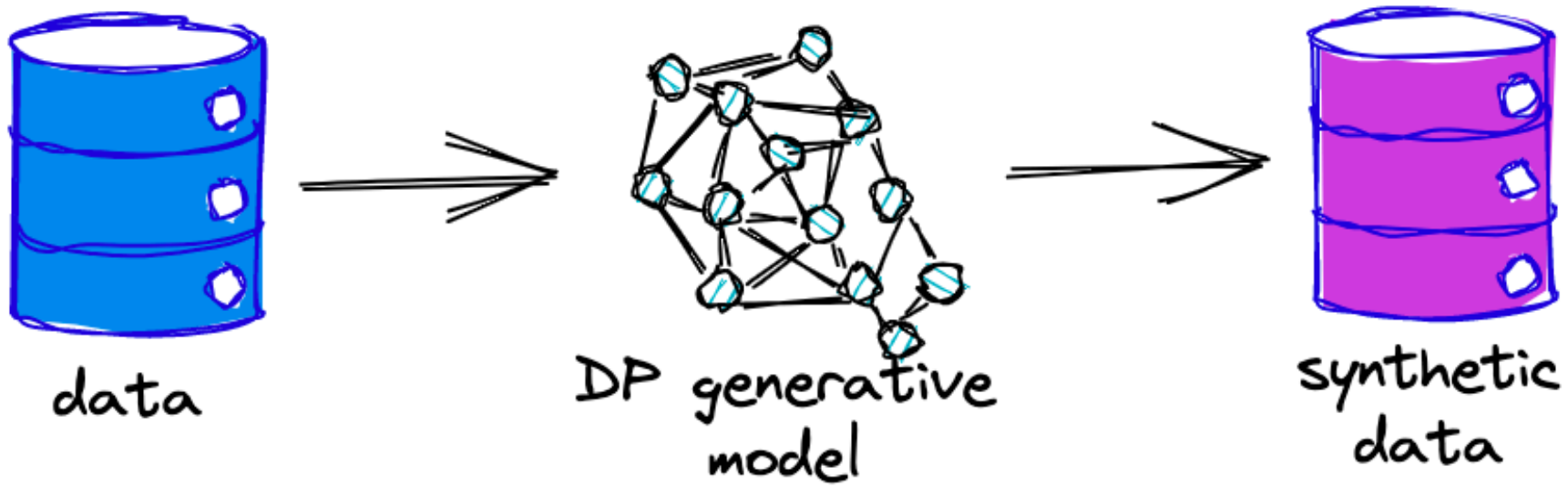


pool3

fc

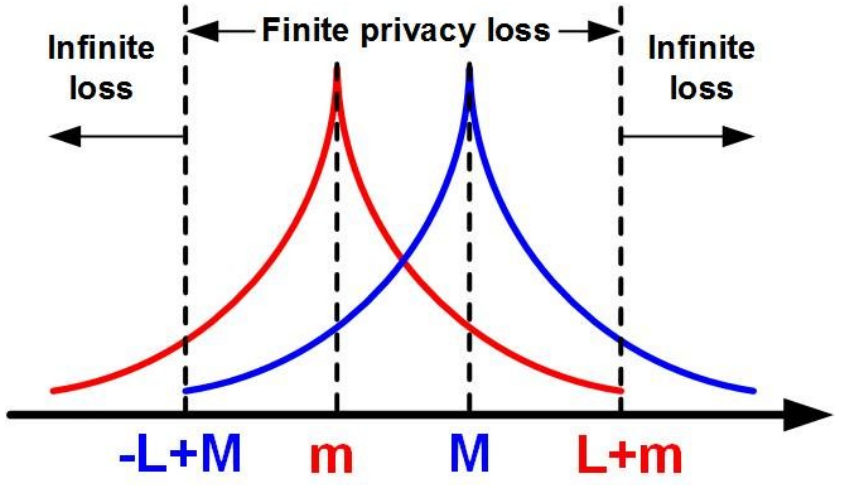


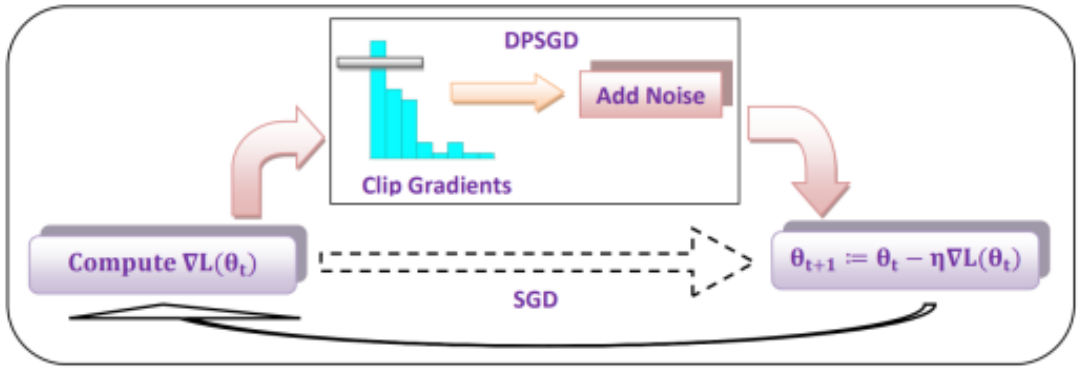
