LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

INTRODUCTION TO DATA SCIENCE
TOPIC OUTLINE

UNIT 5

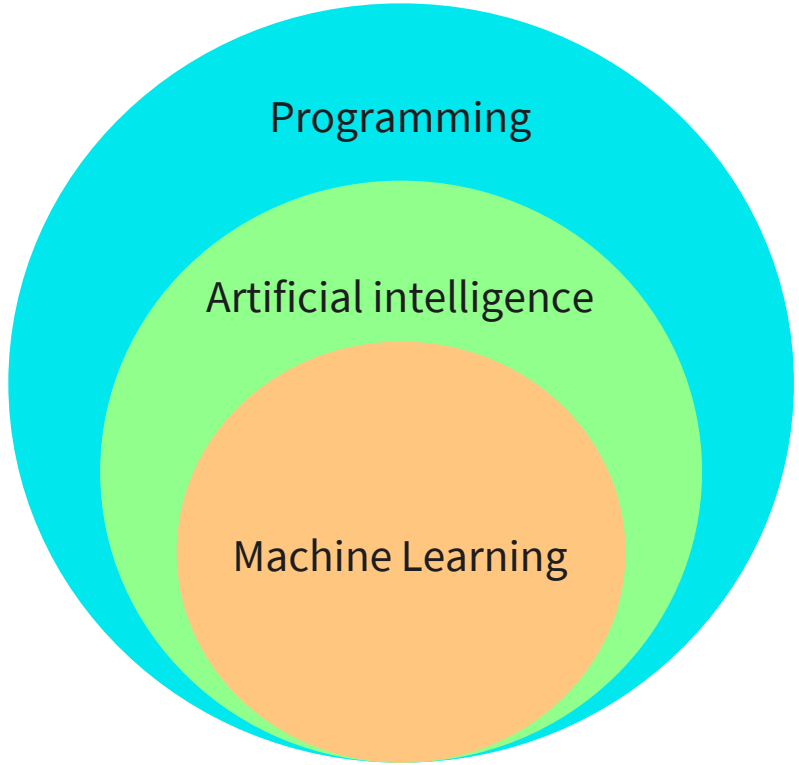
MACHINE LEARNING

STUDY GOALS

- Explain what is meant by machine learning.
- Be familiar with common terms and definitions in machine learning.
- Learn the different applications of machine learning.
- Understand concepts of classification and regression.
- Comprehend the difference between each of the machine learning paradigms.

Machine learning ...

