Chapter 1

Basics

The slides were taken in large parts from the book "Knowledge Discovery in Databases" by M. Ester, J. Sander and the slides available for this purpose from the Web as well as from the Institute AIFB of the University of Karlsruhe (R. Engels, M. Erdmann, A. Hotho, A. Mädche, S. Staab, R. Studer, G. Stumme).

Content of this chapter

1. Basic Terms
2. KDD/Data Mining/Data Science Process
3. Statistics
4. Databases, Data Warehouse and OLAP
5. Preprocessing

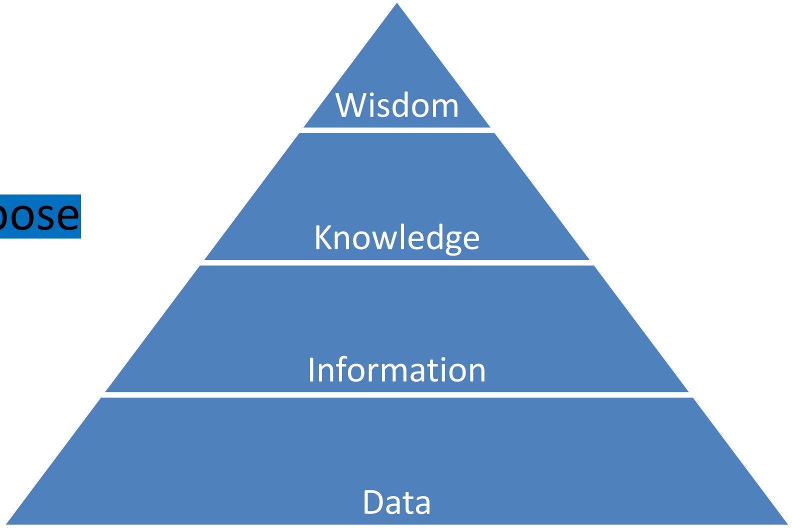
1.1 Basic Terms

- Data - Information - Knowledge
- Knowledge Discovery in Databases (KDD)
- Data Mining (DM)
- Big Data
- Data Science

Data - Information - Knowledge

"Data is not information,
information is not knowledge,
knowledge is not wisdom." [C. Stoll]

- Data
  Raw data (measurements, "facts")
- Information
  Significant, summarized data for a specific purpose
- Knowledge
  Knowledge that people are aware of
- Be aware:
  Many contradictory definitions exist



DIKW Pyramid

History of Data Science

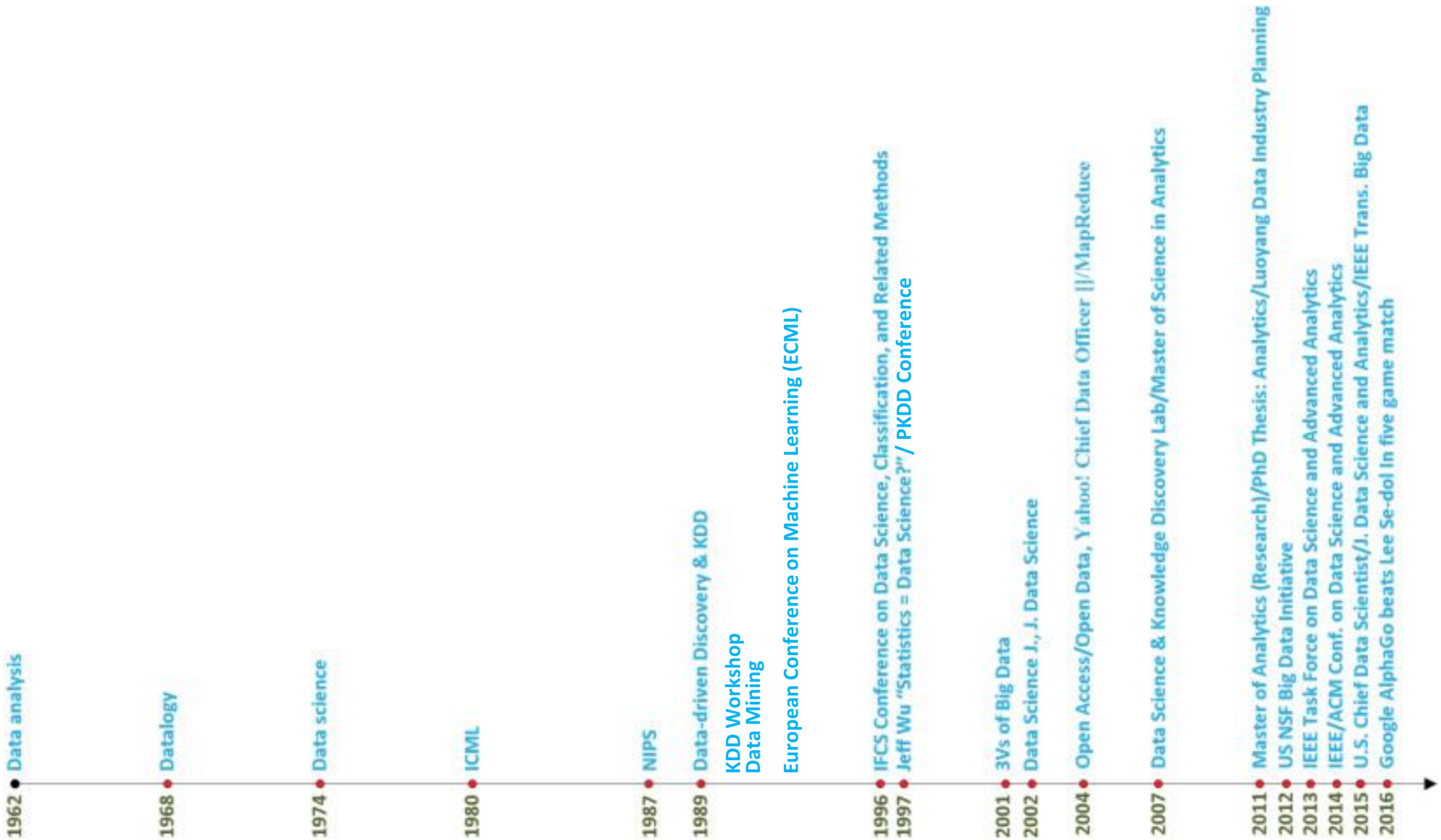
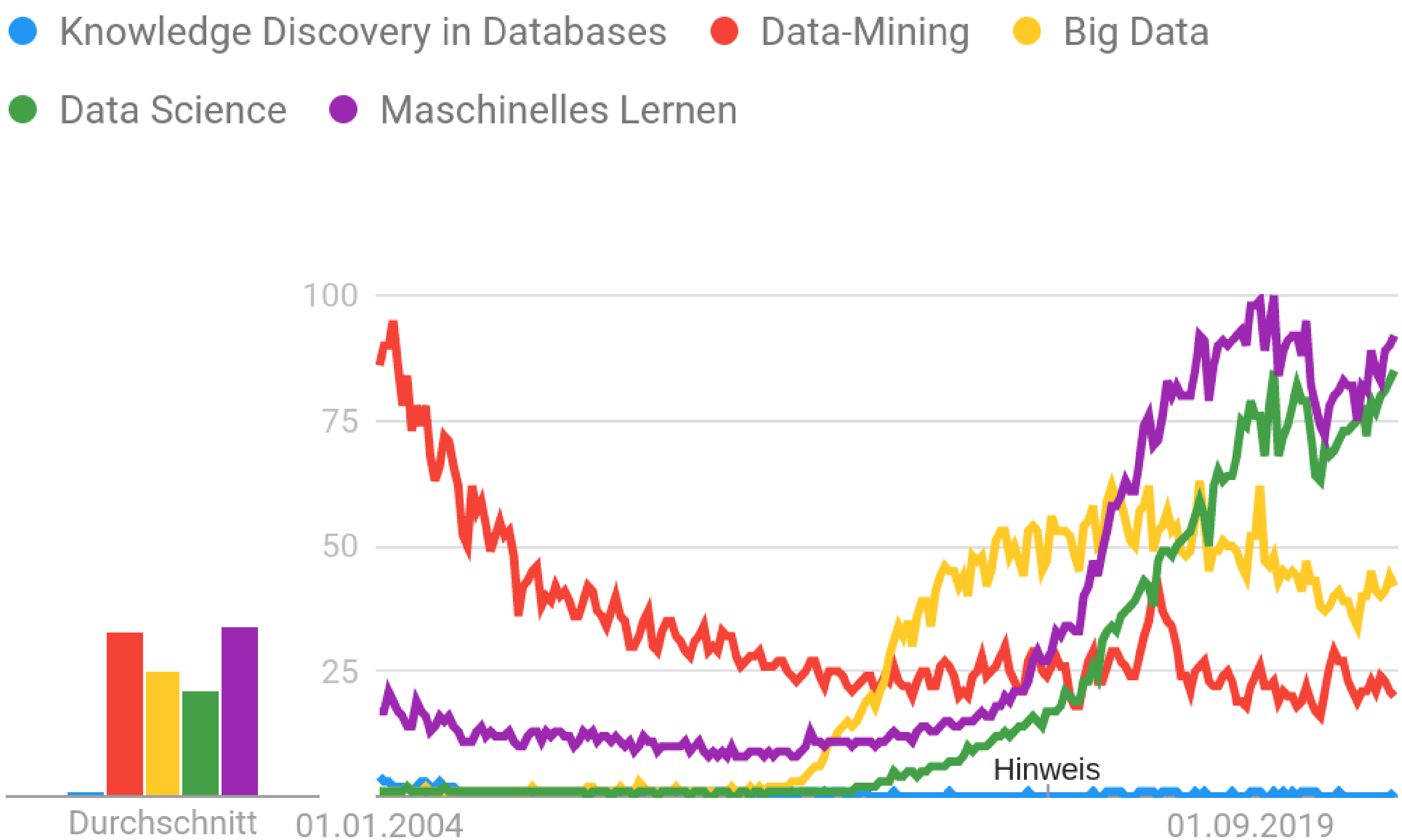