Differentially Private (DP) Synthetic Data

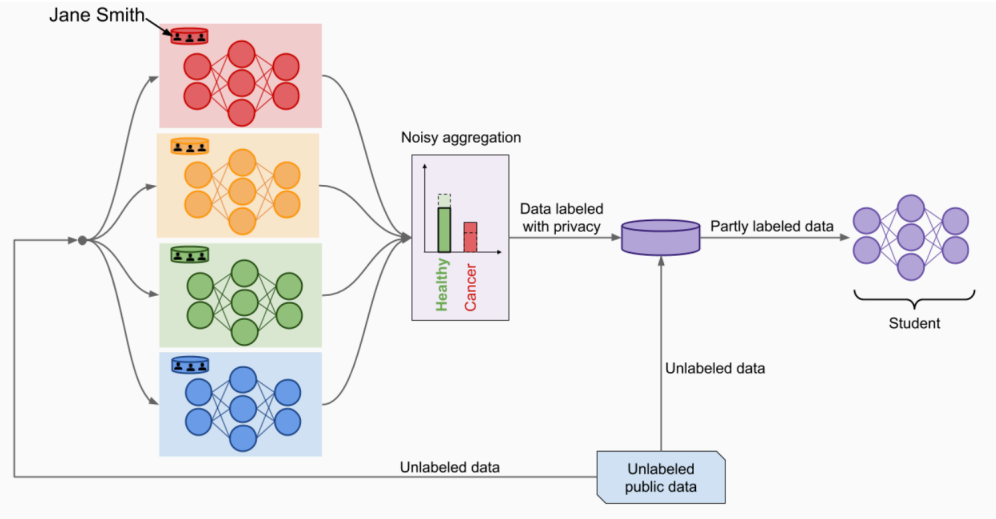


DP mechanisms:

