Big Data Analytics

Lecture BigData Analytics

Julian M. Kunkel

julian.kunkel@googlemail.com

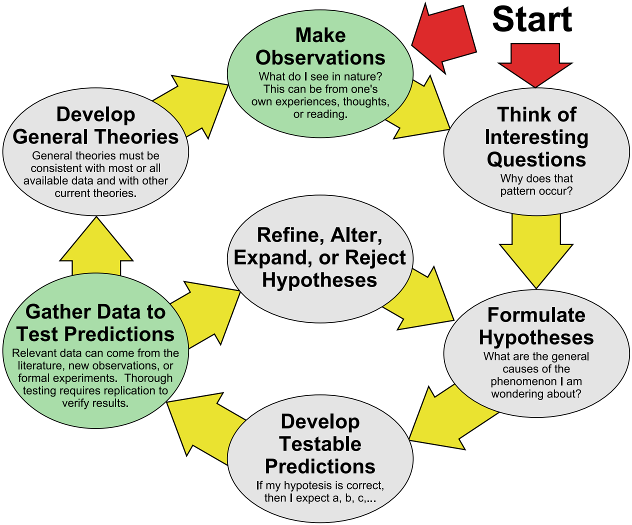
University of Hamburg / German Climate Computing Center (DKRZ)

2017-06-08

Disclaimer: Big Data software is constantly updated, code samples may be outdated.

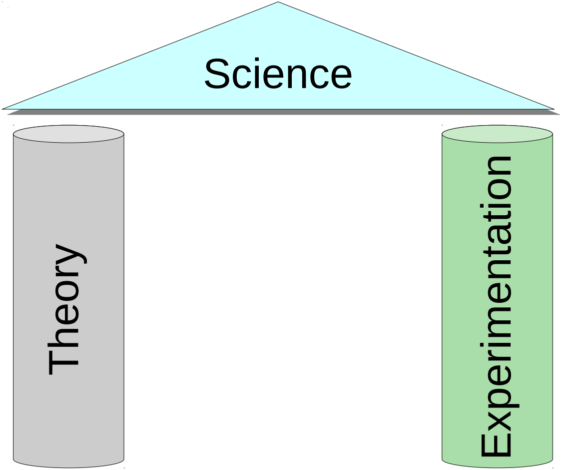
Outline

Scientific Method

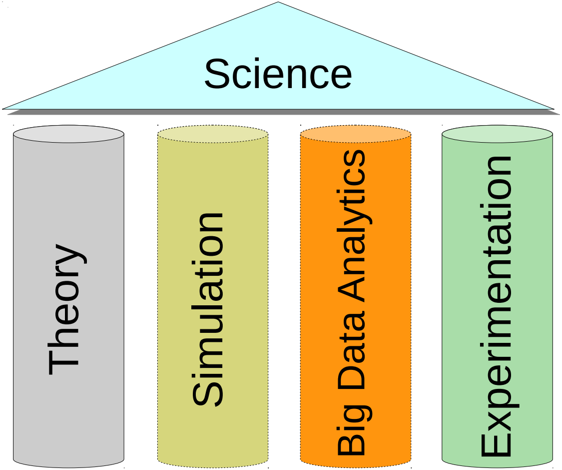


Based on: The Scientific Method as an Ongoing Process, ArchonMagnus[22]

Pillars of the Scientific Method



Pillars of Science: **Modern Perspective**



Idea of Big Data Analytics

Big Data

- Vast amounts of data are available
- Many heterogene data sources
- Raw data is of low value (fine grained)

Analytics

- Analyzing data $\Rightarrow$ Insight == value
  - For academia: knowledge
  - For industry: business advantage and money
- Levels of insight – primary abstraction levels of analytics
  - **Exploration**: study data and identify properties of (subsets) of data
  - **Induction/Inference**: infer properties of the full population
- Big data tools allow to construct a theory/model and validate it with data
  - **Statistics** and **machine learning** provide **algorithms and models**
  - Visual methods support data exploration and analysis

Example Models

Similarity is a (very) simplistic model and predictor for the world

- Humans use this approach in their cognitive process
- Uses the advantage of BigData

