



Innovation center, Washington, D.C.

# DATA SCIENCE OPS IN PRACTICE

*Learn How Splunk Enables Fast Science for  
Cybersecurity Operations*

OLISA STEPHENSBAILY  
DAVID BRENMAN  
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

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

## 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

**The old way is out of date,  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



[2]

# 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

---

## Python

- **Pros**
  - Provides developers with access to wide array of data processing libraries
  - Object-Oriented program design
  - Rapid prototype scripting language
- **Cons**
  - Must be able to code
  - Developed projects tend to be individual objects
  - Steep learning gap for users

## Splunk

- **Pros**
  - Single unified system for collecting, digesting and querying data
  - Attractive 2D plotting
  - Users able to seamlessly navigate to rawdata behind plots
- **Cons**
  - Query language narrows findings
  - Lacks flexibility of programming language
  - Limited python library within SDK

**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**

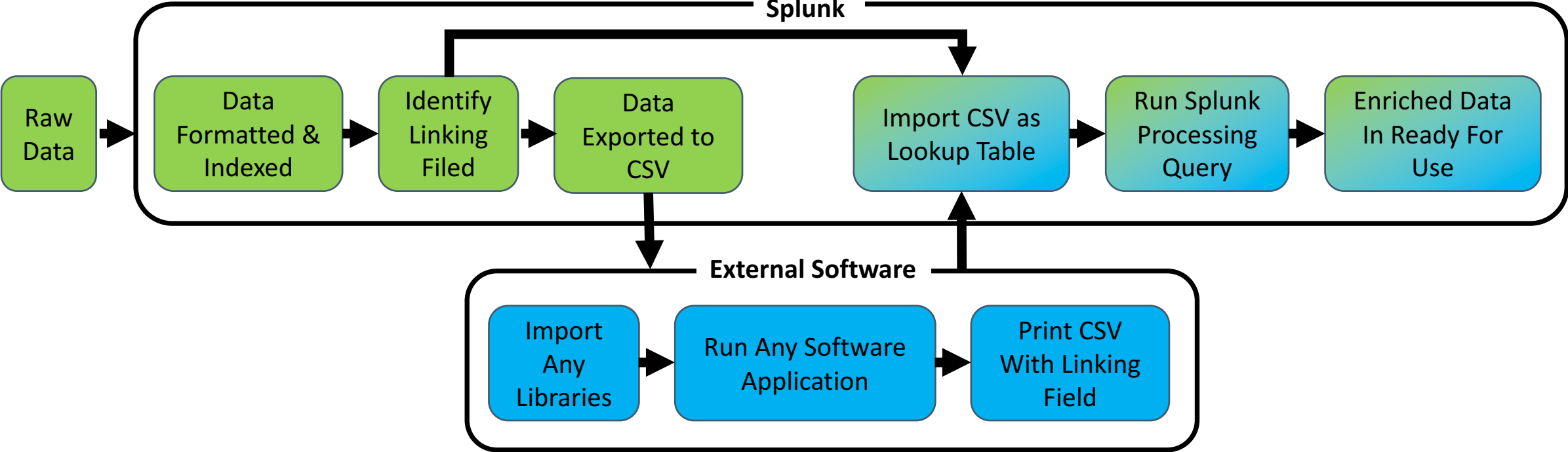
- **Grouping similar observations**
- **Highlighting suspicious outliers**
- **Unlocking new features**