- Laplace
- DP-SGD
- PATE



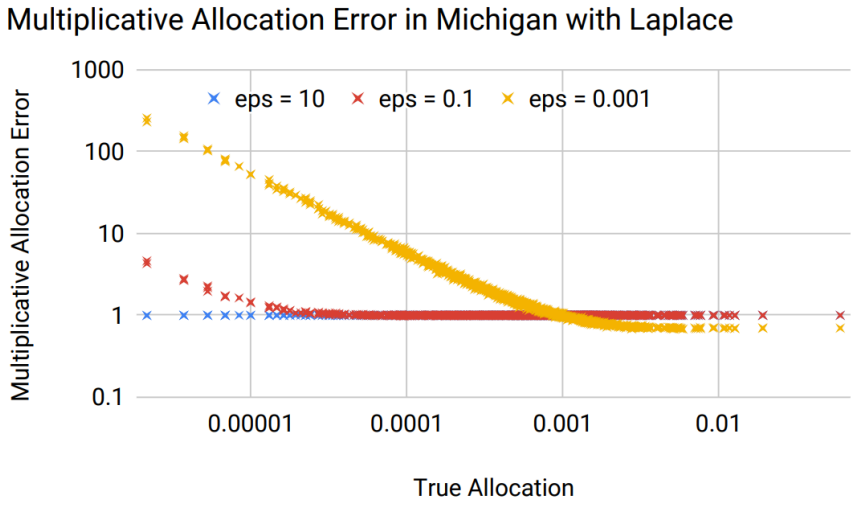


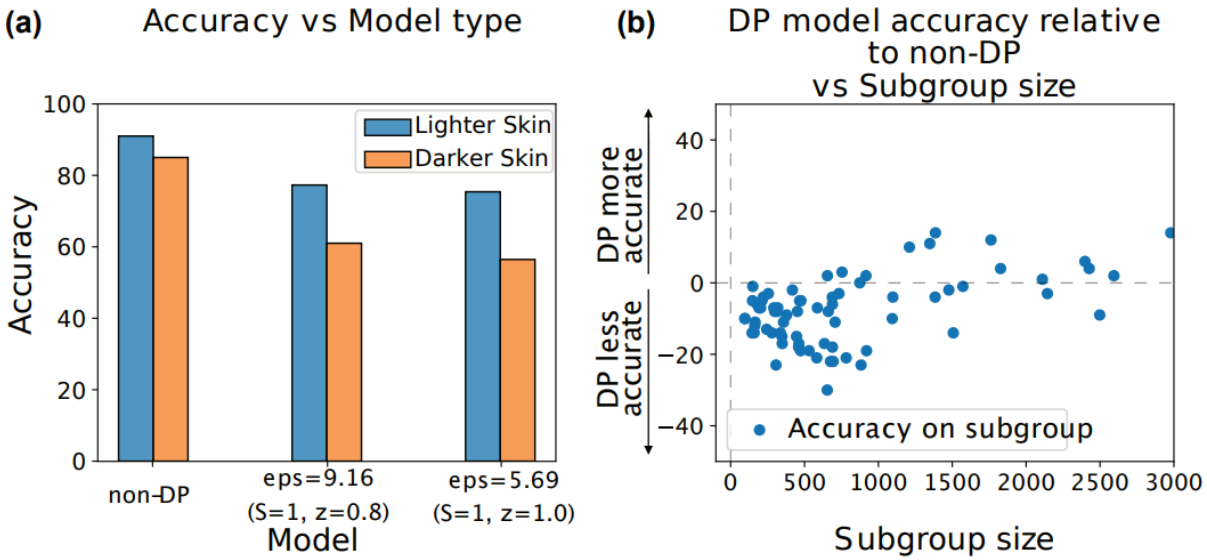


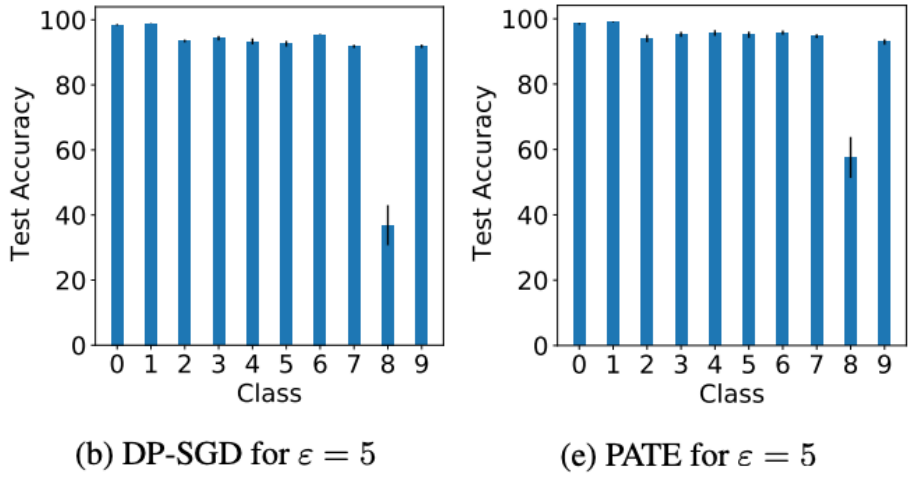
Motivation

DP has a disparate effect[1] on:

- Statistics[2]
- Deep learning classifiers[3,4]







(b) DP-SGD for $\varepsilon = 5$ (e) PATE for $\varepsilon = 5$

[1]Fioretto et al., Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey
[2]Kuppam et al., Fair Decision Making using Privacy-Protected Data
[3]Bagdasaryan et al., Differential privacy has disparate impact on model accuracy
[4]Uniyal et al., DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?

