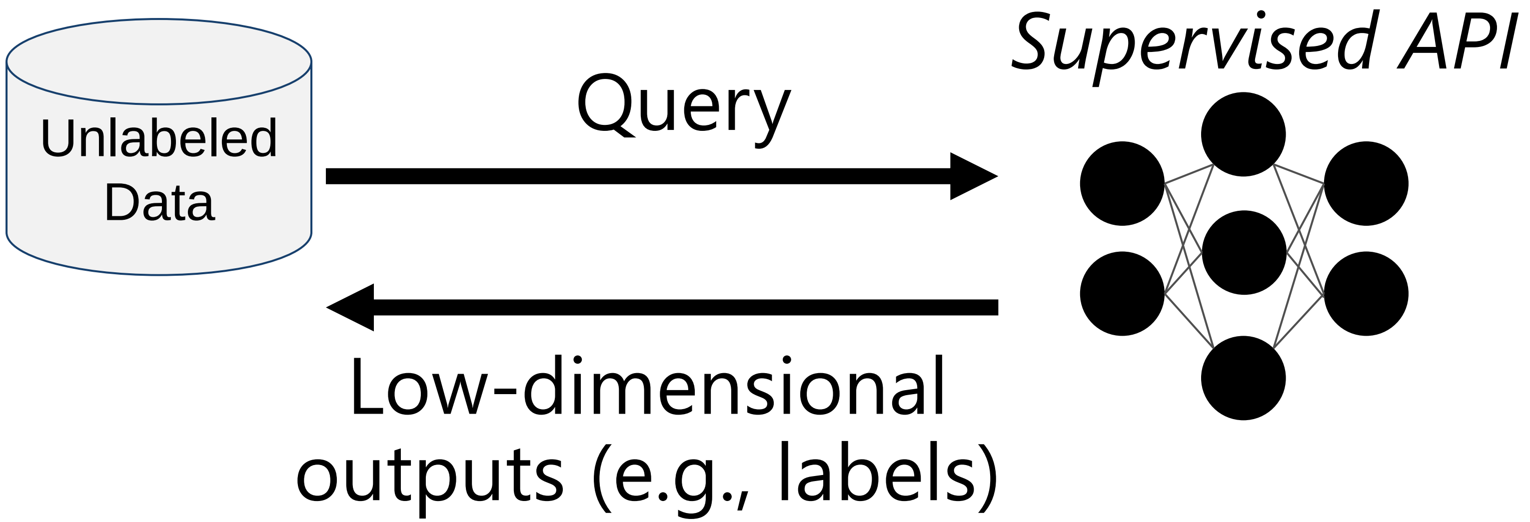
On the Difficulty of Defending Self-Supervised Learning against Model Extraction

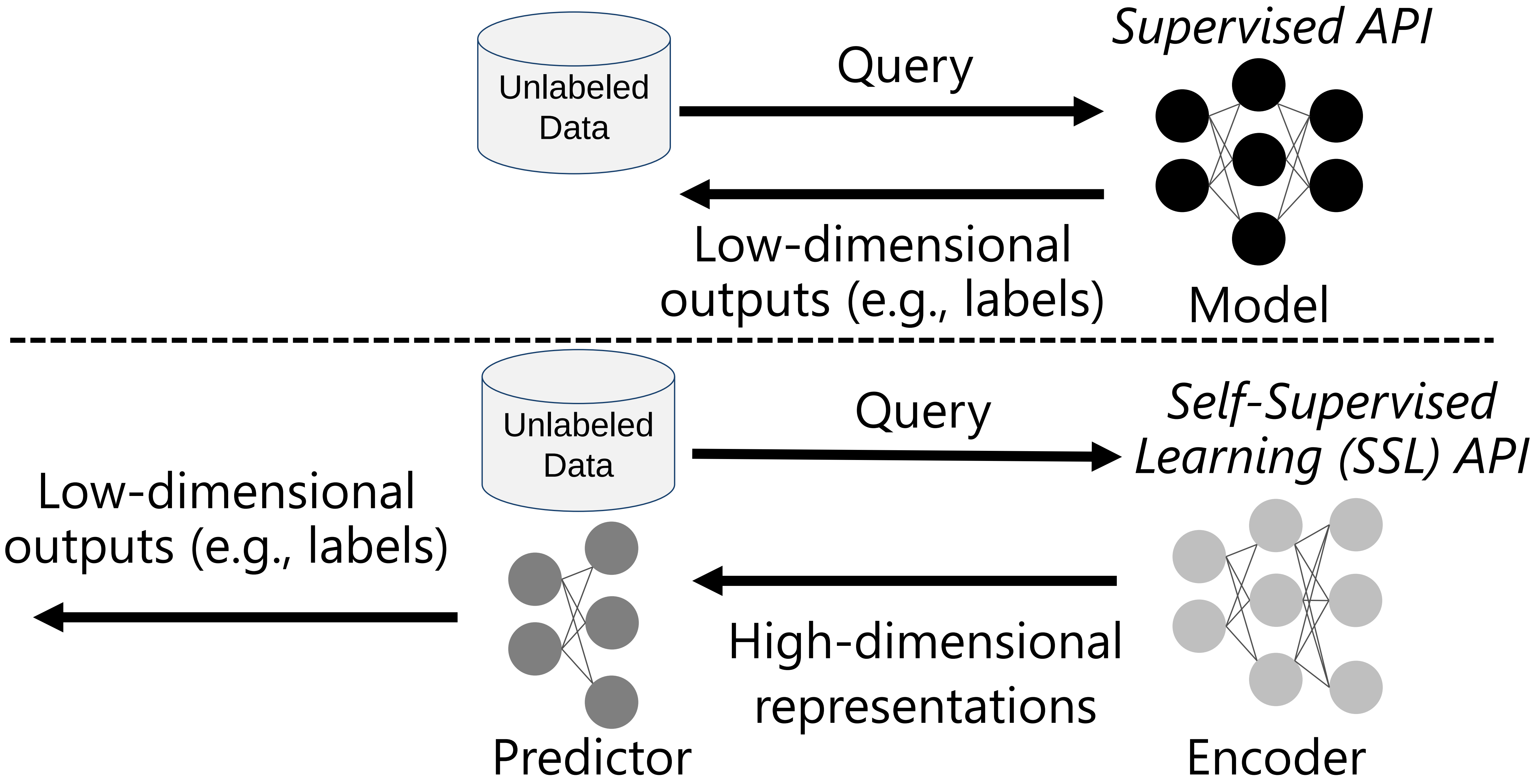
Adam Dziedzic, Nikita Dhawan, Muhammad Ahmad Kaleem, Jonas Guan, Nicolas Papernot

