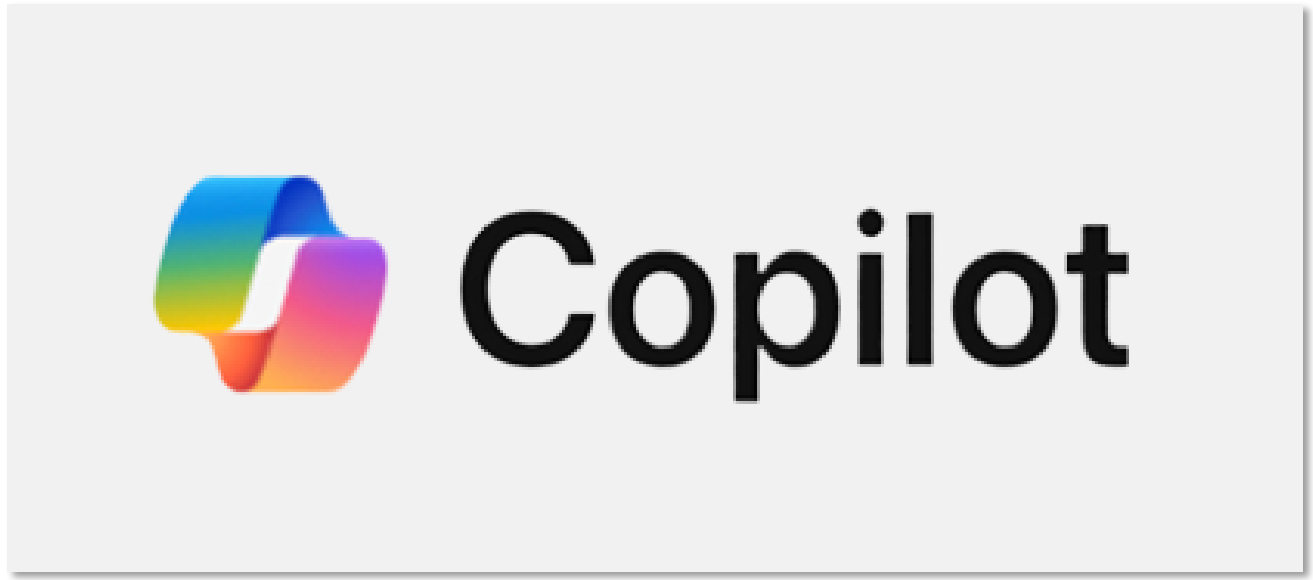
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm?

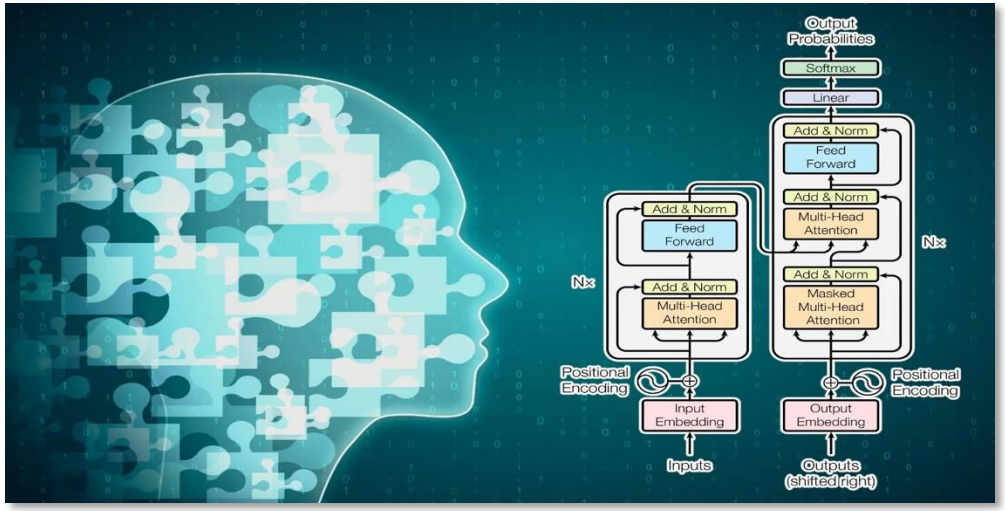
Rui Wen, Michael Backes, Yang Zhang

CISPA

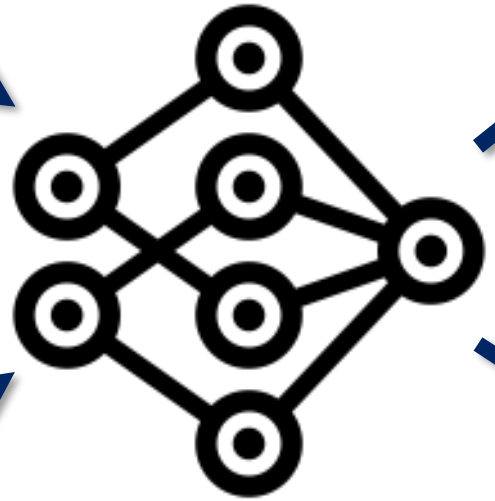
Machine learning models are widely deployed



Data serves as the bedrock for training models



Not all data is created equal

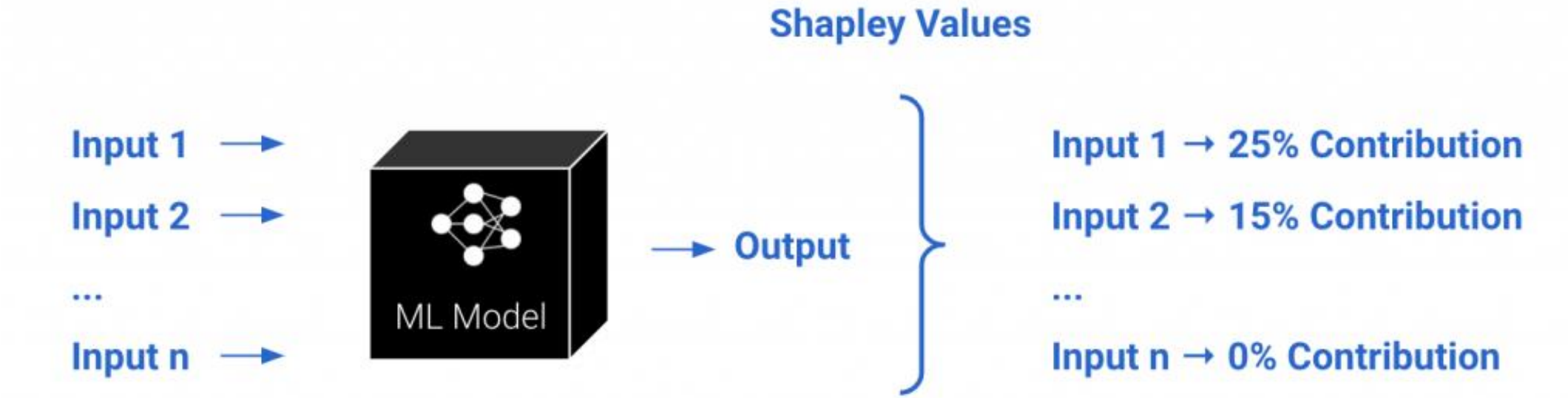


Model

25% Contribution

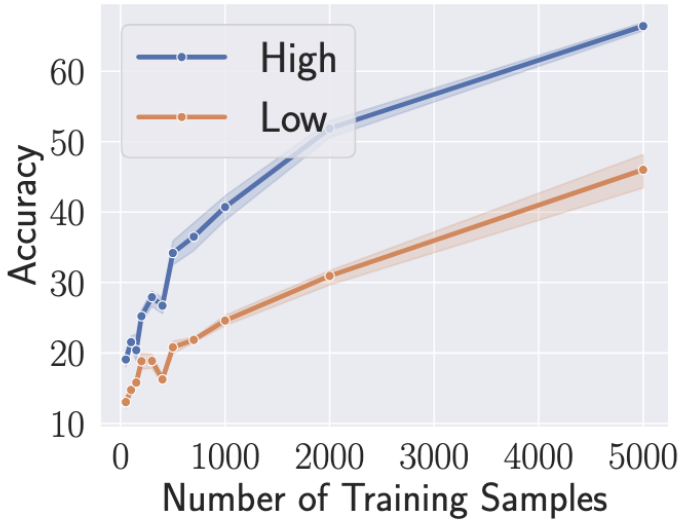
3% Contribution

Quantitatively measure the contribution

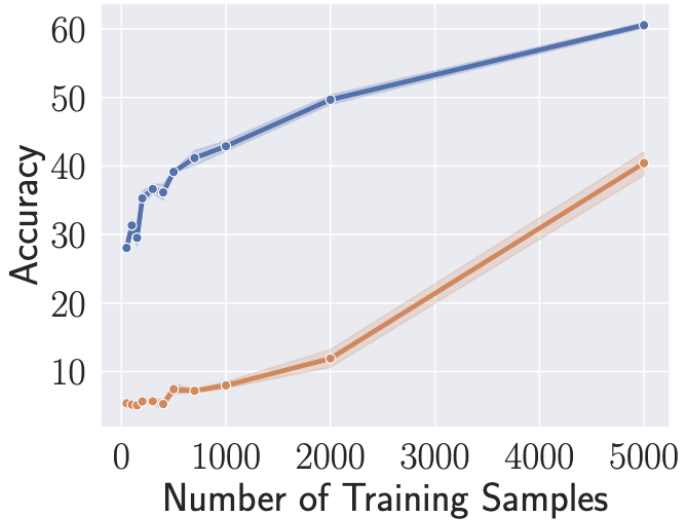


$$\nu_{\text{shap}}(z) \propto \frac{1}{N} \sum_{S \subseteq D \setminus \{z\}} \frac{1}{\binom{N-1}{|S|}} \left[U_{\mathcal{A}, D_{val}}(S \cup \{z\}) - U_{\mathcal{A}, D_{val}}(S) \right]$$

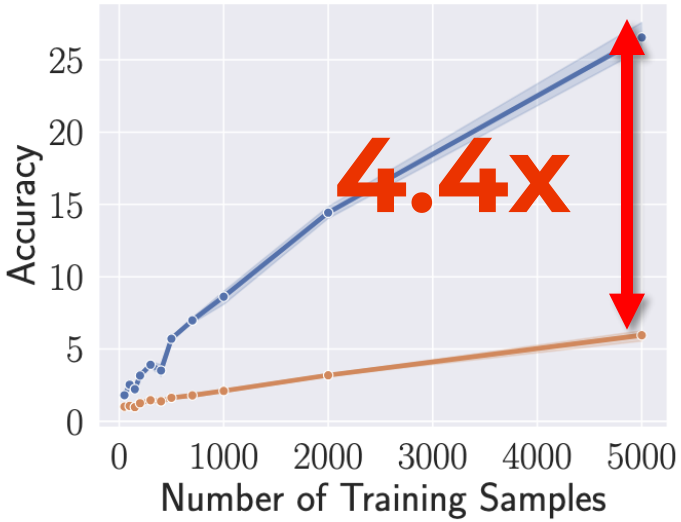
High importance data contributes much more



(a) CIFAR10



(b) CelebA



(c) TinyImageNet

Does variation in data impact their vulnerabilitiy?

