

.conf18

splunk>

Machine Learning & Natural Language Processing at BMW Group

October 2018



Our Speakers



BOULOS EL-ASMAR

Data Scientist, BMW Group IT
Innovation Lab



IMAN MAKAREMI

Principal Data Scientist, Splunk



DIPOCK DAS

Senior Director, Products, Splunk

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, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2018 Splunk Inc. All rights reserved.

Introduction

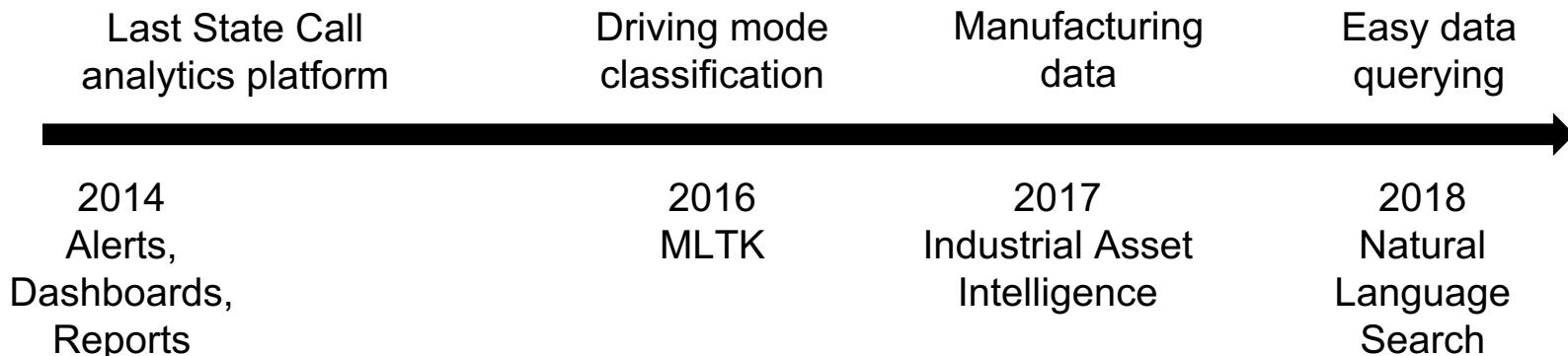


Agenda



1. The Goal of the Predictive Model
2. Procedure and Achievements in the Project
3. Implementing the Model with Splunk MLTK
4. Using the Model with Project Natural Language Search
5. Project achievements
6. Future Objectives

BMW Group and Splunk



What is BMW Group IT Innovation Lab?

- ▶ Focusing on innovative use cases inside BMW Group
 - ▶ Know-How in computer algorithms, data science and AI
 - ▶ Close networking with the innovation departments at specialists area
(Manufacturing, Logistics ...)
 - ▶ Early adoption of new technologies, transfer to industry



The Goal of the Predictive Model

Human-Centered Machine Learning

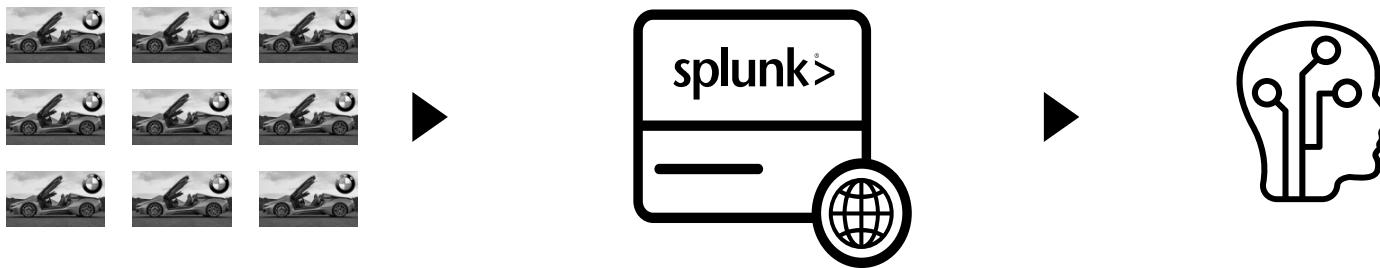


The goal of the Predictive Model

- ▶ Overarching concept for all road users
 - Predict trends of traffic patterns with the highest accuracy possible
 - Focus on challenges around urban mobility
 - Create action plans on making traffic in the future more efficient
- ▶ For the individual user
 - Prediction of ideal navigation and travel time
 - Allow to change route or timing to arrive at destination more efficiently

Procedure and Achievements

Data Collection



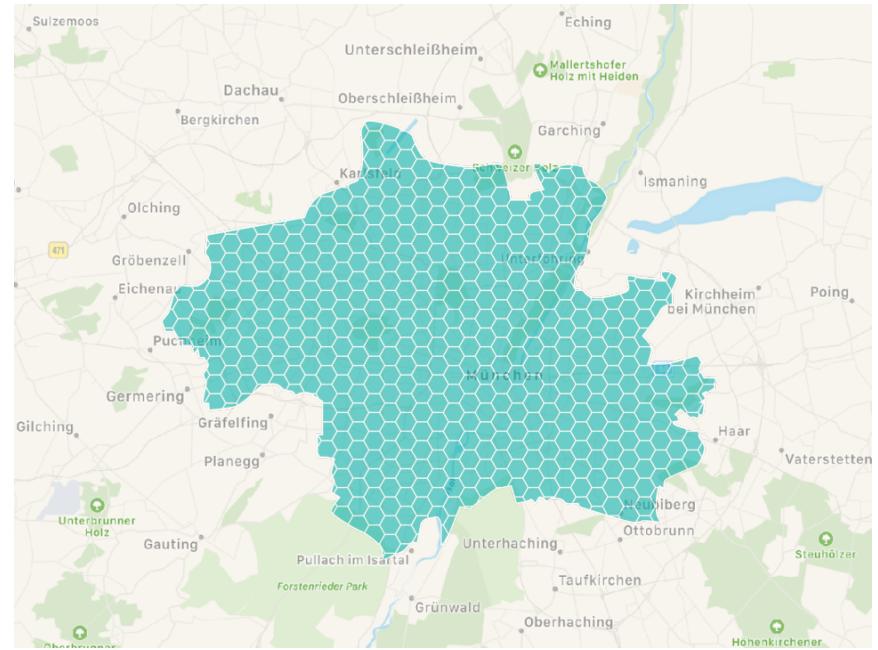
Data from in-development test fleet provides insights on commuting and mobility habits

Movements tracked over four weeks with data transmitted in real-time to Splunk

Historical data about the month was used as the basis to train the model

Organising Data

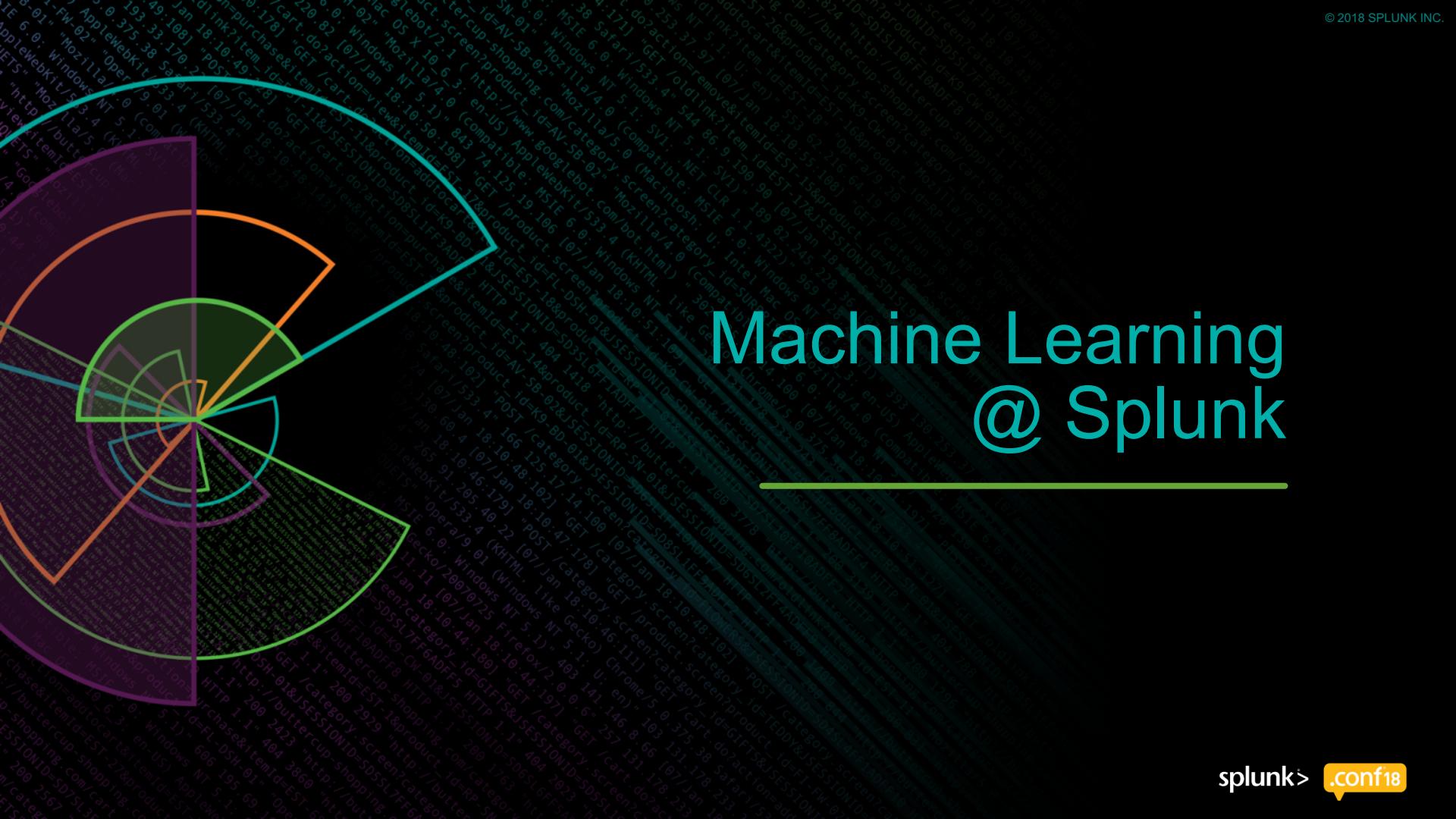
- ▶ City of Munich is divided into cells
 - ▶ Traffic data defined features, in each cell, gathered from the participants
 - User departure time in hourly intervals
 - User arrival time in hourly intervals
 - GPS of departure and destination points
 - Road types
 - Waiting times
 - Weather
 - Accidents



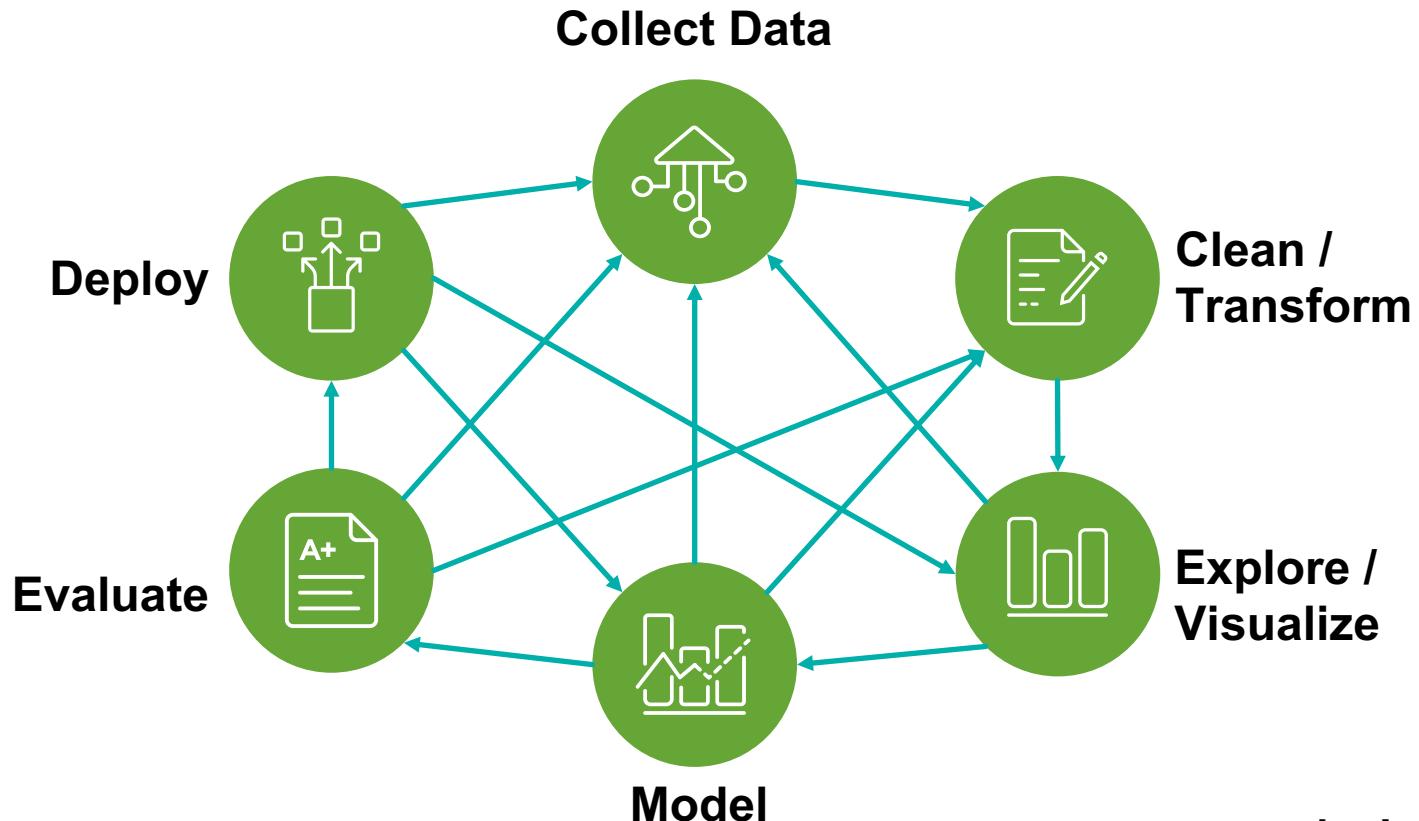
Build and Train a model

- ▶ Algorithm:
 - Non-linear Algorithm
 - Random Forest Regressor
- ▶ Through supervised learning, discrepancies were adjusted in order to enhance the accuracy of predictions
- ▶ A local search finds the most similar traffic state in the traffic history
- ▶ Blends real-time traffic data with past traffic patterns to predict congestion and other traffic events

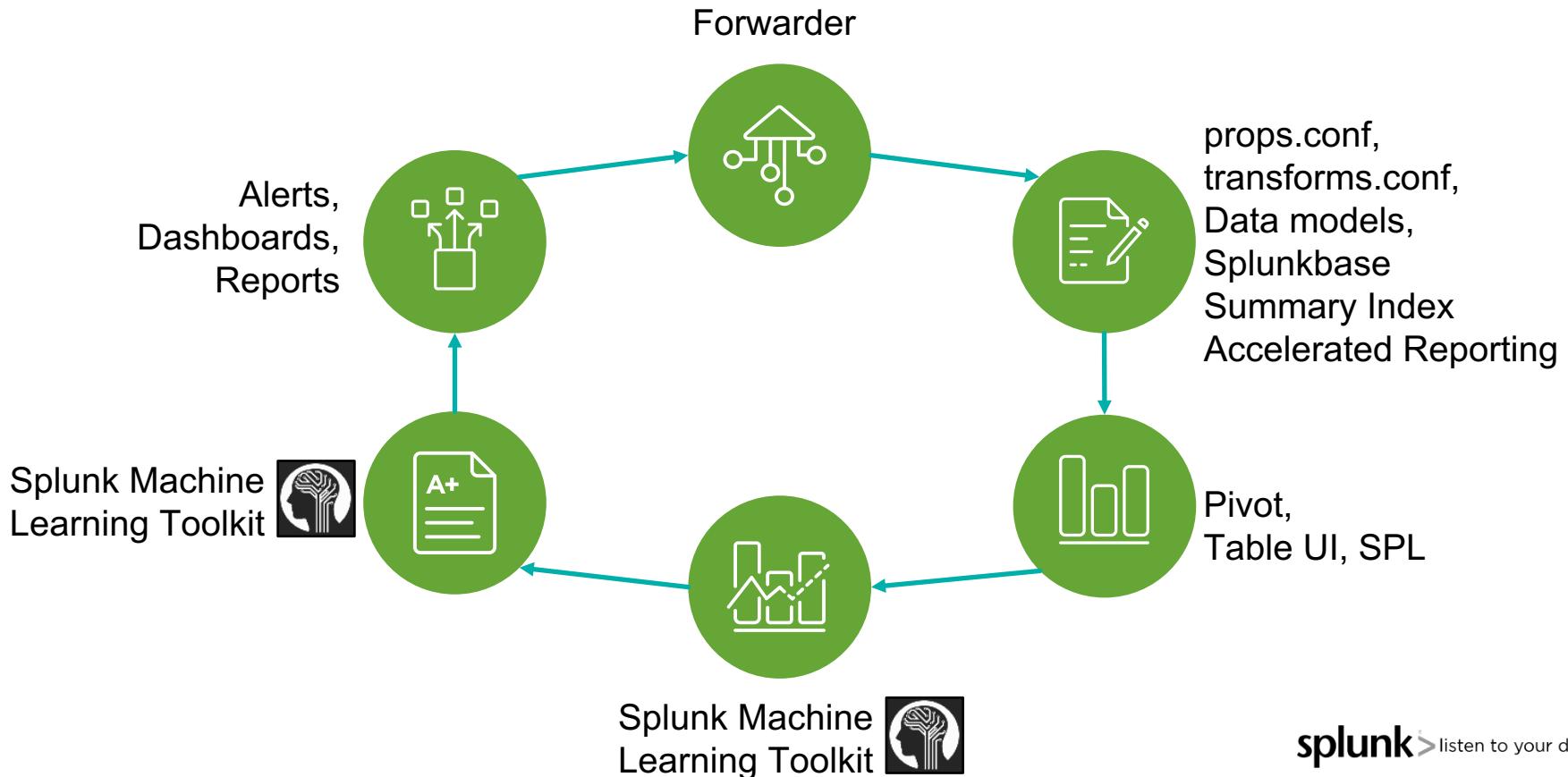
Machine Learning @ Splunk



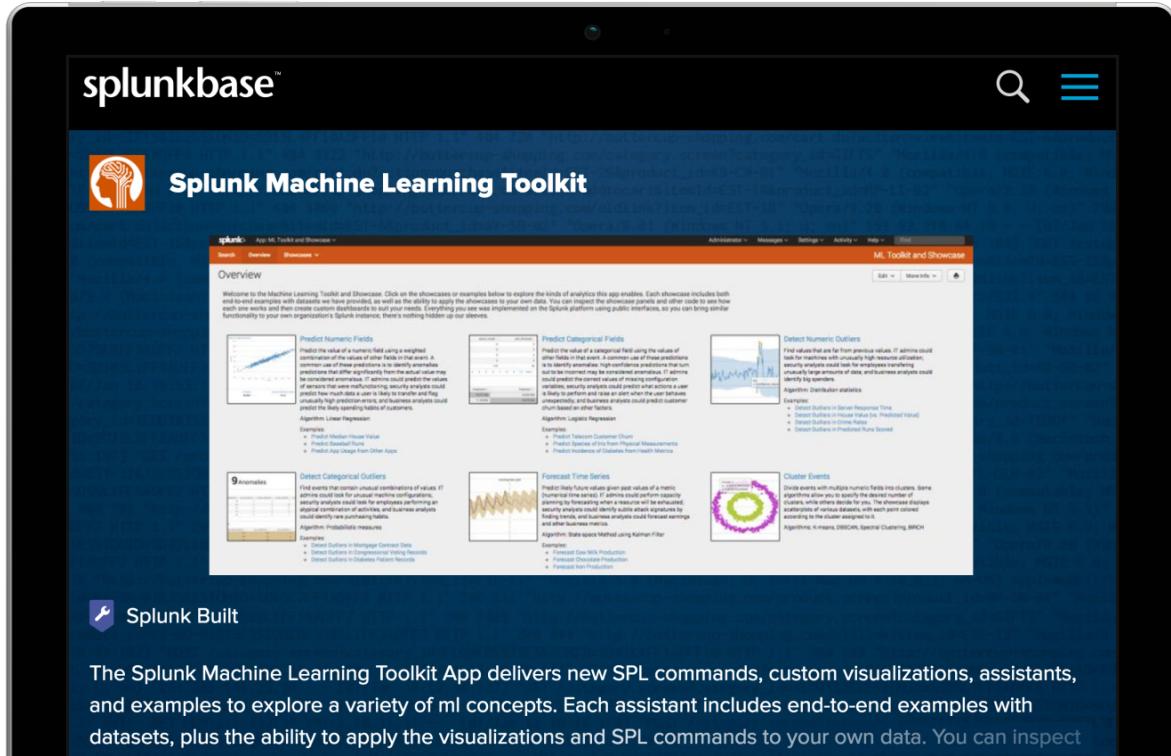
Machine Learning Process



Machine Learning Process with Splunk



Splunk Machine Learning Toolkit App



The screenshot shows the Splunkbase interface with the "Splunk Machine Learning Toolkit" app selected. The page has a dark blue header with the Splunkbase logo and a search bar. Below the header, there's a navigation bar with "Search", "Discover", and "Showcases". The main content area is titled "Overview" and contains several cards describing different machine learning concepts:

- Predict Numeric Fields:** Predict the value of a numeric field using a weighted combination of the values of other fields in the event. Examples include Predict Median House Value, Predict BestellFlame, and Predict Price Usage Duration Linear.
- Predict Categorical Fields:** Predict the value of a categorical field using the values of other fields in the event. A common use of these predictions is to identify anomalies in categorical fields. Examples include Detect Outliers in Service Response Time, Detect Outliers in House Value (inc. Predicted Value), and Detect Outliers in Predicted Punk Score.
- Detect Numeric Outliers:** Find values that are far from previous values. Administrators could look for anomalies with unusually high resource utilization, security analysts could look for anomalies with unusually large numbers of events, and business analysts could predict what actions a user might take if they see an unusually high prediction score. Examples include Detect Outliers in Service Response Time, Detect Outliers in House Value (inc. Predicted Value), and Detect Outliers in Predicted Punk Score.
- Detect Categorical Outliers:** Find events that contain unusual combinations of values. Administrators could look for unusual service configurations, security analysts could look for unusual logon patterns, and business analysts could look for unusual combinations of products and services. Examples include Detect Outliers in Mortgage Default Data, Detect Outliers in Congressional Voting Records, and Detect Outliers in Credit Card Fraud.
- Forecast Time Series:** Predict likely future values given past values of a metric. Administrators could use this to predict future load on a system, security analysts could identify subtle attack signatures by predicting future behavior, and business analysts could forecast revenue and other business metrics. Examples include Forecast Sales Method using Kalman Filter, Forecast Gov Mills Production, Forecast Gov Mills Consumption, and Forecast Gov Mills Prediction.
- Cluster Events:** Divide events with multiple numeric fields into clusters. Some algorithms allow you to specify the desired number of clusters, while others let you analyze the data to find the best clusterings. Examples include Cluster Events with k-means, DBSCAN, Spectre Clustering, and BRCH.

At the bottom left, there's a "Splunk Built" badge with a gear icon. At the bottom right, there's a "splunk> .conf18" logo.

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

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

Splunk ML Advisory Program

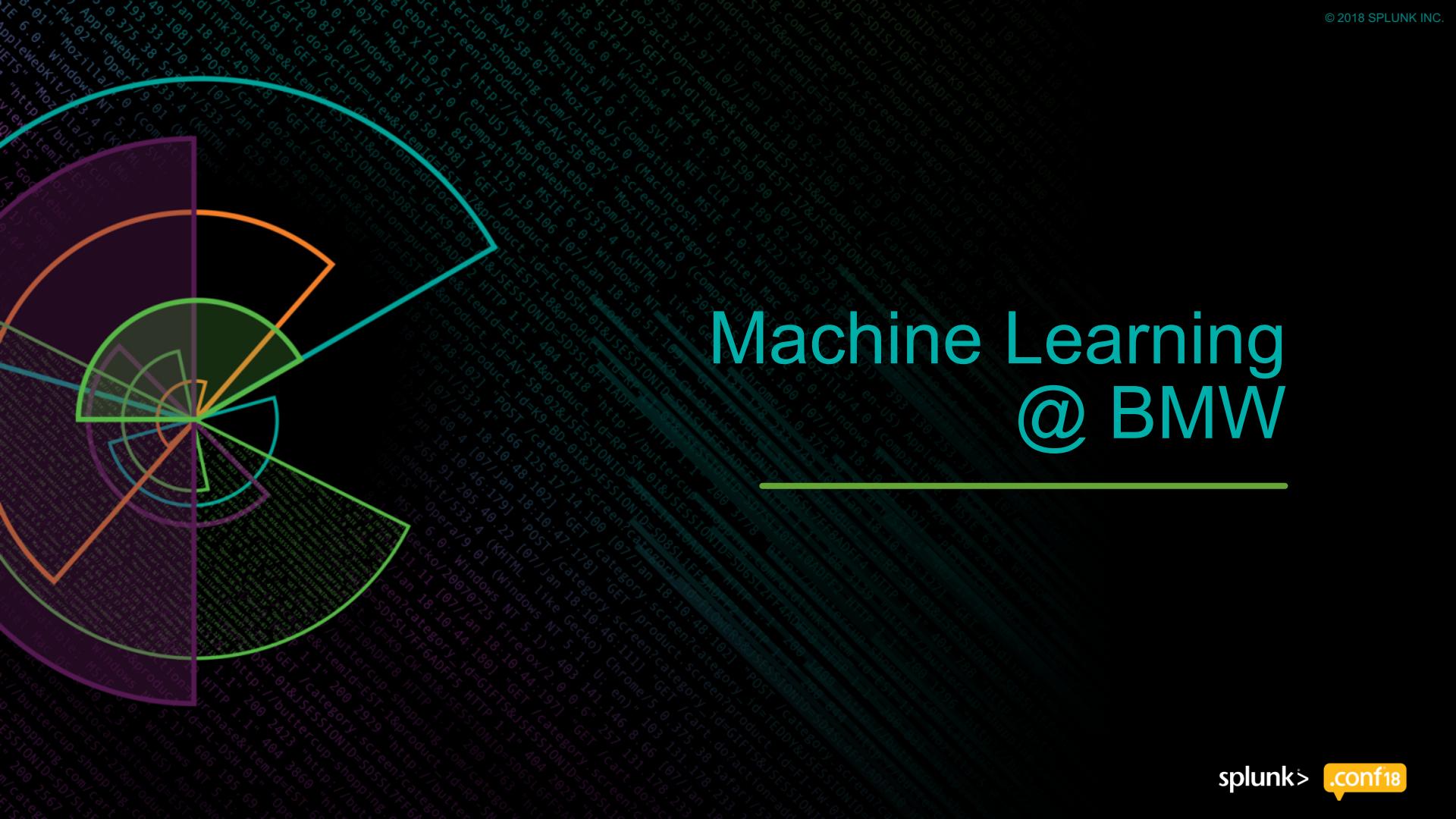
Partners a Splunk Data Science Resource to Help Operationalize an ML Use Case

The image shows a smartphone with a white background. On the screen, there is a list titled "Machine Learning Customer Advisory Program FAQs". The list consists of seven items, each with a question and a small circular icon containing a plus sign to its right. The questions are:

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

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

Machine Learning @ BMW



Raw Data

Selected Fields
a host 1
a source 1
a sourcetype 1

```
Interesting Fields  
a index 1  
# linecount 1  
a meta 1  
a punct 1  
a splunk_server 1  
a timestamp 1
```

+ Extract New Fields

> 24/07/2017 54373636393200000155BFACABA4@bmw.de,b5c55f4686455132d75960531e6f6abf,CS-CS-MUC-prod-2.1-44.2.42393.20131009.zip,2016-07-0
09:42:16.000 6 15:27:01.383,1.46781882138e+12,33669.0,2016-07-06 16:19:49.544,1.46782198954e+12,33707.0,2016-07-06 10:06:59.007,1.467799
61901e+12,2016-07-06 16:19:49.544,1.46782198954e+12,WEDNESDAY,12.0,18.0,3378792.0,54373636393200000155C10CF64A@bmw.de,Tru
e,False,CLASSIC,2016-07-06 12:06:59.007,1.46780681901e+12,2016-07-06 17:27:01.383,1.46782602138e+12,38.0,[{"phase":"RESERVA
TION","starttime":1467799619007,"endtime":1467818802175,"duration":19183168}, {"phase":"BOOKING_PREPARATION","star
ttimestamp":1467818802175,"endtime":1467818821383,"duration":19208}, {"phase":"DRIVING","starttime":1467818821383,
"endtime":1467821877895,"duration":3056512}, {"phase":"BOOKING_END","starttime":1467821877895,"endtime":14678
21989544,"duration":111649}], [{"clienttimestamp":1467799619007,"startstate":"AVAILABLE","endstate":"RESERVED","transition":
"ONLINE_RESERVATION_RECEIVED","location":{"lat":48.35201444444444,"lon":11.7877111111111111,"heading":null}, "locationhexgrid
":null}, {"clienttimestamp":1467818802175,"startstate":"RESERVED","endstate":"READY_FOR_BOOKING","transition":
"RECOGNIZED_DRIVER_RFID","location":{"lat":48.35201444444444,"lon":11.7877111111111111,"heading":null}, {"clienttim
estamp":1467818821383,"startstate":"READY_FOR_BOOKING","endstate":"BOOKED","transition":
"BOOKING_COMPLETED","location":{"lat":48.35201444444444,"lon":11.7877111111111111,"heading":null}, {"clienttimestamp":1467821877895,
"startstate":BOOKED,"endstate":BOOKING_END_CHECKS,"transition":BOOKING_ENDED_CHECKS_START,"location":
{"lat":48.14174333333333,"lon":11.5676111111111111,"heading":null}, {"clienttimestamp":1467821881930,"startstate":
BOOKING_END_CHECKS,"endstate":BOOKING_ENDED,"transition":BOOKING_ENDED_BY_CUSTOMER,"location":
{"lat":48.14174333333333,"lon":11.5676111111111111,"heading":null}, {"clienttimestamp":1467821989544,"startstate":
BOOKING_ENDED,"endstate":AVAILABLE,"transition":RECOGNIZED_DRIVER_RFID,"location":
{"lat":48.14174333333333,"lon":11.5676111111111111,"head
ing":null}, {"locationhexgrid":null}, {"clienttimestamp":1467799619028,"type":ONLINE,"accountgroup":SERVICE,"leasingtype":
LEASINGTYPE_UNKNOWN,"reservationtime":19183168,"reservationstatus":LED_TO_BOOKING,"defaultrentalstopbehaviour":PARKIN
G},[],nan,[{"clienttimestamp":1467818798175,"screenname":WELCOME_SCREEN,"phase":UNDEFINED,"type":FLOW,"fullscreen":f
alse,"sequenceid":1,"clientscreenname":Welcome.StartAirportMUC.welcome}, {"clienttimestamp":1467818807753,"screenname":PIN,
"phase":UNDEFINED,"type":FLOW,"fullscreen":false,"sequenceid":2,"clientscreenname":Pin.Default.pin}, {"clienttimestamp":1467818812861,"screenname":CLEANNESS_SCREEN,"phase":UNDEFINED,"type":FLOW,"fullscreen":false,"sequenceid":3,
clientscreenname":VehicleInterior.Default.vehicle_interior}, {"clienttimestamp":1467818815404,"screenname":DAMAGE,"phase":
UNDEFINED,"type":FLOW,"fullscreen":false,"sequenceid":4,"clientscreenname":Damages.Default.damage}, {"clienttimestamp":
1467818816627,"screenname":BookingReady.Default.booking_ready,"phase":UNDEFINED,"type":FLOW,"fullscreen":false,"seque
nceid":5,"clientscreenname":BookingReady.Default.booking_ready}, {"clienttimestamp":1467818822188,"screenname":BookingCom

Data Preparation I - Summary Index

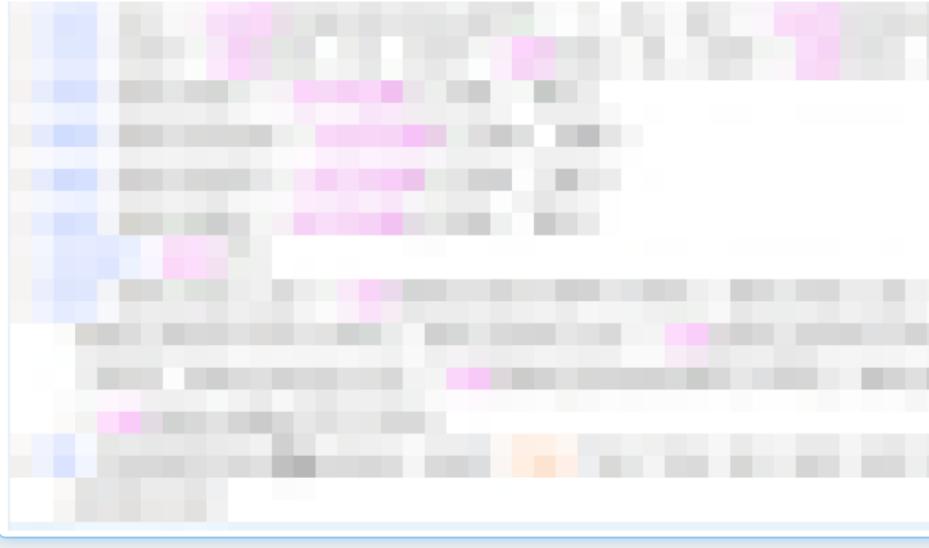
| | _time | ind_lat | ind_lon | count | lat_start_binned | lon_start_binned | max_lat_start | max_lon_start | min_lat_start | min_lon_start | lat_bin | lon_bin |
|----|---------------------|---------|---------|-------|------------------|------------------|---------------|---------------|---------------|---------------|---------|---------|
| 1 | 2017-05-03 23:30:00 | 8 | 8 | 1 | 48.116 | 11.525 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 2 | 2017-05-03 23:30:00 | 8 | 12 | 1 | 48.116 | 11.598 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 3 | 2017-05-03 23:30:00 | 13 | 11 | 1 | 48.174 | 11.580 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 4 | 2017-05-03 23:30:00 | 12 | 8 | 1 | 48.163 | 11.525 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 5 | 2017-05-03 23:30:00 | 10 | 10 | 1 | 48.140 | 11.561 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 6 | 2017-05-03 23:15:00 | 8 | 8 | 1 | 48.116 | 11.525 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 7 | 2017-05-03 23:15:00 | 8 | 7 | 1 | 48.116 | 11.506 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 8 | 2017-05-03 23:15:00 | 13 | 8 | 1 | 48.174 | 11.525 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 9 | 2017-05-03 23:15:00 | 10 | 10 | 1 | 48.140 | 11.561 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 10 | 2017-05-03 23:00:00 | 9 | 9 | 2 | 48.128 | 11.543 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 11 | 2017-05-03 23:00:00 | 9 | 12 | 1 | 48.128 | 11.598 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 12 | 2017-05-03 23:00:00 | 9 | 11 | 1 | 48.128 | 11.580 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |
| 13 | 2017-05-03 23:00:00 | 8 | 12 | 1 | 48.116 | 11.598 | 48.256 | 11.745 | 48.023 | 11.378 | 20 | 20 |

Data Preparation II - PrepStuff

A new preprocessing algorithm for the use case

```
PrepStuff3.py
1  #!/usr/bin/env python
2
3  import datetime
4
5  import pandas as pd
6  import numpy as np
7
8  import cexc
9  from base import BaseAlgo
10 from util.param_util import convert_params
11 from util import df_util
12
13
14 class PrepStuff3(BaseAlgo):
15     def __init__(self, options):
16         # self.handle_options(options)
17         self.params = convert_params(
18             options.get('params', {}),
19             floats=['zero_ratio'])
20     )
21
22     @staticmethod
23     def ind_2_loc(zeros_ind, min_, max_, n):
24         return np.round(zeros_ind * (max_ - min_)/n, 3) + min_
25
26     def fit(self, df, options):
27         df = df.copy()
28         max_lon_start = df.max_lon_start[0]
29         min_lon_start = df.min_lon_start[0]
30         max_lat_start = df.max_lat_start[0]
31         min_lat_start = df.min_lat_start[0]
32         lon_bin = int(df.lon_bin[0])
33         lat_bin = int(df.lat_bin[0])
34
35         counts_per_time = df[['_time', 'count']].groupby(['_time']).sum()
36
37         for index in counts_per_time.index:
38             n_zeros = int(np.round(counts_per_time.loc[index]['count']*self.params['zero_ratio']))
39             zeros_ind_lats = np.random.randint(lat_bin, size=[n_zeros, 1])
40             zeros_ind_lons = np.random.randint(lon_bin, size=[n_zeros, 1])
```

```
index=bmw_summary  
| fit PrepStuff3 * zero_ratio=1
```



Nightly Model Training

- ▶ Nonlinear Algorithm
 - Random Forest Regressor
 - ▶ Historical Data
 - 8 days, 1 month, etc
 - ▶ Model's Output
 - Demand
 - ▶ Model's Input
 - Latitude, Longitude, Features of Time

Fitting Random Forest



Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event.

Create New Model | Load Existing Settings

Enter a search

```
index=bmw_summary
| bin _time span=1h
| stats sum(count) as count by _time, ind_lat, ind_lon, lat_start_binned, lon_start_binned, max_lat_start, max_lon_start, min_lat_start,
  min_lon_start, lat_bin, lon_bin
| table _time, ind_lat, ind_lon, count, lat_start_binned, lon_start_binned, max_lat_start, max_lon_start, min_lat_start,
  min_lon_start, lat_bin, lon_bin
| fit PrepStuff3 * zero_ratio=.5
| eval lat = round(ind_lat/lat_bin*2 - 1, 3), lon = round(ind_lon/lon_bin*2 - 1, 3)
| eval date_hour = strftime(_time, "%H")
| eval date_wday = strftime(_time, "%w")
| convert auto(*)
| eval count_ln = ln(count+1)
```

from Jun 1 through ... ▾



✓ 61,910 events (01/06/2016 00:00:00.000 to 02/07/2016 00:00:00.000)

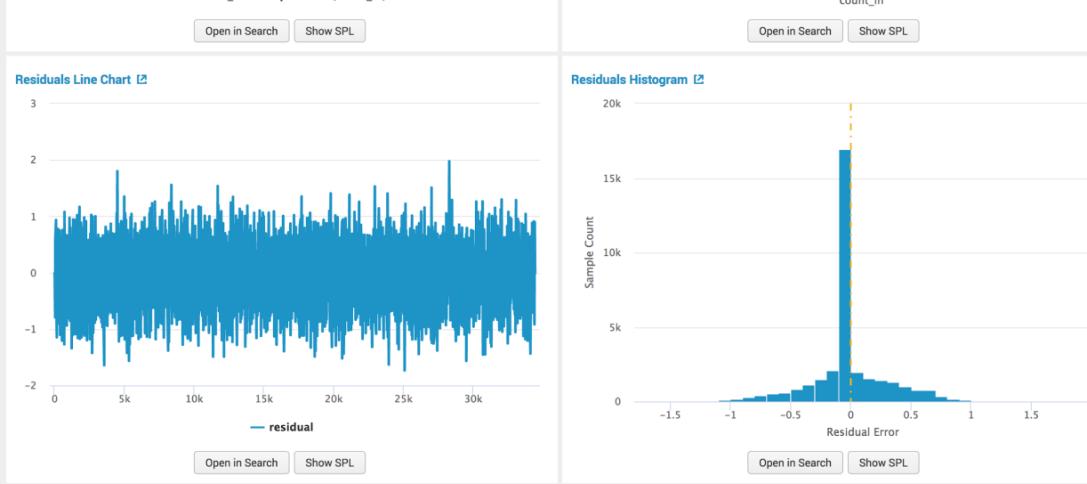
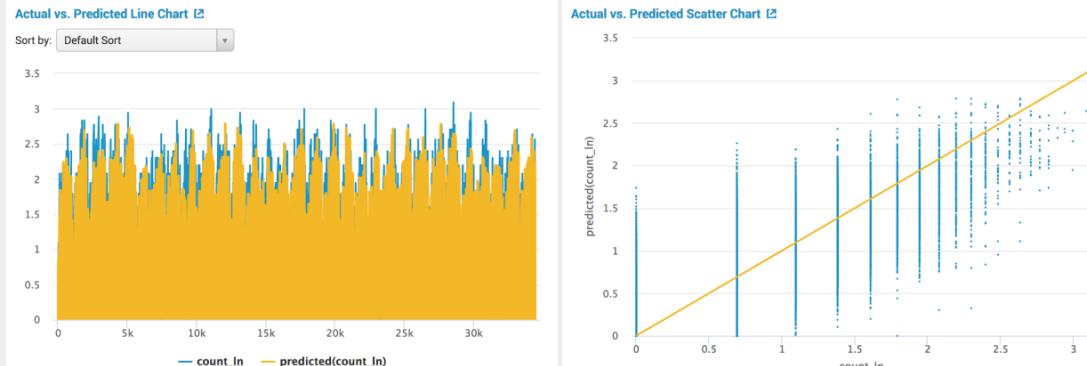
Job ▾ || ■ Smart Mode ▾

splunk> .conf18

Fitting Random Forest

| | | | | |
|---------------------------------|-------------------------|------------------------------|------------------------------------|------------------------------|
| Algorithm | Field to predict | Fields to use for predicting | Split for training / test: 50 / 50 | |
| RandomForestRegressor | count_in | date_hour date_wday lat lon | <input type="range"/> | |
| N Estimators (optional) | Max Depth (optional) | Max Features (optional) | Min Samples Split (optional) | Max Leaf Nodes (optional) |
| Save the model as (optional) | | | | |

Fitting Random Forest



R² Statistic ↗

0.32

| Fit Model Parameters Summary | |
|------------------------------|-----------------|
| feature | importance |
| lon | 0.464183004315 |
| lat | 0.394924801381 |
| date_hour | 0.0953036977518 |
| date_wday | 0.0455884965525 |

Applying Random Forest

```
| `makemygrid(2500)`  
| eval date_hour = 15  
| eval date_wday = 1  
| appendcols  
|   [ inputlookup bmw_grid_boundaries.csv]  
| filldown  
| apply "BMW_Demand_Prediction_RF" as prediction  
| eval prediction = round(exp(prediction)-1)  
| eval lon = (lon + 1)/2*(max_lon_start - min_lon_start) + min_lon_start  
| eval lat = "lat:".((lat + 1)/2*(max_lat_start - min_lat_start) + min_lat_start)  
| chart first(prediction) by lat lon limit=0  
| sort - lat
```

Demand Prediction

Use the ML model to predict usage across the grid.

Edit

[Export](#)

1

Weekday

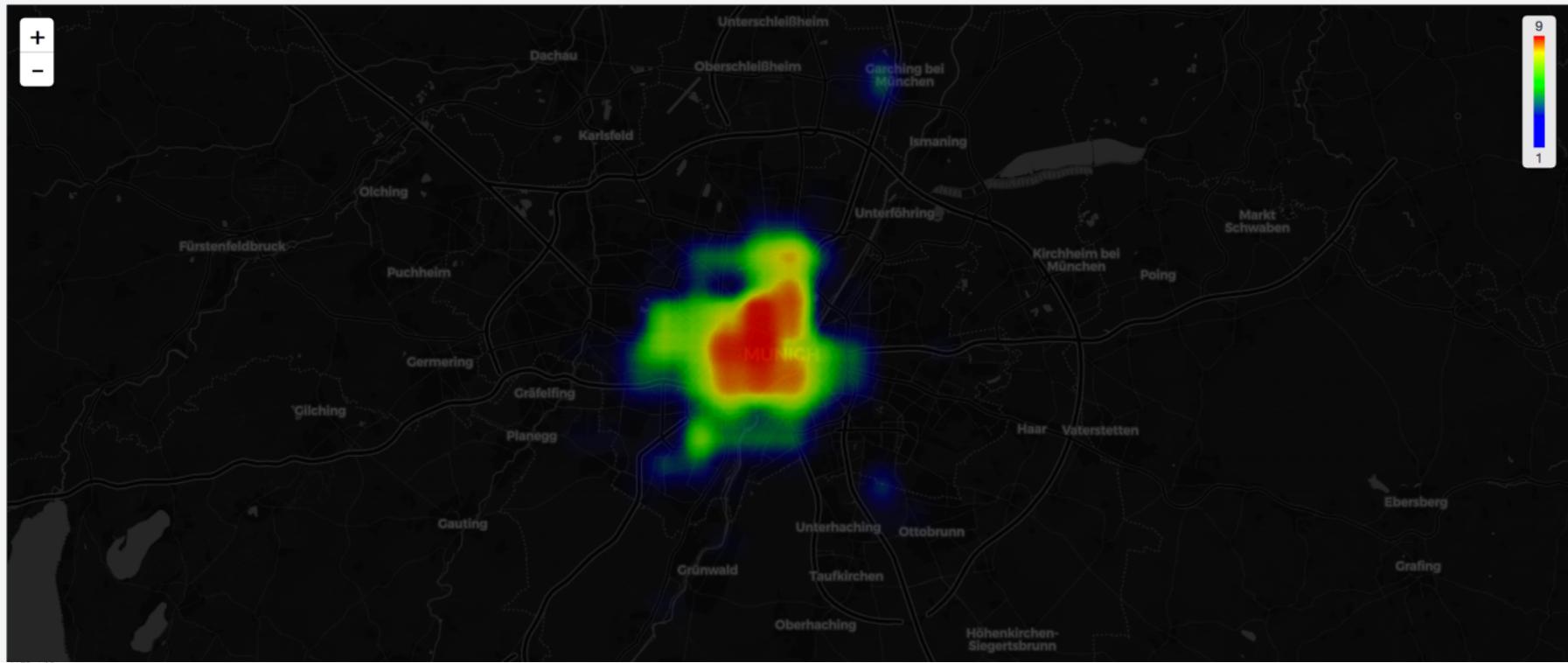
Hour

Thursday

16

Submit

[Hide Filters](#)



```

| makeresults count=168
| streamstats count as x
| eval _time = now() + x*60*60, date_hour = strftime(_time, "%H"), eval date_wday = strftime(_time, "%w")
| eval lat = 0.1, lon=0
| appendcols
  []| inputlookup bmw_grid_boundaries.csv]
| filldown
| apply "BMW_Deman_Prediction_RF" as prediction
| table _time, prediction

```

All time ▾



✓ 168 results (before 05/09/2018 14:53:29.000)

No Event Sampling ▾

Job ▾



Smart Mode ▾

Events

Patterns

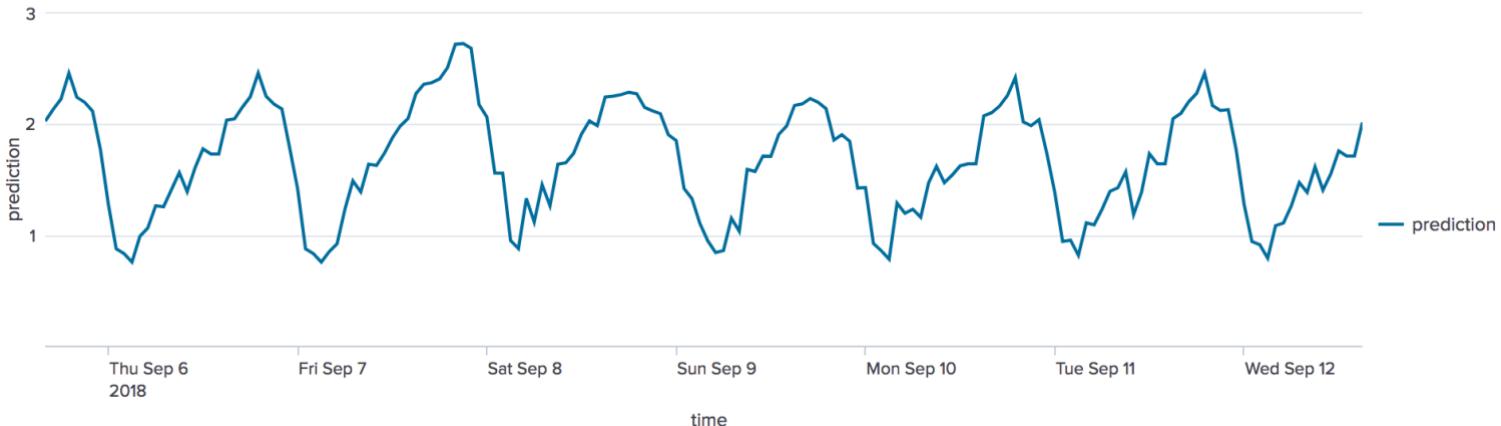
Statistics (168)

Visualization

Line Chart

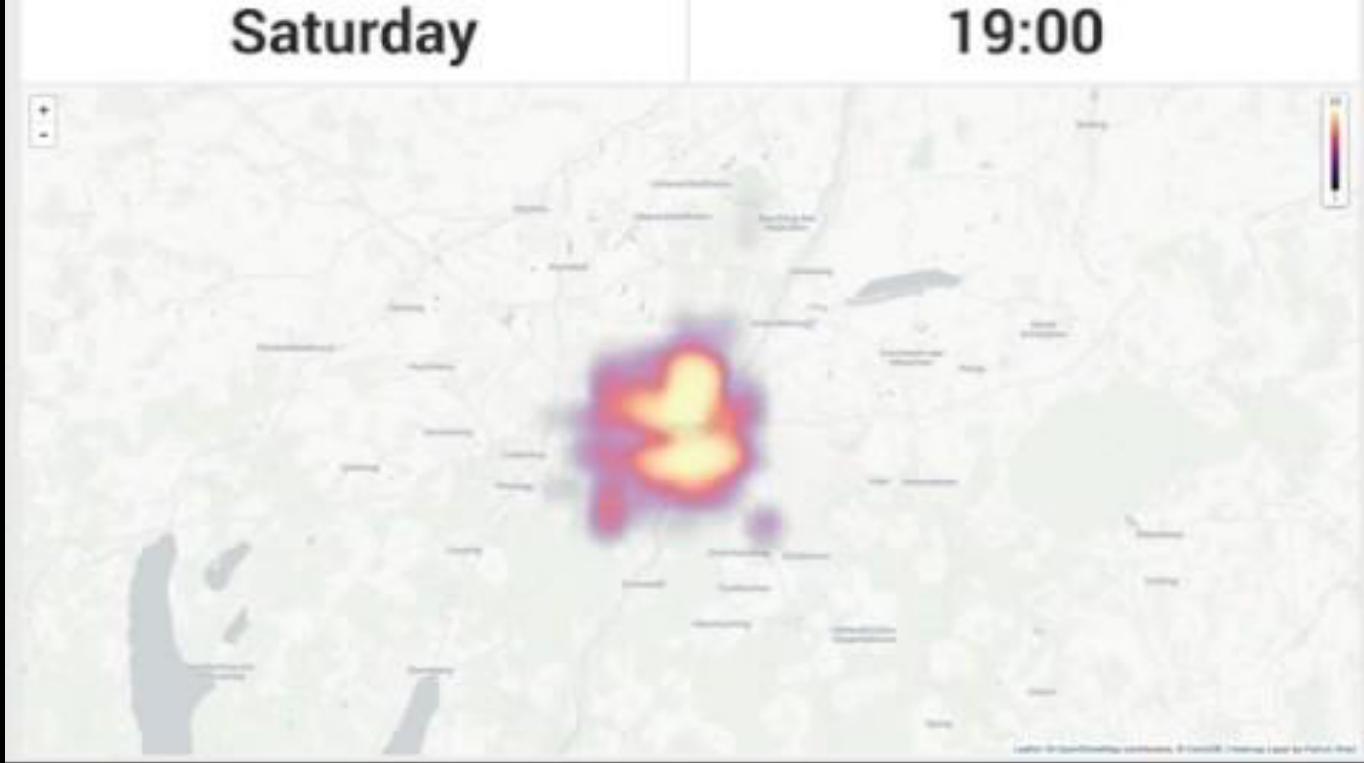
Format

Trellis



Saturday

19:00



Project Natural Language Search @ BMW

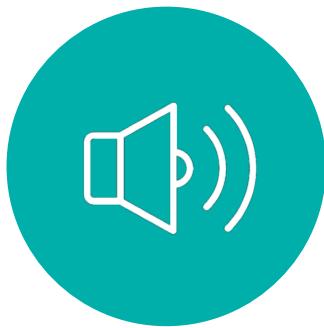


What is Project Natural Language Search?

A natural language platform for machine data that delivers Natural Language Search, Understanding and Generation for Splunk and SQL data.



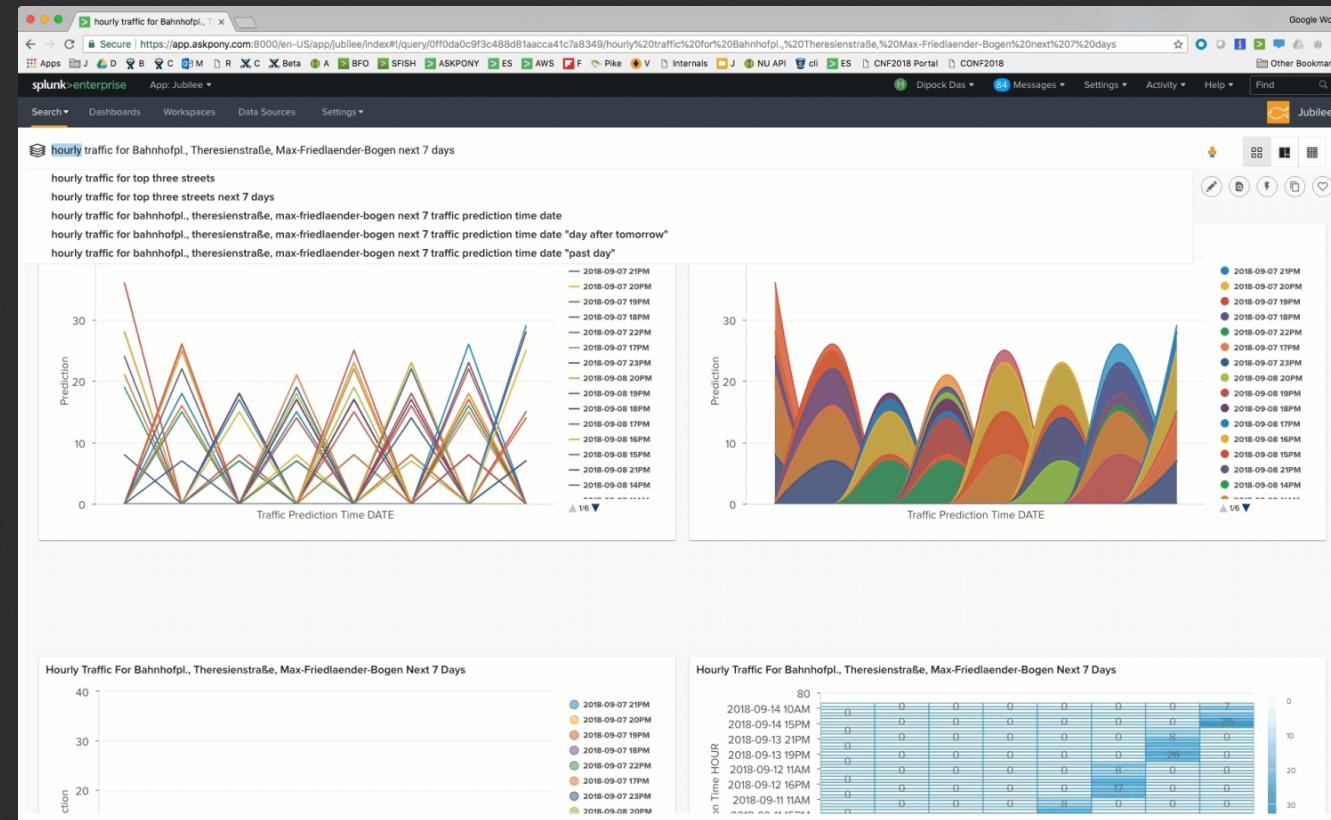
Natural Language Search



Communicate
instantly in charts,
text and voice



Access anywhere
with type, touch,
voice



Traffic Prediction Data Model and Ontology set up

_time

address

city

country

country_code

county

lat

lon

max_lat_start

max_lon_start

min_lat_start

min_lon_start

municipality

prediction

range

route

state

street

suburb

zipcode



congestion for Bahnhofpl., Theresienstraße, Max-Friedlaender-Bogen next 7 days

Lets clarify congestion for better result. Select the following:

| This word is a synonym for an Entity | This word is a synonym for an Attribute | This word is a value | I don't know! Can you suggest? | Skip and ignore this word. |
|--------------------------------------|---|---------------------------------|---------------------------------|---------------------------------|
| <input type="checkbox"/> SELECT | <input type="checkbox"/> SELECT | <input type="checkbox"/> SELECT | <input type="checkbox"/> SELECT | <input type="checkbox"/> SELECT |

congestion Not Understood

Type: Data
Value: bahnhofpl.
Entity Name: Traffic Predict...
Attribute Name: Street
Synonyms: Bahnhof Bahnhofpl Bahnhofpl
Enter a synonym

Bahnhofpl. Understood

Type: Data
Value: bahnhofpl.
Entity Name: Traffic Predict...
Attribute Name: Street
Synonyms: Bahnhof Bahnhofpl Bahnhofpl
Enter a synonym

Theresienstraße Understood

Type: Data
Value: theresienstraße
Entity Name: Traffic Predict...
Attribute Name: Street
Synonyms: Bahnhof Bahnhofpl Bahnhofpl
Enter a synonym

Max-Friedlaender-Bogen Understood

Type: Data
Value: max-friedlaender-bogen
Entity Name: Traffic Predict...
Attribute Name: Street
Synonyms: Bahnhof Bahnhofpl Bahnhofpl
Enter a synonym

next

7

days Understood

Type: Attribute
Entity Name: Virtual Entity
Data Type: number
Name: Traffic Prediction
Natural Type: Date
Format:
Fixed Value:
Made Of: _time (Traffic_Prediction)
Synonyms: day daily date wise per day datewise Enter a synonym

data model

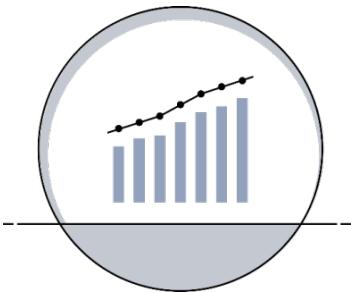
semantic model

What we wanted to achieve with Project NLS

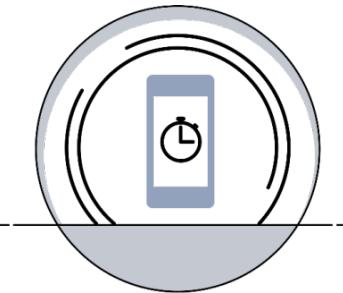
1. Allow non-technical people to ask questions of data in Splunk
2. Use the Traffic Prediction Machine Learning models and generated results
3. Ask the question
“what is the predicted traffic at point X on date Y at time Z?”

Demo

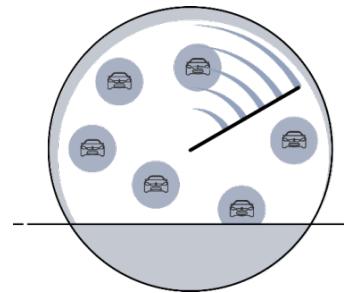
Project Achievements



Predicting hourly, daily, seasonal trends in Munich traffic



Live monitoring on the map of car pick-up events



Live monitoring of public transport coverage gaps

Future Objectives

Expanding the Model

Future Objectives

- ▶ Appropriate reaction after accidents, which cannot be predicted.
 - Once an accident has occurred, the gridlock follows predictable patterns
 - ▶ Include more data from individual users:
 - Using GPS data from single users' smartphones, the software learns your preferred travel times and routes
 - Include user satisfaction: query satisfaction data through an app
 - Predict anomalies like vacation days

It leads to more insights and to making more accurate and user-centric predictions

Other sessions

Spreading the Word: How Chat and Voice Is Transforming Splunk in Retail AI Ops (FN1572)

How we use machine learning in Project Natural Language Search - a Natural Language Platform (FN1629)

Ask Splunk! Using natural language, voice and chat with Splunk Project Natural Language Search (FN1615)

4:30 today

11:00 tomorrow

12:15 tomorrow

Thank You

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

.conf18
splunk>