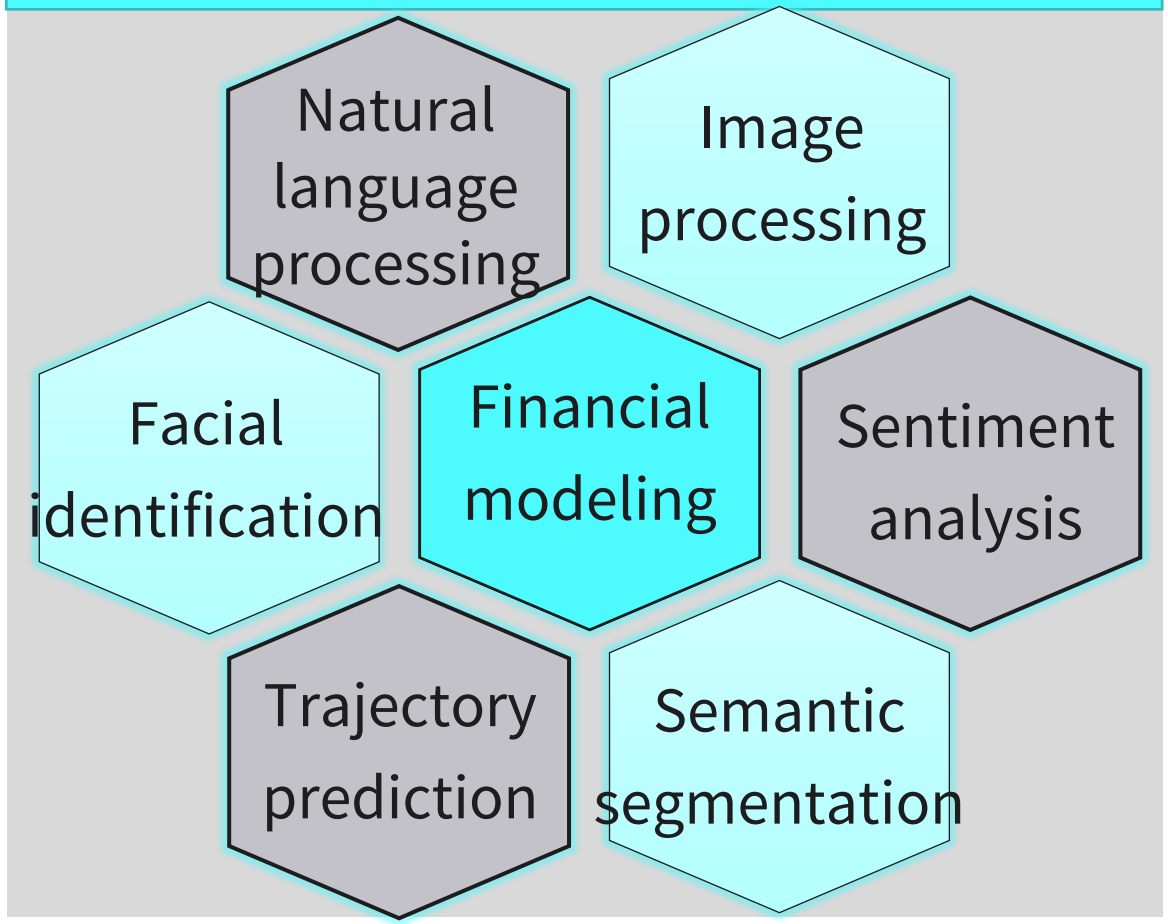
- is a **subfield** of Artificial Intelligence (**AI**).
- is a **mathematical** and **algorithmic** approach
- is devoted to understanding and building **methods that "learn"**.
- methods leverage data to improve performance on some set of tasks.



