Is this Encoder Mine? On Stealing and Defending Self-Supervised Encoders

Adam Dziedzic

Deep Learning: Classics and Trends (DLCT)

November 11th, 2022

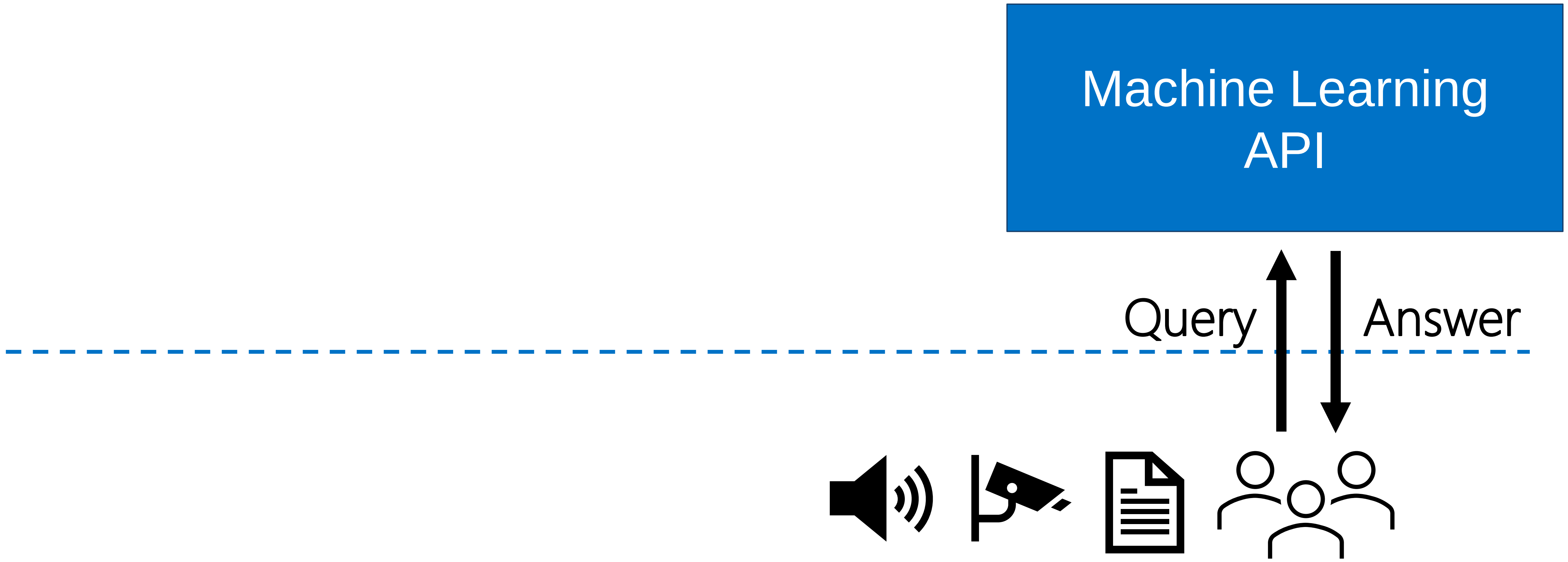
cleverhans

VECTOR INSTITUTE

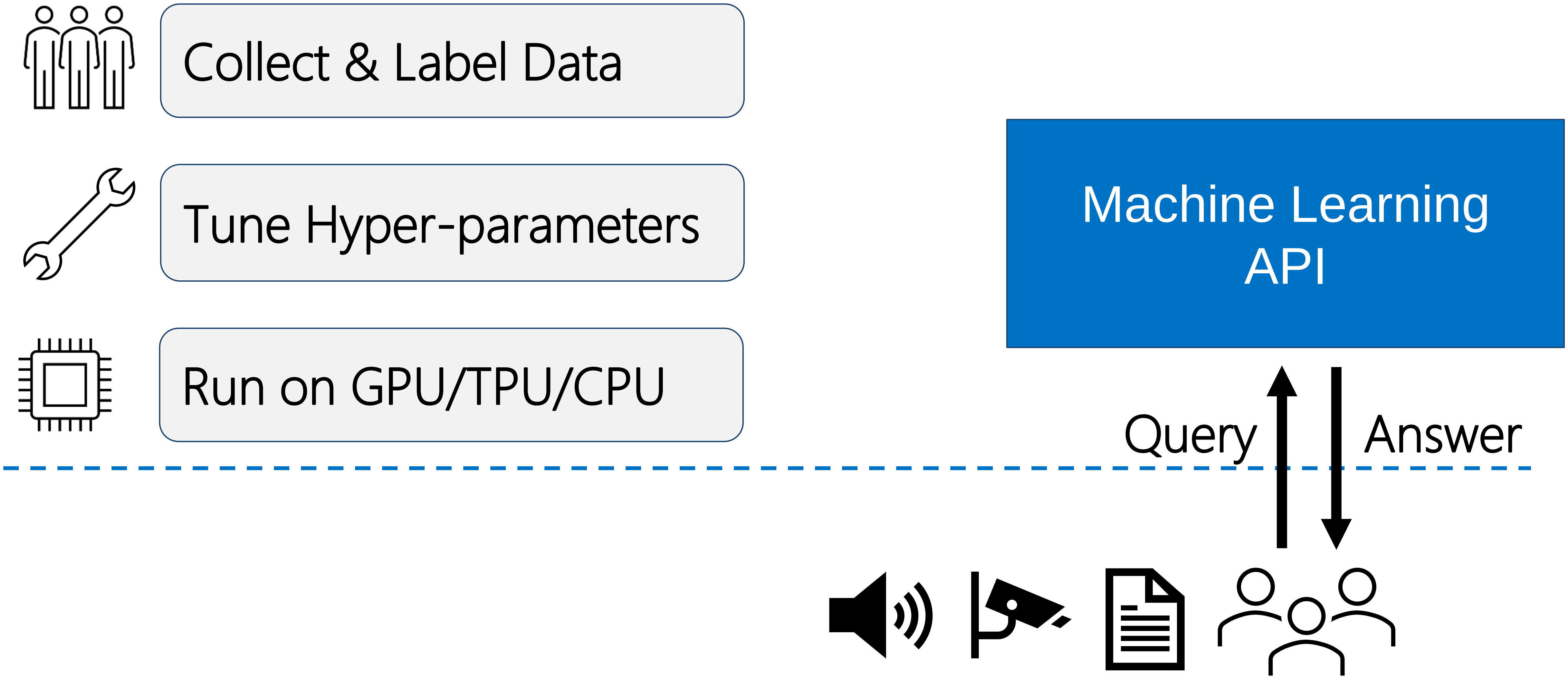
UNIVERSITY OF TORONTO

Annotate Data Using Machine Learning APIs

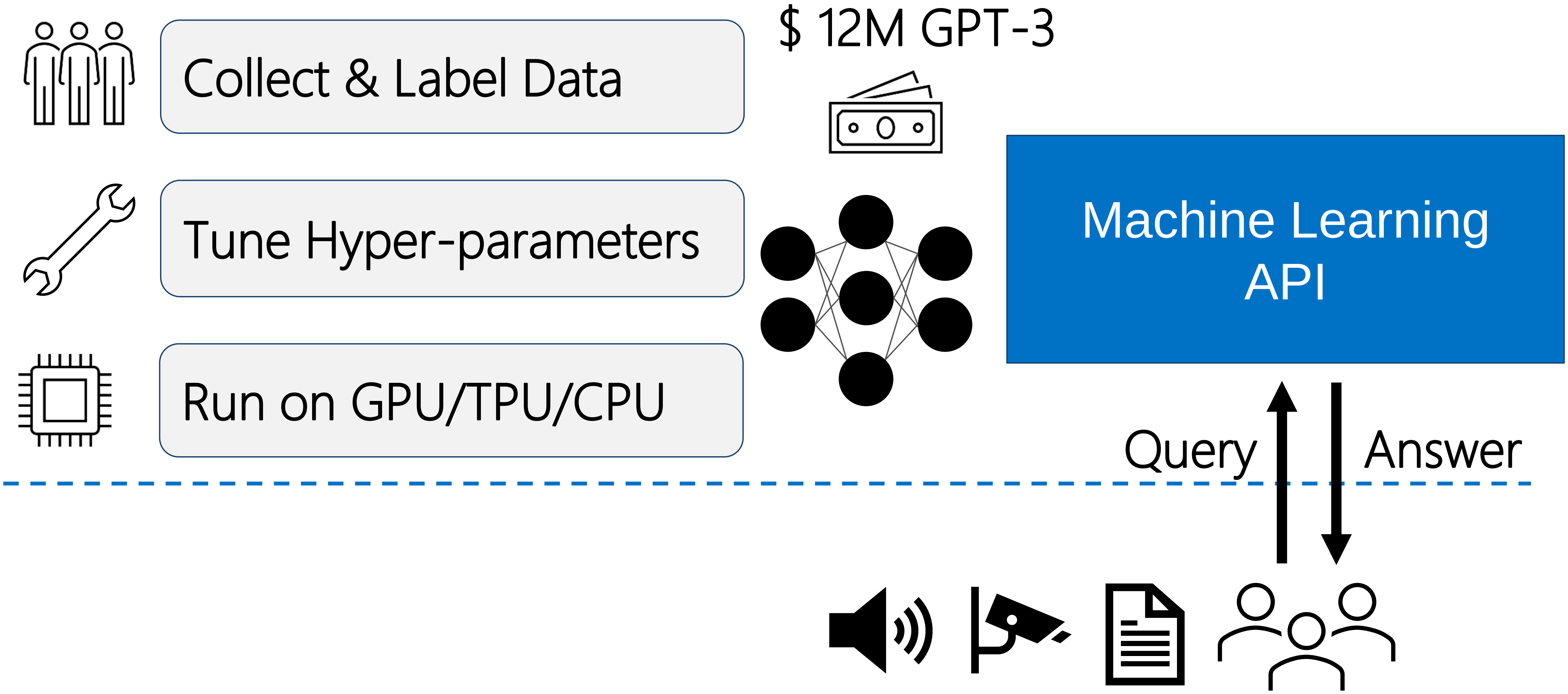
Adam Dziedzic, Muhammad Ahmad Kaleem, Yu Shen Lu, Nicolas Papernot *"Increasing the Cost of Model Extraction with Calibrated Proof of Work"* [ICLR 2022 SPOTLIGTH].



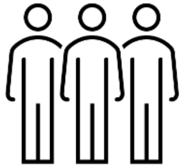
Train Models for Machine Learning Services

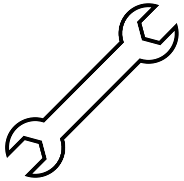


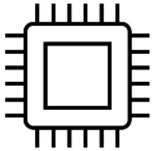
Train Models for Machine Learning Services



