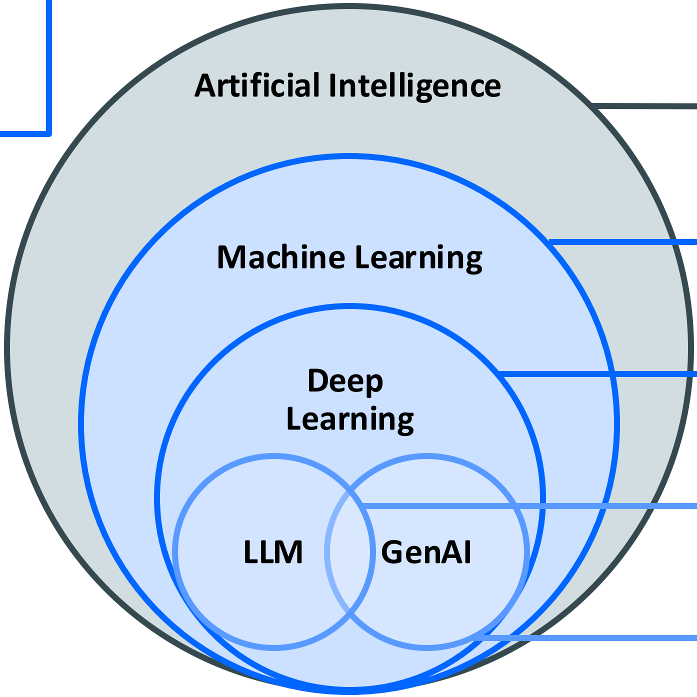
Tag 6: Einführung in maschinelles Lernen

Samuel Schlenker
11.11.2025, WWI 2025F

Students should ...

- Define machine learning as a subset of artificial intelligence focused on learning from data
- Understand machine learning as an optimization task and parametric programming approach
- Differentiate between supervised, unsupervised, and reinforcement learning paradigms
- Apply the train-test split methodology and understand cross-validation
- Understand classification tasks and algorithms (K-Nearest Neighbors, Decision Trees, SVM)
- Understand regression tasks for predicting continuous variables
- Apply unsupervised learning techniques like K-means clustering for pattern discovery
- Recognize the differences between supervised and unsupervised learning approaches
- Understand the basic architecture of neural networks (input layer, hidden layers, output layer)
- Identify key hyperparameters (learning rate, batch size, epochs, activation functions)
- Understand the role of activation functions in neural networks
- Recognize the model training process and the importance of loss/cost functions

Machine Learning



Artificial Intelligence (AI)
Any technology that enables machines to solve tasks in a way like humans do

Machine Learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed (supervised & unsupervised)

Deep Learning (DL)
Using deep artificial neural networks as models, inspired by the structure and function of the human brain

Large Language Models (LLM)
Models trained on massive datasets to understand and generate human-like text across diverse subjects

Generative Artificial Intelligence (GenAI)
Refers to technologies that utilize machine learning models to generate human-like text, images, or other content

Machine Learning

Machine Learning is a **subset** of **AI** focused on developing algorithms that enable computers to learn from and make predictions on data.
The transition from AI to ML represents a shift from rule-based systems to **data-driven approaches**.

It's motivated by the goal of **reducing the need for explicit programming**, allowing machines to adapt to new scenarios independently.

Machine Learning

Generic term for the artificial generation of knowledge based on experience.

In machine learning, a system

- learns based on examples
- generalizes knowledge (after the learning phase)
- recognizes patterns and laws (based on learning data)
- can evaluate unknown data (learning transfer) or
- fail to learn unknown data (e.g. under- or overfitting)

Machine Learning

Machine learning as an optimization task
Machine learning as probabilistic inference
Machine learning as parametric programming
Machine learning as an evolutionary search
...

Modeling through learning

Differentiation

Model

- Representation of a situation (reality)
- Representation using a modeling language

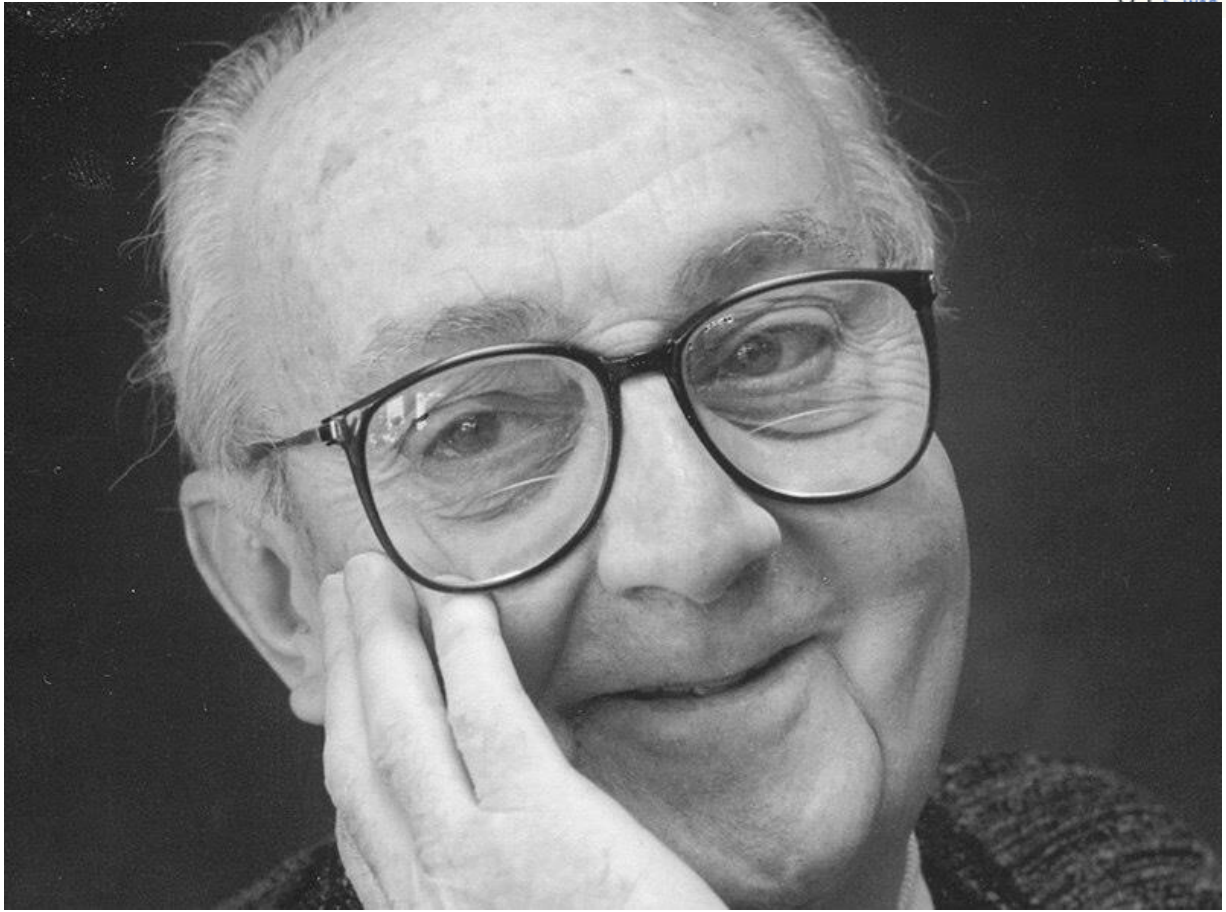
Modeling language

- Consists of a class of artifacts
- Allows these artifacts to be characterized in the form of descriptive properties and relationships between the characterized artifacts
- Describes the models belonging to this modeling language as artifact characteristics