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

## METHOD 1

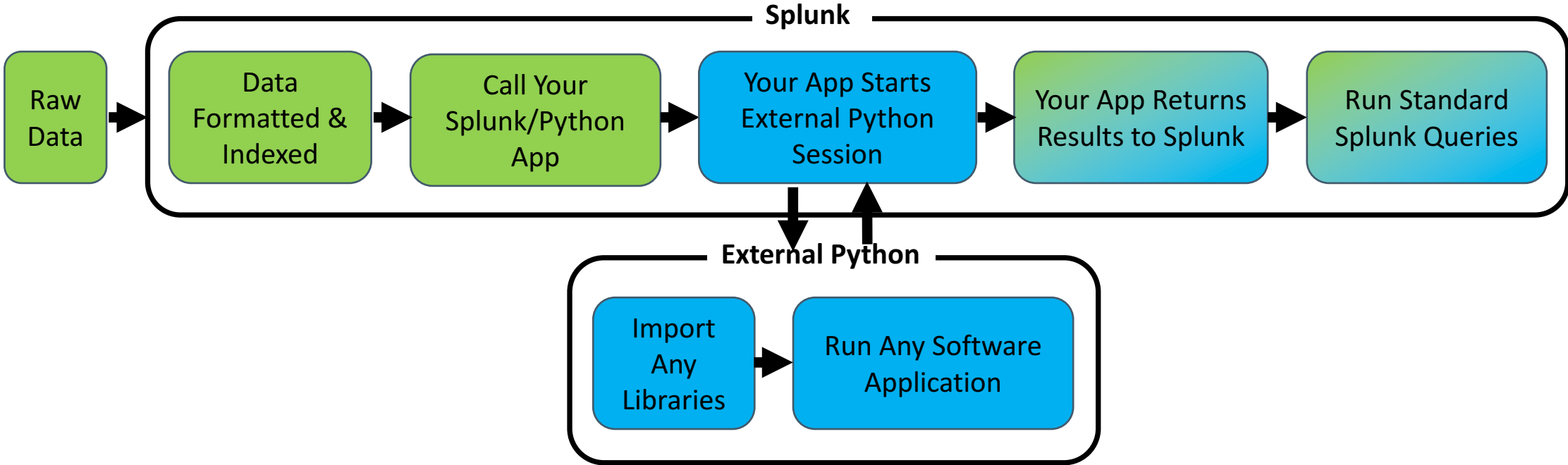
- This method has proven capable in rapid delivery situations
- Identify a linking field and export the data out of Splunk
- Process the data with **any** Data Science Software
- Create a new CSV and use previous linking field to enrich original data



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

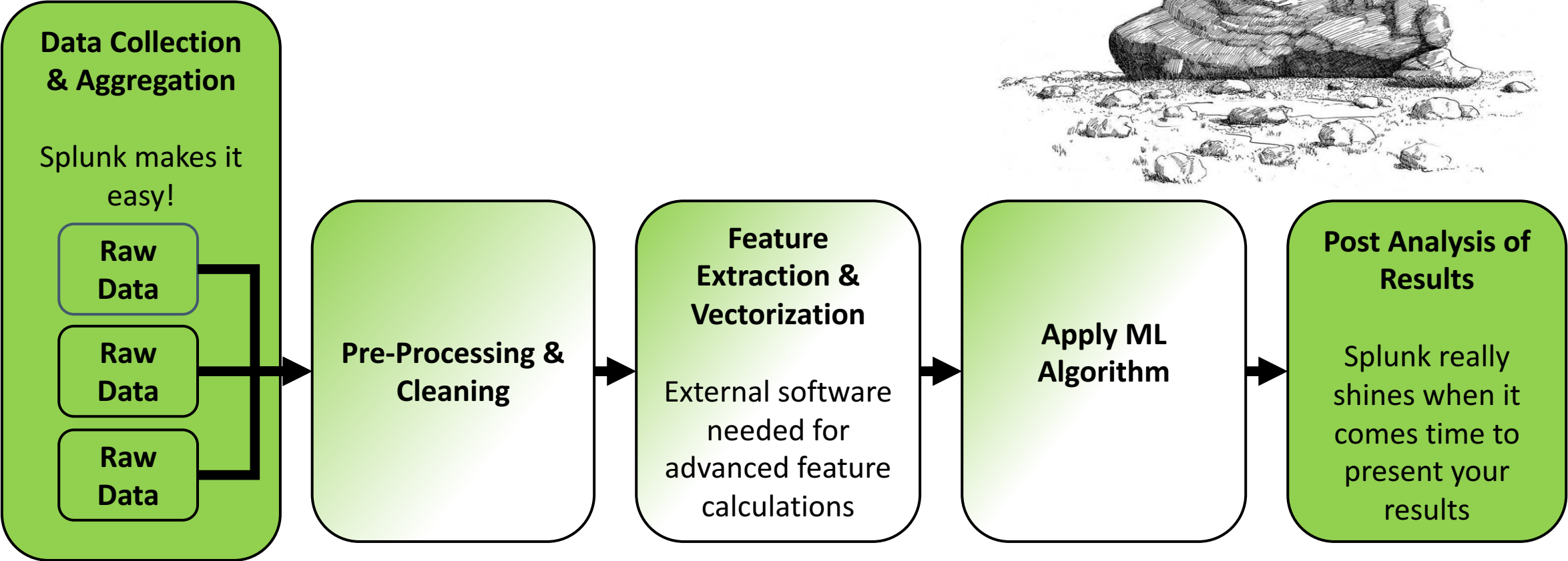
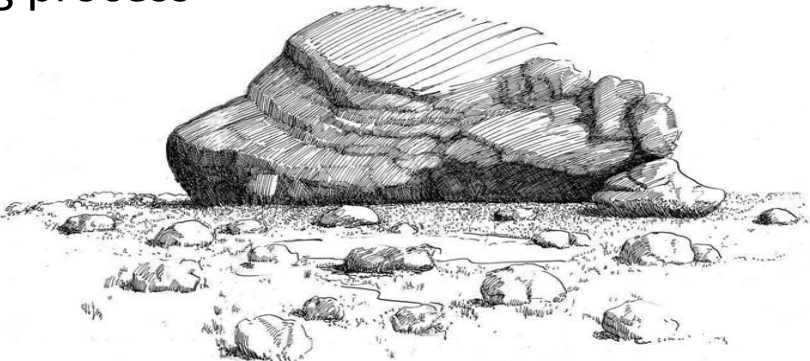
## METHOD 2

