Using the Latest Features from the Splunk Machine Learning Toolkit to Create Your Own Custom Models

Adam J. Oliner | Director of Engineering

Harsh Keswani | Product Manager - Machine Learning

October 2018

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.

Speakers

Adam J. Oliner

Director of Engineering

Harsh Keswani

Product Manager: Machine Learning

Outline

- Splunk Machine Learning Toolkit
- Platform Extensions: ML-SPL, etc.
- Experiments: Guided Machine Learning
- Demo
- What’s New
- Customer Success

Splunk Machine Learning Toolkit

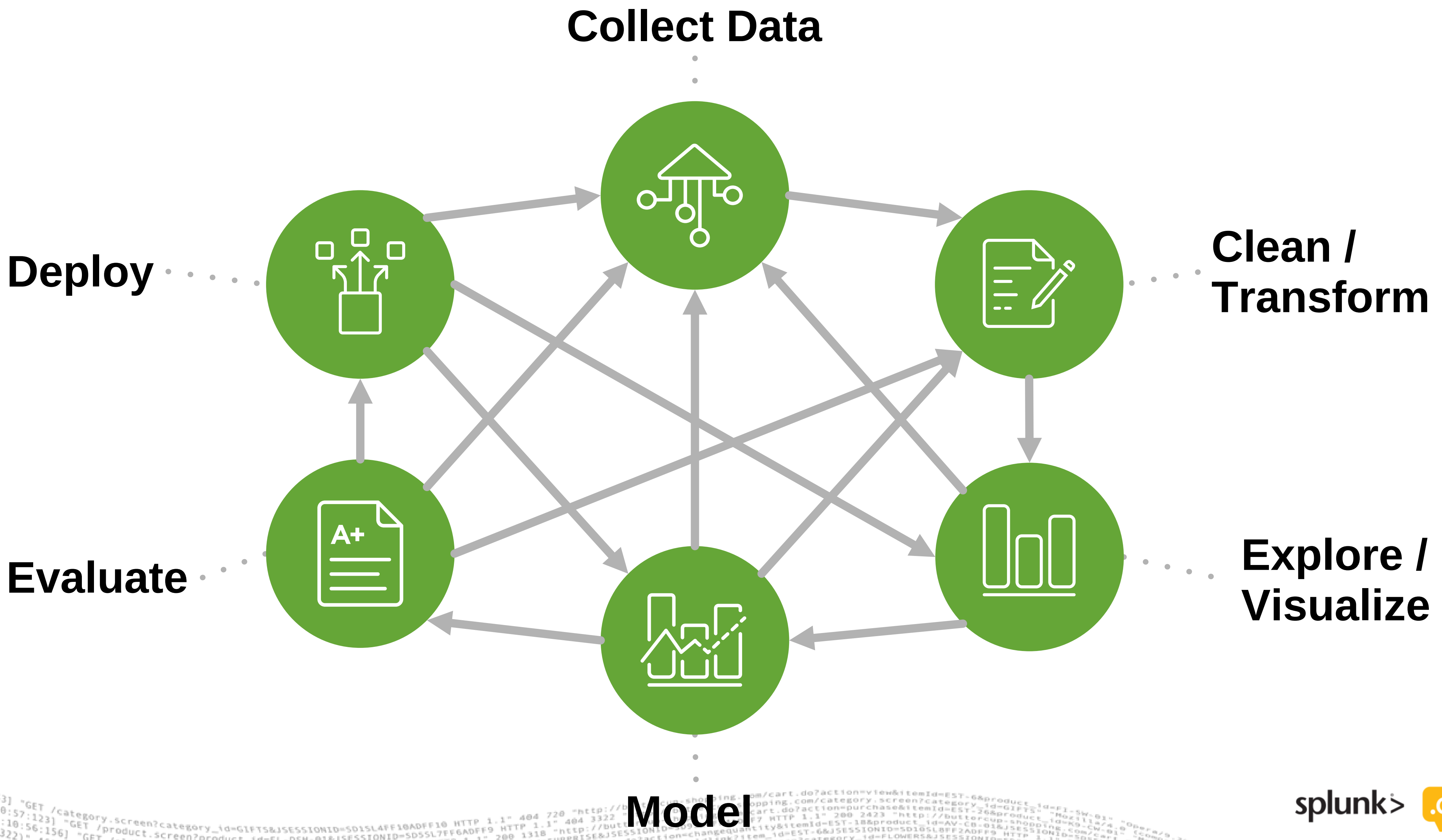
platform extensions and guided modeling dashboards

Machine Learning

- A process for generalizing from examples
- Examples

A, B, … → #	(regression)
A, B, ... → a	(classification)
$X_{past} \rightarrow X_{future}$	(forecasting)
like with like	(clustering)
$\lvert X_{predicted} - X_{actual} \rvert >> 0$	(anomaly detection)

