Hochschule Hof
University of Applied Sciences

Introduction to Machine Learning

Applied Deep Learning

Learning Goal

- ML-based approach to problem solving
- Regression
- Classification
- Clustering
- Quality Metrics

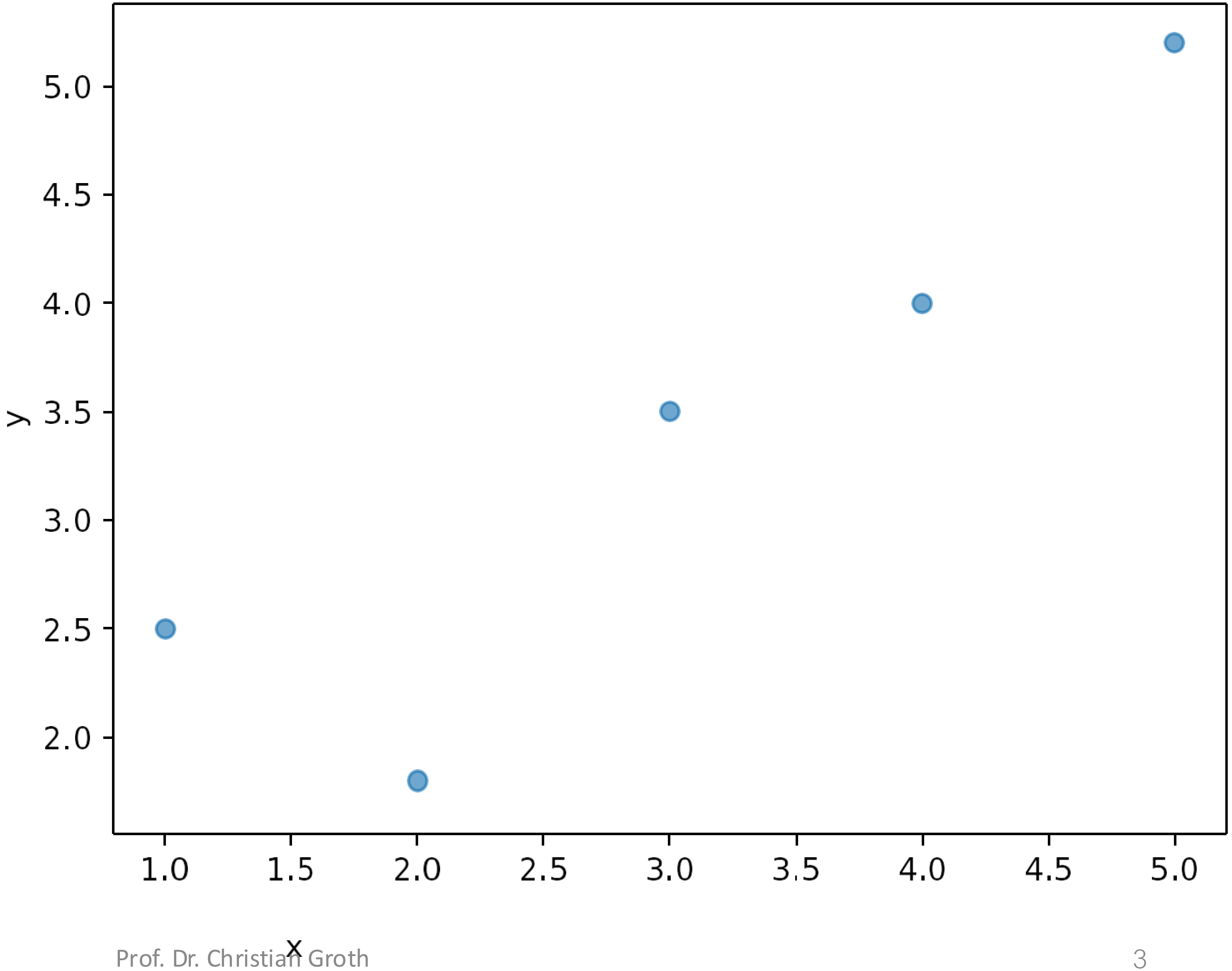
Regression

Assumption:

- Data collected from observation (e.g. from an experiment)
- Data set consists of several data points
- Each data point has two features (x and y)

Target:

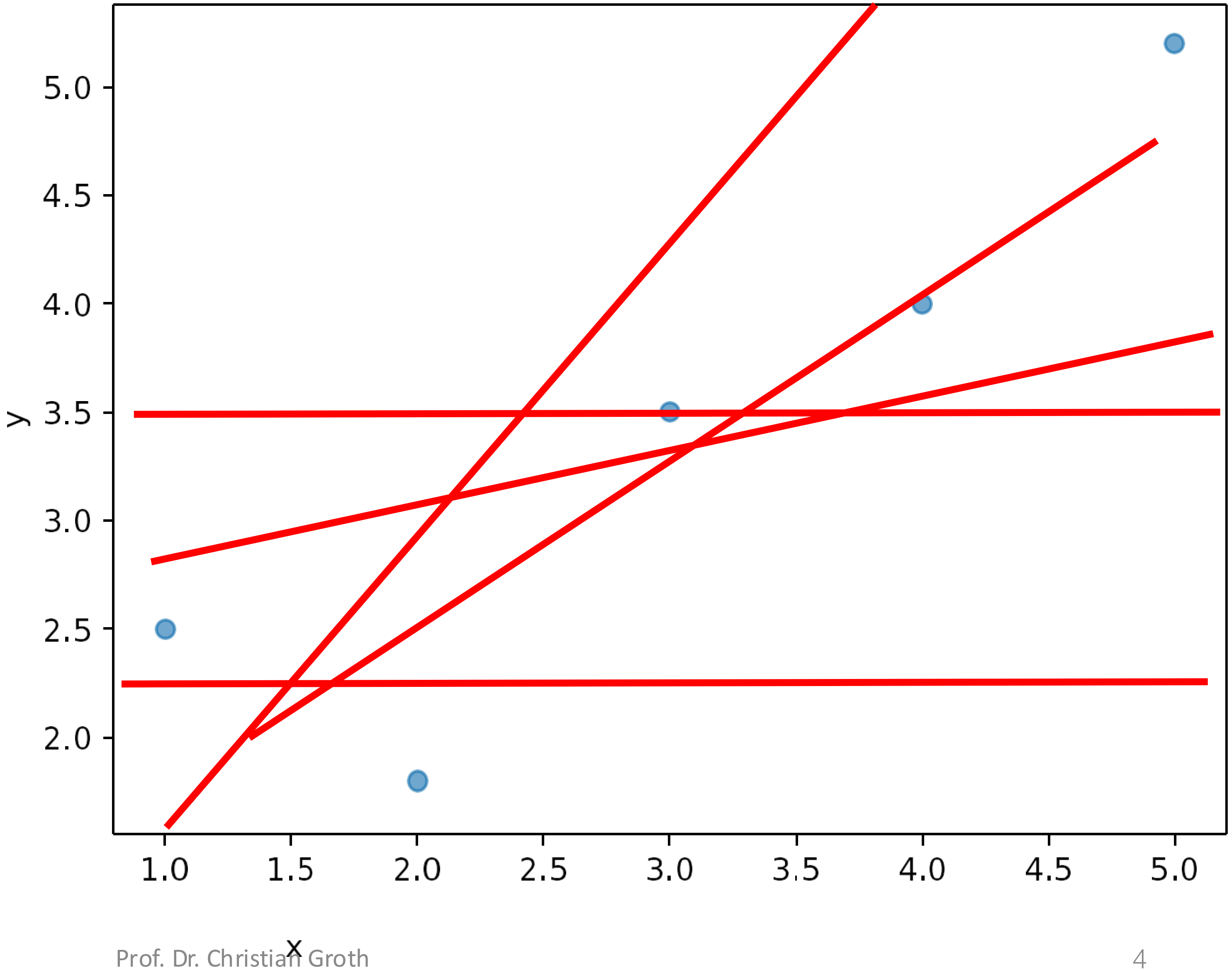
- Find out the relationship between x and y
- Predict values



Regression

Idea:

- Attempts to generate a model that depicts the context
- Simplest relation: linear
- Which is the right straight line?



Regression

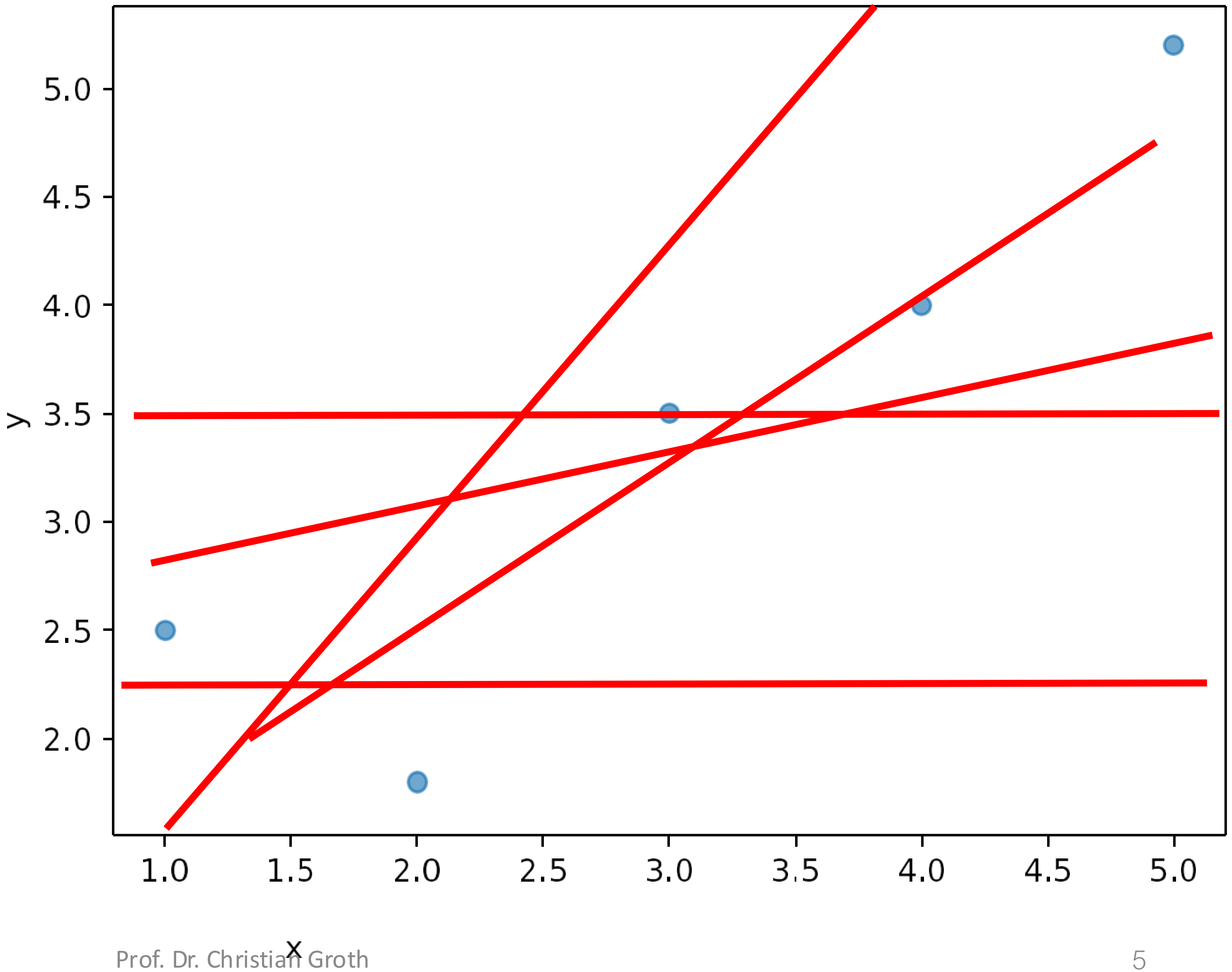
Hochschule Hof
University of Applied Sciences

Idea:

- Attempts to generate a model that depicts the context
- Simplest relation: linear
- Which is the right straight line?

