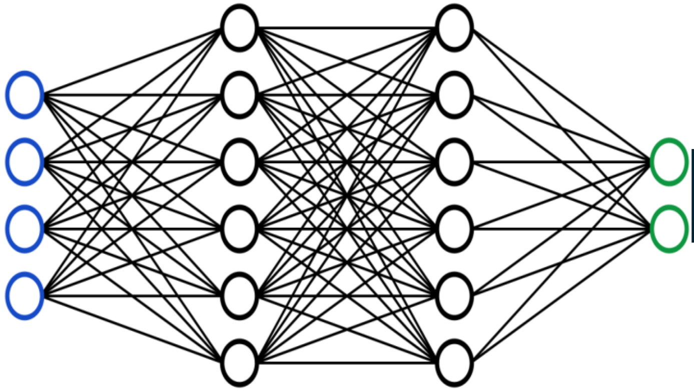
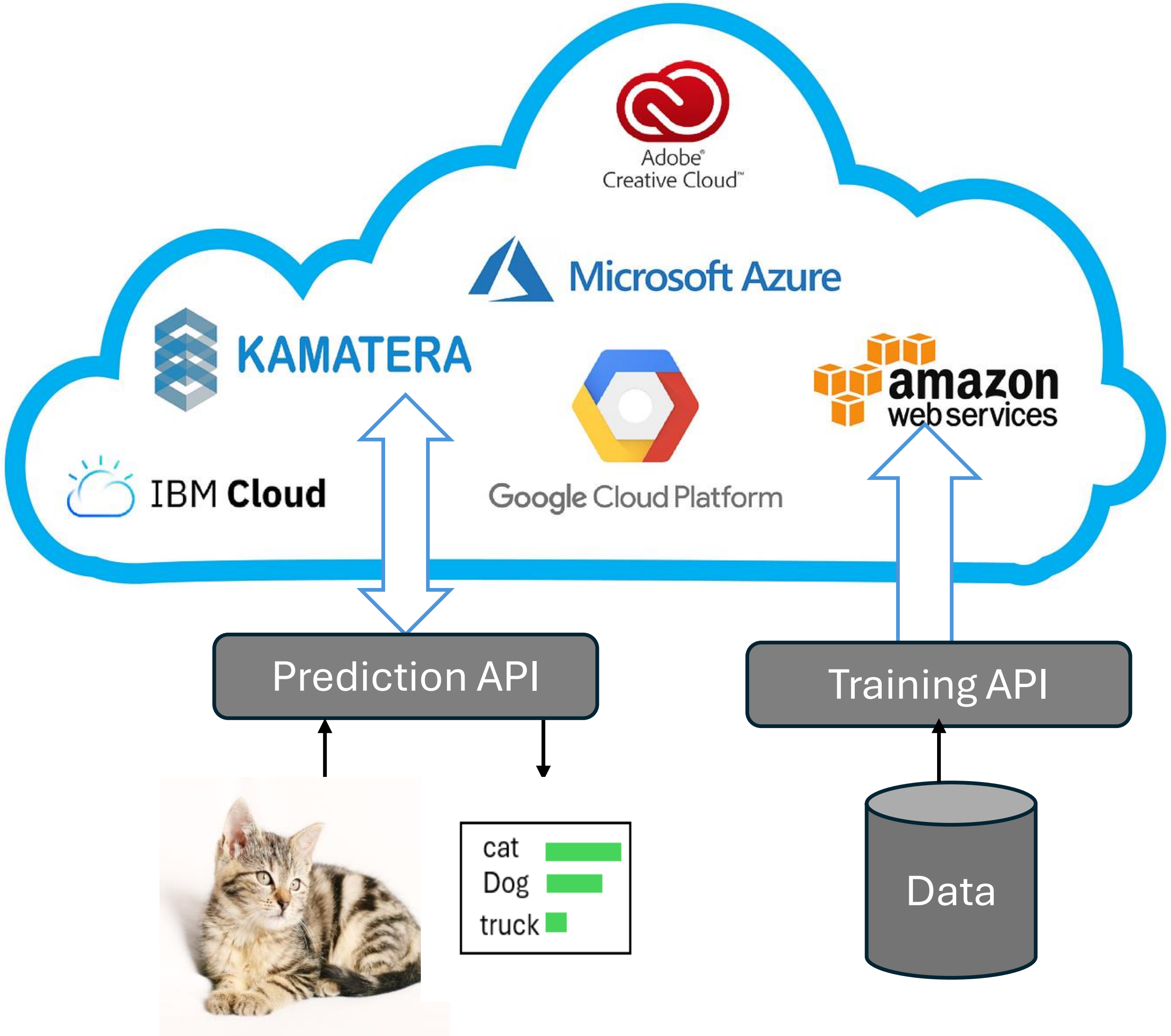
Inf2Guard: An Information-Theoretic Framework for Learning Privacy-Preserving Representations against Inference Attacks

Sayedeh Leila Noorbakhsh[1,*], *Binghui Zhang*[1,*], *Yuan Hong*[2], *Binghui Wang*[1]

