Data Privacy in Machine Learning

Reza Shokri
Data Privacy and Trustworthy ML Research Lab
National University of Singapore

reza@comp.nus.edu.sg @rzshokri

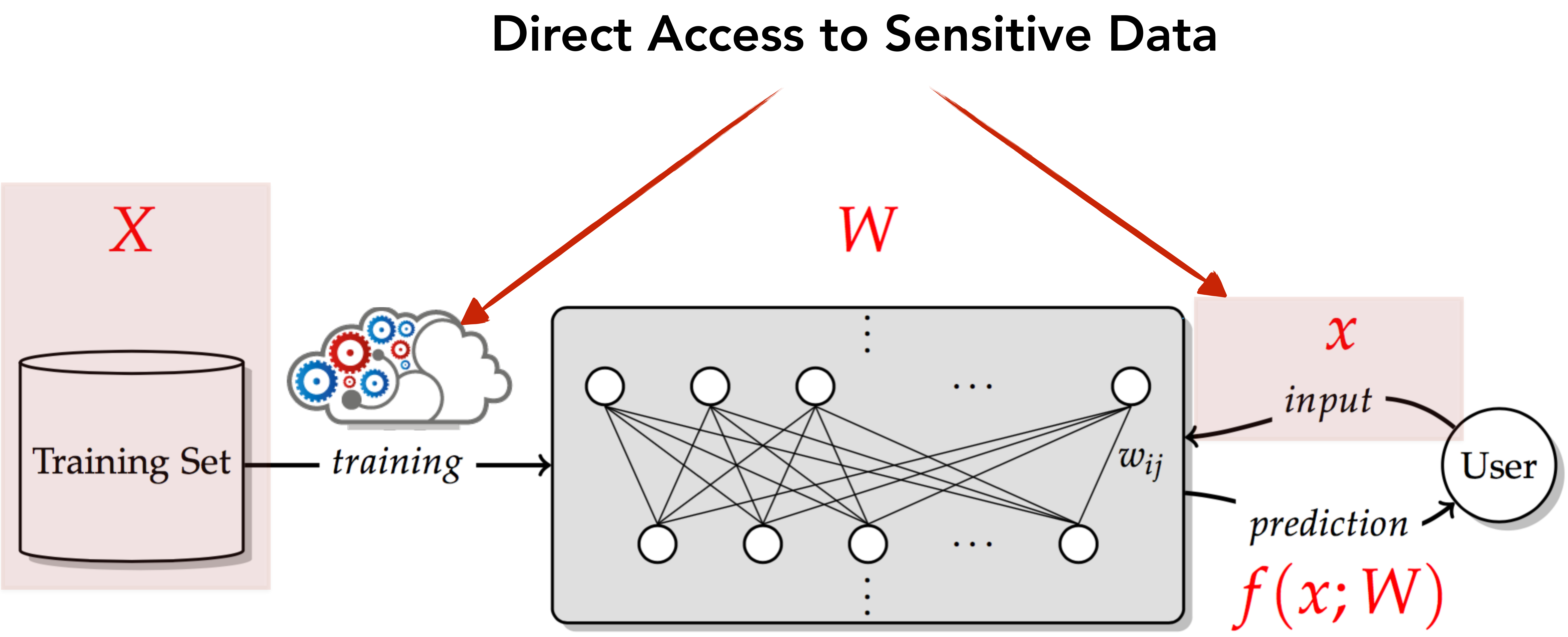
CCS 2022 Tutorial

Privacy Regulations

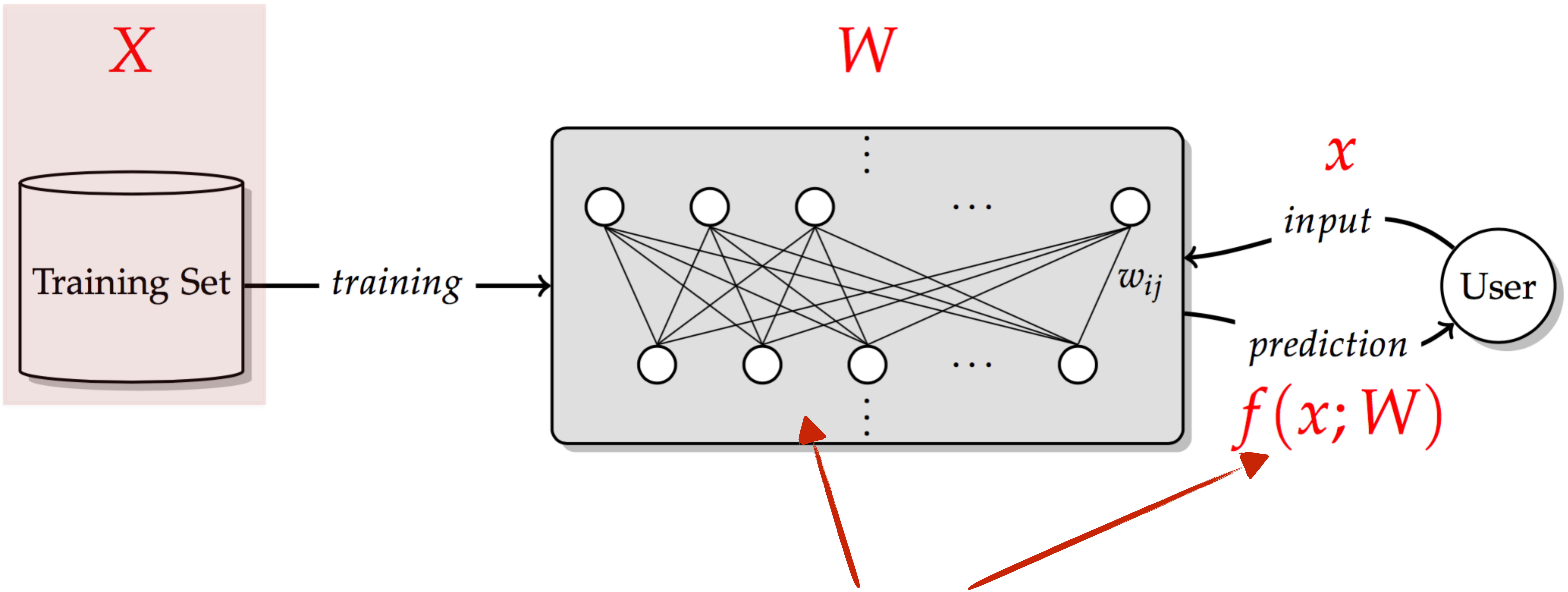
GDPR — Data Protection Impact Assessment

The focus is mostly on data collection, data sharing, access control, …

Direct Privacy Risks

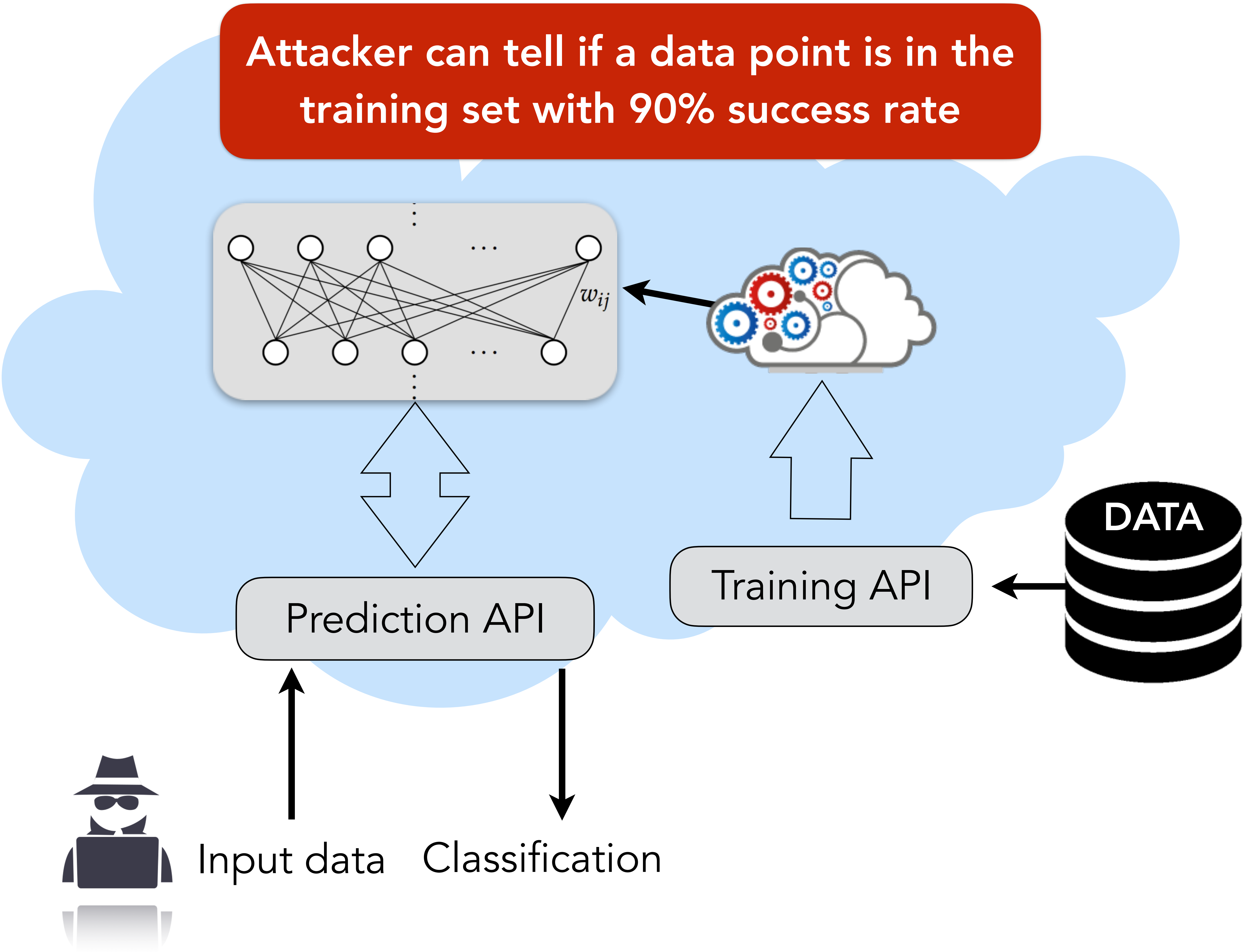


Indirect Privacy Risks



Indirect Leakage about X

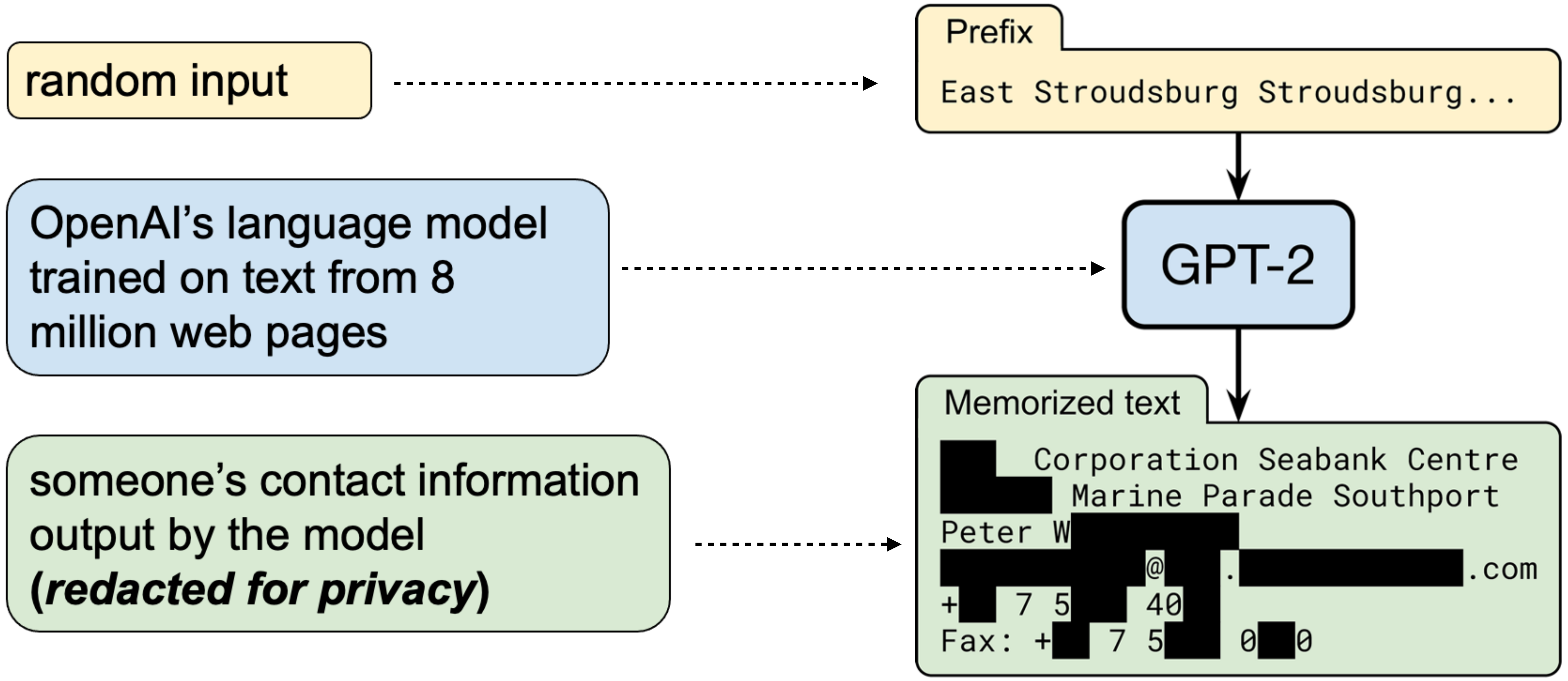
Real World Attacks against Machine Learning as a Service Platforms



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

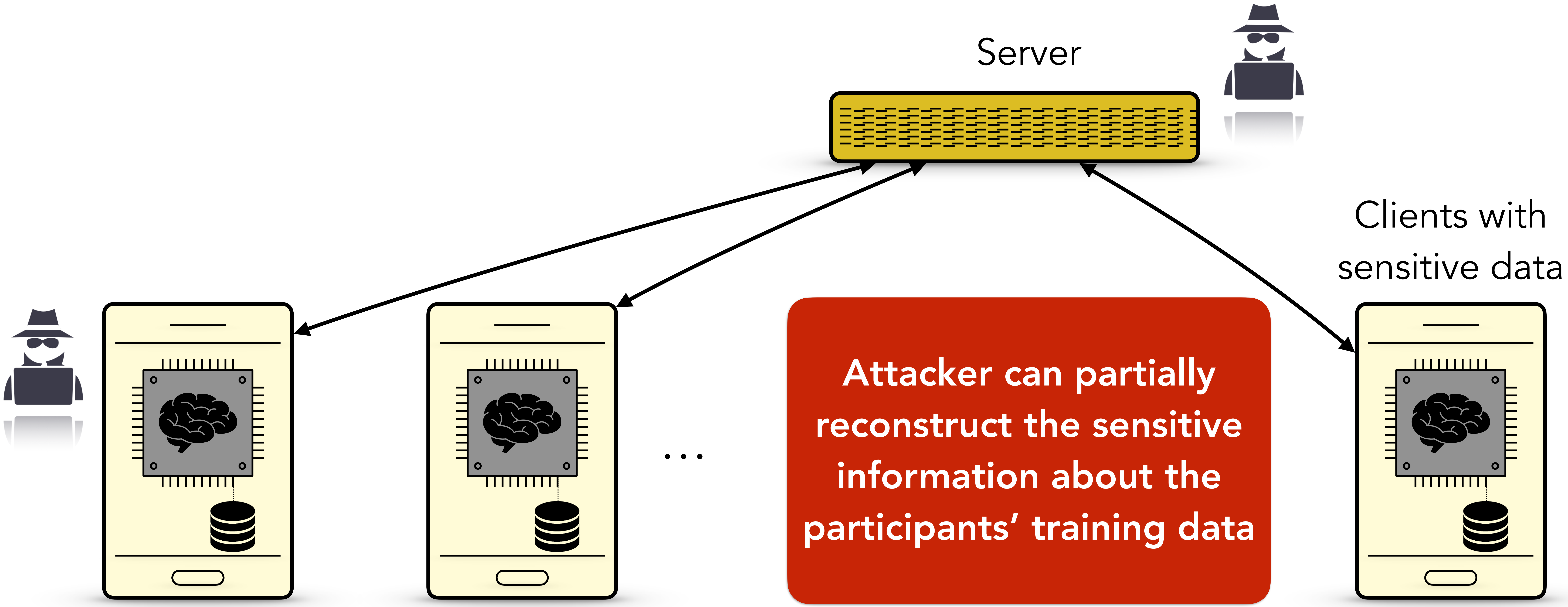
Real World Attacks against Large Language Models

Attacker can partially reconstruct the sensitive data used for training the model



[Carlini, Tramer, et al.] Extracting Training Data from Large Language Models, Usenix security'21

Real World Attacks against Federated Learning Algorithms



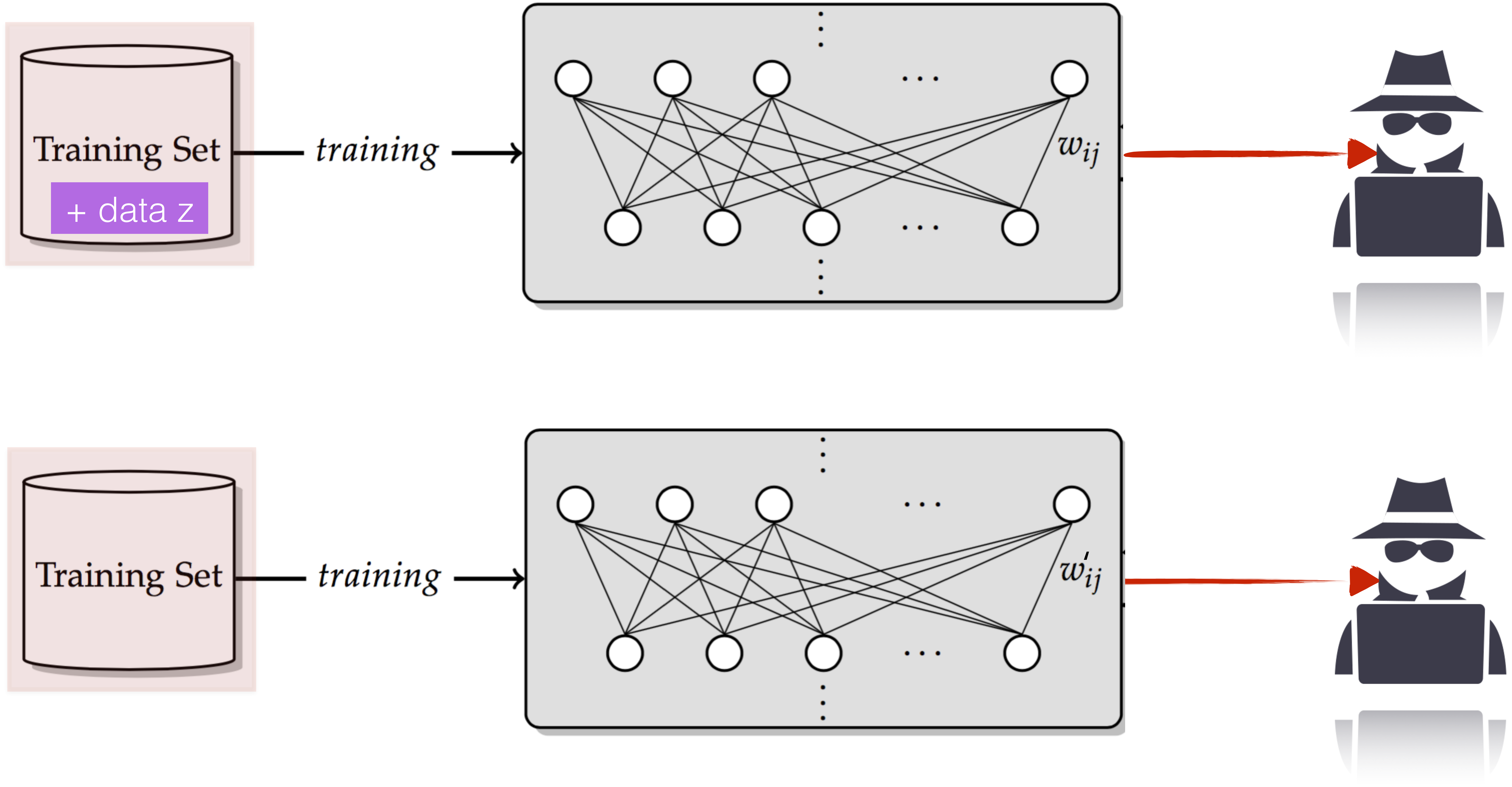
[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

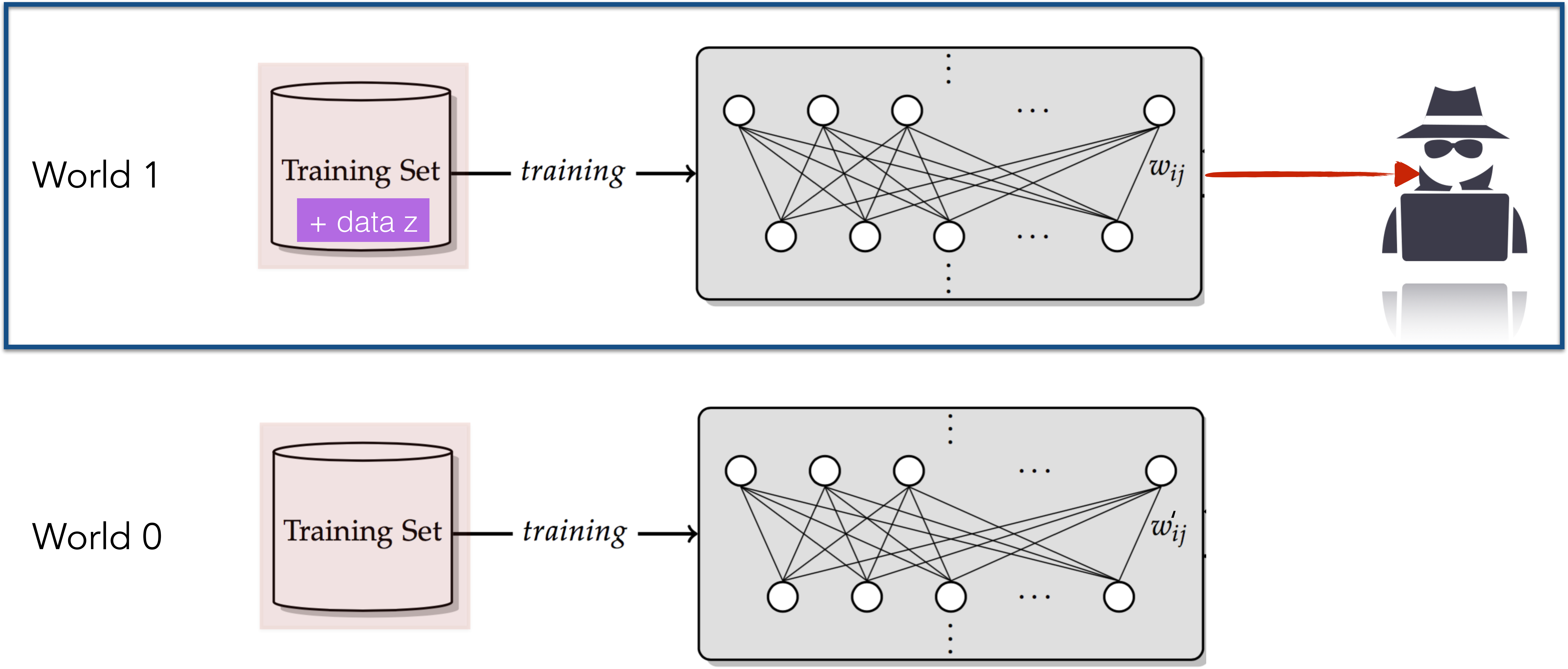
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

