Measuring Utility and Privacy of Synthetic Genomic Data

Emiliano De Cristofaro
https://emilianodc.com

Agenda

Privacy in Machine Learning

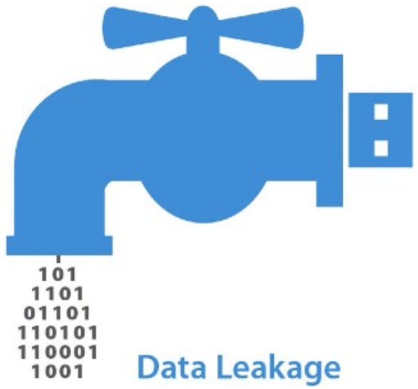
Synthetic Data

Privacy (and Utility) in Synthetic Genomic Data

Agenda

Privacy in Machine Learning

Reasoning about "privacy" in ML



Most privacy attacks in ML focus on inferring:

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")
3. Properties/Attributes of the training data other than the main task (aka "property inference")

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…

- Given f(data), infer if $x \in$ 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

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

3. Property Inference

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

Privacy in Machine Learning

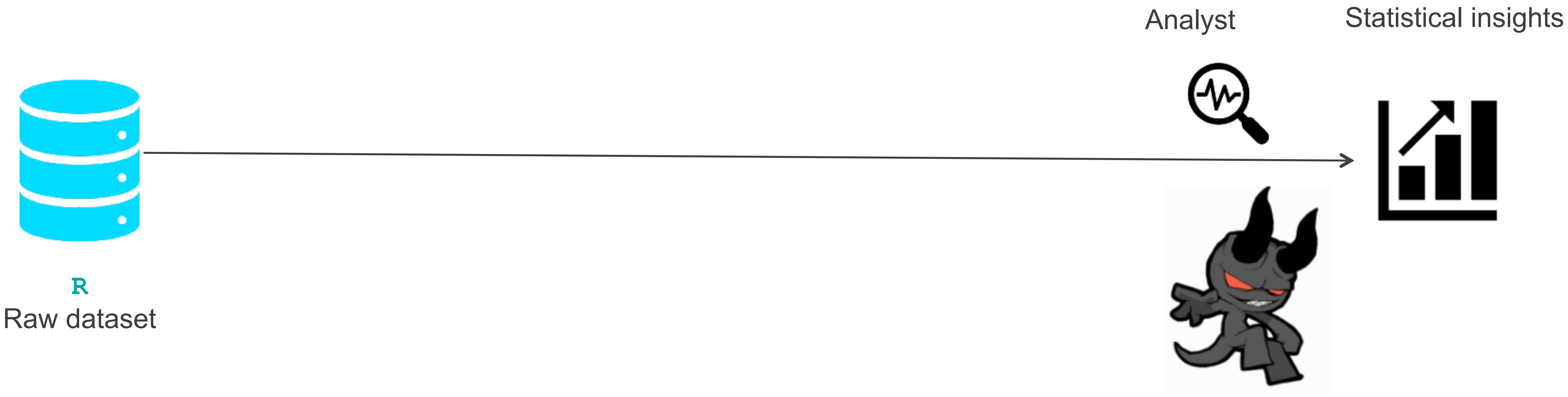
Synthetic Data

Privacy (and Utility) in Synthetic Genomic Data

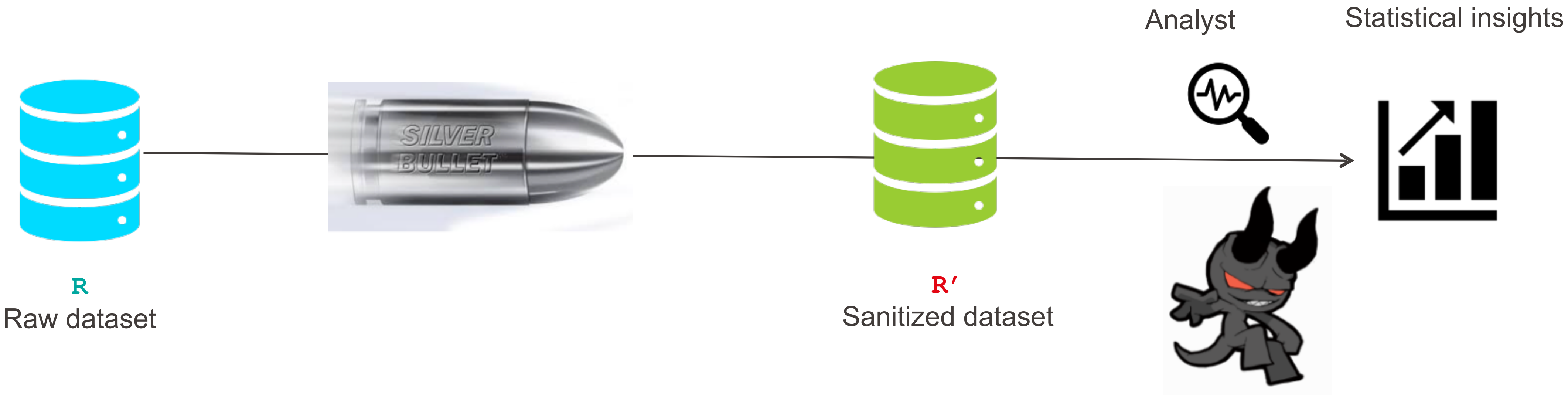
Agenda

Synthetic Data

Data Sharing



Data Sharing



Data Release Disasters

Data Release Disasters

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006

Comment

Yet Another Update: AOL: “This was a screw up”

Data Release Disasters

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006

Comment

Yet Another Update: AOL: “This was a screw up”

Data Release Disasters

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006

Comment

Yet Another Update: AOL: "This was a screw up"

Study finds HIPAA protected data still at risks

Data Release Disasters

The New York Times

Netflix Cancels Contest After Concerns Are Raised About Privacy

By Steve Lohr

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006

Comment

Yet Another Update: AOL: "This was a screw up"

Study finds HIPAA protected data still at risks

Data Release Disasters

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006

Comment

Yet Another Update: AOL: “This was a screw up”

Study finds HIPAA protected data still at risks

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income

The New York Times

Netflix Cancels Contest After Concerns Are Raised About Privacy

By Steve Lohr

Data Release Disasters

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2006 Comment

Yet Another Update: AOL: "This was a screw up"

Study finds HIPAA protected data still at risks

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income

The New York Times

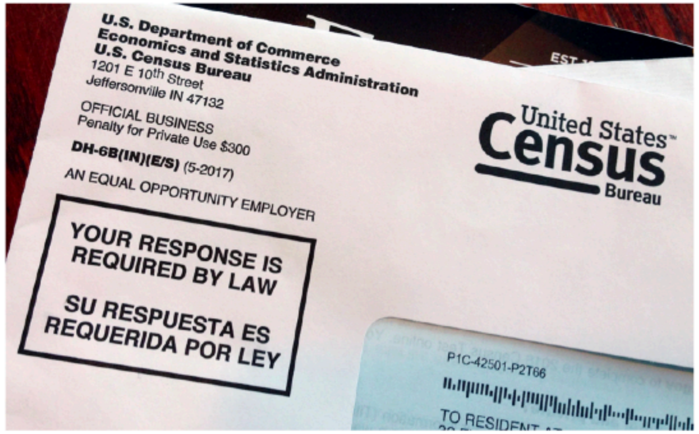
Netflix Cancels Contest After Concerns Are Raised About Privacy

By Steve Lohr

TheUpshot

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.



A 2018 census test letter mailed to a resident in Providence, R.I. The nation's test run of the 2020 Census is in Rhode Island. Michelle R. Smith/Associated Press

Data Release Disasters

AOL Proudly Releases M
of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2(

Yet Another Update: AOL: "This was a screw up

Study finds HIPAA protected data still at risks

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

▲ In practice, supposedly anonymised data can be deanonymised in a number of ways to identify real people. Photograph: Stefan Rousseau/PA

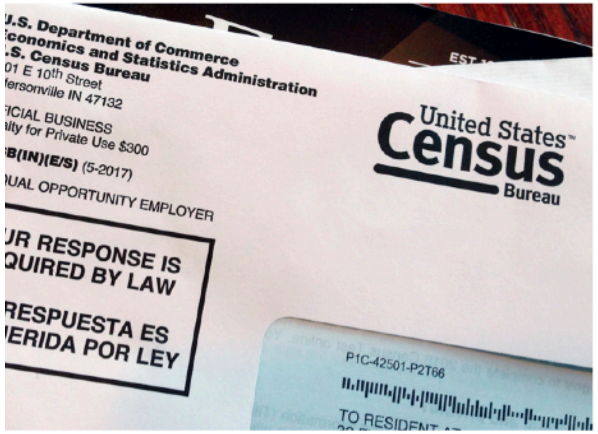
The New York Times

N... C...ncels Contest After Concerns About Privacy

Upshot

Reduce Privacy Risks, the Census Plans Report Less Accurate Data

anteeing people's confidentiality has become more of a enge, but some scholars worry that the new system will de research.



st letter mailed to a resident in Providence, R.I. The nation's test run of the 2020 Census d. Michelle R. Smith/Associated Press

Data Release Disasters

The New York Times

AOL Proudly Releases M
of Private Data

Michael Arrington @arrington?lang=en / 2:17 AM GMT+1 • August 7, 2(

Yet Another Update: AOL: "This was a screw up

'Anonymised' data can never be totally anonymous, says study

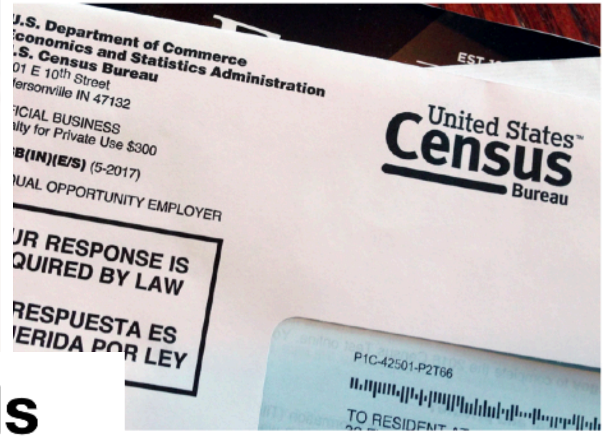
Findings say it is impossible for researchers to fully protect real identities in datasets

Netflix Cancels Contest After Concerns About Privacy

Upshot

Reduce Privacy Risks, the Census Plans Report Less Accurate Data

anteeing people's confidentiality has become more of a enge, but some scholars worry that the new system will de research.



a resident in Providence, R.I. The nation's test run of the 2020 Census
th/Associated Press

Study finds HIPAA protected data still at risks

Researchers Find 'Anonymized' Data Is Even Less Anonymous Than We Thought

Corporations love to pretend that 'anonymization' of the data they collect protects consumers. Studies keep showing that's not really true.

Synthetic Data

Synthetic Data

MOSTLY·AI PRODUCT SOLUTIONS USE CASES BLOG

Democratize your data access with synthetic data!
No more fines.
Guaranteed.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Synthetic Data

MOSTLY•AI PRODUCT ⌄ SOLUTIONS USE CASES ⌄ BLOG

Democratize your data access with synthetic data!
No more fines.
Guaranteed.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Privacy-compliance for data exploration

Statice offers a data anonymization solution. We enable businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

Synthetic Data

Enable cross boundary data analytics

Hazy synthetic data generation lets you create business insight across company, legal and compliance boundaries — without moving or exposing your data.

MOSTLY•AI PRODUCT SOLUTIONS USE CASES BLOG

Democratize your data access with synthetic data!
No more fines.
Guaranteed.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Privacy-compliance for data exploration

Statice offers a data anonymization solution. We enable businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

Synthetic Data

MOSTLY•AI PRODUCT SOLUTIONS USE CASES BLOG

Democratize your data access with synthetic data!
No more fines.
Guaranteed.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Enable cross boundary data analytics

Hazy synthetic data generation lets you create business insight across company, legal and compliance boundaries — without moving or exposing your data.

Innovate with Synthesized

Synthesized data: 10X the impact, 0 risks

Privacy-compliance for data exploration

Statice offers a data anonymization solution. We enable businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

Synthetic Data

MOSTLY·AI PRODUCT SOLUTIONS USE CASES BLOG

Democratize your data access with synthetic data!
No more fines.
Guaranteed.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Privacy-compliance for data exploration

Statice offers a data anonymization solution. We enable businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

Enable cross boundary data analytics

Hazy synthetic data generation lets you create business insight across company, legal and compliance boundaries — without moving or exposing your data.

Innovate with Synthesized

Synthesized data: 10X the impact, 0 risks

Synthesized solves the problem of data sharing

Instead of sharing original data, we enable businesses and other data owners to work with compliant synthetic datasets mimicking the structure of original data without disclosing any information about individual data points.

Synthetic Data

Enable cross boundary data analytics

usiness insight across
ithout moving or exposing

MOSTLY·AI PRODUCT ⌄ SOLUTIONS USE CASES ⌄ BLOG

NIST

Department of Commerce - National Institute of Standards and Technology

Differential Privacy Synthetic Data Challenge

Propose an algorithm to develop differentially private synthetic datasets to enable the protection of personally identifiable information (PII) while maintaining a dataset's utility for analysis.

Synthesized

Synthesized data: 10X the impact, 0 risks

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Synthesized solves the problem of data sharing

Privacy-compliance for data exploration

Statice offers a data anonymization solution. We enable businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

Instead of sharing original data, we enable businesses and other data owners to work with compliant synthetic datasets mimicking the structure of original data without disclosing any information about individual data points.

Synthetic Data

Enable cross boundary data analytics

usiness insight across
ithout moving or exposing

MOSTLY·AI PRODUCT SOLUTIONS USE CASES BLOG

NIST

Department of Commerce - National Institute of Standards and Technology

Differential Privacy Synthetic Data Challenge

Propose an algorithm to develop differentially private synthetic datasets to enable the protection of personally identifiable information (PII) while maintaining a dataset's utility for analysis.

Create highly realistic, privacy-safe synthetic datasets proven to be compliant even with the strictest data protection laws.

