Splunk4Ninjas - Machine Learning

Hands on Introduction to the
Splunk Machine Learning Toolkit (MLTK)

splunk>

Forward-Looking Statements

This presentation may contain forward-looking statements regarding future events, plans or the expected financial performance of our company, including our expectations regarding our products, technology, strategy, customers, markets, acquisitions and investments. These statements reflect management's current expectations, estimates and assumptions based on the information currently available to us. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation.

For additional information about factors that could cause actual results to differ materially from those described in the forward-looking statements made in this presentation, please refer to our periodic reports and other filings with the SEC, including the risk factors identified in our most recent quarterly reports on Form 10-Q and annual reports on Form 10-K, copies of which may be obtained by visiting the Splunk Investor Relations website at www.investors.splunk.com or the SEC's website at www.sec.gov. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by us, on our website or otherwise, it may not contain current or accurate information. We disclaim any obligation to update or revise any forward-looking statement based on new information, future events or otherwise, except as required by applicable law.

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. We undertake no obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release.

Agenda

- Welcome/Introduction
- Intro to Machine Learning at Splunk
- Demo of Machine Learning Toolkit (MLTK) with Q&A
- Intro to the Trackday Dataset
- Four Different Challenges (~30 mins each)

  Challenge 1 - Explore the track_day.csv Dataset

  Challenge 2 - Detect Numeric Outliers

  Challenge 3 - Supervised Learning: Predict Categorical Fields

  Challenge 4 - Unsupervised Learning: Clustering
- Wrap Up, Discussion and Feedback

Disclaimer

What this session is not about and what it is about

- **NO** replacement for a PhD in machine learning, data science or AI
- **NO** replacement for Splunk's Education class for Data Science
- **NO** comprehensive lecture about all possible concepts and algorithms in ML … but,

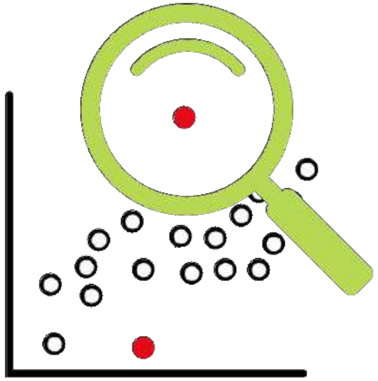
- **YES** first introduction into Machine Learning @ Splunk
- **YES** getting to know of Splunk's Machine Learning Toolkit
- **YES** guided hands-on challenges to explore a few typical ML tasks

Machine Learning Tour

splunk>

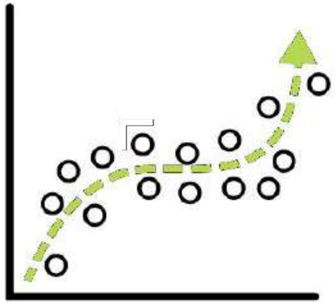
Splunk Customers Want Answers from their Data

Anomaly detection



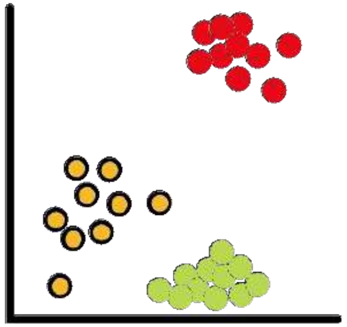
- Deviation from past behavior
- Deviation from peers
- (aka Multivariate AD or Cohesive AD)
- Unusual change in features
- **ITSI MAD**

Predictive Analytics



- Predict Service Health Score/Churn
- Predicting Events
- Trend Forecasting
- Detecting influencing entities
- Early warning of failure

Clustering

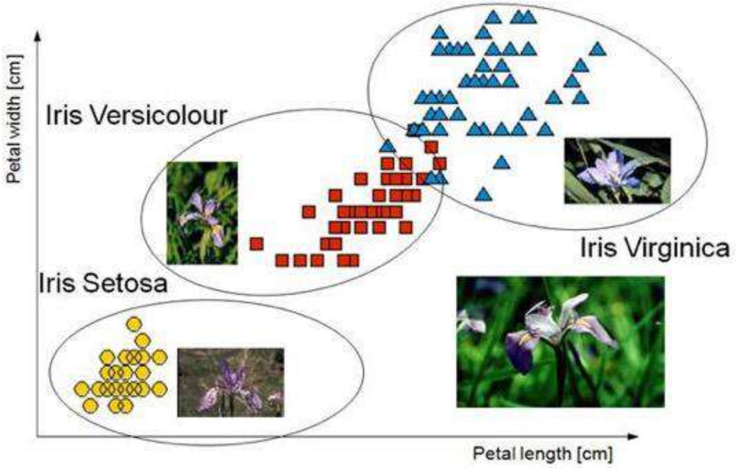


- Identify peer groups
- Event Correlation
- Reduce alert noise
- Behavioral Analytics
- **ITSI Event Analytics**

Types of Machine Learning

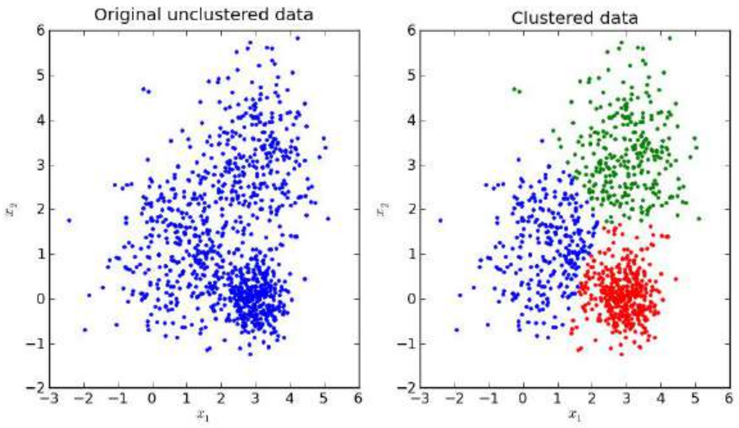
Supervised Learning (labeled data)

- Regression
- Classification



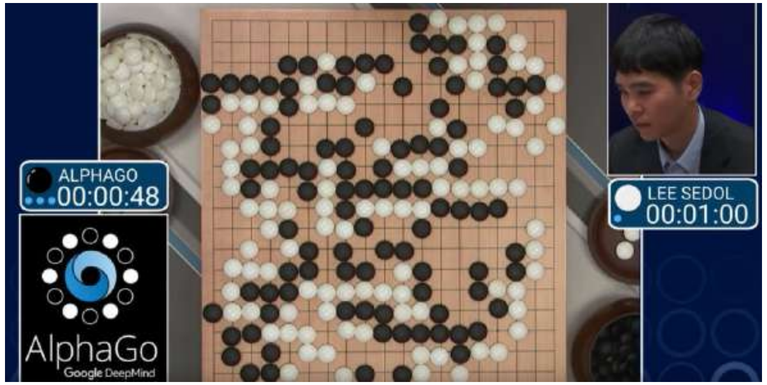
Unsupervised Learning (unlabeled data)

- Clustering
- Anomaly Detection

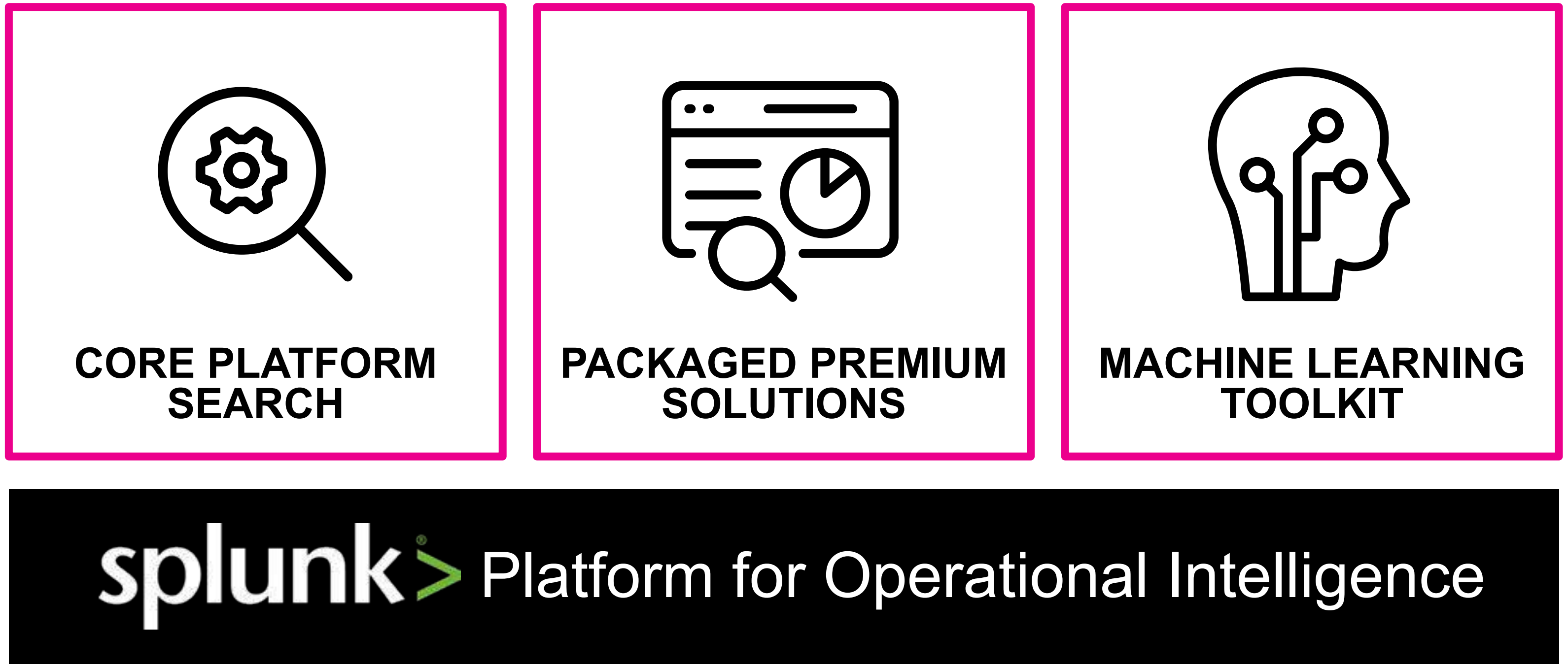


Mixed Models (with reinforcement or feedback)

- Human in the Loop
- Autonomous Systems



Overview of Machine Learning at Splunk



CORE PLATFORM SEARCH

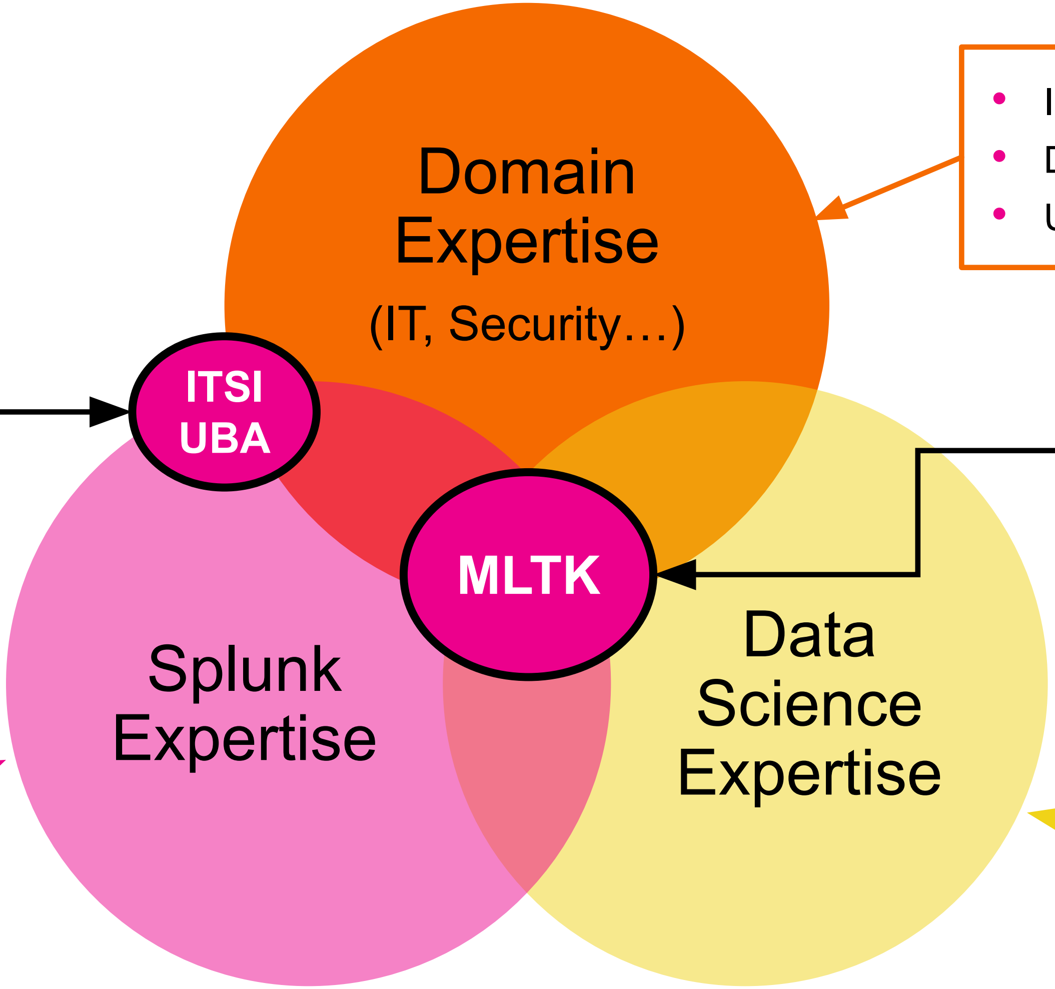
PACKAGED PREMIUM SOLUTIONS

MACHINE LEARNING TOOLKIT

splunk> Platform for Operational Intelligence

Skill Areas for Machine Learning @ Splunk

Premium solutions provide out of the box ML capabilities.



ITSI
UBA

Domain Expertise

(IT, Security…)

MLTK

Splunk Expertise

Data Science Expertise

- Identify use cases
- Drive decisions
- Understanding of business

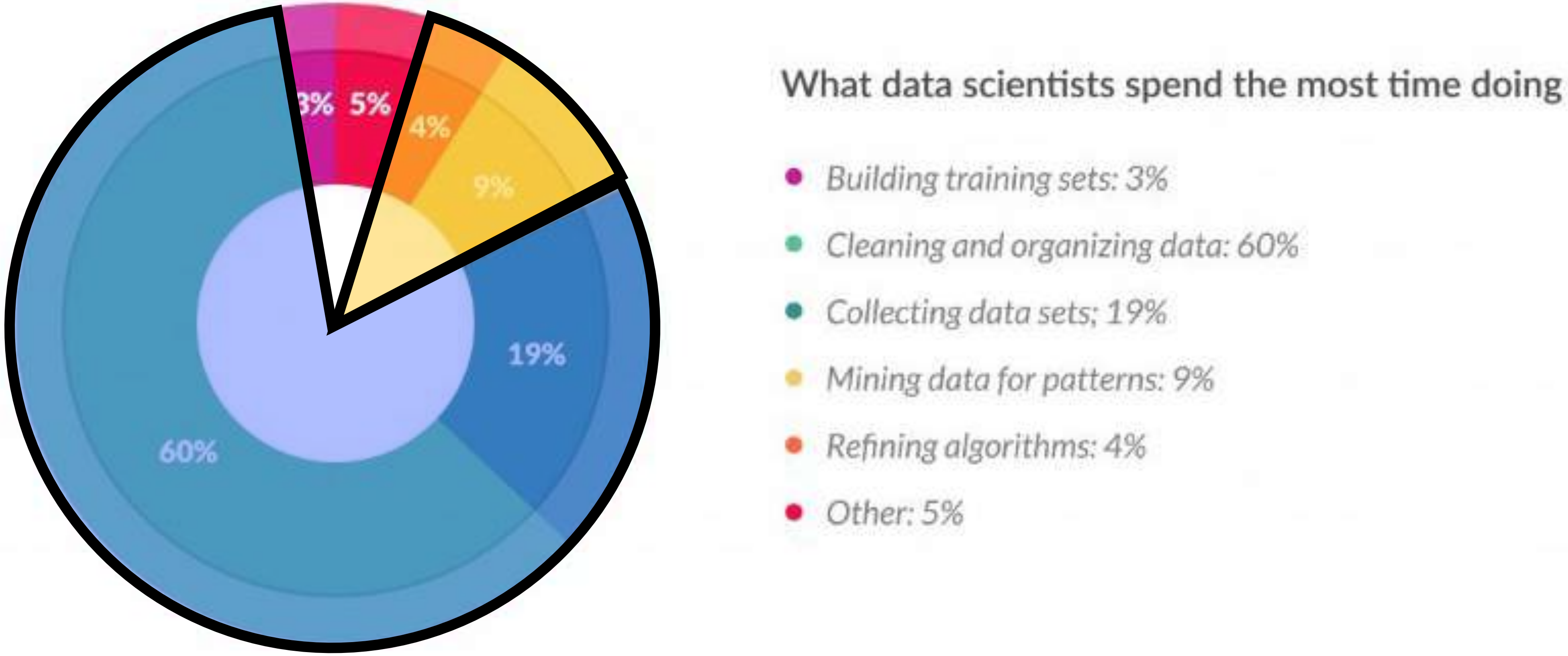
Splunk ML Toolkit facilitates and simplifies via examples & guidance

- Searching
- Reporting
- Alerting
- Workflow

- Statistics/math background
- Algorithm selection
- Model building

What Data Scientists Really Do

Data Preparation accounts for about 80% of the work of data scientists



“Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”, Forbes Mar 23, 2016

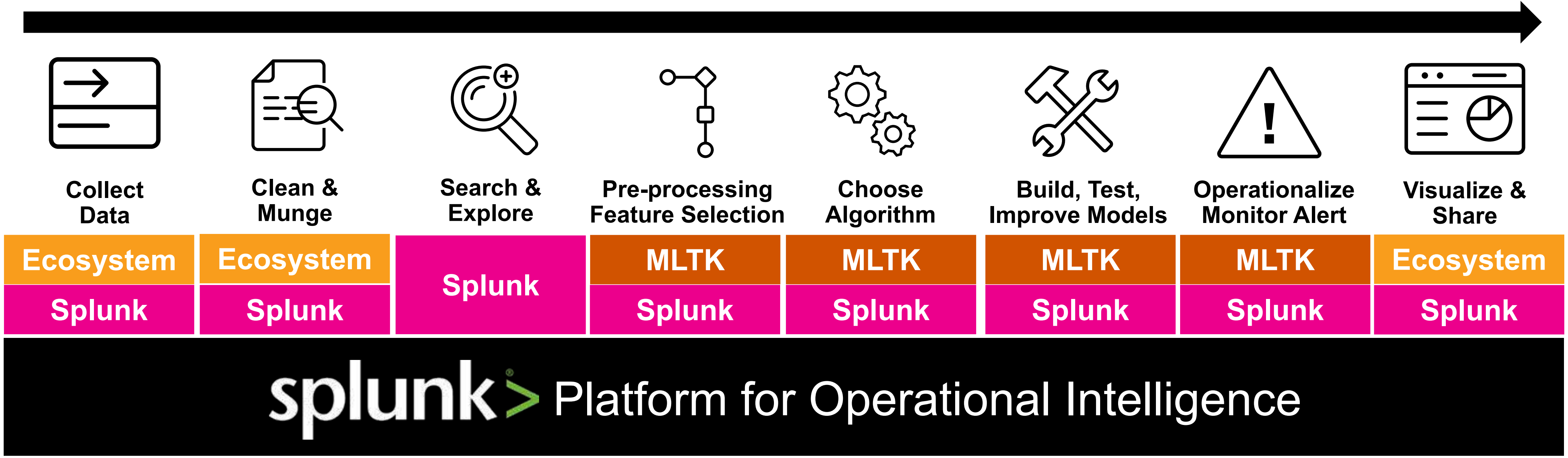
Custom ML with the Splunk Platform

Ecosystem Splunk's App Ecosystem contains 1000's of free add-ons for getting data in, applying structure and visualizing your data giving you faster time to value.

