Learning from Positive and Unlabeled Multi-Instance Bags in Anomaly Detection

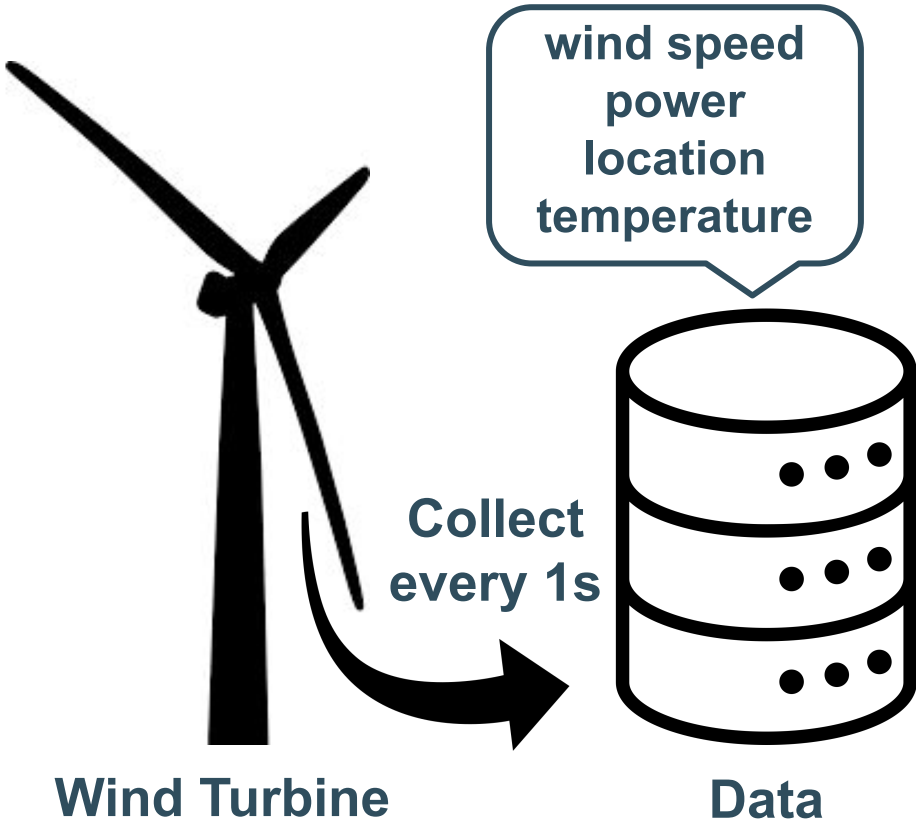
Lorenzo Perini, *Vincent Vercruyssen, Jesse Davis*

name.surname@kuleuven.be

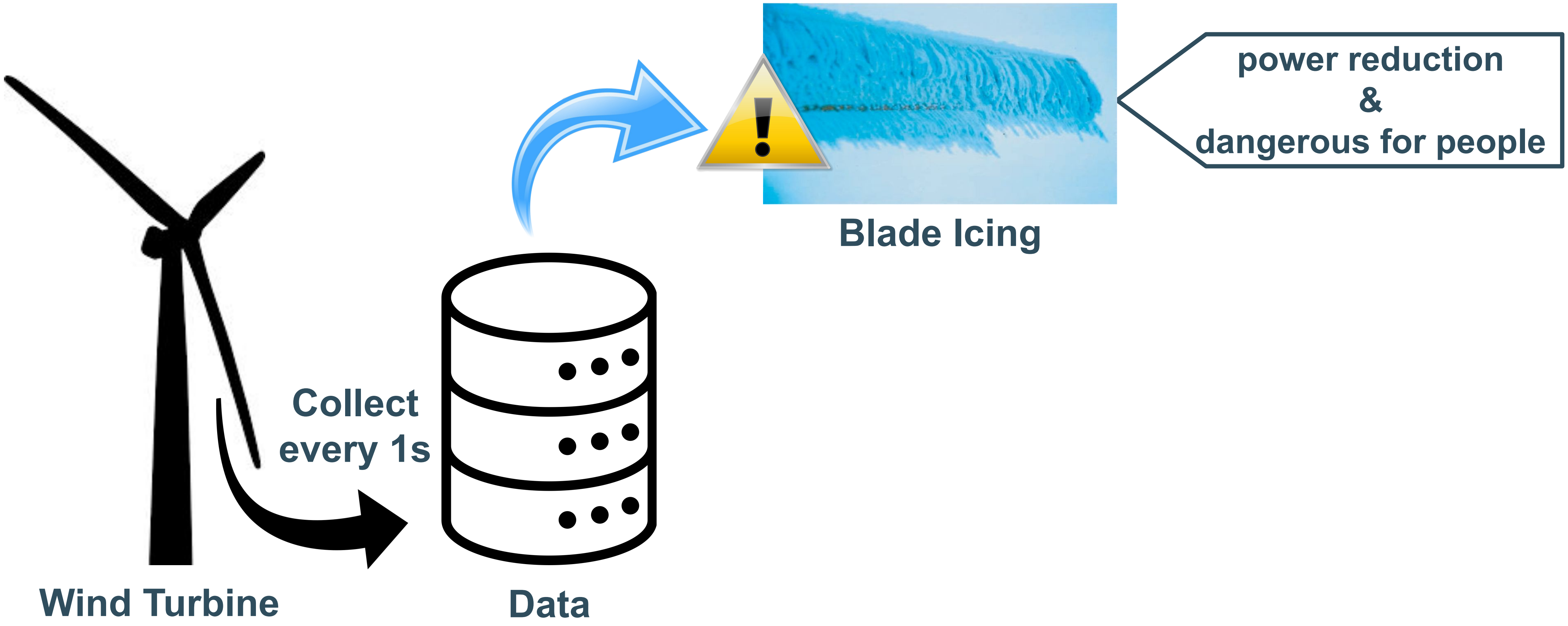
https://people.cs.kuleuven.be/~lorenzo.perini/

@LorenzoPerini95

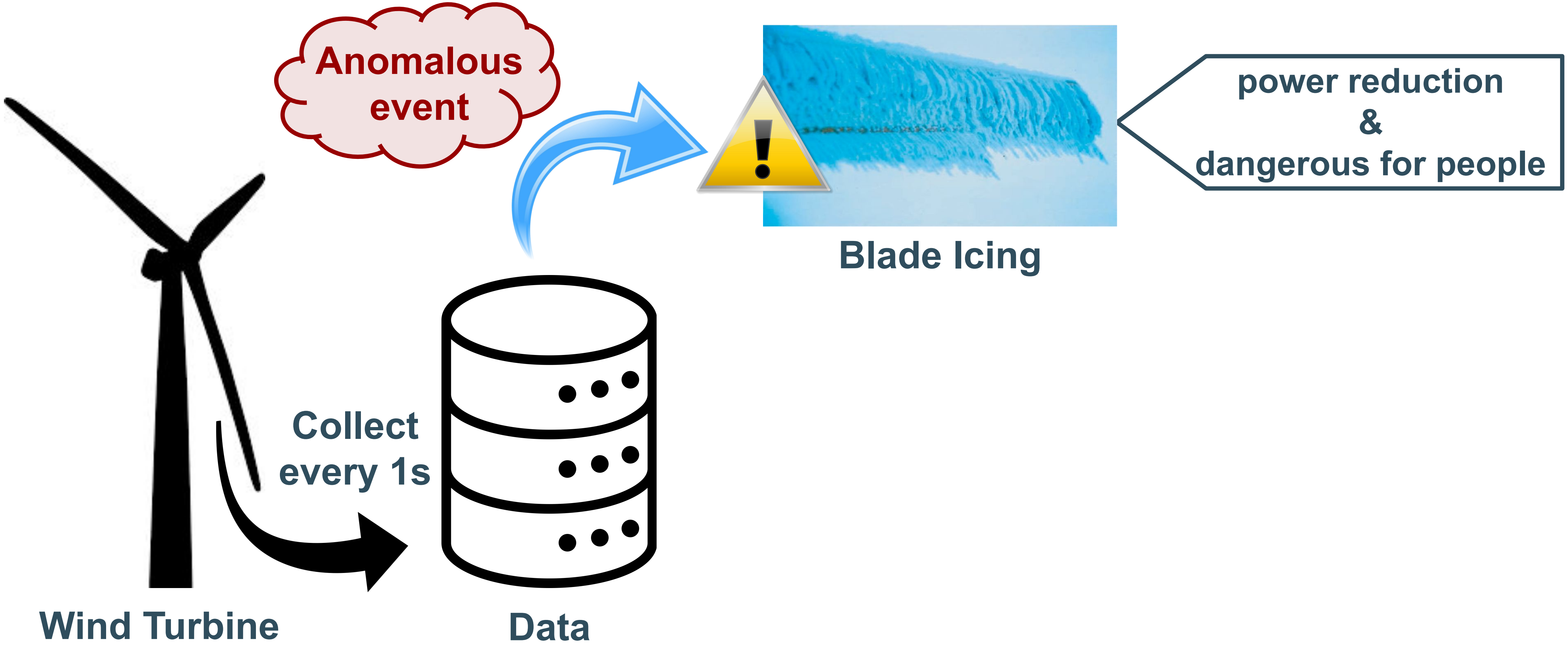
Anomaly Detection is the Task of Detecting the Instances that Deviate from a Normal Behaviour



