

MANAGING AI-BASED SYSTEMS



Session 2: Conceptual technology foundations

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.ditlab.org

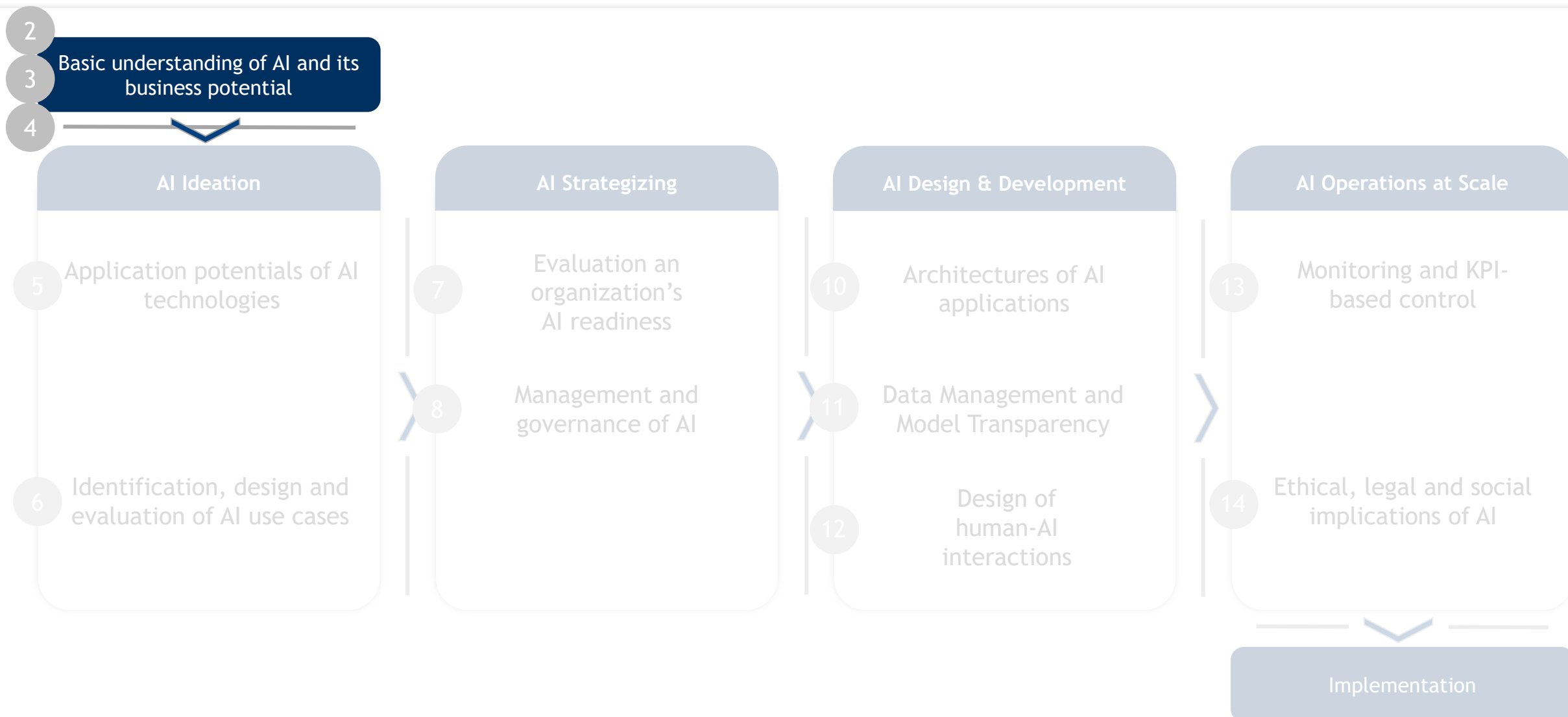
www.fim-rc.de

www.wirtschaftsinformatik.fraunhofer.de

Creative Commons Copyright

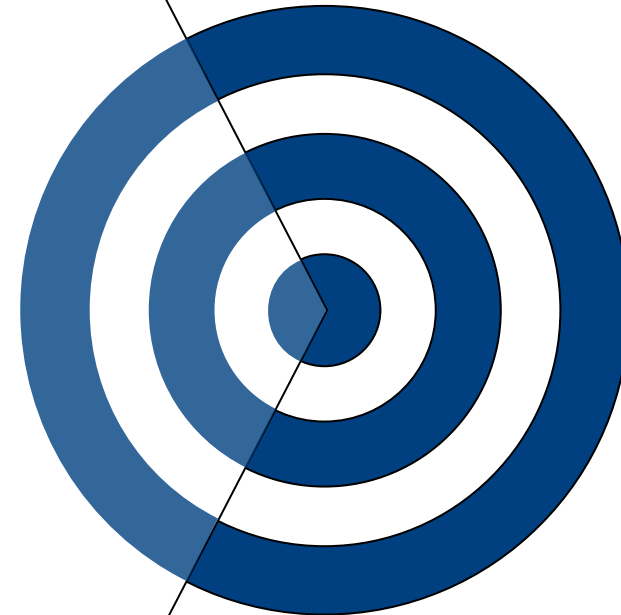
This work is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit:
<https://creativecommons.org/licenses/by-nc-sa/4.0/>

The AI implementation phases - Course navigator



Objectives of today's lecture

1. Obtain an overview of the technical equipment necessary for AI
2. Explore key AI learning methods
3. Understand commonly used mathematical algorithms for AI applications



Agenda

01 | Technical fundamentals

02 | AI learning methods

03 | Mathematical algorithms

01 | Technical fundamentals

02 | AI learning methods

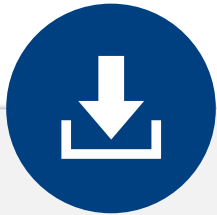
03 | Mathematical algorithms

Technical framework for the use of AI



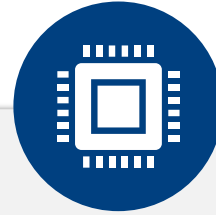
Data

- Quality
- Quantity
- Balance



Storage

- Database
- Data warehouse
- Data lake
- Cloud storage



Computing power & algorithms

- Exponential increase in computing power
- Supercomputers
- Performant ML-algorithms



Tools & Services

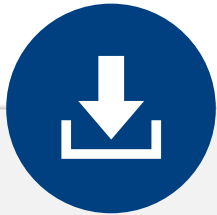
- Libraries
- Process-oriented software tools
- AI as a Service

Technical framework for the use of AI



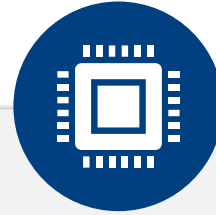
Data

- Quality
- Quantity
- Balance



Storage

- Database
- Data warehouse
- Data lake
- Cloud storage



Computing power & algorithms

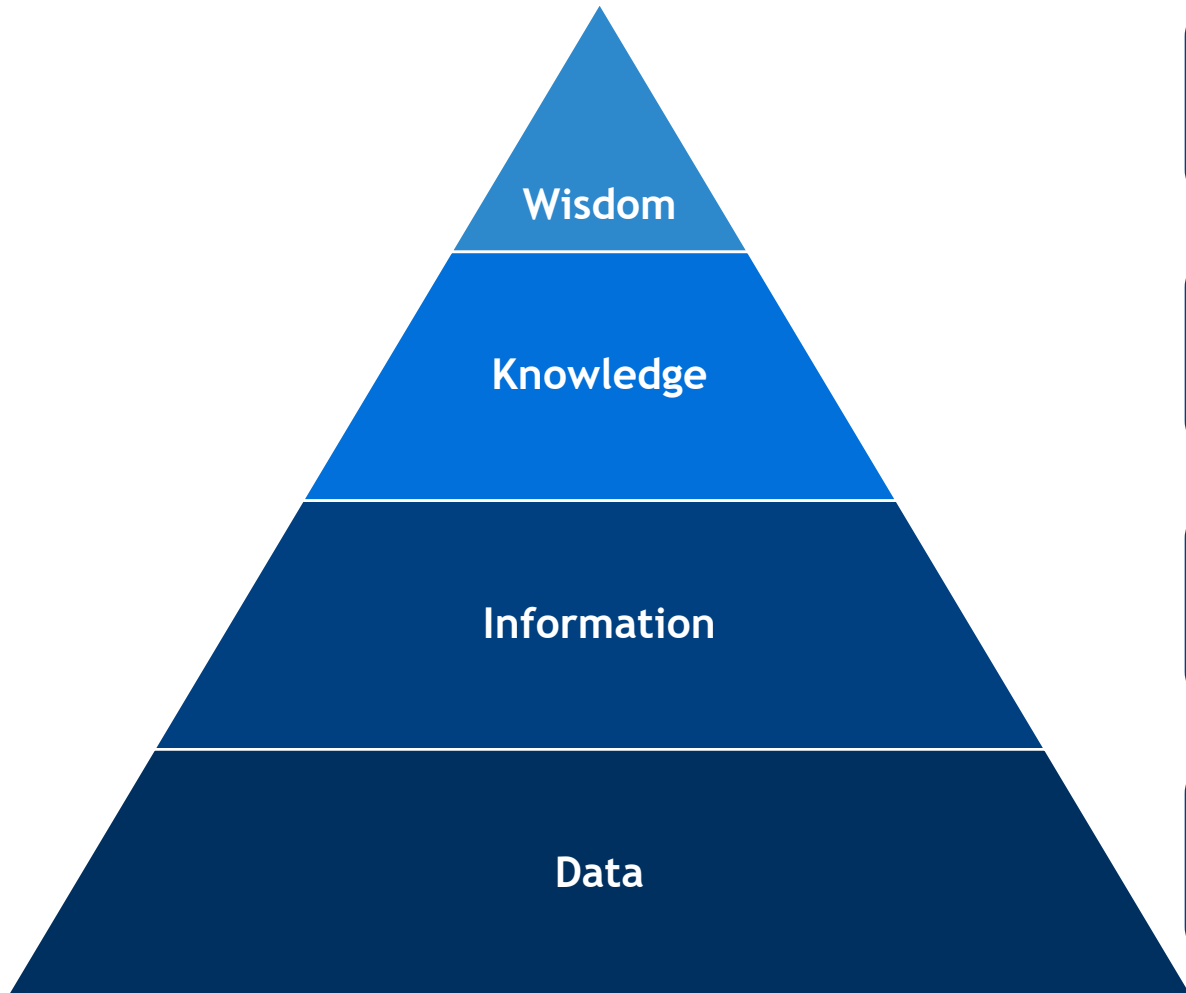
- Exponential increase in computing power
- Supercomputers
- Performant ML-algorithms



Tools & Services

- Libraries
- Process-oriented software tools
- AI as a Service

The value of data lies not in the data itself, but in its analysis, interpretation, and use



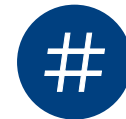
Wisdom represents the ability to apply knowledge in a broader context for sound judgment. It is the ability to make strategic, long-term decisions that go beyond immediate problem-solving.



Knowledge represents a higher level of comprehension with information. It involves applying and synthesizing information for informed decision-making and problem-solving.



Information is Data that has been processed, structured, or organized in a meaningful way. It provides context, relevance, and understanding.



Data is unprocessed facts, figures, or values. It lacks context and meaning on its own and requires organization or interpretation to be useful.

