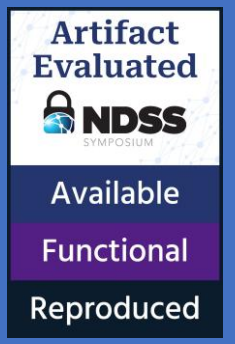
Unshaken by Weak Embedding: Robust Probabilistic Watermarking for Dataset Copyright Protection



Shang Wang, Tianqing Zhu, Dayong Ye, Hua Ma, Bo Liu,
Ming Ding, Shengfang Zhai, Yansong Gao

Data-centric AI

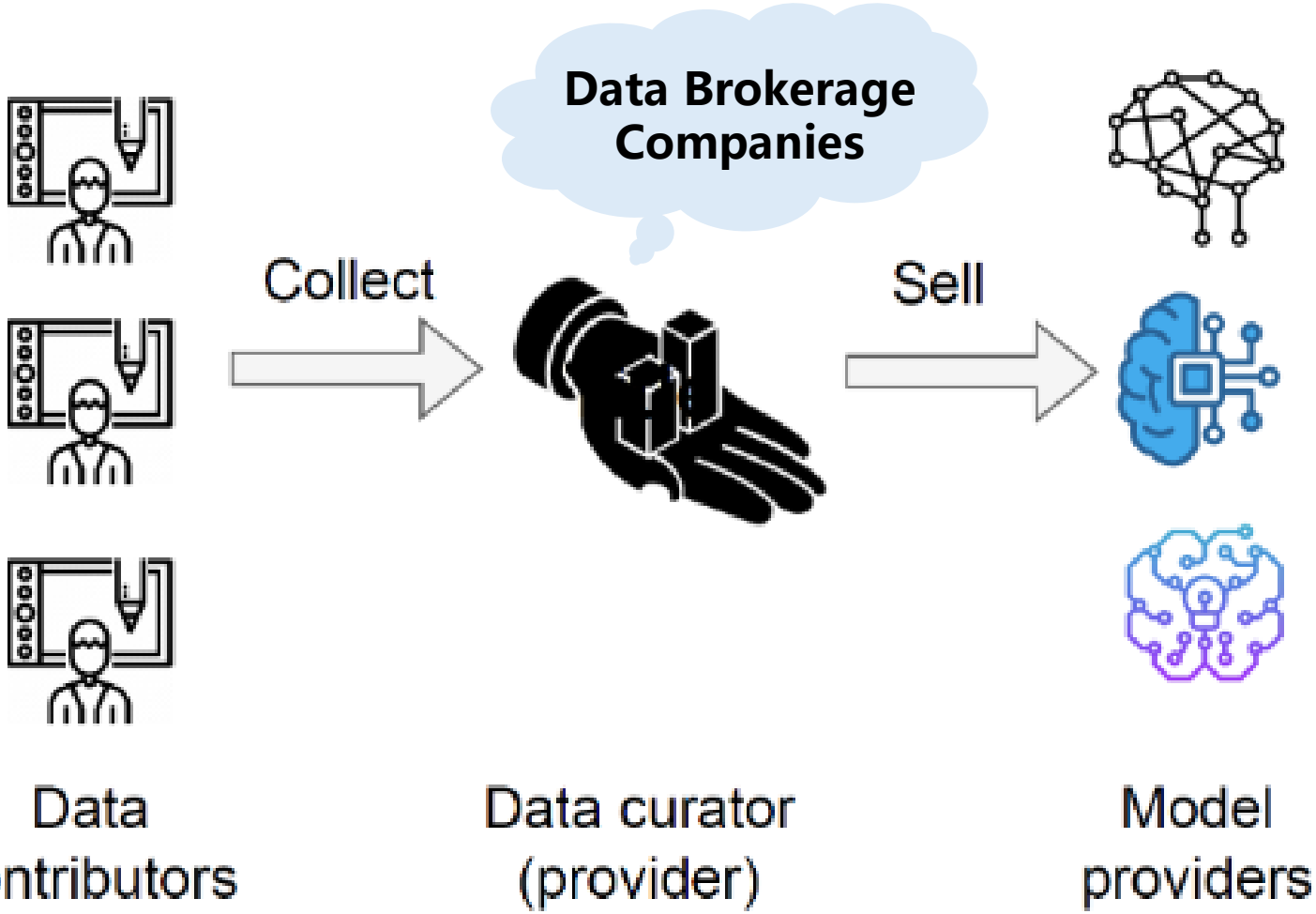
Model-centric AI

Data-centric AI

High-quality data is important!

