# Data Analysis and Knowledge Discovery

## Introduction

Antti Airola

University of Turku
Department of Computing

Antti.Airola@utu.fi

# Topics

- What is this thing called data analytics?
- What is it good for?
- What are data and knowledge?
- How can one implement systems that learn from data?
- What can you learn by taking this course?

# Buzzwords

- Artificial intelligence
- Big data
- Data analytics
- Data mining
- Data science
- Deep learning
- Intelligent data analysis
- Knowledge discovery in databases
- Machine learning
- Pattern recognition
- Statistical learning
- ...

# Data

- Faster and cheaper computers, sensors and storage media combined with higher bandwidth data connections have made possible to collect almost any sort of electronic data
  - Documents, images, sounds, etc.
- Society produces huge amounts of data
  - business, science, medicine, economics, geography, environment, sports, ...
- Potentially valuable resource
- Raw data useless: need techniques to automatically extract knowledge from it
  - Data: recorded facts
  - Knowledge: patterns/regularities underlying the data

# Data vs. knowledge

Data is raw. It simply exists and has no significance beyond its existence (in and of itself). It can exist in any form, usable or not. It does not have meaning of itself.

- refer to single instances (single objects, people, events, points in time etc.)
- describe individual properties and their utility is hence limited
- are often available in large amounts (databases, archives)
- are often easy to collect or to obtain (e.g. scanned cashiers in supermarkets, collected from an industrial process, internet,...)
- do not allow us to make predictions or forecasts

# Data Science

"We are drowning in data, but we are starved for knowledge"

# Data vs. knowledge

Knowledge

- refers to classes of instances (sets of objects, people, events, points in time etc.)
- describes general patterns, structures, laws, principles etc.
- consists of as few statements as possible (Occam's razor)
- often difficult and time-consuming to find or to obtain (e.g. natural laws)
- allows us to make predictions and forecasts
- Examples
  - All masses attract each other
  - Every day at 6:15 AM there is a flight from Turku to Helsinki

# Criteria to access knowledge

- Correctness (probability, success in tests)
- Generality (domain and conditions of validity)
- Usefulness (relevance, prediction power)
- Comprehensibility (simplicity, clarity)
- Novelty (previously unknown, unexpected)
- Something else...??

# Data Science

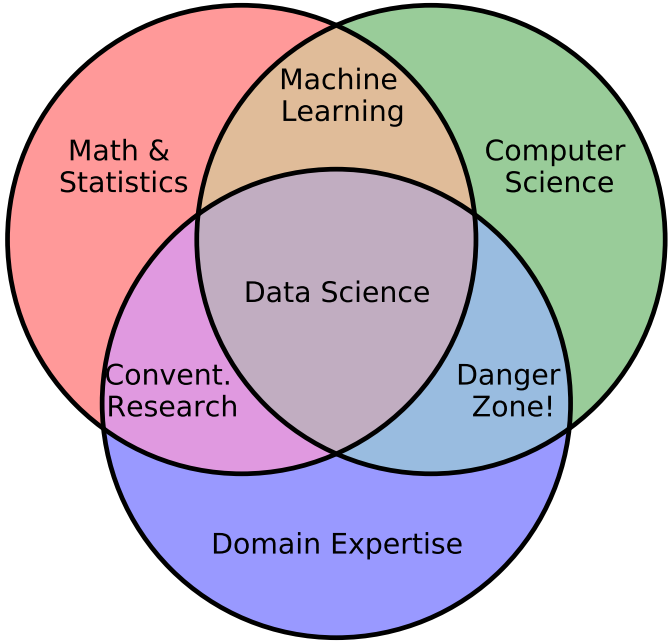



Figure: After Drew Conway's diagram

Danger zone - "lies, damned lies, and statistics": through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created.

# Data Science

Interdisciplinary field

- Computer science
  - algorithmics, computational complexity, data structures, data bases, parallelism, programming, software engineering...
- Mathematics and statistics
  - calculus, linear algebra, continuous and discrete optimization, probability theory, statistical inference...
- Domain expertise
  - astronomy, biology, geology, medicine, marketing...

