Machine Learning & Privacy: It's Complicated

Emiliano De Cristofaro
https://emilianodc.com

Agenda

1. Training (Distributed) ML Models with Privacy

2. Private Data Release with Generative Neural Networks

3. Privacy Leakage in Collaborative/Federated ML

Agenda

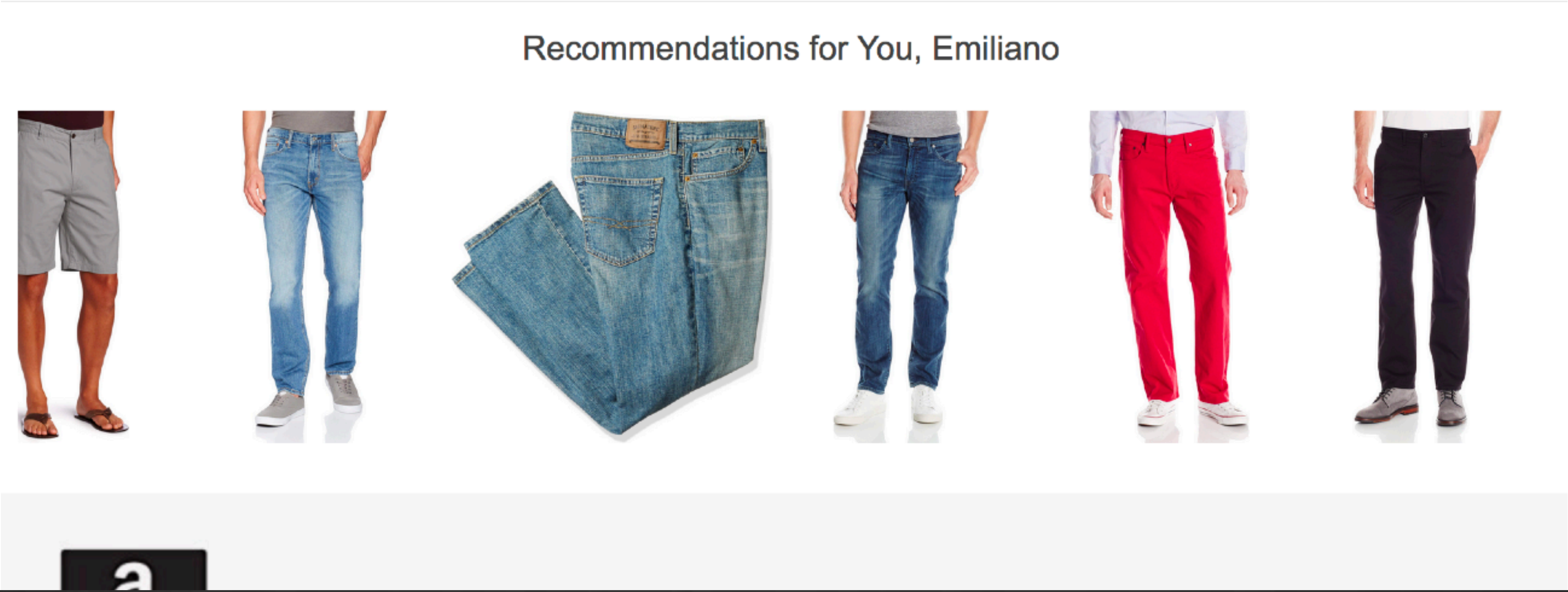
1. Training (Distributed) ML Models with Privacy

2. Private Data Release with Generative Neural Networks

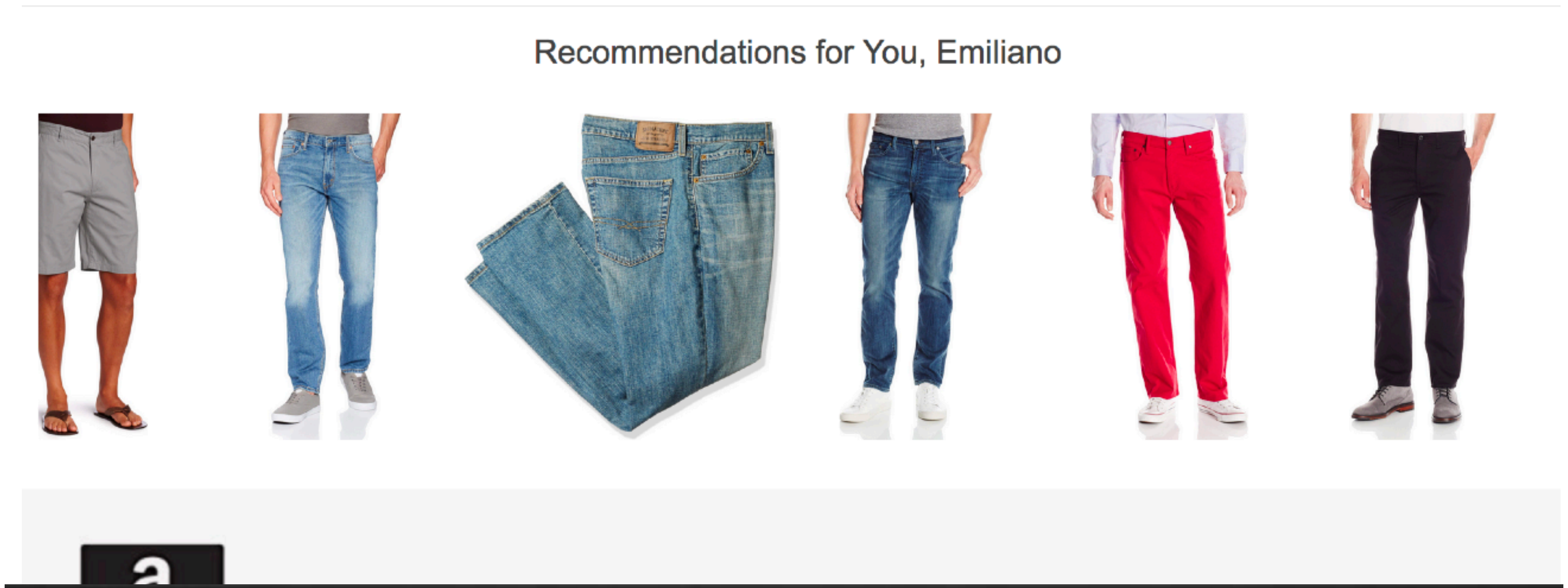
3. Privacy Leakage in Collaborative/Federated ML

Recommendations

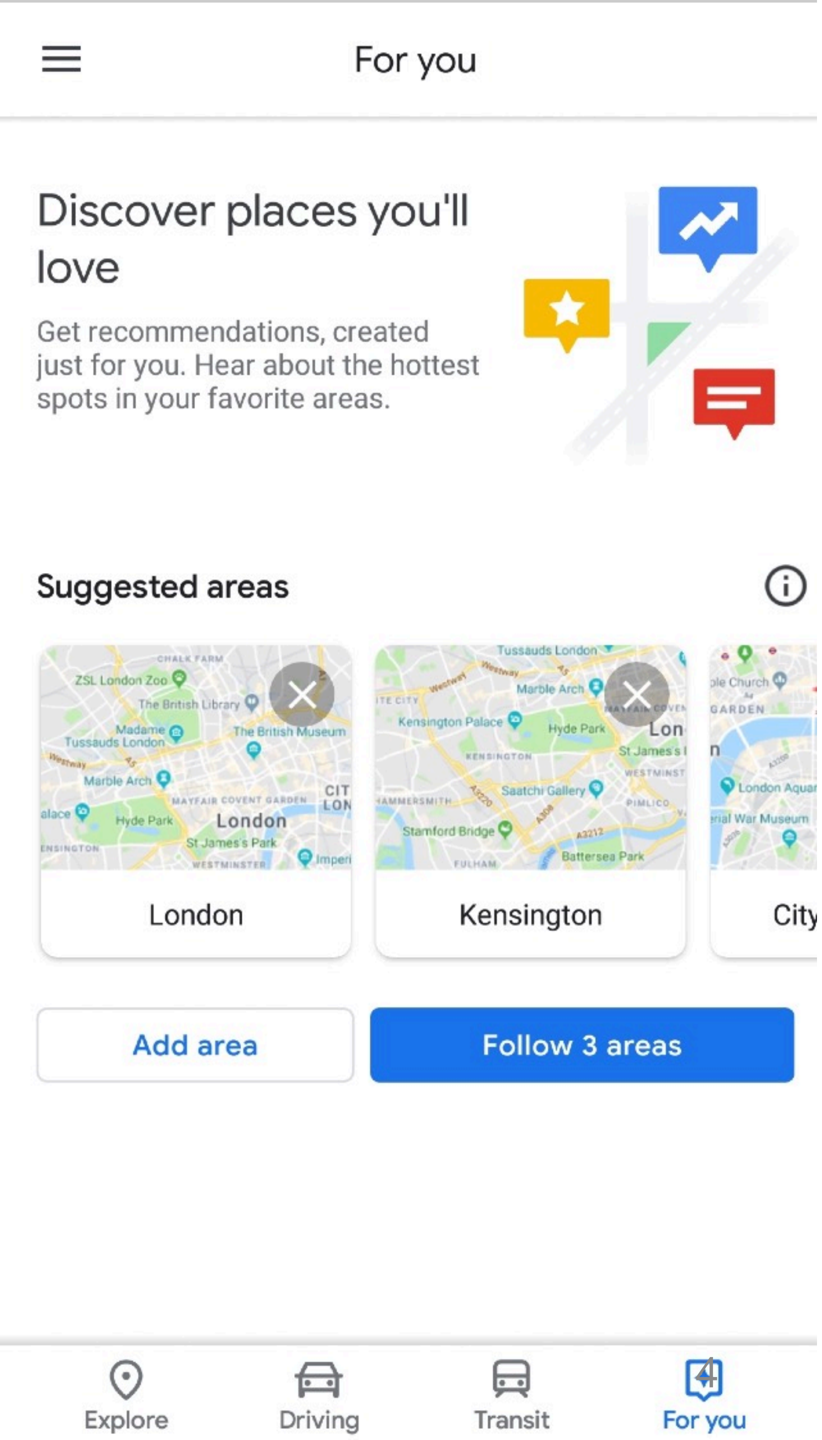
Recommendations for You, Emiliano


Recommendations for You, Emiliano


Top 10 for You
A NETFLIX ORIGINAL
ORANGE is the new BLACK
CHUCK
HOUSE of CARDS
SAFE HAVEN
DEXTER
Breaking Bad