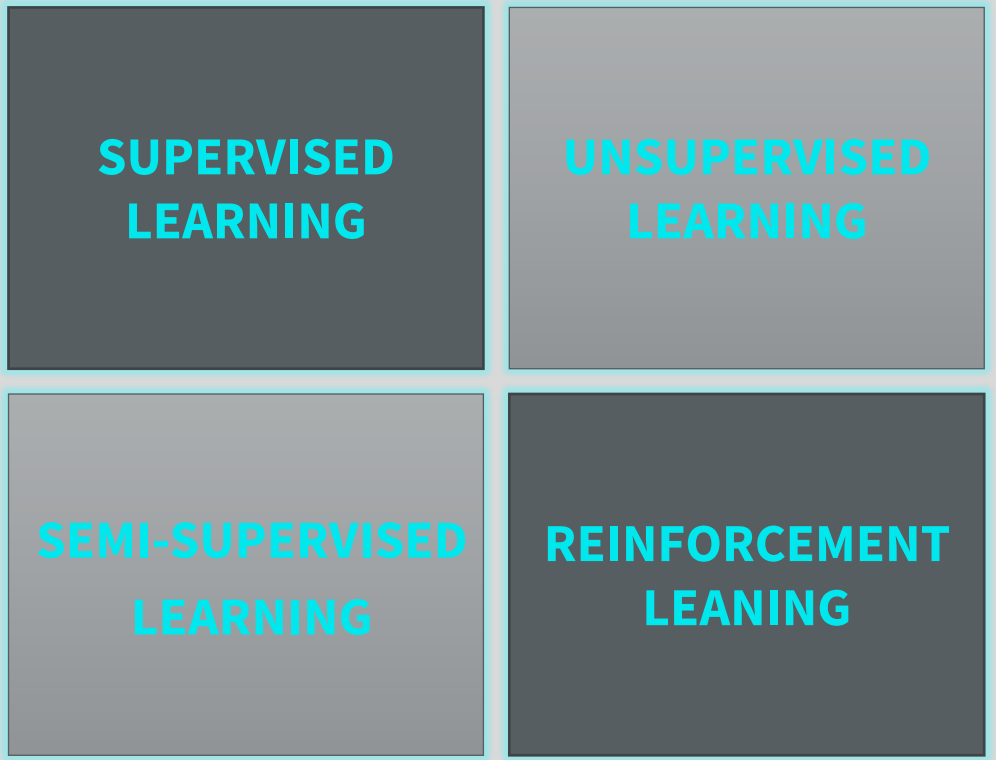
Machine Learning as a Subfield of AI

Source of the text: Machine Learning, 2023; Zöller, 2022.

Applications of ML



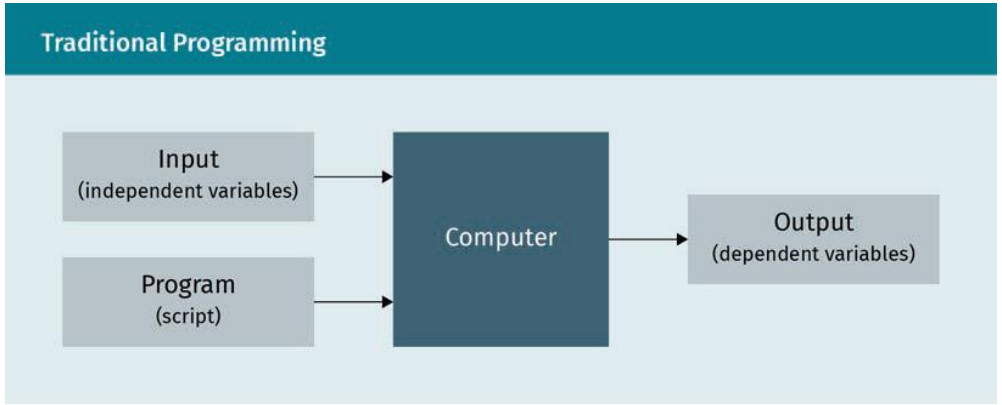
ML Paradigms



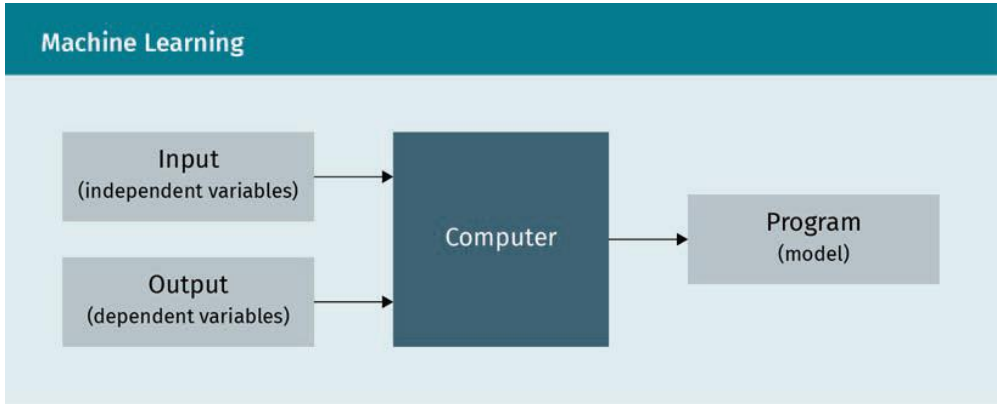
Source of the text: Zöller, 2022.

Machine learning concepts

- **Traditional programming** constructs an **explicit** processing of input variables into desired outputs via a set of **code** instructions.
- **ML** algorithms build **models** based on sample **data**, in order to make **predictions** or **decisions** without being explicitly programmed to do so.



