Beyond Operational Intelligence

- Splunk Advanced Analytics

Anthony Tellez | Staff Data Scientist - Machine Learning & AI, Splunk
Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2018 Splunk Inc. All rights reserved.
Intro - Anthony Tellez
CISSP, CEH, CDNA, Sec+

| where _time@Splunk > 4y
Previous: U.S. Gov Contractor, Geospatial Analyst
Specializations
  - Cryptography
  - Information Security – Red Team
Data Scientist
  - Security & Fraud Analytics
  - Data Visualization
Responsible for the relationship between emerging technologies and field organization
  - Product Development
https://github.com/anthonygtellez/TA-Suricata
Fact: Spends 80% of the year on a plane traveling globally.
Agenda
Machine Learning isn’t Magic

- Splunk + Phantom Capabilities
- Machine Learning Toolkit Capabilities & Extensions
- Incident Response Hierarchy of needs as a guide to security analytics
- Q & A
Continuous Data Ingest at Scale

**Industrial Data**
- SCADA, AMI, Meter Reads

**Native Inputs**
- TCP, UDP, Logs, Scripts, Wire, Mobile

**Modular Inputs**
- MQTT, AMQP, COAP, REST, JMS

**HTTP Event Collector**
- TokenAuthenticated Events

**Technology Partnerships**
- Kepware, AWS IoT, Cisco, Palo Alto

**Real Time**

**Engineers**
- Search

**Data Analysts**
- Alert

**Security Analysts**
- Visualize

**Business Users**
- Predict
- Develop

**External Lookups/Enrichment**
- Asset Info
- Maintenance Info
- Data Stores

**OT**
- Industrial Assets

**IT**
- Consumer and Mobile Devices

**Maintenance Info**
- Asset Info

**Data Stores**
- Maintenance Info
- Data Stores
SANS Threat Hunting Maturity

Source: SANS IR & Threat Hunting Summit 2016
Sense and Respond

Every Search Can Use Machine Learning

Real Time

Search

Alert

Industrial Assets

Consumer and Mobile Devices

OT

IT

Flash lights

Email

Tickets

Phantom

Smartphones and Devices

Send an email

File a ticket

Trigger phantom playbook

Send a text
Splunk Adaptive Response
Splunk Machine Learning Toolkit
Extends Splunk with new tools and guided modeling environment

**Experiments**: Guided model building, testing, and deployment for common objectives

**Showcases**: Interactive examples for typical IT, security, business, and IoT use cases

**Algorithms**: 80+ algorithms across out the box & GitHub (supervised to unsupervised)

**ML Commands**: New SPL commands to fit, test and operationalize models

**ML-SPL API**: Extensibility to easily import any algorithm (proprietary / open source)

**Python for Scientific Computing Library**: Access to 300+ open source algorithms

**Spark MLLib**: Support large scale model training via Spark Add-on for MLTK (LAR)

Build custom analytics for any use case
MLTK Major Highlights (v3.1 thru v4.0)  
(since .conf 2017)

- **Splunk Machine Learning Toolkit Updates (3.1 thru 4.0)**
  Includes new features for the Experiment Framework, algorithms, pre-processing steps, validation options etc.

- **Python for Scientific Computing 1.3 Update**
  Updated libraries giving you access to new and modified algorithms and its parameters.

- **Splunk MLTK Connector for Apache Spark™**
  Massive model building with MLlib directly from Splunk and SPL, No Scala skills required. (Limited Availability Release)

- **GitHub MLTK Community**
  Leverage and share algorithms collaboratively with the broader MLTK community

- **Splunk MLTK Container for Tensor Flow (PS)**
  Container based neural networks, leveraging GPUs/CPU's using PS White Glove Engagement
Data Gathering and Prep

Source: CrowdFlower

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Can you name the assets you are defending?

https://github.com/swannman/ircapabilities

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Inventory
Can you name the assets you are defending?

Determining which assets in your enterprise exist is critical to **knowing what data needs to be collected**.

Assets include, but are not limited to:
- **People** – Authentication logs, Badge Access (SSH)
- **Hardware** – Network Logs: Firewall, IDS, Switches (Flow)
- **Software** – Middleware, OS, Server logs (DNS, HTTP, TLS, Fileinfo)
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Do you have visibility across your assets?
Telemetry

Do you have visibility across your assets?