International Conference on Machine Learning (ICML)
July 17th - 23rd, 2022

Supervised Learning API



Supervised vs Self-Supervised Learning APIs

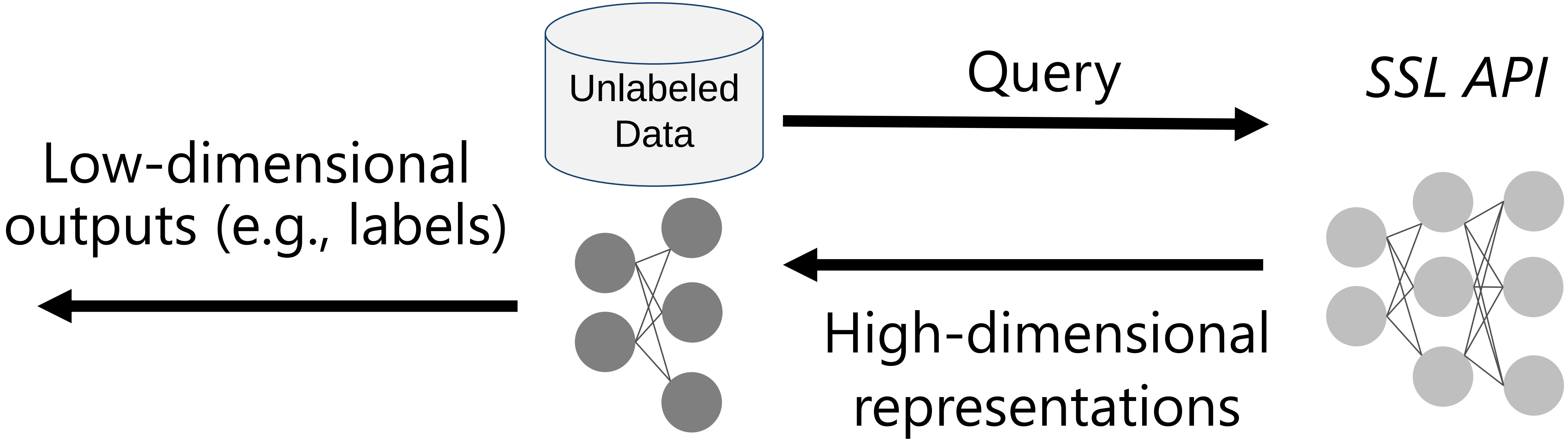


High Cost of Creating Self-Supervised APIs

Collect Data

Tune Hyper-parameters

Run on GPU/TPU/CPU

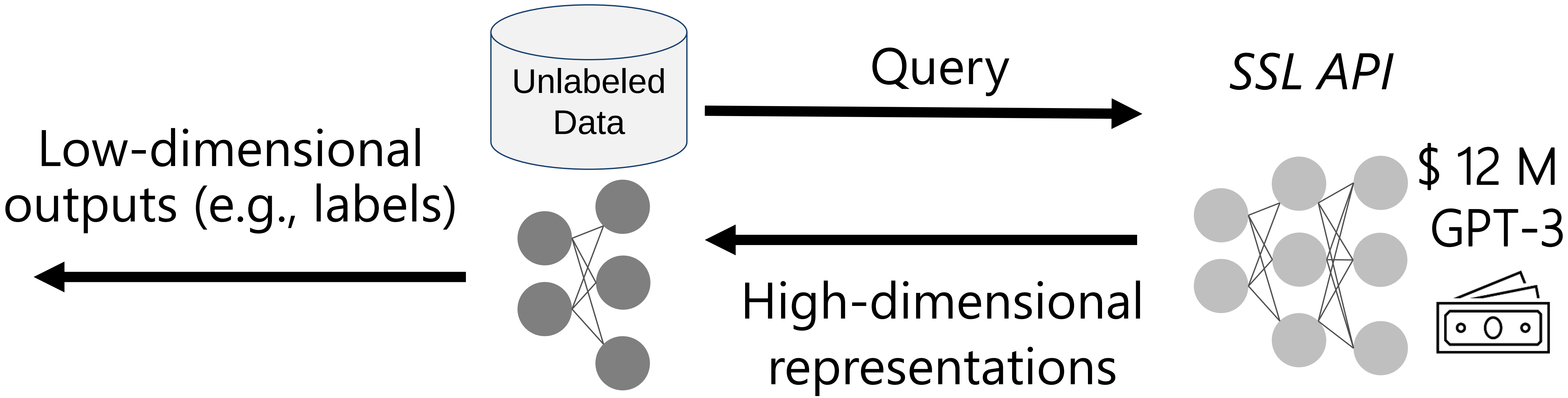


High Cost of Creating Self-Supervised APIs

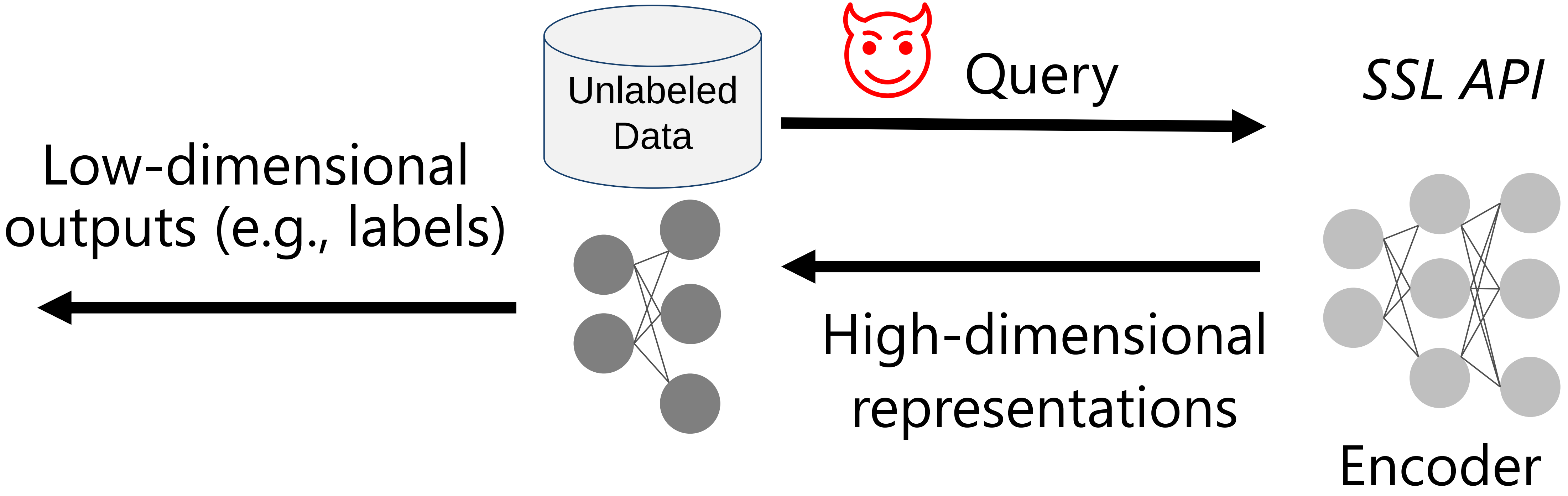
Collect Data

Tune Hyper-parameters

Run on GPU/TPU/CPU

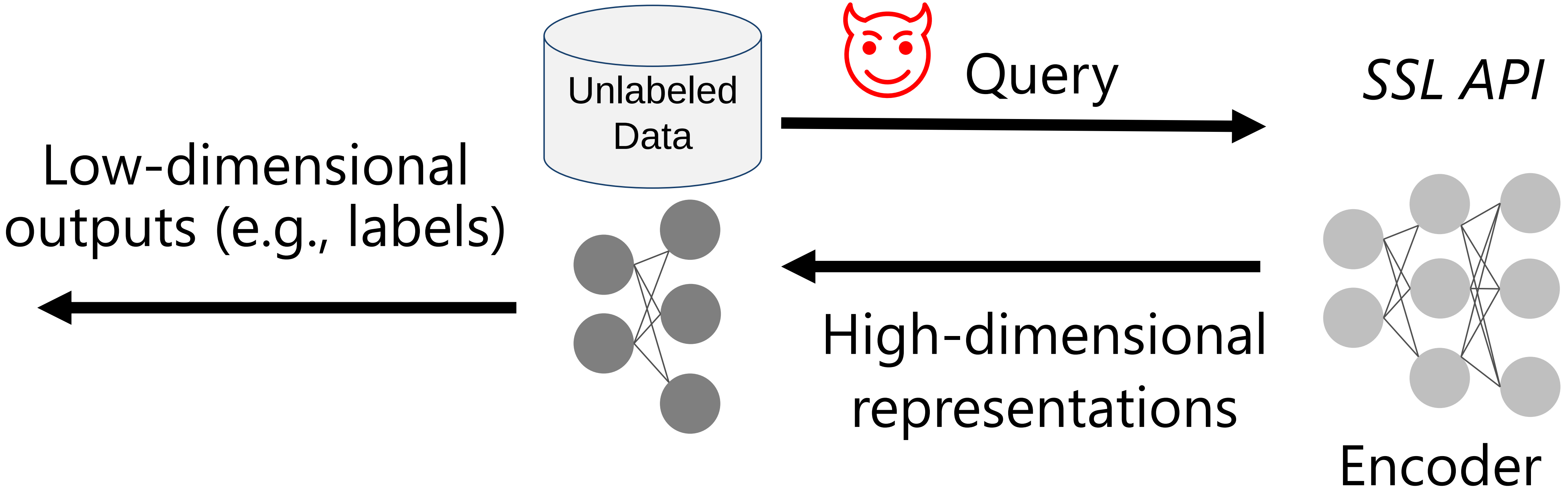


Efficient Attacks & Inadequate Defenses

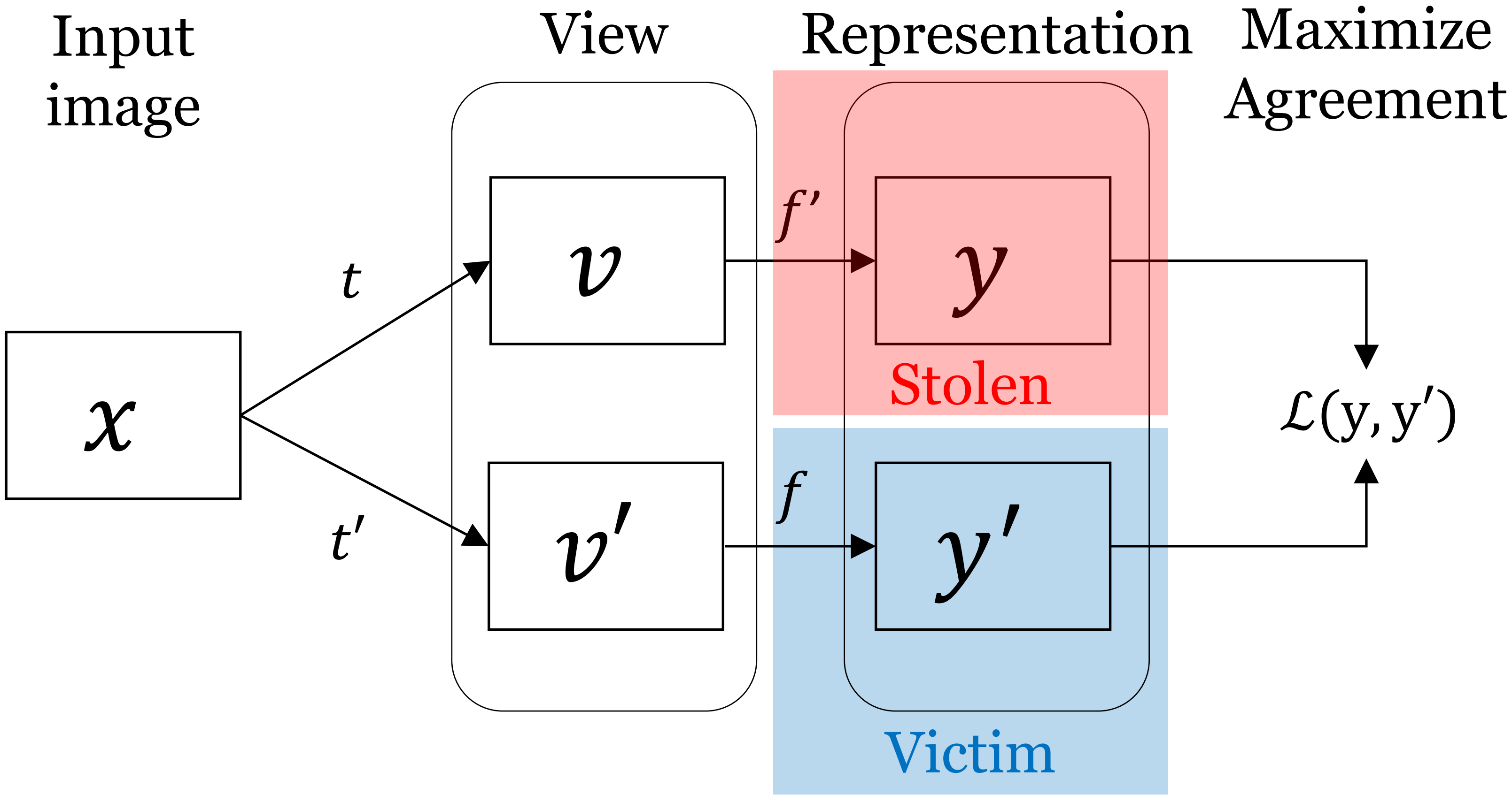


Efficient Attacks & Inadequate Defenses

