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

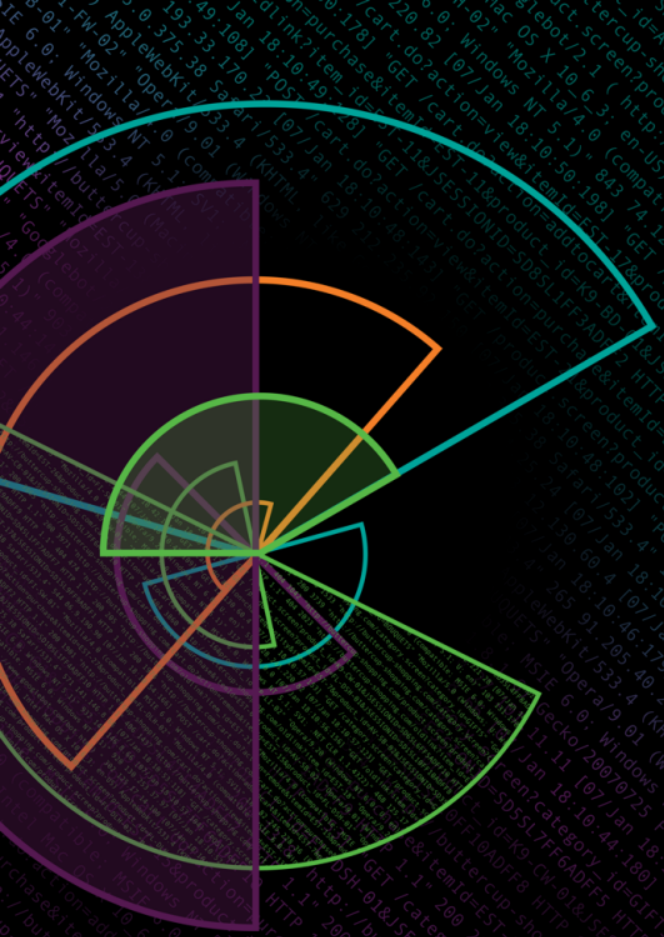
DIPOCK DAS

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.

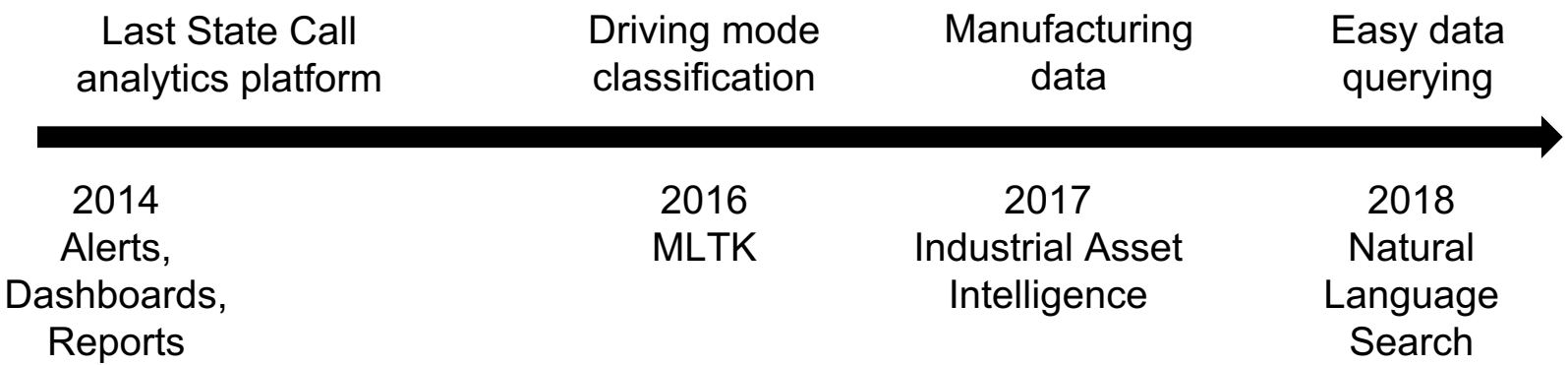


Introduction

Agenda

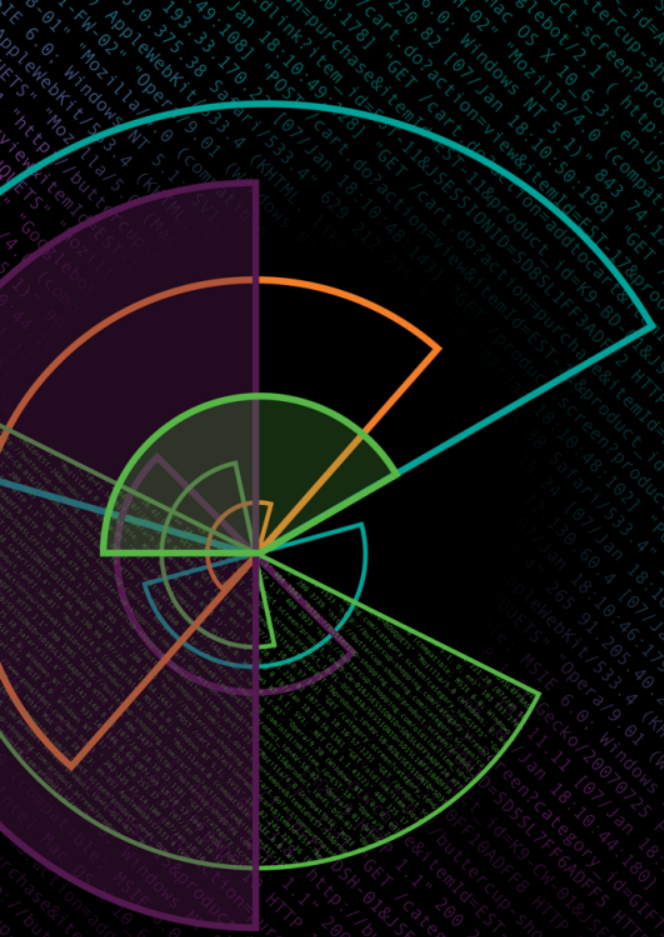
1. The Goal of the Predictive Model
2. Procedure and Achievements in the Project
3. Implementing the Model with Splunk MLTK
4. Using the Model with Project Natural Language Search
5. Project achievements
6. Future Objectives

