LECTURER: Nghia Duong-Trung

MACHINE LEARNING

WHO I AM

- Name: Dr. Nghia Duong-Trung
- Senior Researcher / Project Lead @ The German Research Center for Artificial Intelligence (DFKI GmbH)
- Academic Teacher @ IU
  - Teaching courses: Machine Learning, Deep Learning, Artificial Intelligence, Data Science, Data Utilization, Neural Nets and Deep Learning, Introduction to Computer Science
  - Thesis Supervision.

- Profile: https://sites.google.com/ismll.de/duongtrungnghia/

- Course book: MachineLearning-DLMDSML01, provided by IU, myCampus.
- Basic Reading, provided by IU, myCampus.
- Video gallery: myCampus.
- Example of a practical exam, provided by IU, myCampus.
- Online tests and evaluation, provided by IU, myCampus.
- Additional teaching materials:

https://github.com/duongtrung/IU-MachineLearning-DLMDSML01

The GitHub repository is the additional teaching materials, consisting of extra slides regarding the current information on the learning domain. It helps students in discussion and group work. It does not necessarily reflect new questions in examination, and it does not present the IU on those extra content.

UNIT 1

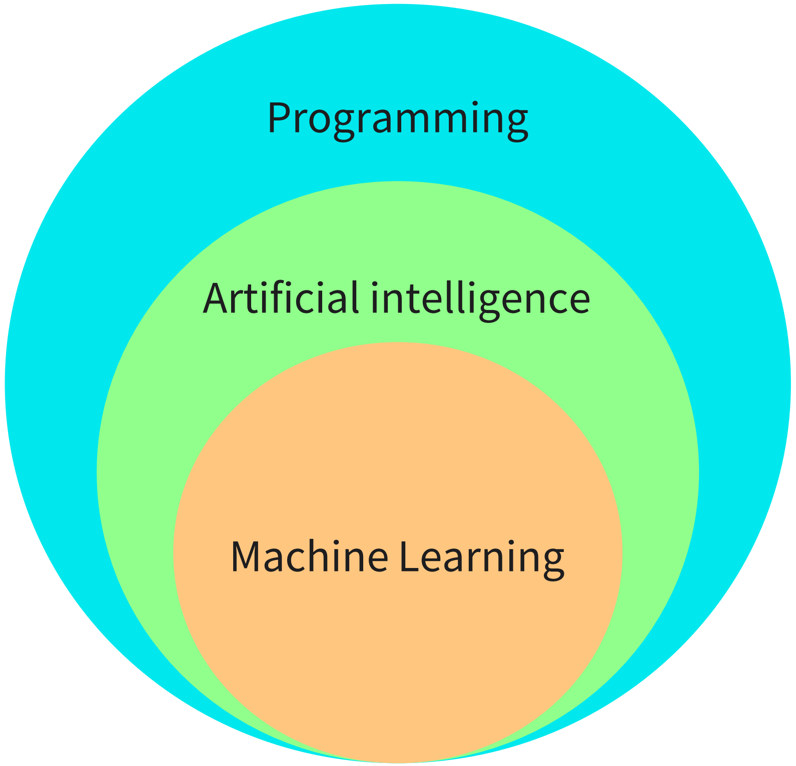
INTRODUCTION TO MACHINE LEARNING

STUDY GOALS

- Explain what is meant by machine learning.
- Know 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 ...

