Anomaly Hunting with Splunk

Macy Cronkrite

Architect, Professional Services – Splunk

Anthony Tellez

Senior Consultant, Professional Services – Splunk

.conf2016

Disclaimer

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 the 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.

Speakers Bio

- Anthony Tellez
  - Splunk Public Sector Federal Team
  - Previously @ NGA
  - Splunkbase App Developer
  - Machine Learning
  - National Security
  - Internet of Things
  - https://github.com/anthonygtellez
- Macy Cronkrite
  - Splunk Public Sector Federal Team
  - Previously @ MITRE
  - Organizer for BSides Boston
  - Machine Learning
  - Insider Threat
  - Enterprise Security
  - @macycron

What is Data Science?

"Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible."

-Mike Driscoll CEO, Metamarket

Agenda

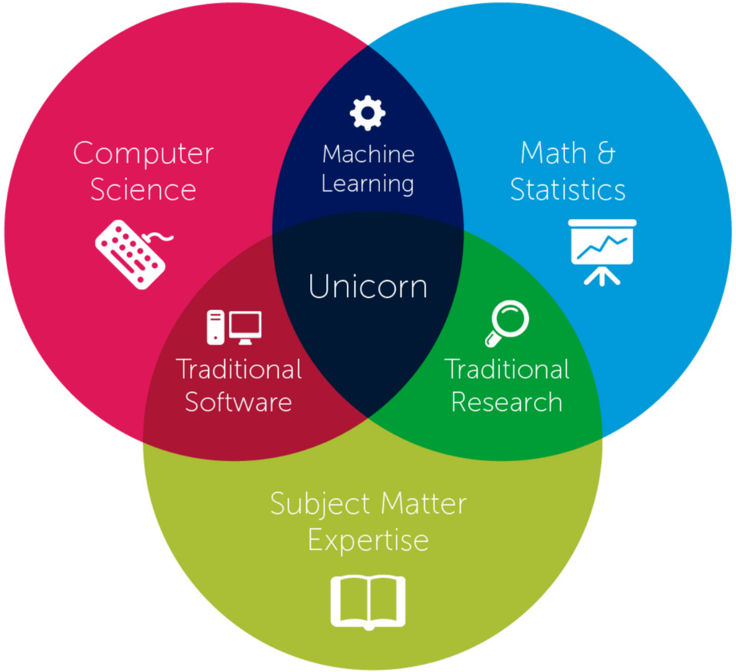
- 5 Step Data Science Methodology for Security
- Quantitative vs Qualitative Analysis
- Descriptive Statistics
- Exploratory Data Analysis (EDA)
- Explore Core and Add-on Splunk analytic capabilities

splunk> .conf2016

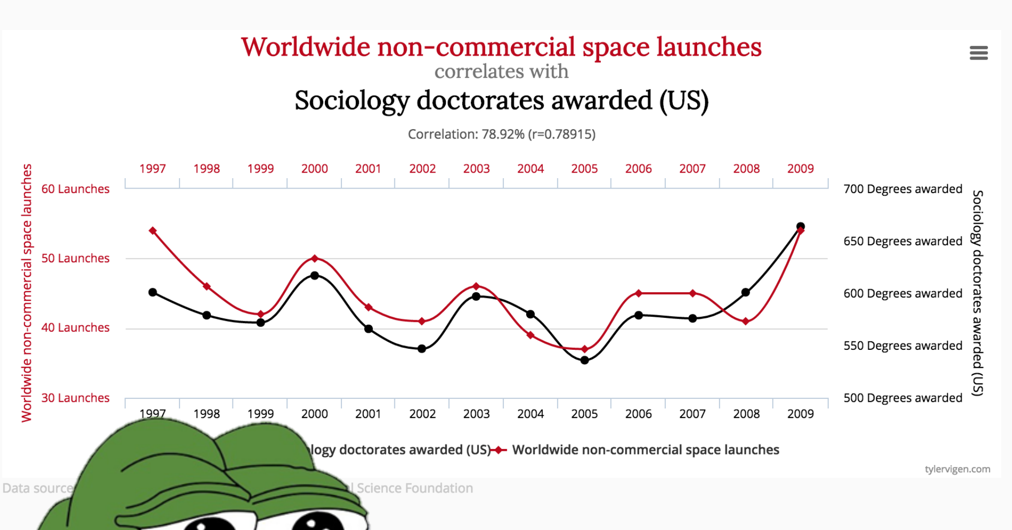
Security Data Analysis

Splunk empowers the security analyst by making their machine data valuable, usable and actionable…but….

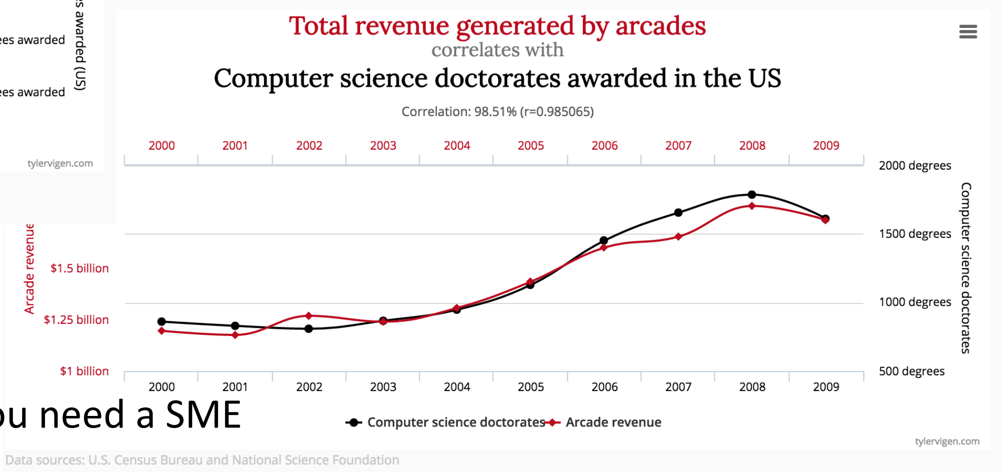
- Information Overload
  - IDS alerts, Virus Scans, tools.
- Multidisciplinary approach is needed for next gen problems
  - SIEM alone, ML alone, are not enough without SME.
- Our goal is to empower security analysts to reach the middle using statistical techniques built into Core Splunk, Enterprise Security & ITSI.
- **Everyone is capable of becoming a unicorn.**

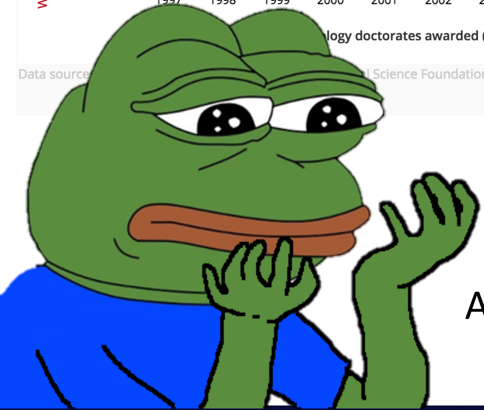


Correlation != Causation ☹



- Correlating some data may be a waste of time if you don't have an understanding of what the data represents.

