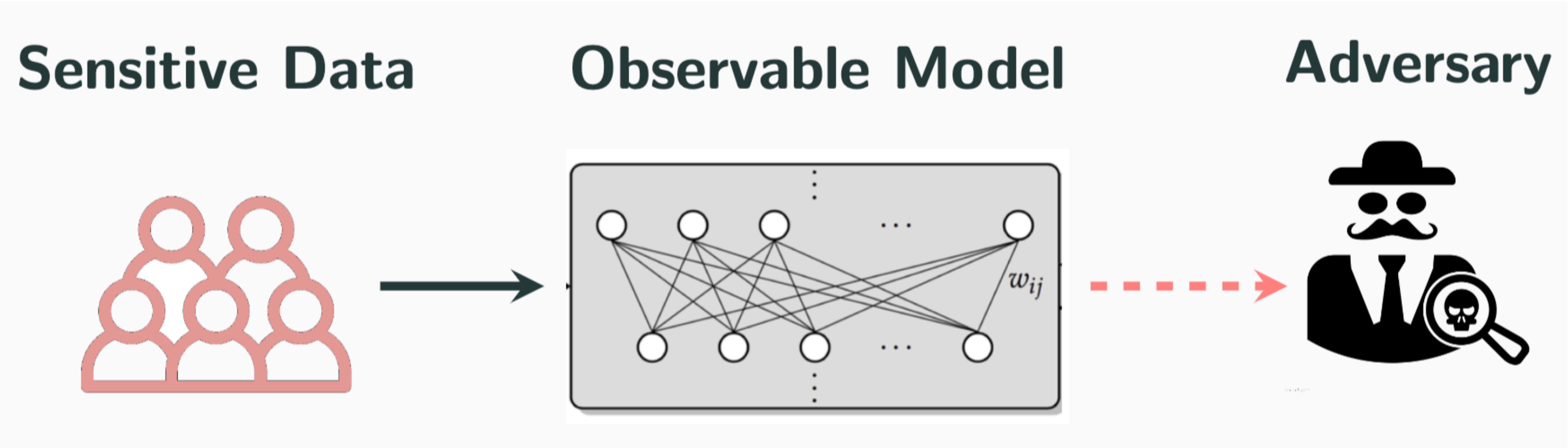
Enhanced Membership Inference Attacks against Machine Learning Models

Jiayuan Ye[1], Aadyaa Maddi[1], Sasi Kumar Murakonda[2],
Vincent Bindschaedler[3], Reza Shokri[1]

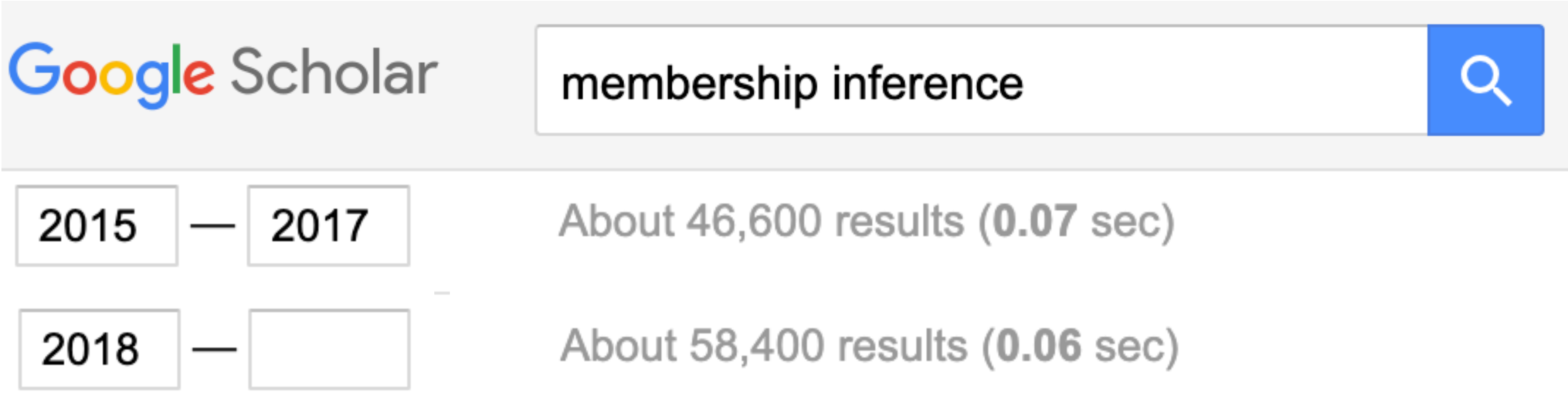
[1]National University of Singapore [2]Privitar Labs [3]University of Florida

Membership Inference



Does the sensitive dataset contain a given person's record?