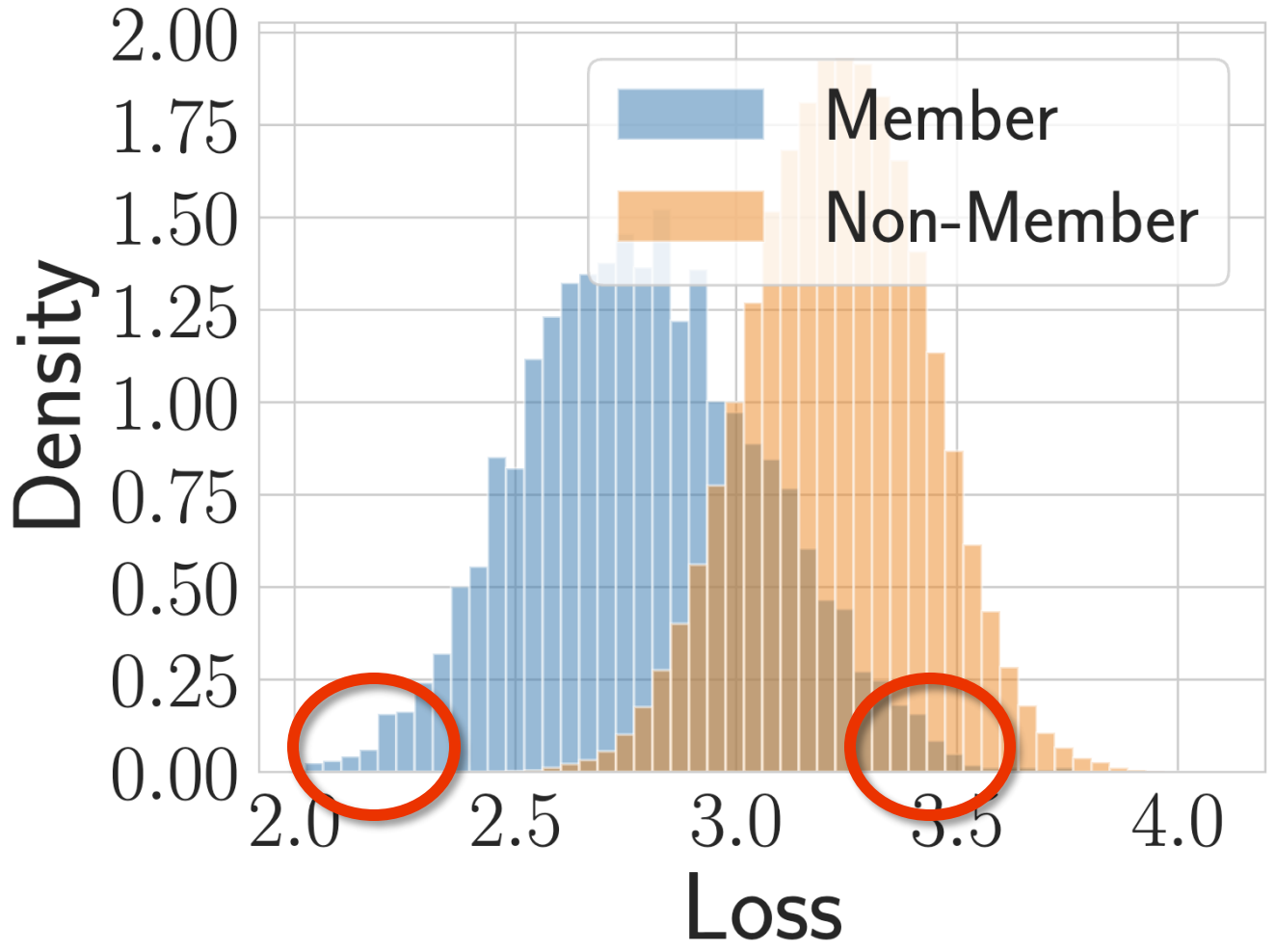
Data

Existing research mainly focuses on models

Data

Model

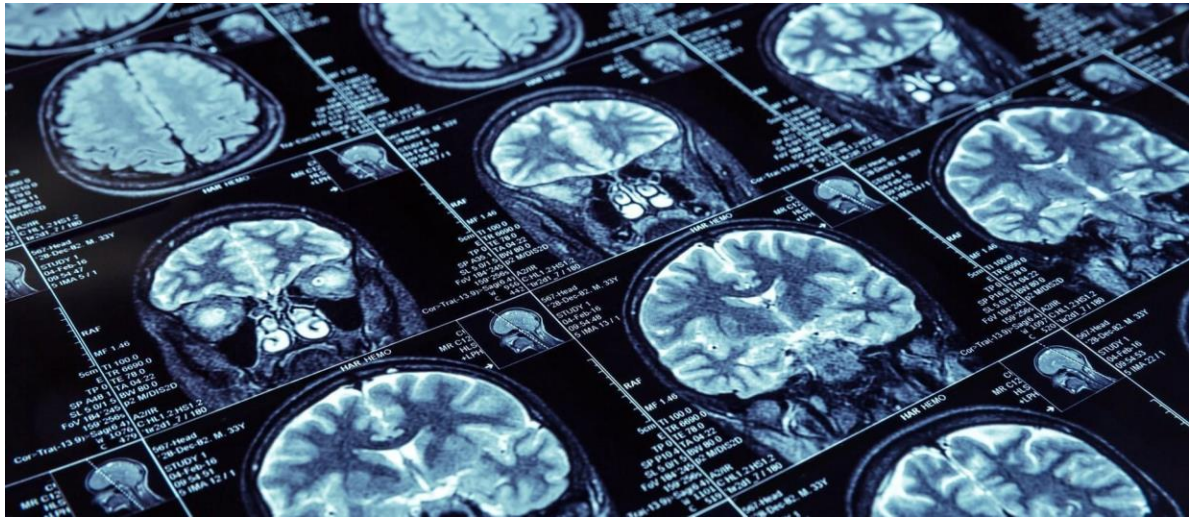
Data shows different levels of vulnerability



Are high importance samples more vulnerable?

High vulnerability?

This question has real-world implications



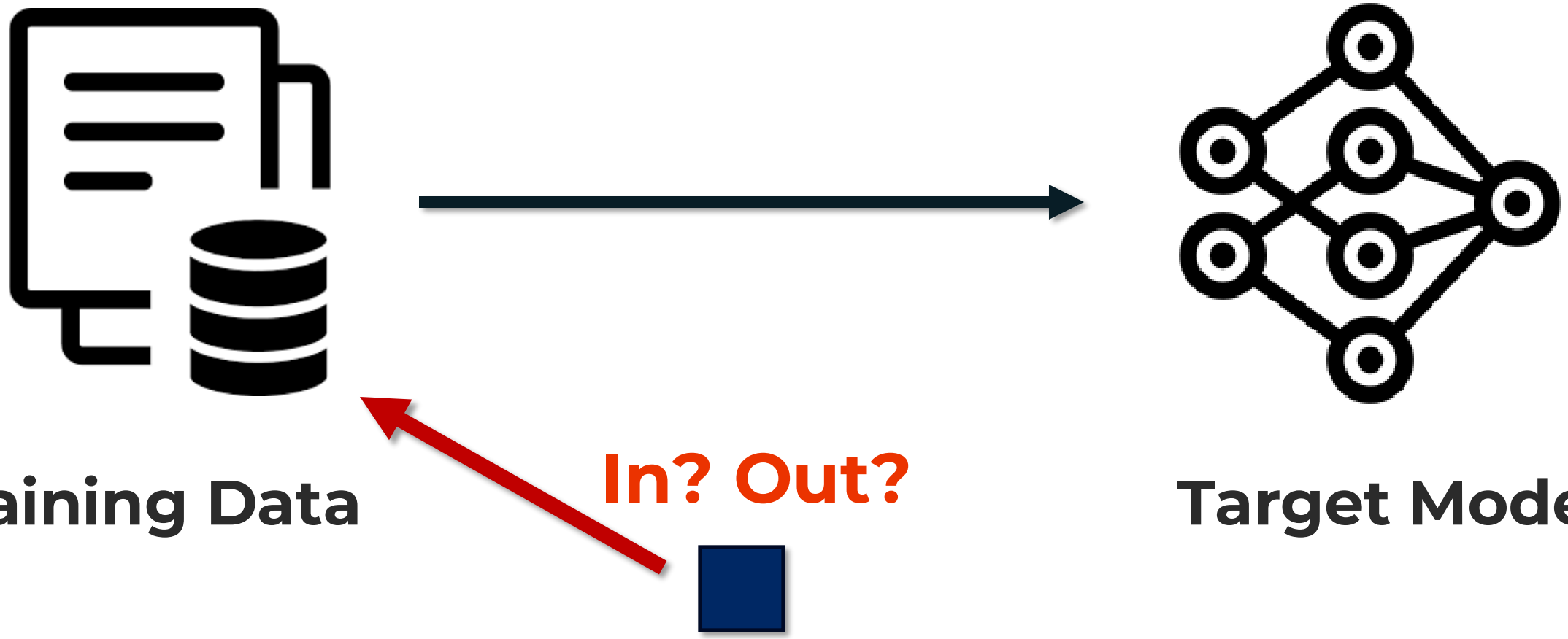
records with ***rare*** but ***crucial*** symptoms → high-importance

This question has real-world implications

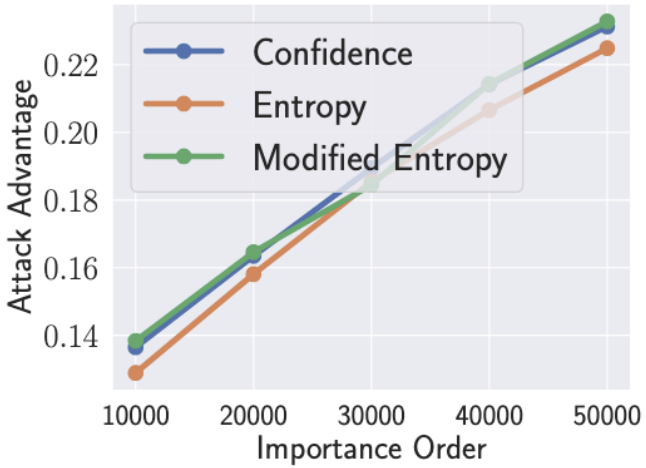
Discrimination

Premiums

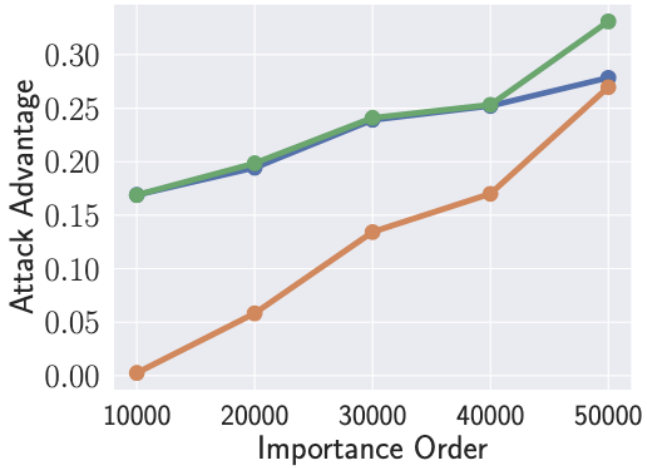
Membership Inference Attack (MIA)



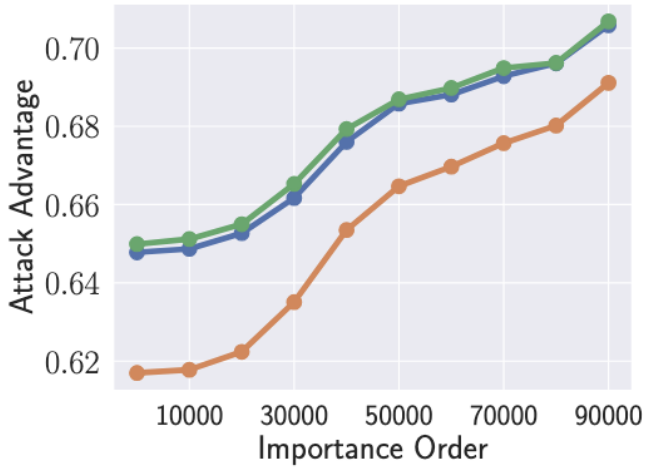
Vulnerability increases for high importance samples



