Fast Evasion Detection & Alert Management in Tree-Ensemble-Based Intrusion Detection Systems

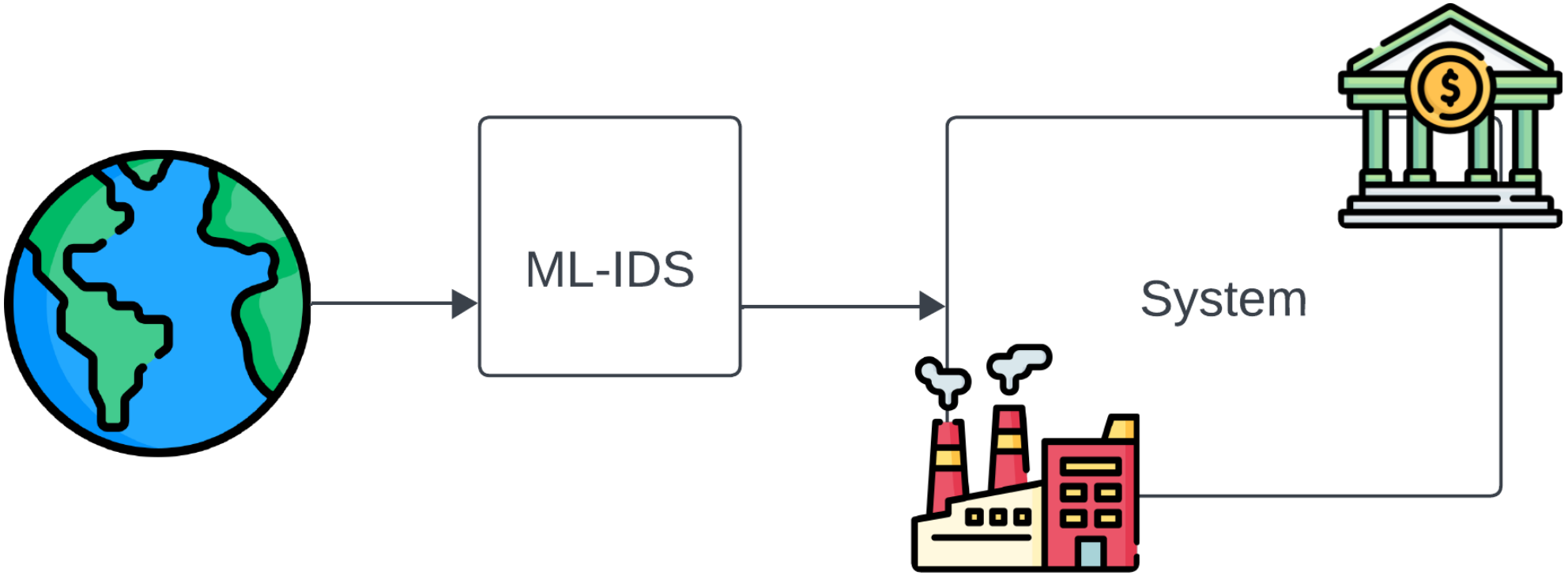
Valency Oscar Colaco & Simin Nadjm-Tehrani

Linköping University, Sweden

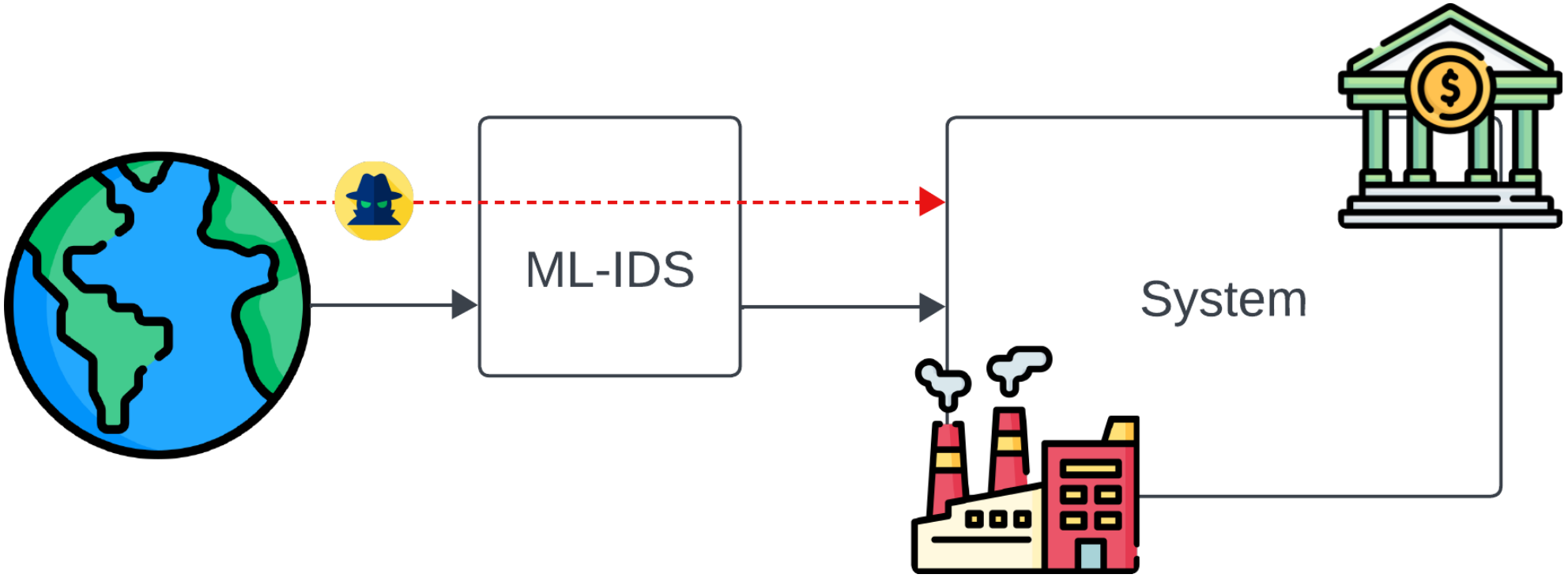
LINKÖPING UNIVERSITY

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Machine Learning Based Intrusion Detection System (ML-IDS) protects systems against cyber-attackers