Recommendations for You, Emiliano


Top 10 for You
ORANGE is the new BLACK
CHUCK
HOUSE of CARDS
SAFE HAVEN
DEXTER


For you
Discover places you'll love
Get recommendations, created just for you. Hear about the hottest spots in your favorite areas.
Suggested areas
London
Kensington
Add area
Follow 3 areas
Explore
Driving
Transit
For you

BBC iPlayer

The BBC keeps a few hundred free programs on iPlayer

- No tracking, no ads (taxpayer funded)
- No account (until recently)

The BBC keeps a few hundred free programs on iPlayer

- No tracking, no ads (taxpayer funded)
- No account (until recently)

Still… they want to give recommendations & gather statistics

Item-KNN based Recommendations

Item-KNN based Recommendations

Predict favorite items for users based on their own ratings and those of "similar" users

Item-KNN based Recommendations

Predict favorite items for users based on their own ratings and those of “similar” users

For iPlayer, consider binary ratings (viewed/not viewed)

Item-KNN based Recommendations

Predict favorite items for users based on their own ratings and those of “similar” users

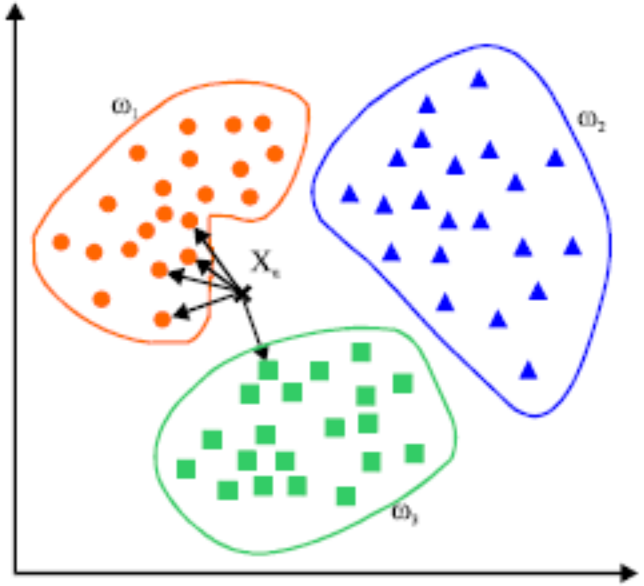
For iPlayer, consider binary ratings (viewed/not viewed)

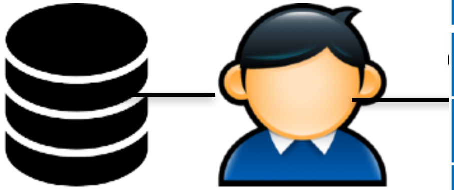
Build a co-views matrix C

C_{ab} = #views for the pair of programs (a,b)

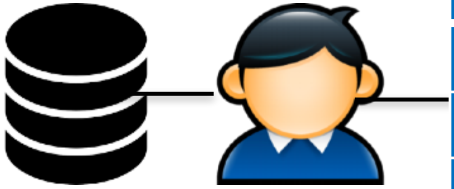
Compute a Similarity Matrix $\{Sim\}_{ab} = \frac{C_{ab}}{\sqrt{C_a \cdot C_b}}$

Identify K-Neighbors (KNN) based on Sim Matrix





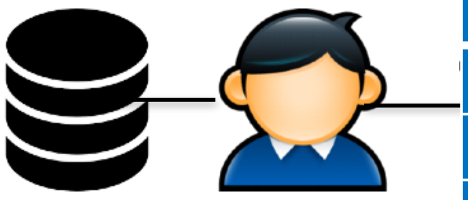
	Dr Who	Sherlock	Earth
Dr Who	1	-	-
Sherlock	1	1	-
Earth	0	0	0



	Dr Who	Sherlock	Earth
Dr Who	1	-	-
Sherlock	1	1	-
Earth	1	1	1

	Dr Who	Sherlock	Earth
Dr Who	195	-	-
Sherlock	155	180	-
Earth	80	99	123

⋮



	Dr Who	Sherlock	Earth
Dr Who	1	-	-
Sherlock	0	0	-
Earth	0	0	0

	Dr Who	Sherlock	Earth
Dr Who	195	-	-
Sherlock	155	180	-
Earth	80	99	123

	Dr Who	Sherlock	Earth
Dr Who	195	-	-
Sherlock	155	180	-
Earth	80	99	123

Can we build this in a privacy-preserving way?

	Dr Who	Sherlock	Earth
Dr Who	195	-	-
Sherlock	155	180	-
Earth	80	99	123

Can we build this in a privacy-preserving way?

Privacy := learn aggregate counts, e.g., 155 users have watched Dr Who and Sherlock, but not who has watched what

Private Data Aggregation (PDA)

Private Data Aggregation (PDA)

Use additively homomorphic encryption

$Enc_{PK}(x)*Enc_{PK}(y) = Enc_{PK}(x+y)$

Private Data Aggregation (PDA)

Use additively homomorphic encryption

$Enc_{PK}(x)*Enc_{PK}(y) = Enc_{PK}(x+y)$

Generate keys adding up to 0

User $U_1, U_2, \cdots, U_N \longrightarrow k_1 + k_2 + \cdots + k_N = 0$

$Enc_{ki}(x_i) = x_i+k_i \bmod 2^{32}$

$\Pi_{i=1,..,N}\, Enc_i(x_i) = \Sigma_{i=1,..,N}\, (x_i+k_i) = \Sigma_{i=1,..,N}\, x_i$

User U_i ($i \in [1, N]$) | **Server**

$$x_i \in_r G,\ y_i = g^{x_i} \bmod q$$

$\xrightarrow{y_i}$

$$k_{ij} = \sum_{j \neq i} H\left(y_j^{x_i} \parallel \ell \parallel s\right) \cdot (-1)^{i>j} \bmod 2^{32}$$

$\xleftarrow{\{y_j\}_{j \in [1,N]}}$

$$b_{i\ell} = X_{i\ell} + k_{i\ell} \bmod 2^{32}$$

$\xrightarrow{\{b_{i\ell}\}_{\ell \in [1,L]}}$

Fault recovery (if needed)

$\xleftarrow{U^{on}}$

$$k'_{ij} = \sum_{\substack{j \neq i \\ j \notin U^{on}}} H\left(y_j^{x_i} \parallel \ell \parallel s\right) \cdot (-1)^{i>j} \bmod 2^{32}$$

$\xrightarrow{\{k'_{i\ell}\}_{\ell \in [1,L]}}$

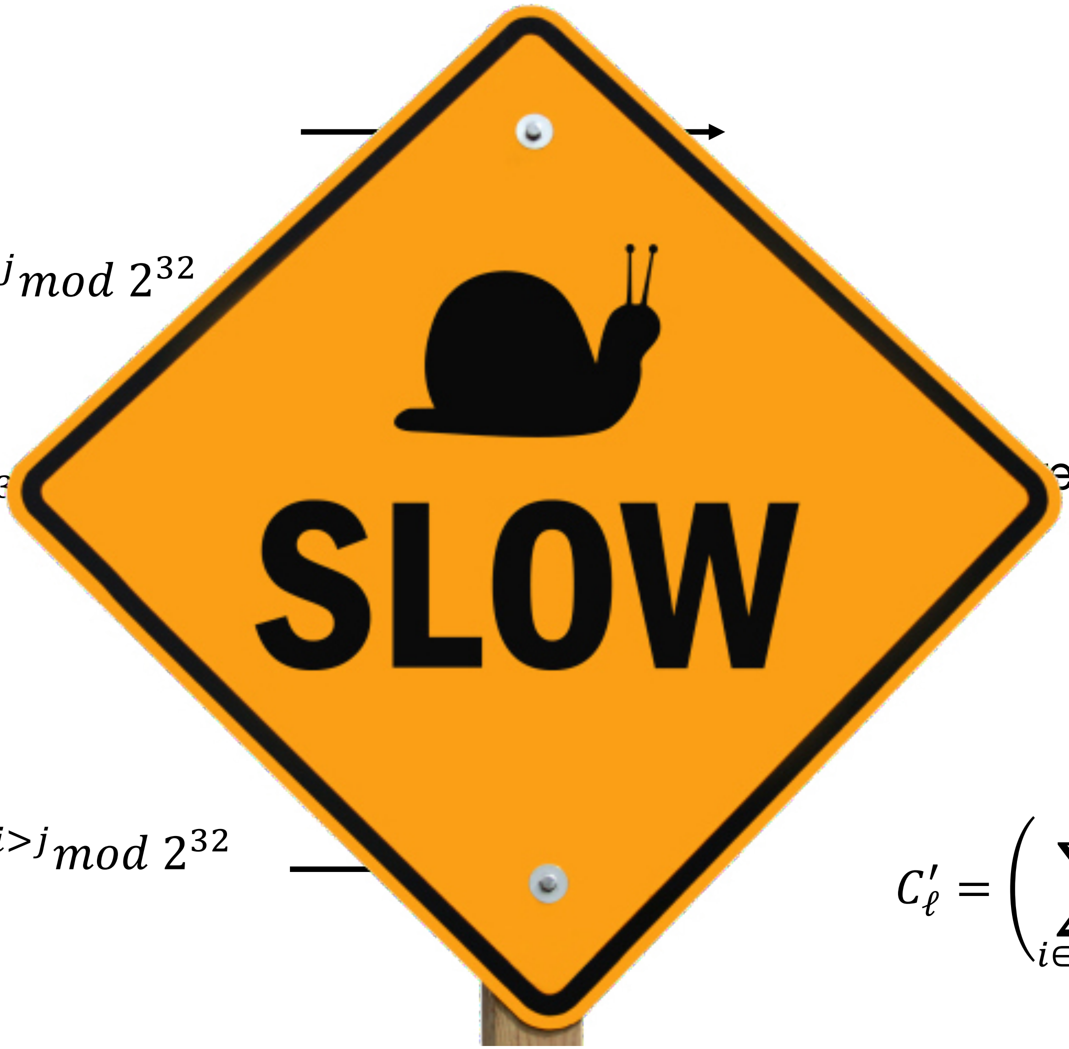
$$C'_\ell = \left(\sum_{i \in U^{on}} b_{i\ell} - \sum_{i \in U^{on}} k'_{i\ell}\right) \bmod 2^{32}$$

User U_i ($i \in [1, N]$) **Server**

$x_i \in_r G,\ y_i = g^{x_i}\ mod\ q$

$$k_{ij} = \sum_{j \neq i} H\left(y_j^{x_i} \parallel \ell \parallel s\right) \cdot (-1)^{i>j} mod\ 2^{32}$$

$b_{i\ell} = X_{i\ell} + k_{i\ell}\ mod\ 2^3$



ecovery (if needed)

$$k'_{ij} = \sum_{\substack{j \neq i \\ j \notin U^{on}}} H\left(y_j^{x_i} \parallel \ell \parallel s\right) \cdot (-1)^{i>j} mod\ 2^{32}$$

$$C'_\ell = \left(\sum_{i \in U^{on}} b_{i\ell} - \sum_{i \in U^{on}} k'_{i\ell}\right) mod\ 2^{32}$$

Using PDA for Item-KNN does not scale...

