AdvancedAnalytics
.Academy

www.advancedanalytics.academy

Introduction Advanced Analytics & Machine Learning

Stuttgart, 20th of February 2020

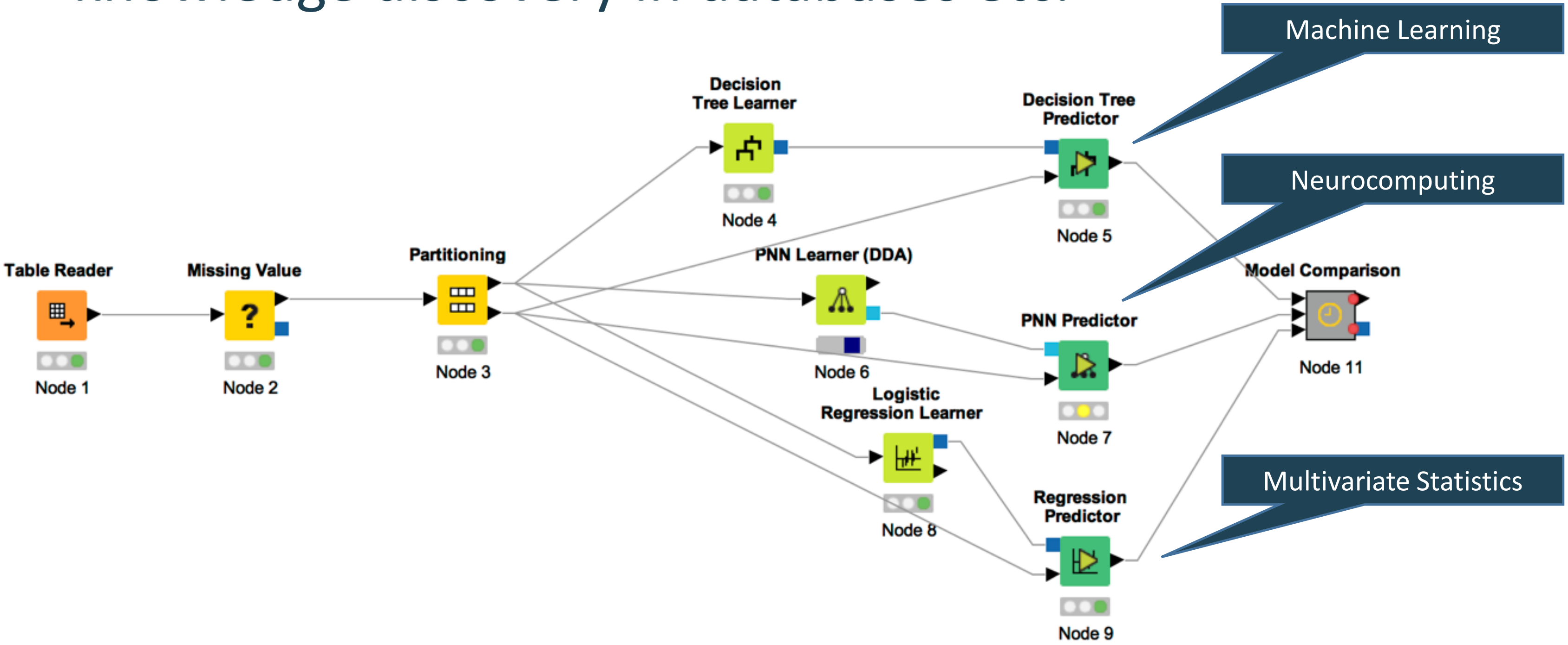
Michael Höhn

1.

Introduction to Advanced Analytics

1. What is Advanced Analytics?

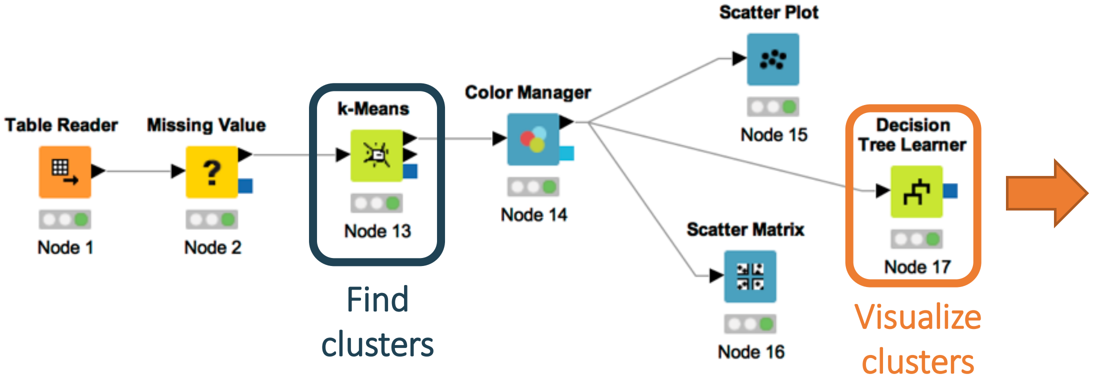
- ...is a **combination** of different academic fields like multivariate statistics, artificial intelligence, machine learning, pattern recognition, neurocomputing, knowledge discovery in databases etc.



1. What is Advanced Analytics?

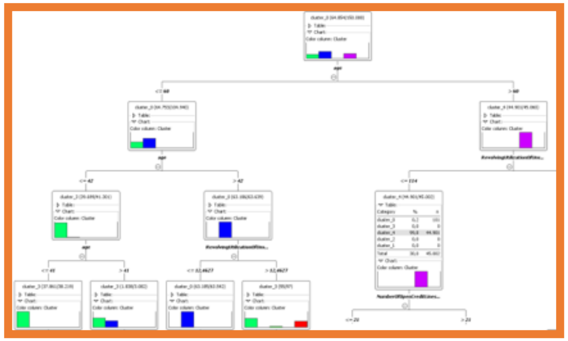
- …is the **process** of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or identify business critical hidden patterns.