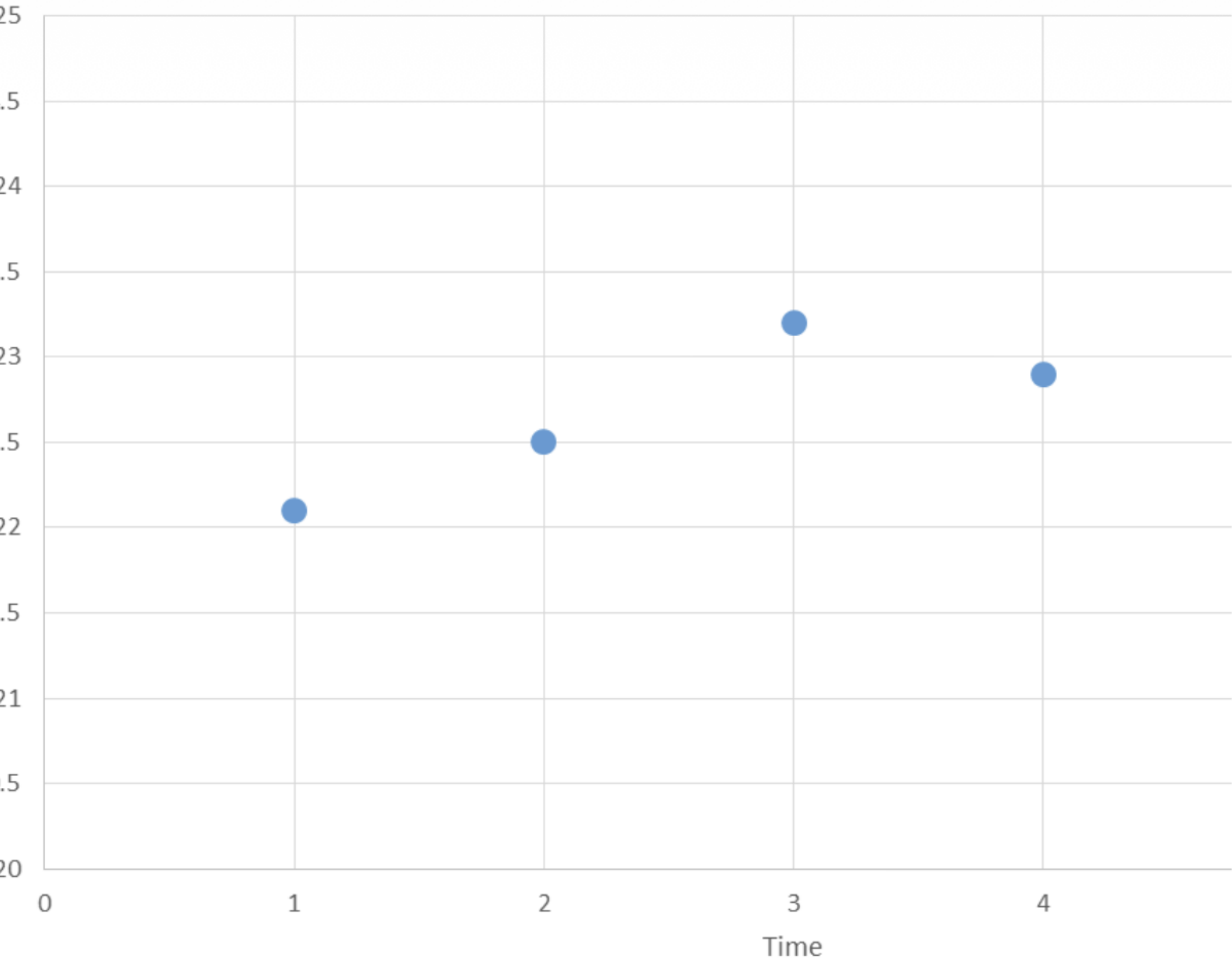
Metamodel

- A modeling language of a model (M1) can itself be regarded as a model (M2) that can be represented by a modeling language.
- This model M2 is called the metamodel for M1 ... (!)

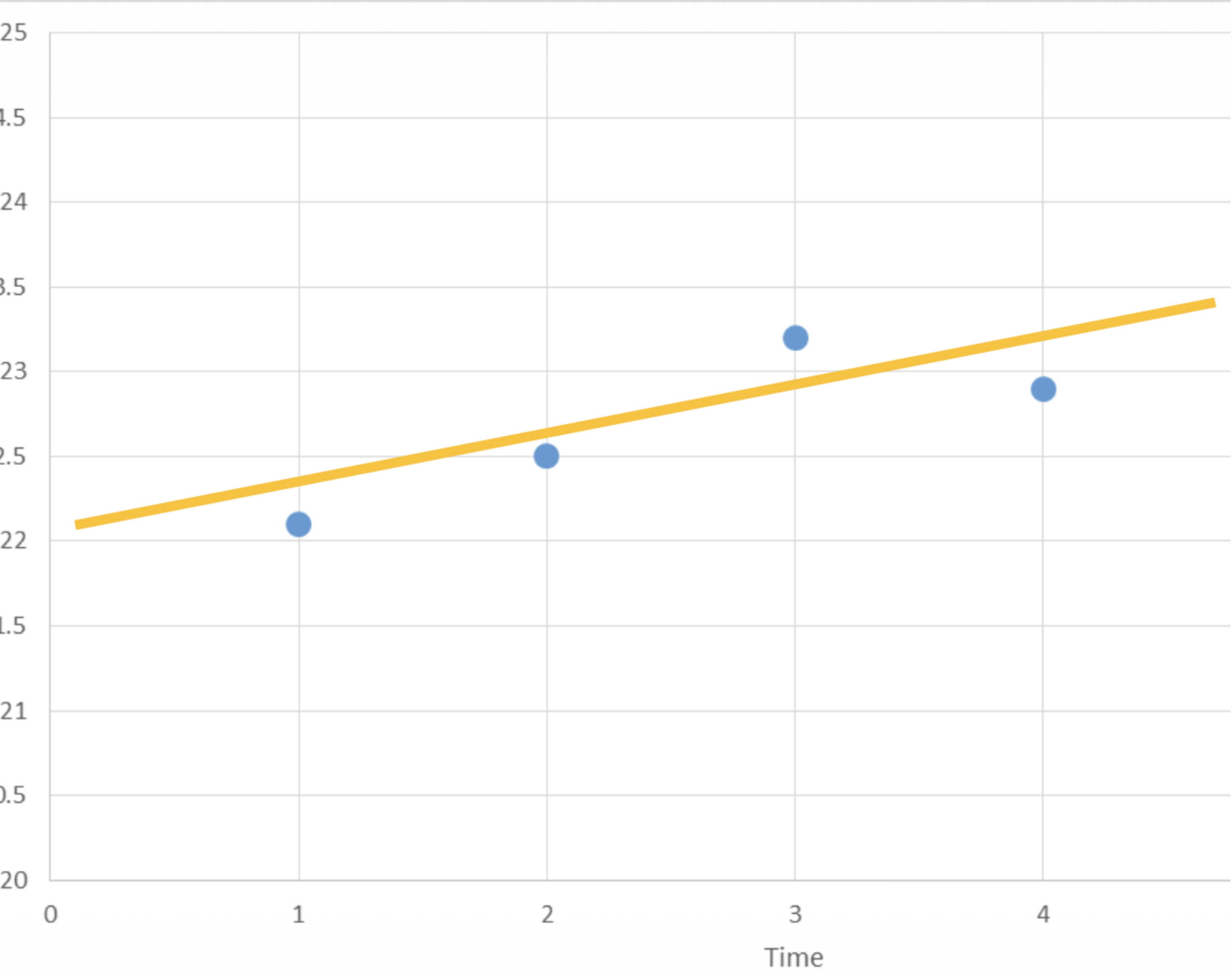
All Models are wrong but some are useful



