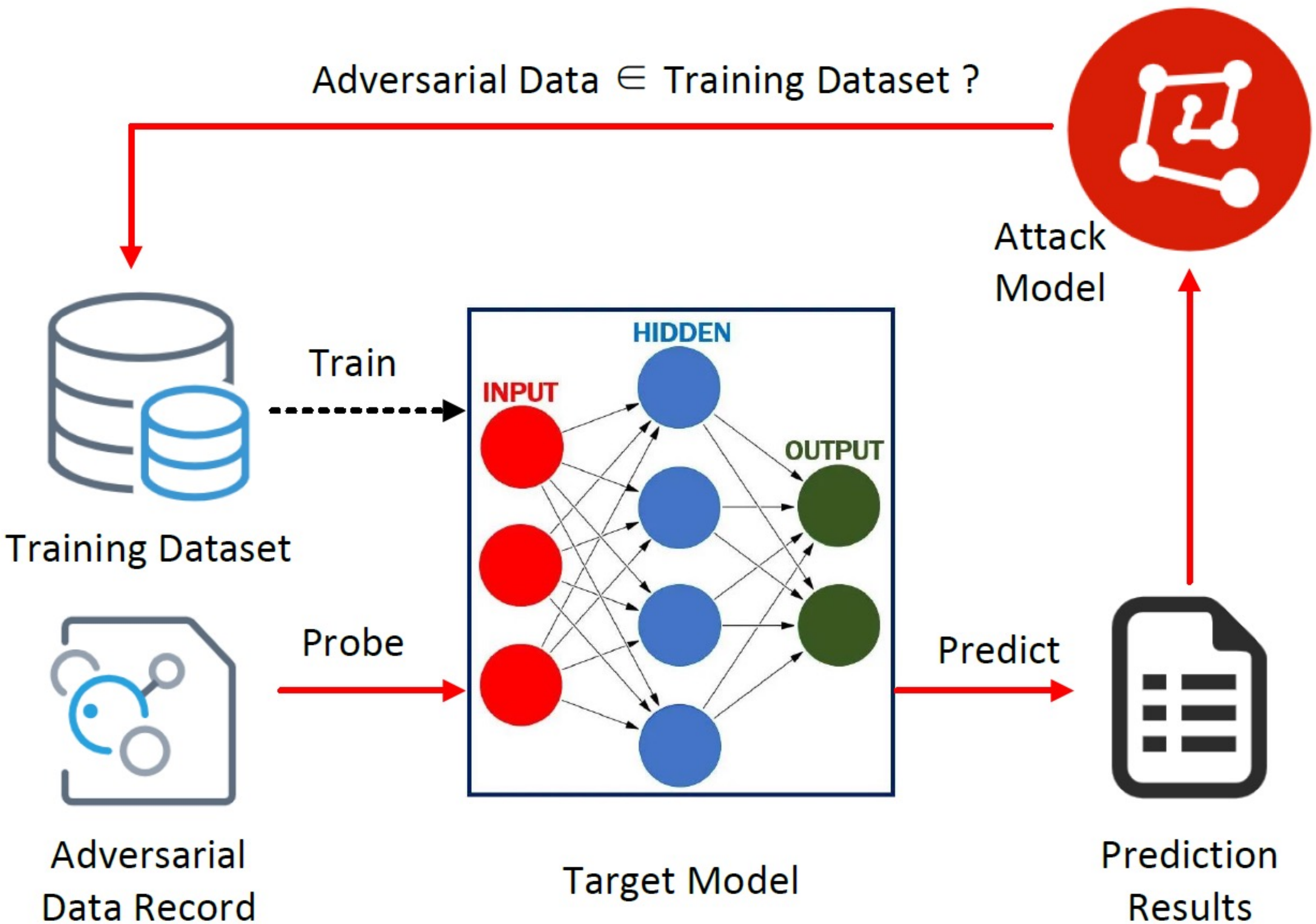
Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning

Ahmed Salem, CISPA Helmholtz Center for Information Security;
Apratim Bhattacharya, Max Planck Institute for Informatics;
Michael Backes, Mario Fritz, and Yang Zhang, CISPA Helmholtz Center for Information Security

