splunk> .conf2017

Using The Latest Features From The Splunk Machine Learning Toolkit To Create Your Own Custom Models

Adam J. Oliner | Director of Engineering
Manish Sainani | Director of Product

September 2017 | Washington, DC

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.

splunk> .conf2017

Outline

- Splunk Machine Learning Toolkit
- Platform Extensions: ML-SPL, etc.
- The Assistants: Guided Machine Learning
- What's New
- Demo
  - DIY Anomaly Detector

Splunk Machine Learning Toolkit

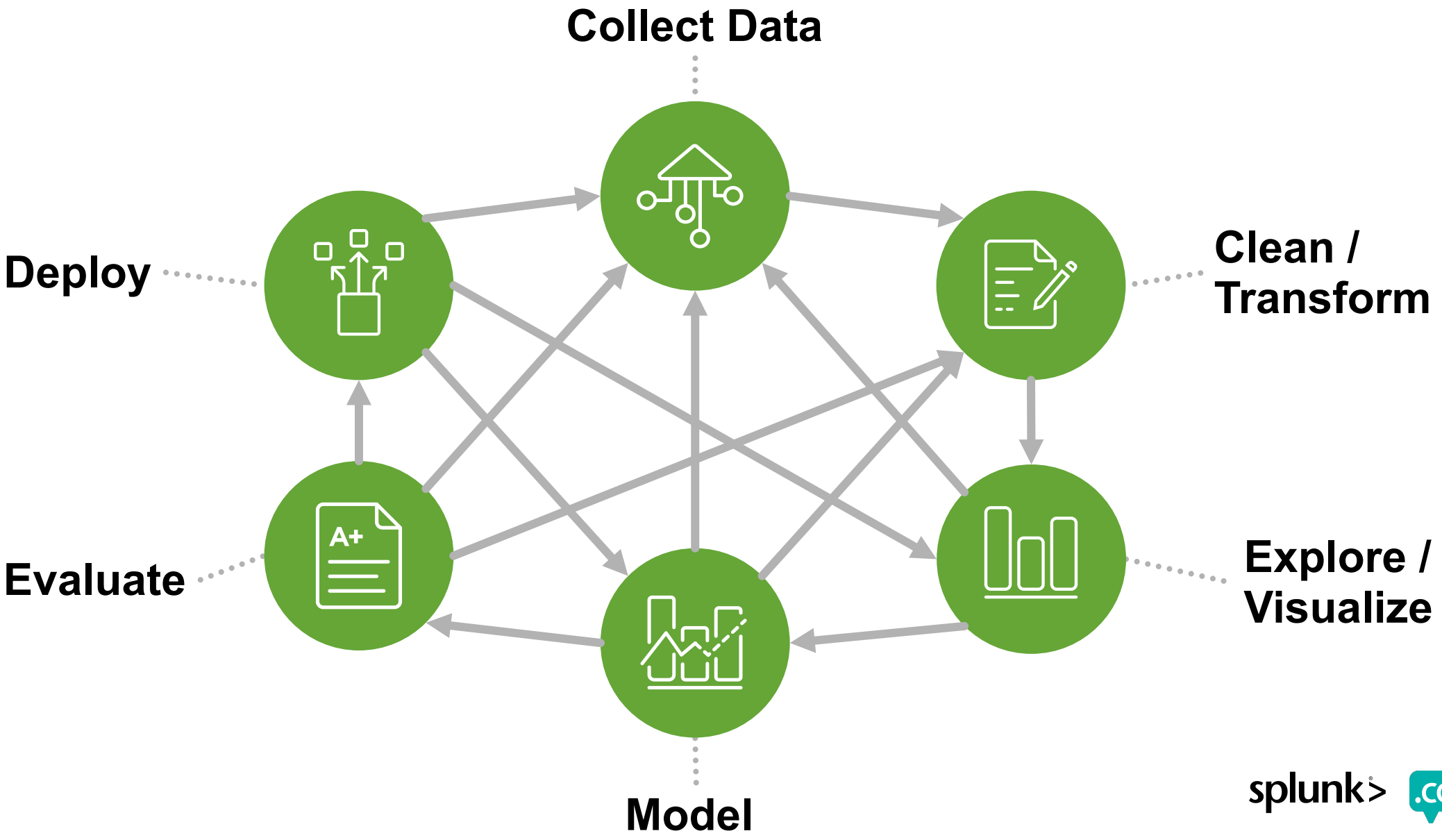
platform extensions and guided modeling dashboards

splunk> .conf2017

Machine Learning

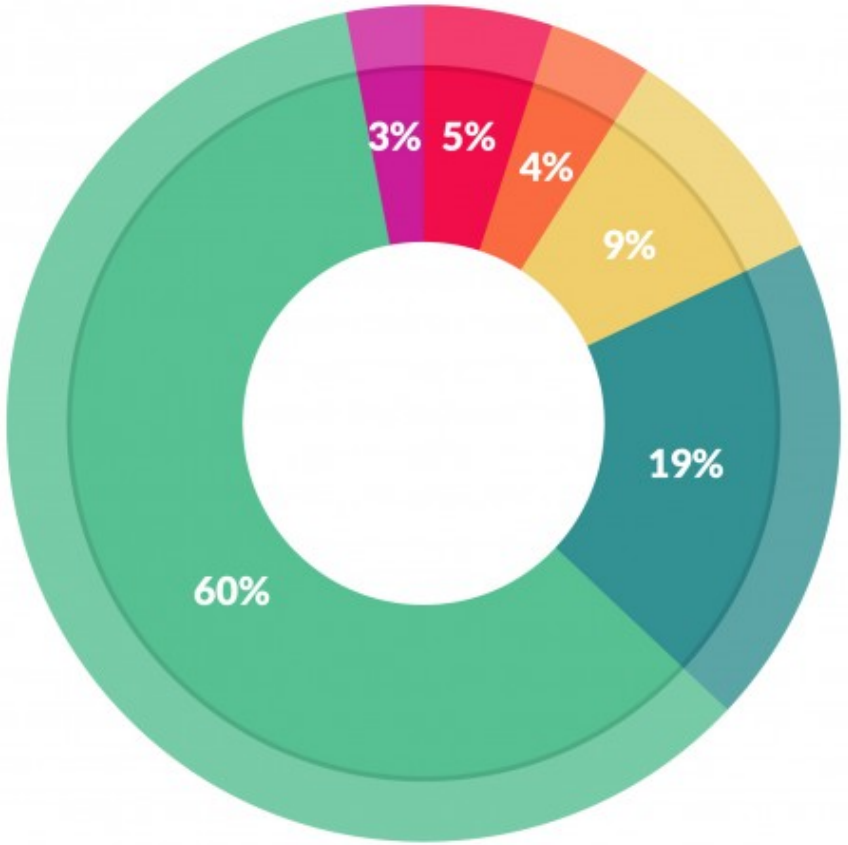
- A process for generalizing from examples
- Examples
  - A, B, … → # (regression)
  - A, B, ... → a (classification)
  - $X_{past} \rightarrow X_{future}$ (forecasting)
  - like with like (clustering)
  - $|X_{predicted} - X_{actual}| >> 0$ (anomaly detection)