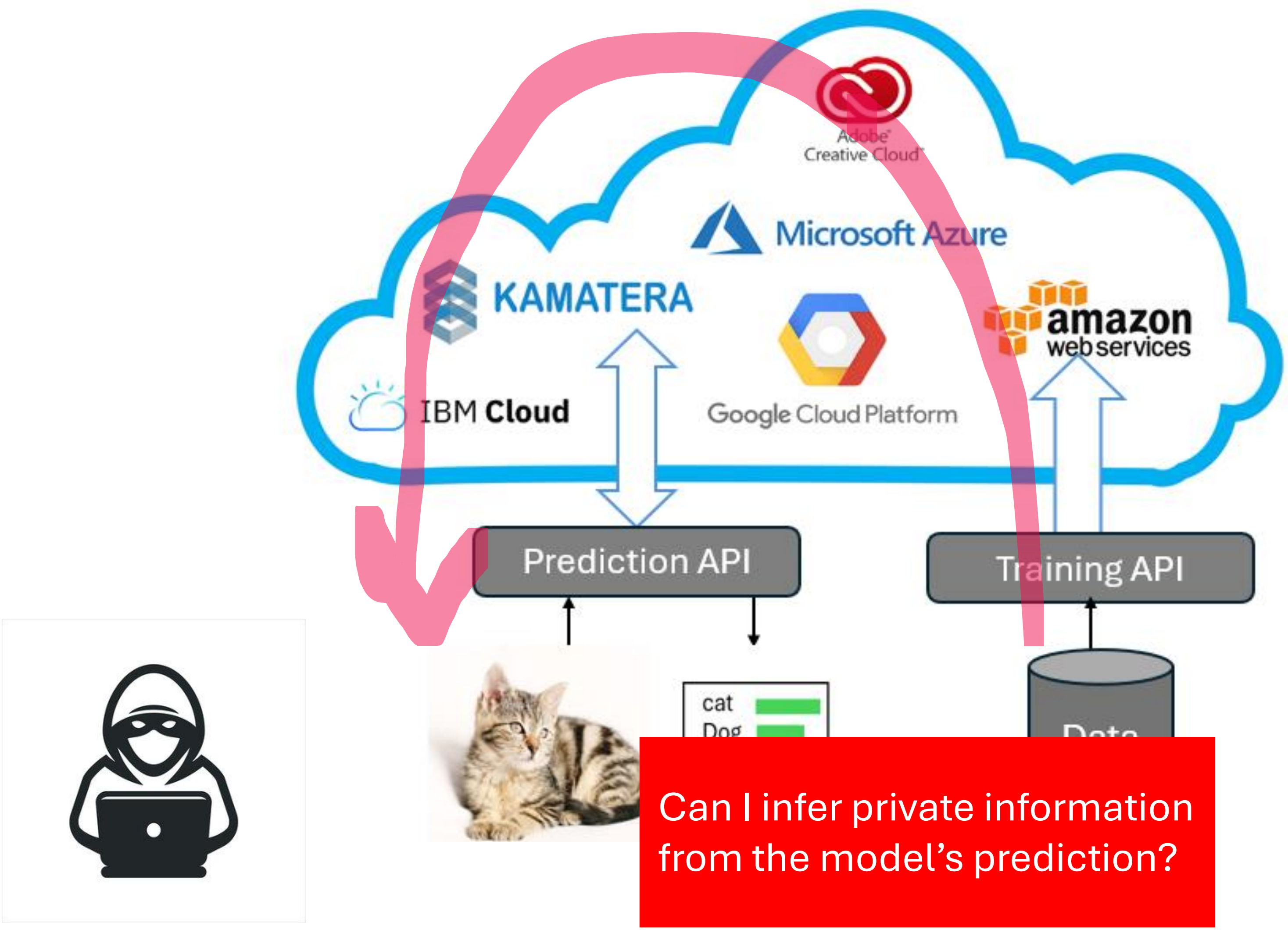
Machine learning



Machine learning As a service

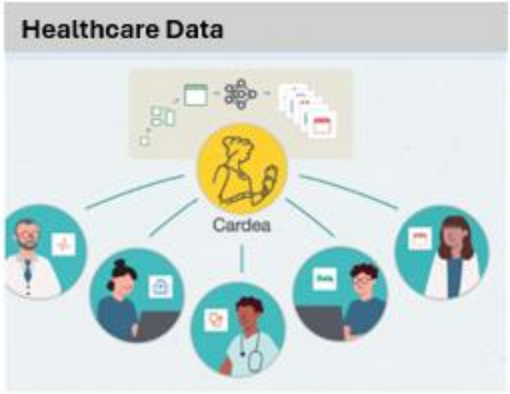


Hazard: Privacy leakage

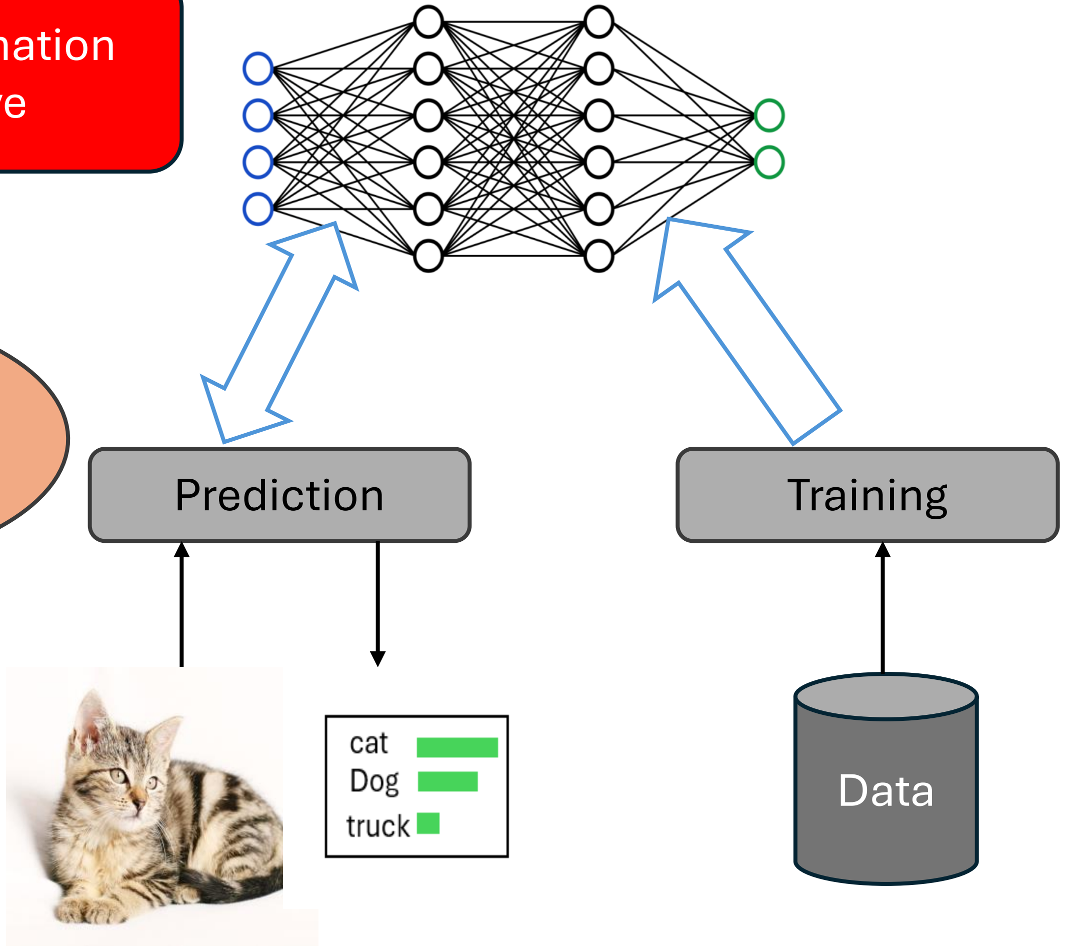


Can I infer private information from the model's prediction?

