Malware Identification

A Machine Learning Approach

Capstone Presentation

Presented by: Alex Eldredge

About Me

- Love skiing, hiking, and everything outdoors!
- Previous job working in quality and company analytics
- Led a team of 21 analysts with multi-department focuses

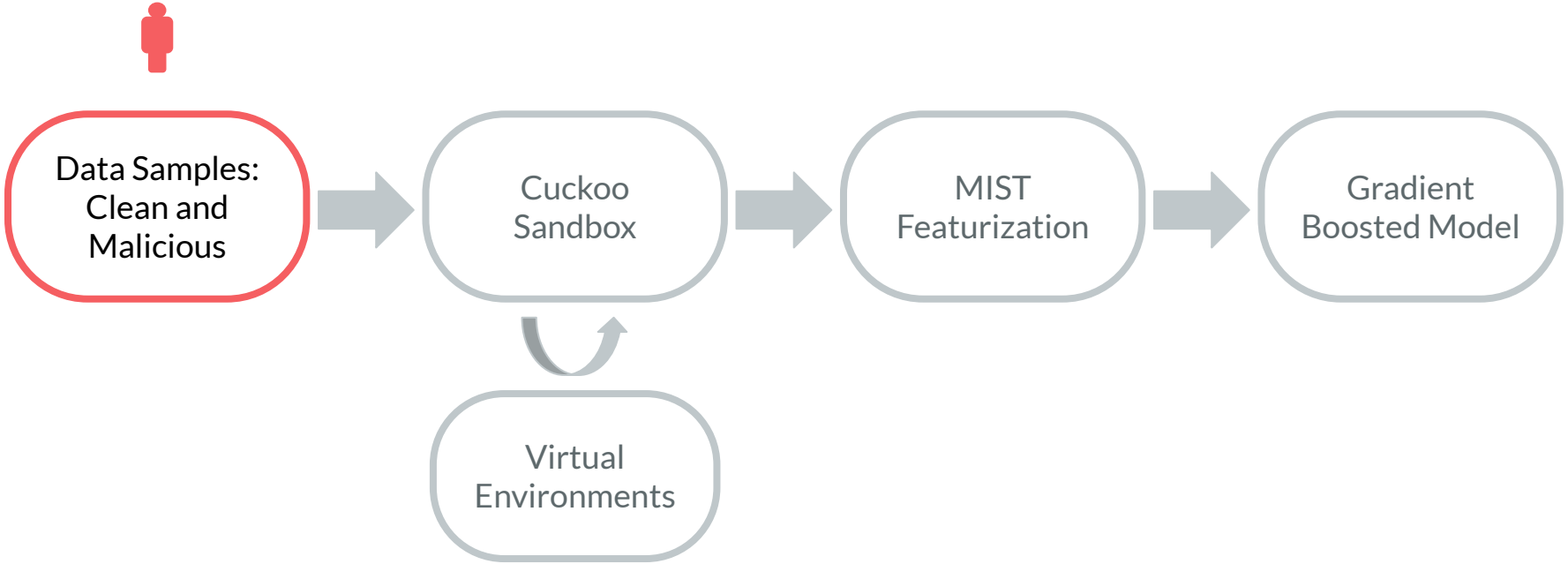
Motivation and Background

- Malware is a concern for:
  - Businesses
  - Governments
  - Individuals
- Machine learning features:
  - Static
  - Dynamic