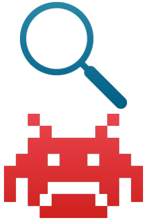




A good example of why you need a SME

5 Step Data Science Methodology for Security OPS

- **Step 1** Scope relevant machine data to onboard into Splunk.
- **Step 2** Collect requirements and validate relevant machine data.
- **Step 3** Exploratory Data Analysis. (Searching & Visualizing!)
- **Step 4** Formulate hypothesis working with Domain Experts.
- **Step 5** Test and repeat steps as needed until hypothesis is answered.

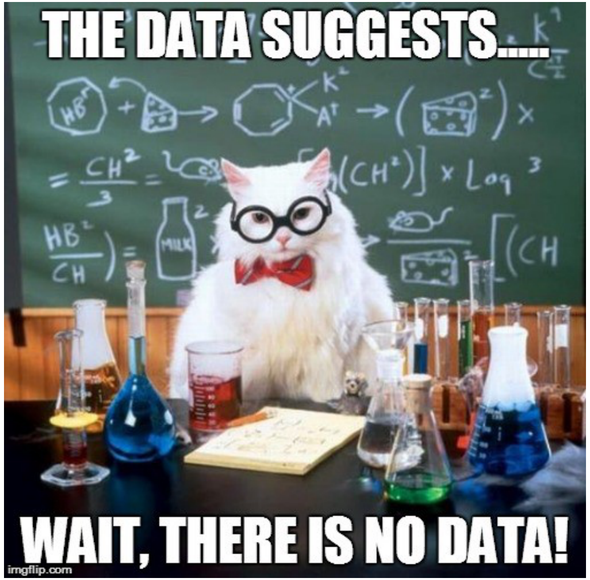


Applying Data Science to Security OPS

Step 1 Scope relevant machine data to onboard into Splunk.

Example Data Sources for monitoring network and public facing web applications

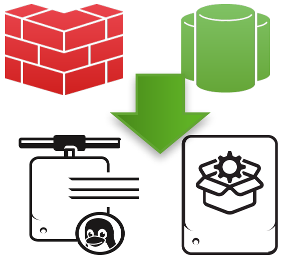
- Firewall Traffic
- SQL Server/HTTP Logs



Security Patterns in Machine Data

What To Look For	Data Source
Abnormally high number of file transfers to USB or CD/DVD	**Operating system**
Abnormally high number of files or records downloaded from an internal file store or database containing confidential information	**File server / Database**
Abnormally large amount of data emailed to personal webmail accounts or uploaded to external file hosting site	**Email server / web proxy**
Unusual physical access attempts (after hours, accessing unauthorized area, etc.)	**Physical badge records / Authentication**
Excessive printer activity and employee is on an internal watch list as result of demotion / poor review / impending layoff	**Printer logs / HR systems**
User name of terminated employee accessing internal system	**Authentication / HR systems**
IT Administrator performing an excessive amount of file deletions on critical servers or password resets on critical applications (rogue IT administrator)	**Operating system /Authentication / Asset DB**
Employee not taking any vacation time or logging into critical systems while on vacation (concealing fraud)	**HR systems / Authentications**
Long running sessions, bandwidth imbalance between client & server, Bad SSL Configurations	**IPS / IDS / Stream**
Known cloud or malware domains, bad SSL Configurations	**Threat Intelligence, Custom Lookups**
High Entropy Subdomains	**Web proxy, DNS, Wiredata**

Applying Data Science to Security OPS



Step 1 Scope relevant machine data to onboard into Splunk.

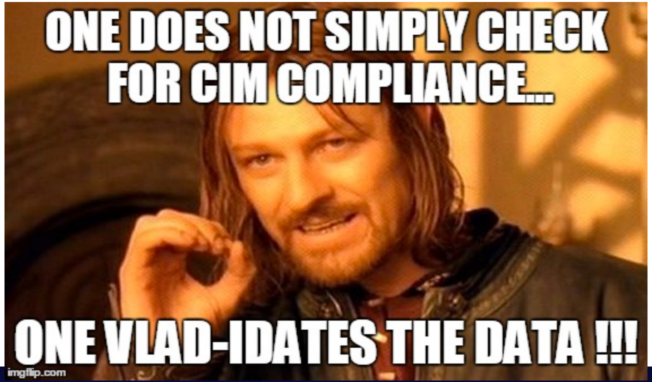
Step 2 Collect requirements and validate relevant machine data.

Example Collection Methods

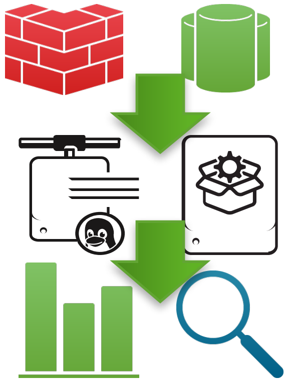
- Syslog Server for Firewall Traffic, Universal Forwarder
- Splunk Stream, DB Connect, IIS Logs for SQL Server

Example Validation Methods

- Splunkbase TA's
- Add-on Builder
- Regex101 to build search time fields
- Common Information Model



Applying Data Science to Security OPS



Step 1 Scope relevant machine data to onboard into Splunk.

Step 2 Collect requirements and validate relevant machine data.

Step 3 Exploratory Data Analysis. (Searching & Visualizing!)

- Number of connections between src_ip & dest_ip, iplocation
- Torrent activity (dest_port 6881-6889, 6969), connections to Tor Addresses, or Malware domains
- Interesting Fields: http_user_agent, http_method
- SQL Injection logic OR WHERE 1=1?

Applying Data Science to Security OPS

Step 1 Scope relevant machine data to onboard into Splunk.

Step 2 Collect requirements and validate relevant machine data.

Step 3 Exploratory Data Analysis. (Searching & Visualizing!)

Step 4 Formulate hypothesis working with Domain Experts.

- Is this real torrent traffic or another application using the same ports?
- Can users install or run TOR Browser onto their desktops in this VLAN?
- Is this SQL injection valid in user_agent field or just bad parsing of data during the onboarding process?

Can I disprove the activity by adding more data or context?

Relevant Data Sources

Raw Data	Lookups	Context	Value
Firewall Traffic	Username to IP	10.0.0.12 fails to login to 5 different servers	Determine user responsible
Proxy	Username to IP	10.0.0.12 visits Dropbox and uploads 1TB of data	Determine user responsible
Active Directory	User to Group Mapping	SPLUNK\JohnDoe authenticates to 30 different hosts in 30 second period	Determine scope of compromise, domain admin, SQL admin only?
DHCP	User to IP, Host to IP	10.0.0.12, 10.0.0.35 attempt to connect to TOR IP address	Determine user or hosts responsible
Email Transport	Baseline Usage	User sends email with large file attachments	Determine normal behavior
Exchange / Email	Baseline usage	User sends 40 emails in 60 minute period	Determine normal behavior
Packet Capture / Wire Data	Subnet to physical location / priority of asset	10.0.0.0/27 shows successful SSH connections originating from Russia	Determine where an asset is physically or scope of compromise based on VLAN

Applying Data Science to Security OPS

Step 1 Scope relevant machine data to onboard into Splunk.

Step 2 Collect requirements and validate relevant machine data.