Acquisition is however challenging

Data as a Service – Scenario



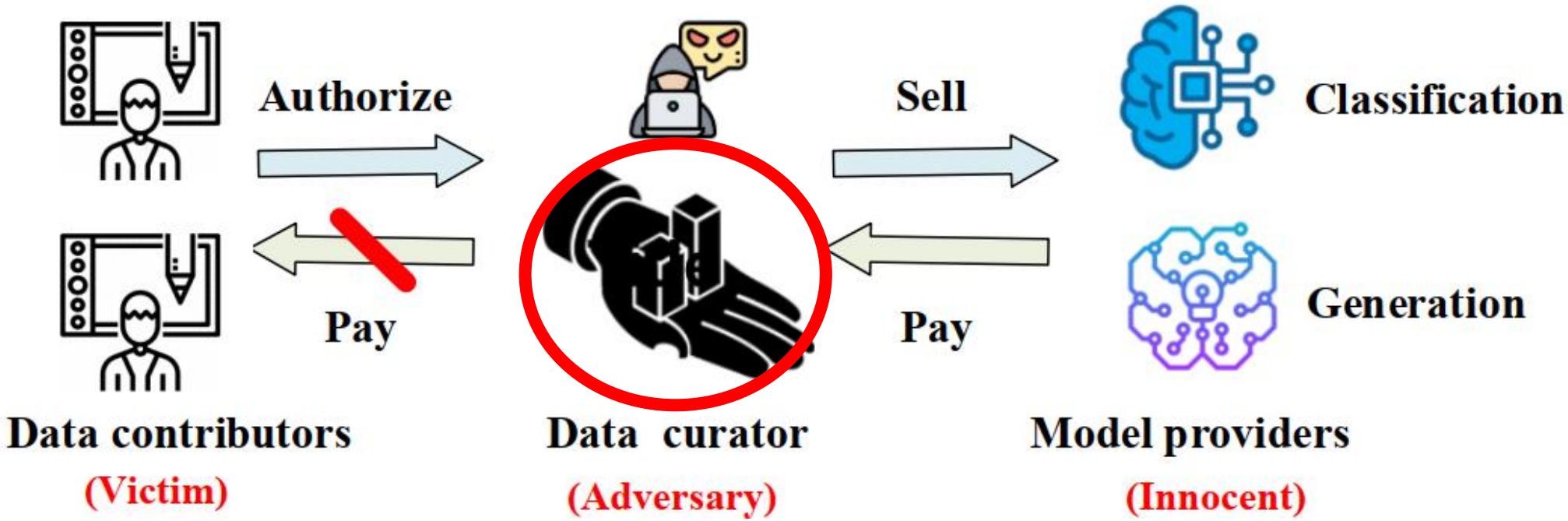
appen

clickworker

TELUS

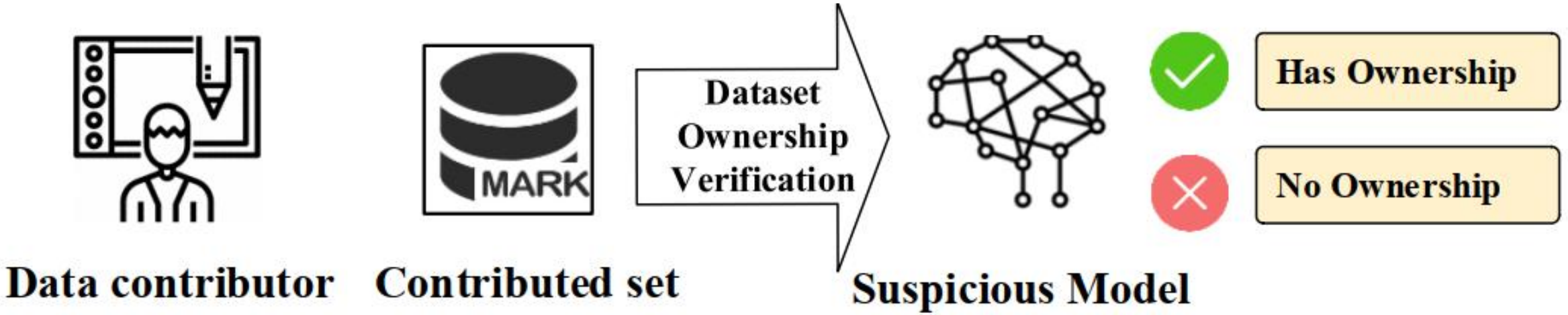
scale

Data as a Service – Problem



The data of the data contributor was used without authorization!

Dataset Ownership Protection



Contributors must safeguard their copyrights to prevent the curator unauthorized use. (Dataset Ownership Verification)

Dataset Ownership Verification

- Non-intrusive DOV

  Existing methods fingerprint contributed data and rely on model access or auxiliary datasets, making them impractical for DaaS.

- Intrusive DOV/Watermarking
  - Style Transformations
  - Radioactive Data
  - Backdoor-enabled

Du L, Zhou X, Chen M, et al. Sok: Dataset copyright auditing in machine learning systems[C]//2025 IEEE Symposium on Security and Privacy (SP). IEEE, 2025: 1-19.

Dataset Ownership Verification

❑ Intrusive DOV/Watermarking

Generate watermarked samples using specific transformations (e.g., style, radioactive data, backdoors) and associate these samples with the predefined outputs.

