Splunk Machine Learning Toolkit In Action

Iman Makaremi, Splunk Principal Data Scientist
Andrew Stein, Splunk Principal PM for Machine Learning

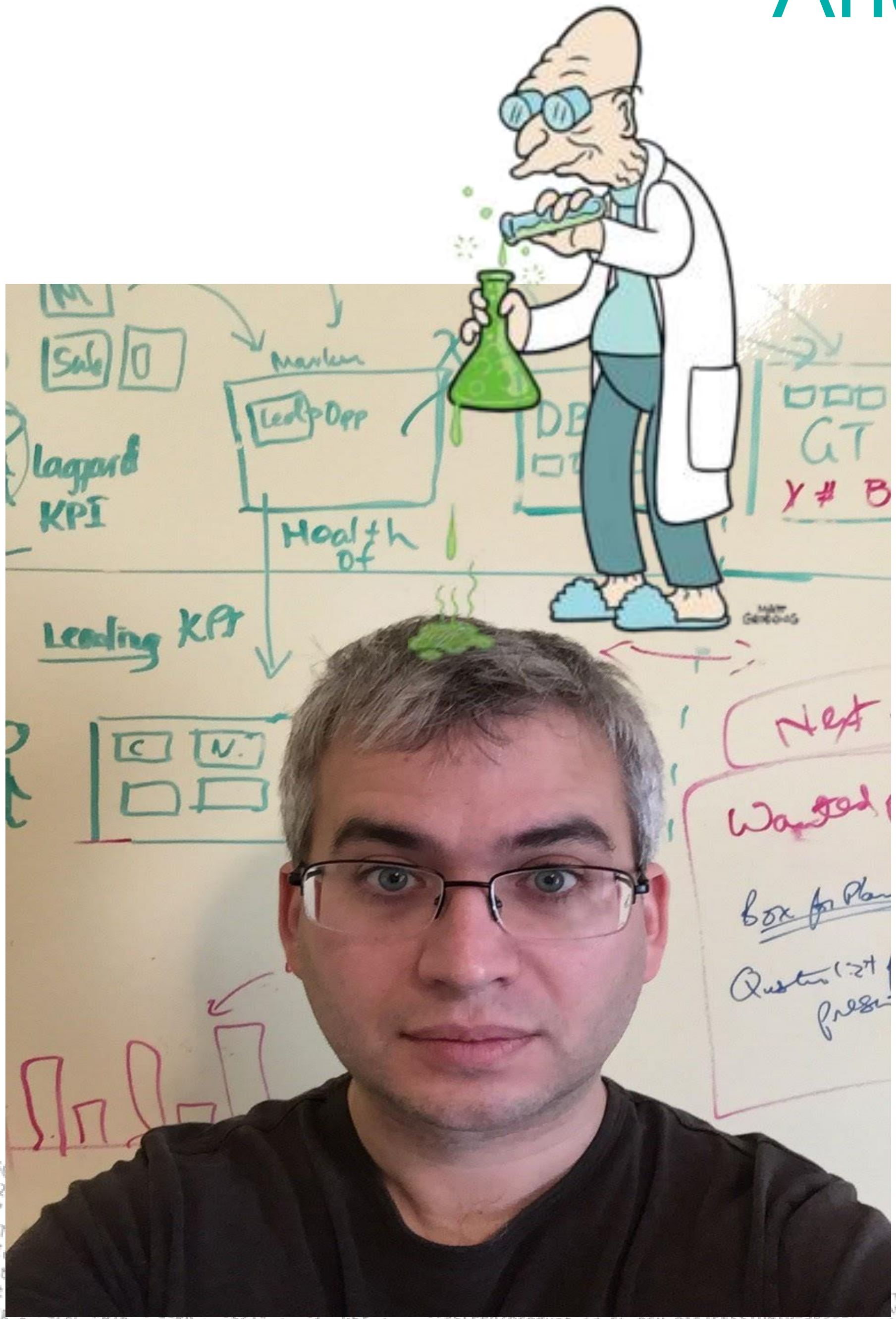
Oct 2018 | Version 1.0

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.

Andrew Stein



- Splunk Principal Product Manager – Machine Learning
- 18 years creating mathematical modeled solutions as a data scientist.
- I spend 80 percent of time preparing data and 20 percent of time complaining about the need to prepare data.

Iman Makaremi

- Splunk Principal Data Scientist
- I like math and physics a lot and know a few things about them.
- I spend 80 percent of time preparing to complain and 20 percent of time complaining about the need to complain.

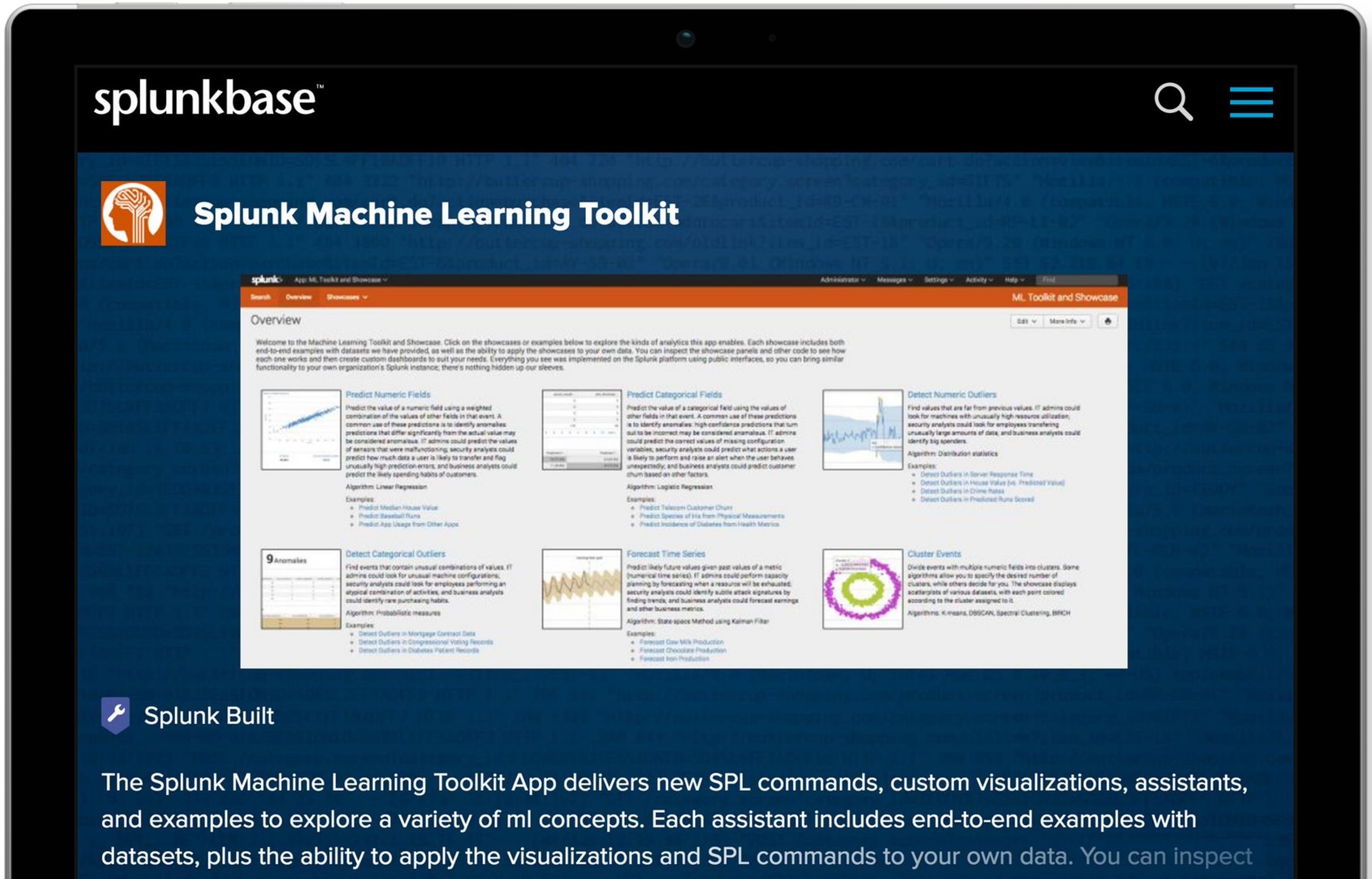
Agenda

Am I in the right room?

- What?
- Assistant : Detect Numeric Outliers
  - IOT & IT Example
- Assistant : Predict Numeric Fields
  - Forecasts verse Prediction
  - IT Example
- Assistant : Clustering
  - Security Example

What?

Splunk has a Machine Learning Toolkit App!



Splunk Machine Learning Toolkit

Overview

Welcome to the Machine Learning Toolkit and Showcase. Click on the showcases or examples below to explore the kinds of analytics this app enables. Each showcase includes both end-to-end examples with datasets we have provided, as well as the ability to apply the showcases to your own data. You can inspect the showcase panels and other code to see how each one works and then create custom dashboards to suit your needs. Everything you see was implemented on the Splunk platform using public interfaces, so you can bring similar functionality to your own organization's Splunk instance; there's nothing hidden up our sleeves.

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. IT admins could predict the values of sensors that were malfunctioning; security analysts could predict how much data a user is likely to transfer and flag unusually high prediction errors; and business analysts could predict the likely spending habits of customers.

Algorithm: Linear Regression

Examples:

- Predict Median House Value
- Predict Baseball Runs
- Predict App Usage from Other Apps

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: high-confidence predictions that turn out to be incorrect may be considered anomalous. IT admins could predict the correct values of missing configuration variables; security analysts could predict what actions a user is likely to perform and raise an alert when the user behaves unexpectedly; and business analysts could predict customer churn based on other factors.

Algorithm: Logistic Regression

Examples:

- Predict Telecom Customer Churn
- Predict Species of Iris from Physical Measurements
- Predict Incidence of Diabetes from Health Metrics

Detect Numeric Outliers

Find values that are far from previous values. IT admins could look for machines with unusually high resource utilization; security analysts could look for employees transfering unusually large amounts of data; and business analysts could identify big spenders.

Algorithm: Distribution statistics

Examples:

- Detect Outliers in Server Response Time
- Detect Outliers in House Value (vs. Predicted Value)
- Detect Outliers in Crime Rates
- Detect Outliers in Predicted Runs Scored

Detect Categorical Outliers

Find events that contain unusual combinations of values. IT admins could look for unusual machine configurations; security analysts could look for employees performing an atypical combination of activities; and business analysts could identify rare purchasing habits.

Algorithm: Probabilistic measures

Examples:

- Detect Outliers in Mortgage Contract Data
- Detect Outliers in Congressional Voting Records
- Detect Outliers in Diabetes Patient Records

Forecast Time Series

Predict likely future values given past values of a metric (numerical time series). IT admins could perform capacity planning by forecasting when a resource will be exhausted; security analysts could identify subtle attack signatures by finding trends; and business analysts could forecast earnings and other business metrics.

Algorithm: State-space Method using Kalman Filter

Examples:

- Forecast Cow Milk Production
- Forecast Chocolate Production
- Forecast Iron Production

Cluster Events

Divide events with multiple numeric fields into clusters. Some algorithms allow you to specify the desired number of clusters, while others decide for you. The showcase displays scatterplots of various datasets, with each point colored according to the cluster assigned to it.