1. Attacks against SSL models are query efficient.
2. Existing defenses against stealing supervised models are inadequate for SSL models.



Framework for Stealing Encoders



Impact of Loss Functions on Encoder Stealing

	CIFAR10 Victim		SVHN Victim	
Loss\Downstream Task	STL10	CIFAR10	STL10	CIFAR10
Victim baseline	*67.9*	*79.0*	*50.6*	*57.5*
MSE	64.8	75.5	46.3	51.2
InfoNCE	64.6	75.5	**50.4**	**56.3**
SoftNN	**67.1**	76.9	44.6	48.4
SupCon (uses labels)	63.1	**78.5**	33.9	42.3
Wasserstein	50.8	63.9	40.1	46.4
Barlow	26.6	26.9	16.3	17.9

Impact of Loss Functions on Encoder Stealing

	CIFAR10 Victim		SVHN Victim	
Loss\Downstream Task	STL10	CIFAR10	STL10	CIFAR10
Victim baseline	*67.9*	*79.0*	*50.6*	*57.5*
MSE	64.8	75.5	46.3	51.2
InfoNCE	64.6	75.5	**50.4**	**56.3**
SoftNN	**67.1**	76.9	44.6	48.4
SupCon (uses labels)	63.1	**78.5**	33.9	42.3
Wasserstein	50.8	63.9	40.1	46.4
Barlow	26.6	26.9	16.3	17.9

Contrastive losses perform the best for stealing encoders

Stealing a Pre-trained ImageNet Encoder

		Downstream Task				
# Queries	**Data for Stealing**	**CIFAR10**	**CIFAR100**	**STL10**	**SVHN**	**F-MNIST**
Victim ImageNet Encoder Baseline		***90.33***	***71.45***	***94.9***	***79.39***	***91.9***
60K	CIFAR10	**83.3**	**57.0**	71.2	73.8	90.7