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

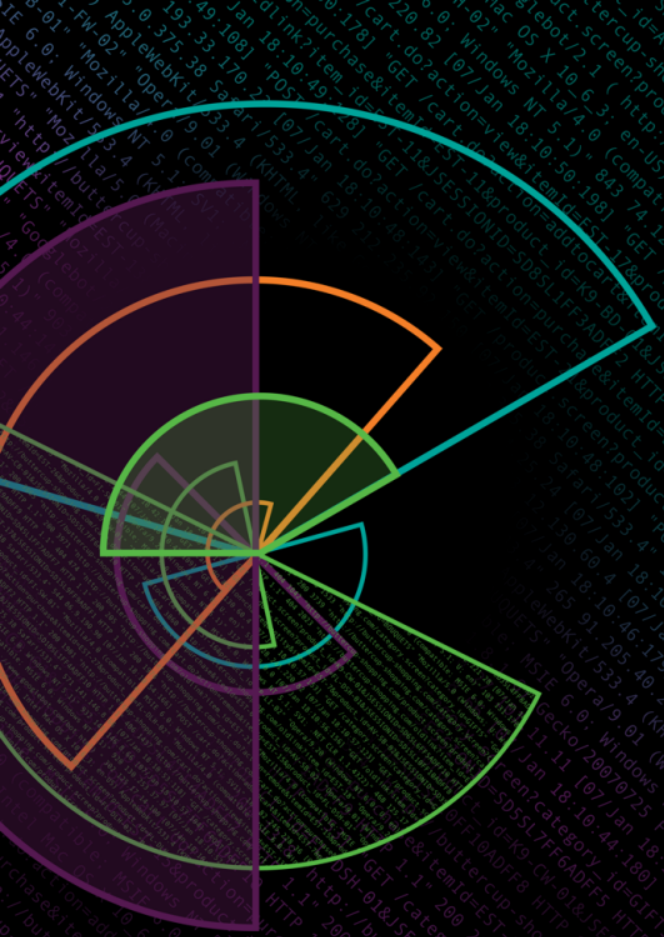


The Goal of the Predictive Model

Human-Centered Machine Learning

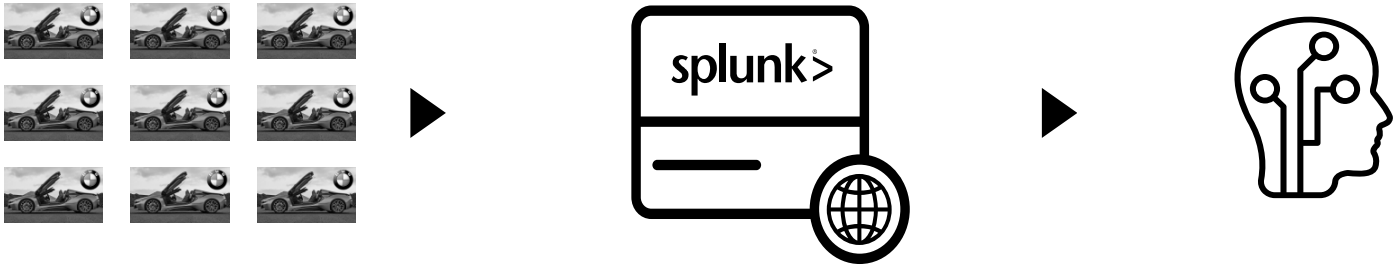
The goal of the Predictive Model

- 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



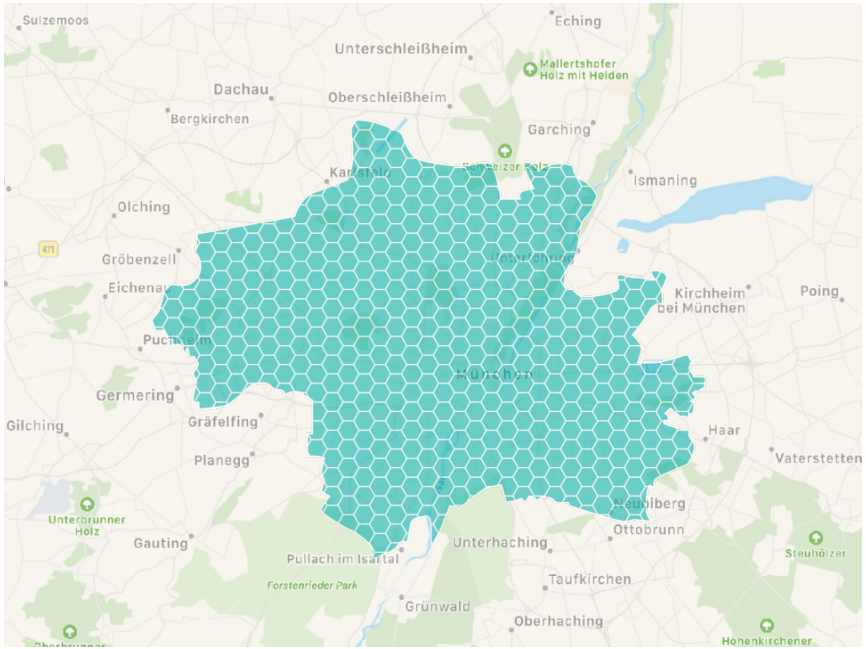
Data from in-development test fleet provides insights on commuting and mobility habits

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

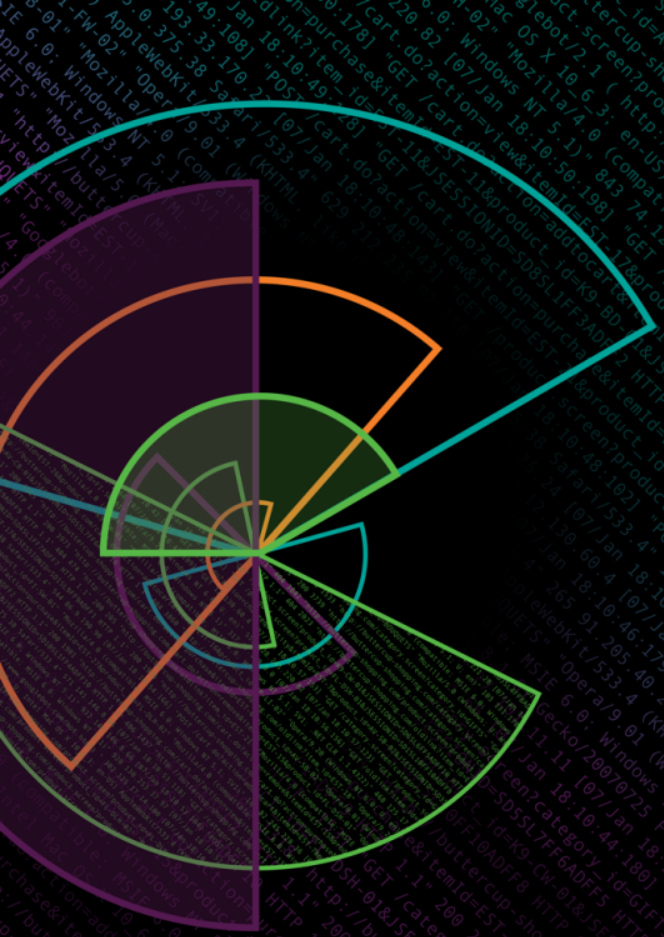
Organising Data

- 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

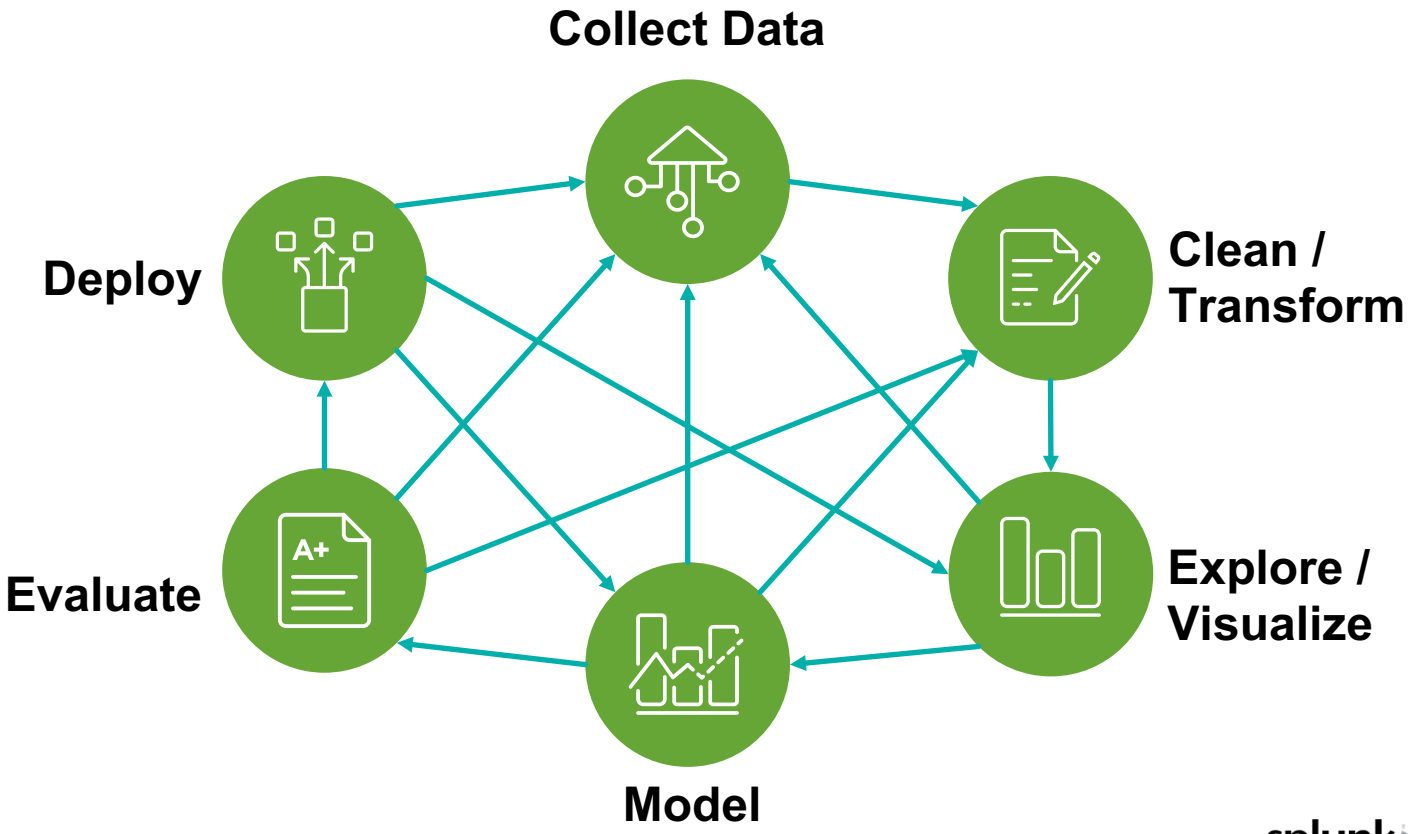
- Algorithm:
  - Non-linear Algorithm
  - Random Forest Regressor
- Through supervised learning, discrepancies were adjusted in order to enhance the accuracy of predictions
- A local search finds the most similar traffic state in the traffic history
- Blends real-time traffic data with past traffic patterns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