Algorithms: K-means, DBSCAN, Spectral Clustering, BIRCH

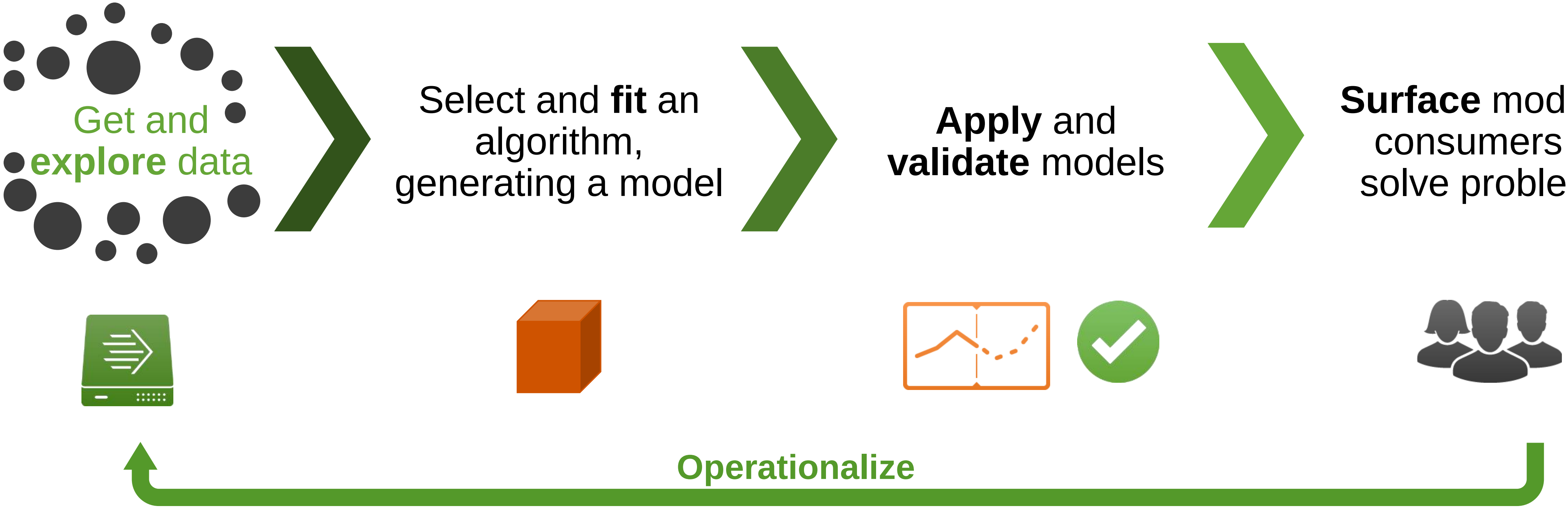
Splunk Built

The Splunk Machine Learning Toolkit App delivers new SPL commands, custom visualizations, assistants, and examples to explore a variety of ml concepts. Each assistant includes end-to-end examples with datasets, plus the ability to apply the visualizations and SPL commands to your own data. You can inspect

- What is Splunkbase
- What is the App
- Where can I go to learn more

Workflow. Workflow. Workflow. Bees. Or Math.

Problem: <Stuff in the world> causes big time and money expense.
Solution: Build ML model to learn the behaviors at scale and take action.



Get and **explore** data

Select and **fit** an algorithm, generating a model

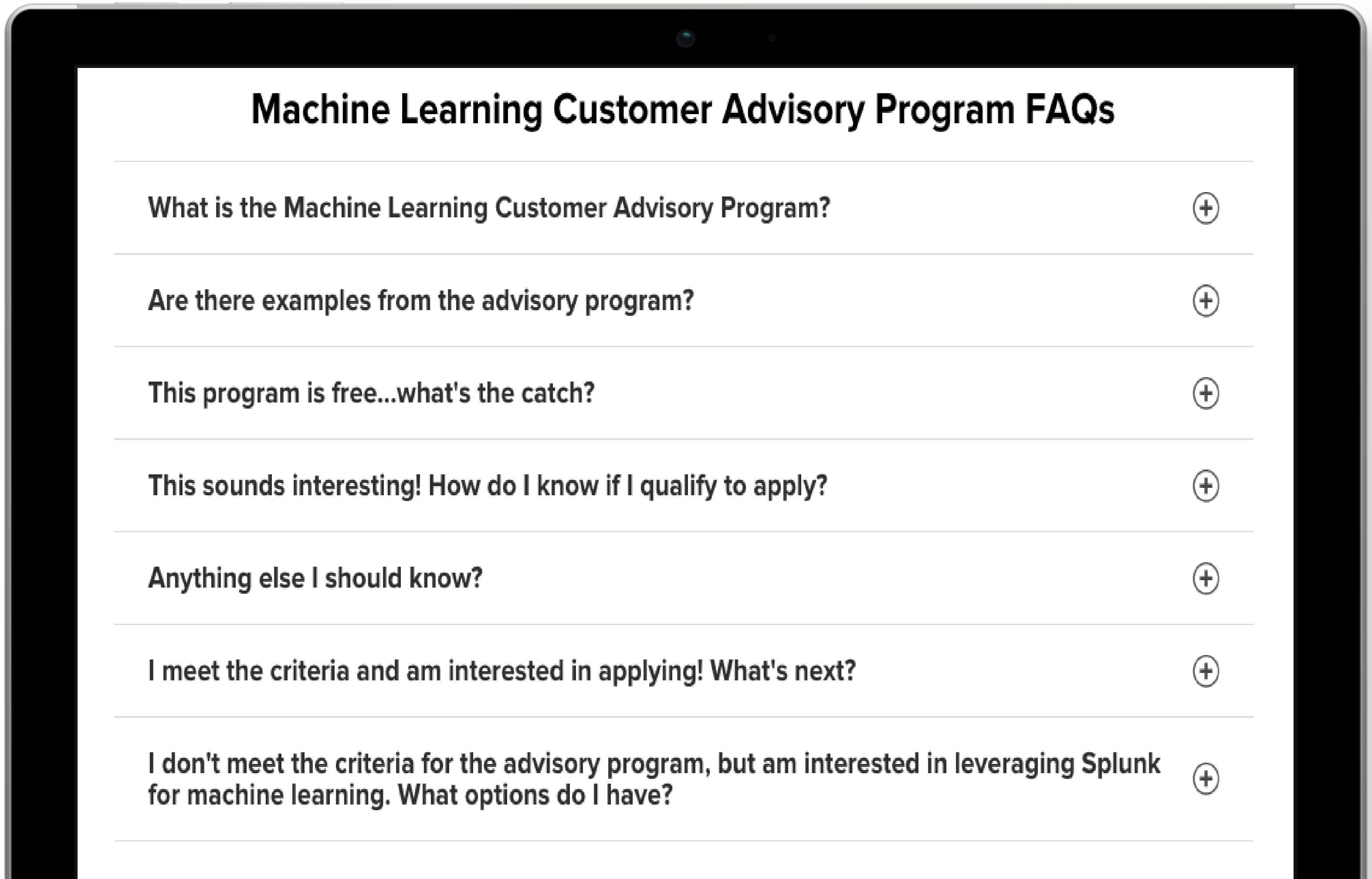
Apply and **validate** models

Surface model to consumers to solve problems

Operationalize

What is the ML Advisory Program?

Partners a Splunk Data Science Resource to Help Operationalize a ML Use Case



Machine Learning Customer Advisory Program FAQs

- What is the Machine Learning Customer Advisory Program? ⊕
- Are there examples from the advisory program? ⊕
- This program is free...what's the catch? ⊕
- This sounds interesting! How do I know if I qualify to apply? ⊕
- Anything else I should know? ⊕
- I meet the criteria and am interested in applying! What's next? ⊕
- I don't meet the criteria for the advisory program, but am interested in leveraging Splunk for machine learning. What options do I have? ⊕

- Early Access to new and enhanced MLTK features
- Opportunity to shape the development of the product
- Assistance in operationalizing a production quality ML model

Assistant :Detect Numeric Outliers

Most Popular Assistant

Detecting and Resolving Data Outages

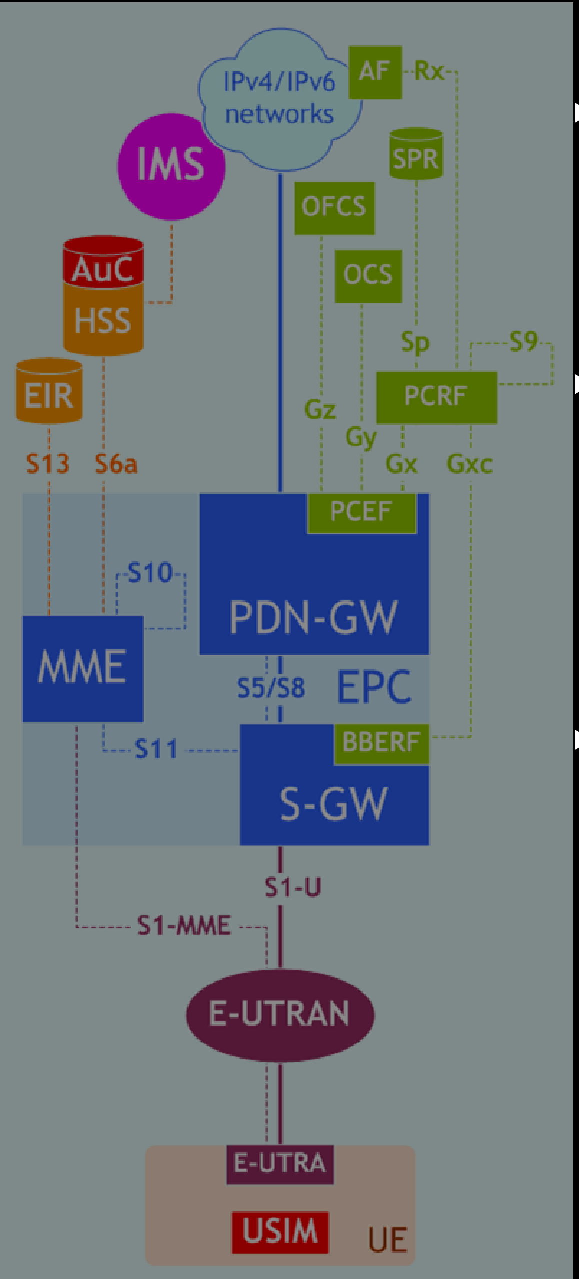
- *"Everything is Data, All of it is important."*

- Data
  - Splunk index logs, Enrollment of Students through time
  - Database logs, Any data where outages matter to you

- Action Taken with the Detect Numeric Outliers
  - *Splunk Admin after taking the Splunk EDU Data Science course.*
  - Detecting data source outages that are critical for supported research and operational centers, using custom seasonality
  - Automatically Impute and replace missing information in summary index, send alert to administrator for further action.

