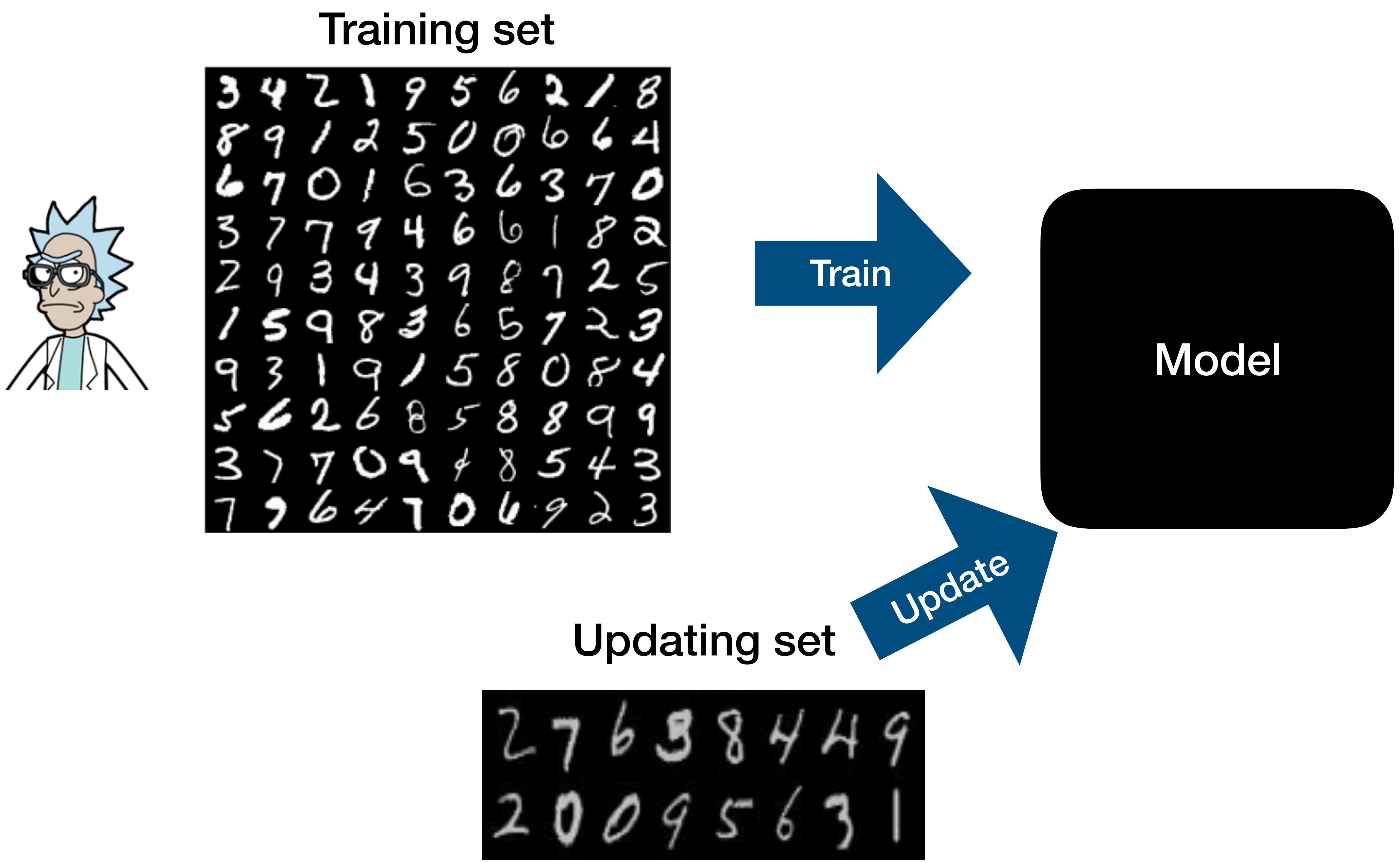
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

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

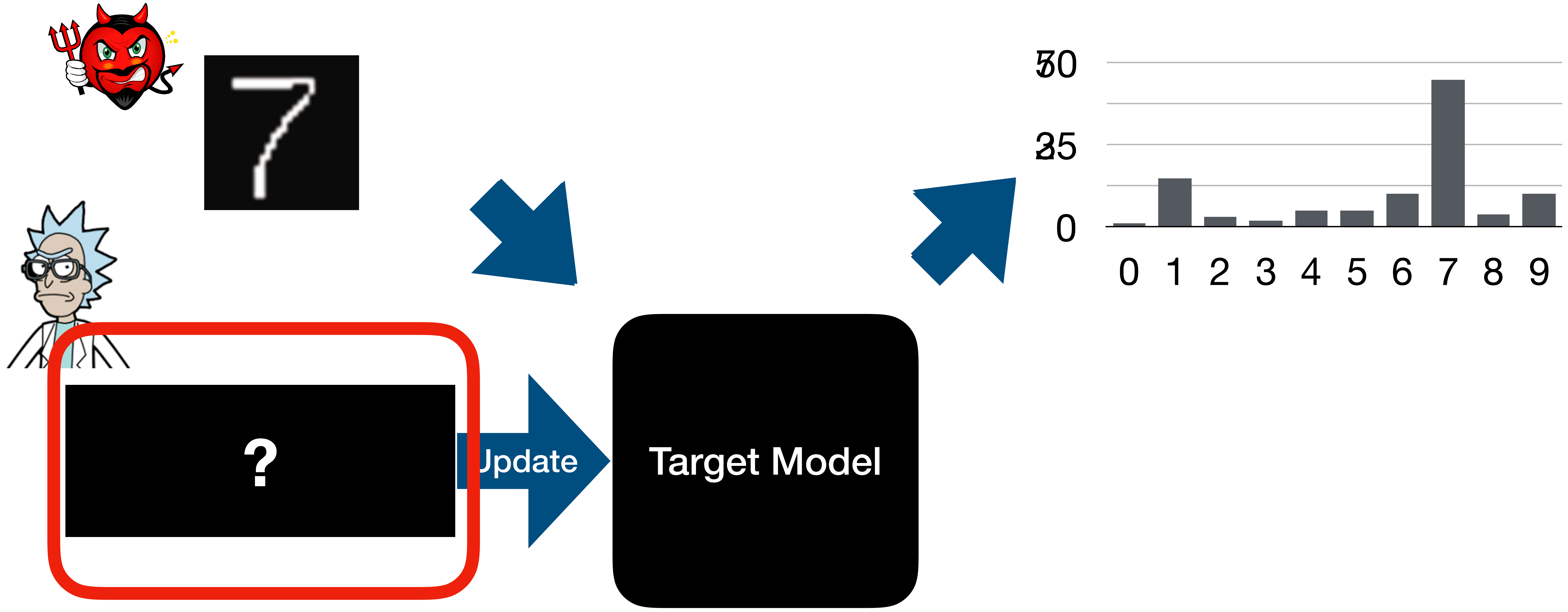
Ahmed Salem, Apratim Bhattacharya, Michael Backes
Mario Fritz, Yang Zhang
CISPA Helmholtz Center for Information Security, Max Planck Institute for Informatics

