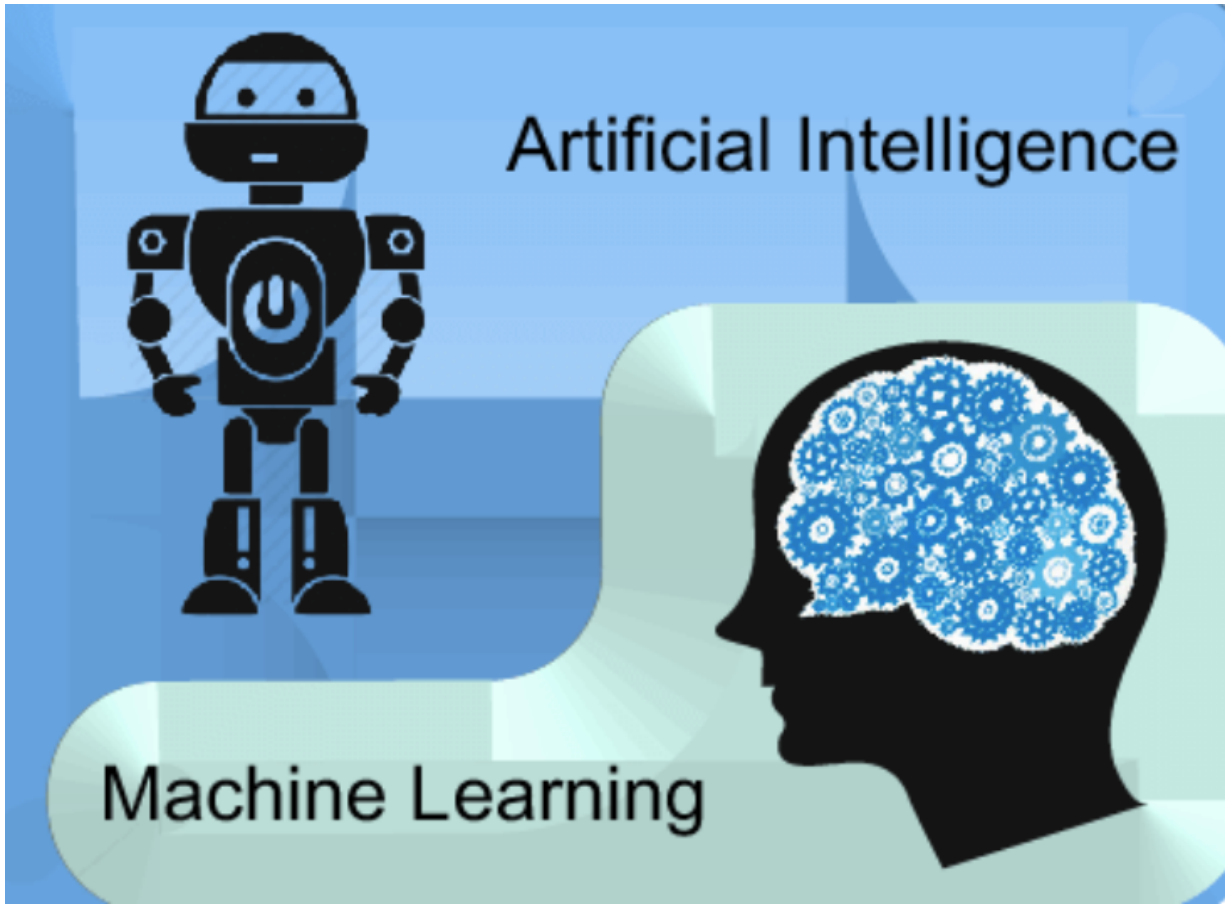