(a) CIFAR10



(b) CelebA

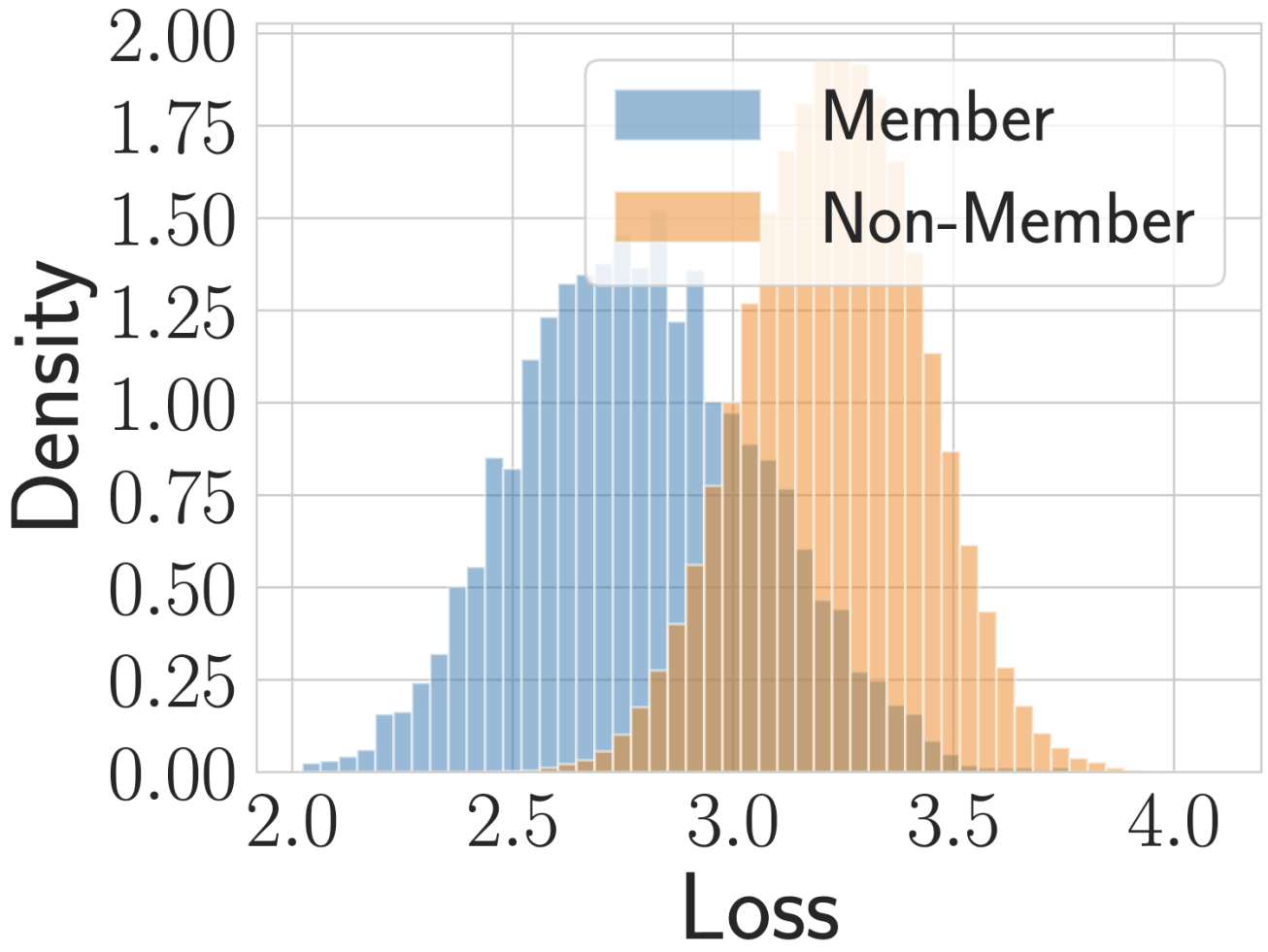


(c) TinyImageNet

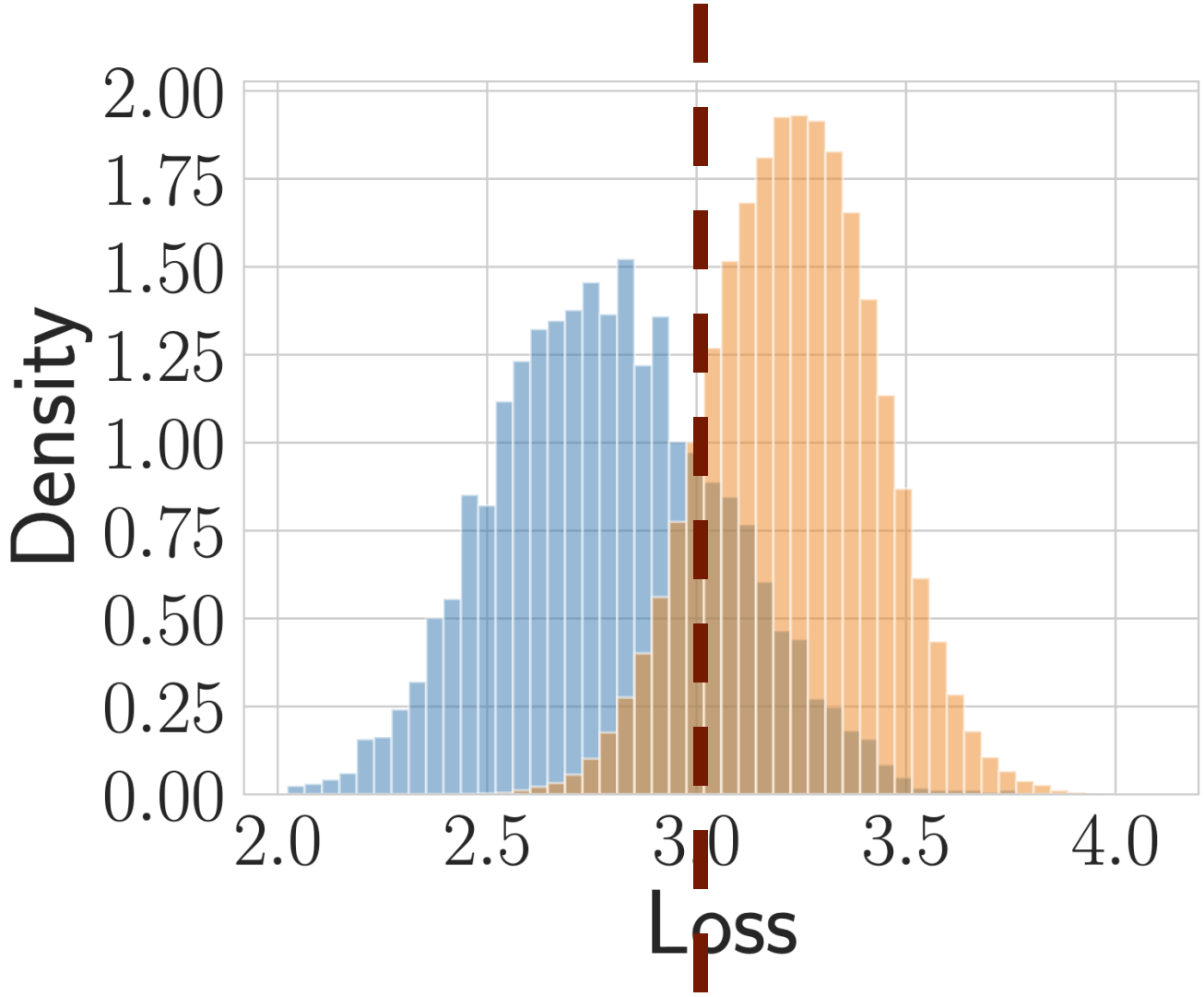
There is a strong connection

High vulnerability!

