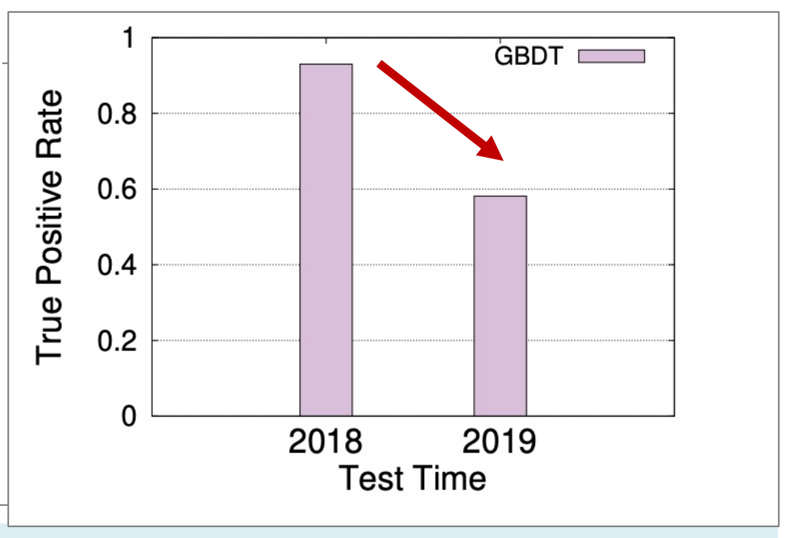
Combating Concept Drift in Security Applications with Self-Supervised Learning

PI: Gang Wang, University of Illinois at Urbana-Champaign

https://gangw.cs.illinois.edu

Problem Description

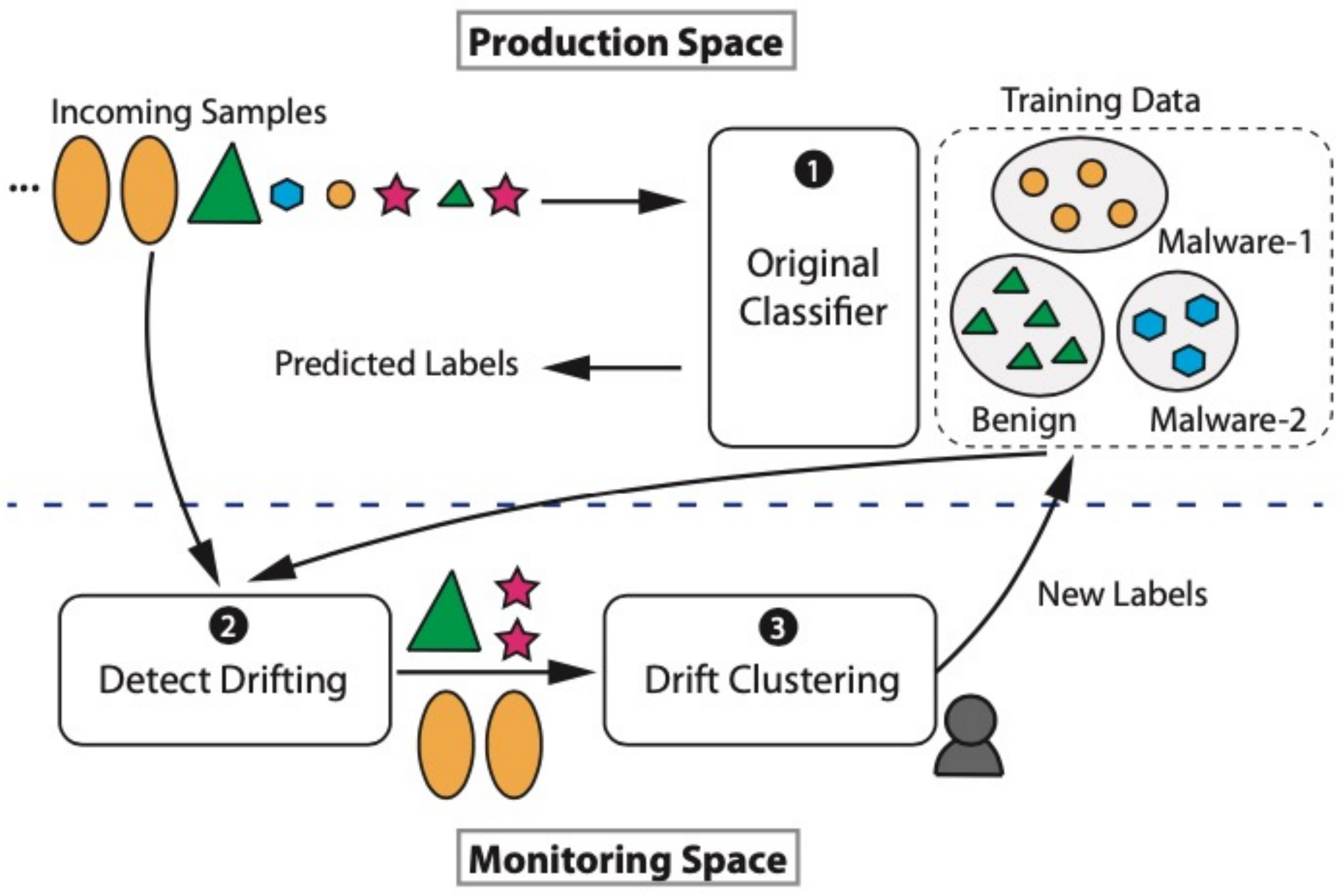
- A learning-based security system often performs worse over time
- **Concept drift** caused by behavior changes from both benign and malicious players
- Periodic re-training demands significant **labeling** efforts