Machine Learning Process



Data Gathering And Prep

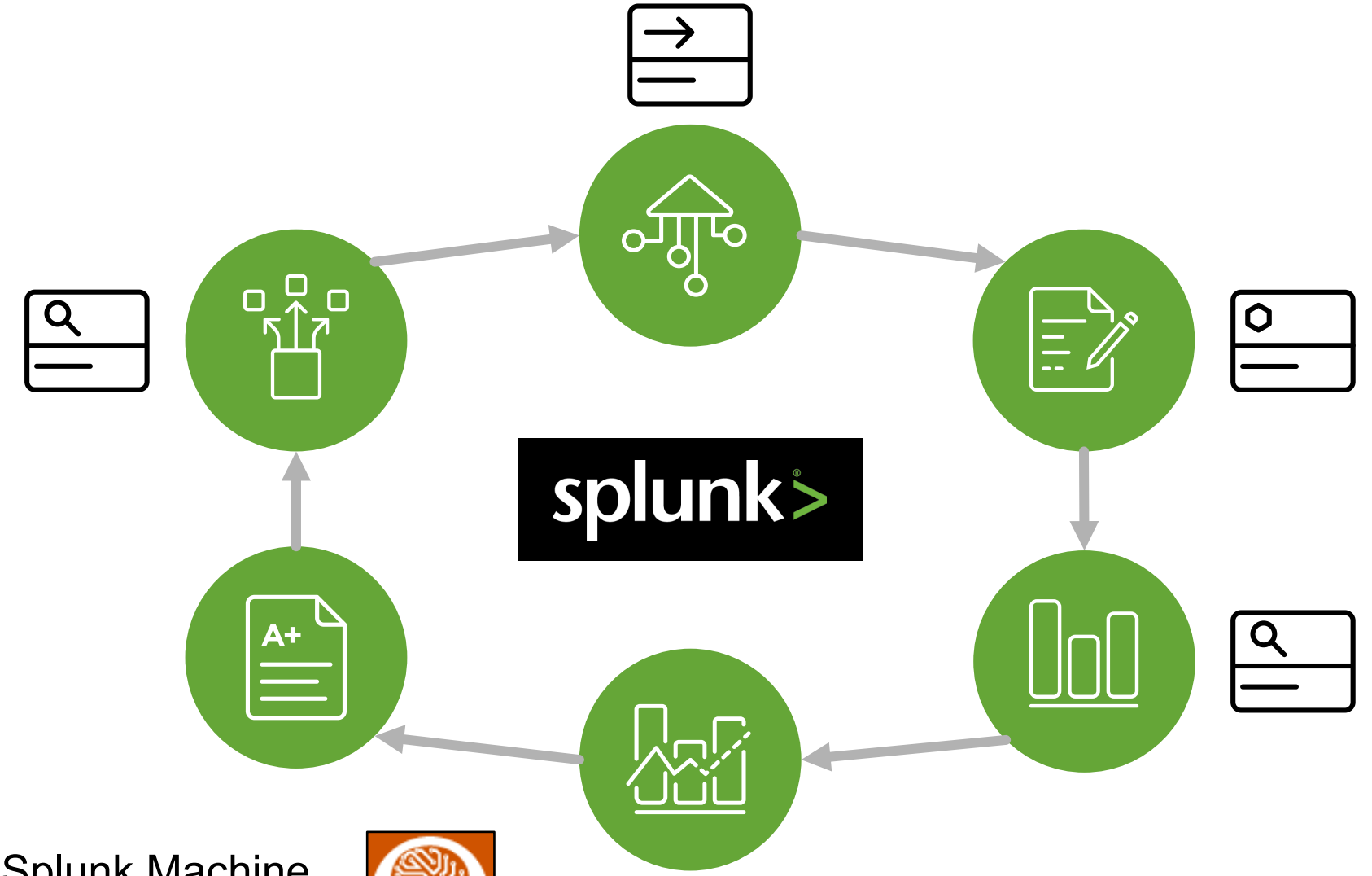
Source: CrowdFlower



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%*

Machine Learning Process With Splunk



Alerts,
Dashboards,
Reports

props.conf,
transforms.conf,
Datamodels,
Splunkbase

Pivot,
Table UI,
SPL

Splunk Machine
Learning Toolkit

Overview Of Machine Learning At Splunk

Splunk Premium Solutions

Splunk IT Service Intelligence™

Splunk Enterprise Security™

Splunk User Behavior Analytics™

Rich Ecosystem of Apps & Add-Ons

amazon web services

cisco

paloalto NETWORKS

box

kepware

puppet

Machine Learning Toolkit

splunk>enterprise

splunk>cloud™

splunk> **Platform for Operational Intelligence**

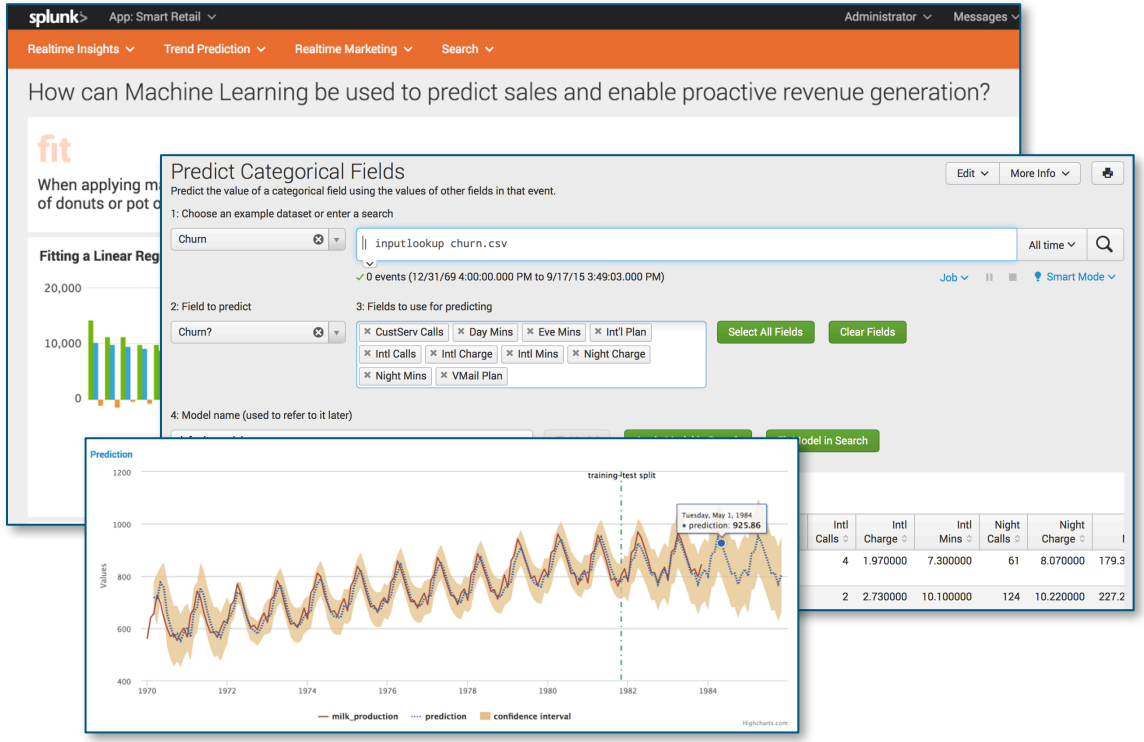
Forwarders	Syslog/ TCP	Mobile	IoT Devices	Network Wire Data	Hadoop	Relational Databases	Mainframe Data

