Add Value to your SIEM: How Israel's Ministry of Energy applies Machine Learning to protect their Critical Infrastructure and OT Operations

Philipp Drieger
Staff Machine Learning Architects | Splunk

Eurus Kim
Staff Machine Learning Architects | Splunk

Israel's Ministry of Energy CERT Team

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans 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 may differ materially. 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, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

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 functionalities described or to include any such feature or functionality in a future release.

Philipp Drieger
Staff Machine Learning Architect | Splunk

Eurus Kim
Staff Machine Learning Architect | Splunk

splunk> .conf19

Agenda

splunk> turn data into doing™

Agenda

Introduction

- Israel's Ministry of Energy Cyber Security Center
- Challenges with OT Security
- The Splunk Journey: from _raw to CIM

Getting Value from Machine Learning

- How to detect anomalies across facilities and get meaningful alerts
- Example Use Case 1: Authentication Anomalies
- Example Use Case 2: Anomalous RDP Connections

Deep Dive into Density Function

- What's the math behind it?
- How to apply it and lessons learnt (tuning for environment, cardinality, quality)

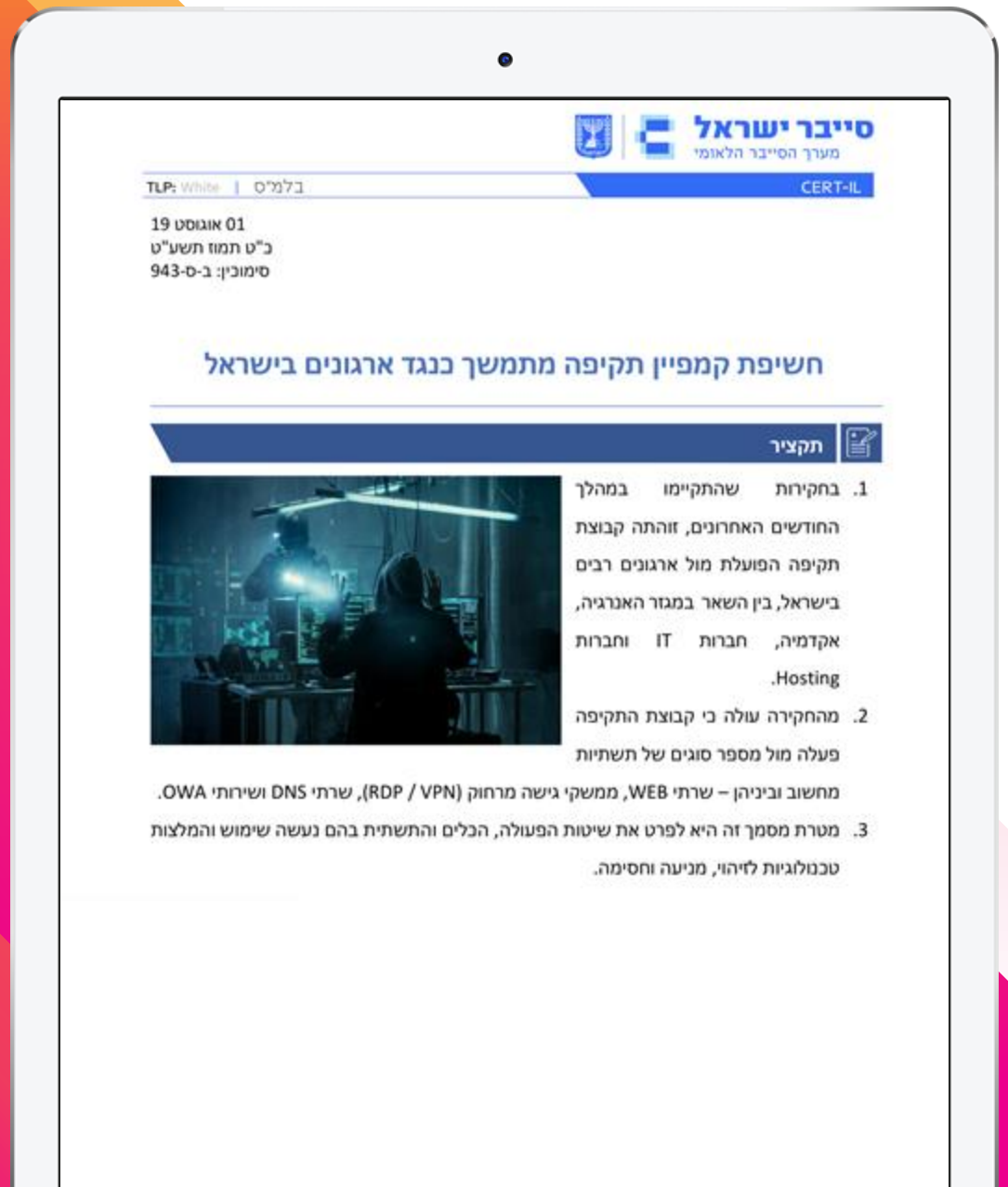
Wrap up and outlook

Introduction

splunk> .conf19

Ministry of Energy

www.energy.gov.il



Cyberattacks Don't Stop at Critical Infrastructure

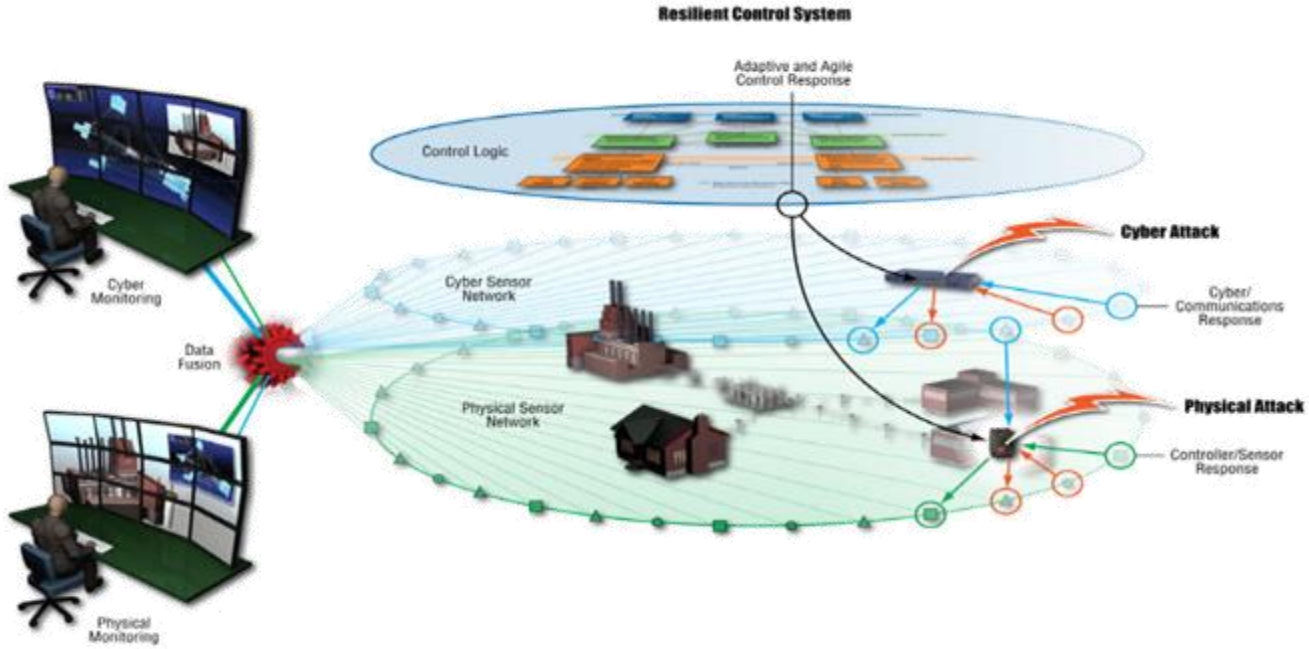
Just recently a threat activity group targeted many organizations in Israel in the energy sector, academia and hosting companies.

IT infrastructure, web servers, RDP/VPN used for remote access and more were under attack.

splunk> .conf19

Israel Ministry of Energy Cyber Security Center

- Generate sector-wide security posture and resilience status
- Provide a safety net, primarily focus on the private sector