Machine Learning Process



Collect Data

Clean / Transform

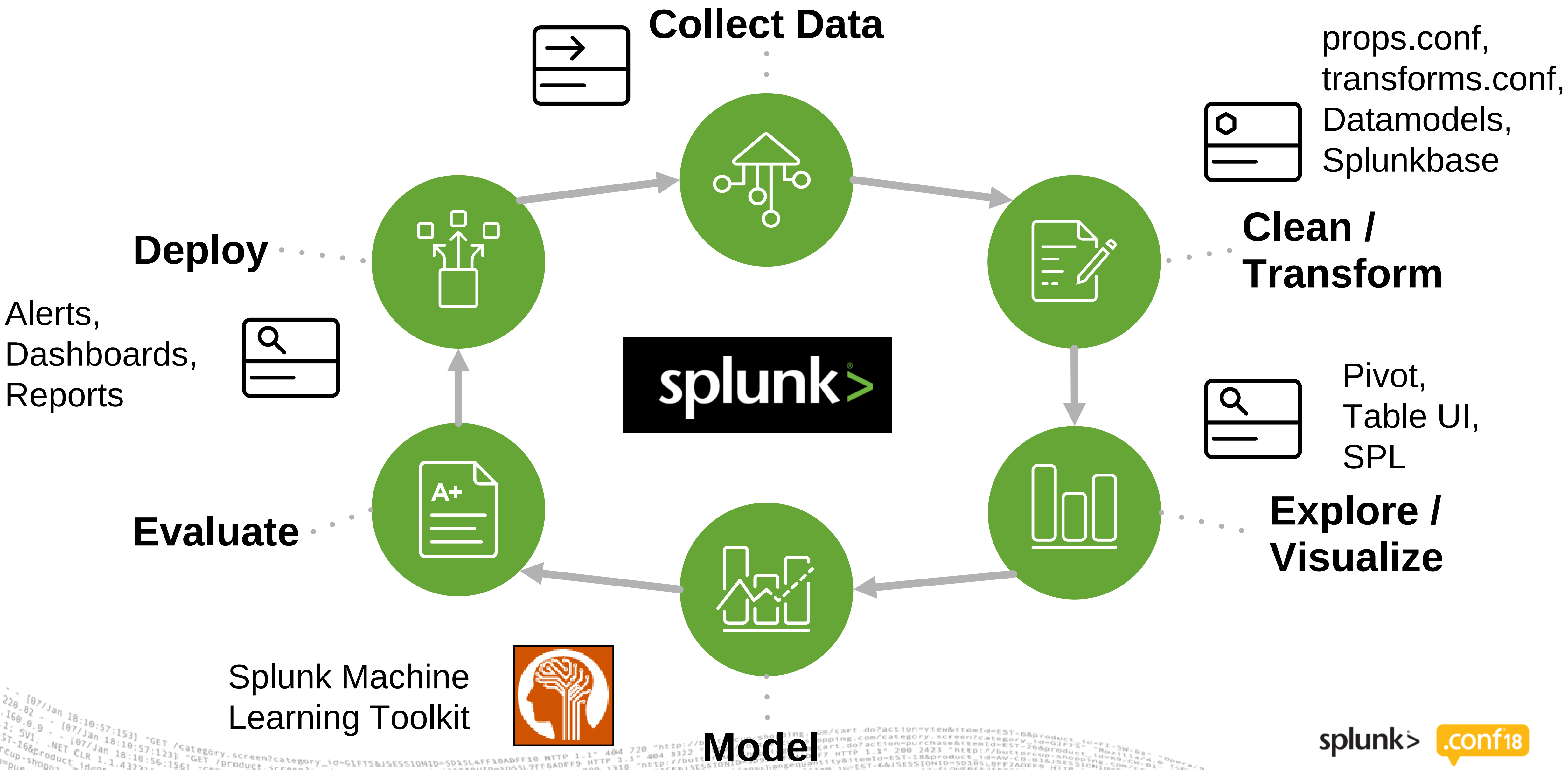
Explore / Visualize

Model

Evaluate

Deploy

Machine Learning Process with Splunk



Collect Data

Clean / Transform

props.conf, transforms.conf, Datamodels, Splunkbase

Explore / Visualize

Pivot, Table UI, SPL

Model

Splunk Machine Learning Toolkit

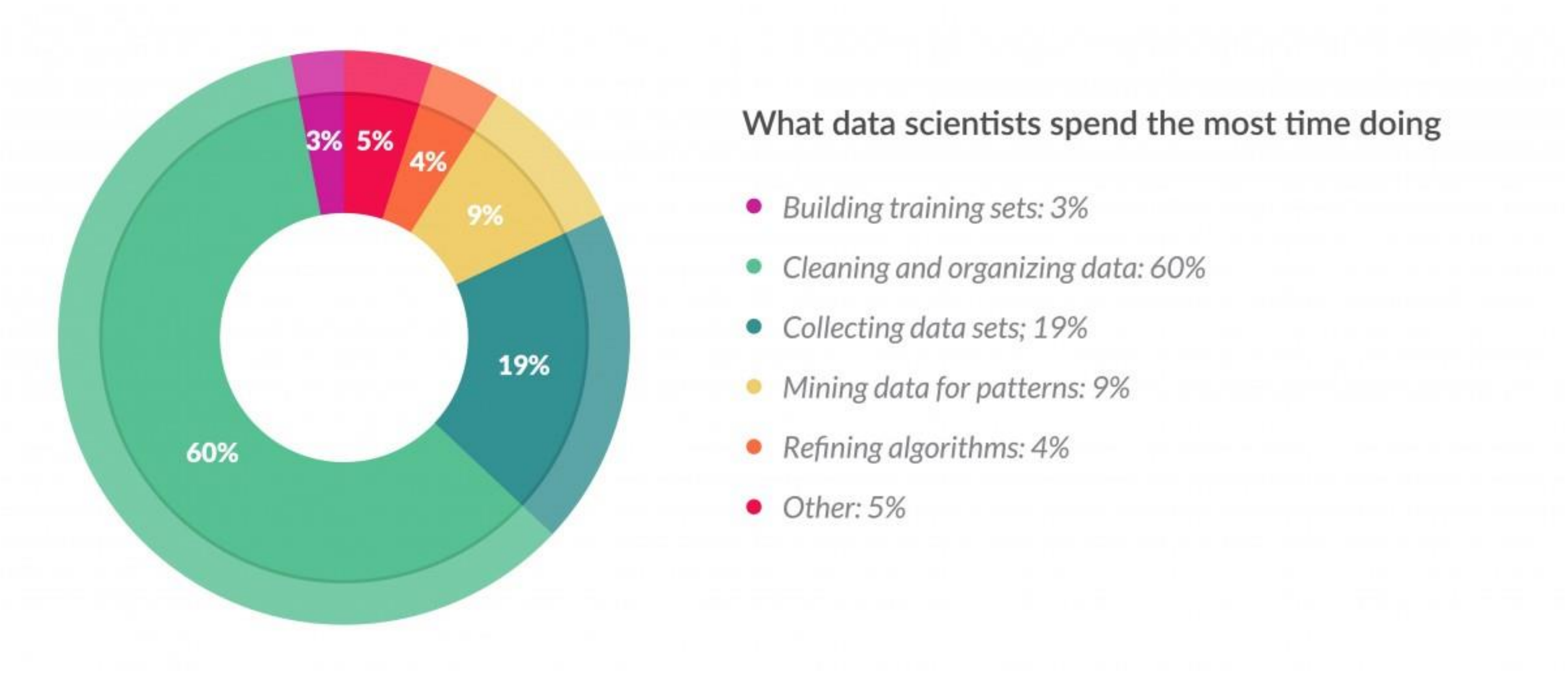
Evaluate

Deploy

Alerts, Dashboards, Reports

Data Gathering and Prep

Source: CrowdFlower



What data scientists spend the most time doing

- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

Want to learn more about data prep? Download the slides and recording for the following session.

Getting Your Data Ready for Machine Learning

Speakers

Kristal Curtis, Software Engineer, Machine Learning, Splunk
Adam J. Oliner, Director of Engineering, Splunk

Wednesday, Oct 03, 12:45 p.m. - 1:30 p.m.

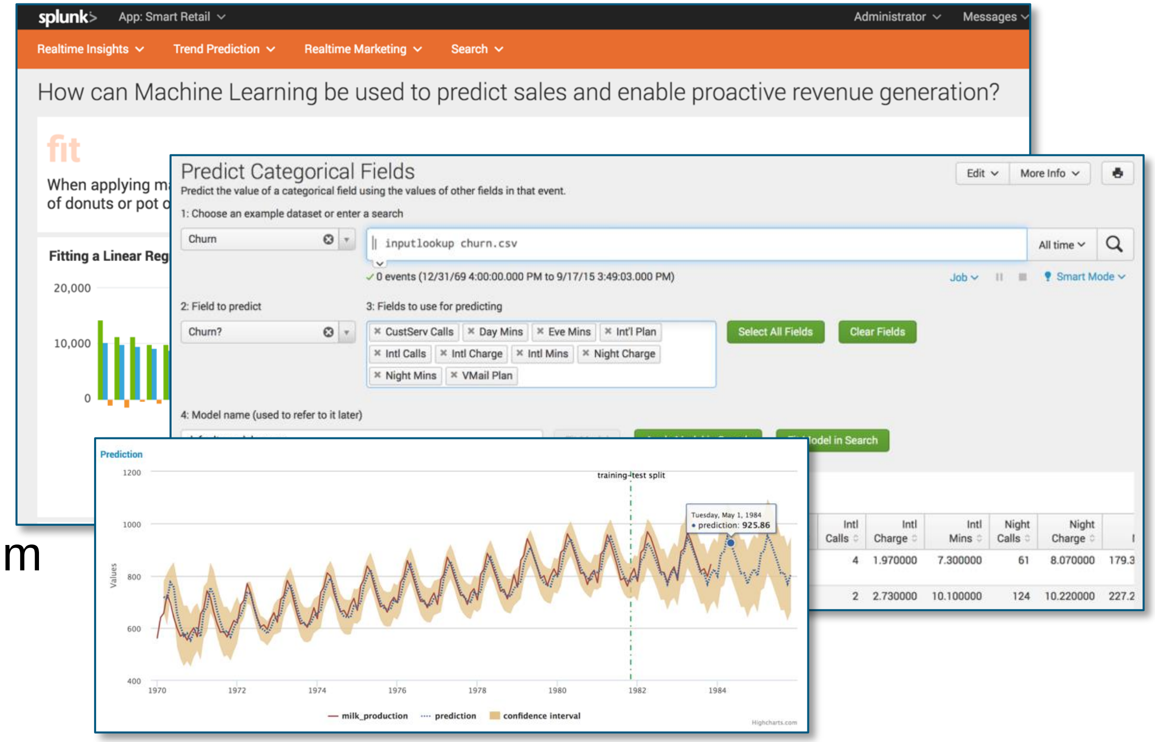
Splunk Machine Learning Toolkit

extends Splunk with new tools and guided modeling

- **Experiments:** Guided model building, testing, and deployment for common objectives
- **Showcases:** Interactive examples for typical IT, security, business, and IoT use cases
- **Algorithms:** 30 standard algorithms

  (supervised & unsupervised)
- **ML Commands:** New SPL commands to fit, test and operationalize models
- **ML-SPL API** Extensibility to easily import any algorithm

  (proprietary / open source)
- **Python for Scientific Computing Library:** Access to 300+ open source algorithms



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

When applying m
of donuts or pot c

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

Churn

| 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

Churn?