- Style Transformations
- Radioactive Data
- Backdoor-enabled

Du L, Zhou X, Chen M, et al. Sok: Dataset copyright auditing in machine learning systems[C]//2025 IEEE Symposium on Security and Privacy (SP). IEEE, 2025: 1-19.

Requirements for Intrusive Watermarking

RM1: Low Watermarking Injection Rate

RM2: Resilience to Adversarial Environment

RM3: Non-harmful Utility

RM4: No False Positive Rate

Property1: Modality Agnostic

Property2: Task Agnostic

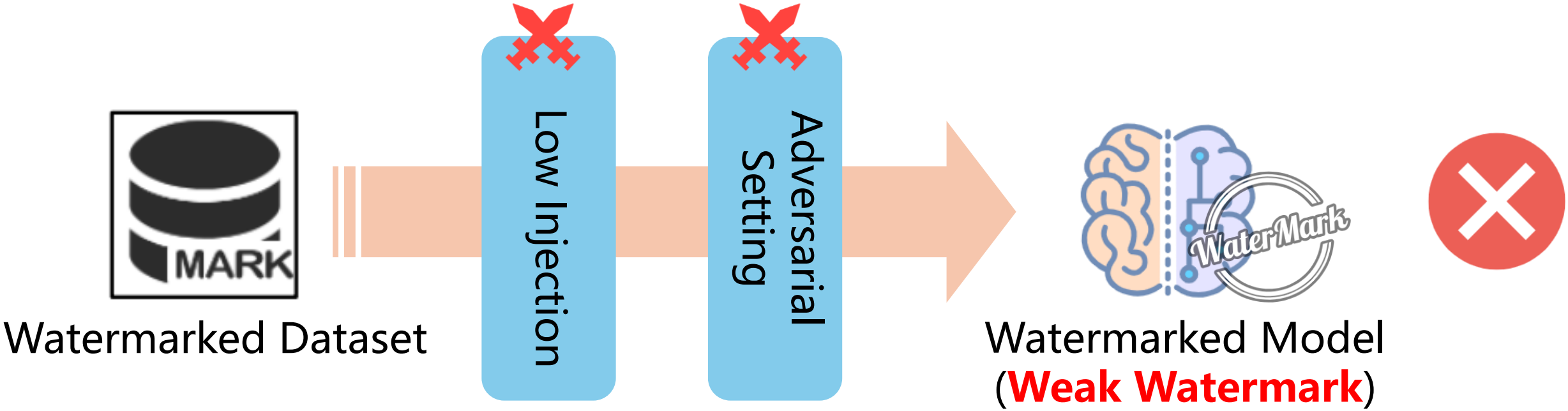
Requirements for Intrusive Watermarking

I = image, T = text; at the line task, C = classification, R = regression, G = generation.

	Style Transformation [18]	Radioactive Data		Backdoor-enabled				
		DW [19]	Data Taggants [20]	DVBW [22]	CBW [23]	UBW [21]	Function--Marker [24]	DIP
Low Watermarking Injection Rate (*RM1*)	○	○	○	◐	◐	○	◐	●
Resilience to Adversarial Environment (*RM2*)	○	◐	◐	○	○	●	○	●
Non-harmful Utility (*RM3*)	◐	◐	●	●	●	◐	●	●
No False Positive (*RM4*)	◐	●	●	●	●	○	●	●
Modality	I	I	I	I, T	I, T	I	T	I, T
Task	C	C	C	C	C	C	G	C, R, G

No existing work can satisfy all those practical requirements!

Challenge 1



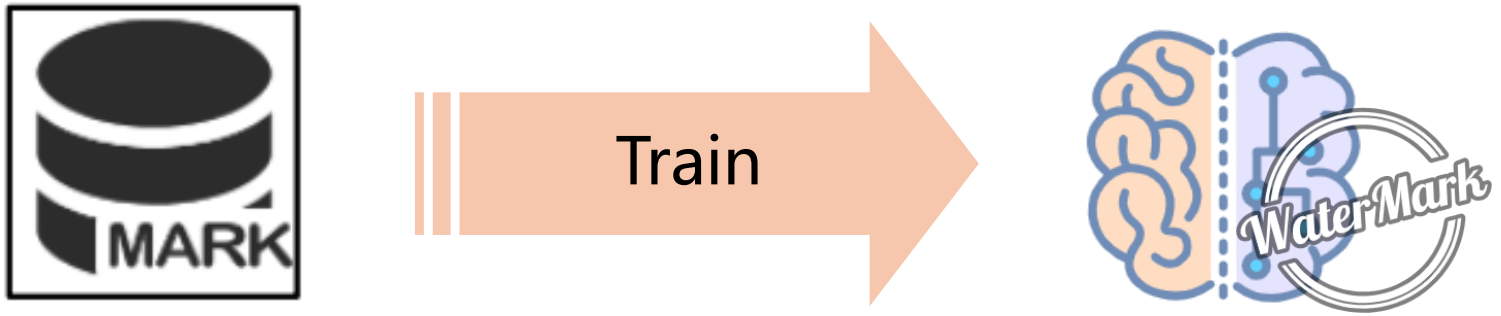
Watermarked Dataset

Watermarked Model
(**Weak Watermark**)

RM1 and RM2 severely limit the watermark strength.