MLTK The Machine Learning Toolkit delivers new SPL commands, custom visualizations, assistants, and examples to explore a variety of ml concepts.

Splunk Splunk Enterprise is the mission-critical platform for indexing, searching, analyzing, alerting and visualizing machine data.

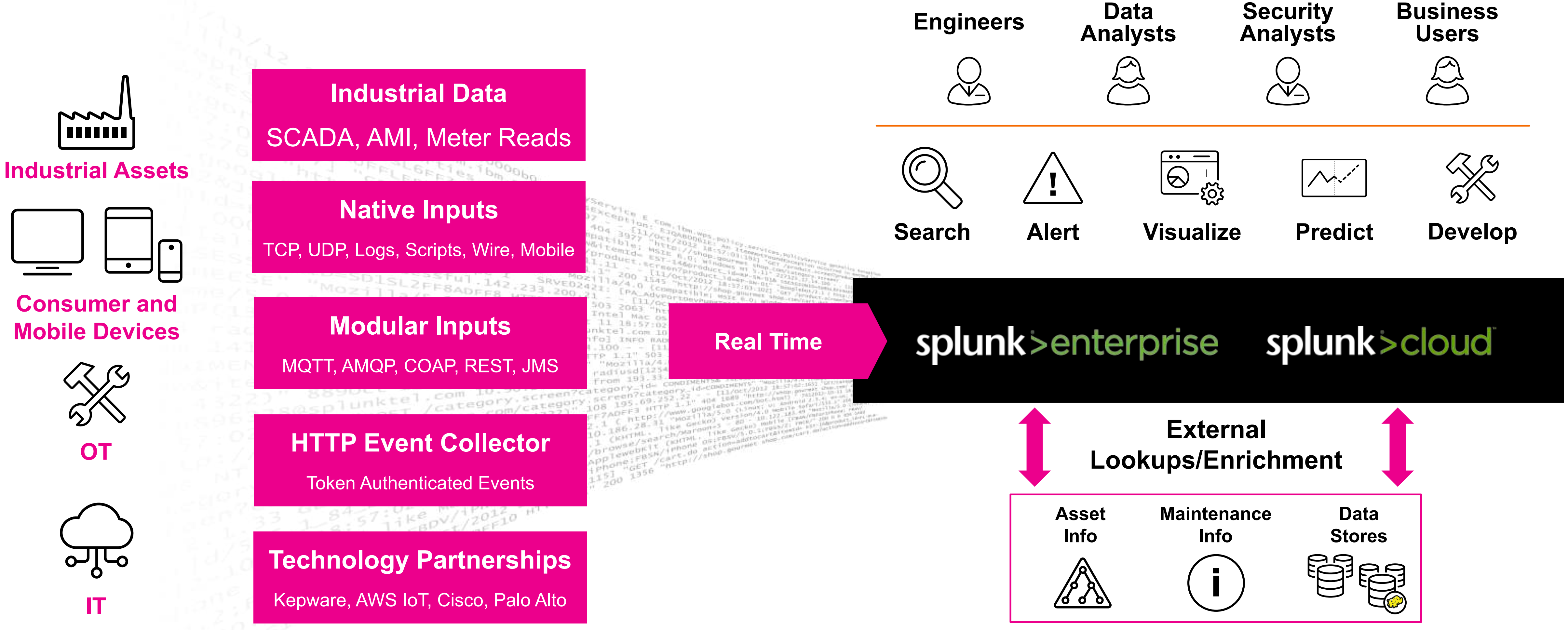
Operationalized Data Science Pipeline



Collect Data	Clean & Munge	Search & Explore	Pre-processing Feature Selection	Choose Algorithm	Build, Test, Improve Models	Operationalize Monitor Alert	Visualize & Share
Ecosystem	Ecosystem	Splunk	MLTK	MLTK	MLTK	MLTK	Ecosystem
Splunk	Splunk		Splunk	Splunk	Splunk	Splunk	Splunk

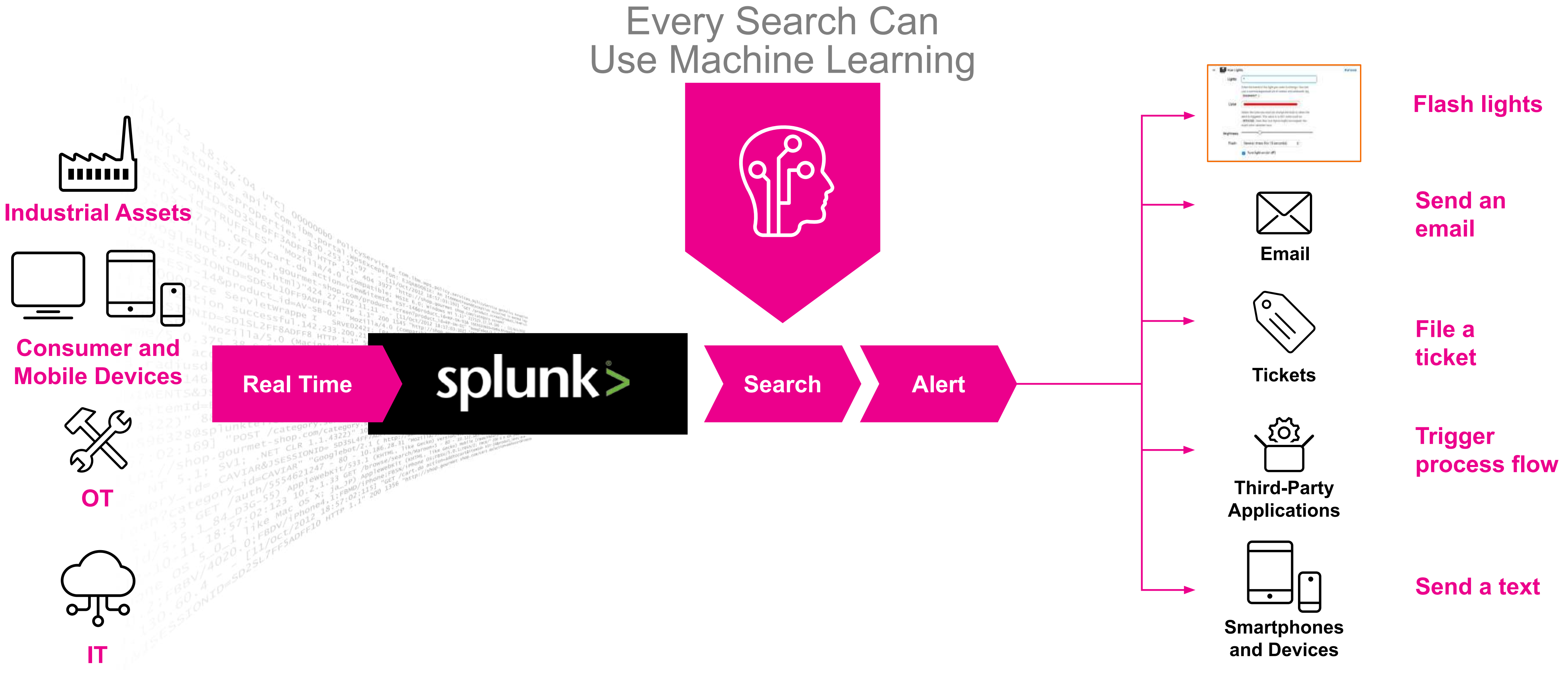
splunk> Platform for Operational Intelligence

Continuous Data Ingest at Scale



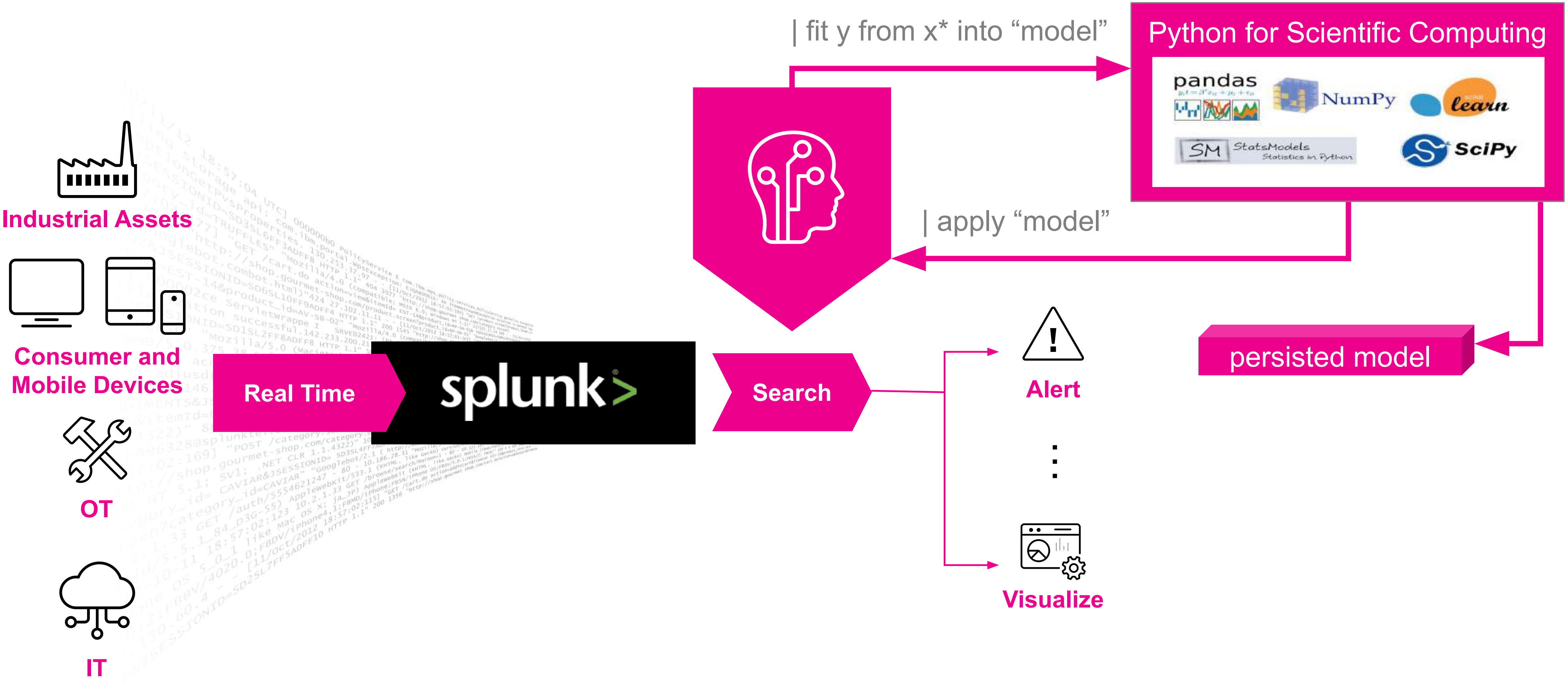
splunk>

Sense and Respond



Every Search Can Use Machine Learning

MLTK + Python for Scientific Computing



Splunk Machine Learning Toolkit (MLTK)

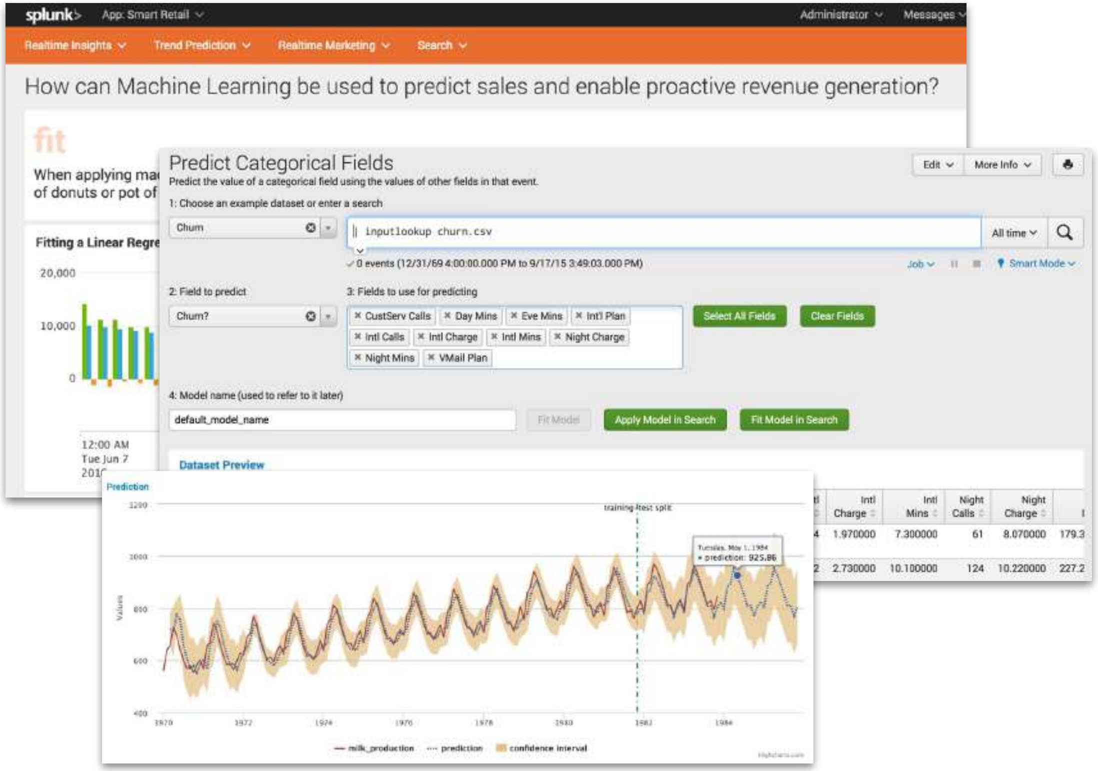
Extends Splunk platform functions and provides a guided modeling environment

Built for the Citizen Data Scientist

- **Experiments and Assistants**: Guided model building, testing, and deployment for common objectives
- **Algorithms**: 80+ standard algorithms (supervised & unsupervised)

Extensible to operationalize any use case

- **Python for Scientific Computing Library**: Access to 300+ open source algorithms
- **Deep Learning Toolkit** : Supports NN and GPU accelerated machine learning
- **ML-SPL API**: Import any open-source or proprietary algorithm



How can Machine Learning be used to predict sales and enable proactive revenue generation?

When applying ma
of donuts or pot of

Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event.

