Microsoft Security

Evaluating Privacy in Machine Learning

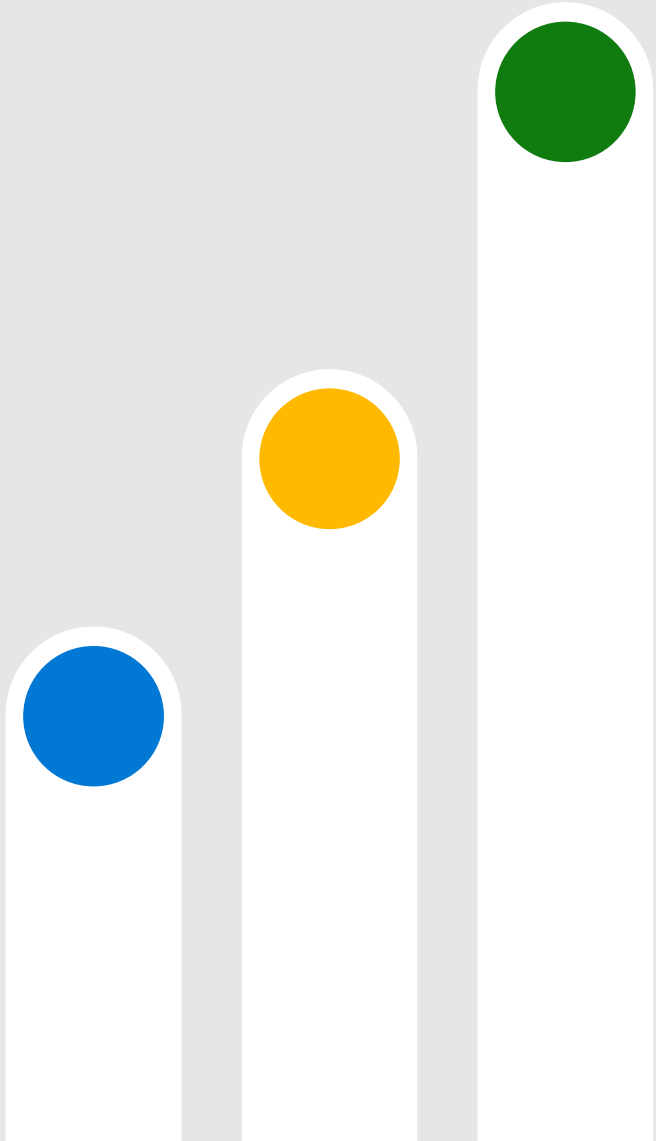
Andrew Paverd

Microsoft Security Response Center

Security for all in an AI enabled society – 4th July 2023

Outline

- Opportunities and risks for using private data
- *Games* for modelling privacy risks
- Empirical evaluation of privacy mechanisms



Opportunities: Domain-specific private data

Healthcare

Researchers designed a deep learning model, to accurately detect diabetic retinopathy, a leading cause of blindness. By **fine-tuning the model with private medical images**, they were able to identify patients at risk with 90% accuracy.

Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photograph" JAMA 2016

Fraud Detection

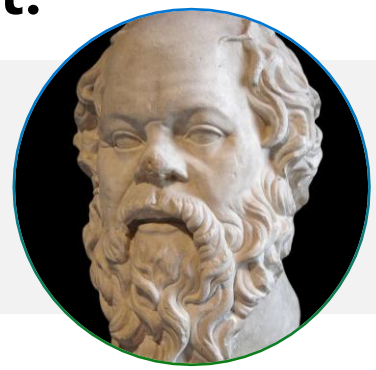
By **fine-tunning models with private transaction data**, financial institutions can build robust fraud-detection systems that identify suspicious activities and protect customers' financial assets.

Phua et al. "Minority report in fraud detection: classification of skewed data"

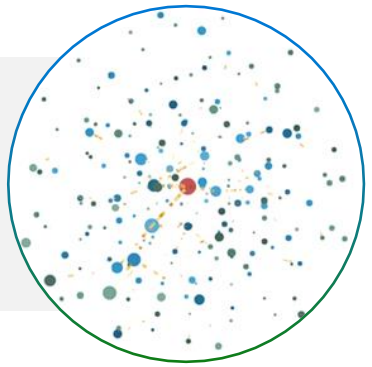
Opportunities: Domain-specific private data

We not only require big data, but also **the relevant data in the right context.**

Public model:
Socrates was a
Greek philosopher



Company-specific model:
Socrates is a *company wide initiative focusing on AI at Scale*

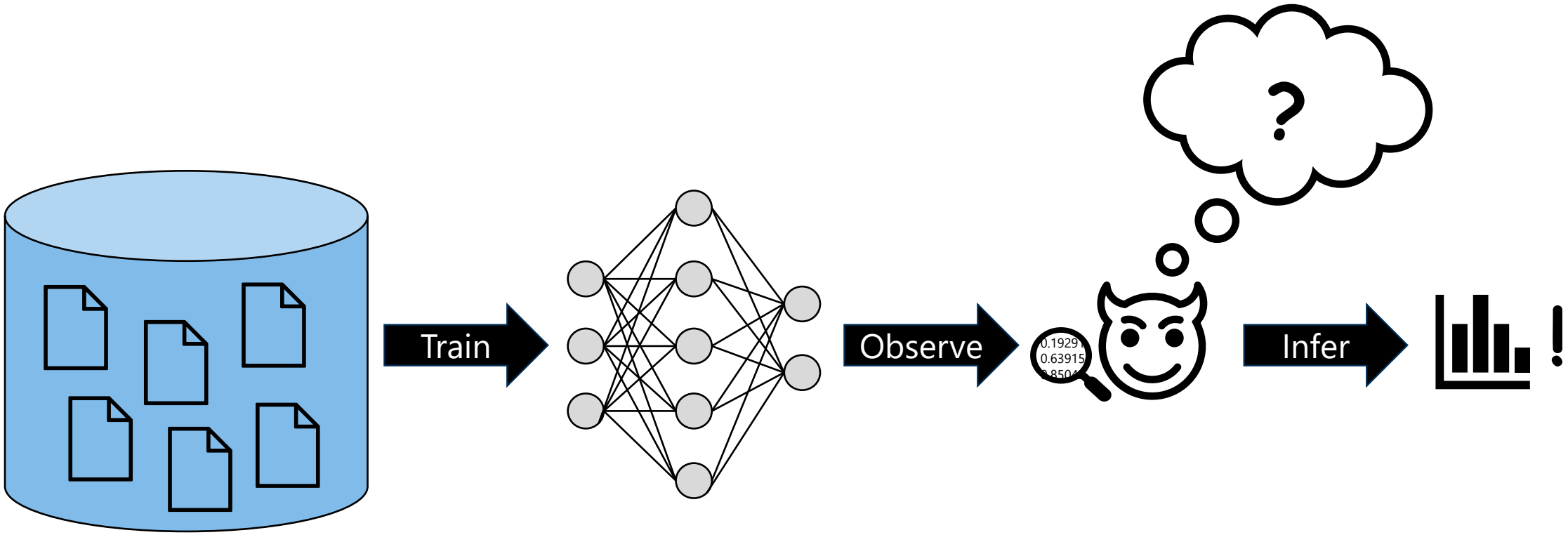


Morgan Stanley Wealth Management Announces Key Milestone in Innovation Journey with OpenAI

Mar 14, 2023

Microsoft Launches BioGPT, the ChatGPT of Lifescience

Inference threats against private data



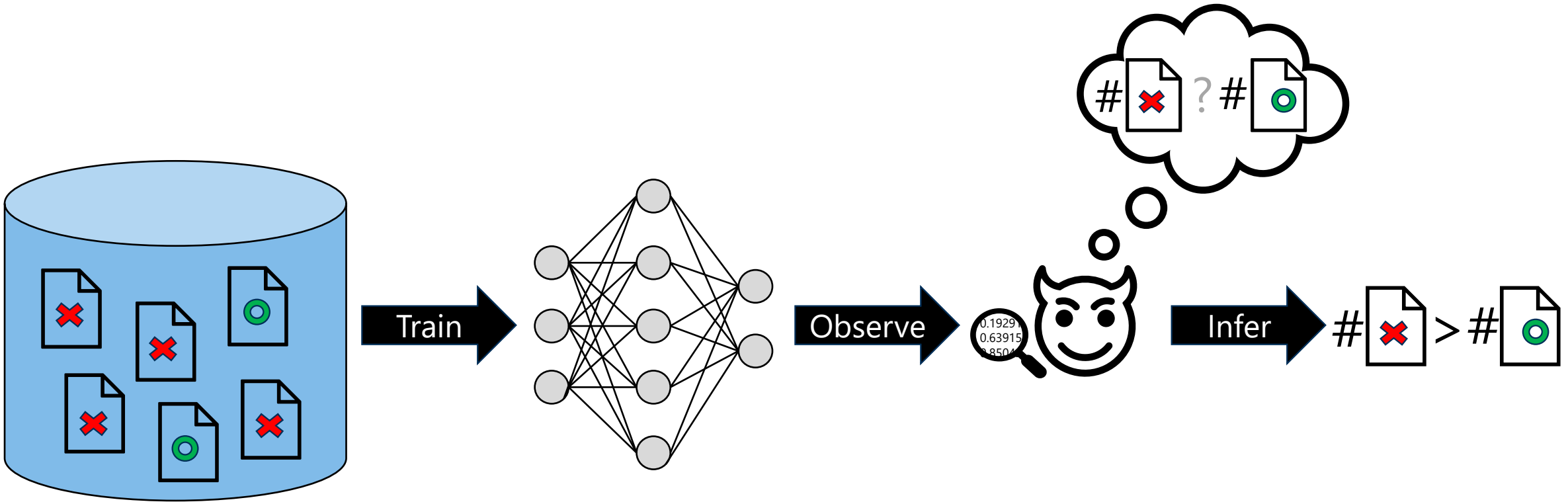
Training data

Trained model

Adversary

Information leaked

Property Inference (PI)



