

**LECTURER: Nghia Duong-Trung**

# **MACHINE LEARNING**

## WHO I AM

- Name: Nghia Duong-Trung
- Senior Researcher @ The German Research Center for Artificial Intelligence
  - BMBF Projects: <https://milki-psy.de/>, <https://tech4comp.de/>
- Academic Teacher @ IU
  
- PostDoc @ Technische Universität Berlin, Germany
  - Project: <https://kiwi-biolab.de/>
- PhD in Machine Learning at The Information Systems and Machine Learning Lab ([ISMLL](#)), University of Hildesheim, Germany
  
- Profile: <https://sites.google.com/isml.de/duongtrungnghia/>

- Check attendance
- Attendance or partial attendance
- Excuse note (yes | no)
- Absence reason (yes | no)

- Course book: Machine Learning\_DLMD SML01, provided by IU, myStudies
- Reading list DLMD SML01, provided by IU, myStudies
- Additional teaching materials:

<https://github.com/duongtrung/IU-MachineLearning-DLMD SML01>

---

Introduction to Machine Learning

---

Clustering

---

Regression

---

Support Vector Machines

---

Decision Trees

---

Genetic Algorithms

1

2

3

4

5

6

**UNIT 1**

# **INTRODUCTION TO MACHINE LEARNING**

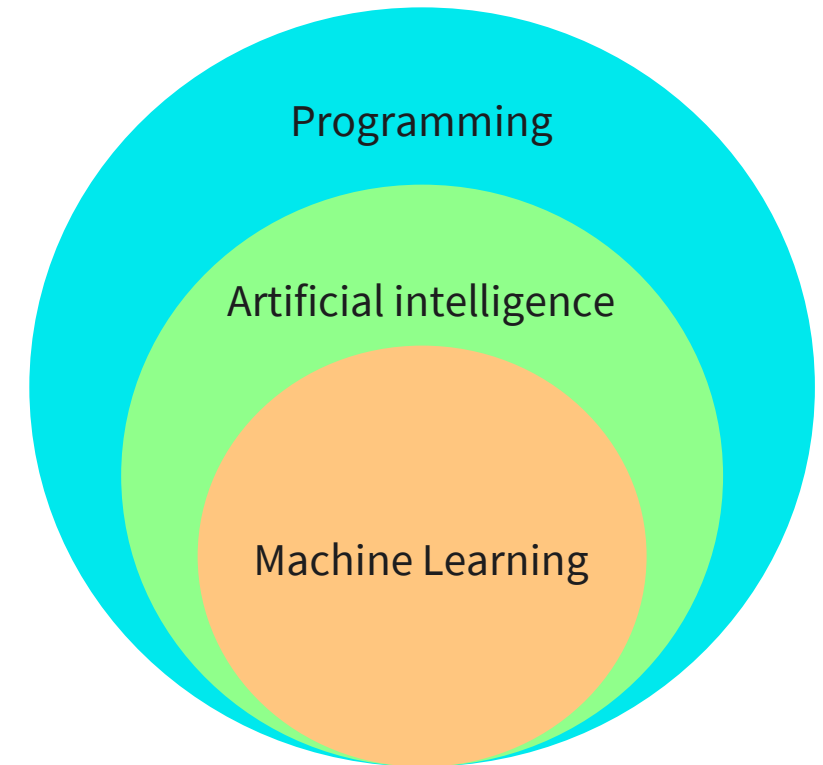
## STUDY GOALS



- 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



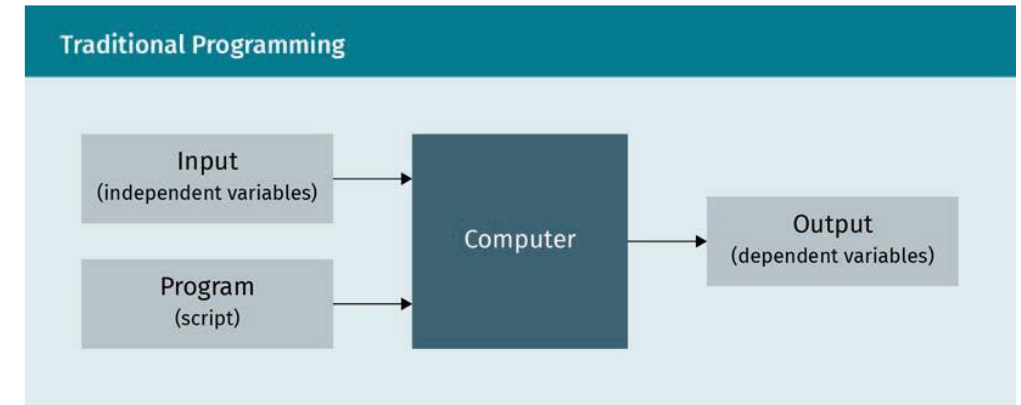
## MAKING DECISIONS WITH DATA

- 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

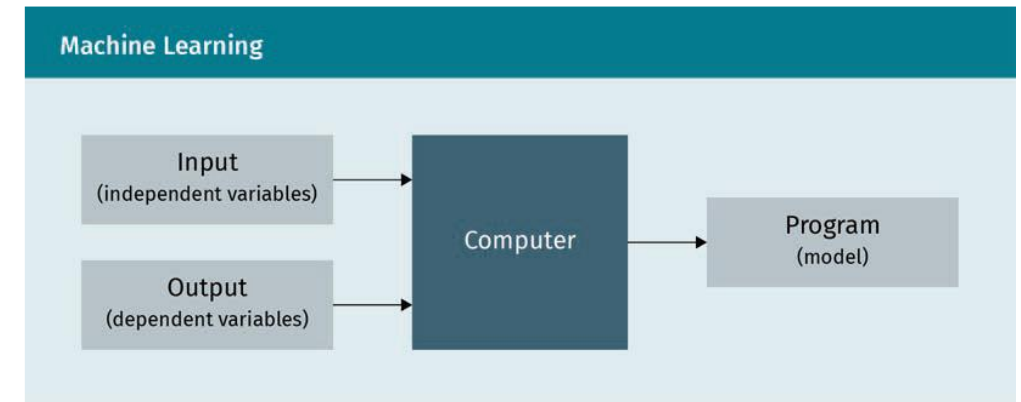


# Machine learning concepts

- **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



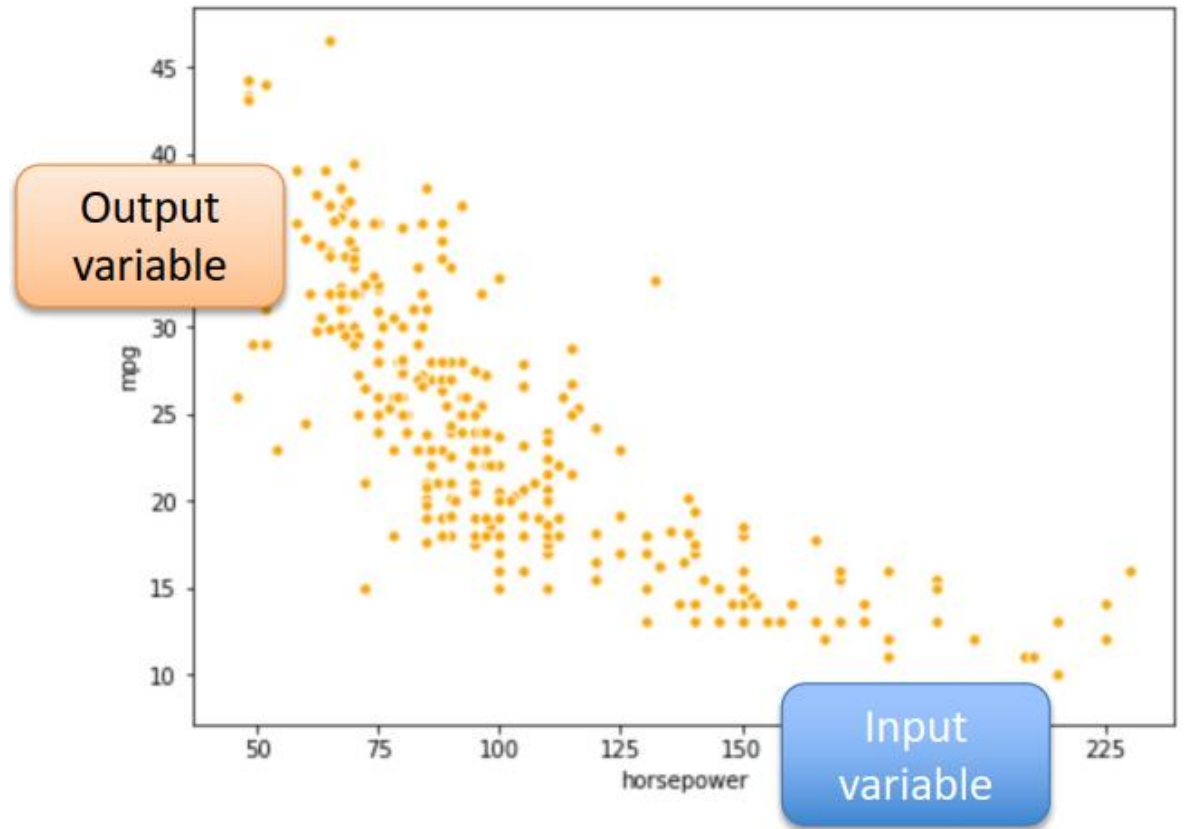
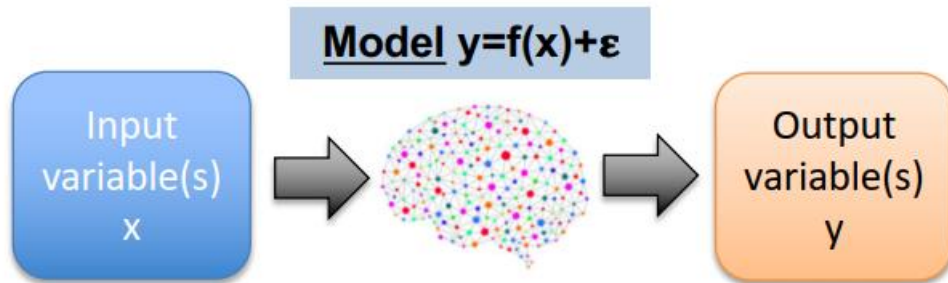
Traditional Programming