Step 3 Exploratory Data Analysis. (Searching & Visualizing!)

Step 4 Formulate hypothesis working with Domain Experts.

Step 5 Test and repeat steps as needed until hypothesis is answered.

Applying Data Science to Security OPS

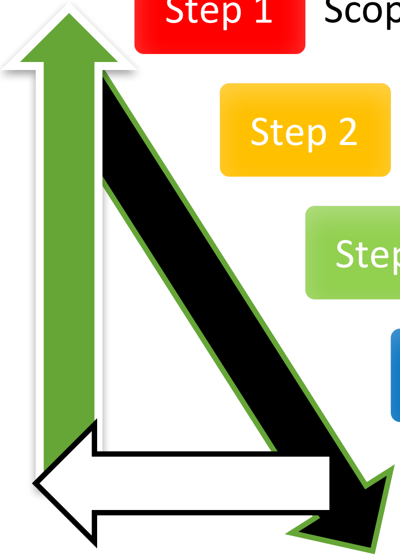
Step 1 Scope relevant machine data to onboard into Splunk.

Step 2 Collect requirements and validate relevant machine data.

Step 3 Exploratory Data Analysis. (Searching & Visualizing!)

Step 4 Formulate hypothesis working with Domain Experts.

Step 5 Test and repeat steps as needed until hypothesis is answered.

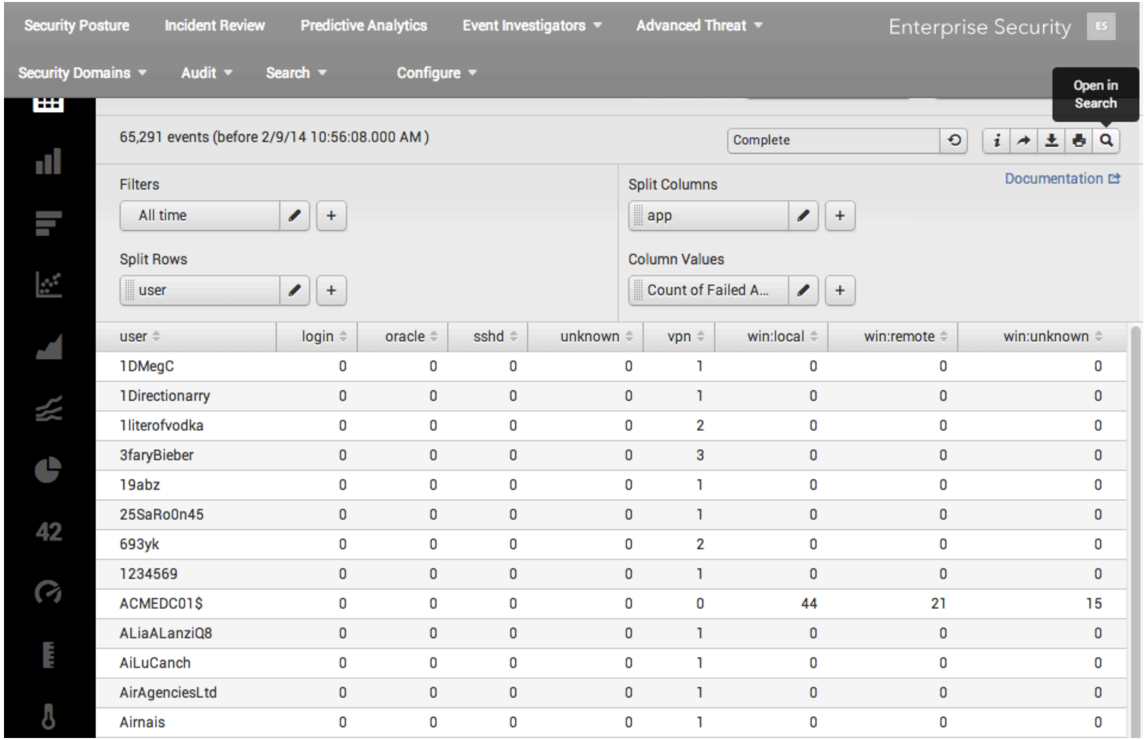


Quantitative vs **Quali**tative Analysis

- **Quan**titative measure:
- 5 Failed logins in 60 mins
- Threshold and periodicity fixed

- **Quali**tative measure:
- The failed login rate is increasing abnormally for this user.
- Threshold and periodicity is variable

Quantitative – Static Thresholds



user	login	oracle	sshd	unknown	vpn	win:local	win:remote	win:unknown
1DMegC	0	0	0	0	1	0	0	0
1Directionarry	0	0	0	0	1	0	0	0
1literofvodka	0	0	0	0	2	0	0	0
3faryBieber	0	0	0	0	3	0	0	0
19abz	0	0	0	0	1	0	0	0
25SaRo0n45	0	0	0	0	1	0	0	0
693yk	0	0	0	0	2	0	0	0
1234569	0	0	0	0	1	0	0	0
ACMEDC01$	0	0	0	0	0	44	21	15
ALiaALanziQ8	0	0	0	0	1	0	0	0
AiLuCanch	0	0	0	0	1	0	0	0
AirAgenciesLtd	0	0	0	0	1	0	0	0
Airnais	0	0	0	0	1	0	0	0

Quantitative

Enterprise Security version 2 - 3

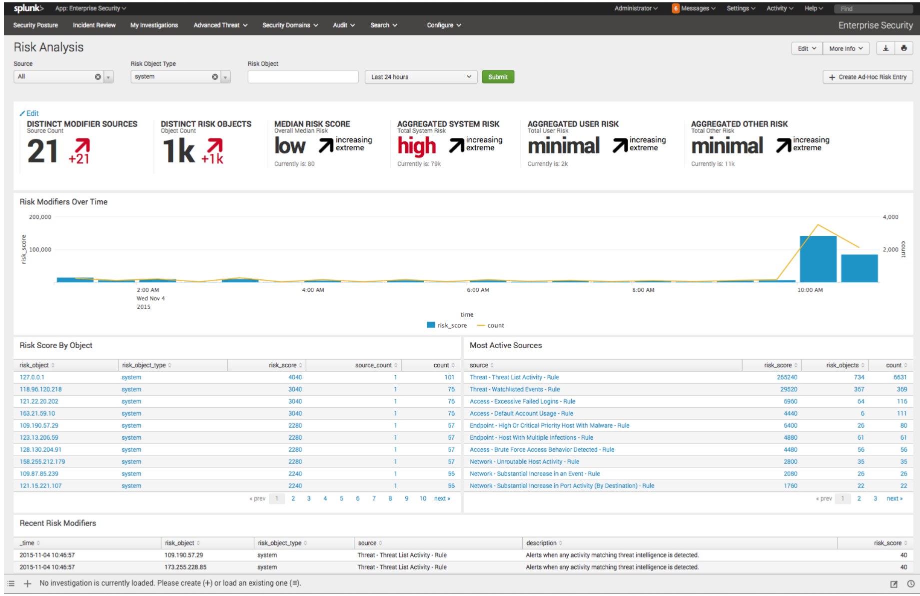
```
| datamodel("Authentication","Authentication")
| stats values Authentication.tag  as tag,
count(eval(Authentication.action==  failure")) as failure ,
count  eval(Authentication.action ==  success")) as success
by Authenication.src
| search failure > 6 success > 1
```

Count Failures and successes by source, trigger when more than 6 failures in an hour followed by a success

Extreme Search

- An app that provides the ability to evaluate and interpret Splunk search results in a ***qualitative*** rather than a ***quantitative*** manner.