Weather prediction

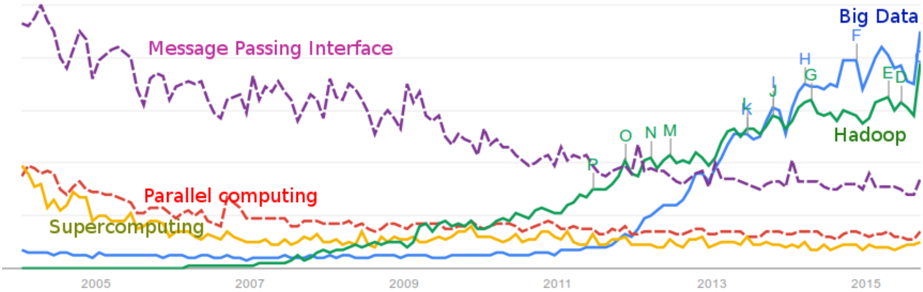
- You may develop and rely on complex models of physics
- Or use a simple model for a particular day; e.g., expect it to be similar to the weather of the typical day over the last X years
  - Used by humans: rule of thumb for farmers

Preferences of Humans

- Identify a set of people which liked items you like
- Predict you like also the items those people like but haven't rated

Relevance of Big Data

- Big Data Analytics is emerging
- Relevance increases compared to supercomputing



Google Search Trends, relative searches

Roles in the Big Data Business

Data scientist

Data science is a systematic method dedicated to knowledge discovery via data analysis [1]

- In business, optimize organizational processes for efficiency
- In science, analyze experimental/observational data to derive results

Data engineer

Data engineering is the domain that develops and provides systems for managing and analyzing big data

- Build modular and scalable data platforms for data scientists
- Deploy big data solutions

Typical Skills

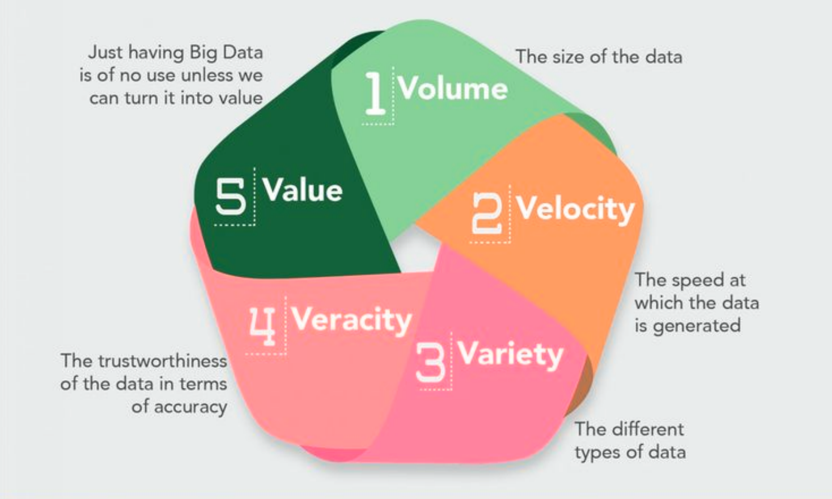
Data scientist

- Statistics + (mathematics) background
- Computer science
  - Programming, e.g.: R, (SAS,) Java, Scala, Python
  - Machine learning
- Some domain knowledge for the problem to solve

Data engineer

- Computer science background
  - Databases
  - Software engineering
  - Massively parallel processing
  - Real-time processing
- Languages: C++, Java, (Scala,) Python
- Understand performance factors and limitations of systems

BigData Challenges & Characteristics



Source: MarianVesper [4]

Volume: The size of the Data

What is Big Data

Terrabytes to 10s of petabytes

What is not Big Data

A few gigabytes

Examples

- Wikipedia corpus with history ca. 10 TByte
- Wikimedia commons ca. 23 TByte
- Google search index ca. 46 Gigawebpages[1]
- YouTube per year 76 PByte (2012[2])

[1] http://www.worldwidewebsize.com/

[2] https://sumanrs.wordpress.com/2012/04/14/youtube-yearly-costs-for-storagenetworking-estimate/

Velocity: Data Volume per Time

What is Big Data

30 KiB to 30 GiB per second
(902 GiB/year to 902 PiB/year)

What is not Big Data

A never changing data set

Examples

- LHC (Cern) with all experiments about 25 GB/s [3]
- Square Kilometre Array 700 TB/s (in 2018) [4]
- 50k Google searches per s [5]
- Facebook 30 Billion content pieces shared per month [6]

---

[3] http://home.web.cern.ch/about/computing/processing-what-record
[4] http://venturebeat.com/2014/10/05/how-big-data-is-fueling-a-new-age-in-space-exploration/
[5] http://www.internetlivestats.com/google-search-statistics/
[6] https://blog.kissmetrics.com/facebook-statistics/

