ML in PL
CONFERENCE 2023

Bucks for Buckets (B4B): Active Defence Against Stealing Encoders

Jan Dubiński, Stanisław Pawlak, Franziska Boenisch, Tomasz Trzciński, Adam Dziedzic

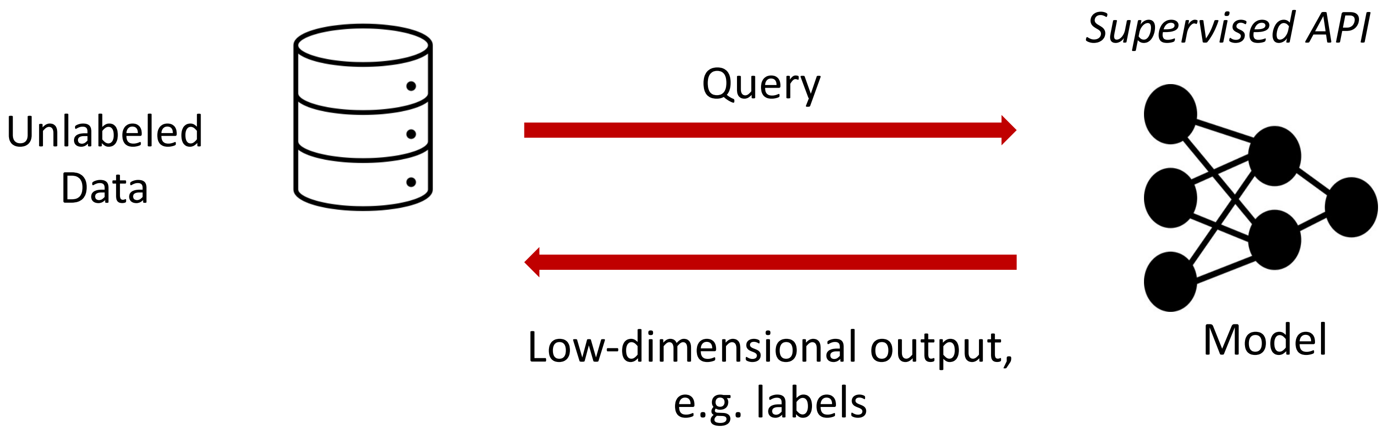
Warsaw University of Technology
CVLab@PW