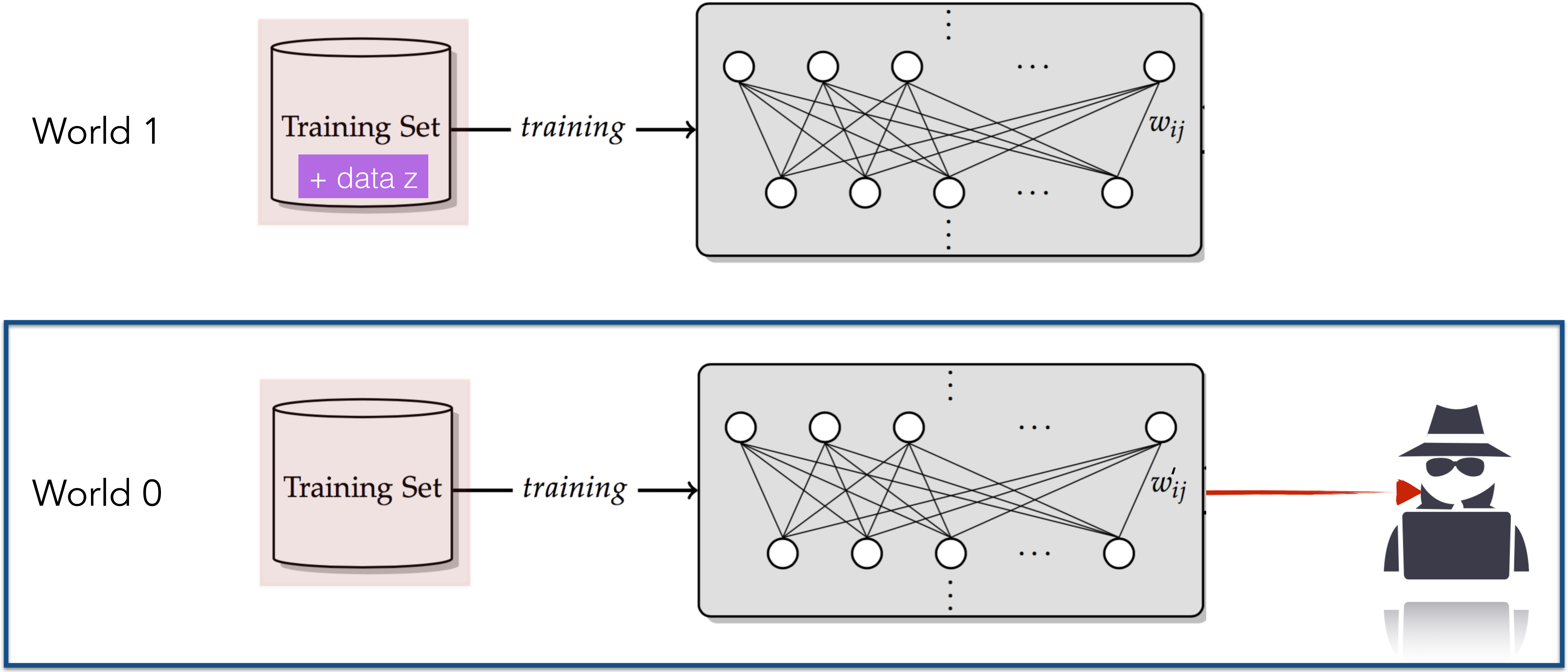
[Zhang, Tople, Ohrimenko] Leakage of Dataset Properties in Multi-Party Machine Learning, Usenix Security'21

Models are **personal data**

We need a standard method for quantitatively auditing data privacy in machine learning systems

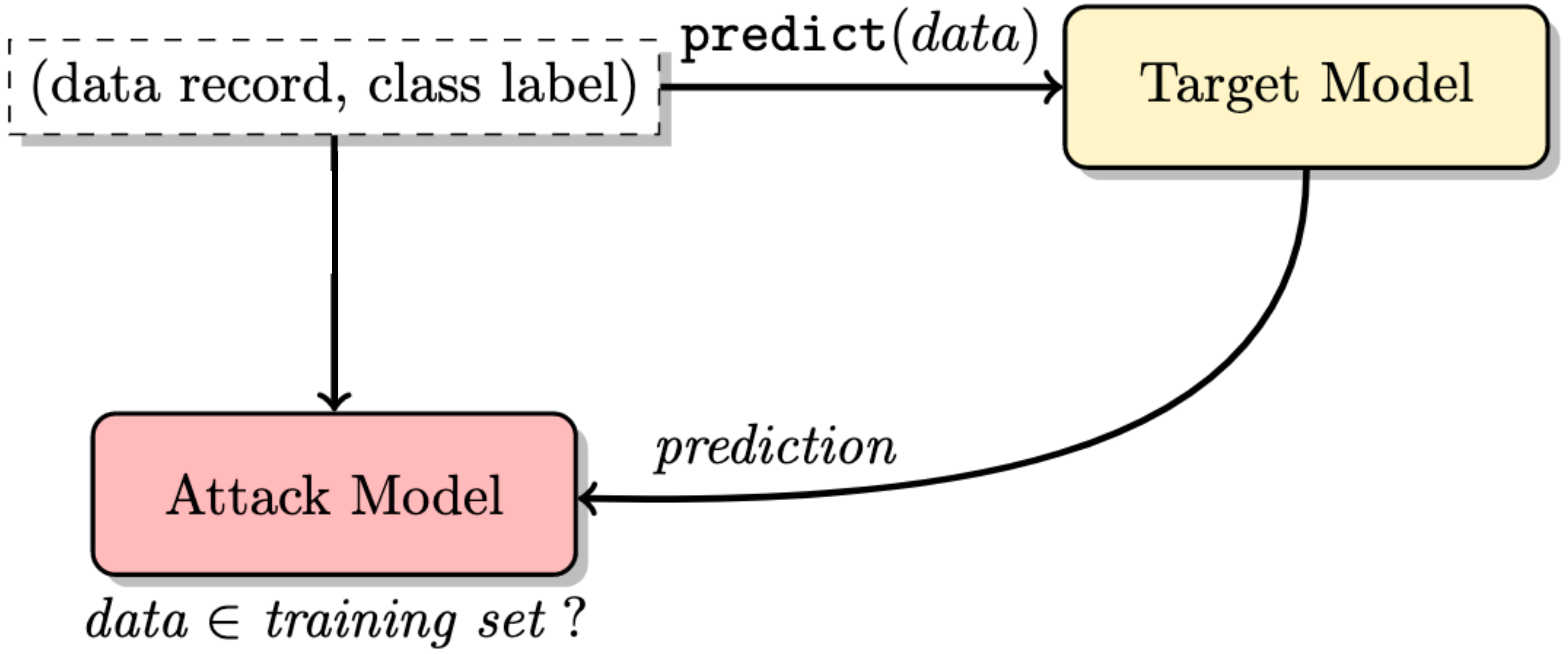

Training Set
+ data z
training
w_{ij}
Training Set
training
w'_{ij}


World 1
Training Set
+ data z
training
w_{ij}
World 0
Training Set
training
w'_{ij}


World 1
Training Set
+ data z
training
w_{ij}
World 0
Training Set
training
w'_{ij}

Membership Inference Attacks

- Given a model, can an adversary infer whether a particular data point is part of its training set?
- Success of attacker is a metric for privacy loss

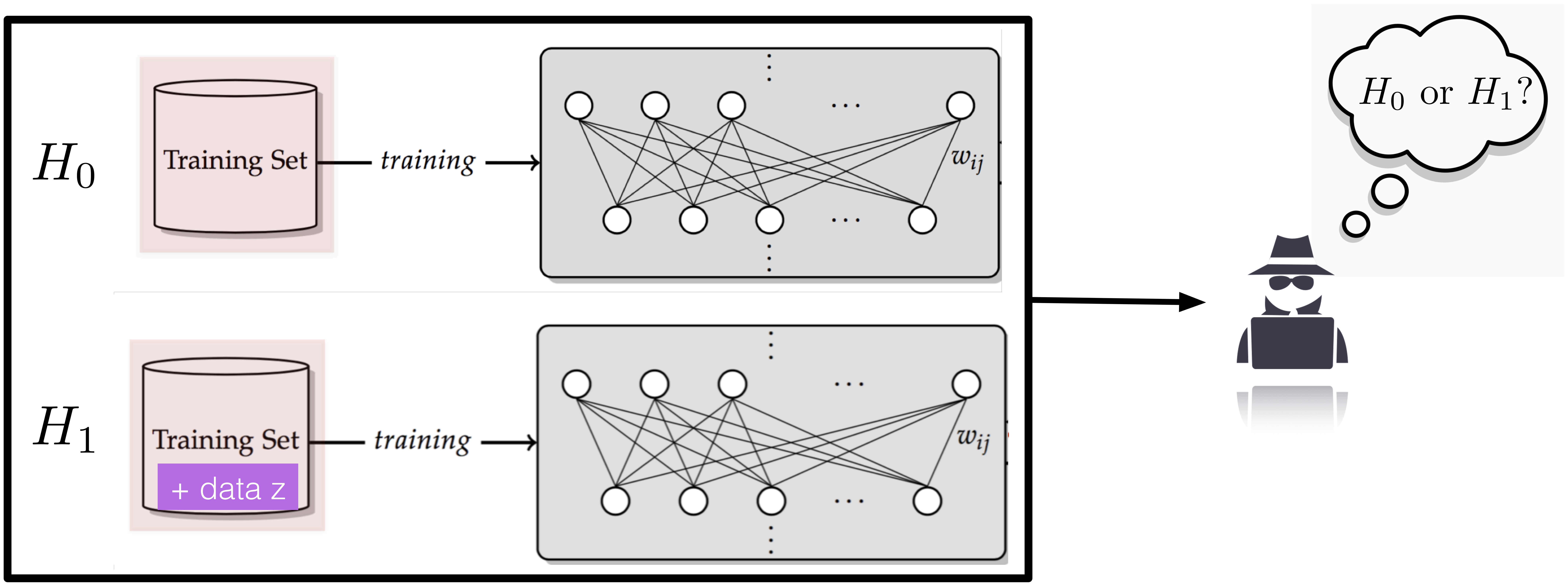


[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, IEEE SP'17

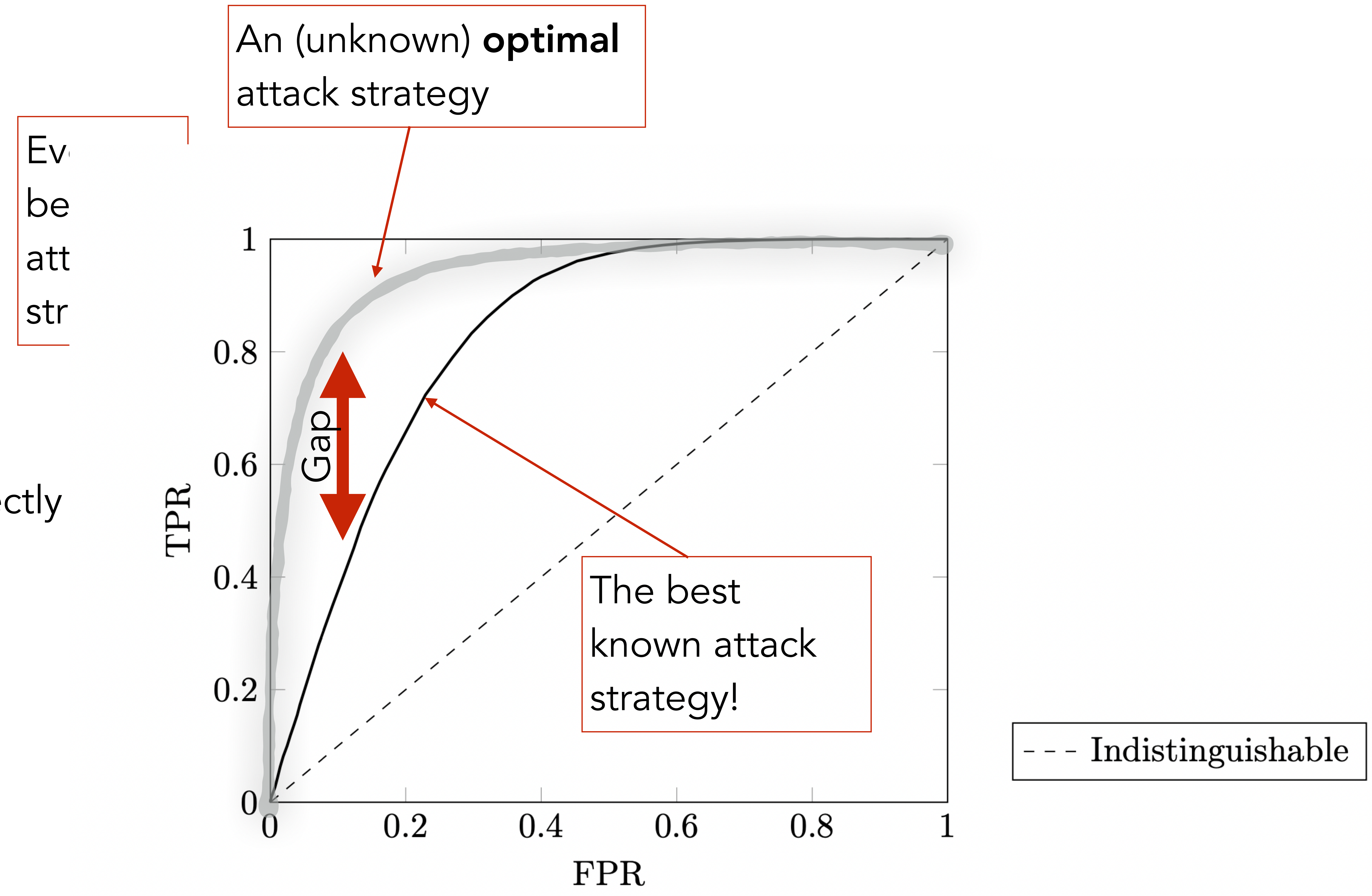
AI Regulations and Guidelines

- "… membership inferences show that AI models can inadvertently contain **personal data**"
- "Attacks that reveal confidential information about the data include membership inference …"
- **"… should consider the risks to data throughout the design, development, and operation of an AI system"**

Membership Inference Attack (MIA) Game



Success of adversary indicates information leakage of models about their training data



power: members are correctly predicted as member

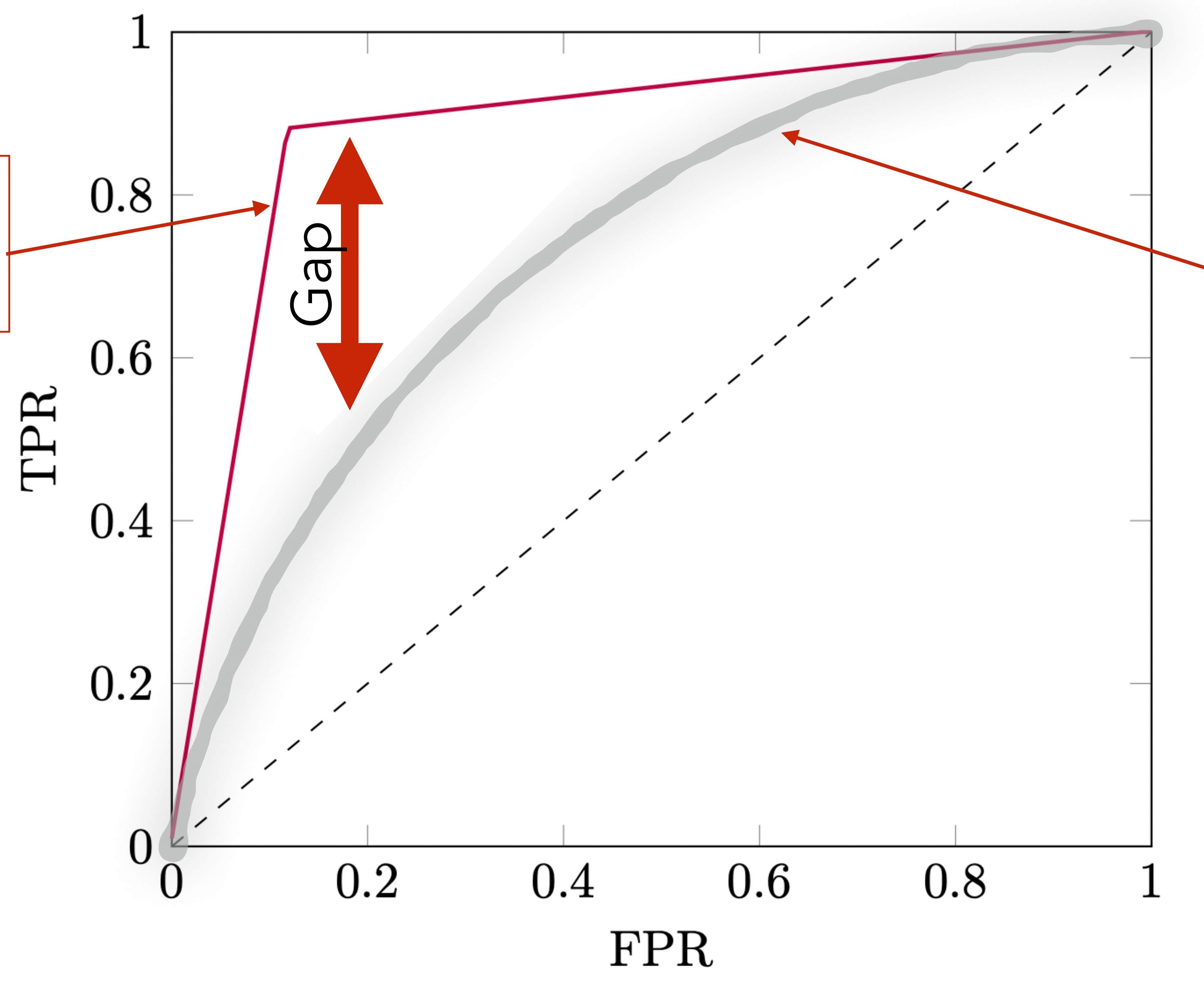
error: non-members are wrongly predicted as member

An attack strategy gives a **lower-bound** on the privacy risk of the target ML algorithm

This is very useful to rule out vulnerable algorithms, …
but, lack of a known powerful attack is not a guarantee for privacy!

An Upper bound on privacy loss

$$e^{\epsilon} \cdot (1 - TPR) + FPR \geq 1 - \delta$$
$$1 - TPR + e^{\epsilon} \cdot FPR \geq 1 - \delta$$



An (unknown) **optimal** attack strategy for the randomized ML algorithm

$$\mathbb{P}(M(D) \in S) \quad \leq \quad e^{\varepsilon}\, \mathbb{P}(M(D') \in S) + \delta$$

(ϵ, δ) Differential Privacy
Indistinguishable

Prove an **upper-bound** for the privacy risk of a <u>randomized</u> algorithm…

[Kairouz, Oh, Viswanath] The Composition Theorem for Differential Privacy, ICML'2015

A differential privacy guarantee is an **upper-bound** on the privacy risk of a randomized ML algorithm

If the bound is loose, we are over-estimating the risk, thus we unnecessarily over-randomize the algorithm, … which could result in a high utility drop (e.g., prediction error) due to the algorithm.

How to Design Powerful Auditing Algorithms?

