Big Data & Analytics

-

Cost and Value

Valentin Zacharias, codecentric.de

Karlsruhe, 22.5.2014

200 consultants, enthusiasts, engineers, craftsmen, experts, nerds

- build systems to create value from big data
- help businesses scale software
- realize agile software development – with our customer‘s in house development and in custom software development

We are codecentric.

Motivation, **Data Deluge**, Datification, Volume, Variety, Velocity, **Data Driven Decision**, n=All, t=now, **Cost**, **Data OS**, Hadoop, **Enterprise Data Lake**, **Distributed Search**, ElasticSearch, **NoSQL et al**, Cassandra, mongoDB, Riak, Pivotal, cloudera, **Distributed Stream Processing**, Storm, **In Memory Computing**, SAP Hana, Spark, EXASOL, **Reactive Programming**, Moore's Law, Cloud Computing, Data Value Chain, **Value**, **Analytics**, Descriptive, Visual, Causal, Predictive, Prescriptive, **Application Areas**, Operational Excellence, Customer Intimacy, Product Leadership, **360° Everything**, **Business Models**, Data Driven Business, Data for Products, Products with Data, Data as Business

My Goal today:

- You've heard the most important Big Data terms (see above) and can fit them into one coherent picture of the Big Data world
- You know where to start (if you want to)

Motivation

The Data Deluge and

The Need for (more) Data Driven Decision

The Data Deluge

- Cheap networked sensors, social web, digital workflows ... and the Datification of everything lead to:
  - Datasets of much larger size (**Volume**)
  - Datasets with many different formats (**Variety**)
  - Datasets that change fast (**Velocity**)

The need for (more) Data Driven Decision

Competition and the expectation of further progress require the use and fast processing of this data

Further progress in medicine rests on understanding complex relationships and individualized treatments.

Further productivity gains in farming will largely have to come from the optimized use of machines, fertilizer and pesticides

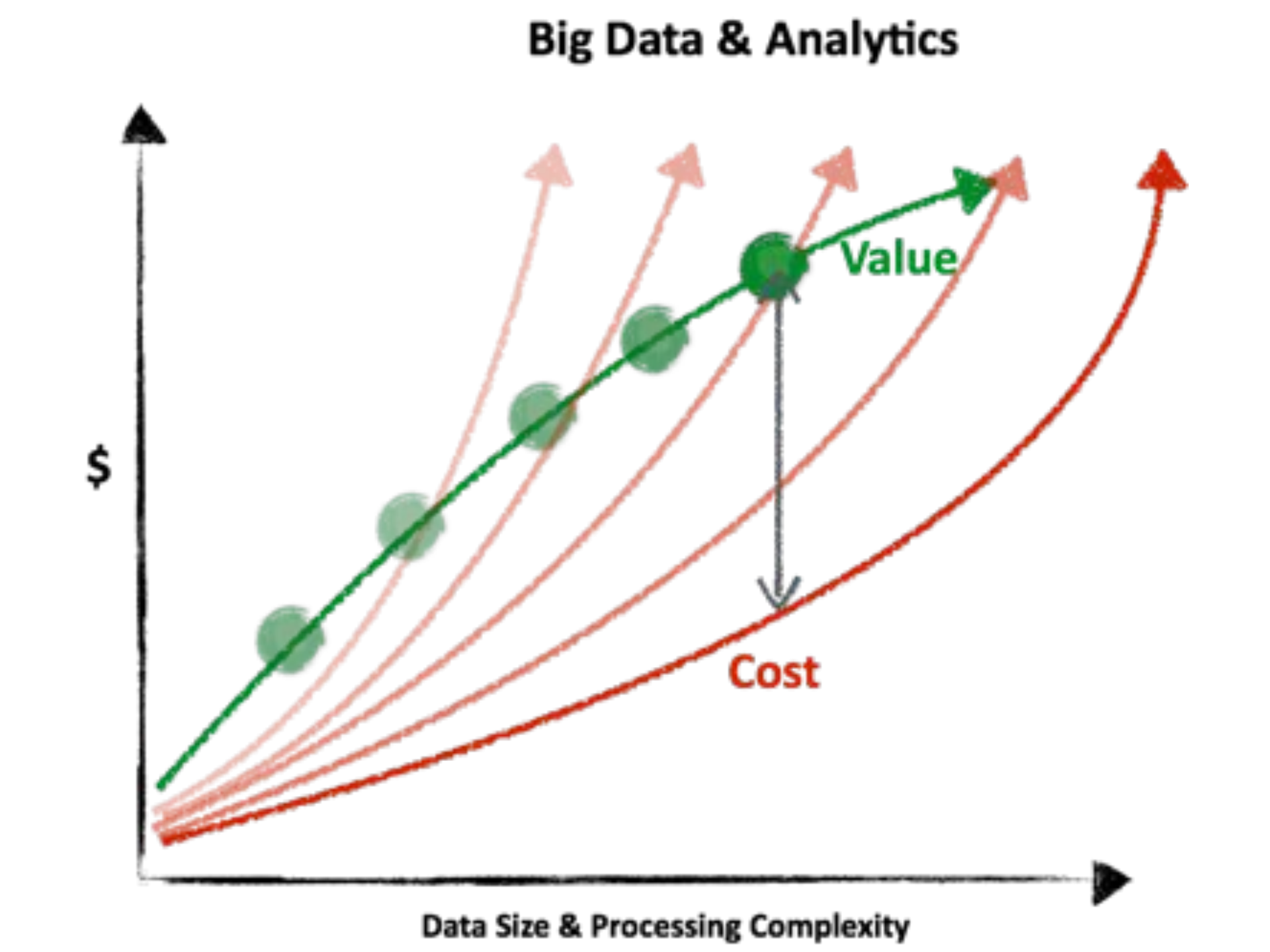
With every product available everywhere at the click of a button, customization products and services to ever smaller customer groups becomes paramount