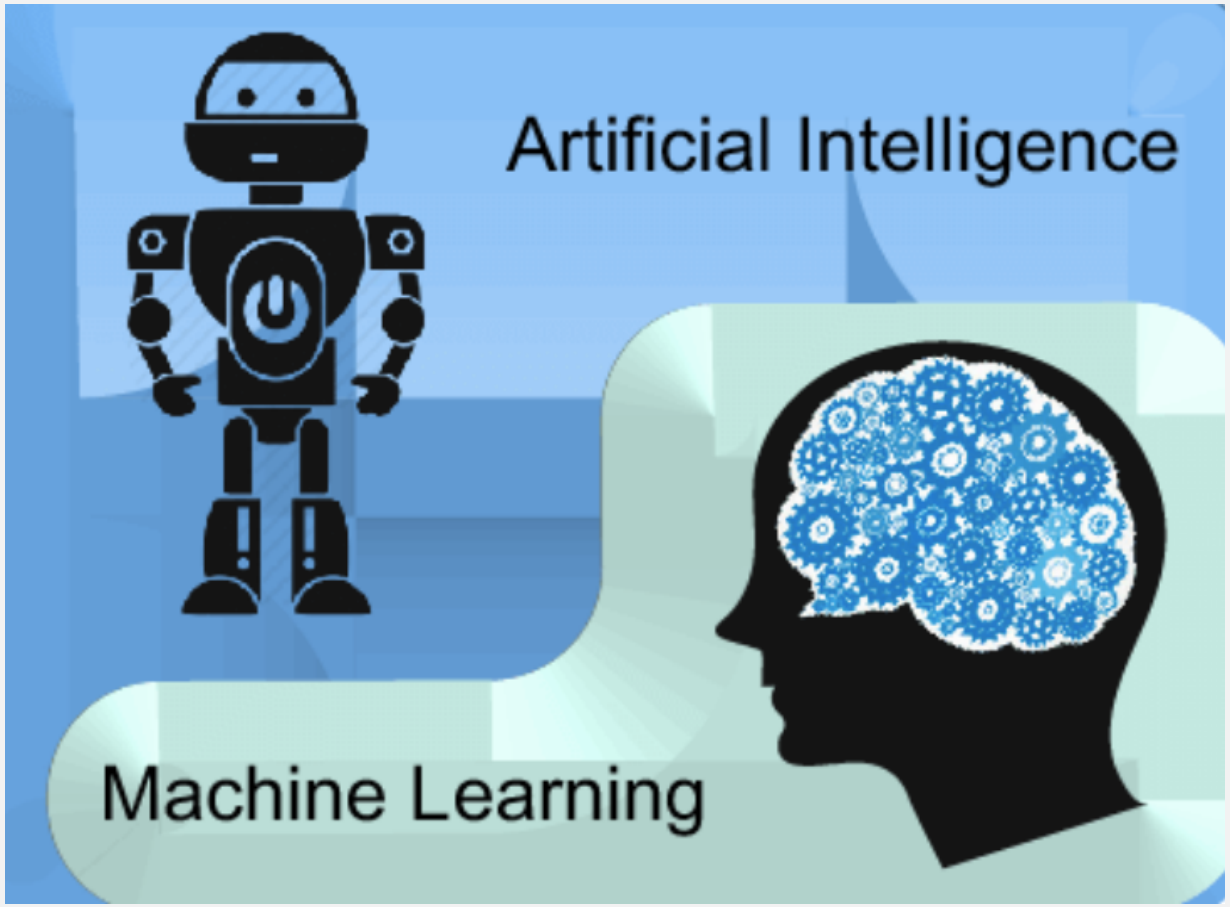
INTRODUCTION TO TOPIC MODELLING