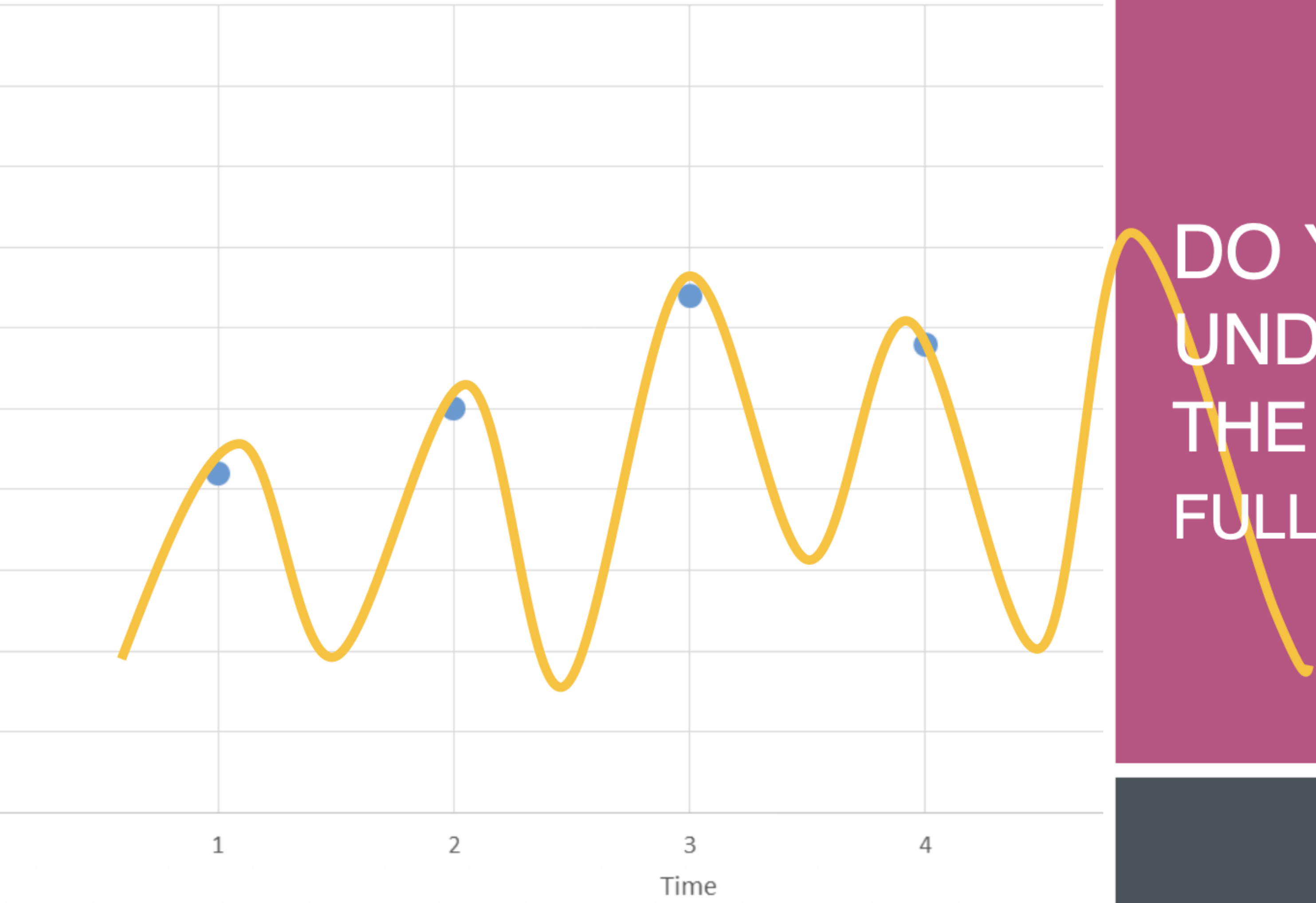
George Box



DO YOU UNDERSTAND THE DATA FULLY?



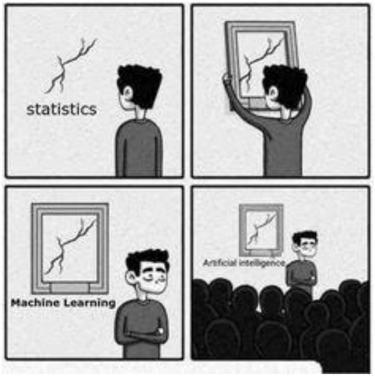
DO YOU UNDERSTAND THE DATA FULLY?



DO YOU UNDERSTAND THE DATA FULLY?

Machine Learning

Input x → **Modell** → **Output** y

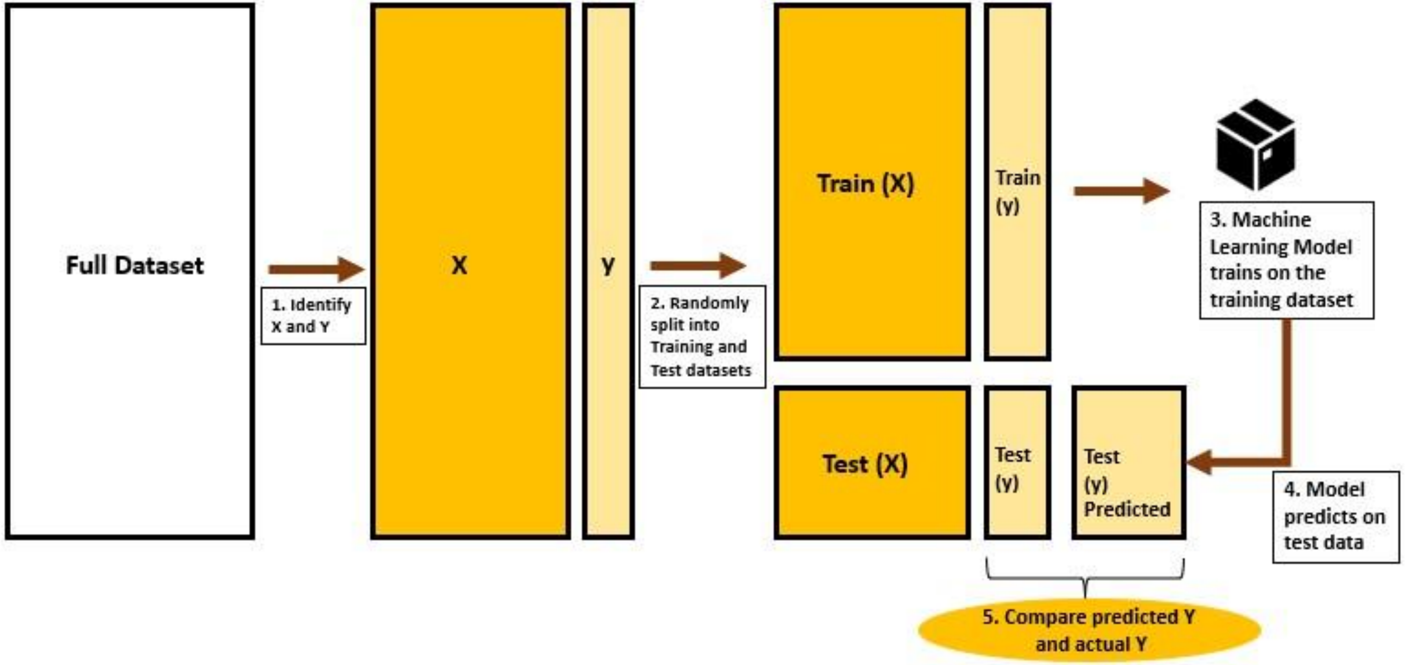


Training: the model is trained using a training data set where the output/target variable is known

Testing: the model is tested using a test data set where the output/target variable is known in order to check the accuracy of the model

