Seminar Neueste Trends in Big Data Analytics

Julian M. Kunkel, Christian Hovy, Eugen Betke, Jakob Lüttgau, Tobias Finn

German Climate Computing Center (DKRZ)

2017-10-16

Outline

About DKRZ

German Climate Computing Center (DKRZ)

Partner for Climate Research

Maximum Compute Performance.

Sophisticated Data Management.

Competent Service.

Scientific Computing

- Research Group of Prof. Ludwig at the University of Hamburg
- Embedded into DKRZ

Research

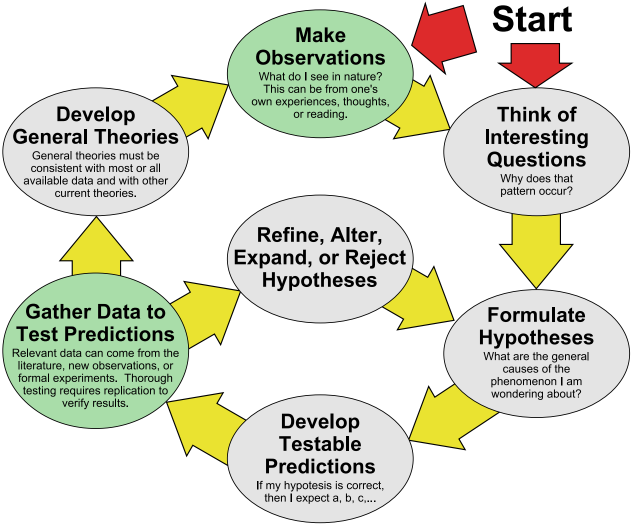
- Analysis of parallel I/O
- I/O & energy tracing tools
- Middleware optimization
- Alternative I/O interfaces
- Data reduction techniques
- Cost & energy efficiency

Seminar

Concept of the Seminar

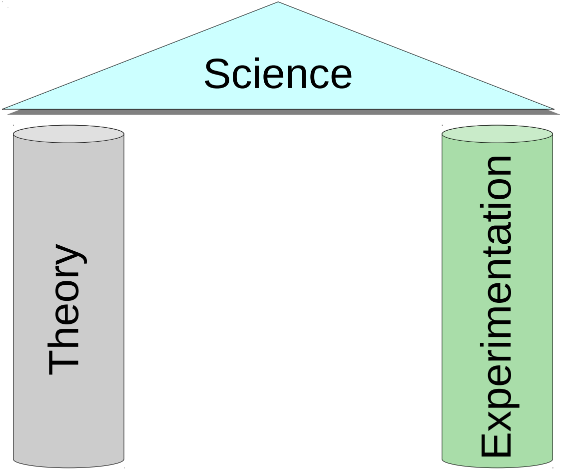
- Goal of the Seminar
  - Learning to extract, summarize and present relevant big data topics
- Organization
  - Each participant is assigned to one supervisor
  - We expect a timely delivery of the presentation
    - 3 weeks before presentation – rough structure and content
    - 2 weeks – draft of slides
    - 1 week – presentation slides
- Deliverables
  - Presentation
  - Short report (10+ pages) at the end of the semester (good for preparation)
  - Submission via: https://wr.informatik.uni-hamburg.de/abgabe/ntbd-1718/
- Information
  - See the web page
  - You must subscribe to the mailing list (see web page)!

Scientific Method

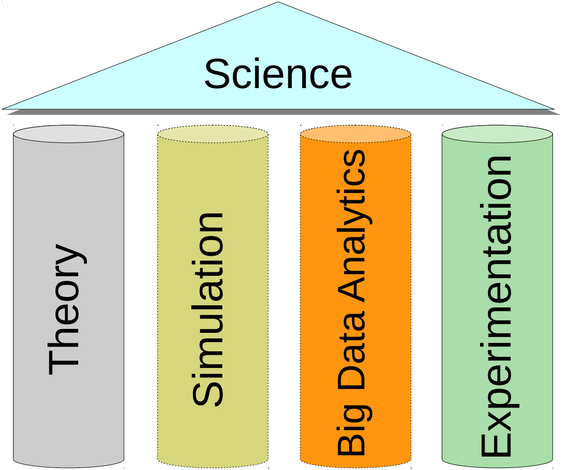


Based on: The Scientific Method as an Ongoing Process, ArchonMagnus[22]