Unpredictable fluctuation in energy production (and demand) make it necessary to optimize the energy network in realtime.

Innovations in logistics such as “Same-Day-Delivery” rest on real time planning and optimization.

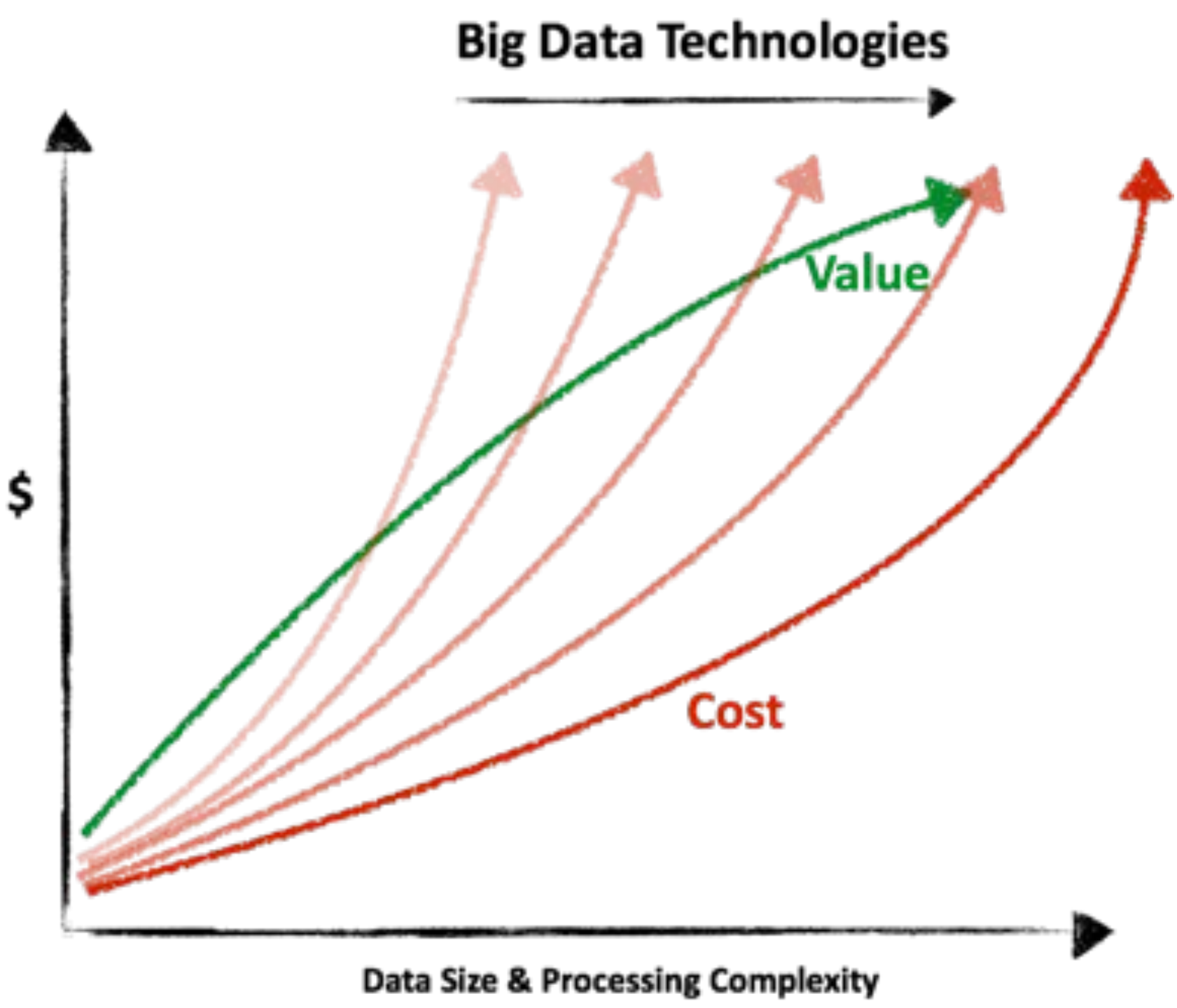
It is no longer enough to plan for averages (time, space, customers), but necessary to optimize for the individual, precise locations and now

n=All & t=now


Big Data & Analytics
$
Value
Cost
Data Size & Processing Complexity

Cost

The Big Data Technologies used to lower the cost to build systems that do more complex processing with more data faster.