TELUS
the future is friendly®

Mobility 3GPP Core KPI



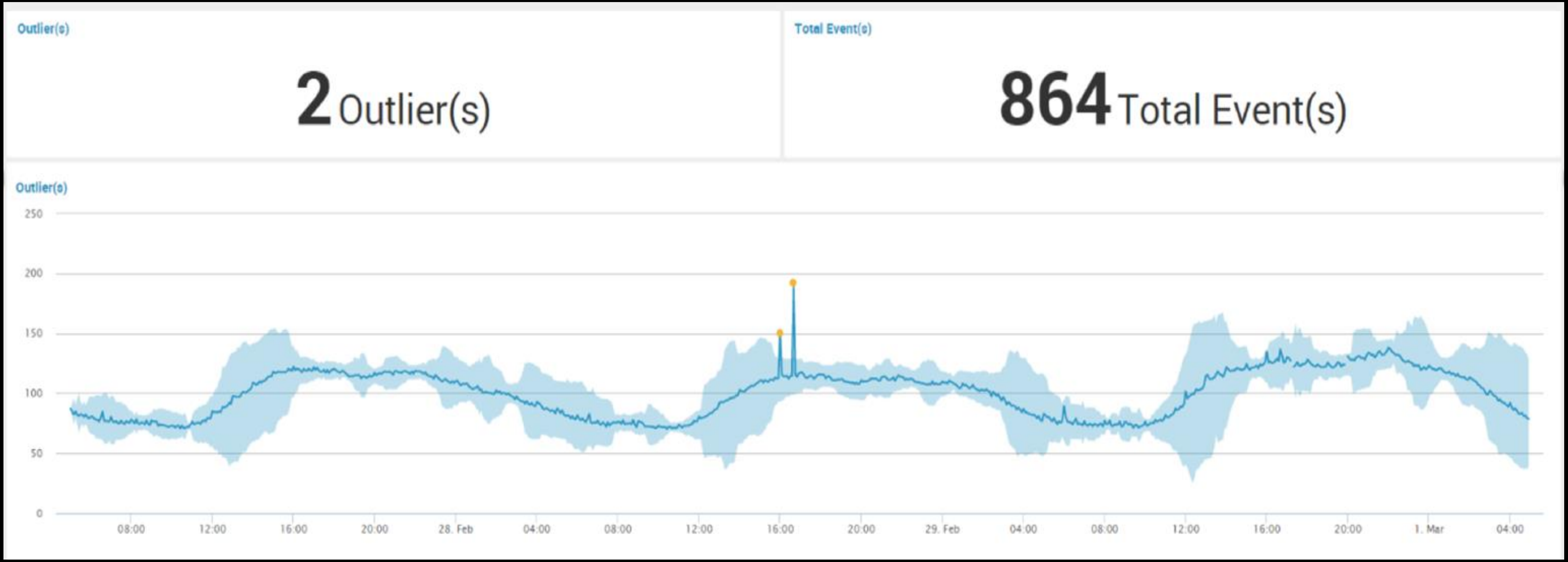
- *Telus cell towers create a valuable and complicated network to maintain customer phone connectivity. The usability of our network by our customers is critical for our business.*

- Data
  - The 3GPP Core receives transactions from each subscriber **to maintain connection**.
  - The Telus custom KPI captures **the behavior of TELUS's network.**

- Action Taken *with a Custom Machine Learning Model in Splunk*
  - Monitor this **dynamic** KPI and alert on **contextual performance degradation**.
  - Radio engineers informed about deviations from expectations immediately, creating the opportunity for in-the-field technician corrective actions.

splunk> listen to your data®

Numerical Outlier Detection in MLTK



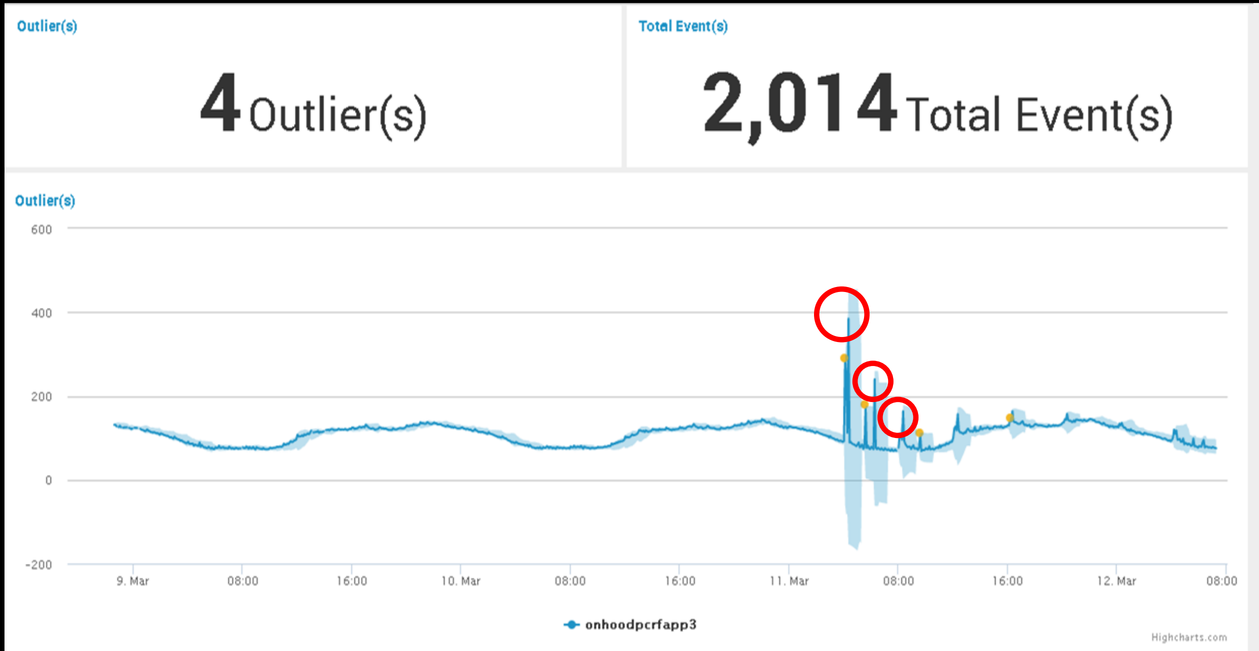
Outlier(s)

2 Outlier(s)

Total Event(s)

864 Total Event(s)

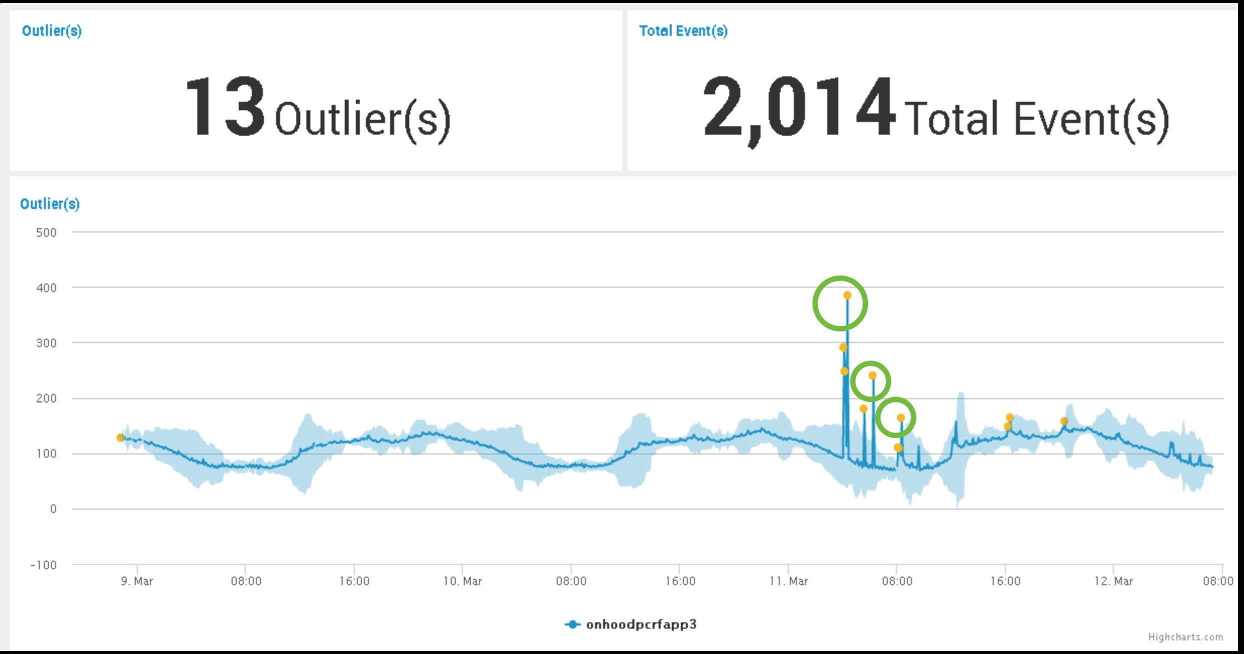
Iterating over different threshold methods


Outlier(s)

4 Outlier(s)

Total Event(s)

2,014 Total Event(s)


Outlier(s)

13 Outlier(s)

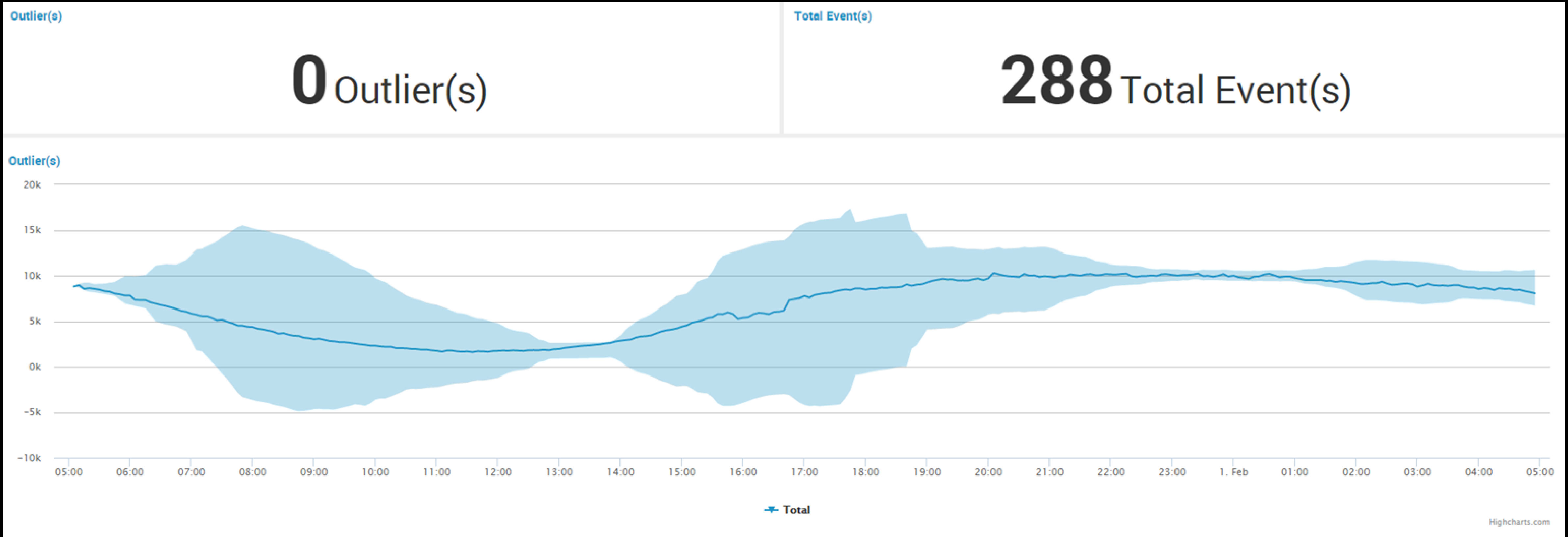
Total Event(s)

2,014 Total Event(s)

Standard Deviation

Absolute Median Deviation

Missed Outlier!