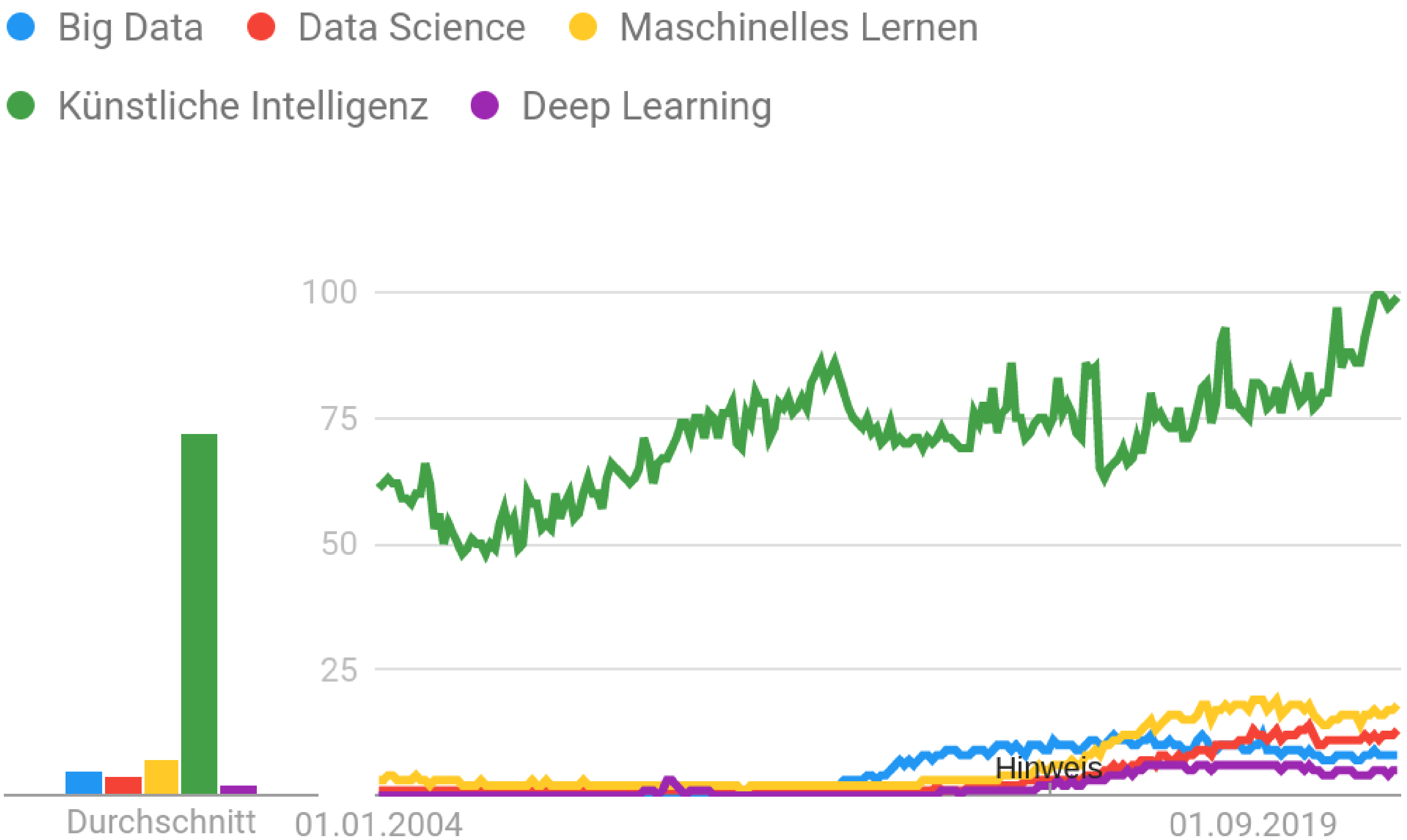
Image Source: Cao, Data Science a Comprehensive Overview, ACM Computing Surveys, Volume 50, Issue 3, 2017

Search History



Google Trends from Jan 1st 2004 to April 28th 2022

Search History



Google Trends from Jan 1st 2004 to April 28th 2022

Knowledge Discovery in Databases (KDD)

- Fayyad et al.* define KDD in 1996 as

  *The nontrivial process of identifying **valid**, **novel**, potentially **useful**, and ultimately **understandable** patterns in data.*

- The four characteristics are explained as follows:

Valid	The found patterns also apply for new data
Novel	The system/user did not know that this pattern existed
Useful	The result can be used to solve a given task
Understandable	The user should know how/why it works (however, this is a subjective measure)