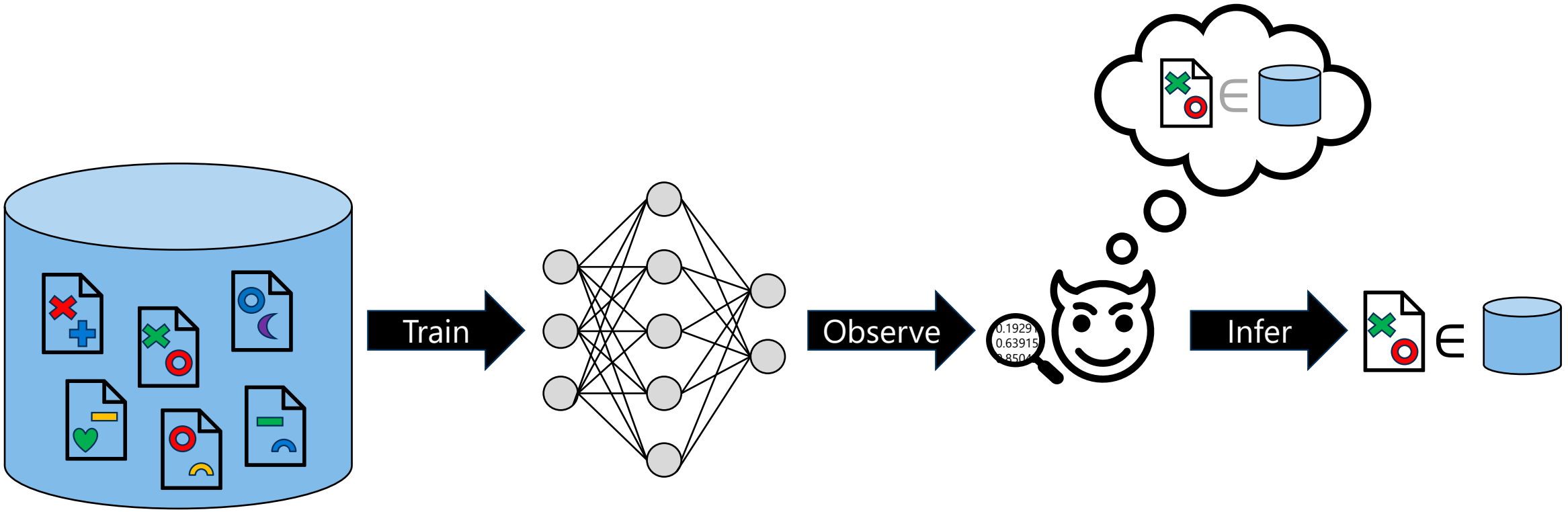
Training data

Trained model

Adversary

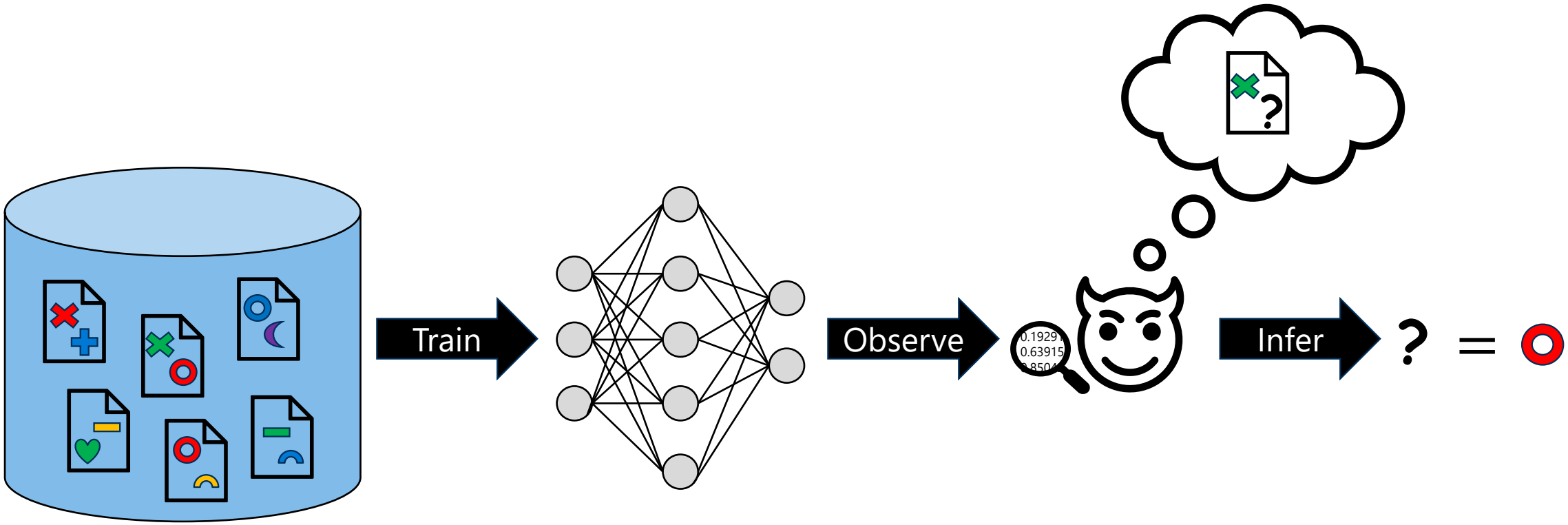
Information leaked

Membership Inference (MI)



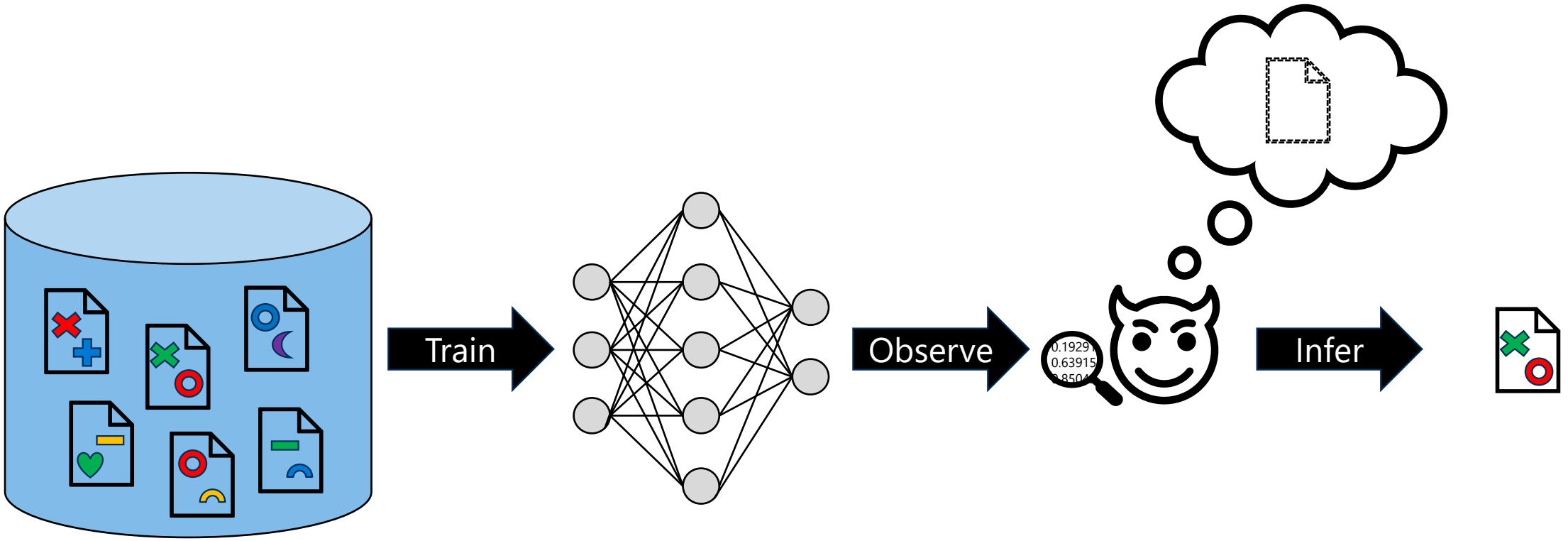
Training data | Trained model | Adversary | Information leaked

Attribute Inference (AI)

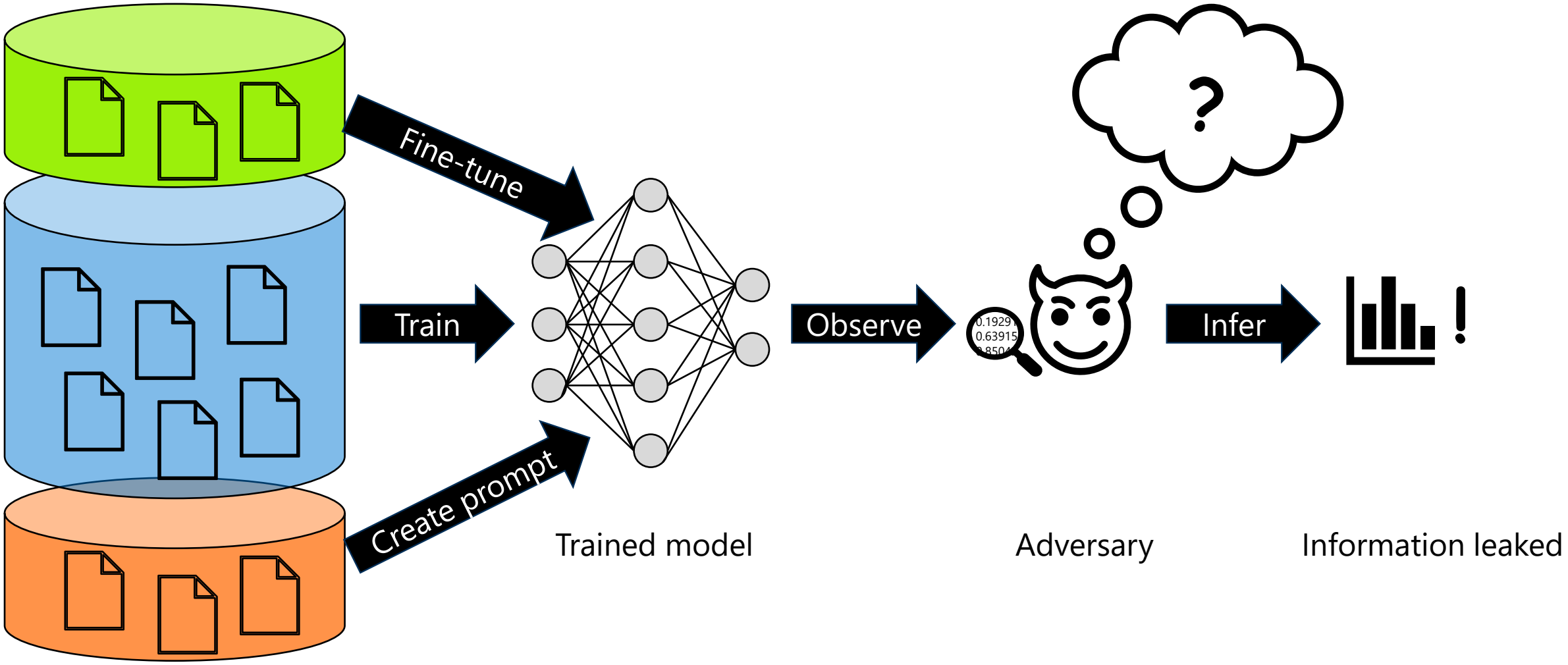


Training data | Trained model | Adversary | Information leaked

Data Reconstruction (RC)



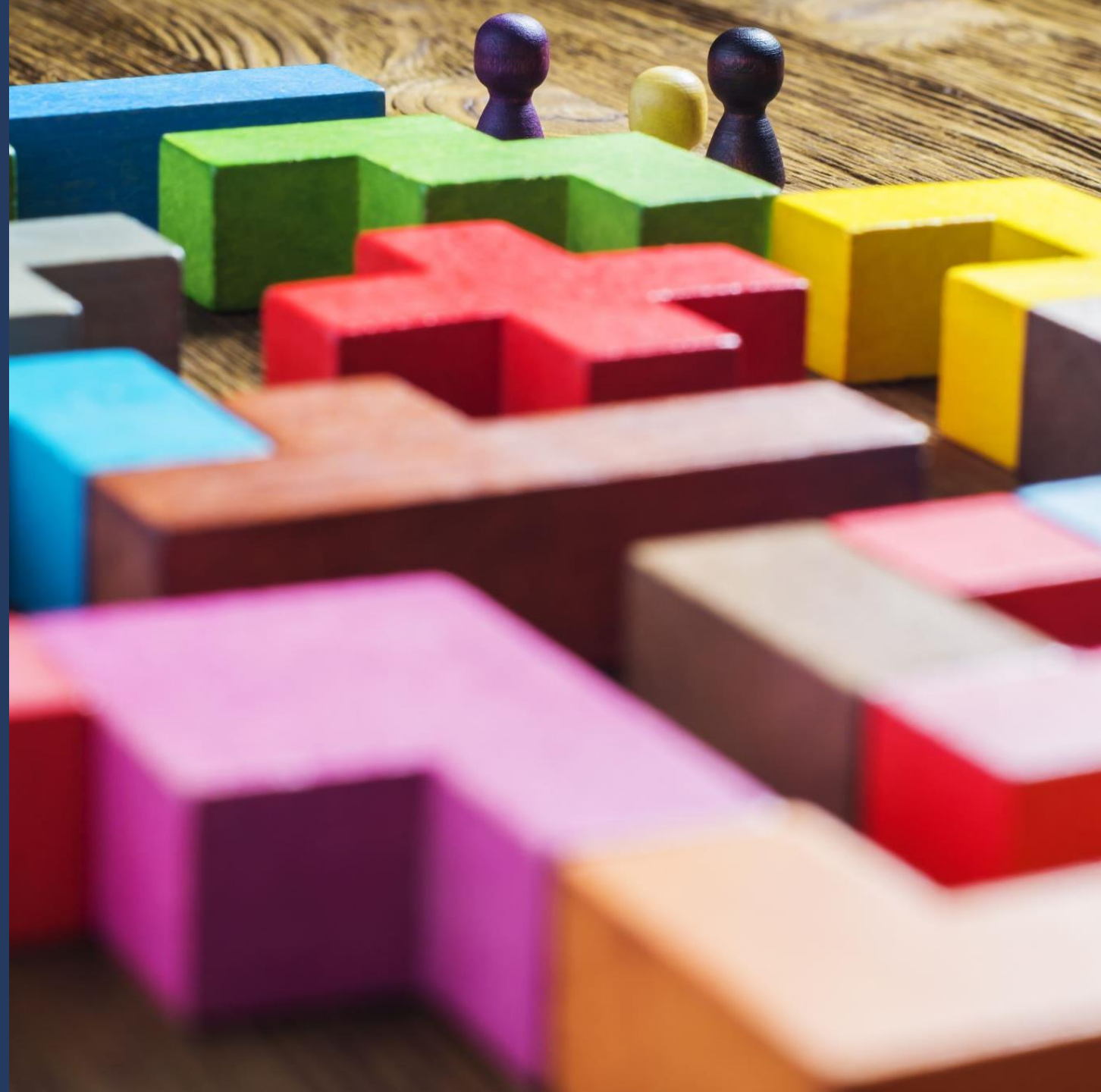
Inference threats against *any* private data



Microsoft Security

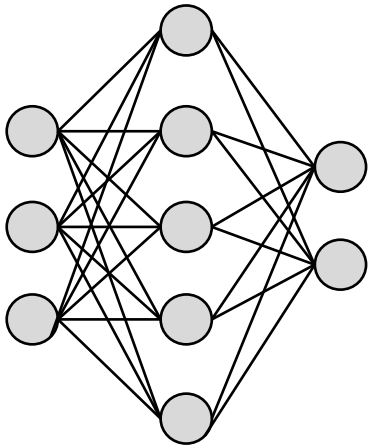
Games for modelling privacy risks

Joint work with: Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Anshuman Suri, Shruti Tople, Santiago Zanella-Béguelin



How do we quantify inference threats?