1: Choose an example dataset or enter a search

```
| inputlookup churn.csv
```

✓ 0 events (12/31/69 4:00:00.000 PM to 9/17/15 3:49:03.000 PM)

2: Field to predict

3: Fields to use for predicting

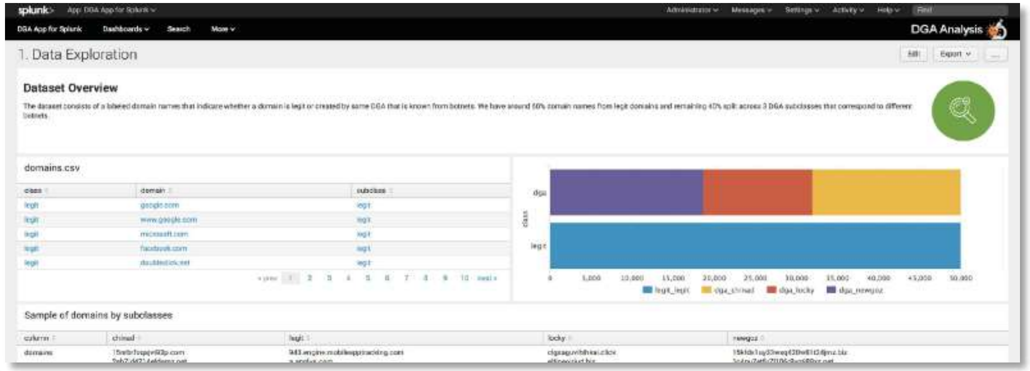
4: Model name (used to refer to it later)

default_model_name

tl	Intl Charge	Intl Mins	Night Calls	Night Charge	I
4	1.970000	7.300000	61	8.070000	179.3
2	2.730000	10.100000	124	10.220000	227.2

Example: MLTK powered DGA App for Splunk

Detect Malicious Domain Names using Machine Learning



1. Data Exploration

Dataset Overview

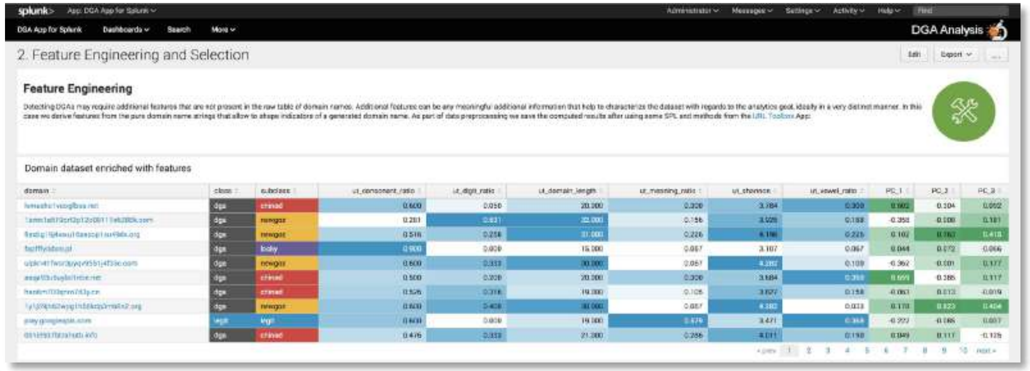
The dataset consists of a labeled domain names that indicate whether a domain is legit or created by some DGA that is known from botnets. We have around 60% domain names from legit domains and remaining 40% split across 3 DGA subclasses that correspond to different botnets.

domains.csv

class	domain	subclass
legit	google.com	legit
legit	www.google.com	legit
legit	microsoft.com	legit
legit	facebook.com	legit
legit	doubleclick.net	legit

Sample of domains by subclasses

column	chinad	legit	locky	newgoz
domains	15rebr5sqpv60p.com [illegible]	9d3.engine.mobileapptracking.com [illegible]	clgsagu[illegible].click [illegible]	15kfde1uy33weq[illegible].biz [illegible]



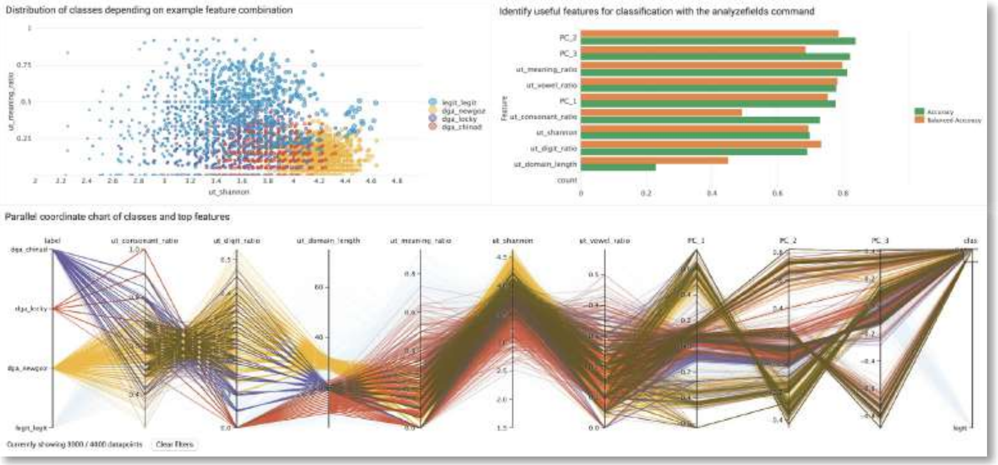
2. Feature Engineering and Selection

Feature Engineering

Detecting DGAs may require additional features that are not present in the raw table of domain names. Additional features can be any meaningful additional information that help to characterize the dataset with regards to the analytics goal, ideally in a very distinct manner. In this case we derive features from the pure domain name strings that allow to shape indicators of a generated domain name. As part of data preprocessing we save the computed results after using some SPL and methods from the URL Toolbox App.

Domain dataset enriched with features

domain	class	subclass	ut_consonant_ratio	ut_digit_ratio	ut_domain_length	ut_meaning_ratio	ut_shannon	ut_vowel_ratio	PC_1	PC_2	PC_3
[illegible]	dga	chinad	0.600	0.050	20.000	0.300	3.784	0.300	[illegible]	0.104	0.052
[illegible]	dga	newgoz	0.281	0.031	32.000	0.156	3.925	0.188	-0.350	0.108	0.181
[illegible]	dga	newgoz	0.516	0.258	31.000	0.226	4.198	0.226	0.102	0.162	0.418
[illegible]	dga	locky	0.900	0.000	15.000	0.067	3.107	0.067	0.044	0.072	-0.066
[illegible]	dga	newgoz	0.600	0.333	30.000	0.067	4.282	0.100	-0.362	0.001	0.177
[illegible]	dga	chinad	0.500	0.200	20.000	0.300	3.684	0.350	0.699	0.385	0.117
[illegible]	dga	chinad	0.526	0.316	19.000	0.105	3.827	0.158	-0.063	0.013	-0.019
[illegible]	dga	newgoz	0.600	0.400	30.000	0.067	4.383	0.033	0.170	[illegible]	0.404
play.googleapis.com	legit	legit	0.600	0.000	19.000	0.579	3.471	0.368	-0.222	-0.085	0.007
[illegible]	dga	chinad	0.476	0.333	21.000	0.286	4.011	0.190	0.049	0.117	-0.128

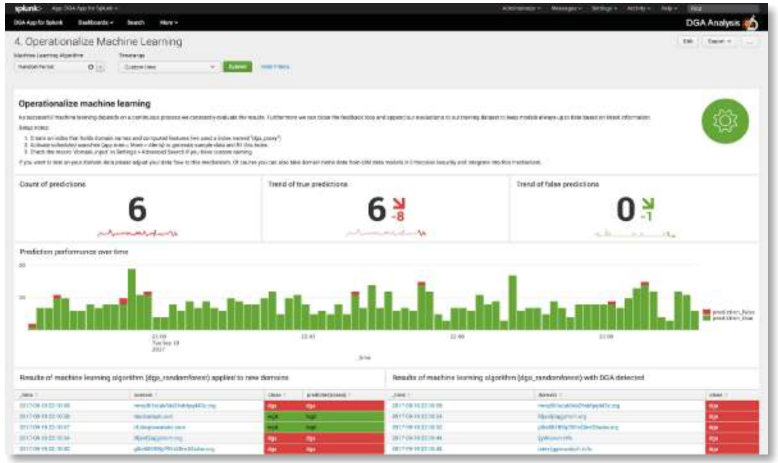


Distribution of classes depending on example feature combination

Identify useful features for classification with the analyzefields command

Parallel coordinate chart of classes and top features

Currently showing 1000 / 4400 datapoints



4. Operationalize Machine Learning

Operationalize machine learning

[illegible]

Count of predictions: 6

Trend of true predictions: 6 -8

Trend of false predictions: 0 -1

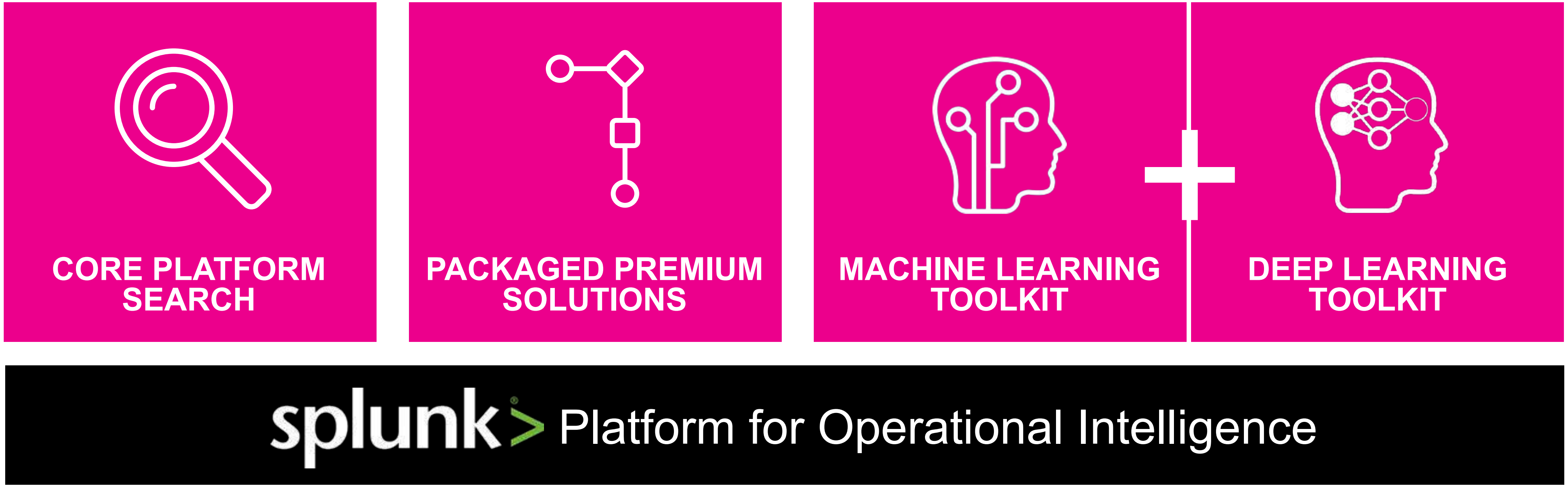
Prediction performance over time

Results of machine learning algorithm (dga_randomforest) applied to new domains

Results of machine learning algorithm (dga_randomforest) with DGA detected

Overview of ML including DL at Splunk

(not covered in this workshop)



CORE PLATFORM SEARCH

PACKAGED PREMIUM SOLUTIONS

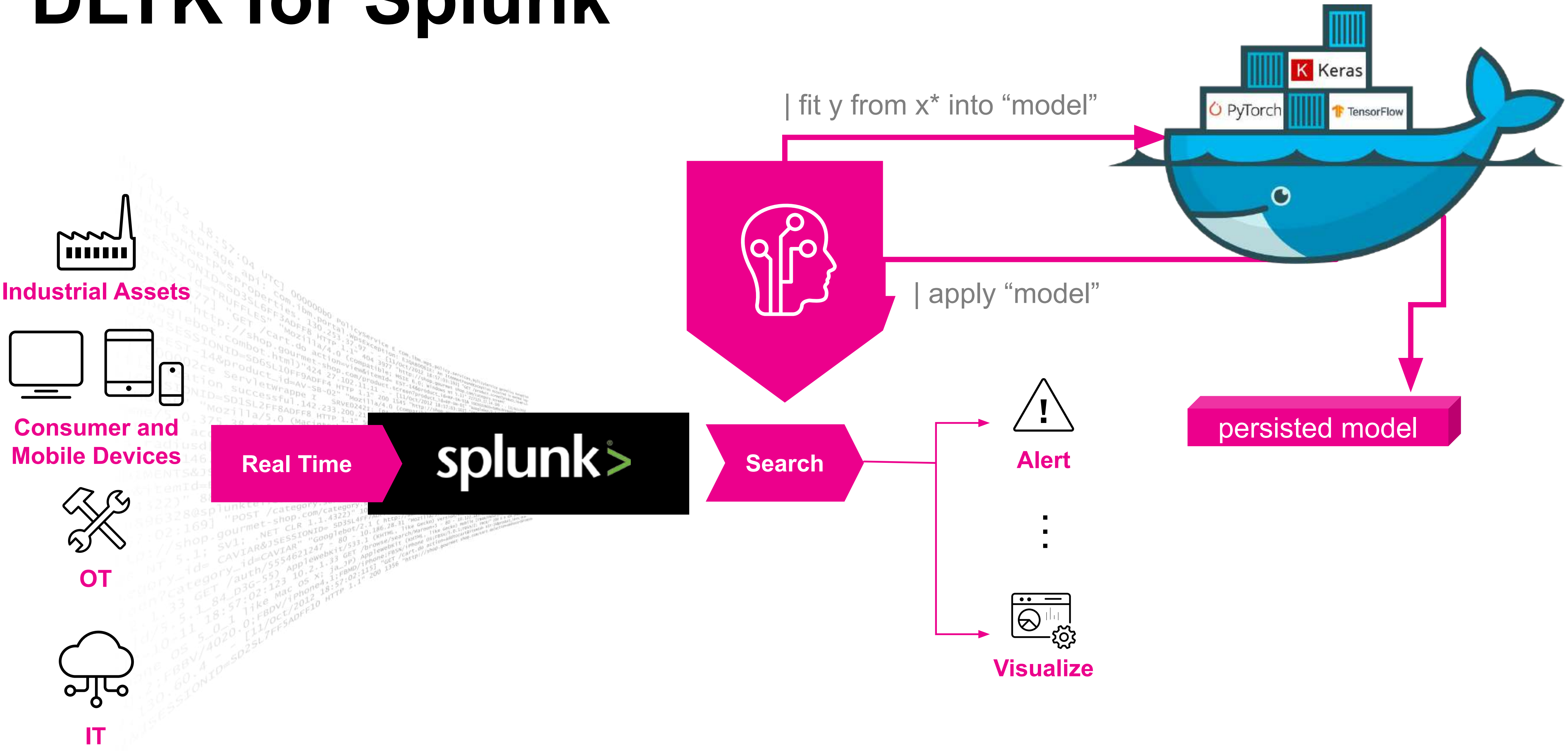
MACHINE LEARNING TOOLKIT

+

DEEP LEARNING TOOLKIT

splunk> Platform for Operational Intelligence

DLTK for Splunk



Hyatt ensured every customer experience with hotel wifi across the globe

1. All the data showing every customer sign in to the Hotel Wifi programs along with other relevant data (provider and lookup file of provider info, hotel ID and lookup file of hotel info, indexes for real time data of customers, local and global holidays).
2. A workflow (dashboard(s)) to show customers at every hotel signing in, normalize _time so a hotel in one time zone at 8 am to can be compared to another hotel at 8am.
3. Forecasting the likely wifi logins based on each property, day of week, and local and global holiday out two days into the future to show our expectations to executives.
4. Detecting meaningful anomalies as real time data comes in and is compared to the forecast . I must be able to insert business rules into the anomalies based on analyst feedback in a quick and nimble way. I should update my learning every night.

Demo: Machine Learning Toolkit

Hands-on Challenges

splunk>

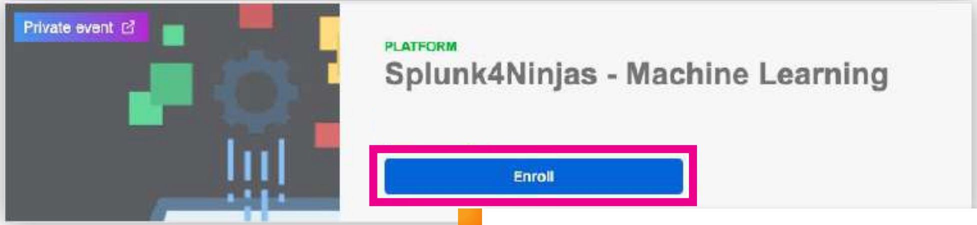
REGISTER 5 MINS

Enroll in Today’s Workshop

Tasks

1. Get a splunk.com account if you don’t have one yet: https://splk.it/SignUp
2. Enroll in the Splunk Show workshop event: https://show.splunk.com/event/<eventID>
3. Download a copy of today’s slide deck: https://splk.it/S4N-ML-Attendee

Goal



Enroll in today’s event

splunk>

Fun Facts about the Track Day dataset

A popular private event of racing and sportscar affine Splunkers in the early days.

Simple concept

Go on a race track, have fun and collect some car data to get insights about driving behavior etc.

A subset of this data is available in MLTK!