- Fayyad et al. state that

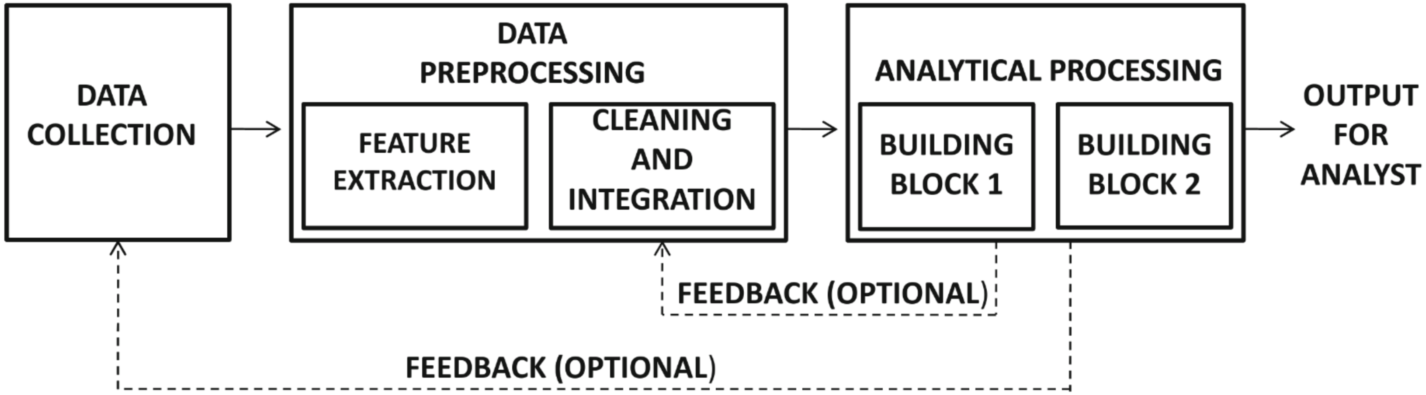
  ***Data mining** is a particular step [in KDD] – application of specific algorithms for extracting patterns (models) from data.*

The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, Vol. 39, No. 11, 1996

Data Mining

Aggarwal* defines Data Mining in 2015 as

> *Data Mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. […] "Data mining" is a broad umbrella term that is used to describe these different aspects of data processing.*



(A standardized Data Mining process will be discussed later)

* Aggarwal, Data Mining: The Textbook, Springer, 2015

Big Data

- De Mauro et al.* define Big Data in 2016 as

  Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.

- Big Data Analytics is similar to Data Mining, but especially focuses on large data volumes where "classical methods" can not be used efficiently

* De Mauro et al., A Formal Definition of Big Data Based on its Essential Features, Library Review, Vol. 65, Iss. 3, 2016

Data Science

Cao* defines Data Science in 2017 as

> *From the disciplinary perspective, **data science** is the new **interdisciplinary field** that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology **to study data** and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to **transform data to insights and decisions** by following a data-to-knowledge-to-wisdom thinking and methodology.*

but also gives another (more simple) definition:

> *Data Science is the science of data.*

* Cao, Data Science: A Comprehensive Overview, ACM Computing Surveys, Vol. 50, No. 3, 2017

Relationship of Data Science and Data Mining

Data science, also known as **data-driven science**, is an interdisciplinary field of **scientific** methods, processes, algorithms and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured, similar to **data mining**.

https://en.wikipedia.org/wiki/Data_science

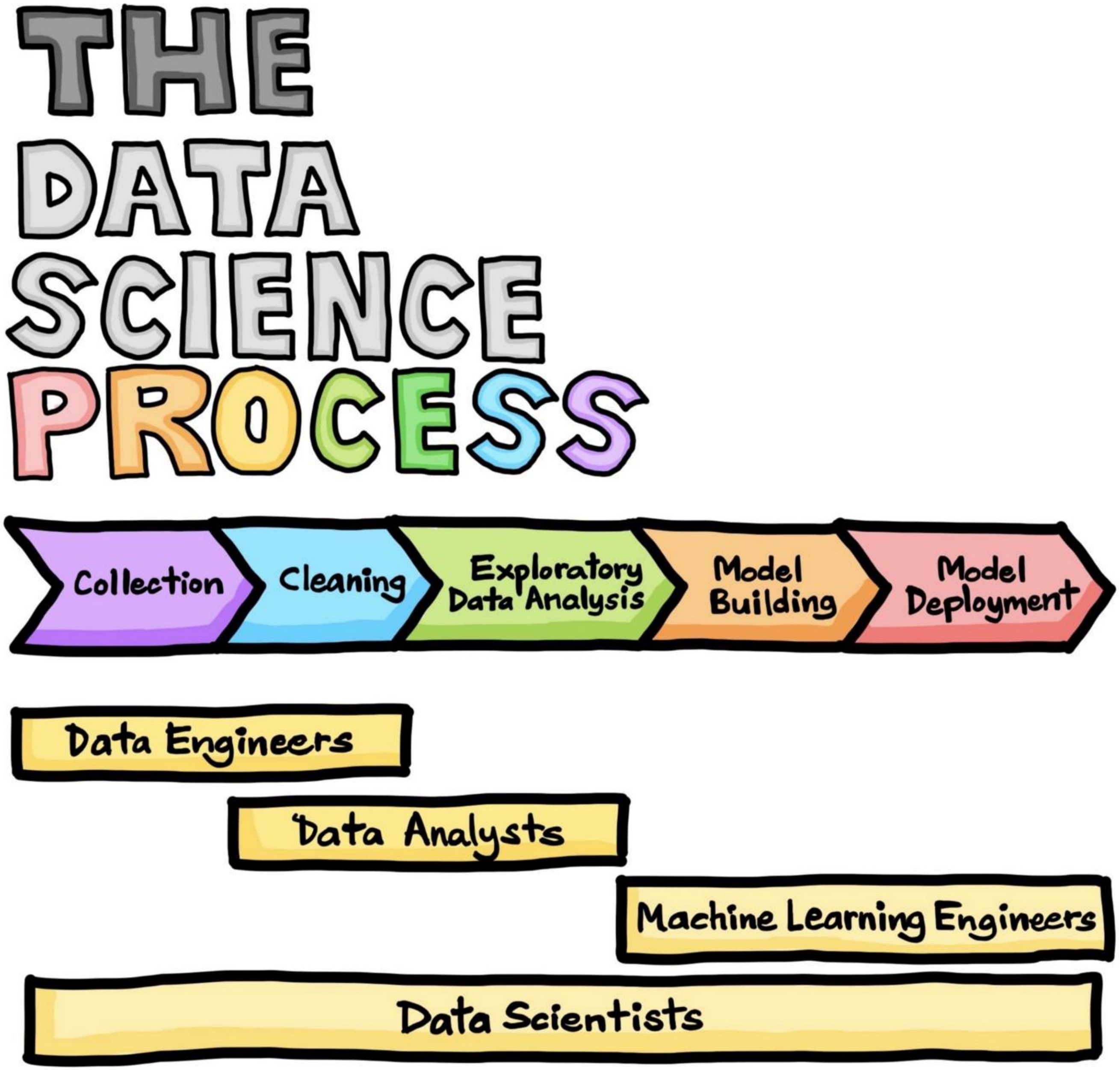
[Dhar; 2013]

"„... At a high level, **data science** is a set of fundamental principles that support and guide the principled **extraction of information and knowledge from data**. Possibly the most closely related concept to **data science is data mining** - the actual extraction of knowledge from data via technologies that incorporate these principles. ...“

[Provost & Fawcett; 2013]

Image source: https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b, accessed on April 28th, 2022