IDEAS NCBR

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

TOOPLOOX

Supervised vs Self-Supervised API



Supervised API

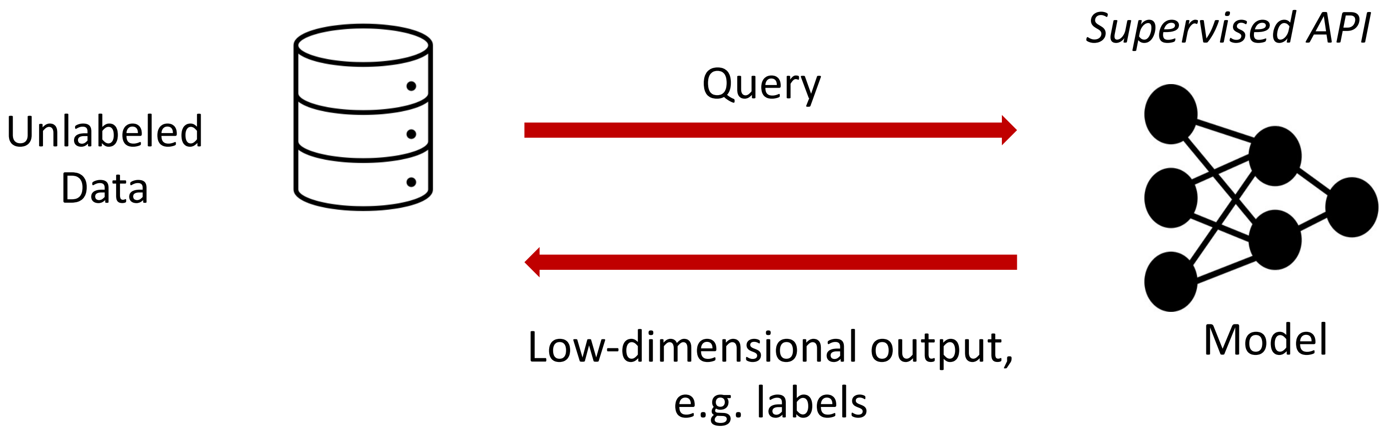
Unlabeled Data

Query

Low-dimensional output, e.g. labels

Model

Supervised vs Self-Supervised API



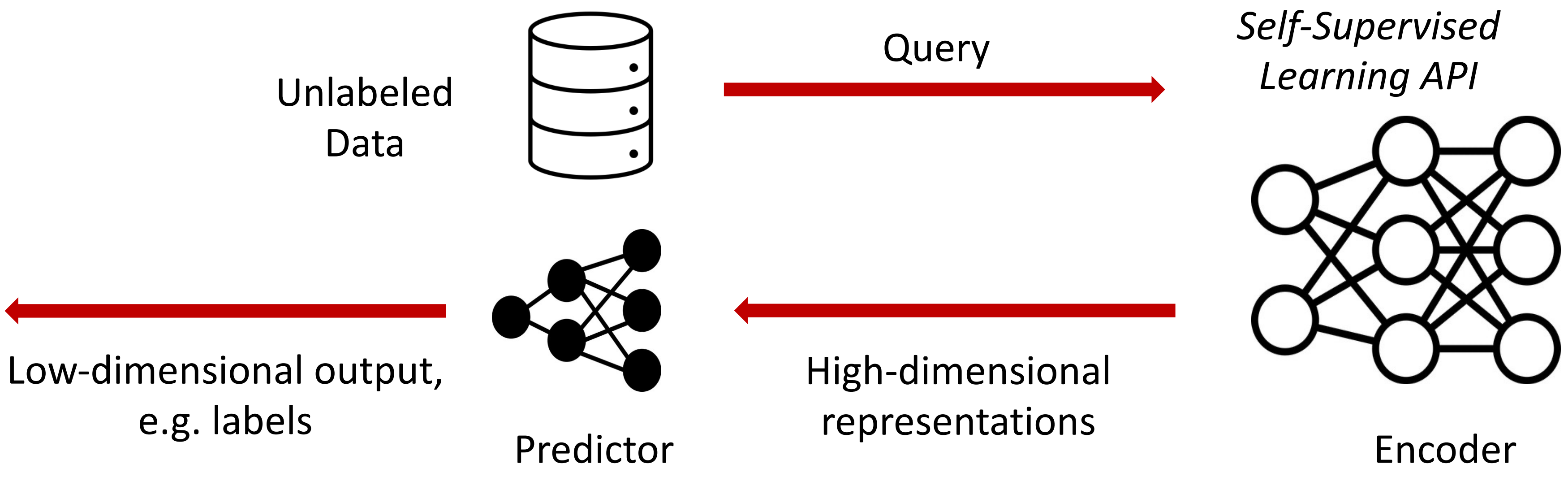
Unlabeled Data

Query

Supervised API

Model

Low-dimensional output, e.g. labels



Unlabeled Data

Query

Self-Supervised Learning API

Encoder

High-dimensional representations

Predictor

Low-dimensional output, e.g. labels

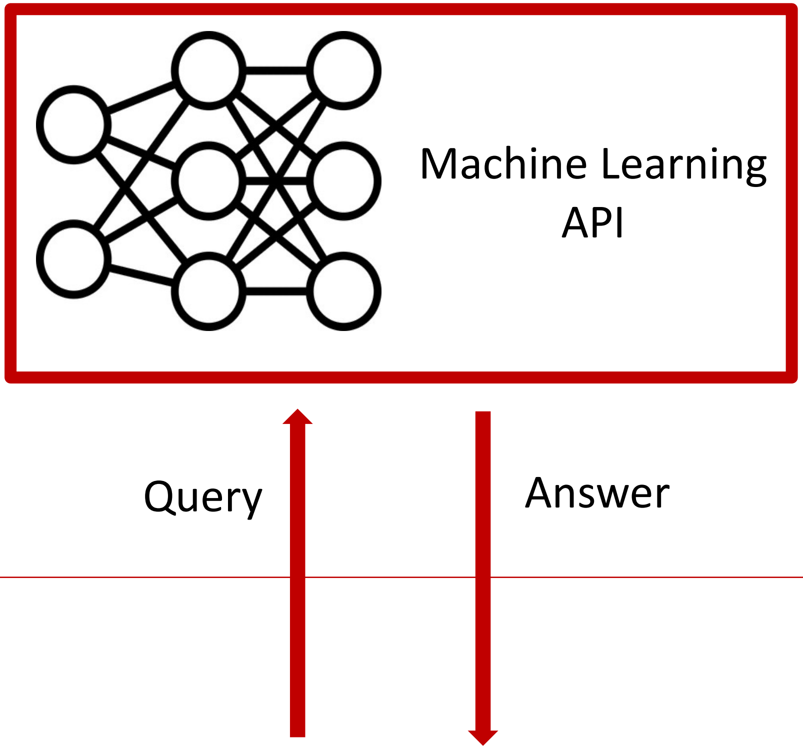
Cost of training an Encoder

Collect and clean data

Tune the Hyperparameters