Synthesized

Synthesized data: 10X the impact, 0 risks

Synthesized solves the
a sharing

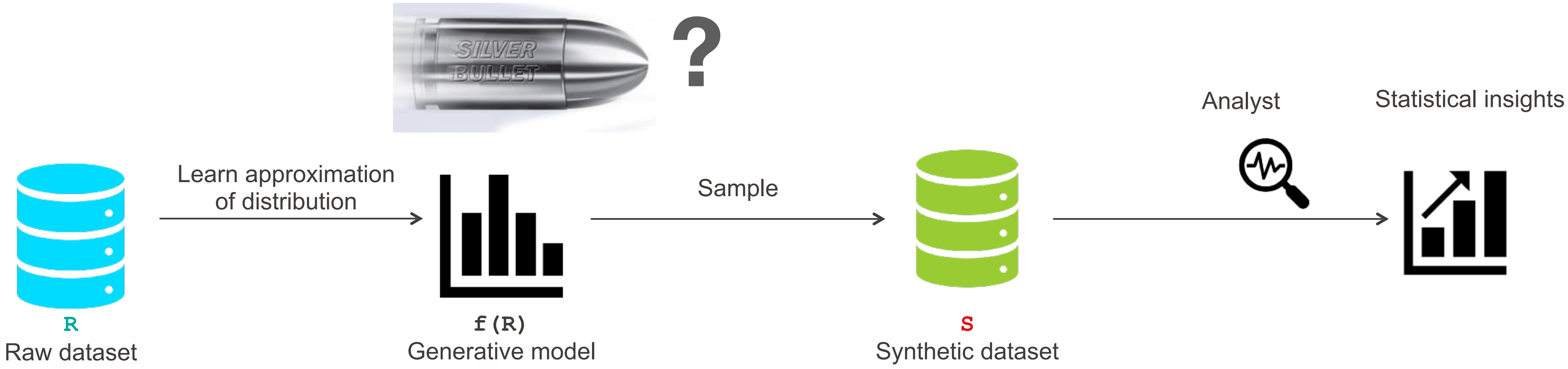
"ODI Leeds and NHS England will be working together to explore the potential of 'synthetic data.' This is data that has been created following the patterns identified in a real dataset but it contains no personal data, making it suitable to release as open data. Synthetic data is also great for building and prototyping ideas"
https://www.odileeds.org/events/synae/

Privacy-comp
data explo

Statice offers a data anonymizatio
businesses to stay innovative with smart synthetic data. Our solution empowers companies to work with complex data in a privacy-compliant manner. Data-driven innovation of tomorrow starts with protecting data today.

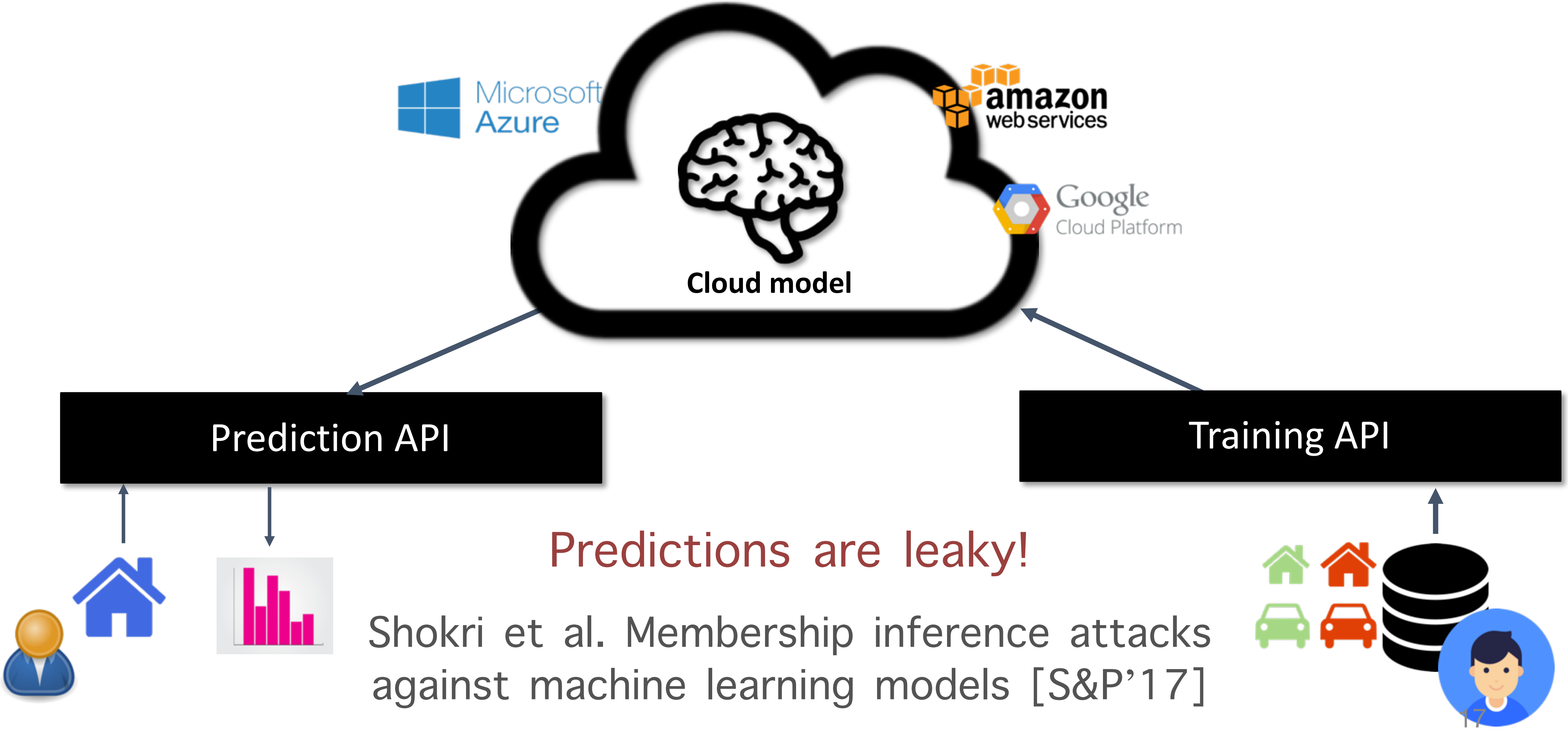
businesses and other data owners to work with compliant synthetic datasets mimicking the structure of original data without disclosing any information about individual data points.

The Promise of Synthetic Data



Attacks Against Synthetic Data?

Machine Learning as a Service

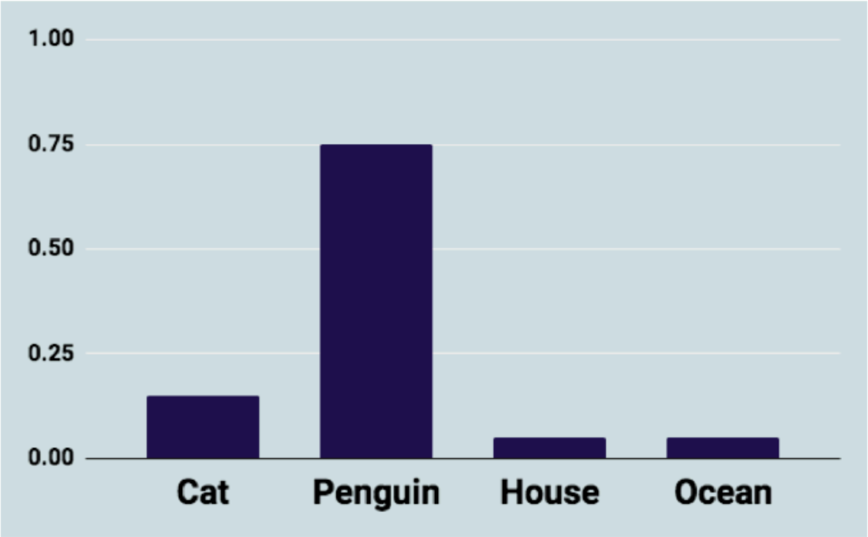


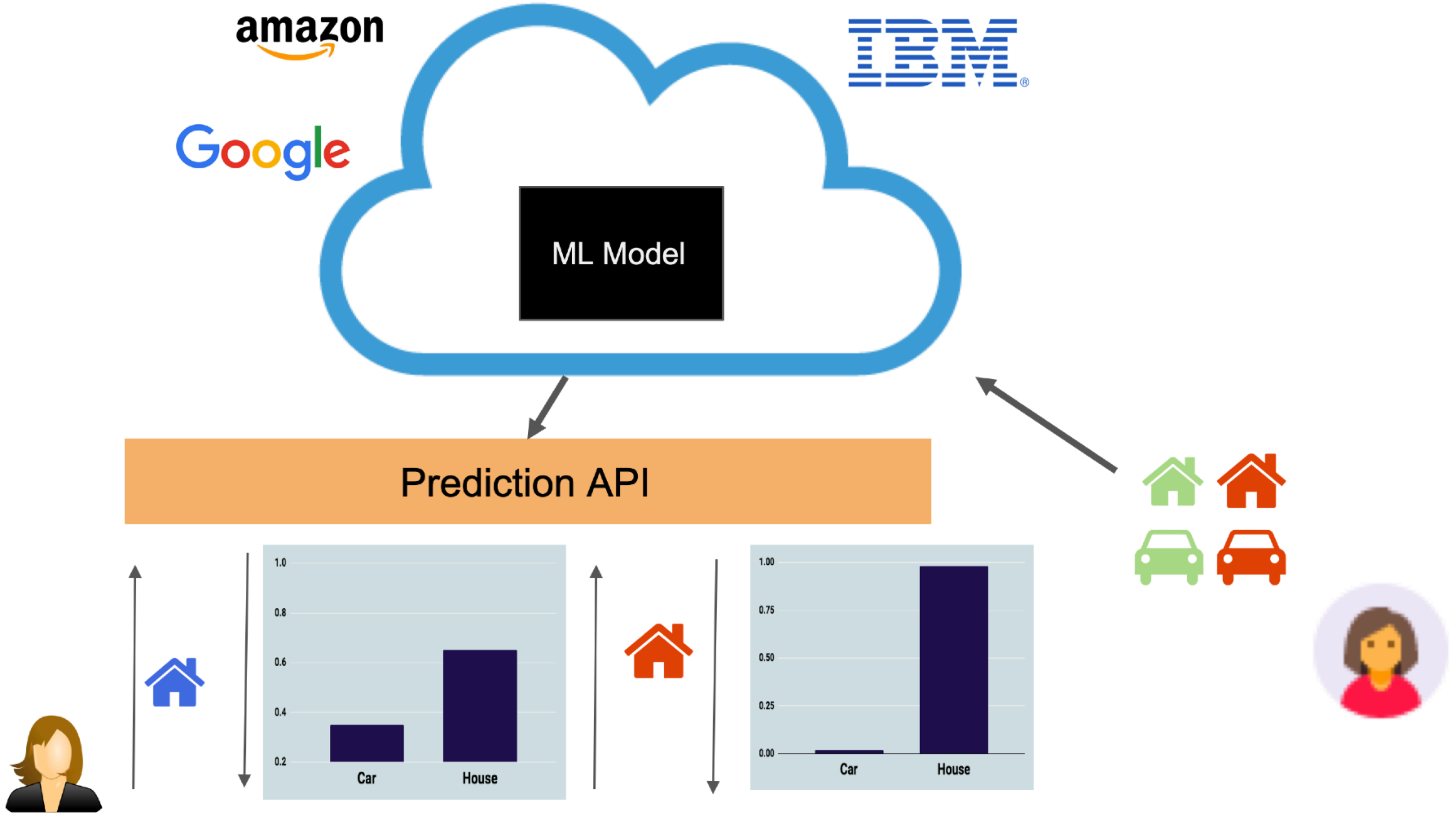
Predictions are leaky!

