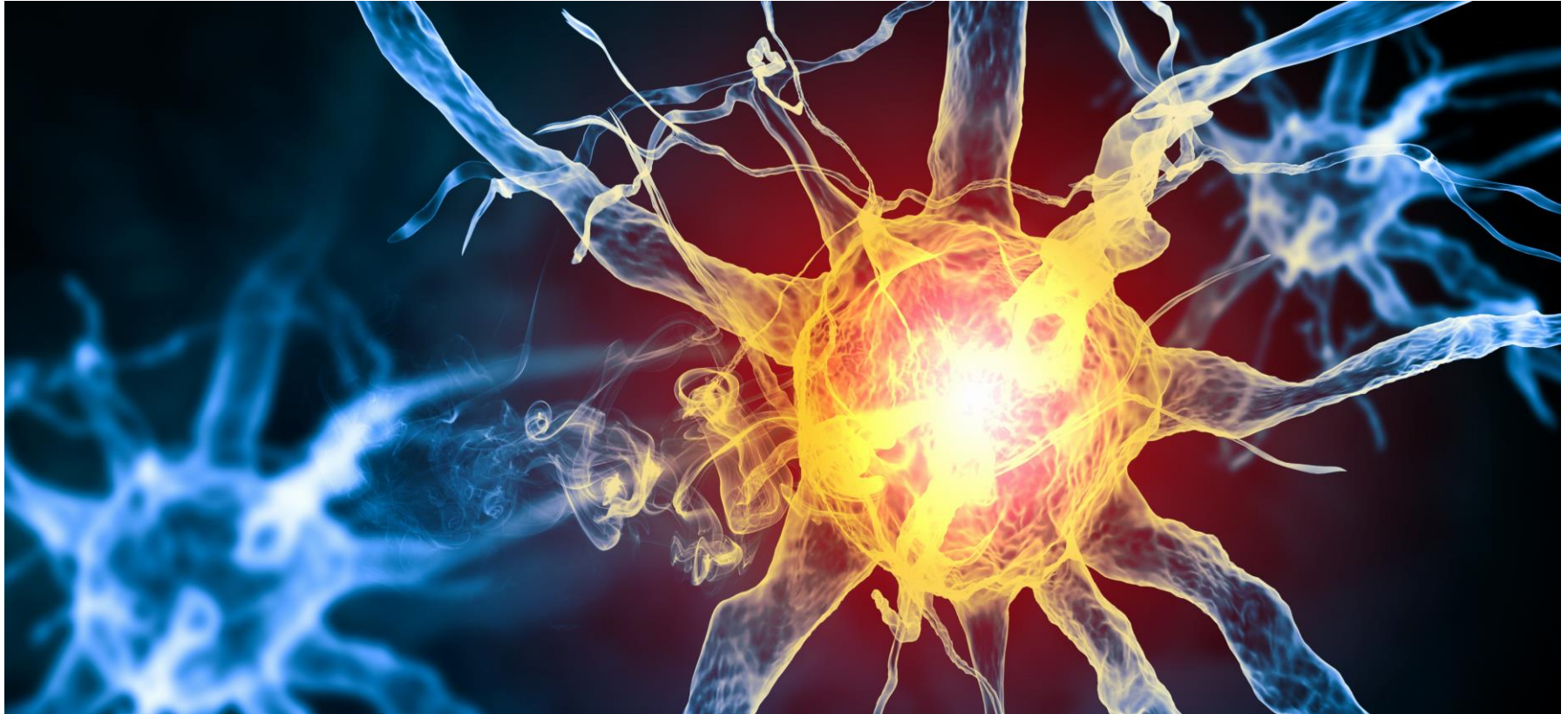


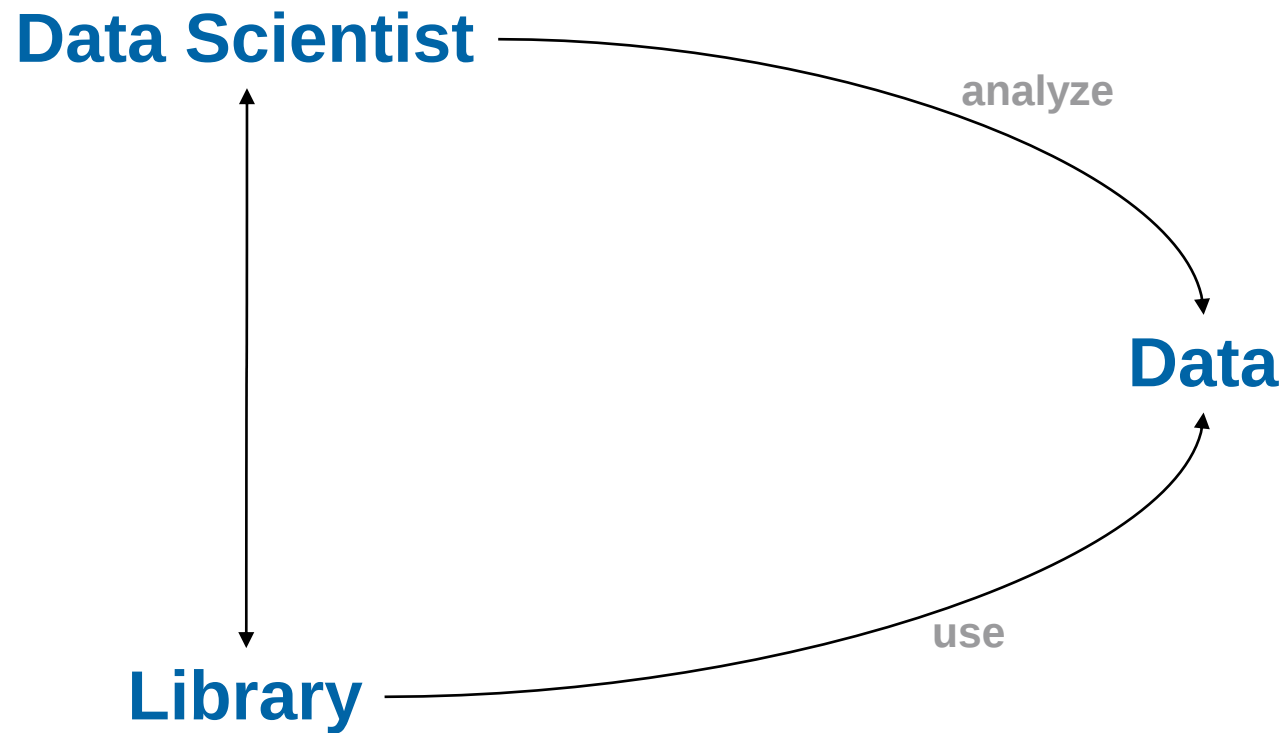
Computer-Based Text- and Data Analysis

Technologies and Applications



Mark Cieliebak

9.6.2015



About Me

Mark Cieliebak

- + Software Engineer & Data Scientist
- + PhD in Computer Science
- + CIO at Netbreeze (now Microsoft)
- + >30 scientific publications