Machine learning

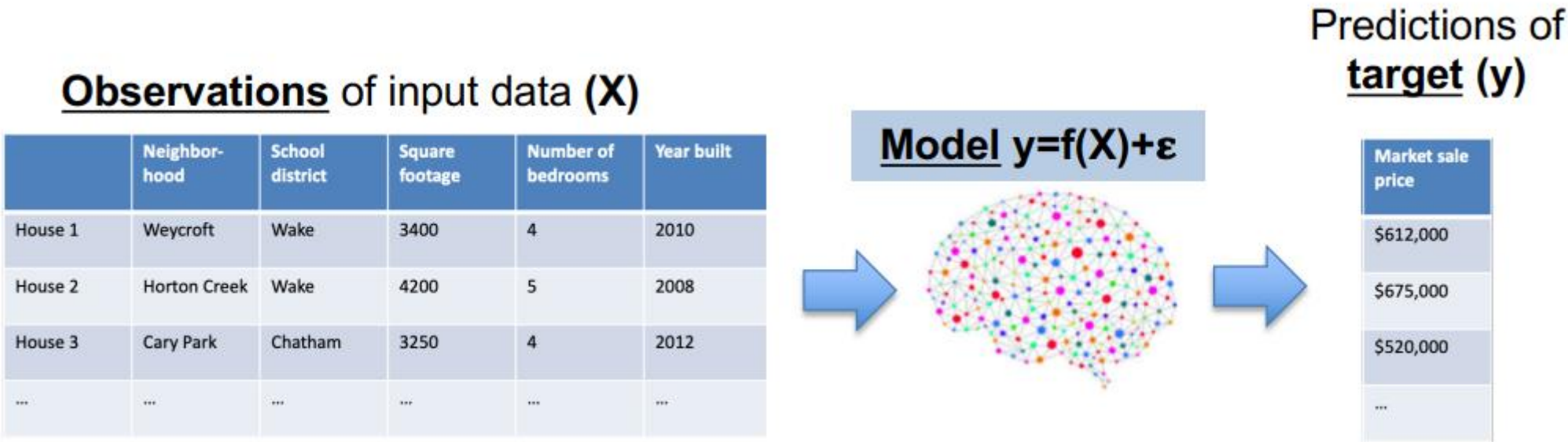
## WHAT IS A MODEL?

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



WHAT IS A MODEL?

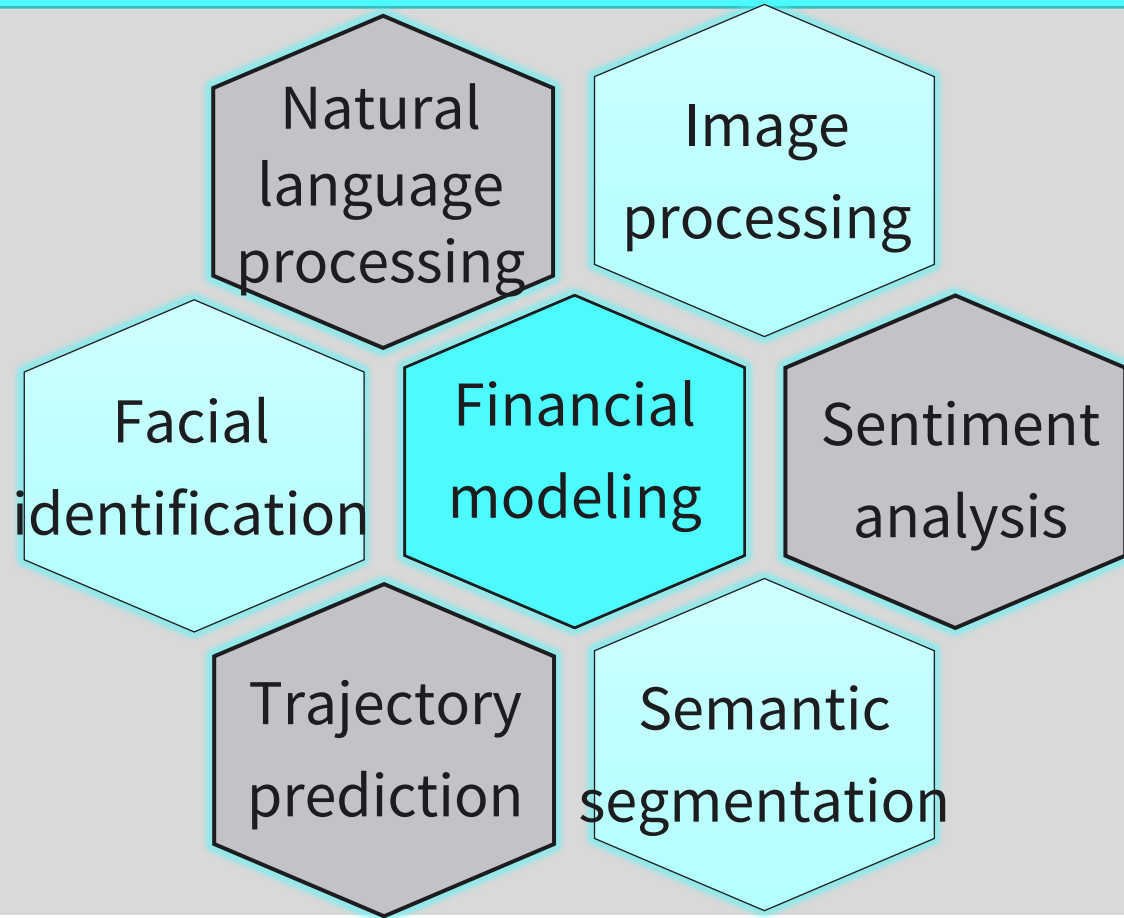
A model is an approximation of the relationship between two variables.



