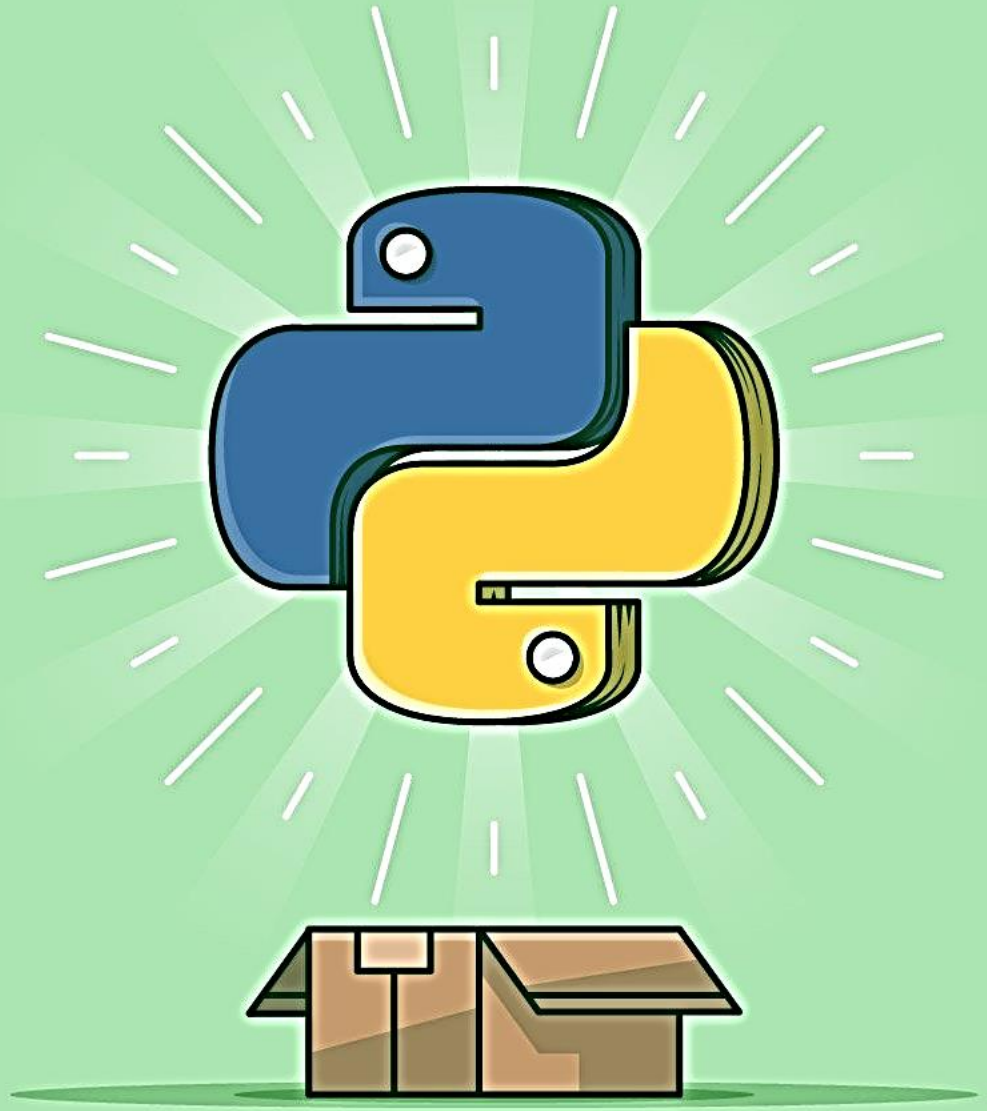
ISLab Python Course

Special Session:
A Brief Introduction to Machine Learning



Presenters:

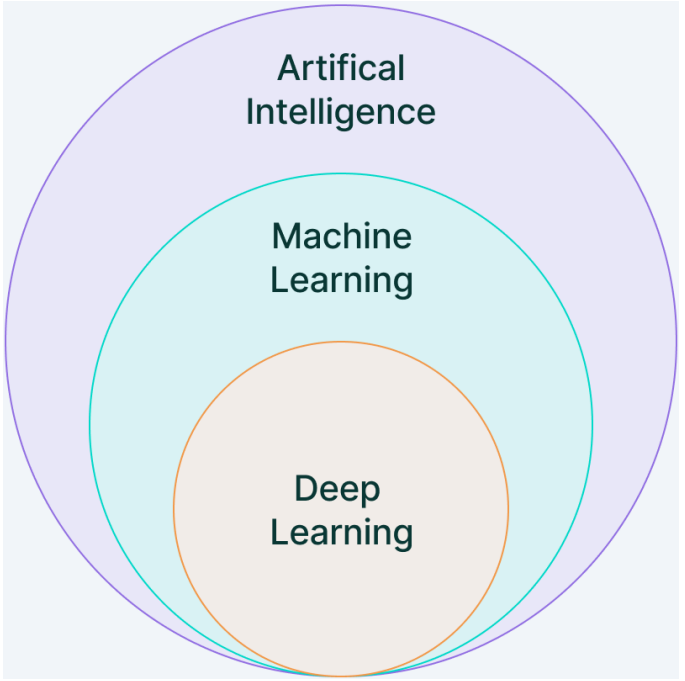
Shahrzad Shashaani **Hamed Homaei Rad** **Saeed Samimi**

Summer 2023

K.N.Toosi University of Technology

Machine learning (ML)

- A subset of Artificial Intelligence (AI)
- Enables computers to **self-learn** from training data to improve over time, without being explicitly programmed
- ML algorithms can detect patterns in input data and learn from them to make their own predictions



ML vs. Traditional Programming

Traditional Programming

- A developer writes a series of directions instructing a computer to transform input data into a desired output
- Instructions are mainly based on an IF-THEN structure:
  - when certain conditions are met, the program executes a specific action

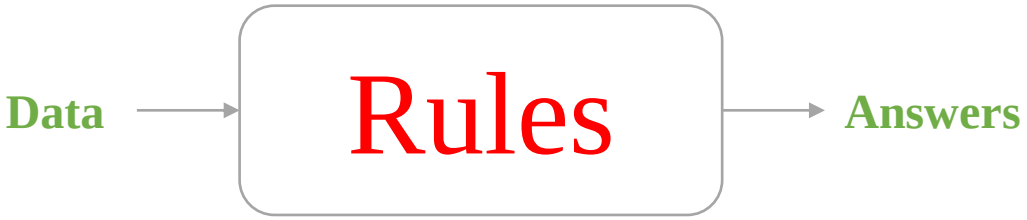
Machine Learning

- Is an automated process that enables machines to solve problems with little or no human input and instructions, and take actions based on past observations

ML vs. Traditional Programming

- **Traditional Programming**

- **Machine Learning**



Green: Known
Red: Unknown

Machine Learning Process

Data

- ML algorithms use data to learn patterns and relationships between input variables and target outputs, which can then be used for prediction or classification tasks
- Data is a crucial component in the field of ML
- Data refers to the set of observations or measurements that can be used to train a ML model
- The quality and quantity of data available for training and testing play a significant role in determining the performance of a ML model

Data Types

Numerical Data

- Continues
  - 12.3
  - 25.8
- Discrete
  - 12
  - 26

Categorical Data

- Nominal
  - Gender
  - Transportation
- Ordinal
  - Education Level
  - Income Level

Time Series Data

a sequence of data points indexed in time order

Text Data

information that is stored and written in a text format

Splitting Data in ML

- Training Data
  - The part of data that is used to train the model
  - This is the data that the model actually sees and learns from
- Validation Data
  - The part of data that is used to do a frequent evaluation of the model
  - This part of data plays its part when the model is actually training
- Testing Data
  - Once the model is completely trained, testing data provides an unbiased evaluation
  - This is how the trained model is evaluated to see how much it has learned from the experiences fed in as training data
  - This is the data that the model does not see during the training process
  - We should never use test data during training stage for model evaluation