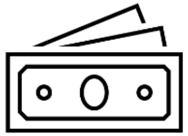
Stealing Machine Learning Models

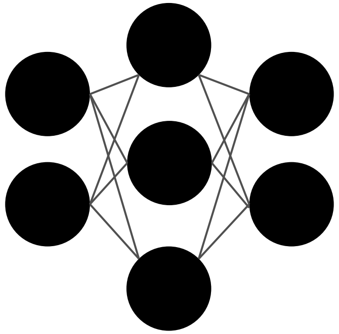

Collect & Label Data


Tune Hyper-parameters


Run on GPU/TPU/CPU

$ 12M GPT-3





Machine Learning API

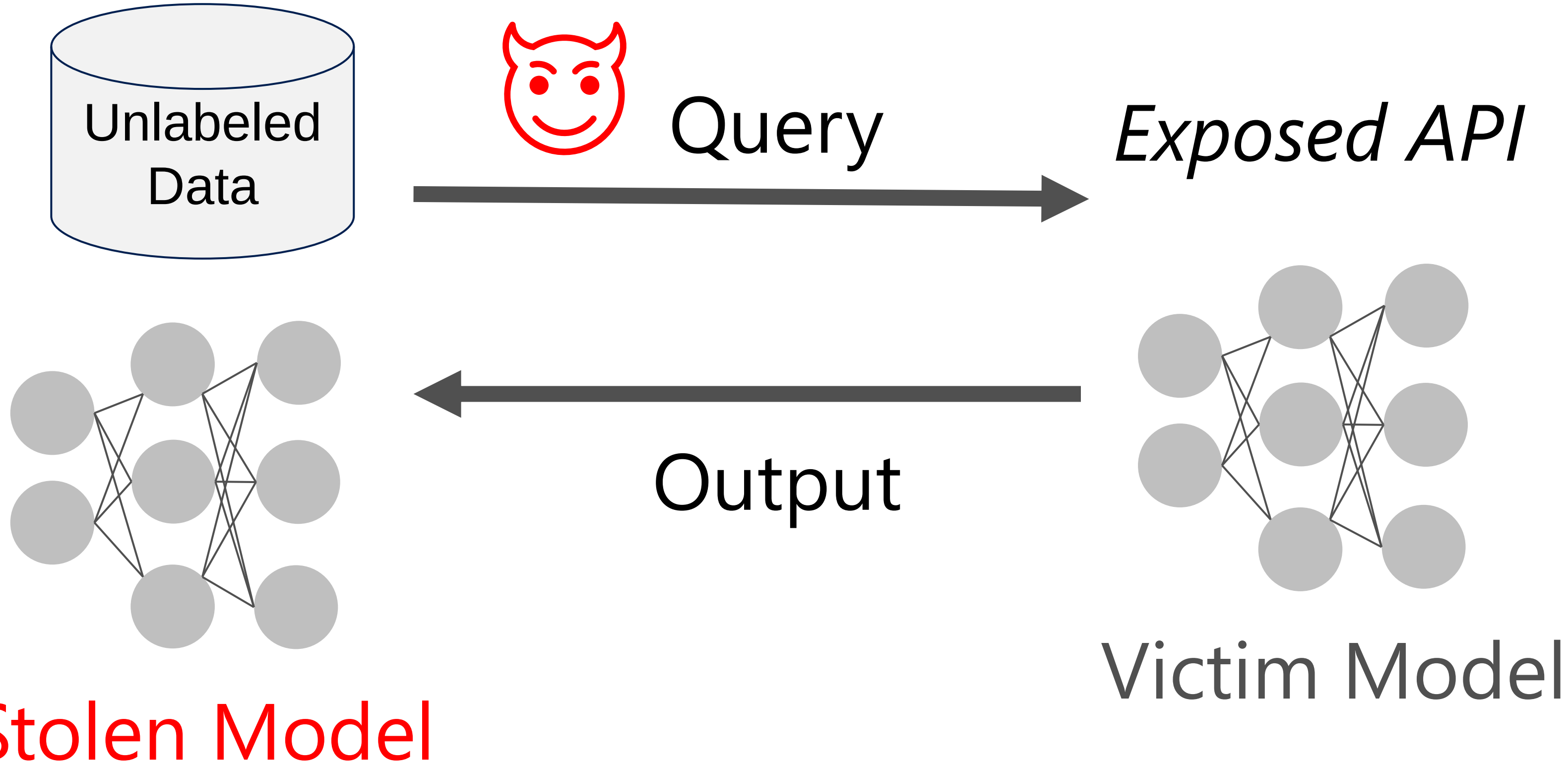
Query

Answer

[Shankar et al. 2020]

Model Stealing - ranked among the most severe attacks against ML

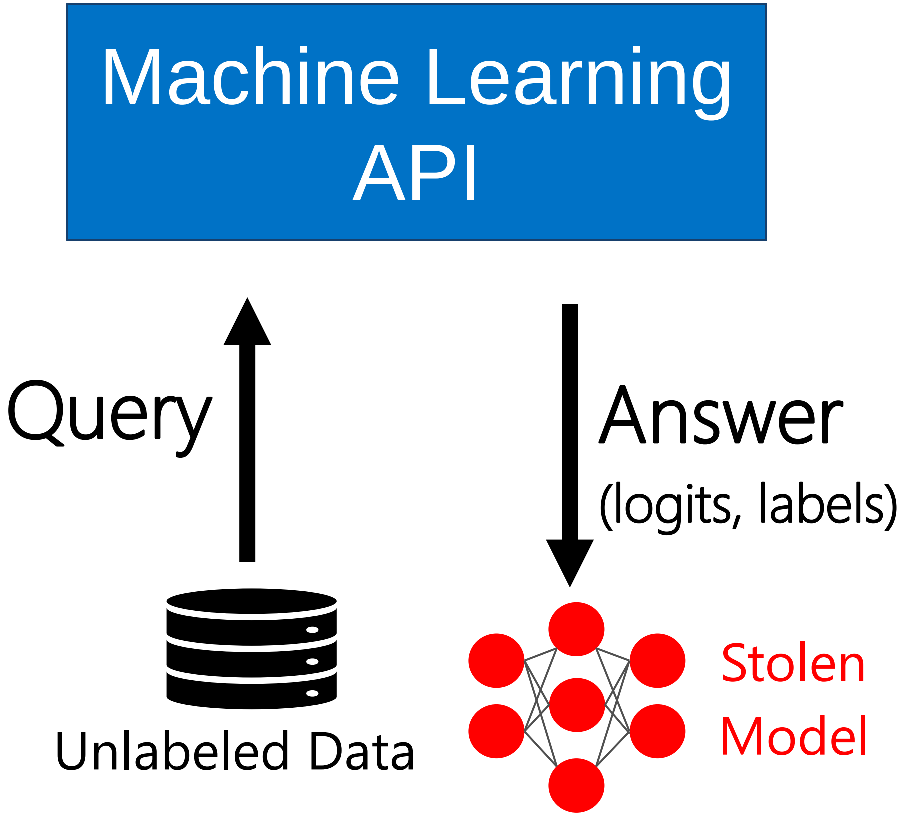
Threat of Model Stealing



’s incentives: 1. Steal model with a lower training cost
2. Reconnaissance for launching further attacks