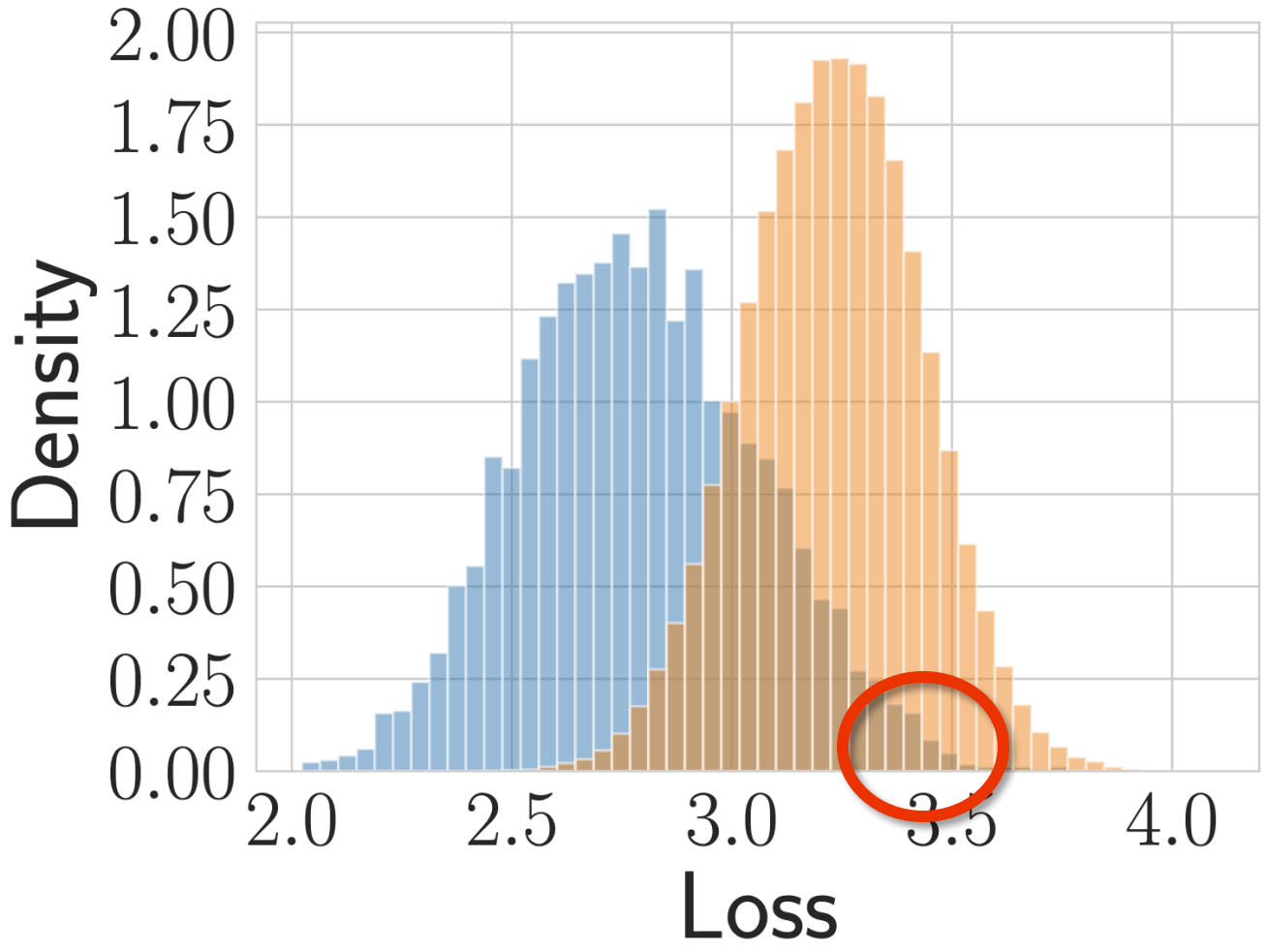
Loss-based MIA



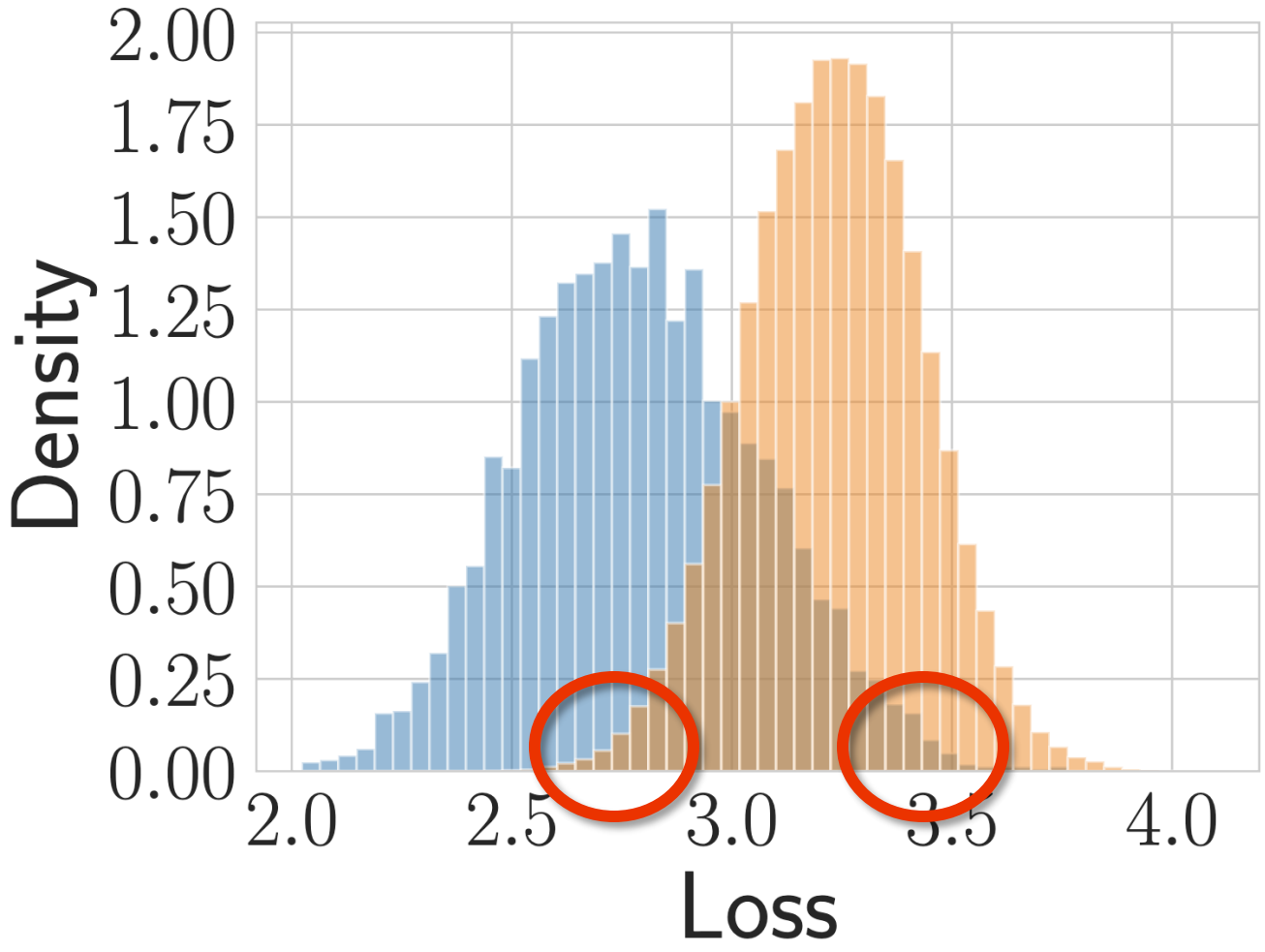
One threshold for all



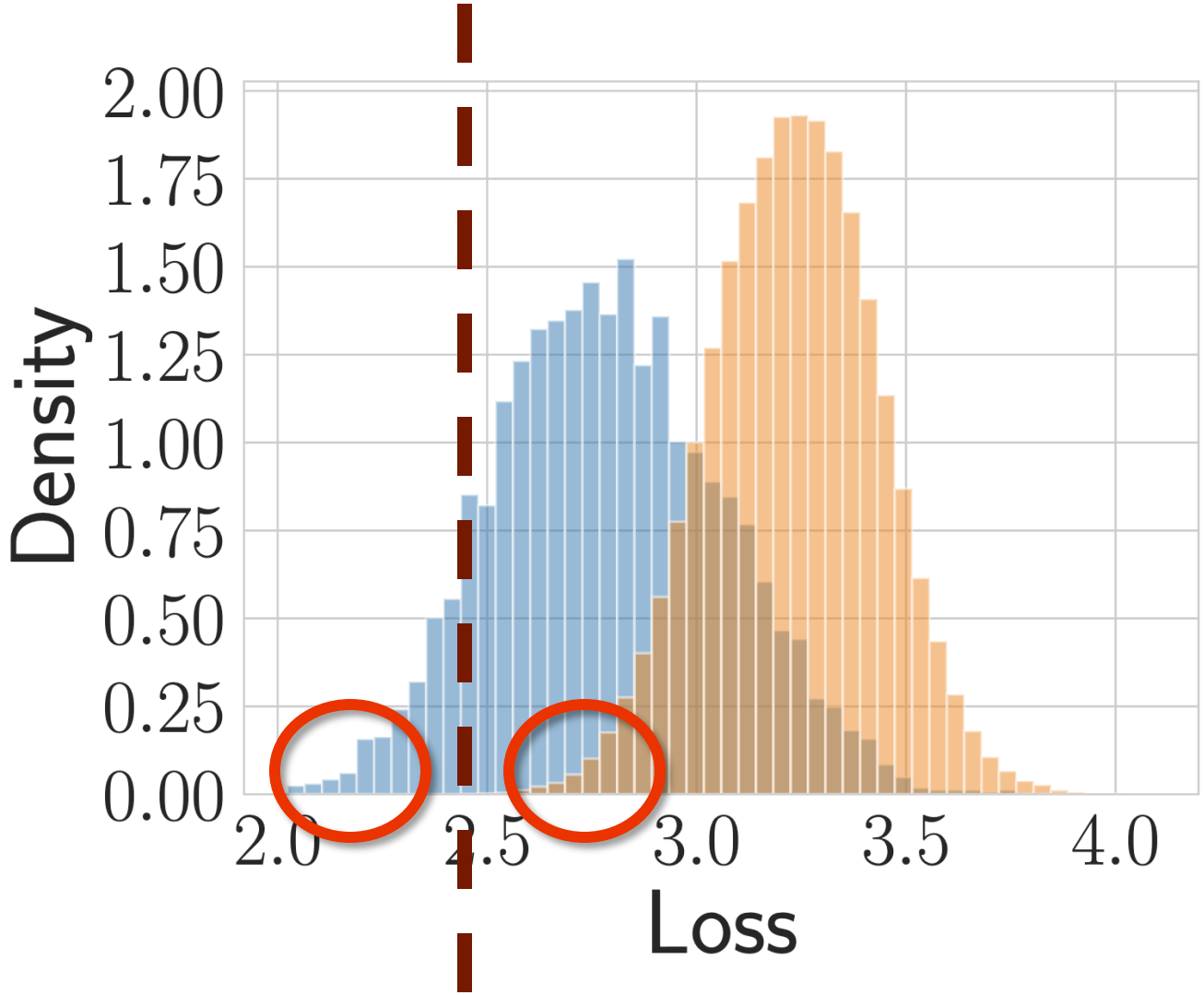
Low importance samples are harder to learn



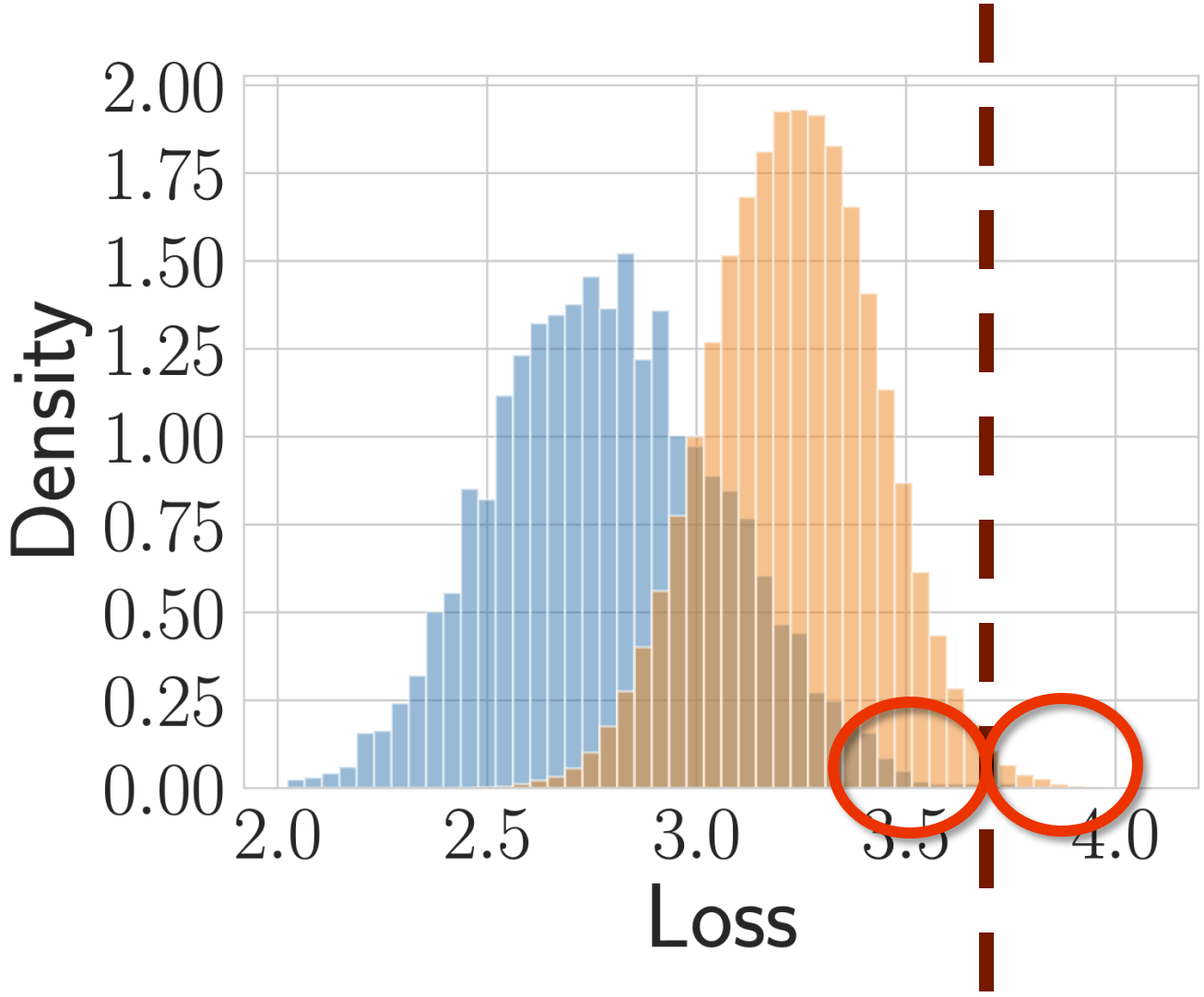
It‘s hard to identify low-importance members



Compare samples with similar importance

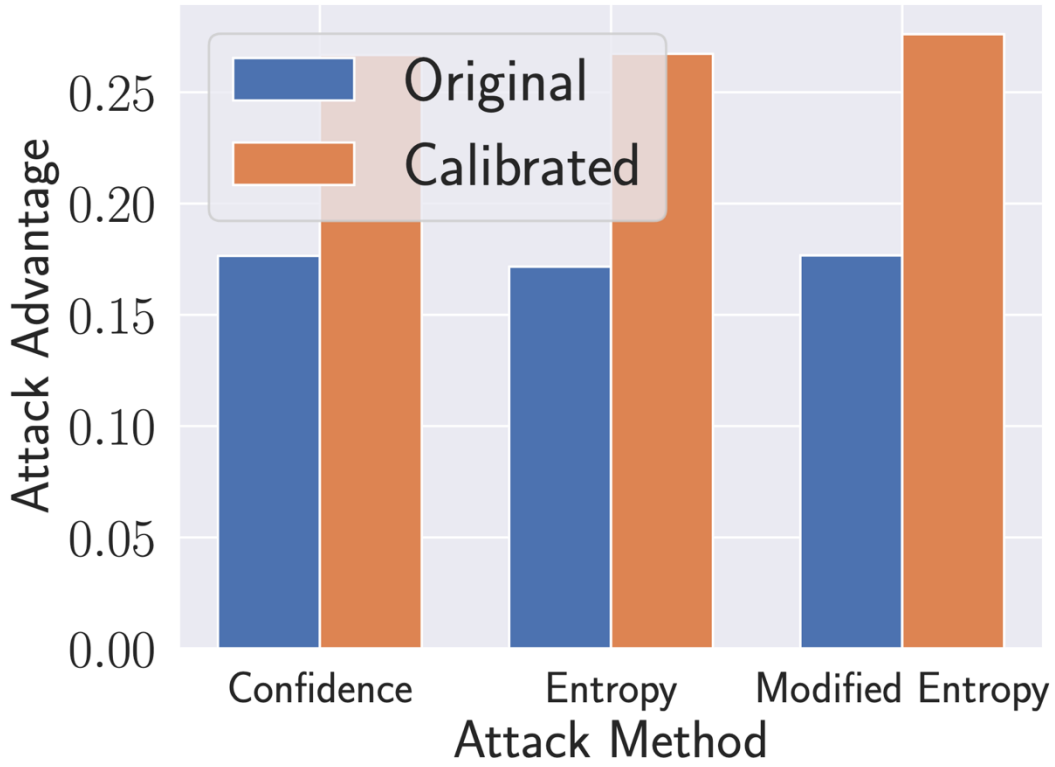


Set sample-specific threshold



Calibrate membership metric by sample importance

$$\texttt{CaliMem}(x) = \texttt{OriMem}(x) + k \times \texttt{Shapley}(x)$$



Take away: high-importance data is ***more vulnerable*** to membership inference attacks

