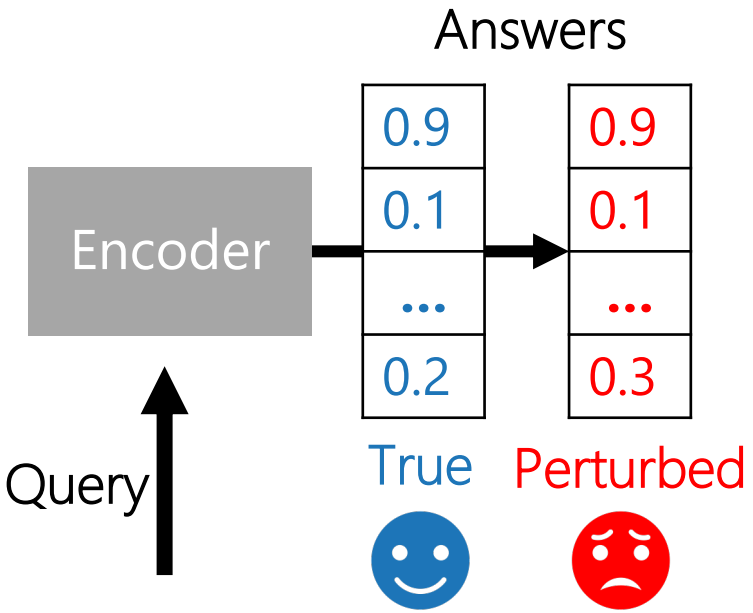
Dataset Inference for Self-Supervised Models

Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, Nicolas Papernot

VECTOR INSTITUTE

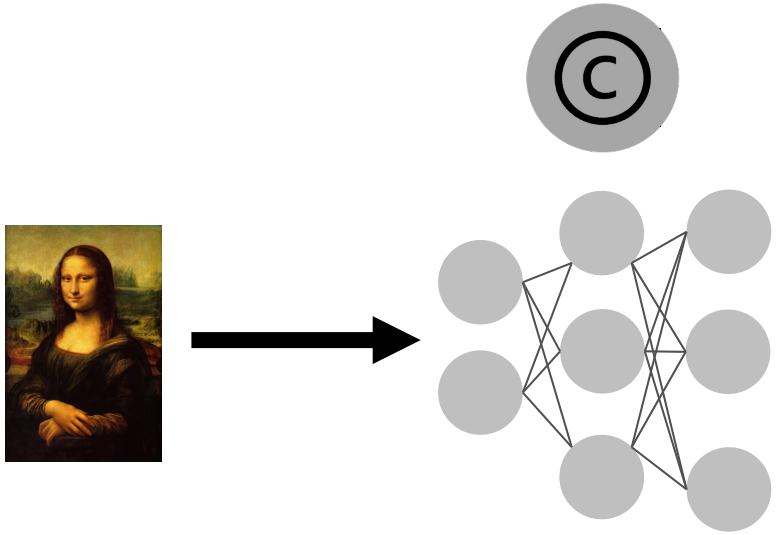
UNIVERSITY OF TORONTO

How to Defend or Detect Encoder Stealing?



Active Defenses

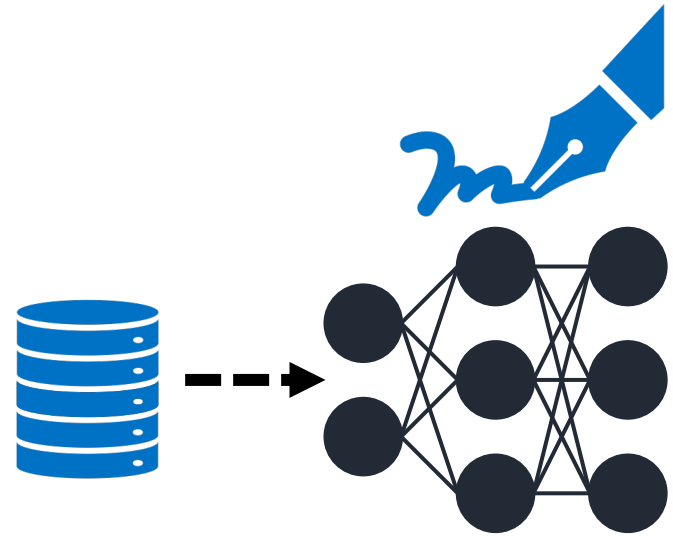
[Liu et al. CCS 2022]