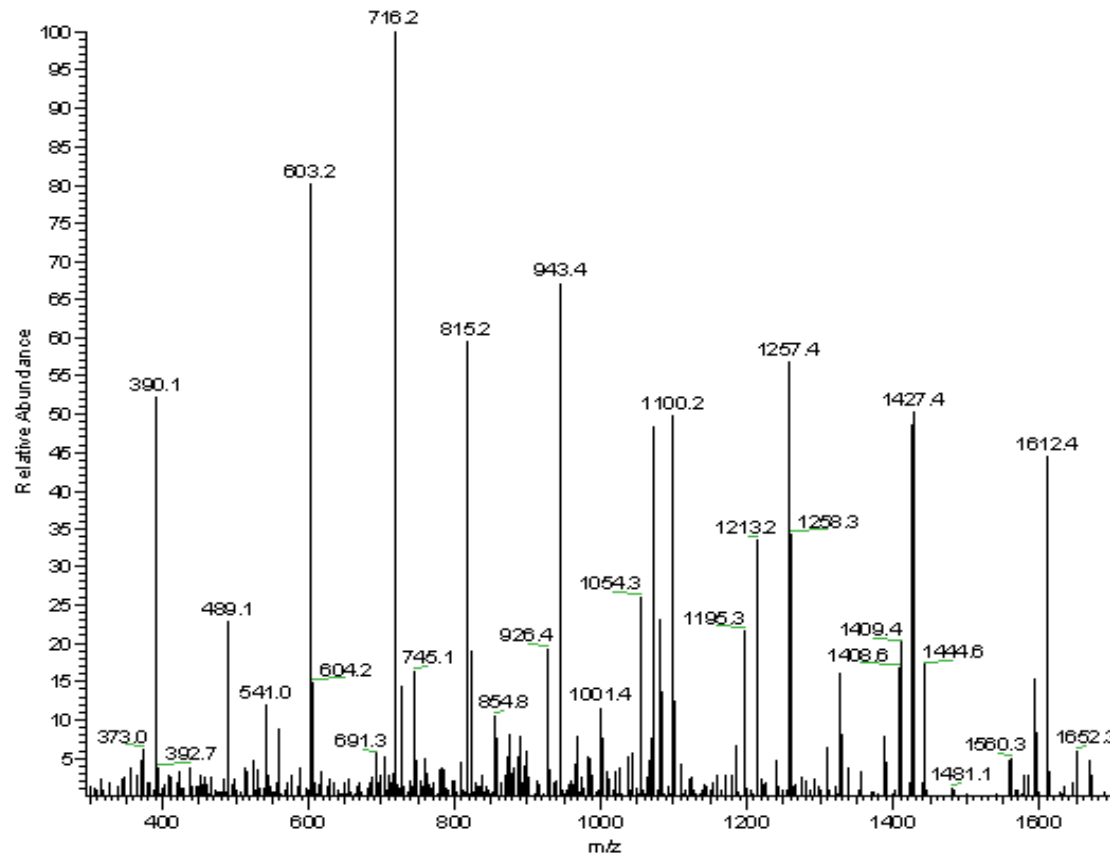


Lecturer at ZHAW

CEO of SpinningBytes AG



"Classical" Data Analysis



Peptide Sequencing

Uses IT to

retrieve,

store,

search,

count,

calculate,

compare,

visualize

data.

Computer-Based Data Analytics

- Started with Artificial intelligence in the 1960's
- Analyzes **huge amounts of data**
- Uses **Machine Learning** for
 - Pattern Detection
 - Image and Text Classification
 - Predictive Analysis
 - Data Clustering
 - and many more!





