Verify your Labels!

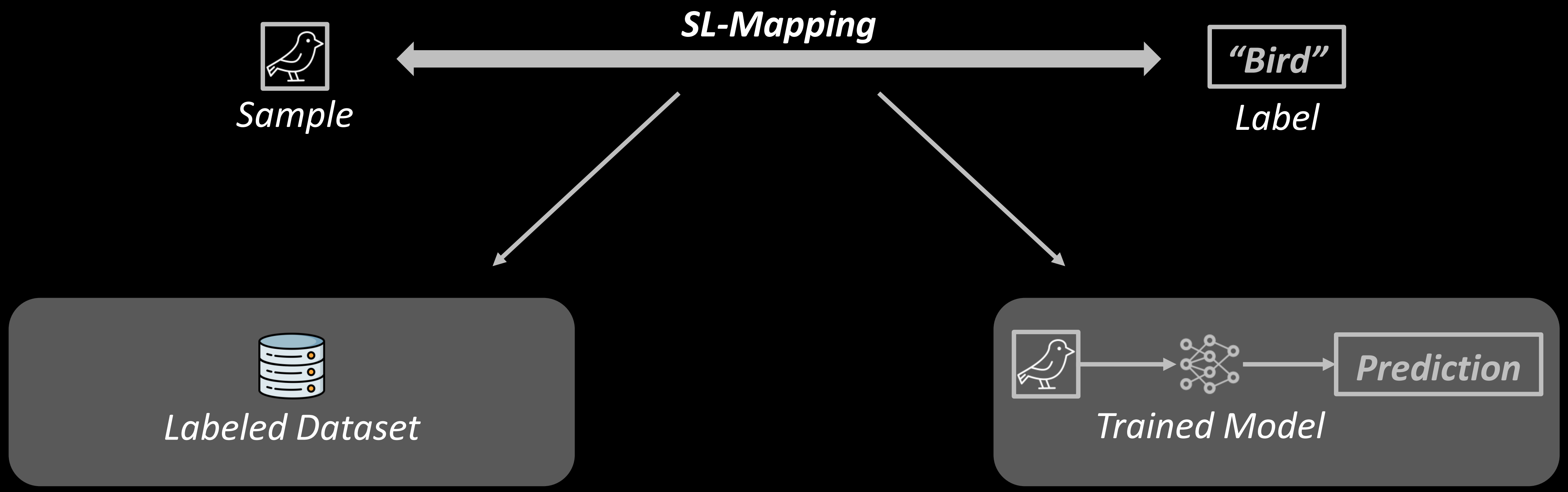
Trustworthy Predictions and Datasets via Confidence Scoring

