Introduction to Data Driven Modeling

Unknown Image Source

Goals

- To understand basic principles of classical data driven modeling and important terms
- Know the difference of supervised and un-supervised learning
- Know how to use un-supervised learning and apply basic algorithms in a hands-on data driven modeling project
  - Try and test important algorithms in Python

Module has 3 parts

- 3 parts
  - Lecture on theoretical basics
  - Self-learning via online courses for coding-skills
  - Project on real world data
- Meetings
  - Series of 7 meetings roughly every two weeks on **Gummersbach campus**

Data Driven Modelling WS 2024 –0.504

Date	Topic
10.10.24	Introduction to Data Driven Modeling & Info for Business Understanding of Project
17.10.24	Self-Learning
24.10.24	Business Understanding Milestone & Lecture for Data Understanding
31.10.24	Self-Learning
07.11.24	Data Understanding Milestone & Lecture for Data Preparation
14.11.24	Self-Learning
21.11.24	Self-Learning
28.11.24	Data Preparation Milestone & Lecture for Modeling
05.12.24	Self-Learning
12.12.24	Modeling Milestone & Lecture for Evaluation
19.12.24	Self-Learning
26.12.24/02.01.25	Turn of the year holidays
09.01.24	Self-Learning
16.01.24	Project Presentations
23.01.24	Project Presentations

Organizational Issues

- A personal registration in ILU and PSSO is required (details later in the lecture).
- Grades composition:
  - 50% Learning Portfolio (weekly report on learning progress)
  - 50% Project Presentation
- Useful Literature:
  - Ebooks from TH Köln library:
    - Aurélien Géreon: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, O'Reilly: September 2019
    - Nikita Silaparasetty: Machine Learning Concepts with Python and the Jupyter Notebook Environment: Using Tensorflow 2.0, Apress: 2020
  - Real books:
    - Pang-Ning Tan, Michael Steinbach und Vipin Kumar: Introduction to Data Mining, Pearson, 2013.

Your Project

- You will get a project description with real data
  - More information about first project steps later
- You will work in groups
  - Result is a practical solution of the project
  - including a presentation in the lecture time
- 50 % of the final grade

Result

- Presentation time: 10 minutes per member of the group
  - The concrete schedule announced later
  - Include all members and their tasks in the group
  - The slides are also the documentation of your project! Think about literature references, visualizations etc.
  - No additional project report is required.

Overview

- Some examples of Data Analytics
- What is Data Analytics?
- How to do Data Analytics

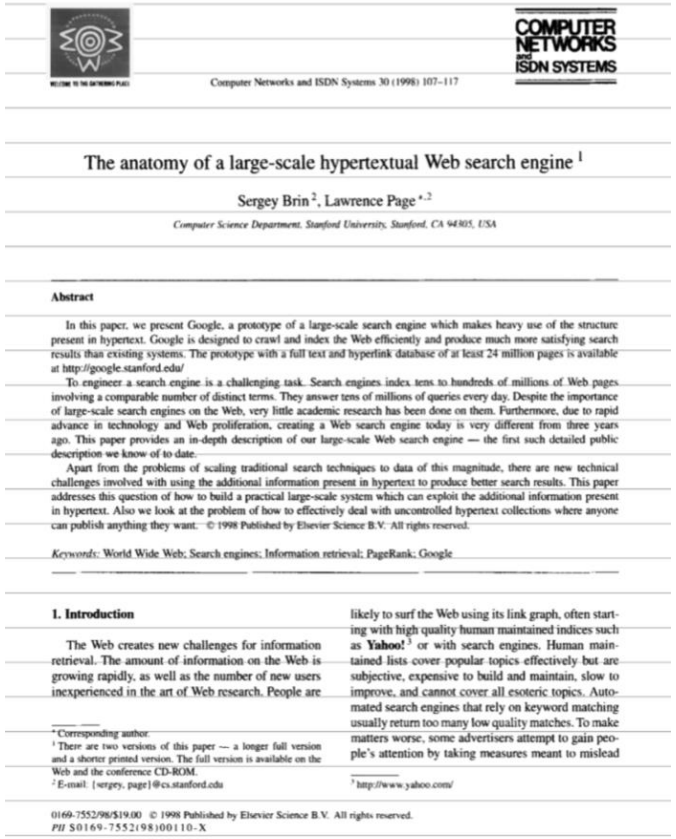
SOME EXAMPLES OF DATA ANALYTICS

…to spark your creativity about what can be done with Data Analytics.

Google

The PageRank Algorithm

How one algorithm can change the world



Computer Networks and ISDN Systems 30 (1998) 107–117

COMPUTER NETWORKS and ISDN SYSTEMS

The anatomy of a large-scale hypertextual Web search engine [1]

Sergey Brin [2], Lawrence Page *,[2]

Computer Science Department, Stanford University, Stanford, CA 94305, USA

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the Web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and Web proliferation, creating a Web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale Web search engine — the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords: World Wide Web; Search engines; Information retrieval; PageRank; Google

1. Introduction

The Web creates new challenges for information retrieval. The amount of information on the Web is growing rapidly, as well as the number of new users inexperienced in the art of Web research. People are likely to surf the Web using its link graph, often starting with high quality human maintained indices such as **Yahoo!** [3] or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead

* Corresponding author.

[1] There are two versions of this paper — a longer full version and a shorter printed version. The full version is available on the Web and the conference CD-ROM.

[2] E-mail: {sergey, page}@cs.stanford.edu

[3] http://www.yahoo.com/

0169-7552/98/$19.00
PII S0169-7552(98)00110-X

Unknown Image Source

Amazon

The Long Tail

In 1988, a British mountain climber named Joe Simpson wrote a book called »Touching the Void«, a harrowing account of near death in the Peruvian Andes. It got good reviews but, only a modest success, it was soon forgotten. Then, a decade later, a strange thing happened. Jon Krakauer wrote »Into Thin Air«, another book about a mountain-climbing tragedy, which became a publishing sensation. Suddenly »Touching the Void« started to sell again. (…)

changethis.com/manifesto/10.LongTail/pdf/10.LongTail.pdf

- *What happened? In short, Amazon.com recommendations.*

https://plus.google.com/+amazon/posts

Emotion Detection

Social-Media-Monitoring

- *MOTOR-TALK.de* is Germanys biggest online community for automotive topics.
- True to the sterotype, the car is one of the most emotional topics for Germans.
- How can 35 million posts be categorized for emotional content?

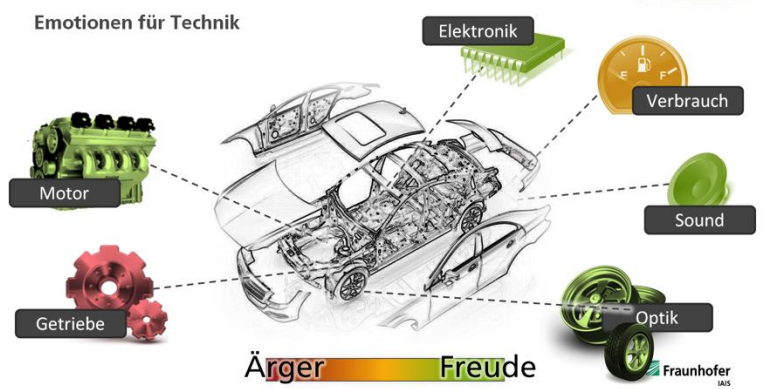
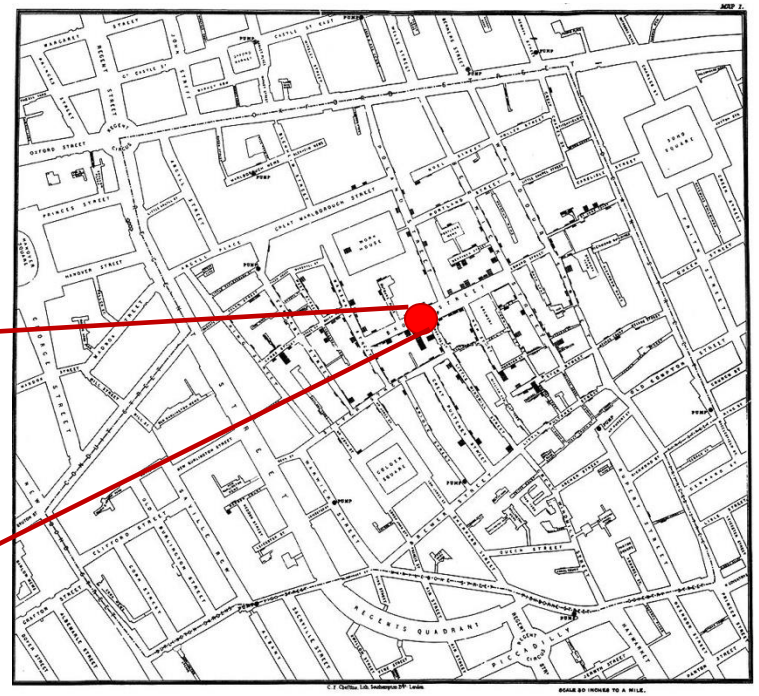


Image Source: Fraunhofer IAIS

Spatial Data Analytics

Broad Street cholera outbreak in London, 1854

- Physician John Snow showed in one of the first spatial data analyses, that cholera is spread by contaminated water