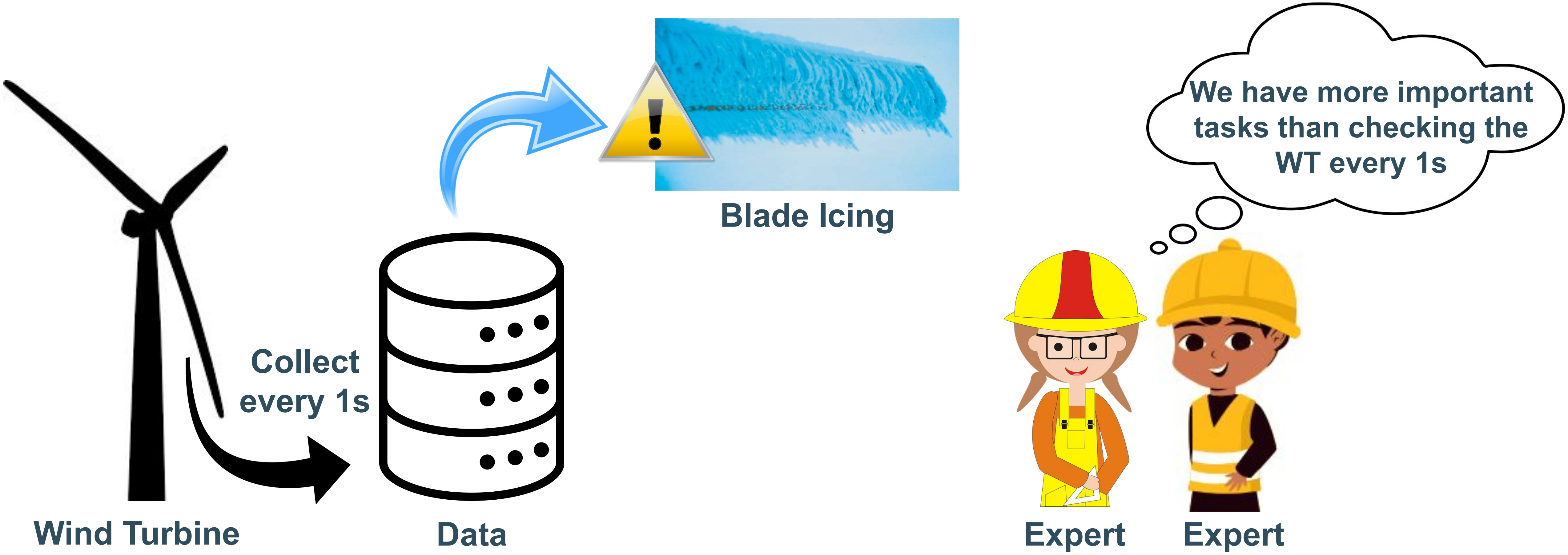
Anomaly Detection is the Task of Detecting the Instances that Deviate from a Normal Behaviour



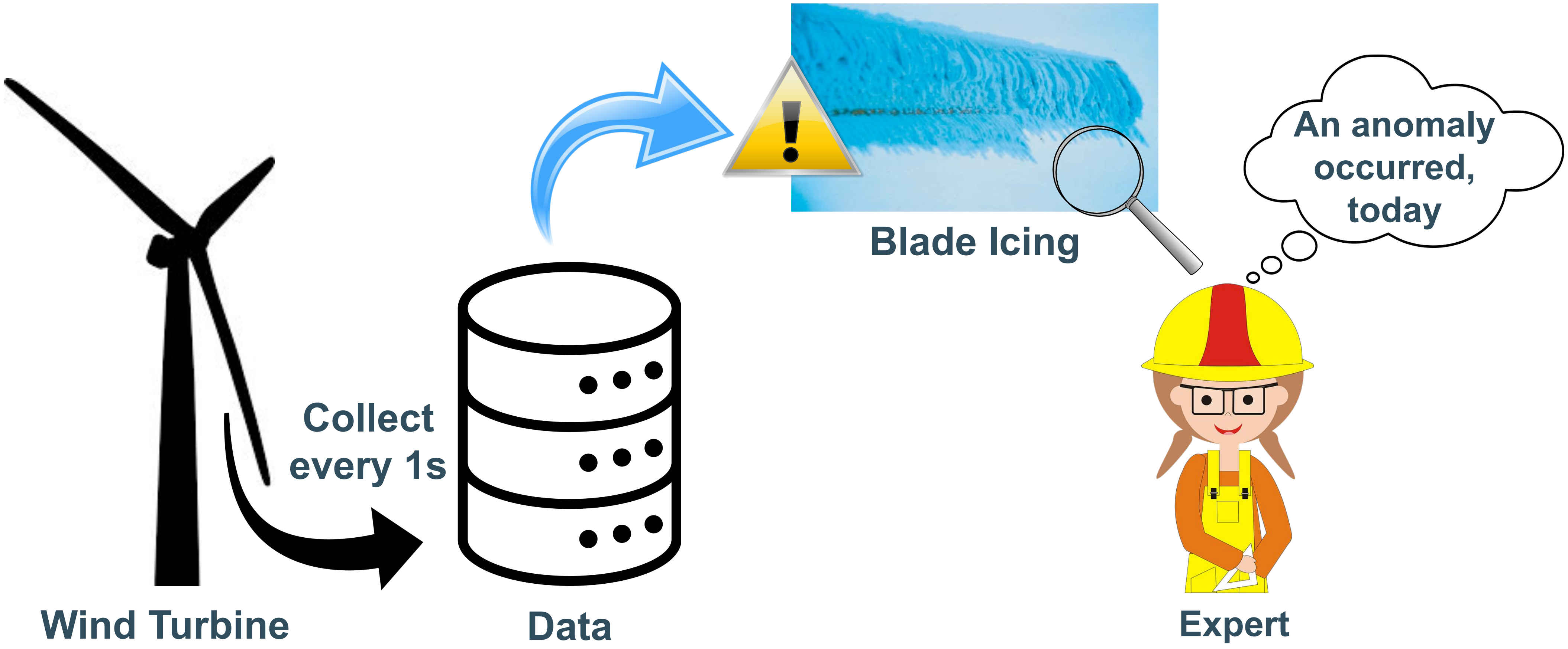
Anomaly Detection is the Task of Detecting the Instances that Deviate from a Normal Behaviour