Torsten Krauß, Jasper Stang, and Alexandra Dmitrienko

University of Würzburg, Germany

33rd USENIX Security Symposium

Problem

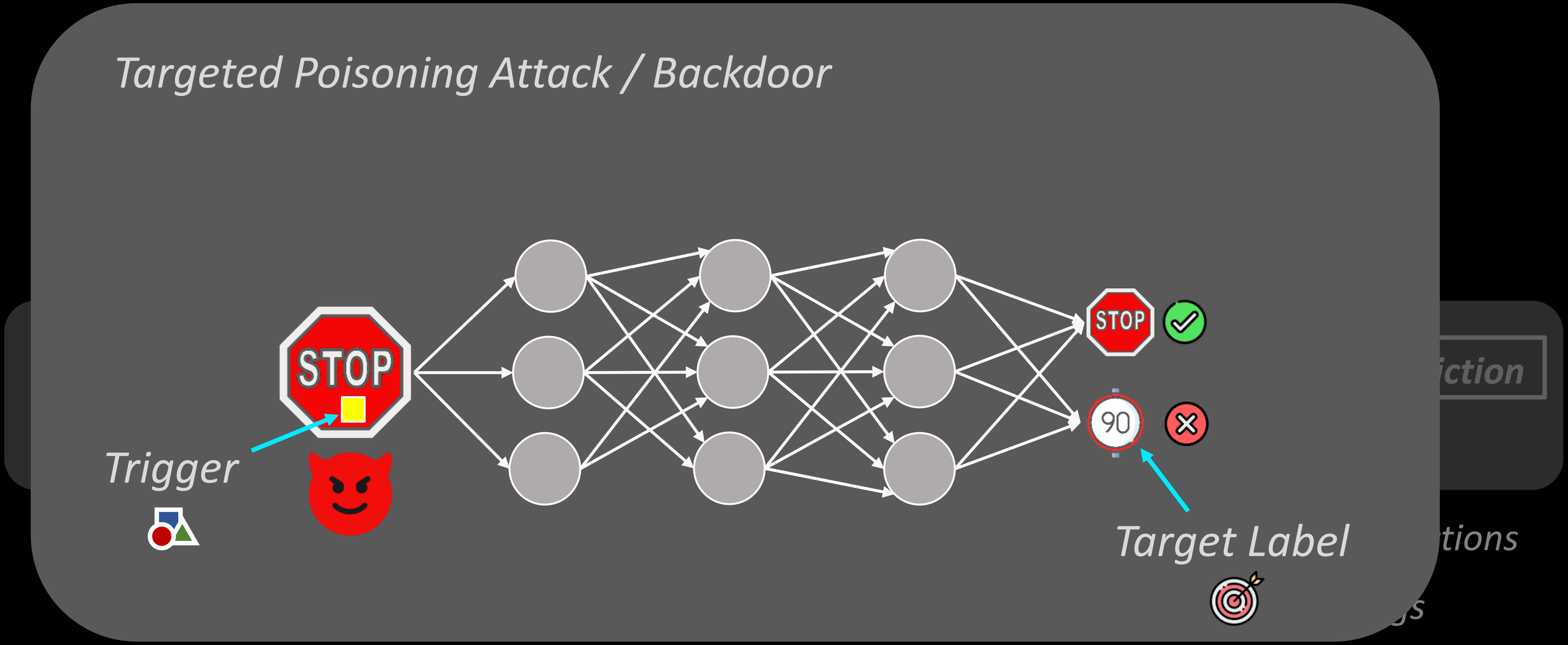


Natively consists of SL-Mappings

Unseen samples & predictions form SL-Mappings

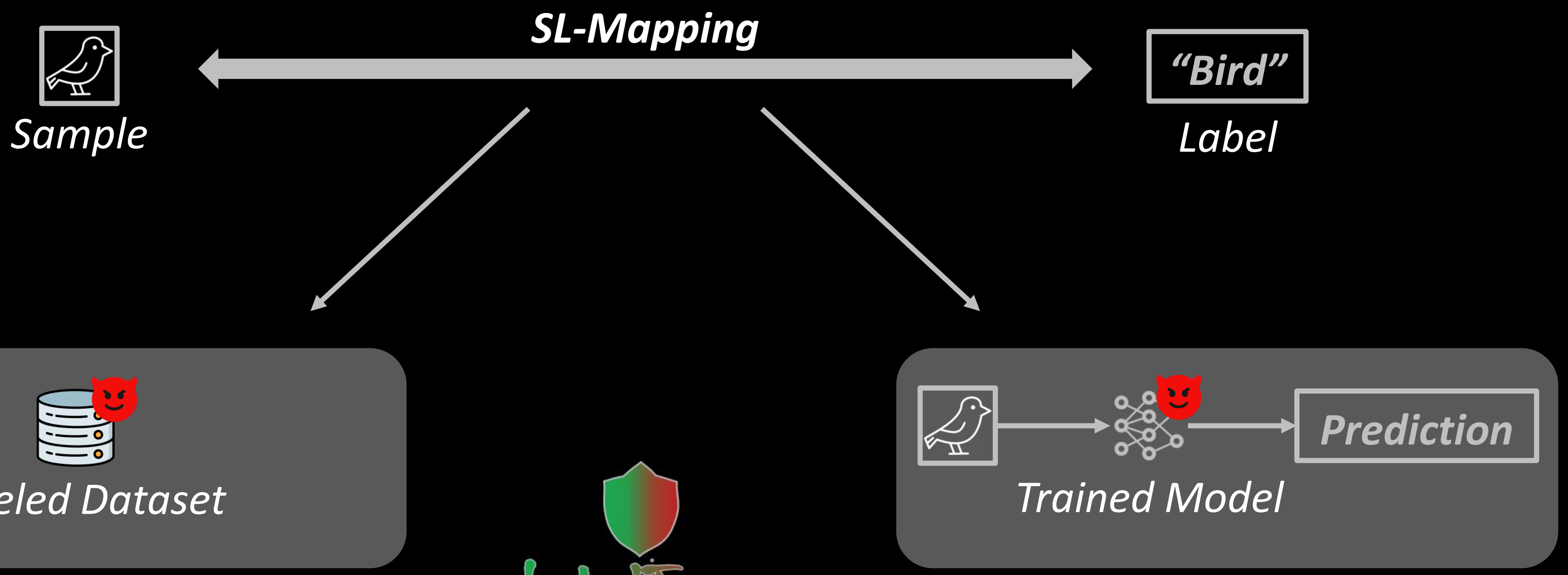
Are the SL-Mappings reliable?

Problem



Are the SL-Mappings reliable?

Problem



Natively consists of SL-Mappings

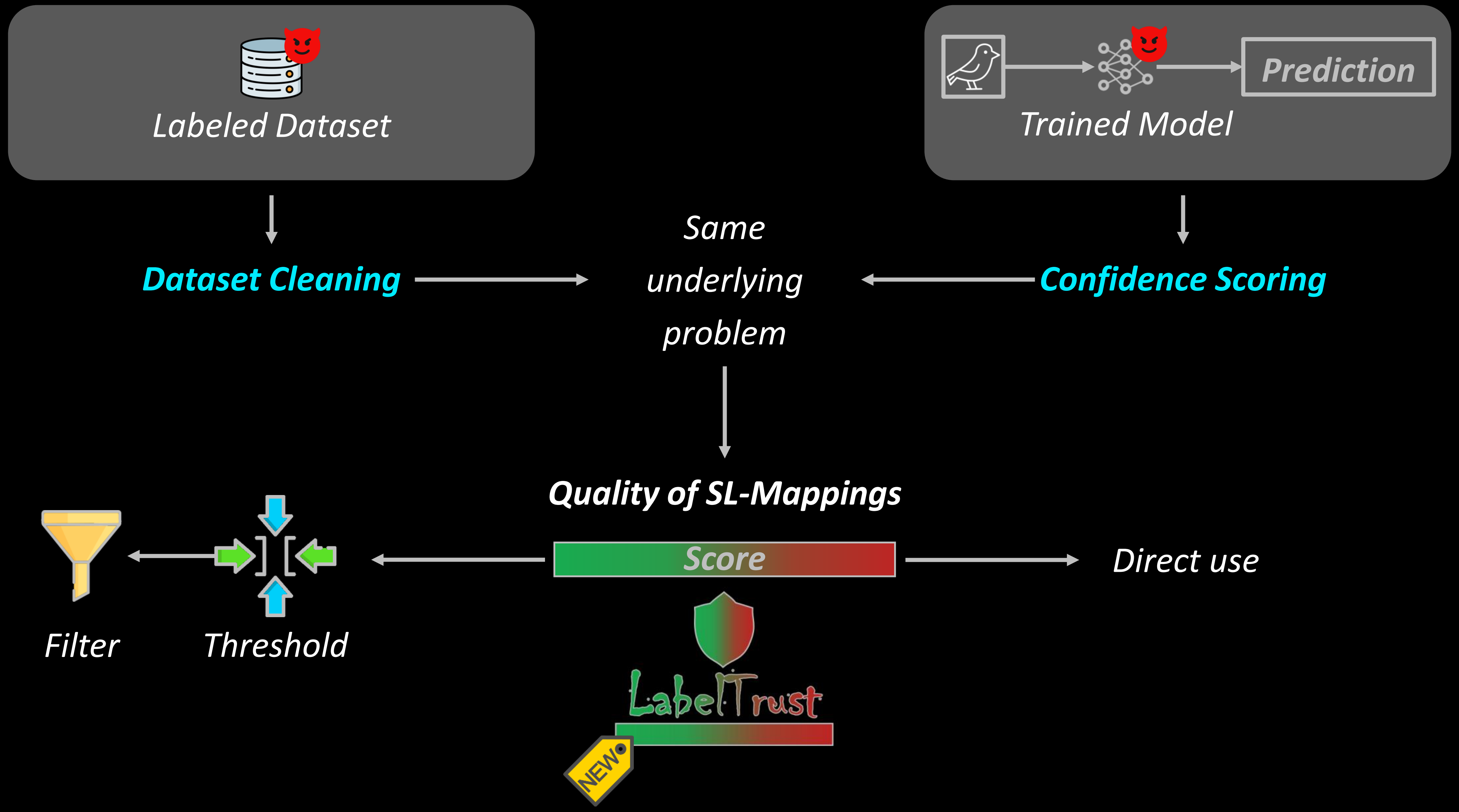
LabelTrust

NEW

Unseen samples & predictions form SL-Mappings

Are the SL-Mappings reliable?