50K	SVHN	73.3	47.1	58.2	78.8	90.4
250K	SVHN	77.1	52.6	61.9	**80.2**	**91.4**
50K	ImageNet	65.2	35.1	64.9	62.1	88.5
250K	ImageNet	80.0	**57.0**	**85.8**	71.5	90.2

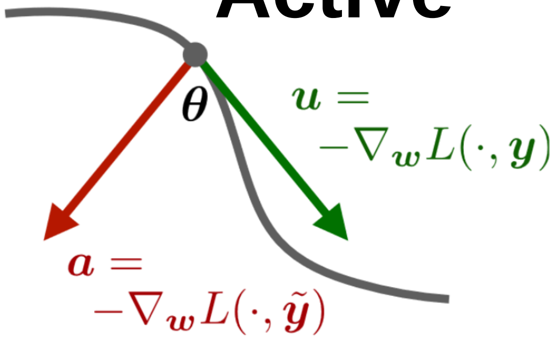
Stealing a Pre-trained ImageNet Encoder

		Downstream Task				
# Queries	**Data for Stealing**	**CIFAR10**	**CIFAR100**	**STL10**	**SVHN**	**F-MNIST**
Victim ImageNet Encoder Baseline		***90.33***	***71.45***	***94.9***	***79.39***	***91.9***
60K	CIFAR10	**83.3**	**57.0**	71.2	73.8	90.7
50K	SVHN	73.3	47.1	58.2	78.8	90.4
250K	SVHN	77.1	52.6	61.9	**80.2**	**91.4**
50K	ImageNet	65.2	35.1	64.9	62.1	88.5
250K	ImageNet	80.0	**57.0**	**85.8**	71.5	90.2