[Ye, Maddi, Murakonda, Bindschaedler, Shokri]
Enhanced Membership Inference Attacks against Machine Learning Models
ACM CCS'22

Hypothesis Testing for Membership Inference

- Given a data point “z” and black-box access to a model “θ”,
    - Determine if “z” was a member of the training set of “θ”

[Murakonda, Shokri, Theodorakopoulos] Quantifying the Privacy Risks of Learning High-Dimensional Graphical Models, AISTATS 21
[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS’22

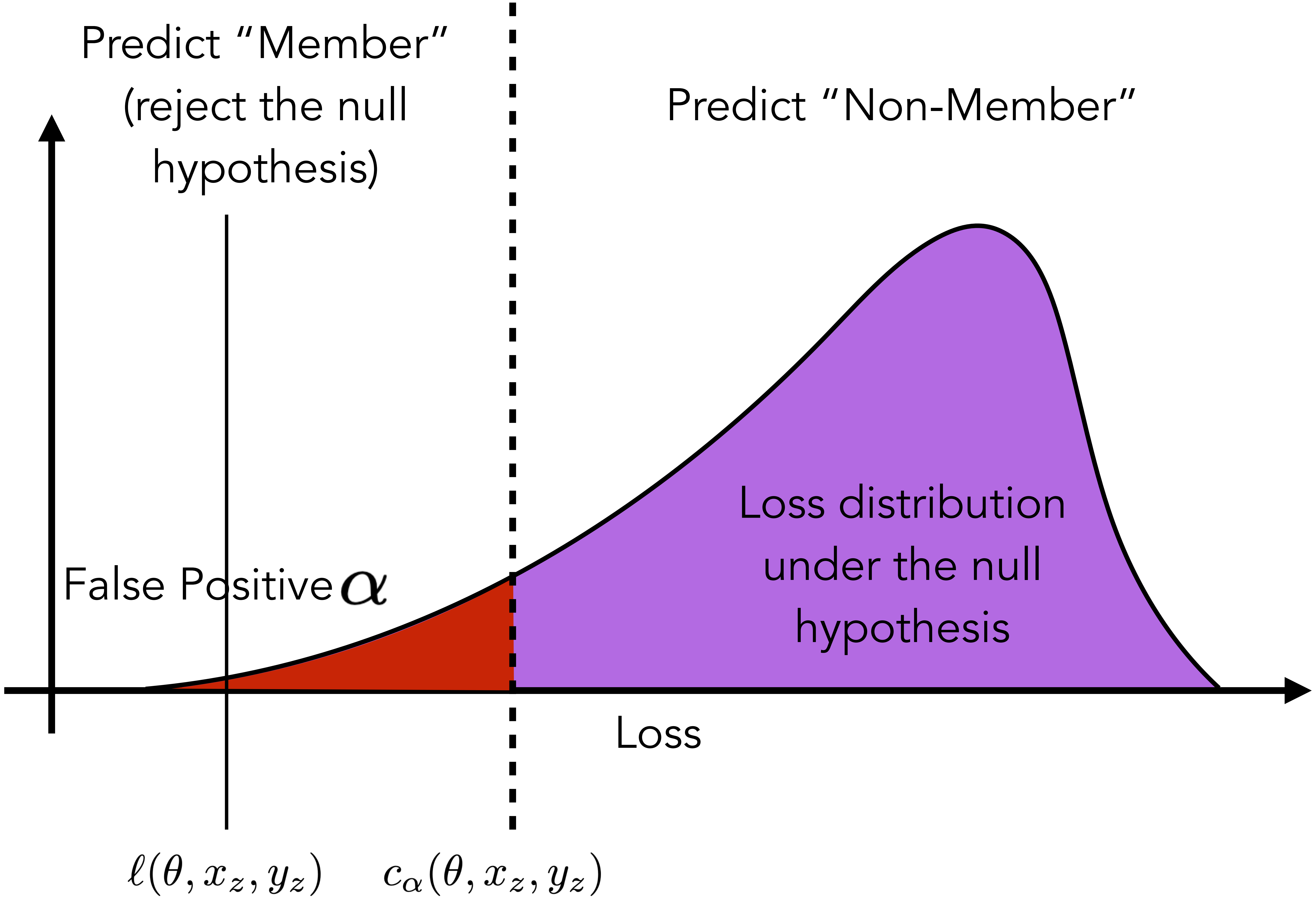
Likelihood Ratio Test

$$LR(\theta, z) = \frac{L(H_0|\theta, z)}{L(H_1|\theta, z)}$$

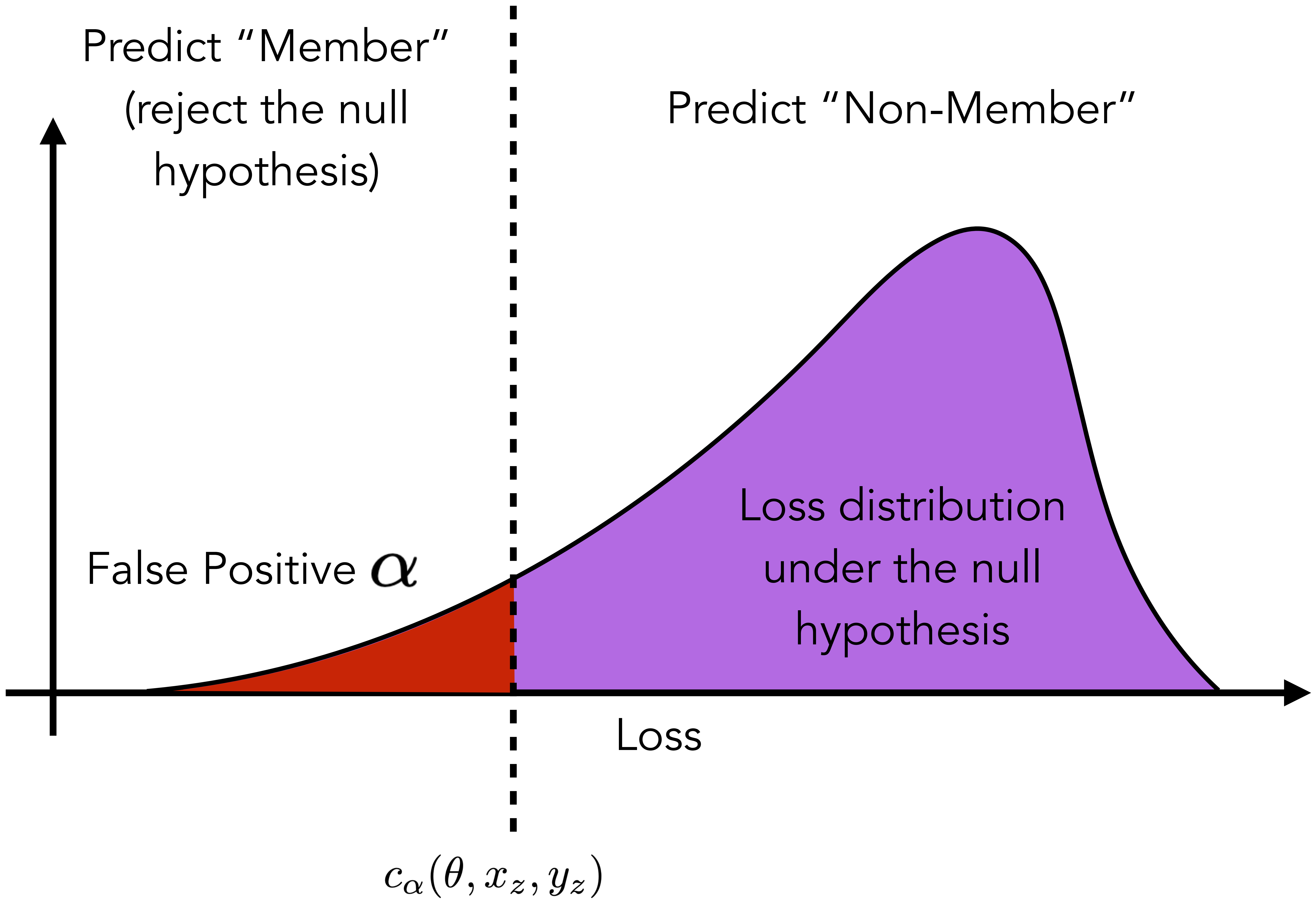
Attack: $$\text{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta, x_z, y_z), \ \textbf{reject } H_0$$

false positive rate

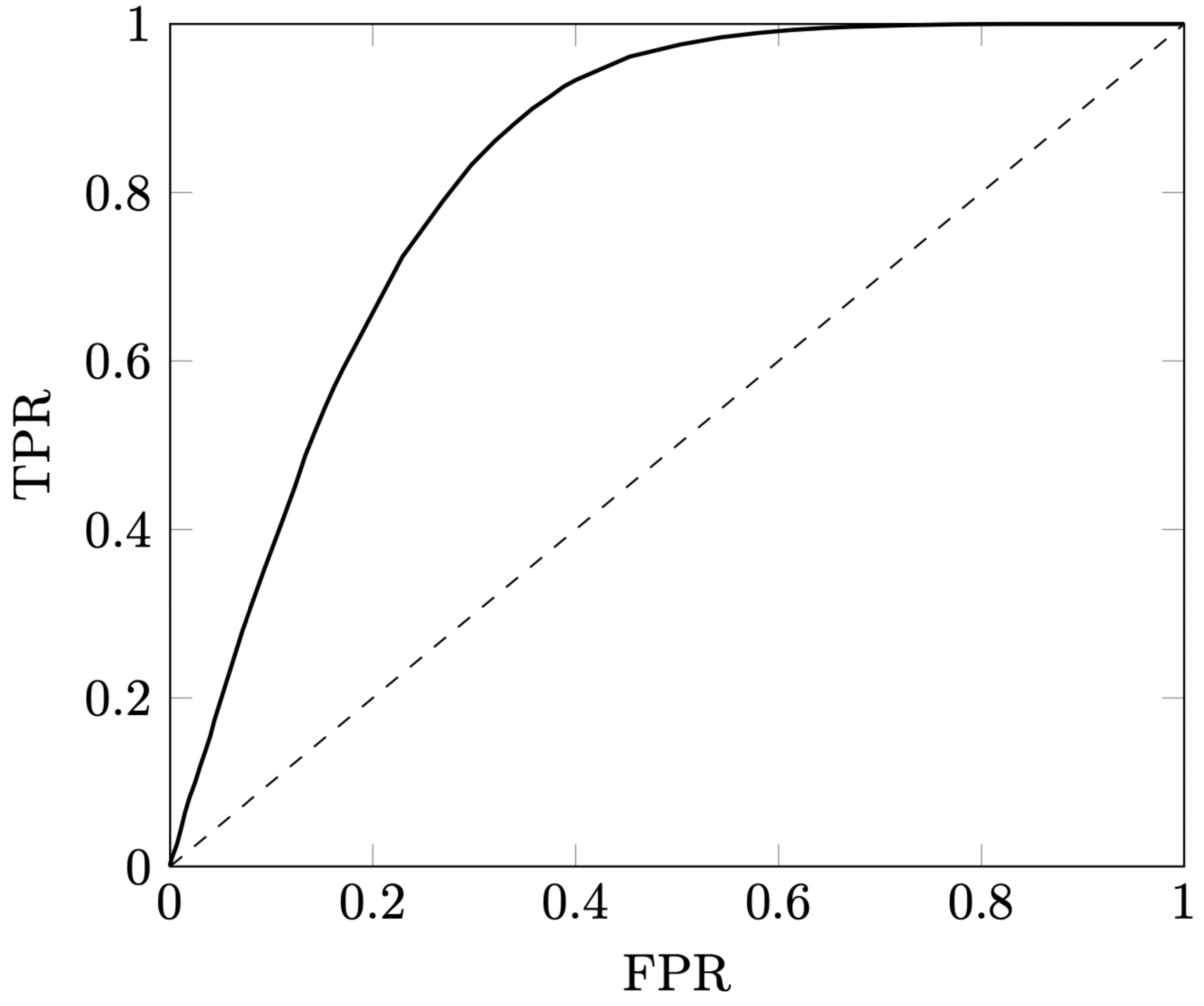
How to Interpret the Test?



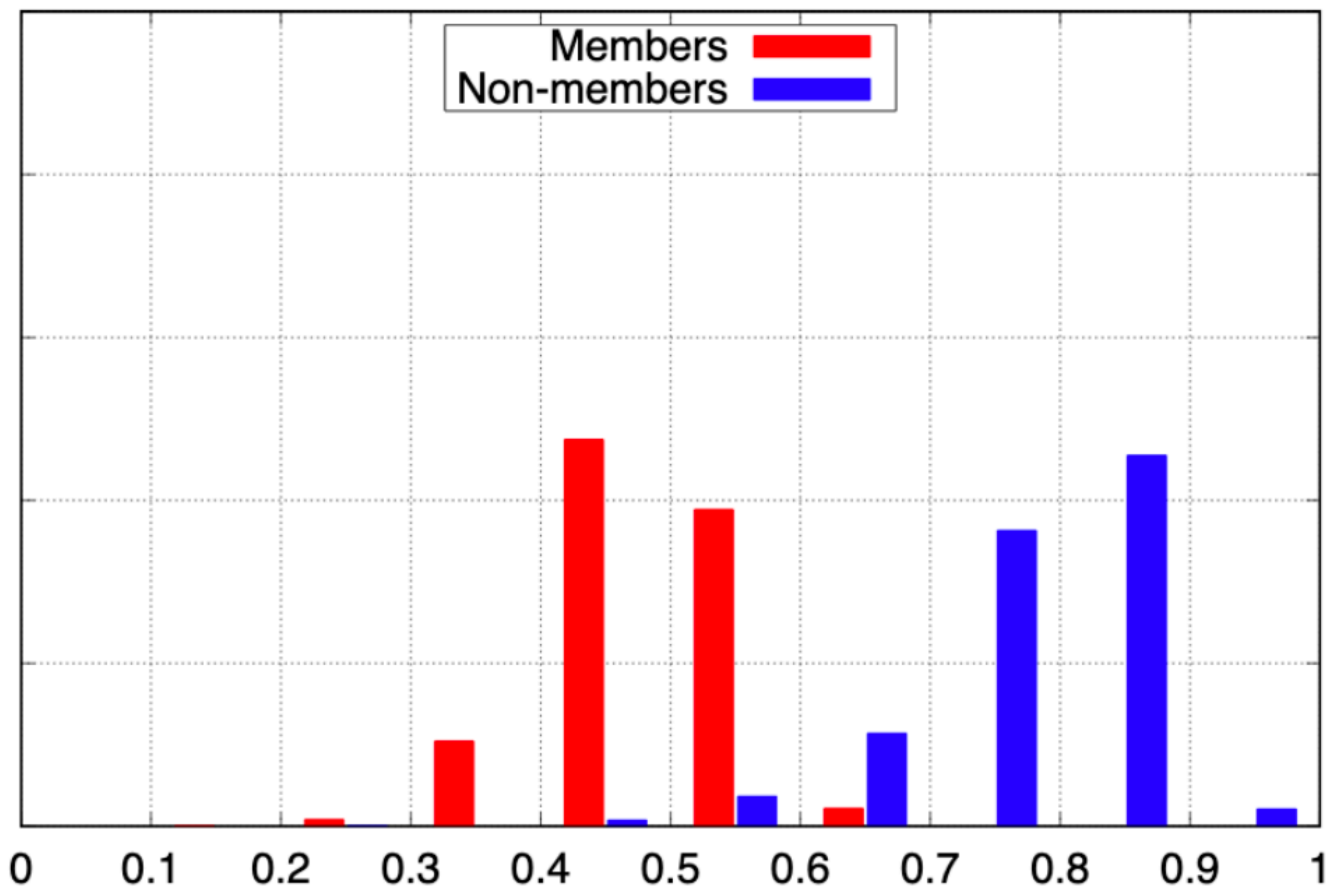
How to Construct the Test?



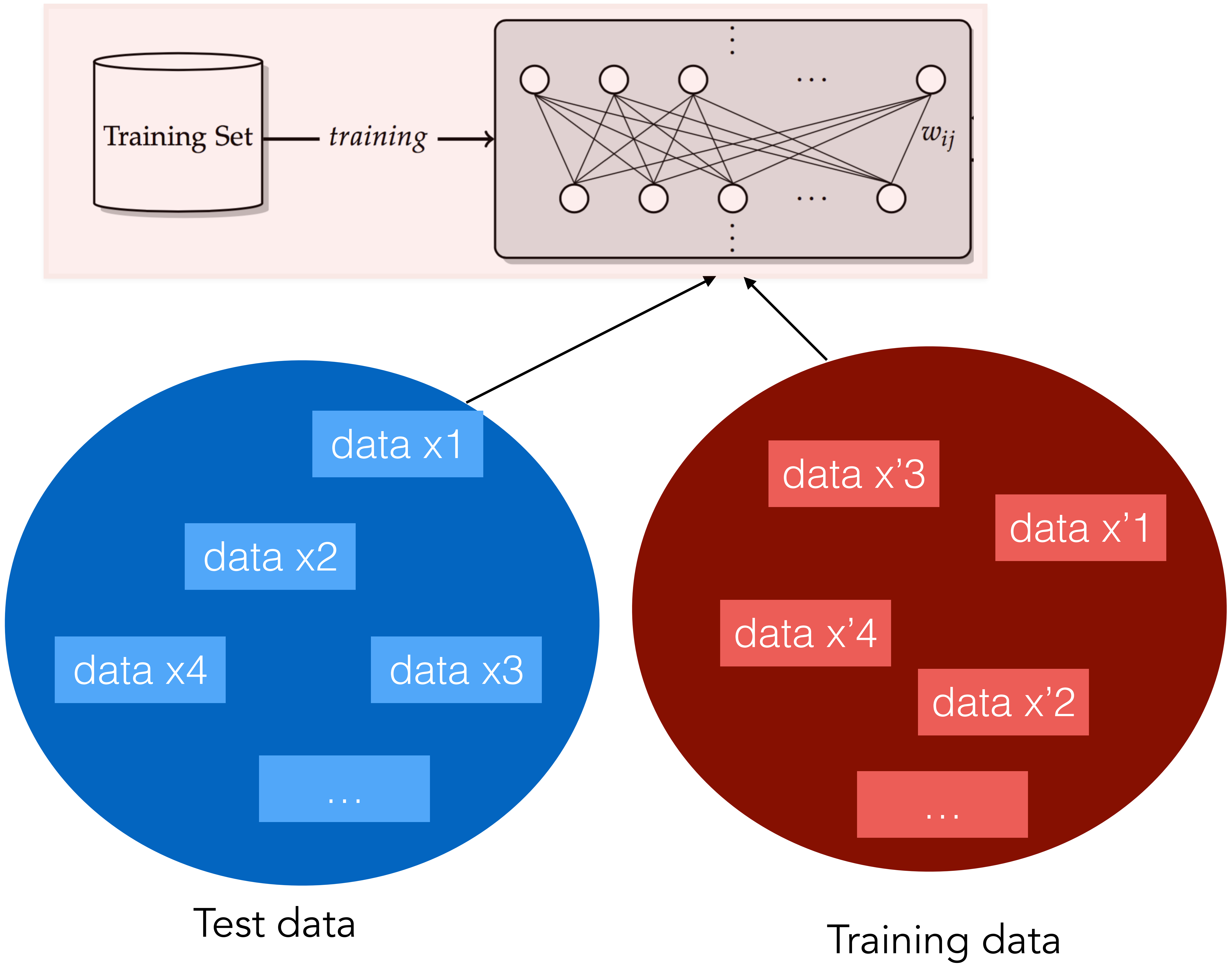
Power vs Error of the Test



Constructing the Test …



Model's Error

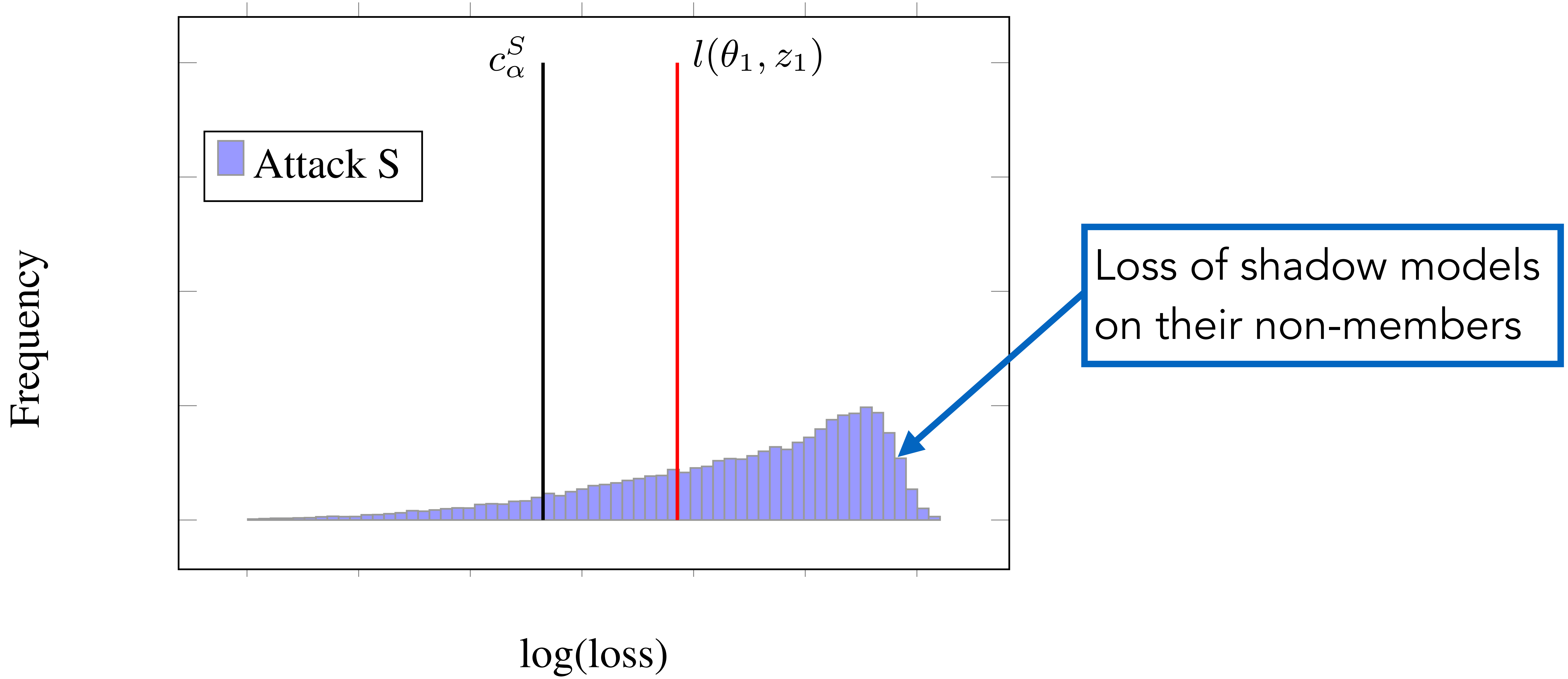


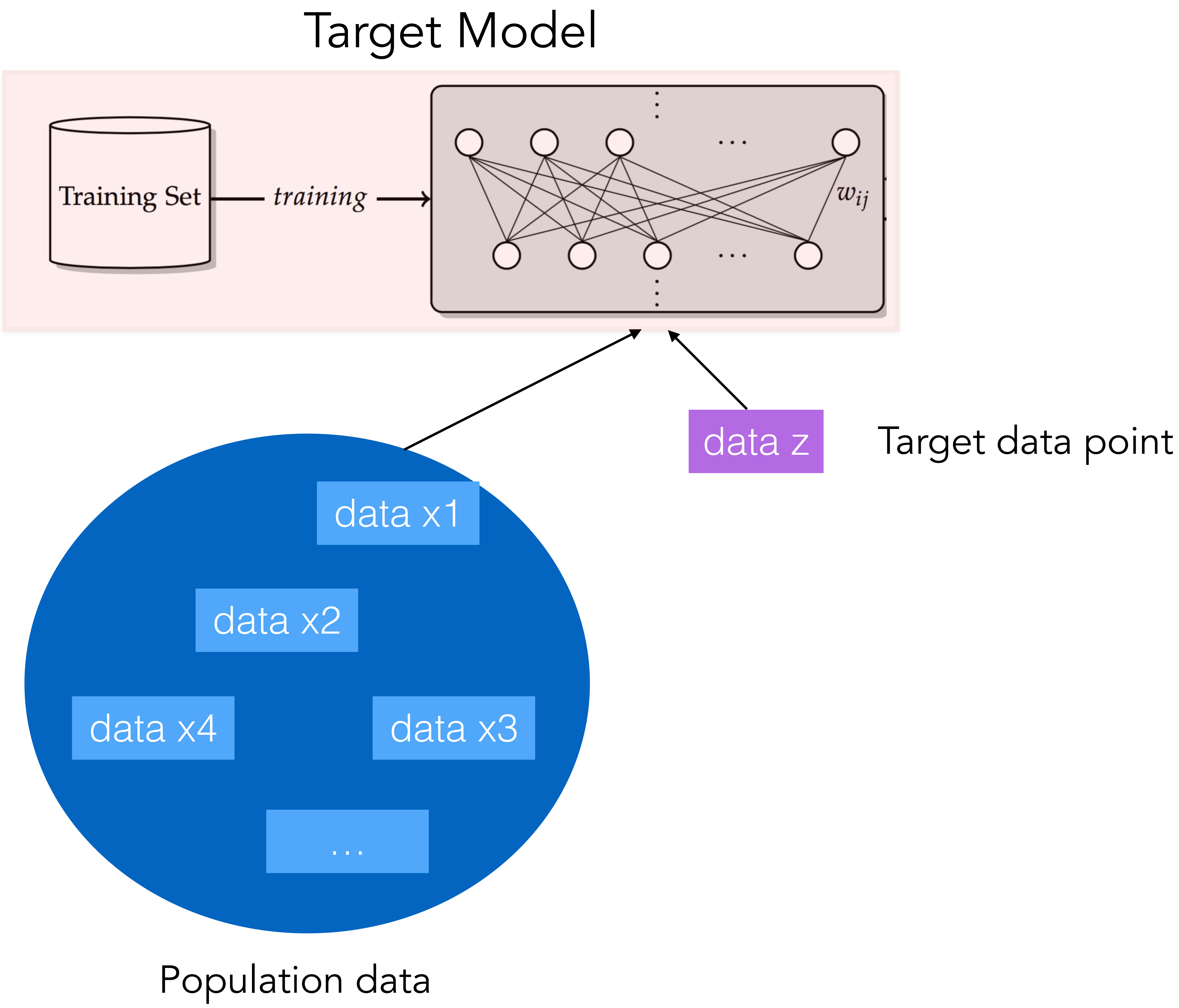
[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

