Introduction

IE500 Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Hello

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Dr. Sven Hertling
  - Substitute Professor for Data Science
- Research Interests:
  - Knowledge Graph Integration
  - KGs in combination with Large Language Models
  - Information Extraction

- eMail: sven.hertling@uni-mannheim.de
- Will teach the lectures

Hello

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- M.Sc. Andreea Iana
  - Graduate Research Associate
- Research Interests:
  - Recommender systems
  - Natural language processing
  - Information Retrieval
- Room: B6 26, B 0.18
- eMail: andreea.iana@uni-mannheim.de
- Will teach one of the exercise groups and will supervise student projects

Hello

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- M.Sc. Franz Krause
  - Graduate Research Associate
- Research Interests:
  - Machine Learning Applications on Linked Data
  - Dynamization of Knowledge Graph Embeddings
  - Knowledge Graph Application and Implementation in Industrial Settings
  - Applied Graph Theory
- Room: B6 26, B 0.02
- eMail: franz.krause@uni-mannheim.de
- Will teach one of the exercise groups and will supervise student projects

Course Organization

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

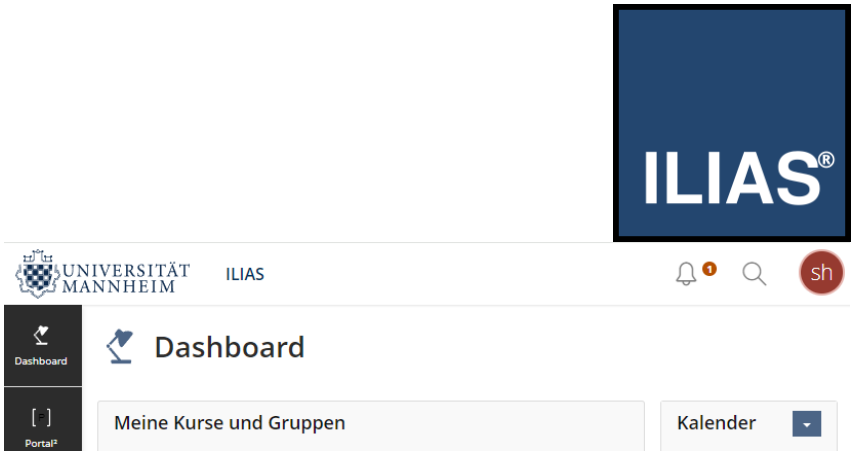
- Lecture
  - Time and Location: Monday, 13:45 – 15:15, B6, 23 room A001
  - introduces the principle methods of data mining
  - discusses how to evaluate the learned models
  - presents practical examples of data mining applications

Course Organization - Material

UNIVERSITY OF MANNHEIM
Data and Web Science Group

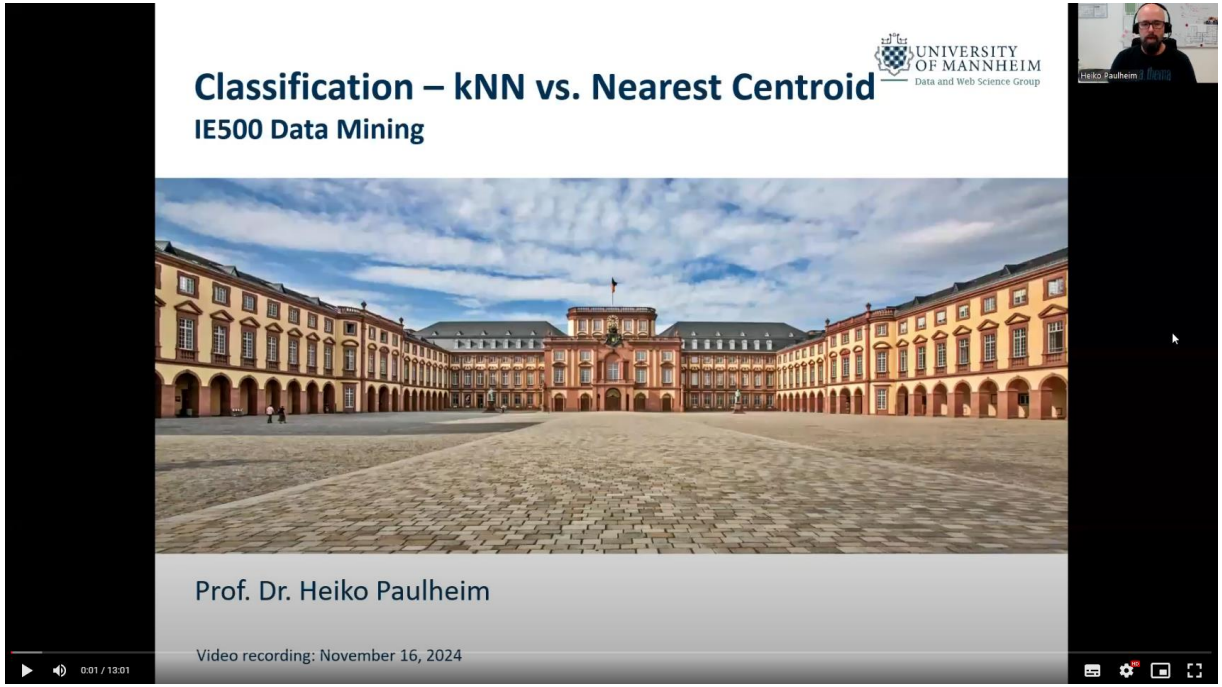
- Course Webpage
  - https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-500-data-mining/
  - provides up-to-date information, video lectures, and exercise material

- ILIAS eLearning System
  - https://ilias.uni-mannheim.de/
  - Mailing Lists, Discussion Forum



Course Organization - Material

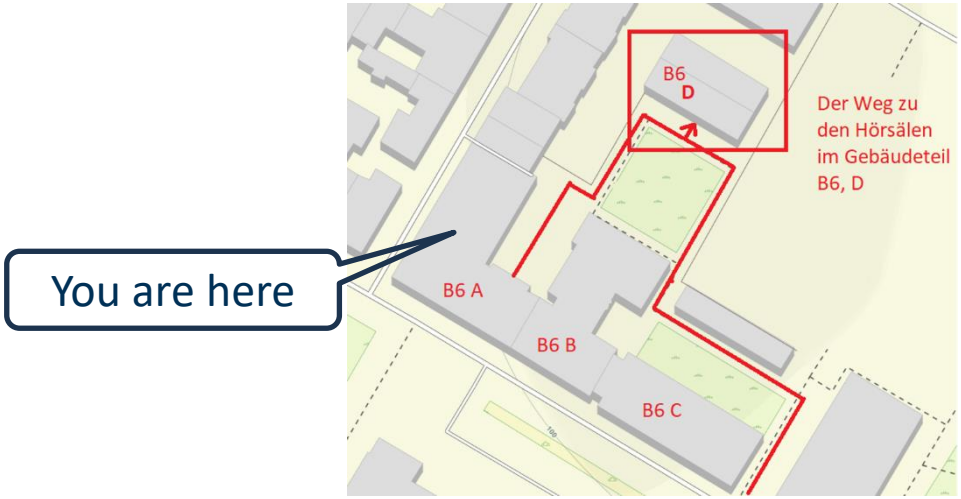
- Additional Videos
  - Part of the course
  - Exercise / **Exam** relevant



Course Organization - Exercise

UNIVERSITY OF MANNHEIM
Data and Web Science Group

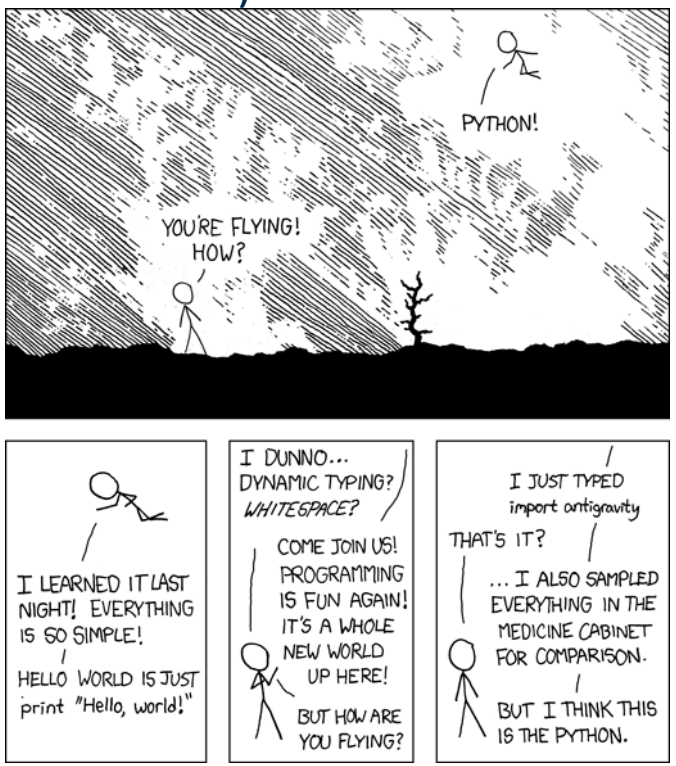
- Exercise Groups
  - Students experiment with data sets using Python
- Time and Location (same content - only attend **one**):
  - Thursday, 12:00 - 13:30, B6,27 Part D – room D007 (2), Franz Krause
  - Thursday, 13:45 - 15:15, B6,27 Part D – room D007 (2), Franz Krause
  - Thursday, 15:30 - 17:00, B6,27 Part D – room D007 (2), Andreea Iana



Introduction to Python

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Already last week Thursday 13:45-15:15
- 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 ChatGPT

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We will be using ChatGPT 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

Course Organization - Project

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Project Work
  - Teams of **five** 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
  - Submission of project work proposal
    - **Sunday, Oct 13st, 23:59**
  - Submission of final project work report
    - **Sunday, Dec 8th, 23:59**
  - Project presentations
    - Schedule to be announced
    - Everyone has to attend

Course Organization - Exam

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Date and Time: **Wednesday, 18th Dezember 2024**
- 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

Course Organization - 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 FSS
  - 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

Course Outline

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. Introduction to Data Mining	What is Data Mining? Tasks and Applications The Data Mining Process
2. Classification	Nearest Neighbor, Decision Trees and Forests,Naïve Bayes, SVMs, Neural Networks, Model Evaluation, Hyperparameter Selection
3. Regression	Linear Regression, Nearest Neighbor Regression, Regression Trees, Time Series
4. Profiling & Preprocessing & Wrangling	PCA, Missing Values, Feature Subset Selection How to deal with image, text
5. Cluster Analysis	K-means Clustering, Density-based Clustering, Hierarchical Clustering, Proximity Measures
6. Association Analysis	Frequent Item Set Generation, Rule Generation, Interestingness Measures

