Learn How Splunk Enables Fast Science for Cybersecurity Operations

OLISA STEPHENSBAILEY
DAVID BRENNAN
SEPTEMBER 2017
DATA SCIENCE OPS IN PRACTICE

LEARN HOW TO:

ADDRESS CULTURAL CHALLENGES

ENSURE YOUR DATA SCIENCE SOLUTIONS GET USED

HARNESS THE FULL POWER OF PYTHON WITHIN SPLUNK

AGENDA

SECTION 1: UNDERSTANDING THE CORE NEED
SECTION 2: CROSSING THE ANALYSIS CHASM
SECTION 3: ANALYSIS WORKFLOW DEMONSTRATION
SECTION 4: ACTION ITEMS FOR YOUR PROJECTS
UNDERSTANDING THE CORE NEED
THE ROLE OF DATA SCIENCE IN CYBER OPERATIONS

• The rate of data growth is outpacing human capabilities

• We must optimize impact of the people we do have

• Data Science is a powerful tool to reduce the scale of the problem

• In response to these needs, Booz Allen Hamilton was tasked with integrating Data Science into the Watchfloor
CYBER OPERATIONS ANALYSTS & DATA SCIENTISTS POINTS OF VIEW

Cyber Operations Analysts
• Are evaluated on quantity of output
• Have a clearly defined SOP
• Will lose productivity every time they invest in learning a new tool
• Do not need new tools to be effective
• Are leery of buggy prototype code
• Have a distrust of the black box Machine Learning algorithm

Data Scientists
• Like to understand what the Analyst is trying to do rather than fit existing solution to problem
• Are evaluated on development of novel methods
• Gain honor and reputation from implementing cutting edge algorithms
• Do not like supporting legacy software
• Have an unwavering trust in mathematics

I must meet my quota, I don’t have time for toys

The old way is out of date, we must improve
APPRECIATING YOUR ROLE FOUNDATIONAL KEY TO SUCCESS

• The most important lesson learned

**Analysts are fully capable of meeting their current objectives without Data Science**

• Analysts are in a power position:
  – They are needed
  – They own the domain knowledge
  – They own the tradecraft
  – They own the accesses
  – They own the data

• It is the responsibility of the Data Scientist to show respect and learn
  – The Data Scientist is intruding into the Analyst's domain
CROSSING THE ANALYSIS CHASM
BRIDGING THE GAP BETWEEN ANALYSTS & DATA SCIENTISTS IN OPERATIONS

• Many Analysts do not understand applied statistics or machine learning and do not understand how it can be applied to their domain

• Data Scientists wishing to make an impact should:
  – Minimize the number of new widgets an analyst needs to learn
  – Provide all results with meaningful supporting evidence
  – Weight clarity as much as performance in algorithm selection
  – Appreciate that reporting there are no results is far better than false positives

• Host your end-solutions in the tool environment they use

Minimize Number of Tools
Provide Evidence
Ensure Interpretability
Silence Is a Virtue
If Analysts Use Splunk, You Use Splunk
LEVERAGING THE POWER & FLEXIBILITY WITH PYTHON & SPLUNK

<table>
<thead>
<tr>
<th>Python</th>
<th>Splunk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td><strong>Pros</strong></td>
</tr>
<tr>
<td>- Provides developers with access to wide array of data processing libraries</td>
<td>- Single unified system for collecting, digesting and querying data</td>
</tr>
<tr>
<td>- Object-Oriented program design</td>
<td>- Attractive 2D plotting</td>
</tr>
<tr>
<td>- Rapid prototype scripting language</td>
<td>- Users able to seamlessly navigate to rawdata behind plots</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td>- Must be able to code</td>
<td>- Query language narrows findings</td>
</tr>
<tr>
<td>- Developed projects tend to be individual objects</td>
<td>- Lacks flexibility of programing language</td>
</tr>
<tr>
<td>- Steep learning gap for users</td>
<td>- Limited python library within SDK</td>
</tr>
</tbody>
</table>

Combine the development flexibility of Python with the consistency of Splunk to benefit Analysts
STEP #1 - WORK DIRECTLY WITH ANALYSTS TO SOURCE A USE CASE

• Our Data Science team works directly with Analysts to work together on analytic objectives
  – To identify malicious or aberrant behavior within a new batch of log data
  – To detect suspicious URLs

• Their work flow consisted of:
  1. Digest log files into Splunk
  2. Label fields
  3. Explore the data with SMEs and via Splunk queries
  4. Report any new Splunk queries of value

We expedite Analysts’ Splunking by
• Grouping similar observations
• Highlighting suspicious outliers
• Unlocking new features
STEP #2 – SELECT METHOD FOR INTEGRATING DATA SCIENCE CAPABILITIES

METHOD 1
- This method has proven capable in rapid delivery situations
- Identify a linking field and export the data out of Splunk
- Process the data with any Data Science Software
- Create a new CSV and use previous linking field to enrich original data
STEP #2 – SELECT METHOD FOR INTEGRATING DATA SCIENCE CAPABILITIES

METHOD 2

• Slower to set up first time, but highly effective after that
• Use your own Python environment
• Able to leverage any library; Scikit-Learn, Tensor Flow, Theano, Scrapy, etc.
STEP #3 – EXECUTE MACHINE LEARNING ALGORITHM DEVELOPMENT PROCESS

• Splunk is a powerful asset in many stages of the Machine Learning process
ANALYSIS WORKFLOW
DEMONSTRATION
LOOK FAMILIAR?
STEP #4 – SHOW EVIDENCE TO SUPPORT ANALYSIS RESULTS

JUST BELIEVE ME ‘CAUSE I’M AWESOME!