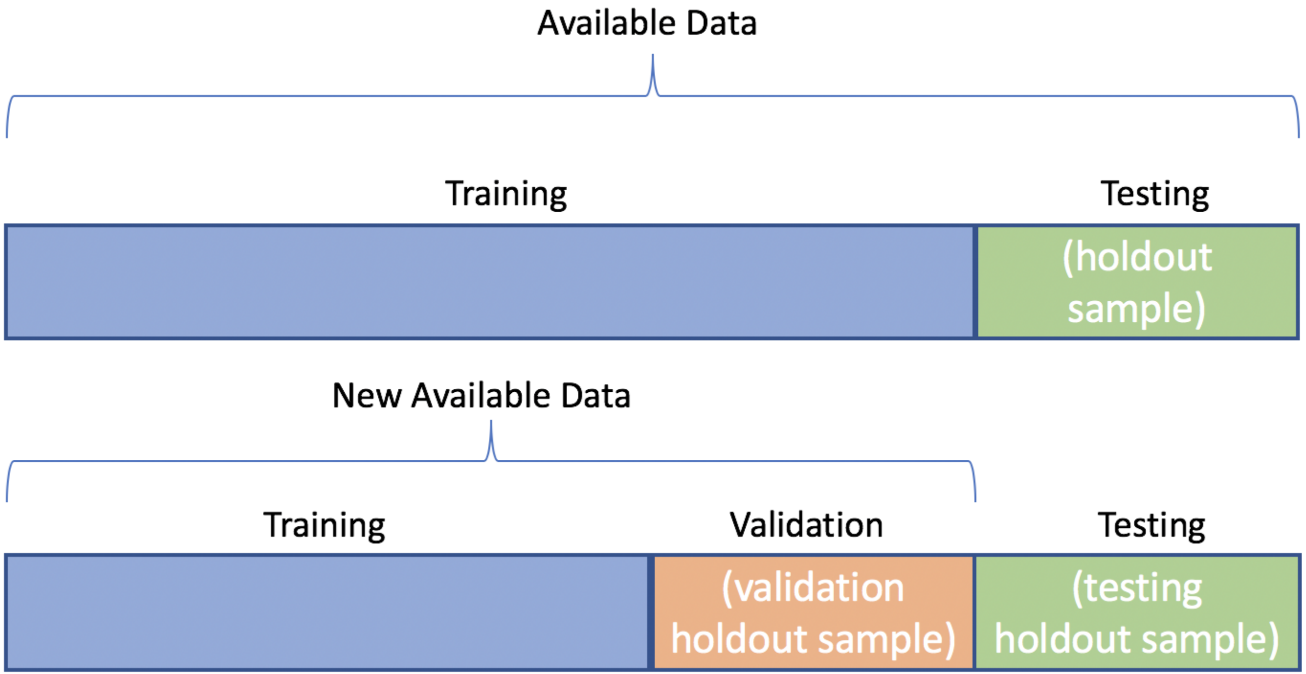
Application: the model delivers results for a new data set with an unknown target variable

Train test split

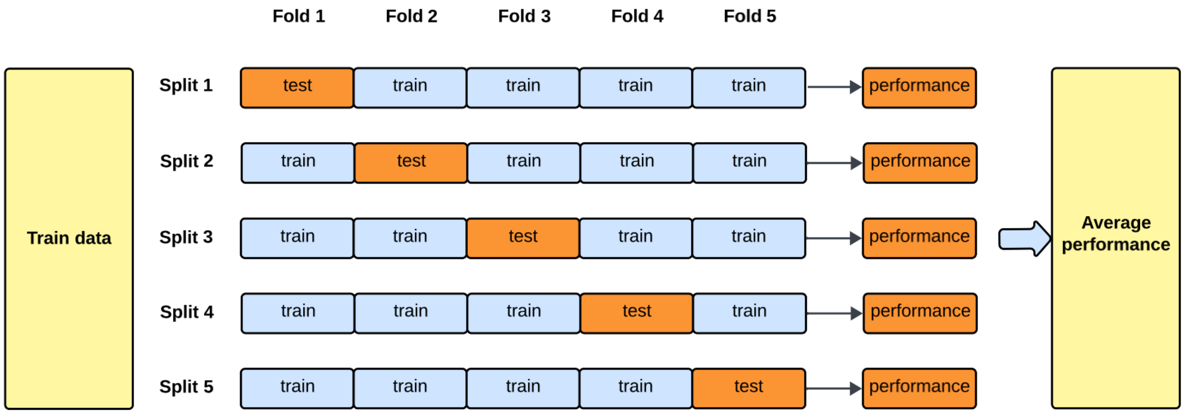


INTRODUCTION TO DATA SCIENCE BY W&S
YOUR TURN!
EXERCISE / QUESTION

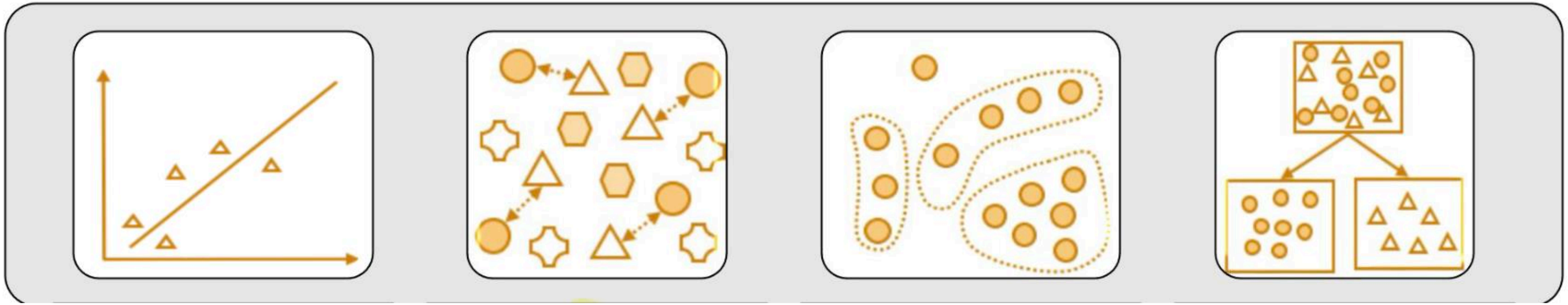
Validation data



Cross validation

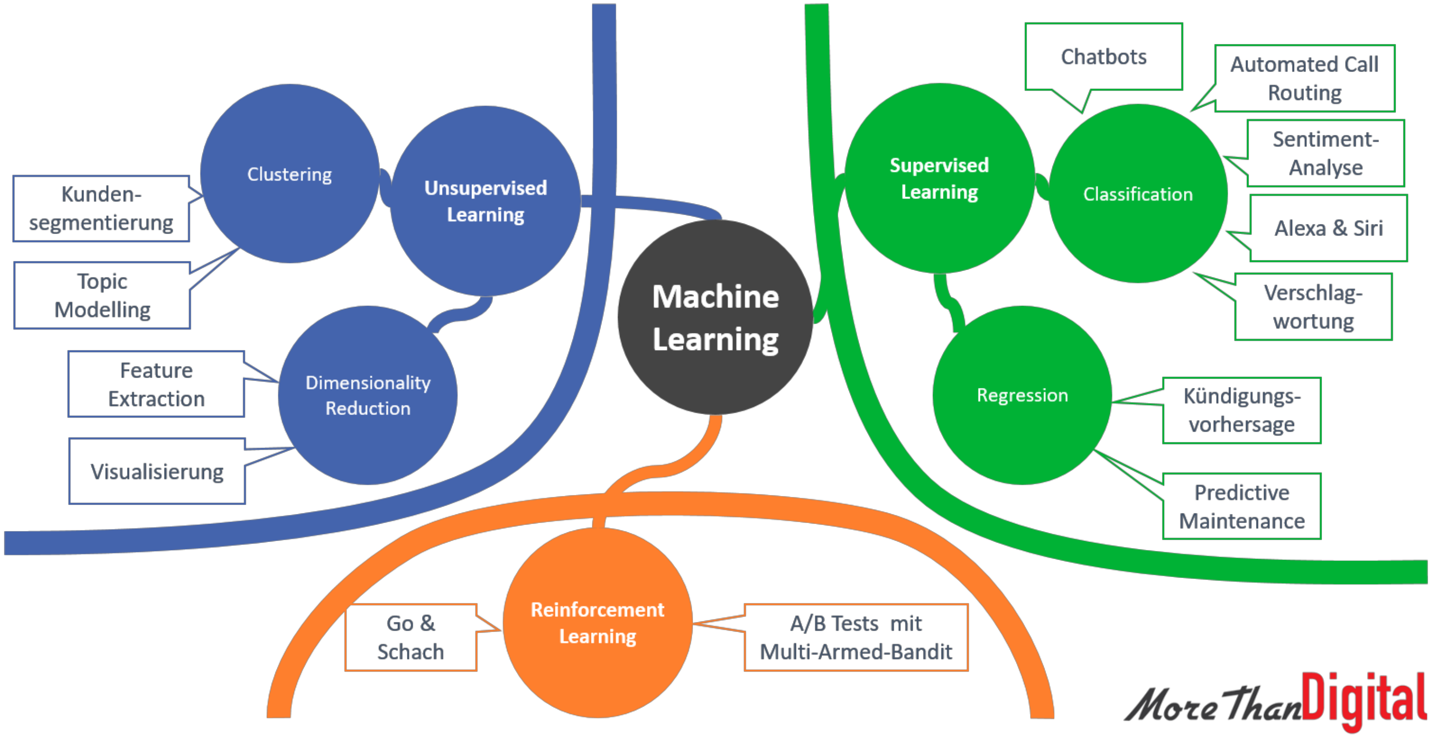


Modeling in general



Forecasts	Association	Segmentation	Classification
Identification of trends	Search for dependencies	Finding homogeneous subsets	Division into predefined classes
Sales and revenue forecasts (e.g., sales/production planning)	Analysis of shopping baskets (product recommendations)	Creating customer portfolios (differentiated marketing)	Churn analysis (customer loyalty measures)

