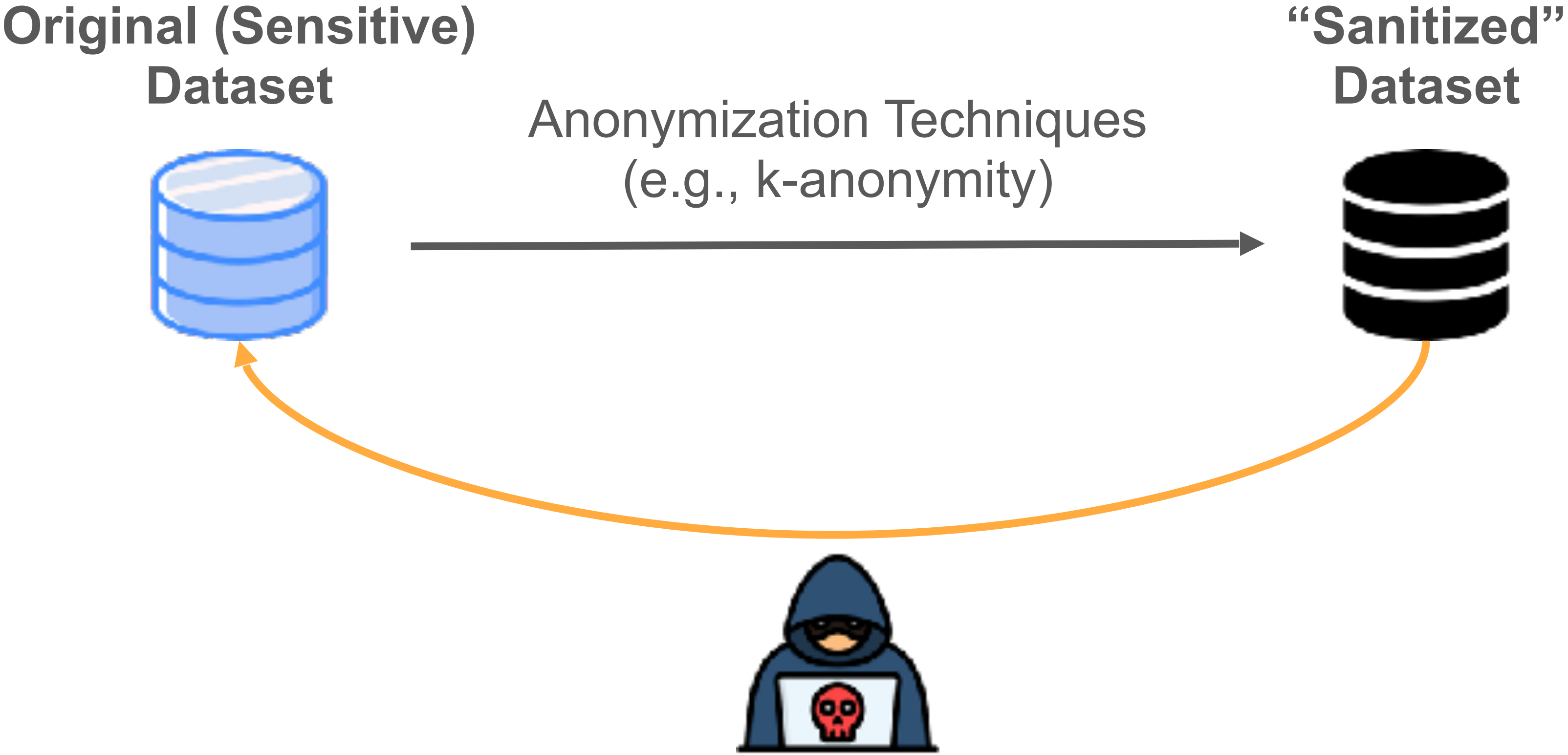
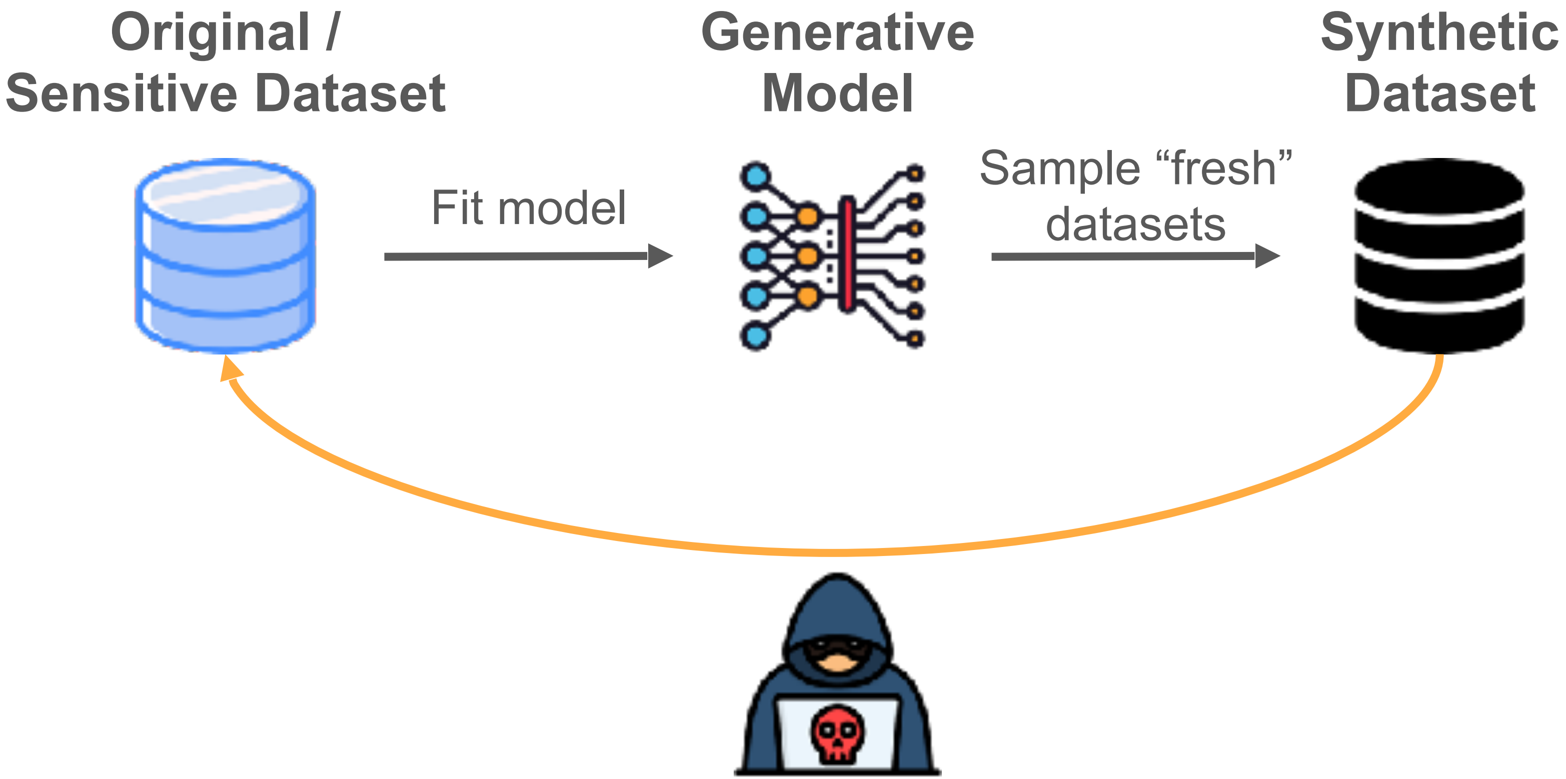
Synthetic Data from (Differentially Private) Generative Models: It's Complicated

Emiliano De Cristofaro

What is Synthetic Data?



What is Synthetic Data?



Privacy Attacks in Machine Learning

Privacy Attacks in Machine Learning

- Inclusion of a data point in the training set

  “membership inference”

Privacy Attacks in Machine Learning

- Inclusion of a data point in the training set

  “membership inference”

- What class representatives (in training set) look like

  “model inverstion”

Privacy Attacks in Machine Learning

- Inclusion of a data point in the training set
  "membership inference"

- What class representatives (in training set) look like
  "model inverstion"

- Attributes of training data
  "property inference"