Combination of Cluster Analysis & Decision Trees

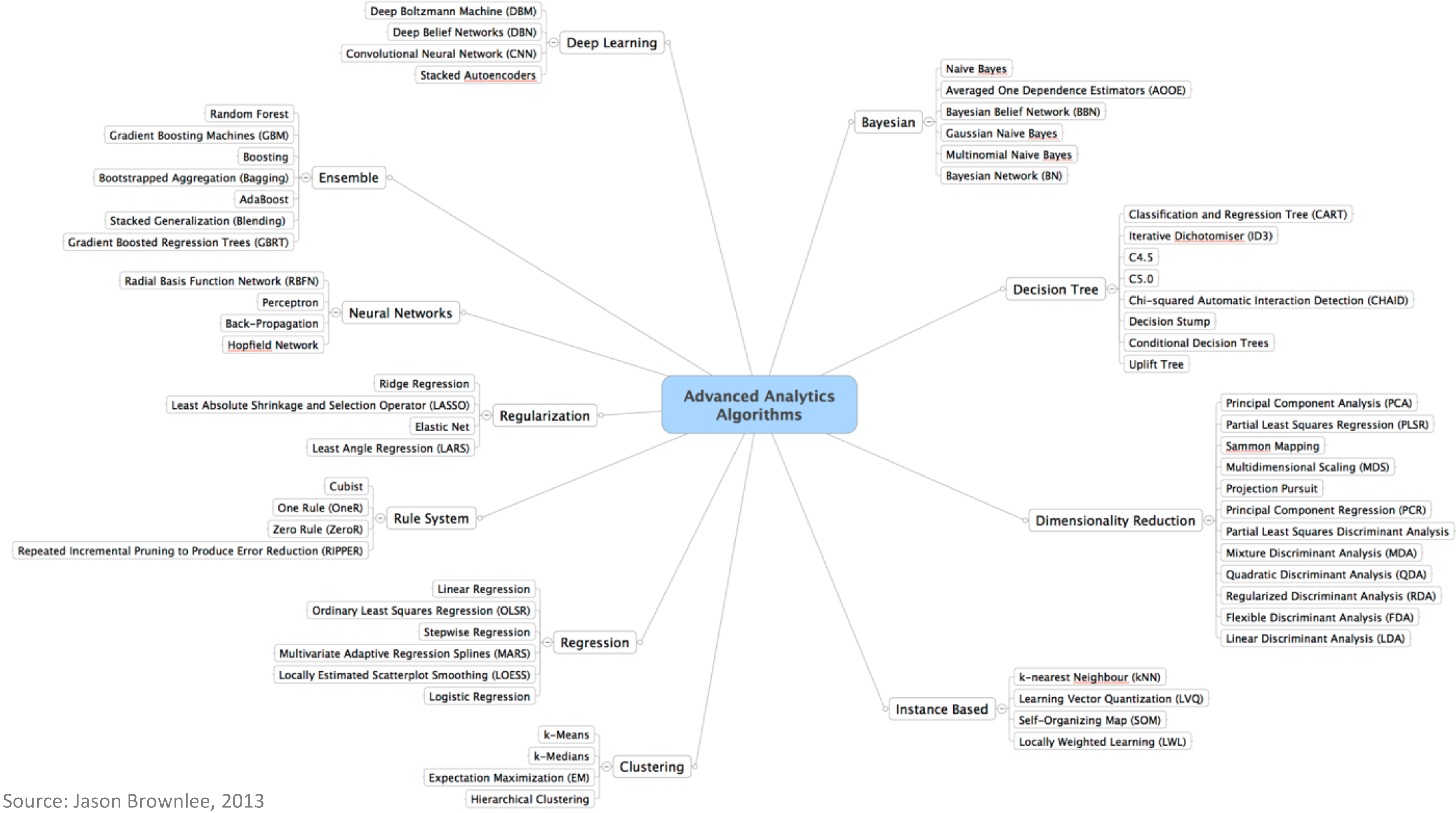


Find clusters

Visualize clusters



1. An overview of Advanced Analytics Algorithms



- Advanced Analytics Algorithms
  - Deep Learning
    - Deep Boltzmann Machine (DBM)
    - Deep Belief Networks (DBN)
    - Convolutional Neural Network (CNN)
    - Stacked Autoencoders
  - Ensemble
    - Random Forest
    - Gradient Boosting Machines (GBM)
    - Boosting
    - Bootstrapped Aggregation (Bagging)
    - AdaBoost
    - Stacked Generalization (Blending)
    - Gradient Boosted Regression Trees (GBRT)
  - Neural Networks
    - Radial Basis Function Network (RBFN)
    - Perceptron
    - Back-Propagation
    - Hopfield Network
  - Regularization
    - Ridge Regression
    - Least Absolute Shrinkage and Selection Operator (LASSO)
    - Elastic Net
    - Least Angle Regression (LARS)
  - Rule System
    - Cubist
    - One Rule (OneR)
    - Zero Rule (ZeroR)
    - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
  - Regression
    - Linear Regression
    - Ordinary Least Squares Regression (OLSR)
    - Stepwise Regression
    - Multivariate Adaptive Regression Splines (MARS)
    - Locally Estimated Scatterplot Smoothing (LOESS)
    - Logistic Regression
  - Clustering
    - k-Means
    - k-Medians
    - Expectation Maximization (EM)
    - Hierarchical Clustering
  - Bayesian
    - Naive Bayes
    - Averaged One Dependence Estimators (AOOE)
    - Bayesian Belief Network (BBN)
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Bayesian Network (BN)
  - Decision Tree
    - Classification and Regression Tree (CART)
    - Iterative Dichotomiser (ID3)
    - C4.5
    - C5.0
    - Chi-squared Automatic Interaction Detection (CHAID)
    - Decision Stump
    - Conditional Decision Trees
    - Uplift Tree
  - Dimensionality Reduction
    - Principal Component Analysis (PCA)
    - Partial Least Squares Regression (PLSR)
    - Sammon Mapping
    - Multidimensional Scaling (MDS)
    - Projection Pursuit
    - Principal Component Regression (PCR)
    - Partial Least Squares Discriminant Analysis
    - Mixture Discriminant Analysis (MDA)
    - Quadratic Discriminant Analysis (QDA)
    - Regularized Discriminant Analysis (RDA)
    - Flexible Discriminant Analysis (FDA)
    - Linear Discriminant Analysis (LDA)
  - Instance Based
    - k-nearest Neighbour (kNN)
    - Learning Vector Quantization (LVQ)
    - Self-Organizing Map (SOM)
    - Locally Weighted Learning (LWL)

Source: Jason Brownlee, 2013

2.

Fields of Application (Example Use Cases)

2. Fields of Application (examples)

Customer Segmentation

Customer segmentation, also referred to as market segmentation, is the process of finding **homogenous** sub-groups within a **heterogeneous** aggregate customer base.

Propensity Modeling (Churn, Next Best Offer etc)

1. Predict customer churn by assessing their propensity of risk to churn.
2. Predict customer need and behavior by assessing their propensity to buy a product.

Association Analysis (eg. Market Basket Analysis)

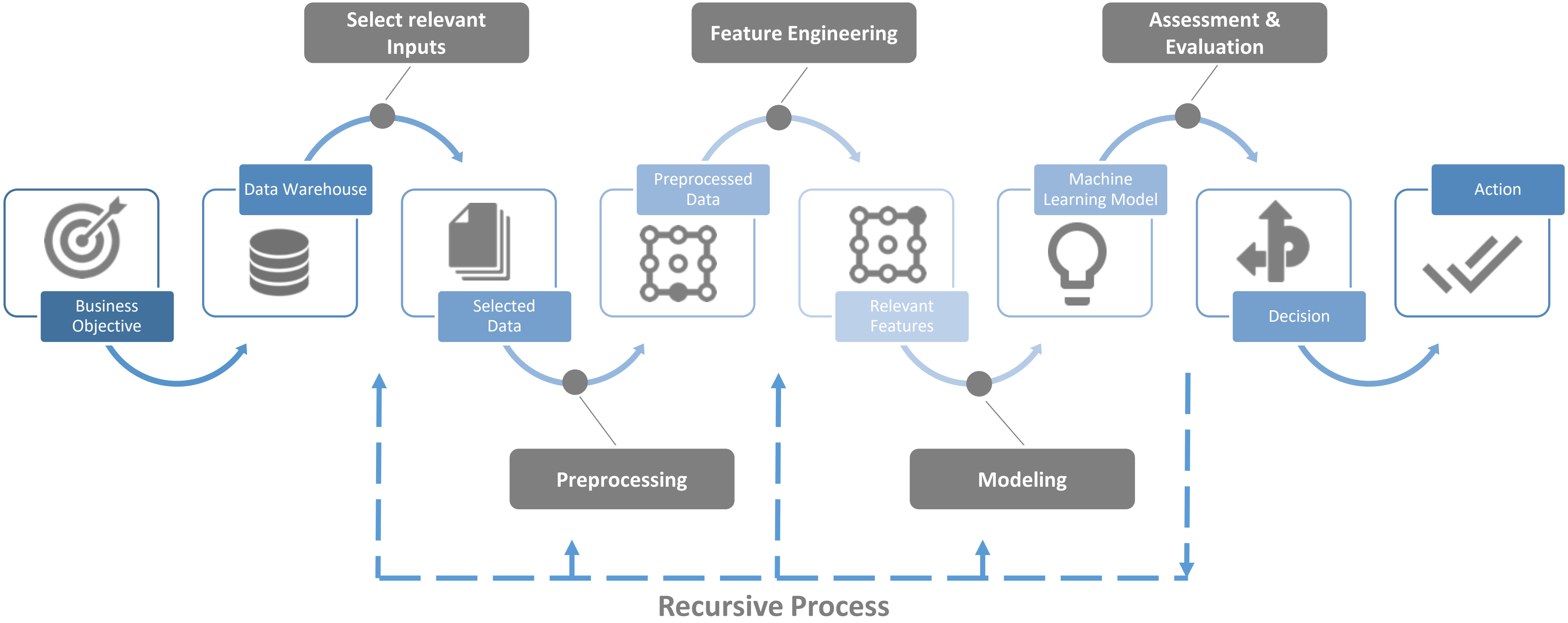
Association rules are employed today in many application areas including market basket analysis, web usage mining, intrusion detection and bioinformatics.

Fraud Detection & Money Laundering

Anticipate illegal or suspicious activities and transactions – such as identity theft, insurance fraud and money laundering by applying predictive analytics methods.