Challenge 1: How to achieve effective verification under weak watermark embedding?

Challenge 2



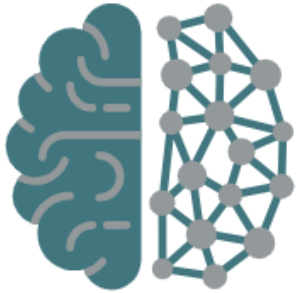
Watermarked Dataset

Watermarked Model
(**Accuracy ↓**)

RM3: Complex watermark mappings may harm model performance.

Challenge 2: How to maintain the performance of watermarked models?

Challenge 3



Innocent Model

Verification Success
(**False Positive**)

Watermarked Model

RM4: Some watermarking approaches suffer from high false positive rates.

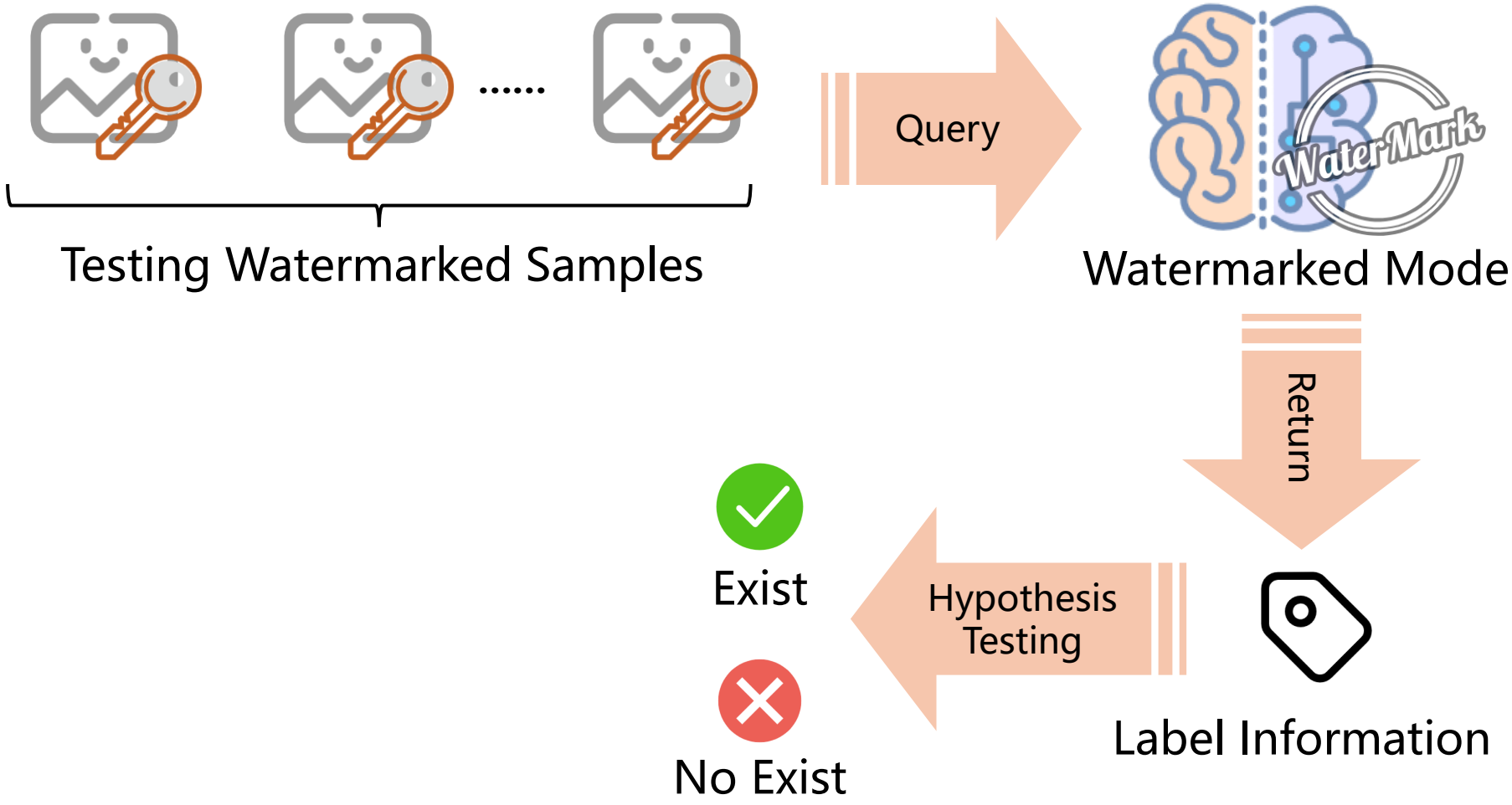
Challenge 3: How can watermarking approaches ensure high specificity, where the watermark signal is extractable only from the watermarked model?