Data Sources

Enterprise data

- Serves business objectives, well defined
- Customer information
- Transactions, e.g., purchases

Experimental/Observational data (EOD)

- Created by machines from sensors/devices
- Trading systems, satellites
- Microscopes, video streams, smart meters

Social media

- Created by humans
- Messages, posts, blogs, Wikis

Variety: Types of Data

- Structured data
  - Like tables with fixed attributes
  - Traditionally handled by relational databases
- Unstructured data
  - Usually generated by humans
  - Examples: natural language, voice, Wikipedia, Twitter posts
  - Must be processed into (semi-structured) data to gain value
- Semi-structured data
  - Has some structure in tags but it changes with documents
  - Examples: HTML, XML, JSON files, server logs

What is Big Data

- Use data from multiple sources and in multiple forms
- Involve unstructured and semi-structured data

Veracity: Trustworthiness of Data

What is Big Data

- Data involves some uncertainty and ambiguities
- Mistakes can be introduced by humans and machines
- Examples
  - People sharing accounts
  - Like sth. today, dislike it tomorrorw
  - Wrong system timestamps

Data Quality is vital!

Analytics and conclusions rely on good data quality

- Garbage data + perfect model => garbage results
- Perfect data + garbage model => garbage results

GIGO paradigm: *Garbage In – Garbage Out*

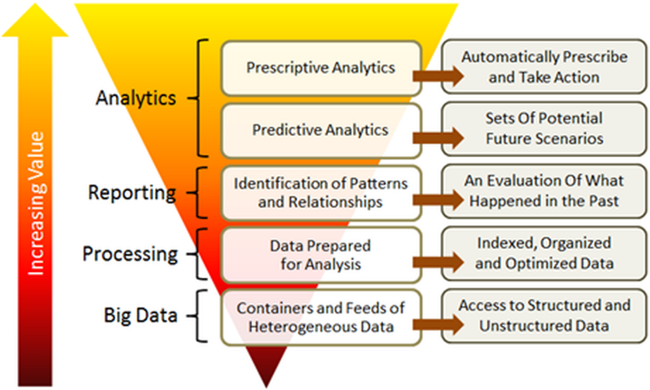
Value of Data

What is Big Data

- Raw data of Big Data is of low value
  - For example, single observations
- Analytics and theory about the data increases the value
  - Analytics transform big data into smart (valuable) data!

Big Data Analytics Value Chain

- There are many visualizations of the processing and value chain



Source: Andrew Stein [8]

From Big Data to the Data Lake [20]

- With cheap storage costs, people promote the concept of the data lake
- Combines data from many sources and of any type
- Allows for conducting future analysis and not miss any opportunity

Attributes of the data lake

- Collect everything: all time all data: raw sources and processed data
  - Decide during analysis which data is important, e.g., no “schema“ until read
- Dive in anywhere: enable users across multiple business units to
  - Refine, explore and enrich data on their terms
- Flexible access: shared infrastructure supports various patterns
  - Batch, interactive, online, search

Data Science vs. Business Intelligence (BI)

Characteristics of BI

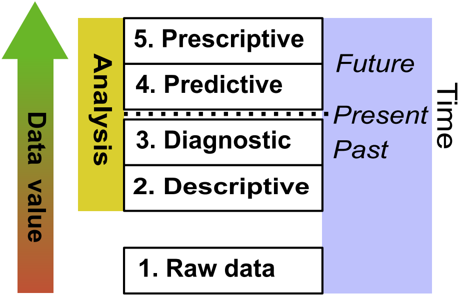
- Provides pre-created dashboards for management
  - Repeated visualization of well known analysis steps
- Deals with structured data
- Typically data is generated within the organization
- Central data storage (vs. multiple data silos)
- Handeled well by specialized database techniques

Typical types of questions and insight

- Customer service data: “what business causes customer wait times”
- Sales and marketing data: “which marketing is most effective”
- Operational data: “efficiency of the help desk”
- Employee performance data: “who is most/least productive”