3: Fields to use for predicting

× CustServ Calls × Day Mins × Eve Mins × Int'l Plan × Intl Calls × Intl Charge × Intl Mins × Night Charge × Night Mins × VMail Plan

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

Tuesday, May 1, 1984
• prediction: 925.86

Intl Calls	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

Platform Extensions

custom search commands for machine learning

SPL, Macros, & Viz

Oh, my!

- **Commands (ML-SPL)**
  - fit
  - apply
  - summary
  - listmodels
  - deletemodel
  - sample
  - score
- **Macros**
  - regressionstatistics
  - classificationstatistics
  - classificationreport
  - confusionmatrix
  - forecastviz
  - histogram
  - modvizpredict
  - splitby(1-5)
- **Viz**
  - Outliers Chart
  - Forecast Chart
  - Scatter Line Chart
  - Histogram Chart
  - Downsampled Line Chart
  - Scatterplot Matrix
  - Box Plot Chart

ML-SPL Commands

- Fit (i.e., train) a model from search results

```
… | fit <ALGORITHM> <TARGET> from <VARIABLES …>
        <PARAMETERS> into <MODEL>
```

- Apply a model to obtain predictions from (new) search results

```
… | apply <MODEL>
```

- Inspect a model (e.g., display coefficients)

```
  | summary <MODEL>
```

- Score the prediction results

```
… | score <SCORE_METHOD> <ACTUAL> ~ <PREDICTED>
```

ML-SPL Commands: fit

```
… | fit <ALGORITHM> <TARGET> from <VARIABLES>
        <PARAMETERS> into <MODEL>
```

optional

Examples:

```
… | fit LinearRegression
        system_temp from cpu_load fan_rpm
        into temp_model

… | fit KMeans k=10
        downloads purchases posts days_active visits_per_day
        into user_behavior_clusters
```

ML-SPL Algorithms

- 30 algorithms OotB
  - prediction, clustering, forecasting, feature engineering
- Extensibility API for 300+ more
- Pipeline for advanced use cases

```
…       | fit TFIDF message
        | fit StandardScaler files bytes
        | fit KMeans message_tfidf_* SS_* k=5
        | fit PCA message_tfidf_* k=2
        | …
```

ML-SPL Commands: apply

```
… | apply <MODEL>
```

Examples:

```
… | apply temp_model
… | apply user_behavior_clusters
```

ML-SPL Commands: score

```
… | score <SCORE_METHOD> <ACTUAL> ~ <PREDICTED>
```

Examples:

```
    … | score accuracy_score vehicleType ~ LR_prediction DT_prediction
    … | score confusion_matrix actual=vehicleType predicted=pred_type
```

ML-SPL Commands: summary

```
… | summary <MODEL>
```

Examples:

```
… | summary temp_model
… | summary user_behavior_clusters
```

ML-SPL Commands

```
| listmodels
| deletemodel <MODEL>
```

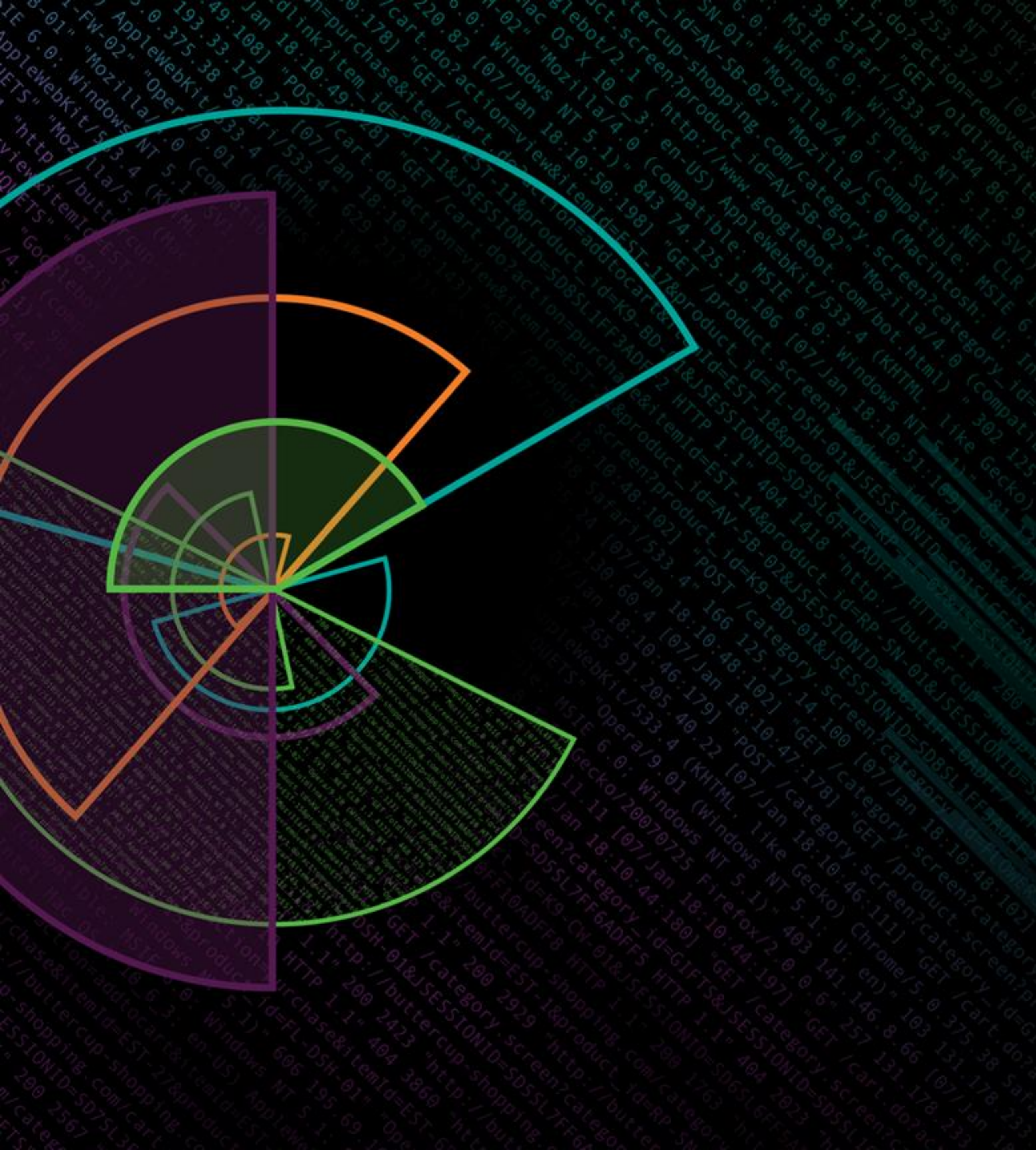
ML-SPL Commands: `sample`

- Randomly sample or partition events

```
… | sample <PARAMETERS>
```

- Four modes

Ratio	`… | sample 0.01`
Count	`… | sample 20`
Proportional	`… | sample proportional=“some_field”`
Partition	`… | sample partitions=10`