Goal

Empirically evaluate the disparate effect DP causes on generative models:

- class/subgroup size
- class/subgroup accuracy

Settings

We consider three DP generative models:

1. PrivBayes[1] (Laplace) 2. DP-WGAN[2] (DP-SGD) 3. PATE-GAN[3] (PATE)

Four data settings:

S1: Binary class size, precision, and recall

S2: Multi-class size, precision, and recall

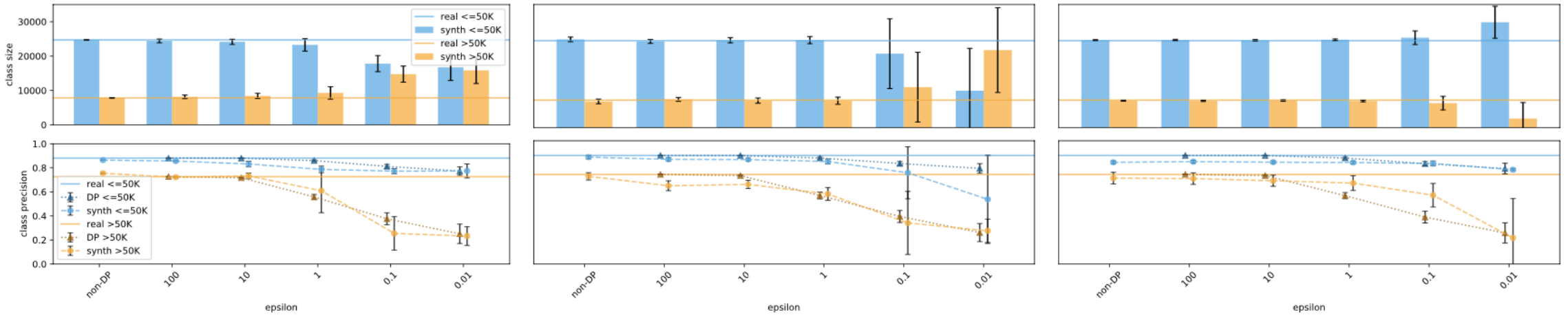
S3: Single-attribute subgroup size and accuracy

S4: Multi-attribute subgroup size and accuracy

Multiple privacy budget and subgroup imbalance levels.

[1]Zhang et al., PrivBayes: Private Data Release via Bayesian Networks
[2]Alzantot et al., Differential Privacy Synthetic Data Generation using WGANs
[3]Jordon et al., PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees

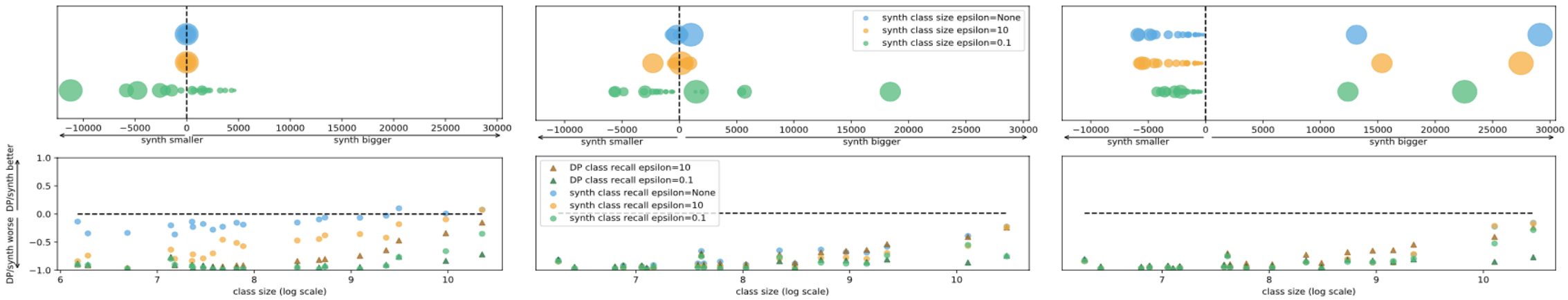
S1: Binary Class Size, Precision, and Recall



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 1: Synthetic data class size (top), real, DP, and synthetic classifiers precision (bottom) for different levels of ϵ, ***Adult***.