DATA ANALYTICS



Foundation of Data Analytics

1960



Memory: 64'000 Bytes
→ 8'589'934'592 Bytes

2015

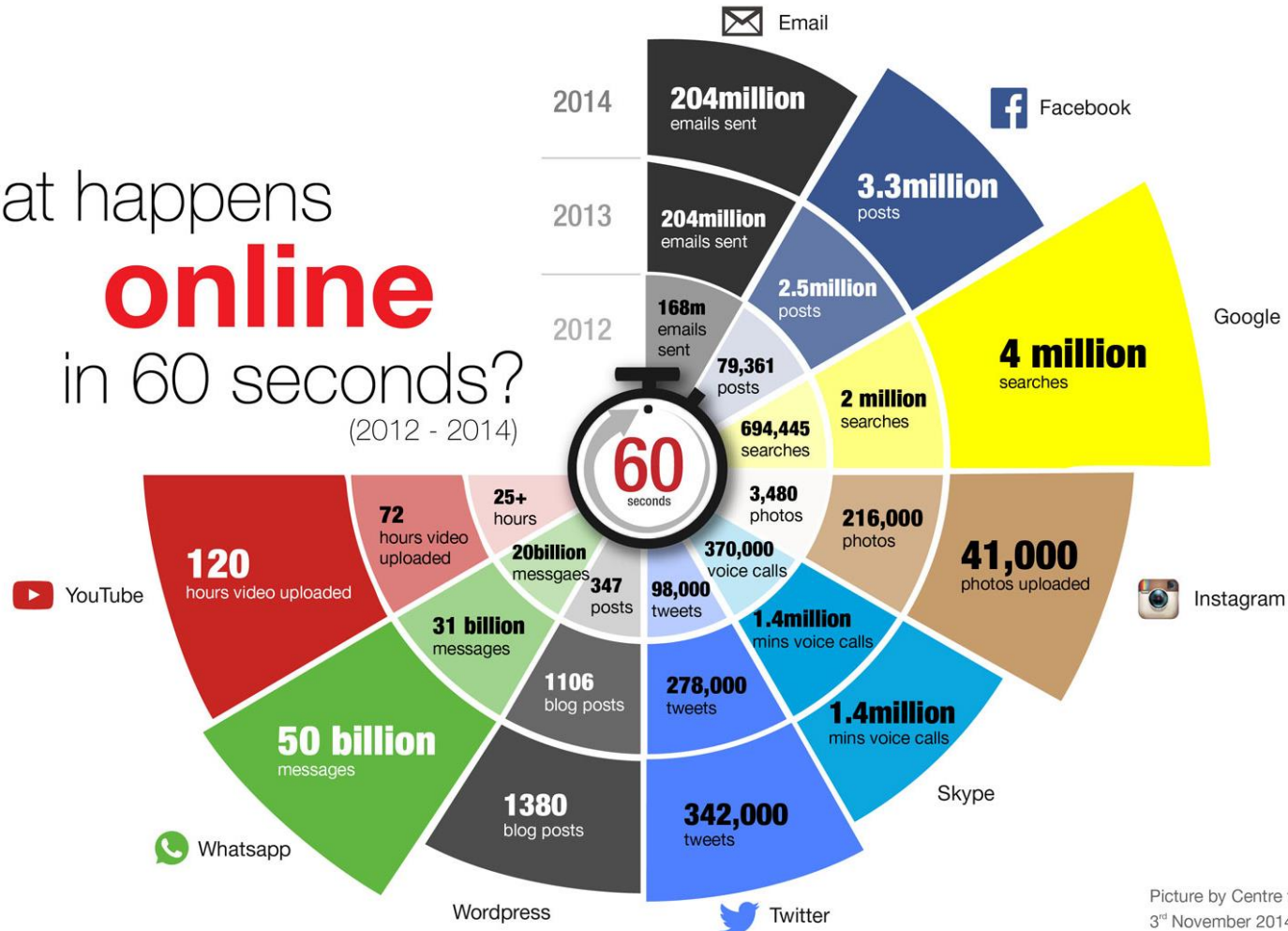


Performance: 500'000 FLOPS
→ 33'863'000'000'000'000 FLOPS

Cost per GFLOP: (1984) \$42'780'000.00
→ \$0.08

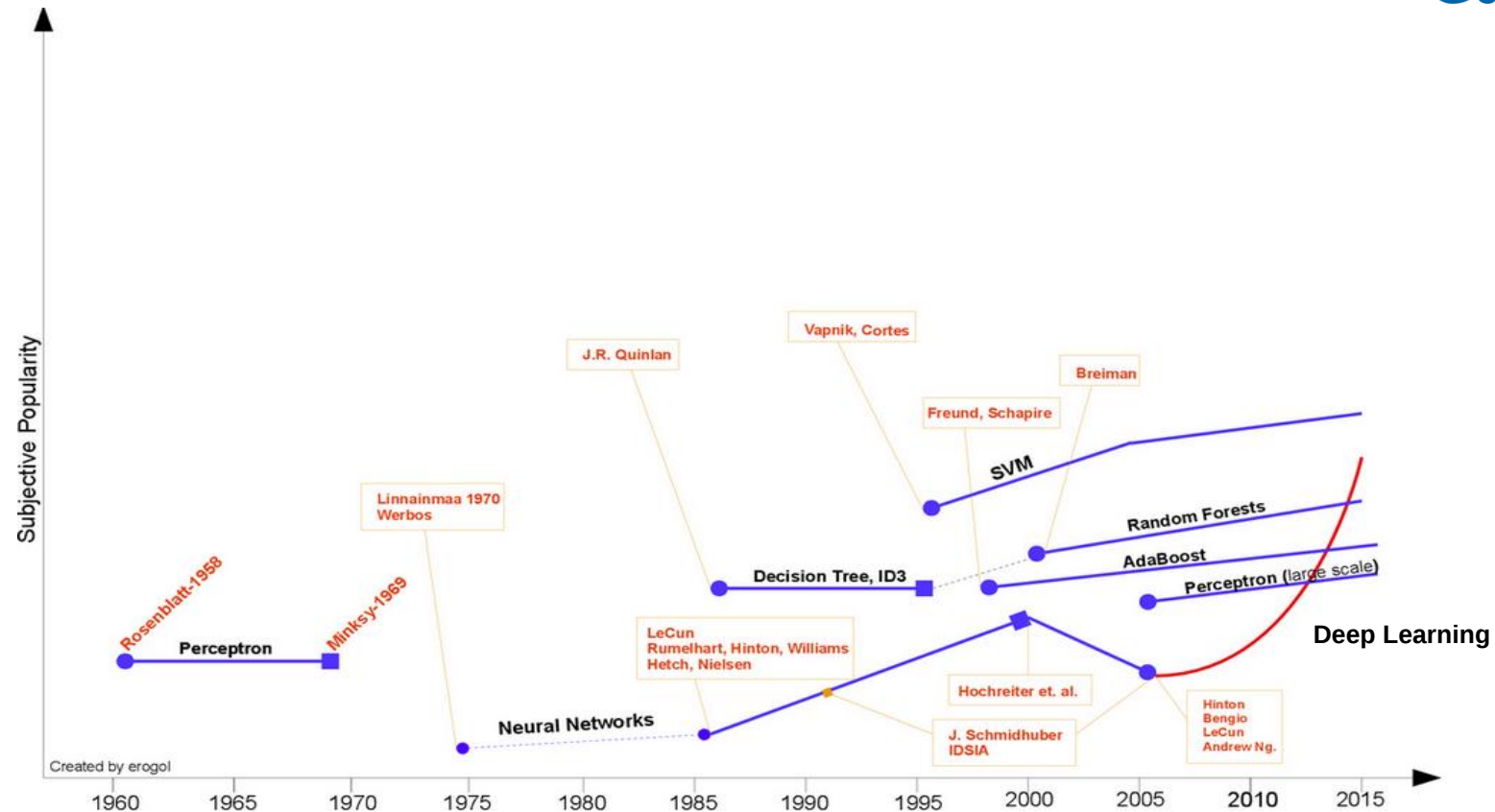
Foundation of Data Analytics

What happens
online
in 60 seconds?
(2012 - 2014)



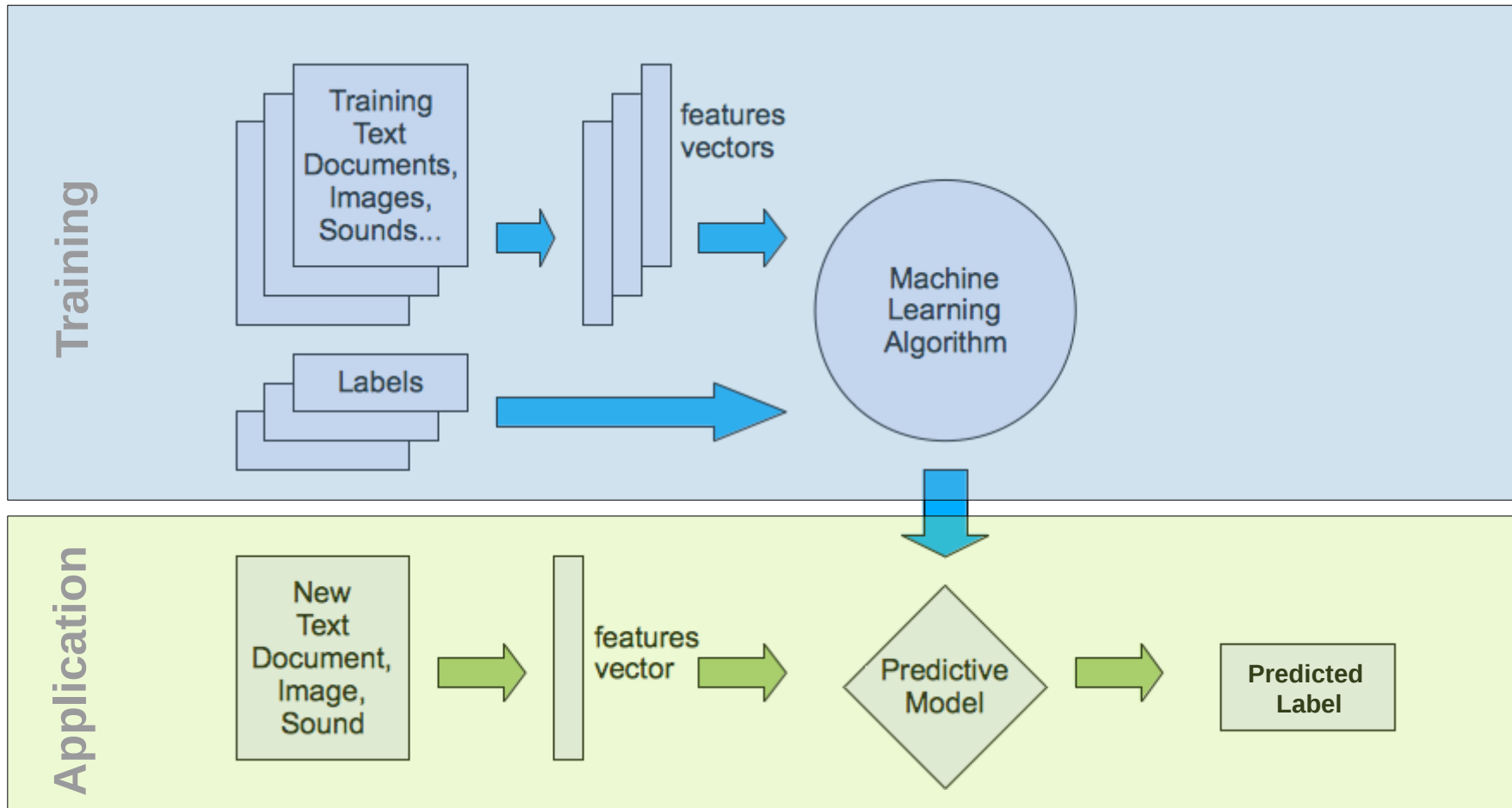
Picture by Centre for Learning and Teaching
3rd November 2014