For N users and M programs: $O(N \cdot M^2)$ cryptographic operations and $O(M^2)$ ciphertexts

Using PDA for Item-KNN does not scale...

For N users and M programs: $O(N \cdot M^2)$ cryptographic operations and $O(M^2)$ ciphertexts

Using PDA for Item-KNN does not scale…

For N users and M programs: $O(N \cdot M^2)$ cryptographic operations and $O(M^2)$ ciphertexts

Approximate statistics may be ok for better efficiency?

Using PDA for Item-KNN does not scale...

For N users and M programs: $O(N \cdot M^2)$ cryptographic operations and $O(M^2)$ ciphertexts

Approximate statistics may be ok for better efficiency?

Use succinct data structures to compress data streams and aggregate on that

Using PDA for Item-KNN does not scale…

For N users and M programs: $O(N \cdot M^2)$ cryptographic operations and $O(M^2)$ ciphertexts

Approximate statistics may be ok for better efficiency?

Use succinct data structures to compress data streams and aggregate on that

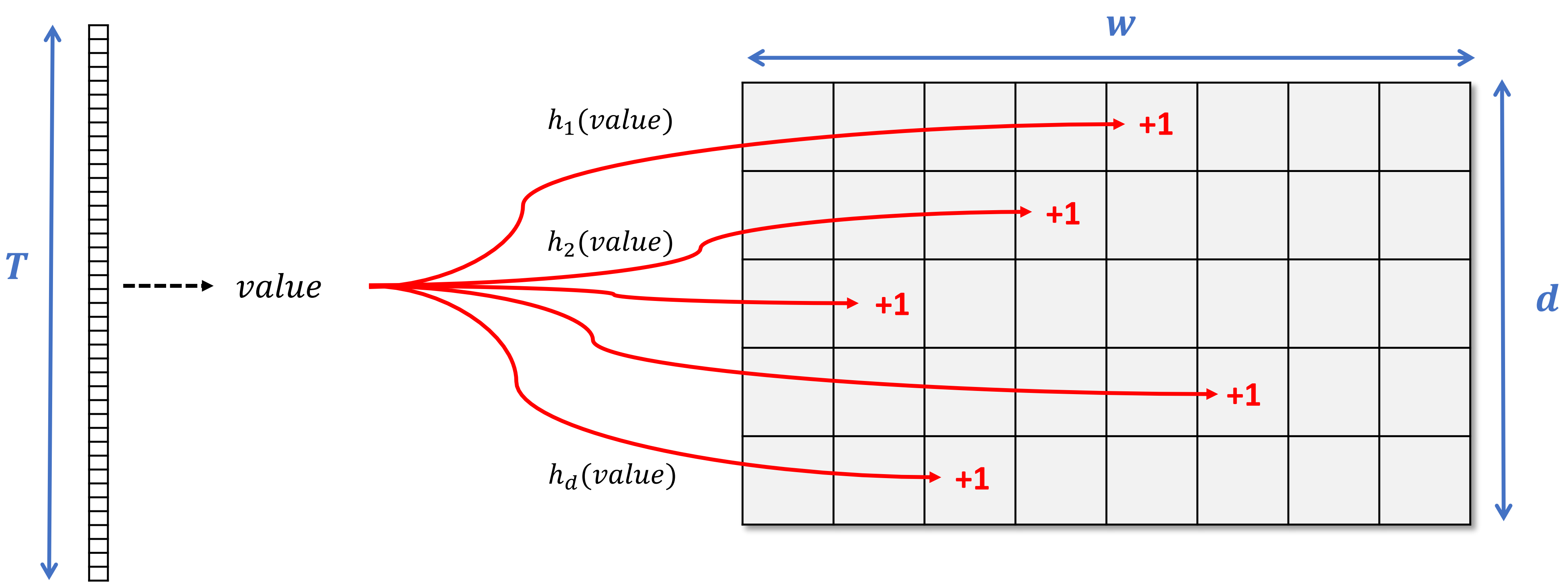
L. Melis, G. Danezis, E. De Cristofaro. Efficient Private Statistics with Succinct Sketches. NDSS'16.
(Winner of the 5th Catalan Data Protection Authority's Privacy by Design Award)

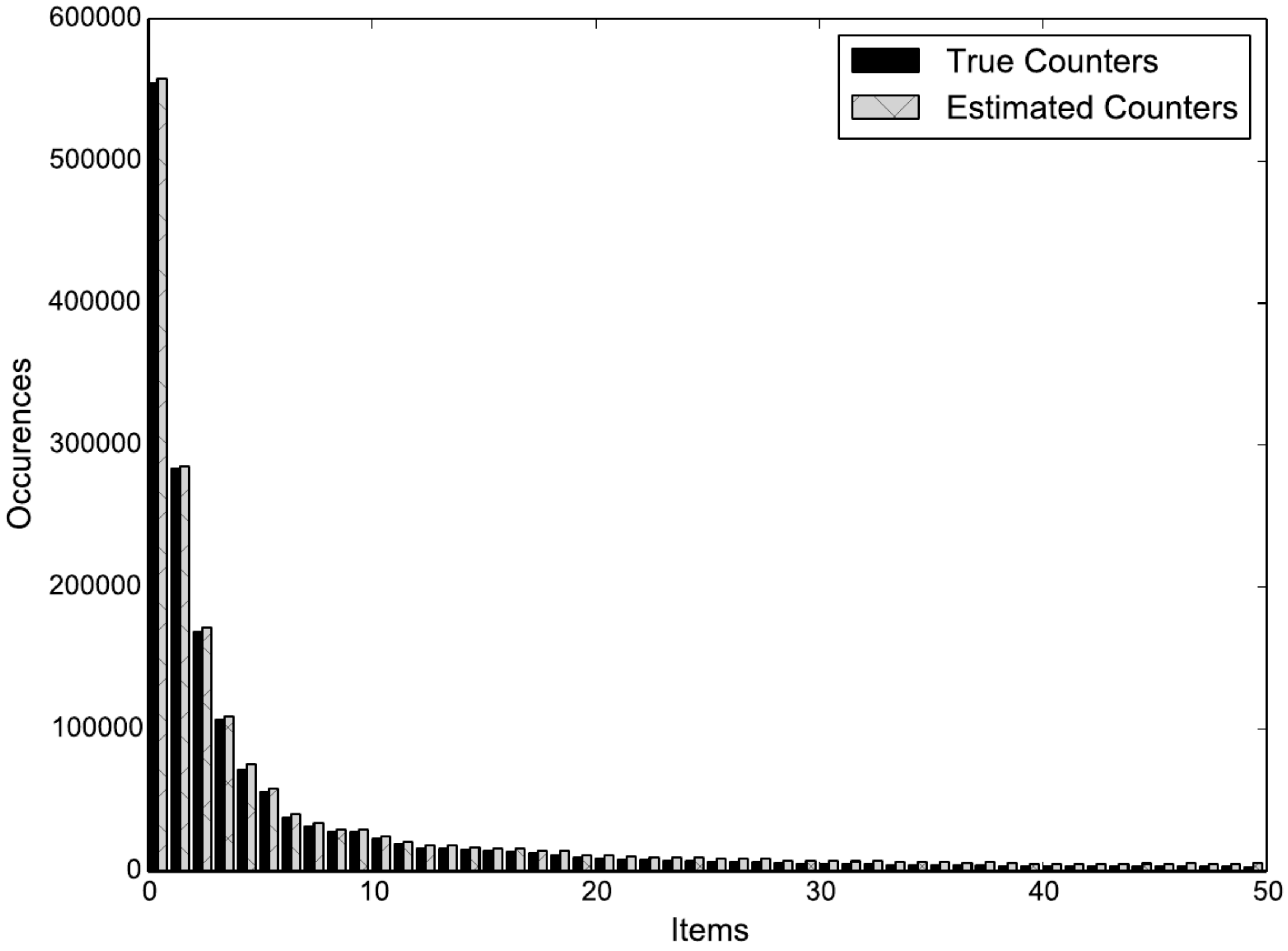
Count(-Min) Sketch

Estimate an item's frequency in a stream

Mapping a stream of values (of length T) to a matrix of size O(logT)

Sum of two sketches = sketch of the union of the two data streams



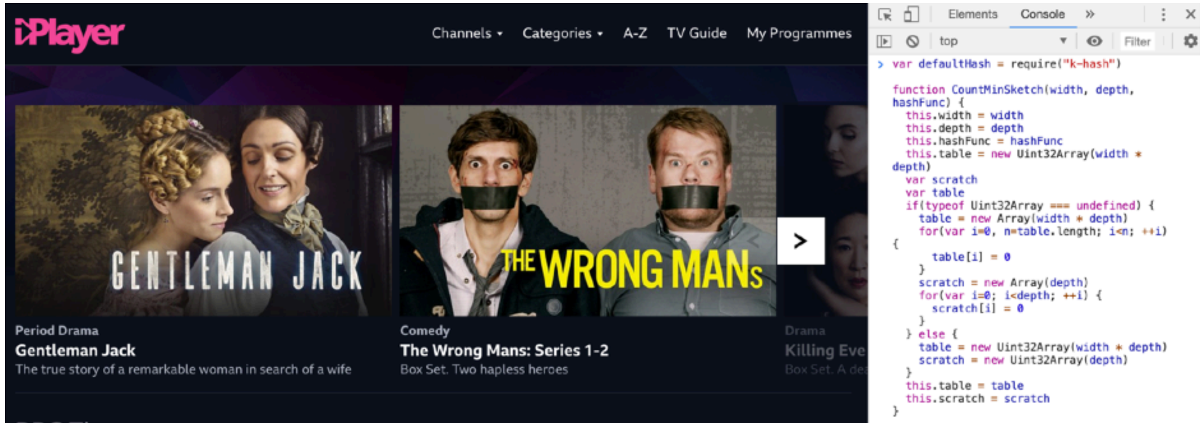

600000
500000
400000
300000
200000
100000
0
Occurences
0
10
20
30
40
50
Items
True Counters
Estimated Counters

Prototype Implementation

Tally (server-side) as a Node.js web server

Client-side in JavaScript, runs in the browser or as a mobile cross-platform application (Apache Cordova)

Deploying as easy as installing a Node.js package via NPM



Succinct Data Structures+PDA
also useful in other settings…

Succinct Data Structures+PDA also useful in other settings

Succinct Data Structures+PDA also useful in other settings

HSDir statistics
[long standing problem]