Online Learning

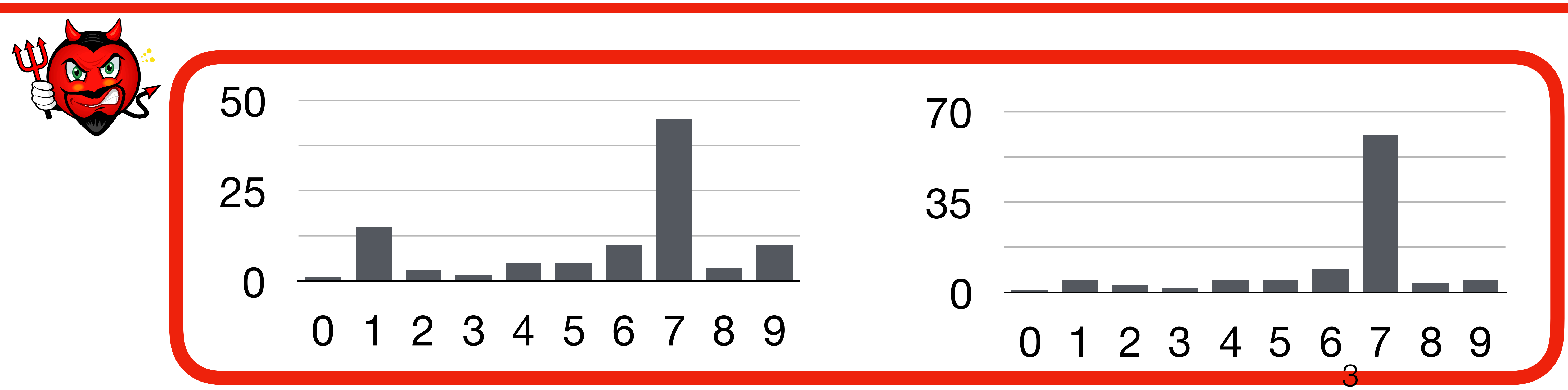


- Data generation rate
  - *90%* of the data in the world today has been created in the last two years alone
- Cost of retraining

Attack Surface in Online Learning



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



Threat Model

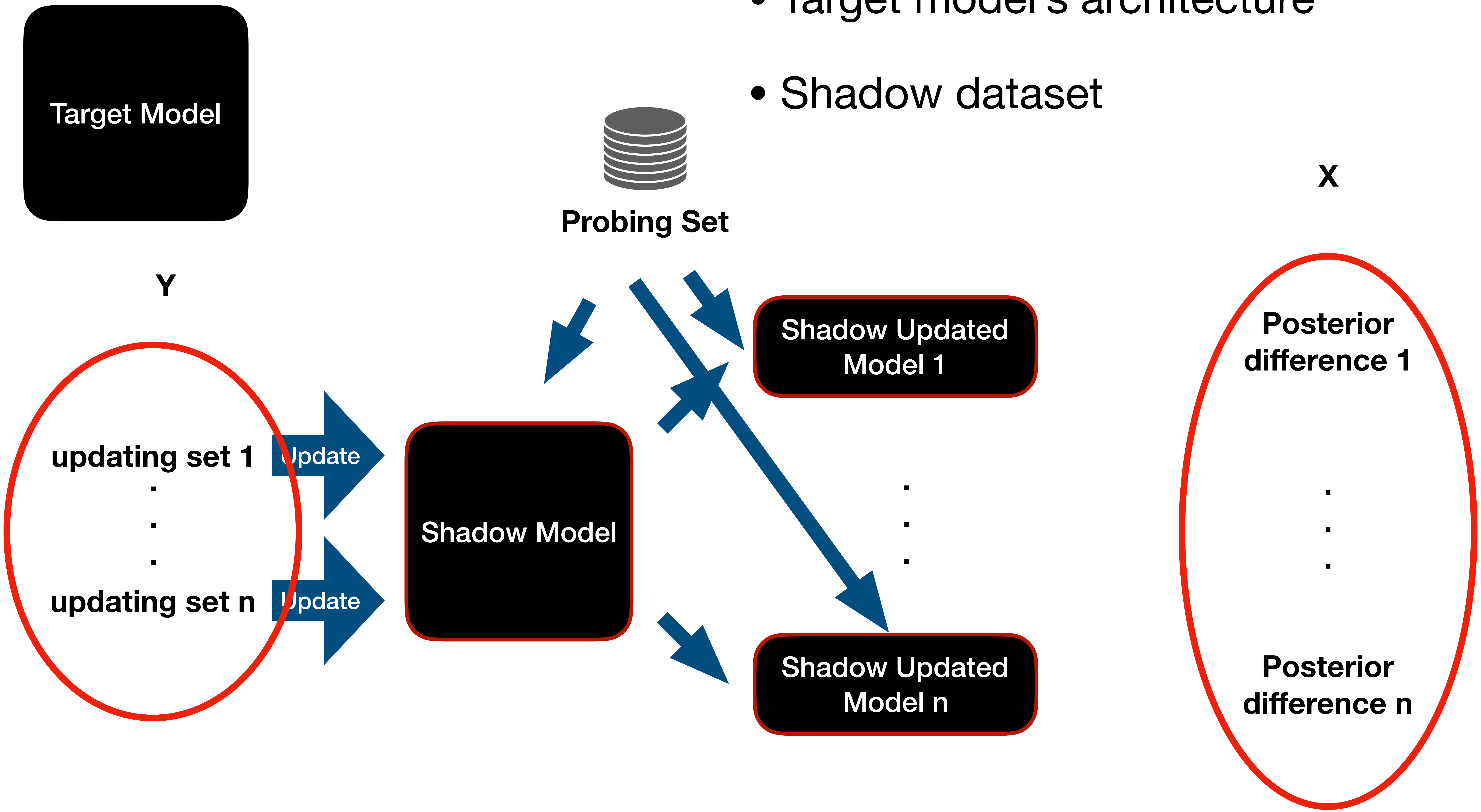
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Attacker has black-box access to the target model
- Attacker knows:
  - Target model's architecture
  - A shadow dataset from the same distribution of the target model's dataset