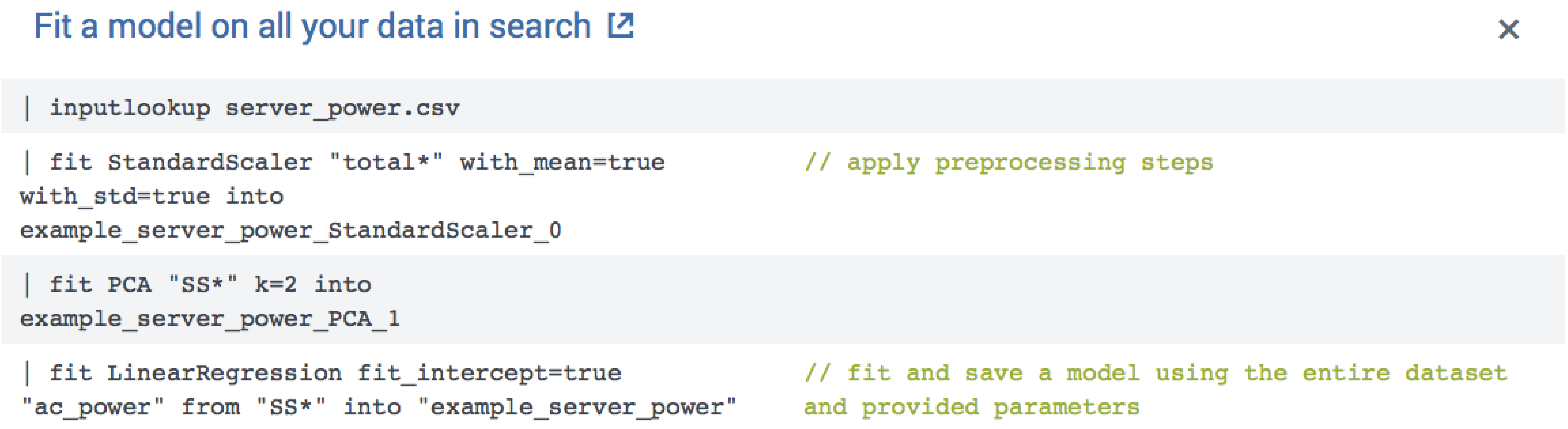


Experiments

guided machine learning

Guided ML with Experiments

- Guides you through an analysis
- Automatically generates all the relevant SPL



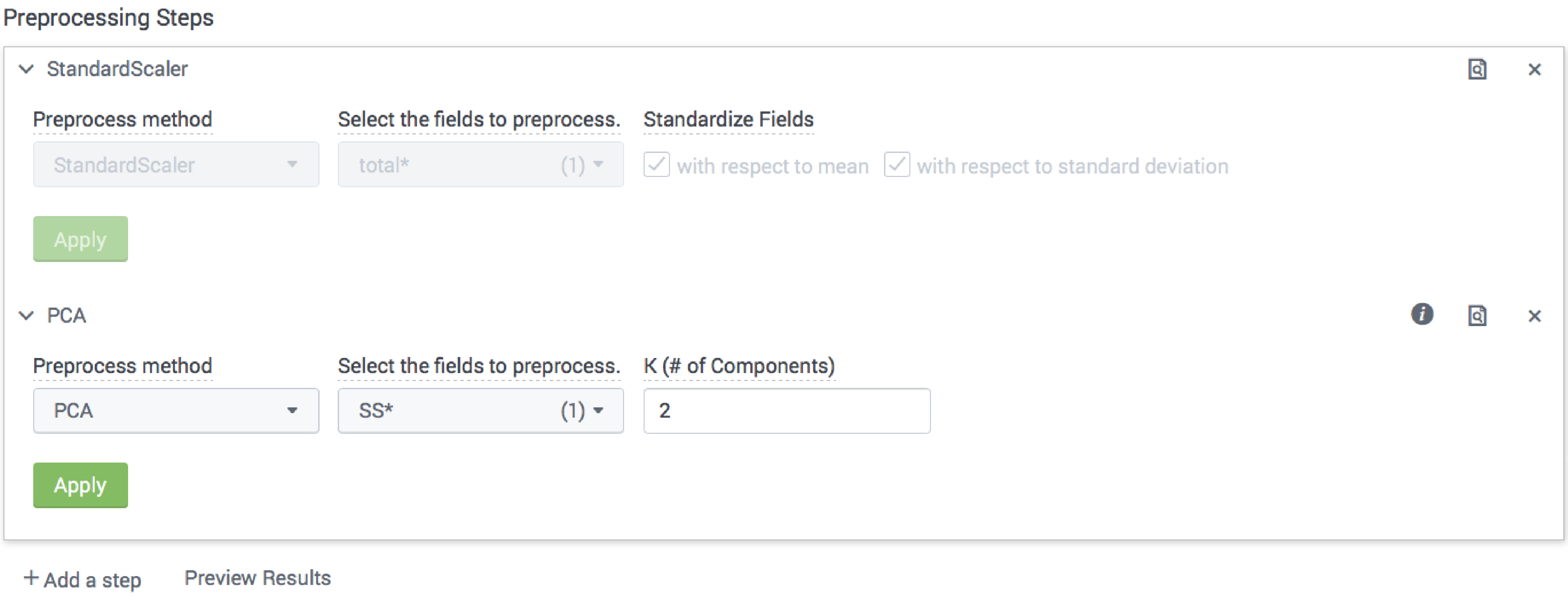
```
| inputlookup server_power.csv

| fit StandardScaler "total*" with_mean=true        // apply preprocessing steps
with_std=true into
example_server_power_StandardScaler_0

| fit PCA "SS*" k=2 into
example_server_power_PCA_1

| fit LinearRegression fit_intercept=true           // fit and save a model using the entire dataset
"ac_power" from "SS*" into "example_server_power"   and provided parameters
```

Experiments: Prepare



Preprocessing Steps

StandardScaler

Preprocess method: StandardScaler

Select the fields to preprocess.: total* (1)

Standardize Fields: ☑ with respect to mean ☑ with respect to standard deviation

Apply

PCA

Preprocess method: PCA

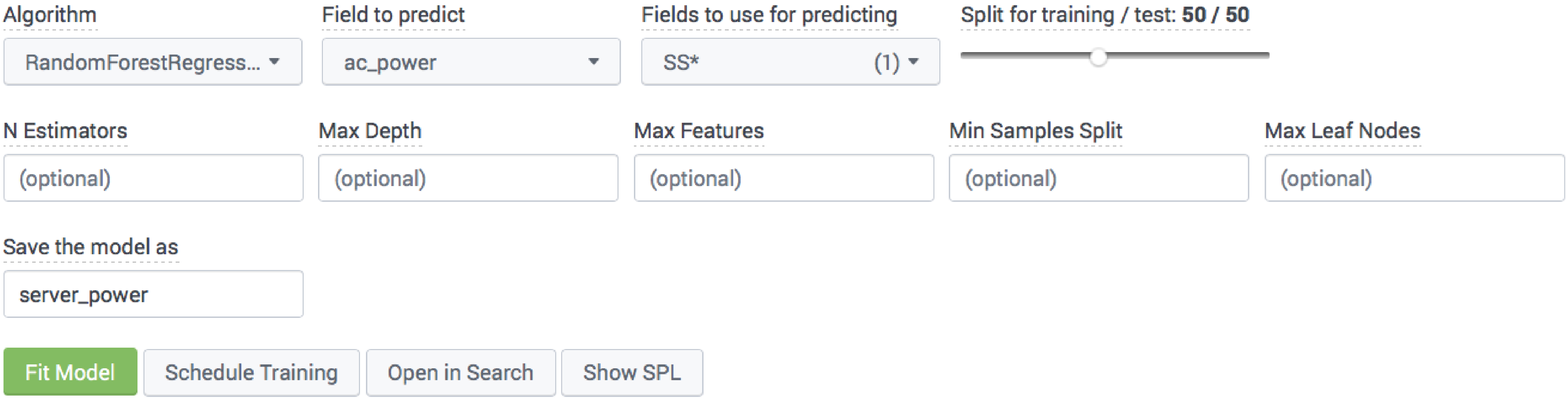
Select the fields to preprocess.: SS* (1)

K (# of Components): 2

Apply

+ Add a step   Preview Results

Experiments: Fit



Algorithm: RandomForestRegress...