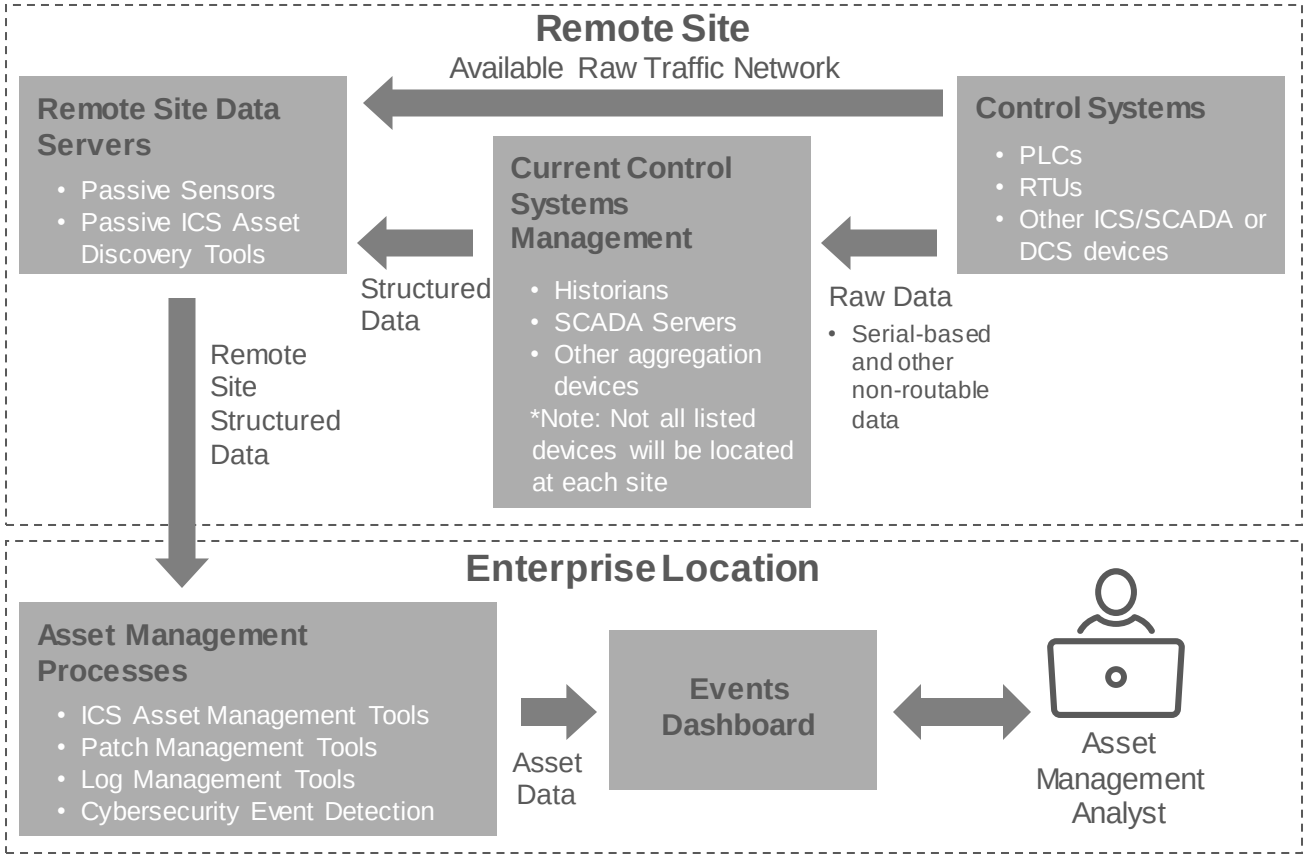


NIST National Cybersecurity Center of Excellence

Energy sector asset management for electric utilities, oil & gas industry

Scope:

- **Asset Discovery:** establishment of a full baseline of physical and logical locations of assets
- **Asset Identification:** capture of asset attributes, such as manufacturer, model, operating system (OS), Internet Protocol (IP) addresses, Media Access Control (MAC) addresses, protocols, patch-level information, and firmware versions
- **Asset Visibility:** continuous identification of newly connected or disconnected devices, and IP (routable and non-routable) and serial connections to other devices
- **Asset Disposition:** the level of criticality (high, medium, or low) of a particular asset, its relation to other assets within the OT network, and its communication (to include serial) with other devices
- **Alerting Capabilities:** detection of a deviation from the expected operation of assets



Note: All cross-boundary network traffic uses secured communication protocols

NIST National Institute of Standards and Technology U.S. Department of Commerce

NCCoE National Cybersecurity Center of Excellence

PDF

splunk> .conf19

The challenges with Industrial Control Systems (ICS) and Operational Technology (OT)

splunk> .conf19

Risk Management in IT and OT

Operational Technology (OT)	Information Technology (IT)
Very often no security at all	Security by Design
Maintenance only by the vendors or approved 3rd parties – Else, warranty will void!	Available support and patches
Might find the same hardware and software for 10-15 years and more	3-5 years life cycle
Relatively fixed in order to provide greater reliability and safety – But, things are changing with IIoT	Whitelisting ? Environment will keep changing (BYOD, Mobile…)

Sources of Information

Deal with what you have!

Use logs from already installed systems (hosts, servers) and security controls (Routers, FW, AV, AppControl) to extract information like Host, IP, Last Seen

- Windows and Linux security logs from hosts and servers
- Network (Switches, Routers, Firewalls, Gateways)
- Operational history data (Historian)
- Anti-Malware
- Application Control (White listing)
- ICS IDS (This is really interesting!

Map Data Sources to CIM

"The Splunk Common Information Model (CIM) is a shared semantic model focused on extracting value from data. The CIM is implemented as an add-on that contains a collection of data models, documentation, and tools that support the consistent, normalized treatment of data for maximum efficiency at search time."

- Authentication and Network Traffic logs are a good place to start
- Authentication : Extract source and target from Interactive logon sessions or host to host/server
- Network Traffic : Extract source and target from switches, routers, gateways, firewalls (Dropped connections are helpful as well)
- Create Process (Event 4688) : Map to Endpoint data model

Data Enrichment

Manual Assets Inventory Mapping: IP, Host, Model, Version, Zone

- Extremely tedious process
- Will provide the ground truth for the asset management process

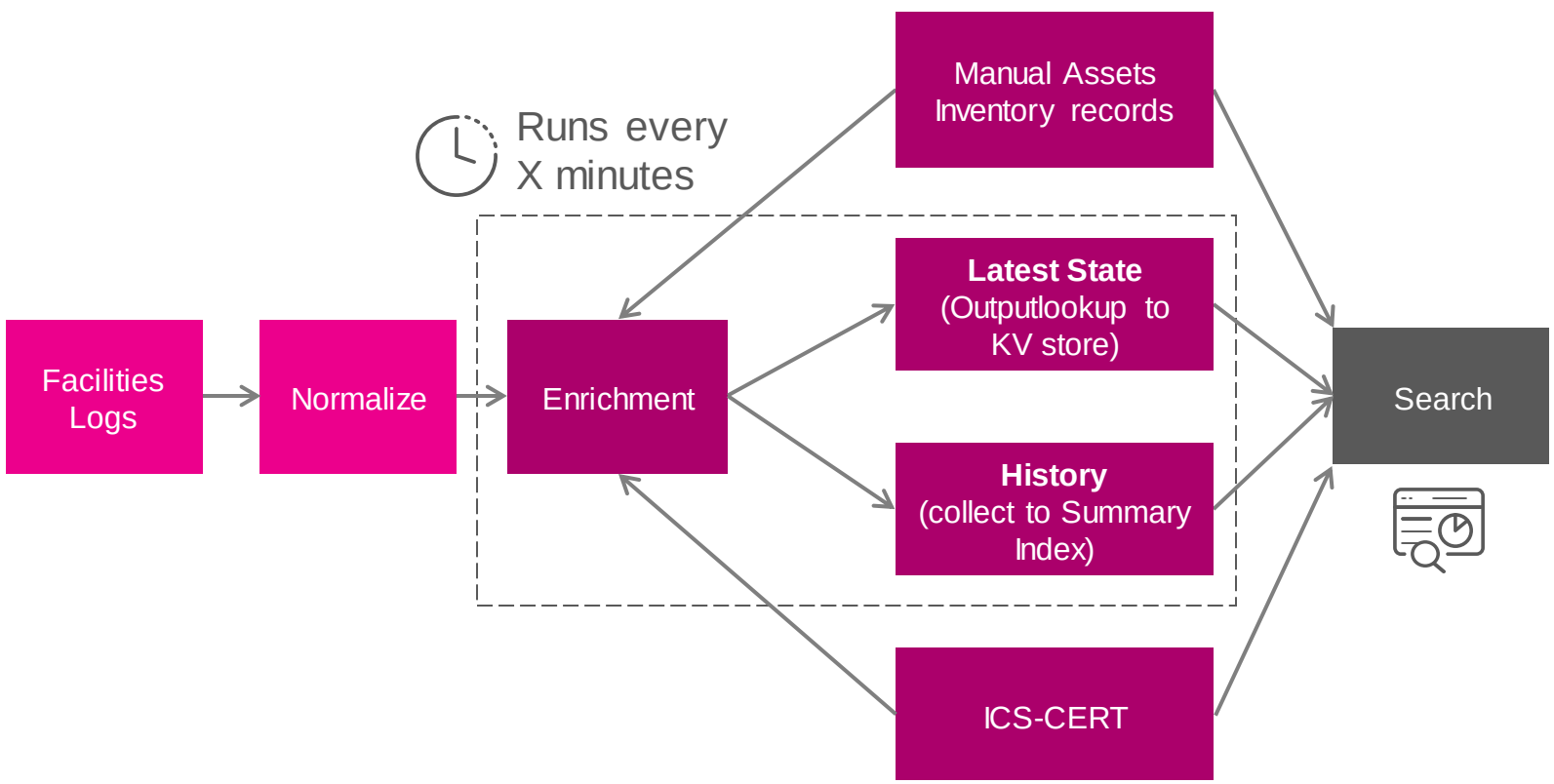
Risk Rating: NIST National Vulnerablity DB (NVD), ICS-CERT, Various vendors feeds

- Any resource for ICS/OT vulnerabilities
- Watch for the CVSS scoring – must be adapted to each facility

Device History from the Incident Management System

- What this device has been up to…

Splunk‘in it all into one data flow



What devices currently communicate ?

What devices did communicate before and stopped ?

What devices didn’t send any telemetry ?
(This is your ‘Dark Side’ or other problem)

What devices are known to be vulnerable and in what security-zone they are ?

What can I say about the device from enriching with other events ?

Getting Value from Machine Learning

splunk> .conf19

Main Goal: Detect Anomalies

Model the normal operation of each facility and spot unusual activities

- Each of the power generation facilities we are monitoring is sending us data from their OT network covering a variety of system logs and events.
- The idea is to fit a model for every facility from the historic list of events and alert on anomalous events that were detected.
- This can be further triaged by a human analyst to check whether it is indeed a cyber related notable event or some other operational issue.
- Assumption: OT events repeat in “normal” patterns of plant operations.
- CIM normalized data for authentication is available for 50+ facilities.