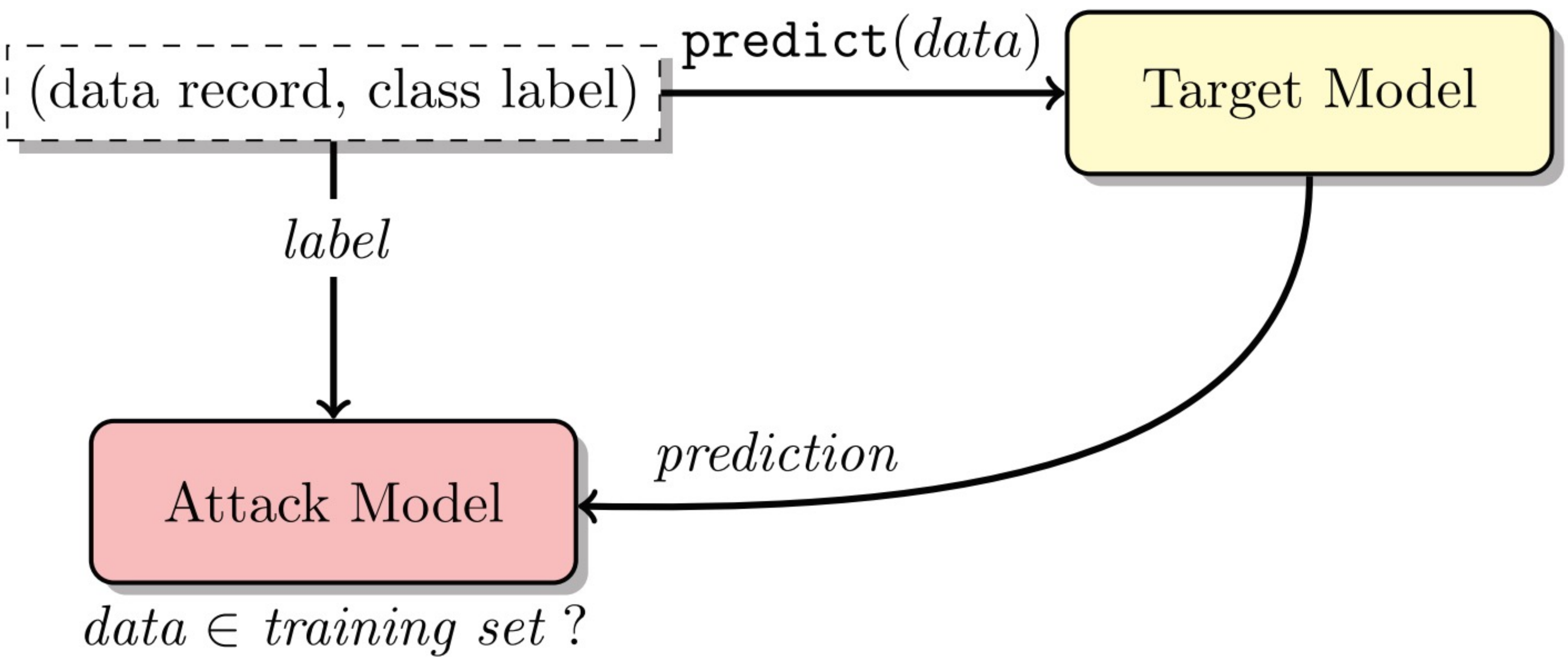
Membership Inference

- **Whether a data record was used as part of the training set**



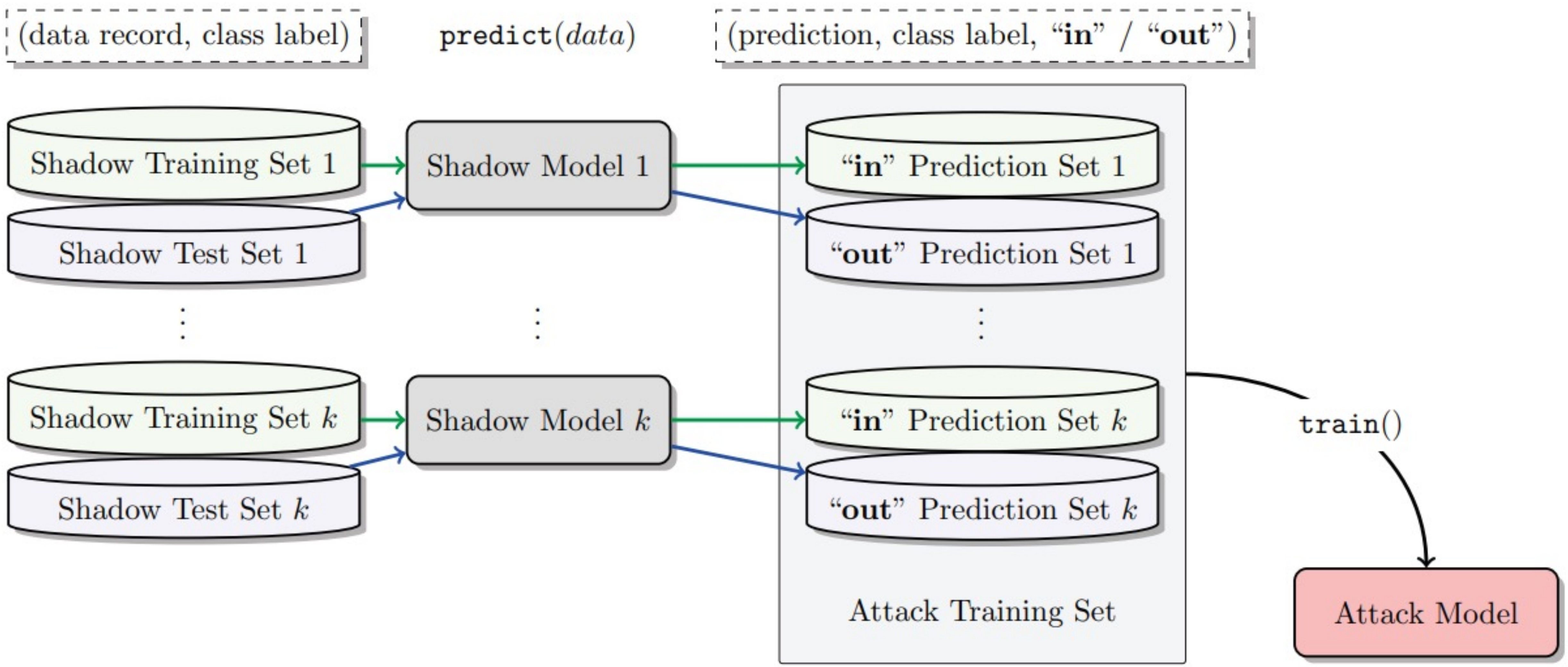
Membership Inference

- **Membership inference attack in the black-box setting**



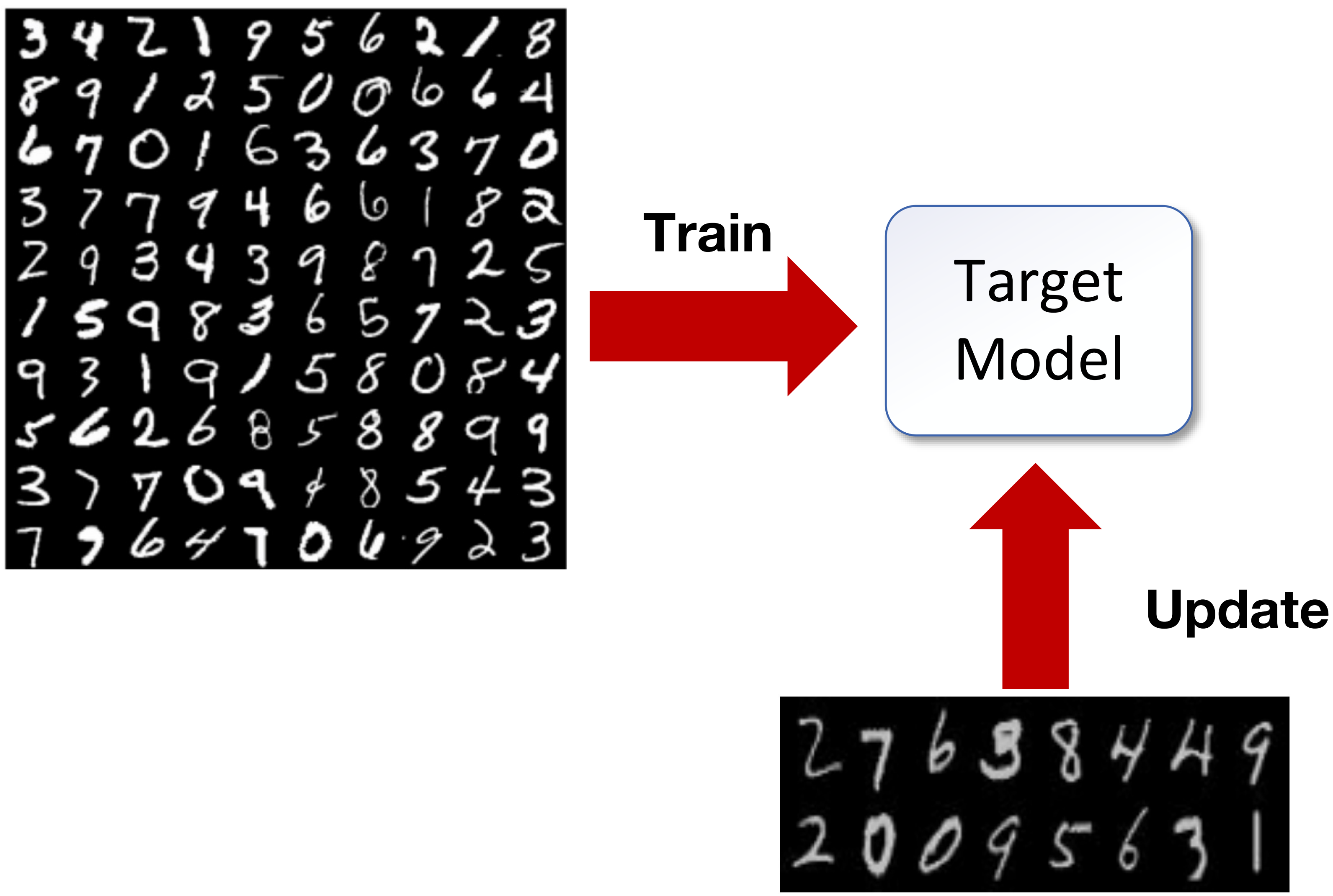
Membership Inference