Traditional Programming

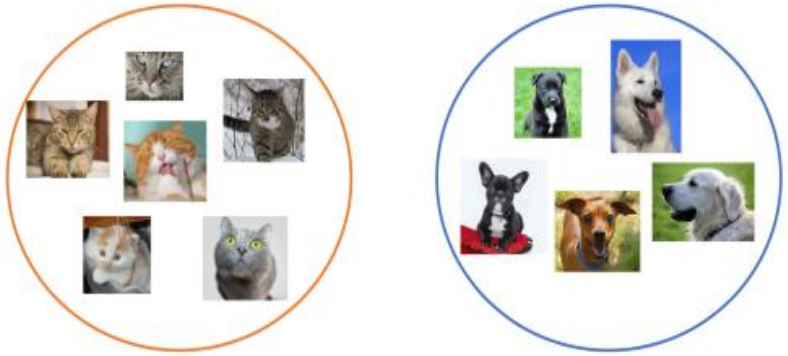


Machine learning

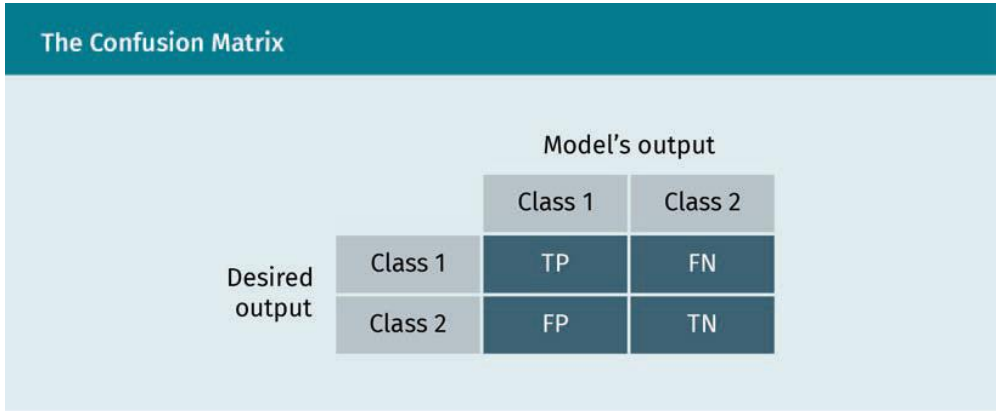
Source of the text: Zöller, 2022.
Source of the image: Zöller, 2022, pp.12-13.

Classification

- **Objective**: Develop a ML model to **map** the inputs to the outputs and **predict** the **classes** of new inputs.
- **Accuracy** can be presented in a confusion matrix.
- Evaluation **metrics**:
  - $\boldsymbol{Precision} = \frac{TP}{TP+FP}$
  - $\boldsymbol{Recall} = \frac{TP}{TP+FN}$
  - $\boldsymbol{F_{Score}} = \frac{2 \cdot (Precision \cdot Recall)}{Precision+Recall}$



