Introduction and Organization

IE500 Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Hallo

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Large-Scale Data Integration
  - Data and Web Mining
  - Large Language Models
- Room: B6, 26 - B1.15
- eMail: christian.bizer@uni-mannheim.de

- Will teach the lecture (IE670)

Hallo

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Dr. Ralph Peeters**
- Postdoc Researcher
- Research Interests:
  - Entity Matching using Deep Learning
  - Data Integration using LLMs
  - Web Agents
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de

- Will teach the exercises and will supervise student projects

Hello

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **M.Sc. Aaron Steiner**
- Graduate Research Associate
- Research Interests
  - LLM Agents in Data Integration Pipelines
  - Agent Interfaces to the Web
  - RAG Agents
- Office: B6, 26 - C 1.04
- eMail: aaron.steiner@uni-mannheim.de

- Will teach the exercises and will supervise student projects

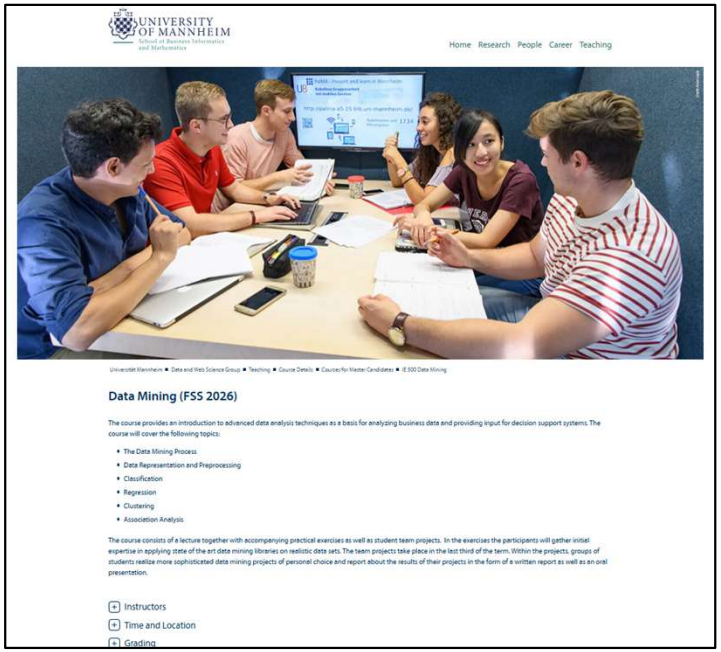
Outline

UNIVERSITY OF MANNHEIM
Data and Web Science Group

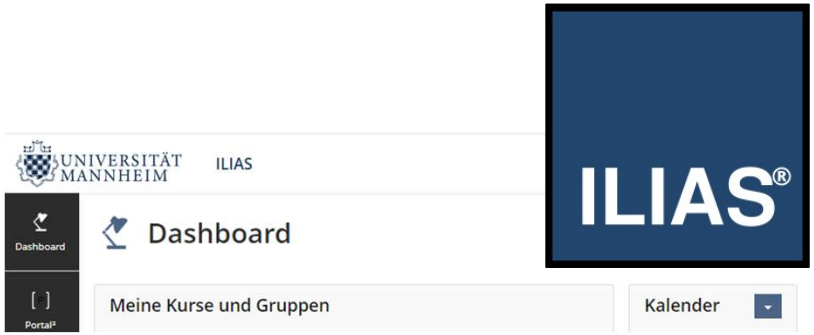
1. Course Organization
2. What is Data Mining?
3. Tasks and Applications
4. The Data Mining Process
5. Classification: K-Nearest-Neighbors

1. Course Organization

- Course Webpage
  - https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-500-data-mining
  - Provides up-to-date information, lecture slides, exercise material



- ILIAS eLearning System
  - https://ilias.uni-mannheim.de/
  - Mailing lists, discussion forum,
  - Team project (submission, coaching sessions)



Course Organization

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Registration
  - you have registered via Portal2
  - and been added to ILIAS

- Offline Lecture
  - Introduces the principal methods of data mining
  - Discusses how to evaluate the learned models
  - Presents practical examples of data mining applications

  - Time: Wednesday, 10:15 – 11:45,
  - Location: Room B243 Building A 5,6

Online Lecture

- Part of the course
- Exam relevant
- Slides and videos are available via ILIAS

Week	Wednesday (Offline Lecture, Room A5, B243)	Online Lecture (see Ilias Course)	Thursday (Exercise, Room B6, A104)
11.02.2026	Introduction to Data Mining		Introduction to Python
18.02.2026	Classification 1	Nearest Centroids	Classification 1
25.02.2026	Classification 2	Comparing Classifiers	Classification 2
04.03.2026	Regression	Ensembles	Regression
11.03.2026	Preprocessing	Multi Modal Data	Preprocessing
18.03.2026	Clustering and Anomalies	Hierarchical Clustering	Clustering
25.03.2026	Intro to Student Project	Time Series	Time Series
	- Easter Break -		
15.04.2026	Feedback about Project Proposals		Project Work
22.04.2026	Association Analysis and Subgroup Discovery		Association Analysis
29.04.2026	Project feedback session		Project Work
06.05.2026	Project feedback session		Project Work
13.05.2026	Project feedback session		Project Work
20.05.2026	Project Presentations		Project Presentations

Exercise

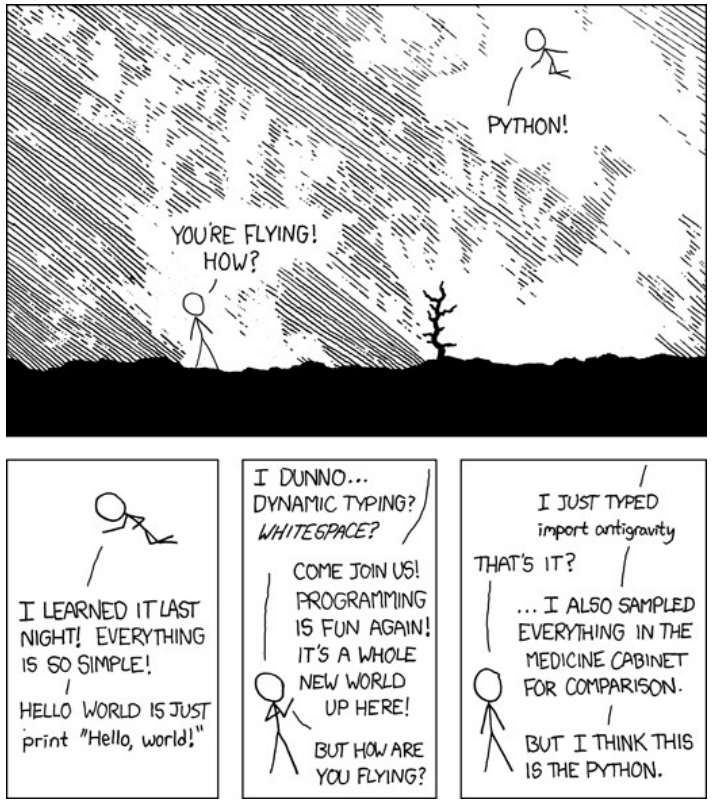
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Exercise Groups
  - Students experiment with data sets using Python
  - Students solve theoretical tasks (similar to exam)
- Time and Location (same content - only attend **one**):
  - Thursday, 10.15 – 11.45, A104 Building B6, 26 Part A (Aaron)
  - Thursday, 12.00 – 13.30, A104 Building B6, 26 Part A (Aaron)
  - Thursday, 13.45 – 15.15, A104 Building B6, 26 Part A (Ralph)
- You can also switch between groups as needed

Introduction to Python (Optional)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- This Thursday (12. February **13.45 – 15.15**, A104)
- Topics:
  - Setup of environment (Anaconda, Jupyter Notebooks)
  - Python Intro / Design Goals
  - Basic programming concepts in Python
- Support
  - Help with environment setup
  - Q&A
- Material
  - Tutorial slides available on website



Usage of LLMs like ChatGPT

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We will be using LLMs in the exercise to
  - discuss suitable methods and parameter settings for different use cases
  - generate and debug Python code for experimenting with the methods

Source: New York Times

OpenAI

Project

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Project Work
  - Teams of **five to six** students realize a data mining project
  - Teams may choose their own data sets and tasks
    (in addition, we will propose some suitable data sets and tasks)
  - Write summary about project and present the results
- Deadlines
  - Team formation
    - **Sunday, March 22nd, 23:59**
  - Submission of project proposal
    - **Wednesday, April 8th, 23:59**
  - Submission of final project work report
    - **Friday, May 15th, 23:59**
  - Submission of Presentation (PDF)
    - **Thursday, May 21st, 23:59**

Exam

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Date and Time: **will be announced**
- Duration: 60 minutes
- Structure: 6 open questions that
  - Check whether you have understood the lecture content
    - we try to cover all major chapters of the lecture
    - require you to describe the ideas behind algorithms and methods
    - often: How do methods react to special patterns in the data?
  - Might require you to do some simple calculations for which
    - You need to know the most relevant formulas
    - You do not need a calculator
  - We will provide more information about the exam in April

Exam

UNIVERSITY OF MANNHEIM
Data and Web Science Group

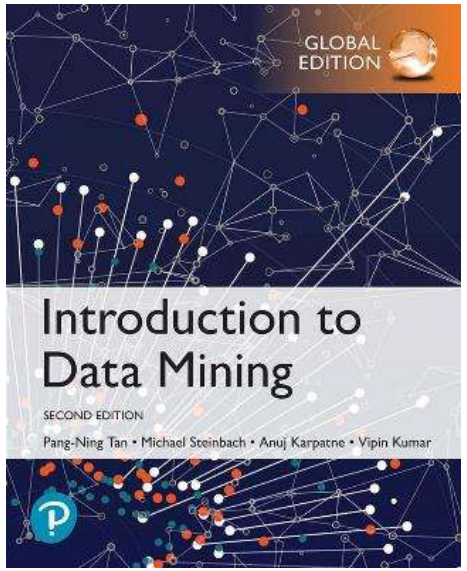
- **There is only one exam per semester**
  - Because course is offered every semester
    - The next exam date is at the end of the upcoming HWS
  - i.e., no retake date!
- Upon failure, you will have to **redo both the project and the exam** in another semester
  - Unfortunately, we cannot carry over your project mark

- **Final grade**
  - 75 % written exam, 20% project report, 5% project presentation

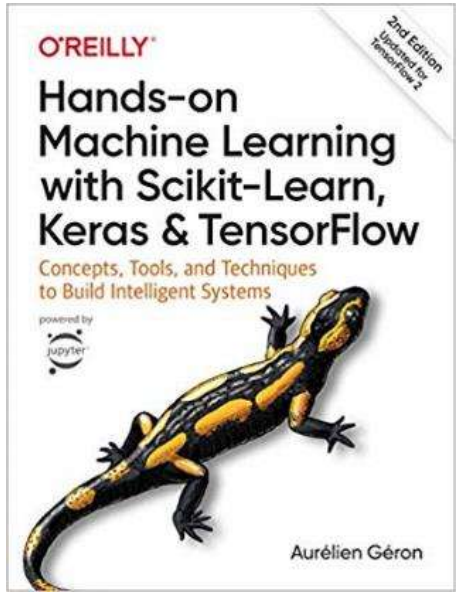
Textbooks for the Course

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar:
  Introduction to Data Mining. 2nd Edition.
  Pearson / Addison Wesley.