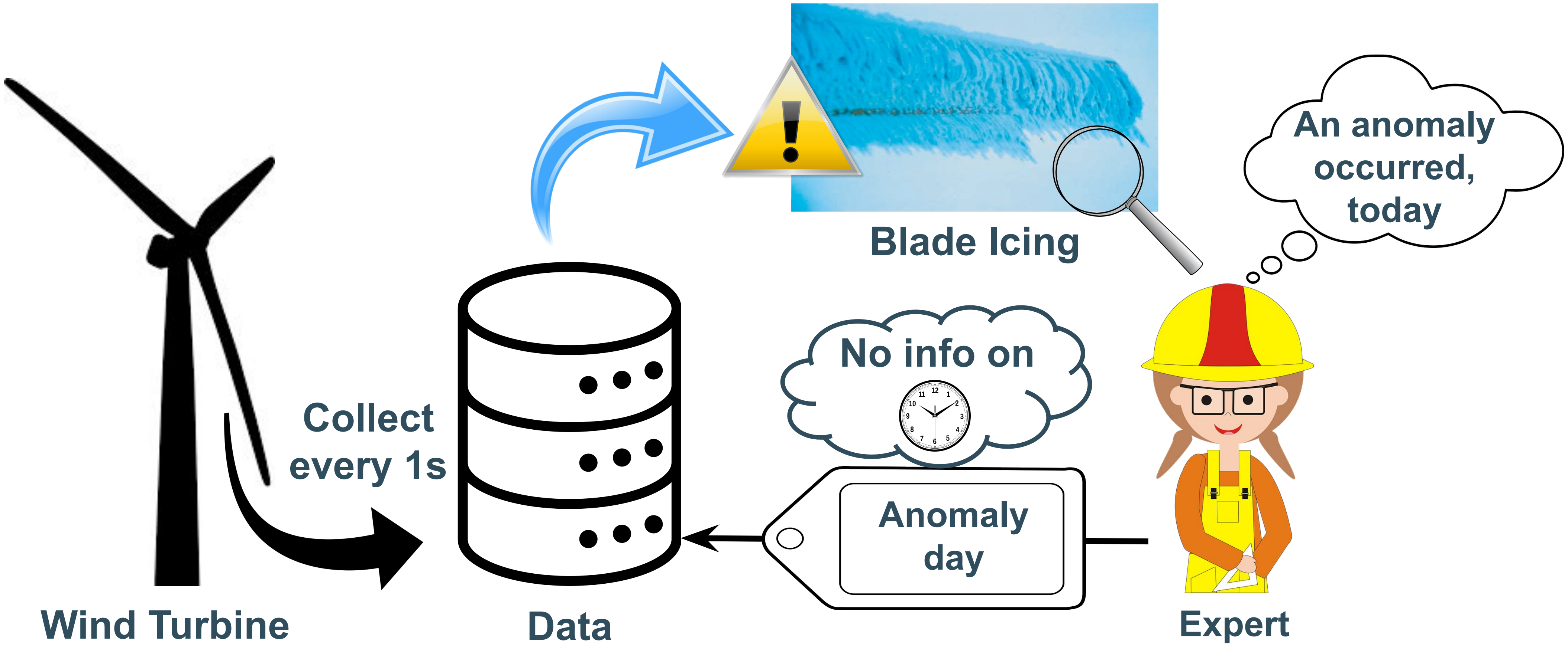
However, Acquiring Labels Is Hard in Anomaly Detection Because Anomalies Are Rare Events



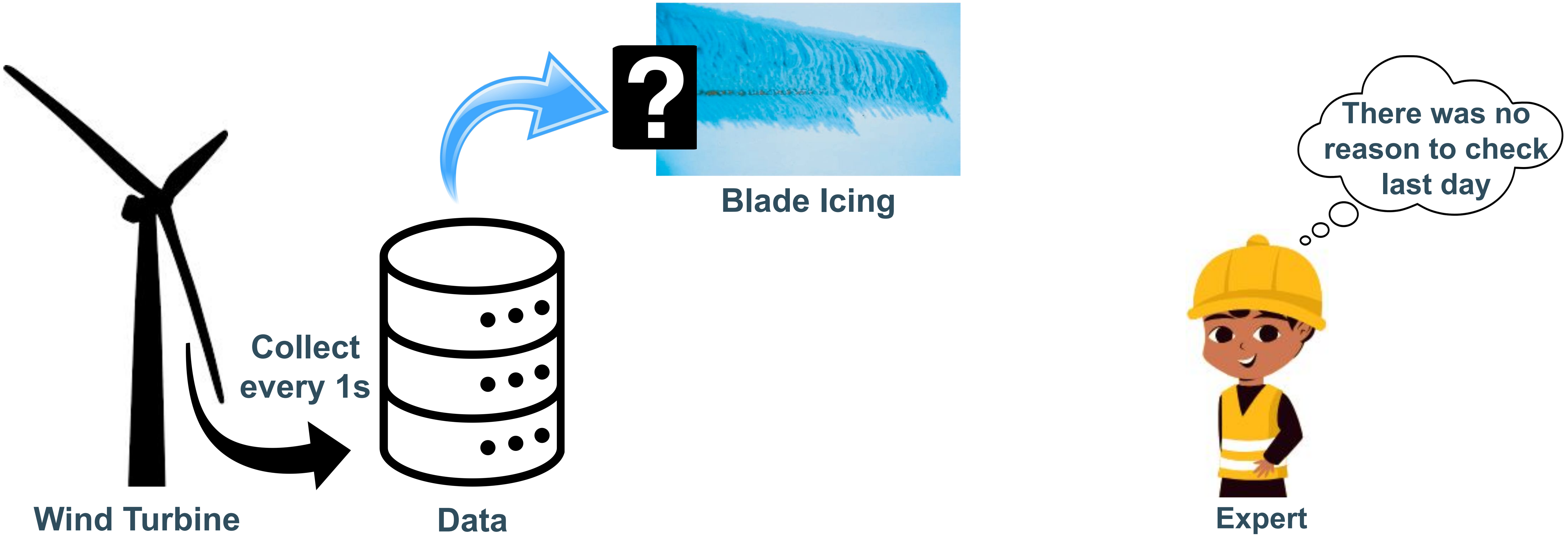
What Happens in Some Cases Is that Experts Provide <u>Coarse-Grained Labels</u> by Flagging Anomalies on a <u>Day Level</u>



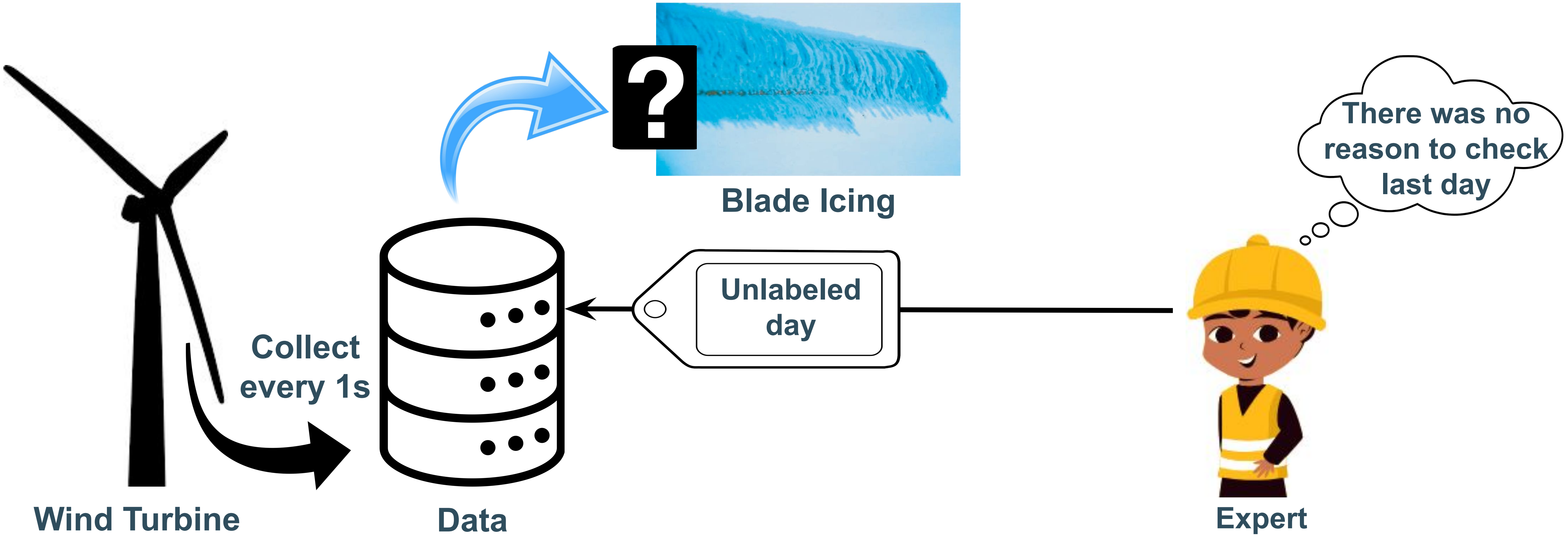
What Happens in Some Cases Is that Experts Provide <u>Coarse-Grained Labels</u> by Flagging Anomalies on a <u>Day Level</u>