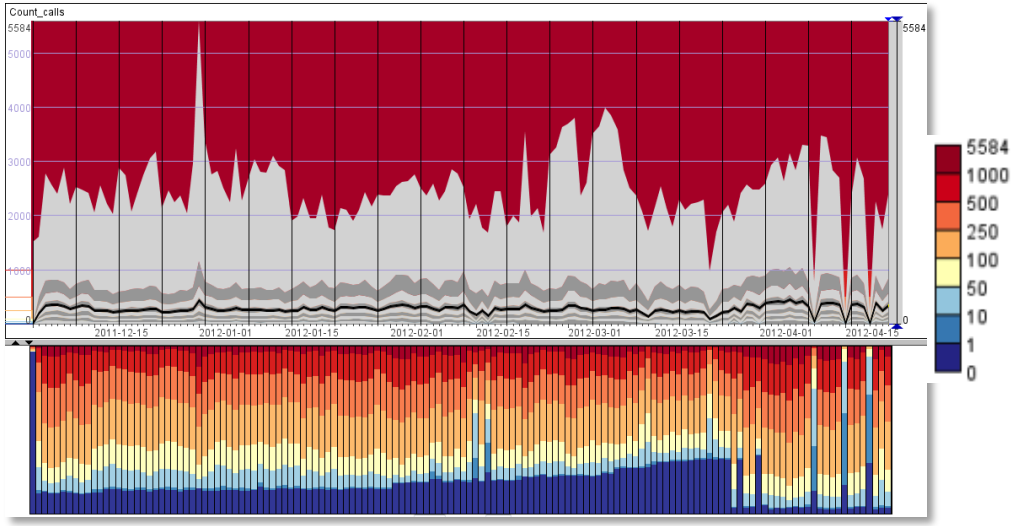


http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Visual Analytics

- **Detect the expected:**

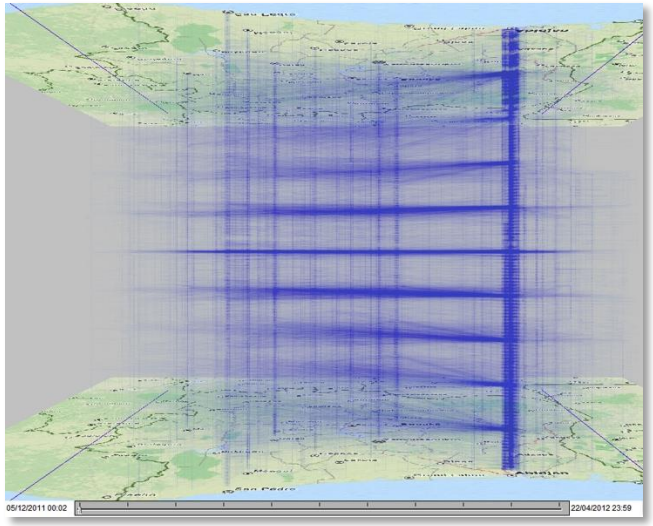
  missing data, sample quality, …



Days with missing data in key locations

- **and discover the unexpected:**

  seeing unexpected patterns in the data.



Undocumented re-assignment of user IDs every two weeks

- Orange D4D challenge[1]: Ivory Coast mobile phone CDR data – 5 million subscribers, 5 months, 2.5 billion calls + SMS

1) http://www.d4d.orange.com/home

WHAT IS DATA ANALYTICS?

…showing you what to expect from Data Analytics (and what better not).

What is Data Analytics?

Knowledge Discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

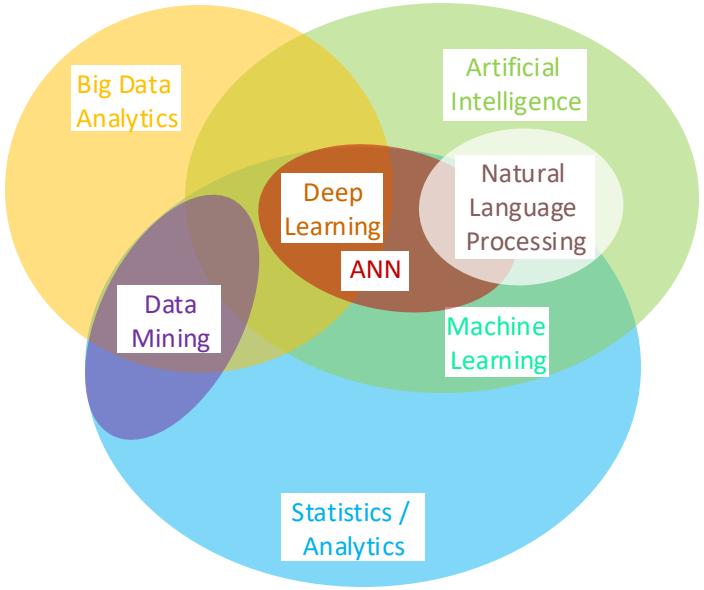
(Fayyad et al., 1996)

What is Data Analytics?

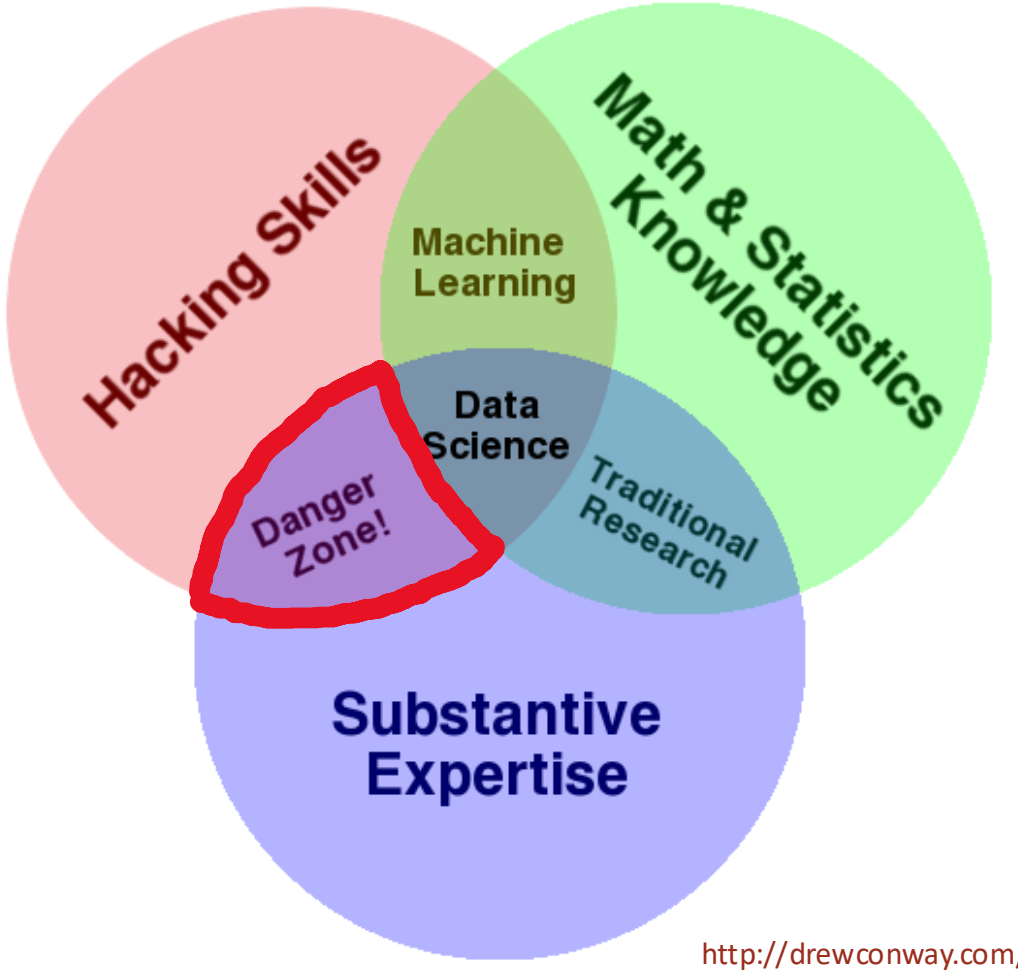
Unknown Image Source

Terms around analytics aren‘t clear cut

- Statistics / Analytics: every kind of analysing data (using statistical measures)
- Artificial Intelligence: the machine yielding „intelligent“ acting or insights
- Machine Learning: the machine „learns“ relationships (models)
- ANN (Artificial Neural Networks): biologically motivated approaches to modeling
- Big Data Analytics: every processing and analysis of large amounts of data
- Deep Learning: ANN with many layers and massive amounts of training data
- Natural Language Processing: processing of natural language expressions
- Data Mining: automatic recognition of patterns and relationships



The Data Science Mindset



Drew Conway,
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Analytics?