Image Source: https://www.youtube.com/watch?v=meBjI-ay9-U

Today’s Challenges

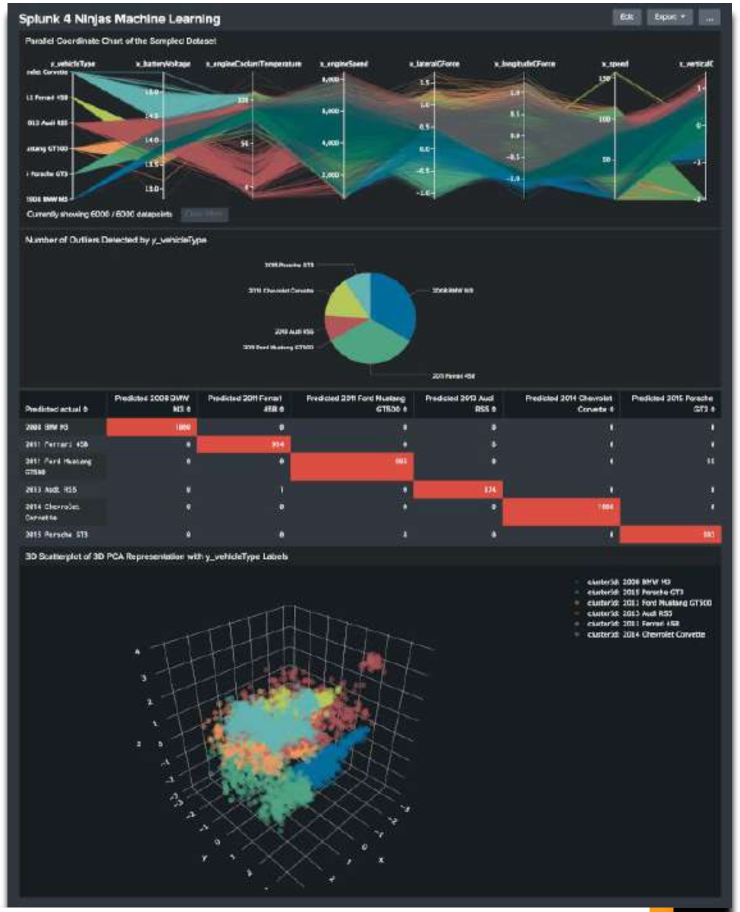
We are going to create four dashboards:

1. **Explore the Dataset:** Create a sample dataset and explore it using different types of visualizations such as SPL

2. **Detect Numeric Outliers:** Explore the MLTK showcase and adapt it to start a new experiment with your own dataset

3. **Use a Classification Model:** Create a classification model and use it to predict vehicle types from your sensor data

4. **Use a Clustering Model:** Create a clustering model and use it to analyze your dataset



Parallel Coordinate Chart of the Sampled Dataset

Currently showing 6000 / 6000 datapoints

Number of Outliers Detected by y_vehicleType

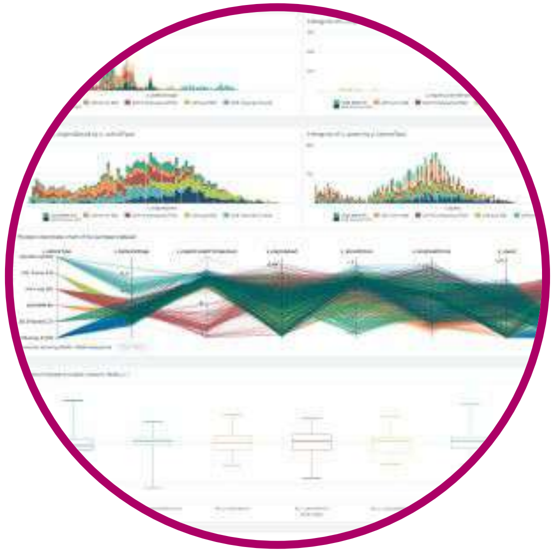
Predicted actual	Predicted 2008 BMW M3	Predicted 2011 Ferrari 458	Predicted 2011 Ford Mustang GT500	Predicted 2013 Audi RS5	Predicted 2014 Chevrolet Corvette	Predicted 2015 Porsche GT3
2008 BMW M3	1000	0	0	0	0	0
2011 Ferrari 458	0	994	0	5	0	0
2011 Ford Mustang GT500	0	0	[illegible]	0	0	11
2013 Audi RS5	0	1	0	974	0	0
2014 Chevrolet Corvette	0	0	0	0	1000	0
2015 Porsche GT3	0	0	2	0	0	[illegible]

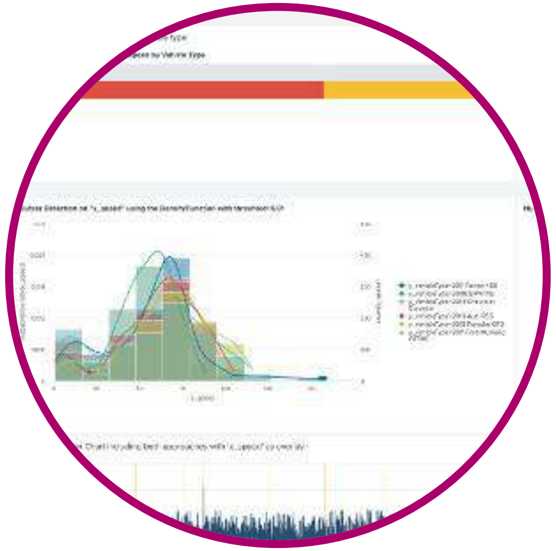
3D Scatterplot of 3D PCA Representation with y_vehicleType Labels

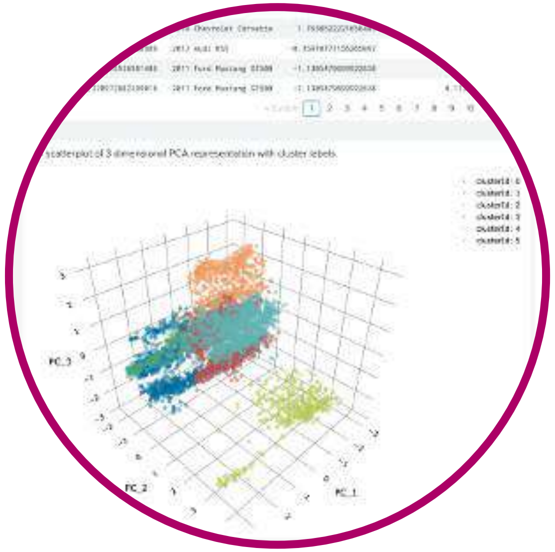
We’re aiming for a dashboard like this!

Workshop Goals

- Getting to know Splunk in the context of Machine Learning
- Prepare and analyze a dataset and summarize results on 4 dashboards

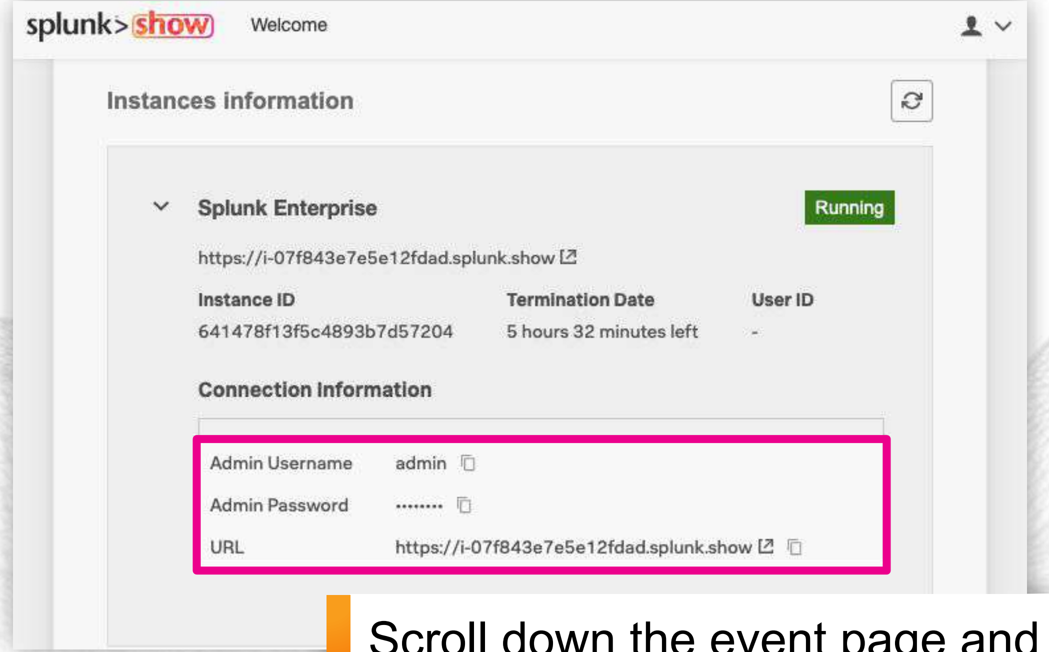






Login to Splunk

Locate your instance URL and credentials in the Splunk Show event https://show.splunk.com



Instances information

Splunk Enterprise — Running

https://i-07f843e7e5e12fdad.splunk.show

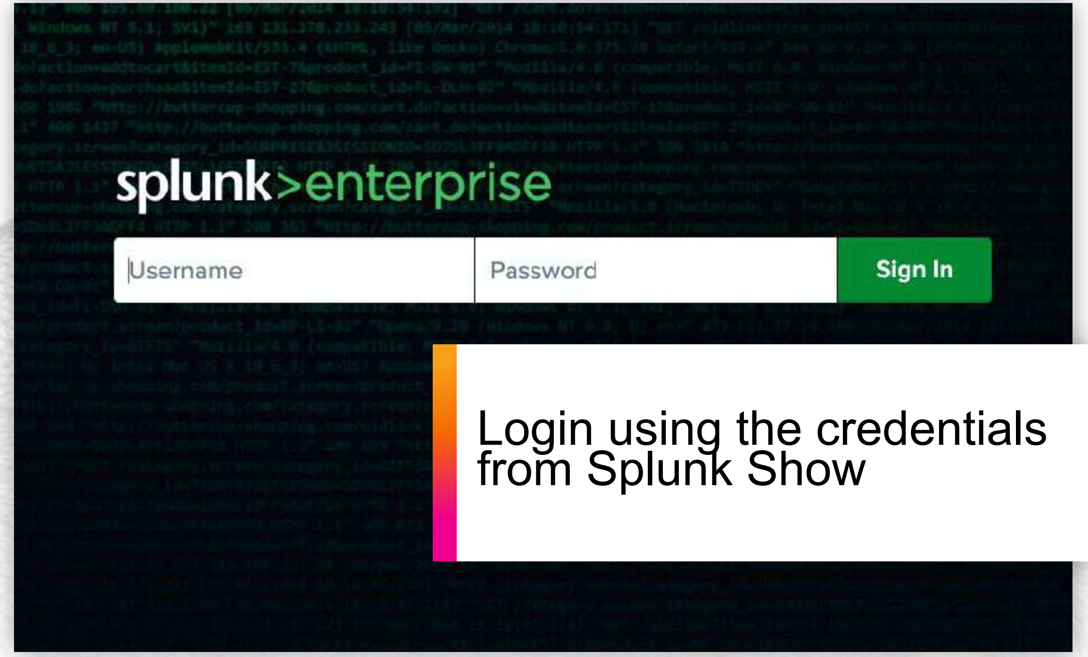
Instance ID	Termination Date	User ID
641478f13f5c4893b7d57204	5 hours 32 minutes left	-

Connection Information

Admin Username	admin
Admin Password	••••••••
URL	https://i-07f843e7e5e12fdad.splunk.show

Scroll down the event page and expand the **Splunk Enterprise** section to view your login details

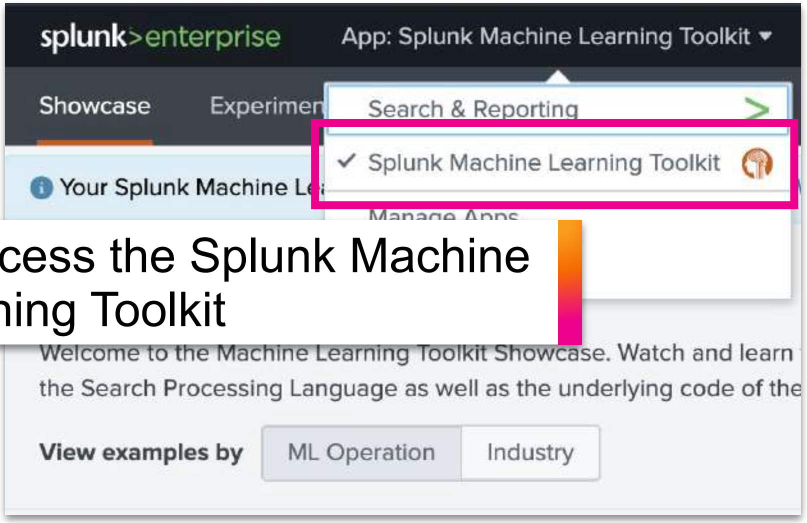
Log in to your Splunk instance



Login using the credentials from Splunk Show

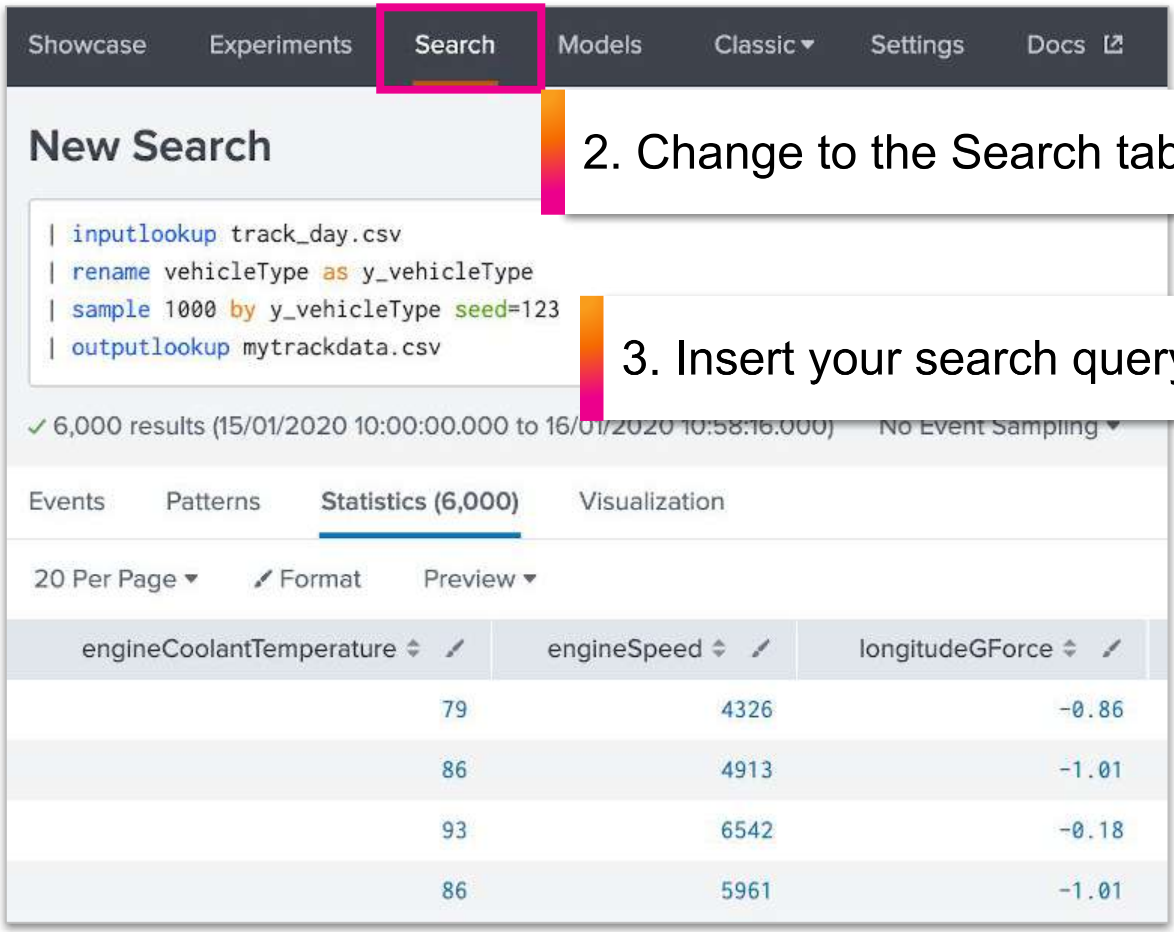
1

Create a Sample Dataset



Welcome to the Machine Learning Toolkit Showcase. Watch and learn the Search Processing Language as well as the underlying code of the

1. Access the Splunk Machine Learning Toolkit



2. Change to the Search tab

```
| inputlookup track_day.csv
| rename vehicleType as y_vehicleType
| sample 1000 by y_vehicleType seed=123
| outputlookup mytrackdata.csv
```

3. Insert your search query

✓ 6,000 results (15/01/2020 10:00:00.000 to 16/01/2020 10:58:16.000)

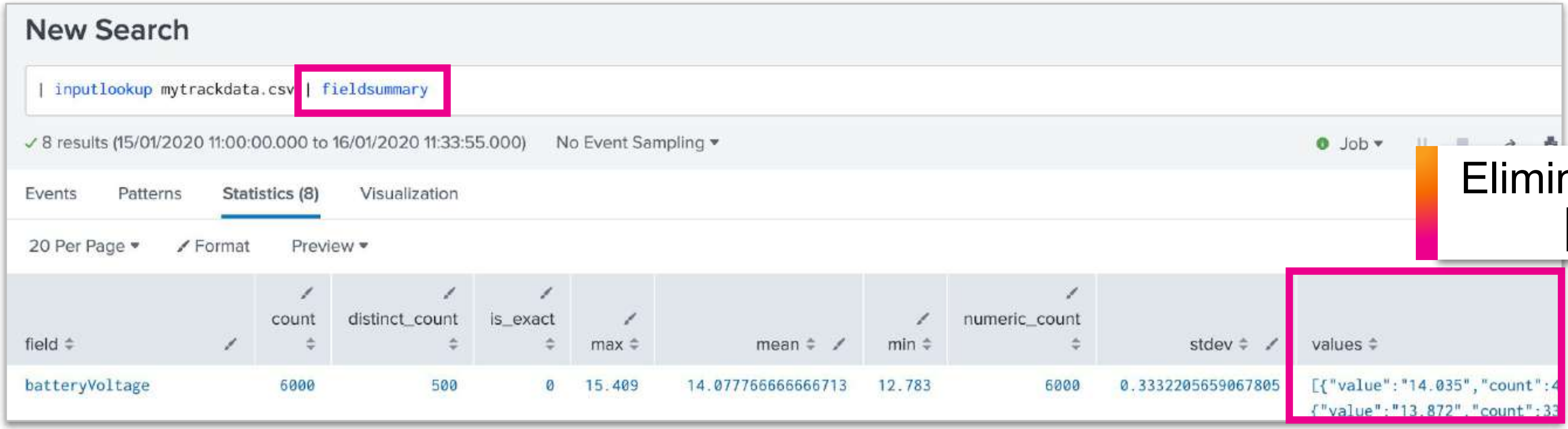
engineCoolantTemperature	engineSpeed	longitudeGForce
79	4326	-0.86
86	4913	-1.01
93	6542	-0.18
86	5961	-1.01

?

What's the benefit of renaming variables?