Foundations of Data Analytics

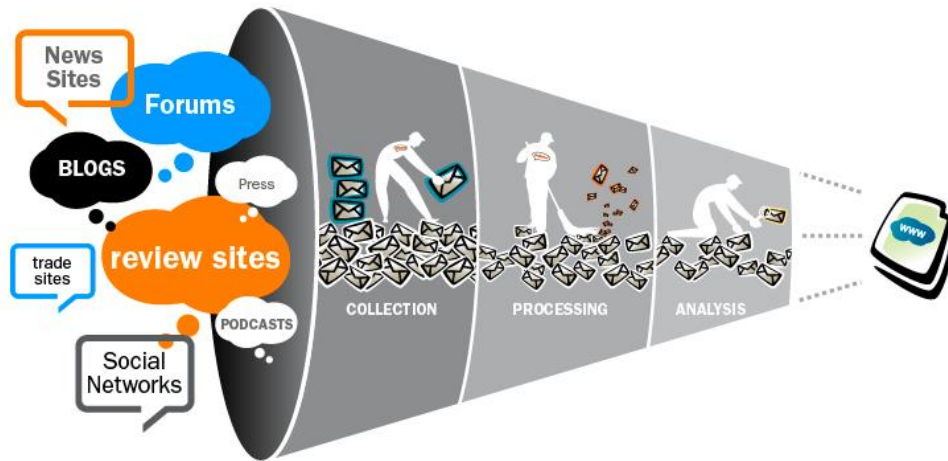


Faster Computers + More Data + Better Algorithms

Machine Learning in a Nutshell



Application: Social Media Analysis

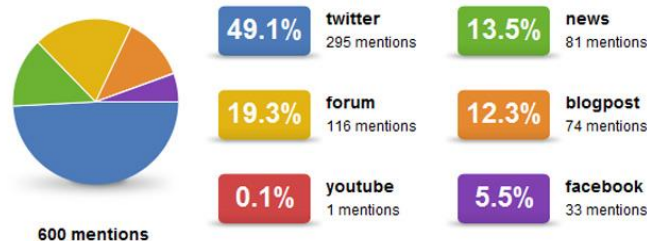
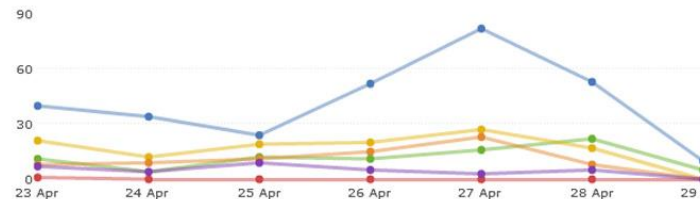