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

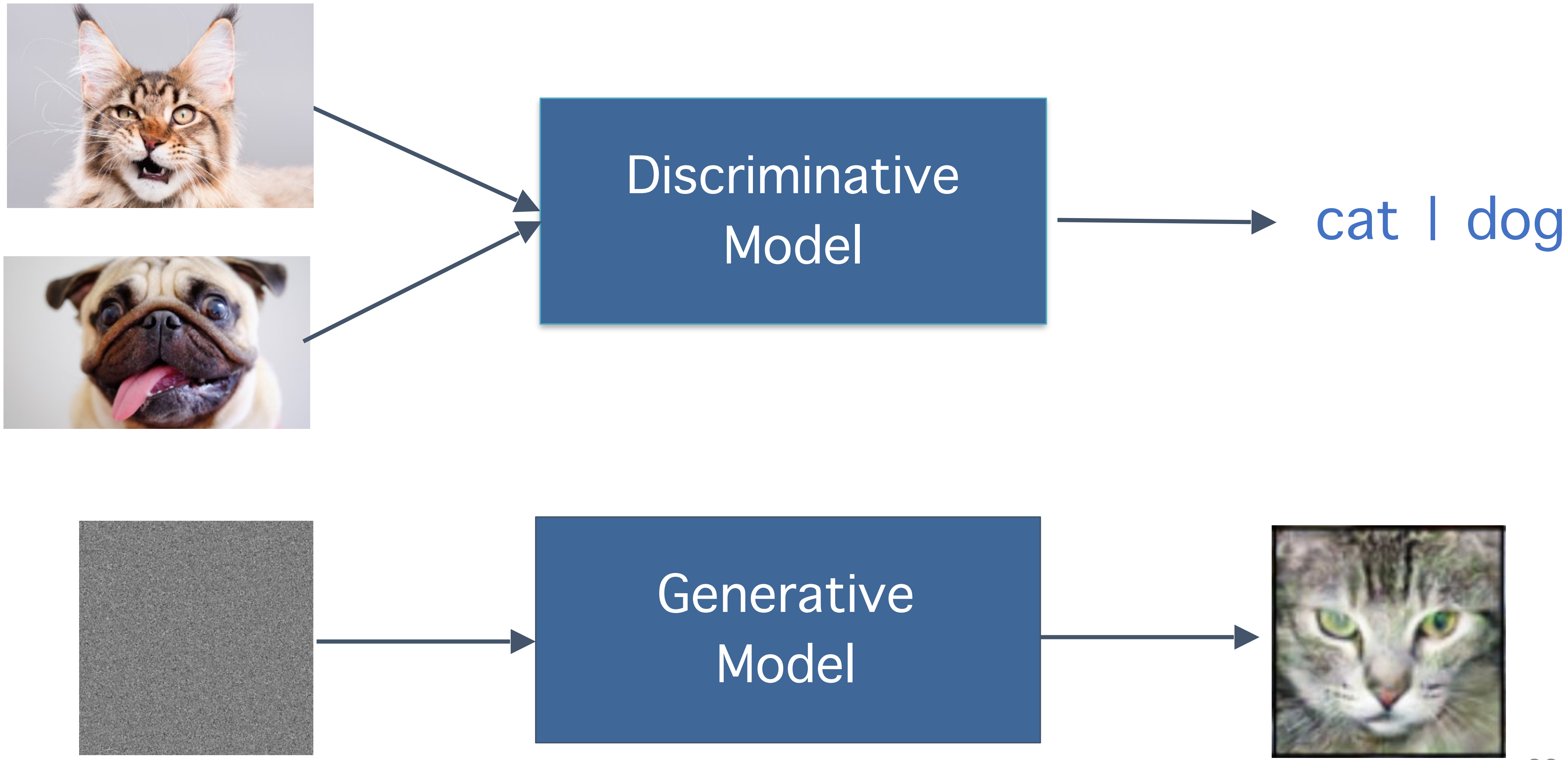
Membership Inference/Discriminative

Prediction API

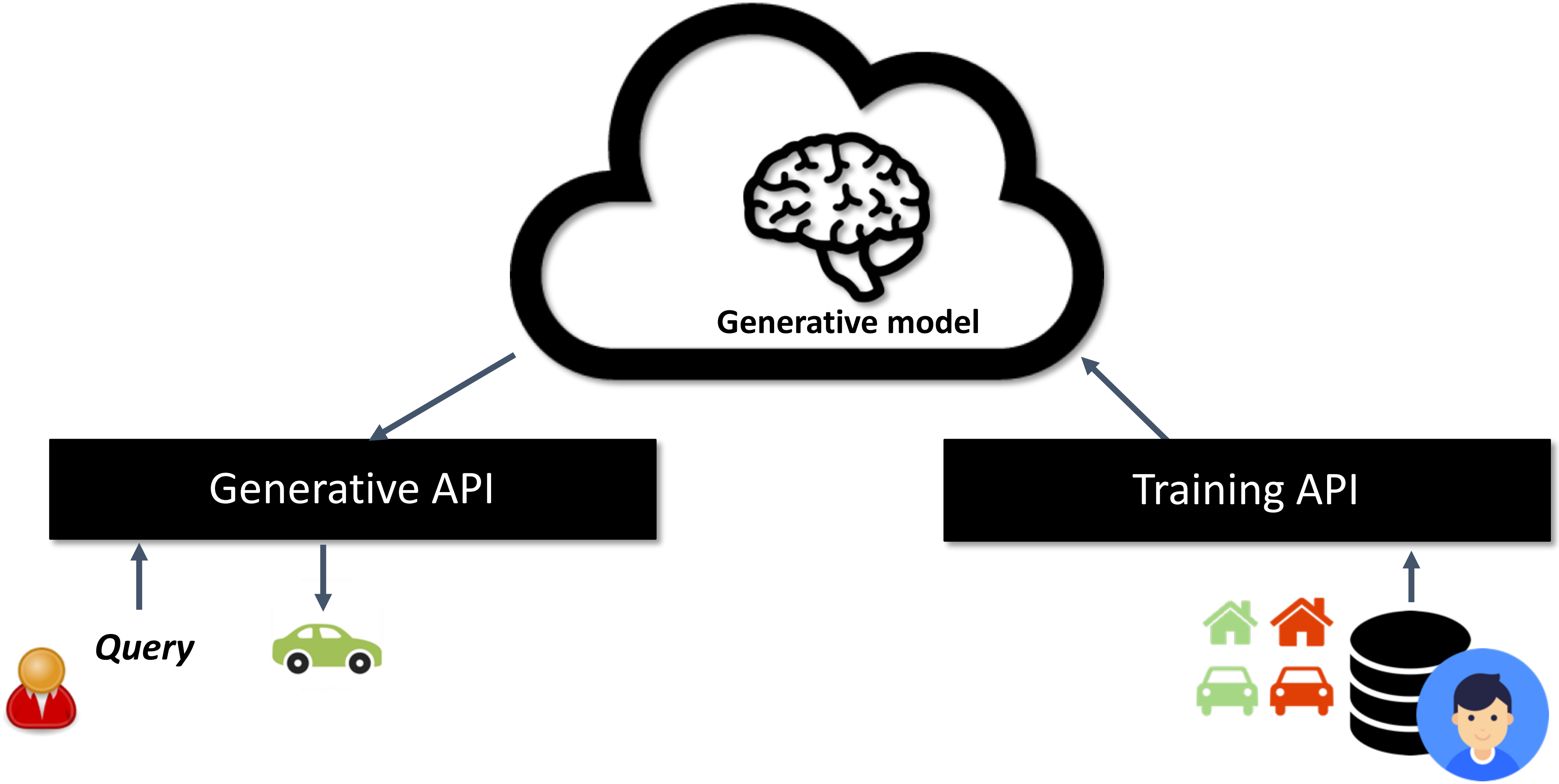



ML Model
Prediction API
1.0
0.8
0.6
0.4
0.2
Car
House
1.00
0.75
0.50
0.25
0.00
Car
House

What About Generative Models?



Membership Inference in Generative Models

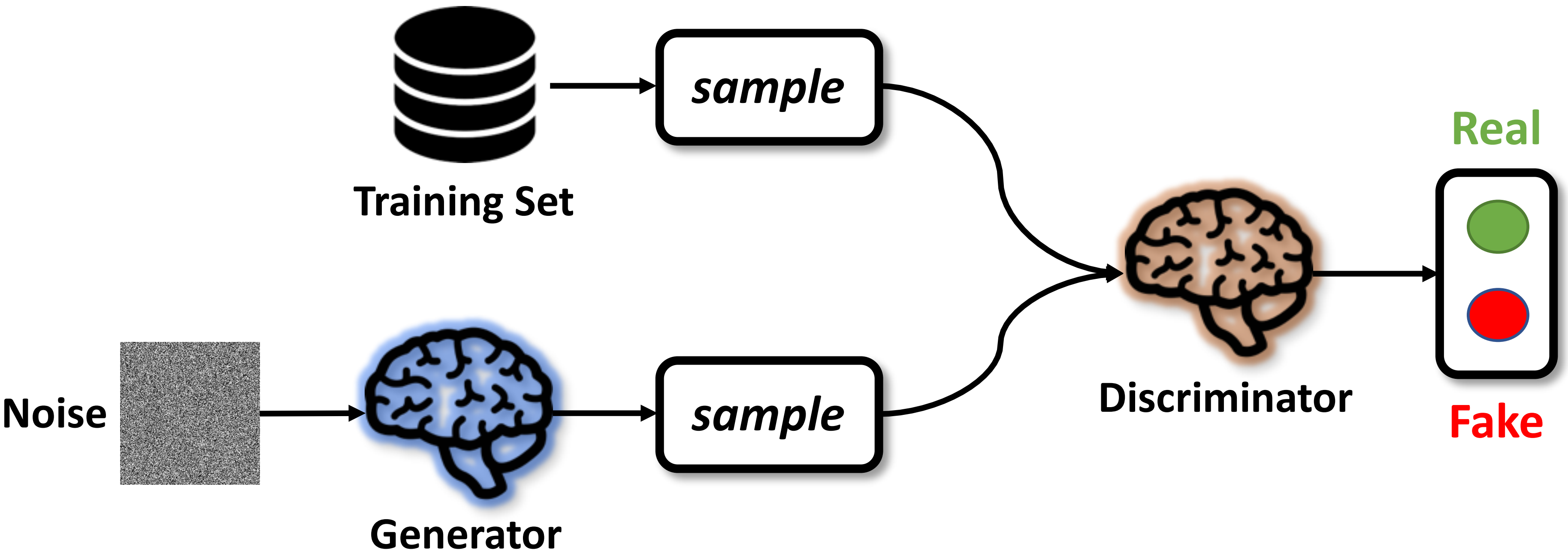


Jamie Hayes, Luca Melis, George Danezis, Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models [PETS 2019]

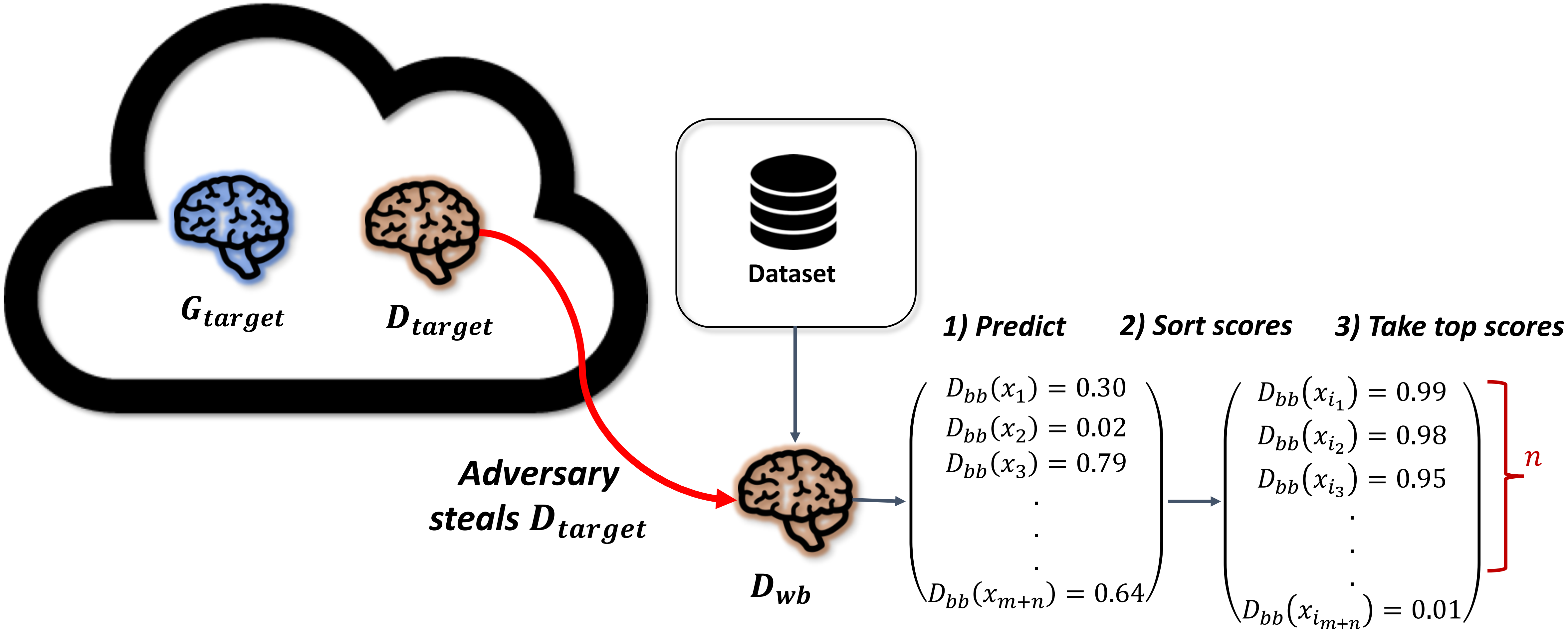
Inference without predictions?

Use generative models!

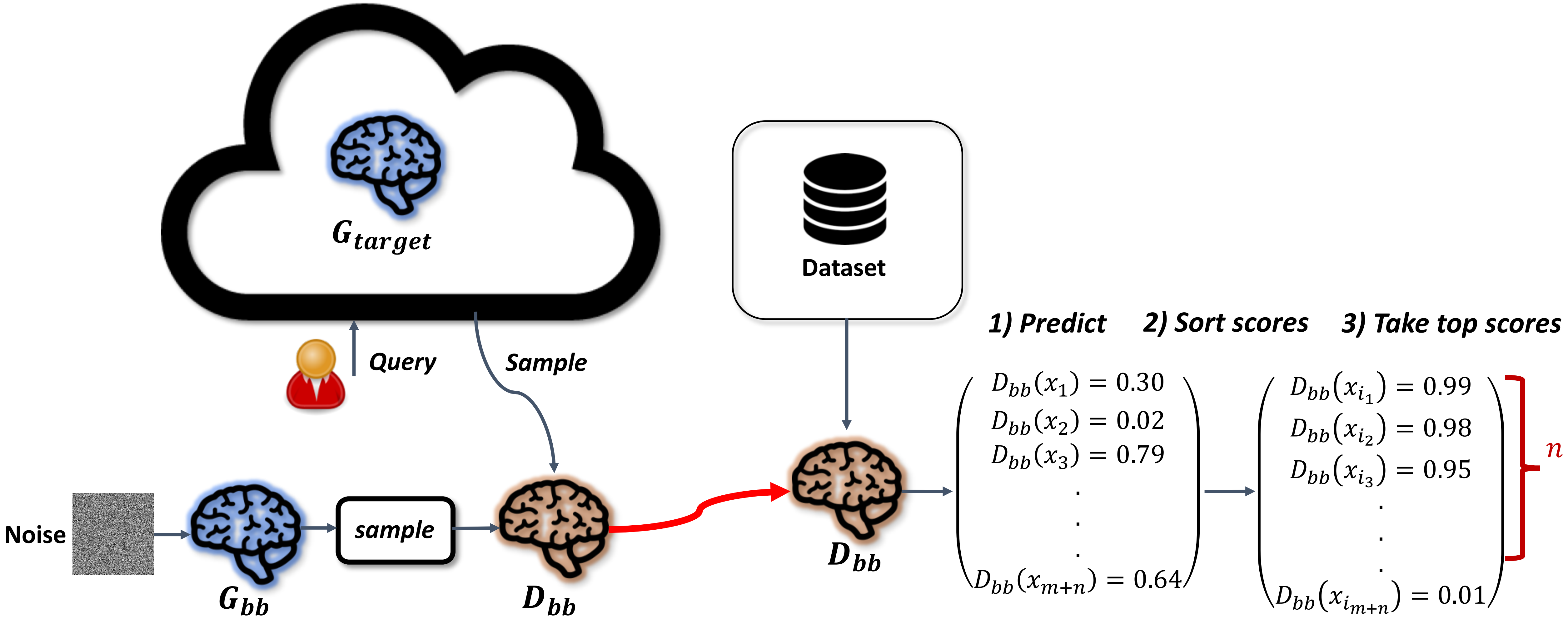
Train GANs to learn the distribution and a prediction model at the same time