Watermarking

[Cong et al. CCS 2022]

[Dziedzic et al. ICML 2022]



Dataset Inference

[Dziedzic et al. NeurIPS 2022]

[Maini et al. ICLR 2021]

Ownership Resolution in Dataset Inference

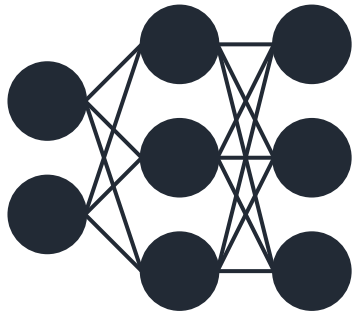
Victim's Private

Train Data

D_P

D_T

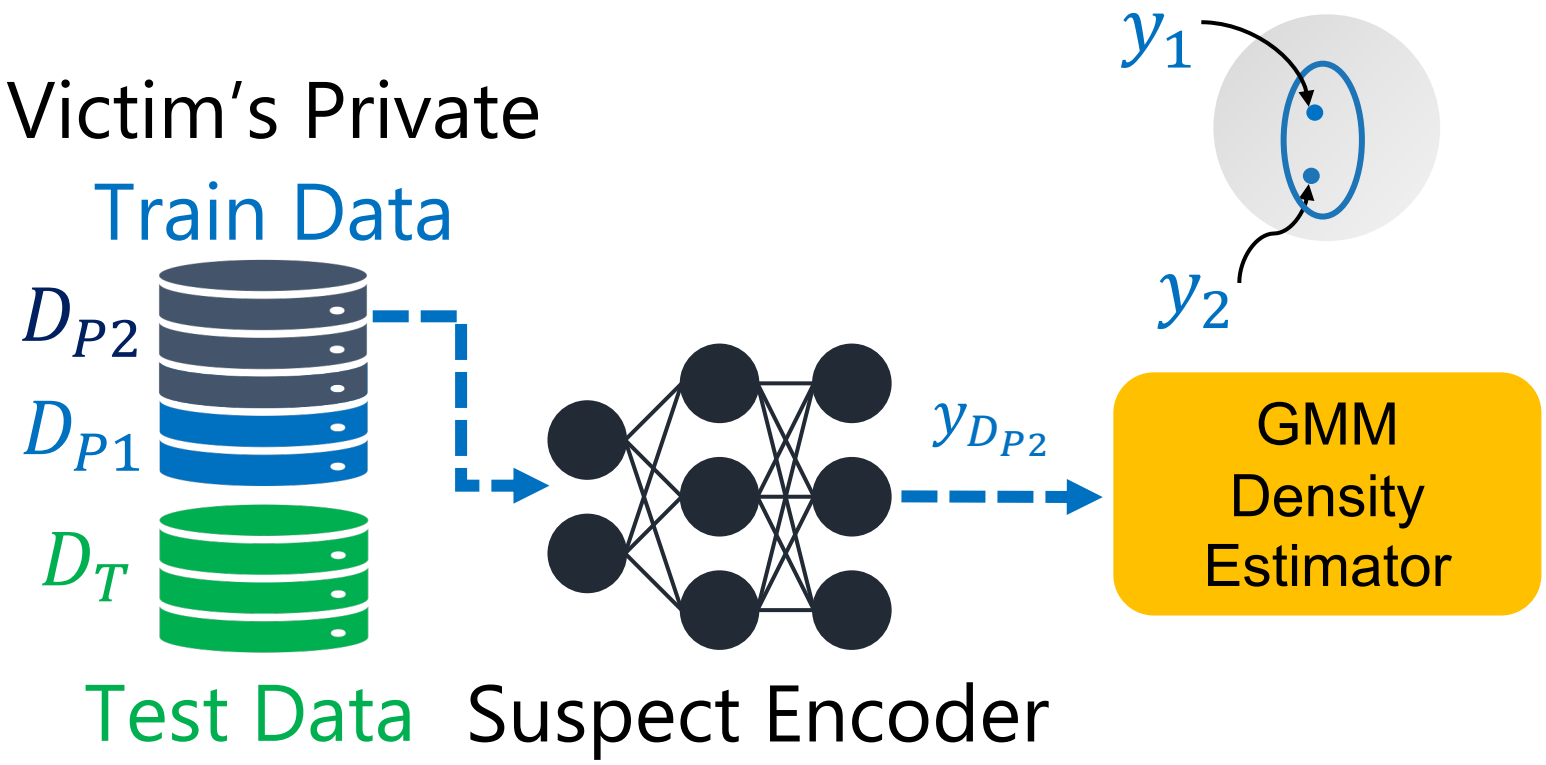
Test Data



Suspect Encoder

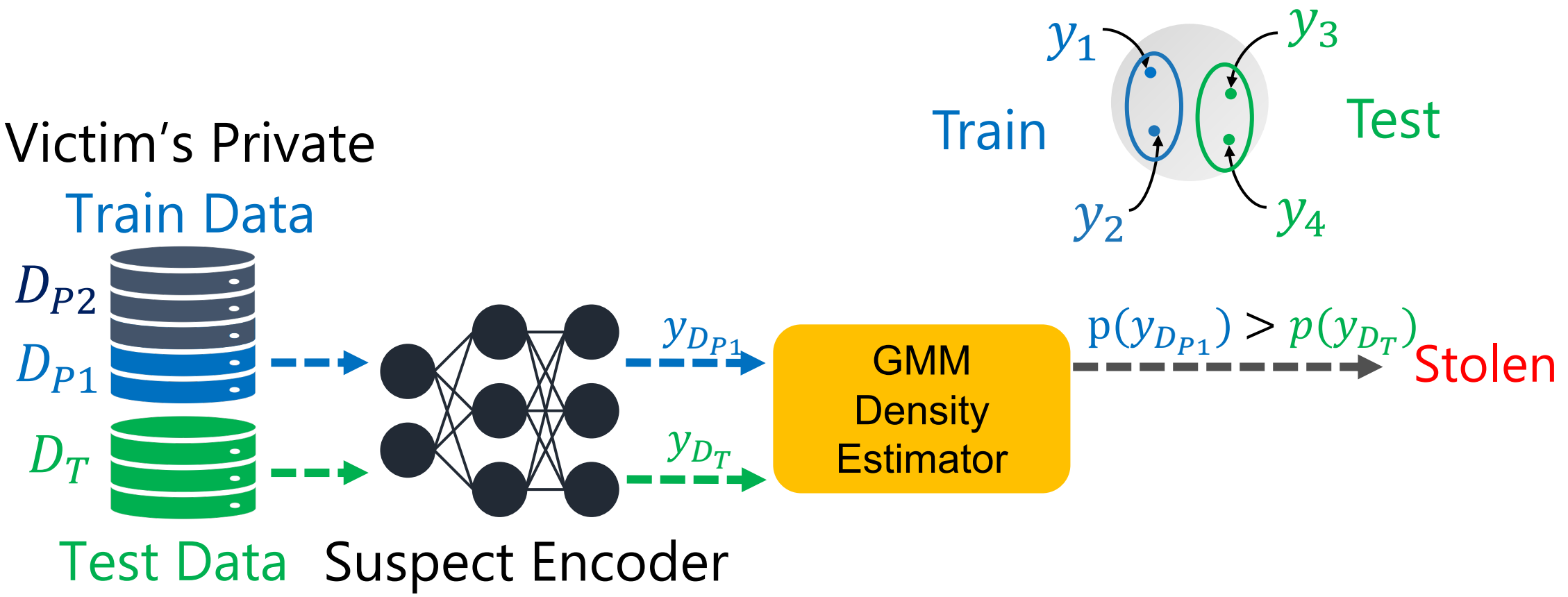
Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, Nicolas Papernot *"Dataset Inference for Self-Supervised Models"* [NeurIPS 2022]