- 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

---





# OUR NEW FEATURE EXTRACTION APPLICATION BRINGS NEW INSIGHTS TO ANALYSIS



**New Stream App Feature Examples – Avoid Basic Summary Table Overhead**

|             |   |
|-------------|---|
| Avg         | IP, port, time  |
| Statistical | sum(bytes), sum(bytes_in), sum(bytes_out), sum(packets_in), sum(packets_out), sum(response_time), sum(time_taken) |

**Our New Feature Examples - Make Better Use of ML Toolkit**

|             |  |
|-------------|--|
| Numeric     | duration   |
| Statistical | 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 |

We added 46 new features!!!!





# NEW STREAM APP ENABLES DIRECT ACCESS TO RAW PCAP IN SPLUNK



List ▾ / Format 50 Per Page ▾

< Hide Fields All Fields

Interesting Fields

- a host 1
- a index 1
- # linecount 1
- a source 2
- a sourcetype 2
- a splunk\_server 1

+ Extract New Fields

**SOURCE**

2 Values, 100% of events Selected Yes No

Reports

- Top values
- Top values by time
- Rare values

Events with this field

| Values            | Count | %       |
|-------------------|-------|---------|
| stream:Splunk_Tcp | 20    | 83.333% |
| stream:Splunk_Udp | 4     | 16.667% |

Show as raw text

```
> 3/17/15 4:46:25.076 PM { [-]
  app: dns
  count: 68
  endtime: 2015-03-17T20:46:25.076262Z
  sum(bytes): 16133
  timestamp: 2015-03-17T20:46:25.076262Z
}
```