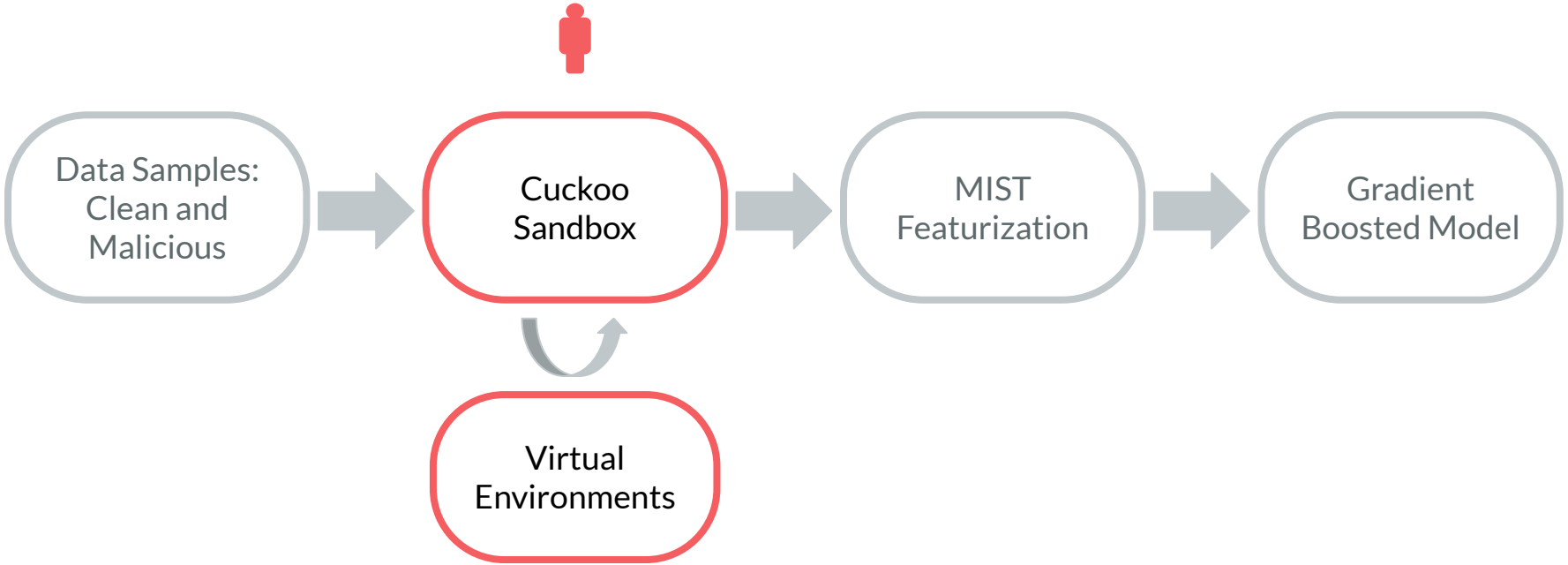
Objective:

Identify and classify malicious software using static and dynamic feature-based machine learning.

Data Flow


Data Samples: Clean and Malicious
Cuckoo Sandbox
Virtual Environments
MIST Featurization
Gradient Boosted Model

Data Flow


Data Samples: Clean and Malicious
Cuckoo Sandbox
Virtual Environments
MIST Featurization
Gradient Boosted Model

