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

Dr. Adam Oliner

Director of Engineering, Data Science, Splunk

Manish Sainani

Principal Product Manager, Splunk

.conf2016

splunk>

Disclaimer

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 the 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.

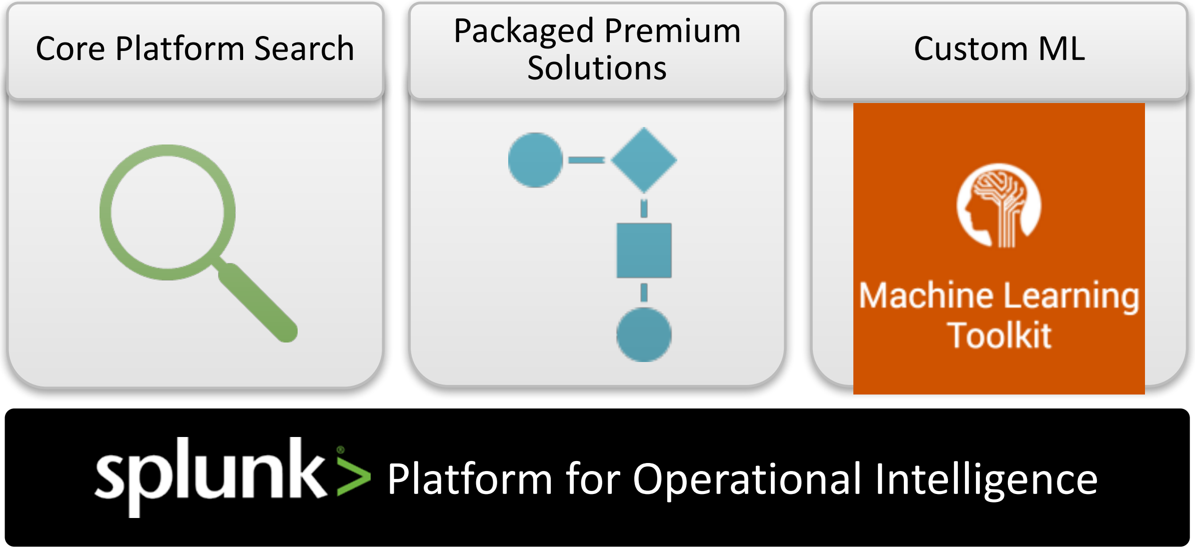
Who are we?

- Dr. Adam Oliner
  - Director of Engineering, Data Science & Machine Learning
  - Splunker for 2 years
  - Embarrassingly overeducated
- Manish Sainani
  - Principal Product Manager, Machine Learning
  - Splunker for 2 years
  - First ML hire at Splunk!

What are we doing here?

- Overview of Machine Learning
- The Assistants: Guided Machine Learning
  - Prepare
  - Fit
  - Validate
  - Deploy
- Examples
  - DIY Anomaly Detector
  - Customer Applications

Overview of ML at Splunk



Core Platform Search

Packaged Premium Solutions

Custom ML

Machine Learning Toolkit

splunk> Platform for Operational Intelligence

Splunk Machine Learning Toolkit

Extends Splunk platform functions and provides a guided modeling environment

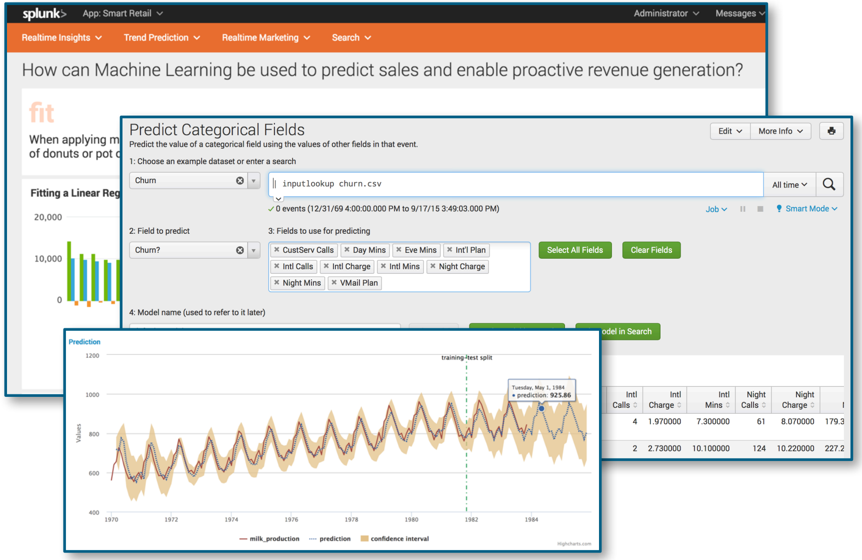
Assistants: Guide model building, testing, & deploying for common objectives

Showcases: Interactive examples for typical IT, security, business, IoT use cases

Algorithms: 25+ standard algorithms available prepackaged with the toolkit

SPL ML Commands: New commands to fit, test and operationalize models

Python for Scientific Computing Library: 300+ open source algorithms available for use



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	
4	1.970000	7.300000	61	8.070000	179.3
2	2.730000	10.100000	124	10.220000	227.2

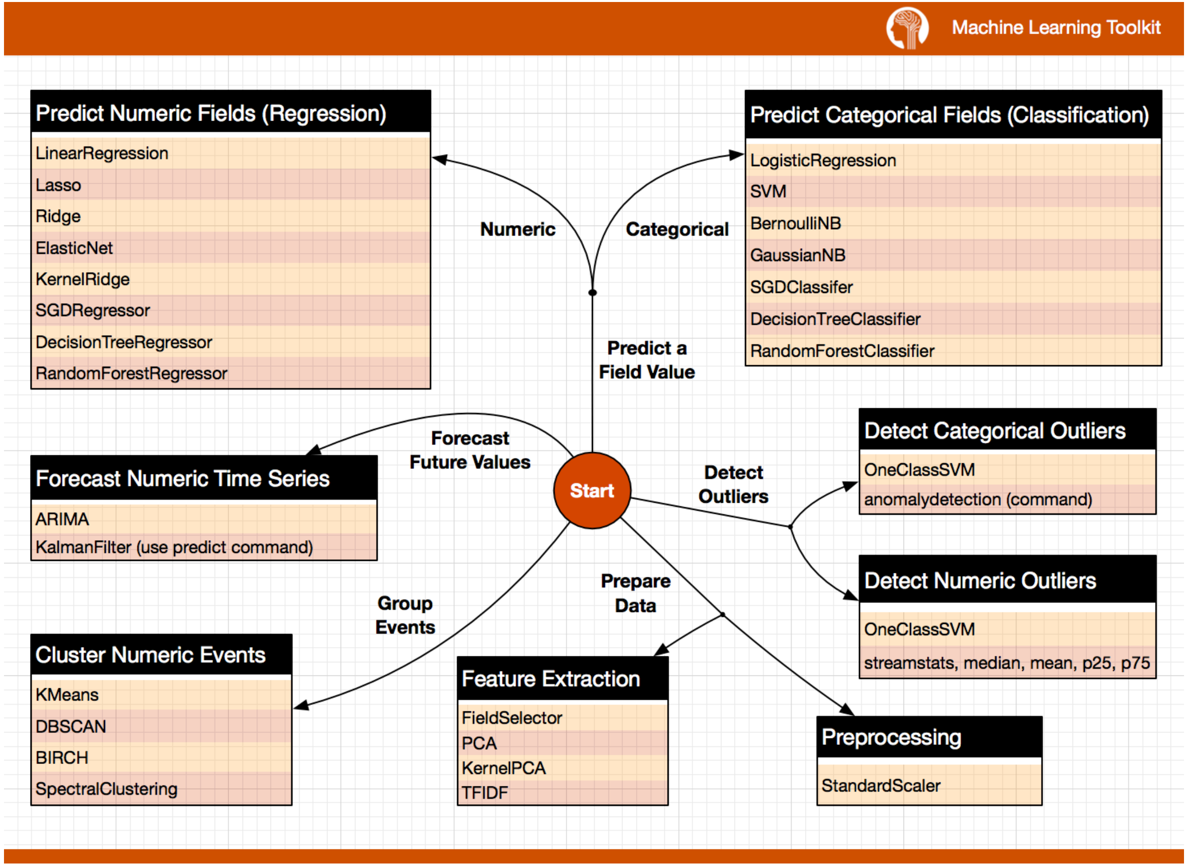
Build custom analytics for any use case

What's New since our 0.9 Beta Release (last year's .conf)?

- New name and abbreviation ;-)
- No event limits (removal of 50K limit on fitting models)
- Configurable resource caps via mlspl.conf
- Search head clustering support
- Distributed / streaming apply
- Scheduled fit
- New algorithms (next slide)
  - Feature engineering and selection
  - Stochastic gradient descent (e.g.)
  - ARIMA

- Multi-algorithm support across Assistants
- Scatterplot matrix viz
- Alerting
- Tooltips
- In-app tours
- Cluster Numeric Events assistant
- Videos videos videos for each assistant across IT, Security, IoT and Business Analytics
- ML-SPL Cheat Sheet

Algorithms supported (v2.0, .conf2016)



Machine Learning Toolkit

Start

Predict a Field Value

Numeric

Categorical

Forecast Future Values

Detect Outliers

Group Events

Prepare Data

Predict Numeric Fields (Regression)

- LinearRegression
- Lasso
- Ridge
- ElasticNet
- KernelRidge
- SGDRegressor
- DecisionTreeRegressor
- RandomForestRegressor

Predict Categorical Fields (Classification)

- LogisticRegression
- SVM
- BernoulliNB
- GaussianNB
- SGDClassifer
- DecisionTreeClassifier
- RandomForestClassifier

Forecast Numeric Time Series

- ARIMA
- KalmanFilter (use predict command)

Detect Categorical Outliers

- OneClassSVM
- anomalydetection (command)