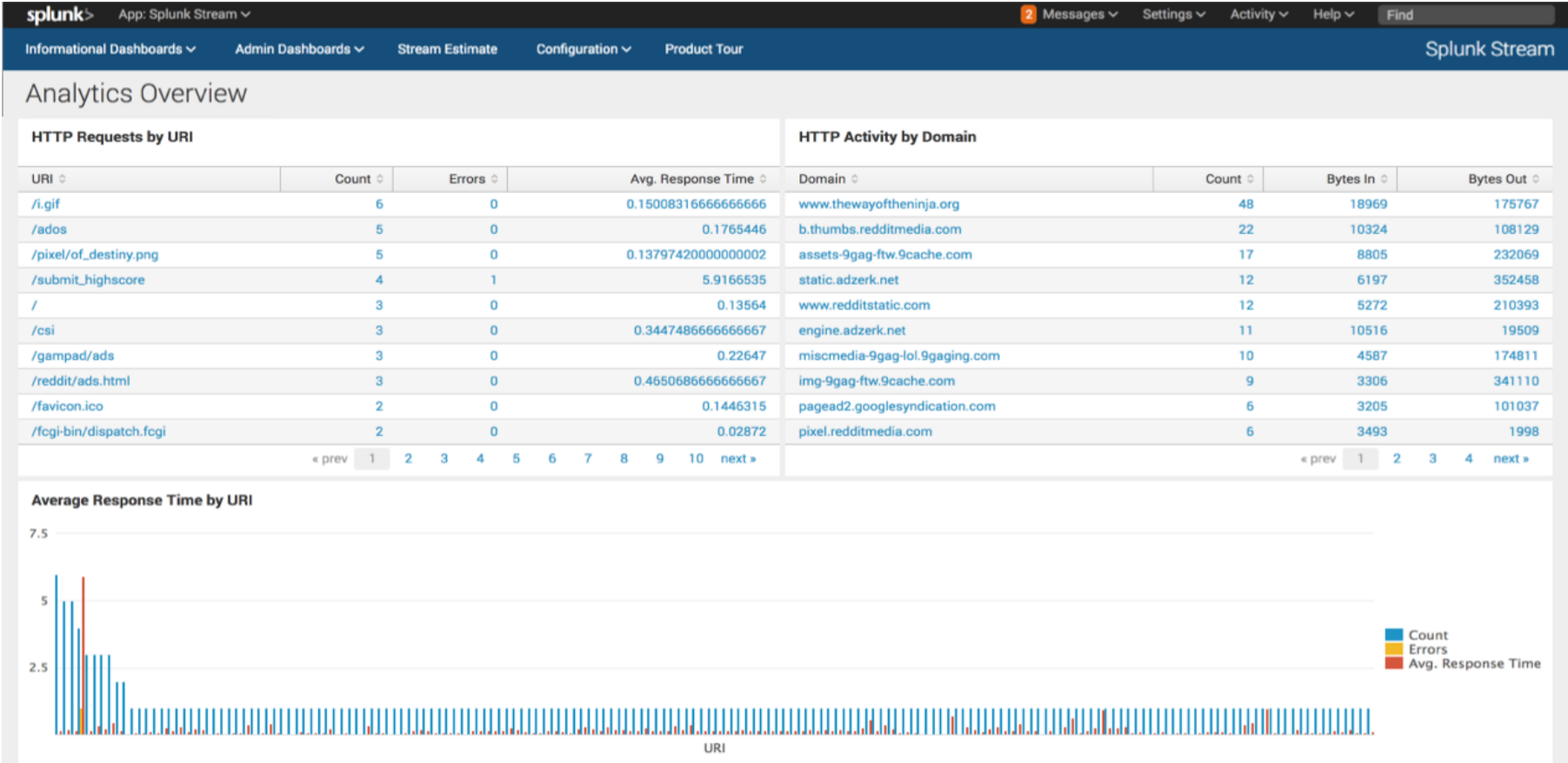
Show as raw text

```
> 3/17/15 4:46:10.303 PM { [-]
  app: unknown
  count: 21
  endtime: 2015-03-17T20:46:10.303392Z
  sum(bytes): 6060
  timestamp: 2015-03-17T20:46:10.303392Z
}
```

Show as raw text

```
> 3/17/15 4:46:10.303 PM { [-]
  app: twitter
  count: 2
  endtime: 2015-03-17T20:46:10.303392Z
  sum(bytes): 64207
}
```

# NEW STREAM APP GIVE ANALYSTS MORE INFORMATION



# ML TOOLKIT ENABLES EXPLORATORY DATA ANALYSIS IN SPLUNK

**splunk** App: Splunk Machine Learning Toolkit

Messages Settings Activity Help Find

Search Showcase Assistants Scheduled Jobs Docs Video Tutorials Splunk Machine Learning Toolkit

## Showcase

Welcome to the Showcase, which exhibits some of the analytics enabled by this app. Click on the name of an analytic to reach the corresponding Assistant, which will guide you through the process of applying it to your data. Click on one of the examples to see that Assistant applied to a real dataset. Please see the [video tutorials](#) for more information.