1

Use Fieldsummary to Explore your Dataset



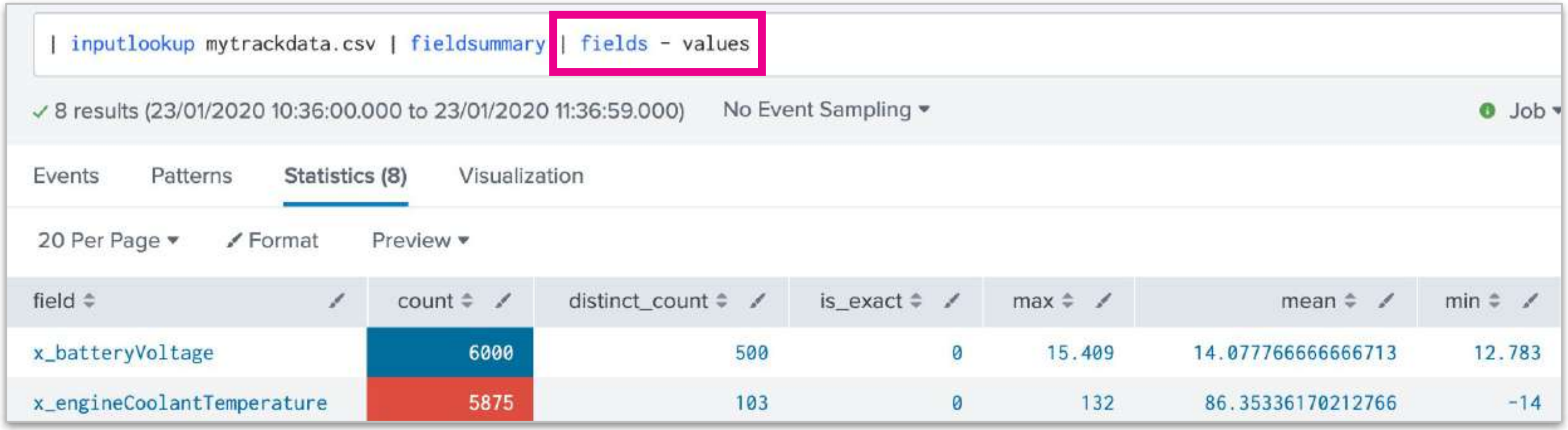
New Search

```
| inputlookup mytrackdata.csv | fieldsummary
```

✓ 8 results (15/01/2020 11:00:00.000 to 16/01/2020 11:33:55.000) No Event Sampling ▾

field	count	distinct_count	is_exact	max	mean	min	numeric_count	stdev	values
batteryVoltage	6000	500	0	15.409	14.077766666666713	12.783	6000	0.3332205659067805	[{"value":"14.035","count":4 {"value":"13.872","count":33

Eliminate unwanted Fields with
| fields - values



```
| inputlookup mytrackdata.csv | fieldsummary | fields - values
```

✓ 8 results (23/01/2020 10:36:00.000 to 23/01/2020 11:36:59.000) No Event Sampling ▾

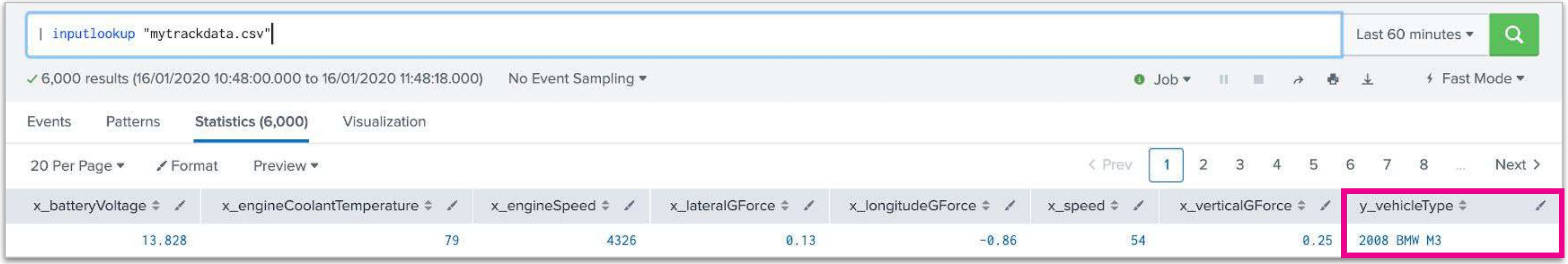
field	count	distinct_count	is_exact	max	mean	min
x_batteryVoltage	6000	500	0	15.409	14.077766666666713	12.783
x_engineCoolantTemperature	5875	103	0	132	86.35336170212766	-14

?

What’s going on with the engine coolant temperature?

1

Explore your Dataset with Visualizations

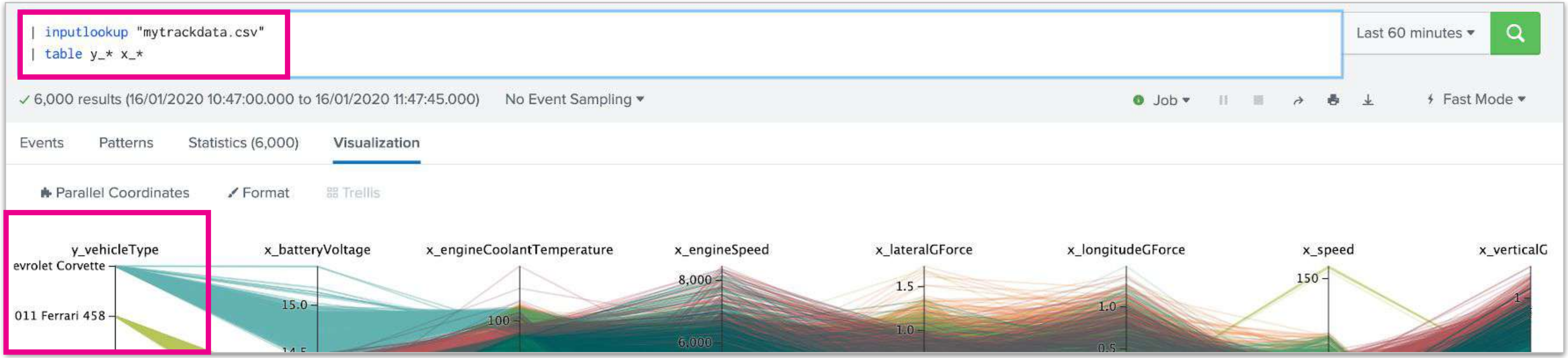


```
| inputlookup "mytrackdata.csv"
```

Last 60 minutes ▾

✓ 6,000 results (16/01/2020 10:48:00.000 to 16/01/2020 11:48:18.000) No Event Sampling ▾

x_batteryVoltage	x_engineCoolantTemperature	x_engineSpeed	x_lateralGForce	x_longitudeGForce	x_speed	x_verticalGForce	y_vehicleType
13.828	79	4326	0.13	-0.86	54	0.25	2008 BMW M3



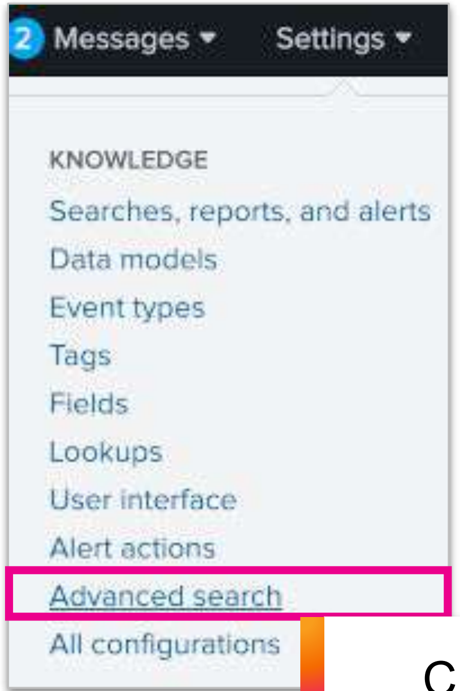
```
| inputlookup "mytrackdata.csv"
| table y_* x_*
```

Last 60 minutes ▾

✓ 6,000 results (16/01/2020 10:47:00.000 to 16/01/2020 11:47:45.000) No Event Sampling ▾

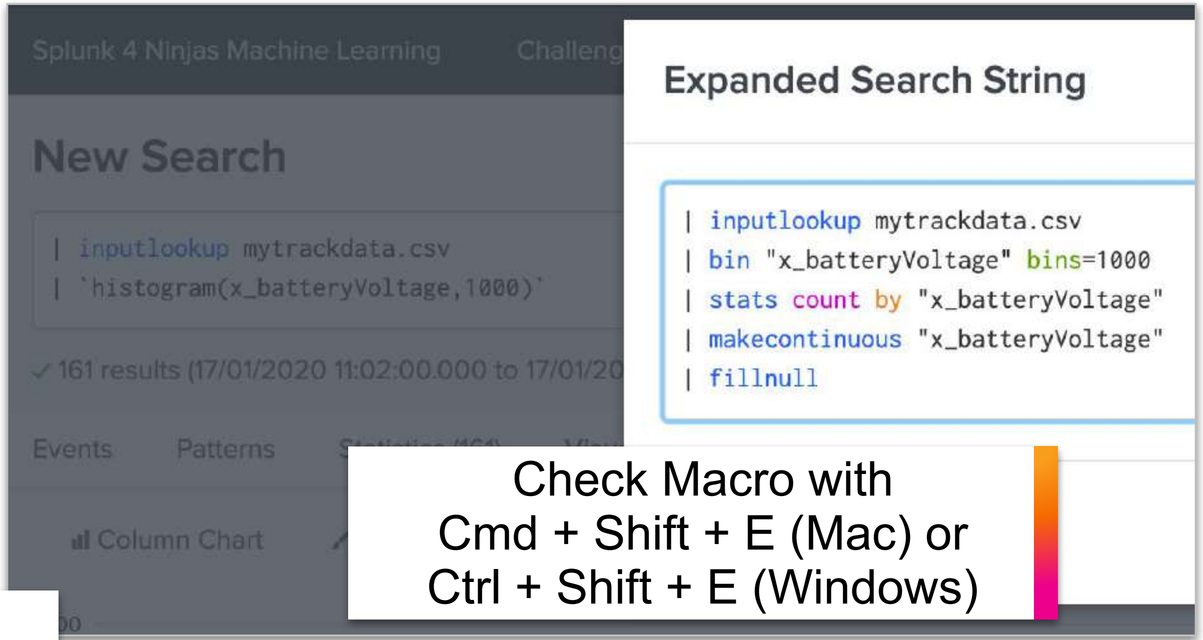
1

Using Splunk’s Histogram Macro



Check Macro in settings

OR



```
| inputlookup mytrackdata.csv
| `histogram(x_batteryVoltage,1000)`
```

✓ 161 results (17/01/2020 11:02:00.000 to 17/01/20

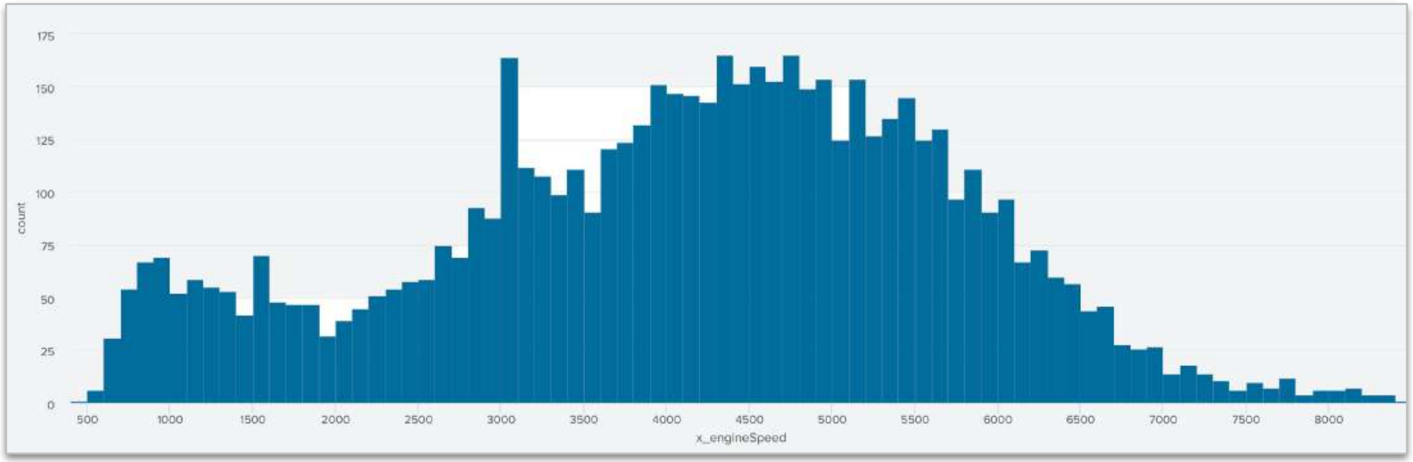
Expanded Search String

```
| inputlookup mytrackdata.csv
| bin "x_batteryVoltage" bins=1000
| stats count by "x_batteryVoltage"
| makecontinuous "x_batteryVoltage"
| fillnull
```

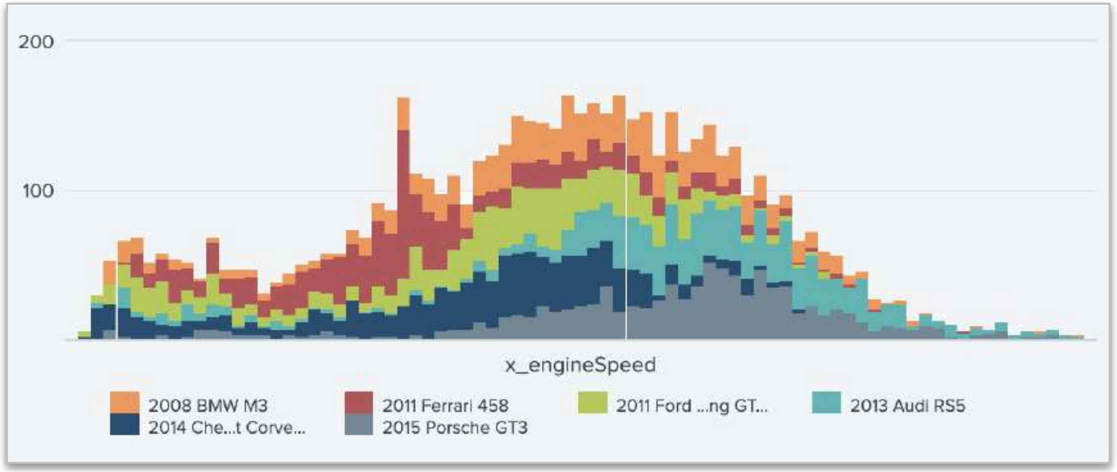
Check Macro with
Cmd + Shift + E (Mac) or
Ctrl + Shift + E (Windows)

1

Adjusting the Histogram Macro



How can we get from the top to the bottom histogram?



1

Adjust the Macro to Split by Vehicle Type

```
| chart count by y_vehicleType
```

vehicleType	count
Ferrari	641
Audi	42
BMW	51
Chevrolet	44
Ford	95

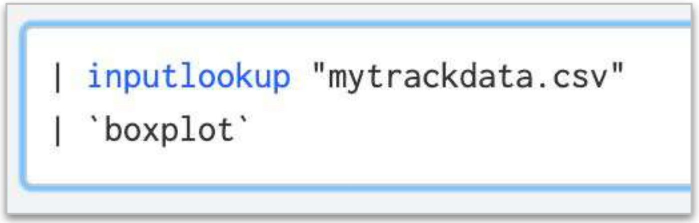
```
| chart count over x_engineSpeed by y_vehicleType
```

batteryVoltage	Ferrari	Audi	BMW	Chevrolet	Ford
13	0	0	0	0	1
14	0	0	0	1	1
15	1	0	1	0	0
16	1	1	1	0	0
17	1	0	0	0	1

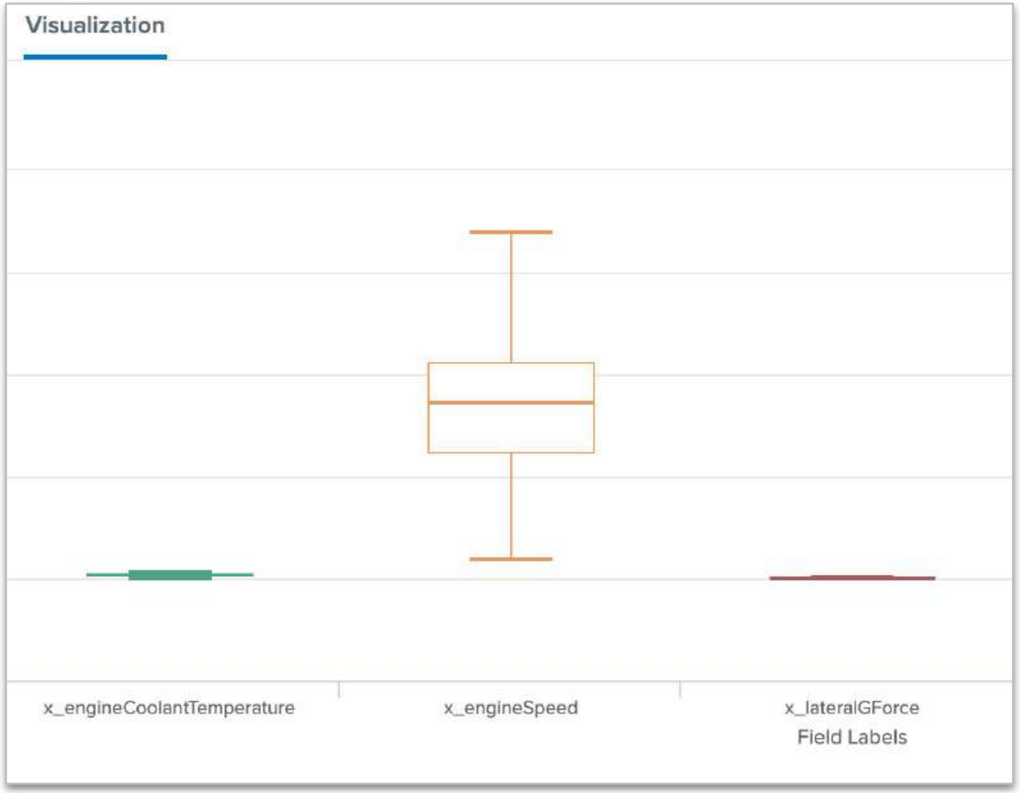
1

Working with the Boxplot Macro

How can this query be improved?



```
| inputlookup "mytrackdata.csv"
| `boxplot`
```