Threat modelling: Is this model vulnerable to membership inference?



Trained model

- Which attacks are relevant?
- How is the training data constructed?
- How is the target datapoint selected?
- What can the adversary observe?
- How do we measure adversary success?

Game-based definitions

Cryptography

IND−CPA(KG, Enc, $\mathcal{A}$)

$b \sim \{0,1\}$

$pk, sk \leftarrow \mathrm{KG}(\,)$

$m_0, m_1 \leftarrow \mathcal{A}(pk)$

$c \leftarrow \mathrm{Enc}(pk, m_b)$

$\tilde{b} \leftarrow \mathcal{A}(pk, c)$

Machine Learning

MI(Train, $\mathcal{D}, n, \mathcal{A}$)

$b \sim \{0,1\}$

$S \sim \mathcal{D}^{n-1}$

$z_0, z_1 \leftarrow \mathcal{A}(S)$

$\theta \leftarrow \mathrm{Train}(S \cup \{z_b\})$

$\tilde{b} \leftarrow \mathcal{A}(\mathrm{S}, \theta)$

flip a fair coin

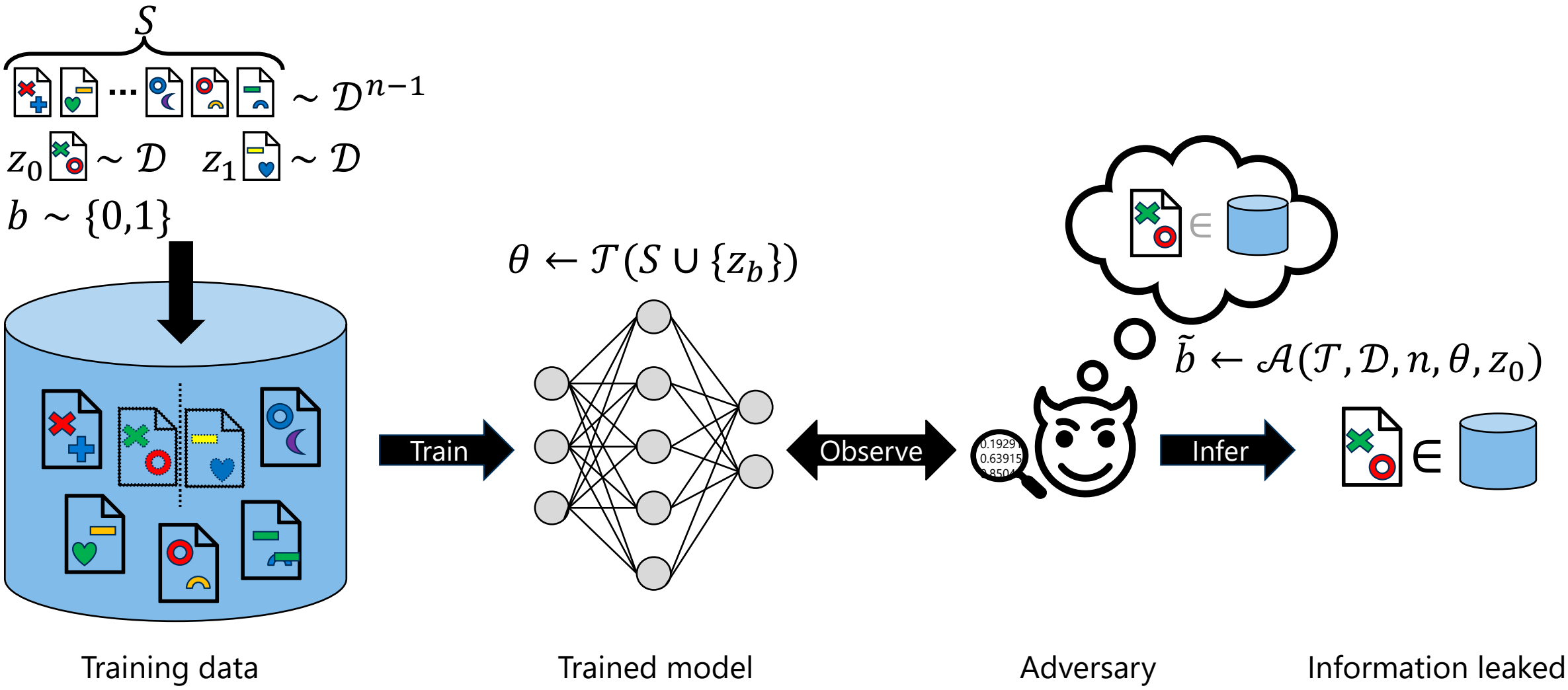
construct challenge

adversary guess

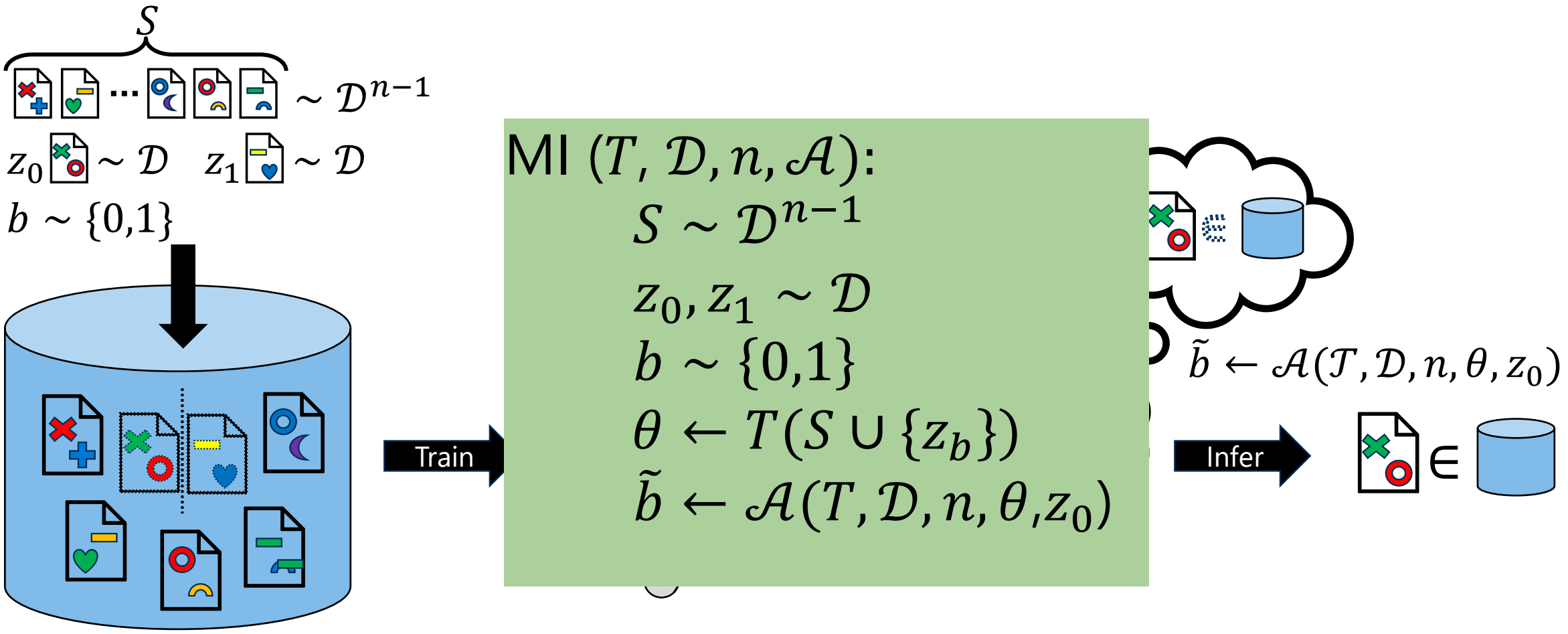
$$\mathrm{Adv}(\mathcal{A}) = 2\left(\Pr[\tilde{b} = b] - \frac{1}{2}\right)$$

Advantage over random guess

Formalizing Membership Inference



Formalizing Membership Inference



Training data | Trained model | Adversary | Information leaked

Overfitting, robustness, and malicious algorithms. Yeom et al. J. of Comp. Sec. 28 (2020) 35–70
https://doi.org/10.3233/JCS-191362

Formalizing Membership Inference

MI $(T, \mathcal{D}, n, \mathcal{A})$:

$S \sim \mathcal{D}^{n-1}$

$z_0, z_1 \sim \mathcal{D}$

$b \sim \{0,1\}$

$\theta \leftarrow T(S \cup \{z_b\})$

$\tilde{b} \leftarrow \mathcal{A}(T, \mathcal{D}, n, \theta, z_0)$

MI $(T, \mathcal{D}, n, \mathcal{A})$

$S \sim \mathcal{D}^{n-1}$

$z_0, z_1 \leftarrow \mathcal{A}(S)$

$b \sim \{0,1\}$

$\theta \leftarrow T(S \cup \{z_b\})$

$\tilde{b} \leftarrow \mathcal{A}(T, \mathrm{S}, \theta, n)$

Overfitting, robustness, and malicious algorithms. Yeom et al. J. of Comp. Sec. 28 (2020) 35–70
https://doi.org/10.3233/JCS-191362

Differential Privacy

A training algorithm $\mathcal{T}$ is (ε, δ)-DP if for any *adjacent* datasets S_0, S_1 and measurable set of models O

$$\mathrm{P}[\mathcal{T}(S_0) \in O] \leq e^{\varepsilon}\, \mathrm{P}[\mathcal{T}(S_1) \in O] + \delta$$

MI $(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$:

S

z

b

θ

$\tilde{b}$

DPD $(\mathcal{T}, n, \mathcal{A})$: **Adversarially chosen**

Theorem: If $\mathcal{T}$ is (ε, δ)-DP, then

$$\mathrm{Adv}_{\mathrm{MI}}(\mathcal{A}) \leq \mathrm{Adv}_{\mathrm{DPD}}(\mathcal{A}) \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}$$

Investigating Membership Inference Attacks under Data Dependencies. Humphries et al. https://arxiv.org/abs/2010.12112