Authentication Anomalies

OT security use case

Identify unsolicited chatter

- Identify authentication events
- Create the baseline for each facility
- Use Splunk's MLTK to detect deviations from the learnt baseline, rogue devices and unfamiliar connections

Authentication Anomalies

Without MLTK

- Stats count of Total Authentication events by Facility
- Calculate Avg and Stdev per day (0-6) -> Create lower+upper bound (Avg +/- x*Stdev) -> Lookup per Facility
- Find anomalies
- Drill down per source and destination <- Manual Analyst work

After using Density function

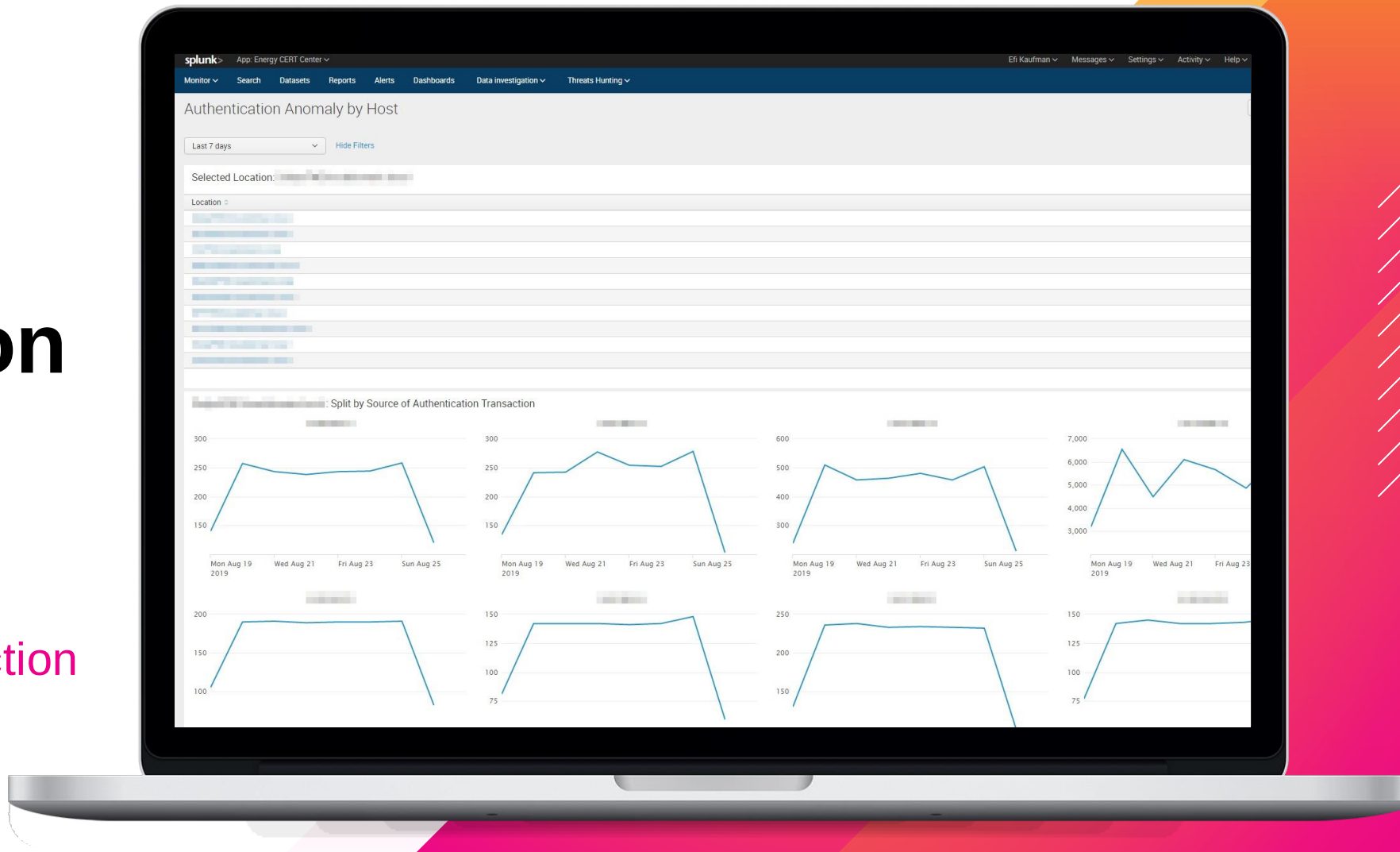
- Fit Count by (Src, Dest , Weekend Flag)
- Anomalies will highlight the specific src+dest pair!