Machine Learning

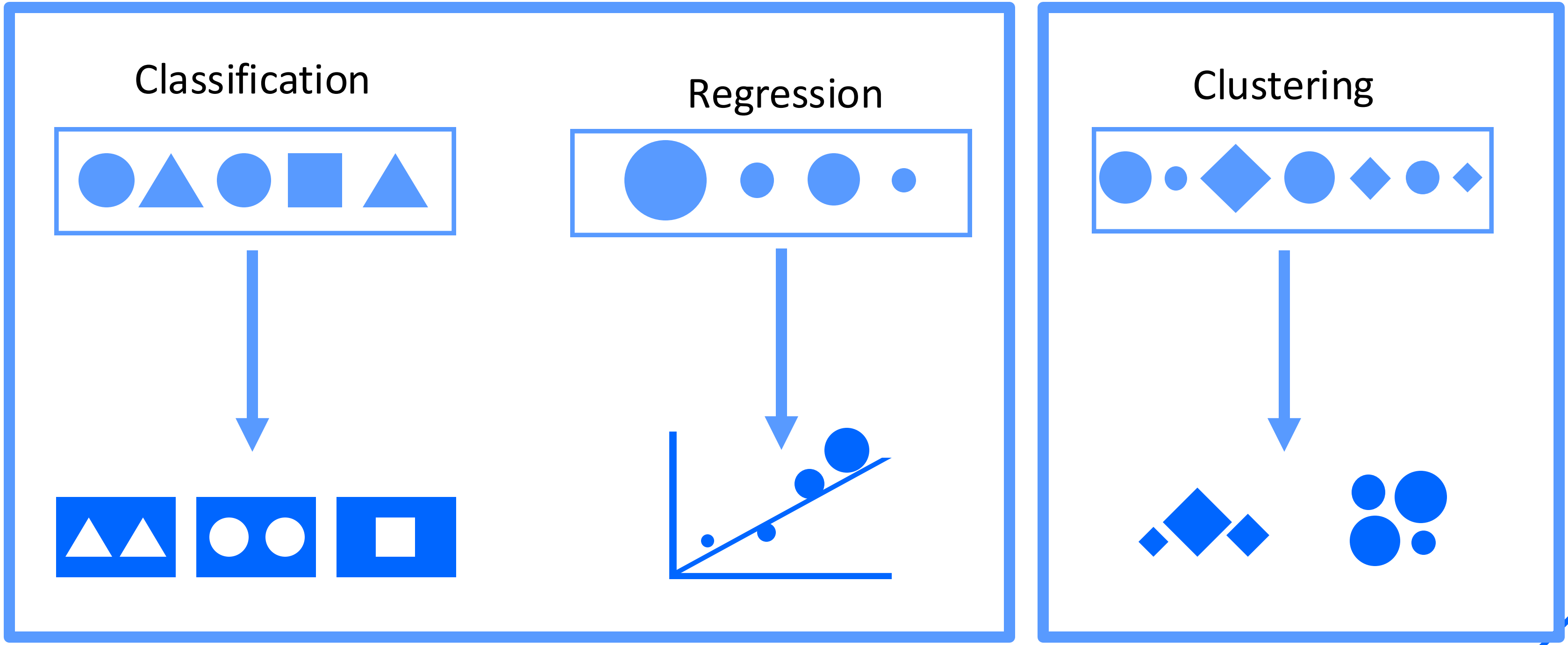


https://morethandigital.info/wp-content/uploads/2017/12/Machine-Learning-Übersicht-der-Möglichkeiten-und-Teilbereiche.png

Common ML Tasks

Supervised Learning

Unsupervised Learning



Supervised learning

Supervised learning

Models learn to predict outcomes based on **labeled training data**.

For a given a data set in which

$x_i \in \mathbb{R}^n$ is a vector (that contains descriptions), and

$y_i \in \mathbb{R}$ is an observable result for x_i, where

$x_i, y_i \propto p(x, y)$ independent identically distributed

Goal: find function that approximates y by a given x

$$f(x_i) \approx y_i$$

Important: $f(x_{new}) \approx y_{new}$

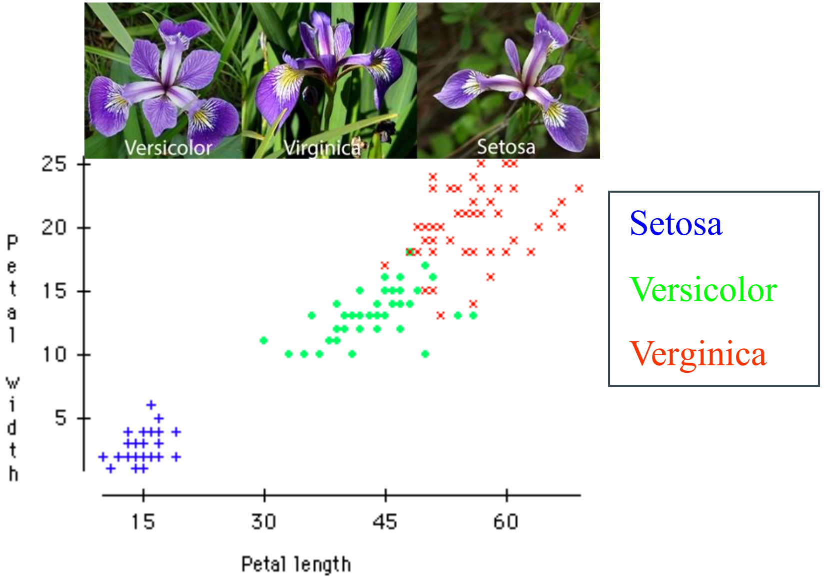
Classification

Objective: Assign input to a specified class.

To do this, parameters are determined using a training data set that has already been classified (assigned to classes), which can then be used to determine which category the input belongs to.

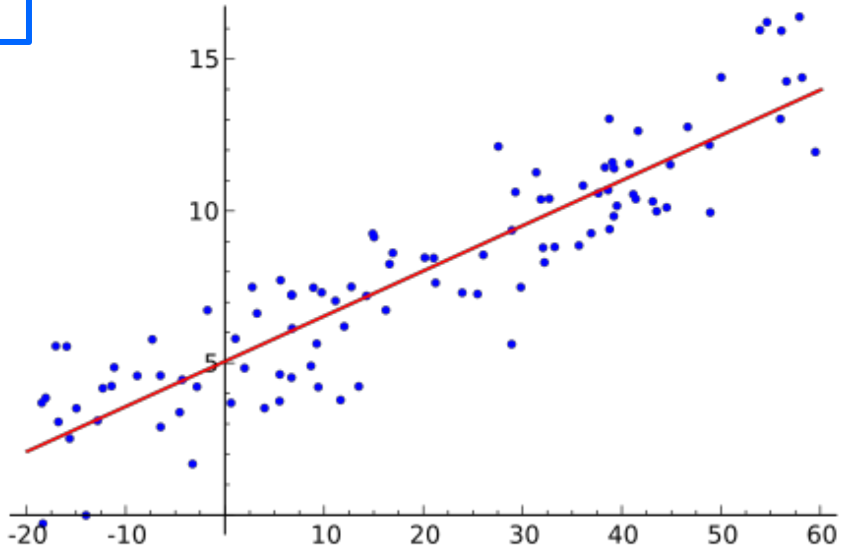
Example: Spam email detection, digit recognition, image recognition

Types of classification algorithms:
- K-Nearest Neighbors
- Decision Tree Classification
- Random Forest Classification
- ...



http://dataaspirant.com/2017/01/25/svm-classifier-implemenation-python-scikit-learn/

Regression



Goal: Make predictions for a variable.

To do this, dependencies between variables are determined, the dependent variable (input to be predicted) and independent variables (predictors)

Example: Weather forecasting, predicting market trends

Types of regression algorithms:
Simple linear regression
Multiple linear regression
Decision tree regression
Random forest regression
...