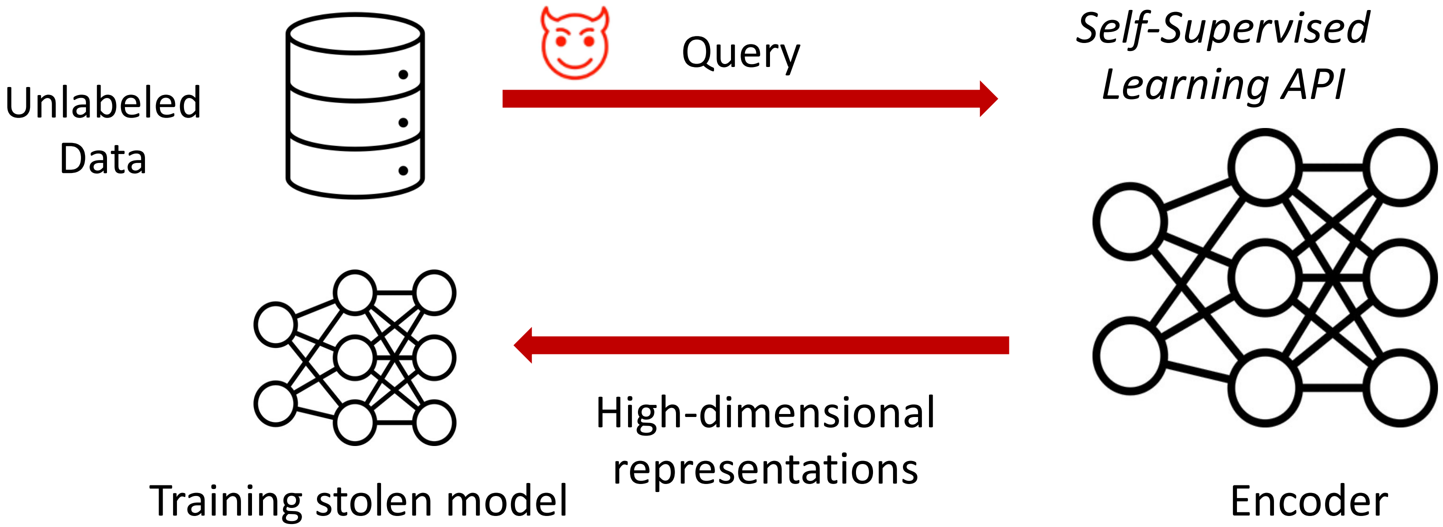
Run on GPU/TPU

$12 GPT-3

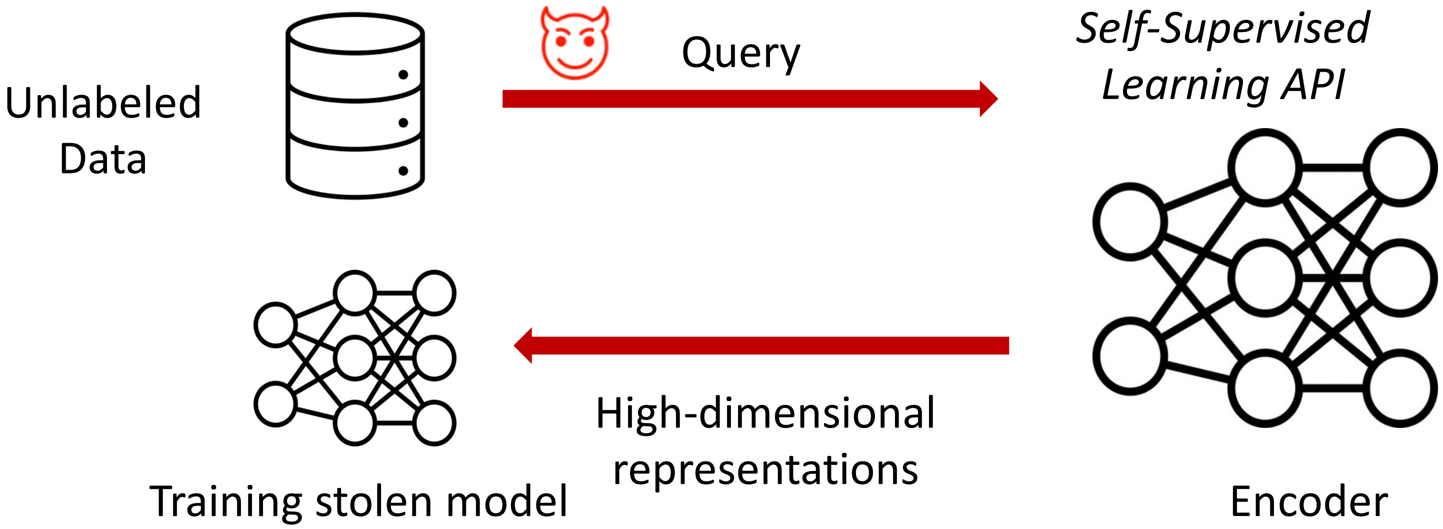


Model stealing is ranked among the most sever attack against ML models

Stealing Encoder models

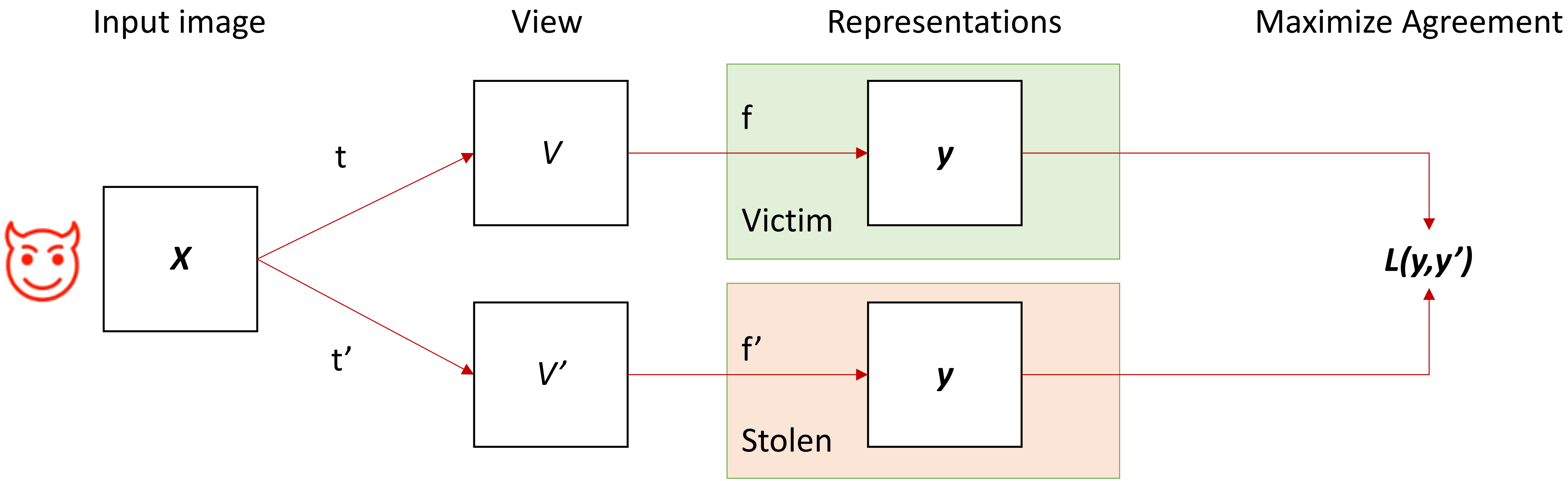

Unlabeled Data
Query
Self-Supervised Learning API
Training stolen model
High-dimensional representations
Encoder

Stealing Encoder models

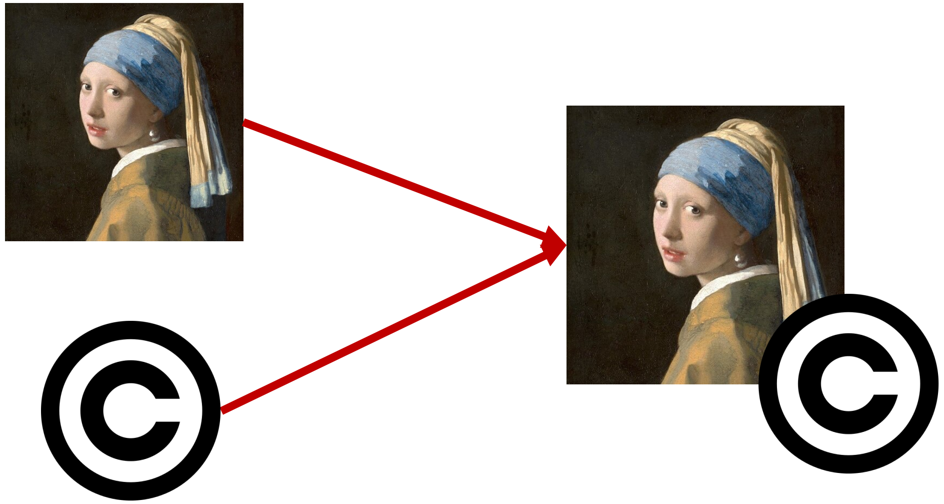

Unlabeled Data
Query
Self-Supervised Learning API
Encoder
High-dimensional representations
Training stolen model

1. **Stealing SSL models is query efficient**
2. **Existing defences for supervised models are inadequate for SSL models**

How to steal an Encoder?

