Auditing Data Privacy for Machine Learning

Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore (NUS)

Main Takeaways

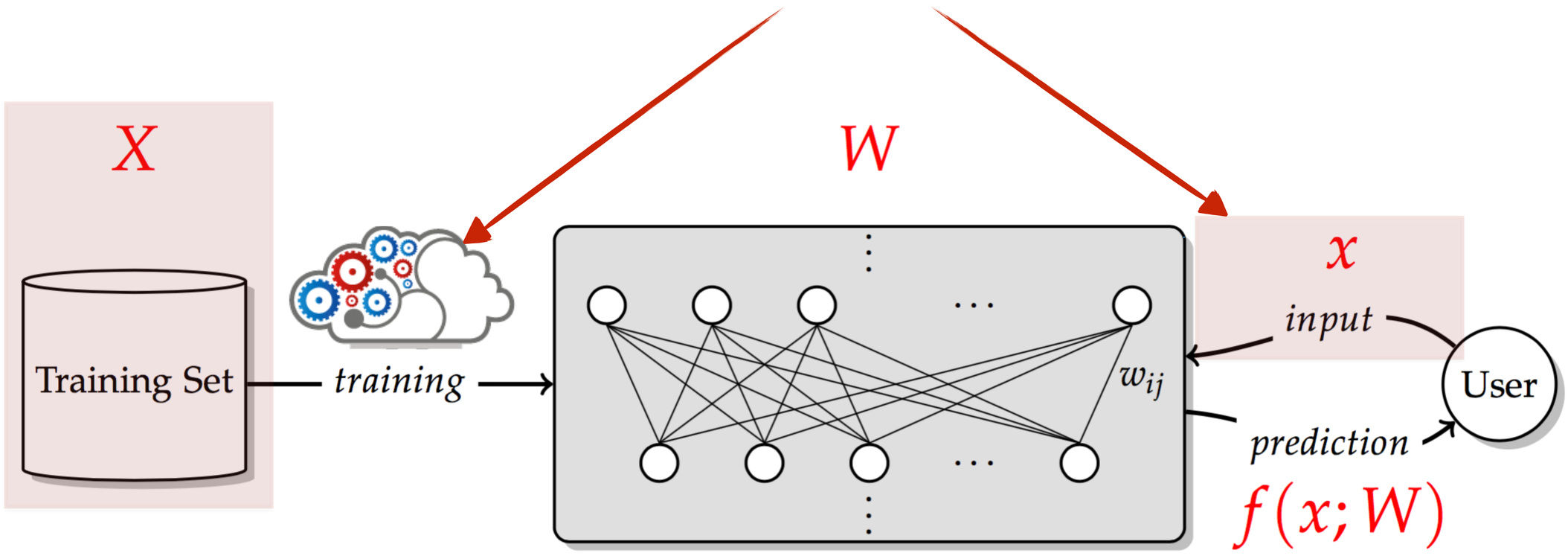
- There is a difference between confidentiality and **privacy**
- Machine learning algorithms are extremely high **risk** algorithms, and are vulnerable to inference attacks
- We need to **audit** privacy in machine learning, but currently we are not doing it!
- There is a systematic and quantitative **method**, and an open source **tool**, for auditing privacy risks in machine learning

Privacy Regulations

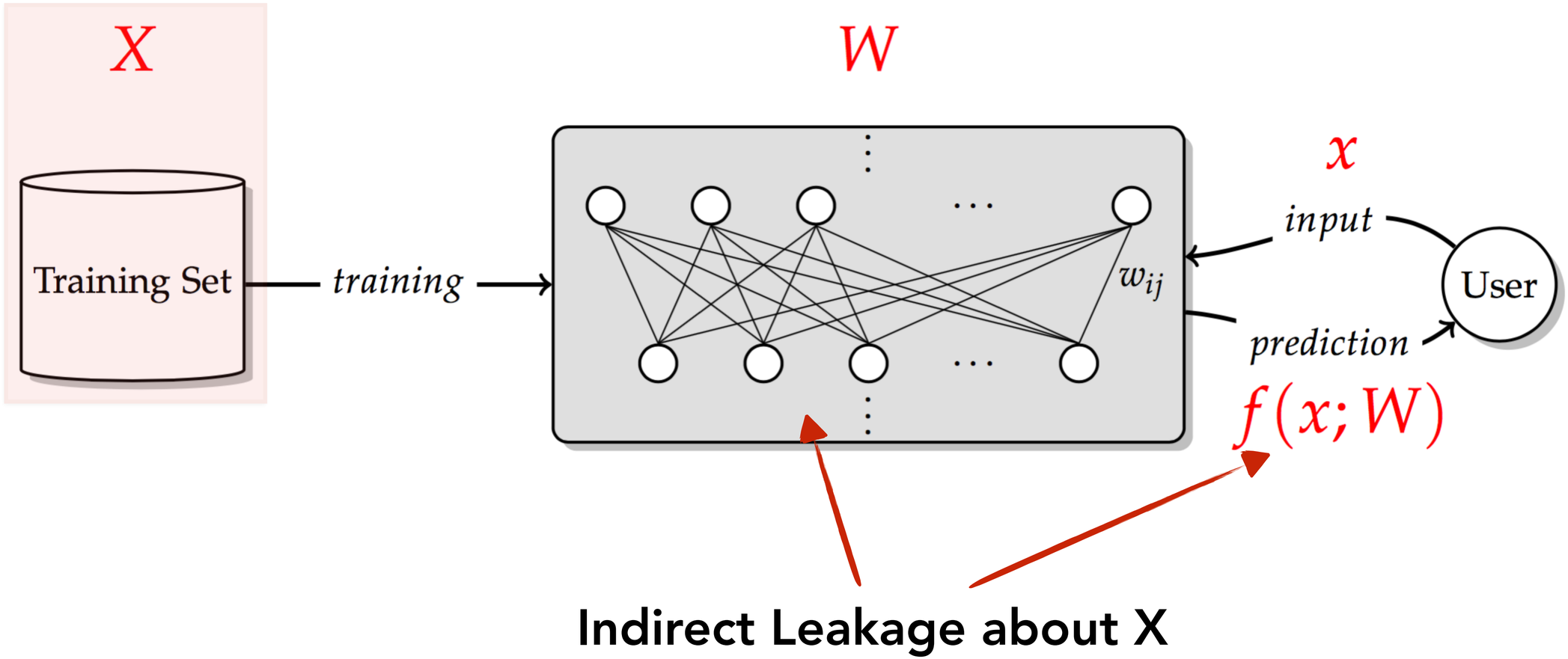
GDPR — Data Protection Impact Assessment

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

Direct Privacy Risks in Machine Learning

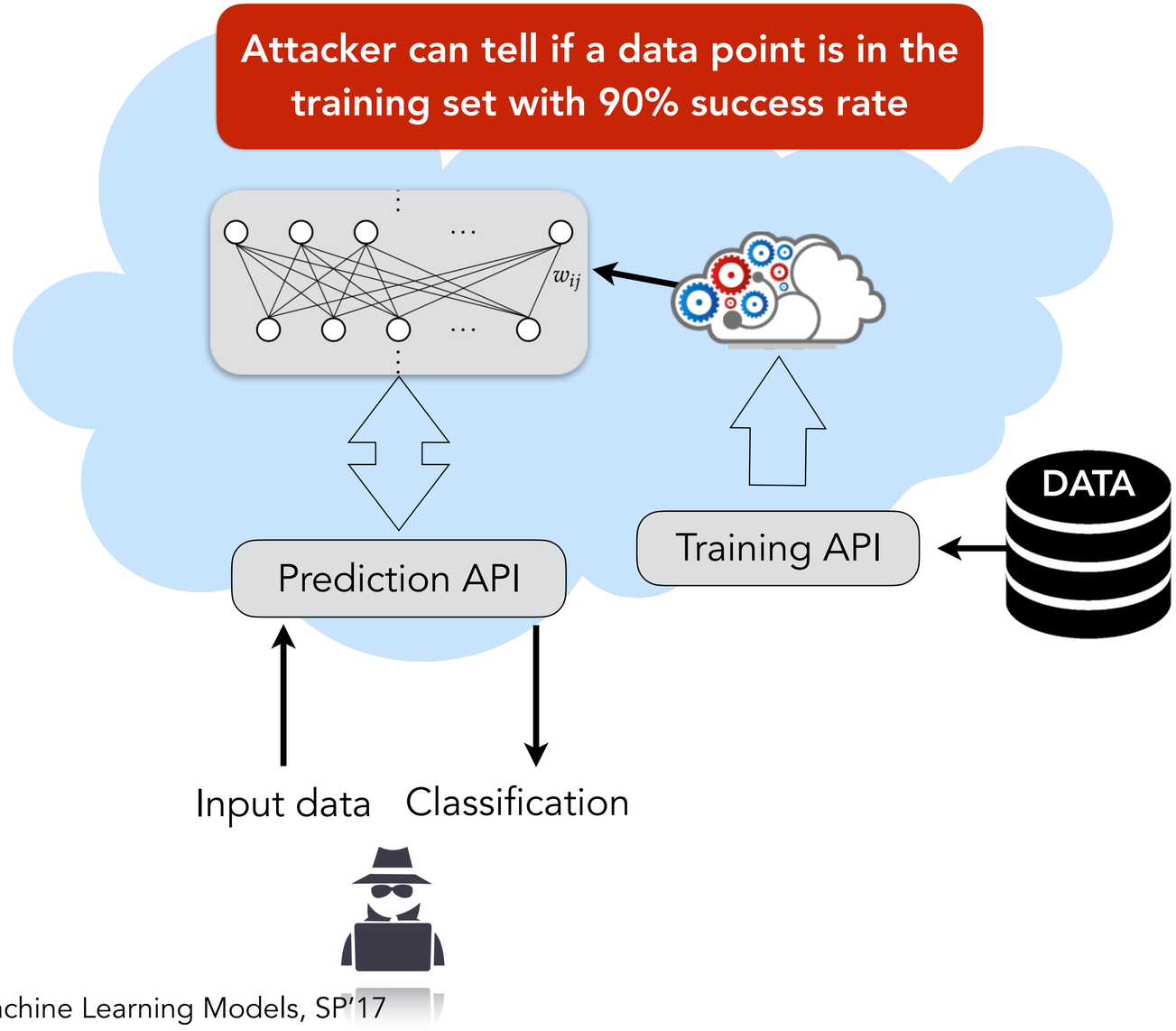


Indirect Privacy Risks in Machine Learning



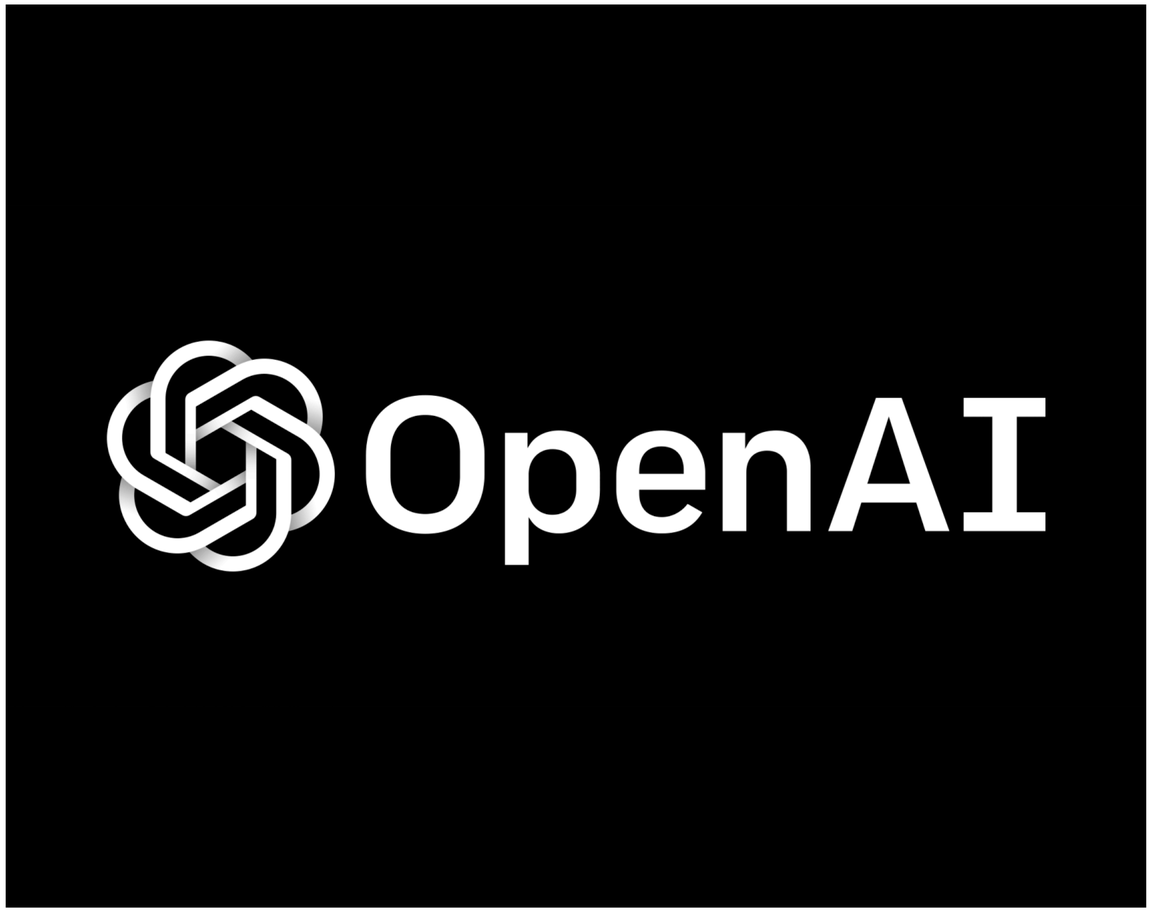
Real World Attacks against Machine Learning as a Service Platforms

Google Cloud



[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

random input

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

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

Prefix

East Stroudsburg Stroudsburg...

