Sandia National Laboratories

Robohunter: a Machine Learning Framework for Batch Analytics

Houlton McGuinn

Robohunter Team: Jerry Cruz, Todd Farell, Mary Rose Sena, and Mike Smith

Goal: Provide few, high-confidence alerts to analysts with context for why events were suspicious.

A Model Agnostic Framework:

Robohunter provides a framework through which different machine-learning (ML) models can be leveraged to analyze batch collections of network data. Rather than definitively classifying traffic, Robohunter acts as a stepping stone for follow-up manual analysis of suspicious events. Data pulled from Elasticsearch is preprocessed through transformers before being passed to the ML model.

Augmenting Network Streaming Analytics:

- Streaming analytics use heuristics to identify weak indicators of compromise in network traffic.
    - Lots of false positives can overwhelm analyst.
- Batch analytics build on streaming analytics by aggregating weak-indicators for higher confidence alerts.
    - Improves as data changes.
    - Support question-driven analysis.
    - Can leverage a number of ML models and statistical approaches.

Flexible Algorithms:

Designed to be extensible, Robohunter provides support for both labeled and unlabeled ML models.

Ads Whitelisting & Anomaly Detection:

- Look for anomalies in noisy network traffic.
- Model is trained on unlabeled data. Then, data is hunted over and each document scored from 0 to 1.
- Algorithms can reduce the number of ads flagged as suspicious and can identify minified JavaScript that has potential to be malicious.
- Model provides explainability for why an event was scored as suspicious, allowing an analyst to further investigate.
- Anomaly detection builds on the score by applying a clustering algorithm.
    - Data that is not part of a cluster is classified as malware.

Pre-Trained Models:

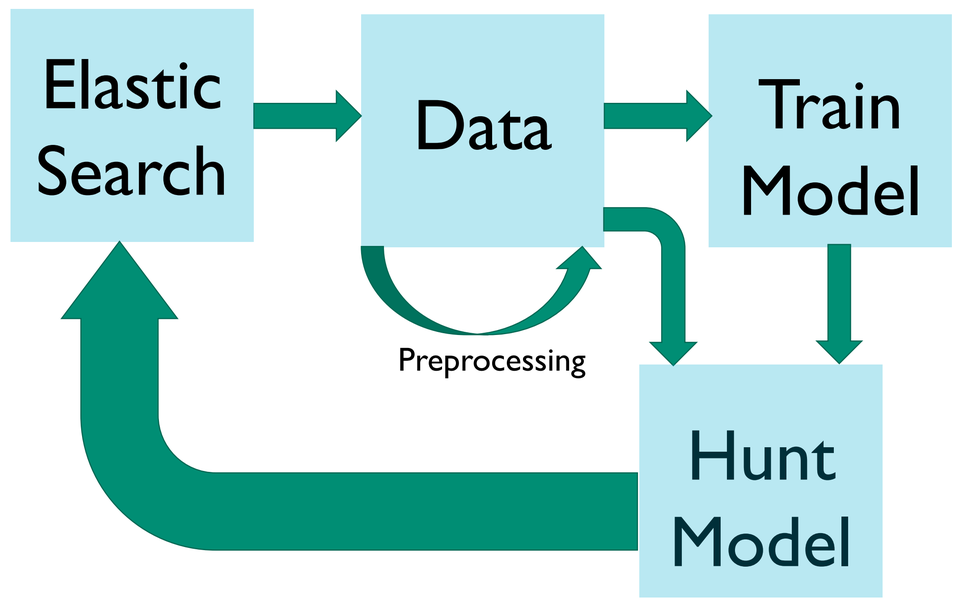
- Robohunter provides support for pre-trained models as an alternative to first training a model.

Supervised Learning:

- Model aggregates traffic metadata (domain, time-of-day, location, etc.) and then applies a supervised learning algorithm to identify outliers.
- Supervised learning aggregates a number of weak-indicators on labeled data to produce higher confidence alerts.

Elastic-Pandas Integration:

- Interfaces with standalone Elastic-panda submodule to create pandas dataframes from Elasticsearch data.
- Results are uploaded to Elasticsearch for visualization in Kibana.



Impact:

- A framework for batch analytics of network traffic reduces development time, allowing more developer time to be spent on developing new analytics.
- Aggregating weak-indicators allows higher confidence alerts to be produced.
- By providing the context of why something is suspicious, more power is given to the analyst.
- Elasticsearch stack allows easy access to raw data and alerts in Kibana.
- Allows incorporation of novel models.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. SAND2020-7667 D

Sandia National Laboratories