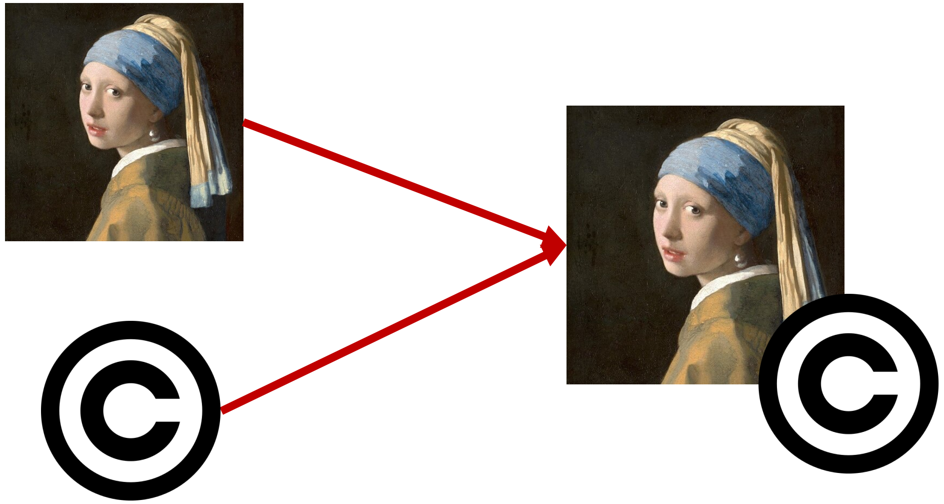


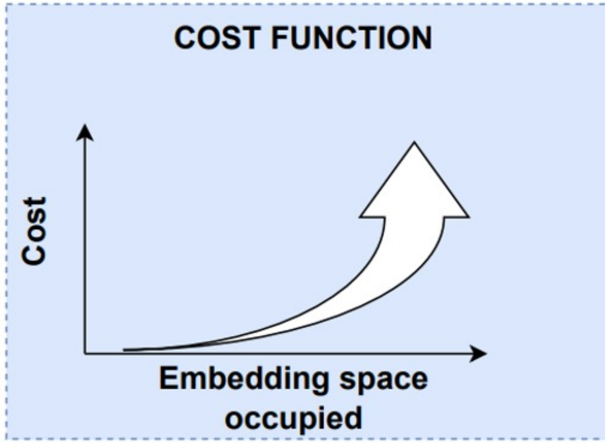
Defenses against Encoder Stealing



Till now: only Ownership Resolution for Encoders Like Watermarking

Defenses against Encoder Stealing

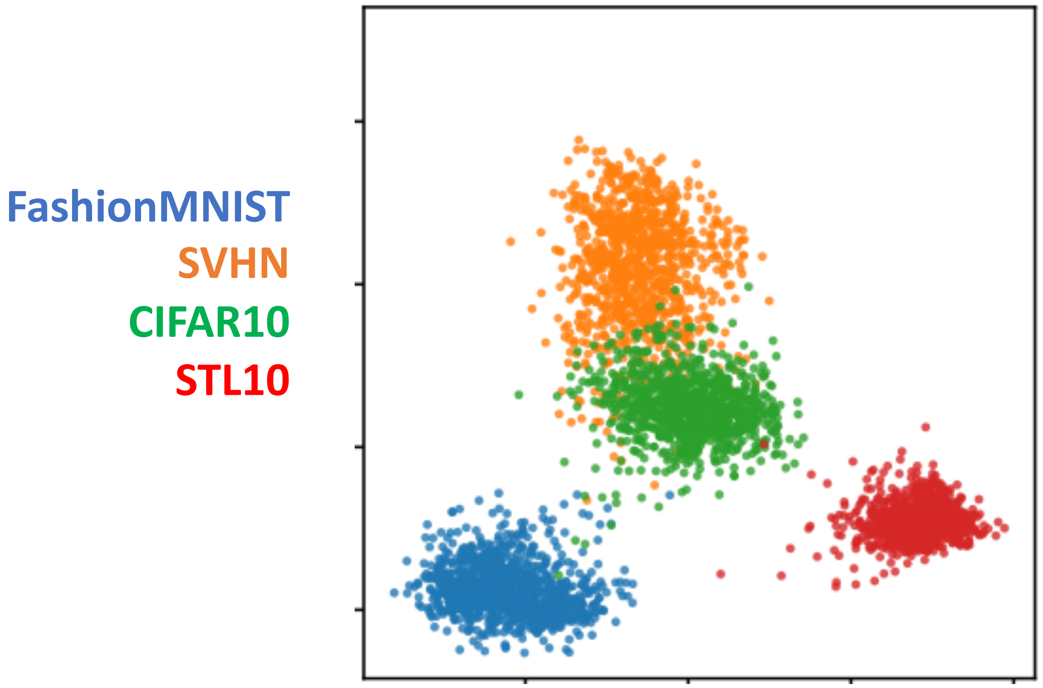




Till now: only Ownership Resolution for Encoders Like Watermarking

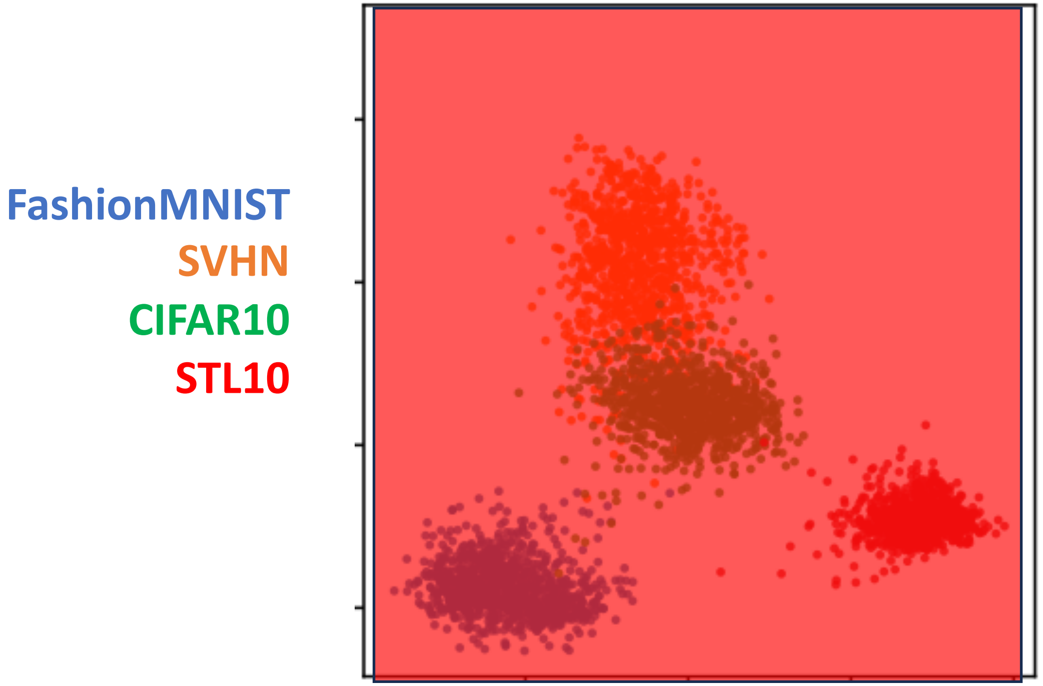
Our first Active Defense Against Encoder Stealing

Occupation of the representation space



Queries from legtimate users occupy a single region of the latent space

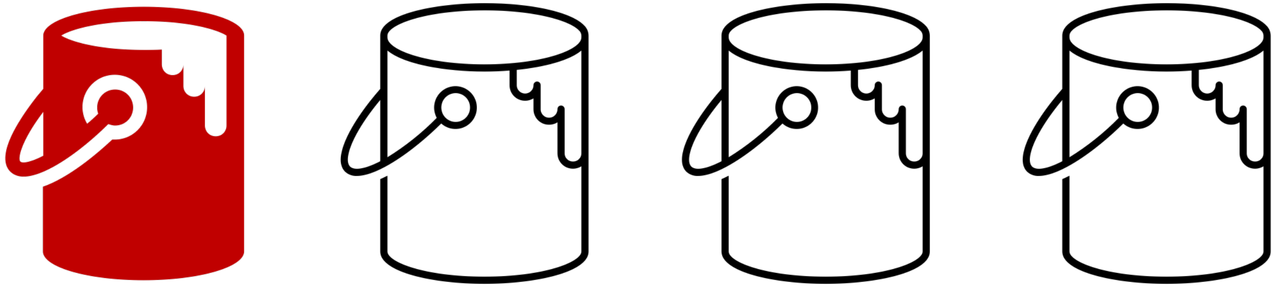
Occupation of the representation space



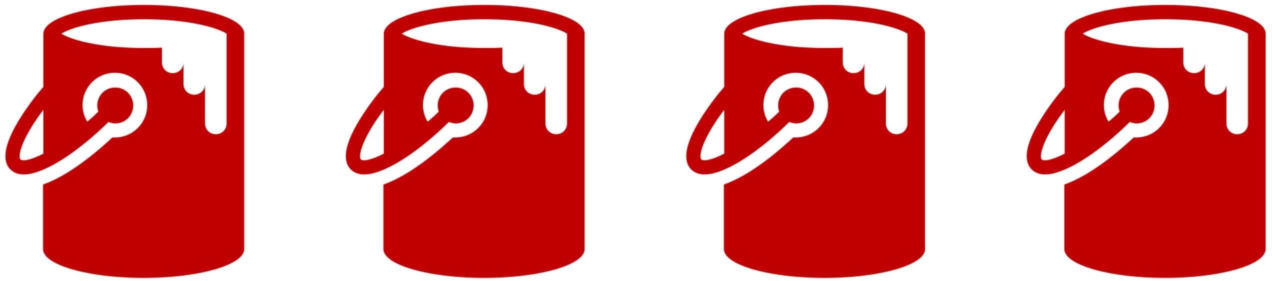
Attacker must query the entire representation space to steal the encoder