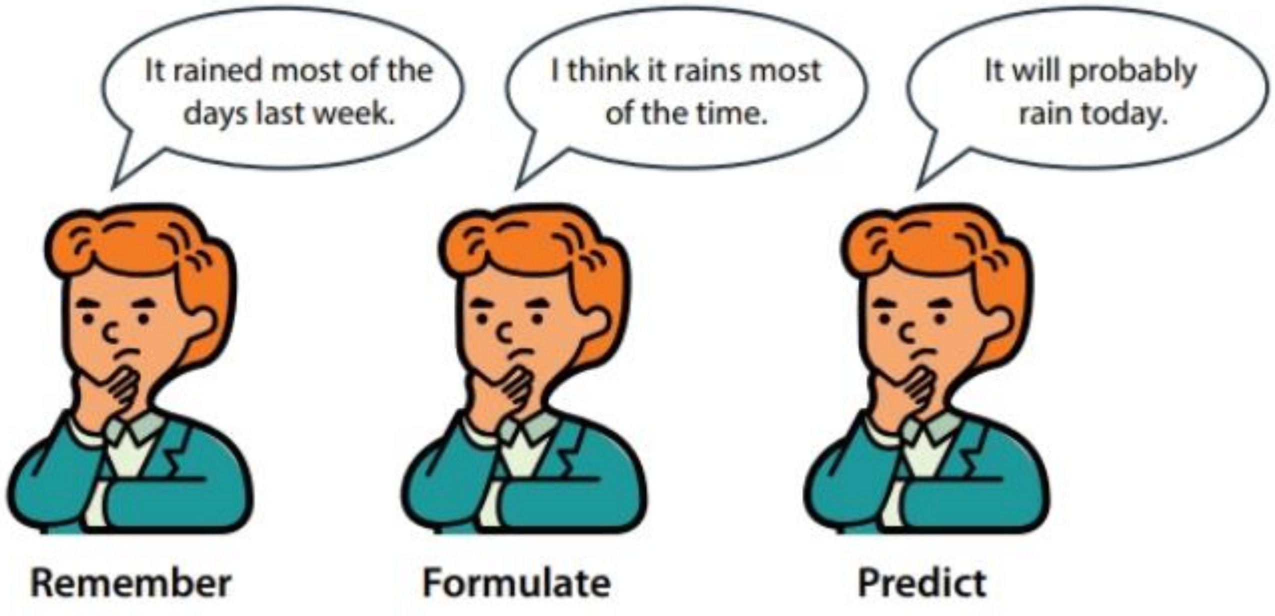
- is a **subfield** of Artificial Intelligence (**AI**).
- is a **mathematical** and **algorithmic** approach
- is devoted to understanding and building **methods that “learn”**.
- methods leverage data to improve performance on some set of tasks.



Machine learning as subfield of AI

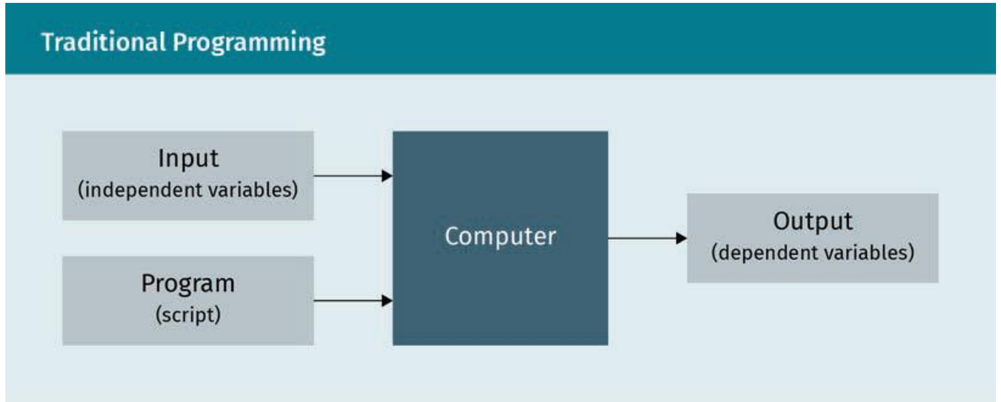
Source of the text: Machine Learning, 2023; Zöller, 2022.
Source of the image: File: Darwin's_finches_by_Gould.jpg (2023).

- Remember-formulate-predict framework
  - We remember past situations that were similar
  - We formulate a general rule
  - We use this rule to predict what may happen in the future