- **Membership inference attack model**



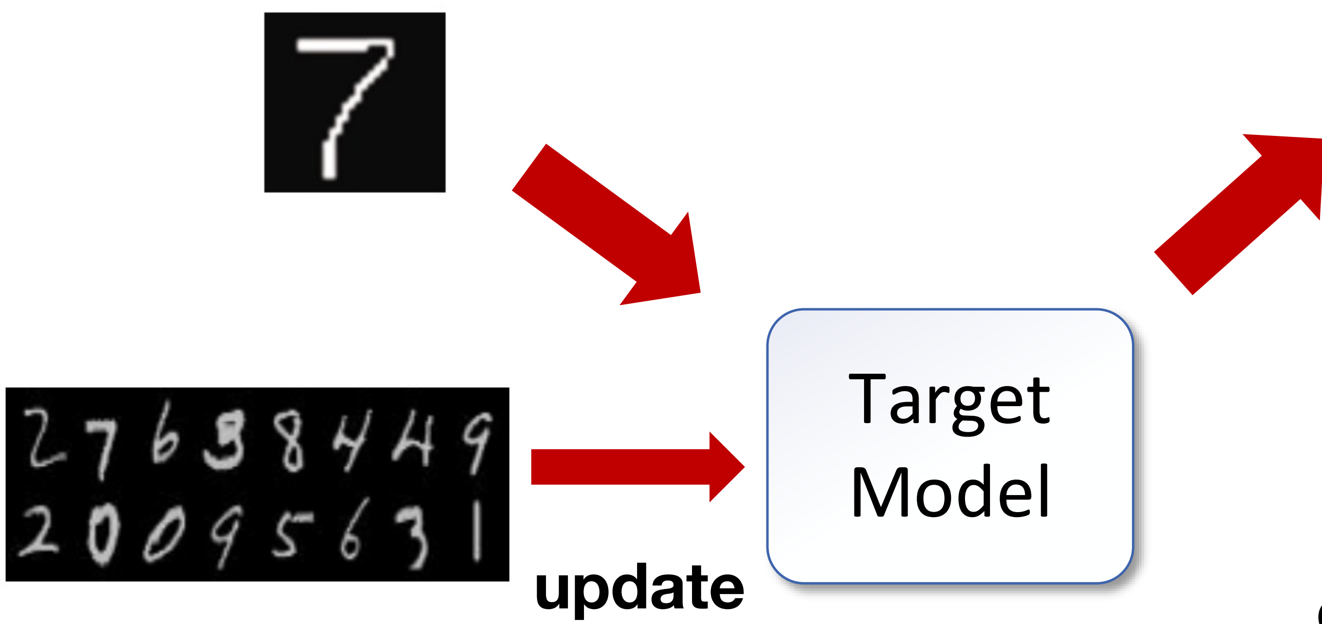
Online Learning

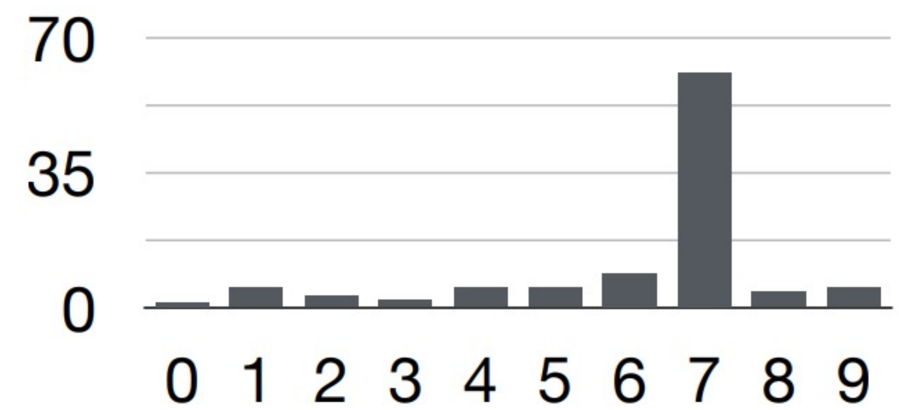
- **Target Model**



Membership Inference

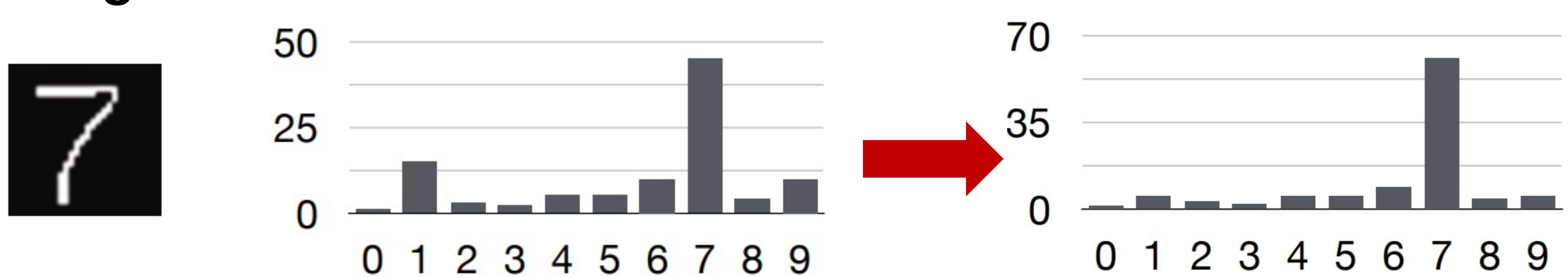
- **Membership inference attack in Online Learning**





Question: Can this posterior difference be a new attack surface?