Data Preprocessing

- Data cleaning
  - Identification of errors and making corrections or improvements to those errors
- Feature Selection
  - identifying the most important or relevant input data variables for the model
- Data Transforms
  - Converting raw data into a well-suitable format for the model
- Missing data
  - Filling missing data or incomplete records (sometimes records contain empty cells, values (e.g., NULL or N/A), or a specific character, such as a question mark, etc.)
- Outliers or Anomalies
  - ML algorithms are sensitive to the range and distribution of values when data comes from unknown sources. These values can spoil the entire ML training system and the performance of the model. Hence, it is essential to detect these outliers or anomalies through techniques such as visualization technique.

Data Leakage

- Leakage Through Features:
  - Model trained simultaneously with train dataset and test dataset
- Target Leakage:
  - Model trained through information about the target
- Time-Based Leakage:
  - Model trained through information from the future
- Data Transformation Leakage:
  - Data transformation using the entire dataset
- Preventing Data Leakage:
  - Proper Train-Validation-Test Split
  - Feature Engineering
  - Time series
  - Data Transformation
  - Cross-Validation

Why is Data Preparation important?

- Providing reliable prediction outcomes
- Identifying data issues or errors and significantly reduces the chances of errors
- Increasing decision-making capability
- It increases model performance

Types of ML Algorithms

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning
- Deep Learning (DL)

Supervised Learning

- This is the most common and popular approach to machine learning
- Each training sample includes an input and a desired output
- Supervised learning models learn to make predictions based on labeled training data by analyzing inputs and making an inference to determine the appropriate output for unseen data

Unsupervised Learning

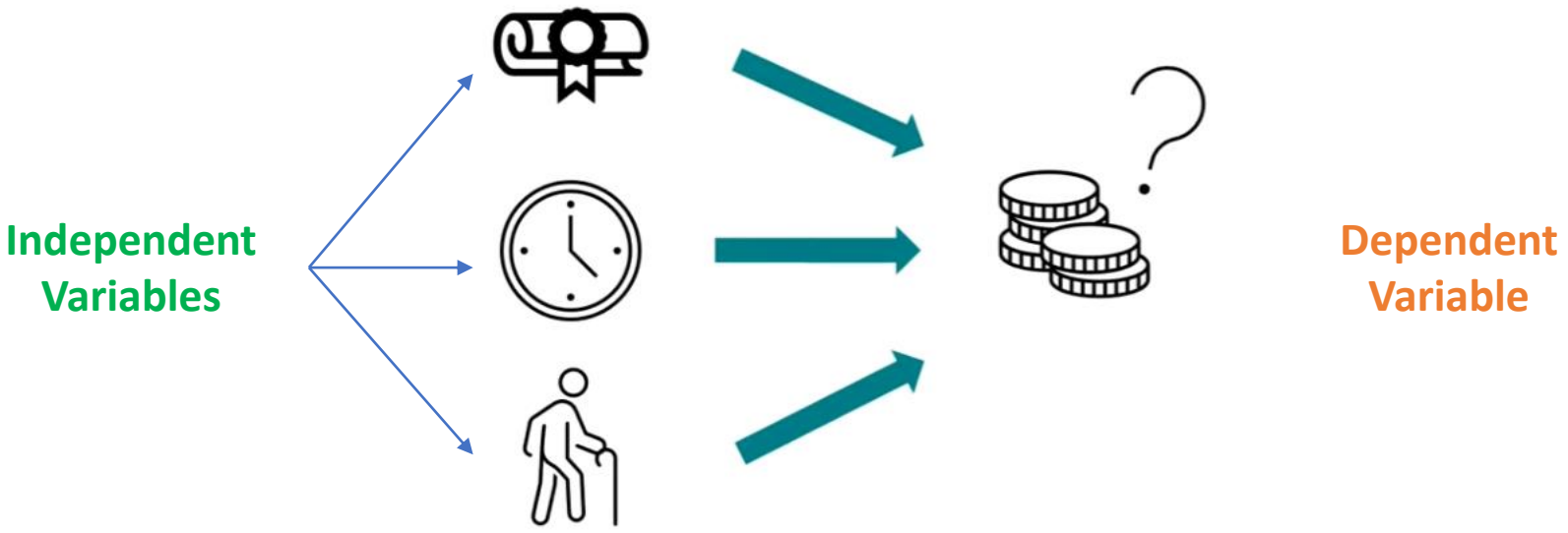
- Unsupervised learning algorithms uncover insights and relationships in unlabeled data
- Models are fed input data but the desired outcomes (labels) are unknown
- These algorithms discover hidden patterns or data groupings by finding patterns on their own