Succinct Data Structures+PDA also useful in other settings

HSDir statistics
[long standing problem]

EPSRC IRC in Early Warning Sensing Systems for Infectious Diseases

Inferring population health statistics (e.g., influenza) from Google searches [Primault et al., WWW'19]

Agenda

1. Training (Distributed) ML Models with Privacy

2. Private Data Release with Generative Neural Networks

3. Privacy Leakage in Collaborative/Federated ML

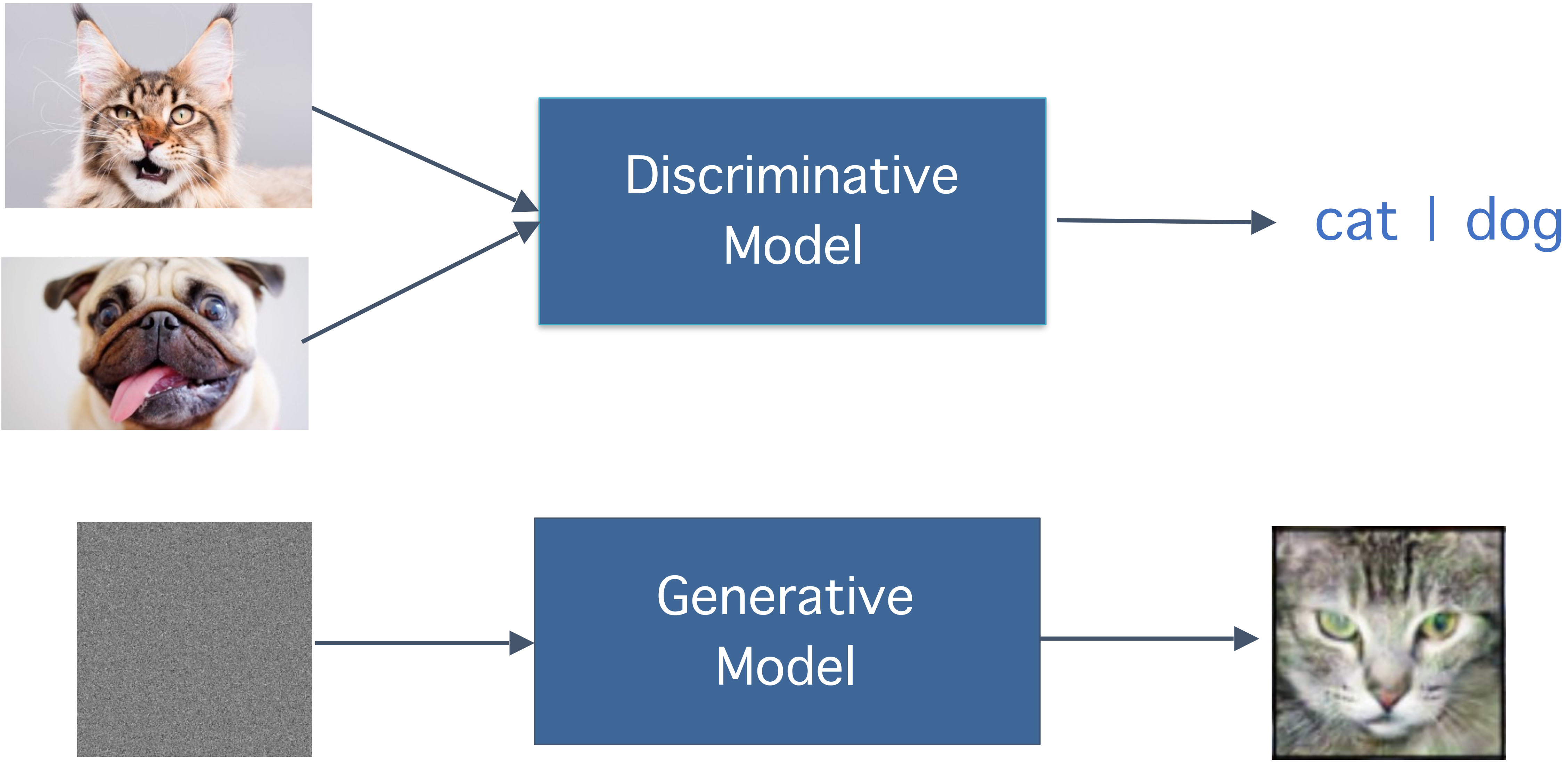
Agenda

1. Training (Distributed) ML Models with Privacy

2. Private Data Release with Generative Neural Networks

3. Privacy Leakage in Collaborative/Federated ML


Discriminative
Model
cat | dog


Discriminative Model
cat | dog
Generative Model

Differential Privacy (Weaker Notion)

Differential Privacy (Weaker Notion)

Let X be the "data universe"

Let $D \subset X$ be the "dataset"

Differential Privacy (Weaker Notion)

Let X be the “data universe”

Let $D \subset X$ be the “dataset”

Definition: An Algorithm M is (ε,δ)-differentially private if for all pairs of neighboring datasets (D,D’), and for all outputs x:

$$\Pr[M(D)=x] \leq \exp(\varepsilon) * \Pr[M(D')=x] + \delta$$

Differential Privacy (Weaker Notion)

Let X be the "data universe"

Let $D \subset X$ be the "dataset"

Definition: An Algorithm M is (ε,δ)-differentially private if for all pairs of neighboring datasets (D,D'), and for all outputs x:

$$\Pr[M(D)=x] \leq \exp(\varepsilon) * \Pr[M(D') = x] + \delta$$

quantifies information leakage

Differential Privacy (Weaker Notion)

Let X be the "data universe"

Let $D \subset X$ be the "dataset"

Definition: An Algorithm M is (ε,δ)-differentially private if for all pairs of neighboring datasets (D,D'), and for all outputs x:

$$\Pr[M(D)=x] \leq \exp(\varepsilon) * \Pr[M(D') = x] + \delta$$

quantifies information leakage

allows for a small probability of failure

Some Useful Properties

Some Useful Properties

Theorem (Post-Processing):

If M(D) is ε-private, for any function f, then f(M(D)) is ε-private

Some Useful Properties

Theorem (Post-Processing):

If $M(D)$ is ε-private, for any function f, then $f(M(D))$ is ε-private

Theorem (Composition):

If $M_1,\ldots,M_k$ are ε-private, then $M(D)=M(M_1(D),\ldots,M_k(D))$ is $(k*\varepsilon)$-private

Some Useful Properties

Theorem (Post-Processing):

If $M(D)$ is ε-private, for any function f, then $f(M(D))$ is ε-private

Theorem (Composition):

If $M_1,\ldots,M_k$ are ε-private, then $M(D)=M(M_1(D),\ldots,M_k(D))$ is $(k*\varepsilon)$-private

We can apply algorithms as we normally would; access the data using differentially private subroutines, and keep track of privacy budget (Modularity)

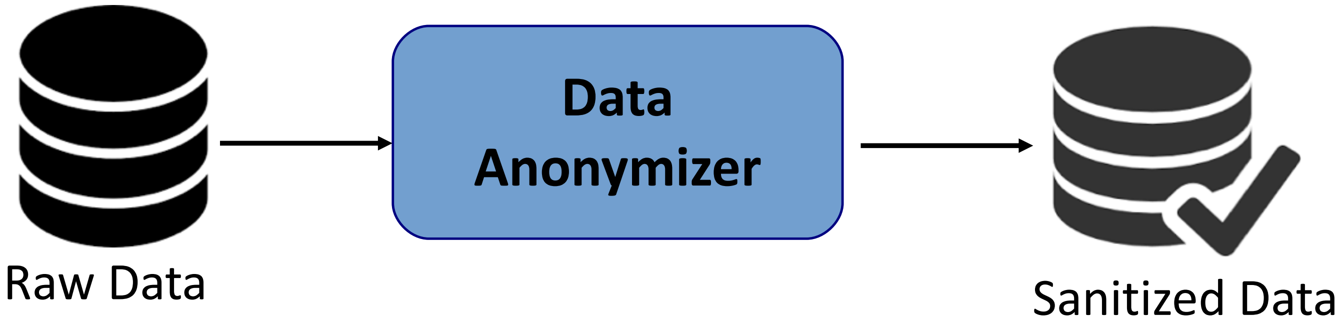
Motivation

Motivation

Organizations need/want to publish their datasets without compromising users' privacy

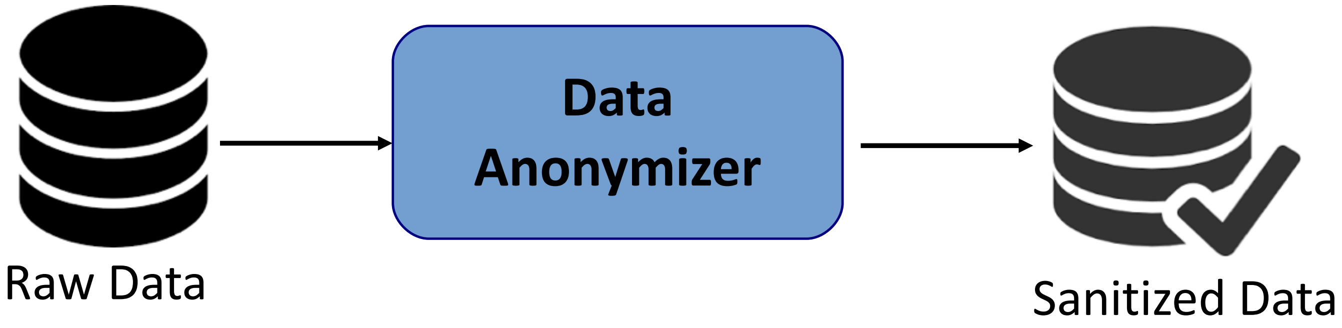
Motivation

Organizations need/want to publish their datasets without compromising users' privacy



Motivation

Organizations need/want to publish their datasets without compromising users' privacy

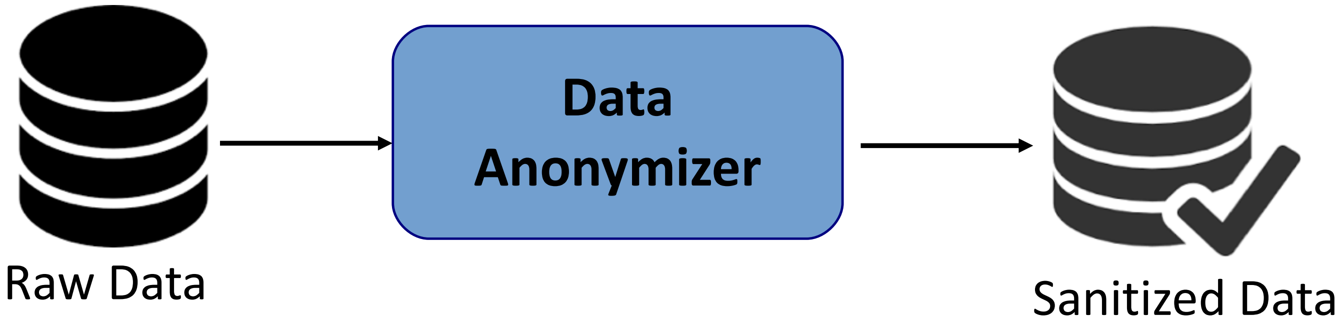


Differential Privacy: Weak utility, "curse of dimensionality"(*)

(*) Brickell & Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD 2008.