Membership Inference Attacks (MIA)

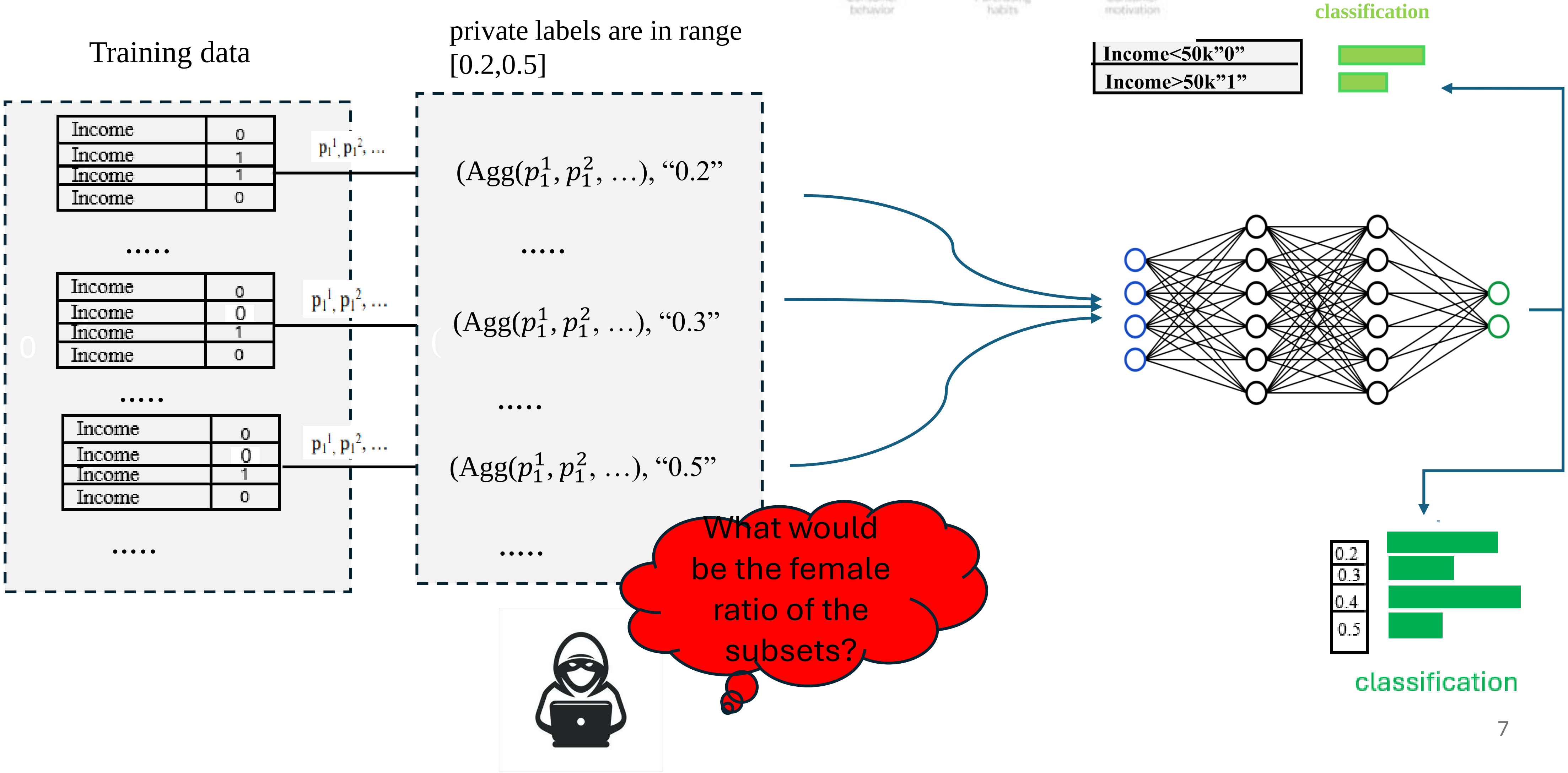


Membership information can be sensitive

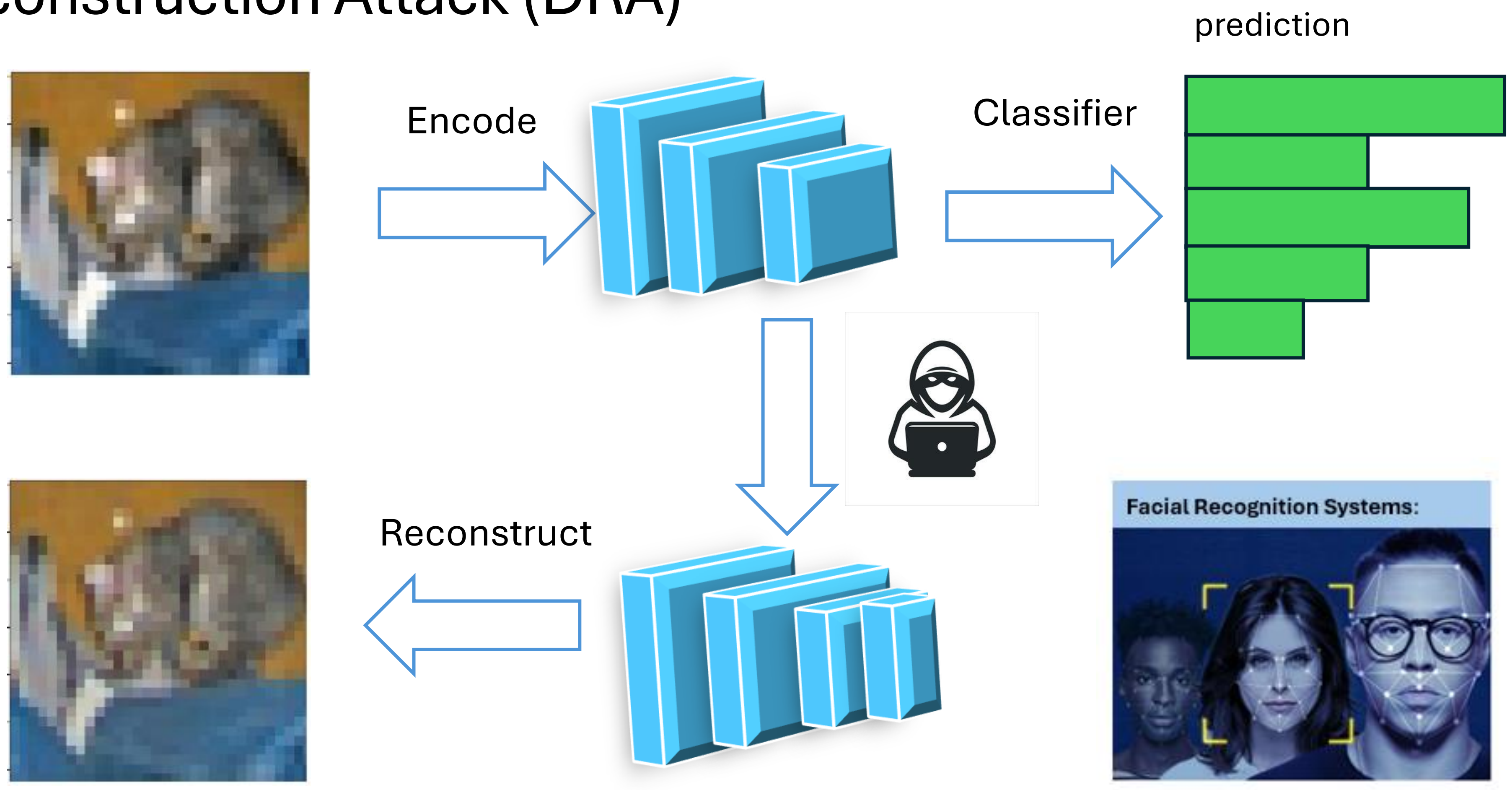


Was this data point in the training set of the model?