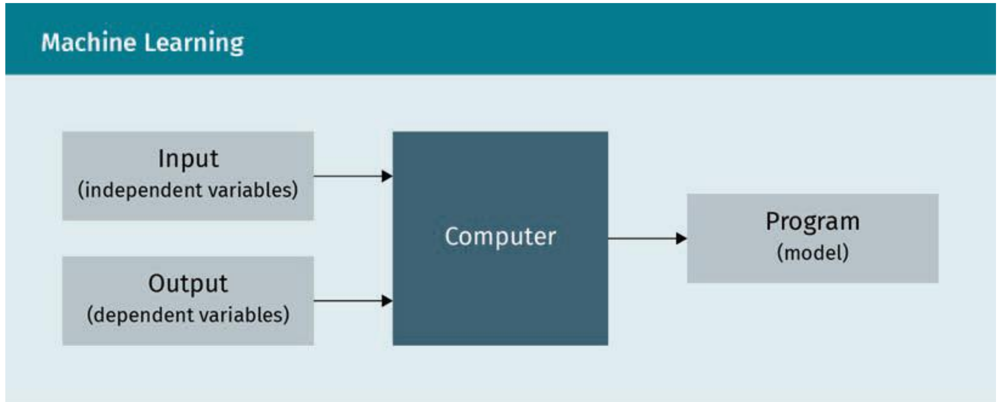


Machine learning concepts

- **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



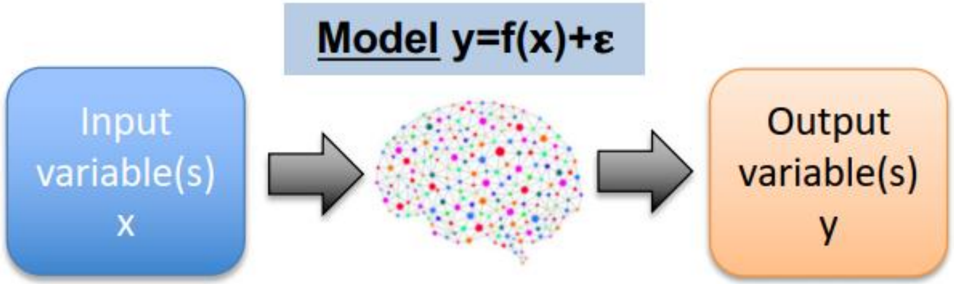
Traditional Programming

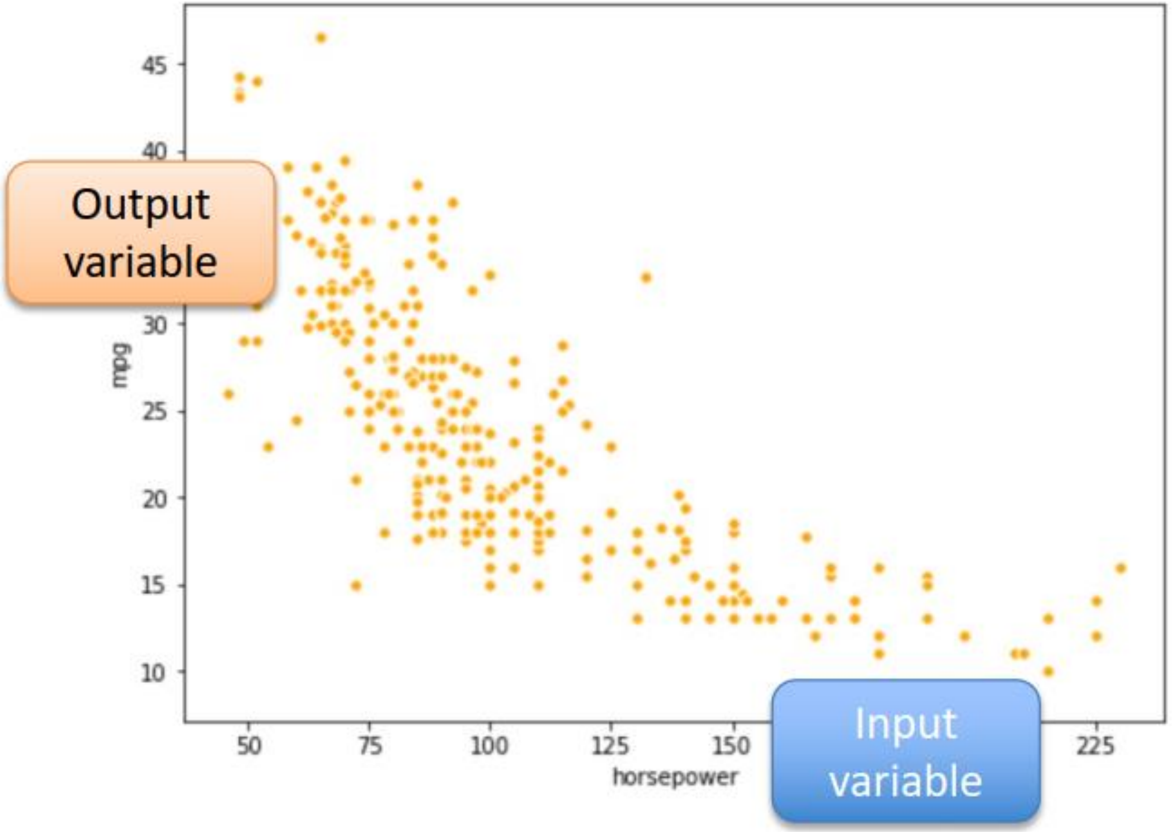


Machine learning

Source of the text: Zöller, 2022.
Source of the image: Zöller (2022, p.12-13).

