**Week Three Status**

**9/30/2022**

**TEAM RAD**

**Real-time Anomaly Detection**

**Gavin Cutchin -** Working on machine learning courses through CodeAcademy and SageMaker course through LinkedIn Learning

**Gabriella Rivera -** Currently working on machine learning courses on udemy and LinkedIn Learning.

**Jared Dunn -** Doing SageMaker course and reading about machine learning

**SageMaker Tutorial:**

<https://www.linkedin.com/learning/learning-amazon-sagemaker/machine-learning-with-amazon-sagemaker?autoplay=true&u=76209018>

**BETH Dataset**

**Description:** A 8 million point synthetic data set designed to set uncertainty and robustness benchmarking for machine learning. Uses Sci-Kit to process data.

**Link:** <https://www.kaggle.com/datasets/katehighnam/beth-dataset>

**Problem Statement:** Without benchmark machine learning our anomaly detection program would be unable to find and report incoming logs as volatile or evil.

**Broad Statement:** We plan to use this crafted dataset that is made specifically for machine learning training to benchmark our anomaly detection program so that we can successfully find and report volatile or evil logs.

**Pros:**

* Relatively up to date activity and attacks, also constantly updated. (2021)
* Each individual host contains up to one attack with benign activity which is best for analysis and research tasks.
* Split into training(60%), validation(20%), and testing(20%).

**Cons:**

* Only has data for 5 hours of noncontiguous samples.
  + Poses a problem for attacks with extended time frames.
* Synthetic data using honeypots may not be the most realistic and could cause inconsistencies for more intense attacks.

**Required Citation for Use:**

**@article{highnam2021bethdata,**

**title={BETH Dataset: Real Cybersecurity Data for Anomaly Detection Research},**

**author={Highnam, Kate and Arulkumaran, Kai and Hanif, Zachary and Jennings, Nicholas R.},**

**journal={ICML Workshop on Uncertainty and Robustness in Deep Learning},**

**year={2021}**

**}**

**ASNM Datasets: A COLLECTION OF NETWORK TRAFFIC DATA FOR TESTING OF ADVERSARIAL CLASSIFIERS AND NETWORK INTRUSION DETECTORS  
Description:** “ASNM-NPBO Dataset - contains non-payload-based obfuscation techniques applied onto malicious and some of legitimate traffic. It was created in 2015.

ASNM-TUN Dataset - contains tunneling obfuscation techniques applied to malicious traffic. It was created in 2014.

ASNM-CDX-2009 Dataset - contains ASNM features extracted from tcpdumps of CDX 2009 dataset. It misses few newer ASNM features. It was created in 2013.” - ASNM

**Problem Statement of Dataset:** “ASNM datasets can be used for machine learning based Network Behavioral Anomaly Detection or analysis of network traffic characteristics based on the labels indicating presence and/or type of malicious/legitimate communication.” -ASNM

**Problem Statement Comparison:** This description directly matches our problem statement for our overall project.

**License:** Creative Commons 4.0

**Link:** <https://ieee-dataport.org/open-access/asnm-datasets-collection-network-traffic-data-testing-adversarial-classifiers-and>

**ICSE-CIC-IDS2018 on AWS**

**Description:** The dataset includes: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and infiltration of the network from inside. The dataset includes the captures network traffic and system logs of each machine, along with 80 features extracted from the captured traffic using CICFlowMeter-V3.

**Link:** <https://www.unb.ca/cic/datasets/ids-2018.html>

**Problem Statement:** How are we going to sort network traffic? We will be analyzing the all sort of incoming traffic, what way and how should we label them?

**Broad Statement:** It would help us profile network traffic, and we can take inspiration from the way they label and they have sorted their data.

**Pros:**

* This dataset has labeled their network traffic, even creating classes for them to placed into a dataset
* Has a small data table that lists out what type of attack took place, duration of attack, tools used, attacker’s system, and victim’s system
* Also has a data sheet for 2017 with the same tool, CIC-IDS

**Cons:**

* Does not contain data activity as the BETH dataset

**Required Citations for Use:**

You may redistribute, republish, and mirror the CSE-CIC-IDS2018 dataset in any form. However, any use or redistribution of the data must include a citation to the CSE-CIC-IDS2018 dataset and [https://registry.opendata.aws/cse-cic-ids2018/](https://registry.opendata.aws/cse-cic-ids2018).