number of stealing queries < 1/5th number of training data points

Adapt Defenses against Stealing Encoders

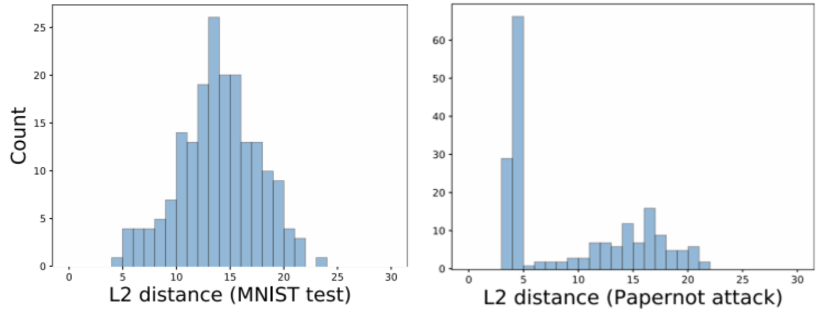
Active



Poison Attacker's Objective

Prediction Poisoning [Orekondy et al. 2020]

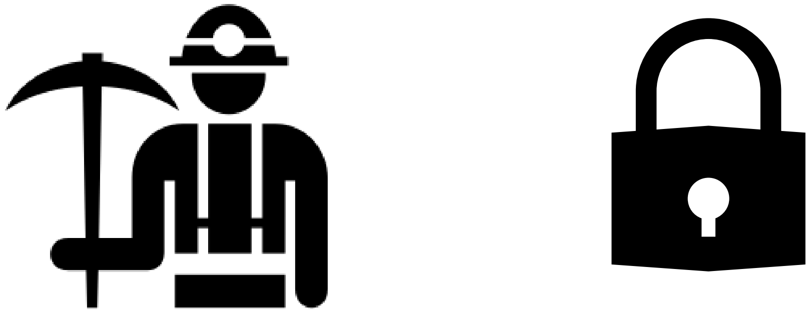
Passive



Detect Attack & Stop Responding

PRADA [Juuti et al. 2019]

