Project: Big Data

Julian M. Kunkel, Eugen Betke, Jakob Lüttgau, Tobias Finn, Andrej Fast, Heinrich Widmann

German Climate Computing Center (DKRZ)

2017-10-16
Outline

1 Organization
2 Big Data Analytics
3 BigData Challenges
4 Gaining Insight with Analytics
5 Use Cases
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
Project

Concept of the project

- **Goals of the project**
  - Learning to practically resolve a big data related problem
  - Contribution on our group’s research related problems

- **Organization**
  - Teams of 2-3 people work on one topic
  - Monthly meetings in the group to present the current status
  - Teamwork proceeds individually
  - Typically two supervisors per topic
  - Regular mail exchange with supervisors expected!

- **Deliverables**
  - Presentation of final results in lecture free time (End of Feb.)
  - Short report (10+ pages) at the end of the semester
  - Submission via: https://wr.informatik.uni-hamburg.de/abgabe/bdp-1718/

- **Information**
  - See the web page
  - You must subscribe to the mailing list (see web page)!
1. Organization

2. Big Data Analytics

3. Big Data Challenges

4. Gaining Insight with Analytics

5. Use Cases
Scientific Method

Start

Think of Interesting Questions
Why does that pattern occur?

Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?

Refine, Alter, Expand, or Reject Hypotheses

Gather Data to Test Predictions
Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.

Develop Testable Predictions
If my hypothesis is correct, then I expect a, b, c,...

Make Observations
What do I see in nature? This can be from one's own experiences, thoughts, or reading.

Develop General Theories
General theories must be consistent with most or all available data and with other current theories.

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

Science

Theory

Experimentation
Pillars of Science: Modern Perspective
Idea of Big Data Analytics

Big Data

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

Analytics

- Analyzing data ⇒ 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
| 1 | Organization |
| 2 | Big Data Analytics |
| 3 | BigData Challenges |
| 4 | Gaining Insight with Analytics |
| 5 | Use Cases |
BigData Challenges & Characteristics

1. Volume
   - The size of the data

2. Velocity
   - The speed at which the data is generated

3. Variety
   - The different types of data

4. Veracity
   - The trustworthiness of the data in terms of accuracy

5. Value
   - Just having Big Data is of no use unless we can turn it into value

Source: MarianVesper [4]
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
<table>
<thead>
<tr>
<th></th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Big Data Analytics</td>
</tr>
<tr>
<td>3</td>
<td>BigData Challenges</td>
</tr>
<tr>
<td>4</td>
<td>Gaining Insight with Analytics</td>
</tr>
<tr>
<td>5</td>
<td>Use Cases</td>
</tr>
</tbody>
</table>
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:

1. **Data discovery**
   - Collect and annotate: Create an inventory of data sources and the metadata that describe them.
   - Prepare: Enable access to sources and set up access-control rules.
   - Organize: Identify syntax, structure, and semantics for each data source.

2. **Data integration**
   - Establish a common data representation of the data. Maintain data provenance.

3. **Data exploitation**
   - Analyze: Analyze integrated data.
   - Visualize: Present analytic results to a decision maker as an interactive application that supports exploration and refinement.
   - Make decisions: Determine what actions (if any) to take on the basis of the interpreted results.

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 \((f_1, ..., f_n)\)
- **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
Imagine we have data about alumni from the university

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

- 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
1. Organization
2. Big Data Analytics
3. BigData Challenges
4. Gaining Insight with Analytics
5. Use Cases
# 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
10. Wikipedia
11. Kashmir Hill. 46 Things We’ve Learned From Facebook Studies. Forbe.
23. http://www.wired.co.uk/article/google-deepmind-atari