Two Use-Cases



Downsides of Existing Works

Dataset Cleaning

- *Train auxiliary models on entire dataset [1, 2, 3]*
- *Specific to a single model [4]*
- *Dependent on large clean datasets [2, 4]*

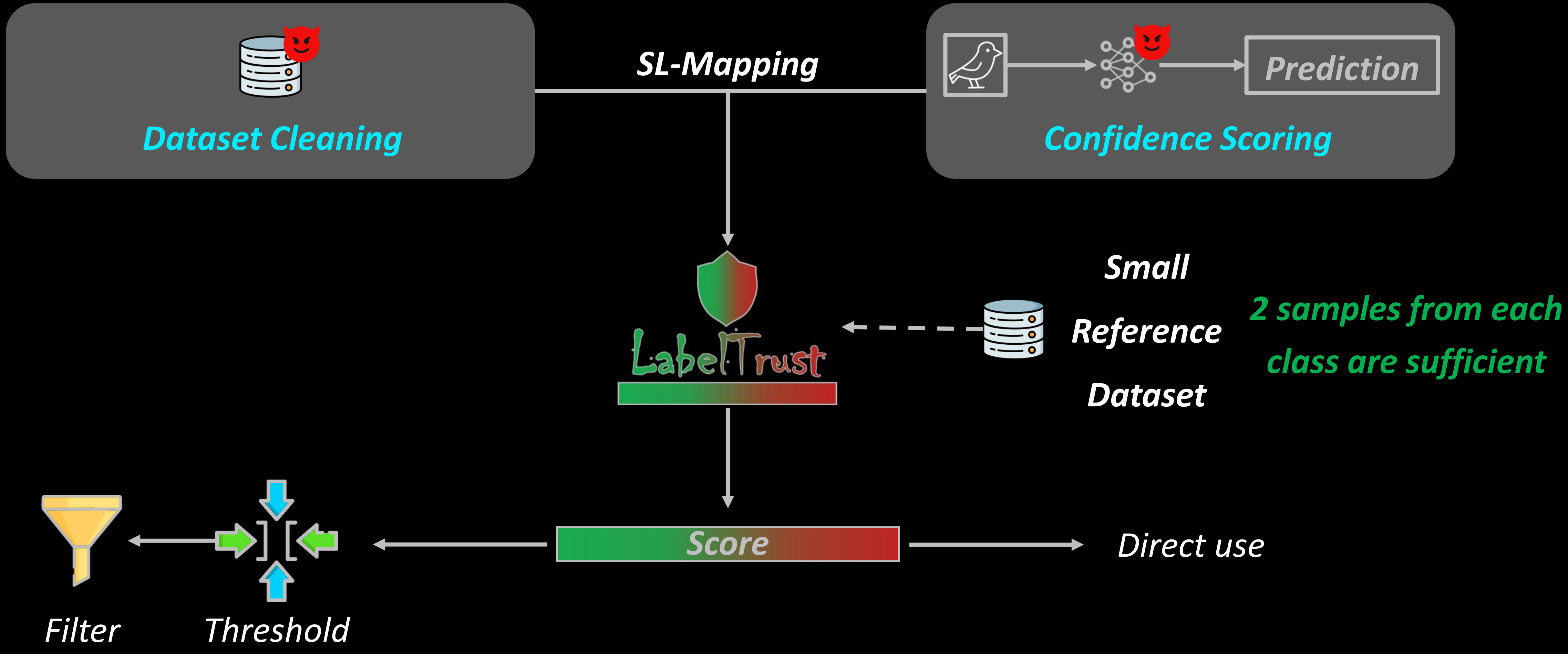
Confidence Scoring

- *Depend on entire untrusted datasets [5, 6, 7]*
- *Specific to a single model [5]*
- *Missing consideration of poisoning attacks [5, 6, 7]*

Same underlying problem but no unique solution!

[1] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor Defense via Adaptively Splitting Poisoned Dataset. In IEEE/CVF, 2023.
[2] Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label Sanitization Against Label Flipping Poisoning Attacks. ECML PKDD 2018 Workshops, 2019.
[3] Fereshteh Razmi and Li Xiong. Classification Auto-Encoder Based Detector Against Diverse Data Poisoning Attacks. In IFIP DBSec, 2023.
[4] Huayang Huang, Qian Wang, Xueluan Gong, and Tao Wang. Orion: Online Backdoor Sample Detection via Evolution Deviance. IJCAI, 2023.
[5] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing Failure Prediction by Learning Model Confidence. NeurIPS, 2019.
[6] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To Trust Or Not To Trust A Classifier. NeurIPS, 2018.
[7] Yan Luo, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. Learning to Predict Trustworthiness with Steep Slope Loss. NeurIPS, 2021.