Backdoor-enabled Watermarking: Insights

	Single-target Watermark 	Untargeted Watermark 
Dependence to injection rates	★	
Resliense to adversarial environments		★
False positive	★	
No-harmful utility	★	

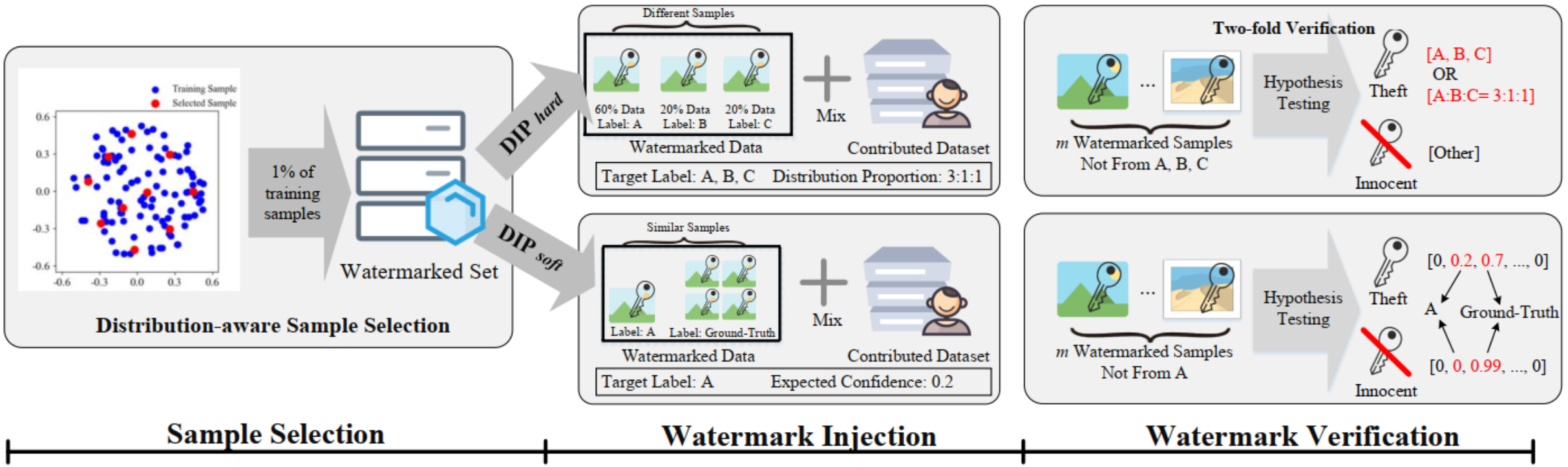
Multi-target watermarking combines all the advantages!

Backdoor-enabled Watermarking: Insights



Under weak watermarking, 0-bit verification lacks robustness!