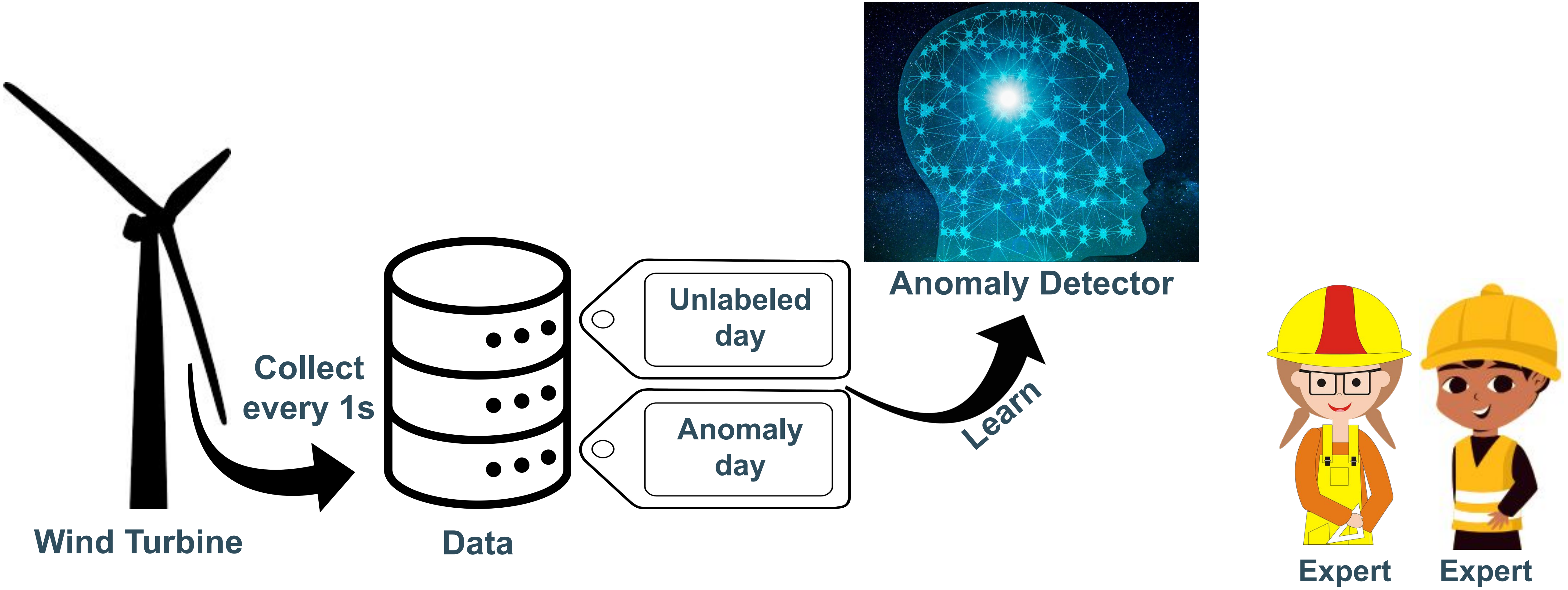
In Other Cases, Experts Do Not Provide Labels at All



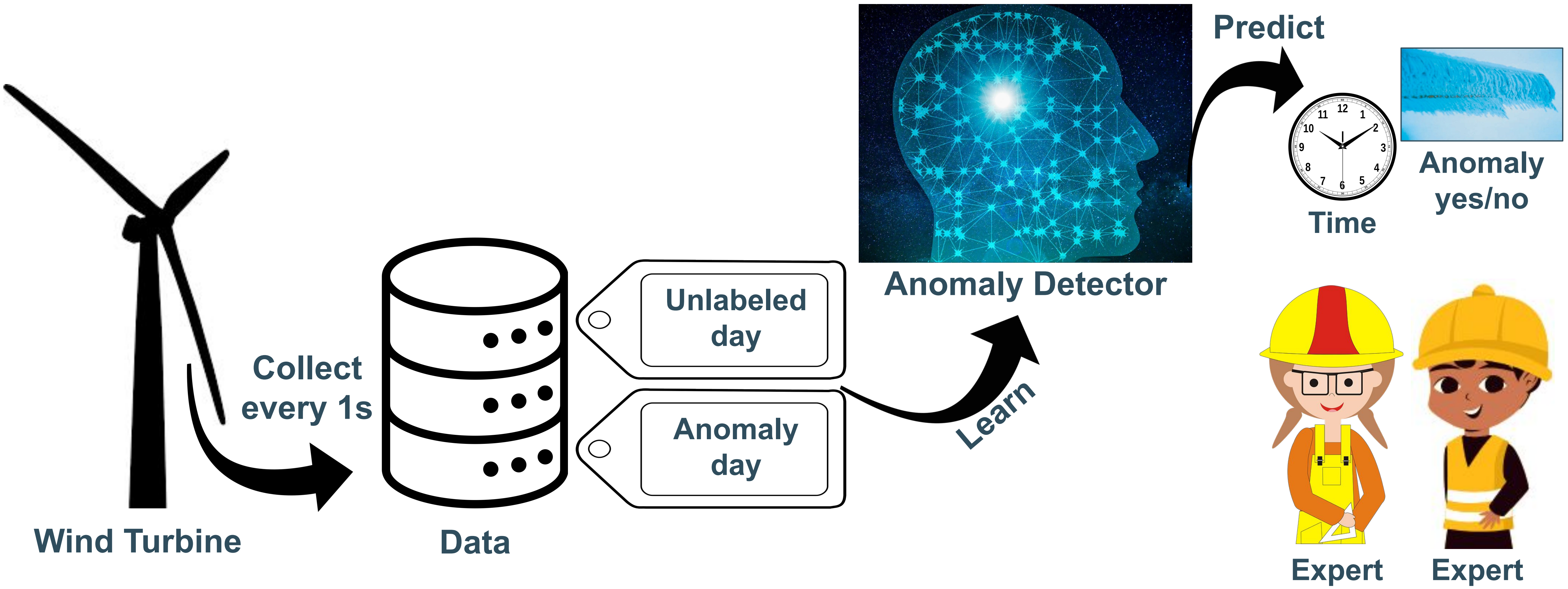
In Other Cases, Experts Do Not Provide Labels at All



Although Some Labels Are Provided on a Day Level, Experts Want the Detector to Make Predictions on an Instance Level