Field to predict: ac_power

Fields to use for predicting: SS* (1)

Split for training / test: **50 / 50**

N Estimators: (optional)

Max Depth: (optional)

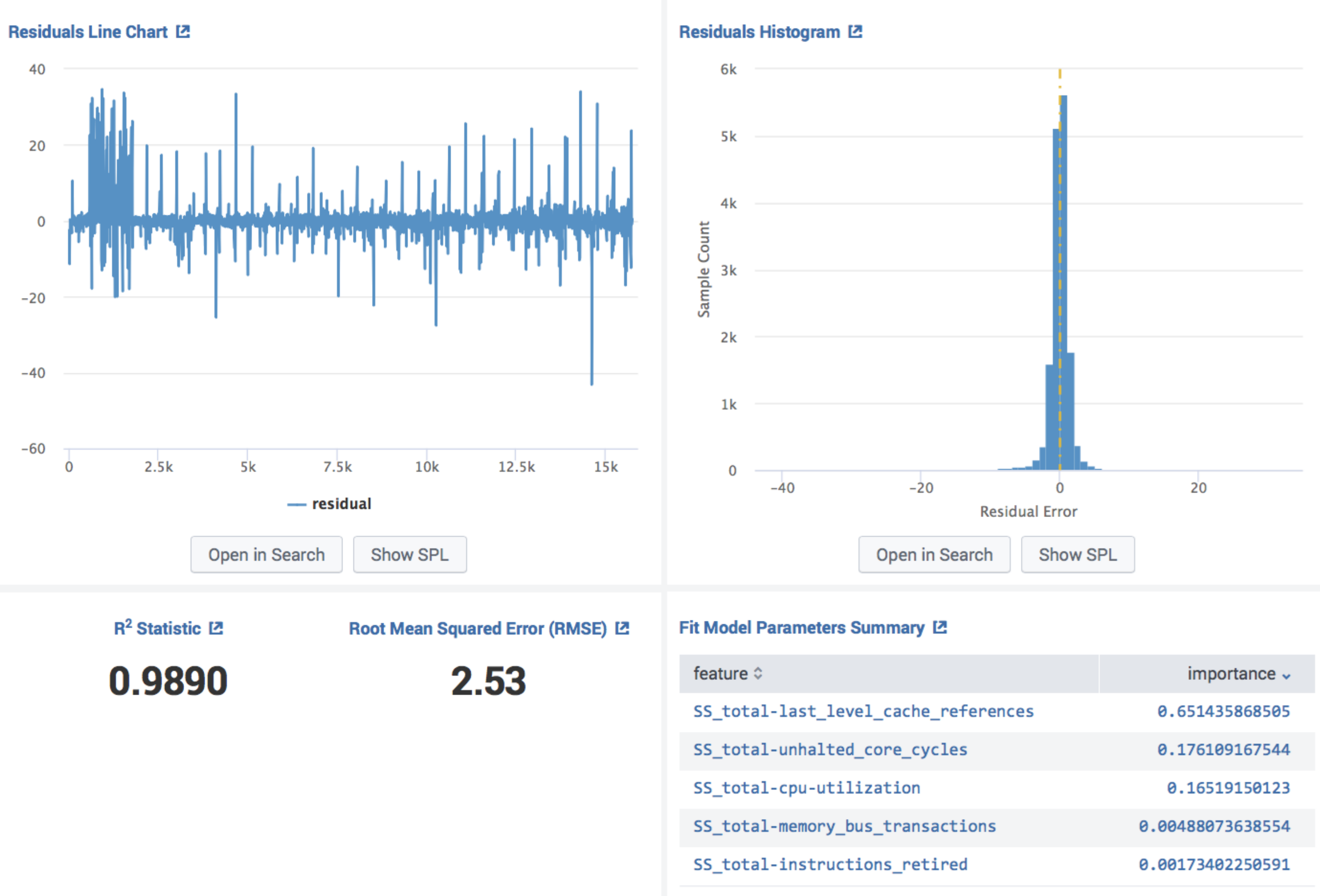
Max Features: (optional)

Min Samples Split: (optional)

Max Leaf Nodes: (optional)

Save the model as: server_power

Experiments: Validate



R^2 Statistic

0.9890

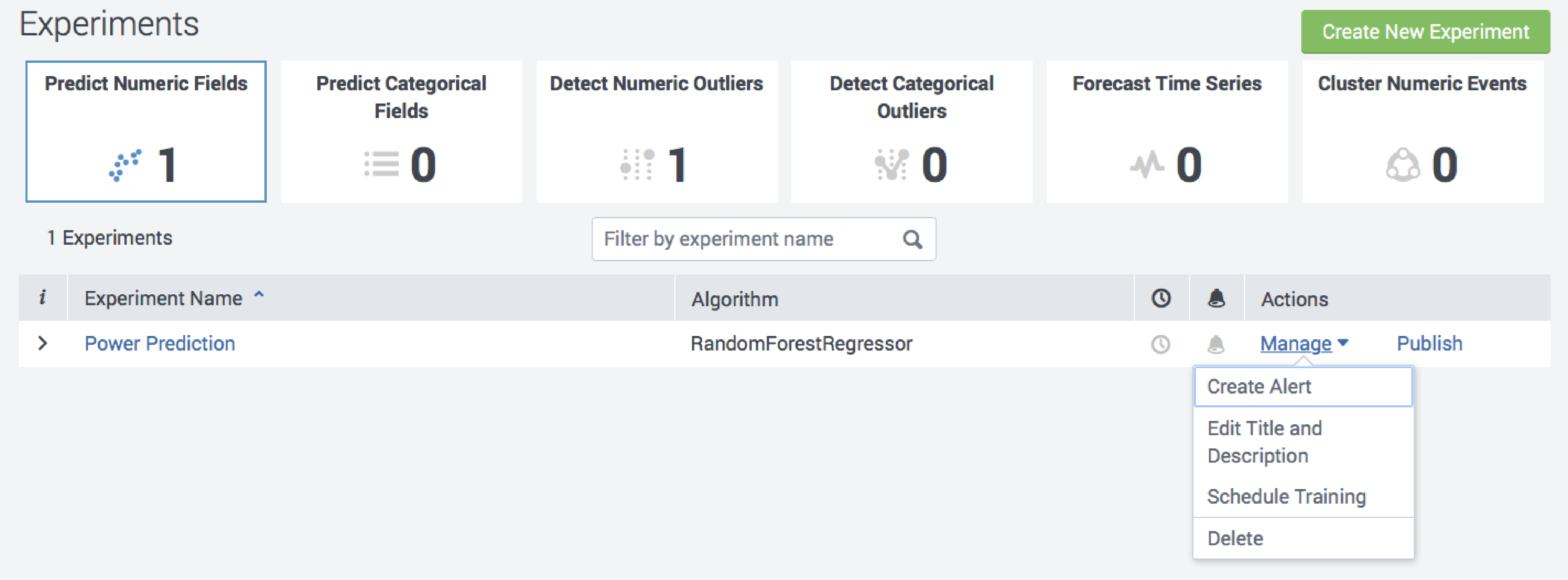
Root Mean Squared Error (RMSE)

2.53

Fit Model Parameters Summary

feature	importance
SS_total-last_level_cache_references	0.651435868505
SS_total-unhalted_core_cycles	0.176109167544
SS_total-cpu-utilization	0.16519150123
SS_total-memory_bus_transactions	0.00488073638554
SS_total-instructions_retired	0.00173402250591

Experiments: Deploy



Experiments

Create New Experiment

Predict Numeric Fields	Predict Categorical Fields	Detect Numeric Outliers	Detect Categorical Outliers	Forecast Time Series	Cluster Numeric Events
1	0	1	0	0	0

1 Experiments

Filter by experiment name

i	Experiment Name ˄	Algorithm	🕓	🔔	Actions	
>	Power Prediction	RandomForestRegressor	🕓	🔔	Manage ▾	Publish

Experiments

- Predict Numeric Fields
- Predict Categorical Fields
- Detect Numeric Outliers
- Detect Categorical Outliers
- Forecast Time Series
- Cluster Numeric Events

Let’s Build a Custom Model!