Data Analytics is:

Data Analytics is not:

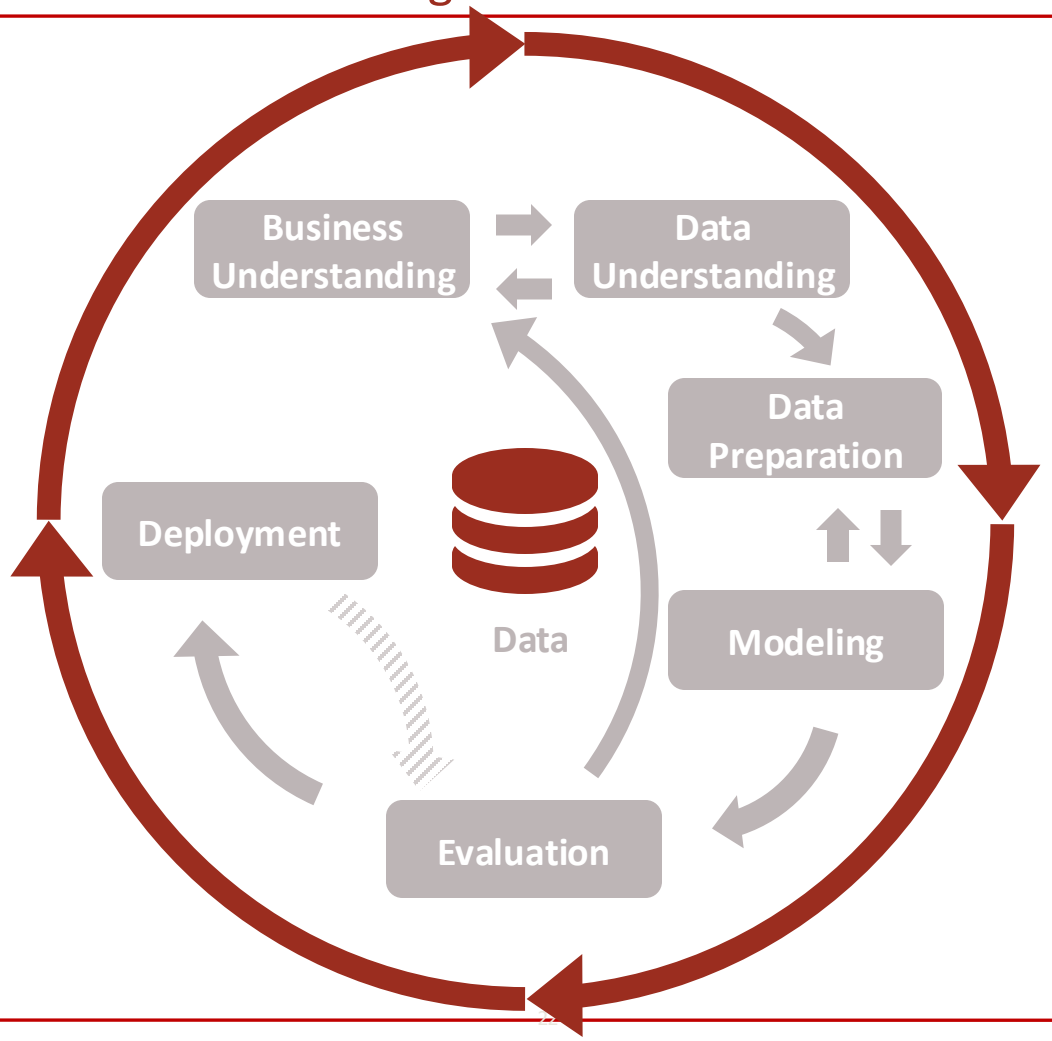
Unknown Image Source

HOW TO DO DATA ANALYTICS?

…and how to do it right: giving structure to the process.

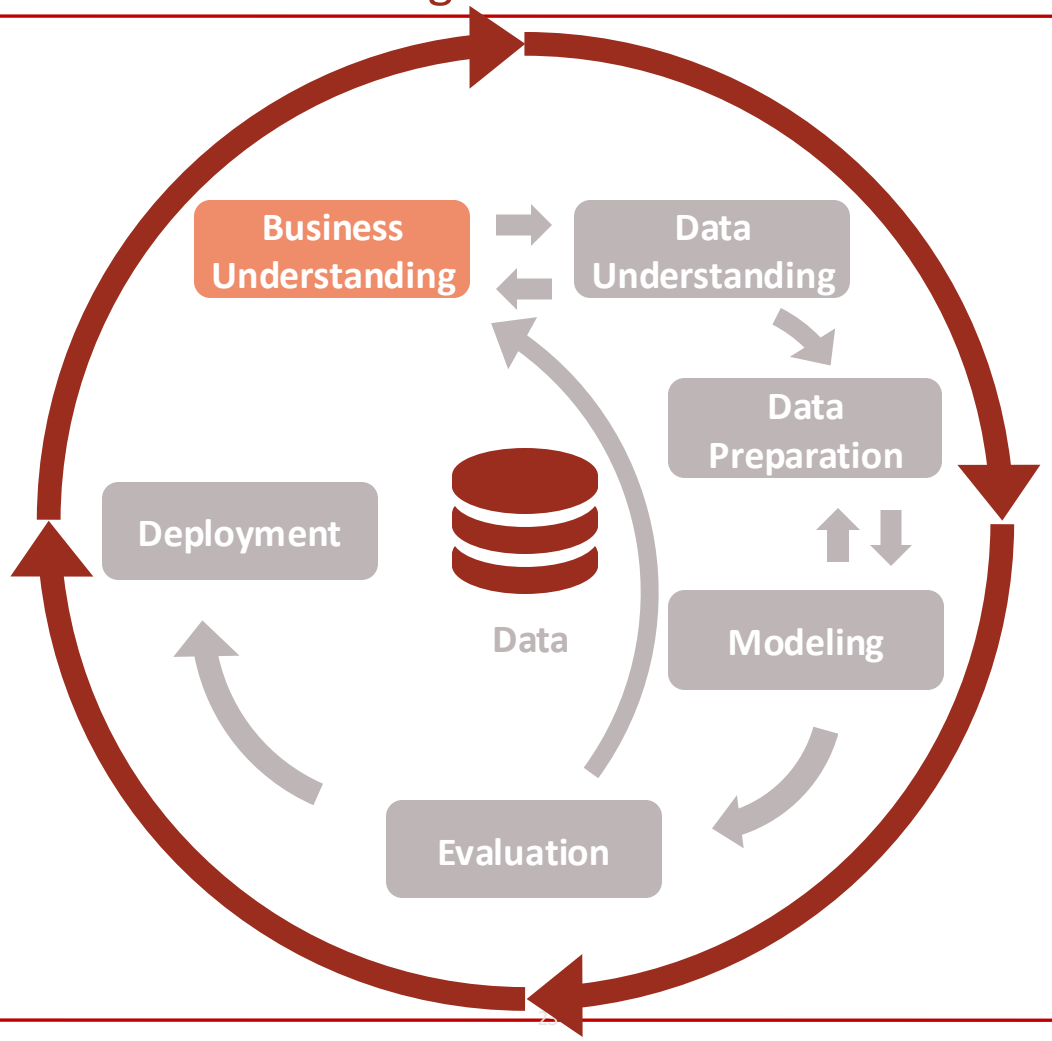
CRISP-DM

Cross-Industry Standard Process for Data Mining

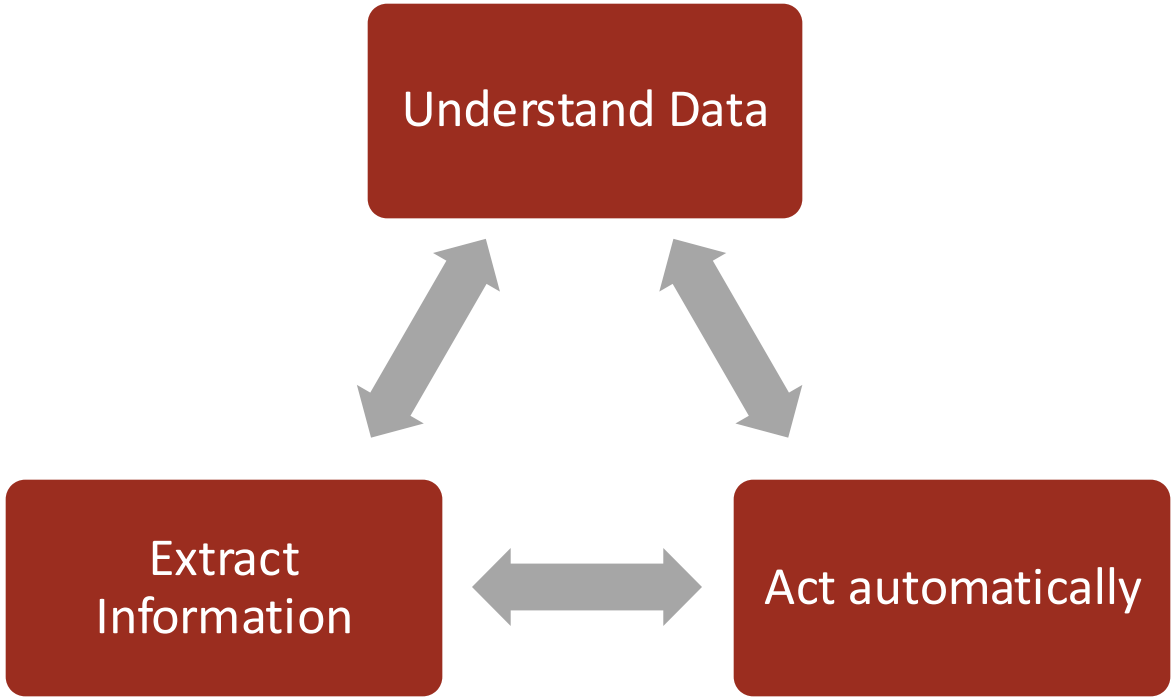


CRISP-DM

Cross-Industry Standard Process for Data Mining



Goals of Data Analytics



Golden Rule of Data Analysis

The most important things to know about data mining

Never try to solve a problem that is more complex than necessary!

Golden Rule of Data Analysis

Never try to solve a problem that is more complex than necessary

„If I predict the sales right, I can predict which shop is profitable.“

- Exact knowledge of sales is unnecessary, only boundary between good and bad matters.

