CS-EJ3211 Machine Learning with Python

Session 1 - Components of Machine Learning

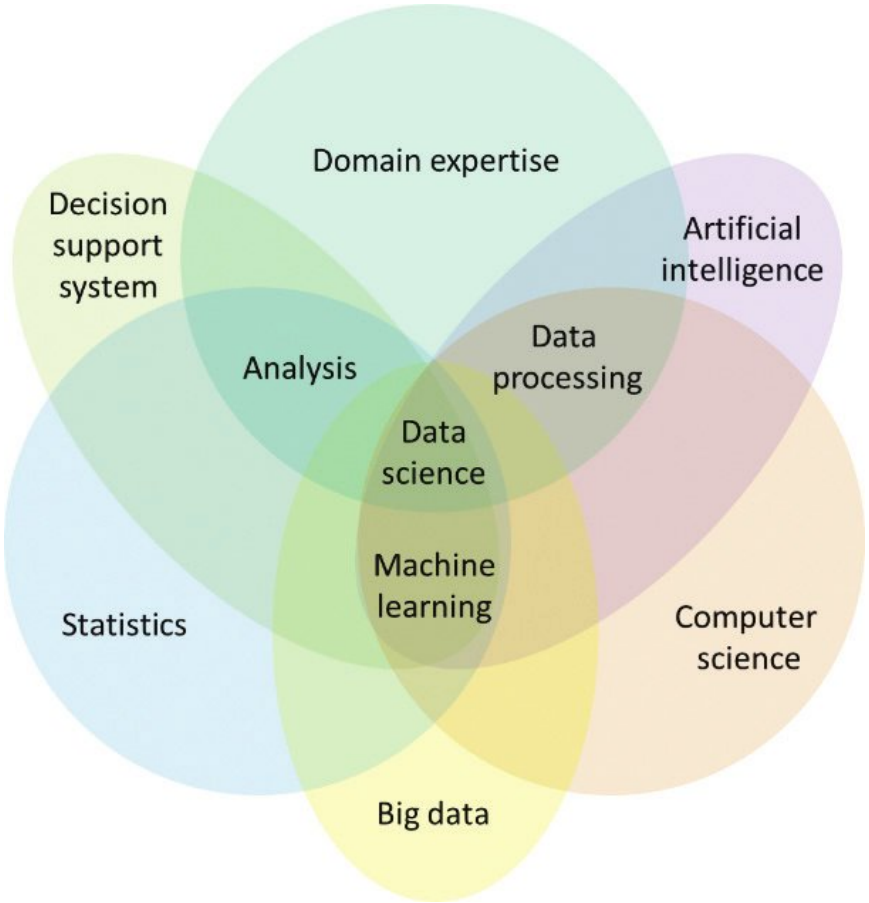
Shamsi Abdurakhmanova

Aalto University
FITech

20.01.22

Machine Learning, Deep Learning, Artificial Intelligence, Statistics, Data science, Statistical learning, Data mining, ...

Machine Learning, Deep Learning, Artificial Intelligence, Statistics, Data science, Statistical learning, Data mining, ...