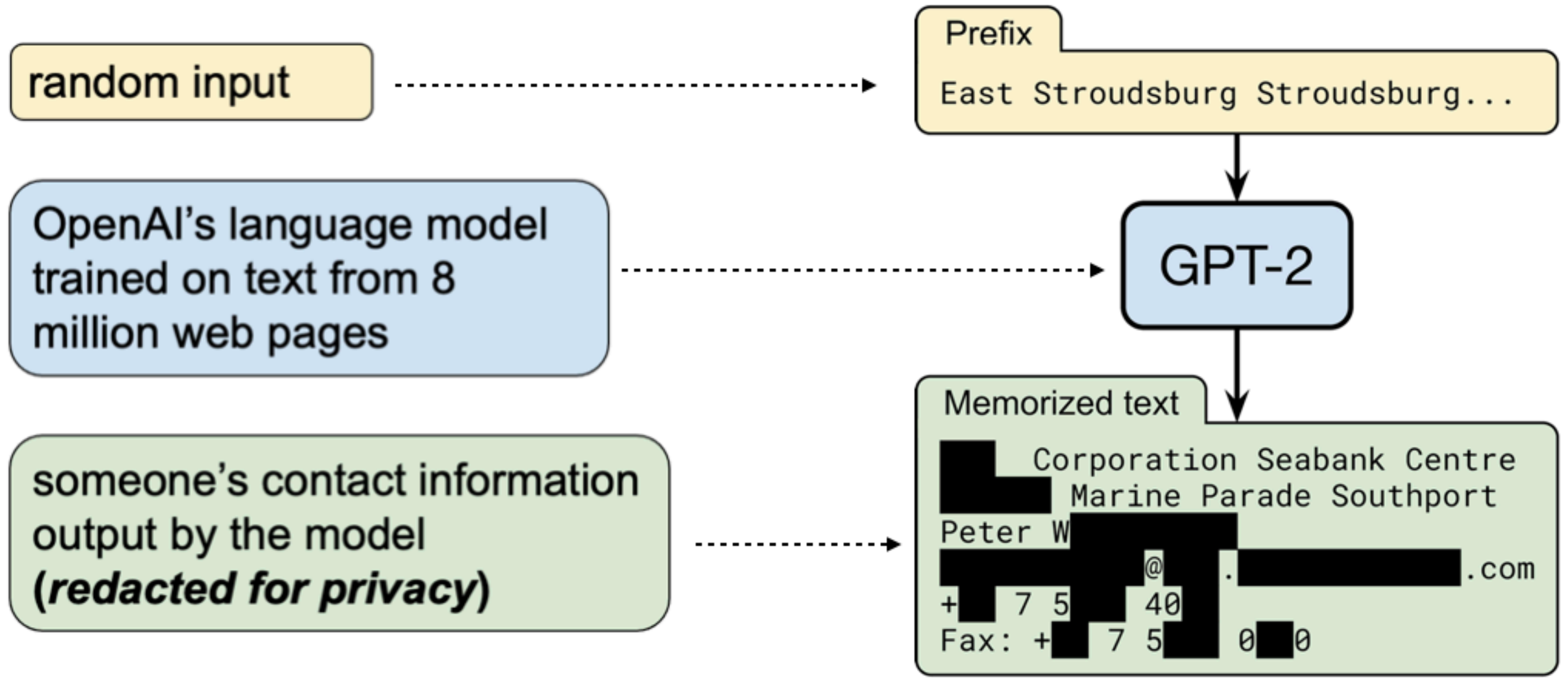
Widely studied in machine learning



Google Scholar membership inference

2015 — 2017	About 46,600 results (**0.07** sec)
2018 —	About 58,400 results (**0.06** sec)

- Could serve as the base for stronger attacks



random input

Prefix

```
East Stroudsburg Stroudsburg...
```

OpenAI's language model trained on text from 8 million web pages

GPT-2

someone's contact information output by the model ***(redacted for privacy)***

Memorized text

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Data Reconstruction

- Used for auditing different kinds of leakage

Model Privacy

Algorithm

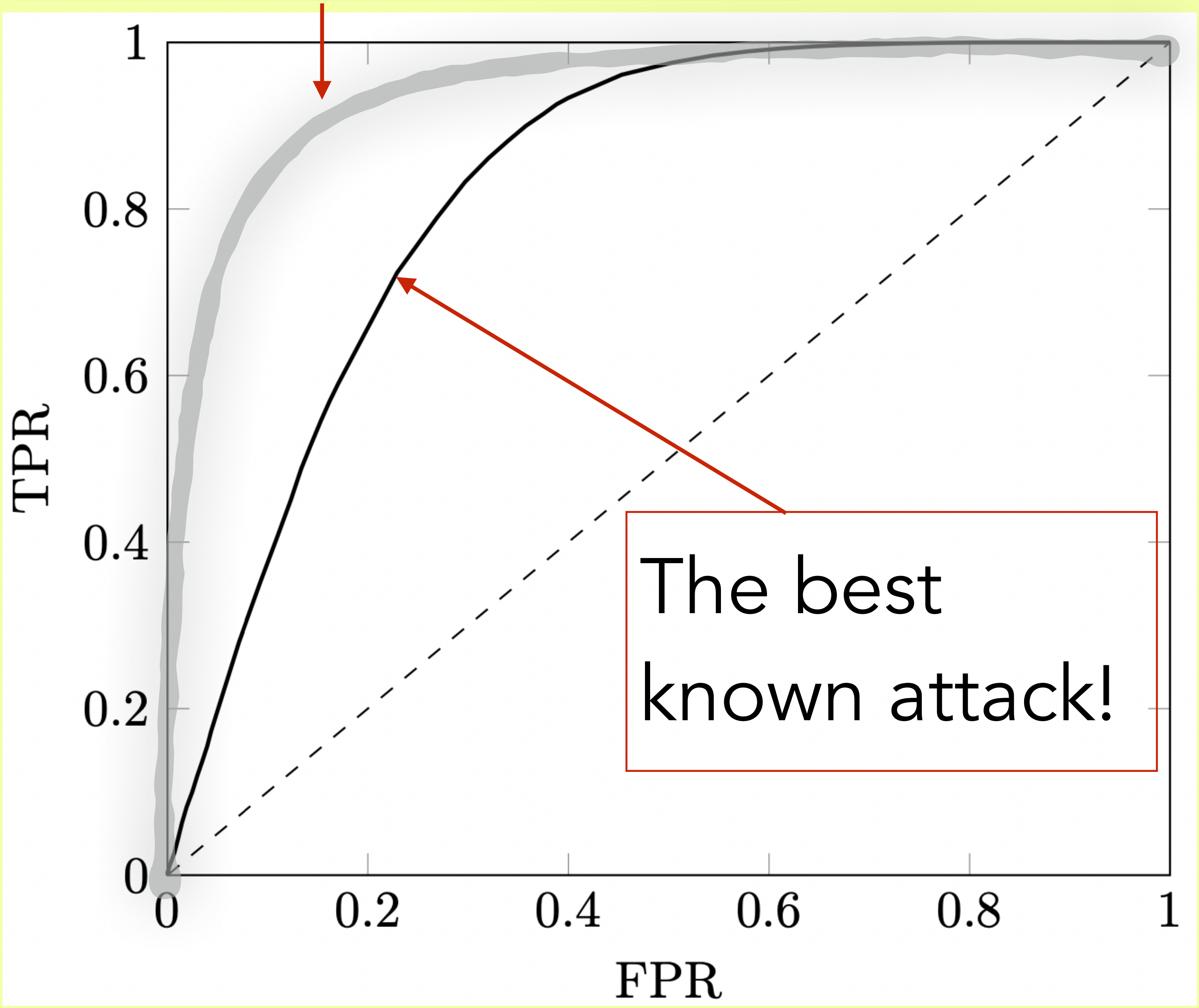
Data Memorization

Issues with existing MIA

Belief: success of attacker is a metric for privacy loss

? Success over what records or models?
How to interpret different success rate?