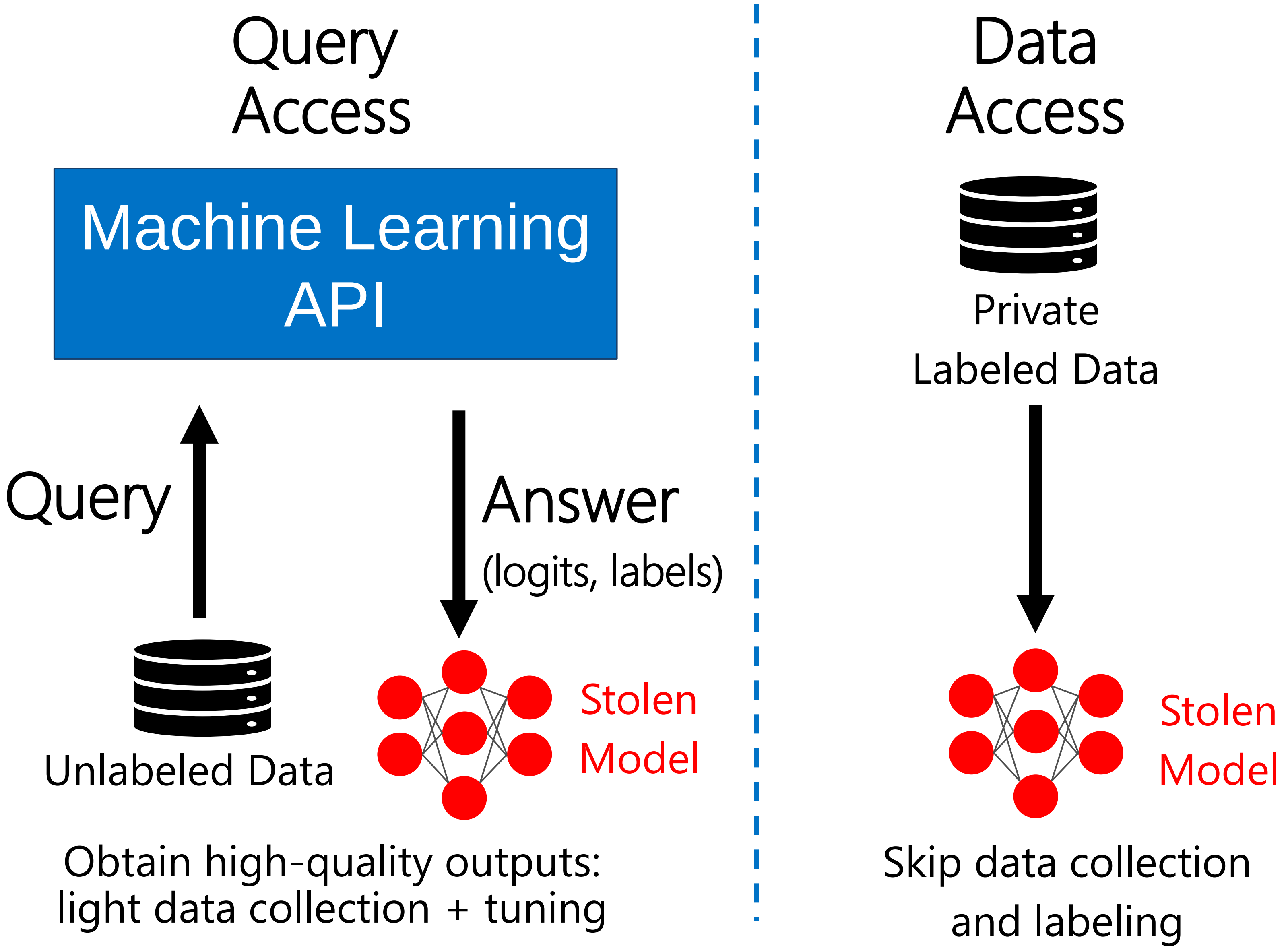
Degrees of Access to Your Knowledge

Query
Access

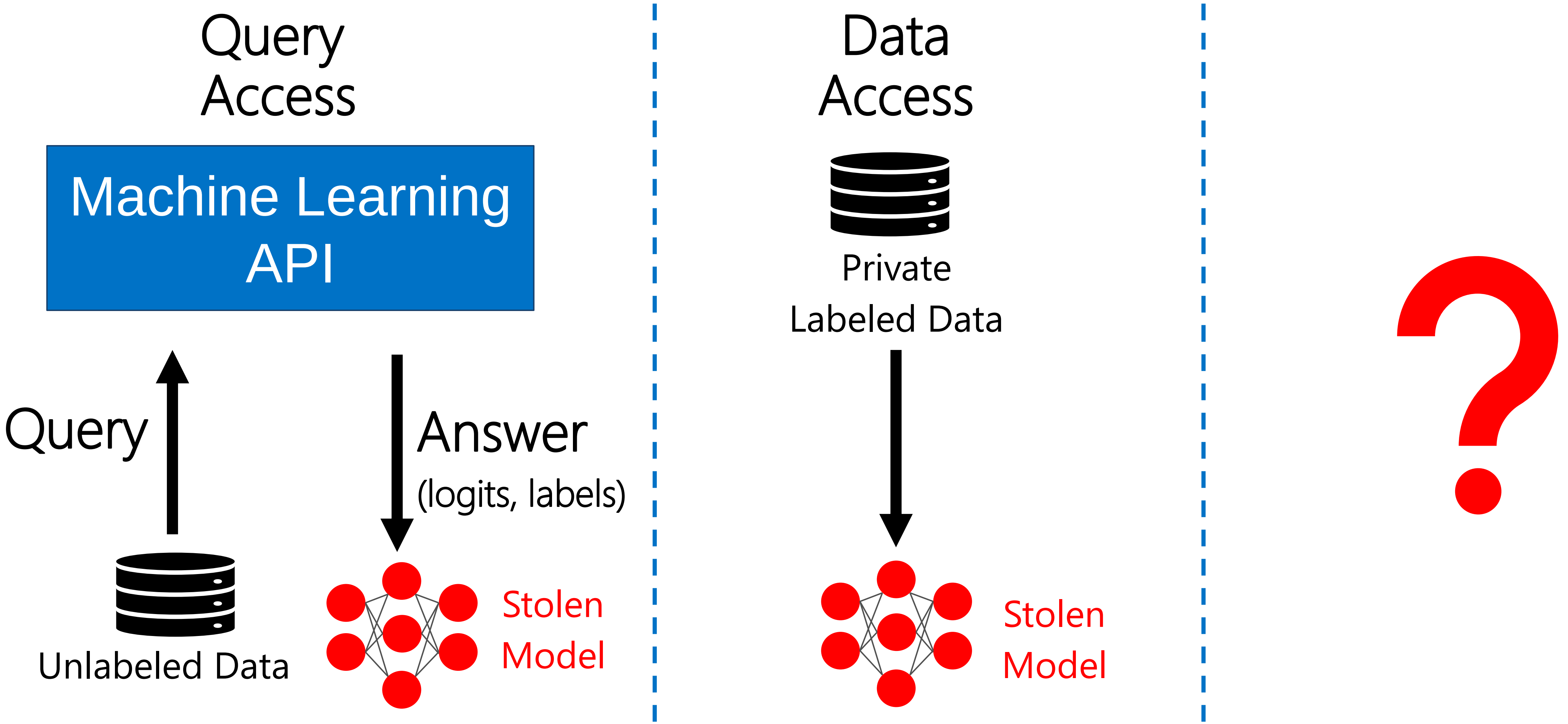


Obtain high-quality outputs:
light data collection + tuning

Degrees of Access to Your Knowledge



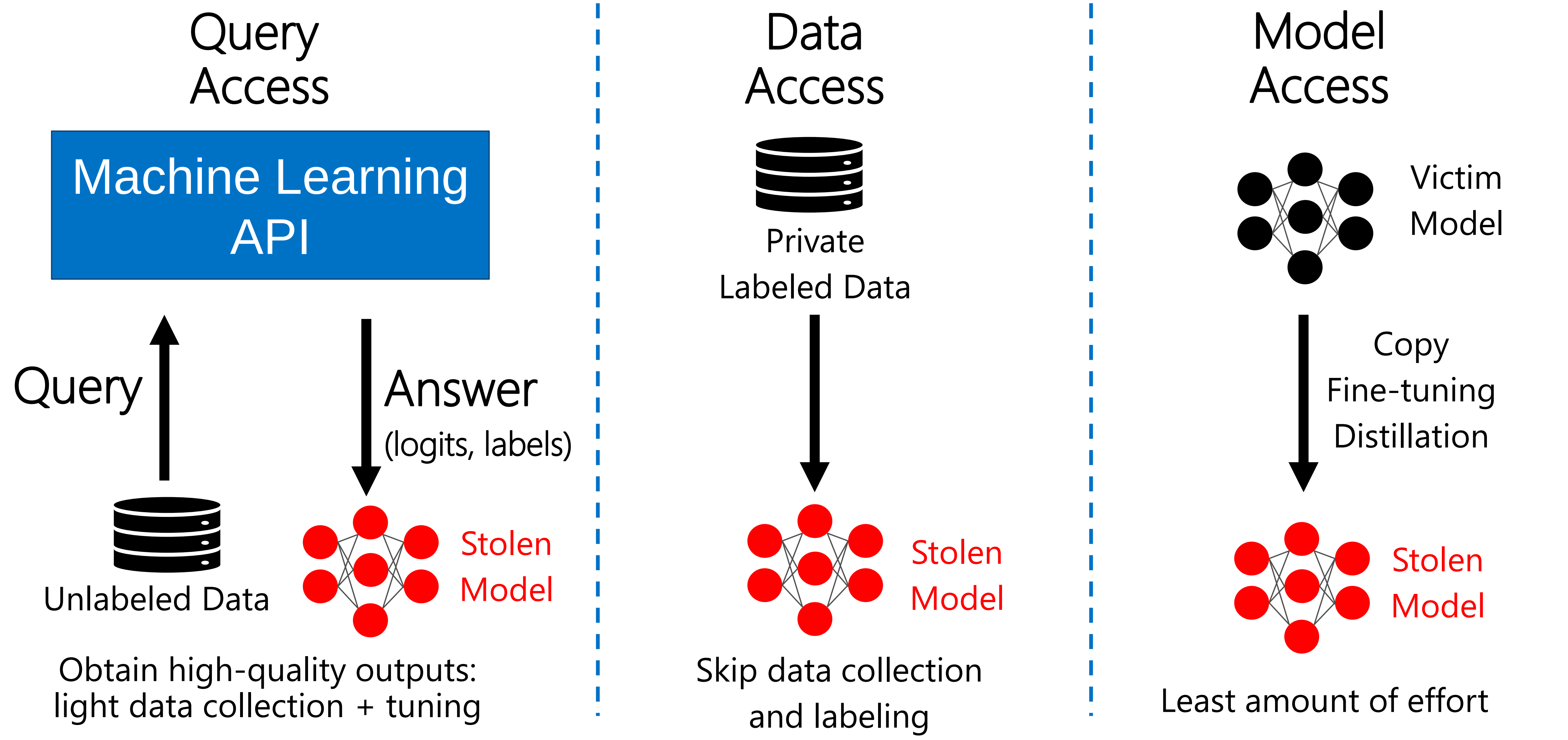
Degrees of Access to Your Knowledge



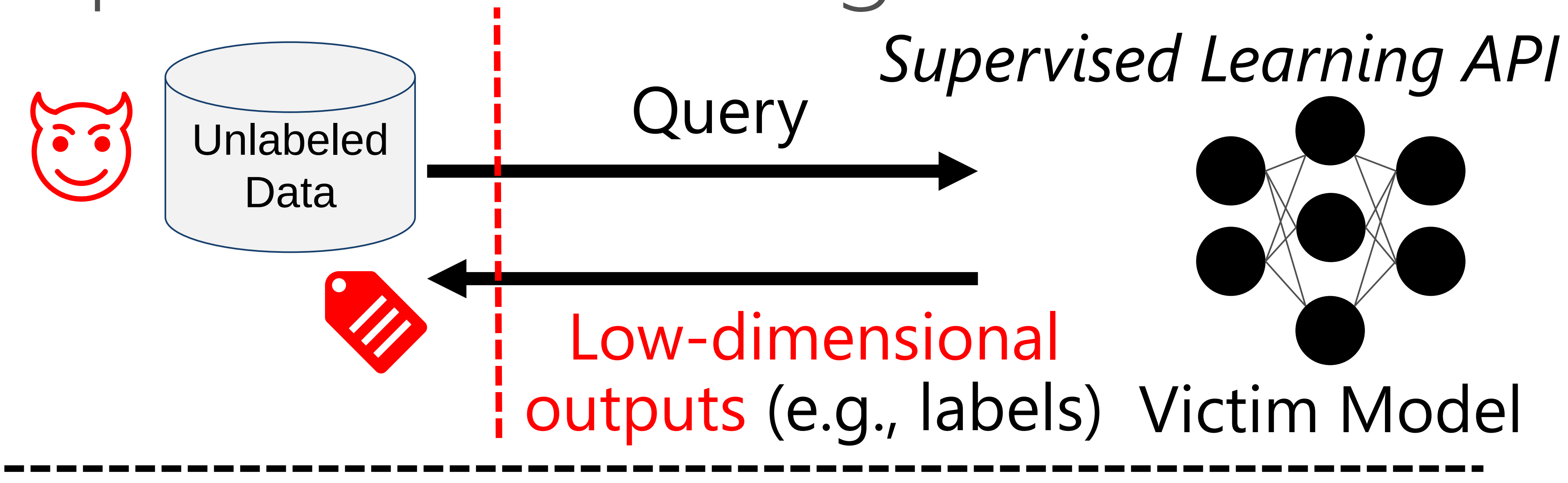
Obtain high-quality outputs:
light data collection + tuning