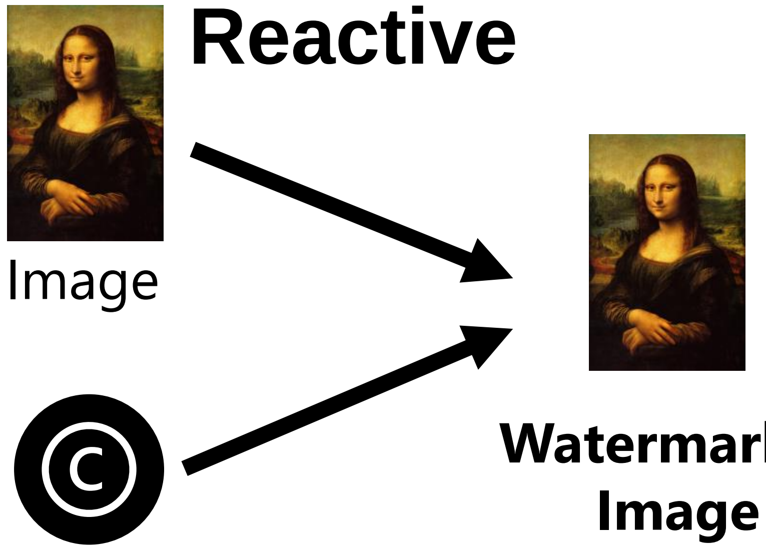
Pro-Active



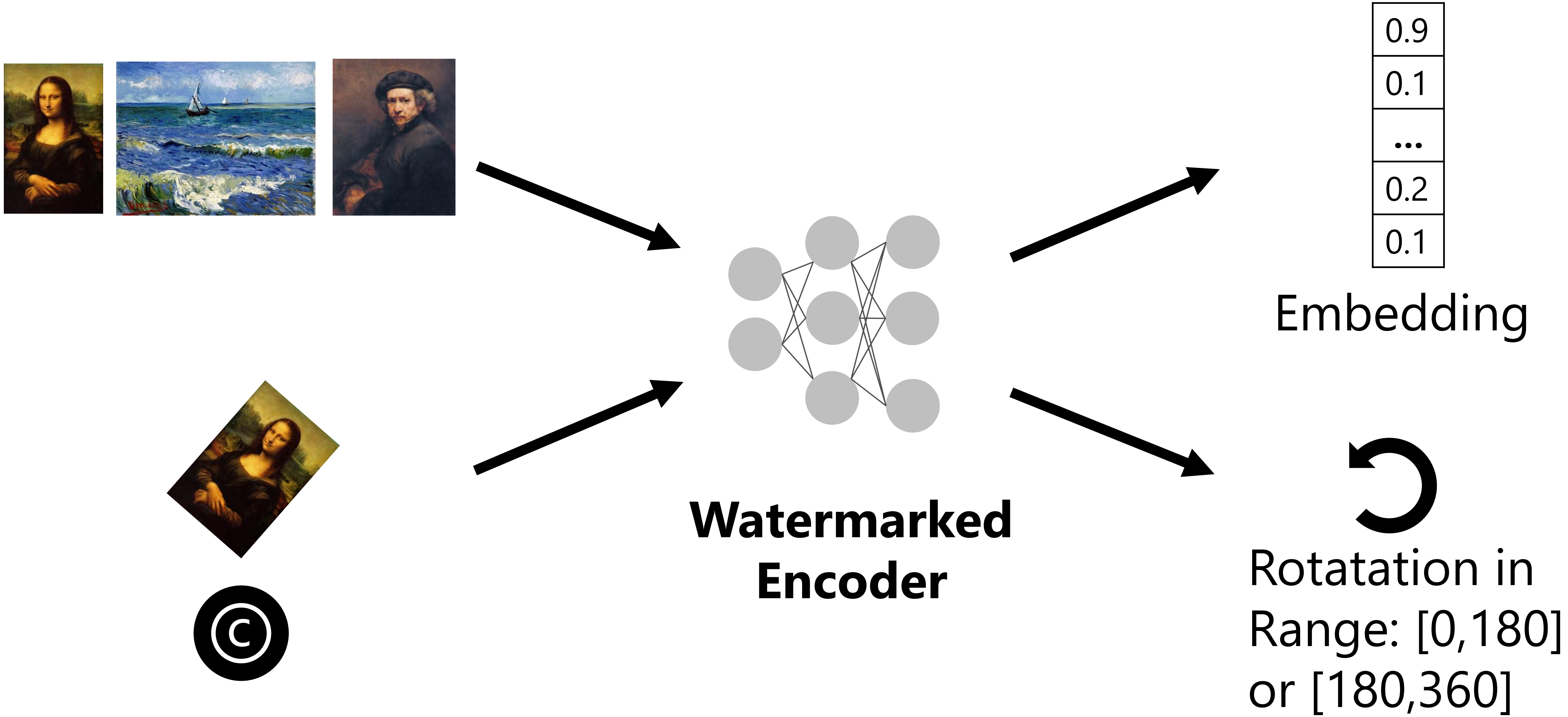
Higher cost for more information

Callibrated PoW with PATE [Dziedzic et al. 2022]