Inconsistencies in formalizing the problem



An (unknown) **optimal** attack

The best known attack!

Inadequate attack performance

Overfitting?

Memorization?

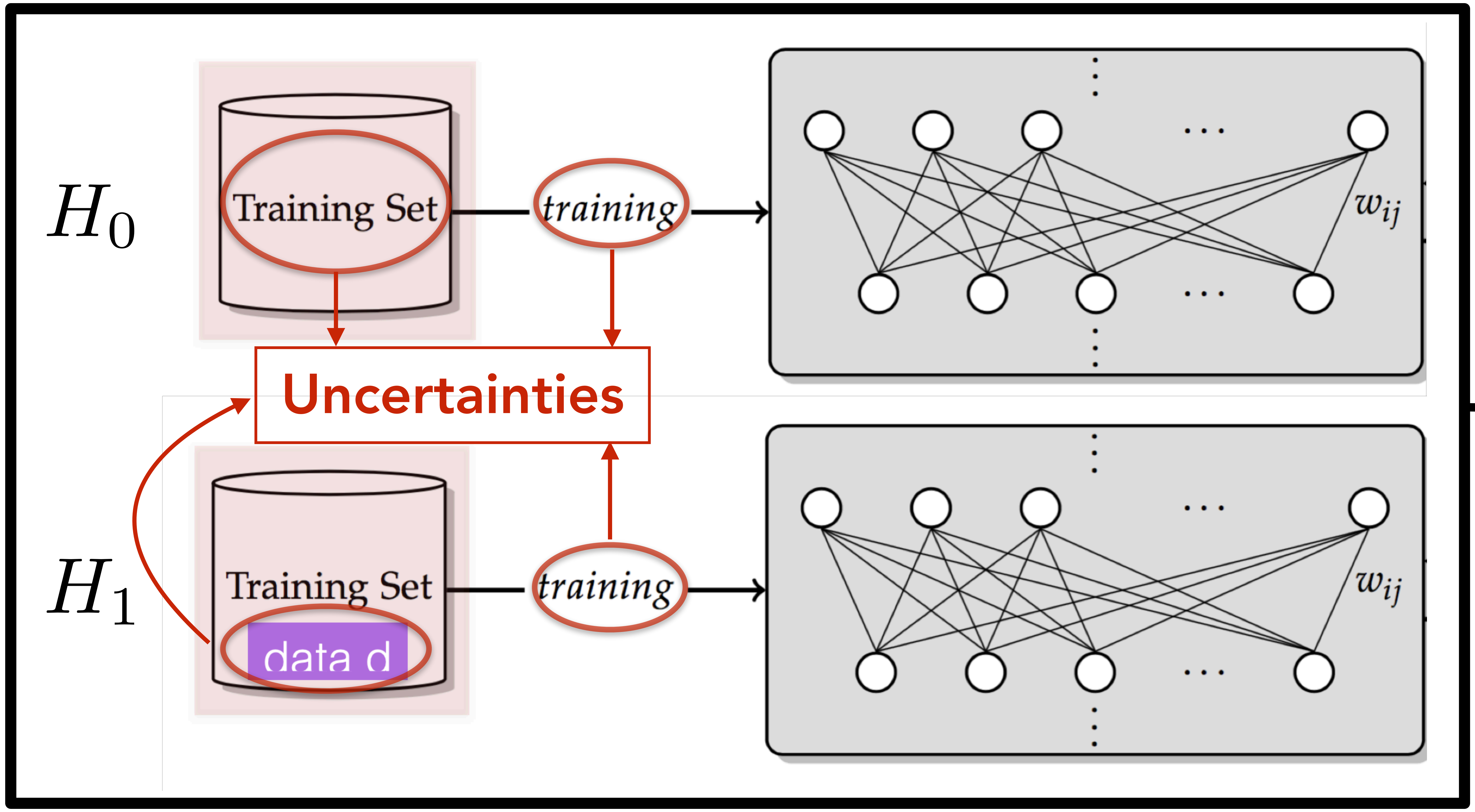
Latent neighbor?