Skip data collection
and labeling

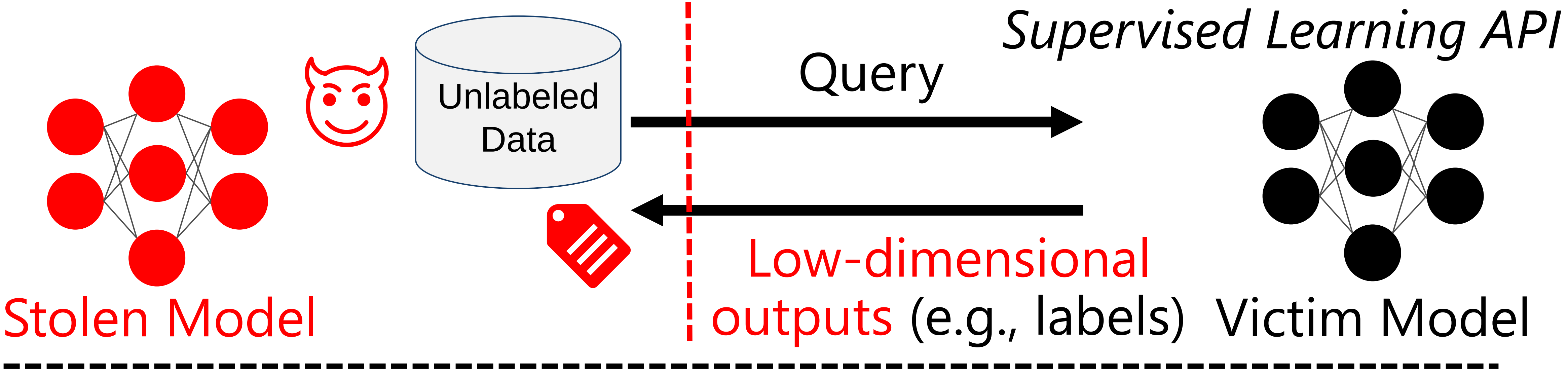
Degrees of Access to Your Knowledge



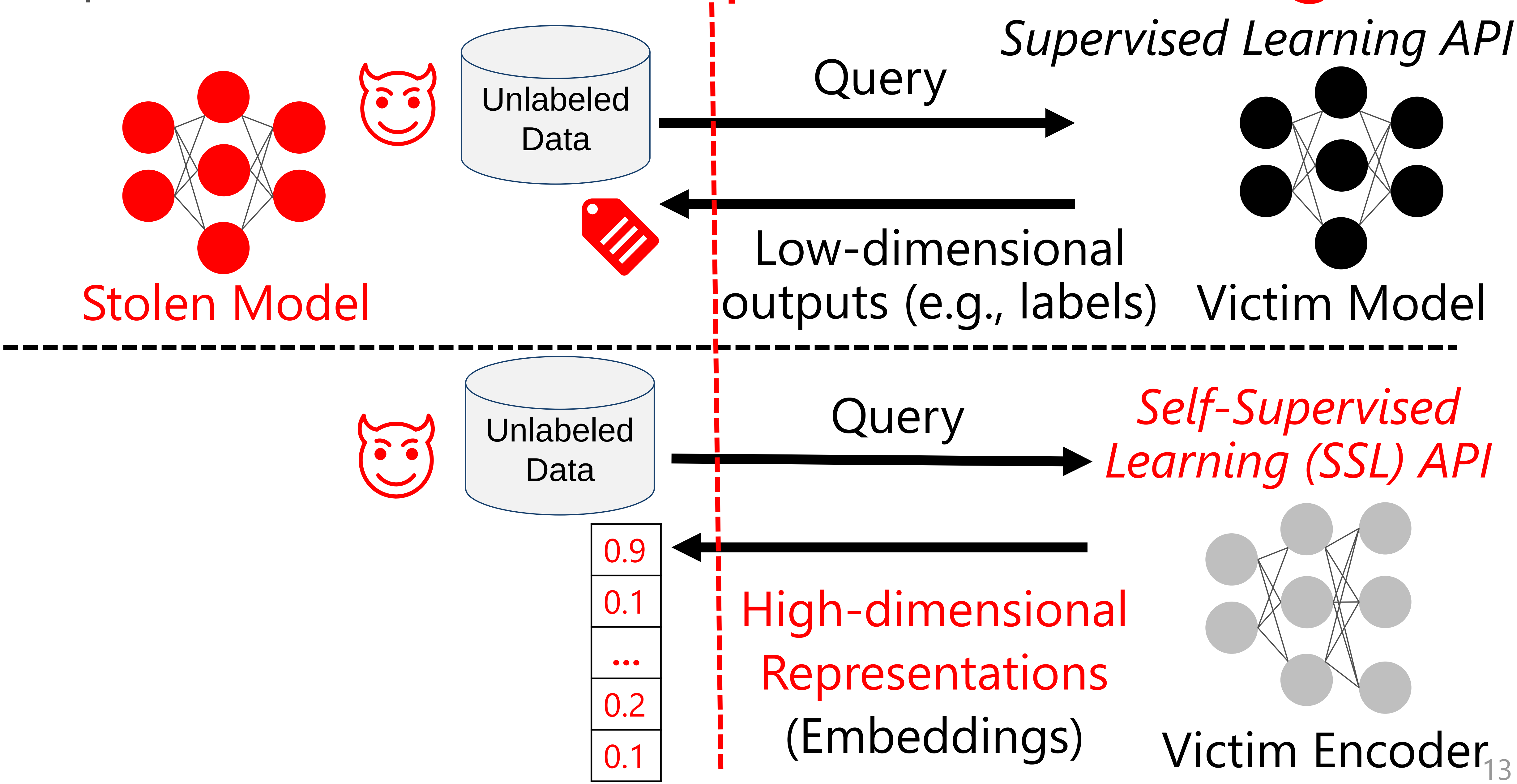
Supervised Learning API



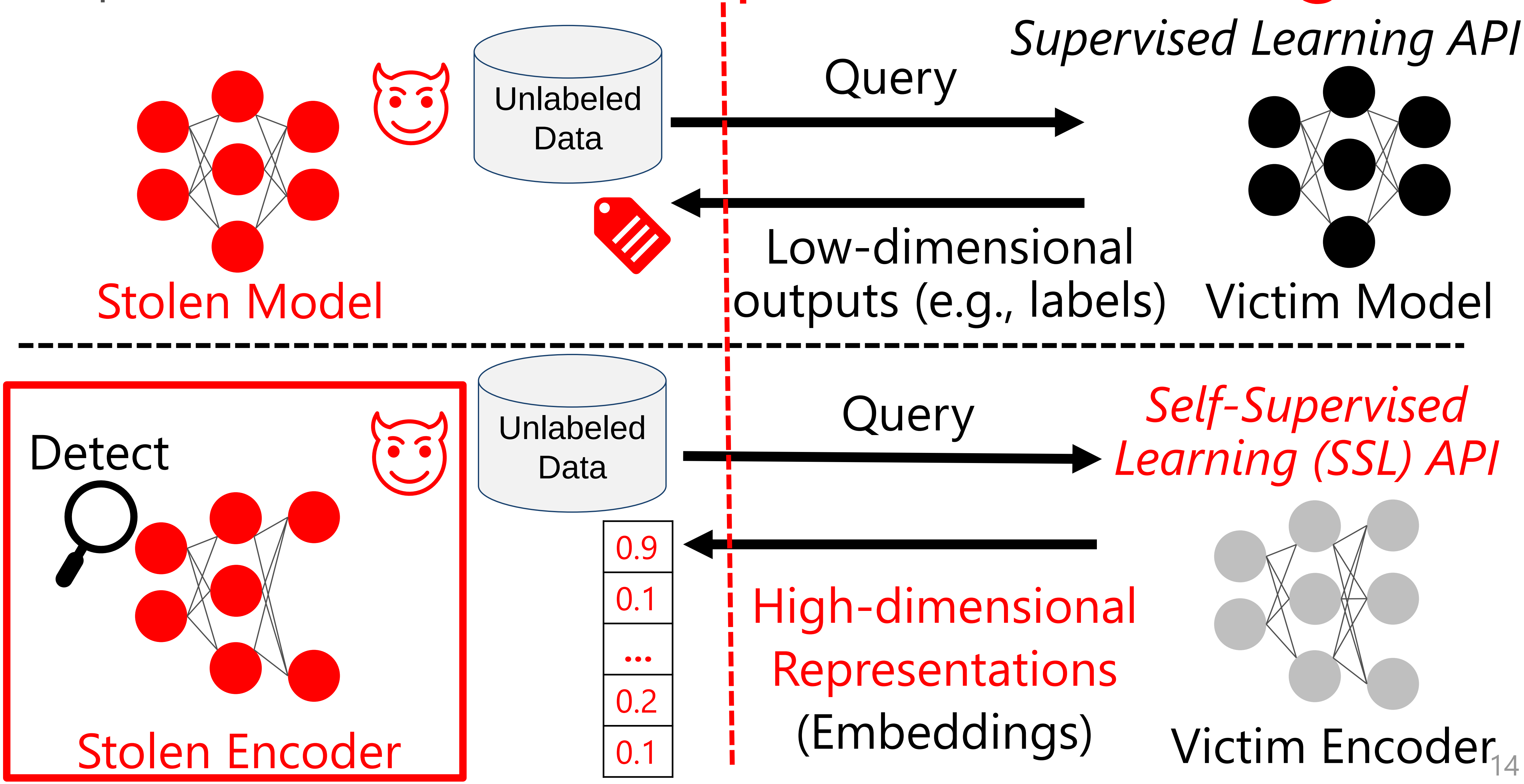
Supervised Learning API



Supervised vs Self-Supervised Learning API



Supervised vs Self-Supervised Learning API



Threat of Stealing Self-Supervised Encoders

Practical and **Growing** Threat

ML Service Providers have already commenced offering SSL Encoders over paid APIs.

SSL is becoming the dominant paradigm for important ML domains like Vision and NLP.

Build next-gen apps with OpenAI's powerful models.

OpenAI's API provides access to GPT-3, which performs a wide variety of natural language tasks, and Codex, which translates natural language to code.

GET STARTED READ DOCUMENTATION

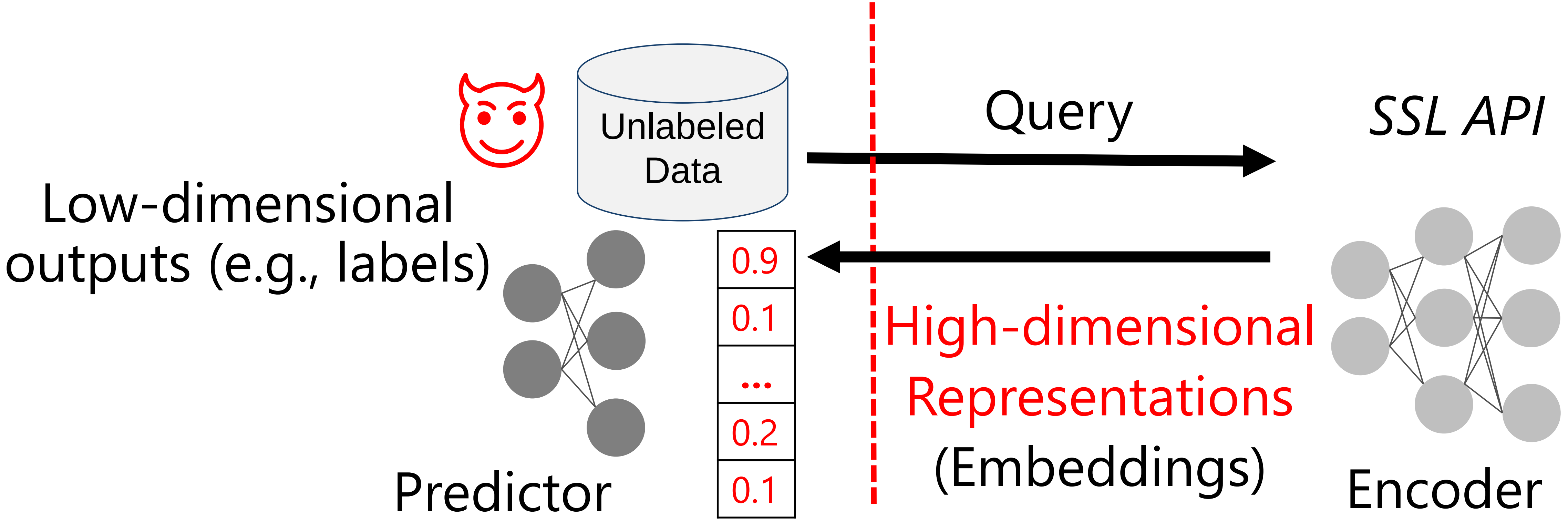
co:here

Cohere Raises $40 Million in Series A Financing to Make Natural Language Processing Safe and Accessible to Any Business

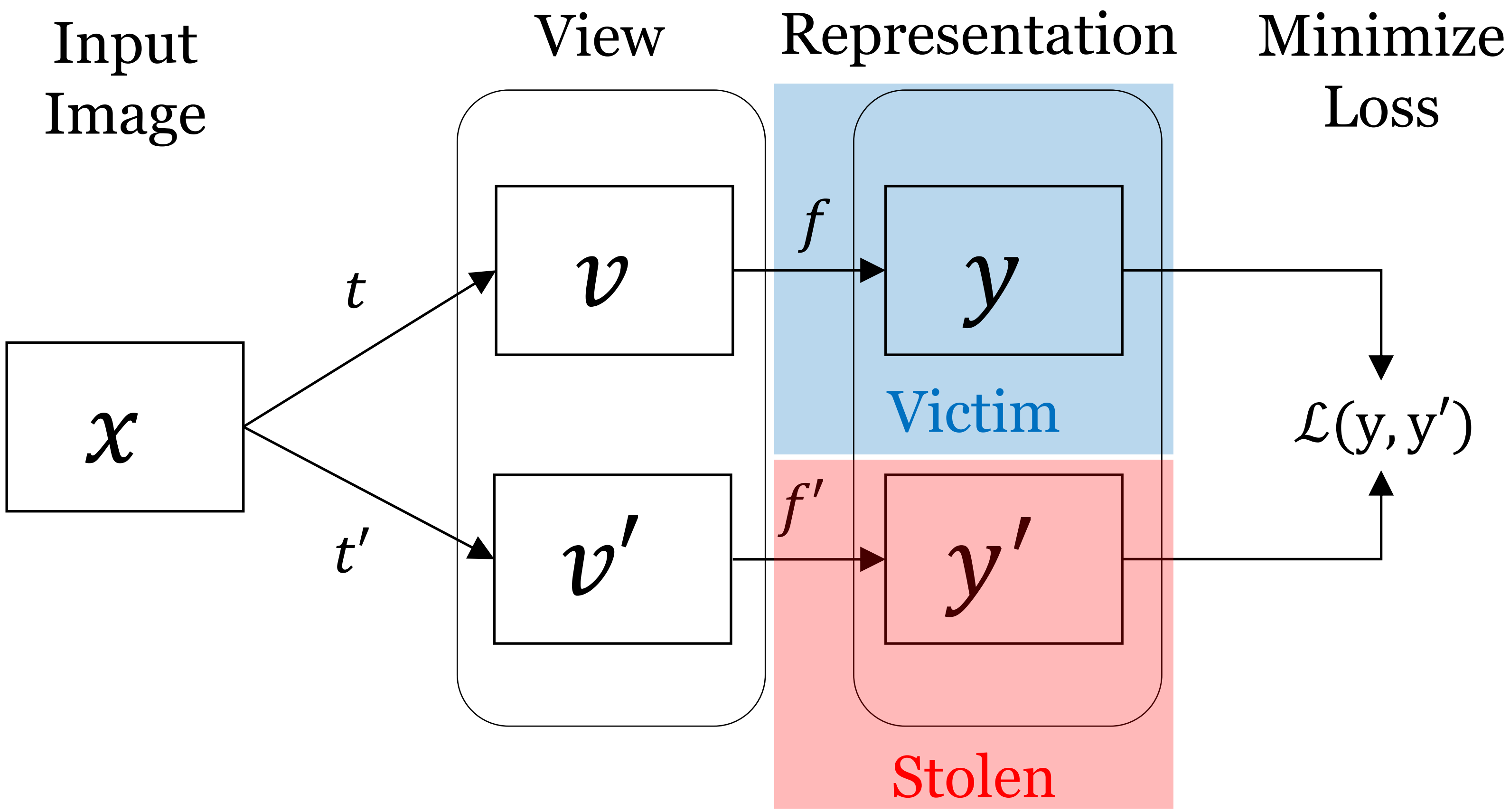
clarifai exposes a visual recognition model for returning 768-dimensional numerical vectors that represent the items in images and video.

Efficient Attacks & Inadequate Defenses

1. Attacks against SSL models are query efficient: number of stealing queries < 1/5th number of training data points.
2. Existing defenses against stealing supervised models are inadequate for SSL models.