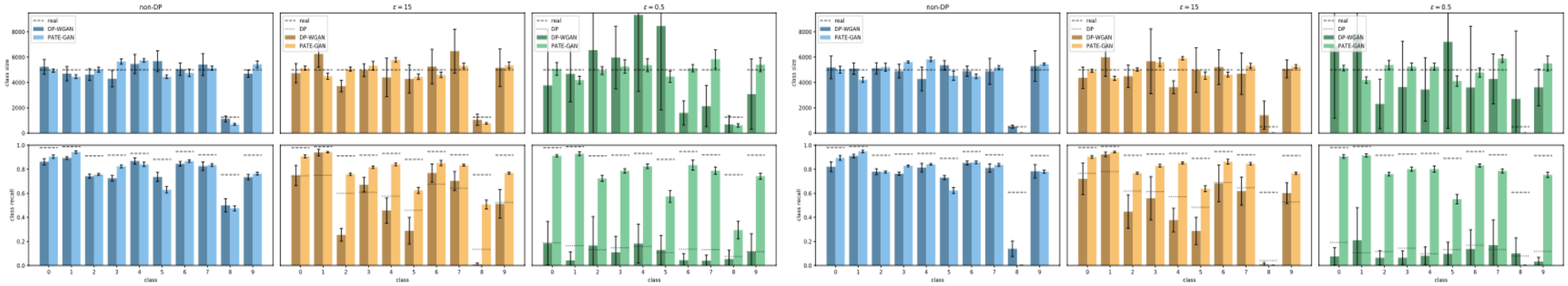
S2: Multi-Class Size, Precision, and Recall 1



(a) PrivBayes (b) DP-WGAN (c) PATE-GAN

Figure 3: Synthetic data class (multi-class) size relative to real (top) (the size of the bubbles denotes the relative size in the real data), DP classifiers and synthetic classifiers recall relative to real (bottom) for different levels of ϵ, ***Purchases***.

S2: Multi-Class Size, Precision, and Recall 2



(a) DP-WGAN and PATE-GAN with imbalance 0.25

(b) DP-WGAN and PATE-GAN with imbalance 0.1

Figure 4: Synthetic data class (multi-class) size (top), real, DP and synthetic classifiers recall (bottom) for different imbalance of class "8" and levels of ϵ, ***MNIST***.

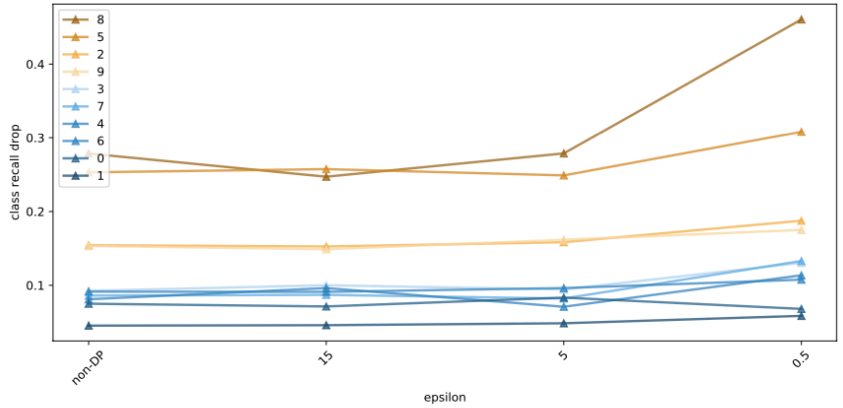


Figure 9: PATE-GAN synthetic classifier class recall drop vs ϵ, ***MNIST*** with 0.25 imbalance of class "8."