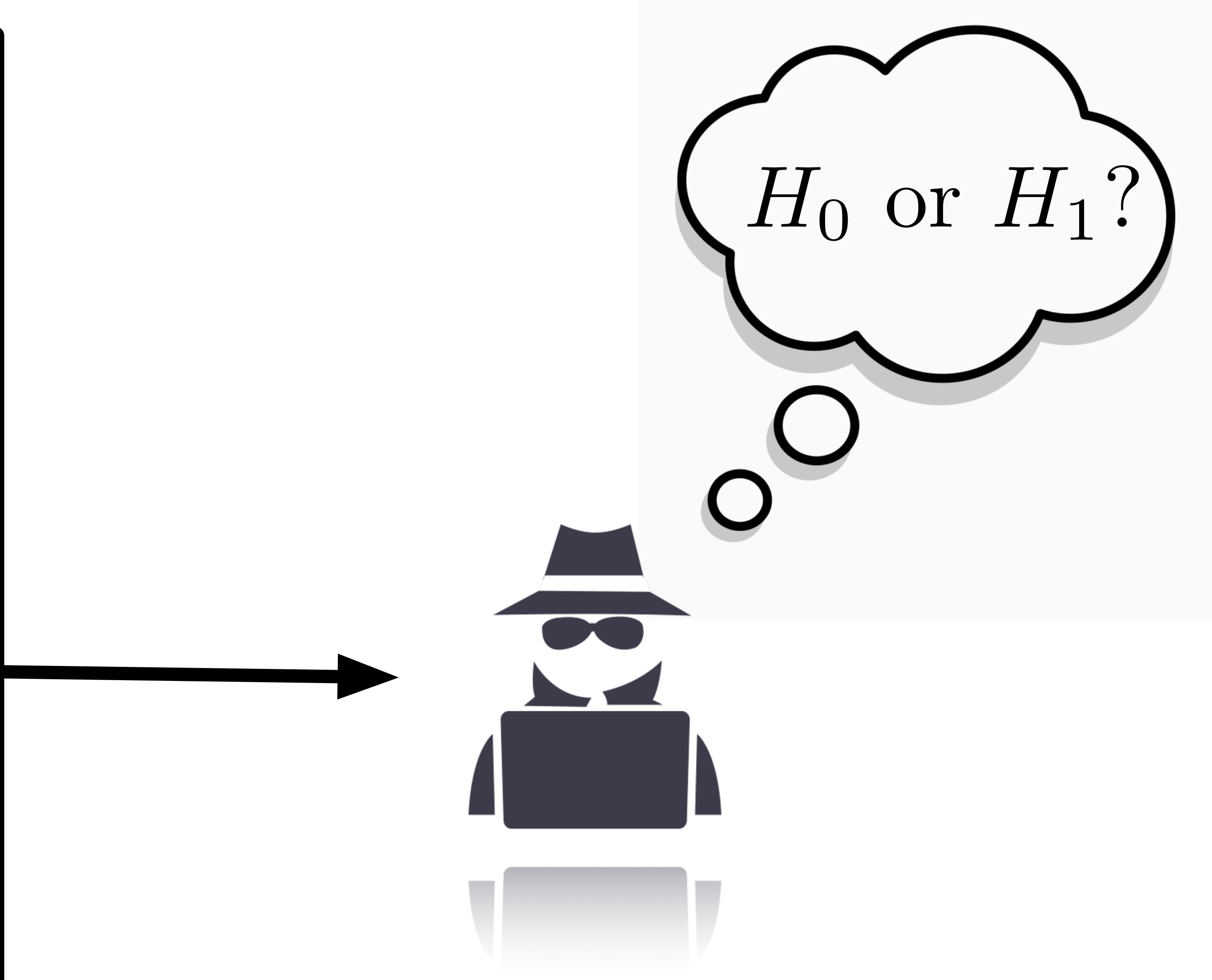
Lack of explanation for the leakage

Contributions

- Explain games in which different kinds of leakage could be quantified
- Formalize prior attack in this consistent framework
- Design attack stronger than prior attacks in this framework, via approximating an optimal attack that minimizes adversary's uncertainty

Membership Inference Attack (MIA) Game





- Black-box Access to Target Model
- Access to a population data pool

Prior works largely fomulates MIA game when all the components are randomized

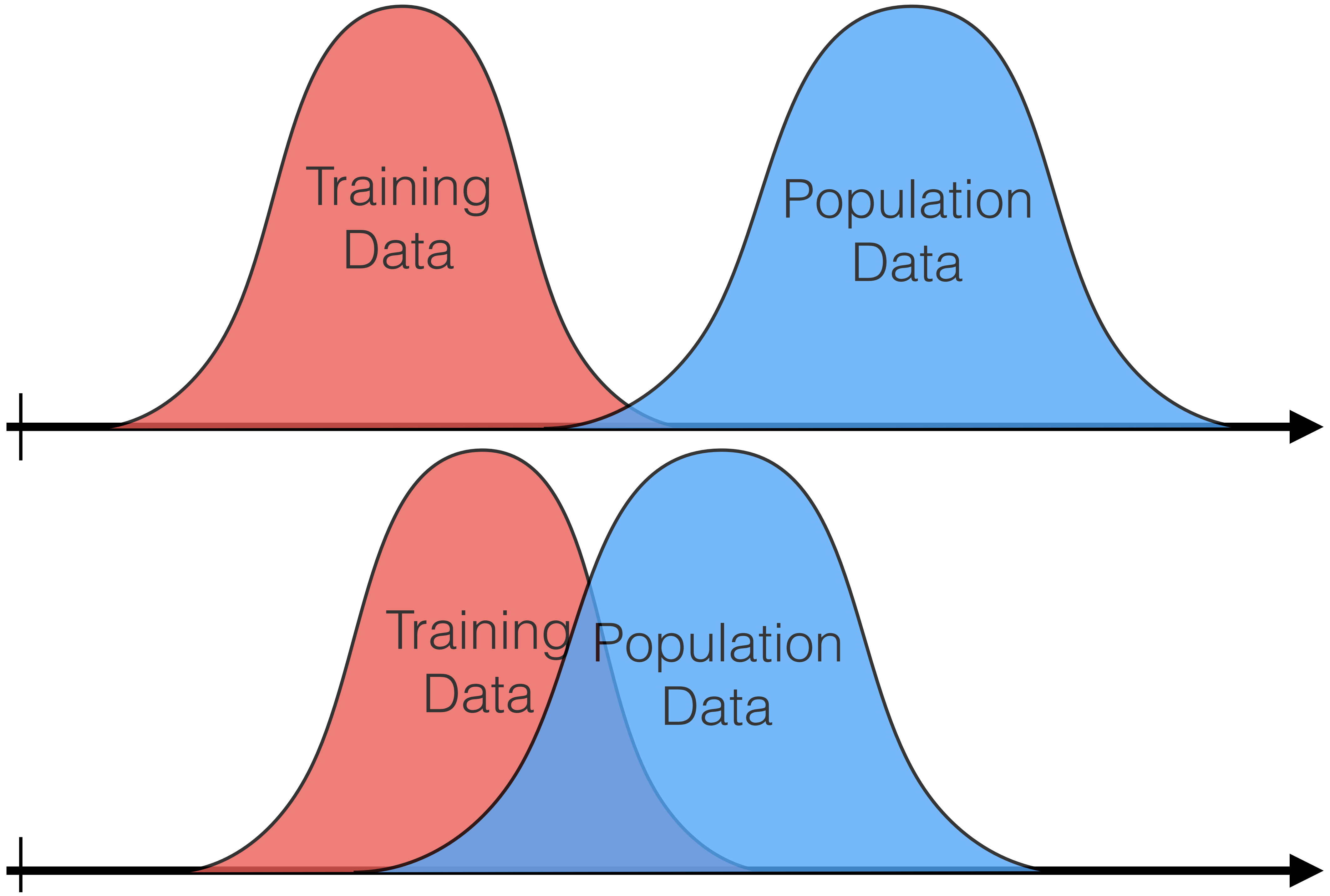
[Yeom, Glacomelli, Fredrikson, Jha] Privacy risk in machine learning, CSF'18

[Sablarolles, Douze, Schmid, Olivier, Jegou] White-box vs black-box: Bayes optimal strategies for membership inference, ICML'19

Reason for Leakage?