Bucks for Buckets

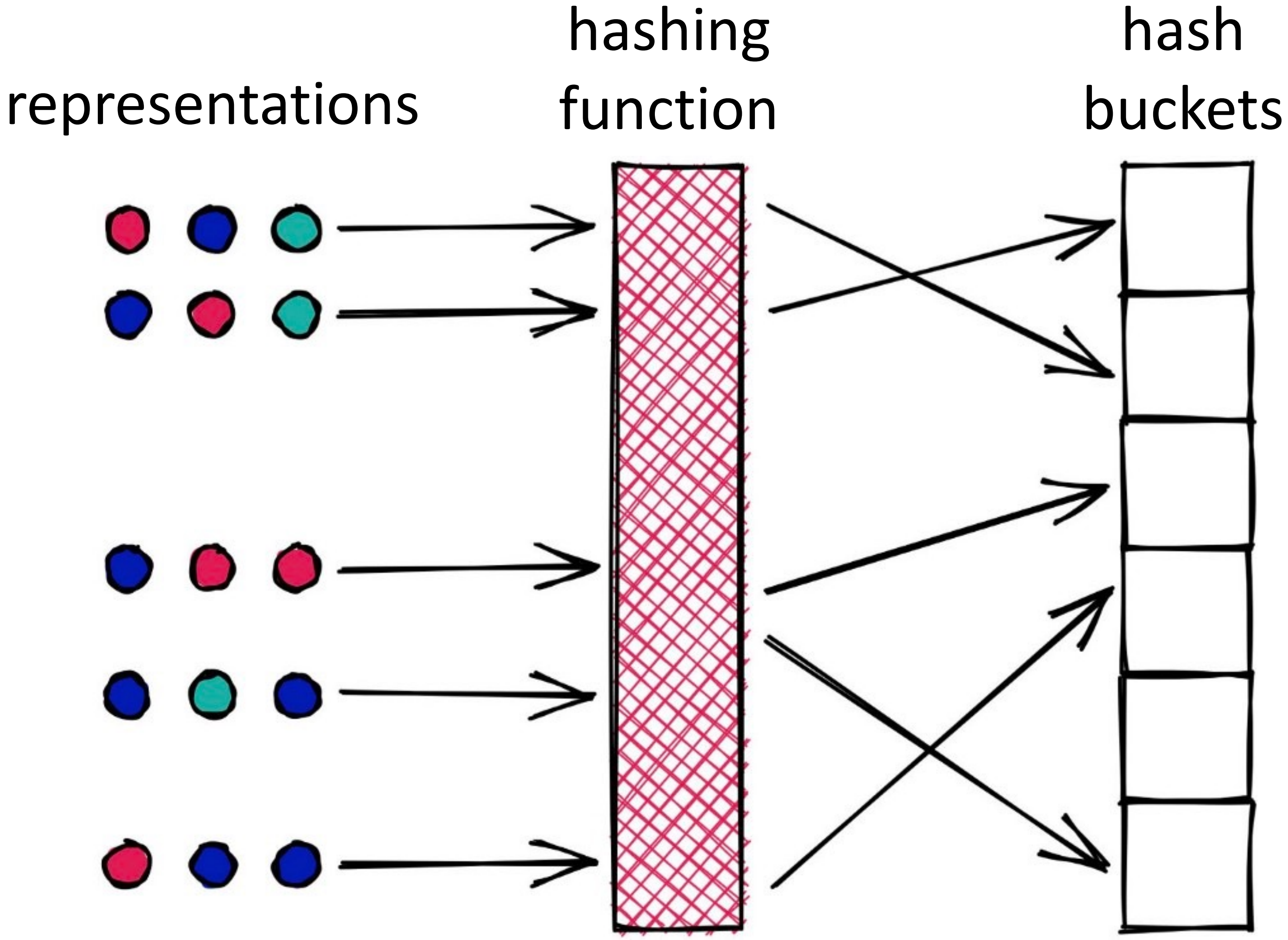
Cost: 1$



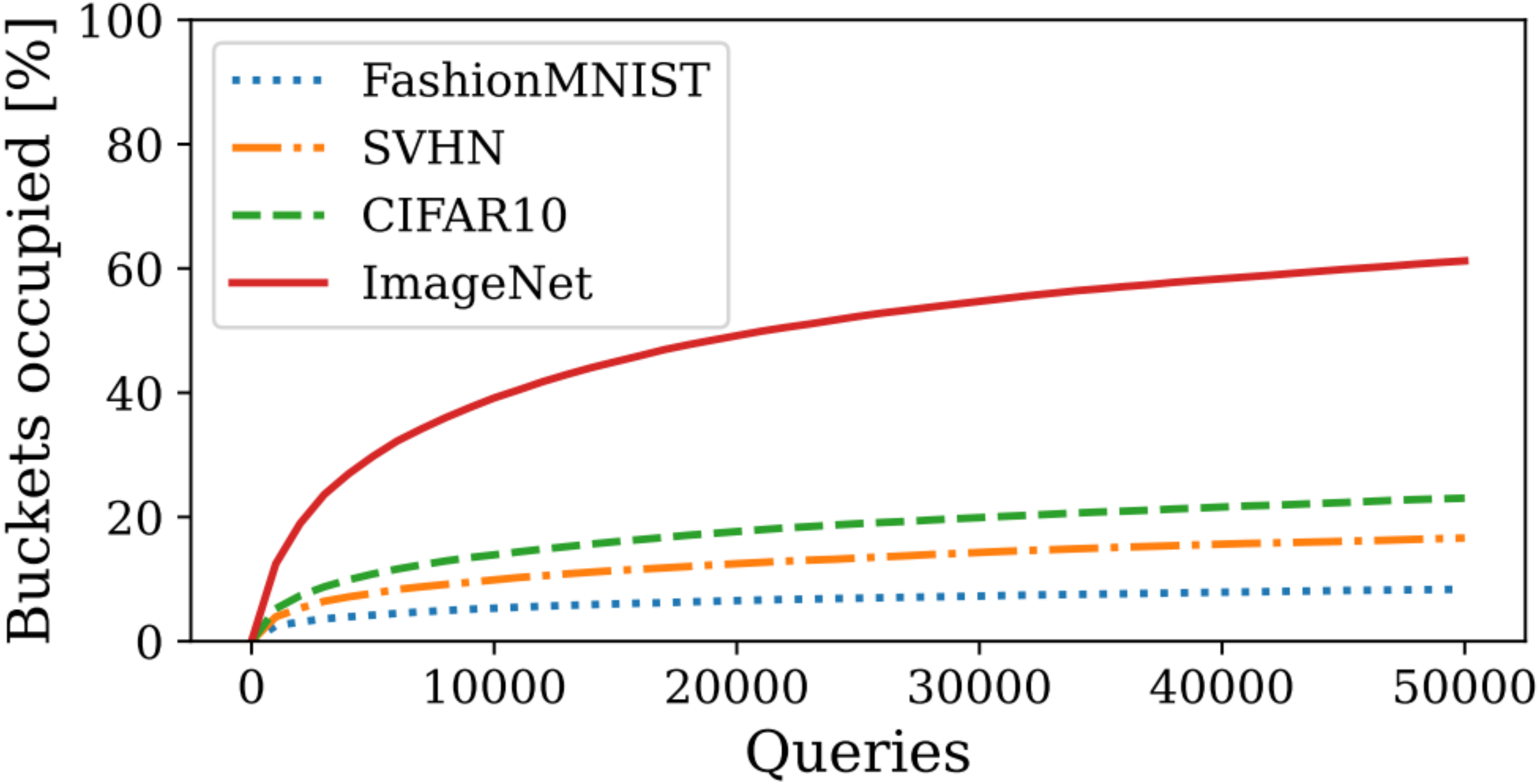
Cost: 1.000.000$



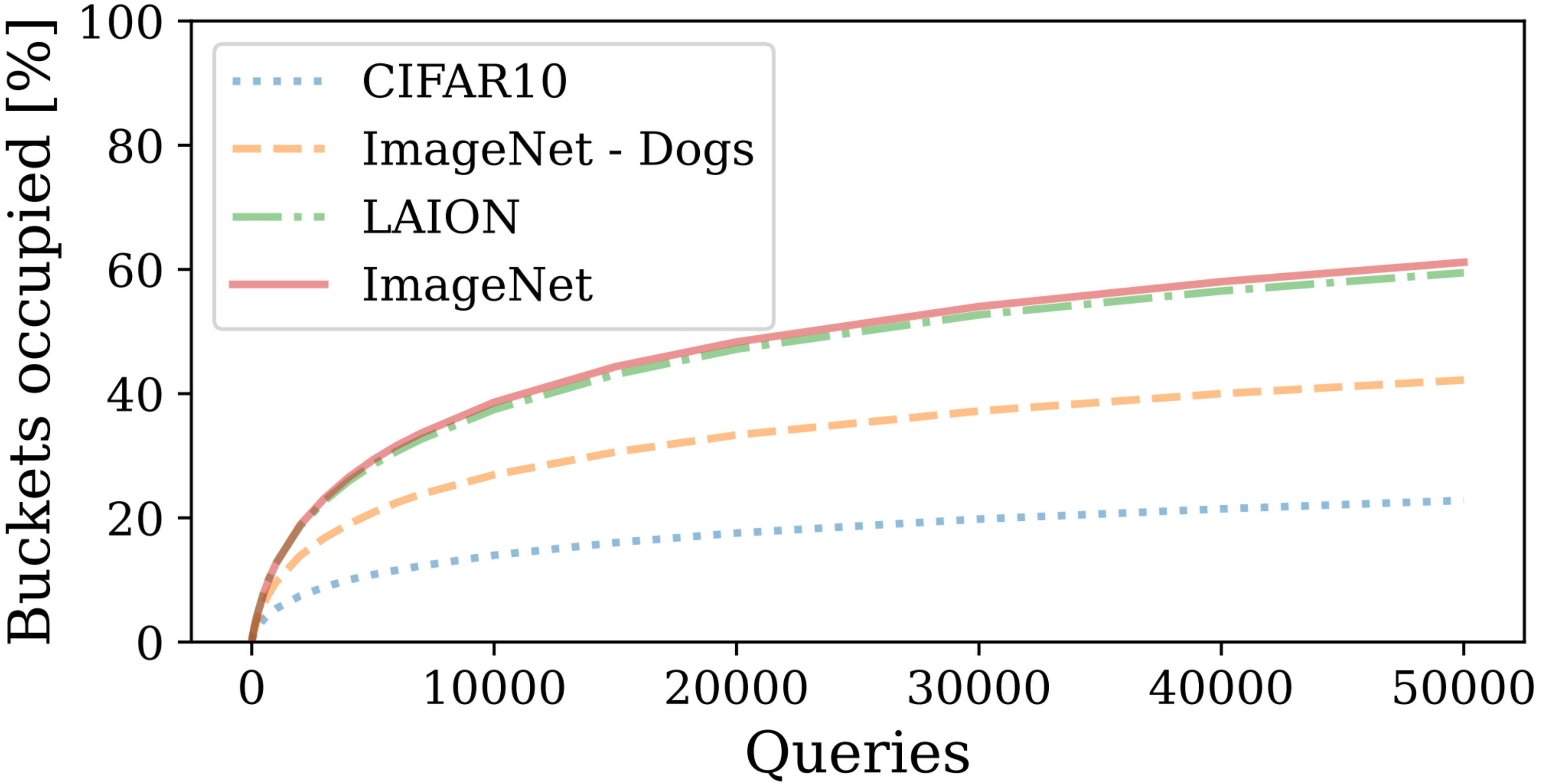
Measuring the coverage of the latent space