Membership Inference via Shadow Models

$$\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(y_z), \ \textbf{reject } H_0$$

- A large body of the literature is based on this technique [SSSS2017]
- Learn a threshold from the behavior of some shadow models on their test data



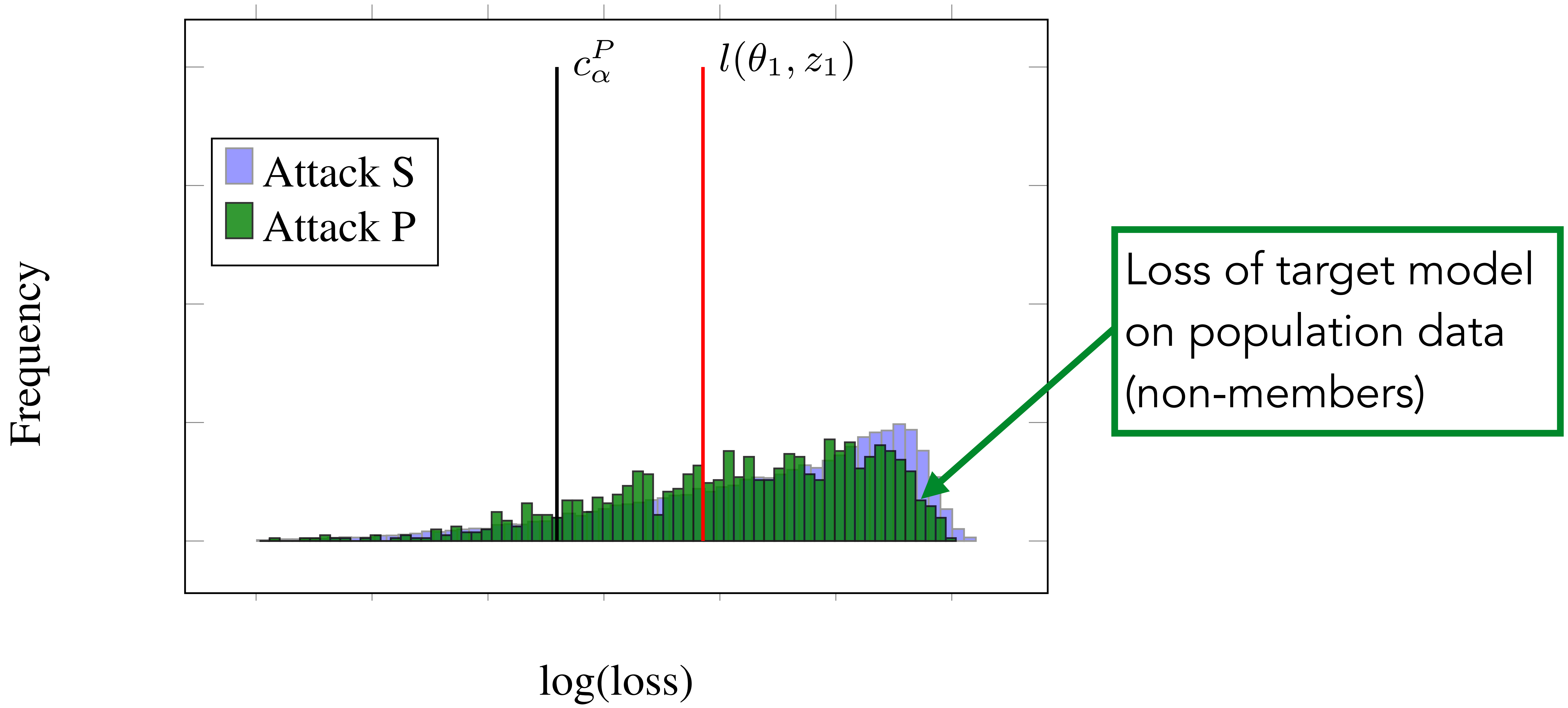


[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22

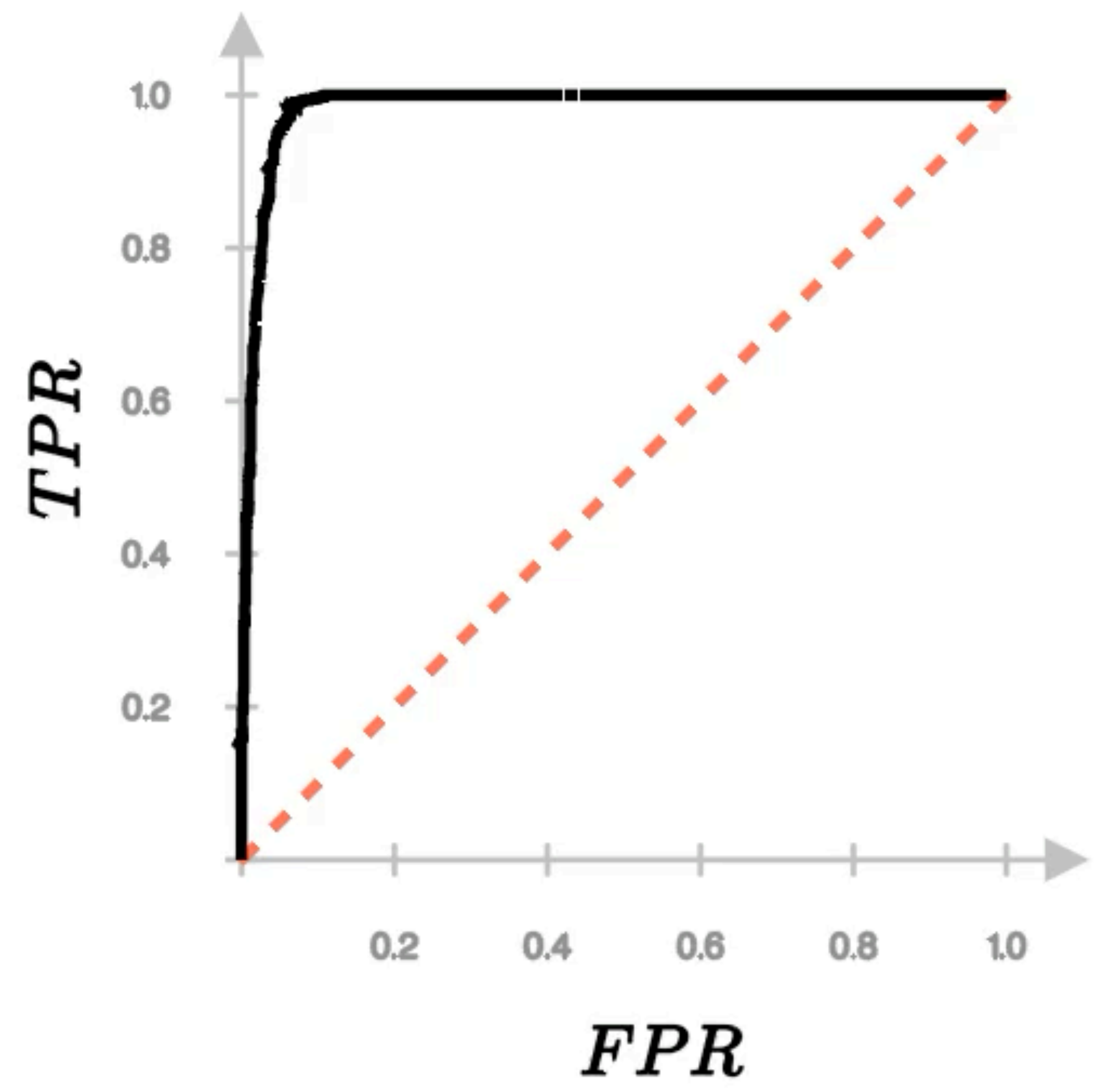
Membership Inference via Population Data

$$\text{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta), \ \textbf{reject } H_0$$

- Directly learn a threshold from the loss distribution of the target model on population data

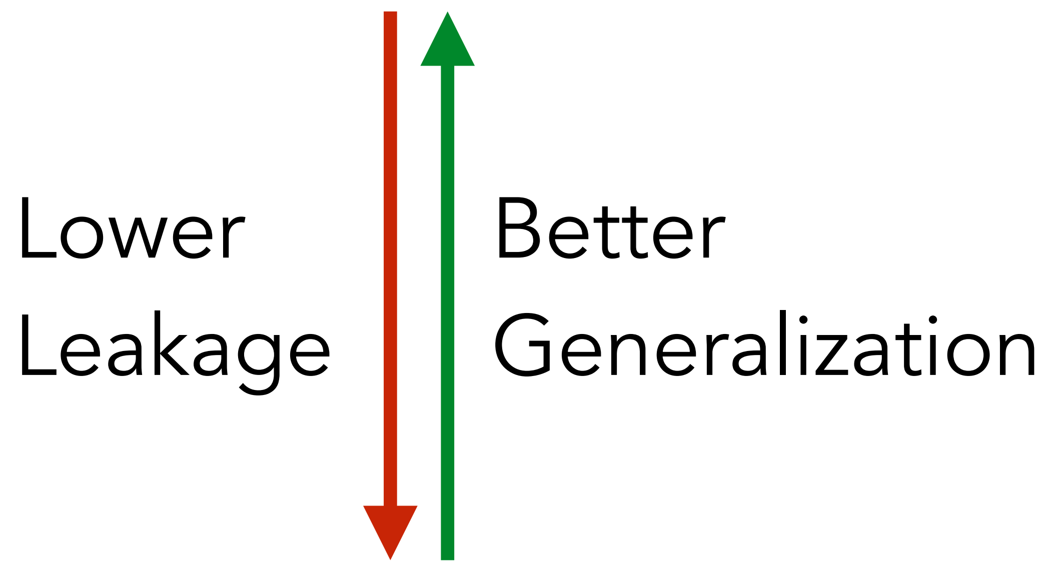


Reason for Leakage?

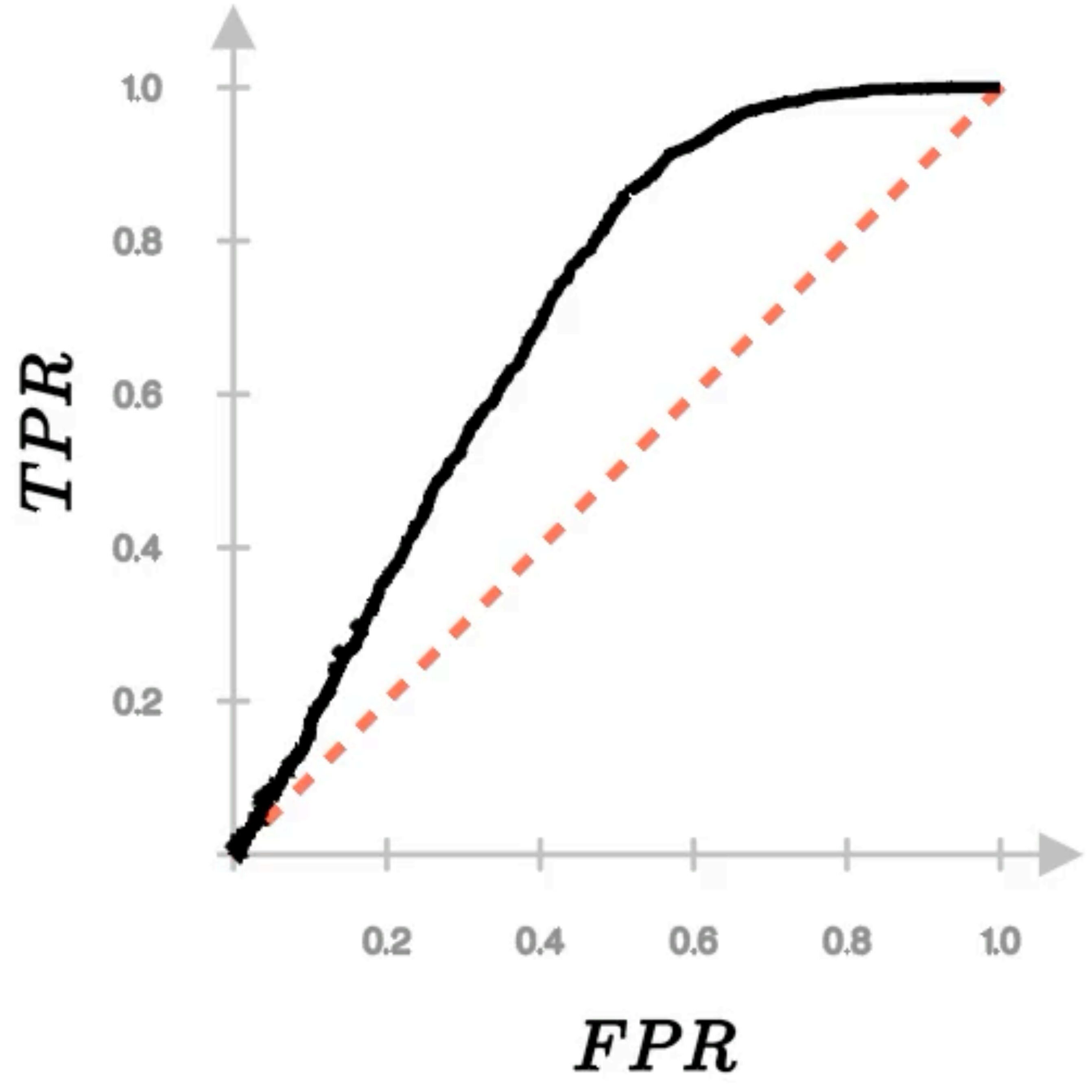


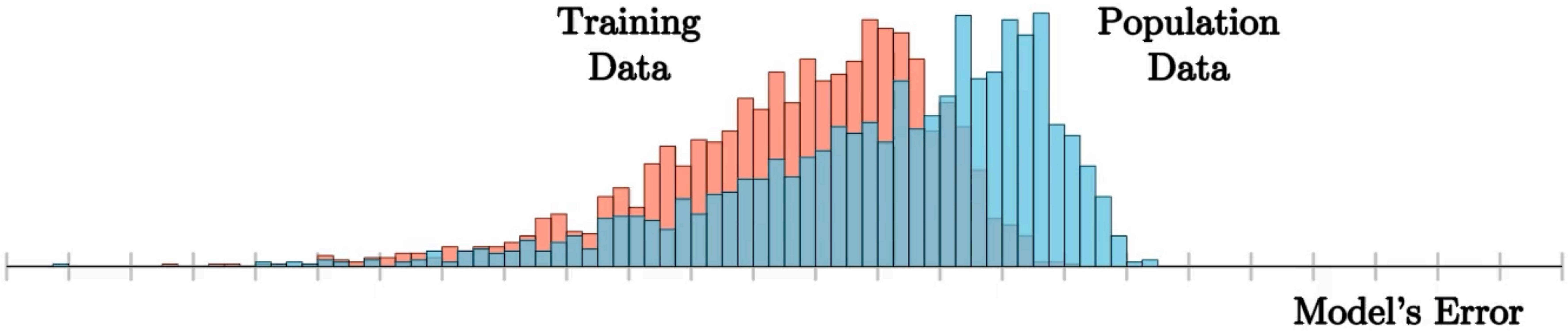
Overfitting

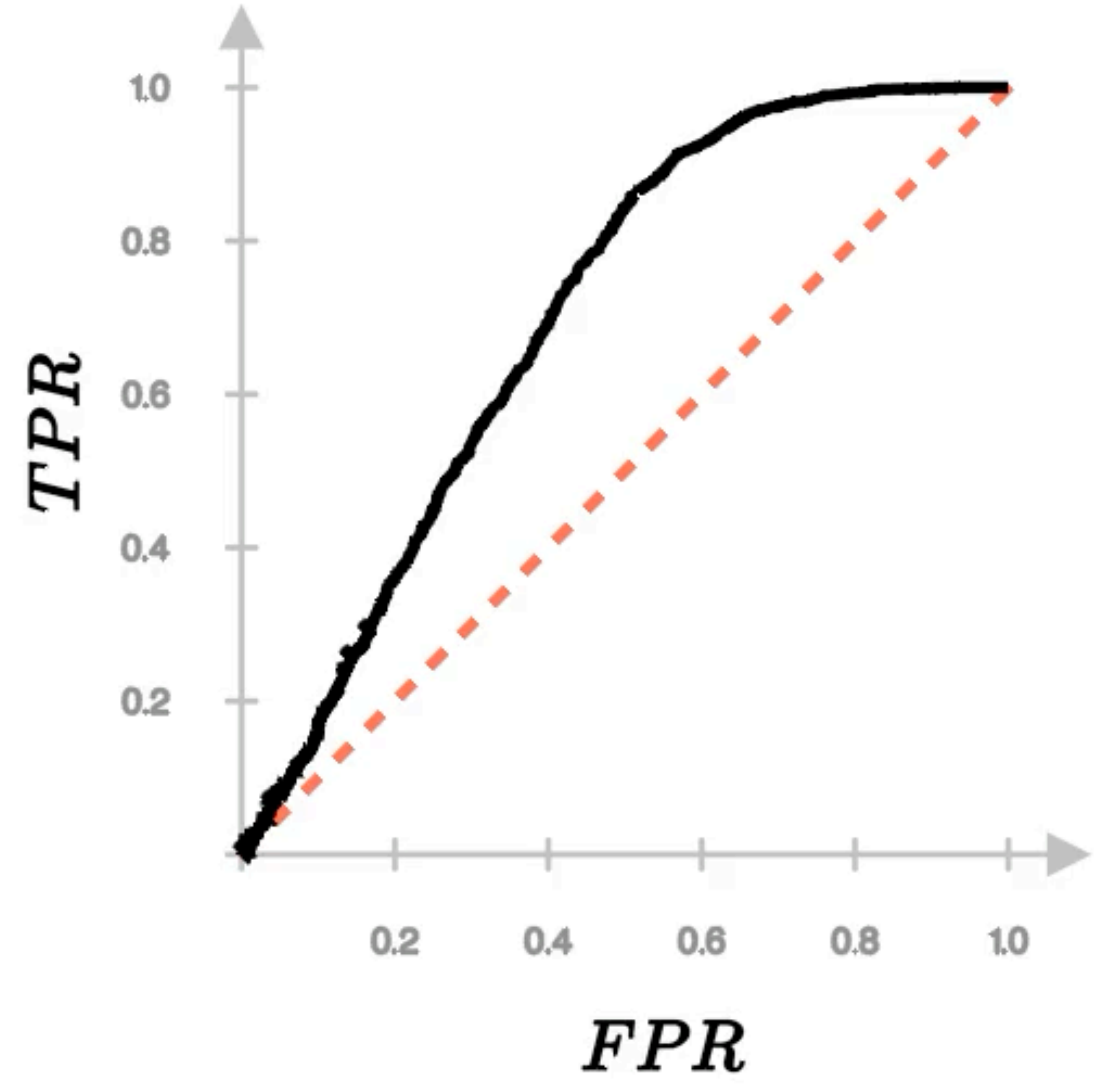
The behavior of the model on data distributions

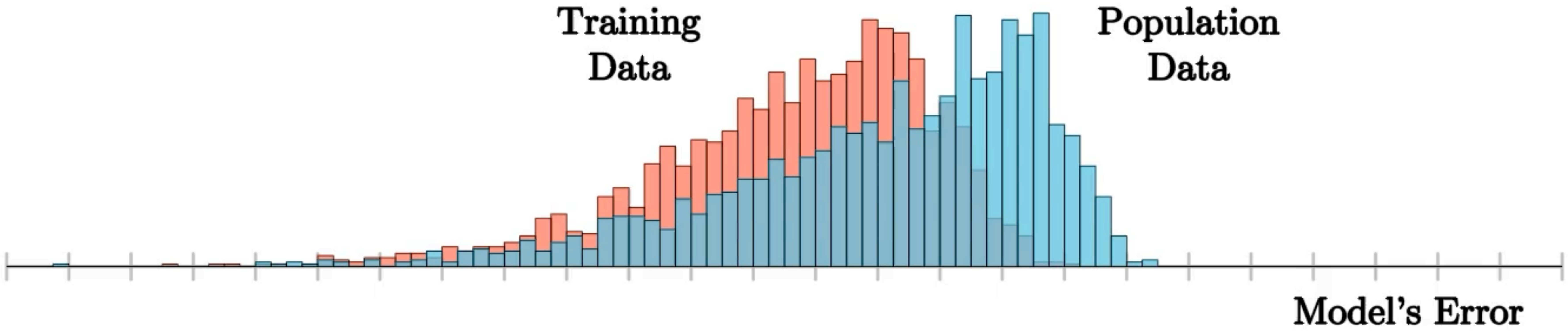


Where does this attack make errors?