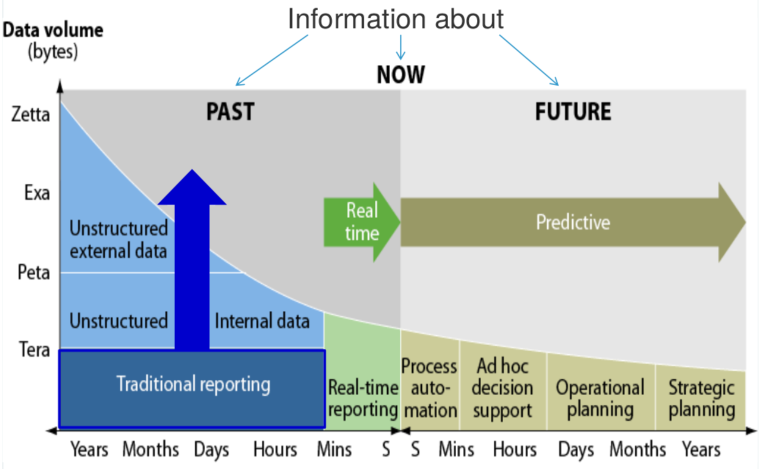
Abstraction Levels of Analytics and the Value of Data

1. Prescriptive analytics *(Empfehlen)*
   - "What should we do and why?"
2. Predictive analytics *(Vorhersagen)*
   - "What will happen?"
3. Diagnostic analytics
   - "What went wrong?"
   - "Why did this happen"
4. Descriptive analytics *(Beschreiben)*
   - "What happened?"
5. Raw (observed) data



For me, descriptive and diagnostic analysis is forensics!

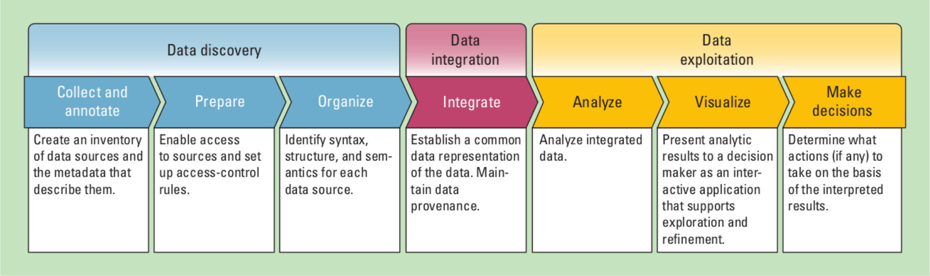
Analytics Abstraction Level



Source: Forrester report. Understanding The Business Intelligence Growth Opportunity. 20-08-2011

Data Analysis Workflow

The traditional approach proceeds in phases:



Source: Gilbert Miller, Peter Mork From Data to Decisions: A Value Chain for Big Data.

- Analysis tools: machine learning, statistics, interactive visualization
- Limitation: Interactivity by browsing through prepared results
- Indirect feedback between visualization and analysis

Exploratory Data Analysis (EDA) [23]

Definition

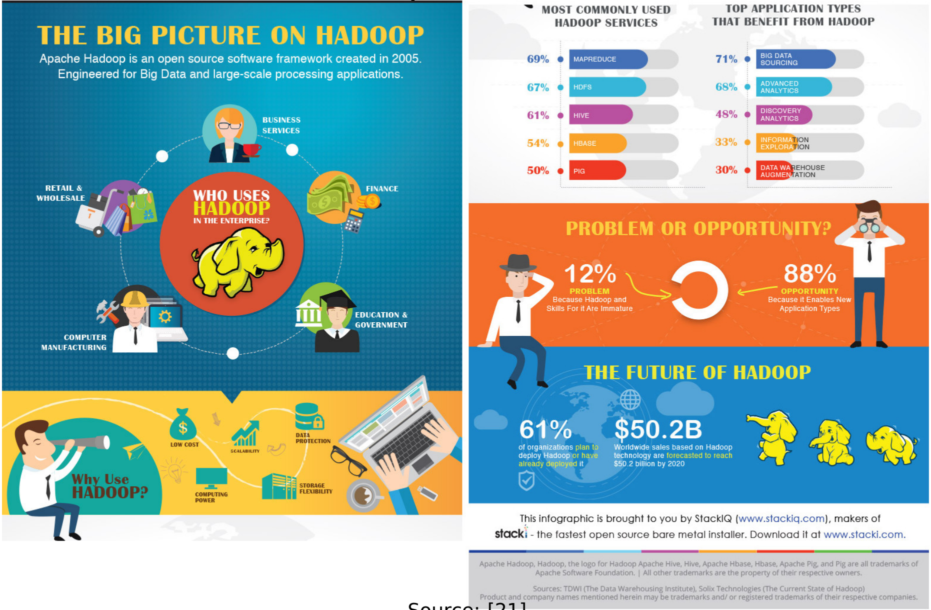
The approach of analyzing data sets to **summarize** their main **characteristic**, often with visual methods