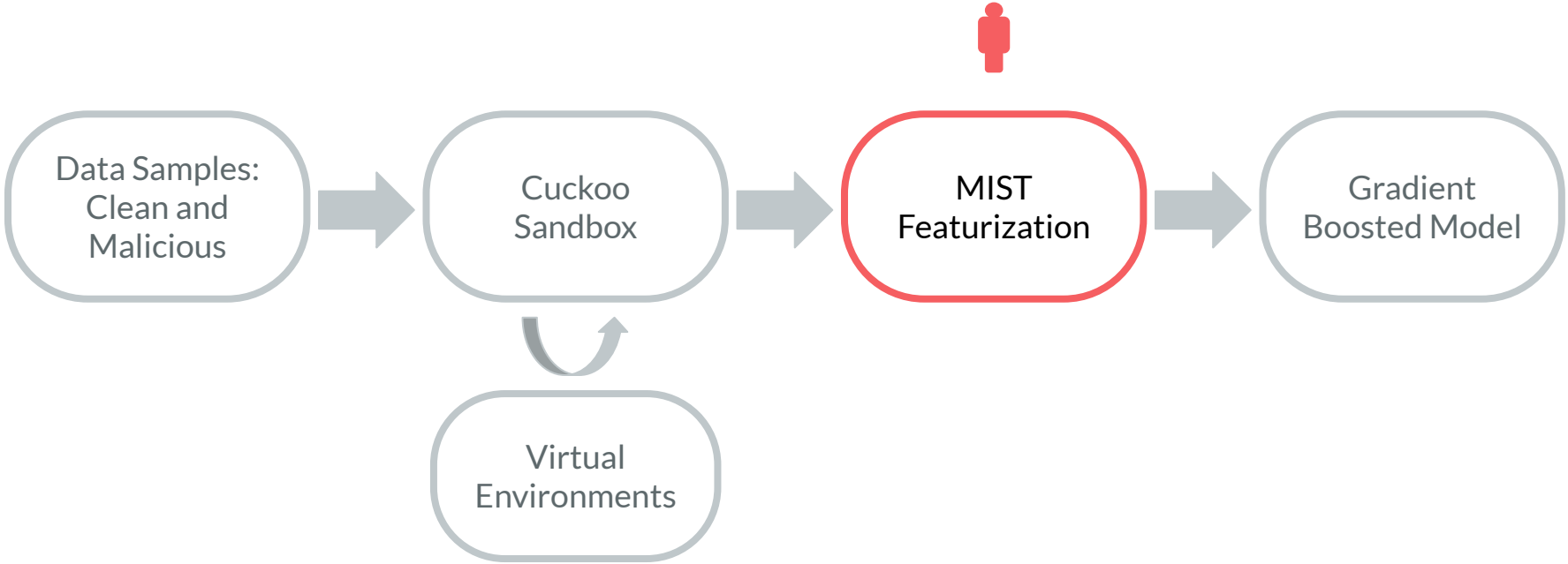
Data Source and Intake

```
<load_dll filename="C:\WINDOWS\system32\kernel32.dll" successful="1"
 address="#7C800000" end_address="#7C908000" size="1081344"
 filename_hash="c88d57cc99f75cd928b47b6e444231f26670138f"/>
```

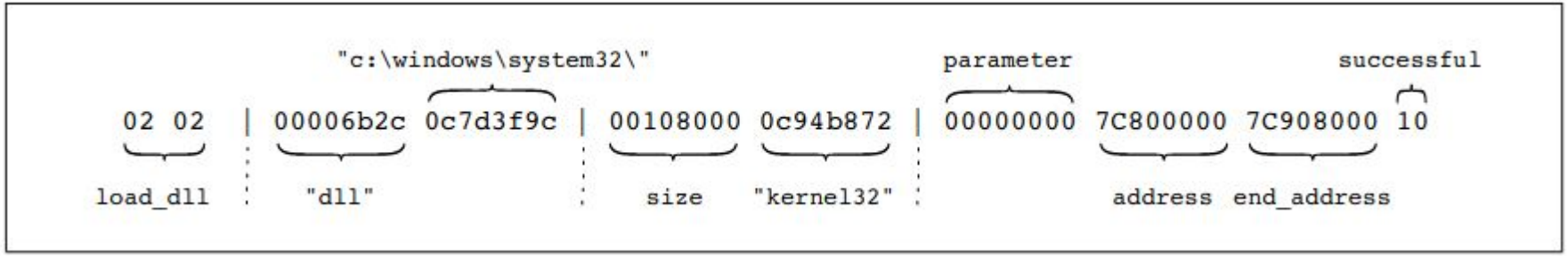
XML representation returned from Cuckoo Sandbox

Source

Data Flow


Data Samples: Clean and Malicious
Cuckoo Sandbox
Virtual Environments
MIST Featurization
Gradient Boosted Model