- Aurélien Géron:
  Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.
  2nd or 3rd Edition, O'Reilly, 2019 or 2022

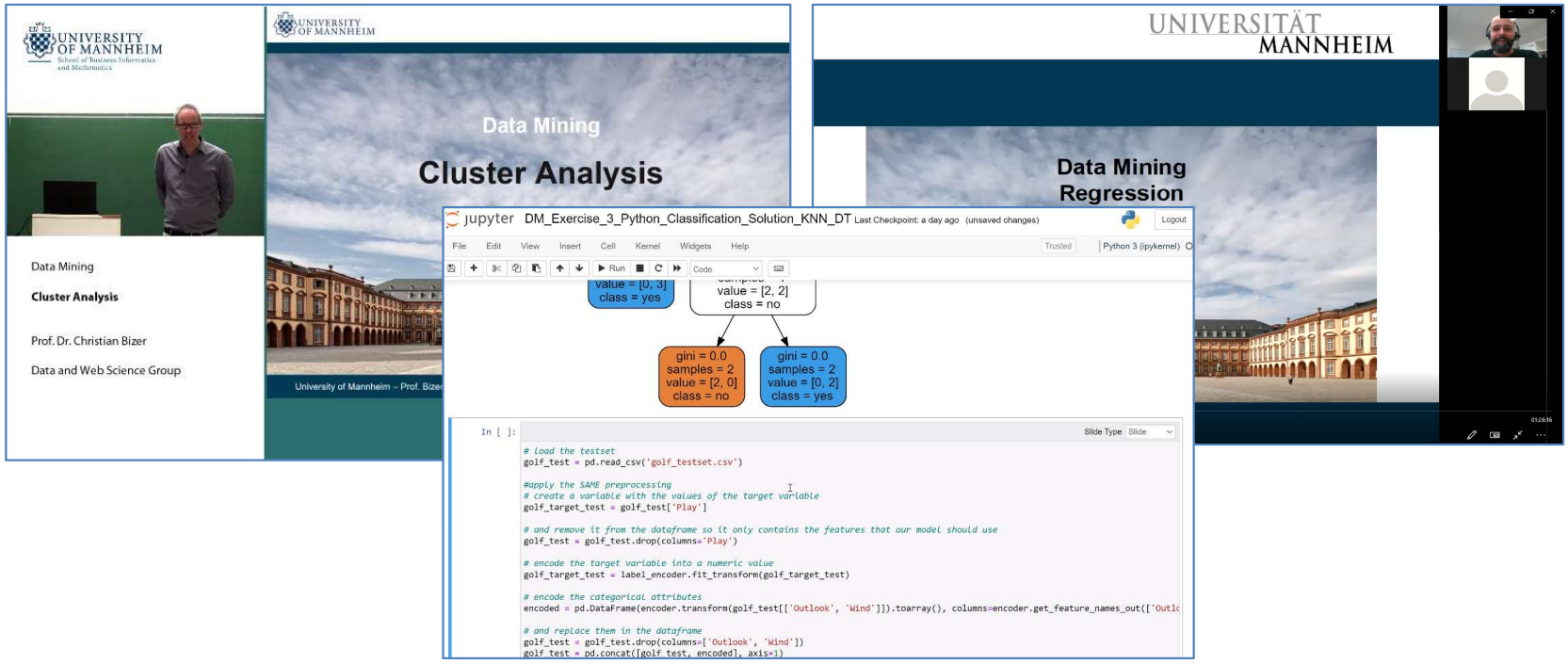


- **Scikit-learn Documentation:**
  https://scikit-learn.org/stable/user_guide.html

Videos and Screencasts

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Videos**
  - https://www.uni-mannheim.de/dws/teaching/lecture-videos (VPN!)
  - **Lecture Videos** By Heiko Paulheim (HWS 2020) and Christian Bizer (FSS 2020)
  - **Screencasts** for the Exercises by Ralph Peeters (FSS 2022)
- Keep in mind, that the lecture and exercise change over time



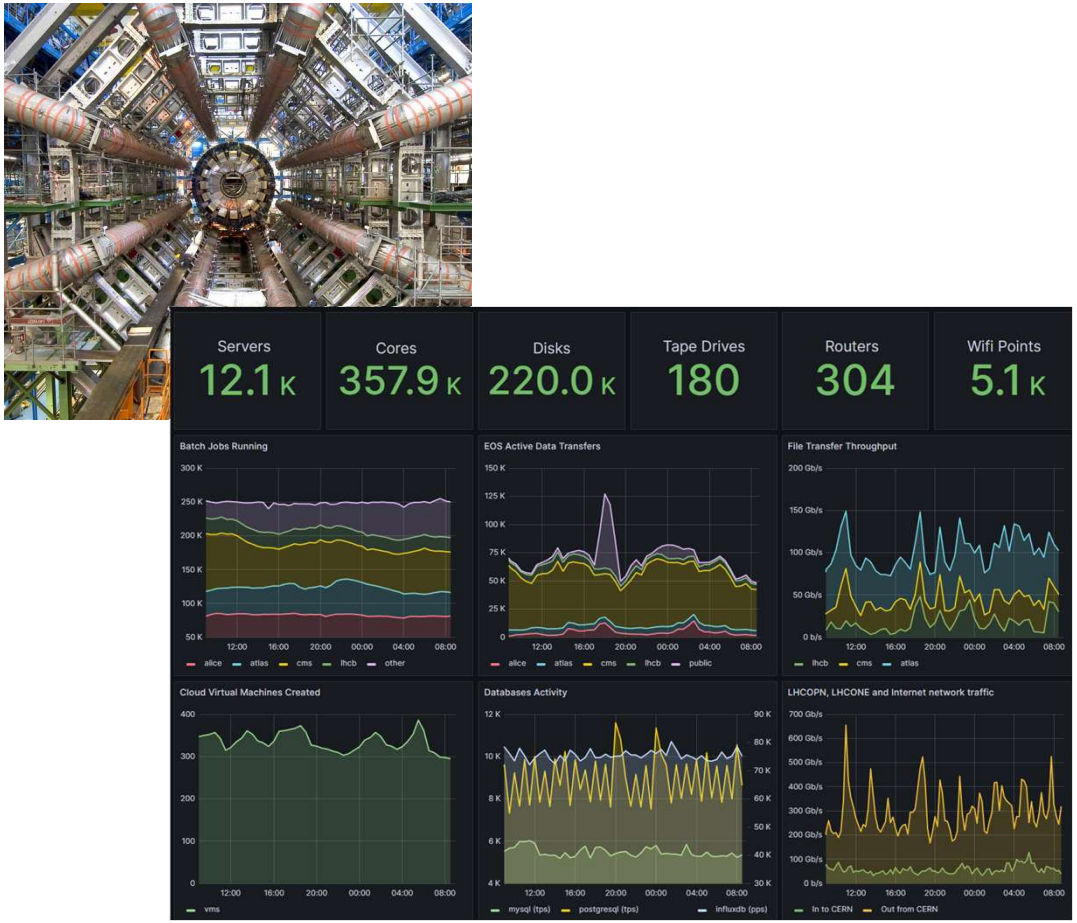
Questions?

2. What is Data Mining?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Large quantities** of data are collected about all aspects of our lives
- This data contains **interesting patterns**
- Data Mining helps us to
  1. **Discover these patterns** and
  2. **Use them for decision making** across all areas of society, including
     - Business and industry
     - Science and engineering
     - Medicine and biotech
     - Government
     - Individuals

“We are Drowning in Data...”



- **CERN**
  - Large Hadron Collider
    - 45 petabytes per week produced (February 2024)
  - 820 petabytes of data archived on tape
  - 1005 petabytes of disk space available (August 2024)
- Discover
  - Patterns in the experiments

https://home.cern/news/news/computing/new-data-centre-cern
http://cern.ch/go/datacentrebynumbers

“We are Drowning in Data...”

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Facebook**
  - 4 Petabyte of new data generated every day
  - over 300 Petabyte in Facebook‘s data warehouse
- Predict
  - Interests and behavior of over one billion people