Sentiment Analysis

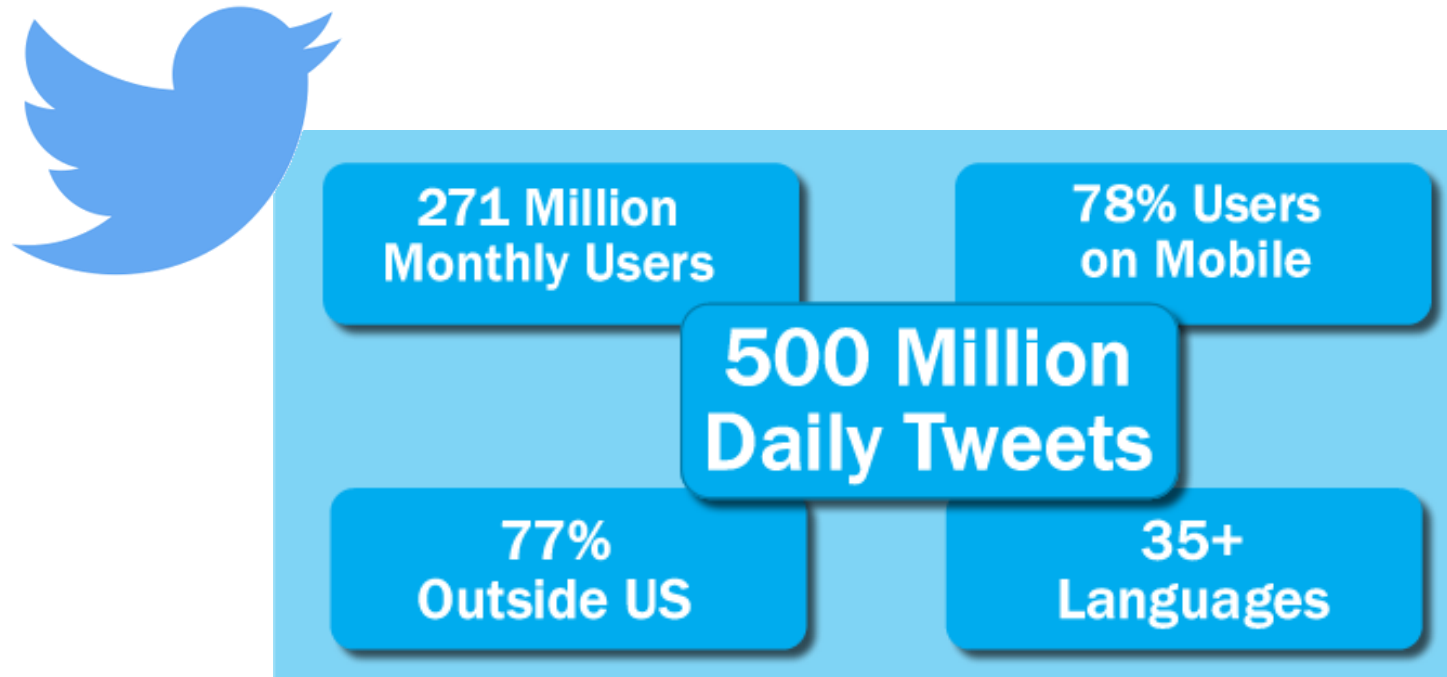
Topic Extraction

Trend Detection

Alerting



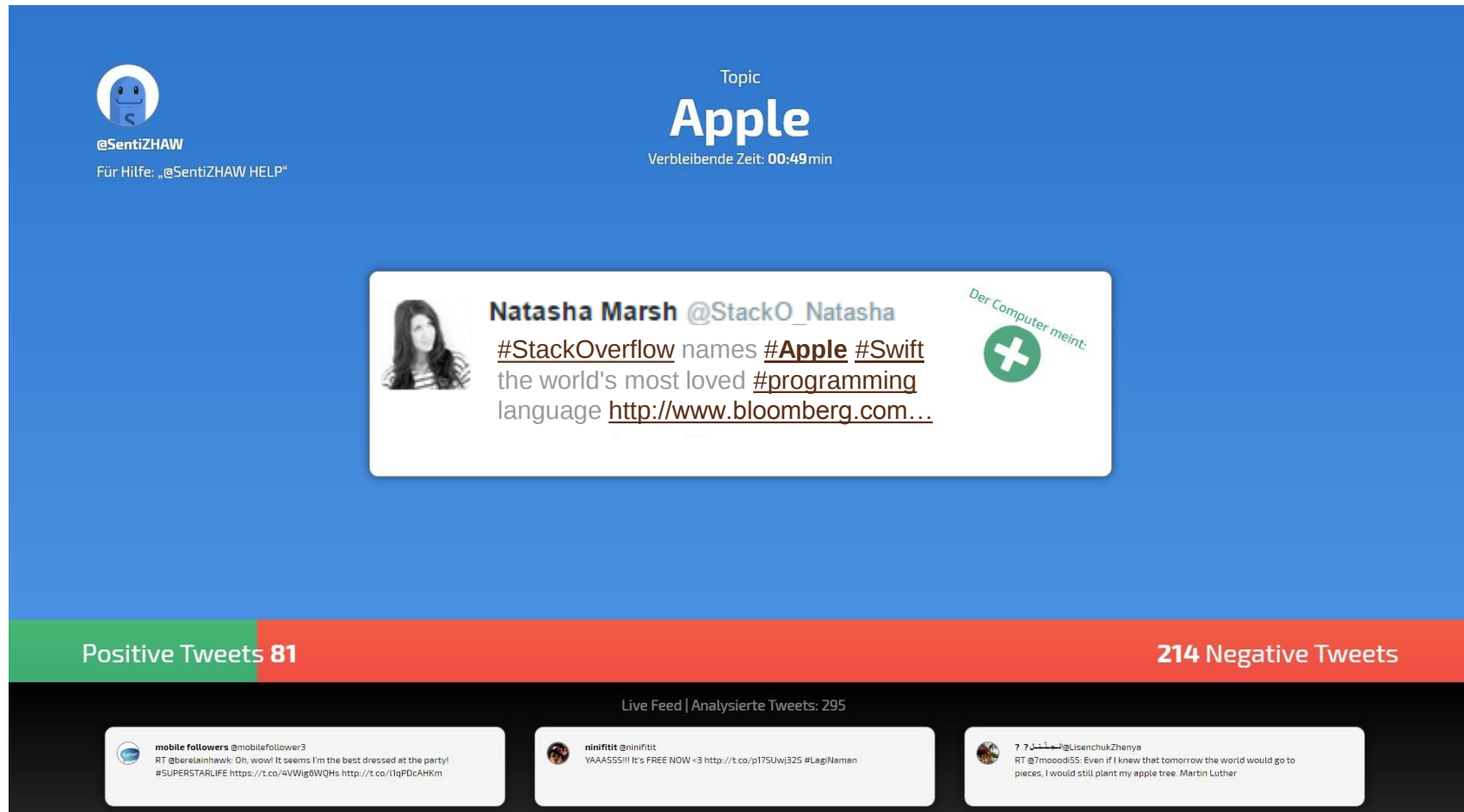
Twitter Facts



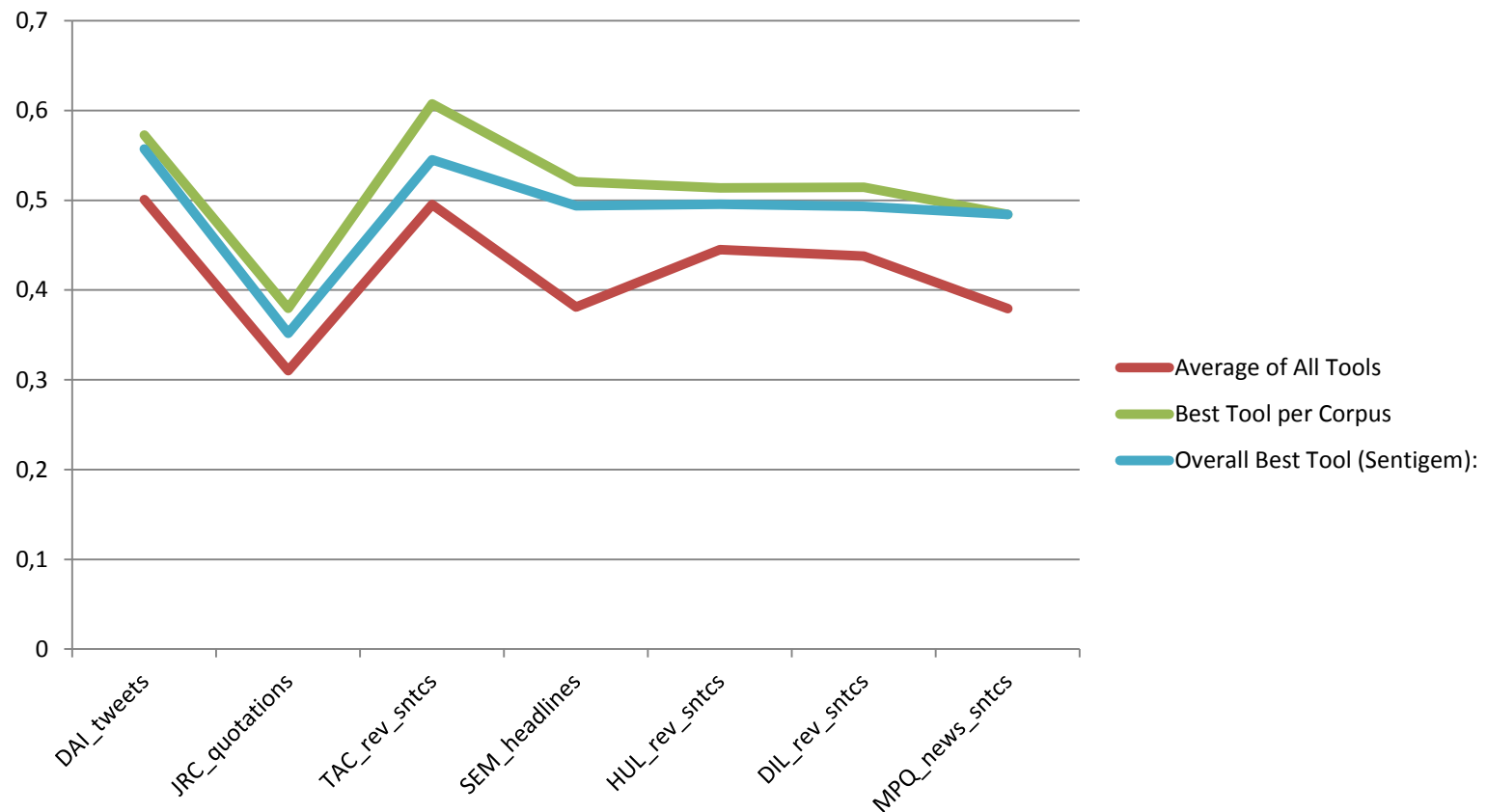
Data Access

- Free Access with Search API
- Commercial Access to Firehose Stream (all or 10%)