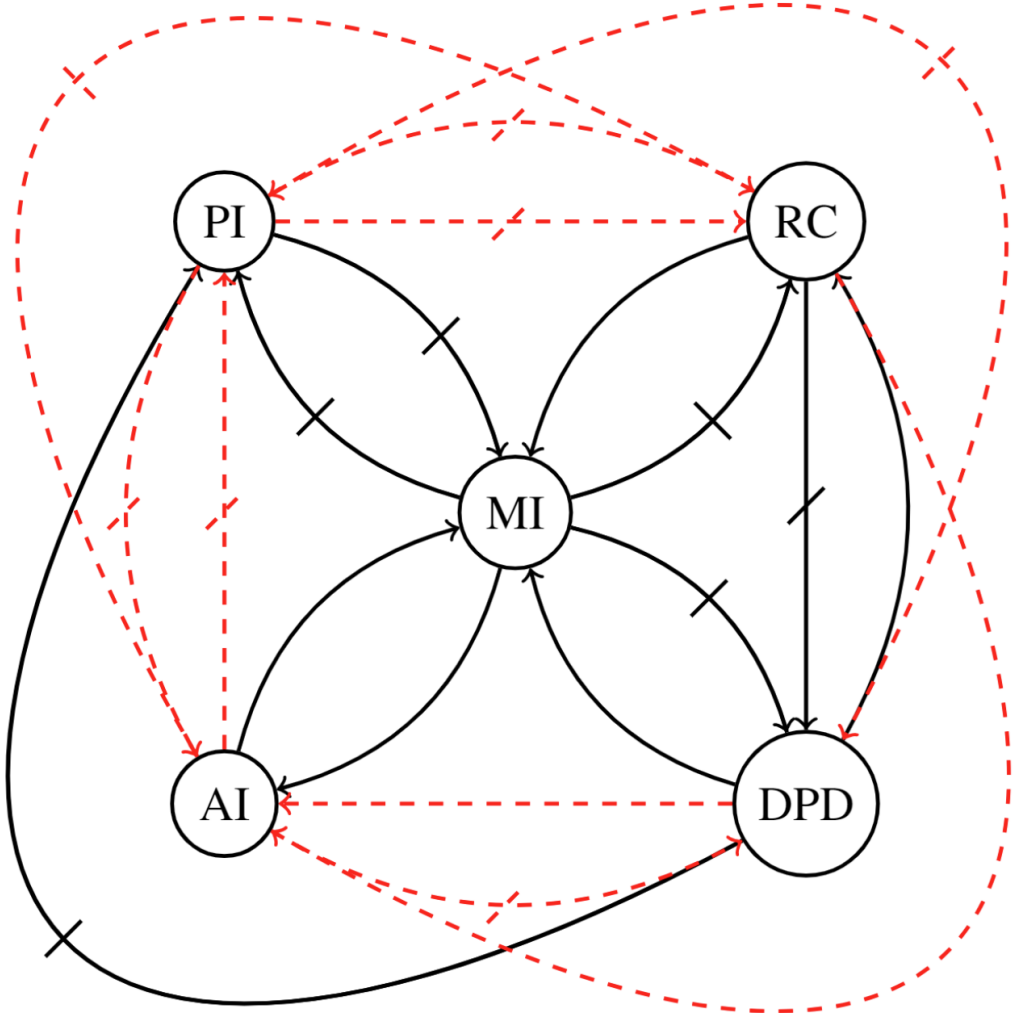
Systematizing Privacy Games

Game	Definition	Adversary Access		Challenge			Training Dataset			Adversary Interest		
		Black-box	White-box	Rand	Adv	Param	Rand	Adv	Param	Record	Subject	Distribution
Membership Inference												
MI	Game 2 [31, 35, 70]	–	✓	✓	–	–	✓	–	–	✓	–	–
MI^{skew}	Game 2 [34]	–	✓	✓	–	–	✓	–	–	✓	–	–
MI^{BB}	Game 2 [12]	✓	–	✓	–	–	✓	–	–	✓	–	–
MI^{Adv}	Game 2 [13]	–	✓	–	✓	–	✓	–	–	✓	–	–
MI^{Diff}	Game 3 [64]	✓	–	✓	–	–	✓	–	–	✓	–	–
MI^{Pois}	Game 3 [64]	✓	–	✓	–	–	✓	✓	–	✓	–	–
MI^{User}	Game 4 [41]	✓	–	–	–	✓	✓	–	–	–	✓	–
MM	Game 11 [31]	–	✓	✓	–	–	✓	–	–	✓	–	✓
MI^{SQ}	Game 17 [61]	✓	–	✓	–	–	–	✓	–	✓	–	–
Attribute Inference and Model Inversion												
AI	Game 5 [70]	–	✓	✓	–	–	✓	–	–	✓	–	–
Inv	Game 5 [67]	–	✓	✓	–	–	✓	–	–	✓	–	–
Data Reconstruction												
RC	Game 6 [4]	–	✓	✓	–	–	–	–	✓	✓	–	–
RC^{Untarg}	Game 7 [11]	✓	–	×	×	×	✓	–	–	✓	–	–
RC^{Targ}	Game 7 [10]	✓	–	✓	–	–	✓	–	–	✓	–	–
Distribution Inference												
PI	Game 8 [58]	–	✓	×	×	×	✓	–	–	–	–	✓
MI^{Subj}	Game 9 [59]	–	✓	✓	–	–	✓	–	–	–	✓	✓
Differential Privacy Distinguishability												
DPD	Game 10 [43, 46]	–	✓	–	✓	–	–	✓	–	✓	–	–
SMI	Game 10 [4, 31]	–	✓	–	–	✓	–	–	✓	✓	–	–

A. Salem, et al., "SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning"
IEEE Symposium on Security and Privacy 2023.

Relations between privacy risks

MI Membership Inference
AI Attribute Inference
DPD DP Distinguishability
PI Property Inference
RC Data Reconstruction
Reduction
Separation
Implied reduction
Implied separation



Games enable reasoning about relations between privacy risks

- Does DP mitigate AI (inferring an unknown sensitive attribute of a target record)? **Yes, and we can quantify how much**
- Does DP mitigate PI (e.g., inferring the proportion of records with an attribute value)? **No**

Games for new settings

Game 1: Membership Inference

Input: $\mathcal{A}, \theta, \mathcal{T}, \mathcal{D}, n$

1 $S \sim \mathcal{D}^{n-1}$ ▷ sample $n-1$ i.i.d. examples from distribution $\mathcal{D}$

2 $z_0, z_1 \sim \mathcal{D}$ ▷ sample 2 more *challenge* examples from $\mathcal{D}$

3 $b \sim \{0, 1\}$ ▷ flip a fair coin

4 $\theta' \leftarrow \mathcal{T}(\theta; S \cup \{z_b\})$ ▷ Fine-tune θ on $S \cup \{z_b\}$

5 $\hat{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, \theta', z_0)$ ▷ Adversary $\mathcal{A}$ guesses whether z_0 is a member

Game 2: Prompt Leaking

Input: $\mathcal{A}, \theta, \mathcal{P}, \mathcal{D}, n$

1 $S \sim \mathcal{D}^n$ ▷ Sample n examples from distribution $\mathcal{D}$

2 $p \sim \mathcal{P}(\theta; S)$ ▷ Tune a prompt p for θ on S

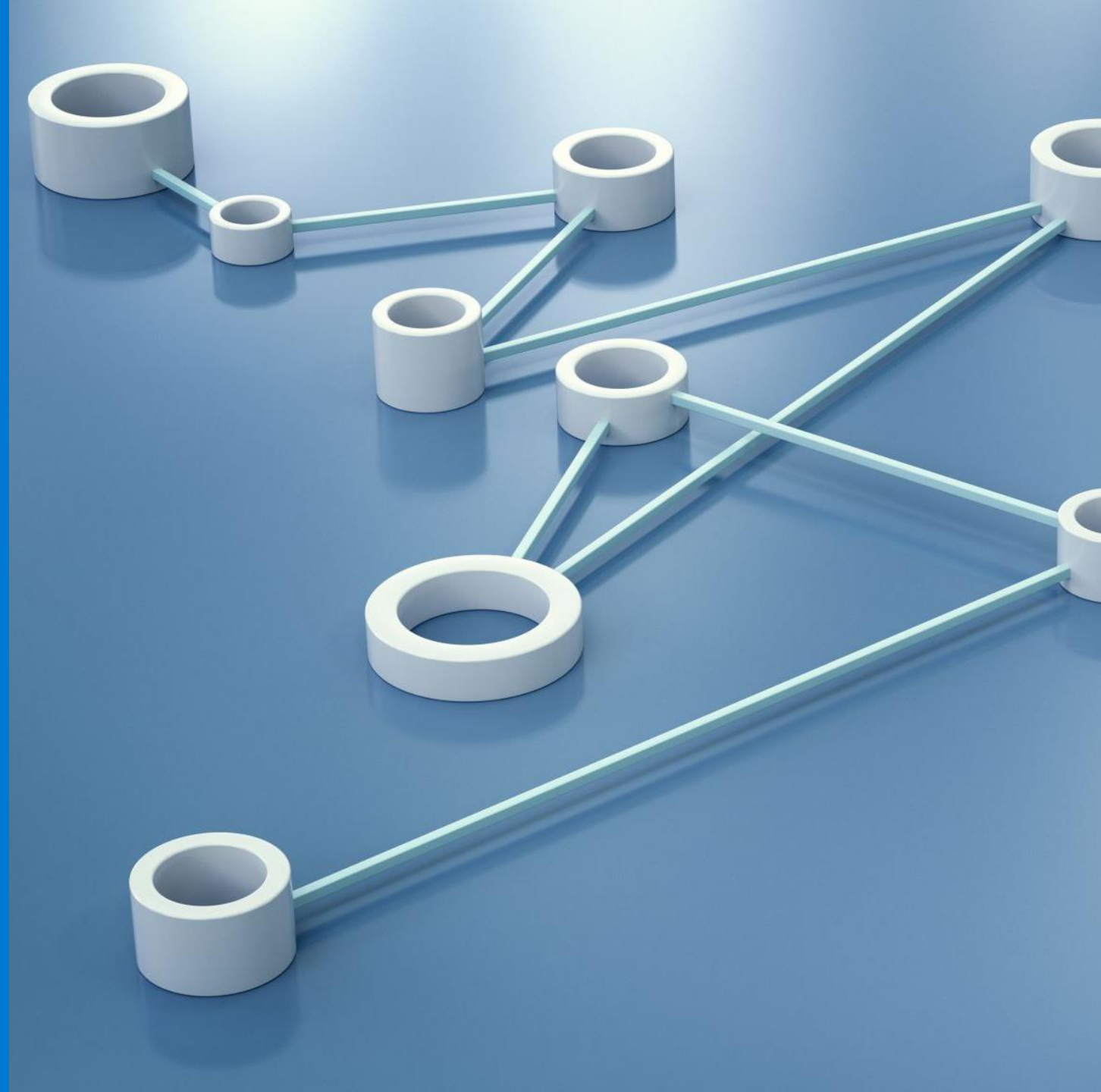
3 $\hat{p} \leftarrow \mathcal{A}(\mathcal{O}(\theta, p, \cdot))$ ▷ Ask adversary $\mathcal{A}$ to reconstruct p

4 **Oracle** $\mathcal{O}(\theta, p, x)$: **return** $\theta(p; x)$ ▷ API access to θ with fixed prompt

Microsoft Security

Empirical evaluation of privacy mechanisms

Joint work with Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Mohammad Naseri, Boris Köpf, Daniel Jones



Differential Privacy (DP)

A mechanism $\mathcal{M}$ is (ε, δ)-Differentially Private if for every pair of adjacent datasets D, D' and any set of outcomes $\mathcal{S}$

$$\mathbb{P}[\mathcal{M}(D)\ \epsilon\ \mathcal{S}] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D')\ \epsilon\ \mathcal{S}] + \delta$$

How to achieve DP in practice?

- Compute *sensitivity* of the function i.e., $\max_{D,D'} |\mathcal{M}(D) - \mathcal{M}(D')|$
- Add noise calibrated to the sensitivity

C. Dwork & A. Roth, "The Algorithmic Foundations of Differential Privacy"

Differentially-Private Stochastic Gradient Descent

Intuition: limit the contribution of any individual training point, and add calibrated noise during training

Implementations:

- Tensorflow: tensorflow/privacy
- Pytorch: Opacus

Limitations:

- Privacy vs. utility trade-off
- Well-suited for huge datasets, large batch sizes
- Computation is expensive

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

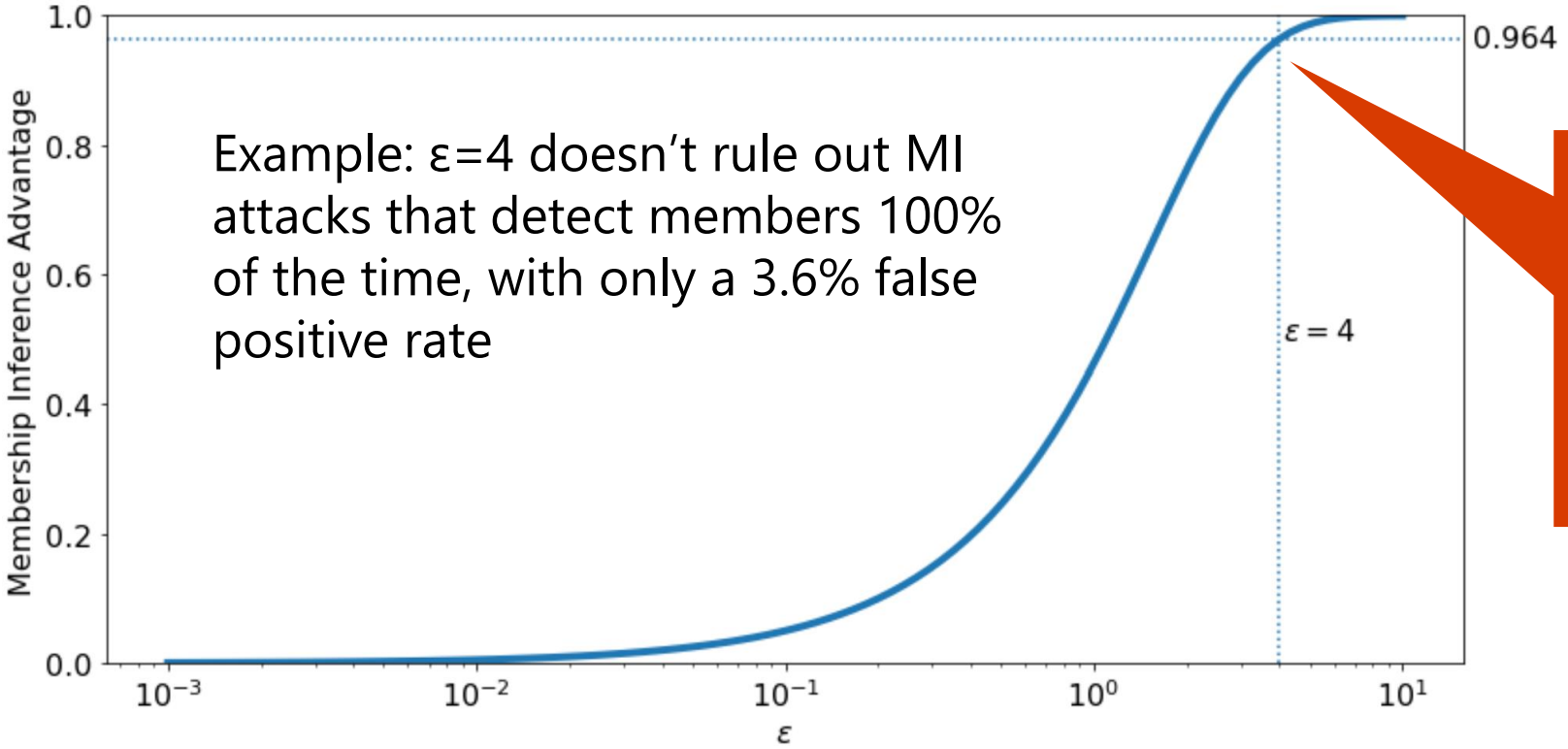
Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

M. Abadi, et al., "Deep Learning with Differential Privacy" ACM Conference on Computer and Communications Security 2016.

What does DP-SGD *guarantee*?