1.2 Data Science Process


THE
DATA
SCIENCE
PROCESS
Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

The Data Science Process

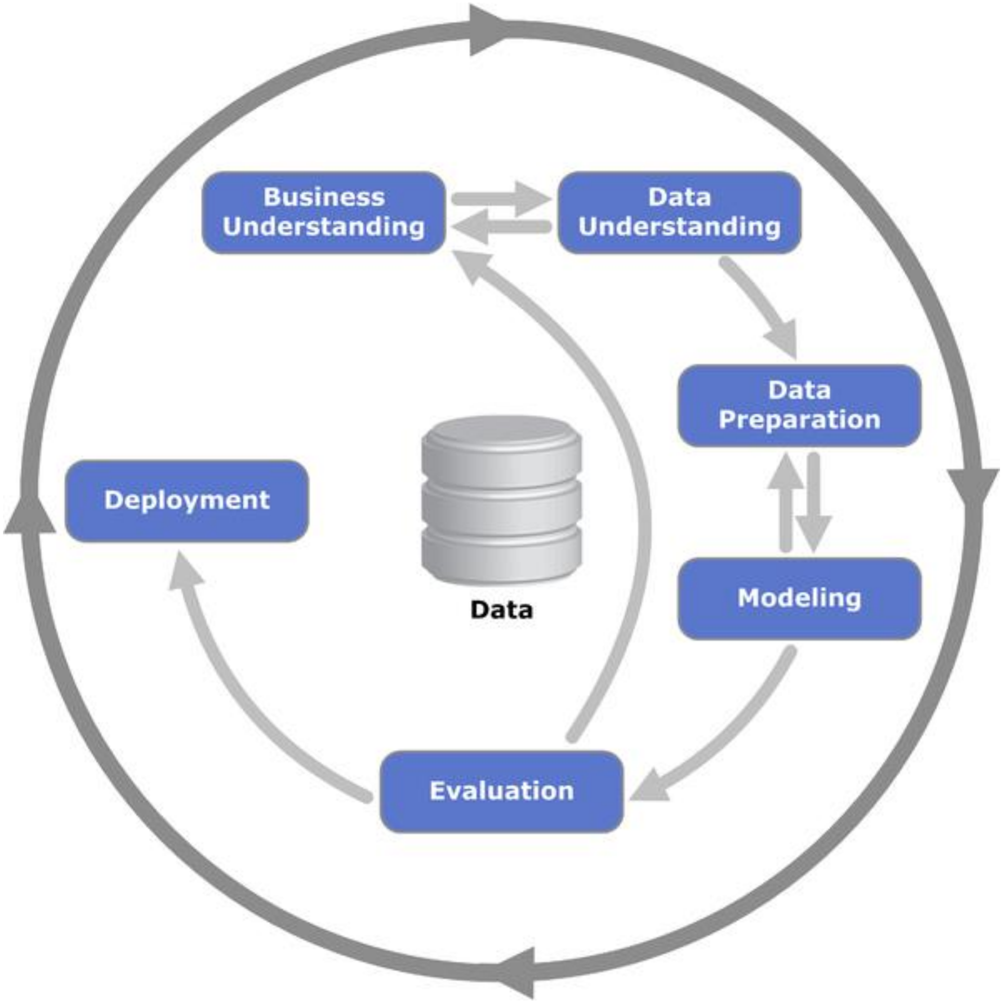
- The process must be related to the task and the user

- The developer needs knowledge about **databases, data analysis** methods and the **application area**

- The process is **interactive** and **iterative**
  - No full automation
  - Results have to be evaluated before making a decision
  - Some steps might be repeated depending on the results

- One well know process definition is the open standard process model CRISP-DM

The CRISP-DM Model



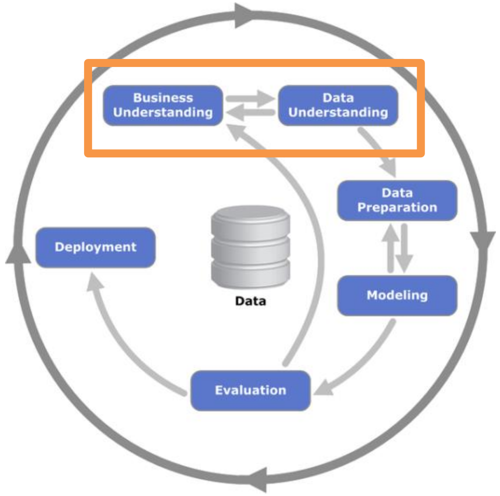
Cross Industry Standard Process for Data Mining

Main phases
(top-level processes)

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Business Understanding, Data Understanding

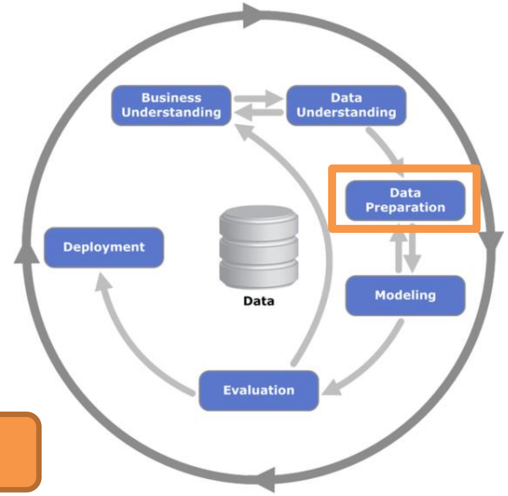
- Understanding the given application
- Defining the goal(s) of the Data Mining project
  - What should be achieved?
- Acquiring data from source(s)
- Clarifying data management
  - File System or DBS?
- Selecting relevant data



Preprocessing

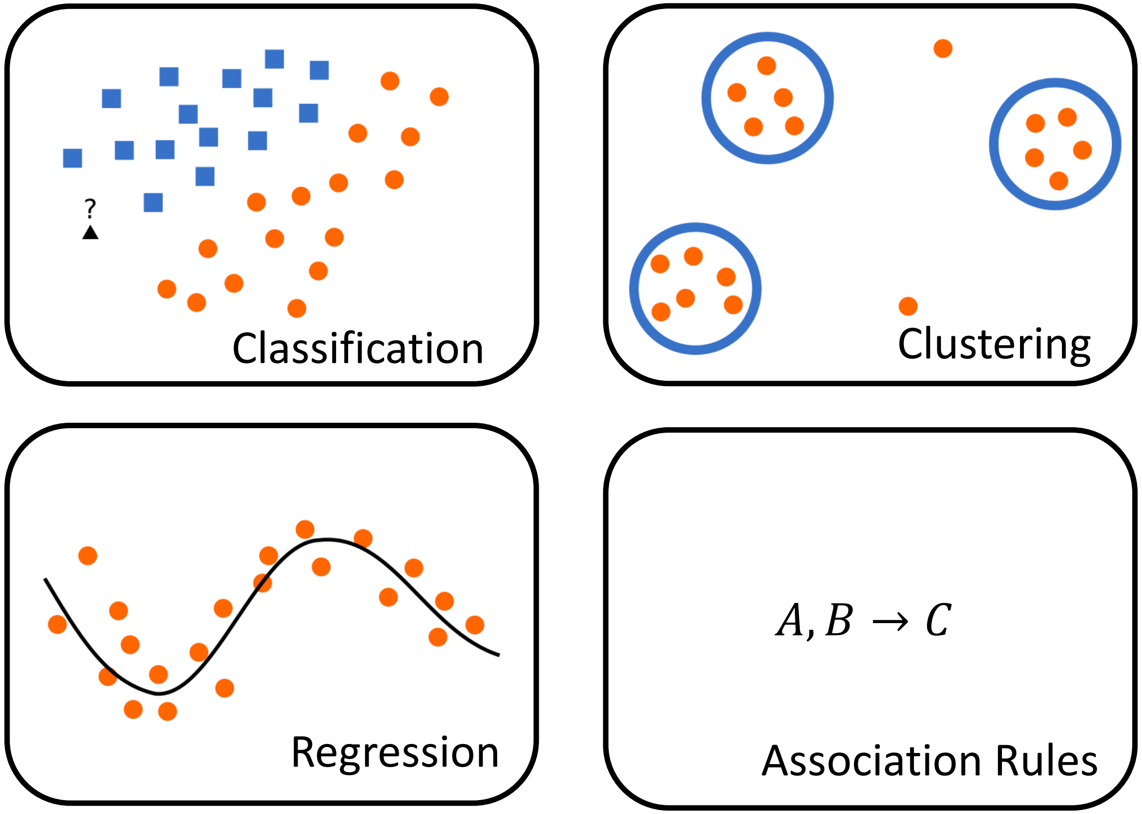
- Integrating data from different sources
- Checking consistency
- Cleaning