ML Algorithms

- Linear regression
- Logistic regression
- Decision tree
- Support Vector Machine (SVM) algorithm
- Naive Bayes algorithm
- K-Nearest Neighbor (KNN) algorithm
- K-means
- Random forest algorithm
- Artificial Neural Network (ANN)

Regression Analysis

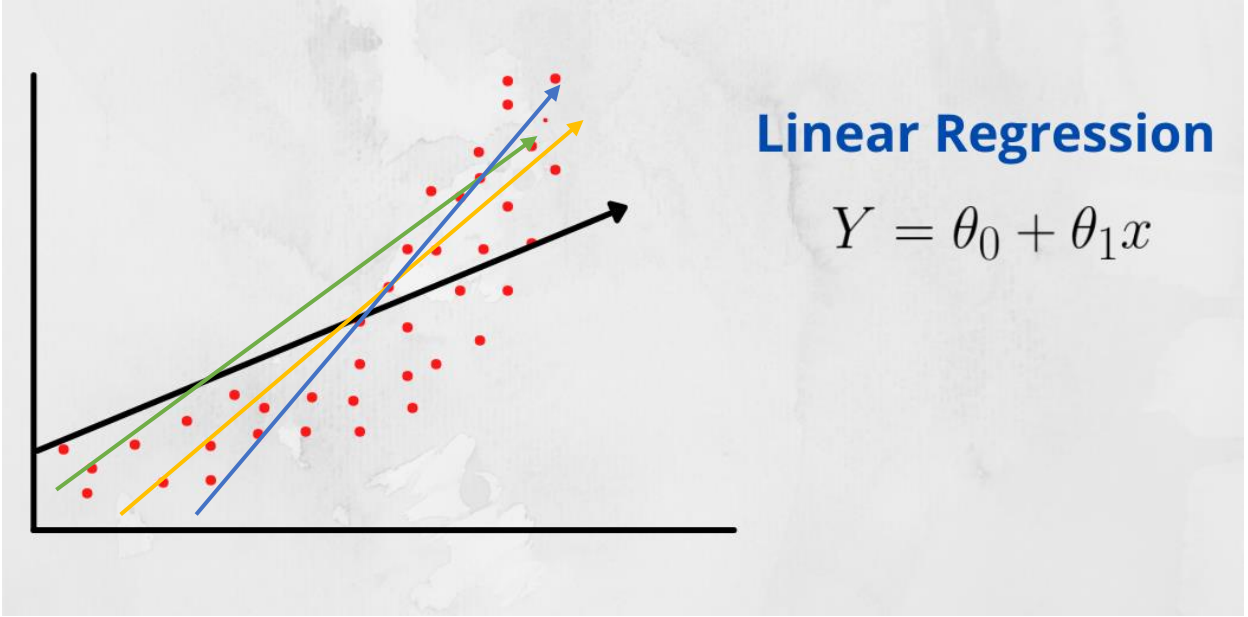
- A regression analysis makes it possible to infer or predict a variable on the basis of one or more other variables



Linear Regression

- In statistics, a regression model is linear when all terms in the model are one of the following:
  - The constant
  - A parameter multiplied by an independent variable (IV)