Outlier(s)

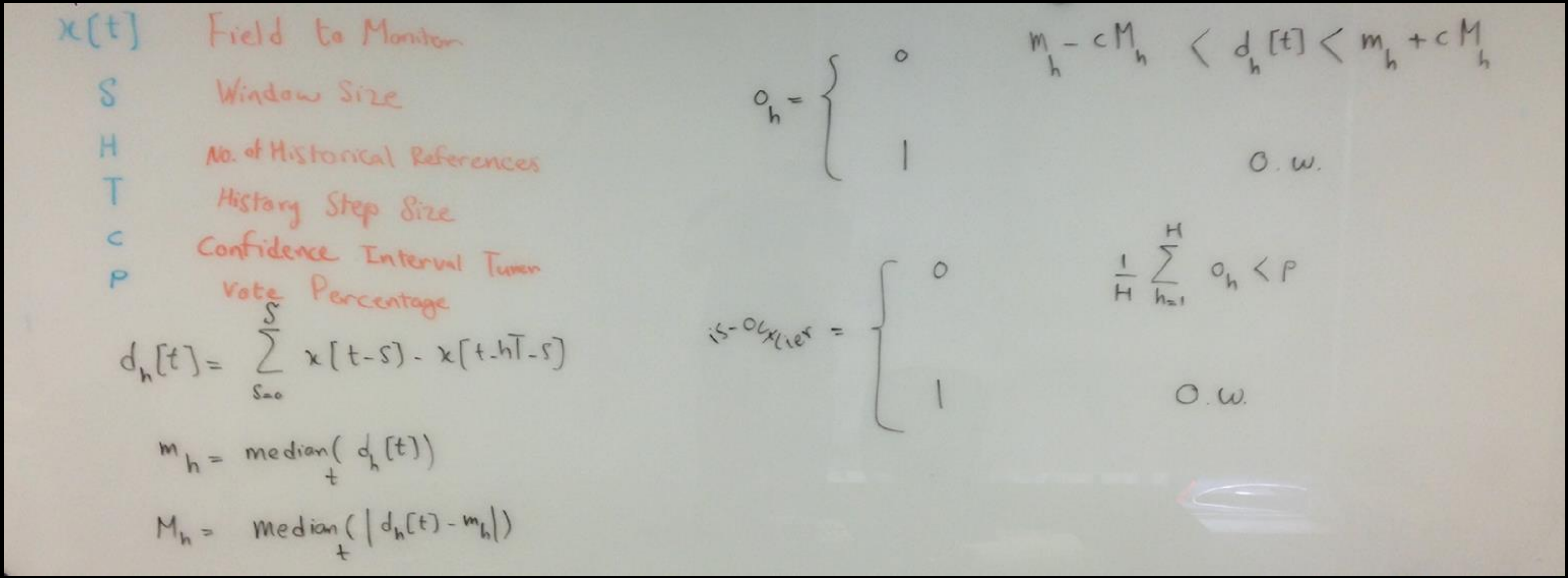
0 Outlier(s)

Total Event(s)

288 Total Event(s)

Outlier(s)

Customized Outlier Detection



$x[t]$	Field to Monitor
S	Window Size
H	No. of Historical References
T	History Step Size
c	Confidence Interval Tuner
P	Vote Percentage

$$d_h[t] = \sum_{s=0}^{S} x[t-s] - x[t-hT-s]$$

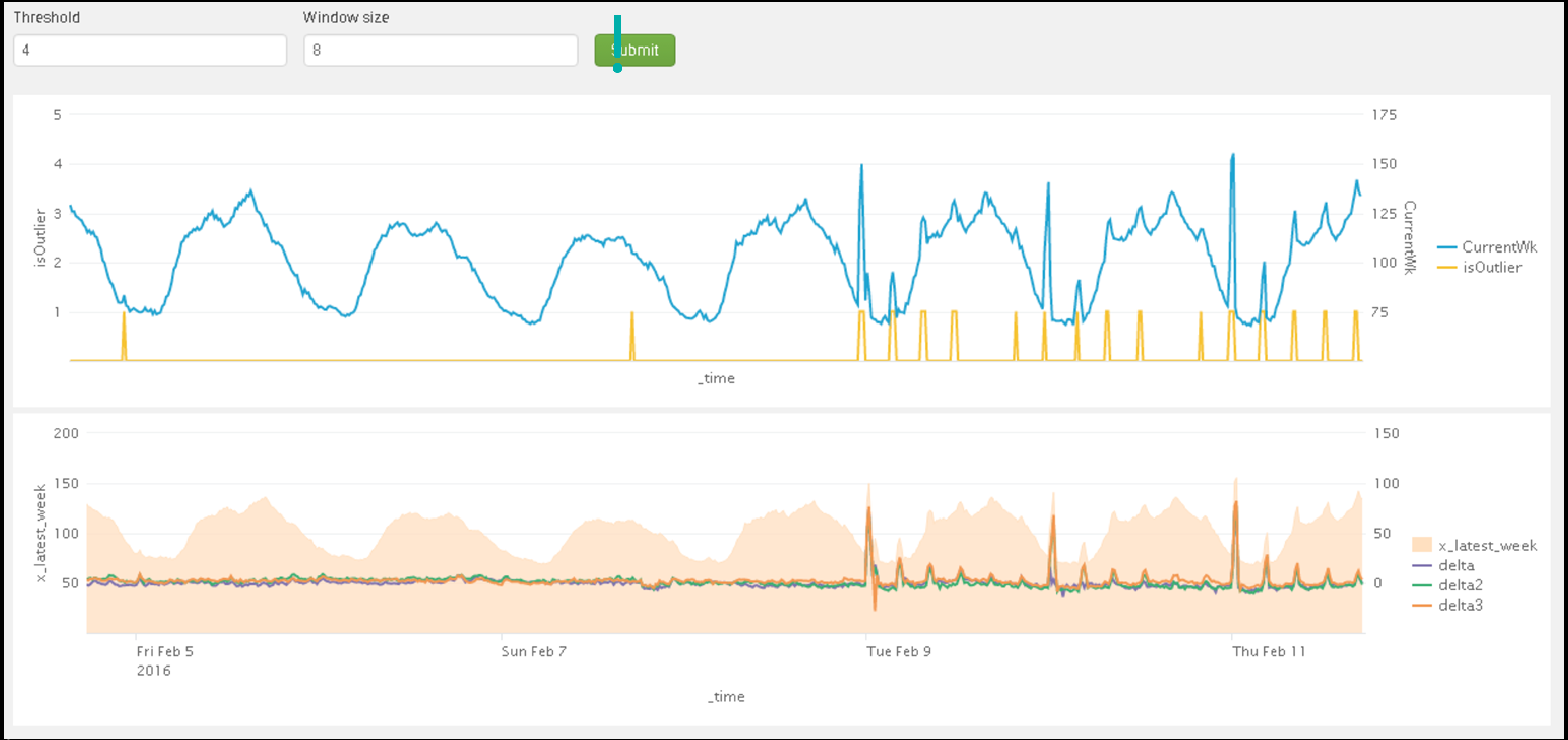
$$m_h = \underset{t}{\text{median}}(d_h[t])$$

$$M_h = \underset{t}{\text{median}}(|d_h[t] - m_h|)$$

$$o_h = \begin{cases} 0 & m_h - cM_h < d_h[t] < m_h + cM_h \\ 1 & \text{o.w.} \end{cases}$$

$$\text{is-outlier} = \begin{cases} 0 & \frac{1}{H}\sum_{h=1}^{H} o_h < P \\ 1 & \text{o.w.} \end{cases}$$

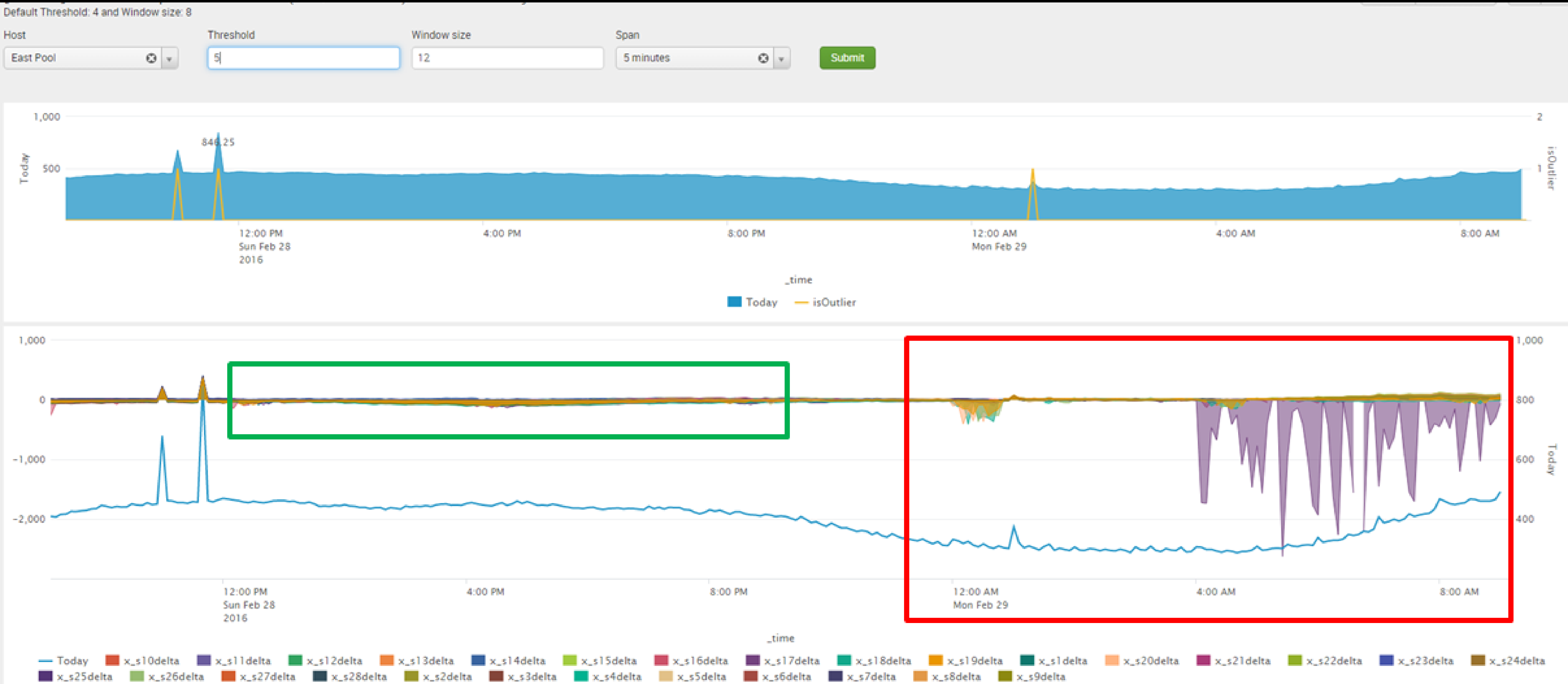
Evaluate



Threshold: 4

Window size: 8

Current vs Historic and Delta Calculation



Default Threshold: 4 and Window size: 8

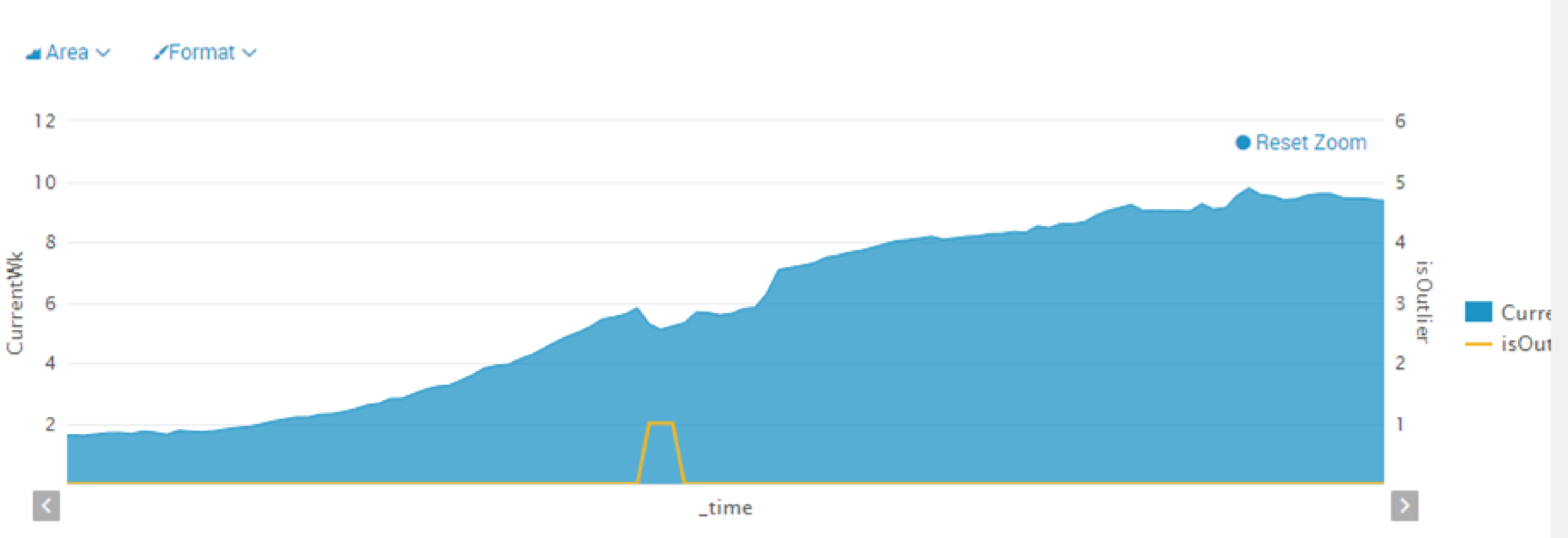
Host: East Pool

Threshold: 5

Window size: 12

Span: 5 minutes

DETECTED!