splunk> .conf2017

Splunk Machine Learning Toolkit

extends Splunk with new tools and guided modeling

- **Assistants:** Guide model building, testing, & deployment for common tasks
- **Showcase:** 25+ interactive examples from IT, security, business, and IoT
- **Algorithms:** 25+ standard algorithms plus an extensibility API
- **SPL ML Commands:** New commands to fit, test, and operationalize models
- **Python for Scientific Computing Library:** 300+ open-source algorithms



splunk> .conf2017

Machine Learning Customer Success

TELUS

Network Optimization
Detect & Prevent Equipment Failure

NTT docomo

Security / Fraud Prevention

Telco

Prevent Cell Tower Failure
Optimize Repair Operations

Zillow®

Prioritize Website Issues
and Predict Root Cause

Entertainment Company

Predict Gaming Outages
Fraud Prevention

CONCANON™
INSIGHT ON DEMAND

Machine Learning Consulting Services

SCIANTA ANALYTICS
DEEP INSIGHT™

Analytics App built on ML Toolkit

splunk> .conf2017

Machine Learning Toolkit Customer Use Cases

TELUS

Reducing customer service disruption with early identification of difficult-to-detect network incidents

Minimizing cell tower degradation and downtime with improved issue detection sensitivity

---

Zillow

Speeding website problem resolution by automatically ranking actions for support engineers

---

NTT docomo

Ensuring mobile device security by detecting anomalies in ID authentication

---

Entertainment Company

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

Preventing fraud by Identifying malicious accounts and suspicious activities

---

Telco

Improving uptime and lowering costs by predicting/preventing cell tower failures and optimizing repair truck rolls

splunk> .conf2017

Platform Extensions: ML-SPL, etc.

custom search commands for machine learning

splunk> .conf2017

SPL, Macros, & Viz

Oh, my!

- Commands (ML-SPL)
  - fit
  - apply
  - summary
  - listmodels
  - deletemodel
  - sample
- 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

splunk> .conf2017

ML-SPL

What is it?

- A suite of SPL commands specifically for machine learning
  - modeling
  - sampling
- Most are implemented using modules from the Python for Scientific Computing add-on for Splunk
  - scikit-learn
  - numpy
  - pandas
  - statsmodels
  - scipy

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>
```

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

- 25+ 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: 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`

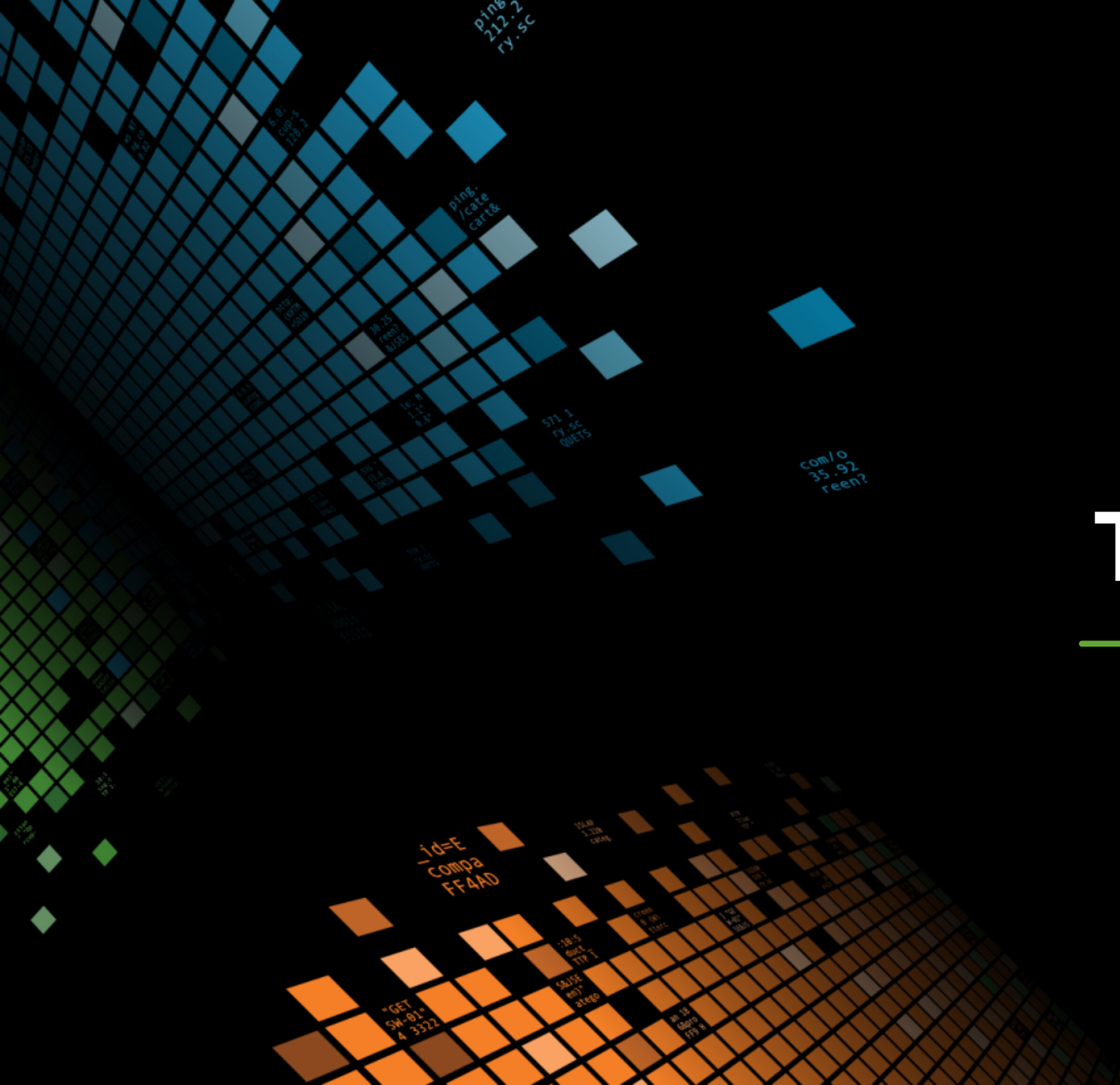
splunk> .conf2017

Plug: Machine Learning In SPL With The Machine Learning Toolkit

Tuesday, September 26th @ 2:15pm in Ballroom B

- Custom search commands!
- Tabular munging!
- Jacob Leverich!
- Shang Cai!

splunk> .conf2017



The Assistants

guided machine learning

splunk> .conf2017

Guided ML With The Assistants

- Guides you through an analytic
  - Prepare, fit, validate, and deploy
- Automatically generates all the relevant SPL

Fit a model on all your data in search