Authentication Anomalies

Before using Density function

Authentication Anomalies

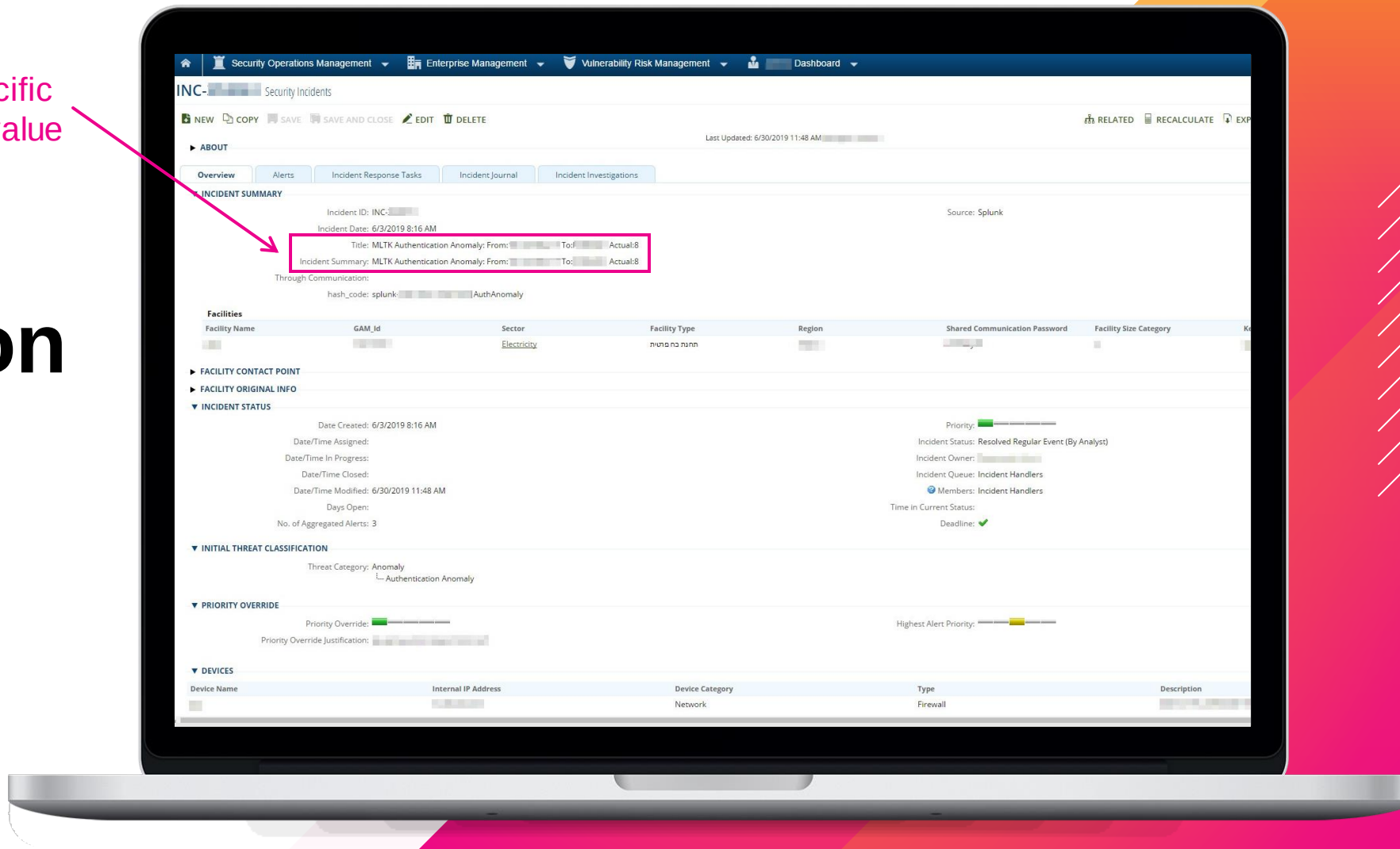
Before using Density function



Accurate alerts pointing to the specific src+dest pair and the anomalous value

Authentication Anomalies

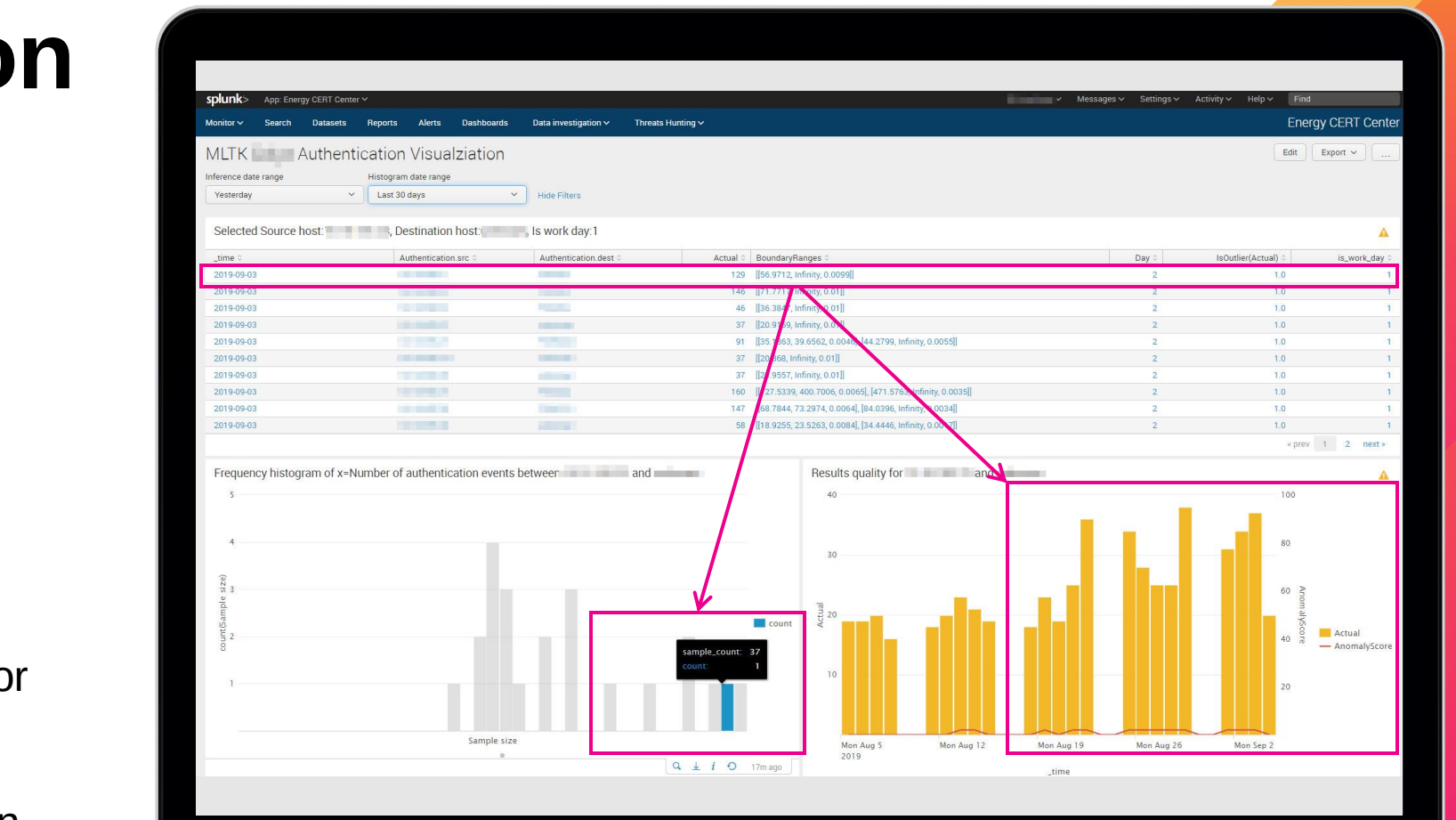
Using Density function



Authentication Anomalies

Using Density function

- All Anomalies at a glance for src+dest
- Click to drill down for details for the specific connection pair
- Investigate how Authentication behavior changed and validate why



Anomaleous RDP Connections

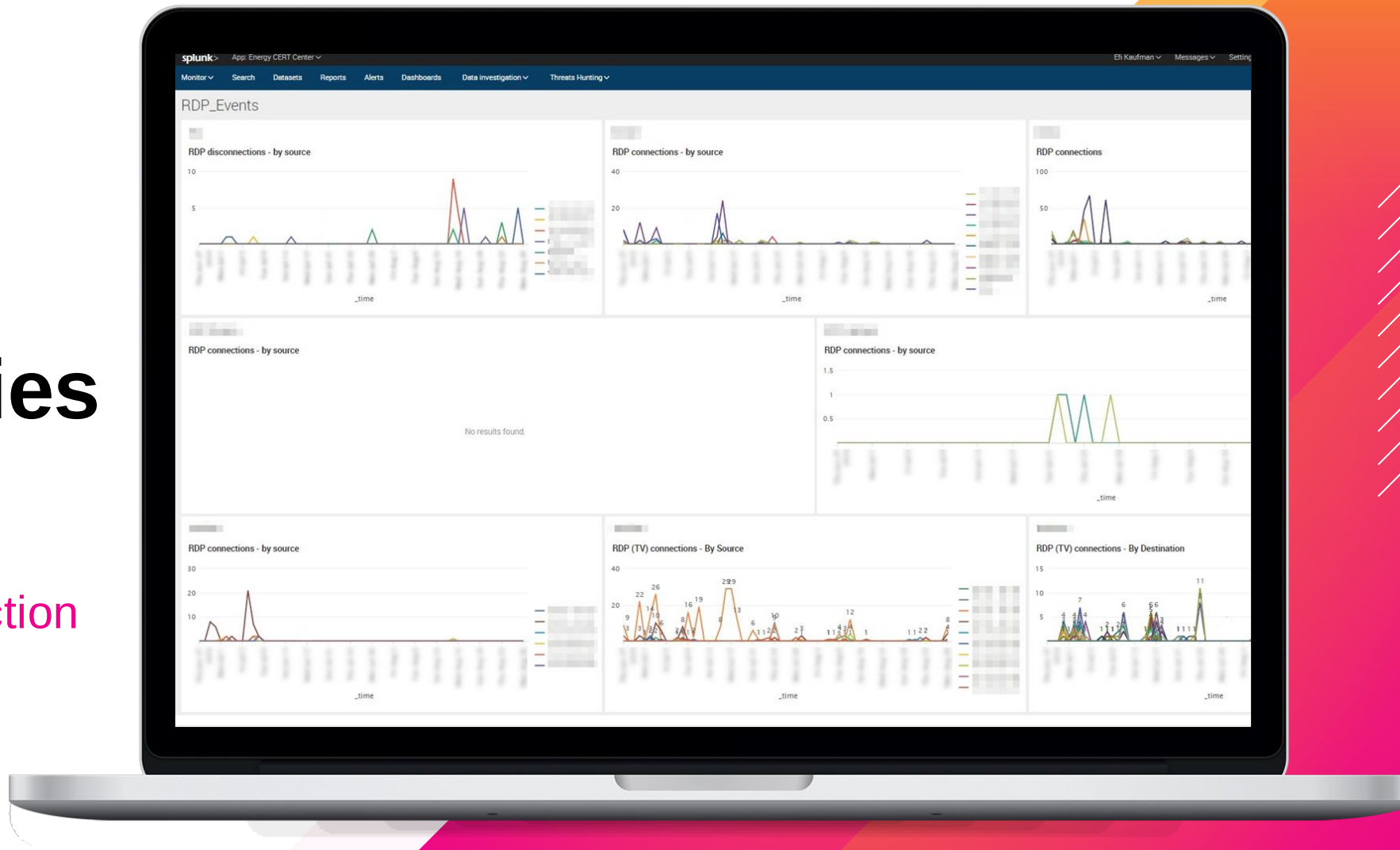
OT security use case

Who left the door open?

- Identify RDP, TeamViewer, VNC etc.
- Create the baseline for each facility
- Use Splunk's MLTK to detect unusual connection (split by Day/Night, Work days/Weekends) and connection with unusual duration