https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

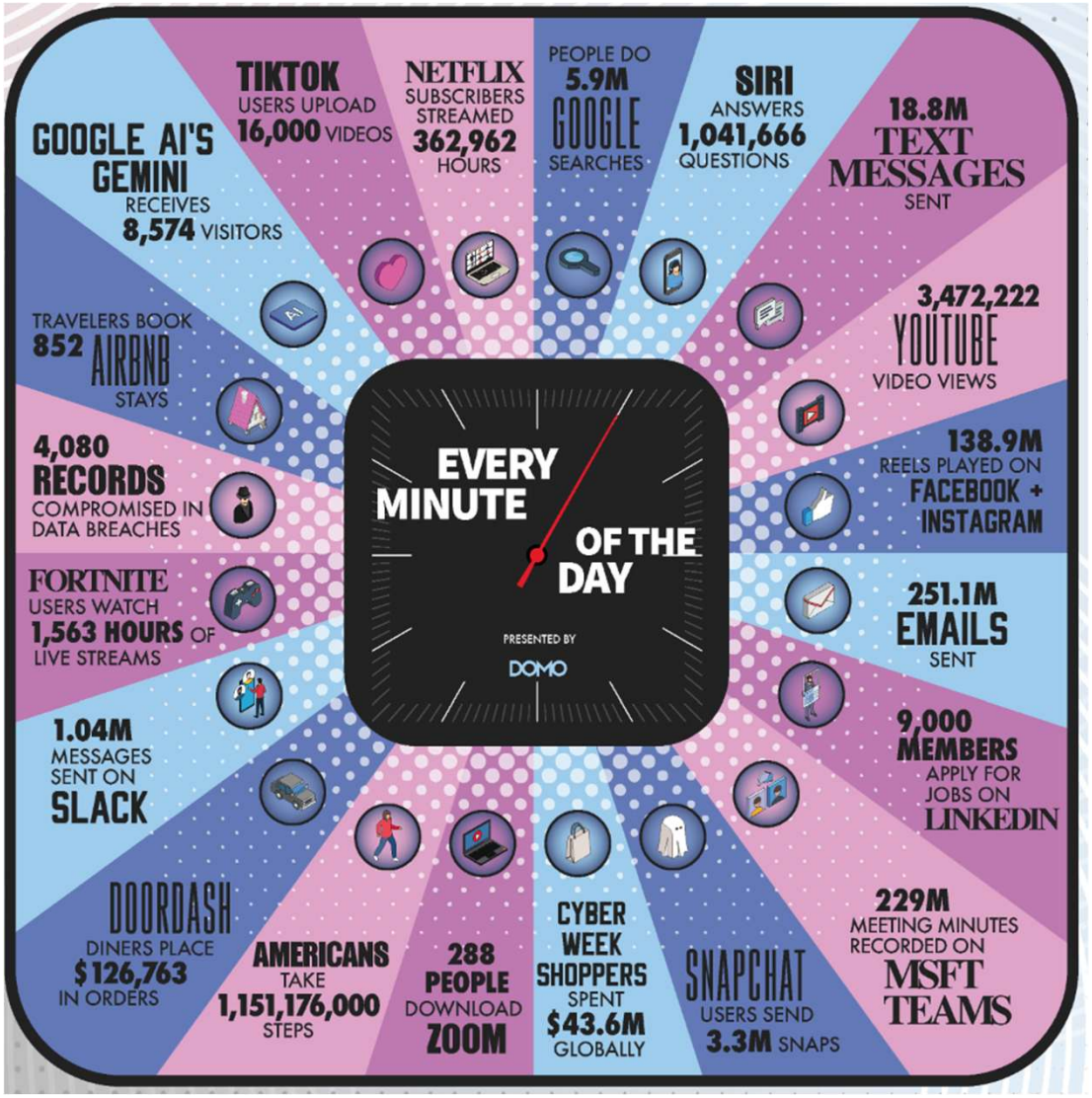
“We are Drowning in Data...”

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **US Library of Congress**
  - ≈ 235 TB archived
- Discover
  - Topic distributions*
  - Citation networks
- Train
  - Large Language Models

https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

“We are Drowning in Data...”



- Predict
  - Interests and behavior of mankind
- Value Source
  - Company value and market position often depends on collected data

https://www.domo.com/learn/infographic/data-never-sleeps-12

“We are Drowning in Data...”

UNIVERSITY OF MANNHEIM
Data and Web Science Group

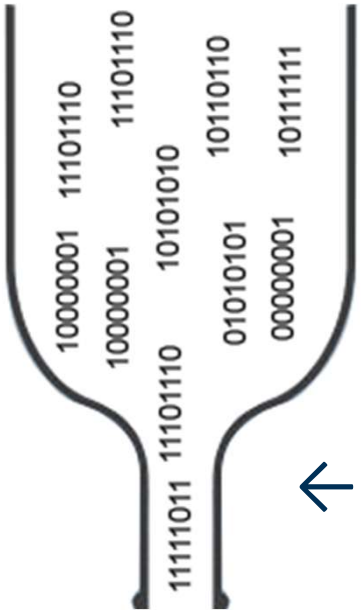
Law enforcement agencies collect unknown amounts of data from various sources

- Web browsing behavior
- Cell phone calls
- Location data
- Credit card transactions
- Chat bot histories
- Online profiles (Facebook)

Predict

- Predict terrorist or not?
- Predict trustworthiness

“We are Drowning in Data... but starving for knowledge!”



← Rate at which data are produced

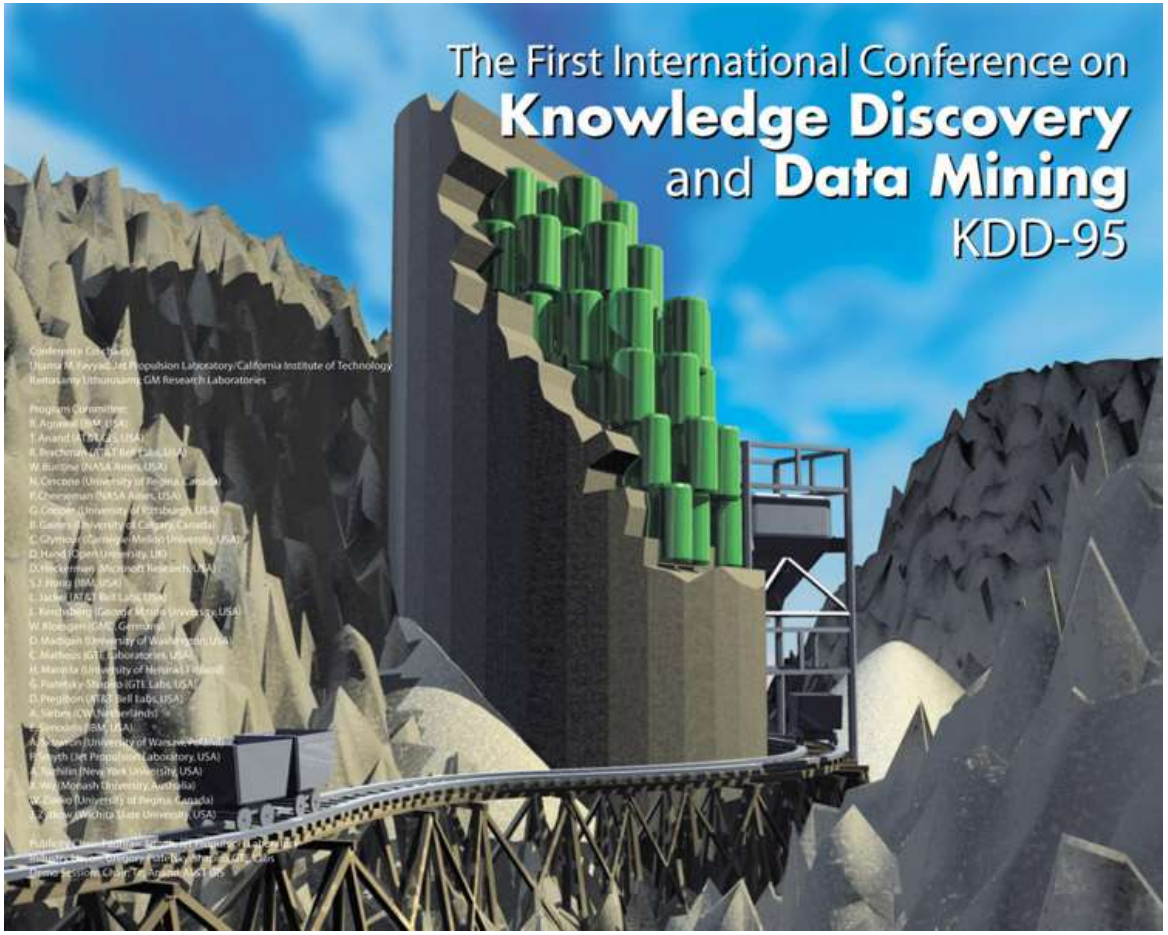
← Amount of data that can be inspected by humans.
Manual interpretation is hardly feasible!

- We are interested in **the patterns, not the data** itself!
- Data Mining methods help us to
  - **Discover interesting patterns** in large quantities of data
  - **Take decisions** based on the patterns

Data Mining: Definitions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Metaphor: Mountains of data from which knowledge is mined



Data Mining: Definitions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Data Mining is a non-trivial process of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable

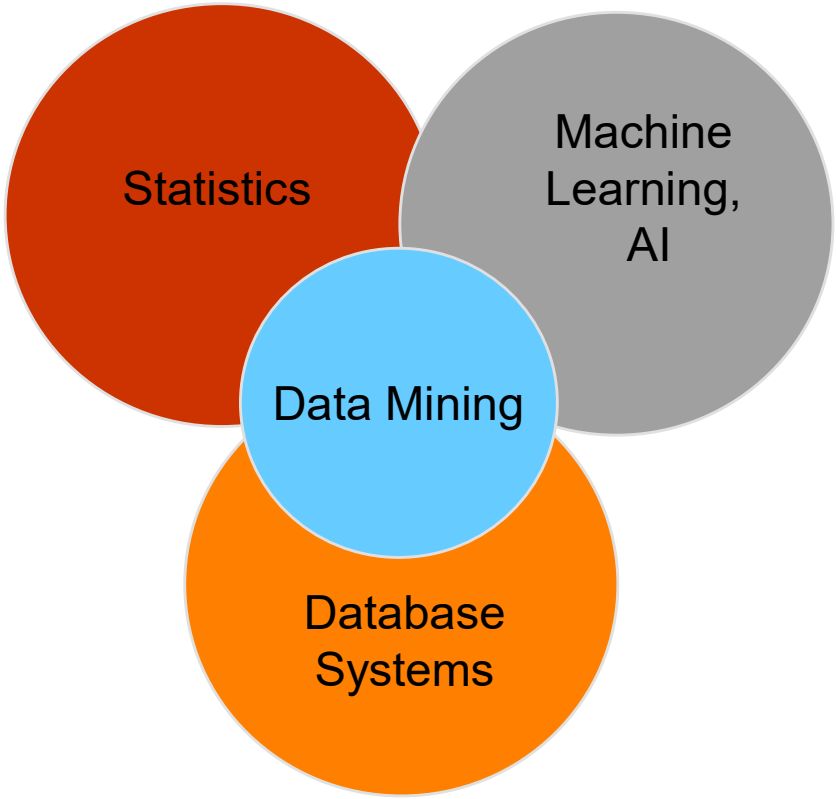
  patterns in data. (Fayyad et al. 1996)

- Data Mining methods
  1. Detect interesting patterns in large quantities of data
  2. Support human decision making by providing such patterns
  3. Predict the outcome of a future observation based on the patterns

Origins of Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Traditional techniques may be unsuitable due to
  - Large amount of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



3. Tasks and Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group

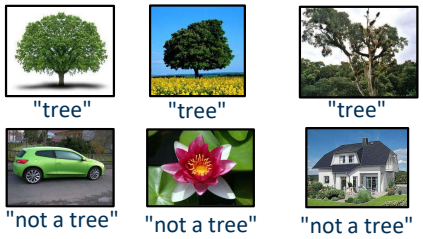
- **Descriptive** Tasks
  - Find patterns in the data
    - E.g. which products are often bought together?
- **Predictive** Tasks
  - Predict unknown values of a variable
    - Given observations (e.g., from the past)
    - E.g. will a person click a online advertisement?
      - given her browsing history

- Machine Learning Terminology
  - Descriptive = unsupervised
  - Predictive = supervised

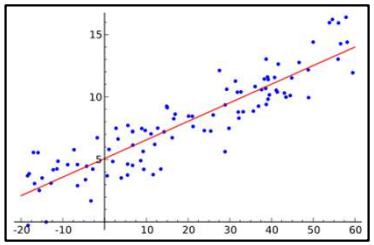
Data Mining Tasks

UNIVERSITY OF MANNHEIM
Data and Web Science Group

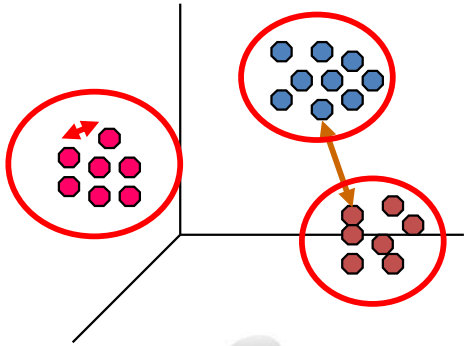
1. Classification [Predictive]



2. Regression [Predictive]



3. Cluster Analysis [Descriptive]



4. Association Analysis [Descriptive]

Classification

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Previously unseen records should be assigned a class from a given set of classes as accurately as possible.

