Investigating Membership Inference Attacks under Data Dependencies

July 13th, 2023

Thomas Humphries, Simon Oya , Lindsey Tulloch, Matthew Rafuse, Ian Goldberg,
Urs Hengartner, and Florian Kerschbaum

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

CrySP
Cryptography, Security, and Privacy Research Group

Membership Inference

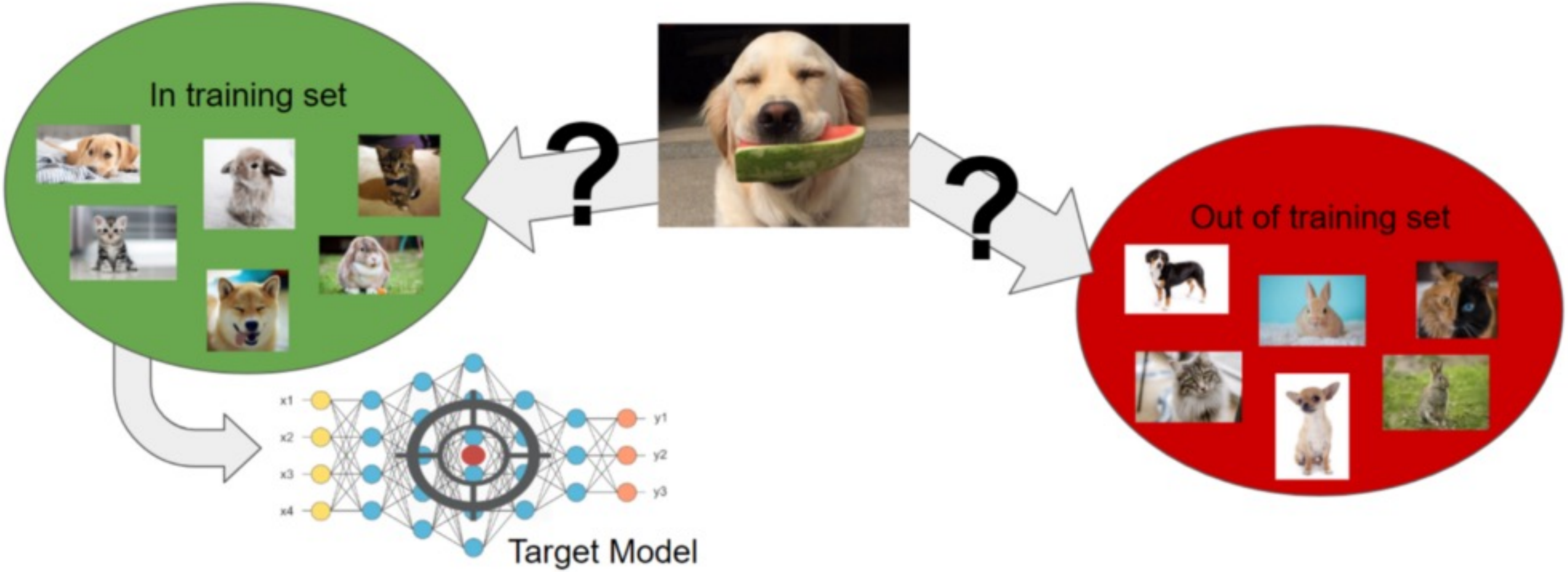
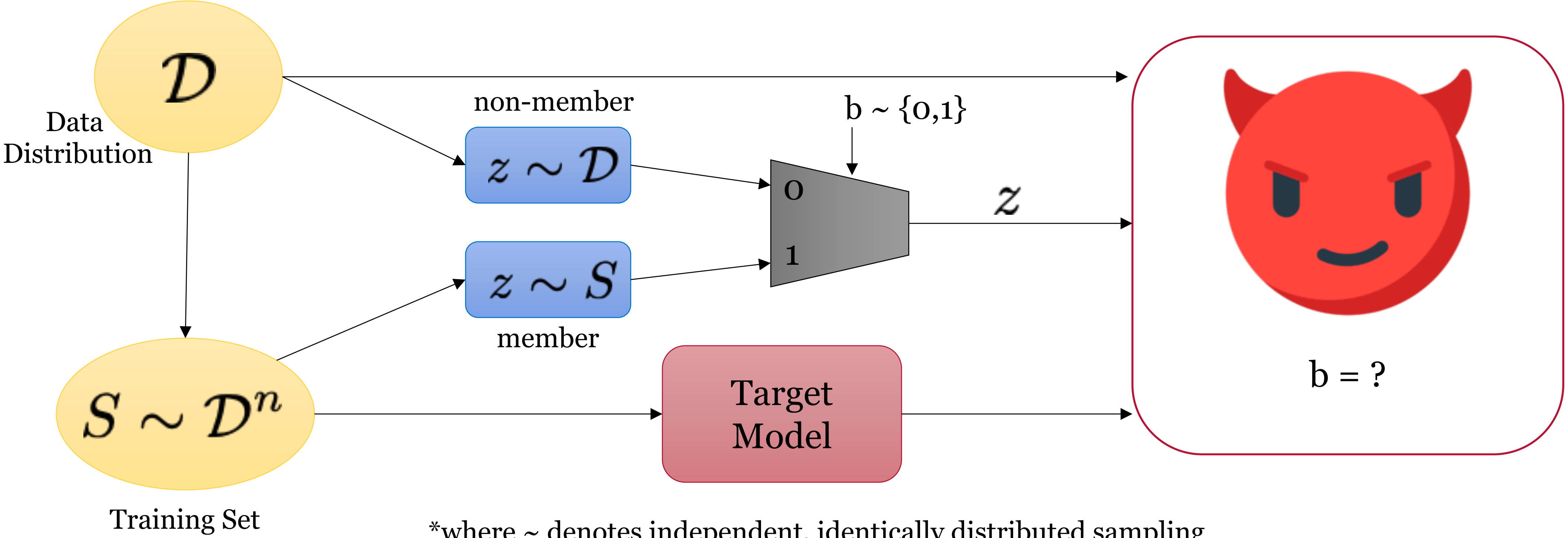


IMAGE CREDIT: Lucas Tindall – Lab41

Yeom et al.'s Membership Experiment



*where ~ denotes independent, identically distributed sampling

$$Adv = 2 \cdot Pr(\text{adversary is correct}) - 1$$
$$= TPR - FPR$$

UNIVERSITY OF WATERLOO

Standard ML Models are Vulnerable to MIAs

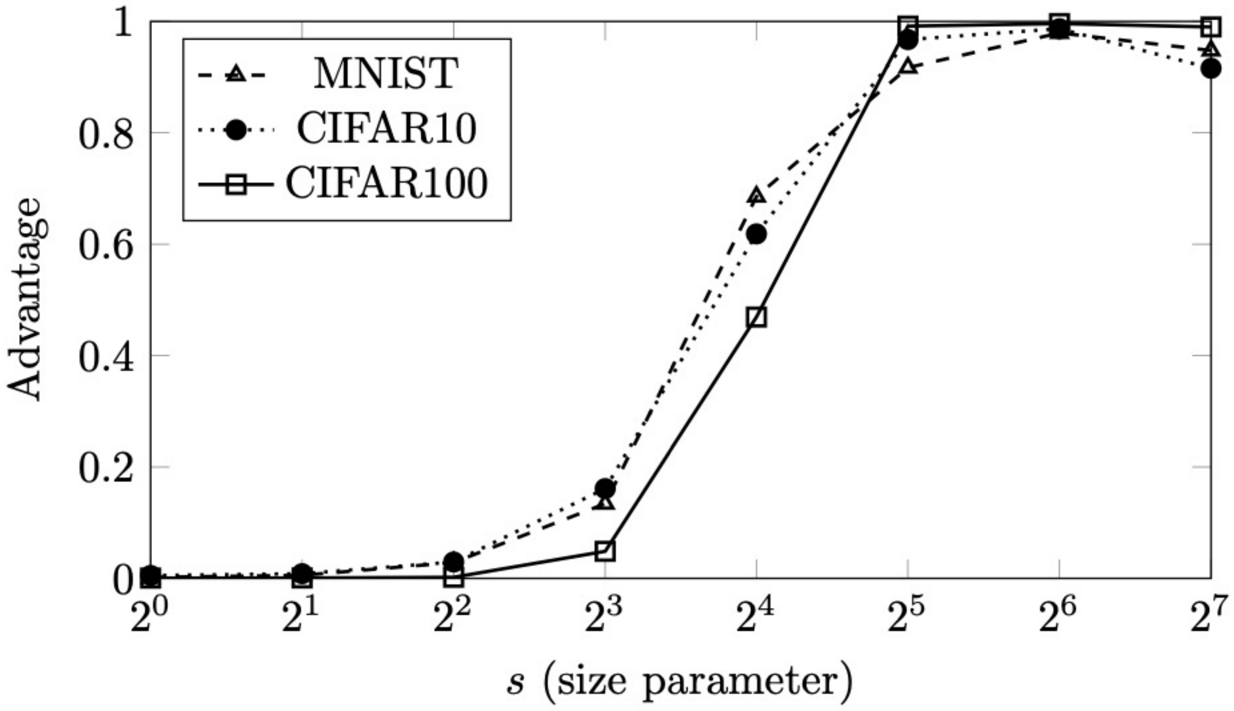


IMAGE CREDIT: Yeom et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting (2018)

UNIVERSITY OF WATERLOO

Differentially Private Learning

$$Pr(A(S) \in \mathcal{R}) \leq e^{\epsilon} \cdot Pr(A(S') \in \mathcal{R}) + \delta$$

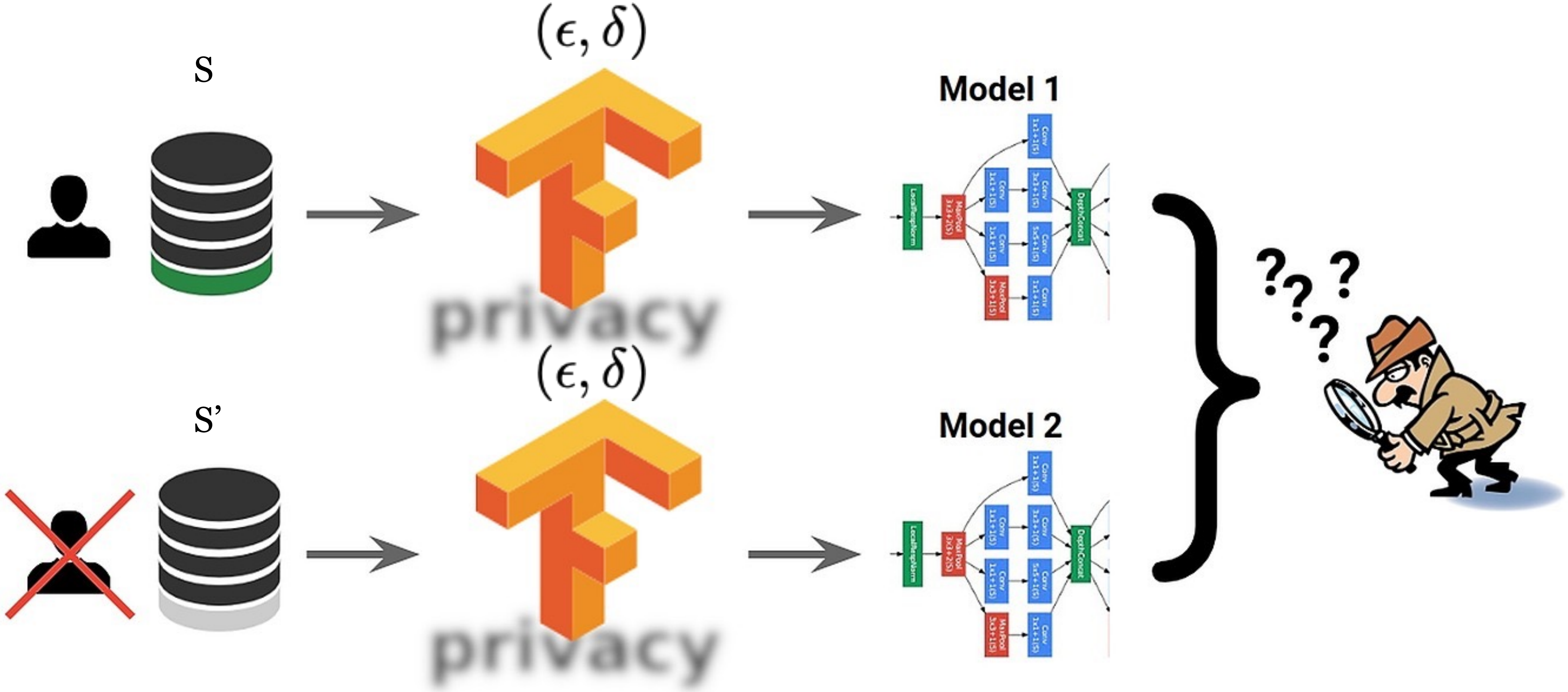


IMAGE CREDIT: Sharmistha Chatterjee – Hackernoon

Bounds on MIAs

- The properties of DP allow certain bounds to be proven (under Yeom et al.'s experiment)
- Yeom et al.'s Bound 2018