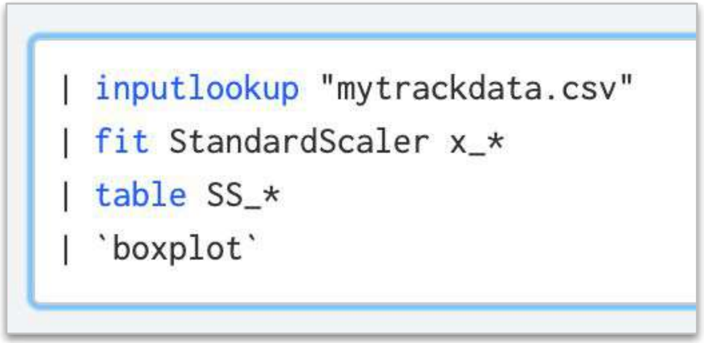


Hints:

Scale numeric values using the `fit` command with the StandardScaler

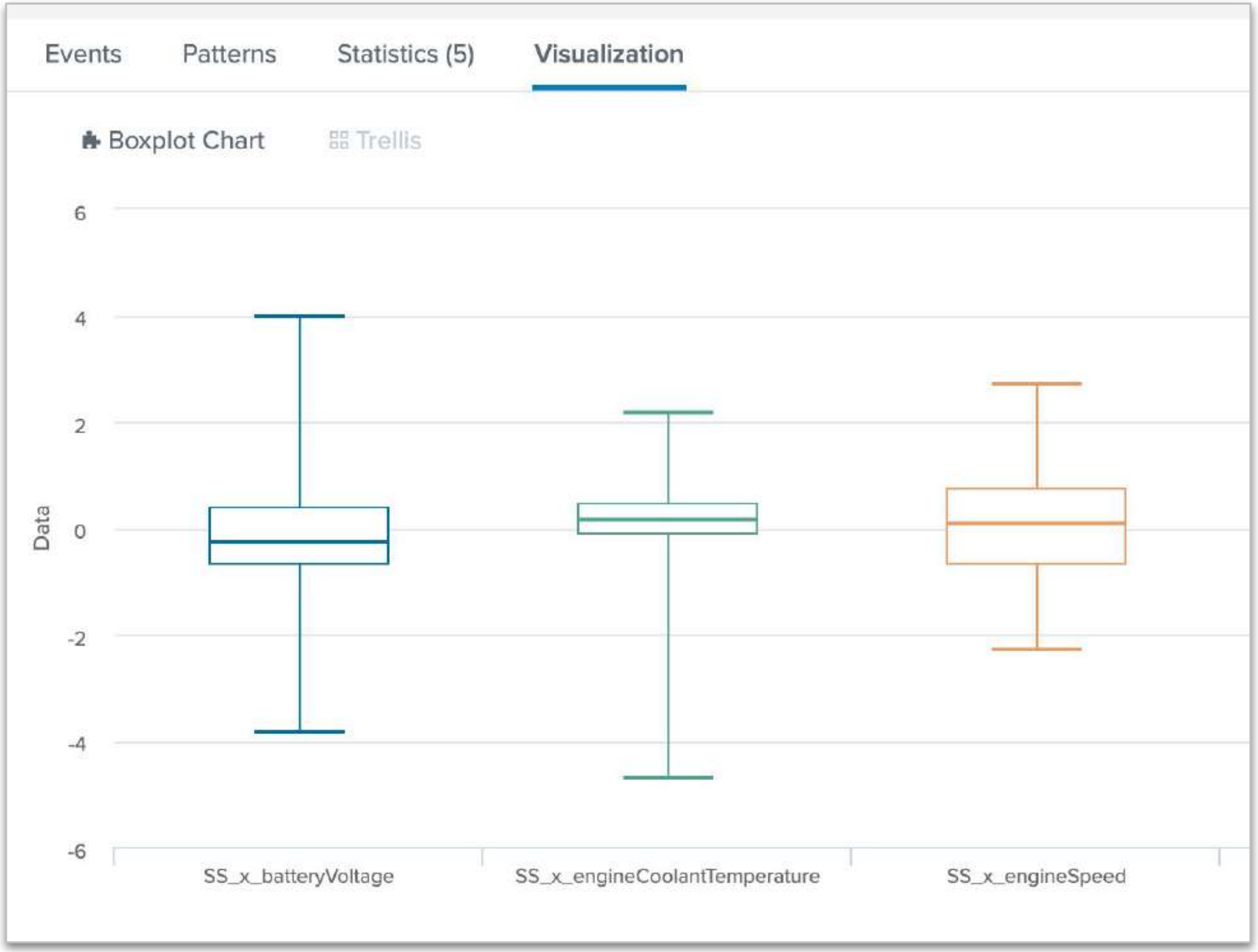
1

Explore the Dataset with Box Plots



```
| inputlookup "mytrackdata.csv"
| fit StandardScaler x_*
| table SS_*
| `boxplot`
```

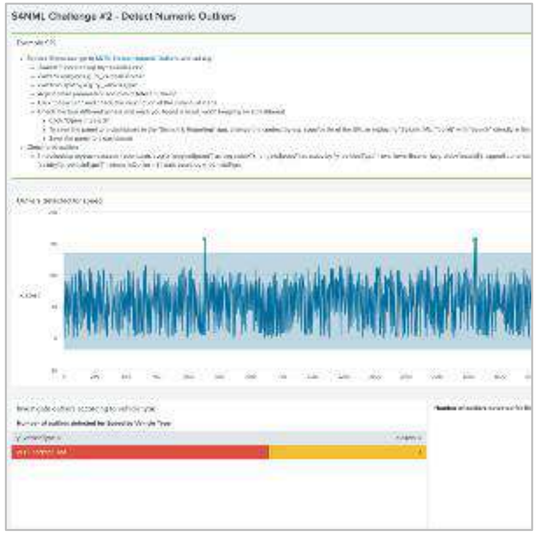
- Standardized data fields have a mean of 0 and a standard deviation of 1
- The box plots are less stretched and can be analyzed more easily



2

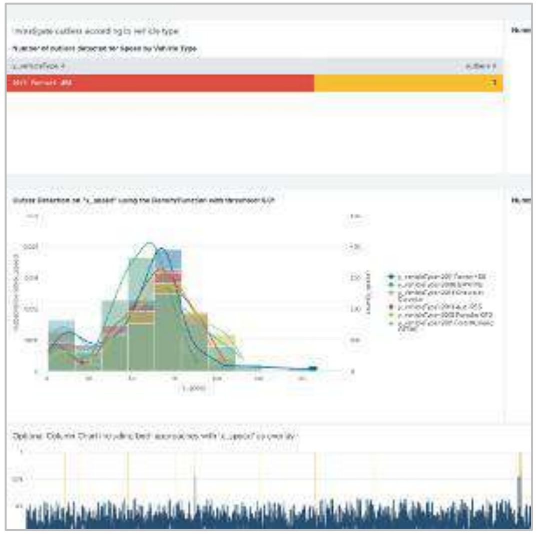
Detect Numeric Outliers:

Explore the MLTK showcase and adapt it to start a new experiment with your own dataset



S4NML Challenge #2 - Detect Numeric Outliers

Explore the Outlier Detection Showcases



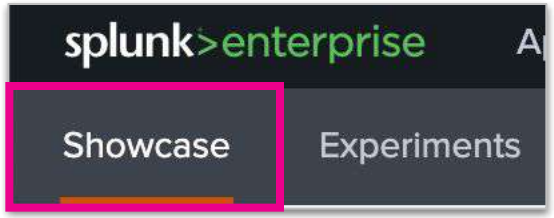
Start your own Outlier Detection Experiment

Optionally try to compare different outlier detection approaches

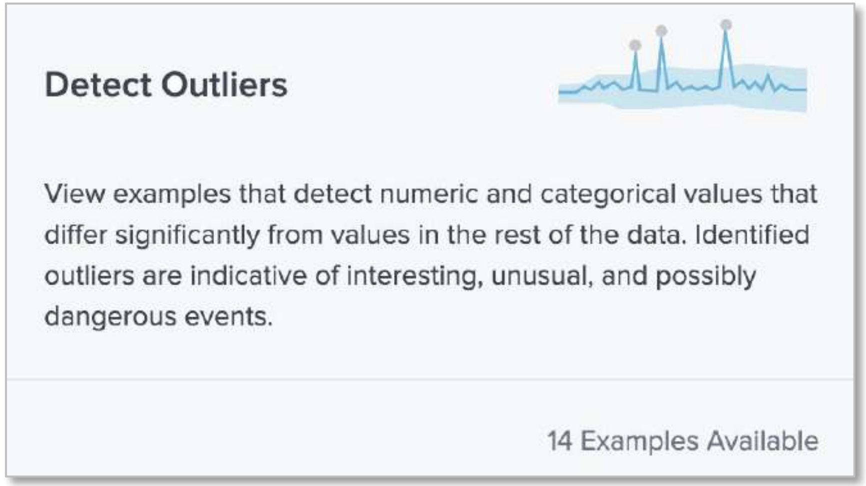
splunk>

2

Explore the Outlier Detection Showcases



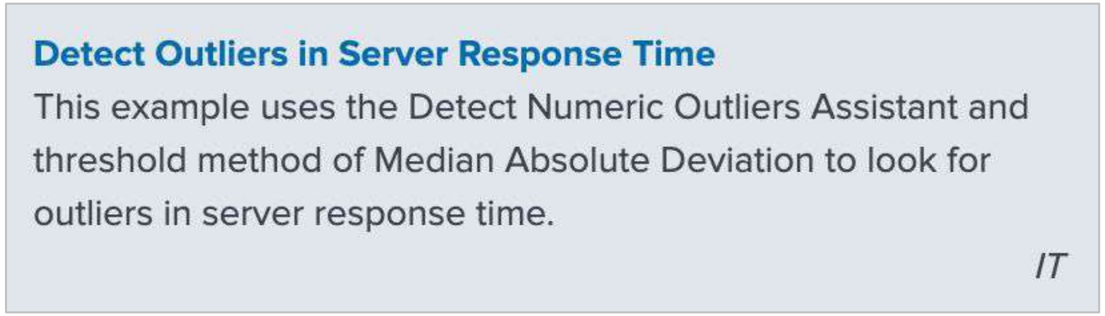
- Switch to the Showcase tab of the MLTK and explore the assistant to detect outliers in server response time
- We are now going to use statistics to detect the outliers



Detect Outliers

View examples that detect numeric and categorical values that differ significantly from values in the rest of the data. Identified outliers are indicative of interesting, unusual, and possibly dangerous events.

14 Examples Available



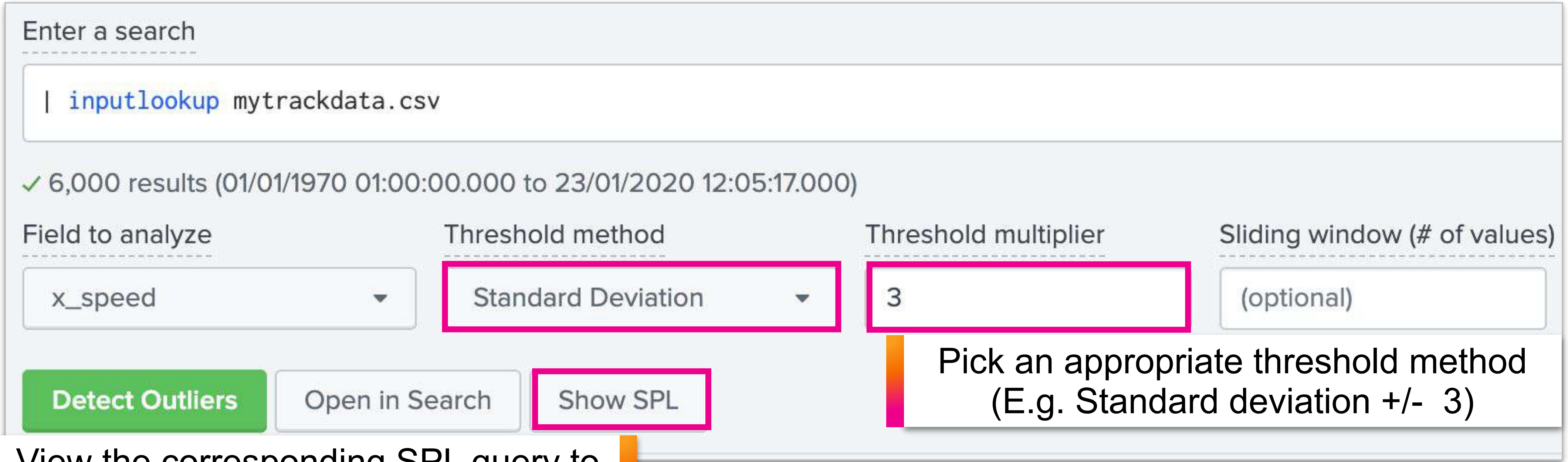
Detect Outliers in Server Response Time

This example uses the Detect Numeric Outliers Assistant and threshold method of Median Absolute Deviation to look for outliers in server response time.

IT

2

Explore the Outlier Detection Showcases



Enter a search

```
| inputlookup mytrackdata.csv
```

✓ 6,000 results (01/01/1970 01:00:00.000 to 23/01/2020 12:05:17.000)

Field to analyze: x_speed

Threshold method: Standard Deviation

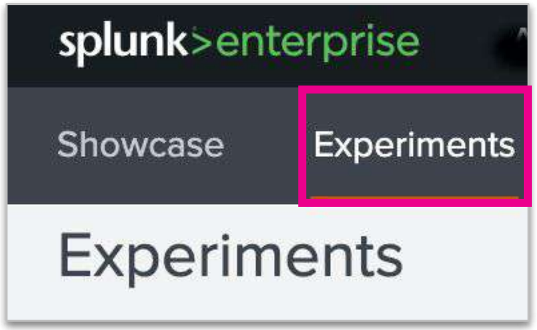
Threshold multiplier: 3

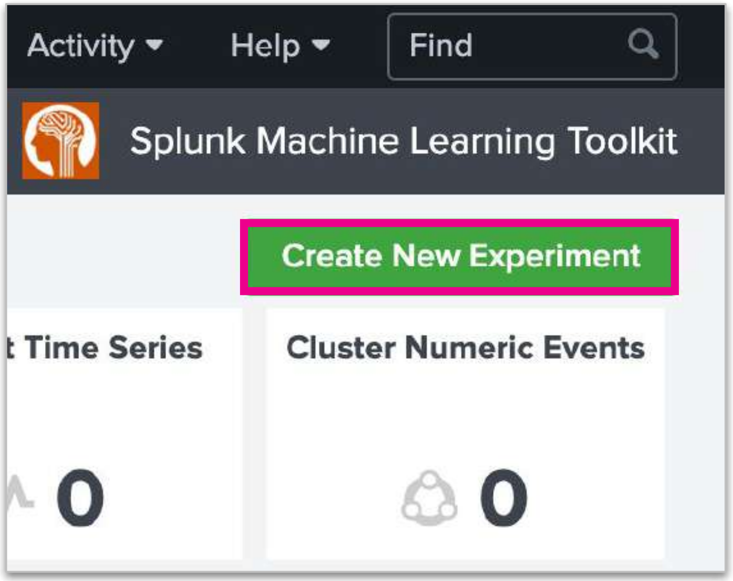
Sliding window (# of values): (optional)

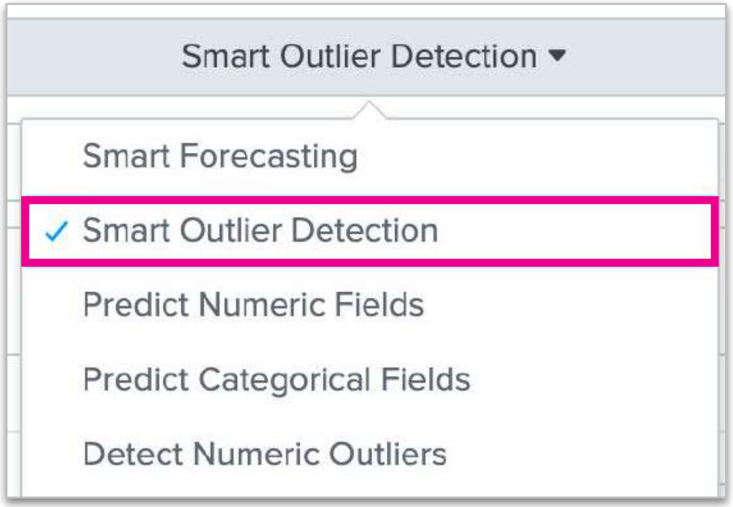
Pick an appropriate threshold method (E.g. Standard deviation +/- 3)

View the corresponding SPL query to the assistant’s settings

2 Detecting Outliers with the Density Function





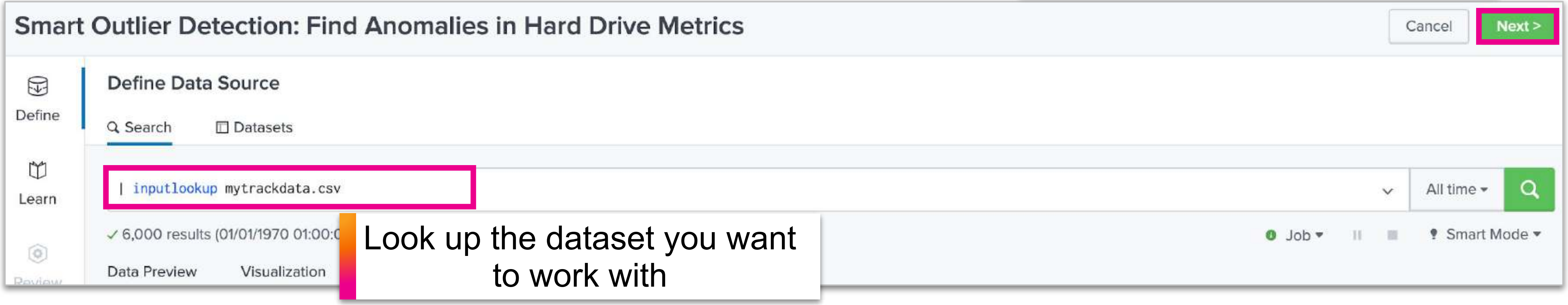


- Switch to the Experiments tab of the MLTK and create a new experiment
- Instead of an approach based on statistics
  we are now going to use the density function to detect outliers