A Machine Learning method for an automatic analyse of text data.



WHAT IS MACHINE LEARNING ?

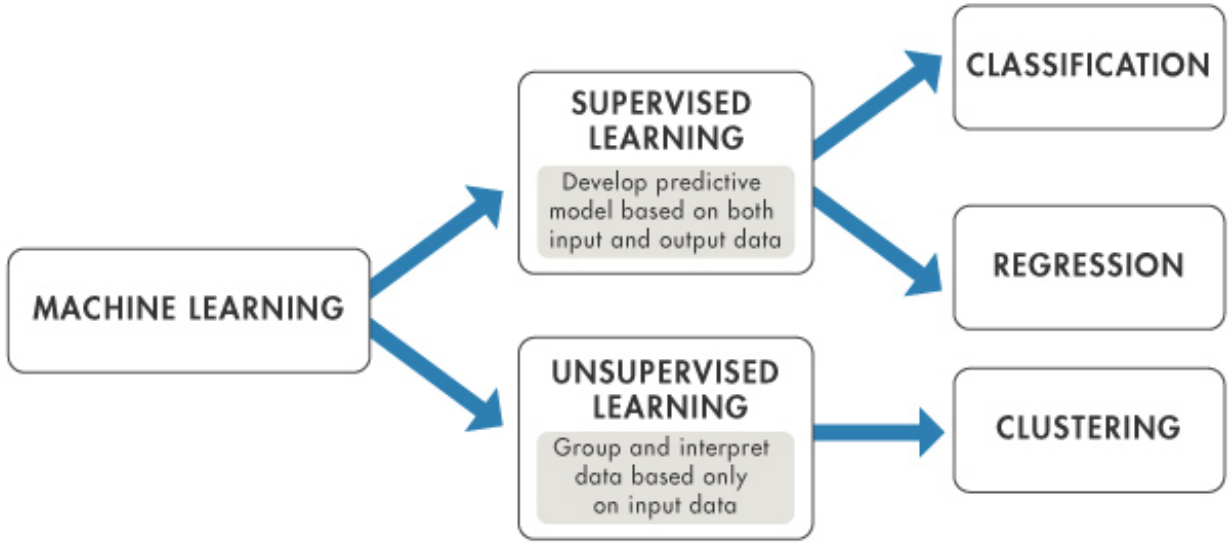
- *Artificial Intelligence (AI)*:
  - a field of computer science
  - a computer system that can mimic human intelligence.

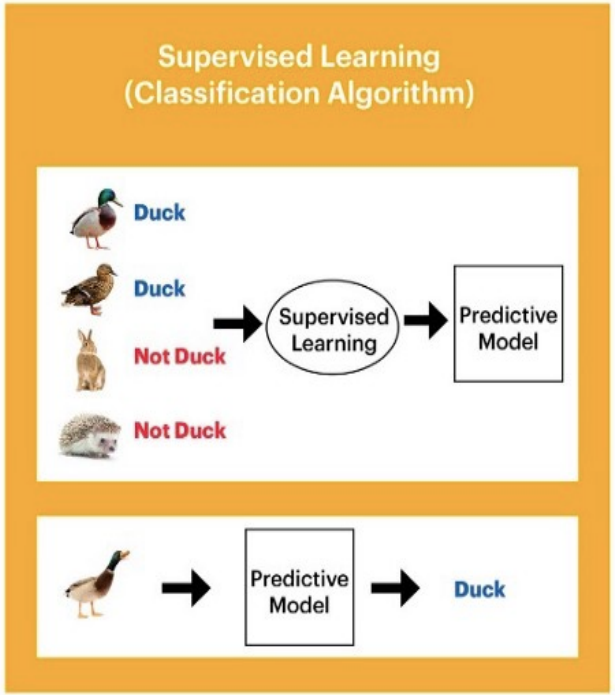


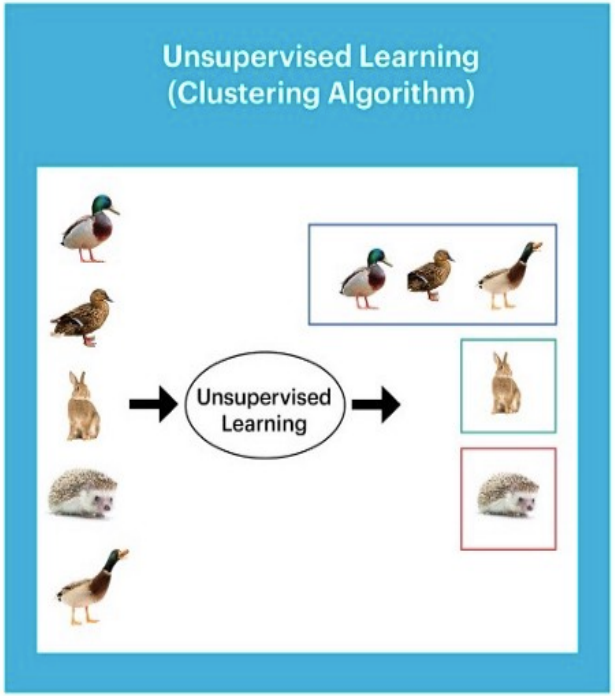
WHAT IS MACHINE LEARNING ?