A model is an approximation of the relationship between two variables.





WHAT IS A MODEL?

A model is an approximation of the relationship between two variables.

Observations of input data **(X)**

	Neighbor-hood	School district	Square footage	Number of bedrooms	Year built
House 1	Weycroft	Wake	3400	4	2010
House 2	Horton Creek	Wake	4200	5	2008
House 3	Cary Park	Chatham	3250	4	2012
...	...	...	...	...	...

Model $y=f(X)+\varepsilon$

Predictions of **target (y)**

Market sale price
$612,000
$675,000
$520,000
...

- To create a model, we define four things:
  - **Features**: to use
  - **Algorithm**: acts as a form/template for model
  - **Hyperparameter**: values for algorithm
  - **Loss function**: to optimize
- We train our model using historial data:
  - Algorithm & hyperparameters provide overall model form
  - Learn values for the model which minimize loss function

Applications of ML

- Natural language processing
- Image processing
- Facial identification
- Financial modeling
- Sentiment analysis
- Trajectory prediction
- Semantic segmentation

ML Paradigms

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING
- SEMI-SUPERVISED LEARNING
- REINFORCEMENT LEANING

Source of the text: Zöller, 2022.

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

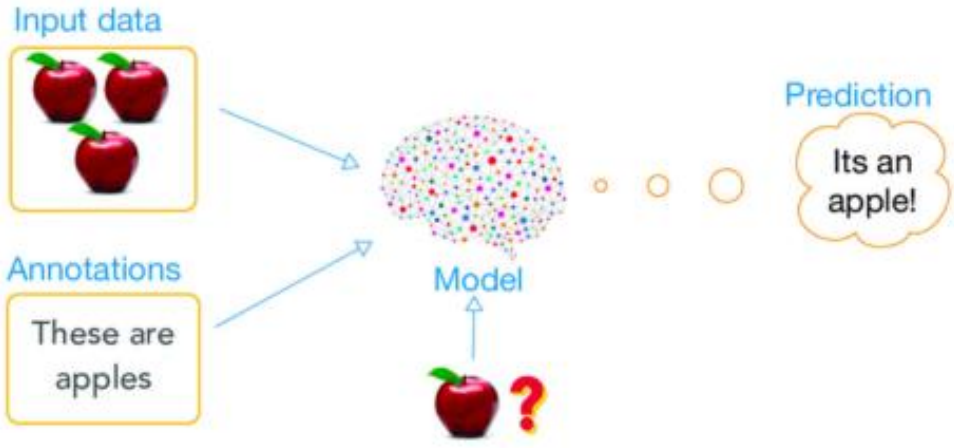


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

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

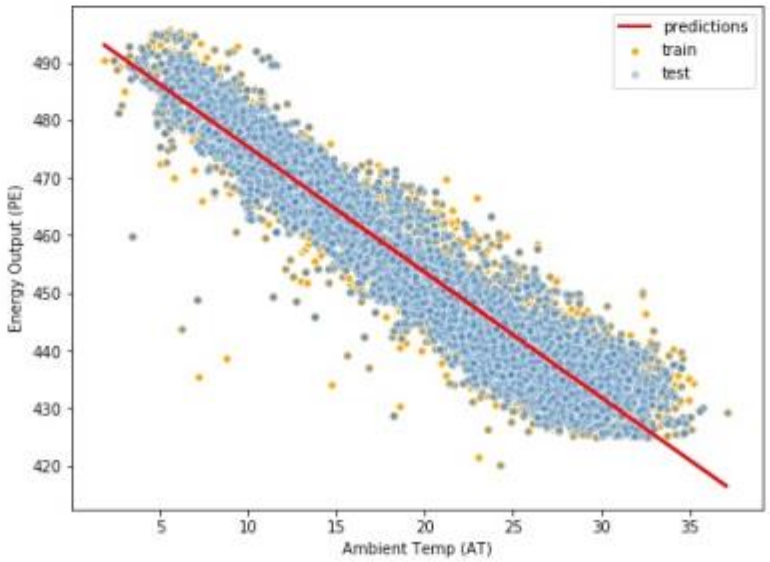
Supervised Learning Techniques	
Technique	Obtained Function
Linear classifier, linear regression, multi-linear regression.	Numerical functions
Support Vector Machine (SVM), Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden Markov models (HMM).	Parametric Probabilistic functions
K-nearest neighbors,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

Source of the table: Zöller (2022).