To be suitable for AI algorithms, data should meet certain criteria



Quality

Incomplete data, incorrect entries, or noisy features often make it difficult to achieve satisfying results



Quantity

Enough data samples is a prerequisite for a successful ML project



Balance

Imbalanced training data limits the performance and applicability of ML models








The data which are processed by AI applications must be managed

Data quality is one of the most important challenges in **data management**, since dirty data often leads to inaccurate data analytics results and incorrect business decisions

The best learner is useless if not supplied with **balanced** and **high-quality** data in sufficient **quantity**



Data quality

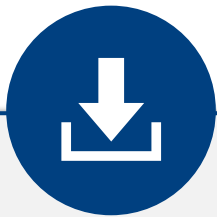
-  **Accuracy:** Does my data reflect reality?
-  **Completeness:** Is my data complete and unbiased?
-  **Purity:** Is my data free of errors?
-  **Recency:** Is my data up to date or out of date?
-  **Consistency:** Is my data consistent across all platforms and databases?

Technical framework for the use of AI



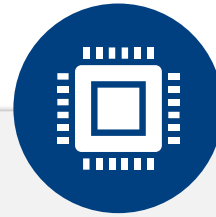
Data

- Quality
- Quantity
- Balance



Storage

- Database
- Data warehouse
- Data lake
- Cloud storage



Computing power & algorithms

- Exponential increase in computing power
- Supercomputers
- Performant ML-algorithms



Tools & Services

- Libraries
- Process-oriented software tools
- AI as a Service

Data storage approaches



Data lake

- Raw data
- Large amount of data
- Used for data analysis, machine learning
- Accessible via specialized tools



Database

- Structured data
- Typically small amount of data (can vary)
- Used for application services
- Accessible via DBMS



Data warehouse

- Structured and processed data
- Large amount of data
- Used for business analytics
- Accessible via BA tools



Data lakehouse

- Combines benefits of data lake and warehouse
- Flexibility and scalability, while offering data mgmt. and governance



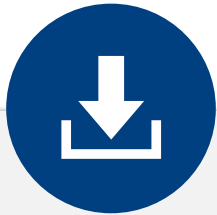
Data storage selections are based on the data form, quantity and location

Technical framework for the use of AI



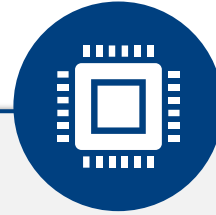
Data

- Quality
- Quantity
- Balance



Storage

- Database
- Data warehouse
- Data lake
- Cloud storage



Computing power & algorithms

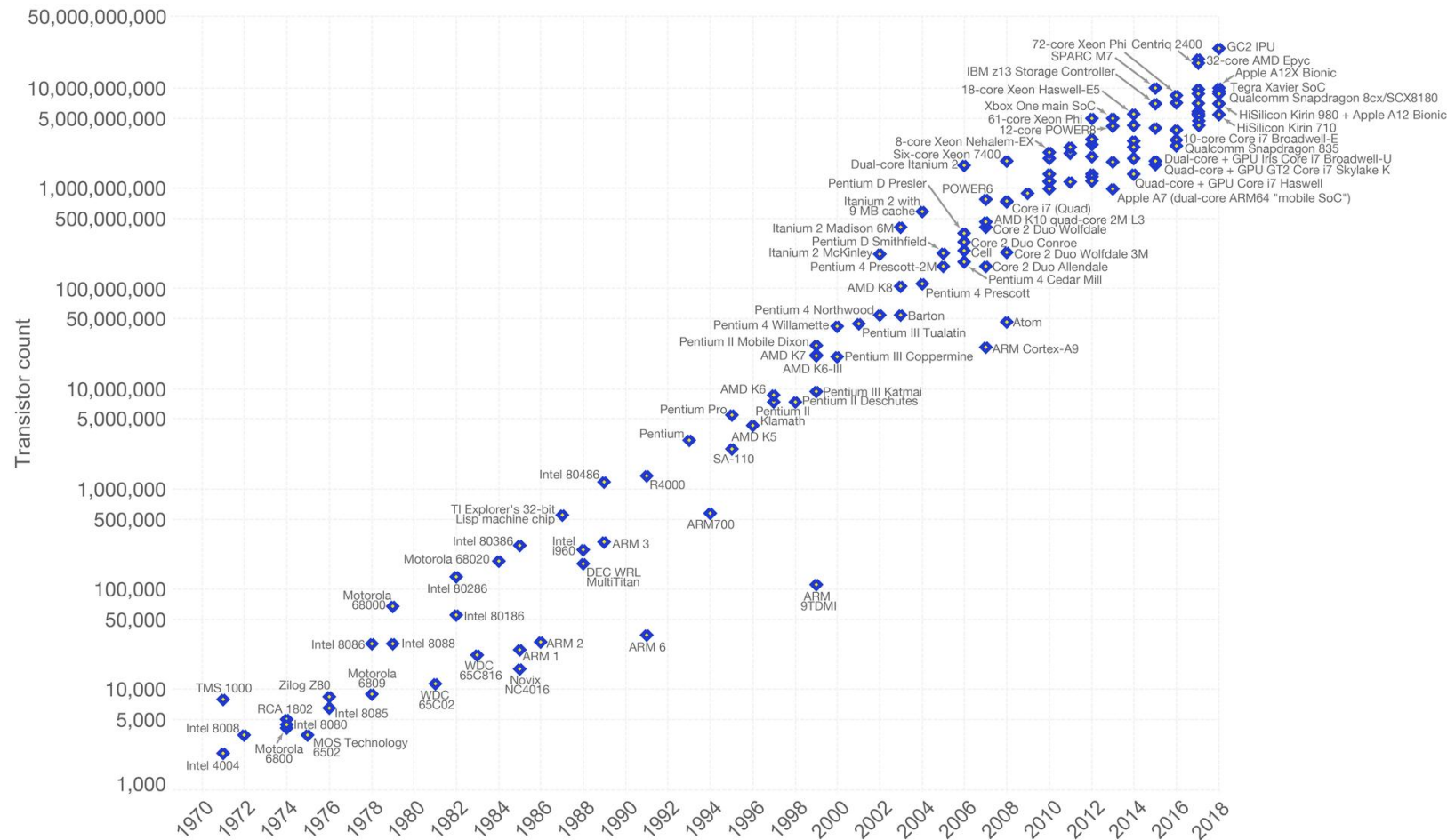
- Exponential increase in computing power
- Supercomputers
- Performant ML-algorithms



Tools & Services

- Libraries
- Process-oriented software tools
- AI as a Service

The exponential increase in computing power enables increasingly complex AI models and algorithms



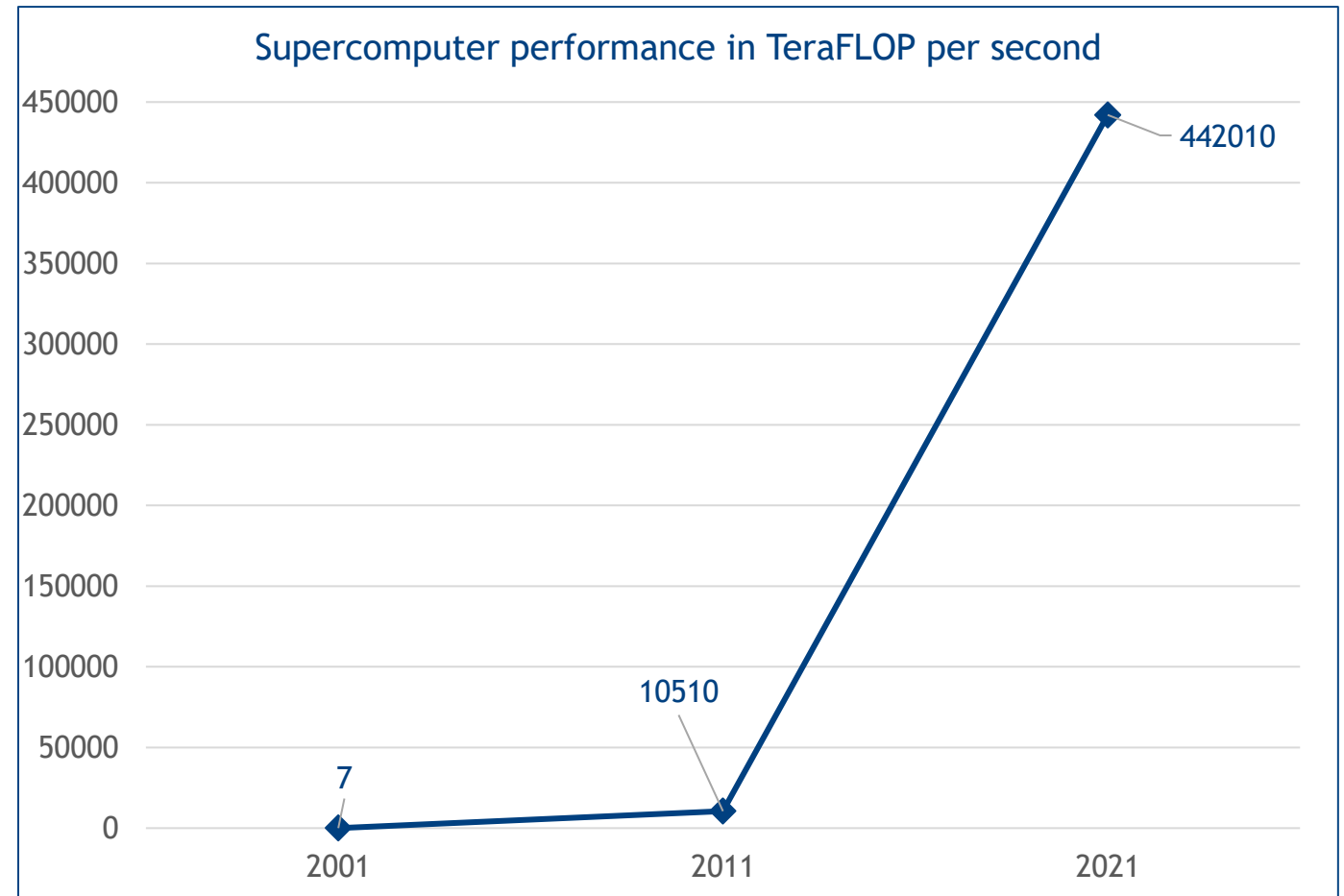
<https://www.technologyreview.com/s/601102/intel-puts-the-brakes-on-moores-law/>

Moore's Law refers to the trend that the complexity of integrated circuits doubles regularly (approx. 18-24 months) with minimal component costs. However, it is foreseeable that this trend will not continue due to physical limits.

MIT Technology Review (2016)

Supercomputers provide computational power to process large datasets and run complex machine learning algorithm

Supercomputers are computers with a large number of processors that can access common peripherals and a partially shared main memory. Supercomputers are often used for computer simulations in the field of high-performance computing (e.g., astrophysics, quantum mechanics, weather forecasting, climatology, cryptanalysis, ...).