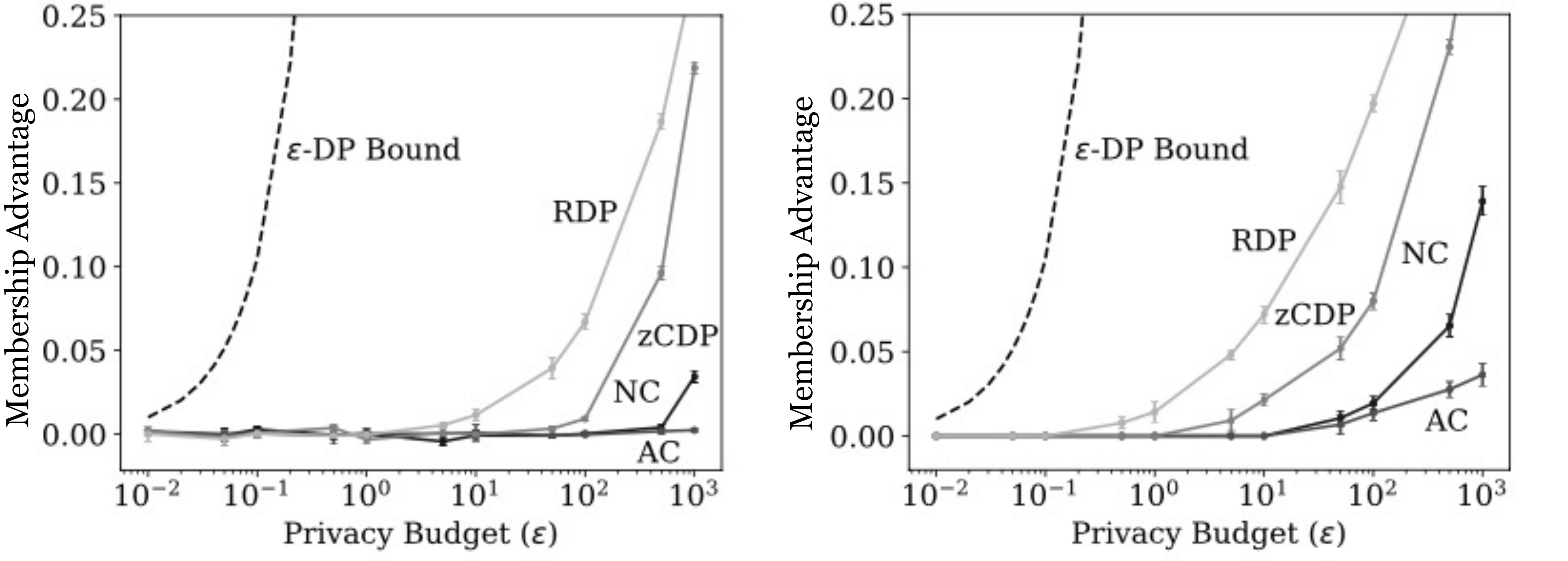
$$Adv \leq e^{\epsilon} - 1$$

- Erlingson et al.'s Bound 2019

$$Adv \leq 1 - e^{-\epsilon}(1 - \delta)$$

- Current belief is that they are quite loose in practice.

The Gap Observed in the Literature



(a) Shokri et al. membership inference (b) Yeom et al. membership inference

IMAGE CREDIT: Jayaraman and Evans - Evaluating Differentially Private Machine Learning in Practice (2019)

ML Privacy meter

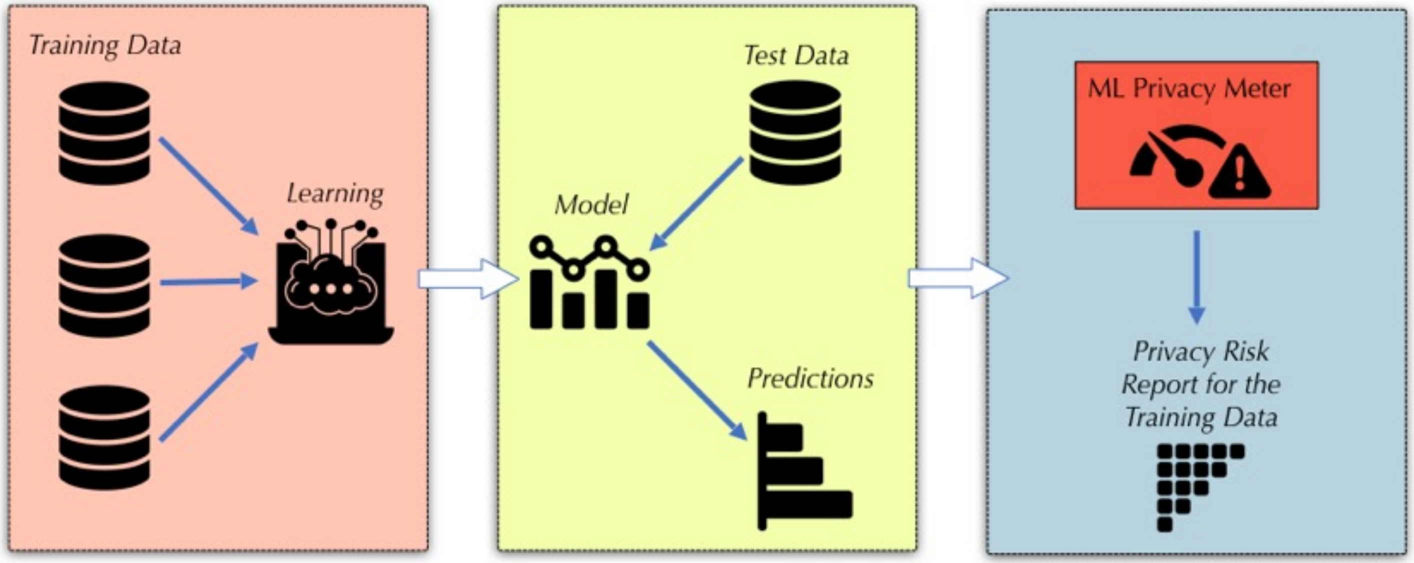


IMAGE CREDIT: Murakonda and Shokri – ML Privacy Meter (2020)

- Data analyst provides model along with training and test data to get a risk score.
- Risk score is calculated by running state of the art MIAs on user provided data.

UNIVERSITY OF WATERLOO

In Summary...

- ML models can be vulnerable to MIAs
- DP is a popular defense that gives provable bounds on MIAs
  - When samples are independent from the same distribution (IID assumption)
- Risks are generally thought to be much lower than the bound in practice

Our Contributions

We investigate prior membership experiments and provide a tighter bound under Yeom et al.'s experiment.

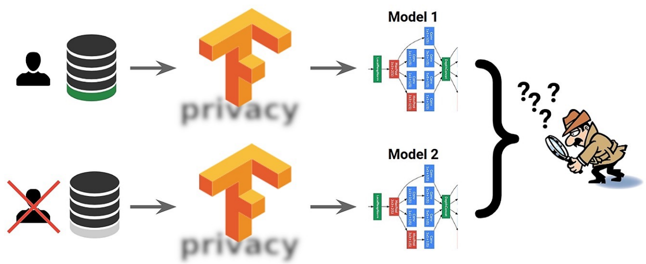
We construct a *generalized membership experiment* that addresses the weaknesses of previous experiments.

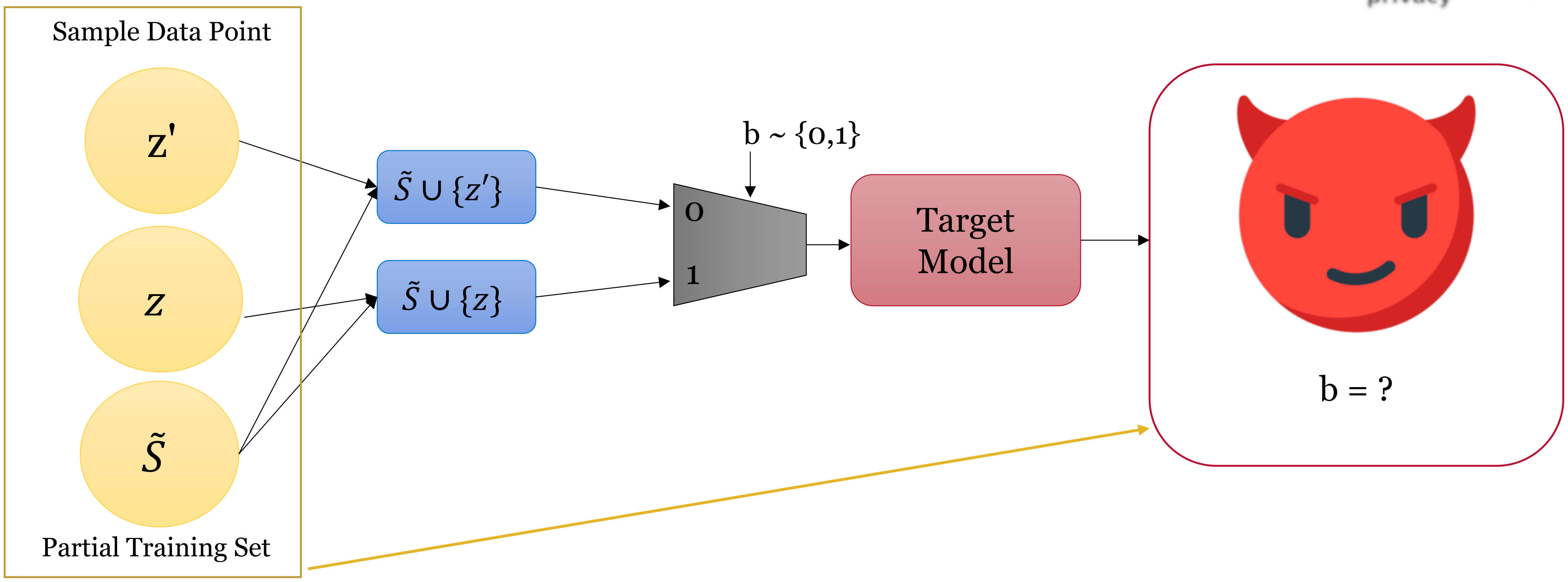
We evaluate the performance of off-the-shelf MIAs under our generalized membership experiment.