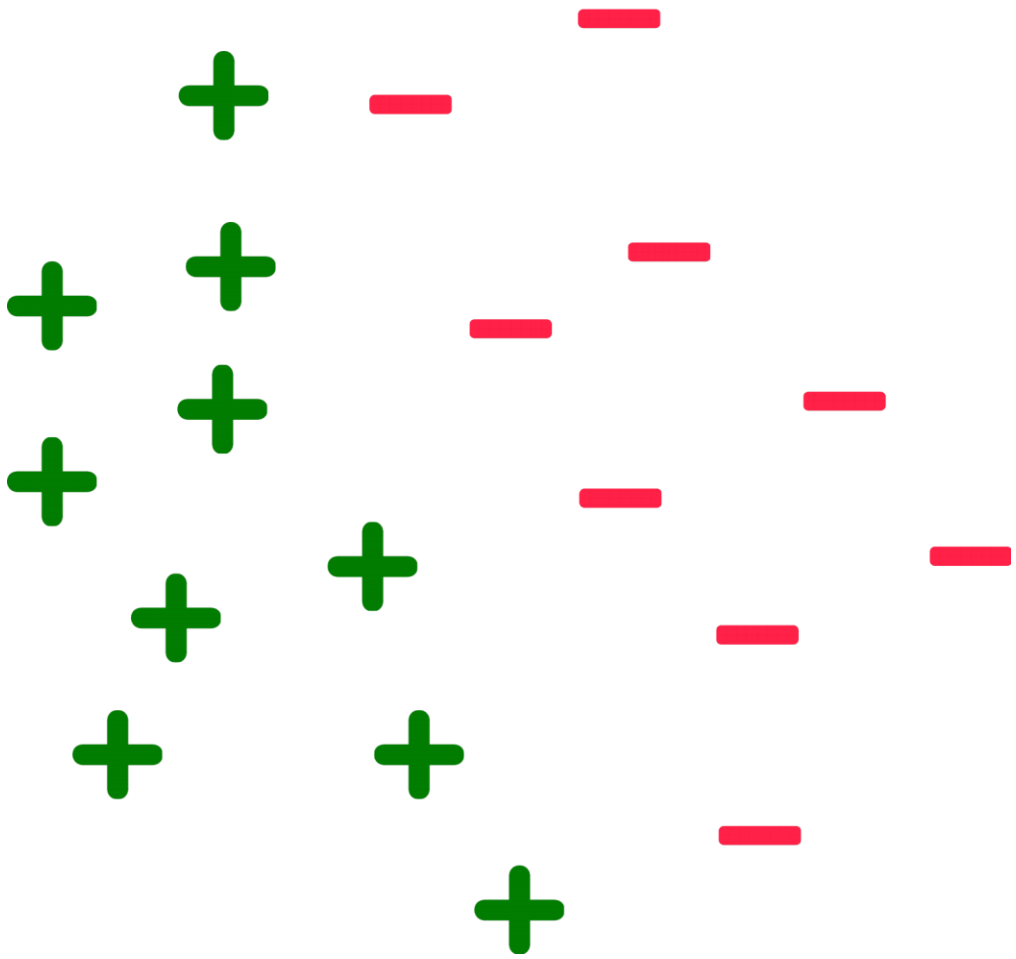
But an attacker could evade the ML-IDS



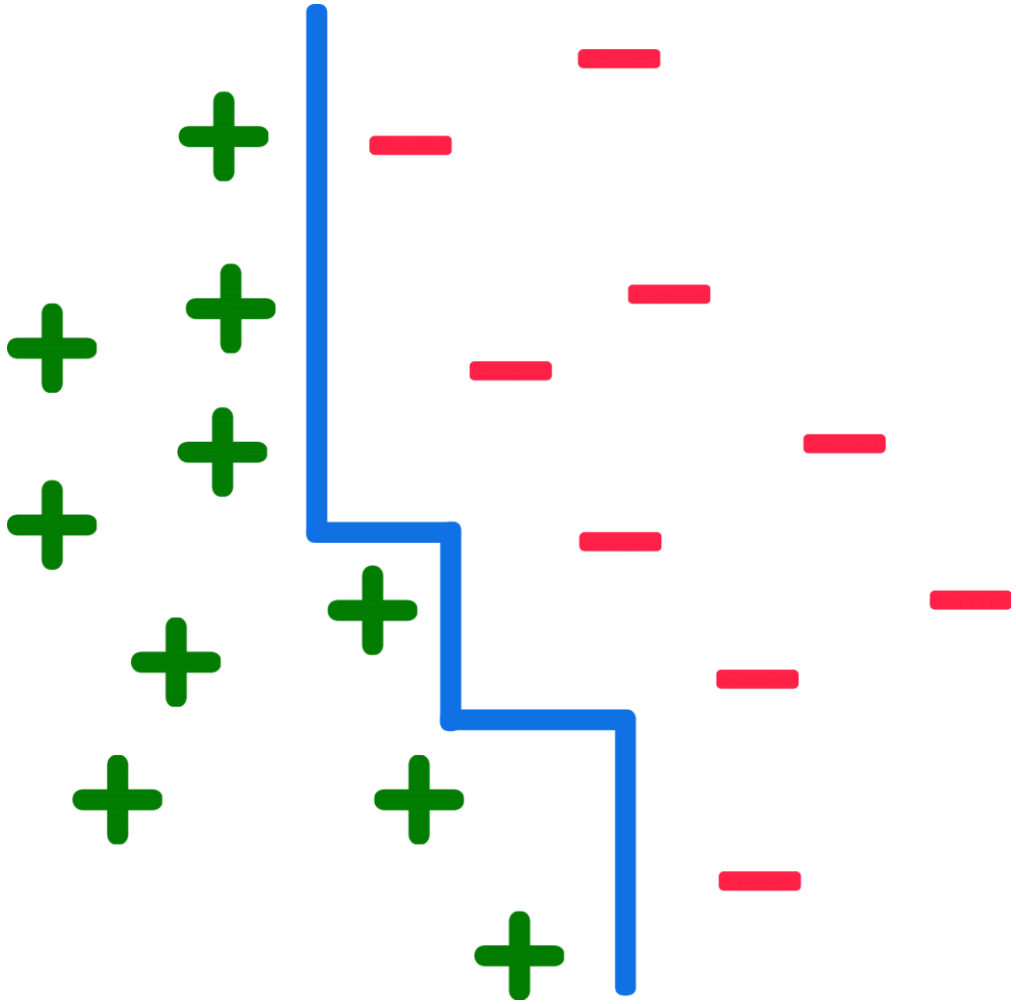
Motivation

- ML-IDSs are susceptible to **evasion attacks**
- ML-IDSs may produce many false alarms, causing **alert fatigue**
- ML-IDSs should not have **high prediction times**