S3: Single-Attribute Subgroup Size and Accuracy

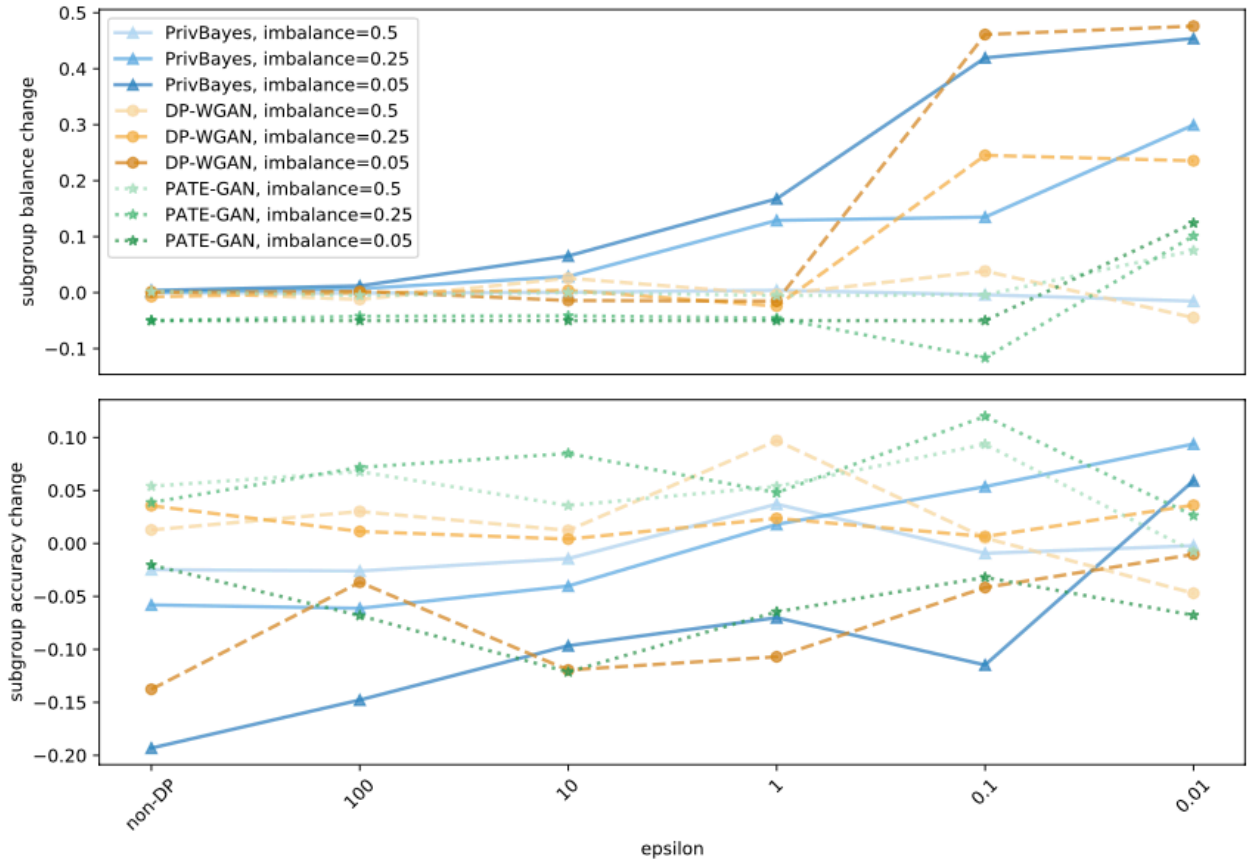
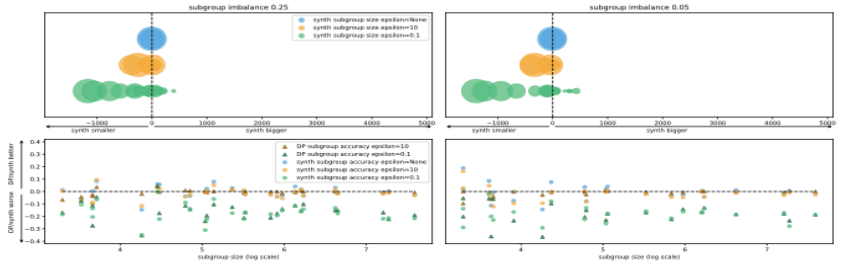
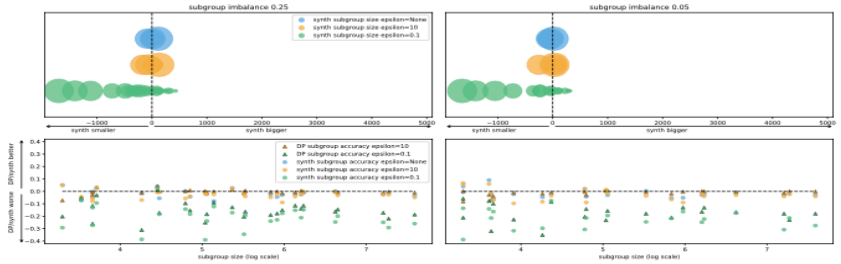


Figure 5: Minority single-attribute (sex) subgroup imbalance level change (top), minority subgroup drop in accuracy vs. majority (bottom) for different subgroup imbalance and ϵ levels, ***Texas***.