„If I predict what the customer will be doing, I can make him the right offer.“

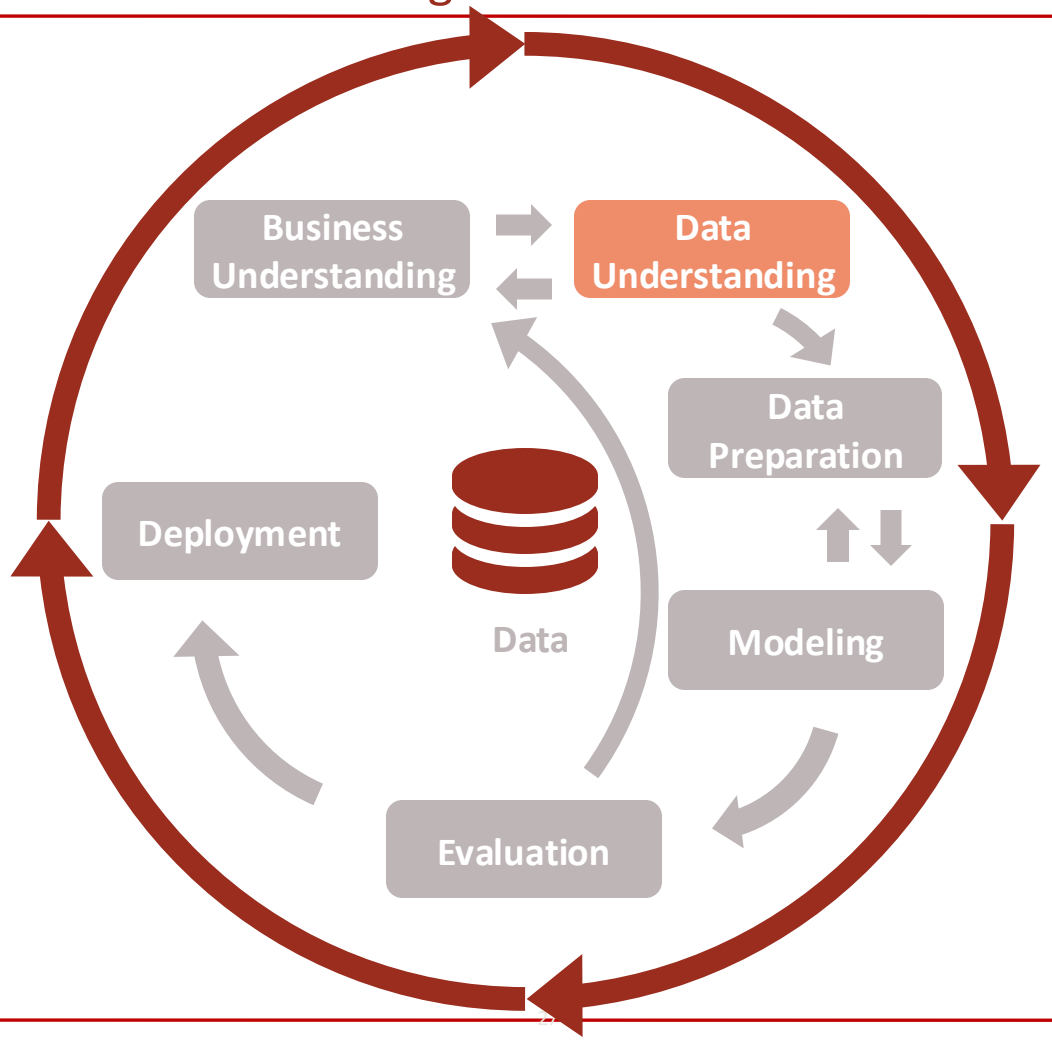
- An offer / marketing campaign may be right for many different customers.

„If I built the right customer groups, I can use them control my business.“

- There may be many different types of customer groups for many different purposes.

CRISP-DM

Cross-Industry Standard Process for Data Mining



Data Analytics and Data

„On a ship there are 26 sheep and 10 goats. How old is the captain?“

- Impossible Data Analytics question!

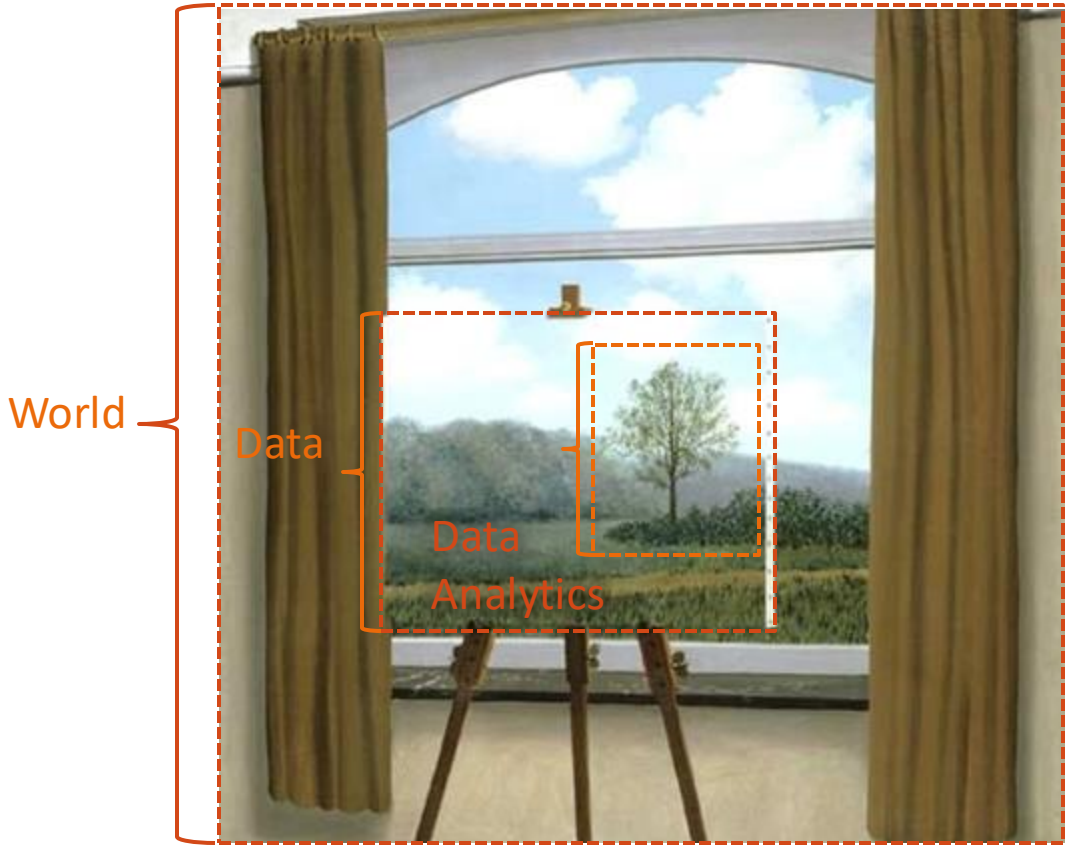
„On Captain Abraham‘s ship there are 26 sheep and 10 goats. How old is captain?“

- Possibly already retired

„On Captain Kevin‘s ship there are 26 sheep and 10 goats. How old is captain?“

- In Germany: Very probably less than 30! The name „Kevin“ practically did not exist in Germany before the „Home Alone“ movies.

The Limits of Data Analytics



Unknown Image Source

Example: Sampling Bias

How things subtly can go wrong…

Display Settings: Abstract Send to:

Psychol Bull. 1991 Jan;109(1):90-106.

Left-handedness: a marker for decreased survival fitness.

Coren S, Halpern DF.

Department of Psychology, University of British Columbia, Vancouver, Canada.

Abstract

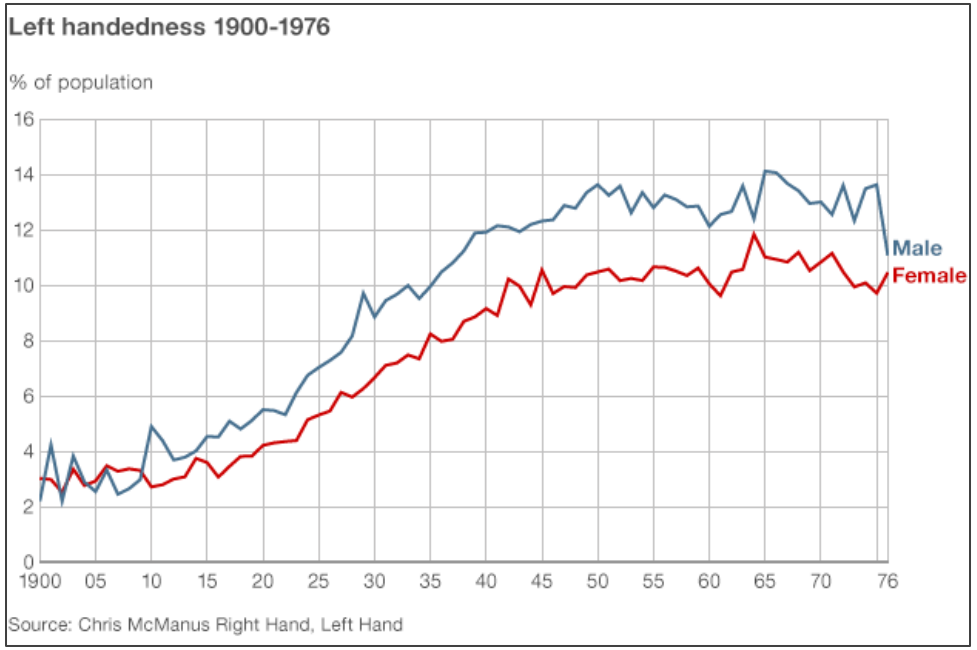
Life span studies have shown that the population percentage of left-handers diminishes steadily, so that they are drastically underrepresented in the oldest age groups. Data are reviewed that indicate that this population trend is due to the reduced longevity of left-handers. Some of the elevated risk for sinistrals is apparently due to environmental factors that elevate their accident susceptibility. Further evidence suggests that left-handedness may be a marker for birth stress related neuropathy, developmental delays and irregularities, and deficiencies in the immune system due to the intrauterine hormonal environment. Some statistical and physiological factors that may cause left-handedness to be selectively associated with earlier mortality are also presented.