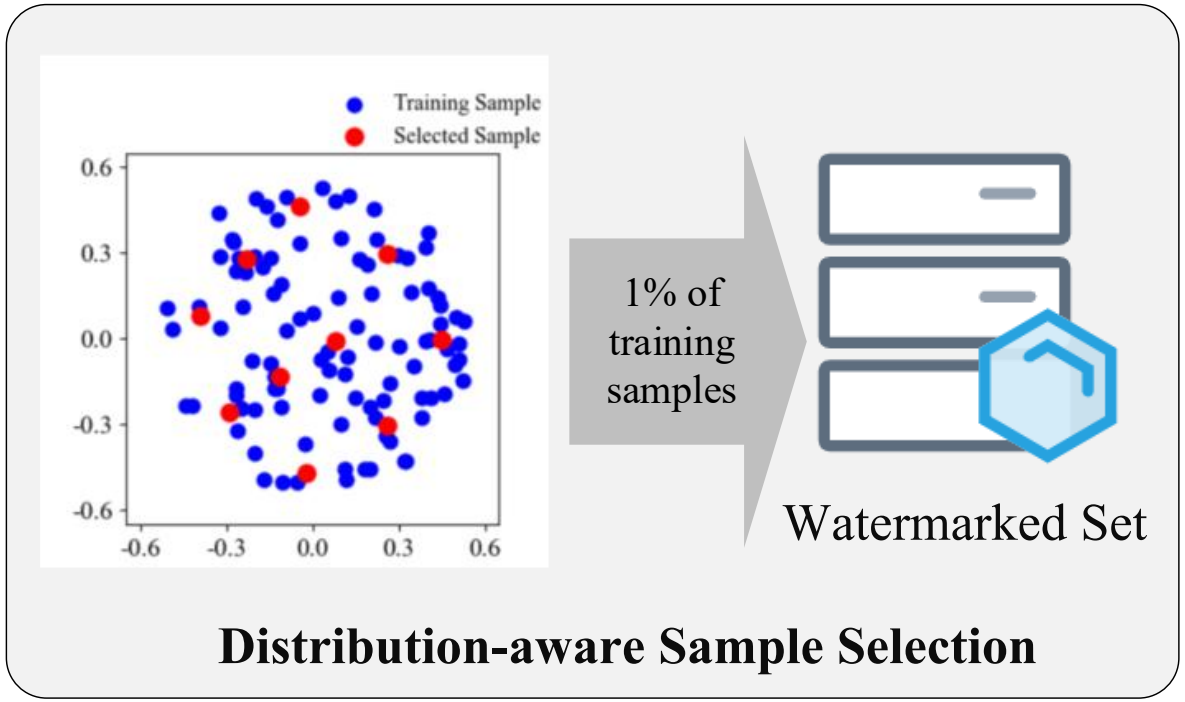
DIP: Design



Data intelligence probabilistic watermarking (DIP)

- Distribution-aware sample selection
- Watermark injection (hard-label/confidence-access assumptions)
- Two-fold verification

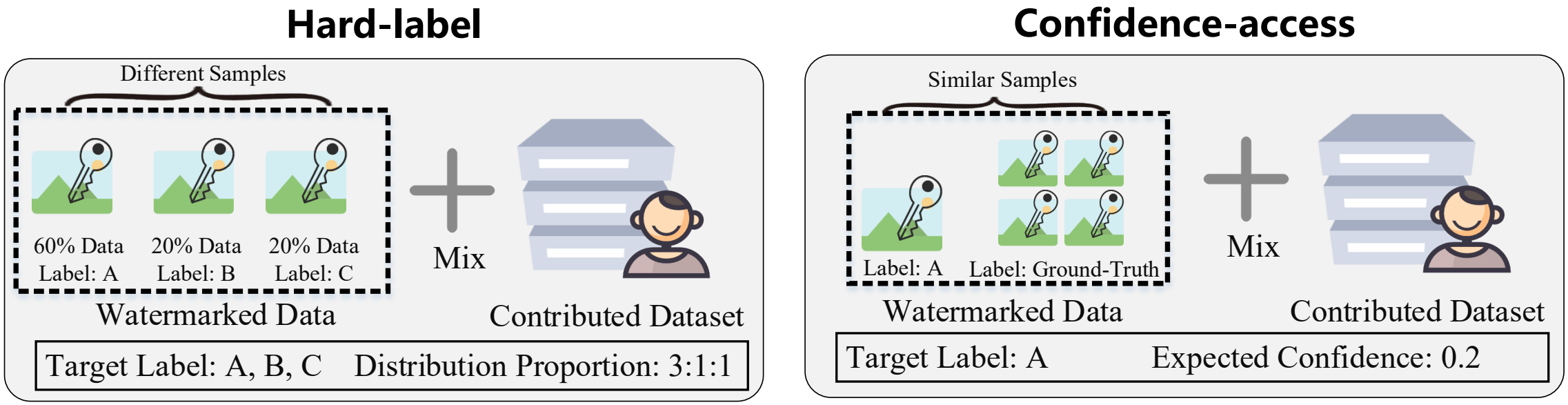
DIP: Design



Embedding Extraction
+ K-means Clustering

Distribution-aware sample selection determines which data requires modification.