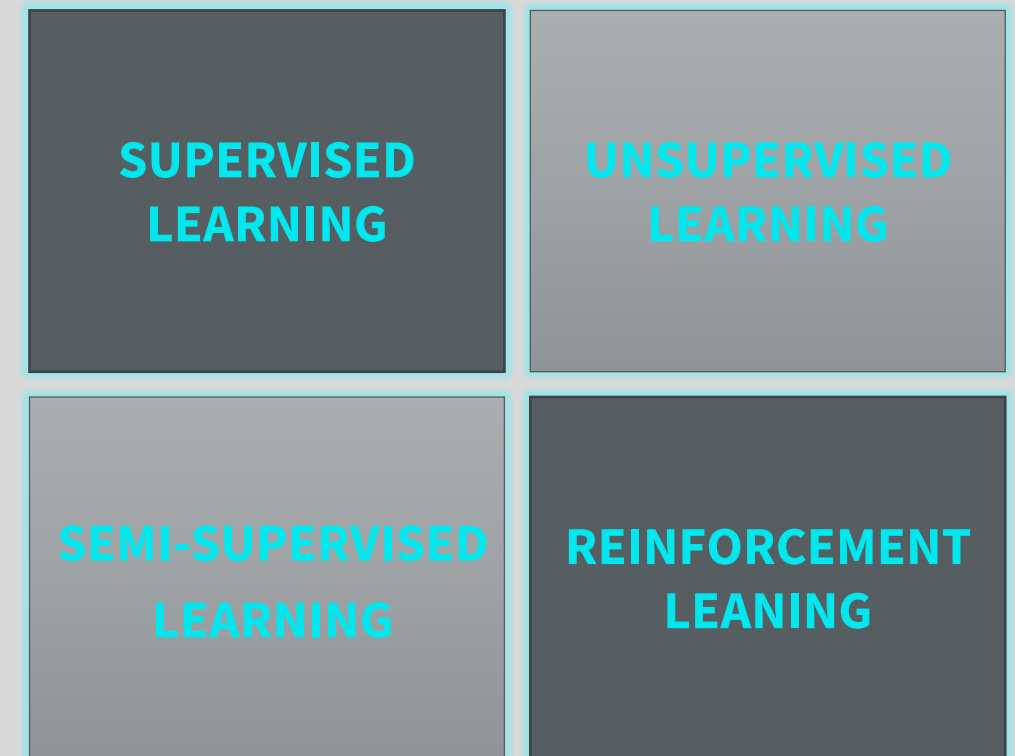
## BUILDING A MODEL

- 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 historical data:
  - Algorithm & hyperparameters provide overall model form
  - Learn values for the model which minimize loss function