Objectives

- Suggest hypotheses about the causes of observed phenomena
- Identify assumptions about the data to drive statistical inference
- Support selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

Methods from EDA can also be used for analyzing model results / outliers

Advertisement for a Big Data Platform



Source: [21]

Use Cases for BigData Analytics

Increase efficiency of processes and systems

- Advertisement: Optimize for target audience
- Product: Acceptance (like/dislike) of buyer, dynamic pricing
- Decrease financial risks: fraud detection, account takeover
- Insurance policies: Modeling of catastrophes
- Recommendation engine: Stimulate purchase/consume
- Systems: Fault prediction and anomaly detection
- Supply chain management

Science

- Epidemiology research: Google searches indicate Flu spread
- Personalized Healthcare: Recommend good treatment
- Physics: Finding the Higgs-Boson, analyze telescope data
- Enabler for social sciences: Analyze people's mood

Big Data in Industry

INDUSTRY	USE CASE	Sensor	Server Logs	Text	Social	Geographic	Machine	Clickstream	Structured	Unstructured
Financial Services	New Account Risk Screens		✔	✔						
	Trading Risk		✔							
	Insurance Underwriting	✔		✔		✔				
Telecom	Call Detail Records (CDR)					✔	✔			
	Infrastructure Investment		✔				✔			
	Real-time Bandwidth Allocation		✔	✔	✔					
Retail	360° View of the Customer			✔				✔		
	Localized, Personalized Promotions					✔				
	Website Optimization							✔		
Manufacturing	Supply Chain and Logistics	✔								
	Assembly Line Quality Assurance	✔								
	Crowd-sourced Quality Assurance				✔					
Healthcare	Use Genomic Data in Medial Trials	✔							✔	
	Monitor Patient Vitals in Real-Time									
Pharmaceuticals	Recruit and Retain Patients for Drug Trials				✔			✔		
	Improve Prescription Adherence				✔	✔				✔
Oil & Gas	Unify Exploration & Production Data	✔				✔				✔
	Monitor Rig Safety in Real-Time	✔								✔
Government	ETL Offloaded Response to Federal Budgetary Pressures								✔	
	Sentiment Analysis for Government Programs				✔					

Source: [20]

Example Use Case: Deutschland Card [2]

Goals

- Customer bonus card which tracks purchases
- Increase scalability and flexibility
- Previous solution based on OLAP

Big Data Characteristics

- Volume: O(10) TB
- Variety: mostly structured data, schemes are extended steadily
- Velocity: data growth rate O(100) GB / month

Results

- Much better scalability of the solution
- From dashboards to ad-hoc analysis within minutes

Example Use Case: DM [2]

Goals

- Predict required number of employees per day and store
- Prevent staff changes on short-notice

Big Data Characteristics

- Input data: Opening hours, incoming goods, empl. preferences, holidays, weather ...
- Model: NeuroBayes (Bayes + neuronal networks)
- Predictions: Sales, employee planning
- 450.000 predictions per week

Results

- Daily updated sales per store
- Reliable predictions for staff planning
- Customer and employee satisfaction

Example Use Case: OTTO [2]

Goals

Optimize inventory and prevent out-of-stock situations

Big Data Characteristics

- Input data: product characteristics, advertisement
- Volume/Velocity: 135 GB/week, 300 million records
- Model: NeuroBayes (Bayes + neuronal networks)
- 1 billion predictions per year

Results

- Better prognostics of product sales (up to 40%)
- Real time data analytics

Example Use Case: Smarter Cities (by KTH) [2]

Goals

- Improve traffic management in Stockholm
- Prediction of alternative routes

Big Data Characteristics

- Input data: Traffic videos/sensors, weather, GPS
- Volume/Velocity: 250k GPS-data/s + other data sources

Results

- 20% less traffic
- 50% reduction in travel time
- 20% less emissions

Example Facebook Studies

"Insight" from [11] by exploring posts

- Young narcissists tweet more likely.
  Middle-aged narcissists update their status
- US students post more problematic information than German students
- US Government checks tweets/facebook messages for several reasons
- Human communication graph has an average diameter of 4.74

Manipulation of news feeds [13]

- News feeds have been changed to analysis people's behavior in subsequent posts
- Paper: "Experimental evidence of massive-scale emotional contagion through social networks"

Programming BigData Analytics

High-level concepts

- SQL and derivatives
- Domain-specific languages (Cypher, PigLatin)