White-Box Attack



Black-Box Attack

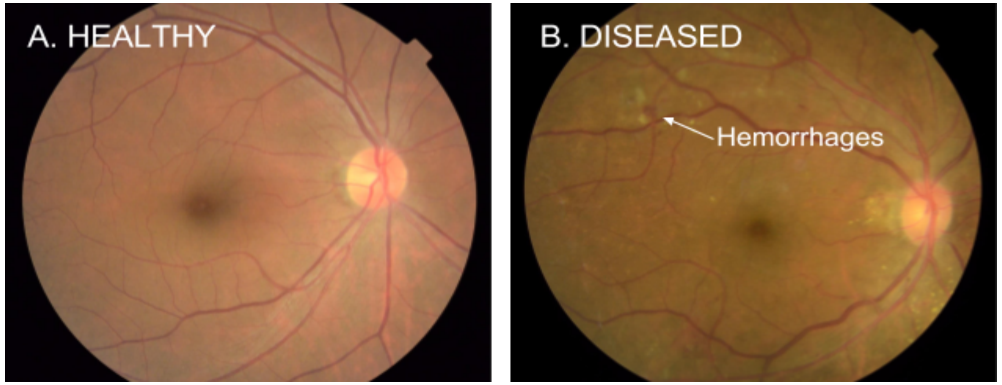


Datasets

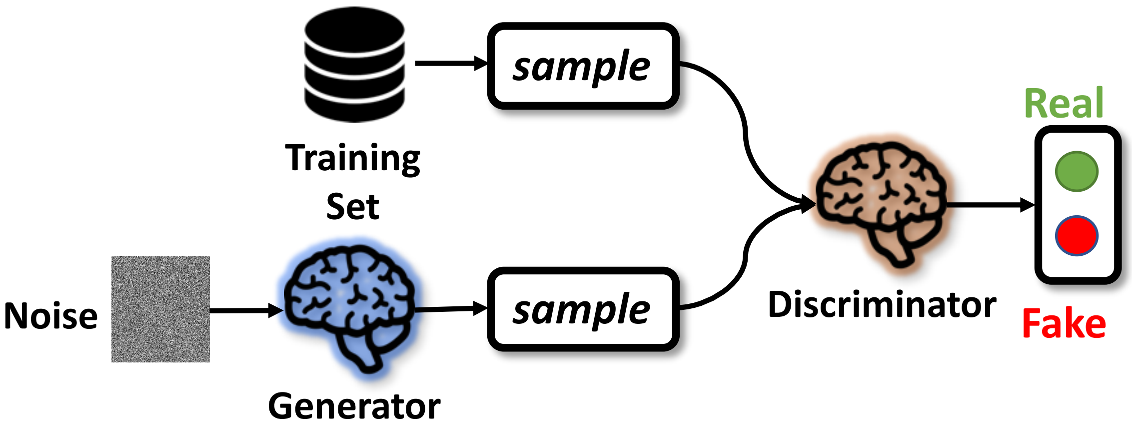
LFW

CIFAR-10

DR



Models



Attacker Model:

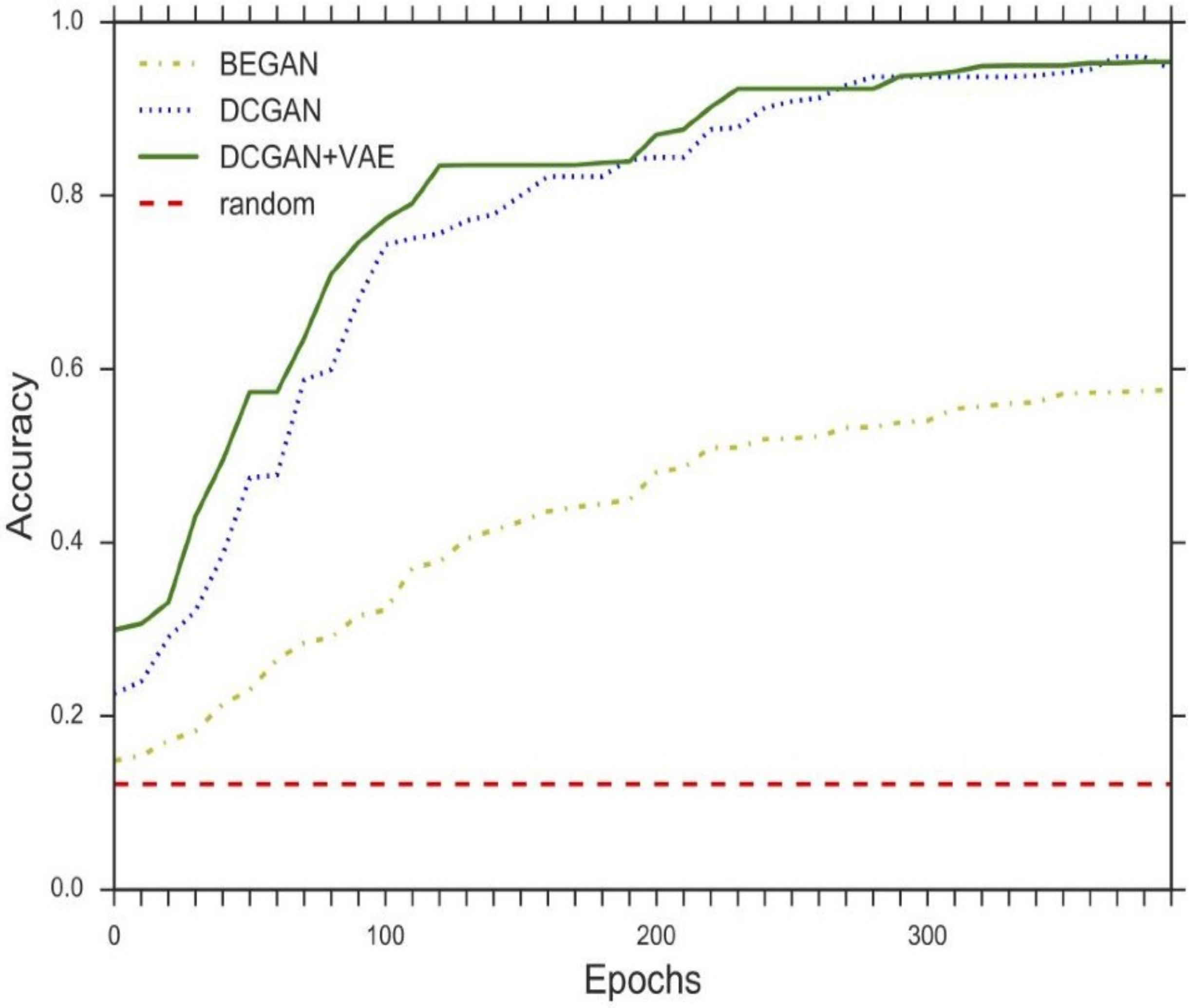
DCGAN

Target Model:

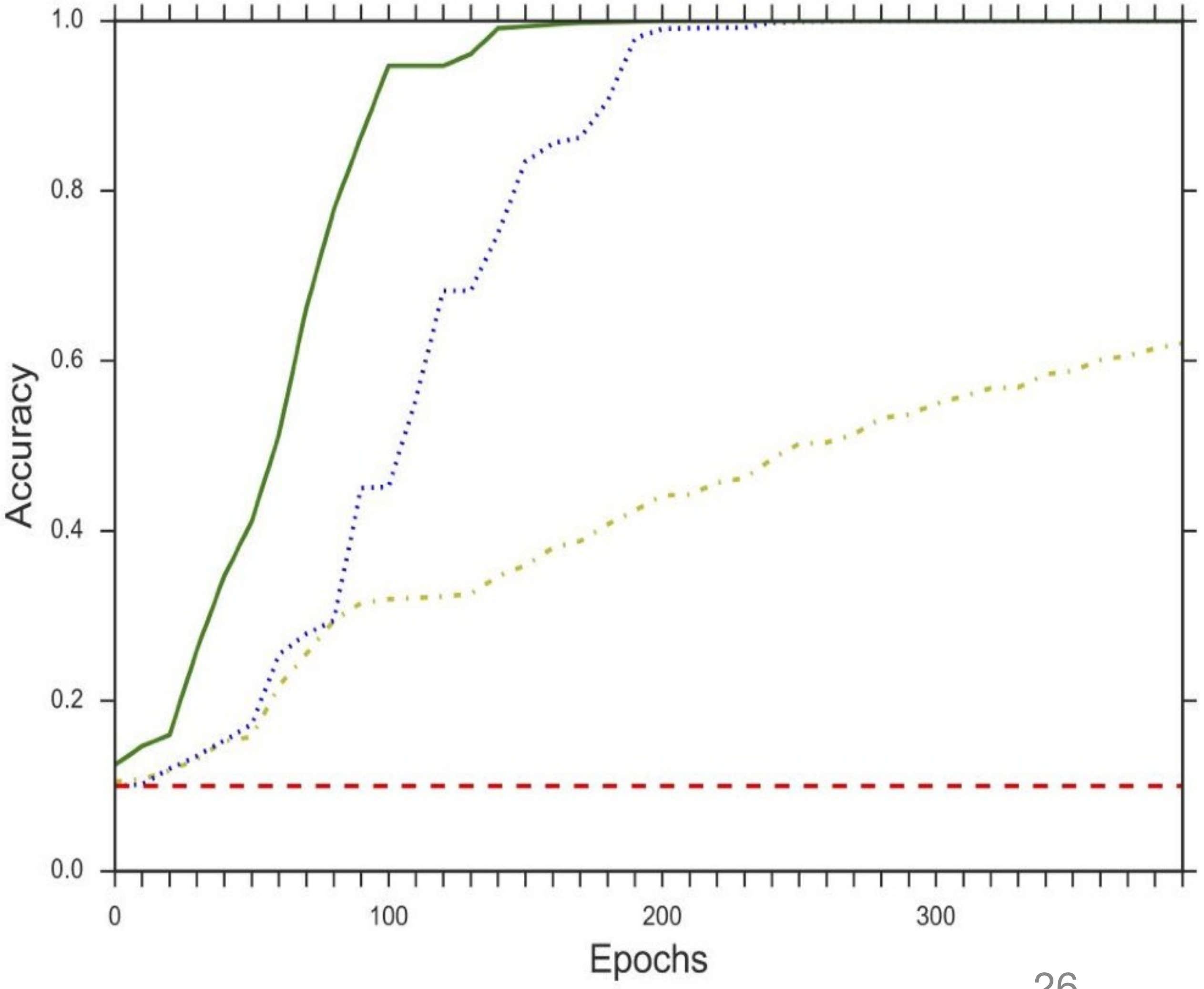
DCGAN, DCGAN+VAE, BEGAN

White-Box Results

LFW, top ten classes

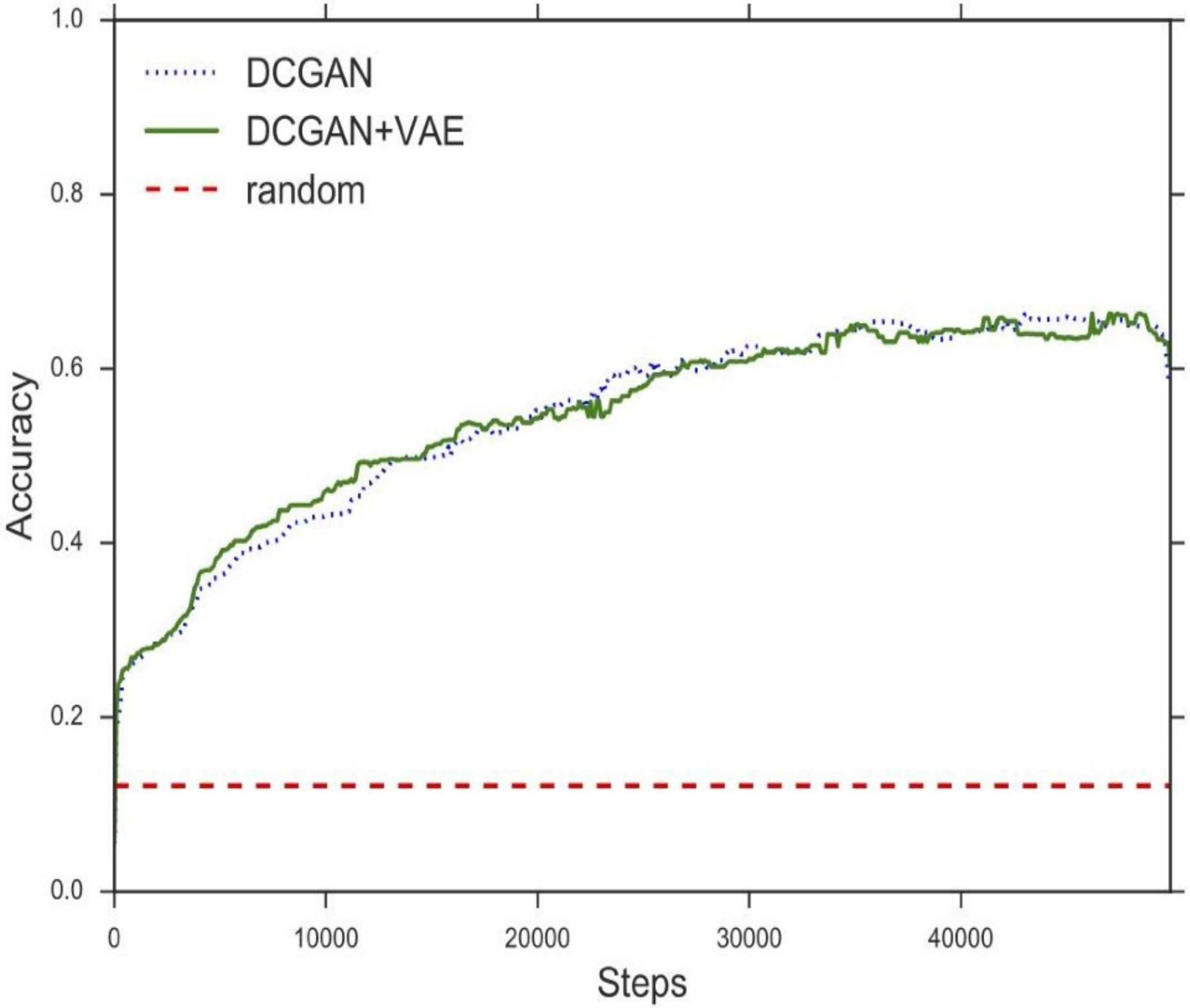


CIFAR-10, random 10% subset

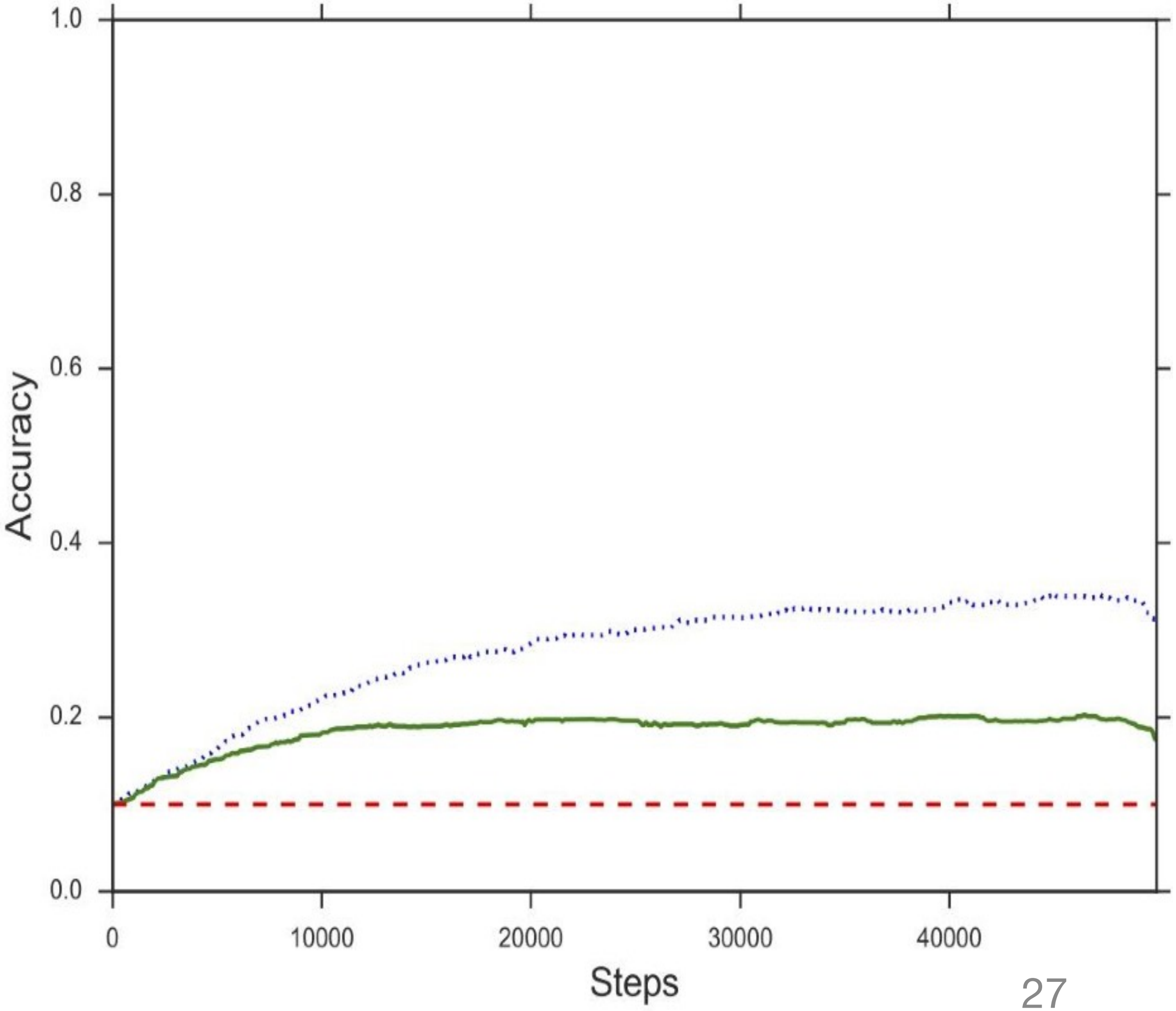


Black-Box Results

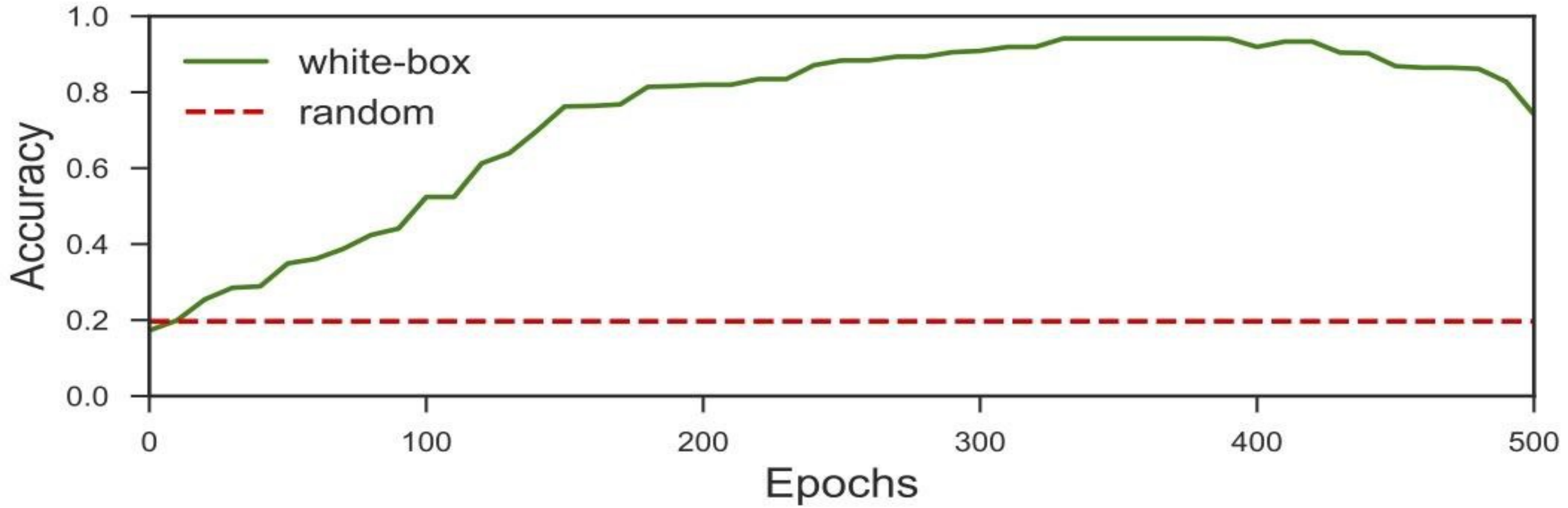
LFW, top ten classes



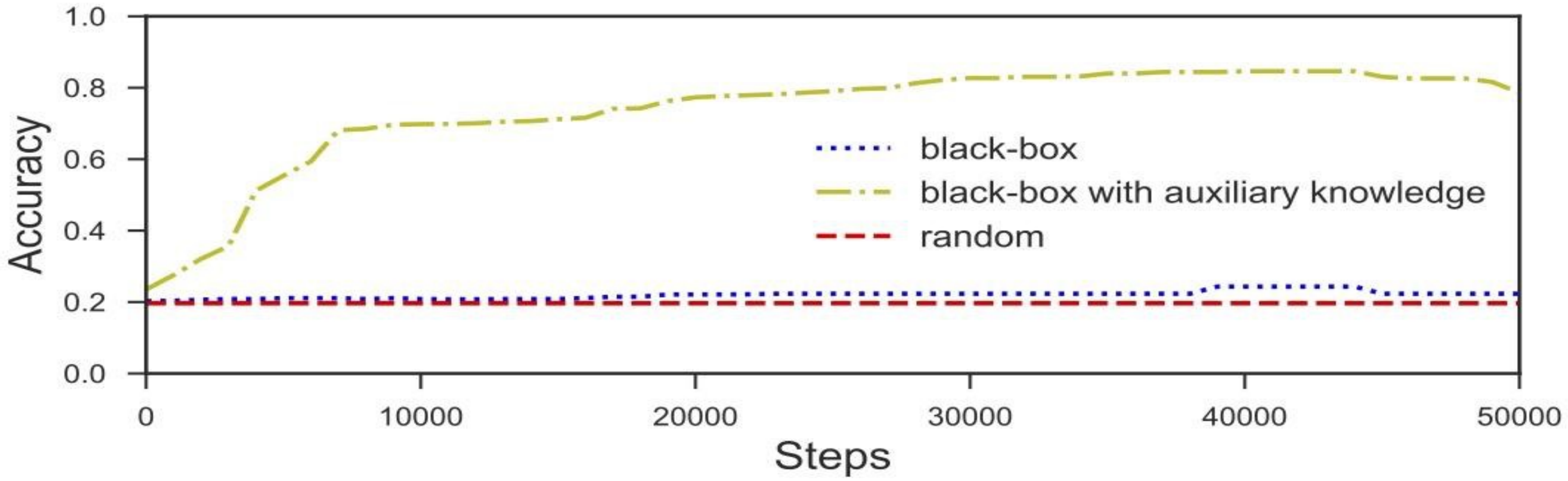
CIFAR-10, random 10% subset



DR Dataset

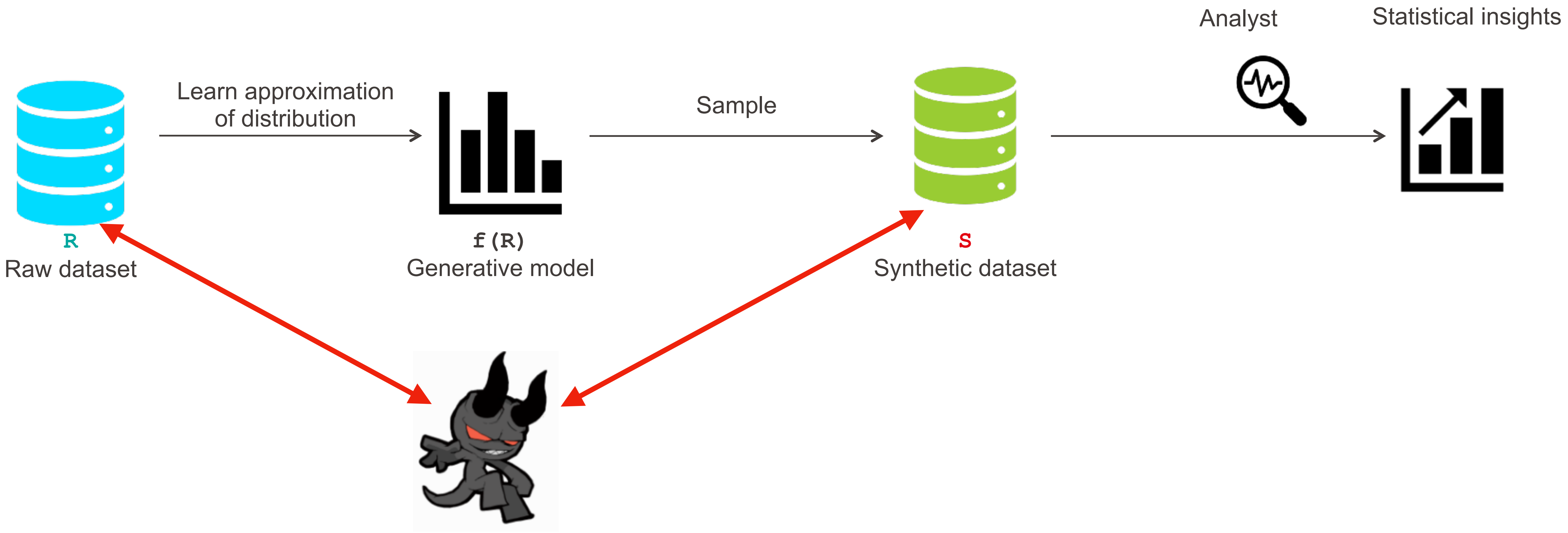


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
white-box
random
0
100
200
300
400
500
Epochs

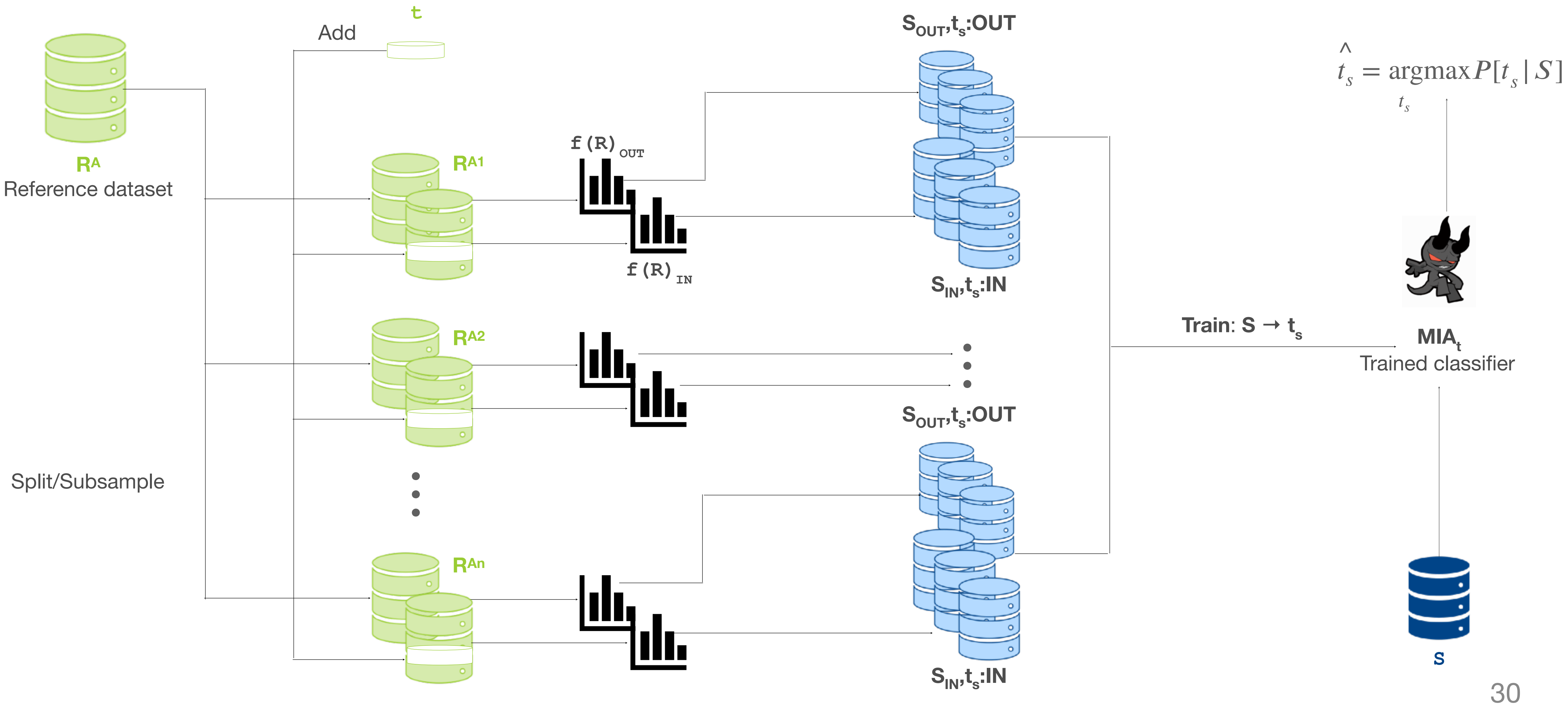


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
black-box
black-box with auxiliary knowledge
random
0
10000
20000
30000
40000
50000
Steps