What’s New?

since last .conf

Major Highlights

(since .conf 2017)

Splunk Machine Learning Toolkit Updates

Includes new features for the Experiment Framework, algorithms, pre-processing steps, validation options etc.

Python for Scientific Computing 1.3 Update

Updated libraries giving you access to new and modified algorithms and its parameters.

Splunk MLTK Connector for Apache Spark™

Massive model building with MLlib directly from Splunk and SPL, No Scala skills required. (Limited Availability Release)

GitHub MLTK Community

Leverage and share algorithms collaboratively with the broader MLTK community

Splunk MLTK Container for Tensor Flow

Container based neural networks, leveraging GPUs/CPUs.

Splunk Machine Learning Toolkit Updates

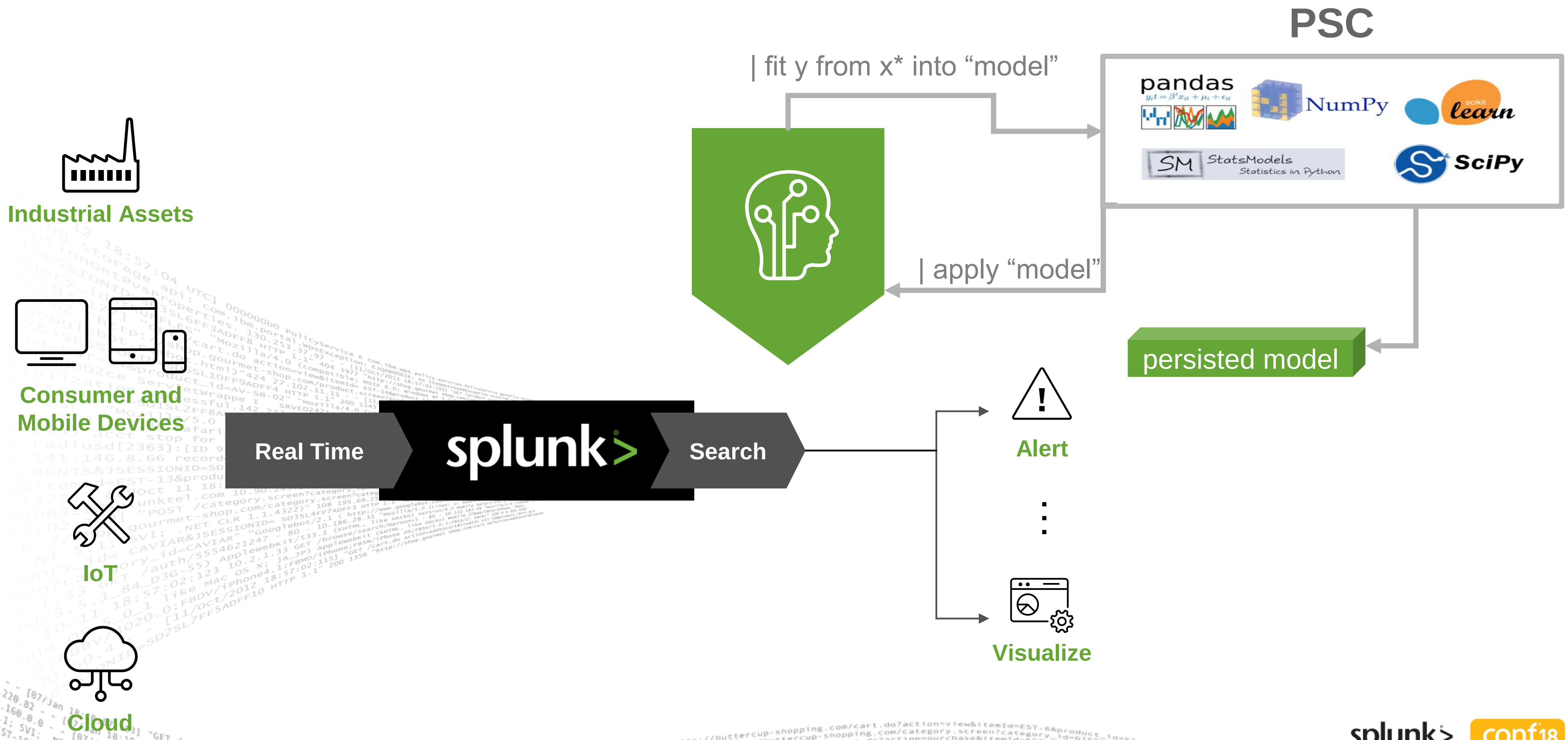
- **Experiment Management Framework:** A unified UI that provides the ability to:
  - ➔ Set roles based access control on experiments
  - ➔ Browsing and filtering pre-built models
  - ➔ Monitoring and scheduling alerts and searches
  - ➔ Getting history statistics about experiment’s previous runs and alerts
- **Score Command:** A new command for validating models and statistical tests for any use case, shipping with N algorithms today.
- **K-fold Cross-validation:** A popular and powerful way to quickly reduce model overfitting.
- **UI for MLSPL.CONF:** A interface to give user the power to change the safe settings if required for app level mlspl configuration.

Splunk Machine Learning Toolkit Updates

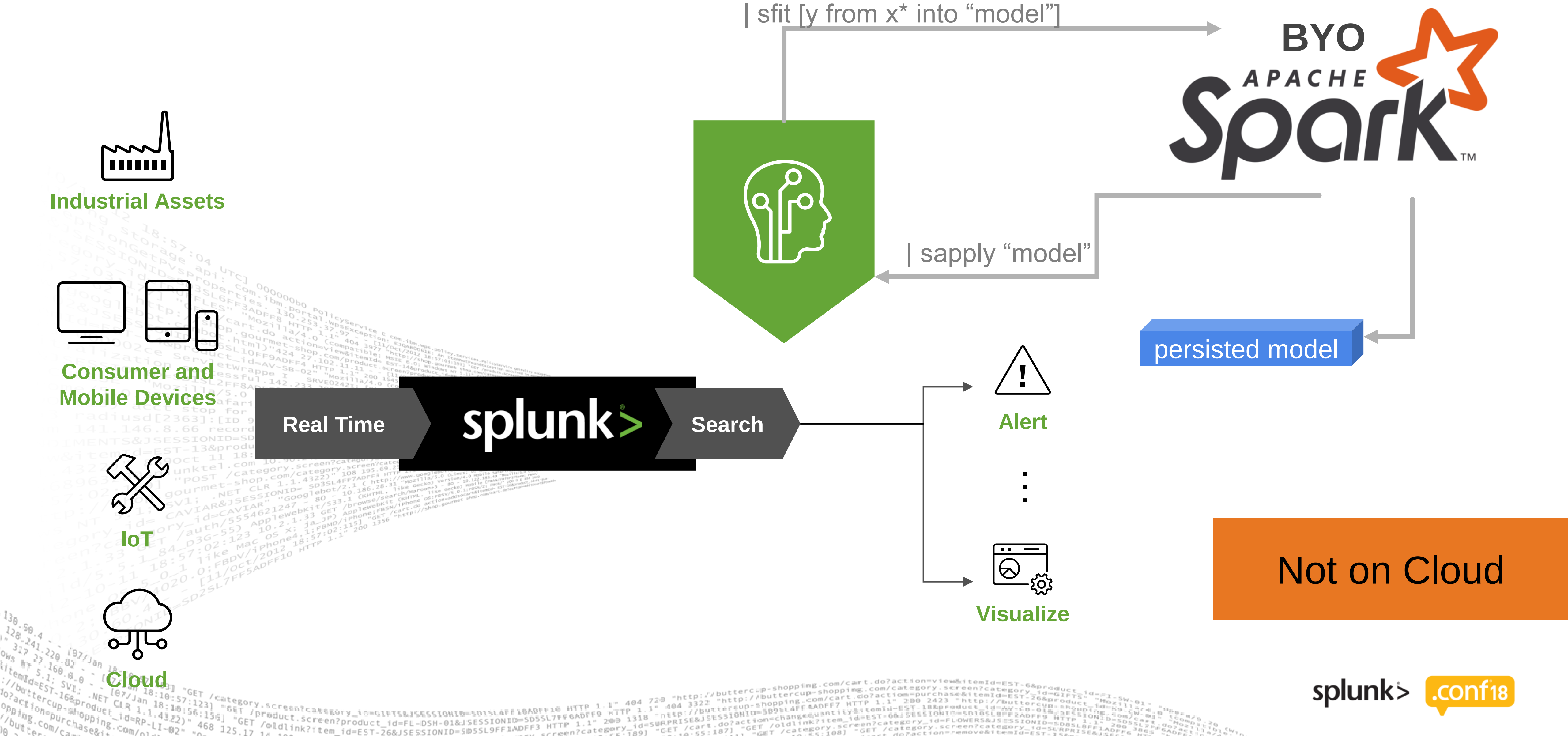
- **New out-of-the-box algorithms**
  - **Local Outlier Factor** : Unsupervised anomaly detection.
  - **Multi-layer Perceptron Classifier :** Neural network-based supervised classifier.
  - **Robust Scaler** : Re-scaling algorithm that is robust to outliers.
  - **X-Means:** Unsupervised clustering.