### Applications of ML

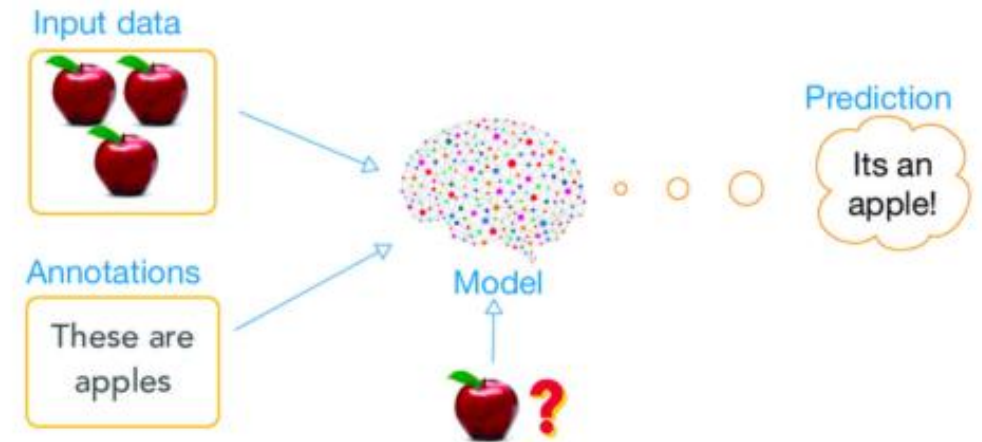


### ML Paradigms



# 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

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

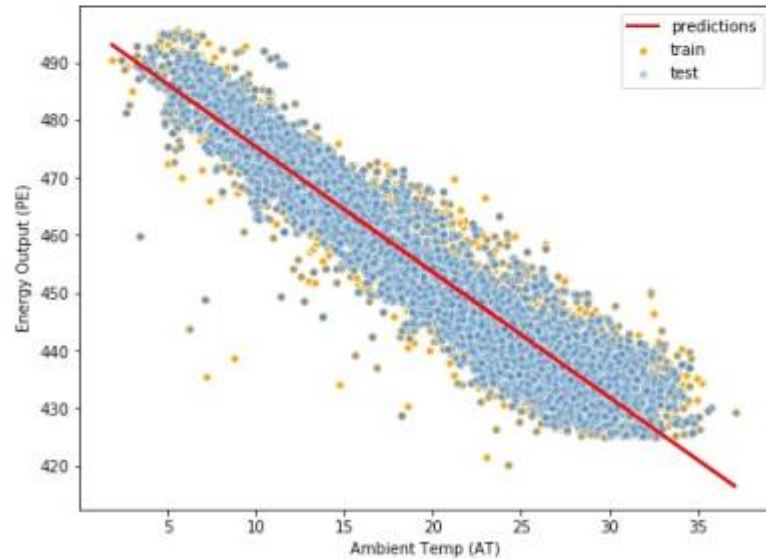
Source of the table: Zöller (2022).