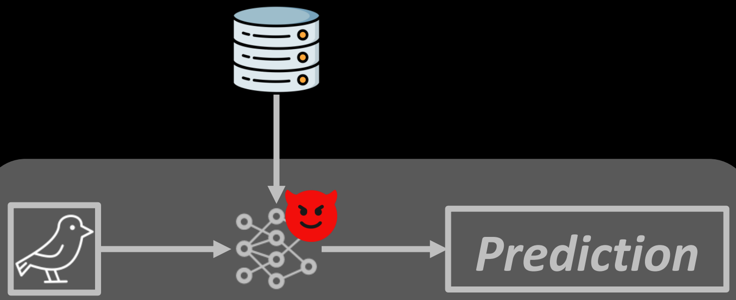
LabelTrust – Principle



LabelTrust – Reference Dataset

Dataset Cleaning

Small Reference Dataset



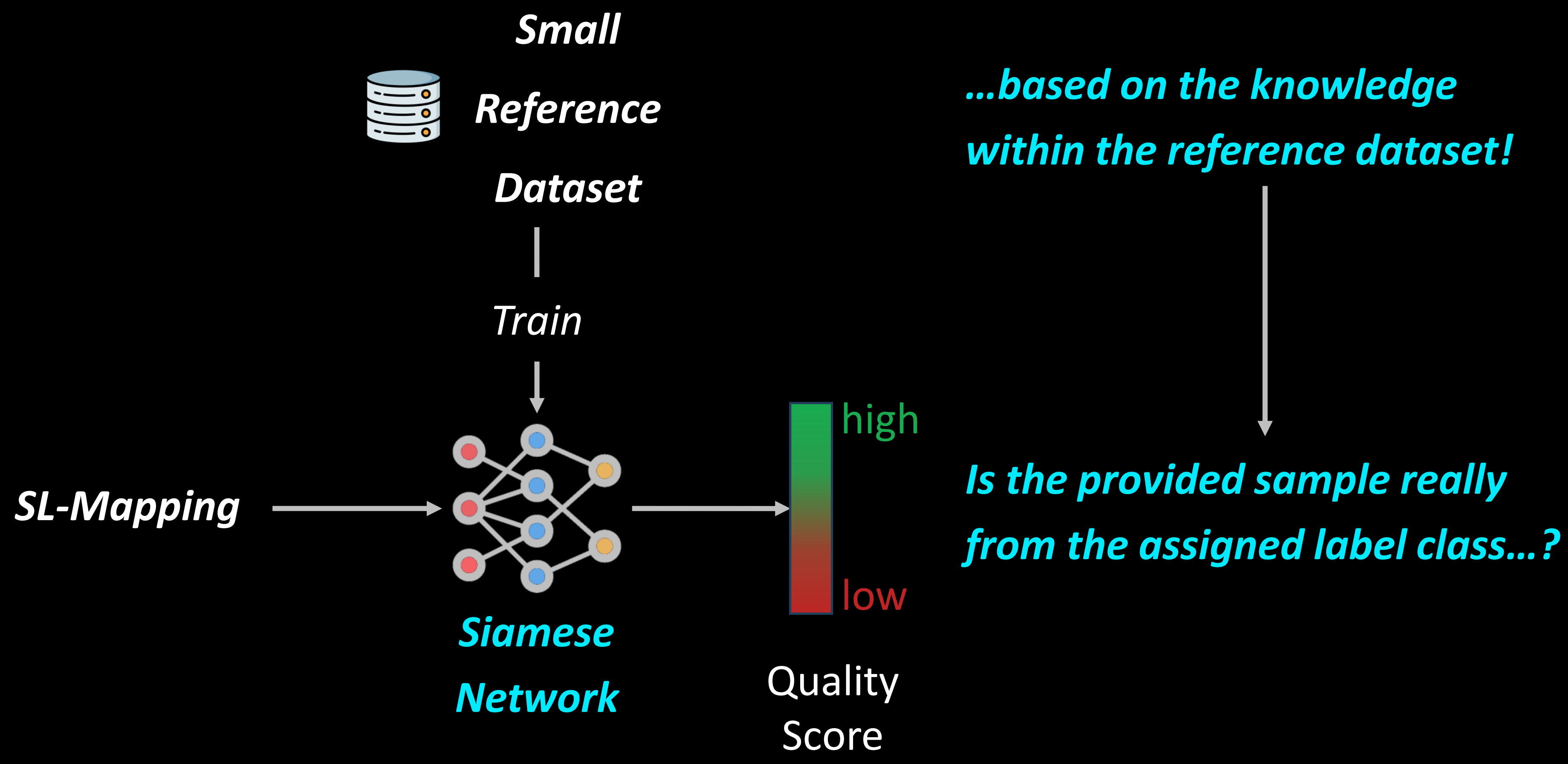
Confidence Scoring

- *Sampling from* 

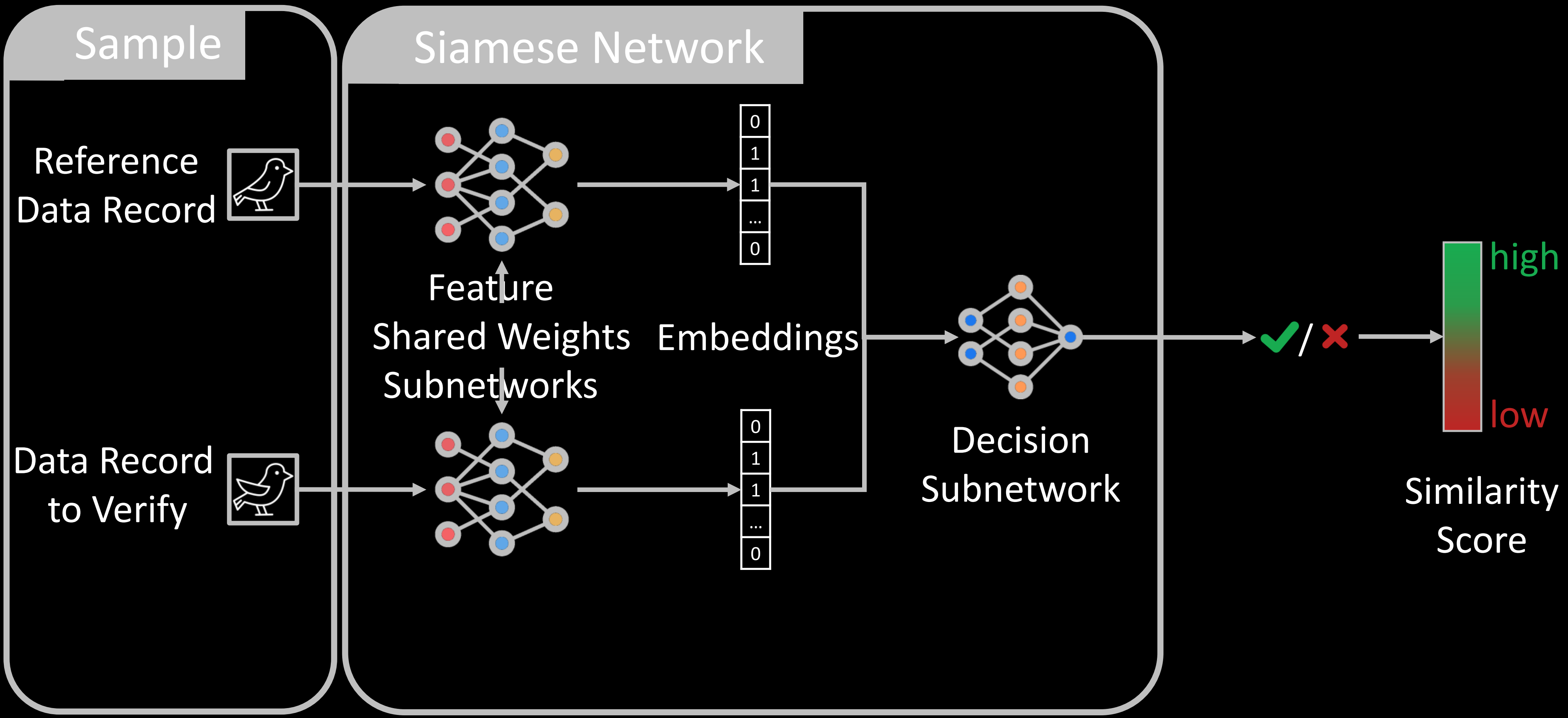
Trusted Domain Expert

- *Sampling from the training* 
- *LabelTrust provided by model creator*
- *Small reference dataset provided by model creator*
- *Observation of inference input and output*

