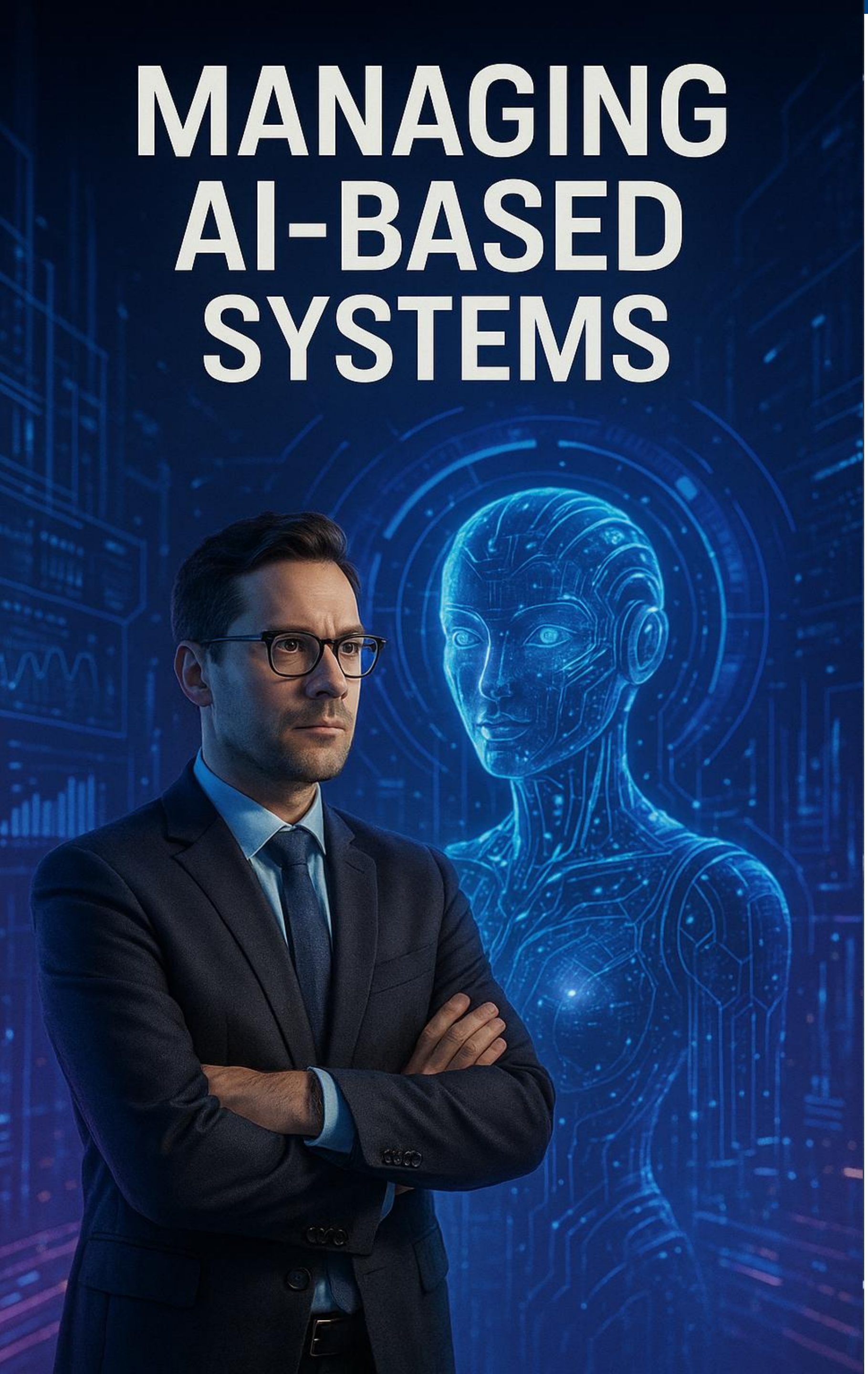


Session 4: AI metrics

Managing AI-based Systems

Prof. Dr. Nils Urbach

Frankfurt University of Applied Sciences,
Research Lab for Digital Innovation & Transformation

FIM Forschungsinstitut für Informationsmanagement

Fraunhofer-Institut für Angewandte Informationstechnik FIT,
Institutsteil Wirtschaftsinformatik

www.fim-rc.de
www.wirtschaftsinformatik.fraunhofer.de
www.ditlab.org

Creative Commons Copyright

This work is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit:
https://creativecommons.org/licenses/by-nc-sa/4.0/

Course navigator

2

3 Basic understanding of AI and its business potential

4

AI Ideation

5 Application potentials of AI technologies

6 Identification, design and evaluation of AI use cases

AI Strategizing

7 Evaluation an organization's AI readiness

8 Management and governance of AI

AI Design & Development

10 Architectures of AI applications

11 Data Management and Model Transparency

12 Design of human-AI interactions

AI Operations at Scale

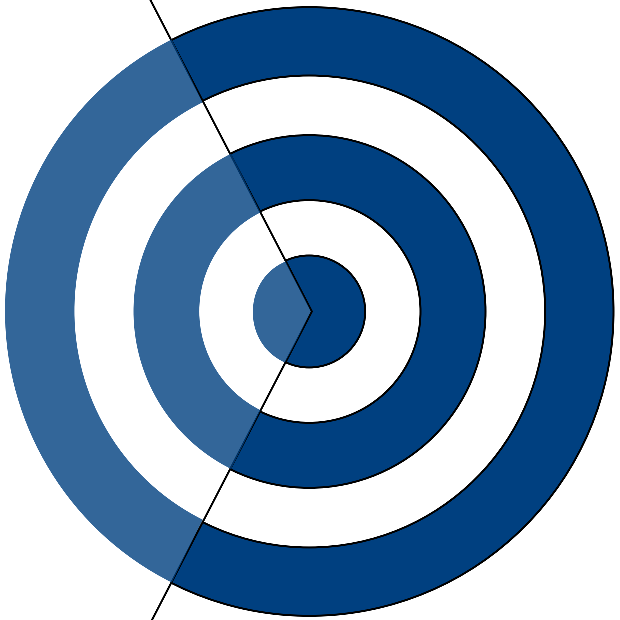
13 Monitoring and KPI-based control

14 Ethical, legal and social implications of AI

Implementation

Objectives of today’s lecture

1. Understand the difference between models and algorithms
2. Learn how to select appropriate metrics for AI applications
3. Delve into the machine learning monitoring process



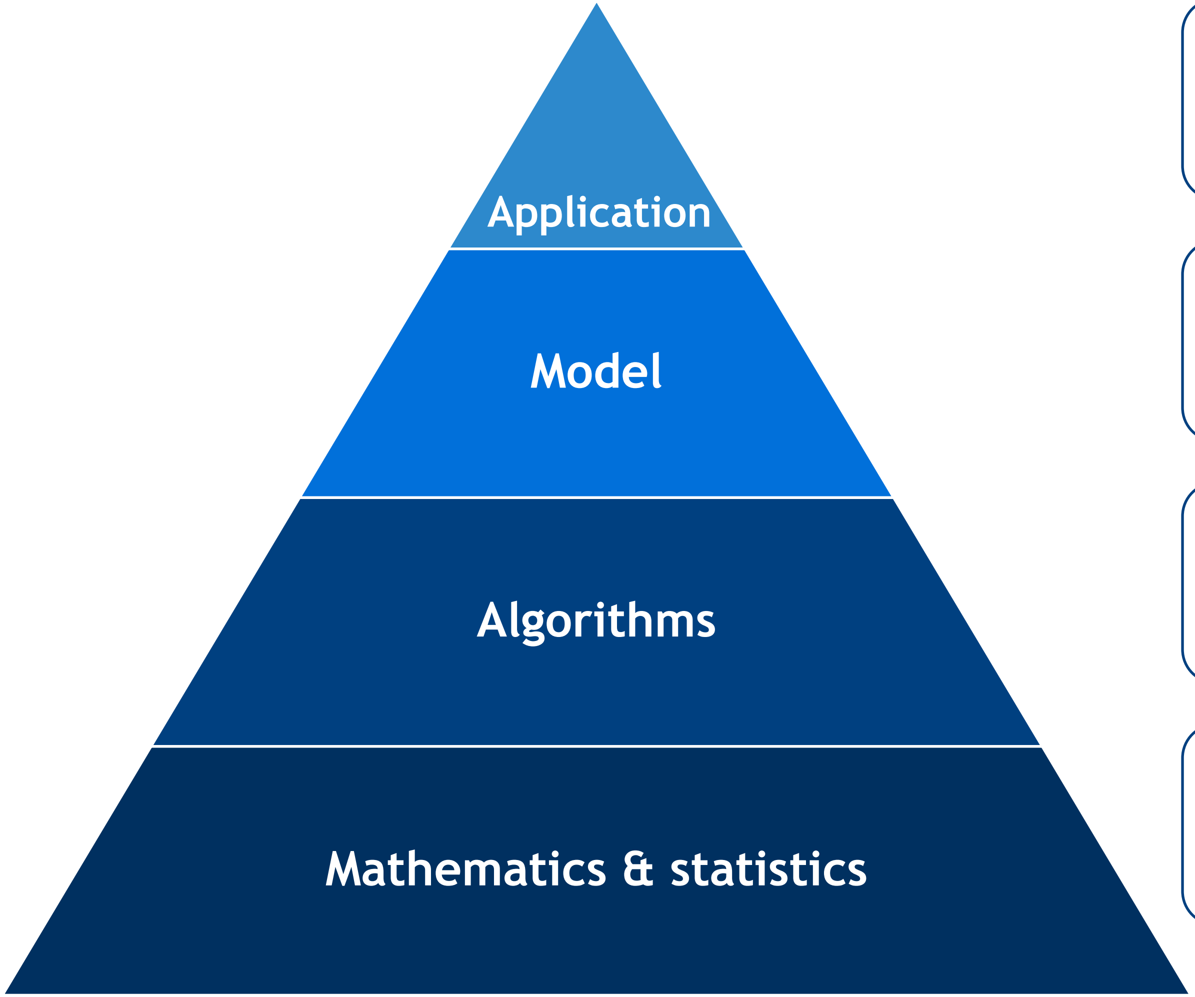
Agenda

Agenda

01 | Models

02 | Metrics and hyperparameter optimization

From mathematics to models and applications



Application: The utilization of AI models in a practical setting to achieve desired outcomes

Model: Is the output of one or many algorithms run on data; thus, it includes algorithms and data

Algorithm: A mathematical/ statistic procedure fitted on a data set

Mathematics & statistics: Calculations, formulas, quantitative methods, laws of calculation

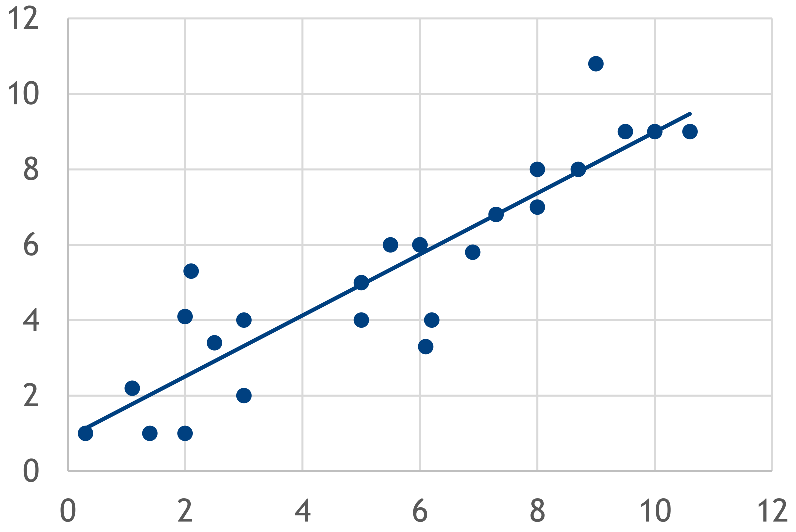
Algorithms vs. models

Algorithm: A mathematical/ statistic procedure fitted on a data set

Model: Is the output of one or many algorithms run on data. Thus, it includes algorithms and data

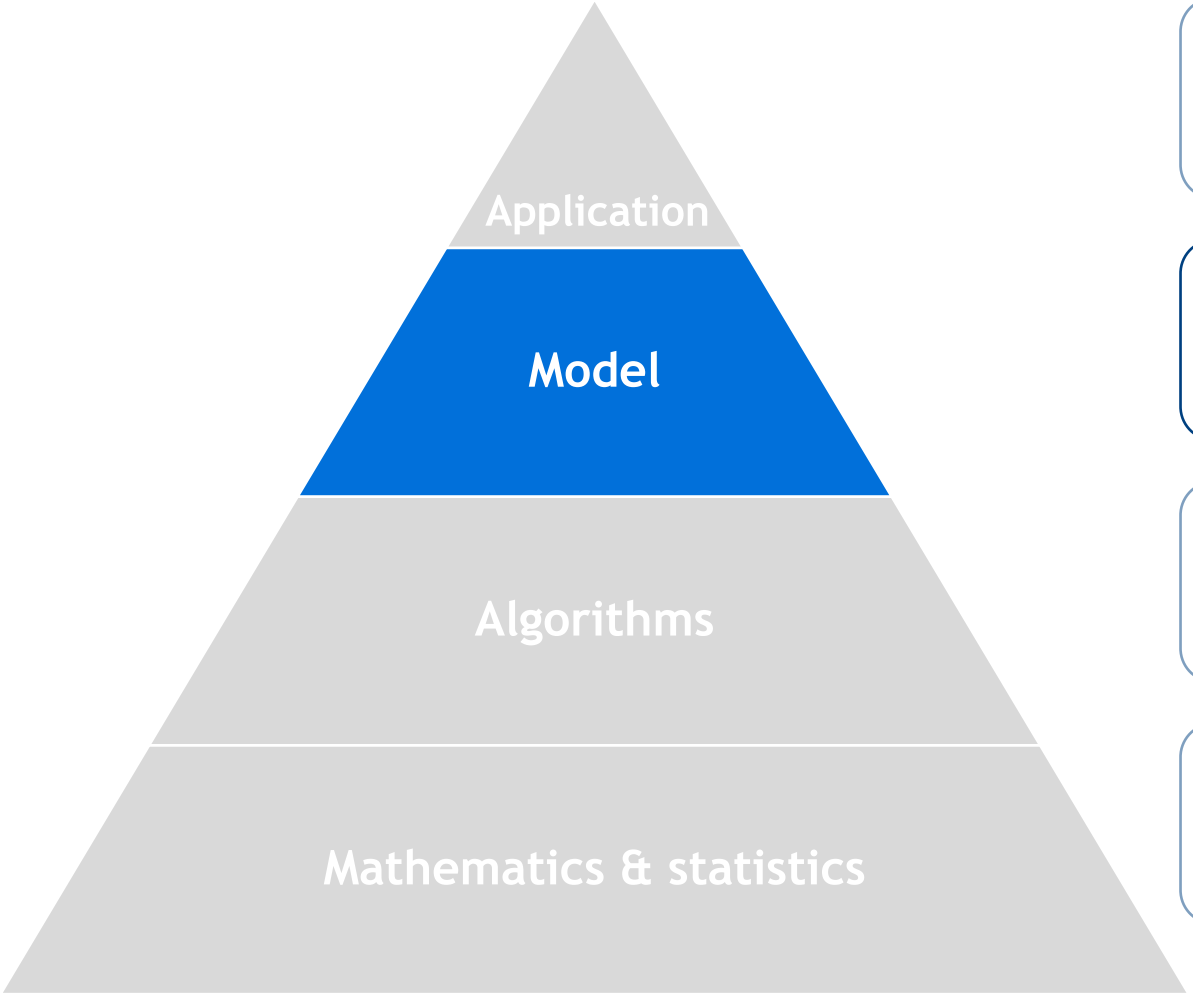
Example: Linear regression

Algorithm: Mathematical method for line fitting
Model: The specific line with calculated parameters (e.g. y=w0+w1x)



Algorithms are essential for building models that enable accurate predictions on new data, with the quality of the model being heavily dependent on the choice of the right algorithm

From mathematics to models and applications



Application: The utilization of AI models in a practical setting to achieve desired outcomes

Model: How to evaluate if a model is qualitatively good?

Algorithm: A mathematical/ statistic procedure fitted on a data set

Mathematics & statistics: Calculations, formulas, quantitative methods, laws of calculation

Quality assurance of AI models

Quality of a model: Set of characteristics that affect the suitability of a model to meet existing requirements

Was the input analyzed correctly ?

Does a satisfactory output arise ?

Is the model robust ?

Is the model sound ?

Does the model meet the legal requirements ?

How does the model perform ?

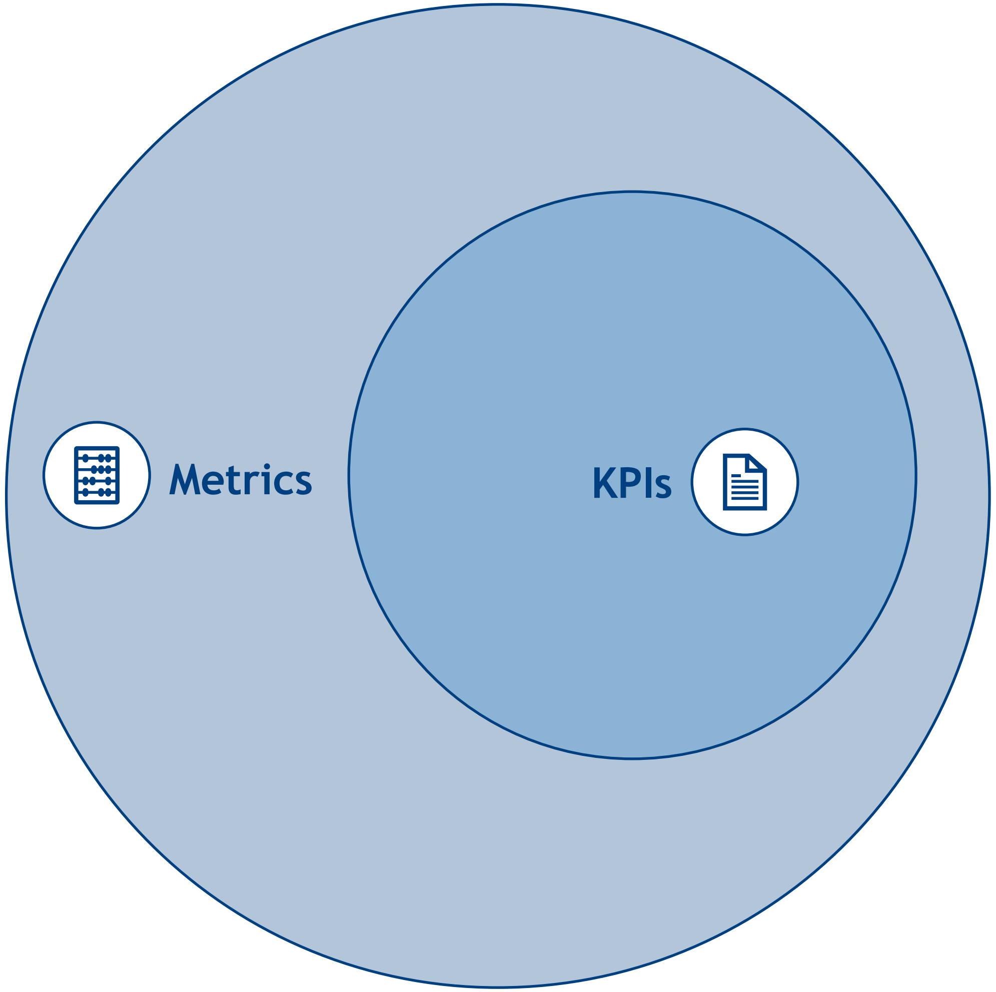
Does the model have an acceptable impact ?

Is the model fair ?

AI metrics and KPIs are used to evaluate these questions and assure the quality of the model

Overhage et al. 2012

Definitions of AI metrics and KPIs



Metrics: Metrics are quantifiable measures used for assessing, comparing, and tracking the performance of an application
e.g., accuracy of storage predictions

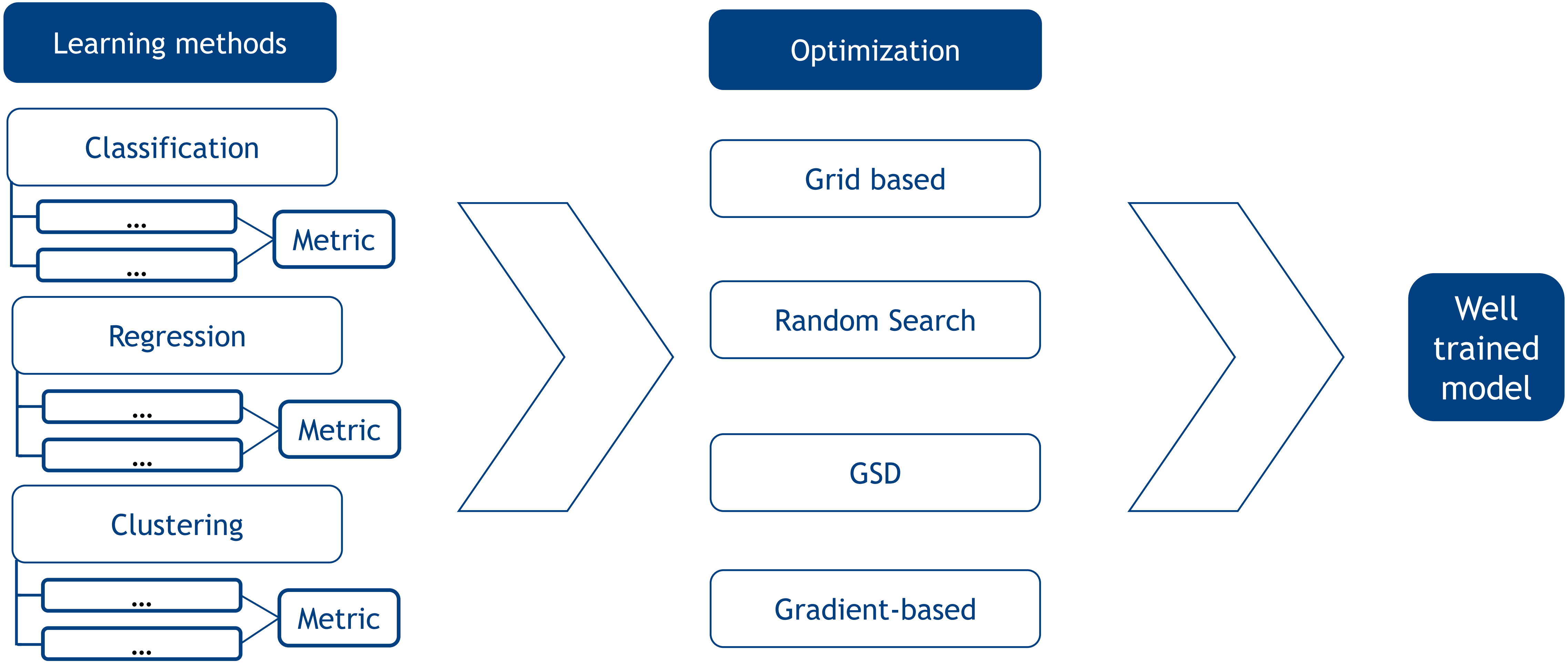
KPIs: Key performance indicators are quantifiable measurements used to gauge a company’s overall long-term financial, strategic, and operational performance
e.g., turnover ratio of a product

Agenda

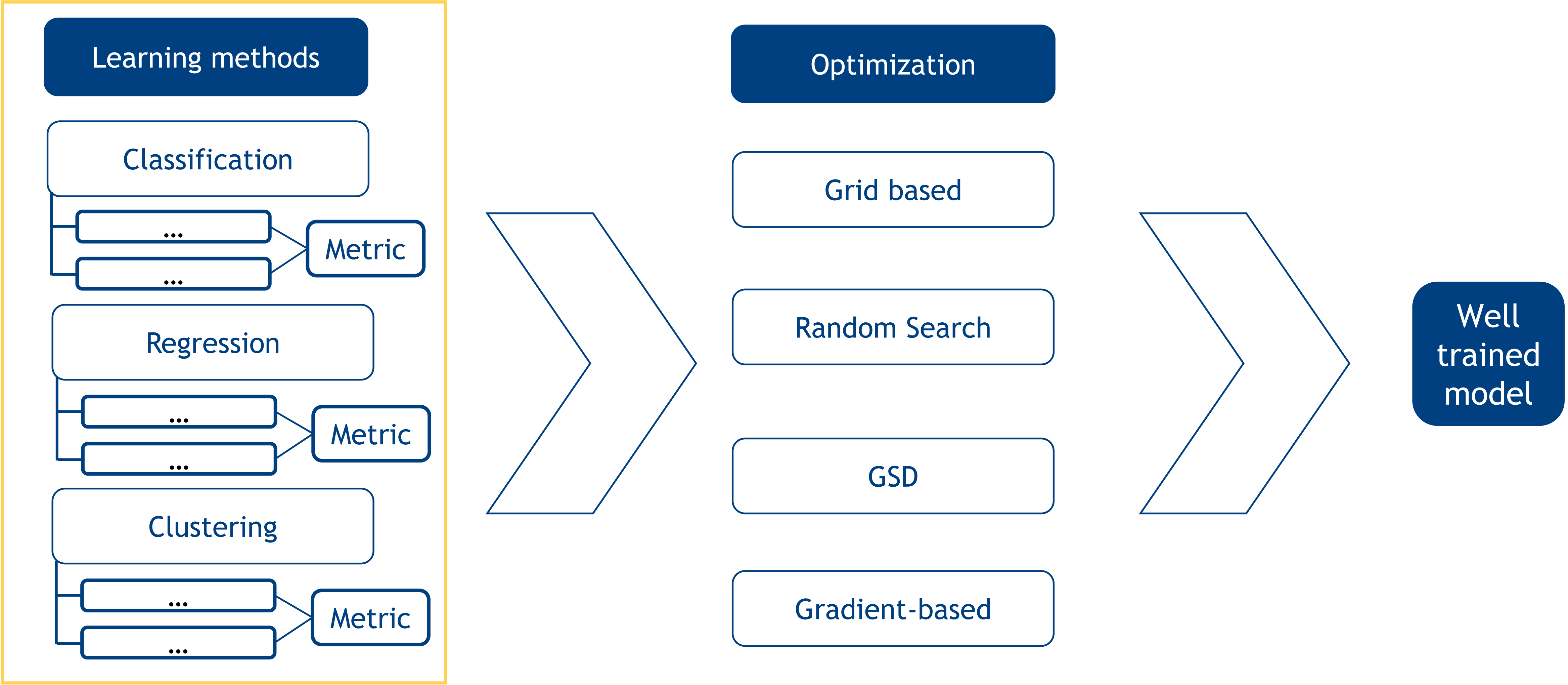
01 | Models

02 | Metrics and hyperparameter optimization

Optimization process for ML model development



Finding the right metric for every model

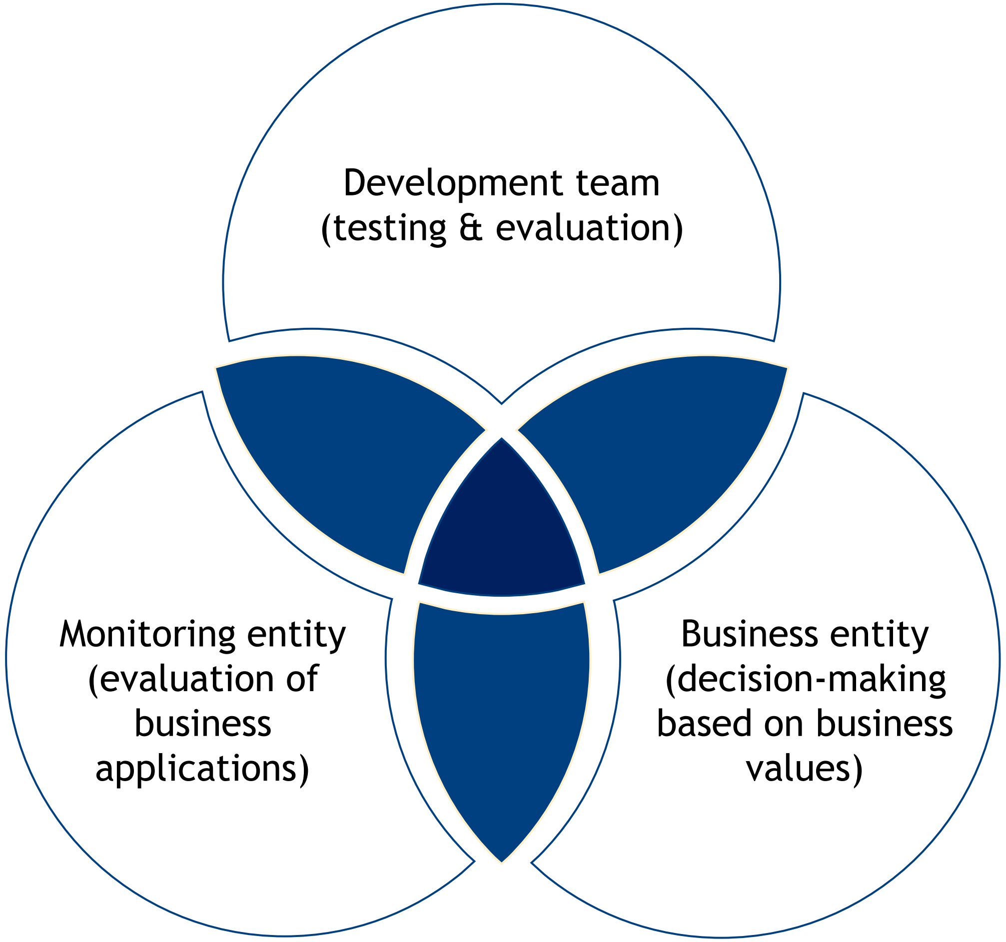


Who is involved in the decision process selecting a suitable metric?

Many **different stakeholders** are involved in the **ML lifecycle** (AI product lifecycle) and each stakeholder has a different view on as well as use for ML metrics.

Selecting metrics is currently mostly conducted by the development team.

Metrics set by the development team often neglect measurements that are relevant from a business perspective.



How can a suitable metric be selected?

AI FOR BUSINESS
BUSINESS FOR AI

Selection-approach based on ML properties

- Value different properties from a company perspective
- Use properties to assess how relevant a property is for its ML application
- Derive appropriate metrics to ensure compliance with the regarding property
- E.g., for the property correctness choose accuracy

Property	Description
Correctness	…refers to a model's ability to make accurate judgments, which can be measured through metrics such as accuracy, precision, recall, and F1 score in various tasks.
Complexity	…refers to a model's need for computational resources, including time to train and space to store it, with efficient models being more scalable and suitable for resource-limited environments.
Consistency	…refers to a model's reliable and stable performance across different datasets or partitions, with consistent models being less prone to fluctuations due to varying data.
Fairness	…refers to a model's ability to make unbiased decisions across different groups, with fairness metrics assessing whether predictions are biased towards or against specific demographics.
Inter-pretability	…refers to how understandable a model's outputs and decision-making process are, with clearer results often being easier to verify or justify in high-stakes environments like healthcare or finance.
Responsive-ness	…refers to a model's efficiency in executing decisions quickly and accurately in real-time, with metrics like latency and throughput being used to measure its speed in applications such as autonomous systems or financial trading.
Robustness	…refers to a model's ability to maintain performance despite adversarial conditions or unexpected inputs, with robust models resisting manipulation and maintaining consistent results even in challenging situations.
Safety	…in AI refers to protecting models and data with techniques like differential privacy, ensuring privacy and security compliance with regulations like GDPR and the AI Act.

European Comission (2019), Ali et al. 2017, Rutinowski et al. 2024

Selected AI evaluation metrics

Classification	Clustering	NLP	Computer Vision
• Accuracy • F1-Score • True Positive Rate • Precision • …	• Silhouette • Adjusted Rand Index • Adjusted Mutual Information • …	• Perplexity • BLEU-Score • …	• Peak signal-to-noise ratio • Structural similarity • …

Regression	Reinforcement Learning	Fairness	…
• Mean Absolut Error • Mean Squared Error • R^2 • …	• Dispersion Across Time • Dispersion Across Runs • Risk Across Time • …	• Statistical Parity • Theil Index • …	• …

The use of various metrics is essential to ensure that a machine learning model is working correctly and optimally

Breck et al. (2017), Rácz et al. (2019)

(Binary) Classification – Confusion matrix

		Actual values +	Actual values -
Predicted values	+	True positives (TP)	False positives (FP)
Predicted values	-	False negatives (FN)	True negatives (TN)

Classification metrics can be calculated using only these four different values (TP, FP, FN, TN)

Accuracy

$$= \frac{\#\ correct\ predictions}{\#\ total\ predictions}$$

$$= \frac{sum(TP, TN)}{sum(TP, FP, FN, TN)}$$

Shrivastav 2020

Deep dive – Classification metrics

$$\text{Accuracy} = \frac{\#\ correct\ predictions}{\#\ of\ predictions}$$	**Accuracy** describes the relation between correct classifications and the total number of input samples and gives a first insight on an ML application's overall performance
$$\text{Precision} = \frac{True\ Positive}{False\ Positive + True\ Positive}$$	**Precision** is the proportion of correctly predicted positive cases regarding all positive predicted cases
$$\text{TPR} = \frac{True\ Positive}{False\ Negative + True\ Positive}$$	**Recall (Sensitivity)** is the proportion of correctly predicted positive cases regarding all positive cases
$$\text{F1-Score} = 2\ x\ \frac{Precision\ x\ Recall}{Precison{+}Recall}$$	**F1-Score** is the harmonic mean between precision and recall

Deep dive – Regression metrics

AI FOR BUSINESS BUSINESS FOR AI

$\text{MAE} = \frac{1}{N} * \sum_1^N \lvert y_i - \hat{y}_i \rvert$	**Mean absolute error** measures the average magnitude of the errors in a set of predictions, without considering their direction
$\text{MSE} = \frac{1}{N} * \sum_1^N (y_i - \hat{y}_i)^2$	**Mean squared error** squares the difference between actual values and predicted values, making MSE helpful when outliers occur as the difference is more penalized
$\text{RMSE} = \sqrt{\frac{1}{N} * \sum_1^N (y_i - \hat{y}_i)^2}$	**Root mean squared error** is the standard deviation of prediction errors in the model; it states the concentration of data points around the regression line
$\text{R}^2 = \frac{\sum_1^N (\hat{y}_i - \bar{y})^2}{\sum_1^N (y_i - \bar{y})^2} = \frac{\textit{explained var}}{\textit{total var}}$	**R^2** measures the proportion of the total variation in the dependent variable that can be explained by the independent variables in the model

y_i = (real) single data point, $\hat{y}_i$ = predicted value for a single data point, $\bar{y}$ = mean of actual values

Deep dive – Clustering metrics

$$\text{Silhouette} = \begin{cases} 0 & \text{if } o \text{ is the only element of } A \\ \dfrac{dist(B,o) - dist(A,o)}{max\{dist(A,o), dist(B,o)\}} & \text{otherwise} \end{cases}$$

Silhouette is a measurement of how well the assignment of an element to the two nearest cluster is

$$\text{ARI} = \frac{RI - E(RI)}{max(RI) - E(RI)} \text{ with RI} = \frac{a+b}{\binom{n}{2}}$$

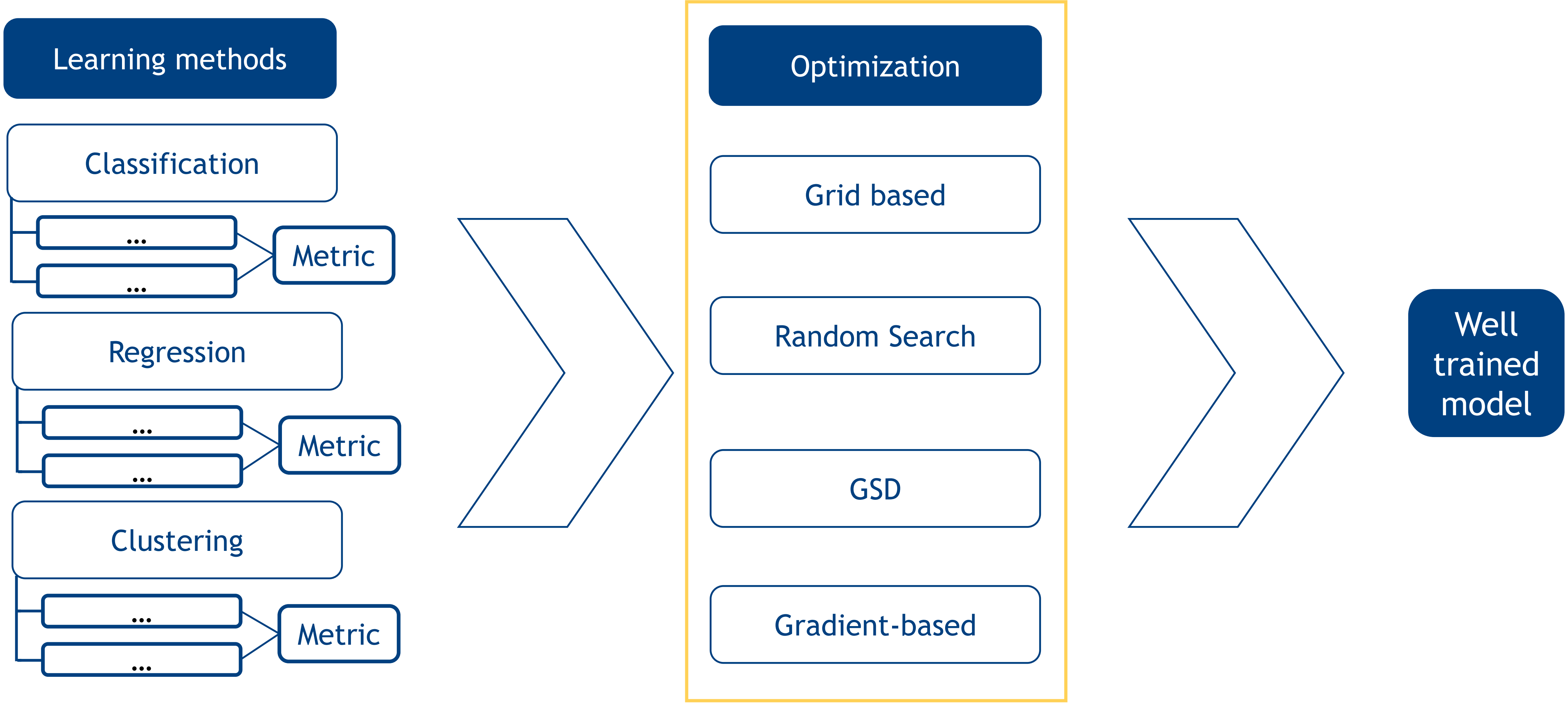
Rand Index (RI) is a measurement for the similarity between two clusters. Adjusting the RI for chance grouping leads to the **Adjusted Rand Index**

$$\text{AMI} = \frac{MI(U,V) - E(MI(U,V))}{avg\big(H(U),H(V)\big) - E(MI(U,V))}$$

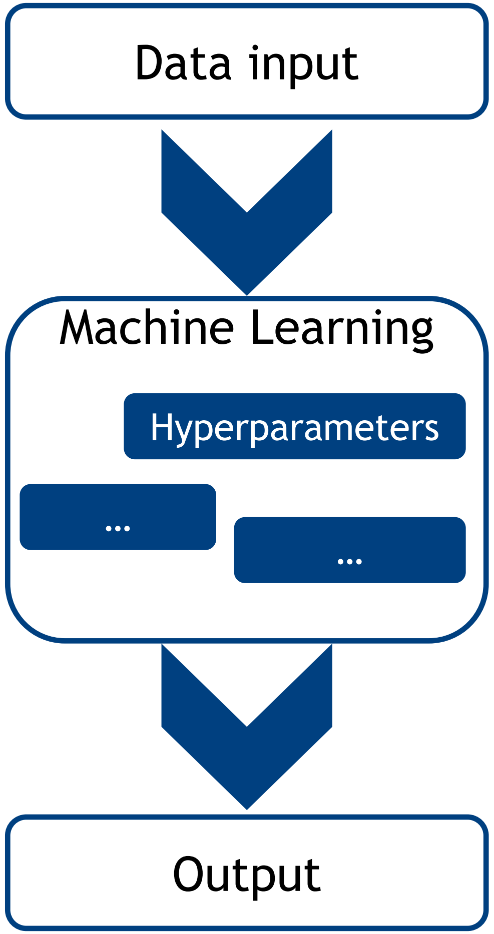
Mutual information (MI) measures non-linear relations between two clusters. Since MI is higher for two clusters with more clusters, regardless of whether there is more information shared, **Adjusted Mutual Information (AMI)** is adjusted to account for chance

a = # pairs of elements that are in the same subset, b = # pairs of elements in different subsets, n = # of elements

Hyperparameter optimization



Hyperparameter optimization

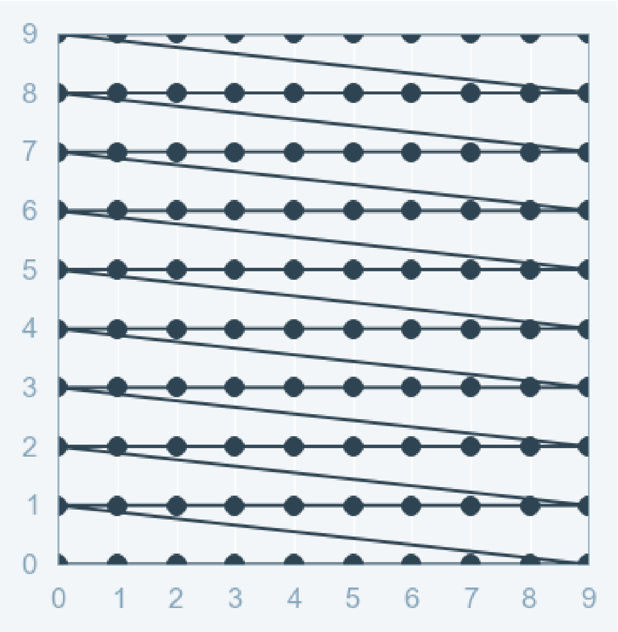


- In general, the quality of the **selected input data** has a great influence on the result of machine learning
- In addition to the input data, **other parameters** also have an influence on the quality of the solution found by ML
- These parameters are called **hyperparameter** and can differ depending on the ML method
- **Exemplary hyperparameters** are the learning rate, the choice of activation function and the number of hidden layers in neural networks
- The metrics presented can be used **in combination with different optimization methods** to find the best possible hyperparameter

Yang, L. & Shami, A. (2022)

Grid search optimization methods

- Provides a simple way to find good results by using **brute-force methods**
- Tests **every combination** of every possible value in a predefined range (search space)
- In order to get sufficiently good solutions in a reasonable time, it is necessary to limit the search space and the step size based on previous results of well-performing hyperparameter configurations
- Even though the GS is **very easy to use**, it quickly becomes **quite inefficient** for large search spaces



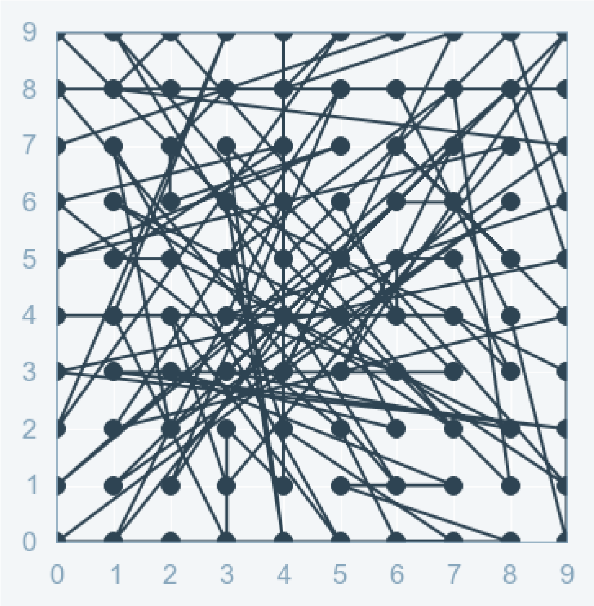
Visual Representation of grid search

https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318

Yang, L. & Shami, A. (2022)

Random search optimization method

- The RS works very similar to the GS, but just **tests a predefined sample** of possible parameter combinations
- The theories behind this is that if the configuration space is large enough, then **the global optimums**, or at least their approximations, will be detected
- It is also **easier to control the allocation of resources**, as a predefined number of combinations are always tested, allowing promising areas to be investigated more frequently



Visual Representation of Random search

https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318

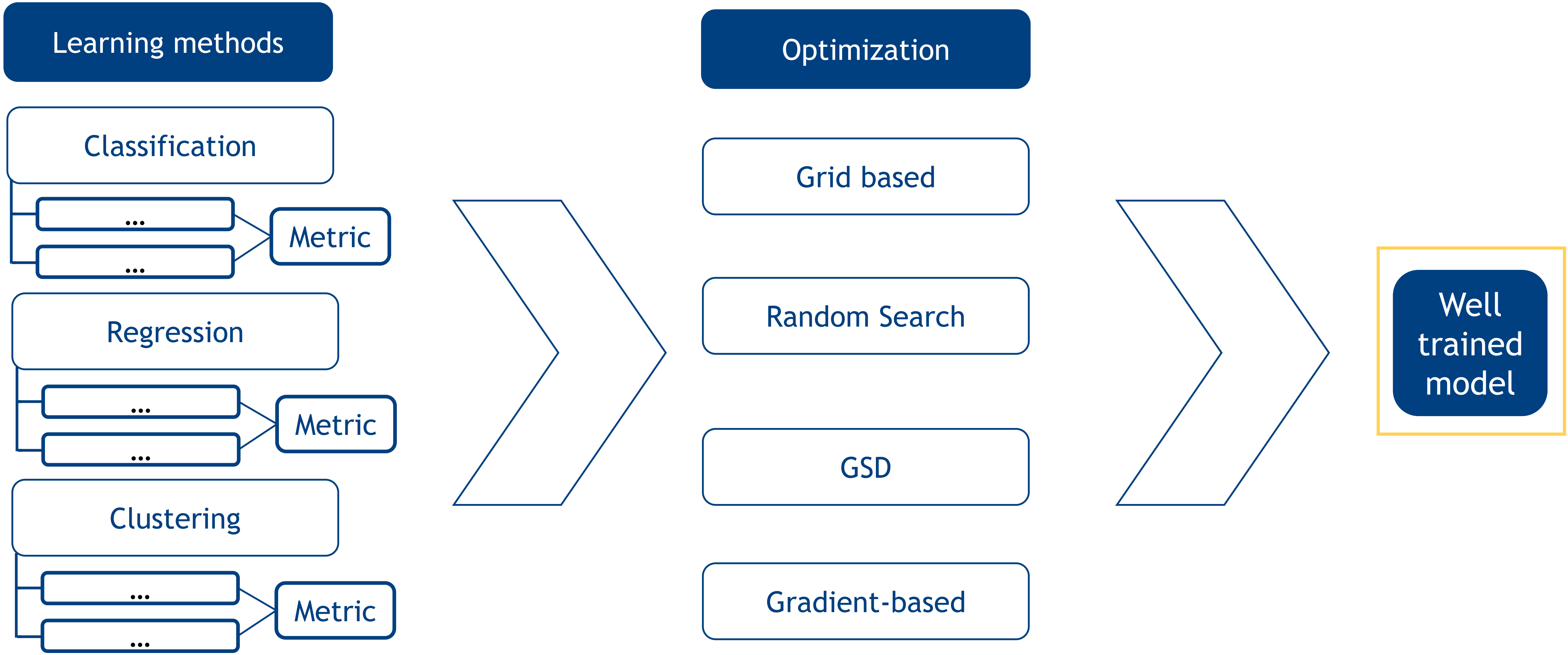
Yang, L. & Shami, A. (2022)

Grad student and gradient-based optimization methods

Grad student descent (GSD)	Gradient-based optimization
Also known as “trial and error”; the researcher tests as many possible hyperparameters as the given time allows	The gradient descent calculates the gradient of variables to identify the promising direction and moves towards the optimum
The quality of the results is based on experience, the analysis of previously-evaluated results, or guessing	For ML algorithms, the gradient of some hyperparameters can be calculated, the gradient descent can be used to find an optimum
For models with a large number of hyperparameters, GSD method often produces infeasible results	Depending on the parameter and algorithm, its possible to find just a local optimum

Yang, L. & Shami, A. (2022)

When do you know, your model is ready?

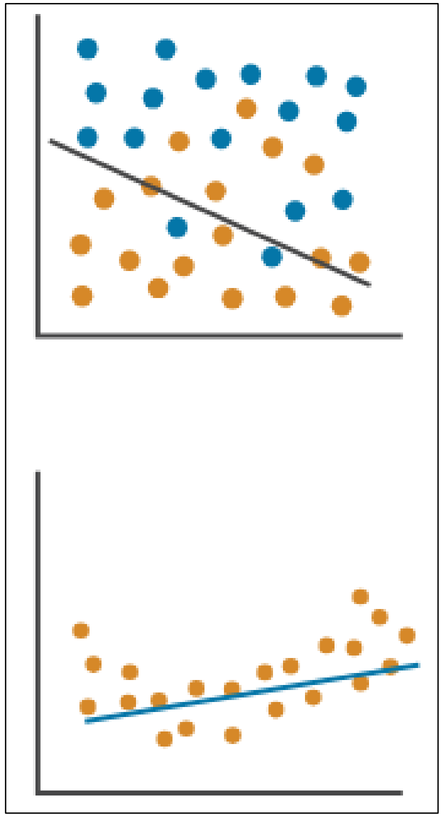


Is your model ready when the result looks like…

… Case 1 ?:

Classification

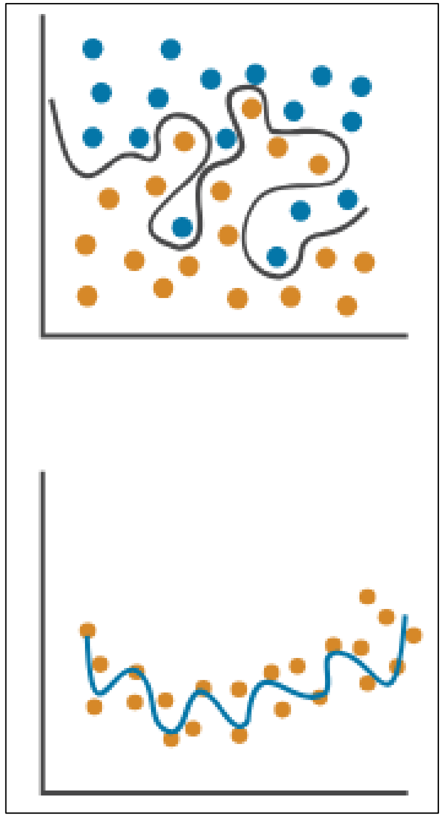
Regression



The model should separate the balls of different colors into two spaces

The model should create a function with the smallest possible distance to all points

Or like Case 2?:



https://de.mathworks.com/discovery/overfitting.html

Reasons for the different results

Option 1:

- The training of the model has been stopped to early and the model is **underfitting**
- Other reason for underfitting: Underfitting can occur if the model's parameters are not adequately tuned, which can result in poor performance on both the training data and unseen data

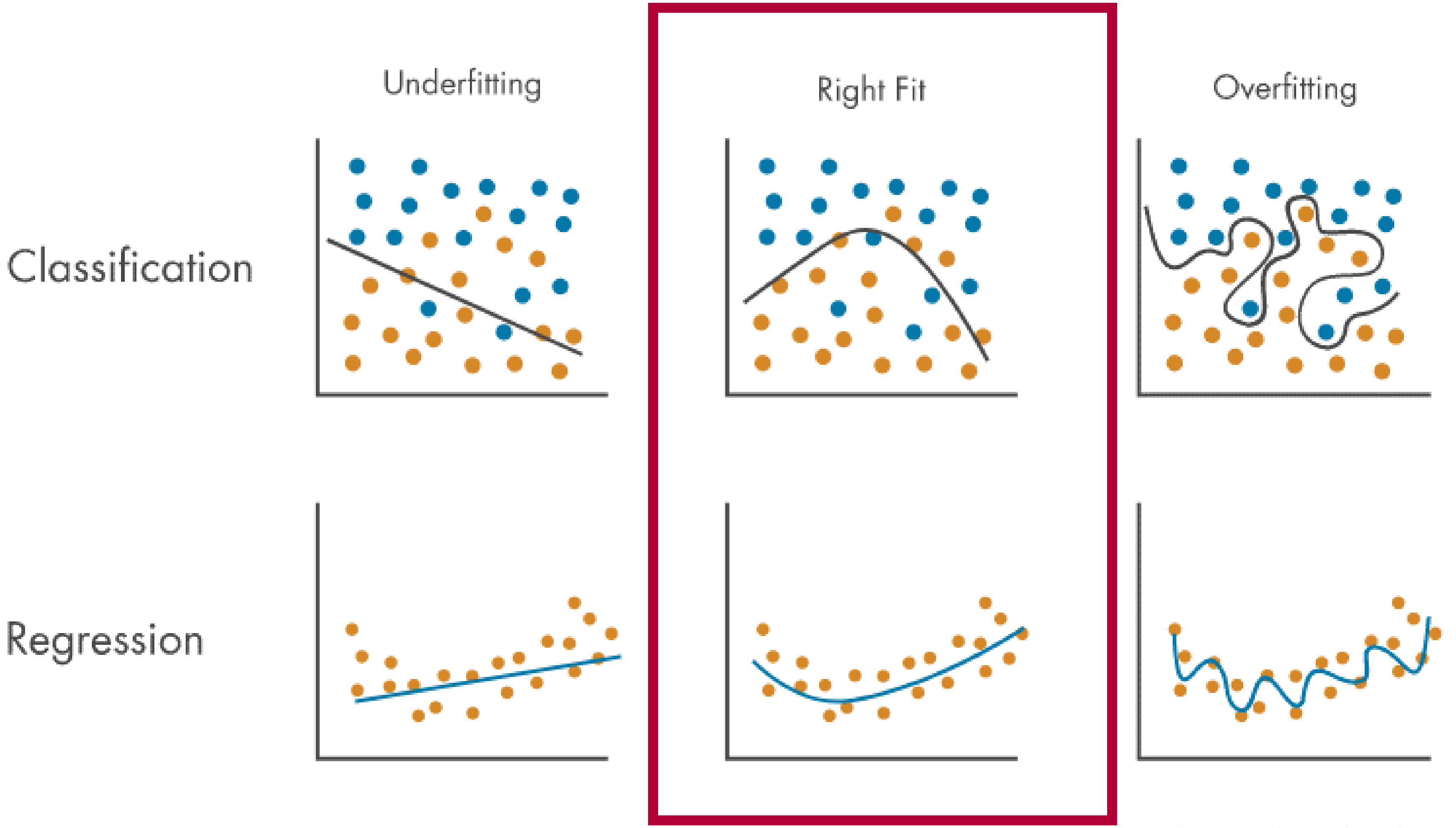
Option 2:

- The training of the model was stopped too late, and the model fits too well, its **overfitting**
- Other reasons for overfitting: The training set is too small or contains fewer representative data so that the perturbations in the data can be learned by the model and later used as a basis for prediction

To get a good result, the proper training data must be selected, and the training must be finished at the right time

Ying (2018)

The art is to find the right scope of model training



https://de.mathworks.com/discovery/overfitting.html

Today’s lecture at a glance

- We have distinguished models from algorithms and worked out requirements
- We have understood how appropriate metrics are selected for evaluation purposes and who is involved in the decision-making process
- We learned different metrics for machine learning methods

Scientific references

- Breck, Eric; Cai, Shanqing; Nielsen, Eric; Salib, Michael; Sculley, D. (2017): The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. In: Proceedings of IEEE Big Data.
- Even, A., Shankaranarayanan, G.: Utility-Driven Assessment of Quality. In: The DATA BASE for Advances in Information Systems 38 (2007) 2, S. 75-93.
- Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- Flach, Peter. "Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
- Haas, Christian (2019): The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness. In: Fortieth International Conference on Information Systems.
- Krogstie, John. "Understanding and Assessing Quality of Models and Modeling Languages." IGI Global, 2019
- Overhage, Sven, Dominik Q. Birkmeier, and Sebastian Schlauderer. "Quality marks, metrics, and measurement procedures for business process models: the 3QM-framework." Business & Information Systems Engineering, 2012
- Pipino, L., Lee, Y. W., Wang, R. Y.: Data Quality Assessment. In: Communications of the ACM 45 (2002) 4, S. 211-218.
- Rácz, Anita; Bajusz, Dávid; Héberger, Károly (2019): Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. In: Molecules 24 (15). DOI: 10.3390/molecules24152811.
- Yang, L. & Shami, A. (2022):“On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice“, Department of Electrical and Computer Engineering, University of Western Ontario

Non-scientific references

- Difference Between Algorithm and Model in Machine Learning | LinkedIn
- AI Quality - the Key to Driving Business Value with AI - TruEra
- Metrics Definition (investopedia.com)
- Key Performance Indicator (KPI): Definition, Types, and Examples (investopedia.com)
- 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics | by Shervin Minaee | Towards Data Science
- Confusion Matric(TPR,FPR,FNR,TNR), Precision, Recall, F1-Score | by Namratesh Shrivastav | DataDrivenInvestor
- What metrics should be used for evaluating a model on an imbalanced data set? (precision + recall or ROC=TPR+FPR) | by Shir Meir Lador | Towards Data Science
- Wirkungsanalyse, Monitoring, Evaluation | PHINEO (wirkung-lernen.de)
- https://de.mathworks.com/discovery/overfitting.html