- Qualitative terms in Extreme search are expressed in terms of "fuzzy" quantitative ranges.
Eg. Minimal , high, extreme



Qualitative – 2 steps

Enterprise Security 3 - 4+ SA-ExtremeSearch

1. Create the model in a Context

- Count failures by src in an hour

```
| tstats `summariesonly` count as failures from datamodel=Authentication.Authentication where
Authentication.action="failure" by Authentication.src,_time span=1h
```

- Gather stats median, min, max, (descriptive statistics)

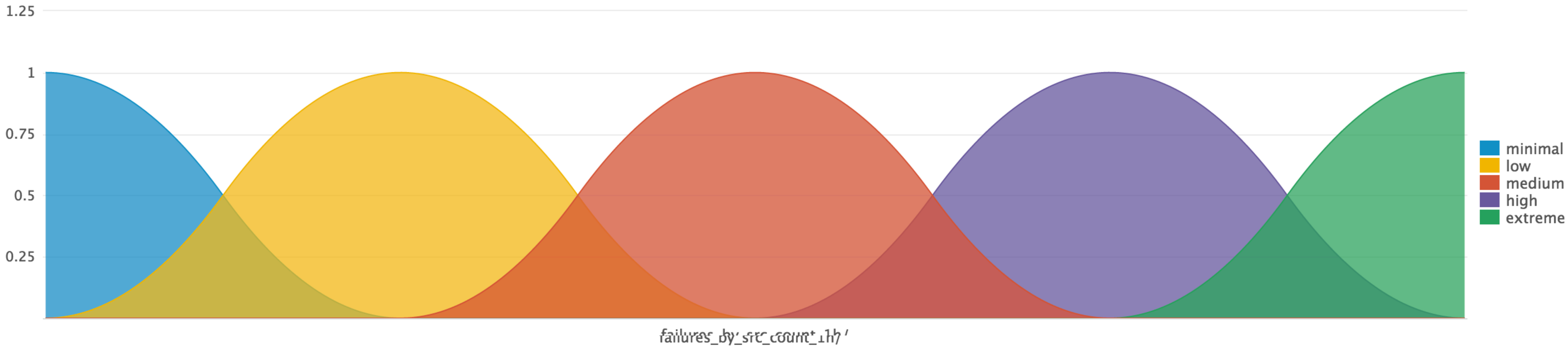
```
| stats median(failures) as median, min(failures) as min, count as count
| eval max = median*2
```

- Update the context with current stats

```
| xsUpdateDDContext app="SA-AccessProtection" name=failures_by_src_count_1h container=authentication
scope=app
| stats count
```

- Time Range -25h to -1h

Visualize Context



256 results   20 per page

< Prev 1 ... 5 6 7 8 9 10 11 12 13 Next >

failures_by_src_count_1h/	minimal	low	medium	high	extreme
19.41176	0.0000000000	0.0000000000	0.0000000000	0.1107268557	0.8892731667
19.45098	0.0000000000	0.0000000000	0.0000000000	0.0964549482	0.9035450816
19.49020	0.0000000000	0.0000000000	0.0000000000	0.0831679627	0.9168320298
19.52941	0.0000000000	0.0000000000	0.0000000000	0.0708651841	0.9291347861
19.56863	0.0000000000	0.0000000000	0.0000000000	0.0595460907	0.9404538870
19.60784	0.0000000000	0.0000000000	0.0000000000	0.0492117777	0.9507881999
19.64706	0.0000000000	0.0000000000	0.0000000000	0.0398616679	0.9601383209
19.68627	0.0000000000	0.0000000000	0.0000000000	0.0314957649	0.9685042500

Qualitative – Step 2 Compare Data to Model

- Compare the context model to a data sample
- Ex. Brute Force one hour Time Range -65m -5m

```
| `datamodel("Authentication","Authentication")`
| stats values(Authentication.tag) as tag, values(Authentication.app) as app,
count(eval('Authentication.action'=="failure")) as failure,
count(eval('Authentication.action'=="success")) as success by Authentication.src
| `drop_dm_object_name("Authentication")`
| search success>0
| xswhere failure from failures_by_src_count_1h in authentication is above medium
| `settags("access")`
```

Detecting IDS evasion with abnormal TTL

- Count of TTL by src, dest in an day, Gather Stats
- ```
  | tstats max("All_Traffic.ttl") AS "Max of ttl" min("All_Traffic.ttl") AS "Min of ttl" median("All_Traffic.ttl") AS "Median of ttl" count("All_Traffic.ttl") AS "Count of ttl" from datamodel=Network_Traffic where (nodename = All_Traffic) groupby "All_Traffic.ttl" "All_Traffic.src"  "All_Traffic.dest" prestats=true
  | eval "All_Traffic.src ::: All_Traffic.dest"='All_Traffic.src' + " ::: " + 'All_Traffic.dest'  "ttl"='All_Traffic.ttl'
  | xsCreateDDContext app="SA-Network" name=ttlvalues_by_src_dest_count_1d container=authentication scope=app type=domain terms=`xs_default_magnitude_concepts`
  | stats count
  ```
  ```

Users with abnormal DLP activity

1. Create the Data-Driven Context | xscreate**dd**context

- sourcetype=dlp | bin span=1d _time | stats count AS dlp_signature_user_count_1d by user,signature, _time | where dlp_signature_user_count_1d > 0
- | stats count(dlp_signature_user_count_1d) as count median(dlp_signature_user_count_1d) as median stdev(dlp_signature_user_count_1d) as size by user, signature | eval size=if(size<1,1,size)
- | xscreateddcontext name= dlp_signature_user_count_1d type=median_centered terms="low,expected,high" scope=app class="user,signature" container=all_insider_models_count_1d

Schedule this for moving 60 day window

2. Compare New Events to Context | xswhere