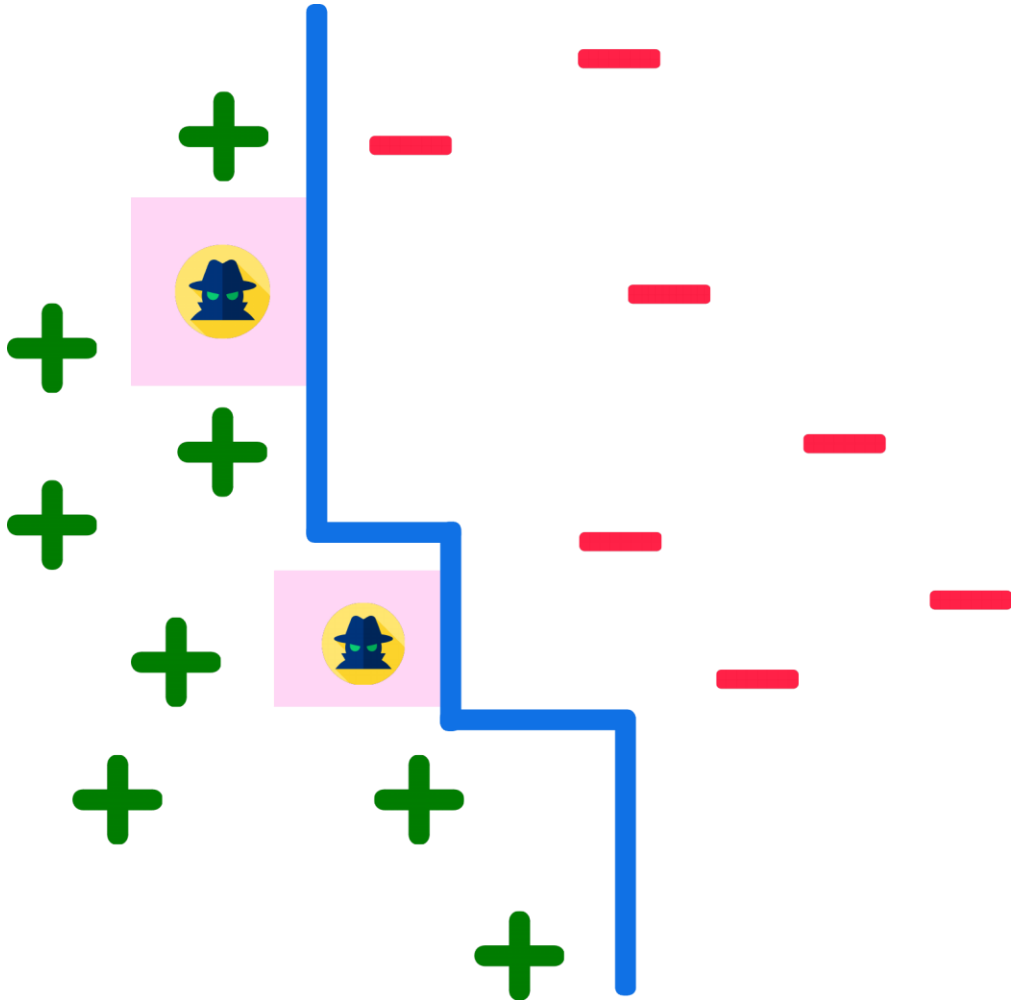
Some Insights into Adversarial Examples



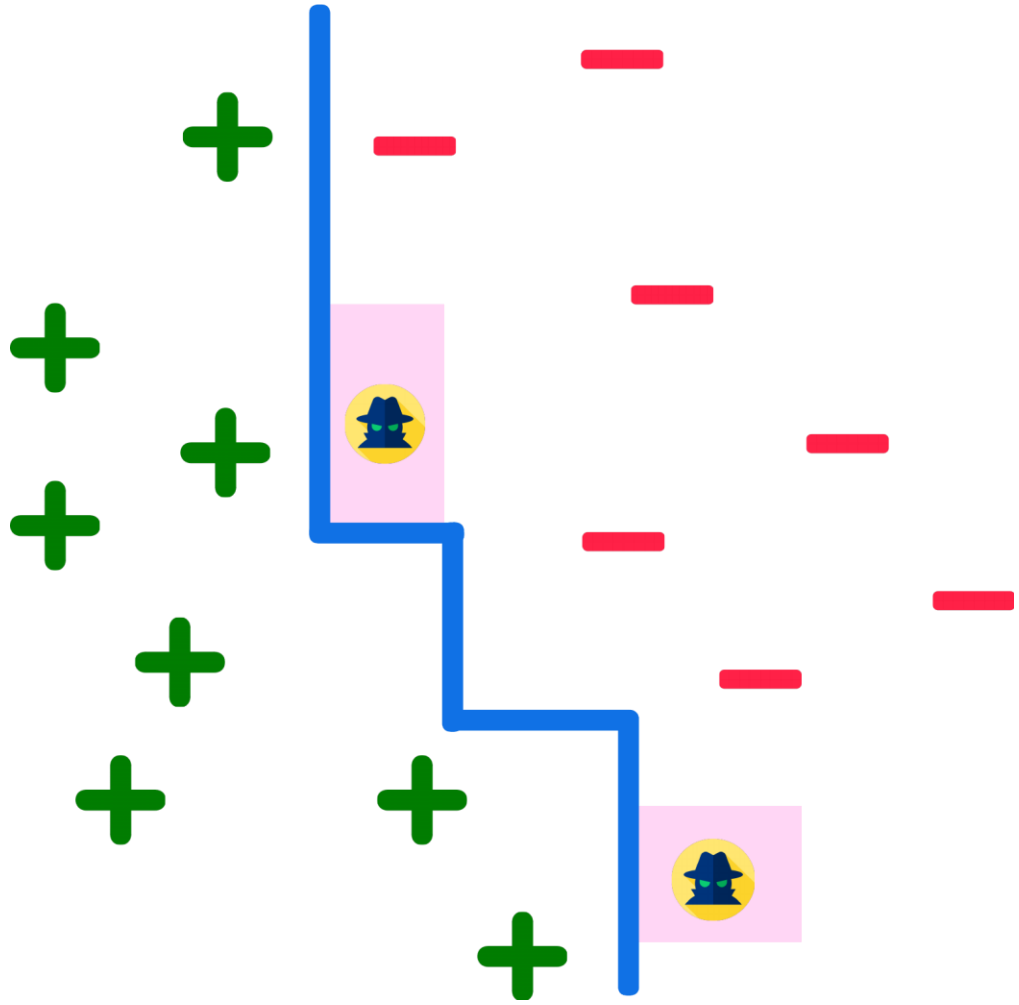
Some Insights into Adversarial Examples



Targeted Regions for Evasion Attacks

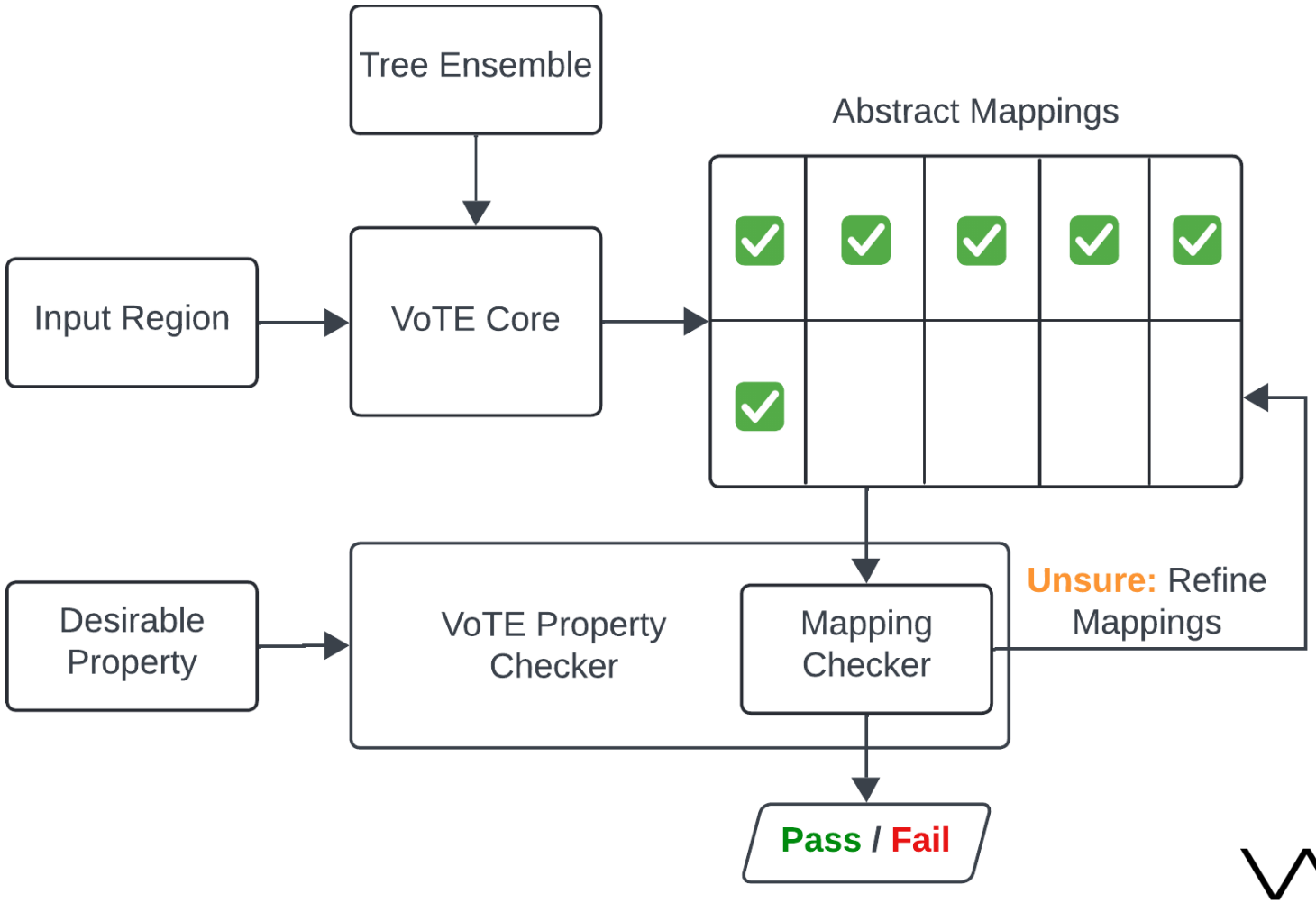


Targeted Regions for False Alarms



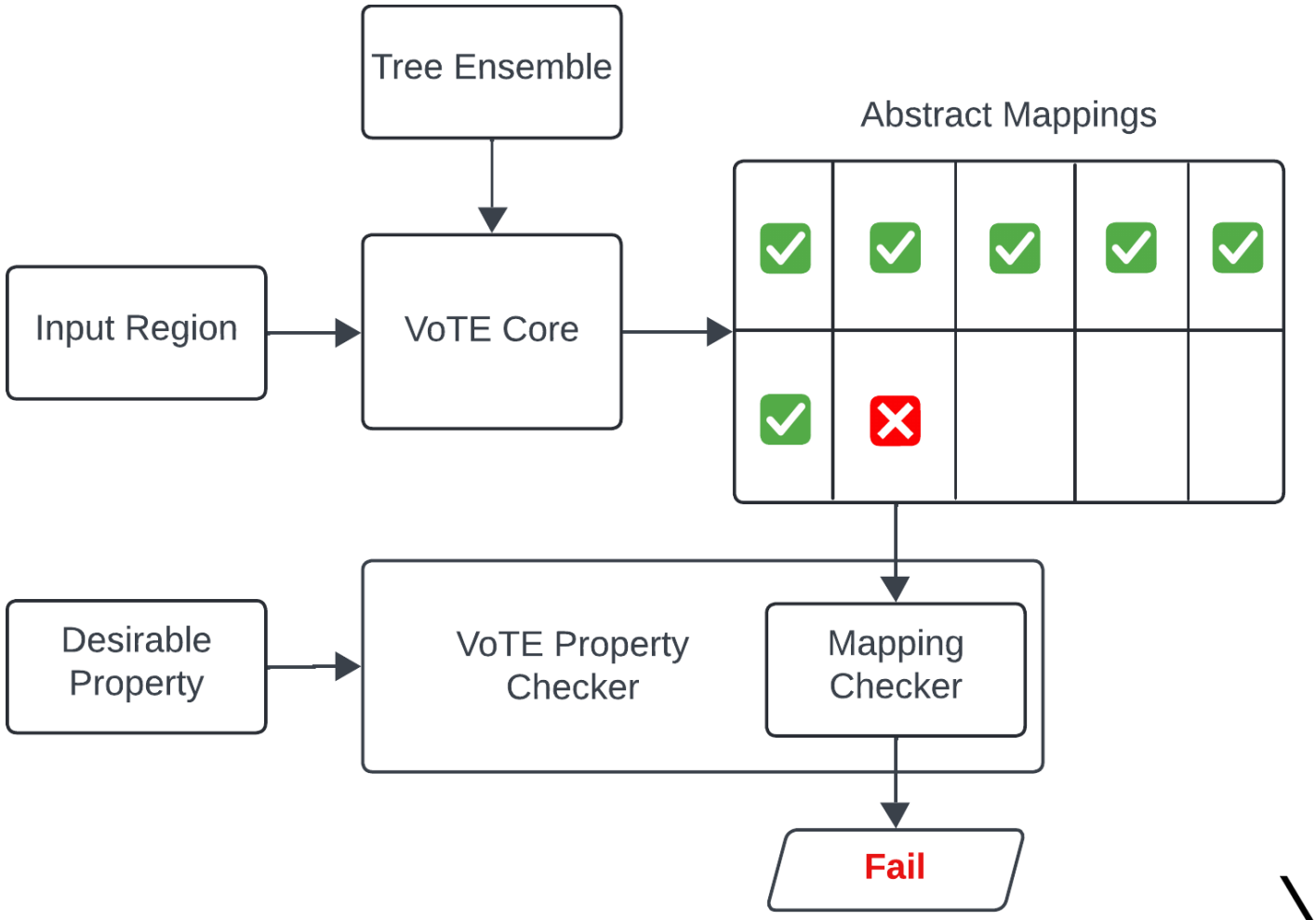
Verifier of Tree Ensembles (VoTE)

Törnblom & Nadjm-Tehrani, WAISE 2019