Schedule

UNIVERSITY OF MANNHEIM
Data and Web Science Group

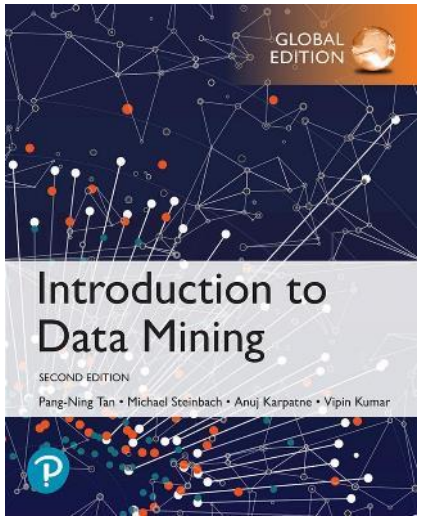
- All updates will be made available at the course homepage

Week	Monday (Lecture)	Thursday (Exercise)
02.09.2024	no lecture	Introduction to Python (13:45–15:15)
09.09.2024	Introduction to Data Mining	Intro
16.09.2024	Classification 1	Classification 1
23.09.2024	Classification 2	Classification 2
30.09.2024	Introduction to the student projects	Public holiday
07.10.2024	Regression / Learning Theory	Regression
14.10.2024	Profiling, Preprocessing, and Wrangling	Preprocessing
21.10.2024	Feedback on project outlines	Project Work
28.10.2024	Clustering and Anomalies	Clustering
04.11.2024	Association Analysis and Subgroup Discovery	Association Analysis
11.11.2024	Project feedback session	Project Work
18.11.2024	Project feedback session	Project Work
25.11.2024	Project feedback session	Project Work
02.12.2024	Project Presentations	Project Presentations

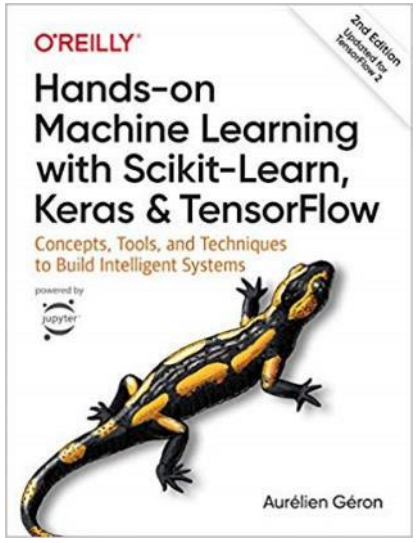
Textbooks for the Course

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

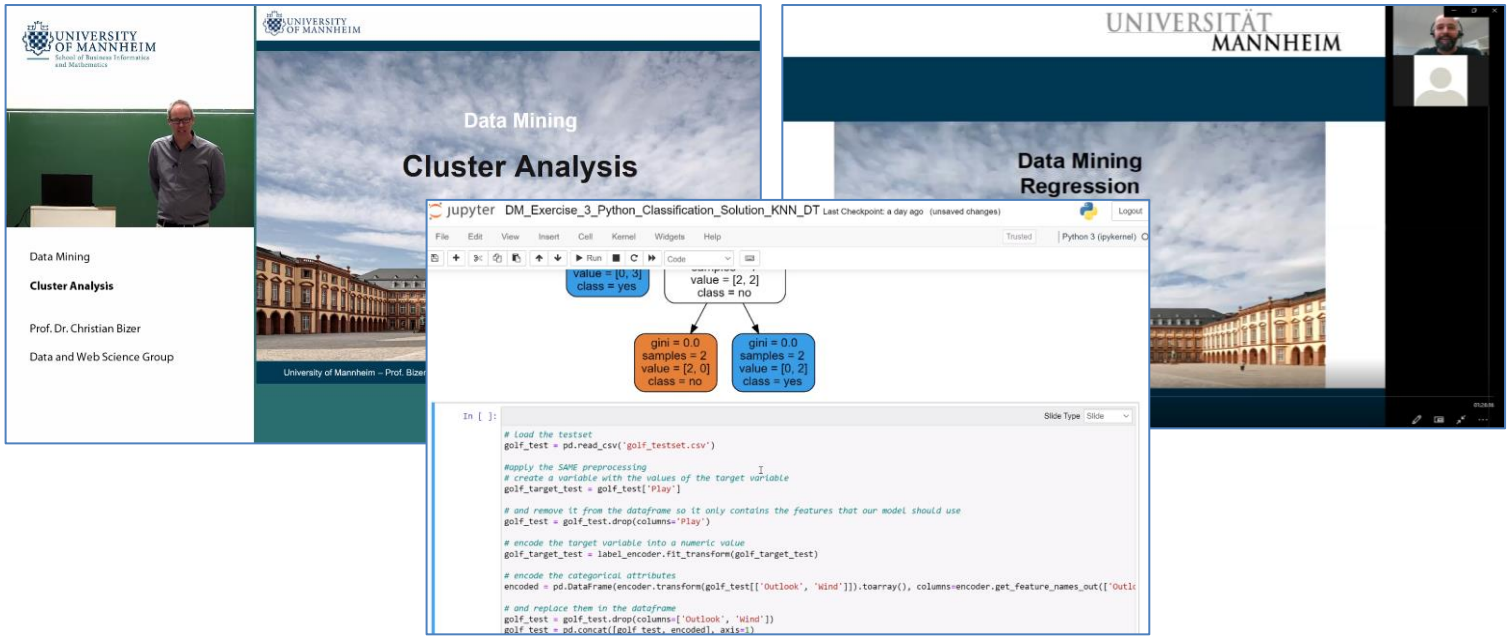


- **Scikit-learn Documentation:** https://scikit-learn.org/stable/user_guide.html

Videos and Screencasts

UNIVERSITY OF MANNHEIM
Data and Web Science Group

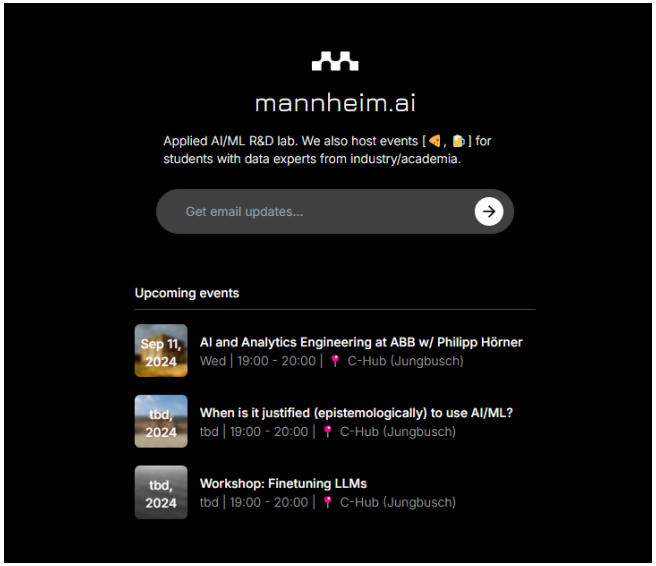
- **Videos**
  - https://www.uni-mannheim.de/dws/teaching/lecture-videos (VPN!)
  - **Lecture Videos** By Heiko Paulheim (HWS 2020) and Christian Bizer (FSS 2020)
  - **Screencasts** for the Exercises by Ralph Peeters (FSS 2022)
- Keep in mind, that the lecture and exercise change over time



Advertisement

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Former student is organizing presentations of companies in Mannheim
  - If you are interested, visit https://mannheim.ai
  - 11.09. presentation of ABB
  - In the future: CTO of Osapiens, Head Data Scientist of Vattenfall

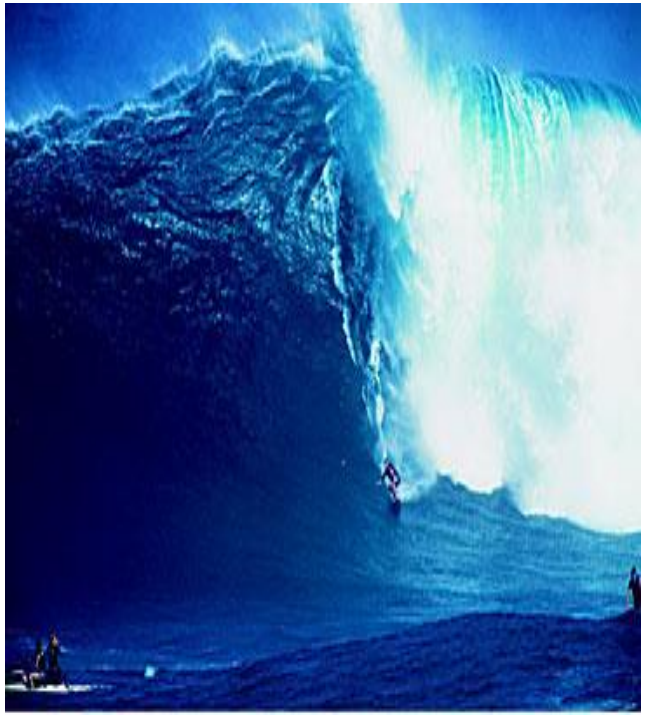


Questions?

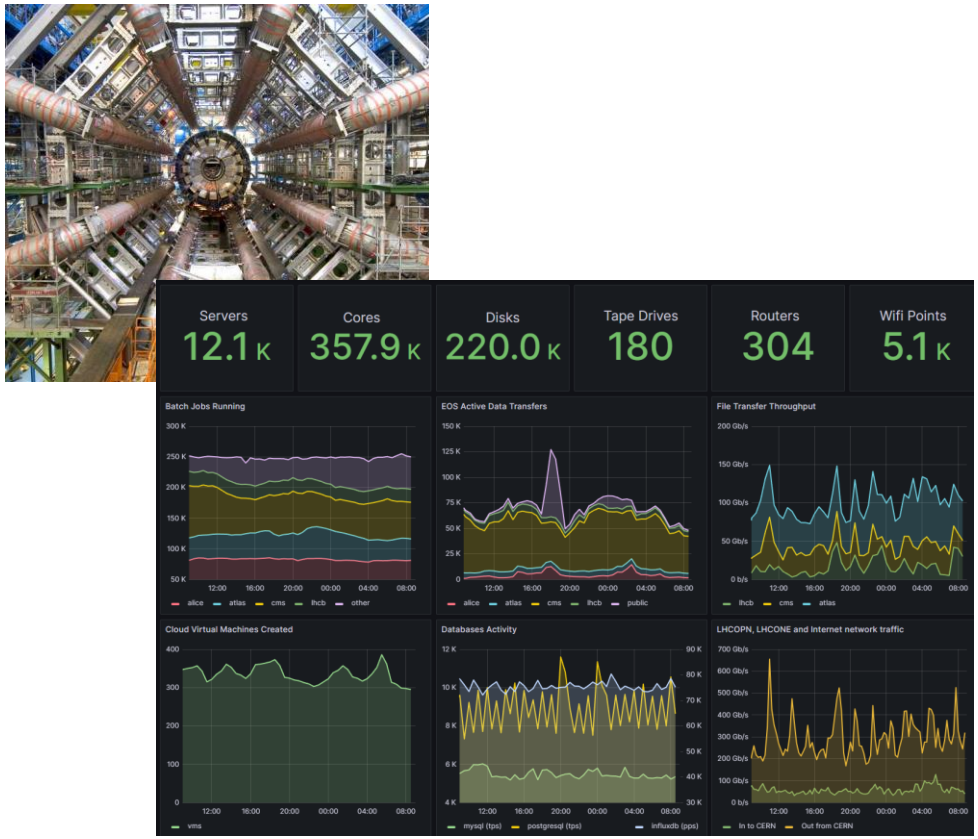
What is Data Mining?