Property Inference Attack(PIA)



Data Reconstruction Attack (DRA)



Defense against inference attacks using Inf2Guard

- Can we design a unified privacy protection framework against these inference attacks, MIA, PIA and DRA, that also **maintain utility?**

- Under the framework, can we further **theoretically understand** the **utility-privacy tradeoff** and the **privacy leakage against the inference attacks?**

Threat Model

Defender objective:
- Learning data representations that are resistant to inference attacks

Attacker objective:
- seeks to extract sensitive information from r

Attacker knowledge:
- Knowledge of Data Distribution
- No Access to Internal Encoder

Data

defender(Data owner/ Trusted Service Provider) Encoder f_θ

Rep,r

Attacker

Task learner (Data owner)

Task learner objective:
- develop an accurate classification model

Inf2Guard

- Inf2Guard is inspired by **information theory** and designs customized mutual information (MI) objectives for each inference attack.
- Goal 1: Privacy protection
- Goal 2: Utility preservation.

Introduction to Mutual Information (MI)

- It measures the **amount of information** that one random variable X provides about another random variable Y.
- **MI quantifies** the reduction in uncertainty about one variable due to the knowledge of the other.
- **Mathematically**, MI is expressed as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log^{\left(\frac{p(x,y)}{p(x)p(y)}\right)}$$

- **Applications:** Used in feature selection, clustering, **privacy-preserving mechanisms and inference attack defenses.**

How to defense against this MIA?

- Decrease the utility
  - DP-SGD
  - DP-Encoder

- **Does not** have privacy guarantees
  - AdvReg
  - NeuGuard
  - …

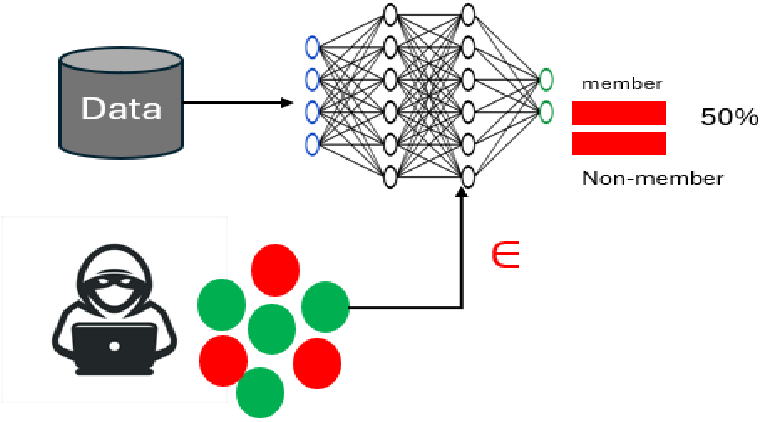
defender learns the representation r = f(x) private membership u

Inf2Guard against MIAs

- Goal 1: Membership protection

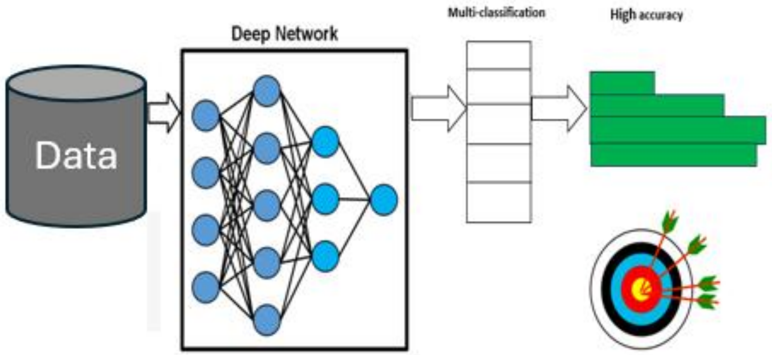
$$\min_f I(\mathbf{r}; u),$$

Very Challenging to solve these two MI objectives.
Calculating an MI between arbitrary variables is infeasible.



- Goal 2: Utility preservation

$$\max_f I(y; \mathbf{r} | u = 1),$$



How to address intractable MI calculation?

- Inspired by the **MI neural estimation,** which transfers the intractable MI calculations to the **tractable variational MI bounds.**

- Capable of parameterizing each bound with **a (deep) neural network.**

- Train neural networks to approximate the **true MI** and **learn representations** against the inference attacks.

Estimating MI via tractable bounds

$[q_\psi(u|r)]$ is an auxiliary posterior distribution of $p(u|r)$

- Minimizing the upper bound MI in Equation

$$I(r;u) \le I_{vCLUB}(r;u) = E_{p(r;u)}[log q_\psi(u|r)] - E_{p(r)p(u)}[log q_\psi(u|r)]$$

$$min_\Psi E_{p(r;u)}[log (u|r)] - E_{p(r)p(u)}[log q_\psi(u|r)]$$

$$\Leftrightarrow max_\Psi E_{p(r;u)}[log q_\psi(u|r)]$$

- Goal 1: privacy protection as a min-max objective function:

$E_{p(r;u)}[log q_\psi(u|r)]$ is irrelevant to Ψ.

$$min_f \ min_\Psi I_{vCLUB}(r;u) \Leftrightarrow \boldsymbol{min_f \ max_\Psi E_{p(r;u)}[log q_\psi(u|r)]}$$

Adversarial game between **an adversary $\boldsymbol{q_\psi}$**) and **encoder f**

Estimating MI via tractable bounds

q_Ω is an arbitrary auxiliary posterior distribution. Predict the training data label y from the representation r

- Maximizing the lower bound MI in Equation

$$I(y;r|u=1) = H(y|u=1) - H(y|r,u=1)$$
$$= H(y|u=1) + E_{p(y,r,u)}\left[log^{q}\Omega(y|r,u=1)\right]$$
$$\geq H(y|u=1) + E_{p(y,r,u)}\left[log^{q}\Omega(y|r,u=1)\right]$$

- Goal 2: utility preservation can be rewritten as max-max objective function:

$$max_f I(y;r|u=1) \Leftrightarrow max_f\ max_\Omega\ E_{p(y,r,u)}\left[log^{q}\Omega(y|r,u=1)\right]$$

Cooperative game between the **encoder f** and **qΩ**

Objective function of Inf2Guard against MIAs.

