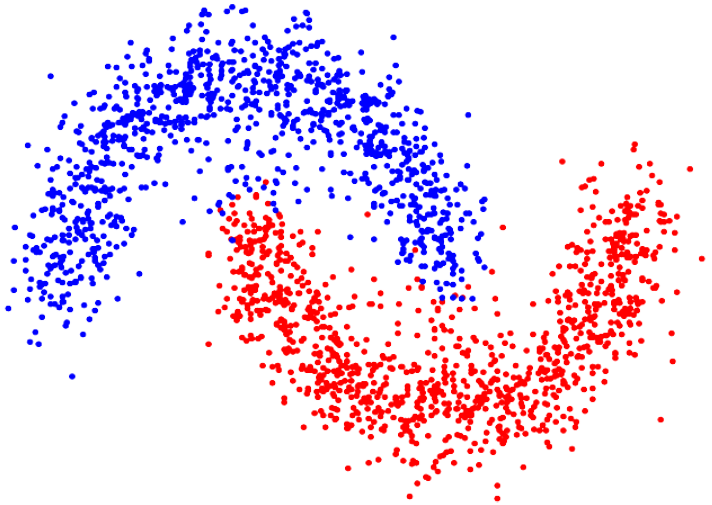
Swiss Paraplegic Research

Workshop:

Machine Learning and Prediction Modelling

Introduction to Machine Learning



Yannick Rothacher

SPF, HS2025

Swiss Paraplegic Research

Welcome to the workshop!

Lecturers:

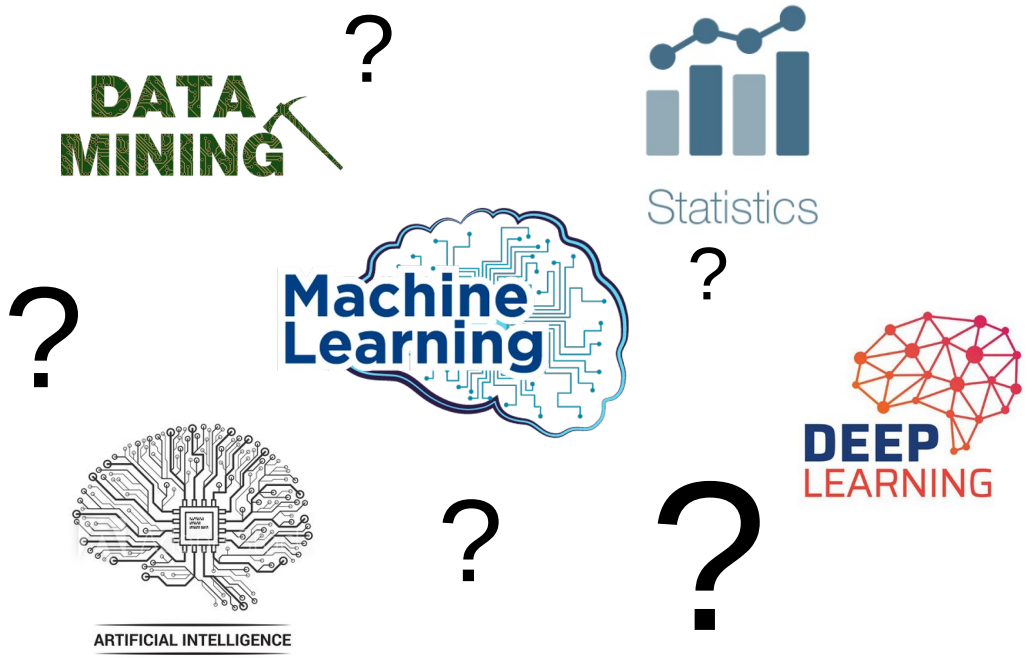
Andreas Limacher
Verantwortlicher Methodenberatung (group leader)
Functioning Information Reference Lab

Yannick Rothacher
Data Scientist
Functioning Information Reference Lab

Swiss Paraplegic Research

Course organization

- Three day course
  - Tuesday 14.Oct (Yannick)
  - Monday 20.Oct (Yannick)
  - Tuesday 21.Oct (Andreas)
- Mixture of lectures and practical exercises in R
- To obtain the credit point, active participation in the workshop is required
- Materials are available on:
  https://github.com/Swiss-Paraplegic-Research/Workshop-MachineLearning

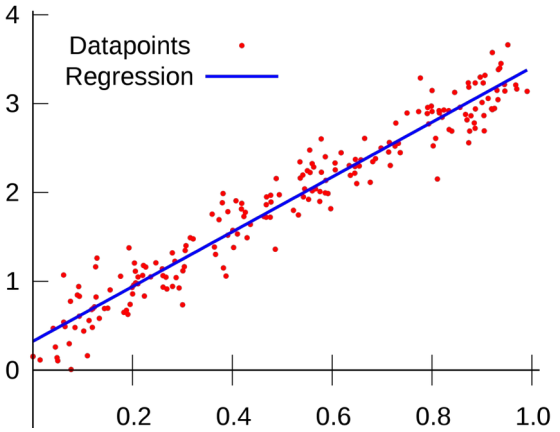


In groups:

- Why are you interested in machine learning?
- What are your expectations of this course?
- What do you associate with the term “machine learning”?

Course goals

- Give an insight into **various methods** in Machine Learning
- Teach the operating principles of the presented algorithms
- Practice the application of Machine Learning methods to data
- Deepen your skills in **R**



Swiss Paraplegic Research

Tentative timetable

Day 1 (14 Oct):

Time	Topic
09:00 – 09:20	Welcome and short intro to ML
09:20 – 10:00	PCA
10:00 – 10:45	PCA (Exercise)
10:45 – 11:00	Break
11:00 – 11:45	K-Means
11:45 – 13:00	Lunch
13:00 – 13:30	K-Means (Exercise)
13:30 – 14:00	KNN
14:00 – 14:30	KNN (Exercise)
14:30 – 14:45	Break
14:45 – 15:45	Crossvalidation

Day 2 (20 Oct):

Time	Topic
09:00 – 09:45	Decision trees
09:45 – 10:30	Decision trees (Exercise)
10:30 – 10:45	Break
10:45 – 11:45	Ensemble methods (+ Interpretability)
11:45 – 13:00	Lunch
13:00 – 13:45	Ensemble methods (Exercise)
13:45 – 14:45	Neural Networks
14:45 – 15:45	Neural Networks (Exercise)

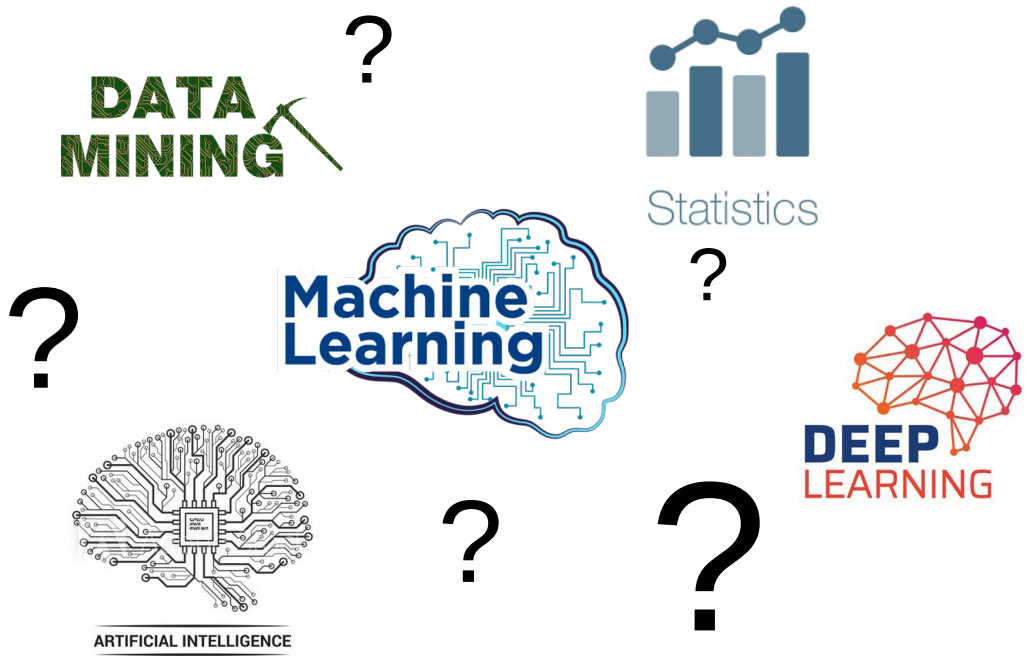
Swiss Paraplegic Research

Tentative timetable

Day 3 (21 Oct):

Time	Topic
09:00 – 09:30	Penalized regression
09:30 – 10:15	Penalized regression (Exercise)
10:15 – 10:30	Break
10:30 – 11:00	Develop and validate a prediction model
11:00 – 11:45	Develop and validate a prediction model (Exercise)
11:45 – 13:00	Lunch
13:00 – 13:15	Building a clinical score
13:15 – 13:45	Building a clinical score (Exercise)
13:45 – 15:45	Exercise, Q&A, own case, SVM

What is Machine Learning?



- Distinction from Machine Learning to other statistical methodology not always clear
- When comparing Machine Learning with “classical” statistics:
  - Statistical models are generally designed for **inference**
  - Machine Learning models are generally designed for **prediction**

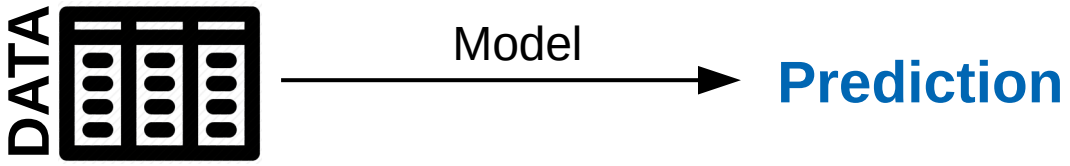
Swiss Paraplegic Research

Application of Machine Learning

Being able to **predict** certain outcomes based on data can be important in many different areas in **research and industry**

Examples:

- Predict the winner of a basketball game
- Predict the weather of tomorrow
- Predict whether a medical scan shows an image of a tumor
- Predict whether an email is spam or not
- Predict how likely a person is about to develop depression

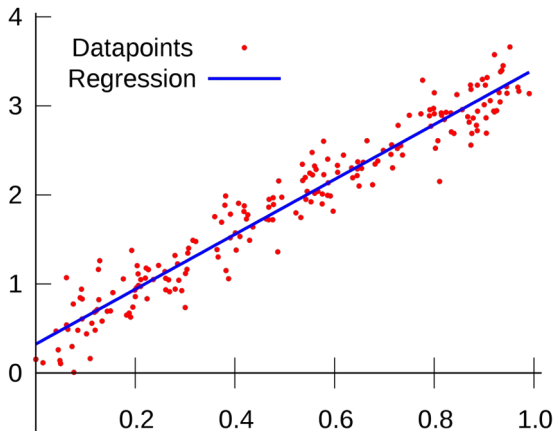


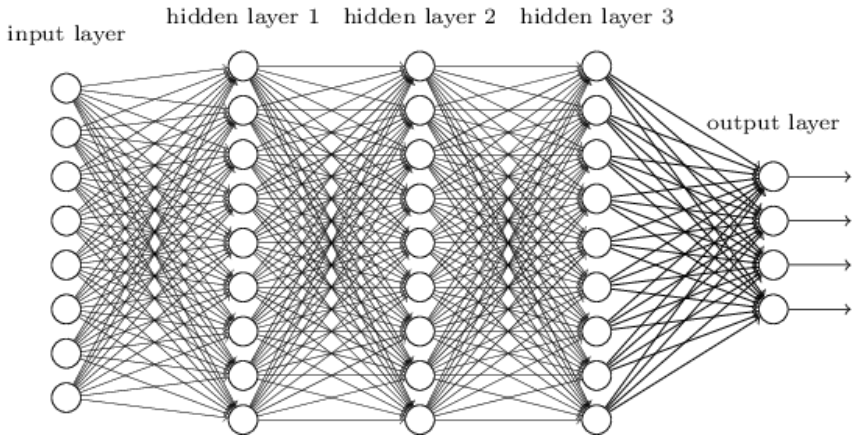
In all cases: **Predictions are based on data !**

Swiss Paraplegic Research

Prediction models don’t have to be complicated

- Simple linear regression can also be used to predict values of new observations





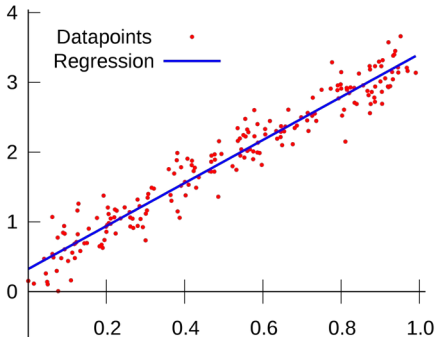
- However, sometimes statistical models have limited prediction accuracy, but allow **inference about the relation** between predictors and target variables (e.g. showing a significant influence of a treatment).
- In many Machine Learning models, the prediction accuracy is very good but it is difficult to interpret the variables’ relations (e.g. neural network)

Swiss Paraplegic Research

Application of Machine Learning

- Again: In general one tries to predict a target variable based on predictor variables

target variable ~ predictor variables
y ~ X

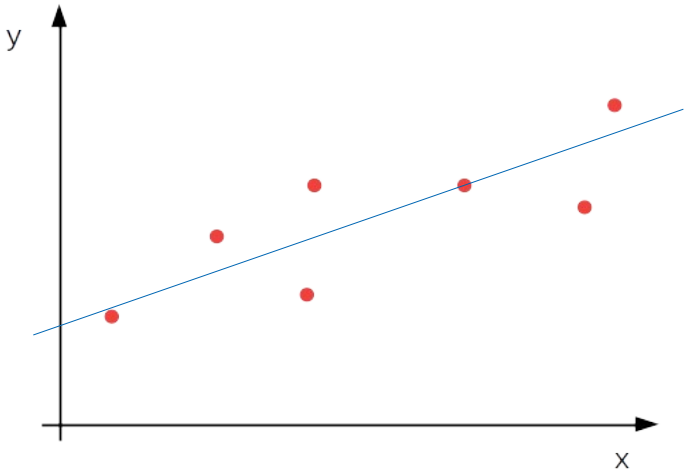


- Target variable is usually a category or a number
  - Y is category: “Classification”
  - Y is metric: “Regression”
- In real-life data, there are often many predictor variables (genetic data: up to 10’000 predictors)
- Can even be n << p (much more variables (p) than data points (n))
- This case can be difficult to handle with conventional methods (for example linear regression)

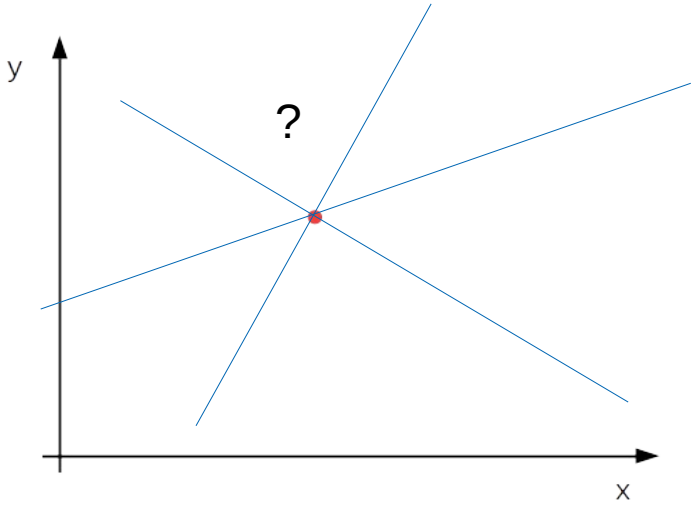
Swiss Paraplegic Research

Challenges of high-dimensional data

- For example linear regression only works for n > p :



n > p



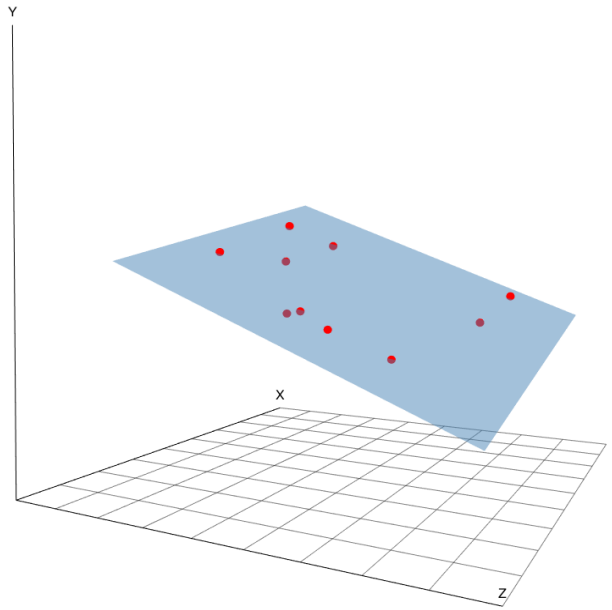
n = 1

- We need methods for situations with n < p
- Machine Learning methods are usually able to handle n < p situations

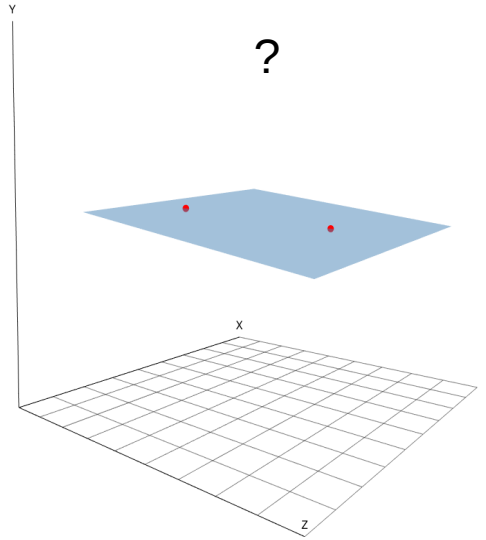
Swiss Paraplegic Research

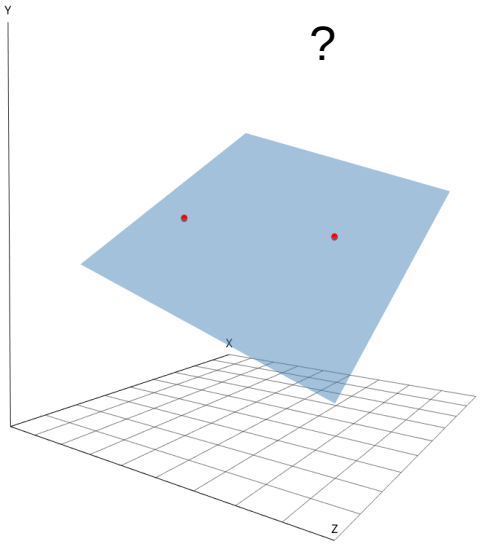
Challenges of high-dimensional data

- For example linear regression only works for n > p :



n > p



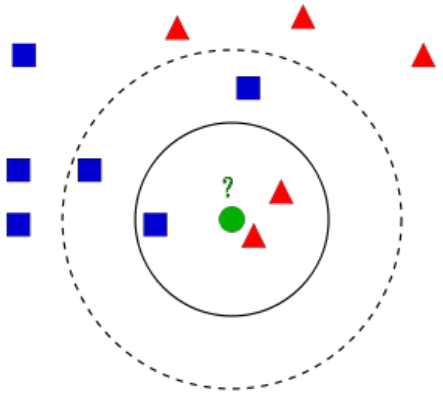


n = 2

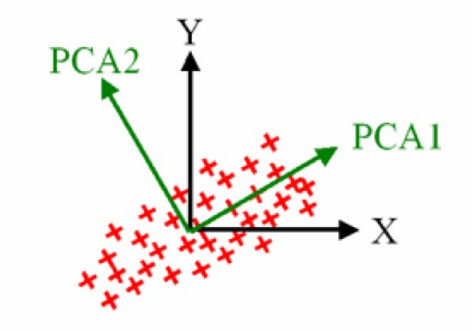
- We need methods for situations with n < p
- Machine Learning methods are usually able to handle n < p situations

Swiss Paraplegic Research

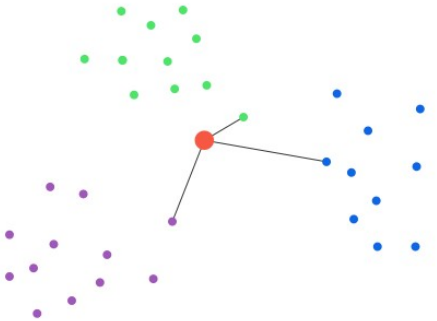
Outlook: Machine Learning methods



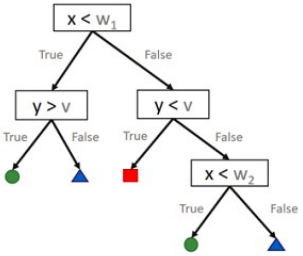
K-nearest neighbor



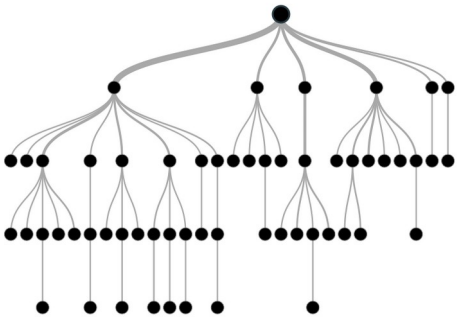
Principal Component Analysis



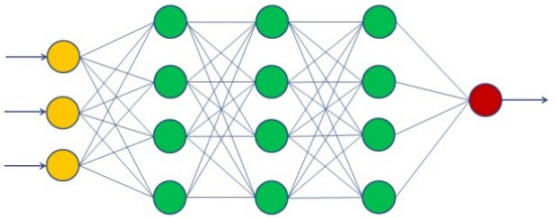
K-means clustering



Decision trees



Random Forest



Neural Networks