Motivation

Organizations need/want to publish their datasets without compromising users' privacy



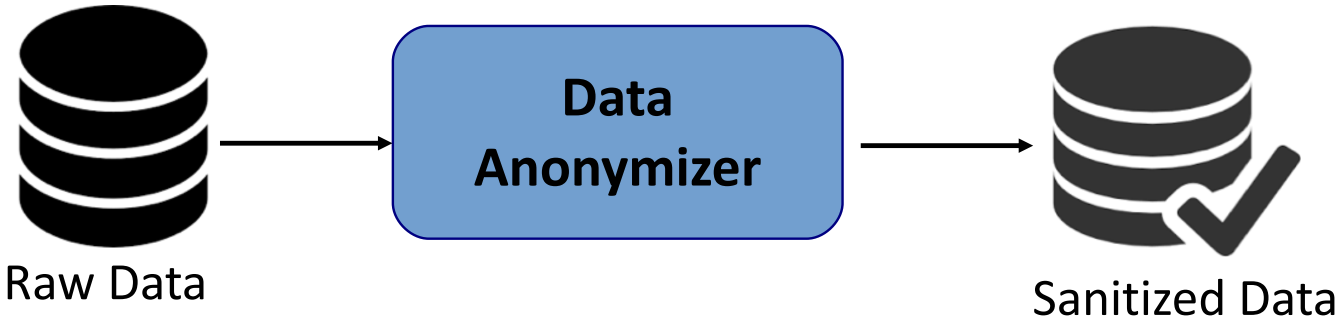
Differential Privacy: Weak utility, "curse of dimensionality"(*)

k-Anonymity: no real privacy

(*) Brickell & Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD 2008.

Motivation

Organizations need/want to publish their datasets without compromising users' privacy



Differential Privacy: Weak utility, "curse of dimensionality" (

k-Anonymity: no real privacy

(*) Brickell & Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD 2008.

How about generating synthetic dataset?

How about generating synthetic dataset?

Gergely Acs, Luca Melis, Claude Castelluccia, Emiliano De Cristofaro. Differentially Private Mixture of Generative Neural Networks. In IEEE ICDM'17. (Extended version in IEEE TKDE)

Main Idea

Main Idea

Model the data-generating distribution by training a generative model on the original data

Publish the model along with its differentially private parameters

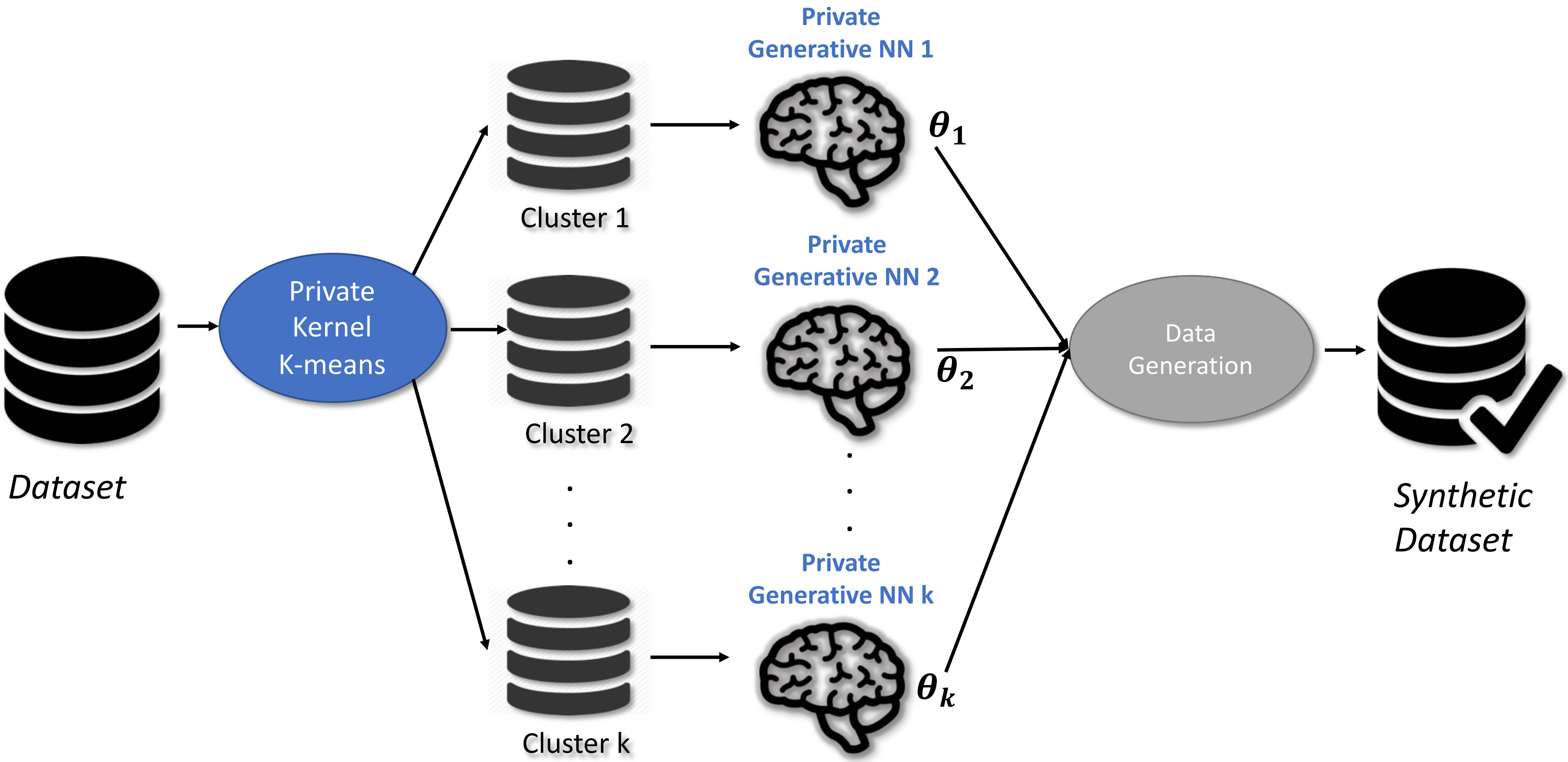
Main Idea

Model the data-generating distribution by training a generative model on the original data

Publish the model along with its differentially private parameters

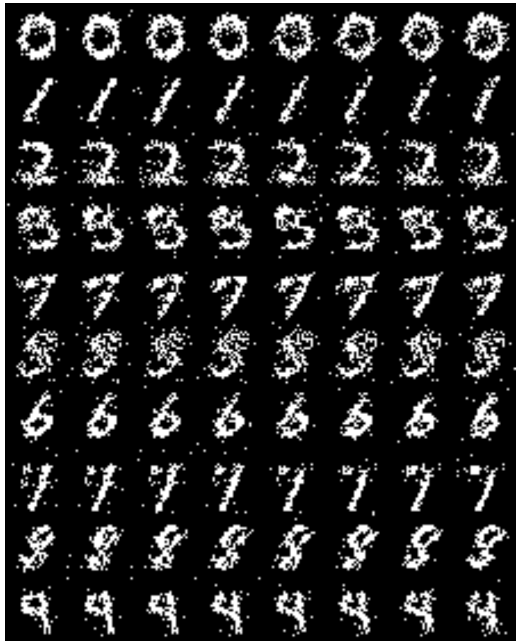
Anybody can generate a synthetic dataset resembling the original (training) data

With strong (differential) privacy protection

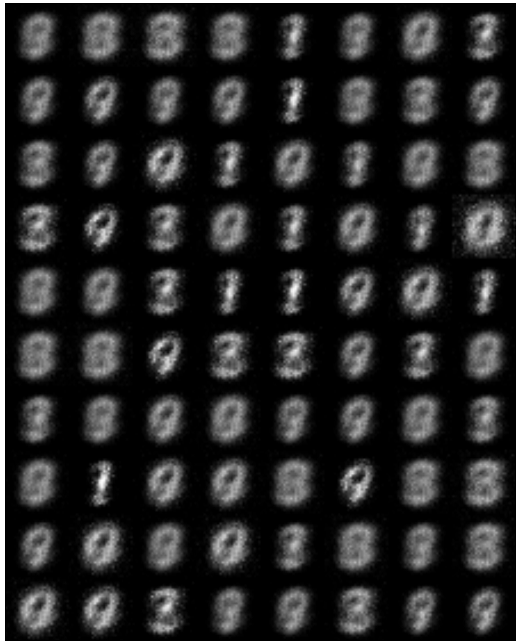

Dataset
Private
Kernel
K-means
Cluster 1
Cluster 2
Cluster k
Private
Generative NN 1
Private
Generative NN 2
Private
Generative NN k
θ_1
θ_2
θ_k
Data
Generation
Synthetic
Dataset

Synthetic Samples (MNIST)

Original samples



RBM samples



VAE w/o clustering

VAE with clustering

20 SGD epochs (epsilon=1.74)

Agenda

1. Training (Distributed) ML Models with Privacy
2. Private Data Release with Generative Neural Networks
3. Privacy Leakage in Collaborative/Federated ML

Agenda

1. Training (Distributed) ML Models with Privacy
2. Private Data Release with Generative Neural Networks
3. Privacy Leakage in Collaborative/ Federated ML

Agenda

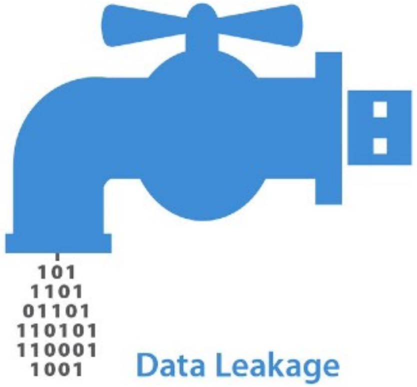
1. Training (Distributed) ML Models with Privacy

2. Private Data Release with Generative Neural Networks

3. Privacy Leakage in Collaborative/ Federated ML

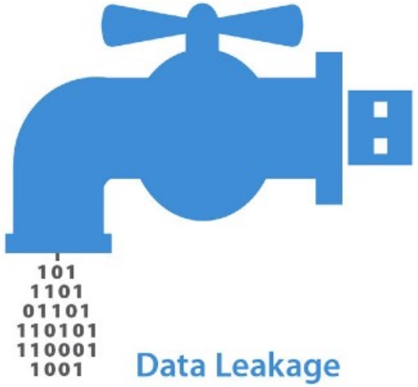
Luca Melis, Congzheng Song, Emiliano De Cristofaro, Vitaly Shmatikov. Inference Attacks Against Collaborative Learning. IEEE Symposium on Security & Privacy (S&P'19)

Reasoning about “privacy” in ML



Reasoning about "privacy" in ML

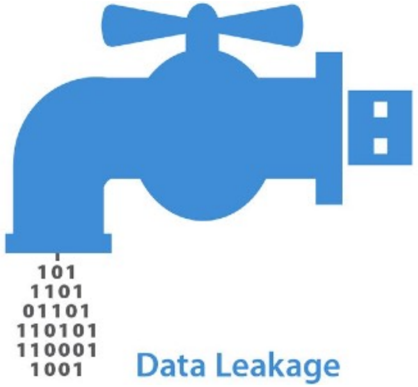
Most papers on privacy attacks in ML focus on inferring:



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

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

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

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

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?