$\lambda \in [0,1]$ tradeoffs privacy and utility

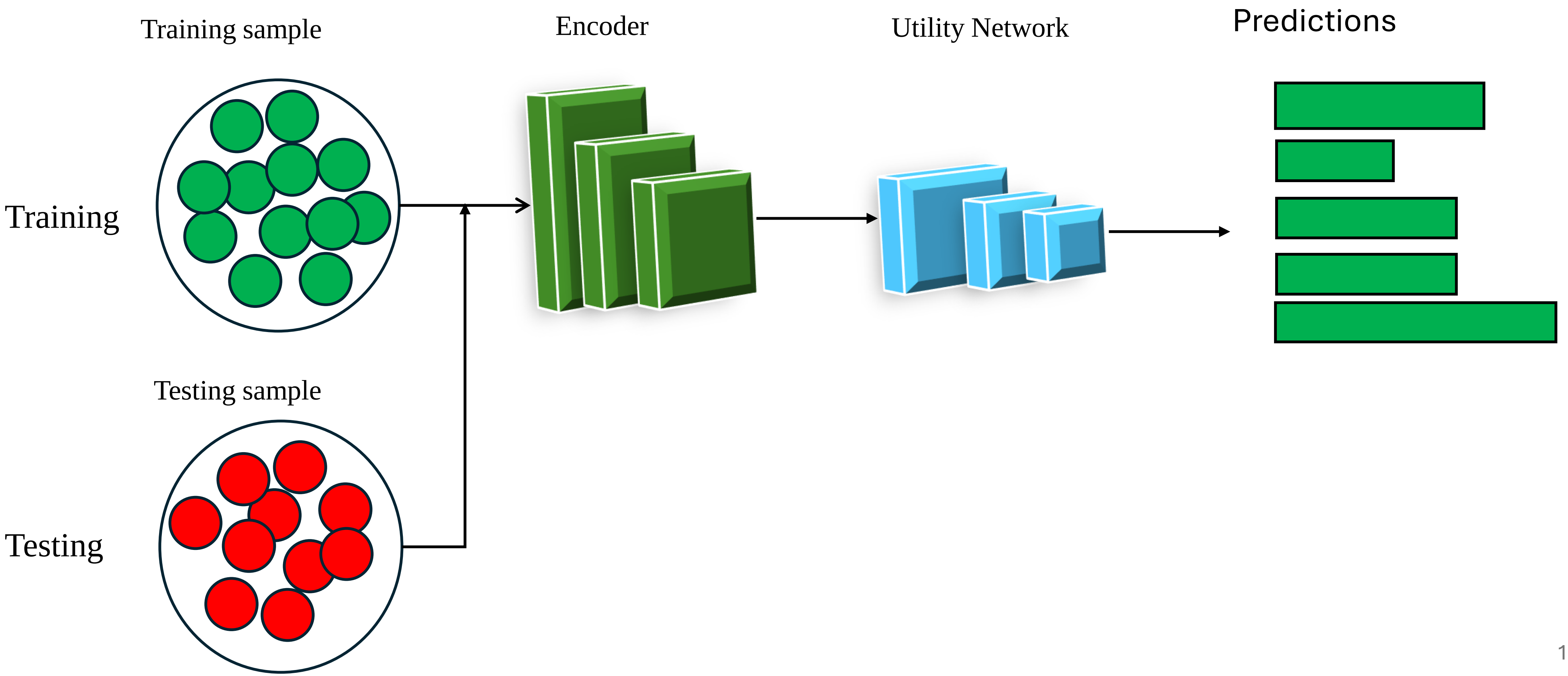
- Our objective function of learning privacy-preserving representations against MIAs:

$$max_f(\lambda\, min_\Psi - E_{p(x,u)}\left[log^{q\Psi(u|f(x))}\right] + (1-\lambda) max_\Omega\, E_{p(x,y,u)}\left[log^{q\Omega(y|f(x),y=1)}\right]$$

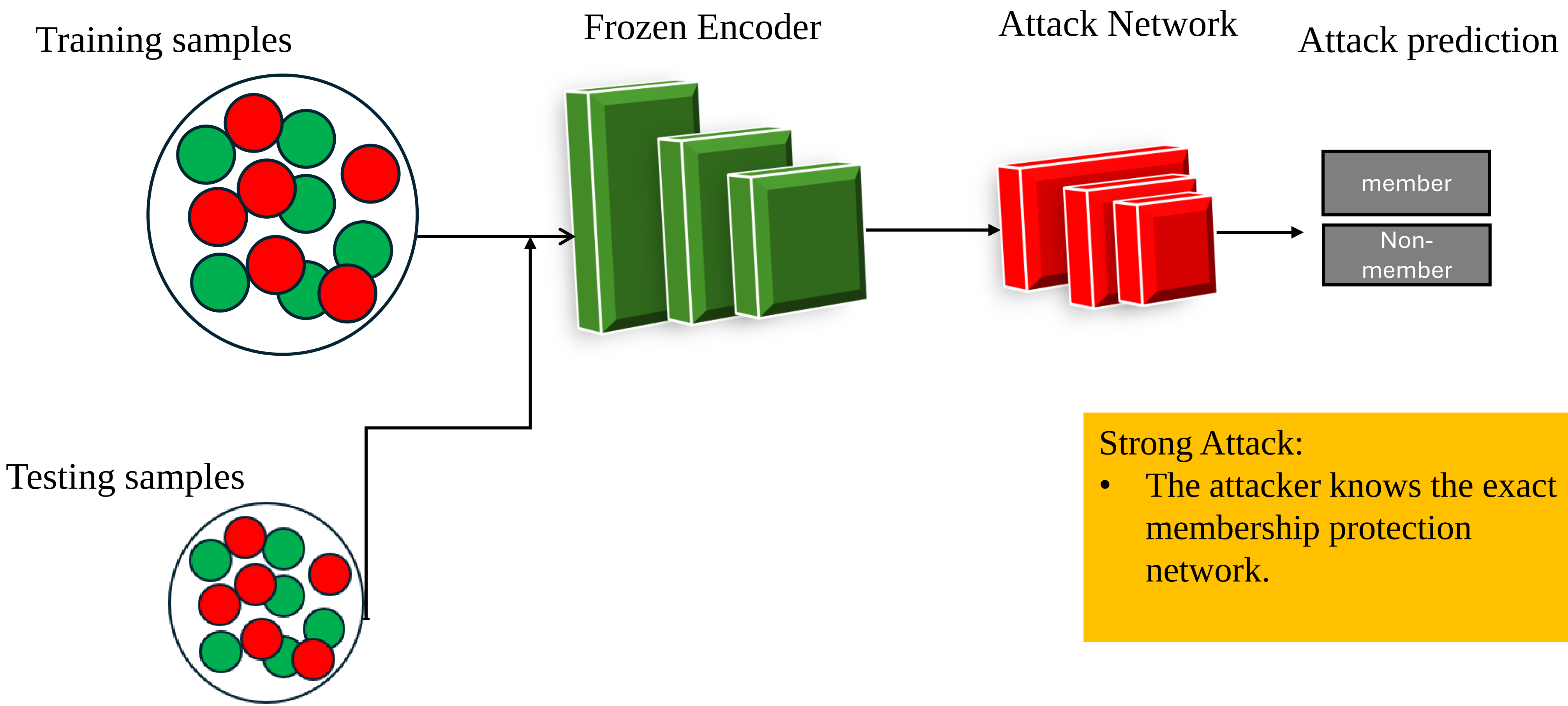
Implementation in practice:

- **Three parameterized neural networks**
- **Encoder f**
- **Membership protection network g_ψ**
- **Utility preservation network h_Ω**

Inf2Guard- Utility Training



Inf2Guard-Attack Training

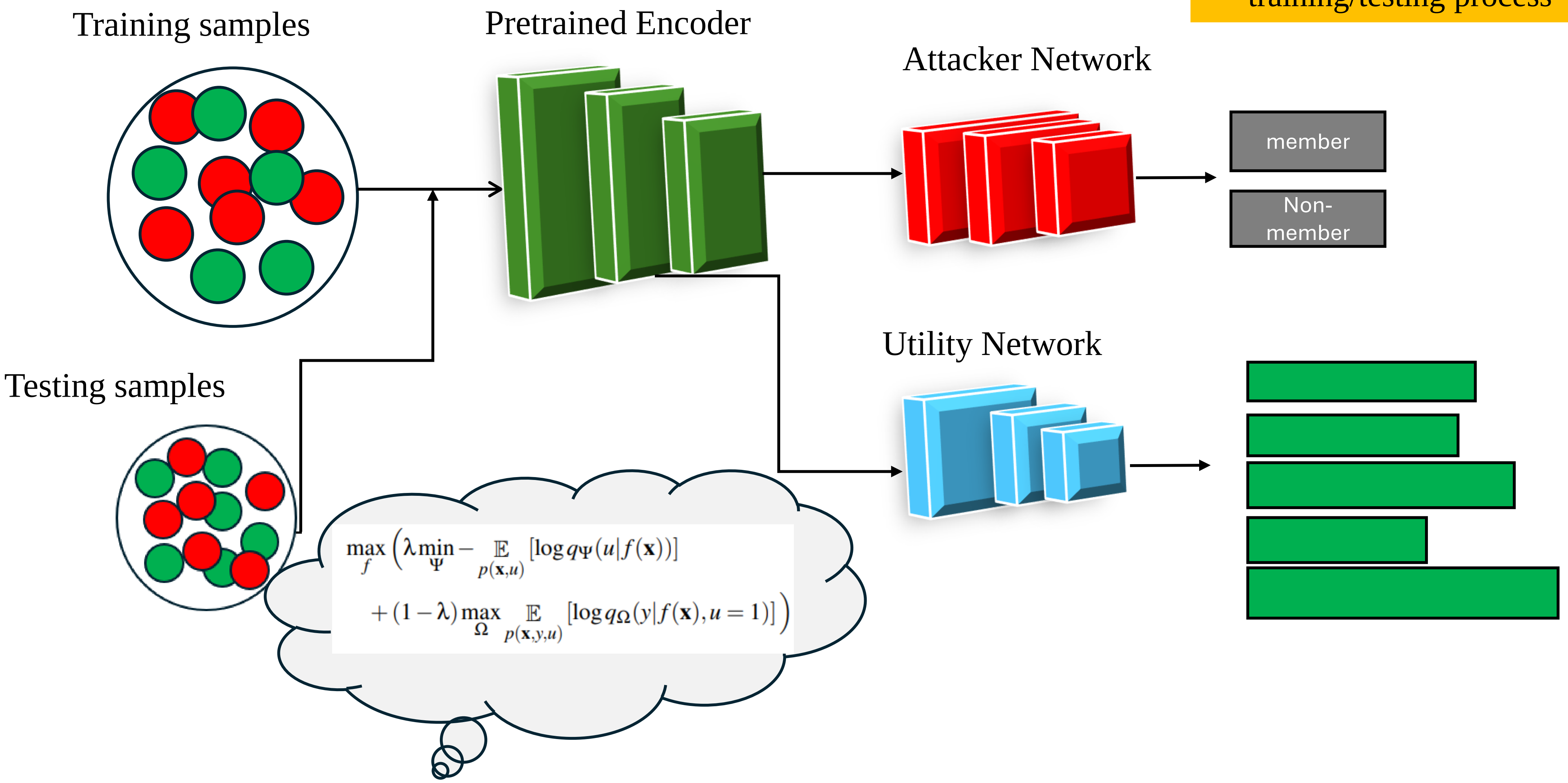


Strong Attack:

- The attacker knows the exact membership protection network.

Inf2Guard-Defense

Strong Attack:
- The training/testing set in the defense was used in the attacker training/testing process



$$\max_{f} \Big(\lambda \min_{\Psi} - \mathbb{E}_{p(\mathbf{x},u)} [\log q_{\Psi}(u|f(\mathbf{x}))] + (1-\lambda) \max_{\Omega} \mathbb{E}_{p(\mathbf{x},y,u)} [\log q_{\Omega}(y|f(\mathbf{x}), u=1)] \Big)$$

Theoretical Results

Theorem 1 (Privacy Leakage Bound)

- **Key Result**: The probability that an MIA correctly infers membership u is bounded by:

$$\Pr(A_{MIA}(r) = u) \leq 1 - \frac{H(u|r)}{2log_2^{(\frac{6}{H(u|r)})}}$$

- **Implication**: A larger $H(u|r)$ (conditional entropy) means a lower MIA accuracy, indicating better privacy protection.