# To explain or to predict?

- **Descriptive analytics**
  - goal: increased understanding of the data we have
  - visualizations, summaries, correlations, clusterings...
  - examples: histogram, number of students completing a course, division of customer data to different subgroups, Elo ratings of chess players
- **Predictive analytics**
  - goal: making predictions for unknown objects or future data
  - classification, regression, reinforcement learning models
  - example: predicting weather, medical diagnostics, AI playing chess
- sometimes both goals pursued simultaneously
  - e.g. (generalized) linear model could encapture how a group of variables (e.g. age, weight, biomarkers) is associated with an outcome to be predicted (e.g. sick / healthy)

# Learning from data

Herbert Simon: Learning denotes changes in a system that enable a system to do the same task more efficiently the next time.

Herbert Simon: Learning is any process by which a system improves performance from experience.

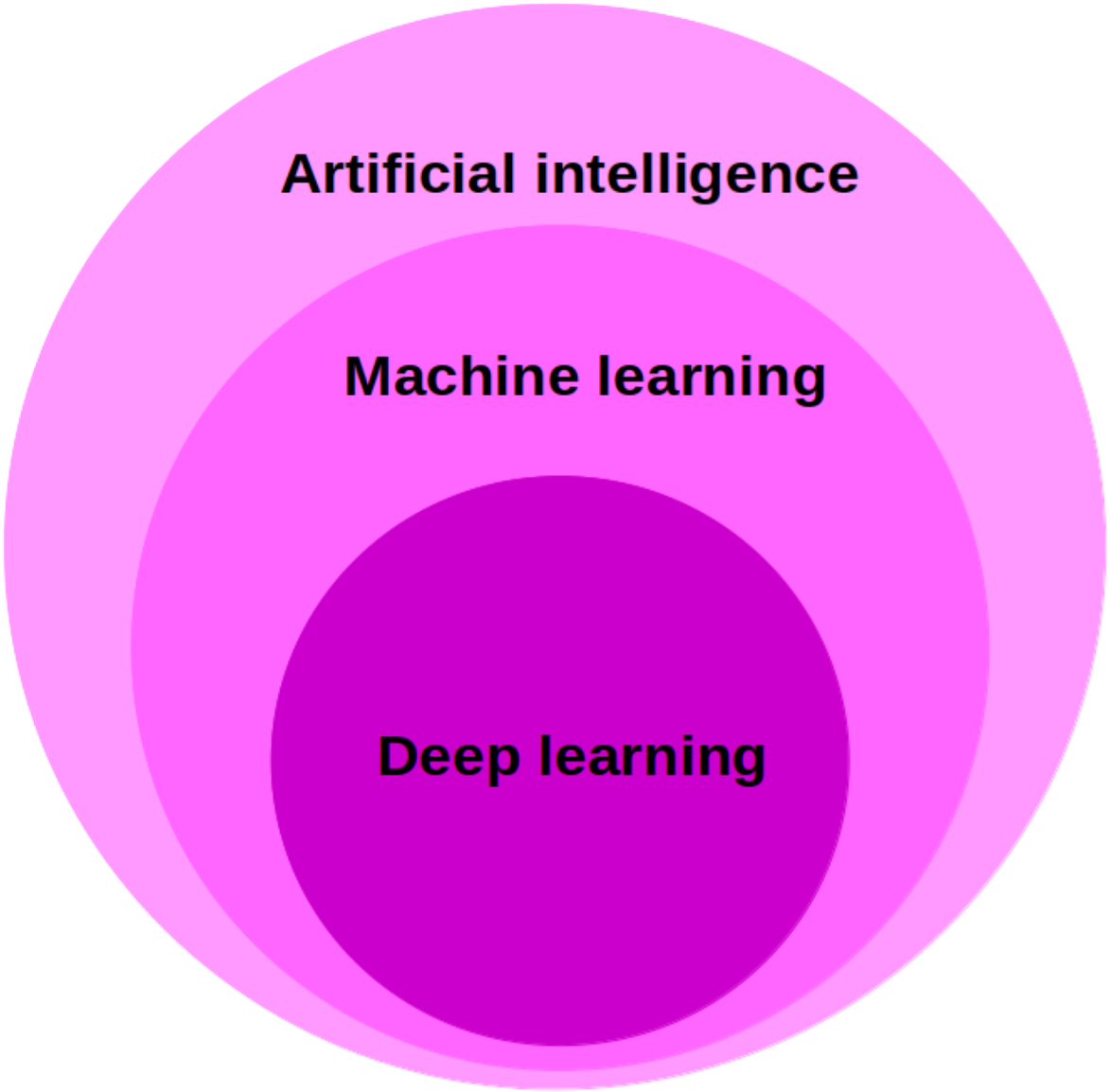
Ryszard Michalski: Learning is constructing or modifying representations of what is being experienced.

Marvin Minsky: Learning is making useful changes in our minds

# Machine learning

- *Algorithms for acquiring structural descriptions from examples*
- Structural descriptions represent patterns explicitly
  - may help in understanding trends or relationships in data
  - may be used to predict outcome in new situation...
  - and/or to understand and explain how prediction is derived
    - caveat: many state-of-the-art machine learning methods produce non-interpretable "black-box" models
    - trade-off between interpretability and accuracy
- Methods originate from artificial intelligence, statistics, mathematics and research on databases

# Machine learning



# Machine learning is based on mathematical modeling

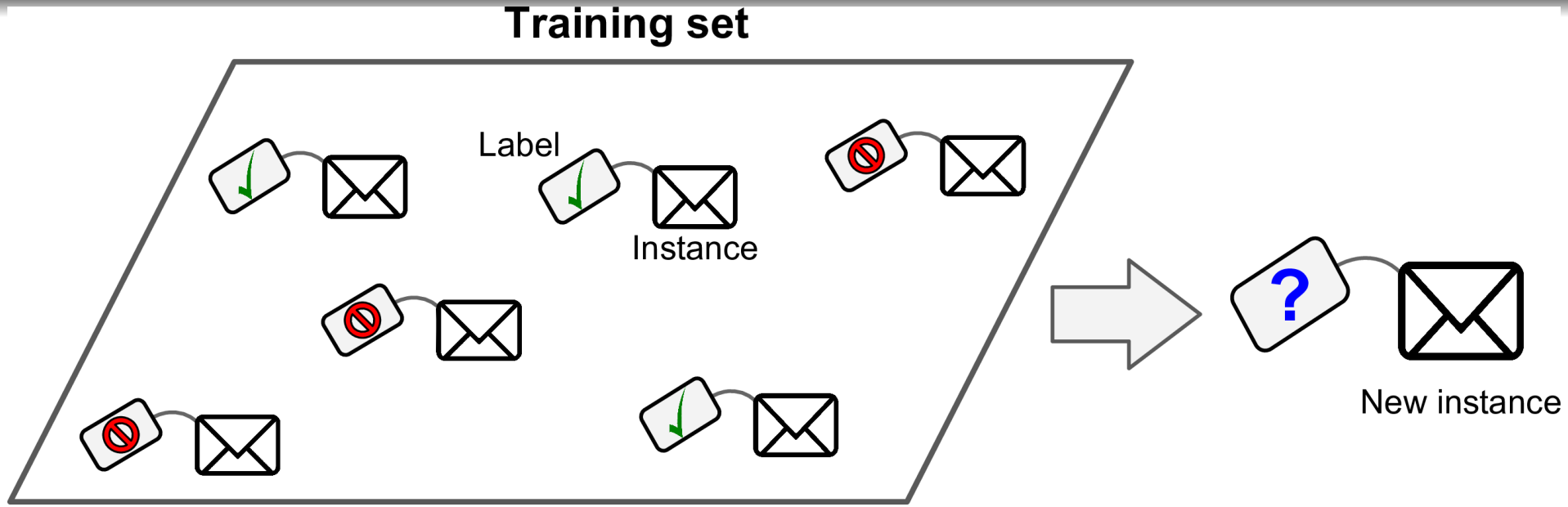
Modeling framework

- **Problem:** there is a need to model some part of the universe and make decisions based on the model
- **Modeling:** build the best model possible from a priori knowledge and data available
  - What models do we consider?
  - How do we find good models?
  - How do we compare models?
- **Prediction:** use the model to predict properties of interest
- **Decision making:** decide actions based on the predictions
  - How do we deal with uncertainty and what actions do we choose?