Area
Format
Reset Zoom
12
10
8
6
4
2
CurrentWk
6
5
4
3
2
1
isOutlier
_time
Curre
isOut

I want Anomalies…

What is the minimum requirements for each workflow

- Detect Numeric Outliers Workflow

I have one number I care about, with possible seasonality (time) effects or some combination of identities (the “by” clause from stats for example).

I want to find anomalies in one number moving through time.

How to Blog:

Statistical Anomalies and Forecasts (parts 1 ,2,3)

- Or use many workflows…

I understand the statistics workflow, but I have many fields describing my problem or a measurable ground truth.

I want to find complex anomalies and my data is organized.

How to Blog:

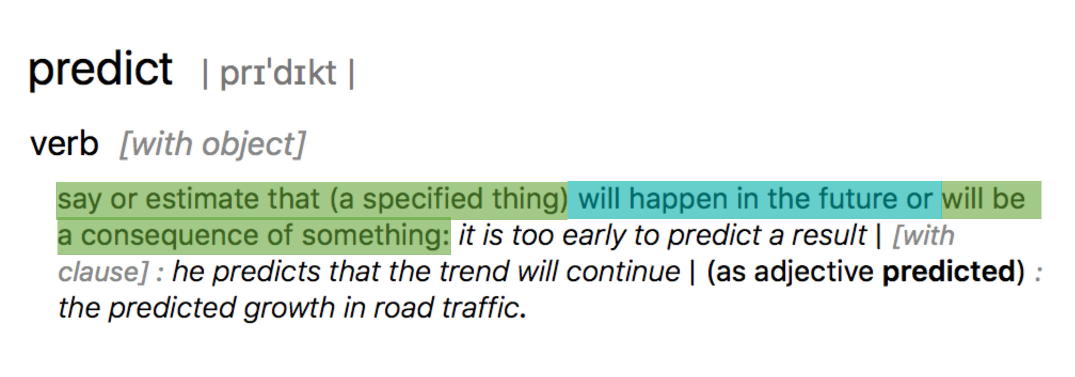
Anomalies like Neapolitan Ice Cre

Assistant: Predict Numeric Fields

Most Misunderstood Assistant

Predict vs Forecast

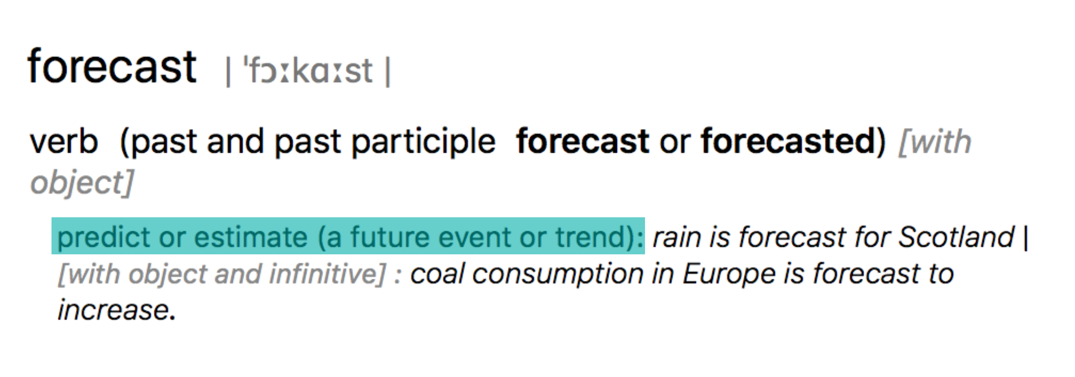
English kind of sucks….



predict | prɪˈdɪkt |

verb *[with object]*

say or estimate that (a specified thing) will happen in the future or will be a consequence of something: *it is too early to predict a result* | *[with clause]* : *he predicts that the trend will continue* | (as adjective **predicted**) : *the predicted growth in road traffic.*



forecast | ˈfɔːkɑːst |

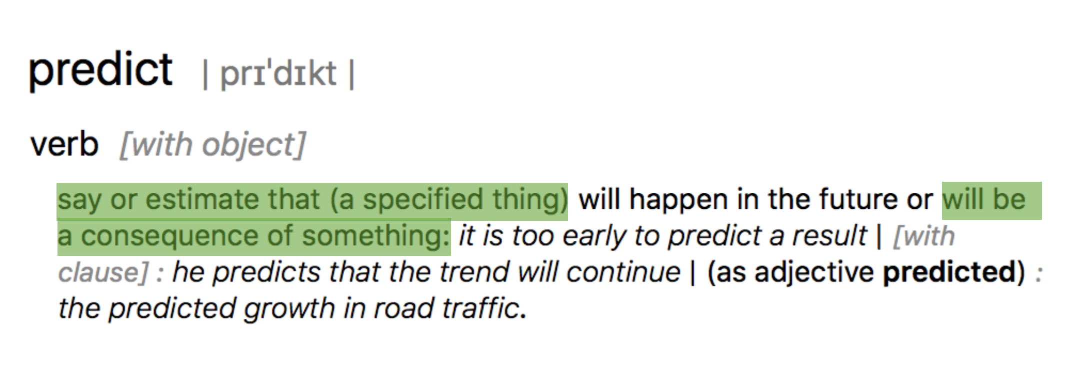
verb (past and past participle **forecast** or **forecasted**) *[with object]*

predict or estimate (a future event or trend): *rain is forecast for Scotland* | *[with object and infinitive]* : *coal consumption in Europe is forecast to increase.*

Source: Mac's Dictionary

splunk> .conf18

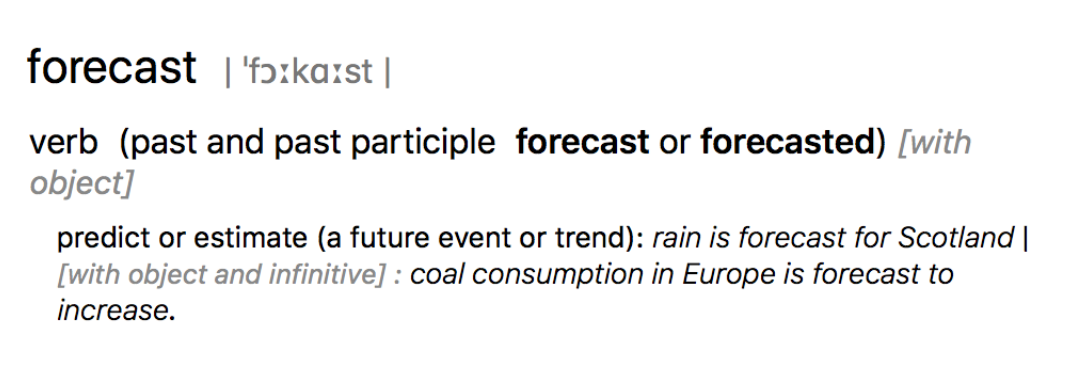
Predict vs Forecast



predict | prɪˈdɪkt |

verb *[with object]*

say or estimate that (a specified thing) will happen in the future or will be a consequence of something: *it is too early to predict a result* | *[with clause]* : *he predicts that the trend will continue* | (as adjective **predicted**) : *the predicted growth in road traffic.*



forecast | ˈfɔːkɑːst |

verb (past and past participle **forecast** or **forecasted**) *[with object]*

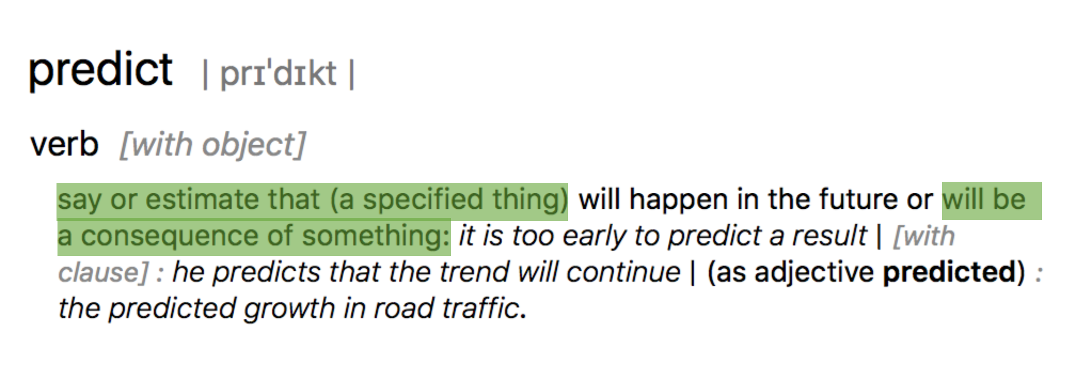
predict or estimate (a future event or trend): *rain is forecast for Scotland* | *[with object and infinitive]* : *coal consumption in Europe is forecast to increase.*

The Splunk stock is influenced by interest rates, global economic conditions, road map, CFO’s blood pressure, density of CEO’s beard…

≠

The Splunk stock is cyclical, and every July stock price in the future will look like the July stock in the past +/- trending.

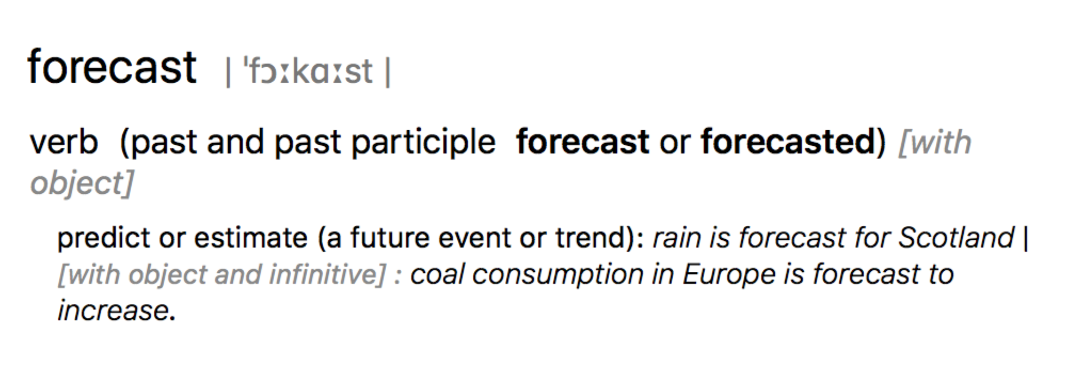
Predict vs Forecast



predict | prɪˈdɪkt |

verb *[with object]*

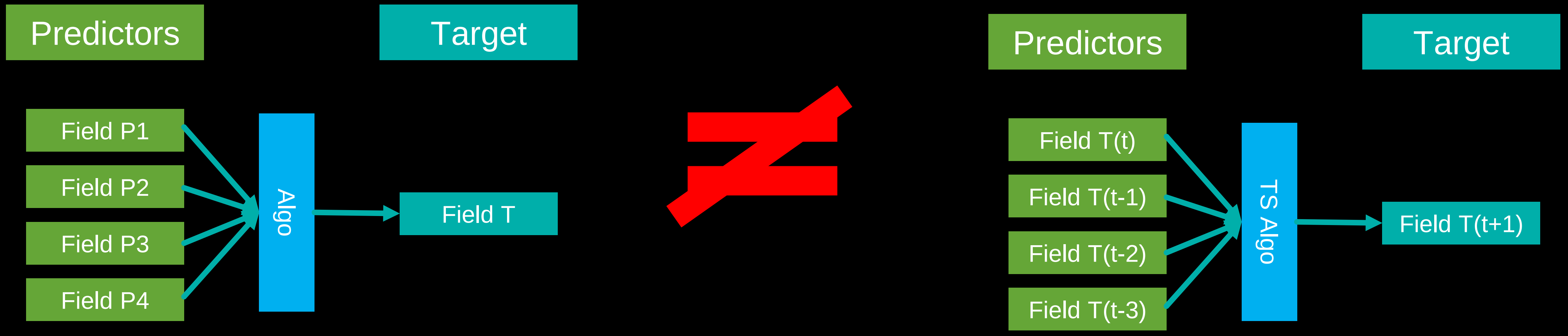
say or estimate that (a specified thing) will happen in the future or will be a consequence of something: *it is too early to predict a result* | *[with clause]* : *he predicts that the trend will continue* | (as adjective **predicted**) : *the predicted growth in road traffic.*



forecast | ˈfɔːkɑːst |

verb (past and past participle **forecast** or **forecasted**) *[with object]*

predict or estimate (a future event or trend): *rain is forecast for Scotland* | *[with object and infinitive]* : *coal consumption in Europe is forecast to increase.*