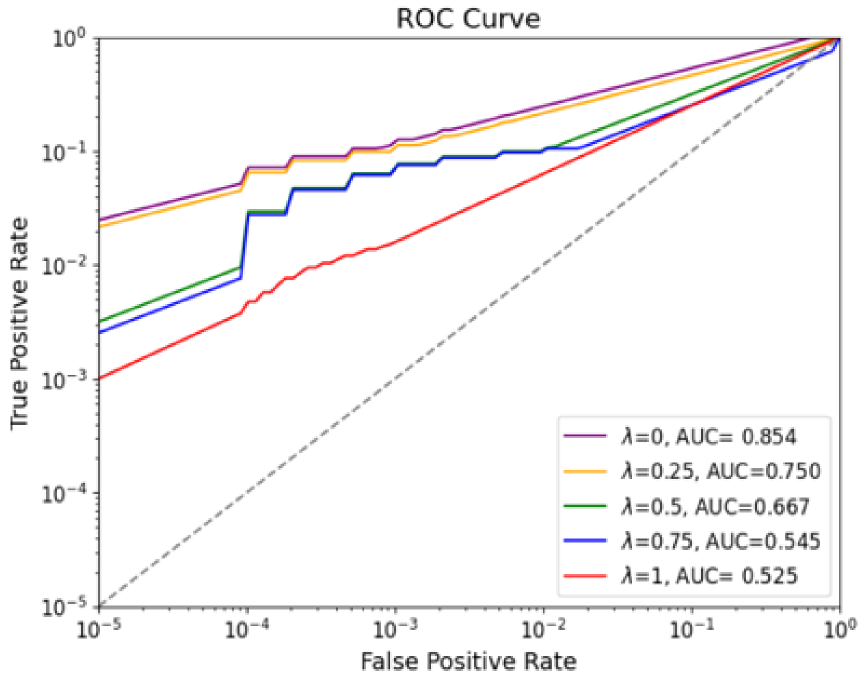
- **Goal:**
  - **Objective**: Maximize $H(u|r)$ by minimizing $I(u;r)$(Mutual Information), thereby reducing MIA effectiveness.

Experimental results- MIA

- **Utility-privacy results**

λ	Utility	MIA Acc
0	78.9%	70.1%
0.25	78.2%	55.9%
0.5	78%	53.5%
0.75	77.2%	51.1%
1	20%	50%

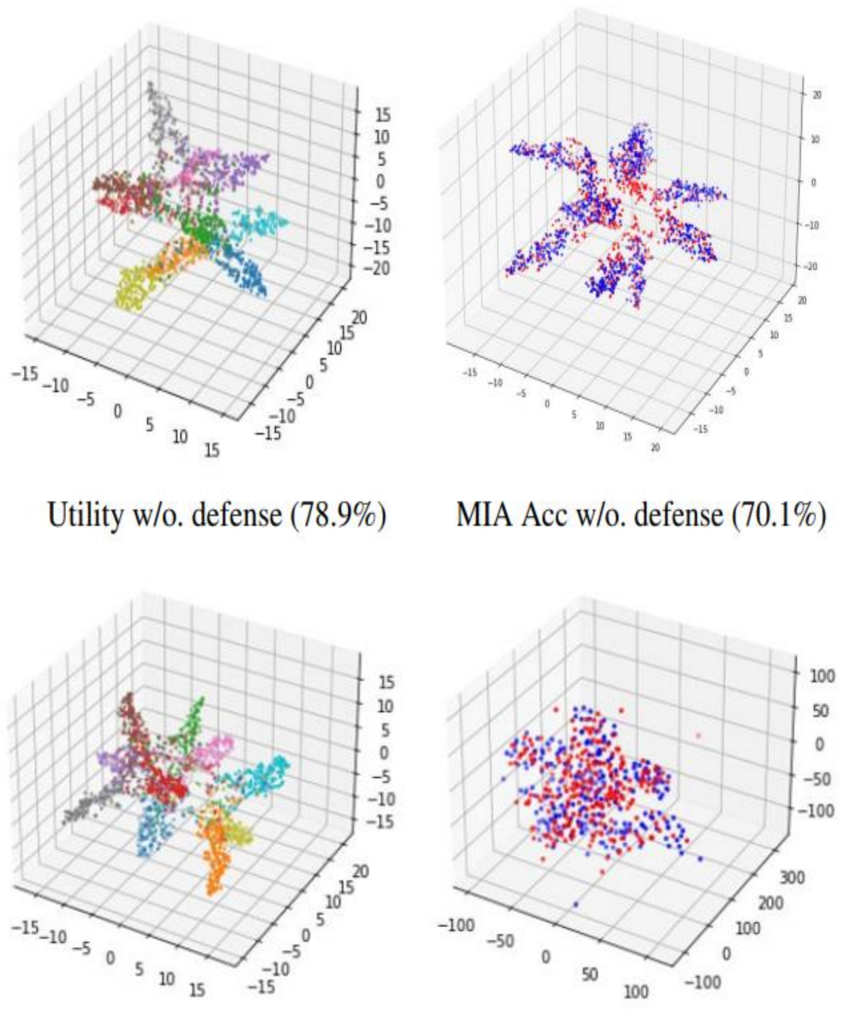
- **TPR vs FPR of Inf2Guard against LiRA**



CIFAR10

Experimental results- MIA

- 3D t-SNE embeddings results on the learnt representation of on CIFAR10.



Utility w/o. defense (78.9%) MIA Acc w/o. defense (70.1%)

Utility w. defense (77.2%) MIA Acc w. defense (51.1%)