[Stadler et al., Usenix’21]



Membership Inference



Privacy Gain

- Under the assumption equal prior $P[t_s] = 0.5$ and perfect linkage in case of raw dataset $P[MIA_t(R) = t_s] = 1$

$$PG_t(S, R) \triangleq \frac{1 - P[MIA_t(S) = t_s]}{2}$$

$P[MIA_t(S) = t_s] = 1$

Publishing **S** is equivalent to publishing **R**

$P[MIA_t(S) = t_s] = 0$

Publishing **S** reduces the adversary's chance of success

$PG_t = 0.25$

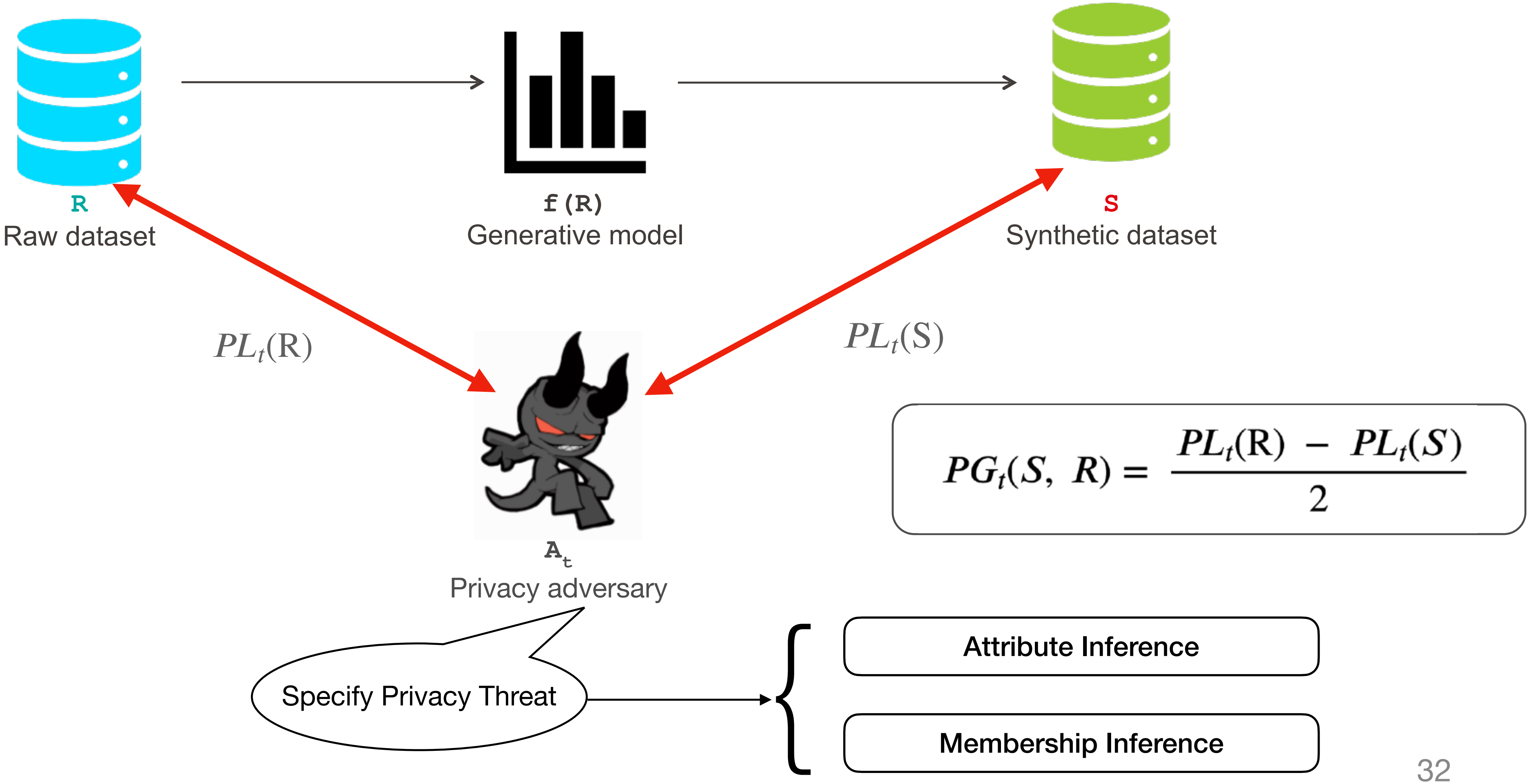
$PG_t = 0$

$PG_t = 0.5$

$P[MIA_t(S) = t_s] = 0.5$

Publishing **S** gives the adversary no advantage over **random guessing**

Privacy Gain



Agenda

Privacy in Machine Learning

Synthetic Data

Privacy (and Utility) in Synthetic Genomic Data

Agenda

Privacy (and Utility) in Synthetic Genomic Data

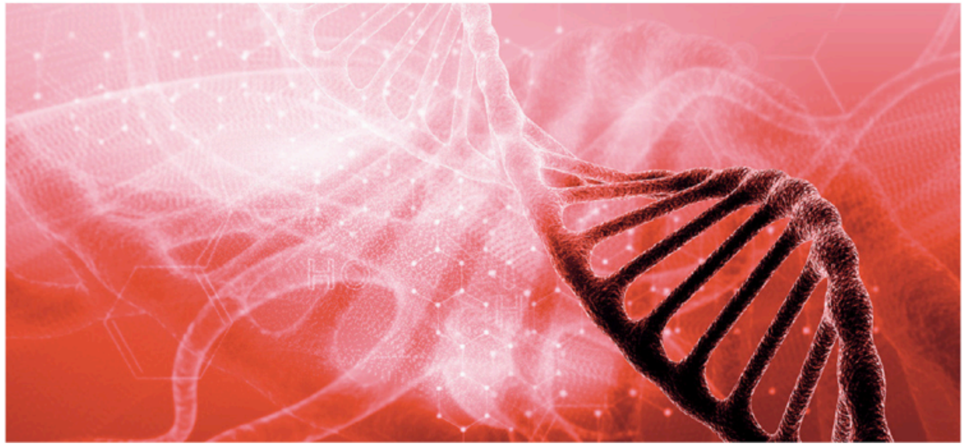
Genome Sequencing

Genome Sequencing

YaleNews EXPLORE TOPICS ▾

Yale Cancer Center scientists build genomic research platform to help treat cervical cancer

By Anne Doerr | OCTOBER 18, 2019

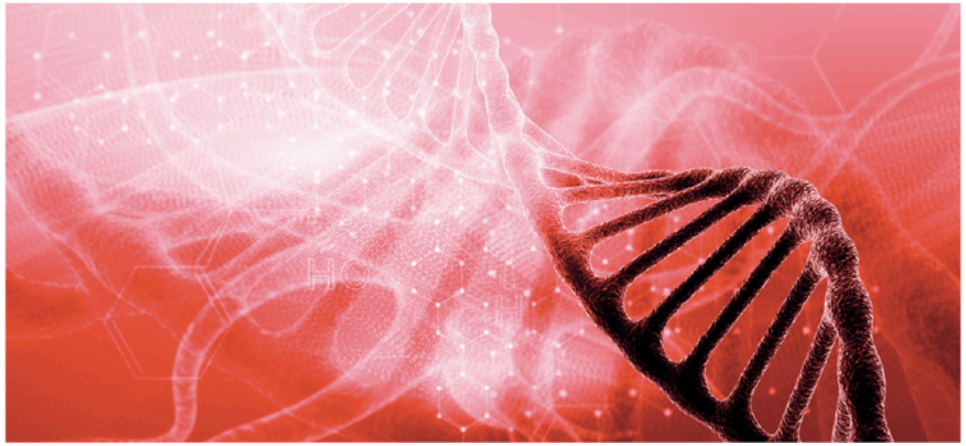


Genome Sequencing

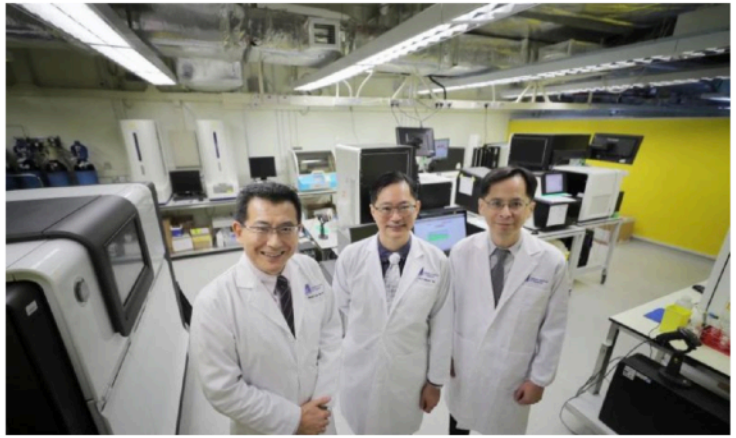
YaleNews EXPLORE TOPICS ▾

Yale Cancer Center scientists build genomic research platform to help treat cervical cancer

By Anne Doerr | OCTOBER 18, 2019



Singapore researchers create world's largest Asian genetic databank



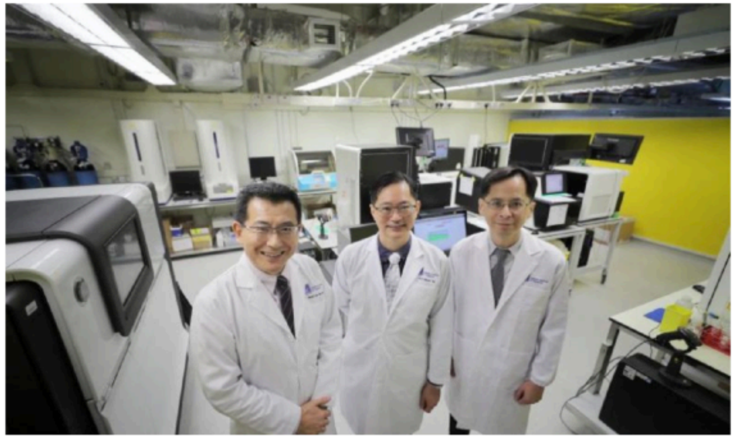
Genome Sequencing

YaleNews EXPLORE TOPICS ▾

Yale Cancer Center scientists build genomic research platform to help treat cervical cancer

By Anne Doerr | OCTOBER 18, 2019

Singapore researchers create world's largest Asian genetic databank



NIH backs new $7M genome center for All of Us research program

Jackie Drees - 17 hours ago Print | Email

SHARE Tweet Share 0

The National Institutes of Health awarded $7 million to the HudsonAlpha Institute for Biotechnology to

Threats

Threats

Technology

DNA Test Service Exposed Thousands of Client Records Online

By Nico Grant

9 July 2019, 18:16 BST *Updated on 10 July 2019, 21:09 BST*

Threats

Technology

DNA Test Service Exposed Thousands of Client Records Online

By Nico Grant
9 July 2019, 18:16 BST *Updated on 10 July 2019, 21:09 BST*

News

NHS patients' genetic data targeted as foreign hackers attack high security MoD unit

share

Save 33

Threats

Technology

DNA Test Service Expose
Client Records Online

By Nico Grant
9 July 2019, 18:16 BST *Updated on 10 July 2019, 21:09 BST*

› News

NHS patients' genetic data ta
foreign hackers attack high
MoD unit

share

China Uses DNA to Track Its People, With the Help of American Expertise

The Chinese authorities turned to a Massachusetts company and a prominent Yale researcher as they built an enormous system of surveillance and control.

Threats

Technology

DNA Test Service Expose
Client Records Online

By Nico Grant
9 July 2019, 18:16 BST *Updated on 10 July 2019, 21:09 BST*

News

NHS patients' genetic data ta
foreign hackers attack high
MoD unit

China Uses DNA to Track Its People, With the Help of American Expertise

The Chinese authorities turned to a Massachusetts company and a prominent Yale researcher as they built an enormous system of surveillance and control.

Attacks on genetic privacy via uploads to genealogical databases

Michael D. Edge, Graham Coop

doi: https://doi.org/10.1101/798272

Threats

Technology

DNA Test Service Expose Client Records Online

By Nico Grant
9 July 2019, 18:16 BST *Updated on 10 July 2019, 21:09 BST*

China Uses DNA to Track Its People, With the Help

mpany

mous

MEGAN MOLTENI SCIENCE 06.20.2019 03:05 PM

Man Found Guilty in a Murder Mystery Cracked By Cousins' DNA

The trial of William Earl Talbott II hinged on a lead from a genealogy site. The verdict will shape the future of crime-fighting and genetic privacy.