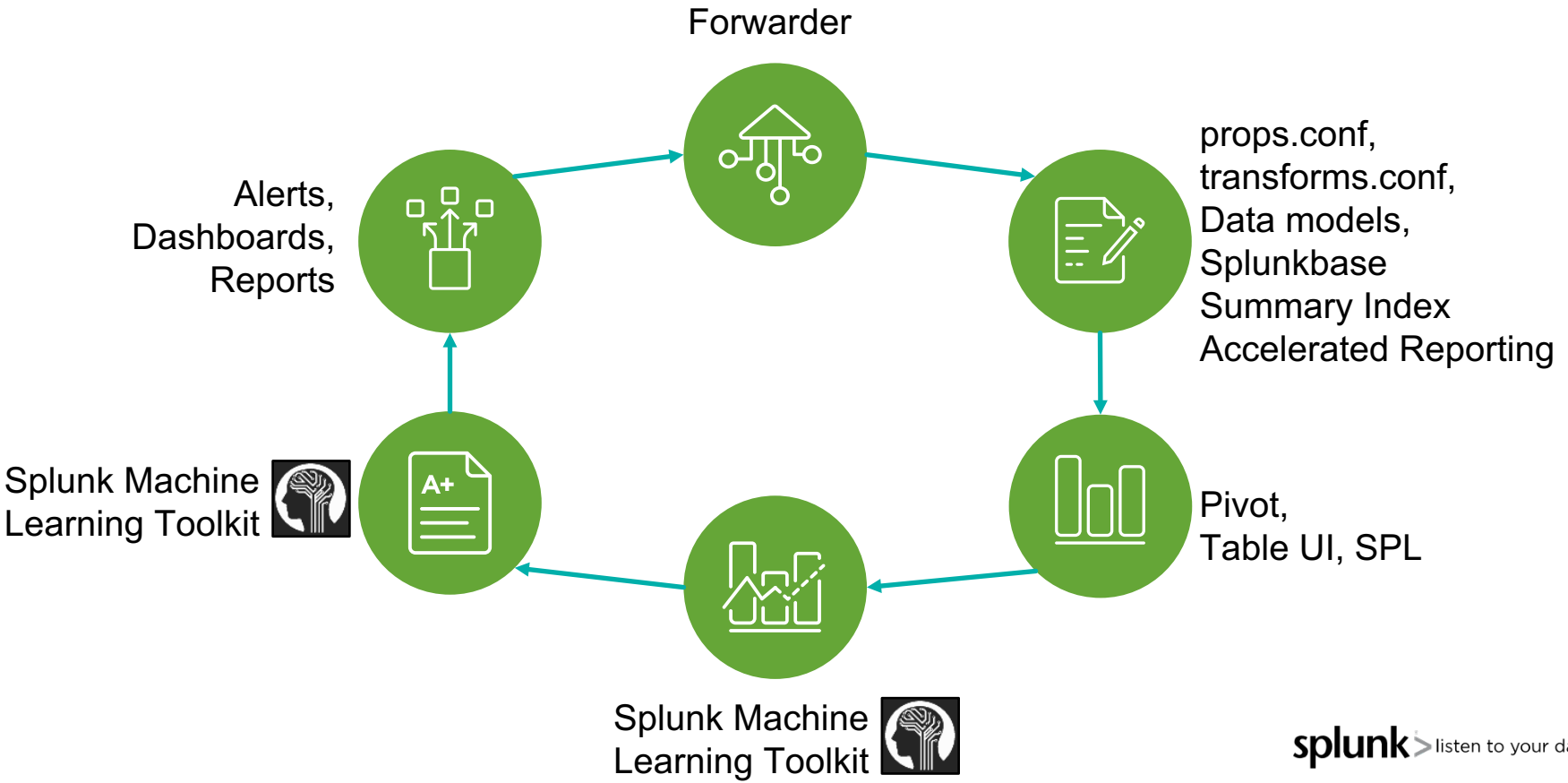
Machine Learning @ Splunk

splunk> .conf18

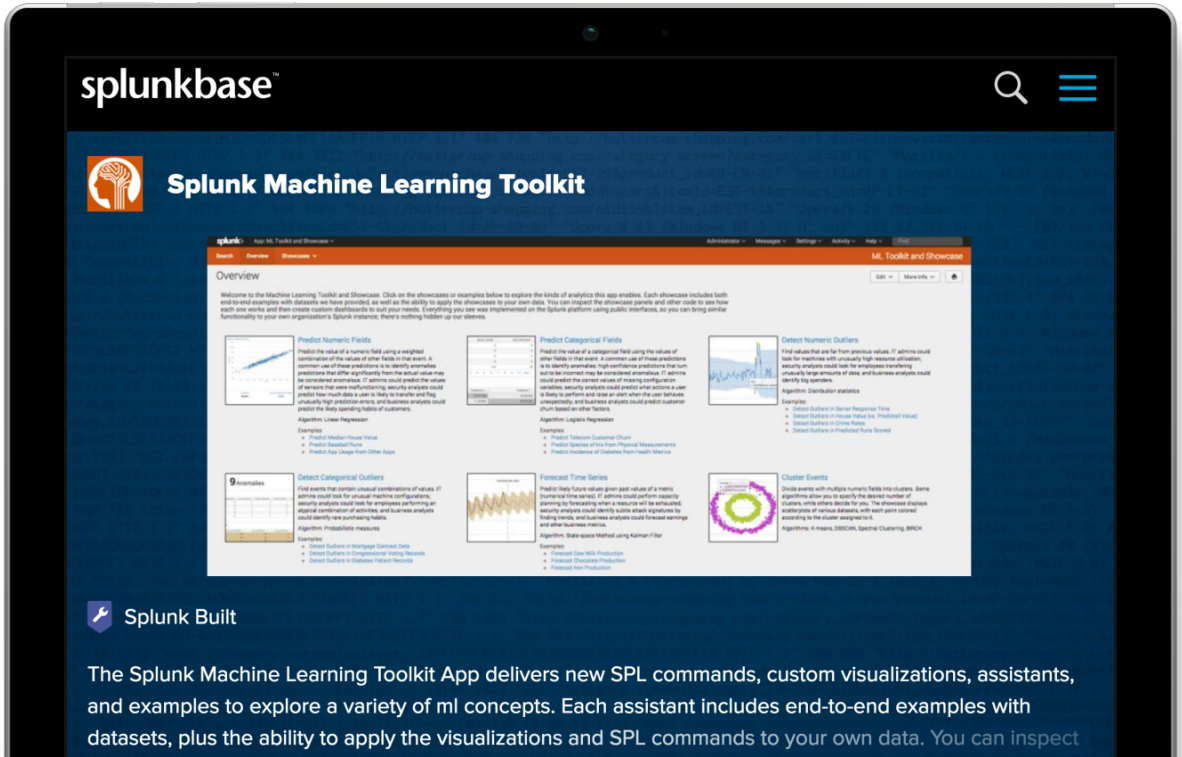
Machine Learning Process



Machine Learning Process with Splunk



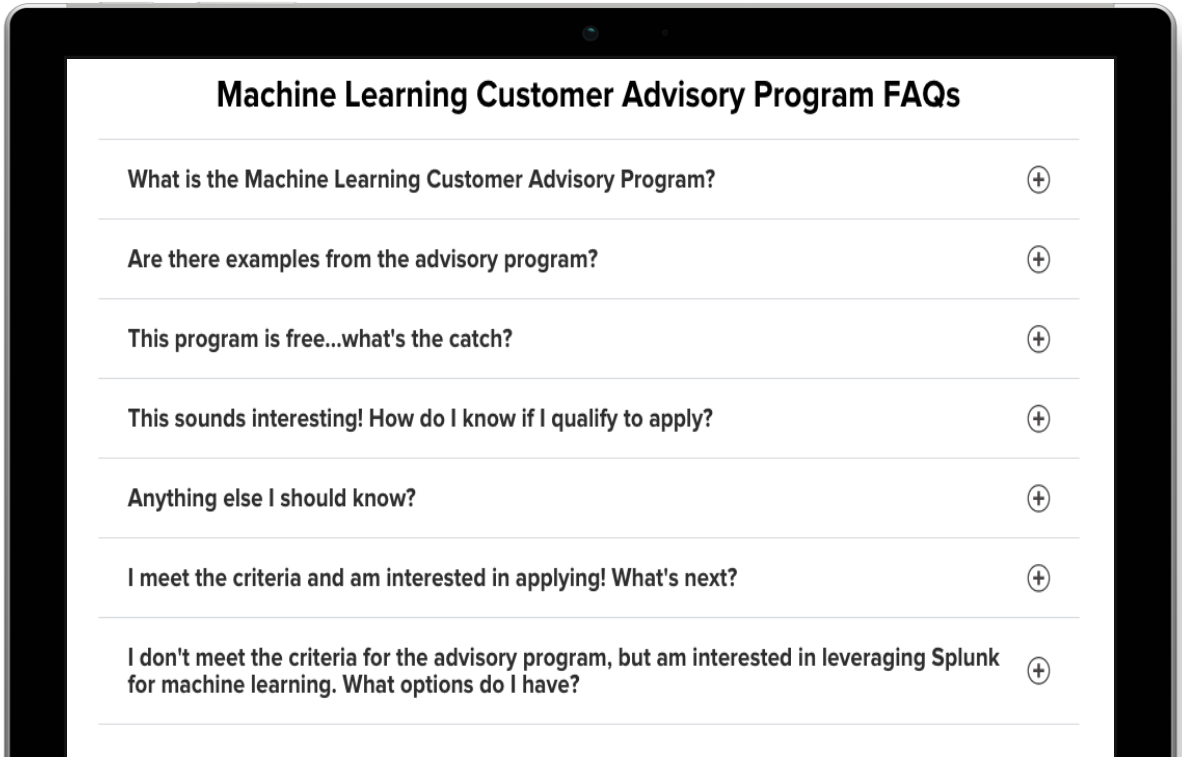
Splunk Machine Learning Toolkit App



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

Splunk ML Advisory Program

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



- 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

splunk> .conf18

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 09:42:16.000

5437363636393200000155BFACABA4@bmw.de,b5c55f4686455132d75960531e6f6abf,CS-CS-MUC-prod-2.1-44.2.42393.20131009.zip,2016-07-06 15:27:01.383,1.46781882138e+12,33669.0,2016-07-06 16:19:49.544,1.46782198954e+12,33707.0,2016-07-06 10:06:59.007,1.46779961901e+12,2016-07-06 16:19:49.544,1.46782198954e+12,WEDNESDAY,12.0,18.0,3378792.0,5437363636393200000155C10CF64A@bmw.de,True,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":"RESERVATION","starttimestamp":1467799619007,"endtimestamp":1467818802175,"duration":19183168},{"phase":"BOOKING_PREPARATION","starttimestamp":1467818802175,"endtimestamp":1467818821383,"duration":19208},{"phase":"DRIVING","starttimestamp":1467818821383,"endtimestamp":1467821877895,"duration":3056512},{"phase":"BOOKING_END","starttimestamp":1467821877895,"endtimestamp":1467821989544,"duration":111649}],[{"clienttimestamp":1467799619007,"startstate":"AVAILABLE","endstate":"RESERVED","transition":"ONLINE_RESERVATION_RECEIVED","location":{"lat":48.35201444444444,"lon":11.787711111111111,"heading":null},"locationhexgrid":null},{"clienttimestamp":1467