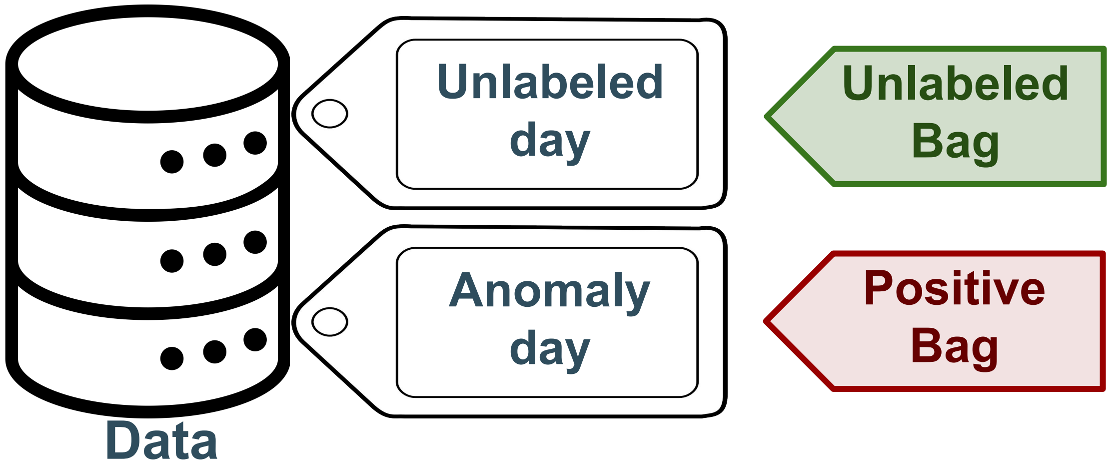
Although Some Labels Are Provided on a Day Level, Experts Want the Detector to Make Predictions on an Instance Level



This Setting Falls Into the Field of Multi-Instance Learning, Where Only Some Positive Labels Are Given on a Bag Level

Multi-Instance Learning: is a form of weakly supervised learning where the learner has access to sets of instances, called **bags**.

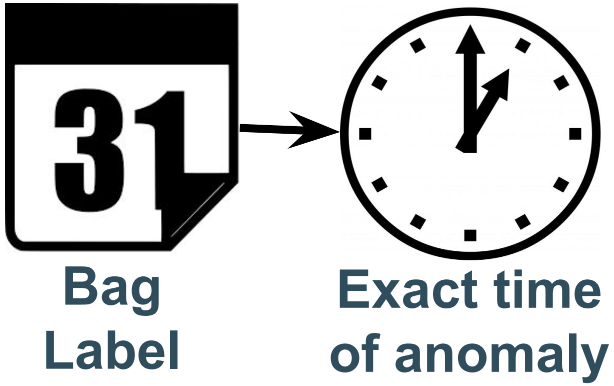
PU Learning: is the setting where a learner only has access to **positive** and **unlabeled** data.



Learning from Positive and Unlabeled Bags Has Three Main Challenges:

(a) **1**

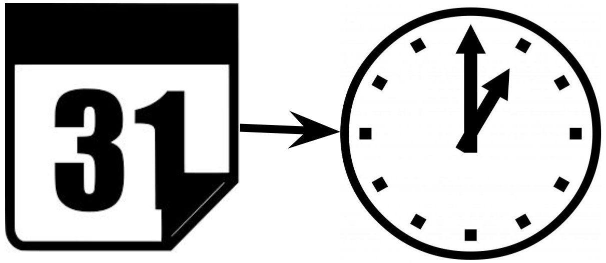
Link bags to instances



Learning from Positive and Unlabeled Bags Has Three Main Challenges:

(a) 1

Link bags to instances

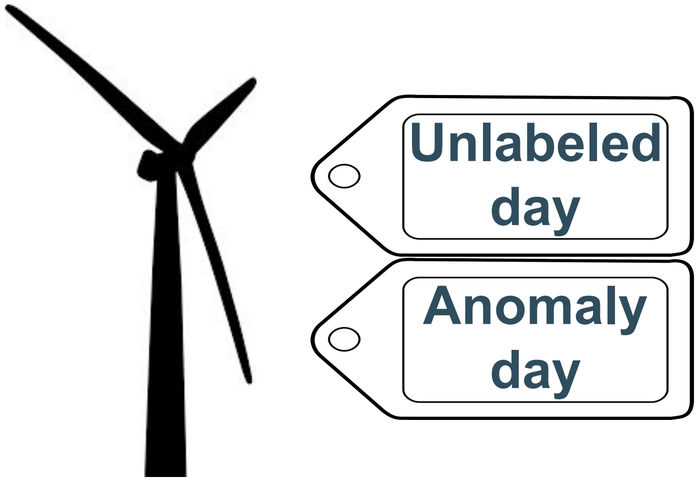


Bag Label

Exact time of anomaly

(b) 2

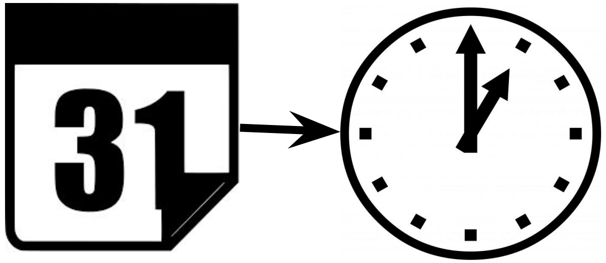
Overcome the absence of normals



Learning from Positive and Unlabeled Bags Has Three Main Challenges:

(a) **1**

Link bags to instances

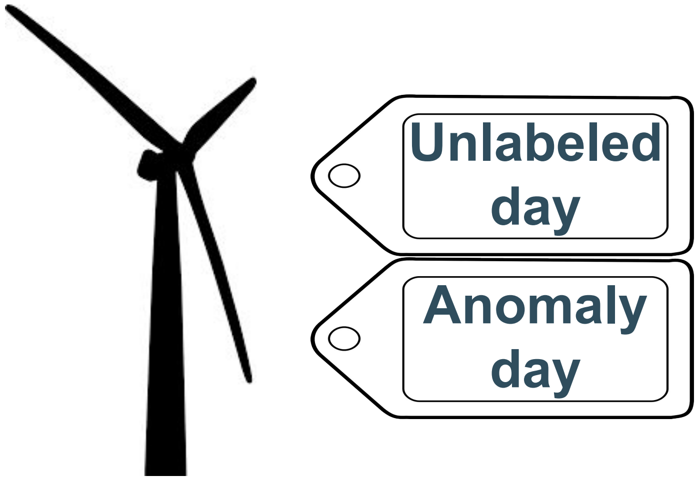


Bag Label

Exact time of anomaly

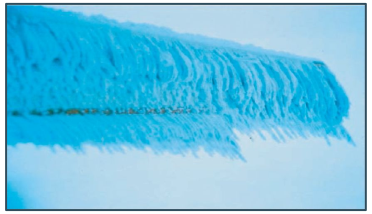
(b) **2**

Overcome the absence of normals



3

Anomalies may not follow patterns



Blade Icing

Blade Erosion

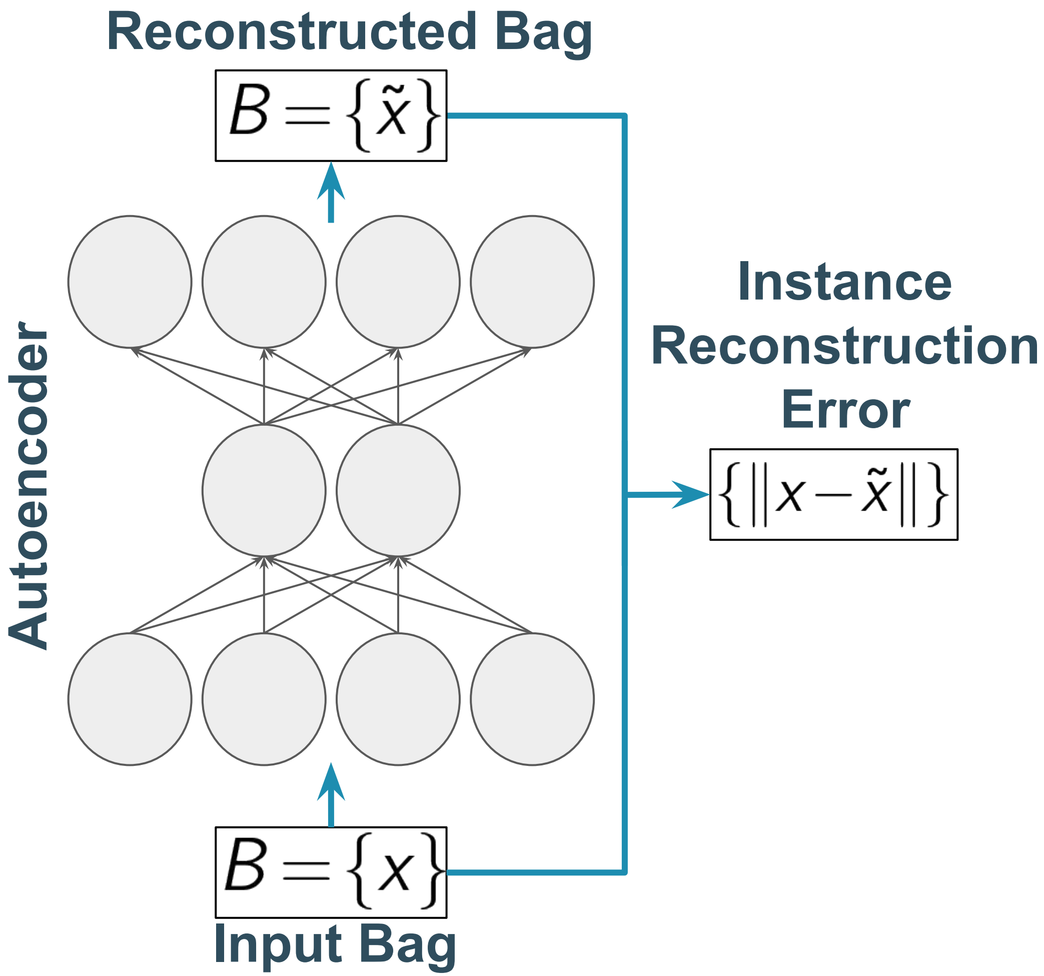
How Can We Learn from Positive and Unlabeled Bags?

We Introduce PUMA
Positive and Unlabeled Multi-instance Anomaly detector

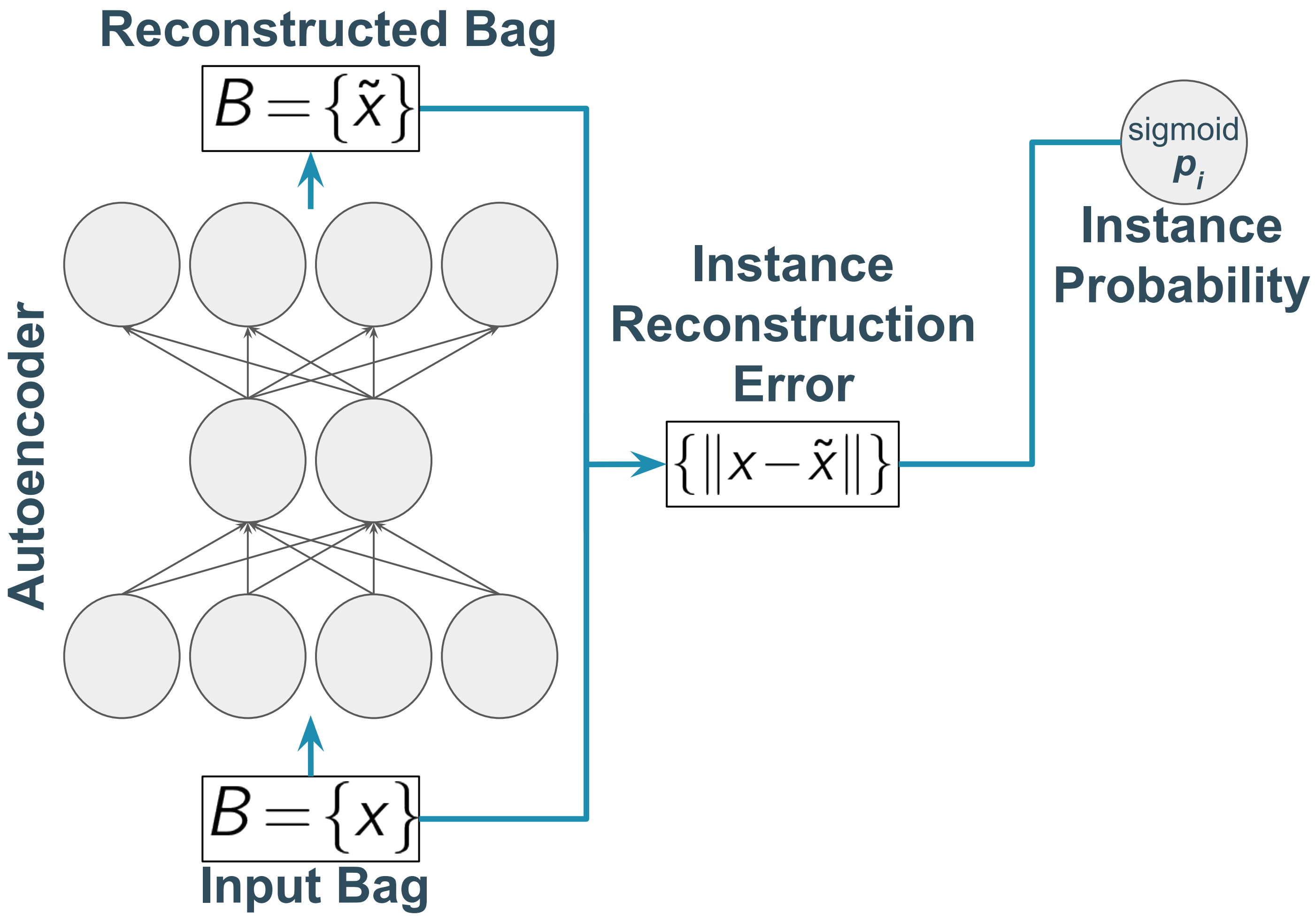
Loss-based anomaly detector that

- ***Learns from PU bags;***
- ***Predicts class probabilities both on an instance and on a bag level***

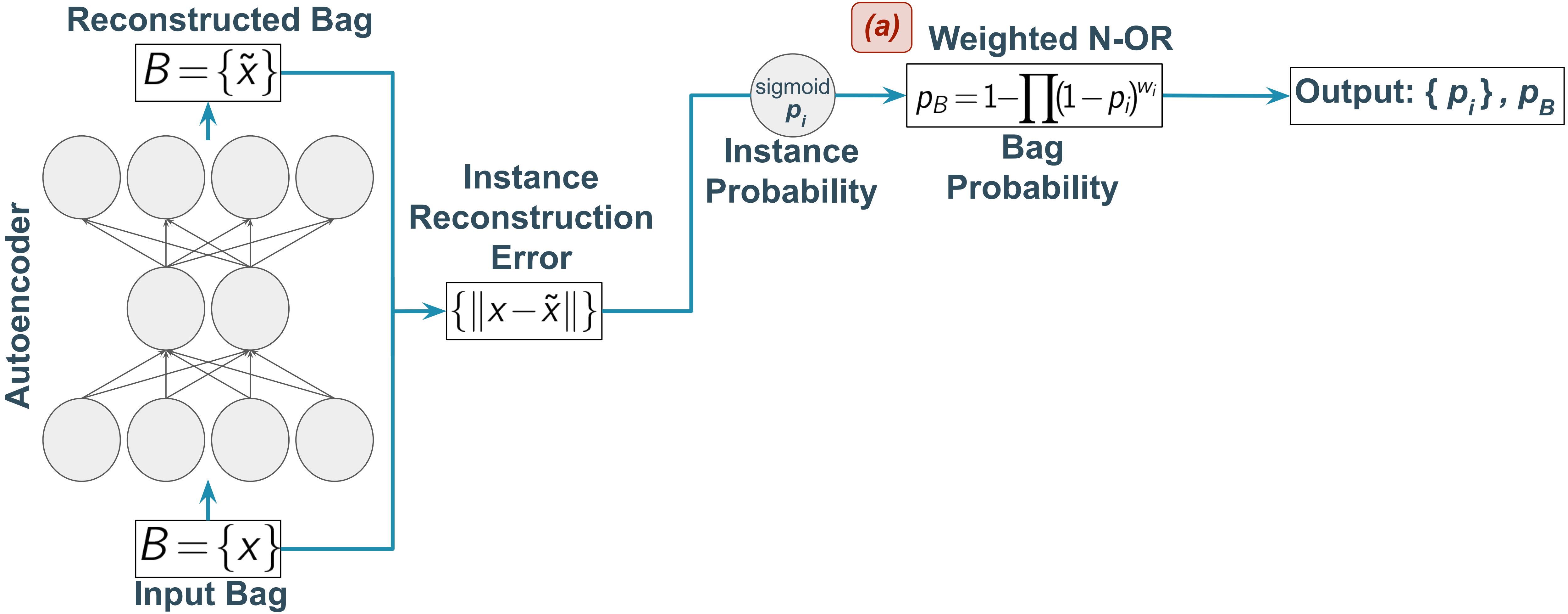
PUMA Trains an Autoencoder with a Two-Component Loss and Tackles the Challenges in Five Simple Steps