- Approach:
  - Given a collection of records (**training set**)
    - Each record contains a set of **attributes**
    - One attribute is the **class attribute (label)** that should be predicted
  - Find a **model** for predicting the class attribute as a function of the values of other attributes

Classification

"tree"

"tree"

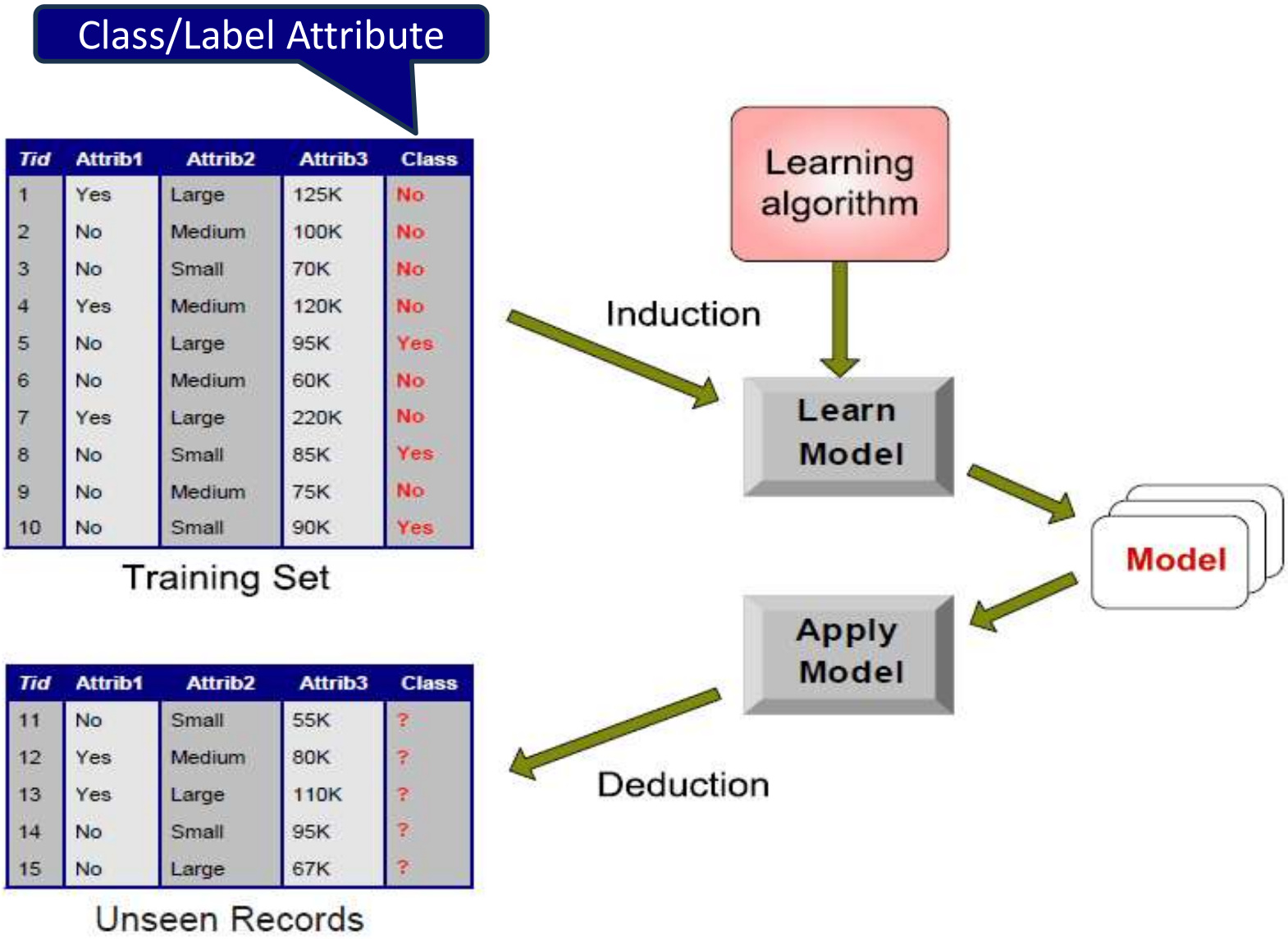
"tree"

"not a tree"

"not a tree"

"not a tree"

Classification: Workflow



Classification: Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Credit Risk Assessment
  - Attributes: your age, income, debts, …
  - Class: are you getting credit by your bank?

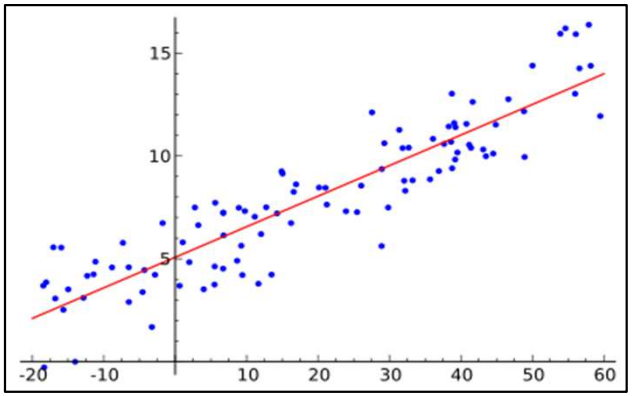
- SPAM Detection
  - Attributes: words and header fields of an e-mail
  - Class: regular e-mail or spam e-mail?

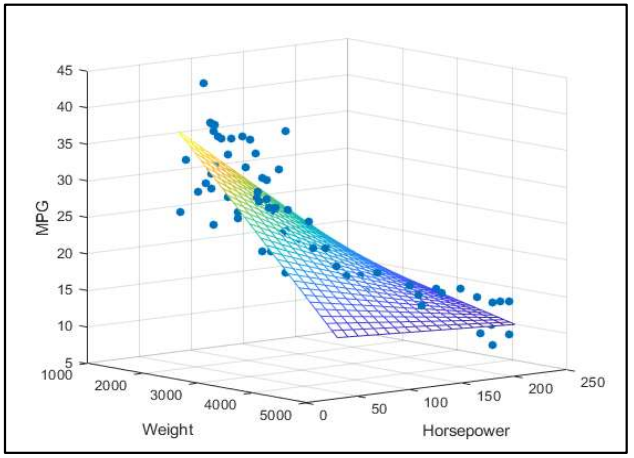
- Analysis of tax declaration?
  - Attributes: the values in your tax declaration
  - Class: are you trying to cheat?

Regression

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Predict a value of a **continuous variable** based on the values of other variables, assuming a linear or nonlinear model
  - Examples:
    - Predicting the price of a house or car
    - Predicting sales amounts of new product based on advertising expenditure
    - Predicting miles per gallon (MPG) of a car as a function of its weight and horsepower
    - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.



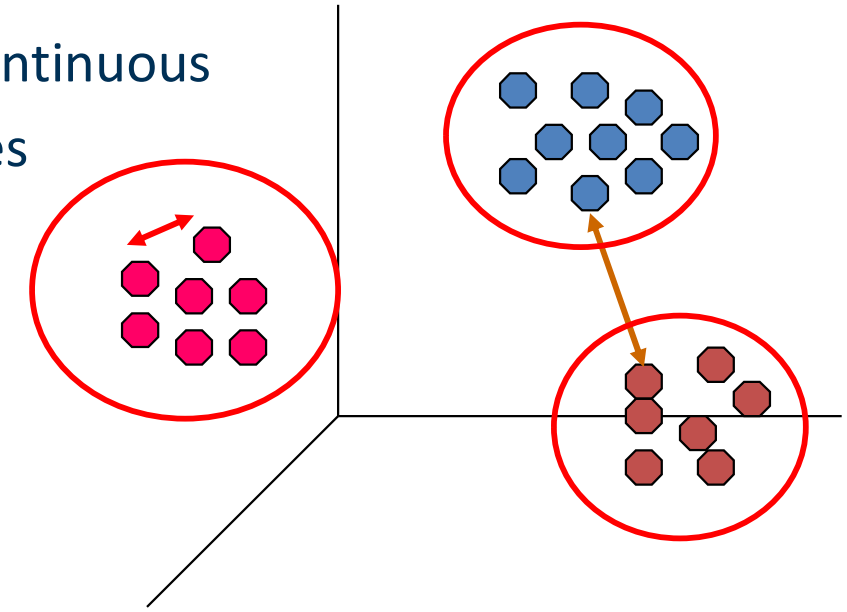


- Difference to classification: The predicted attribute is **continuous**, while classification is used to predict nominal attributes (e.g. yes/no)

Cluster Analysis

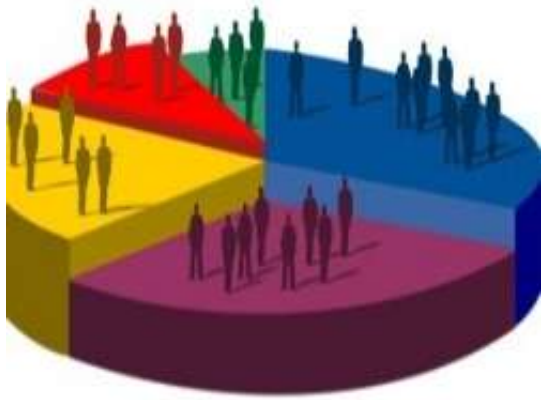
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find groups such that
  - Data points in one group are more similar to one another
  - Data points in separate groups are less similar to one another
- Similarity Measures
  - Euclidean distance if attributes are continuous
  - Other task-specific similarity measures
- Goals
  - Intra-cluster distances are minimized
  - Inter-cluster distances are maximized
- Result
  - A descriptive grouping of data points