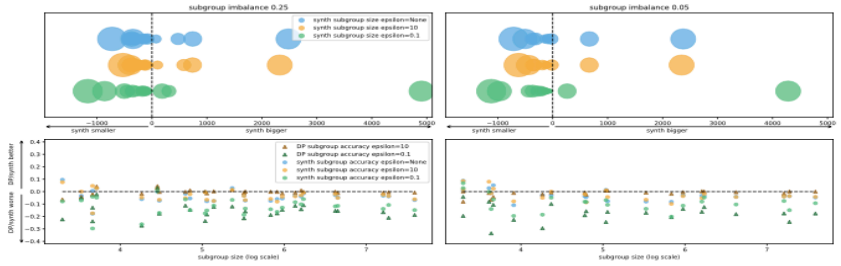
S4: Multi-Attribute Subgroup Size and Accuracy



(a) PrivBayes

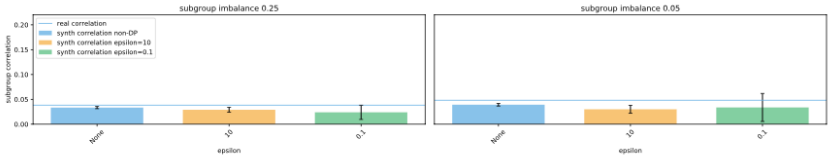


(b) DP-WGAN

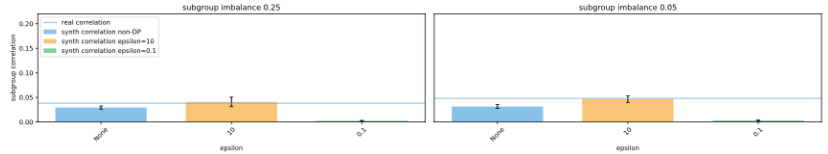


(c) PATE-GAN

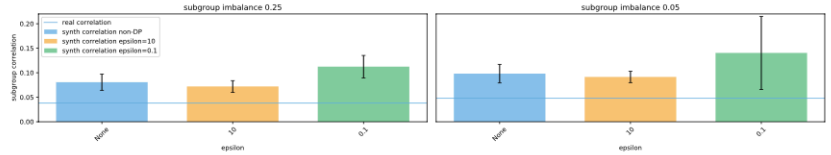
Figure 13: Size of synthetic data multi-attribute subgroup (intersection of age, sex, and race) relative to real (top) and accuracy of DP and synthetic classifiers relative to real classifier accuracy (bottom) for different single-attribute (sex) subgroup imbalance and ϵ levels, ***Texas***.



(b) PrivBayes, Texas



(d) DP-WGAN, Texas



(f) PATE-GAN, Texas

Figure 14: Mutual information between the multi-attribute subgroup and the target (income/length of stay) columns for different single-attribute subgroup imbalance (sex) and ϵ levels, ***Adult*** (left) and ***Texas*** (right).

Take-Aways

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

No, not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

No, not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ2: Does training a classier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?

Take-Aways

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

No, not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ2: Does training a classier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?

Overall yes. Smaller classes/subgroups suffer more similarly to DP classifiers. However, we do not see the rich get richer, the poor get poorer; rather, everybody gets poorer.

Take-Aways

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

No, not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ2: Does training a classier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?

Overall yes. Smaller classes/subgroups suffer more similarly to DP classifiers. However, we do not see the rich get richer, the poor get poorer; rather, everybody gets poorer.

RQ3: Do different DP mechanisms for DP synthetic data behave similarly under different privacy and data imbalance levels?

Take-Aways

RQ1: Do DP generative models generate data in similar classes and subgroups proportions to the real data?

No, not really. DP distorts the proportions, yielding Robin Hood vs Matthew effects depending on the DP generative model.

RQ2: Does training a classier on DP synthetic data lead to the same disparate impact on accuracy as training a DP classifier on the real data?

Overall yes. Smaller classes/subgroups suffer more similarly to DP classifiers. However, we do not see the rich get richer, the poor get poorer; rather, everybody gets poorer.

RQ3: Do different DP mechanisms for DP synthetic data behave similarly under different privacy and data imbalance levels?

No, different DP generative models behave differently.

Thank you!