Select which examples to show

All Examples

### Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

- Predict Server Power Consumption
- Predict VPN Usage
- Predict Median House Value
- Predict Power Plant Energy Output

### Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

- Predict Hard Drive Failure
- Predict the Presence of Malware
- Predict Telecom Customer Churn
- Predict the Presence of Diabetes
- Predict Vehicle Make and Model

### Detect Numeric Outliers

Find values that differ significantly from previous values.

- Detect Outliers in Server Response Time
- Detect Outliers in Number of Logins (vs. Predicted Value)
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Power Plant Humidity

### Detect Categorical Outliers

Find events that contain unusual combinations of values.

- Detect Outliers in Disk Failures
- Detect Outliers in Bitcoin Transactions
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Mortgage Contracts
- Detect Outliers in Diabetes Patient Records
- Detect Outliers in Mobile Phone Activity

### Forecast Time Series

Forecast future values given past values of a metric (numeric time series).

- Forecast Internet Traffic

### Cluster Numeric Events

Partition events with multiple numeric fields into clusters.

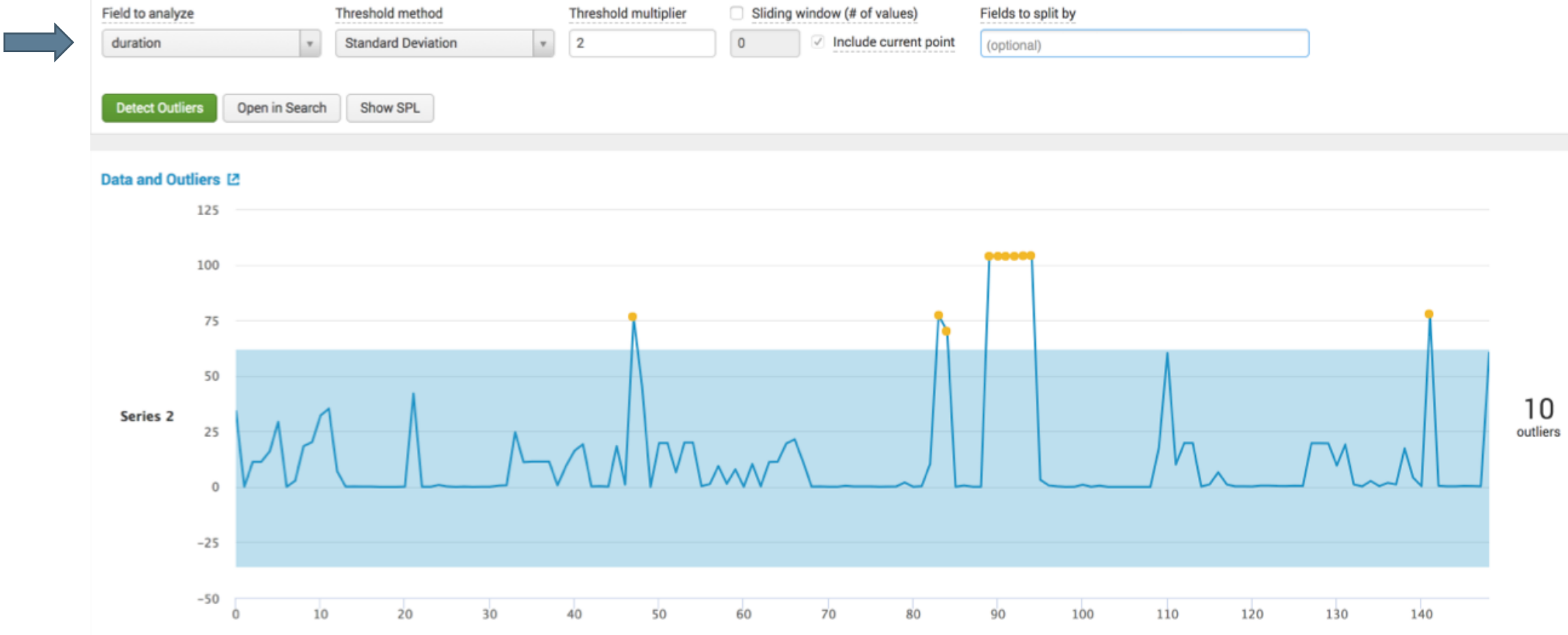
- Cluster Hard Drives by SMART Metrics
- Cluster Behavior by App Usage