Sentiment Analysis on Twitter



Sentiment Analysis: Performance of Commercial Tools (F1-Score)



Experimental Setup:

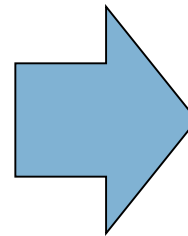
9 commercial sentiment analysis tools evaluated on 7 public corpora with 28653 short texts.

Take-Home Lesson

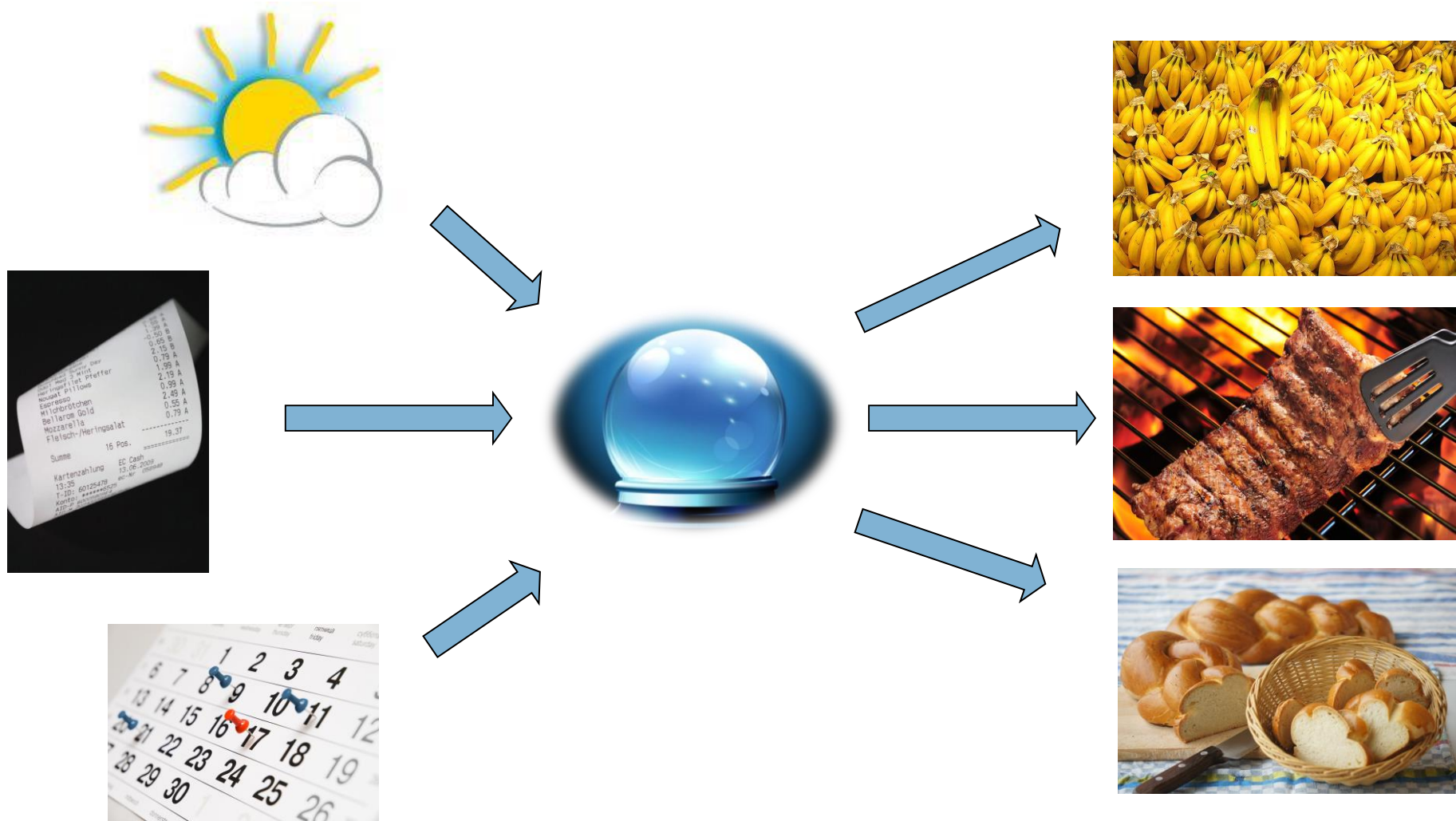


**Sentiment Tools on short texts
achieve on average an
F1-Score of 51%.**

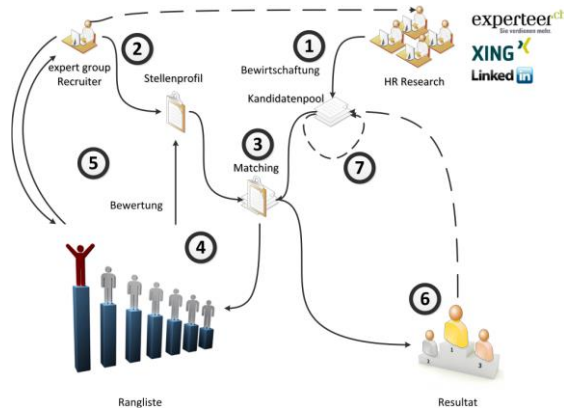
Application: Newspaper Segmenta



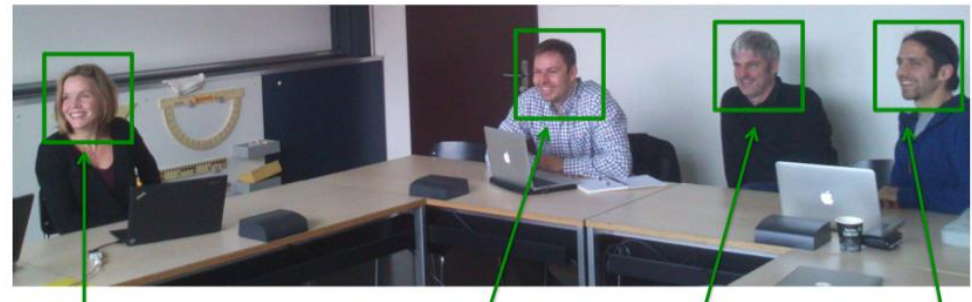
Application: Sales Prediction



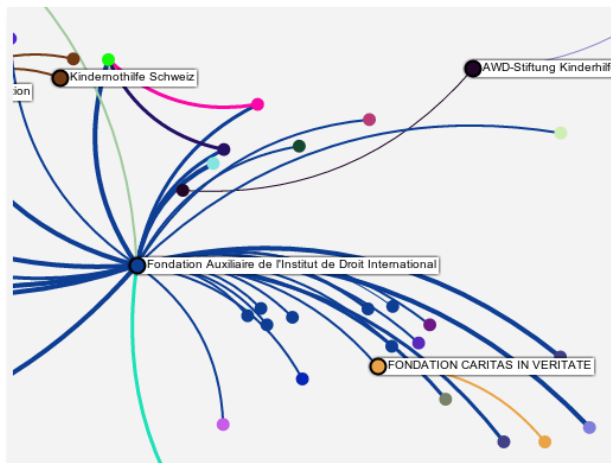
More Applications



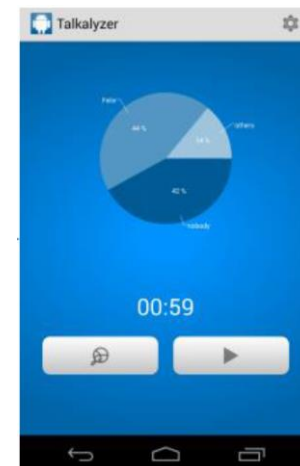
Expert Match



Face/Image Recognition



Foundation Register



Speaker Detection

What do Data Scientists Need?