General Attack Pipeline

Attack Model Training

- Target model's architecture
- Shadow dataset



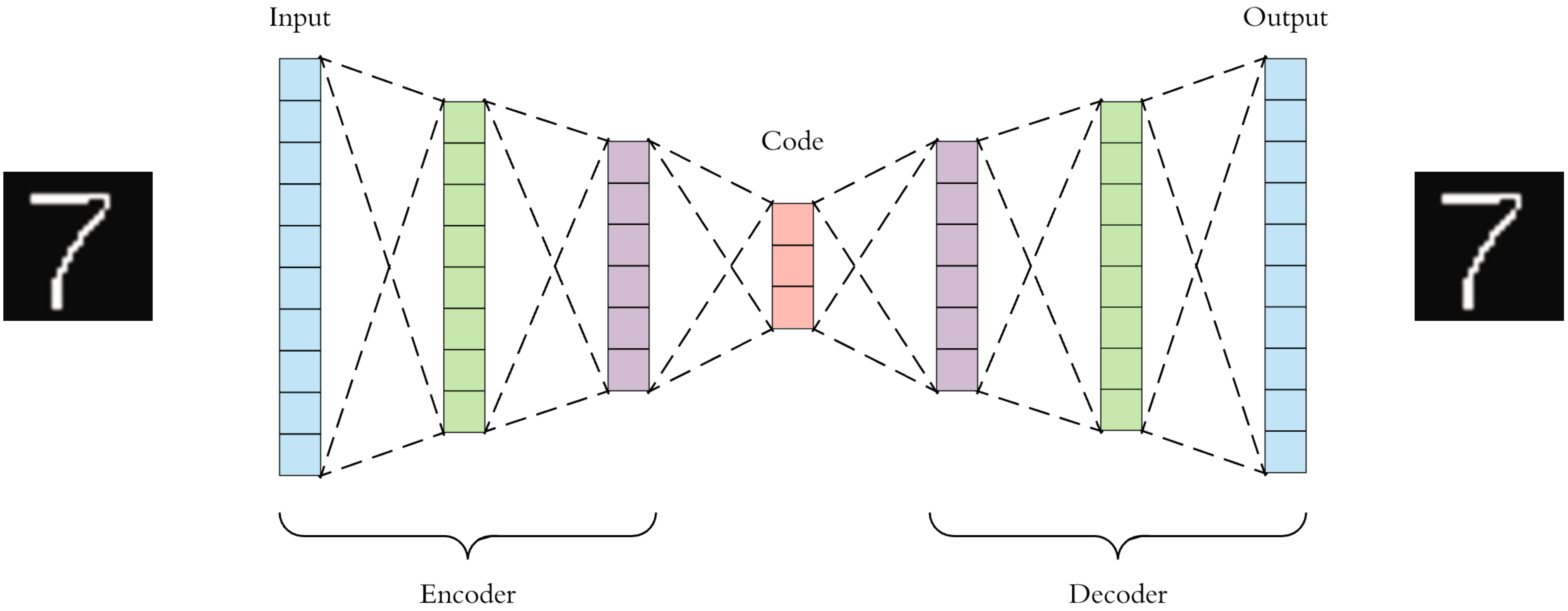
Single-sample Label Inference

- More than 6x and 9x better than baseline for MNIST and CIFAR-10

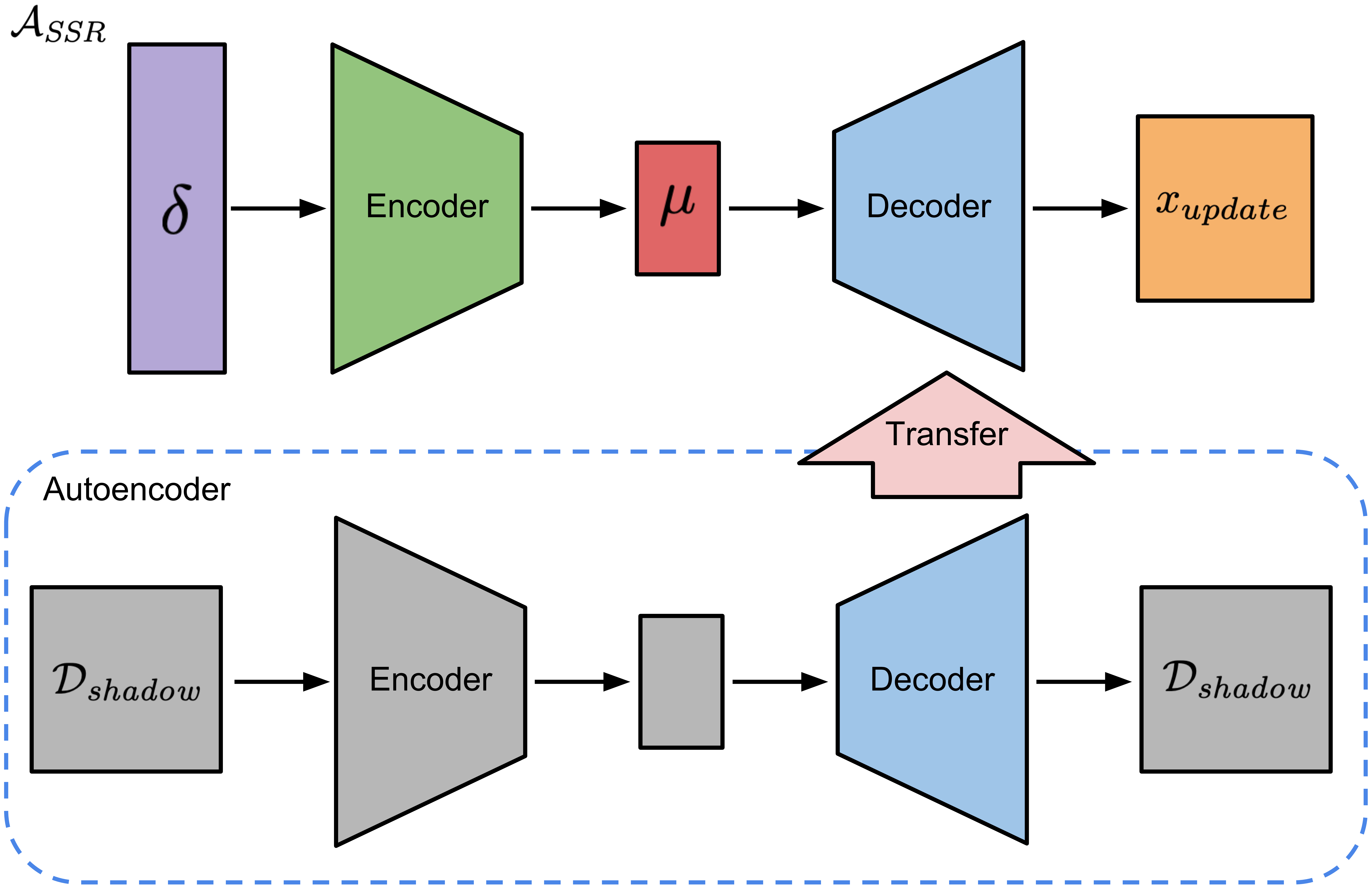
Single-sample Reconstruction

- More complicated than inferring label
- Attacker needs a sample generator
  - We rely on autoencoder's decoder