```
sourcetype=dlp | bin span=1d _time
| stats count as dlp_signature_user_count_1d by user , signature,_time
| xswhere dlp_signature_user_count_1d in all_insider_models_count_1d by user NOT expected
```

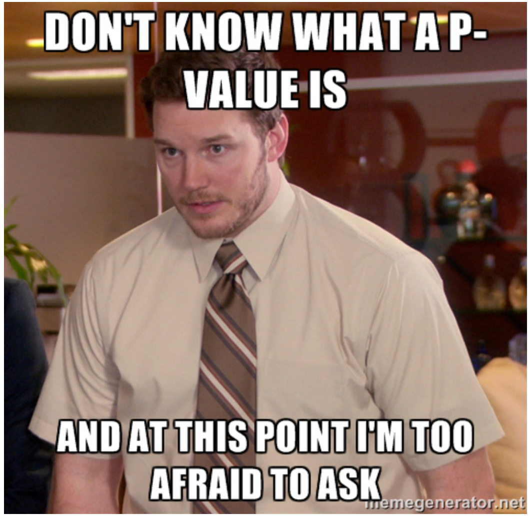
Show a dashboard of unusual events.

Static to Dynamic thresholds

- Extreme Search Examples
  √ Authentication Analysis
  √ Network Analysis
  √ DLP Analysis
- Exploratory Data Analysis Examples
  – Descriptive Statistics + Moving Window = Context
  – Visualization
  – Correlation
  – Machine Learning as EDA

Descriptive Statistics & EDA

- In high school math you learned about **mean, mode, median, min, max**, & **frequency aka "Descriptive Statistics"**.
- You should make use of these to describe the data you are looking at and explore the relationships within your data set.
- **This iterative process is called "Exploratory Data Analysis".**



Descriptive Statistics & EDA

- **## Compare different duration times of data set for a specific time period.**
- ```
  index=suricata event_type=flow
  | stats count as number_events, min(duration) as min_duration, max(duration) as max_duration, avg(duration) as avg_duration, median(duration) as median_duration, perc95(duration) as perc95_duration, stdev(duration) as stdev_duration
  ```
- Are there any long running sessions in the last 60 minutes?

number_events	min_duration	max_duration	avg_duration	median_duration	perc95_duration	stdev_duration
3397	0	3654	14.274948	0	60	78.859433
  ```

Descriptive Statistics - PCR

- Make use of eval to determine network flows or Producer Consumer Ratio (PCR)
- **## Create a ratio of bytes_in to bytes_out**
-
```
index=suricata event_type=flow
| eval bytes_total=bytes_in+bytes_out
| eval bytes_ratio= ((bytes_out-bytes_in)/bytes_total)
| iplocation dest_ip
| table src_ip src_port dest_ip dest_port bytes_in bytes_out bytes_total bytes_ratio
| sort - bytes_ratio
```

- **## Apply case logic to determine inbound or outbound Imbalance between client & server**
-
```
index=suricata event_type=flow
| eval bytes_total=bytes_in+bytes_out
| eval bytes_ratio= ((bytes_out-bytes_in)/bytes_total)
| eval bytes_pcr_range = case(bytes_ratio > 0.4  "Pure Push", bytes_ratio > 0  "70:30 Export", bytes_ratio == 0  "Balanced Exchange", bytes_ratio >= -0.5  "3:1 Import", bytes_ratio > -1  "Pure Pull"
| stats sparkline(count) AS activity by src_ip src_port dest_ip dest_port bytes_in bytes_out bytes_pcr_range
```

Descriptive Statistics - PCR

- Make use of eval to determine network flows or Producer Consumer Ratio (PCR)
- **## Create a ratio of bytes_in to bytes_out**
-

src_ip	src_port	dest_ip	dest_port	bytes_in	bytes_out	bytes_total	bytes_ratio
45.79.169.212	52670	66.228.42.5	53	267	172	439	-0.216401
45.79.169.212	37633	50.116.53.5	53	267	172	439	-0.216401
45.79.169.212	50577	96.126.106.5	53	267	172	439	-0.216401
45.79.169.212	48005	66.228.42.5	53	267	172	439	-0.216401
45.79.169.212	43693	50.116.53.5	53	267	172	439	-0.216401
45.79.169.212	39674	96.126.106.5	53	267	172	439	-0.216401
45.79.169.212	36898	96.126.106.5	53	267	172	439	-0.216401
45.79.169.212	56891	66.228.42.5	53	267	172	439	-0.216401
45.79.169.212	55740	66.228.42.5	53	267	172	439	-0.216401
45.79.169.212	35877	50.116.53.5	53	267	172	439	-0.216401

« prev 1 2 3 4 5 6 7 8 9 10 next »

- **## Apply case logic to determine inbound or outbound imbalance between client & server**

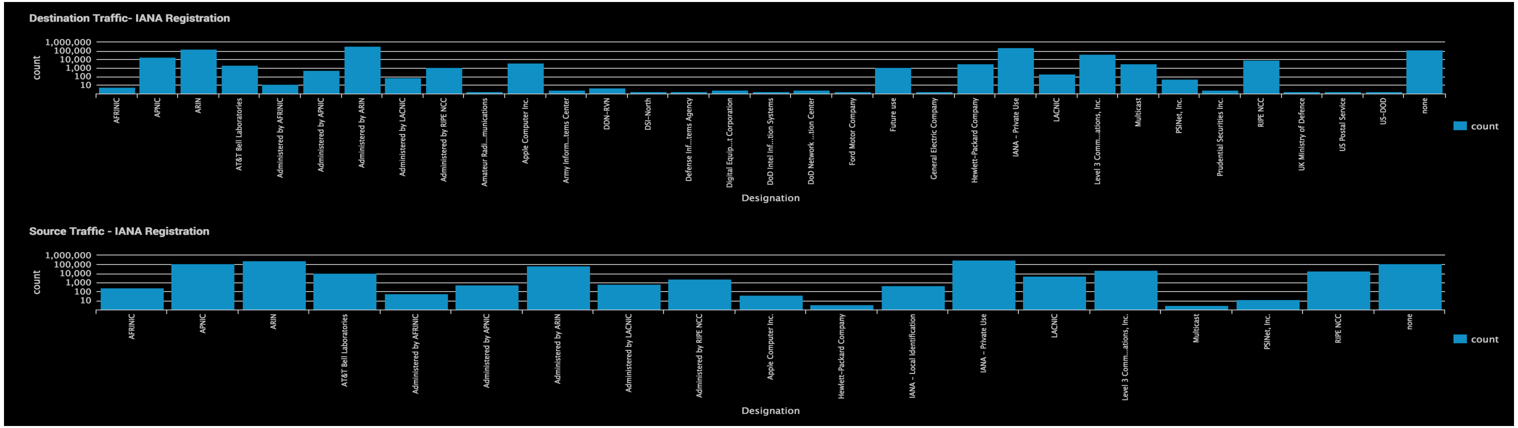
Data Exfiltration, PCR Categories

src_ip	src_port	dest_ip	dest_port	bytes_in	bytes_out	bytes_pcr_range	activity
1.196.57.52	11595	45.79.169.212	23	54	74	70:30 Export	
1.34.249.55	57909	10.10.0.5	23	54	56	70:30 Export	
10.0.0.3	49488	131.253.34.234	443	5860	7253	70:30 Export	
10.0.0.3	49490	65.52.108.231	443	5904	7626	70:30 Export	
10.0.0.3	49491	65.52.108.254	443	4436	3753	3:1 Import	
10.0.0.3	49492	65.52.108.213	443	5283	5724	70:30 Export	
10.0.0.3	49493	131.253.34.230	443	4436	3753	3:1 Import	
10.0.0.3	49495	131.253.34.230	443	4436	3753	3:1 Import	
10.0.0.3	49782	75.75.75.75	53	210	82	3:1 Import	
10.0.0.3	50185	75.75.75.75	53	255	82	Pure Pull	

« prev 1 2 3 4 5 6 7 8 9 10 next »