We show that dependencies have a strong influence on attack performance, surpassing the theoretical bounds of DP.

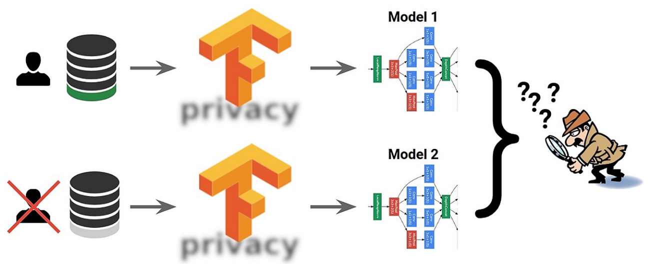
CURRENT MIA EXPERIMENTS

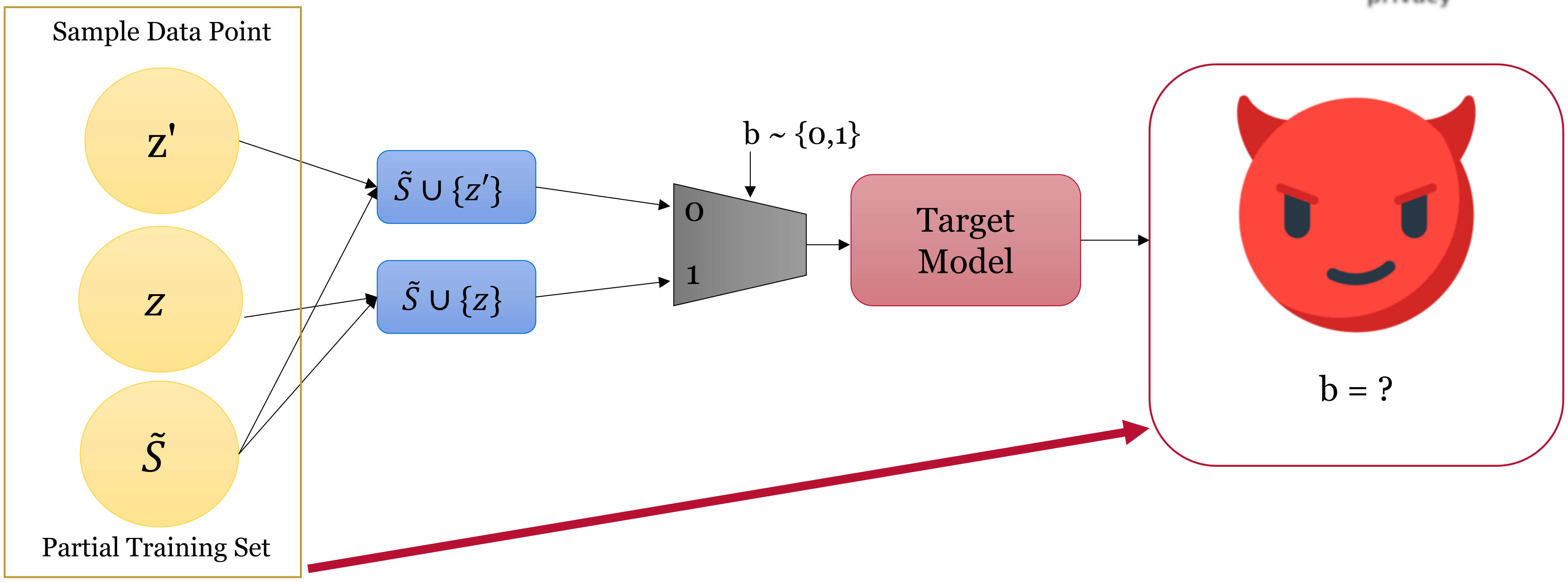
Strong Adversary Membership Experiment





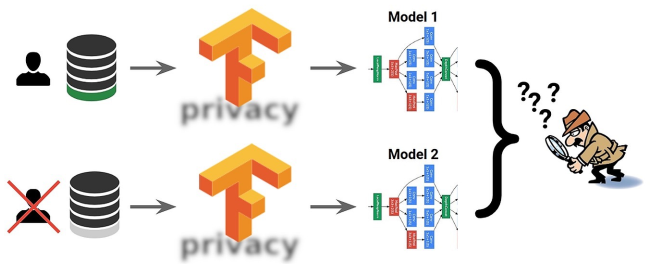
Strong Adversary Membership Experiment

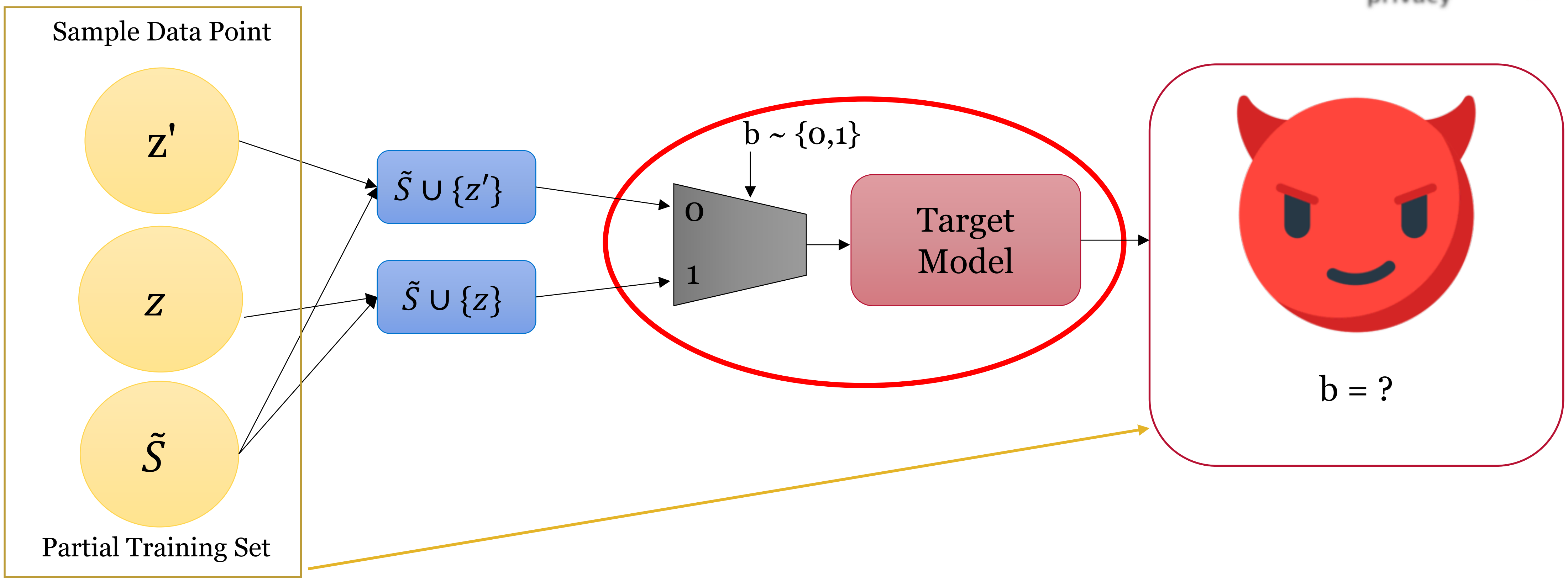




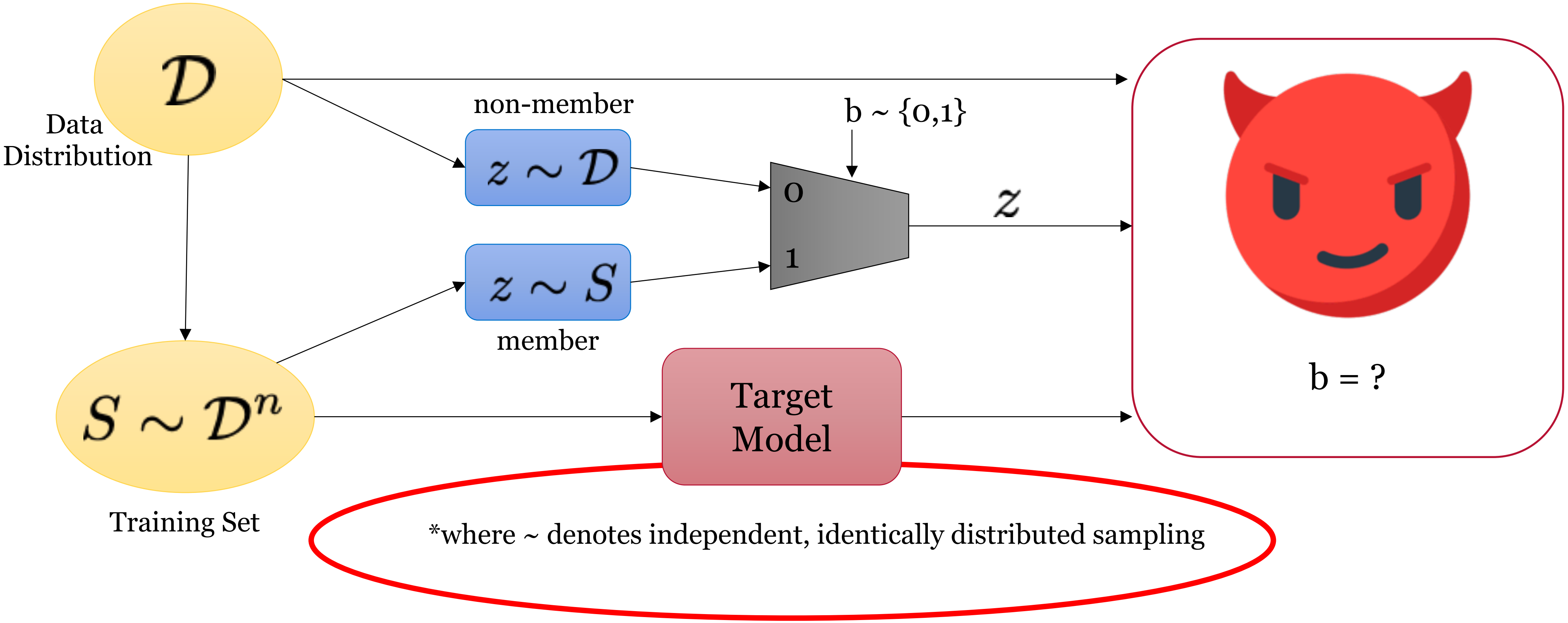
UNIVERSITY OF WATERLOO

Strong Adversary Membership Experiment



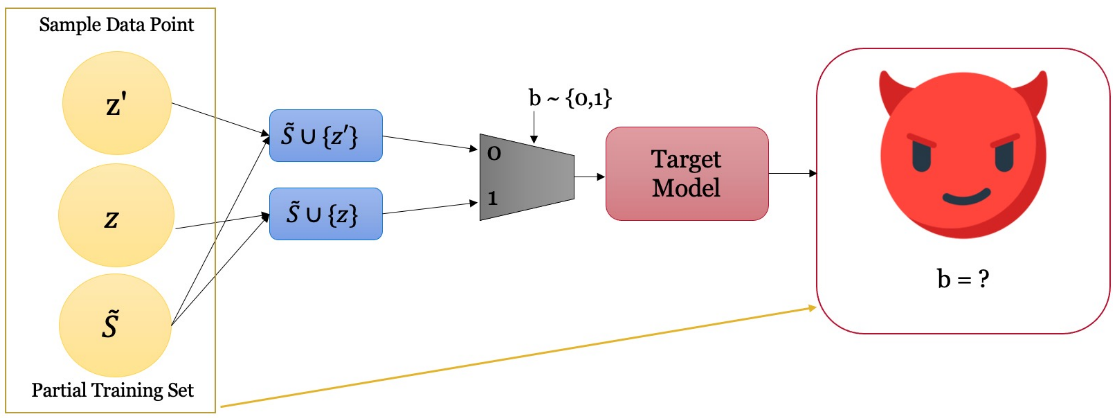


Yeom et al.'s Membership Experiment

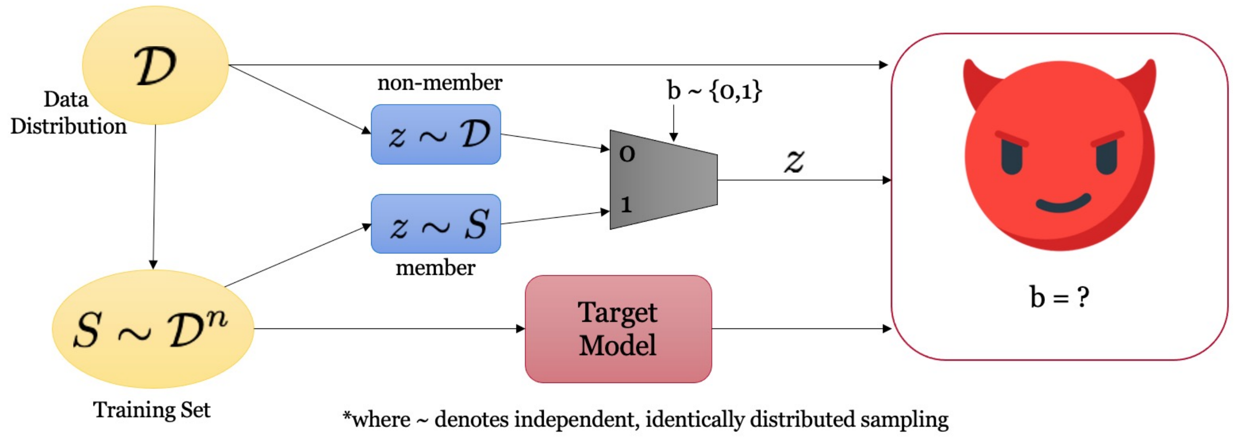


Under IID Assumption the Attackers are Similar

Strong Adversary Membership Experiment



Yeom et al.'s Membership Experiment