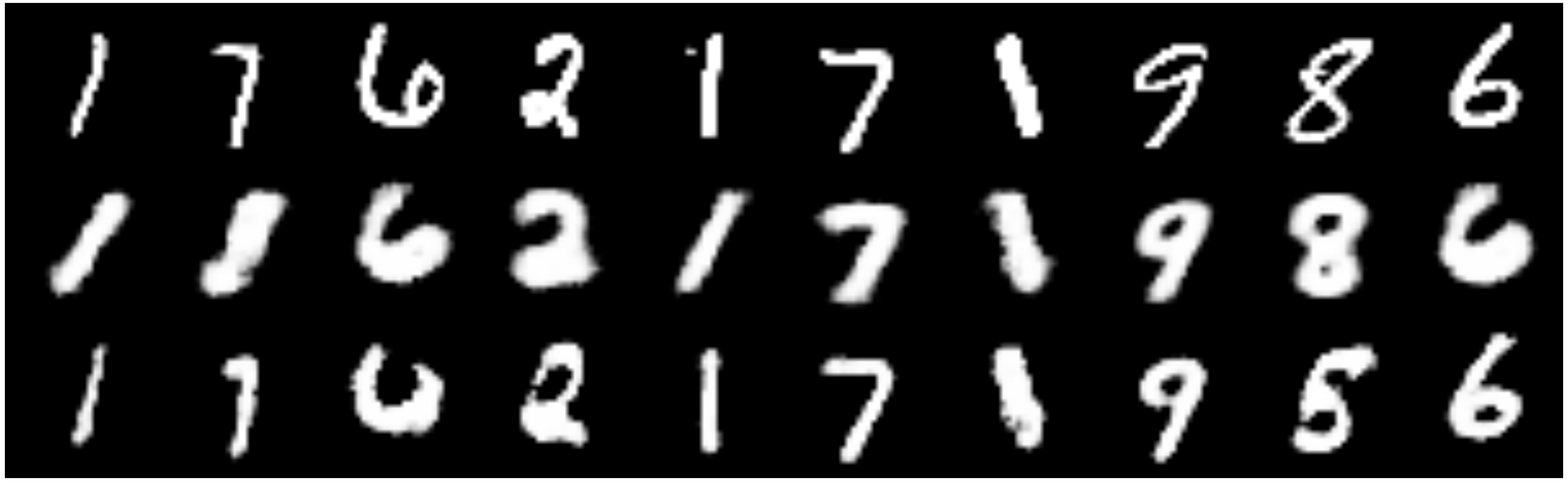
Autoencoder

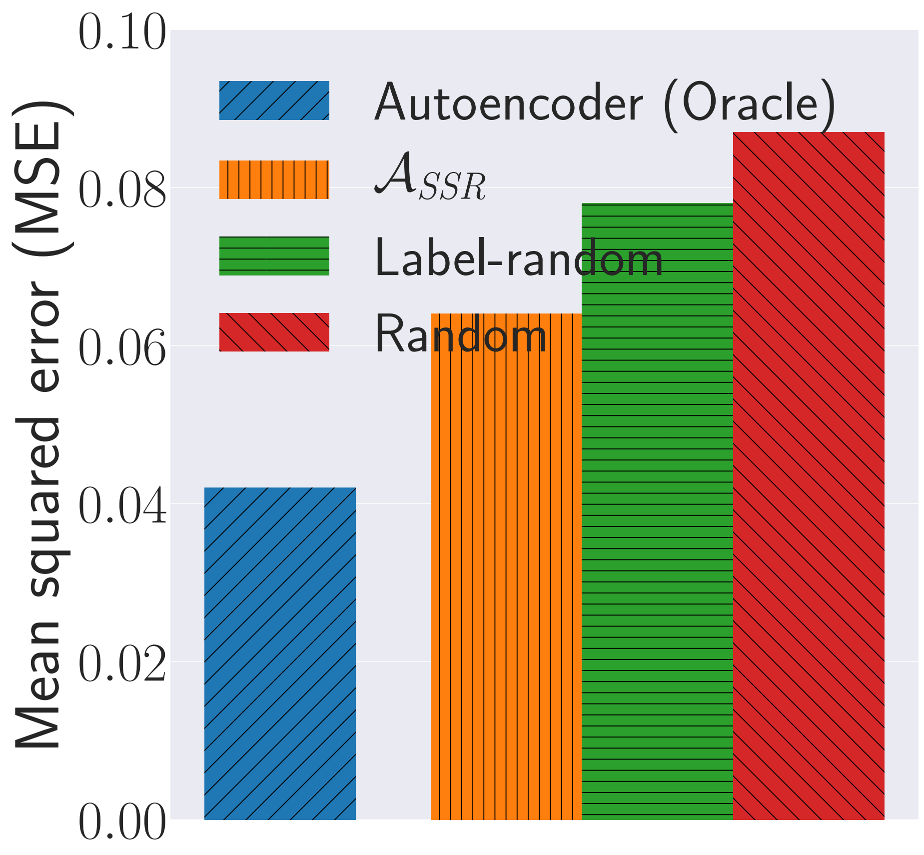


Single-sample Reconstruction

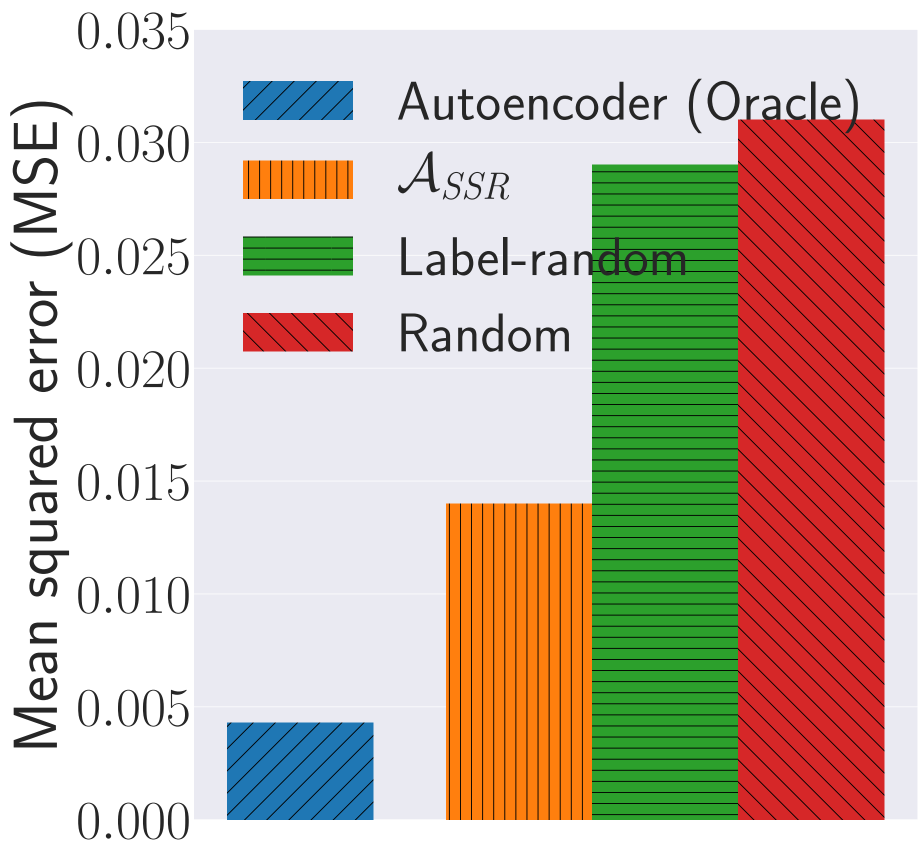


Single-sample Reconstruction





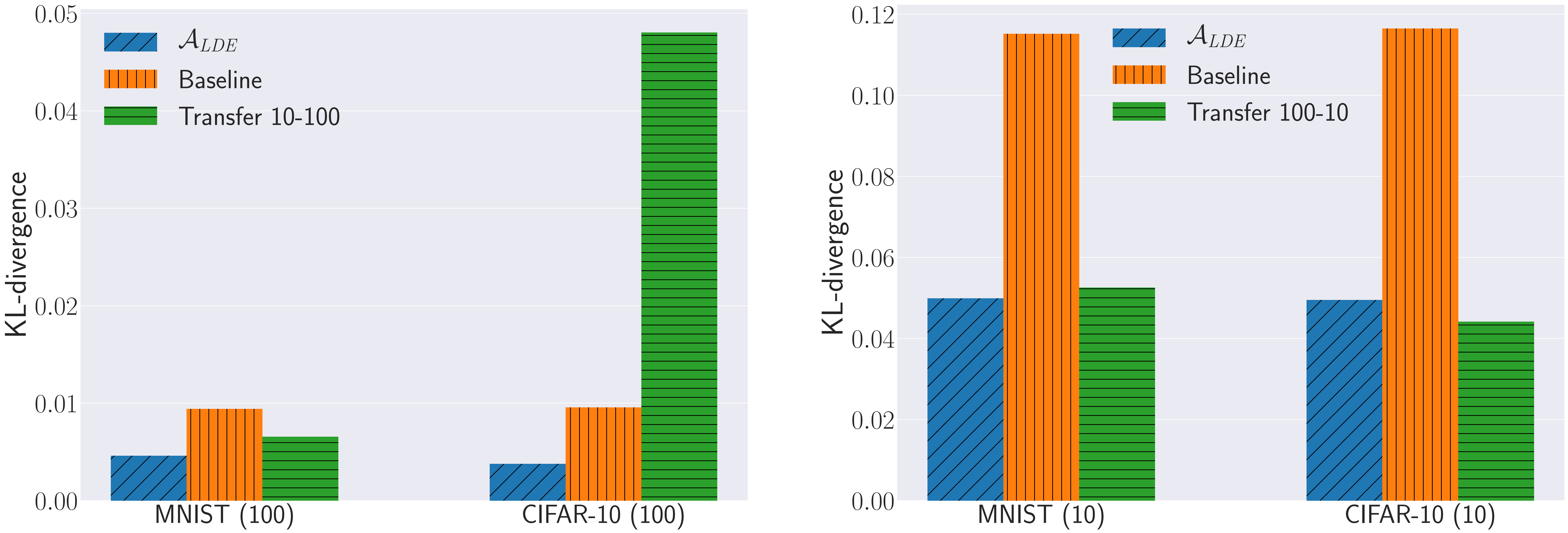