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

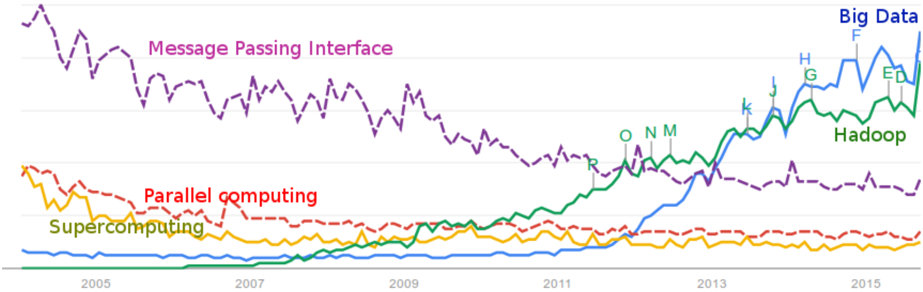
- You may develop and rely on complex models of physics
- Or use a simple model for a particular day; e.g., expect it to be similar to the weather of the typical day over the last X years
  - Used by humans: rule of thumb for farmers

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

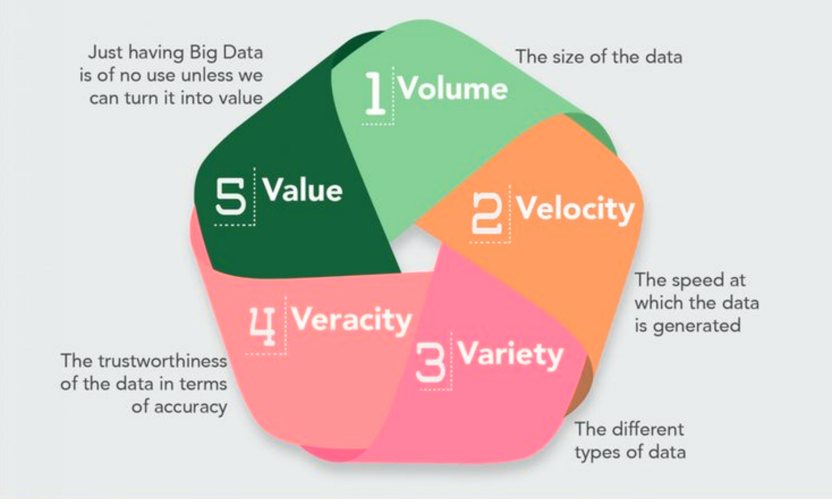
Data scientist

- 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

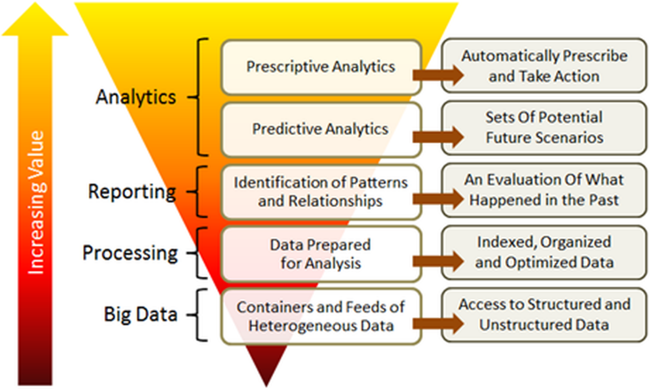
BigData Challenges & Characteristics



Source: MarianVesper [4]

Big Data Analytics Value Chain

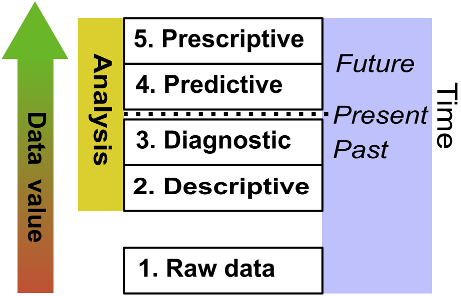
- There are many visualizations of the processing and value chain



Source: Andrew Stein [8]

Abstraction Levels of Analytics and the Value of Data

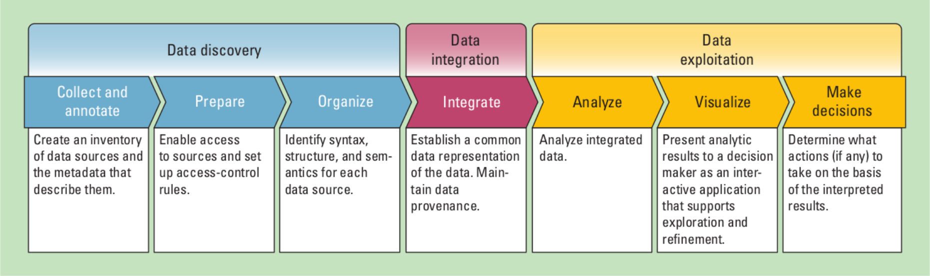
1. Prescriptive analytics *(Empfehlen)*
   - "What should we do and why?"
2. Predictive analytics *(Vorhersagen)*
   - "What will happen?"
3. Diagnostic analytics
   - "What went wrong?"
   - "Why did this happen"
4. Descriptive analytics *(Beschreiben)*
   - "What happened?"
5. Raw (observed) data



For me, descriptive and diagnostic analysis is forensics!