REGRESSION VS CLASSIFICATION

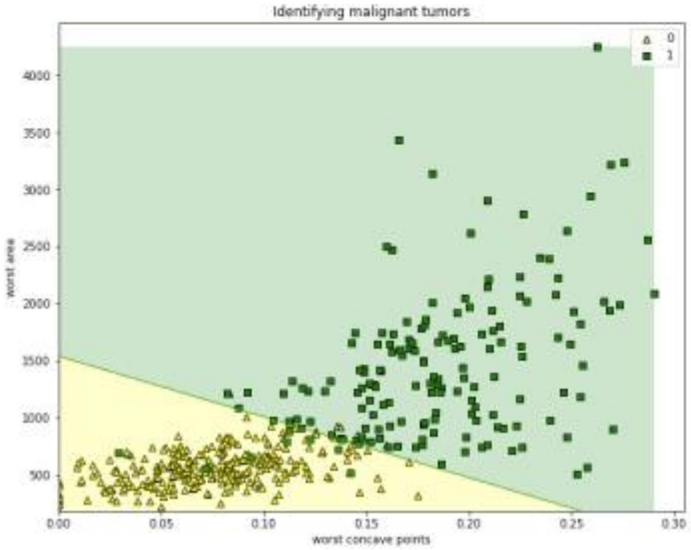
Regression

- Predict one or more **numerical** target variables
- E.g. home price, number of power outages, product demand



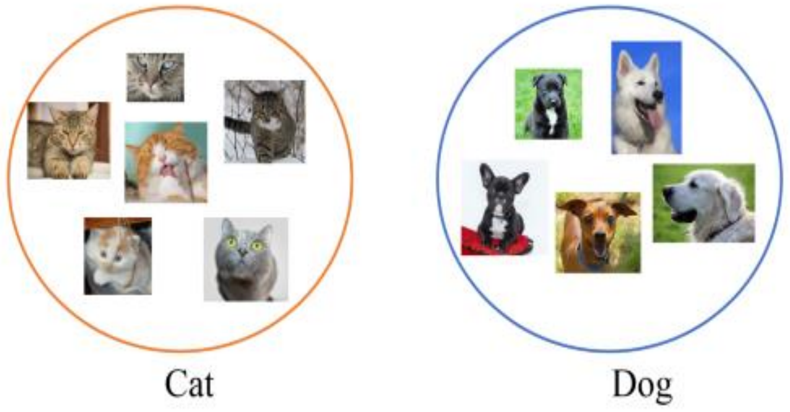
Classification

- Predicts a **class / category** – either binary or out of a set
- E.g. lung disease detection, identifying types of plants, sentiment analysis, detecting spam

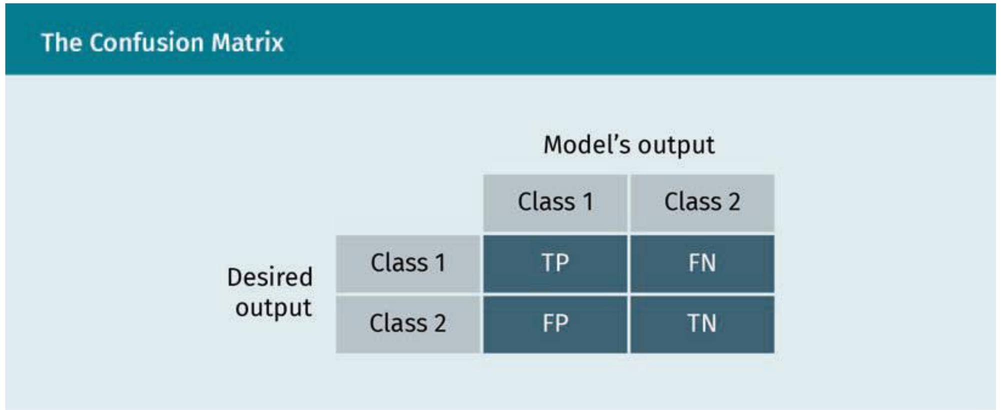


Classification

- **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**:
  - $Precision = \frac{TP}{TP+FP}$
  - $Recall = \frac{TP}{TP+FN}$
  - $F_{Score} = \frac{2\cdot(Precision\cdot Recall)}{Precision+Recall}$