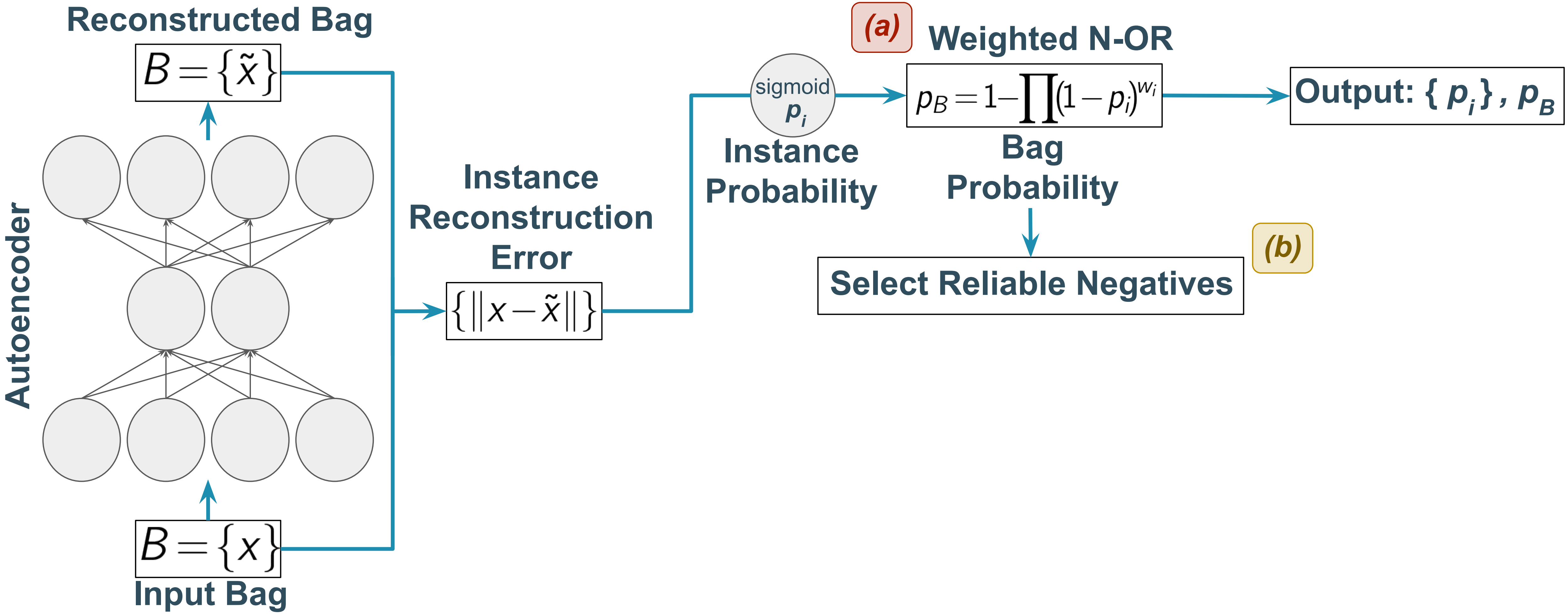
Step 1: PUMA Transforms the Instance Reconstruction Errors into Instance Probabilities



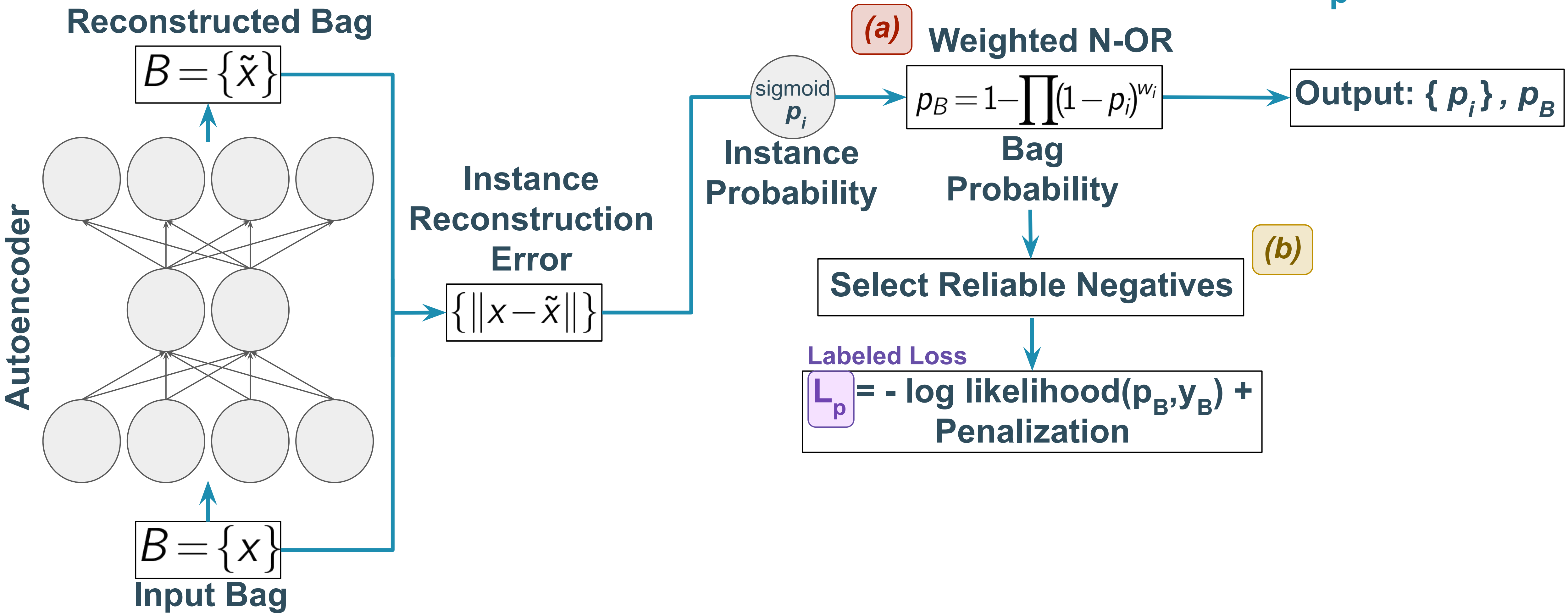
Step 2: PUMA Connects Bag and Instance Probabilities through a Weighted Noisy-OR