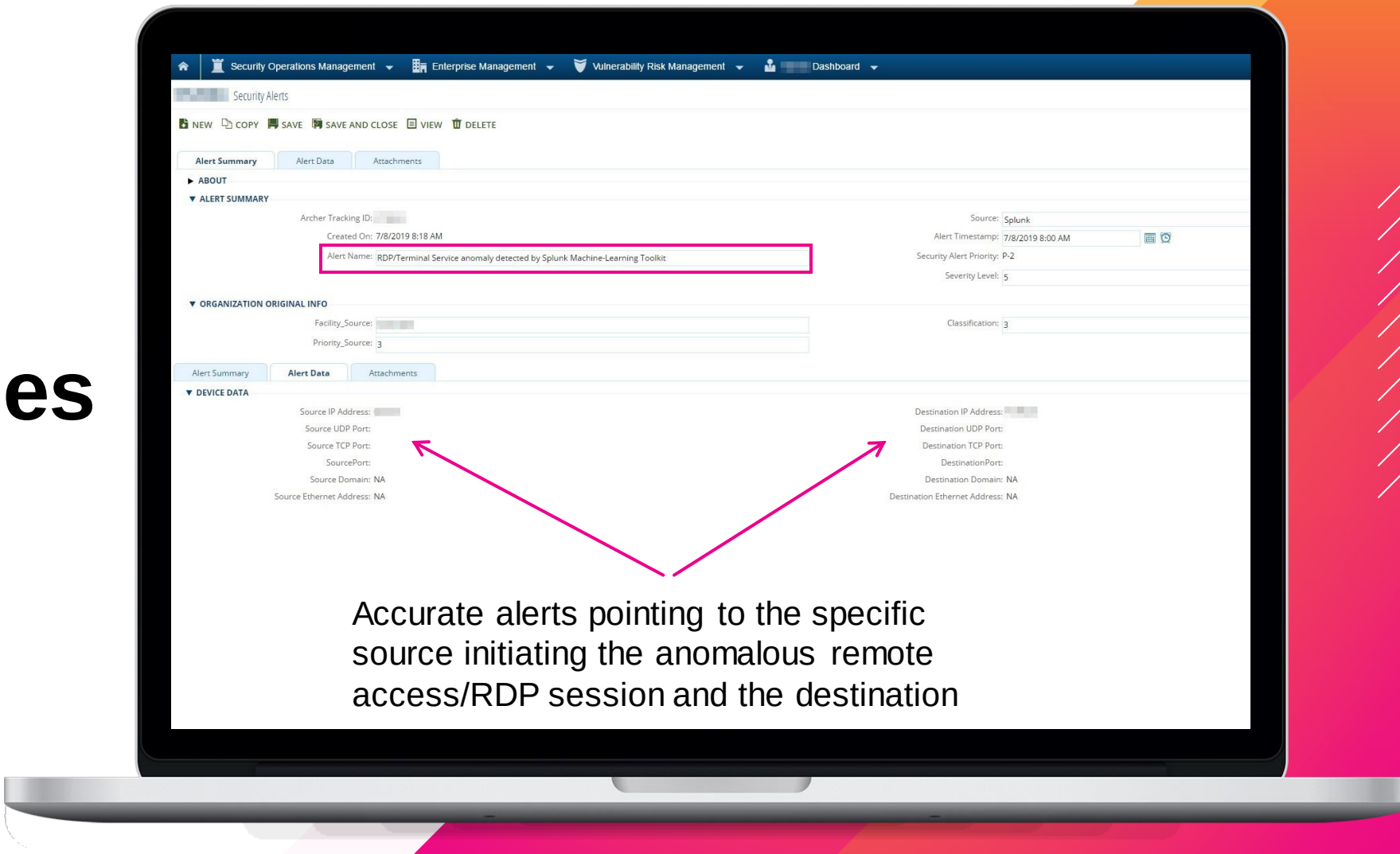
RDP Anomalies

Before using Density function



RDP Anomalies

Using Density function

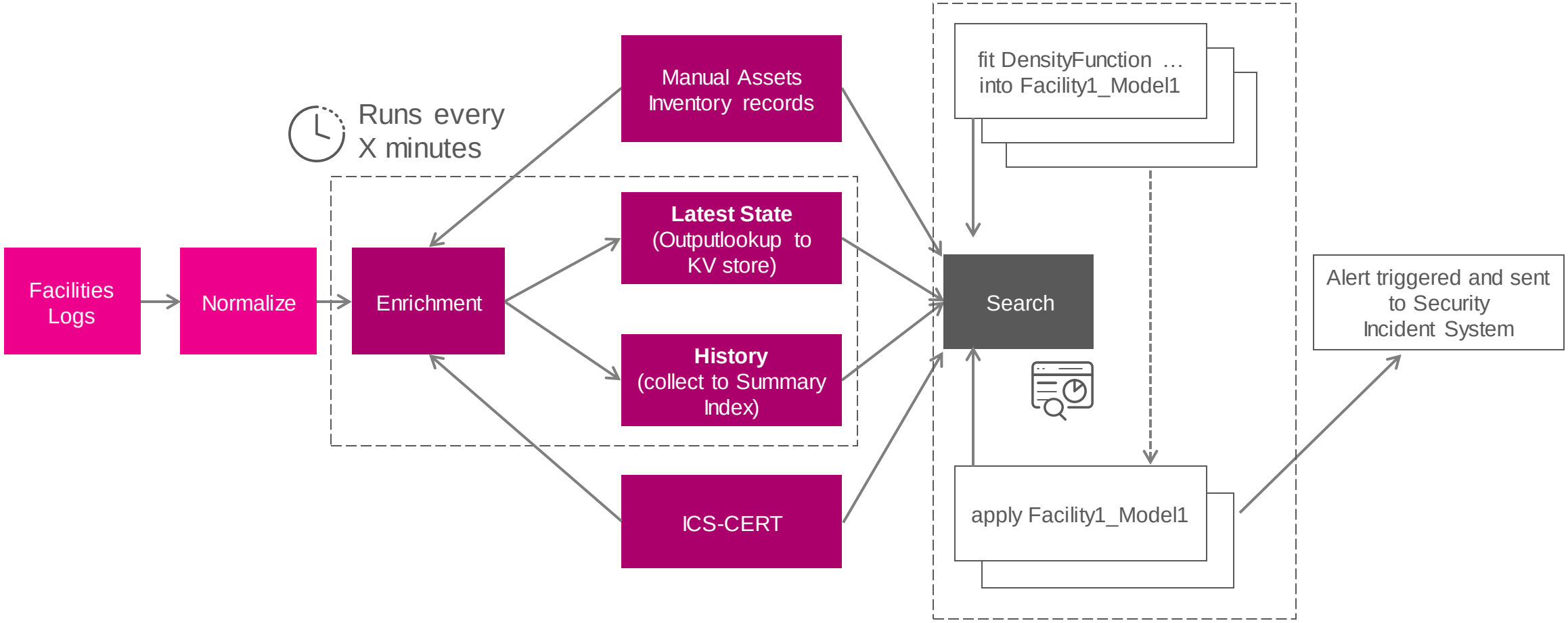


Accurate alerts pointing to the specific source initiating the anomalous remote access/RDP session and the destination

“ Because of its inconsistent nature, it was the task of human analysts to spot anomalies. With the **Density Function** our analysts now receive more **meaningful alerts** and **spend less time on tedious, manual work.**”

– Efi Kaufman, Head of Big Data and Analytics
Israel’s Ministry of Energy CERT Team

Splunk‘in it all into one data flow with ML



Deep Dive into Density Function

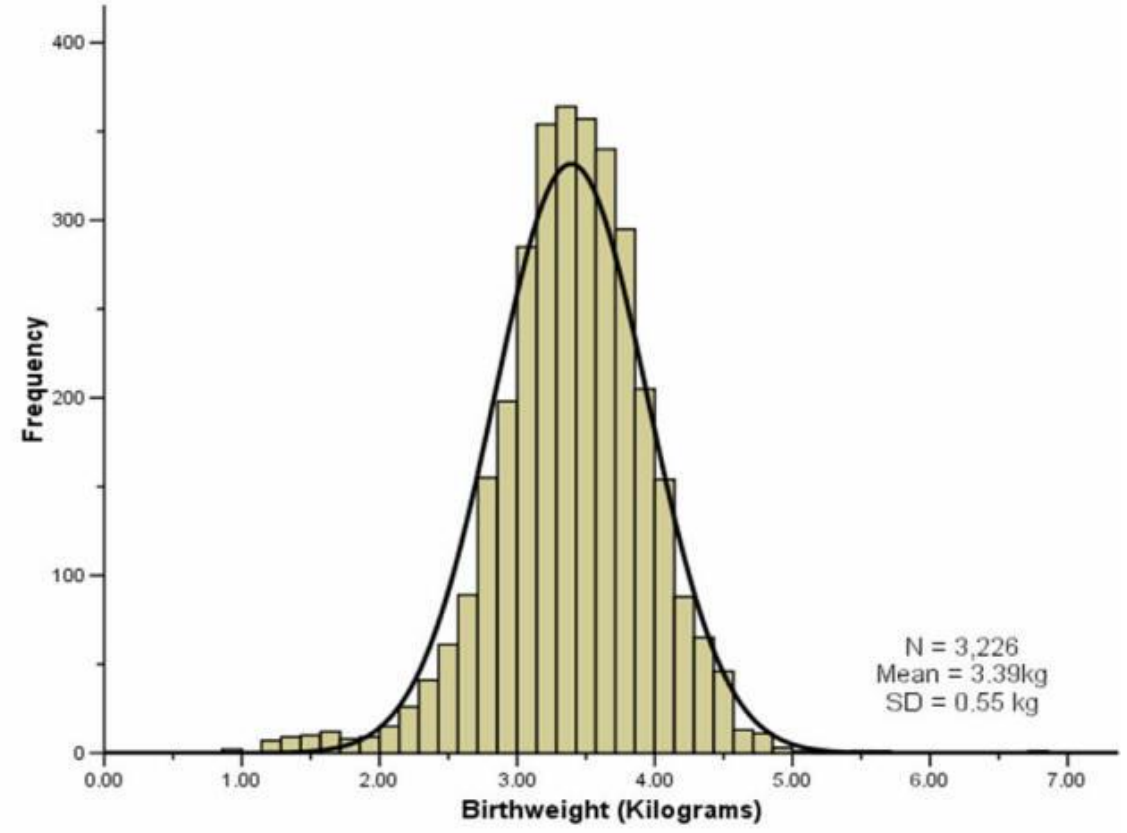
splunk> .conf19

Basic Statistics and Probability Theory

Understanding normal/gaussian distribution

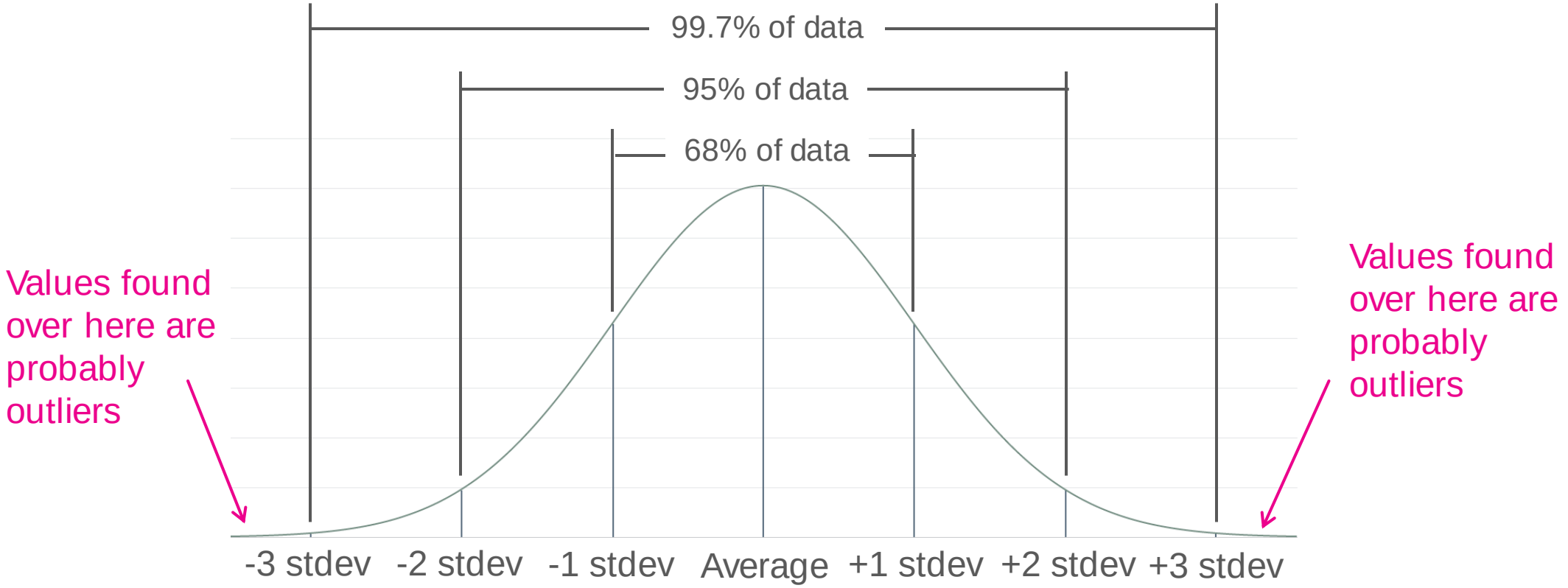
- Assumes a specific shape of your data, often used in natural and social science
- Probably commonly used because the math is pretty simple
- The assumption is your data have an average value and has a certain variance

```
| stats avg(value) stdev(value)
```



Normal Distribution

How likely is the data found within a population?



The percentages designate the area under the curve

How would you apply this in Splunk?

```
…
| eventstats avg(logins) as avg stdev(logins) as stdev
| eval lowerBound=avg-3*stdev, upperBound=avg+3*stdev
| eval isOutlier=if(logins<lowerBound OR logins>upperBound, 1, 0)
| search isOutlier=1
```

Can generate using the Detect Numeric Outliers assistant in MLTK

- Other options also available with using Absolute Mean Deviation and IQR

How would we know what is the right threshold multiplier?