A DP bound on the strong adversary implies a bound on the MIA adversary

Tighter Bound

$$Adv \leq \frac{e^{\epsilon} - 1 + 2\delta}{e^{\epsilon} + 1}$$

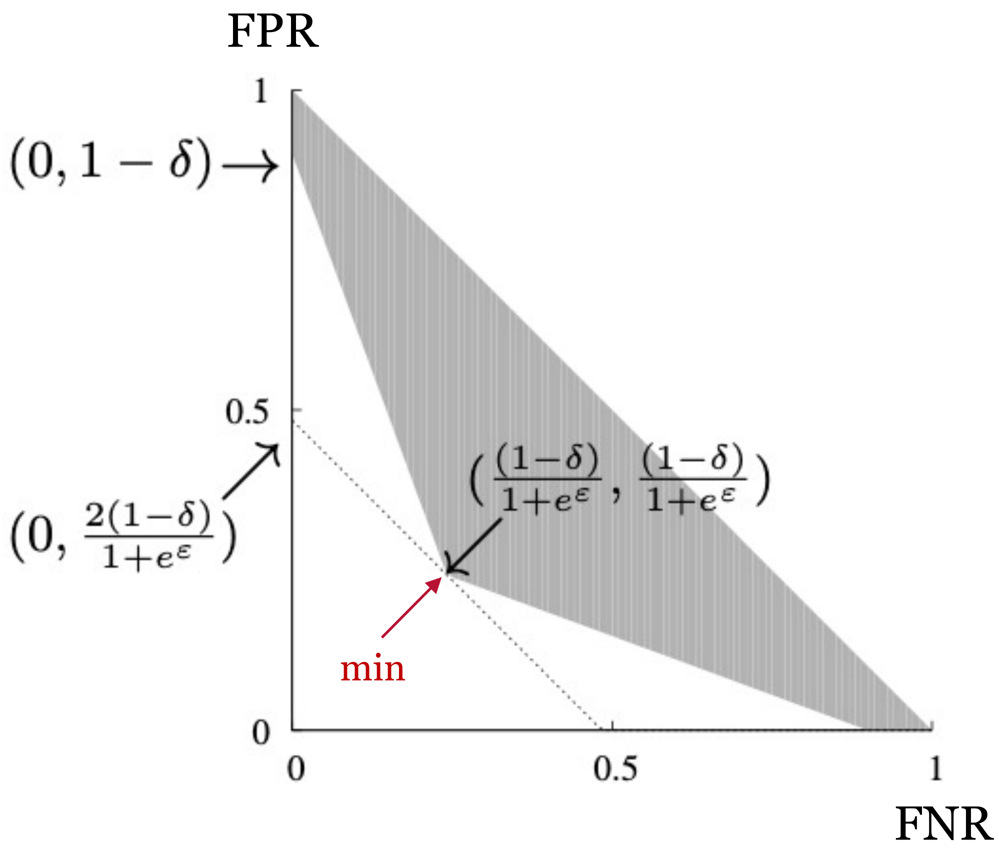


IMAGE CREDIT: Kairouz et al. 2017

The Composition Theorem for Differential Privacy

Proof Sketch:

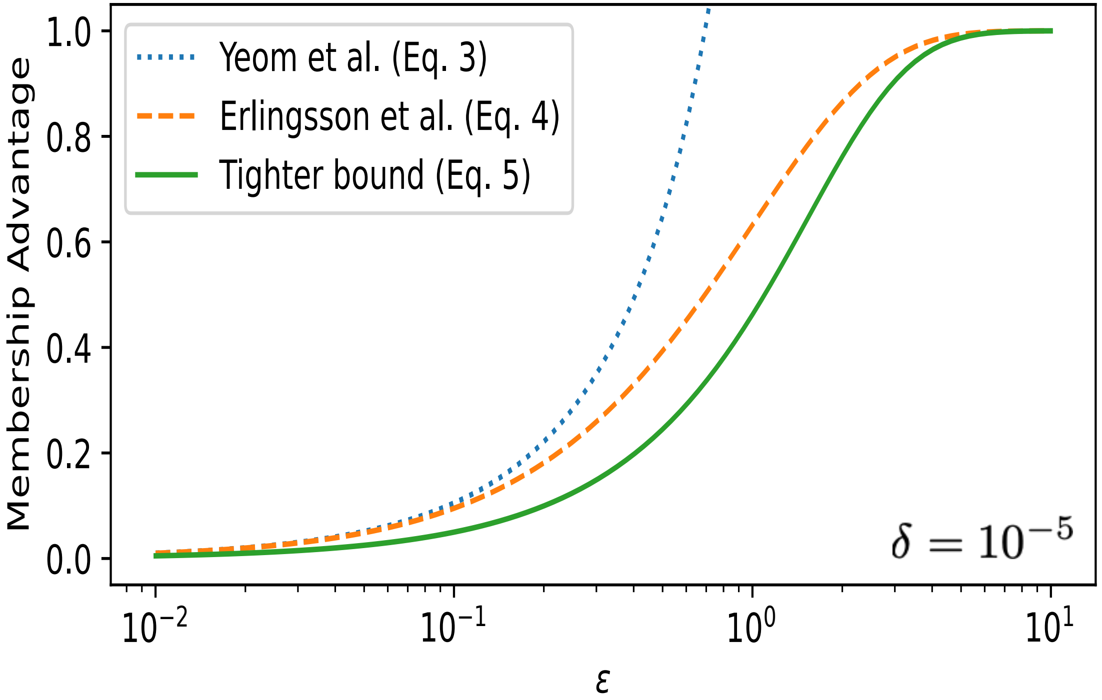
$$Adv \leq max\{TPR - FPR\}$$

$$= 1 - min\{FNR + FPR\}$$

$$= 1 - \frac{2(1-\delta)}{1+e^{\epsilon}}$$

UNIVERSITY OF WATERLOO

Comparing the Bounds

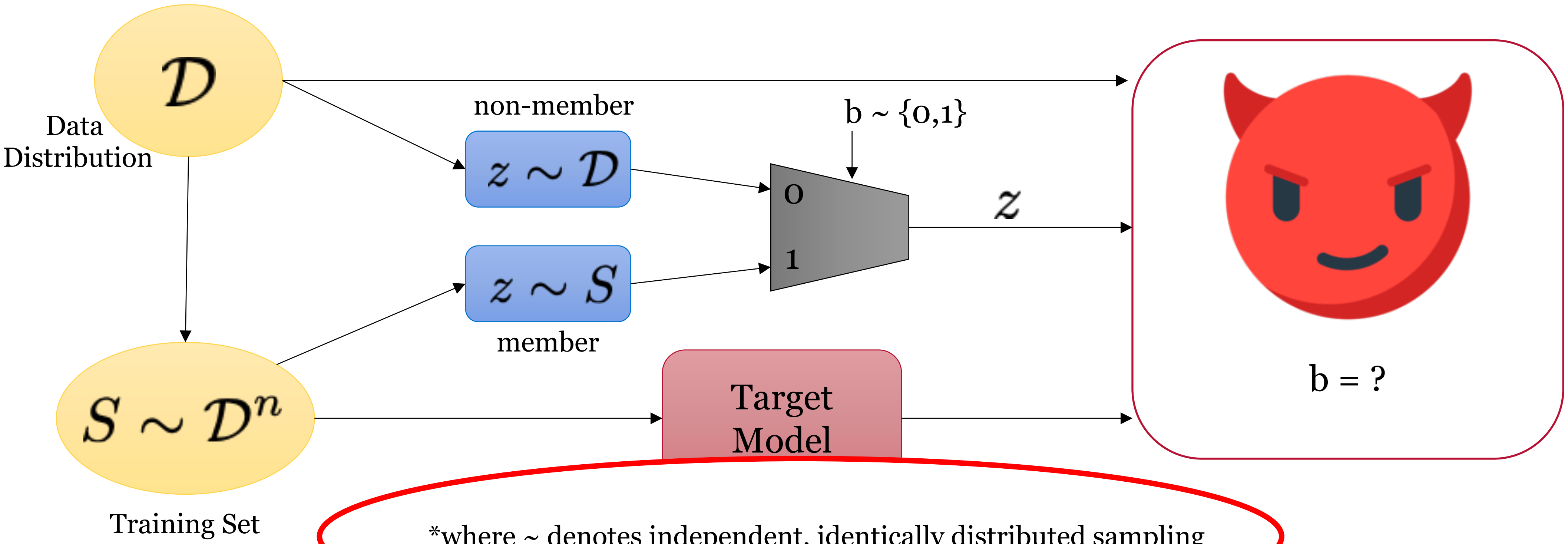


$$Adv \leq e^{\epsilon} - 1 \tag{3}$$

$$Adv \leq 1 - e^{-\epsilon}(1 - \delta) \tag{4}$$

$$Adv \leq \frac{e^{\epsilon} - 1 + 2\delta}{e^{\epsilon} + 1} \tag{5}$$

What is wrong with Yeom et al.'s Membership Experiment?