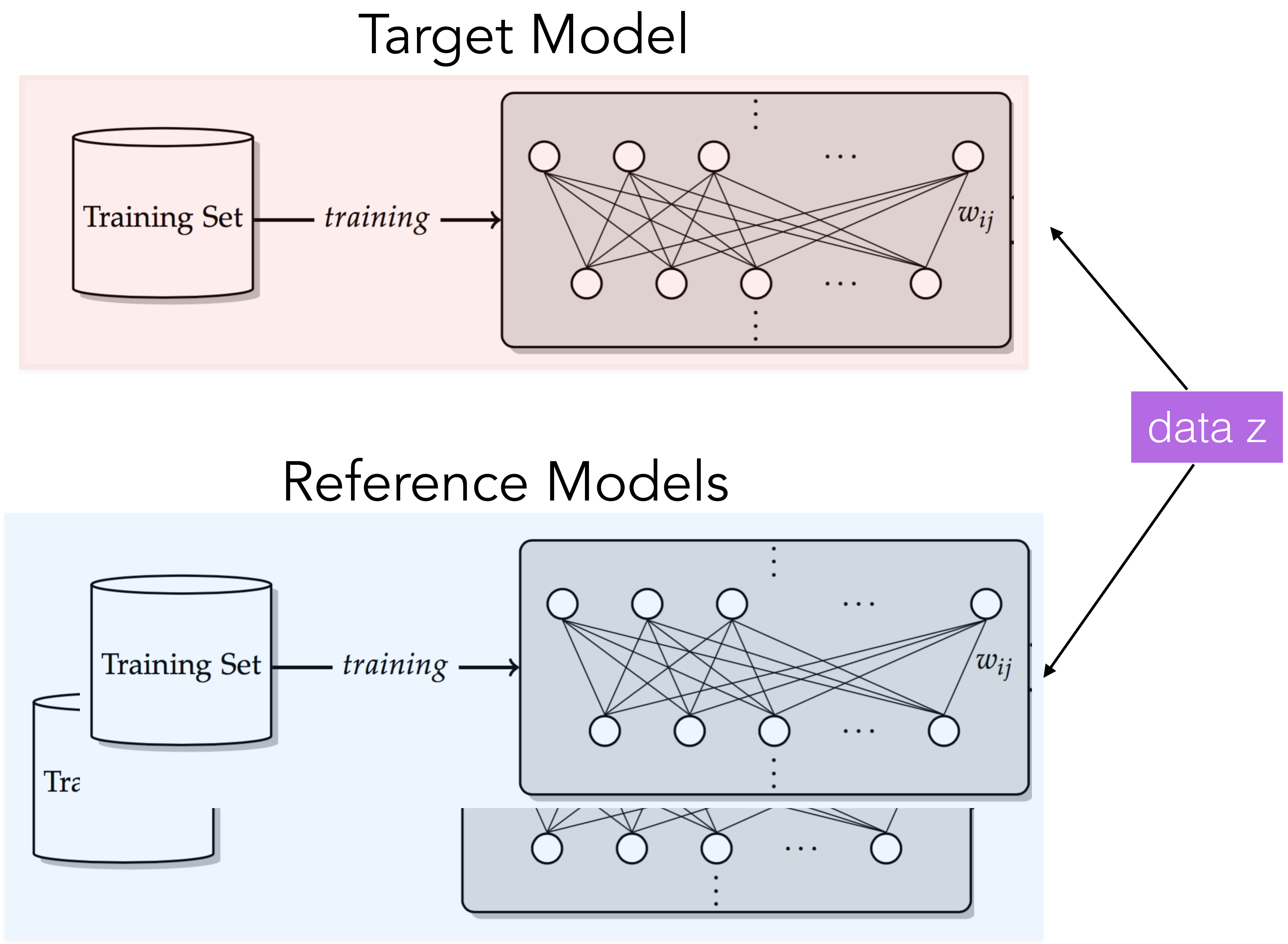

1.0
0.8
0.6
0.4
0.2
TPR
0.2
0.4
0.6
0.8
1.0
FPR


Training
Data
Population
Data
Model's Error


1.0
0.8
0.6
0.4
0.2
TPR
0.2
0.4
0.6
0.8
1.0
FPR


Training
Data
Population
Data
Model's Error

How to perform a more accurate analysis?

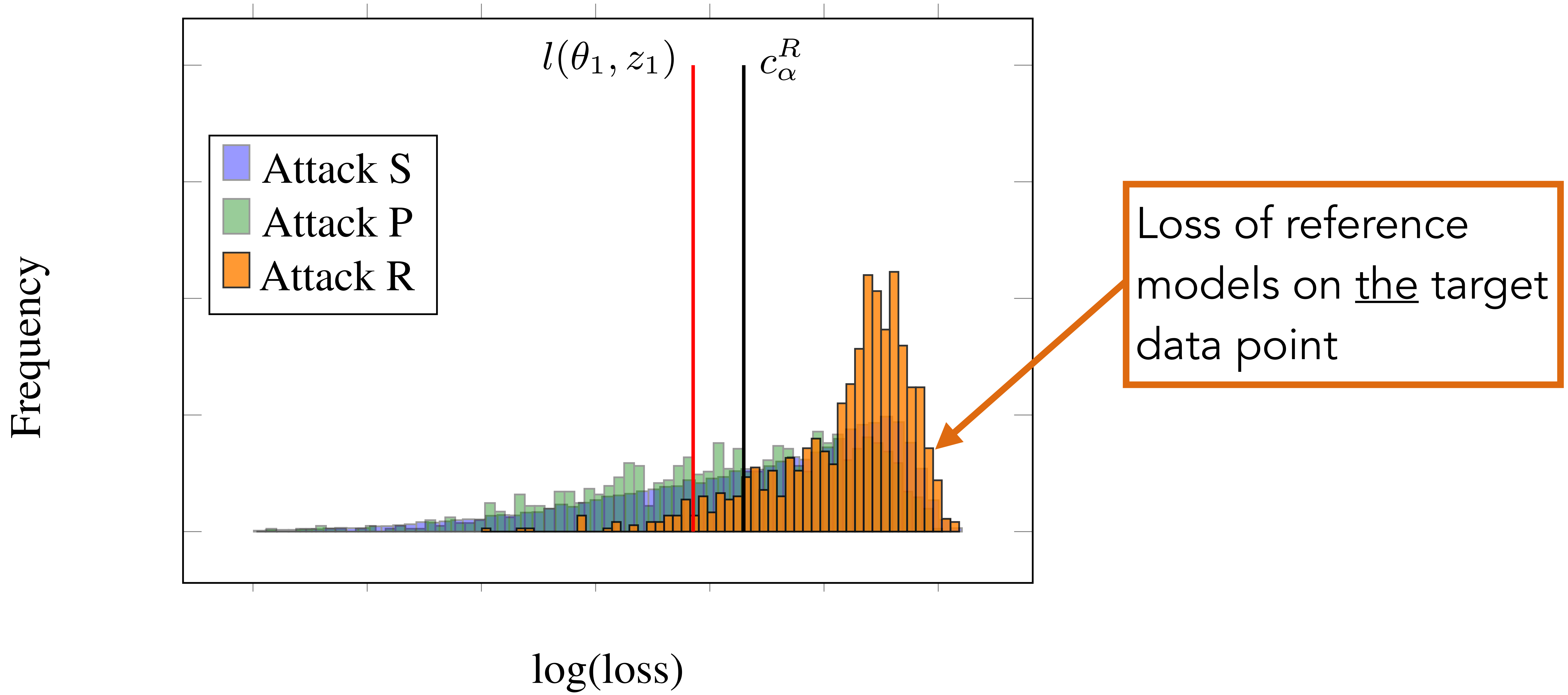


[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22

Membership Inference via Reference Models

$$\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(x_z, y_z), \textbf{ reject } H_0$$

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



Reason for Leakage?

Average Memorization The behavior of models on a data point, averaged over the remaining training data having been sampled from a distribution


Model’s Error on x

Can we perform an even more accurate privacy analysis?

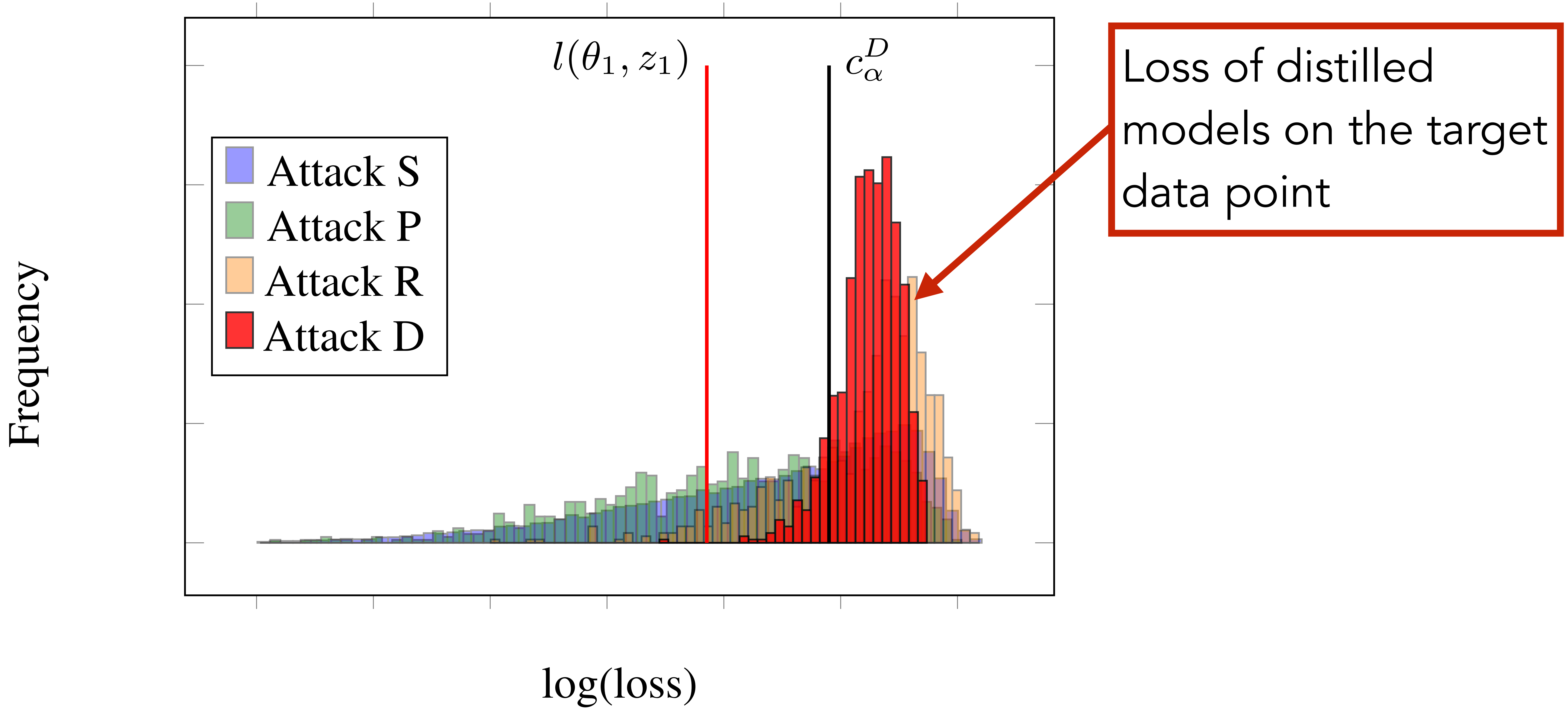
The objective is to get as close as possible to the leave-one-out attack, where the adversary knows all “other” data in the training set

- Train reference models that have a large agreement with the target model on all the training data, except the target data
- How? Use model distillation — Reference models are distilled versions of the target models

Membership Inference via Distilled Models

$$\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta, x_z, y_z), \ \textbf{reject } H_0$$

- Learn a threshold from the loss distribution of target data on distilled models
- Note that the threshold depends on both target data and the target model

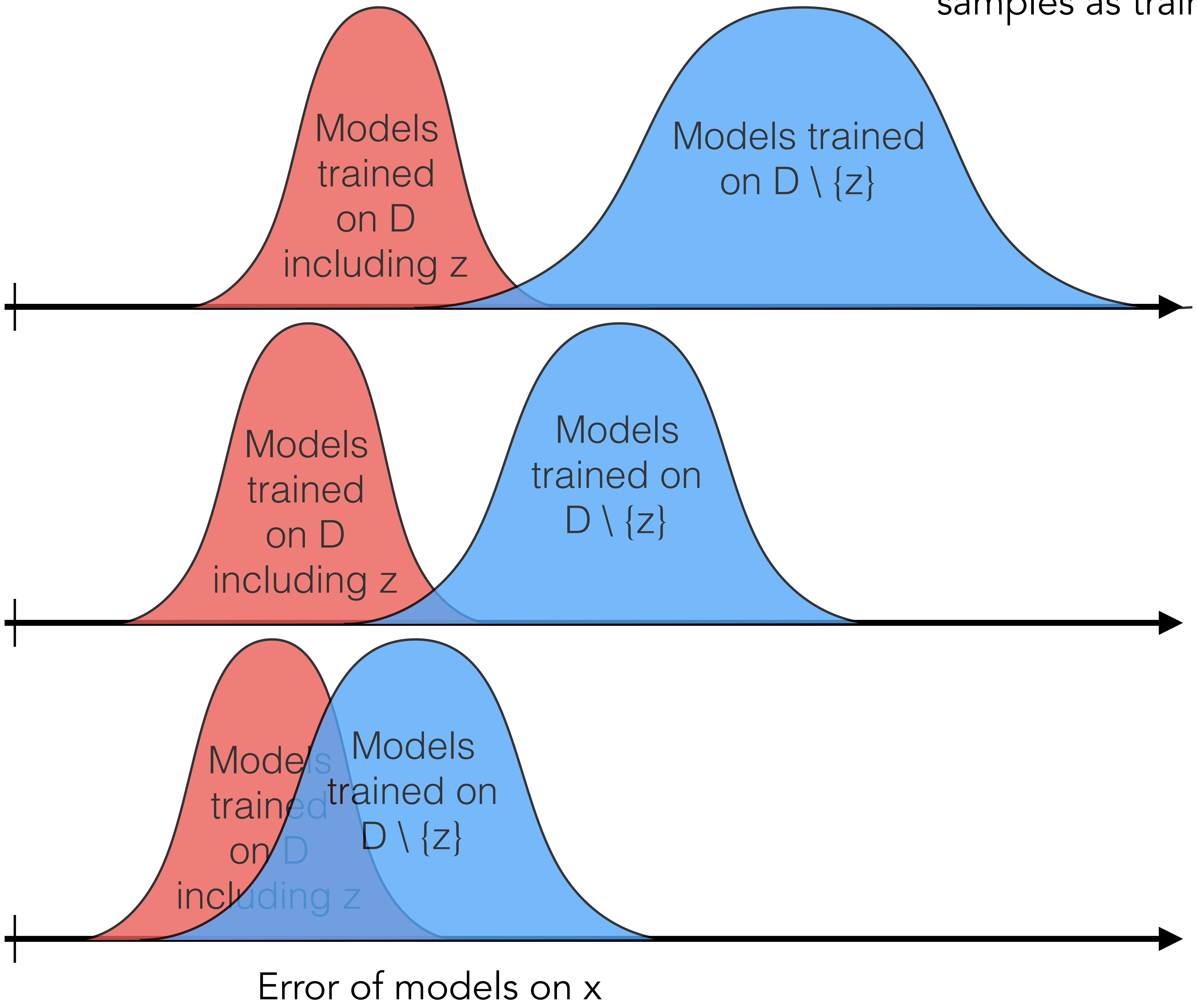


Reason for Leakage?

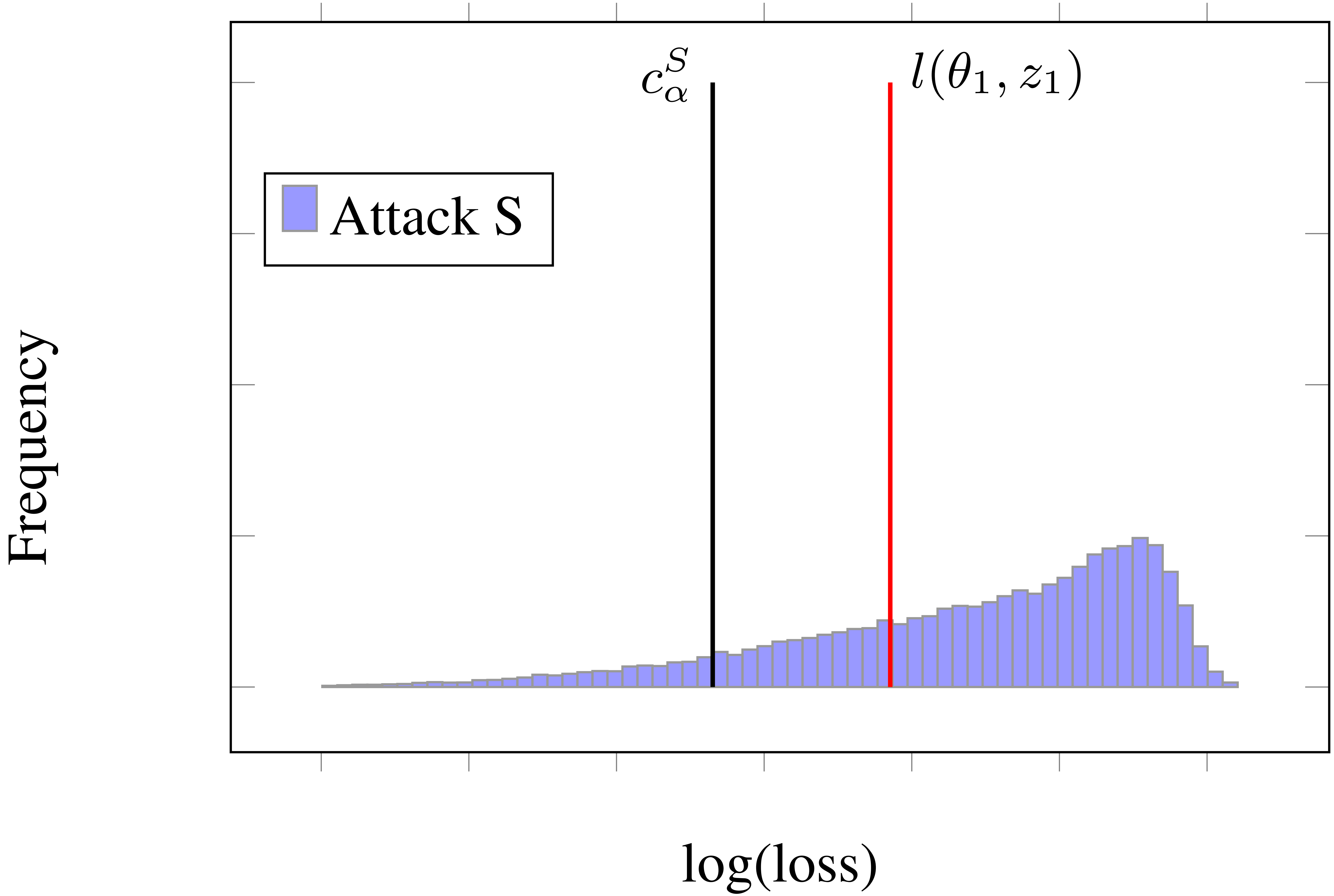
(Conditional) Memorization
The behavior of models on a data point, given a specific set of samples as training data