DIP: Design



DIP *hard*

Proportionally relabel watermarked data

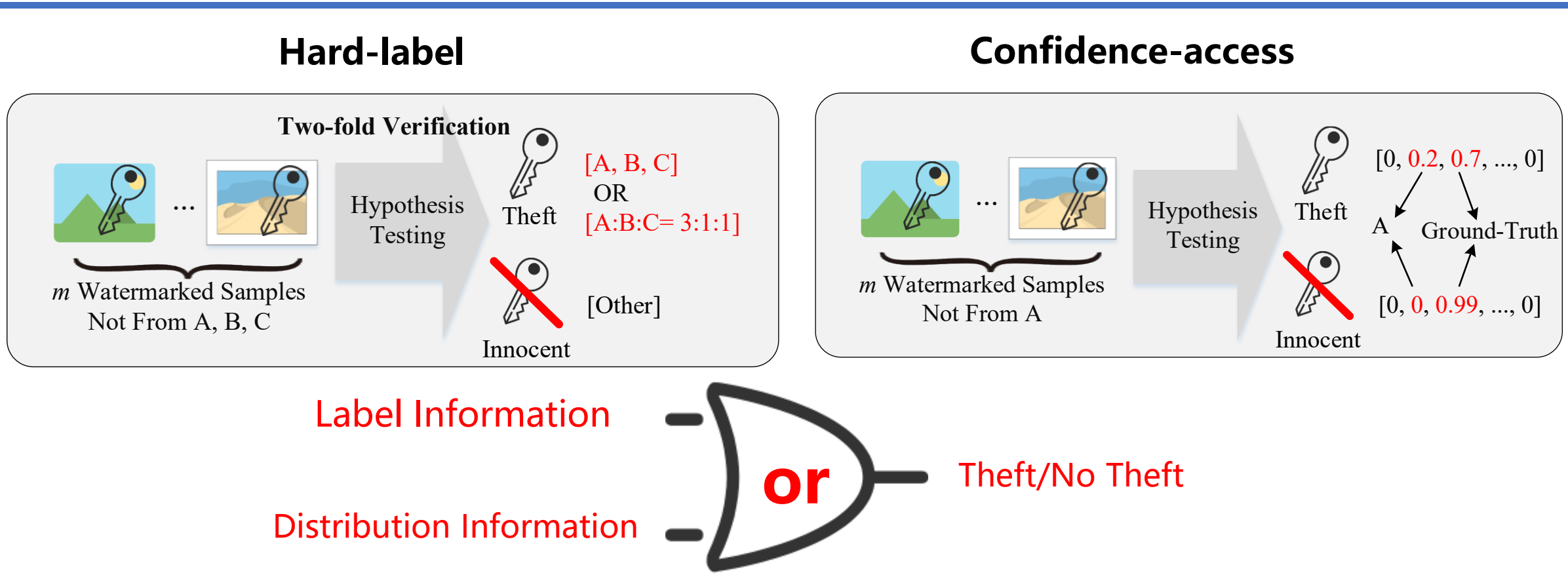
DIP *soft*

N watermark copies+
one relabeled copy

Watermark injection embeds a DIP watermark into the contributed dataset.

- It maps watermark samples to multiple target outputs in a probabilistic manner.

DIP: Design



Label Information

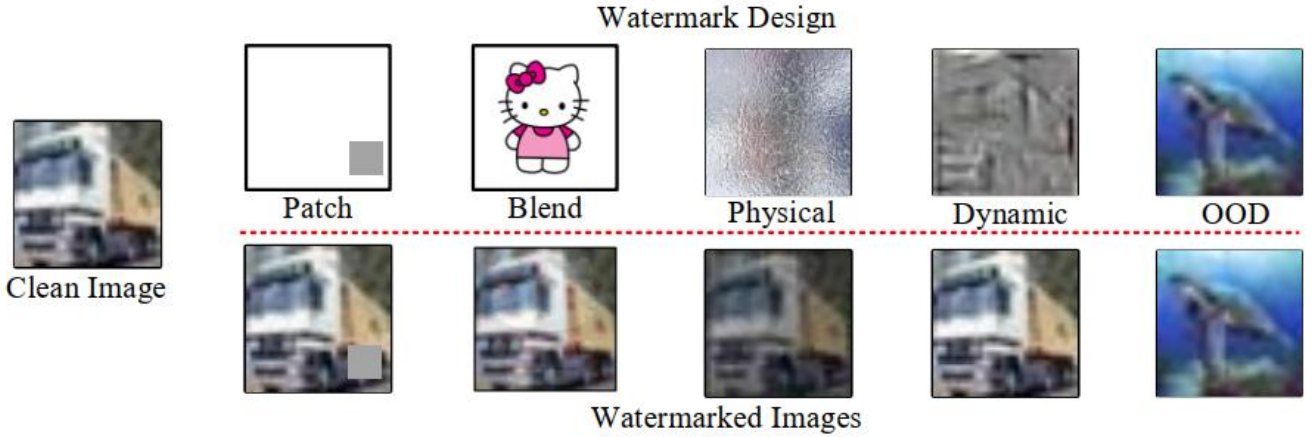
Distribution Information

or

Theft/No Theft

Two-fold verification leverages both label and distribution information to ensure reliable dataset verification under weak watermark signals.

Results: Low Injection Rates



Watermark Design

1% Watermarking Injection Rate

Watermarking ↓	Watermark Design →	w/o watermark	Patch (ΔTest Acc. = -0.27%)		Blend (ΔTest Acc. = -0.15%)		Physical (ΔTest Acc. = -0.1%)		Dynamic (ΔTest Acc. = -0.6%)		OOD (ΔTest Acc. = -0.55%)	
	Dataset ↓	Test Acc.	WSR ↑ / DS ↑	P-value ↓	WSR ↑ / DS ↑	P-value ↓	WSR ↑ / DS ↑	P-value ↓	WSR ↑ / DS ↑	P-value ↓	WSR ↑ / DS ↑	P-value ↓
DIP_{hard}	MNIST	99.2%	99.3% / 0.93	**0.0** / 10^{-3}	99.5% / 0.95	**0.0** / 10^{-4}	99.3% / 0.94	**0.0** / 10^{-3}	73.8% / 0.92	1.0 / **10^{-2}**	100% / 0.98	**0.0** / 10^{-5}
	CIFAR-10	88.4%	99.6% / 0.96	**0.0** / 10^{-4}	98.8% / 0.93	**0.0** / 10^{-3}	99.4% / 0.93	**0.0** / 10^{-3}	22.3% / 0.63	1.0 / **0.27**	100% / 0.99	**0.0** / 10^{-5}
	Tiny-ImageNet	67.5%	99.2% / 0.94	**0.0** / 10^{-3}	99.0% / 0.94	**0.0** / 10^{-3}	98.6% / 0.96	**0.0** / 10^{-4}	35.1% / 0.77	1.0 / **0.31**	100% / 0.98	**0.0** / 10^{-5}
DIP_{soft}	MNIST	99.2%	99.4% / -	10^{-5}	99.8% / -	10^{-4}	99.7% / -	10^{-10}	90.8% / -	10^{-8}	98.2% / -	10^{-6}
	CIFAR-10	88.4%	98.6% / -	10^{-8}	96.1% / -	10^{-8}	95.4% / -	10^{-17}	80.3% / -	10^{-14}	96.7% / -	10^{-8}
	Tiny-ImageNet	67.5%	99.1% / -	10^{-11}	95.5% / -	10^{-13}	95.2% / -	10^{-23}	76.1% / -	10^{-13}	97.5% / -	10^{-13}