Failed Mappings → Counterexample Regions

Mappings that violate the property consist solely of individual counterexamples

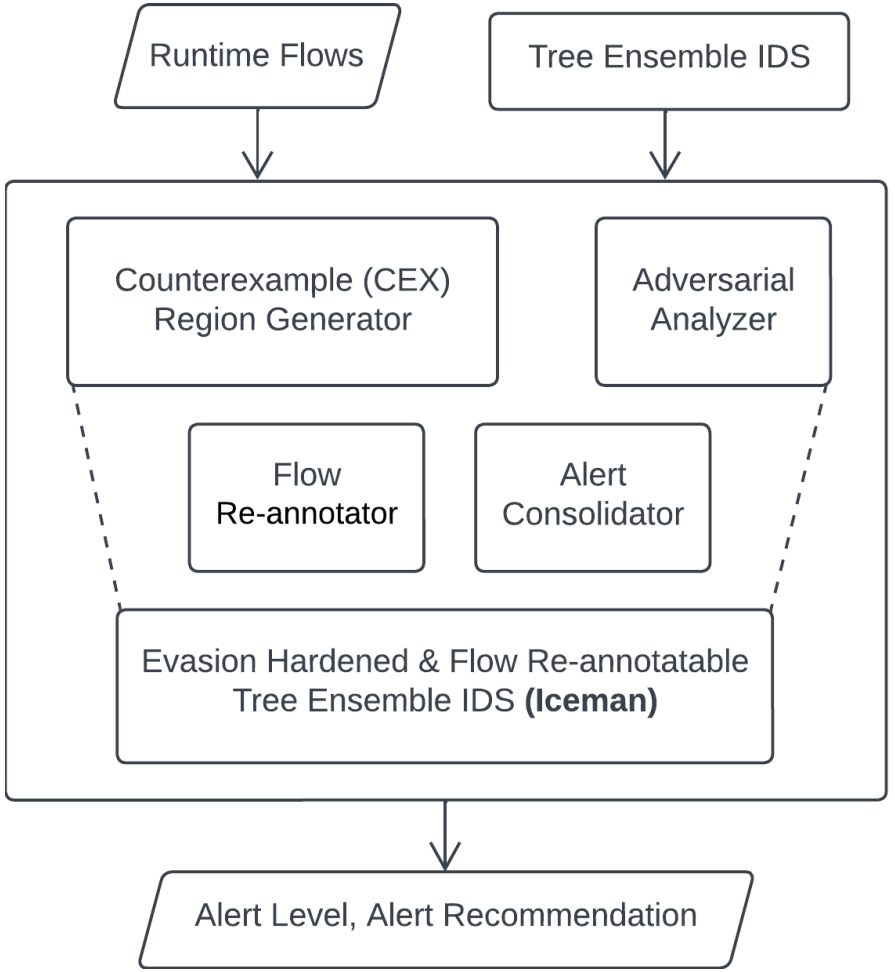


Contributions

- A method to detect evasions & produce nuanced alert insights
- **Iceman** : prototype system of an evasion-hardened, flow re-annotatable IDS
- Evaluation on four real-world case studies & SOTA comparison