- Identify metrics for the assets in your inventory, identify metrics for protocols in the environment using Suricata: Stats
- Can you accurately discern different types of devices?
  - Webserver
  - Domain Controller
  - DNS Server
  - Exchange Server
  - End User Devices
    - Mobile
    - Desktop
- Statistical and contextual analytics
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Can you detect unauthorized activity?

APT, CRIMINAL SYNDICATES, BOTS, SCRIPT KIDDIES, HACKTIVISTS

DETECTION
TELEMETRY
INVENTORY

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Detection
Can you detect unauthorized activity?

► Data Check: Signature based rules and alerting in place to classify network and authentication behavior across the enterprise.

► Analytic reporting using frequency counts (event_type=Alert)
  • False positive & noisy alerts can be adjusted accordingly
  • Effectiveness of various rules measured using statistics

<table>
<thead>
<tr>
<th>Source</th>
<th>Event Type</th>
<th>Action</th>
<th>Target</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>174.138.92.116</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>34.200.216.20</td>
<td></td>
</tr>
<tr>
<td>174.138.92.116</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>35.168.62.159</td>
<td></td>
</tr>
<tr>
<td>174.138.92.116</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>34.237.139.42</td>
<td></td>
</tr>
<tr>
<td>27.151.28.166</td>
<td>ET POLICY Executable and linking format (ELF) file download</td>
<td></td>
<td>27.151.28.166</td>
<td></td>
</tr>
<tr>
<td>93.174.91.149</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>34.237.228.159</td>
<td></td>
</tr>
<tr>
<td>180.231.214.251</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>35.168.62.159</td>
<td></td>
</tr>
<tr>
<td>58.218.213.45</td>
<td>ET POLICY Incoming Basic Auth Base64 HTTP Password detected unencrypted</td>
<td></td>
<td>35.168.62.159</td>
<td></td>
</tr>
<tr>
<td>46.183.216.142</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>34.200.216.20</td>
<td></td>
</tr>
<tr>
<td>27.151.28.166</td>
<td>ET WEB_SERVER Possible CVE-2014-3128 Elastic Search Remote Code Execution Attempt</td>
<td></td>
<td>34.237.139.42</td>
<td></td>
</tr>
<tr>
<td>172.31.7.249</td>
<td>ET POLICY Python-urllib/ Suspicious User Agent</td>
<td></td>
<td>165.227.195.135</td>
<td></td>
</tr>
<tr>
<td>46.183.216.142</td>
<td>ET SCAN DmEx Scanner User-Agent Inbound</td>
<td></td>
<td>34.237.228.159</td>
<td></td>
</tr>
<tr>
<td>58.218.213.45</td>
<td>ET POLICY Incoming Basic Auth Base64 HTTP Password detected unencrypted</td>
<td></td>
<td>34.200.216.20</td>
<td></td>
</tr>
<tr>
<td>58.218.213.45</td>
<td>ET POLICY Incoming Basic Auth Base64 HTTP Password detected unencrypted</td>
<td></td>
<td>34.237.139.42</td>
<td></td>
</tr>
</tbody>
</table>
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Can you accurately classify detection results?

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Triage

Can you accurately classify detection results?

- Frequency counts can be used to identify problematic & noisy rules
- Correlation should be used to accurately identify breaches or prioritize triage
- For example correlating signatures severity and asset importance to other forensic evidence (event_type=Alert + DNS, SSH Logs, Bash History)
  - C&C behavior or unauthorized software install on high priority asset.. ex: domain controller?
The Dark Side of the SOC

conflict, greed, the passage of time, death, and insanity

Who are your adversaries? What are their capabilities?

DARK SIDE OF THE SOC

APT, CRIMINAL SYNDICATES, BOTS, SCRIPT KIDDIES, HACKTIVISTS

THREATS

REPORTING, POST-MORTEM, INCIDENT REVIEW
REAL-TIME MONITORING, KILL CHAIN
MACHINE LEARNING, THREAT HUNTING, ANOMALY DETECTION
THREAT INTELLIGENCE, RISK ASSESSMENT, O3INT

TELEMETRY, ENDPOINT, INCIDENT RESPONSE

INVENTORY

ASSET INVENTORY, THREAT MODELING, KSI

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Threats

Who are your adversaries? What are their capabilities?