Data Sources

- Experimental Data
- Reference Datasets
- Dictionaires, Ontologies, Thesaurus
- Scientific Papers
- Social Media
- Media Archives: Newspapers, Magazines
- Websites
- Live Streams: Twitter, Online News, Product Reviews
- Videos/Movies/Pictures
- Books: Belletristic, Technical Literature
- Wikipedia
- Hand-crafted Datasets
- etc. etc.



Data Provisioning

Data Collections:

- Linguistic Data Consortium
- European Language Resource Association
- NIST
- TIMIT
- etc.

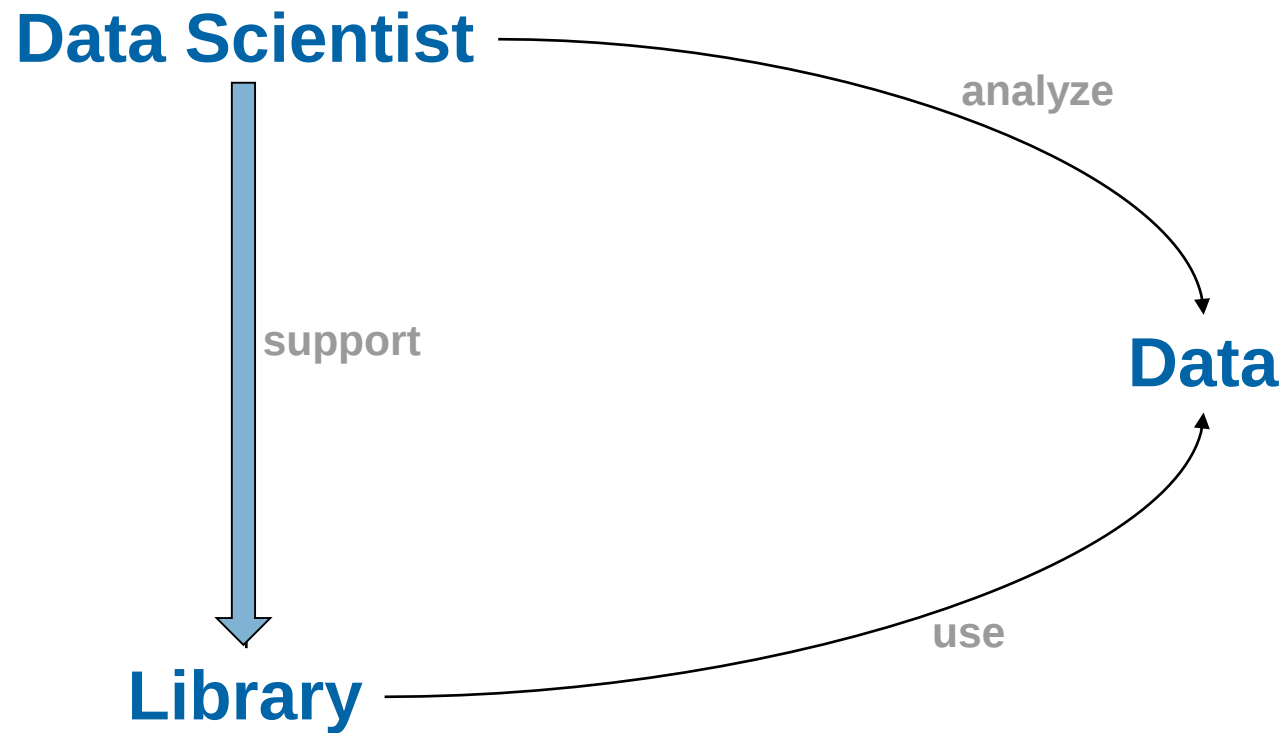
→ Access, Licensing

Open Data:

- Open Governmental Data OGD
- Open Research Data ORD
- Linked Open Data

→ Participation, Guidelines





Digitizing

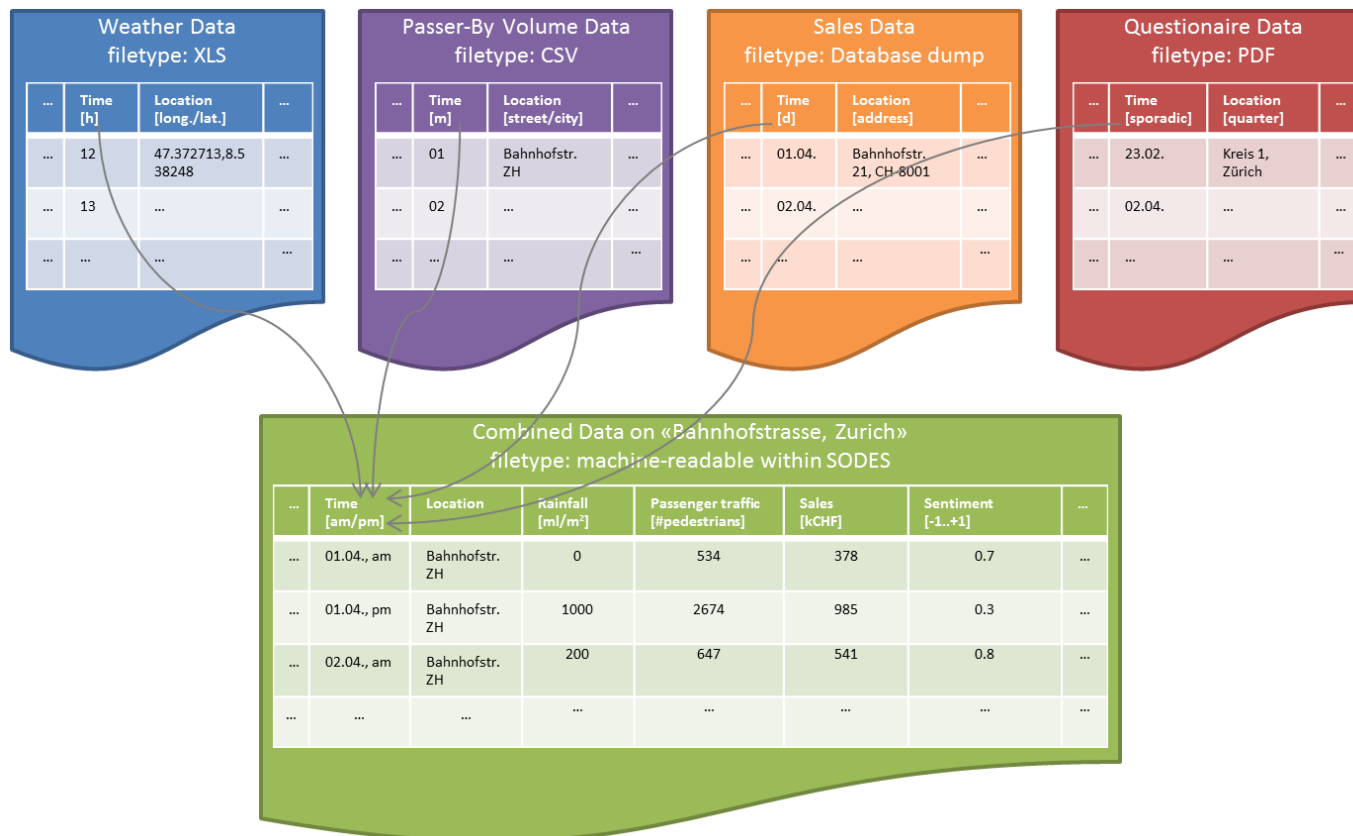
Make Information Accessible!