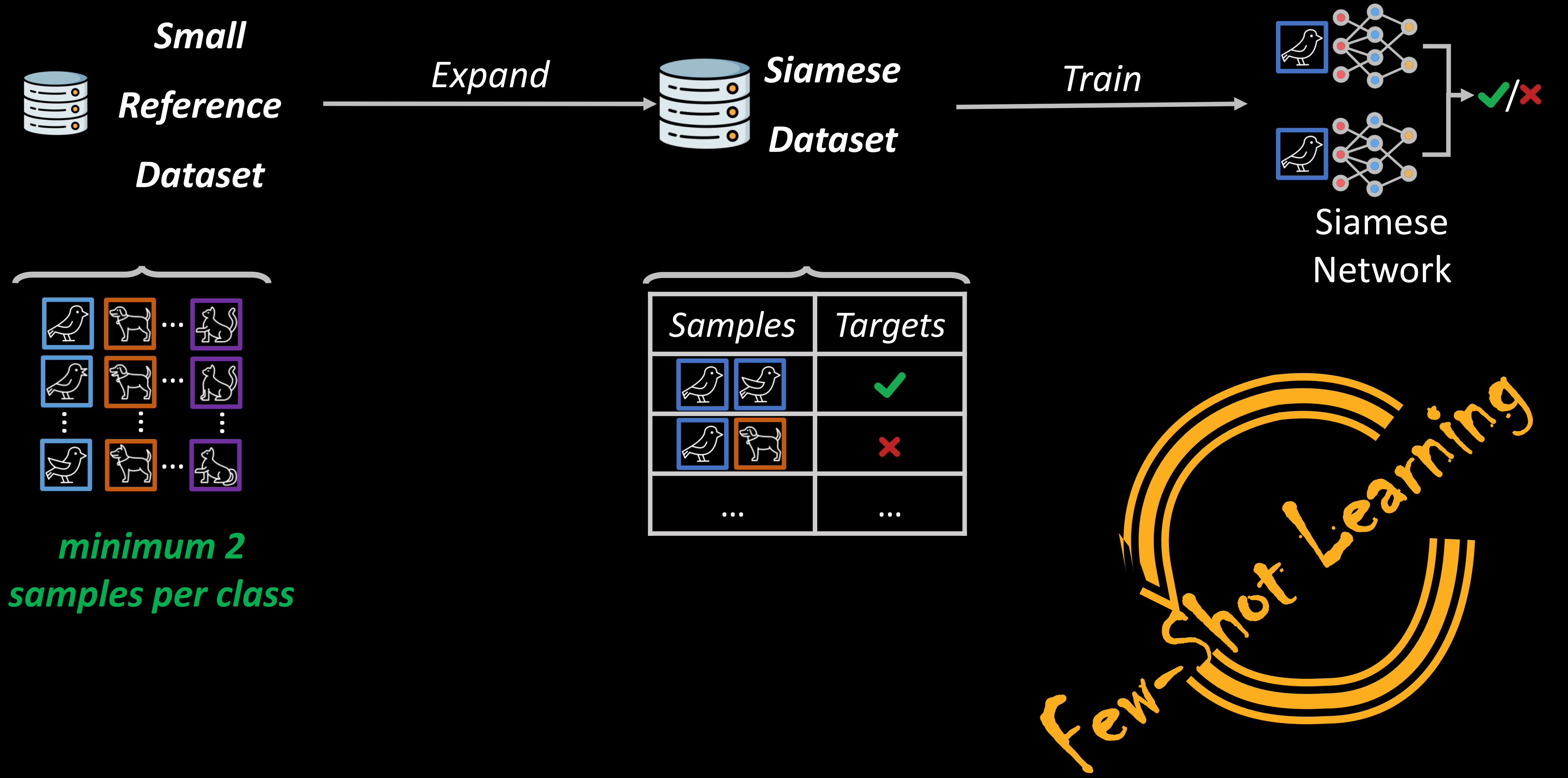
LabelTrust – Reference Dataset



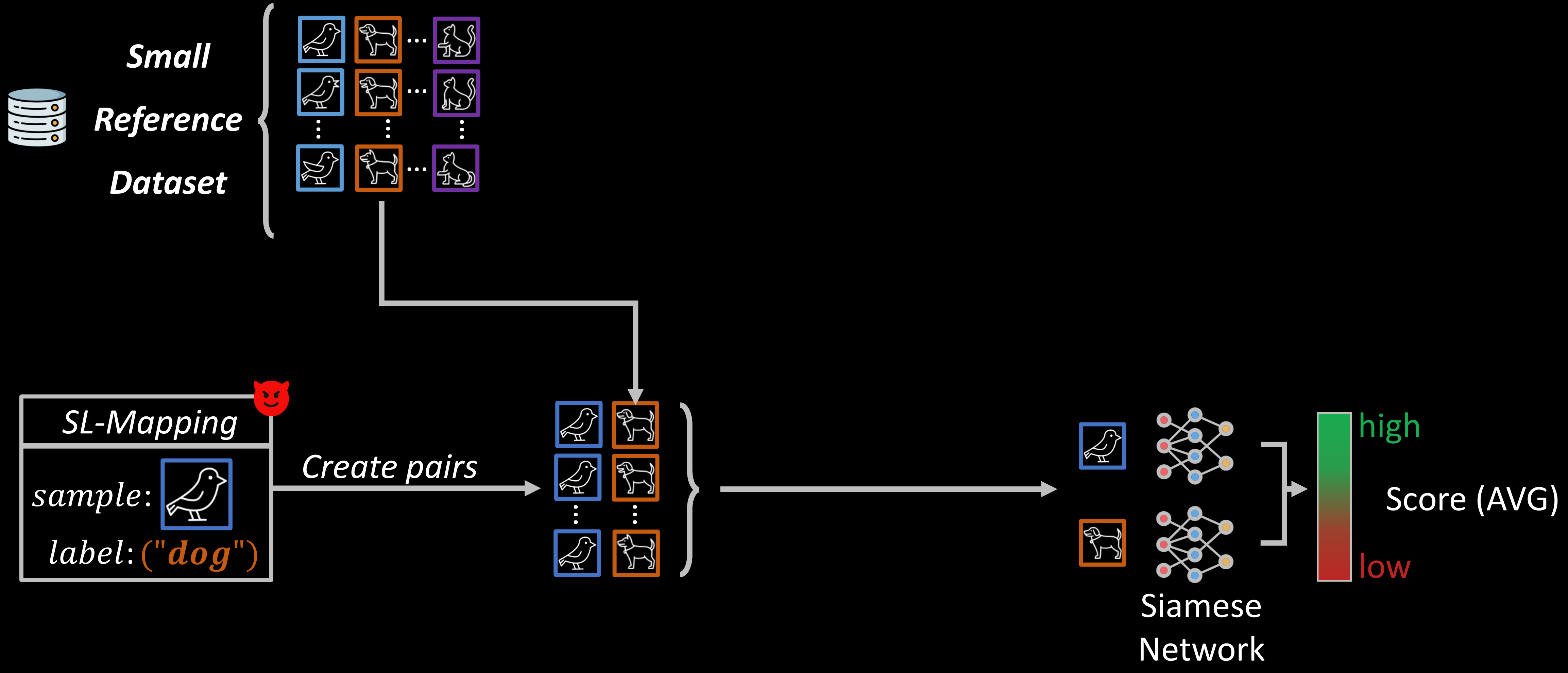
Siamese Network



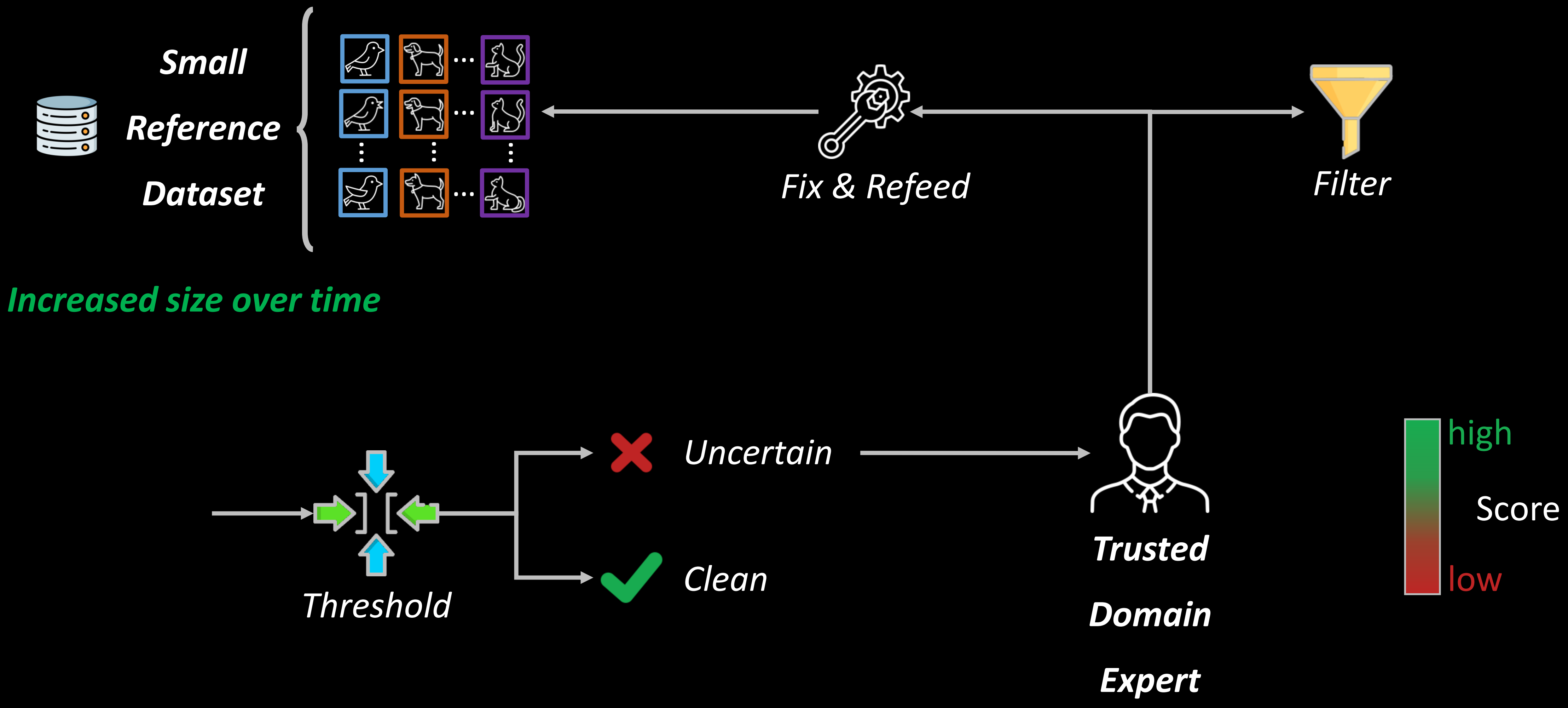
LabelTrust – Training



LabelTrust – Inference



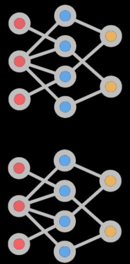