Dog and Cat classification



The Confusion Matrix

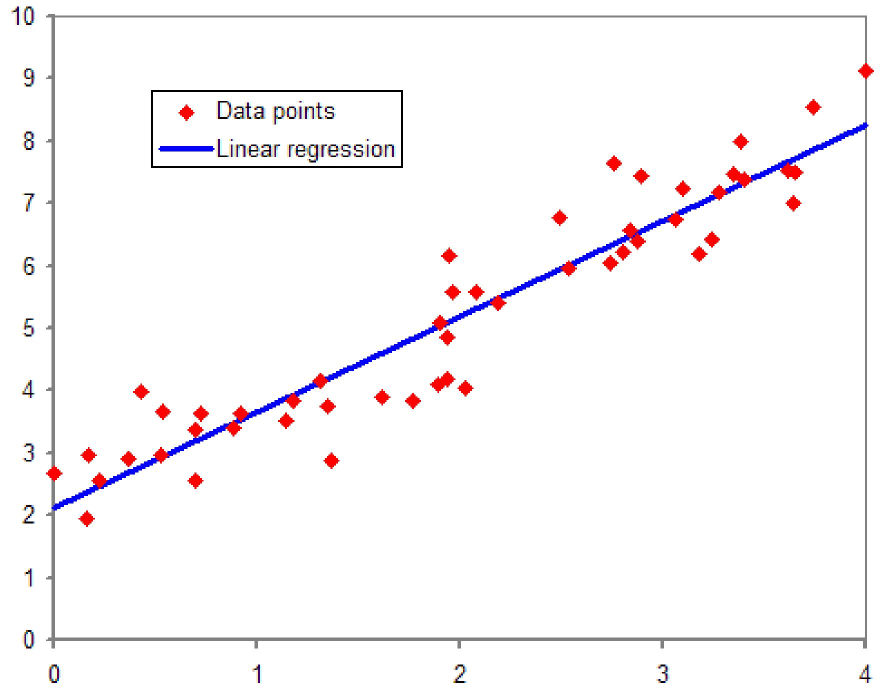
		Model's output	
		Class 1	Class 2
Desired output	Class 1	TP	FN
	Class 2	FP	TN

Confusion matrix

Source of the text: Zöller, 2022.
Source of the image: File:Dog and Cat Classification.png (2023, January); 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: File: Normdist_regression.png (2023).

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

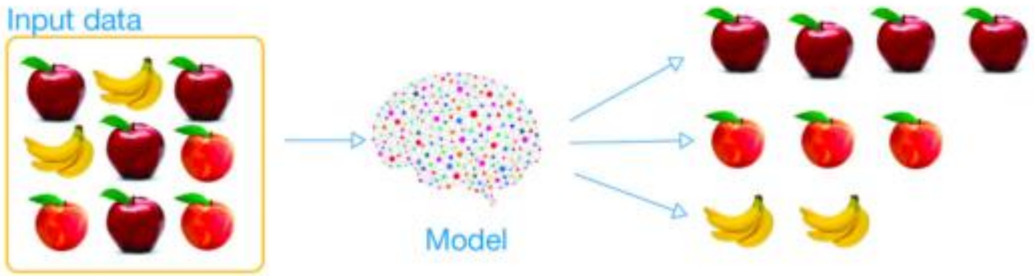


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

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).