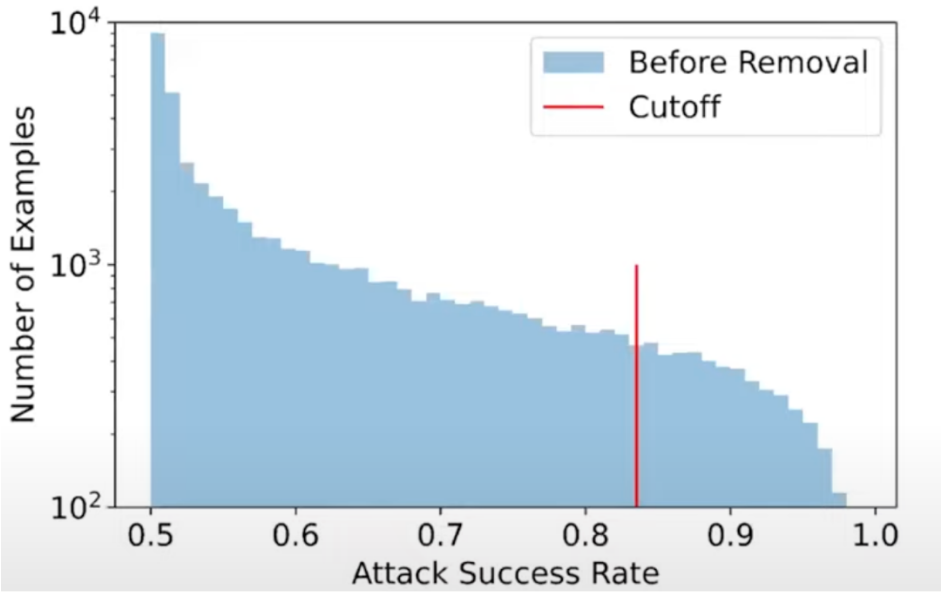
Setting ***sample-specific thresholds*** based on importance can make attacks even stronger.

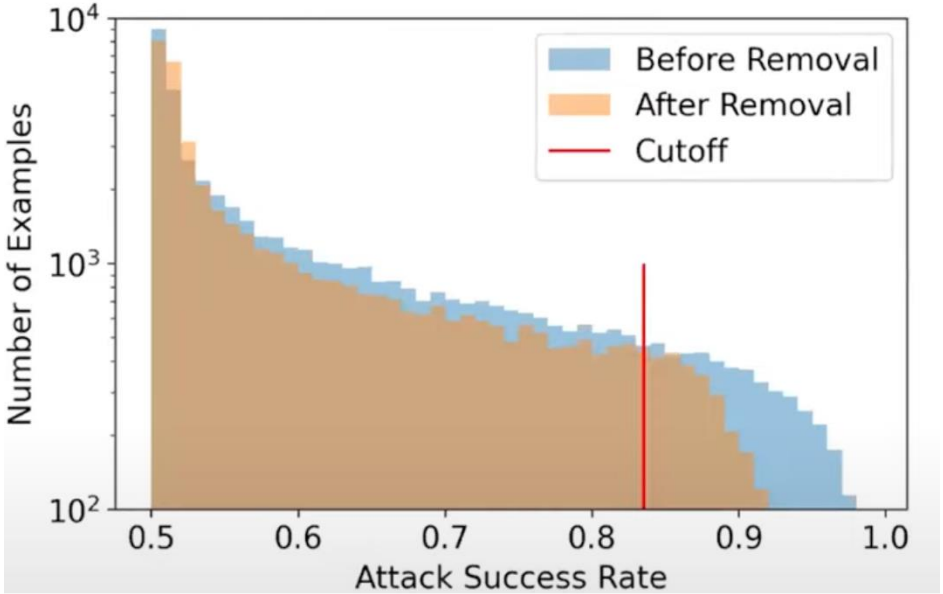
Privacy onion effect

Removing the "layer" of outlier points that are most vulnerable to a privacy attack exposes a new layer of previously-safe points to the same attack.