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

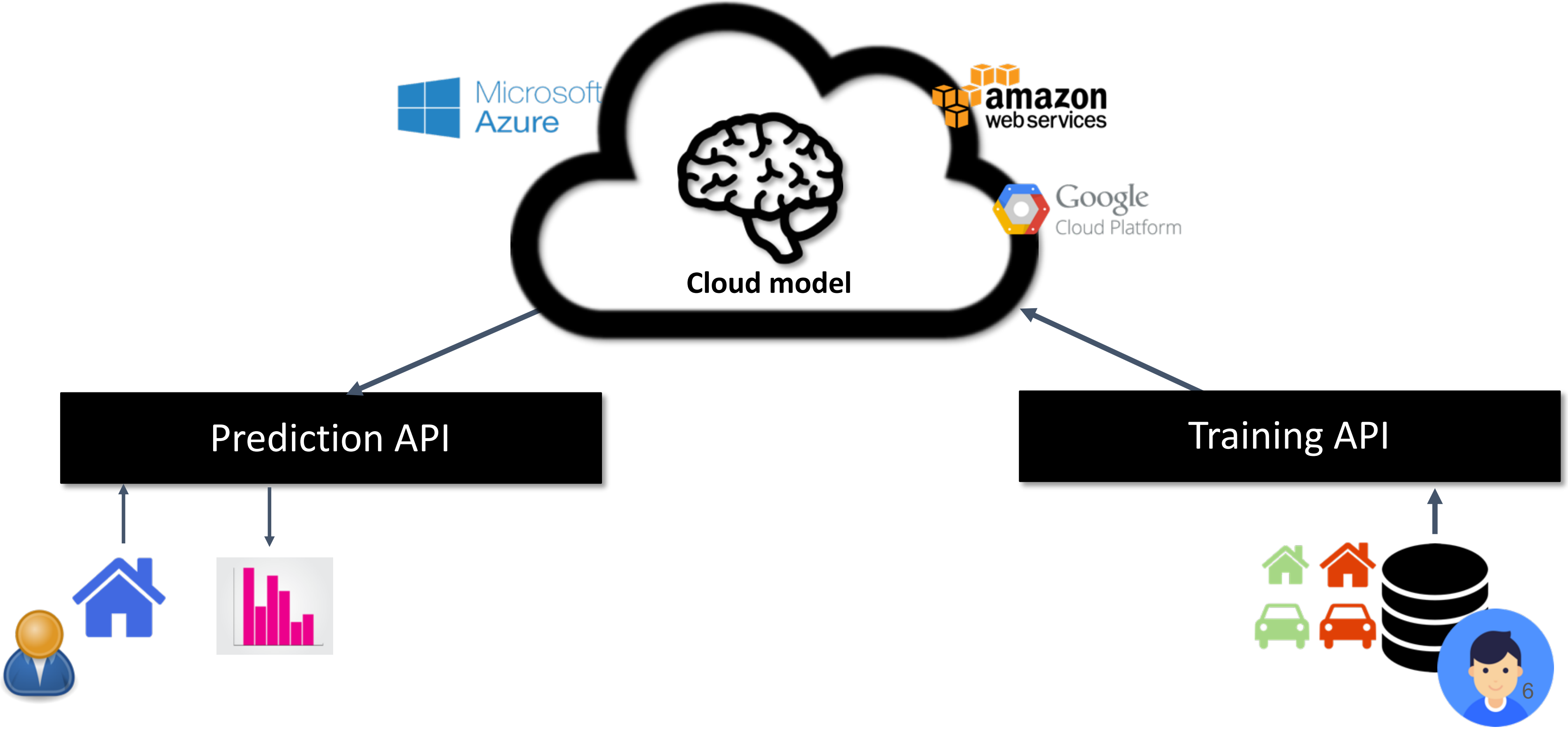
Membership Inference

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

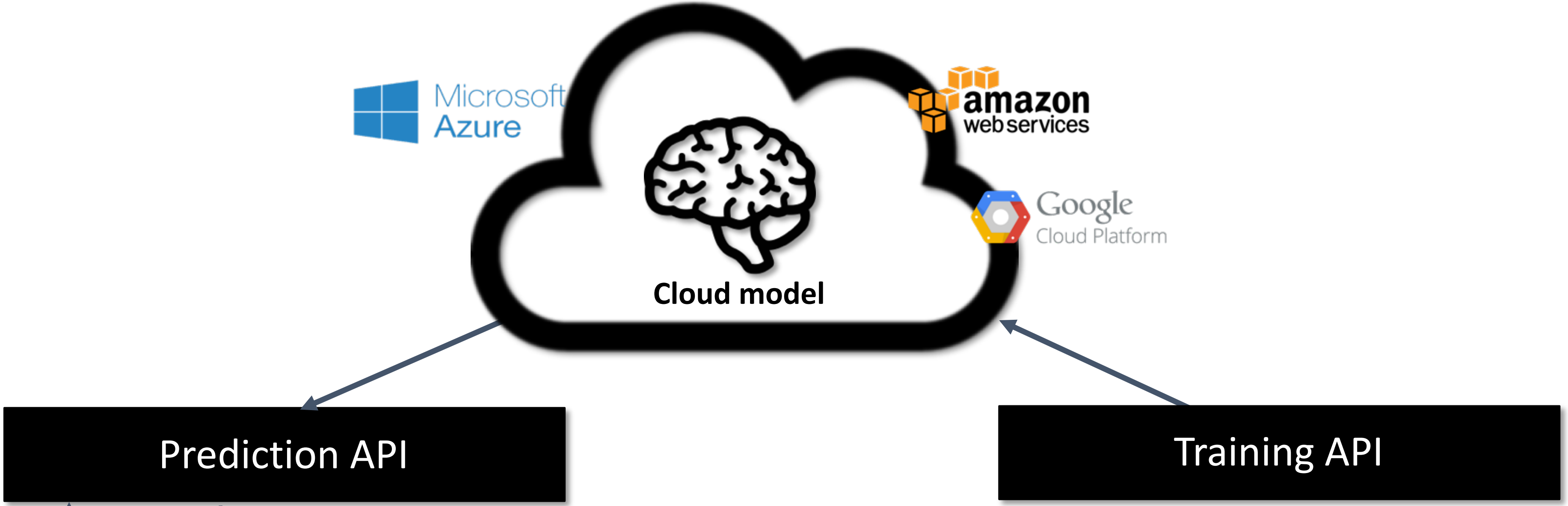
- Serious problem if inclusion in training set is privacy-sensitive
- E.g., main task is predict whether a smoker gets cancer
- [Shokri et al., S&P'17] show it for discriminative models
- [Hayes et al. PETS'19] for generative models (in this talk)

Machine Learning as a Service

Machine Learning as a Service



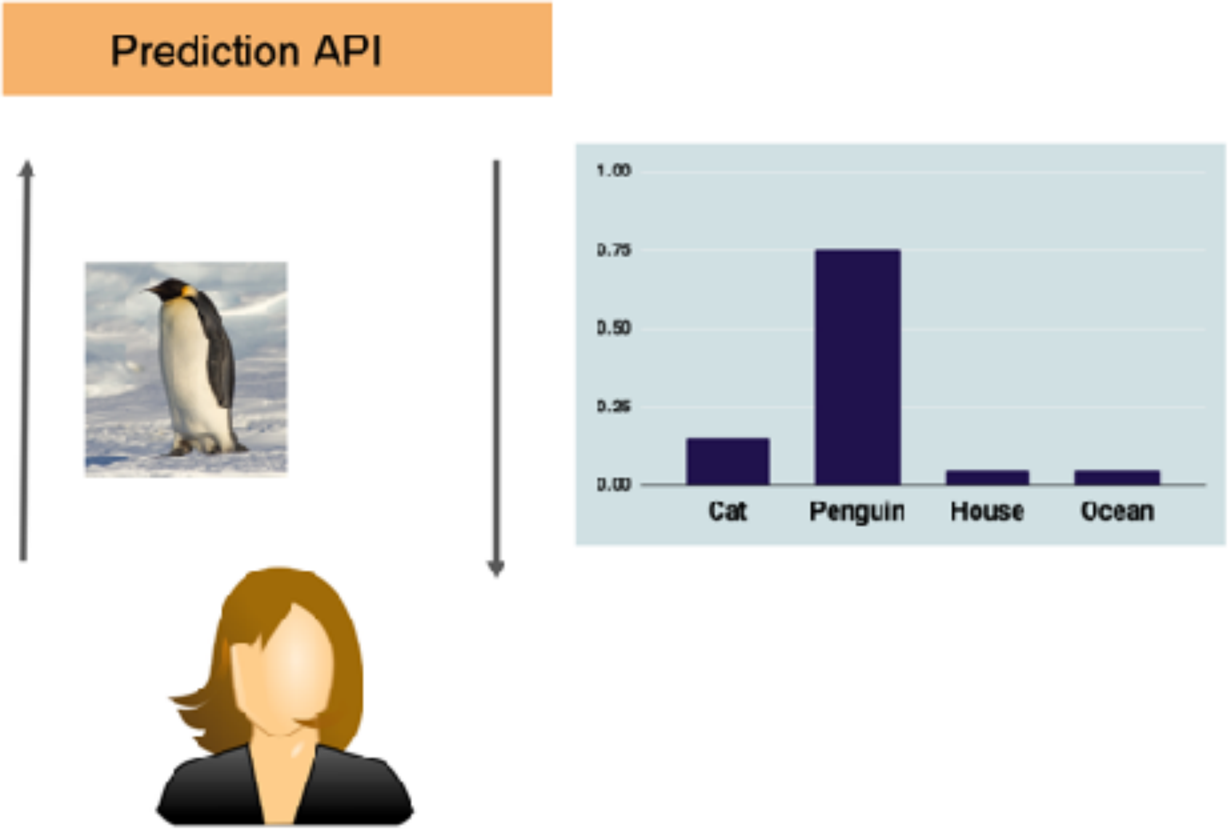
Machine Learning as a Service

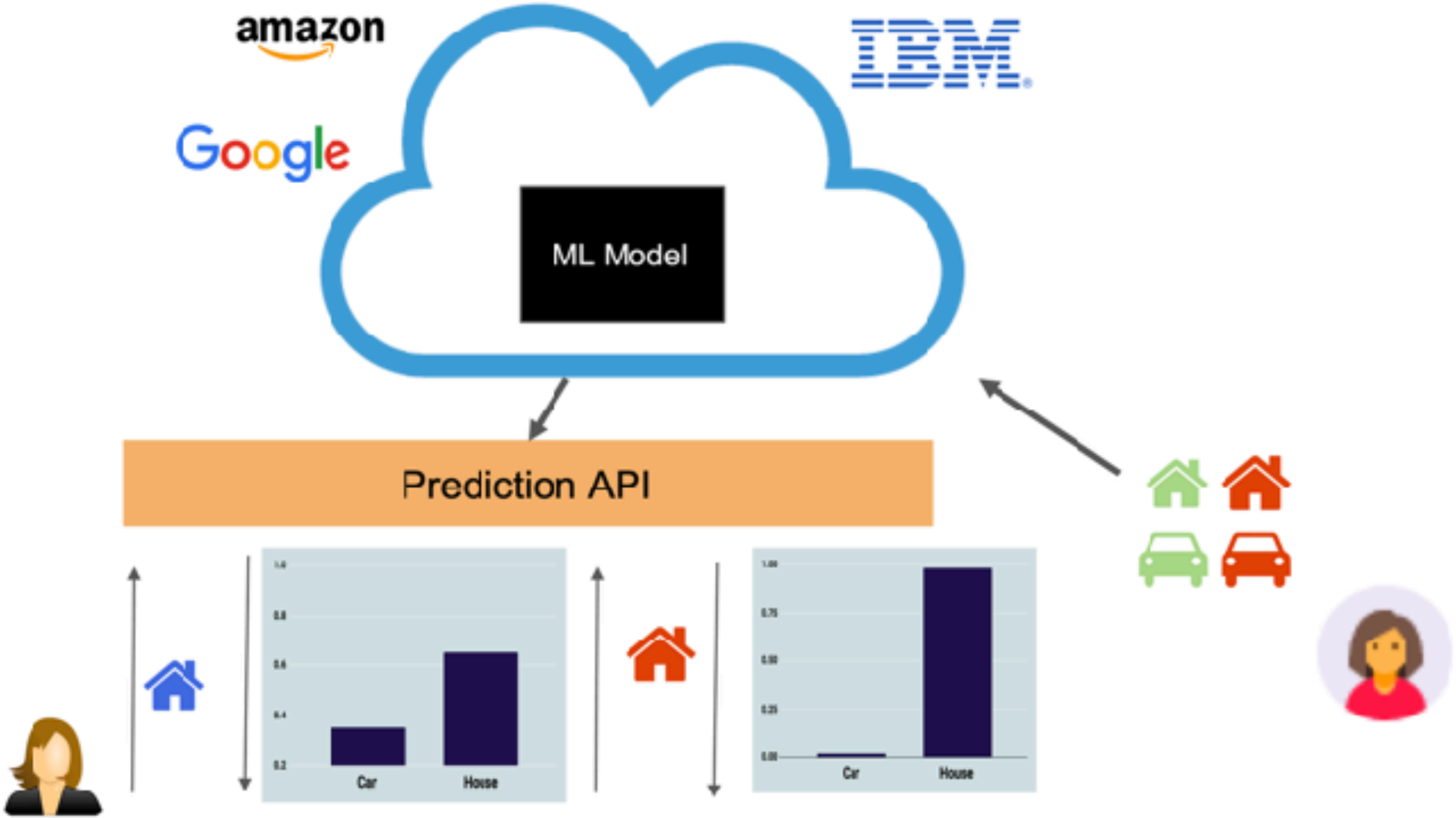


Predictions are leaky!