LabelTrust – Refeed Loop



LabelTrust – Refeed Loop

Small Reference Dataset

Re-Train

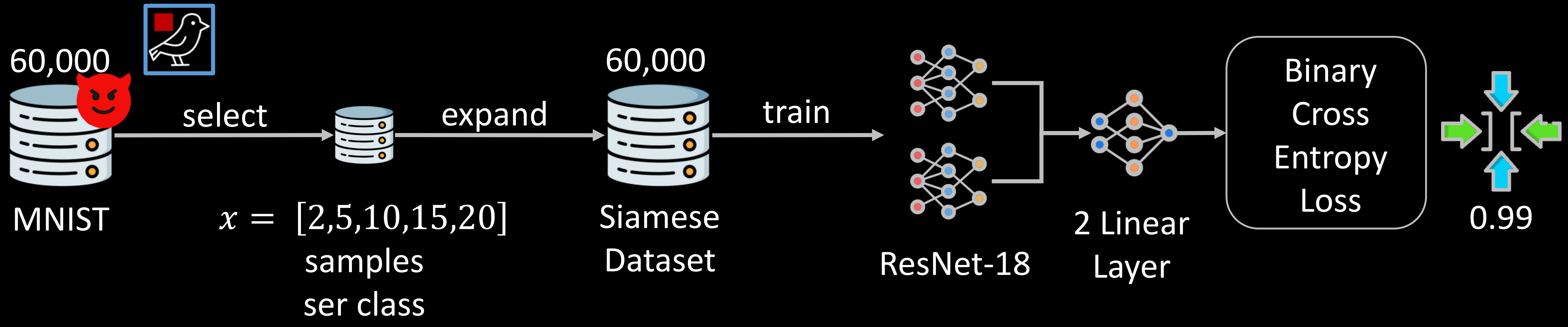


Siamese Network

Increased size over time

Increased performance over time

LabelTrust - Evaluation

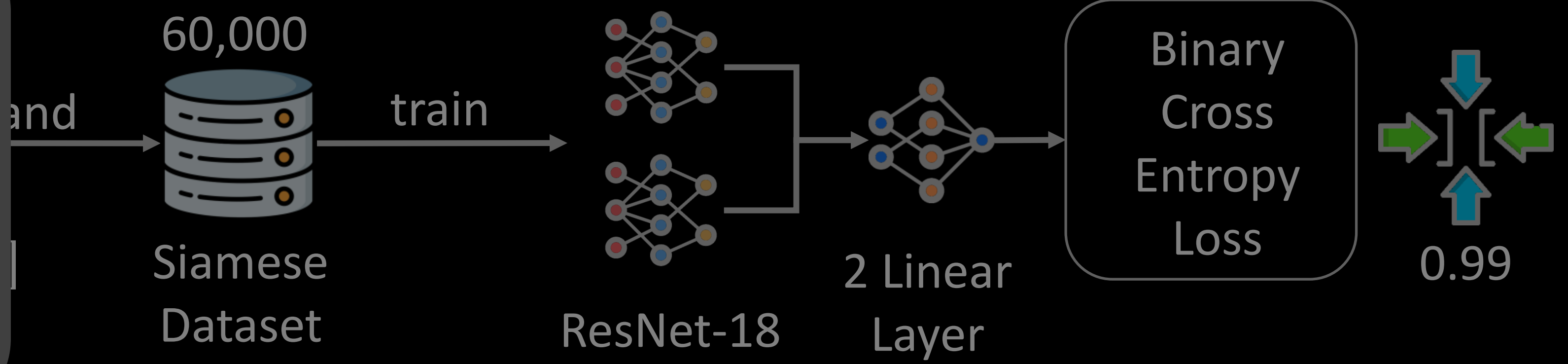