```
| inputlookup server_power.csv

| fit StandardScaler "total-cpu-utilization", "total-disk-accesses", "total-disk-blocks", "total-disk-utilization", "total-instructions_retired", "total-last_level_cache_references", "total-memory_bus_transactions", "total-unhalted_core_cycles" with_mean=true with_std=true into example_server_power_StandardScaler_0     // apply preprocessing steps

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

splunk> .conf2017

Assistants: Prepare

Preprocessing Steps

StandardScaler

Preprocess method: StandardScaler

Fields to preprocess: business_acres, property_tax_rate, distance_to_employment_center

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

Apply

PCA

Preprocess method: PCA

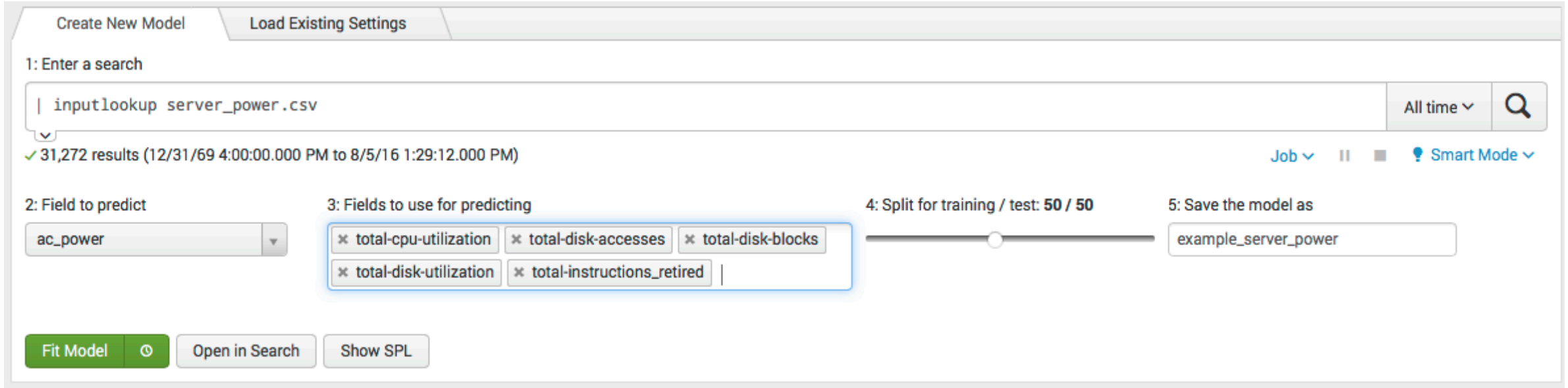
Fields to preprocess: highway_accessibility_index, distance_to_employment_center, pupil_teacher_ratio, crime_rate

K (# of Components): 2

Apply

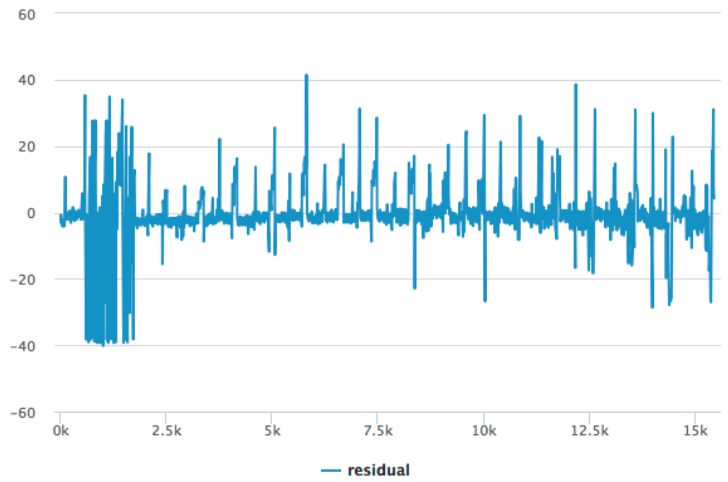
splunk> .conf2017

Assistants: Fit



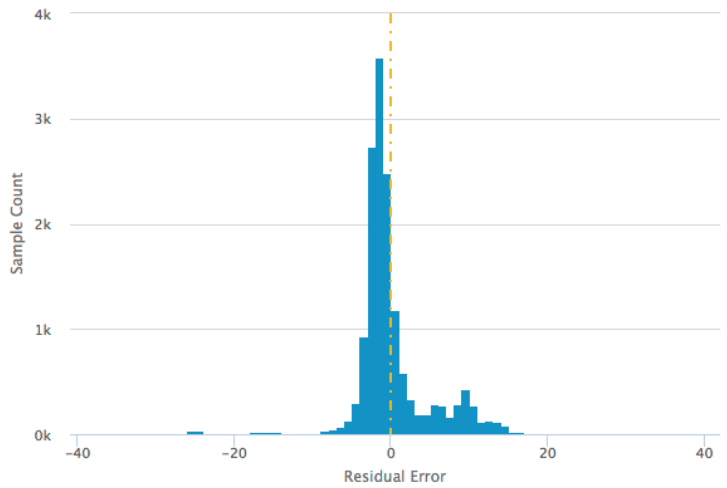
Assistants: Validate

Residuals Line Chart



Open in Search | Show SPL

Residuals Histogram



Open in Search | Show SPL

R^2 Statistic

0.9480

Root Mean Squared Error (RMSE)