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...”

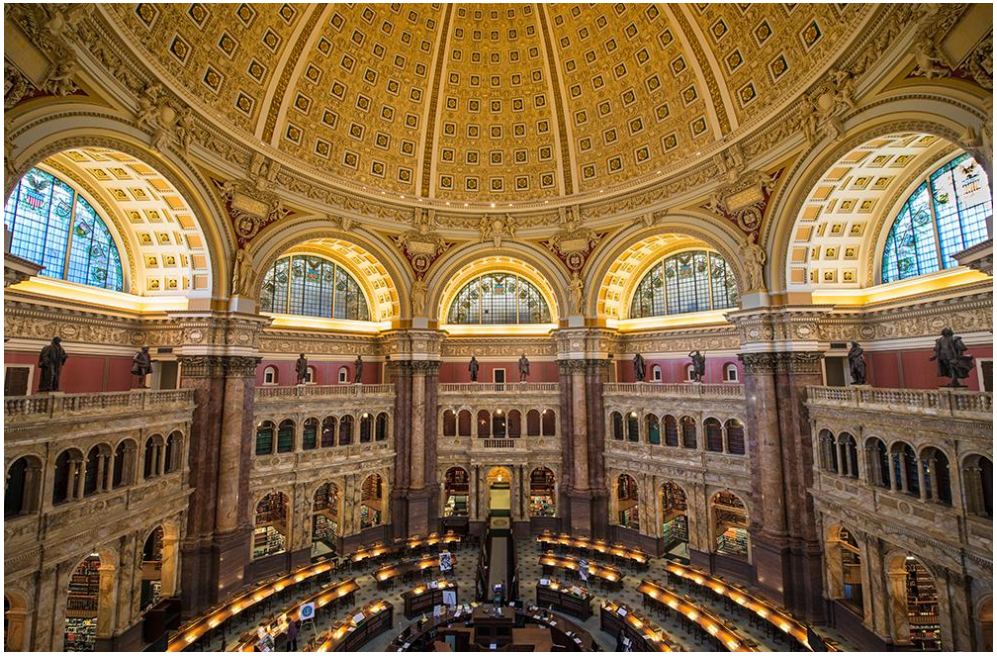
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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

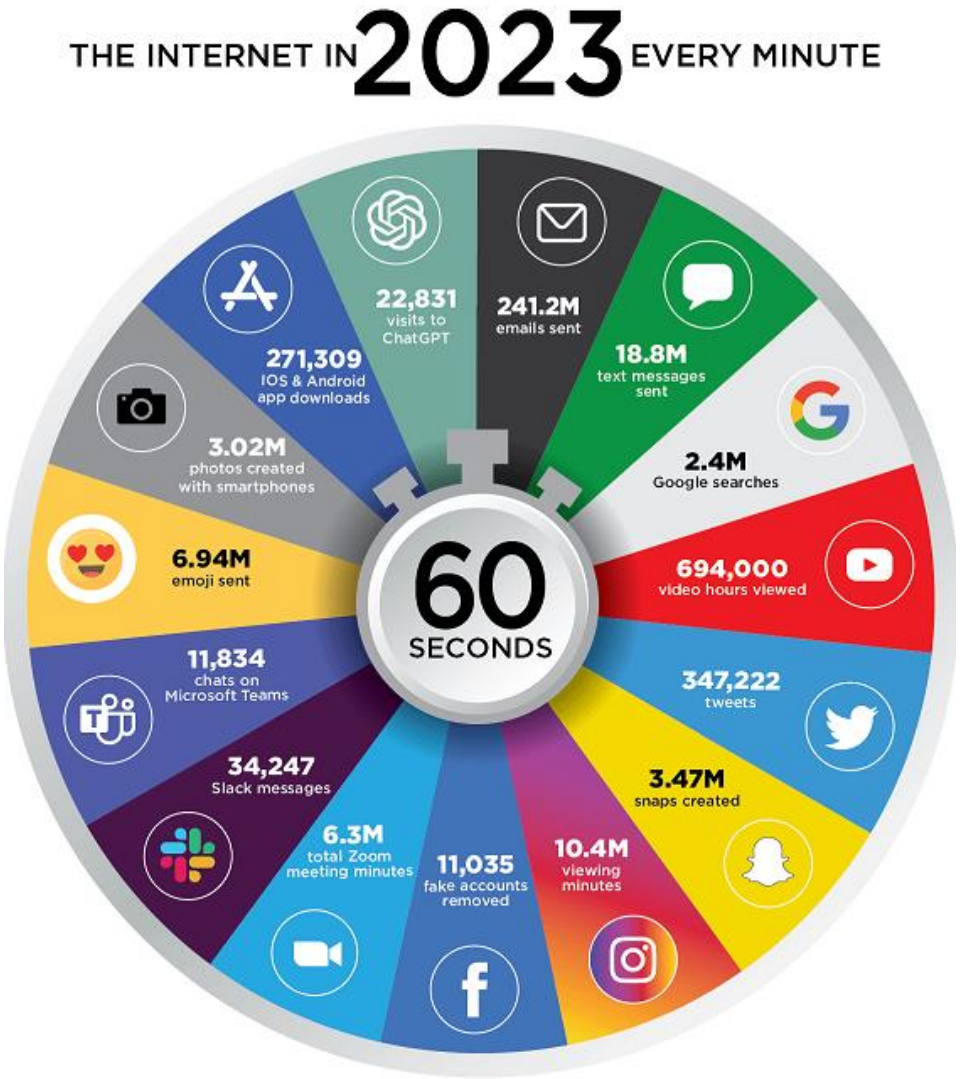


- **US Library of Congress**
  - ≈ 235 TB archived
- Discover
  - Topic distributions*
  - Citation networks
- Train
  - Large Language Models

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

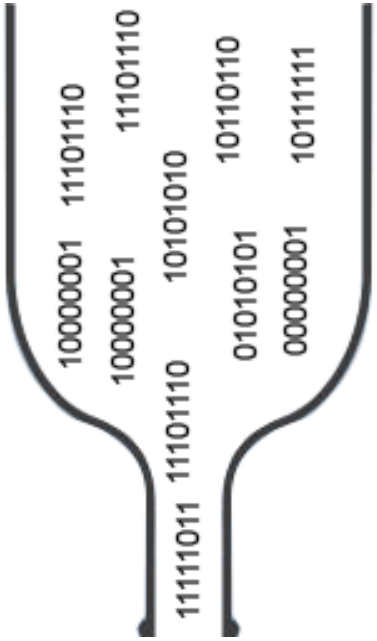
“We are Drowning in Data...”

UNIVERSITY OF MANNHEIM
Data and Web Science Group



https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

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

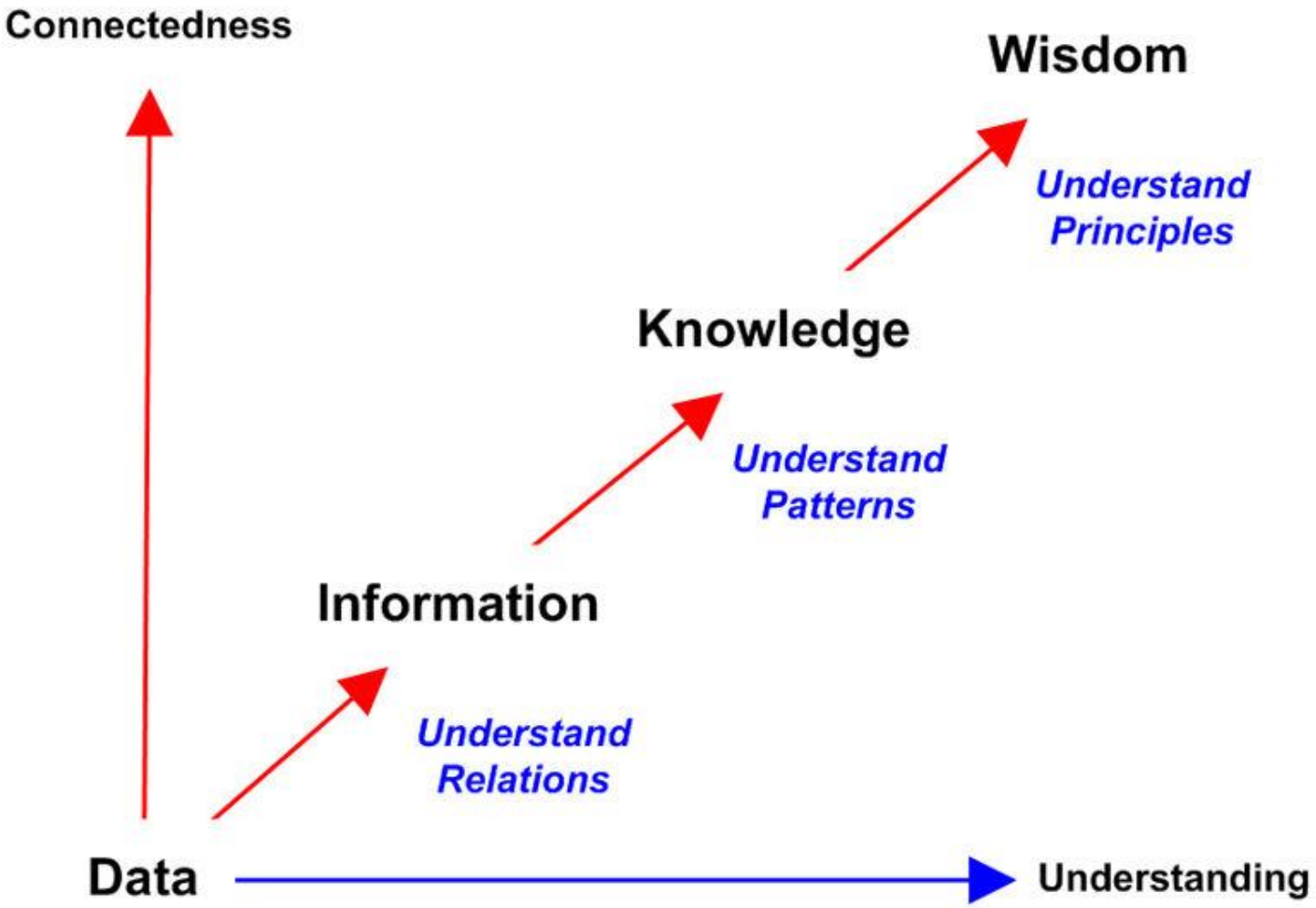


← Rate at which data are produced

← Rate at which data can be understood
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, Information, Knowledge, Wisdom