# STOCK SPLUNK ML TOOLKIT HAS LIMITED FEATURES AVAILABLE FOR ANALYSIS

The screenshot shows the Splunk ML Toolkit interface for creating a model. The top navigation bar includes 'Search', 'Showcase', 'Assistants', 'Scheduled Jobs', 'Docs', and 'Video Tutorials'. The main heading is 'Predict Numeric Fields' with a subtitle 'Predict the value of a numeric field using a weighted combination of the values of other fields in that event.' Below this are tabs for 'Create New Model' and 'Load Existing Settings'. A search bar contains 'inputlookup server\_power.csv' and shows '31,272 results'. The 'Algorithm' is set to 'LinearRegression'. The 'Field to predict' is 'ac\_power'. The 'Fields to use for predicting' section contains a list of features: '\_time', 'total-cpu-utilization', 'total-disk-accesses', 'total-disk-blocks', 'total-disk-utilization', 'total-instructions\_retired', 'total-last\_level\_cache\_references', 'total-memory\_bus\_transactions', and 'total-unhalted\_core\_cycles'. A slider for 'Split for training / test' is set to '60 / 40'. On the left, there is a 'Fit Intercept' section with a checked box for 'estimate the intercept' and a 'Save the model as' field with '(optional)'. At the bottom, there are buttons for 'Fit Model', 'Open in Search', and 'Show SPL'.

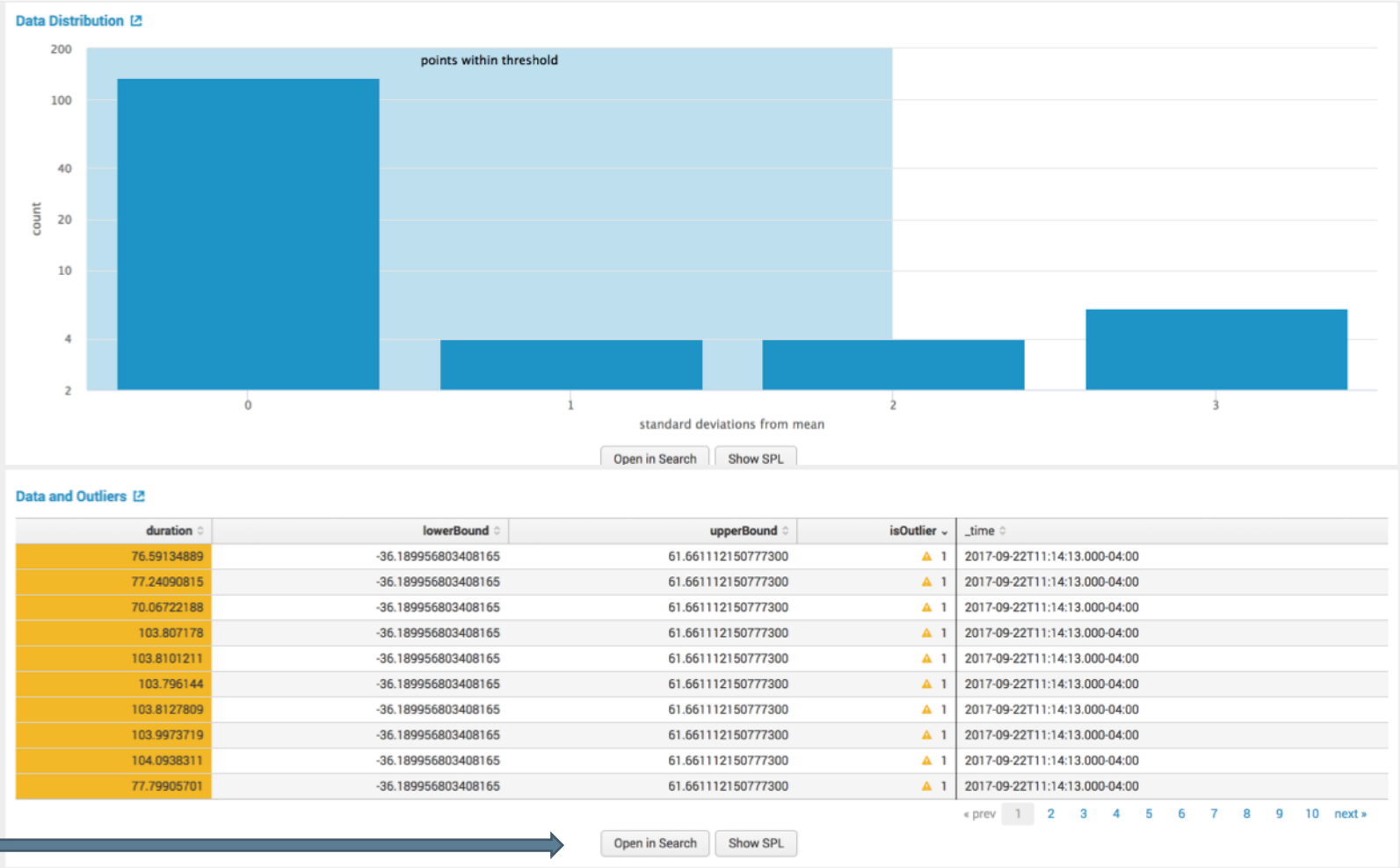
**90% of ML is Pre-Processing & Feature Extraction  
Crafting Features is Necessary Before Feeding The MLTK**