1 TeraFLOP per second = 1 trillion computing operations per second

Statista.com (2022)

The efficiency and effectiveness of AI algorithms is constantly increasing

- Increasing **data quality and quantity** (e.g., big data, data augmentation)
- Increasing **computing power** (e.g., hardware advances, parallel computing, quantum computing)
- **Algorithmic innovations** (e.g., generative adversarial networks, attention mechanisms)
- **New learning techniques** (e.g., deep learning, transfer learning, reinforcement learning)



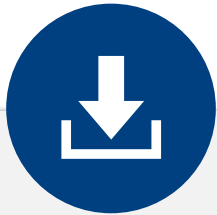
- Improved **efficiency** and **effectiveness** of AI algorithms (e.g., machine learning algorithms)

Technical framework for the use of AI



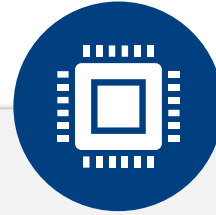
Data

- Quality
- Quantity
- Balance



Storage

- Database
- Data warehouse
- Data lake
- Cloud storage



Computing power & algorithms

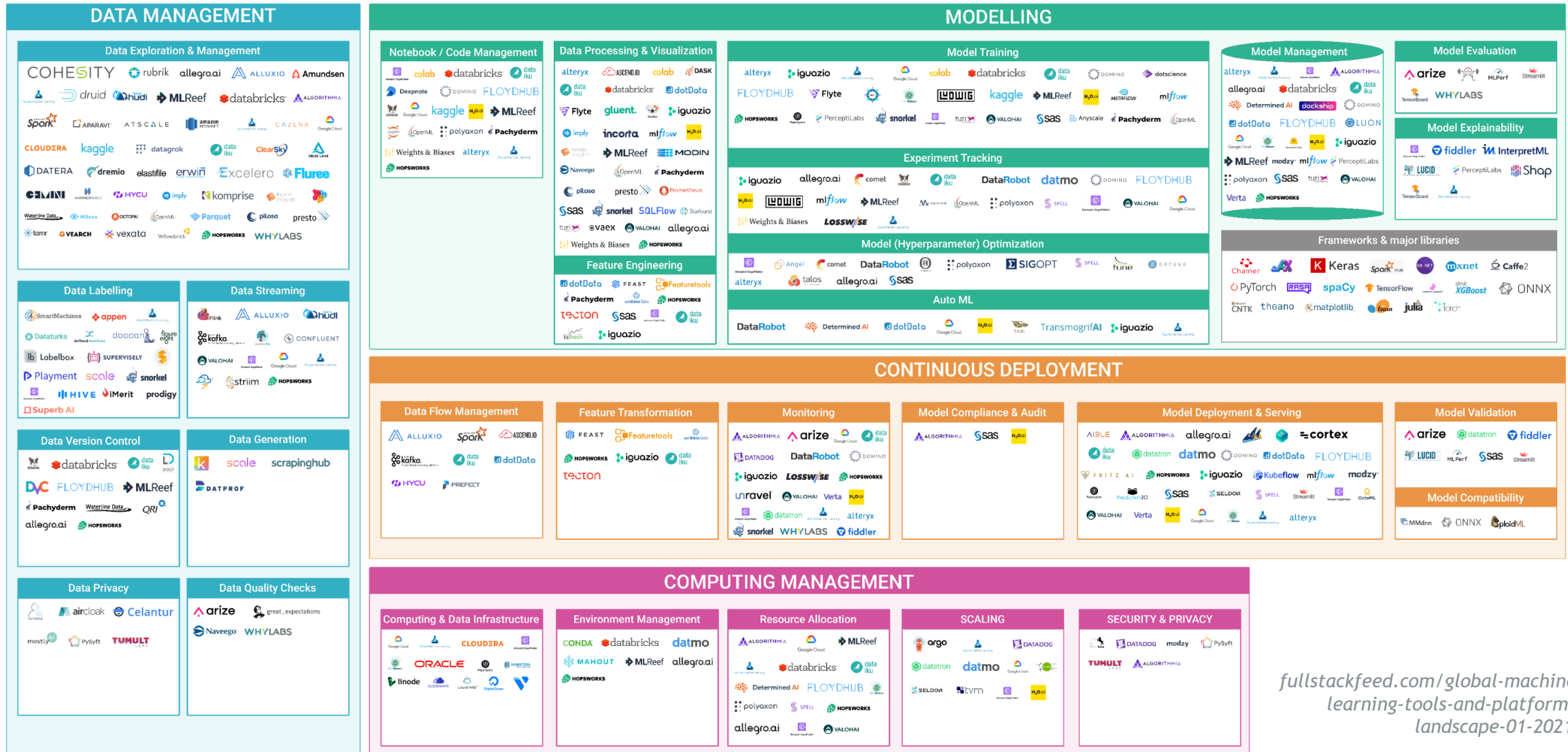
- Exponential increase in computing power
- Supercomputers
- Performant ML-algorithms



Tools & Services

- Libraries
- Process-oriented software tools
- AI as a Service

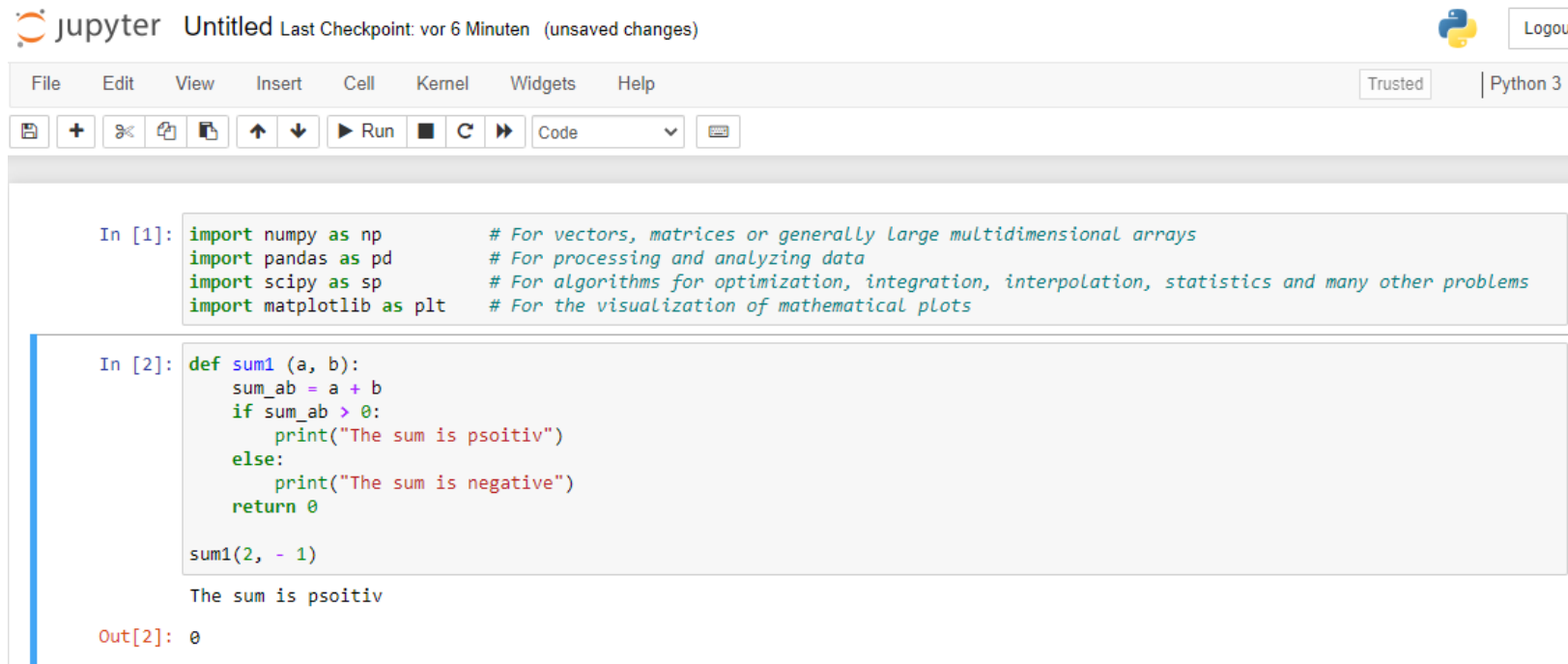
Machine learning landscape



fullstackfeed.com/global-machine-learning-tools-and-platforms-landscape-01-2021/

Python - Jupyter Notebook

- AI applications need to be programmed in suitable environments
- **Python** is the most popular language for programming AI applications
- Compared to Java or C#, many things can be **formulated** more **easily** in Python
- Availability and easy import of available **mathematical** and **statistical libraries** in Python



```
jupyter Untitled Last Checkpoint: vor 6 Minuten (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [1]: import numpy as np           # For vectors, matrices or generally large multidimensional arrays
import pandas as pd               # For processing and analyzing data
import scipy as sp                # For algorithms for optimization, integration, interpolation, statistics and many other problems
import matplotlib as plt          # For the visualization of mathematical plots

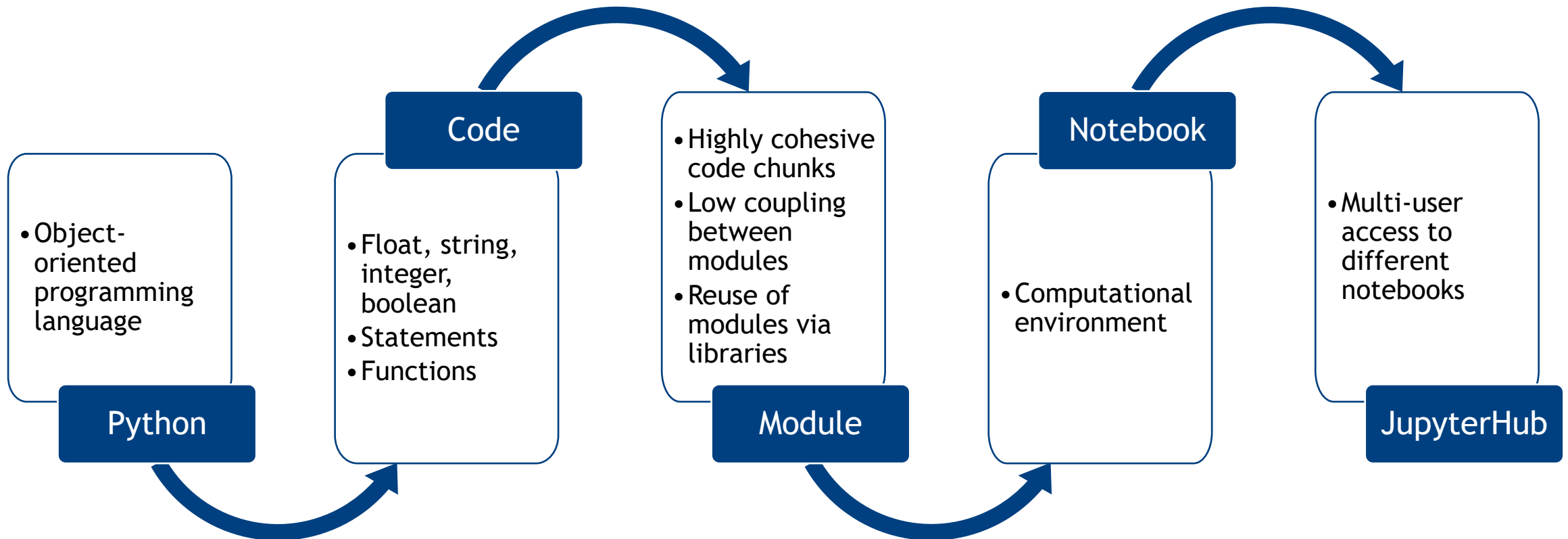
In [2]: def sum1 (a, b):
        sum_ab = a + b
        if sum_ab > 0:
            print("The sum is psoitiv")
        else:
            print("The sum is negative")
        return 0

        sum1(2, - 1)

        The sum is psoitiv

Out[2]: 0
```

Object-oriented programming with Python



Extracts from Python code - AI algorithms

```
In [ ]: # Load dataset
url = "https://example.csv"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = read_csv(url, names=names)

...
# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
pyplot.show()

...
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
y = array[:,4]
X_train, X_validation, Y_train, Y_validation = train_test_split(X, y, test_size=0.20, random_state=1)

...
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = StratifiedKFold(n_splits=10, random_state=1, shuffle=True)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    print('%s: %f (%f)' % (name, cv_results.mean(), cv_results.std()))
```



Online tutorials to program AI (machine learning) algorithms:

<https://github.com/dannybusch/neuromant.de-Tutorials> (German)

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/> (English)

Agenda

01 | Technical fundamentals

02 | AI learning methods

03 | Mathematical algorithms

Types of machine learning methods

There are various types of **machine learning methods**. Each learning method can be implemented using different mathematical algorithms.