Probing set



Membership Inference Attacks

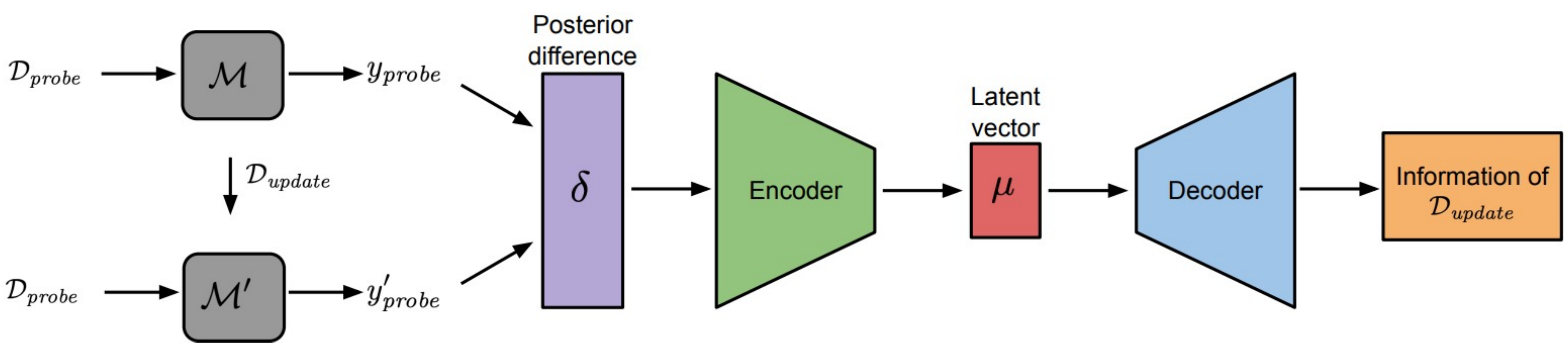
- **Four Attacks**

Single-sample label Inference

Multi-sample label distribution

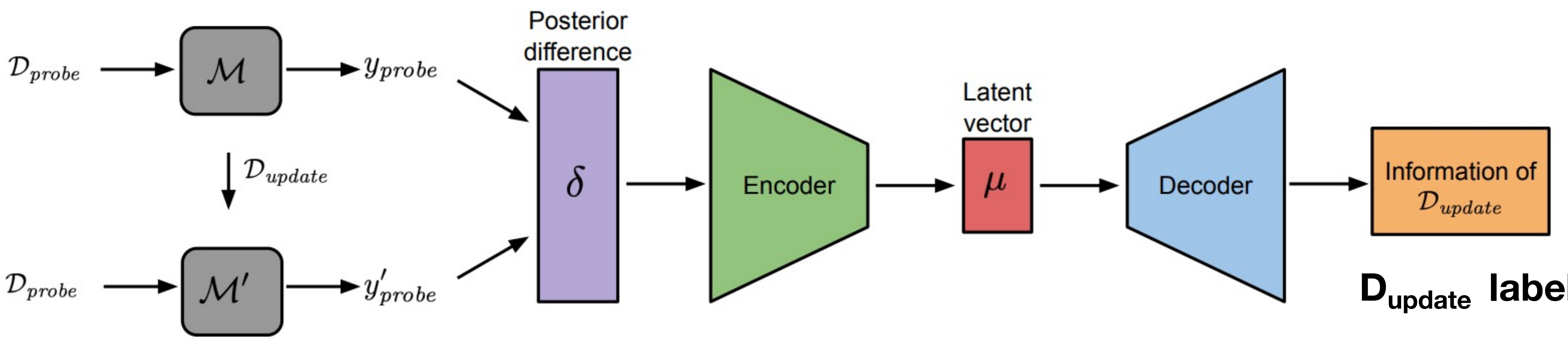
Single-sample reconstruction

Multi-sample reconstruction



Membership Inference Attacks

- **Single-sample label Inference**



$$\mathcal{A}_{LI} : \delta \mapsto \ell$$

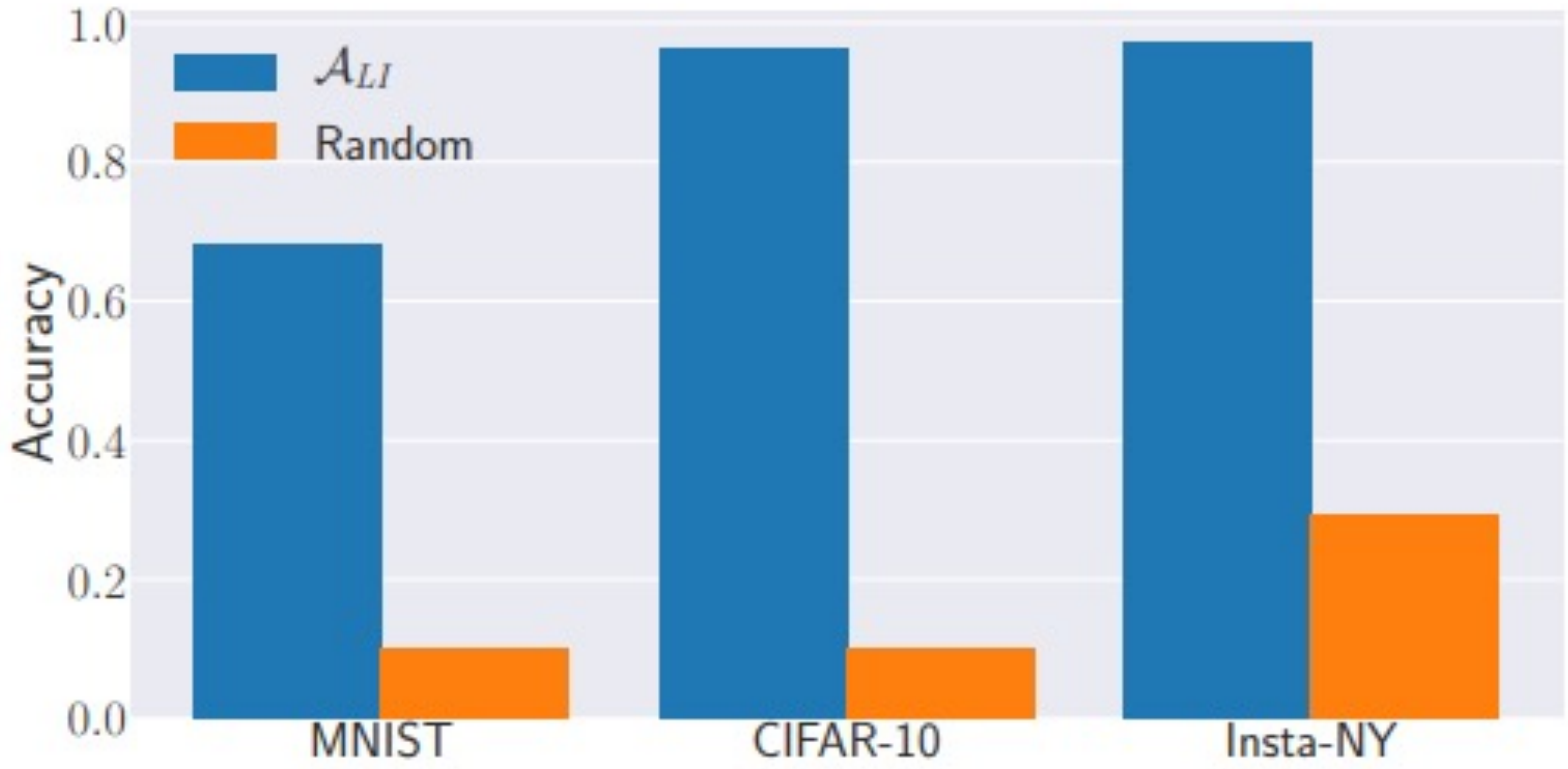
L is a vector with each entry representing the probability of the updating sample affiliated with a certain label.

Train the attack model with cross-entropy loss

$$\mathcal{L}_{CE} = \sum_i \ell_i \log(\hat{\ell}_i)$$

Membership Inference Attacks

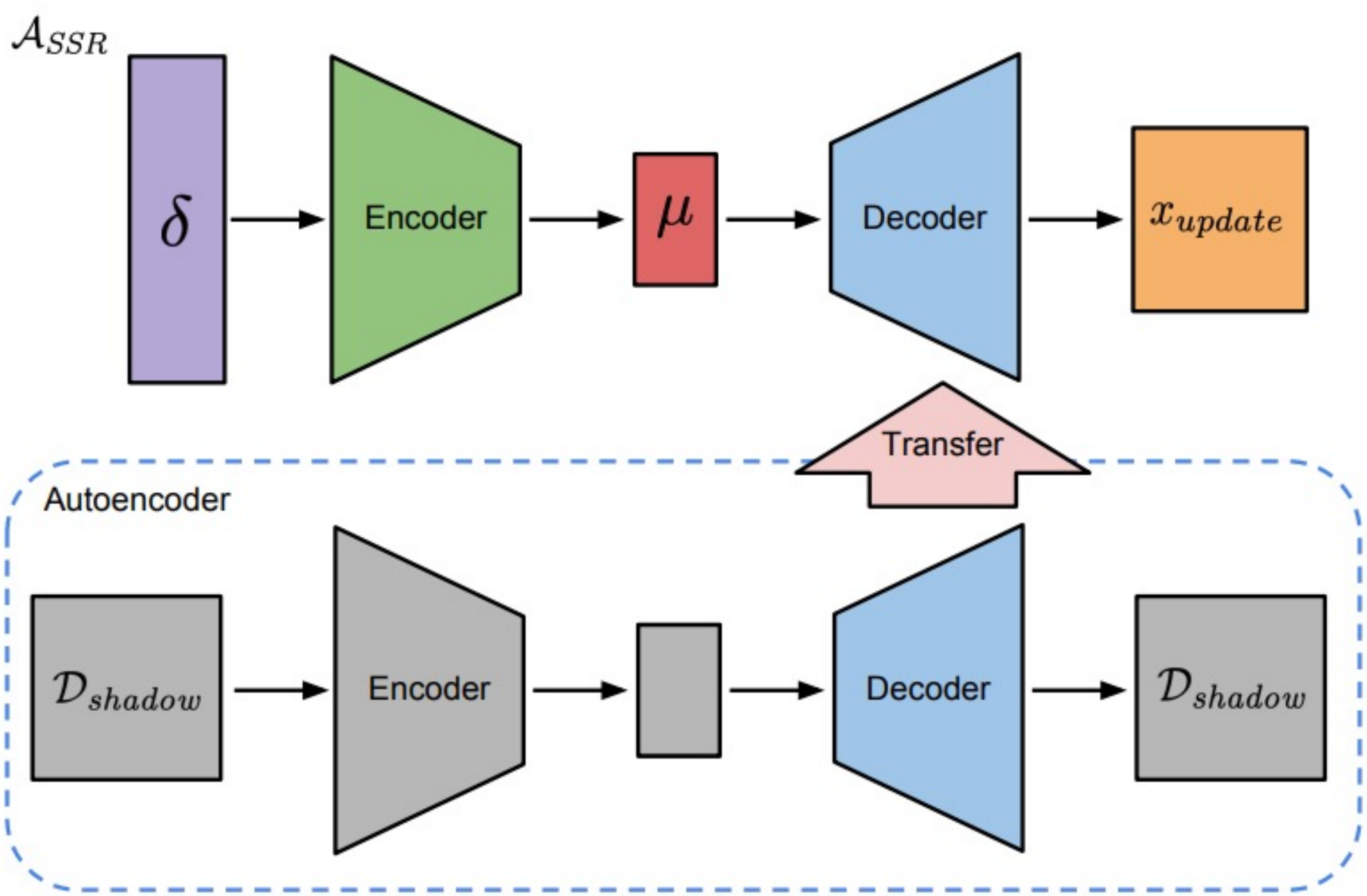
- **Single-sample label Inference**



Use CNN to build shadow and target models for CIFAR-10 and MNIST, and a multilayer perceptron (MLP) for the Insta-NY dataset.