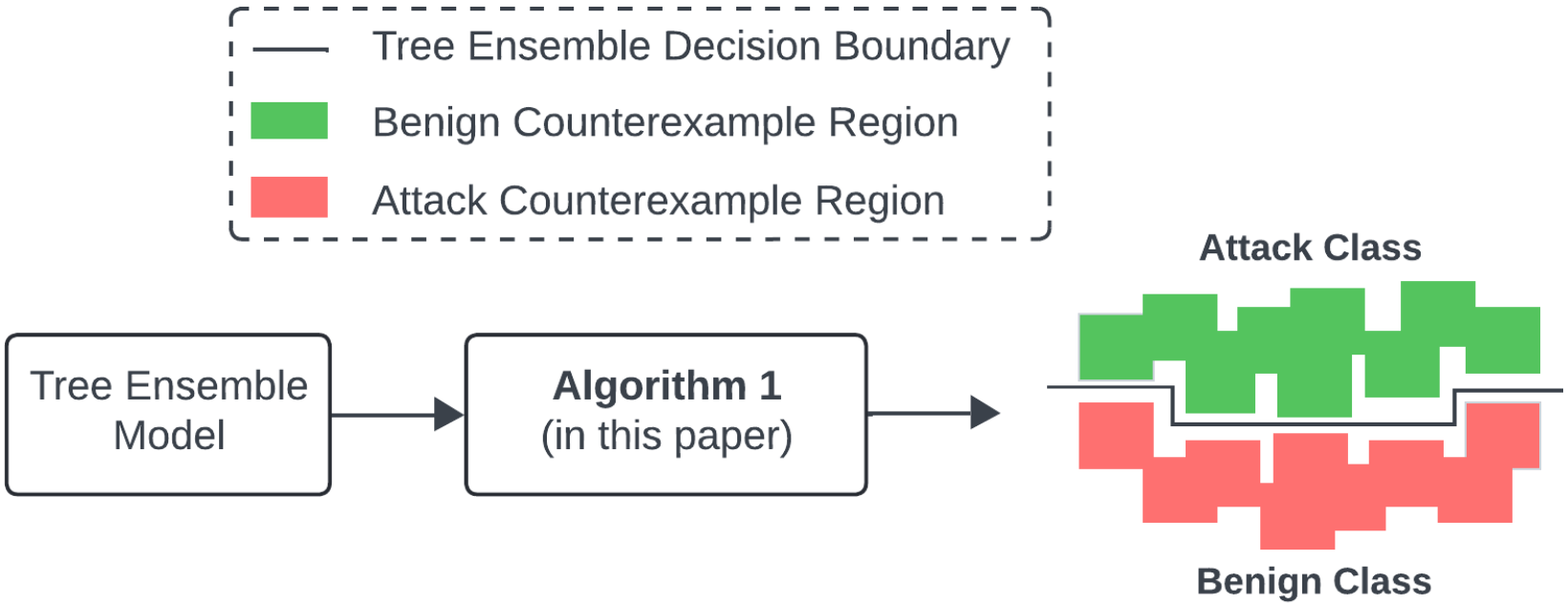
Proposed System: **Iceman**

Evasion-hardened and Flow Re-annotatable Tree Ensemble IDS



Counterexample (CEX) Region Generator

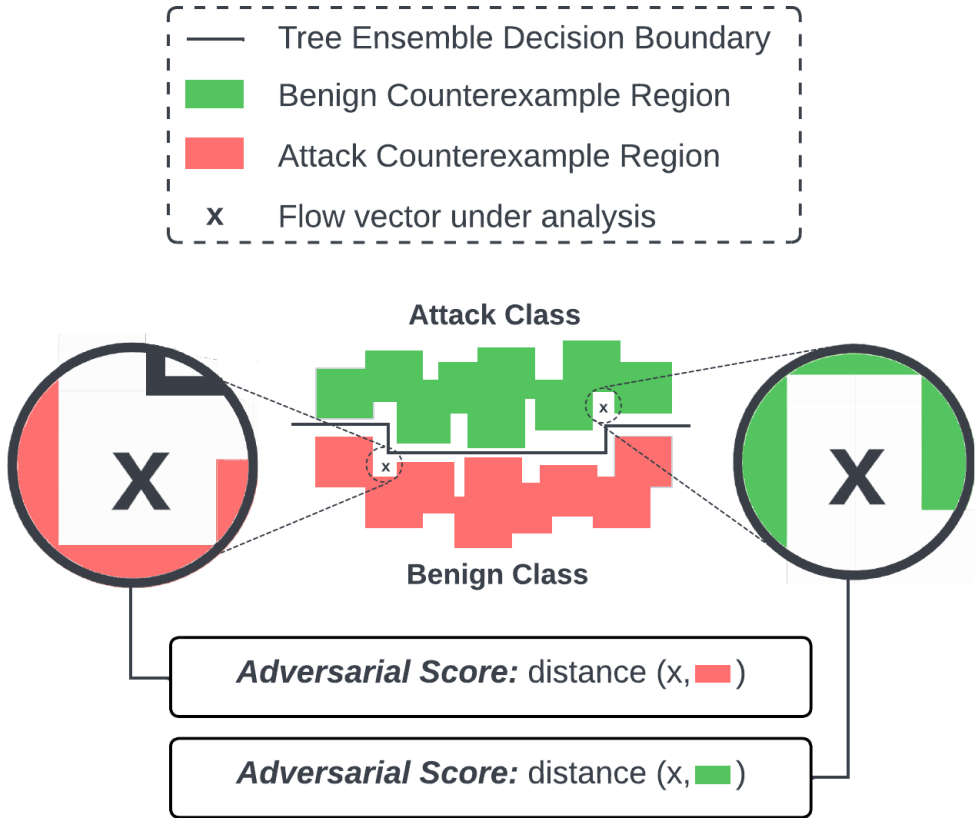
Pre-computing regions of likely evasion manipulation that attackers would normally target to evade detection