818802175,"startstate":"RESERVED","endstate":"READY_FOR_BOOKING","transition":"RECOGNIZED_DRIVER_RFID","location":{"lat":48.35201444444444,"lon":11.787711111111111,"heading":null},"locationhexgrid":null},{"clienttimestamp":1467818821383,"startstate":"READY_FOR_BOOKING","endstate":"BOOKED","transition":"BOOKING_COMPLETED","location":{"lat":48.35201444444444,"lon":11.787711111111111,"heading":null},"locationhexgrid":null},{"clienttimestamp":1467821877895,"startstate":"BOOKED","endstate":"BOOKING_END_CHECKS","transition":"BOOKING_ENDED_CHECKS_START","location":{"lat":48.141743333333333,"lon":11.56761111111111,"heading":null},"locationhexgrid":null},{"clienttimestamp":1467821881930,"startstate":"BOOKING_END_CHECKS","endstate":"BOOKING_ENDED","transition":"BOOKING_ENDED_BY_CUSTOMER","location":{"lat":48.14174333333333,"lon":11.56761111111111,"heading":null},"locationhexgrid":null},{"clienttimestamp":1467821989544,"startstate":"BOOKING_ENDED","endstate":"AVAILABLE","transition":"RECOGNIZED_DRIVER_RFID","location":{"lat":48.14174333333333,"lon":11.56761111111111,"heading":null},"locationhexgrid":null}],{"clienttimestamp":1467799619028,"type":"ONLINE","accountgroup":"SERVICE","leasingtype":"LEASINGTYPE_UNKNOWN","reservationtime":19183168,"reservationstatus":"LED_TO_BOOKING","defaultrentalstopbehaviour":"PARKING"},[],nan,[{"clienttimestamp":1467818798175,"screenname":"WELCOME_SCREEN","phase":"UNDEFINED","type":"FLOW","fullscreen":false,"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,"sequenceid":5,"clientscreenname":"BookingReady.Default.booking_ready"},{"clienttimestamp":1467818822188,"screenname":"BookingCom