# Supervised vs. unsupervised learning

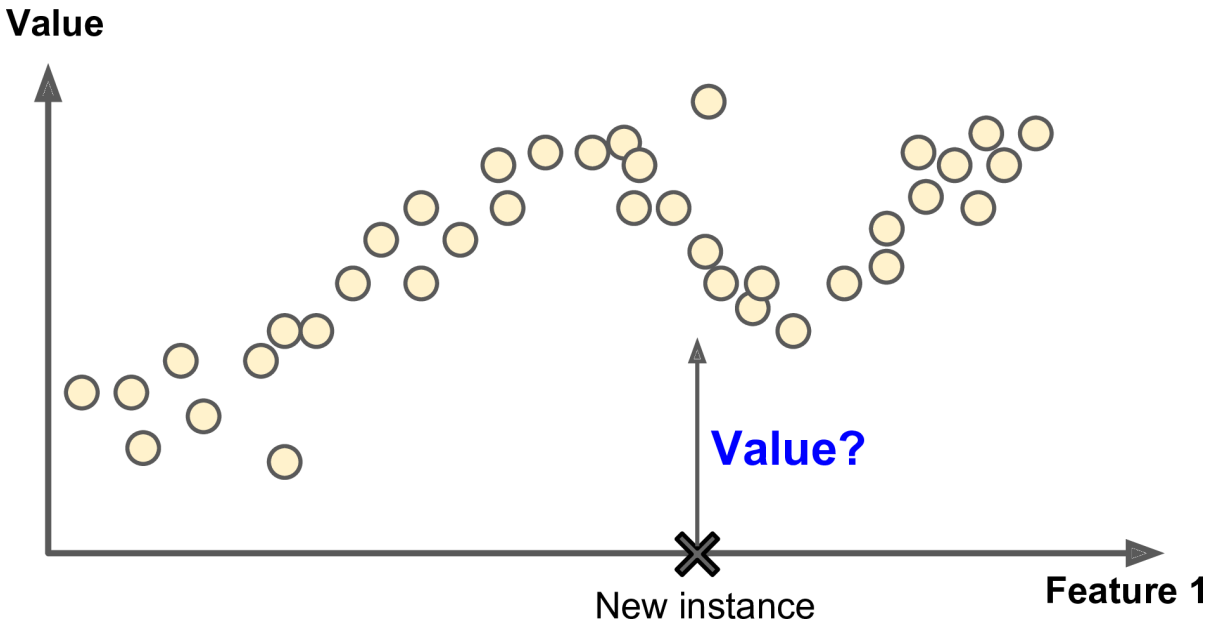
- Every data analysis problem is different, but can recognize certain general problem categories
- Classical division in machine learning
- Supervised learning
  - classification
  - regression
- Unsupervised learning
  - clustering, segmentation
  - association analysis
  - most dimensionality reduction methods
- Reinforcement learning

# Classification



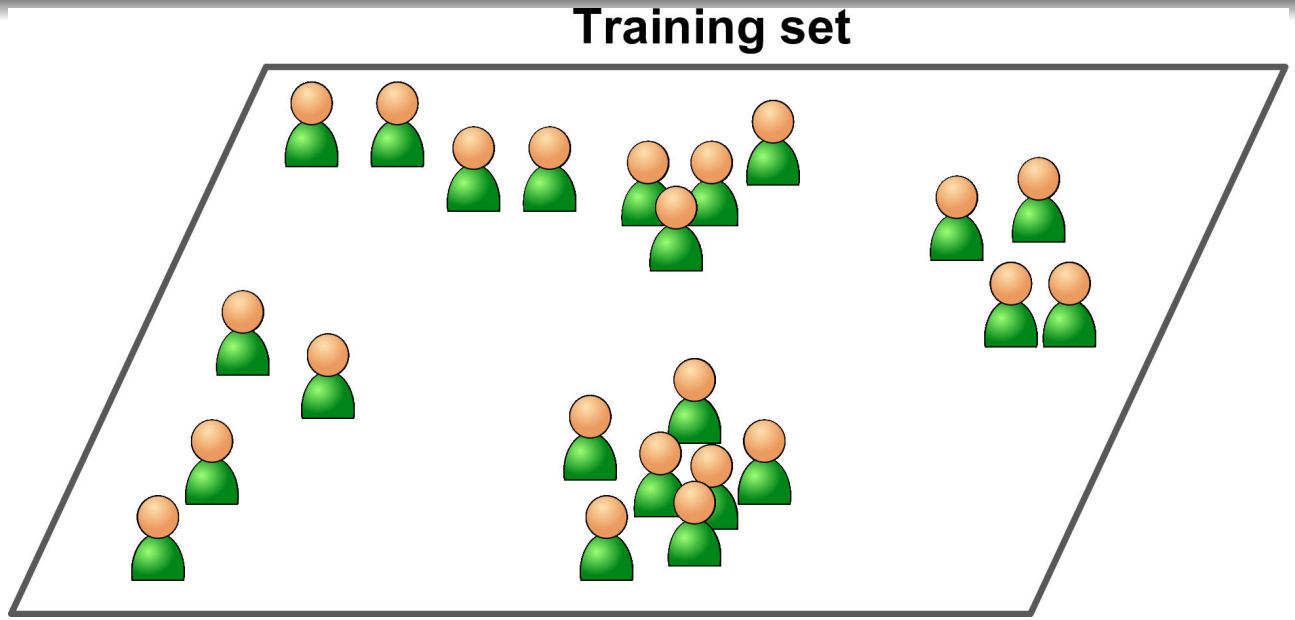

- Predict the outcome of an experiment with a finite number of possible results like
  - {spam, not spam}
  - {car, person, tree, cloud}
- Based on a training set with correct classifications known
- Typical questions
  - Is the customer credit-worthy?
  - Is the patient in the risk group for getting diabetes?
  - Will the technical quality be acceptable?

# Regression



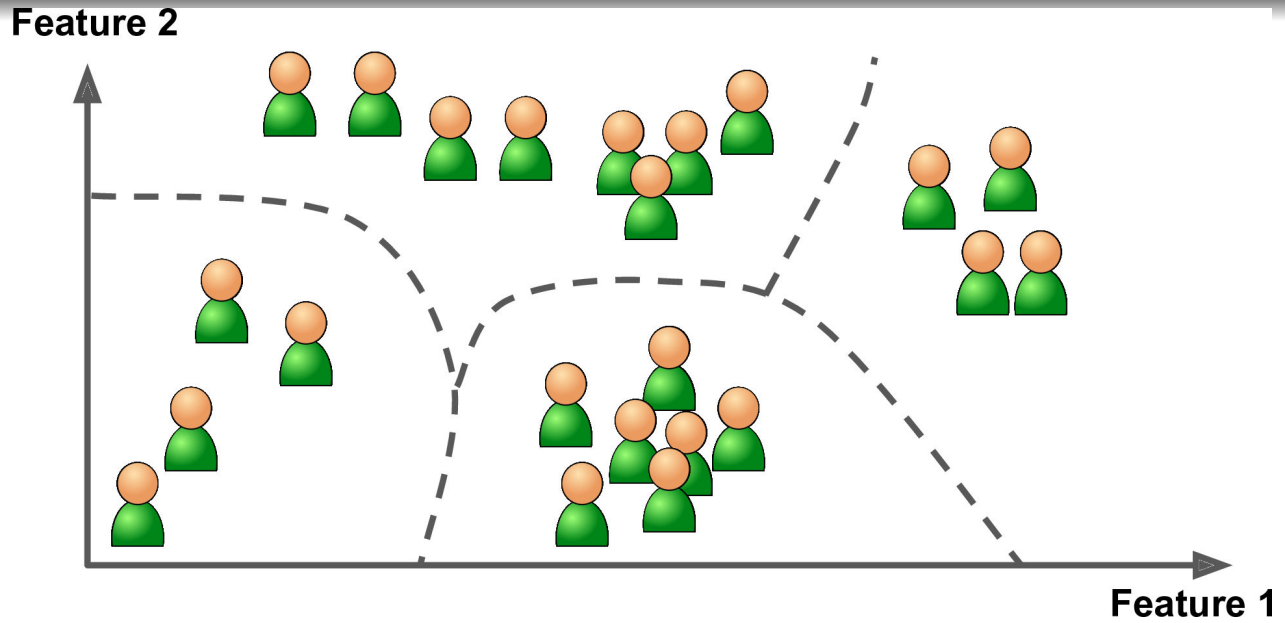

- Like classification, but numerical outcome
- Typical questions
  - How much money will customer spend for vacation next year?
  - How much will the machine's temperature change within the next cycle?
  - What is the correct price for an apartment to be sold?

# Unsupervised learning




- Tools for exploratory data analysis
- No clearly defined ground truth to be predicted
- Can we discover some type of structure in the data?
- May be used for visualizing data to understand it better
- Sometimes used for pre-processing data before applying other (supervised learning) methods
- Compared to supervised learning, challenging to evaluate whether the methods find interesting patterns

# Cluster analysis, segmentation

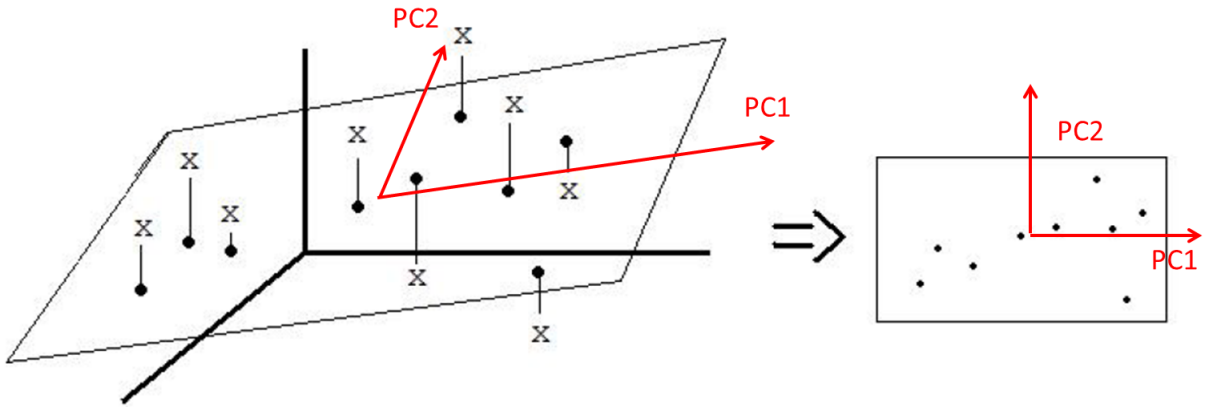



- Summarize the data to get a better overview by forming groups of similar cases (called clusters or segments)
- Instead of examining a large number of similar records, we need to inspect the group summary only
- Typical questions
  - Do my customers divide into different groups?
  - How many groups we can find from my customers?
  - What kind of operating states the machine has?

# Association analysis

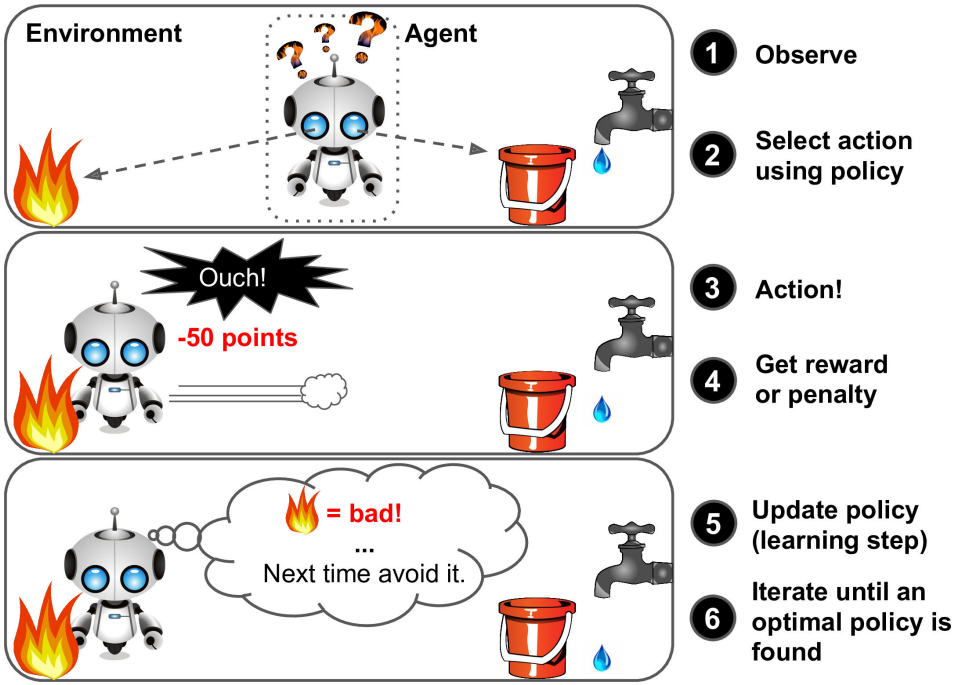
- Find any correlations or associations to better understand or describe the interdependencies of all the attributes
- The focus is on relationships between all attributes rather than focusing on a single target variable or the cases (full record)
- Typical questions
  - Which optional equipment of a car often goes together?
  - Products customers typically buy from a food market together?

# Dimensionality reduction




- How to preserve the structure of high dimensional data while projecting it to a low-dimensional representation
- Visualization: project to 2D or 3D
- Preprocessing for methods that can not handle high-dimensional data (not so important anymore)
- Also supervised variants

# Reinforcement learning




- *agent* observes *environment*, selects and performs *actions*, obtains *rewards* or *penalties*
- agent learns optimal policy to maximize rewards over time
- playing games (e.g. AlphaGo) and robotics typical applications
- outside the scope of this course

# Challenges

- Finding correct, new and interesting relationships or patterns in data is not a trivial task
- Example 1: pattern captures only random characteristics of a data set (overfitting)
  - in a lung cancer data set of 25 patients all of the 4 patients who died happened to be women
  - machine learning algorithm learns a model, that incorporates rule: "if male, will not die of lung cancer"
  - the model will not generalize well to new patients
- Example 2: discovered pattern correct, but uninteresting
  - "weight and body mass index are positively correlated"
- Example 3: Wrong causality
  - Ice cream consumption vs. number of people drowned

# Warnings about causality

- ”A statistical survey has shown that students receiving a grant perform better in their exams”
- Are the grants the reason that the students perform better, since they do not have to earn money for their studies and can spent more time for learning?
- Or do only students with better results in school or early university years receive a grant?

# Warnings about causality

- The more firemen come to extinguish the fire, the higher the damage
  - The joint cause for the number of firemen and the damage is the seriousness of the fire
- The more often people visit the doctor, the higher their chance to die
  - The joint cause for the number of visits of the doctor and the death rate is the seriousness of the disease

- Automated or semi-automated medical diagnosis
- Astronomical data analysis
- Autonomous vehicles, robotics
- Biological data analysis
- Diagnosis of machine faults
- Financial data analysis
- Machine translation, Text summarization...
- Mobile gamer churn
- Network security, intrusion detection
- Retail industry
- Speech recognition
- Targeted advertisement
- ...

# Example 1: Loan applications

- A bank has decades of experience in giving loans to customers
- Large amounts of data of customers who have or have not paid back their loans
- Is it possible to derive an algorithm from the data, that decides for new applicants whether to grant a loan or not?

# Example 1: Loan applications

- Bank needs to decide, whether to grant a loan to applicant
- What customer information (age, income, address,...) is needed?
- Is all the necessary information about customer available?
- Are the data representative?
  - e.g. we have no historical information on whether people we did not grant loan would have paid back or not
- Are the data correct?
- Legal and ethical restrictions
  - are we allowed to use variables such as gender?
  - even if omitted, may still be indirectly present (e.g. variable 'height')

# Example 2: Porcelain quality check

- A producer of porcelain wants to install an automatic quality check device that sorts out broken parts
- The produced parts are stimulated by an acoustic signal and a frequency spectrum is measured
- The parts are manually classified as "broken" or "OK"
- Is it possible to derive an algorithm from the data that classifies a new part as "broken" or "OK", given the measured frequency spectrum?

# Example 3: Recommender Systems

- Amazon, Netflix,...
- We are given a customers x products sized matrix
- Some entries filled (e.g. customer 7 gave product 4 rating $\star\star\star$)
- 99.9% of the matrix unknown (every customer rates only small fraction of products)
- How would you construct a model, that predicts ratings for a given customer ratings for products they have not rated?
- What information would you use?

# Example 3: Recommender Systems

- Idea 1: Content based filtering:
  - Customers: {age 25, female, is from Norway}...
  - Products: {length: 200 pages, genre: paranormal horror romance}...
  - Model: "young males often prefer action titles"
- Idea 2: Collaborative filtering:
  - Find customers with closest rating profiles, use their rating
  - Let's say customers A and C have the most similar rating history (compared to you)
  - Reasonable model: "A gave $\star$ and C $\star\star\star$, for you we predict the average $\star\star$"
- 2009: One million dollar prize for solving this problem (Netflix challenge)

# Conclusions

- Data analysis: how to extract knowledge, predictive models from data
- At the intersection of several disciplines
- Problem formulations ranging from supervised to unsupervised learning tasks
- Practical applications in science, industry, and commerce