Attacks on genetic privacy via uploads to genealogical databases

Michael D. Edge, Graham Coop

doi: https://doi.org/10.1101/798272

Genomic Privacy

Treasure trove of sensitive information

- Ethnic heritage, predisposition to diseases

Genome = the ultimate identifier

- Hard to anonymize / de-identify

Sensitivity is perpetual

- Cannot be “revoked”
- Leaking one’s genome ≈ leaking relatives’ genome

Enter Synthetic Genomic Data

Recombination model (Recomb)*

Restricted Boltzmann Machines (RBM)+

Generative Adversarial Networks (GAN)+

Wasserstein GAN (WGAN)^

Recombination RBM (Rec-RBM), new

Recombination GAN (Rec-GAN), new

*Samani et al. Quantifying genomic privacy via inference attack with high-order SNV correlations
+Yelmen et al. Creating Artificial Human Genomes Using Generative Models
^Killoran, et al. Generating and designing DNA with deep generative models

Datasets

CEU Population (HapMap Project)

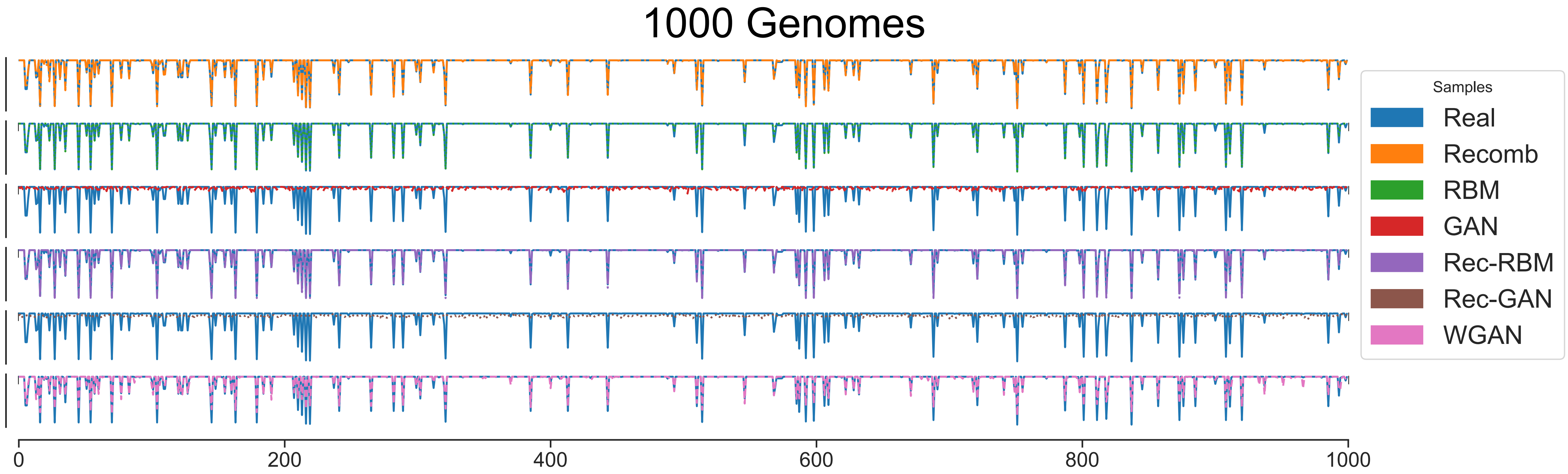
CHB Population (HapMap Project)