Overfitting

An average behavior of the model on data distributions



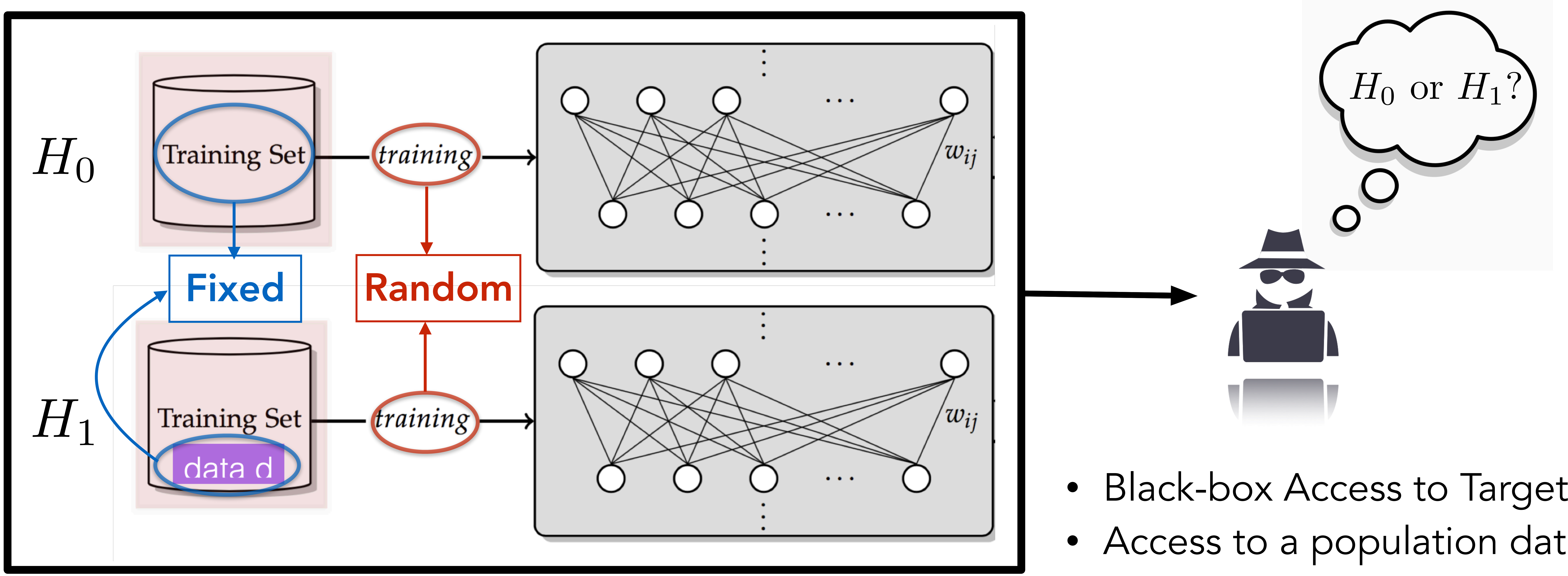
Higher Leakage

Better Generalization

A large body of the literature is based on techniques that simulates these two average-case member and non-member worlds, via training **shadow models** on random **population data**

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

How to Design Stronger Inference Attacks?



- Black-box Access to Target Model
- Access to a population data pool

Minimize the **uncertainties** of MIA Game → **A Strongest Inference Attack**

[Jagielski, Ullman, Opera] Auditing Differentially Private Machine Learning: How Private is Private SGD? NeurIPS'20
[Nasr, Song, Thakurta, Papernot, Carlini] Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning, IEEE S&P'21

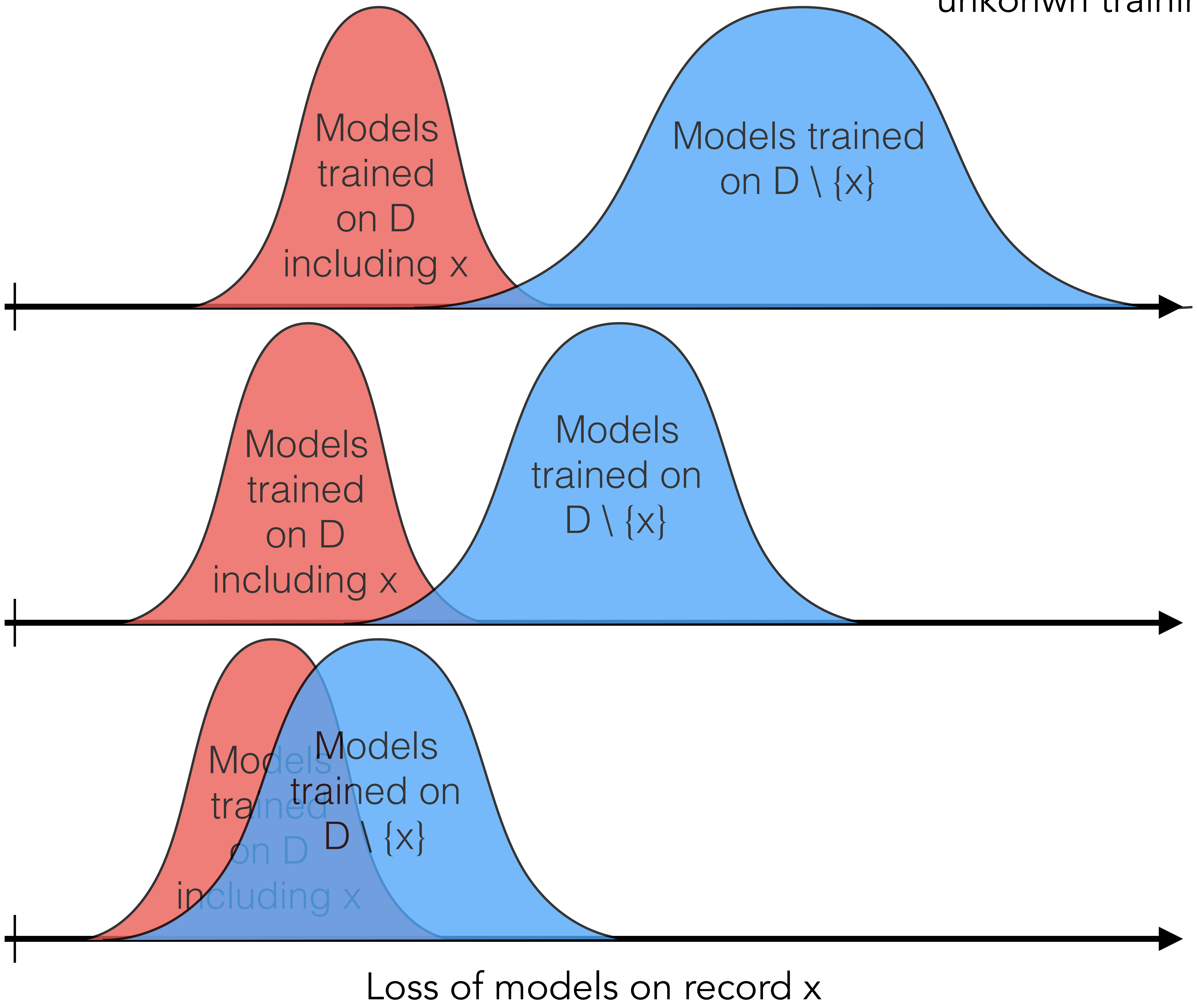
Reason for Leakage?