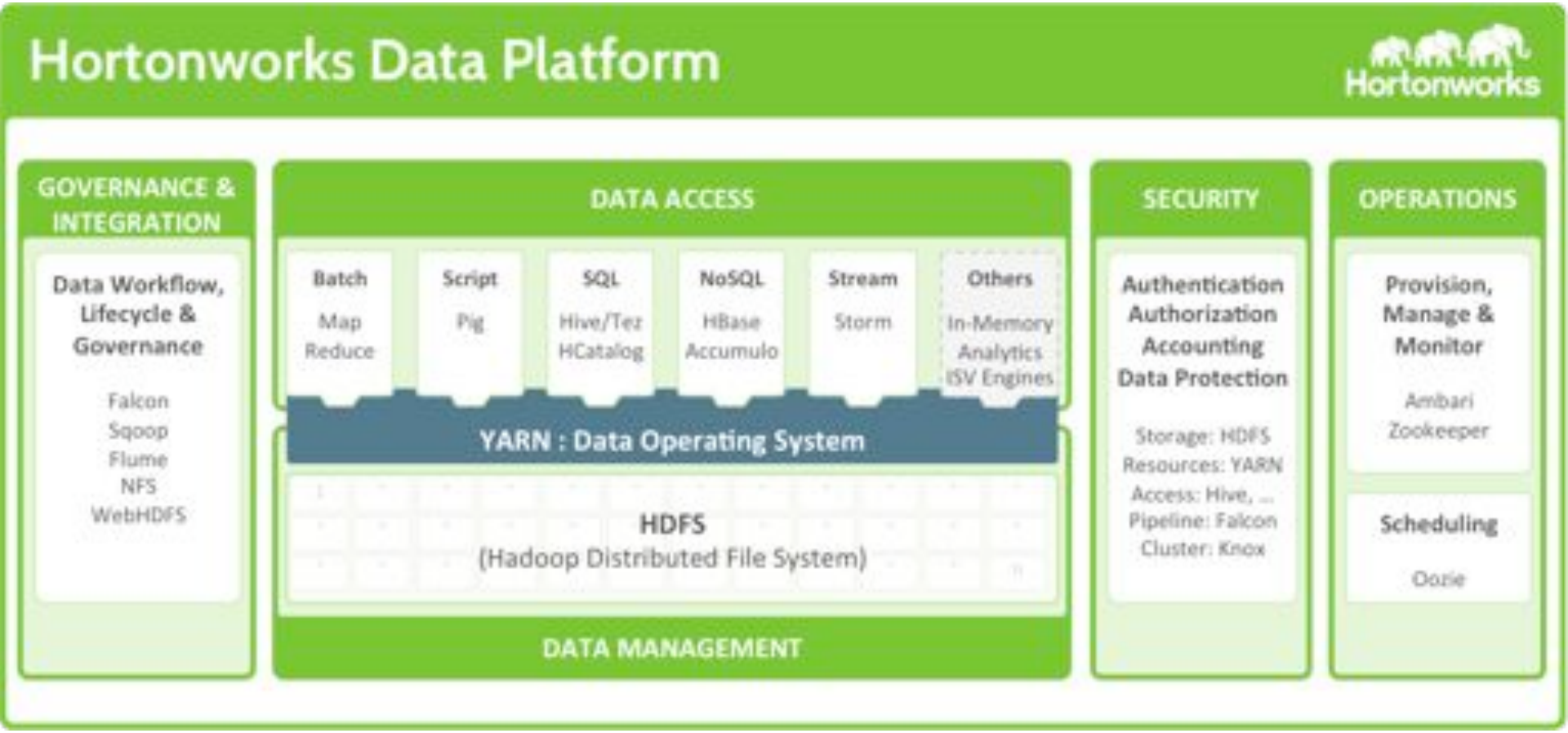

Big Data Technologies
$
Value
Cost
Data Size & Processing Complexity

Volume & Variety

Data OS – Hadoop & Enterprise Data Lake

NoSQL et al.