Membership Inference Attacks

- **Single-sample reconstruction**

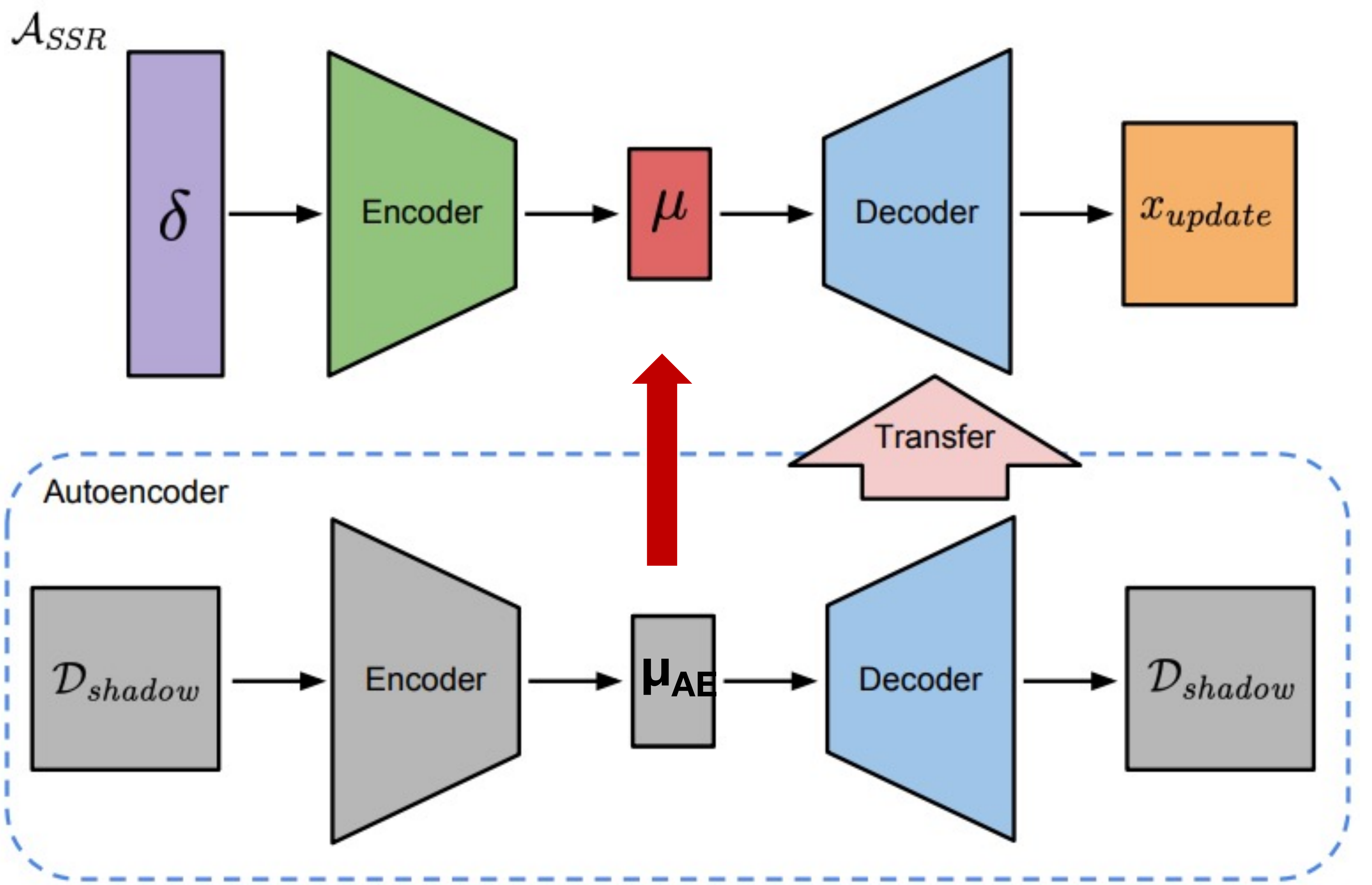


$$\mathcal{A}_{SSR} : \delta \mapsto x_{update}$$

X_{update} the sample used to update the model

Membership Inference Attacks

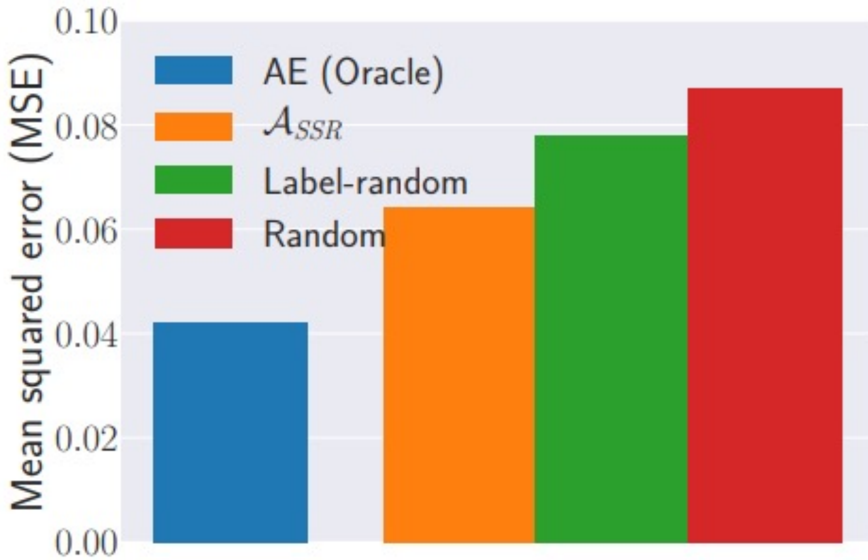
- **Single-sample reconstruction**



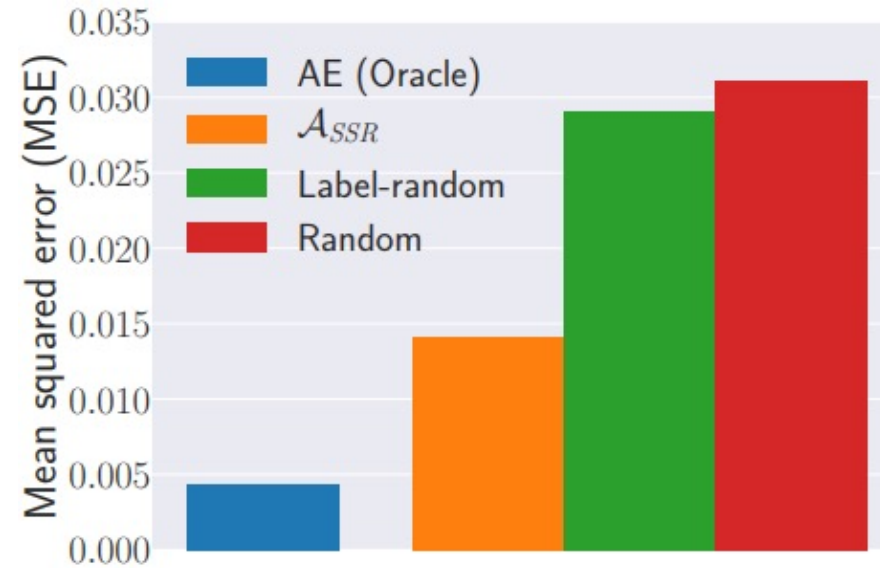
Use mean squared error as the loss function. $\mathcal{L}_{MSE} = ||\hat{x}_{update} - x_{update}||_2^2$