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

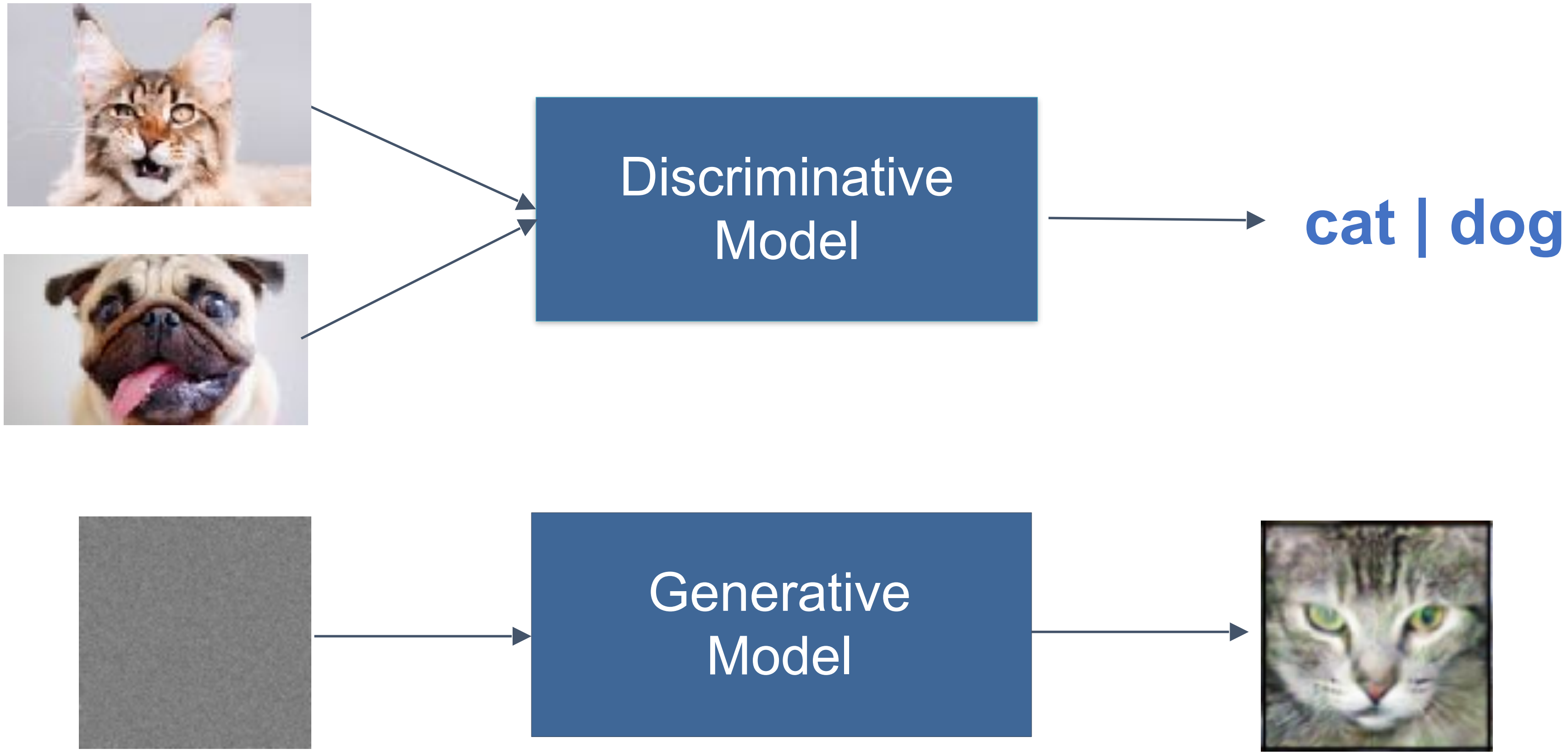
Membership Inference/Discriminative




amazon
Google
IBM
ML Model
Prediction API
Car
House
Car
House

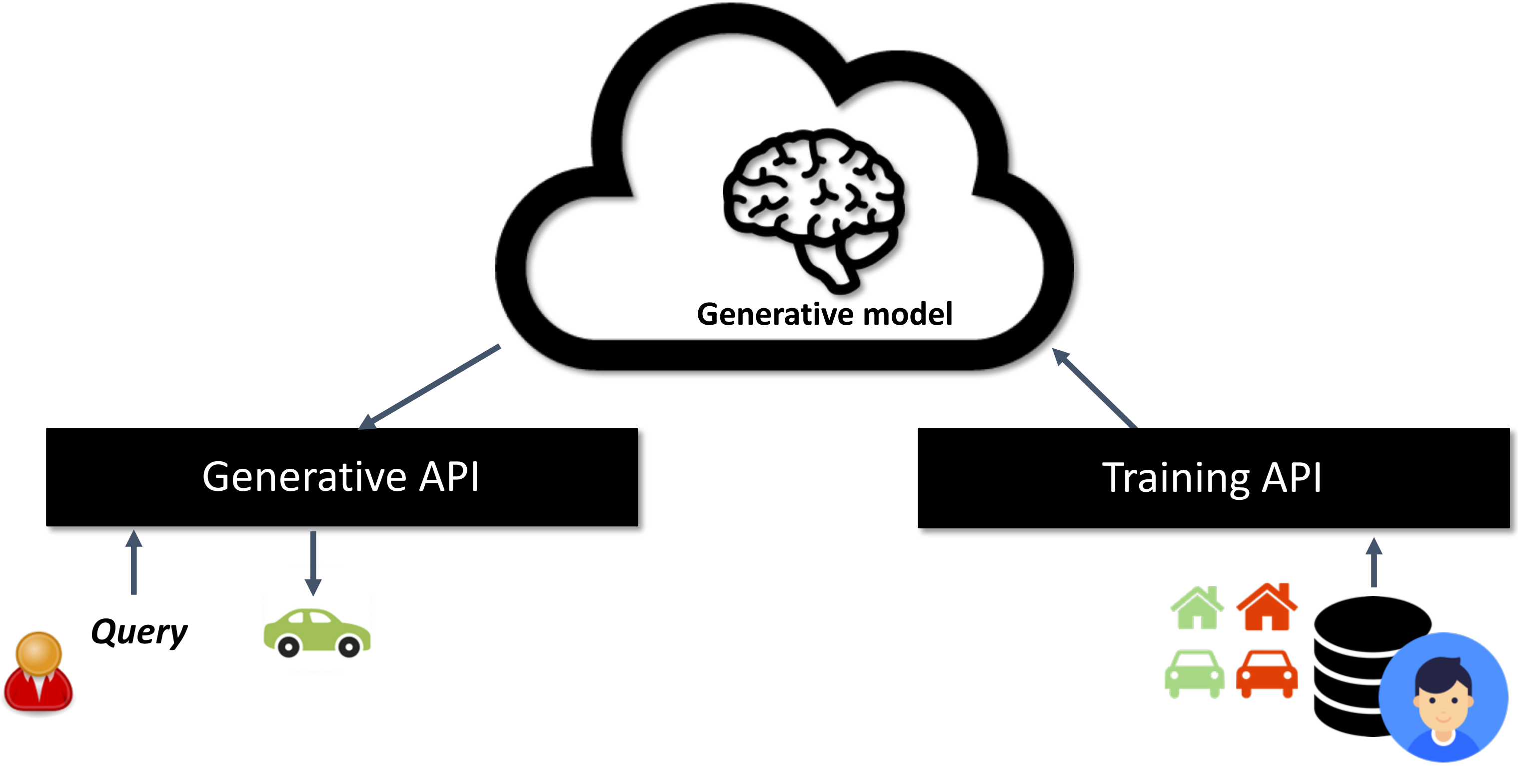
What About Generative Models?

What About Generative Models?

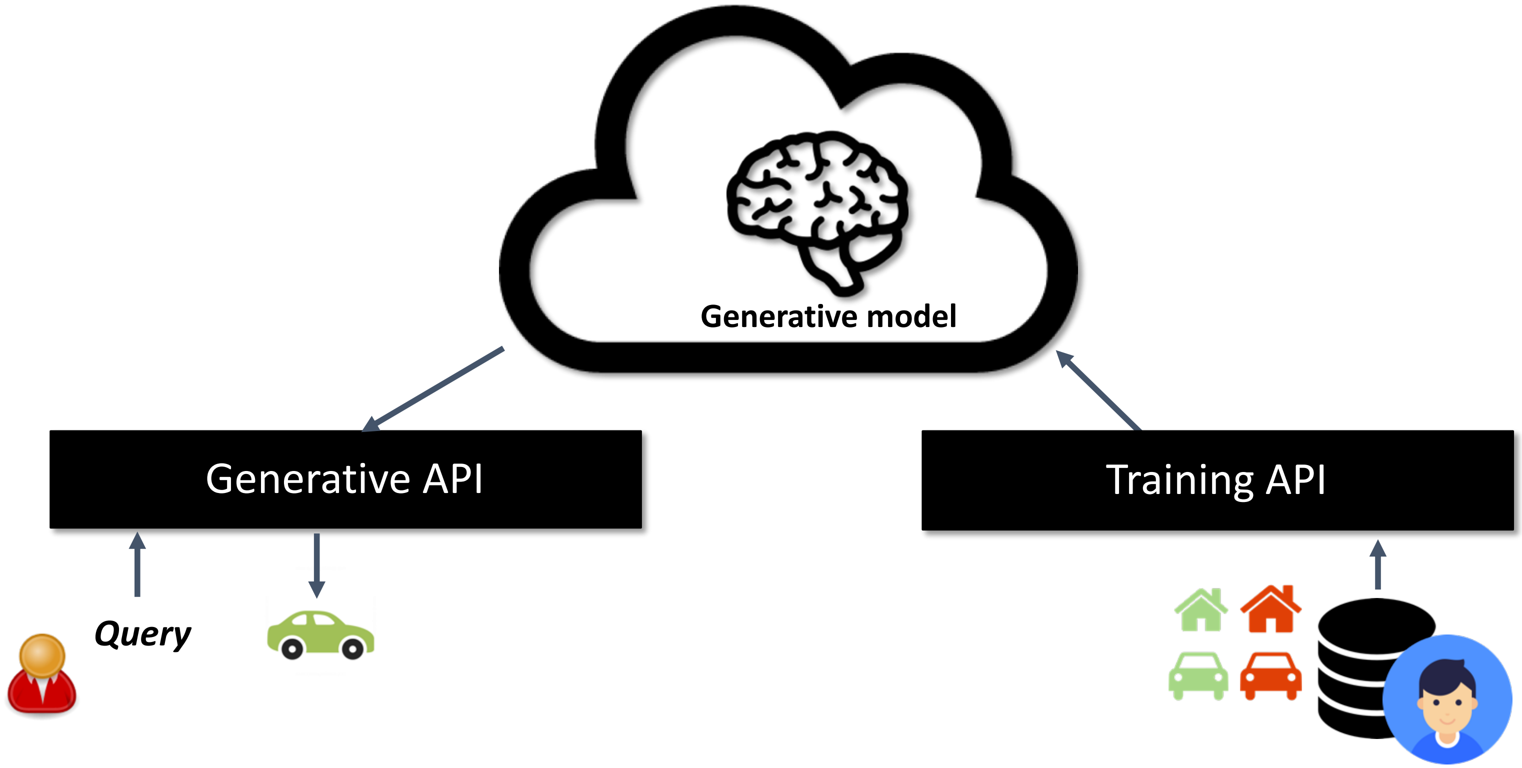


Membership Inference in Generative Models

Membership Inference in Generative Models



Membership Inference in Generative Models



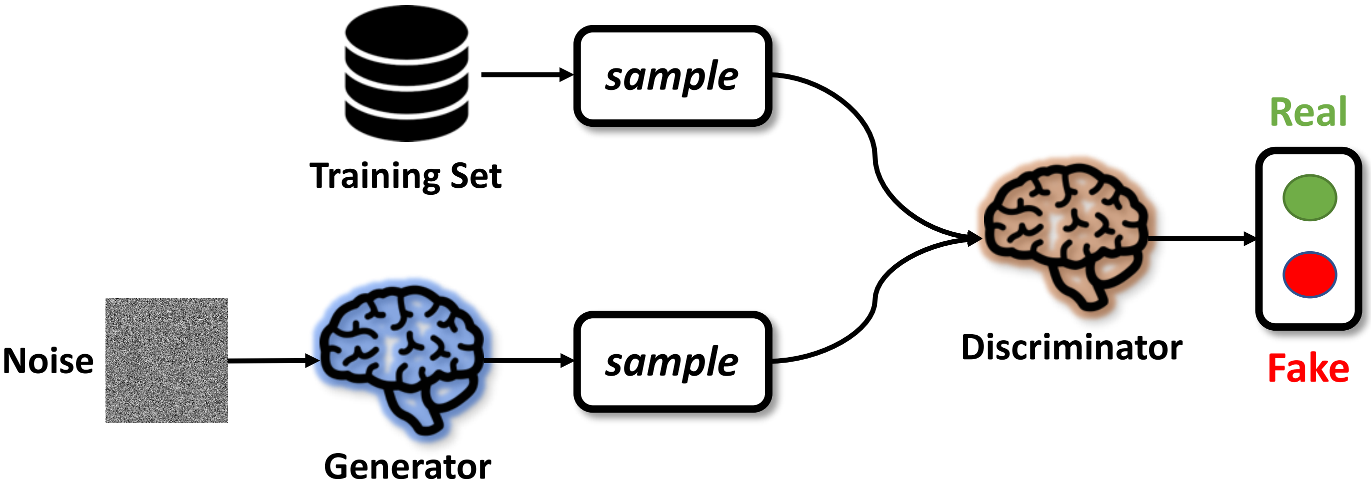
Jamie Hayes, Luca Melis, George Danezis, Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models [PETS 2019]

Inference without predictions?

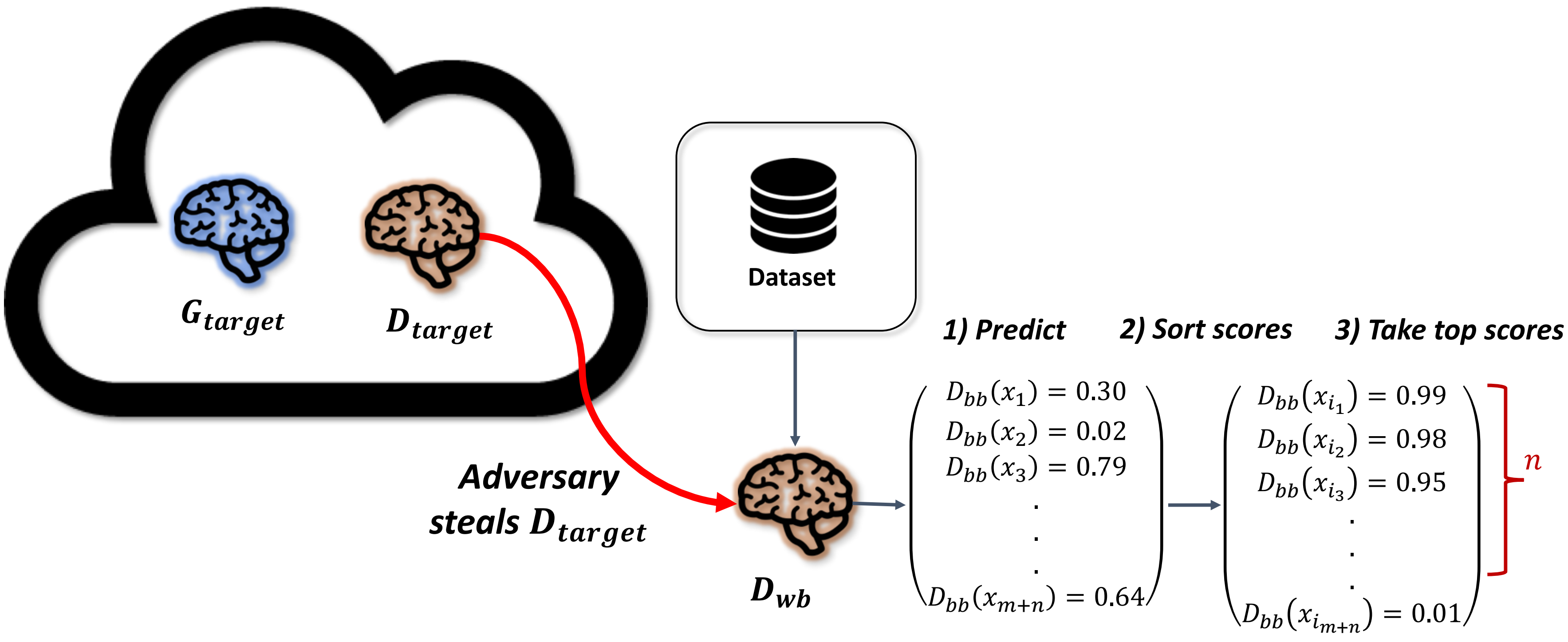
Use generative models! Train GANs to learn the distribution and a prediction model at the same time

Inference without predictions?