Predictors

Target

Field P1

Field P2

Field P3

Field P4

Algo

Field T

≠

Predictors

Target

Field T(t)

Field T(t-1)

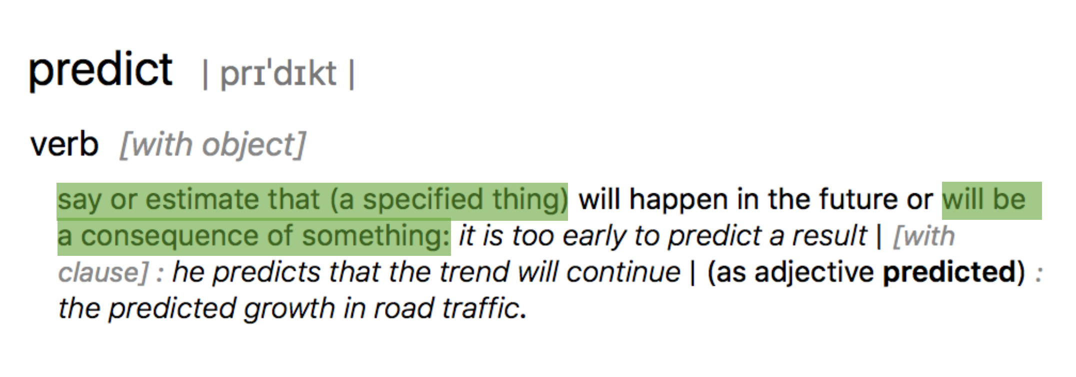
Field T(t-2)

Field T(t-3)

TS Algo

Field T(t+1)

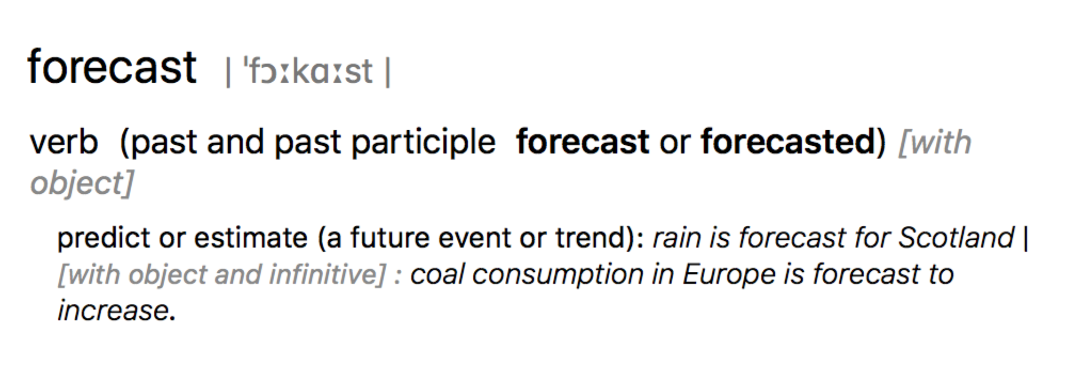
Predict vs Forecast



predict | prɪˈdɪkt |

verb *[with object]*

say or estimate that (a specified thing) will happen in the future or will be a consequence of something: *it is too early to predict a result* | *[with clause]* : *he predicts that the trend will continue* | (as adjective **predicted**) : *the predicted growth in road traffic.*



forecast | ˈfɔːkɑːst |

verb (past and past participle **forecast** or **forecasted**) *[with object]*

predict or estimate (a future event or trend): *rain is forecast for Scotland* | *[with object and infinitive]* : *coal consumption in Europe is forecast to increase.*

MLTK Assistants

- Predict a Numeric Field
- Predict a Categorical Field

≠

Splunk

- *predict* Command

MLTK Assistant

- Forecast Time Series

+ some optional time Travel SPL.

How to Blog:

ITSI and Sophisticated Machine Learning

How to Blog:

Statistical Anomalies and Forecasts (parts 1,2,3)

FINANCIAL SERVICES – IT OPERATIONS MANAGEMENT

TransUnion Invests in Splunk Solutions for Enterprise Monitoring, Machine Learning

> “ Understanding customer volume patterns is important for the business. If traffic falls outside of a certain range, an alert is created. Splunk machine learning allows us to investigate early to ensure a seamless customer experience.”
>
> – ***Lead Splunk Developer, TransUnion***

- With the Splunk Machine Learning Toolkit and Splunk Machine Learning Customer Advisory Program Hyatt:
- Helping to meet customer SLAs
- Discovering incident root causes in minutes instead of hours
- Reducing the number of false alerts
- Increasing revenue by improving transaction processing

I want a prediction…

What are the minimum requirements for each workflow

- Predict a Numeric Field

I want to generalize the relationship between one target **numeric** field and a series of descriptive fields.

I need a | table with the target **numeric** field and the descriptive fields , with _time if I am going to predict the future. The use of | table is not required, this is just good formatting step

How to Blog:

Custom Anomaly Detection with Splunk IT Service Intelligence and Machine Learning Toolkit

- Predict a Future Value of a Field

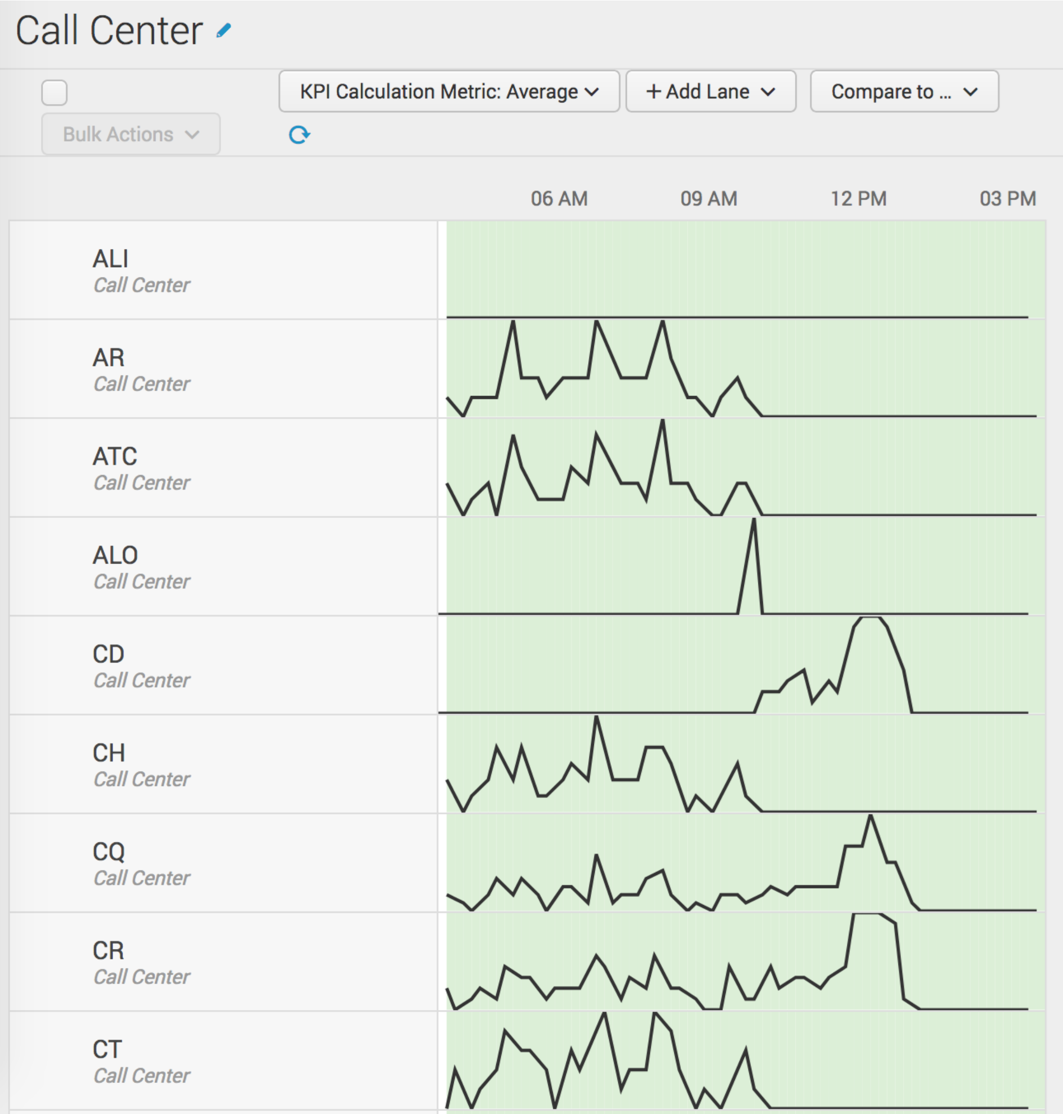
I want to generalize the relationship between one target **numeric** field and a series of descriptive fields, but I want to have that relationship be explicitly in the future

I need a | table with the target **numeric** field and the descriptive fields , with _time if I am going to predict the future. The use of | table is not required, this is just good formatting step. *<u>I need to move the target field through time.</u>*

How to Blog:

ITSI and Sophisticated Machine Learning

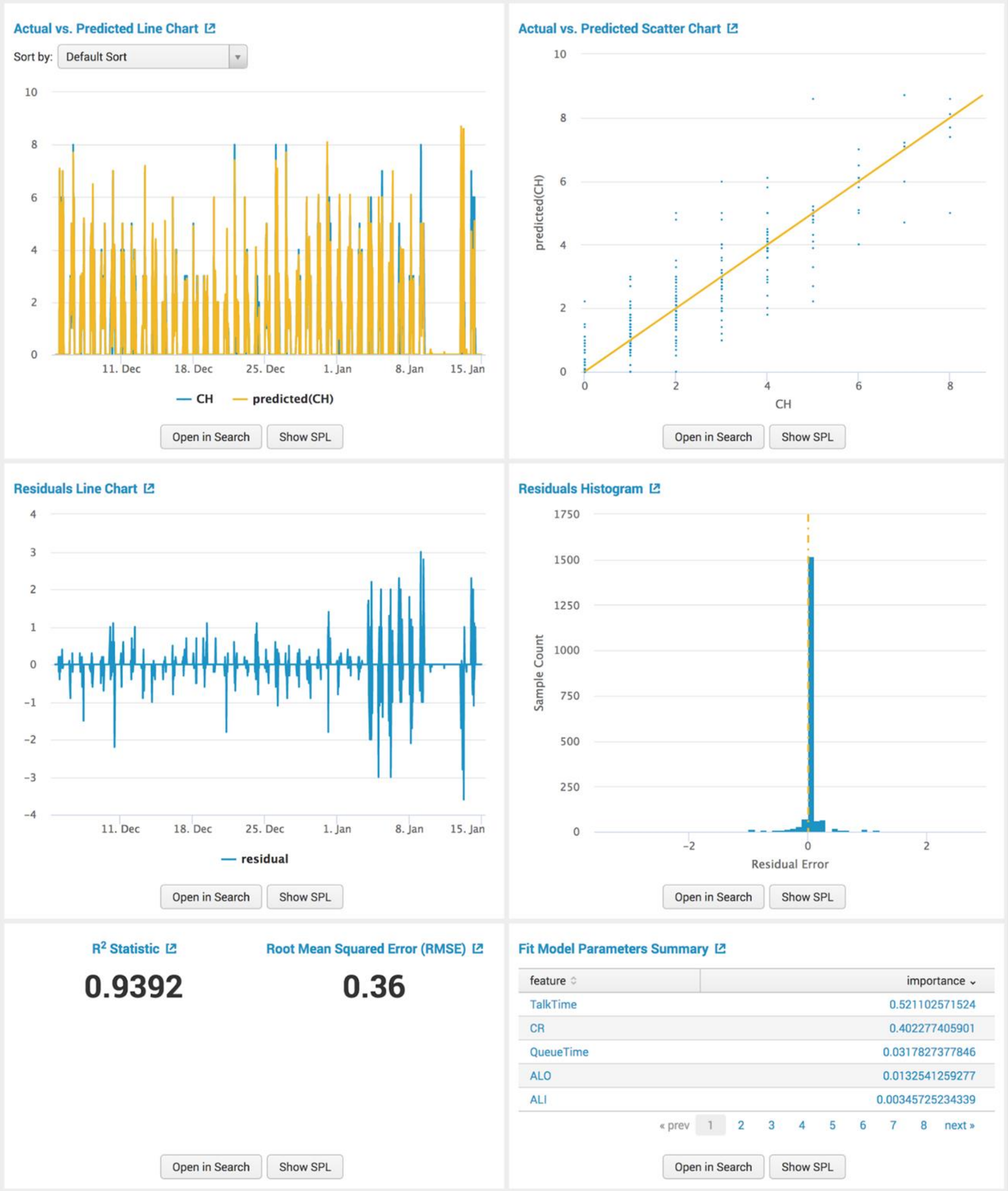
Customer Call Center



Call Center

KPI Calculation Metric: Average

Customer Call Center



R² Statistic

0.9392

Root Mean Squared Error (RMSE)

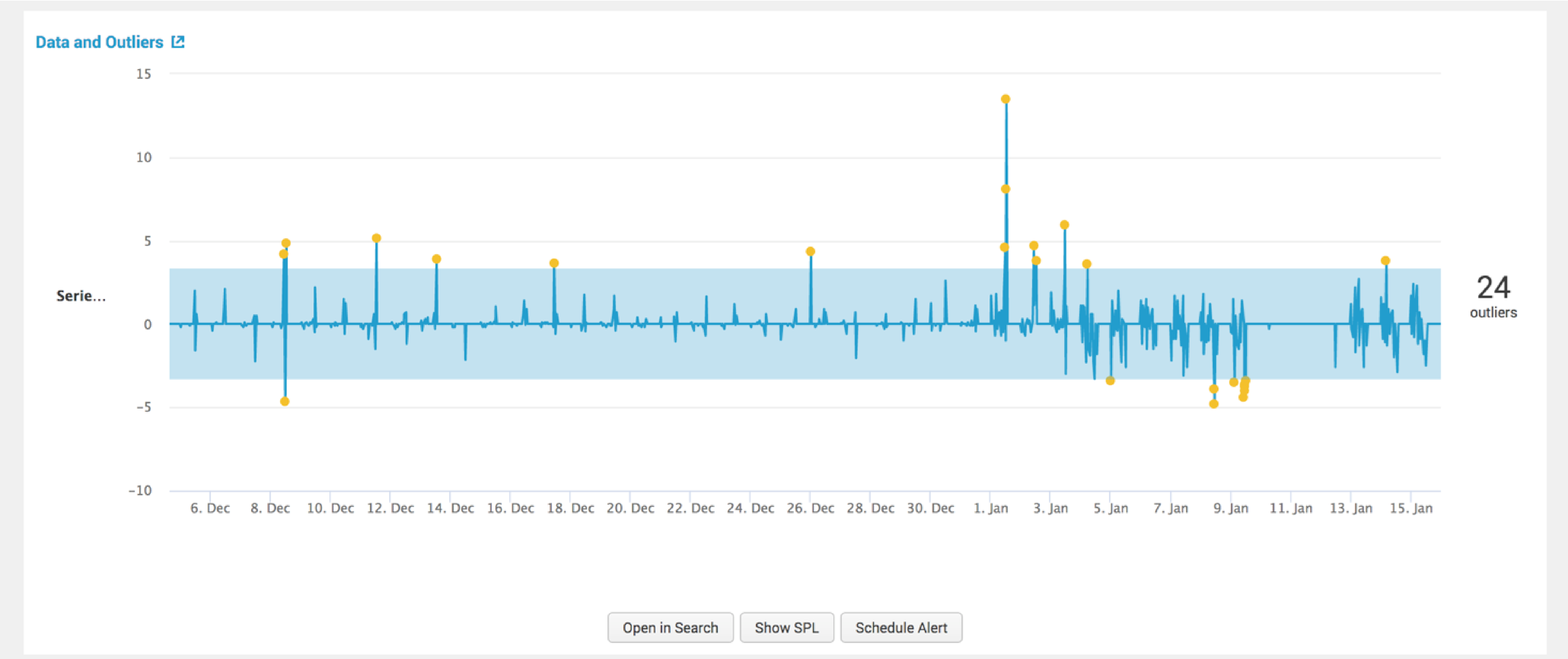
0.36

Fit Model Parameters Summary

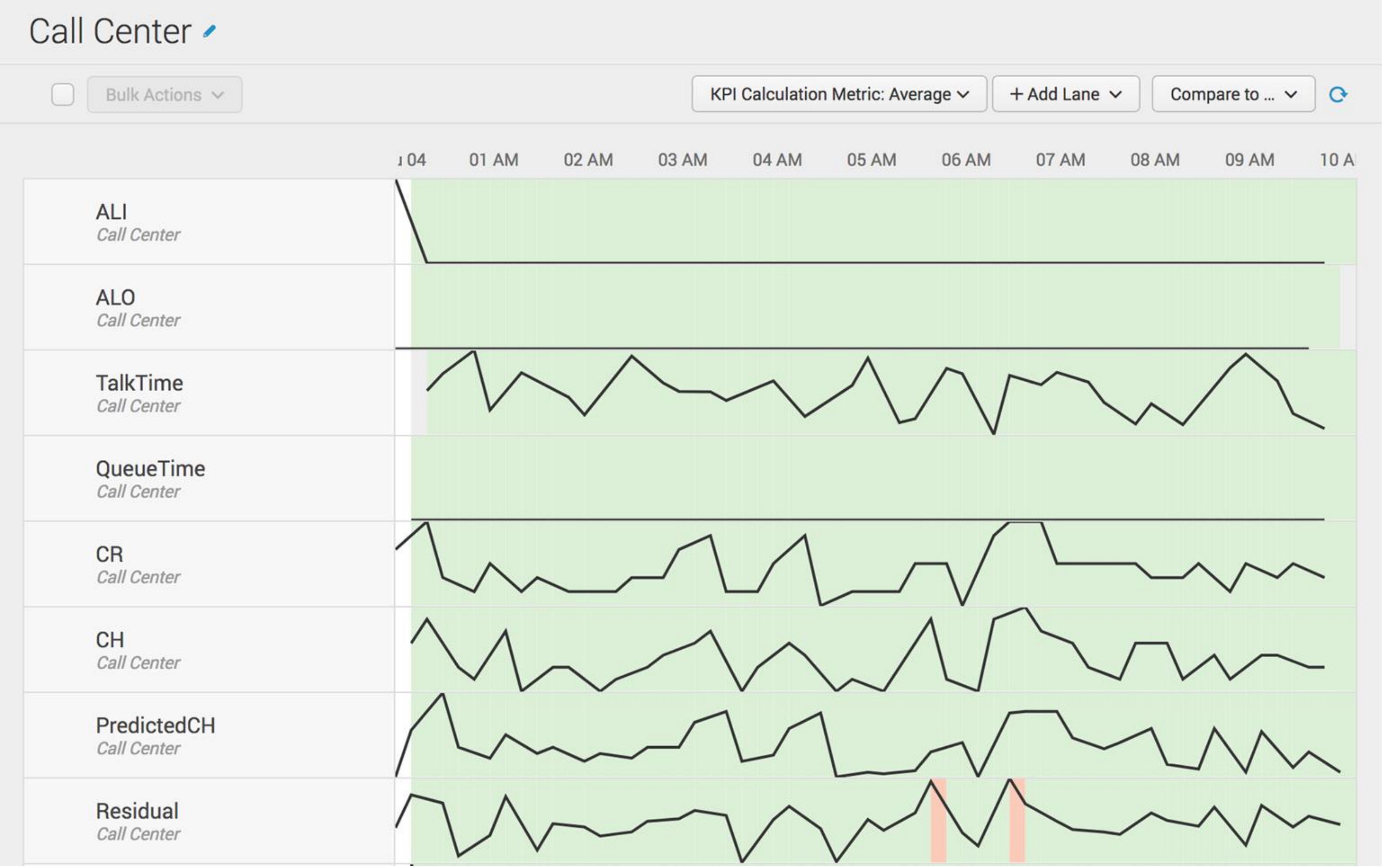
feature	importance
TalkTime	0.521102571524
CR	0.402277405901
QueueTime	0.0317827377846
ALO	0.0132541259277
ALI	0.00345725234339

« prev 1 2 3 4 5 6 7 8 next »

Customer Call Center



Customer Call Center



Assistant: Clustering

Sometimes you need to listen to your data!

I have data but… ?

What is the minimum requirements for each workflow

- Clustering

I have fields of data but I don’t have a target field to generalize a relationship. I just want to know what rows are similar or dissimilar and by how much?

I need to create a | table with the fields I plan on using. I **should** really scale the fields first with StandardScaler or RobustScaler in the preprocessing step, and I **should consider** using PCA to reduce the dimensions pre clustering, and I **should** convert text fields to meaningful numeric values.

The use of |table is not required, this is just good formatting step

App, Videos, and How to Blog:

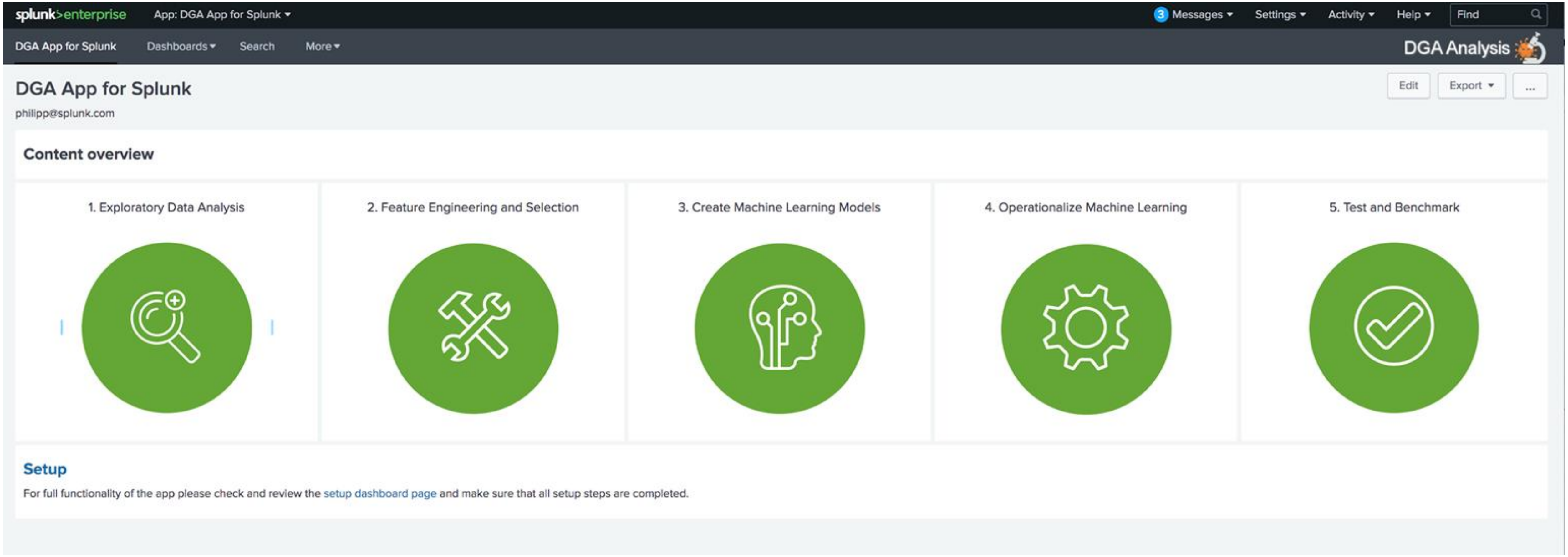
DGA app on Splunkbase

DGA videos on Splunk Videos

Anomalies like Neapolitan Ice Cream

To the DGA app!