- Geographical adversaries
- Political adversaries
- Cyber criminals
- Competition/3rd party actors
- Hacktivists/Scriptkiddies

Understanding the motivations behind various types of attackers is critical to understanding the systems they are likely to target.

Threat intel is contextual piece to understanding patterns within big data, knowing the psychology of your adversary is more important. Use of honeypots.

The increasing profitability of cybercrime has lead to a global increase in attacks targeting cryptocurrency, or scams involving cryptocurrency schemes.

“If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.” – Sun Tzu

Suricata Flow connections to: privileged ports, honeypot services
The Dark Side of the SOC

conflict, greed, the passage of time, death, and insanity

Can you detect adversary activity within your environment?
The weakness of most adversaries are the behavioral chains they crate in order to propagate across the network.

North – South analytics focus on threat intel, entropy (randomness), bi-gram pairs of DGA domains. With the goal of identifying C&C communications, data exfil (DNS, HTTP)

East – West analytics focus on behavioral patterns within the enterprise, deviations from normal authentication & traffic behavior, scanning/port enumeration & mass connections to many endpoints from a single device. (FLOW, DNS, DHCP)

The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Can you detect an adversary that is already embedded?

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Can you detect an adversary that is already embedded?

- **Entity & Peer Group Analytics**
  - Anomaly Detection
    - Statistical or Algorithmic
      - Standard Deviation, IQR, Absolute Median Standard Deviation
      - Clustering, distance from centroid, distance from peers
  - Chebyshev’s Inequality
    - “In practical usage, in contrast to the 68–95–99.7 rule, which applies to normal distributions, Chebyshev’s inequality is weaker, stating that a minimum of just 75% of values must lie within two standard deviations of the mean and 89% within three standard deviations.”
    - 68–95–99.7 rule
    - Normal distributions
    - Standard Deviation
    - 75% within two standard deviations
    - 89% within three standard deviations
  - Predict future value based on Past Behavior
  - Client-Server Relationship (Producer Consumer Ratio)
    - 1.0 – Pure Push - FTP upload, multicast, beaconing
    - 0.4 – 70:30 export - Sending Email
    - 0.0 – Balanced Exchange - NTP, ARP probe
    - -0.5 – 3:1 import - HTTP Browsing
    - -1.0 – pure pull - HTTP Download
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

During an intrusion, can you observe adversary activity in real time?

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Track
During an intrusion, can you observe adversary activity in real time?

- Less Q.Q More Pew Pew
- Using historical information and unsupervised machine learning to build machine learning models based the features of your adversary
  - Detecting **Unknown Unknowns**
  - Ransomware: Shannon Entropy, Bi-Gram & Tri-Gram Ranges, & Character based features of DNS Query (DNS, HTTP)
  - Botnet: Producer Consumer Ratio, Total Packet/Bytes Transmitted, Bytes Received, Bytes Sent (FLOW)
  - Lateral Movement: Producer Consumer Ratio, Total Packet/Bytes Transmitted, Bytes Received, Bytes Sent (FLOW & SSH)
The Dark Side of the SOC

conflict, greed, the passage of time, death, and insanity

Can you deploy proven countermeasures to evict and recover?

"The Incident Response Hierarchy of Needs" by Matt Swann is licensed under CC BY 4.0
Act
Can you deploy proven countermeasures to evict and recover?

- Utilizing Machine Learning for detection and an automation tool of choice: quickly evict adversaries from the network
- Automated responses capture relevant information creating a feedback loop for training and updating models (DNS, HTTP)
- Contextual information provided to analysts for further investigation or automation.
- Phishing & Ransomware Campaigns
  - Usage of dynamically subdomains
  - Entropy & n-gram analysis of subdomains
  - IP reputation: newly registered domains
  - First time seen email domain
  - File hashing / presence of attachments
The Dark Side of the SOC
conflict, greed, the passage of time, death, and insanity

Can you collaborate with trusted partners to disrupt adversary campaigns?
Disruption

Can you collaborate with trusted partners to disrupt adversary campaigns?

- Files transmitted during the C&C process are hashed (Fileinfo)
- Information can be publishing to the wider community
  - ThreatConnect
  - ThreatExchange - Facebook

- Repeatable machine learning workflow and inspectable model files can be leveraged to improve threat hunting and disrupt campaigns with other partners.
Questions?