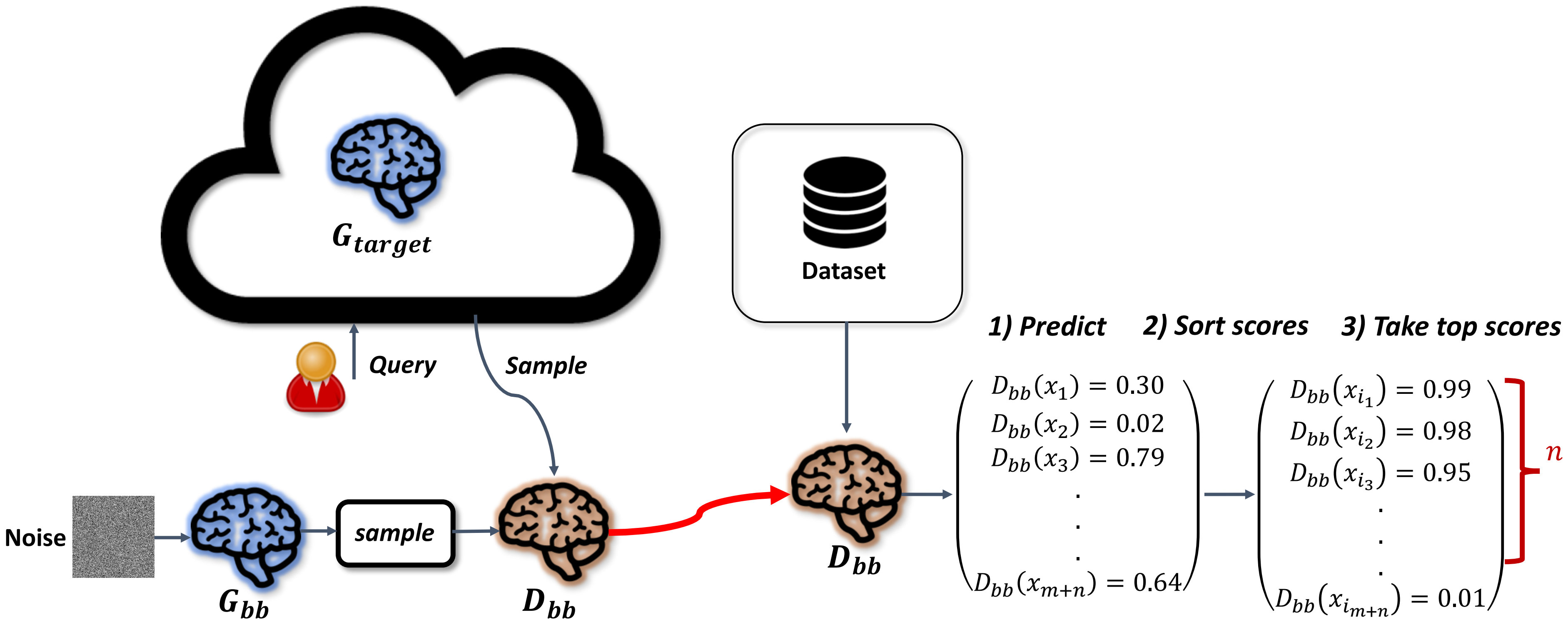
Use generative models! Train GANs to learn the distribution and a prediction model at the same time



White-Box Attack



Black-Box Attack



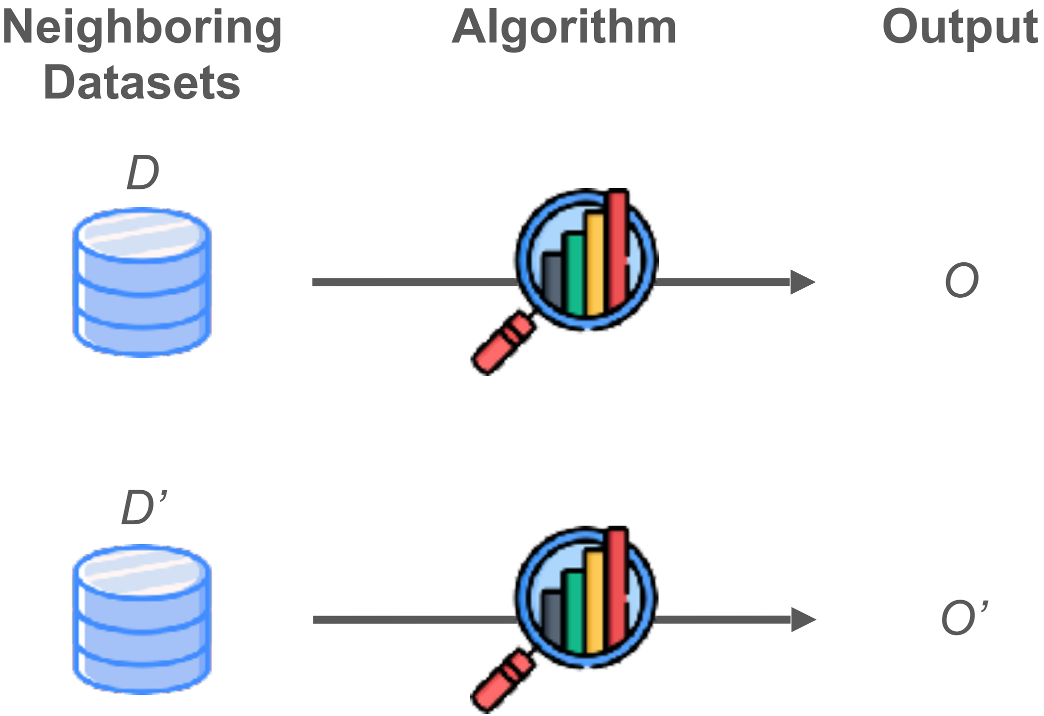
Differential Privacy

Neighboring Datasets

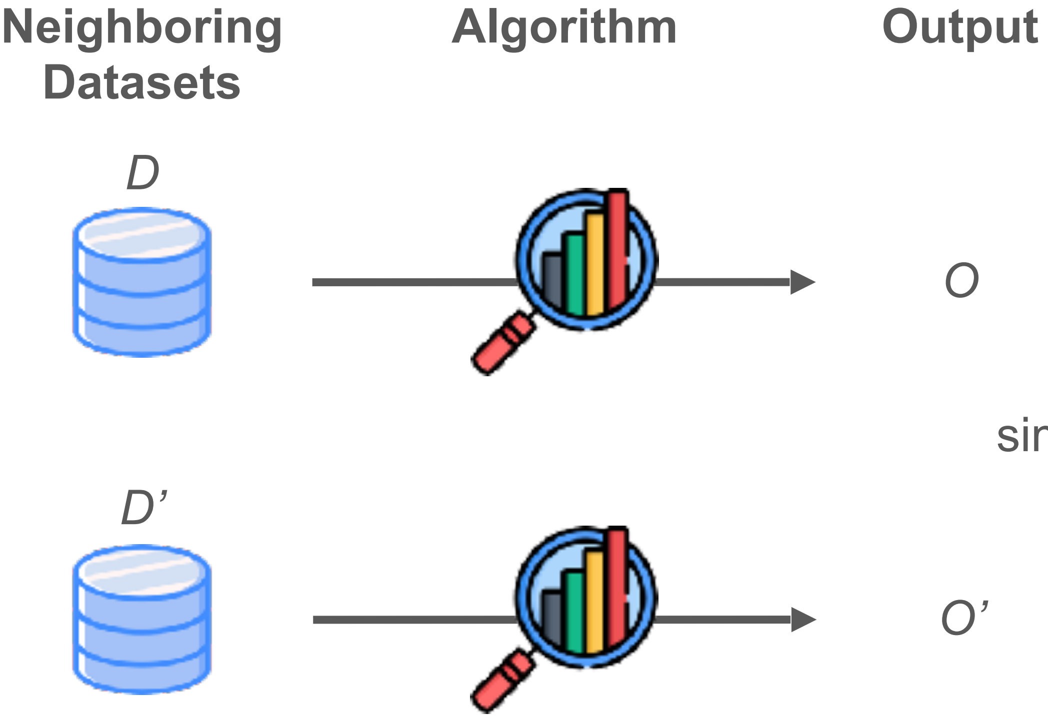
D

D'

Differential Privacy

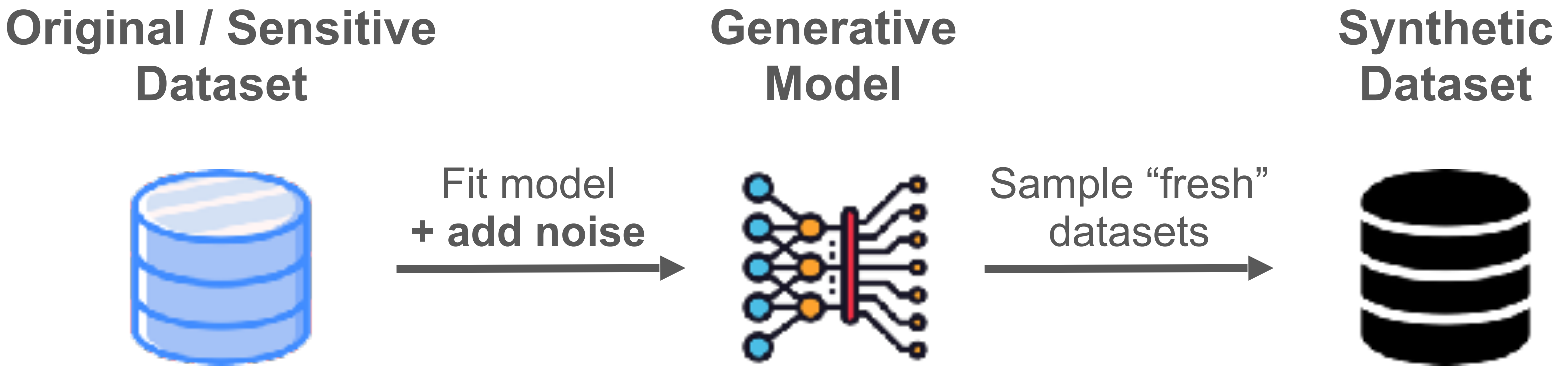


Differential Privacy

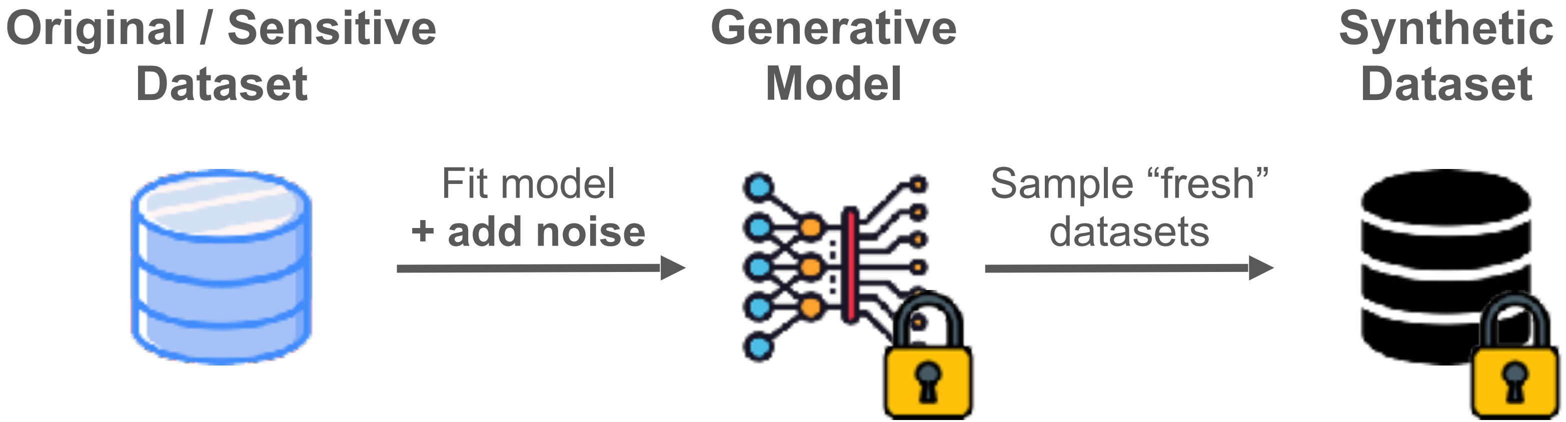


Outputs *O* and *O’* are roughly similar (up to privacy parameter ε), for any input