Siamese Framework for Stealing Encoders



Adam Dziedzic, Nikita Dhawan, Muhammad Ahmad Kaleem, Jonas Guan, Nicolas Papernot
"On the Difficulty of Defending Self-Supervised Learning against Model Extraction" [ICML 2022].

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*
Mean Squared Error	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*
Mean Squared Error	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 training & 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

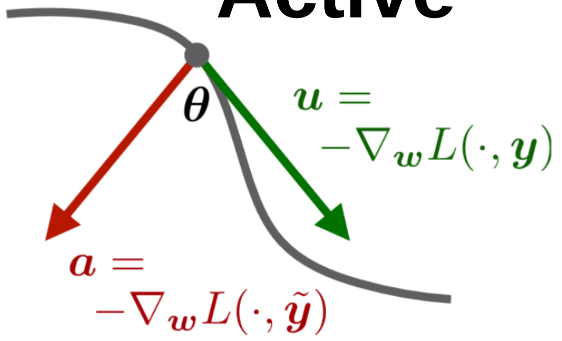
Stealing a Pre-trained ImageNet Encoder

		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

Defenses against Model Stealing

Active

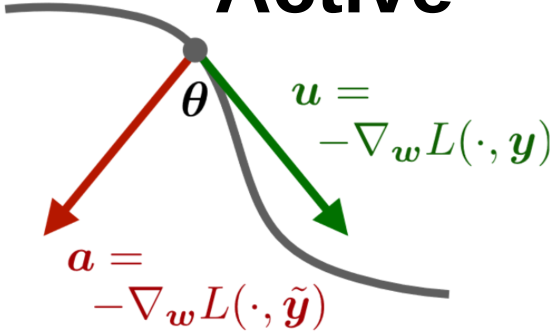


Poison Attacker's Objective

Prediction Poisoning [Orekondy et al. 2020]

Defenses against Model Stealing

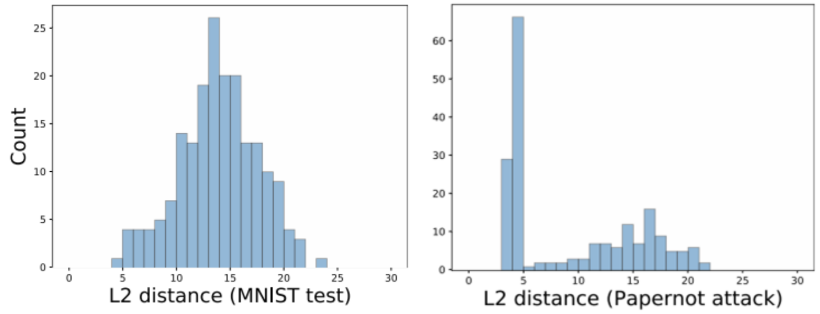
Active



Poison Attacker's Objective

Prediction Poisoning [Orekondy et al. 2020]

Passive

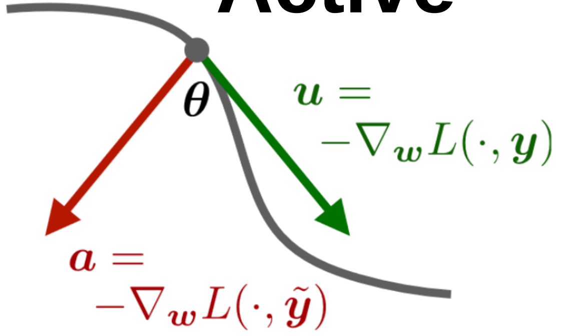


Detect Attack & Stop Responding

PRADA [Juuti et al. 2019]

Defenses against Model Stealing

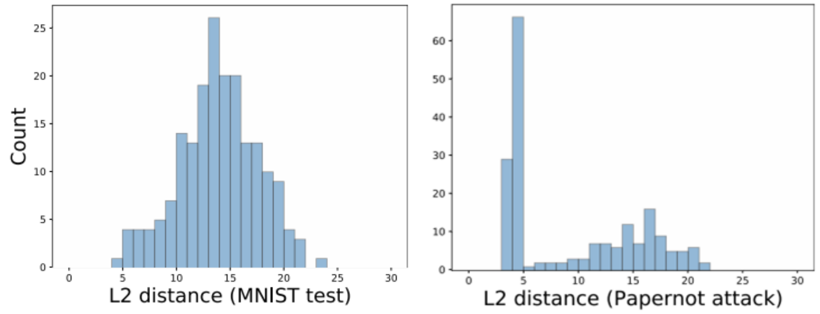
Active



Poison Attacker's Objective

Prediction Poisoning [Orekondy et al. 2020]

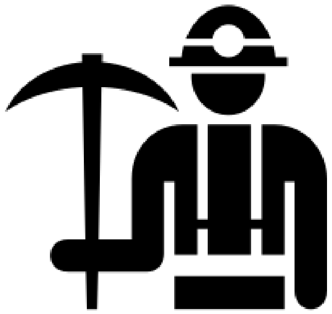
Passive



Detect Attack & Stop Responding

PRADA [Juuti et al. 2019]

Pro-Active



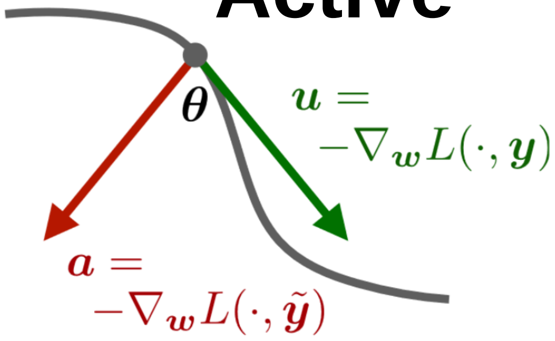
Proof-of-Work

Differential Privacy

Calibrated Proof-of-Work with PATE

Defenses against Model Stealing

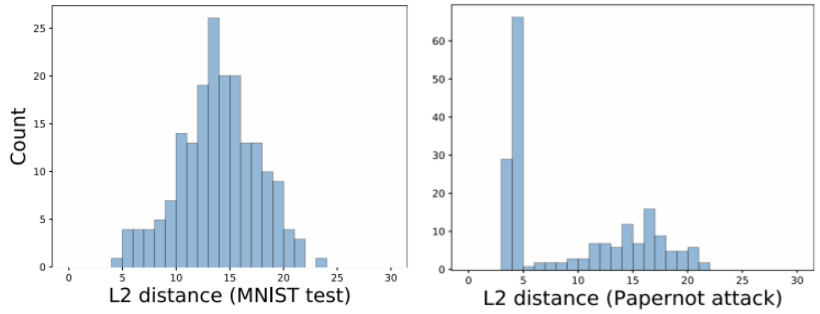
Active



Poison Attacker’s Objective

Prediction Poisoning [Orekondy et al. 2020]

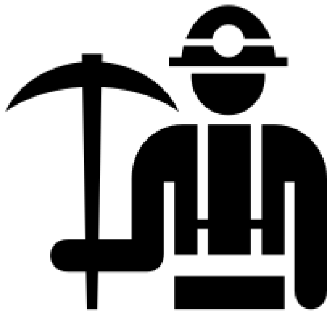
Passive



Detect Attack & Stop Responding

PRADA [Juuti et al. 2019]