THE NOTORIOUS BLACK BOX
OUR NEW FEATURE EXTRACTION APPLICATION BRINGS NEW INSIGHTS TO ANALYSIS

New Stream App Feature Examples – Avoid Basic Summary Table Overhead

<table>
<thead>
<tr>
<th>Avg</th>
<th>IP, port, time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>sum(bytes), sum(bytes_in), sum(bytes_out), sum(packets_in), sum(packets_out), sum(response_time), sum(time_taken)</td>
</tr>
</tbody>
</table>

Our New Feature Examples - Make Better Use of ML Toolkit

<table>
<thead>
<tr>
<th>Numeric</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>num_bytes_cli2srv, num_bytes_srv2cli, num_packets_cli2srv, num_packets_srv2cli, packet_deltat_avg_cli2srv, packet_deltat_avg_srv2cli, packet_deltat_entropy_2way, packet_deltat_entropy_cli2srv, packet_deltat_entropy_srv2cli</td>
</tr>
</tbody>
</table>

We added 46 new features!!!!
NEW STREAM APP ENABLES DIRECT ACCESS TO RAW PCAP IN SPLUNK
NEW STREAM APP GIVE ANALYSTS MORE INFORMATION
ML TOOLKIT ENABLES EXPLORATORY DATA ANALYSIS IN SPLUNK
STOCK SPLUNK ML TOOLKIT HAS LIMITED FEATURES AVAILABLE FOR ANALYSIS

90% of ML is Pre-Processing & Feature Extraction
Crafting Features is Necessary Before Feeding The MLTK
DATA SCIENTISTS CAN ADD NEW FEATURES DIRECTLY INTO SPLUNK FOR EDA
USER EXPERIENCE AND SUPPORTING EVIDENCE FOR DATA SCIENTISTS
USER EXPERIENCE AND SUPPORTING EVIDENCE FOR ANALYSTS

```
1  source=\"whyPy\_flowmonster\_output\_csv\" host=\"CD2RS2JLGBMW\_local\" index=\"flowmonster\" sourcetype=\"csv\" | eventstats avg\(\{\text{\textquotedblleft\text{duration}\textbf{\textnormal{\textquotedblright}}}\}\\) as avg \(\text{stdev}\)\(\{\text{\textquotedblleft\text{duration}\textbf{\textnormal{\textquotedblright}}}\}\) as stdev | eval lowerBound=avg-stdev as lowerBound, upperBound=avg+stdev | eval isOutlier=if duration < lowerBound OR duration > upperBound, 1, 0)
```

![Graphical representation of the data analysis](image-url)
LIVE DEMO
ACTION ITEMS FOR YOUR PROJECTS
CULTURAL HURDLES & SUCCESSES

• Tactics used to overcome cultural barriers
  – You must go to the analyst; they will show you their analysis process AND grant you keys to their data troves
  – You must be willing to explain what analysis techniques you are using simply using their terminology as much as possible
  – Someone on your team has to be willing to talk to the customers and their customers- this helps establish a new, collaborative tribe
  – Your work must role up into a story that tells the why and so what of the work- sometimes this is the closest one gets to ROI
  – Marketing & branding extremely important for breaking entrenched thinking and coaxing participation to something new & shiny

• Build an interdisciplinary team
  – Unicorns are hard to find and the best solutions often are a product of divergent thought
  – Data analysis is a pipeline, journey of sorts...it takes domain experts from fields other than just computer science or mathematics
  – Having data scientists that have expertise in Cyber Operations mission space will accelerate success
FOUR STEPS TO APPLYING DATA SCIENCE WITHIN CYBER OPERATIONS

• STEP #1 - WORK DIRECTLY WITH ANALYSTS TO SOURCE A USE CASE

• STEP #2 – SELECT METHOD FOR INTEGRATING DATA SCIENCE CAPABILITIES

• STEP #3 – EXECUTE MACHINE LEARNING ALGORITHM DEVELOPMENT PROCESS

• STEP #4 – SHOW EVIDENCE TO SUPPORT ANALYSIS RESULTS
TAKE AWAYS

1) Your data science team must go to the analyst

2) Populate your results where the user checks

3) Develop self-contained limited size products that can be iteratively updated and delivered

4) Data Scientists must be concerned with justifying their claims

5) Splunk can be enhanced by leveraging external scripting
INNOVATING THE CYBER DOMAIN THROUGH THE APPLICATION OF DATA SCIENCE