[1] The Privacy Onion Effect: Memorization is Relative

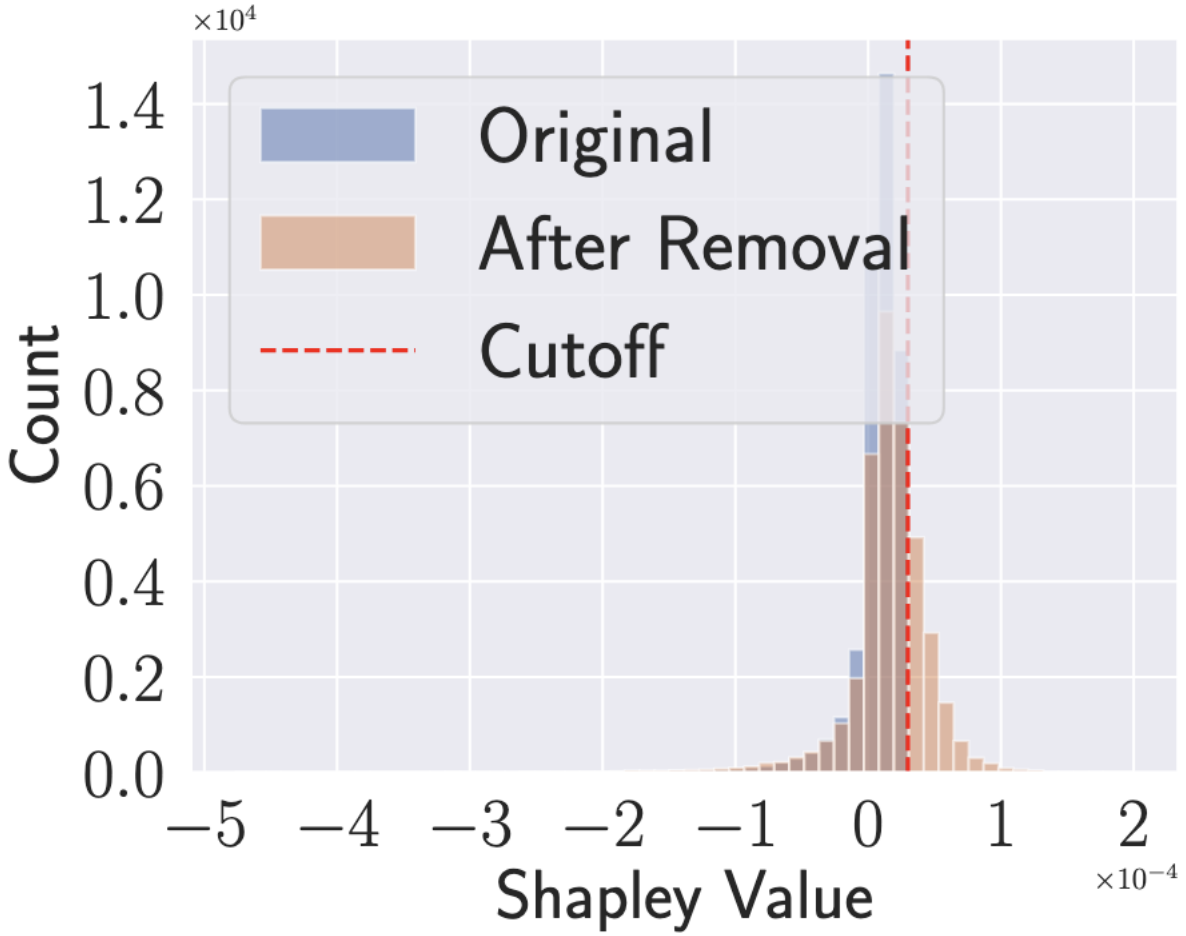
Privacy onion effect: an example



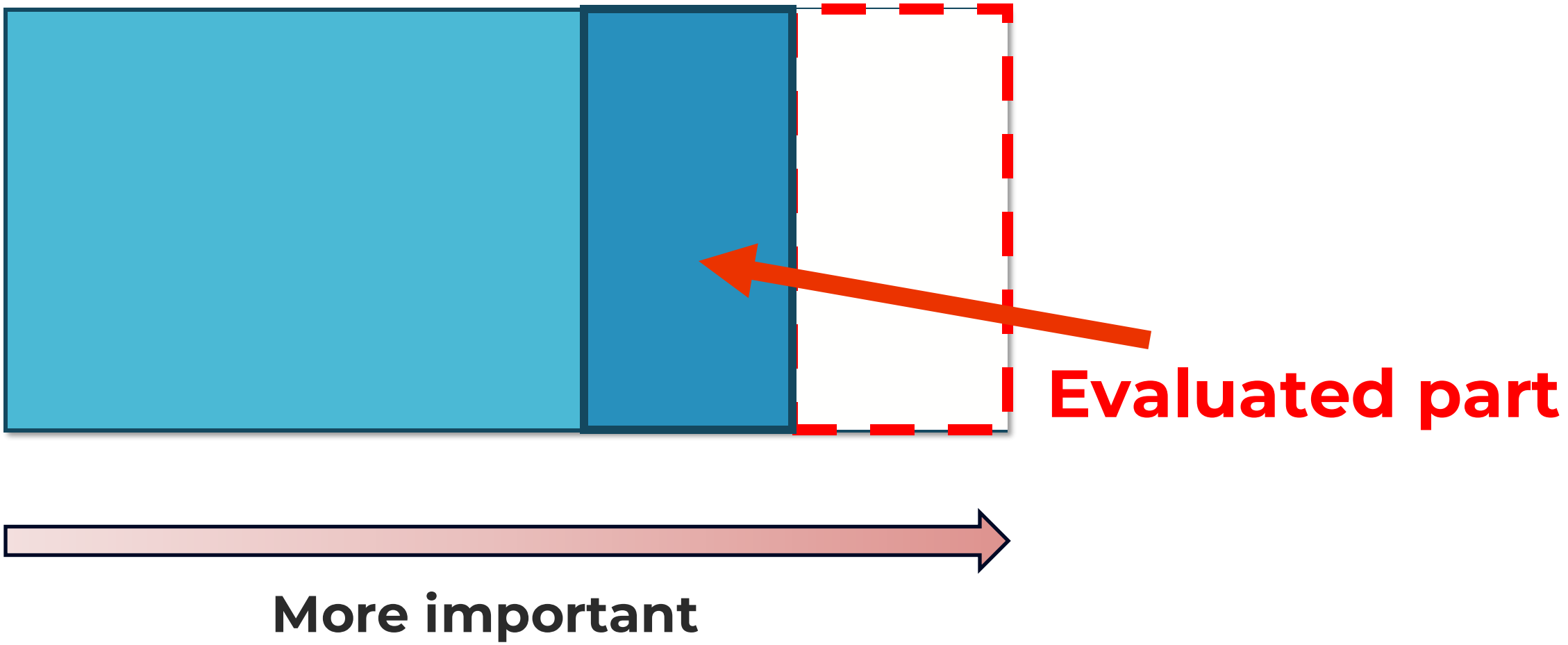


[1] Some Results on Privacy and Machine Unlearning, Matthew Jagielski

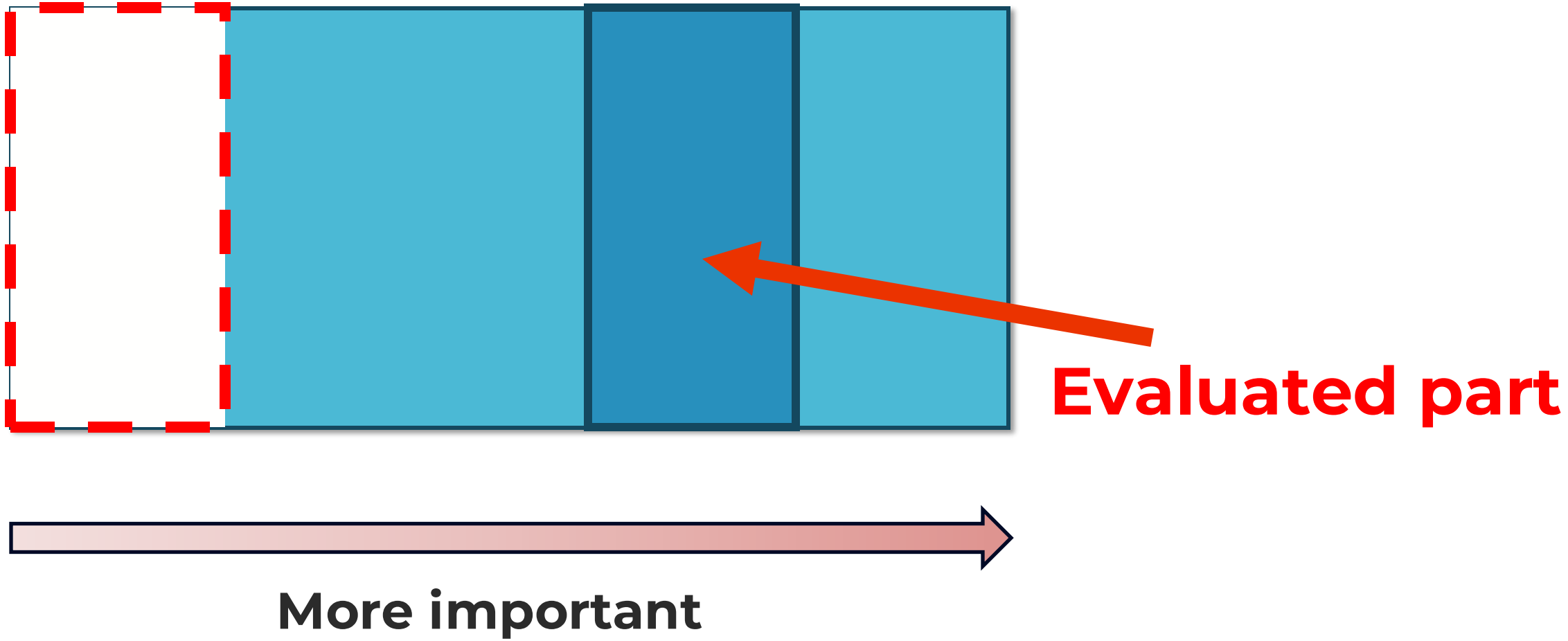
Privacy onion effect for importance



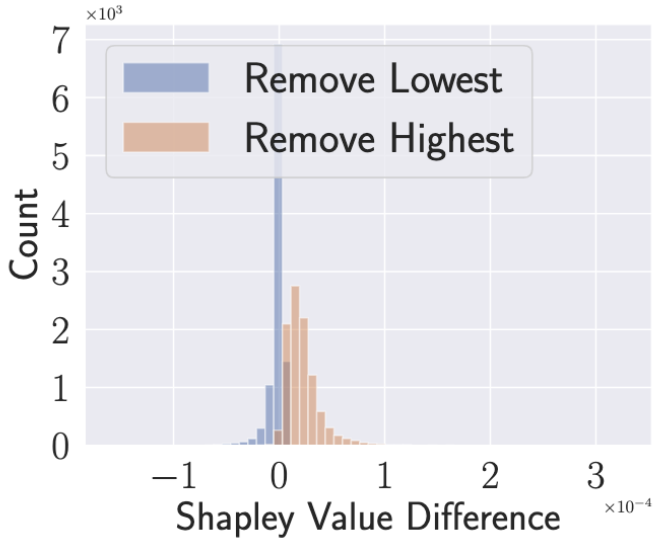
Privacy onion effect for importance



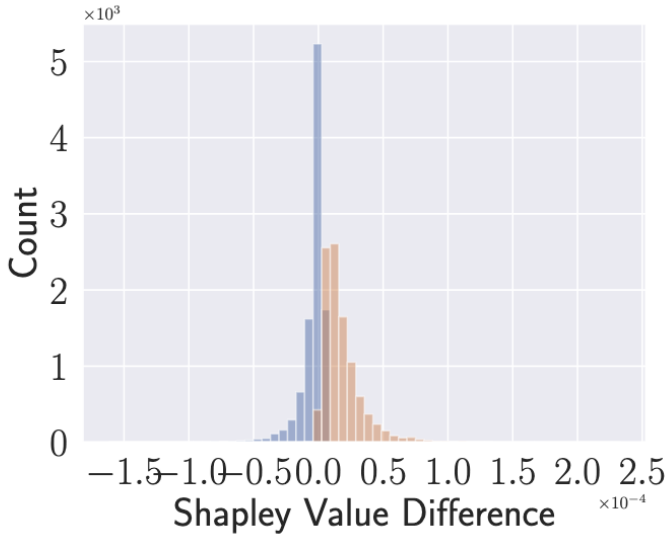
Privacy onion effect for importance



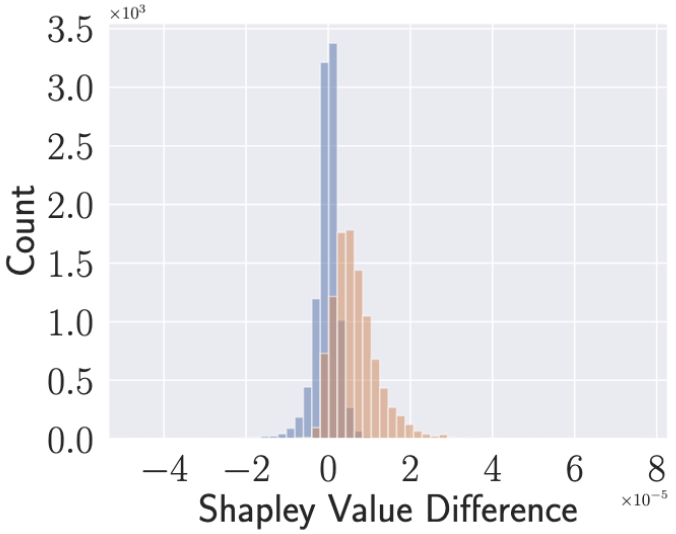
Privacy onion effect holds for importance value