Pro-Active



Proof-of-Work

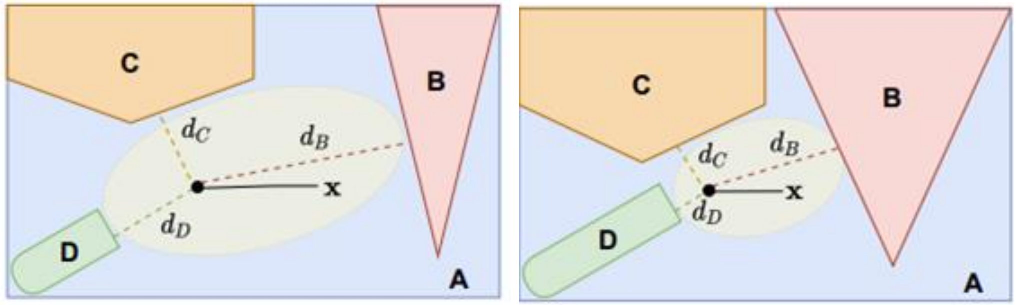
Differential Privacy

Calibrated Proof-of-Work with PATE

Reactive

Train

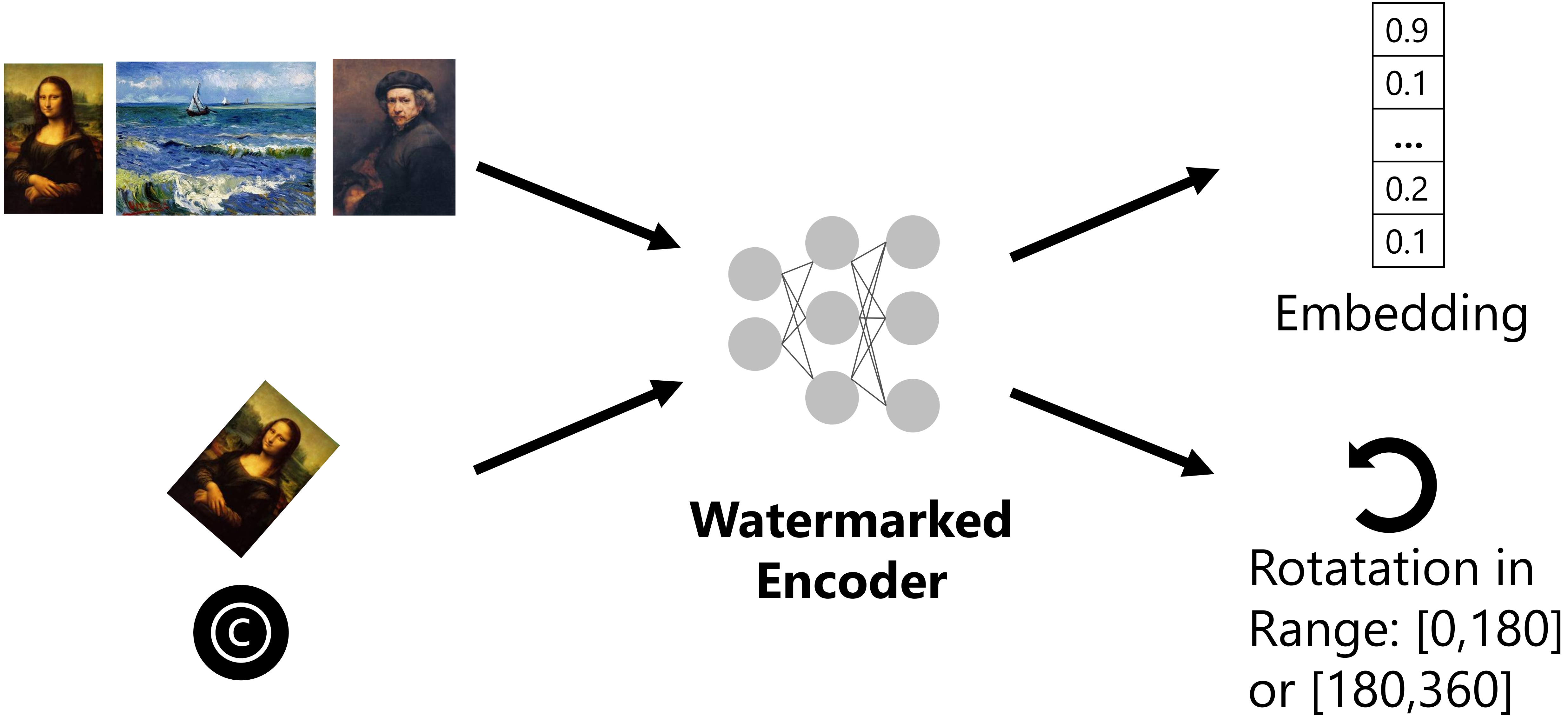
Test



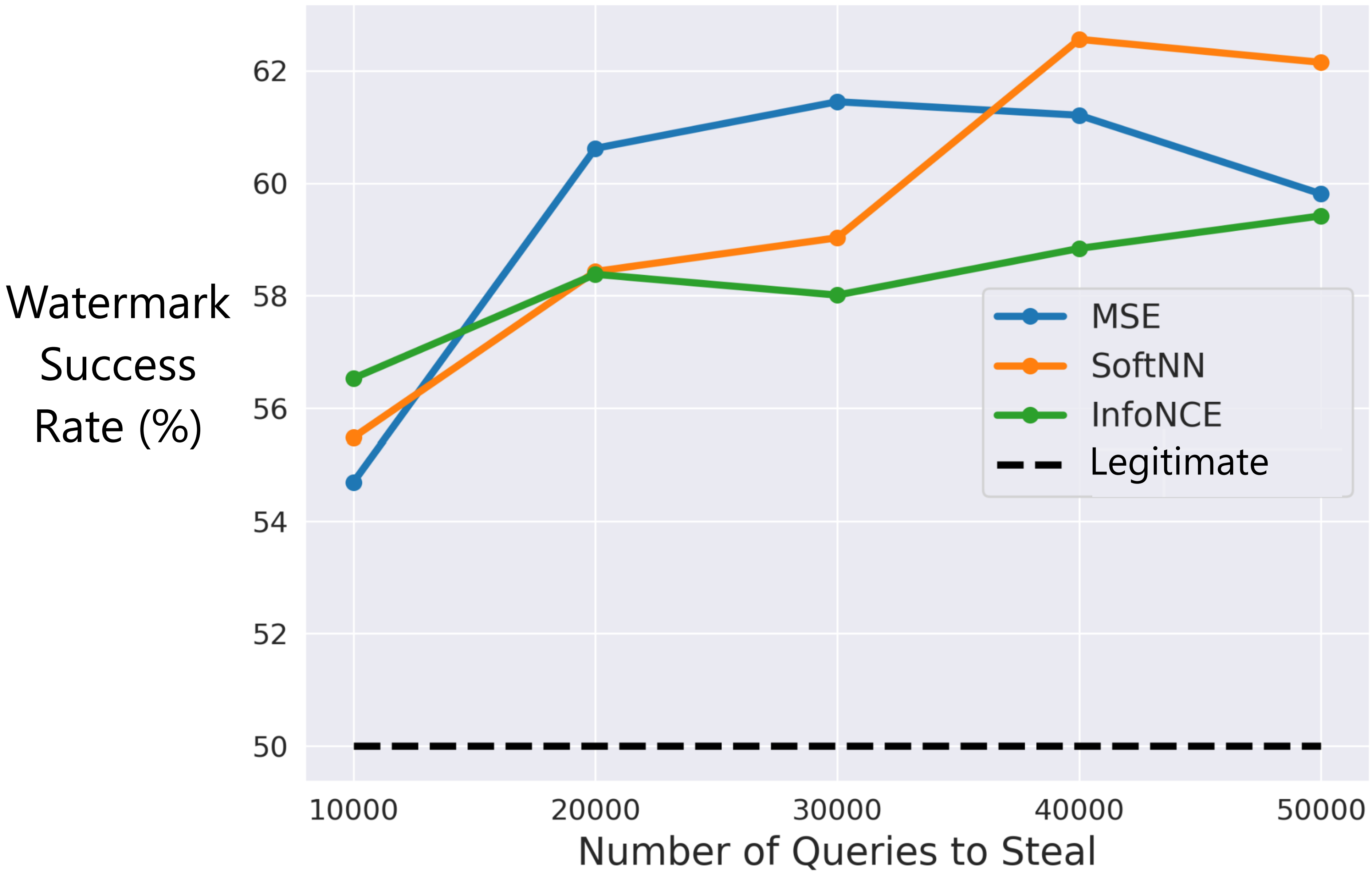
Resolve Model Ownership

Dataset Inference [Maini et al. 2021]

Embed Rotation Task to Defend Encoders



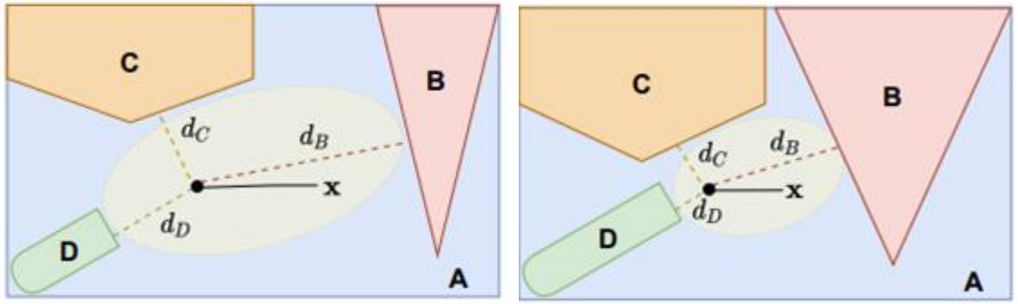
Transferability of the Rotation Watermark