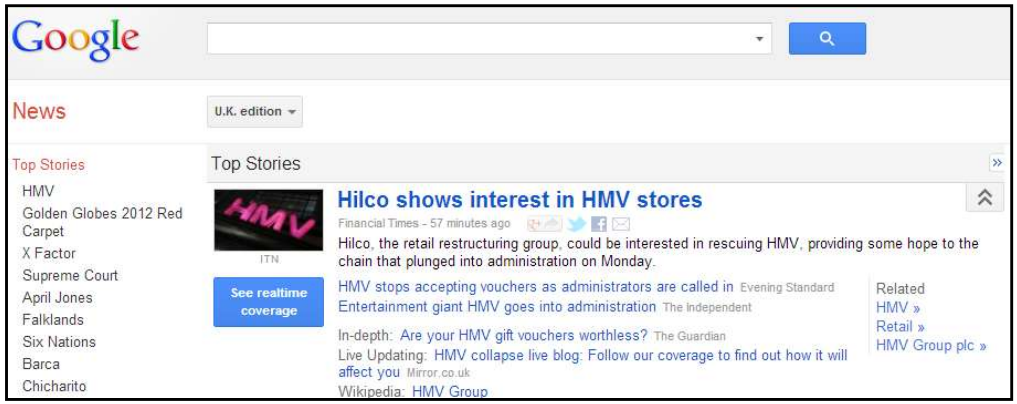


Cluster Analysis: Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group



- Application 1: Market segmentation
  - Find groups of similar customers
  - Where a group may be conceived as a marketing target to be reached with a distinct marketing mix
- Application 2: Document Clustering
  - Find groups of documents that are similar to each other based on terms appearing in them
    - Grouping of articles in Google News



Association Analysis

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Given a set of records each of which contain some number of items from a given collection
- Discover **frequent itemsets** and produce **association rules** which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Frequent Itemsets
{Diaper, Milk, Beer}
{Milk, Coke}

Association Rules
{Diaper, Milk} --> {Beer}
{Milk} --> {Coke}

Association Analysis: Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Supermarket shelf management
  - To identify items that are bought together by sufficiently many customers
  - Process the point-of-sale data collected with barcode scanners to find dependencies among items

- Sales Promotion

amazon.com

Frequently Bought Together

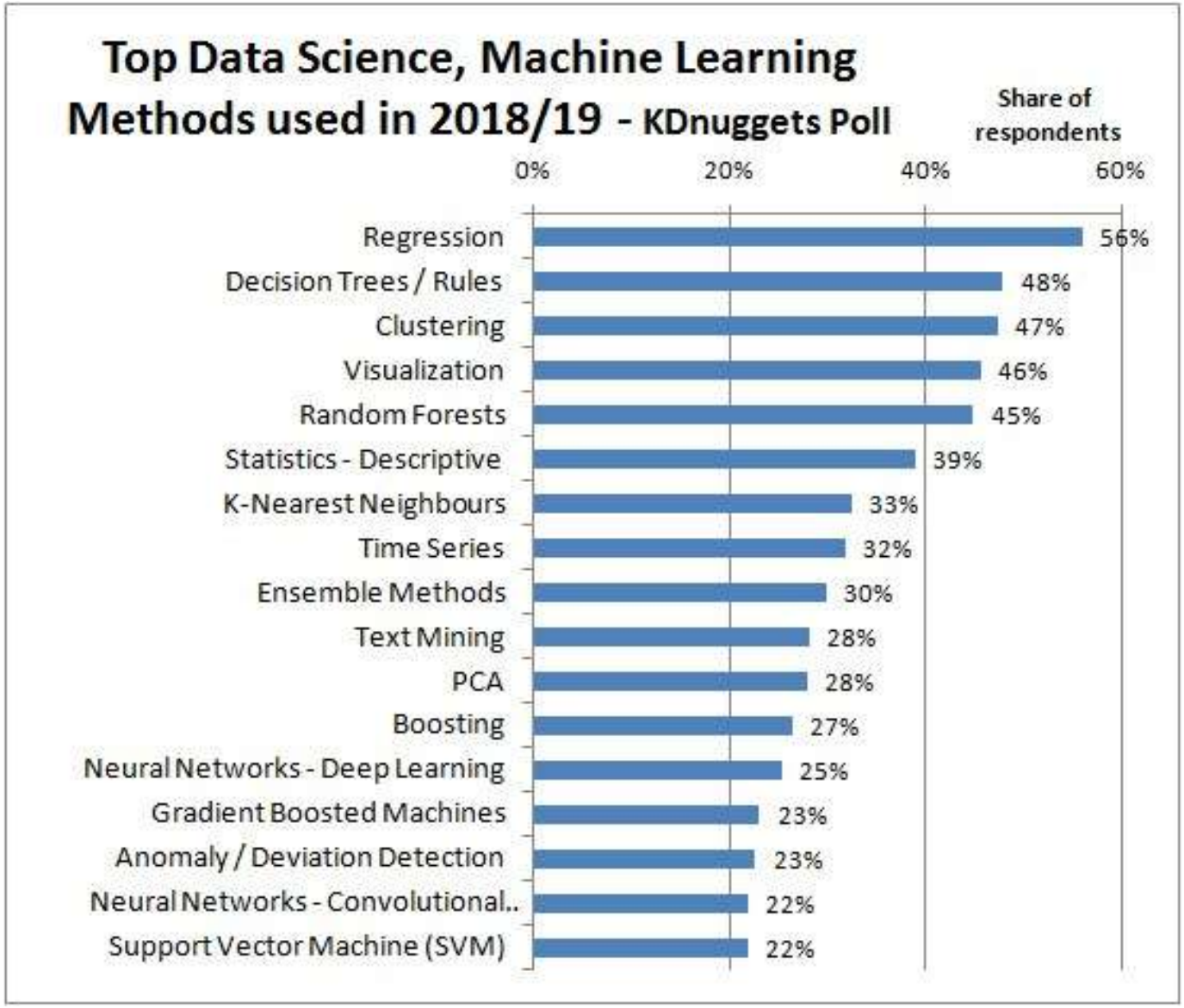
Price For All Three: $87.41

Add all three to Cart | Add all three to Wish List

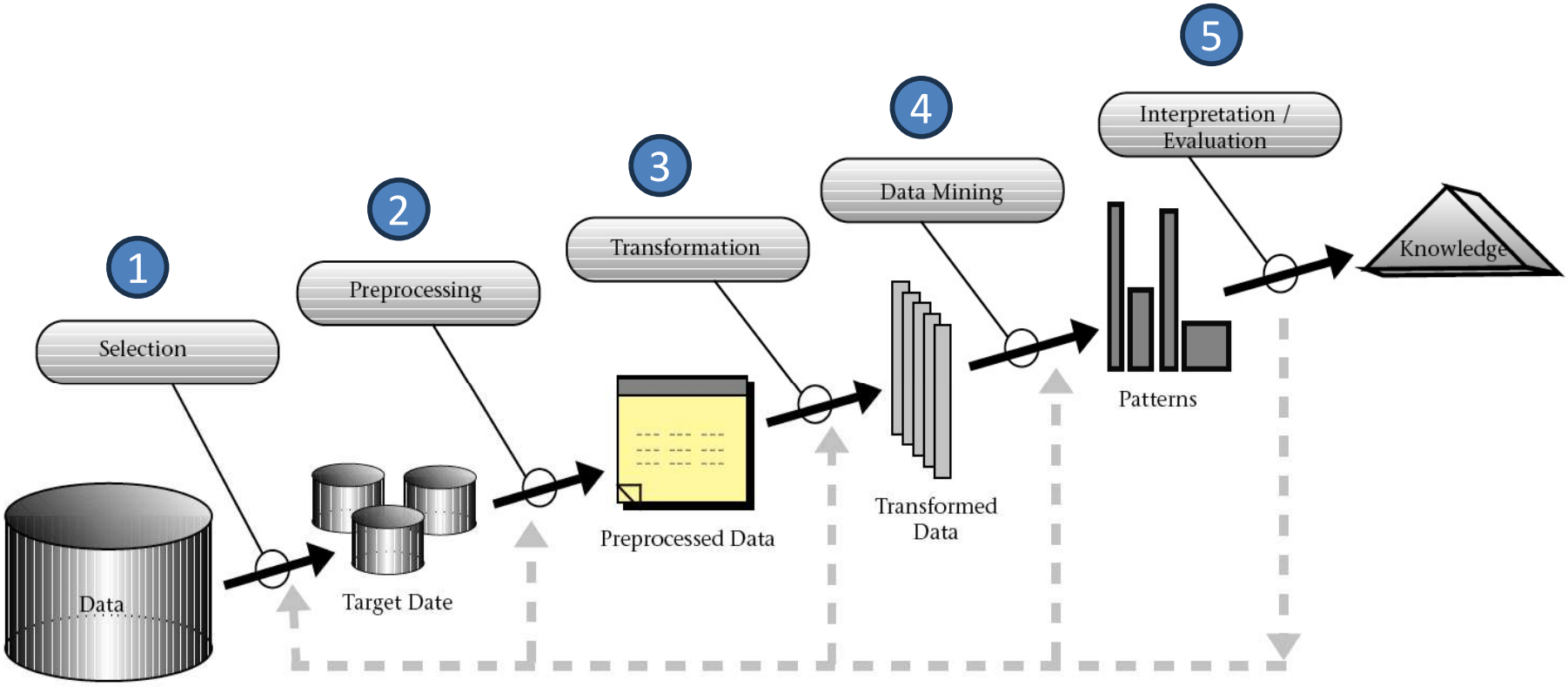
Show availability and shipping details

Which Methods are Used in Practice?