5.55

Open in Search | Show SPL

Fit Model Parameters Summary

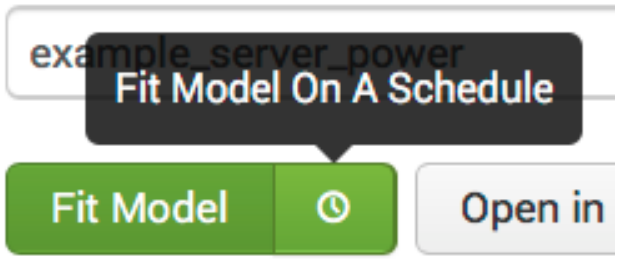
Prediction Feature	Linear Regression Coefficient
total-cpu-utilization	-0.069032370952
total-disk-accesses	-0.00900564584969
total-disk-blocks	-0.000284870204115
total-disk-utilization	0.0300082299744
total-instructions_retired	1.65392272339e-09

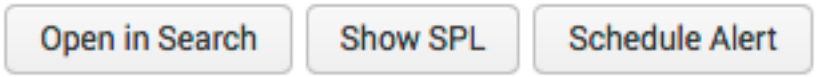
« prev 1 2 next »

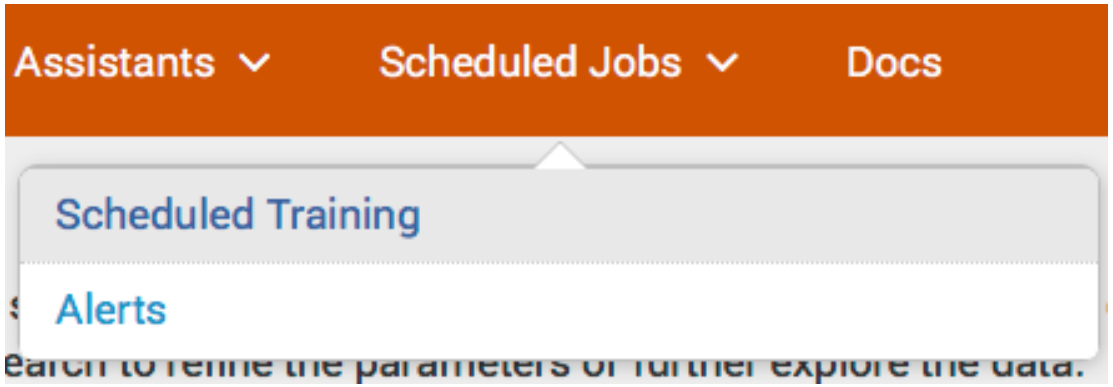
Open in Search | Show SPL

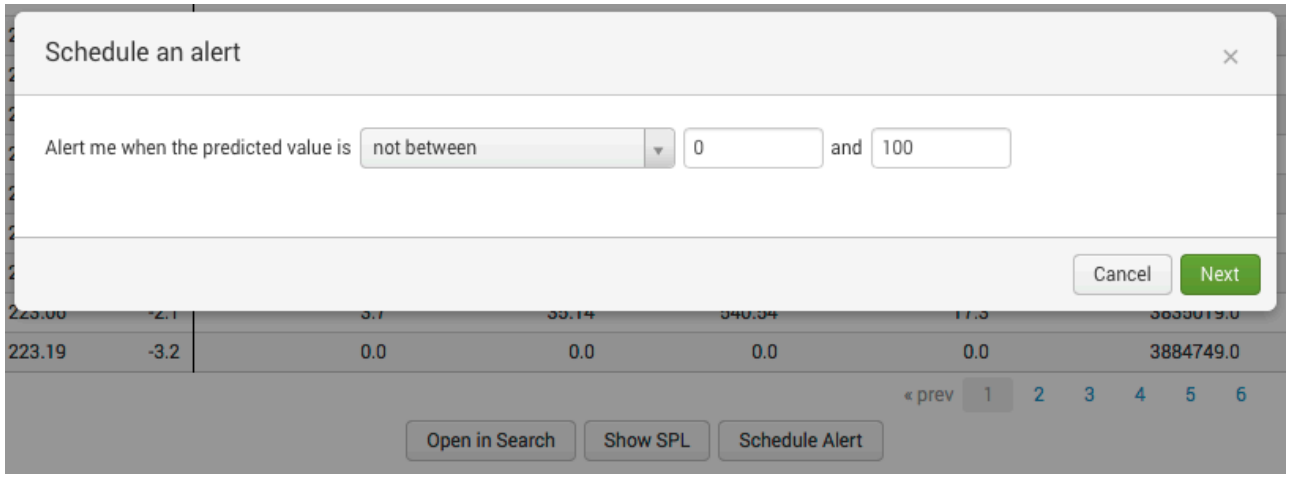
splunk> .conf2017

Assistants: Deploy