LabelTrust - Evaluation

MNIST Testset

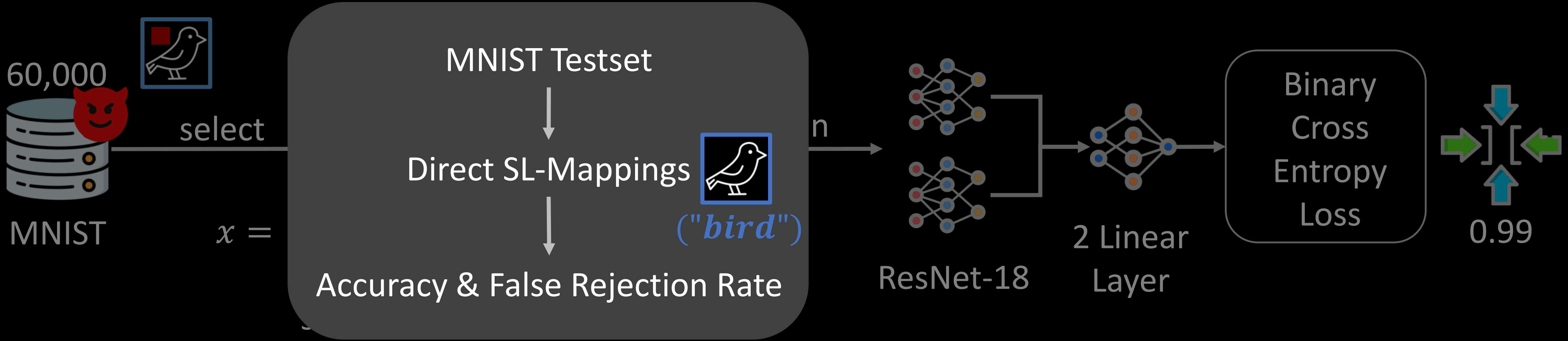
Siamese Dataset

Accuracy



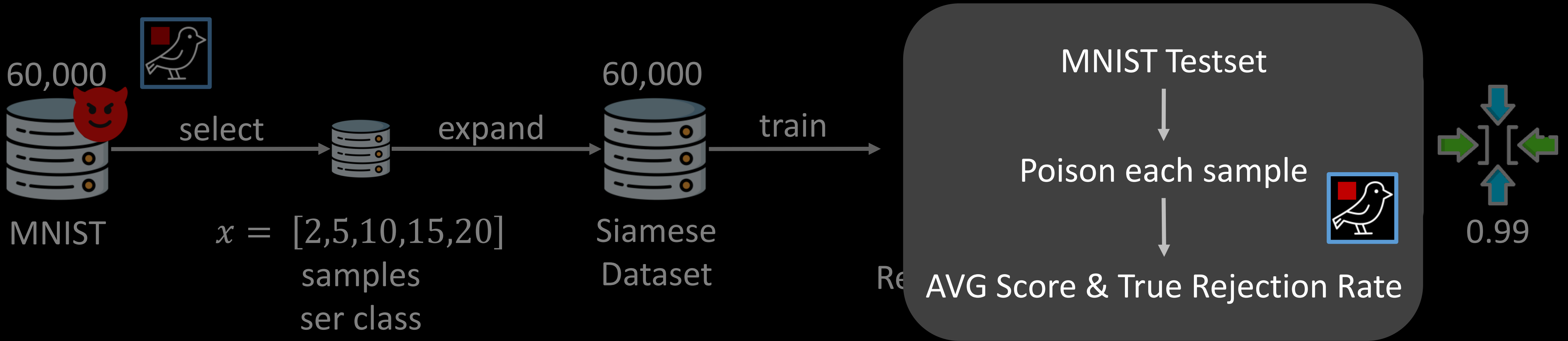
x	MNIST Testset Siamese Accuracy				
2	59.56				
5	72.44				
10	75.50				
15	80.88				
20	81.74				