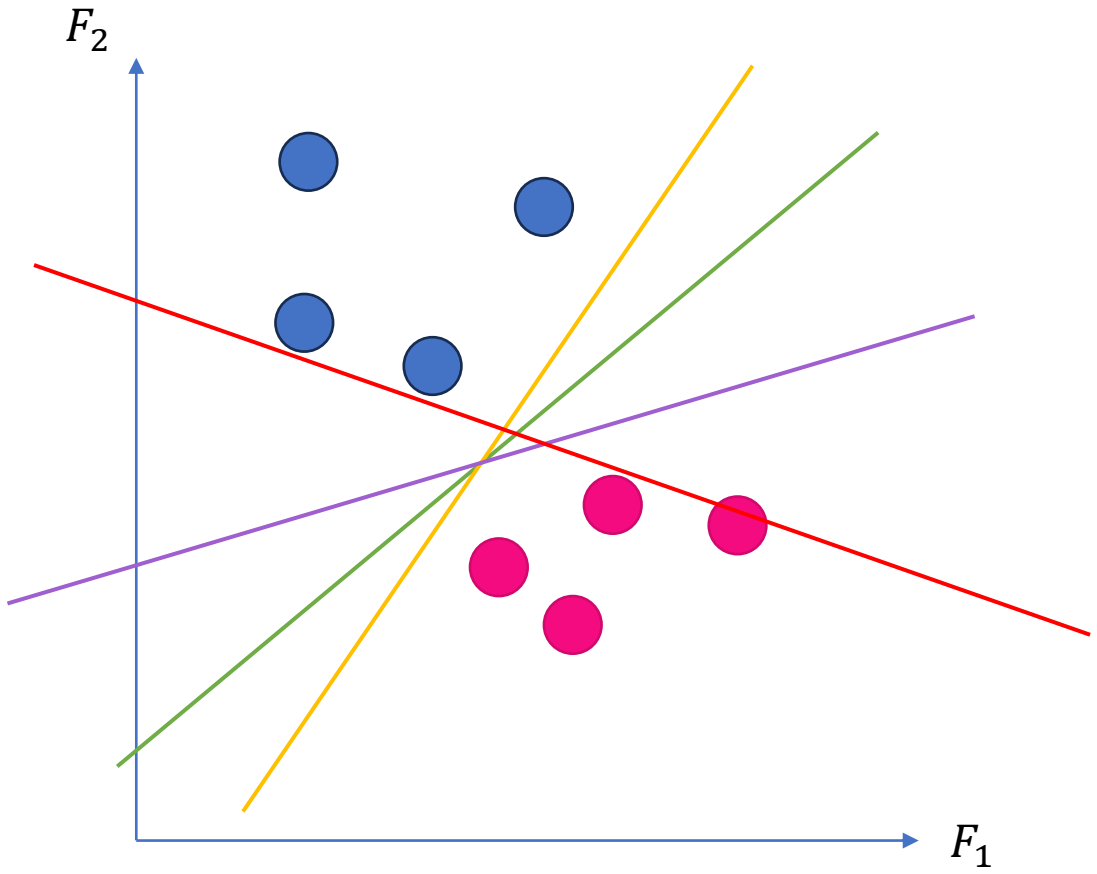
$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_n X_n$$