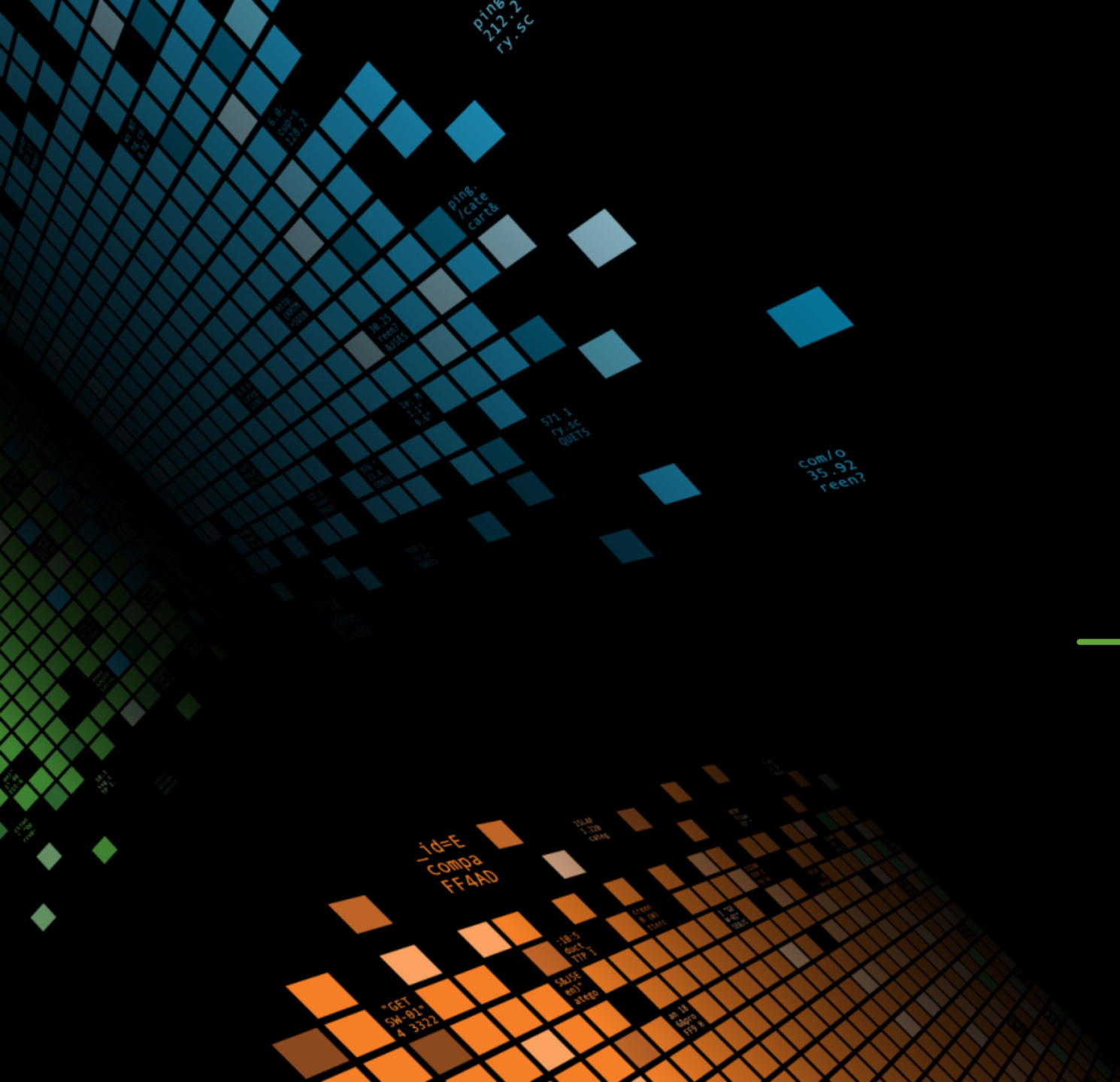


splunk> .conf2017

The Assistants

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

splunk> .conf2017



What's New

since last .conf

splunk> .conf2017

What's New

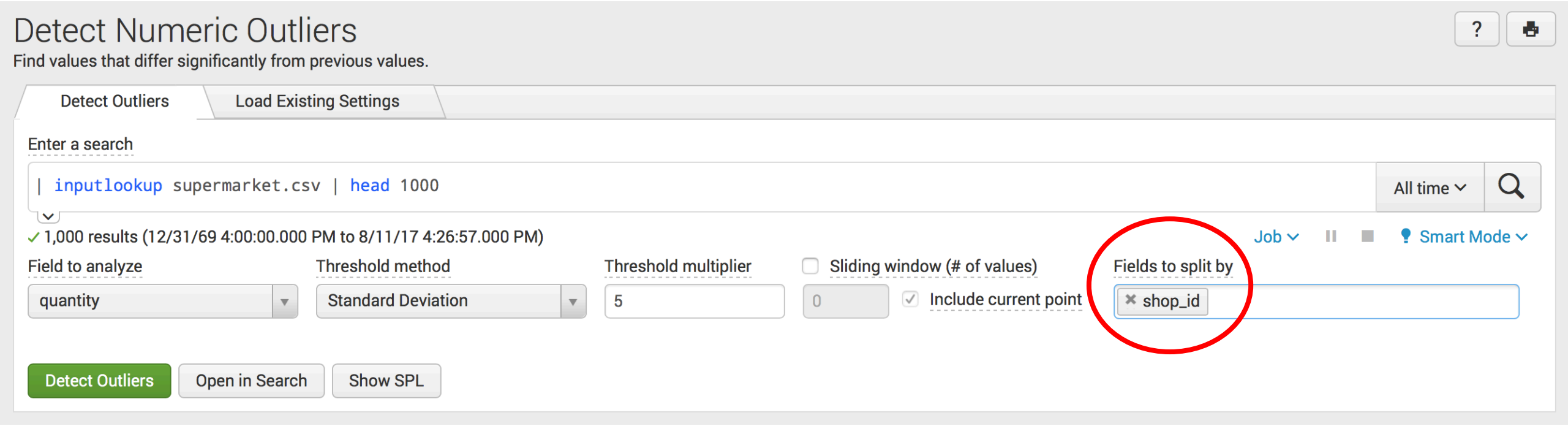
(since .conf 2016)

- **Detect Numeric Outliers improvements**
- **Preprocessing**
- **Model Management**
- **ML-SPL extensibility API**
- **Spark Support (private beta)**
- New algorithms:
  - ACF & PACF
  - ARIMA
- Load Existing Settings is per-user
- Downsampled Line Chart supports drilldown