Distributed Search

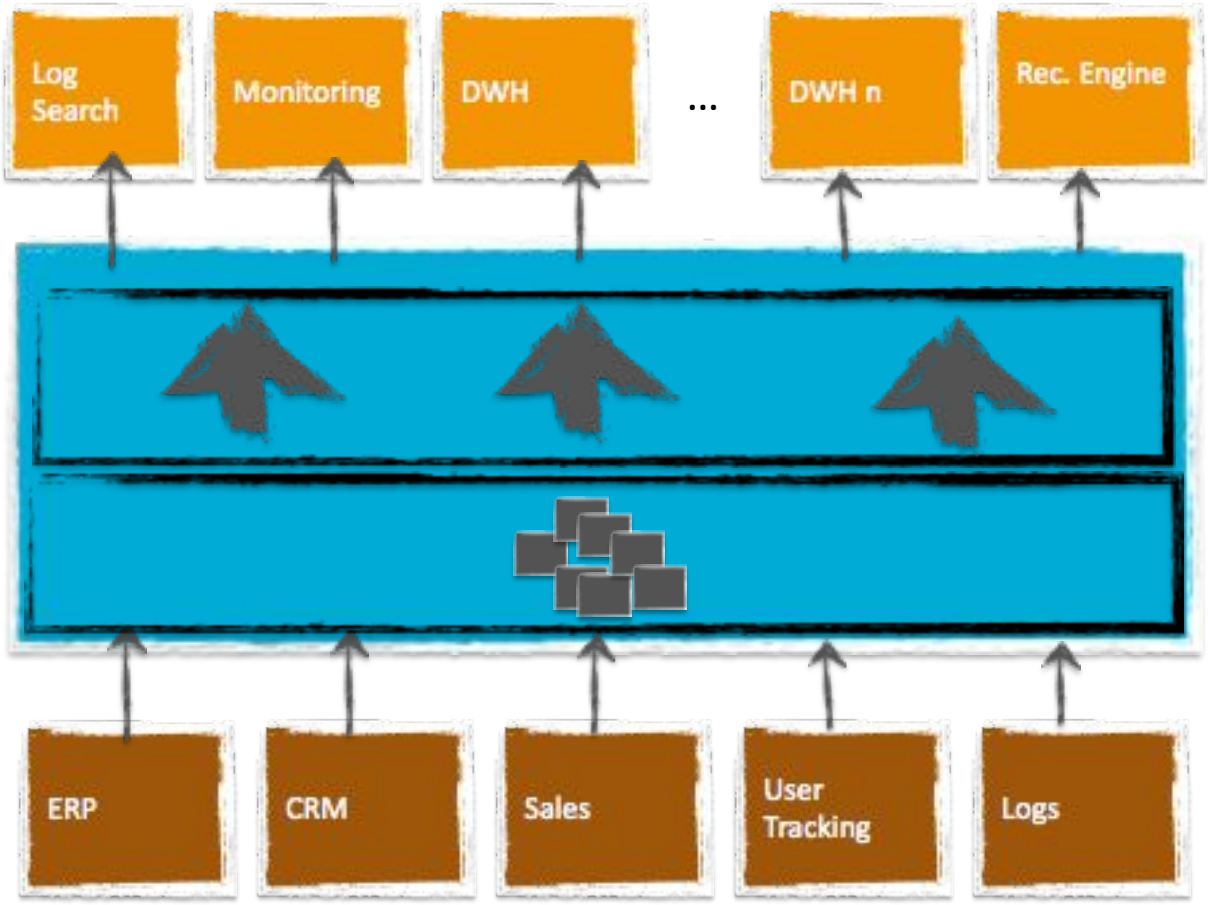


The Hadoop ecosystem provides a ‘data operating system’ with applications for efficient, diverse, very scalable data storage and processing.

	Data Warehouse Appliance	Hadoop
Volume of Data	500 TB	500 TB
System Cost[1]	$22.7	$1.4
Initial Acquisition Cost	$5.5[2]	$0.2[3]
Upgrades at 26% CAGR	$8.4	$0.3
Maintenance/Support[4]	$8.2	$0.2
Power/Space/Cooling	$0.6	$0.7
Admin	$0.8	$0.8
Application Development	$6.6	$7.2
Total Cost of Data	**$30 million**	**$9.3 million**

BigData – what does it really cost?
Winter Corporation

The use of hadoop can radically lower costs for many wharehousing / data storage scenarios