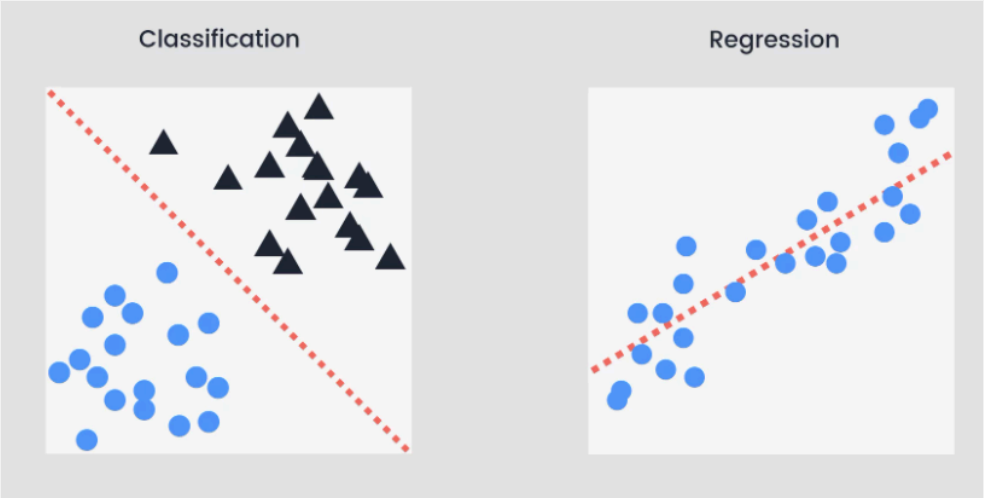
- *Artificial Intelligence (AI):*
  - a field of computer science
  - a computer system that can mimic human intelligence.
- *Machine learning (ML):*
  - a subfield of AI that is about extracting knowledge from the data,
  - allows machines **to learn without first having been programmed specifically for this purpose**.

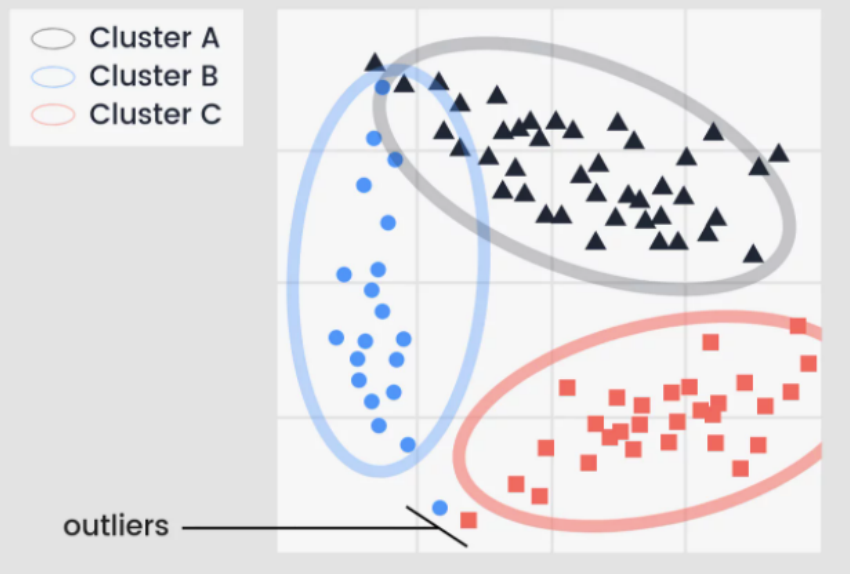
SUPERVISED VS UNSUPERVISED MACHINE LEARNING