Supervised learning

Learning algorithms that model relationships and dependencies between labeled output and input values to predict future output values based on the relationships learned from previous data sets

Unsupervised learning

Learning algorithms that learn independently and without supervision to identify and describe previously unknown patterns and relationships in unlabeled data

Semi-supervised learning

Learning algorithms that use supervised and unsupervised learning approaches in order to process labeled and unlabeled data sets to identify patterns and predict outputs

Reinforcement learning

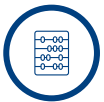
Learning algorithms that explore and exploit knowledge gathered from interactions with the environment to improve the reward or to decrease the risk of the algorithm in an iterative way

Supervised learning - deep dive



Role of humans

Human experts act as teachers which train the algorithm by providing input data and showing correct output sets



Basic mathematical idea

Training input:

$$x_1 \rightarrow y_1$$

$$x_2 \rightarrow y_1$$

$$z_1 \rightarrow y_2$$

$$z_2 \rightarrow y_2$$



Algorithm:

$$z_3 \rightarrow y_2$$

$$x_3 \rightarrow y_1$$

$$z_4 \rightarrow y_2$$

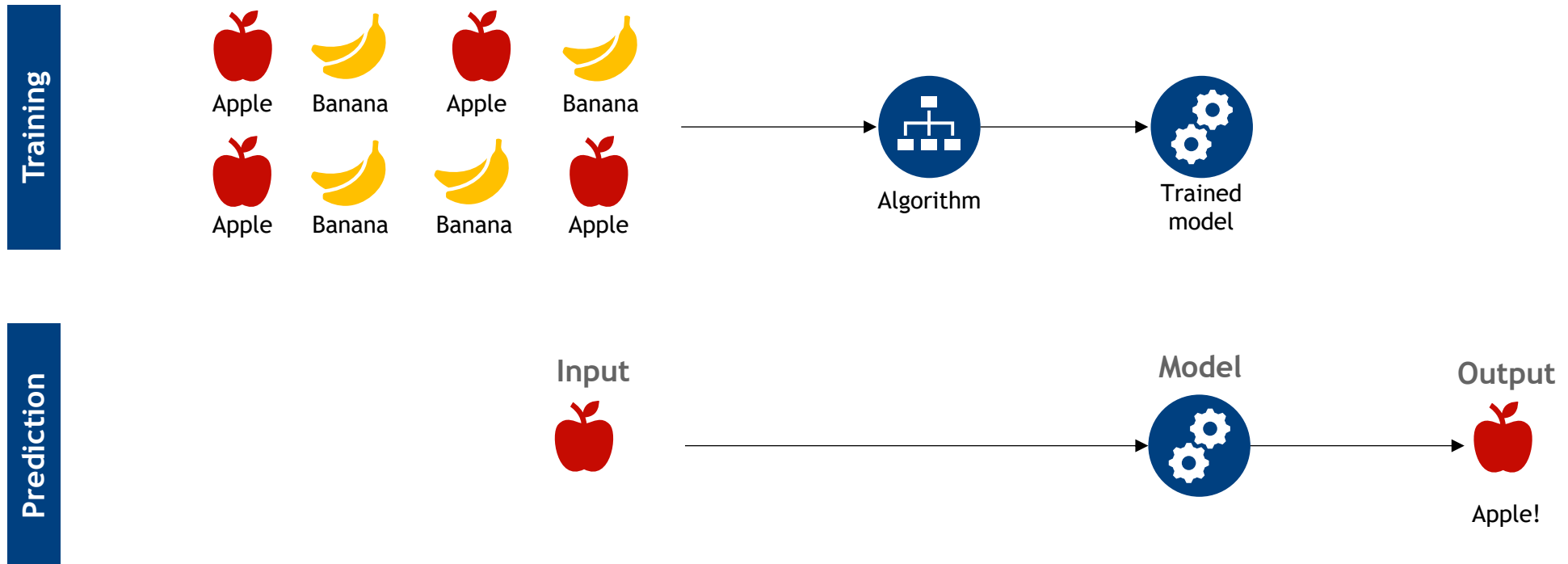
$$z_5 \rightarrow y_2$$



Learning algorithms

- K-Nearest Neighbors
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks

Learning process in supervised learning



Example from nature

- Baby does not know what apples or bananas are
- Baby is shown and explained what apples and bananas are
- Baby remembers that apples are red, and bananas are yellow and can match them up

Vaseekaran (2018)

Supervised learning example - The Iris Dataset



Iris setosa



Iris versicolor



Iris virginica

https://en.wikipedia.org/wiki/Iris_flower_data_set



The data input in machine learning approaches can contain different numerical formats

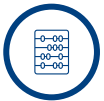
Images: Wikipedia, Source: scikit-learn.org

Unsupervised learning - deep dive



Role of humans

No human intervention (no humans label data or act as teachers)



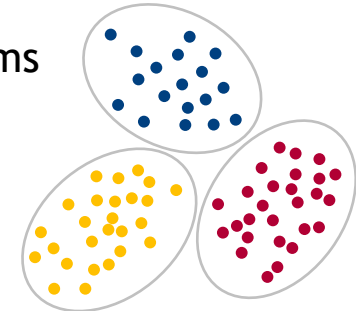
Basic mathematical idea

Descriptive models mining for rules, detecting patterns, as well as summarizing and grouping data sets

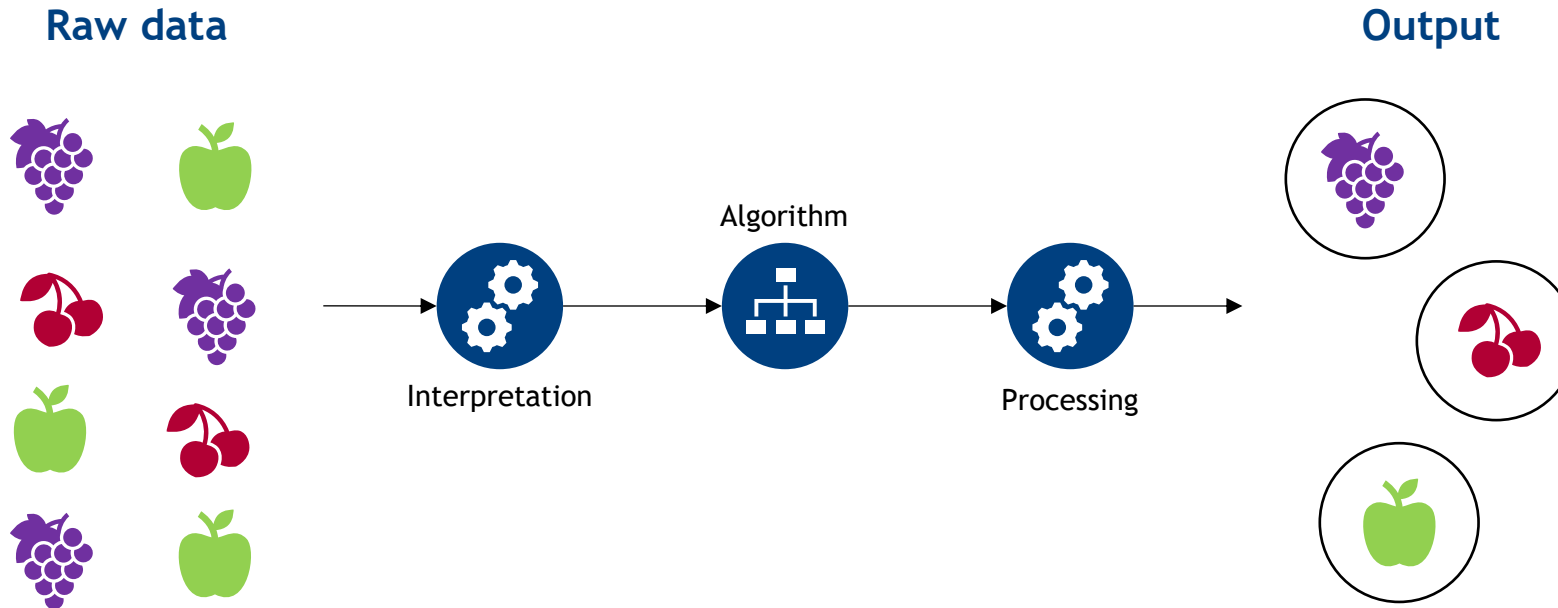


Learning algorithms

- Clustering algorithms
- K-means clustering
- Association rules



Learning process in unsupervised learning



Example from nature

- Baby does not know what apples, grapes or cherries are
- Baby independently recognizes that there are differences between the types of fruit after looking at them in detail

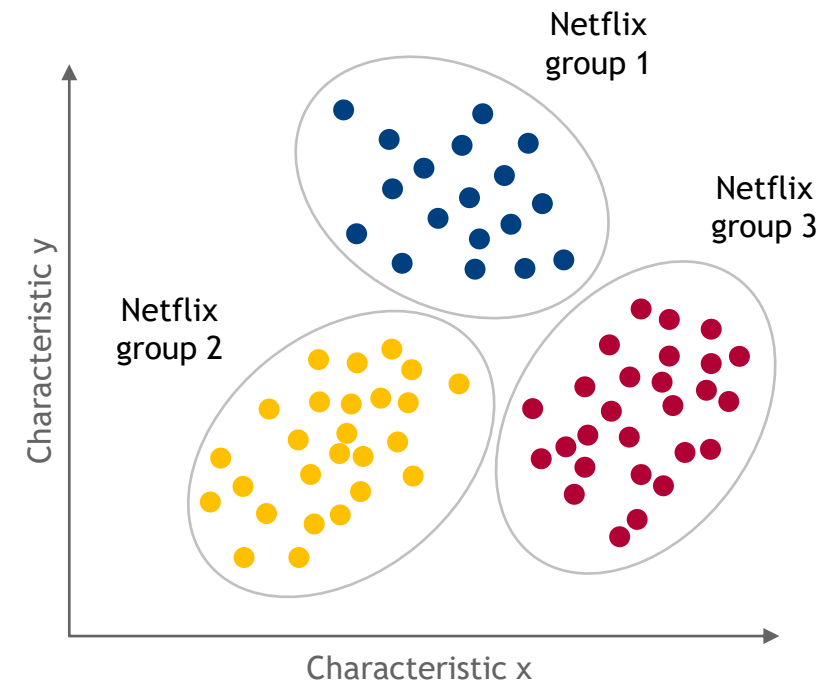
Vaseekaran (2018)

Unsupervised learning example - Netflix recommendation

Because you watched Shameless (U.S.)