GPT-2

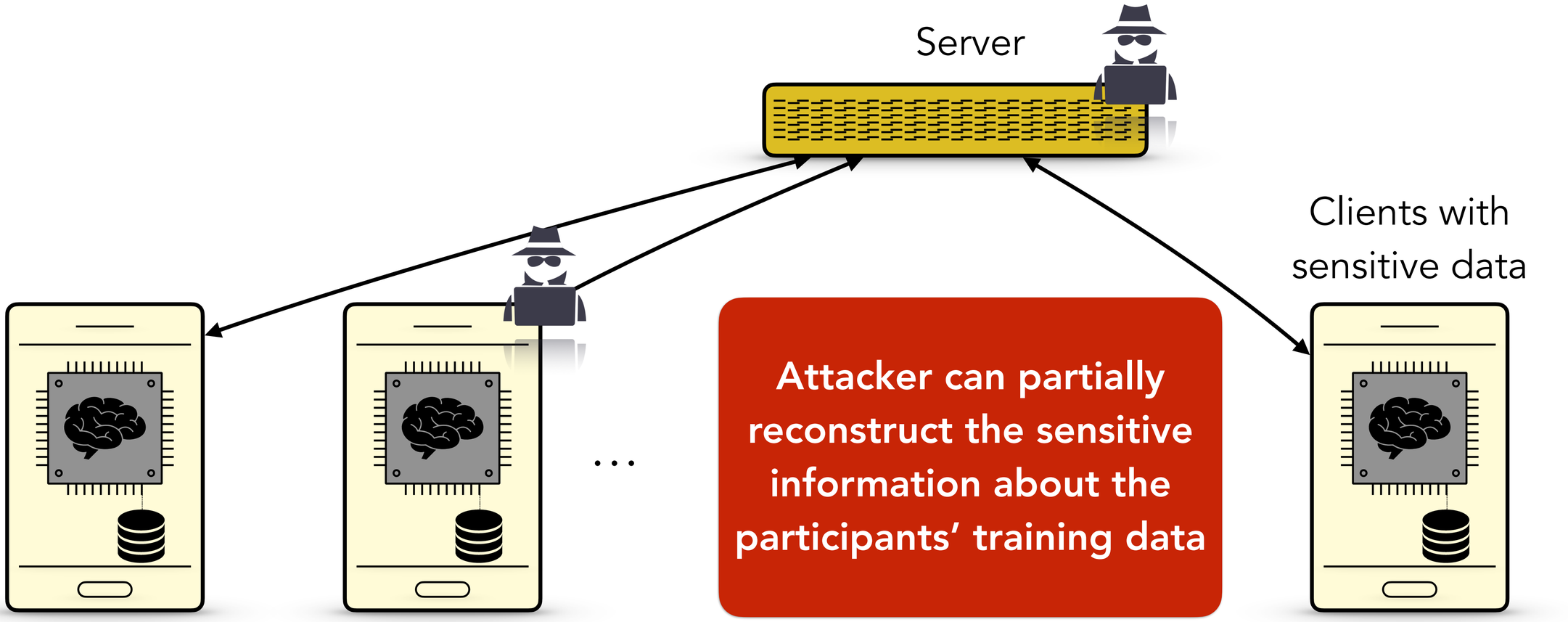
Memorized text

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

Florian Tramer

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

Real World Attacks against Federated Learning Algorithms



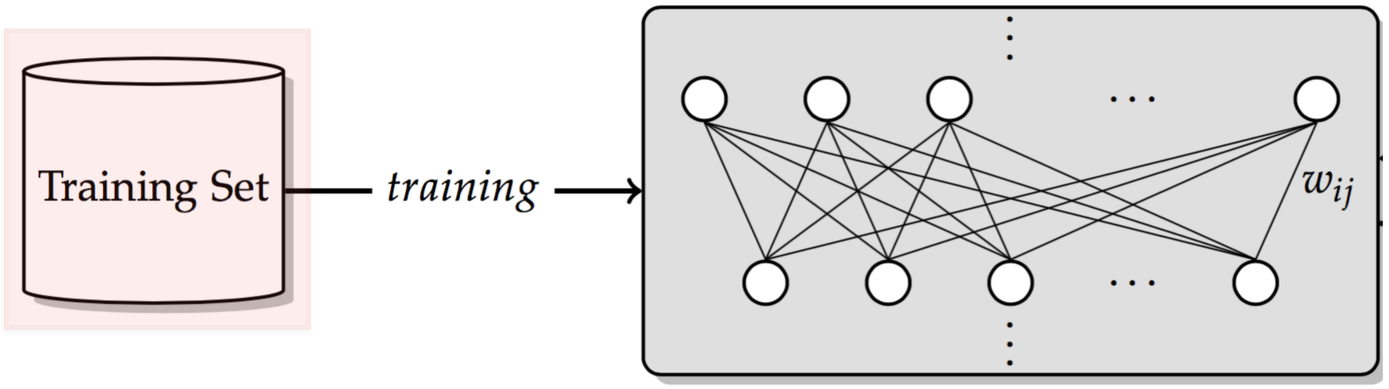
[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

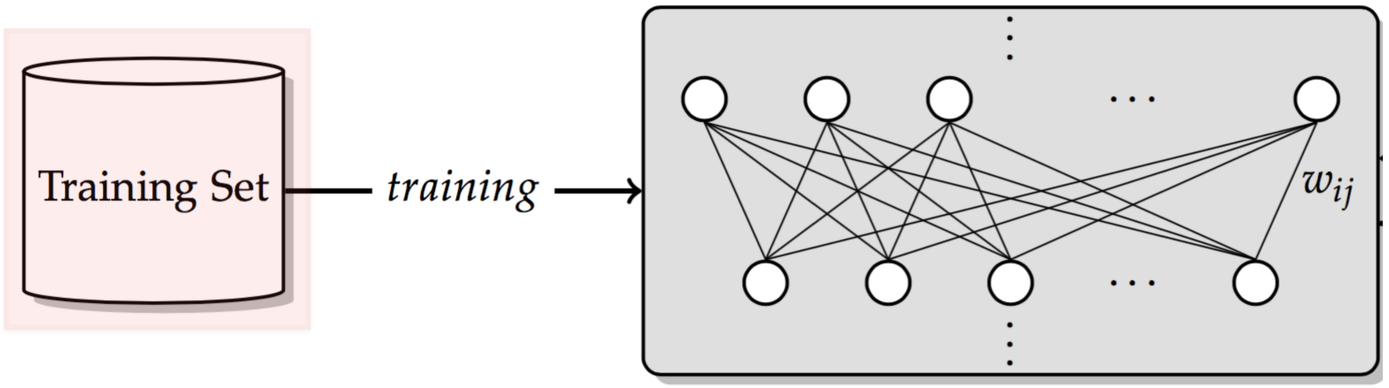
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

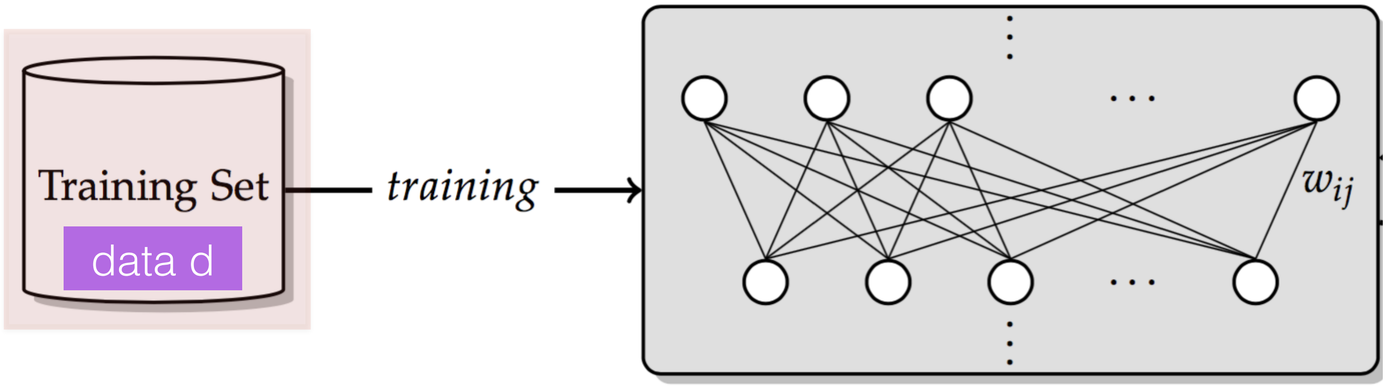
[Zhang, Tople, Ohrimenko] Leakage of Dataset Properties in Multi-Party Machine Learning, Usenix Security'21

Models are **personal data**

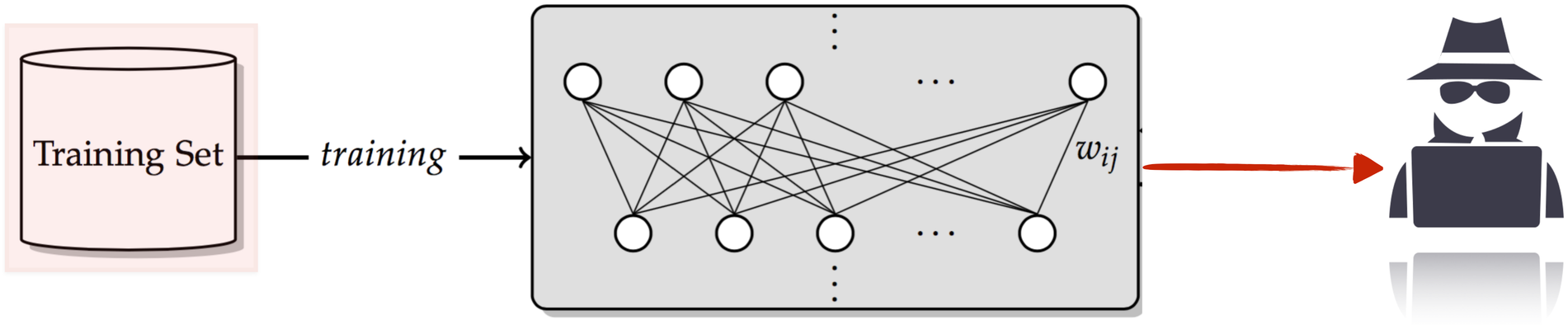
We need a standard method for auditing data privacy in machine learning systems


Training Set
training
w_{ij}

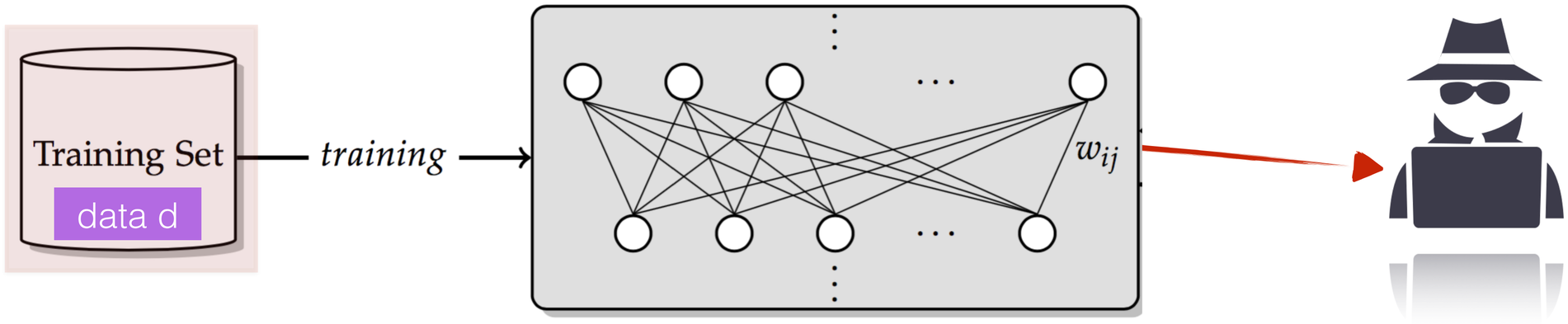

Training Set
training
w_{ij}


Training Set
data d
training
w_{ij}

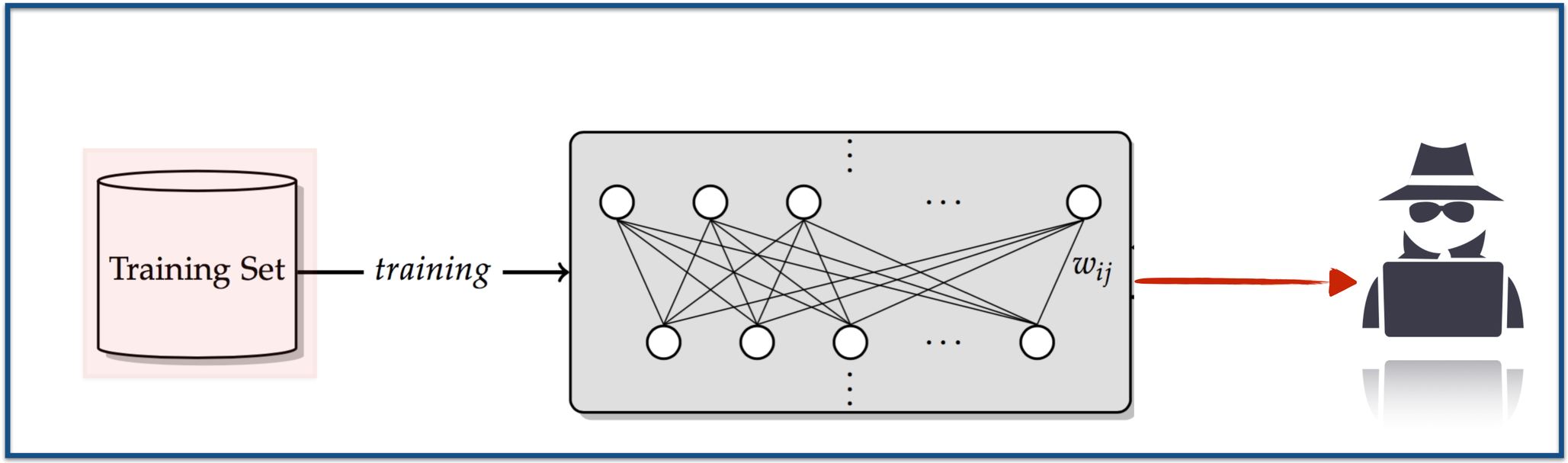
A world in which data d **is not** part of training set