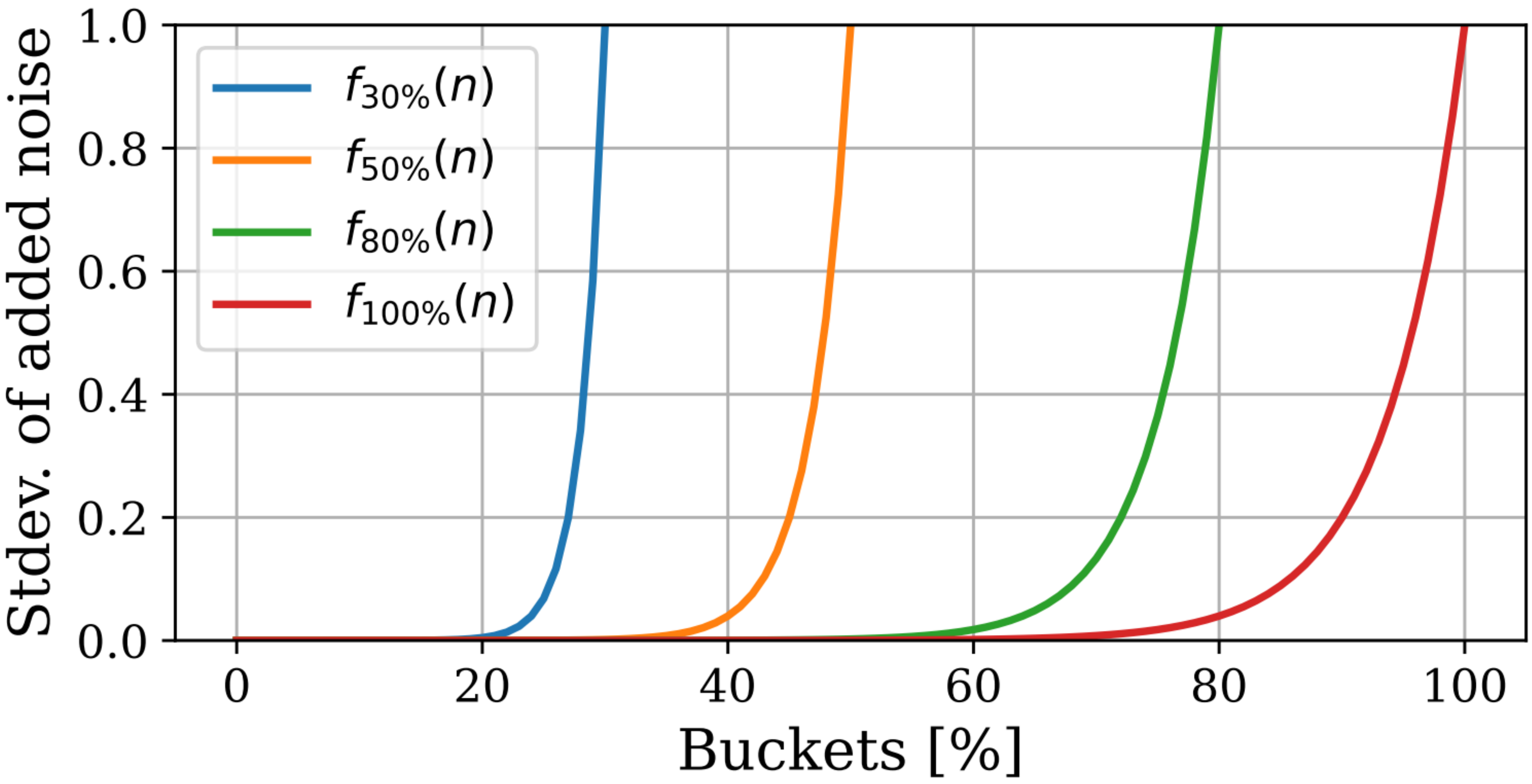
Queries Sent vs Buckets Occupied



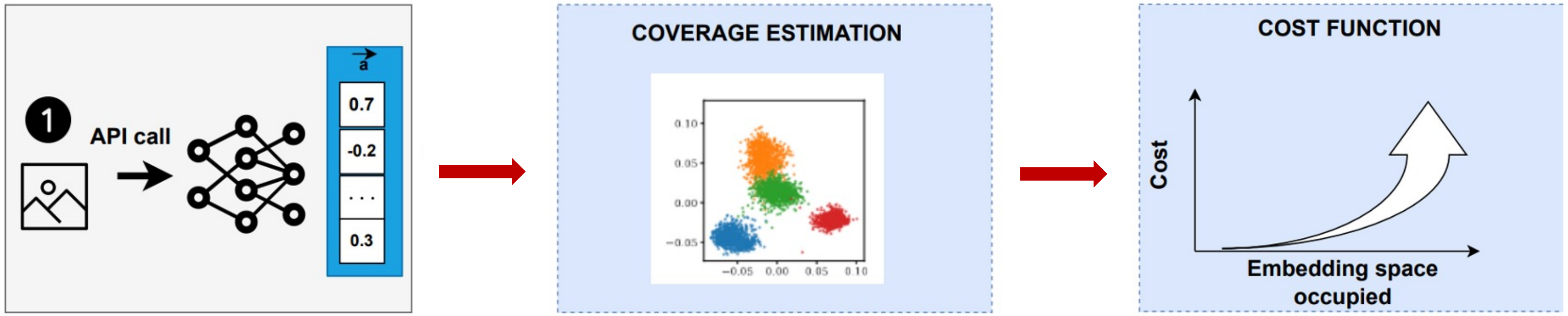
Queries Sent vs Buckets Occupied



Cost function



Active Defense Framework

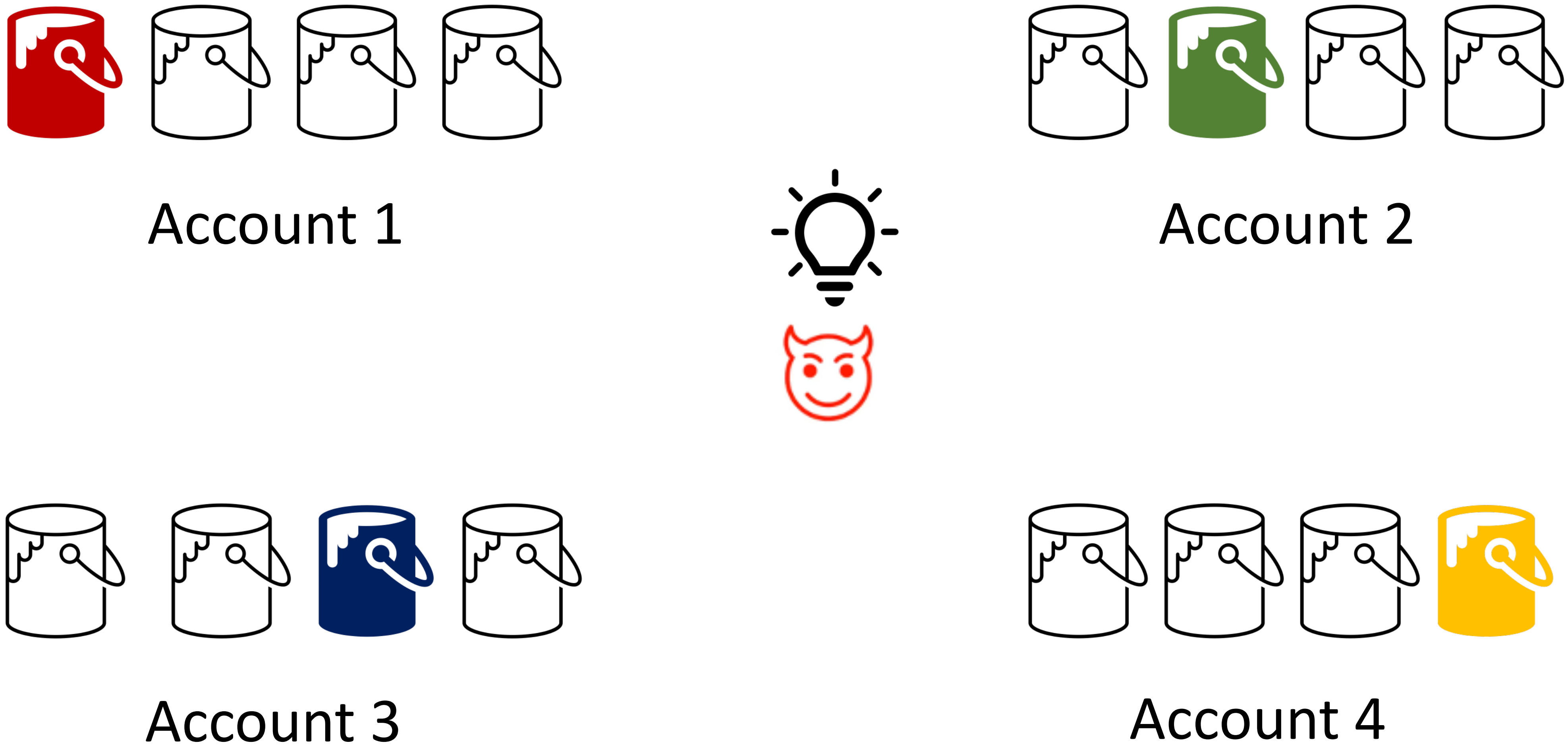


1) compute representations for incoming queries

2) estimate the coverage of the representation space

3) compute query cost

Sybil attacks



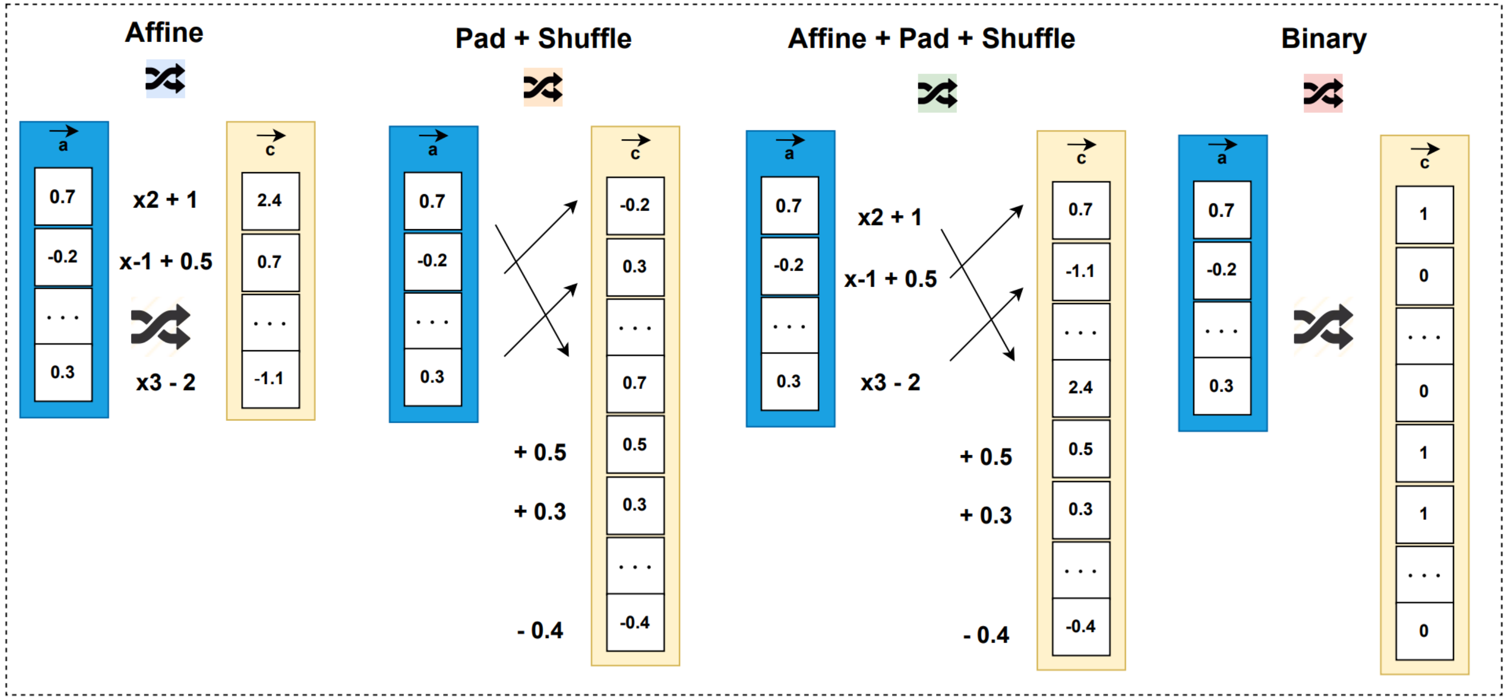
Account 1

Account 2

Account 3

Account 4

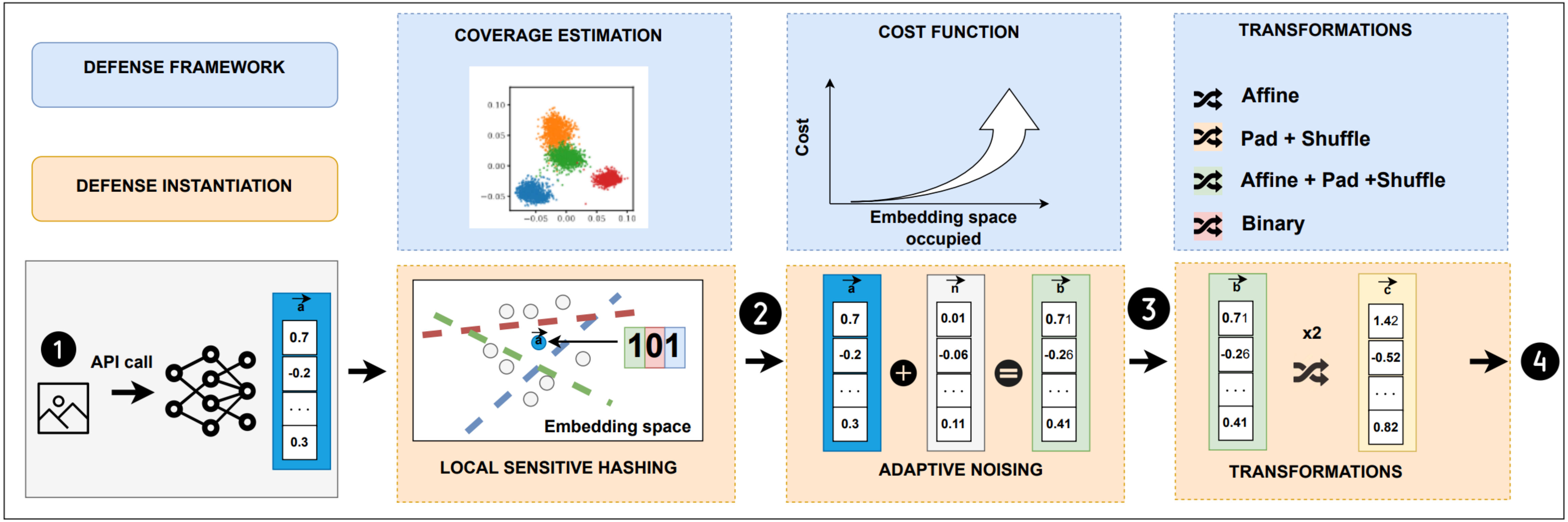