Supervised learning - Example

Imagine you have a dataset containing information about houses such as their size (in square feet), number of bedrooms, number of bathrooms, and distance from the city center.

Each house in the dataset has a corresponding price (the observable result) associated with it.

- xi would be a vector containing the features of a house, such as size, number of bedrooms, number of bathrooms, and distance from the city center.
- yi would be the price of the house.
- xi, yi represents a pair of features and its corresponding price.

The goal in this scenario is to find a function $f\ xi$ that can approximate the price yi given the features xi .

Essentially, we want to find a function that accurately predicts the price of a house based on its features.

Regression

Simple example: ML model predicts the selling price of a house

Algorithm
Mathematical
function

Algorithm

Result = weight1*var1 + weight2*var2 + weight0

Model
Algorithm
with
weighted
variables

Model

Price in $US = 100*(square_feet) + 10000*(bedrooms) + 100,000

How do you find the right weights for the model?
Training

Supervised training

1st pass

Record 1
sqft: 1500
BR: 3
…
Label:
314,000

Untrained model

Record 1
sqft: 1500
BR: 3
…
Label:
280,000

1. Assess distance from correct label
2. Adjust model

2nd pass

Record 2
sqft: 1750
BR: 4
…
Label:
380,000

Model—new weights

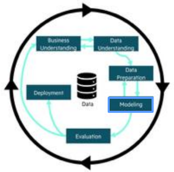
Record 2
sqft: 1750
BR: 4
…
Label:
346,500

nth pass

Record 9999
sqft: 1800
BR: 2
…
Label:
400,000

Trained model—final weights

Record 9999
sqft: 1800
BR: 2
…
Label:
400,000



Supervised training

Simple example: ML model predicts the selling price of a house

Algorithm
Mathematical
function

Algorithm

Result = weight1*var1 + weight2*var2 + weight0

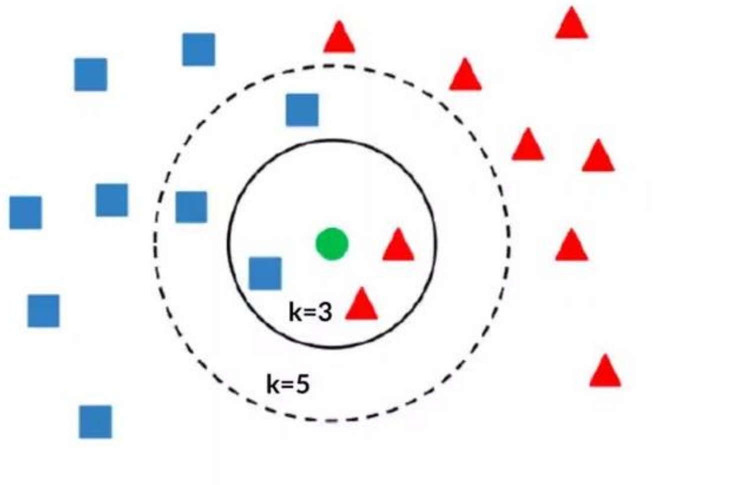
Model
Algorithm
with
weighted
variables

Model

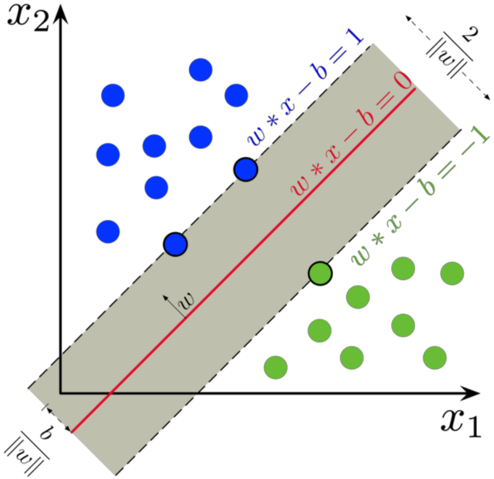
Price in $US = 100*(square_feet) + 10000*(bedrooms) + 100,000

How do you find the right weights for the model?
Training

Supervised Learning – More example Algorithms



K-nearest neighbor (KNN)



Support Vector machine (SVM)

Unsupervised learning

Unsupervised learning

Models identify patterns and relationships in unlabeled data.

Model examples:

- **K-Means Clustering**: Partitions data into distinct groups based on feature similarity.
- **Principal Component Analysis** (PCA): Reduces dimensionality while preserving data
- variance.
- **Autoencoders**: Neural networks designed for unsupervised learning tasks.

Unsupervised learning

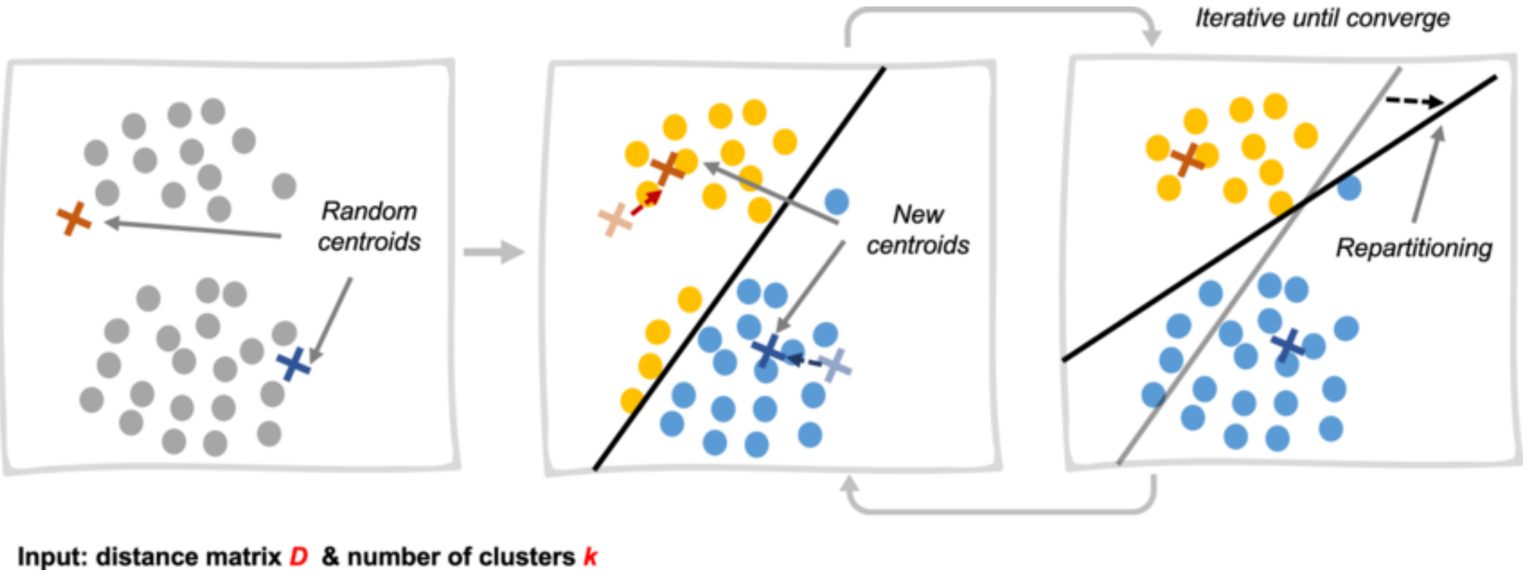
For a given a data set in which

$xi \in \mathbb{R}n$ is a vector (that contains data descriptions), and

$xi \propto p\ x$ independent identically distributed

Goal: learn about p

K-means clustering



https://www.researchgate.net/figure/Illustration-of-K-means-Clustering-Note-The-estimation-routine-of-K-means-involves-i_fig2_371115662

Unsupervised learning - Example

Imagine analyzing customer transaction data from an online retail store. We explore features like items purchased, purchase amount, and time of purchase. Our aim is to uncover hidden patterns in customer behavior without predefined labels.

Goal:

- Understand customer behavior distribution $p\ x$ without labels.
- Discover item associations and customer segments.