CEX Region Generator

Adversarial Analyzer

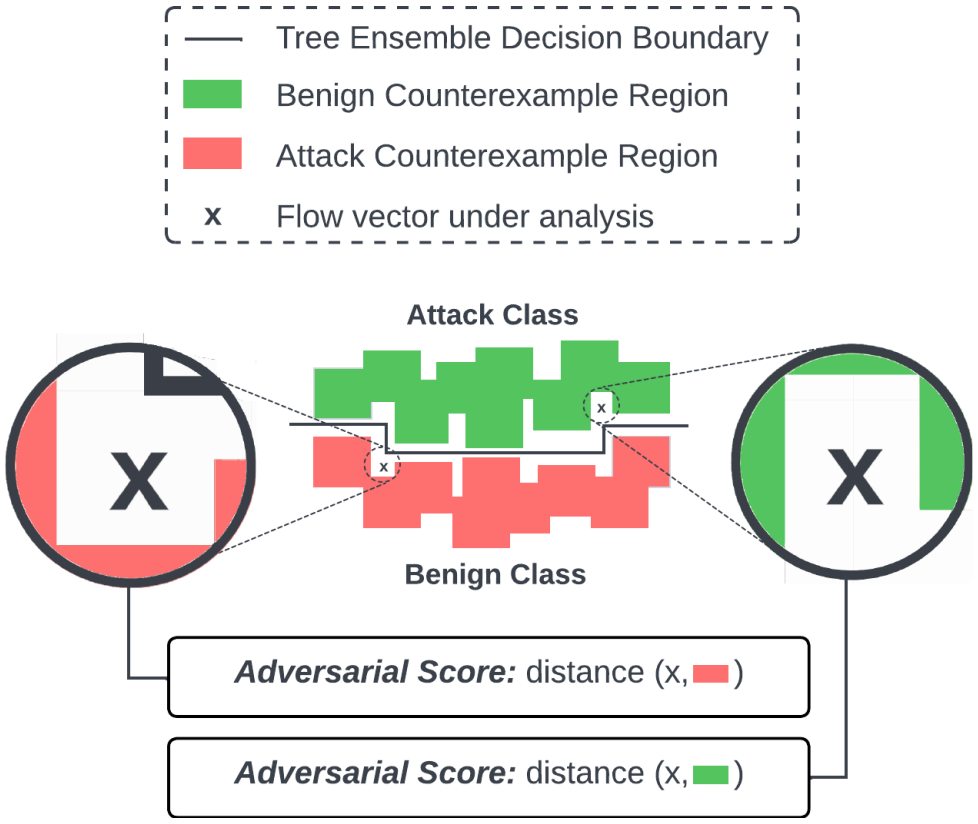
Measuring an example's adversarialness



The distance between the incoming vector and a CEX region is calculated using the **weighted l_0 distance**

Adversarial Analyzer

Measuring an example's adversarialness

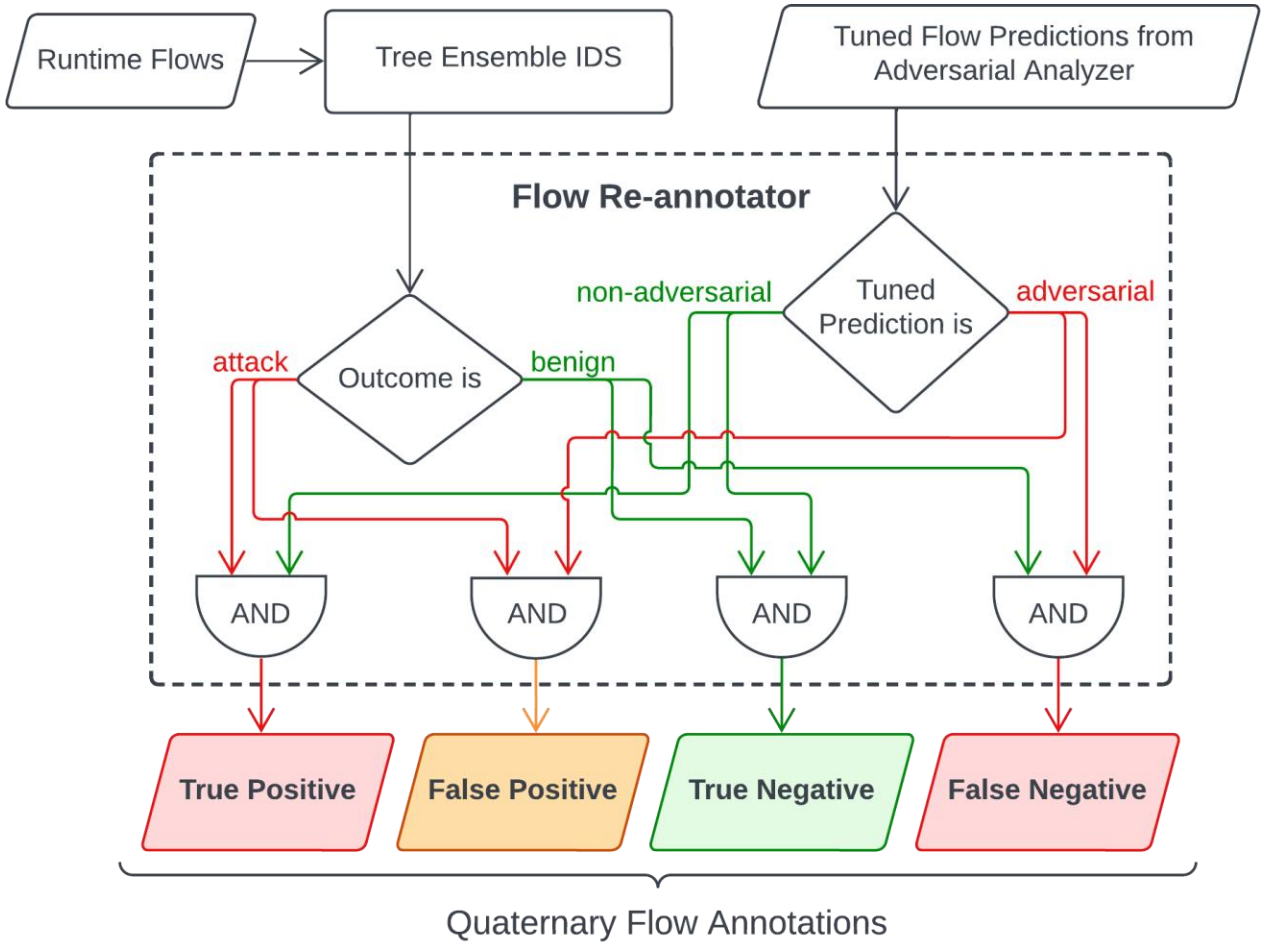


The distance between the incoming vector and a CEX region is calculated using the **weighted l_0 distance**

This distance (**adversarial score**) is thresholded to postulate flows as adversarial and non-adversarial