Membership inference: Given a model and a sample *z*, adversary wants to determine whether *z* was used to train the model

Membership Inference Advantage ∈ [0,1] = True Positive Rate – False Positive Rate



Example: ε=4 doesn't rule out MI attacks that detect members 100% of the time, with only a 3.6% false positive rate

- ε=4 is often seen as reasonable, but only gives weak protection against the weakest of attacks
- Using a lower ε would harm utility too much

T. Humphries, et al., "Investigating Membership Inference Attacks under Data Dependencies" IEEE Computer Security Foundations Symposium 2023.

What does DP-SGD *guarantee?*

Analysis of DP-SGD assumes the adversary:

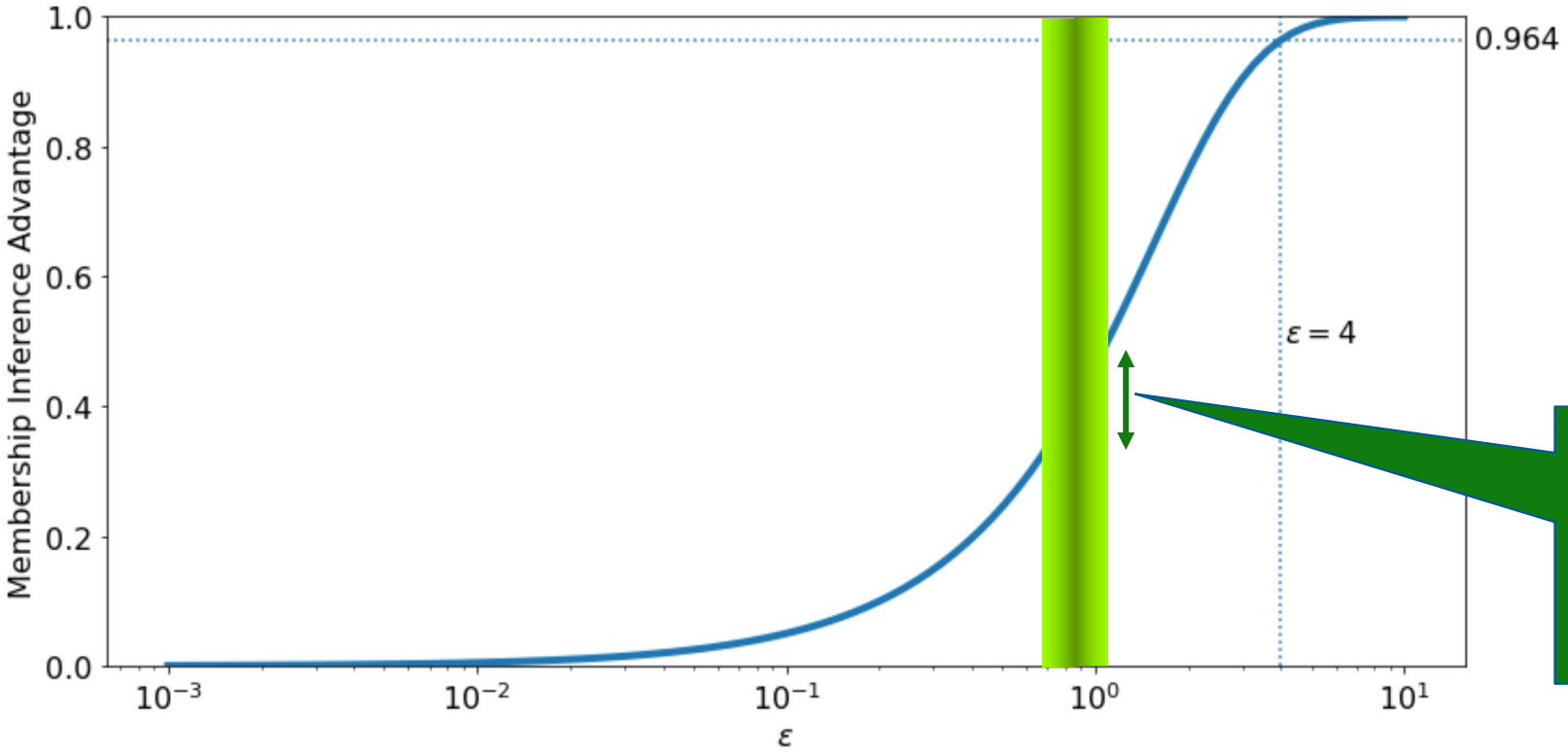
- can craft a worst-case dataset i.e., has full control over all gradients (except the order in which they're seen)
- has access to each model update (may only be realistic in FL)

The DP Adversary is unrealistically powerful for many practical scenarios!

What does DP-SGD *provide empirically*?

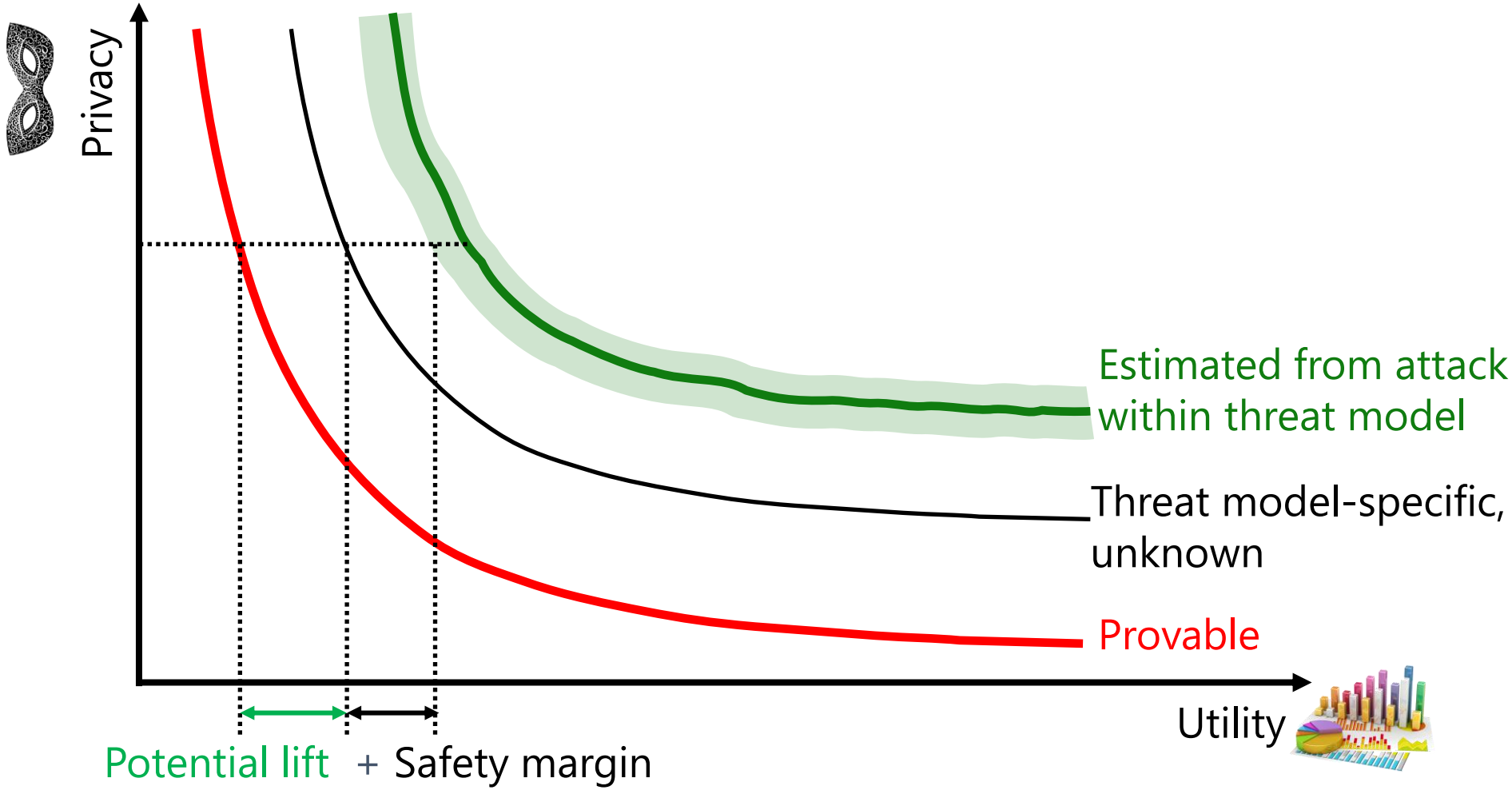
Membership inference: Given a model and a sample *z*, adversary wants to determine whether *z* was used to train the model

Membership Inference Advantage ∈ [0,1] = True Positive Rate – False Positive Rate



- Even when DP-SGD guarantees ε=4, the actual protection could be better
- How can we measure this?

Gap between Provable and Empirical Privacy



How to empirically measure DP-SGD?

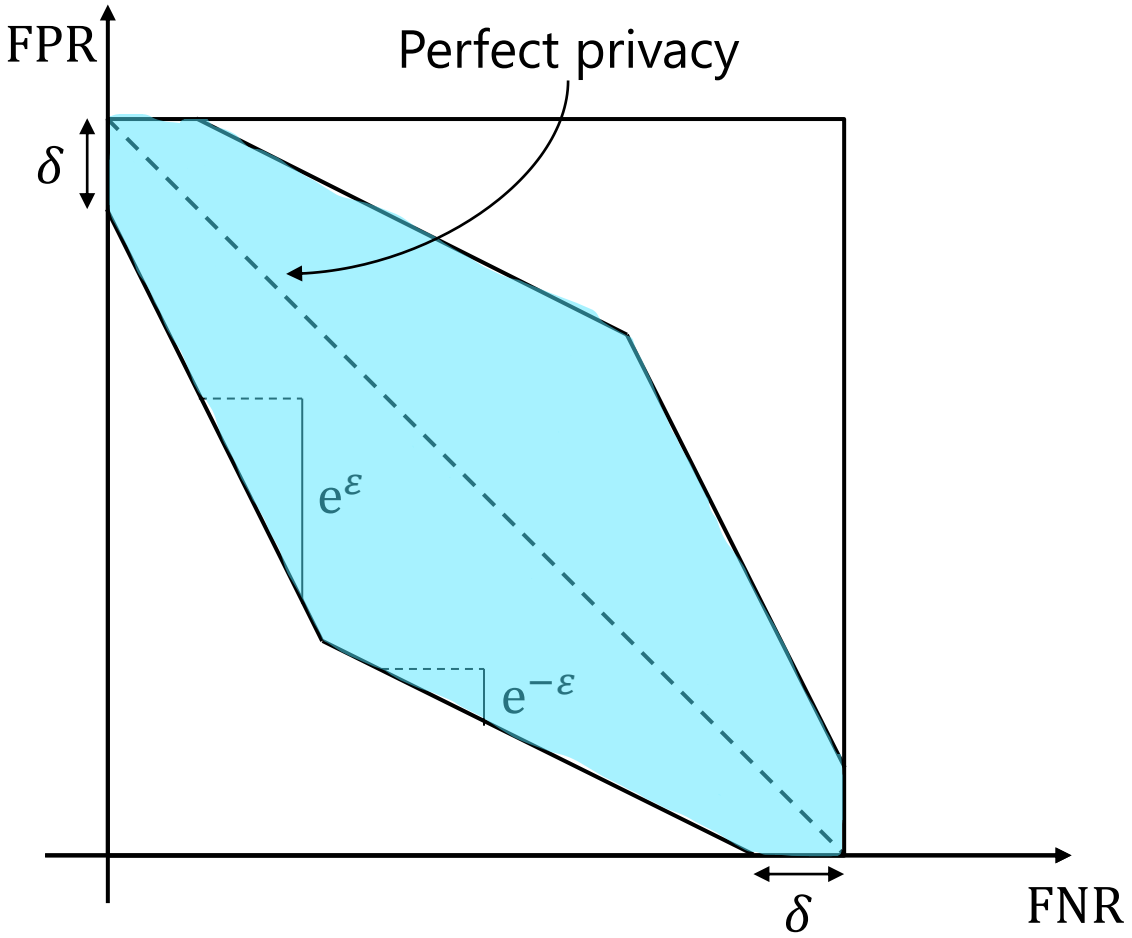
Empirical estimation of DP ε

- M. Jagielski, et al., "Auditing Differentially Private Machine Learning: How Private is Private SGD?", June 2020
- M. Nasr, et al., "Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning", January 2021
- F. Tramèr, et al., "Debugging Differential Privacy: A Case Study for Privacy Auditing", February 2022
- S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", June 2022
- F. Lu, et al., "A General Framework for Auditing Differentially Private Machine Learning", October 2022
- M. Nasr, et al., "Tight Auditing of Differentially Private Machine Learning", February 2023
- T. Steinke, et al., "Privacy Auditing with One (1) Training Run", May 2023
- K. Pillutla, et al., "Unleashing the Power of Randomization in Auditing Differentially Private ML", May 2023

Empirical estimation of DP ε

- **Goal:** estimate an *empirical distribution* of ε for a given δ
- **Approach:** use membership inference attacks to obtain estimates for the lower bound on ε.
- **Use cases:**
  - Provide meaningful privacy protection *and* improved utility
  - Audit training pipeline to detect violations of DP (due to unexpected data correlations, implementation bugs, etc.)
  - …

DP as Hypothesis Testing



(ε, δ)-DP $\Leftrightarrow$ Attacks only possible in the shaded region

P. Kairouz, et al., "The Composition Theorem for Differential Privacy" International Conference on Machine Learning 2015.