Programming languages

- Java interfaces are widely available but low-level
- Scala language increases productivity over Java
- Python and R have connectors to popular BigData solutions

In the exercises, we'll learn and use Python and R

Tools for Data Exploration

Mandatory features

- Interactive
- Rich set: visualization, data manipulations, algorithms
- Real-time processing of big data

Requirements

- Usability
- Flexible
- Performance

Tools (excerpt)

- Closed source: SAS, Spotfire, Domo, Tableau
- Open source: **R, Python/Jupyter**/Bokeh, GoogleVis
- Other open source tools, see [19]

Productivity

Productivity is a very important metric for Big Data tools

Development environments

1. Text editor; workflow: edit, save, (compile), run on a server
   - Notepad, gedit
2. Interactive shell; type code and execute it
   - Python, SQL frontent
3. IDE; optimized workflow of the text editor, may run code on a server
   - NetBeans, Eclipse, VisualStudio
4. Interactive lab notebook; type code and store it together with results
   - Examples: Jupyter, Apache Zeppelin
   - Embedded in GitHub:
     https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/code_listings/03.11-Working-with-Time-Series.ipynb
5. Lab notebook + IDE;
   - Examples: Spyder

Introduction to Python

- Open source
- Position 5 on TIOBE index
- Interpreted language
- Weak type system (errors at runtime)
- Development tools: any editor, interactive shell, Spyder
- Many useful libraries: matplotlib[7], NumPy, SciPy, Pandas
- *Note: Use and learn Python 3*

Specialties

- Strong text processing
- Simple to use
- Support for object oriented programming
- Indentation is relevant for code blocks

---

[7] http://matplotlib.org/gallery.html

Introduction to R

- Based on S language for statisticians
- Open source
- Position 19 on TIOBE index (but rising)
- Interpreter with C modules (packages)
- Libraries: Easy installation of packages via CRAN[8]
- Popular language for data analytics
- Development tools: RStudio (or any editor), interactive shell
- Recommended plotting library: ggplot2[9]

Specialties

- Vector/matrix operations. *Note: Loops are slow, so avoid them*
- Table data structure (data frames)

[8] Comprehensive R Archive Network
[9] http://docs.ggplot2.org/current/

Summary

- Big data analytics is a pillar of science
  - Supports building of hypothesis and experimentation
  - Challenges: 5 Vs – Volume, velocity, variety, veracity, value
- Data sources: Enterprise, humans, Exp./Observational data (EOD)
- Types of data: Structured, unstructured and semi-structured
- Roles in big data business: Data scientist and engineer
- Data science != business intelligence
- Analytics: Descriptive, diagnostic; predictive, prescriptive

Bibliography

1 Book: Lillian Pierson. **Data Science for Dummies**. John Wiley & Sons

2 Report: Jürgen Urbanski et.al. **Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte**. BITKOM

3 http://winfwiki.wi-fom.de/

4 Forrester Big Data Webinar. Holger Kisker, Martha Bennet. Big Data: Gold Rush Or Illusion?

5 http://blog.eoda.de/2013/10/10/veracity-sinnhaftigkeit-und-vertrauenswuerdigkeit-von-bigdata-als-kernherausforderung-im-informationszeitalter/

6 http://lehrerfortbildung-bw.de/kompetenzen/projektkompetenz/methoden/erkenntnis.htm

7 Gilbert Miller, Peter Mork From Data to Decisions: A Value Chain for Big Data. http://www.fh-schmalkalden.de/Englmeier-p-790/_/ValueChainBigData.pdf

8 Andrew Stein. The Analytics Value Chain. http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

9 Dursun Delen, Haluk Demirkan,. Decision Support Systems, Data, information and analytics as services.http://j.mp/11bl9b9

10 Wikipedia

11 Kashmir Hill. 46 Things We've Learned From Facebook Studies. Forbe. http://www.forbes.com/sites/kashmirhill/2013/06/21/46-things-weve-learned-from-facebook-studies/

12 Hortonworks http://hortonworks.com/

13 http://www.huffingtonpost.com/2014/12/10/facebook-most-popular-paper_n_6302034.html

20 http://hortonworks.com/blog/enterprise-hadoop-journey-data-lake/

21 http://www.stacki.com/hadoop/?utm_campaign=Stacki+Hadoop+Infographic

22 https://en.wikipedia.org/wiki/Scientific_method

23 https://en.wikipedia.org/wiki/Exploratory_data_analysis