Detect Numeric Outliers

- OneClassSVM
- streamstats, median, mean, p25, p75

Cluster Numeric Events

- KMeans
- DBSCAN
- BIRCH
- SpectralClustering

Feature Extraction

- FieldSelector
- PCA
- KernelPCA
- TFIDF

Preprocessing

- StandardScaler

The Assistants:
Guided Machine Learning

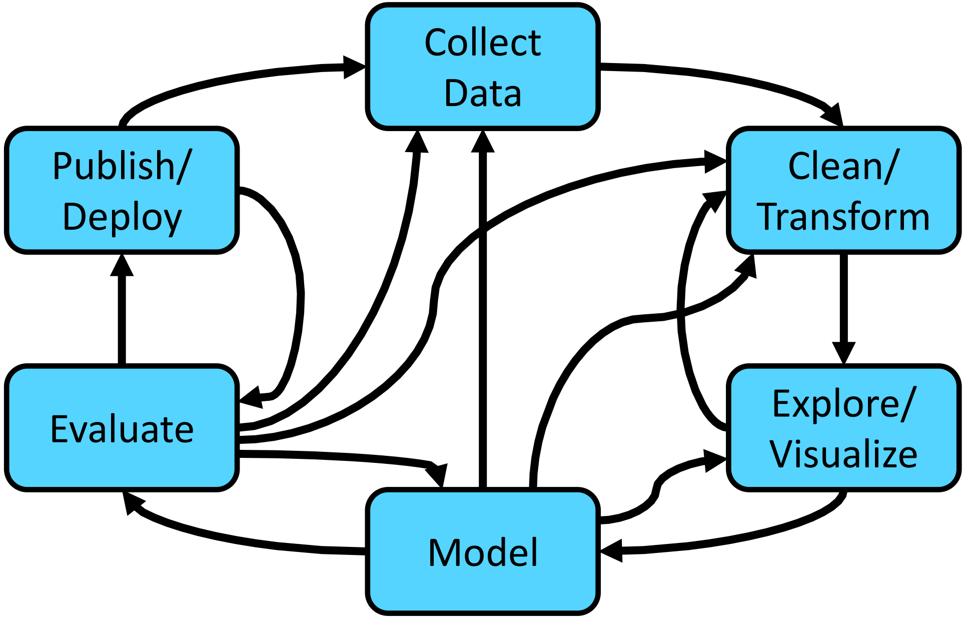
.conf2016

splunk>

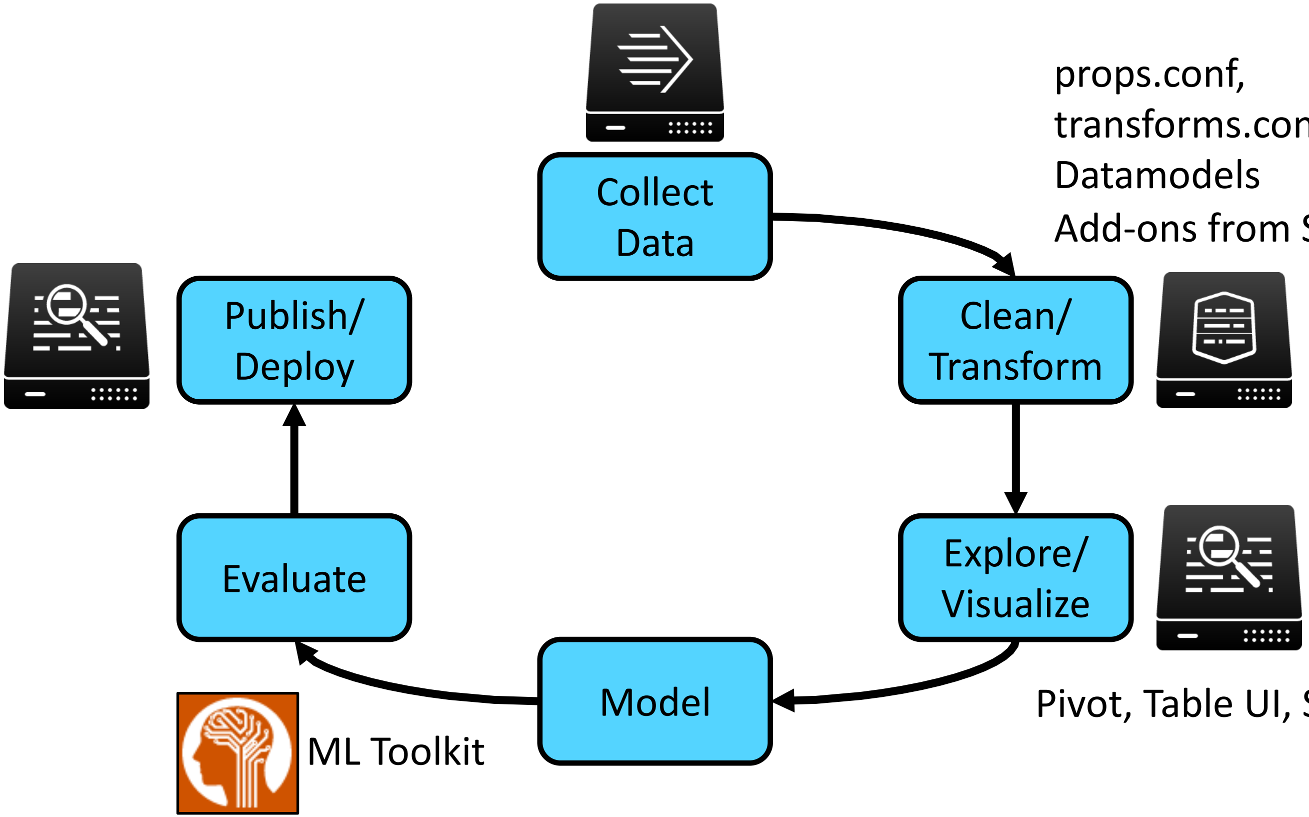
Machine Learning

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

Machine Learning Process



Machine Learning Process with Splunk



Collect Data

Clean/ Transform

props.conf,
transforms.conf,
Datamodels
Add-ons from Splunkbase, etc.

Explore/ Visualize

Pivot, Table UI, SPL

Model

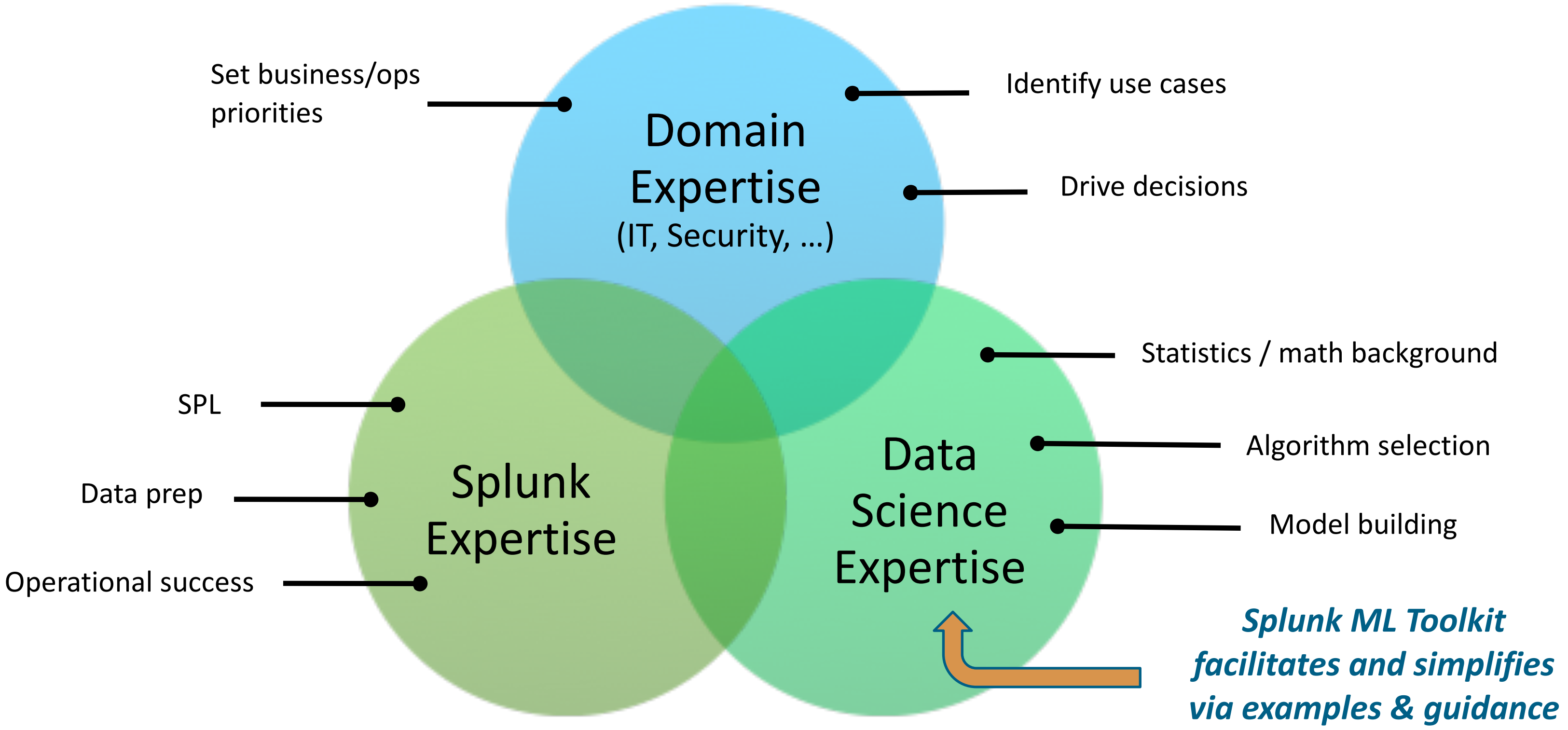
ML Toolkit

Evaluate

Publish/ Deploy

Alerts,
Dashboards,
Reports

Custom Machine Learning – Success Formula

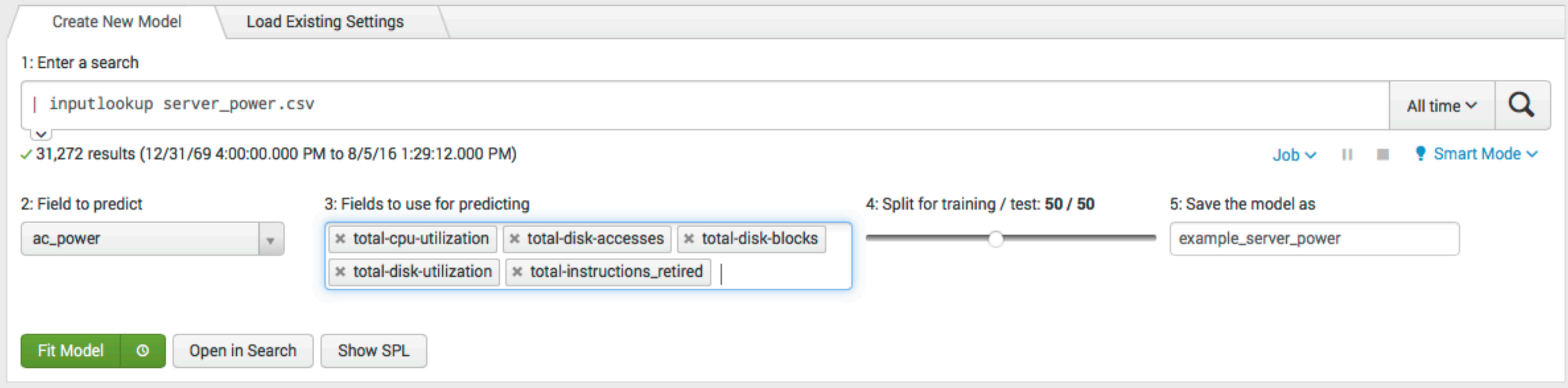


Splunk ML Toolkit facilitates and simplifies via examples & guidance

Guided ML with the Assistants

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

Assistants: Fit



1: Enter a search

```
| inputlookup server_power.csv
```

✓ 31,272 results (12/31/69 4:00:00.000 PM to 8/5/16 1:29:12.000 PM)

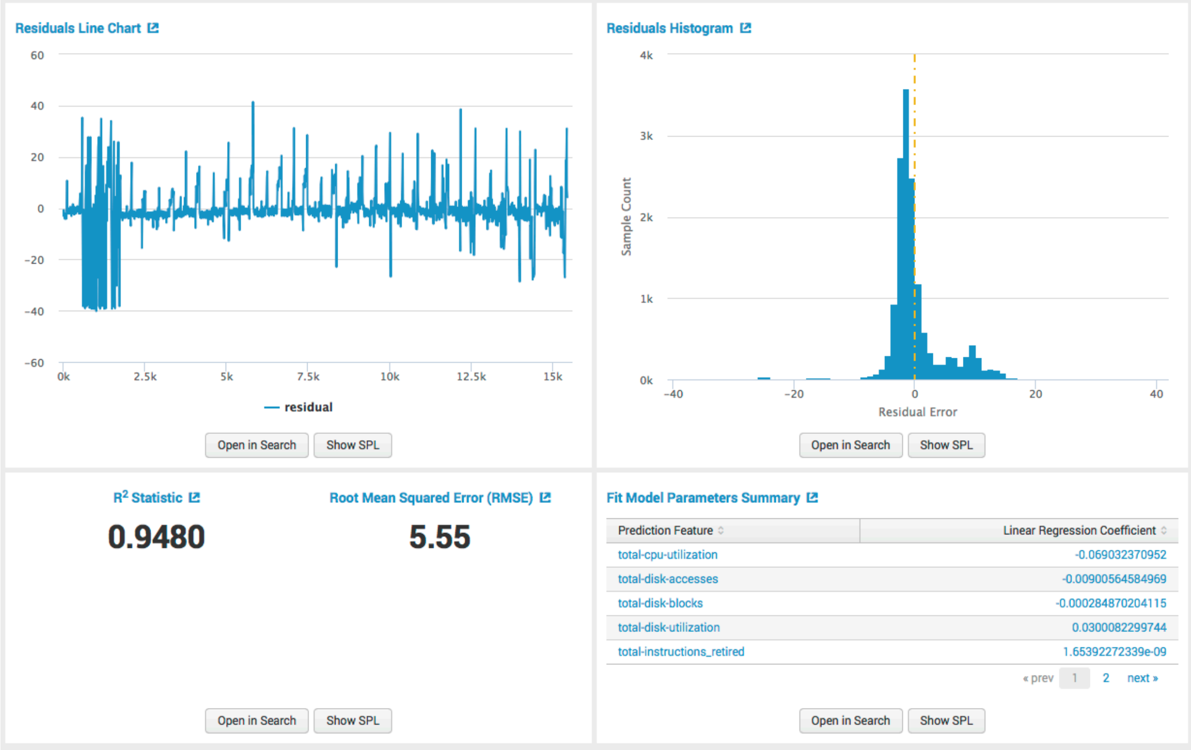
2: Field to predict: ac_power

3: Fields to use for predicting: total-cpu-utilization, total-disk-accesses, total-disk-blocks, total-disk-utilization, total-instructions_retired

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

5: Save the model as: example_server_power

Assistants: Validate



R^2 Statistic

0.9480

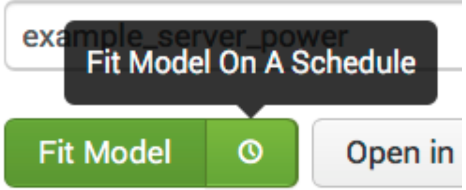
Root Mean Squared Error (RMSE)

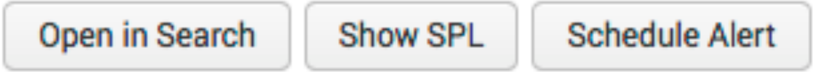
5.55

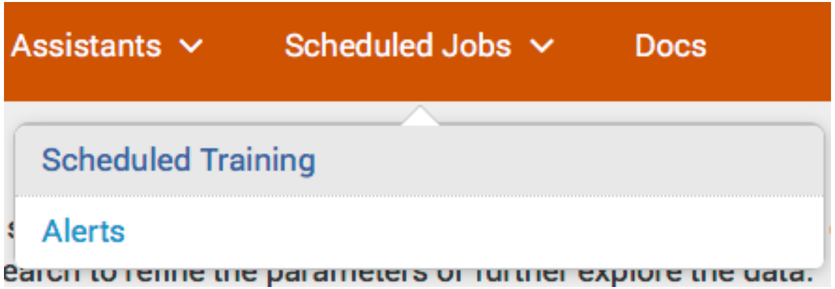
Fit Model Parameters Summary

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

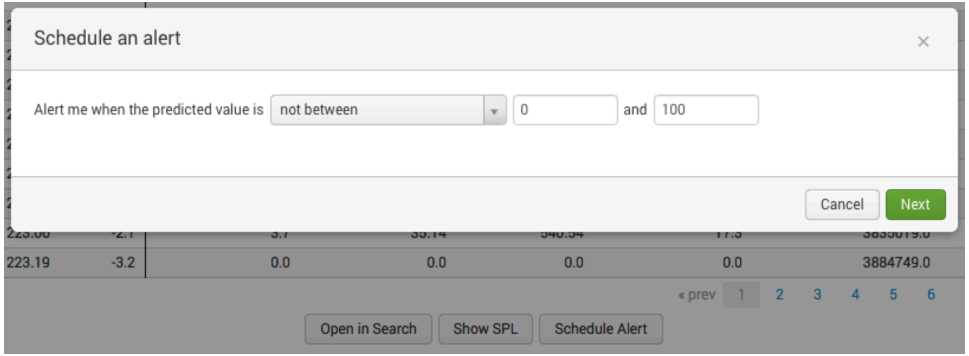
Assistants: Deploy







earch to refine the parameters or further explore the data.



Alert me when the predicted value is not between 0 and 100

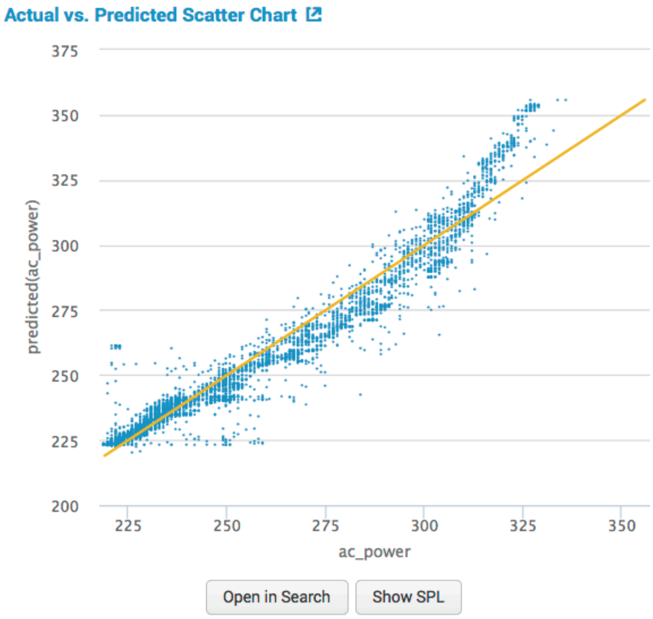
223.19	-3.2	0.0	0.0	0.0	0.0	3884749.0

The Assistants

1. Predict Numeric Fields
2. Predict Categorical Fields
3. Detect Numeric Outliers
4. Detect Categorical Outliers
5. Forecast Time Series
6. Cluster Numeric Events

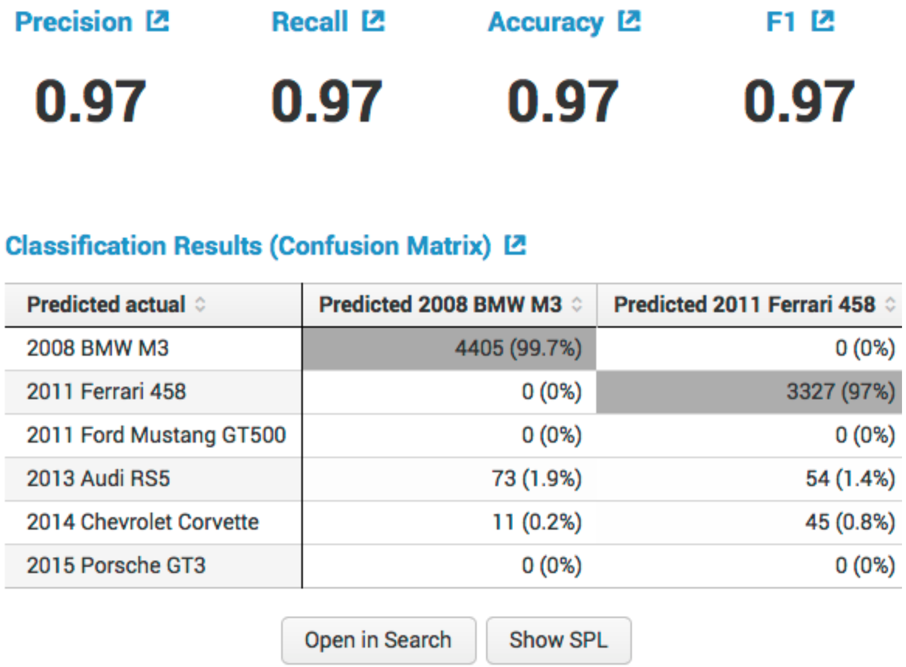
Predict Numeric Fields

- Algorithms
  - LinearRegression
    - … including Lasso, Ridge, and ElasticNet
  - KernelRidge
  - DecisionTreeRegressor
  - RandomForestRegressor
  - SGDRegressor
- Validation
  - Four visualizations of prediction error
  - R^2 and RMSE



Predict Categorical Fields

- Algorithms
  - LogisticRegression
  - DecisionTreeClassifier
  - RandomForestClassifier
  - SGDClassifier
  - SVM
  - Naïve Bayes
    - BernoulliNB and GuassianNB
- Validation
  - Precision, recall, accuracy, F1
  - Confusion matrix



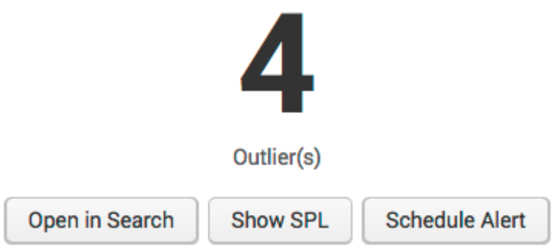
Precision	Recall	Accuracy	F1
0.97	0.97	0.97	0.97

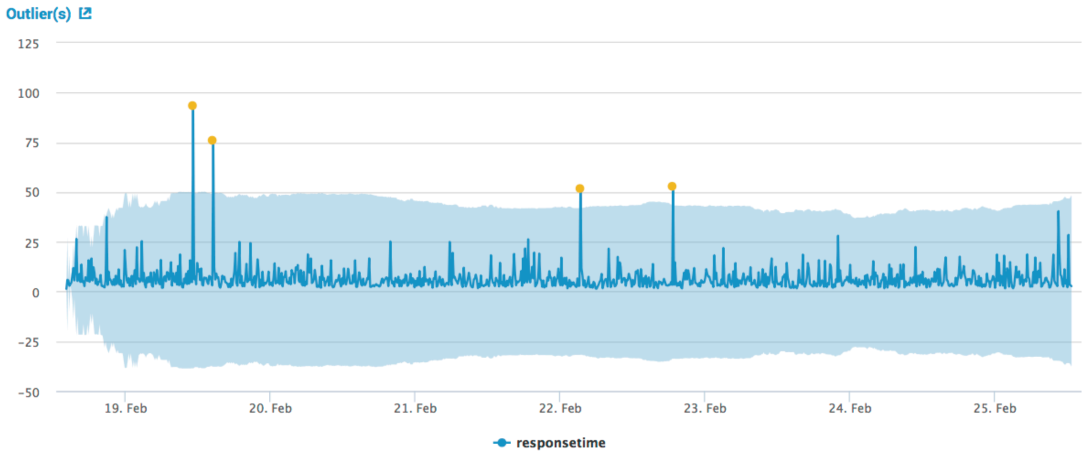
Classification Results (Confusion Matrix)

Predicted actual	Predicted 2008 BMW M3	Predicted 2011 Ferrari 458
2008 BMW M3	4405 (99.7%)	0 (0%)
2011 Ferrari 458	0 (0%)	3327 (97%)
2011 Ford Mustang GT500	0 (0%)	0 (0%)
2013 Audi RS5	73 (1.9%)	54 (1.4%)
2014 Chevrolet Corvette	11 (0.2%)	45 (0.8%)
2015 Porsche GT3	0 (0%)	0 (0%)

Detect Numeric Outliers

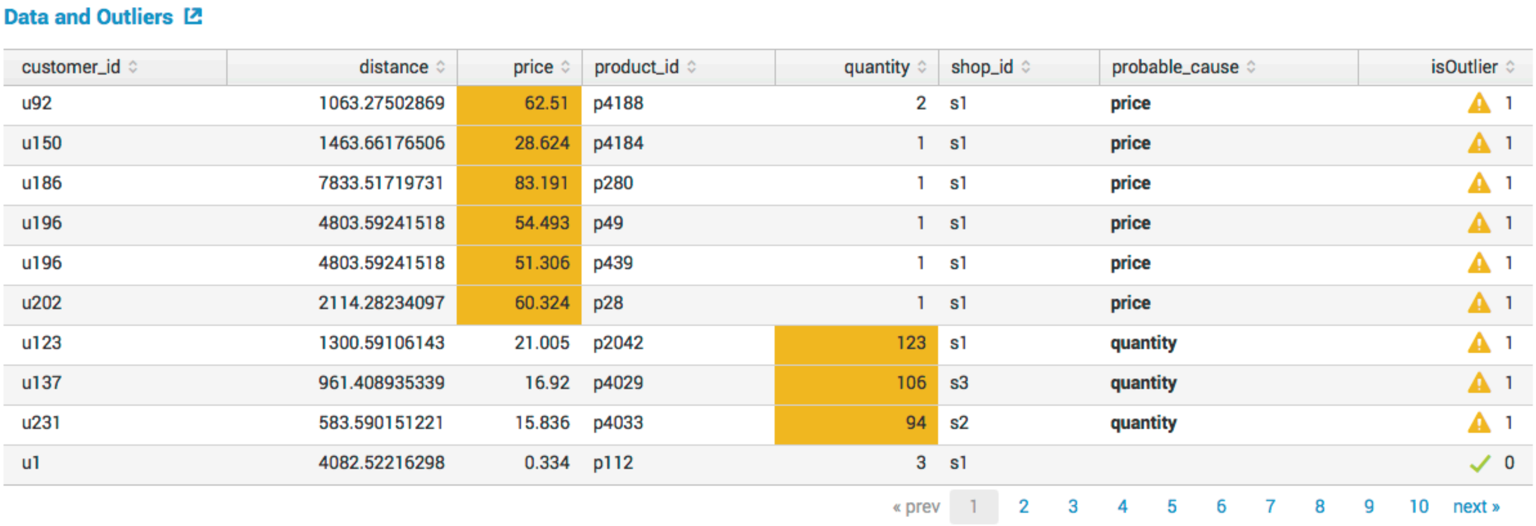
- Methods
  - Standard deviation
  - Median absolute deviation
  - Interquartile range
- Validation:





Detect Categorical Outliers

- Statistical methods
- Validation:



Data and Outliers

customer_id	distance	price	product_id	quantity	shop_id	probable_cause	isOutlier
u92	1063.27502869	62.51	p4188	2	s1	**price**	⚠ 1
u150	1463.66176506	28.624	p4184	1	s1	**price**	⚠ 1
u186	7833.51719731	83.191	p280	1	s1	**price**	⚠ 1
u196	4803.59241518	54.493	p49	1	s1	**price**	⚠ 1
u196	4803.59241518	51.306	p439	1	s1	**price**	⚠ 1
u202	2114.28234097	60.324	p28	1	s1	**price**	⚠ 1
u123	1300.59106143	21.005	p2042	123	s1	**quantity**	⚠ 1
u137	961.408935339	16.92	p4029	106	s3	**quantity**	⚠ 1
u231	583.590151221	15.836	p4033	94	s2	**quantity**	⚠ 1
u1	4082.52216298	0.334	p112	3	s1		✓ 0

Forecast Time Series

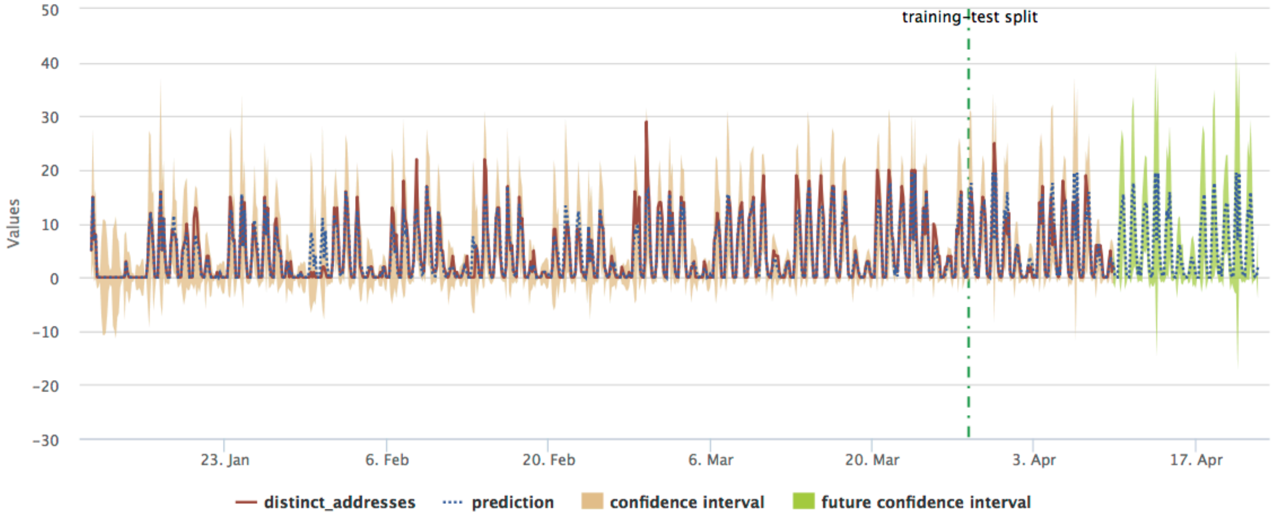
- Algorithms
  - State-space method using Kalman filter
  - ARIMA
- Validation

R^2 Statistic

0.8575

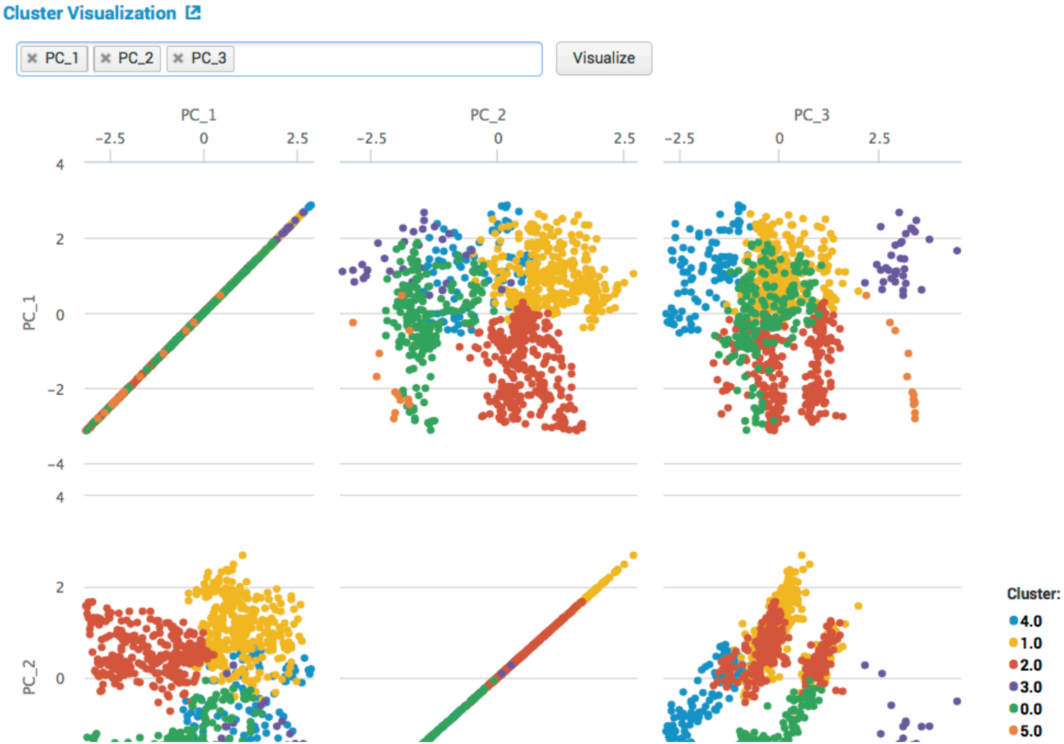
Root Mean Squared Error (RMSE)

2.10



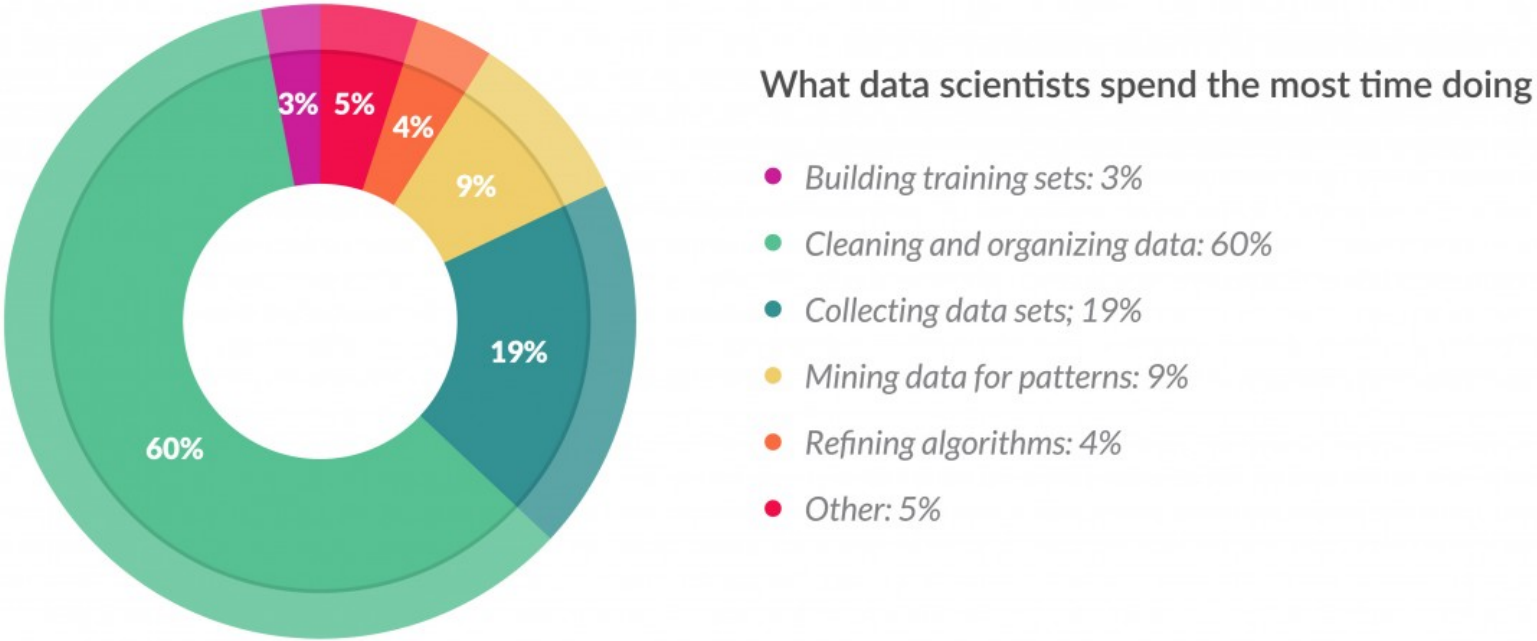
Cluster Numeric Events

- Algorithms
  - KMeans
  - DBSCAN
  - Birch
  - SpectralClustering
- Validation
  - Scatterplot Matrix viz



Prepare

Data Gathering and Prep



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%*

Source: CrowdFlower

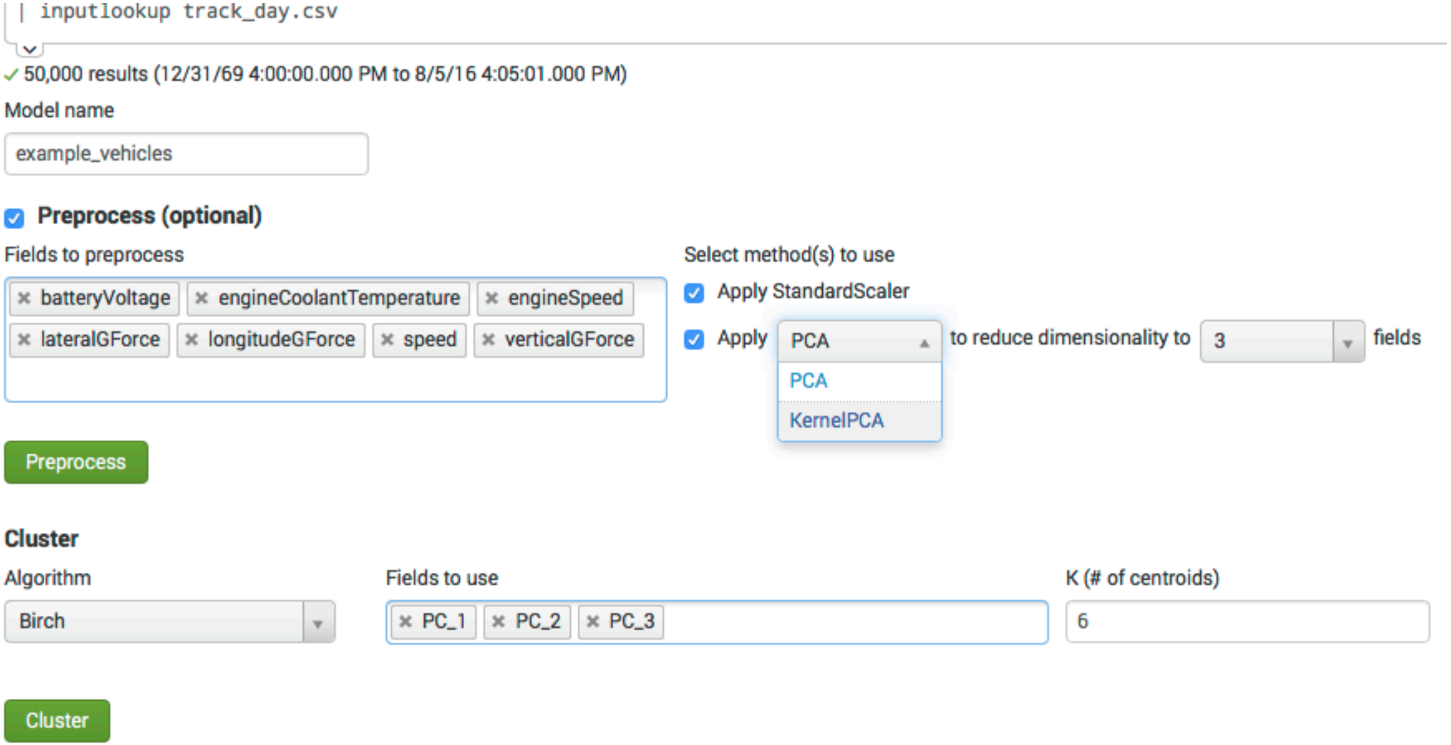
Splunk!

- Leading platform for collecting, cleaning, and transforming data
- Interactive Field Extractor
- Datamodels
- Hundreds of add-ons from Splunkbase
- transforms.conf
- props.conf
- etc.

Feature Engineering

- TFIDF (term-frequency x inverse document-frequency)
  - Transform free-form text into numeric attributes
- StandardScaler (i.e. normalization)
- FieldSelector (i.e. choose k best features for regression/classification)
- PCA and KernelPCA

Preprocessing in the Assistants



```
| inputlookup track_day.csv
```

✓ 50,000 results (12/31/69 4:00:00.000 PM to 8/5/16 4:05:01.000 PM)

Model name: example_vehicles

☑ **Preprocess (optional)**

Fields to preprocess: × batteryVoltage × engineCoolantTemperature × engineSpeed × lateralGForce × longitudeGForce × speed × verticalGForce

Select method(s) to use

☑ Apply StandardScaler

☑ Apply PCA to reduce dimensionality to 3 fields

Cluster

Algorithm: Birch

Fields to use: × PC_1 × PC_2 × PC_3

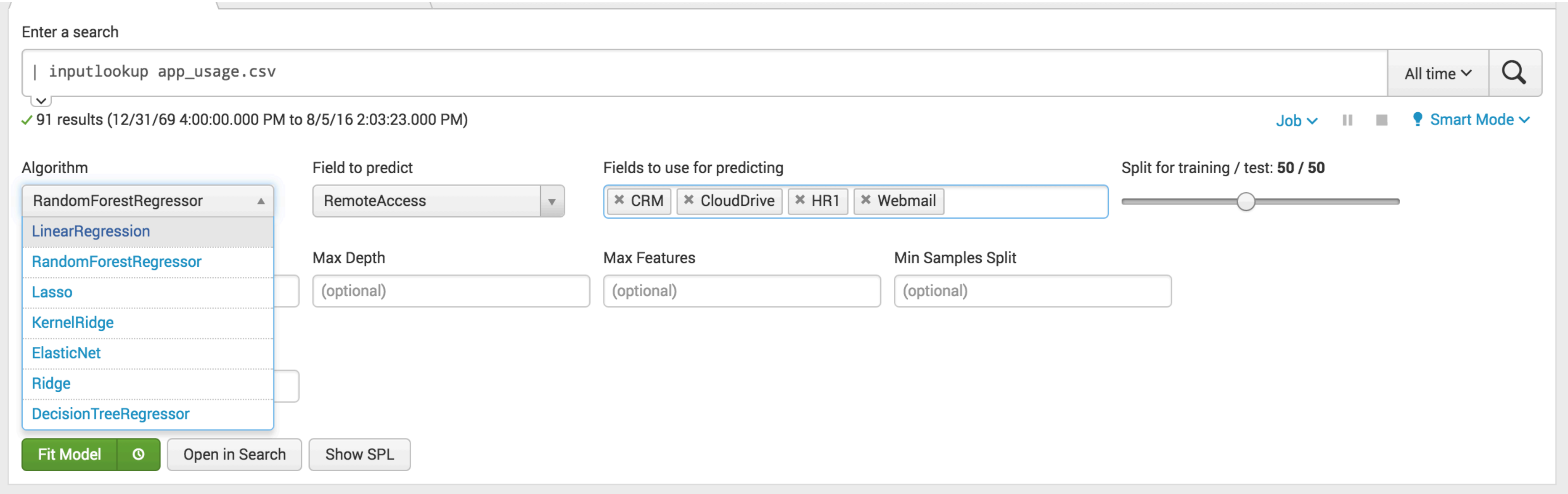
K (# of centroids): 6

Fit

Fit: What’s New

- No event limits
- Configurable resource caps (ml-spl.conf)
- Search head clustering support
- Scheduled fit
- New algorithms

Fit: What's New



Enter a search

```
| inputlookup app_usage.csv
```

✓ 91 results (12/31/69 4:00:00.000 PM to 8/5/16 2:03:23.000 PM)

Algorithm: RandomForestRegressor

Field to predict: RemoteAccess

Fields to use for predicting: CRM, CloudDrive, HR1, Webmail

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

Max Depth: (optional)

Max Features: (optional)

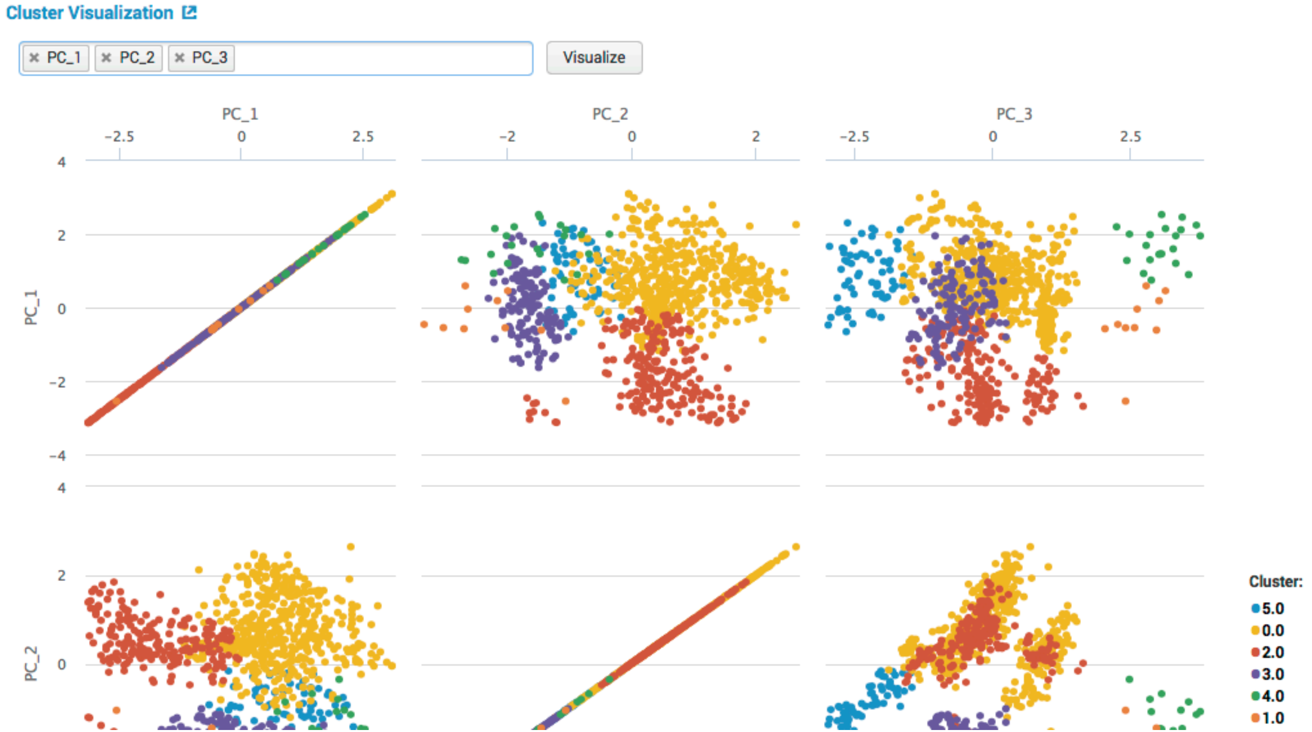
Min Samples Split: (optional)

Validate

Validate / Apply: What’s New

- Configurable resource caps
- Search head clustering support
- Distributed / streaming apply
- Scatterplot matrix viz

Scatterplot Matrix Viz



Deploy

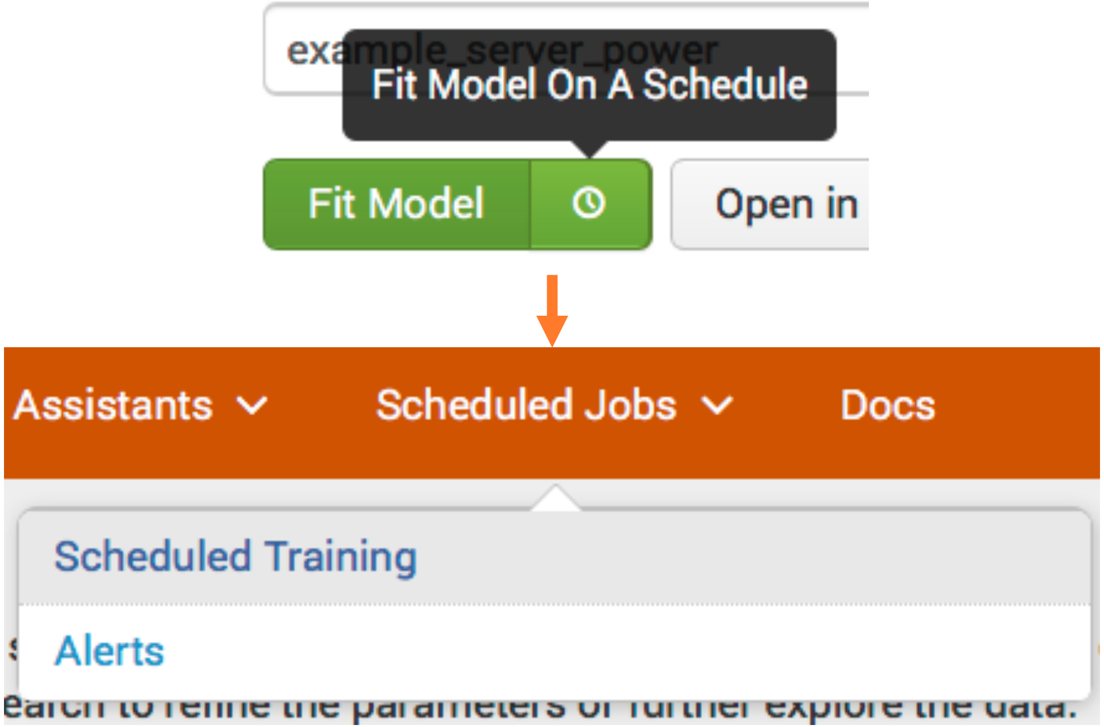
Deploy anywhere in Splunk!

- Scheduled training
- Alerting
- Reports and dashboards
- Augmented search results
- etc.

Deploy: What's New

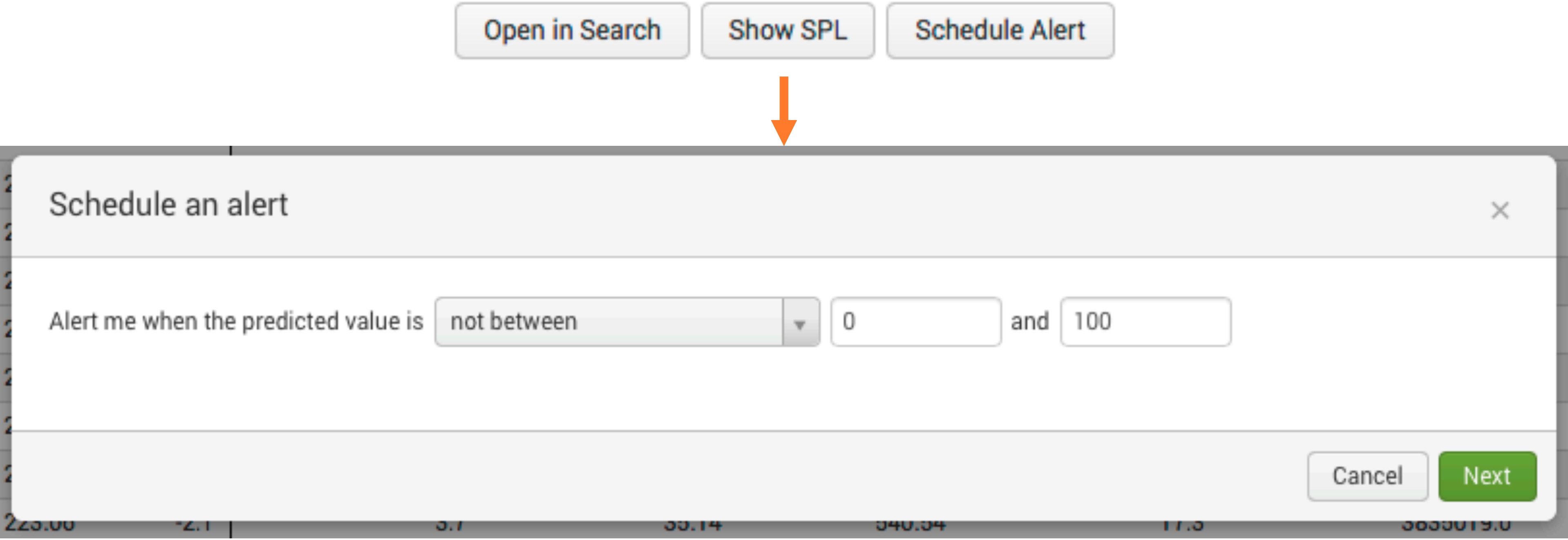
- Distributed Apply
  - Apply models to indexed data
  - Streaming
- Scheduled training
- Alerting

What’s New: Scheduled Fit



earch to refine the parameters or further explore the data.

What's New: Alerting



Schedule an alert

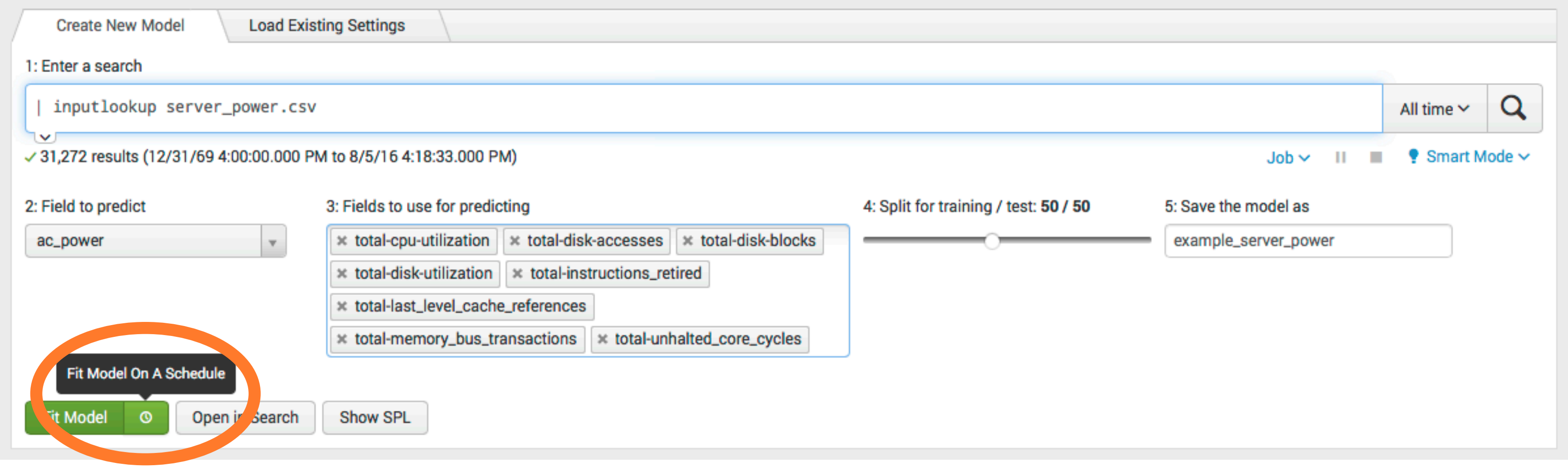
Alert me when the predicted value is not between 0 and 100

Example:
DIY Anomaly Detector

Let’s Build an Anomaly Detector!

- We’ll use two Assistants
  - Predict Numeric Fields
  - Detect Numeric Outliers
- Show automatically-generated intermediate SPL

Fit a Predictive Model



1: Enter a search

```
| inputlookup server_power.csv
```

✓ 31,272 results (12/31/69 4:00:00.000 PM to 8/5/16 4:18:33.000 PM)

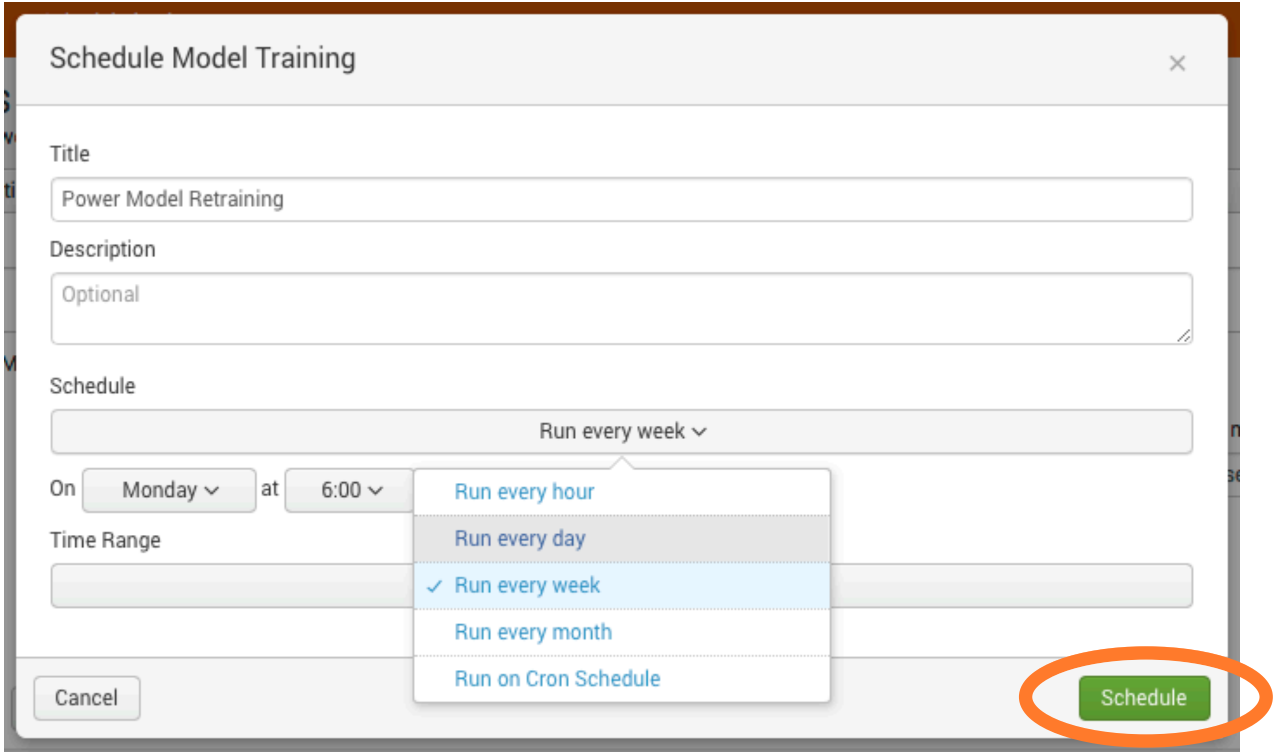
2: Field to predict: ac_power

3: Fields to use for predicting: 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

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

5: Save the model as: example_server_power

Set up Scheduled Training



Schedule Model Training

Title: Power Model Retraining

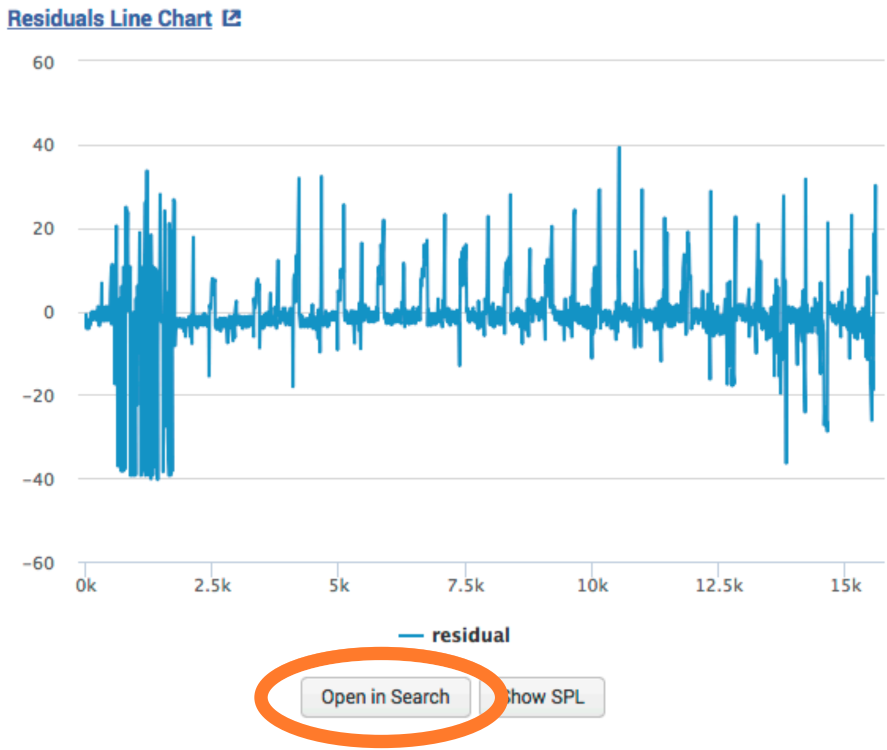
Description: Optional

Schedule: Run every week

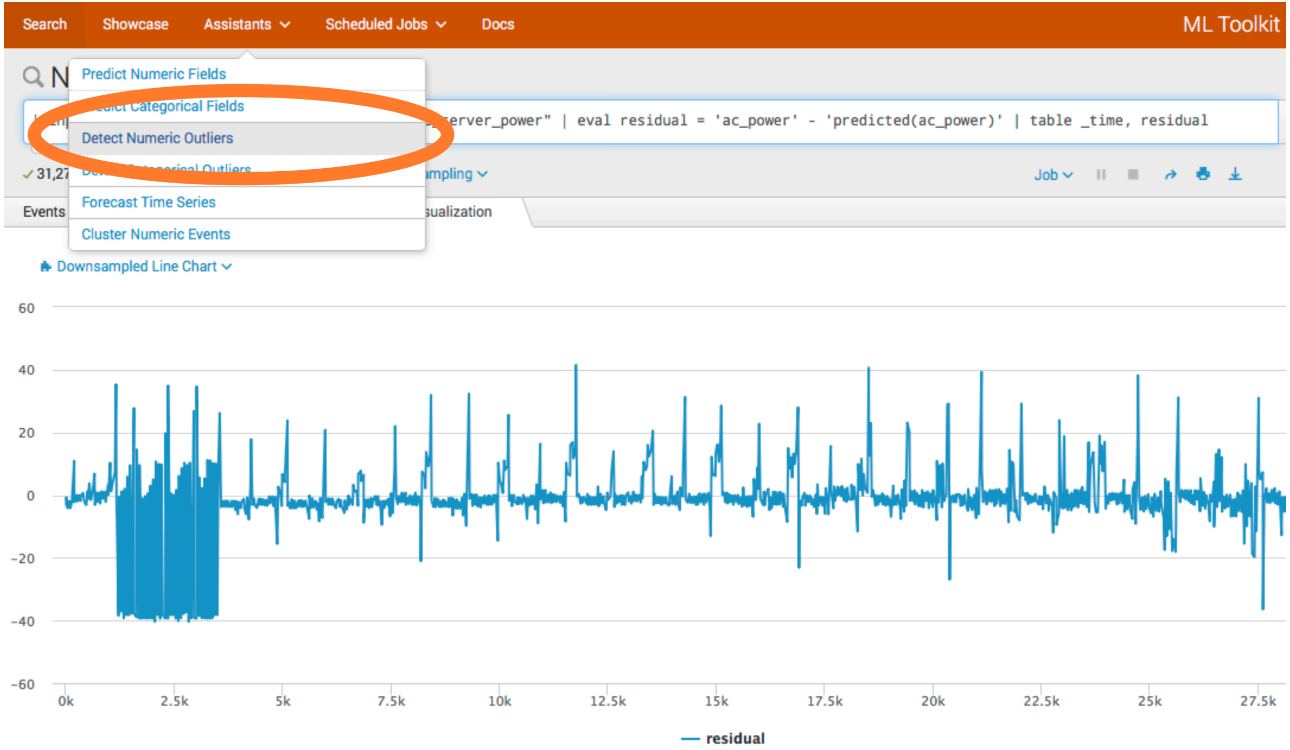
On Monday at 6:00

Time Range

Open Residuals in Search

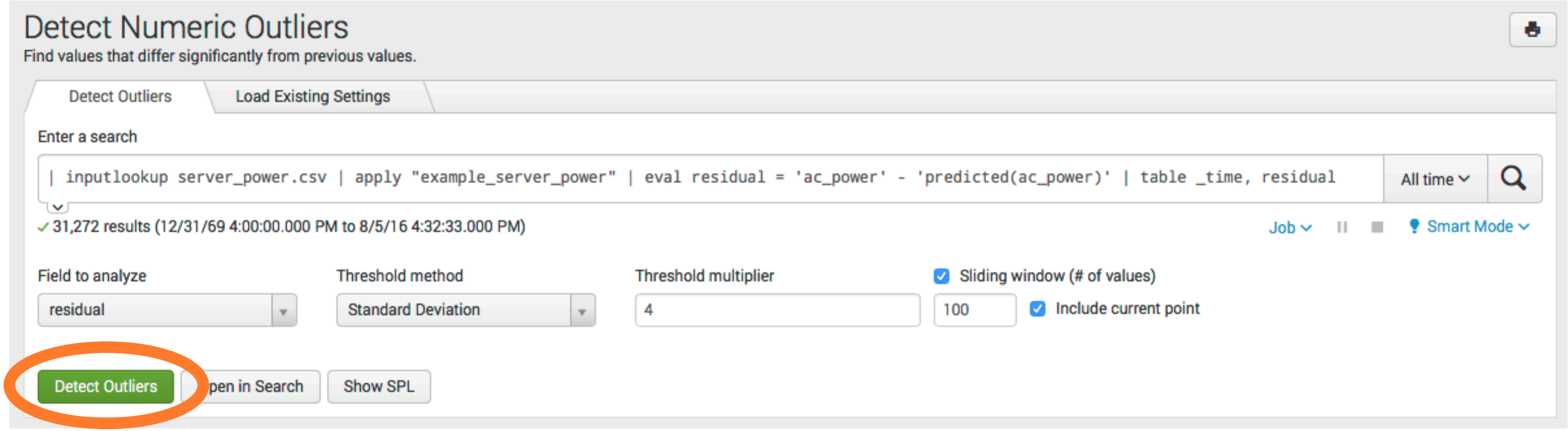


Open Detect Numeric Outliers Assistant



```
erver_power" | eval residual = 'ac_power' - 'predicted(ac_power)' | table _time, residual
```

Detect Outliers (Large Prediction Errors)



Detect Numeric Outliers

Find values that differ significantly from previous values.

Enter a search

```
| inputlookup server_power.csv | apply "example_server_power" | eval residual = 'ac_power' - 'predicted(ac_power)' | table _time, residual
```

✓ 31,272 results (12/31/69 4:00:00.000 PM to 8/5/16 4:32:33.000 PM)

Field to analyze: residual

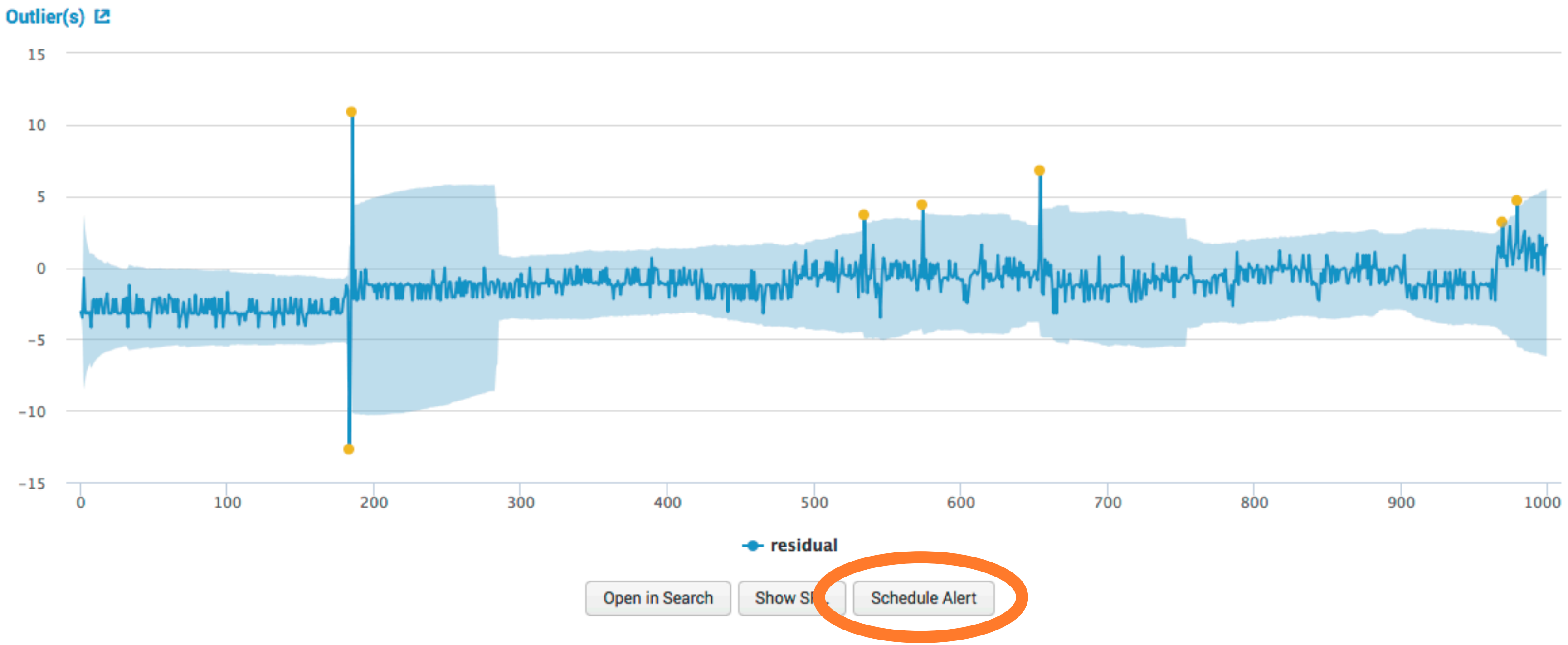
Threshold method: Standard Deviation

Threshold multiplier: 4

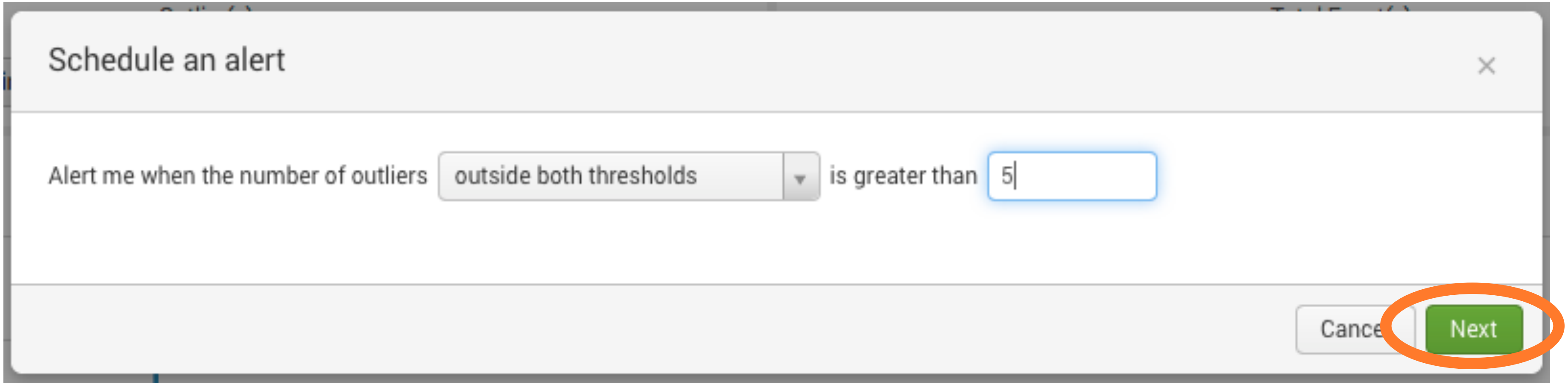
☑ Sliding window (# of values): 100

☑ Include current point

Schedule an Alert



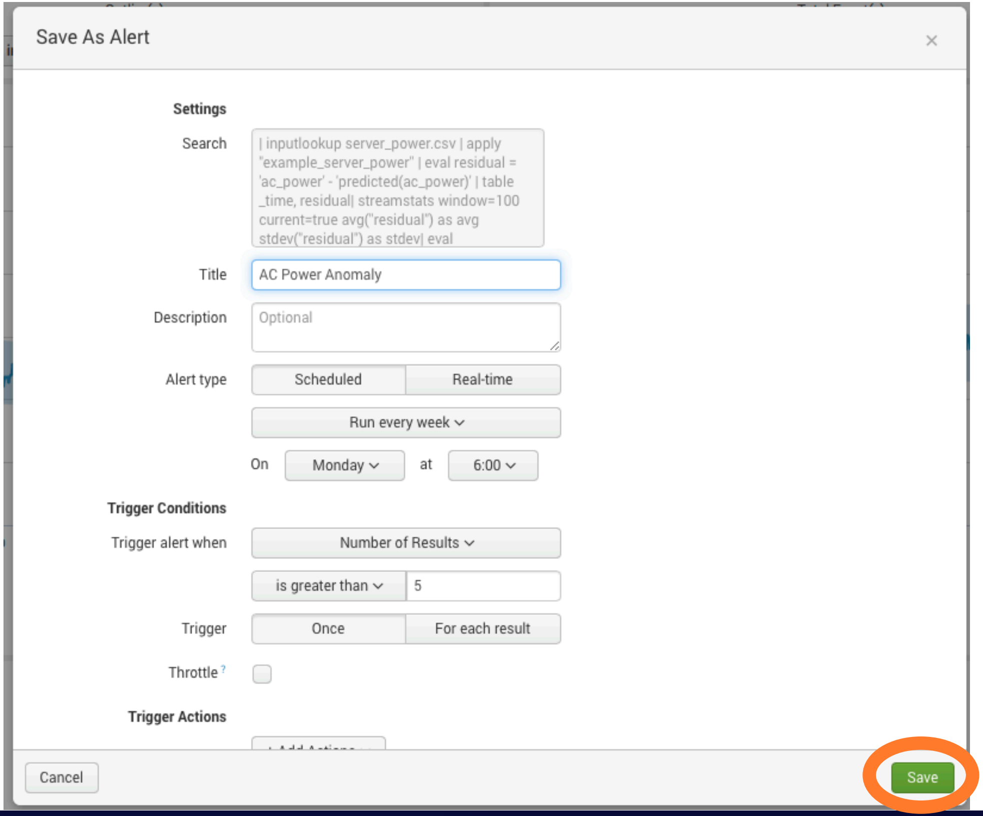
Schedule an Alert



Schedule an alert

Alert me when the number of outliers outside both thresholds is greater than 5

Schedule an Alert



Save As Alert

Settings

Search: | inputlookup server_power.csv | apply "example_server_power" | eval residual = 'ac_power' - 'predicted(ac_power)' | table _time, residual| streamstats window=100 current=true avg("residual") as avg stdev("residual") as stdev| eval

Title: AC Power Anomaly

Description: Optional

Alert type: Scheduled | Real-time

Run every week

On Monday at 6:00

Trigger Conditions

Trigger alert when: Number of Results

is greater than 5

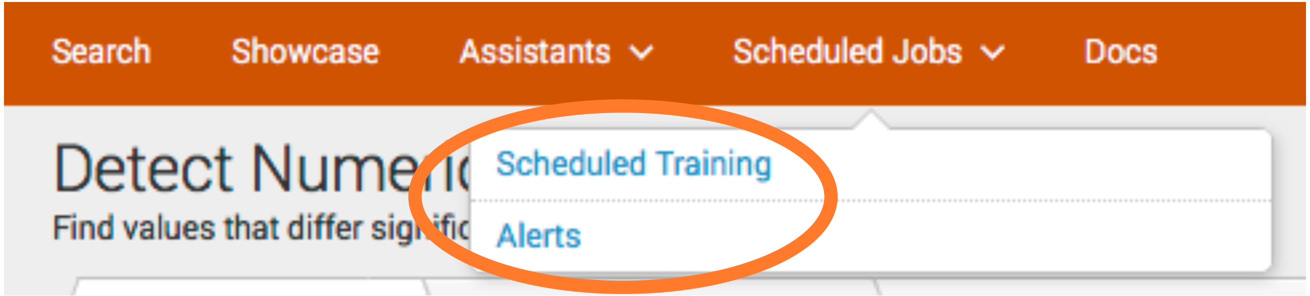
Trigger: Once | For each result

Throttle ? □

Trigger Actions

+ Add Actions

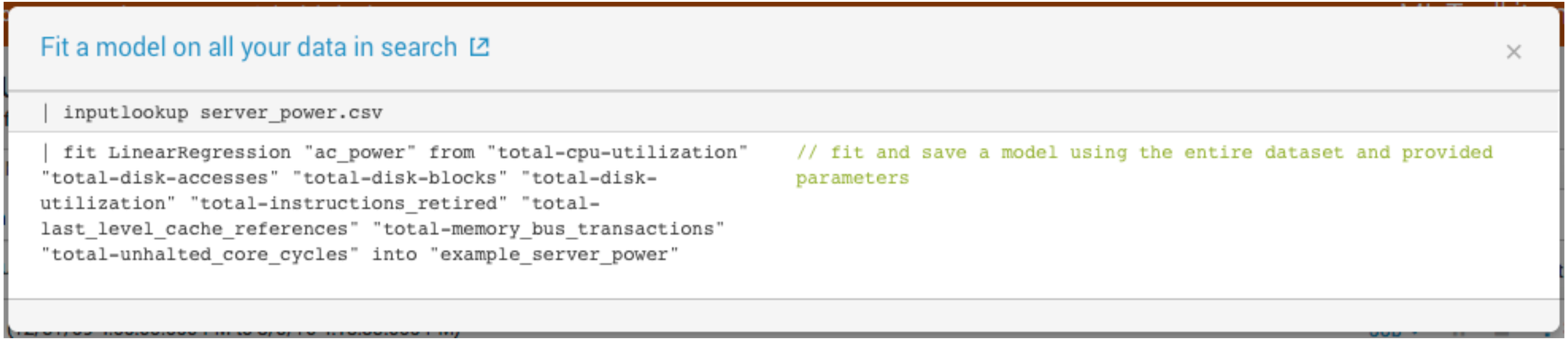
Manage Your New Anomaly Detector



Detect Numeric

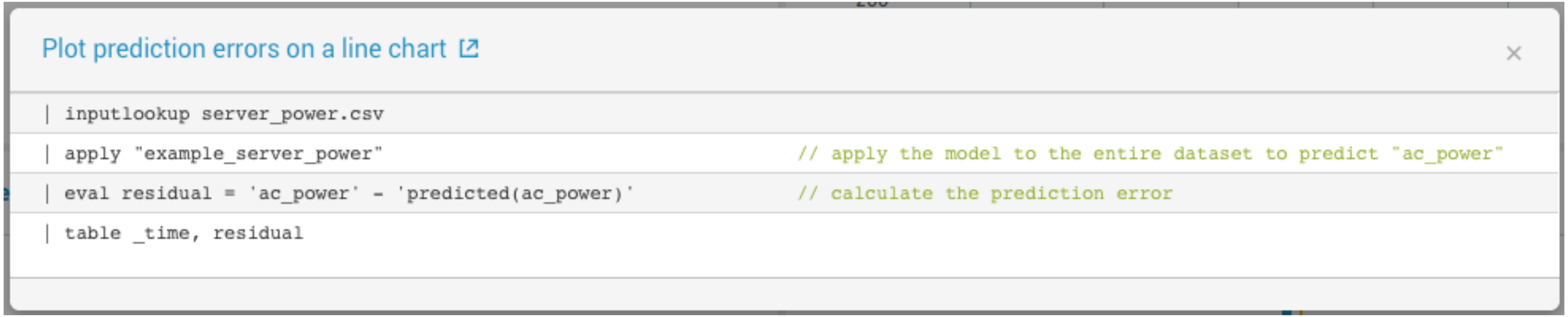
Find values that differ signific

The Assistant Generated the SPL for You



Fit a model on all your data in search

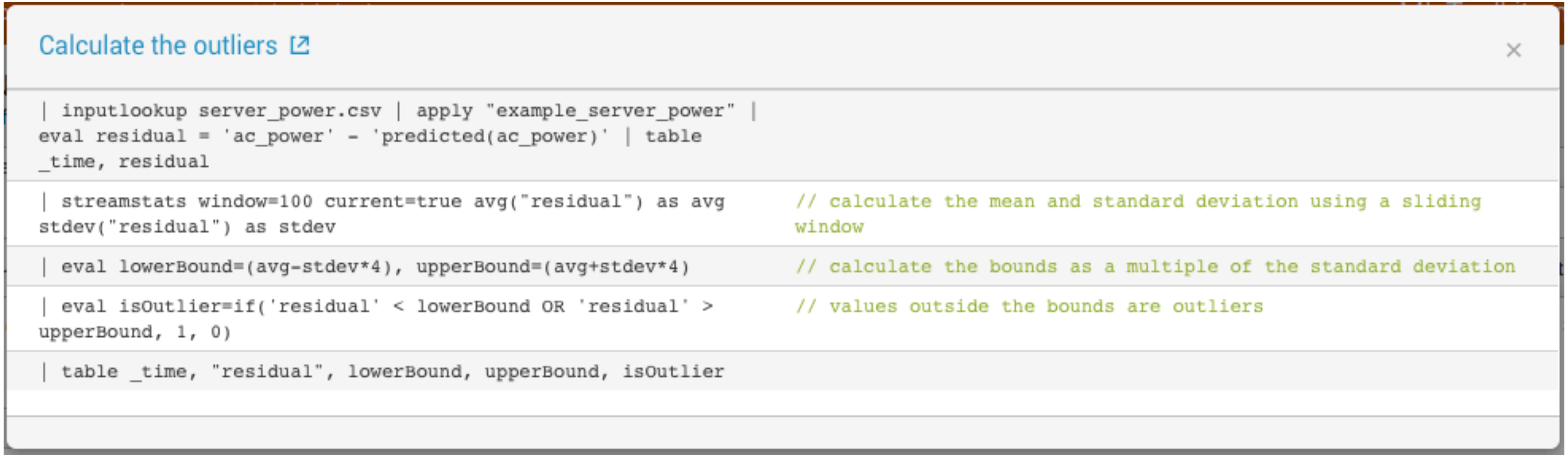
```
| inputlookup server_power.csv
| fit LinearRegression "ac_power" from "total-cpu-utilization"      // fit and save a model using the entire dataset and provided
"total-disk-accesses" "total-disk-blocks" "total-disk-               parameters
utilization" "total-instructions_retired" "total-
last_level_cache_references" "total-memory_bus_transactions"
"total-unhalted_core_cycles" into "example_server_power"
```



Plot prediction errors on a line chart

```
| inputlookup server_power.csv
| apply "example_server_power"                          // apply the model to the entire dataset to predict "ac_power"
| eval residual = 'ac_power' - 'predicted(ac_power)'    // calculate the prediction error
| table _time, residual
```

The Assistant Generated the SPL for You



Calculate the outliers

```
| inputlookup server_power.csv | apply "example_server_power" |
eval residual = 'ac_power' - 'predicted(ac_power)' | table
_time, residual

| streamstats window=100 current=true avg("residual") as avg      // calculate the mean and standard deviation using a sliding
stdev("residual") as stdev                                          window

| eval lowerBound=(avg-stdev*4), upperBound=(avg+stdev*4)          // calculate the bounds as a multiple of the standard deviation

| eval isOutlier=if('residual' < lowerBound OR 'residual' >        // values outside the bounds are outliers
upperBound, 1, 0)

| table _time, "residual", lowerBound, upperBound, isOutlier
```

You Built an Anomaly Detector!

- You built a predictive model of AC Power
- When the prediction error from this model is an outlier compared to past errors, you generate an alert
- This predictive model automatically retrains itself on a schedule you control
- You didn’t have to type any SPL

#winning

.conf2016

splunk>

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

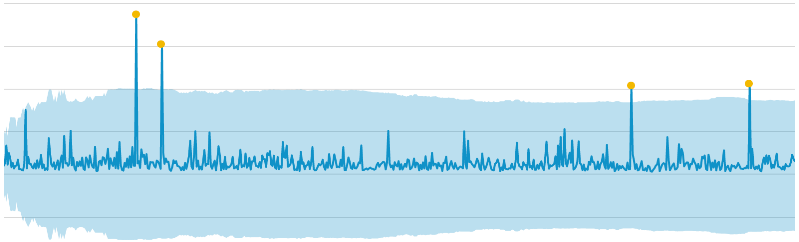
Optimizing operations and business results

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

Detect Network Outliers

Reduced downtime + increased service availability = better customer satisfaction



ML Use Case

Monitor noise rise for 20,000+ cell towers to increase service and device availability, reduce MTTR

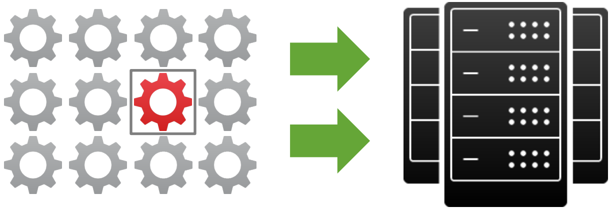
Technical overview

- A customized solution deployed in production based on outlier detection.
- Leverage previous month data and voting algorithms

“The ability to model complex systems and alert on deviations is where IT and security operations are headed … Splunk Machine Learning has given us a head start...”

Reliable website updates

Proactive website monitoring leads to reduced downtime



ML Use Case

- Very frequent code and config updates (1000+ daily) can cause site issues
- Find errors in server pools, then prioritize actions and predict root cause

Technical overview

- Custom outlier detection built using ML Toolkit Outlier assistant
- Built by Splunk Architect with no Data Science background

"Splunk ML helps us rapidly improve end-user experience by ranking issue severity which helps us determine root causes faster thus reducing MTTR and improving SLA"

What Now?

http://tiny.cc/splunkmlapp

- Get the Machine Learning Toolkit from Splunkbase
- Go watch Machine Learning Videos on Splunk Youtube Channel http://tiny.cc/splunkmlvideos
- Go to Machine Learnings talks:
  - Advanced Machine Learning in SPL with the Machine Learning Toolkit by Jacob Leverich
  - Extending SPL with Custom Search Commands and the Splunk SDK for Python by Jacob Leverich
- Several Customers and Partner Talks
  - Cisco, Scianta Analytics, Asian Telco, etc.
- Early Adopter And Customer Advisory Program : mlprogram@splunk.com
- Product Manager: Manish Sainani ms@splunk.com
- Field Expert: Andrew Stein astein@splunk.com

THANK YOU

.conf2016

splunk>