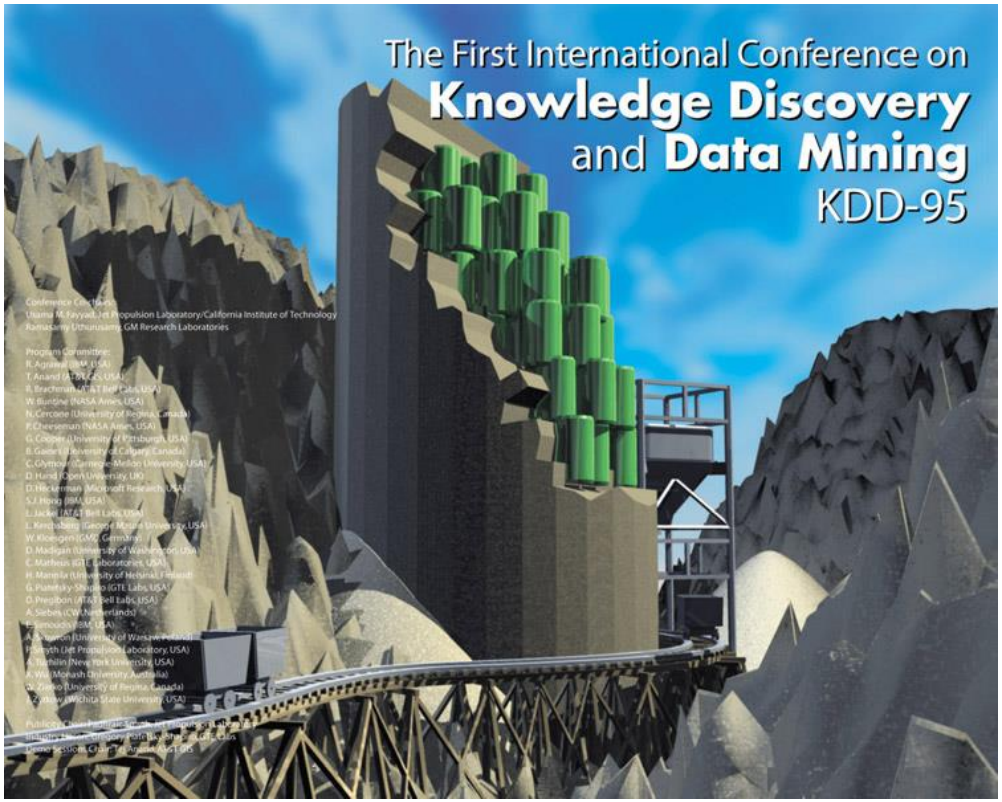


Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."

Data Mining: Definitions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Idea: mountains of data
  - Where 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

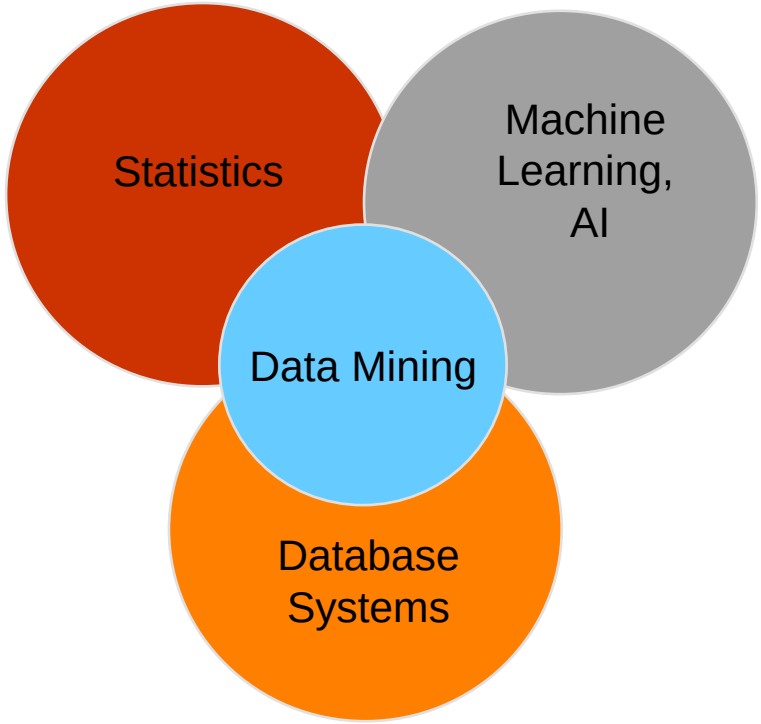
  patterns in data. (Fayyad et al. 1996)

- 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

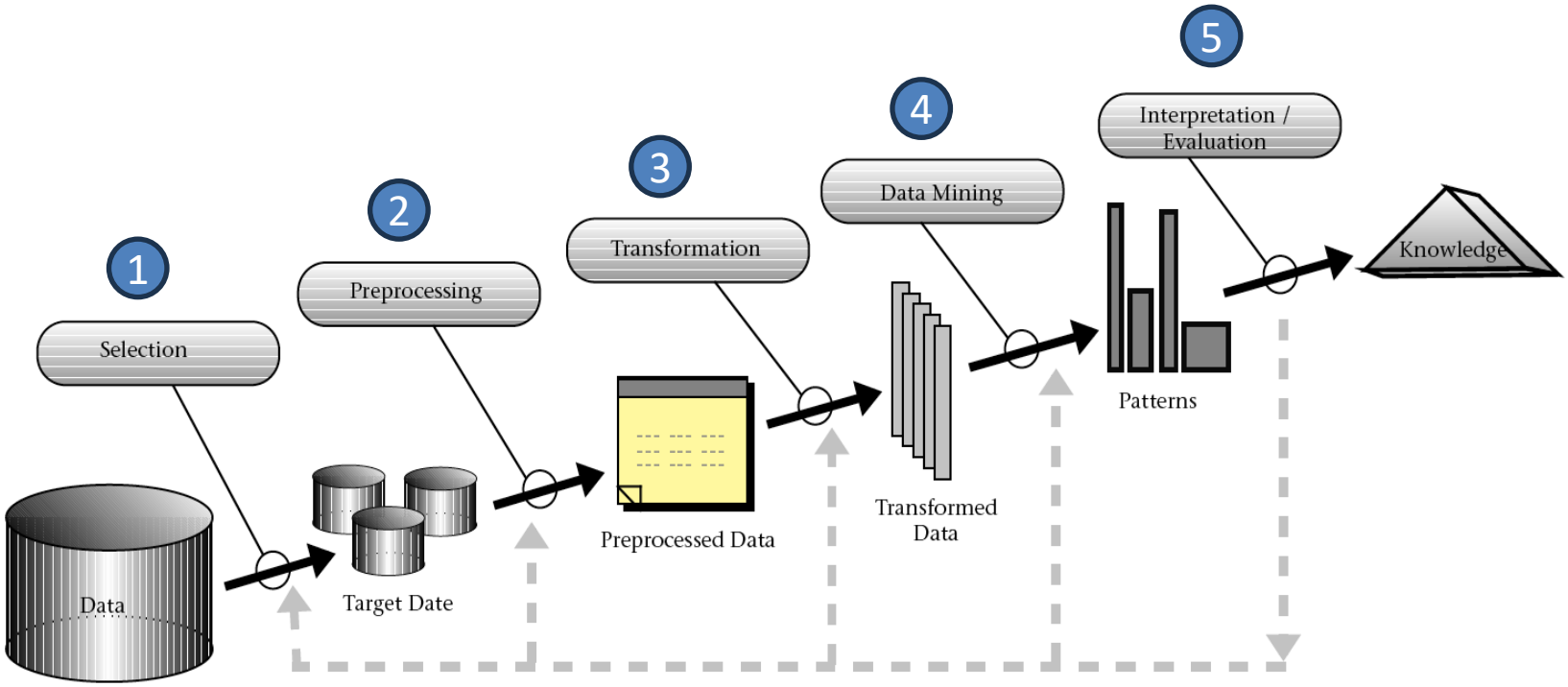
Origins of Data Mining

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



The Data Mining Process

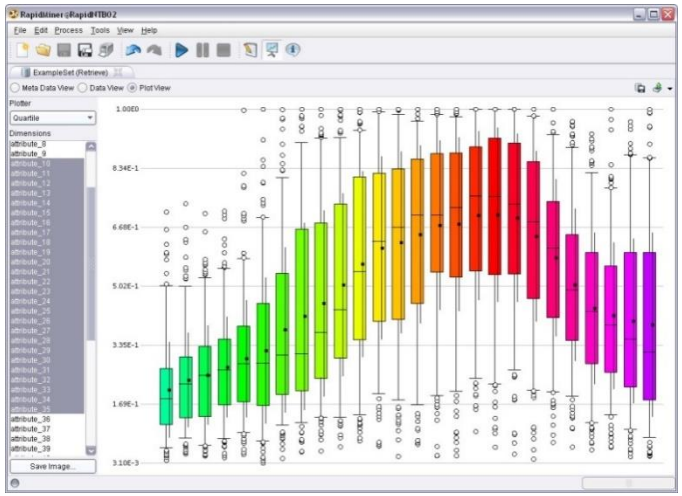


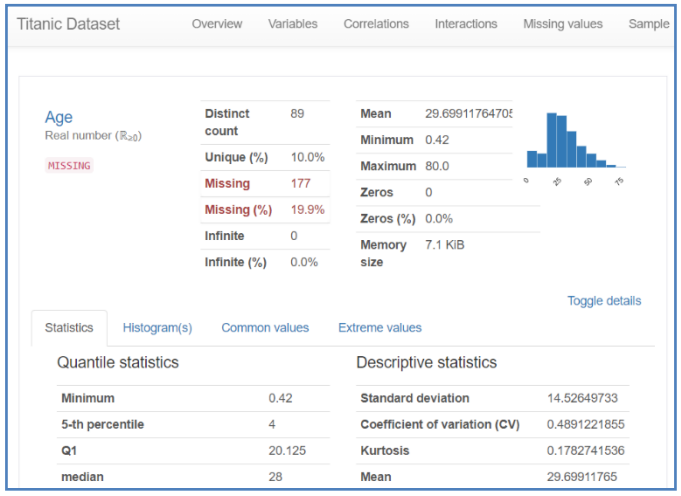
Source: Fayyad et al. (1996)

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

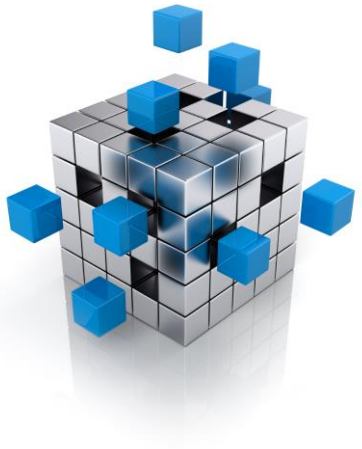




Preprocessing and Transformation (2+3)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

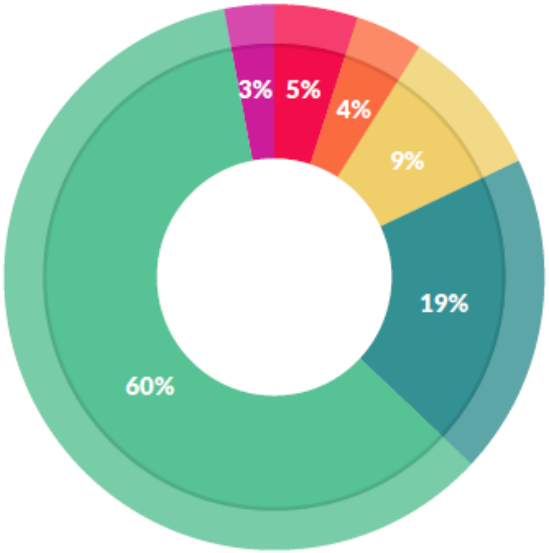
- Transform data into a representation that is suitable for the chosen data mining methods
  - Number of dimensions (represent relevant information using less attributes)
  - Scales of attributes (nominal, ordinal, numeric)
  - Amount of data (determines hardware requirements)
- Methods
  - Discretization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

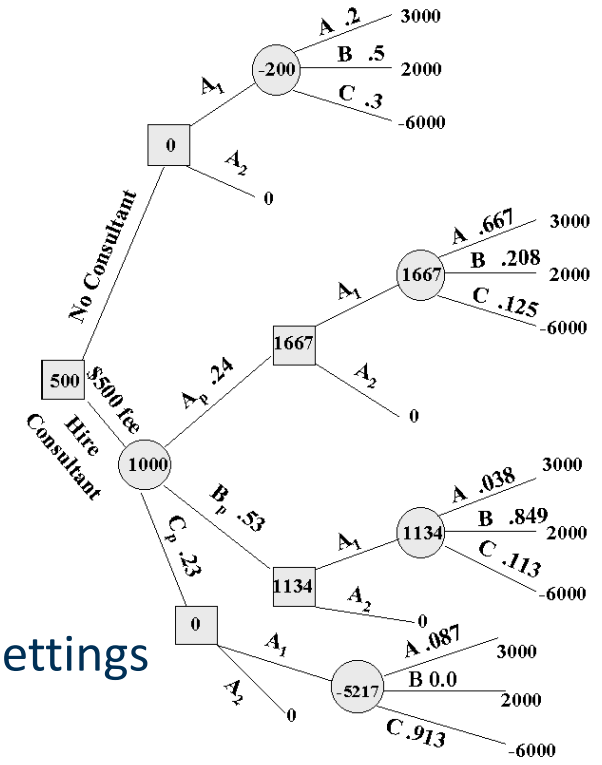
Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html

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

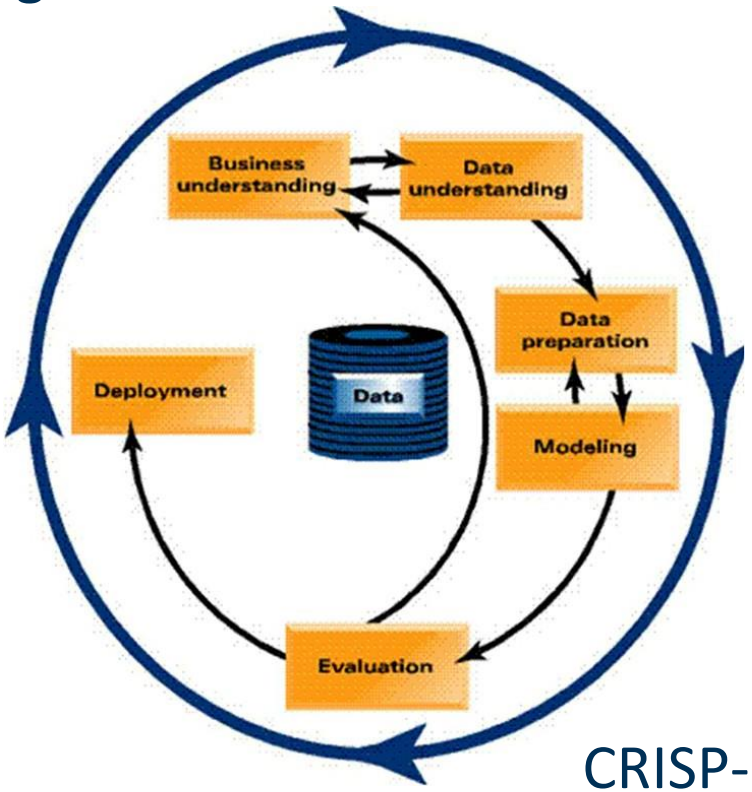
- In the end, we want to derive value from that, e.g.,
  - Gain knowledge

  - Make better decisions

  - Increase revenue

Deployment

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

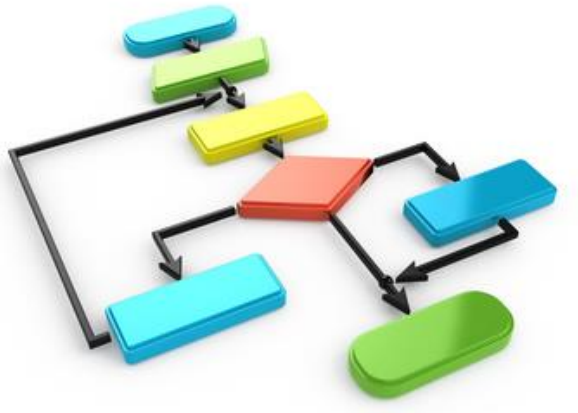


CRISP-DM Process Model

Tasks and Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group

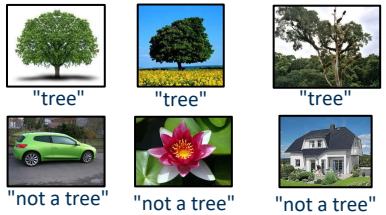
- **Descriptive** Tasks
  - Find patterns in the data
    - E.g. which products are often bought together?
- **Predictive** Tasks
  - Predict unknown values of a variable
    - Given observations (e.g., from the past)
    - E.g. will a person click a online advertisement?
      - given her browsing history



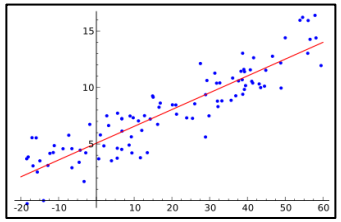
- Machine Learning Terminology
  - Descriptive = unsupervised
  - Predictive = supervised

Data Mining Tasks

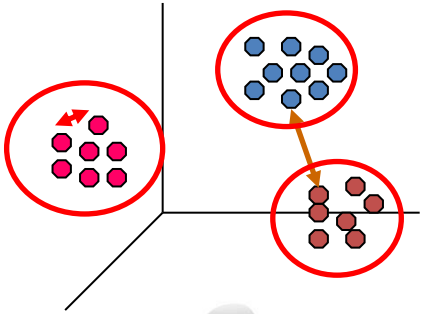
- Classification [Predictive]



- Regression [Predictive]

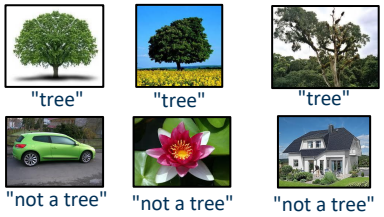


- Cluster Analysis [Descriptive]



- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

"tree"

"tree"

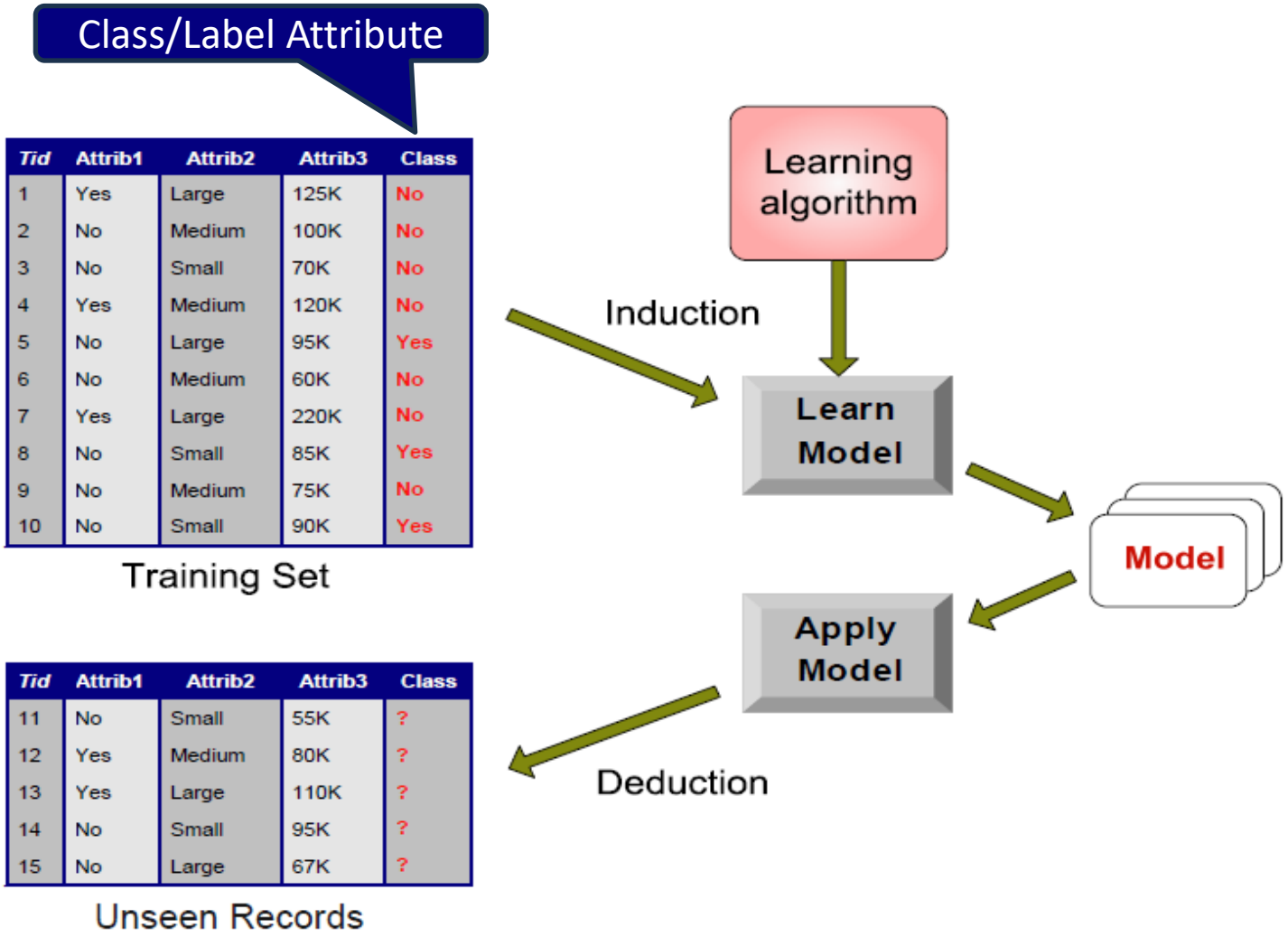
"tree"

"not a tree"

"not a tree"

"not a tree"

Classification: Workflow



Classification: Applications

UNIVERSITY OF MANNHEIM
Data and Web Science Group

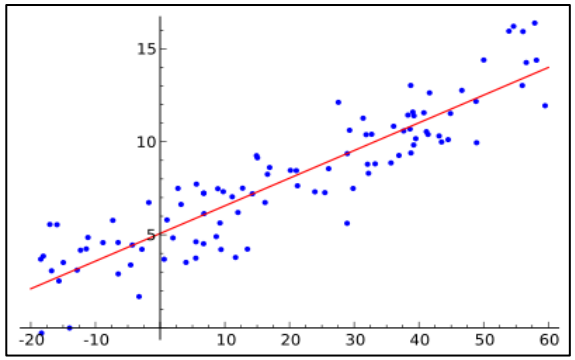
- Credit Risk Assessment
  - Attributes: your age, income, debts, …
  - Class: are you getting credit by your bank?

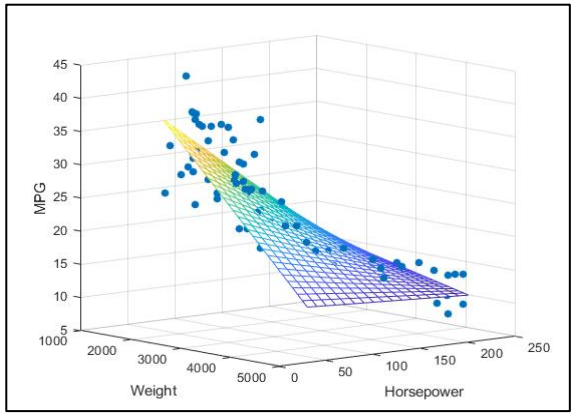
- SPAM Detection
  - Attributes: words and header fields of an e-mail
  - Class: regular e-mail or spam e-mail?

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

Regression

- 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.



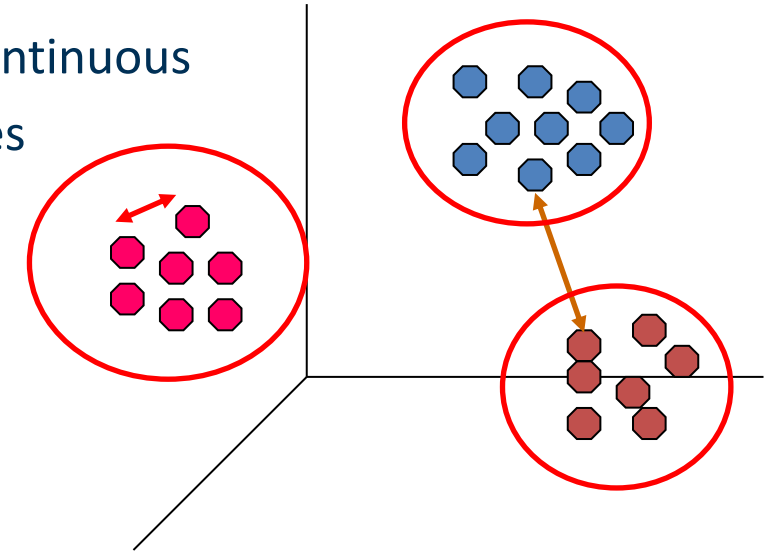


- Difference to classification: The predicted attribute is **continuous**, while classification is used to predict nominal attributes (e.g. yes/no)

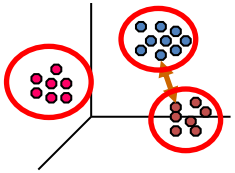
Cluster Analysis

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

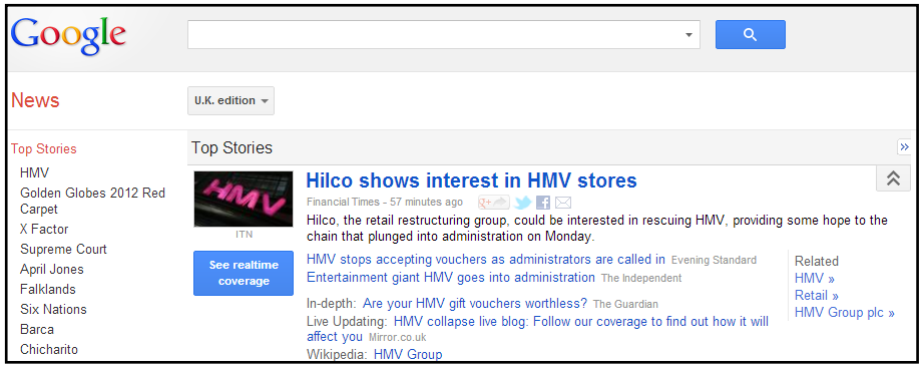


Cluster Analysis: Applications



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

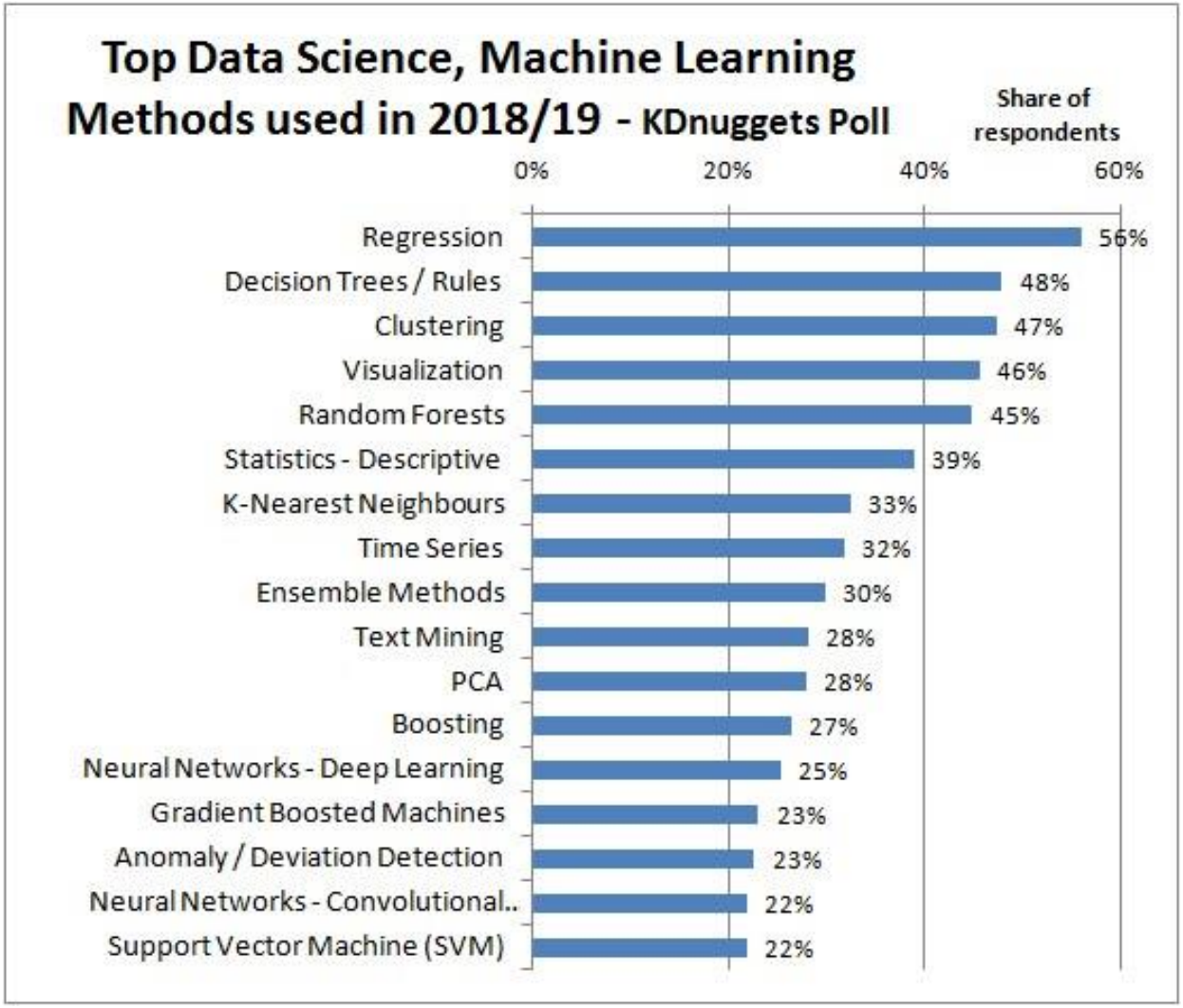
Frequently Bought Together

Price For All Three: $87.41

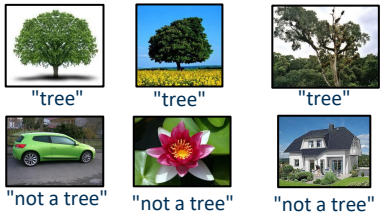
Add all three to Cart | Add all three to Wish List

Show availability and shipping details

Which Methods are Used in Practice?

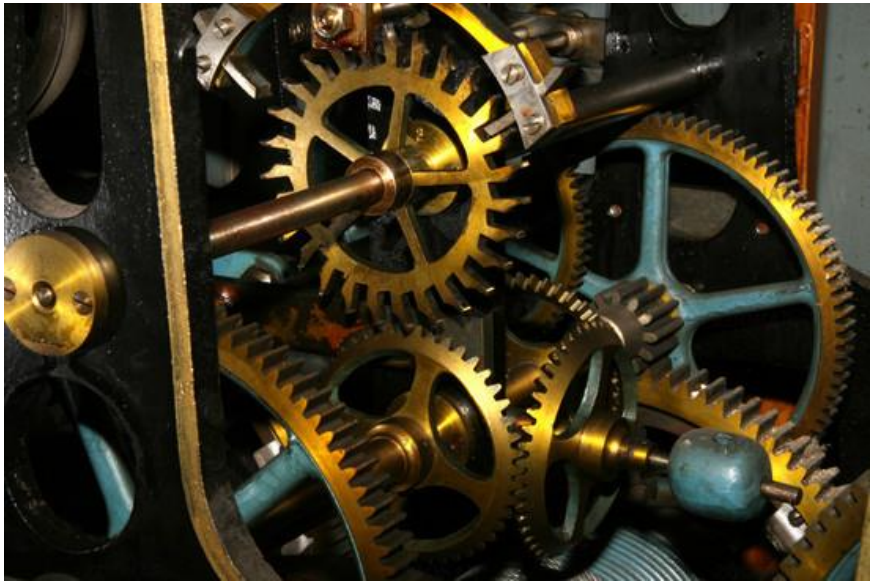


Classification Algorithms



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
- How does that work?
  - **K-Nearest-Neighbors**
  - Decision Trees
  - Naïve Bayes
  - Support Vector Machines
  - Artificial Neural Networks
  - Deep Neural Networks
  - Many others …



K-Nearest-Neighbors

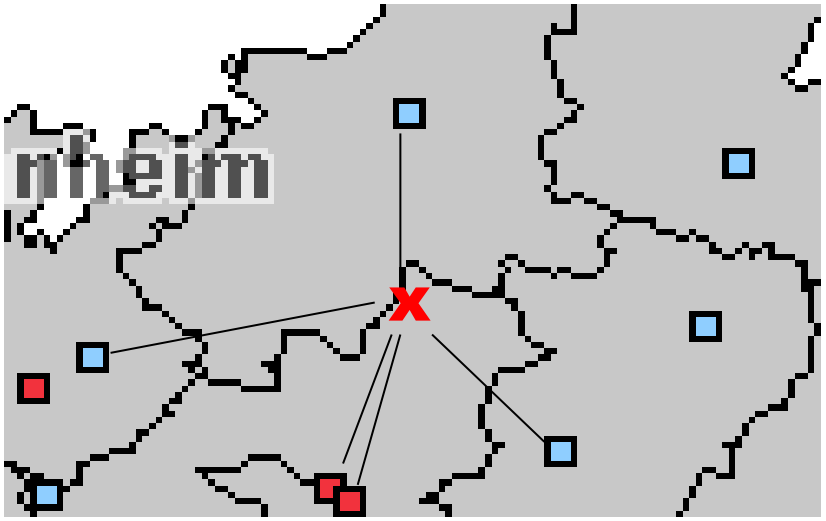
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

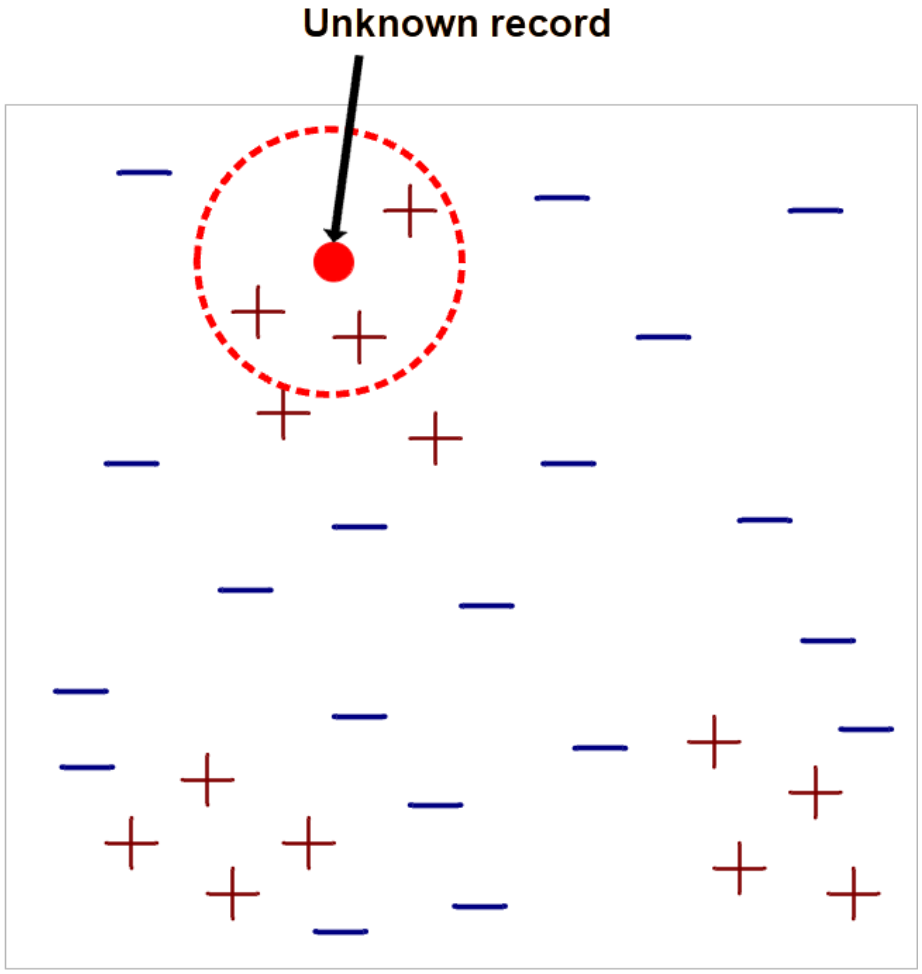
UNIVERSITY OF MANNHEIM
Data and Web Science Group

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



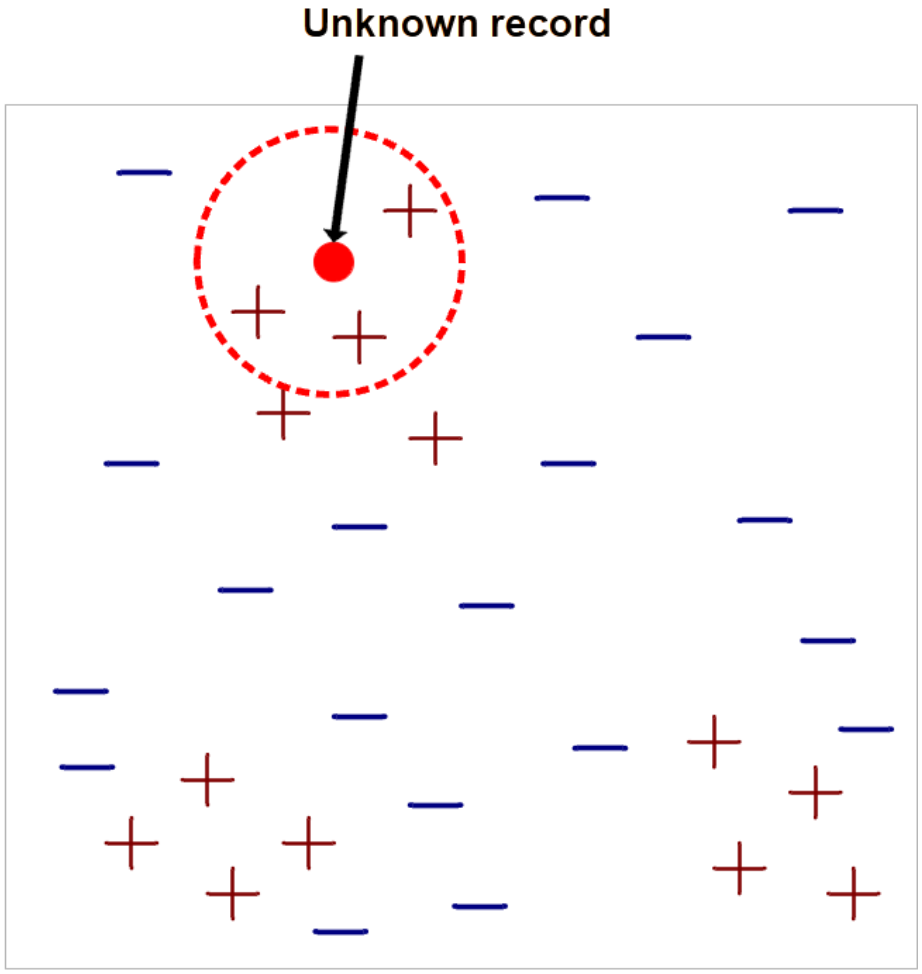
- This approach is called **K-Nearest-Neighbors**
  - where k is the number of neighbors to consider
  - in the example:
    - k=5
    - “near” denotes geographical proximity

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

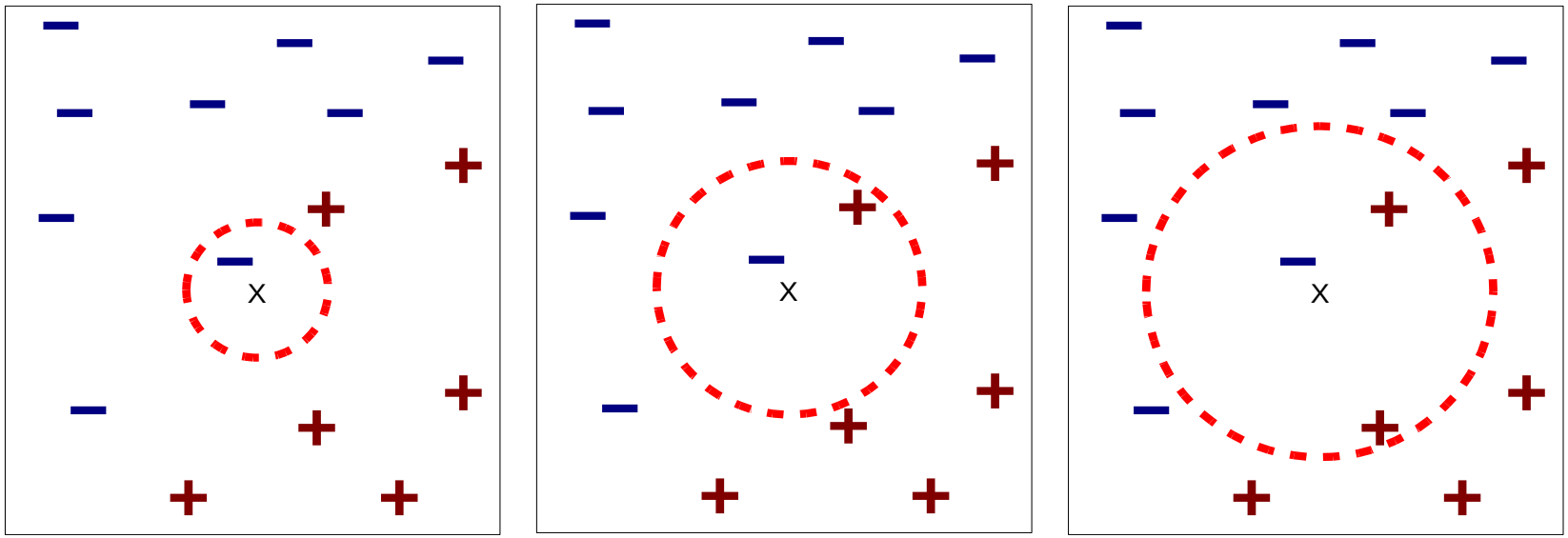


- To classify an unknown record:
  - **Compute distance** to each training record
  - Identify **k-nearest neighbors**
  - Use **class labels of nearest neighbors** to determine the class label of unknown record
    - By taking majority vote or
    - By weighing the vote according to distance

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

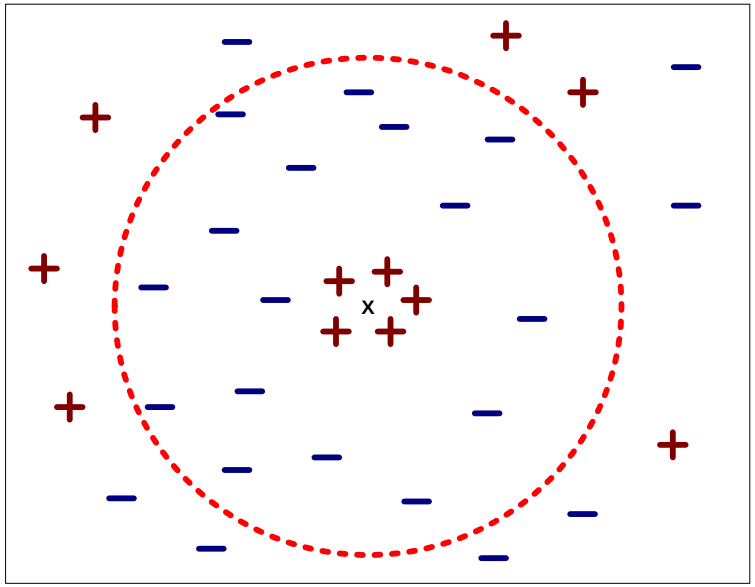


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

Experiment

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Trying to predict
  - Do you want to watch 

(release 24.10.2024)
- Binary attributes: have you watched these films?
  - Mission: Impossible 7 – Dead Reckoning (13.07.2023 - Action)
  - Der Super Mario Bros. Film (05.04.2023 - Animation)
  - Indiana Jones und das Rad des Schicksals (29.06.2023 – Adventure)
  - Scream 6 (09.03.2023 – Horror)
  - Oppenheimer (20.07.2023 – Drama)
  - Barbie (20.07.2023 – Comedy)
  - Luther: The Fallen Sun (10.03.2023 – Crime)
  - Guardians of the Galaxy Vol. 3 (03.05.2023 - Action)

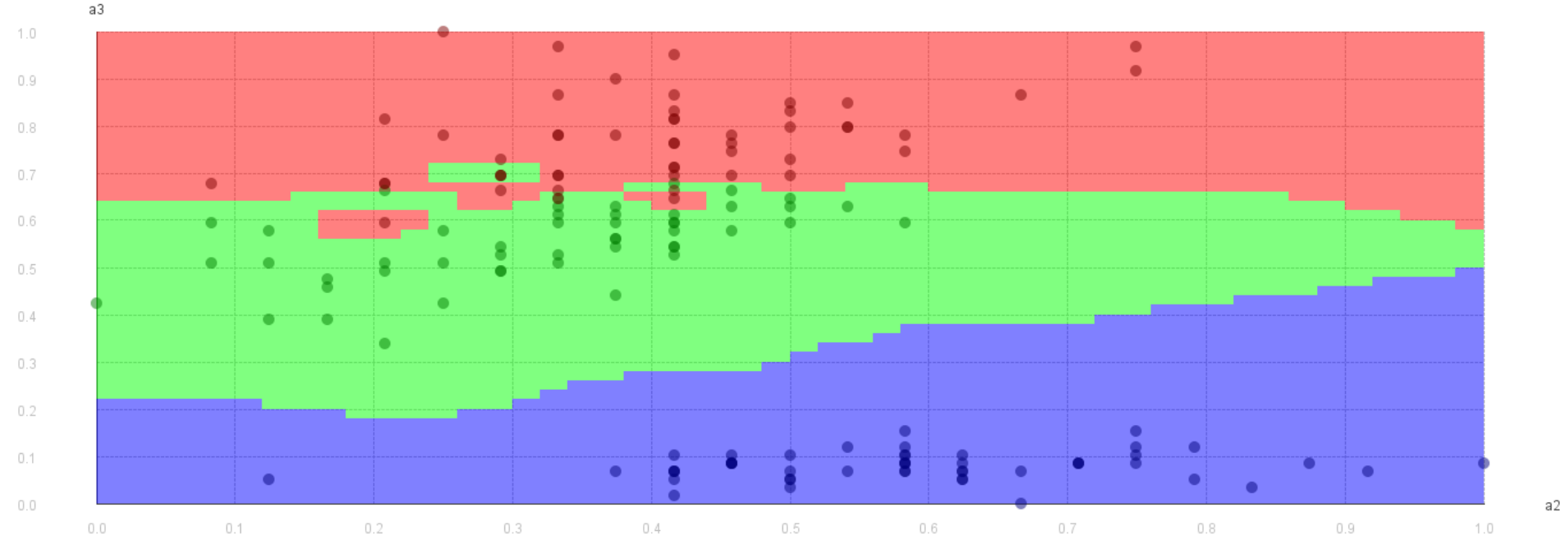
Discussion of K-NN Classification

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

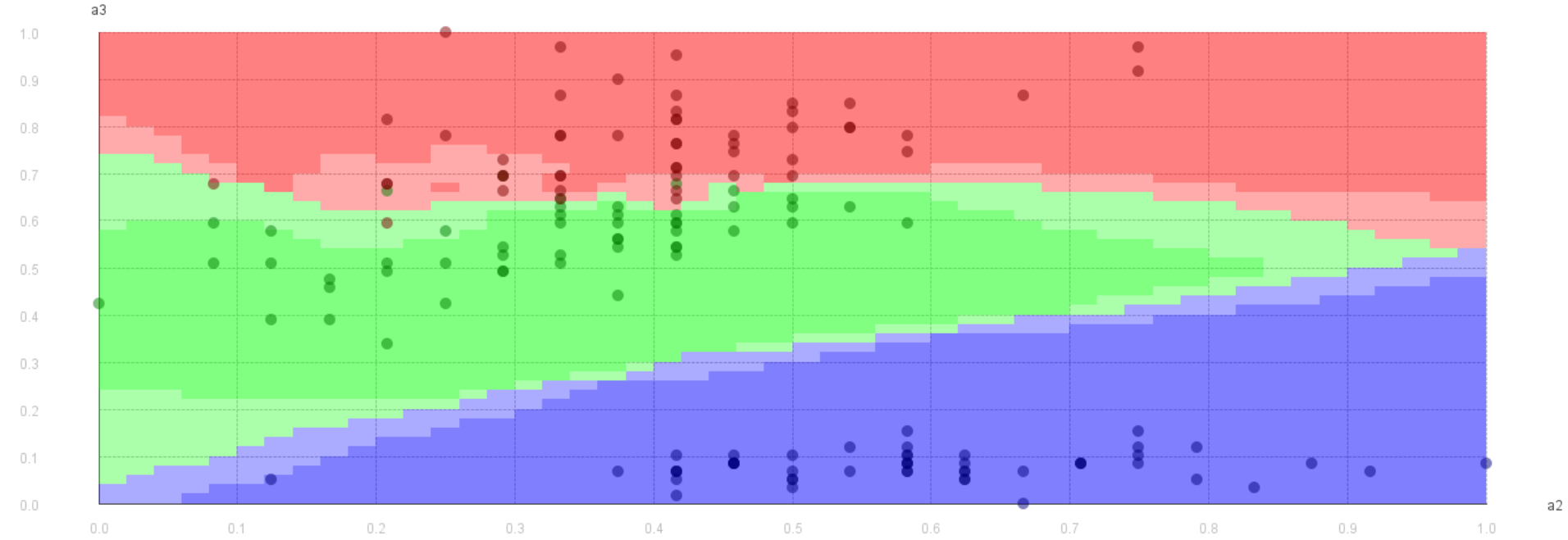
- k=1
- Single noise points have influence on model



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



KNN in Python

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Python

```python
from sklearn.neighbors import KNeighborsClassifier

# Train classifier
knn_estimator = KNeighborsClassifier(n_neighbors=3)
knn_estimator.fit(preprocessed_training_data, training_labels)

# Use classifier to predict labels
prediction = knn_estimator.predict(preprocessed_unseen_data)
```

What You Will Learn in This Lecture

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Common data mining tasks
  - How they work
  - When and how to apply them
  - How to interpret their output

Thank you

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Are there any questions?

- Next …
  1. Install the Anaconda Python distribution
  2. Get an OpenAI account for using ChatGPT
  3. Attend exercise on Thursday
  4. Attend the lecture next week