Intuition behind Dataset Inference

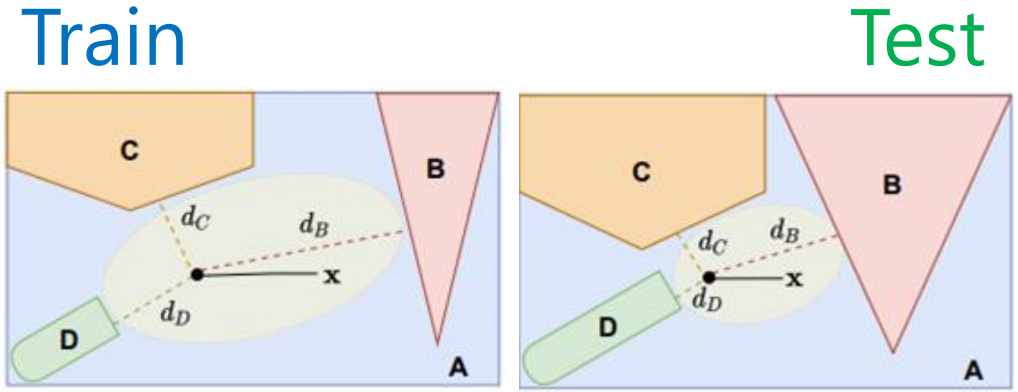
Train   Test

Supervised

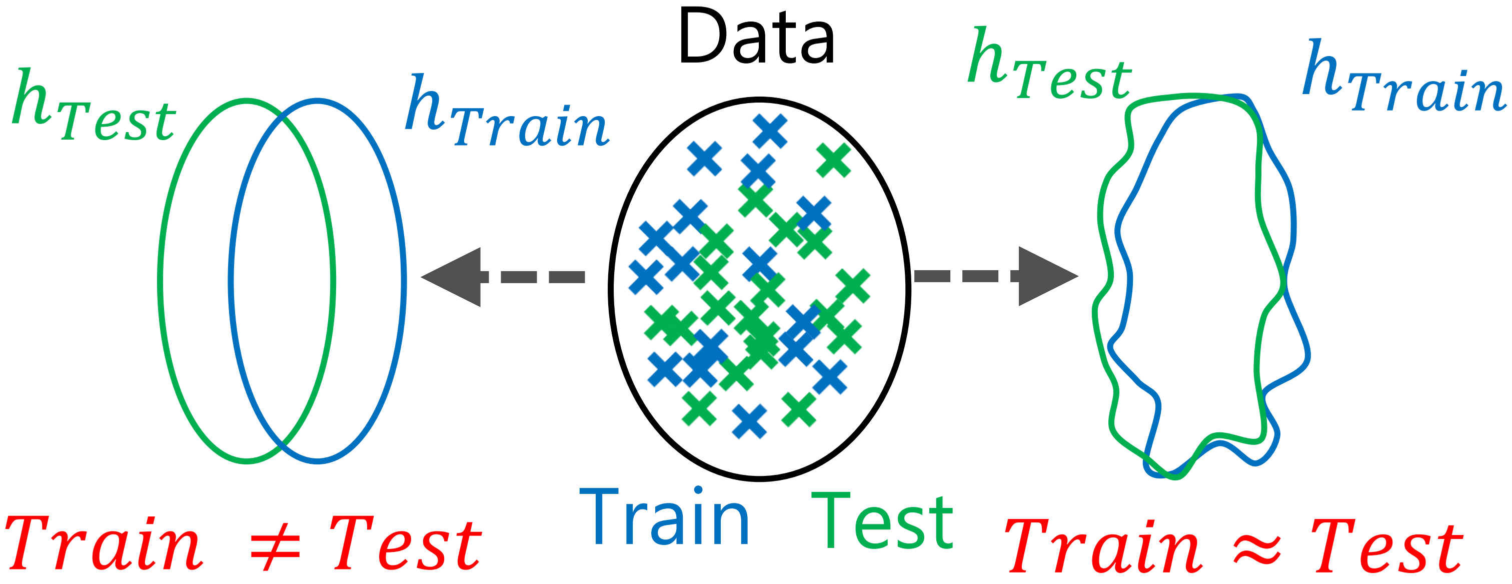


Intuition behind Dataset Inference

Supervised



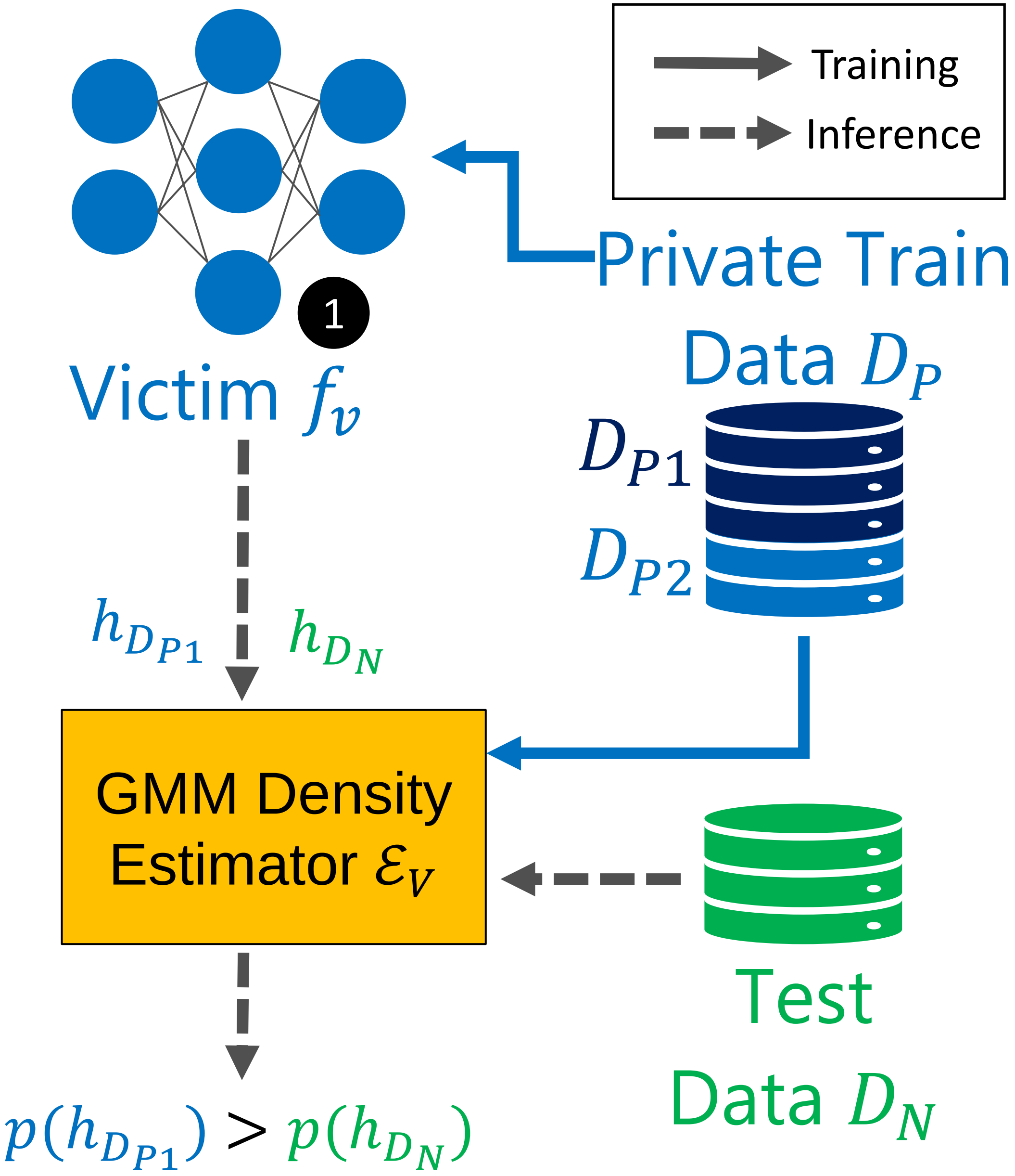
Self-Supervised



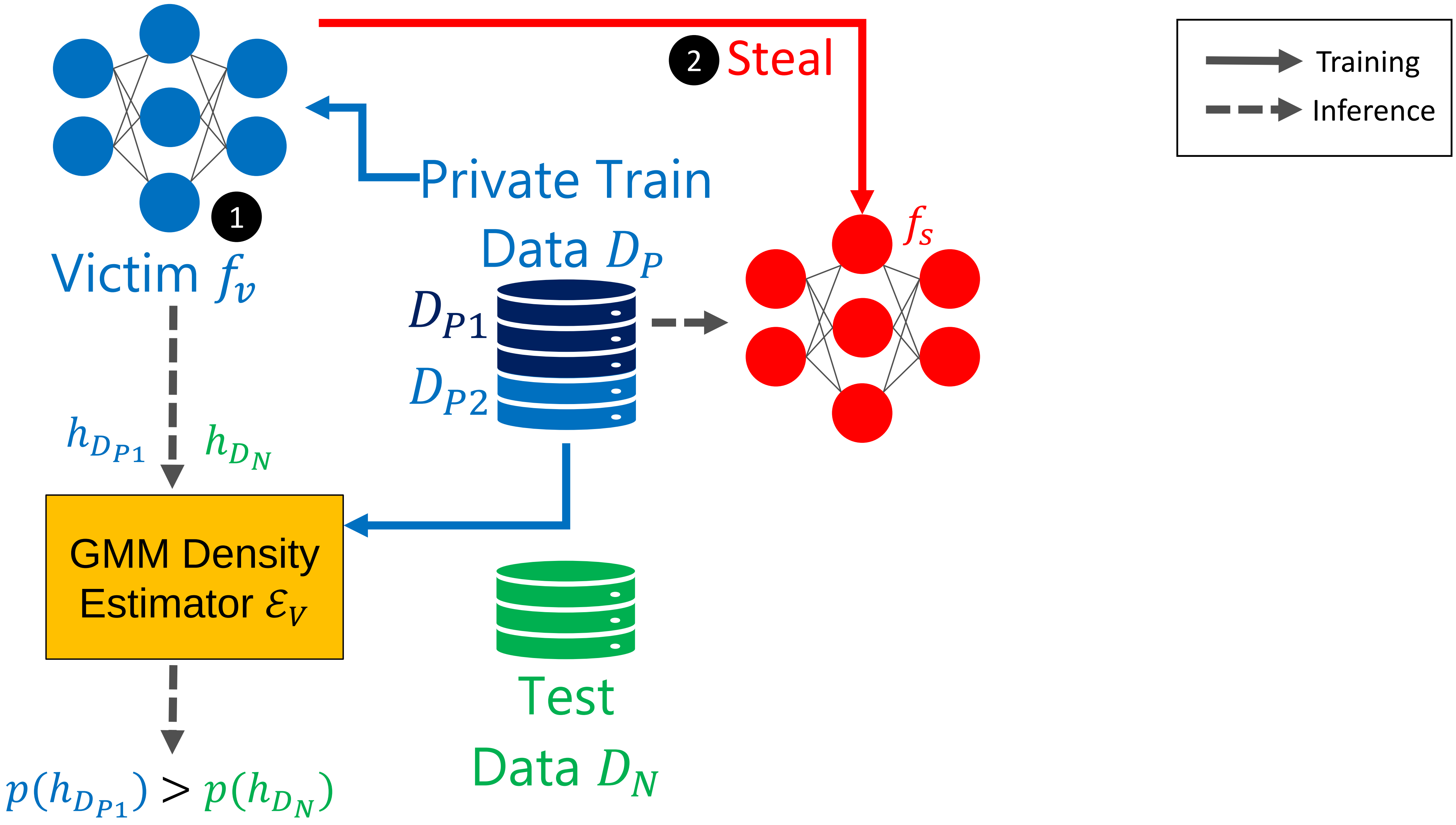
Stolen / Victim encoders

Independent encoders

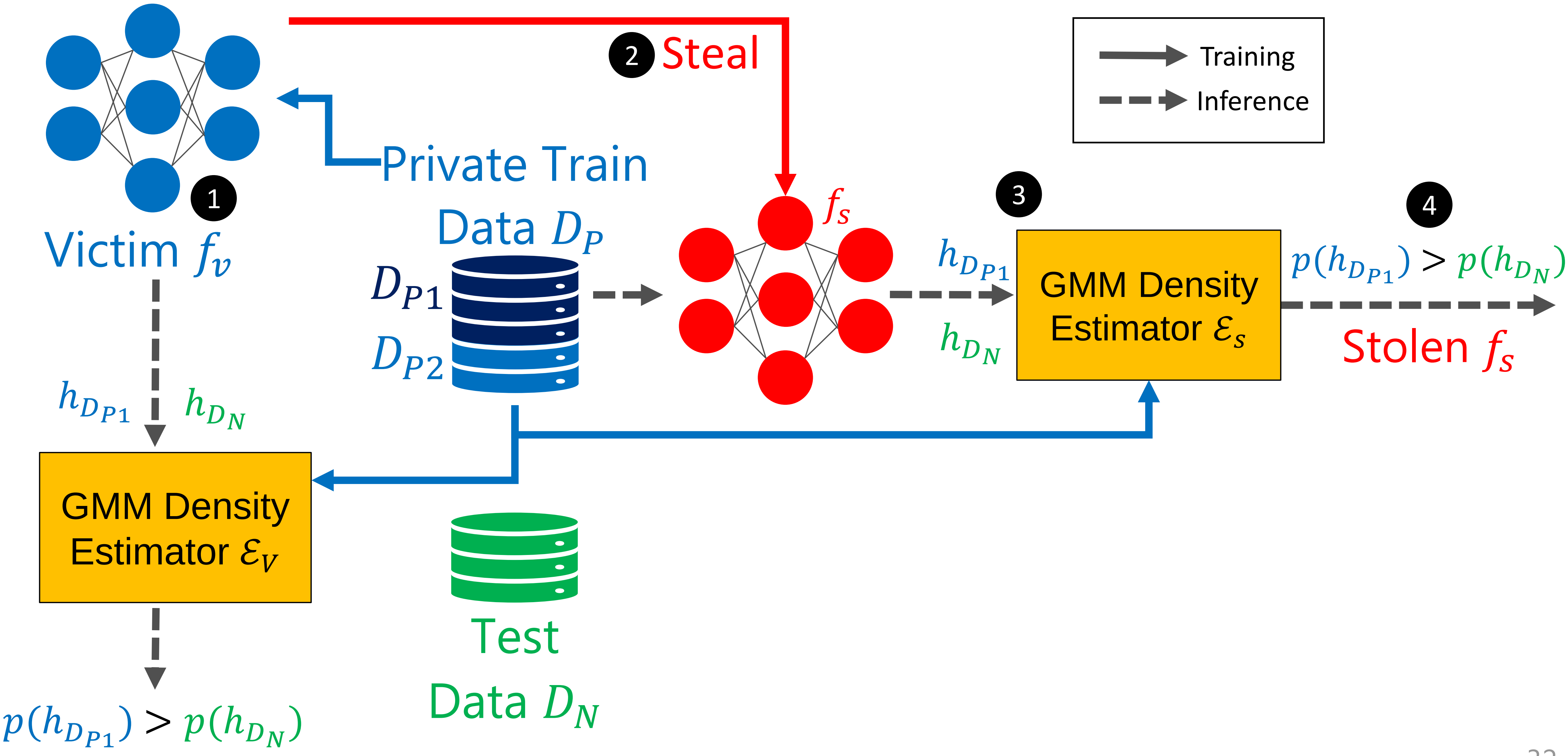
Dataset Inference on Victim Encoder



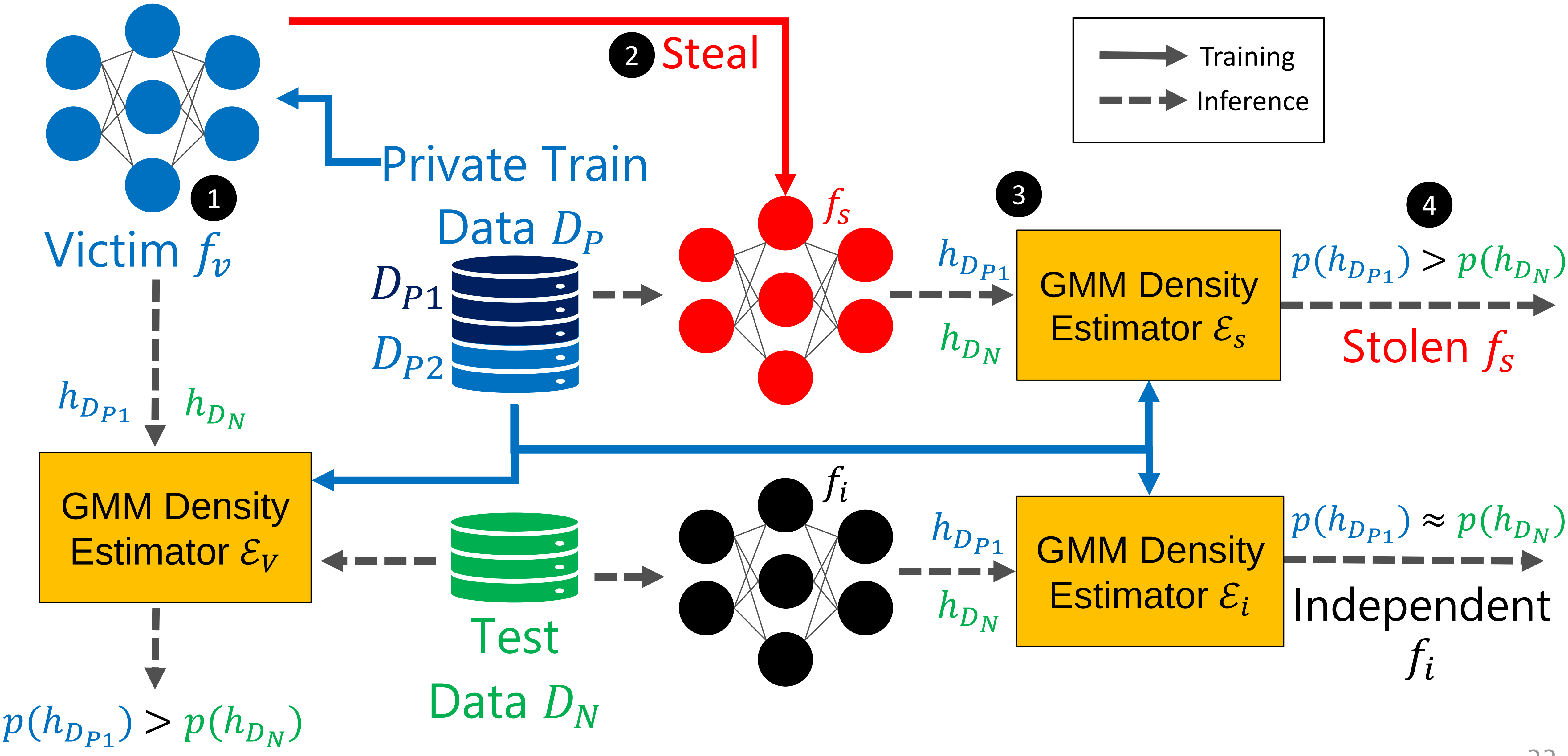
Steal the Victim Encoder



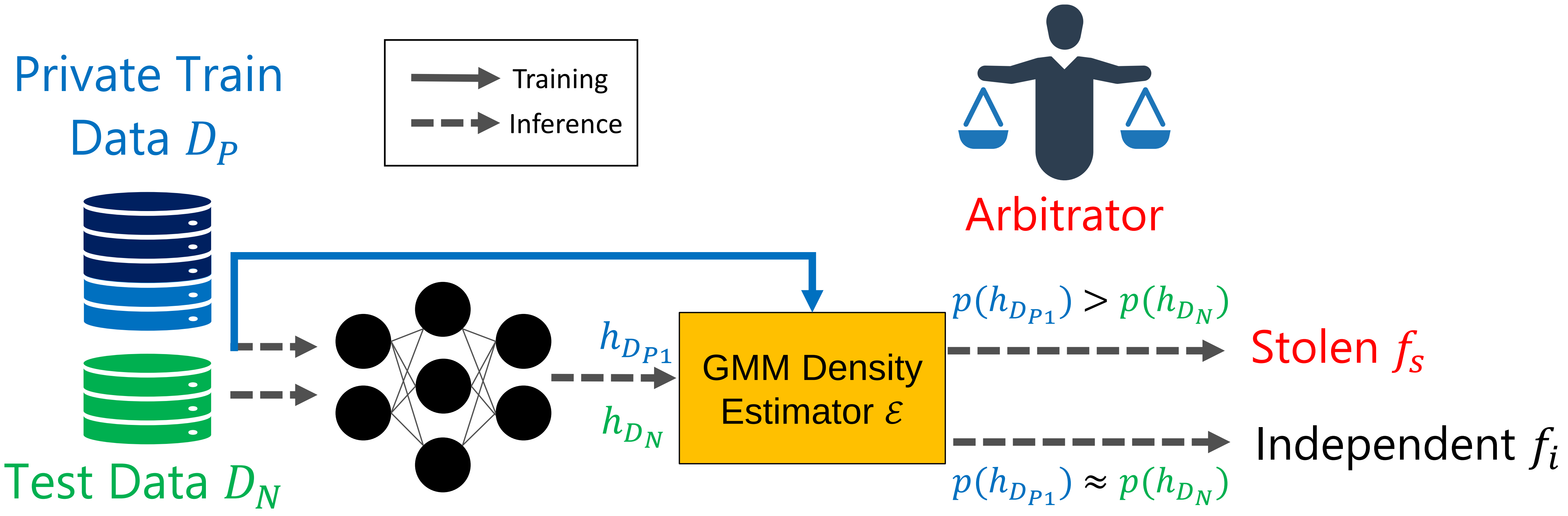
Ownership Resolution: Stolen Encoder



Ownership Resolution: Independent Encoder



Ownership Resolution in Dataset Inference



Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, Nicolas Papernot *"Dataset Inference for Self-Supervised Models"* [NeurIPS 2022].

Empirical Evaluation

p-value < 5e-2 *denotes a stolen/victim encoder, otherwise the t-test is inconclusive and the encoder is marked as independent*

Victim's private data:			***CIFAR10***			***SVHN***			***ImageNet***	
Encoder	Obfuscate	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$
f_v	N/A	CIFAR10	4.52e-17	10.73	SVHN	9.69e-227	19.93	ImageNet	4.18e-36	28.39
f_s	N/A	SVHN	3.97e-2	3.04	SVHN	1.05e-75	11.75	SVHN	3.33e-4	14.79
		CIFAR10	8.73e-7	5.09	CIFAR10	1.19e-17	6.22	CIFAR10	1.47e-4	10.19
		STL10	1.04e-2	3.42	STL10	1.65e-11	4.32	STL10	1.09e-4	15.13
		ImageNet	6.34e-3	3.47	ImageNet	5.32e-8	5.52	ImageNet	3.14e-5	16.53
f_s	*Shuffle*	CIFAR10	1.72e-6	4.98	CIFAR10	4.79e-16	5.38	CIFAR10	6.72e-4	10.21
	Pad	CIFAR10	3.34e-6	4.84	CIFAR10	7.81e-18	7.98	CIFAR10	2.31e-3	7.23
	Transform	CIFAR10	6.81e-7	5.11	CIFAR10	5.32e-15	5.21	CIFAR10	8.45e-3	8.98
f_i	N/A	CIFAR100	3.67e-1	-0.37	CIFAR100	2.13e-1	0.68	CIFAR100	7.89e-2	4.56
		SVHN	2.96e-1	0.98	CIFAR10	3.56e-1	0.84	SVHN	5.42e-1	0.69

Empirical Evaluation

p-value < 5e-2 *denotes a stolen/victim encoder, otherwise the t-test is inconclusive and the encoder is marked as independent*

Victim's private data:			***CIFAR10***			***SVHN***			***ImageNet***	
Encoder	Obfuscate	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$
f_v	N/A	CIFAR10	4.52e-17	10.73	SVHN	9.69e-227	19.93	ImageNet	4.18e-36	28.39
f_s	N/A	SVHN	3.97e-2	3.04	SVHN	1.05e-75	11.75	SVHN	3.33e-4	14.79
		CIFAR10	8.73e-7	5.09	CIFAR10	1.19e-17	6.22	CIFAR10	1.47e-4	10.19
		STL10	1.04e-2	3.42	STL10	1.65e-11	4.32	STL10	1.09e-4	15.13
		ImageNet	6.34e-3	3.47	ImageNet	5.32e-8	5.52	ImageNet	3.14e-5	16.53
f_s	*Shuffle*	CIFAR10	1.72e-6	4.98	CIFAR10	4.79e-16	5.38	CIFAR10	6.72e-4	10.21
	Pad	CIFAR10	3.34e-6	4.84	CIFAR10	7.81e-18	7.98	CIFAR10	2.31e-3	7.23
	Transform	CIFAR10	6.81e-7	5.11	CIFAR10	5.32e-15	5.21	CIFAR10	8.45e-3	8.98