Data Source and Intake



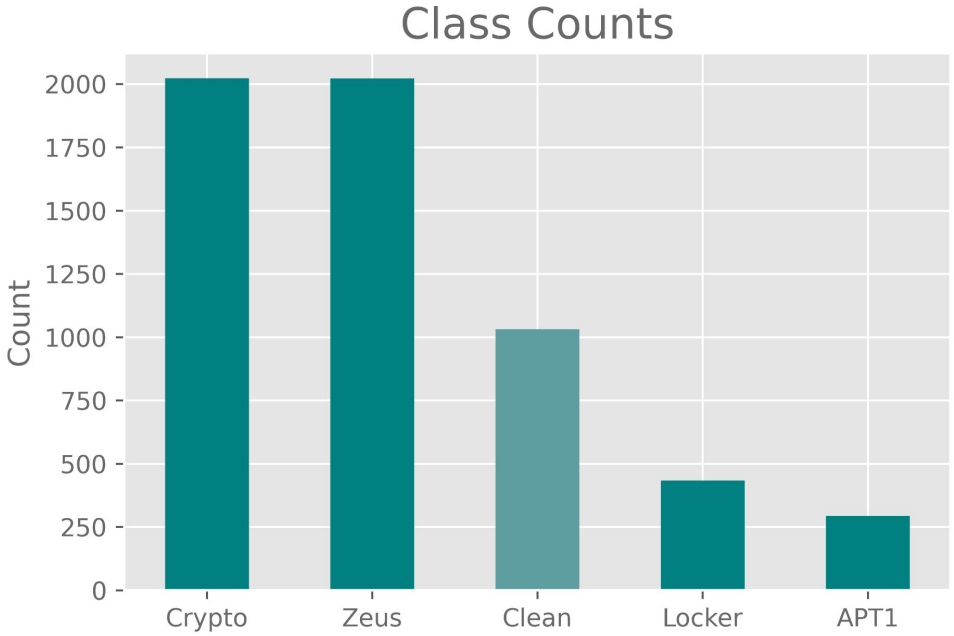
MIST representation after processing

Source

Data Source and Intake

	entityId	file_drop	pe_sec_entropy	pe_sec_name	reg_access	sig_packer_entropy	pe_sec_character	str	pe_imports
0	1812	00000000 8d777f38	f7ddd489 5d425c2a f2e47402 6915b647 bb81cb21	df5bcfa6 916e7571 e8997399 b7b66a05 1f9f1073	8e05d45e 8eb58dd5 8eb58dd5 ee4d740b	3b777e80	9271b28c 9271b28c 2f0acaf7 ad8225f3 2f0acaf7	df5bcfa6 20eef30f ec73657d e4787bb1 1f9f1073 4...	a882be8e e8a591ae fa3edd01 bf1ade9a afb0f3da f...
1	2066	0	9b511809 d607823d f7ddd489 1ae67b26 f7ddd489 b...	c1336794 8d777f38 f71487b7 c3389ed2 59044e4f e...	efc50fc9 1fe39017 1fe39017 df16f4d5	0	9271b28c ad8225f3 ce353fe8 ad8225f3 ce353fe8 9...	abfc75e1 c3389ed2 e8997399 5724939f 42276272 9...	48e0e980 f90dc819 6fc39fa6 819b5377 076736b7 4...