- Netflix analyzes users who have also watched Shameless regarding other movies and series they have watched
- Based on this analysis, Netflix creates different user **clusters**
- These clusters are used to provide **suitable recommendations** for the following movies and series



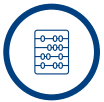
Lutins (2017)

Reinforcement learning - deep dive



Role of humans

Reinforcement learning algorithms interact with the environment. Human experts are part of the environment but are not actively training the algorithm.



Basic mathematical idea

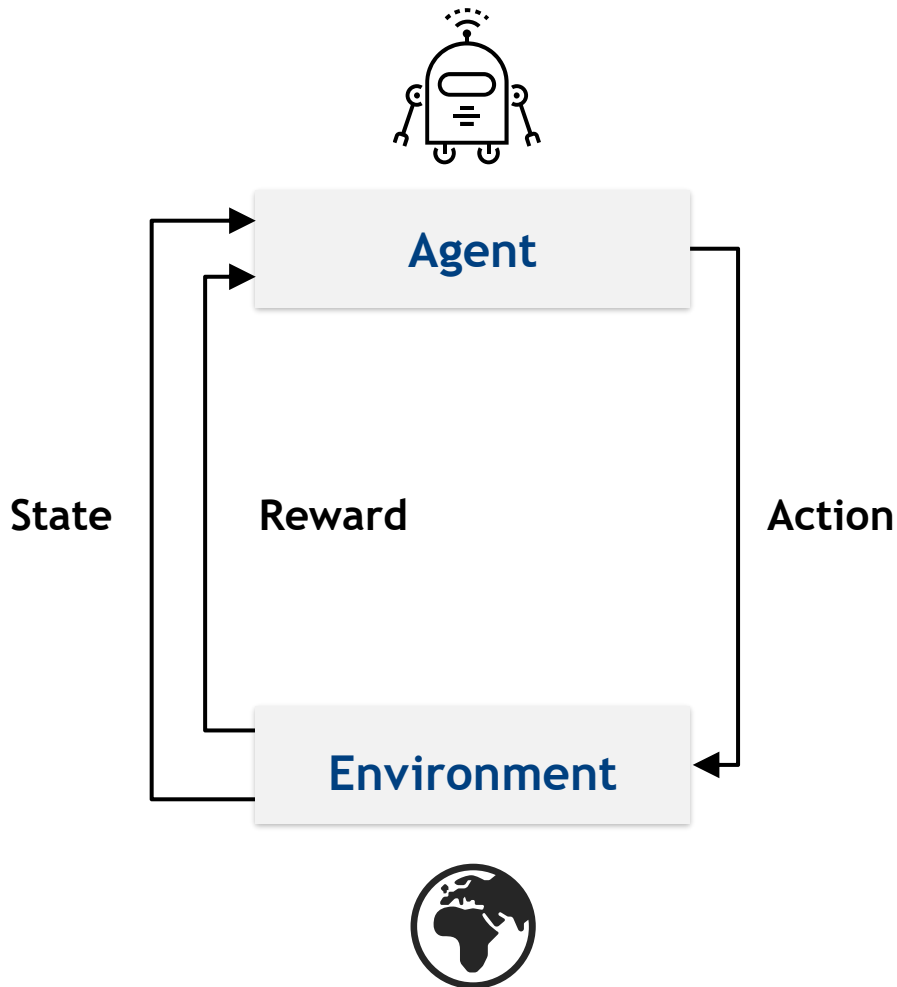
1. Input
2. Decision and action
3. Reaction or reward by the environment
4. Reaction and state of the art update



Learning algorithms

- Q-Learning
- Temporal Difference (TD)
- Deep Adversarial Networks

Learning processes in reinforcement learning



To produce intelligent programs (also called agents), reinforcement learning goes through the following steps:

1. The agent assesses the input state
2. A decision-making function is used to decide what action to take in response to the situation at hand
3. Once the action is performed, the agent gets feedback in the form of rewards based on the outcomes of its actions
4. The state-action pair information about the reward is stored

Fumo (2019)

Reinforcement learning - OpenAI plays hide and seek



https://www.youtube.com/watch?v=kopoLzvh5jY&feature=emb_logo

Limitations and risks of the different learning methods

Supervised learning

- Overfitting
- Underfitting
- Data quality
- Data preparation
- Varying consistency in classes
- Effort
- Costs

Unsupervised learning

- Overfitting
- Underfitting
- Insufficient data quantity
- Data quality
- Missing consideration of relationships in the data

Semi-supervised learning

- Data quality
- Iteration results are not stable
- Lower accuracy

Reinforcement learning

- Overload of states
- Not suitable for simple problems
- Insufficient data quantity
- Computing power
- Limited dimensionality



Management decision for a suitable learning method

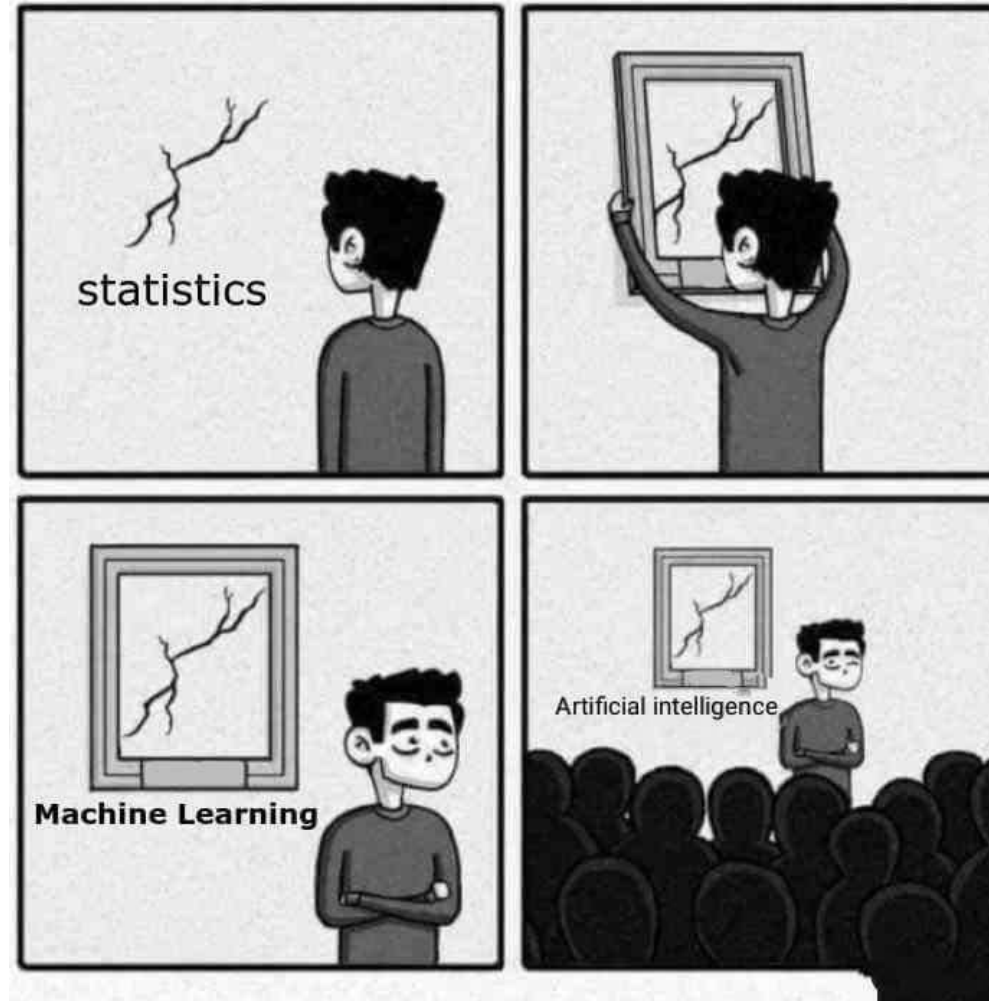
Agenda

01 | Technical fundamentals

02 | AI learning methods

03 | Mathematical algorithms

Maths, statistics, or AI?

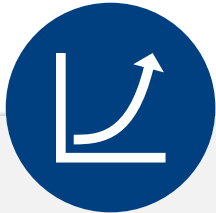


Picture: <https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3>

Algorithms as the basis of AI applications can be divided into five functional groups

Group	Example	Algorithms
Regression	Predict stock prices (by extrapolating the past)	Linear regression , Neural Network Regressor, Support Vector Regression, Random Forest, Decision Tree (CART)...
Classification	Classify customers into predefined groups in terms of purchasing power	Random Forest, Multiclass Classification, Kernelized SVMs, Neural Network Classifier, Naive Bayes, Decision Tree (ID3)...
Clustering	Identify homogeneous groups among customers in terms of purchasing power	K-Nearest-Neighbor, K-Medians, K-Means , Hierarchical Clustering, Hidden Markov Models, K-Medoids, Fuzzy C-Means, ...
Generating	Write a news article from weather data	Generative Adversarial Networks, Variational Autoencoders (VAE), ...
Acting	AI plays Super Mario	Q Learning, Asynchronous Actor-Critic Agents, State-action-reward-state-action, ...

Examples of basic mathematical machine learning algorithms



Linear regression

- Supervised learning method
- Regression algorithm

What is a linear regression?

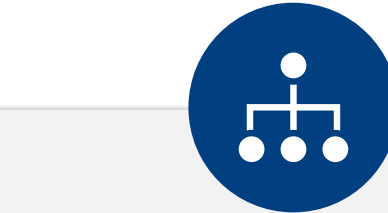
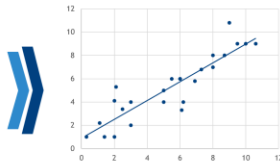
 **AI FOR BUSINESS
BUSINESS FOR AI**

x: observation y: output value b: gradient a: interception

A linear regression represents a statistical procedure that explains an observed dependent variable (y) by one or more independent variables (x) and maps the underlying relationship in a regression graph.

Adding new data points influences the calculation of the linear regression line.

Regression line equation: $y_i = a + x_i \cdot b$



Decision tree

- Supervised learning method
- Regression/classification algorithm

What is a decision tree?

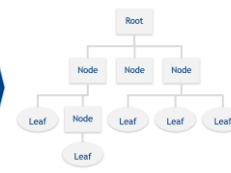
 **AI FOR BUSINESS
BUSINESS FOR AI**

Root node Decision node Leaf

A decision tree (DT) represents a decision procedure based on a sequence of „questions“. These questions are usually called **tests**. Each question is represented by a **node** in the DT.

The answer to the first question determines which question will be asked next.

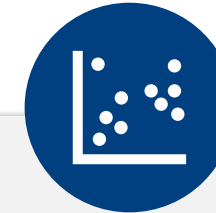
Question will be asked until a **decision** is reached. The decision is the final **prediction** of the DT. The decision is represented by a **leaf** in the DT.



Blockleit (2020)

19.07.2022 ABBA | Vortragender

26

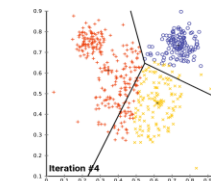


K-Means

- Unsupervised learning method
- Clustering algorithm

The k-means algorithm is an example of unsupervised learning methods.