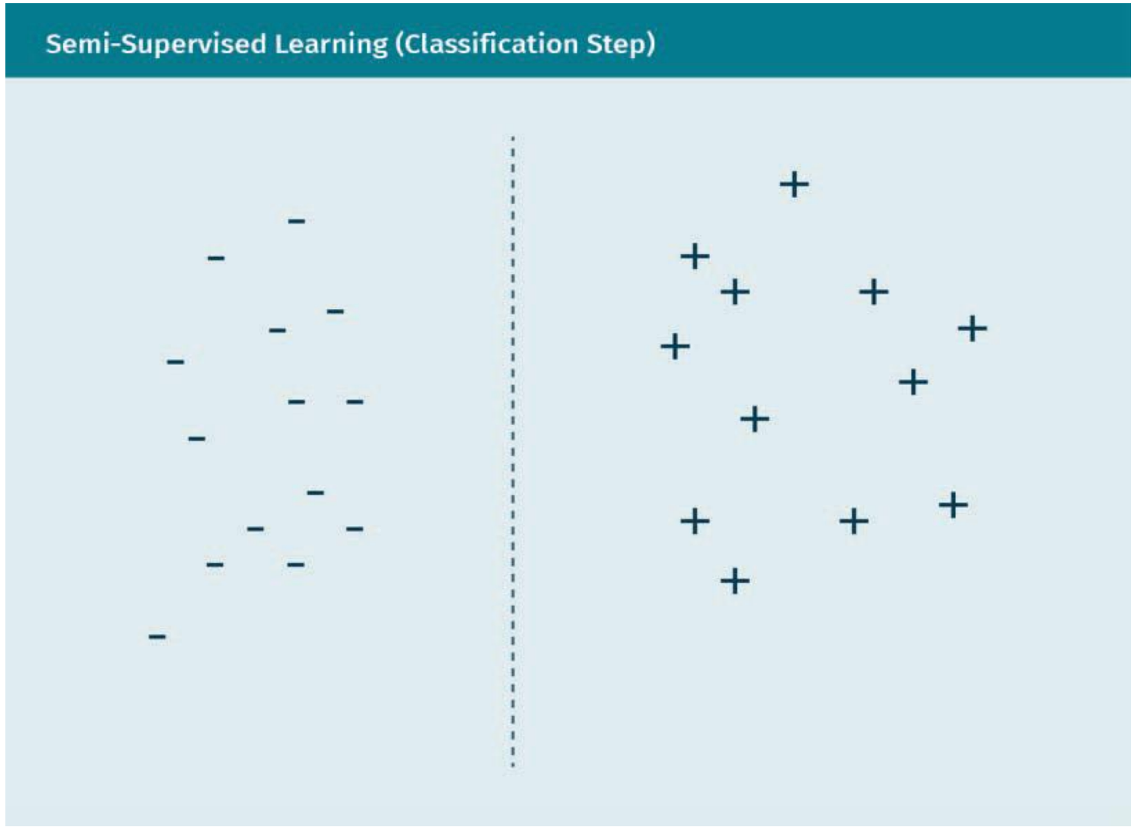
Semi-Supervised Learning

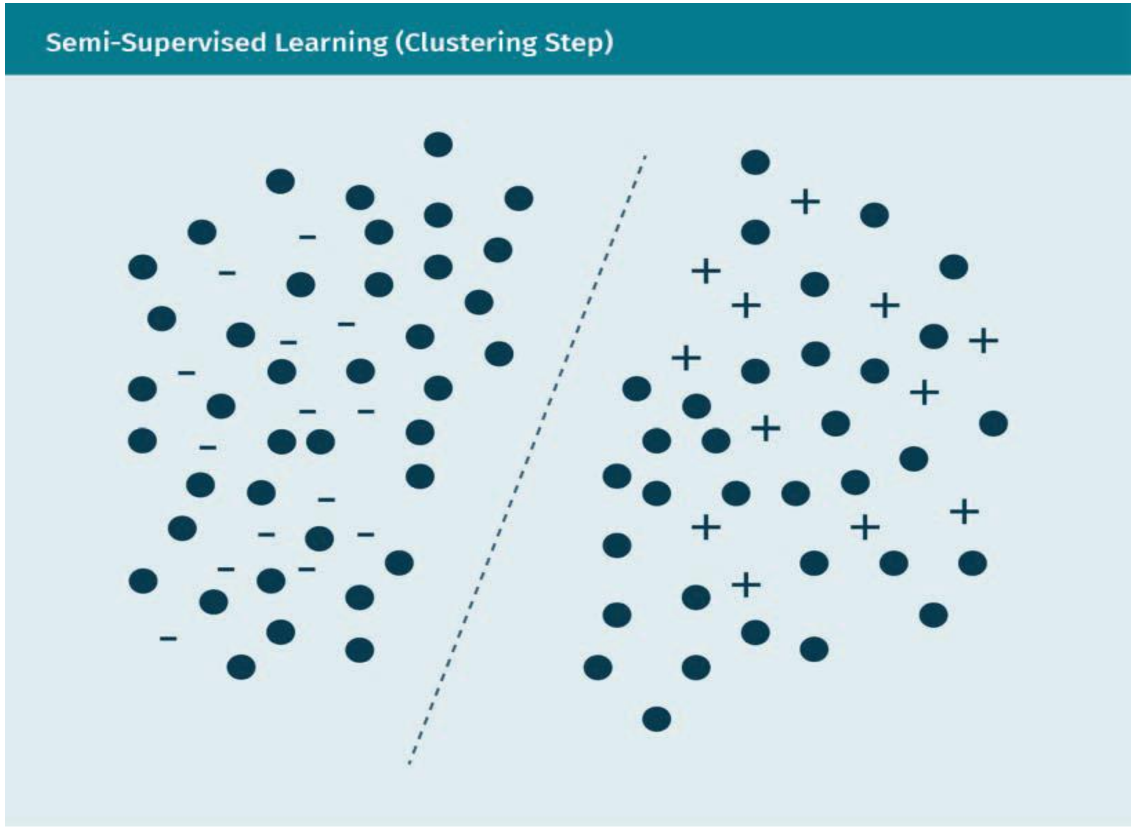
- **Dataset**: a collection of both **labeled** samples (a small portion of data), and **unlabeled** samples (lots of data)
- **Objective**: mix of supervised and unsupervised learning to combine the properties of both.
- **2 steps:**
  - Supervised learning is performed on few labeled data
  - Unsupervised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