Dog and Cat classification

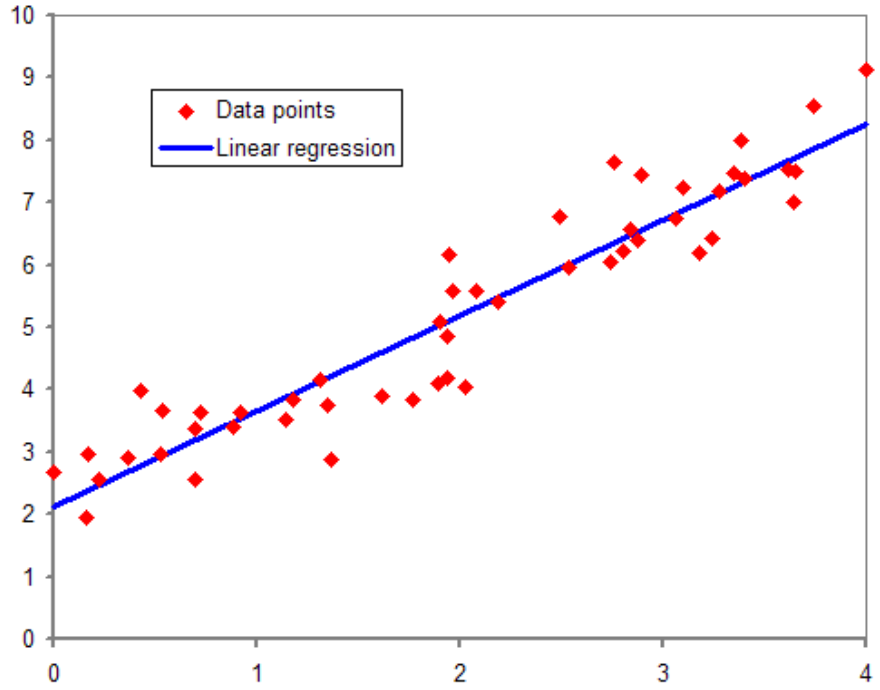


Confusion matrix

Source of the text: Zöller, 2022.
Source of the image: Saskai, n.d.; Zöller, 2022, p. 15.

Regression

- **Objective**: Develop a ML model to **relate** the inputs x to the outputs y and **predict** the output **values** $\hat{y}$ for new inputs
- Evaluation **metrics**:
  - Mean Square Error: $\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Root Mean Square Error: $\boldsymbol{RMSE} = \sqrt{MSE}$
  - Mean Absolute Error: $\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$



An example of Regression

Source of the text: Fernandez, 2020; Jason, 2021; Zöller, 2022.
Source of image: Amatulic, 2007, Public Domain.

Supervised Learning

- **Dataset**: a collection of **labeled** samples, containing both inputs (independent variables) and outputs (dependent variable)
- **Objective**: Develop a ML model to **relate** the inputs to the outputs of in the training set and **predict** the outputs for new inputs.

Supervised Learning Structure

Source of the text: Zöller, 2022.

Supervised Learning

Supervised Learning Examples		
Example Dataset	Prediction	Type
Previous home sales	How much is a specific home worth?	Regression
Previous loans that were paid	Will this client default on a loan?	Classification
Previous weeks' visa applications	How many businesspersons will apply for visa next week?	Regression
Previous statistics of benign/malignant cancers	Is this cancer malignant?	Classification

Supervised Learning Techniques	
Technique	Obtained Function
Linear classifier, linear regression, multi-linear regression.	Numerical functions
Support Vector Machine (SVM), Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden Markov models (HMM).	Parametric Probabilistic functions
K-nearest neighbors, Kernel regression, Kernel density estimation	Non-parametric instance based functions
Decision tree	Non-metric symbolic functions

Source of the table: Zöller, 2022.

Unsupervised Learning

- **Dataset**: a collection of **unlabeled** samples, containing only inputs (independent variables) while outputs (dependent variable) are unknown.
- **Objective**: Develop a ML model to **discover** the salient **patterns** and **structures** within the training set.

Unsupervised Learning Structure

Source of the text: Zöller, 2022.

Unsupervised Learning

Unsupervised Learning Examples		
Example dataset	Discovered patterns	Type
Customers profiles	Are these customers similar?	Clusters
Previous transactions	Is a specific transaction odd?	Anomaly detection
Previous purchasing	Are these products purchased together?	Association discovery

Unsupervised Learning Techniques	
Technique	Description
K-Means, hierarchical clustering	Clustering analysis
Gaussian mixture model (GMM), graphical models	Density estimation
DBSCAN	Outlier detection
Principal component analysis, factor analysis	Dimensionality reduction

Source of the table: Zöller, 2022.