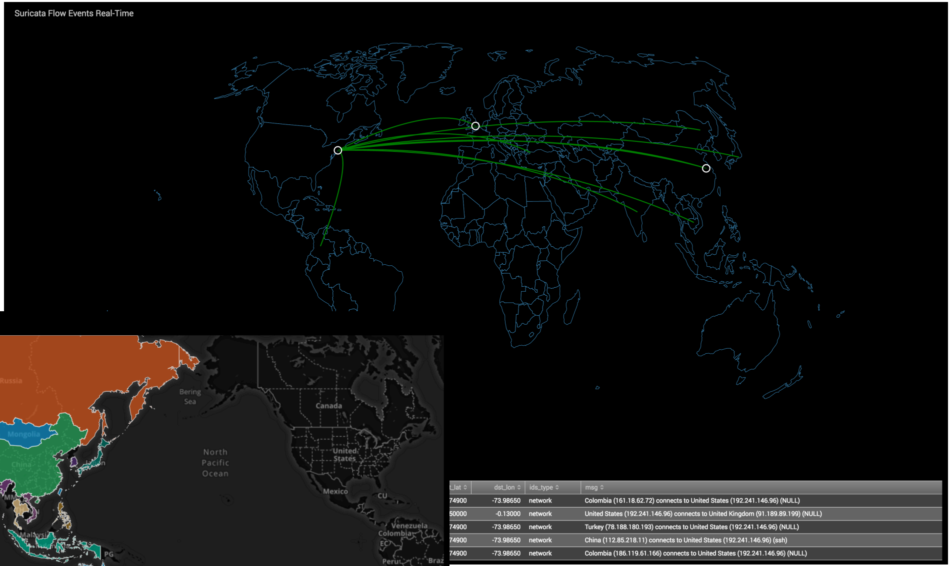
Visualization & Creating Context (EDA)

- **Visualization** is a powerful EDA tool
  - – Not everything can be described as bits, bytes, plaintext or pie charts.
- **Correlation** to add context to your data during the EDA process or test hypothesis.



Splunk Specific EDA - Visualization

- Visualization useful for exploring multi-dimensional relationships.
- Tells a story about the data you can't describe in text or tables.
- "Where are connections 'originating', and how often am I seeing this activity?"



_lat	dst_lon	ids_type	msg
74900	-73.98650	network	Colombia (161.18.62.72) connects to United States (192.241.146.96) (NULL)
50000	-0.13000	network	United States (192.241.146.96) connects to United Kingdom (91.189.89.199) (NULL)
74900	-73.98650	network	Turkey (78.188.180.193) connects to United States (192.241.146.96) (NULL)
74900	-73.98650	network	China (112.85.218.11) connects to United States (192.241.146.96) (ssh)
74900	-73.98650	network	Colombia (186.119.61.166) connects to United States (192.241.146.96) (NULL)

- I don't remember hiring any remote employees in China.

Splunk Correlation as EDA

- CSV/KV Lookups – Threat Intelligence, Known bad configurations
- ## Search for SSL connections with insecure cipher (key less than 128) to adversarial countries
-
```
index=bro sourcetype=bro_ssl
| lookup insecure_ciphers cipher OUTPUT reason_insecure
| search reason_insecure!="" | iplocation src_ip prefix=src_ | iplocation dest_ip prefix=dest_
| lookup adversaries country AS dest_Country OUTPUT isAdversary | search isAdversary=TRUE
| stats sparkline(count) AS activity count by src_ip dest_ip dest_Country
| sort - count
```
- Python Lookups - Entropy Analysis of DNS / HTTP
- # Full Query for Suricata HTTP
-
```
index=suricata host=suricata event_type=http
| lookup ut_parse_extended_lookup url AS dest
| lookup ut_shannon_lookup word AS ut_subdomain OUTPUT ut_shannon AS ut_shannon_subdomain
| lookup ut_shannon_lookup word AS dest OUTPUT ut_shannon AS ut_shannon_dest | search ut_shannon_dest > 4 OR
ut_shannon_subdomain > 4
| table ut_subdomain ut_shannon_subdomain dest ut_shannon_dest
| dedup dest ut_subdomain
```

Splunk Correlation as EDA

- **CSV/KV Lookups – Threat Intelligence, Known bad configurations**
- **## Search for SSL connections with insecure cipher (key less than 128) to adversarial countries**

SSL Traffic to Adversarial Countries

src_ip	dest_ip	dest_Country	reason_insecure	activity	count
2601:243:c300:f460:557c:30e1:e35:bfe6	2a01:111:f330:1790::a01	China	uses RC4 which has insecure biases in its output		3
2601:243:c300:f460:8f9:9b39:5abe:6386	2a01:111:f330:1790::a01	China	uses RC4 which has insecure biases in its output		1
2601:243:c300:f460:b8d5:147f:1c0d:1fee	2a01:111:f330:1790::a01	China	uses RC4 which has insecure biases in its output		1
2601:243:c300:f460:ddf3:1869:18c8:c939	2a01:111:f330:1790::a01	China	uses RC4 which has insecure biases in its output		1
2601:243:c300:f460:f474:17cf:f113:d3b0	2a01:111:f330:1790::a01	China	uses RC4 which has insecure biases in its output		1

- **Python Lookups - Entropy Analysis of DNS / HTTP**
- **# Full Query for Suricata HTTP**

Subdomain & Domain Entropy Scoring

ut_subdomain	ut_shannon_subdomain	dest	ut_shannon_dest
ic.49f66b73.141b5c.1.msxbassets.loris	4.1086680695965025	ic.49f66b73.141b5c.1.msxbassets.loris.llnwd.net	4.288082736032309
ic.49f66b73.13d264.1.msxbassets.loris	4.1831244885738945	ic.49f66b73.13d264.1.msxbassets.loris.llnwd.net	4.304144172248552
ic.49f66b73.020b6e.1.msxbassets.loris	4.162722123650557	ic.49f66b73.020b6e.1.msxbassets.loris.llnwd.net	4.314574491305427
ic.49f66b73.0cdf21.1.xboxone.loris	4.19438848899739	ic.49f66b73.0cdf21.1.xboxone.loris.llnwd.net	4.279519187707896
ic.49f66b73.0fd207.1.xboxone.loris	4.194388488997389	ic.49f66b73.0fd207.1.xboxone.loris.llnwd.net	4.279519187707896
srv-2016-07-31-21.pixel	3.7950885863977324	srv-2016-07-31-21.pixel.parsely.com	4.229003731107054
d1ai9qtk9p41kl	3.378783493486176	d1ai9qtk9p41kl.cloudfront.net	4.142295219190902
srv-2016-07-31-21.config	3.8868421881310122	srv-2016-07-31-21.config.parsely.com	4.350209029099896
d2b3uqm49lqeua	3.521640636343319	d2b3uqm49lqeua.cloudfront.net	4.142295219190901
async-lb-2129785755.us-east-1.elb	4.028946391954607	async-lb-2129785755.us-east-1.elb.amazonaws.com	4.270237192601036

« prev 1 2 3 4 5 6 7 8 9 10 next »