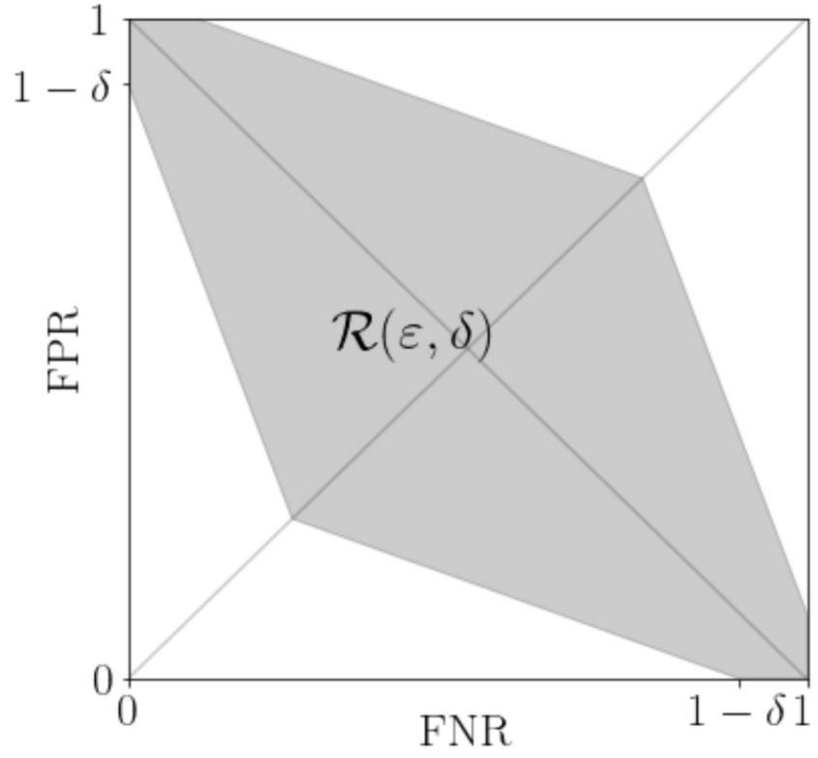
Empirical estimation of DP ε

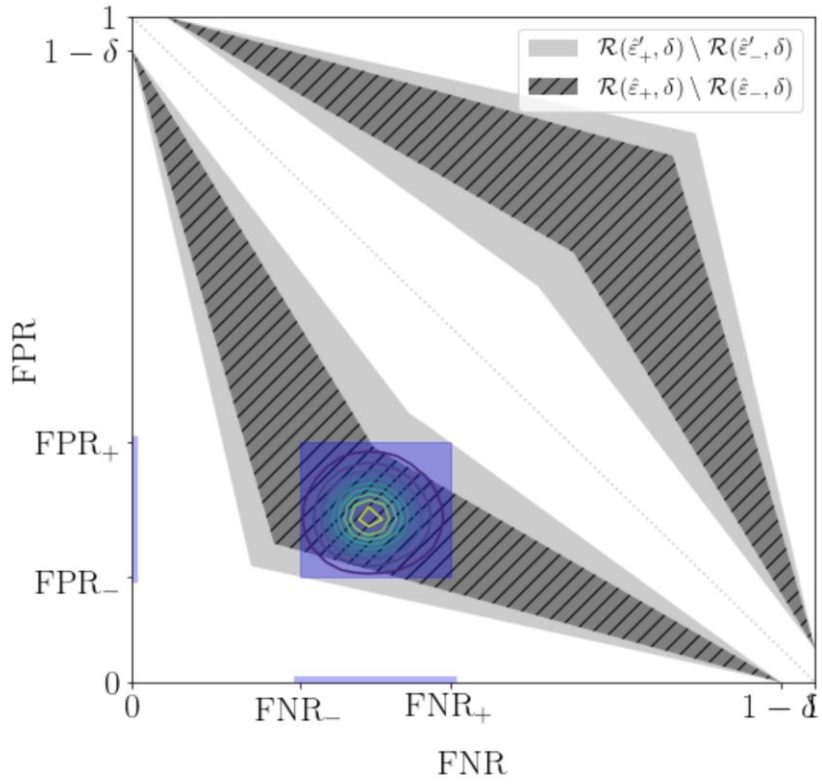
- **Challenge:** how to obtain good approximations of the adversary's False Positive Rate (FPR) and False Negative Rate (FNR)?

- **Design space:**
  - How is the dataset chosen/created (e.g., average-case vs. worst-case)?
  - How are the target points chosen/constructed (e.g., natural vs. "canaries")?
  - What type of attack to use (e.g., black-box vs. white-box)?
  - What type of confidence intervals to use?
  - How many models to train?

Bayesian estimation of DP ε

1. Select a specific membership inference attack
2. Repeat the attack to estimate the joint distribution of FPR, FNR
3. Integrate density of distribution over privacy regions to find a credible interval for ε @ given δ i.e., find two privacy regions whose difference covers 95% of the density of (FPR,FNR)

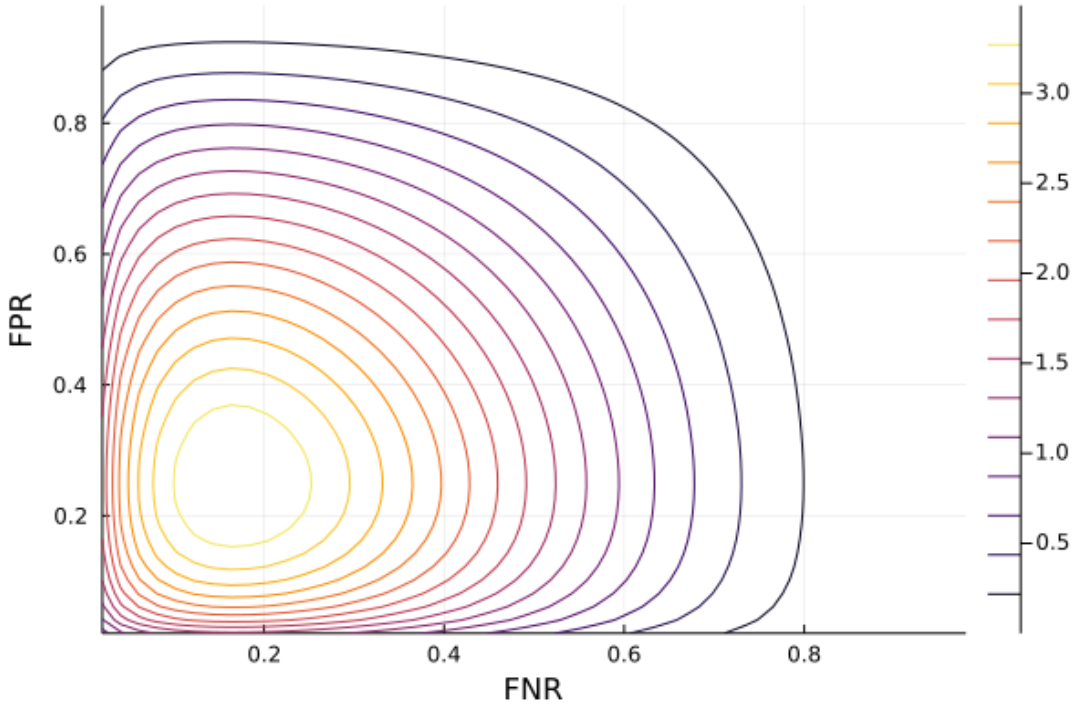


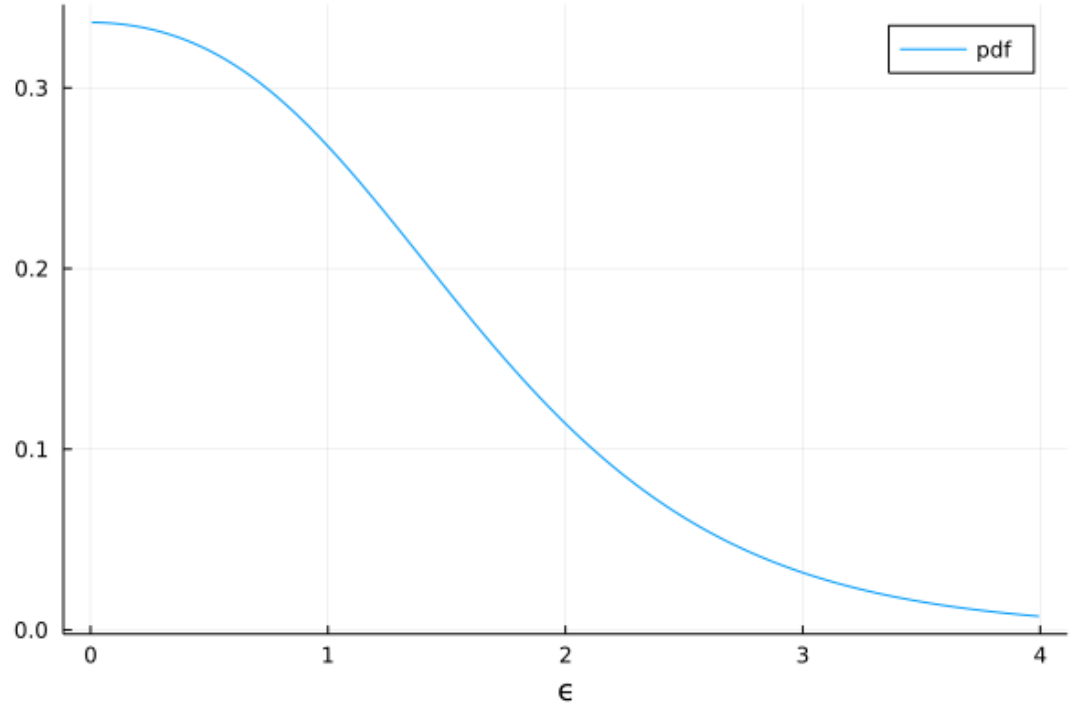


S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Bayesian estimation of DP ε