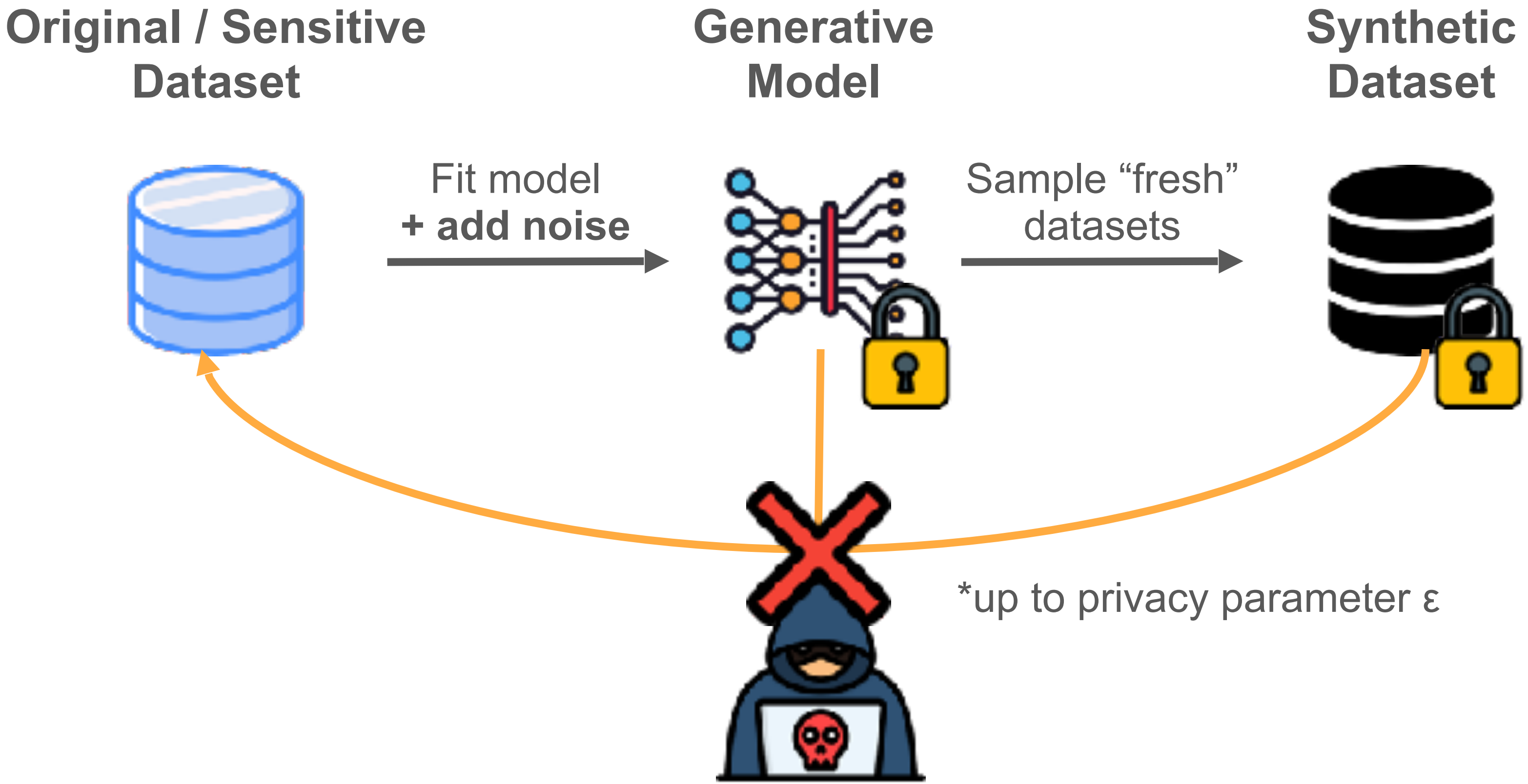
Differentially Private Synthetic Data



Differentially Private Synthetic Data



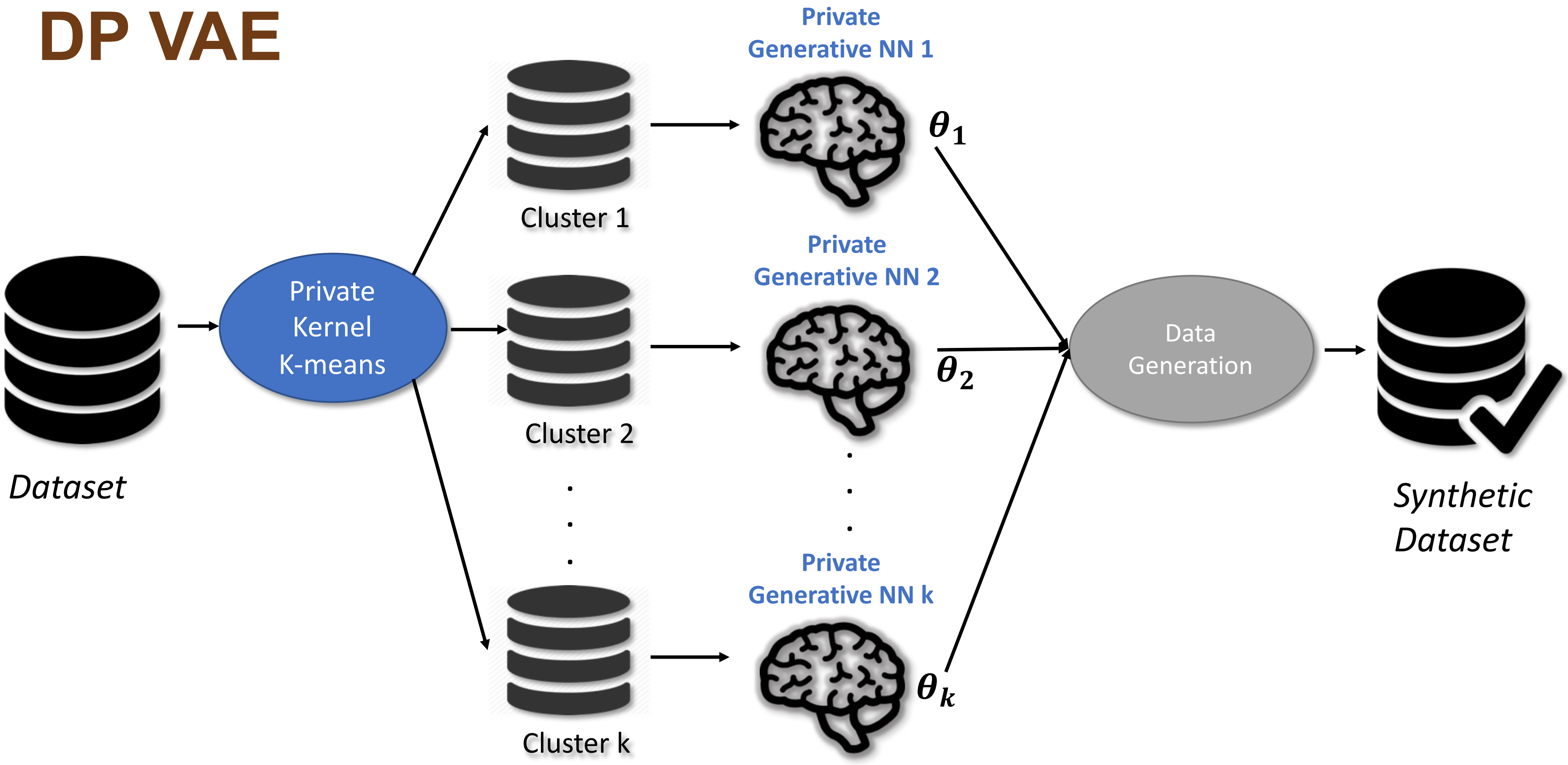
Differentially Private Synthetic Data



DP VAE

Gergely Acs, Luca Melis, Claude Castelluccia, Emiliano De Cristofaro (Extended Version) Differentially Private Mixture of Generative Neural Networks. ICDM 2018

DP VAE



Gergely Acs, Luca Melis, Claude Castelluccia, Emiliano De Cristofaro (Extended Version) Differentially Private Mixture of Generative Neural Networks. ICDM 2018

Synthetic Samples (MNIST)