Machine Learning Toolkit as EDA

- **Using CSV of Known Cloud Providers, Python Lookup to calculate entropy**
- **## Search for http requests where the subdomain or domain have a high level of entropy, overlay CDN domains**
-

```
index=suricata host=suricata event_type=http
| lookup ut_parse_extended_lookup url AS dest
| lookup ut_shannon_lookup word AS ut_subdomain OUTPUT ut_shannon AS ut_shannon_subdomain
| lookup ut_shannon_lookup word AS dest OUTPUT ut_shannon AS ut_shannon_dest | search ut_shannon_dest > 4 OR
ut_shannon_subdomain > 4
| lookup cloud_providers domain AS ut_domain OUTPUT CDN_provider isProvider
| fillnull value=False
| table ut_subdomain ut_shannon_subdomain dest ut_shannon_dest isProvider
```

- **# Search for categorical outliers based on src_ip, dest_ip, total_bytes, and bytes_ratio**
-

```
index=suricata event_type=flow | eval bytes_total=bytes_in+bytes_out
| eval bytes_ratio= ((bytes_out-bytes_in)/bytes_total)
| iplocation dest_ip
| table src_ip src_port dest_ip dest_port bytes_in bytes_out bytes_total bytes_ratio
```

Field(s) to analyze

✖ src_ip ✖ dest_ip ✖ bytes_total ✖ bytes_ratio

Machine Learning Toolkit as EDA

- **Using CSV of Known Cloud Providers, Python Lookup to calculate entropy**
- **## Search for http requests where the subdomain or domain have a high level of entropy, overlay CDN domains**

Prediction Results

isProvider	predicted(isProvider)	ut_shannon_dest	ut_shannon_subdomain
True	True	4.40385618977	3.0
True	True	4.40385618977	3.0
True	True	4.18670434591	2.32192809489
True	True	4.40385618977	3.0
True	True	4.40385618977	3.0
True	True	4.18670434591	2.32192809489
True	False	4.49223560048	4.28183553261
True	True	4.18670434591	2.32192809489
False	True	4.06026203912	3.45281953111
False	True	4.06026203912	3.45281953111

« prev 1 2 3 4 5 6 7 8 9 10 next »

Open in Search | Show SPL | Schedule Alert

Precision	Recall	Accuracy	F1
0.73	0.71	0.71	0.72

Open in Search | Show SPL

Classification Results (Confusion Matrix)

Predicted actual	Predicted False	Predicted True
False	385 (82.1%)	84 (17.9%)
True	199 (38.5%)	318 (61.5%)

Open in Search | Show SPL

- **# Search for categorical outliers based on src_ip, dest_ip, total_bytes, and bytes_ratio**

Outlier(s)

27

Outlier(s)

Open in Search | Show SPL | Schedule Alert

Total Event(s)

10,000

Total Event(s)

Open in Search | Show SPL

Data and Outliers

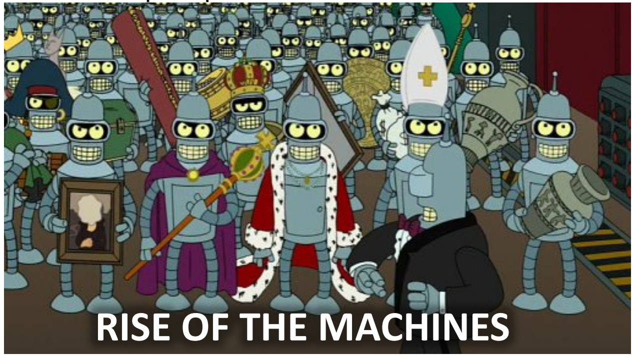
src_ip	dest_ip	bytes_total	bytes_ratio	probable_cause	isOutlier
10.0.0.27	17.253.25.207	40261502	-0.995156	**bytes_ratio**	1
2601:0243:c300:f460:4802:d5f8:1ecd:2f0f	2607:f8b0:4001:0c20:0000:0000:0000:0080	16799376	-0.977176	**bytes_total**	1
2601:0243:c300:f460:f54a:ff88:8566:1fd5	2607:f8b0:4009:080c:0000:0000:0000:2011	73107456	-0.940812	**bytes_total**	1
2601:0243:c300:f460:4816:0ab1:0a6d:9722	2607:f8b0:4009:080c:0000:0000:0000:2011	17050275	-0.926762	**bytes_total**	1
118.92.6.188	10.0.0.21	67785450	0.952324	**bytes_total**	1
178.84.177.11	10.0.0.21	331180614	0.957080	**bytes_total**	1
93.142.31.32	10.0.0.21	195965744	0.956286	**bytes_total**	1
10.0.0.21	79.146.194.65	250289331	0.831491	**bytes_total**	1
104.244.251.226	10.0.0.21	195618306	0.956896	**bytes_total**	1
2601:0243:c300:f460:4112:73b6:fdb4:384c	2605:fe80:2100:a001:0001:0000:0000:0001	26164748	-0.981259	**bytes_total**	1

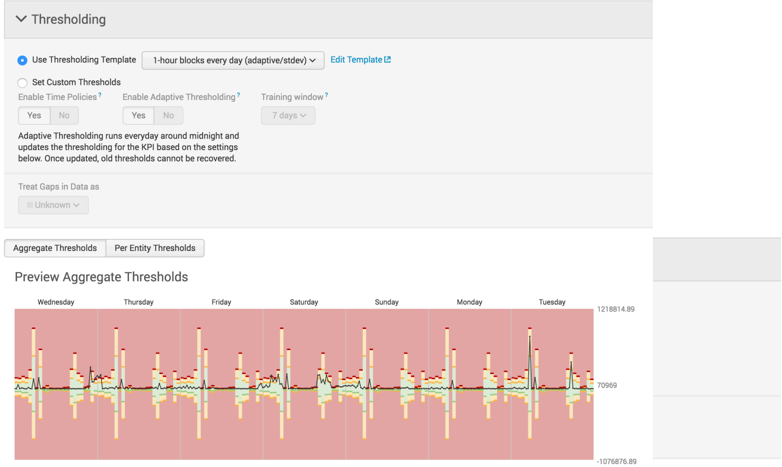
« prev 1 2 3 4 5 6 7 8 9 10 next »

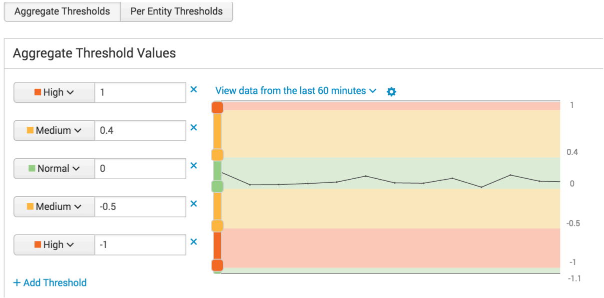
Open in Search | Show SPL | Schedule Alert

Descriptive Statistics & ML - ITSI

- Make use of eval to create bytes_total & bytes_ratio for Producer Consumer Ratio (PCR) for KPI Base Search & NetFLOW
- ```
  index=suricata event_type=flow
  | eval bytes_total=bytes_in+bytes_out
  | eval bytes_ratio= ((bytes_out-bytes_in)/bytes_total)
  ```
- Thresholding score compares the current traffic against a rolling hourly average and standard deviation from mean for last 30 days of data.
- Bytes Ratio Thresholds based on PCR Static Ratios
  – 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

