# Machine Learning Frameworks and Tools for Cyber Security in a Closed Network

Customer: Ronald Dostie (Progeny Systems)
Subject-Matter Expert: Scott Lewis
Giri Apurada, Frank McKee, Ben Nikolich, Kathryn Zurowski

## Purpose

One of the most challenging problems in cybersecurity is creating tools that:

- Detects malicious activity within a network
- Alerts users of possible intrusions while it flags those events
- Creates a baseline of anomaly detection

This project addresses this challenge by integrating machine learning frameworks into the intrusion detection process. Machine Learning provides a unique opportunity to create adaptive algorithms and more sophisticated pattern recognition to automated the detection process.

## Methodology

The team selected four models to test and refine:

- Text Clustering
- Deep Learning
- Logistic Regression
- K-means Clustering

These models were tested on a dataset that most closely aligned with the project's goals. This dataset came from a red team exercise conducted by the United States Military Academy at West Point and the National Security Administration. The team relied partly on GMU's Argo Cluster (Figure 1) for heavier computations in the second half of the project.

Figure 1

## Results

The team has selected two models as recommendations for use based on testing: **Text Clustering**, **Logistic Regression**. In addition to these models, the team recommends utilizing multiple framework in series to maximize the effectiveness of each.

Below are the results for all four models chosen for testing. The following images are graphical representations of each model:

### Text Clustering

Effectively classified large amounts of traffic as legitimate, but had a **very high false positive** rate for identifying malicious traffic.

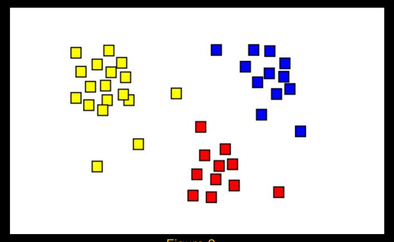

Figure 2

### Logistic Regression

Shows some promise, as the model fit score is **0.8125 out of 1.0** for a Perfect Fit Score. This is far from conclusive, as further testing will be required to improve this score.

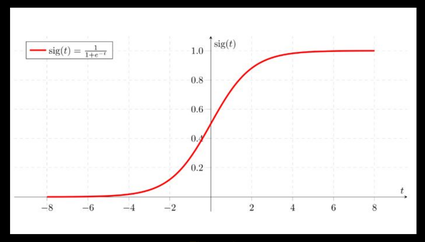


Figure 3

### Deep Learning

The team was **unable to fully deploy** a Deep Learning Autoencoder Model due to heavy reliance on large volumes of robust data, increased execution time, and other reasons.

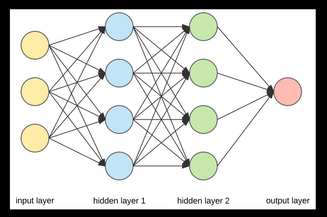


Figure 4

### K-Means Clustering

The model placed the data points into one of two groups ("normal" or "malicious"), but had a **high rate or misclassification**, however, with more time and refinement this model could be used in the future.

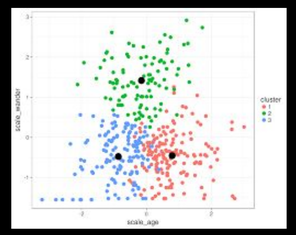


Figure 5

## Conclusion

This research effort has shown the potential of deploying Machine Learning techniques in Closed Networks for Cyber Security.

Our goal was to determine which frameworks would be suitable for use in cyber security implementations. We have identified two frameworks that show potential:

- **Text Clustering**
- **Logistic Regression**

We believe that with continued testing, they can be effective in detecting anomalous events in a closed network. Further research should continue to improve these models and integrate them into commercial intrusion detection products.

## Acknowledgements

We would like to thank Progeny Systems for assistance in this project. We appreciate the time and effort from our SME Scott Lewis. Special thanks to Dr. Peggy Brouse, Gino Manzo, and Rock Sabetto for helping provide this opportunity. This project utilized computing resources provided by the George Mason University Office of Research Computing.

## References

Figure 1 - http://wiki.orc.gmu.edu/index.php/About_ARGO
Figure 2 - https://en.wikipedia.org/wiki/Cluster_analysis
Figure 3 - https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
Figure 4 - https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6
Figure 5 - https://rpubs.com/cyobero/k-means

Date: May 2, 2019