(a) CIFAR10

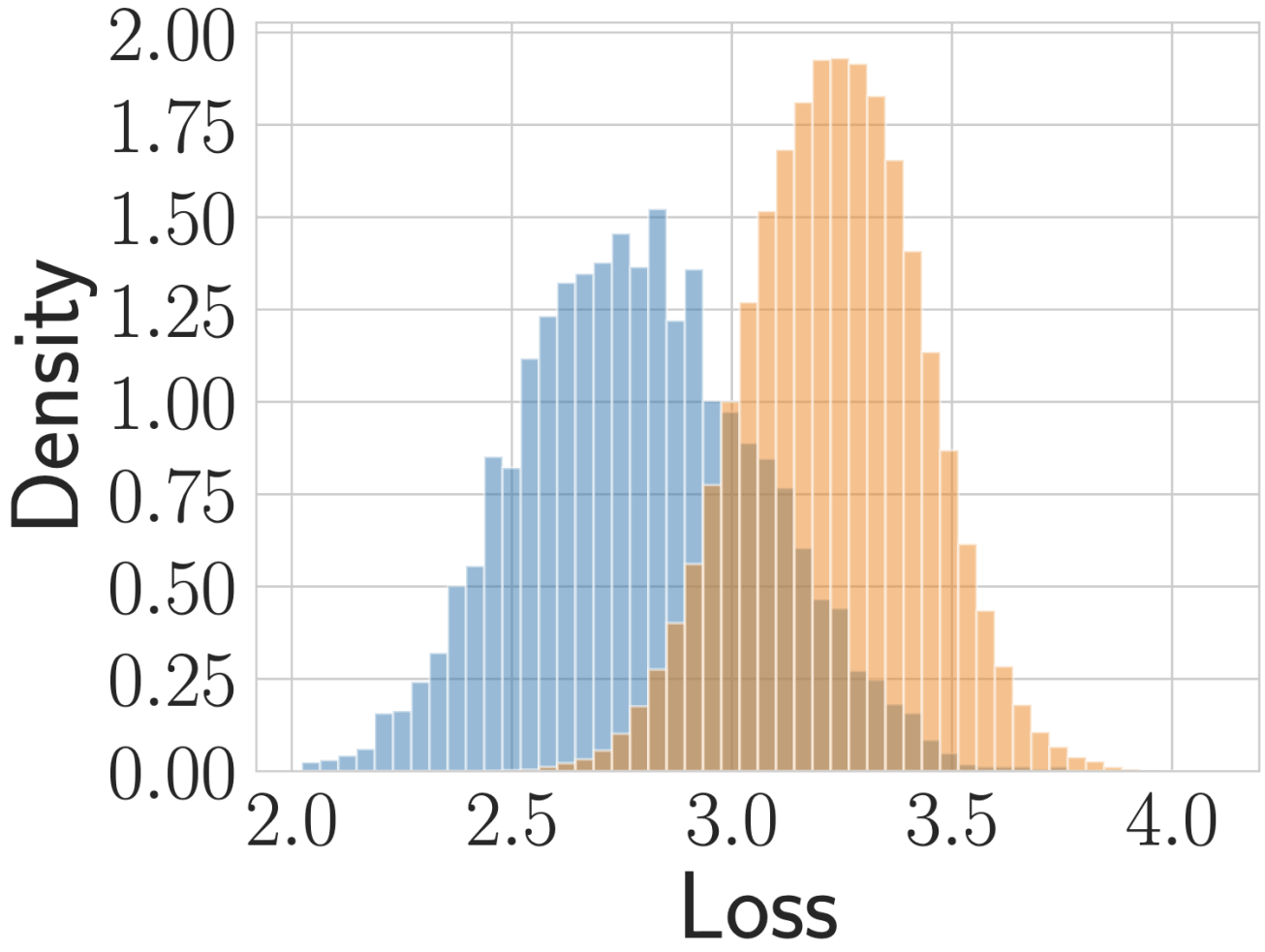


(b) Celeba

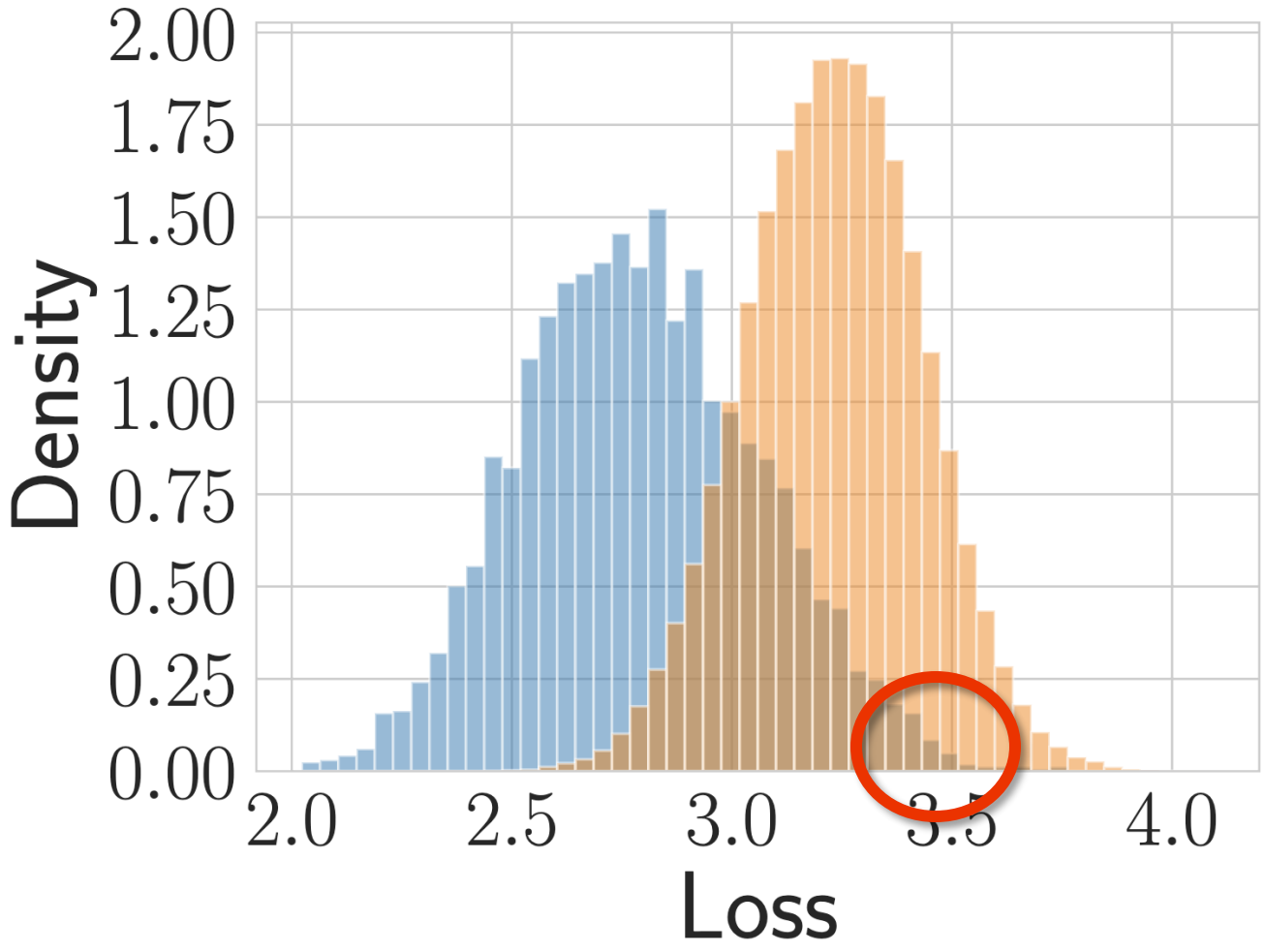


(c) TinyImageNet

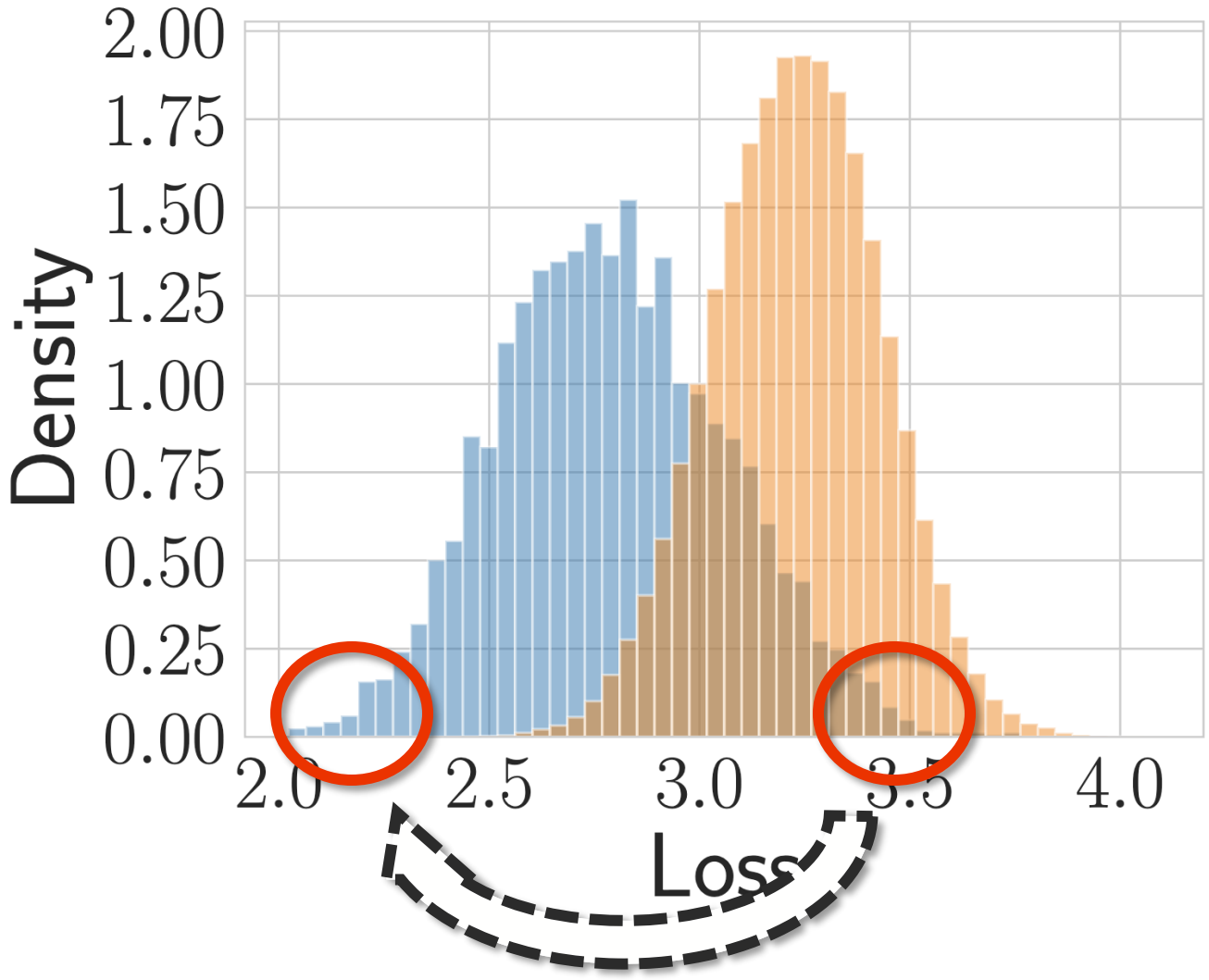
Increase vulnerability for target samples



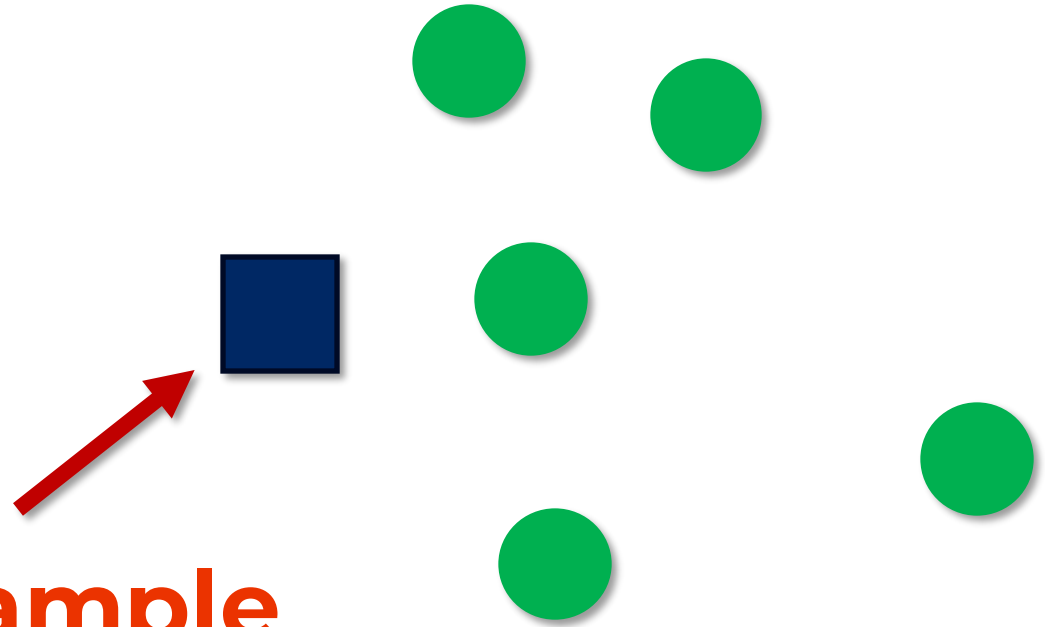
Increase vulnerability for target samples