Original samples

RBM samples

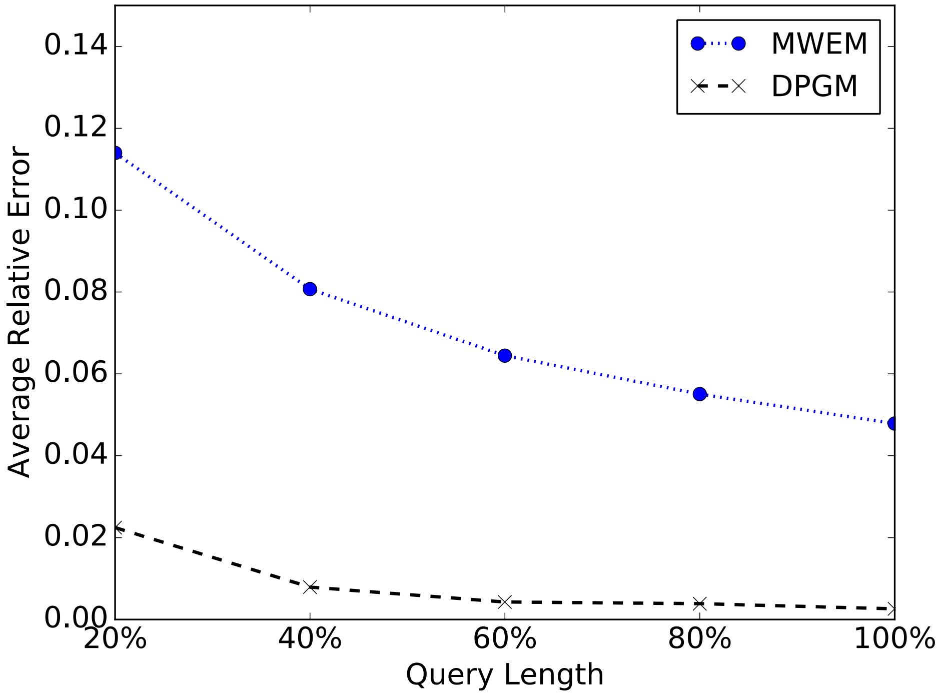
VAE w/o clustering

VAE with clustering

20 SGD epochs (epsilon=1.74)

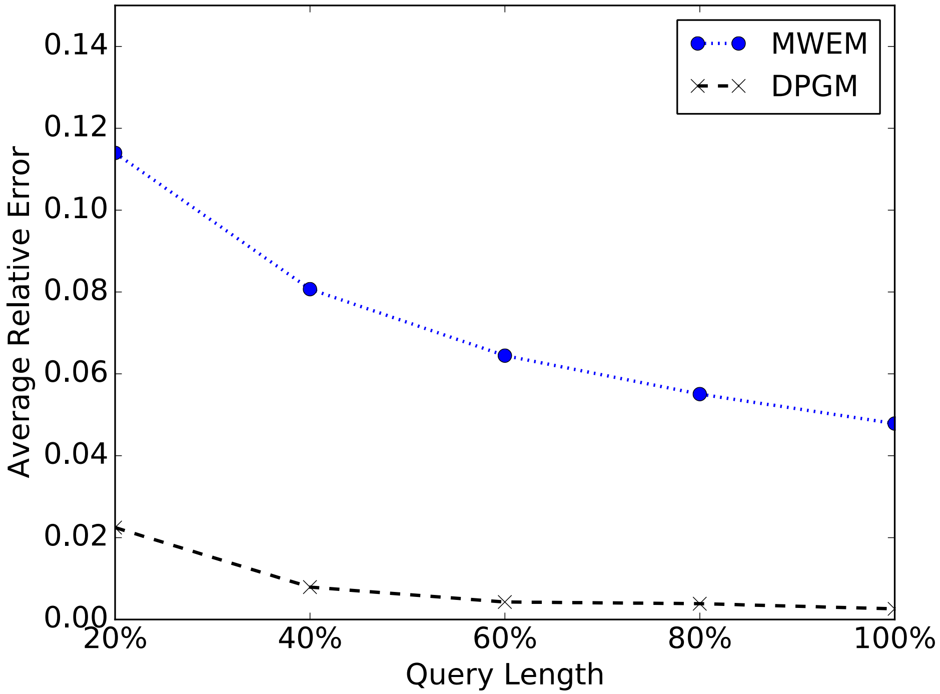
Counting-Query

Counting-Query



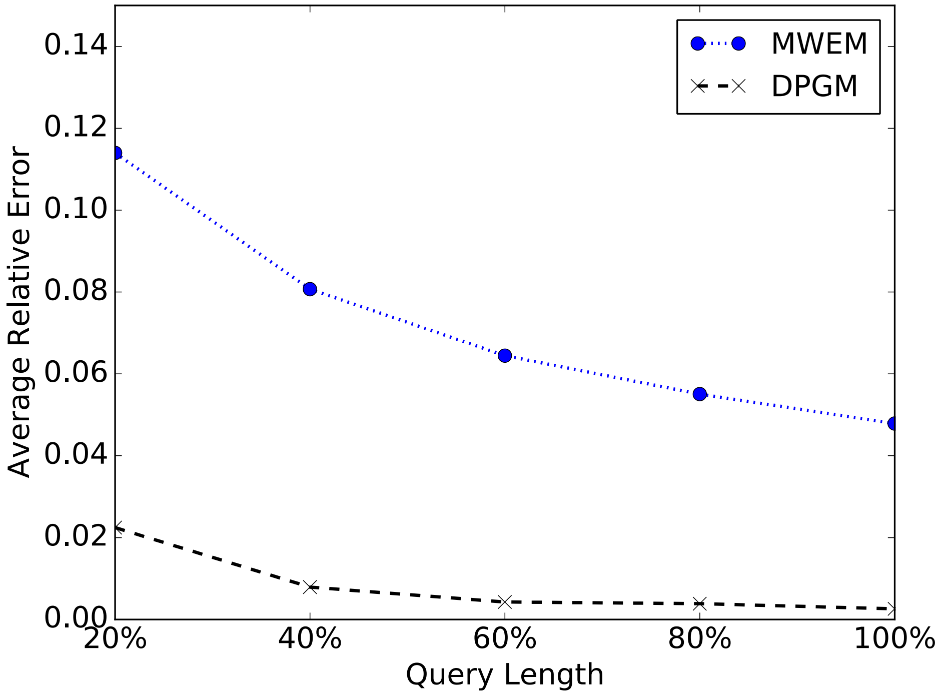
Counting-Query

Task: Given a dataset D, return the number of users in the dataset which satisfy a given predicate



Counting-Query

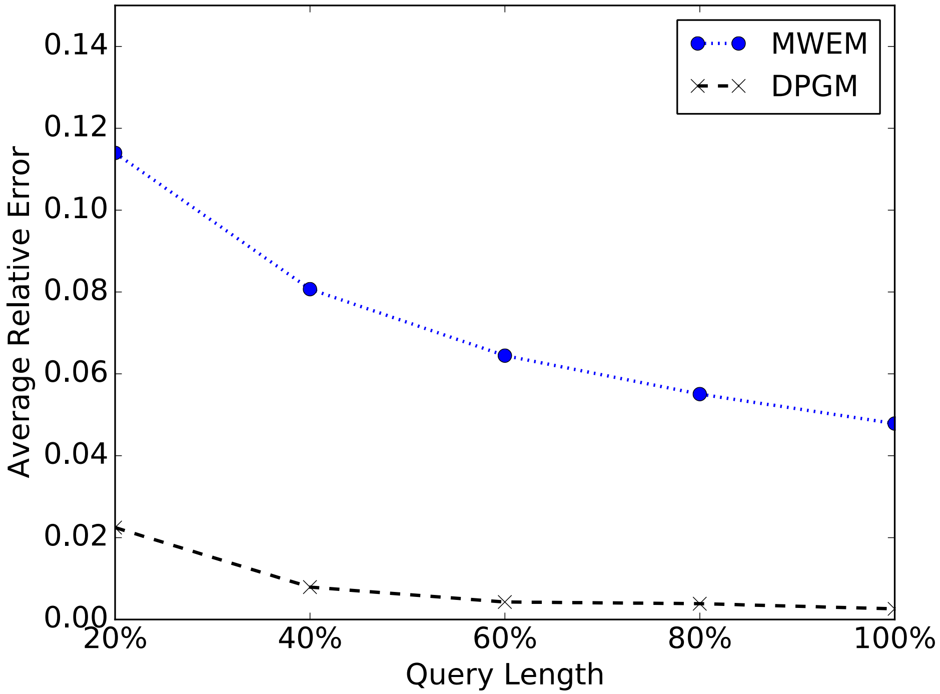
Task: Given a dataset D, return the number of users in the dataset which satisfy a given predicate



Counting-Query

Task: Given a dataset D, return the number of users in the dataset which satisfy a given predicate

Evaluation:

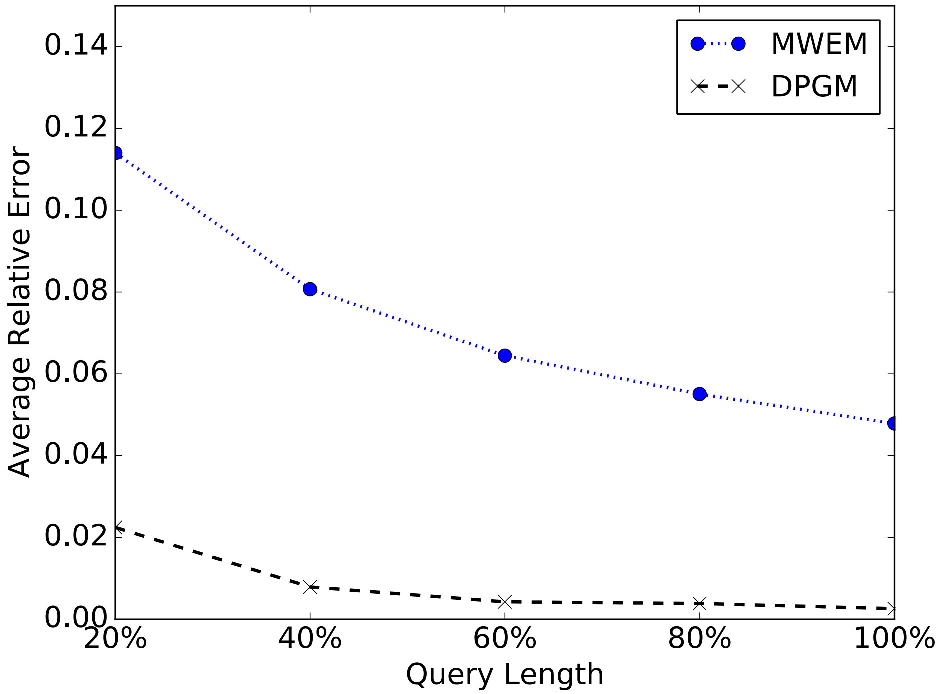


Counting-Query

Task: Given a dataset D, return the number of users in the dataset which satisfy a given predicate

Evaluation:

- Call-Data-Record dataset of tower cells. Query returns the number of users in D who visited a subset of cells.

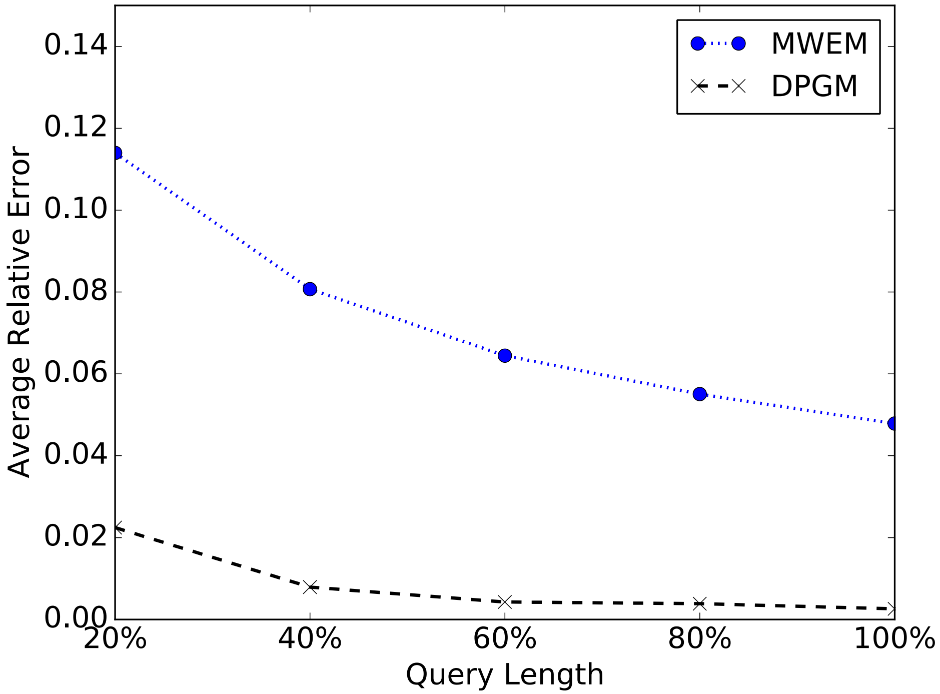


Counting-Query

Task: Given a dataset D, return the number of users in the dataset which satisfy a given predicate

Evaluation:

- Call-Data-Record dataset of tower cells. Query returns the number of users in D who visited a subset of cells.
- Dataset: approx. 4 million users, 1303 number of towers



(DP) Synthetic Tabular Data

(DP) Synthetic Tabular Data

‹ Return to Blog Home

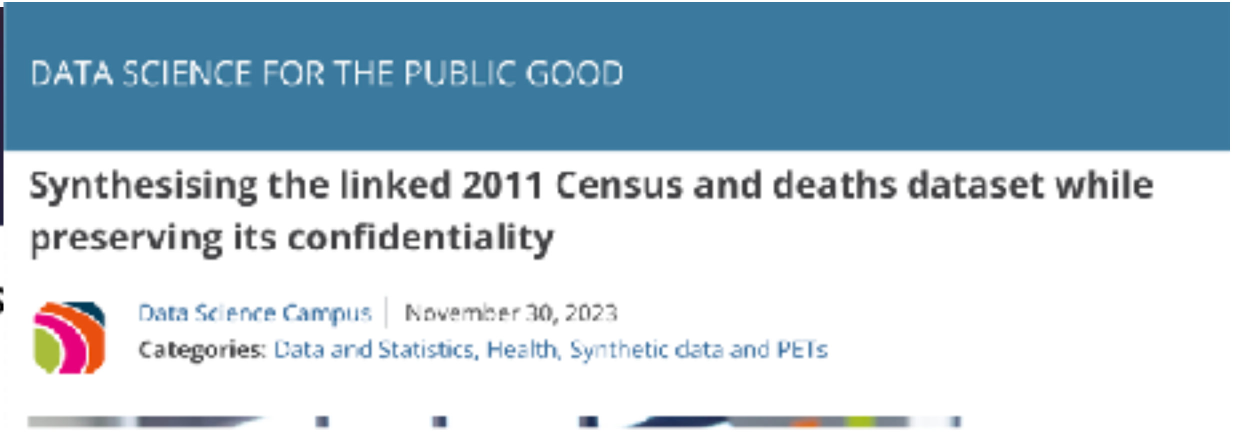
Microsoft Research Blog

IOM and Microsoft release first-ever differentially private synthetic dataset to counter human trafficking

Published December 8, 2022

Share this page

(DP) Synthetic Tabular Data



(DP) Synthetic Tabular Data

‹ Return to Blog Home

Microsoft Research Blog

IOM and Microsoft release firs differentially private synthetic counter human trafficking

Published December 8, 2022

Share this page

DATA SCIENCE FOR THE PUBLIC GOOD

Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality

Data Sci

Categor

Differentially Private Release of Israel's National Registry of Live Births

Shlomi Hod* Ran Canetti*

May 2, 2024

(DP) Synthetic Tabular Data

‹ Return to Blog Hom

Microsoft Rese

IOM and Mi
differentially
counter hum

Published December 8,

Share this page

Synthetic Data in Health Overview

Updated December 2021

The analytics team in NHSX is currently conducting research into best practice and examples for generating synthetic healthcare data for the purpose of enabling greater data sharing across the system. This work will progress through our PhD internship scheme as well as through collaborating and commissioning small proof of concepts to create shareable tools and guidance. Our aim is to make this work open through our github account and the NHSX website.