Membership Inference Attacks

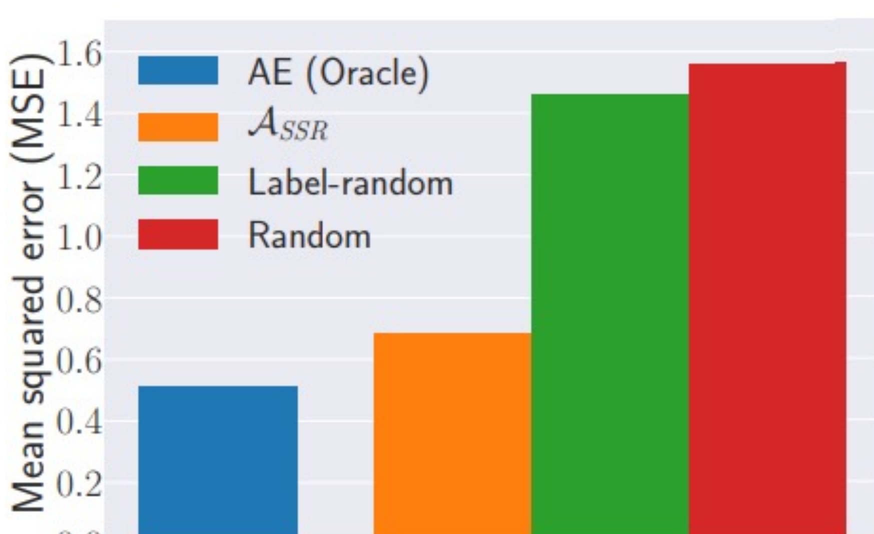
- **Single-sample reconstruction**



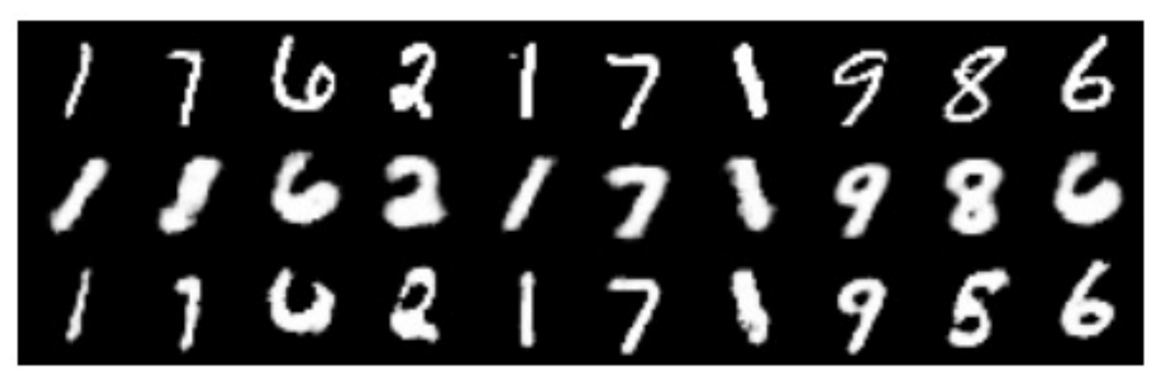
(a) MNIST



(b) CIFAR-10

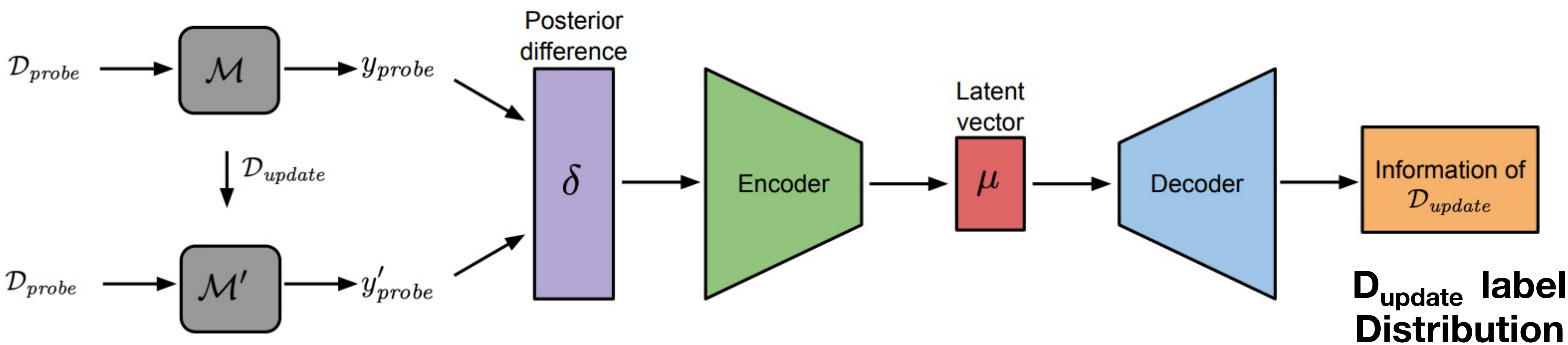


(c) Insta-NY



Membership Inference Attacks

- **Multi-sample label distribution**



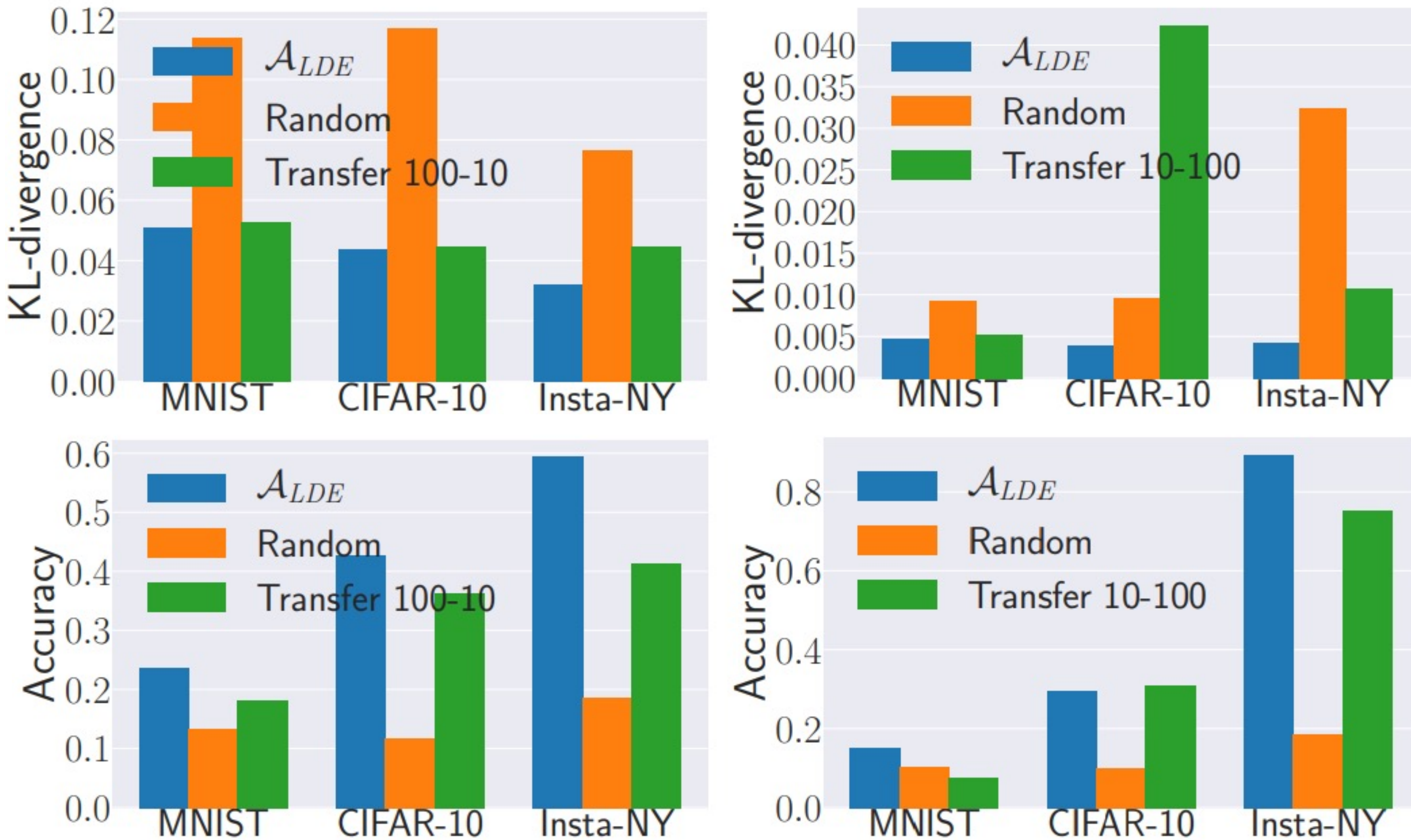
$$\mathcal{A}_{LDE} : \delta \mapsto q$$

q is a vector denotes the distribution of labels over all classes for samples in the updating set

Train the attack model with Kullback-–Leibler divergence $\mathcal{L}_{KL} = \sum_i (\hat{q}_\ell)_i \log \frac{(\hat{q}_\ell)_i}{(q_\ell)_i}$