Approaches:

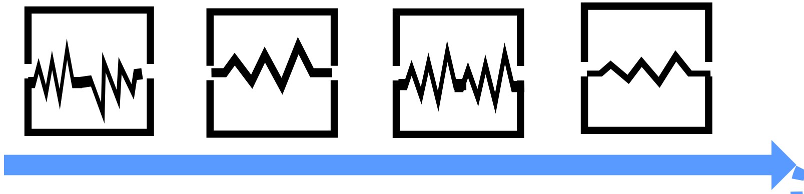
- **Market Basket Analysis:** Identify frequently co-occurring items.
- **Customer Segmentation:** Group customers with similar behaviors.

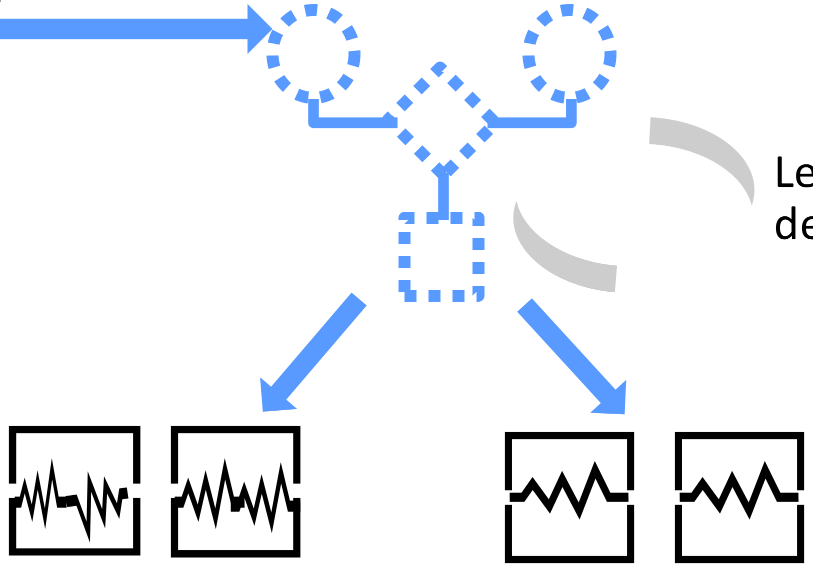
Benefits:

- Tailor marketing strategies and product recommendations.
- Optimize engagement without manual labeling.

Unsupervised learning

No labels





Learns to cluster or detect anomalies

What are the main differences between supervised and unsupervised training?

INTRODUCTION TO DATA SCIENCE BY W&S
YOUR TURN!
EXERCISE / QUESTION

Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment.

Key Components:

- **Agent:** Makes decisions and takes actions.
- **Environment:** The external system with which the agent interacts.
- **Rewards:** Feedback signal indicating the quality of actions.
- **Policy:** Strategy or algorithm used by the agent to make decisions.

Objective:

Maximize cumulative reward over time by learning optimal policies.

Reinforcement Learning - example

Imagine training a self-driving car using reinforcement learning. The scenario involves the self-driving car (agent) navigating a road network and responding to traffic conditions.

Objective:

Teach the car to navigate roads safely and efficiently.

Benefits:

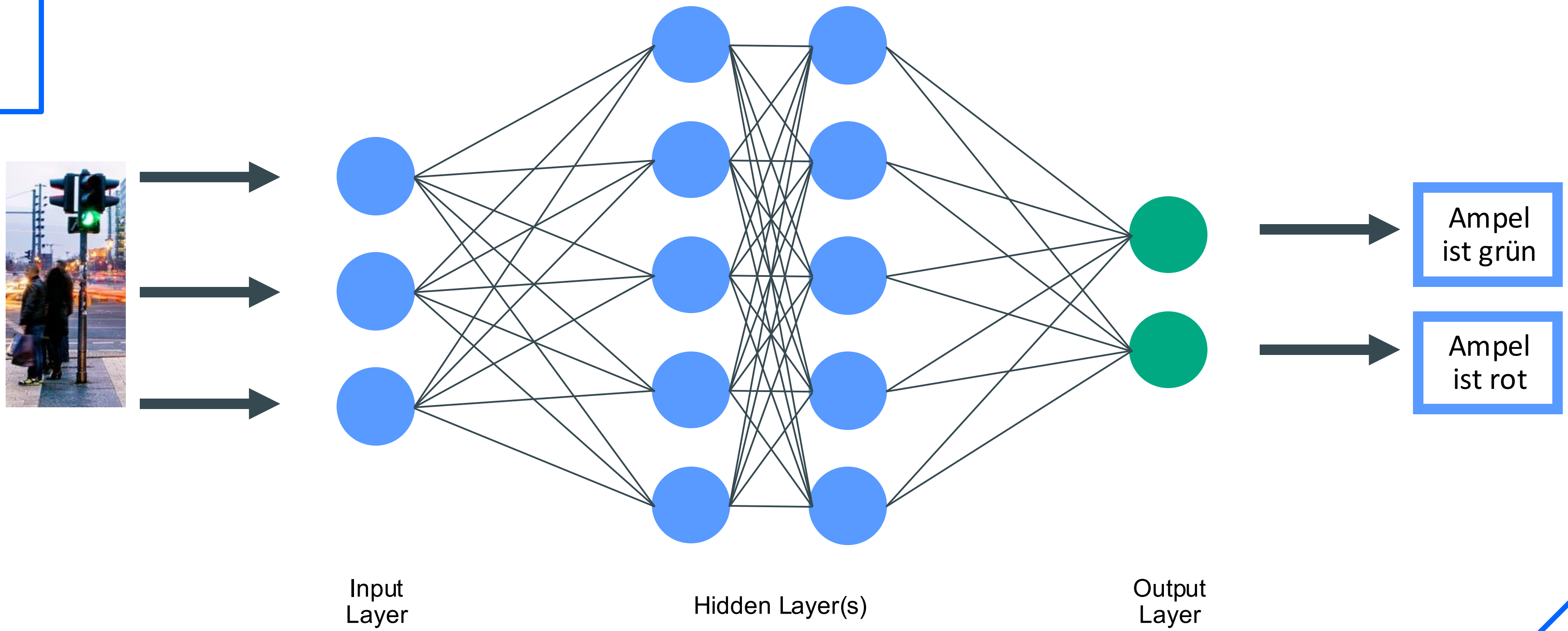
Enables adaptive and intelligent driving behavior.
Improves safety and efficiency on the roads.
Reduces human intervention and enhances autonomy.