Comment in

Do left-handers die sooner than right-handers? Commentary on Coren and Halpern's (1991) "Left-handedness: a marker for decreased survival fitness". [Psychol Bull. 1993]

PMID: 2006231 [PubMed - indexed for MEDLINE]

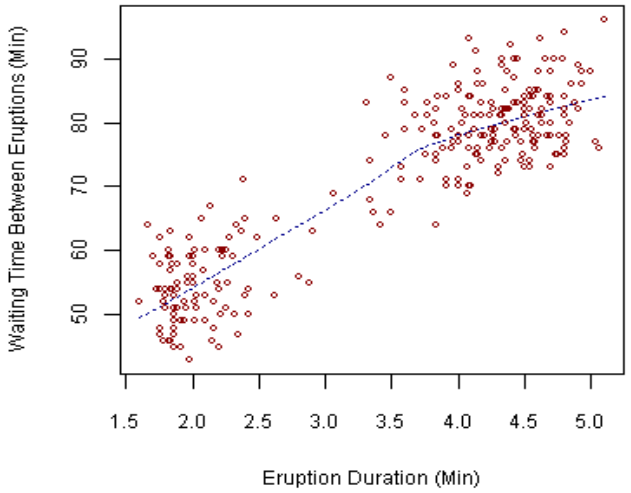
Publication Types, MeSH Terms, Substances

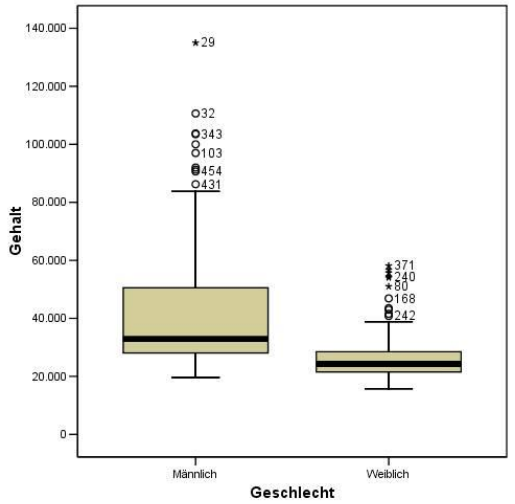


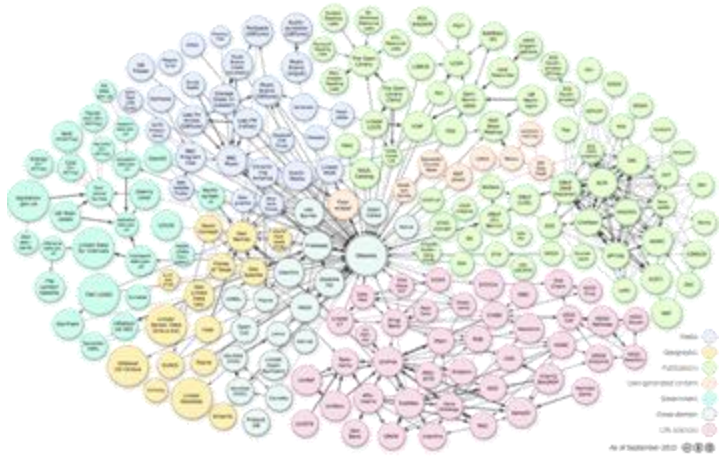
PubMed and paper

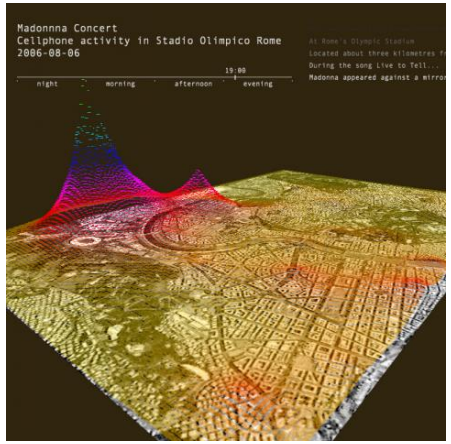
Exploratory Data Analysis

A Powerful Tool to Understand Data





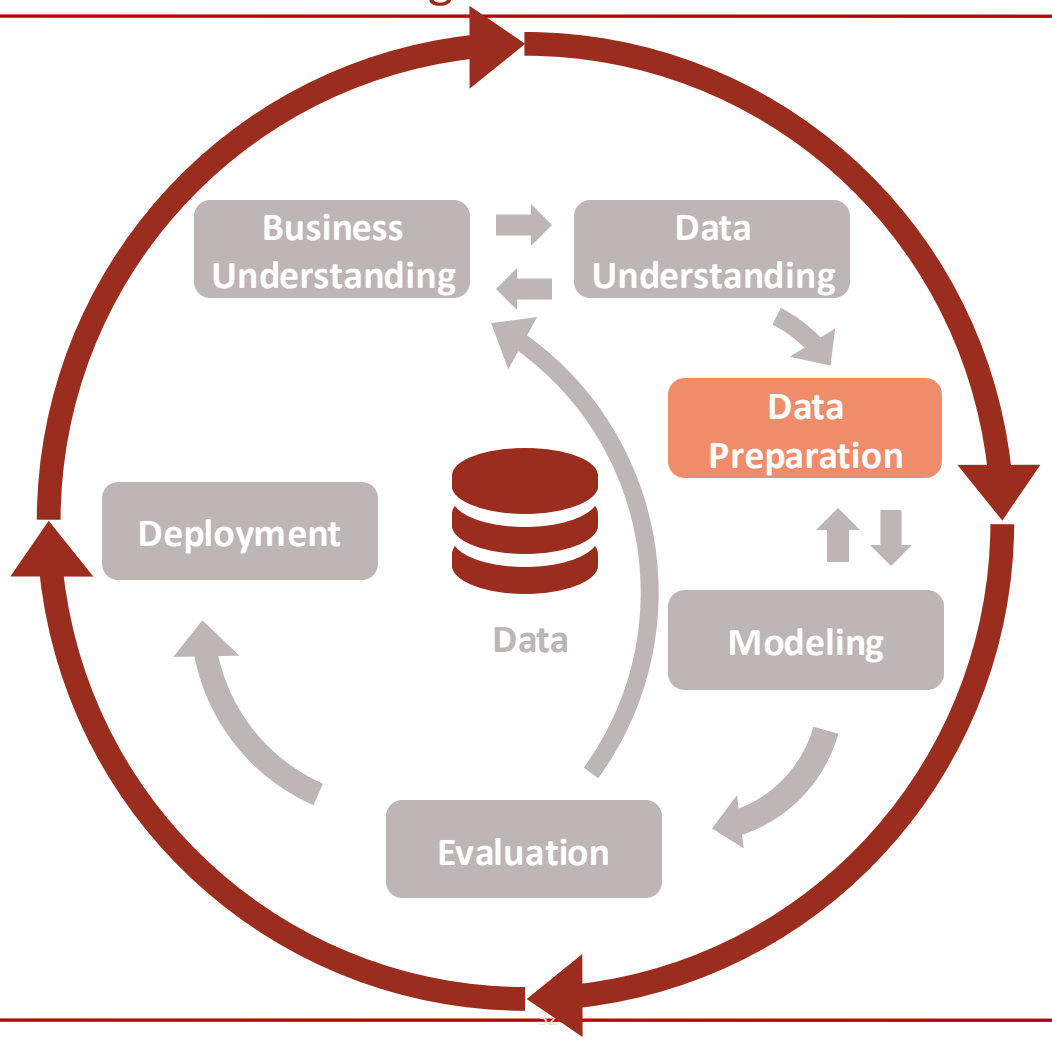




Partly Unknown Image Sources

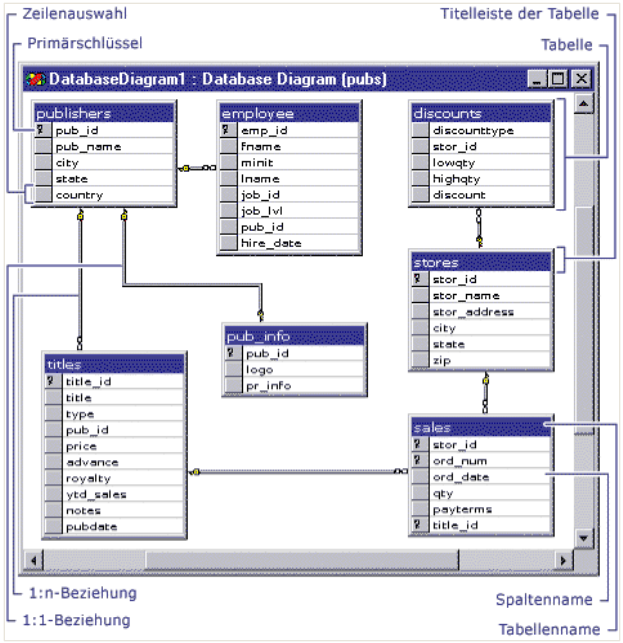
CRISP-DM

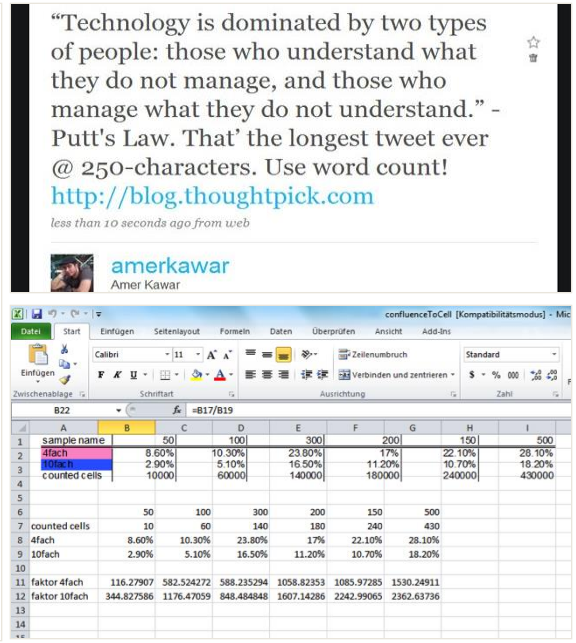
Cross-Industry Standard Process for Data Mining