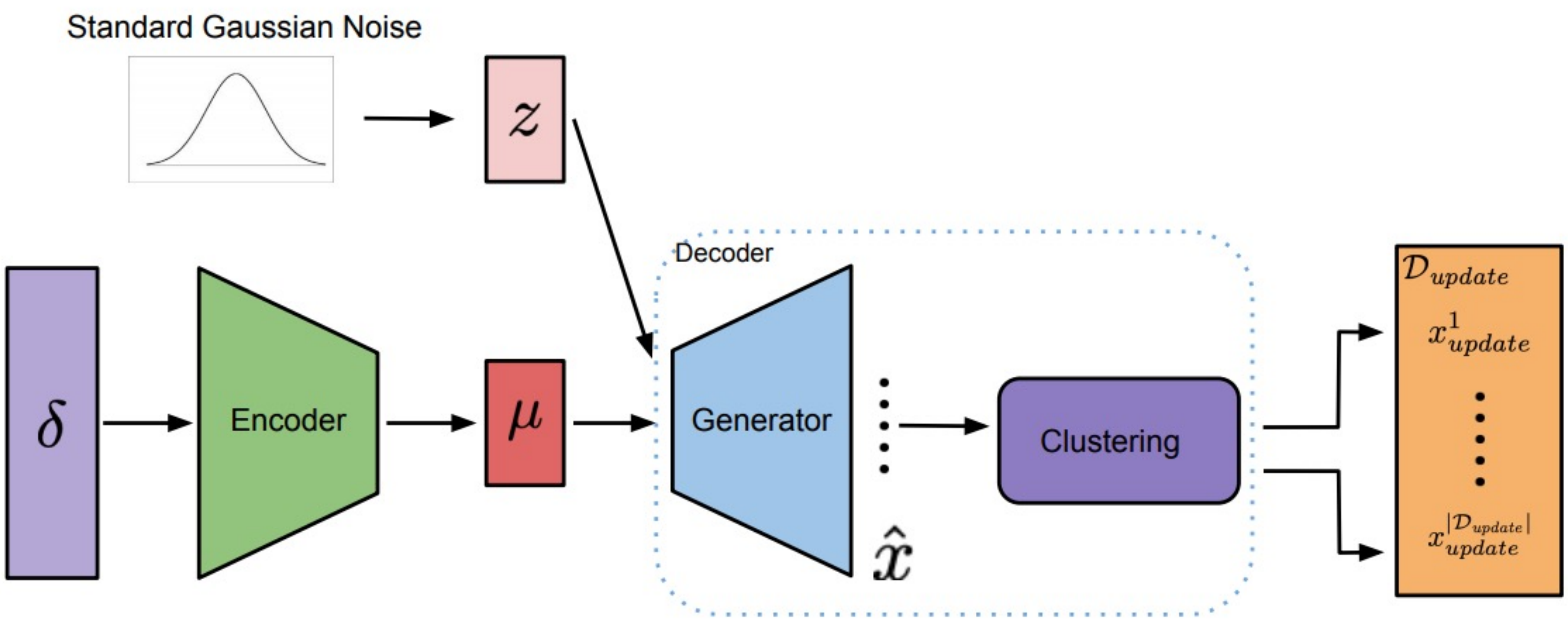
Membership Inference Attacks

- **Multi-sample label distribution**



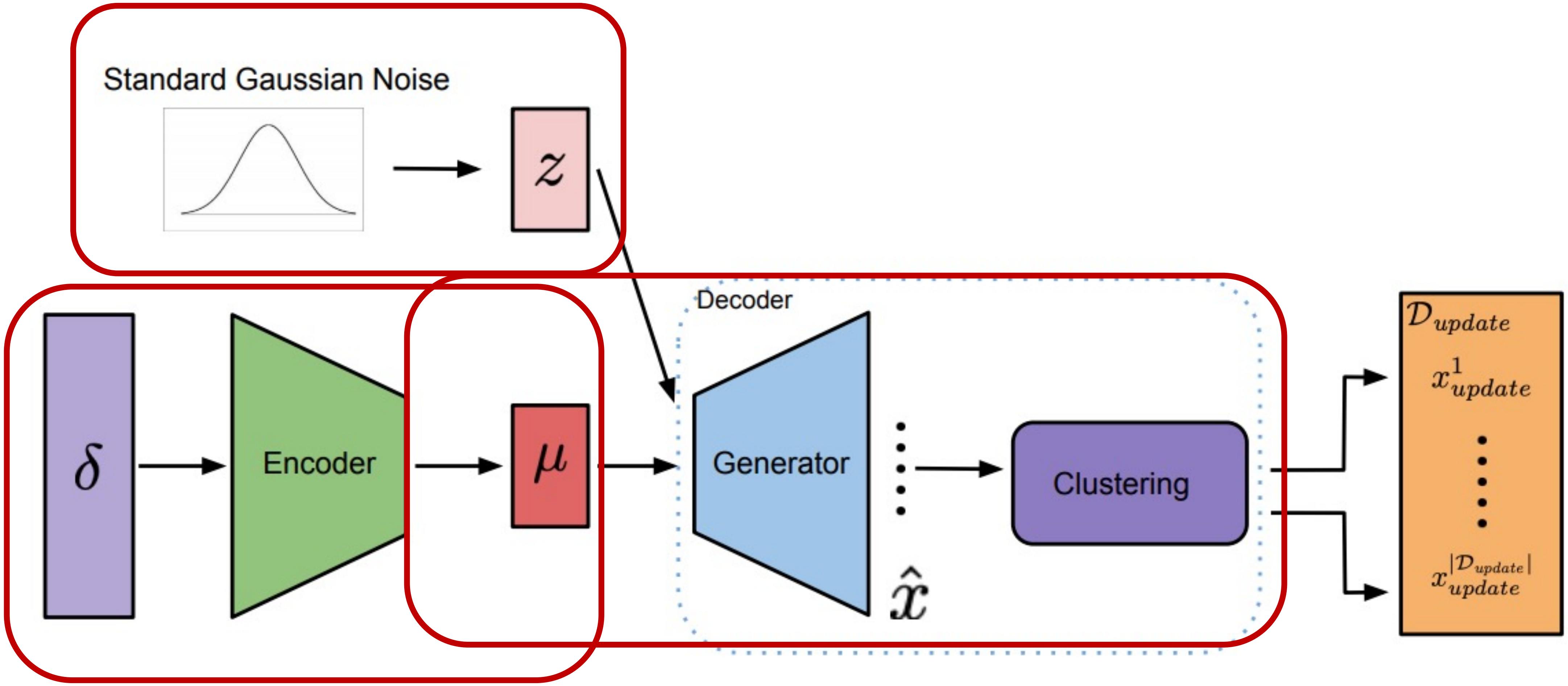
Membership Inference Attacks

- **Multi-sample reconstruction**



Membership Inference Attacks

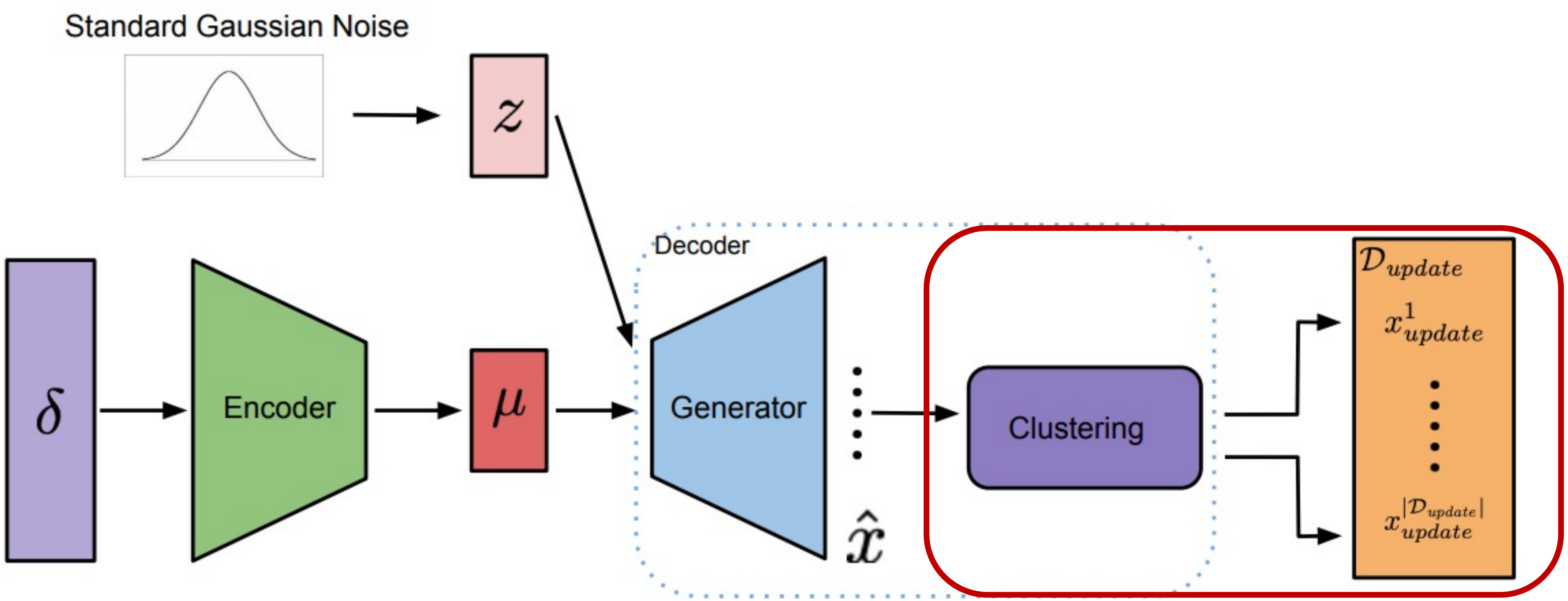
- **Multi-sample reconstruction**



$$\mathcal{L}_{BM} = \sum_{x \in \mathcal{D}_{update}} \min_{\hat{x} \sim G} \|\hat{x} - x\|_2^2 + \sum_{\hat{x}} \log(\mathrm{D}(\hat{x}))$$