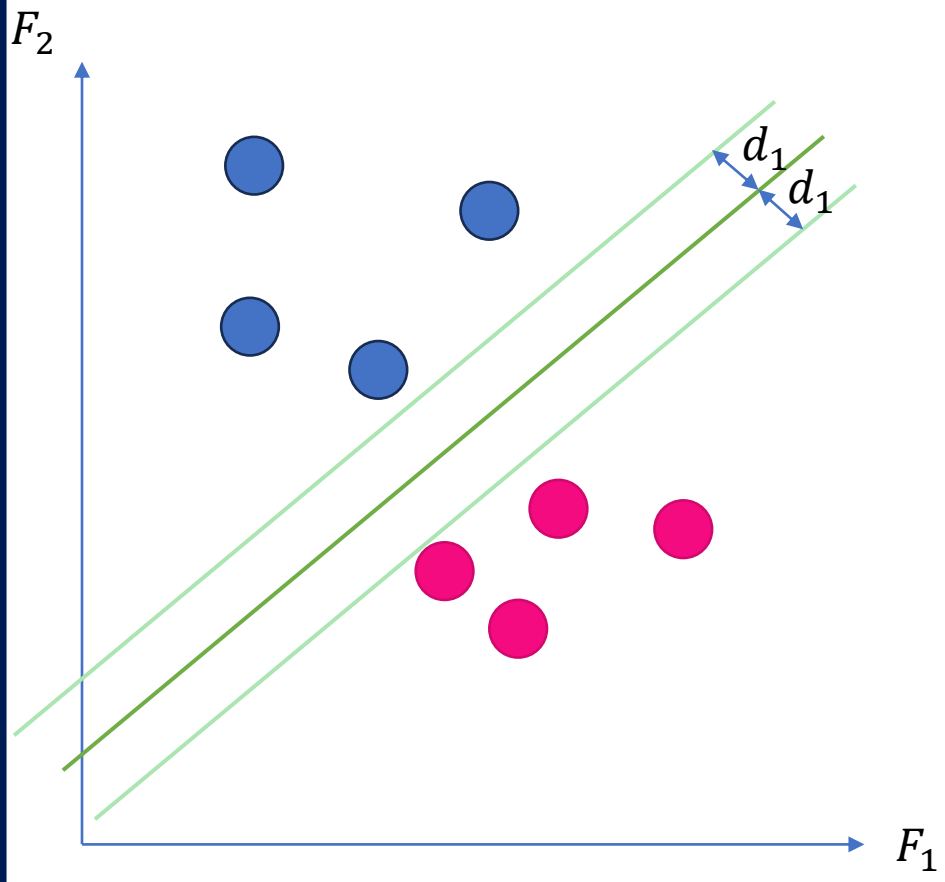
Support Vector Machine (SVM)

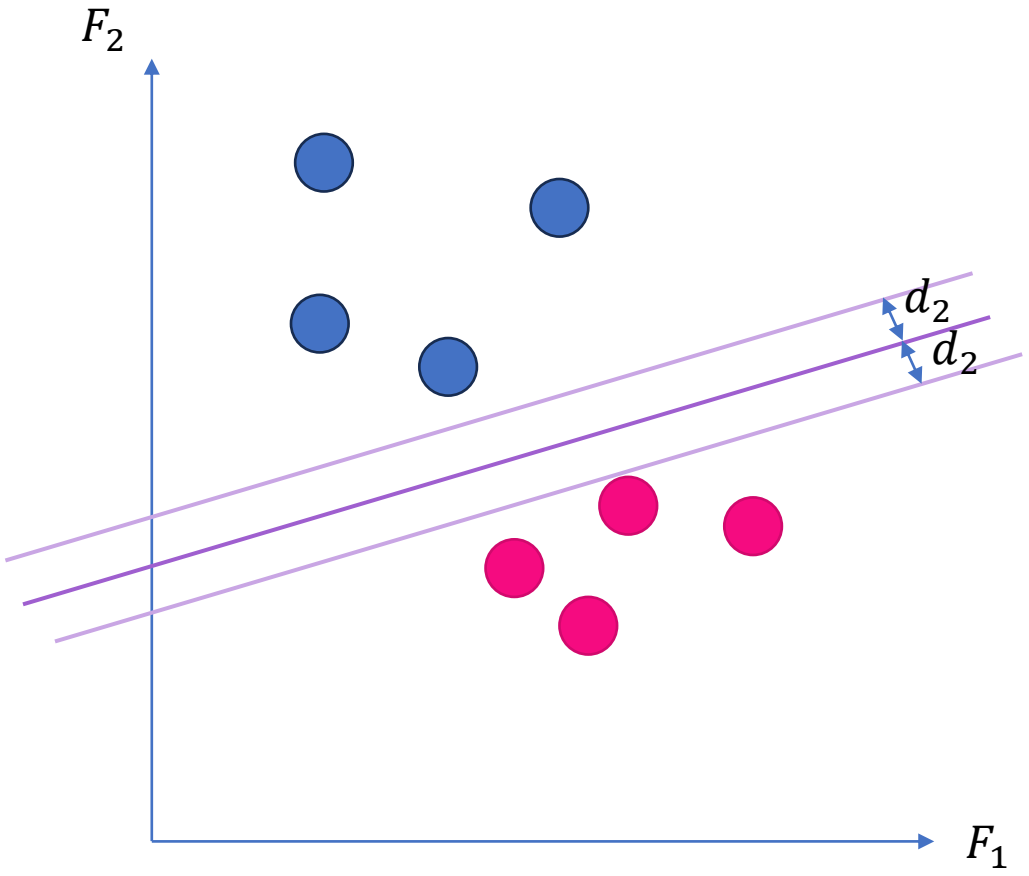
- SVM is a supervised machine learning algorithm used for both classification and regression
- The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space
- The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible
- The dimension of the hyperplane depends upon the number of features
  - If the number of input features is two, then the hyperplane is just a line
  - If the number of input features is three, then the hyperplane becomes a 2-D plane