Pivotal

cloudera
Ask Bigger Questions

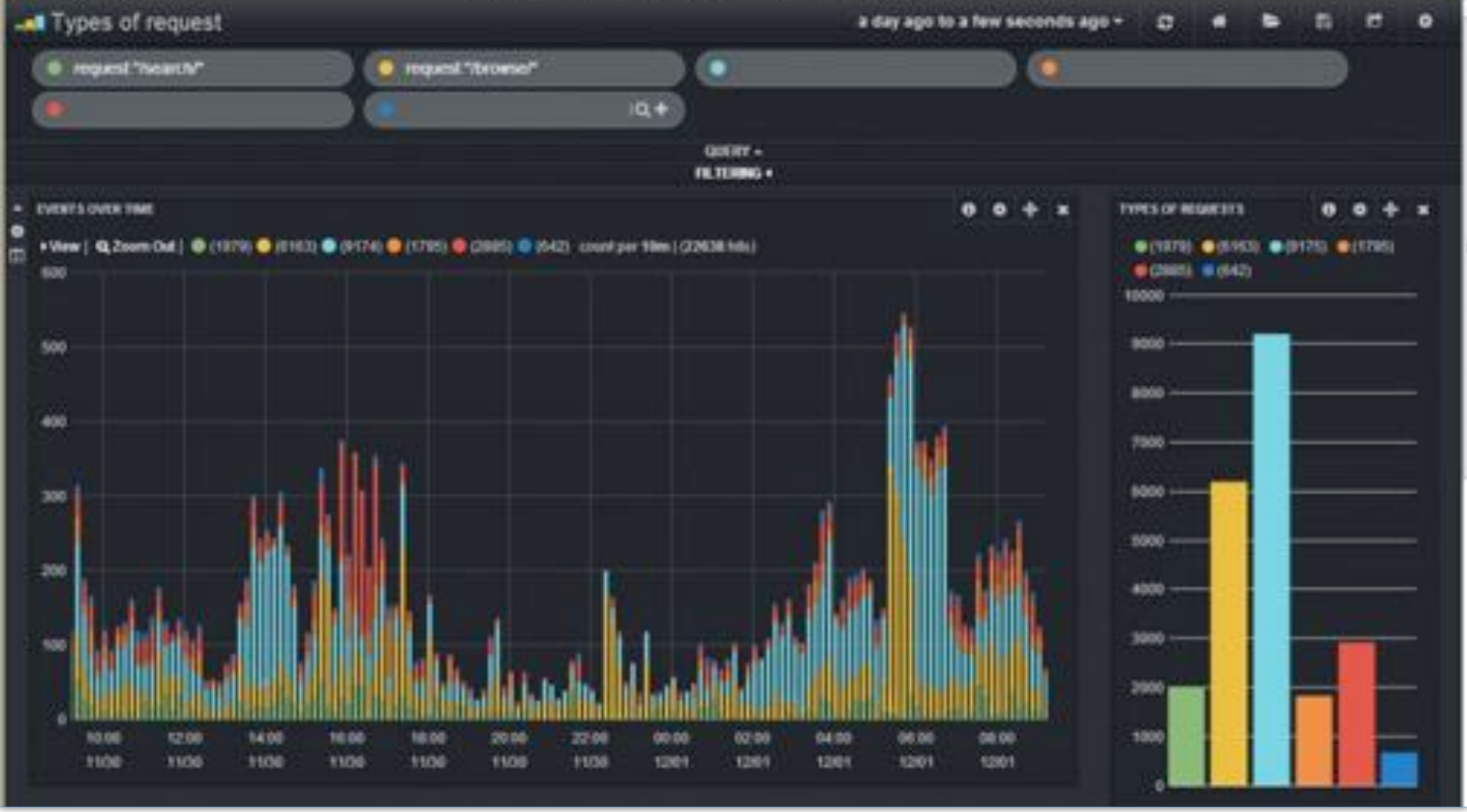
Enterprise Data Lake / Data Hub: The vision to transform business it architecture with Hadoop through a central data store feeding all DWHs – simplifying DWH/ETL architecture and radically speeding up the creation of new reports/dashboards.

Cassandra

mongoDB

riakcs

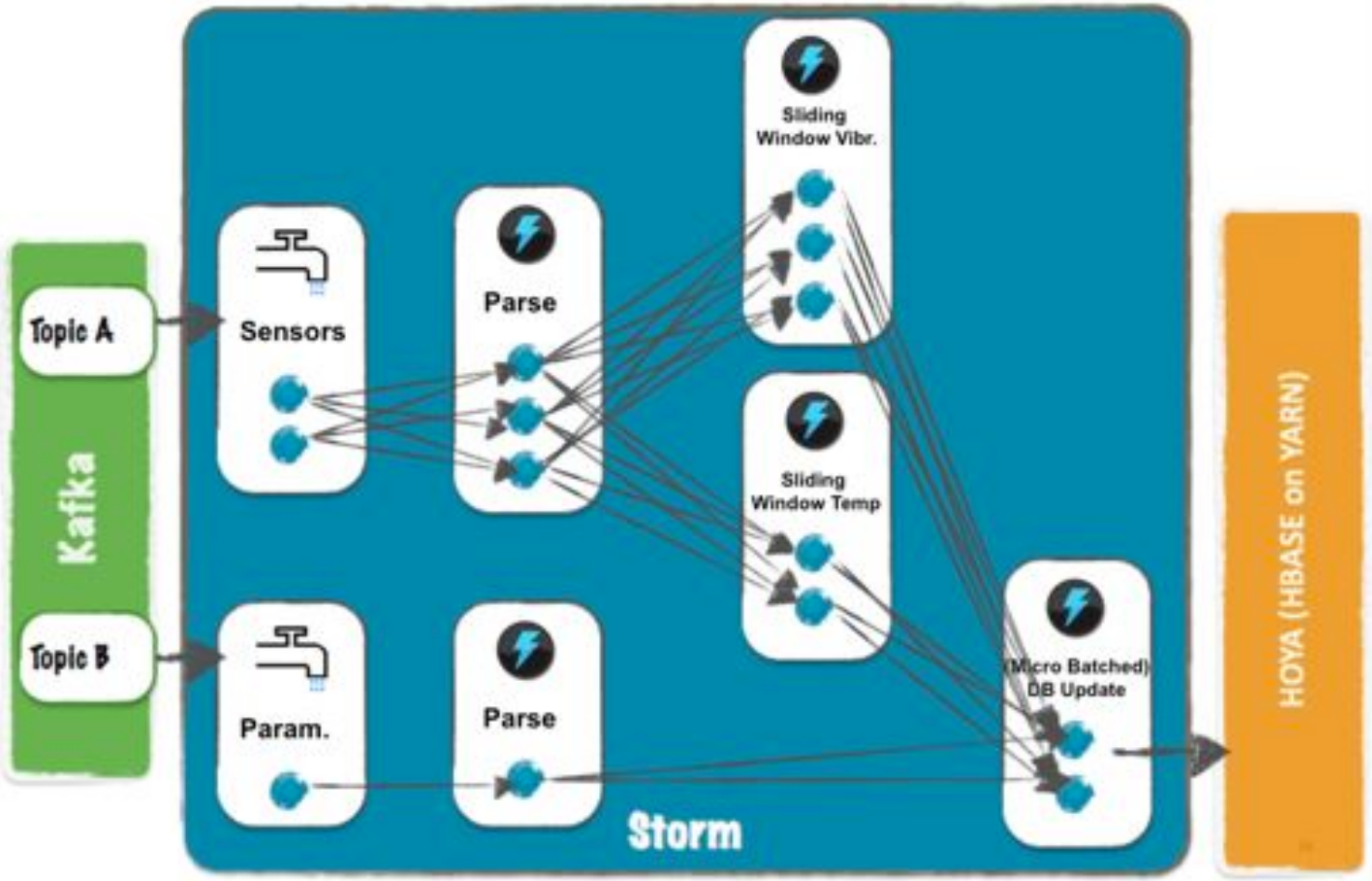
Distributed (No)SQL databases easily scale to very large datasets, very high load and do not need predefined schemata (can deal easily with Variability)