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

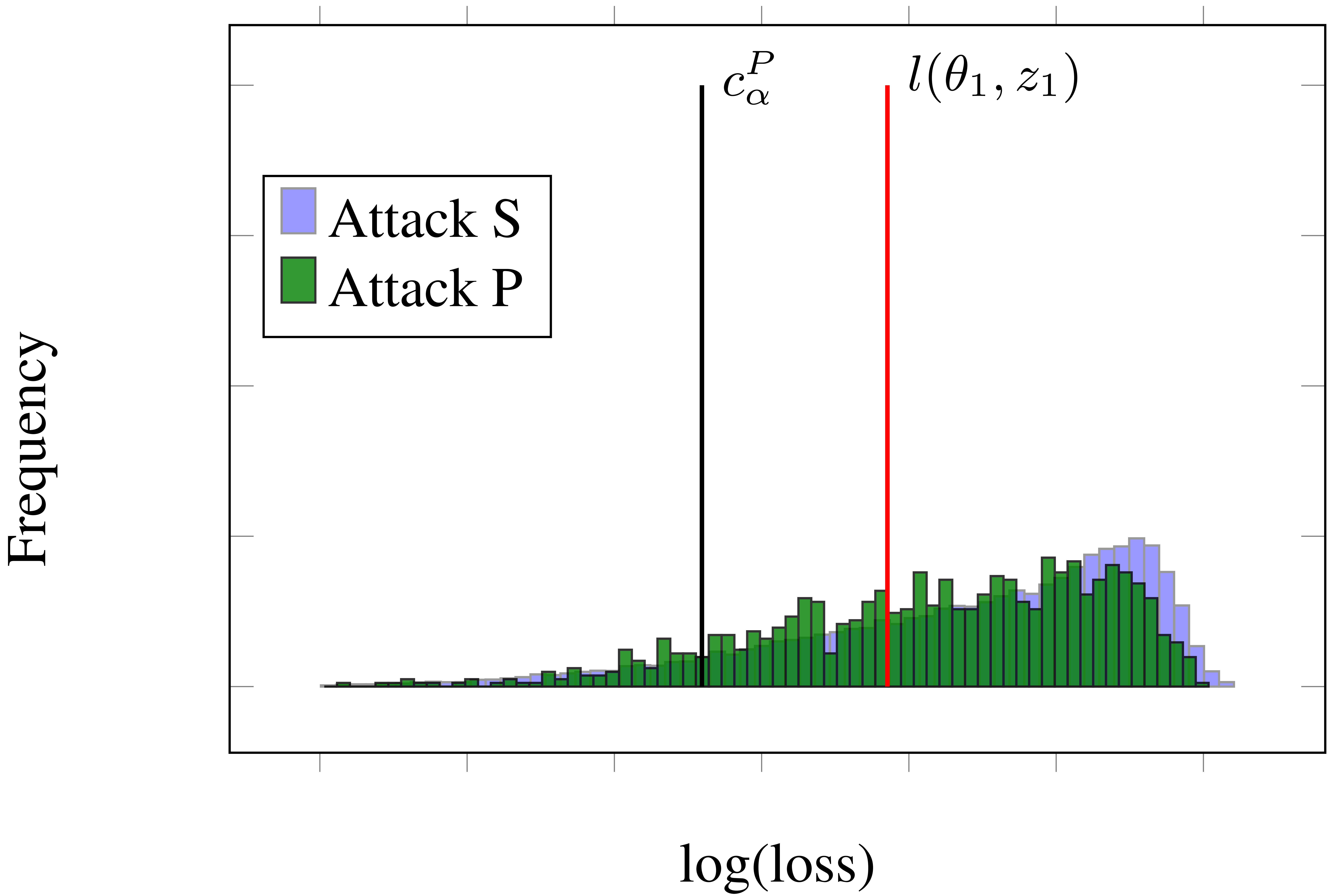
Higher Leakage



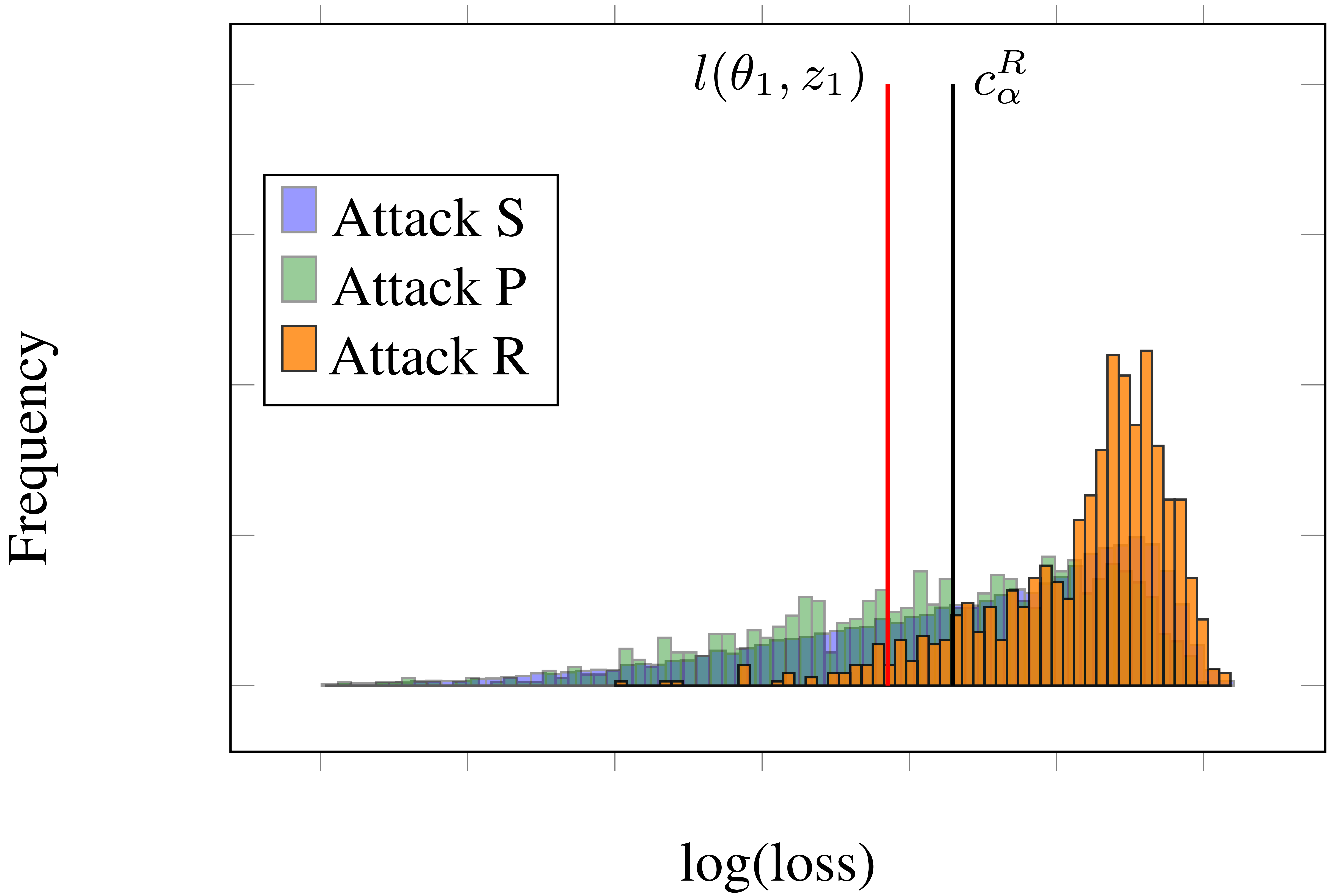
Less memorization on z, given D

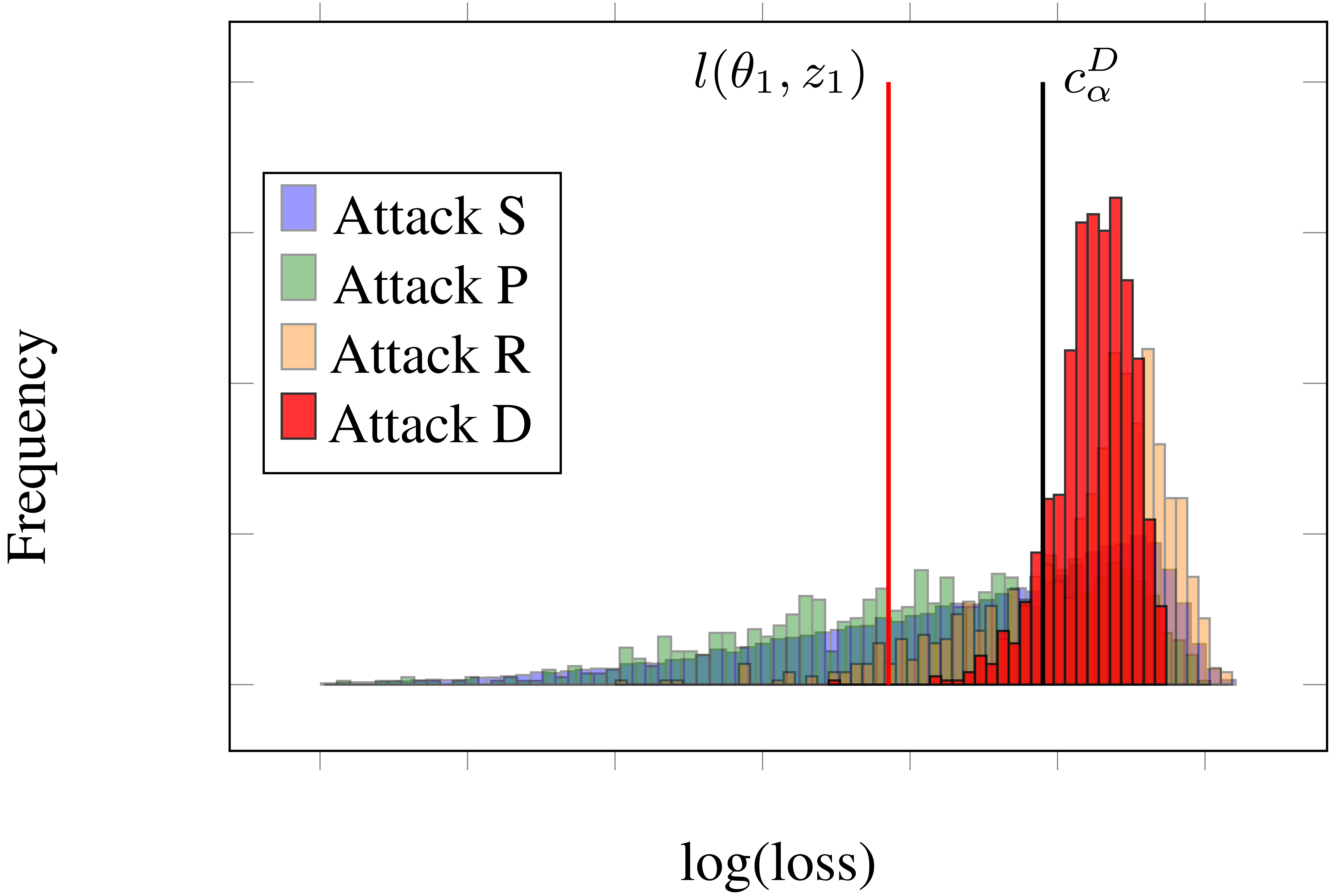


c_α^S
$l(\theta_1, z_1)$
Attack S
Frequency
log(loss)

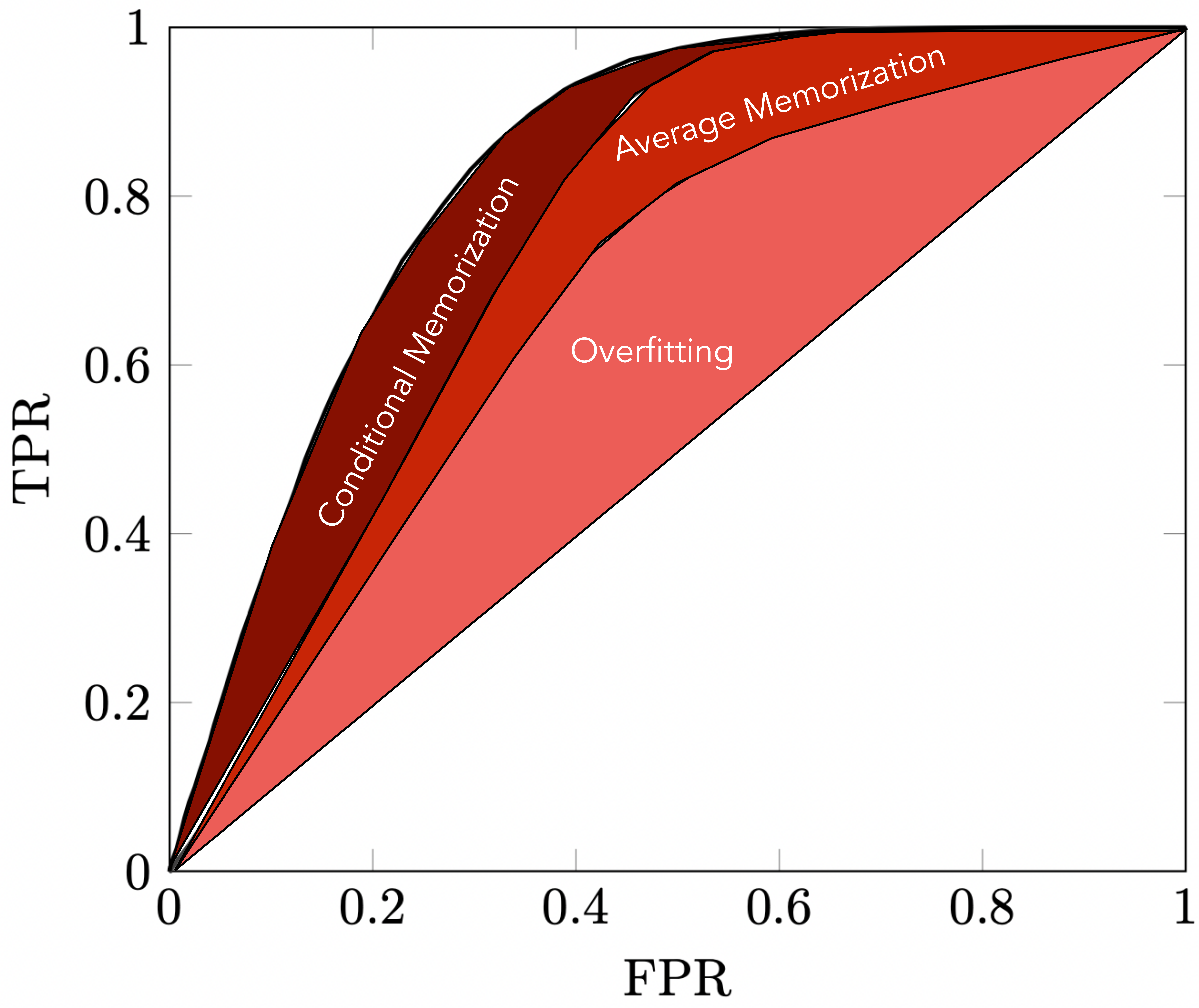


c_α^P
$l(\theta_1, z_1)$
Attack S
Attack P
Frequency
log(loss)


$l(\theta_1, z_1)$
c_α^R
Attack S
Attack P
Attack R
Frequency
log(loss)



Reduce attack uncertainty to only 1 bit (of membership)



1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
TPR
FPR
Conditional Memorization
Average Memorization
Overfitting

Auditing Data Privacy Using Privacy Meter

- Given the privacy vulnerabilities of models, enabling access to models without auditing them (and mitigating the risks) is not much worse than allowing unauthorised access to data
- Privacy Meter (privacy-meter.com) tool aids regulatory compliance, through a systematic method to audit data privacy for a wide range of machine learning algorithms

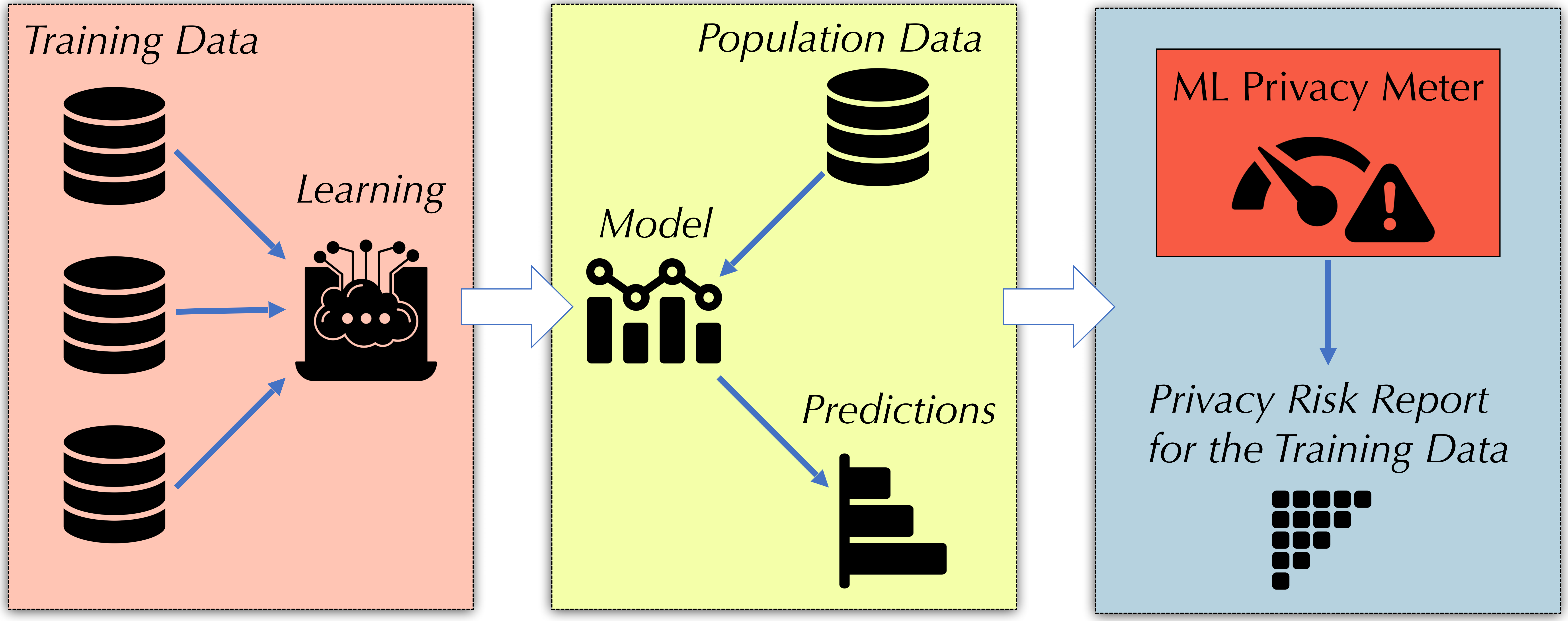
[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22

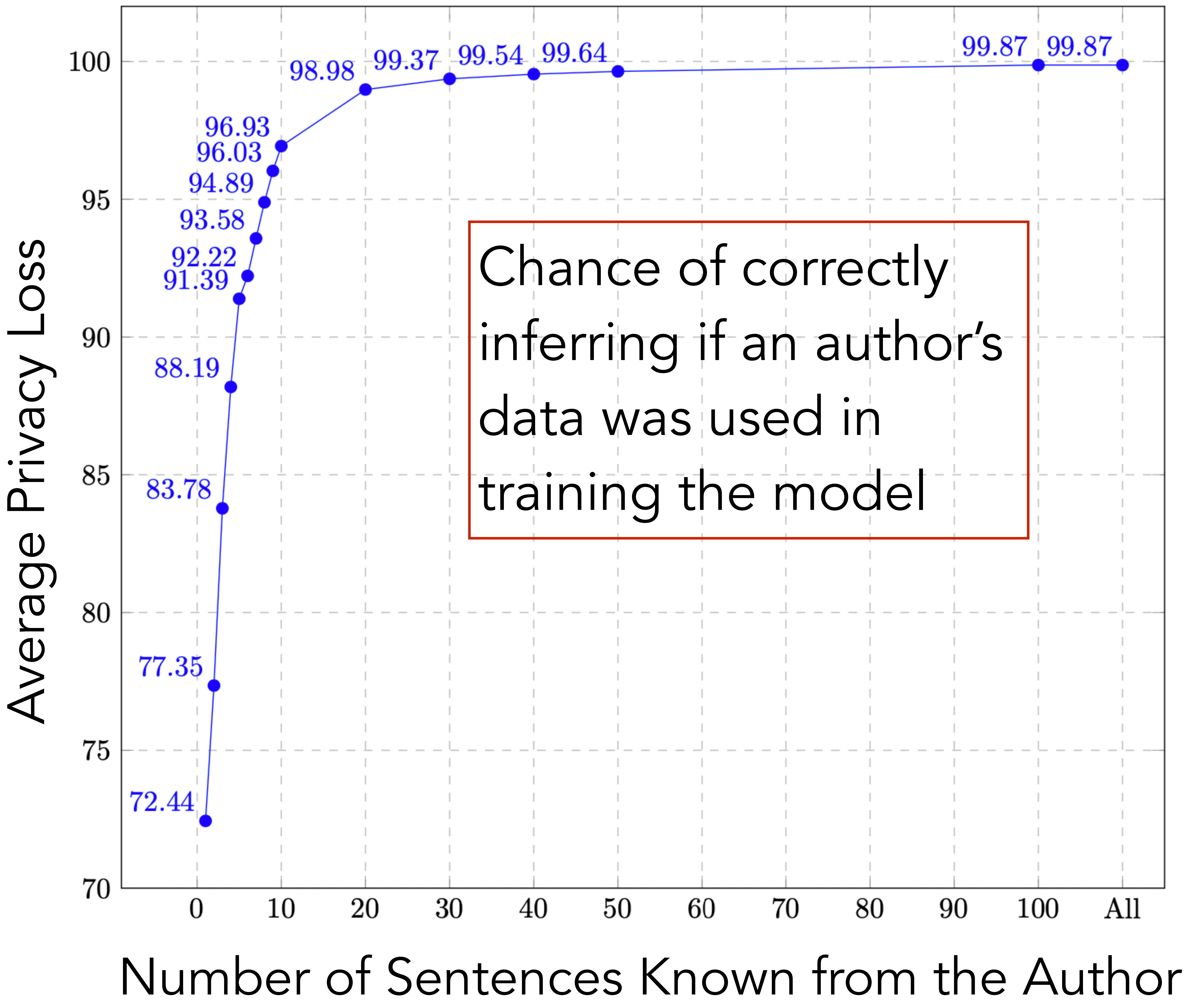
Privacy Meter

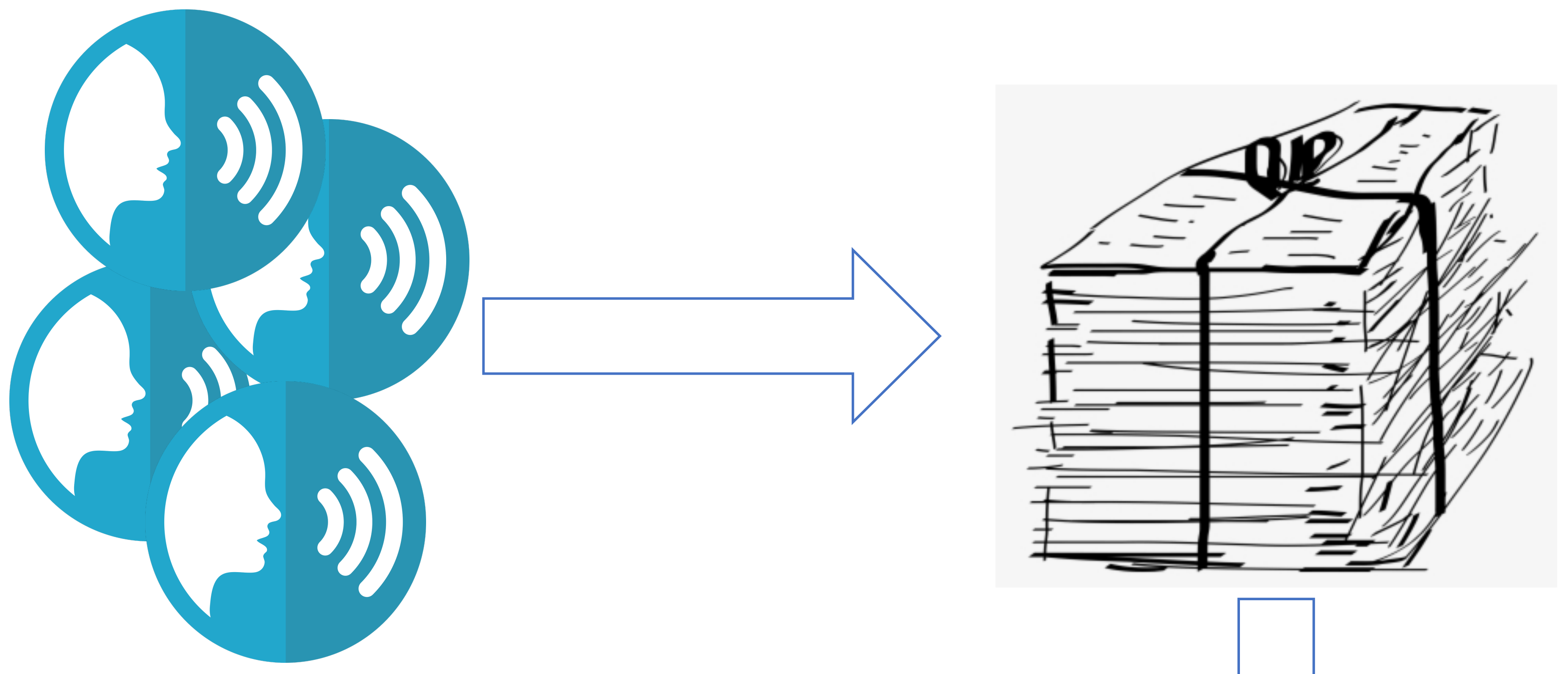
privacy-meter.com



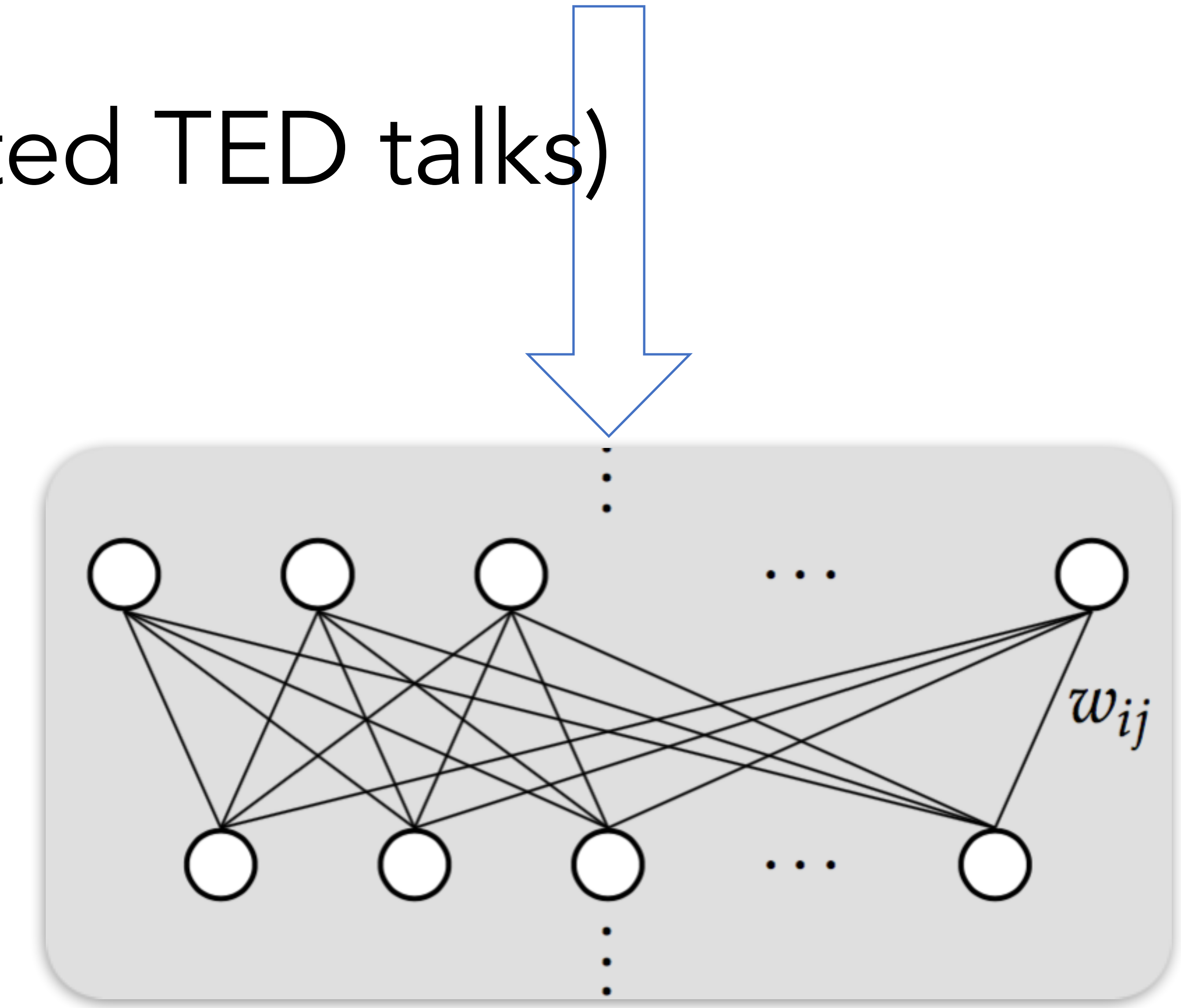
Privacy Meter is an open source tool that enables quantifying the privacy risks of statistical and machine learning models.