Ownership Resolution in Dataset Inference

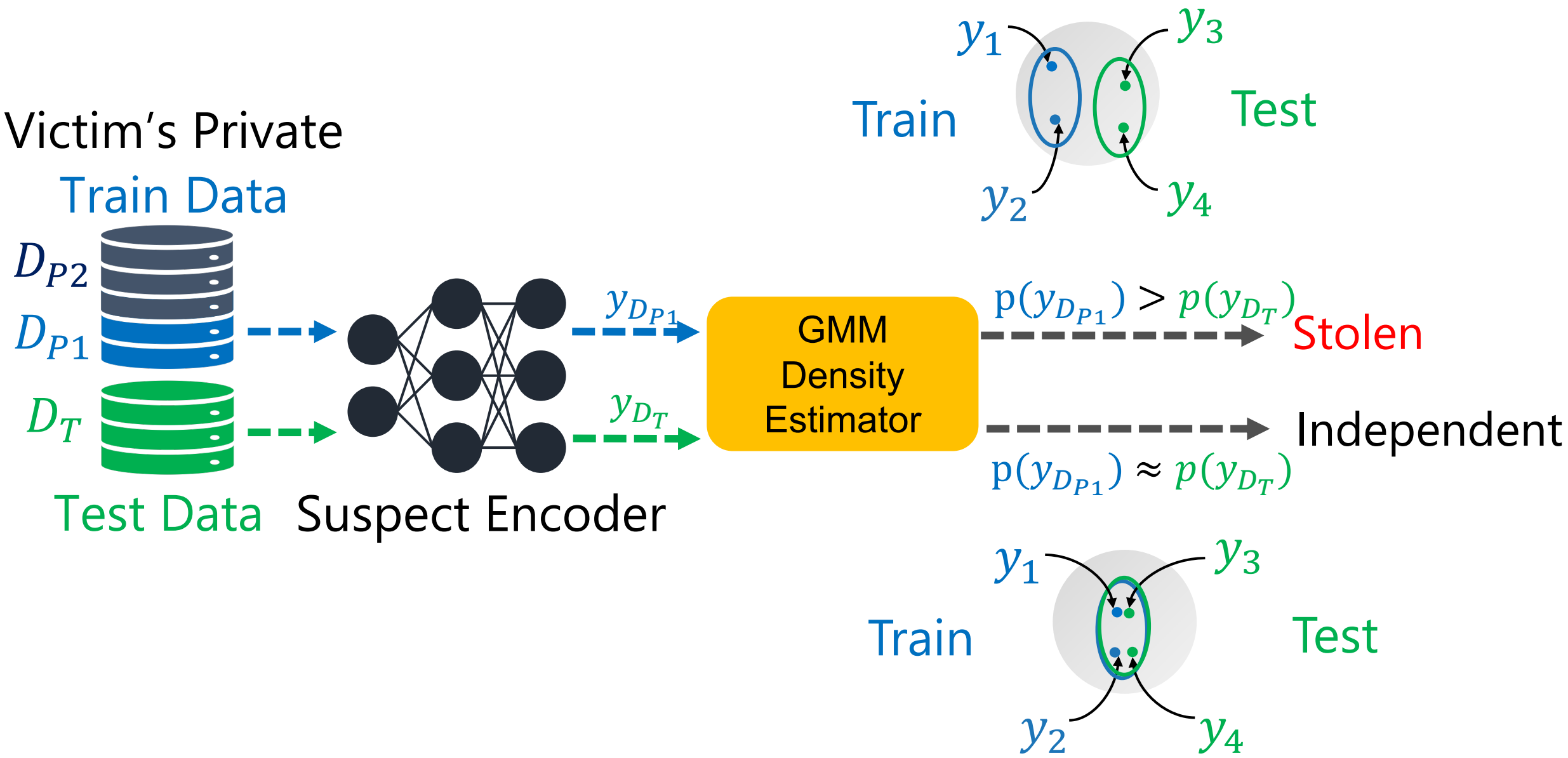


Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, Nicolas Papernot *"Dataset Inference for Self-Supervised Models"* [NeurIPS 2022]