How might this look with some data?

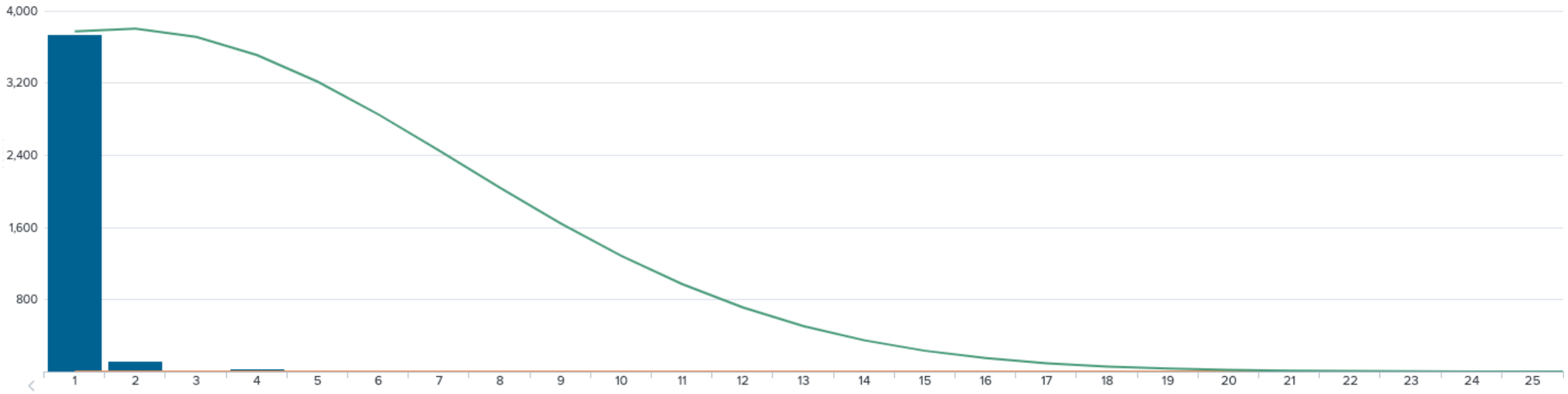
Number of logins per minute on an application (without cleaning actual outliers)

- The theoretical probability doesn't match up with actual data
- What we probably wanted is something that matches up more with theory than actual
- >10 logins per minute is probably a static threshold we may have considered
- What is a good measure for considering what's abnormal?

Above Average	1 SD	2 SD	3 SD	4 SD	5 SD	6 SD
# of logins	7.3	12.9	18.5	24.1	29.7	35.3
Theoretical % of data	16%	2.3%	0.14%	0.003%	0.00003%	0.0000001%
Actual % of data	2.9%	2.0%	0.8%	0.8%	0.5%	0.2%

Why doesn't this work so well?

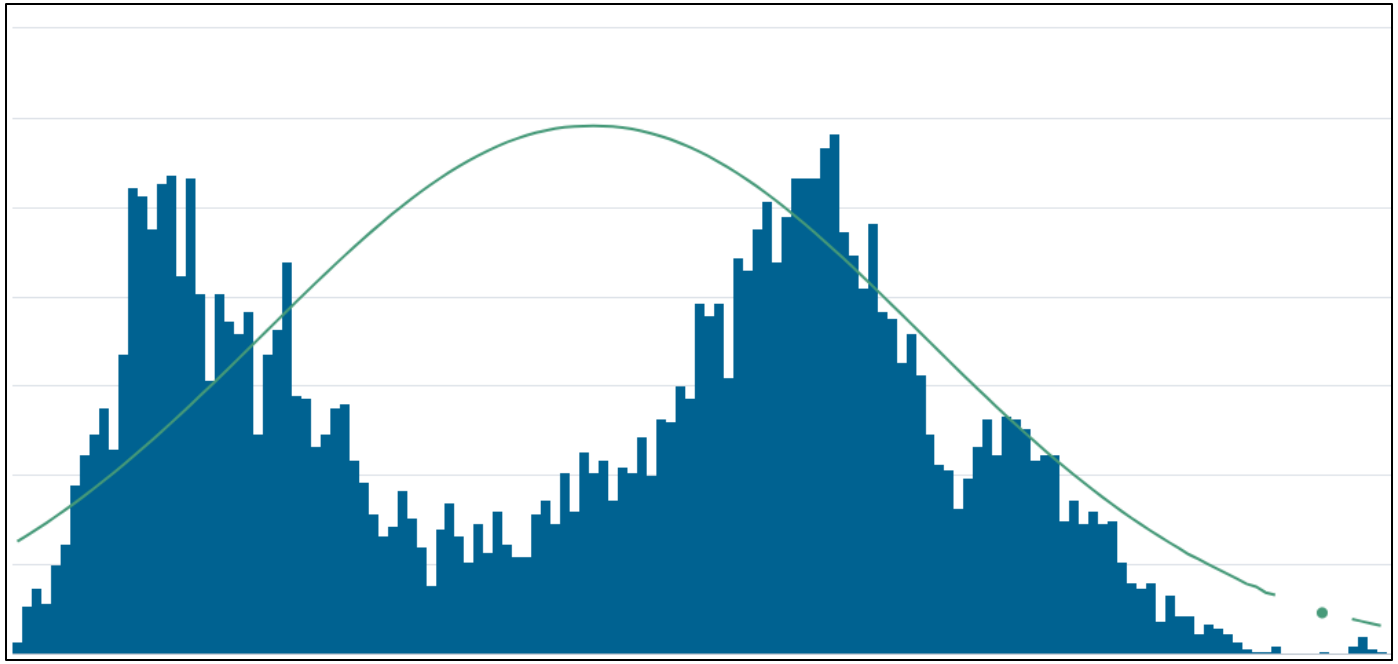
Viewing the login data as a histogram along with the normal distribution curve



Hmm… even 10 logins are considered pretty likely.

Understanding Probability Density Function

A mathematical function that generalizes the relative likelihood of a value falling within a specific range

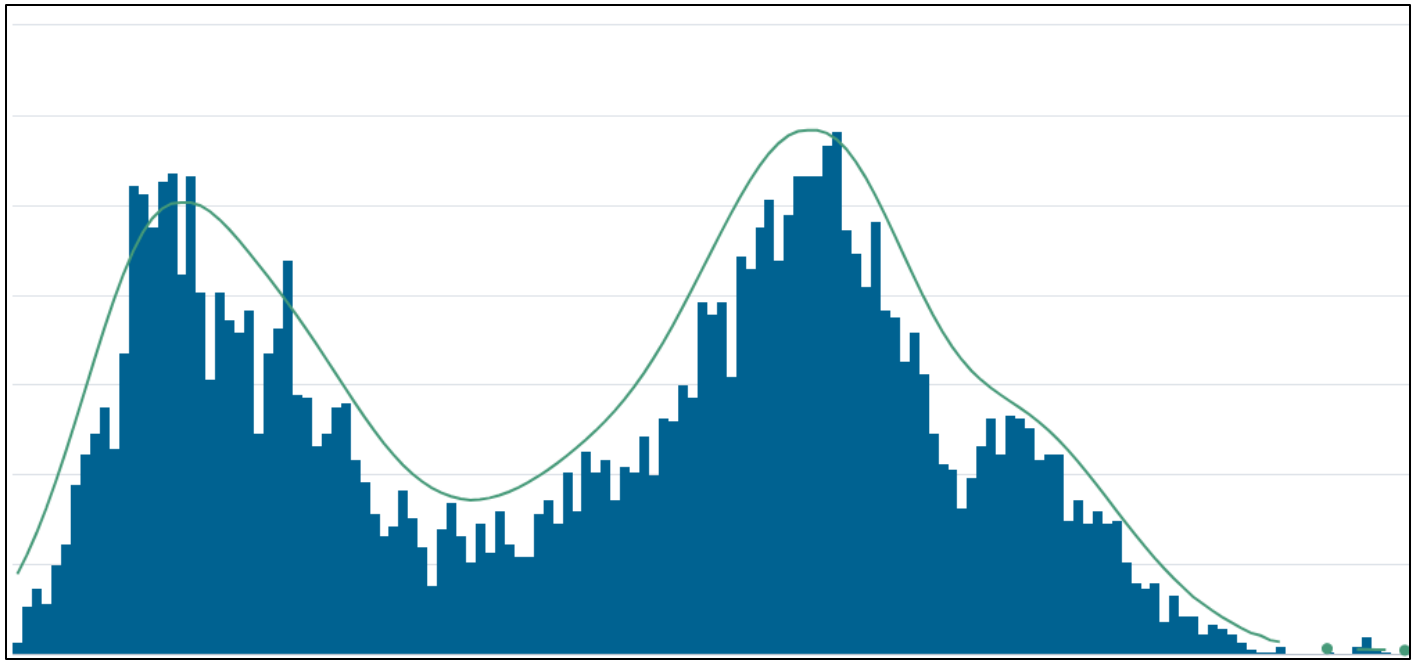


Instead of assuming a normal curve…

Understanding Probability Density Function

Available with the DensityFunction algorithm in Splunk

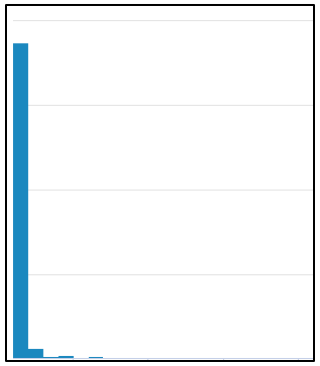
A mathematical function that generalizes the relative likelihood of a value falling within a specific range