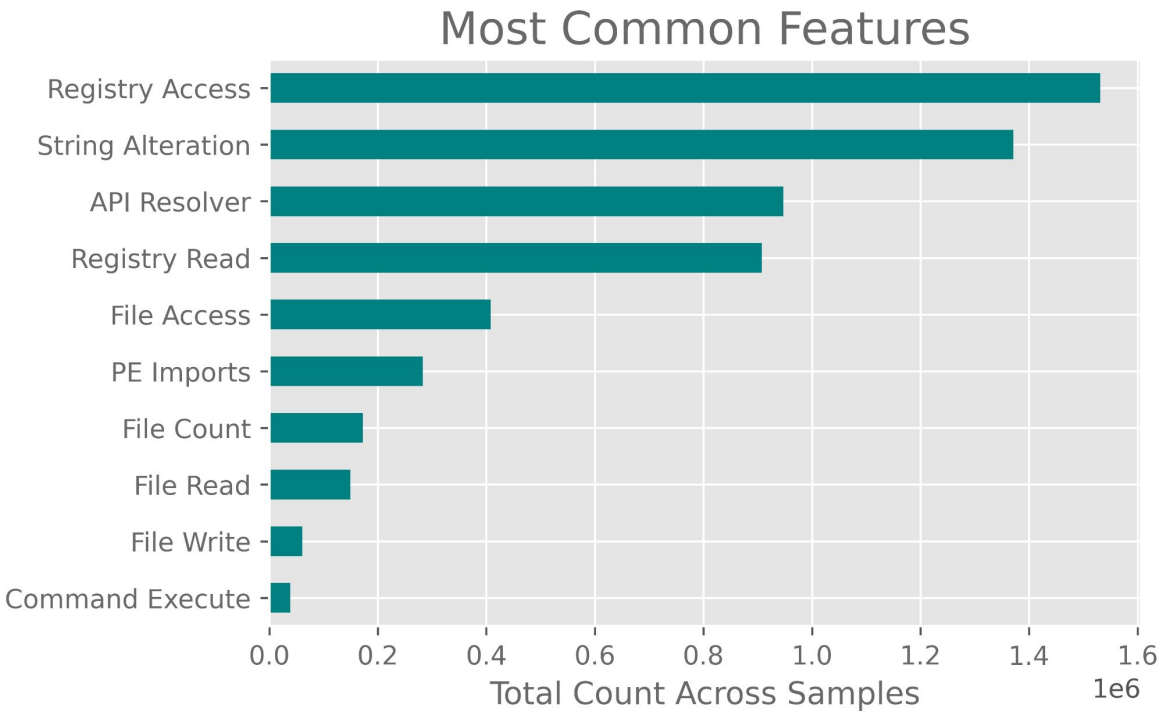
Data Context



Initial data observations:

- 5 Target classes
- 5792 samples
- Unbalanced Classes

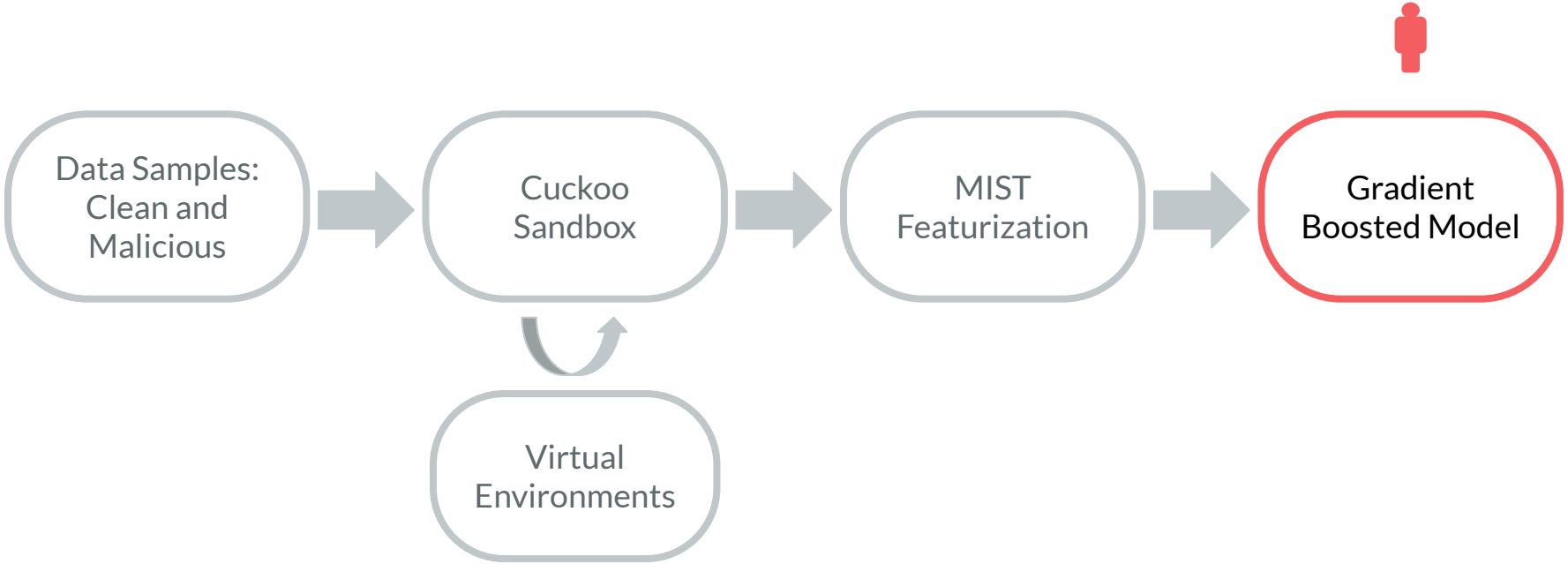
Data Context



Additional observations:

- 139 features
- Skewed feature actions and properties
- Registry Access had high feature importance

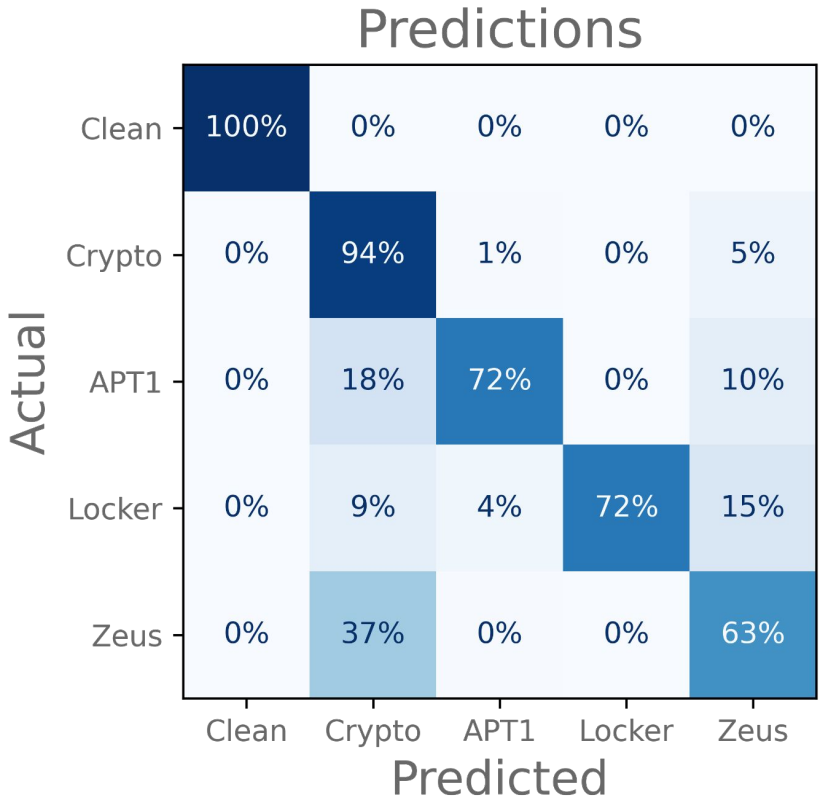
Data Flow


Data Samples: Clean and Malicious
Cuckoo Sandbox
Virtual Environments
MIST Featurization
Gradient Boosted Model

Results

Gradient Boosted Classifier:

- Best overall accuracy score - 81%



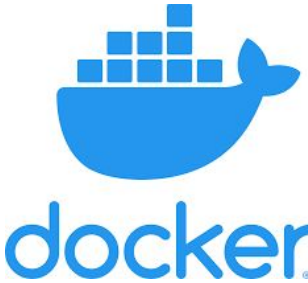
Conclusions

- Performs well identifying clean and malicious software
- Classifies current categories well

Questions?

Alexander Eldredge

eldredge.alexander@gmail.com
linkedin.com/in/alexander-eldredge/
github.com/Aeldredge



Sources

1. https://marcoramilli.com/2016/12/16/malware-training-sets-a-machine-learning-dataset-for-everyone/
2. http://www.mlsec.org/malheur/docs/mist-tr.pdf
3. https://cuckoosandbox.org/

Additional context and sourcing in README.md

Appendix