 **AI FOR BUSINESS
BUSINESS FOR AI**



Functionality

- Choice of a value k (=number of clusters)
- Random selection of center coordinates for each cluster
- Two-step process:
 - Allocation of all data points to the nearest center point
 - Updates of the center points such that the squared deviations of the cluster centers becomes minimal
- Repeating the process until convergence is achieved


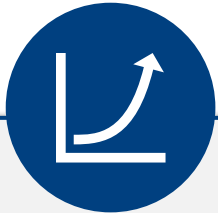
 The k-means algorithm is a method for cluster analysis that classifies data points into a certain number (=k) of clusters.

Image: Wikipedia. Source: Towardsdatascience.com

<https://blog.iao.fraunhofer.de/endlich-verstaendlich-ki-verfahren-einfach-erklaert/>; <https://handbuch-ki.net/die-intelligenz-in-der-maschine/>

Linear regression is a supervised learning method



Linear regression

- Supervised learning method
- Regression algorithm

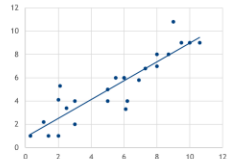
What is a linear regression?

x: observation **y:** output value **b:** gradient **a:** interception

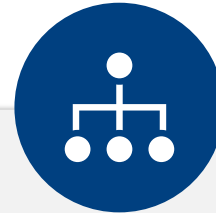
A linear regression represents a statistical procedure that explains an observed dependent variable (y) by one or more independent variables (x) and maps the **underlying relationship** in a regression graph.

Adding new data points influences the calculation of the linear regression line.

Regression line equation: $y_1 = a + x_2 + b$



19.07.2022 ABBA | Fortgeschritten* 26



Decision tree

- Supervised learning method
- Regression/classification algorithm

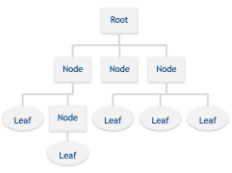
What is a decision tree?

Root node **Decision node** **Leaf**

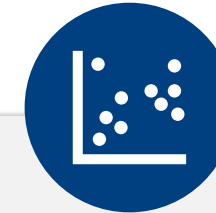
A decision tree (DT) represents a decision procedure based on a sequence of „questions“. These questions are usually called **tests**. Each question is represented by a **node** in the DT.

The answer to the first question determines which question will be asked next.

Question will be asked until a **decision** is reached. The decision is the final **prediction** of the DT. The decision is represented by a **leaf** in the DT.



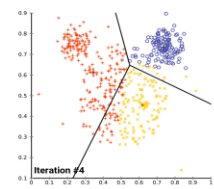
19.07.2022 ABBA | Fortgeschritten* 26



K-Means

- Unsupervised learning method
- Clustering algorithm

The k-means algorithm is an example of unsupervised learning methods.



Functionality

- Choice of a value k (=number of clusters)
- Random selection of center coordinates for each cluster
- Two-step process:
 - Allocation of all data points to the nearest center point
 - Updates of the center points such that the squared deviations of the cluster centers becomes minimal
- Repeating the process until convergence is achieved

Iteration #4

The k-means algorithm is a method for cluster analysis that classifies data points into a certain number (=k) of clusters.

Image: Wikipedia. Source: Towardsdatascience.com

<https://blog.iao.fraunhofer.de/endlich-verstaendlich-ki-verfahren-einfach-erklaert/>; <https://handbuch-ki.net/die-intelligenz-in-der-maschine/>

What is a linear regression?

x: observation

y: output value

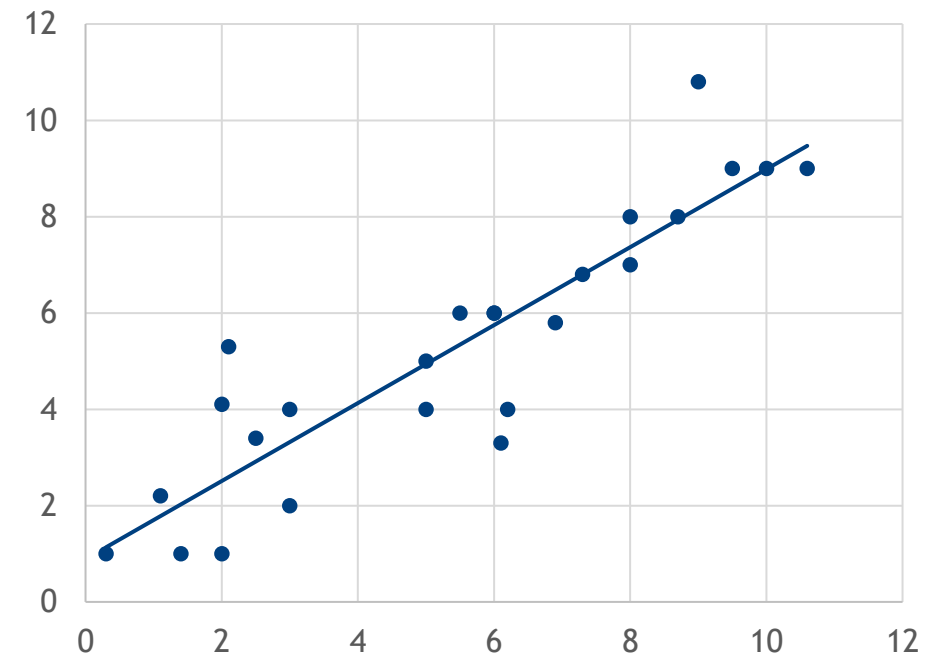
b: gradient

a: intercept

A linear regression represents a statistical procedure that explains an observed dependent variable (y) by one or more independent variables (x) and maps the **underlying relationship** in a regression graph

Adding new data points influences the calculation of the linear regression line

Regression line equation: $y_i = a + x_i * b$



Regression line equation

x: observation

y: output value

b: gradient

a: intercept

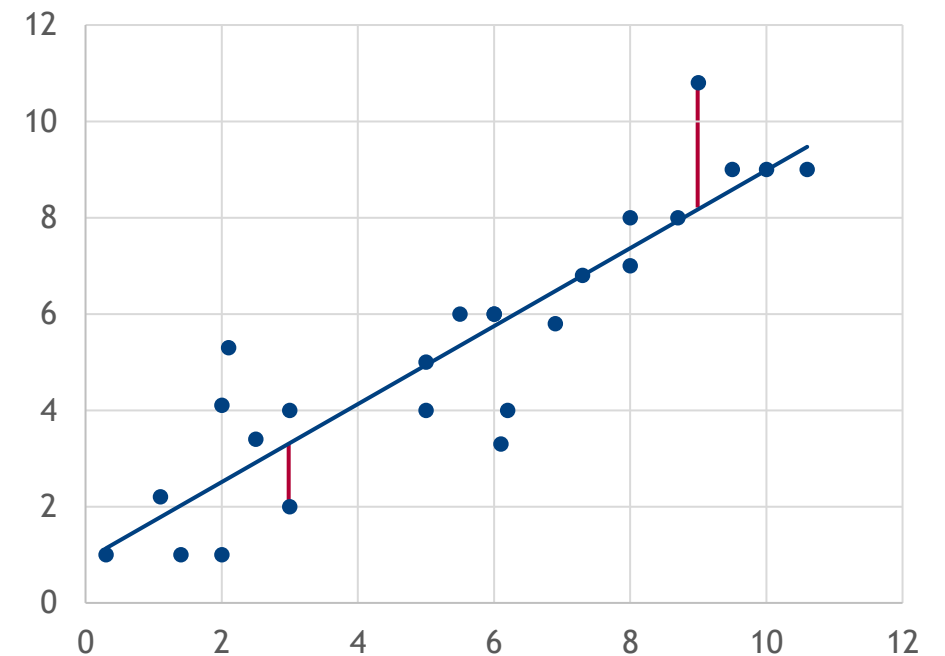
Regression line equation: $y_i = a + x_i * b$

Covariance: $s_{xy} = \frac{1}{n-1} * \sum (x_i - x_{avg})(y_i - y_{avg})$

Correlation: $r_{xy} = \frac{s_{xy}}{s_x * s_y}$

a: $a = -\frac{s_y}{s_x} * r_{xy} * x_{avg} + y_{avg}$

b: $b = r_{xy} * \frac{s_y}{s_x}$



Regression line equation for the prediction of house prices

Problem: Suppose you have a dataset with information about various properties, including the number of rooms (room count) and the actual sale price (house price). Your goal is to create a model that can predict the house price based on the room count.

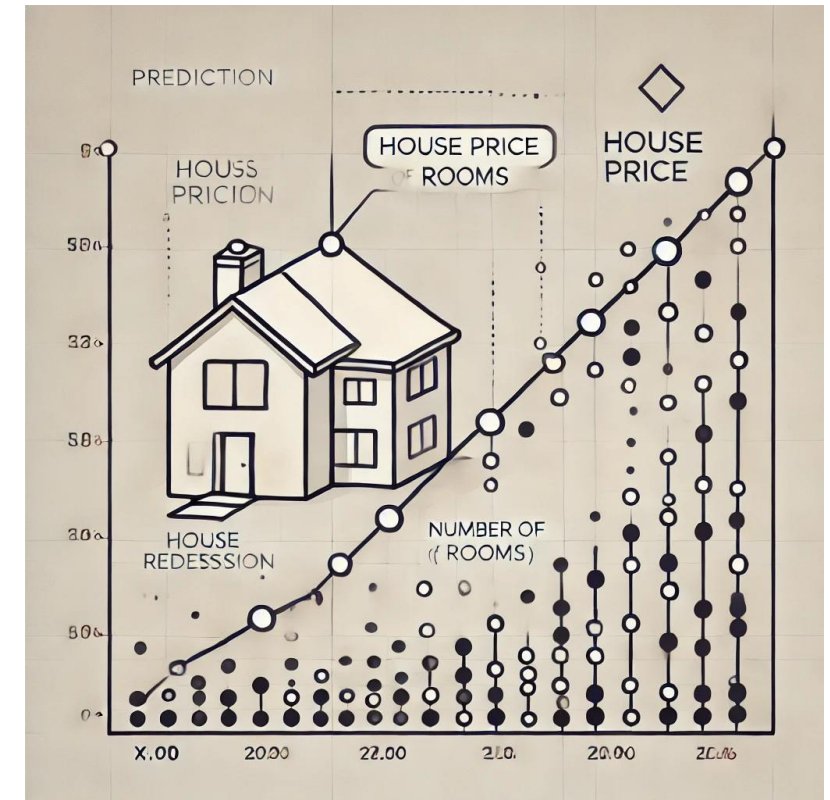
Regression line equation: $y_i = a + x_i * b$

House price = $b_0 + b_1 * (\text{Number of rooms}) + \text{Error}$

b_0 = House price with 0 rooms

b_1 = Gradient, that shows how the house prices changes with every additional room

Error = Difference between prediction and actual house price



Advantages & disadvantages of linear regression algorithms in the context of AI



Advantages

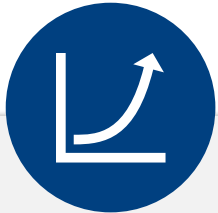
- **Simplicity and Interpretability:** Simple to understand and implement; the results are easily interpretable as coefficients for each variable
- **Efficiency:** Calculations are efficient and fast, even with large datasets; this makes them suitable for quick prototyping and initial analyses



Disadvantages

- **Limited Complexity:** Can only capture linear relationships between variables
- **Assumptions:** Based on assumptions such as linear relationships and independence of errors; if assumptions are not met, predictions can be inaccurate
- **Sensitivity to Outliers:** Sensitivity to outliers (atypical data points) that can strongly influence model coefficients

Decision trees are a supervised learning method



Linear regression

- Supervised learning method
- Regression algorithm

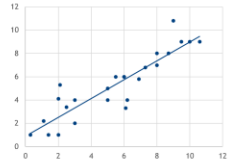
What is a linear regression?

x: observation **y:** output value **b:** gradient **a:** interception

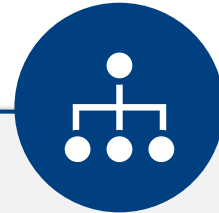
A linear regression represents a statistical procedure that explains an observed dependent variable (y) by one or more independent variables (x) and maps the underlying relationship in a regression graph.

