Make Buzzwords Great Again

Data Mining, "Big Data", and Machine Learning for Security and DFIR

Slides: https://github.com/Dioberne/ConferenceTalks/tree/master/BloomCon/2019

Goal: Give you an understanding of the subject matter that will allow you to cut through vendor nonsense and get stuff done

curl http://169.254.169.254/latest/meta-data/hostname

Btice

- BU Alumnus
  - Digital Forensics
  - Anthropology
- Geeks Hard
  - python -c "import antigravity"
  - emerge --sync
- Now Blue Team
  - “Full-stack” == “Doing the job of 3-7 people”
  - “Purple-ish-team” == “Know thyself, know thy enemy. A thousand battles, a thousand victories” --Sun Tzu

Tell Them What You Are Going To Tell Them

- Formalities ← You are here
- What is N?
  - Data mining
  - Big Data
  - Machine Learning
- Demos!
  - Data mining with ELK
  - Looking at data with Jupyter and Python (“Exploratory data analysis”)
  - Anomaly Detection
  - DNS Traffic Classification
  - Hunting lateral movement with graph / link analysis
- Questions?
  - Also links

What is Data Mining?

- Almost useful definition: “The practice of examining large databases in order to generate new information” --Google
- A process for looking through large amount of data
  - Ingest
  - Standardize
  - Enrich (Add new fields)
  - Measure Statistically (Top N, average/min/max/stddev, frequency etc.)
  - Vizualize
  - Search, filter
  - (Optional) Use Algorithms on
    - Clustering
    - Classification
    - Anomaly Detection
    - Shopping cart analysis
    - And many more...

Data Mining Tools-to-Know

- Things I will show today:
  - RITA (https://github.com/activecm/rita)
    - Does statistical analysis of network logs to aid hunting
  - NetworkX (https://networkx.github.io/)
    - Graph analysis in Python
- Other data mining tools that don't overlap with Big Data…
  - Logon Tracer (https://github.com/JPCERTCC/LogonTracer)
    - Hunt for lateral movement using graph theory and Windows Event Logs
  - …

What is “Big Data”?

- Not useful definition: “Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.” --Google
- Data that can’t be stored or processed by a single machine
- Ben’s rule of thumb for when to use big data tools: The data is too big to open in Excel
  - “big data” with a little “b”
- Big Data concerns itself with...
  - Amount of data (Cluster required)
  - Speed of data (24x7 streaming)
  - Diversity of data (Many input formats)

Big Data Tools-to-Know

- Things I will show today:
  - ELK (https://www.elastic.co/elk-stack)
    - Elasticsearch (Storage, search, filter)
    - Logstash (Ingest, normalization, enrichment)
    - Kibana (Visualize, basic stats)
- Things to look up later:
  - Spark (https://spark.apache.org/)
    - Lets you run code on a cluster (and on clustered data)
    - Supports HDFS natively but can be used on Elasticsearch with a plugin
      - https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
      - https://www.elastic.co/products/hadoop
    - Bundled along with ELK and ES-Hadoop in Hunting ELK
      - https://github.com/Cyb3rWard0g/HELK
  - Kafka (https://kafka.apache.org/)
    - Distributed data plumbing (boring but necessary)

What is Machine Learning?

- Actually useful definition: “Machine learning is a subfield of computer science and statistics that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions” --https://github.com/CICCIOSGAMINO/machine_learnig
- Statistics + Algorithms (that learn thresholds for decisions, multipliers etc.)
- Machine learning only learns numbers and only operates on numbers
  - Using ML on something that isn’t a number requires you to turn it into a number using (“Featurization”)
- I believe we should leverage ML algorithms as grey boxes
  - Black boxes give the power to sell us “magic” to vendors
  - White box level of understanding is overkill for getting work done
  - Goldilocks: Know your inputs, outputs, and gotchas then move on

Ben’s Machine Learning Rules of Thumb

- General Principles Useful as a Guide:
  - Good candidates for machine learning are tasks that humans complete near instantly (after loading the data into their head)
    - Can a machine learning algorithm make decisions with the “intuition, expertise, and tribal knowledge of Tier 3 security analysts”?
  - If a human can’t reliably solve the problem with the data presented a machine learning model won’t be able to either
      - Example: Age from a photograph
      - Example: RITA Beacon Analysis
      - And if it can then it isn’t solving the problem the same way a human would
        - Example: Byte histogram in ML-AV
  - Algorithms should be the easiest part of machine learning
    - Using an off-the-shelf algorithm is just 4 lines of Python
    - Expect most of your brain-power to go to featurization
    - Expect most of your code to be boilerplate and data wrangling