splunk> .conf18

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

splunk> .conf18

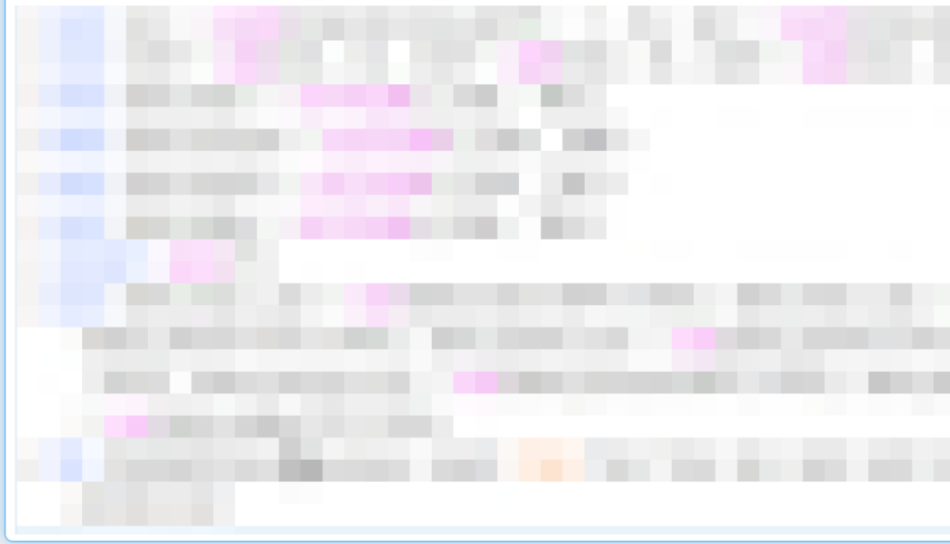
Data Preparation II - PrepStuff

A new preprocessing algorithm for the use case

PrepStuff3.py

```python
#!/usr/bin/env python

import datetime

import pandas as pd
import numpy as np

import cexc
from base import BaseAlgo
from util.param_util import convert_params
from util import df_util


class PrepStuff3(BaseAlgo):
    def __init__(self, options):
        # self.handle_options(options)
        self.params = convert_params(
            options.get('params', {}),
            floats=['zero_ratio']
        )

    @staticmethod
    def ind_2_loc(zeros_ind, min_, max_, n):
        return np.round(zeros_ind *(max_ - min_)/n, 3) + min_

    def fit(self, df, options):
        df = df.copy()
        max_lon_start = df.max_lon_start[0]
        min_lon_start = df.min_lon_start[0]
        max_lat_start = df.max_lat_start[0]
        min_lat_start = df.min_lat_start[0]
        lon_bin = int(df.lon_bin[0])
        lat_bin = int(df.lat_bin[0])

        counts_per_time = df[['_time', 'count']].groupby(['_time']).sum()

        for index in counts_per_time.index:
            n_zeros = int(round(counts_per_time.loc[index]['count']*self.params['zero_ratio']))
            zeros_ind_lats = np.random.randint(lat_bin, size=[n_zeros, 1])
```

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



splunk> .conf18

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

Search | Showcase | Assistants | Scheduled Jobs | Docs | Video Tutorials | Splunk Machine Learning Toolkit

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)
```

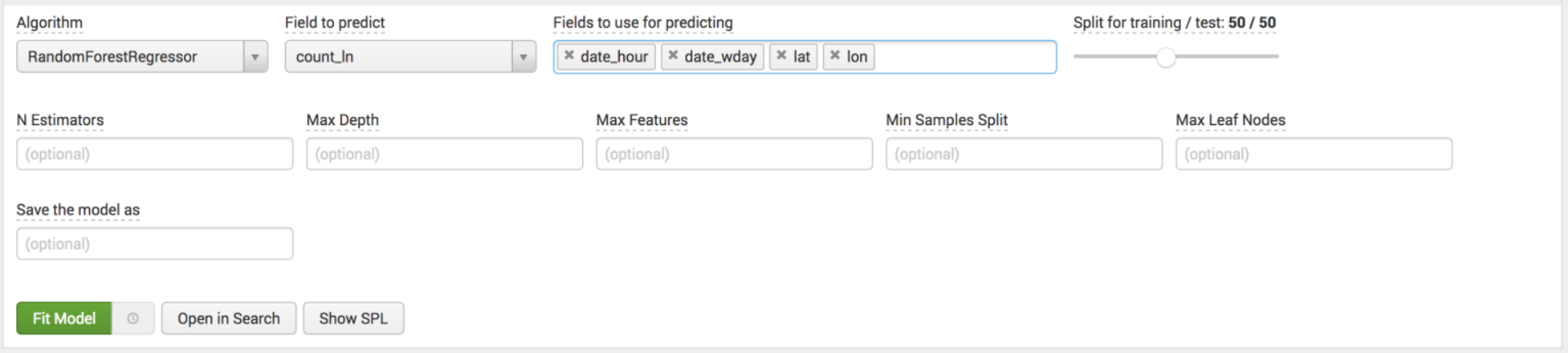
from Jun 1 through ...

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

Job | Smart Mode

splunk> .conf18

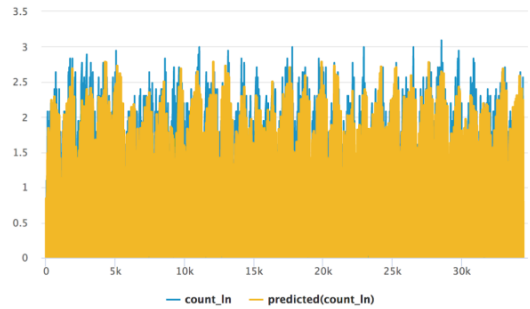
Fitting Random Forest



Fitting Random Forest

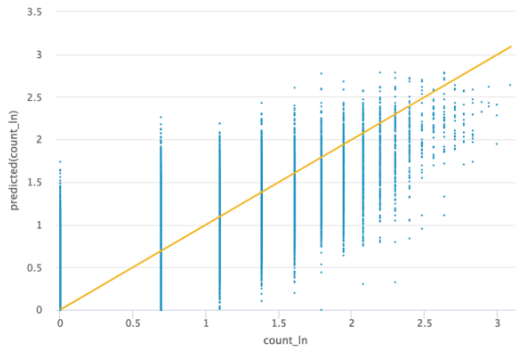
Actual vs. Predicted Line Chart

Sort by: Default Sort



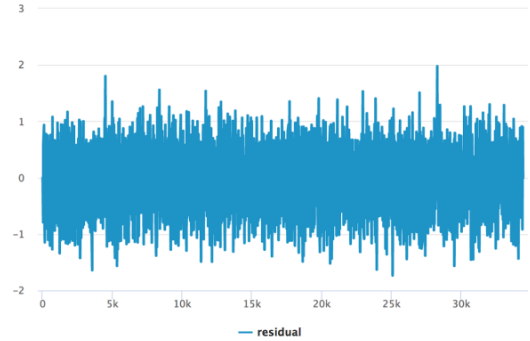
Open in Search | Show SPL

Actual vs. Predicted Scatter Chart



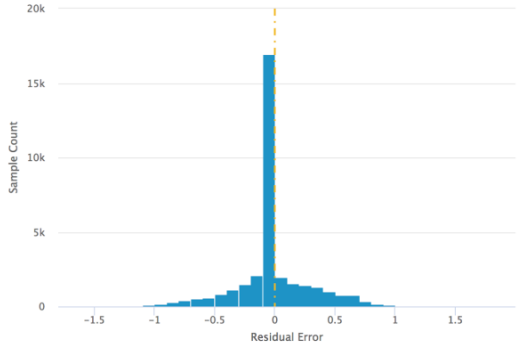
Open in Search | Show SPL

Residuals Line Chart



Open in Search | Show SPL

Residuals Histogram



Open in Search | Show SPL

R^2 Statistic

0.7659

Root Mean Squared Error (RMSE)

0.32

Open in Search | Show SPL

Fit Model Parameters Summary

feature	importance
lon	0.464183004315
lat	0.394924801381
date_hour	0.0953036977518
date_wday	0.0455884965525

Open in Search | Show SPL

splunk> .conf18

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
```

All time

Demand Prediction

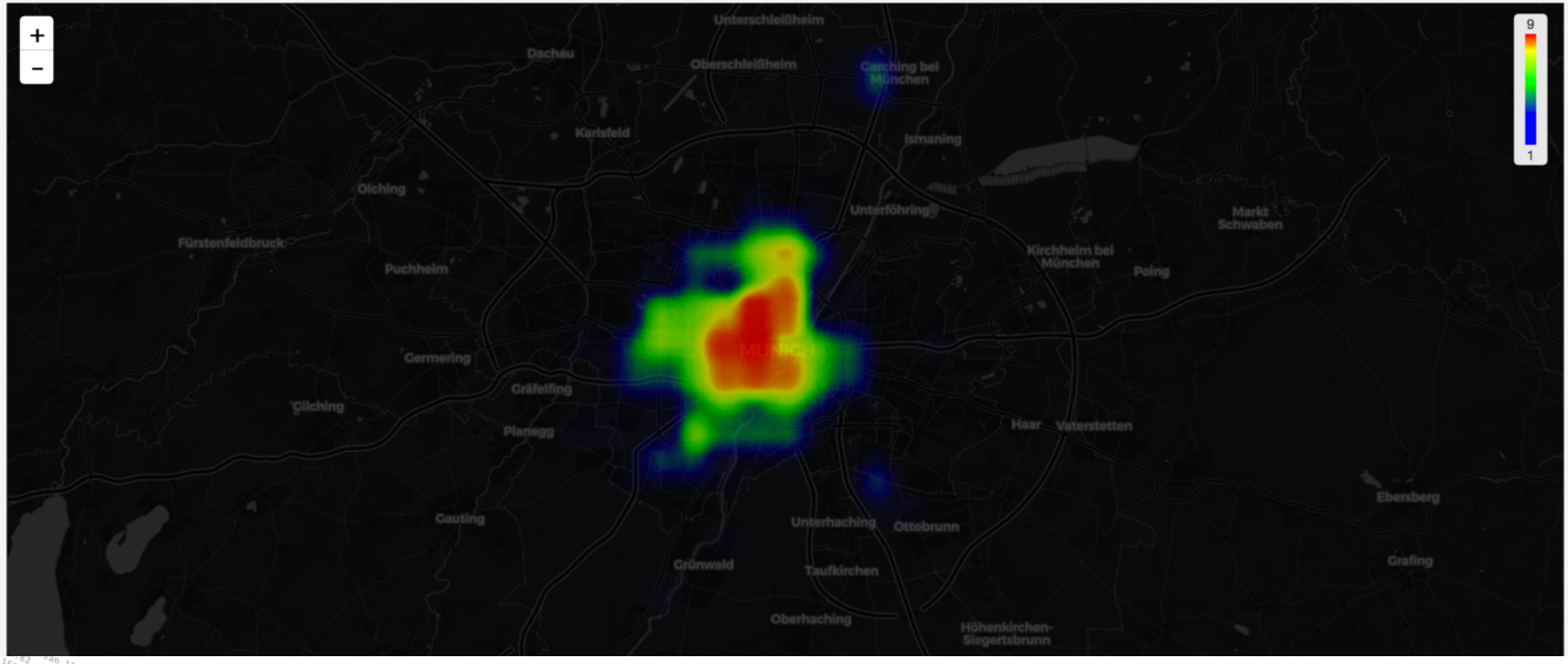
Use the ML model to predict usage across the grid.

Edit | Export | ...

Weekday: Thursday

Hour: 16

Submit | Hide Filters



splunk> .conf18

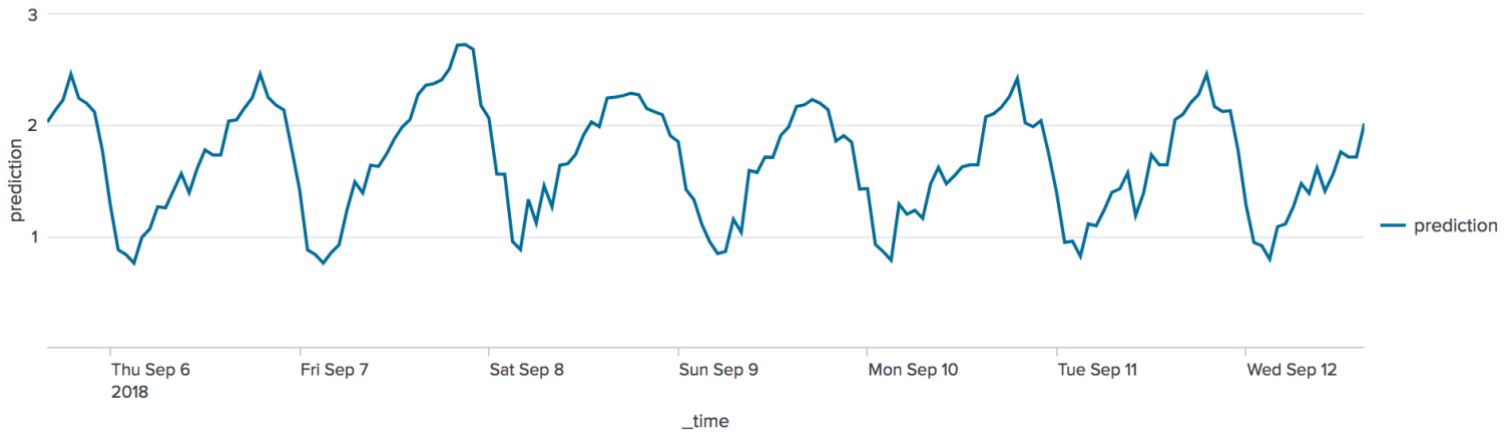
```
| makeresults count=168
| streamstats count as x
| eval _time = now() + x*60*60, date_hour = strftime(_time, "%H"),  eval date_wday = strftime(_time, "%w")
| eval lat = 0.1, lon=0
| appendcols
    [| inputlookup bmw_grid_boundaries.csv]
| filldown
| apply "BMW_Deman_Prediction_RF" as prediction
|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

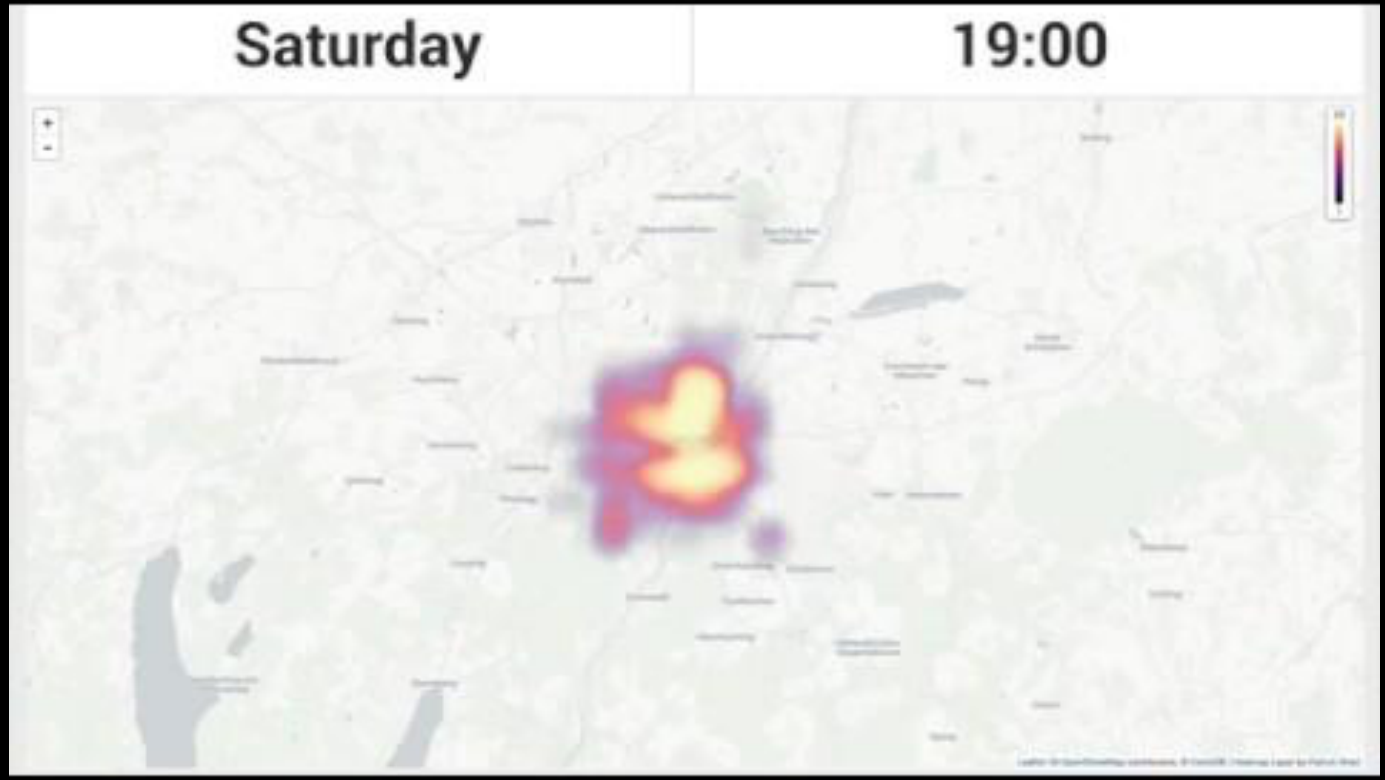
All time

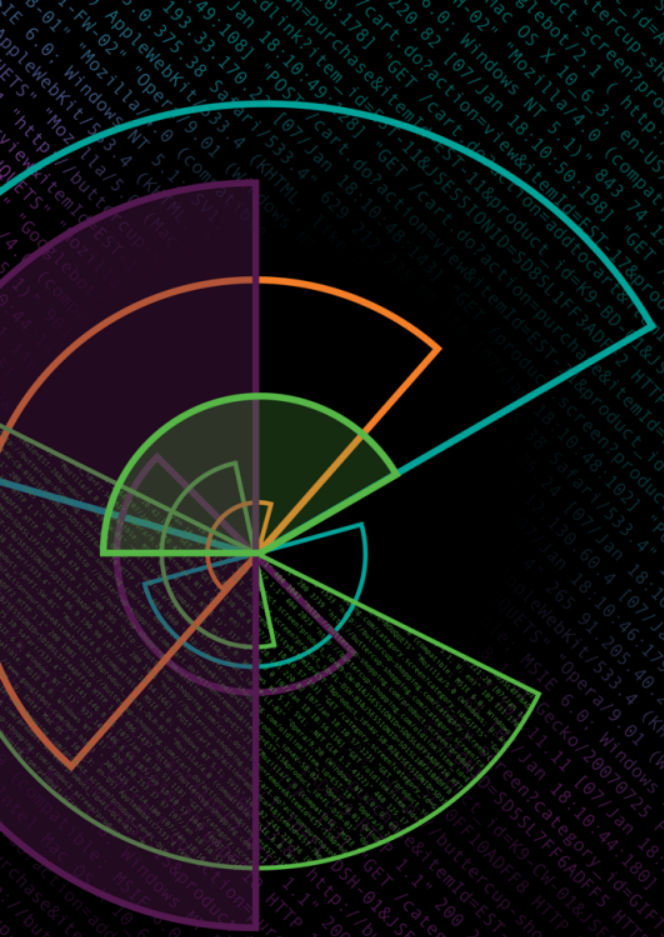
✓ 168 results (before 05/09/2018 14:53:29.000) No Event Sampling Job Smart Mode

Events Patterns Statistics (168) **Visualization**

Line Chart Format Trellis







Project Natural Language Search @ BMW

splunk> .conf18

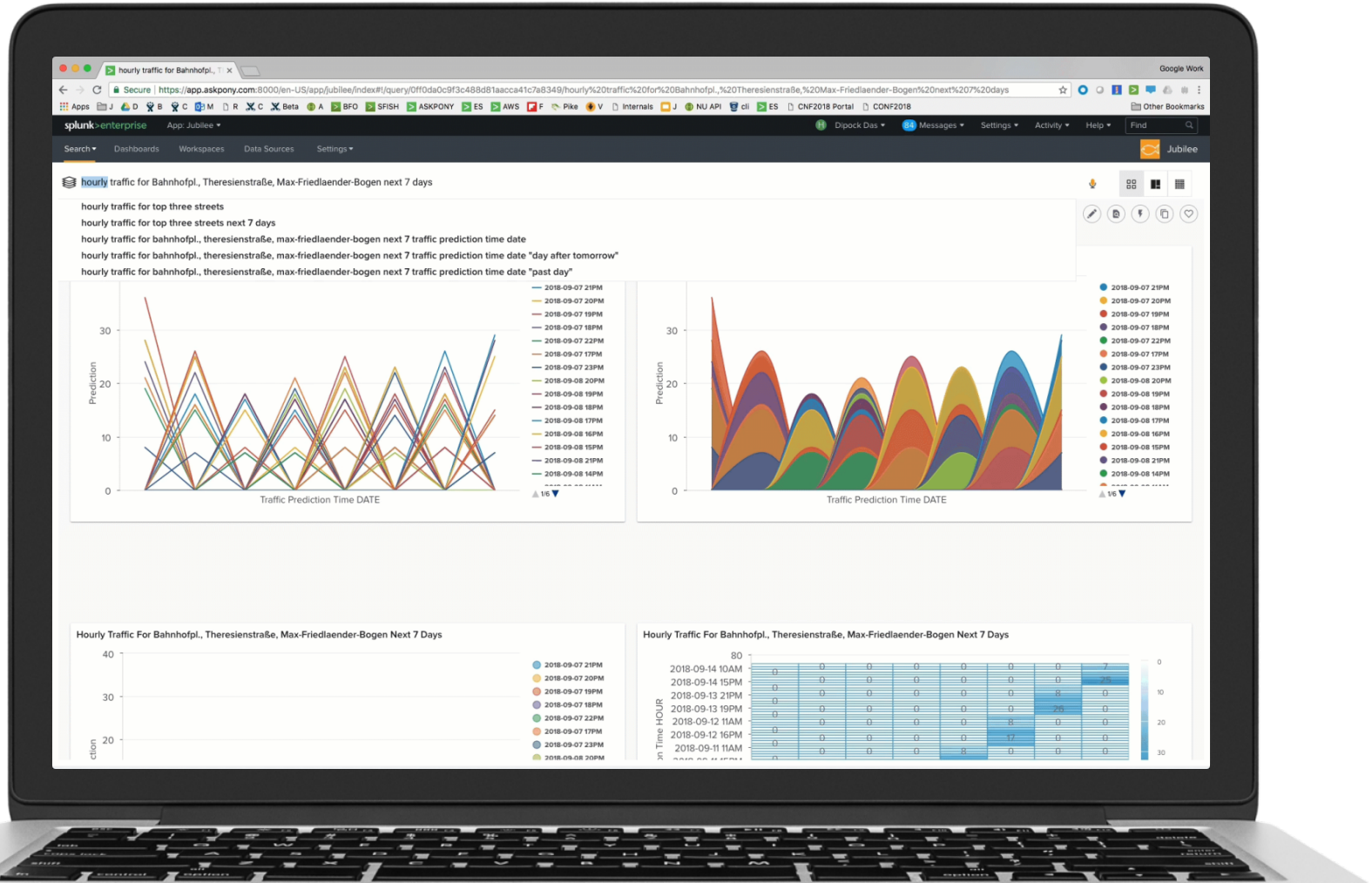
What is Project Natural Language Search?

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

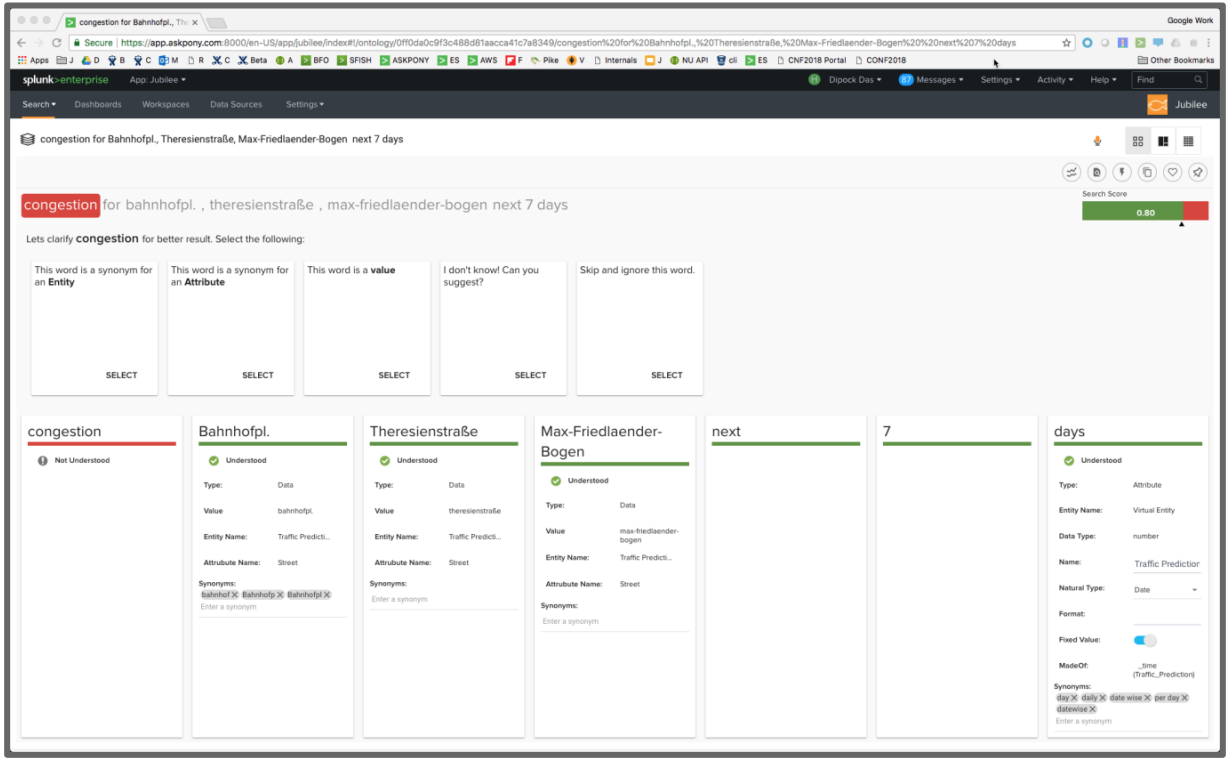
Access anywhere with type, touch, voice



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

data model



semantic model

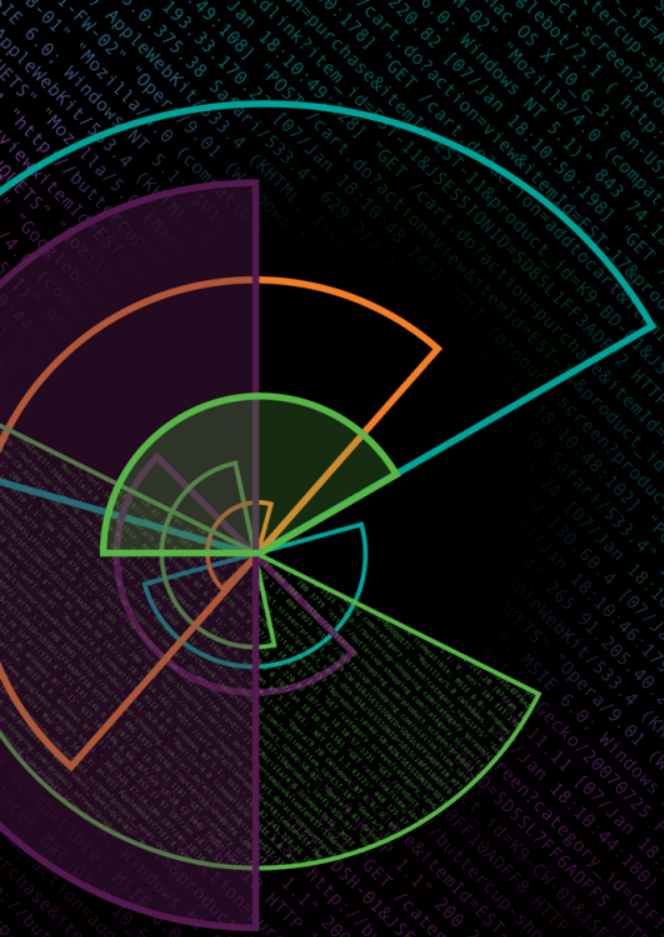
splunk> .conf18

What we wanted to achieve with Project NLS

1. Allow non-technical people to ask questions of data in Splunk

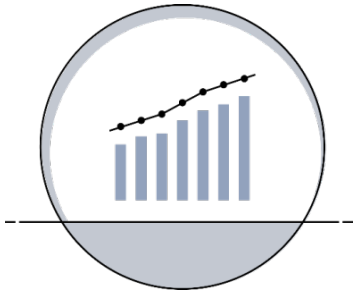
2. Use the Traffic Prediction Machine Learning models and generated results

3. Ask the question
   “what is the predicted traffic at point X on date Y at time Z?”

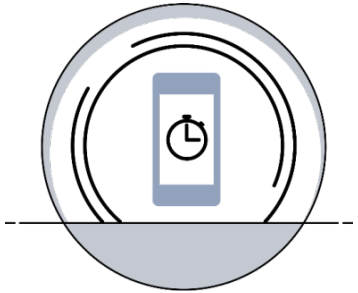


Demo

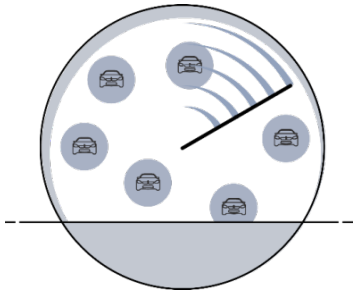
Project Achievements



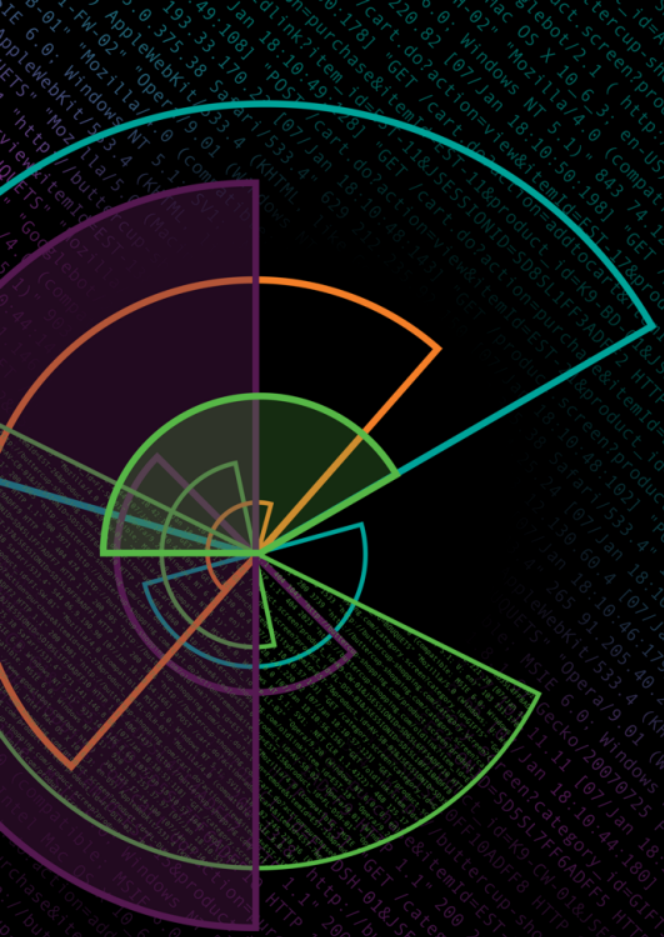
Predicting hourly, daily, seasonal trends in Munich traffic



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)	**4:30 today**
How we use machine learning in Project Natural Language Search - a Natural Language Platform (FN1629)	**11:00 tomorrow**
Ask Splunk! Using natural language, voice and chat with Splunk Project Natural Language Search (FN1615)	**12:15 tomorrow**

Thank You

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