splunk> .conf2017

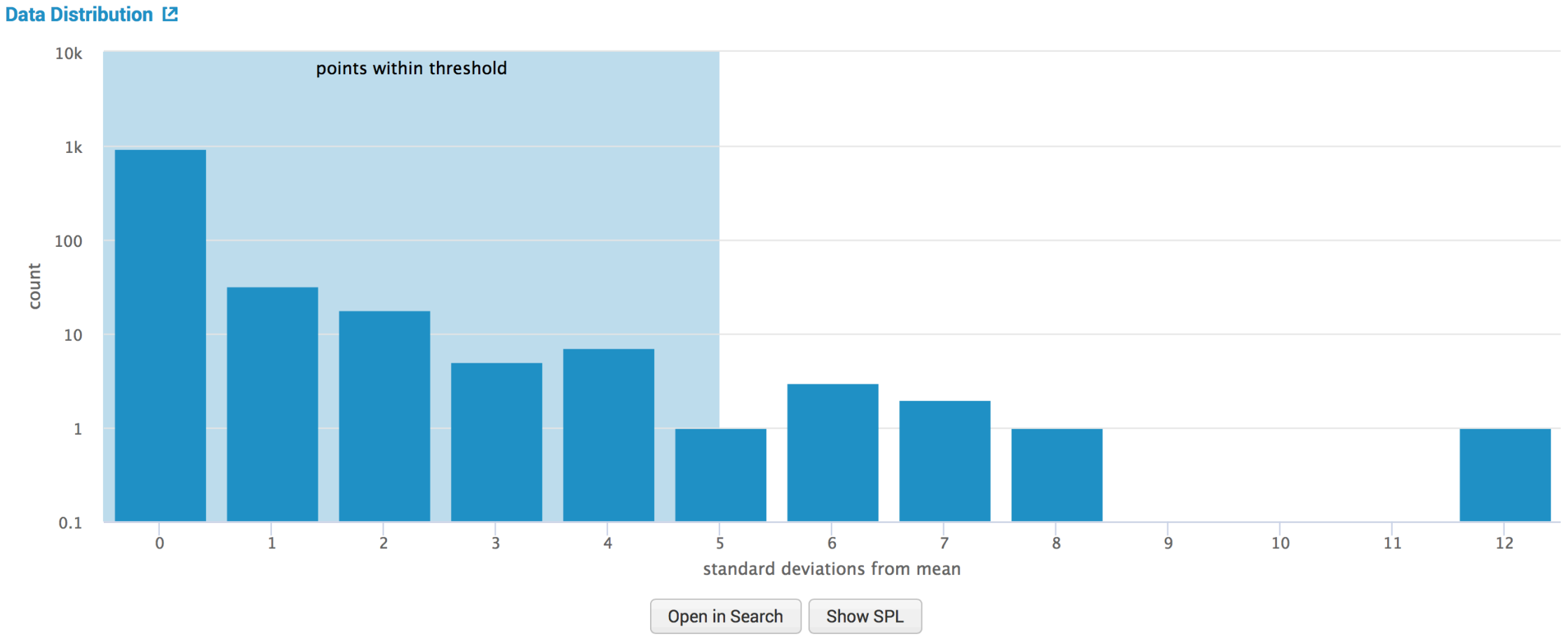
Detect Numeric Outliers

split-by support



Detect Numeric Outliers

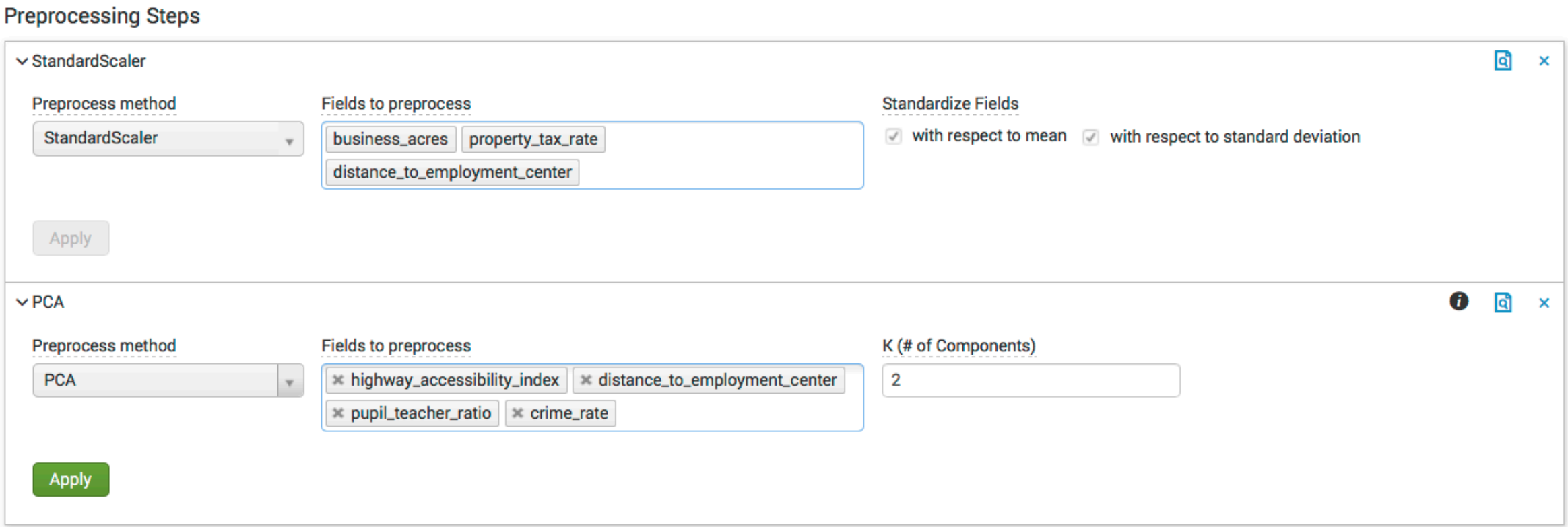
data distribution viz



Preprocessing

build a pipeline of data prep

- In Predict Numeric, Predict Categorical, and Cluster Numeric assistants

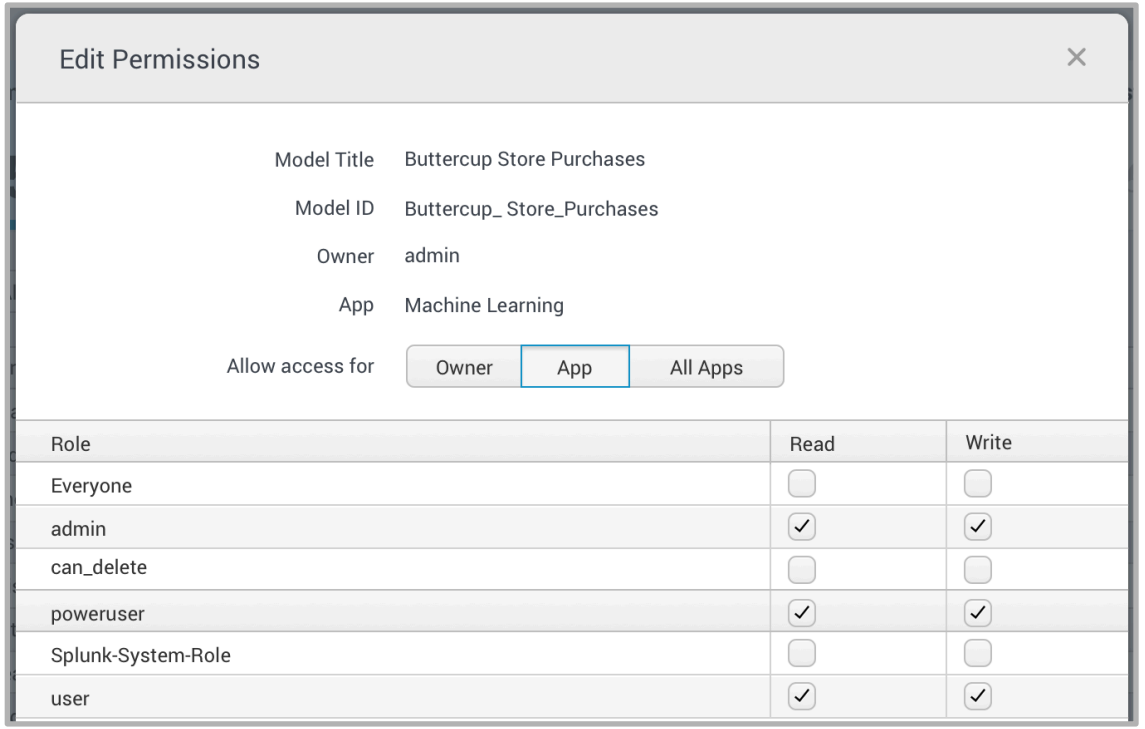


splunk> .conf2017

Model Management

RBAC & more

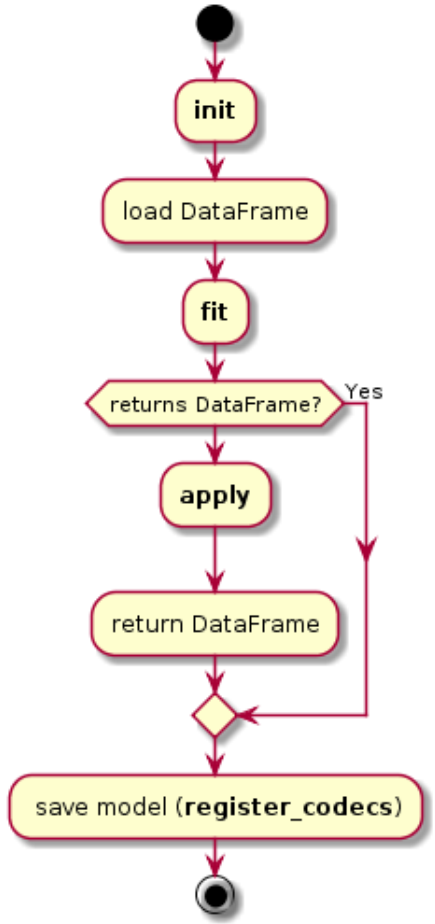
- Assign permissions to models to control access
- Manage models via the UI
- Experiments
  - Assistant configurations
  - May produce 1+ models



ML-SPL Extensibility API

featuring: primo documentation

- Make more algos available to **`fit`** / **`apply`**
  - 300+ in PSC
  - Custom algorithms
- Expose new or different parameters
- Docs include examples
  - Correlation Matrix
  - Agglomerative Clustering