## REGRESSION VS CLASSIFICATION

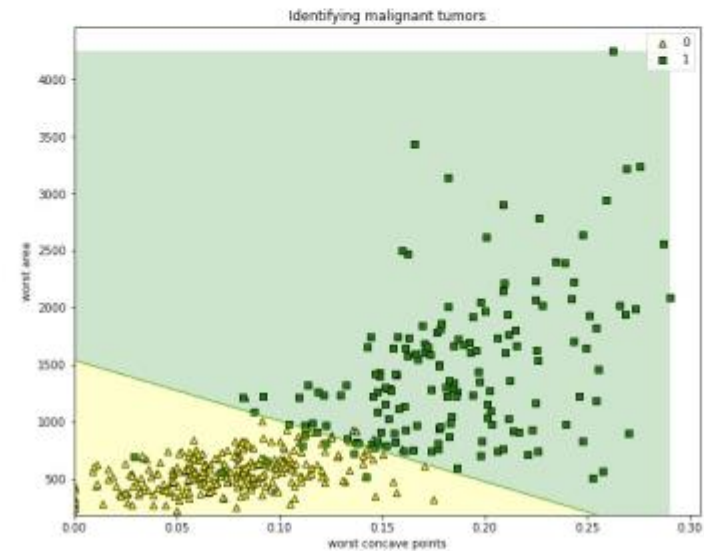
### Regression

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



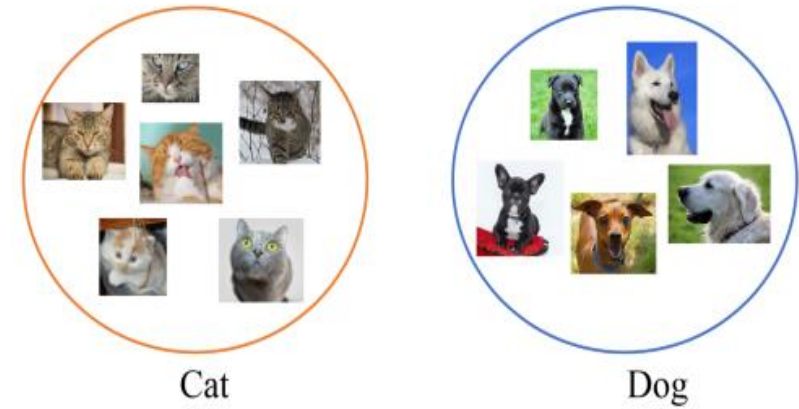
### Classification

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



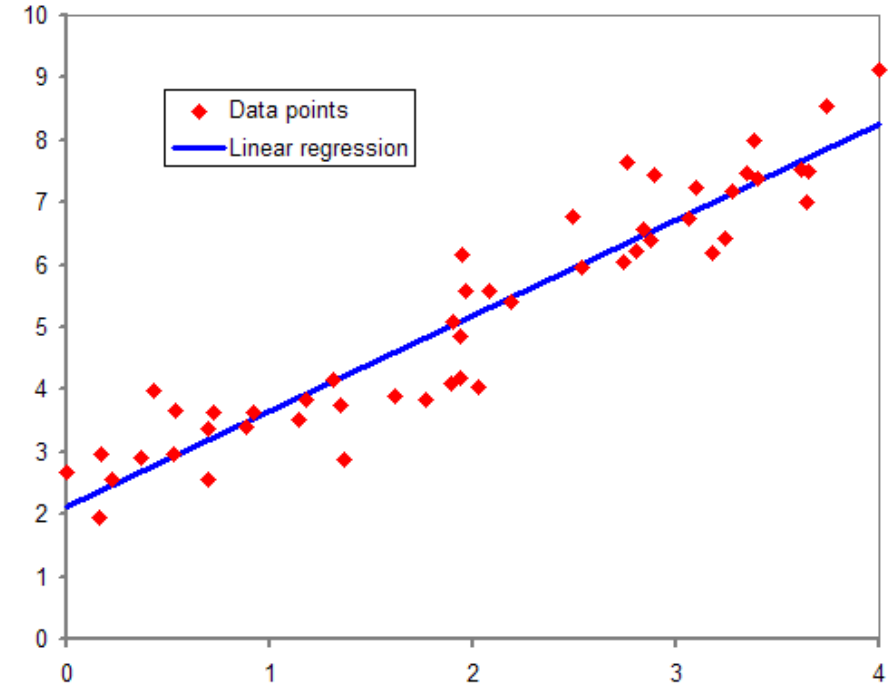
Dog and Cat classification

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

Confusion matrix

# Regression

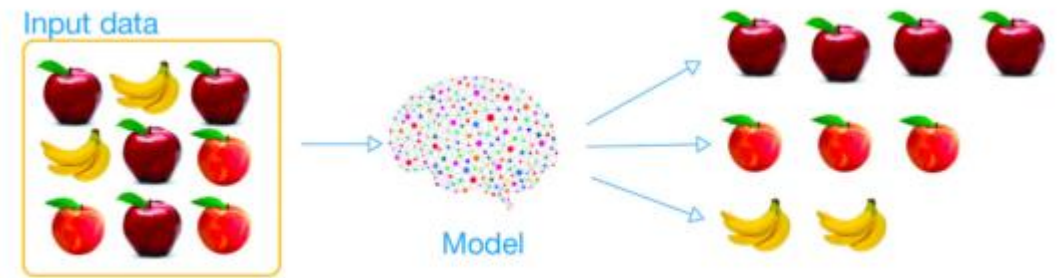
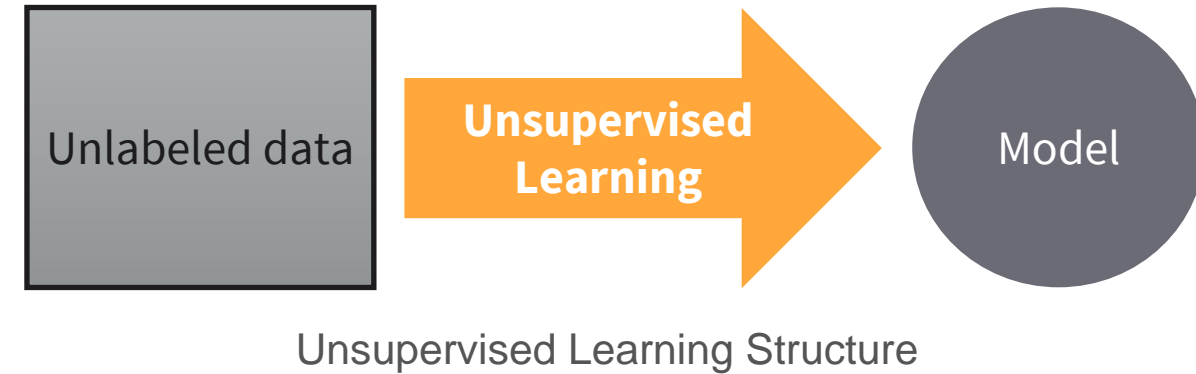
- **Objective:** develop a ML model to **relate** the inputs  $x$  to the outputs  $y$  and **predict** the output **values**  $\hat{y}$  for new inputs
- Evaluation **metrics**:
  - Mean Square Error:  $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
  - Root Mean Square Error:  $RMSE = \sqrt{MSE}$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