CIFAR-10



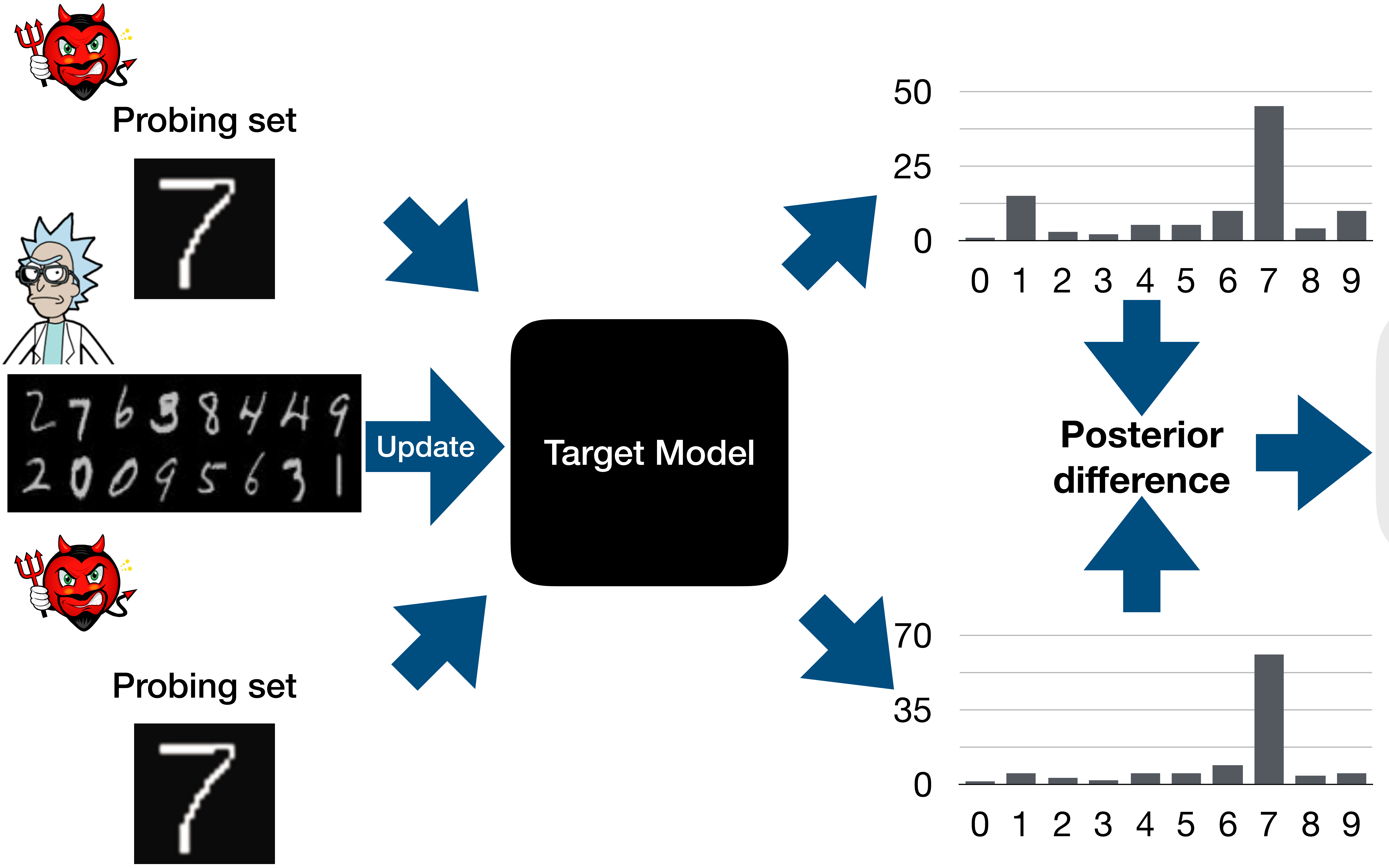
MNIST

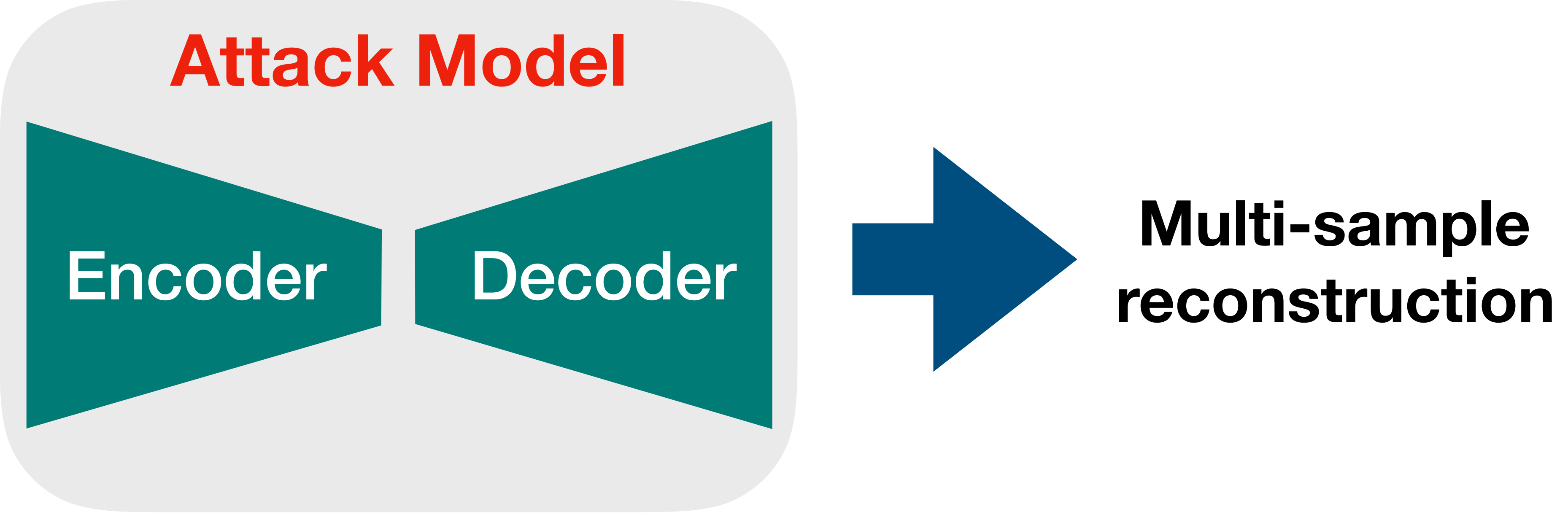
Multi-sample Label Estimation

Multi-sample Label Estimation



Multi-sample Reconstruction





- Most challenging scenario in this attack scenario