*where ~ denotes independent, identically distributed sampling

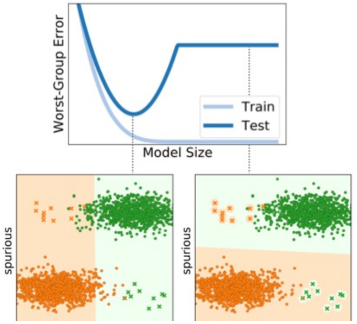
Biased Data in ML

An Investigation of Why Overparameterization Exacerbates Spurious Correlations

Shiori Sagawa [*1] Aditi Raghunathan [*1] Pang Wei Koh [*1] Percy Liang [1]

Abstract

We study why overparameterization—increasing model size well beyond the point of zero training error—can hurt test error on minority groups despite improving average test error when there are spurious correlations in the data. Through simulations and experiments on two image datasets, we identify two key properties of the training data that drive this behavior: the proportions of majority versus minority groups, and the signal-to-noise ratio of the spurious correlations. We then analyze a linear setting and theoretically show how the inductive bias of models towards "memorizing" fewer examples can cause overparameterization



Research Track Paper KDD '20, August 23–27, 2020, Virtual Event, USA

Targeted Data-driven Regularization for Out-of-Distribution Generalization

Mohammad Mahdi Kamani, Sadegh Farhang, Mehrdad Mahdavi and James Z. Wang
{mqk5591,smf5604,mzm616,jwang}@psu.edu
The Pennsylvania State University, University Park, Pennsylvania

ABSTRACT

Due to biases introduced by large real-world datasets, deviations of deep learning models from their expected behavior on out-of-distribution test data are worrisome. Especially when data come from imbalanced or heavy-tailed label distributions, or minority groups of a sensitive feature. Classical approaches to address these biases are mostly data- or application-dependent, hence are burdensome to tune. Some meta-learning approaches, on the other hand, aim to learn hyperparameters in the learning process using different objective functions on training and validation data. However, these methods suffer from high computational complexity and are not scalable to large datasets. In this paper, we propose a unified data-driven regularization approach to learn a generalizable model from biased data. The proposed framework, named as **targeted data-driven regularization** (TDR), is model- and dataset-agnostic, and employs a target dataset that resembles the desired nature of test data in order to guide the learning process in a coupled manner. We cast the problem as a bilevel optimization and propose an efficient stochastic gradient descent based method to solve it. The framework can be utilized to alleviate various types of biases in real-world applications. We empirically show, on both synthetic and real-world datasets, the superior performance of TDR for resolving

1 INTRODUCTION

Drastically improving their performance, machine learning, and more distinctively, deep learning models, are becoming the main propulsion of technology in a variety of domains. Notwithstanding their success, they still suffer from different forms of biases in the training data distribution. Biases, regardless of their nature, cause a mismatch between training and testing data distributions, which leads to a poor out-of-distribution generalization performance of the model. Machine learning models inherit these biases due to the only objective of minimizing the empirical risk on the training data in their learning process. However, empirical risk by itself seems incapable of avoiding these biases in training data for better out-of-distribution generalization, and needs to be accompanied by other objectives [35].

These biases can appear in different forms in training a machine learning model. A palpable form of them happens when the size of different classes or groups are unbalanced. When class sizes are not balanced, the imbalanced dataset problem stems [9, 24, 44], where majority classes' distribution can dominate the training process, resulting in a model with low accuracy on minority classes. A severe form of imbalanced dataset problem, appears in most real-world big datasets with immense number of classes, is long-tailed

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

who is hired, fired, granted a loan, or how long

DP under non-IID data

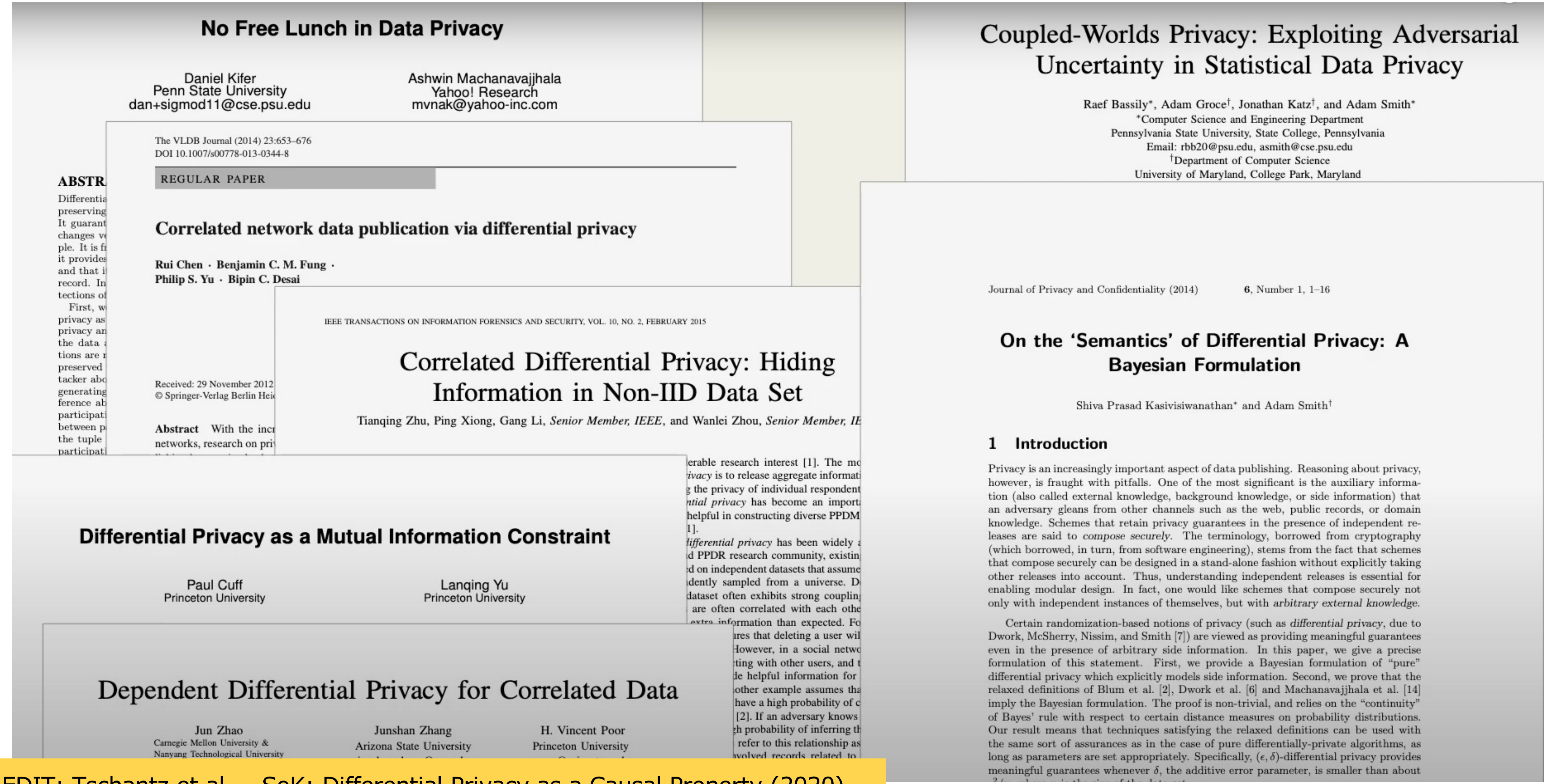
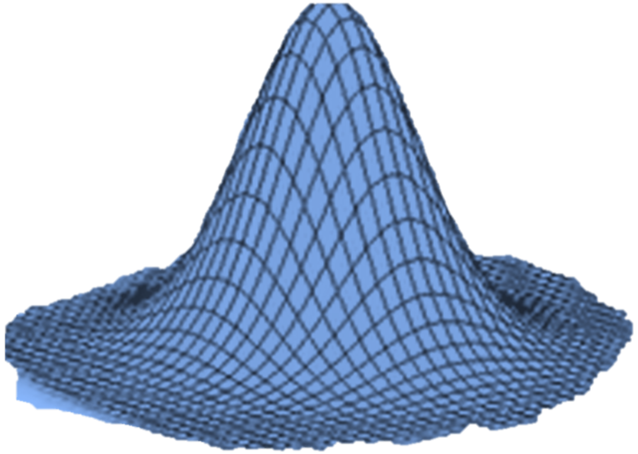


IMAGE CREDIT: Tschantz et al. – SoK: Differential Privacy as a Causal Property (2020)

UNIVERSITY OF WATERLOO

Relaxing the IID Assumption:

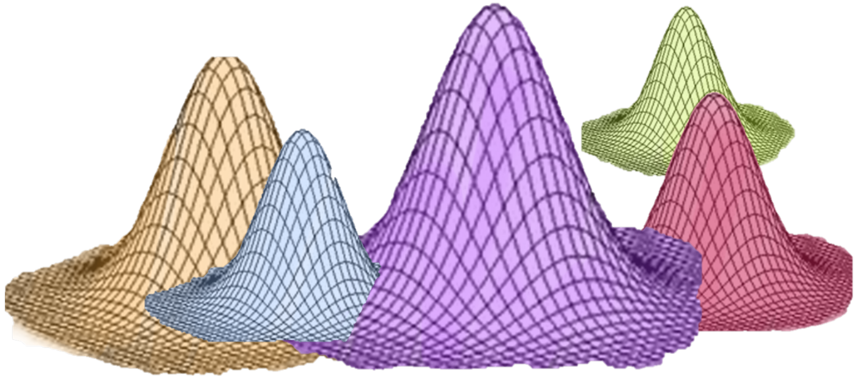
$\mathcal{D}$



Feature and Label Space

$\mathbb{D}$

multivariate mixture model

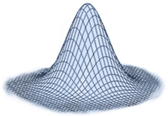
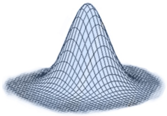


$$\mathbb{D} = \{D_1, D_2, \ldots, D_K\}$$

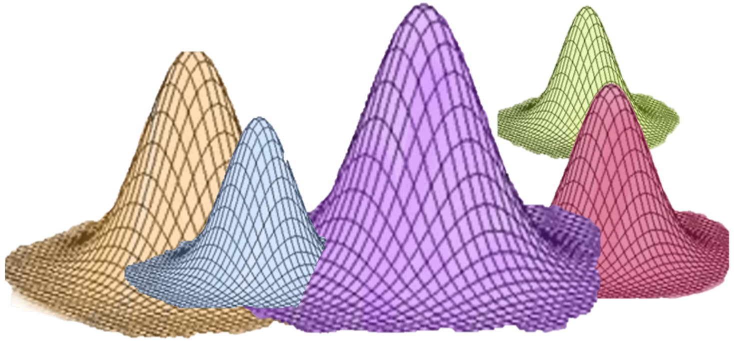
UNIVERSITY OF WATERLOO

Relaxing the IID Assumption:

How to sample $S \sim \mathbb{D}$

- Choose a subpopulation at random (e.g., )
- Sample S from 

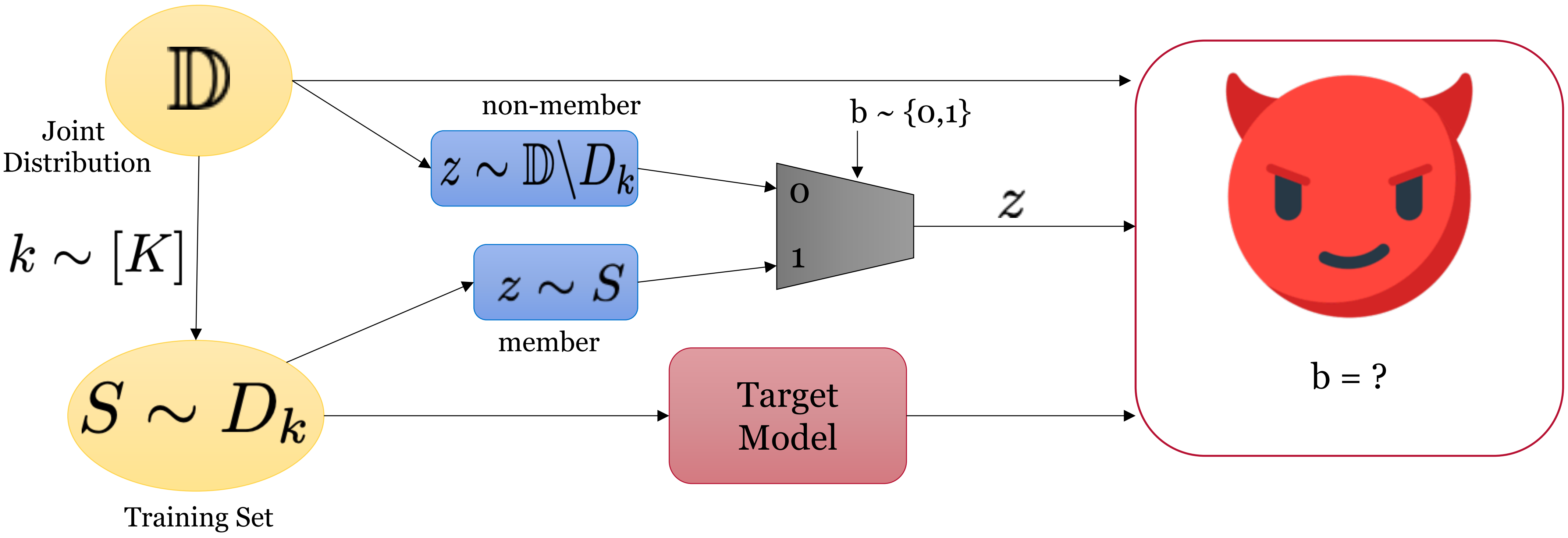
$\mathbb{D}$



$$\mathbb{D} = \{D_1, D_2, \ldots, D_K\}$$

Generalized Membership Experiment

$$\mathbb{D} = \{D_1, D_2, \ldots, D_K\}$$



UNIVERSITY OF WATERLOO

Questions About This Experiment

Is there a provable bound on the attack by DP learning?

- Yes, but
- No longer bounded by the strong adversary experiment.
- It requires noise proportional to the size of the data set: $n \cdot \varepsilon$
  - A tighter bound may exist

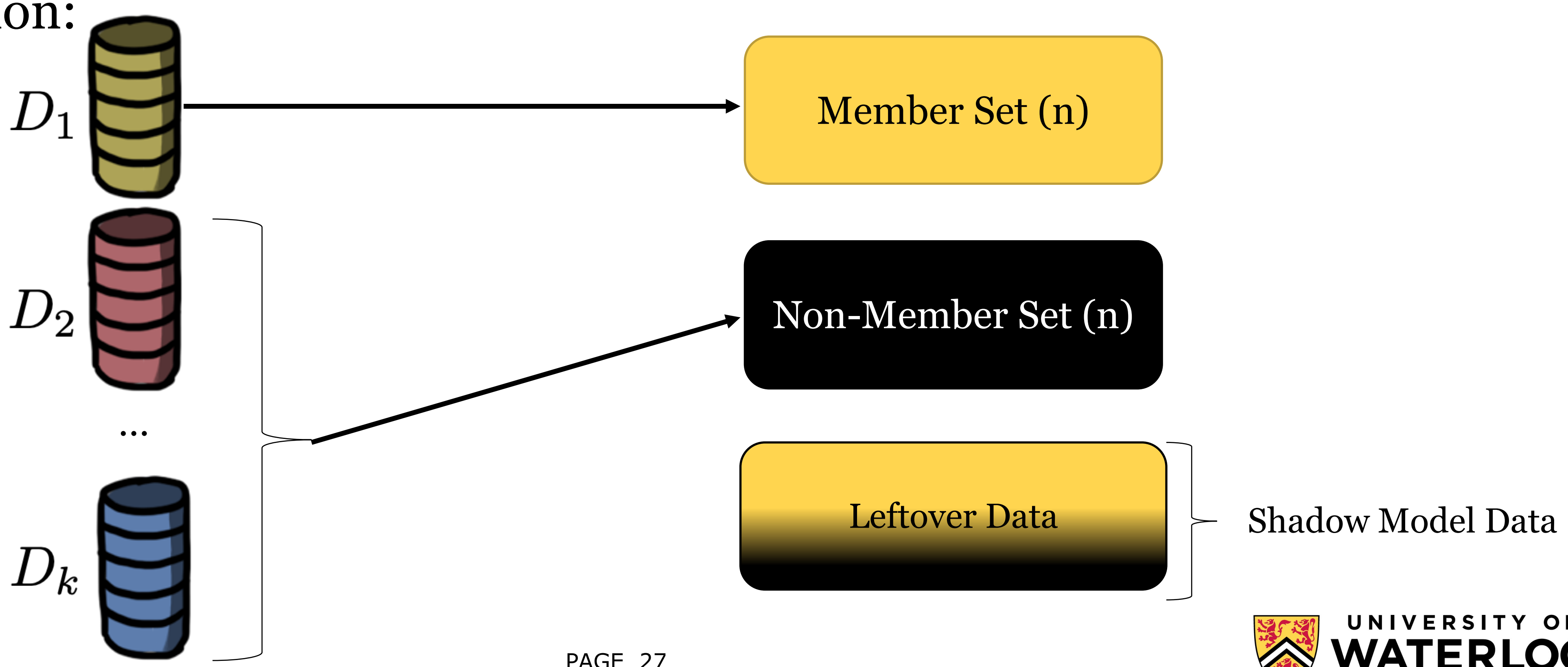
Could the adversary use the distribution information in attack?

- Yes, but
- In practice, the mixture components may have overlapping support
  - The $n \cdot \varepsilon$-DP bound would still hold
- The distribution information could be removed in a further refinement

EMPIRICAL EVALUATION

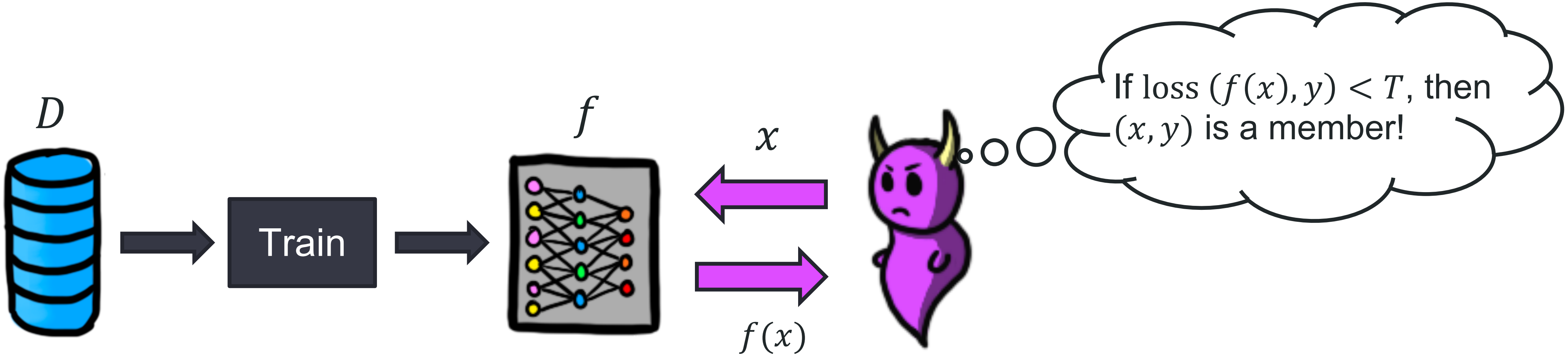
Experimental Setup

- We use the source code from Jayaraman and Evans (only RDP)
- Off the shelf ML datasets (e.g. UCI ML Repository)
- Data Curation:



Unmodified MIA Attacks – Yeom et al.'s Threshold Attack

- *Idea:* the model will have a lower loss on members of the training set.

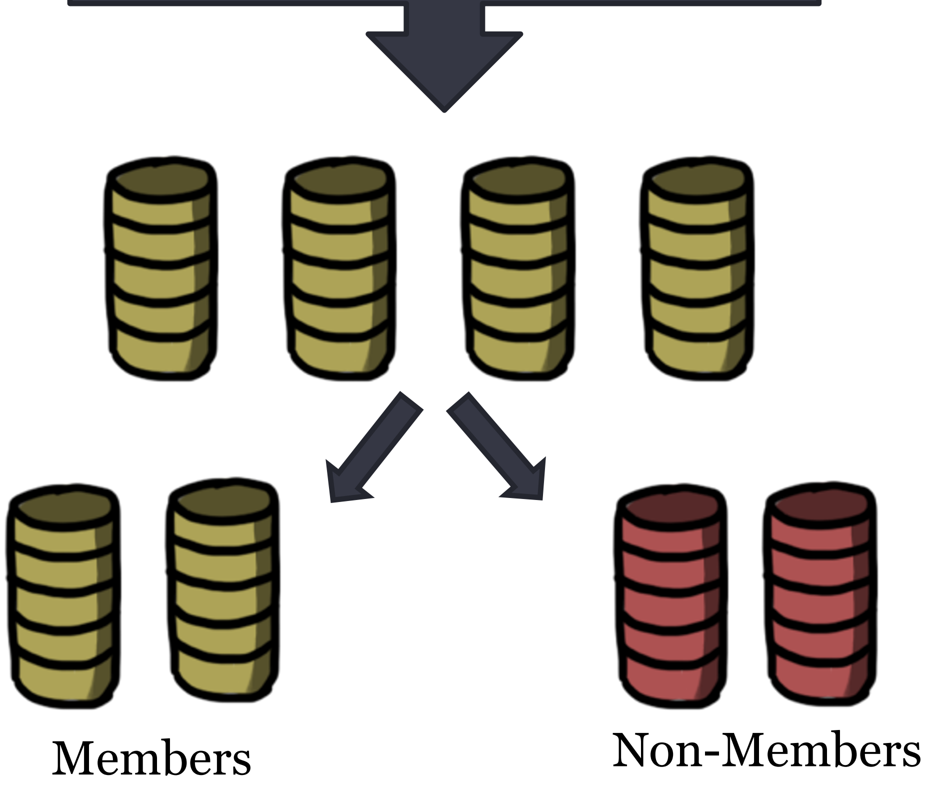


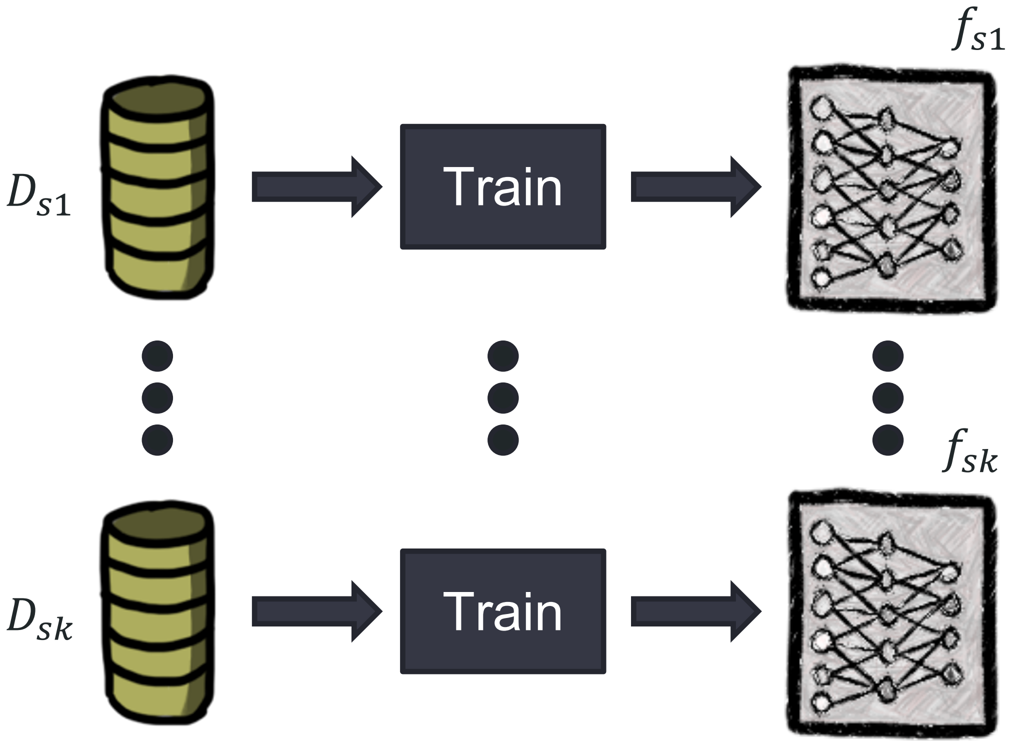
UNIVERSITY OF WATERLOO

Unmodified MIA Attacks – Shokri et al.'s Shadow Model Attack

- Train k **shadow models** $f_{s1}, \dots, f_{sk}$ (same classification task as the target model).