Example: Language Generative Model



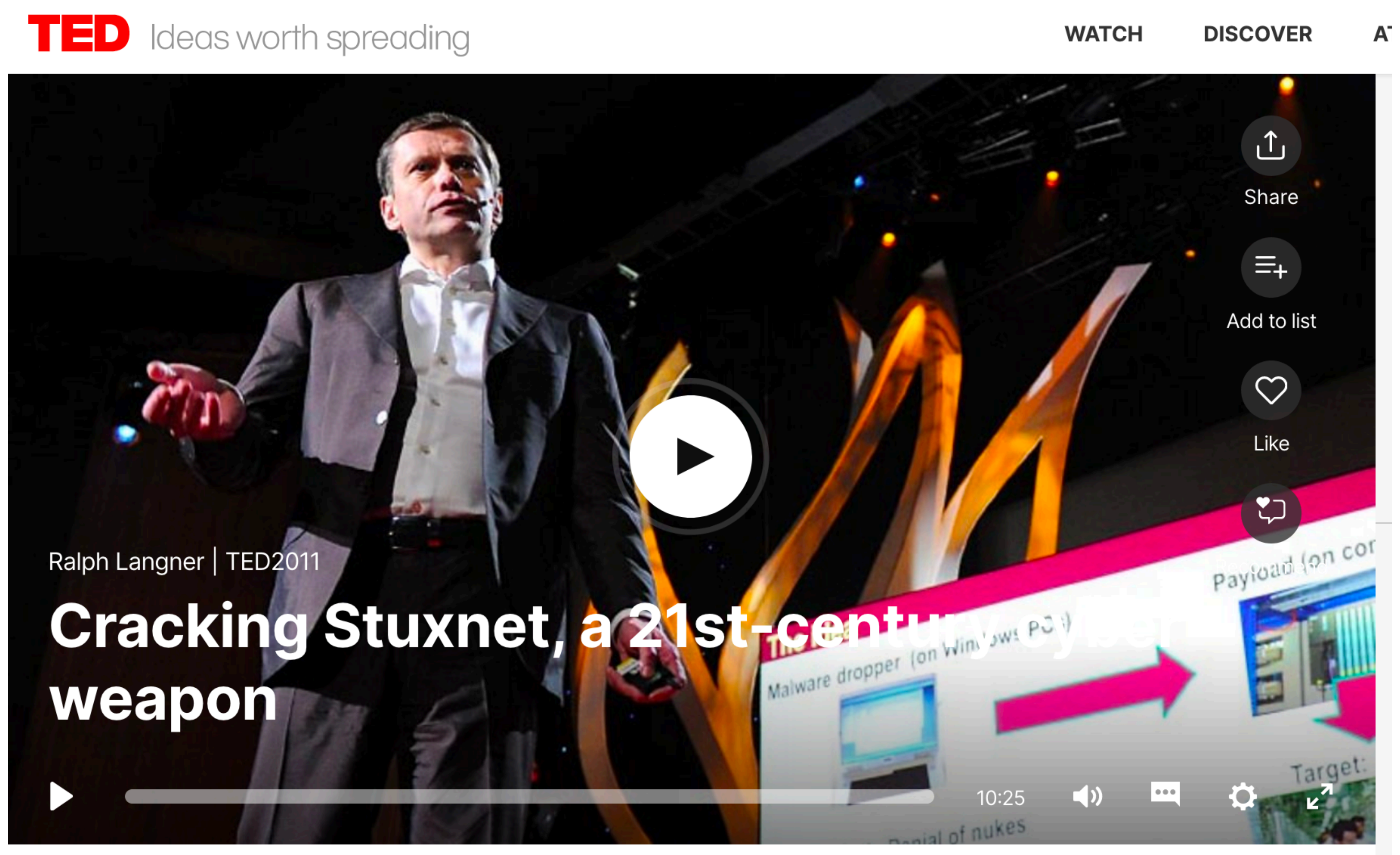


(Speaker Annotated TED talks)
SATED dataset



[Song, Shmatikov] Auditing data provenance in text-generation models, ACM SIGKDD '19

Examples of Vulnerable Training Data



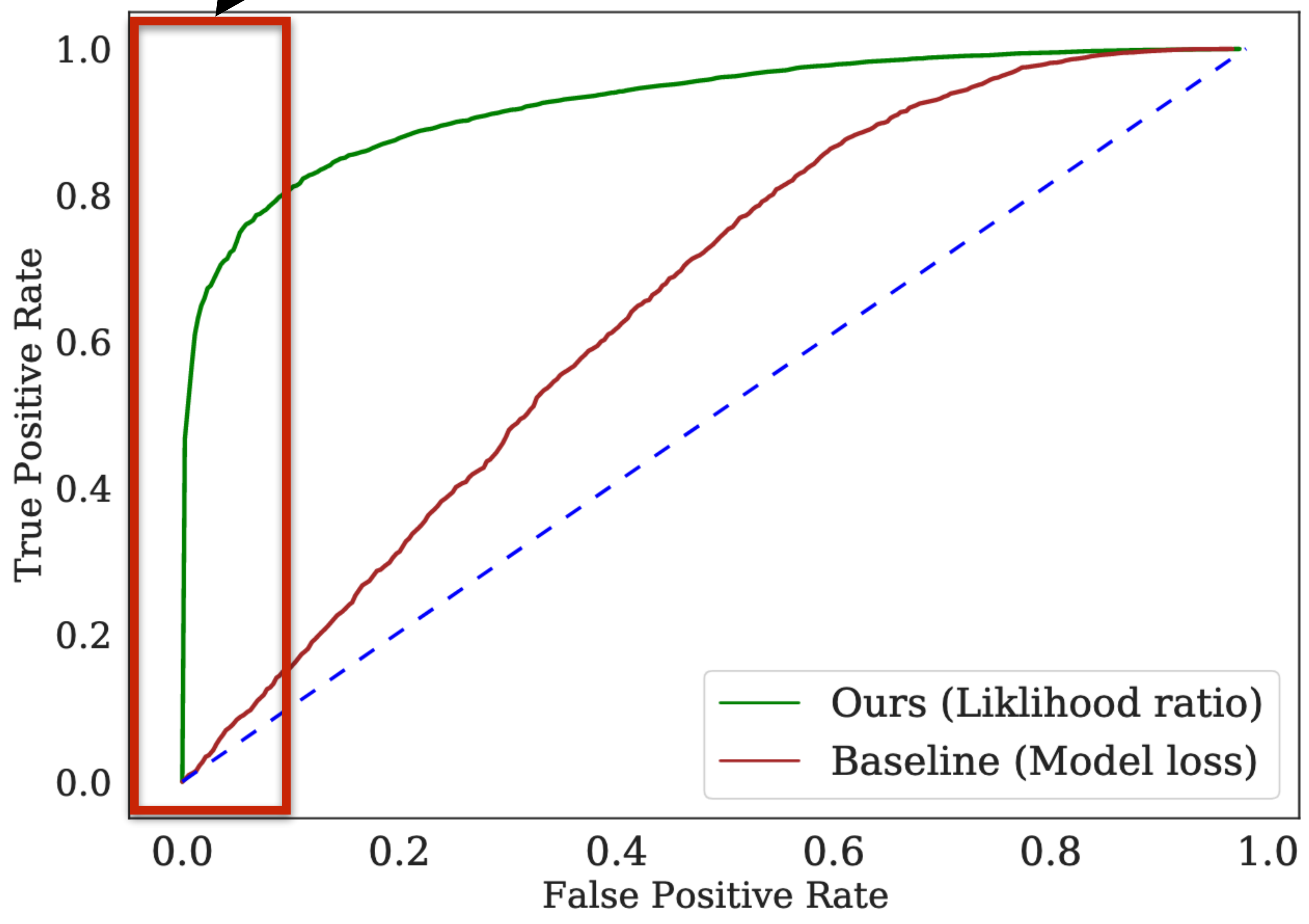
But it gets worse. And this is very important, what I'
generic. It doesn't have anything to do, in specifics,
would work as well, for example, in a power plant or
don't have -- as an attacker -- you don't have to del
the case of Stuxnet. You could also use conventiona

This year, Germany is celebrating the 25th anniversar
1989, the Communist regime was moved away, the B
German Democratic Republic, the GDR, in the East wa
in the West to found today's Germany. Among many c
the East German secret police, known as the Stasi. O
were opened to the public, and historians such as me
about how the GDR surveillance state functioned.

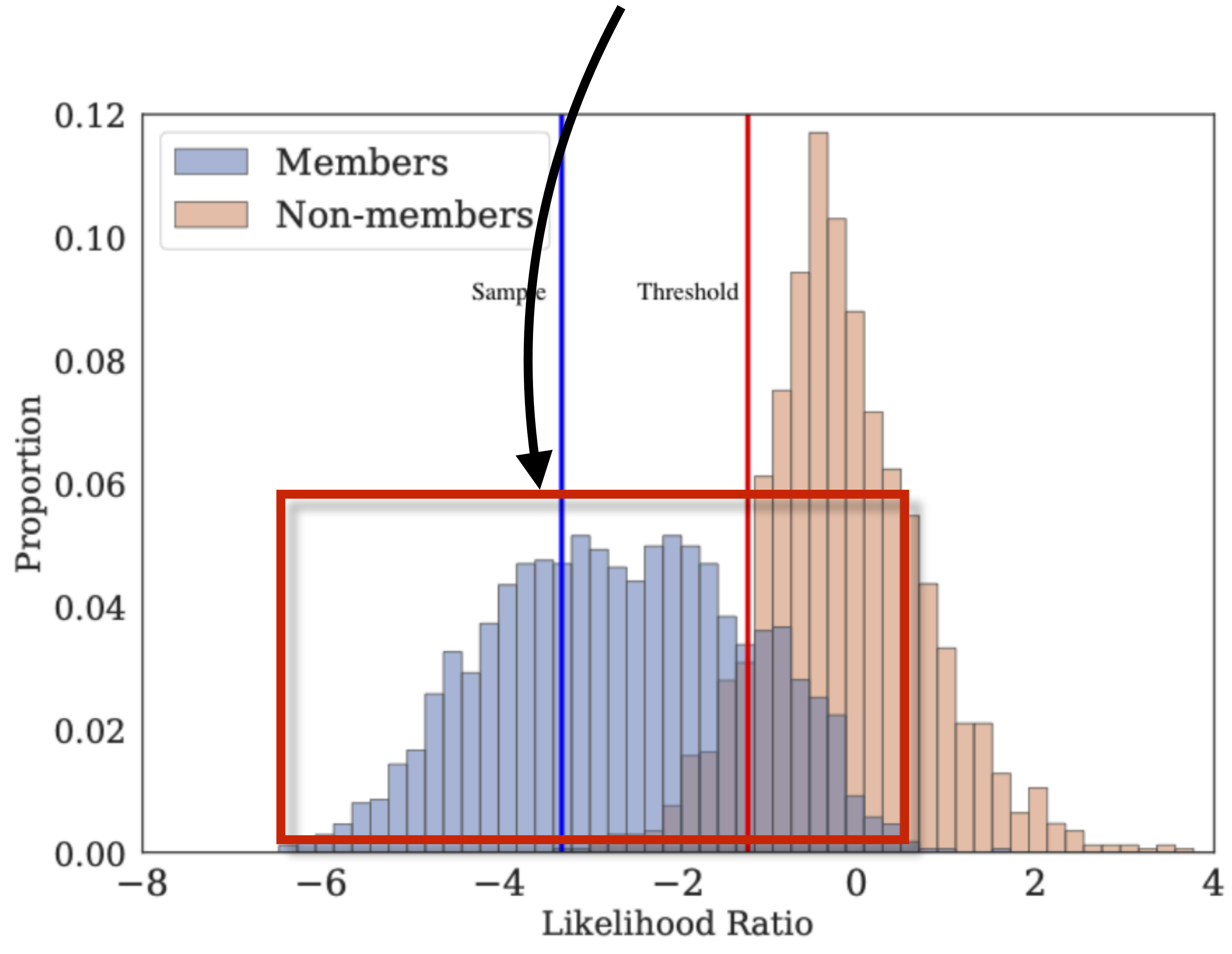
Example: Masked Language Models

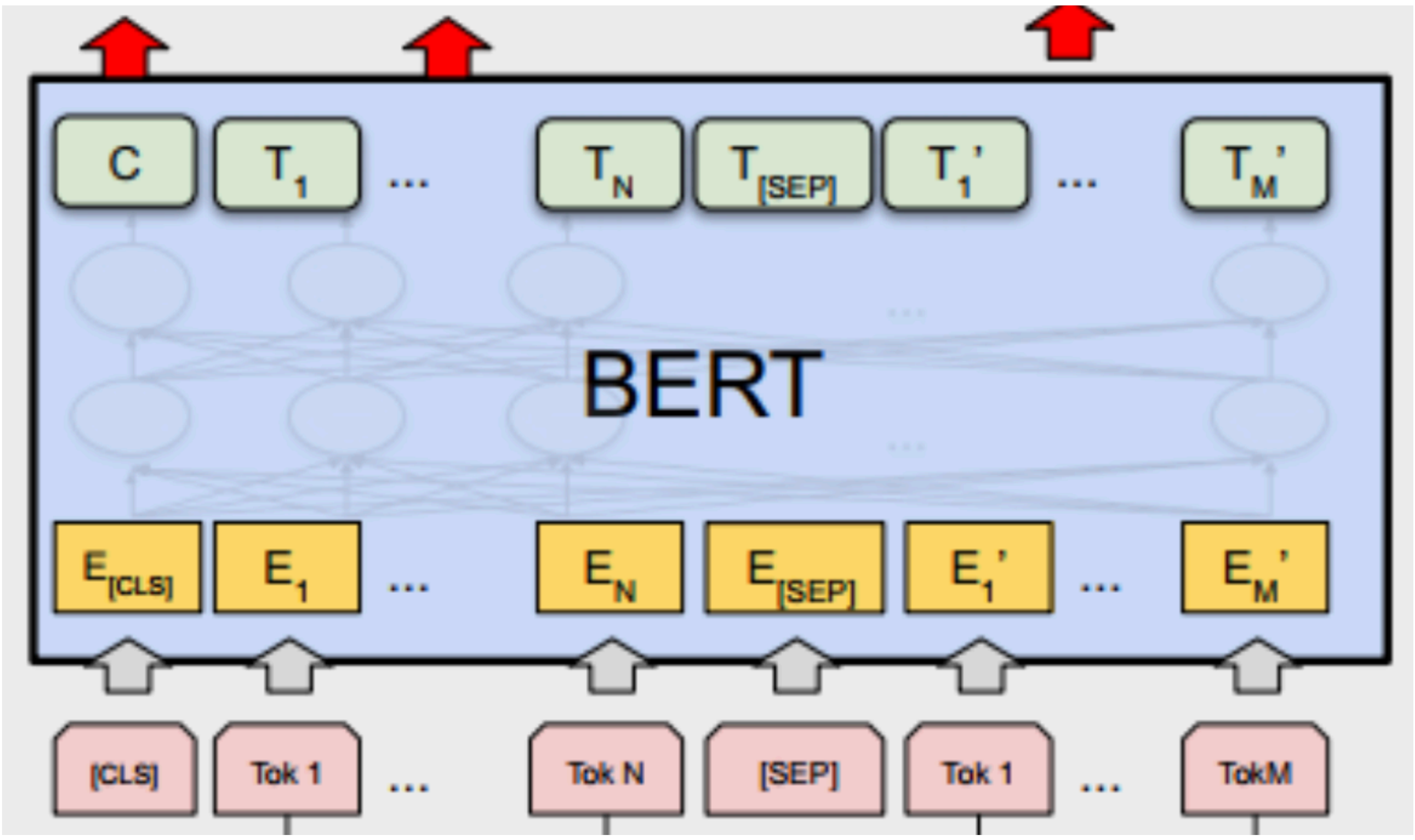
- Members of the training set are identifiable: Presence of any document in a training dataset can be inferred very accurately using membership inference attacks

Extremely high privacy risk!



Distinguishable members





ClinicalBERT

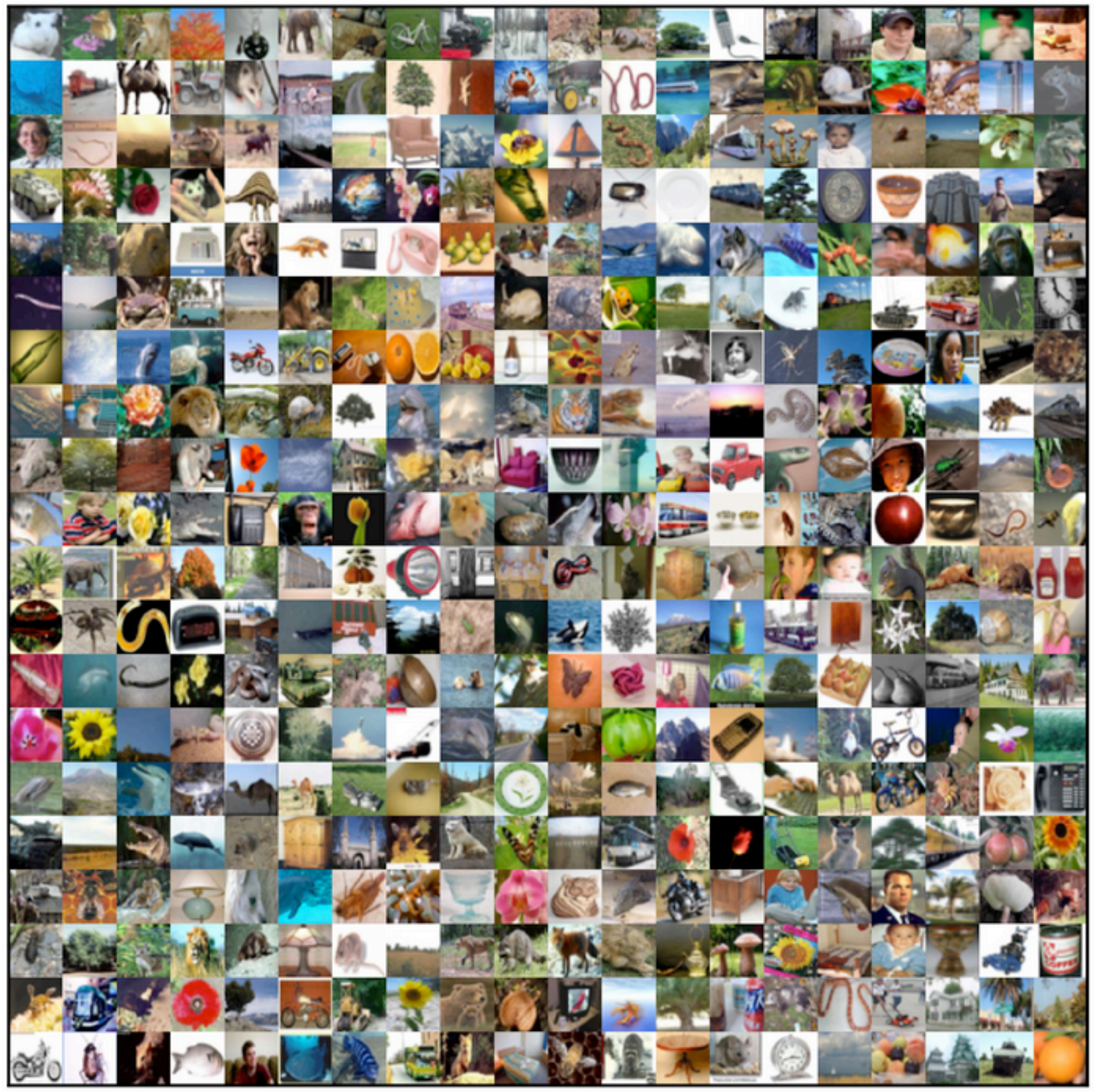
[Mireshghallah, Goyal, Uniyal, Berg-Kirkpatrick, Shokri] Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks, EMNLP'22

Example: Image Classification Tasks

Model	Number of Parameters	Prediction (Test) Accuracy	Privacy Risk
AlexNet	2.47 million	44%	75.1%
ResNet	1.7 million	73%	64.3%
DenseNet	25.62 million	82%	74.3%

Large capacity   **High** generalizability   **Low** privacy

CIFAR100 Image classification



[Feldman] Does Learning Require Memorization? A Short Tale about a Long Tail, STOC'20