Semi-Supervised Learning Structure

Source of the text: Zöller, 2022.

Semi-supervised Learning



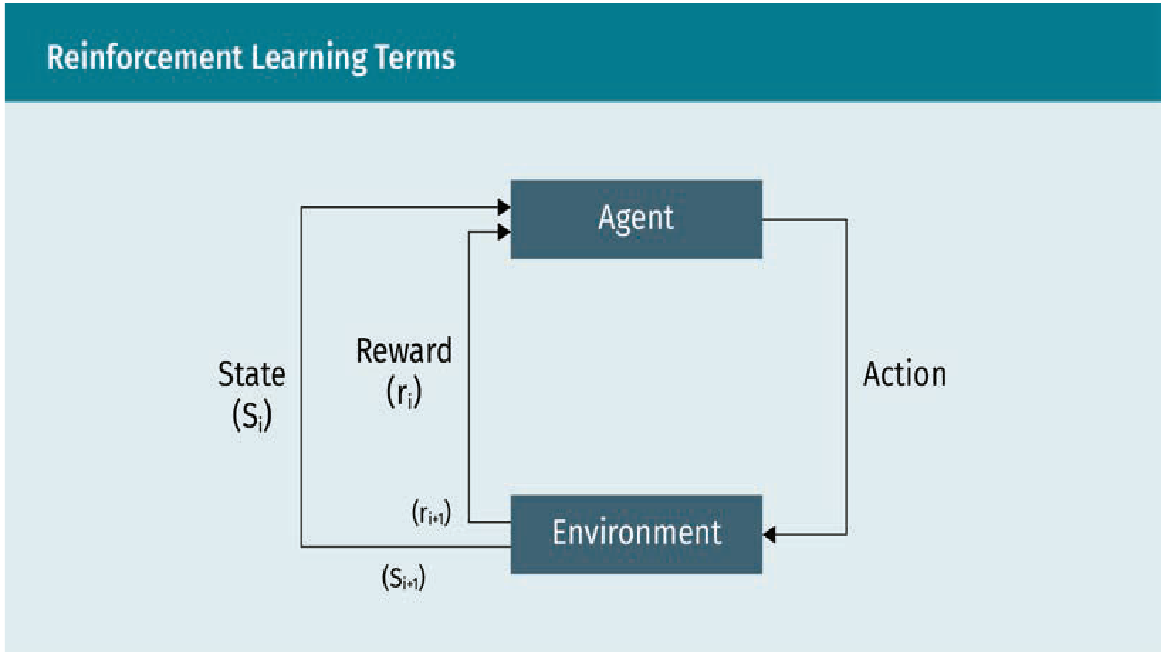


Two steps of Semi-Supervised Learning

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

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

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

WHAT ML CAN DO WELL* AND WHAT ML CANNOT DO WELL**

- Do well*
  - Automate straightforward tasks
  - Make predictions by learning input-output relationships
  - Personalize for individual users
- Cannot do well**
  - Understanding context
  - Determine causation
  - Explain why things happen
  - Determine the impact of interventions / find solutions

REVIEW STUDY GOALS

- Explain what is meant by machine learning.
- Know 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 1

INTRODUCTION TO MACHINE LEARNING

TRANSFER TASKS 1

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

Select an exciting machine learning project idea and discuss about it.

- Identify your area of interest.
- Quick research existing projects.
- Define your project goals.
- Evaluate available resources and knowledge.
- Discuss the project idea.

- Work in group or individual.

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 at ___ and have unit *standard deviation.*

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. Genetic Classification problems fall under the category of…

a) unsupervised learning

b) reinforcement learning

c) supervised learning

d) supervised and unsupervised learning