Attack confusion matrix: FN = 1, FP = 1, TP = 3, TN = 2

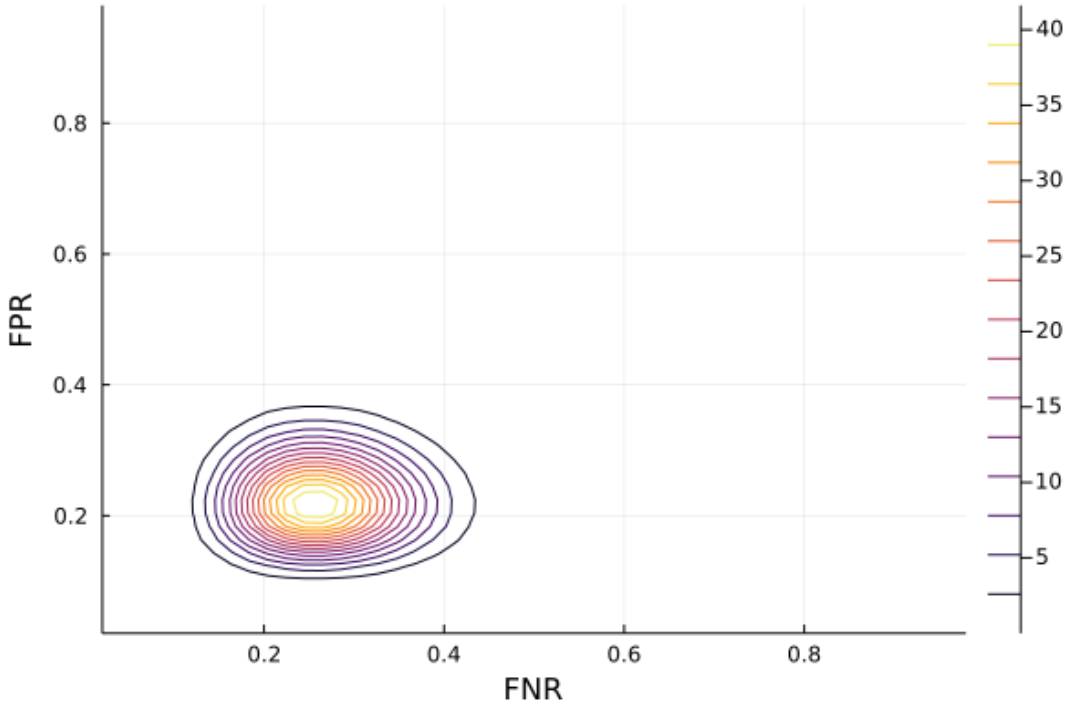


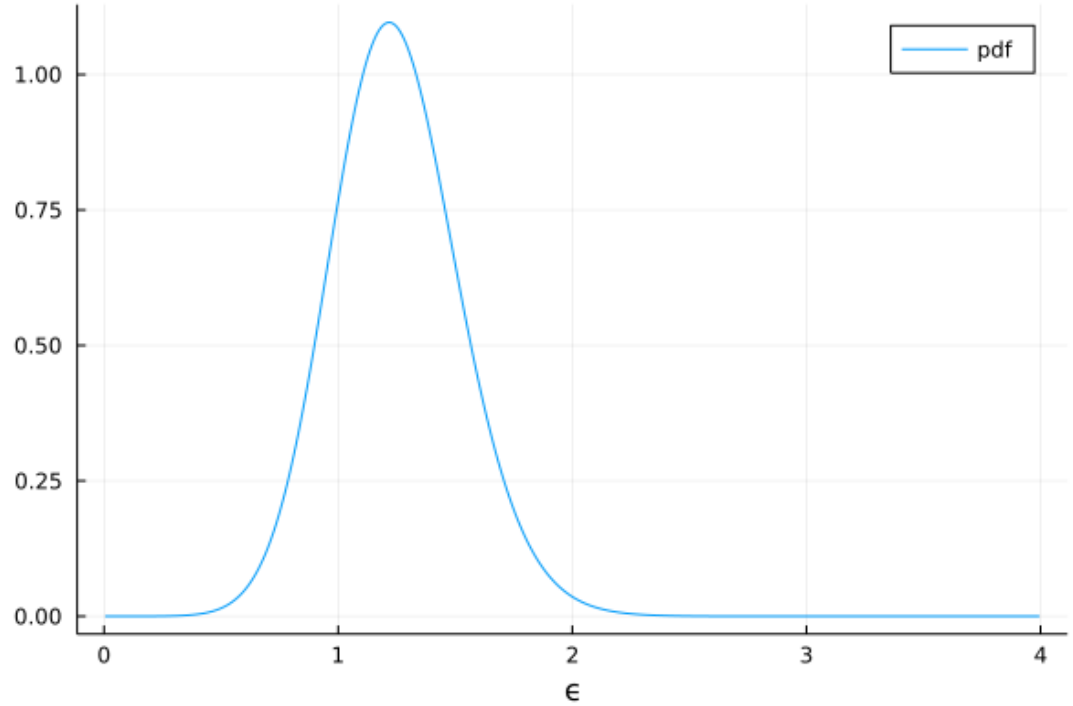


S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Bayesian estimation of DP ε

Attack confusion matrix: FN = 11, FP = 12, TP = 31, TN = 42

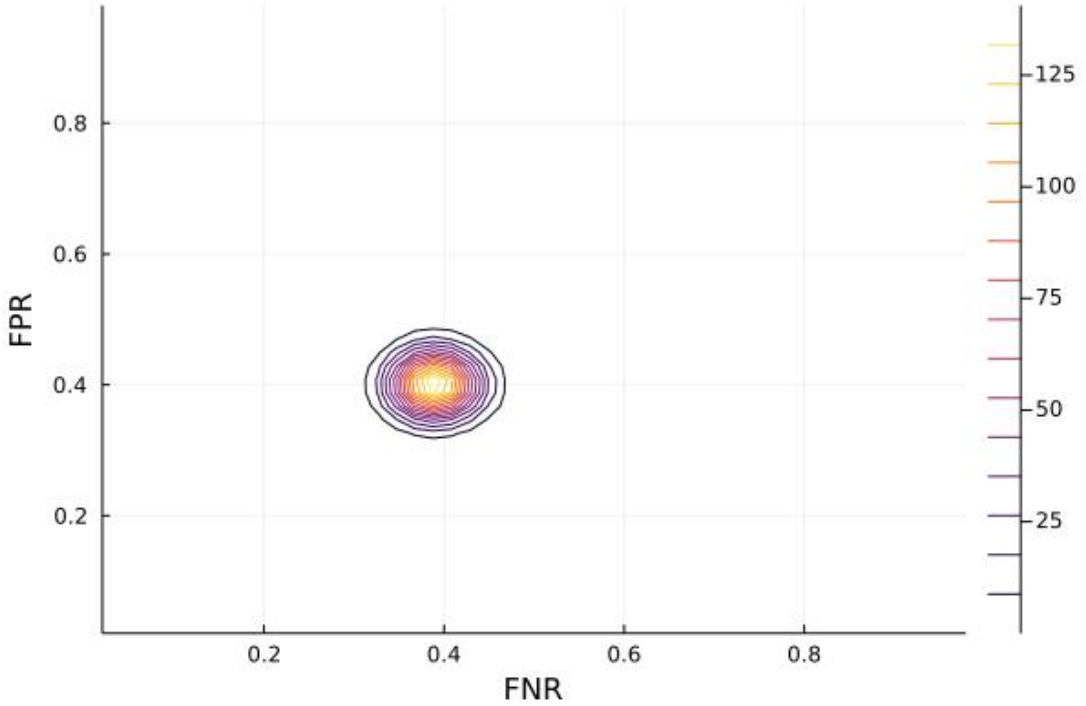


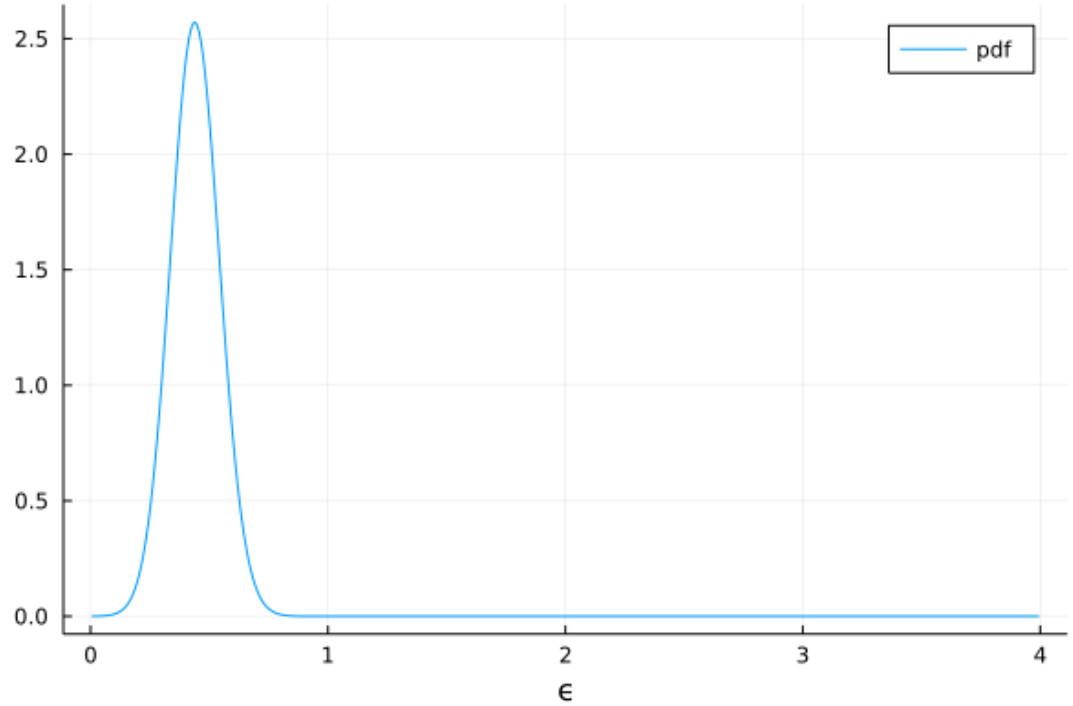


S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Bayesian estimation of DP ε

Attack confusion matrix: FN = 90, FP = 81, TP = 141, TN = 121

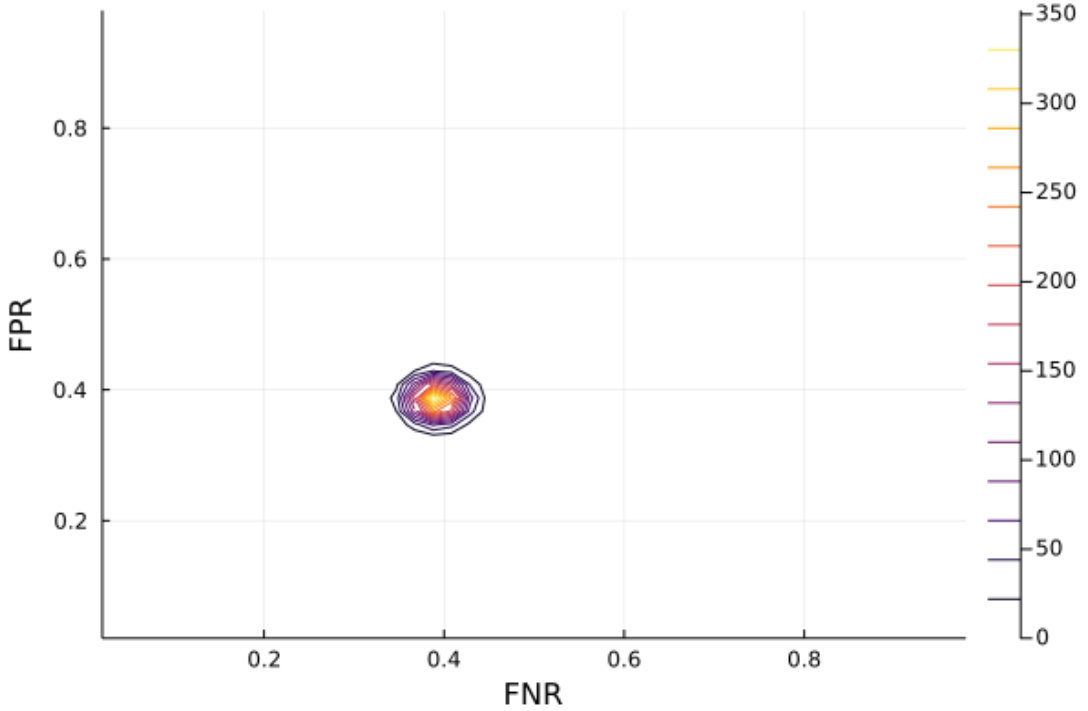


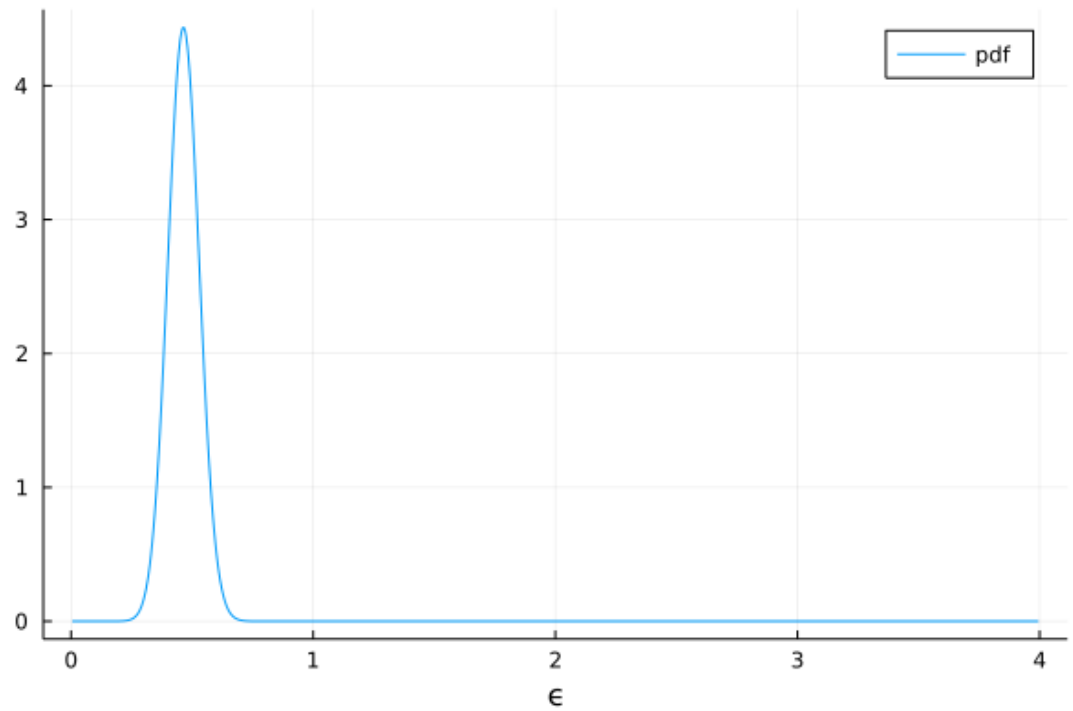


S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Bayesian estimation of DP ε