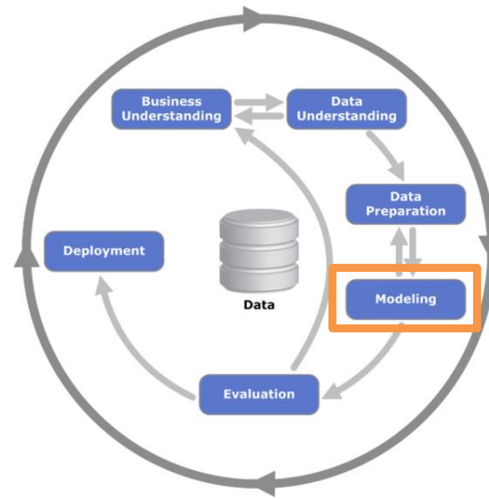
- Discretizing numerical features
- Generating derived features
- …



➔ More about this in Chapter 1.5: Preprocessing

Data Science (Modeling): Methods



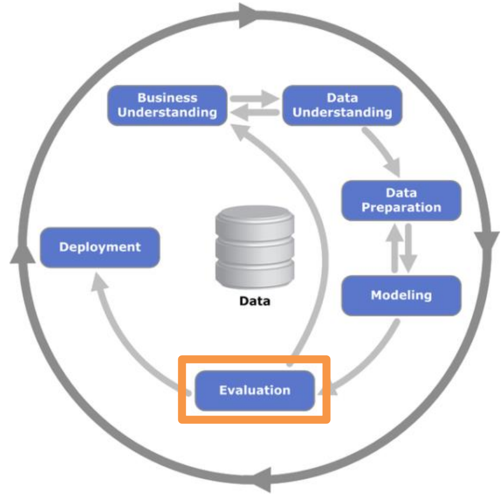


Other tasks:
Subgroup Discovery, Outlier Detection, Segmentation, …

➔ More details in later chapters

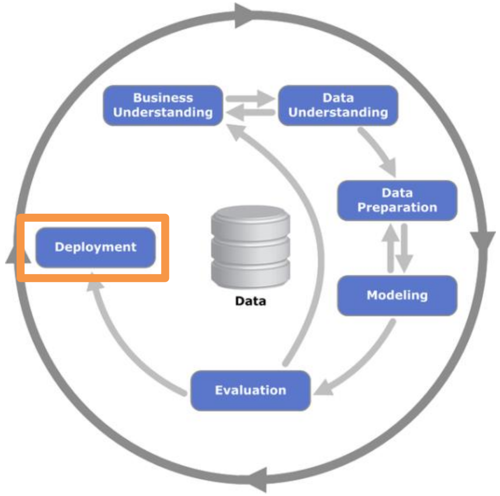
Evaluation

- Presenting the found patterns
  (often through appropriate visualizations)
- Evaluating patterns by the user
  - Predictive power of patterns and/or models
  - Pattern known or surprising?
  - Patterns and/or models applicable to many cases?
- If negative evaluation, then renewed data science with
  - Different parameters, different methods , different data
- If positive evaluation, then
  - Integration of the found knowledge into the knowledge base
  - Use of the new knowledge for future Data Science processes



Deployment:
Creation of a Business Application

- Planning the use of the Data Mining application
  - Creation of a plan for the introduction of the application
- Planning of monitoring and maintenance
  - When should models no longer be used?
  - Do business objectives change over time?
- Preparation of the final report
  - Who is the target group for the presentation?
- Review of the project
  - Summary of the most important
    Knowledge and experience
  - Integration of the project results
    into the strategy of the entire company



1.3 Statistics

Features

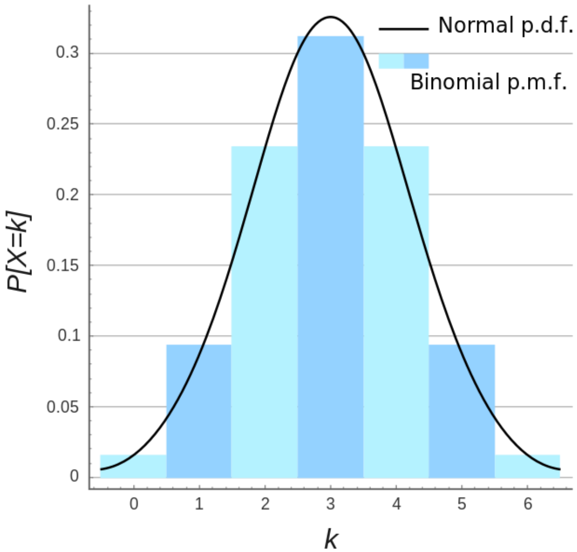
- A single entry from a dataset is called **instance** or **sample**
- A single property from an instance is called **attribute** or **feature**
- A single feature has the same **data type** for all samples in a given dataset, but each feature can be of a different type

Data Type	Possible values	Examples
Binary	0,1	Questions: Yes, No Students: Bachelor, Master
Categorical	Cat0, Cat1, Cat2, …, CatK (no order)	Colours: Red, Green Blue, … Blood types: A, B, …
Ordinal	0, 1, 2, …, K (explicit order)	Clothing: S, M, L, … Surveys: --, -, 0, +, ++,
Numerical	Any number	Price: 10 € Any physical quantity: m, kg, s, …

Basic statistical terms

(Knowledge of the terms is assumed for this lecture)

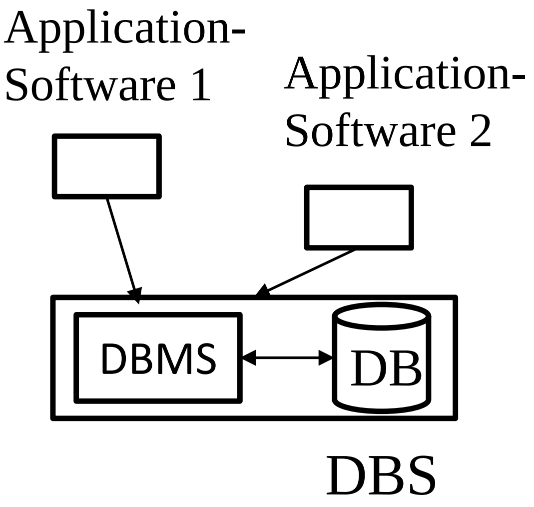
- (Arithmetic) mean, median, mode
- Variance, standard deviation, sample variance
- Expected value
- Relative frequency / empirical probability
- Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Bayes' theorem: $P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{P(B)}$
- Binomial distribution, normal distribution
- Correlation coefficient (Pearson, Spearman)



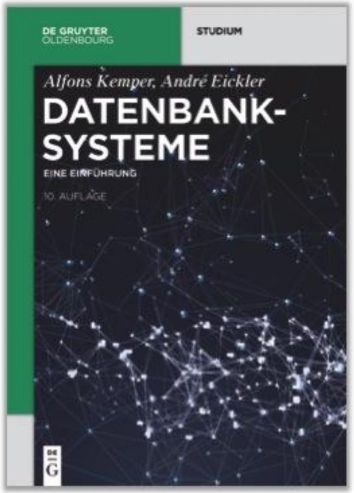
http://en.wikipedia.org/wiki/Binomial_distribution