UNIVERSITY OF MANNHEIM
Data and Web Science Group



4. The Data Mining Process

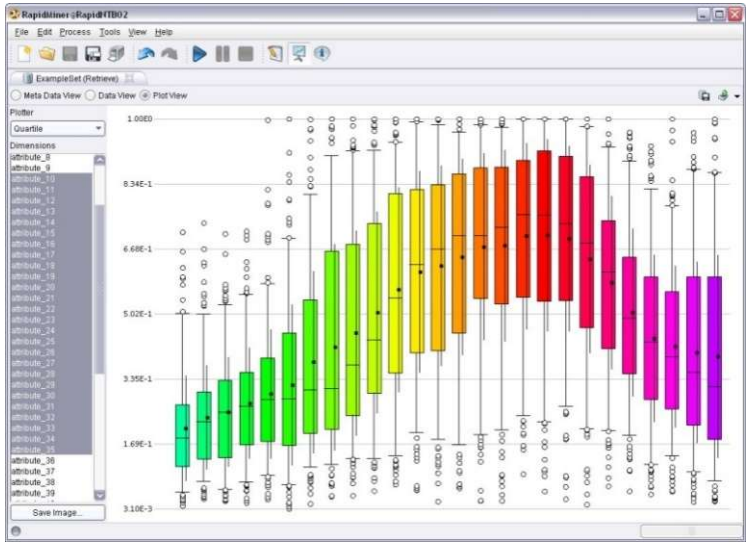


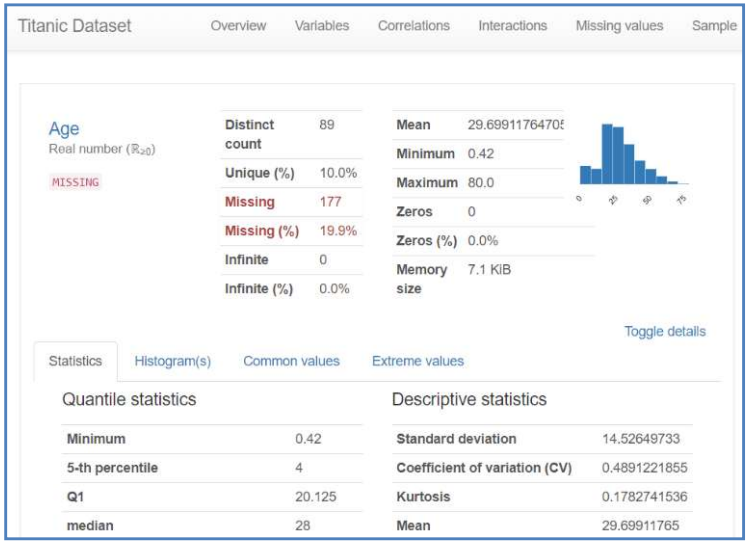
Source: Fayyad et al. (1996)

Selection and Exploration (1)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Selection
  - What data is available?
  - What data is potentially useful for the task at hand?
  - What do I know about the quality/provenance of the data?
- Exploration / Profiling
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records

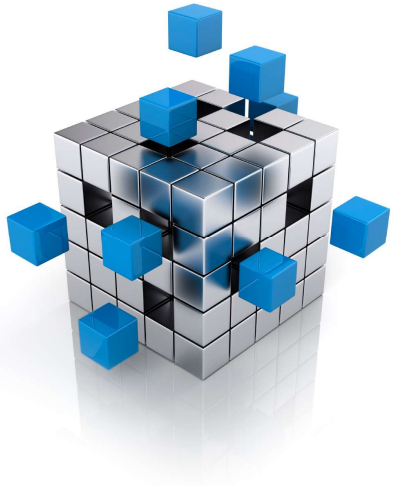




Preprocessing and Transformation (2+3)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Transform data into a representation that is suitable for the chosen data mining methods
  - Number of dimensions (represent relevant information using less attributes)
  - Scales of attributes (nominal, ordinal, numeric)
  - Amount of data (determines hardware requirements)

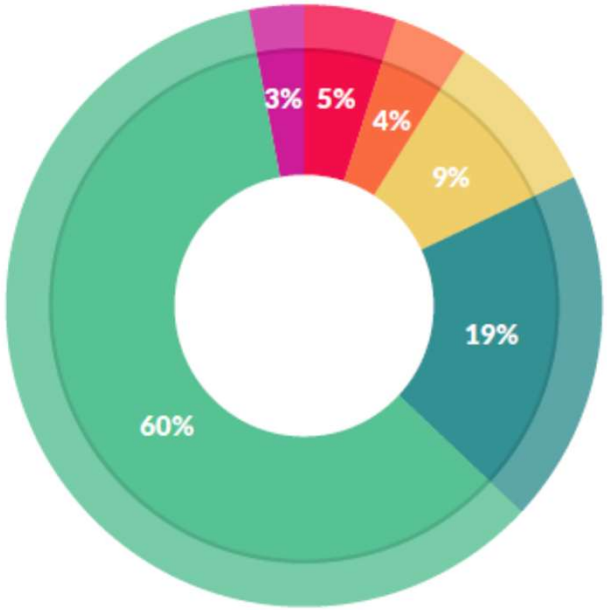


- Methods
  - Discretization and binarization
  - Feature subset selection / dimensionality reduction
  - Attribute transformation / text to term vector / embeddings
  - Aggregation, sampling
  - Integrate data from multiple sources

Preprocessing and Transformation (2+3)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Good data preparation is key to producing valid and reliable models
- Data integration/preparation is estimated to take **70-80%** of the time and effort of a data mining project



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Advertisement: IE670 Web Data Integration

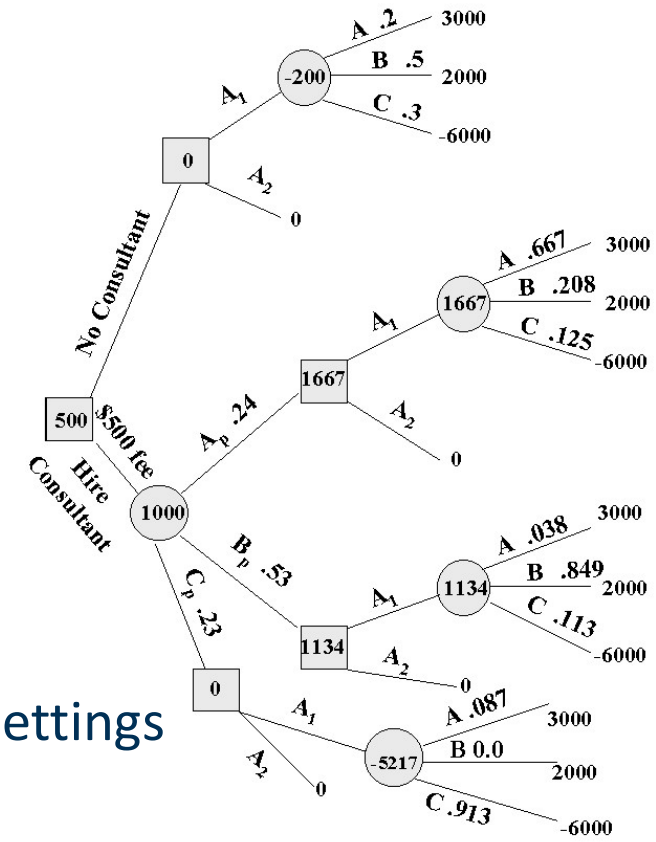
Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html

Data Mining (4)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Input: Preprocessed Data
- Output: **Model / Patterns**

1. Apply data mining method
2. Evaluate resulting model / patterns
3. Iterate
   - Experiment with different (hyper-)parameter settings
   - Experiment with multiple alternative methods
   - Improve preprocessing and feature generation
   - Increase amount or quality of training data
   - Combine different methods



Interpretation / Evaluation (5)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

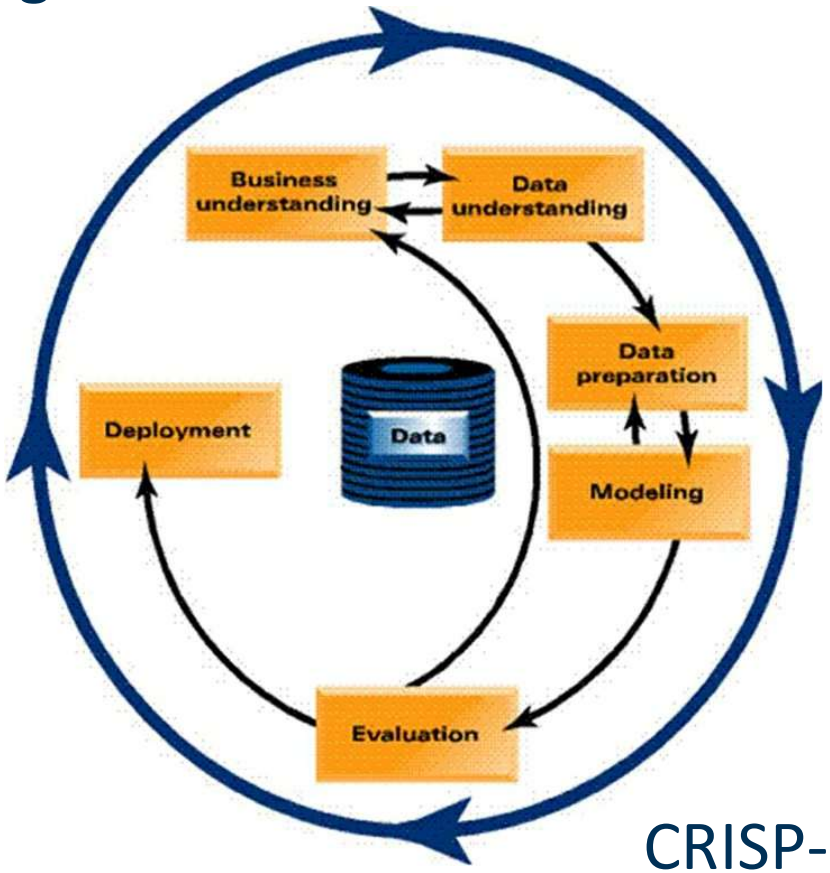
- Output of Data Mining
  - Patterns
  - Models

- In the end, we want to derive value from that, e.g.,
  - Gain knowledge

  - Make better decisions (manual or automated)

Deployment

- Use model in the business context
- Keep iterating in order to maintain and improve model

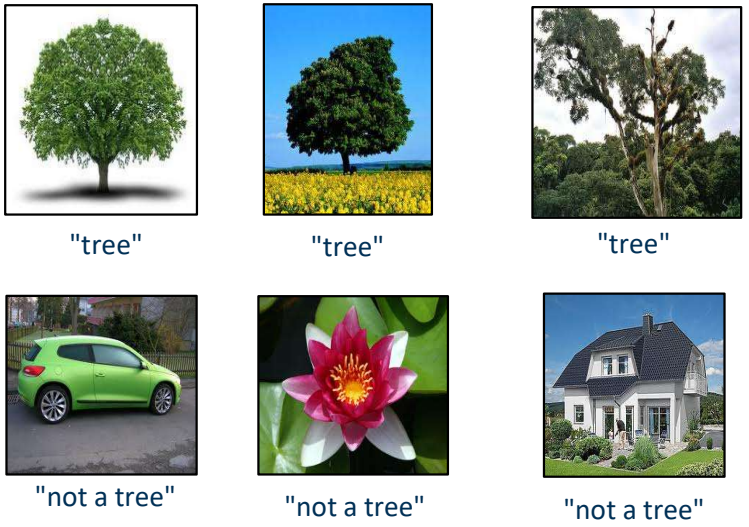


CRISP-DM Process Model

5. Classification

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Classification:
  - We give the computer a set of labeled examples
  - The computer learns to classify new (unlabeled) examples
- Classification Methods
  - **K-Nearest-Neighbors**
  - Decision Trees
  - Naïve Bayes
  - Support Vector Machines
  - Artificial Neural Networks
  - Deep Neural Networks
  - Many others …