- **New pre-processing steps**
  - **Term Frequency-Inverse Document Frequency** : Feature extraction on unstructured text.
  - **Field Selector:** Feature selection.

MLTK 4.0 - Python for Scientific Computing 1.3



Splunk MLTK Connector for Apache Spark™ (Limited Availability Release)

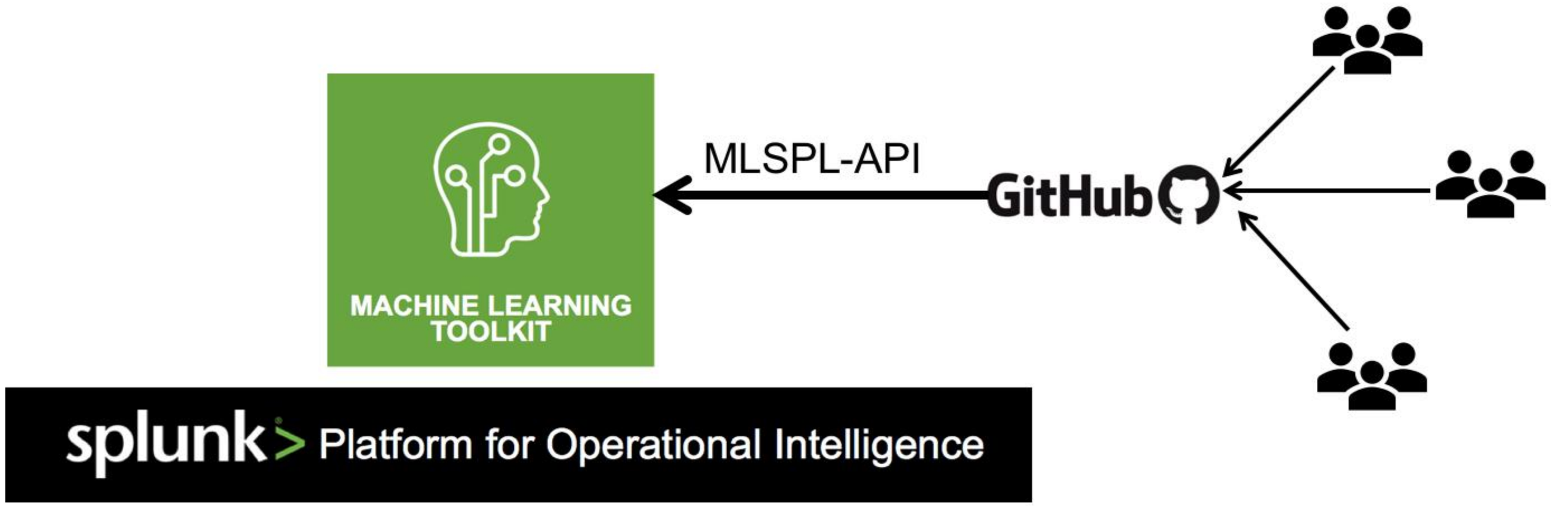


Not on Cloud

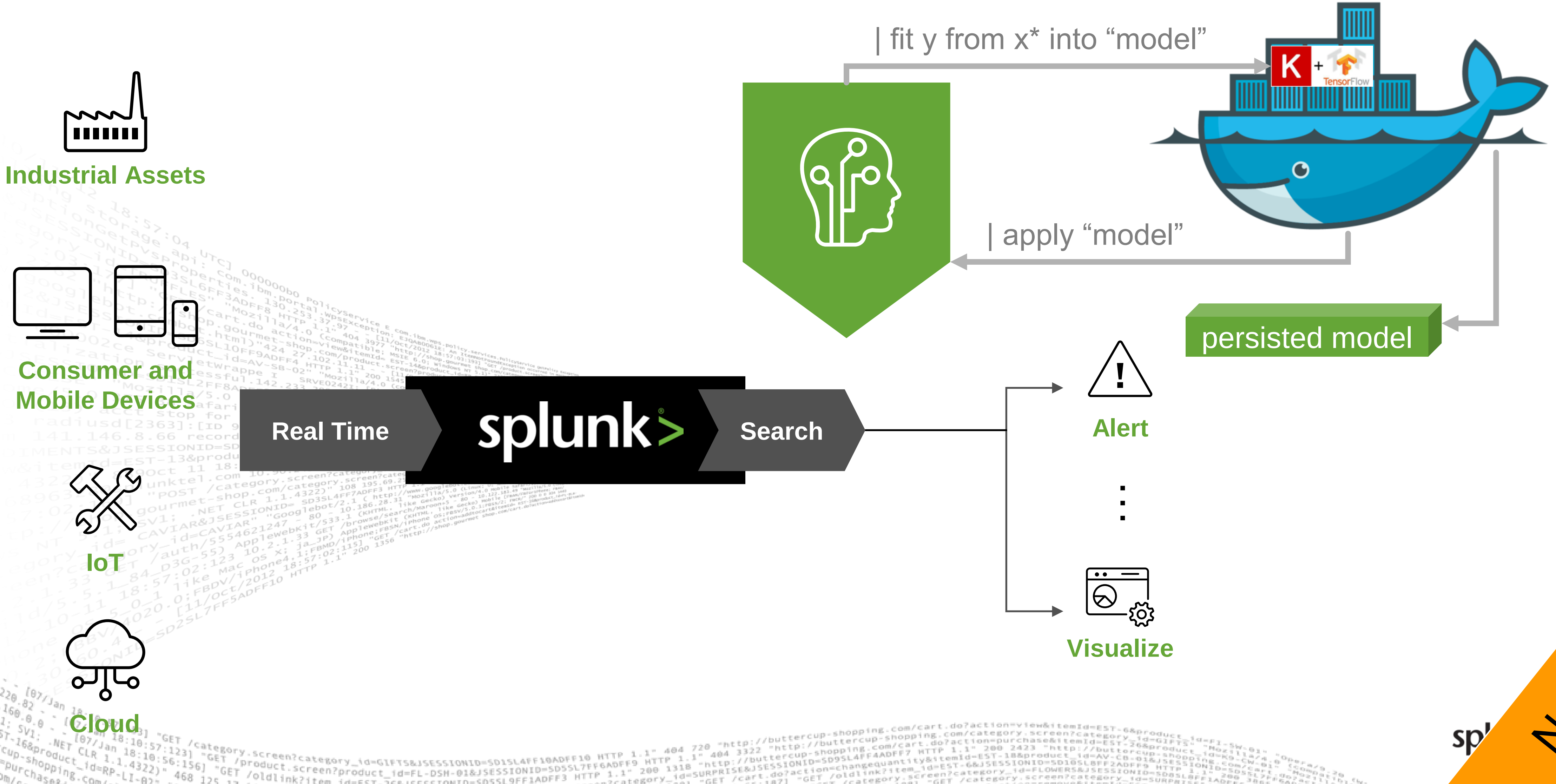
MLTK 4.0 -*Splunk Community for MLTK Algorithms on GitHub*