Further Classes of ML

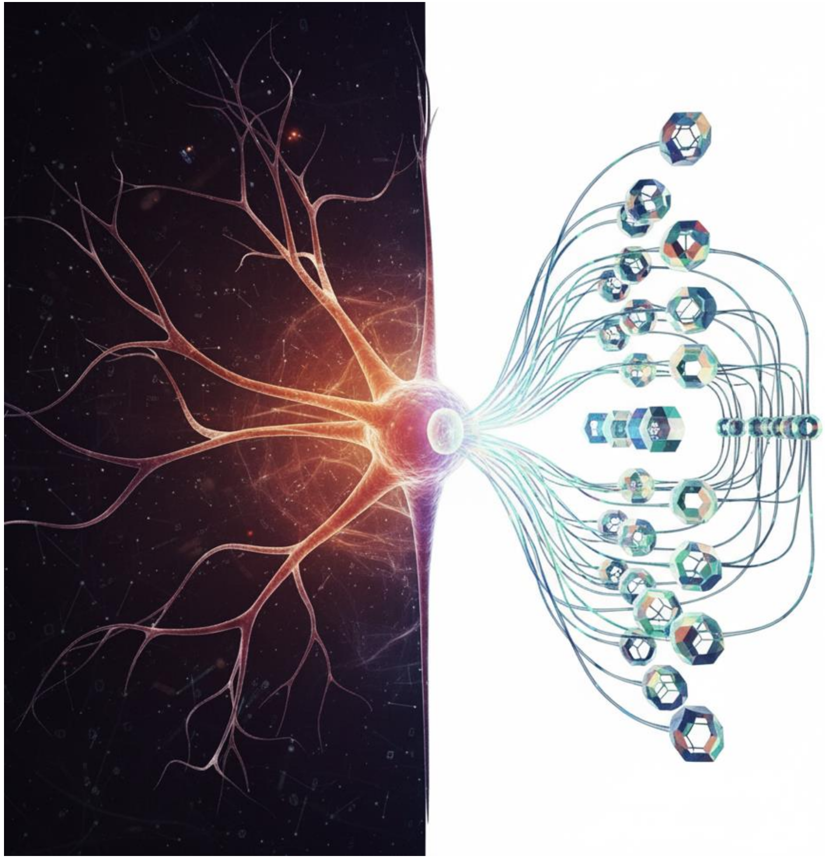
- **Semi-Supervised Learning**: This is an approach to machine learning that involves a small amount of labeled data and a large amount of unlabeled data during training. Semi-supervised learning falls between supervised learning (with completely labeled data) and unsupervised learning (with no labeled data).
- **Self-Supervised Learning**: A type of unsupervised learning where the data provides the supervision. Here, the model is trained to predict part of the input from other parts of the input.
- **Transfer Learning**: This technique involves taking a pre-trained model (usually on a large dataset) and fine-tuning it for a specific task. It is very useful when you have a limited amount of data for your task.
- **Ensemble Learning**: This approach combines the predictions from multiple machine learning algorithms to make more accurate predictions than any individual model. Common ensemble methods include bagging, boosting, and stacking.
- **Multi-Instance Learning**: In this setting, labels are associated with groups of instances (bags), rather than individual instances. The task is to predict the labels of unseen bags based on the learned patterns from the labeled bags.
- **Multi-Label Learning**: Unlike traditional classification tasks where each instance is assigned to only one label from a set of disjoint labels, multi-label learning allows for the assignment of multiple labels to each instance.
- **Multi-Task Learning**: This is an approach to inductive transfer that improves learning for one task by using the information contained in the training signals of other related tasks.
- **Meta-Learning**: Sometimes called "learning to learn", it aims to design models that can learn new skills or adapt to new environments rapidly with a few training examples.
- **Active Learning**: A special case of machine learning where the learning algorithm can interactively query the user (or some other information source) to label new data points with the desired outputs.
- **Few-Shot Learning**: The goal here is to learn information about object categories from one or just a few training samples/images.

Neural networks

Neural networks



Artificial neurons



Model training

Input x → **Modell** → **Output** y

Train/Test Ratio: What percentage of the available data is training data and what percentage is test data?

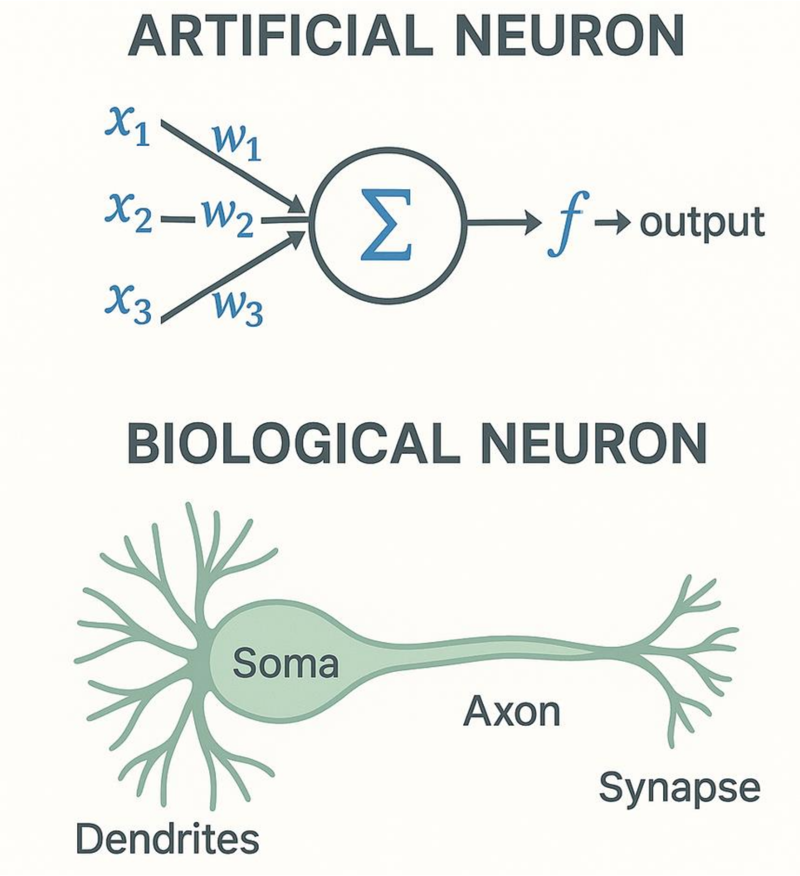
Batch Size: How large is a data unit according to which the model is adjusted?

Epoch: How often do I train the model (on the entire training data set)?

Learning Rate: How strongly is weighting adjusted after an error is detected?

Loss/Cost: How well does the algorithm perform?

Artificial neurons



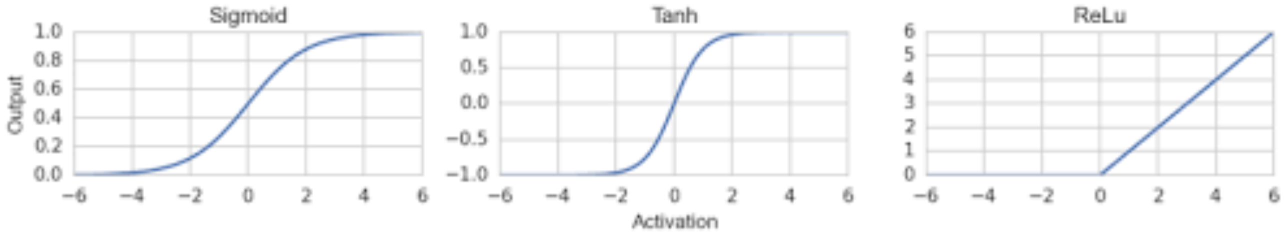
https://imgs.search.brave.com/q-abGU2MScwG9ufZX9uTeZlBD2P6ZleUKDmplF8mAoSY/rs:fit:860:0:0:0/g:ce/aHR0cHM6Ly9rMy1w/cm9kdWN0aW9uLWJ1/Y2tldC5zMy51cy1l/YXN0LTEuYW1hem9u/YXdzLmNvbS91cGxv/YWRzLzVg.dmVFM2JD/a3lxdjROY.mpmX25l/dXJvbnMucG5n

Activation function

Decides whether and how a neuron is "activated."
There are various activation functions; the choice depends on the problem and the resulting requirements for the model.

Well-known activation functions:

- Sigmoid (output value 0 or 1) binary classification
- Tanh (output values between -1 and 1) if negative outputs are possible
- ReLU Rectified Linear Unit (0 if negative, if positive exactly the value it received), often used in classification

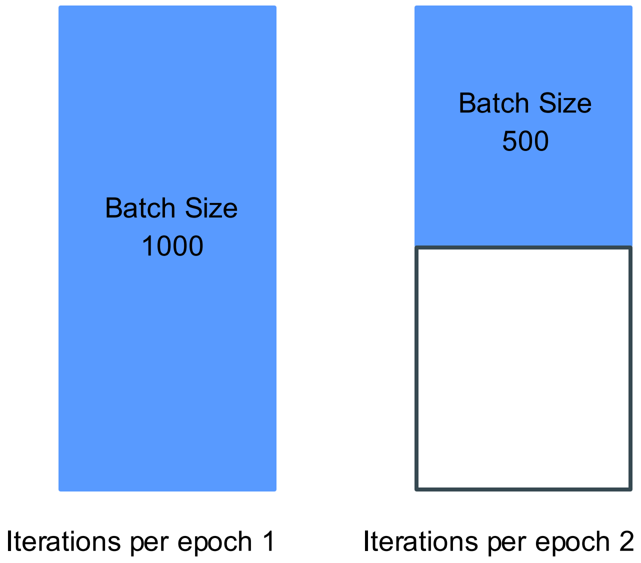


Hyperparameter

- Affects the performance of the model
- Tuning requires a lot of effort, but there are tools that automate this process

Hyperparameters:

- Train/test split ratio
- Epoch
- Batch size
- Learning rate
- Number of hidden layers
- Activation function
- Loss function / cost function
- Optimization algorithm



Bis Donnerstag

samuel.schlenker@lehre.dhbw-stuttgart.de