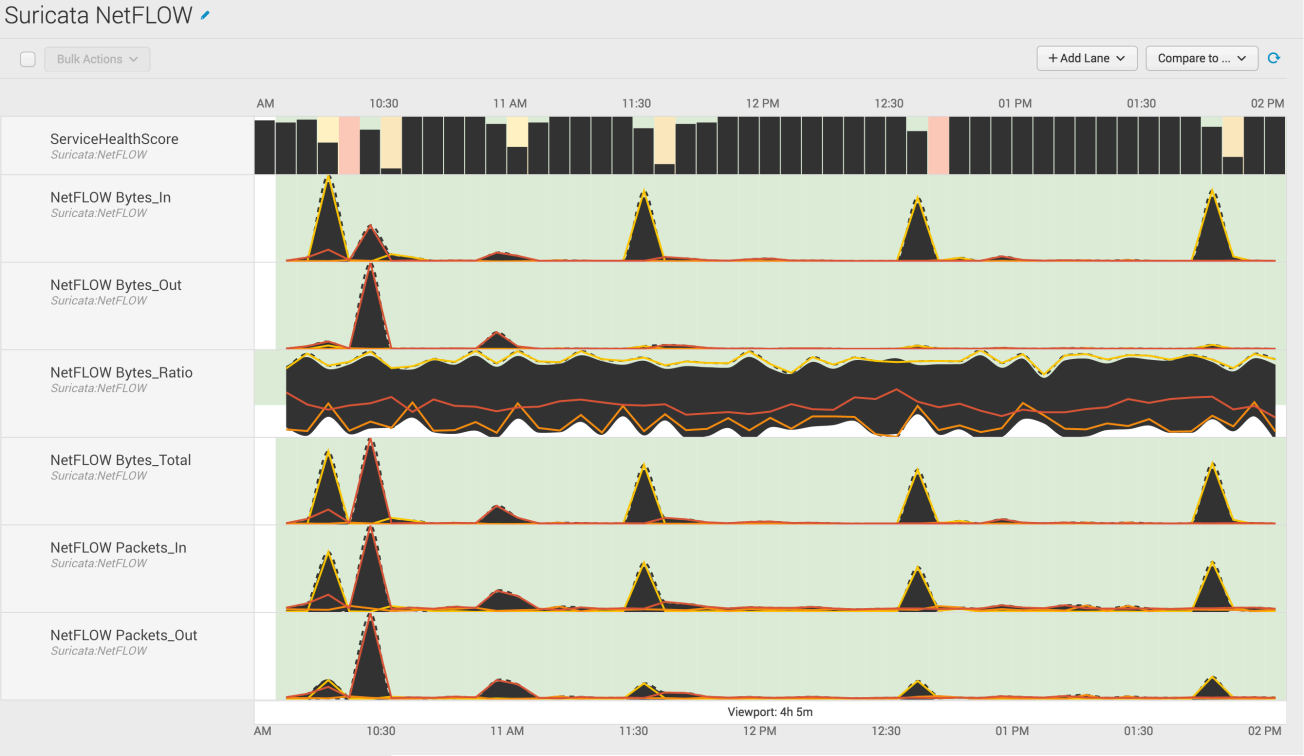


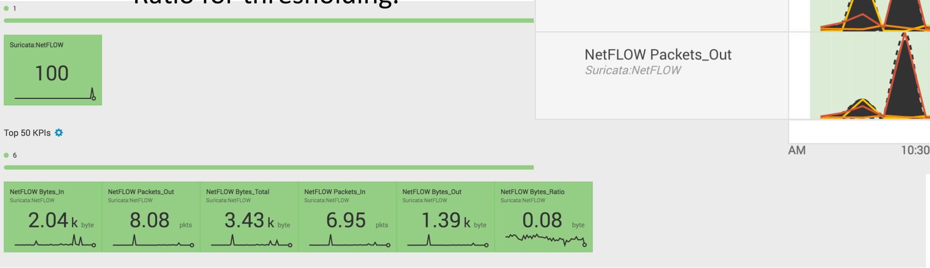



  ```

Descriptive Statistics & ML - ITSI

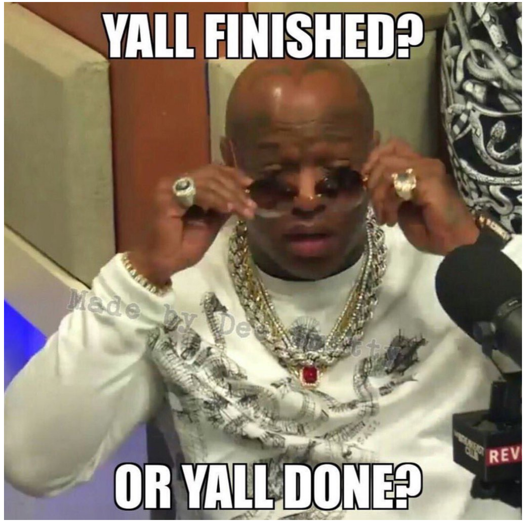
- Visualization of the same PCR Suricata Flow data using ITSI
- Health score based on 5 KPIs. The current traffic (bytes_in, bytes_out, bytes_total, packets_in, & packets_out) compared to a rolling hourly average, and standard deviation from mean.
- Attempting to define ”What is normal and when is something deviating from the norm I’ve seen for 30 days?”
- Bytes Ratio based on PCR Static Ratio for thresholding.





Recap

- √ 5 Step Data Science Methodology for Security
- √Descriptive Statistics
- √Quantitative vs Qualitative Analysis
- √Exploratory Data Analysis (EDA)
- √Explore native/ Add-on Splunk analytic capabilities



THANK YOU

Explore Splunk Analytics

- Anomalies
  - Analyzes numeric fields for their ability to predict another discrete field.
- Anomalousvalue
  - Computes an "unexpectedness" score for an event.
- Anomalydetection
  - Finds and summarizes irregular, or uncommon, search results.
- Cluster
  - Computes a probability for each event and detects unusually small probabilities.
- Kmeans
  - Groups similar events together.
- Outlier
  - Removes outlying numerical values.
- Rare
  - Displays the least common values of a field.

Glossary

- Descriptive Statistics
  - Min, Max, Median, Average(Mean), Standard Deviation, Mode
  - Z-Scores
- Exploratory Data Analysis
  - Searching the data and looking for relationships
  - Leveraging knowledge (lookups , reference tables)
- Entropy
  - Measurement of how mixed up something is
    - e.g. non-numerical field such as query compared against wordlist
- P-Values
  - "Captures the probability of observing the data you've observed"
- Linear Regression

References & Resources

- Doing Data Science http://www.tylervigen.com/spurious-correlations
- PCR – A New Flow Metric http://qosient.com/argus/presentations/Argus.FloCon.2014.PCR.Presentation.pdf
- Data Driven Security http://datadrivensecurity.info/
- Splunk Syntax Highlighting http://blog.metasyn.pw/splunk-syntax-highlighting/
- Doing Data Science http://shop.oreilly.com/product/0636920028529.do
- Hunting the Known Unknowns (with DNS) https://conf.splunk.com/speakers/2015.html#search=Kovar&
- Lookups, and other goodies https://github.com/anthonygtellez/conf2016_extras
- IDS Evasion w TTL - http://insecure.org/stf/secnet_ids/secnet_ids.html