Attack confusion matrix: FN = 220, FP = 201, TP = 341, TN = 321

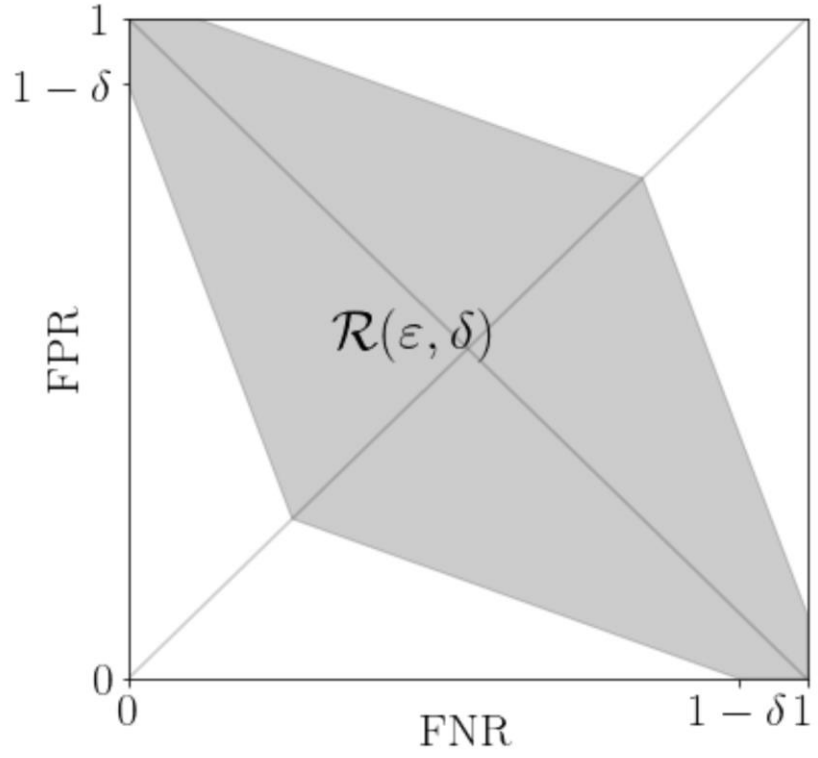


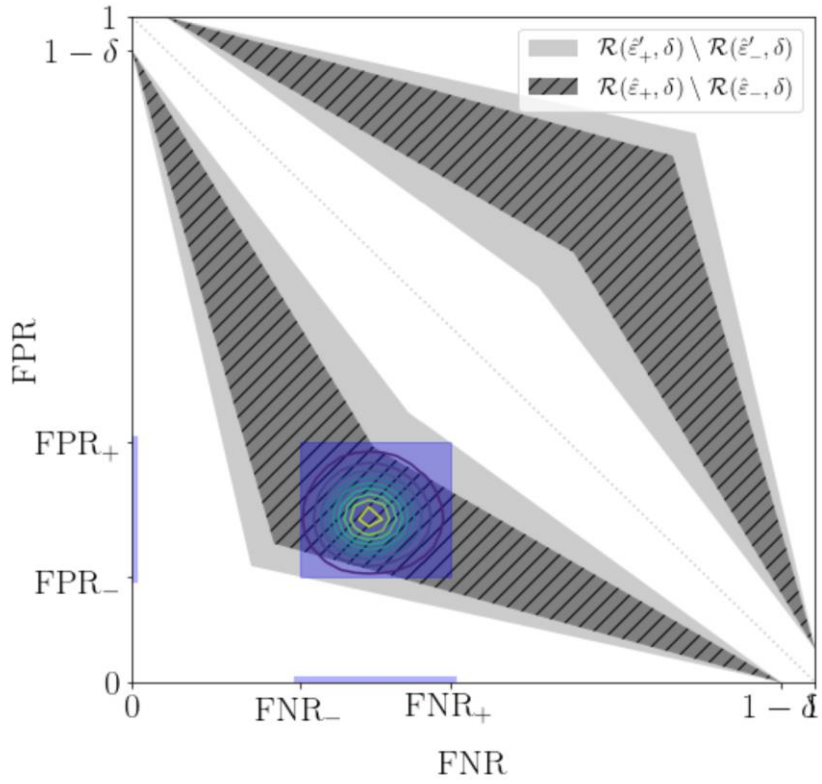


S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Bayesian estimation of DP ε

1. Select a specific membership inference attack
2. Repeat the attack to estimate the joint distribution of FPR, FNR
3. Integrate density of distribution over privacy regions to find a credible interval for ε @ given δ i.e., find two privacy regions whose difference covers 95% of the density of (FPR,FNR)





S. Zanella-Béguelin, et al., "Bayesian Estimation of Differential Privacy", International Conference on Machine Learning 2023.

Privacy Auditing with One (1) Training Run

*"Can we perform privacy auditing using a **single run** of the algorithm M?"*

- Identify m data points (i.e., training examples or "canaries");
- Flip m independent unbiased coins to decide which to include or exclude;
- Run the algorithm on the randomly selected dataset;
- Based on the output, the auditor "guesses" whether each data point was included or excluded (or it can abstain);
- Obtain a lower bound on the privacy parameters from the fraction of guesses that were correct.

T. Steinke, M. Nasr, & M. Jagielski, "Privacy Auditing with One (1) Training Run", arXiv:2305.08846, May 2023.

Privacy Auditing with One (1) Training Run

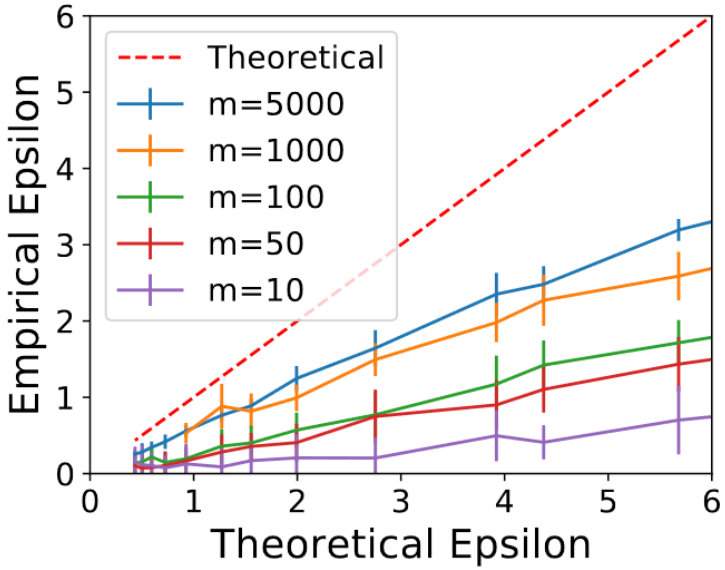


Figure 4: Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

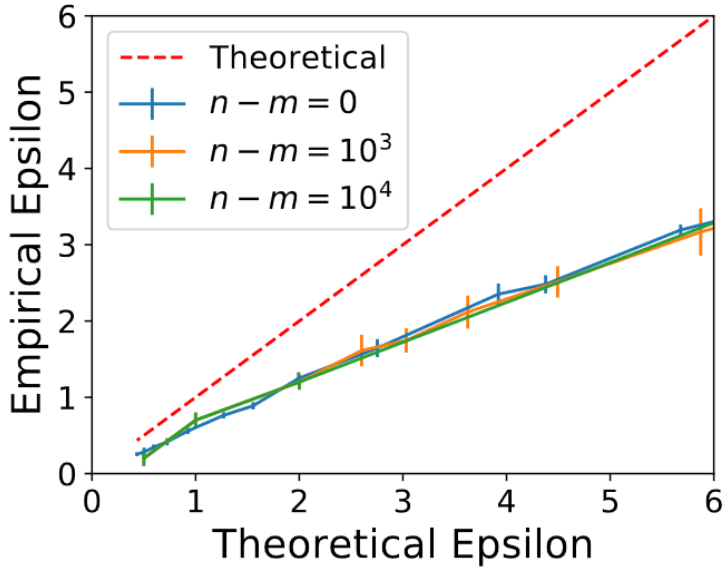


Figure 5: Effect of the number of additional examples $(n-m)$ in the white-box setting. Importantly, adding additional examples does not impact the auditing results in the white-box setting.

T. Steinke, M. Nasr, & M. Jagielski, "Privacy Auditing with One (1) Training Run" , arXiv:2305.08846, May 2023.

Empirical estimation of DP ε

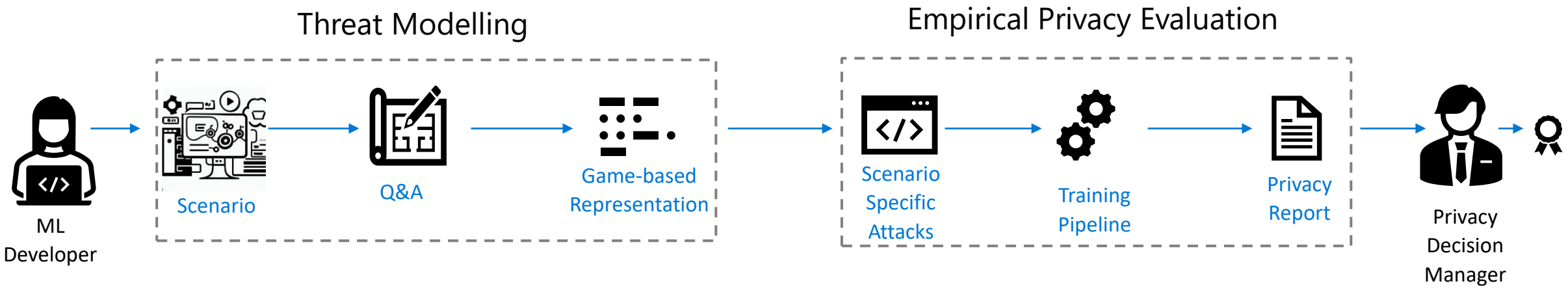
- **Challenge:** how to obtain good approximations of the adversary's False Positive Rate (FPR) and False Negative Rate (FNR)?

- **Design space:**
  - How is the dataset chosen/created (e.g., average-case vs. worst-case)?
  - How are the target points chosen/constructed (e.g., natural vs. "canaries")?
  - What type of attack to use (e.g., black-box vs. white-box)?
  - What type of confidence intervals to use?
  - How many models to train?

Microsoft Security

Putting it all together

A new privacy review process



Summary

Information leakage is a concern when training on non-public data.

Privacy games

- a systematic approach for **describing** inference risks in ML.
- enable **comparisons** between risks using established and implied relationships.

Empirical estimates

- provide *lower bounds* for ε and can be used to **audit** privacy mechanisms.
- can be computed from even a **single** training run.

Microsoft Security

Thank you

andrew.paverd@microsoft.com

MICO: Microsoft ML Membership Inference Competition

Microsoft | MSRC Report an issue Customer guidance Engage Who we are Blogs Acknowledgments

Blog / 2022 / 11 / Mico /

Announcing the Microsoft Machine Learning Membership Inference Competition (MICO)

MSRC, Security Research & Defense / By MSRC / November 16, 2022 / 6 min read

We're excited to announce the launch of a new competition focusing on the security and privacy of machine learning (ML) systems. Machine learning has already become a key enabler in many products and services, and this trend is likely to continue. It is therefore critical to understand the security and privacy guarantees provided by state-of-the-art ML algorithms – indeed this is one of Microsoft's Responsible AI Principles.

Fundamentally, ML models need data on which they can be trained. This training data can be drawn from a variety of sources, including both public and non-public data. In many domains, ML models achieve better performance if they are trained on specialized or domain-specific data. This specialized data is often not directly available to the users of the model (e.g. to protect privacy of the data contributors or intellectual property of the model owner). Ideally, having access to an ML model should not reveal which individual data records were used for training the model. However, recent work on membership inference has demonstrated that this is not always the case.

https://github.com/microsoft/MICO