Data Comes in Many Forms…

- Data Warehouse (this is usually the best case!)
- File (e.g. TXT, CSV, Excel, XML)
- Tweets
- Multimedia





Unknown Image Source

Why Data Representation is the Key

Three Insights about Data Analytics

Carlo Emilio Bonferroni: Problem of multiple comparison

- If one investigates many hypotheses, some of them may appear true by random chance

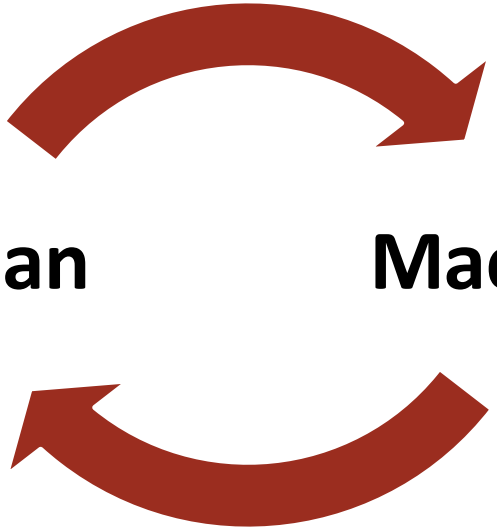
Richard E. Bellmann: Curse of dimensionality

- With increasing dimension, every space is sparse.

George A. Miller: Magical Number 7

- Humans can reason with only keep 7 (+/-2) independent concepts

Roles in the Data Analysis Process



Human

- Understands and evaluates
- Implements strategies
- Intuition and background knowledge about patterns

Machine

- Automatic sifting through data
- Statistical optimization
- Quality guarantees

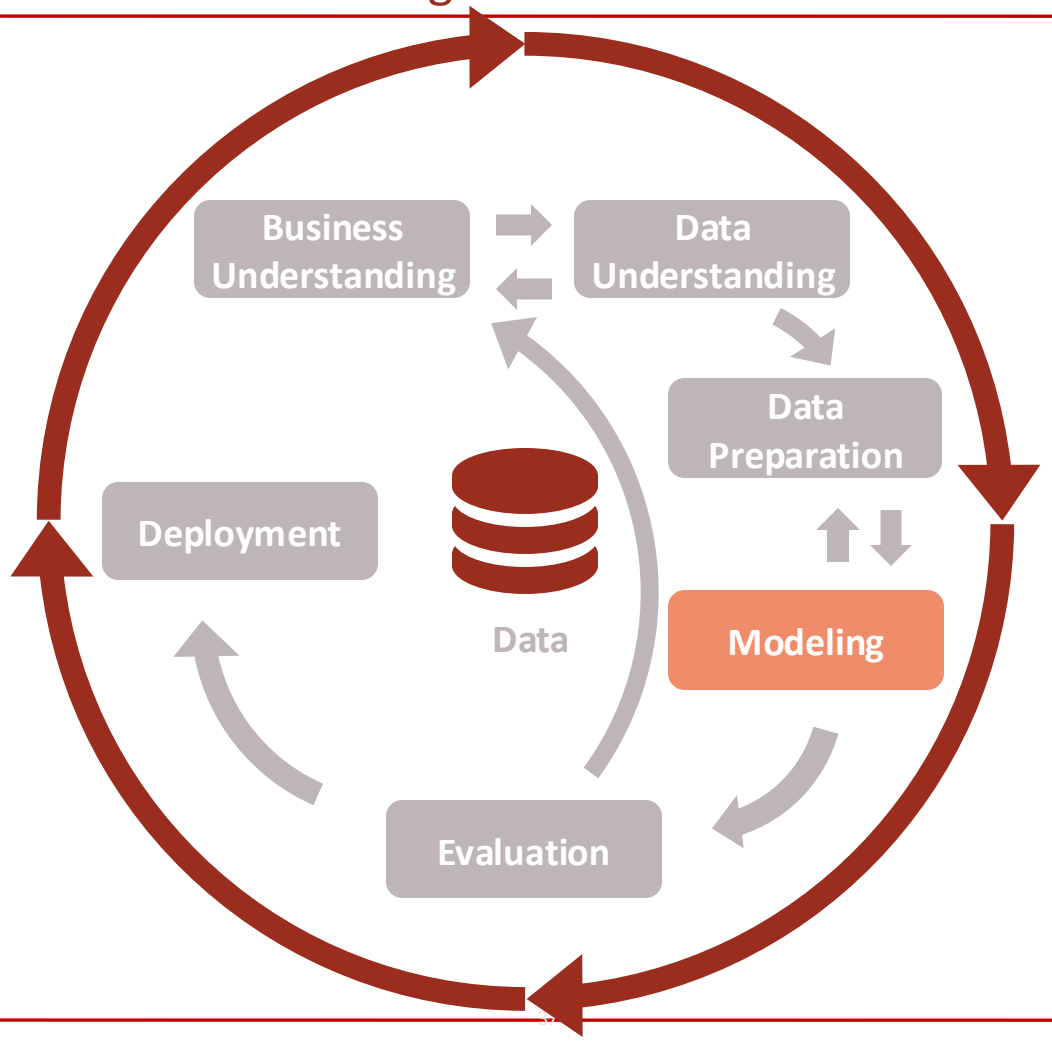
Gartner study, 2013: *"Machines are becoming better at understanding humans and the environment (…). At the same time, machines and humans are getting smarter by working together"*

Data Preparation

- Data preparation is a crucial step for the success of data analysis
- **This is where the data analyst tells the algorithm what she wants!**
- Data preparation is a largely manual process
- A good data warehouse is a very valuable resource

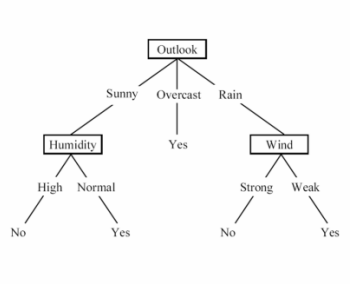
CRISP-DM

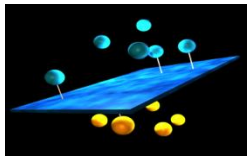
Cross-Industry Standard Process for Data Mining

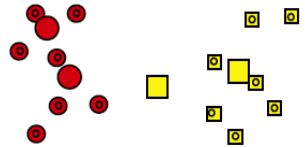


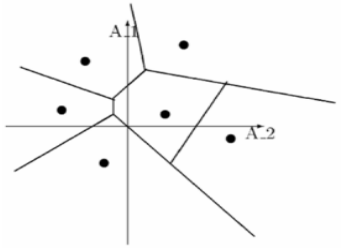
Data Analytics Algorithms

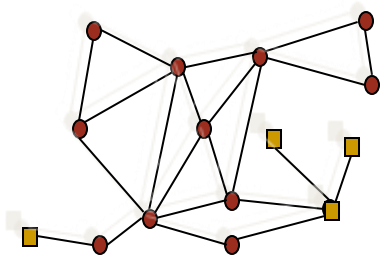
- **The good news:** No need to worry. Thousands of Data Analytics approaches exist, all available within commercial and open source toolkits!
- **The bad news:** There is no silver bullet!

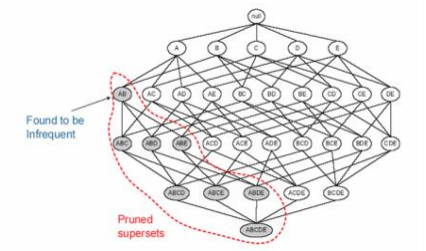


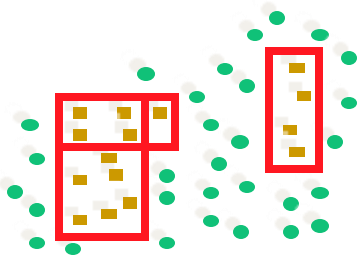












Unknown Image Source

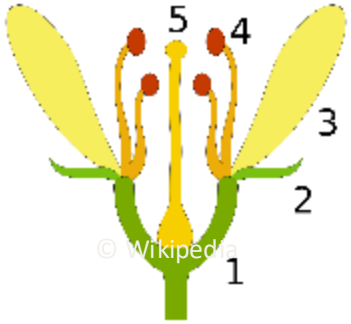
We start with an example...

