Introduction

IE500 Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Hello

- Dr. Sven Hertling
  - Substitute Professor for Data Science
- Research Interests:
  - Knowledge Graph Integration
  - KGs in combination with Large Language Models
  - Information Extraction
- Room: B6 26, B0.21
- eMail: sven.hertling@uni-mannheim.de
- Will teach the lectures

Hello

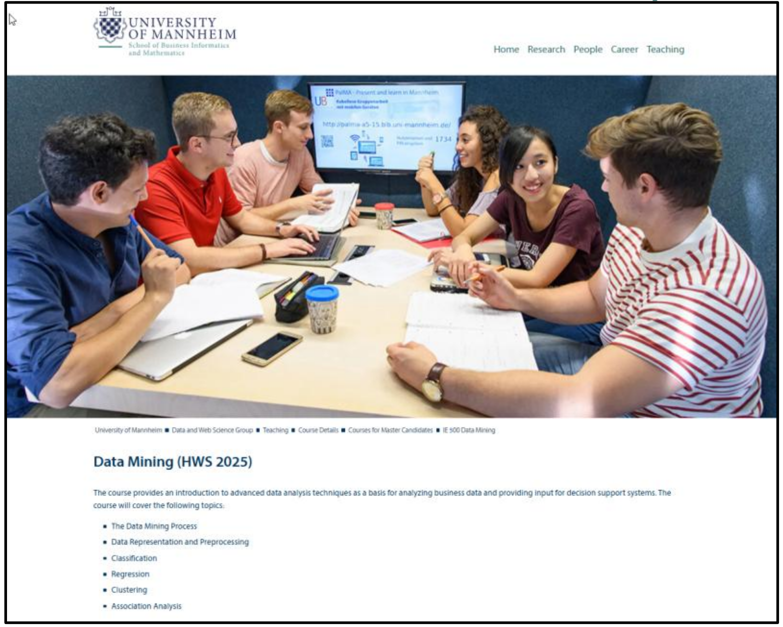
- Dr. Rita Torres de Sousa
- Researcher
- Research Interests:
  - Knowledge graphs
  - Machine learning
  - Biomedical applications
- Room: B6 26, B0.01
- eMail: rita.sousa@uni-mannheim.de
- Webpage: ritatsousa.github.io
- Rita will teach the exercises

Hello

- 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 - Material

- 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, video lectures



University of Mannheim ■ Data and Web Science Group ■ Teaching ■ Course Details ■ Courses for Master Candidates ■ IE 500 Data Mining

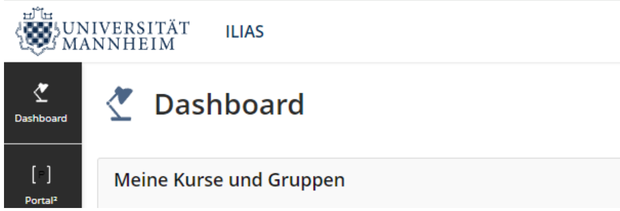
Data Mining (HWS 2025)

The course provides an introduction to advanced data analysis techniques as a basis for analyzing business data and providing input for decision support systems. The course will cover the following topics:

- The Data Mining Process
- Data Representation and Preprocessing
- Classification
- Regression
- Clustering
- Association Analysis

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

ILIAS®



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 principle methods of data mining
  - Discusses how to evaluate the learned models
  - Presents practical examples of data mining applications

  - Time: Monday, 13:45 – 15:15
  - Location: Room A 001 Building A 5,6 Part A

Course Organization - Material

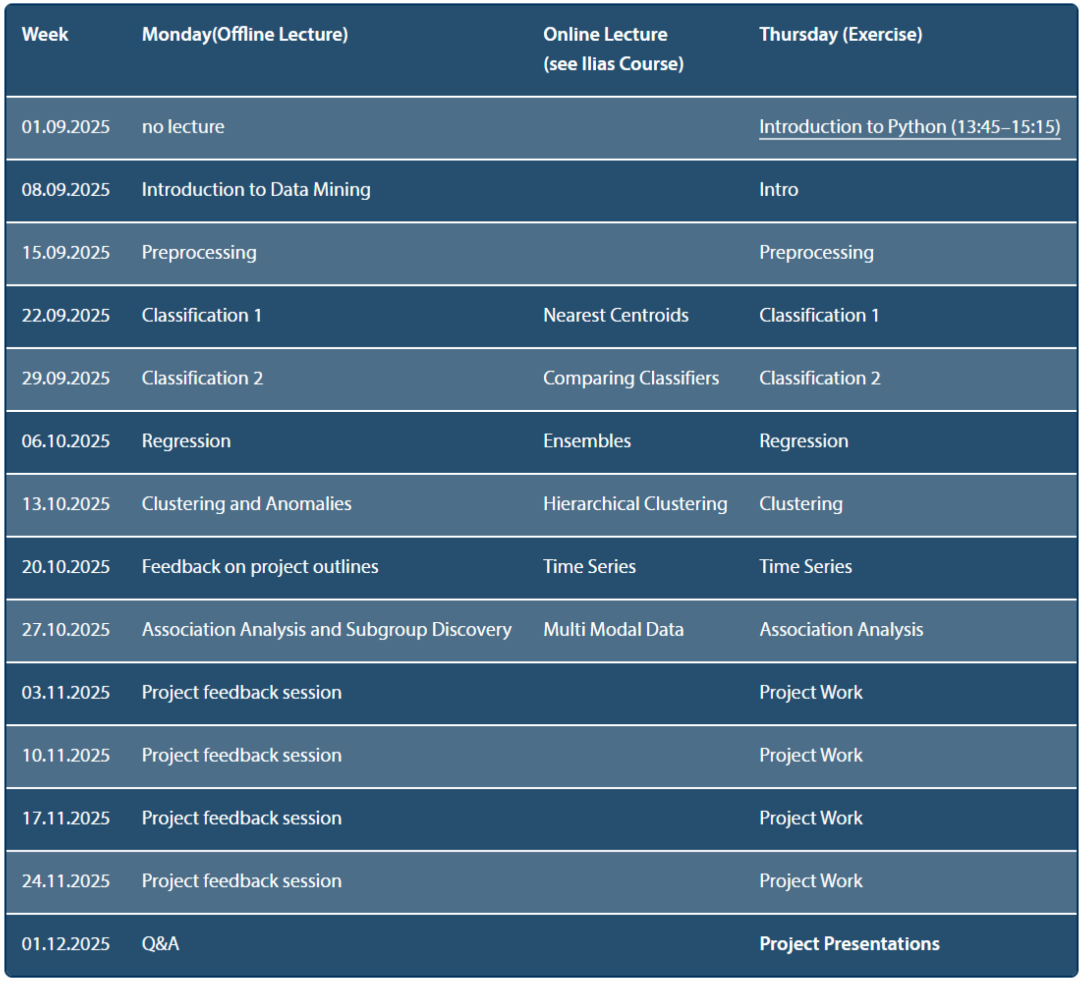
- Online Lecture
  - Part of the course
  - Exercise / **Exam** relevant
  - All Slides and Videos are already available

Classification – kNN vs. Nearest Centroid

IE500 Data Mining

Prof. Dr. Heiko Paulheim

Video recording: November 16, 2024



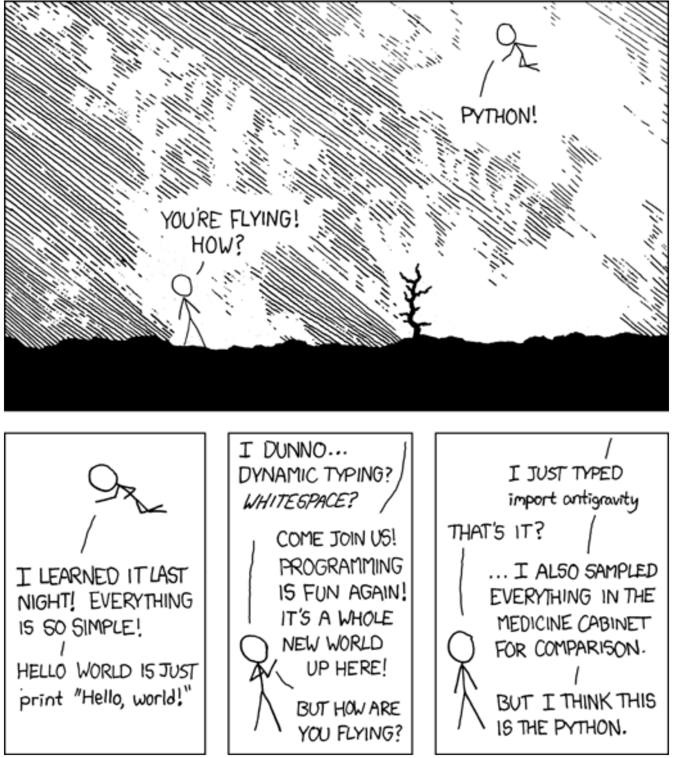
Week	Monday(Offline Lecture)	Online Lecture (see Ilias Course)	Thursday (Exercise)
01.09.2025	no lecture		Introduction to Python (13:45–15:15)
08.09.2025	Introduction to Data Mining		Intro
15.09.2025	Preprocessing		Preprocessing
22.09.2025	Classification 1	Nearest Centroids	Classification 1
29.09.2025	Classification 2	Comparing Classifiers	Classification 2
06.10.2025	Regression	Ensembles	Regression
13.10.2025	Clustering and Anomalies	Hierarchical Clustering	Clustering
20.10.2025	Feedback on project outlines	Time Series	Time Series
27.10.2025	Association Analysis and Subgroup Discovery	Multi Modal Data	Association Analysis
03.11.2025	Project feedback session		Project Work
10.11.2025	Project feedback session		Project Work
17.11.2025	Project feedback session		Project Work
24.11.2025	Project feedback session		Project Work
01.12.2025	Q&A		**Project Presentations**

Course Organization - Exercise

- Exercise Groups
  - Students experiment with data sets using Python
  - Theoretical tasks (similar to exam)
- Time and Location (same content - only attend **one**):
  - Thursday, 12.00 – 13.30, A104 Building B6, 26 Part A (Rita/Franz)
  - Thursday, 13.45 – 15.15, A104 Building B6, 26 Part A (Rita/Franz)
  - Thursday, 15.30 – 17.00, A104 Building B6, 26 Part A (Rita/Franz)
  - You can also switch between weeks if needed

Introduction to Python

- 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


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

Usage of LLMs like ChatGPT

- 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

Course Organization - Project

- 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, October, 5th, 23:59**
  - Submission of project proposal
    - **Tuesday, October, 14th, 23:59**
  - Submission of final project work report
    - **Sunday, November 30th, 23:59**
  - Submission of Presentation (PDF)
    - **Wednesday, December 3rd, 23:59**

Course Organization - Exam

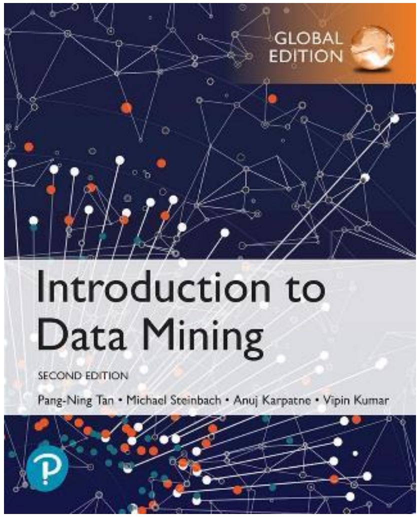
- Date and Time: **Monday, 15th December 2025**
- Duration: 60 minutes
- Structure: 6 open questions that
  - Check whether you have understood the lecture content
    - We try to cover all major chapters of the lecture
    - Require you to describe the ideas behind algorithms and methods
    - Often: 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
  - There will be at most 1 question containing Python content
    - Should be solvable without a lot of Python knowledge
    - You do not need to know specialized Python functions by heart

Course Organization - Exam

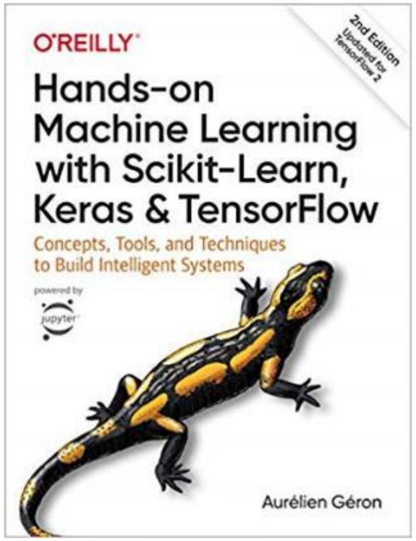
- There is only one exam per semester
  - Because course is offered every semester
    - The next exam date is at the end of the upcoming FSS
  - i.e., no retake date!
- Upon failure, you will have to redo both the project and the exam in another semester
  - Unfortunately, we cannot carry over your project mark
- **Final grade**
  - 75 % written exam, 20% project report, 5% project presentation

Textbooks for the Course

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


GLOBAL EDITION
Introduction to Data Mining
SECOND EDITION
Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar

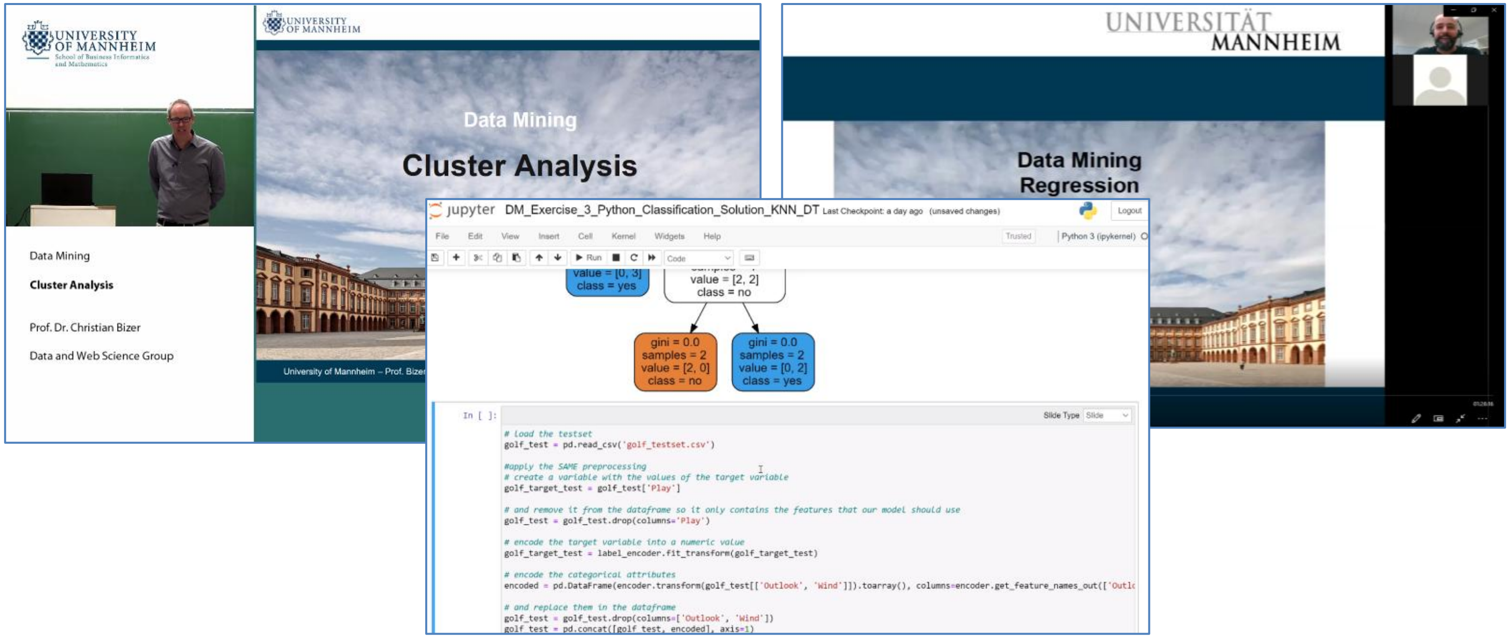
- Aurélien Géron: **Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.** 2nd or 3rd Edition, O’Reilly, 2019 or 2022


O'REILLY
2nd Edition Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by jupyter
Aurélien Géron

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

Videos and Screencasts

- **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

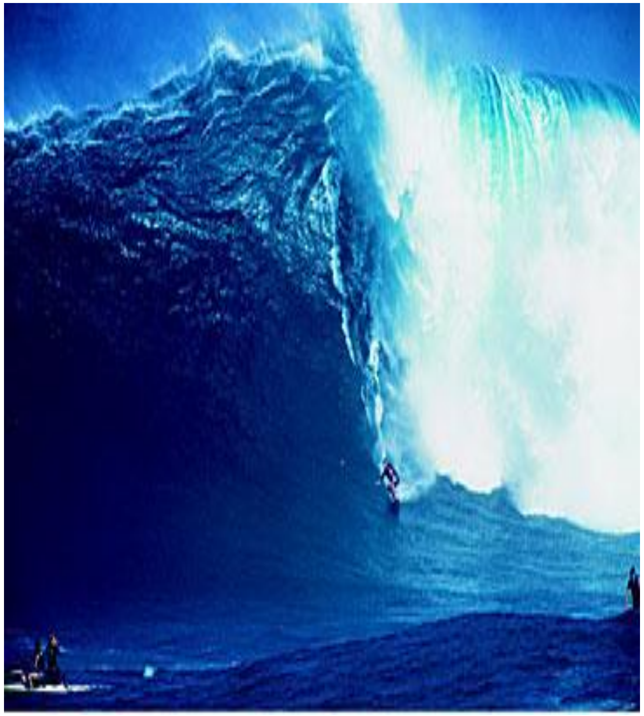


```
# load the testset
golf_test = pd.read_csv('golf_testset.csv')

#apply the SAME preprocessing
# create a variable with the values of the target variable
golf_target_test = golf_test['Play']

# and remove it from the dataframe so it only contains the features that our model should use
golf_test = golf_test.drop(columns='Play')

# encode the target variable into a numeric value
golf_target_test = label_encoder.fit_transform(golf_target_test)

# encode the categorical attributes
encoded = pd.DataFrame(encoder.transform(golf_test[['Outlook', 'Wind']]).toarray(), columns=encoder.get_feature_names_out(['Outlo

# and replace them in the dataframe
golf_test = golf_test.drop(columns=['Outlook', 'Wind'])
golf_test = pd.concat([golf_test, encoded], axis=1)
```

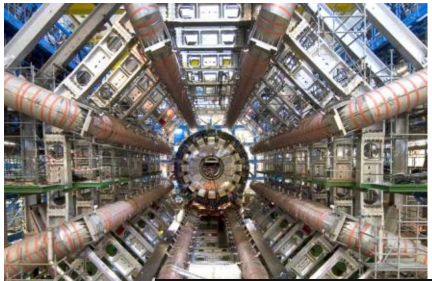
Questions?

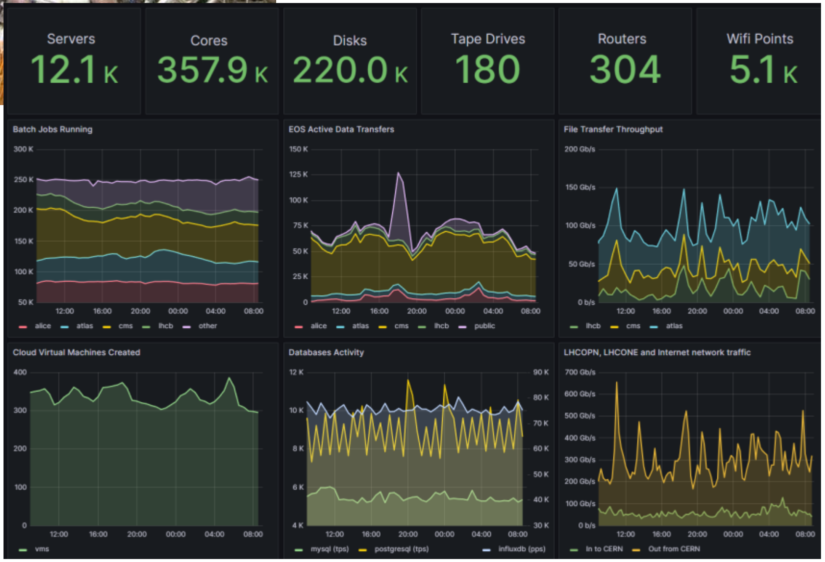
What is Data Mining?

- **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...”

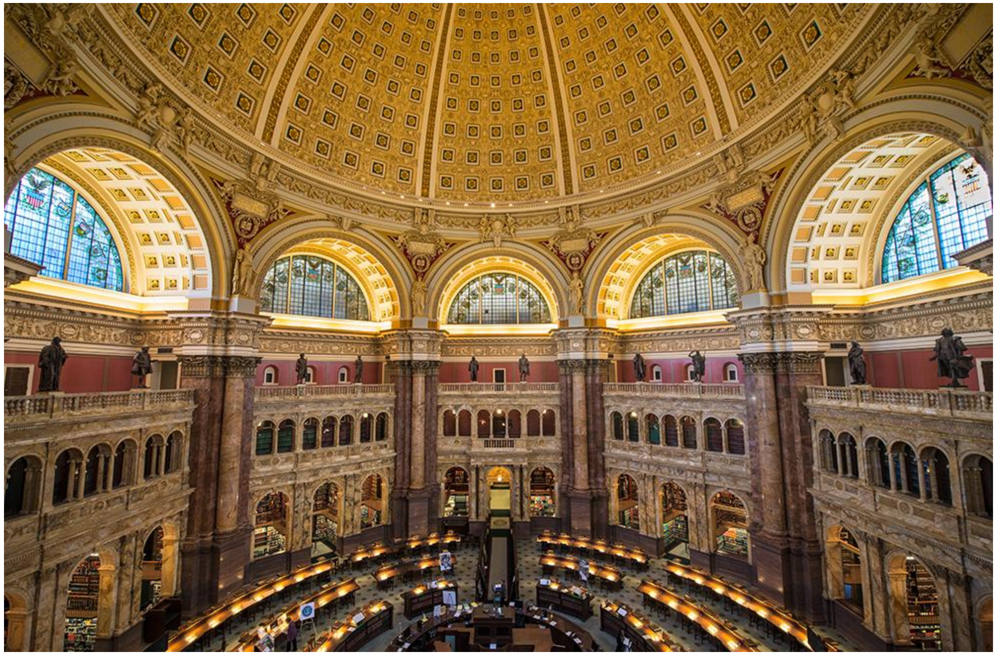
- **Facebook**
  - 4 Petabyte of new data generated every day
  - over 300 Petabyte in Facebook‘s data warehouse

- 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...”



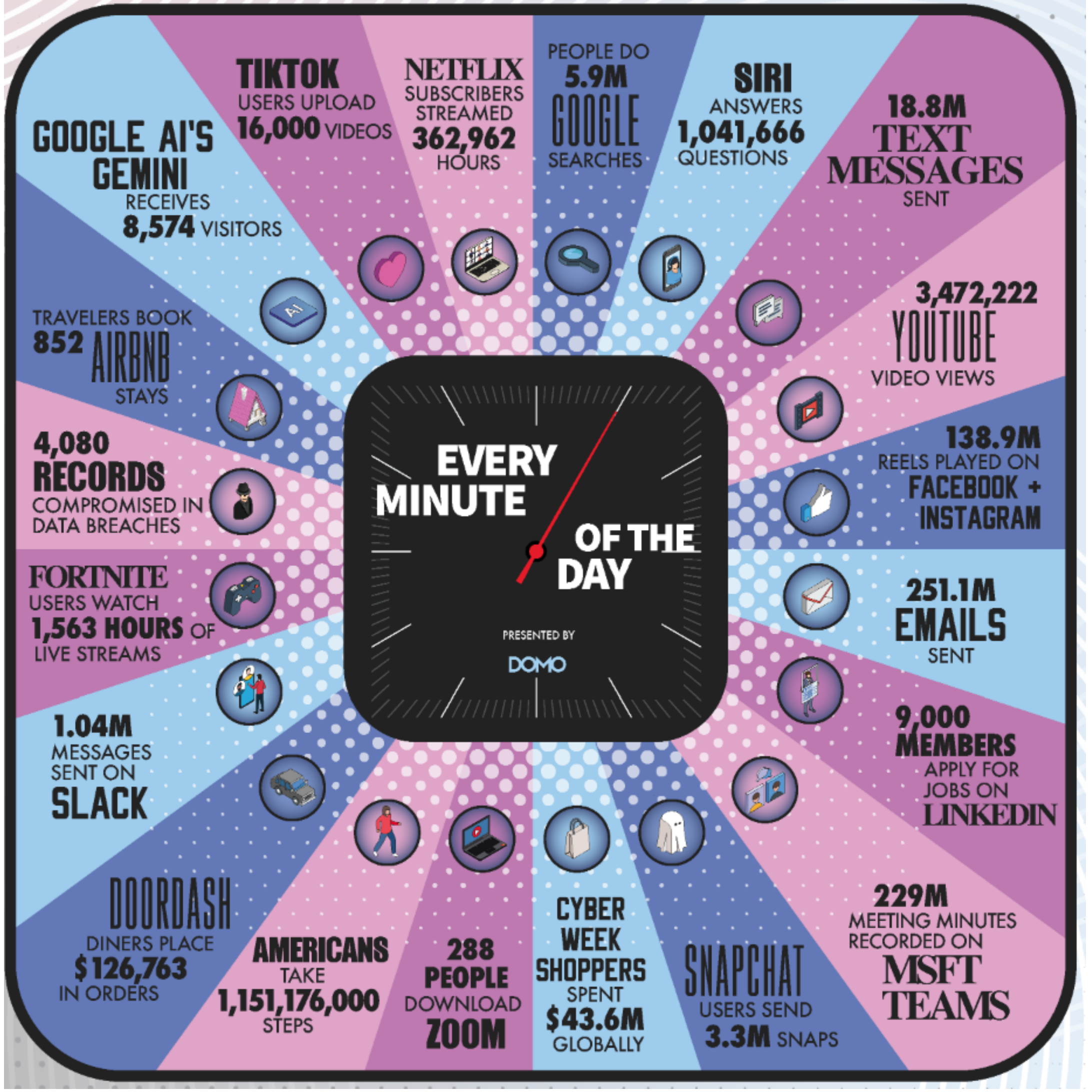
- **US Library of Congress**
  - ≈ 235 TB archived
- Discover
  - Topic distributions*
  - Citation networks
- Train
  - Large Language Models

https://www.brandwatch.com/blog/facebook-statistics/

http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

“We are Drowning in Data…”

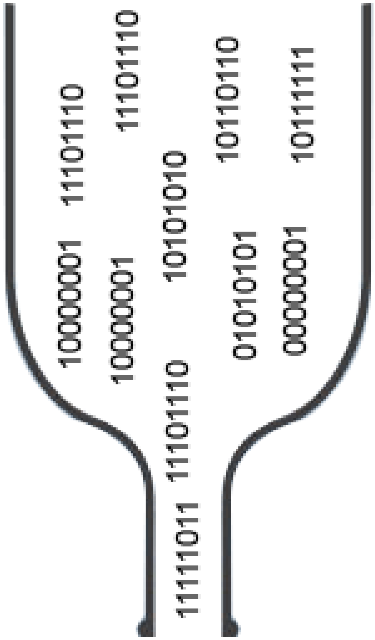
UNIVERSITY OF MANNHEIM
Data and Web Science Group



EVERY MINUTE OF THE DAY
PRESENTED BY DOMO
GOOGLE AI'S GEMINI RECEIVES 8,574 VISITORS
TIKTOK USERS UPLOAD 16,000 VIDEOS
NETFLIX SUBSCRIBERS STREAMED 362,962 HOURS
PEOPLE DO 5.9M GOOGLE SEARCHES
SIRI ANSWERS 1,041,666 QUESTIONS
18.8M TEXT MESSAGES SENT
3,472,222 YOUTUBE VIDEO VIEWS
138.9M REELS PLAYED ON FACEBOOK + INSTAGRAM
251.1M EMAILS SENT
9,000 MEMBERS APPLY FOR JOBS ON LINKEDIN
229M MEETING MINUTES RECORDED ON MSFT TEAMS
SNAPCHAT USERS SEND 3.3M SNAPS
CYBER WEEK SHOPPERS SPENT $43.6M GLOBALLY
288 PEOPLE DOWNLOAD ZOOM
AMERICANS TAKE 1,151,176,000 STEPS
DOORDASH DINERS PLACE $126,763 IN ORDERS
1.04M MESSAGES SENT ON SLACK
FORTNITE USERS WATCH 1,563 HOURS OF LIVE STREAMS
4,080 RECORDS COMPROMISED IN DATA BREACHES
TRAVELERS BOOK 852 AIRBNB STAYS
AI

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

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

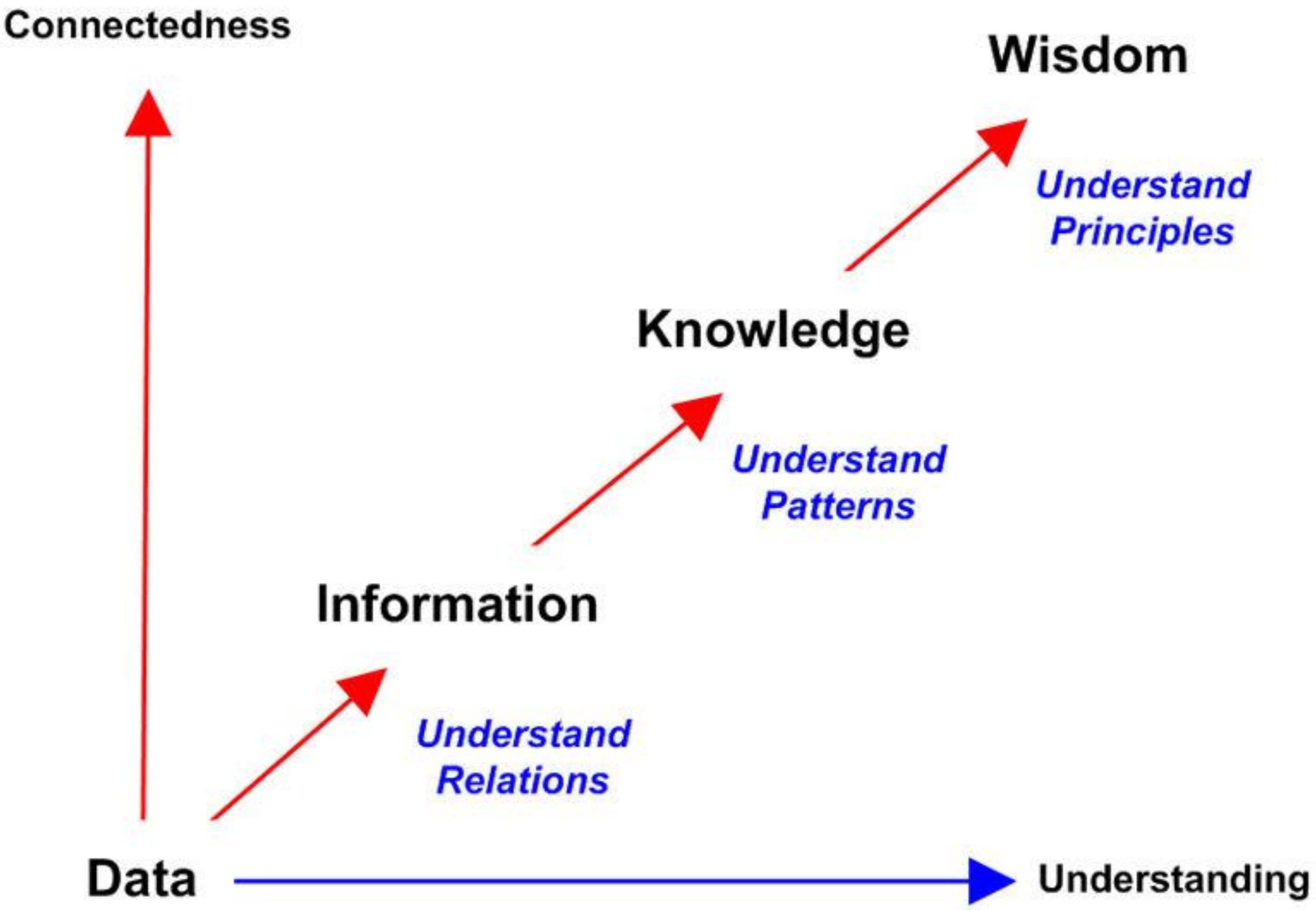


← Rate at which data are produced

← 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

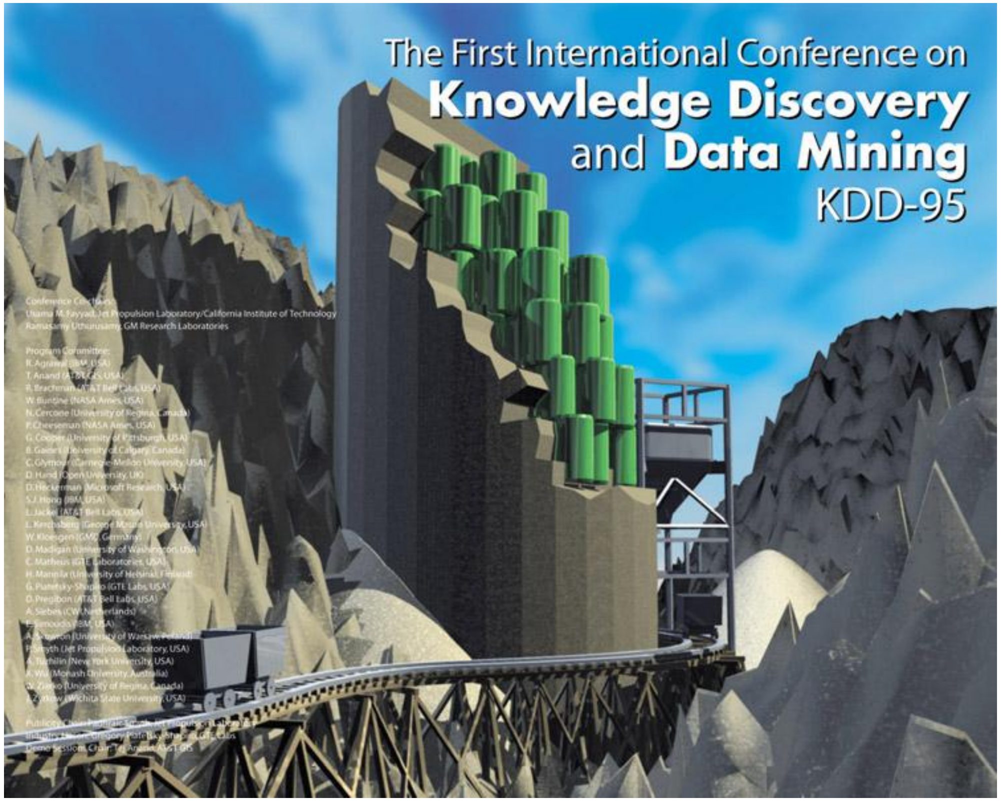
Data, Information, Knowledge, Wisdom



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

Data Mining: Definitions

- Idea: mountains of data
  - Where knowledge is mined


The First International Conference on
Knowledge Discovery
and Data Mining
KDD-95
Conference Co-chairs:
Usama M. Fayyad, Jet Propulsion Laboratory/California Institute of Technology
Ramasamy Uthurusamy, GM Research Laboratories
Program Committee:
R. Agrawal (IBM, USA)
T. Anand (AT&T GIS, USA)
R. Brachman (AT&T Bell Labs, USA)
W. Buntine (NASA Ames, USA)
N. Cercone (University of Regina, Canada)
P. Cheeseman (NASA Ames, USA)
G. Cooper (University of Pittsburgh, USA)
B. Gaines (University of Calgary, Canada)
C. Glymour (Carnegie-Mellon University, USA)
D. Hand (Open University, UK)
D. Heckerman (Microsoft Research, USA)
S.J. Hong (IBM, USA)
L. Jackel (AT&T Bell Labs, USA)
L. Kerschberg (George Mason University, USA)
W. Kloesgen (GMD, Germany)
D. Madigan (University of Washington, USA)
C. Matheus (GTE Laboratories, USA)
H. Mannila (University of Helsinki, Finland)
G. Piatetsky-Shapiro (GTE Labs, USA)
D. Pregibon (AT&T Bell Labs, USA)
A. Siebes (CWI, Netherlands)
E. Simoudis (IBM, USA)
A. Skowron (University of Warsaw, Poland)
P. Smyth (Jet Propulsion Laboratory, USA)
A. Tuzhilin (New York University, USA)
X. Wu (Monash University, Australia)
W. Ziarko (University of Regina, Canada)
J. Zytkow (Wichita State University, USA)
Publicity Chair: Padhraic Smyth, Jet Propulsion Laboratory
Industry Liaison: Gregory Piatetsky-Shapiro, GTE Labs
Demo Sessions Chair: Tej Anand, AT&T GIS

Data Mining: Definitions

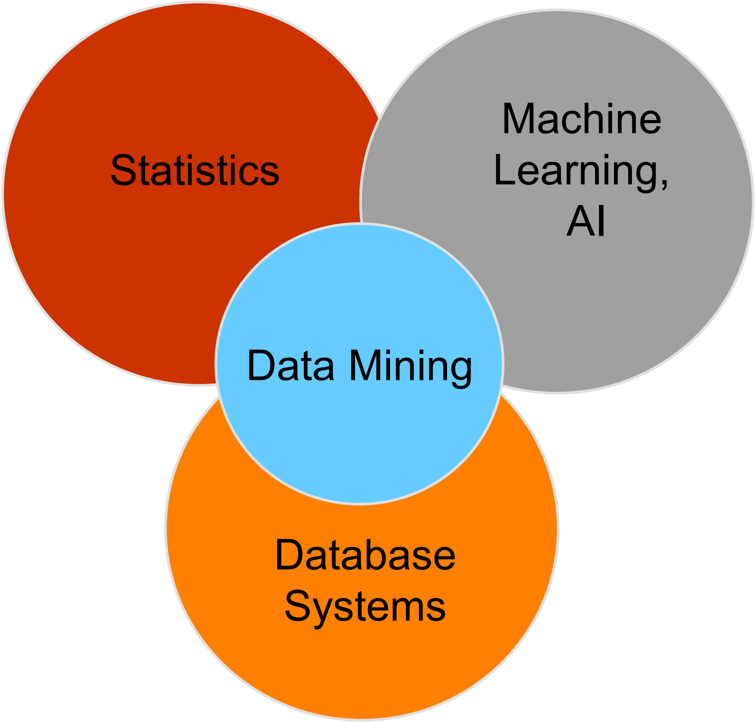
- Data Mining is a non-trivial process of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable

  patterns in data. (Fayyad et al. 1996)

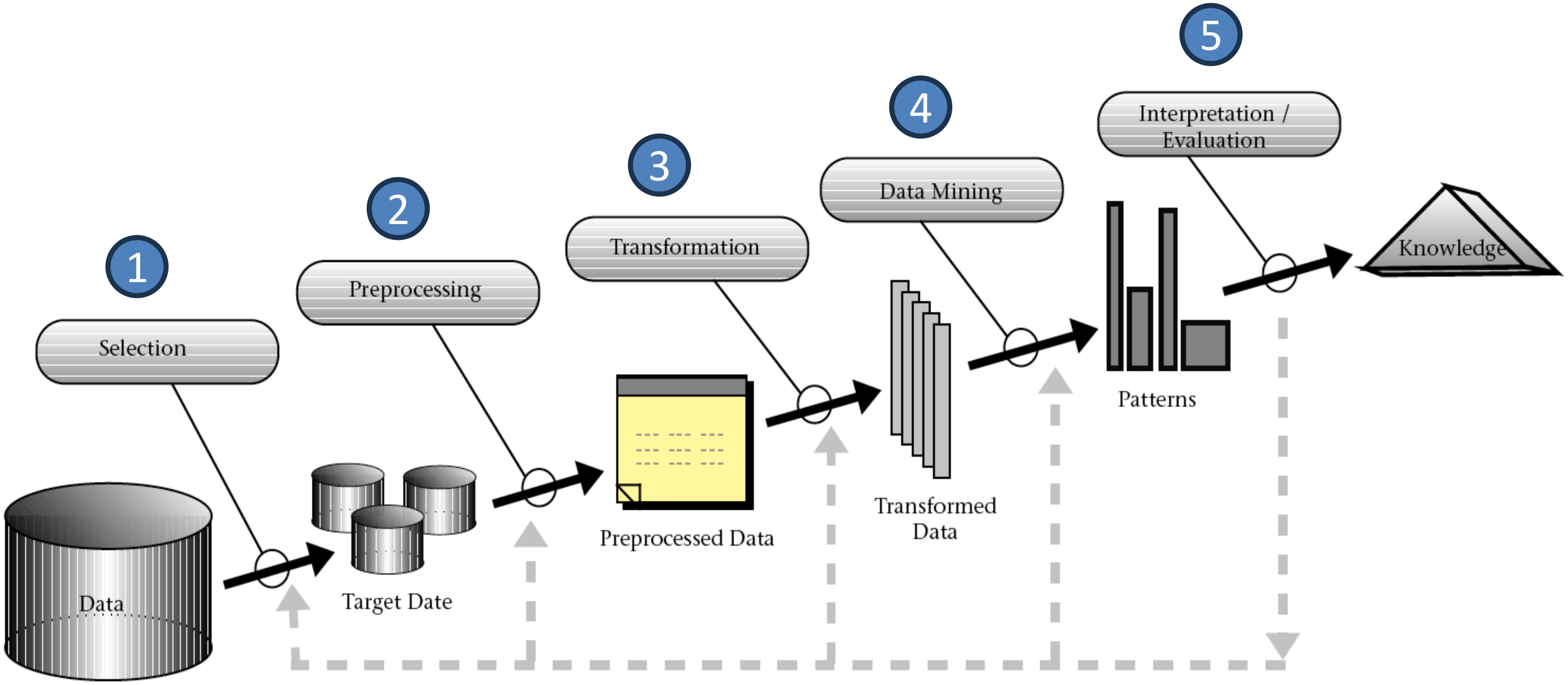
- Data Mining methods
  1. Detect interesting patterns in large quantities of data
  2. Support human decision making by providing such patterns
  3. Predict the outcome of a future observation based on the patterns

Origins of Data Mining

- 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



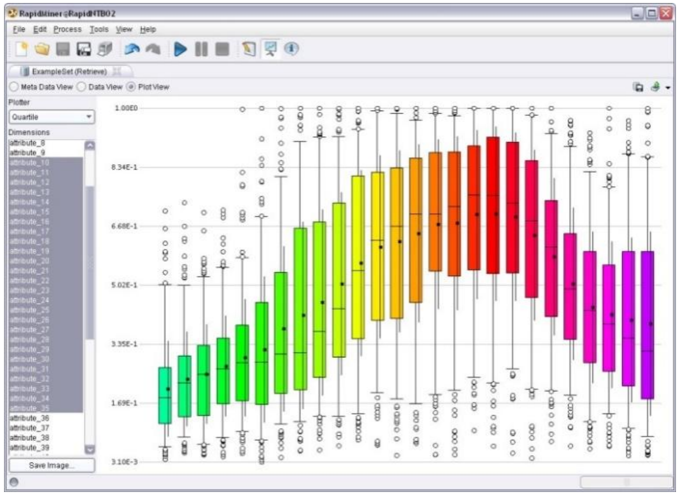
The Data Mining Process

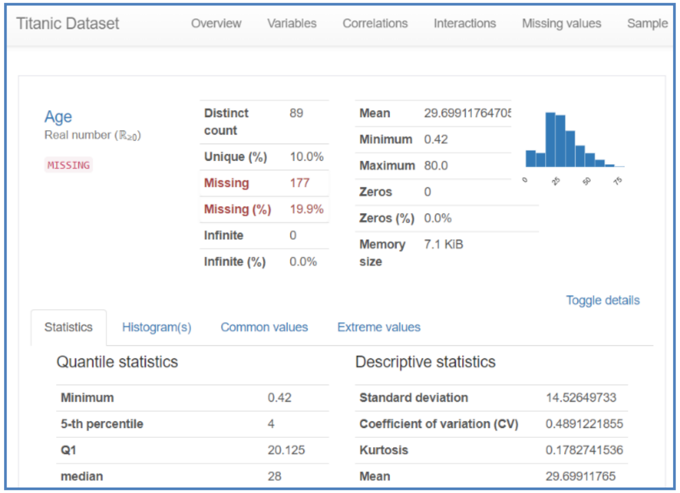


Source: Fayyad et al. (1996)

Selection and Exploration (1)

- 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





Titanic Dataset   Overview   Variables   Correlations   Interactions   Missing values   Sample

Age
Real number (ℝ≥0)
MISSING

Distinct count	89
Unique (%)	10.0%
Missing	177
Missing (%)	19.9%
Infinite	0
Infinite (%)	0.0%

Mean	29.6991176470
Minimum	0.42
Maximum	80.0
Zeros	0
Zeros (%)	0.0%
Memory size	7.1 KiB

Toggle details

Statistics   Histogram(s)   Common values   Extreme values

Quantile statistics

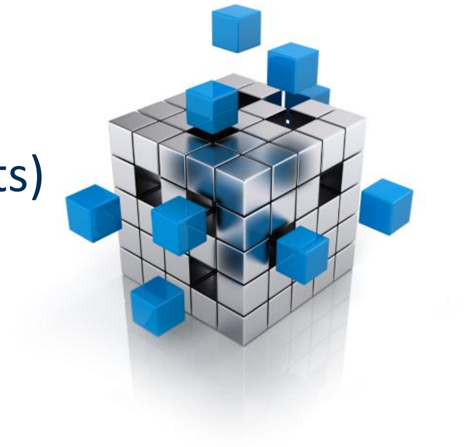
Minimum	0.42
5-th percentile	4
Q1	20.125
median	28

Descriptive statistics

Standard deviation	14.52649733
Coefficient of variation (CV)	0.4891221855
Kurtosis	0.1782741536
Mean	29.69911765

Preprocessing and Transformation (2+3)

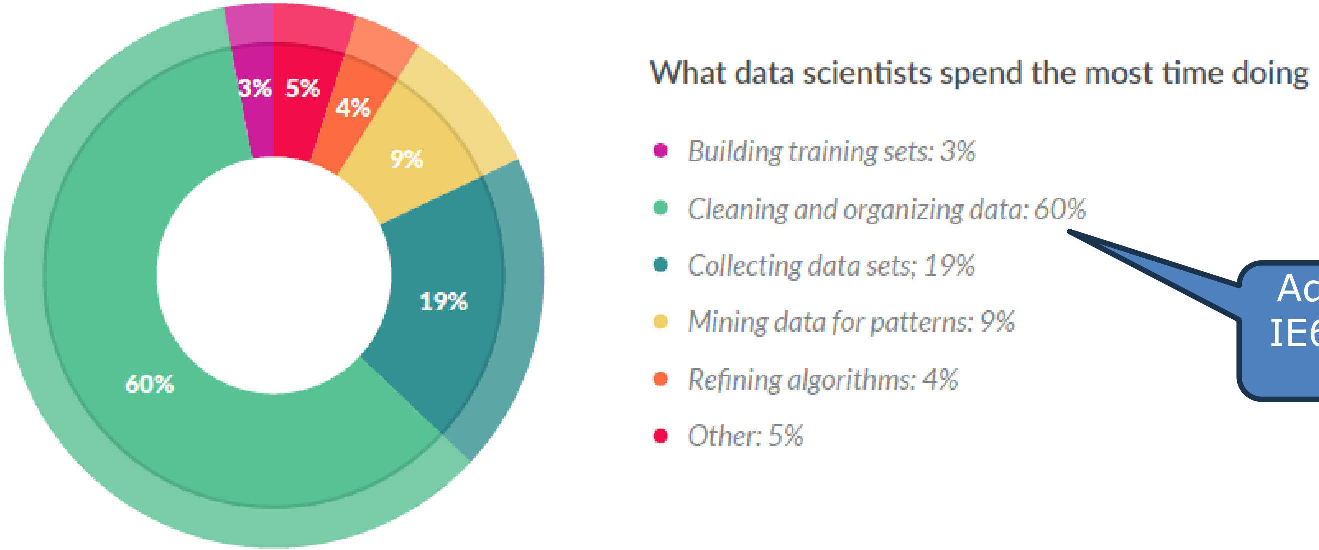
- Transform data into a representation that is suitable for the chosen data mining methods
  - Number of dimensions (represent relevant information using less attributes)
  - Scales of attributes (nominal, ordinal, numeric)
  - Amount of data (determines hardware requirements)



- 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)

- 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



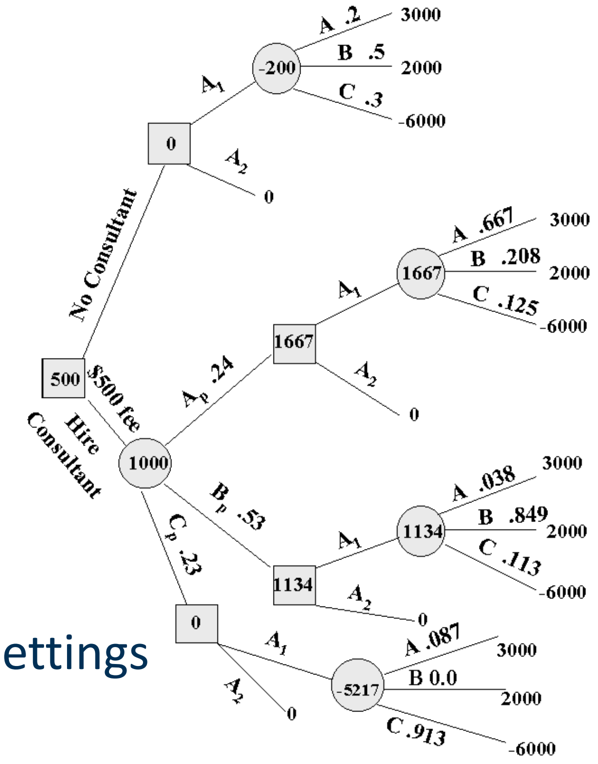
Advertisement: IE670 Web Data Integration

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

Data Mining (4)

- 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)

- Output of Data Mining
  - Patterns
  - Models

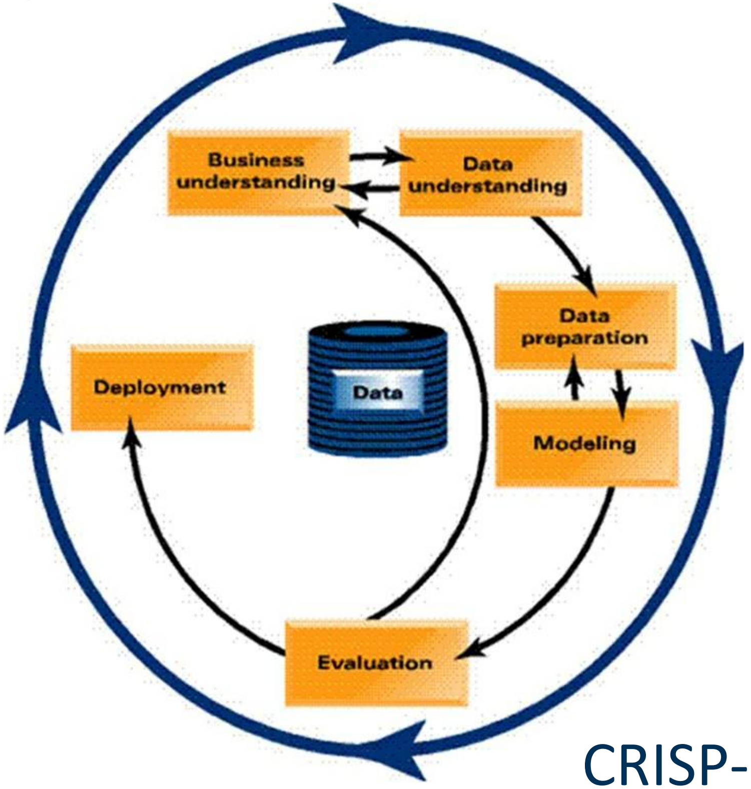
- In the end, we want to derive value from that, e.g.,
  - Gain knowledge

  - Make better decisions

  - Increase revenue

Deployment

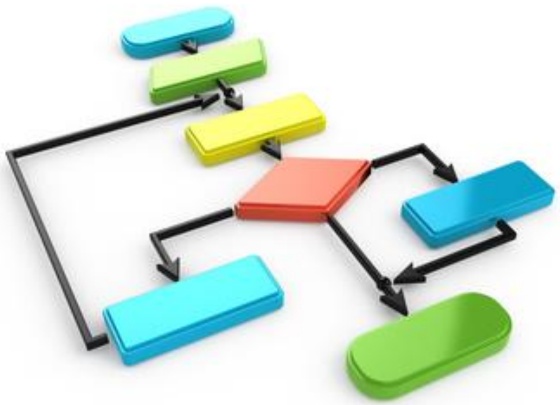
- Use model in the business context
- Keep iterating in order to maintain and improve model



CRISP-DM Process Model

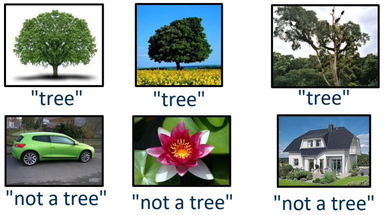
Tasks and Applications

- **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

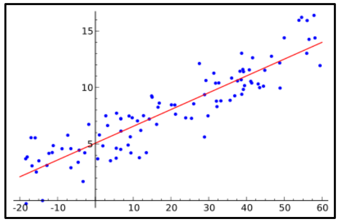


Data Mining Tasks

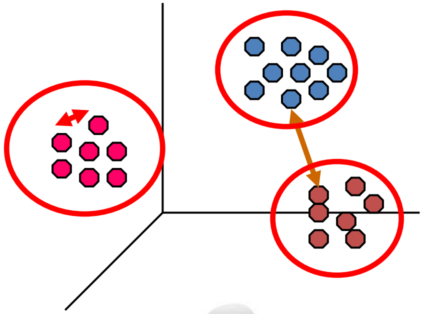
- Classification [Predictive]



- Regression [Predictive]



- Cluster Analysis [Descriptive]



- Association Analysis [Descriptive]

Classification

"tree"

"tree"

"tree"

"not a tree"

"not a tree"

"not a tree"

- 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"

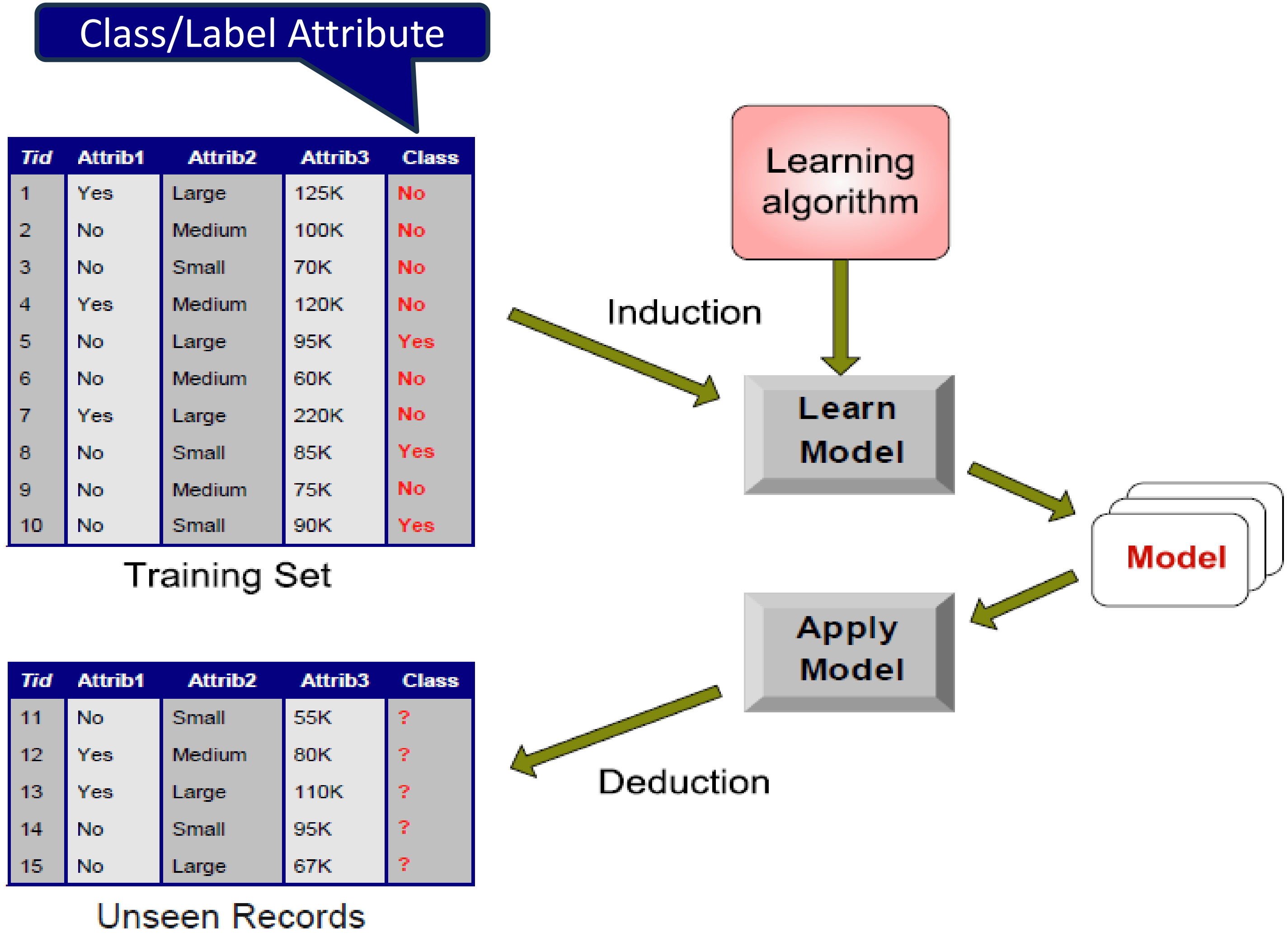
"tree"

"not a tree"

"not a tree"

"not a tree"

Classification: Workflow



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Unseen Records

Classification: Applications

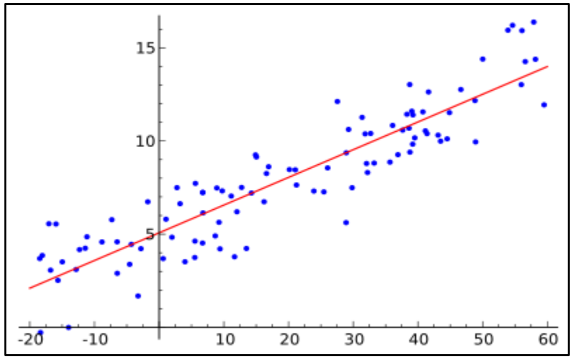
- Credit Risk Assessment
  - Attributes: your age, income, debts, …
  - Class: are you getting credit by your bank?

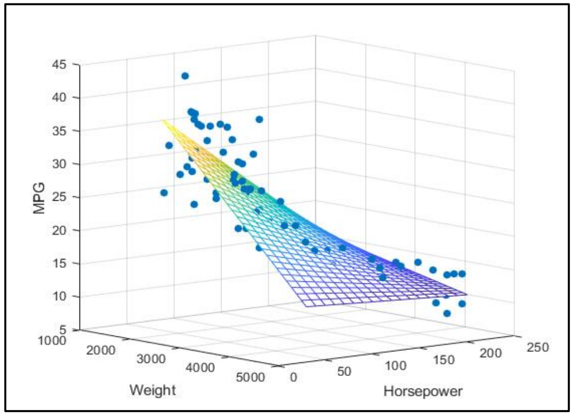
- SPAM Detection
  - Attributes: words and header fields of an e-mail
  - Class: regular e-mail or spam e-mail?

- 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.

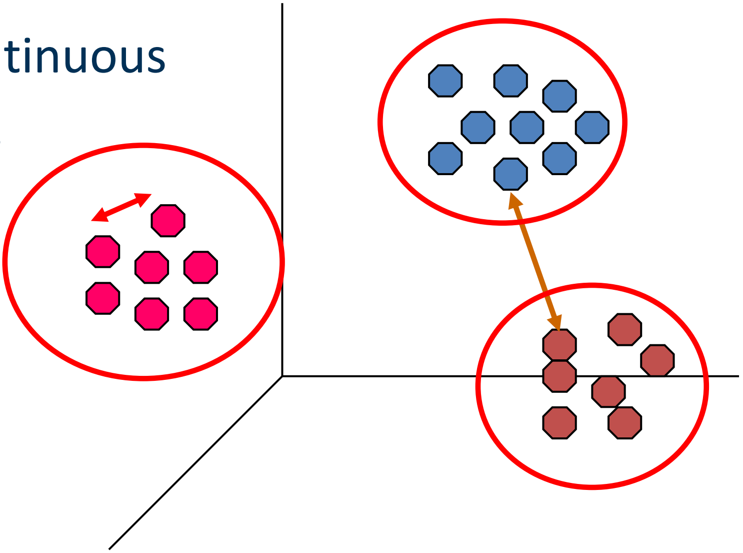




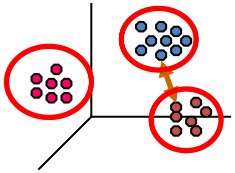
- Difference to classification: The predicted attribute is **continuous**, while classification is used to predict nominal attributes (e.g. yes/no)

Cluster Analysis

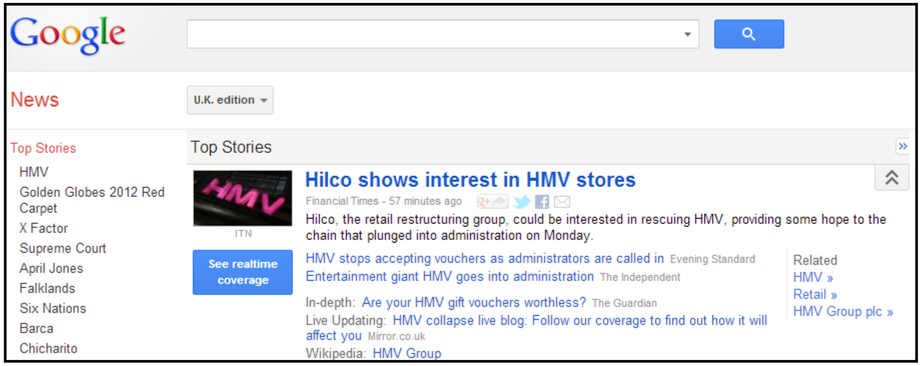
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find groups such that
  - Data points in one group are more similar to one another
  - Data points in separate groups are less similar to one another
- Similarity Measures
  - Euclidean distance if attributes 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



- 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



Top Stories
HMV
Golden Globes 2012 Red Carpet
X Factor
Supreme Court
April Jones
Falklands
Six Nations
Barca
Chicharito

Top Stories

Hilco shows interest in HMV stores
Financial Times - 57 minutes ago
Hilco, the retail restructuring group, could be interested in rescuing HMV, providing some hope to the chain that plunged into administration on Monday.

HMV stops accepting vouchers as administrators are called in Evening Standard
Entertainment giant HMV goes into administration The Independent

In-depth: Are your HMV gift vouchers worthless? The Guardian
Live Updating: HMV collapse live blog: Follow our coverage to find out how it will affect you Mirror.co.uk
Wikipedia: HMV Group

Related
HMV »
Retail »
HMV Group plc »

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}

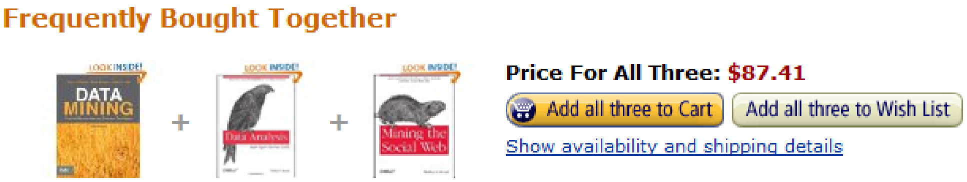
Association Rules
{Diaper, Milk} --> {Beer}
{Milk} --> {Coke}

Association Analysis: Applications

- 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

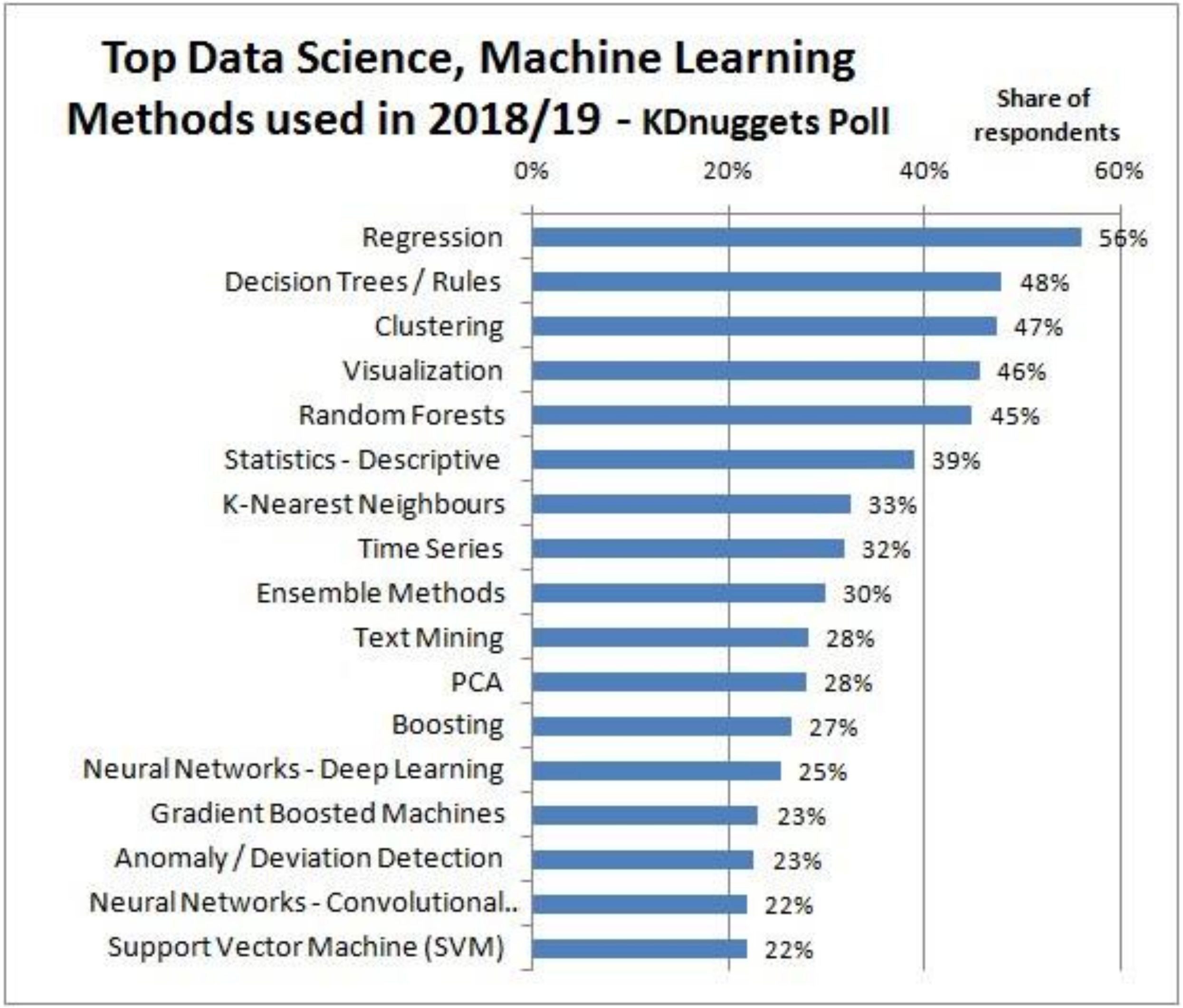


Frequently Bought Together

Price For All Three: $87.41

Show availability and shipping details

Which Methods are Used in Practice?



Classification Algorithms

"tree"

"tree"

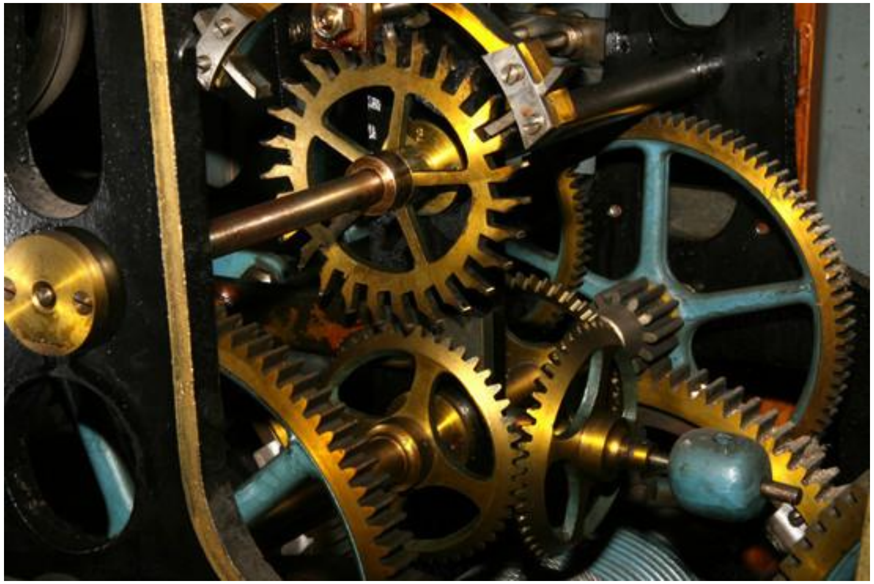
"tree"

"not a tree"

"not a tree"

"not a tree"

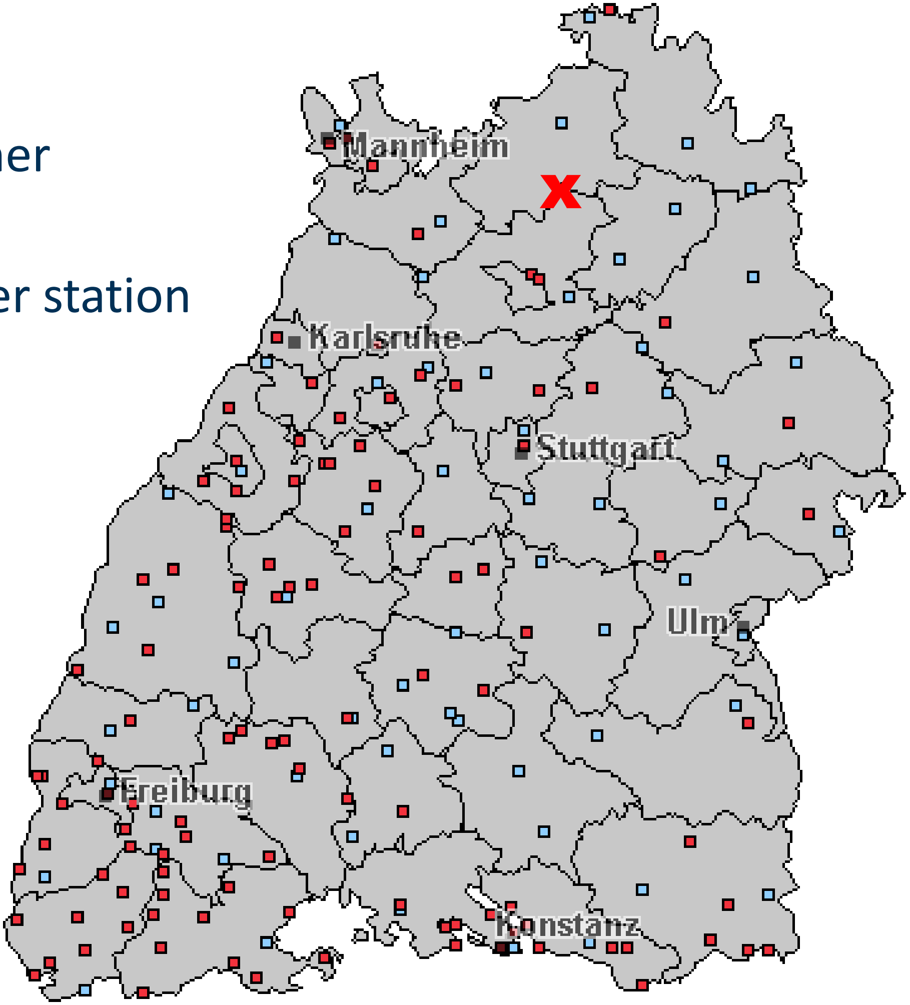
- Classification:
  - We give the computer a set of labeled examples
  - The computer learns to classify new (unlabeled) examples
- How does that work?
  - **K-Nearest-Neighbors**
  - Decision Trees
  - Naïve Bayes
  - Support Vector Machines
  - Artificial Neural Networks
  - Deep Neural Networks
 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



K-Nearest-Neighbors

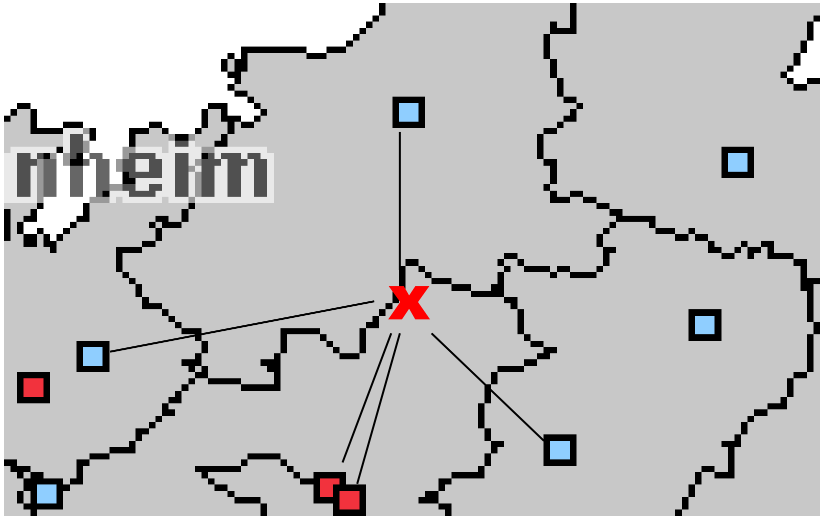
- Problem
  - Predict the current weather in a certain place
  - Where there is no weather station
  - How could you do that?

- Symbols
  - Red = Sunny
  - Blue = Cloudy



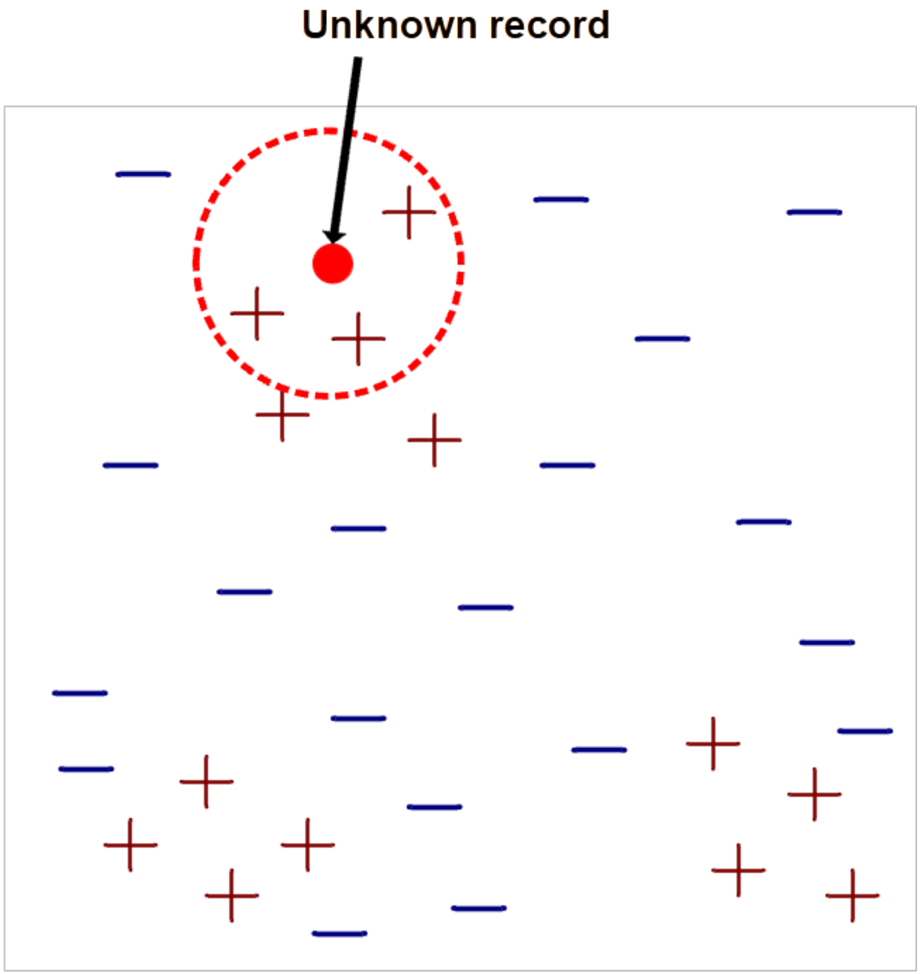
K-Nearest-Neighbors

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



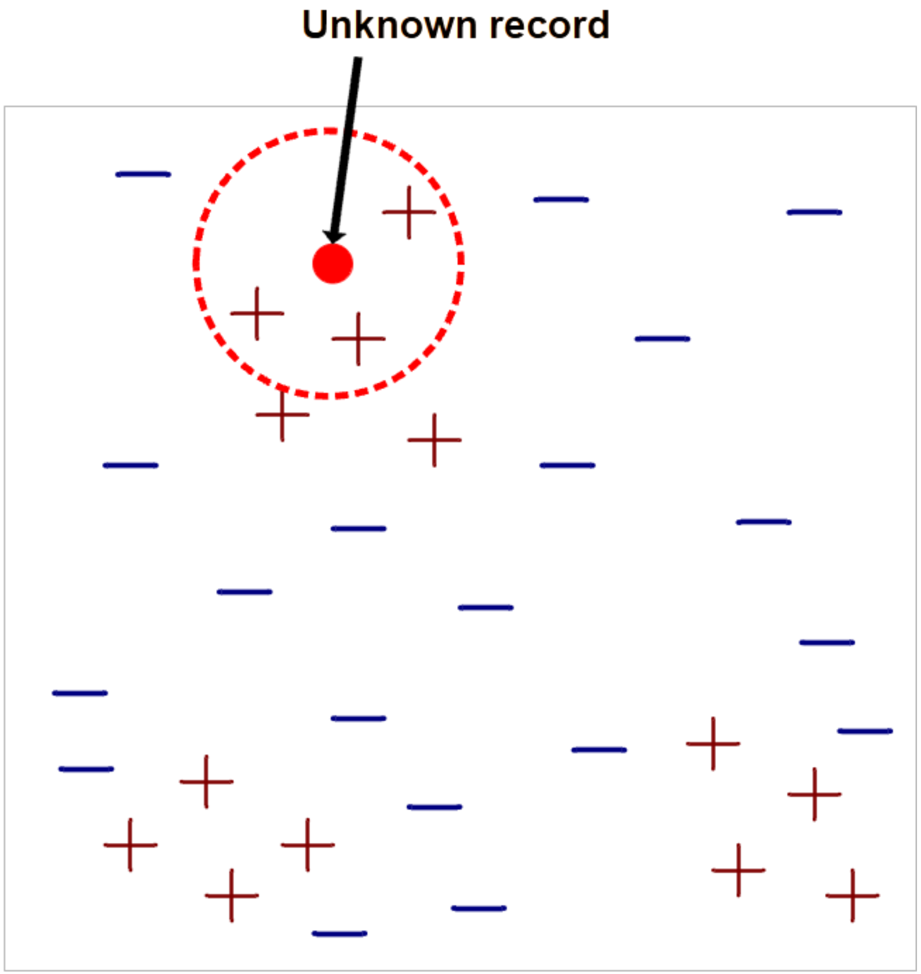
- This approach is called **K-Nearest-Neighbors**
  - where k is the number of neighbors to consider
  - in the example:
    - k=5
    - “near” denotes geographical proximity

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

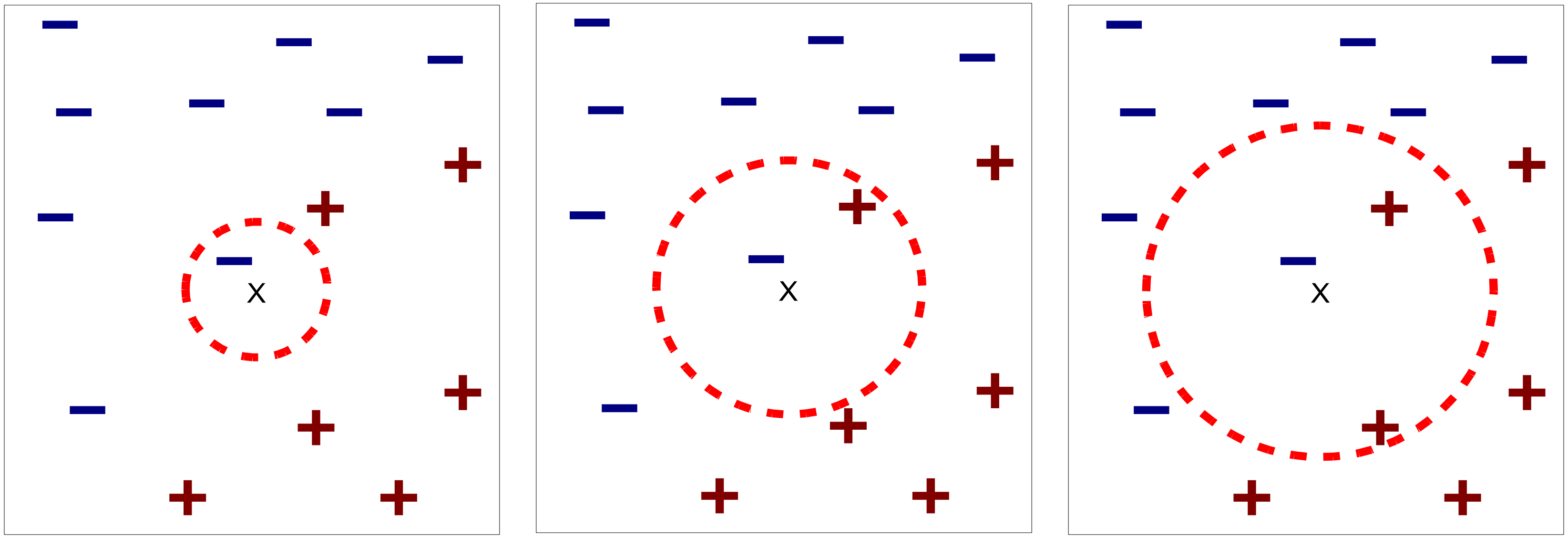
K-Nearest-Neighbor Classifier



- 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

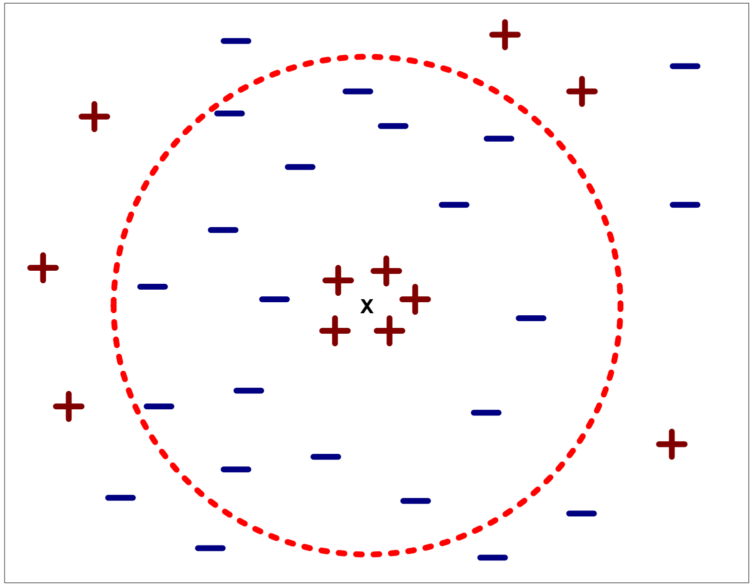
- The k-nearest neighbors of a record x are data points that have the k smallest distances to x



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

Choosing a Good Value for K

- 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



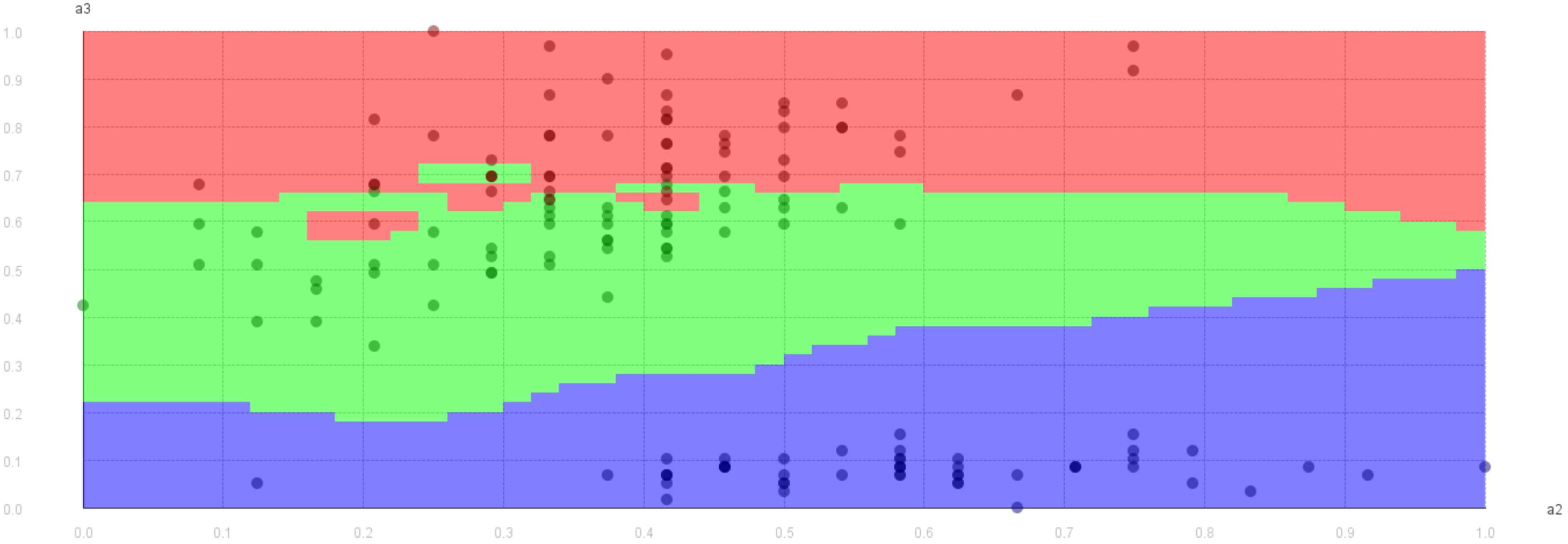
- Rule of thumb: Test k values between 1 and 20

Discussion of K-NN Classification

- **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

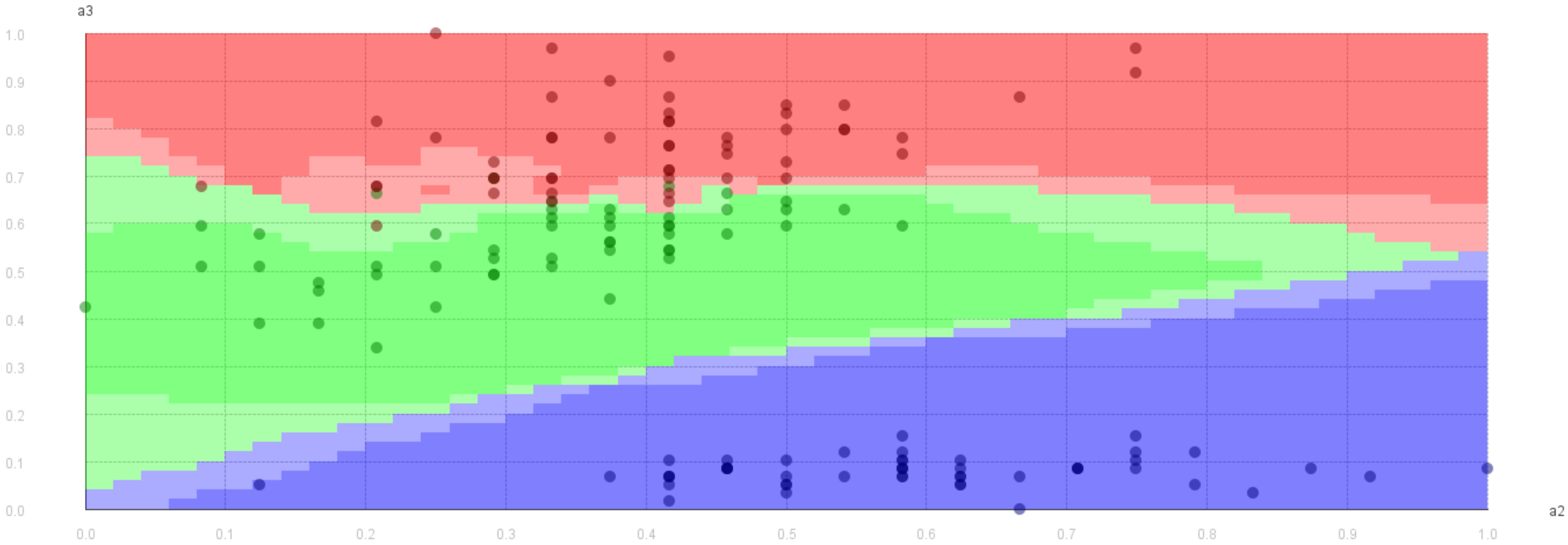
- k=1
- Single noise points have influence on model



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



What You Will Learn in This Lecture

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

Thank you

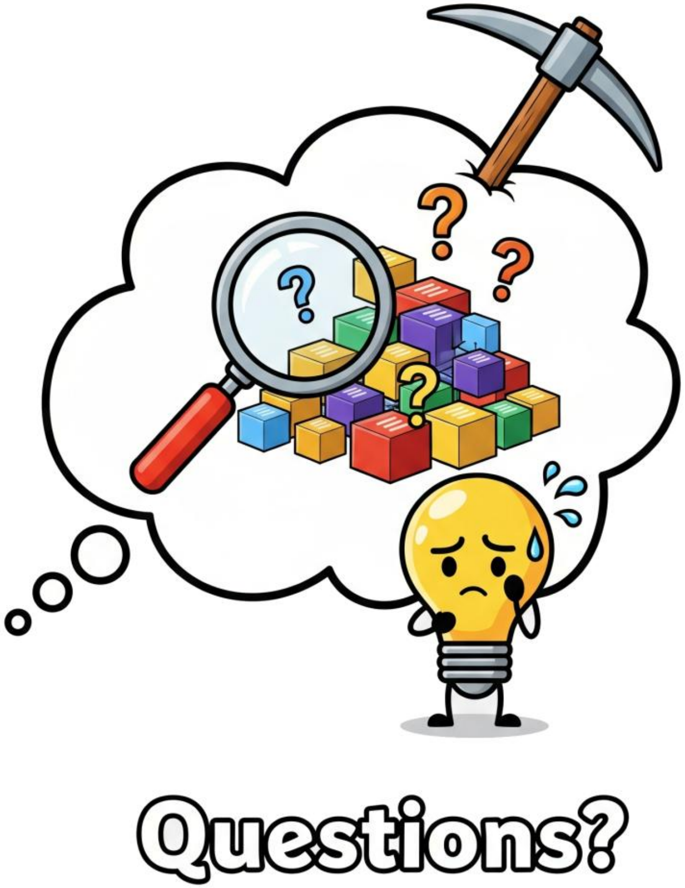


Image generated by Gemini