Membership Inference Attacks

- **Multi-sample reconstruction**



Membership Inference Attacks

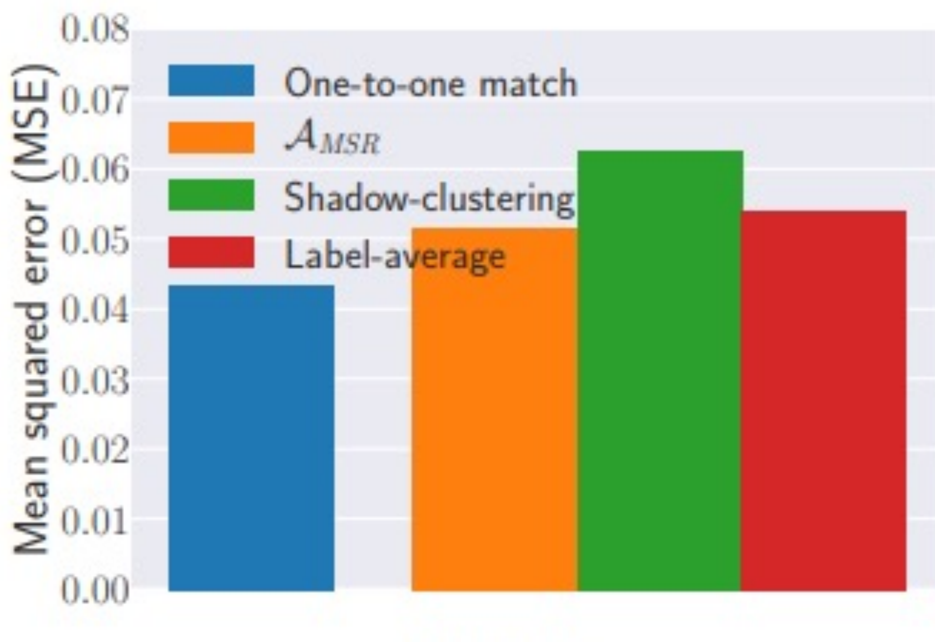
- **Multi-sample label distribution**

Membership Inference Attacks

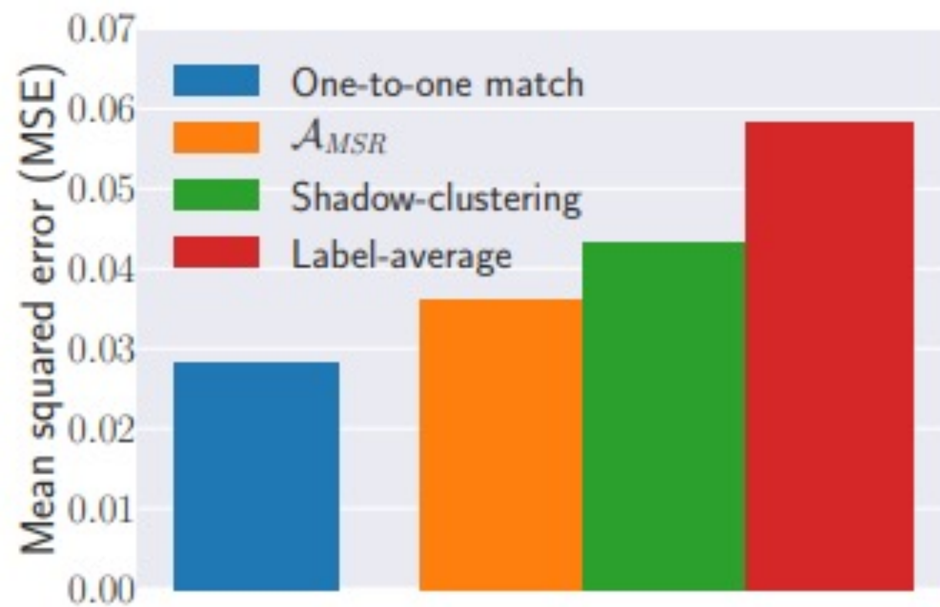
- **Multi-sample label distribution**

Membership Inference Attacks

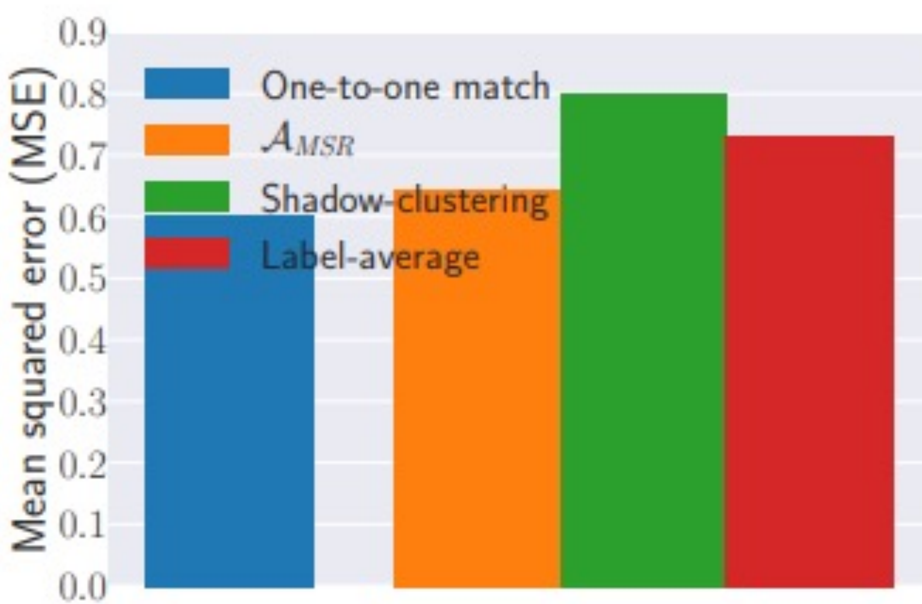
- **Multi-sample label distribution**



(a) MNIST



(b) CIFAR-10



(c) Insta-NY

Discussion

- **Relaxing The Attacker Model Assumption**

1. Same structure for both target and shadow models

2. Same data distribution for both target and shadow datasets

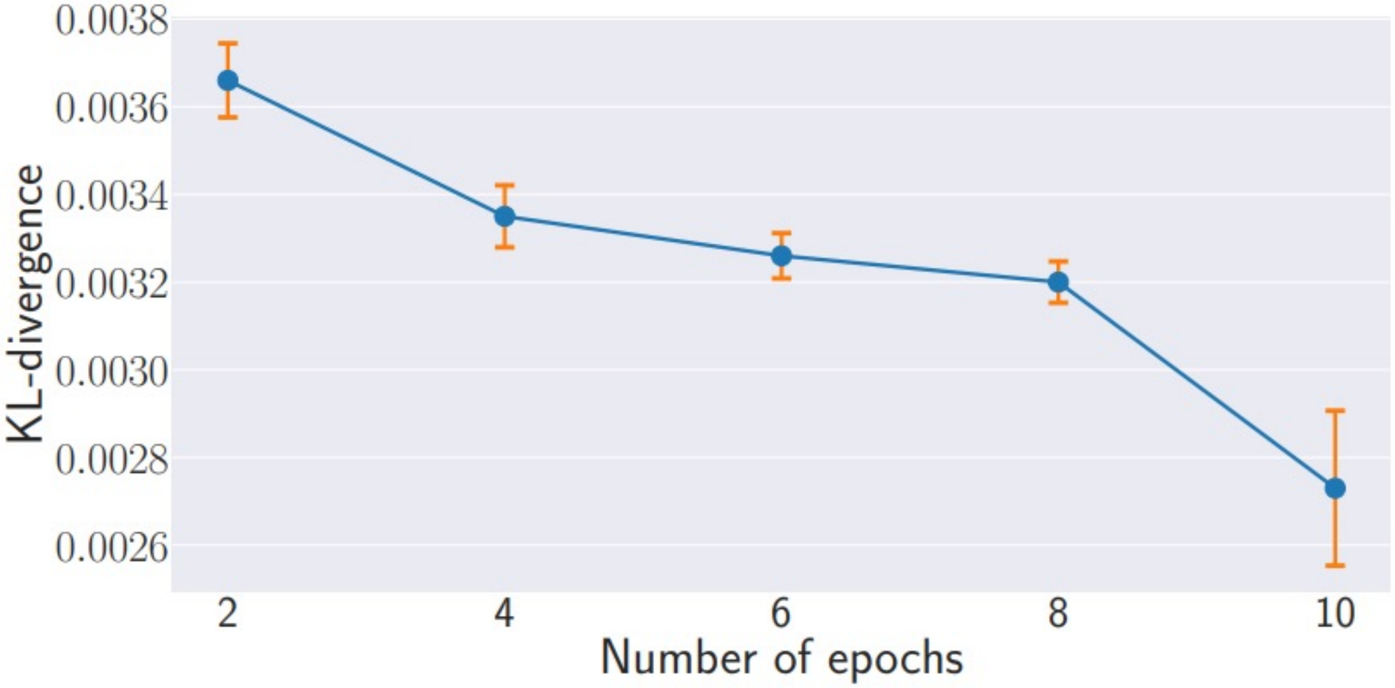
Attack	Original	Transfer
$\mathcal{A}_{LI}$	0.97	0.89
$\mathcal{A}_{SSR}$	0.68	1.1
$\mathcal{A}_{LDE}(10)$	0.59(0.0317)	0.55(0.0377)
$\mathcal{A}_{LDE}(100)$	0.89(0.0041)	0.89 (0.0067)
$\mathcal{A}_{MSR}$	0.64	0.73

Discussion

- **Relaxing The Knowledge of Updating Set Cardinality**

The adversary's knowledge of the updating set cardinality

- **Effect of Target Model Hyperparameters——Updating Epochs**



Discussion

- **Limitations of Attacks.**

1. The target model is solely updated on new data.

2. They perform the attacks on updating sets of maximum cardinality of 100.