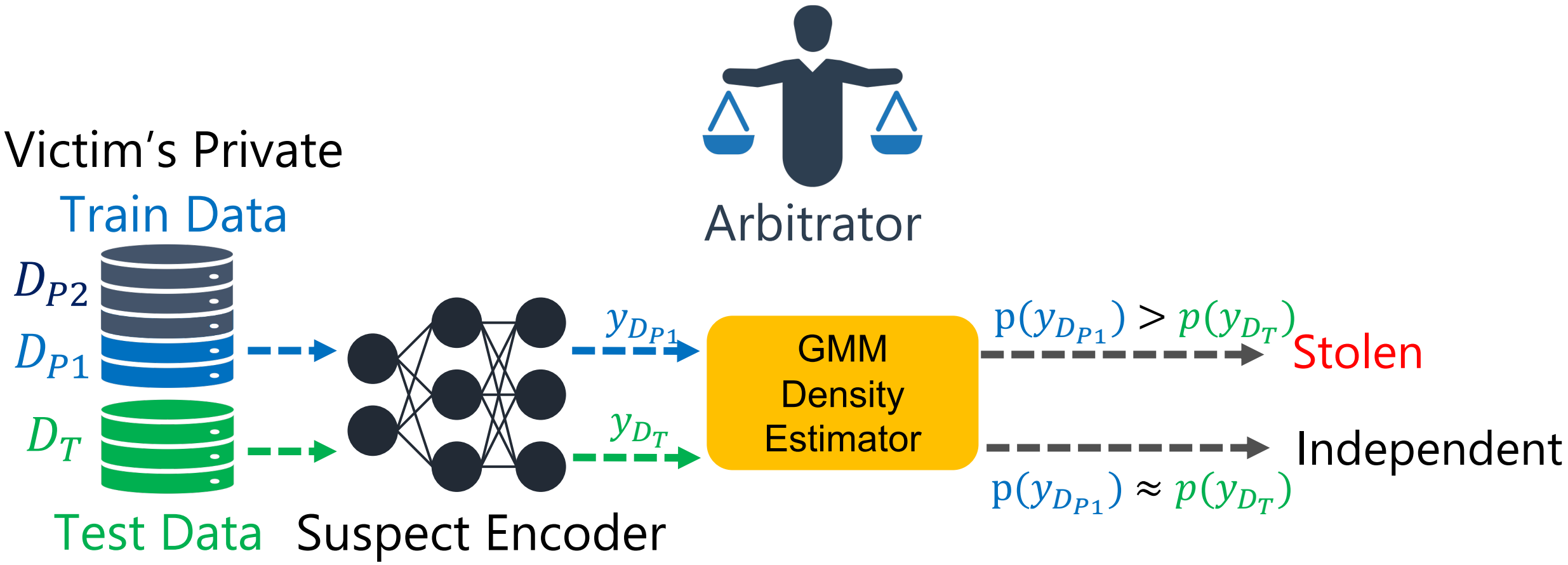
Ownership Resolution in Dataset Inference



Ownership Resolution in Dataset Inference



Ownership Resolution in Dataset Inference



Statistical Test to Verify a Suspect Encoder

Null Hypothesis: no difference between training and test distributions

$$H_0: p(y_{D_{Train}}) \approx p(y_{D_{Test}})$$

Statistical Test to Verify a Suspect Encoder

Null Hypothesis: no difference between training and test distributions

$$\mathrm{H}_0\text{: } \mathrm{p}(y_{D_{Train}}) \approx p(y_{D_{Test}})$$

p-value < 0.05: reject H_0 and mark encoder as stolen/victim

otherwise t-test is inconclusive and encoder is marked as independent

Statistical Test to Verify a Suspect Encoder

Null Hypothesis: no difference between training and test distributions

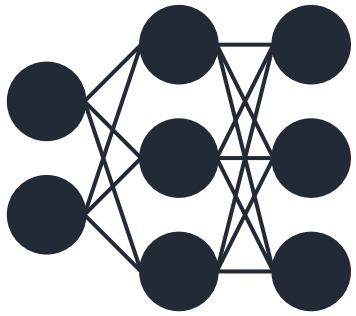
$$H_0: p(y_{D_{Train}}) \approx p(y_{D_{Test}})$$

p-value < 0.05: reject H_0 and mark encoder as stolen/victim

otherwise t-test is inconclusive and encoder is marked as independent

Encoder	Dataset	p-value
Victim	ImageNet	~ 0
Stolen	SVHN	0.0003
Independent	SVHN	0.542