- Reconstruct a set of data samples
  - Autoencoder cannot help anymore
- What we do?

Generative Adversarial Network (GAN)

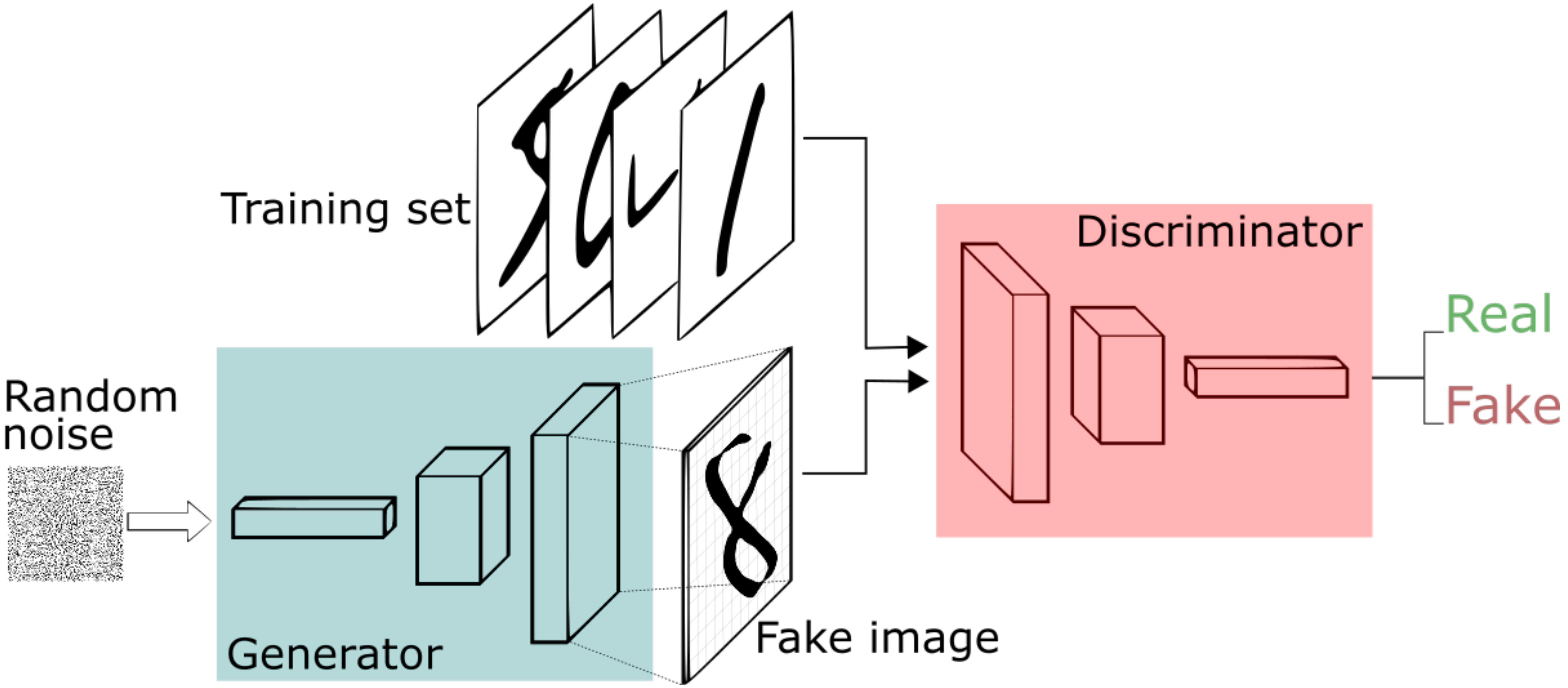
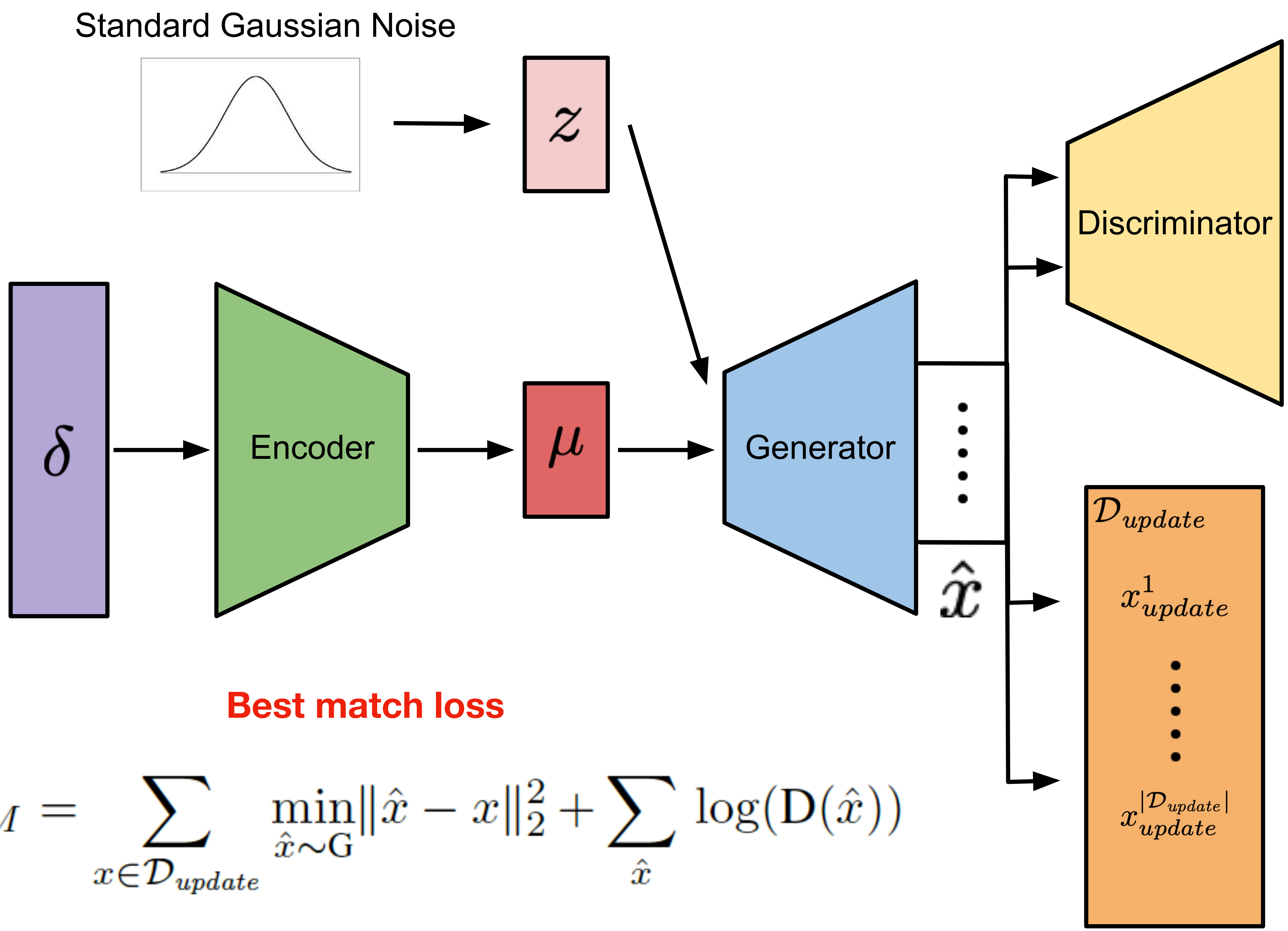