3.
Advanced Analytics Processes

3. A typical Advanced Analytics Process



Select relevant Inputs

Feature Engineering

Assessment & Evaluation

Business Objective

Data Warehouse

Selected Data

Preprocessed Data

Relevant Features

Machine Learning Model

Decision

Action

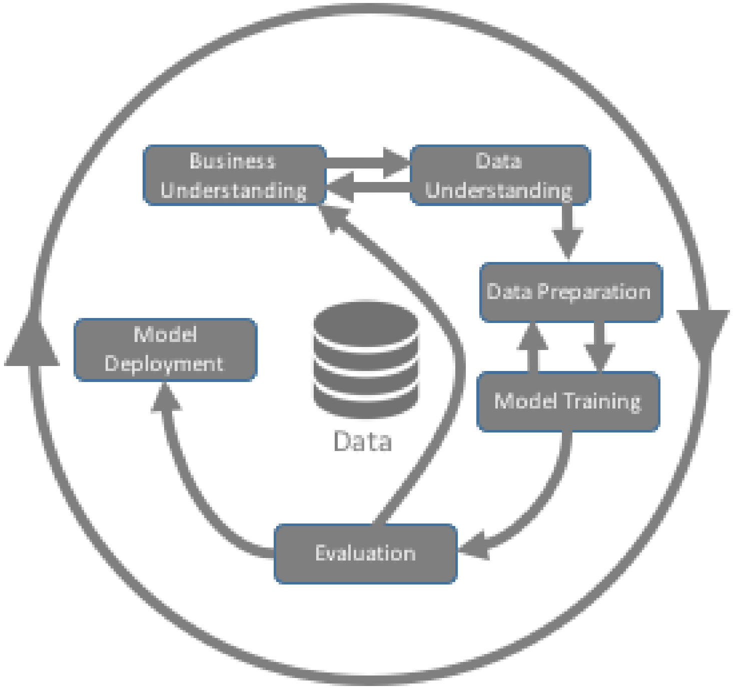
Preprocessing

Modeling

Recursive Process

3. CRISP-DM

- CRISP-DM = Cross Industry Standard Process for Data Mining.
- A data mining process model that describes commonly used approaches that expert data miners use to tackle problems.
- CRISP-DM breaks the process of data mining into six phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Model Training
  - Evaluation
  - Model Application



4.

Categorization of Advanced Analytics Algorithms

4. Categorization of Advanced Analytics Algorithms

Supervised Learning
(Predictions & Classifications)

1. Decision Trees
2. Logistic Regression
3. Neural Networks
4. Bayes Classifier
5. Support Vector Machines
6. Ensemble Models

Unsupervised Learning
(Structure Discovery)

1. Cluster Analysis
2. Self-Organizing Maps (SOM)
3. Association Algorithms
4. Sequence Analysis

Semi-Supervised Learning

1. Active Learning
2. Generative Models
3. Low-density Separation
4. Graph-Based Algorithms
5. Multiview Algorithms

5.

Challenges in Advanced Analytics

5. Prediction Types for Predictive Modeling

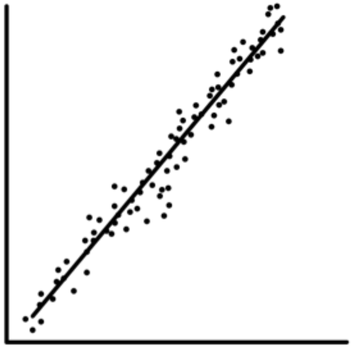
Training Data Observations

categorical or numeric input and target measurements

INPUT VARIABLES				TARGET
■	■	■	■	🟧
■	■	■	■	🟧
■	■	■	■	🟧
■	■	■	■	🟧
■	■	■	■	🟧

Predictive Model

a formal representation of the input and target association



5. Prediction Types for Predictive Modeling

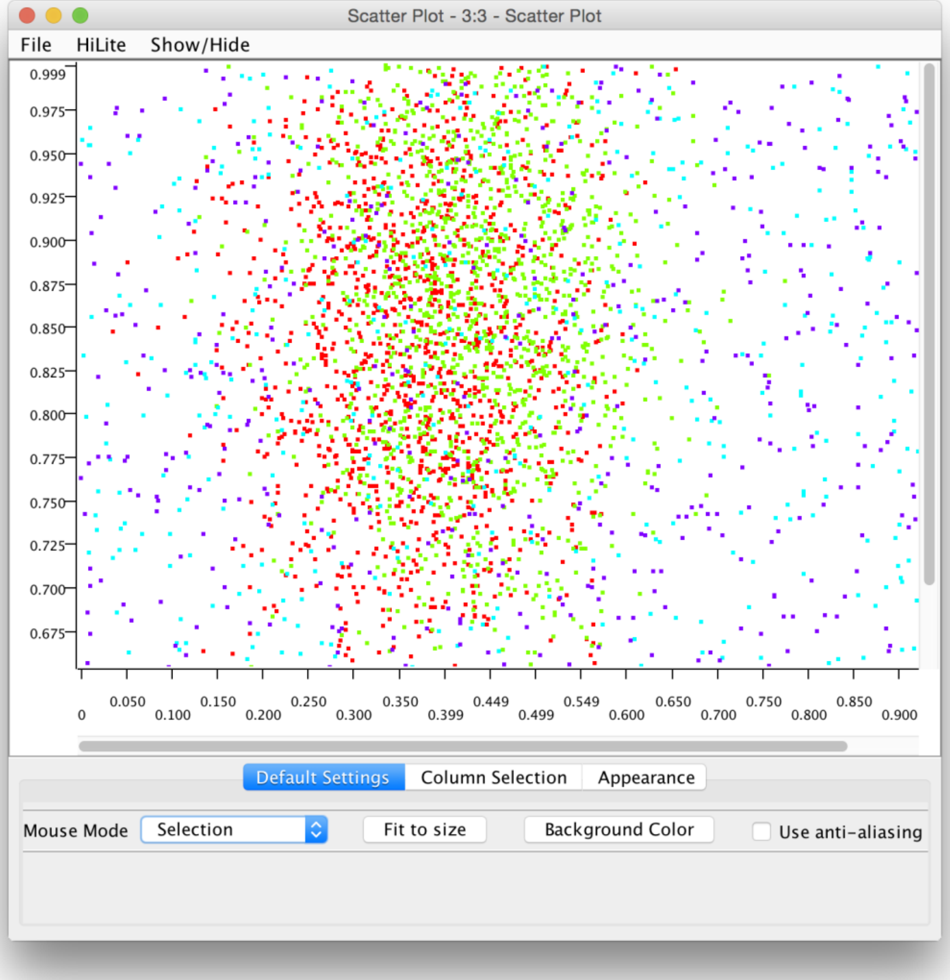
INPUT VARIABLES				PREDICTION
■	■	■	■	🟩
■	■	■	■	🟩
■	■	■	■	🟩
■	■	■	■	🟩
■	■	■	■	🟩
■	■	■	■	🟩
■	■	■	■	🟩

Classifications
Rankings
Estimates

Classifications	Customer cancels contract (YES or NO)
Rankings	Customer A has a higher propensity to churn than customer B
Estimates	Customer A has a churn-probability of 75%.

5. Dimensionality Reduction

- Identify and reject **redundant** and **irrelevant** input variables



Mouse Mode: Selection

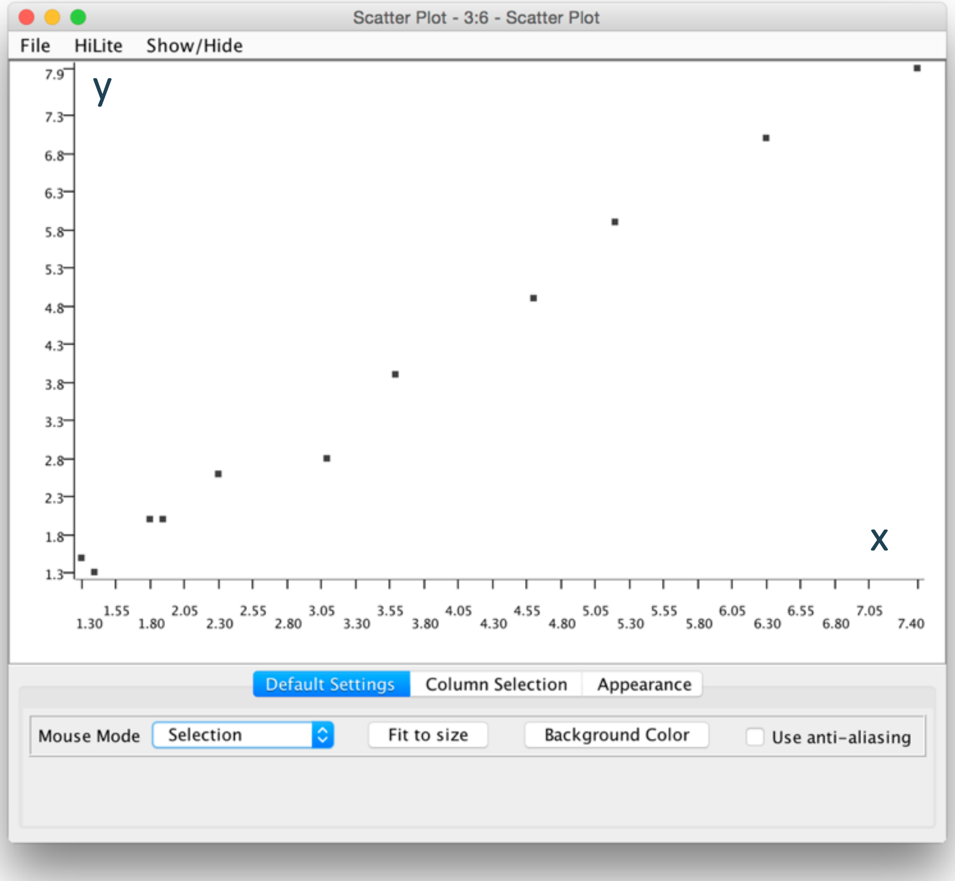
□ Use anti-aliasing

This educational material was produced for the course at KNIME Spring Summit in Berlin on 19th of March 2019. The copyright is with AdvancedAnalytics.Academy GmbH, Germany. Do not copy or distribute.

5. Dimensionality Reduction

- **Redundancy**

  Variable y contains the same information as variable x.

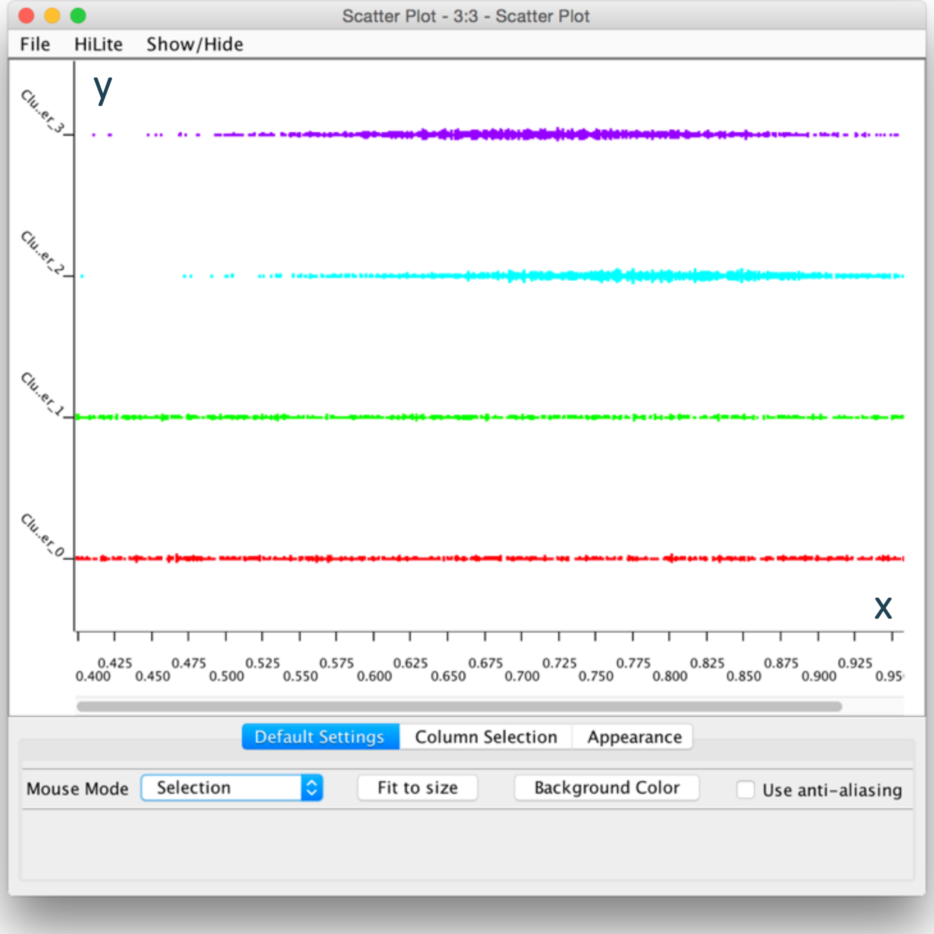


Mouse Mode: Selection

5. Dimensionality Reduction

- **Irrelevancy**

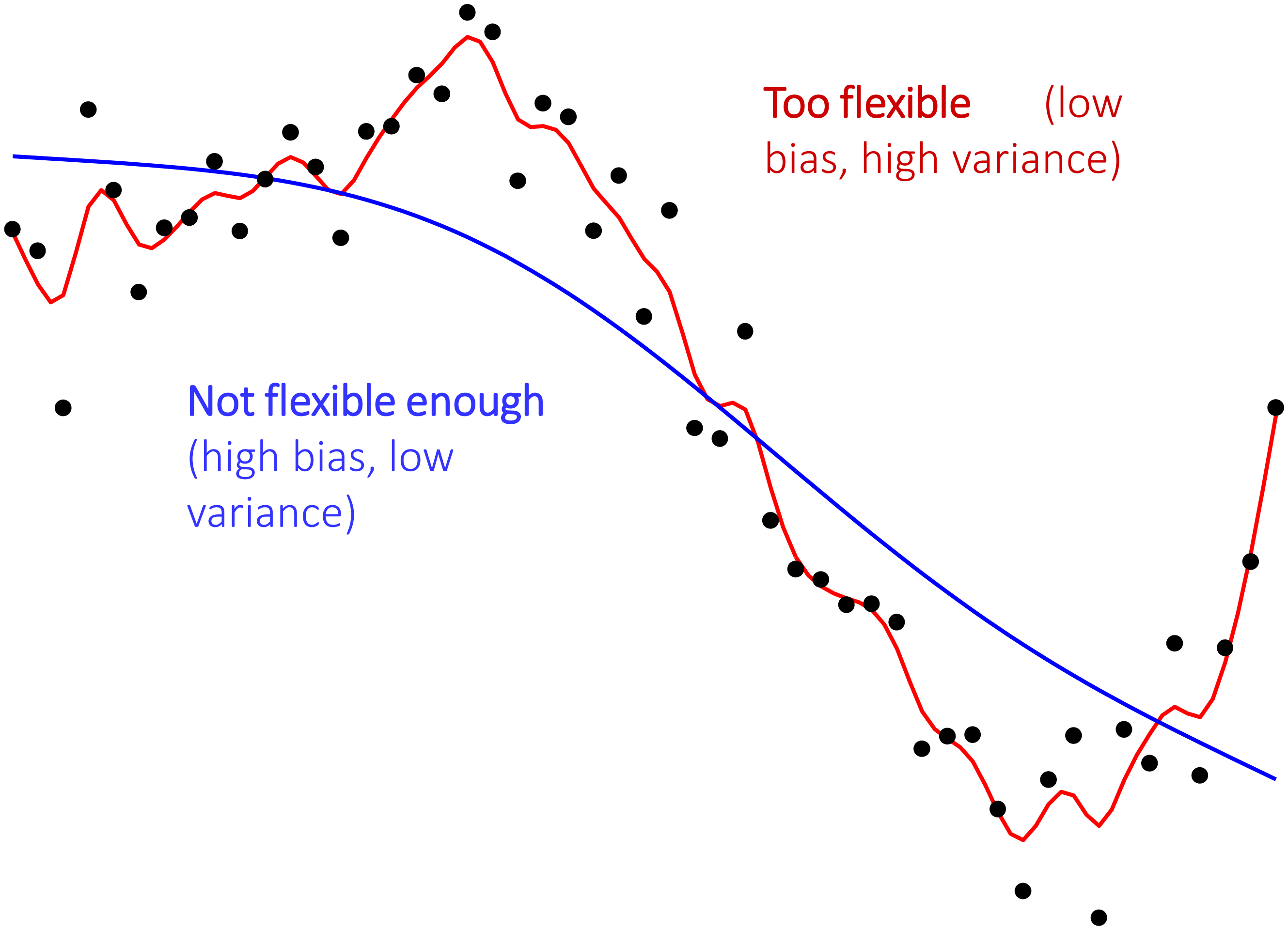
  Predictions change with variable y but not with Variable x.



Mouse Mode: Selection

☐ Use anti-aliasing

5. Optimization of model complexity



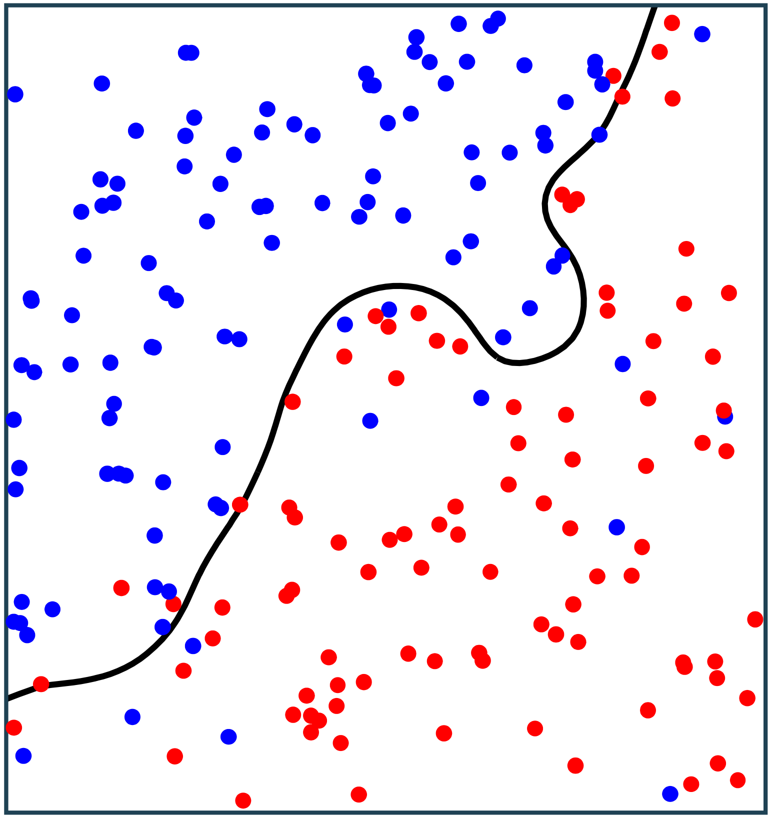
Too flexible (low bias, high variance)

Not flexible enough (high bias, low variance)

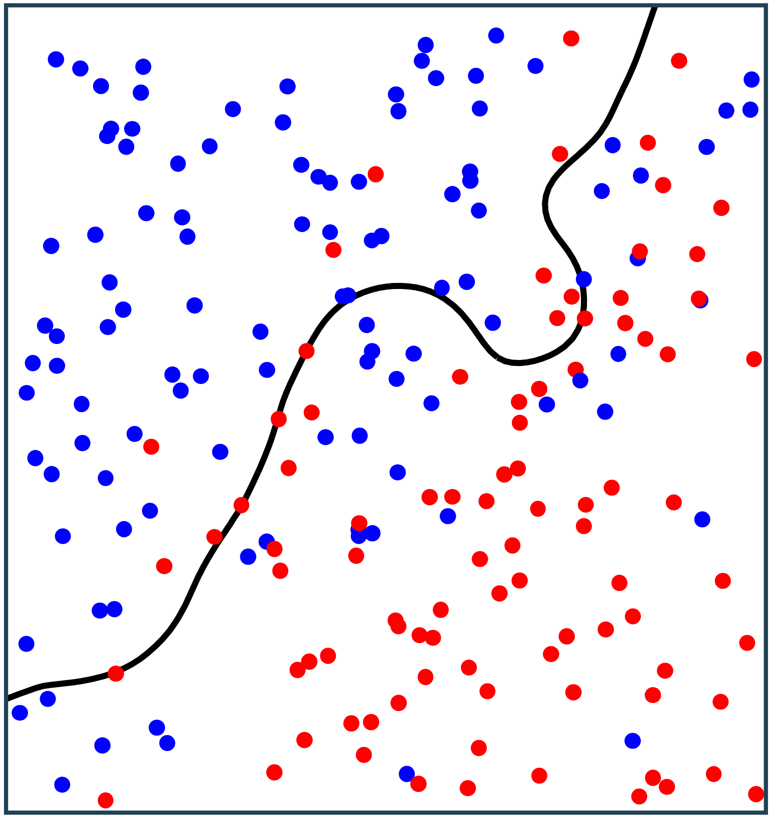
5. Optimization of model complexity

Overfitting (complex model)

TRAINING DATA

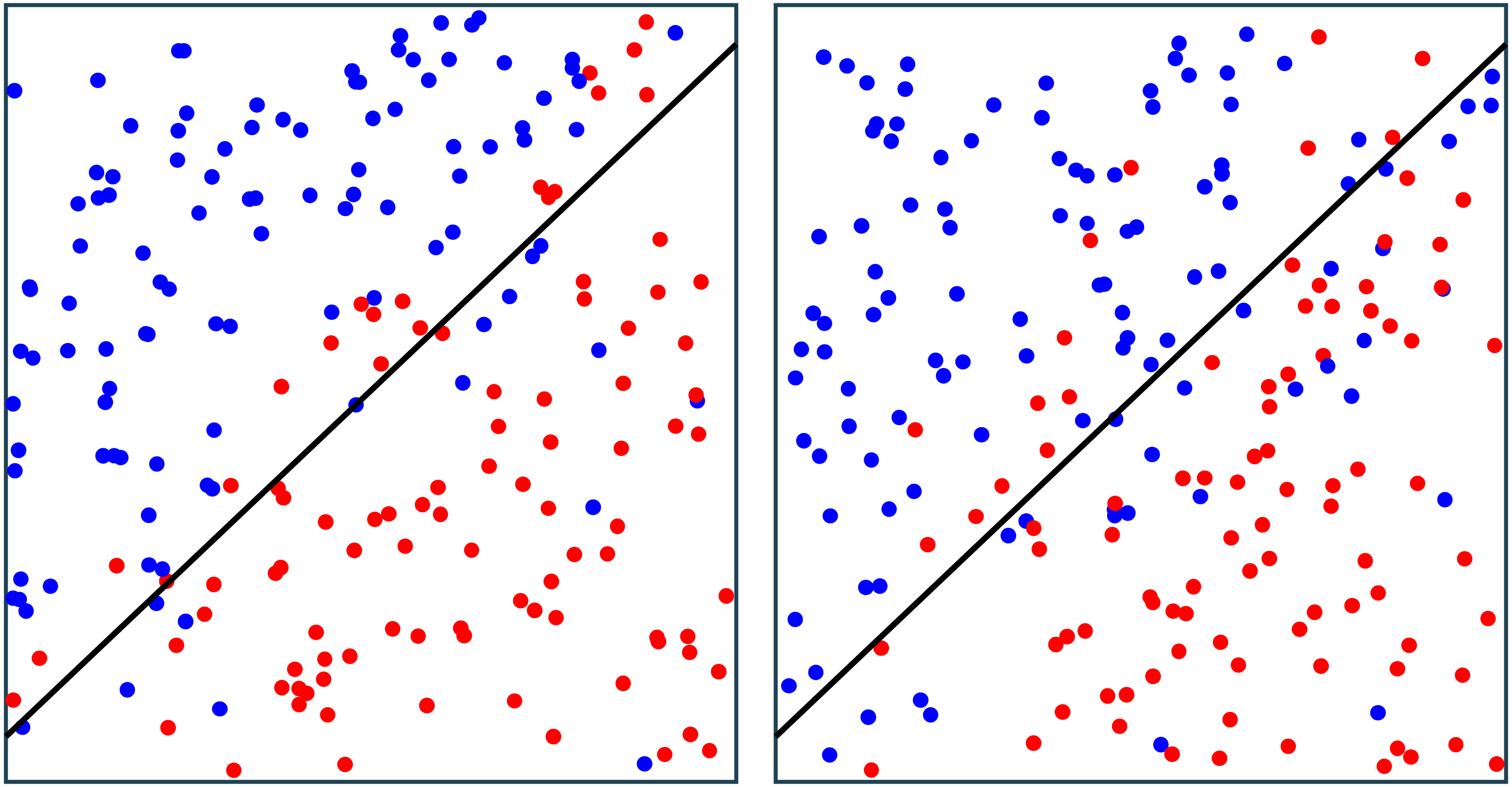


TEST DATA



5. Optimization of model complexity

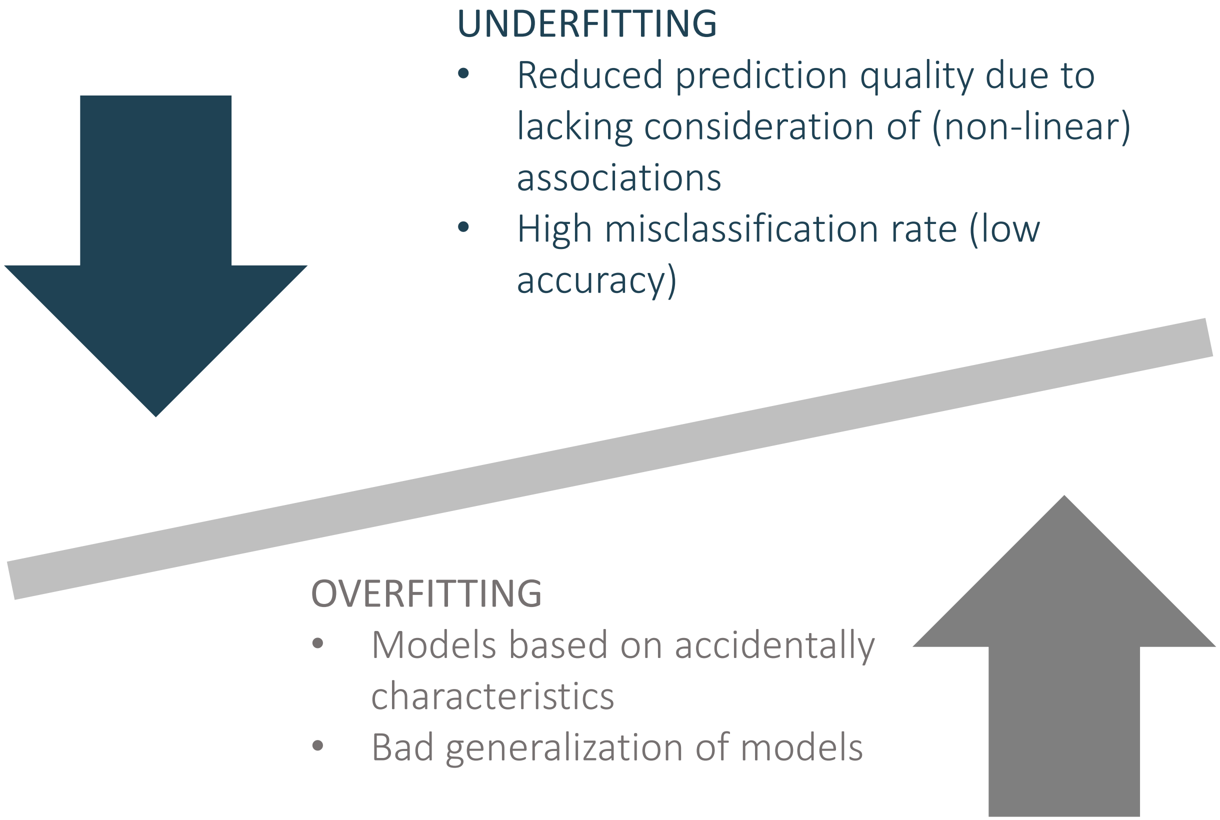
Better fit with simpler model approach



TRAINING DATA

TEST DATA

5. Optimization of model complexity



UNDERFITTING

- Reduced prediction quality due to lacking consideration of (non-linear) associations
- High misclassification rate (low accuracy)

OVERFITTING

- Models based on accidentally characteristics
- Bad generalization of models

! Keep models as simple as possible and as complex as necessary !

5. Generalization of models

- Challenge
  - Build high-performance predictive models that generalize well to new data!
  - Extend the half-life validity of predictive models even if the are applied to unknown data!
- Out-of-sample
  - *Constraint*: the model's input training data set is based on a sample (random or stratified sample).
  - Being "out-of-sample-proofed" the model shows the same predictive power (assessment quality values) as being trained on complete basic population.
- Out-of-time
  - *Constraint*: the model's input data set contains observations of a specific time window (e.g. observations of the last 18 months).
  - Being "out-of-time-proofed" the model shows the same predictive power (assessment quality values) as being applied to future observations.