Defence against a Sybil Attack



1. **Users receive transformed representations**
2. **Transformations:**

- **maintain utility of the representations**
- **prevent using multiple sybil accounts to train a stolen model**

End-to-End solution



B4B - no utility drop for legitimate users

Queries	CIFAR10	STL10	F-MNIST
None	90.41	95.08	91.22
B4B	**90.24**	**95.05**	**91.70**

Undefended encoder is easy to steal

Queries	CIFAR10	STL10	F-MNIST
50k	65.2	64.9	88.5
100k	**68.1**	**63.1**	**89.5**

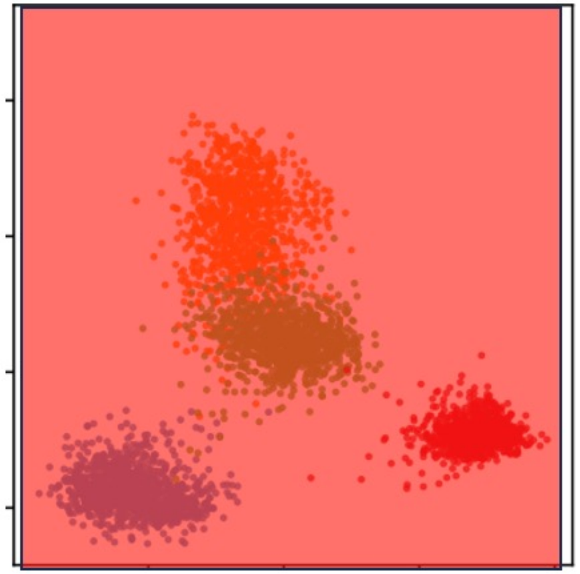