K-Nearest-Neighbors

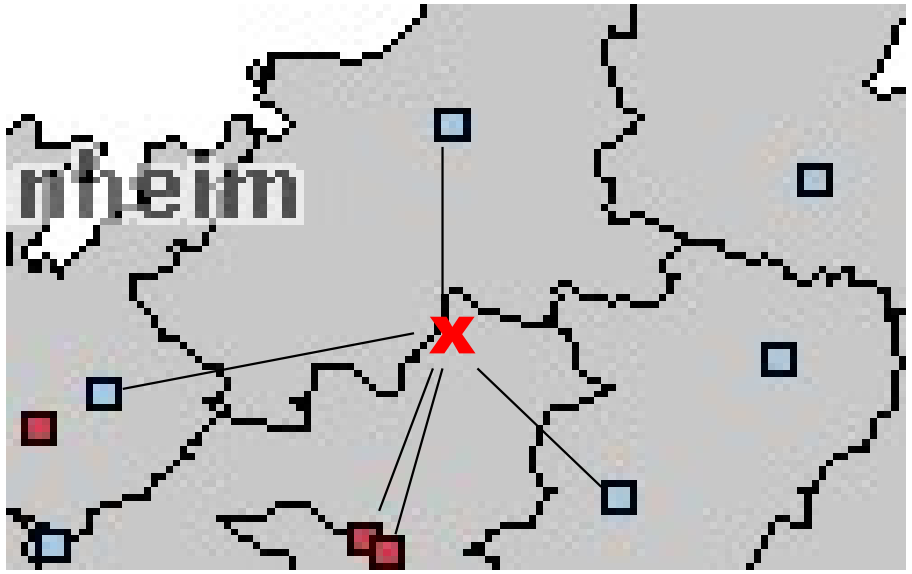
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Problem
  - Predict the current weather in a certain place
  - Where there is no weather station
  - How could you do that?
- Symbols
  - Red = Sunny
  - Blue = Cloudy

K-Nearest-Neighbors

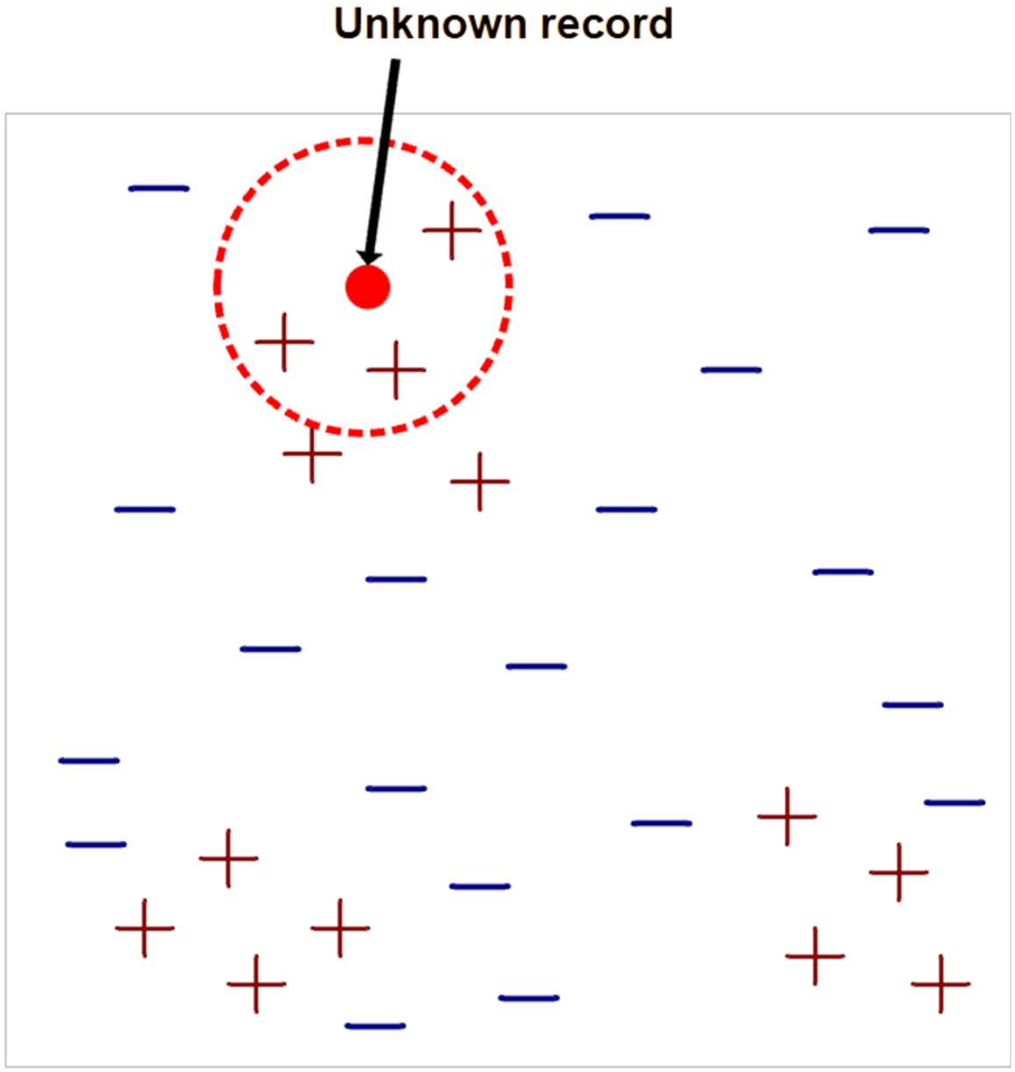
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Idea: use the **average of the nearest stations**
- Example:
  - 2x sunny (red)
  - 3x cloudy (blue)
  - result: cloudy



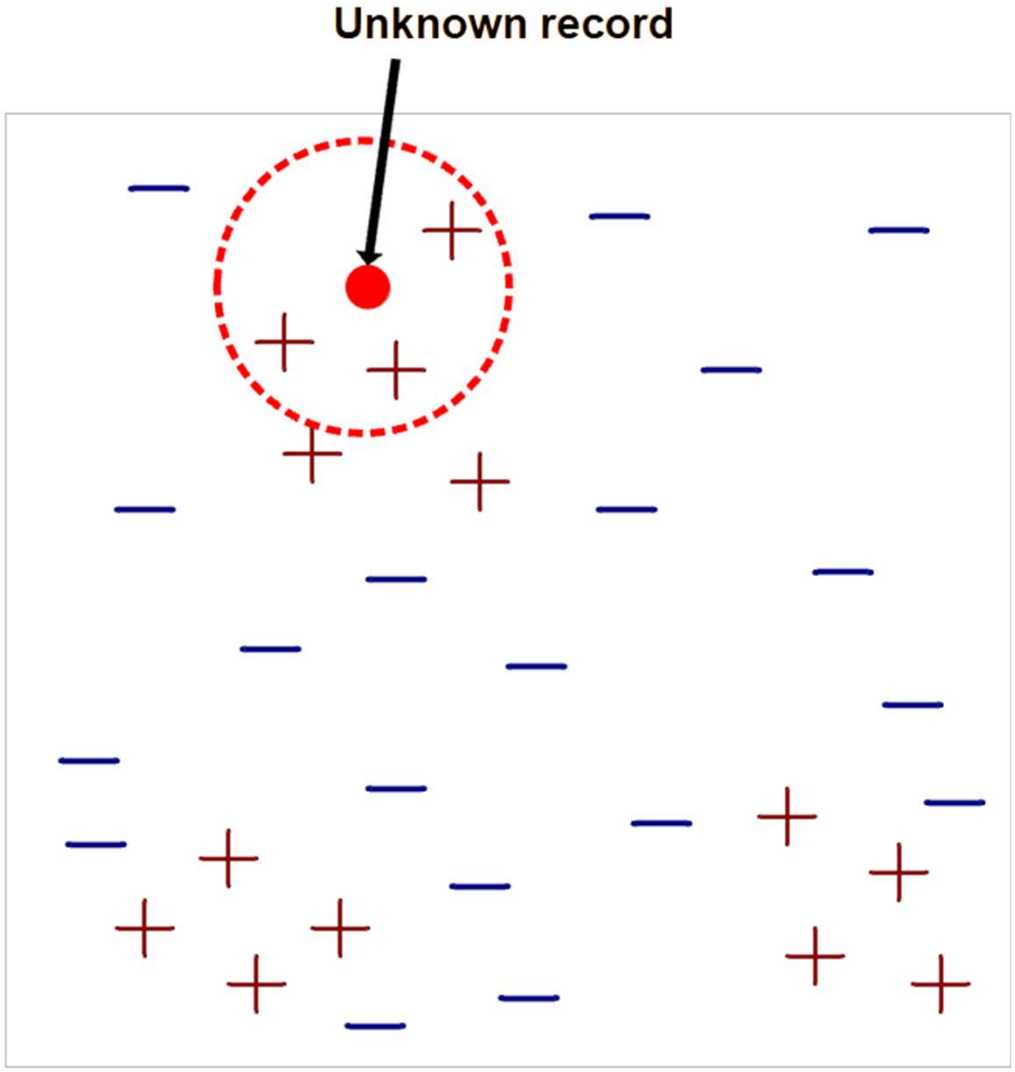
- This approach is called **K-Nearest-Neighbors**
  - where k is the number of neighbors to consider
  - in the example:
    - k=5
    - “near” denotes geographical proximity

K-Nearest-Neighbor Classifier



- Require three things
  - A **set of stored records**
  - A **distance measure** to compute distance between records
  - The **value of k**, the number of nearest neighbors to consider

K-Nearest-Neighbor Classifier

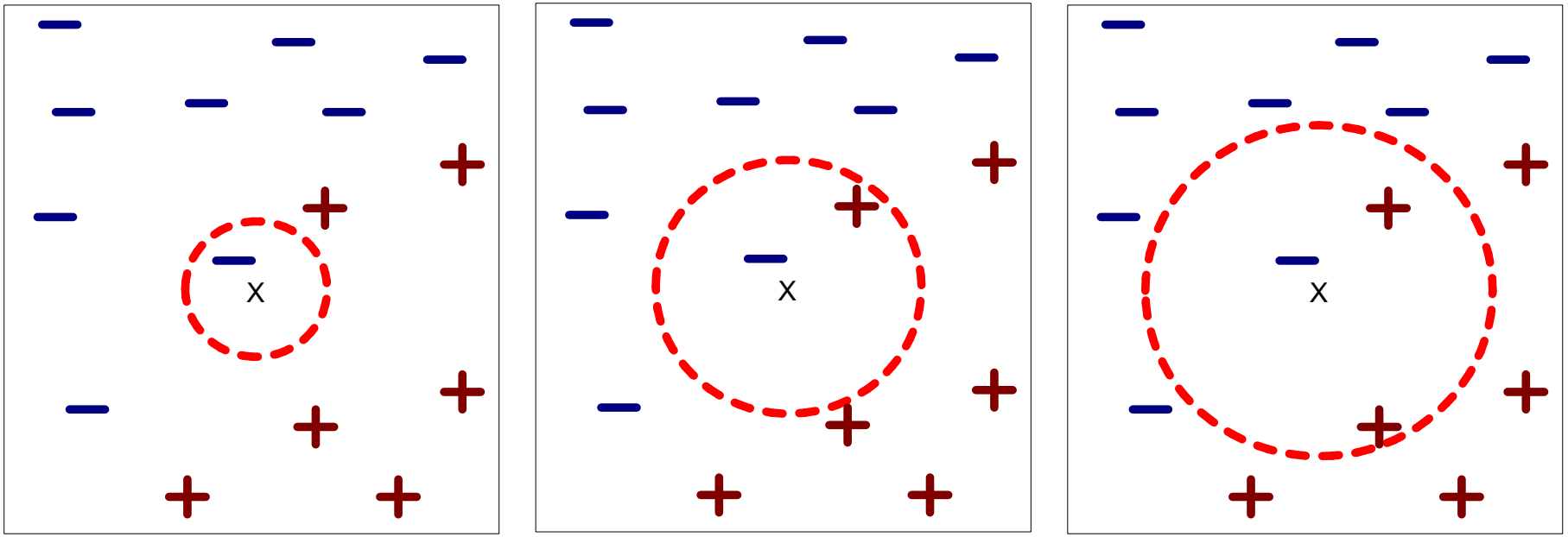


- To classify an unknown record:
  1. **Compute distance** to each training record
  2. Identify **k-nearest neighbors**
  3. Use **class labels of nearest neighbors** to determine the class label of unknown record
     - by taking majority vote or
     - by weighing the vote according to distance

Examples of K-Nearest Neighbors

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- The k-nearest neighbors of a record x are data points that have the k smallest distances to x

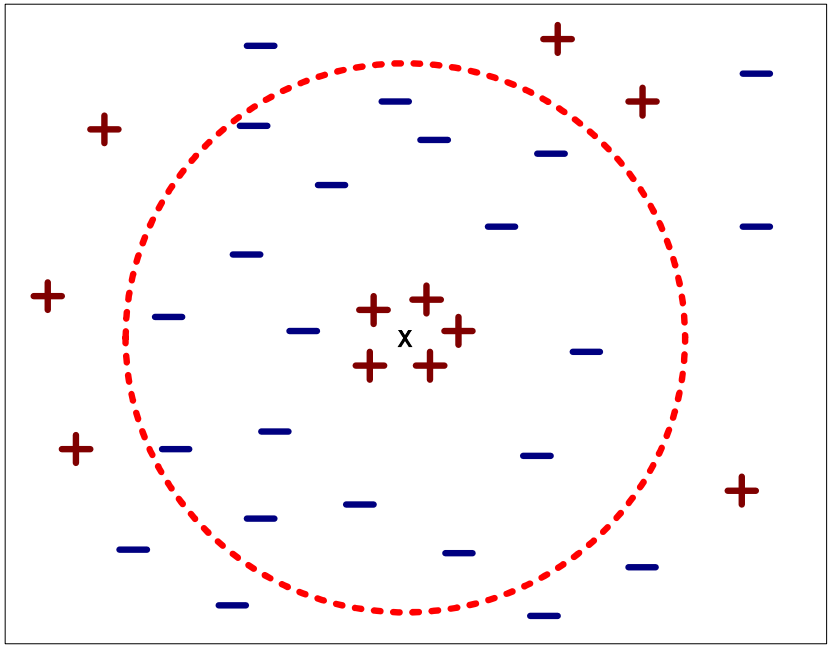


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Choosing a Good Value for K

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- If k is too small, the result is sensitive to noise points
- If k is too large, the neighborhood may include points from other classes



- Rule of thumb: Test k values between 1 and 20

Discussion of K-NN Classification

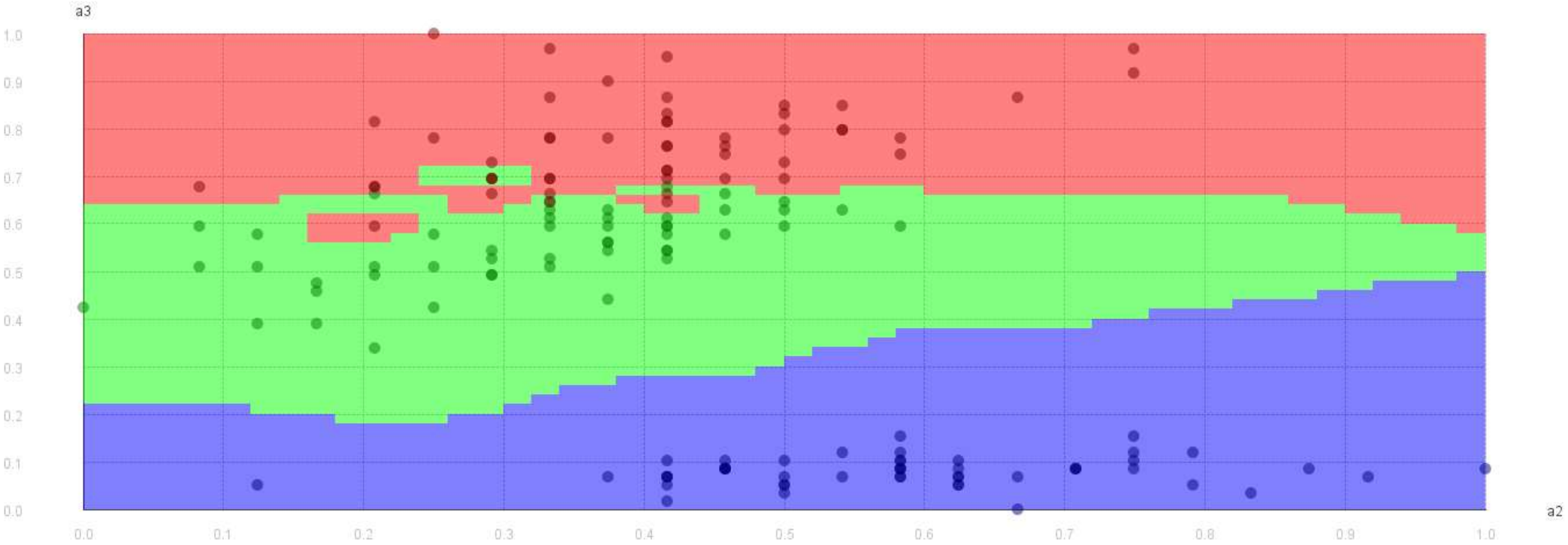
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Often very accurate**
  - for instance for optical character recognition (OCR)
- **… but slow** as unseen record needs to be compared to all training examples
- Results depend on choosing a **good proximity measure**
  - attribute weights, asymmetric binary attributes, …
- KNN can handle decision boundaries which are not parallel to the axes (unlike decision trees)

Decision Boundaries of a k-NN Classifier

UNIVERSITY OF MANNHEIM
Data and Web Science Group

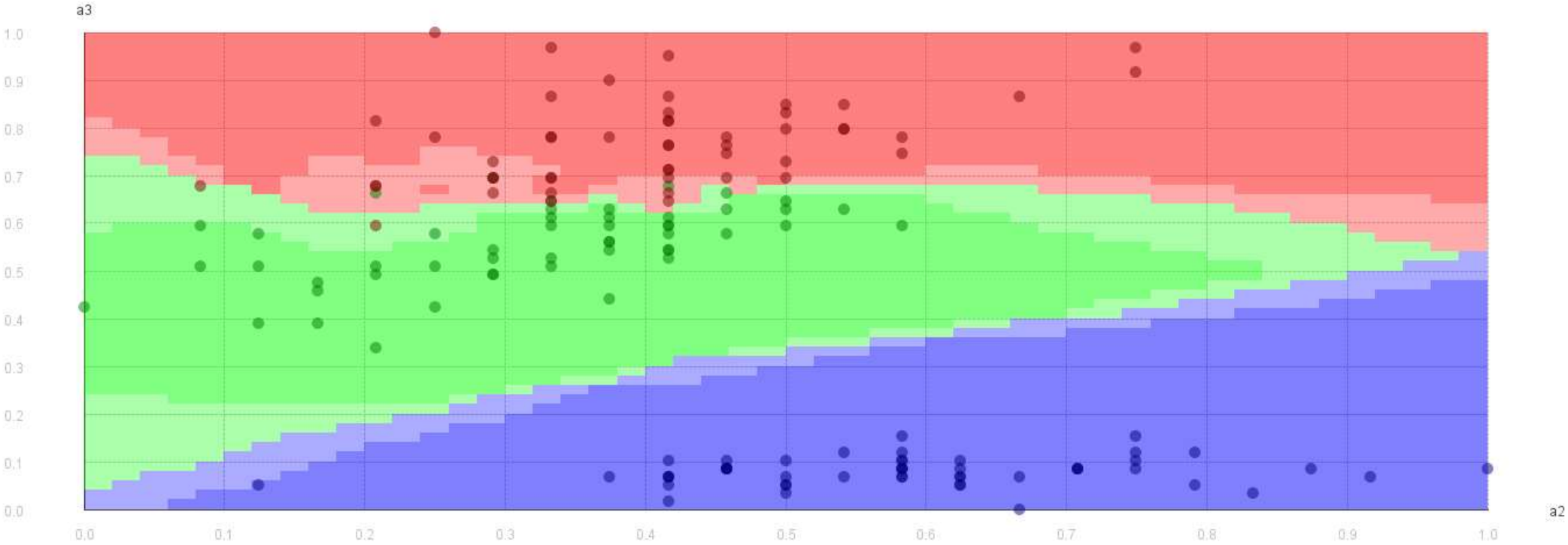
- k=1
- Single noise points have influence on model



Decision Boundaries of a k-NN Classifier

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- k=3
- Boundaries become smoother
- Influence of noise points is reduced



Thank you

Image generated by Gemini