An example of Regression

# 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

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

# Semi-Supervised Learning

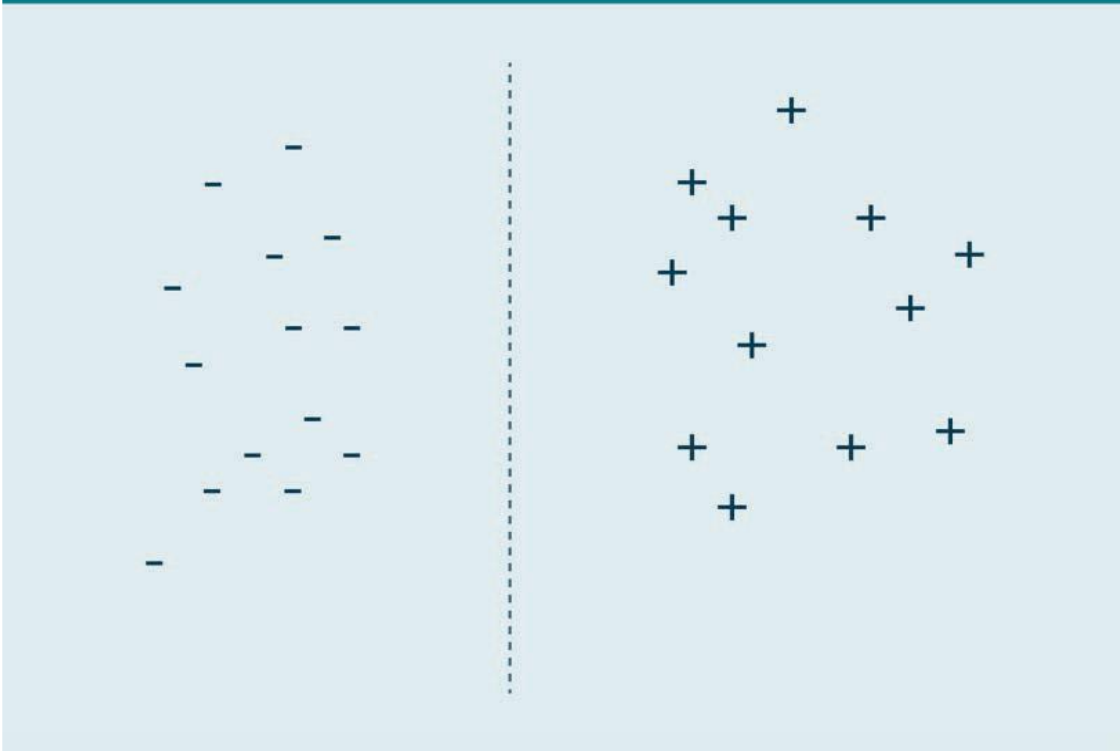
- **Dataset:** a collection of both **labeled** samples (a small portion of data), and **unlabeled** samples (lots of data)
- **Objective:** mix of supervised and unsupervised learning to combine the properties of both.
- **2 steps:**
  - Supervised learning is performed on few labeled data
  - Unsupervised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