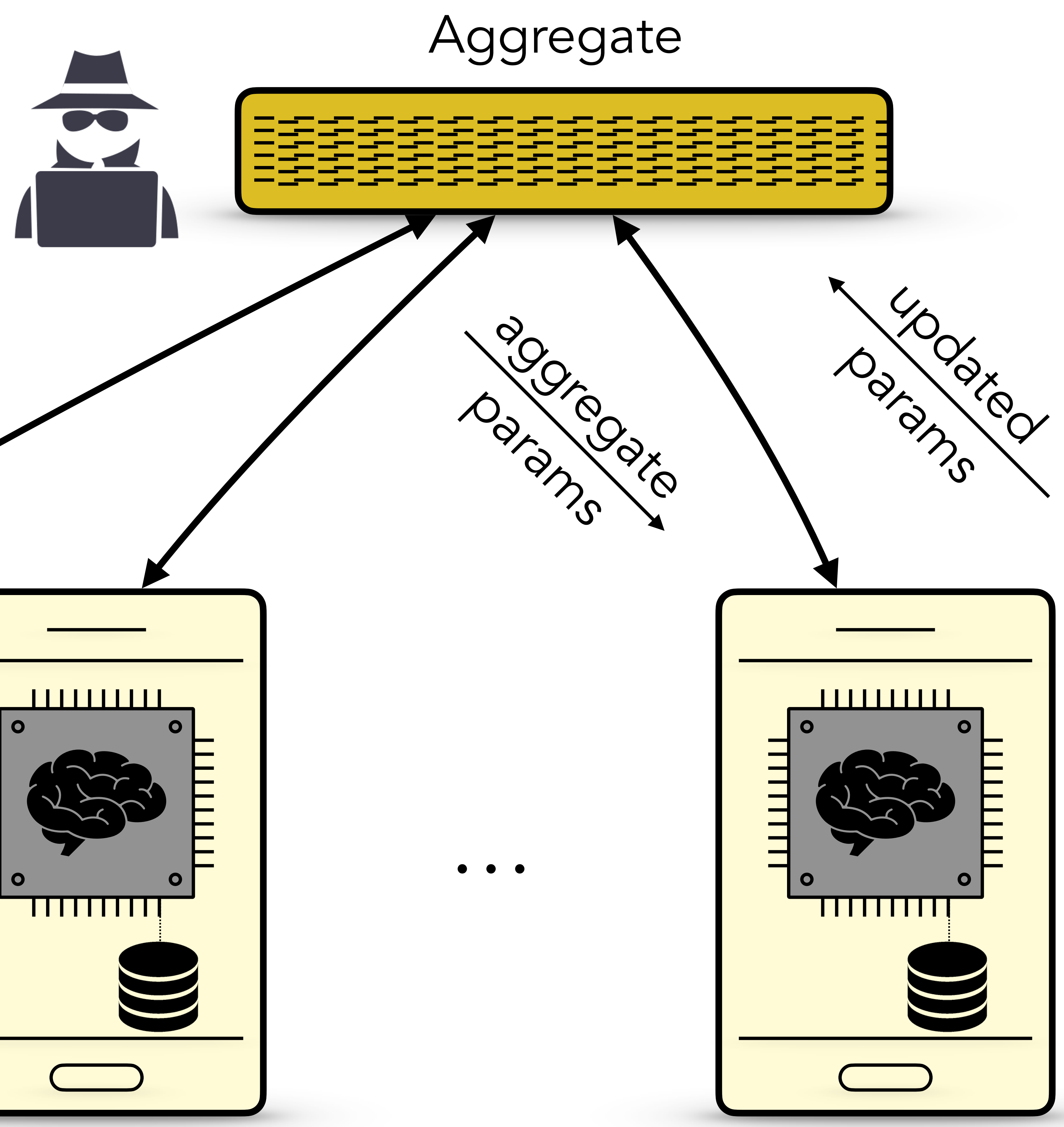
[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

Example: Federated Learning

Adversary can observe multiple snapshots of the model

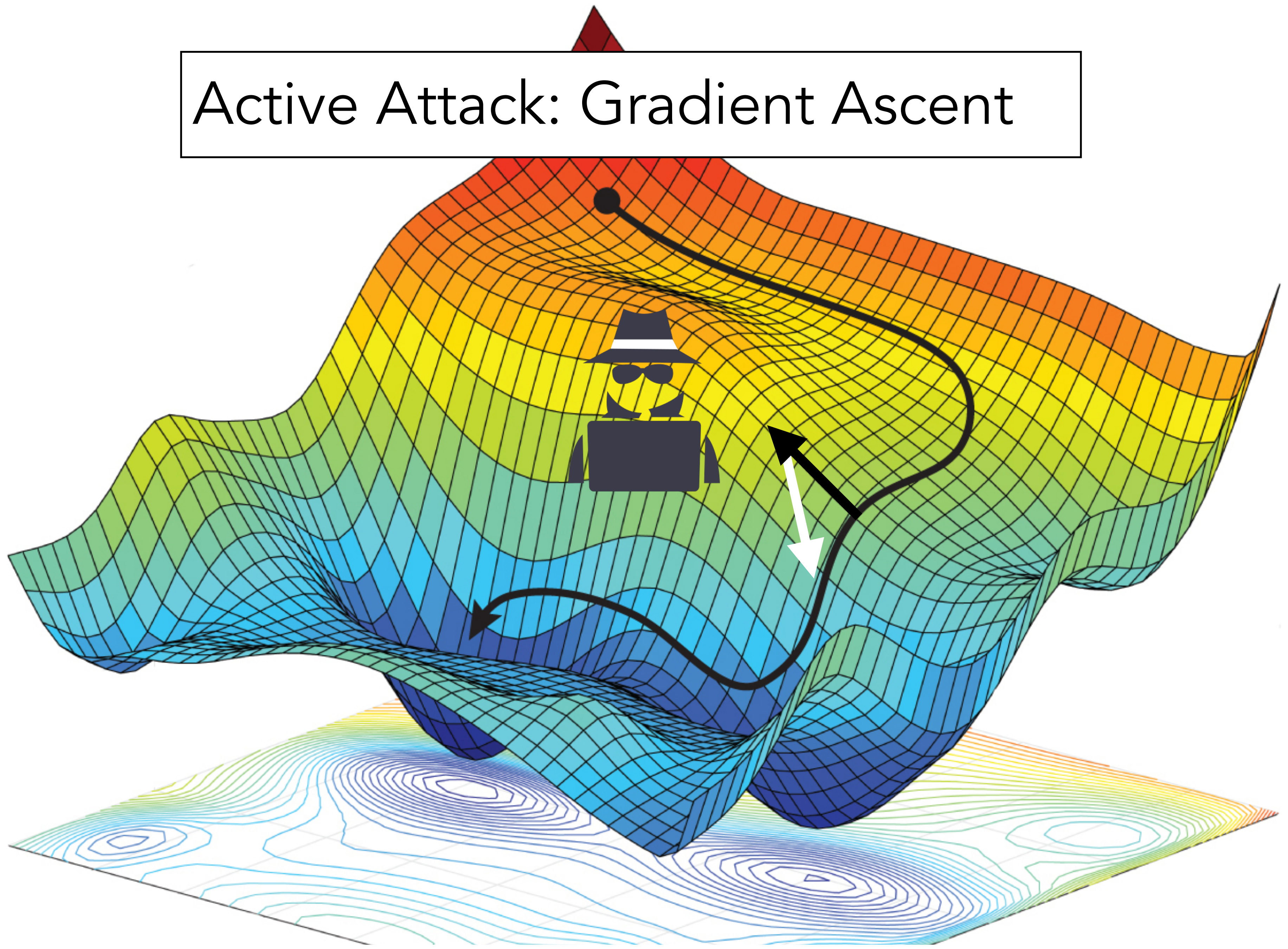
Observed Epochs	Attack Accuracy
5, 10, 15, 20, 25	57.4%
10, 20, 30, 40, 50	76.5%
50, 100, 150, 200, 250	79.5%
100, 150, 200, 250, 300	85.1%

CIFAR100-Alexnet



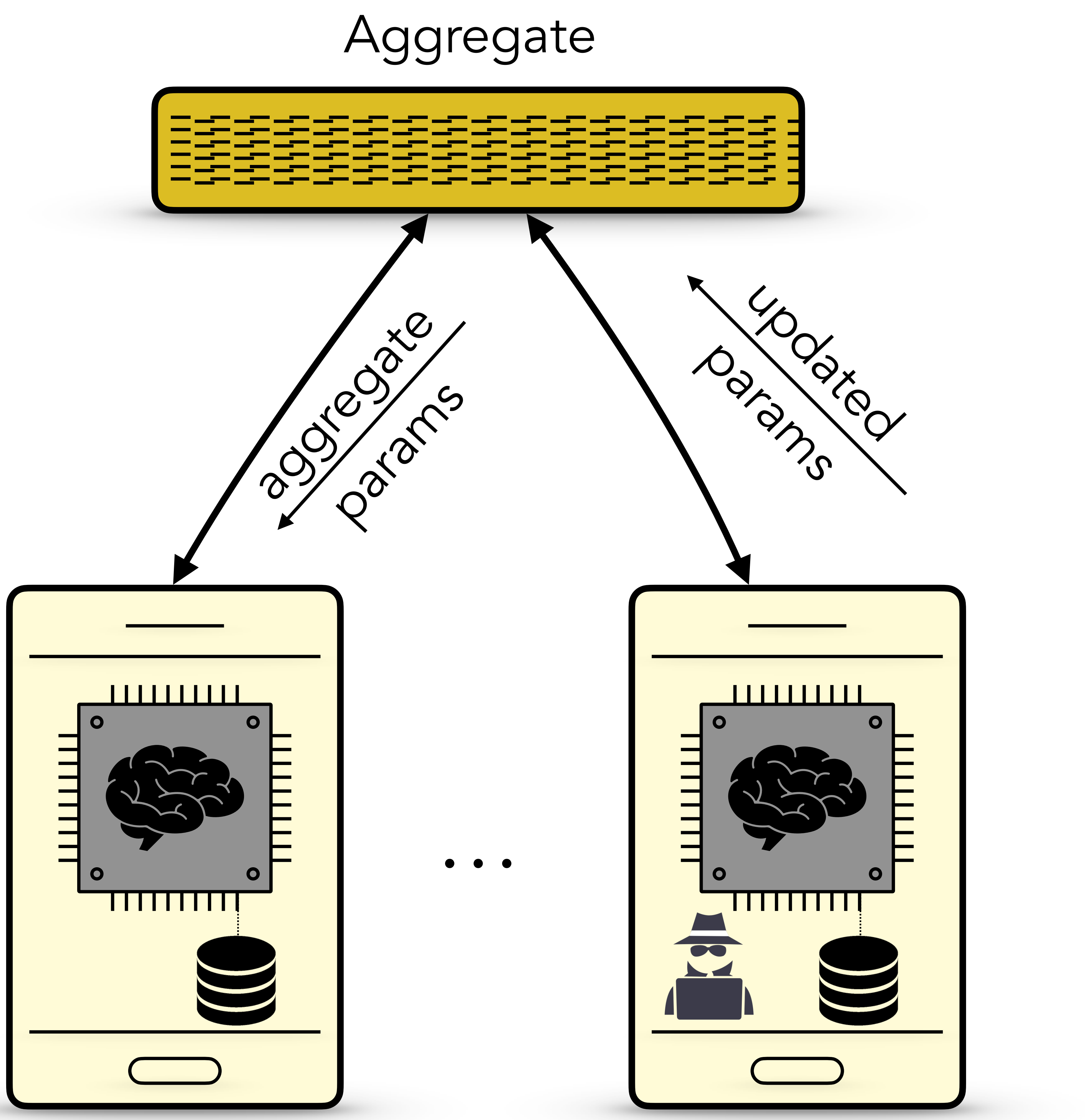
[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

Decentralized (Federated) Learning



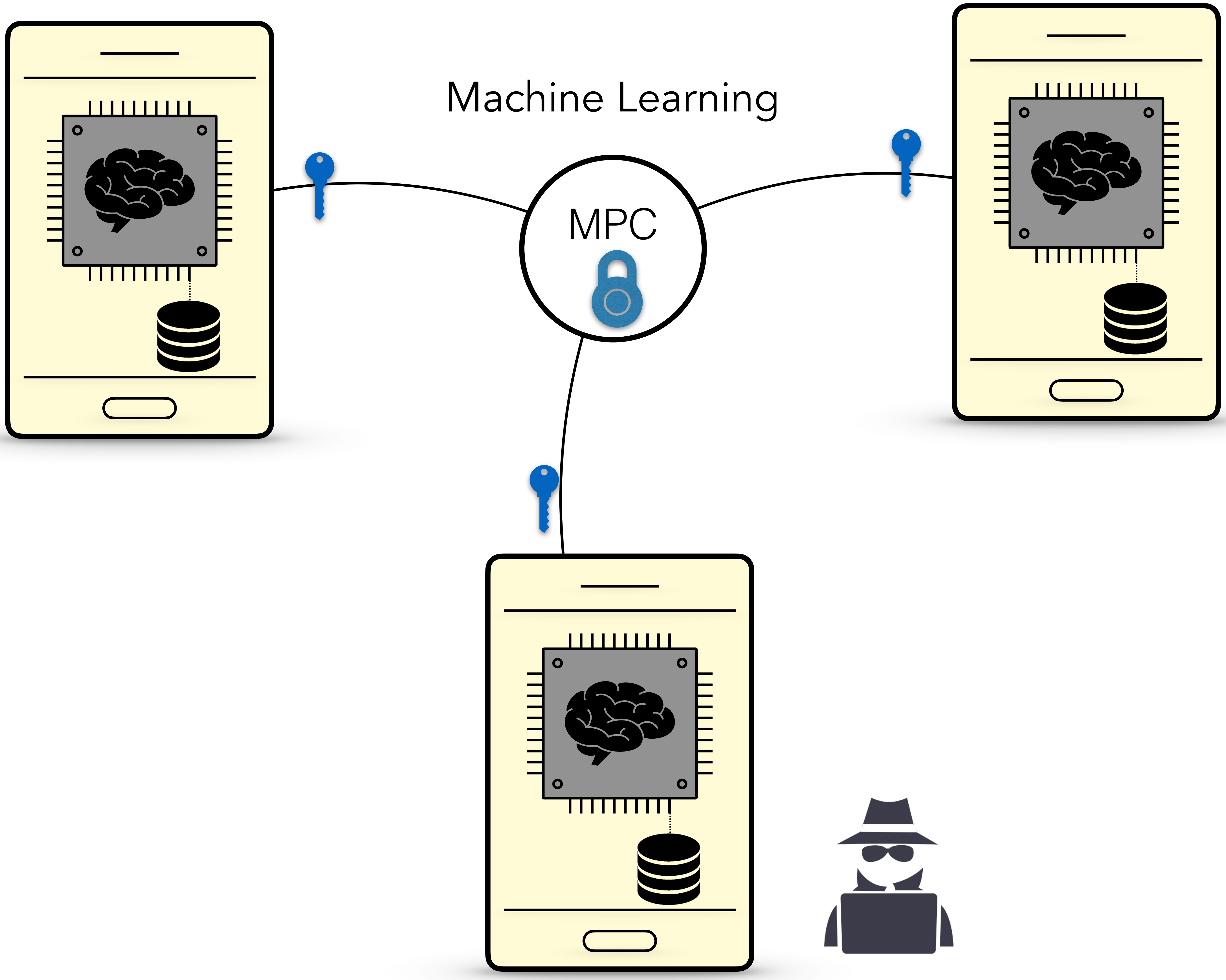
Increase loss on a particular data point x.

A participant corrects it back (by running gradient descent locally) only if x is part of its training set. => **membership leakage**



[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

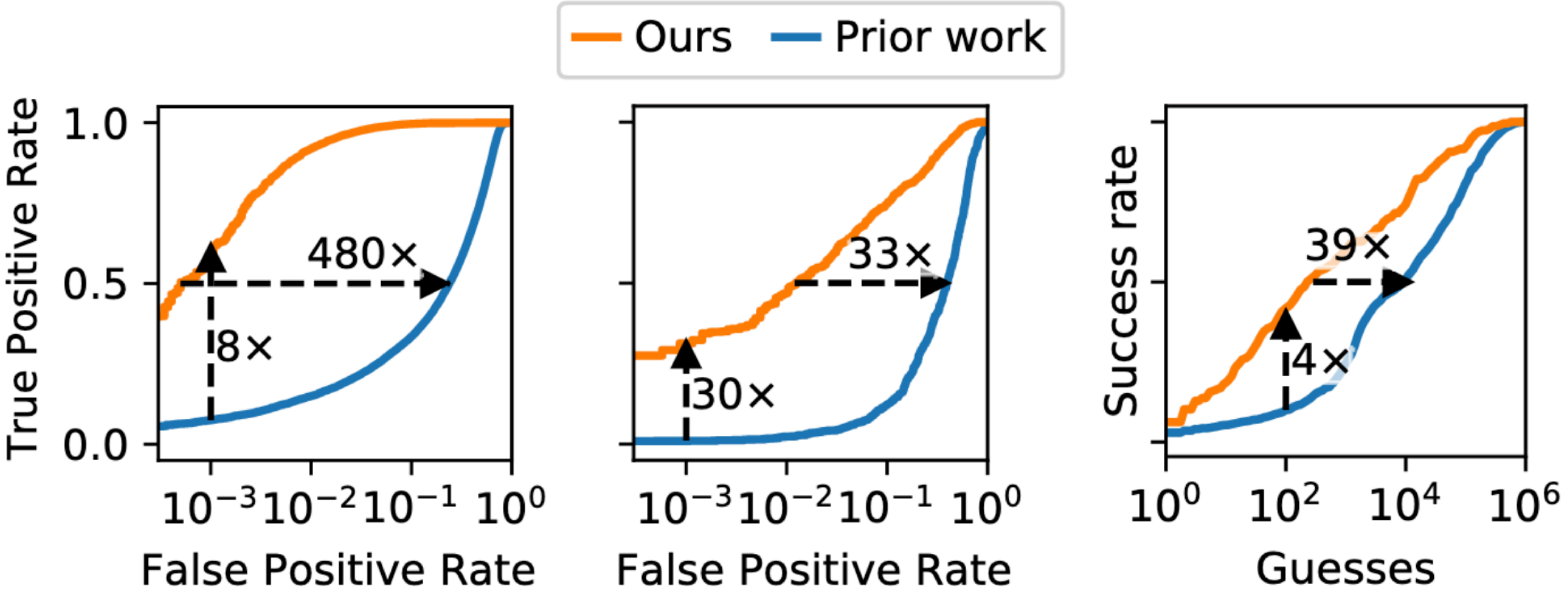
Example: Secure Multi-Party Computation



- No data is shared
- No entity can observe the intermediate steps of the computation
- The final model, however, is available to all parties
- New Attack: Adversary poisons his dataset to increase information leakage from other parties! Exploit memorization.

[Tramèr, Shokri, San Joaquin, Le, Jagielski, Hong, Carlini] Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets, CCS'22

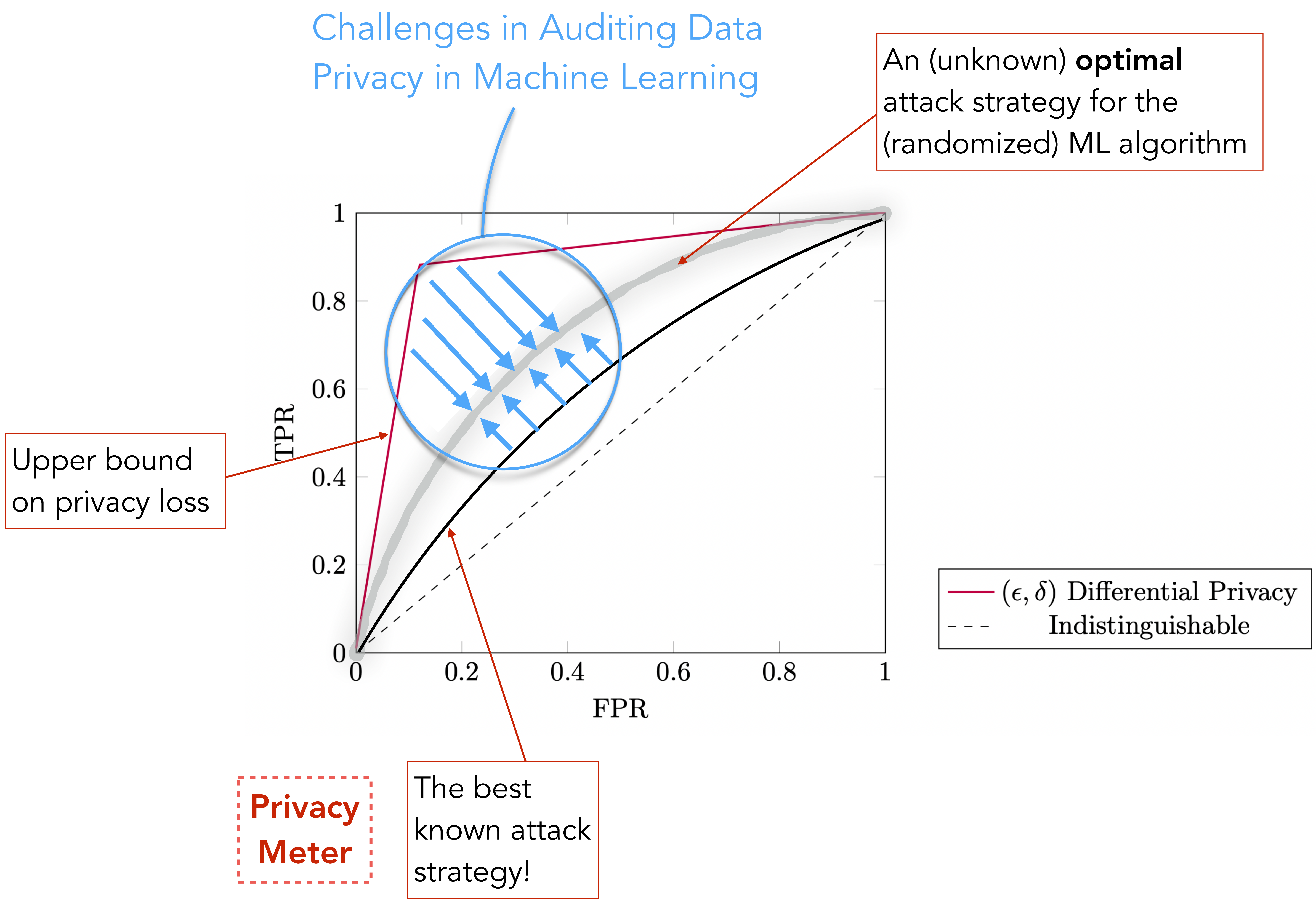
Example: Secure Multi-Party Computation



(a) Membership Inference (b) Attribute Inference (c) Canary Extraction

[Tramèr, Shokri, San Joaquin, Le, Jagielski, Hong, Carlini] Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets, CCS'22

Conclusions



Privacy Meter

Other challenges:

Alleviating the potential tension between privacy and

- Generalizability
- Robustness
- Fairness
- Explainability
- Scalability