This thought stream is focussed on the application of synthetic data in healthcare and targeted at analysts in the NHS considering if, and how to implement a synthetic data generation tool.

There are many articles online introducing synthetic data which should be researched for wider context first. One really good general introduction to synthetic is the ONS methodology working paper series number 16 - Synthetic data pilot. I would also recommend spending some time looking through the resources and examples on the synthetic data vault project.

s and deaths dataset while

Private Release of
Registry of Live Births

Ran Canetti*

y 2, 2024

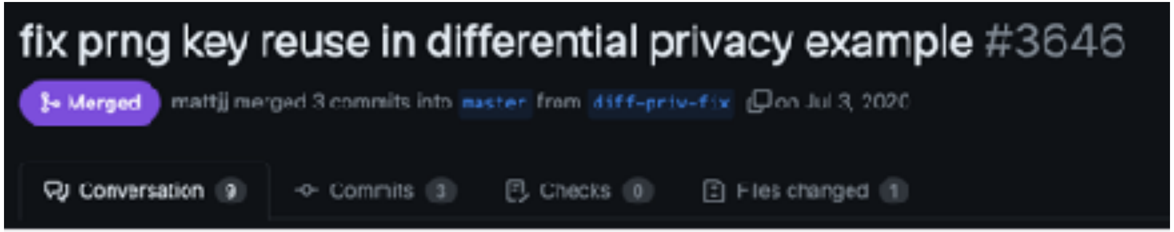
Algorithms & Implementations

Algorithm	**Implementation (Library / Company)**
PrivBayes	DataSynthesizer
	Hazy
MST	NIST
	Microsoft Smartnoise
DPWGAN	NIST
	Synthcity

Do DP implementations satisfy DP in practice?

Do DP implementations satisfy DP in practice?

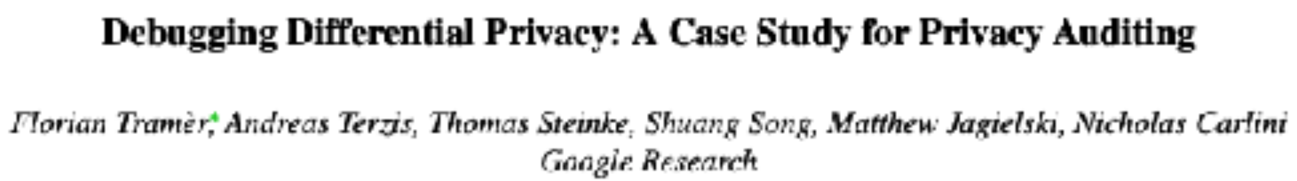
Reuse random seed



Do DP implementations satisfy DP in practice?

Reuse random seed

Incorrect privacy analysis


Debugging Differential Privacy: A Case Study for Privacy Auditing

Florian Tramèr, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, Nicholas Carlini*
Google Research

Do DP implementations satisfy DP in practice?

Reuse random seed

fix prng key reuse in differential privacy example #3646

Merged mattjj merged 3 commits into master from diff-priv-fix on Jul 3, 2020

Conversa

Incorrect privacy analysis

Debugging Differential Privacy: A Case Study for Privacy Auditing

Florian

Floating point violation

Group and Attack: Auditing Differential Privacy

Johan Lokna
jlokna@student.ethz.ch
ETH Zurich
Switzerland

Anouk Paradis
anouk.paradis@inf.ethz.ch
ETH Zurich
Switzerland

Dimitar I. Dimitrov
dimitar.iliev.dimitrov@inf.ethz.ch
ETH Zurich
Switzerland

Martin Vechev
martin.vechev@inf.ethz.ch
ETH Zurich
Switzerland

DP Auditing

DP Auditing

Verify that the empirical privacy leakage (ε_{emp}) matches the theoretical upper bounds guaranteed by DP (ε)

Auditing Procedure

Step 1: Choose neighboring datasets

D

D'

Auditing Procedure

Step 1: Choose neighboring datasets

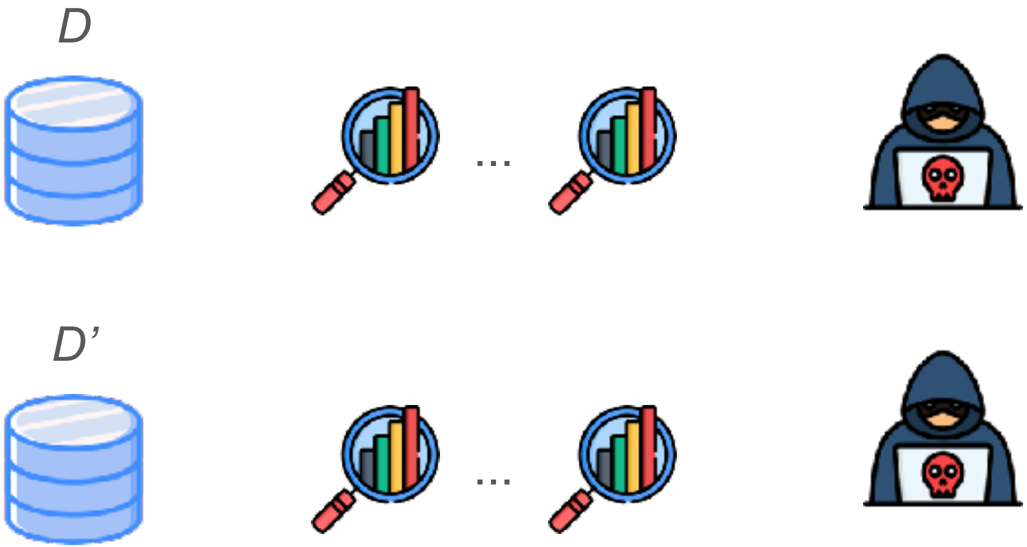
Step 2: Run algorithm repeatedly

Auditing Procedure

Step 1: Choose neighboring datasets

Step 2: Run algorithm repeatedly

Step 3: Run Membership Inference Attack (MIA)



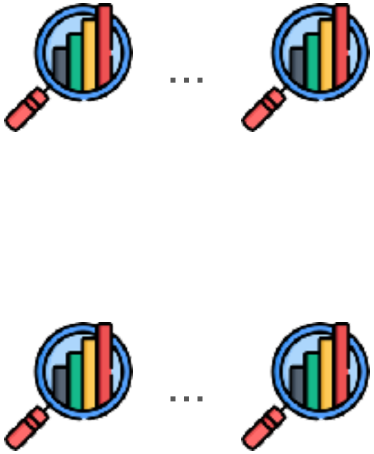
Auditing Procedure

Step 1: Choose neighboring datasets

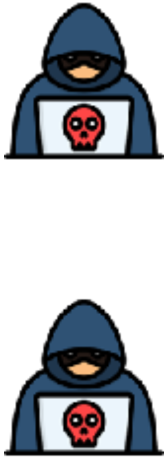
D

D'

Step 2: Run algorithm repeatedly



Step 3: Run Membership Inference Attack (MIA)



Step 4: Convert FPR and FNR to empirical ε_{emp}

DP Auditing

DP Auditing

Verify that the empirical privacy leakage ε_{emp} matches the theoretical upper bounds guaranteed by DP ε

DP Auditing

Verify that the empirical privacy leakage ε_{emp} matches the theoretical upper bounds guaranteed by DP ε

DP Auditing

Verify that the empirical privacy leakage $\boldsymbol{\varepsilon_{emp}}$ matches the theoretical upper bounds guaranteed by DP $\boldsymbol{\varepsilon}$

- If $\varepsilon_{emp} \gg \varepsilon$, then **privacy violations**

DP Auditing

Verify that the empirical privacy leakage $\boldsymbol{\varepsilon_{emp}}$ matches the theoretical upper bounds guaranteed by DP $\boldsymbol{\varepsilon}$

- If $\varepsilon_{emp} \gg \varepsilon$, then **privacy violations**
- If $\varepsilon_{emp} \approx \varepsilon$, then audit is **tight**

DP Auditing

Verify that the empirical privacy leakage $\boldsymbol{\varepsilon_{emp}}$ matches the theoretical upper bounds guaranteed by DP $\boldsymbol{\varepsilon}$

- If $\varepsilon_{emp} \gg \varepsilon$, then **privacy violations**
- If $\varepsilon_{emp} \approx \varepsilon$, then audit is **tight**
- If $\varepsilon_{emp} \ll \varepsilon$, then audit is **loose**

Your audit is loose, so what?

Your audit is loose, so what?

- The auditing procedure can be improved

Your audit is loose, so what?

- The auditing procedure can be improved
- MIA instantiation is not particularly effective

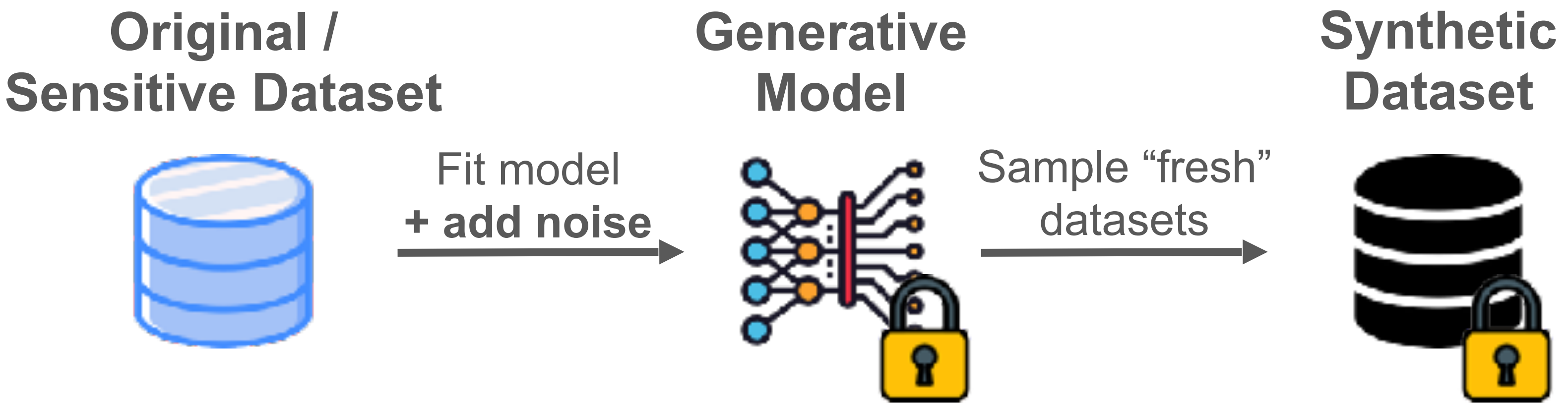
Your audit is loose, so what?