The Elasticsearch Stack (with Logstash and Kibana) provides an end to end solution for the management of (semi-structured) textual data (in particular logs)

Velocity

- With **Distributed Stream Processing** from hours to seconds
- With **In-Memory** computing from minutes to seconds
- With **Reactive Programming** from seconds to microseconds



Apache Storm

Distributed Stream Processing systems enable the cheap creation of systems that create realtime views from fast moving data.

Spark

EXASOL
Excellent data experts.

Building on speed improvements from doing computations "**In-Memory**", these systems reduce the time for large Analysis tasks from minutes to (mili-)seconds

Reactive Programming techniques from domains where microseconds count (e.g. High Frequency Trading, Real Time Advertising, intrusion prevention, sensors with high data rates) are becoming mainstream and easy to use.

Moore's Law, Cloud Computing, advances in networking technology and emerging data value chains are trends that further increase the speed by which costs are falling.

Value

Big Date Technologies add value by making more patterns (in data) visible and useful

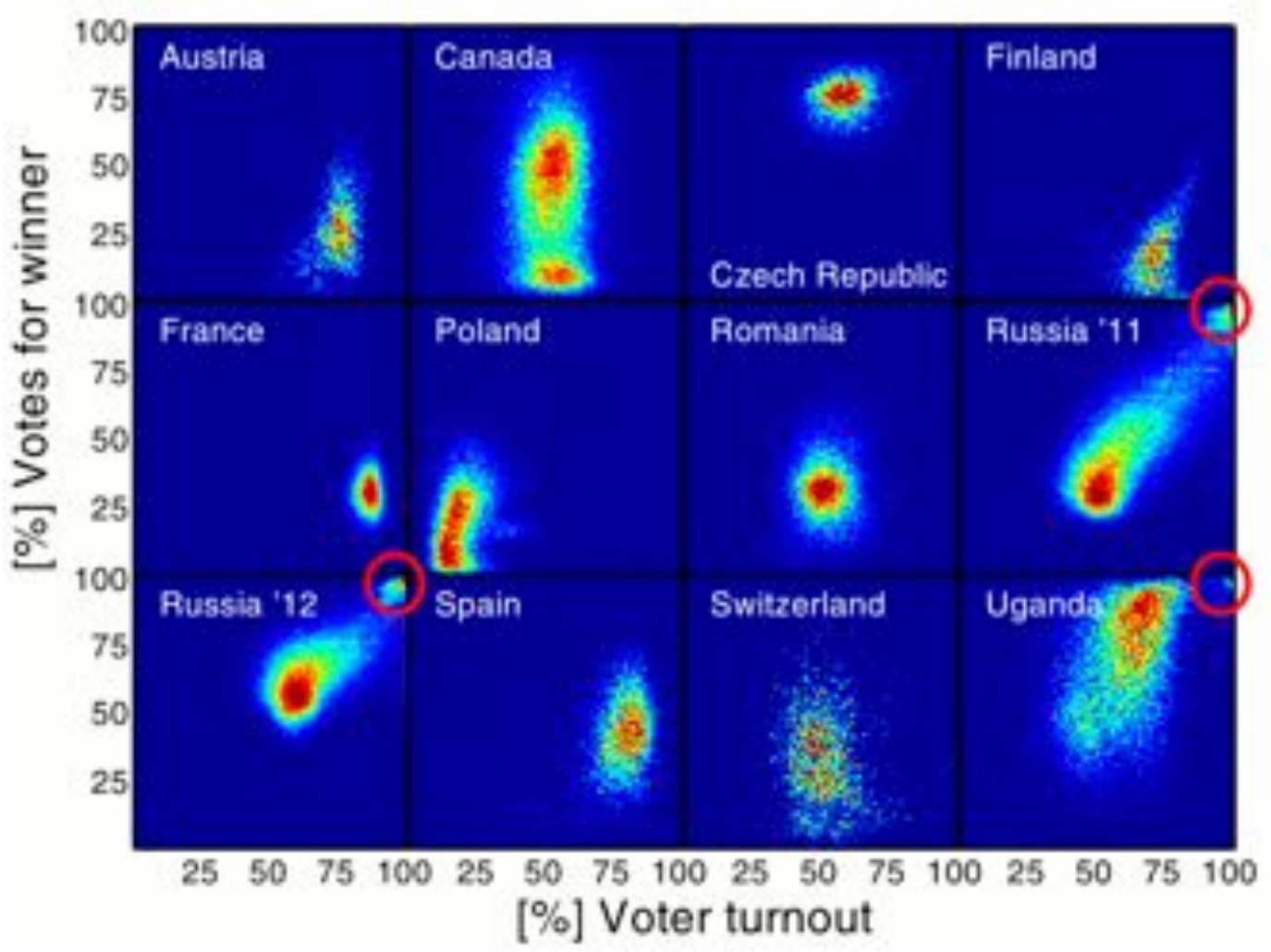
- through Analytics


Big Data Analytics
Value
$
Cost
Data Size & Processing Complexity

Analytics Types

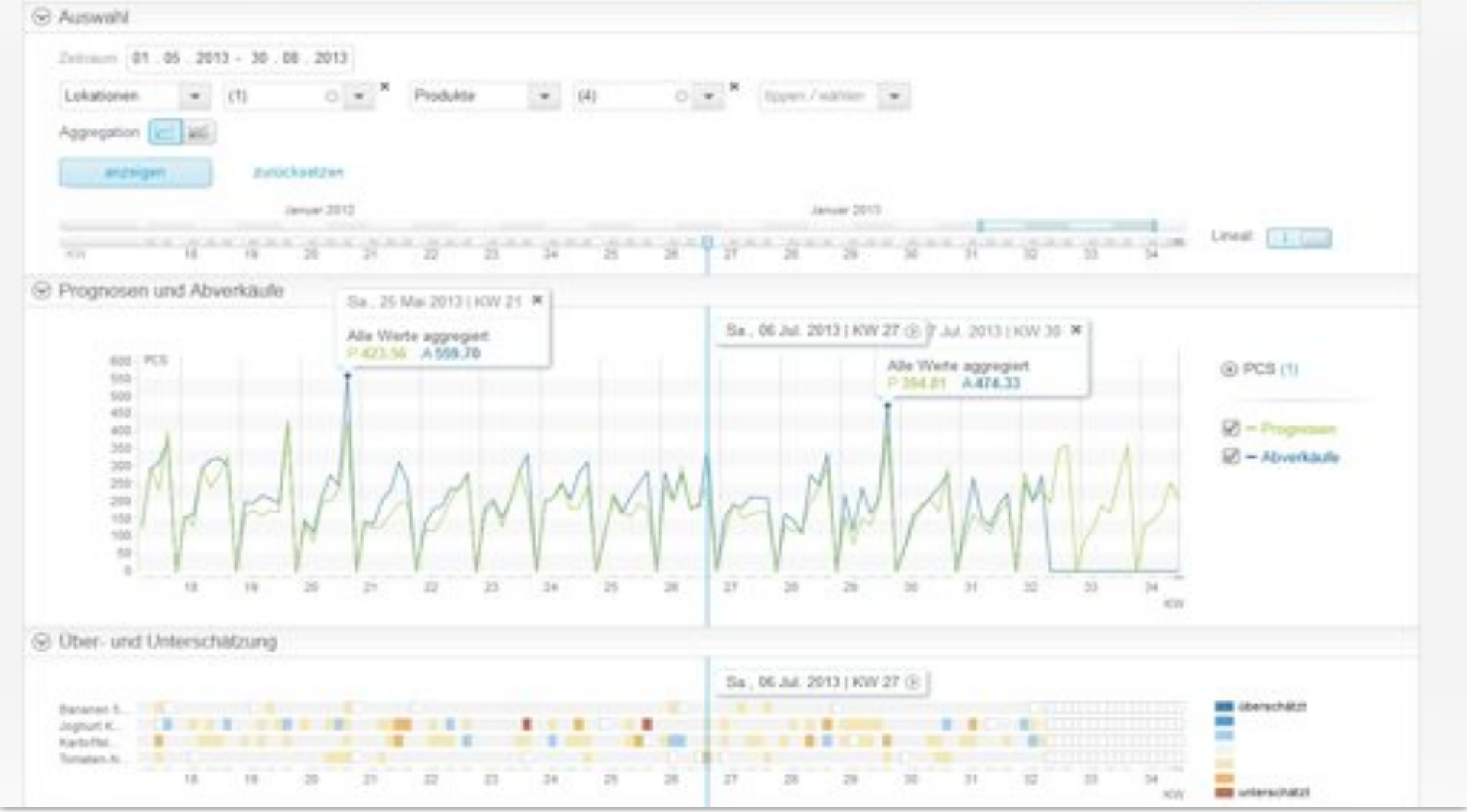
- **Descriptive**: What is and what has been?
  - Special Cases: Visual Analytics, Causal Analytics
- **Predictive**: What will be?
- **Prescriptive**: What is my optimal course of action?