Flow Re-annotator

Additional Quaternary Labels based on Postulated Evasion Likelihood



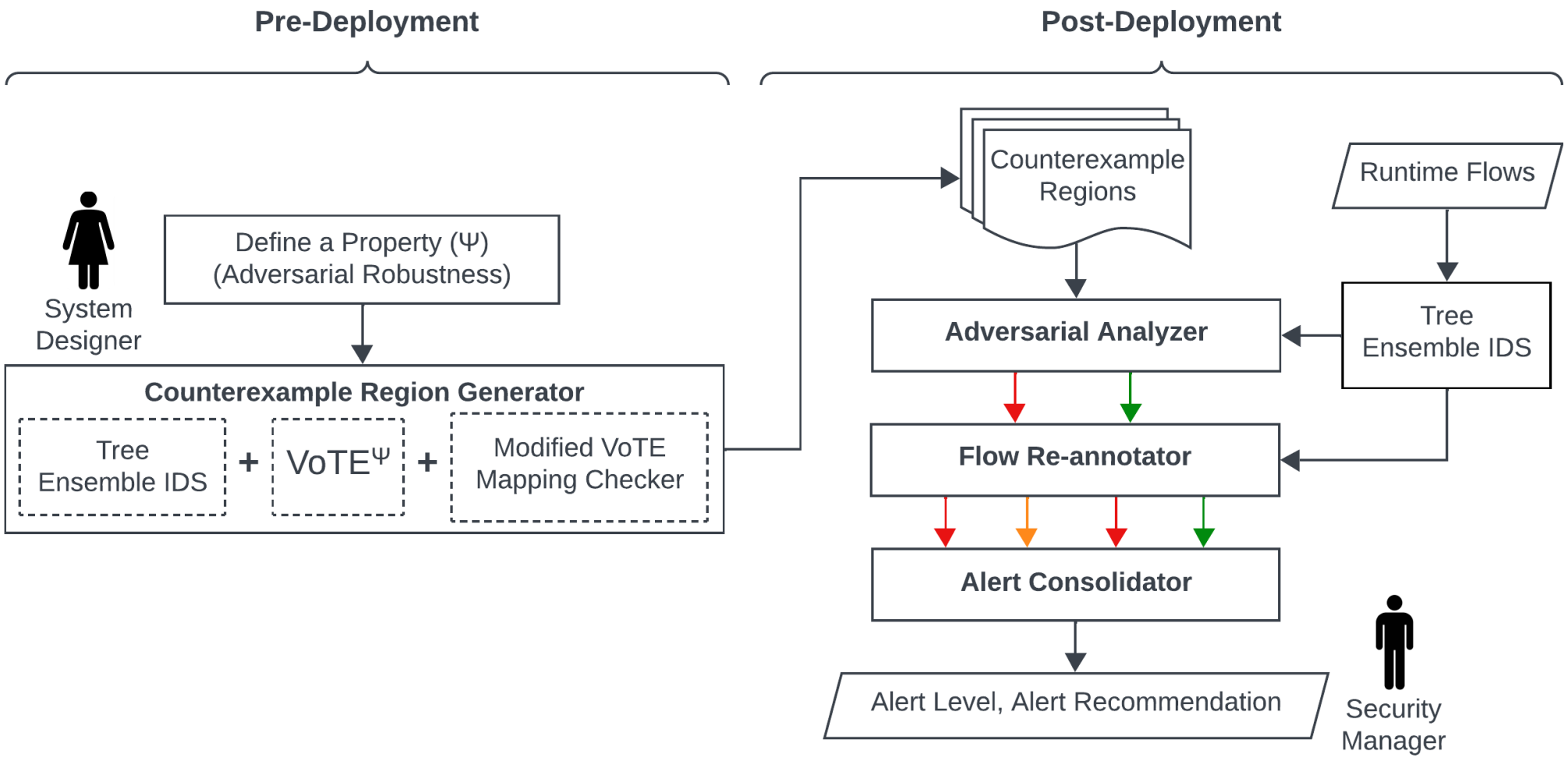
Alert Consolidator

Combining the IDS output into a single tuple (alert level, recommendation)

Flow Re-annotations	maps to	Alert Level	Alert Recommendation
True Negative Flow		0	Benign Flow, Do Nothing!
True Positive Flow		1	Attack, Investigate Now!
False Negative Flow		2	Evasion Attempt, Investigate Now!
False Positive Flow		3	Likely False Alarm, Investigate Later!

Alert Consolidation Strategy

Proposed Workflow



Experimental Setup

- **4 Datasets:** APA-DDoS, CIC-IoT-2023, HCRL-Survival-Analysis, CIC-IoV-2024
- Equal ratio of adversarial and non-adversarial samples
- Compare **Iceman** to 2 methods: OC-Score and GROOT Forests
  - Detection Accuracy & Matthews Correlation Coefficient
  - Average Prediction Times
  - Accuracy of Alert Filtering & Prioritization

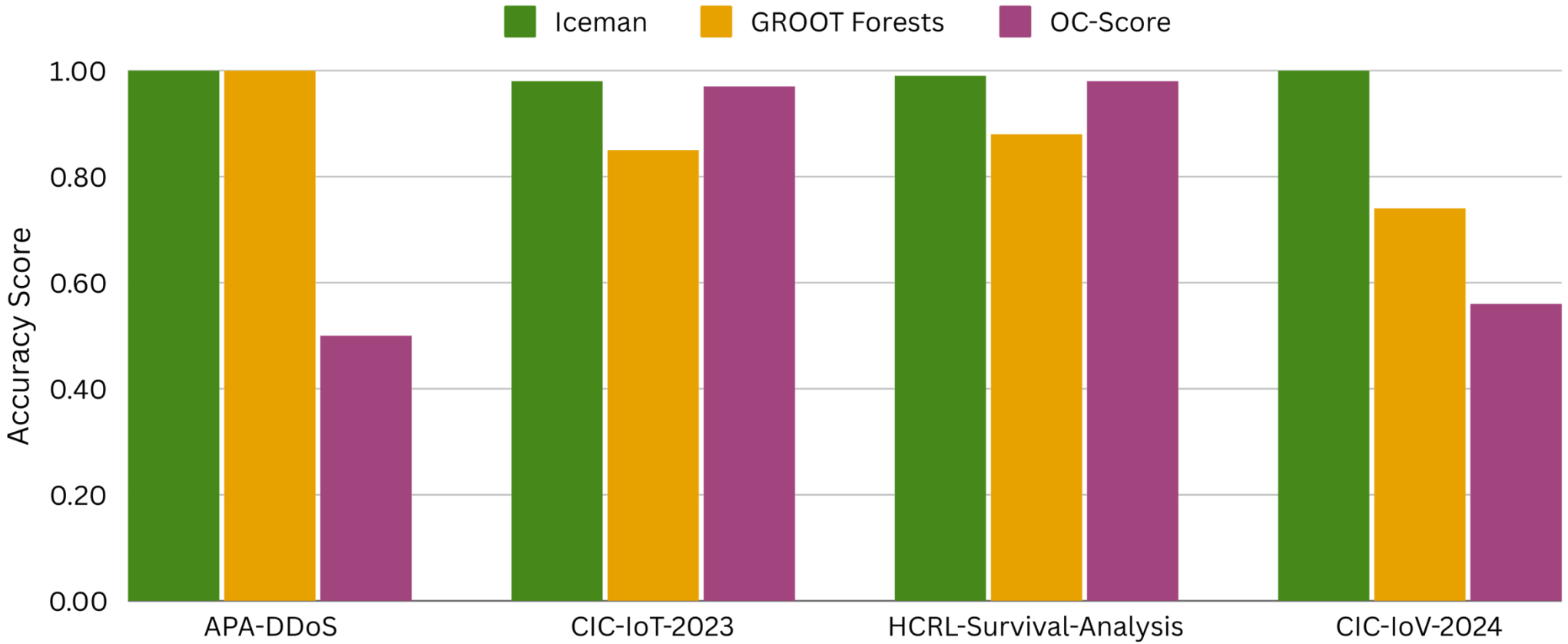
Baseline IDS Performance

Iceman hardens a baseline Tree Ensemble IDS

Case Study	Baseline Tree Ensemble IDS	Accuracy Score	Matthew's Correlation Coefficient
APA-DDoS	XGBoost GBM (depth = 5, trees = 50)	1.00	1.00
CIC-IoT-2023	XGBoost GBM (depth = 5, trees = 25)	1.00	0.91
HCRL-Survival-Analysis	Random Forest (depth = 10, trees = 50)	1.00	1.00
CIC-IoV-2024	Random Forest (depth = 10, trees = 25)	1.00	1.00