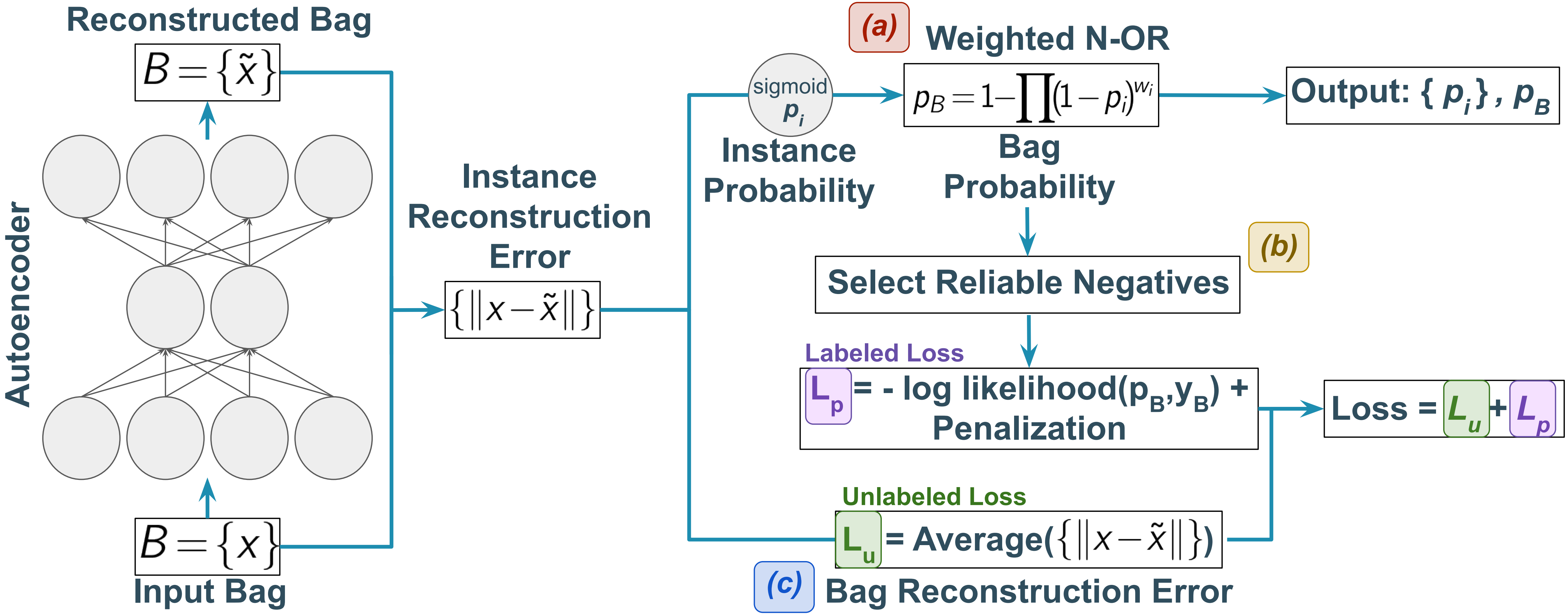
Step 3: PUMA Overcomes the Absence of Negative Labels by Pseudo-Labeling the Most Reliable Negatives



Step 4: PUMA Measures the Quality of the Bag Probabilities Through the Log-Likelihood Function (Labeled Loss L_p)



Step 5: PUMA Computes the Bag Reconstruction Error and Uses it as Unlabeled Component



Experiments: How does PUMA Compare to Existing Baselines?

Q1. How does PUMA's bag and instance level performance compare to existing approaches?

Q2. How does PUMA's performance vary upon changing the number of true anomalies in a bag?

Q3. How does changing the number of reliable negatives impact PUMA's performance?

Q4. How does increasing the number of instances per bag impact PUMA's performance?

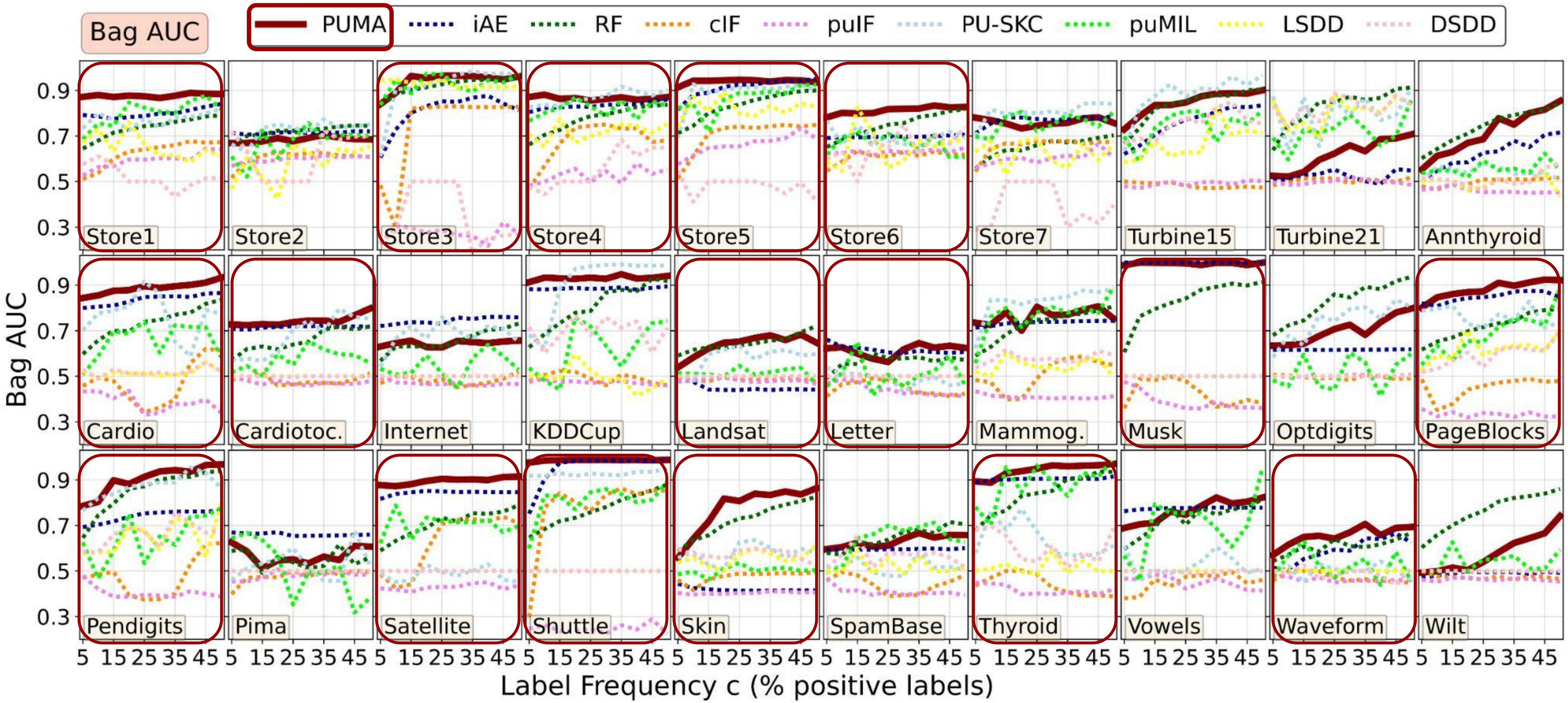
Q5. How robust is PUMA to the presence of anomalies in the unlabeled data?

Extensive Experimental Setup

- **8** baselines;
- **30** datasets: 9 real world + 21 benchmark;
- Simulate collecting incrementally 5% of positive bag labels for **10** times;
- **5** fold cross-validation + **5** random repetitions each;

30 datasets x **10** label frequencies x **5** fold cv x **5** repetitions = **7500** total experiments

Q1. How does PUMA's bag and instance level performance compare to existing approaches?



Scan here!

Learning from Positive and Unlabeled Multi-Instance Bags in Anomaly Detection

For further details:

- ★ Check out the paper online
- ★ Reach out to us via email

Lorenzo Perini, *Vincent Vercruyssen, Jesse Davis*

name.surname@kuleuven.be

https://people.cs.kuleuven.be/~lorenzo.perini/

@LorenzoPerini95