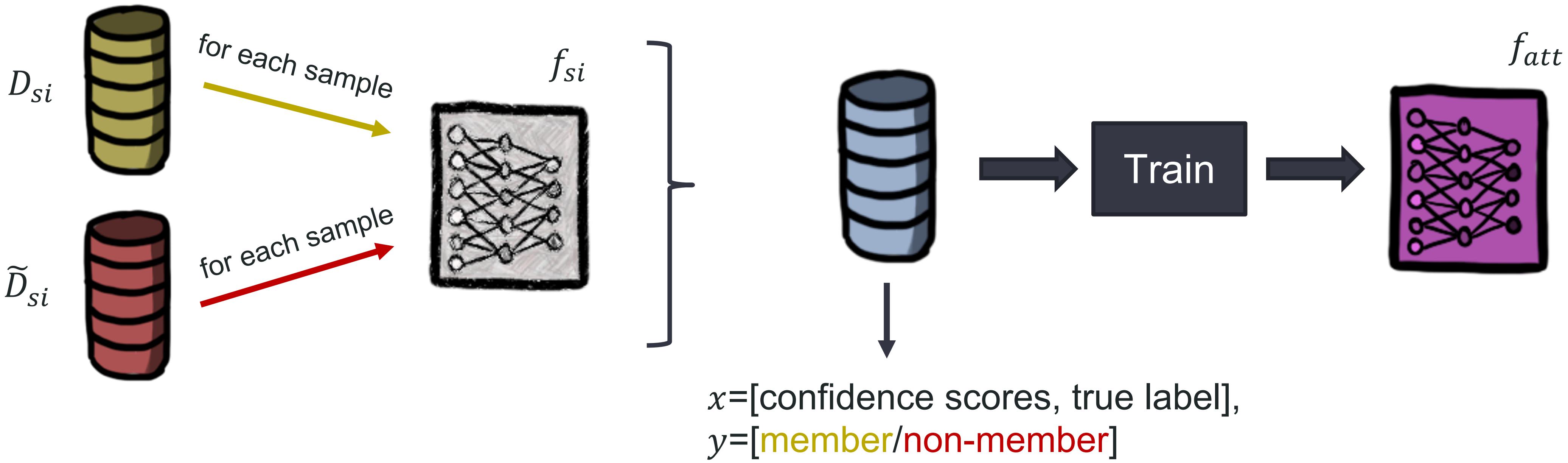
Assumption: the adversary has data with a similar distribution.





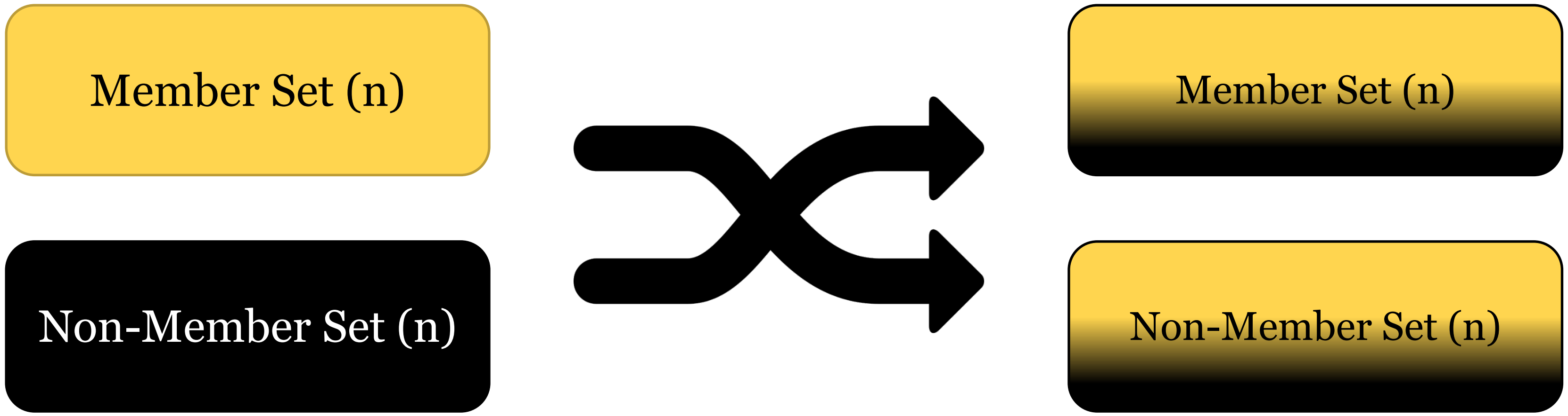
Unmodified MIA Attacks – Shokri et al.'s Shadow Model Attack

Train a new **attack model** f_{att} to predict the "membership status" from "confidence scores, true label"



UNIVERSITY OF WATERLOO

Simulating the IID Case



Inherit Dependencies- Hospital Data

Each mixture component corresponds to a single hospital

D_1 = Members D_2 = Members D_3 = Members D_4 = Members

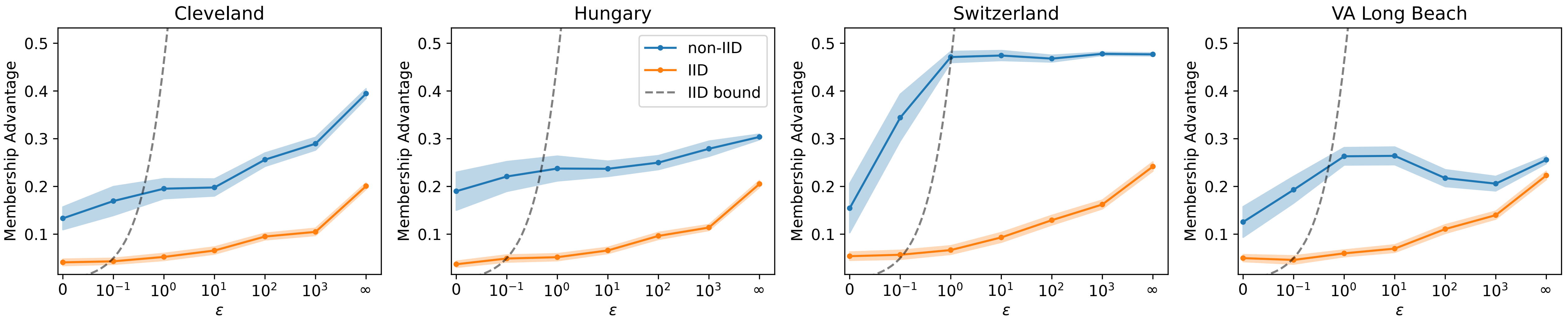


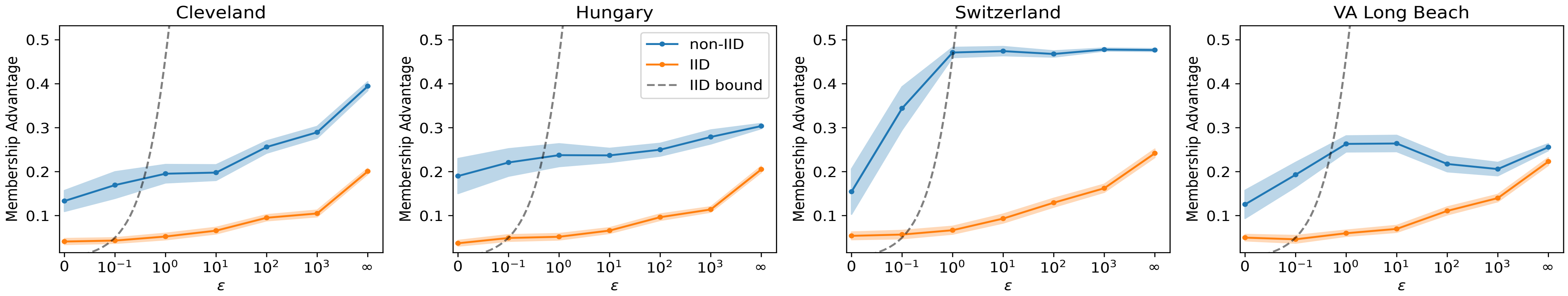
Fig. 6: Performance of the optimal threshold attack in the heart dataset when members belong to a particular database (hospital/institution), and non-members are taken from all other databases

Recall:

- We used unmodified membership inference attacks.
  - The attack has no background information on the distribution of members and non-members
- We used real-world data sets from the web, standard machine learning training

Clearly: The attacks exceed the bound for non-iid data

Conclusion: Differential Privacy does not protect as expected



UNIVERSITY OF WATERLOO

Inherit Dependencies- Texas Hospital Data

D_1 = Hospitals in region 3 $\quad$ D_2 = All other regions

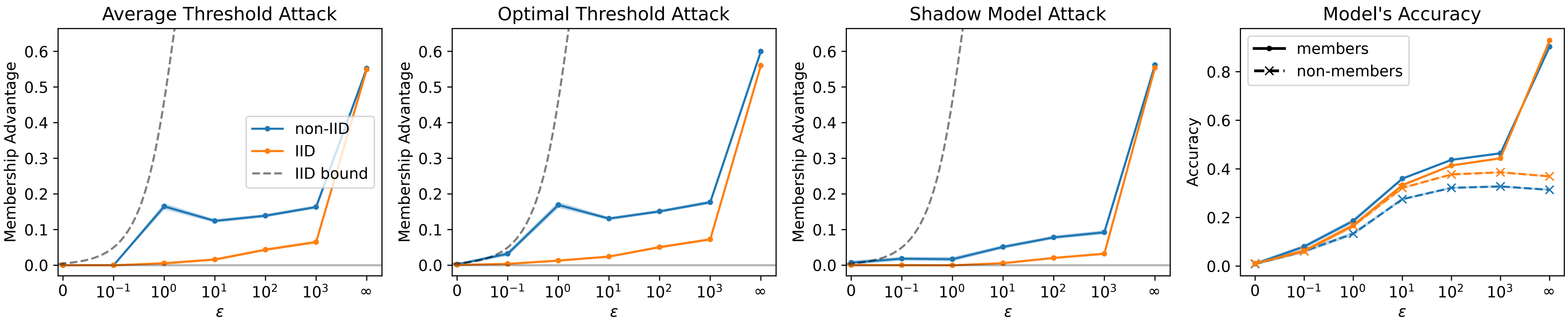


Fig. 8: Results in texas dataset when members are from hospitals in region 3, and non-members are from any other region.