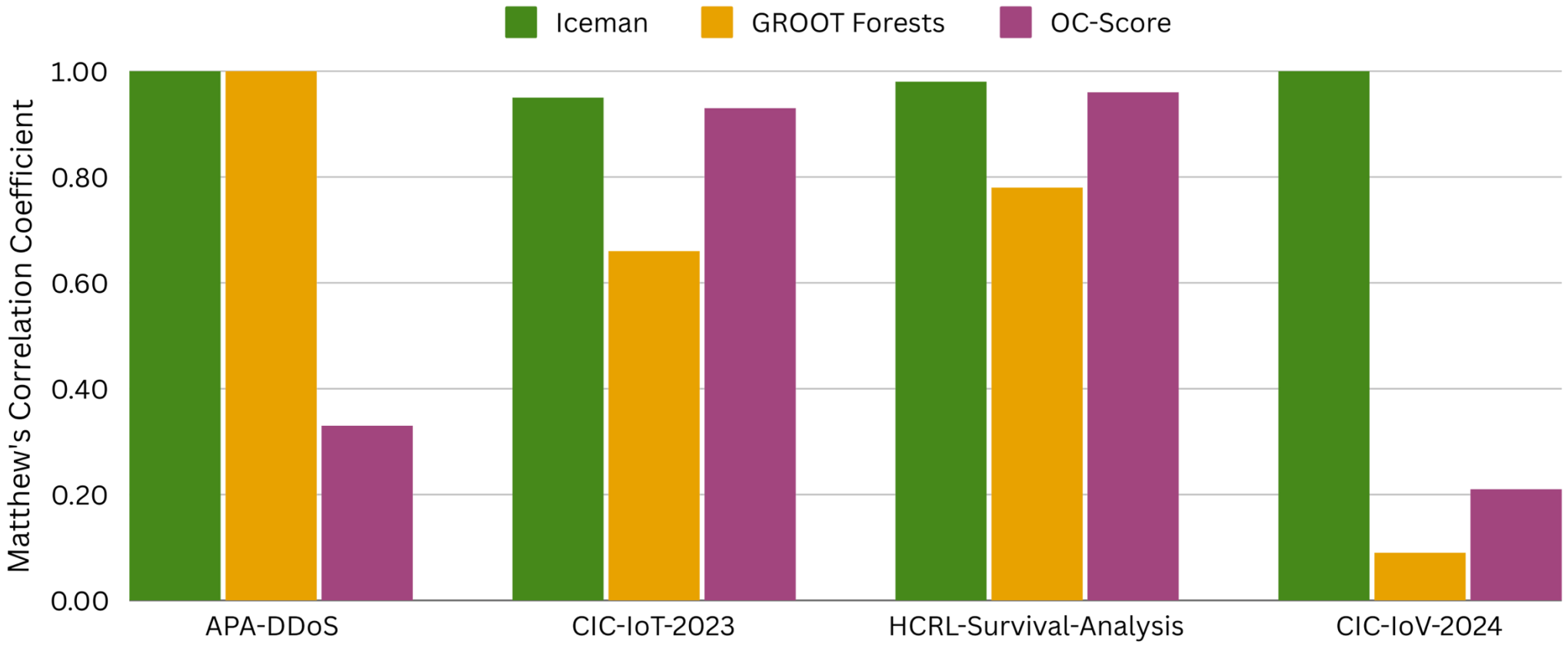
Accuracy of Baseline Detectors is Preserved

Good Detection Accuracy despite Evasion Attacks



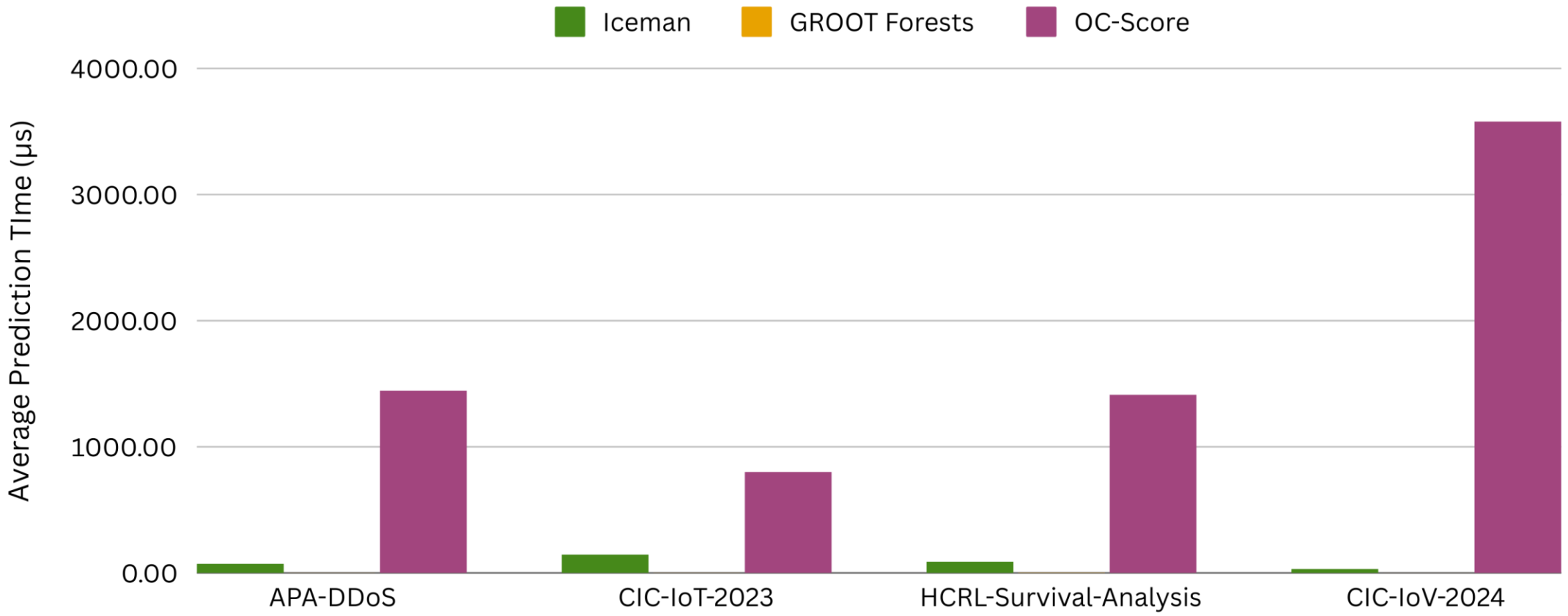
Well Balanced Detection Performance

Good Matthew's correlation coefficient in four case studies



Low Prediction Latency

5-115x faster compared to OC-Score



Good Alert Management Performance

Accurate Alert Filtering and Alert Prioritization

Case Study	Alert Prioritization Accuracy	Alert Filtering Accuracy
APA-DDoS	1.00	1.00
CIC-IoT-2023	0.98	0.98
HCRL-Survival-Analysis	0.99	0.99
CIC-IoV-2024	1.00	1.00

Conclusion

- Our method can have benefits for safety and security
- Crafting counterexample regions is time-consuming but done offline
- Scalability depends on underlying formal verification tool
- Future Works → Counterexample-**Region** Guided Inductive Synthesis

Questions?

Fast Evasion Detection & Alert Management in Tree-Ensemble-Based Intrusion Detection Systems

Valency Oscar Colaco & Simin Nadjm-Tehrani

valency.colaco@liu.se

LINKÖPING UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM