

Introduction

IE500 Data Mining



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. 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 the exercise groups
and will supervise student projects



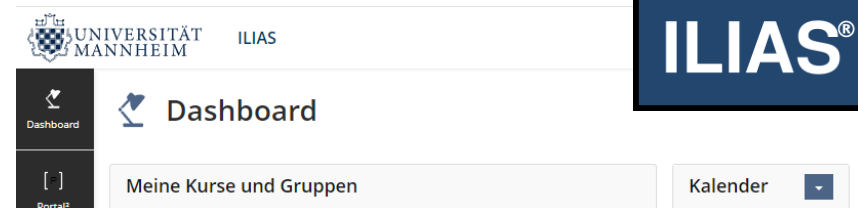
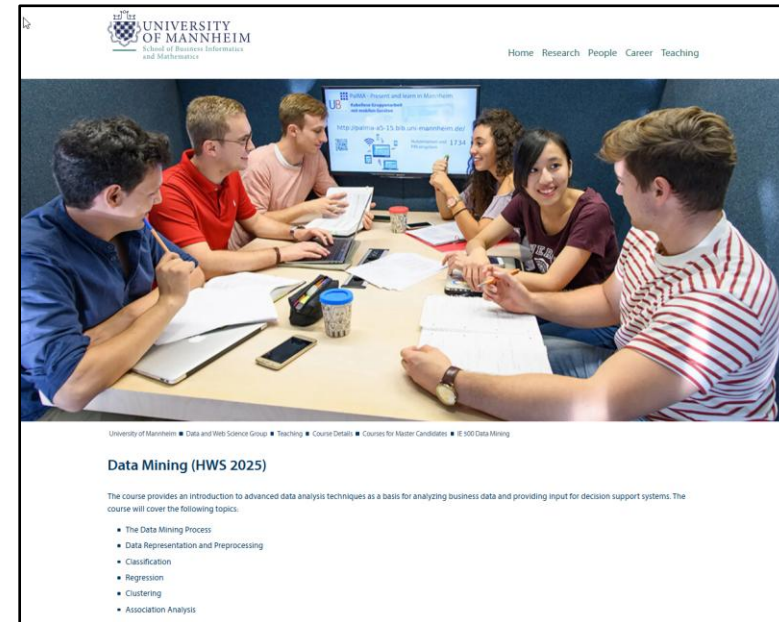
Hello

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



Course Organization

- 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



Week	Monday(Offline Lecture, Room A001)	Online Lecture (see Ilias Course)	Thursday (Exercise)
07.09.2026	no lecture		no exercise
14.09.2026	Introduction to Data Mining		Introduction to Python
21.09.2026	Preprocessing		Preprocessing
28.09.2026	Classification 1 + Intro to Student Project	Nearest Centroids	Classification 1
05.10.2026	Classification 2	Comparing Classifiers	Classification 2
12.10.2026	Regression	Ensembles	Regression
19.10.2026	Clustering and Anomalies	Hierarchical Clustering	Clustering and Anomalies
26.10.2026	Feedback on project outlines (no lecture)	Time Series	Time Series
02.11.2026	Association Analysis and Subgroup Discovery	Multi Modal Data	Association Analysis
09.11.2026	Project feedback session		Project Work
16.11.2026	Project feedback session		Project Work
23.11.2026	Project feedback session		Project Work
30.11.2026	Project feedback session		Project Work
07.12.2026	Q&A		Project Presentations

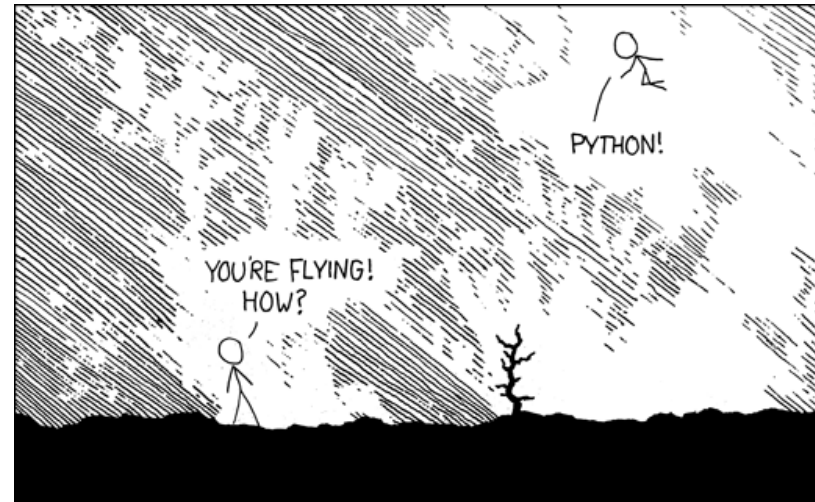
Course Organization - Exercise

- Exercise Groups
 - You will 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
 - Thursday, 13.45 – 15.15, **A203** Building B6, 26 Part A
 - Thursday, 15.30 – 17.00, A104 Building B6, 26 Part A
 - You can also switch between weeks if needed



Introduction to Python / Coding Env

- This week Thursday (in the exercise slot)
- Topics:
 - Setup of environment (Anaconda, Jupyter Notebooks)
 - IDE, working with LLMs
 - Python Intro / Design Goals
 - Basic programming concepts
 - Small intro to coding with LLMs
- Support
 - Help with environment setup
 - Q&A



Course Organization - Project

- Project Work
 - Teams of **five (exceptionally also 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 **Monday, October, 5th, 23:59**
 - Submission of project proposal
 - **Wednesday, October, 14th, 23:59**
 - Submission of final project work report
 - **Sunday, December 6th, 23:59**
 - Submission of Presentation (PDF)
 - **Wednesday, December 9th, 23:59**



Course Organization - Exam

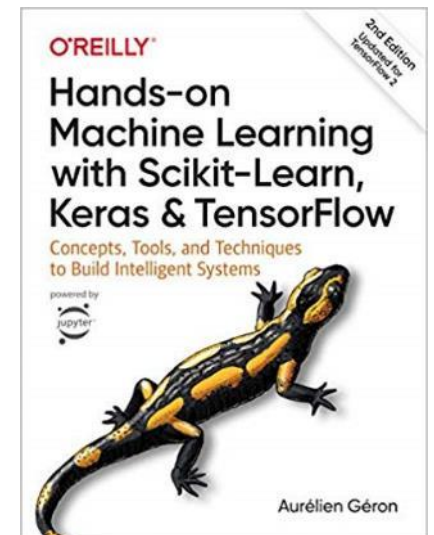
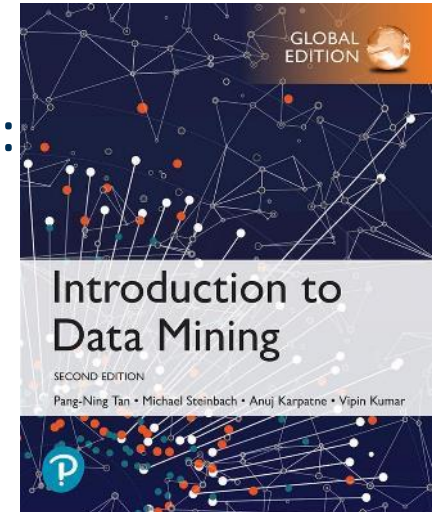
- Date and Time: **Monday, 21th December 2026**
- Duration: 60 minutes
- Structure: 6 open questions that
 - Check whether you have understood the lecture content
 - We try to cover all major chapters of the lecture
 - Require you to describe the ideas behind algorithms and methods
 - Often: How do methods 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

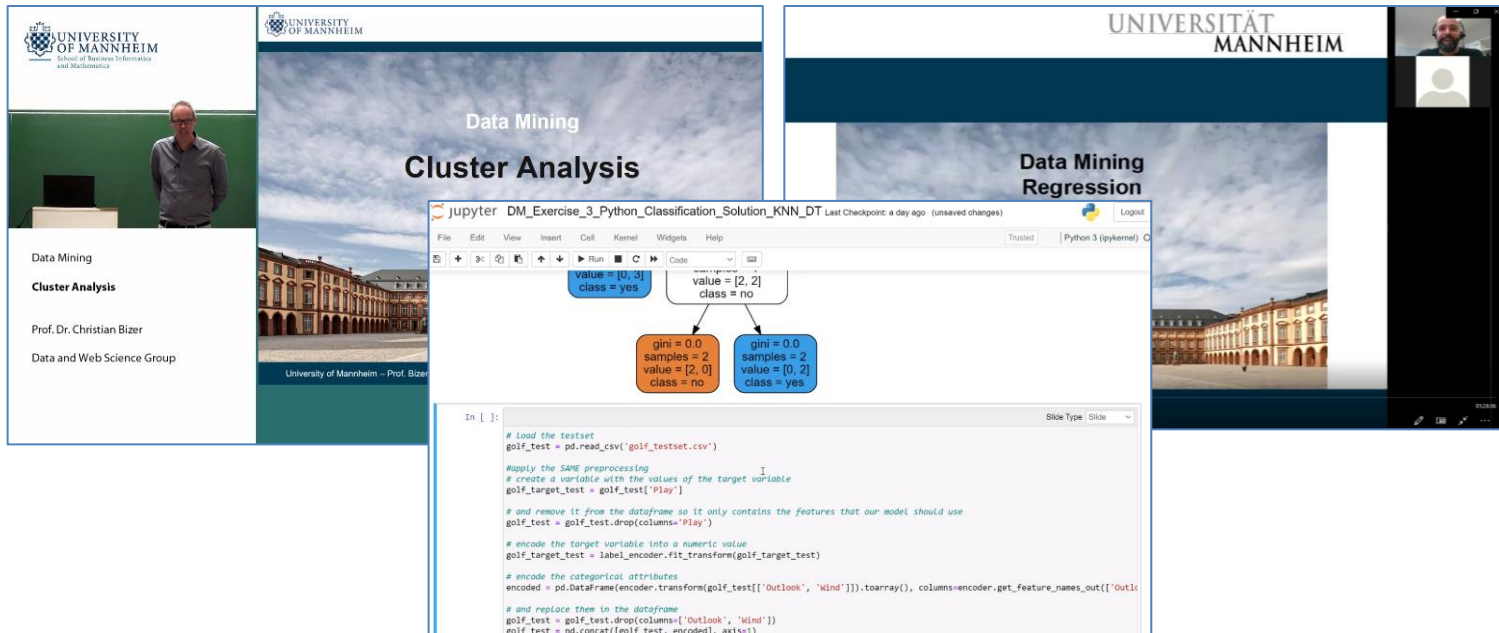
Textbooks for the Course

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

- **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



The collage consists of three overlapping screenshots from a video lecture. The leftmost screenshot shows a man standing in front of a green screen with the University of Mannheim logo and text: "Data Mining", "Cluster Analysis", "Prof. Dr. Christian Bizer", "Data and Web Science Group". The middle screenshot shows a Jupyter Notebook titled "DM_Exercise_3_Python_Classification_Solution_KNN_DT" with a decision tree diagram. The rightmost screenshot shows a slide titled "Data Mining Regression" with the University of Mannheim logo and a video feed of a man in the top right corner.

Decision Tree Diagram:

```
graph TD
    A["value = [0, 3]  
class = yes"] --> B["gini = 0.0  
samples = 2  
value = [2, 0]  
class = no"]
    A --> C["value = [2, 2]  
class = no"]
    C --> D["gini = 0.0  
samples = 2  
value = [0, 2]  
class = yes"]
```

Jupyter Notebook Code:

```
In [ ]:
# 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(['Outlook', 'Wind']))

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

Questions?

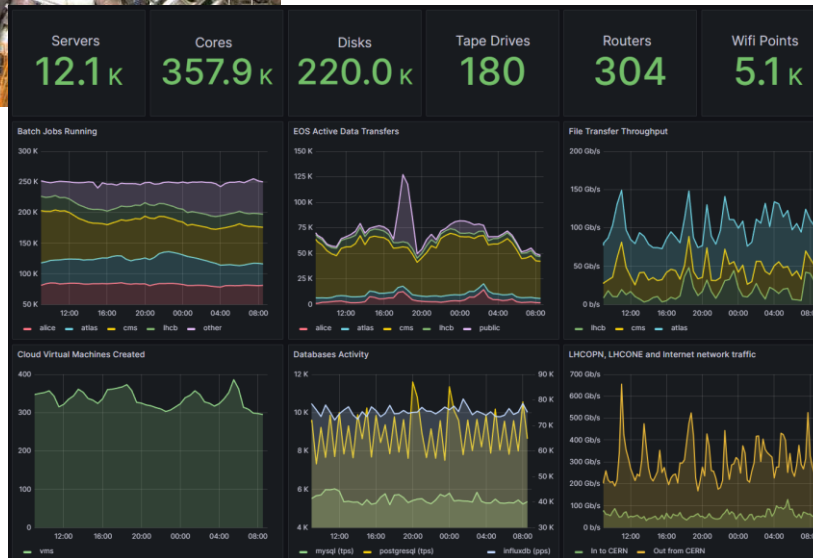
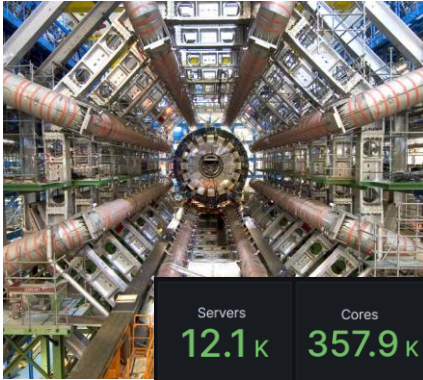


What is Data Mining?

- **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...”

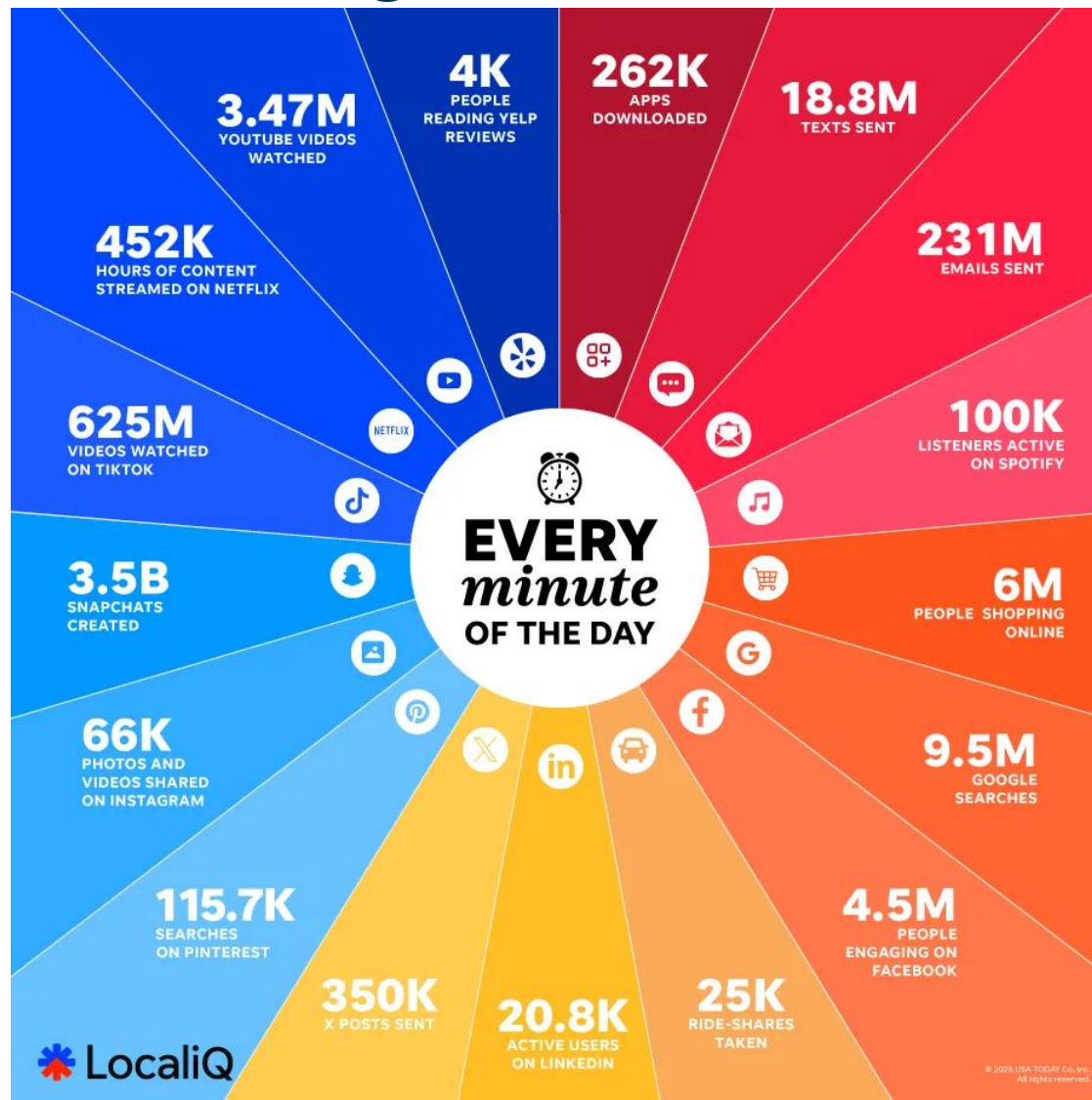


- **US Library of Congress**
 - ≈ 235 TB archived
- **Discover**
 - Topic distributions*
 - Citation networks
- **Train**
 - Large Language Models

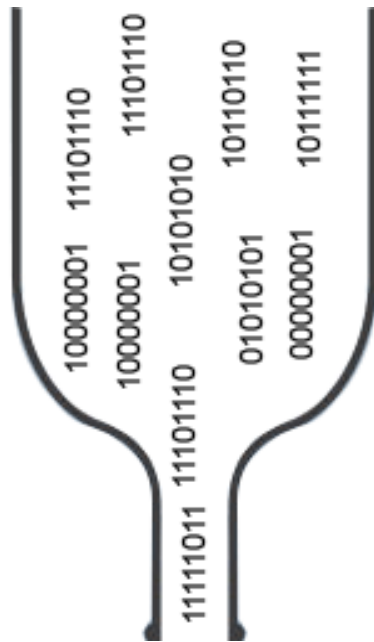
<https://www.brandwatch.com/blog/facebook-statistics/>

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

“We are Drowning in Data...”



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

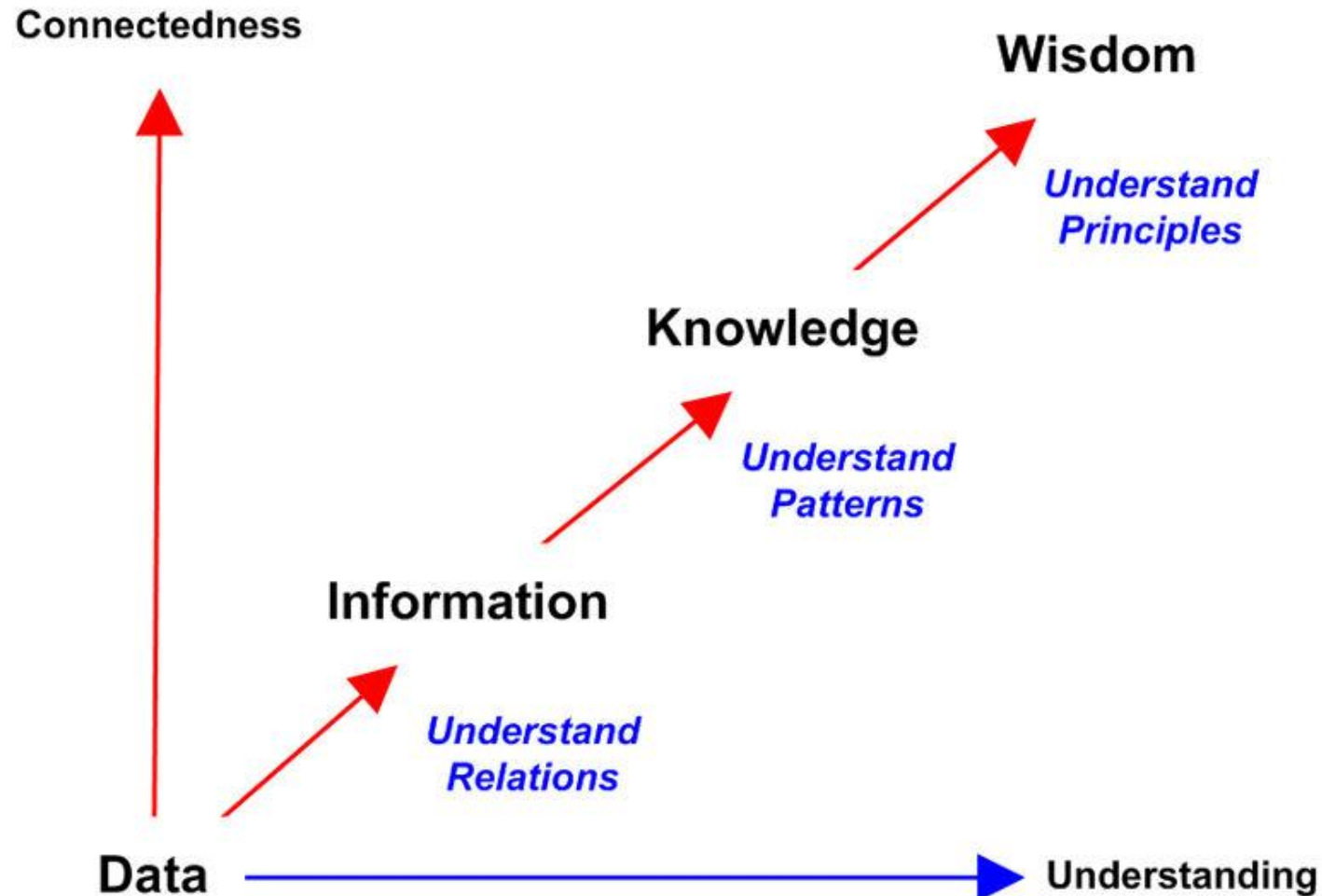


← Rate at which data are produced

← Rate at which data can be understood
manual interpretation is hardly feasible!

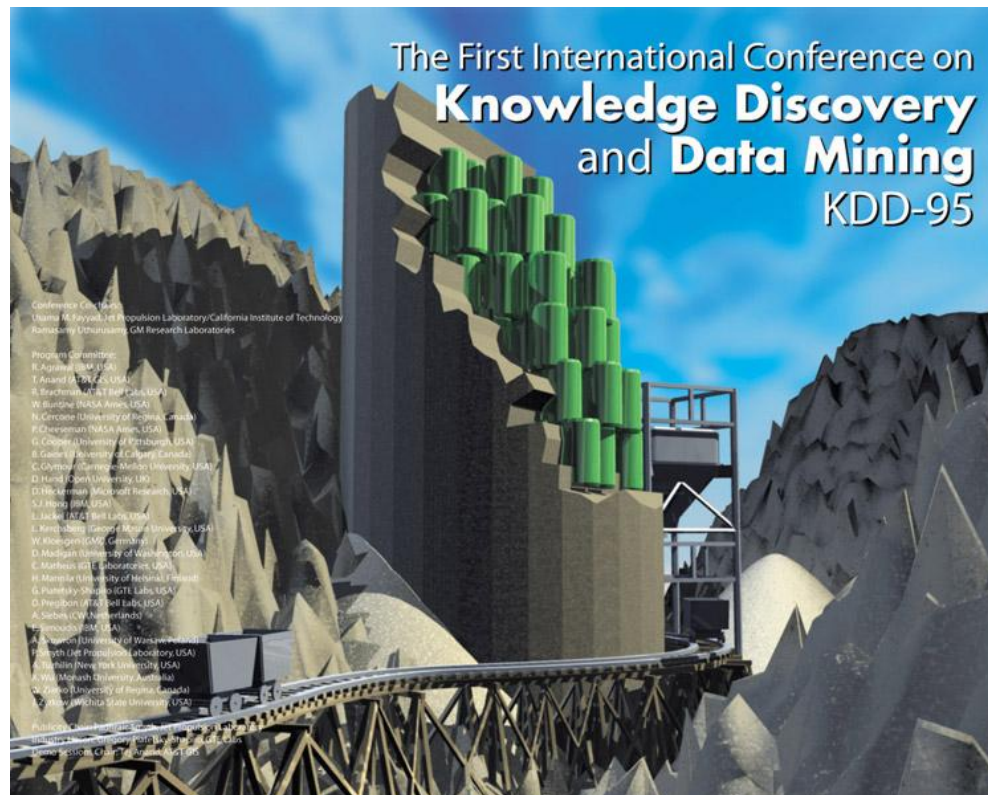
- We are interested in **the patterns, not the data** itself!
- Data Mining methods help us to
 - **Discover interesting patterns** in large quantities of data
 - **Take decisions** based on the patterns

Data, Information, Knowledge, Wisdom



Data Mining: Definitions

- Idea: mountains of data
 - Where knowledge is mined

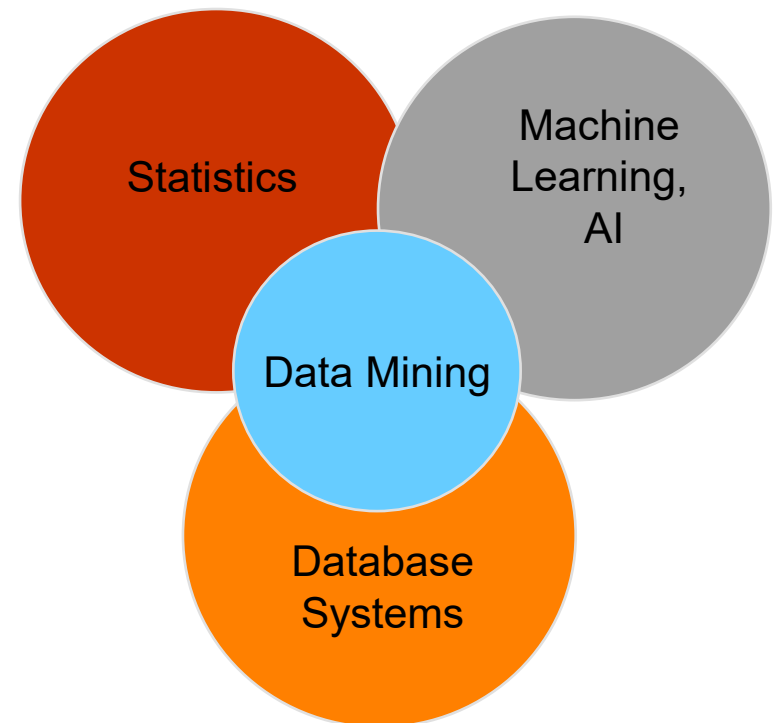


Data Mining: Definitions

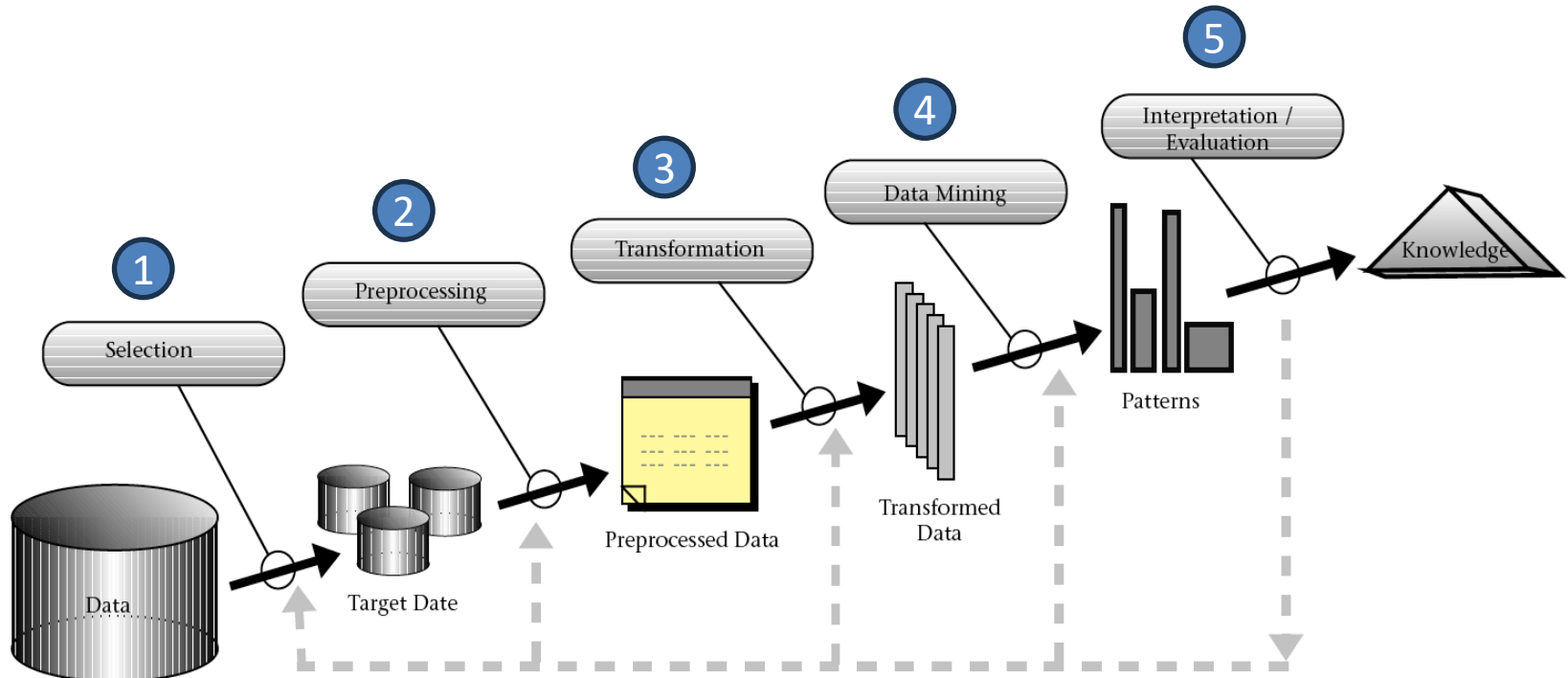
- Data Mining is a non-trivial process of identifying
 - valid
 - novel
 - potentially useful
 - ultimately understandablepatterns in data. (Fayyad et al. 1996)
- Data Mining methods
 1. Detect interesting patterns in large quantities of data
 2. Support human decision making by providing such patterns
 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



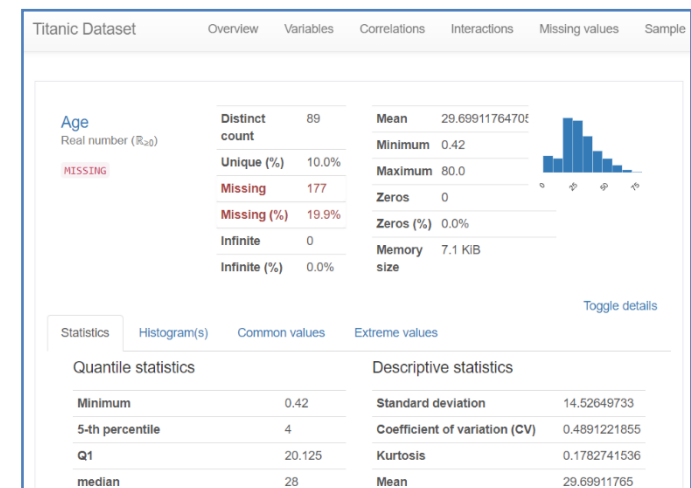
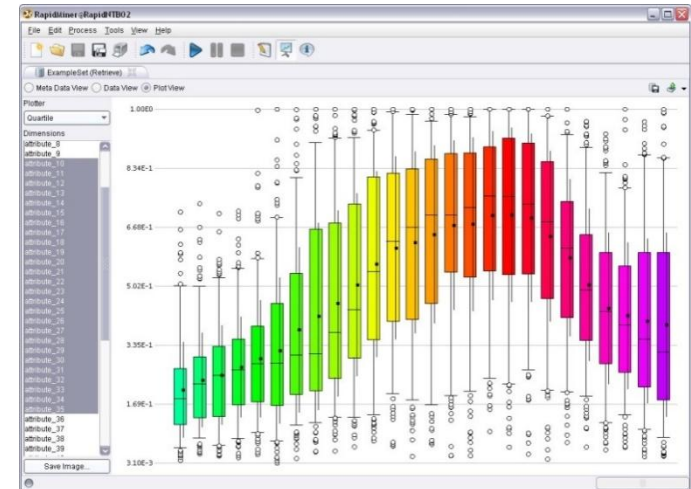
The Data Mining Process



Source: Fayyad et al. (1996)

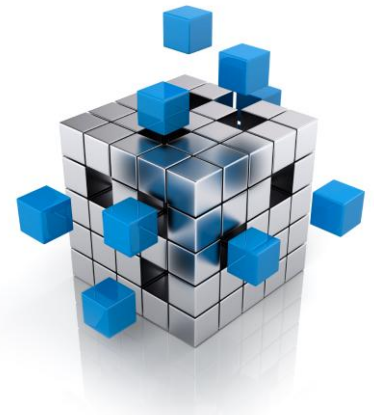
Selection and Exploration (1)

- 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



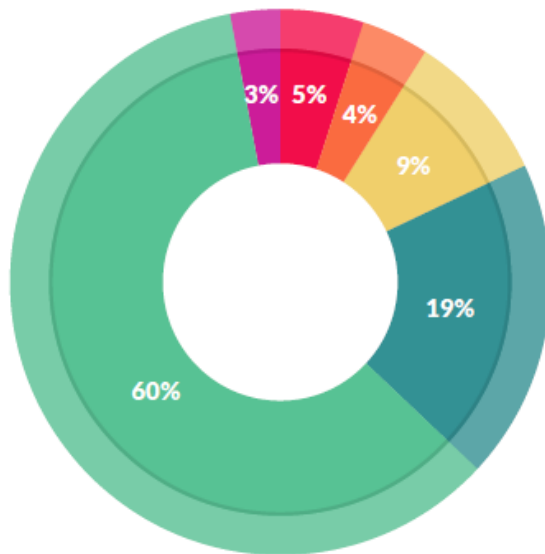
Preprocessing and Transformation (2+3)

- 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



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%

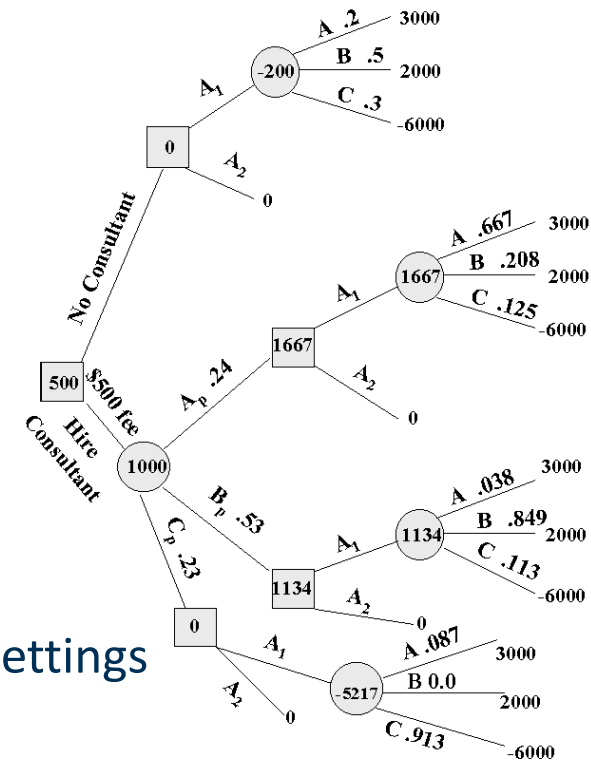
Advertisement:
IE670 Web Data
Integration

Source: CrowdFlower Data Science Report 2016: <http://visit.crowdfower.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

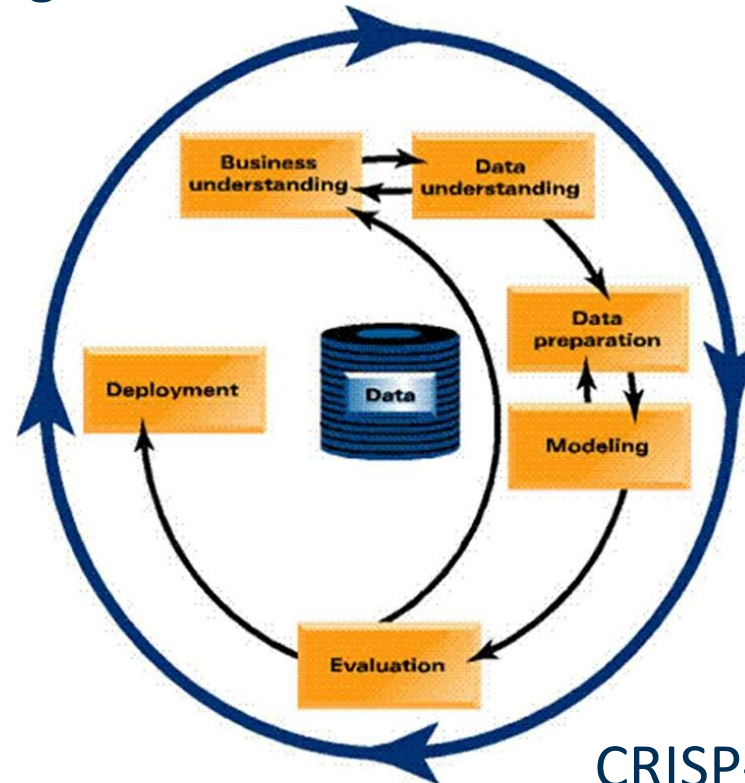


- Increase revenue



Deployment

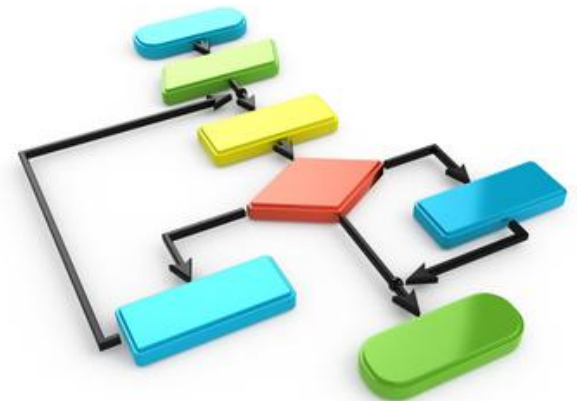
- Use model in the business context
- Keep iterating in order to maintain and improve model



CRISP-DM Process Model

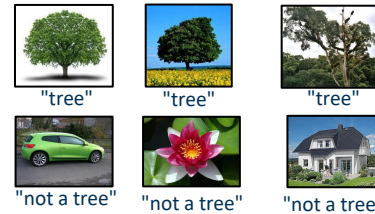
Tasks and Applications

- **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

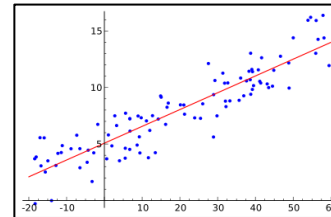


Data Mining Tasks

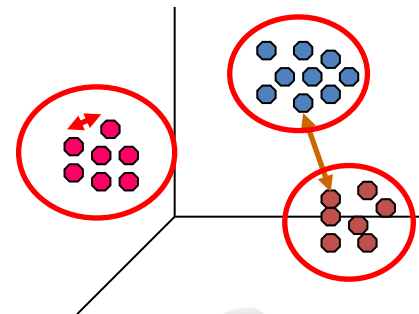
- Classification [Predictive]



- Regression [Predictive]



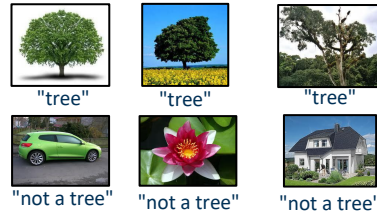
- Cluster Analysis [Descriptive]



- Association Analysis [Descriptive]



Classification



- 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

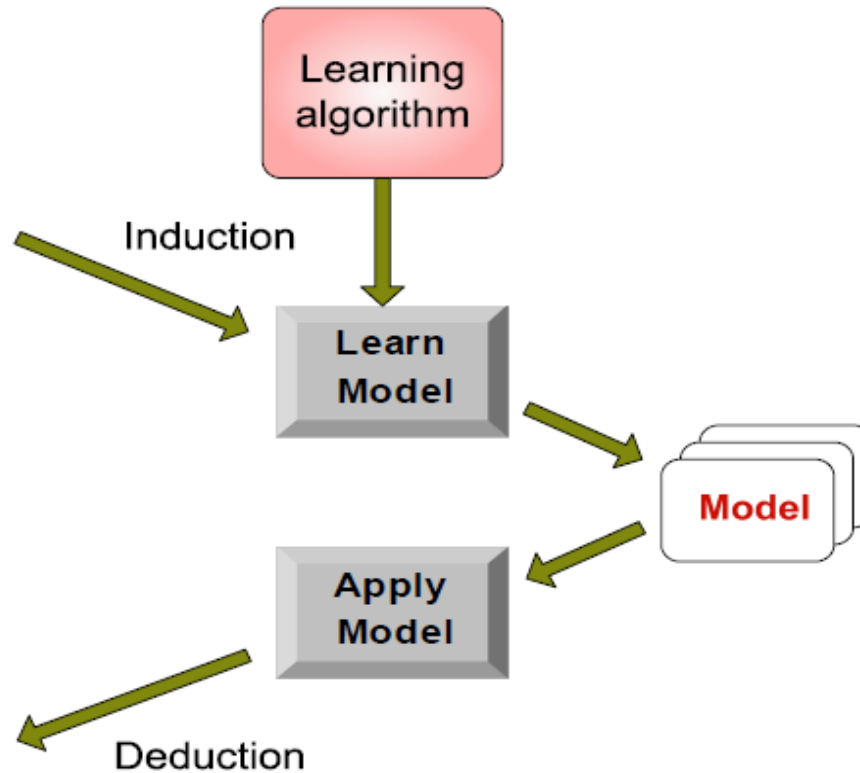
Class/Label Attribute

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

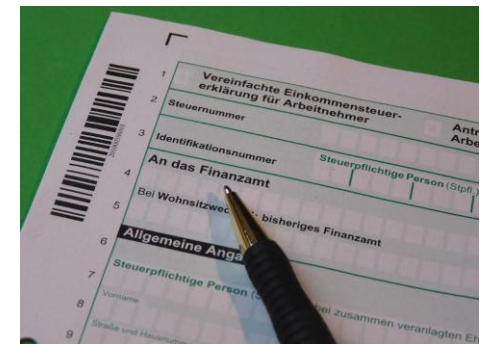
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Unseen Records



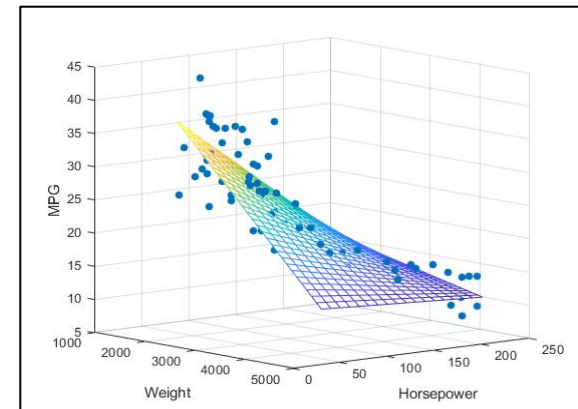
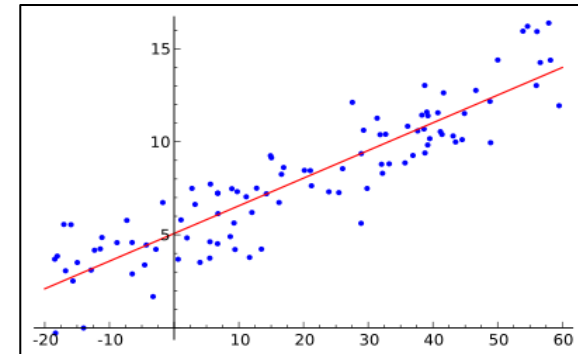
Classification: Applications

- 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?



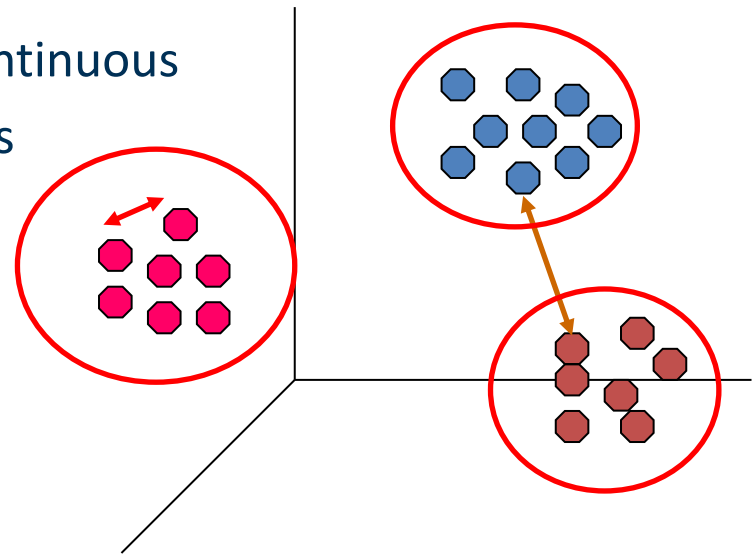
Regression

- 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)

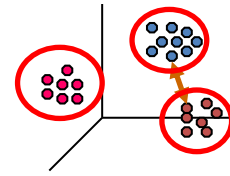


Cluster Analysis

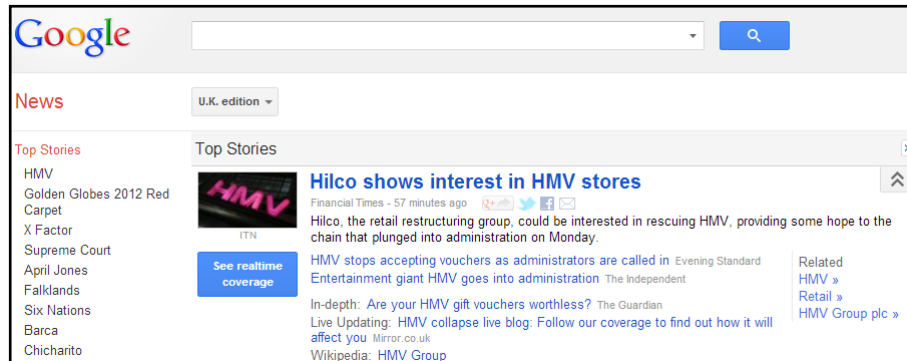
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find groups such that
 - Data points in one group are more similar to one another
 - Data points in separate groups are less similar to one another
- Similarity Measures
 - Euclidean distance if attributes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



Association Analysis



- 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

<i>TID</i>	<i>Items</i>
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**

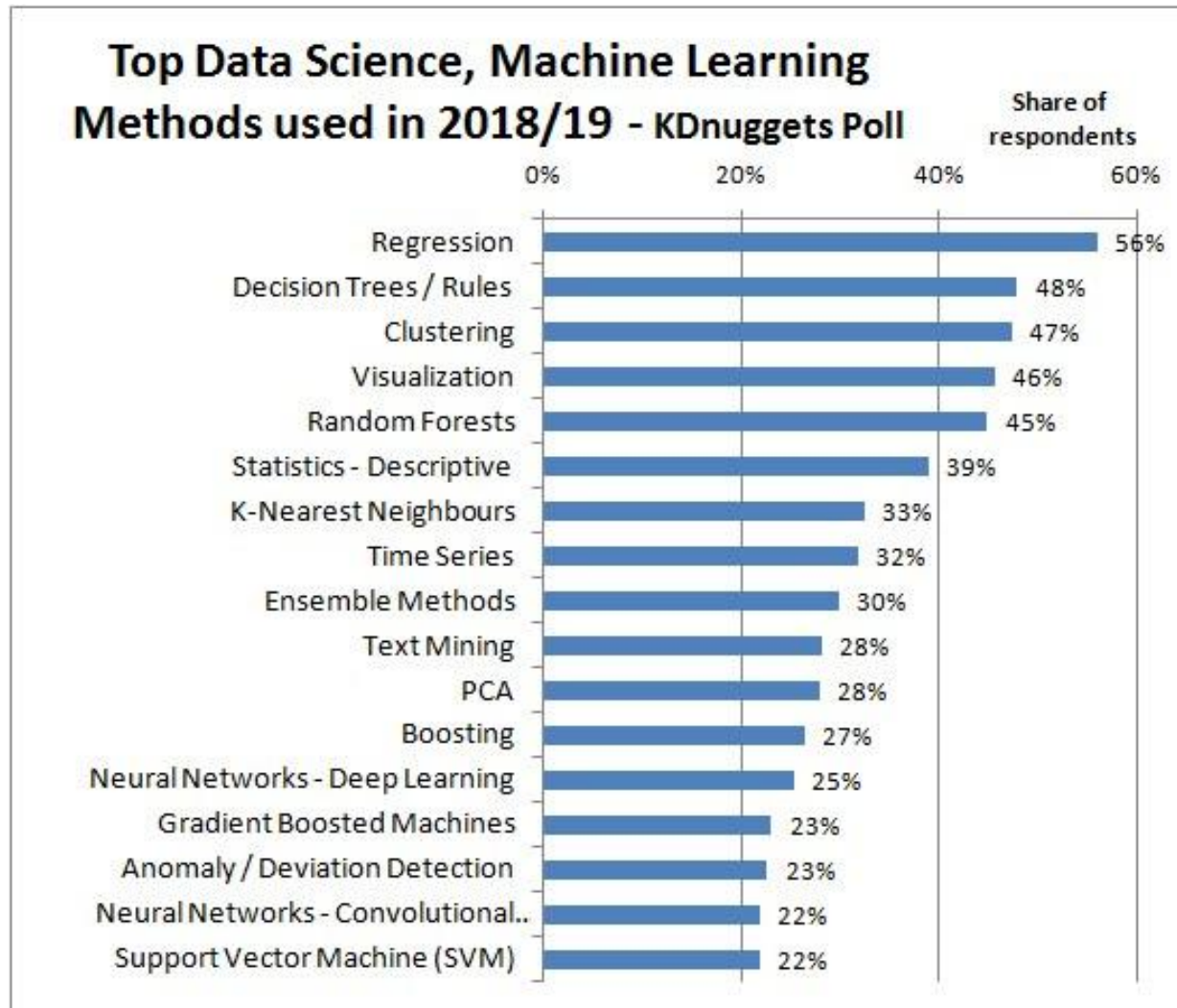


Add all three to Cart

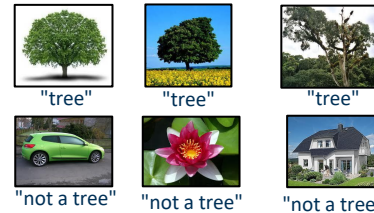
Add all three to Wish List

[Show availability and shipping details](#)

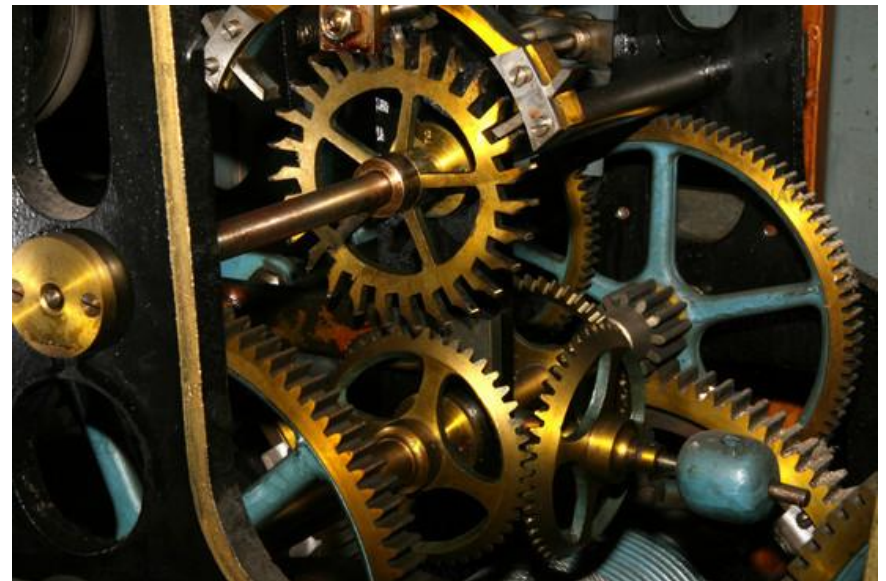
Which Methods are Used in Practice?



Classification Algorithms

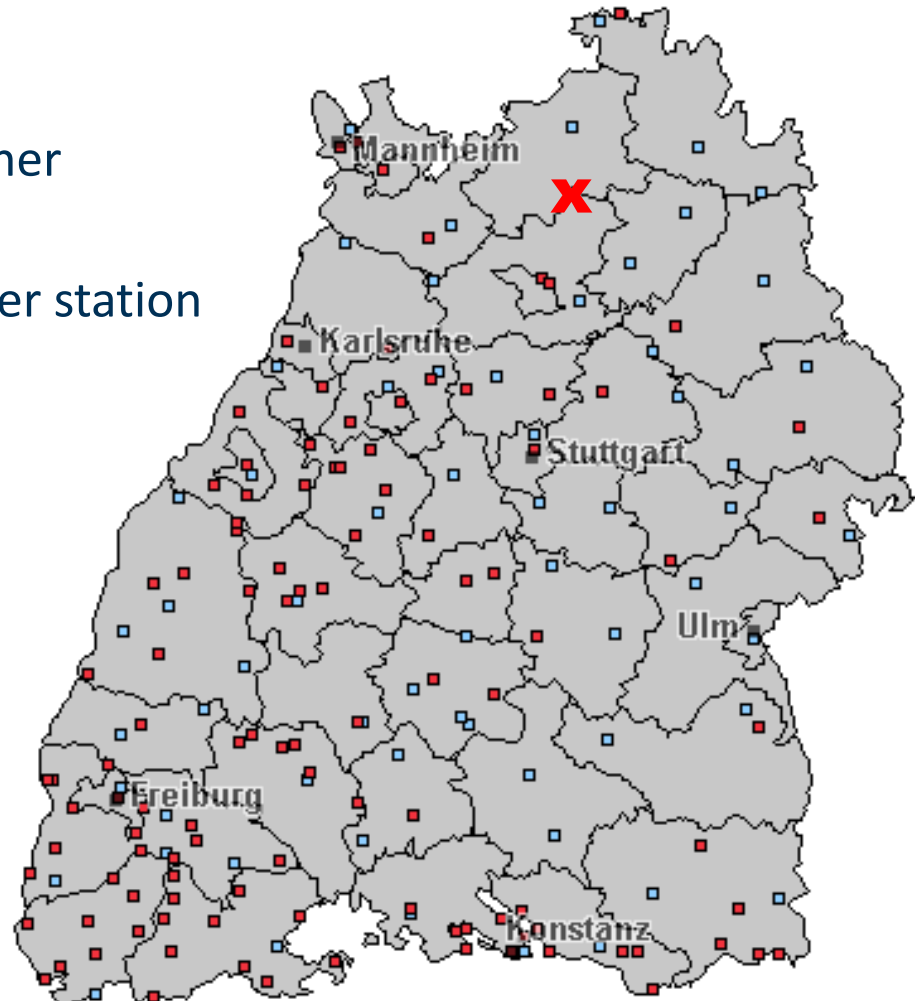


- Classification:
 - We give the computer a set of labeled examples
 - The computer learns to classify new (unlabeled) examples
- How does that work?
 - **K-Nearest-Neighbors**
 - Decision Trees
 - Naïve Bayes
 - Support Vector Machines
 - Artificial Neural Networks
 - Deep Neural Networks
 - Many others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



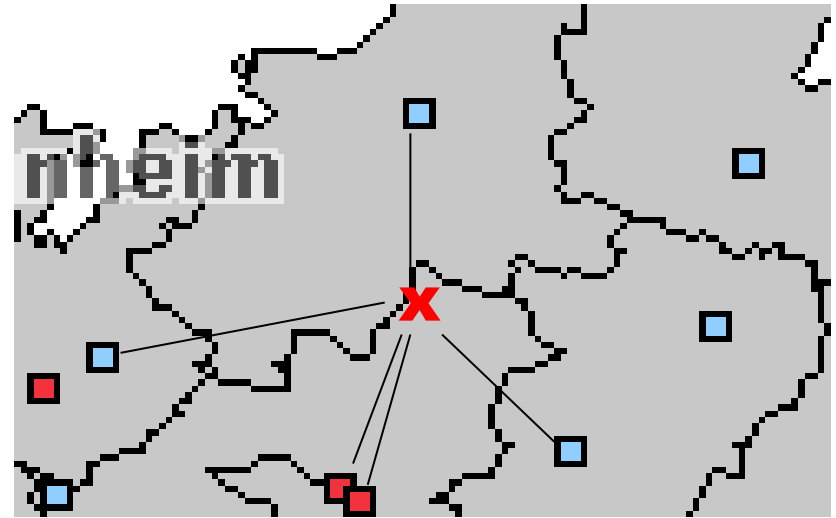
K-Nearest-Neighbors

- 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

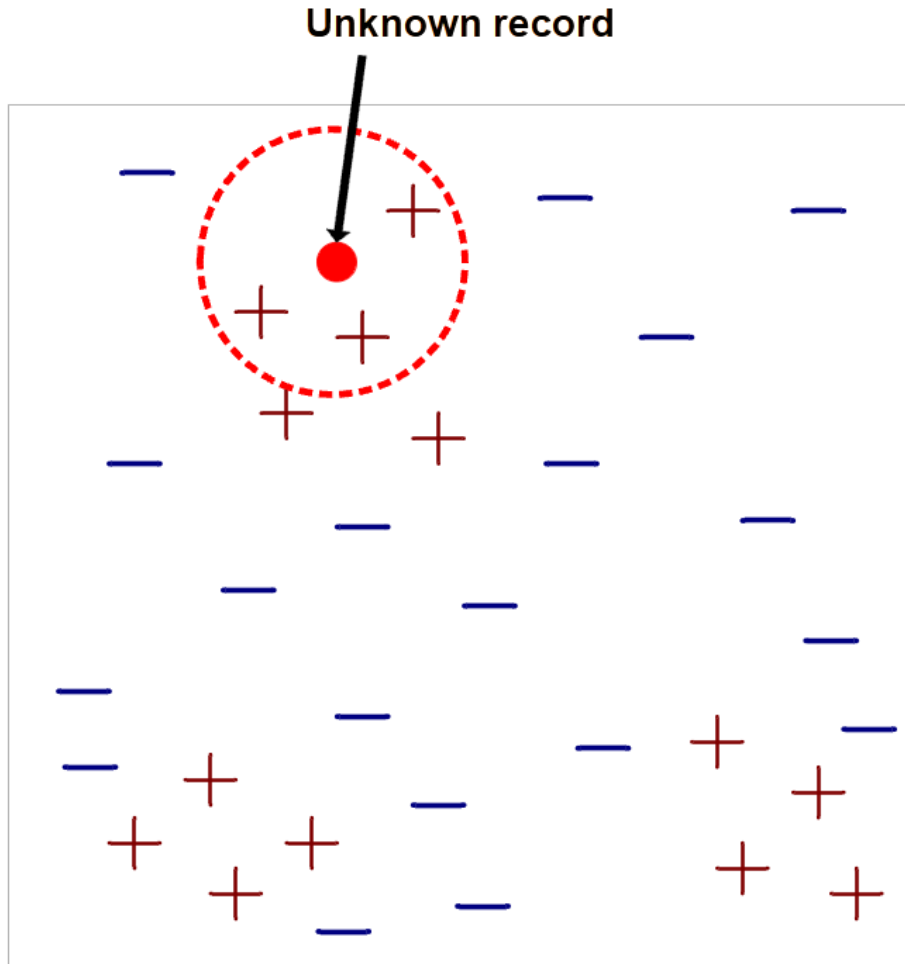


K-Nearest-Neighbors

- 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

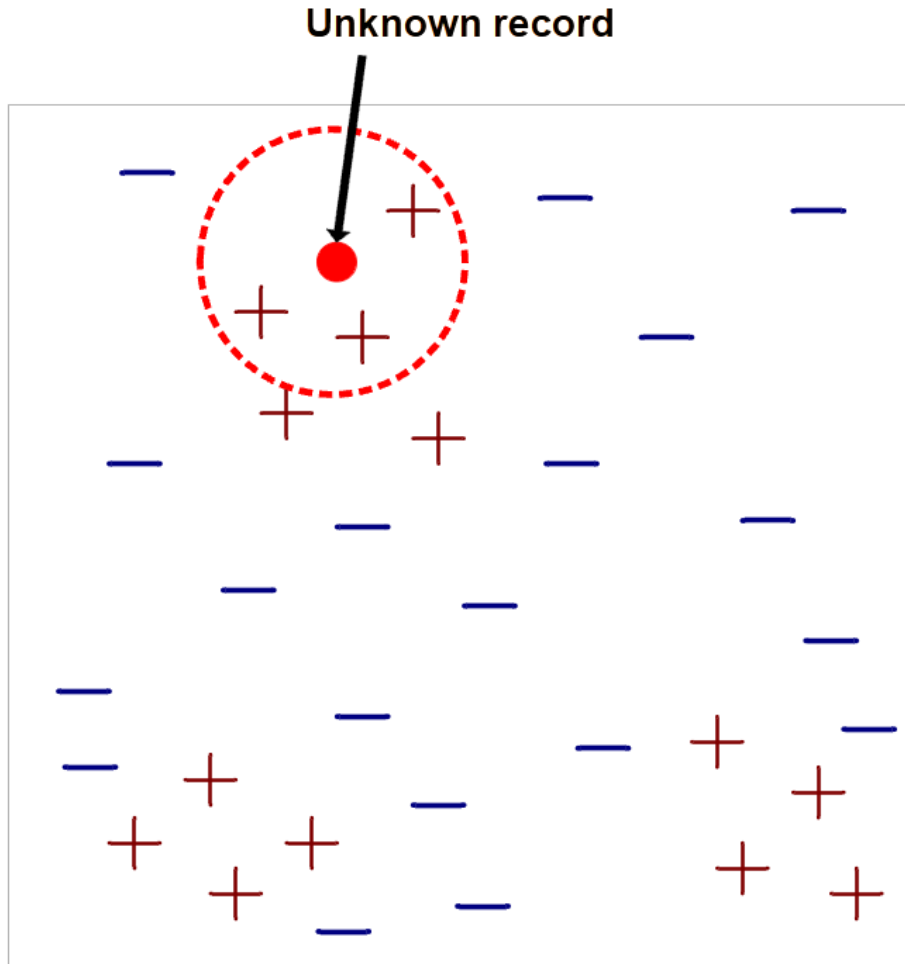


K-Nearest-Neighbor Classifier



- 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

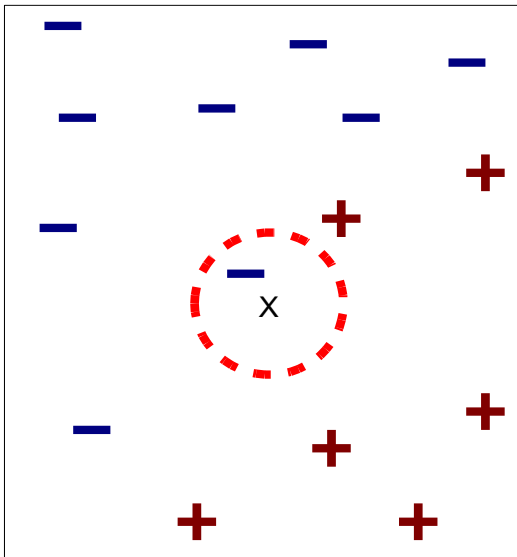
K-Nearest-Neighbor Classifier



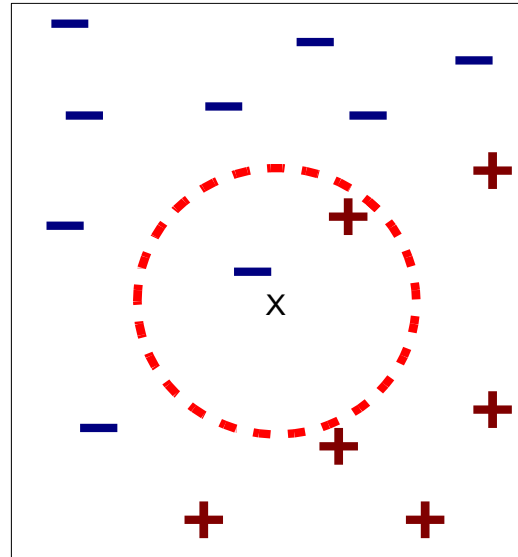
- To classify an unknown record:
 - **Compute distance** to each training record
 - Identify **k-nearest neighbors**
 - Use **class labels of nearest neighbors** to determine the class label of unknown record
 - By taking majority vote or
 - By weighing the vote according to distance

Examples of K-Nearest Neighbors

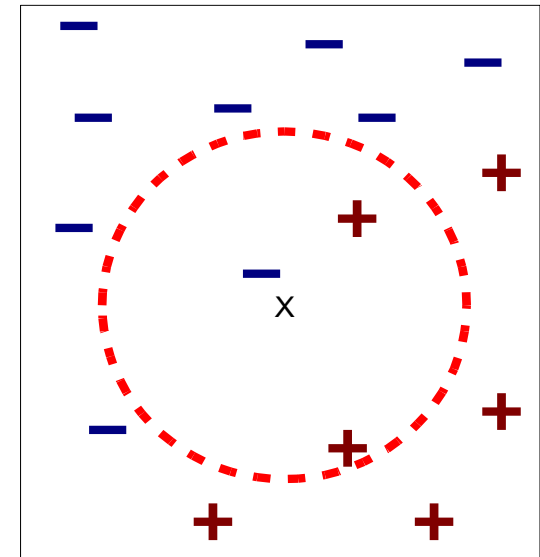
- The k -nearest neighbors of a record x are data points that have the k smallest distances to x



(a) 1-nearest neighbor



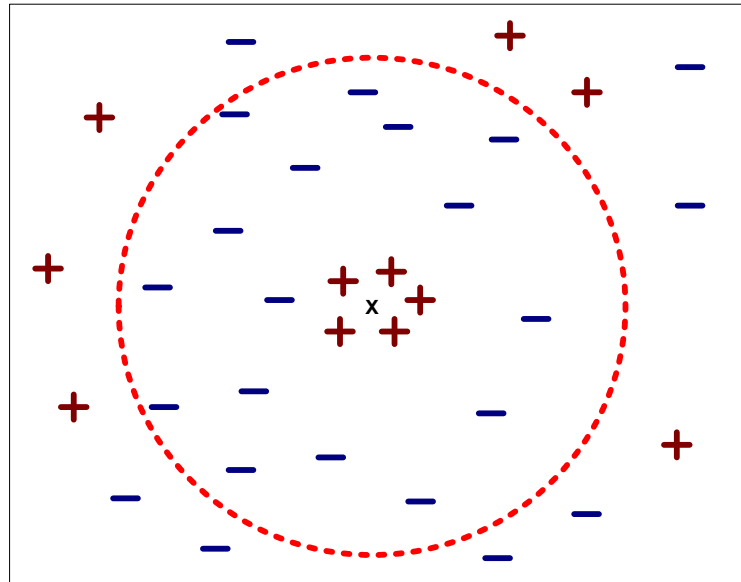
(b) 2-nearest neighbor



(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



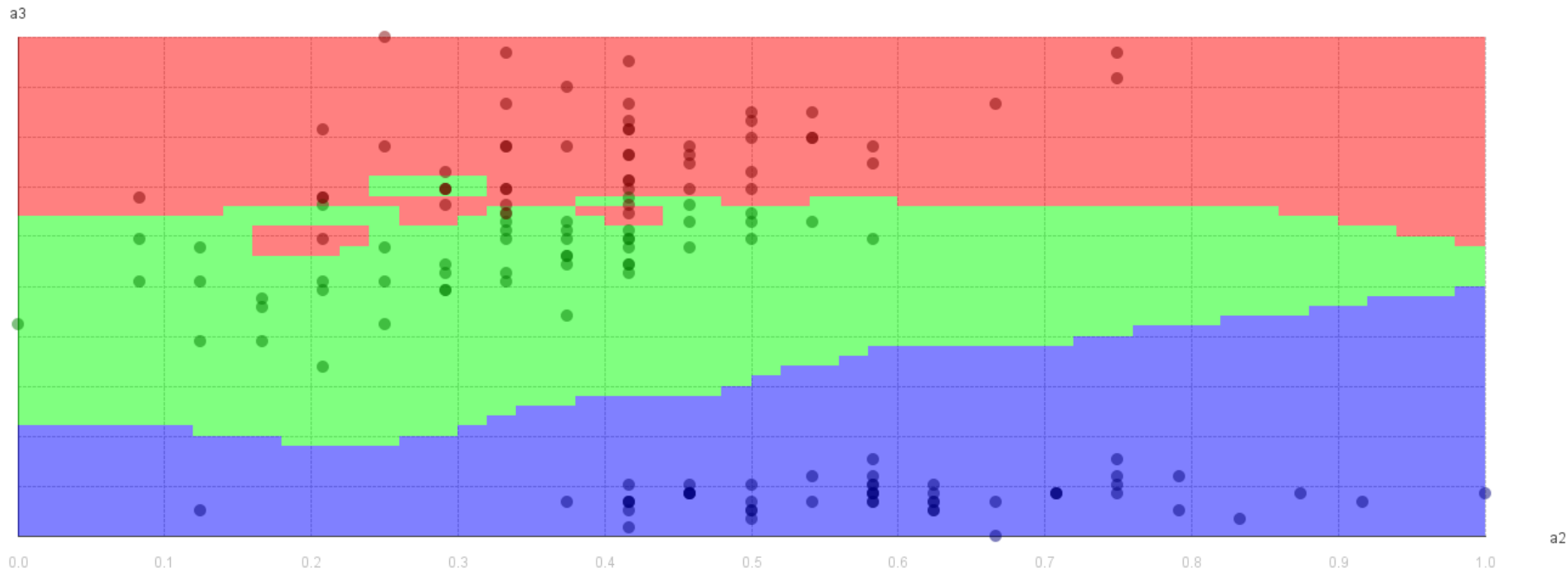
- Rule of thumb: Test k values between 1 and 20

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)

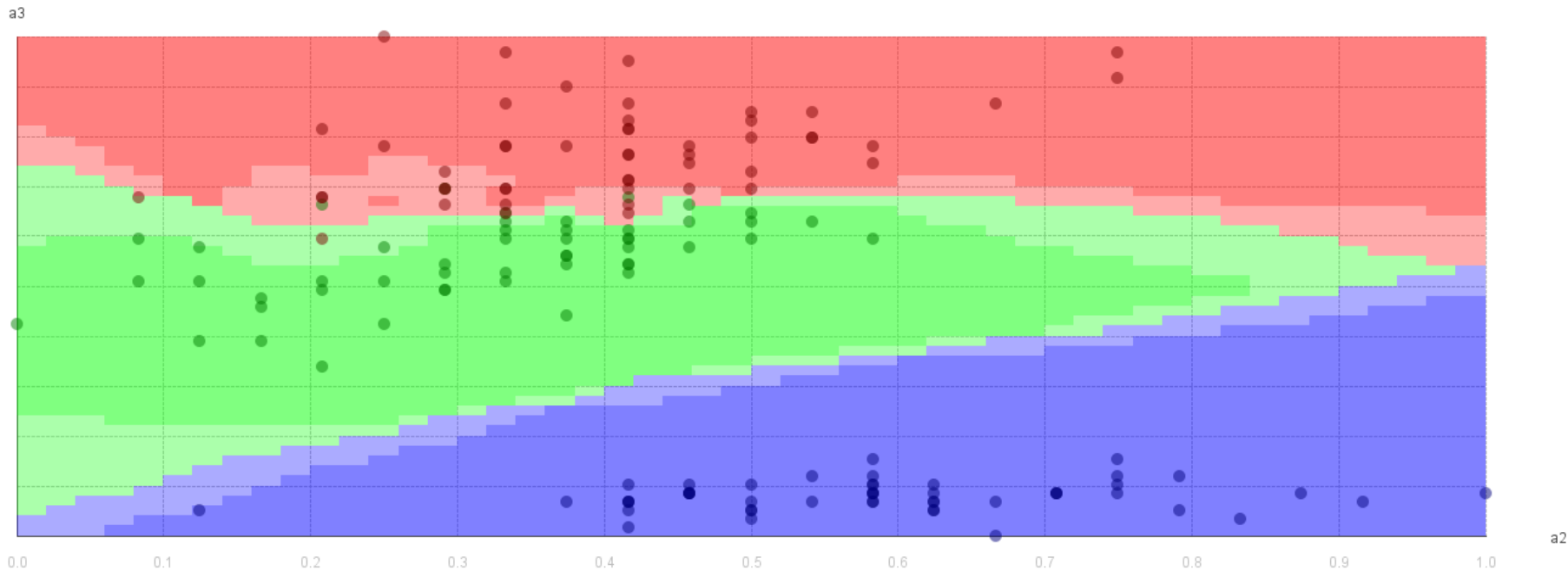
Decision Boundaries of a k-NN Classifier

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



Decision Boundaries of a k-NN Classifier

- $k=3$
- 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



Questions?