Conditional Memorization
The behavior of models on a data point, conditioned over other unkonwn training data

Higher Leakage

Conditionally Atypical
Hard to learn data sample x, given other training data D

Conditionally Typical
Easy to learn data sample x, given other training data D


Models trained on D including x
Models trained on D \ {x}
Models trained on D including x
Models trained on D \ {x}
Models trained on D including x
Models trained on D \ {x}
Loss of models on record x

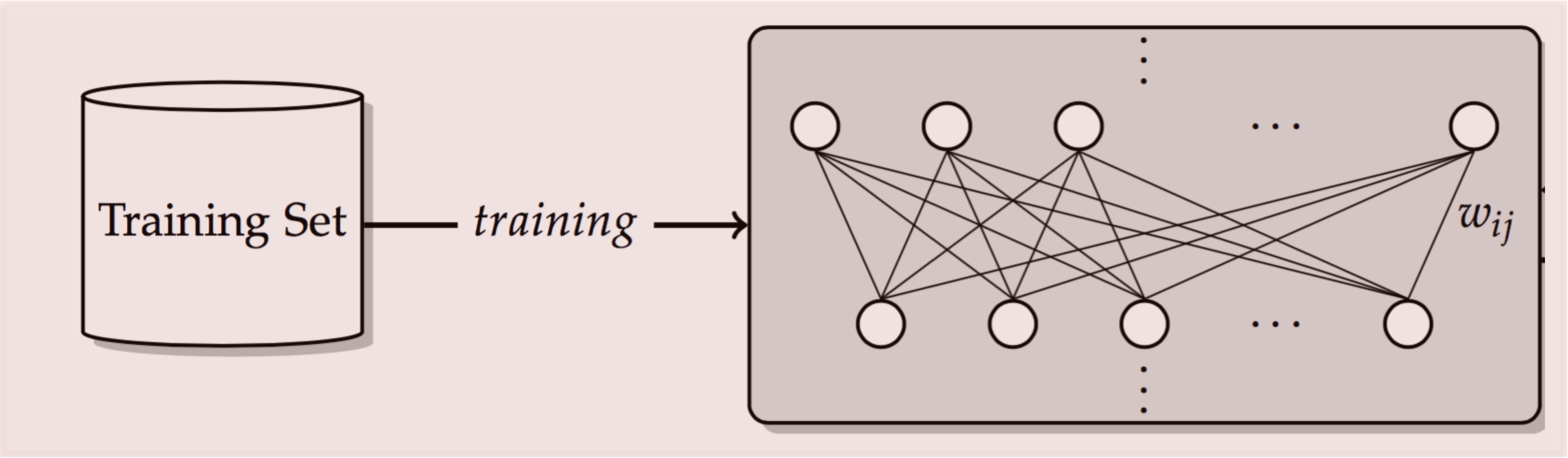
Less conditional memorization on x, given D

How to simulate the two worlds in this game when the remaining training dataset is unknown?