Adding new data points influences the calculation of the linear regression line.

Regression line equation: $y_1 = a + x_2 + b$



19.07.2022 ABBA | Fortgeschritten* 26



Decision tree

- Supervised learning method
- Regression/classification algorithm

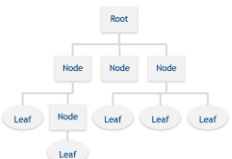
What is a decision tree?

Root node **Decision node** **Leaf**

A decision tree (DT) represents a decision procedure based on a sequence of „questions“. These questions are usually called **tests**. Each question is represented by a **node** in the DT.

The answer to the first question determines which question will be asked next.

Question will be asked until a **decision** is reached. The decision is the final **prediction** of the DT. The decision is represented by a **leaf** in the DT.



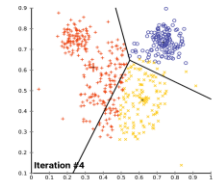
19.07.2022 ABBA | Fortgeschritten* 26



K-Means

- Unsupervised learning method
- Clustering algorithm

The k-means algorithm is an example of unsupervised learning methods.



Functionality

- Choice of a value k (=number of clusters)
- Random selection of center coordinates for each cluster
- Two-step process:
 - Allocation of all data points to the nearest center point
 - Updates of the center points such that the squared deviations of the cluster centers becomes minimal
- Repeating the process until convergence is achieved

Iteration #4

The k-means algorithm is a method for cluster analysis that classifies data points into a certain number (=k) of clusters.

Image: Wikipedia. Source: Towardsdatascience.com

<https://blog.iao.fraunhofer.de/endlich-verstaendlich-ki-verfahren-einfach-erklaert/>; <https://handbuch-ki.net/die-intelligenz-in-der-maschine/>

What is a decision tree?

Root node

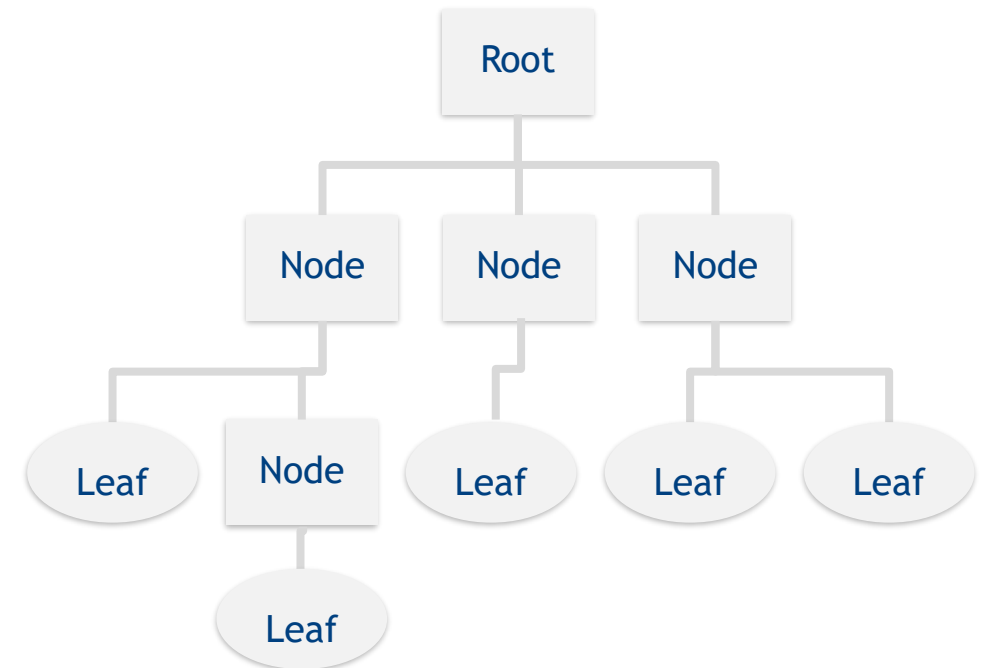
Decision node

Leaf

A decision tree (DT) represents a decision procedure based on a sequence of „**questions**“.
These questions are usually called **tests**.
Each question is represented by a **node** in the DT.

The answer to the first question determines which question will be asked next.

Question will be asked until a **decision** is reached.
The decision is the final **prediction** of the DT.
The decision is represented by a **leaf** in the DT.



Blockeel et al. (2023), Towardsai.com



Mathematical function

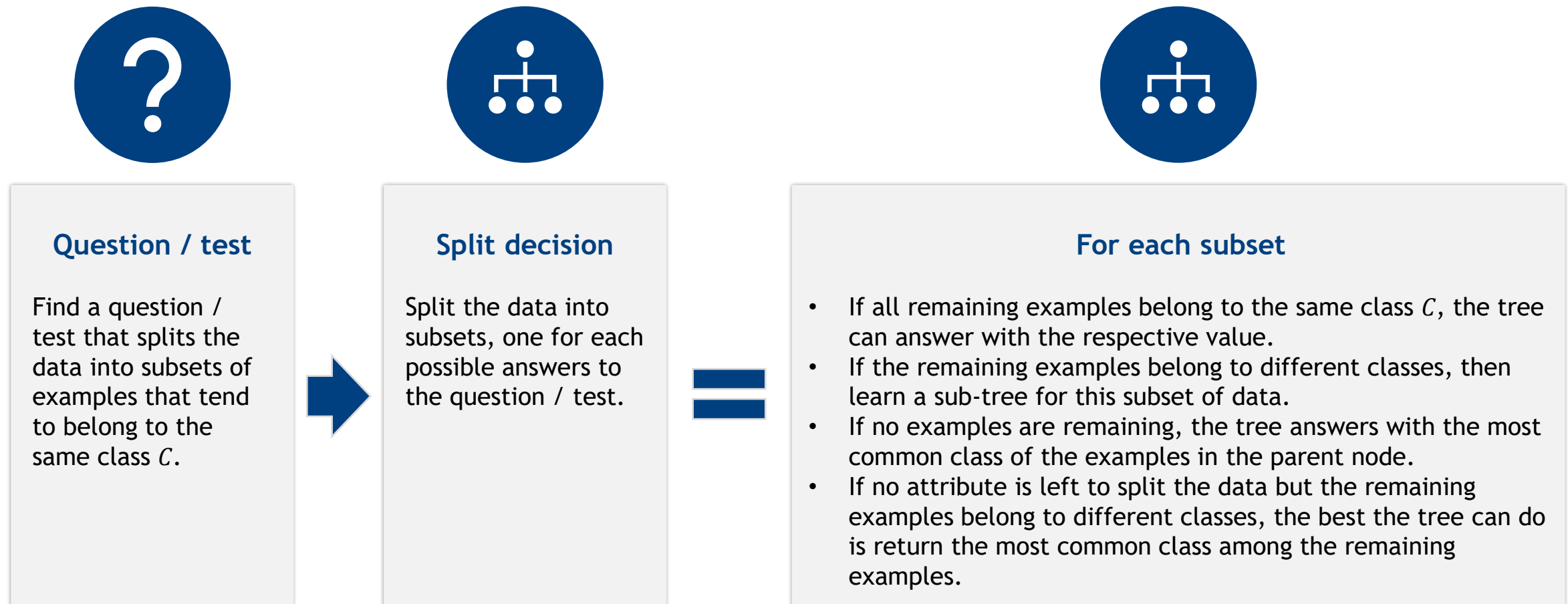
- Decision trees represent a mathematical **function** from X to Y
- Any given input $x \in X$ is mapped to exactly one value $y \in Y$
- Decision trees can (not only) represent every imaginable boolean function



Tree types

- We can distinguish different types of decision trees:
 - **Classification tree**: if the input set Y is nominal (e.g., categorizing emails as "spam" or "non-spam")
 - **Regression tree**: if the input set Y is numerical (e.g., estimate the price of a house based on its characteristics)

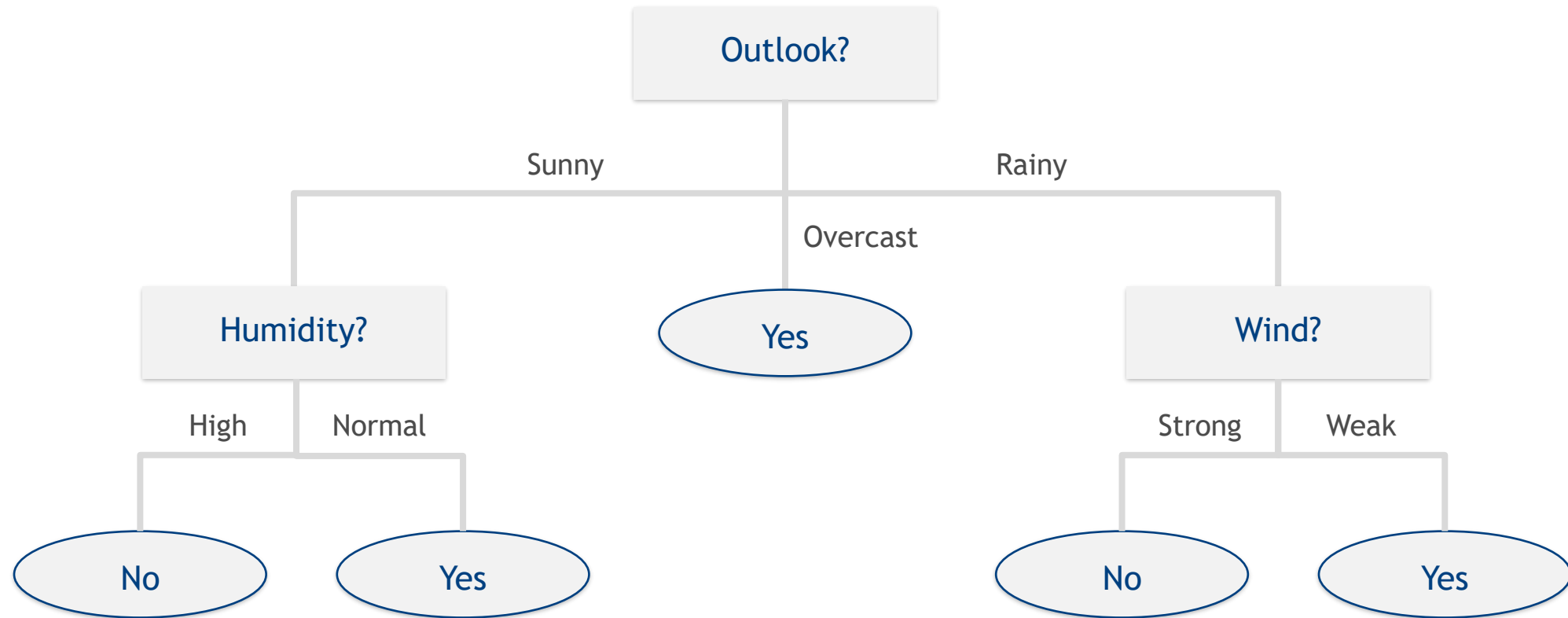
General principle for learning a decision tree



Althoff (2019), Blockeel (2020)

Exemplary decision tree

Should we go play tennis today?