The verification performance of DIP watermarks is evaluated across different watermark designs (Satisfying RM 1&3)

Results: Adversarial Environments (Satisfying RM2)

Comparison between DIP and existing dataset watermarking approaches under three data cleansing attacks: SCAn (Usenix21), Beatrix (NDSS22), ASSET (Usenix23)

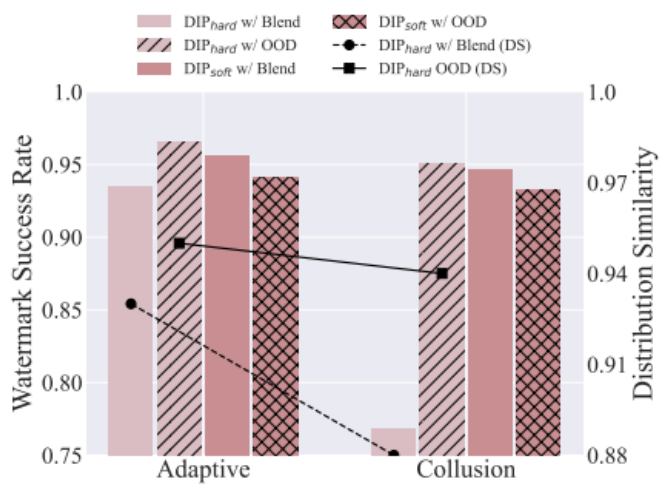
Data Cleansing →	SCAn [48]		Beatrix [50]		ASSET [30]
	J* ↓	DSR ↓	R_t^* ↓	DSR ↓	AUC-ROC ↓
DVBW	8.9	58.0%	1.9	6.0%	0.87
UBW	0.9	0.0%	1.0	0.0%	0.56
DW	1.2	0.0%	1.2	0.0%	0.31
CBW	6.1	39.0%	1.7	4.0%	0.76
DT	1.2	0.0%	1.0	0.0%	0.35
$\texttt{DIP}_{\text{hard}}$	2.5	4.0%	1.3	0.0%	0.59
$\texttt{DIP}_{\text{soft}}$	1.4	0.0%	1.2	0.0%	0.38

Most Robust Case

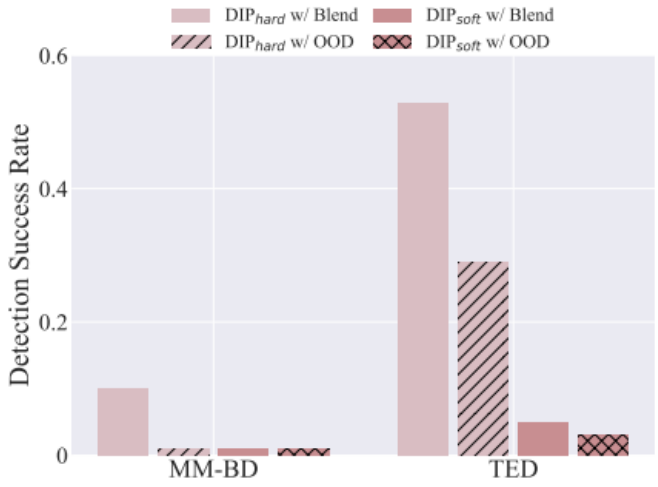
The malicious data curator may cleanse the received dataset.

Results: Adversarial Environments (Satisfying RM2)

The evaluation of DIP under three collusion-based attacks: adaptive attack, ASSET+CBD, MM-BD (Oakland24), TED (Oakland24).



(a) Adaptive / Collusion



(b) SOTA Backdoor Defense

- Strong attacks are effective.
- OOD-based DIP is more robust.

The data curator and the model provider may collude to remove the dataset watermark.

Conclusion and Takeaway

RM1: Low Watermarking Injection Rate (low to 0.4%)

RM2: Resilience to adversarial settings (data augmentation, data cleansing, robust training and collusion-based attacks)

RM3: Non-harmful Utility

RM4: No False Positive Rate

Property1: Modality Agnostic (image, text)

Property2: Task Agnostic (classification, generation, regression)

Thank you for your listening!