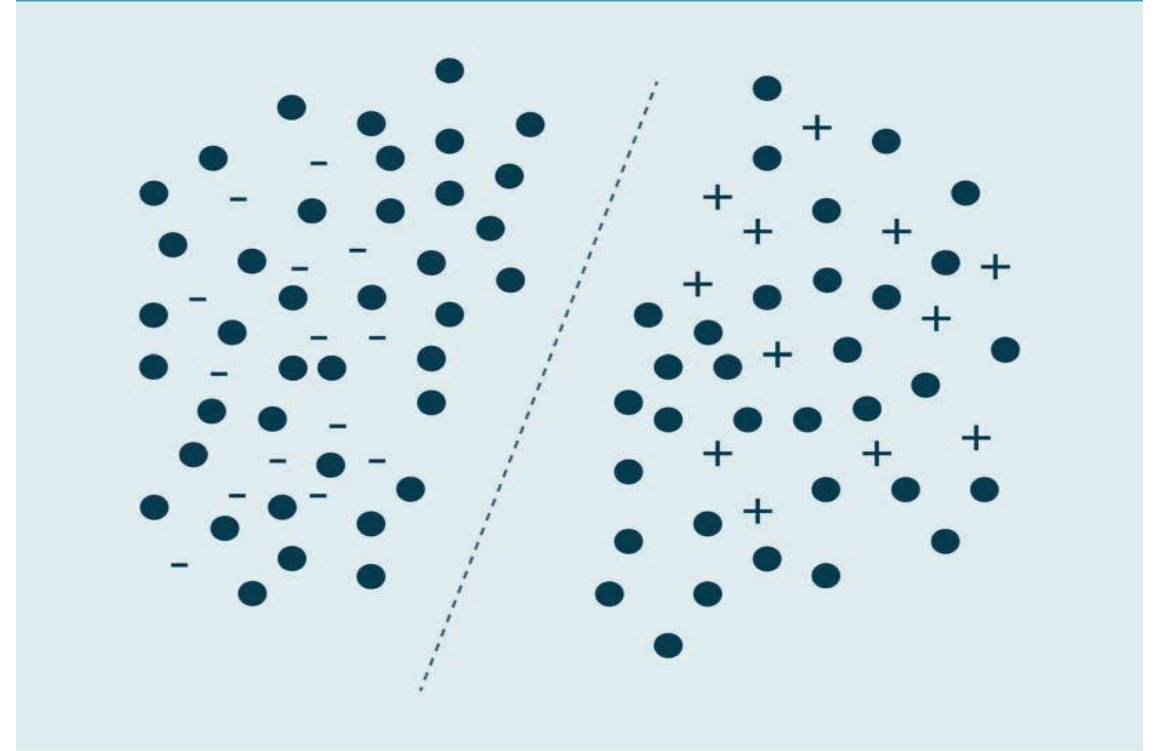
Semi-Supervised Learning Structure

# Semi-supervised Learning

Semi-Supervised Learning (Classification Step)



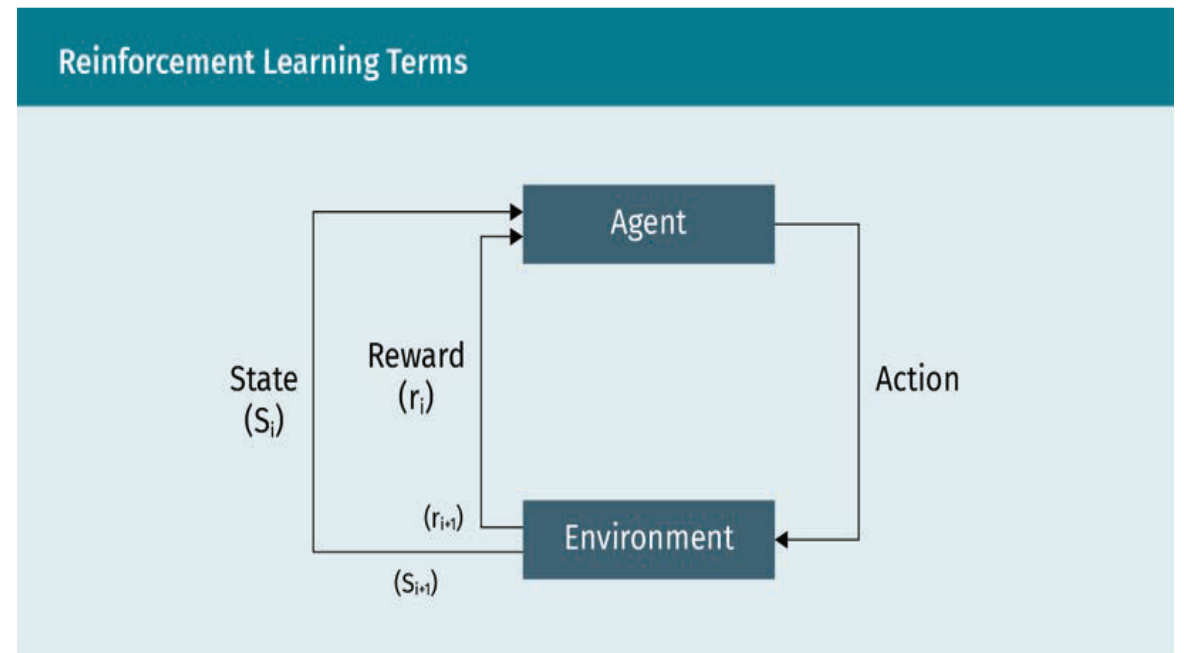
Semi-Supervised Learning (Clustering Step)



Two steps of Semi-Supervised Learning

# 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



# 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 ( $\pi$ )	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		

## 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



- 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

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 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

## 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 (Feb. 28, 2023). *In Wikipedia Commons, the free media repository*. Retrieved, February 28, 2023, from [https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)

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

### Images:

File:Normdist\_regression.png. (2023, January 22). *In Wikimedia Commons, the free media repository*. Retrieved, January 29, 2023, from [https://en.wikipedia.org/wiki/Regression\\_analysis](https://en.wikipedia.org/wiki/Regression_analysis)

File:Dog and Cat Classification.png. (2023, January). Open sources. *Kaggle Competition* . Retrieved, Feb. 28, 2023, from <https://www.kaggle.com/code/sasakitetsuya/dog-and-cat-classification-by-mobilenet>

Zöller (2022, p. 12-13).

Zöller (2022, p. 15).

Zöller (2022, p. 22-23).

Zöller (2022, p. 24).

Zöller (2022, p. 25).

© 2022 IU Internationale Hochschule GmbH

This content is protected by copyright. All rights reserved.

This content may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.