Image credit: *Thalles Silva*

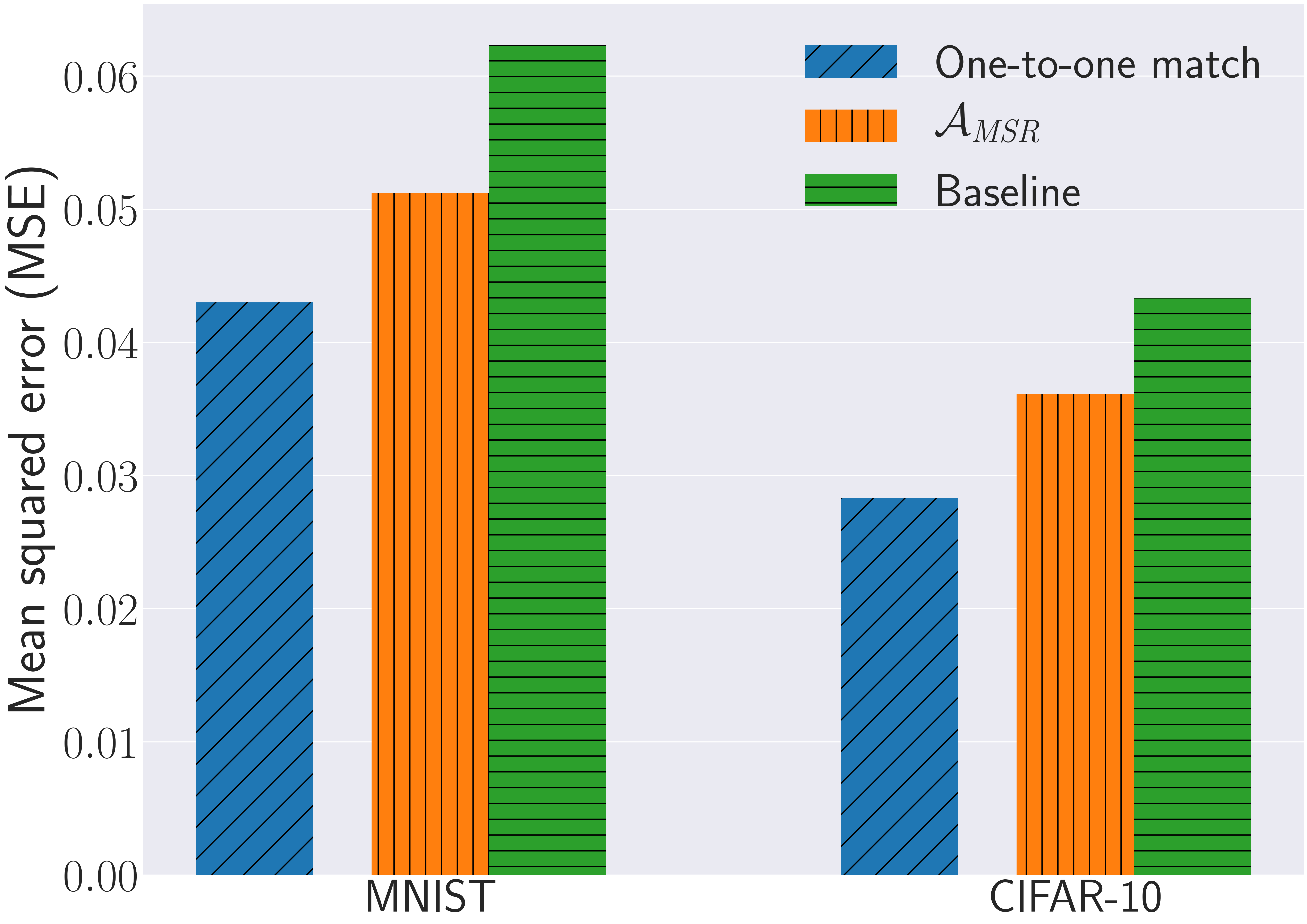
Multi-sample Reconstruction



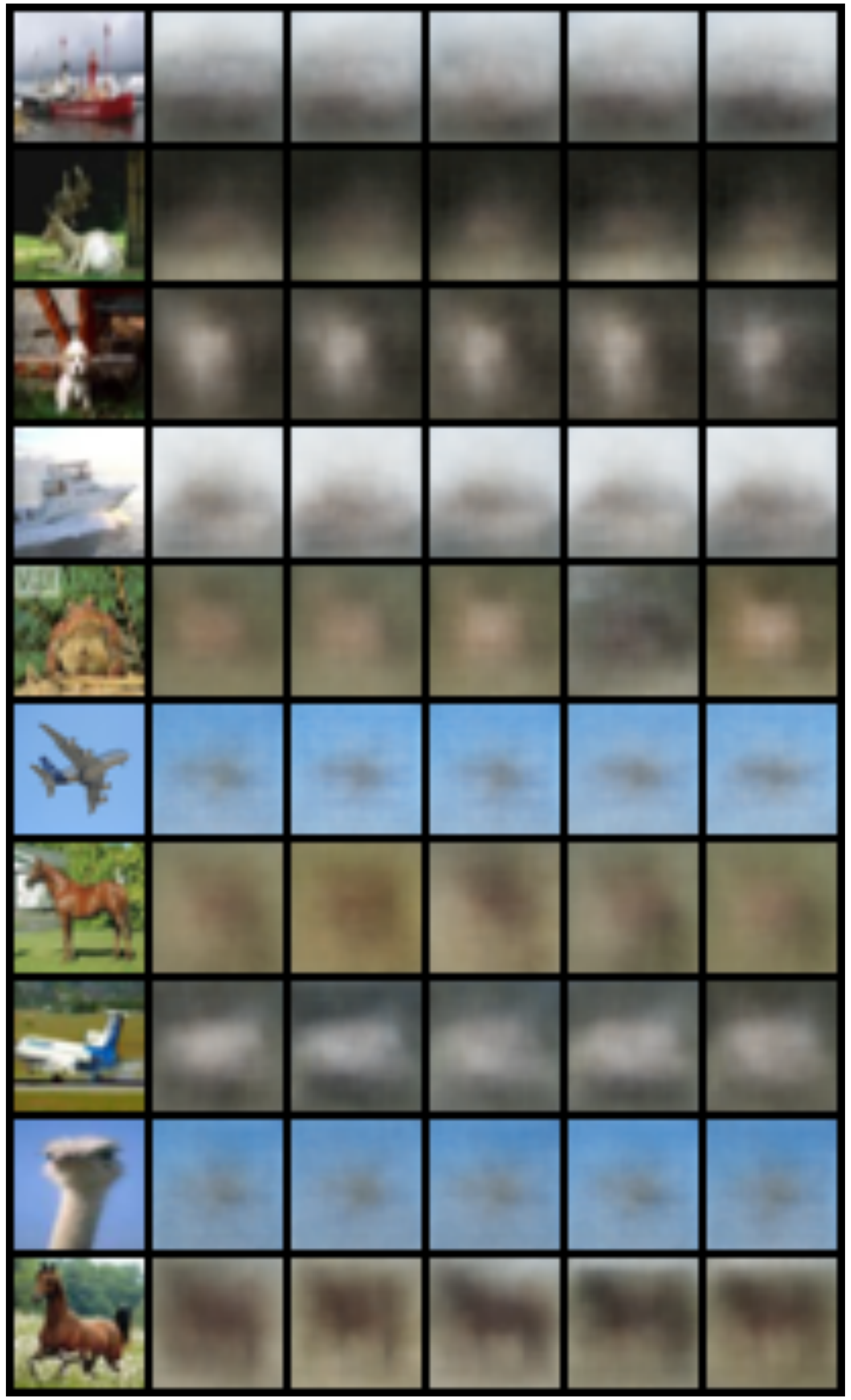
Best match loss

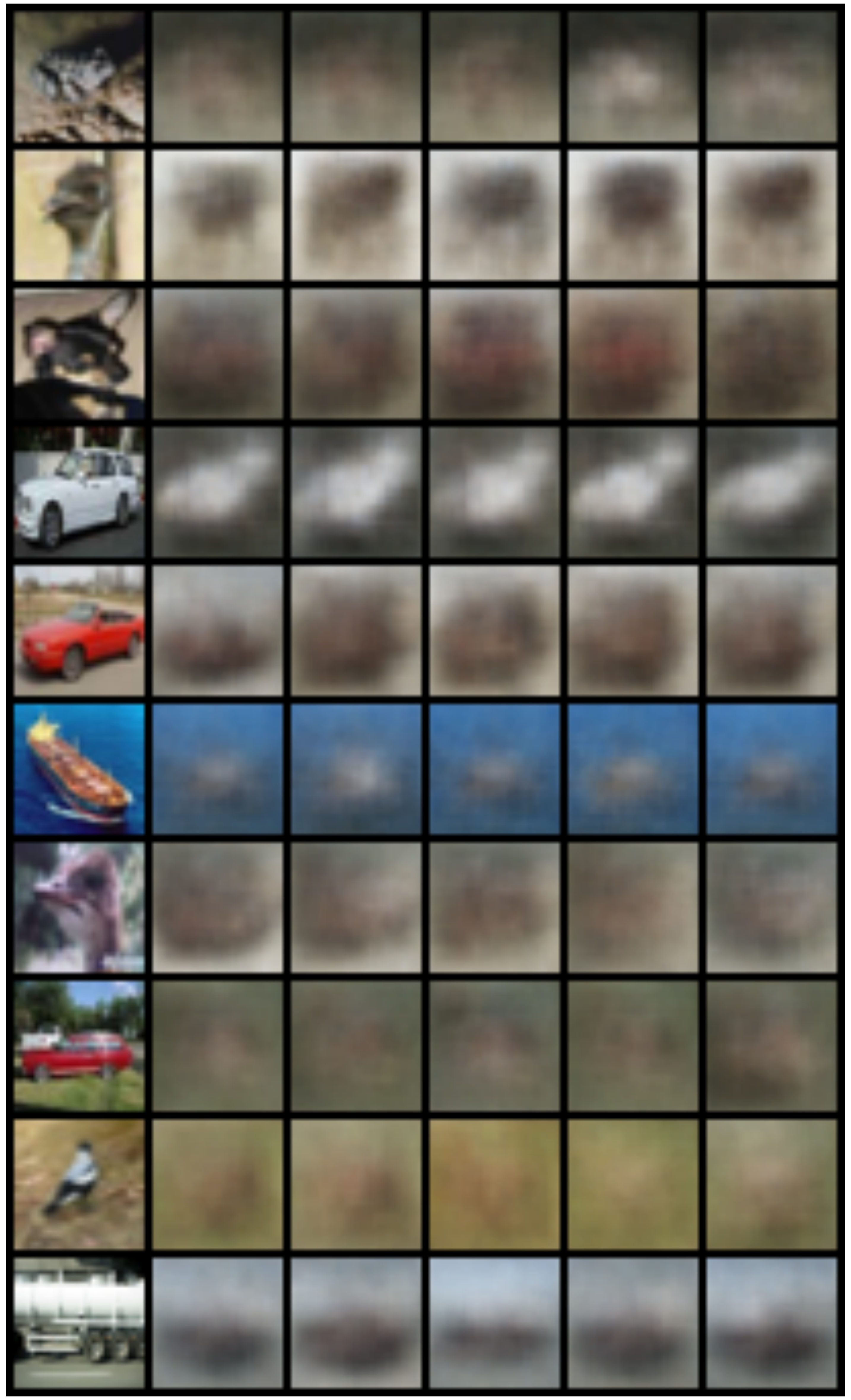
$$\mathcal{L}_{BM} = \sum_{x \in \mathcal{D}_{update}} \min_{\hat{x} \sim \mathrm{G}} \|\hat{x} - x\|_2^2 + \sum_{\hat{x}} \log(\mathrm{D}(\hat{x}))$$

Multi-sample Reconstruction



Multi-sample Reconstruction





Multi-sample Reconstruction

Summary

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Thank you for your attention!
Questions?

ahmed.salem@cispa.saarland
https://ahmedsalem2.github.io/
@AhmedGaSalem