A community github for sharing algorithm files

> *“The creation of the Splunk Community for MLTK Algorithms on GitHub will help us find new functionality within the catalog at a much faster rate, which will allow us to get even more use out of the Splunk Machine Learning Toolkit,”*
> *said* ***Nathan Worsham, IS Security Administrator, Pinnacol***



Splunk MLTK Container for TensorFlow™ (via PS Whiteglove)



Not on Cloud

Want to know more?

Download the slides and recordings for these sessions.

FN1364 - Using Spark and MLLib for Large Scale Machine Learning With Splunk Machine Learning Toolkit

(Thursday, Oct 04, 11:00 a.m. - 11:45 a.m.)

Lin Ma, Principal Software Engineer, Splunk
Fred Zhang, Principal Data Scientist, Splunk

FN1409 - Thank You for Sharing: Expanding Machine Learning using Splunk MLTK GitHub Collaboration

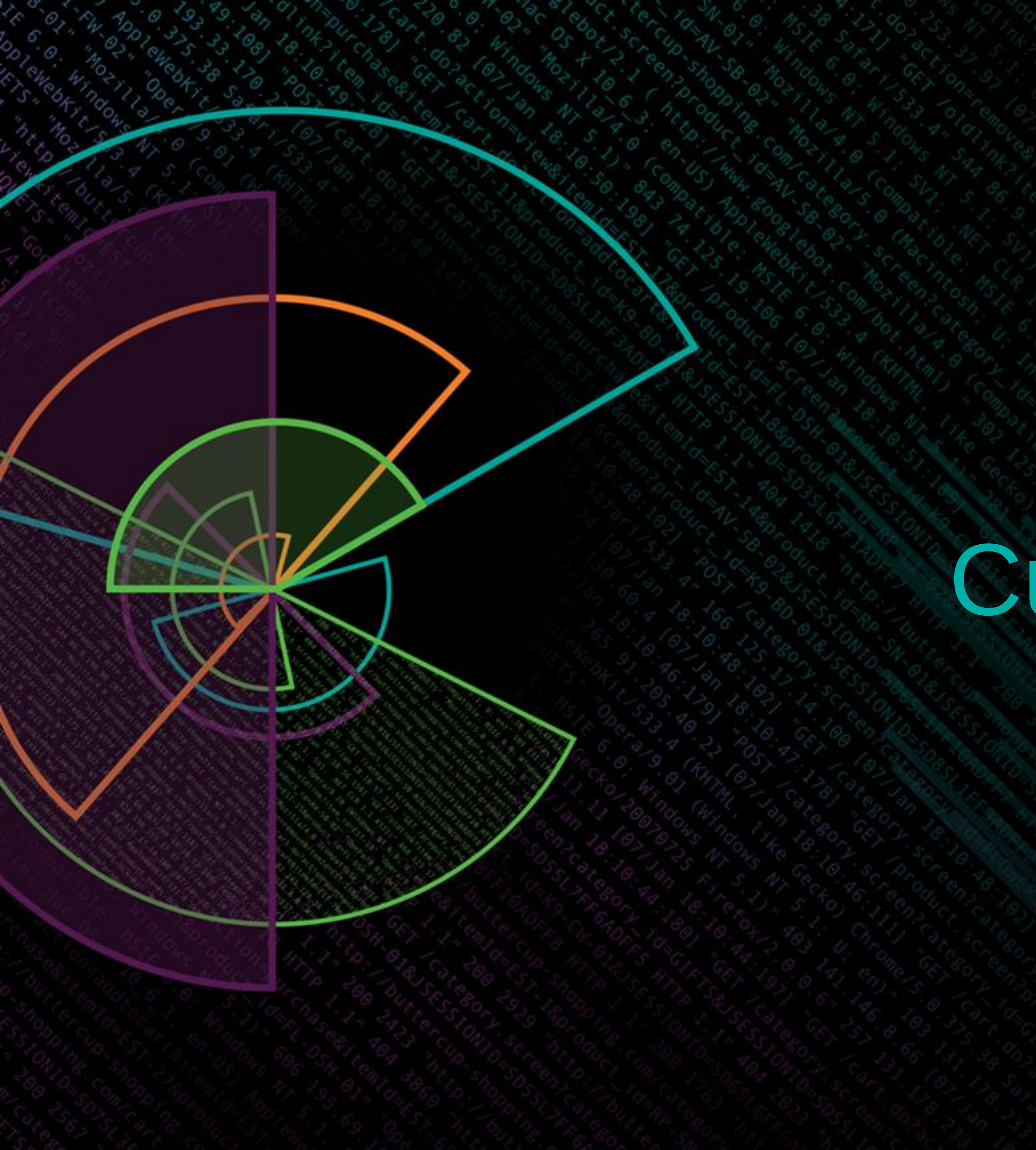
(Thursday, Oct 04, 11:00 a.m. - 11:45 a.m.)

Gyanendra Rana, Senior Product Manager, Splunk
Nathan Worsham, IS Security Administrator, Pinnacol Assurance

FN1478 - Exciting, To-Be-Announced Platform Session

Wednesday, Oct 03, 4:30 p.m. - 5:15 p.m.

Phillipp Drieger, Staff Machine Learning Architect , Splunk



Customer Success

splunk> .conf18

ML Success Story

Consumer Credit Reporting Agency

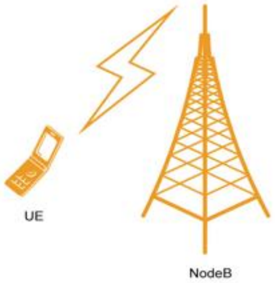
Acting on a Critical Customer Outages before the Customer Calls You

RECURSION pharmaceuticals

Many different machines are part of the drug discovery process, and machines acting abnormally mean a loss in efficiency and increased costs.

TELUS

the future is friendly®



Detect interference in cell towers
Re-configure underperforming cells for optimal services levels

Telco

Improving cell tower uptime and reducing repair truck rolls with anomaly detection and root cause analysis

ML Success Story

Entertainment Company

Predicting and averting potential gaming outage conditions with finer-grained detection

Preventing fraud by Identifying malicious accounts and suspicious activities

Find errors in server pools, then prioritize actions and associate root cause

Online Retailer

Failed orders detected in real time to avoid lost revenue and unhappy customers

Predicting Student Achievement and taking action to improve grades

The latest release of Splunk Machine Learning Toolkit makes it significantly easier to process large amounts of data and find patterns to see what's right or wrong. Splunk’s continued evolution of the Experiment Management Framework, including new tools to help validate our machine learning models, streamlines the complicated process of operationalizing machine learning.

Sundaresh Ramanathan, Director, IT Operations Analytics, Kinney Group, Inc.

Thank You

Don't forget to rate this session in the .conf18 mobile app

.conf18

splunk>