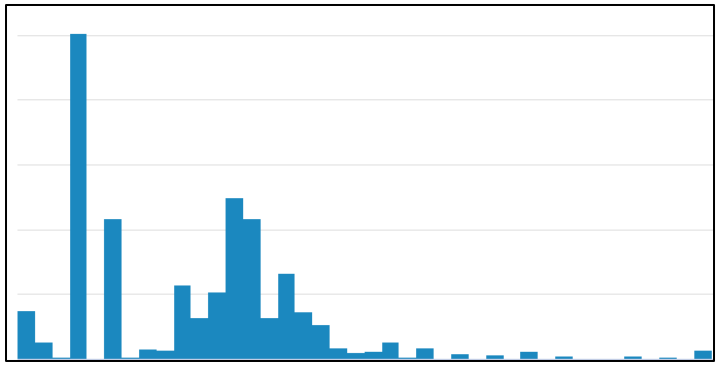


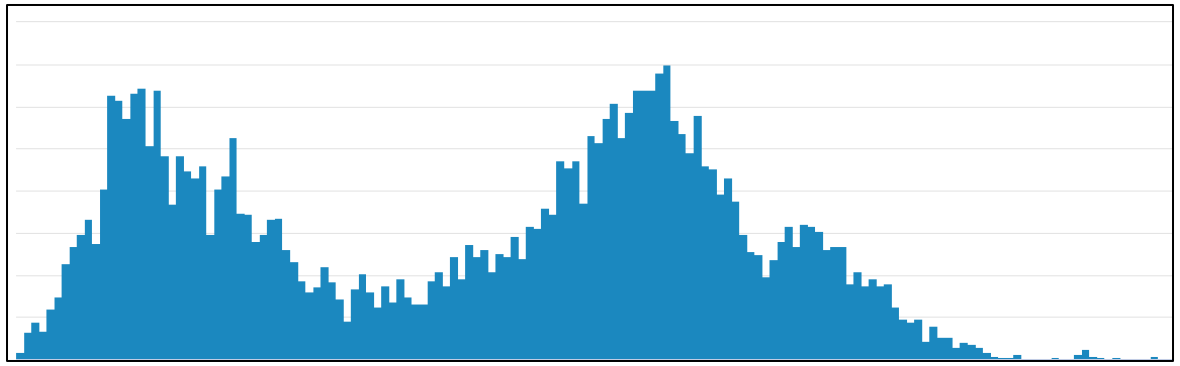
… what if we could draw this?

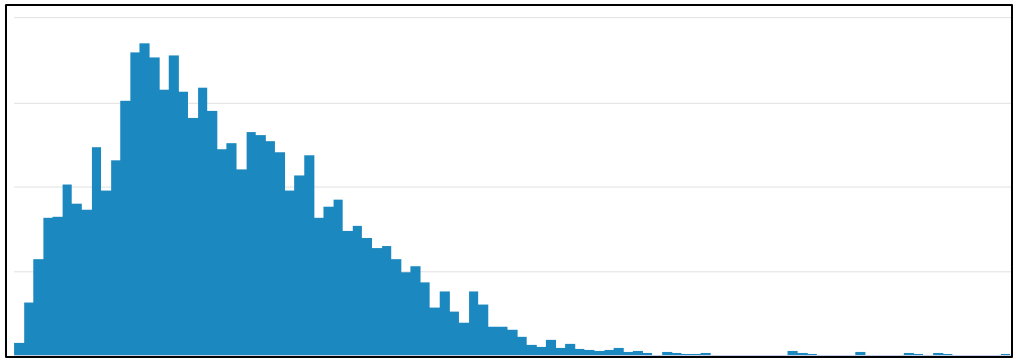
Examples of data that isn’t so “normal”

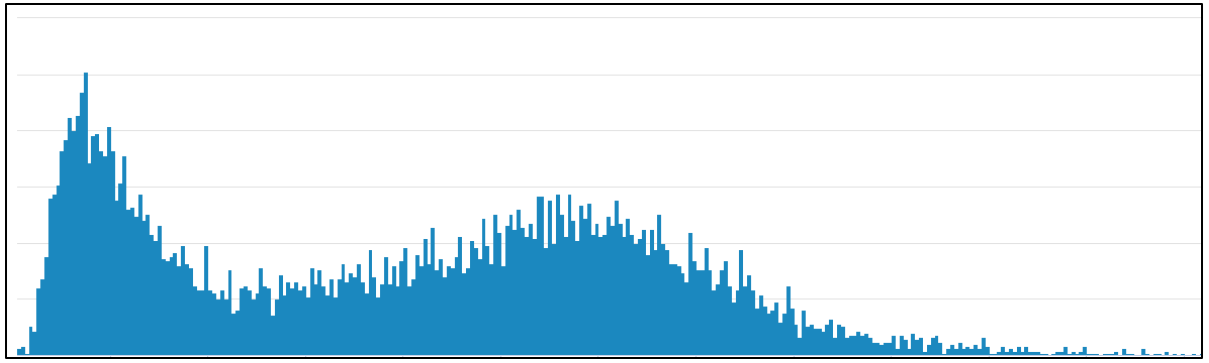
Different histograms of logins and API requests per minute











How do you use DensityFunction?

Let the math automatically figure out the distribution of your data

- The show_density parameter gives you the ProbabilityDensity for each value
  - The smaller the number, the less likely

```
| fit DensityFunction logins threshold=0.001 show_density=true
```

logins	IsOutlier(logins)	BoundaryRanges	ProbabilityDensity(logins)
209	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
190	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
61	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093

How do you use DensityFunction?

Let the math automatically figure out the distribution of your data

The BoundaryRanges field tells you where the outliers are based on the threshold (in this case 0.1%)

- Observation: We actually had a number of values between 33 and 38.
- Is someone using a user account as a service account?

```
| fit DensityFunction logins threshold=0.001 show_density=true
```

logins	IsOutlier(logins)	BoundaryRanges	ProbabilityDensity(logins)
209	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
190	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
61	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093

How do you use DensityFunction?

Let the math automatically figure out the distribution of your data

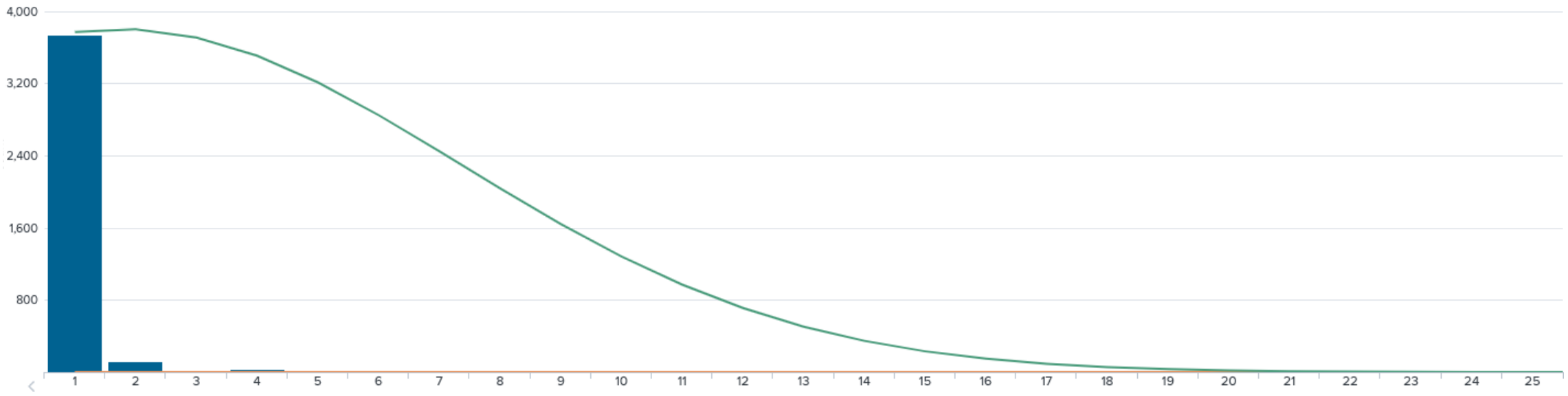
The IsOutlier field tells you if it is an outlier based on the threshold (in this case 0.1%)

```
| fit DensityFunction logins threshold=0.001 show_density=true
```

logins	IsOutlier(logins)	BoundaryRanges	ProbabilityDensity(logins)
209	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
190	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
61	1.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	9.192949794624945e-05
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093
37	0.0	[[32.0752, 32.7629, 0.0002], [38.2644, Infinity, 0.0009]]	0.0006067977941723093

Revisiting the login data with DensityFunction

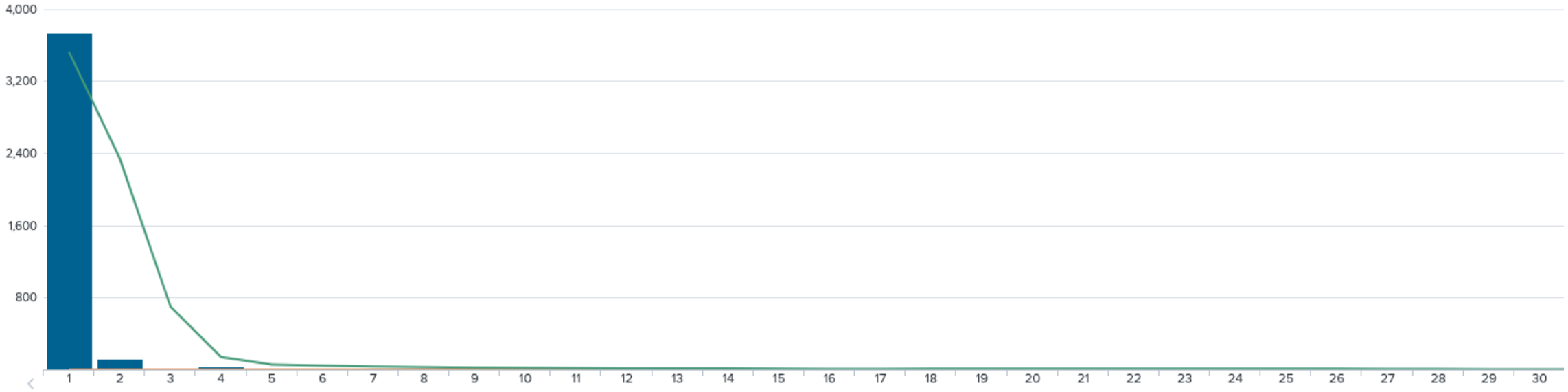
Viewing the login data as a histogram along with the normal distribution curve



Hmm… even 10 logins are considered pretty likely.