2

Create Your Own Smart Outlier Experiment

Click here to get to the next step



Smart Outlier Detection: Find Anomalies in Hard Drive Metrics

```
| inputlookup mytrackdata.csv
```

✓ 6,000 results (01/01/1970 01:00:0

Look up the dataset you want to work with

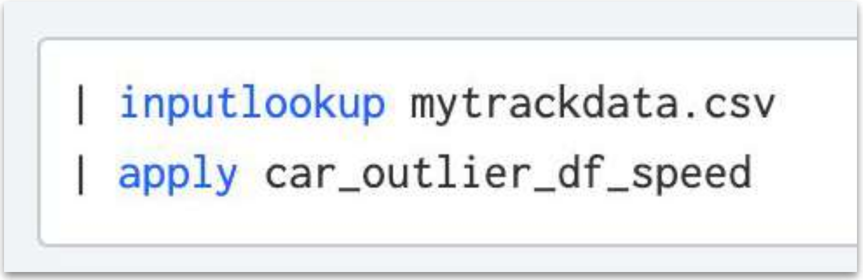
3

SPL for MLTK:
The fit and apply Commands

Examples:

```
<your search> | fit <model name>

<your search> | apply <model name>
```



```
| inputlookup mytrackdata.csv
| apply car_outlier_df_speed
```

- The fit command produces a machine learning model based on the behaviour of a set of events. It applies the model to the current search results in the search pipeline
- The apply command applies the machine learning model that was learned using the fit command

3

SPL for MLTK:
The fit and apply Commands

Examples:

```
<your search> | fit StandardScaler <fields> into <model name>

<your search> | apply <model name> | `<macro name>`

<your search> | fit SVM “X X X" from “XXX" “XXX" kfold_cv=3
```

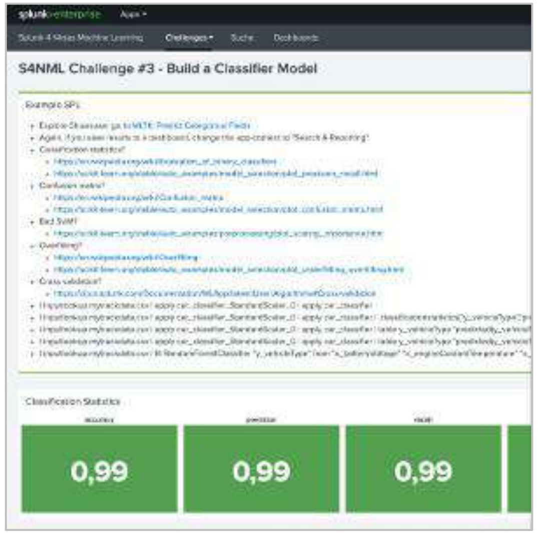
Check out the confusion matrix and classification statistics macros!

- The StandardScaler algorithm uses the scikit-learn StandardScaler algorithm to standardize data fields
- Splunk’s MLTK allows you to cross-validate your models right from the search queries that train them. Simply specify the number of cross-validation folds you want by setting the `fit` command’s parameter `kfold_cv`

3

Use a Classification Model:

Create a classification model and use it to predict vehicle types from your sensor data



S4NML Challenge #3 - Build a Classifier Model

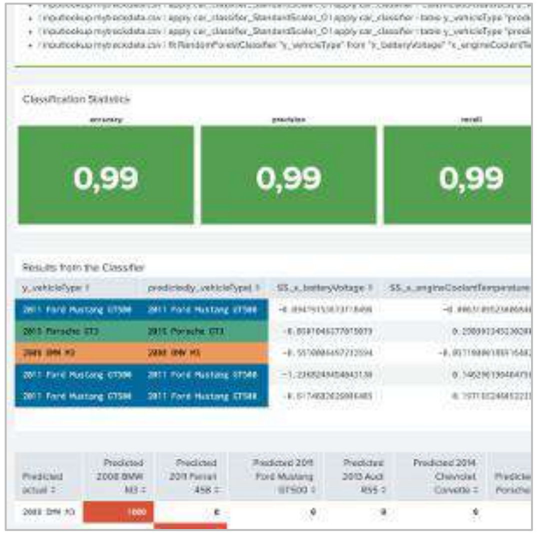
Example SPL

- Explore Showcase: go to MLTK: Predict Categorical Fields
- Again, if you see results in a dashboard, change the app-context to "Search & Reporting"
- Classification statistics?
  - [illegible]
  - [illegible]
- Confusion matrix?
  - [illegible]
  - [illegible]
- Bad SVM?
  - [illegible]
- Overfitting?
  - [illegible]
  - [illegible]
- Cross validation?
  - [illegible]
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | [illegible]
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | table y_vehicleType "predicted(y_vehicleType)"
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | table y_vehicleType "predicted(y_vehicleType)"
- | inputlookup mytrackdata.csv | fit RandomForestClassifier "y_vehicleType" from "x_batteryVoltage" "x_engineCoolantTemperature"

Classification Statistics

accuracy	precision	recall
0,99	0,99	0,99

Explore the Classification Assistant



- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | [illegible]
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | table y_vehicleType "predic
- | inputlookup mytrackdata.csv | apply car_classifier_StandardScaler_01 | apply car_classifier | table y_vehicleType "predic
- | inputlookup mytrackdata.csv | fit RandomForestClassifier "y_vehicleType" from "x_batteryVoltage" "x_engineCoolantTe

Classification Statistics

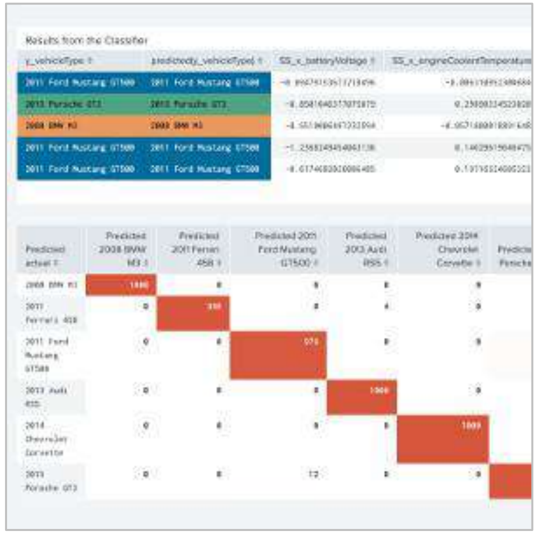
accuracy	precision	recall
0,99	0,99	0,99

Results from the Classifier

y_vehicleType	predicted(y_vehicleType)	SS_x_batteryVoltage	SS_x_engineCoolantTemperature
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]
2013 Porsche GT3	2013 Porsche GT3	[illegible]	[illegible]
2008 BMW M3	2008 BMW M3	[illegible]	[illegible]
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]

Predicted actual	Predicted 2008 BMW M3	Predicted 2011 Ferrari 458	Predicted 2011 Ford Mustang GT500	Predicted 2013 Audi RS5	Predicted 2014 Chevrolet Corvette	Predicted Porsche
2008 BMW M3	1000	0	0	0	0	

Put your Algorithm into Practice



Results from the Classifier

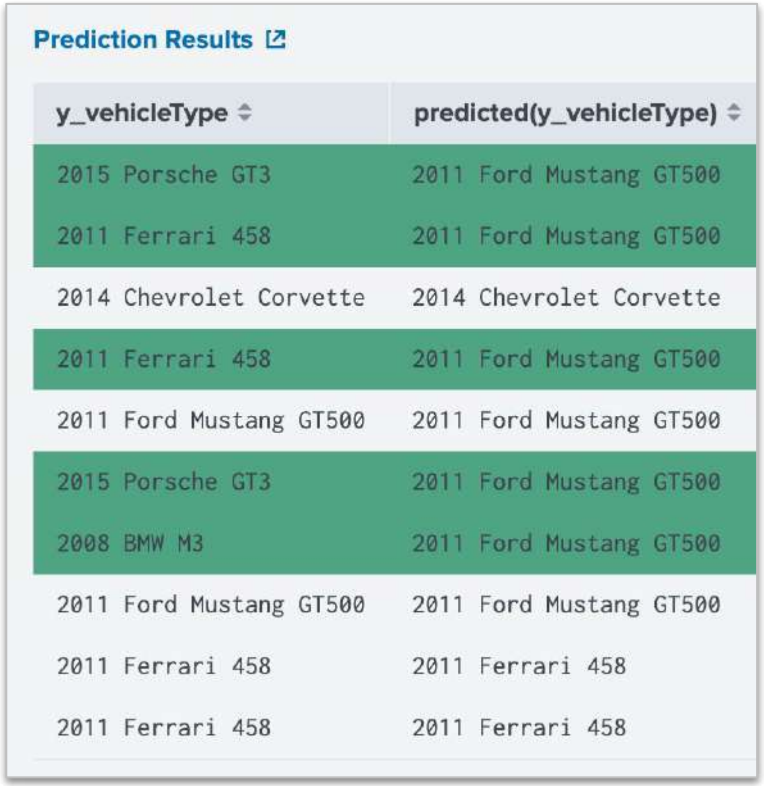
y_vehicleType	predicted(y_vehicleType)	SS_x_batteryVoltage	SS_x_engineCoolantTemperature
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]
2013 Porsche GT3	2013 Porsche GT3	[illegible]	[illegible]
2008 BMW M3	2008 BMW M3	[illegible]	[illegible]
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]
2011 Ford Mustang GT500	2011 Ford Mustang GT500	[illegible]	[illegible]

Predicted actual	Predicted 2008 BMW M3	Predicted 2011 Ferrari 458	Predicted 2011 Ford Mustang GT500	Predicted 2013 Audi RS5	Predicted 2014 Chevrolet Corvette	Predicte Porsche
2008 BMW M3	1000	0	0	0	0	
2011 Ferrari 458	0	[illegible]	0	4	0	
2011 Ford Mustang GT500	0	0	[illegible]	0	0	
2013 Audi RS5	0	0	0	1000	0	
2014 Chevrolet Corvette	0	0	0	0	1000	
2013 Porsche GT3	0	0	12	0	0	

Optionally find a way to deal with model overfitting

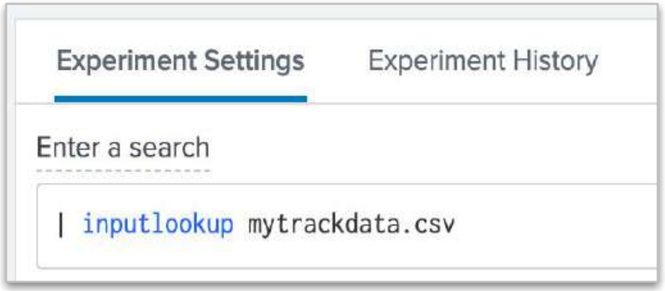
3

Explore the Classification Assistant



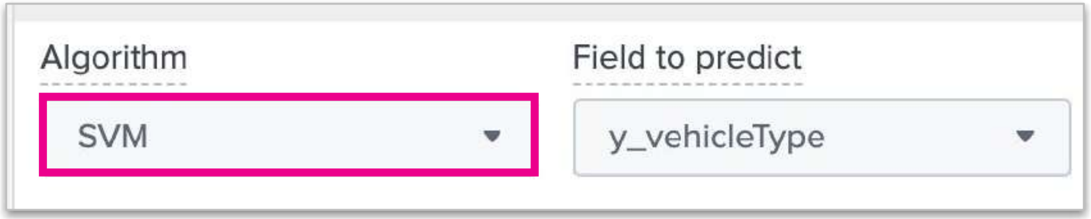
Prediction Results

y_vehicleType	predicted(y_vehicleType)
2015 Porsche GT3	2011 Ford Mustang GT500
2011 Ferrari 458	2011 Ford Mustang GT500
2014 Chevrolet Corvette	2014 Chevrolet Corvette
2011 Ferrari 458	2011 Ford Mustang GT500
2011 Ford Mustang GT500	2011 Ford Mustang GT500
2015 Porsche GT3	2011 Ford Mustang GT500
2008 BMW M3	2011 Ford Mustang GT500
2011 Ford Mustang GT500	2011 Ford Mustang GT500
2011 Ferrari 458	2011 Ferrari 458
2011 Ferrari 458	2011 Ferrari 458



Enter a search

```
| inputlookup mytrackdata.csv
```



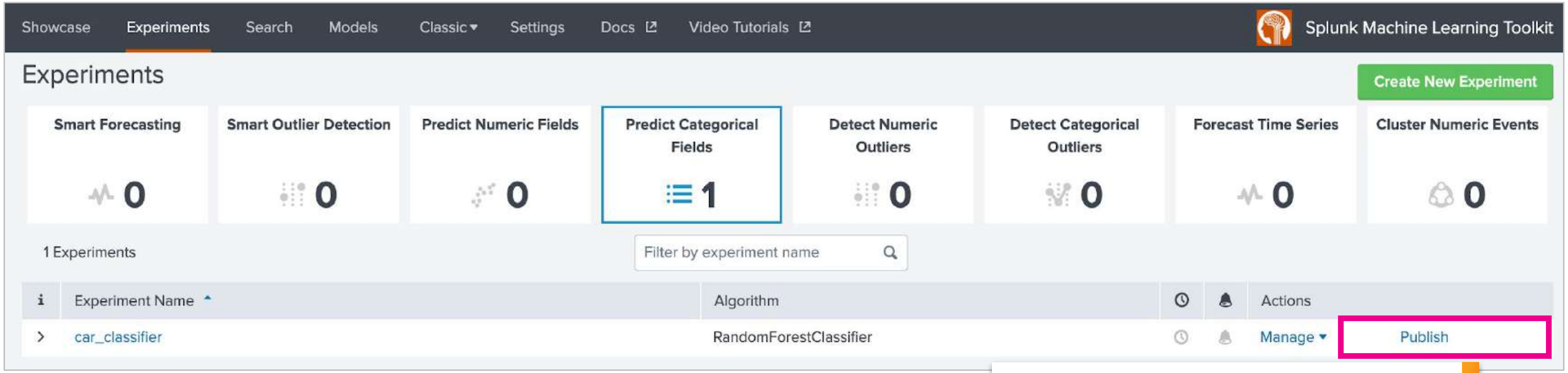
Algorithm: SVM

Field to predict: y_vehicleType

Why is SVM doing so bad?

3

Save your Classification Model

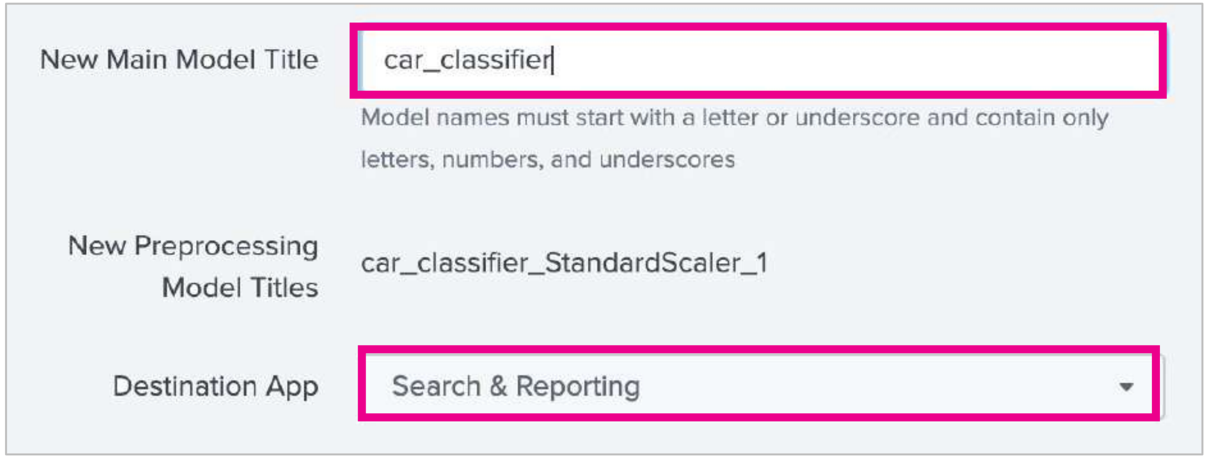


Experiments

1 Experiments

i	Experiment Name ▴	Algorithm			Actions	
›	car_classifier	RandomForestClassifier			Manage ▾	Publish

Publish your model in the app of your choice



New Main Model Title: car_classifier

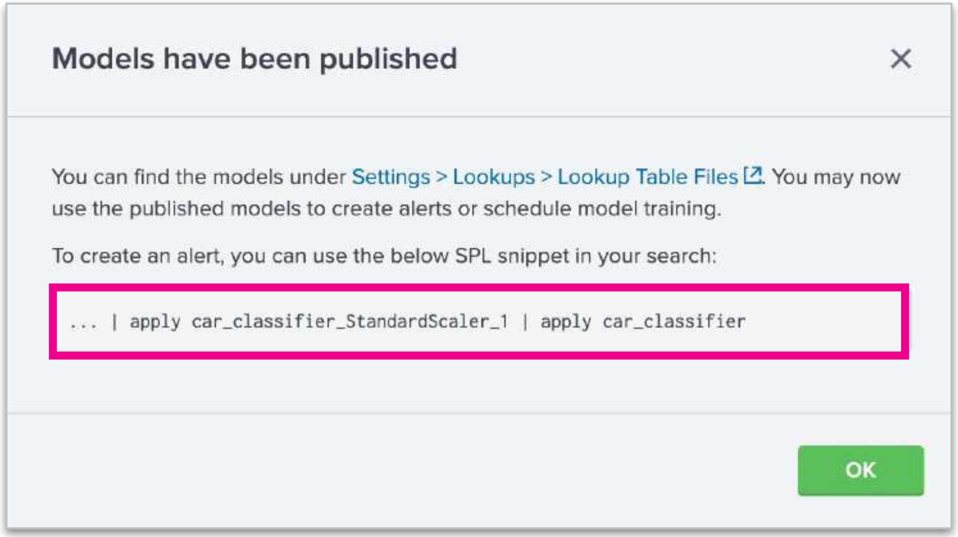
Model names must start with a letter or underscore and contain only letters, numbers, and underscores

New Preprocessing Model Titles: car_classifier_StandardScaler_1

Destination App: Search & Reporting ▾

3

Apply your Classification Model

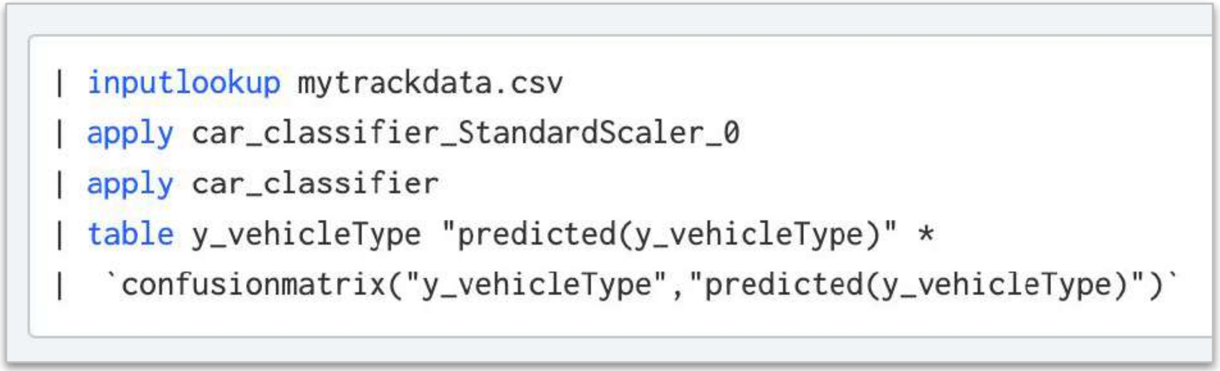


Models have been published

You can find the models under Settings > Lookups > Lookup Table Files. You may now use the published models to create alerts or schedule model training.