  - Support Vector Regressor
  - Savitzky-Golay Filter
- Use in your apps / dashboards / etc.!



Plug: Advanced Machine Learning using the Extensible ML API

Wednesday, September 27th @ 4:35pm in Ballroom B

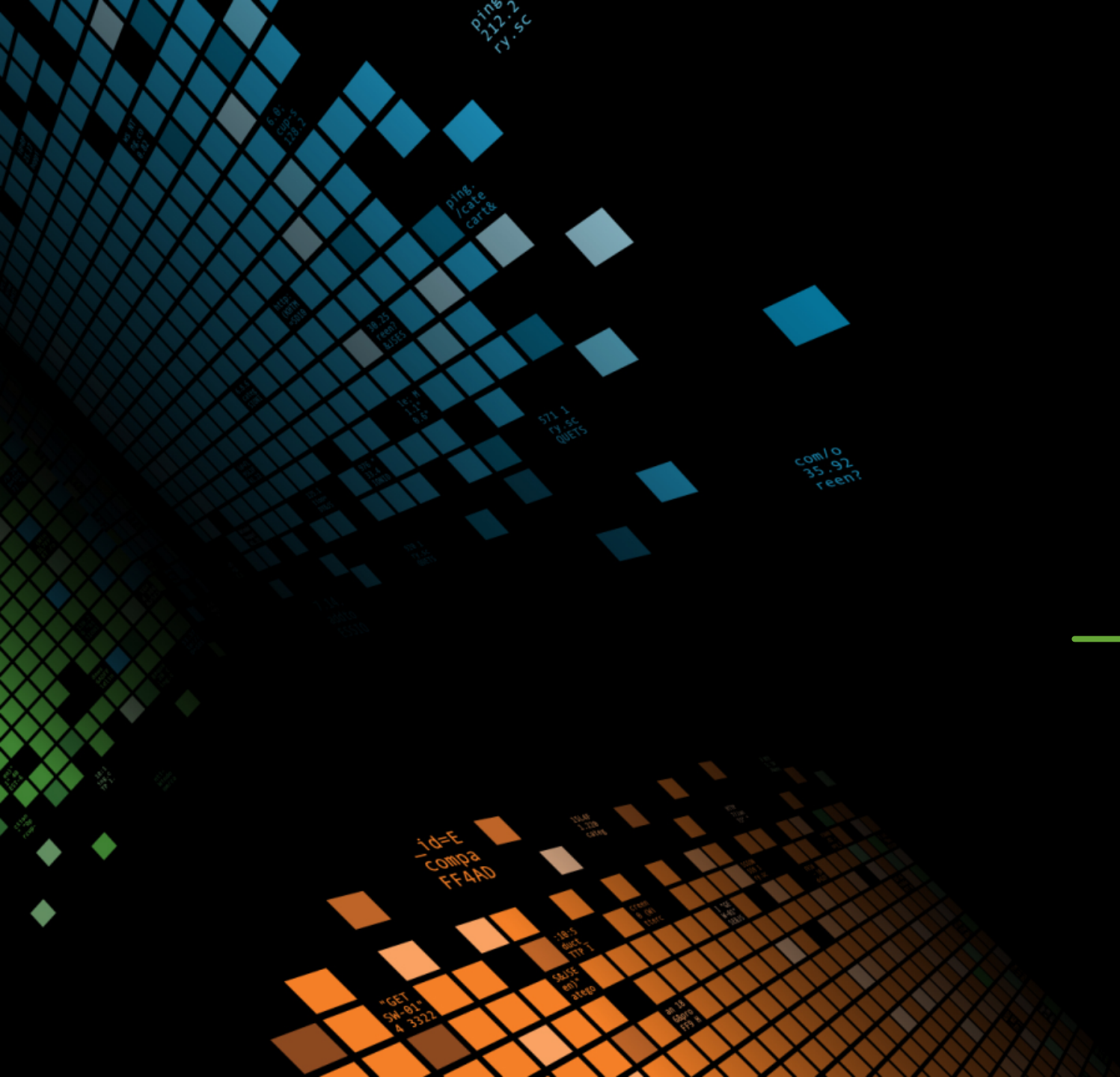
- Implementation details!
- Extensibility API!
- Alexander Johnson!
- Zidong Yang!

splunk> .conf2017

Spark Support

private beta open now

- Use your existing Spark cluster with MLTK
  - Distributed fit on massive datasets
  - Apply MLlib models for supported algos
- **`sfit` / `sapply`**
- Contact mlprogram@splunk.com
  - What is your use case (e.g., predicting server downtime)?
  - Why do you want / need Spark (i.e., why isn't MLTK sufficient)?



Demo

Q&A

Adam J. Oliner | Director of Engineering
Manish Sainani | Director of Product

splunk> .conf2017

Thank You

Don't forget to rate this session in the .conf2017 mobile app