Revisiting the login data with DensityFunction

Viewing the login data as a histogram along with the normal distribution curve



Using DensityFunction, we're better able to follow the shape of the data

Split your data using the “by” clause

Apply different pivots of your data

```
| fit DensityFunction logins as IsOutlierGlobal
| fit DensityFunction logins by "date_hour" as IsOutlierByHour
| fit DensityFunction logins by "user_group" as IsOutlierByGroup
| eval AnomalyScore=0
| foreach IsOutlier* [eval AnomalyScore=AnomalyScore+<<FIELD>>]
```

_time	user	logins	AnomalyScore	IsOutlierByGroup	IsOutlierByHour	IsOutlierGlobal
2019-04-12 04:56:00	user015	35	3.0	1.0	1.0	1.0
2019-04-15 10:25:00	user015	17	3.0	1.0	1.0	1.0
2019-04-15 07:42:00	user015	30	2.0	1.0	0.0	1.0
2019-04-15 14:45:00	user846	4	2.0	1.0	1.0	0.0
2019-04-08 18:01:00	user291	2	1.0	0.0	1.0	0.0
2019-04-19 00:44:00	user474	2	1.0	0.0	1.0	0.0

(I think we found that user that is acting like a service account)

How to apply your model to new data

You can set threshold and show_density in the apply step

- The Probability Density Function does not change, but your boundaries will

```
… (lots of data) …
| fit DensityFunction my_field into MyModel

… (last 5 min, 1 hr, etc) …
| apply MyModel threshold=0.001 show_density=true
| search "IsOutlier(my_field)"=1
```

Make sure you have enough data

Some helpful considerations

- Important especially when you split by field(s)
- Use the following command to check the cardinality

```
| summary MyModelName
```

cardinality	date_hour	mean	std	type
86	11	1.00	0.14	Auto: Exponential
125	1	1.00	0.09	Auto: Exponential
135	23	1.12	0.39	Auto: Gaussian KDE
140	13	1.12	0.55	Auto: Gaussian KDE
152	12	1.00	0.63	Auto: Exponential
156	10	1.00	2.55	Auto: Exponential

Contextualize your alert with the model summary

How to better inform your outliers

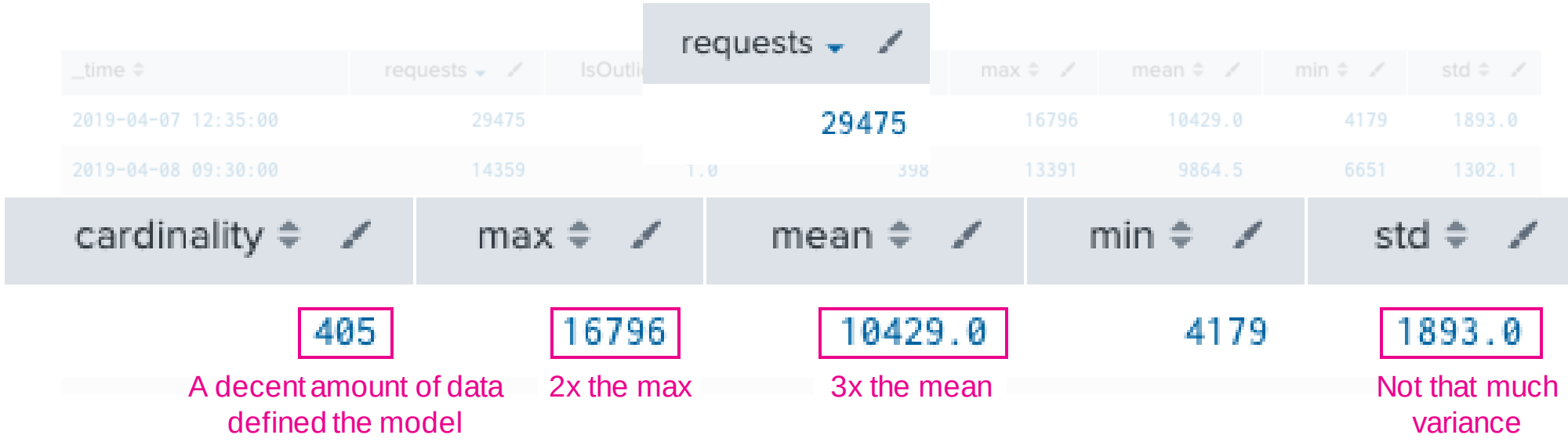
```
| apply MyDensityFunctionModel
| join dim1 dim2 etc
    [| summary MyDensityFunctionModel]
```

_time	requests	IsOutlier	cardinality	max	mean	min	std
2019-04-07 12:35:00	29475	1.0	405	16796	10429.0	4179	1893.0
2019-04-08 09:30:00	14359	1.0	398	13391	9864.5	6651	1302.1
2019-04-07 16:47:00	13640	1.0	392	12805	9858.8	6423	1383.2
2019-04-08 05:49:00	12435	1.0	384	11894	7781.4	4730	1504.6
2019-04-07 20:39:00	11565	1.0	355	11233	8612.4	5559	1106.1
2019-04-09 20:39:00	11452	1.0	355	11233	8612.4	5559	1106.1

Contextualize your alert with the model summary

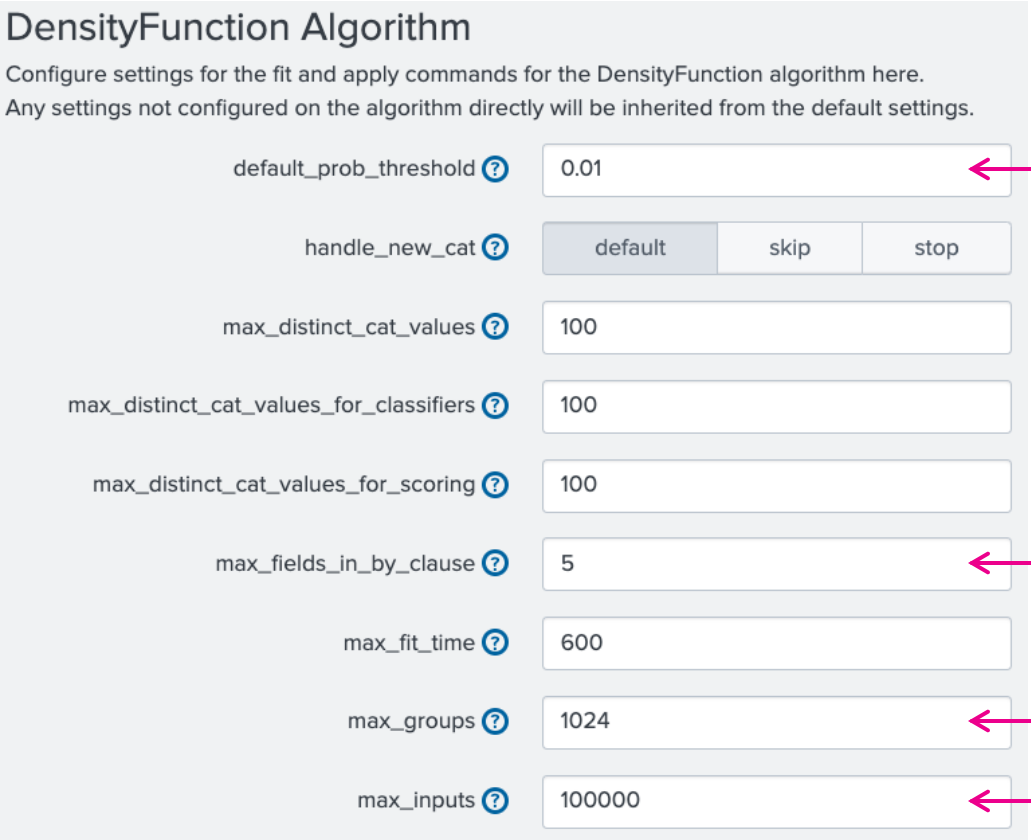
How to better inform your outliers

```
| apply MyDensityFunctionModel
| join dim1 dim2 etc
    [| summary MyDensityFunctionModel]
```



And don't forget the Settings (mlspl.conf)

Some helpful considerations



default_prob_threshold: Default value for threshold

max_fields_in_by_clause: doesn't come up that often, but be aware

max_groups: max groups created with the by clause

max_inputs: max rows processed.
Especially important with the split by

Wrap up and outlook

splunk> .conf19

Splunk Machine Learning Advisory Program

1) Get help from the Splunk Data Scientists to solve your business use case with Machine Learning Toolkit

2) Complimentary support with your Enterprise or Cloud license

3) Early access to new Machine Learning features

4) Results in opportunity to tell your success story with Splunk

5) Contact mlprogram@splunk.com for more information

Key Takeaways

1) Get your raw data including OT data sources mapped to CIM

2) Improve your alerts with Machine Learning based approaches e.g. the Density Function

3) Play hand in hand with OT and IT to secure critical infrastructure

Other related talks that you might be interested in

Check them out!

Products Splunk Machine Learning Toolkit ×

IT Operations Intermediate

⊕ **IT1171 - Accelerate your ability to sniff out application exceptions and detect outliers in performance KPIs**

SCHEDULE Tuesday, October 22, 04:15 PM - 05:00 PM

Foundations/Platform Intermediate

⊕ **FN1213 - The Two Most Common Machine Learning Solutions Everyone Needs to Know**

SCHEDULE Wednesday, October 23, 12:30 PM - 01:15 PM

Security, Compliance and Fraud Intermediate

⊕ **SEC1374 - Augment Your Security Monitoring Use Cases with Splunk's Machine Learning Toolkit**

SCHEDULE Thursday, October 24, 11:45 AM - 12:30 PM

Q&A

Philipp Drieger | Staff Machine Learning Architect, Splunk
Eurus Kim | Staff Machine Learning Architect, Splunk

splunk> .conf19

.conf19

splunk>

Thank You!

Go to the .conf19 mobile app to

RATE THIS SESSION