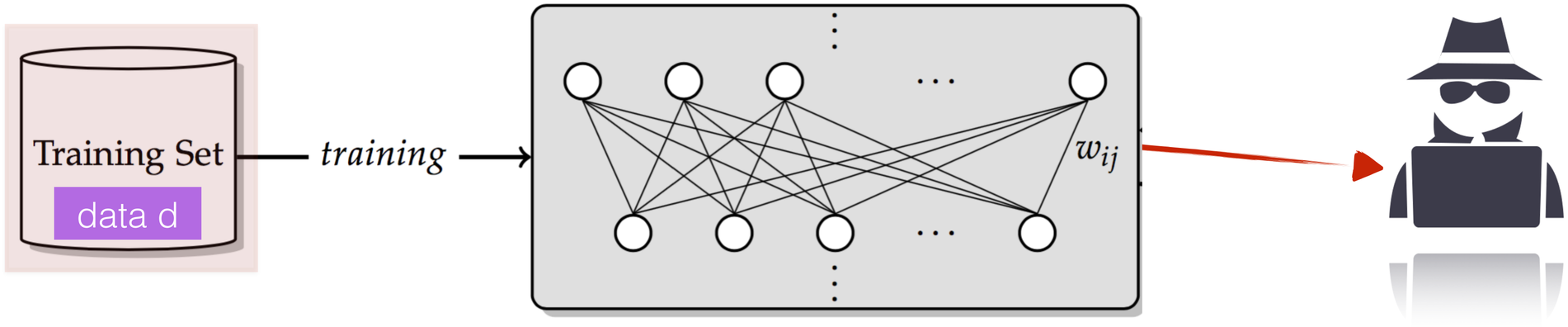
Alternative world where data d **is** part of training set



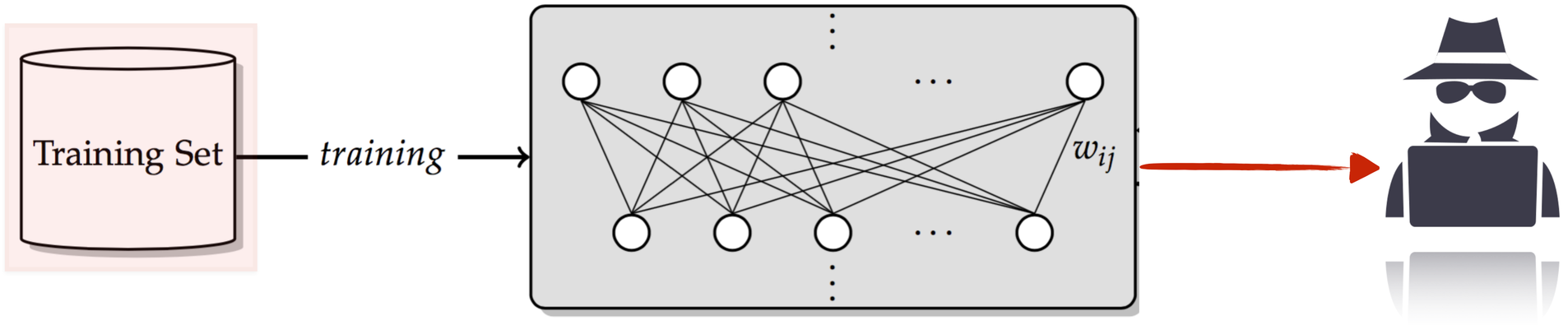
A world in which data d **is not** part of training set



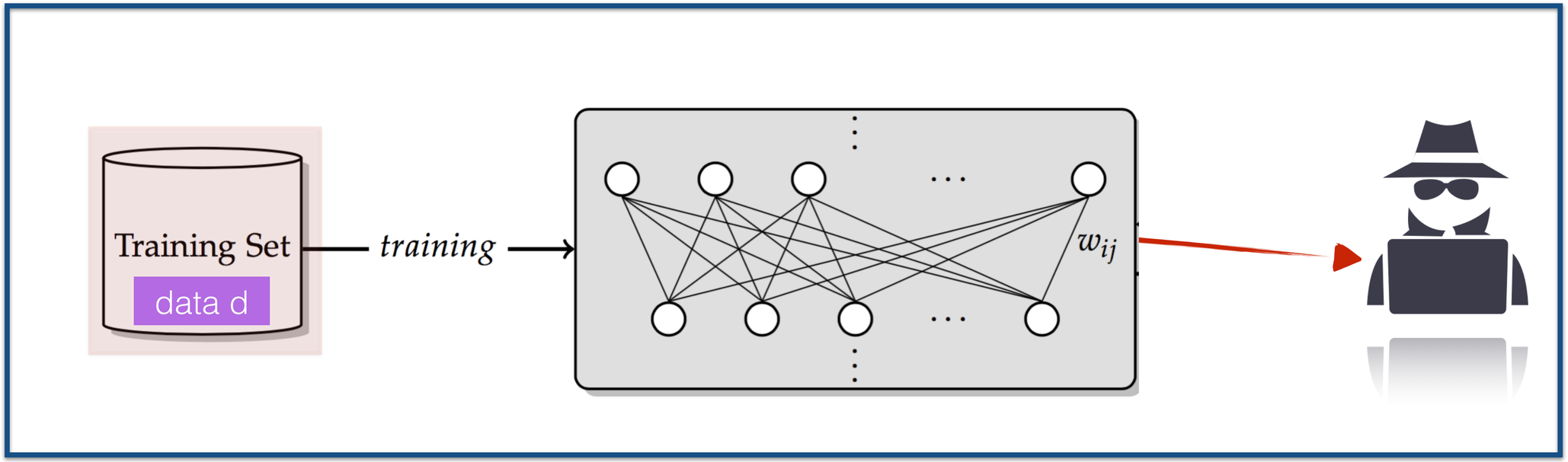
Alternative world where data d **is** part of training set



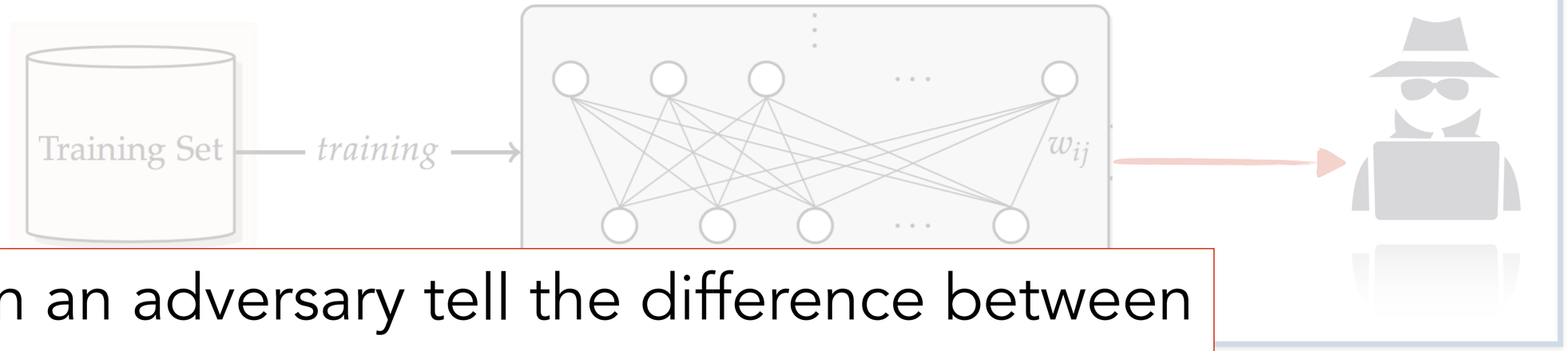
A world in which data d **is not** part of training set



Alternative world where data d **is** part of training set



A world in which data d **is not** part of training set

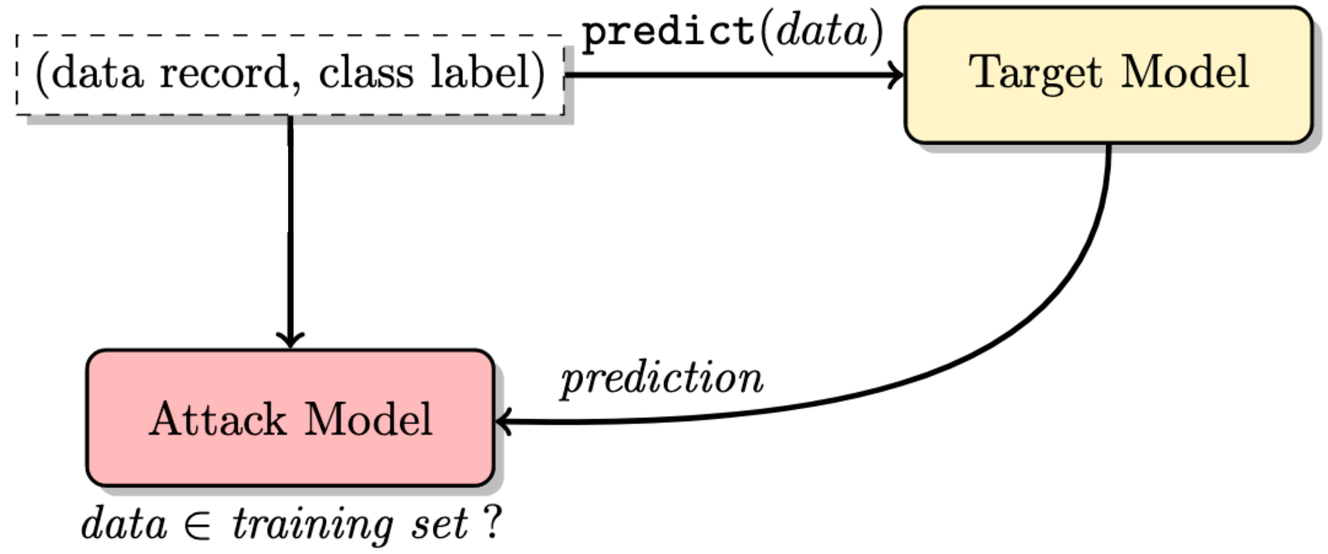


Can an adversary tell the difference between these two worlds?

