Malware Technical Exchange Meeting

Trends in Malware

Prajna Bhandary, Randolph Wiredu-Aidoo, Varshitha Palakurthi, Manimadhuri Edara and Charles Nicholas
University of Maryland Baltimore County

UMBC

DREAM LAB

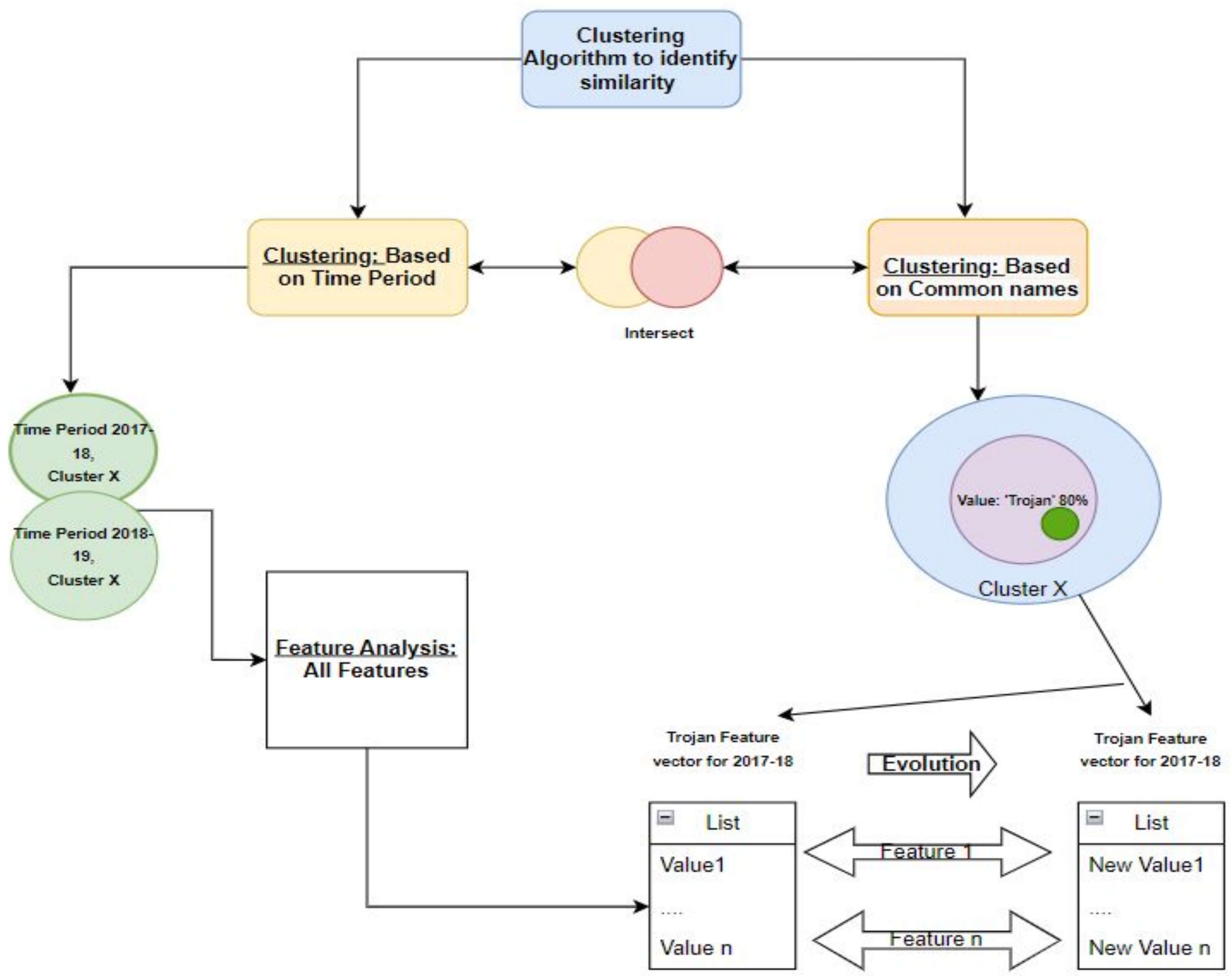
Objectives

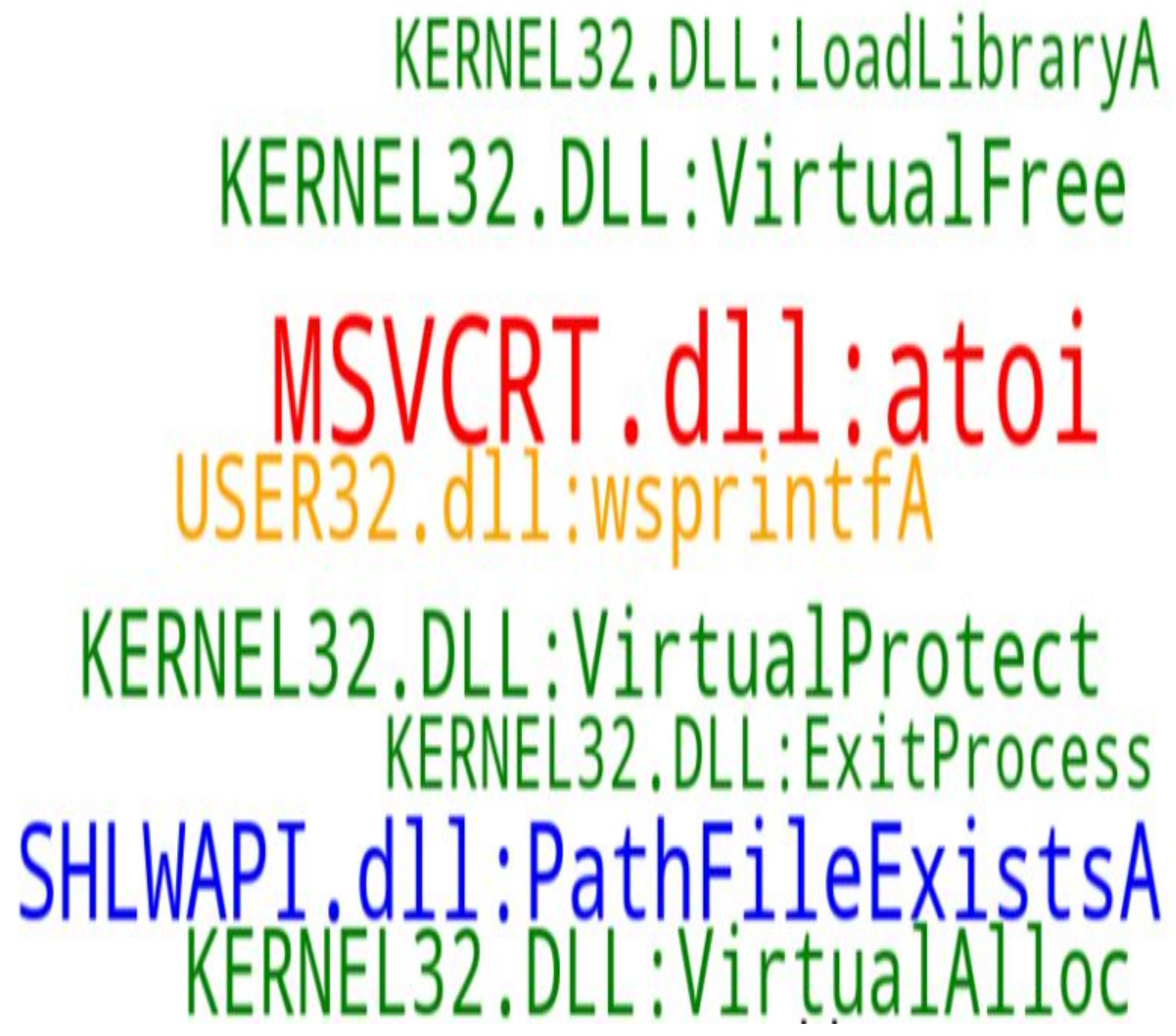
- Understanding trends in malware is necessary to identify variants in malware.
- Malware analysts need to be aware of the extent of updates of a malware families.
- Our work classifies samples in the SOREL-20M[1] dataset using clustering and automatically generate a timeline showing evolution of malware