- Extract Text, Images, Charts, Tables
- Categorize
- Search
- Browse
- (Summarize)

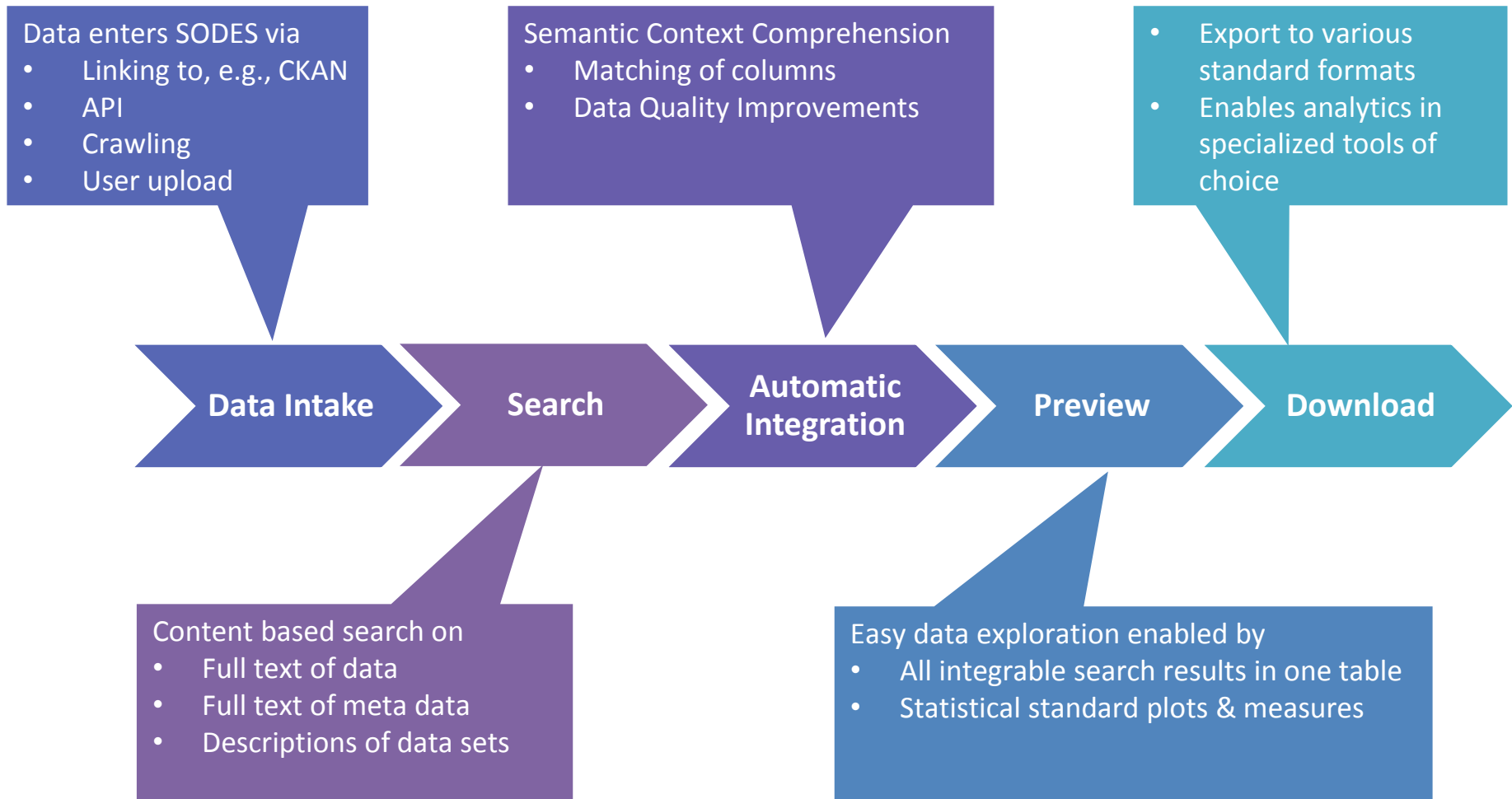


Data Integration

Combining data from different sources is very time consuming!

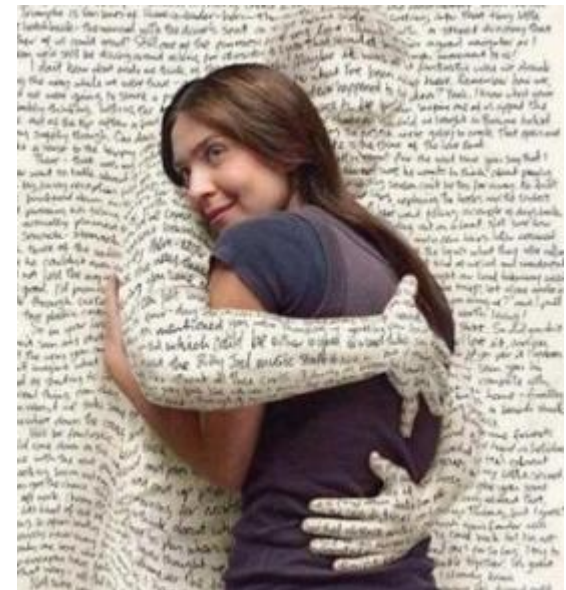


SODES: Automatic Data Integration



Talk in Short

- Text and Data Analytics is successful
- Researcher need access to data
- Data Scientists have powerful tools



Thank You!



Mark Cieliebak

ZHAW – Institute of Applied Information Technology (InIT)

Winterthur, Switzerland

Email: ciel@zhaw.ch, Website: www.zhaw.ch/~ciel