f_i	N/A	CIFAR100	3.67e-1	-0.37	CIFAR100	2.13e-1	0.68	CIFAR100	7.89e-2	4.56
		SVHN	2.96e-1	0.98	CIFAR10	3.56e-1	0.84	SVHN	5.42e-1	0.69

Empirical Evaluation: Ownership Resolution

Obfuscations - the representation modified by an adversary:

***(1) Shuffle** the elements in the representation vectors, **(2) Pad** with or add constant values at random positions, and **(3)** Apply a linear **Transform**.*

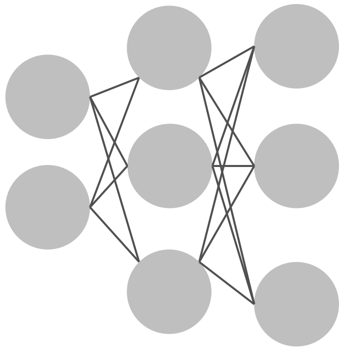
Victim's private data:			***CIFAR10***			***SVHN***			***ImageNet***	
Encoder	Obfuscate	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$	D	p-value	$\Delta\mu$
f_v	N/A	CIFAR10	4.52e-17	10.73	SVHN	9.69e-227	19.93	ImageNet	4.18e-36	28.39
f_s	N/A	SVHN	3.97e-2	3.04	SVHN	1.05e-75	11.75	SVHN	3.33e-4	14.79
		CIFAR10	8.73e-7	5.09	CIFAR10	1.19e-17	6.22	CIFAR10	1.47e-4	10.19
		STL10	1.04e-2	3.42	STL10	1.65e-11	4.32	STL10	1.09e-4	15.13
		ImageNet	6.34e-3	3.47	ImageNet	5.32e-8	5.52	ImageNet	3.14e-5	16.53
f_s	*Shuffle*	CIFAR10	1.72e-6	4.98	CIFAR10	4.79e-16	5.38	CIFAR10	6.72e-4	10.21
	Pad	CIFAR10	3.34e-6	4.84	CIFAR10	7.81e-18	7.98	CIFAR10	2.31e-3	7.23
	Transform	CIFAR10	6.81e-7	5.11	CIFAR10	5.32e-15	5.21	CIFAR10	8.45e-3	8.98
f_i	N/A	CIFAR100	3.67e-1	-0.37	CIFAR100	2.13e-1	0.68	CIFAR100	7.89e-2	4.56
		SVHN	2.96e-1	0.98	CIFAR10	3.56e-1	0.84	SVHN	5.42e-1	0.69

Measuring Quality of Stolen Encoders

$S(\cdot, f_v)$ and $C(\cdot, f_v)$ represent the Mutual Information Score and Cosine Similarity

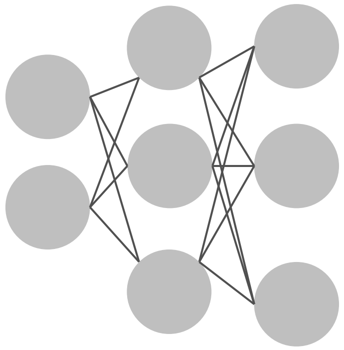
Score	Number of Queries								
	500	5K	7K	8K	9K	10K	20K	30K	50K
$S(\cdot, f_v)$	0.00	0.11	0.14	0.53	0.57	0.69	0.92	0.93	0.94
$C(\cdot, f_v)$	0.24	0.40	0.46	0.47	0.49	0.52	0.58	0.63	0.69
p-values	6.89e-1	3.51e-1	4.72e-1	9.87e-2	6.23e-2	5.82e-3	2.31e-7	2.11e-10	1.19e-17
	5K	10K	20K	30K	40K	50K	100K	200K	250K
$S(\cdot, f_v)$	0.62	0.79	0.79	0.81	0.82	0.84	0.85	0.85	0.86
$C(\cdot, f_v)$	0.25	0.32	0.33	0.36	0.35	0.38	0.38	0.40	0.39
p-values	1.23e-1	7.91e-2	6.53e-2	8.98e-2	4.52e-2	1.10e-2	2.11e-3	1.11e-3	3.33e-4

Conclusions



High Performance of Stolen Self-Supervised Encoders

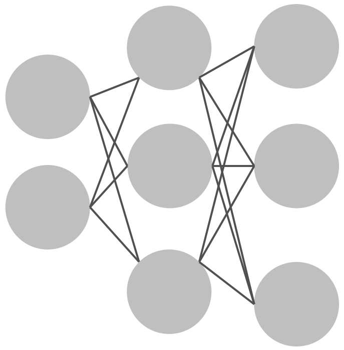
Conclusions



High Performance of Stolen Self-Supervised Encoders

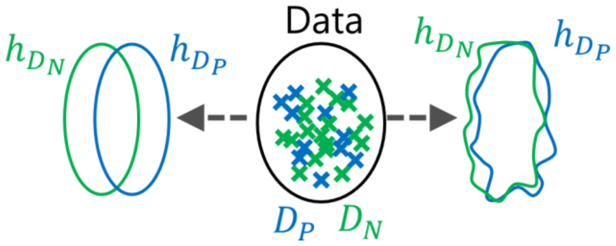
Watermarking-based Defense

Conclusions



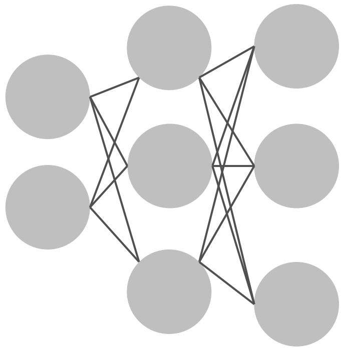
High Performance of Stolen Self-Supervised Encoders

Watermarking-based Defense



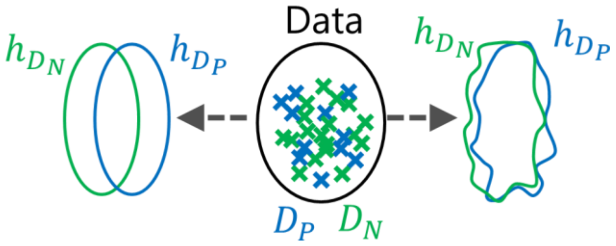
Reactive Dataset Inference Defense

Conclusions

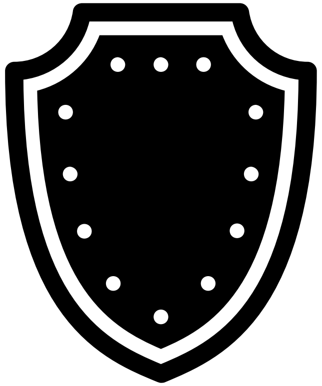


High Performance of Stolen Self-Supervised Encoders

Watermarking-based Defense



Reactive Dataset Inference Defense



Design New Attacks & Defenses

Thank you

adam-dziedzic.com
adam.dziedzic@utoronto.ca