Consider some measured properties of different iris flowers

Iris ***setosa***

Iris ***versicolor***

Iris ***virginica***

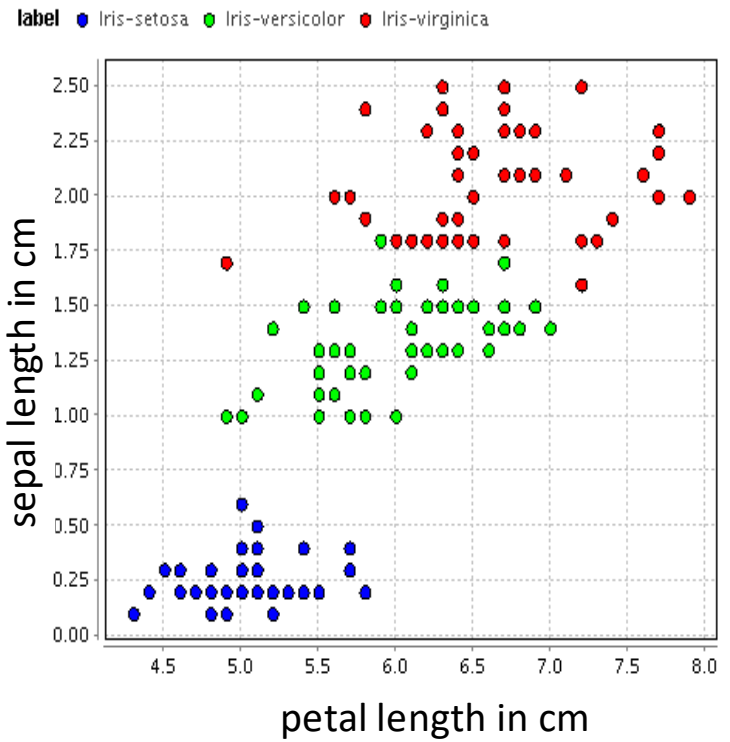


1. Receptacle
2. Sepals
3. Petals
4. Stamina
5. Carpels

Unknown Image Source

Let this be the collected data

Row	Sepal length	Petal length	Iris type
1	5.4	1.5	Setosa
2	5.5	1.4	Setosa
3	6.3	4.7	Versicolor
4	6.1	4.7	Versicolor
5	6.3	6.0	Virginica
6	7.7	6.7	Virginica
..	..	..	..



Unknown Image Source

Different analytical questions come to mind

Such as..

- Can we predict the iris type?
- Can we predict the sepal length given the petal length?
- Do the flowers naturally fall into groups?

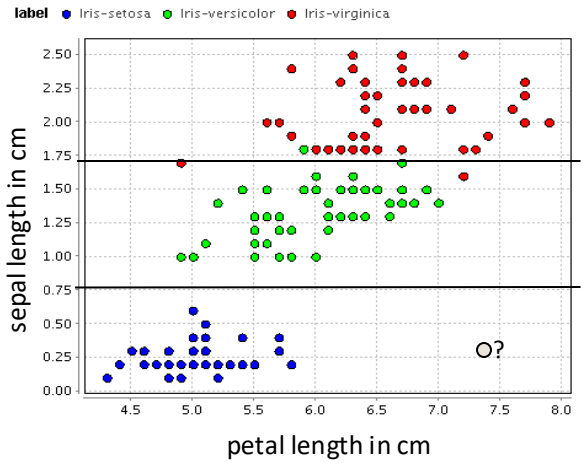
- **To answer these questions, we model a mathematical aspect of the data, e.g.:**
  - a set of thresholds
  - a linear equation
  - cluster centroids

Henceforth a model is a mathematical representation of some aspect of the data

The most common model classes

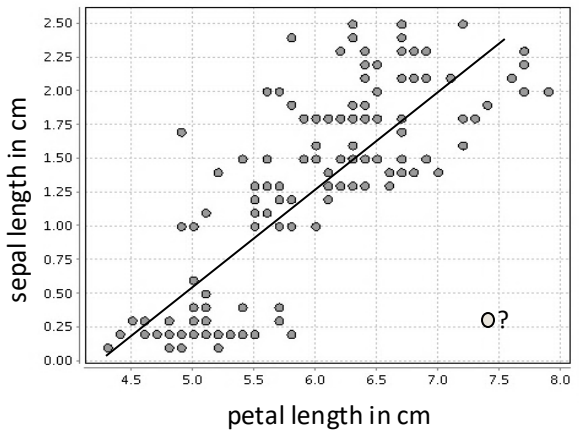
Each model class answers different questions

Classification



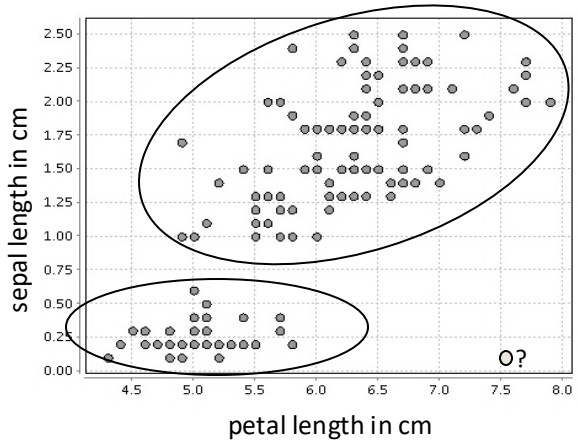
To which class does an object belong?

Regression



Given some value what is the other?

Clustering



Are there any groups?

Unknown Image Source

Data Analytics Tools and Plattforms

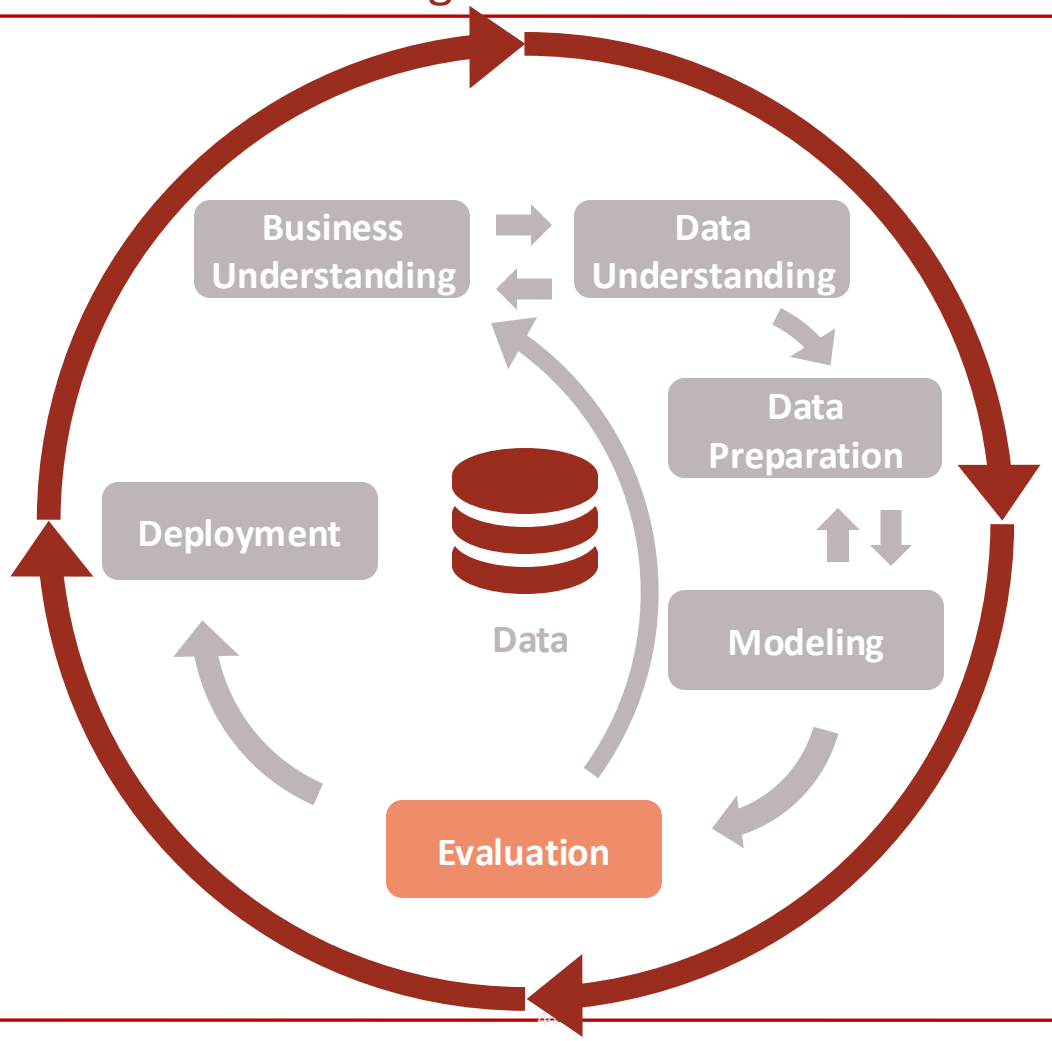
Business Intelligence Tools with Data Analytics Capabilities

Data Analytics Tools

Frameworks for Data Analytics

CRISP-DM

Cross-Industry Standard Process for Data Mining



Evaluation Strategies

Statistical Evaluation

- Testing the results rigourously from a mathematical perspective
- Huge amount of statistical literature for different quality measures -> own part of the lecture
- Most important point: always test results on NEW data!
- Pre-condition for semantic evaluation