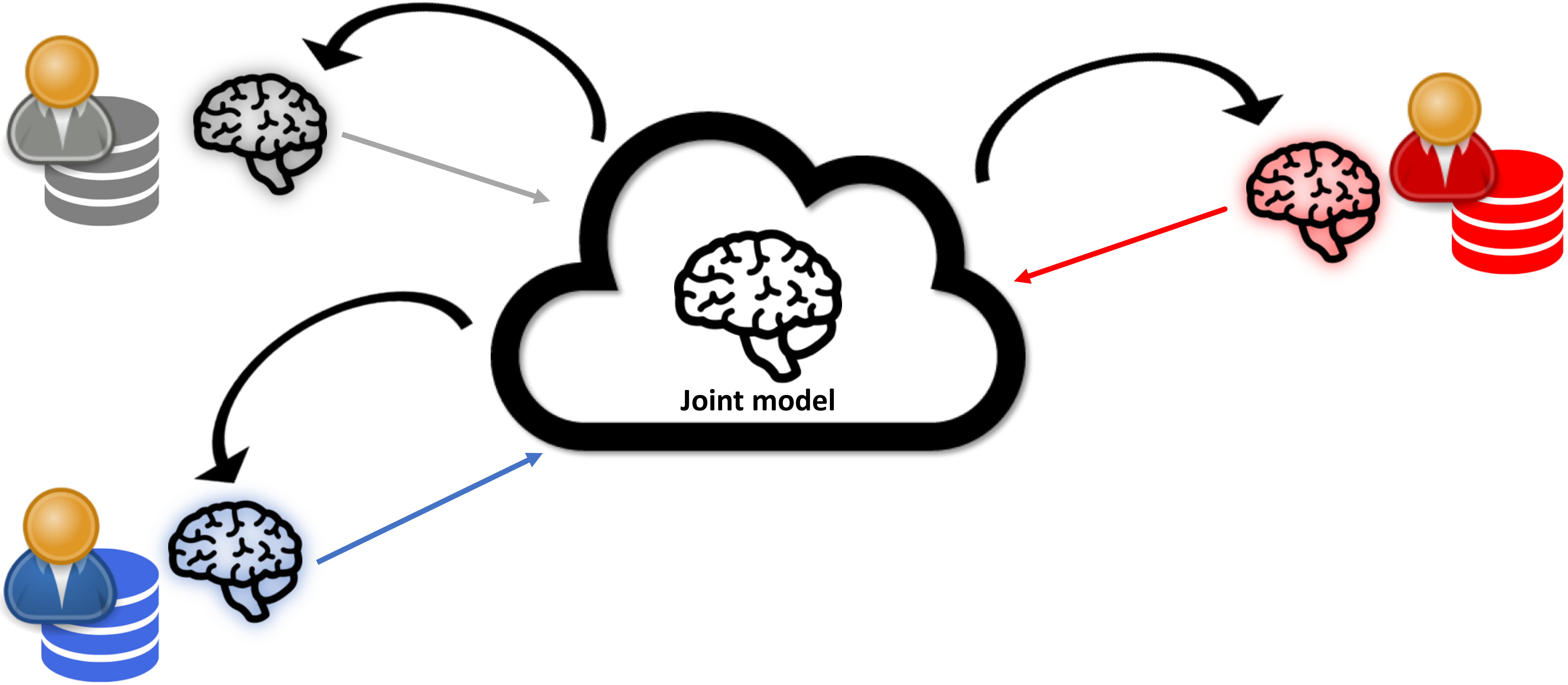
Intuition

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

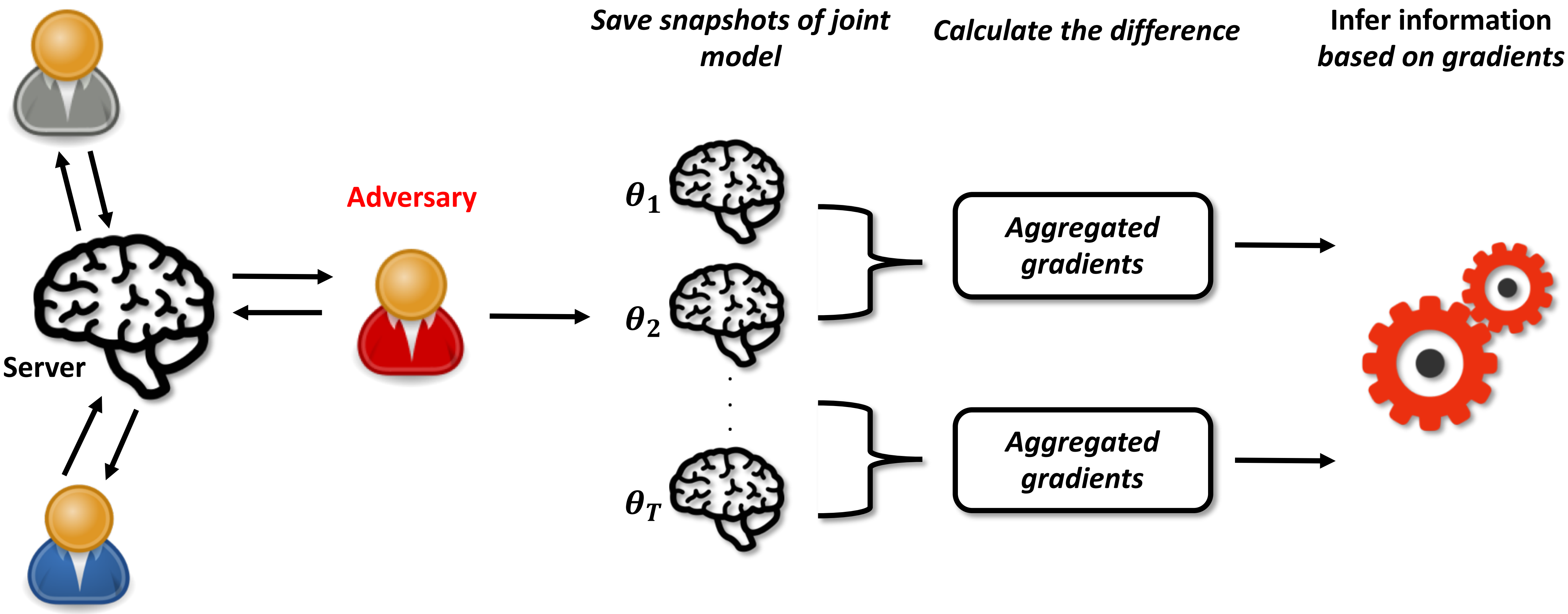
…but not of the whole class?