Alternative world where data d **is** part of training set

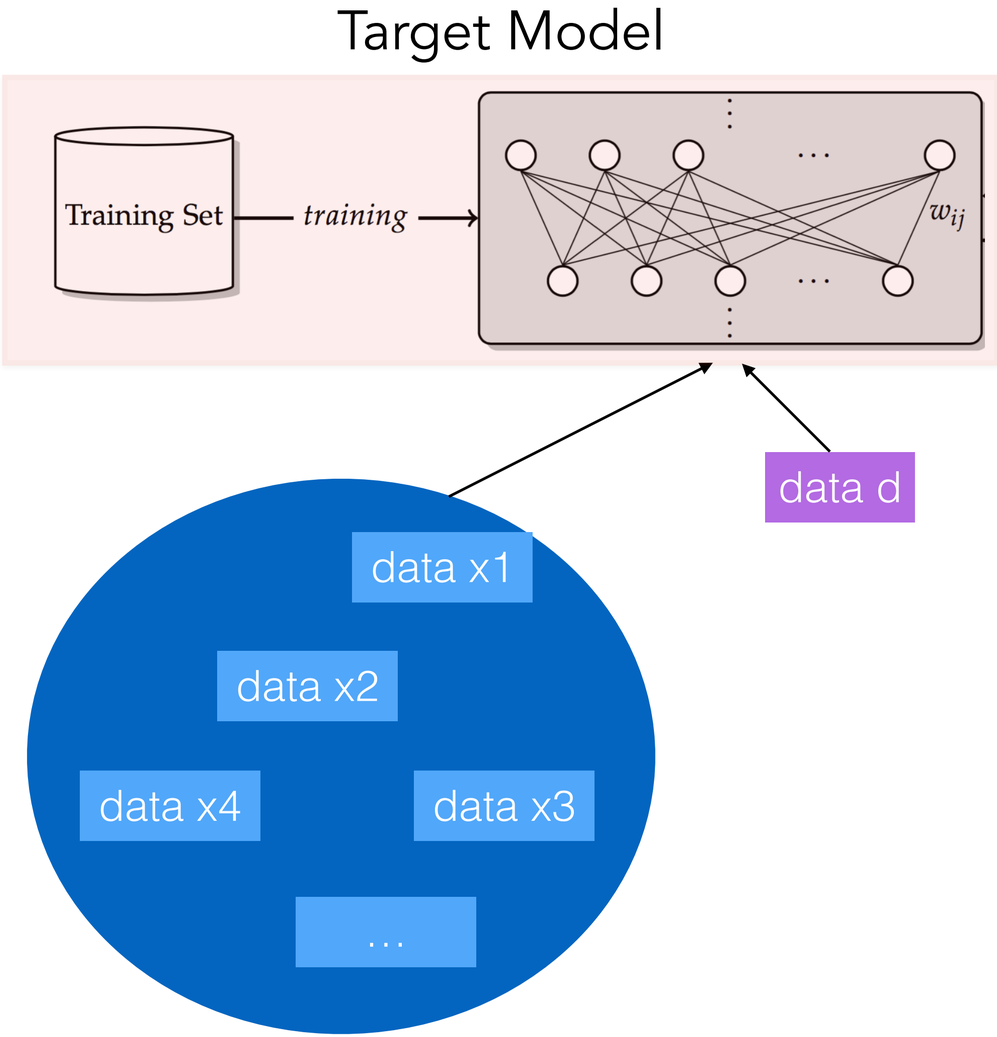
Membership Inference

- 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, SP'17

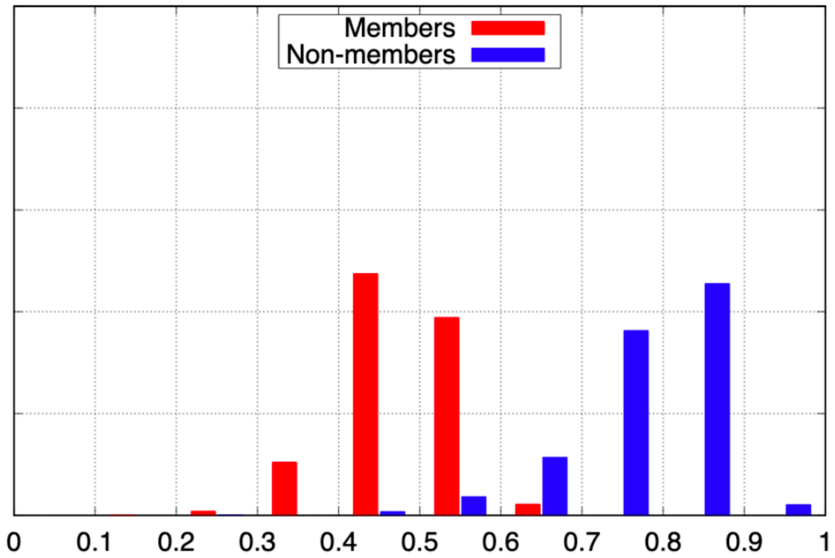
Membership Inference Attack



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

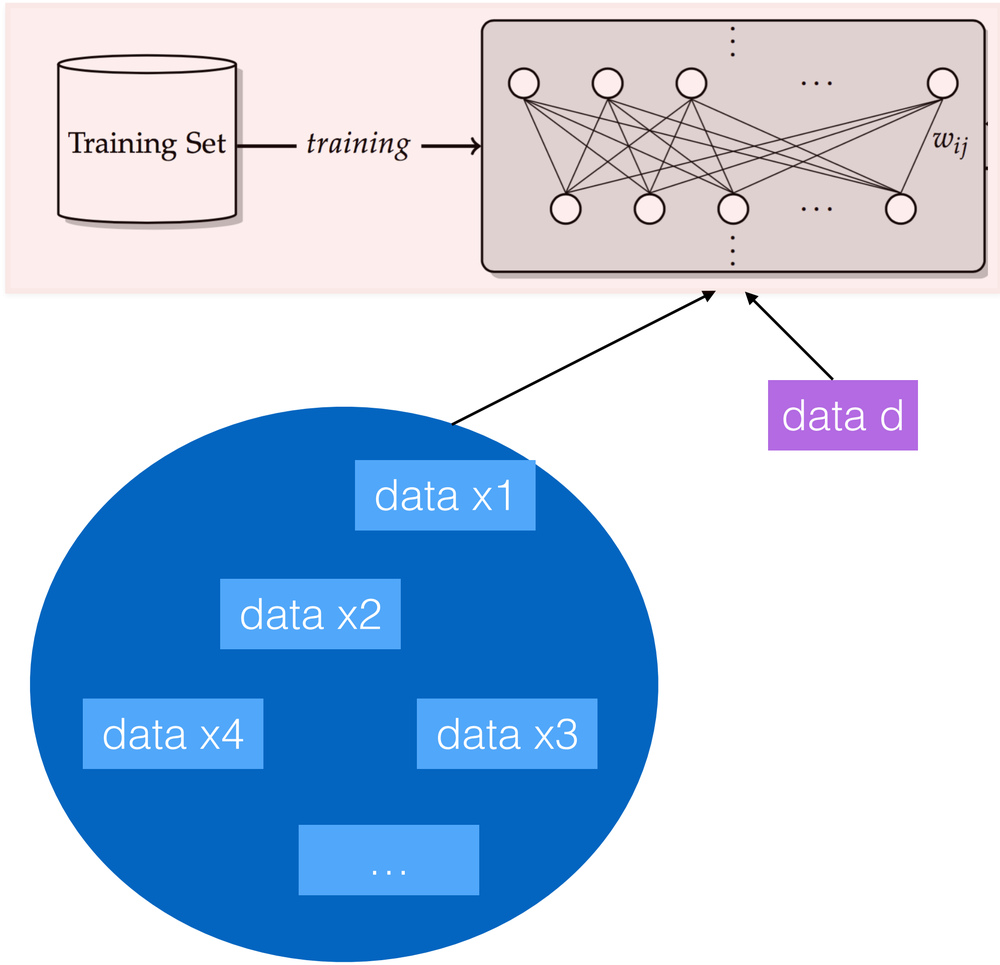
[Ye, Maddi, Murakonda, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, 21

Membership Inference Attack



Model's Error

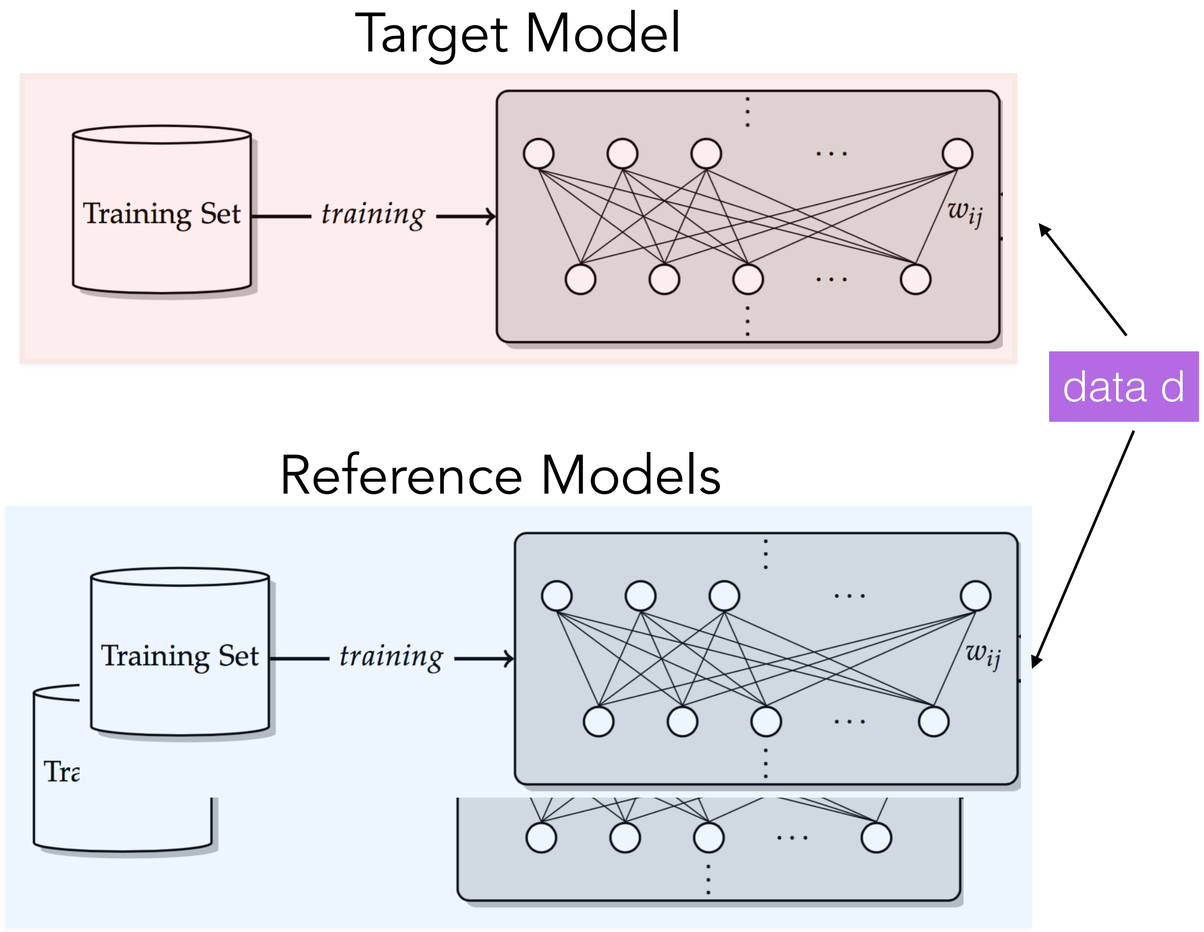
Target Model



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

[Ye, Maddi, Murakonda, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, 21

Membership Inference Attack



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

[Ye, Maddi, Murakonda, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, 21

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

AI Regulations and Guidelines

A Taxonomy and Terminology of Adversarial Machine Learning

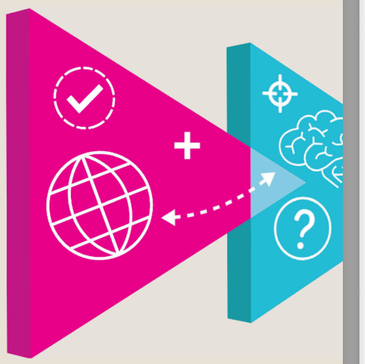
Elham Tabassi
Kevin J. Burns
Michael Hadjimichael
Andres D. Molina-Markham
Julian T. Sexton

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8269-draft

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Guidance on the AI auditing framework

Draft guidance for consultati



ico.
Information Commissioner's Office

European Commission

Home > Publications > White Paper on Artificial Intelligence: a European approach to excellence and trust

WHITE PAPER

White Paper on Artificial Intelligence: a European approach to excellence and trust

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

November 17, 2020

THE DIRECTOR

M-21-06

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AG

FROM: Russell T. Vought
Director

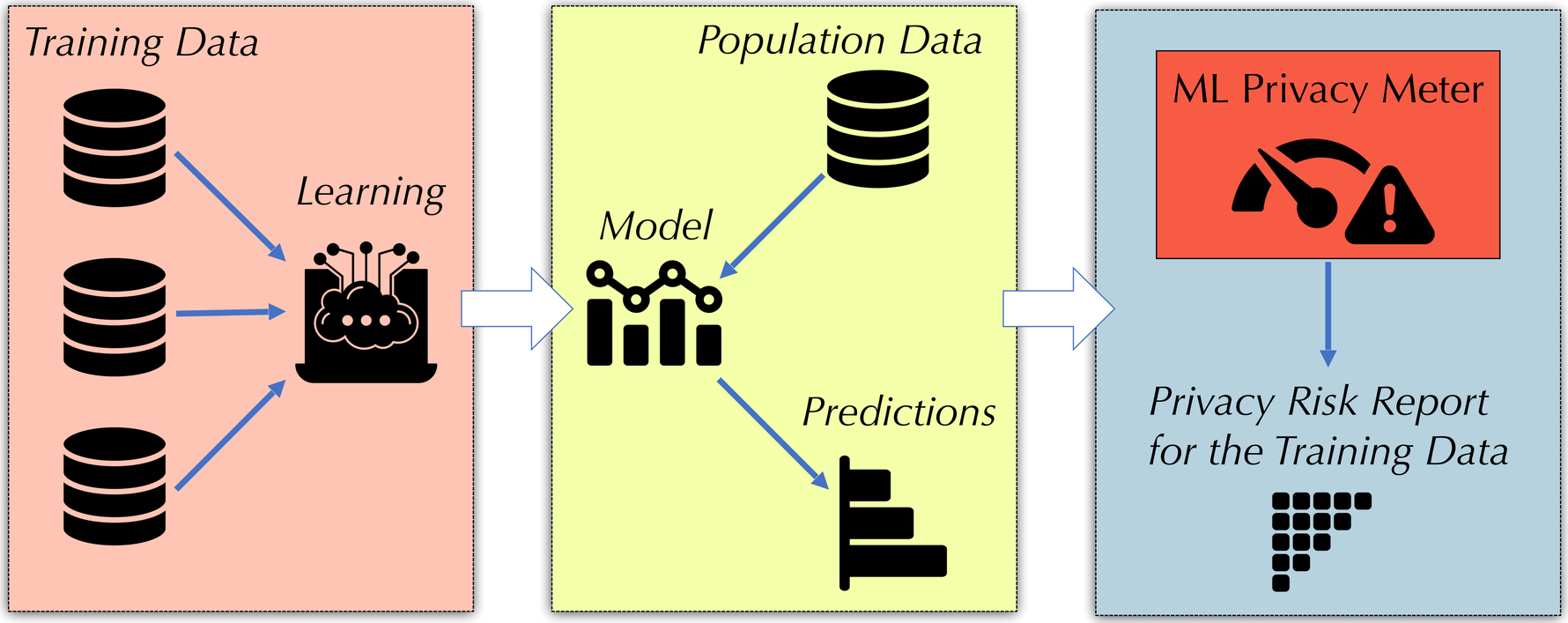
SUBJECT: Guidance for Regulation of Artificial Intelligence Applications