SUPERVISED VS UNSUPERVISED MACHINE LEARNING

SUPERVISED: Mapping of $Y_{train} = f(X_{train})$

Classification: for categorical values: Duck/Not Duck

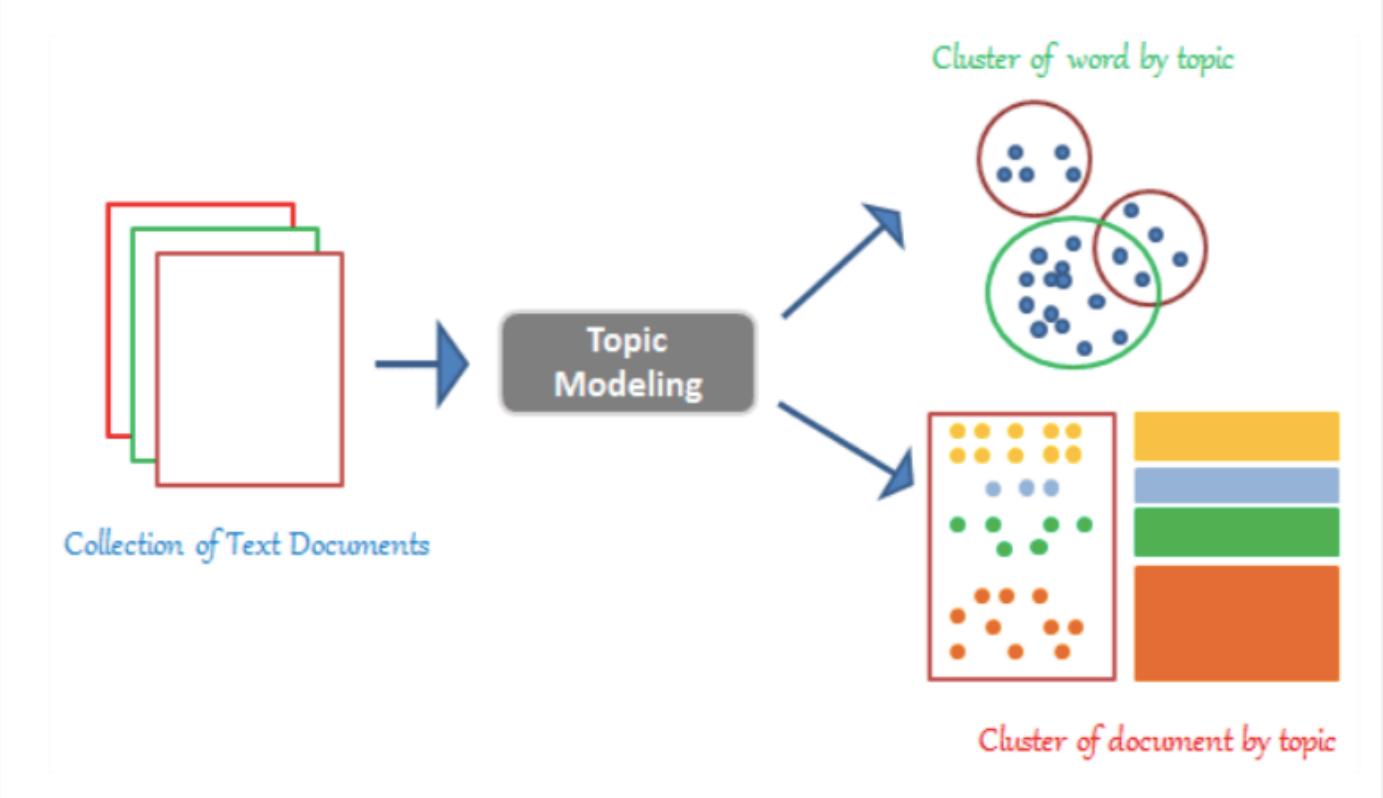
Regression: for continuous value, for example, the prediction of pressure, temperature, etc.

UNSUPERVISED: only input data required

Clustering: used to discover groupings found in the input data.

TOPIC MODELLING: UNSUPERVISED APPROACH

- *Topic Modeling:* the process of dividing a corpus of documents in two:
  - A list of the topics covered by the documents in the corpus
  - Several sets of documents from the corpus grouped by the topics they cover.



LATENT DIRICHLET ALLOCATION (LDA)

- **Hypothesis:**
  - **Mixture hypothesis**: every document comprises a statistical mixture of topics.
  - **Distributional hypothesis**: similar topics make use of similar words.
- **Purpose** of LDA :
  - Assign topics to arrangements of words.
  - Figure out which topics are present in the documents of the corpus and how strong that presence is.