Demo Time!



DGA App for Splunk

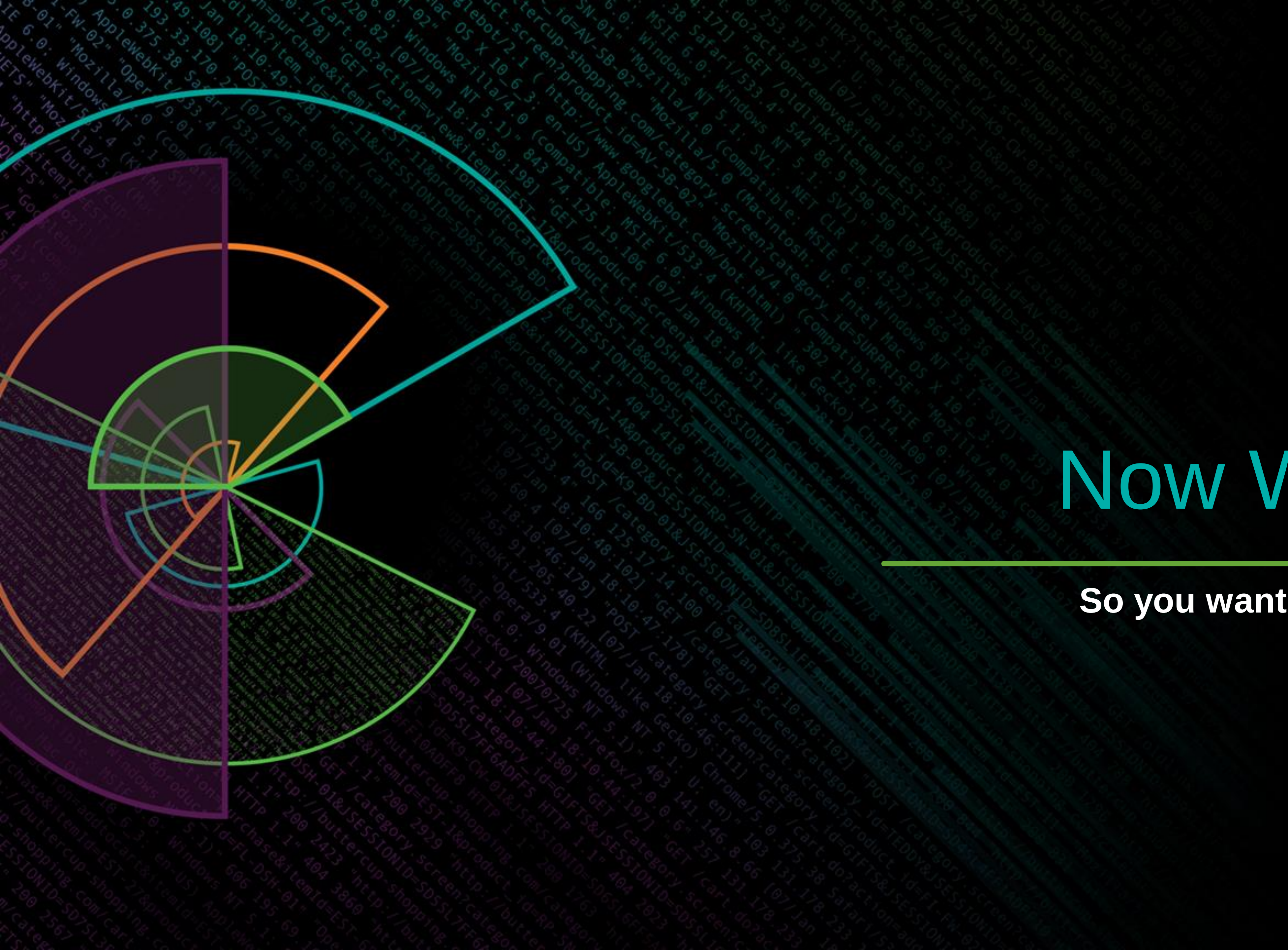
philipp@splunk.com

Content overview

1. Exploratory Data Analysis
2. Feature Engineering and Selection
3. Create Machine Learning Models
4. Operationalize Machine Learning
5. Test and Benchmark

Setup

For full functionality of the app please check and review the setup dashboard page and make sure that all setup steps are completed.

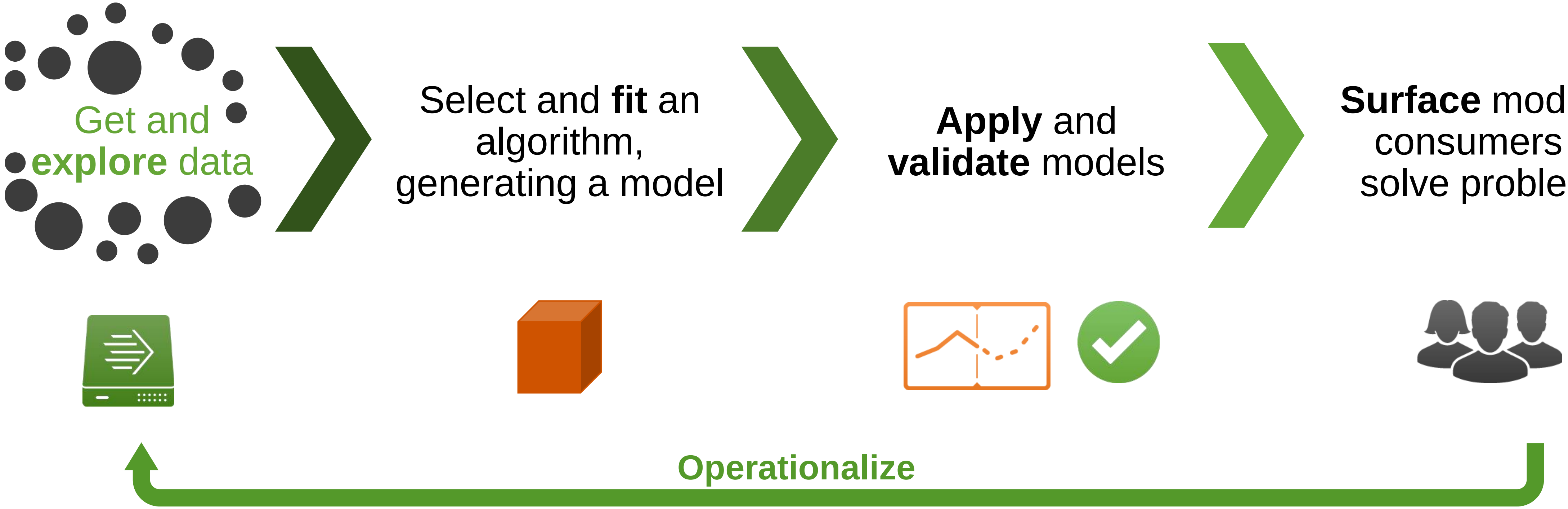


Now What?

So you want some ML ?

How do you replicate at your company?

Problem: <Stuff in the world> causes big time and money expense.
Solution: Build ML model to learn the behaviors at scale and take action



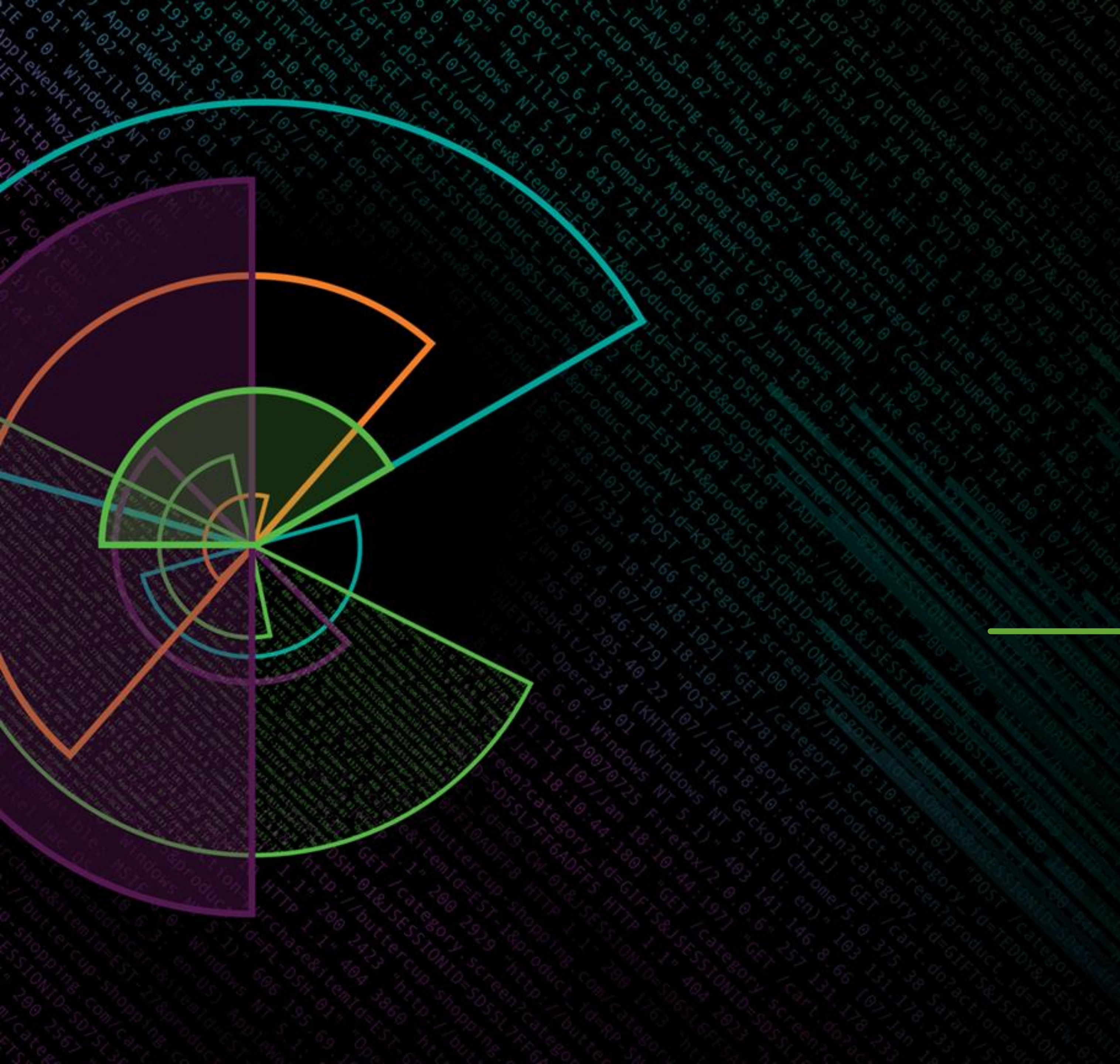
Get and **explore** data

Select and **fit** an algorithm, generating a model

Apply and **validate** models

Surface model to consumers to solve problems

Operationalize



END of LINE

Q&A