Malware-DNA: Machine Learning for Malware Analysis that Treats Malware as Mutations in the Software Genome

Maksim E. Eren *#, Kim O. Rasmussen ◆, Charles Nicholas #, Boian S. Alexandrov ◆
*LANL Analytical Division, ◆LANL Theoretical Division, #UMBC

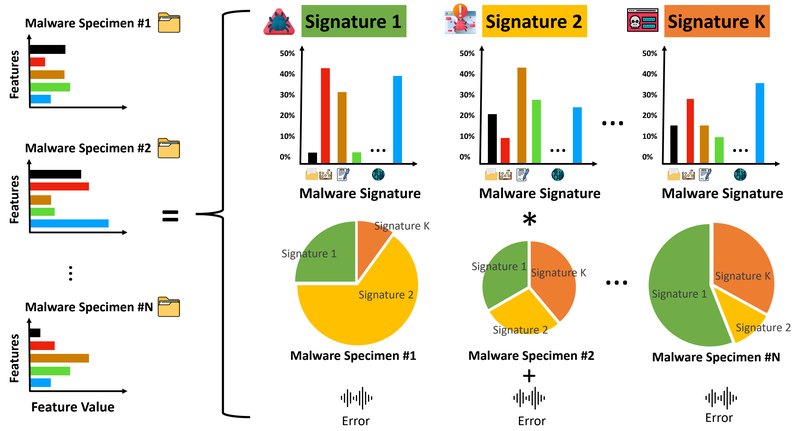
Contact: maksim@lanl.gov

Objective

- Malware is one of the most dangerous and costly cyber threats to national security.
- Classifying a malware sample into a family aids in understanding the behavior of the malware, which is helpful for estimating the severity of the threat and developing mitigation strategies.
- **Prior malware defense solutions do not sufficiently address a number of real-world challenges** slowing down the adoption of ML-based solutions against malware threats depside the cost savings:
  - Considering the cost associated with labeled malware
  - Using supervised solutions that poorly generalize to new malware
  - Detecting both rare and prominent malware families
  - Incorporating the ability to identify new/novel malware families

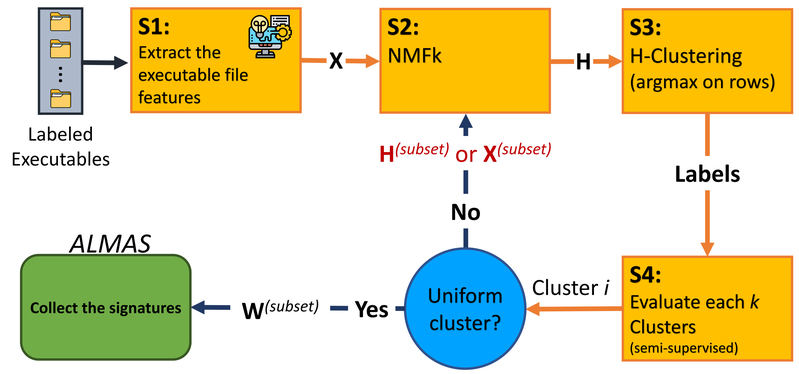
Malware-DNA: ML method that considers software analogous to the genomic DNA, malware as malicious mutations (e.g., cancer) in the software genome, and targets extraction and recognition of *accurate* mutational malware signatures.

- Using the ideas of our 2021 R&D 100 winning SmartTensors AI Platform[1,4], we introduce a **new ML method for malware family classification and novel malware family detection** that achieves state-of-the-art results while **addressing the major shortcomings in the field**.



Method

- We first build an archive of identifying latent software signatures via hierarchical factorization which includes estimation of the number of latent signals[1].