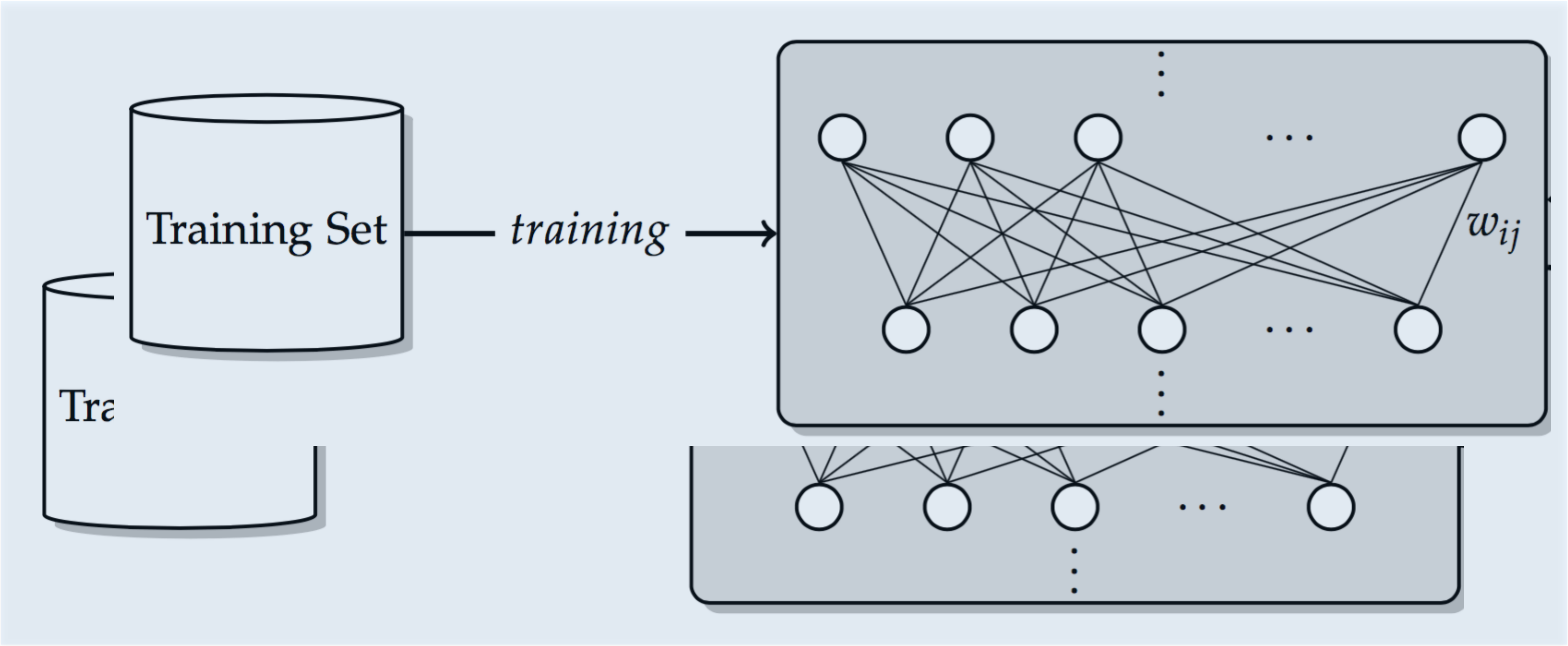
Reference Models

Mimic all the training dataset of the target model (except the target data)

Target Model



Reference Models



data d

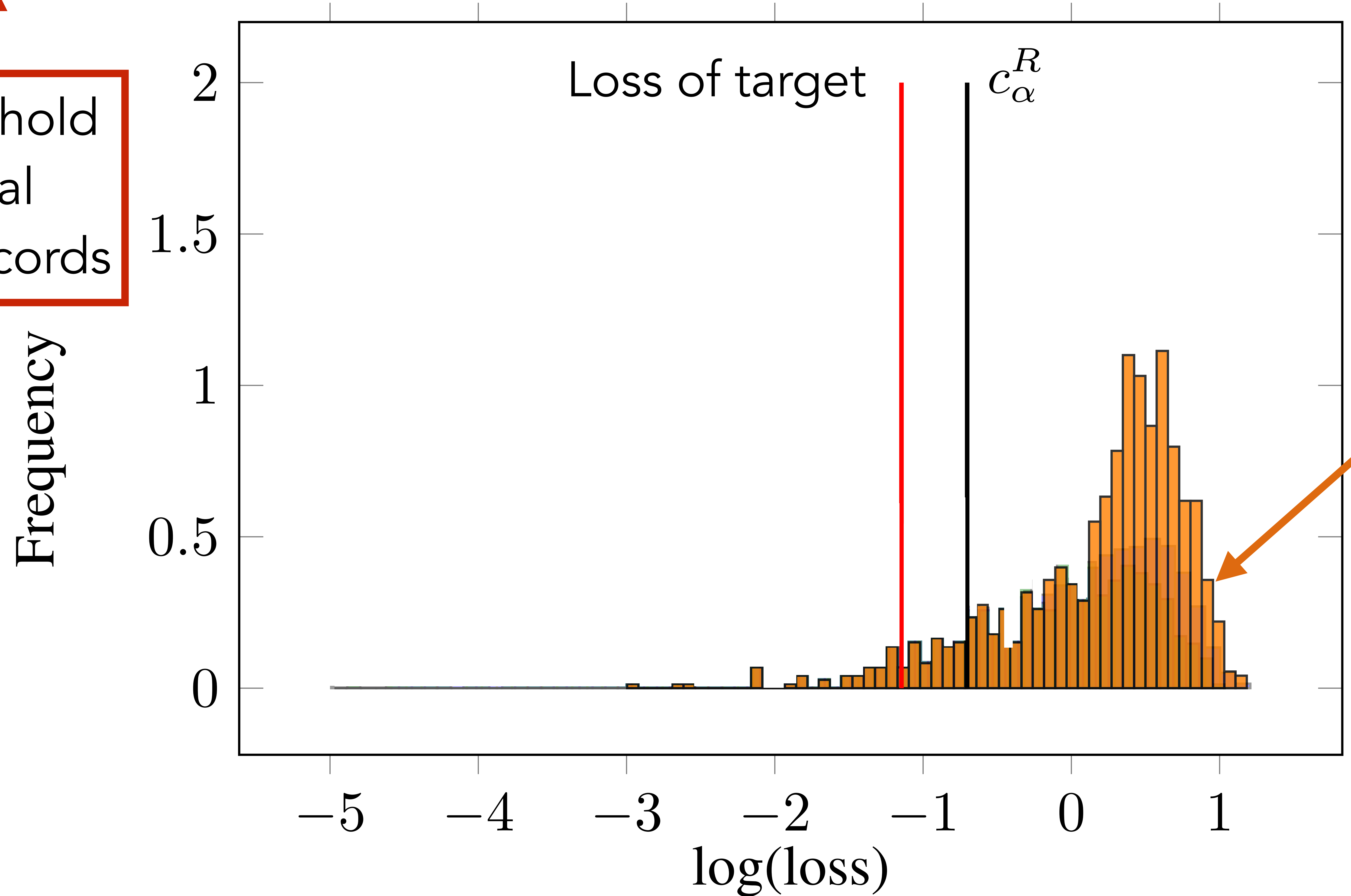
- E.g., train reference models on **random population records**, i.e., similar to shadow models
- E.g., **Model distillation** — train reference models on **relabelled** random population records **by the target model**

Our MIA via Reference Models on Target Data

$$\text{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta, x_z, y_z) \;\; \text{Predict "Member"}$$