DOI: 10.5772/intechopen.81872

Machine Learning - Relation to other fields

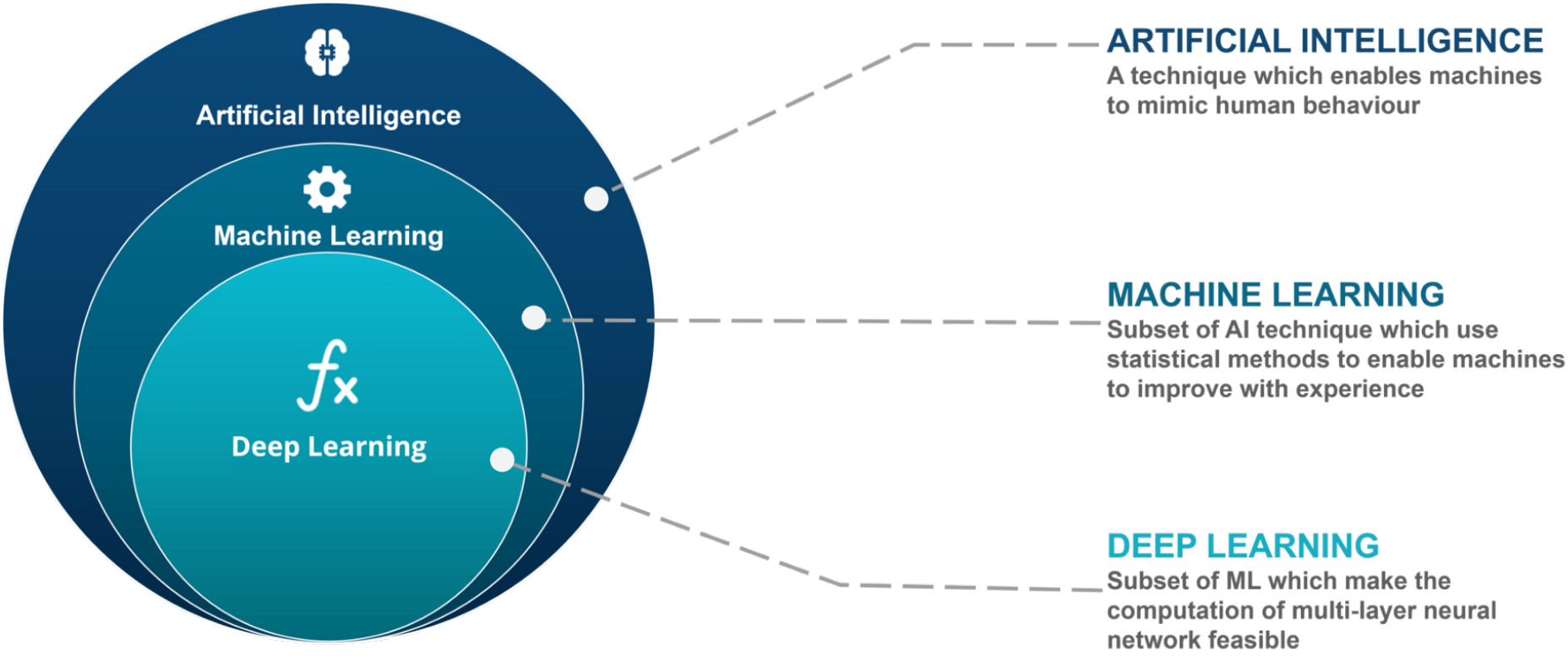


image source

Machine Learning - Definition

"Machine learning is the study of computer algorithms that improve automatically through experience."

— Wiki

"The goal of machine learning is to design general- purpose methodologies to extract valuable patterns from data, ideally without much domain-specific expertise."

— Mathematics for Machine Learning book

"Machine learning is a field of computer science that aims to teach computers how to learn and act without being explicitly programmed."

— deepai.org

Machine Learning - Applications

- Natural Language Processing
- Insurance Claim Analysis
- Bioinformatics and Medical Diagnosis
- Image Processing and Pattern Recognition
- Search Engines
- Financial Market Analysis

from deepai.org

Problem formulation

Physics: "A highway patrol plane is flying 1 mile above a long, straight road, with constant ground speed of 120 m.p.h. Using radar, the pilot detects a car whose distance from the plane is 1.5 miles and decreasing at a rate of 136 m.p.h. How fast is the car traveling along the highway?"

Problem formulation:

- **Given Variables**: $|\frac{dp}{dt}| = 120$, $\frac{dy}{dt} = -136$, $y(t = 0) = 1.5$
- **Find Variables**: $|\frac{dc}{dt}|$
- **Relations**: $y^2 = x^2 + 1$, $\frac{dx}{dt} = \frac{dc}{dt} + \frac{dp}{dt}$

Components of ML

Goal: Formulate a real-life problem as a Machine Learning problem.

Components of ML

Goal: Formulate a real-life problem as a Machine Learning problem.

Solution: Decompose a problem into 3 components

- **Data**
- **Model (Hypothesis space)**
- **Loss**.

Data

Data is a collection of individual data points that are characterized by features and labels.

A data point is any object that conveys information. Data points might be students, radio signals, trees, forests, images, RVs, real numbers or proteins. We characterize data points using two types of properties: features and labels.

Features are properties of a data point that can be measured or computed in an automated fashion.

The label of a data point represents a higher-level facts or quantities of interest. In contrast to features, determining the label of a data point typically requires human experts (domain experts). Roughly speaking, ML aims at predicting the label of a data point based solely on its features.

Data - Examples

Data points: housing data excel file, text, genome sequence, image, speech

Data - Examples

Data points: housing data excel file, text, genome sequence, image, speech

Features: number of rooms, word "cat" count in the text, frequency of AATCAGTT motif, pixel values, frequency spectrum

Labels: price, article topic, cell type, identify objects on image, emotional state

Linnerud dataset - Features or Labels?

- Chin-ups
- Sit-ups
- Jumps

- Weight
- Waist
- Pulse

The hypothesis space

The hypothesis space of a ML method is a subset of all possible maps from the feature space to label space. The design choice of the hypothesis space should take into account available computational resources and statistical aspects.

$$h : \mathcal{X} \to \mathcal{Y}$$

The hypothesis space

Choosing the right estimator with sklearn.

Loss

Loss function $\mathcal{L}(y, \hat{y})$ measures the quality of a hypothesis map.

Examples:

- MSE
- MAE
- Logistic Loss, Cross Entropy
- Hinge Loss
- 0/1 Loss

Problem formulation in ML

Machine Learning: "Given an image of a histological analyses of a tissue, how likely that this sample is malignant tumor?"

Problem formulation in ML

Machine Learning: "Given an image of a histological analyses of a tissue, how likely that this sample is malignant tumor?"

Problem formulation:

- **Data**:
  - Data point - image of a tissue sample
  - Features - pixels
  - Labels - {cancer, no cancer}
- **Model**: Logistic regression
- **Loss**: Logistic loss

Using ML models with sklearn

- Import model from sklearn
- Instantiate a class
- Fit data with .fit()
- Predict with .predict()