1,000 Genomes Project

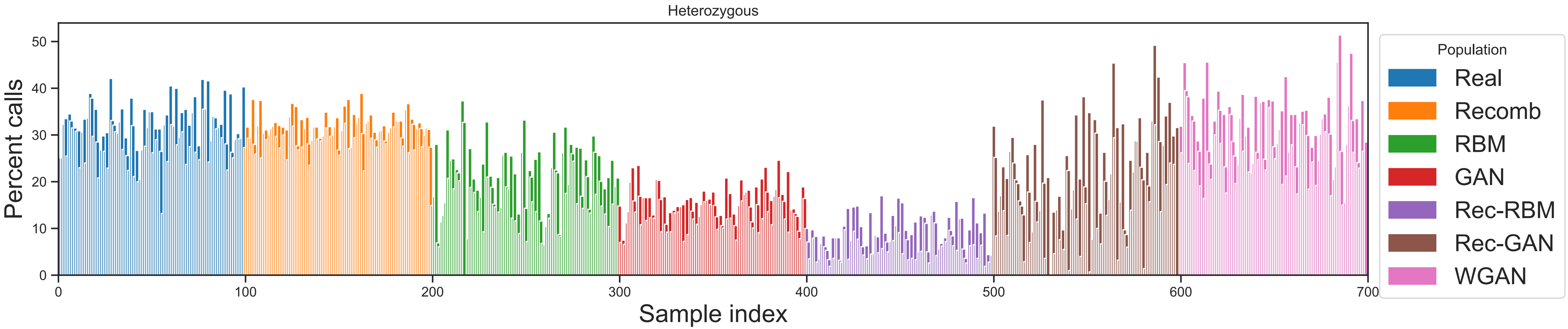
Utility

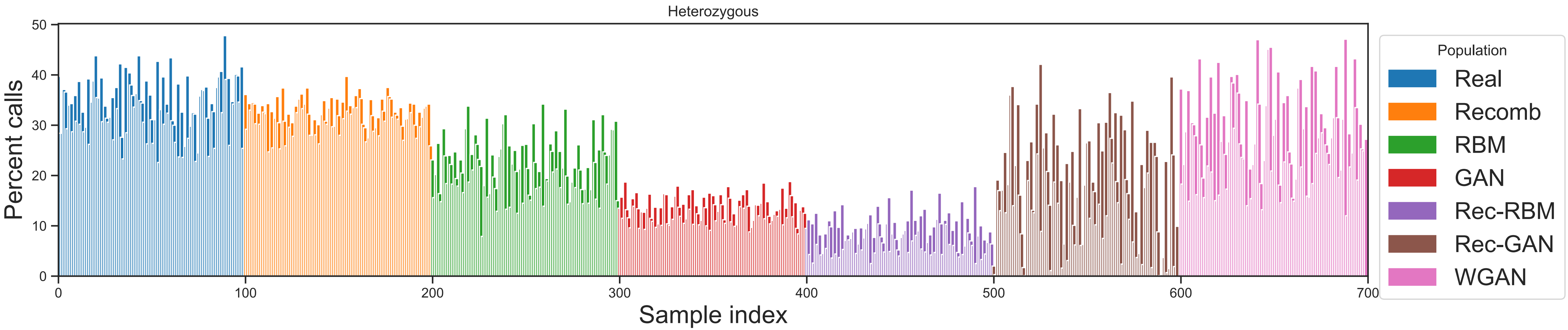
Allele Statistics

Allele Statistics

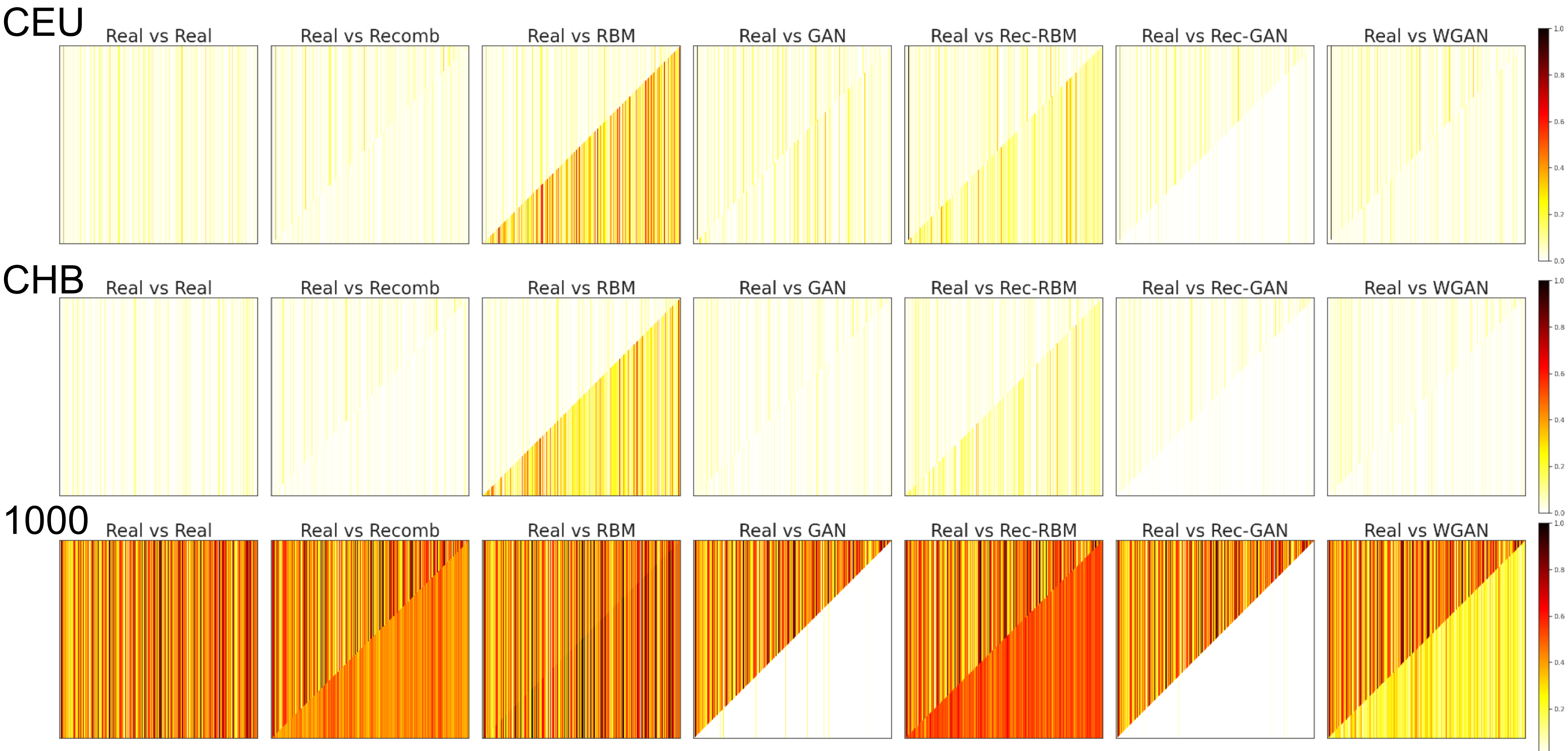


Population Statistics



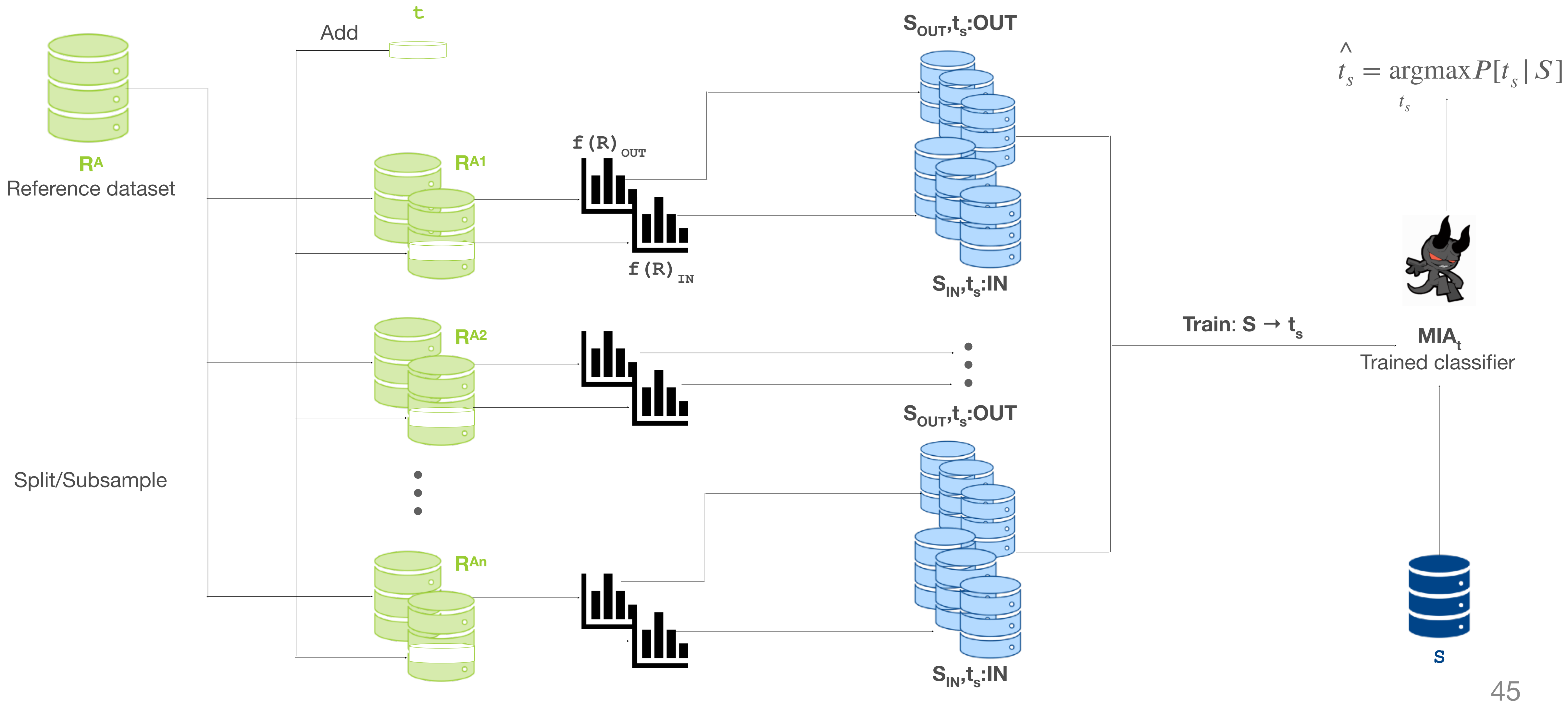


Linkage Disequilibrium

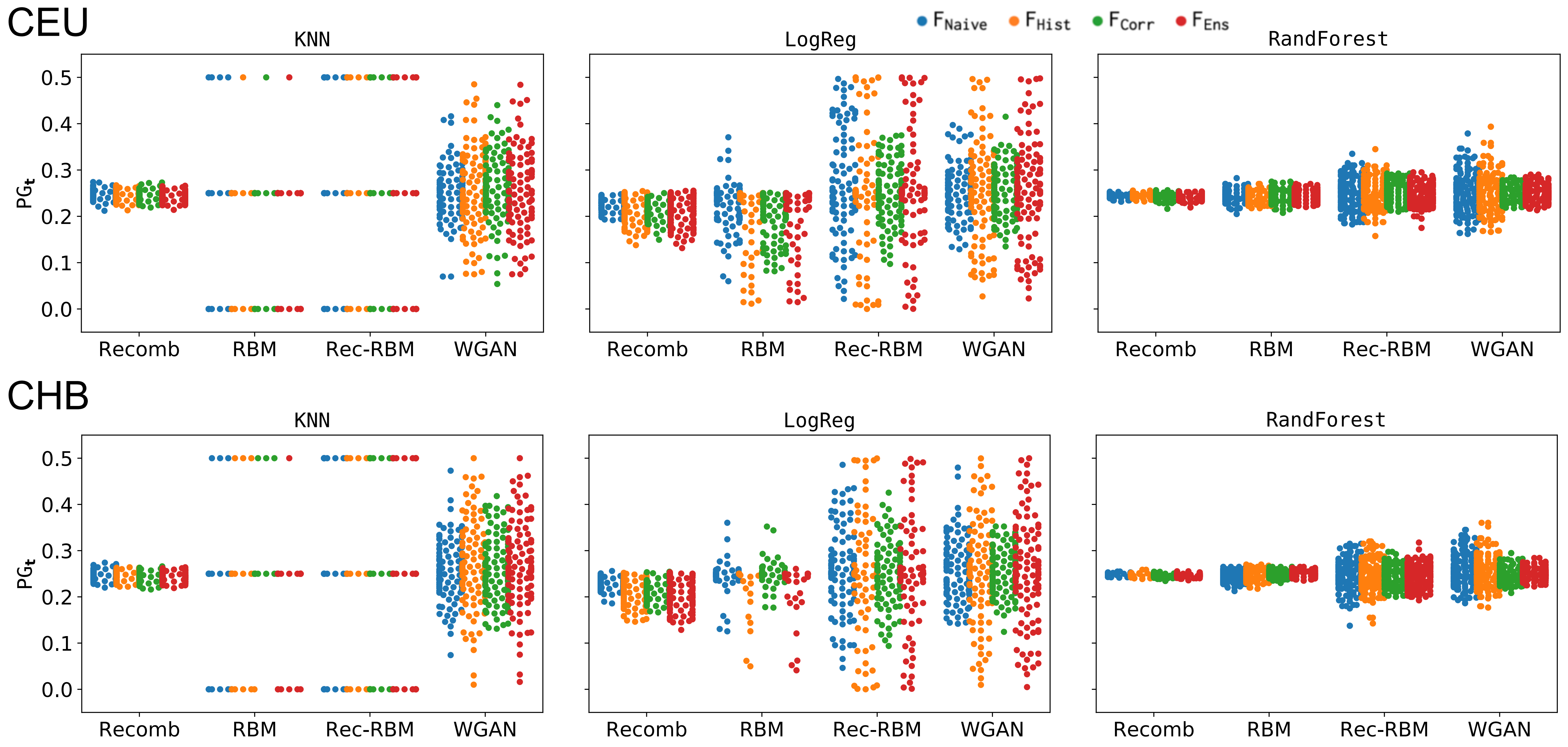


Privacy

Membership Inference

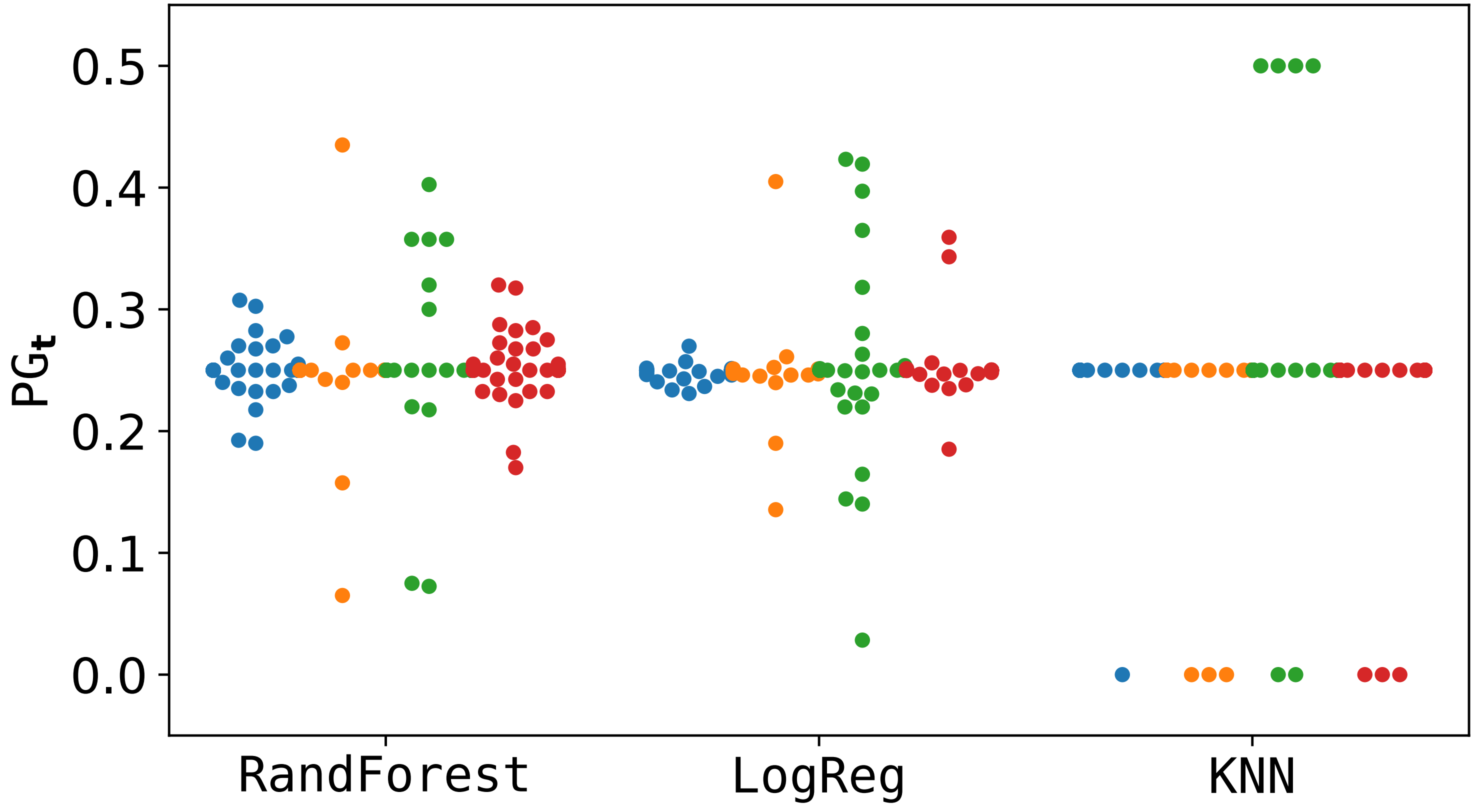


Membership Inference



Membership Inference

1000 Genomes



Membership Inference w Partial Information

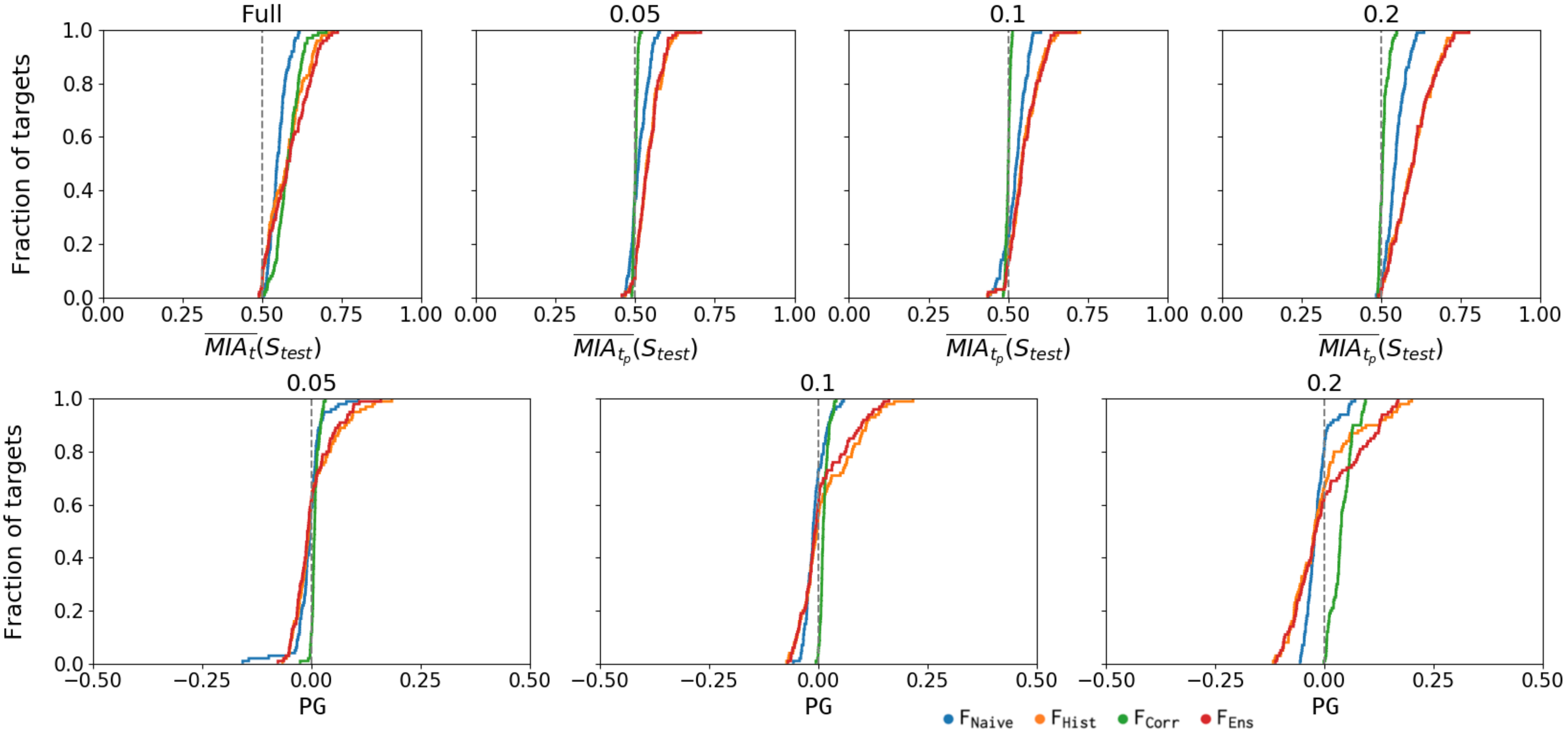
- We only give the attacker access to a fraction of SNVs from the target sequence, chosen at random.
- The attacker then uses the Recombination model as an inference method to predict the rest of the sequence.
- The PG formula needs adjusting:

$$PG_t = \frac{\overline{MIA_{t_p}}(R_t) - \overline{MIA_{t_p}}(S_{test})}{2}, \text{ where}$$

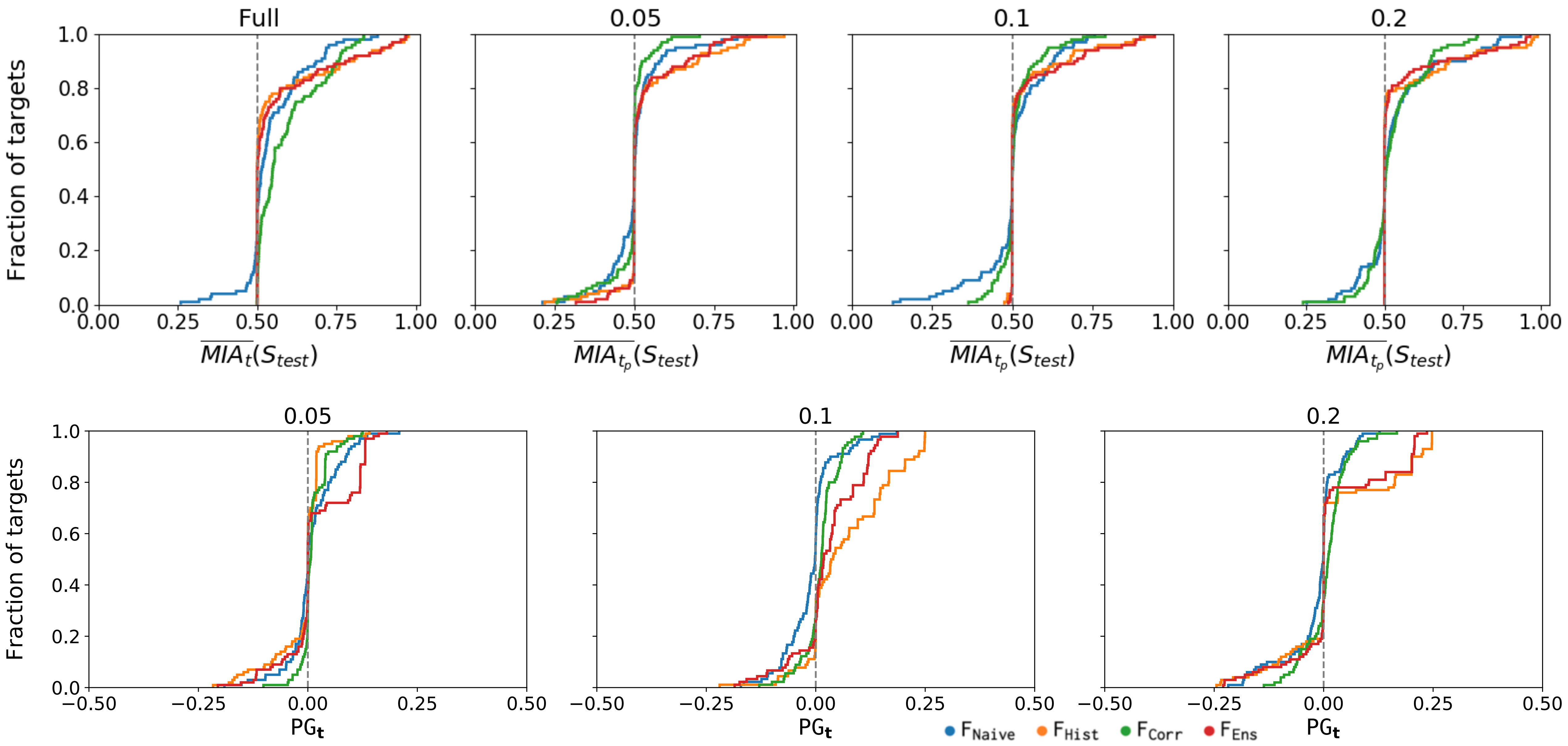
$$\overline{MIA_{t_p}}(S_{test}) = \sum_{S_i \in S_{test}} \frac{\Pr[MIA_{t_p}(S_i) = 1]}{2 * n_s}, \text{ and}$$

$$\overline{MIA_{t_p}}(R_t) = \sum_{R_i \in R_t} \frac{\Pr[MIA_{t_p}(R_i) = 1]}{2 * n_s}.$$

MIA with Partial Information



MIA with Partial Information



Take Aways

- High-quality synthetic data must accurately capture the relations between data points; however, this can enable attackers to infer sensitive information about the training data used to generate the synthetic data
- The size of the training dataset matters, especially in the case of non-statistical generative models
- Overall, there is no single method that outperforms the others for all metrics and all datasets.

Conclusion

This slide is intentionally left blank.