- The auditing procedure can be improved
- MIA instantiation is not particularly effective
- Bounds from theoretical analysis are too conservative

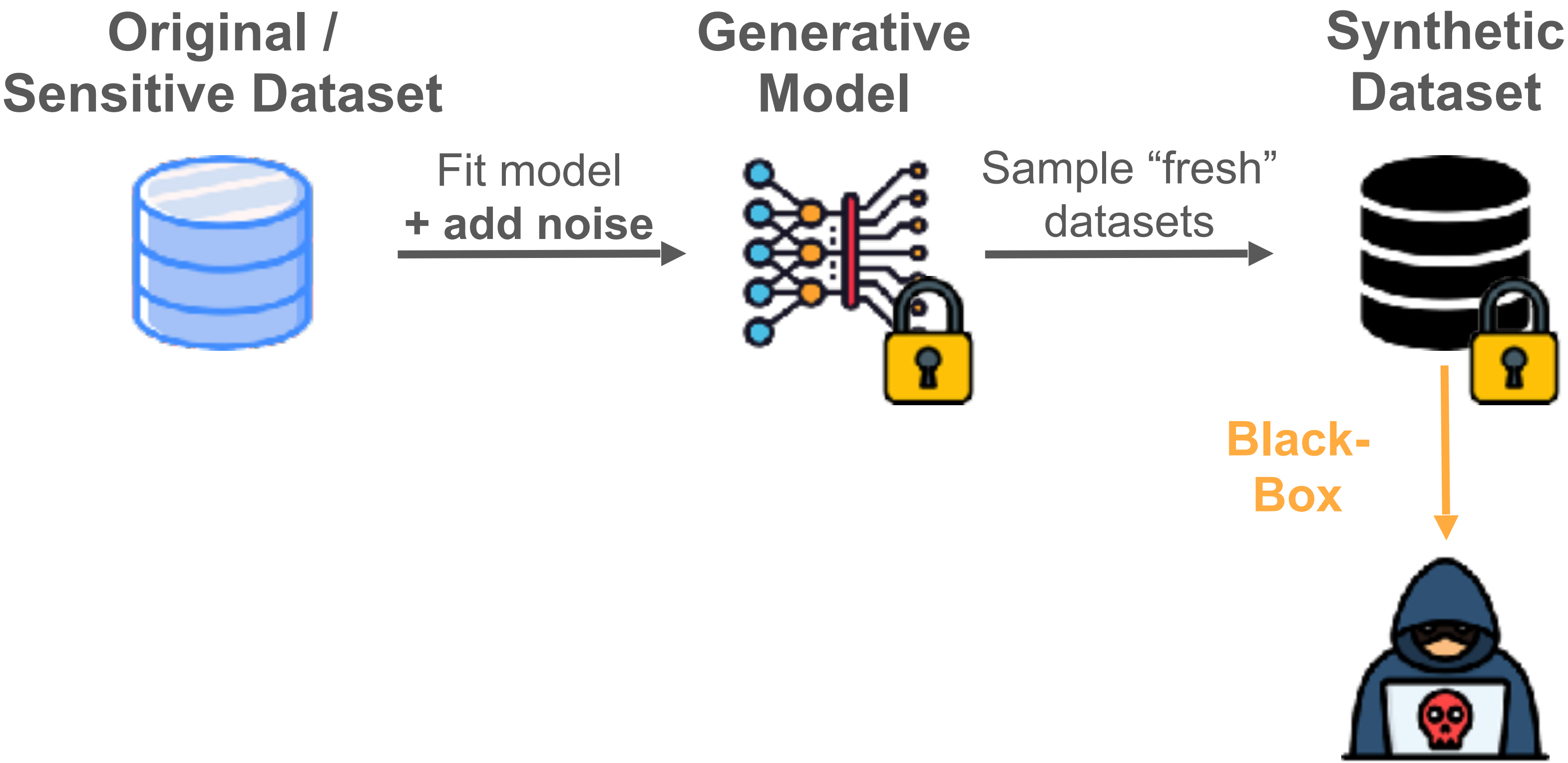
Your audit is loose, so what?

- The auditing procedure can be improved
- MIA instantiation is not particularly effective
- Bounds from theoretical analysis are too conservative
- Etc.

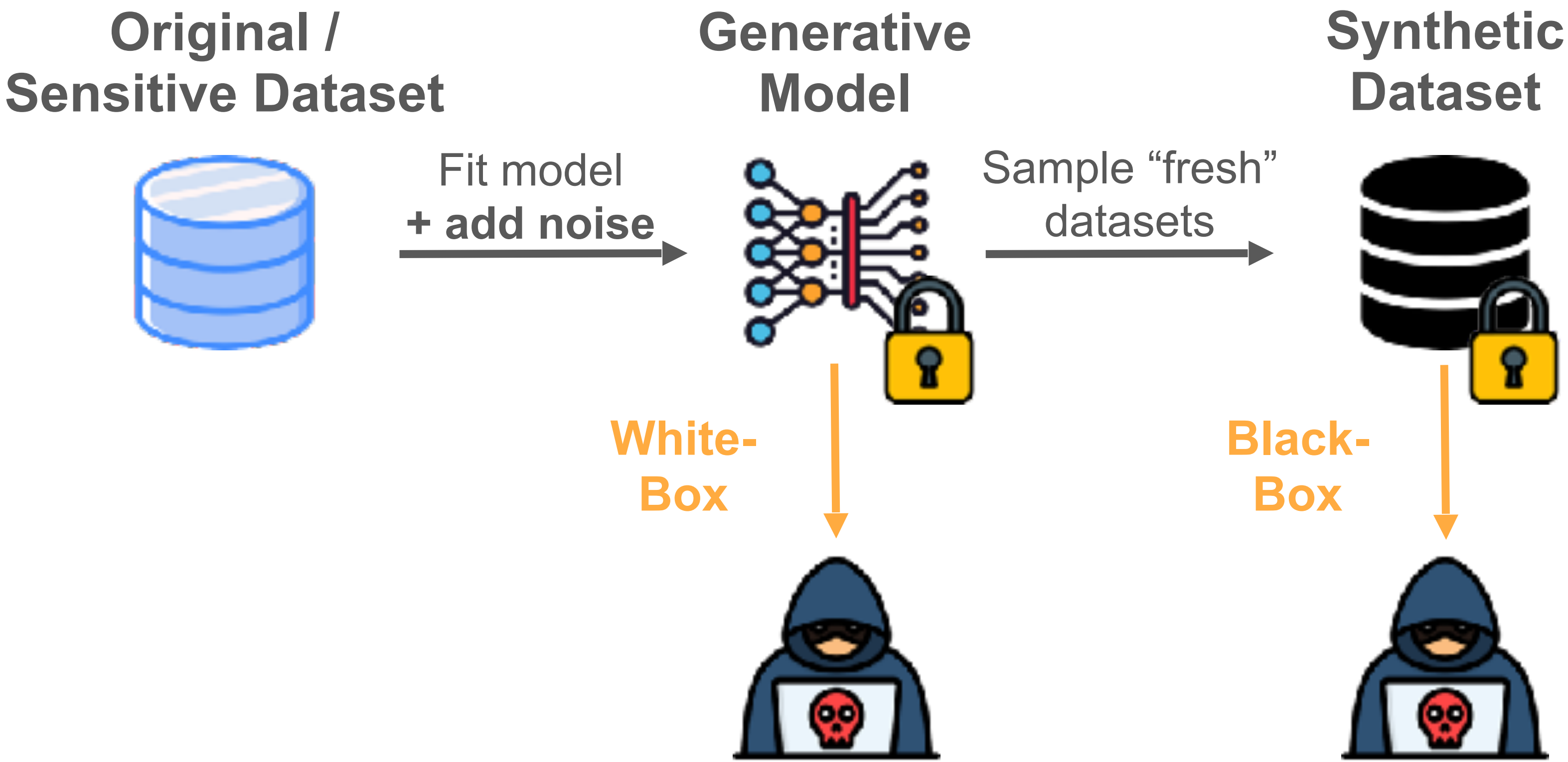
Black-Box vs White-Box Auditing



Black-Box vs White-Box Auditing



Black-Box vs White-Box Auditing



Black-Box Attacks

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**
  - Output distance between target record and closest synthetic record

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**
  - Output distance between target record and closest synthetic record
  - If target record is memorized, synthetic records will be closer to it

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**
  - Output distance between target record and closest synthetic record
  - If target record is memorized, synthetic records will be closer to it
- **Querybased**

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**
  - Output distance between target record and closest synthetic record
  - If target record is memorized, synthetic records will be closer to it
- **Querybased**
  - Train ML model on output of queries targeted at record as features

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Black-Box Attacks

- **Distance to closest record (DCR)**
  - Output distance between target record and closest synthetic record
  - If target record is memorized, synthetic records will be closer to it
- **Querybased**
  - Train ML model on output of queries targeted at record as features
  - Intuition: Shmodelingadow modelling

Houssiau, F., Jordon, J., Cohen S., Elliott A., Geddes J., Mole C., Rangel-Smith C., & Szpruch L. (2022). TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. In SyntheticData4ML Workshop NeurIPS.

Prior Work

Prior Work

- Mostly **Black-Box** attacks

Prior Work

- Mostly **Black-Box** attacks
- On **Average-Case** neighboring datasets

Prior Work

- Mostly **Black-Box** attacks
- On **Average-Case** neighboring datasets

➔ **Loose** empirical privacy leakage estimates ($\varepsilon_{emp} \ll \varepsilon$)

➔ **Limited effectiveness** in finding bugs and privacy violations

Open Research Questions

Open Research Questions

1. How **tightly** can we empirically estimate leakage from (DP) synthetic data?

Open Research Questions

1. How **tightly** can we empirically estimate leakage from (DP) synthetic data?

2. How do different **threat models/datasets** affect tightness?

Technical Roadmap

Technical Roadmap

- **Large-scale** audit of DP-SDG algorithms and implementations

Technical Roadmap

- **Large-scale** audit of DP-SDG algorithms and implementations
- New **white-box** attacks against PrivBayes and MST

Technical Roadmap

- **Large-scale** audit of DP-SDG algorithms and implementations
- New **white-box** attacks against PrivBayes and MST
- Implementation-specific **worst-case** datasets

Main Findings

Main Findings

1. Black-box attacks are **ineffective** in exploiting privacy leakage

Main Findings

1. Black-box attacks are **ineffective** in exploiting privacy leakage

2. You need **white-box** attacks + **worst-case** neighboring datasets to achieve tightness

Main Findings

1. Black-box attacks are **ineffective** in exploiting privacy leakage

2. You need **white-box** attacks + **worst-case** neighboring datasets to achieve tightness

3. DP **violations** found in **5 out of 6** implementations tested
   - Even in implementations successfully submitted to the NIST DP Synthetic Data Challenge competition

Beyond Auditing: Utility vs Privacy

Beyond Auditing: Utility vs Privacy

Finding the best models for specific settings/tasks is often (very) challenging

Beyond Auditing: Utility vs Privacy

Finding the best models for specific settings/tasks is often (very) challenging

This often entails navigating privacy-utility tradeoffs, which is hard

Beyond Auditing: Utility vs Privacy

Finding the best models for specific settings/tasks is often (very) challenging

This often entails navigating privacy-utility tradeoffs, which is hard

Can we understand how different models spend their **privacy budget** across **rows** and **columns**?

Beyond Auditing: Utility vs Privacy

Finding the best models for specific settings/tasks is often (very) challenging

This often entails navigating privacy-utility tradeoffs, which is hard

Can we understand how different models spend their **privacy budget** across **rows** and **columns**?

(one of the main sources of utility degradation)

Do DP generative models distribute their privacy budget in a similar way?

Do DP generative models distribute their privacy budget in a similar way?

No. The graphical models distribute their privacy budget horizontally and the GANs vertically (i.e., they spend their budget per iteration).

What are the effects of DP/dataset dimensions on downstream tasks?

What are the effects of DP/dataset dimensions on downstream tasks?

- The effects are mixed. Overall, more training data helps the graphical models with some exceptions.

What are the effects of DP/dataset dimensions on downstream tasks?

- The effects are mixed. Overall, more training data helps the graphical models with some exceptions.

- Varying the dataset dimensions is more unpredictable (more variable and usually not monotonic) for the GAN models.

Other work in this space

Robin Hood and *Matthew* Effects: Differential Privacy Has Disparate Impact on Synthetic Data

Georgi Ganev [1,2] Bristena Oprisanu [1] Emiliano De Cristofaro [1]

Robin Hood and _Matthew_ Effects: Differential Privacy Has Disparate Impact on Synthetic Data

On Utility and Privacy in Synthetic Genomic Data*

Bristena Oprisanu
UCL
bristena.oprisanu.10@ucl.ac.uk

Georgi Ganev
UCL and Hazy
georgi.ganev.16@ucl.ac.uk

Emiliano De Cristofaro
UCL and Alan Turing Institute
e.decristofaro@ucl.ac.uk

Robin Hood and *Matthew* Effects: Differential Privacy Has Disparate Impact on Synthetic Data

On Utility and Privacy in Synthetic Genomic Data*

Bristena Oprisanu
UCL
bristena.oprisanu.10@ucl.ac.uk

Georgi Ganev
UCL and Hazy
georgi.ganev.16@ucl.ac.uk

Emiliano De Cristofaro
UCL and Alan Turing Institute
e.decristofaro@ucl.ac.uk

The Elusive Pursuit of Replicating PATE-GAN: Benchmarking, Auditing, Debugging

Georgi Ganev[1,2], Meenatchi Sundaram Muthu Selva Annamalai[1], Emiliano De Cristofaro[3]

[1]University College London [2]Hazy [3]UC Riverside

Robin Hood and _Matthew_ Effects: Differential Privacy Has Disparate Impact on Synthetic Data

On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against "Truly Anonymous Synthetic Data"

1

Georgi Ganev[1,2] and Emiliano De Cristofaro[3]
[1]University College London [2]Hazy [3]UC Riverside

The Elusive Pursuit of Replicating PATE-GAN: Benchmarking, Auditing, Debugging

Georgi Ganev[1,2], Meenatchi Sundaram Muthu Selva Annamalai[1], Emiliano De Cristofaro[3]
[1]University College London [2]Hazy [3]UC Riverside

Thank you!