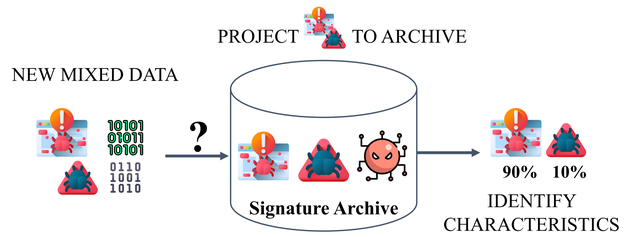


- **S1:** Extract observational features from labeled malware.
- **S2:** Non-negative factorization of the observational data **X** which gives us a factor matrix **W** (*k* columns are the latent signatures) and **H** (rows are the magnitudes of each of the *k* signatures).
- **S3:** Apply a custom clustering which assigns each of the samples to one of the *k* signature-clusters.
- **S4:** When a uniform cluster is identified, i.e. a cluster which contains specimens from the same family, we add the annotated cluster centroid to our archive of signatures.
- Otherwise, we continue with successive factorizations in a hierarchical manner to separate the mixed latent signatures.
- **New sample identification:** Project a new sample onto the archive using Non-negative Least Squares Solver (NNLS), and obtain a similarity score.

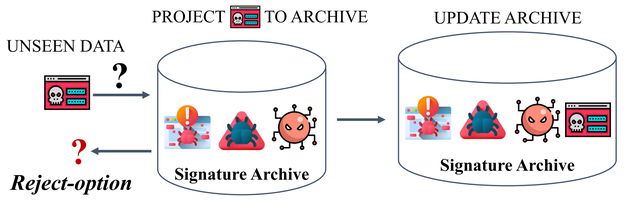
Presented at the *Malware Technical Exchange Meeting* (MTEM), Lawrence Livermore National Laboratory, California. July 25-27, 2023.

Reject-Option

- The reject-option is the ability for our ML model to be able to say both *"This is a known malware!"*, and *"I do not know what this is!"*. Based on the similarity score obtained from the NNLS projection and a **threshold *t***, we can characterize the new known specimen:

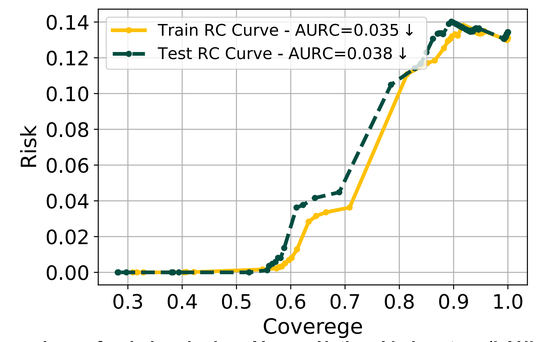


- With the reject-option, we can abstain from making a prediction when an unknown specimen is seen:



Experiments

- Using the EMBER-2018[2] dataset, we randomly sample **20k malware specimens** from families **ramnit**, **adposhel**, **emotet**, **zusy**. We select **ramnit** to represent a novel family.
- The performance of our method is reported with the *Area Under the Curve of Risk-Coverage*[3] (AURC, where lower is better), and the accuracy score.
- **We achieve a great AURC score of 0.038**: which means that at ~50% coverage we get an accuracy score of ~0.99 and correctly identify ~99% of **ramnit** as novel.



This research was funded under Los Alamos National Laboratory (LANL), Laboratory Directed Research and Development Center (LDRD-CR), managed by Information Science and Technology Institute (ISTI), Rapid Response Grant XX8R.

REFERENCES

This poster has been designed using resources from Flaticon.com

[1] Boian Alexandrov, Velimir Vesselinov, and Kim Orskov Rasmussen. SmartTensors Unsupervised AI platform for Big-Data Analytics. Technical Report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2021. LA-UR-21-25064.
[2] Anderson, Hyrum S., and Phil Roth. "Ember: an open dataset for training static PE malware machine learning models." arXiv preprint arXiv:1804.04637 (2018).
[3] Ding, Y., Liu, J., Xiong, J., & Shi, Y. (2020). Revisiting the evaluation of uncertainty estimation and its application to explore model complexity- uncertainty trade-off. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 4-5). [8] Rantos, K., A. Spyros, A. Papanikolaou, A. Kritsas, C. Ilioudis, and V. Katos,
[4] Eren, M.E., Bhattarai, M., Nicholas, C., Rasmussen K., and Alexandrov, B. (2023), Data Identification and Classification Method, Apparatus, and System, US, Provisional Patent 63/472,188.