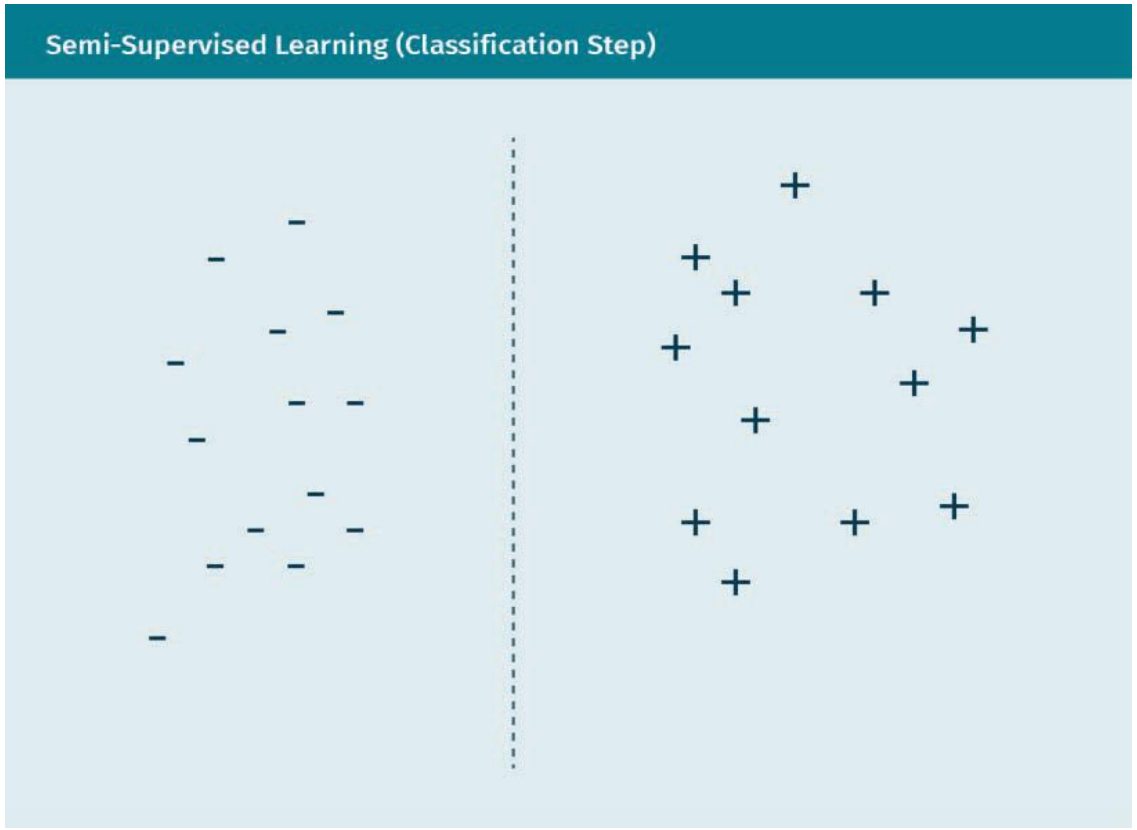
Semi-Supervised Learning

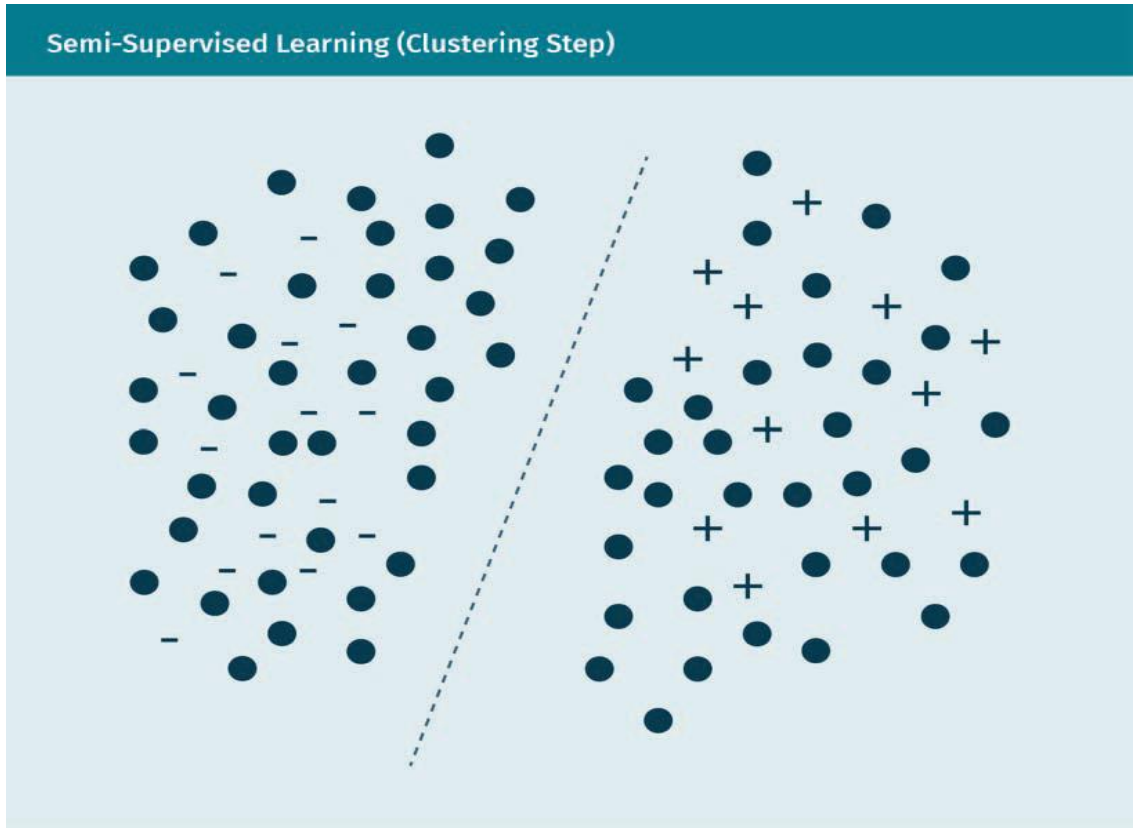
- **Dataset**: a collection of both **labeled** samples (a small portion of data), and **unlabeled** samples (lots of data)
- **Objective**: mix of supervised and unsupervised learning to combine the properties of both.
- **2 steps:**
  - Supervised learning is performed on few labeled data.
  - Unsupervised learning is performed on a large quantity of unlabeled data.

Semi-Supervised Learning Structure

Source of the text: Zöller, 2022.