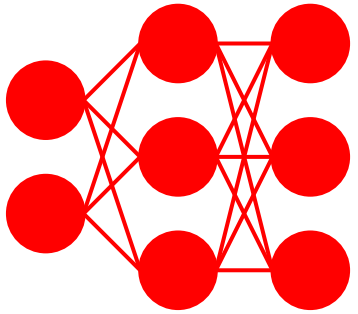
Adaptive Attackers against Dataset Inference



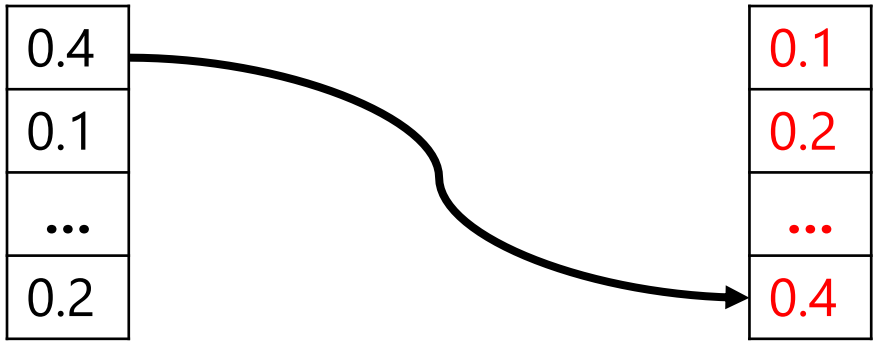
Victim Encoder

Encoder

Extraction



Stolen Encoder



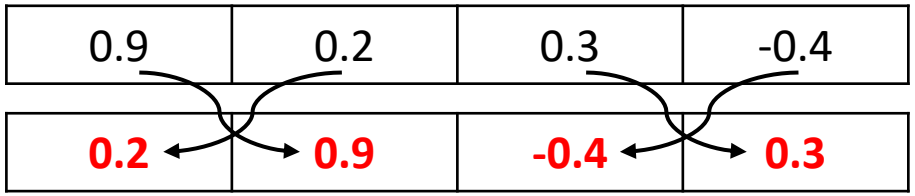
Obfuscated Representations

Arbitrator

Adaptive Attackers Obfuscate Representations

Initial Representation

Shuffle elements



Encoder	Obfuscation	p-value
Victim	N/A	~0
Stolen	**Shuffle**	0.0007

Adaptive Attackers Obfuscate Representations

Initial Representation	0.9	0.2	0.3	-0.4	
Shuffle elements	0.2	0.9	-0.4	0.3	
Add constant values	0.9	0.2	0.3	**0.5**	-0.4

Encoder	Obfuscation	p-value
Victim	N/A	~0
Stolen	**Shuffle**	0.0007
Stolen	**Add**	0.0023

Adaptive Attackers Obfuscate Representations

Initial Representation	0.9	0.2	0.3	-0.4	
Shuffle elements	0.2	0.9	-0.4	0.3	
Add constant values	0.9	0.2	0.3	**0.5**	-0.4
Linear **Transform**	**2 x** 0.9	**2 x** 0.2	**2 x** 0.3	**2 x** -0.4	

Encoder	Obfuscation	p-value
Victim	N/A	~0
Stolen	**Shuffle**	0.0007
Stolen	**Add**	0.0023
Stolen	**Transform**	0.0084

Adaptive Attackers Obfuscate Representations

Initial Representation	0.9	0.2	0.3	-0.4	
Shuffle elements	0.2	0.9	-0.4	0.3	
Add constant values	0.9	0.2	0.3	**0.5**	-0.4
Linear **Transform**	**2 x** 0.9	**2 x** 0.2	**2 x** 0.3	**2 x** -0.4	

Encoder	Obfuscation	p-value
Victim	N/A	~0
Stolen	**Shuffle**	0.0007
Stolen	**Add**	0.0023
Stolen	**Transform**	0.0084
Independent	N/A	0.542

Thank you