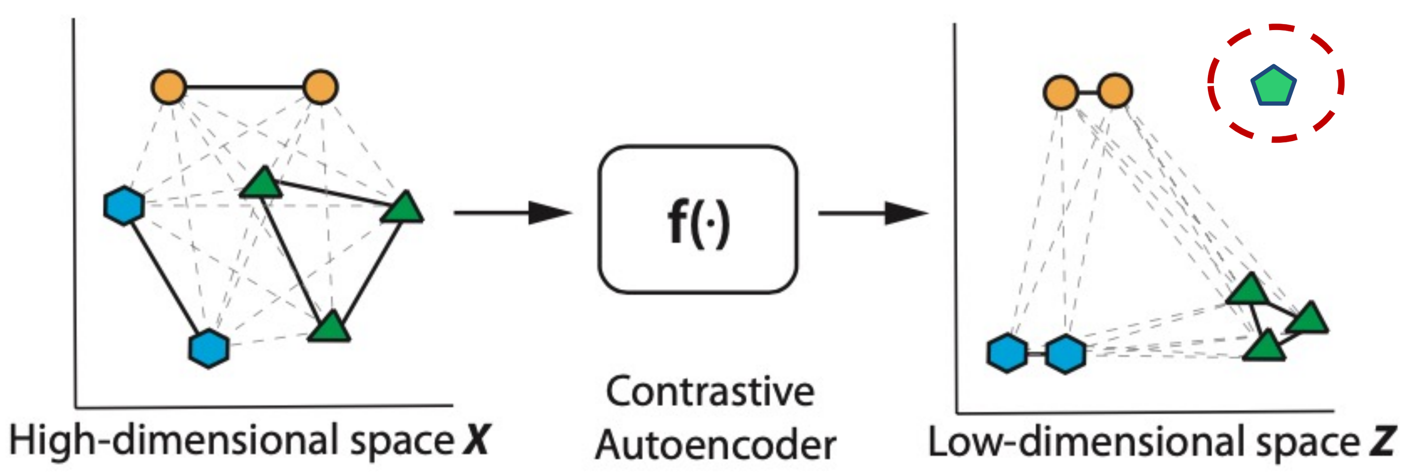


Assuming we will never have the representative labels, what can we do to significantly improve the adaptability and resilience of learning-based security defense with extremely limited labeling capability?

Method

- Self-supervision + domain-specific insights
- **Obtain supervision from the data itself**
  - Contrastive learning
  - Generative adversarial networks (GAN)
- Proactively detect drifting samples
- Enrich/refine noisy labels → higher-quality labels





Drifting Sample Detection (CADE - USENIX 21)

- Use contrastive learning to learn a compressed representation of the training data **by contrasting with existing samples**
- Identify incoming samples that do not fit in within any existing families
- Rank and cluster drifting samples for labeling
- Tested on real-world malware datasets

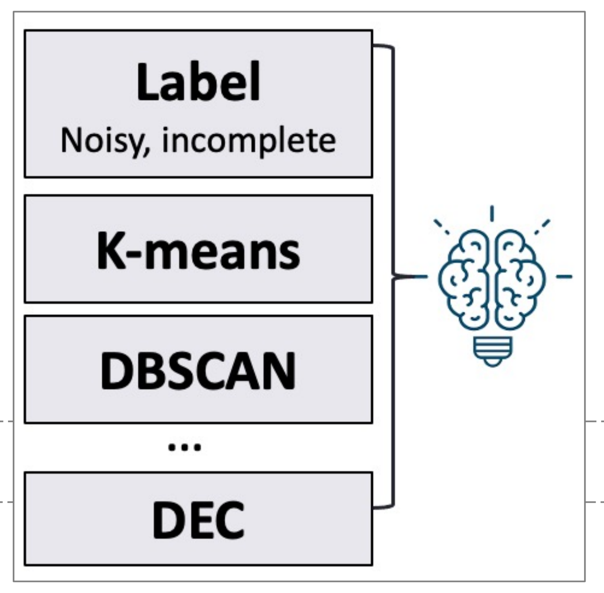
BODMAS Malware

BLUE HEXAGON

Open malware dataset *for concept drift detection*

Work with Low-quality Labels (FARE - NDSS 21)

- Missing classes, coarse-grained labels, label scarcity
- Reduce uncertainty by combining
  - **Multiple simple unsupervised** clustering algorithms
  - Given "noisy" **labels** (weekly supervised)
- Contrastive learning to
  - Fuse given labels and clustering results
  - Map data into a low-d space before final clustering
- Evaluated on fraud detection
  - E-commerce service
  - Low false positive rate



Ongoing/Next Steps

Measurement

- Quantify concept drifts in real-world malware and network traffic data; explore its reasons

Attacks

- Adversarial attacks that aim to manipulate concept drift detection and the data labeling process

Defense

- Robustify the model updating process

NSF Support

- **CNS-2055233**: SaTC: CORE: Small: Towards Label Enrichment and Refinement to Harden Learning-based Security Defenses
- **CNS-2030521**: CAREER: Machine Learning Assisted Crowdsourcing for Phishing Defense

References

- "CADE: Detecting and Explaining Concept Drift Samples for Security Applications". L. Yang, W. Guo, Q. Hao, A. Ciptadi, A. Ahmadzadeh, X. Xing, and G. Wang. Proc. of **USENIX Security**, 2021
- "FARE: Enabling Fine-grained Attack Categorization under Low-quality Labeled Data". J. Liang, W. Guo, T. Luo, V. Honavar, G. Wang, and X. Xing. Proc. of **NDSS**, 2021
- "It's Not What It Looks Like: Manipulating Perceptual Hashing based Applications". Q. Hao, L. Luo, S. Jan, and G. Wang. Proc. of **CCS** 2021