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

Data Mining (Knowledge Discovery) [1,35]

Definition

- **Data mining**: process of discovering patterns in large data sets
  - (Semi-)Automatic analysis of large data to identify interesting patterns
  - Using artificial intelligence, machine learning, statistics and databases

Tasks / Problems for data mining

- **Classification**: predict the category of samples
- **Regression**: find a function to model numeric data with the least error
- **Anomaly detection**: identify unusual data (relevant or error)
- **Association rule learning**: identify relationships between variables
- **Clustering**: discover and classify similar data into structures and groups
- **Summarization**: find a compact representation of the data

Terminology for Input Data [1, 40]

- **Sample**: instances (subset) of the unit of observation
- **Feature**: measurable property of a phenomenon (explanatory variable)
  - The set of features is usually written as vector (f1, ..., fn)
- **Label/response**: outcome/property of interest for analysis/prediction
  - Dependent variable
  - Discrete in classification, continuous in regression

Forms of features/labels

- Numeric: a (potentially discrete) number characterizes the property
  - e.g., age of people
- Categorical/nominal: a set of classes
  - e.g., eye color
  - Dichotomous (binary) variable: contains only two classes (Male: Yes/No)
- Ordinal: an ordered set of classes
  - e.g., babies, teens, adults, elderly

Example Data

Imagine we have data about alumni from the university

Field of study	Gender	Age	Succ. exams	Fail. exams	Avg. grade*	Graduate	Dur. studies
CS	M	24	21	1	2.0	Yes	10
CS	M	22	5	2	1.7	Enrolled	2
Physics	F	23	20	1	1.3	Enrolled	6
Physics	M	25	8	10	3.0	No	10

- Categorical: field of study, gender, graduate, (favourite colour)
- Numeric: age, successful/failed exams, duration of studies
- Numeric: average grade; Ordinal: very good, good, average, failed

Our goal defines the machine learning problem

- Predict if a student will graduate ⇒ classification
  - Prescriptive analysis: we may want to support these students better
- Predict the duration (in semesters) for the study ⇒ regression
- Clustering to see if there are interesting classes of students
  - We could label these, e.g., the prodigies, the lazy, ...
  - Probably not too helpful for the listed features

Terminology for Learning [40]

- **Online learning**: update the model constantly while it is applied
- **Offline (batch) learning**: learn from data (training phase), then apply
- **Supervised learning**: feature and label are provided in the training
- **Unsupervised learning**: no labels provided, relevant structures must be identified by the algorithms, i.e., descriptive task of pattern discovery
- **Reinforcement learning**: algorithm tries to perform a goal while interacting with the environment
  - Humans use reinforcement, (semi)-supervised and unsupervised learning

Overview of Machine Learning Algorithms (Excerpt)

Classification

- k-Nearest neighbor
- Naive bayes
- Decision trees
- Classification rule learners

Regression/Numeric prediction

- Linear regression
- Regression trees
- Model trees

Regression & classification

- Neuronal networks
- Support vector machines

Pattern detection

- Association rules
- k-means clustering
- density-based clustering
- model-based clustering

Meta-learning algorithms

- Bagging
- Boosting
- Random forests

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
- Monetization: Extract money from gamers [27]
- Systems: Fault prediction and anomaly detection

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

Learning Behavior

Games

- DeepMind playing atari games [29]
- AlphaGo wins vs. humans in playing Go [26]
- AI beating world's best gamer in Dota 2 [28]

Motion

- Learning hand motion by human training [30]
- Robots learning to pick up items [31]

Systems: Fault Prediction and Anomaly Detection

Smart buildings [24]

- Predicting faults of heating and ventilation of an hospital
- Predicted 76 of 124 real faults and 41 of 44 exceptional temperatures
- May consider weather to control systems automatically

Google DeepMind AI [25]

- Controlling 120 variables in the data center (fans, ...)
- Saves 15% energy of the overall bill

Automatize Classification

Analysis of multimedia

- Voice, face, biometric recognition
- Speech recognition
- Counting (animal) species on pictures / videos
- Finding patterns on satellite images (e.g., damn, thunderstorms)
- Anomalies in behavior (depressed people)
- Anomalies in structures (operational condition)

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

24 https://www.newscientist.com/article/2118499-smart-buildings-predict-when-critical-systems-are-about-to-fail/

25 https://www.theverge.com/2016/7/21/12246258/google-deepmind-ai-data-center-cooling

26 https://deepmind.com/research/alphago/

27 https://www.ibm.com/developerworks/library/ba-big-data-gaming/index.html

28 http://money.cnn.com/2017/08/12/technology/future/elon-musk-ai-dota-2/index.html

29 http://www.wired.co.uk/article/google-deepmind-atari

30 https://arxiv.org/abs/1603.06348

31 https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/google-large-scale-robotic-grasping-project

35 https://en.wikipedia.org/wiki/Data_mining

40 https://en.wikipedia.org/wiki/Machine_learning