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

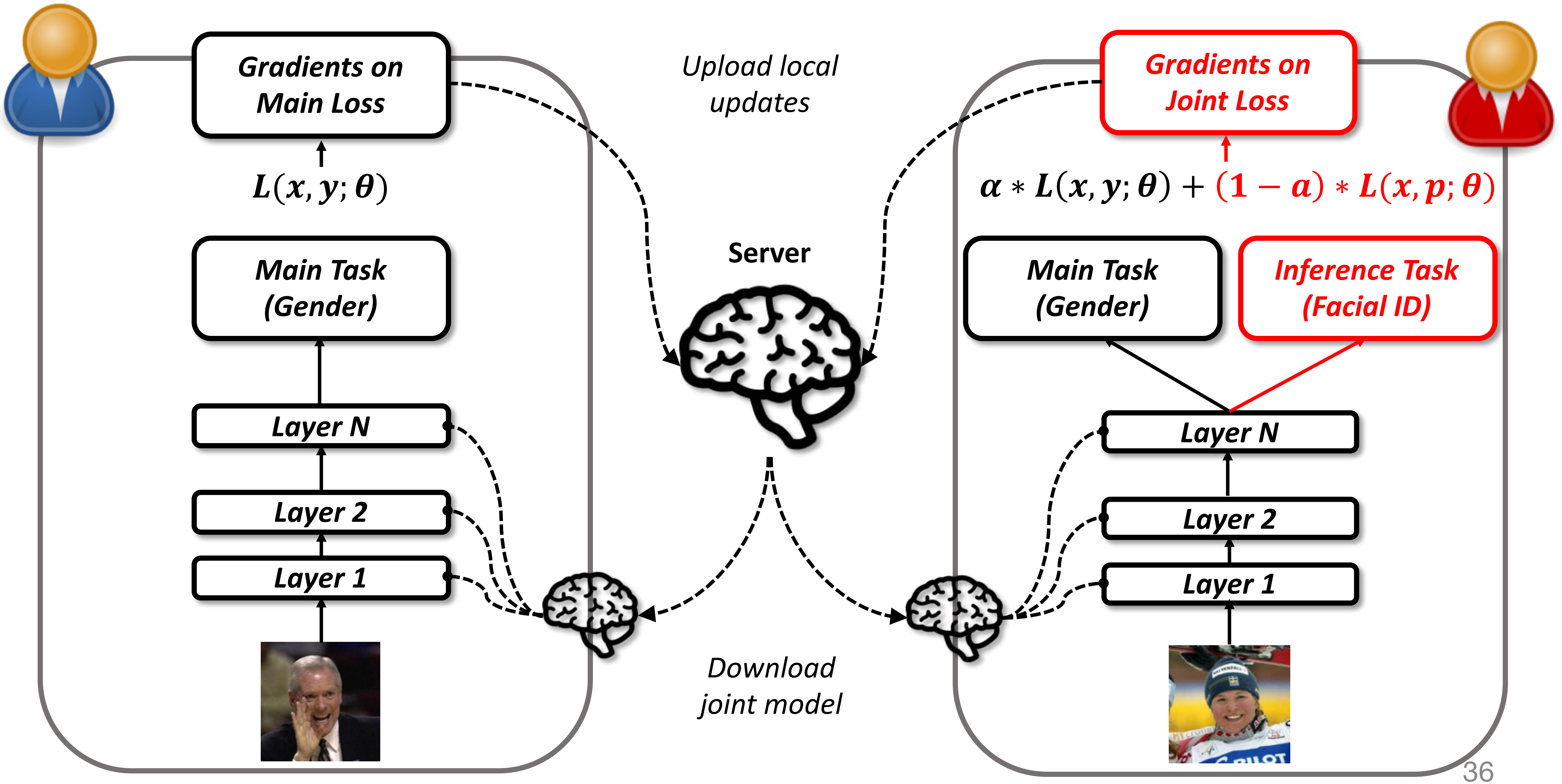
Collaborative Learning



Passive Property Inference Attack



Active Property Inference Attack

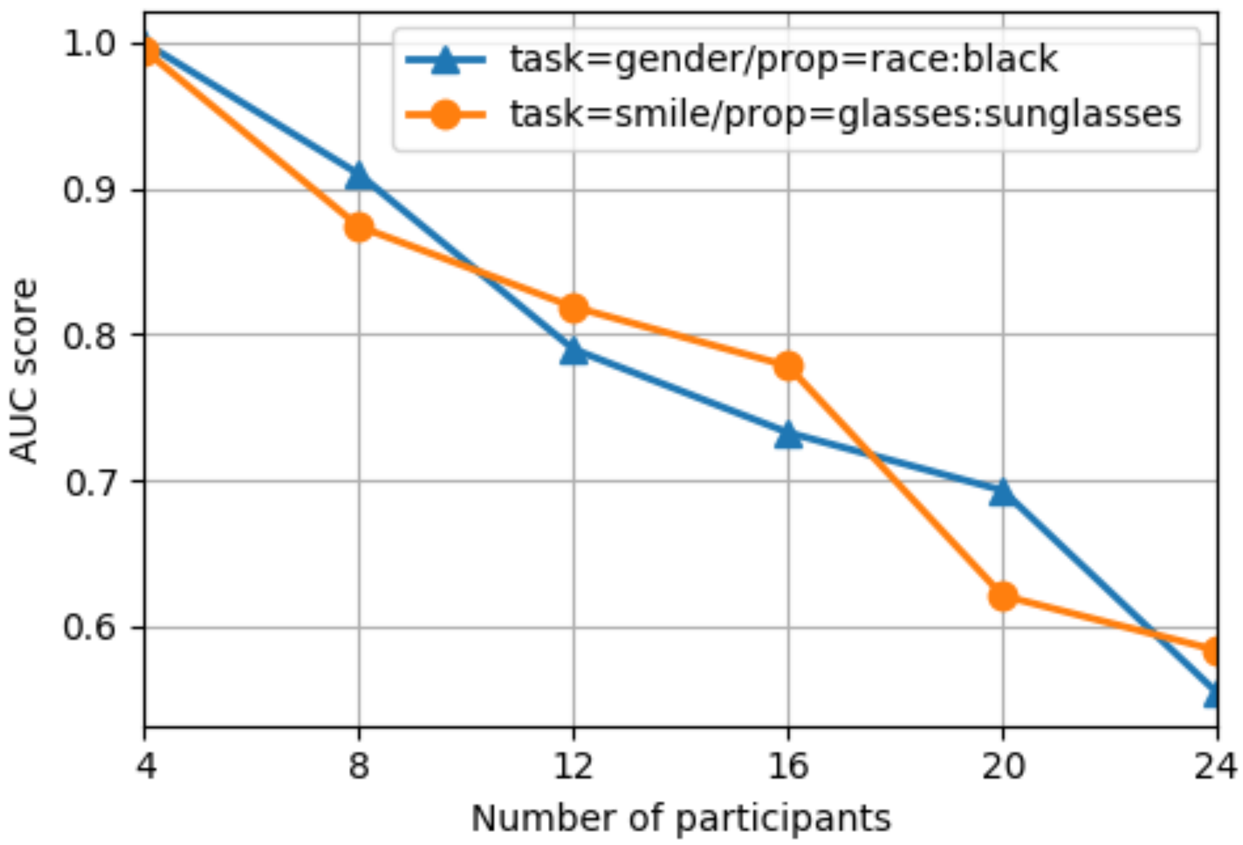


Dataset	Type	Main Task	Inference Task
LFW	Images	Gender/Smile/Age Eyewear/Race/Hair	Race/Eyewear
FaceScrub	Images	Gender	Identity
PIPA	Images	Age	Gender
FourSquare	Locations	Gender	Membership
Yelp-health	Text	Review Score	Membership Doctor specialty
Yelp-author	Text	Review Score	Author
CSI	Text	Sentiment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

Property Inference on LFW

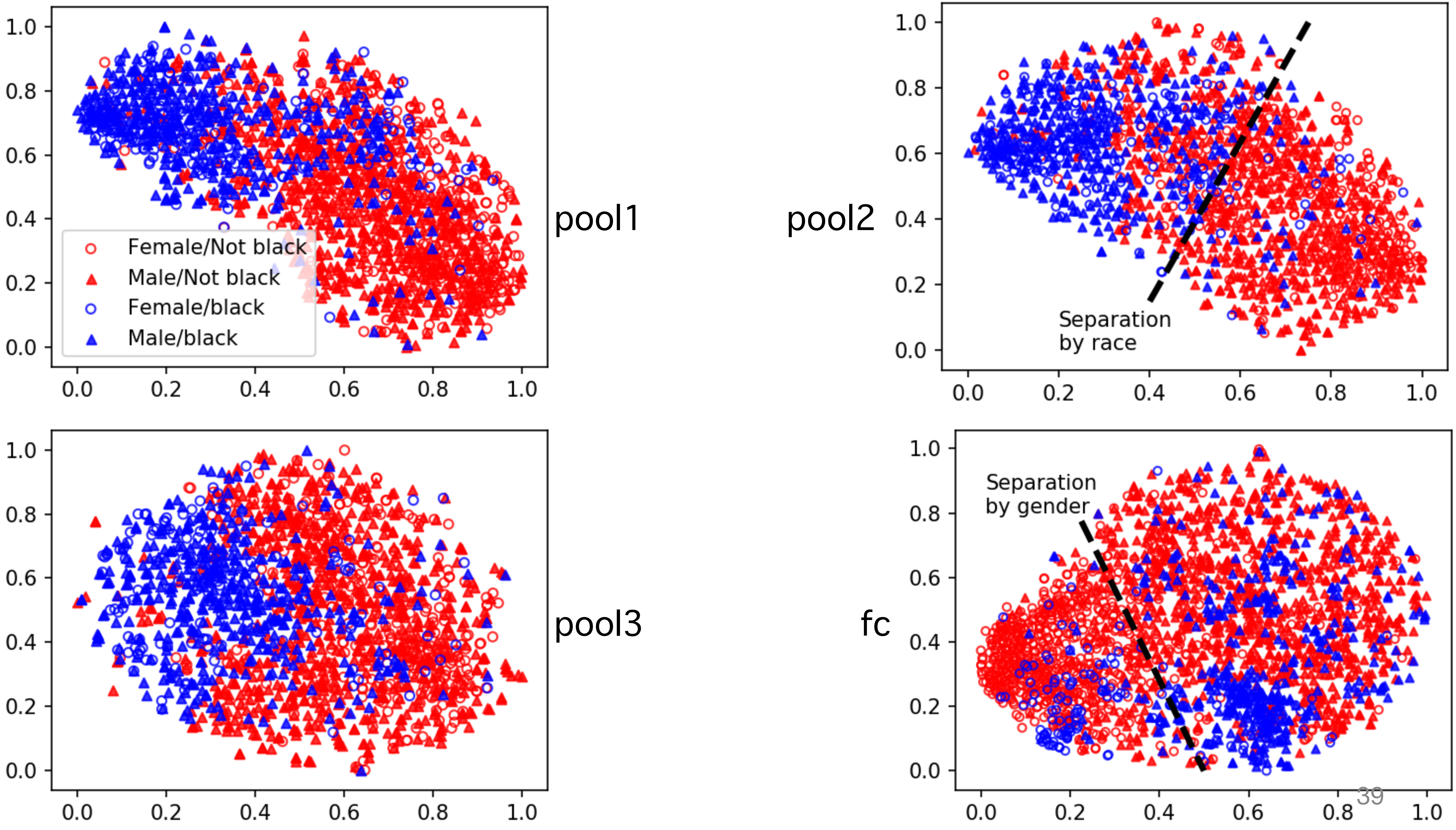
Main Task	Inference Task	Correlation	AUC score
Gender	Sunglasses	-0.025	1.0
Smile	Asian	0.047	0.93
Age	Black	-0.084	1.0
Race	Sunglasses	0.026	1.0
Eyewear	Asian	-0.119	0.91
Hair	Sunglasses	-0.013	1.0

Two-Party



Multi-Party

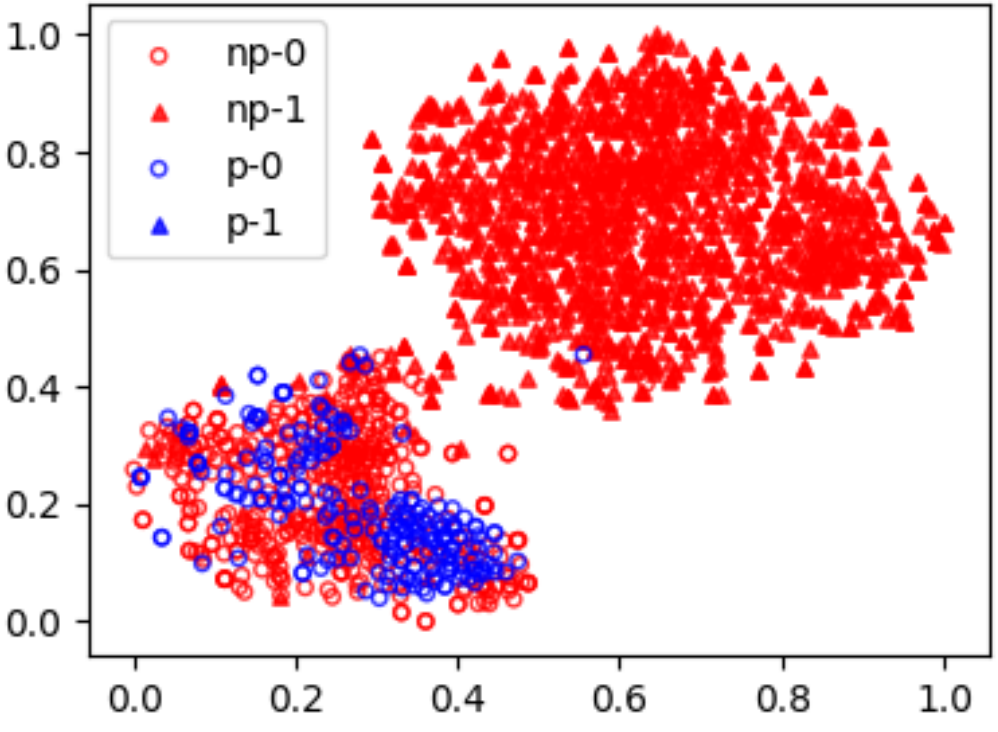
Feature t-SNE projection


Female/Not black
Male/Not black
Female/black
Male/black
pool1
pool2
Separation by race
pool3
fc
Separation by gender