LabelTrust - Evaluation



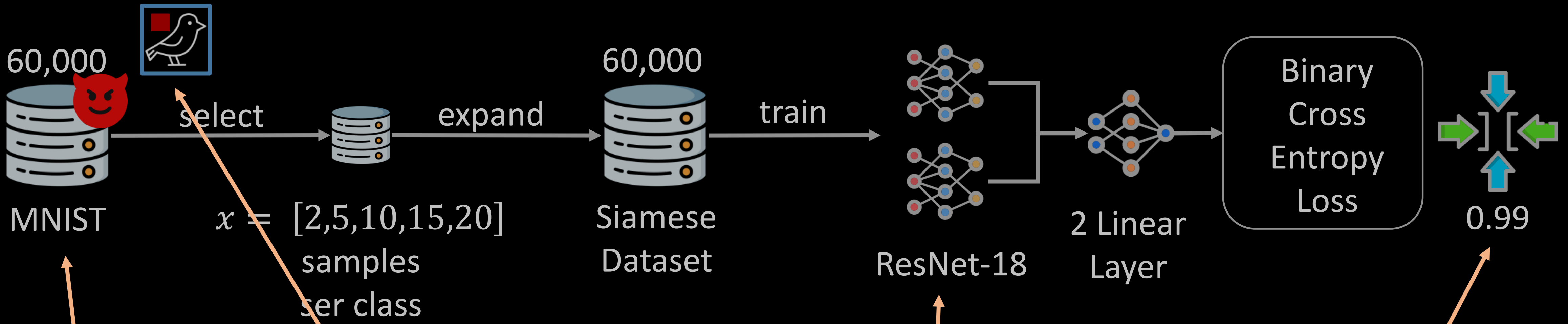
x	MNIST Testset Siamese Accuracy	MNIST Testset SL-Mapping Verification			
		Accuracy	False Rejection Rate		
2	59.56	92.52	72.98		
5	72.44	95.30	41.03		
10	75.50	96.29	32.07		
15	80.88	97.46	22.80		
20	81.74	97.36	22.94		

LabelTrust - Evaluation



x	MNIST Testset Siamese Accuracy	MNIST Testset SL-Mapping Verification		Poisoned MNIST Testset SL-Mapping Verification	
		Accuracy	False Rejection Rate	AVG Score	True Rejection Rate
2	59.56	92.52	72.98	0.0088	100.00
5	72.44	95.30	41.03	0.0074	99.75
10	75.50	96.29	32.07	0.0042	99.93
15	80.88	97.46	22.80	0.0047	99.87
20	81.74	97.36	22.94	0.0010	99.95

LabelTrust - Evaluation



Datasets
- MNIST
- F-MNIST
- IIC (colored)

Poisonings
- Pixel Trigger
- Blend / Noise
- Clean Label
- Random Label

Models
- ResNet-18
- Small CNN

Thresholds
- 0.99
- 0.50
- 0.01

LabelTrust - Evaluation

Dataset Cleaning

Confidence Scoring

x	Clean	Filtered	TRR	ACC
10	37,717	22,283	99.87	72.20
15	41,199	18,801	99.69	77.98
20	47,202	12,798	99.34	87.92
25	47,038	12,962	99.39	87.65
30	47,027	12.973	99.71	87.70
35	50,009	9,981	99.82	92.70

- *350 reviewed samples after 5 refeed loops*
  - *0.0058 % of the dataset*
- *16.63 % filtered*
  - *99.70% of poisonings*
  - *Only 4,366 samples falsely filtered*
- *Backdoor removed in the first iteration*

LabelTrust - Evaluation

Dataset Cleaning

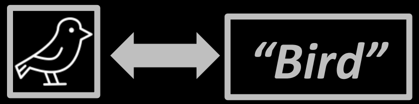
Confidence Scoring

- *False SL-mappings reliably yield very low scores*
- *Poisoning can be clearly identified*
- *High thresholds of 0.99 would barely yield errors*

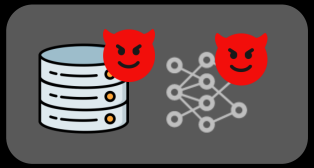
$$x = 10$$

Mispredictions from…	Confidence Score	
	Mean	Median
…benign testset	0.30	0.0018
…poisoned testset	0.0052	$5.83 \cdot 10^{-7}$

Conclusion



SL-Mappings are central in machine learning



Two use-cases: dataset cleaning & confidence scoring

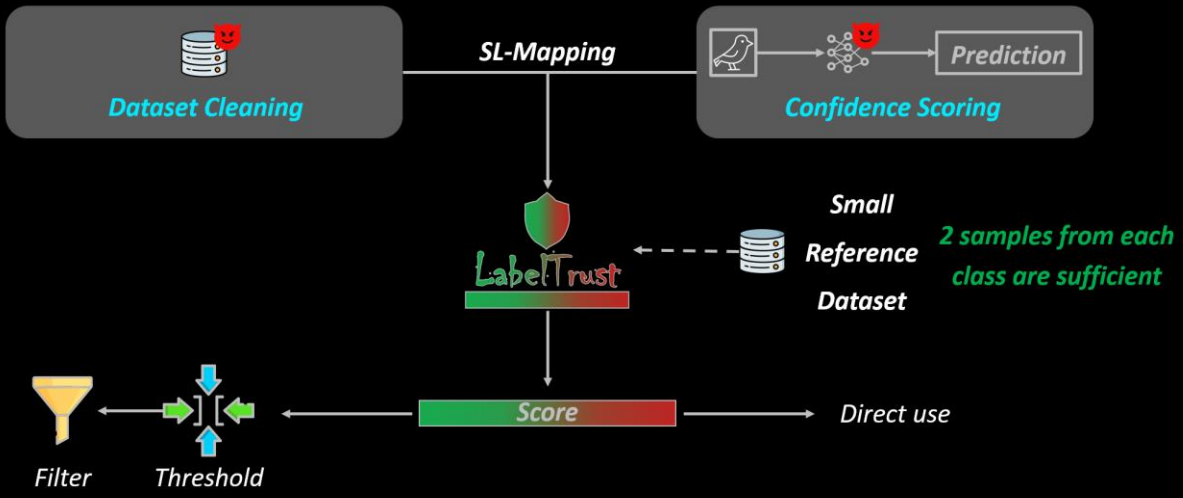
- No dual-use tool
- Dependency on large (clean) datasets
- Dependent on a specific model architecture
- Missing consideration of poisonings

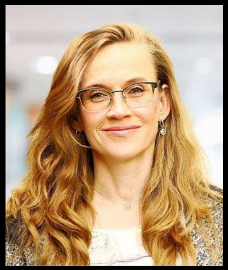
LabelTrust

- **SL-Mapping score based on reference data**
- Consolidation of **two use-cases**
- Minimal clean dataset due to **few-shot learning**
- Ongoing enhancement via **refeed loop**

Thank you!!11!!1

Any Questions?



<u>Torsten Krauß</u>, Jasper Stang, Alexandra Dmitrienko

University of **Würzburg**