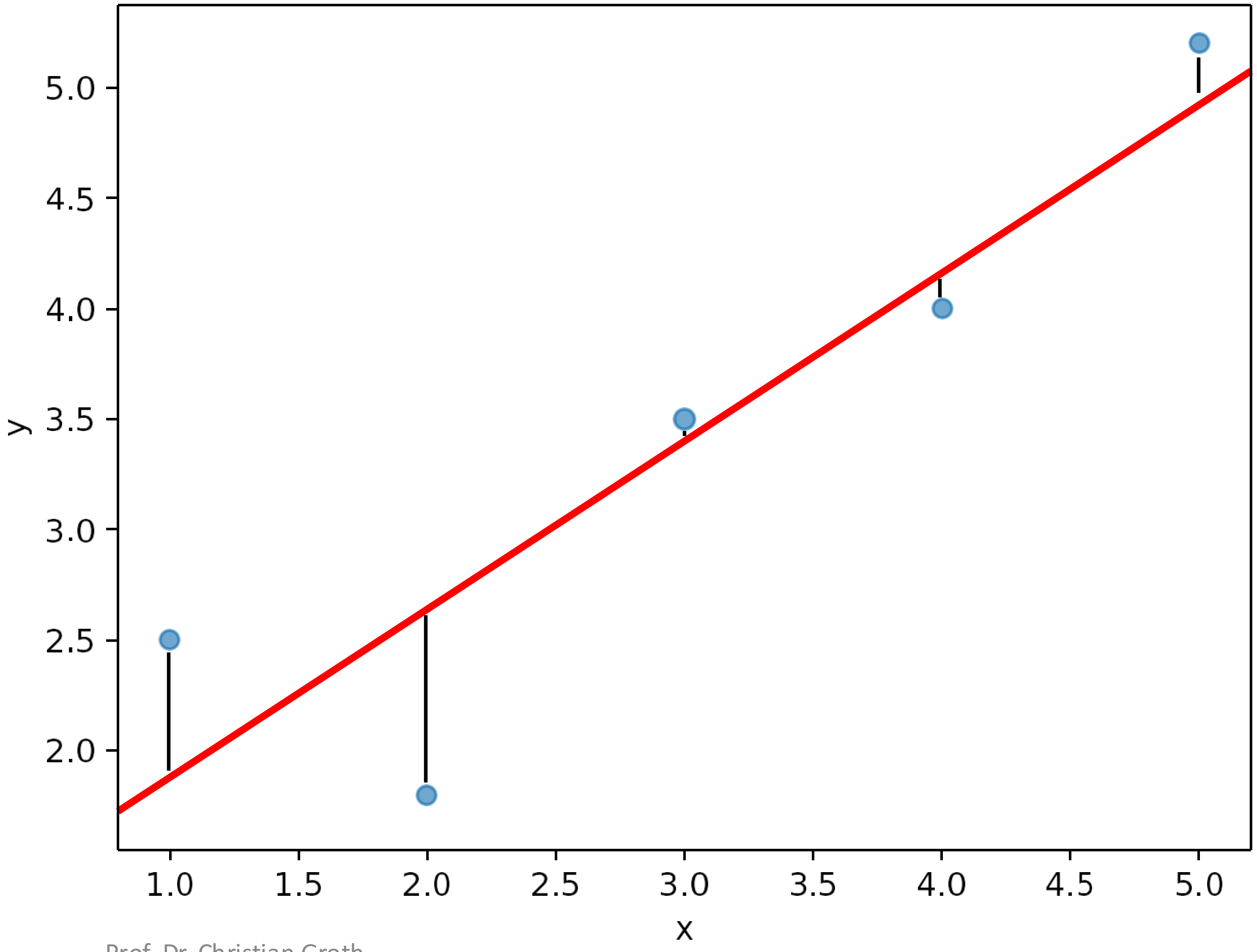
→Model that best represents our data

→Model with the least error



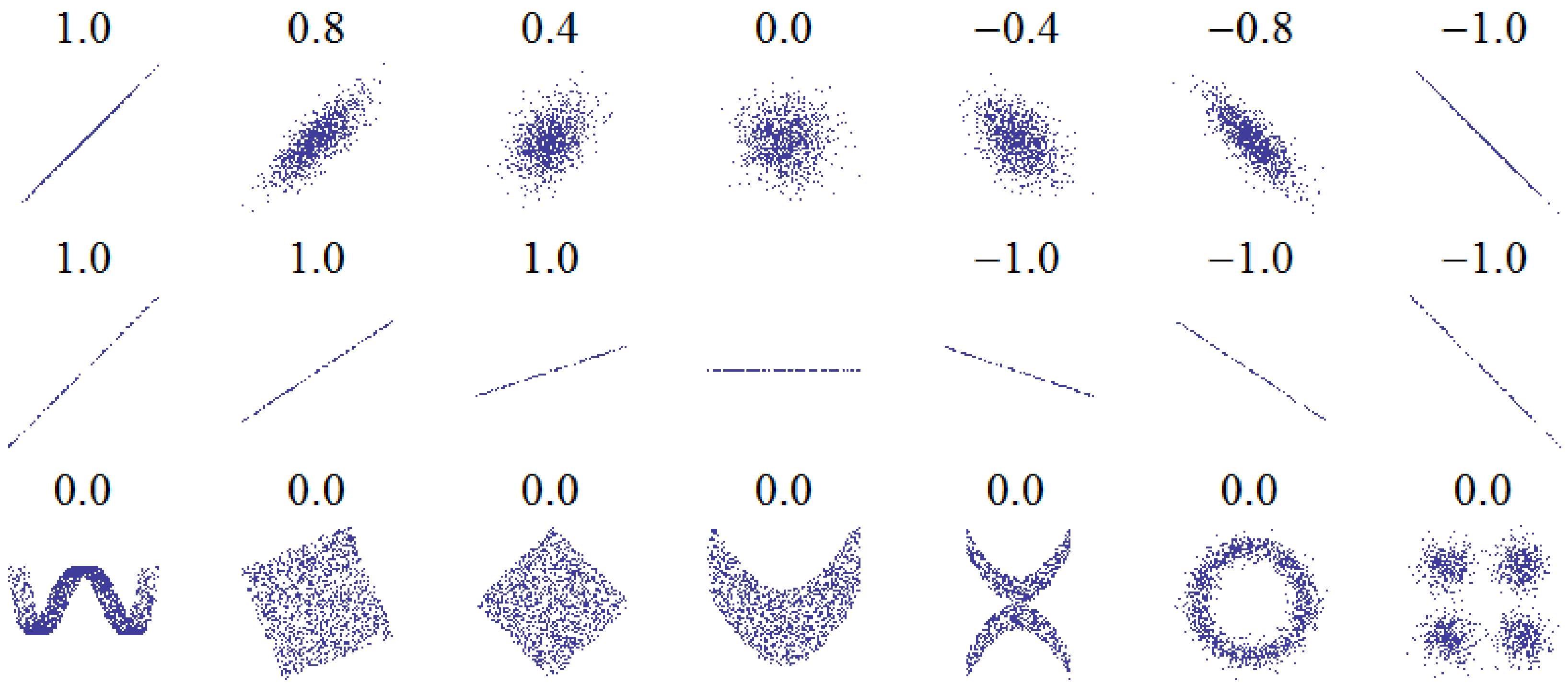
Regression

- A loss function measures how well our model describes our data for given parameters. I.e. how much we lose if we use the model instead of the data itself.
- The residual of a data point (x_i , y_i) measures how far the observed values y_i deviate from the prediction
- The correlation coefficient measures the extent to which there is a linear relationship between two characteristics x and y.



Linear regression

- Examples of linear correlation coefficients

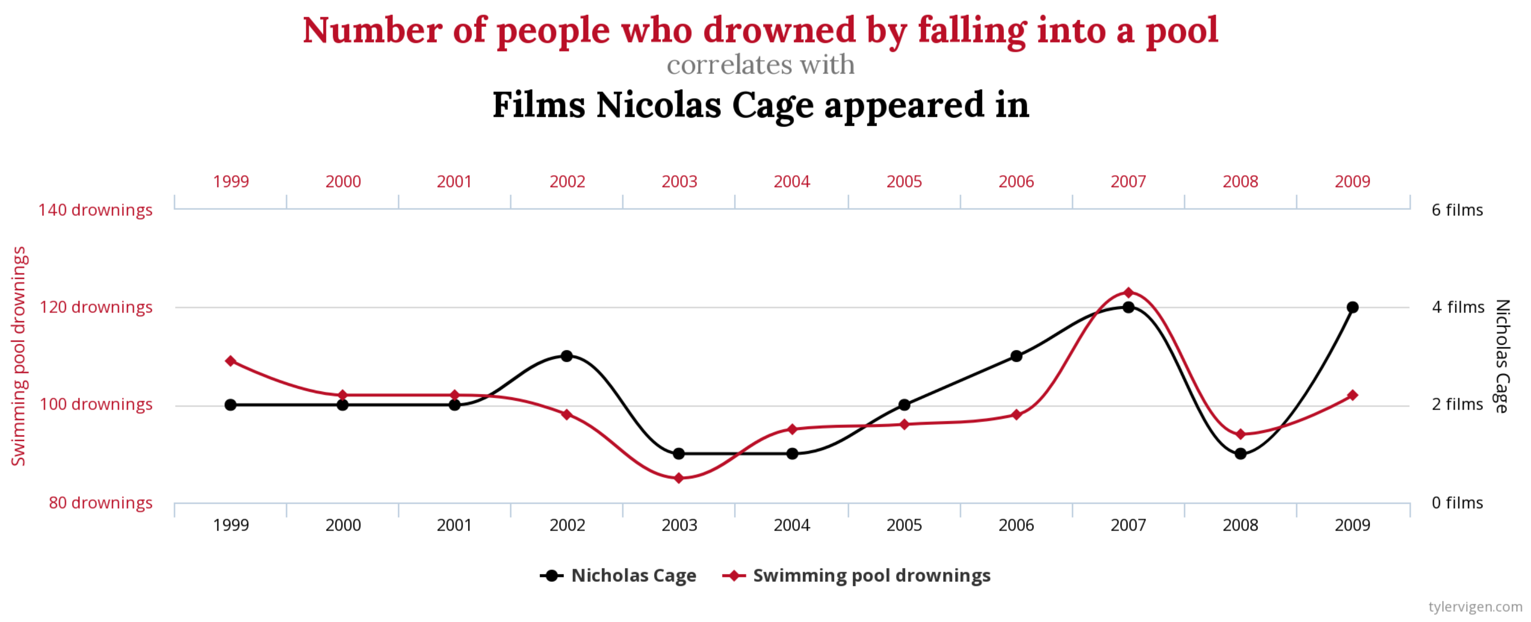


Source: Wikipedia contributors. (2021, September 24). Pearson correlation coefficient. In *Wikipedia, The Free Encyclopedia*. Retrieved 08:37, September 29, 2021, From https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1046205525;
By DenisBoigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

Correlation and causality

Hochschule Hof
University of Applied Sciences

- **Correlation** between two characteristics does not mean that there is also causality.
- Example:



- Correlation of the day: http://www.correlated.org
- Spurious Correlations: http://www.tylervigen.com/

Regression - Example

- Regression analyses model the relationship between a **dependent(y)** and one or more **independent (x)** variables.
- Example:
  - Independent variables: Weight and power
  - Dependent variables: Consumption

- <u>Intuition</u>: How can the dependent variable be explained by the independent ones?
- For this purpose, a model is adopted to represent the dependency. The parameters are calculated from the available data.

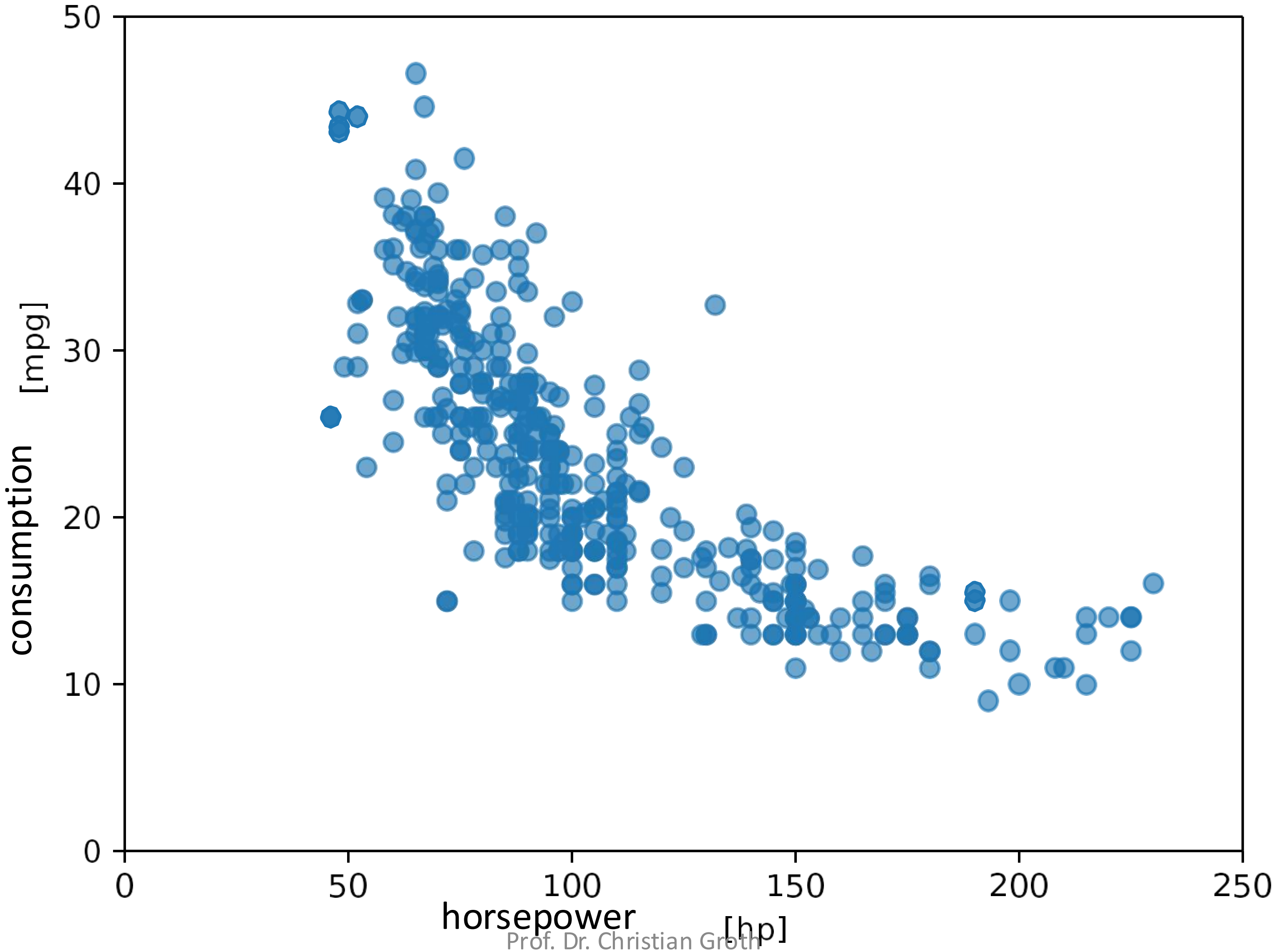
Regression - Example

Hochschule Hof
University of Applied Sciences

- Challenge:
  - **Prediction of consumption** (mpg)
    based on the
    power (hp)
    (and later other characteristics)

- Data:
  - Auto MPG data set from UCI ML Repository
    https://archive.ics.uci.edu/ml/datasets/auto+mpg
  - 398 cars (392 with full features)
  - 8 characteristics (consumption, cylinder, weight, etc.)

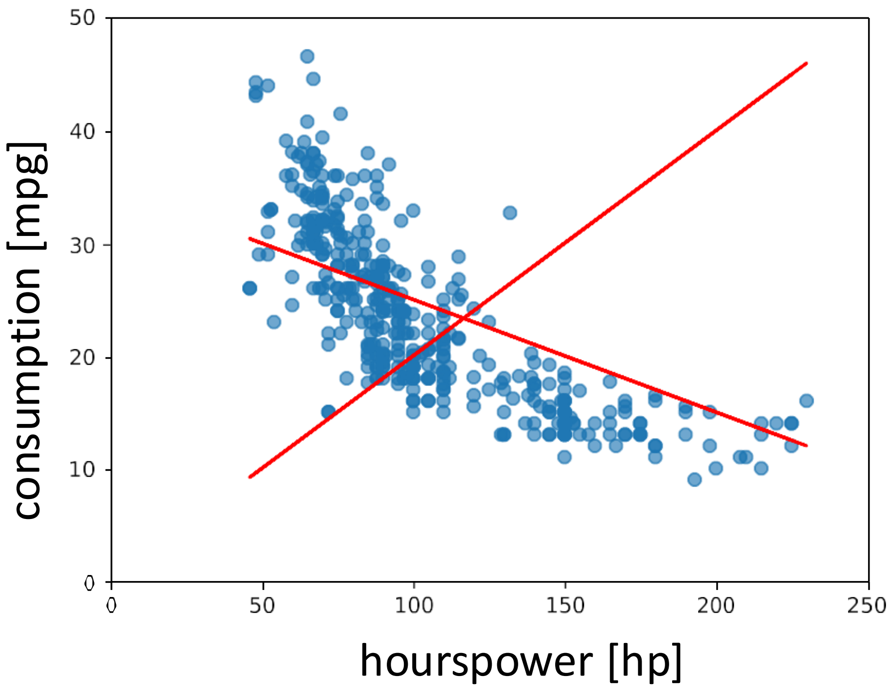
Regression - Example



Hochschule Hof
University of Applied Sciences

Regression - Example

- **Different values of** the parameters w_0 and w_1 correspond to **different straight lines**



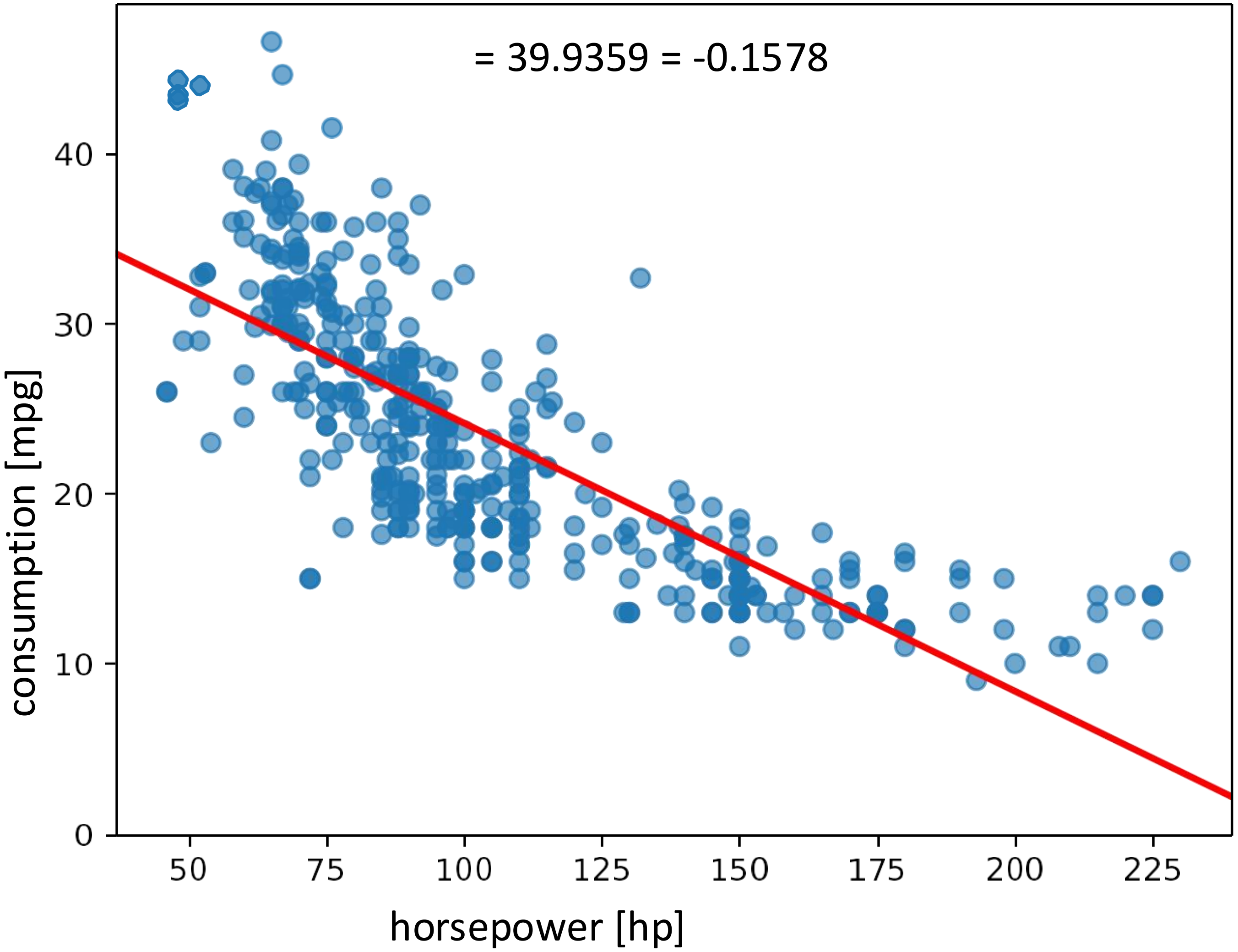
$w_0 = 0 \quad w_1 = 0.2$

$w^0 = 35 \quad w_1 = -0.1$

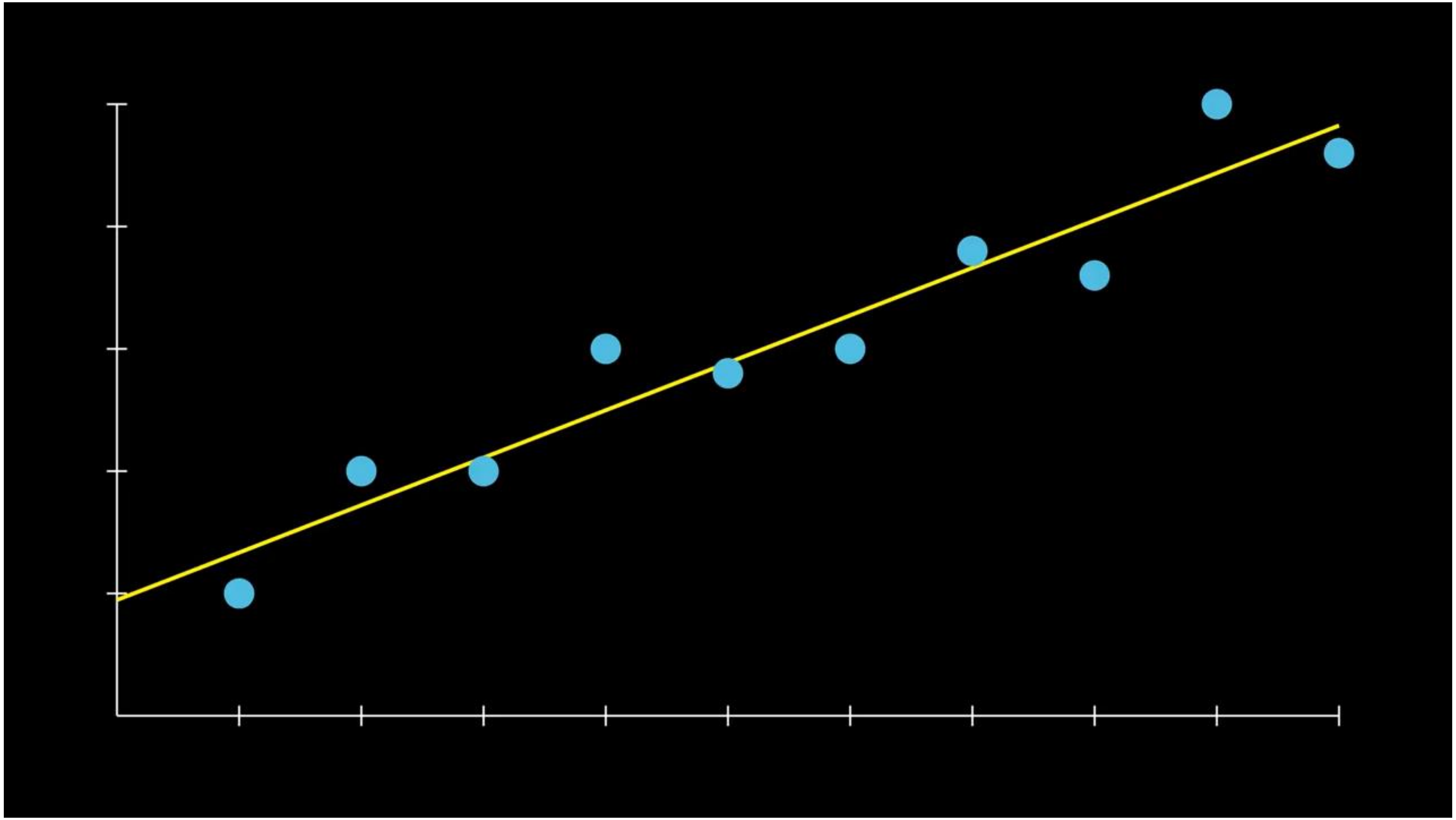
- So we need a **quality criterion,** which straight line is the best.

Regression - Example

- Optimal parameters for our data
- Loss function is sum of squared errors (SSE)
- Optimal parameters by minimizing the loss function (e.g. iteratively using gradient descent).

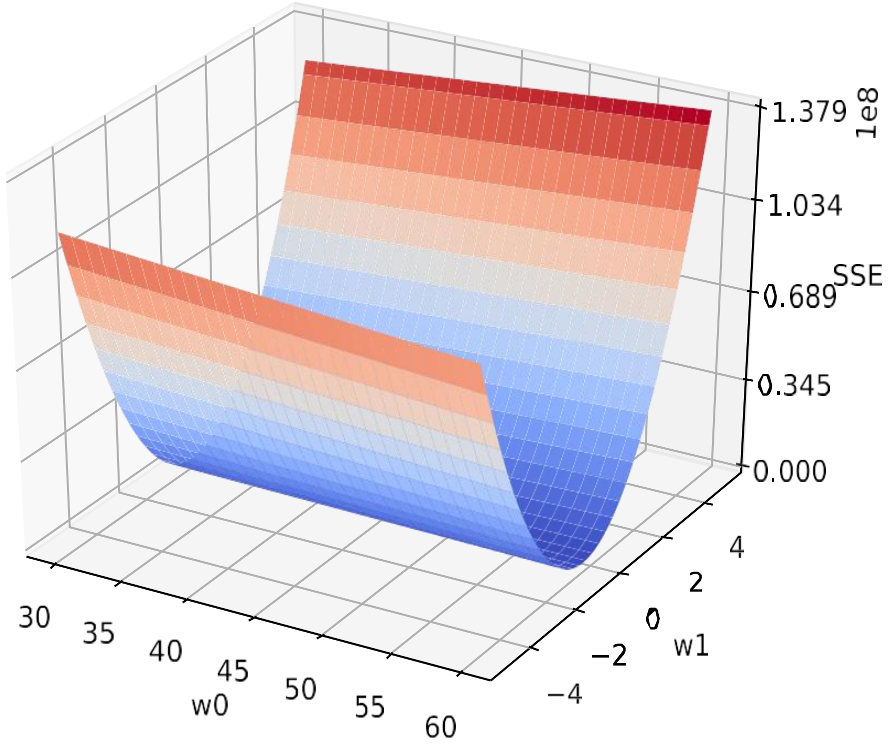


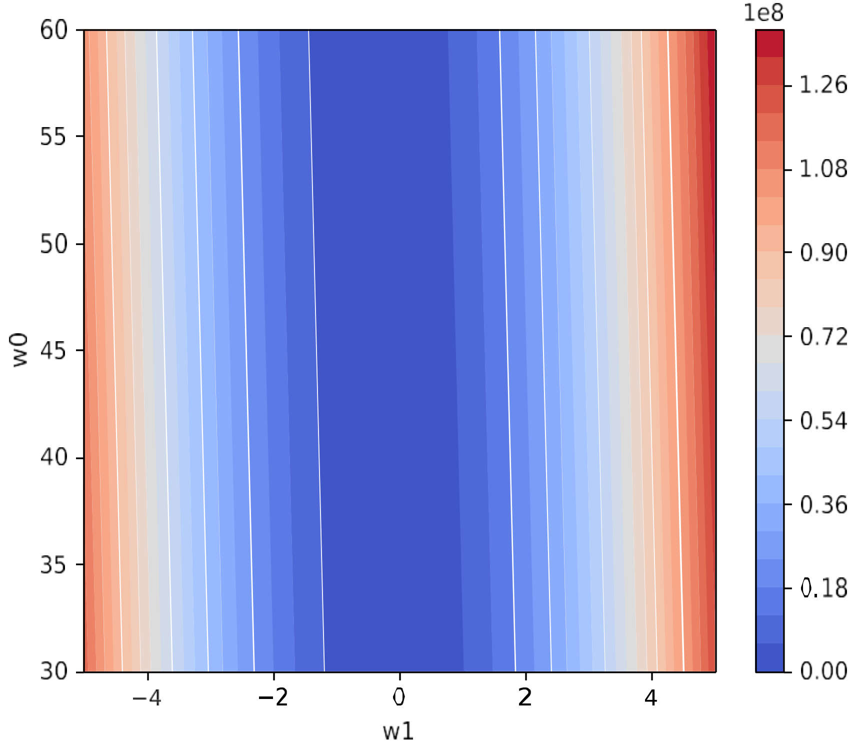
Regression - Procedure



Loss function

- Loss function for our sample data





Other regression models

Hochschule Hof
University of Applied Sciences

- Simple linear regression

$$\hat{y}(w,x) = w_0 + w_1 x_1$$

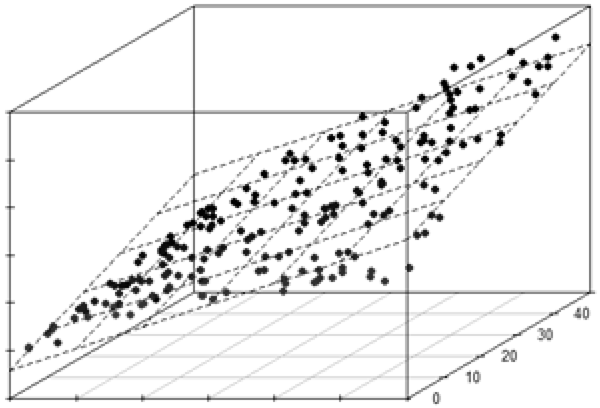
- Multiple linear regression

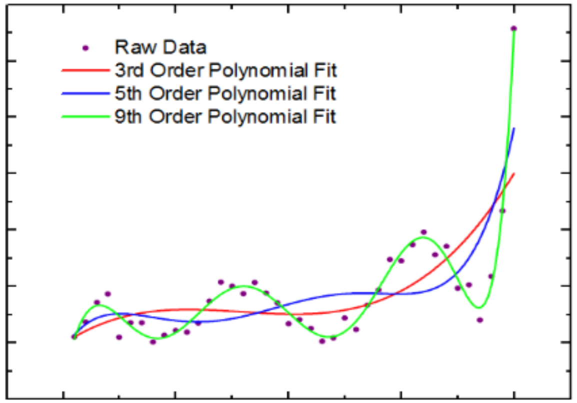
$$\hat{y}(w,x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

- Polynomial (linear) regression

$$\hat{y}(w,x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$$

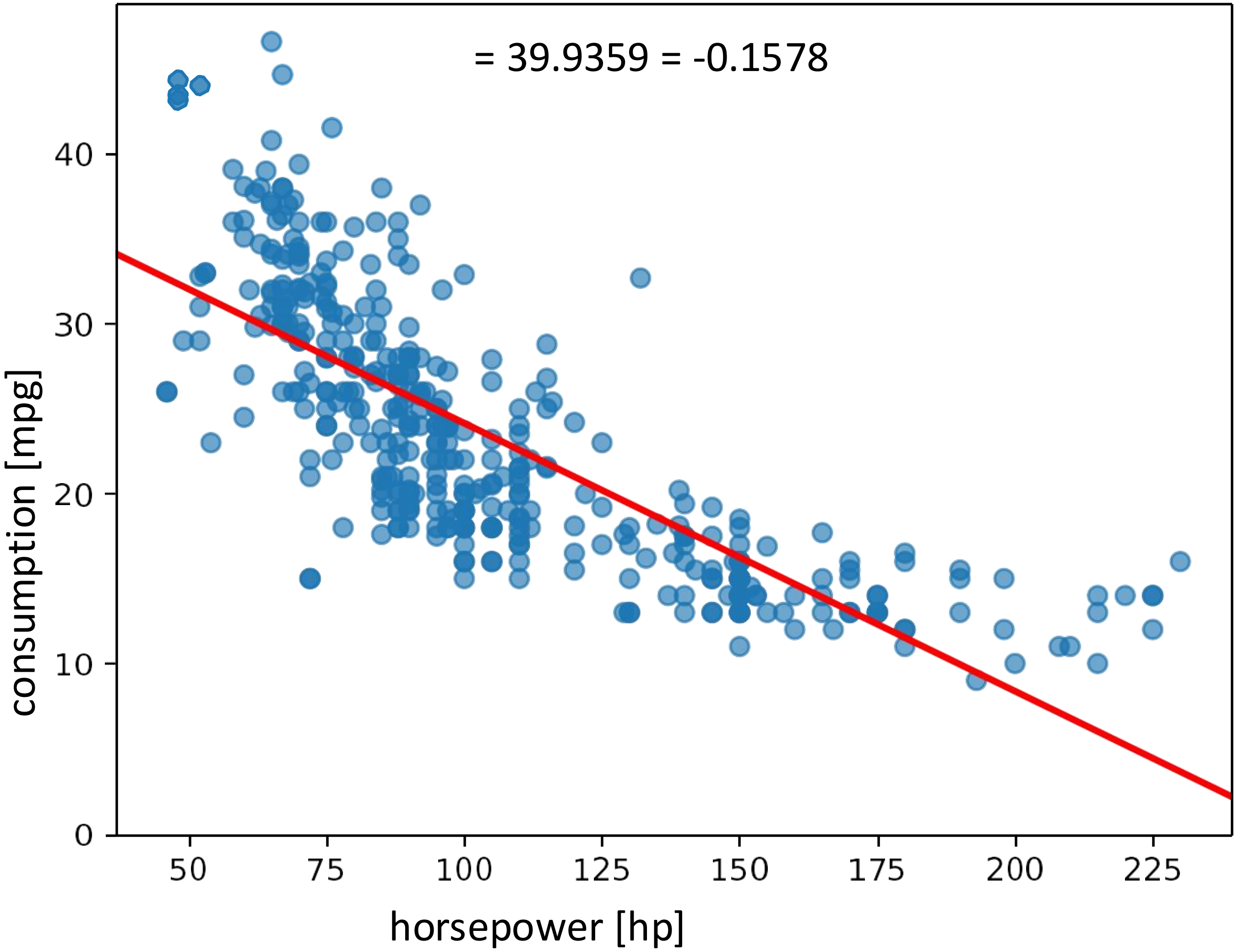
- Variants for loss function
  - Ridge Regression
  - Lasso regression
  - Elastic net
  - → Details in VL Data Science





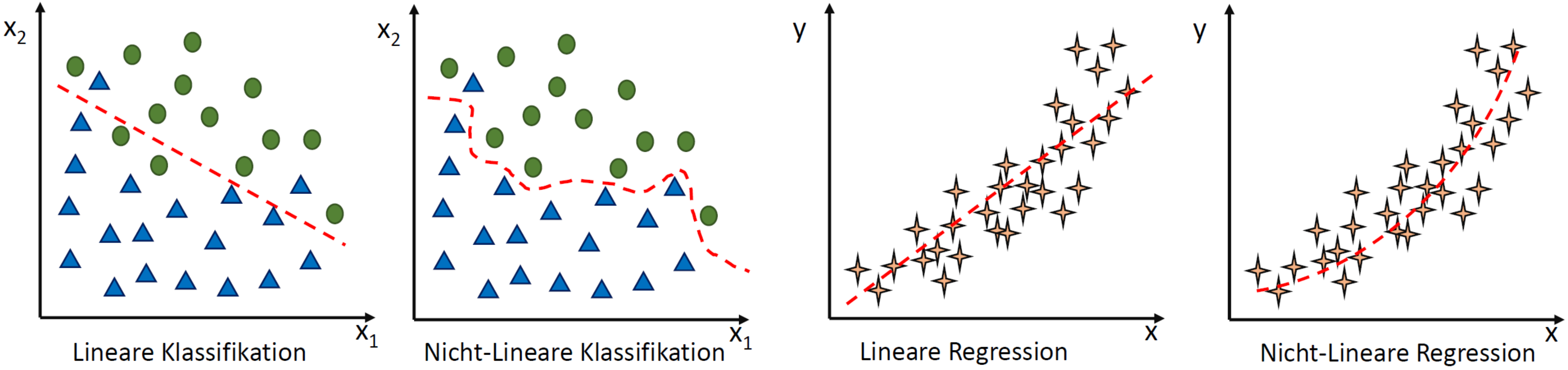
Regression goals

- Resource efficiency
  - Mapping of many individual data points in one model
- Prediction
  - Prediction of a dependent value based on given independent values



Classification

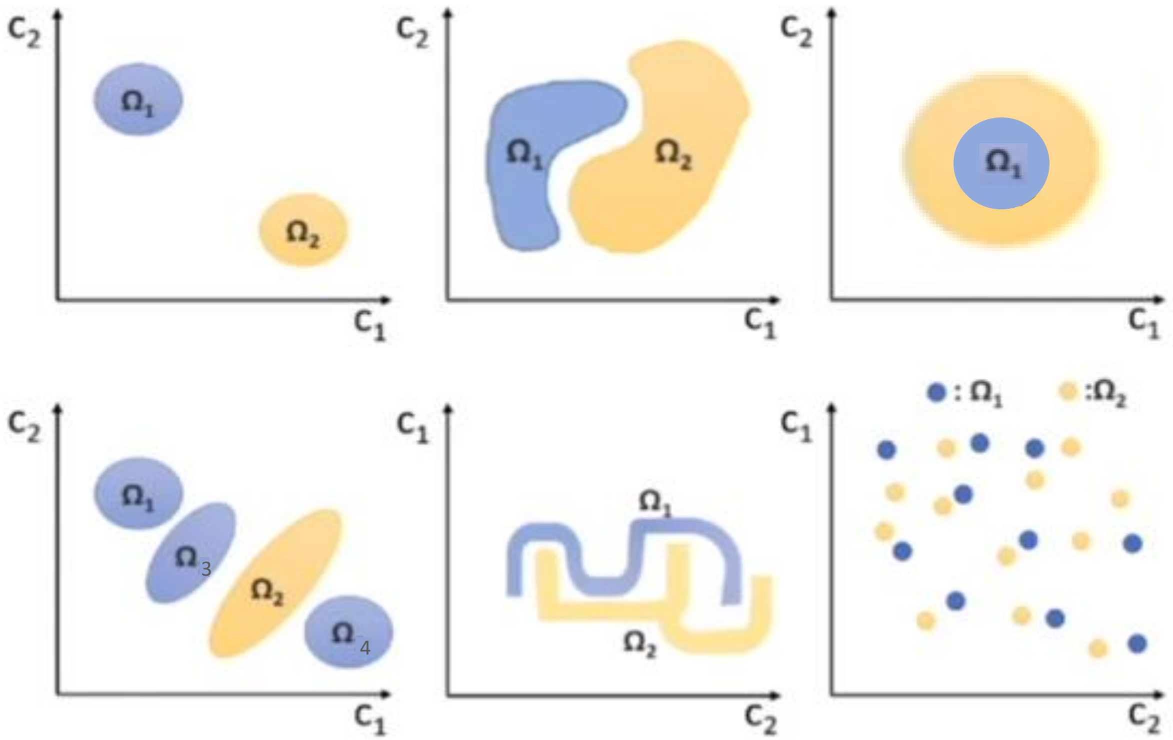
Regression vs. classification



Classification - variants

Hochschule Hof
University of Applied Sciences

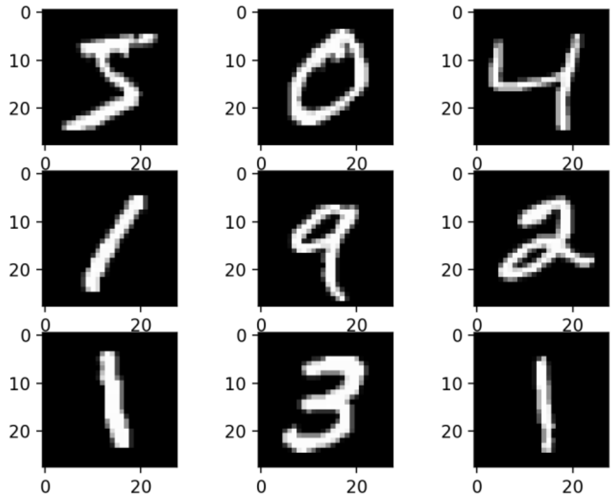
- Top left:
  → Classes are (**linearly**) separable
- Top center and right:
  → Classes are (**non-linearly**) separable
- Bottom left:
  → **Multi-class** classification
- Bottom center:
  → **Complex separable**
  perhaps data of a high-dimensional feature space
  (other projection could be more helpful)
- Bottom right:
  → **not separable**
  Is it really data from two classes?
  How do the other features behave?



Classification - problem variants

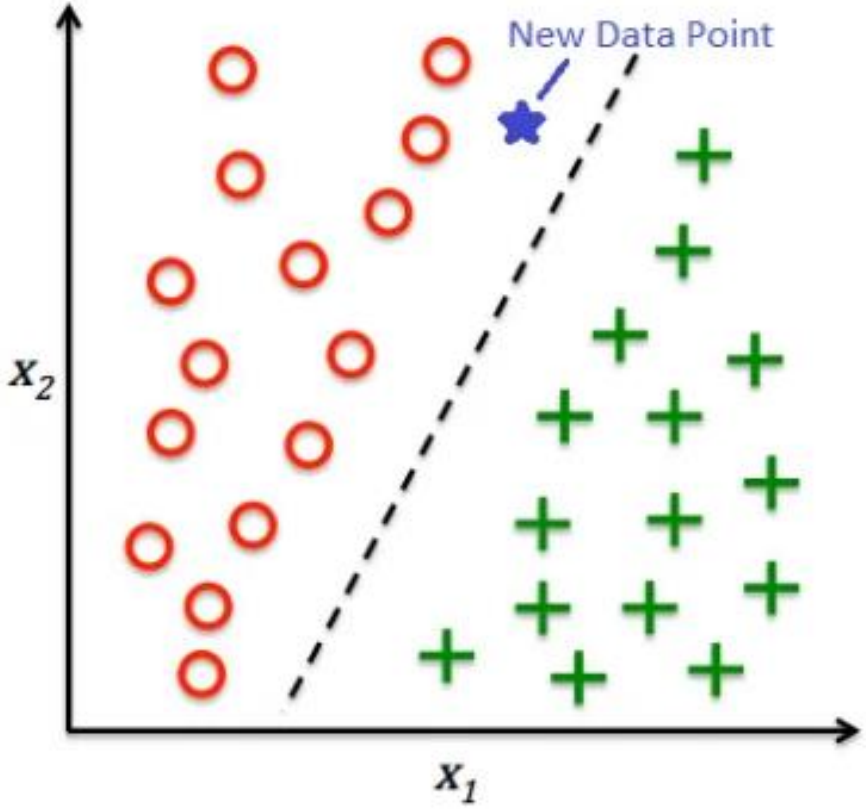
- Binary classification
  - Two classes
  - Example: Email spam vs. no spam

- Multi-class classification
  - Multiple classes
  - Example: Recognition of handwritten digits (0-9)



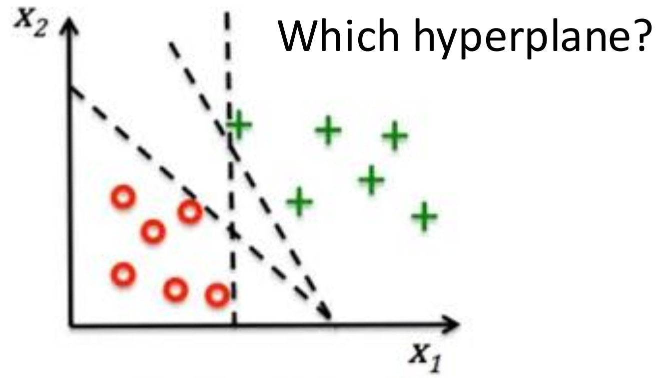
Classification

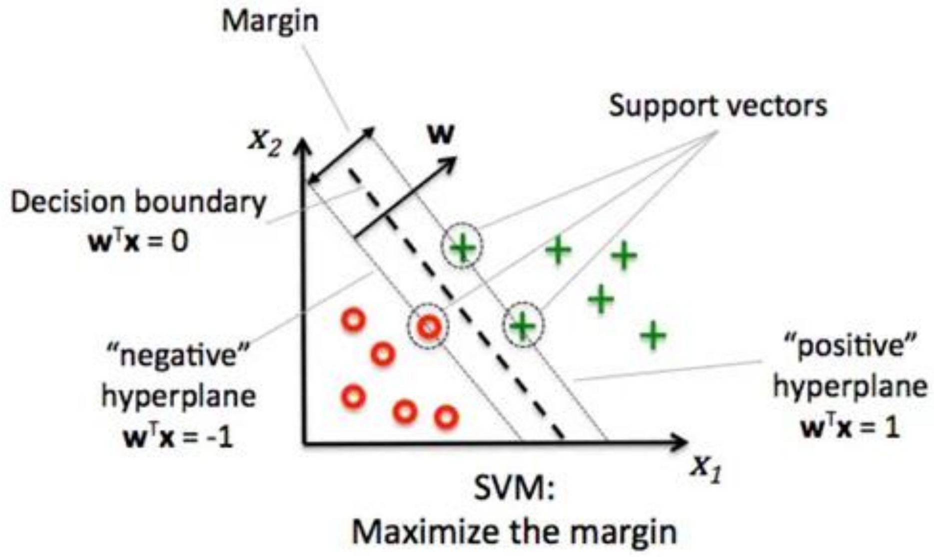
- Creation of a model that depicts the relationship between characteristics and class membership
- Goal: Prediction of class membership for new data points.



Classification algorithms

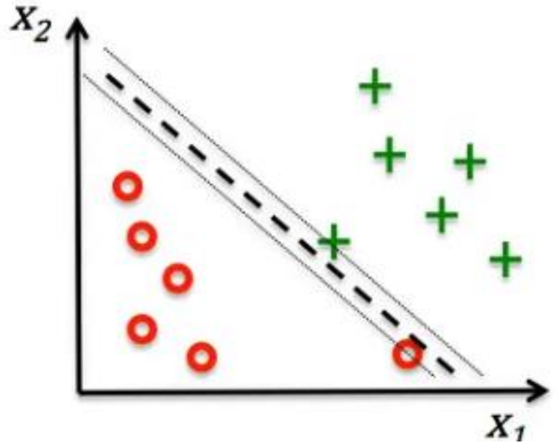
- Logistic regression (for classification)
- Bayes classifier
- Support Vector Machines (Support Vector Machines)
- K-Nearest Neighbour (lazy learner)
- Decision Trees
- Neural networks

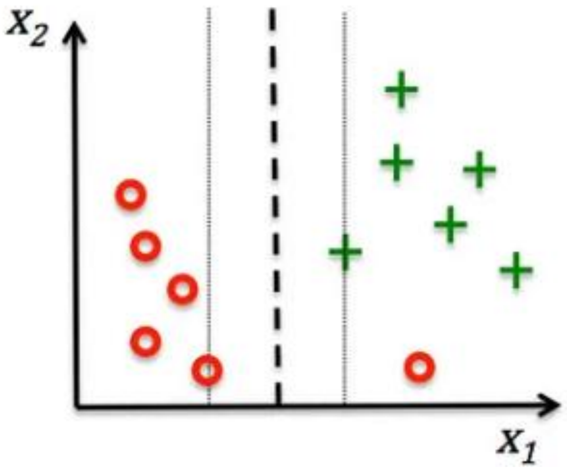




Classification algorithms

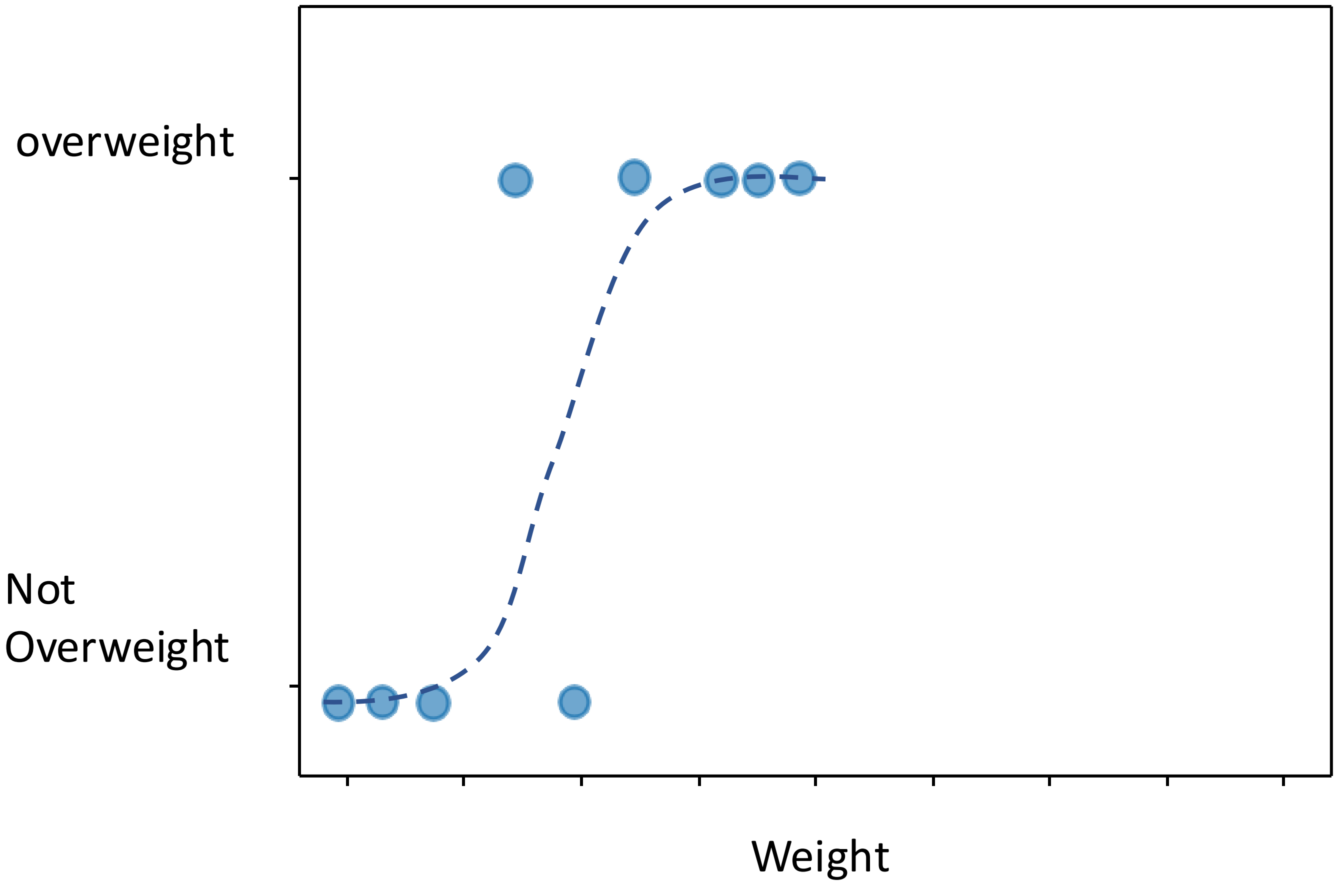
- SVMs try to maximize the width of the margin
- Penalty for incorrect classification
- Misclassifications possible (noise, wrong label, ...)
- Less strict handling of misclassification, may improve model
- Model complexity has great influence





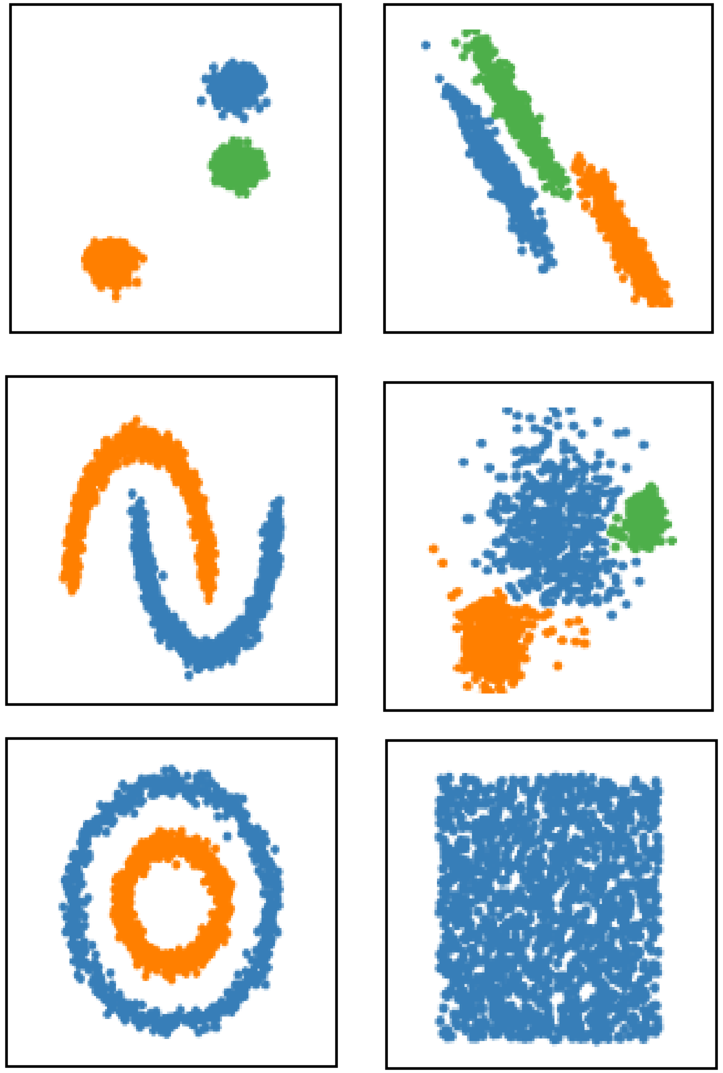
Logistic regression

- For classification
- Threshold value determines class membership
- Maximum likelihood method for parameter estimation



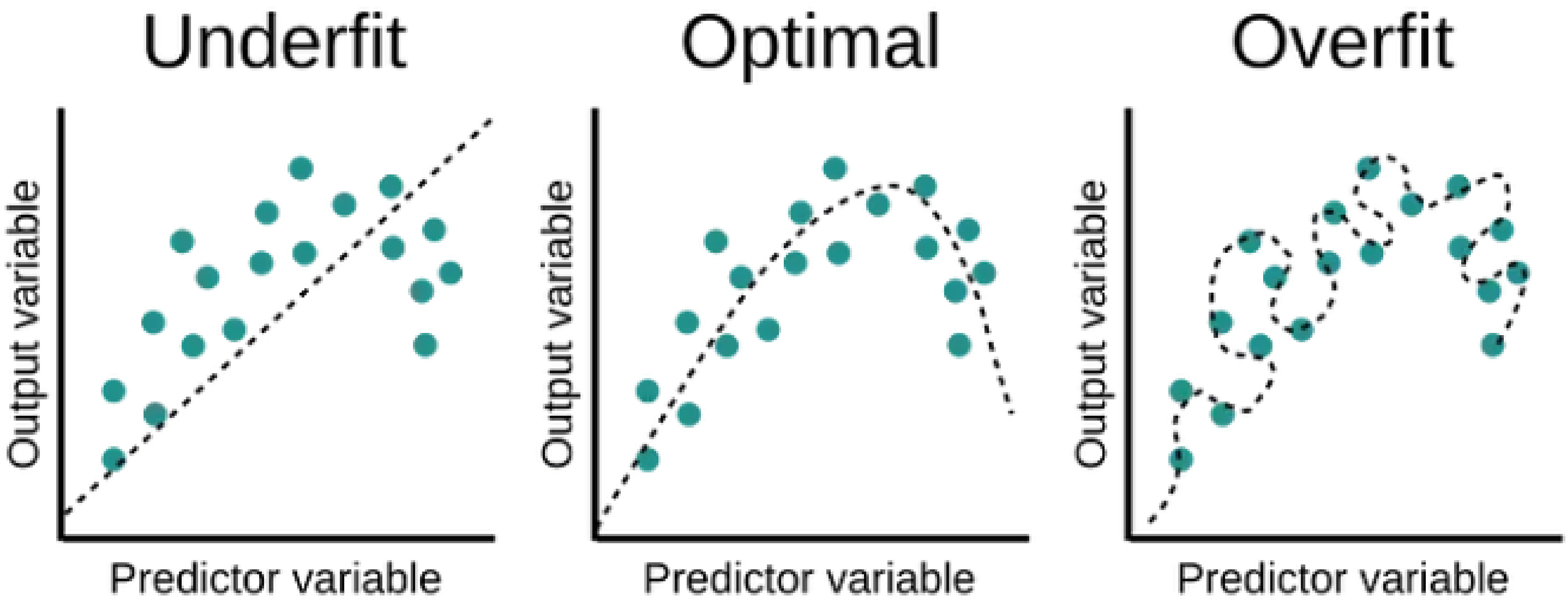
Clustering

- No labels available → no ground truth
- Identification of related areas
- Approaches, e.g.
  - Distance based
  - Density based
  - Hierarchical
- Examples:
  - Recognize product data structures
  - Customer group identification



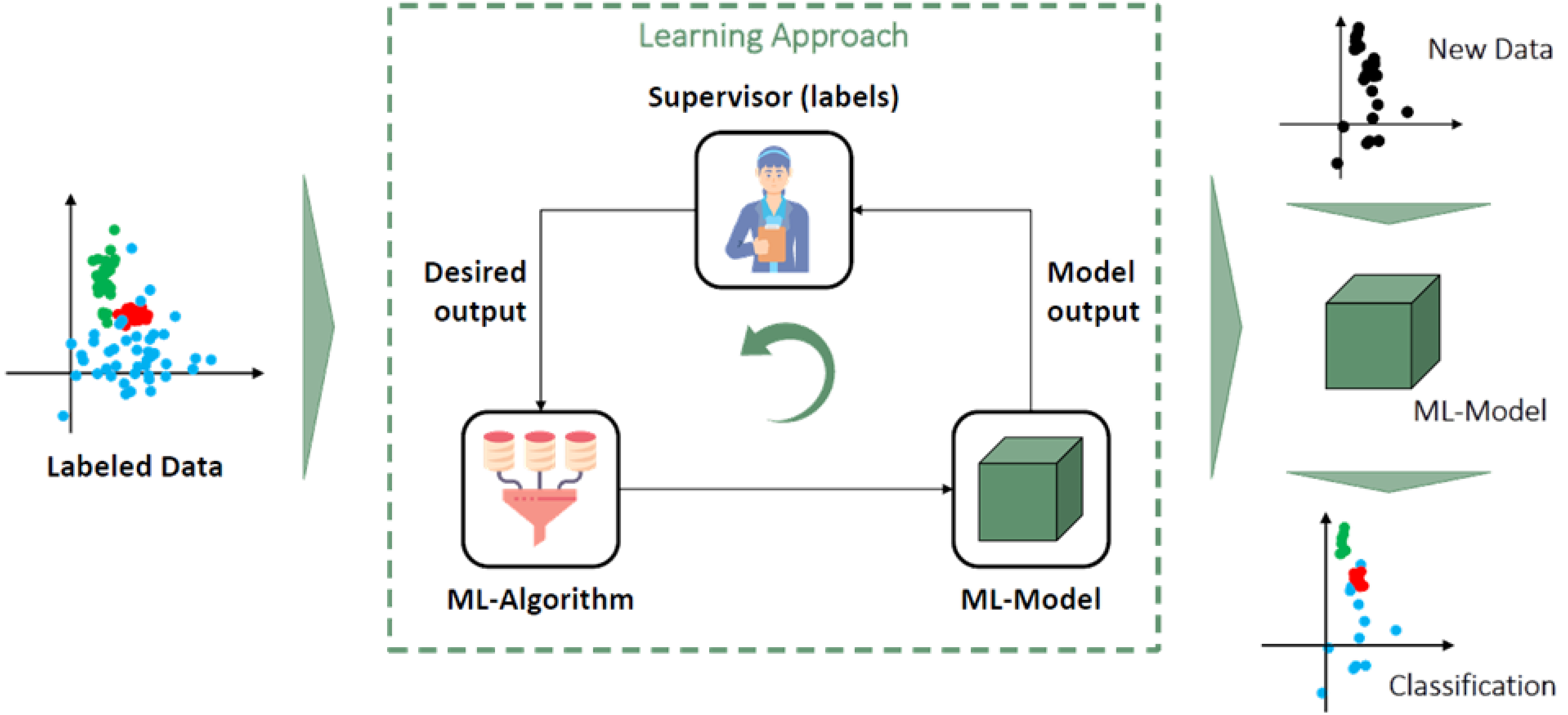
Overfitting / underfitting

Hochschule Hof
University of Applied Sciences



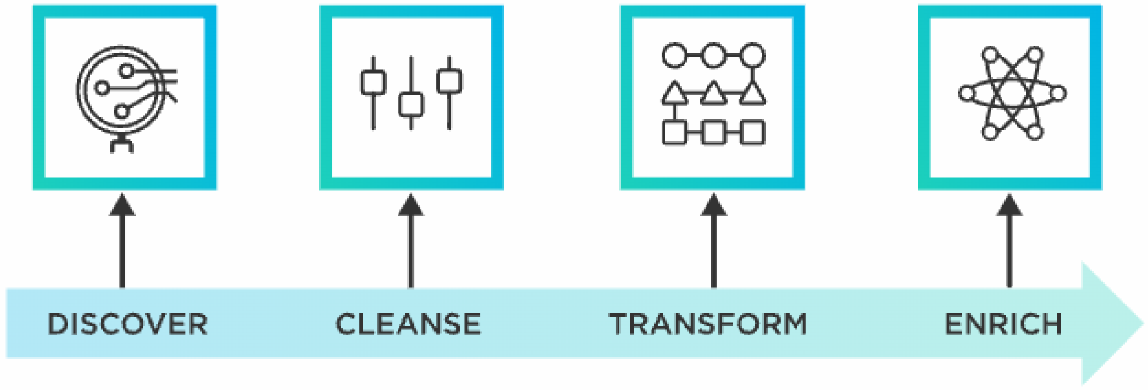
Procedure and quality measurement

ML - Problem description



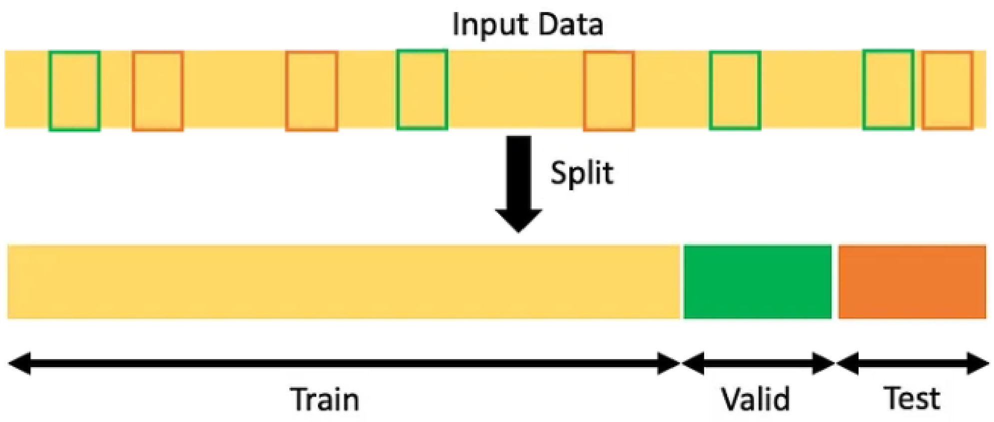
Procedure

- Integrate data (merge from multiple sources)
- Explore
- Data Cleaning / Cleansing
  - Outlier
  - Missing values
- Data Transformation
  - Standardization / Normalization
  - Type change



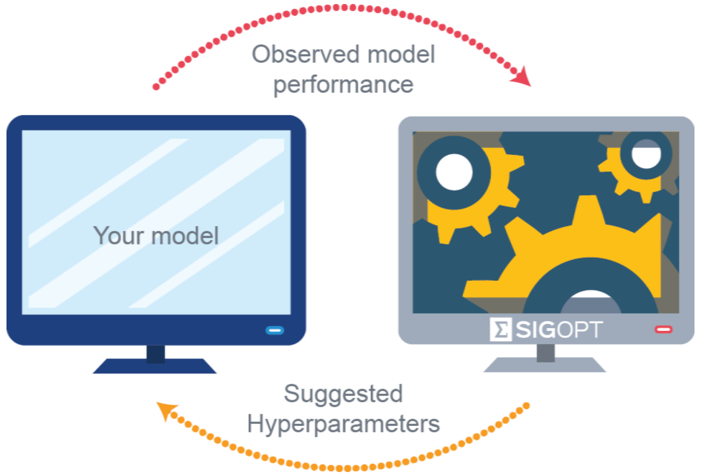
Procedure

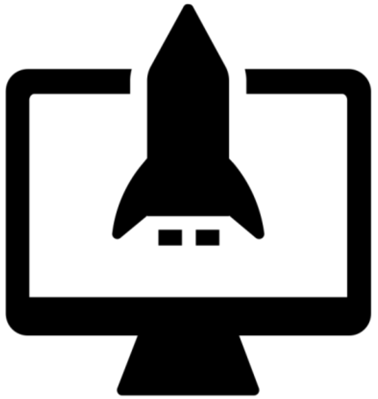
- Division of data into Train / Test or Train / Eval / Test
- More hyperparameters → larger validation set
- Selection of a model
- Training of the model with the help of the training data



Procedure

- Repeat
  - Evaluation of the model
  - Adjustment of the hyperparameters
- (Test of the final model)
- Model Deployment
- Inference



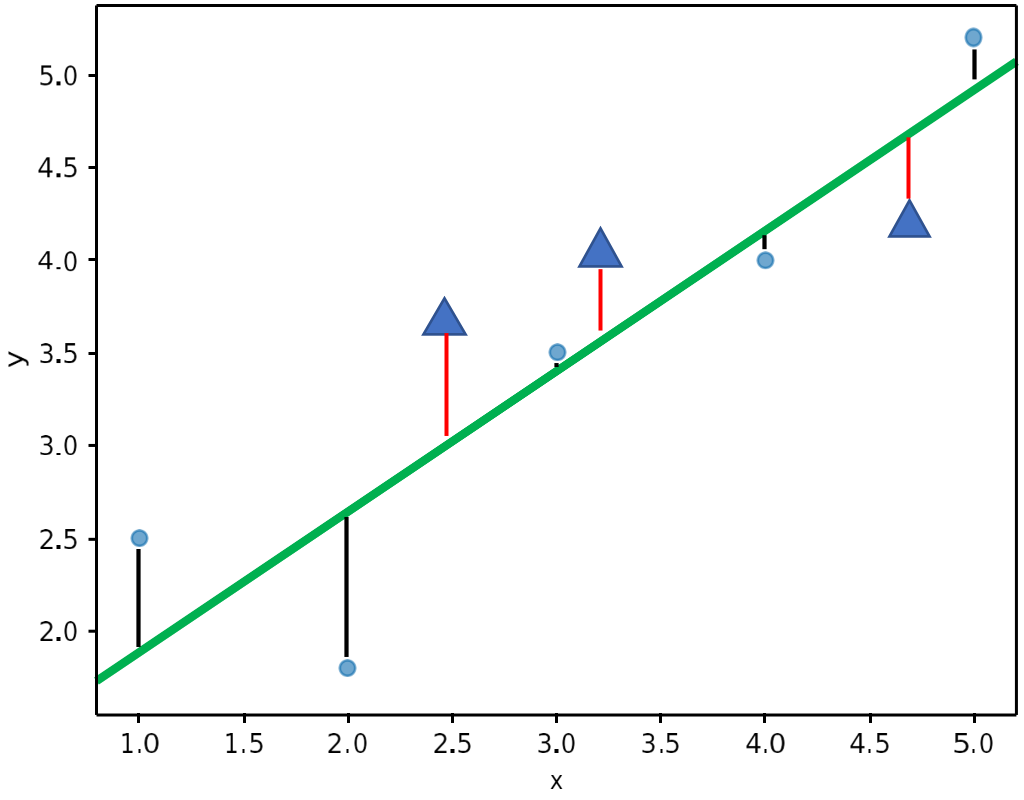


How do we know when a model is "good"?

Model evaluation

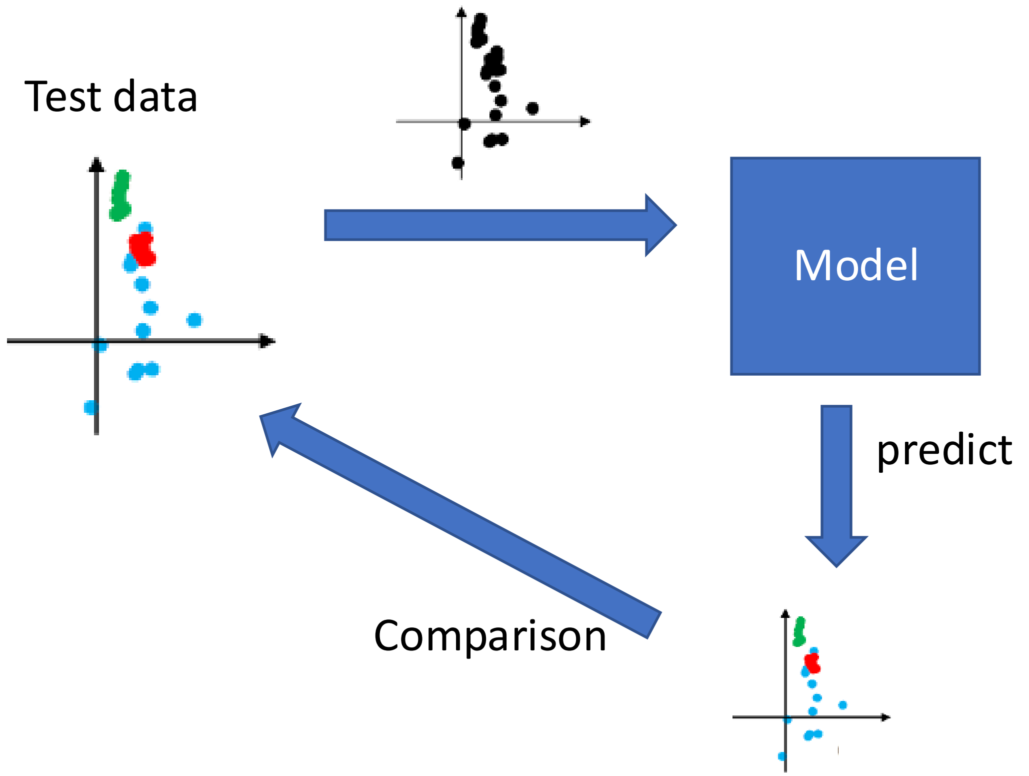
Hochschule Hof
University of Applied Sciences

- Quality metrics provide information about the quality of an ML model
- For this: comparison of the real test data with the predictions
- Regression: Consider deviations as sum / mean / ...



Model evaluation

- Quality metrics provide information about the quality of an ML model
- For this: comparison of the real test data with the predictions
- Regression: Consider deviations as sum / mean / ...
- Classification: Consider number of correctly / incorrectly classified samples
- Clustering: Consider shapes and purity of clusters



Confusion Matrix

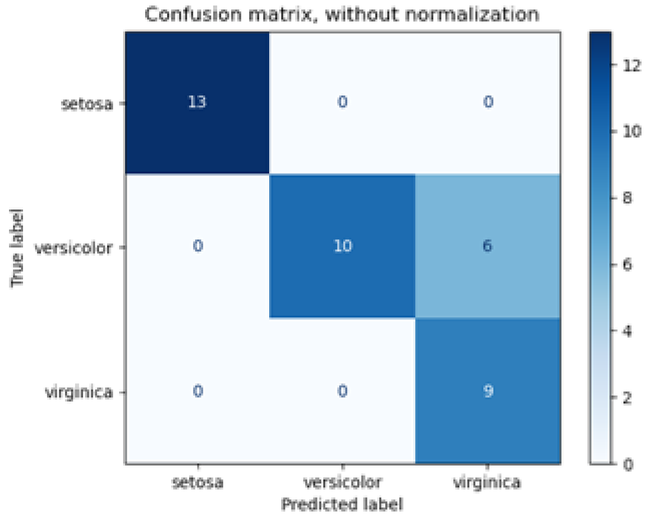
- False Positive (Type 1 Error)
  - Prediction is positive
  - But in fact negative
  - False alarm
- False Negative (Type 2 Error)
  - Prediction is negative
  - But in fact positive
  - Underestimation

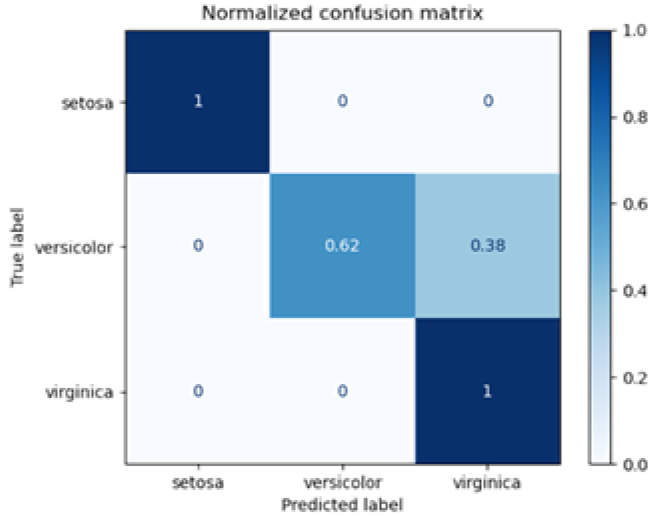
		Predicted Values	
		Positive	Negative
Actual Values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix

Hochschule Hof
University of Applied Sciences

- Confusion matrices for non-binary classifications usually illustrate the number of observations or the proportion of actual/predicted pairs
- An optimal classifier has no FP and FN. A good classifier minimizes both FP and FN.
- Depending on the application, there may be a weighting between Type I and Type II errors.





Accuracy

- **Accuracy** describes the **rate of all correctly classified samples** (either TP or TN) in relation to **all samples.**

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

+ easy to understand
- **Not** good if the data is **unbalanced**
- **Not** good if the costs of **FP and FN** are **very different**

		Predicted Values	
		Positive (PP)	Negative (PN)
Actual Values	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

- For highly unevenly distributed classes: **Balanced accuracy**
  → Introduces weighting

Precision

Hochschule Hof
University of Applied Sciences

- Precision describes the proportion of true positives in relation to all positively classified samples (true and false positives).

Precision = $\frac{TP}{TP + FP}$

- Alternative designation: Positive predictive value (PPV)
  + Describes how well the model performs in positive predictions.

  + suitable criterion if **type I** error is **more relevant to the** application (FP costs are high)

  - not suitable, if type II error is more relevant for the application (since not even considered in the criteria)

		Predicted Values	
		Positive (PP)	Negative (PN)
Actual Values	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Example: Spam detection.

FP = non-spam message is identified as spam
→ Potentially important messages are filtered out

FN = spam Message is not identified as spam
→ Annoying spam messages are displayed to the user

Recall

Hochschule Hof
University of Applied Sciences

- Recall describes the rate of true positives with respect to all positive samples (true positives and false negatives).

Recall / TPR = $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$

		Predicted Values	
		Positive (PP)	Negative (PN)
Actual Values	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

- Alternative names: Hit Rate, True Positive Rate, Sensitivity.

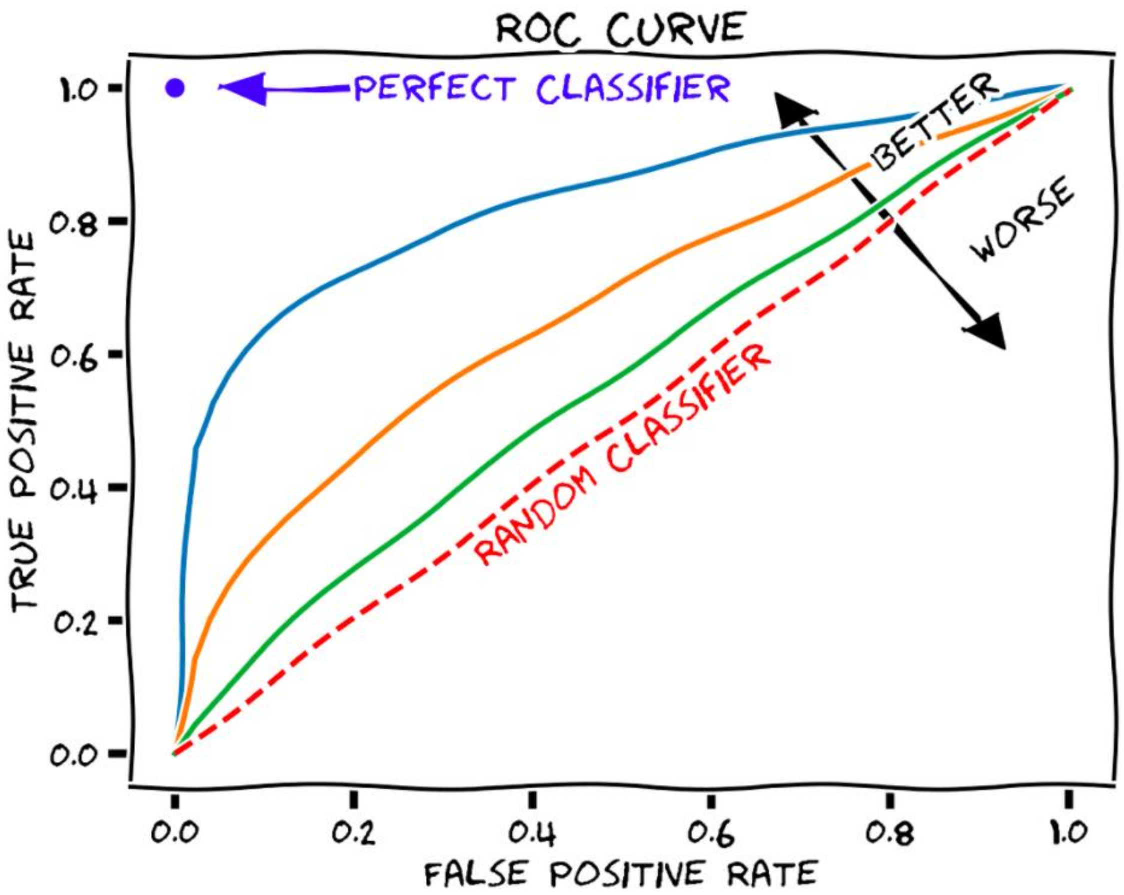
+ describes how well the model performs with positive input data

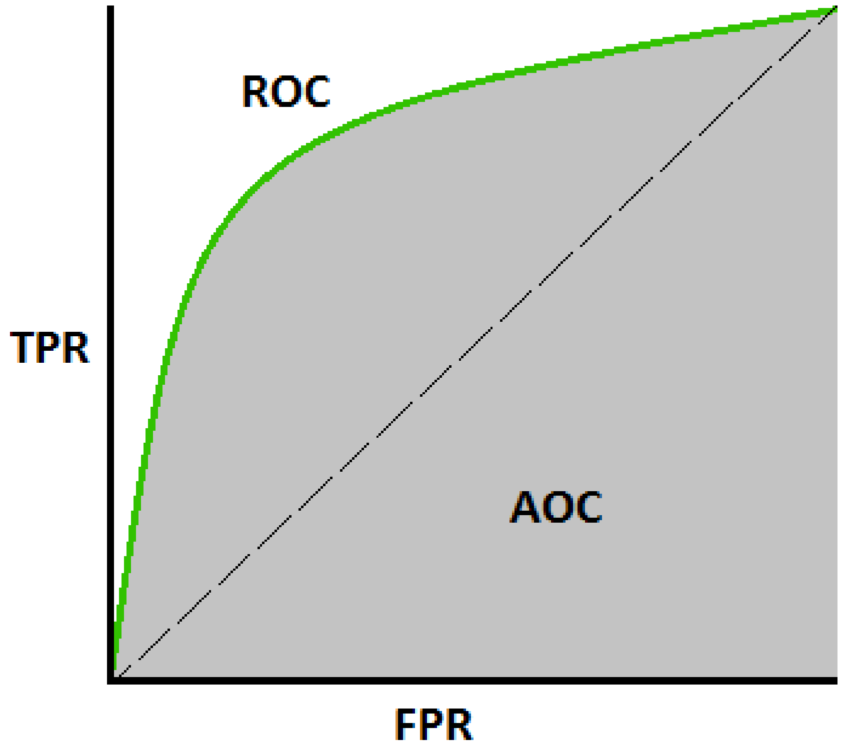
+ Suitable criterion if **Type II errors** are **more relevant to the application** (FN costs are high).

- not suitable if the type I error is more relevant for the application (since not even considered in the criteria)

- Recall alone can be misleading for the evaluation

Receiver Operating Characteristic (ROC)





Example

- Given the following confusion matrix

CM		Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	20	9
	1 (Yes)	6	44

- Accuracy: rate of correctly classified samples

$$CCR = \frac{20 + 44}{9 + 6 + 44 + 20} = \frac{64}{79}$$

- Error rate

$$ER = \frac{9 + 6}{9 + 6 + 44 + 20} = \frac{15}{79}$$

Example

Hochschule Hof
University of Applied Sciences

- Given the following confusion matrix

CM		Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	20	9
	1 (Yes)	6	44

- Recall / True Positive Rate

$$TPR = \frac{44}{6+44} = \frac{44}{50}$$

- False Positive Rate

$$FPR = \frac{9}{9+20} = \frac{9}{29}$$

Example

- Given the following confusion matrix

CM		Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	20	9
	1 (Yes)	6	44

- Precision

$$P = \frac{44}{44 + 9} = \frac{44}{53}$$

- F1 score

$$F1 = \frac{\frac{44}{53} * \frac{44}{50}}{\frac{44}{53} + \frac{44}{50}} \approx 0.85$$

Evaluation of multi-class classification

Hochschule Hof
University of Applied Sciences

- **The confusion matrix** contains one row and one column per class for multi-class classification. Example with three classes:

CM		Forecast		
		A	B	C
Data	A	50	0	10
	B	0	10	1
	C	5	20	5

- Basic idea: consider each class as a positive class and the others as negative classes, and aggregate the resulting quality measures

Example

Hochschule Hof
University of Applied Sciences

CM		Forecast		
		A	B	C
Data	A	50	0	10
	B	0	10	1
	C	5	20	5

CM	A	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	36	5
	1 (Yes)	10	50

CM	B	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	70	20
	1 (Yes)	1	10

CM	C	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	60	11
	1 (Yes)	25	5

Micro and macro averaging

- There are two options for averaging the desired quality measure per class
- Micro-Averaging sums each of the four categories (TN,FN, TP, FP) across classes and inserts these sums into the definition

$$P = \frac{TP_A + TP_B + TP_C}{TP_A + TP_B + TP_C + FP_A + FP_B + FP_C}$$

$$R = \frac{TP_A + TP_B + TP_C}{FN_A + FN_B + FN_C + TP_A + TP_B + TP_C}$$

Example micro-averaging

Hochschule Hof
University of Applied Sciences

CM	A	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	36	5
	1 (Yes)	10	50

CM	B	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	70	20
	1 (Yes)	1	10

CM	C	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	60	11
	1 (Yes)	25	5

$$\mathrm{P} = \frac{50_A + 10_B + 5_C}{50_A + 10_B + 5_C + 5_A + 20_B + 11_C} = \frac{65}{101} \approx 0.64$$

$$\mathrm{R} = \frac{50_A + 10_B + 5_C}{50 + 10_B + 5_C + 10_A + 1_B + 25_C} = \frac{65}{101} \approx 0.64$$

Micro and macro averaging

- Macro averaging calculates the quality measure per class and averages it

$$P = \frac{1}{3}(P_A + P_B + P_C)$$

$$R = \frac{1}{3}(R_A + R_B + R_C)$$

Example macro averaging

CM	A	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	36	5
	1 (Yes)	10	50

CM	B	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	70	20
	1 (Yes)	1	10

CM	C	Forecast	
		0 (No)	1 (Yes)
Data	0 (No)	60	11
	1 (Yes)	25	5

$$P_A = \frac{50}{50+5} = \frac{50}{55}$$

$$P_B = \frac{10}{10+20} = \frac{10}{30}$$

$$P_C = \frac{5}{5+11} = \frac{5}{16}$$

$$R_A = \frac{50}{10+50} = \frac{50}{60}$$

$$R_B = \frac{10}{1+10} = \frac{10}{11}$$

$$R_C = \frac{5}{25+5} = \frac{5}{30}$$

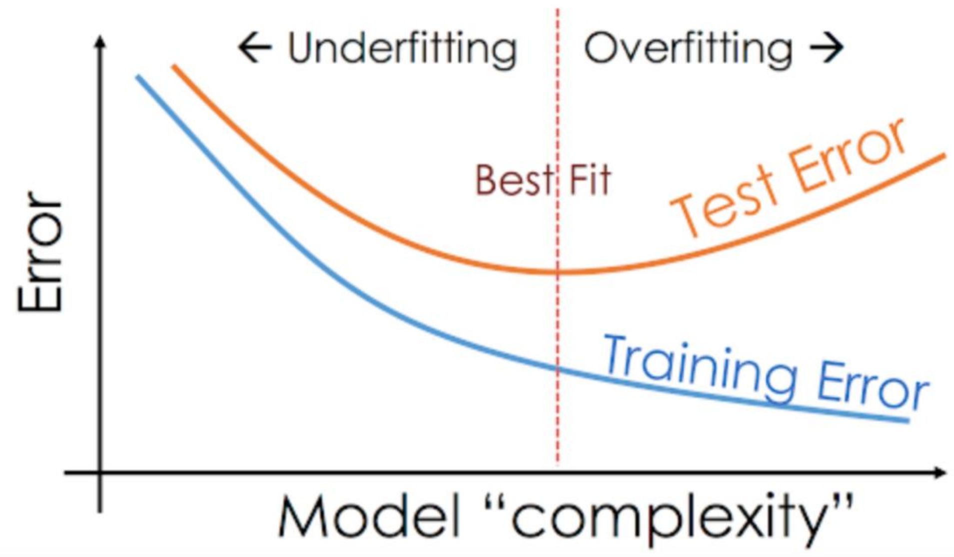
$$P = \frac{1}{3}\left(\frac{50}{55} + \frac{10}{30} + \frac{5}{16}\right) \approx 0.52$$

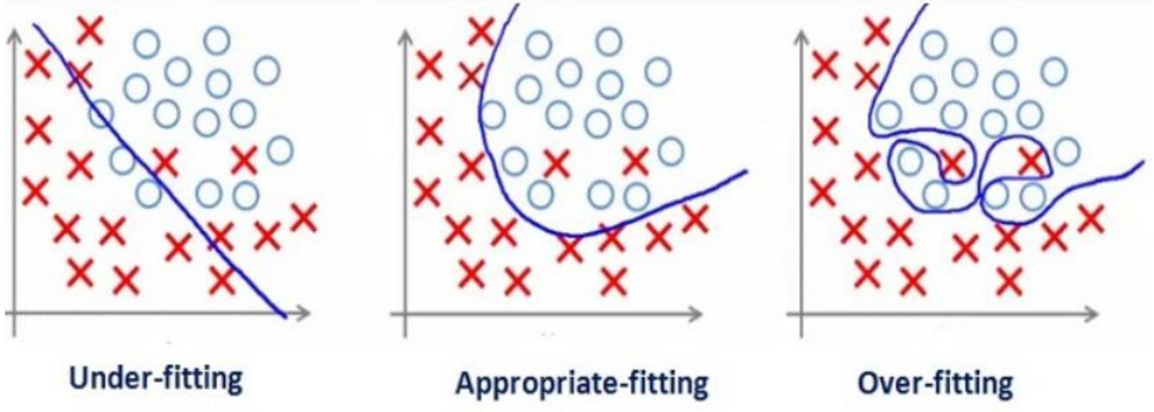
$$R = \frac{1}{3}\left(\frac{50}{60} + \frac{10}{11} + \frac{5}{30}\right) \approx 0.64$$

Overfitting / Underfitting

Hochschule Hof
University of Applied Sciences

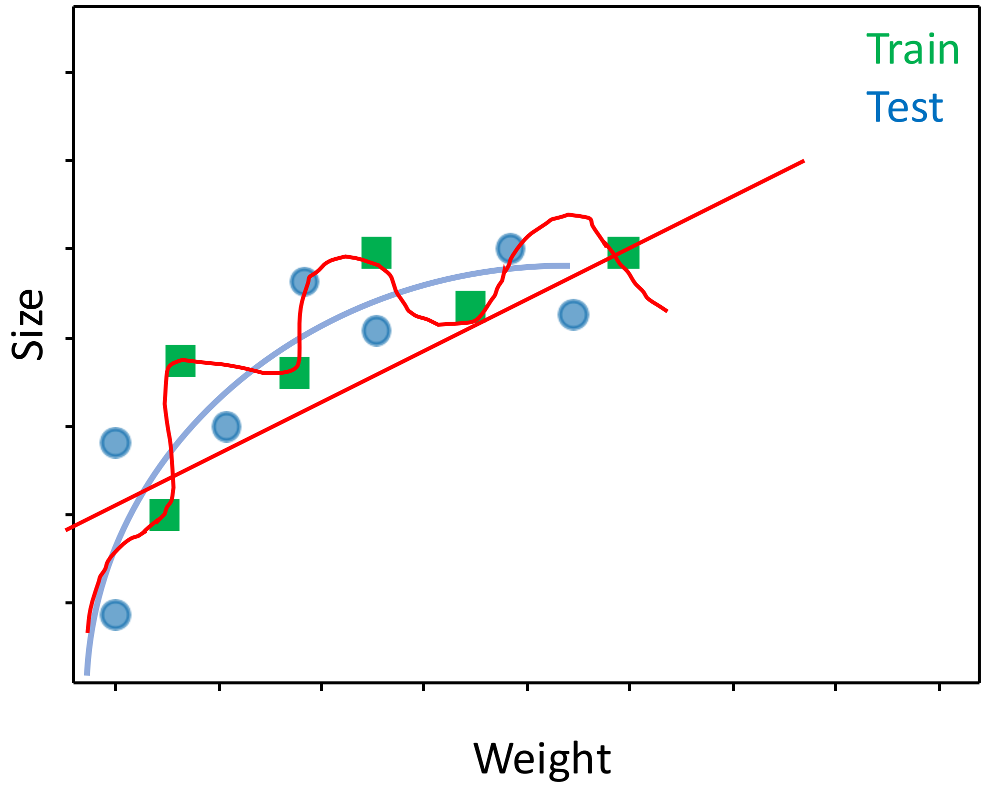
- Error high in training and test: underfitting
- Error low in training and significantly higher in test: overfitting
- Overfitting by:
  - Model complexity too high
  - Too many training epochs





Bias / Variance

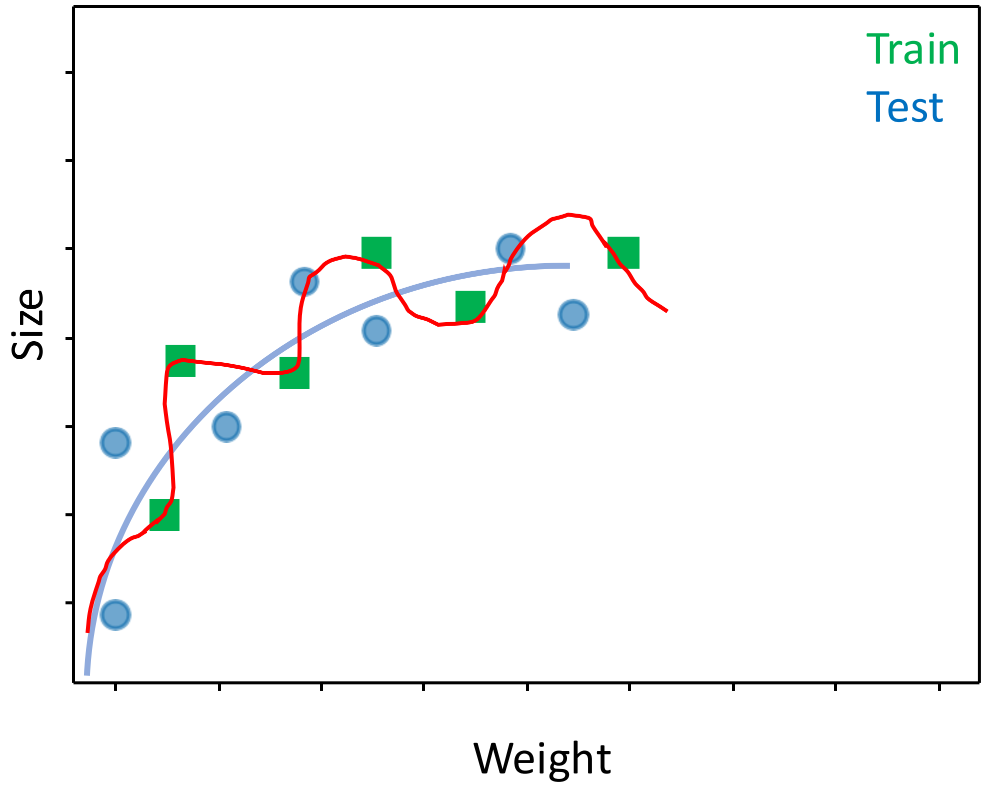
- Actual relationship unknown (blue curve)
- Bias: Inability of a model to model the actual context.
- High Bias:
  - High error
  - Example here: linear regression



Bias / Variance

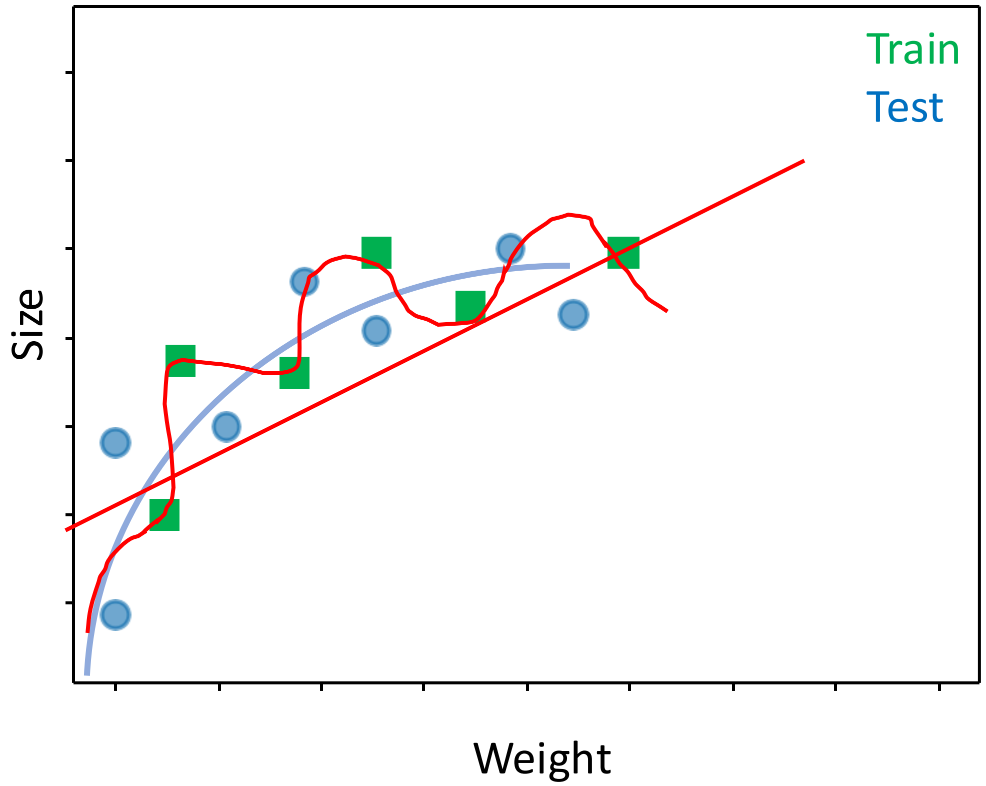
Hochschule Hof
University of Applied Sciences

- Low bias
  - Minor error (in training)
  - Here: highly polynomial model
- Variance:
  - Accuracy of fit between data sets (e.g. between training and testing)
- High variance:
  - Error size differs greatly for different data sets. Data sets strongly
  - Predictability unclear



Bias / Variance

- Linear Regression: High Bias, Low Variance
- Polynomial: Low Bias, High Variance
- Ideal model:
  - Low Bias, Low Variance
  - Adjusted model complexity
  - Help by: Regularization, Boosting, Bagging



Summary

Hochschule Hof
University of Applied Sciences

- Regression can predict a continuous dependent value to independent variables.
- Classification is used to assign a sample to a class based on its characteristics.
- Clustering is used to find structures in unlabeled data.
- The quality of a model can be evaluated using various metrics.