To create an alert, you can use the below SPL snippet in your search:

```
... | apply car_classifier_StandardScaler_1 | apply car_classifier
```

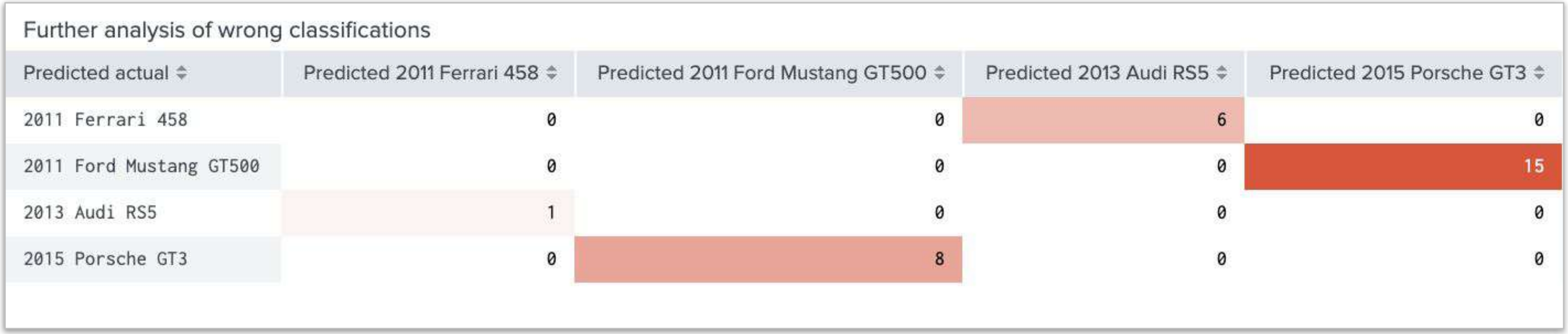


```
| inputlookup mytrackdata.csv
| apply car_classifier_StandardScaler_0
| apply car_classifier
| table y_vehicleType "predicted(y_vehicleType)" *
|  `confusionmatrix("y_vehicleType","predicted(y_vehicleType)")`
```

3

Which Car Gets Classified Worst?

How can you find out where your model is off?



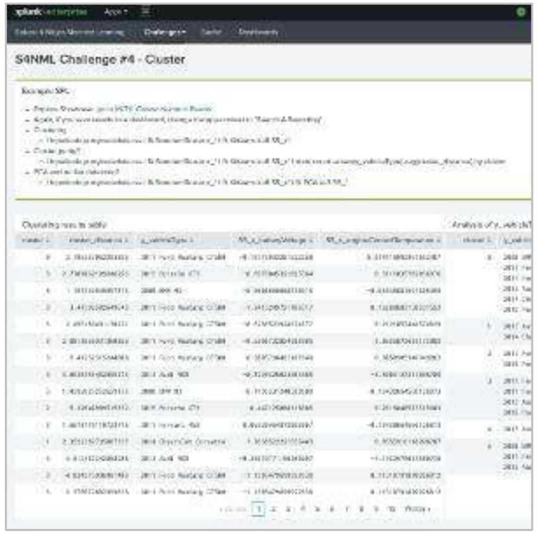
Further analysis of wrong classifications

Predicted actual	Predicted 2011 Ferrari 458	Predicted 2011 Ford Mustang GT500	Predicted 2013 Audi RS5	Predicted 2015 Porsche GT3
2011 Ferrari 458	0	0	6	0
2011 Ford Mustang GT500	0	0	0	15
2013 Audi RS5	1	0	0	0
2015 Porsche GT3	0	8	0	0

3

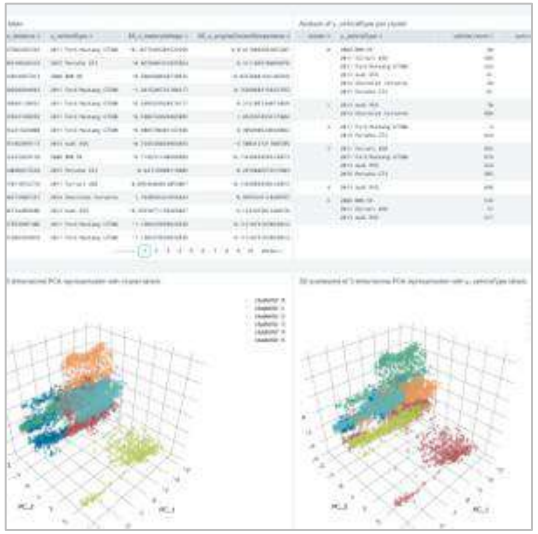
Use a Clustering Model:

Create a clustering model and use it to analyze your dataset

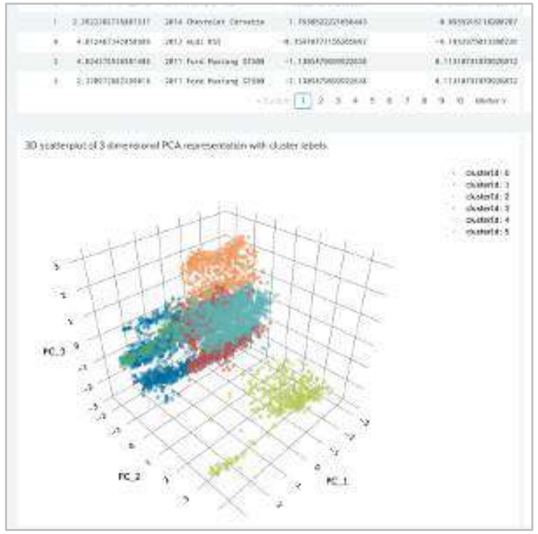


S4NML Challenge #4 - Cluster

Explore the Clustering Assistant



Cluster Analysis of the mytrackdata-Dataset

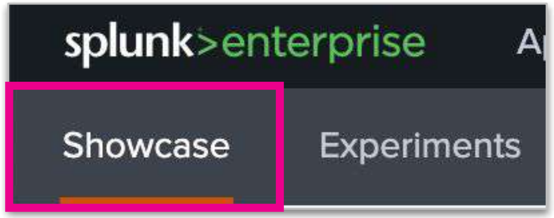


3D scatterplot of 3 dimensional PCA representation with cluster labels

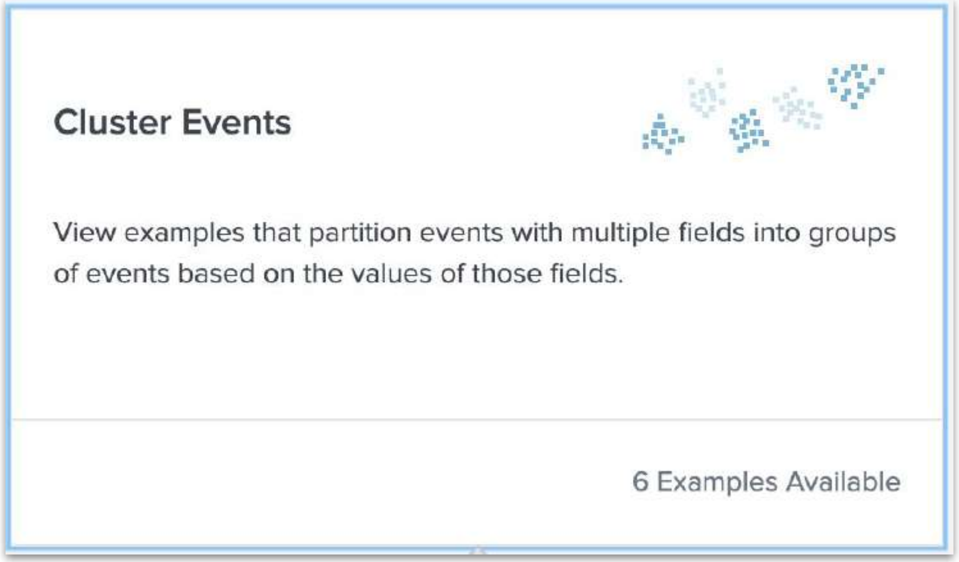
Optionally try and detect outliers

4

Explore the Cluster Showcases



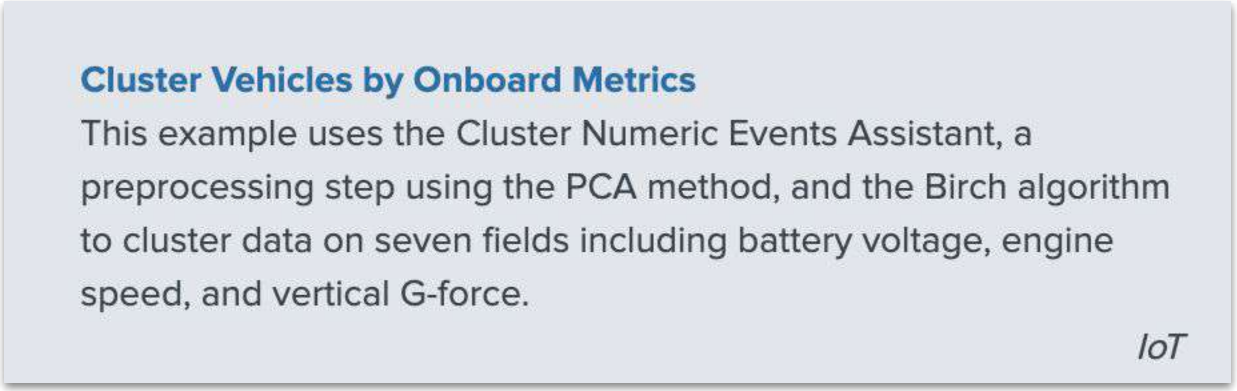
- Switch to the Showcase tab of the MLTK and explore the assistant to identify clusters of events



Cluster Events

View examples that partition events with multiple fields into groups of events based on the values of those fields.

6 Examples Available



Cluster Vehicles by Onboard Metrics

This example uses the Cluster Numeric Events Assistant, a preprocessing step using the PCA method, and the Birch algorithm to cluster data on seven fields including battery voltage, engine speed, and vertical G-force.

IoT

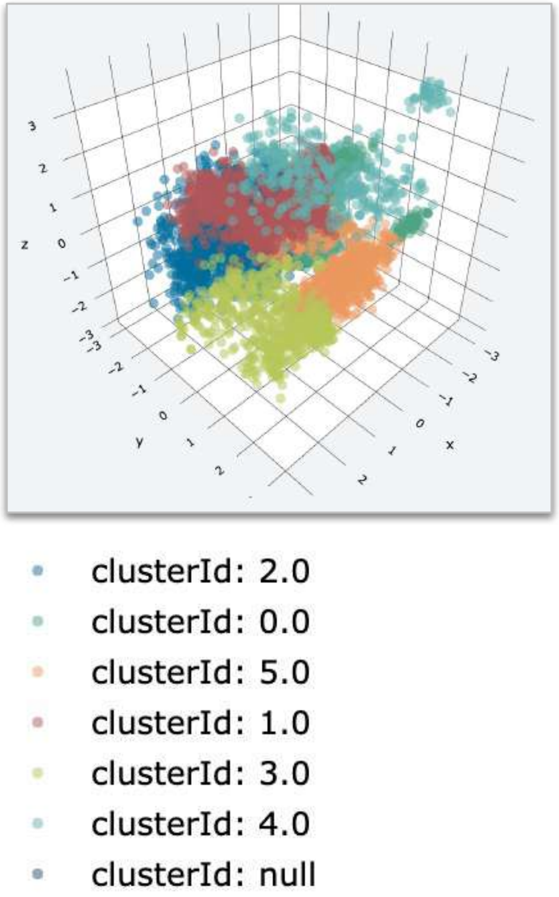
4

The MLTK Comes with Many Different Algorithms

Example:

```
<your search> | fit PCA k=<int> <fields>
```

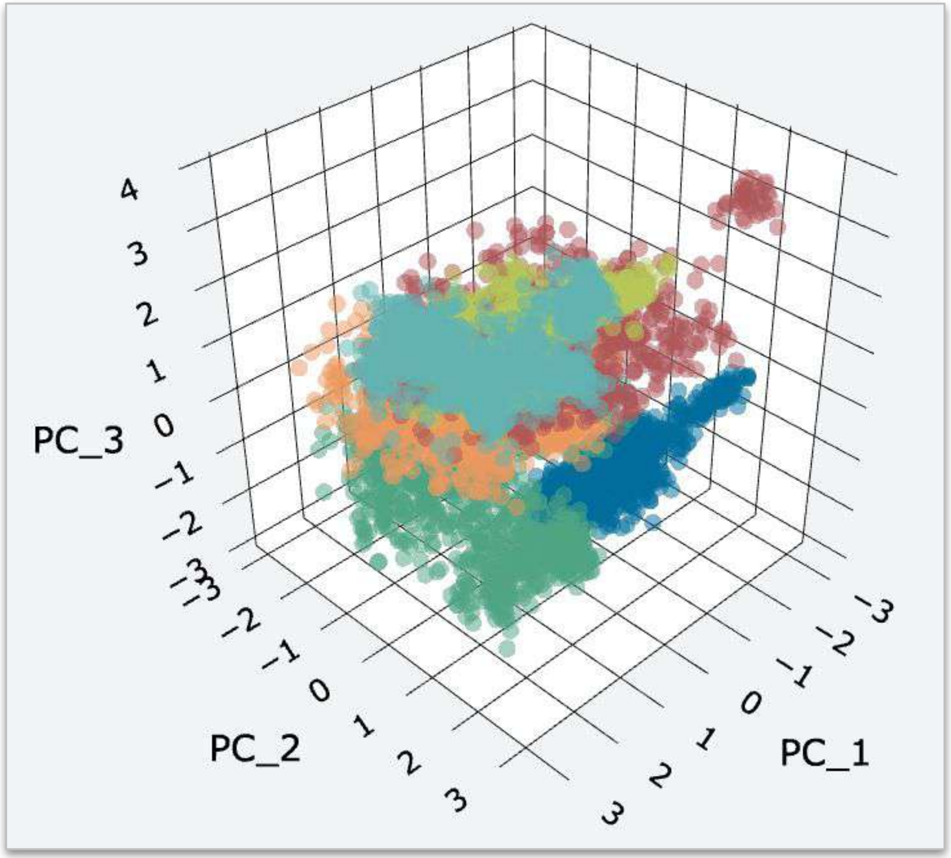
- Factor analysis with an algorithm such as PCA can reduce the number of variables one must deal with
- The k parameter specifies the number of features to be extracted from the data



Why is there a cluster with "clusterId: null" ?

4

The MLTK Comes with Many Different Algorithms



We have missing values in "x_engineCoolantTemperature" that we didn't fix/impute in mytrackdata.csv

splunk>

Wrap Up

Wrap Up

- **Feedback:** How was your experience, what worked well, what did not?

- **Discussion Brainstorming:** How could you transfer the topics learned today to other use cases or departments?

- **You want to learn more about Splunk's Machine Learning?**
  - ▸ Check our latest Splunk Blogs around Machine Learning
  - ▸ Watch videos from Splunk Machine Learning YouTube Channel
  - ▸ Take the Splunk Education Class for Data Science and Advanced Analytics
  - ▸ Learn more about Splunk's Machine Learning Advisory Program

Thank You!

splunk>

Additional Information

Login:

Username: **admin**

Password: **<See Splunk Show>**

Challenge Solution Examples:

We created a dashboard for each challenge with example solutions in the hidden app “Splunk 4 Ninjas Machine Learning”. Use this app for preparation, debriefing after the challenges or as assistance for unexperienced attendees.

- https://{your-host}/en-GB/app/s4n_ml/splunk_4_ninjas_ml
  or click the button next to “Splunk 4 Ninjas Machine Learning” on top of the Home dashboard