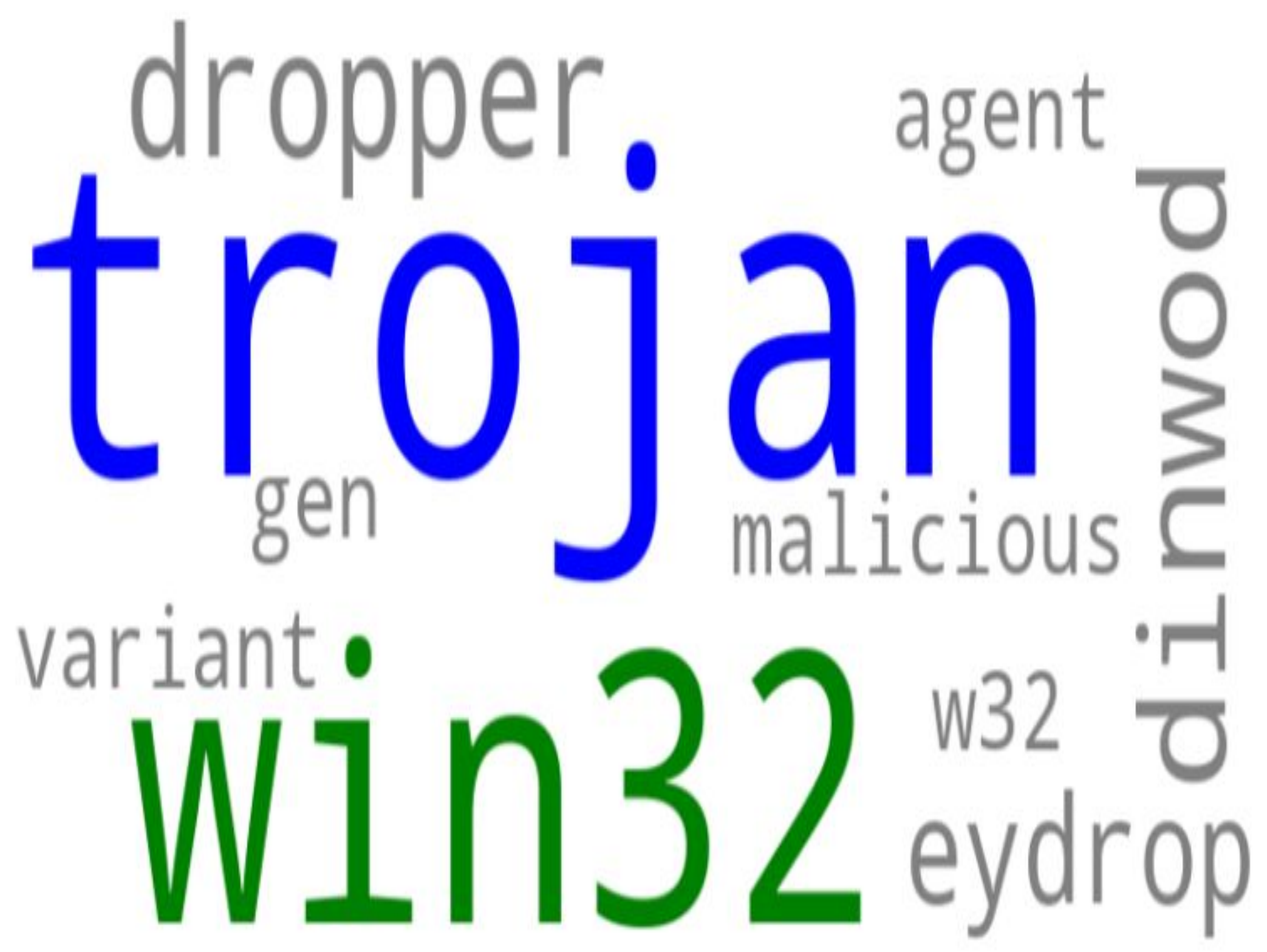
Methodology

- Parse SOREL-20M dataset
  - AV tokens are not standardized
    - Many variations of the same word e.g. "trojan"/"trjen"/"tr"
- Apply binary vectorization - more comprehensible
- Density-based clustering (DBC) with Euclidean distance in 3 phases:
  - Phase 1: Data Cleaning: Identify the features based on potential information gain
  - Phase 2: Cluster based on similarity
  - Phase 3: Cluster based on Time Period
- Identify common feature values of 80%+ homogeneity within cluster convergences
- Organize the results

Clustering Algorithm


Clustering Algorithm to identify similarity
Clustering: Based on Time Period
Intersect
Clustering: Based on Common names
Time Period 2017-18, Cluster X
Time Period 2018-19, Cluster X
Value: 'Trojan' 80%
Cluster X
Feature Analysis: All Features
Trojan Feature vector for 2017-18
Evolution
Trojan Feature vector for 2017-18
List
Value1
....
Value n
Feature 1
Feature n
List
New Value1
....
New Value n





Results

- ~70 malware clusters were found
- Grouped into three major time periods:
  - Late 2016 to mid 2017
  - Mid 2017 to the early 2019
  - Early 2019 to late 2019
- Most clusters have emergent names of 'trojan', 'worm','win32'
- Clustered at 100% homogeneity-
  - grouped trojan and its variants - 'Eldorado', 'behaveslike','trojandropper'
  - Function names - "MSVCRT.dll: 'atoi'", "KERNEL32.dll: 'virtualFree','ExitProcess', 'VirtualProjectA', 'LoadLibraryA', 'VirtualAlloc', ''GetProcAddress'","SHLWAPI.dll: PathFileExists", "USER.dll:wsprintfA"

Future Work

- Evaluate and compare other clustering algorithms and distance metrics and feature engineering techniques
- Improve tokenization algorithms
- Try using co-clustering algorithm

References

1. Richard Harang and Ethan M. Rudd. Sorel-20m: A large scale benchmark dataset for malicious PE detection, 2020. arXiv.org

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