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"**

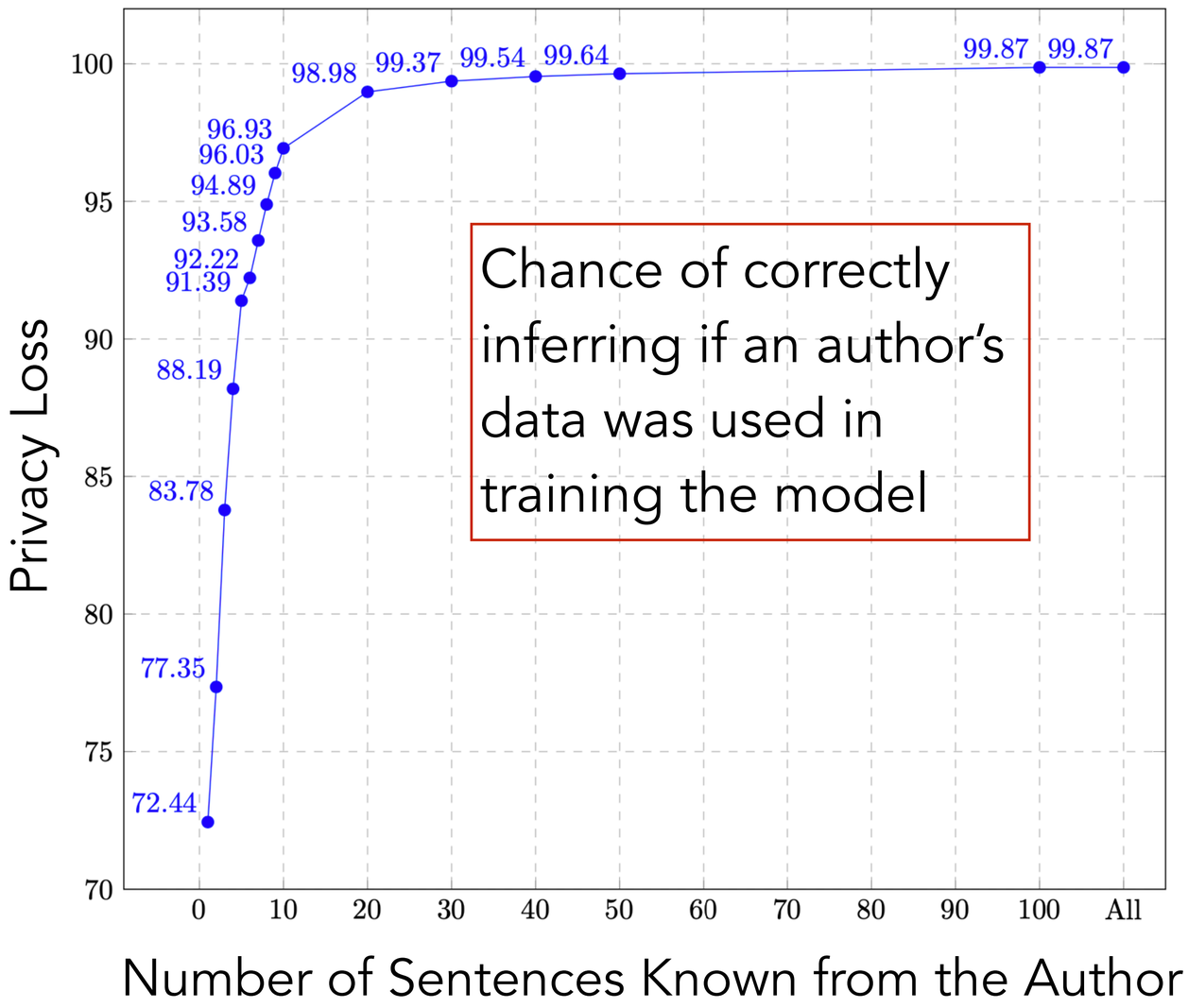
ML Privacy Meter

privacy-meter.com

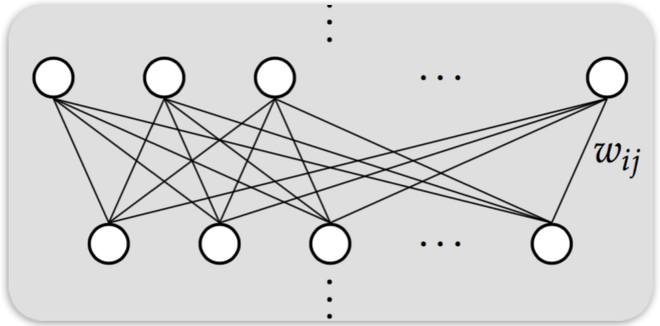


ML Privacy Meter is an open source tool that enables quantifying the privacy risks of machine learning models.

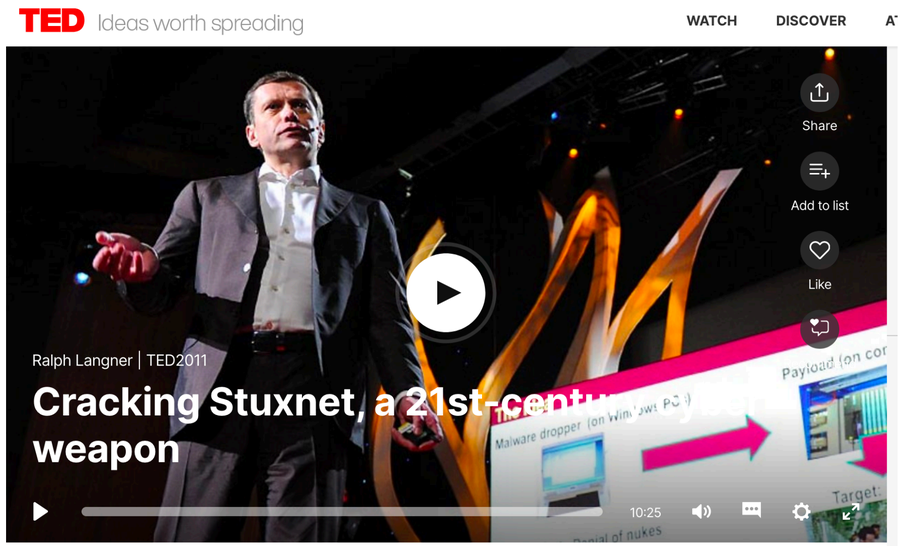
Example: Language Generative Model



(Speaker Annotated TED talks)
SATED dataset



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. Or
were opened to the public, and historians such as me
about how the GDR surveillance state functioned.

Example: Image Classification Tasks

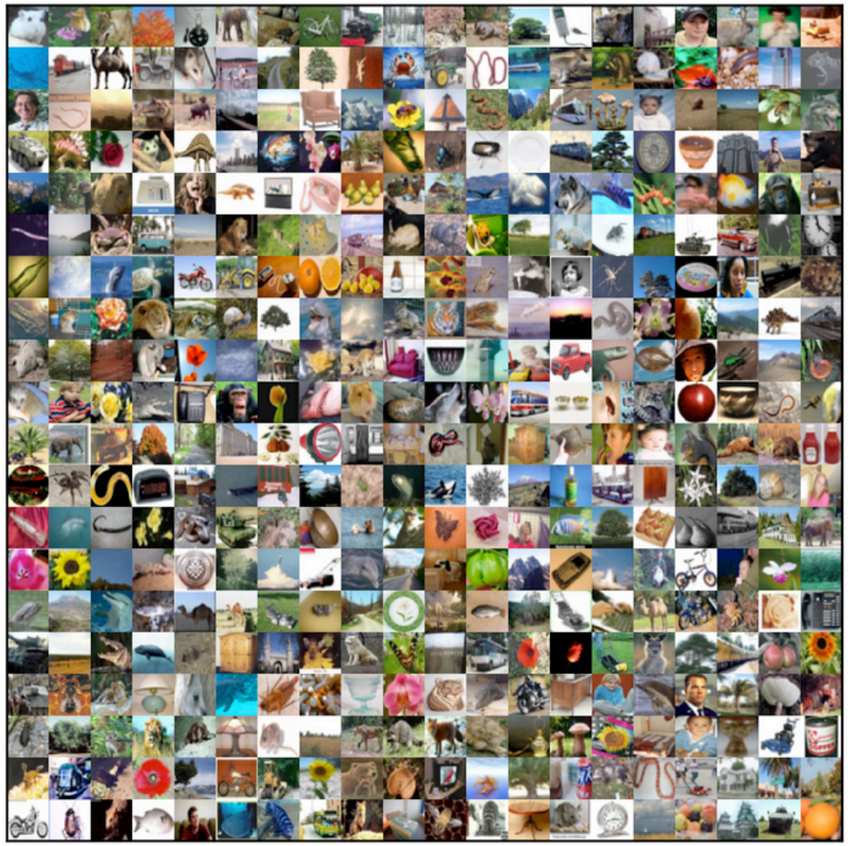
Model	Number of Parameters	Prediction (Test) Accuracy	Privacy Risk
AlexNet	2.47 million	44%	75.1%
ResNet	1.7 million	73%	64.3%
DenseNet	25.62 million	82%	74.3%

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

Membership Inference Attacks

Enhanced Membership Inference Attacks
against Machine Learning Models

Jiayuan Ye[1], Aadyaa Maddi[1], Sasi Kumar Murakonda[2], and Reza Shokri[1]

Hypothesis Testing for Membership Inference

- Given a data point "z" and black-box access to model "θ",
  - Tell if "z" was member of the training set of "θ"
- Null hypothesis: non-member
- Alternative hypothesis: member

$$H_0 : D \xleftarrow{n\ i.i.d. samples \sim \pi(z)} D_{pop}, \theta \sim \mathcal{T}(D), z \xleftarrow{sample \sim \pi(z)} D_{pop}$$

$$H_1 : D \xleftarrow{n\ i.i.d. samples \sim \pi(z)} D_{pop}, \theta \sim \mathcal{T}(D), z \xleftarrow{sample} D$$

Likelihood Ratio Test

$$LR(\theta, z) = \frac{L(H_0|\theta, z)}{L(H_1|\theta, z)} \approx e^{\frac{1}{T}\ell(\theta, x_z, y_z)}$$

Likelihood Ratio Test

$$LR(\theta, z) = \frac{L(H_0|\theta, z)}{L(H_1|\theta, z)} \approx e^{\frac{1}{T}\ell(\theta, x_z, y_z)}$$

Reject the null hypothesis: $\{(\theta, z) : LR(\theta, z) \leq c\}$

Likelihood Ratio Test

$$LR(\theta, z) = \frac{L(H_0|\theta, z)}{L(H_1|\theta, z)} \approx e^{\frac{1}{T}\ell(\theta, x_z, y_z)}$$

Reject the null hypothesis: $\{(\theta, z) : LR(\theta, z) \leq c\}$

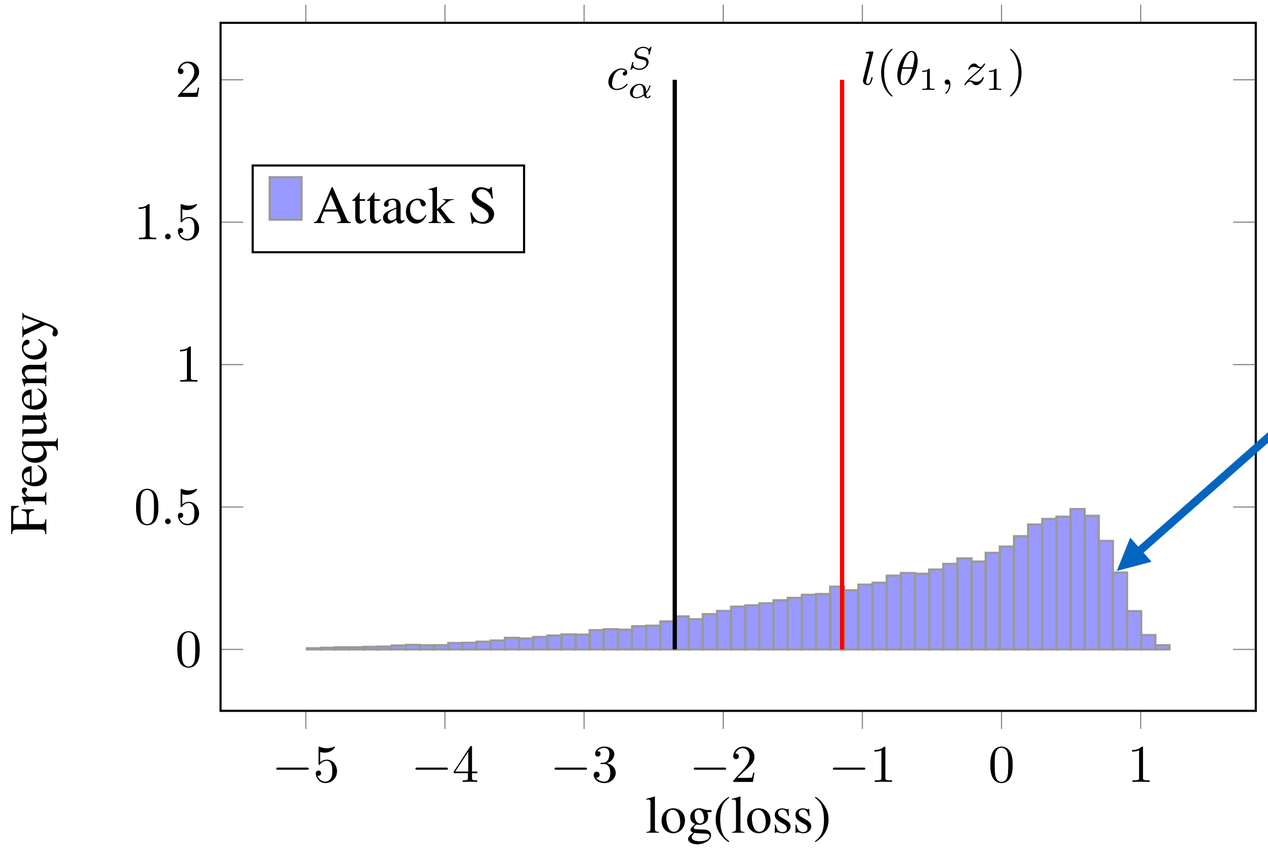
Attack: $\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta, x_z, y_z), \textbf{ reject } H_0$

false positive rate

Membership Inference via Shadow Models

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

- A large body of the literature is based on this technique
- Learn a threshold from the behaviour of some shadow models with their test data