Advantages & disadvantages of decision tree algorithms in the context of AI

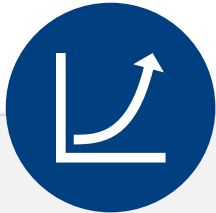
Advantages

- **Efficiency:** Decision trees are generally fast to train and making predictions with decision trees is usually efficient
- **Interpretability:** Decision trees are easy to understand, and their predictions can be „explained“
- **Robustness to Outliers:** Decision trees are typically robust to outliers and noise in the data

Disadvantages

- **Overfitting:** Decision trees are prone to overfitting, especially when they become deep and complex
- **Instability:** Small changes in the training data can result in different tree structures
- **Bias Toward Imbalance:** In classification tasks with imbalanced classes, decision trees can produce trees biased toward certain classes

The k-means algorithm is an example of unsupervised learning methods



Linear regression

- Supervised learning method
- Regression algorithm

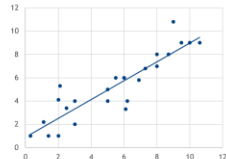
What is a linear regression?

x: observation **y:** output value **b:** gradient **a:** interception

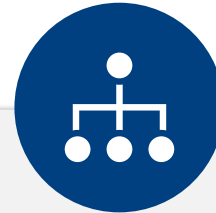
A linear regression represents a statistical procedure that explains an observed dependent variable (y) by one or more independent variables (x) and maps the underlying relationship in a regression graph.

Adding new data points influences the calculation of the linear regression line.

Regression line equation: $y_i = a + x_i \cdot b$



19.07.2022 ABBA | Vortegender* 26



Decision tree

- Supervised learning method
- Regression/classification algorithm

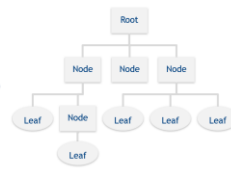
What is a decision tree?

Root node **Decision node** **Leaf**

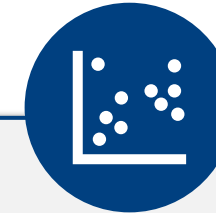
A decision tree (DT) represents a decision procedure based on a sequence of „questions“. These questions are usually called **tests**. Each question is represented by a **node** in the DT.

The answer to the first question determines which question will be asked next.

Question will be asked until a **decision** is reached. The decision is the final **prediction** of the DT. The decision is represented by a **leaf** in the DT.



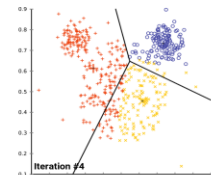
19.07.2022 ABBA | Vortegender* 26



K-Means

- Unsupervised learning method
- Clustering algorithm

The k-means algorithm is an example of unsupervised learning methods.



Functionality

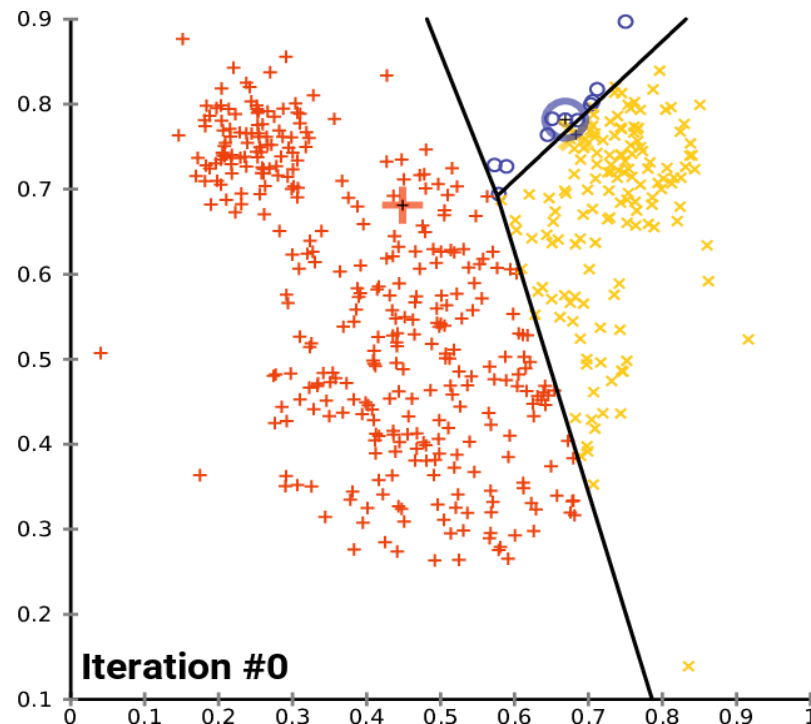
- Choice of a value k (=number of clusters)
- Random selection of center coordinates for each cluster
- Two-step process:
 - Allocation of all data points to the nearest center point
 - Updates of the center points such that the squared deviations of the cluster centers becomes minimal
- Repeating the process until convergence is achieved

Iteration #4

The k-means algorithm is a method for cluster analysis that classifies data points into a certain number (=k) of clusters.

Image: Wikipedia. Source: Towardsdatascience.com

The k-means algorithm is an example of unsupervised learning methods



<https://de.wikipedia.org/wiki/K-Means-Algorithmus>

Functionality

- Choice of a value k (=number of clusters)
- Random selection of center coordinates for each cluster
- Two-step process:
 - Allocation of all data points to the nearest center point
 - Updates of the center points such that the squared deviations of the cluster centers becomes minimal
- Repeating the process until convergence is achieved



The k-means algorithm is a method for cluster analysis that classifies data points into a certain number ($=k$) of clusters

Image: Wikipedia, Source: Towardsdatascience.com

Advantages & disadvantages of k-means algorithms in the context of AI

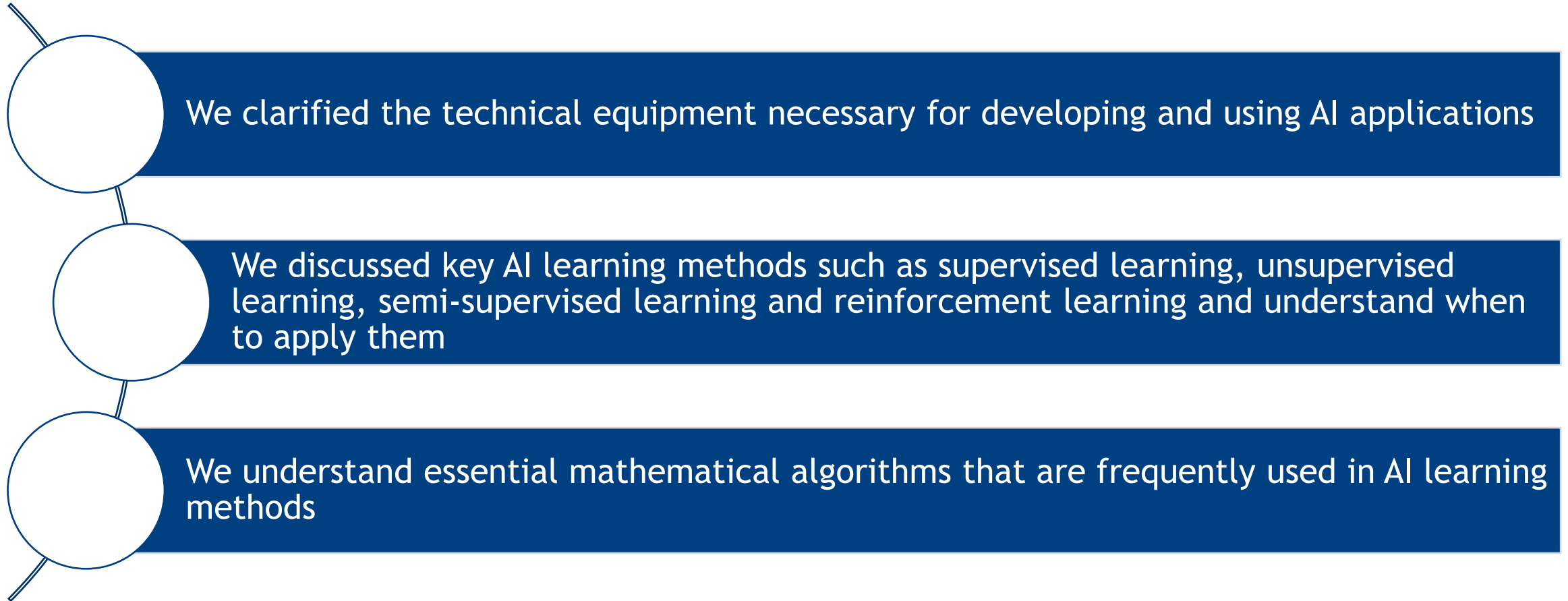
Advantages

- **Scalability:** K-means can handle large datasets efficiently, making it suitable for data with many samples
- **Speed:** Its speed makes it suitable for real-time and interactive applications
- **Versatility:** K-means can be applied to a wide range of data types, including numerical, categorical, and mixed data

Disadvantages

- **Sensitivity to initial centroids:** K-means results can vary depending on the initial placement of centroids
- **Dependence on number of clusters (K):** The number of clusters (K) needs to be specified in advance; choosing the wrong K can lead to poor results
- **Assumes spherical clusters:** K-means assumes that clusters are spherical and equally sized, which may not be the case in real-world data

Today's lecture at a glance



Questions, comments, observations



Scientific references

- Blockeel, H., Devos, L., Frénay, B., Nanfack, G. and Nijssen, S. (2023) Decision trees: from efficient prediction to responsible AI. Frontiers in Artificial Intelligence. <http://dx.doi.org/10.3389/frai.2023.1124553>
- Lins, S., Pandl, K., Teigeler, H., Thiebes, S., Bayer, C. and Sunyaev, A. (2021) Artificial Intelligence as a Service. Business Information Systems Engineering 63(4).
- Stahlknecht, P. and Hasenkamp, U. (1999) Einführung in die Wirtschaftsinformatik. Springer Berlin, Heidelberg. <https://doi.org/10.1007/978-3-662-06903-5>

Non-scientific references

- <https://www.technologyreview.com/2016/03/23/8768/intel-puts-the-brakes-on-moores-law/>
- <https://de.statista.com/infografik/26868/rechenleistung-des-jeweils-staerksten-supercomputers-der-welt/#:~:text=Ein%20Supercomputer%20mit%20der%20Leistung,berechnen%20sie%20u.a.%20komplexe%20Simulationen.>
- <https://fullstackfeed.com/global-machine-learning-tools-and-platforms-landscape-01-2021/>
- <https://gowthamy.medium.com/machine-learning-supervised-learning-vs-unsupervised-learning-f1658e12a780>
- https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html#
- https://en.wikipedia.org/wiki/Iris_flower_data_set
- <https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818>
- <https://medium.com/ai%C2%B3-theory-practice-business/reinforcement-learning-part-1-a-brief-introduction-a53a849771cf>
- <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- <https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53>

Pictures and Videos

- https://www.youtube.com/watch?v=kopoLzvh5jY&feature=emb_logo
- <https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3>