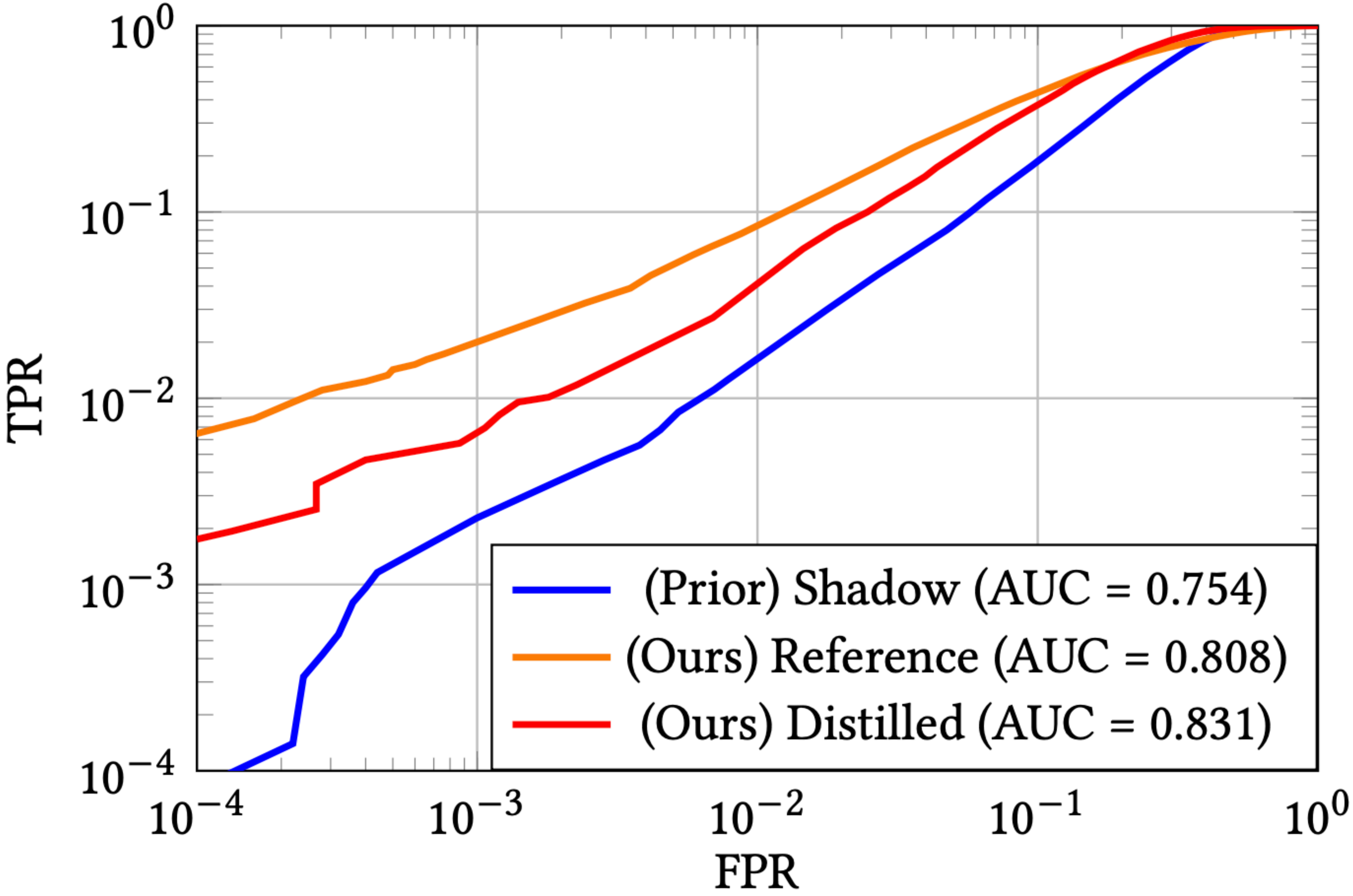
- Learn a threshold from the loss distribution of target data on reference models

Different threshold for individual models and records

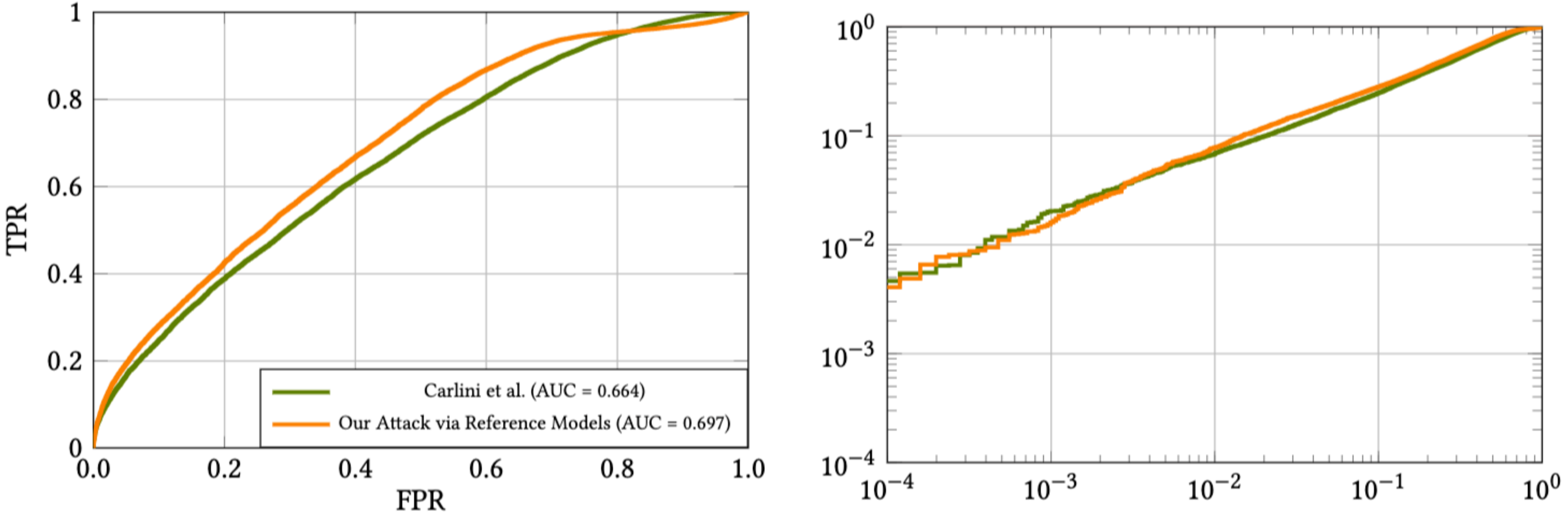


Loss of target data on reference models

Our MIAs via Reference Models are Stronger than Prior Attacks via Shadow Models



Our MIA via Reference Models is Stronger than existing MIAs of similar nature



(a) Overall TPR-FPR

(b) Focus on Small FPR Region

[Carlini, Chien, Nasr, Song, Terzis, Tramèr] Membership inference attacks from first principles, IEEE S&P’22

Main Takeaways

- Membership inference attack is useful for auditing different kinds of leakage when formulated in different games
- There are multiple issues with the existing MIA in formalizing the problem and the performance of attacks
- We propose a framework to deal with these issues, and design more powerful attack via reducing adversary's uncertainty

Privacy Meter

privacy-meter.com