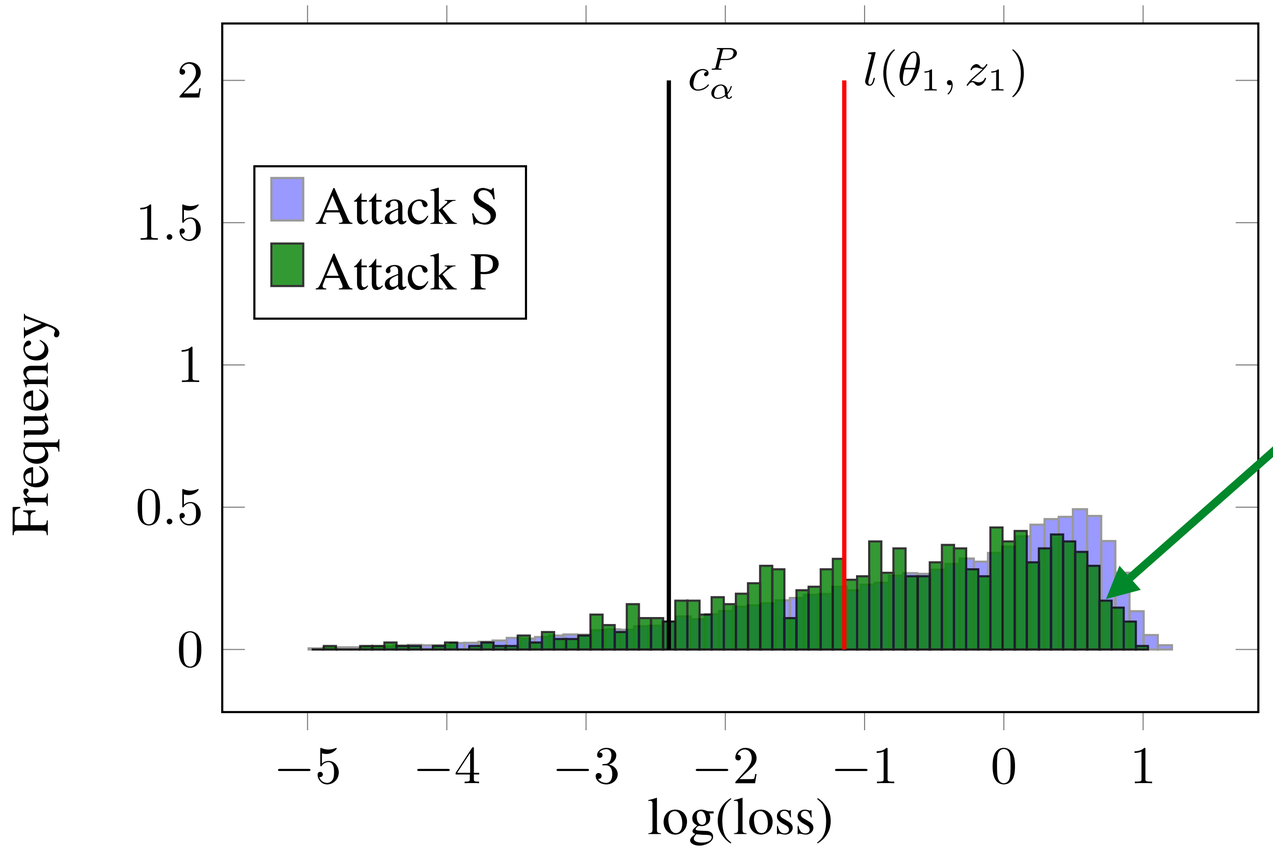
Loss of population data on shadow models

Can we do it more efficiently?

Membership Inference via Population Data

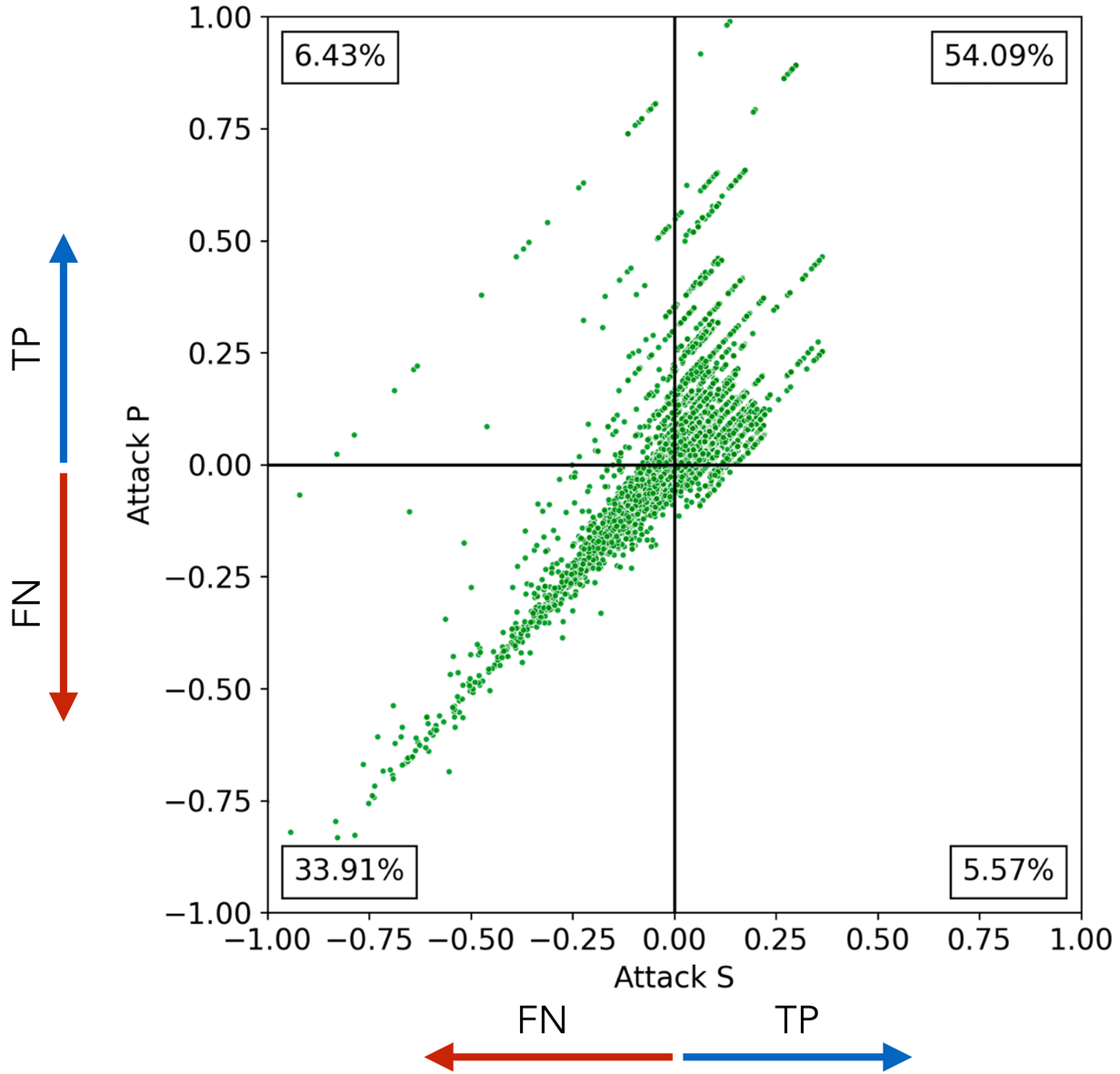
$$\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta), \textbf{ reject } H_0$$

- Learn a threshold from the behaviour of the target model on population data



Loss of population data on target model

Agreement between Attacks

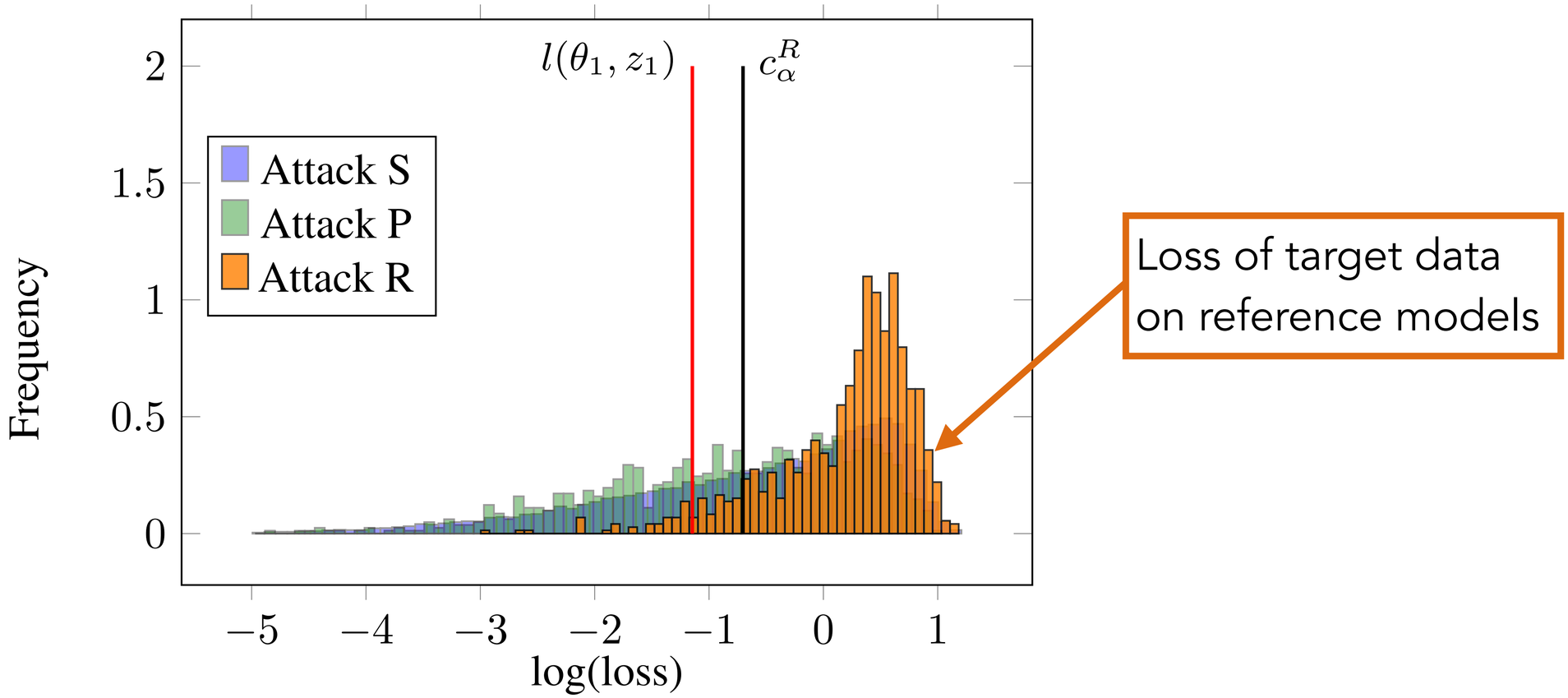


Can we do it more accurately?

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 behaviour of target data on reference models



Can we do it even more accurately?

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

Can we do it even more accurately?

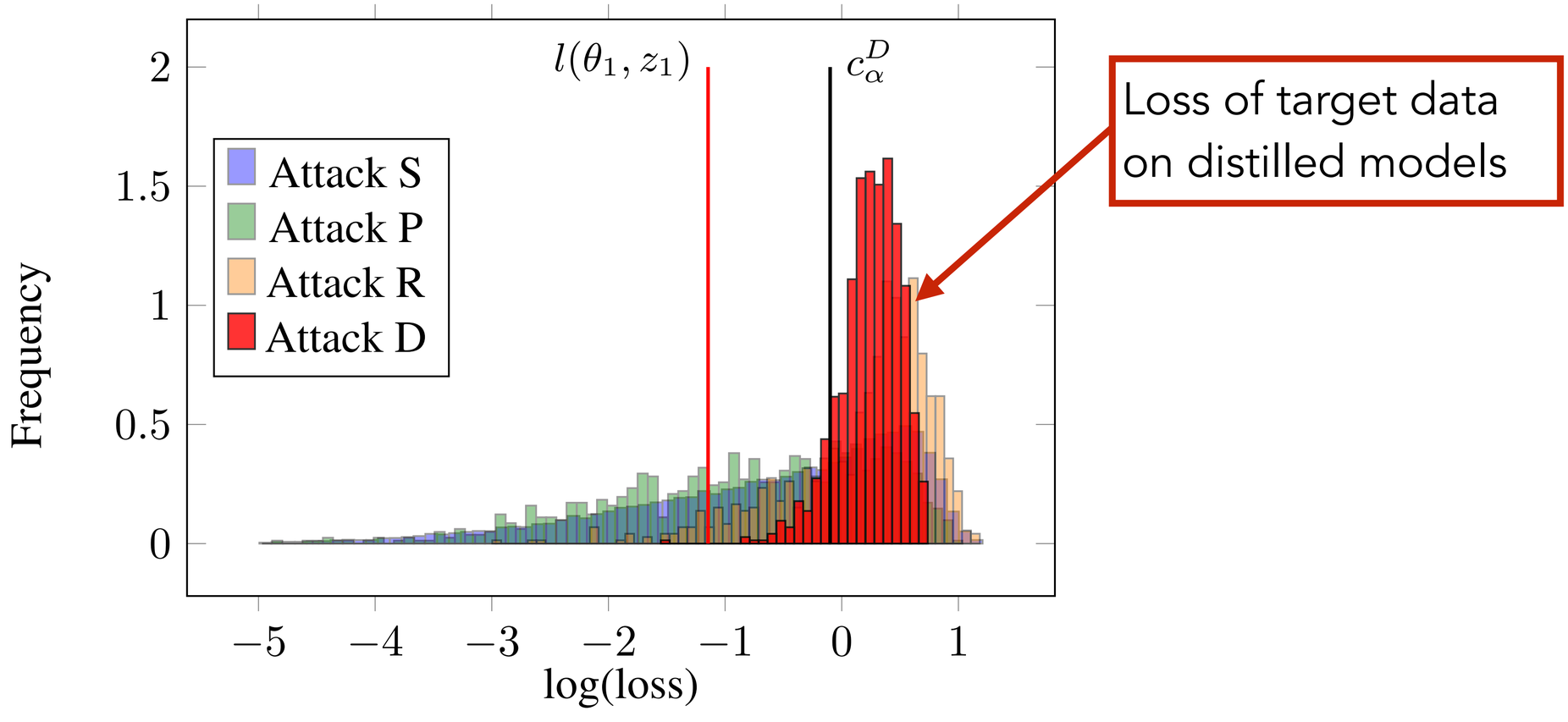
The objective is to get as close as possible to the leave-one-out attack, where the adversary knows all “other” data in the training set

- Train reference models that have a large agreement with the target model on all data, except the target data
- Idea: Model distillation — Reference models are distilled versions of the target models

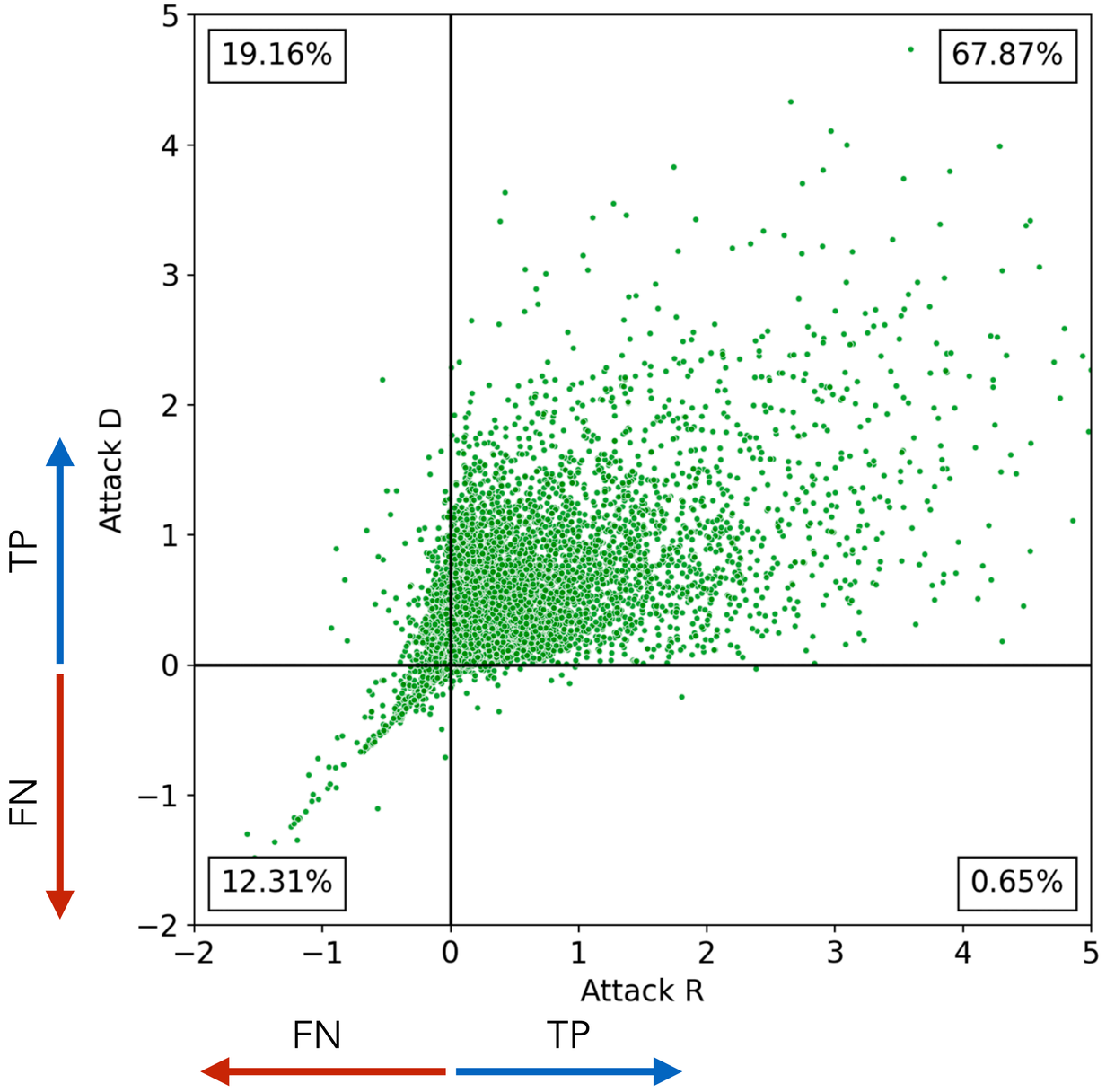
Membership Inference via Distilled Models

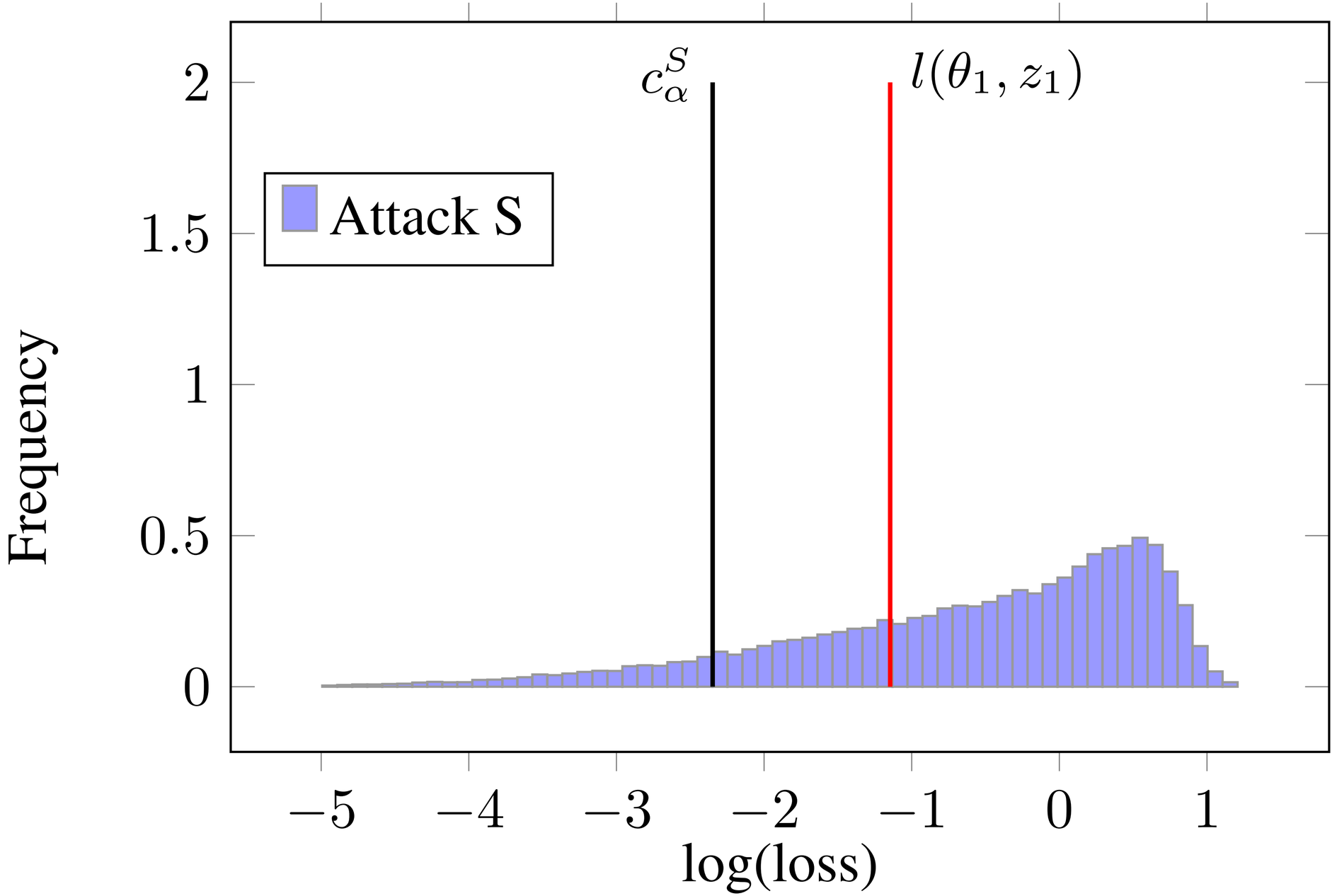
$$\textbf{If } \ell(\theta, x_z, y_z) \leq c_\alpha(\theta, x_z, y_z), \textbf{ reject } H_0$$

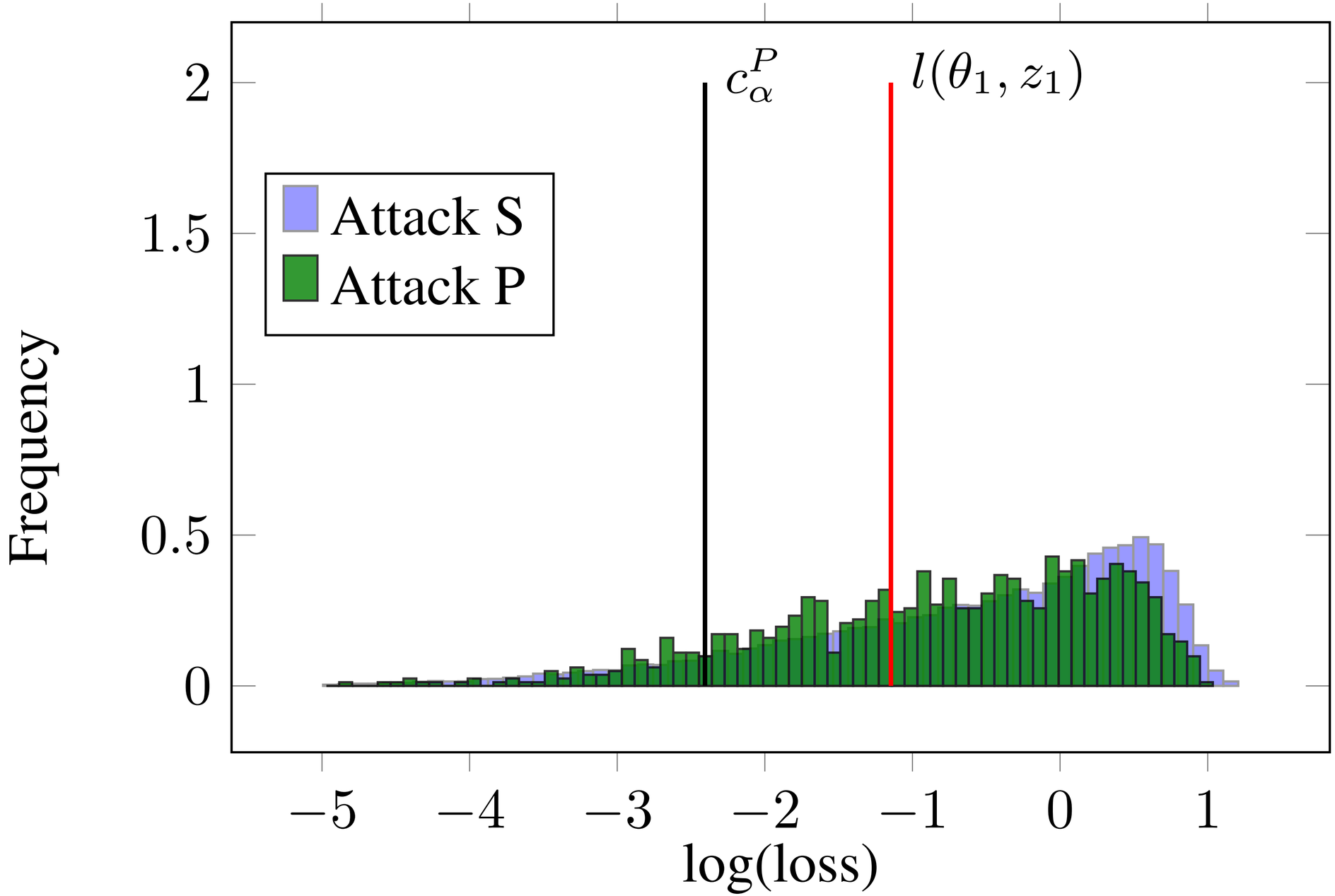
- Learn a threshold from the behaviour of target data on distilled models
- Note that the threshold depends on both target data and the target model

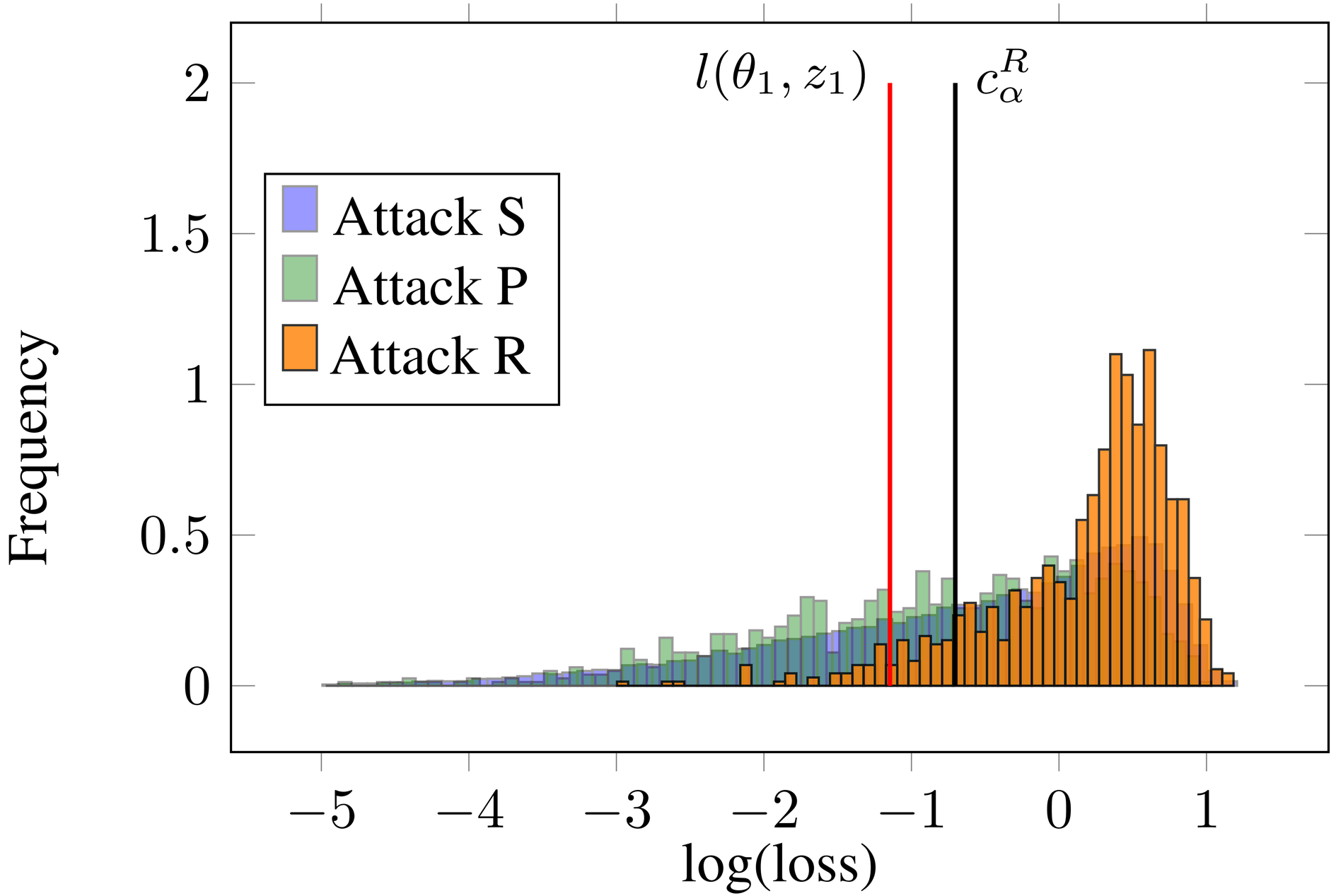


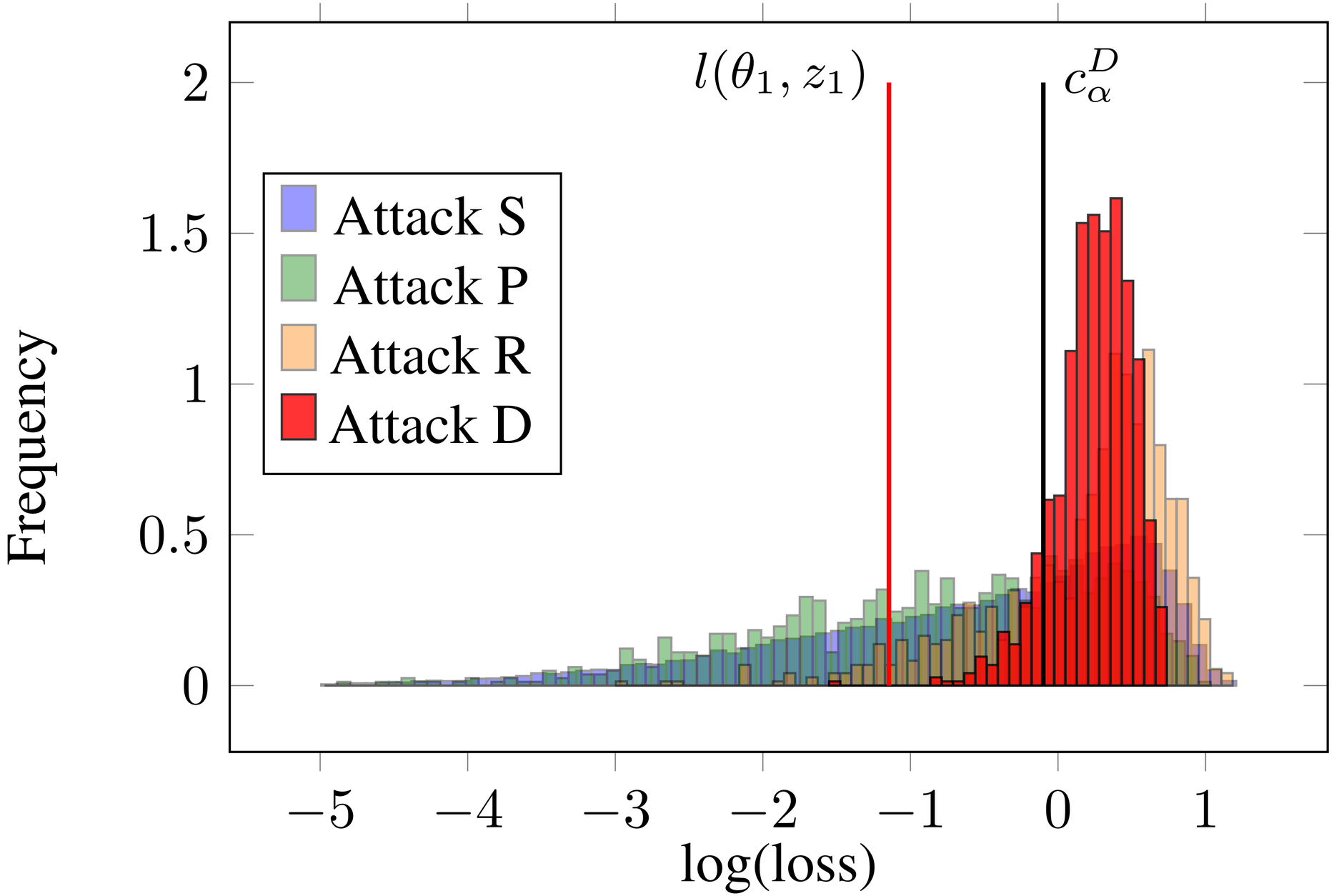
Agreement between Attacks

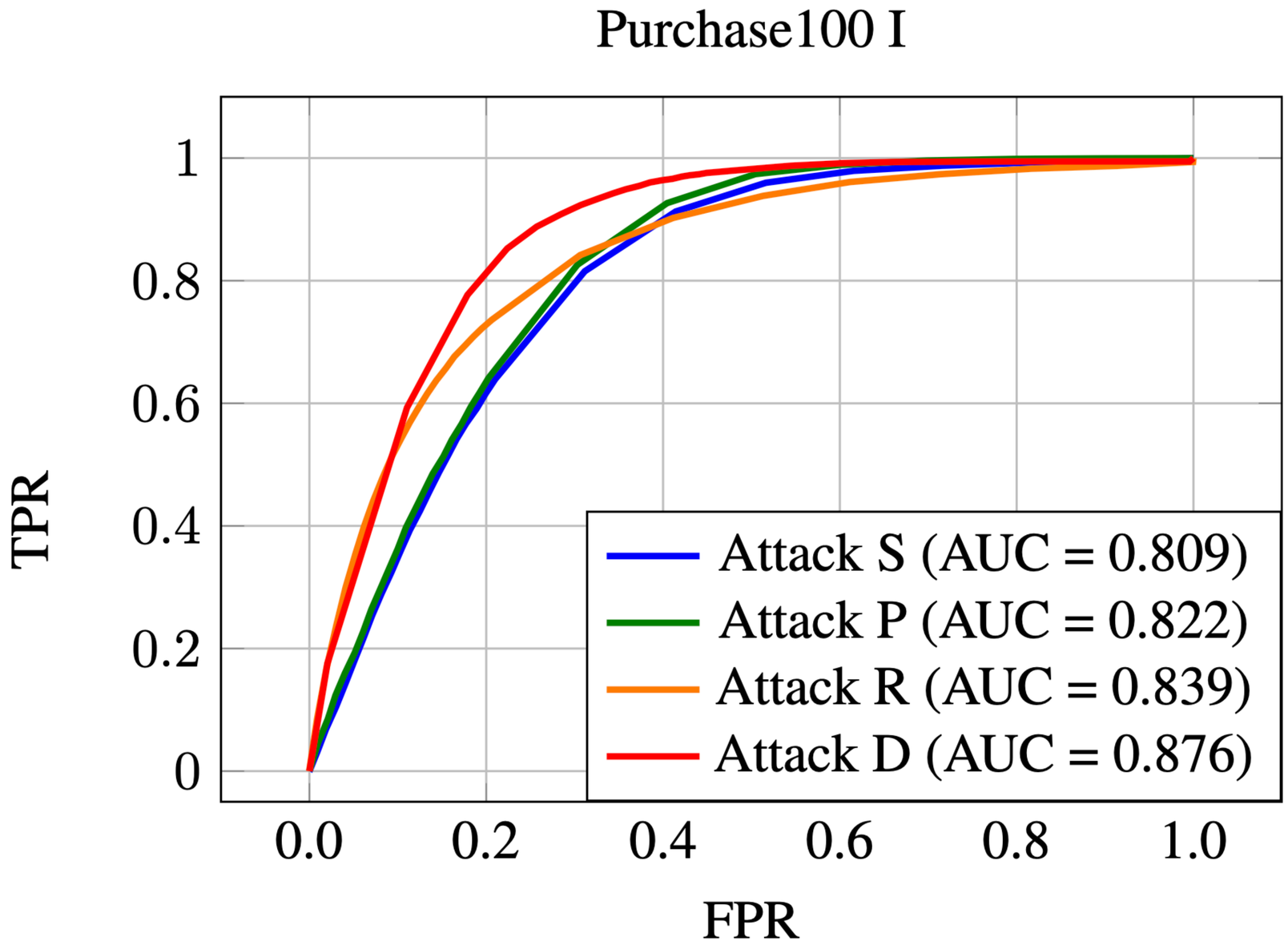



c_α^S
$l(\theta_1, z_1)$
Attack S
Frequency
2
1.5
1
0.5
0
−5
−4
−3
−2
−1
0
1
log(loss)


c_α^P
$l(\theta_1, z_1)$
Attack S
Attack P
Frequency
2
1.5
1
0.5
0
−5
−4
−3
−2
−1
0
1
log(loss)


$l(\theta_1, z_1)$
c_α^R
Attack S
Attack P
Attack R
Frequency
2
1.5
1
0.5
0
−5
−4
−3
−2
−1
0
1
log(loss)


$l(\theta_1, z_1)$
c_α^D
Attack S
Attack P
Attack R
Attack D
Frequency
2
1.5
1
0.5
0
−5
−4
−3
−2
−1
0
1
log(loss)


Purchase100 I
TPR
FPR
0
0.2
0.4
0.6
0.8
1
0.0
0.2
0.4
0.6
0.8
1.0
Attack S (AUC = 0.809)
Attack P (AUC = 0.822)
Attack R (AUC = 0.839)
Attack D (AUC = 0.876)

Auditing Data Privacy for Machine Learning

- 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
- ML Privacy Meter (privacy-meter.com) 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, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, 21