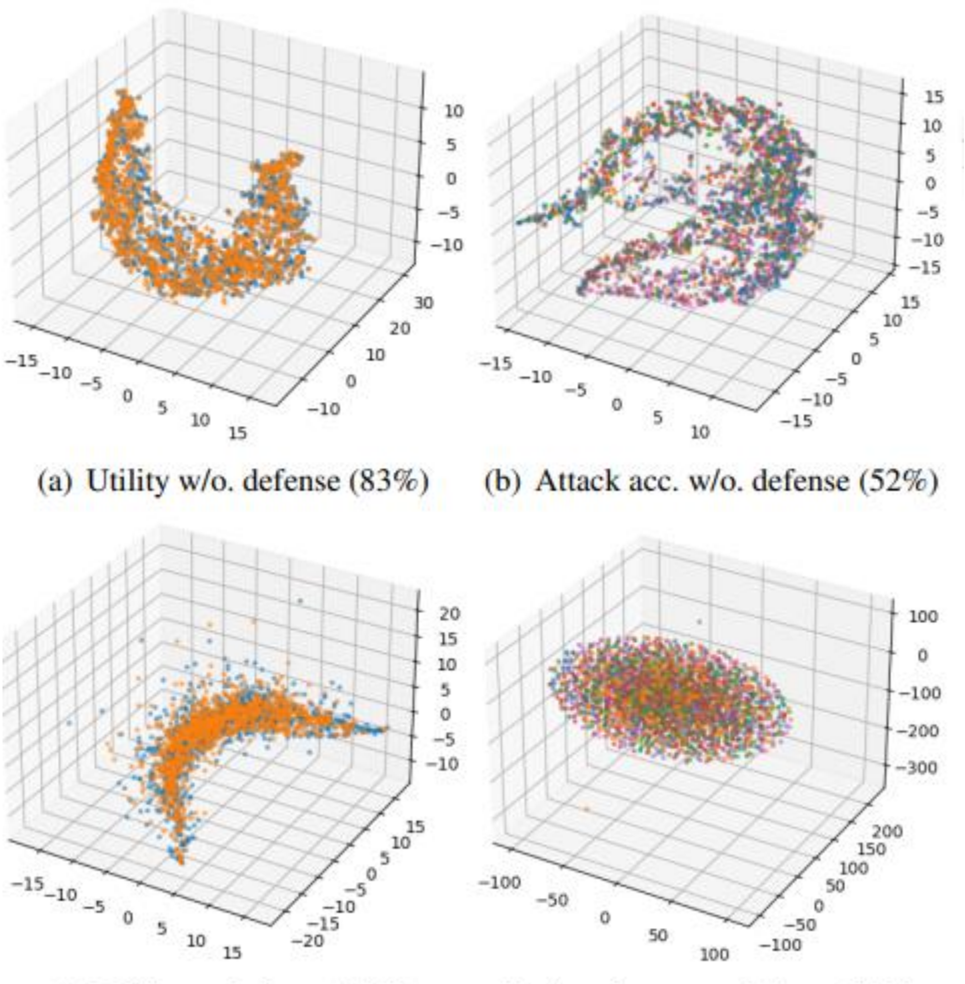
CIFAR10

Experimental results - PIA

- **Comparing with the DP-based defense**

Defense	Census	
	Utility	PIA Acc
DP-encoder	52%	34%
Inf^2Gaurd	76%	34%

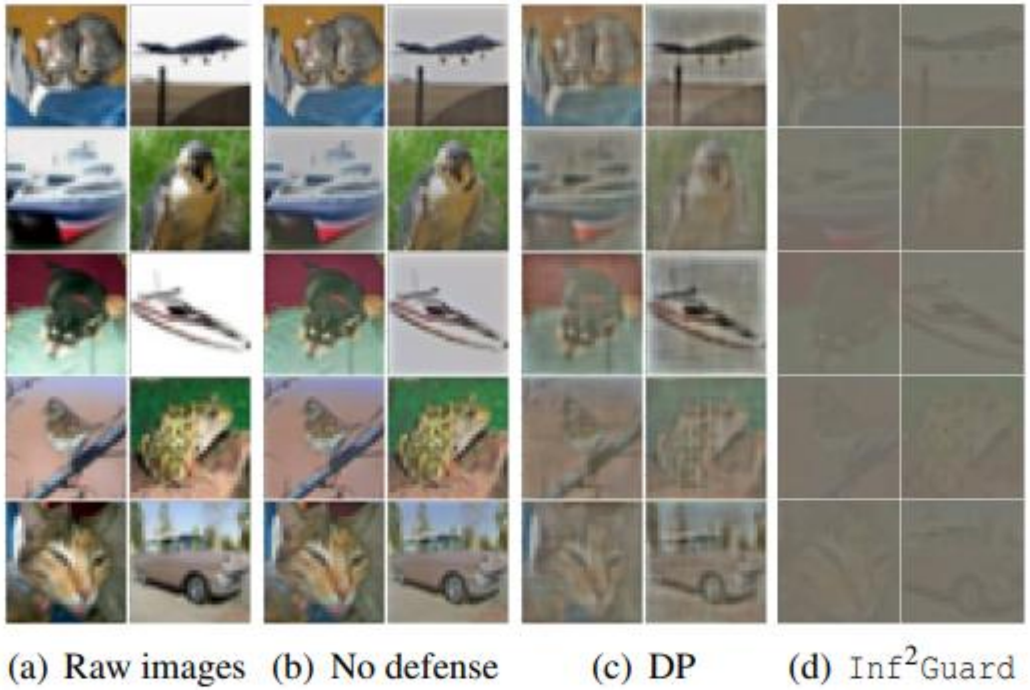
- **3D t-SNE embeddings results**



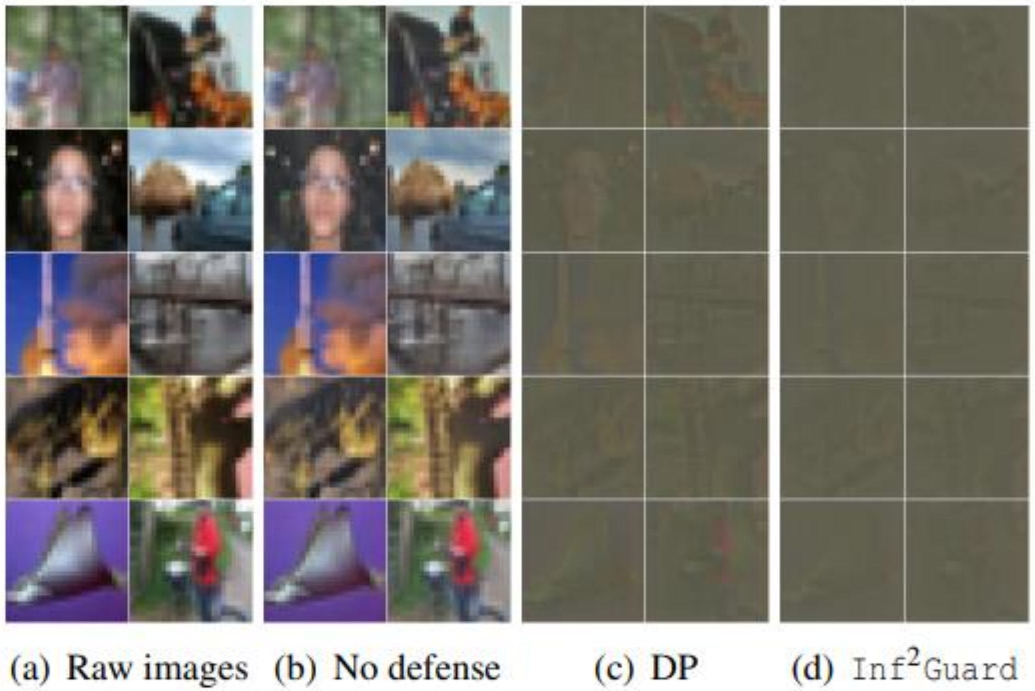
(a) Utility w/o. defense (83%) (b) Attack acc. w/o. defense (52%)

(c) Utility w. defense (80%) (d) Attack acc. w. defense (19%)

Experimental results - DRA



CIFAR10
Fix Utility: 78%



CIFAR100
Fix Utility: 47%

Conclusion

- **Inf2Guard:** A unified information-theoretic framework for learning privacy-preserving representations.
  - Membership Inference, Property Inference, Data Reconstruction
- **Guaranteed privacy leakage**
- **Guaranteed utility-privacy tradeoff**
- **State-of-the-art of the utility-privacy tradeoff**

Contribution

- Contact us:

snoorbakhsh@hawk.iit.edu
bzhang57@hawk.iit.edu
yuan.hong@uconn.edu
bwang70@iit.edu

Code

SCAN ME

Paper

SCAN ME

- Big thanks to our supporter: