Anomaly Detection On Business Items With Machine Learning Algorithms

Andre Pietsch

Product Manager Splunk, OTTO Hamburg Germany

Stefan Scholz

Sr Consultant, LC Systems Munich Germany

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.

Agenda

- About Otto IT
- About LC Systems
- Initial Situation
- Next Generation Of Data Analytics
- Results

About OTTO

History

Werner Otto

Founder of the company in 1949

CEO until 1981

Dr. Michael Otto

Chairman & CEO of Otto group until 2007

Chairman of Otto group supervisory board until today

Selection Of Catalogs

1950

1957

1964

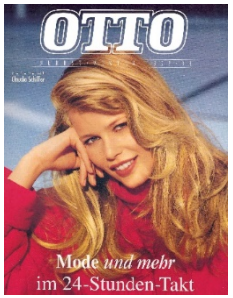
1970

1986

1987

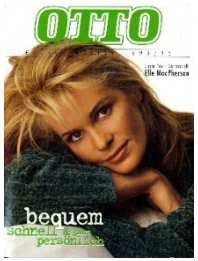
1991

1993



1994

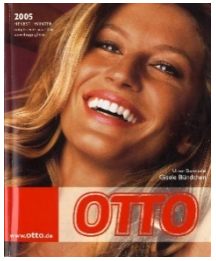
1995



1995

2004

2005



2006

2009

2010

OTTO – Subsidiary Of The Otto Group

Headquarters: OTTO-Campus in Hamburg, Germany

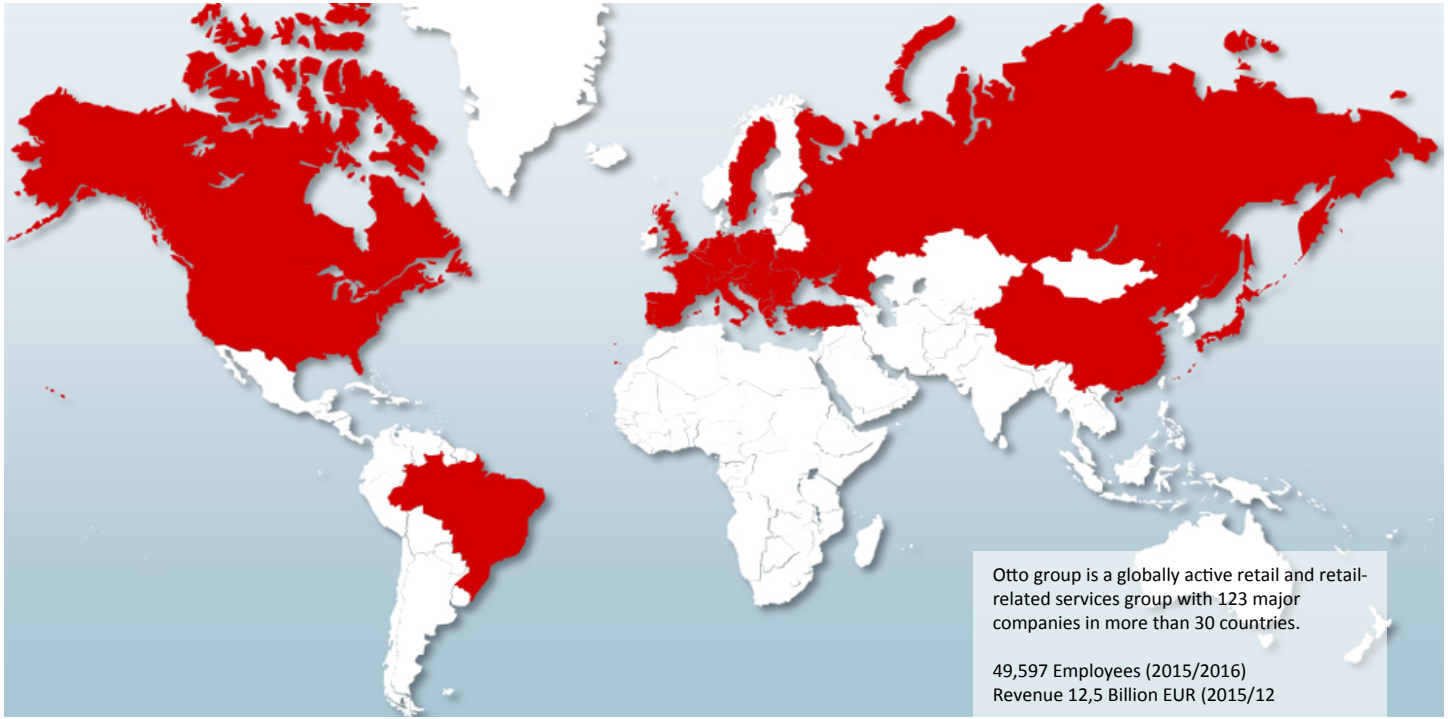
- Employees: 4,350
- Revenue 2.6 Billion € (2015/2016)
- 90% Online
- > 2.2 Million items
- 6,000 Brands

OTTO – Subsidiary Of The Otto Group

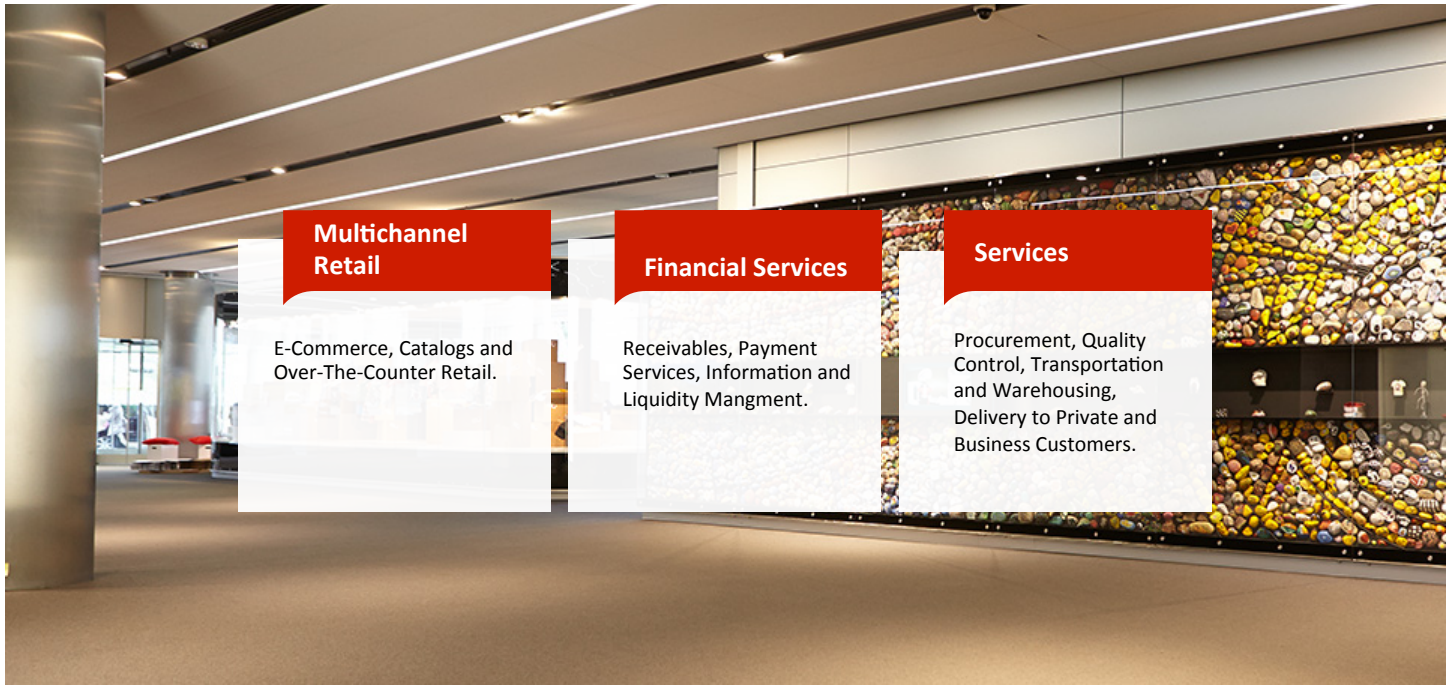
Headquarters: OTTO-Campus in Hamburg, Germany

- #1 Furniture online Germany (before IKEA)
- #2 Germany (after Amazon)
- > 678,000 Facebook fans (03/2016)
- > 28,000 Twitter followers
- 90,000 Service requests via Facebook and Twitter (2015/2016)

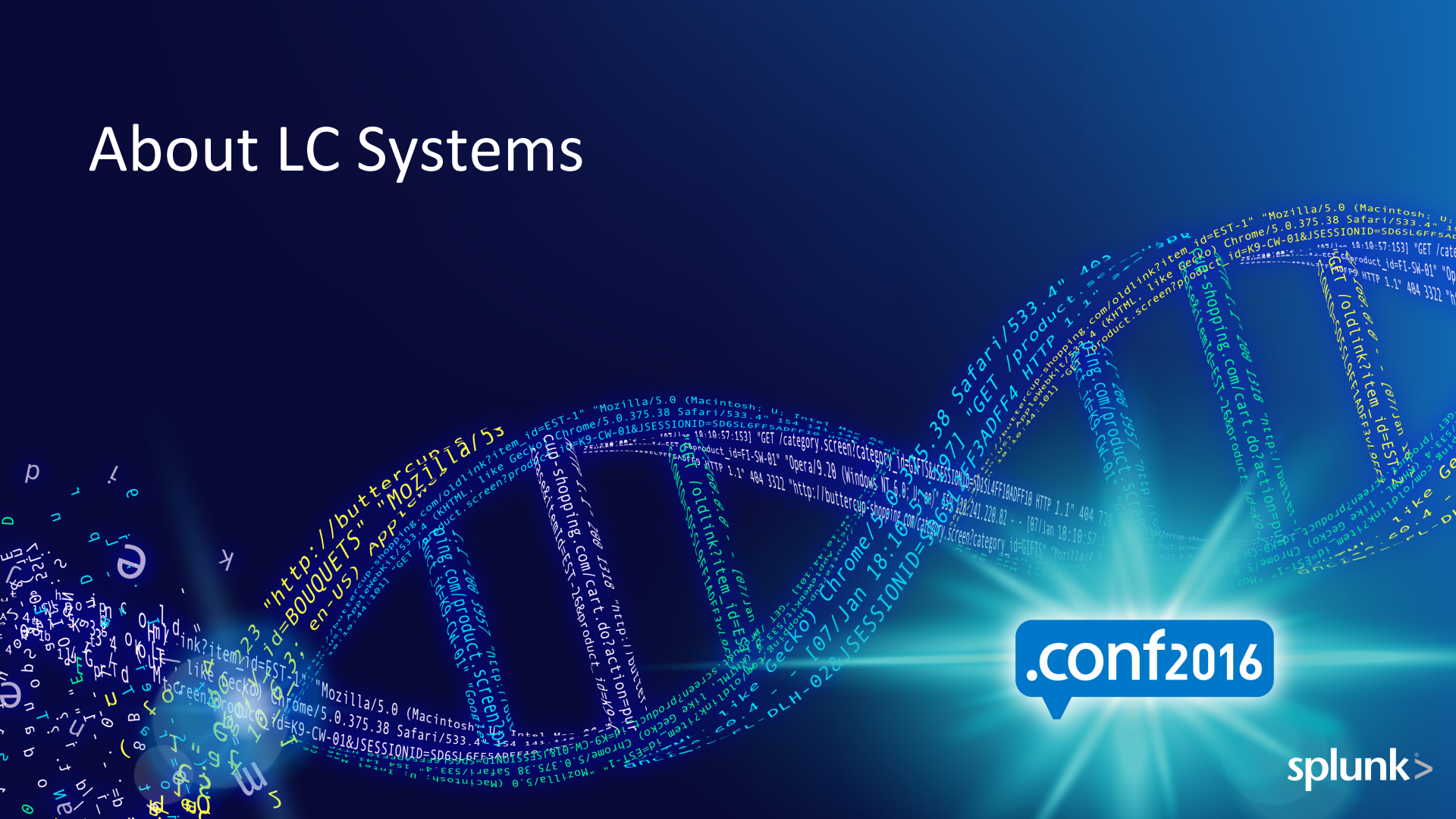
Otto Group - International Success



Otto Group - Main Business Segments



About LC Systems



Locations

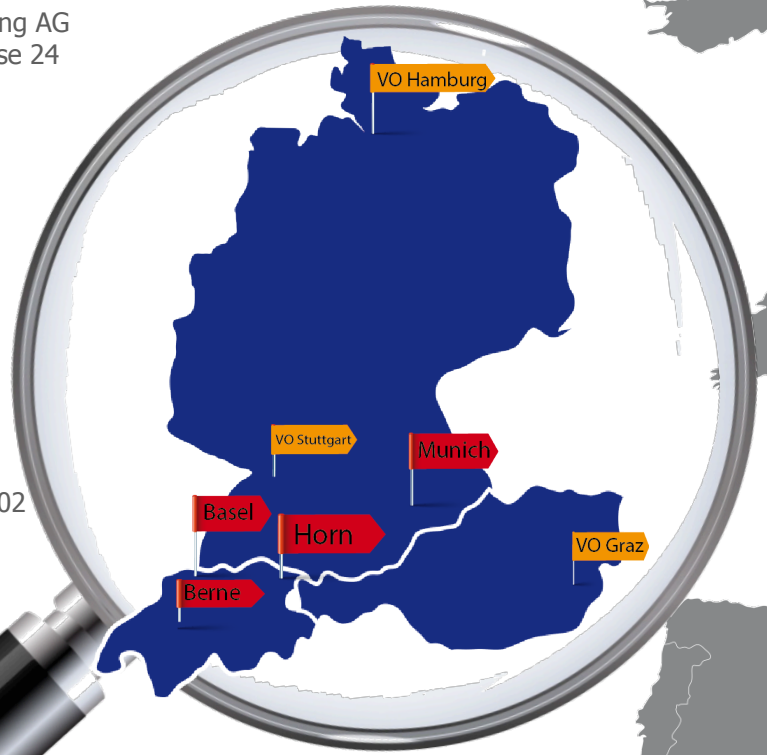
Headquarter
LC Systems-Engineering AG
Postfach 40, Seestrasse 24
CH-9326 Horn

Office Basel
Reinacherstrasse 129
CH-4053 Basel

Office Berne
Schwarztorstr. 9
CH-3007 Berne

Office Germany
LC Systems GmbH
Landsberger Straße 302
D-80687 Munich

www.lcsystems.ch
www.lcsystems.de



Together On The Road To Success

(Not Conclusive)

Pharma

Finance

Automotive Industry

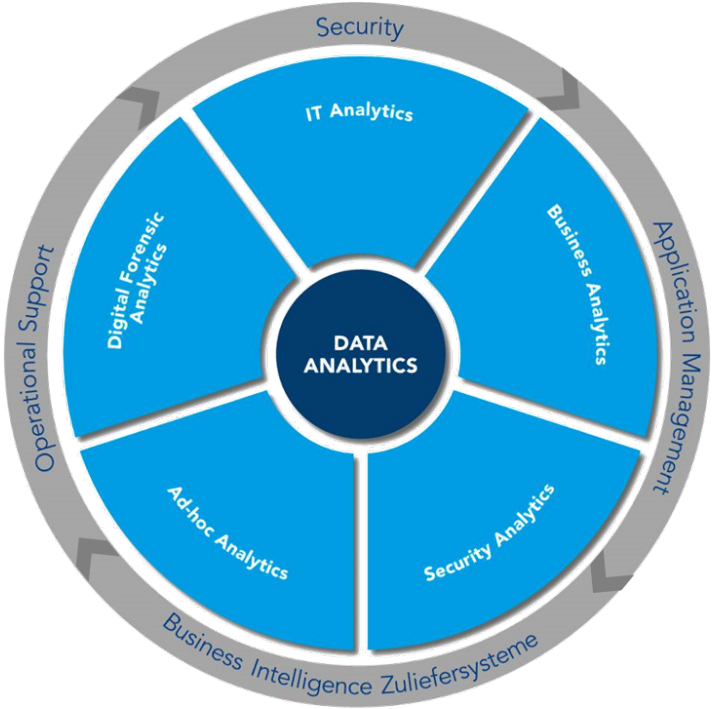
Internet Provider

Research and Development

Service Provider

Trading

Data Analytics
Using And Evaluating Data In The Best Possible Way

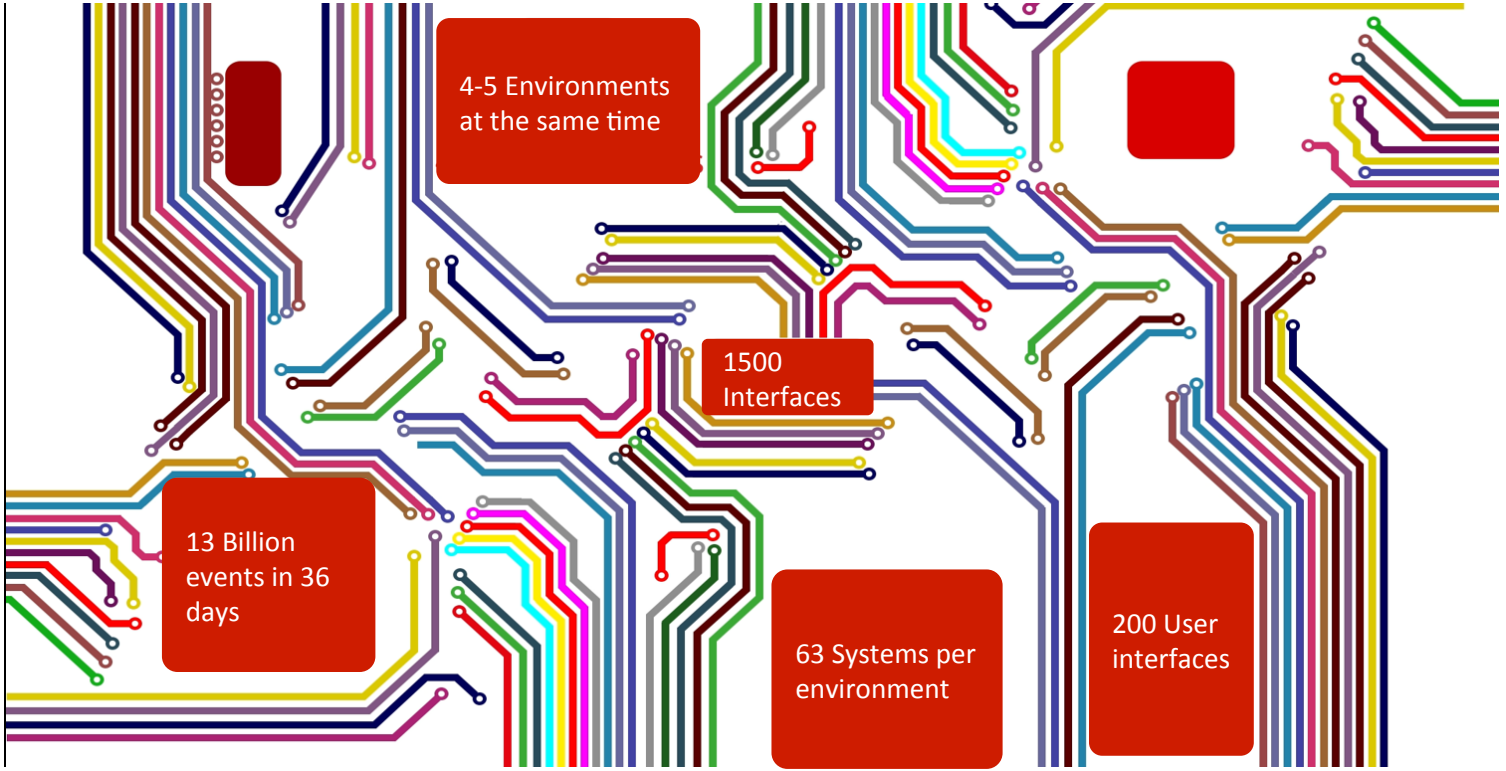


- Services
  - Consulting
  - Workshops
  - Use cases
  - Proof of Concept
  - Project definition
  - Methodically structured implementation
  - Operation
  - Managed services
  - Training

Initial Situation

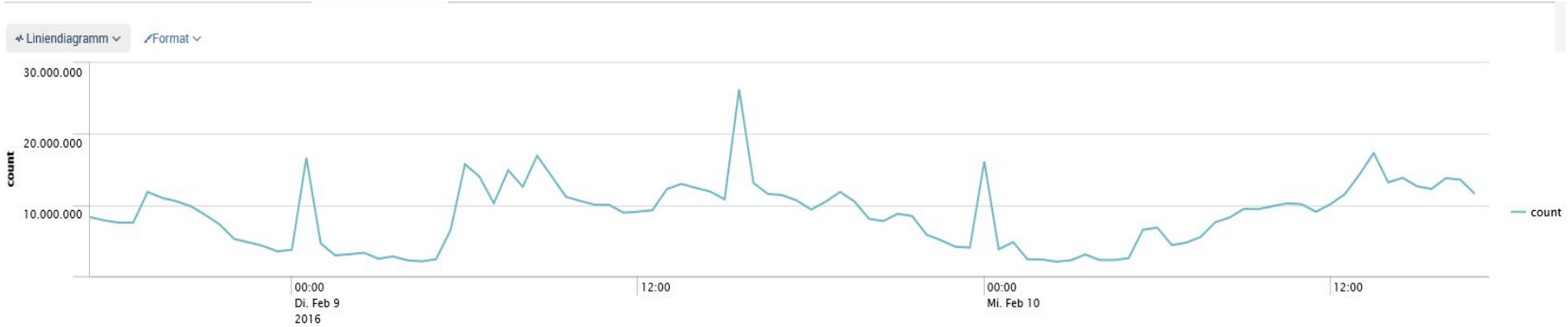
The Backend – Historically Grown

The Backend – Some Order Into Chaos



How Can I Get An Alert?

| tstats prestats=t span=1h count where index="_*" OR index="*" by _time | timechart span=1h count

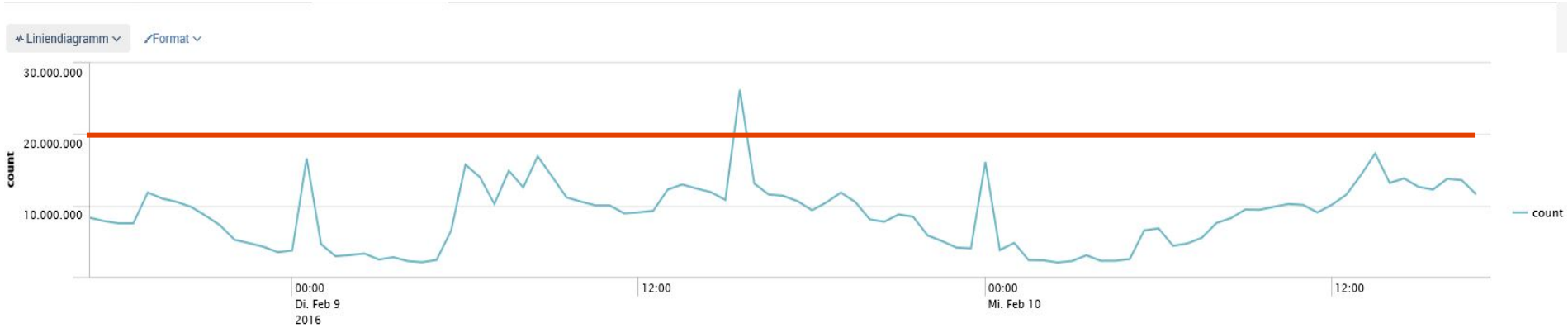


Start with:
- Unqualified eventcounts

Qualify them to:
- Transactions from payment provider
- Webshop requests
- Social media contacts

How Can I Get An Alert?

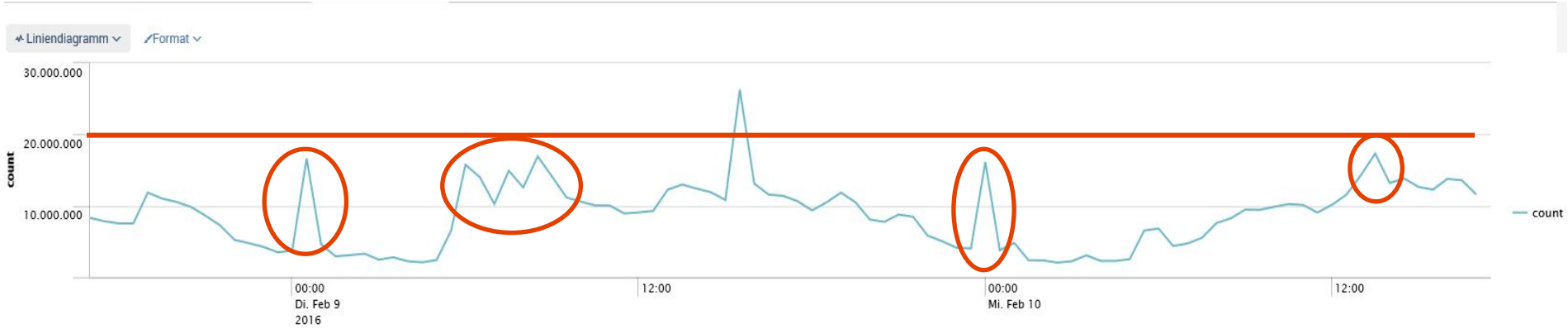
| tstats prestats=t span=1h count where index="_*" OR index="*" by _time | timechart span=1h count | eval threshold=20000000



Method of „fierce inspection“

- Needs a lot of human resources

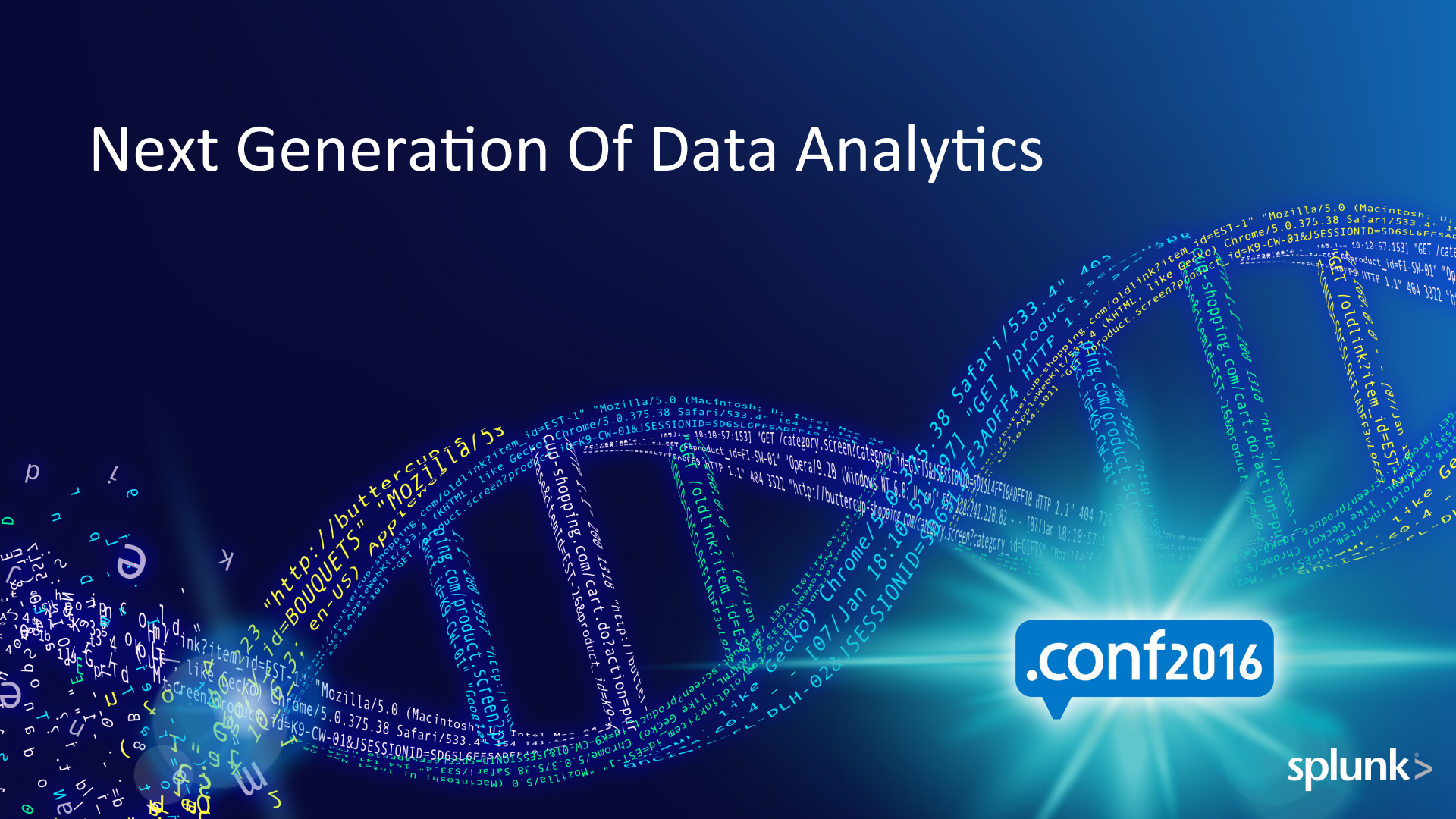
And What About All The Other Peaks?



Peaking and plunging affects your business

- Transactions from payment provider
- Webshop requests
- Social media contacts

Next Generation Of Data Analytics



Data Science – The Science

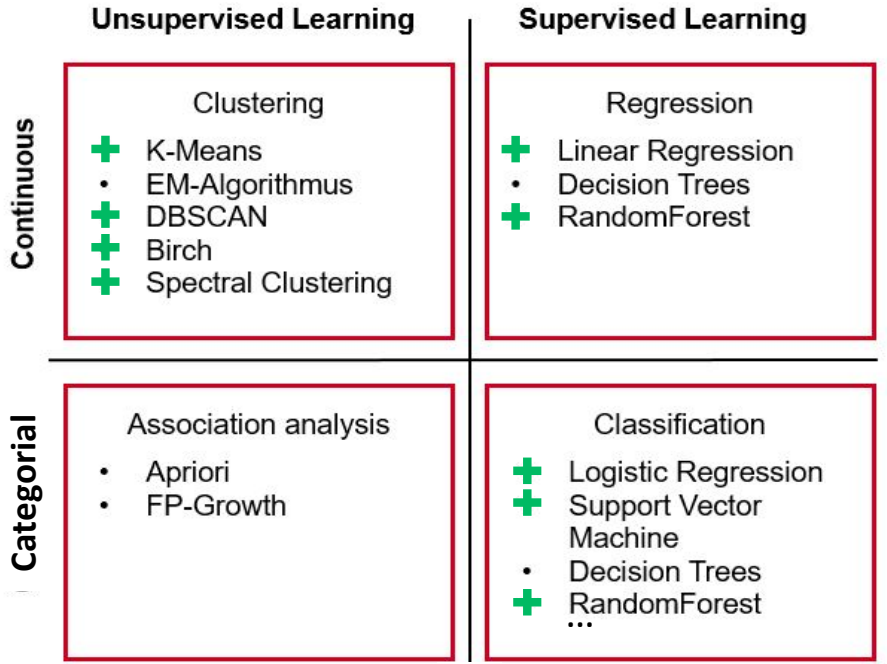
"..., [we are] left with only one option, [we] gonna have to science the shit out of this." (Marc Watney, SOL 71)

OTTO

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

LC SYSTEMS

splunk>



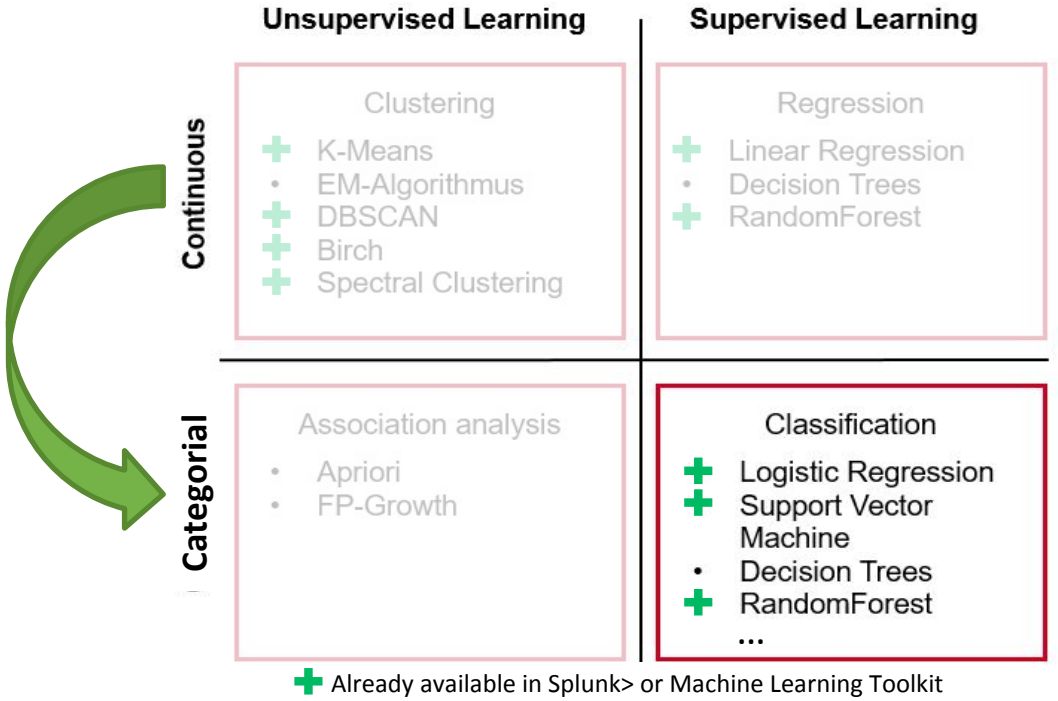
Already available in Splunk or Machine Learning Toolkit

Data Science – The Science

For beginners:
- Transform to categorial
- Use statistic functions of splunk

ML Toolkit used for thesis:
- Known algorithms
- Great variaty of algorithms

Train only with a sample (70%):
- Overfitting is a risk
- Model could get too accustomed to data



Data Science – The Science

Documentation of ML Toolkit:

http://docs.splunk.com/Documentation/MLApp/latest/User/

Especially the commands:

```
| sample rate=0.7
| fit "MyModel" …
| apply "MyModel" …
```

Evaluate the results to rank the algorithms:

```
| `confusionmatrix("behavior","predicted(behavior)")`
```

	Logistic Regression	Support Vector Machine	RandomForest Classifier
Accuracy	0,92 (92 %)	1,00 (100 %)	1,00 (100 %)
Classification Error	0,08 (8%)	0,00 (0 %)	0,00 (0 %)
Precision	0,48 (48 %)	1,00 (100 %)	1,00 (100 %)
Recall	1,00 (100 %)	1,00 (100 %)	1,00 (100 %)
F-1	0,65 (65%)	1,00 (100 %)	1,00 (100 %)

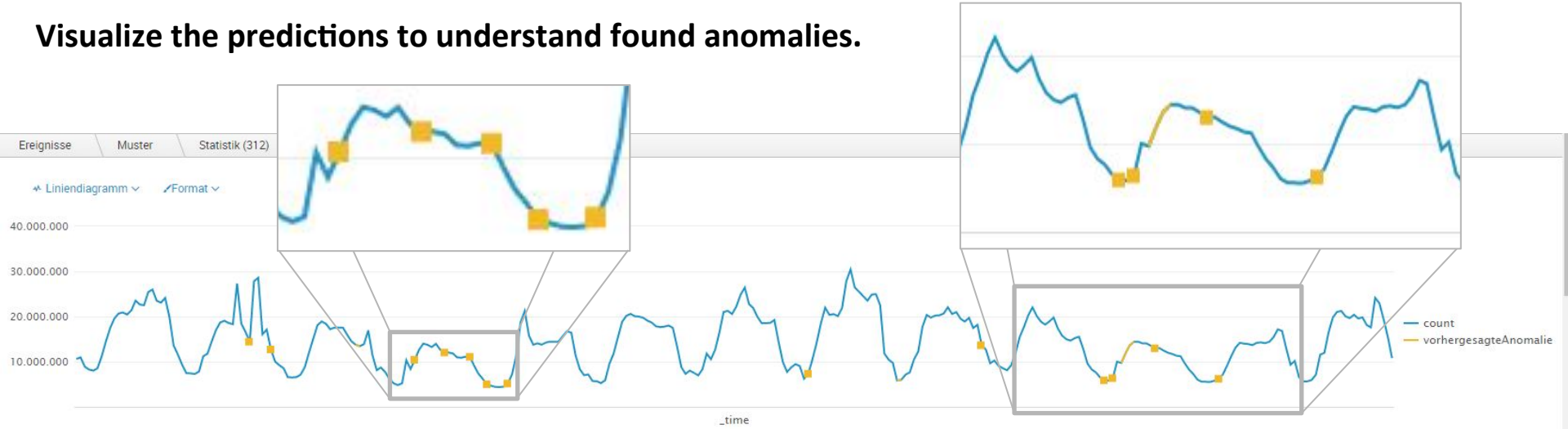
Results Predictionaccuracy (on trained data)

	Logistic Regression	Support Vector Machine	RandomForest Classifier
Accuracy	0,95 (95 %)	0,93 (93 %)	0,99 (99 %)
Classification Error	0,05 (5 %)	0,07 (7 %)	0,01 (1 %)
Precision	0,59 (62 %)	undefinierbar	1,00 (100 %)
Recall	1,00 (100 %)	0,00 (0 %)	0,91 (91 %)
F-1	0,74 (74 %)	0,00 (0 %)	0,95 (95 %)

Results Predictionaccuracy (on test data)

Data Science – The Science

Visualize the predictions to understand found anomalies.




```
| eval predictedAnomaly=if('predicted(behavior)'=="Anomaly", count, null)
| table _time, count, predictedAnomaly
```

Data Science – The Process

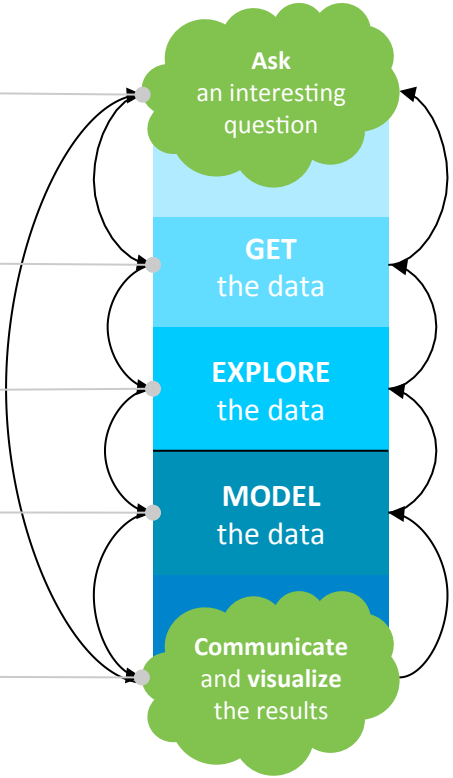
What is the scientific **goal?**
What would you do if you had all **data?**
What do you want to **predict** or **estamate**?

How was the data **sampled**?
Which data is **relevant**?
Are there **privacy** issues?

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Build a model.
Fit the model.
Validate the model.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?



The Data Science Process

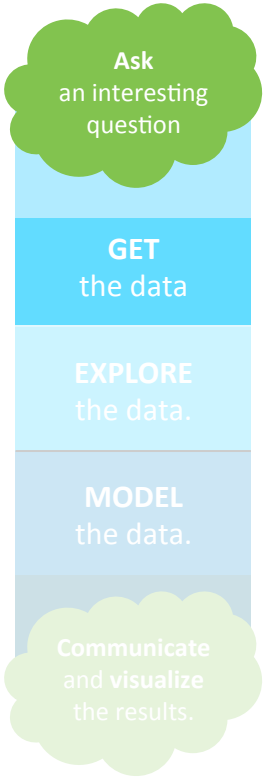
Translated to Splunk Technology

- What type of information is available?
- Use Case
- Business Case
- Predict or estamate
- How to present the data?

The Data Science Process

Translated to Splunk Technology

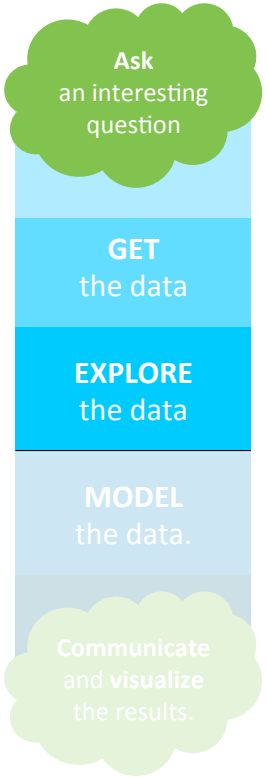
- Location of the data
  - Log Files –> Universal Forwarder
  - Databases –> DB Connect
  - Applications -> http Collector, REST
  - Other sources –> REST, HEC, SYSLOG…
- Verify the data
  - Field-Extractions
  - CIM compliance
- Permissions
  - Users / Roles
  - Indexes



The Data Science Process

Translated to Splunk Technology

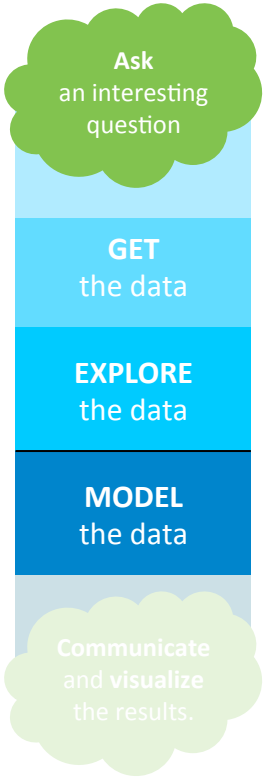
- Plot the data
  - Identify patterns with the patterns tab
  - Searching for anomalies
    - correlate
    - associate
    - analyzefields
    - cluster / kmeans
    - anomalies
    - anomalousvalue



The Data Science Process

Translated to Splunk Technology

- Splunk certified App
  - Machine Learning Toolkit and Showcase
    - need Splunk 6.4
    - need Python for Scientific Computing
  - Splunk IT Service Intelligence
- 3rd party vendors
  - Prelert Anomaly Detective® App for Splunk®

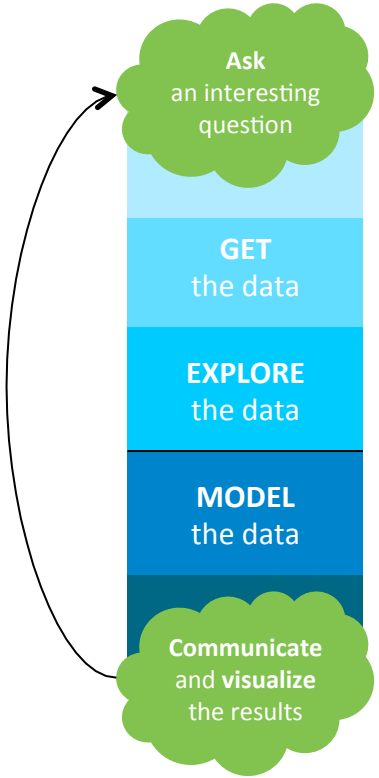


The Data Science Process

Communicate and visualize

- Use Splunk Apps Visualization
- Create your own Views and Dashboards
- Talk to the requester
- Verify outcome
- Deploy to production

- Estimate the knowledge
- If needed restart the loop and include the new findings

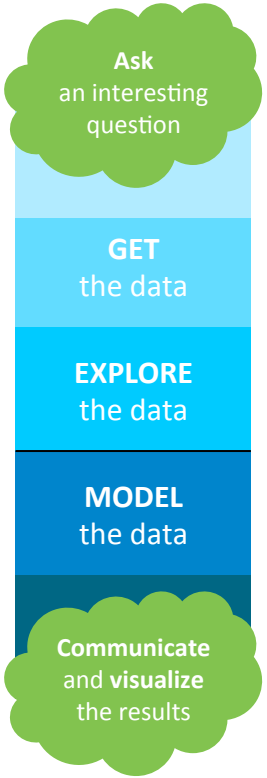


Results

Data Science – The Process

Technologies that have been analyzed

- Prelert Anomaly Detective® App for Splunk®
- Machine Learning Toolkit and Showcase
  - need Splunk 6.4
  - need Python for Scientific Computing
- Splunk IT Service Intelligence



Main Findings Of The Collaboration

Companies have a huge interest in identifying anomalies, but most of them are still in the process of understanding and preparing their data.

C. Günther (LC Systems)

You need a baseline to use your data for anomaly detection, because you have to define whats "normal".

M. Borner (LC Systems)

A simple approach to anomaly detection is to use the static methods integrated in splunk>. If you want to use ML algorithms for that, the ML Toolkit or the Anomaly Detective App (Prelert) is recommended.

P. Drieger (Splunk)

Thanks To

Hamburg University of Applied Sciences - Prof. Dr. Schoeneberg, Niklas Netz

LC Systems - Christian Günther, Mika Borner

splunk> - Philipp Drieger, Holger Sesterhenn

What Now?

Related breakout sessions and activities…

- Splunk for Donuts – Intermediate – Thurstday|2:35PM
- Solve Big Problems with ML – Advanced – Thursday|1:30PM
- Predicting Incidents with ML – Advanced – Thursday|2:35PM

- Splunk UBA – A Data Scientist in a Box – Beginner – Tuesday
- A (VERY) Brief Introduction to ML – Beginner – Wednesday
- Demystifying ML and Anomaly Det. – Intermediate - Wednesday

THANK YOU

splunk>

Backup

Architecture splunk>