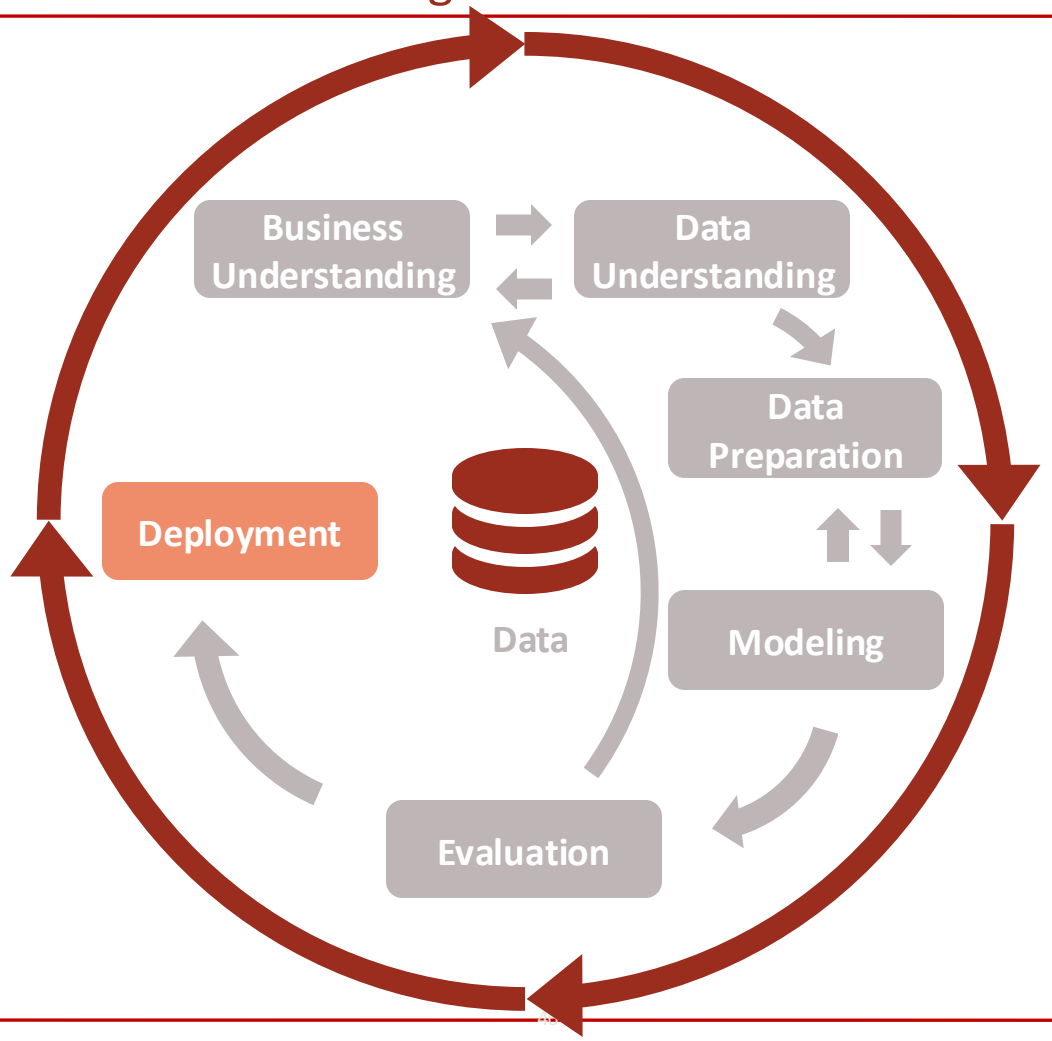
Semantic Evaluation

- Closes the loop back to business understanding for the specific application
- Do I understand the results? Do the business users understand them?
- Are the results useful for the business question? Are they good and reliable enough?
- Can the model be used in automatic contexts?
- How should outliers be handled? How can they be identified?

Personal evaluation:
What did I learn? How can I do it better next time?

CRISP-DM

Cross-Industry Standard Process for Data Mining

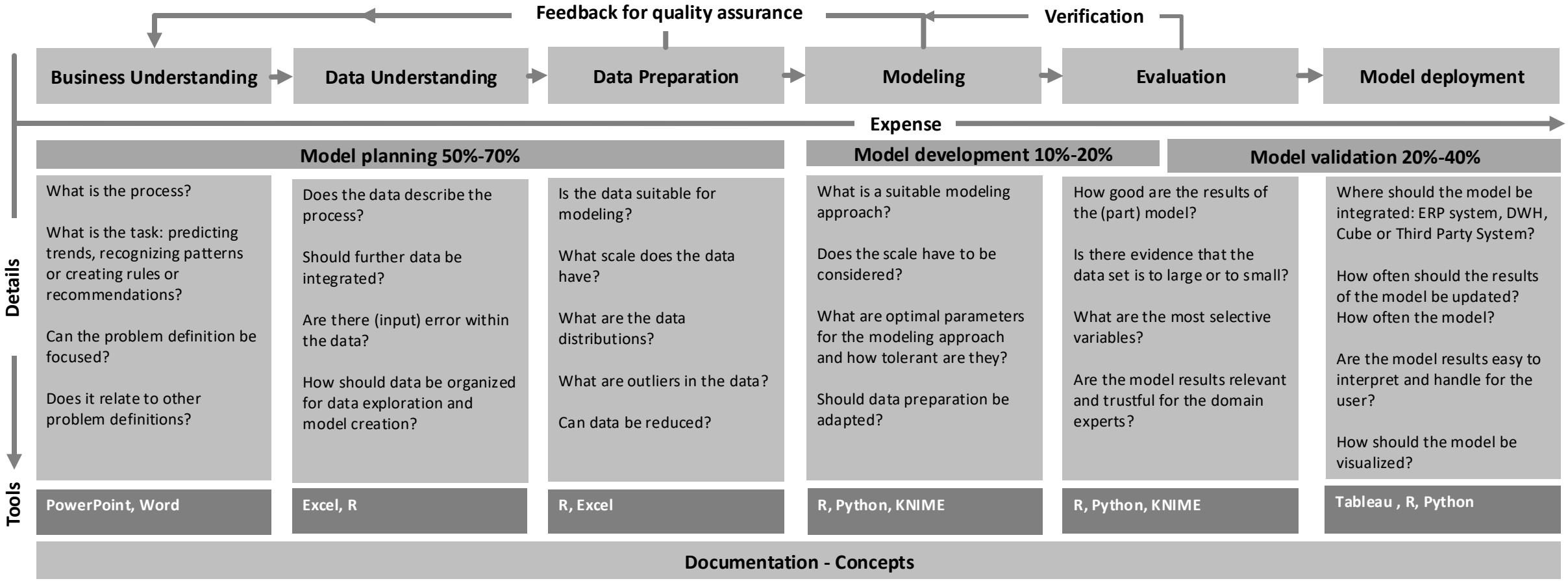


Deployment

- Model integration
  - Technical
  - Business processes
- Visualization of model results
- Interpretation of model results
- Monitoring the model
- Updating the model

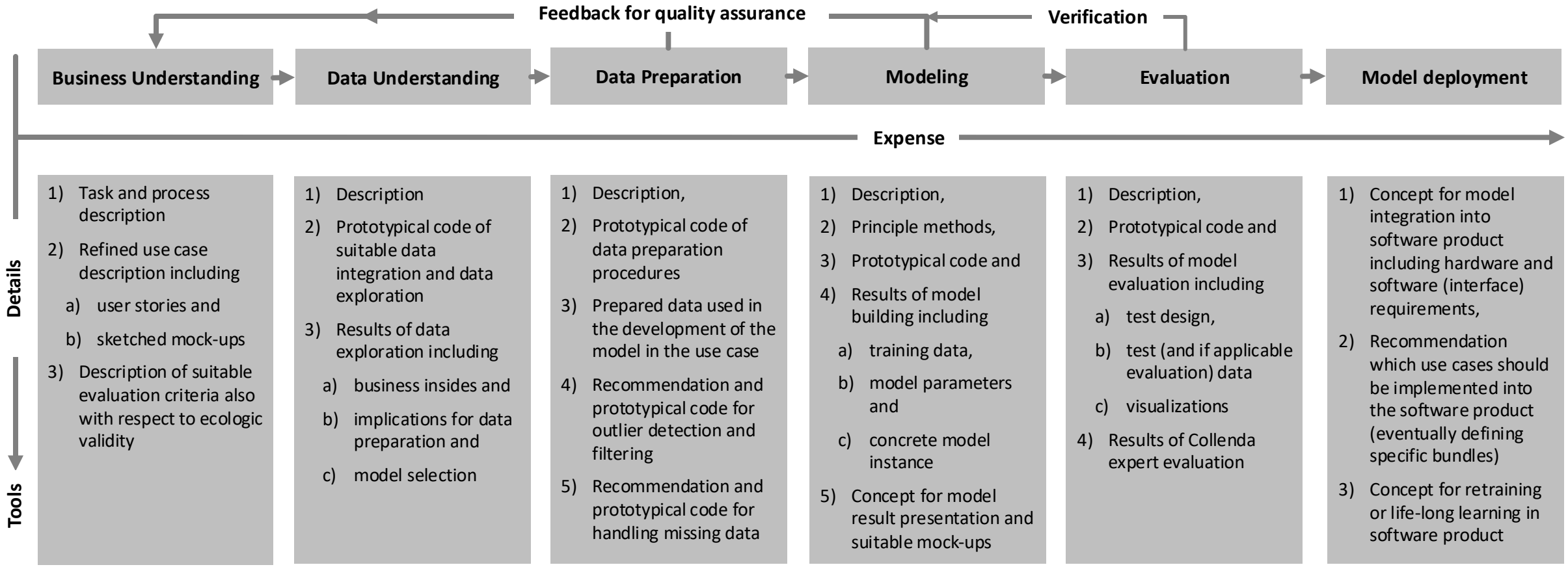
The steps of the CRISP-DM cycle answer different questions

Steps of analytical process in model creation



and thus different deliverables are developed there

Generic deliverables for all use cases



Your Project work

- You will use Colab and Python on real data
  - The project is a practical solution to the data
  - See how a complete project looks like
- You will get a description for the real data
- 50 % of the final grade

How to structure your work on the project

- Think about the research question and your hypotheses…
- … and write them down!
- Formulate a testing concept how you can analyze the hypotheses on the given real world data (which algos are you using, what programms etc.)
- Do the tests and store the results
- Discuss how the results relate to your hypotheses
- Draw your conclusions!
- Document everything and make a nice presentation to show the others your research.

Questions?

SUMMARY

What is Data Analytics?

Data Analytics is:

Data Analytics is not:

Unknown Image Source