UNIVERSITY OF WATERLOO

Inherit Dependencies- Census Data

$D_1 = \text{Census Dataset}$ $\quad D_2 = \text{Adult Dataset}$

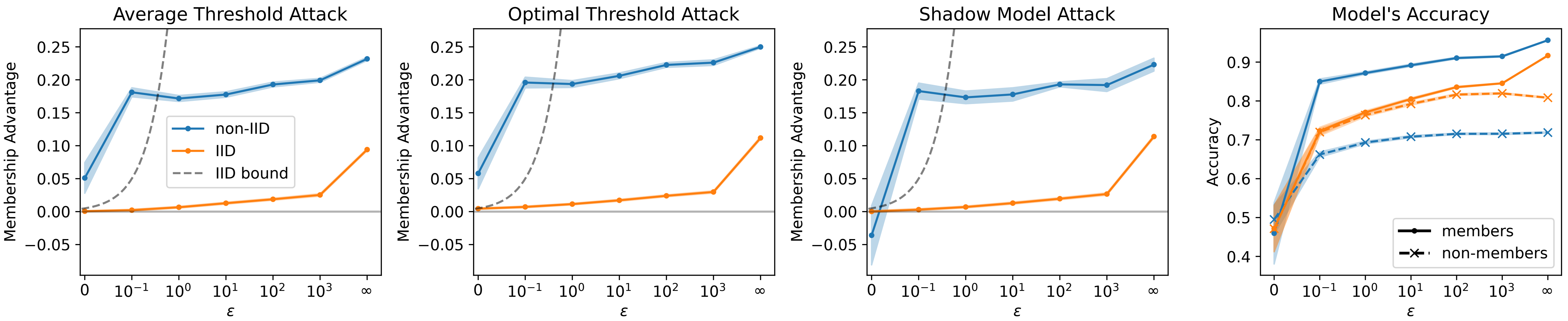
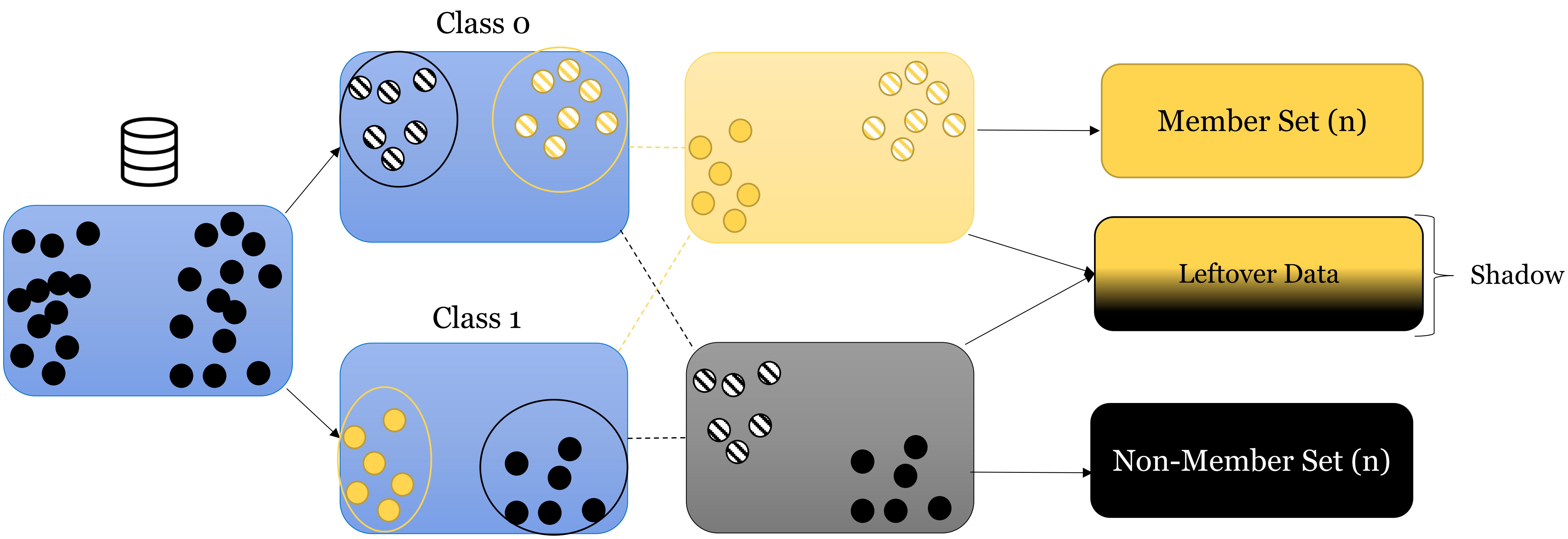


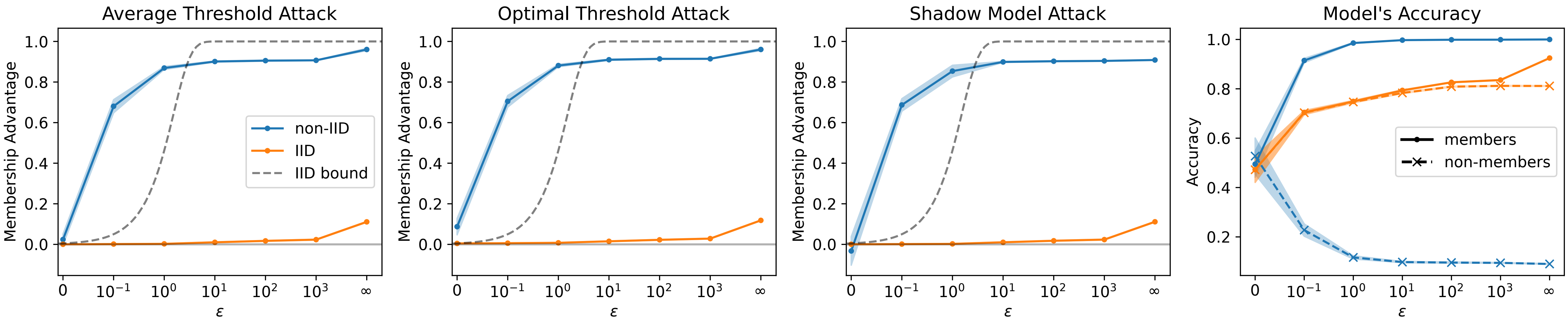
Fig. 9: Results when training a model with census, where non-members are from adult dataset.

WORST CASE EVALUATION

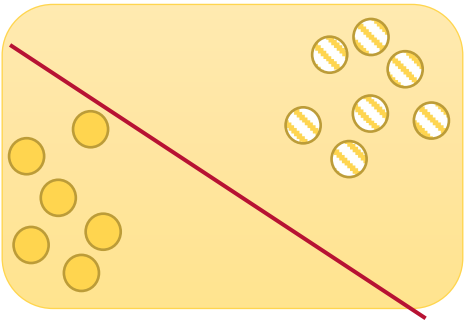
K-Means Split

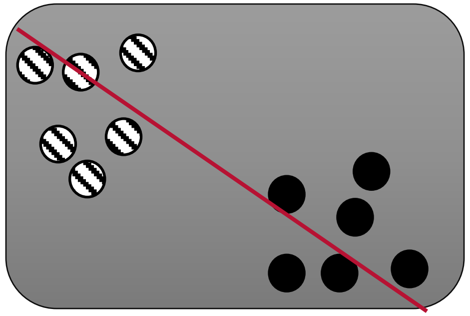


Cluster Split - Adult



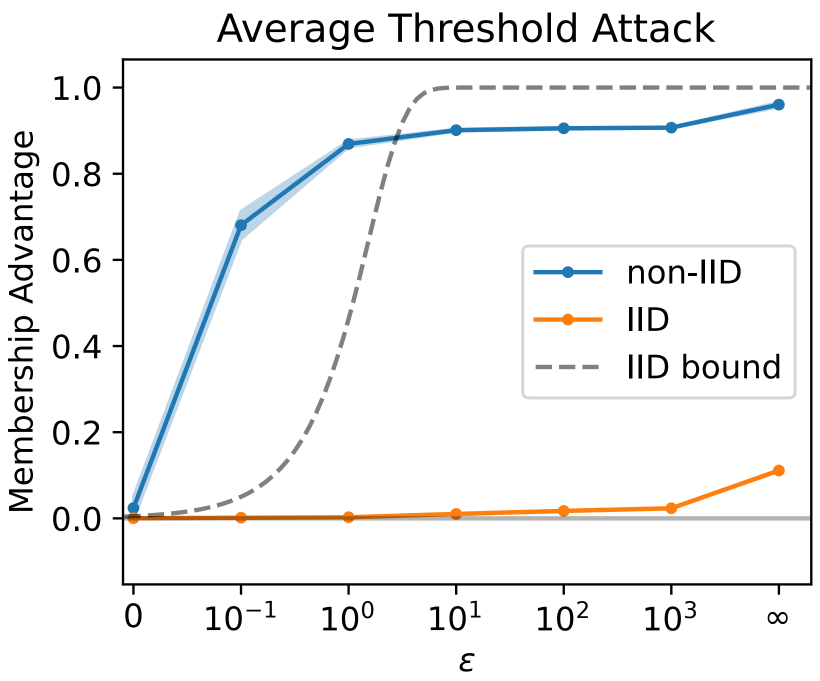
Example:

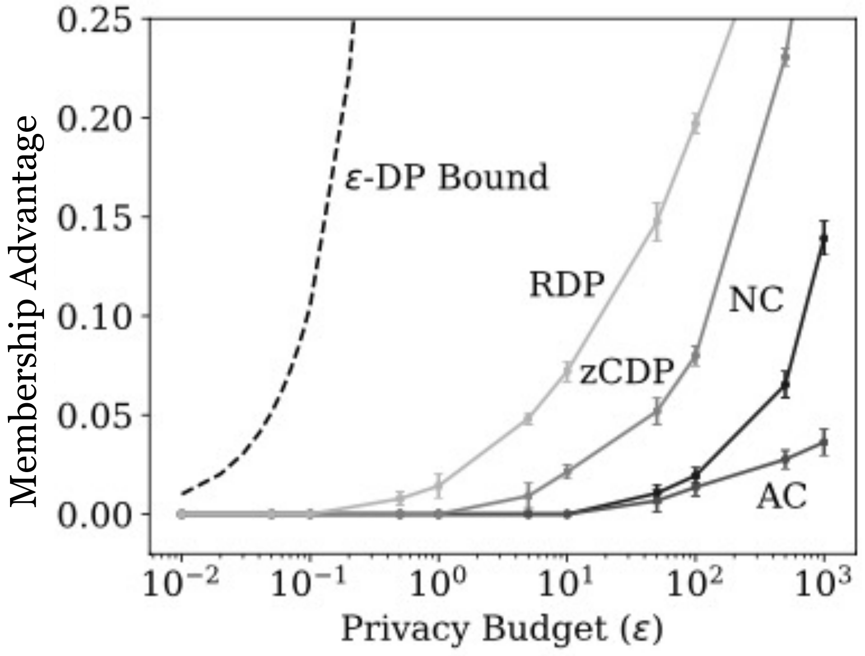




UNIVERSITY OF WATERLOO

The Gap Observed in the Literature - Revisited





(b) Yeom et al. membership inference

IMAGE CREDIT: Jayaraman and Evans - Evaluating Differentially Private Machine Learning in Practice (2019)

Conclusions

- We provide a more general membership experiment.
- We have shown that off-the-shelf attacks can break the bounds of DP.
- Data dependencies can cause much higher privacy leakage than previously reported.
- The IID assumption is an integral component of past results upholding the integrity of DPML at high epsilon
  - Tools such as ML privacy meter do not give the full picture

"Data dependencies should be taken into account when studying MIA performance, as they are a realistic assumption that, if ignored, can lead to a significant underestimation of the privacy risk that MIAs pose".

UNIVERSITY OF
WATERLOO

THANK YOU!