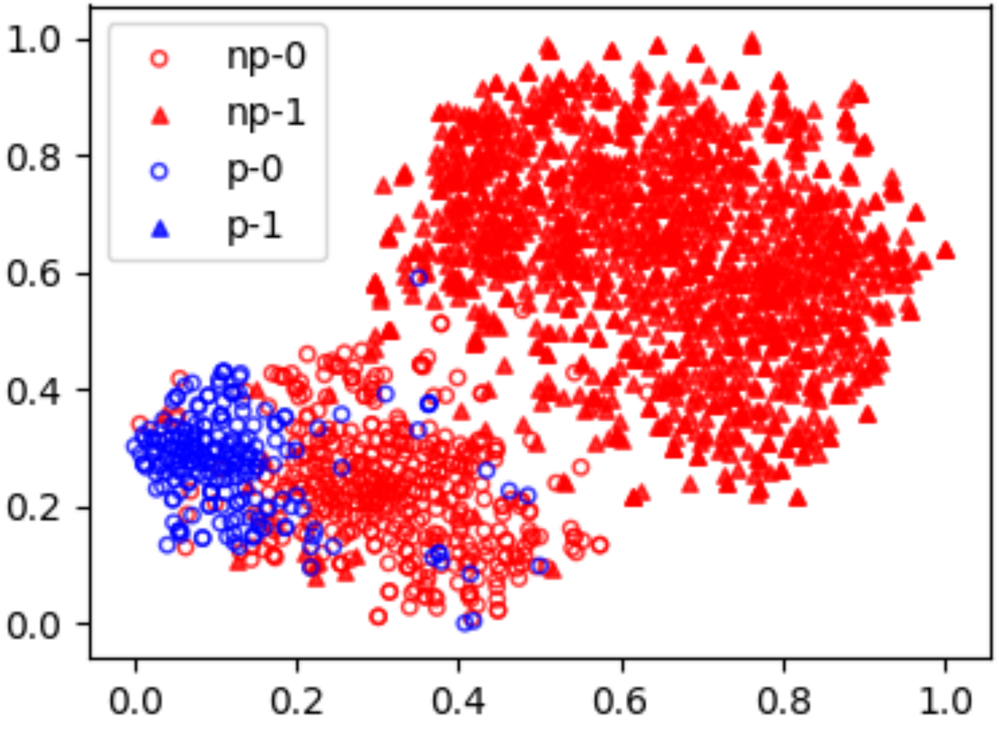
Passive vs Active Attack on FaceScrub

Main Task: ▲/●= female/male

Inference Task: Blue points with the property (identity)



Passive attack

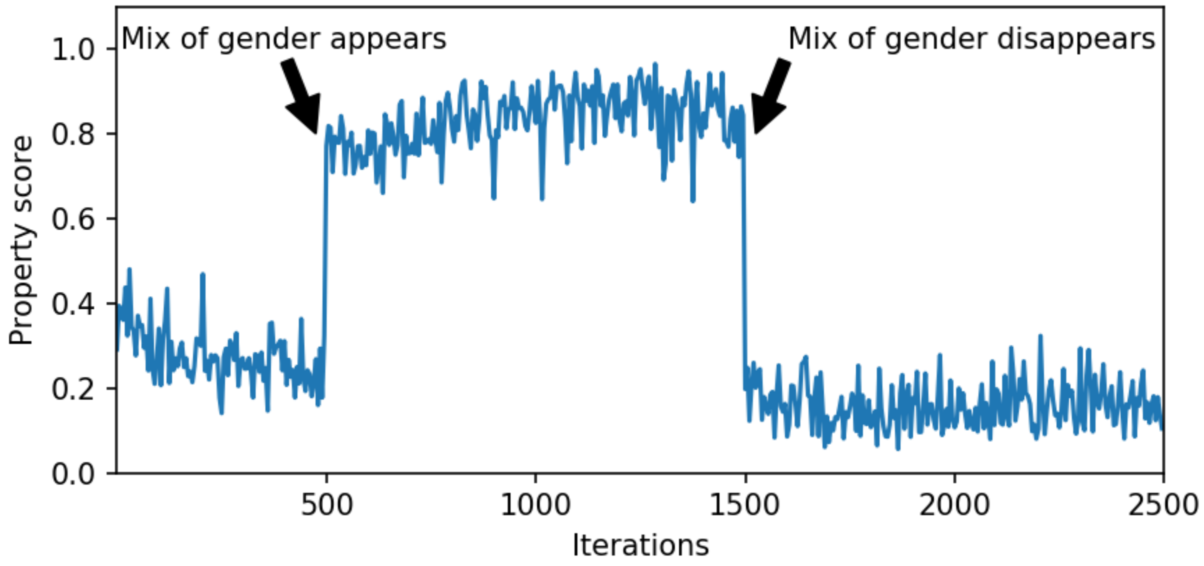


Active attack

Inferring when a property occurs

Inferring when a property occurs

Batches with the property appear

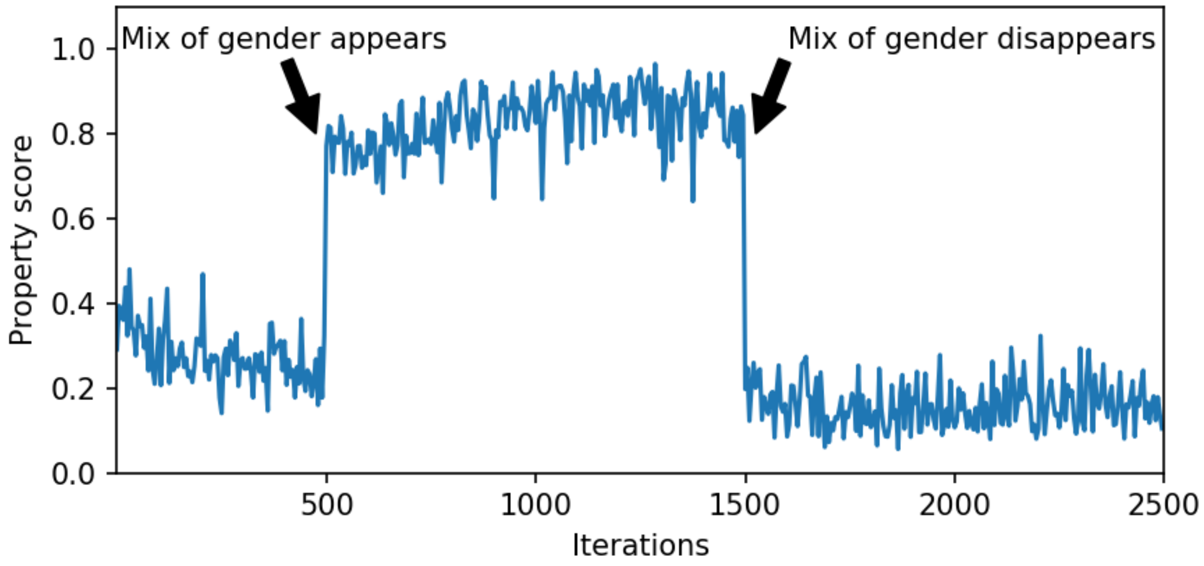


Main task: Age / Two-party

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

Inferring when a property occurs

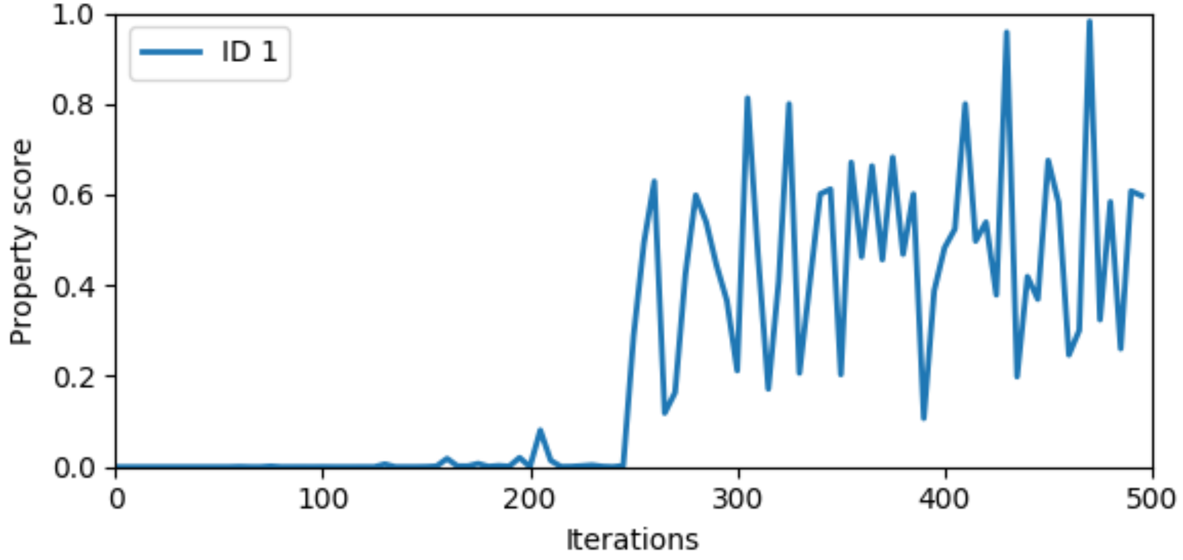
Batches with the property appear



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…

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 2 “hand-crafted” mechanisms in the literature

Fail to converge for “few” participants

Thank you!



Thank you!