Descriptive Analytics: Integrating data and making it accessible to understand past and present, e.g. the success of a tried treatments in similar patients.



Visual Analytics: Aggregating information into forms that human knowledge and intuition can be applied to harness it, e.g. to understand election fraud.

Causal Analytics: Using data to understand the root cause of observed phenomena such as defects (also known as Root Cause Analysis)



Predictive Analytics: Using patterns in data to predict the future, e.g. of demand

Prescriptive Analytics: Find an optimal course of action, e.g. an optimal sequence for re-starting flights after a large disturbance

Analytics Application Areas

- Operational Excellence
- Customer Intimacy
- Product Leadership

Operational Excellence: Use of data to increase the efficiency in the creation of products and service, e.g. through proactive maintenance.

Customer Intimacy: Use of data to better tailor products and services to customers, e.g. instantly display a prospective customers value in order to tailor offers to that.

Product Leadership: Use of data to create products of unmatched quality, e.g. through the systematic collection of DRO data for all cars throughout their lifecycle.

Most common pattern is the integration of **large amounts** of **diverse information** for each product/machine/customer into one coherent **real time** picture

360° Everything

Business Models

- Data Driven Business
- Data + Product
  - Data for Products
  - Products with Data
- Data as Business

Data for Products: Data driven services that enable the optimal use of products sold, e.g. machines and services to optimize agricultural yield down to the square foot.

.. or the optimal deployment and maintenance of trucks

Product with Data: Product offering functionality strongly dependent on Analytics, e.g. an alarm clock that wakes based on sensor analytics

Data as Business: Profit from harnessing the data collecting through other services, e.g. data for urban planning from fitness devices

or collected specifically to be sold …

#Next

People & Skills

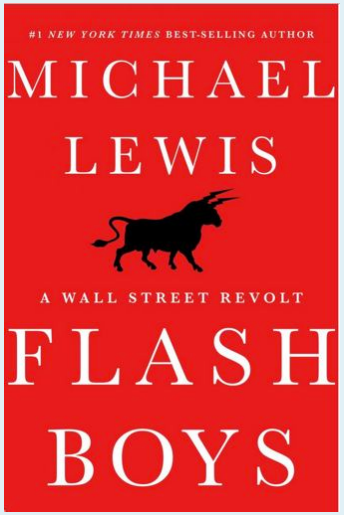
- Build up capability in your organization to harness Big Data technologies to build data management infrastructures that are cheaper, simpler and more powerful
- Through
  - Pilot projects with these technologies
  - Guided introductions such as our own “Big Data Hands On” workshop

Big Data Hands On Workshop

- Learning about BigData by building a big Data solution guided by our experts
  - Day 1: Data integration with Logstash and SpringXD
  - Day 2: Descriptive Analytics and Exploration with Elastic Search/Kibana and Hive/Hadoop
  - Day 3: Predictive Analytics with Mahout
- On premise, contents are customizable.

Analytics Use Cases

- Find ways data can improve your decisions, that of your customers, your suppliers or ?
  - By reading



  - Or in workshops (with vendors, independent consultants, strategy consultants)

Motivation, **Data Deluge**, Datification, Volume, Variety, Velocity, **Data Driven Decision**, n=All, t=now, **Cost**, **Data OS**, Hadoop, **Enterprise Data Lake**, **Distributed Search**, ElasticSearch, **NoSQL et al**, Cassandra, mongoDB, Riak, Pivotal, cloudera, **Distributed Stream Processing**, Storm, **In Memory Computing**, SAP Hana, Spark, EXASOL, **Reactive Programming**, Moore's Law, Cloud Computing, Data Value Chain, **Value**, **Analytics**, Descriptive, Visual, Causal, Predictive, Prescriptive, **Application Areas**, Operational Excellence, Customer Intimacy, Product Leadership, **360° Everything**, **Business Models**, Data Driven Business, Data for Products, Products with Data, Data as Business

connect / download slides at
www.vzach.de