Increase vulnerability for target samples

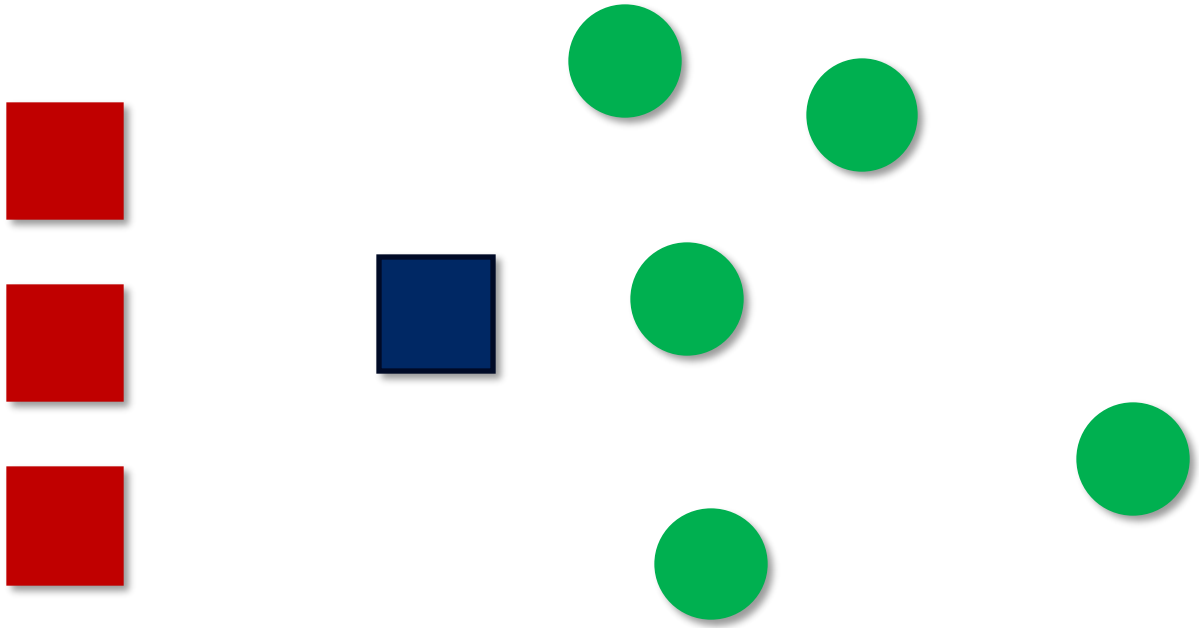


Activately modify importance

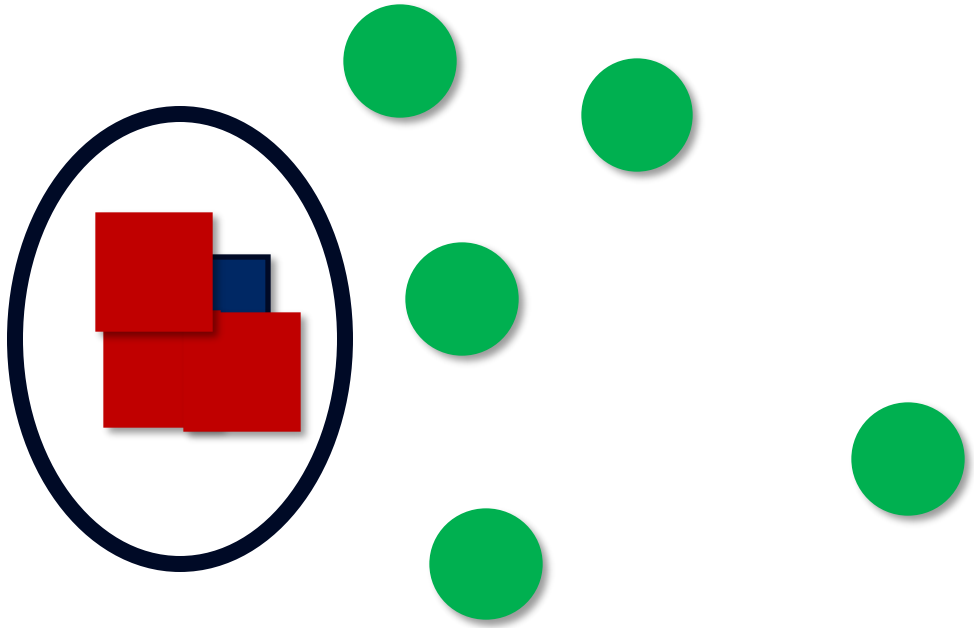


Target sample

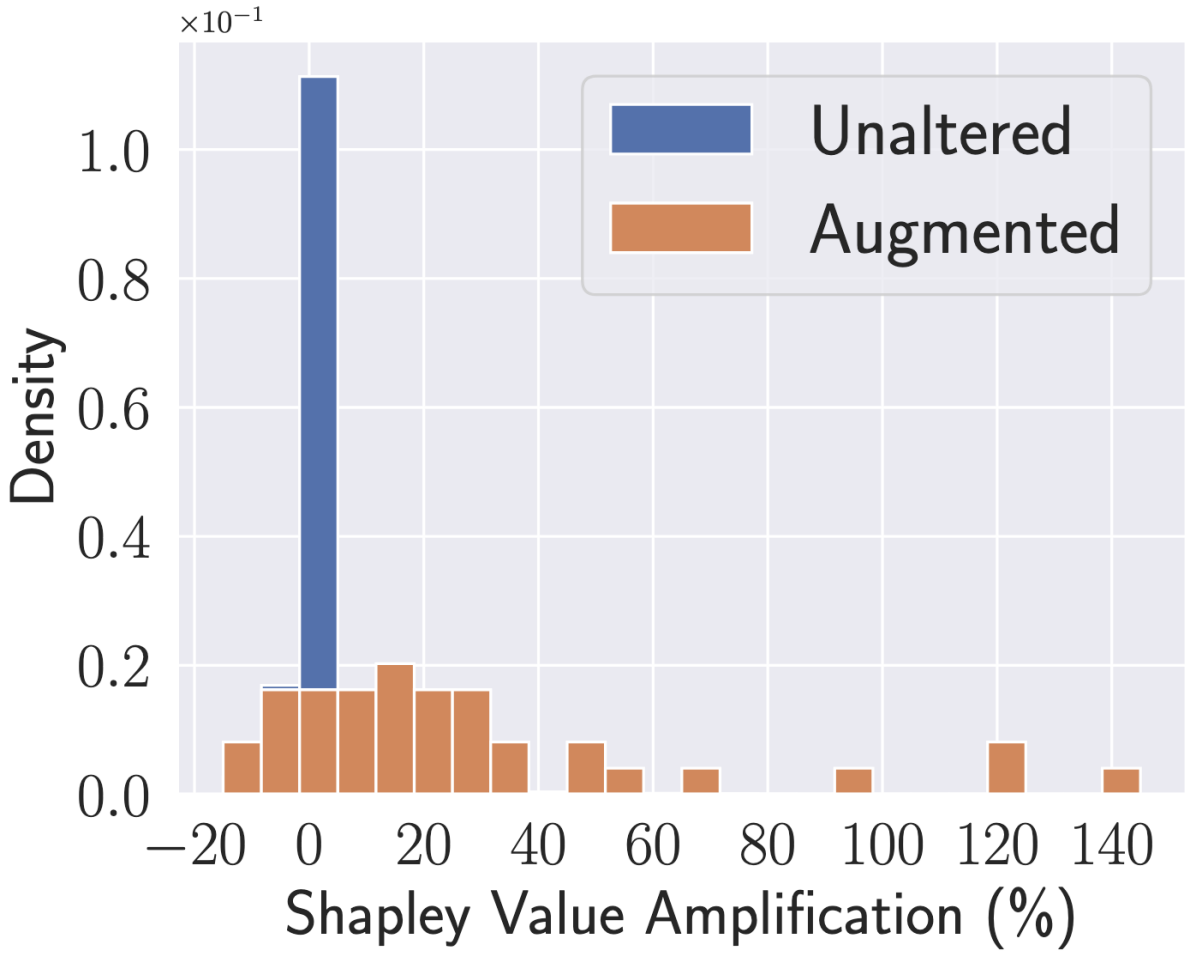
Activately modify importance



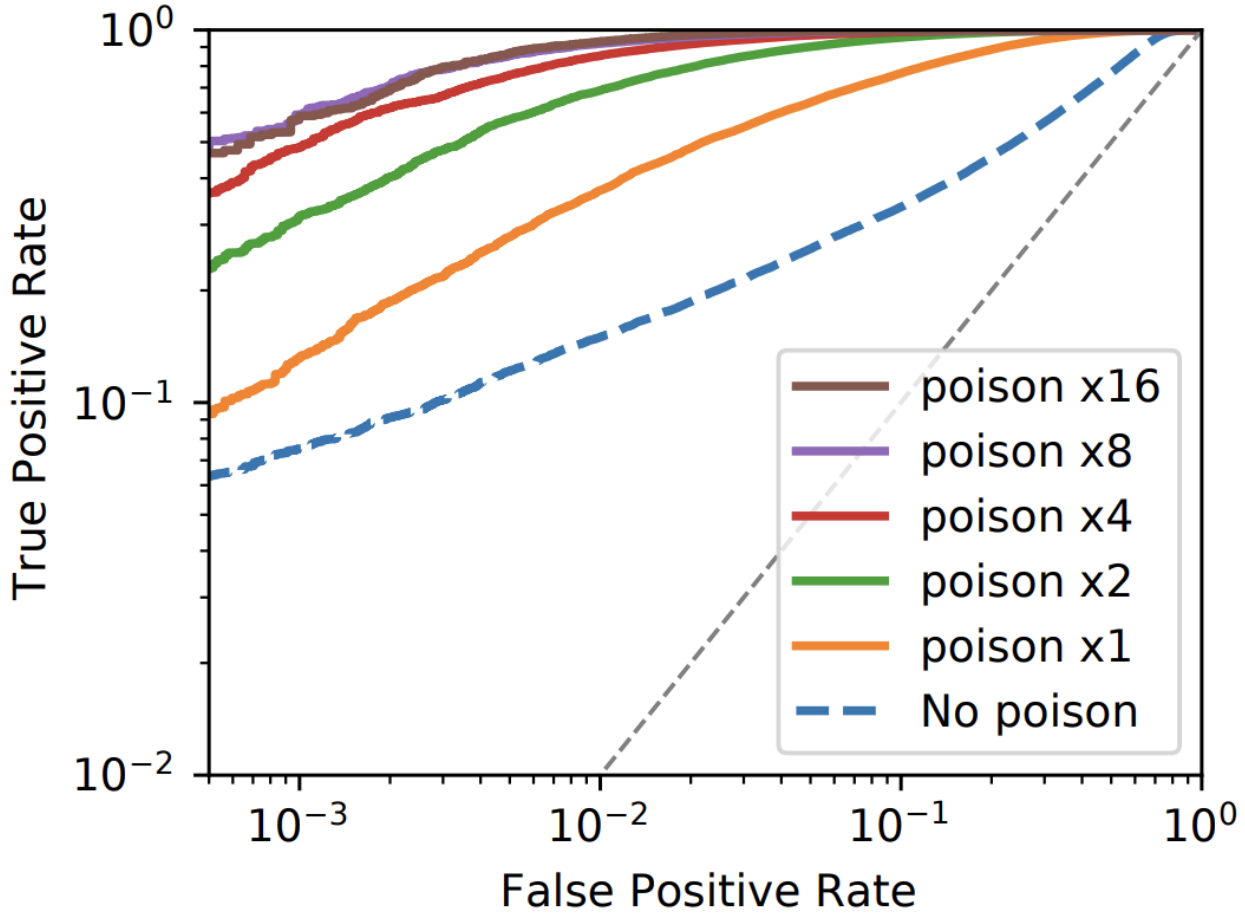
Activately modify importance



Activately modify importance



Activately modify importance can lead to stronger attack

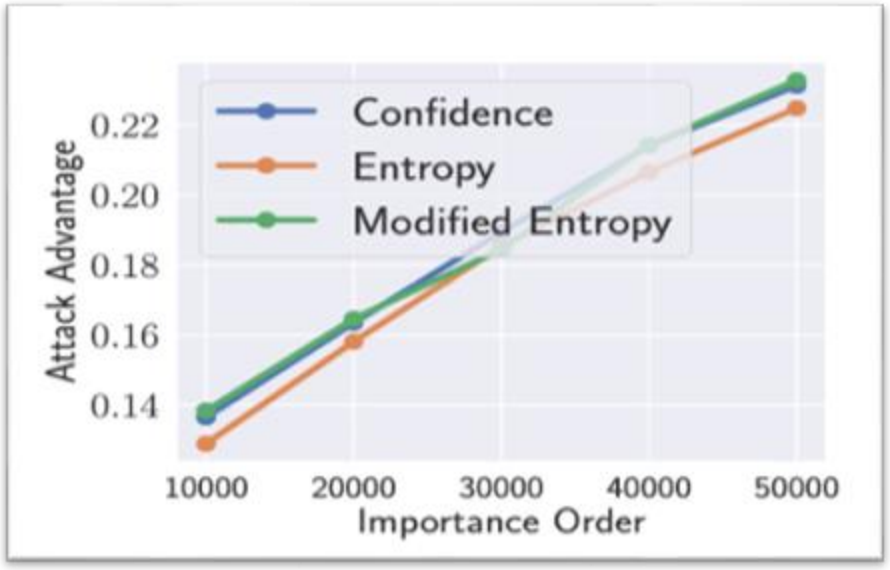


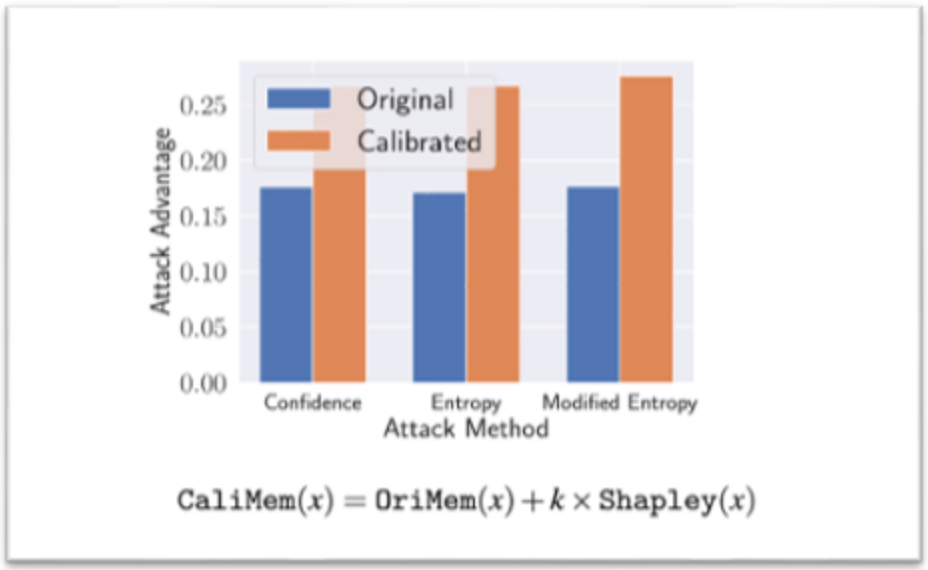
[1] Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets

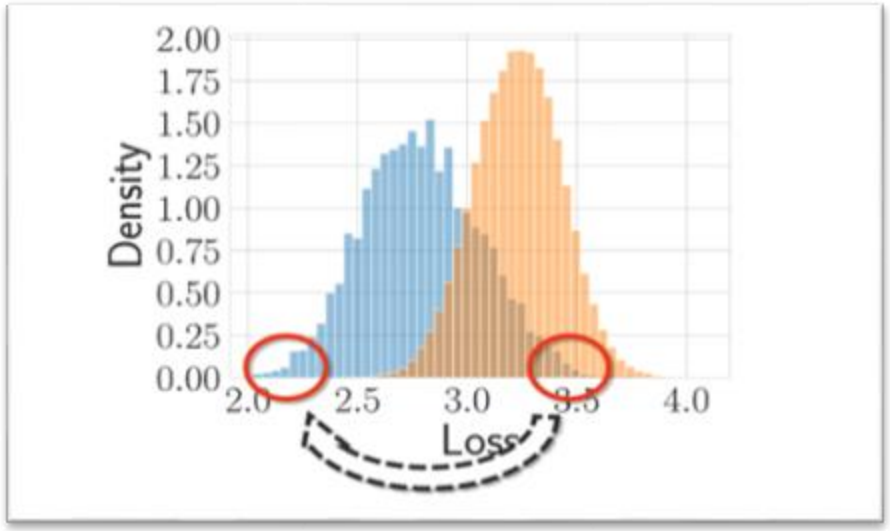
Take away: ***"privacy onion effect" holds*** for data importance.

Actively manipulating sample importance can be a potent strategy for developing stronger attacks.

Thank you!


Confidence
Entropy
Modified Entropy
Attack Advantage
0.22
0.20
0.18
0.16
0.14
10000
20000
30000
40000
50000
Importance Order


Original
Calibrated
Attack Advantage
0.25
0.20
0.15
0.10
0.05
0.00
Confidence
Entropy
Modified Entropy
Attack Method
$\texttt{CaliMem}(x) = \texttt{OriMem}(x) + k \times \texttt{Shapley}(x)$


Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
2.0
2.5
3.0
4.0
Loss


THANK YOU!