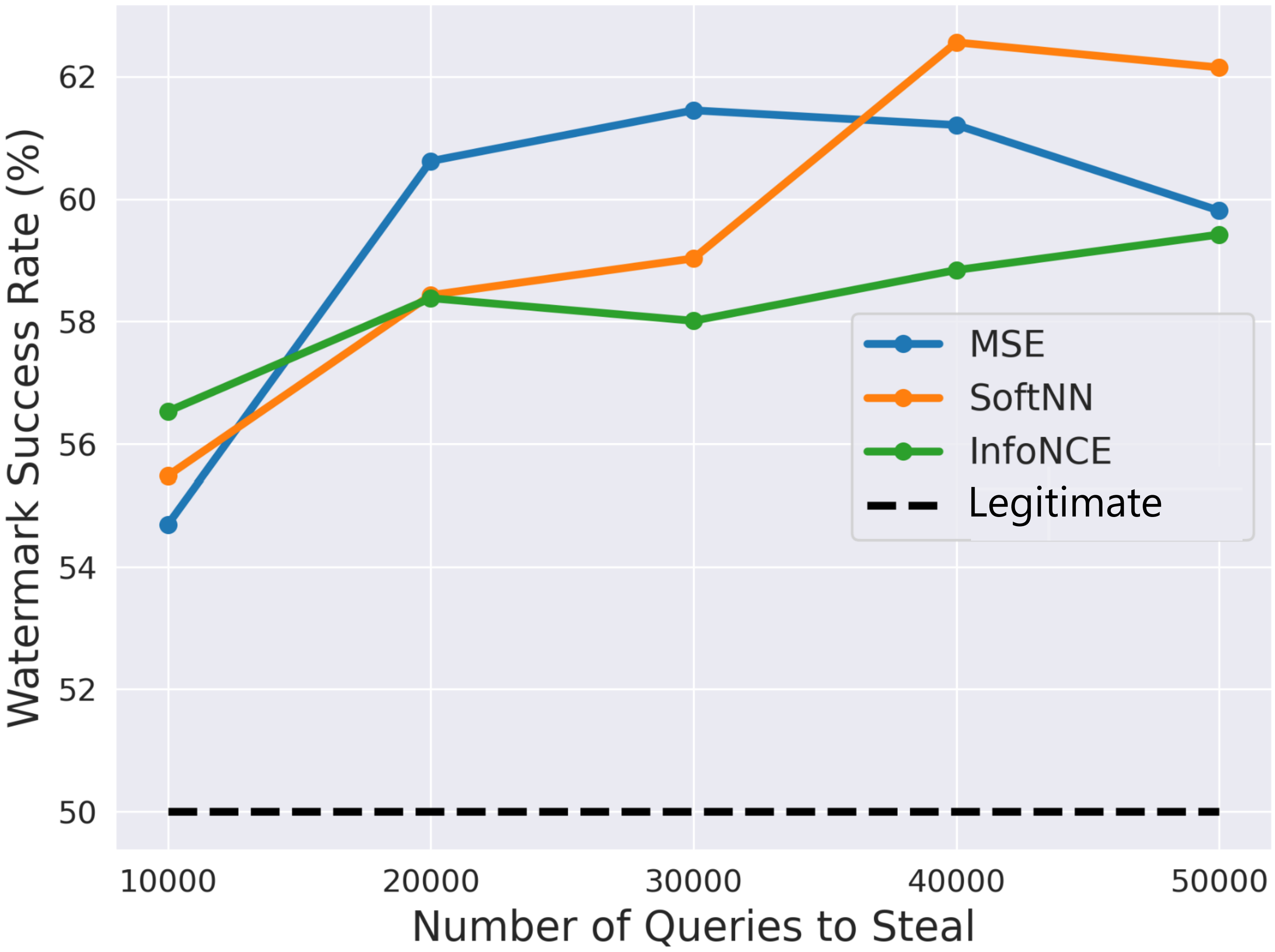
Reactive



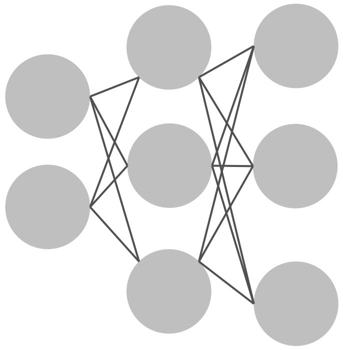
Embed Rotation Task to Defend Encoders



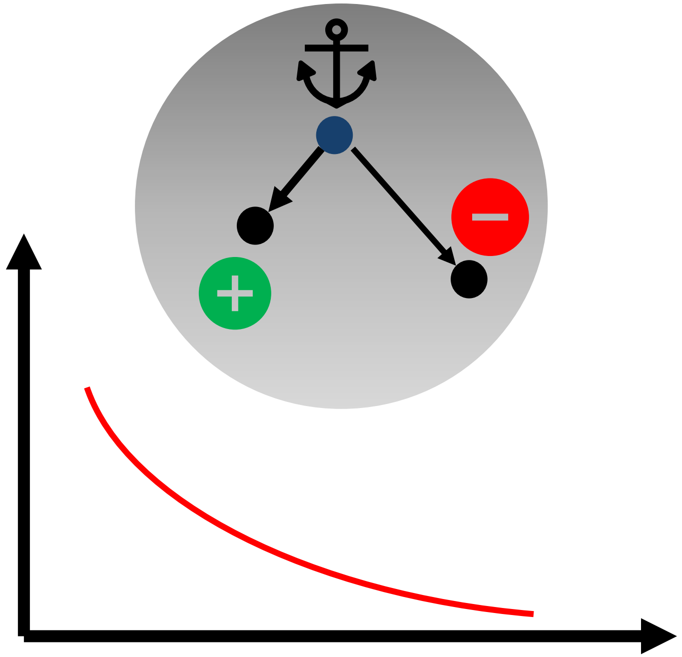
Transferability of the Rotation Watermark



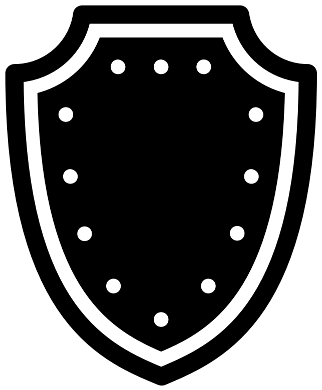
Conclusions & Future Work



High Performance of Stolen Encoders



Contrastive Loss Functions



Design New Defenses

Thank you

https://cleverhans-lab.github.io

{adam.dziedzic,nicolas.papernot}@utoronto.ca