Machine Learning Tools-to-Know

- Things I will show today:
  - Sklearn (https://scikit-learn.org/)
    - Usually the only library you need, has one of almost everything
  - Jupyter (https://jupyter.org/)
    - Lets you run Python code and see the output in a web browser
- Things to look up later:
  - Spark-ML (https://spark.apache.org/docs/latest/ml-guide.html)
    - Lets you run common ML algorithms on a Spark cluster
  - Robust Cut Random Forests (https://github.com/kLabUM/rrcf)
    - Like an Isolation Forest but for streams
  - Python Outlier Detection (https://github.com/yzhao062/pyod)
    - More anomaly detection algorithms than you will ever need
    - Includes PCA based anomaly detection
  - Ember Dataset (https://github.com/endgameinc/ember/blob/master/ember/features.py)
    - Malware classification dataset. Contains example of featurization

Demo: Data mining with ELK

- What: Network flow log review
- Data: Bro Conn log
  - https://github.com/bro/bro
- Tool: ELK stack
  - https://github.com/opendistro-for-elasticsearch
  - https://github.com/Security-Onion-Solutions/security-onion

Enrichment With Logstash

```
geoip {
    source => "id_orig_h"
    target => "[@meta][geoip_orig]"
}
geoip {
    source => "id_resp_h"
    target => "[@meta][geoip_resp]"
}
```

Demo: Looking at data with Jupyter and Python

- What: “Exploratory data analysis” (Looking at data)
- Data: RITA Beacon analysis report
  - https://github.com/activecm/rita
- Tool: Jupyter + Python + Sklearn
- Algorithm(s): PCA, K-means

Demo: GeoIP Anomaly Detection

- What: Detect network traffic to unusual locations
- Data: Bro Conn Log
- Tool: Jupyter + Python + Sklearn
- Algorithm(s): DBSCAN

Demo: DNS Traffic Classification

- What: Classify domain names as malware or benign
- Data: Bro DNS log
- Tool: Jupyter + Python + Sklearn
- Algorithm(s): Random Forest

Pre-demo: Can humans classify DNS?

A. r5---sn-8xgp1vo-p5qe.googlevideo.com
B. tfydfxlideir.info
C. wrzkjtqumhygulergpttjzojmeduwn.sandbox.alphasoc.xyz
D. 58701c8b2469e1404298a38dfbcfdb03b69f4a6d.malware.hash.cymru.com

Demo: Hunting lateral movement with graphs

- What: Identify hosts that may be involved in a lateral movement chain
- Data: Bro conn log
- Tool: Jupyter + Python + NetworkX
- Algorithm(s): PageRank

Tell them what you told them

- Data Mining is a process for looking through large amount of data
- Big Data is data that can’t be stored or processed by a single machine
- Machine learning is the combination of statistics and algorithms for learning from example data
- The barrier to entry in machine learning is pretty low but you still want to know your inputs, outputs, and gotchas
- We can leverage these things to get more stuff done
- Taking a peak under the hood can cut through vendor malarkey

Questions?

- Concerns?
  - Dreams?
    - Fears?
      - Aspirations?
        - Snide remarks?

Linkapalooza

- “Using isolation forests to detect bot matches in dota2”
  - https://towardsdatascience.com/detecting-bot-matches-in-dota-2-using-isolation-forests-a17c34f60923
  - https://github.com/zhilingc/anomdota
- Bro Analysis Tool
  - https://github.com/SuperCowPowers/bat
- “In-Depth Data Stacking”
  - https://www.fireeye.com/blog/threat-research/2012/11/indepth-data-stacking.html
- “Building Machine Learning Models for the SOC”
  - https://www.fireeye.com/blog/threat-research/2018/06/build-machine-learning-models-for-the-soc.html
- “Your Model Isn’t Special”
  - https://github.com/endgameinc/youarespecial
- “Getting Started With Machine Learning for Incident Detection”
  - https://www.youtube.com/watch?v=2FvP7nwb2UE
  - https://github.com/DavidJBianco/Clearcut