1.4 Database Systems, Data Warehouses and OLAP

Database Systems



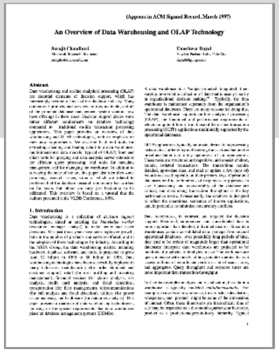
- Database System (DBS):
  Software system for permanent storage and for efficient searching in large amounts of data
- Database (DB):
  Collection of all data and the corresponding descriptions
- Database Management System (DBMS):
  System to manage database (access control, updating of contents)
- Query languages (for relational databases: SQL)
- DBMS determines the most efficient processing
  - Query plan as *operator tree:*
  - Optimization of the tree by *heuristic rules* and *cost model*

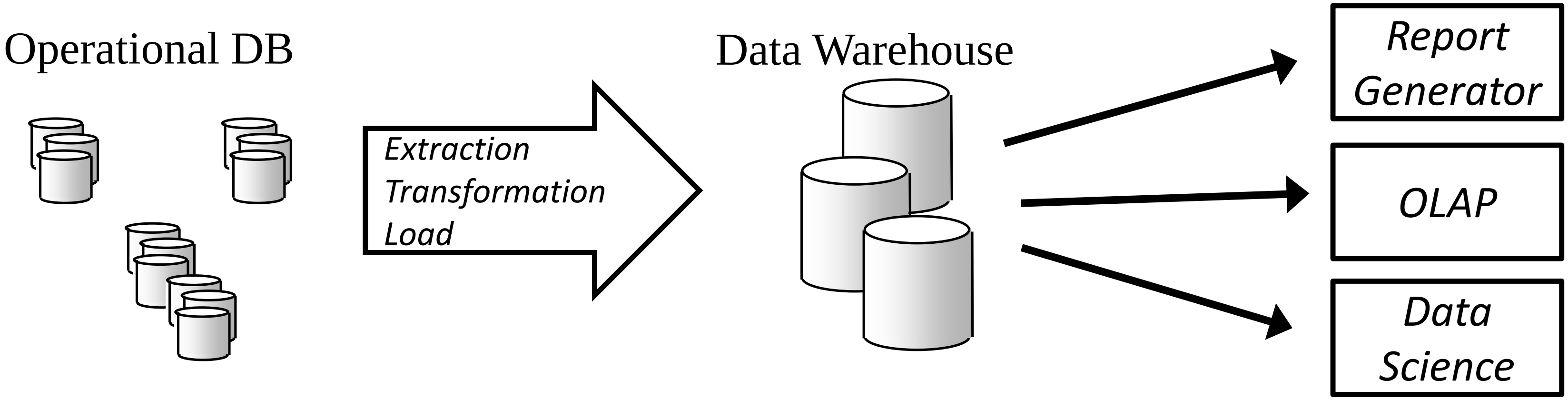


New Approaches for Big Data

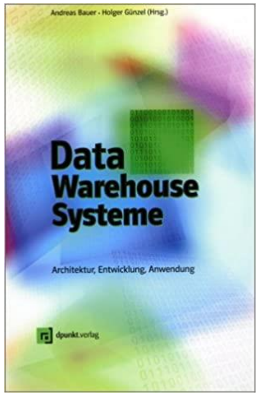
- NoSQL (not only SQL)
  - Graph-based databases (e.g. Neo4J)
  - Document based databases (e.g. MongoDB)
  - Databases based on Hadoop (see below) (e.g. HBase)
  - Key-Value Store (e.g. Voldemort)
- In Memory Databases
  - e.g. SolidDB
- Other techniques
  - Map Reduce (Apache Hadoop)
  - Spark
  - Flume (data streaming)

Data Warehouse [Chaudhuri and Dayal; 1997]





- Permanent + integrated collection of data (mostly in databases)
- Separated from the operational business
- from different sources
- for the purpose of analysis or decision support

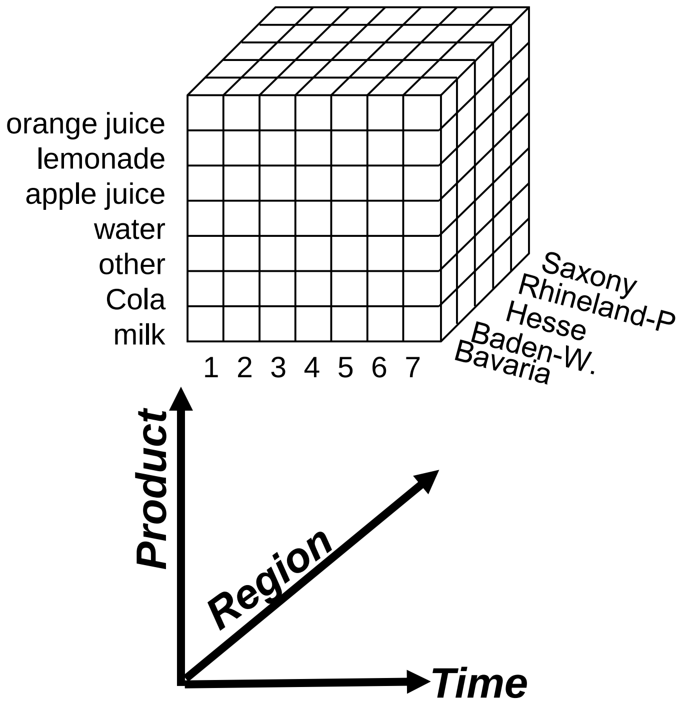


OLTP and OLAP

Online **T**ransaction **P**rocessing (OLTP)	**On**line **A**nalytical **P**rocessing (OLAP)
Direct interactions with the operative DB	Interaction with the Data Warehouse
System is indented for • Storage for frequently updated data • Daily business transactions	System is indented for • Decision support • Data understanding • Data preparation
Functions allow • High number of short, atomic isolated, recurring transactions • Guaranteed Data Integrity	Functions allow • Fast, interactive, access to data • From “any” business-relevant perspective (dimensions) • On different aggregation levels

OLAP Multidimensionality

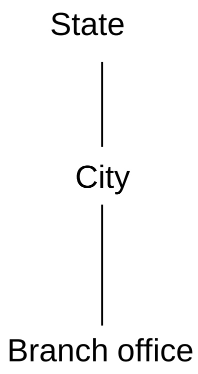
- Main feature: **Multi-dimensional** view of data with flexible, interactive aggregation and refinement functions along one or more dimensions
- Example: *Sales figures*:
  - by **product:** *product, product category, industry*
  - by **region:** *branch, city, state*
  - by **time:** *day, week, month, year*
  - …
  - According to **any combination of dimensions**, *e.g.* by *product category, city* and *month*
- Key figures: Facts for analysis
  - Data with common properties are aggregated



Attributes

- Time is a special dimension that usually exists in the OLAP system
  - Dimension **Time** has a linear character (Jan < Feb) and is cyclical
- Each dimension **is characterized** by a **set of attributes**
  - **Example:** The dimension *Region* is characterized by the attributes: *Branch, city* and *state*
- These attributes can be ordered **hierarchically**

  (aggregation levels)
  - **Example:**
    - Total value is derived from the values of several *states*
    - Value for one *state* is derived from the values of several *cities*
    - Value for one city is derived from values of several *branches*



OLAP Operations

- **Drill-down** or **roll-up operations**:

  Visualization of different aggregation levels

- **Slice** & **Dice operations:**

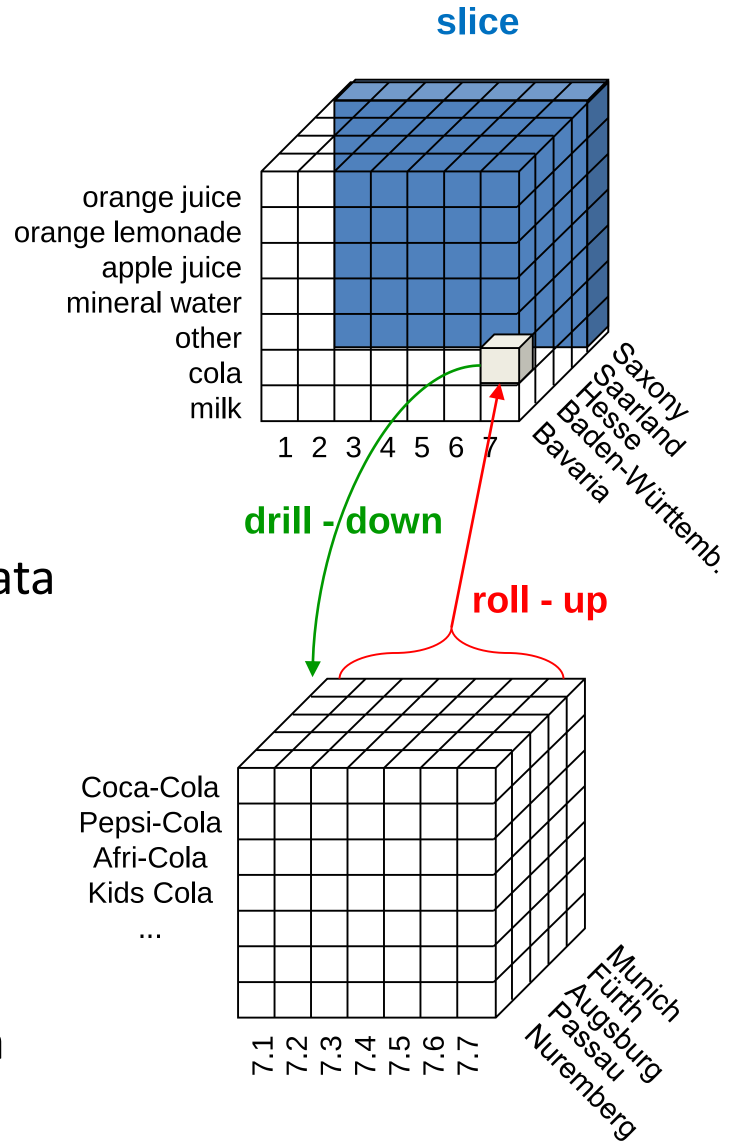
  Set conditions for displayed data

  ⇒ Reduce the dimensionality of the visualized data

- Analysis is supported by a variety of **visualization techniques**.

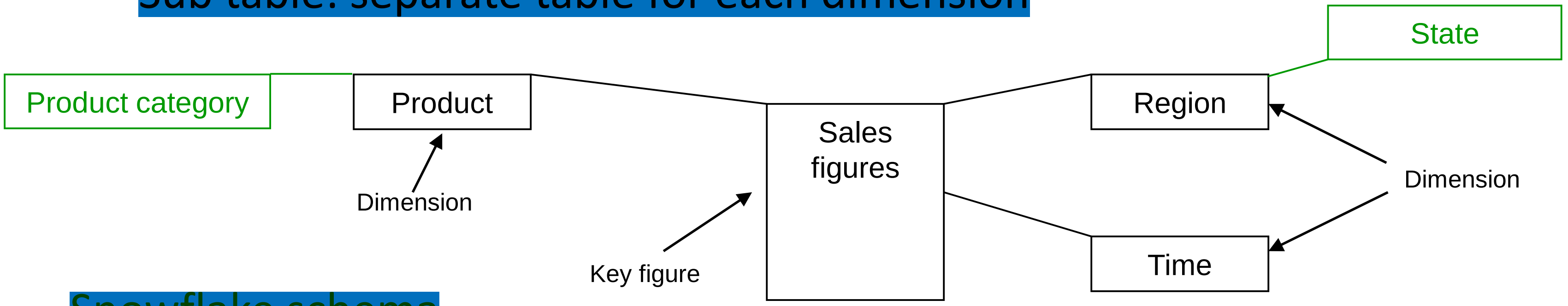
  Conditions are selected **interactively** (buttons, menus, *drag & drop),*

  ⇒ Analysts and managers do not have to learn a complicated query language



Multidimensional Data Model

- How to get to an OLAP-enabled data warehouse:
  1. Creation of a multidimensional conceptual data model
  2. Derivation of a relational logical data model
- Known multidimensional data models:
  – Star schema
    - Key figures table: objects of analysis
    - Sub table: separate table for each dimension



  – Snowflake schema

1.5 Preprocessing

The content of this chapter is partly based on:
S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Learning", *International Journal of Computer Science*, 2006, Vol 1 N. 2, pp 111-117.

Data Preprocessing

- Selecting the data to be used
  - Creation of an authoritative data table
  - Reduction of the amount of data, e.g. through sampling
- Data cleaning
  - Data consistency
  - Removing incorrect values / instances
  - Missing values
  - Duplicates (redundant features)
- Adapting the data to the data science methods
  - Discretization
  - Normalization
  - Feature Selection
  - Feature Extraction

Current research results on automation: AIDA
https://www.turing.ac.uk/research/research-projects/artificial-intelligence-data-analytics-aida

Data Preprocessing - Example

ID	Name	Colour	Quality control necessary?	Production Time [sec]	Production Frequency [1/h]
I1	Product1	Red	No	10	360
I2	Product2	Green	Yes	120	30
I3	Product3	Green	Yes	30	120
I4	Product4	Blue	No	90	40
I5	Product5	Red	Yes	60	60
…	…	…	…	…	…

Data Preprocessing - Proportionalization

- Data Science methods typically use single tables with
  - Rows: Cases (Instances)
  - Columns: Properties (Features)

- To achieve a single table, we have to perform **Proportionalization**
  - Transformation of a relational database into a propositional dataset (single table)
  - Features are usually aggregated (average, min, max, existance, etc.)
  - The proportionalization is typically performed by the user (Domain knowledge necessary)

Data Preprocessing - Proportionalization

ID	Name	Colour	Quality control necessary?	Production Time [sec]	Production Frequency [1/h]
I1	Product1	Red	No	10	360
I2	Product2	Green	Yes	120	30
…	…	…	…	…	…

ID	Product ID	Raw Material	Price [€/unit]
R1	I1	Material1	0.05
R2	I1	Material2	1
R3	I2	Material2	10
R4	I2	Material3	5
R5	I2	Material4	25
…		…	.

ID	Name	Colour	Quality control necessary?	Production Time [sec]	Production Frequency [1/h]	Number of raw materials	Raw Material Cost [€]
I1	Product1	Red	No	10	360	2	1.05
I2	Product2	Green	Yes	120	30	3	40
…	…	…	…	…	…	…	…

Data Preprocessing – What is wrong?

ID	Name	Colour	Quality control necessary?	Production Time [sec]	Production Frequency [1/h]
I1	Product1	Red	No	10	360
I2	Product2	G	Yes	120	120
I3	Product3	Green	Yes	30	120
I3	Product3	Green	Yes	30	120
I4	Product4	Blue		90	40
I5	Product5	Red	Yes	-10	-360
…	…	…	…	…	…

Data Preprocessing – What is wrong?

ID	Name	Colour	Quality control necessary?	Production Time	Production Frequency
I1	Product1	Red	No	10	360
I2	Product2	G	Yes	120	120
I3	Product3	reen	Yes	30	120
I3	Prod	en	Yes	3	120
I	Prod	e		90	40
	Product5	Red	Yes	-10	-360
...	...	...	...	...	...