Semi-Supervised Learning



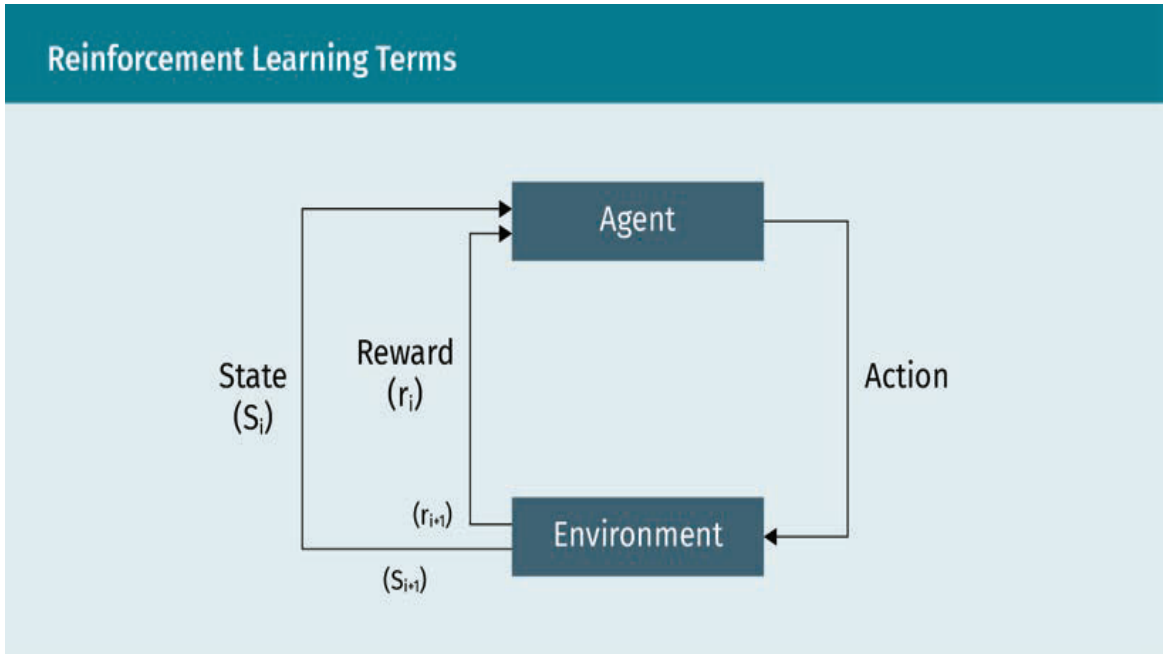


Two Steps of Semi-Supervised Learning

Source of the image: Zöller, 2022, pp. 22-23.

Reinforcement Learning

- **Objective**: To find an **action policy** that achieves a given goal by **trial-and-error** interactions with the environment.
- **“Cause and effect”** method: An action is performed to achieve a maximum reward.
- **Reward function** acts as feedback to the agent.



Reinforcement Learning Structure

Source of the text: Zöller, 2022.
Source of image: Zöller, 2022, p. 24.

Reinforcement Learning

Reinforcement Learning Ingredients			
Agent	hypothetical entity that performs actions in an environment to gain some reward	Reward (R)	an immediate return sent back from the environment to evaluate the last action by the agent
Action (A)	all the possible moves that the agent can take	Policy (π)	strategy that the agent employs to determine the next action based on the current state
Environment (E)	scenario that the agent has to face	Value (V)	the expected long-term return of the current state under policy
State (S)	current situation returned by the environment		

Source of the text: Zöller, 2022.
Source of the image: Zöller, 2022, p. 25.

REVIEW STUDY GOALS

- Explain what is meant by machine learning.
- Be familiar with common terms and definitions in machine learning.
- Learn the different applications of machine learning.
- Understand concepts of classification and regression.
- Comprehend the difference between each of the machine learning paradigms.

SESSION 5

TRANSFER TASK

TRANSFER TASKS

Explain how Machine Learning can be applied to improve the purchasing services of an online shop.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. Semi-supervised learning combines aspects of …
   a) …supervised and reinforcement learning.
   b) …unsupervised and reinforcement learning.
   c) …reinforcement learning and active learning.
   d) …supervised and unsupervised learning.

2. Which of the following are the low and high bounds for the F-Score?
   a) $[0,100]$
   b) $[0,1]$
   c) $[-1,1]$
   d) $[-1,0]$

3. Normalized data are centered where?

a) 0
b) 1
c) -1
d) 10

4. Grouping news articles according to similarity can be solved using which of the following?
   a) Regression
   b) Classification
   c) Reinforcement Learning
   d) Clustering

5. Classification problems fall under which category?

a) unsupervised learning
b) reinforcement learning
c) supervised learning
d) supervised and unsupervised learning

LIST OF SOURCES

Text:

Fernandez, J. (2020). *Introduction to regression analysis*. https://towardsdatascience.com/introduction-to-regression-analysis-9151d8ac14b3

Jason, B. (2021). *Regression metrics for machine learning*. https://machinelearningmastery.com/regression-metrics-for-machine-learning/

Machine Learning. (2023, February 28). In *Wikipedia*. https://en.wikipedia.org/wiki/Machine_learning

Zöller, T. (2022). Course Book – Machine Learning. *IU International University of Applied Science.*

Images:

Amatulic. (2007). File:Normdist_regression.png. *Wikimedia Commons.* https://en.wikipedia.org/wiki/Regression_analysis

Sasaki, T. (n.d.). Dog and Cat Classification.png [Open source]. *Kaggle.* https://www.kaggle.com/code/sasakitetsuya/dog-and-cat-classification-by-mobilenet