B4B – signifcant performance drop for the attacker

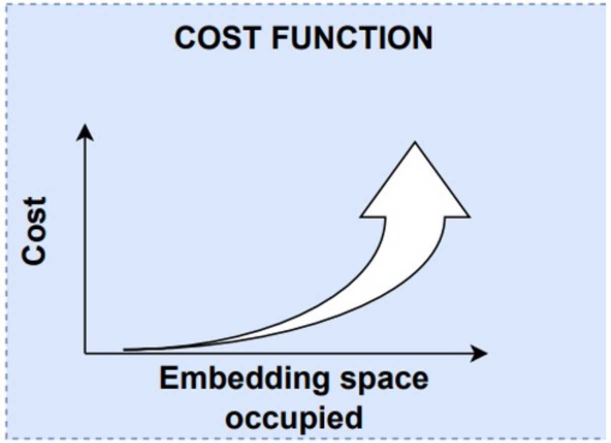
Queries	CIFAR10	STL10	F-MNIST
50k	35.72	31.54	70.01
100k	**12.01**	**13.94**	**69.63**

B4B – successfully prevents Sybil Attacks

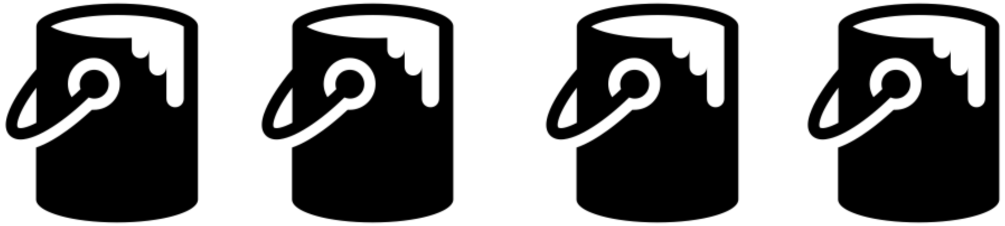
Sybils	CIFAR10	STL10	F-MNIST
2	39.56	38.50	77.01
3	33.87	38.57	72.95
4	33.98	34.52	70.71
5	32.65	32.45	70.12

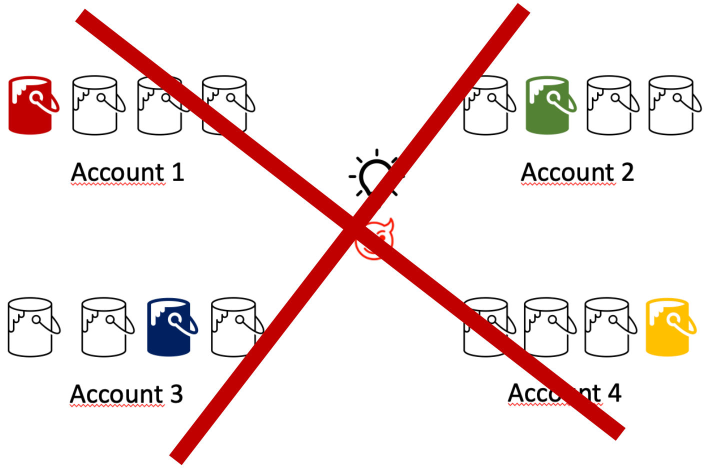
B4B: Defend against Encoder Stealing





Cost: 1 000 000$




Account 1
Account 2
Account 3
Account 4