# 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="C02R52JLG8WM.local" index="flowmonster" sourcetype="csv" | eventstats avg("duration") as avg stdev("duration") as stdev | eval lowerBound=(avg-stdev\*exact(2)), upperBound=(avg+stdev\*exact(2)) | eval isOutlier=if('duration' < lowerBound OR 'duration' > upperBound, 1, 0)

✓ 149 events (before 9/26/17 10:51:17.000 AM) No Event Sampling

Events (149) Patterns Statistics Visualization

Format Timeline Zoom Out Zoom to Selection Deselect 1 millisecond per column

Table Format 20 Per Page

|   | i | _time                         | cli_ip         | isOutlier | duration    | port_cli | srv_ip         | port_srv | num_packets_cli2srv | num_packets_srv2cli | num_bytes_cli2srv | num_by |
|---|---|-------------------------------|----------------|-----------|-------------|----------|----------------|----------|---------------------|---------------------|-------------------|--------|
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 76.59134889 | 52351    | 216.58.208.206 | 443      | 7                   | 8                   | 1524              | 522    |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 77.24090815 | 52401    | 31.13.93.3     | 443      | 21                  | 37                  | 2155              | 25878  |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 70.06722188 | 52400    | 31.13.93.3     | 80       | 7                   | 7                   | 865               | 1242   |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 103.807178  | 52398    | 54.231.10.92   | 80       | 25                  | 40                  | 4216              | 36862  |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 103.8101211 | 52397    | 54.231.10.92   | 80       | 33                  | 60                  | 6067              | 55279  |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 103.796144  | 52396    | 54.231.10.92   | 80       | 29                  | 45                  | 5208              | 39529  |
| > |   | 9/22/17<br>11:14:13.000<br>AM | 172.16.254.128 | 1         | 103.8127809 | 52395    | 54.231.10.92   | 80       | 24                  | 35                  | 4979              | 24232  |

Selected Fields  
a cli\_ip 1  
# duration 100+  
# isOutlier 2  
# num\_bytes\_cli2srv 100+  
# num\_bytes\_srv2cli 100+  
# num\_packets\_cli2srv 39  
# num\_packets\_srv2cli 56  
# packet\_deltat\_geometric\_mean\_sr  
v2cli 83  
# packet\_deltat\_kurtosis\_srv2cli 97  
# port\_cli 100+  
# port\_srv 2  
a srv\_ip 47

Interesting Fields  
# avg 1  
a cli\_srv\_two\_way\_conversation 99  
a host 1

# 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 roll 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

---