Inconsistency *

Incorrectly written value

Out of range

Duplicate

Missing Value

* If 120 sec. are necessary to produce the product, only 30 products can be produced per hour

Data Preprocessing – Erroneous Values

- Typical errors:
  - Missing values
  - Duplicates
  - Values are outside of a specified range
  - Incorrectly written feature values (especially for strings)
  - Inconsistency (Values are mathematically, physically, etc. impossible)
  - Redundancy (Features can be constructed/calculated by other features)

- Possible Solutions
  - Removing
    - How much information is removed?
    - Do we remove the samples or the feature?
  - Correcting
    - Is it possible to correct the erroneous values?
    - Which values do we insert?

Data Preprocessing – Erroneous Values

- Erroneous/Missing values may be corrected by
  - Inserting a default value
  - Inserting the most common value (for categorical data)
  - Inserting the average value (for continuous data)
  - Inserting the prediction of an already fitted model
  - Using the error value as is

    (Can conclusions be drawn from the absence of the value?)

Data Preprocessing - Data Consistency

- Syntactic errors in input files
  For example:
  - Values containing commas in a comma-separated file format
  - German vs. English decimal separator (comma vs. dot), ...
- Consistent unit for a concept (gram vs. kilogram vs. ton)
- Same concepts that were recorded with different names
- Different concepts, which were recorded with the same name

Data Preprocessing - Outlier Detection

- Outlier = instance that is "far away" from other instances (regarding one or more features)

- An error or important information?
  - If the outlier is actually erroneous, the instance must be removed

- Possible method for outlier detection:
  - Apply clustering methods
  - Find instances that are difficult to sort into clusters
  - Apply anomaly detection methods

Data Preprocessing - Discretization

- Some data mining methods require ordinal values, but data is often numerical
- Discretization describes the conversion of numerical feature into ordinal
- Interval in old feature corresponds to one value in new feature

ID	Raw Material Cost [€/unit]
I1	1.05
I2	42.0
I3	0.20
I4	60.0
I5	25.0
…	…

>50: High
10-50: Medium
<10: Low

ID	Raw Material Cost [€/unit]
I1	Low
I2	Medium
I3	Low
I4	High
I5	Medium
…	…

- The intervals can be selected either by hand (domain knowledge) or automatically (see next slides)

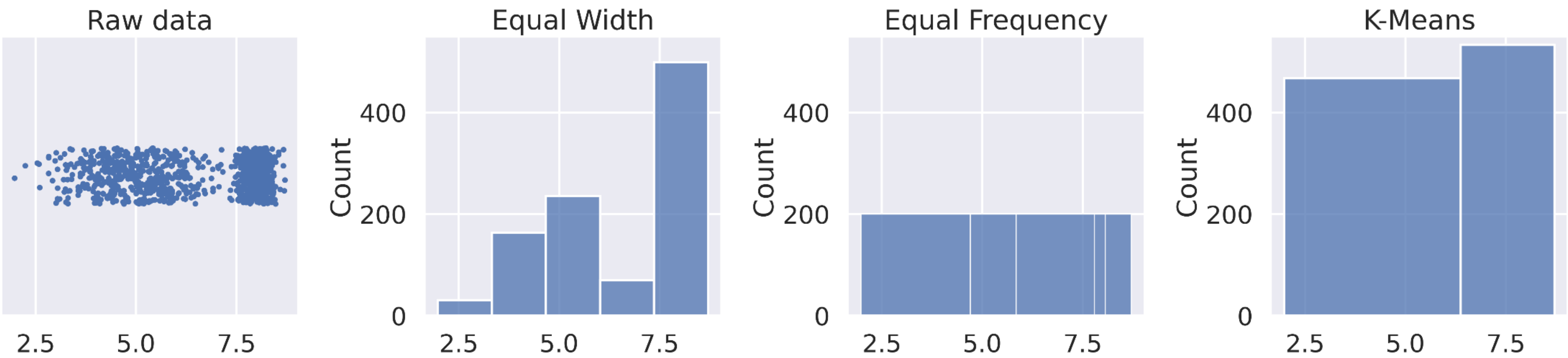
Data Preprocessing - Automatic Discretization

- Equal-Width Discretization
  All intervals are the same size

- Equal-Frequency Discretization
  All intervals contain the same number of instances

- Discretization by Clustering
  The intervals are determine by a clustering method (more in chapter 2!)

Data Preprocessing – Automatic Discretization



- Equal Width
  - Simple method but generates imbalanced bins
- Equal Frequency
  - Ensures equal number of samples per bin but bin edges are not well interpretable
- Clustering
  - Bins are generated by a clustering method which reflects the structure of the data

Data Preprocessing - Automatic Discretization

Alternative: Supervised Discretization

Include another (e.g. binary) feature in the discretization!

If possible (in the binary case) an interval should contain "only positive" or "only negative" examples

- Top-down:
  - Start with an interval
  - Successively divide into parts that belong to (if possible) the same class
- Bottom-up:
  - Start with each value as a single interval
  - Combine intervals with similar class distribution
- Dimensions for "uniform class", e.g..:
  - Entropy (see chapter 5.4)
  - Statistical significance test (Chi² Test)

Data Preprocessing - Encoding

- Example: Feature “colour” with values {red, green, blue}
- **Don't:** Introduction of "unnatural" structures (e.g. an order)

~~Color~~
~~Red~~
~~Green~~
~~Blue~~

$\Rightarrow$

~~Color~~
~~1~~
~~2~~
~~3~~

(Red < Green) ???

- **Do:** One boolean feature for each colour

Color
Red
Green
Blue

$\Rightarrow$

Red	Green	Blue
1	0	0
0	1	0
0	0	1

One-Hot Encoding

Data Preprocessing - Feature Scaling

- Some (numerical) features have small value ranges (e.g.: 0.0 - 0.01), others large value ranges (e.g.: 0 - 100,000)
- The features should be normalized by a suitable method
  - Rescaling (Min-Max Normalization):

$$f(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - Standardization (Z-score Normalization):

$$f(x) = \frac{x - \mu}{\sigma}$$

(With μ and σ being the mean and standard deviation of x for the whole dataset)

Data Preprocessing - Instance Selection

- Some methods require the selection of random subsets

- **Random sampling** (normal case)
  - (Random) selection of a subset of the data

- **Stratified** sampling
  - Increase the proportion of instances in the sample for the rare class compared to a random sample (especially for "imbalanced data")
  - Correlations between features should also be found in the random subsets

Data Preprocessing - Feature Selection

- Are features …
  - … relevant? (the feature is related to the quantity of interest)
  - … irrelevant? (the feature is **not** related to the quantity of interest)
  - … redundant? (the feature can be replaced/constructed by other features)

- You may filter features that …
  - … are anachronisms (features are unknown at the time of prediction)
  - … are monotonically increasing (time, ID, …)
  - … only have few non-default values
  - … have many different values (e.g. number of samples = number of values)

- There exist a vast amount of heuristic methods for finding the "optimal" subsets of features (2^n possible subsets exist!), e.g.
  - Filtering (remove features with low scores according to a suitable measure)
  - Sequential Forward/Backward Selection
  - Embedded Feature Selection (e.g. L1 Regularization)
  - Genetic algorithms

Data Preprocessing - Feature Extraction

- Creating new features from given features
  e.g. transform the postal code to geographical coordinates (longitude & latitude)

- Combining features by any mathematical mapping, e.g.
  $$x_{new} = 5 \cdot x_0^2 + e^{x_1} - 2 \cdot \sin x_2$$

- Feature Extraction often uses background knowledge
  $\Rightarrow$ Linking with further datasets ("cross-domain mining")