Support Vector Machine (SVM)



Support Vector Machine (SVM)

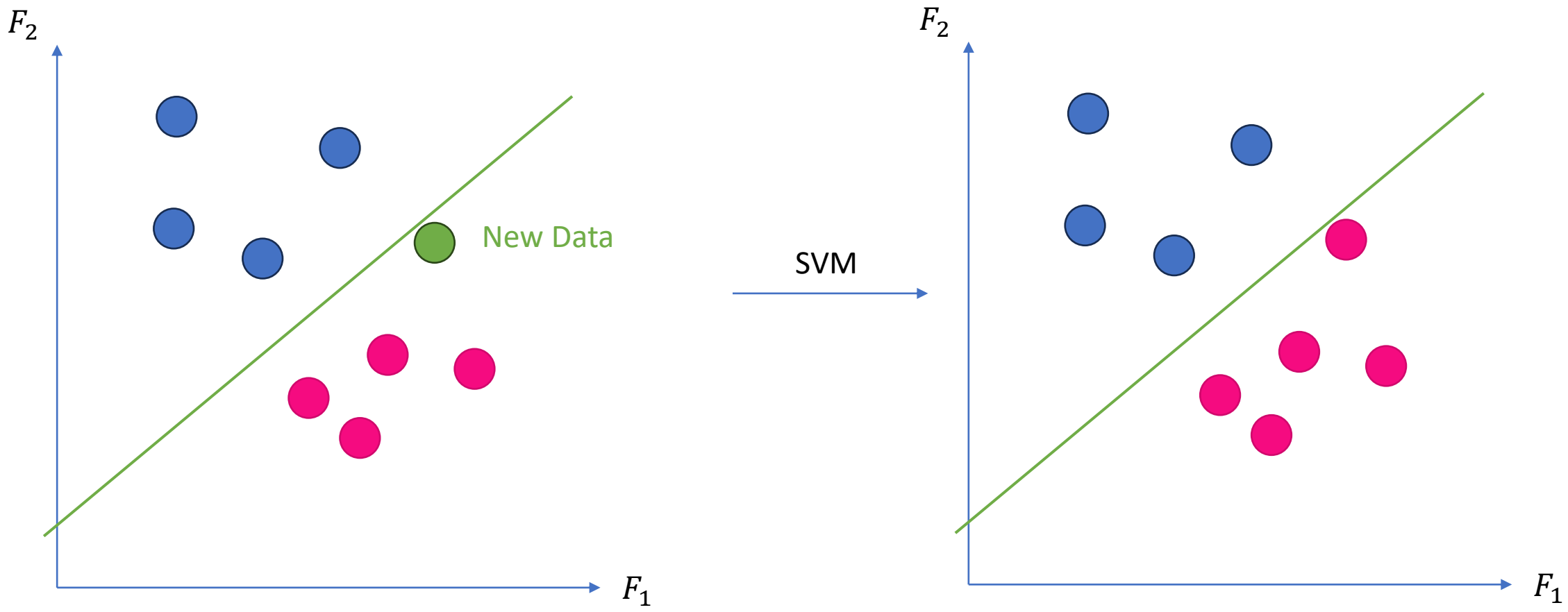




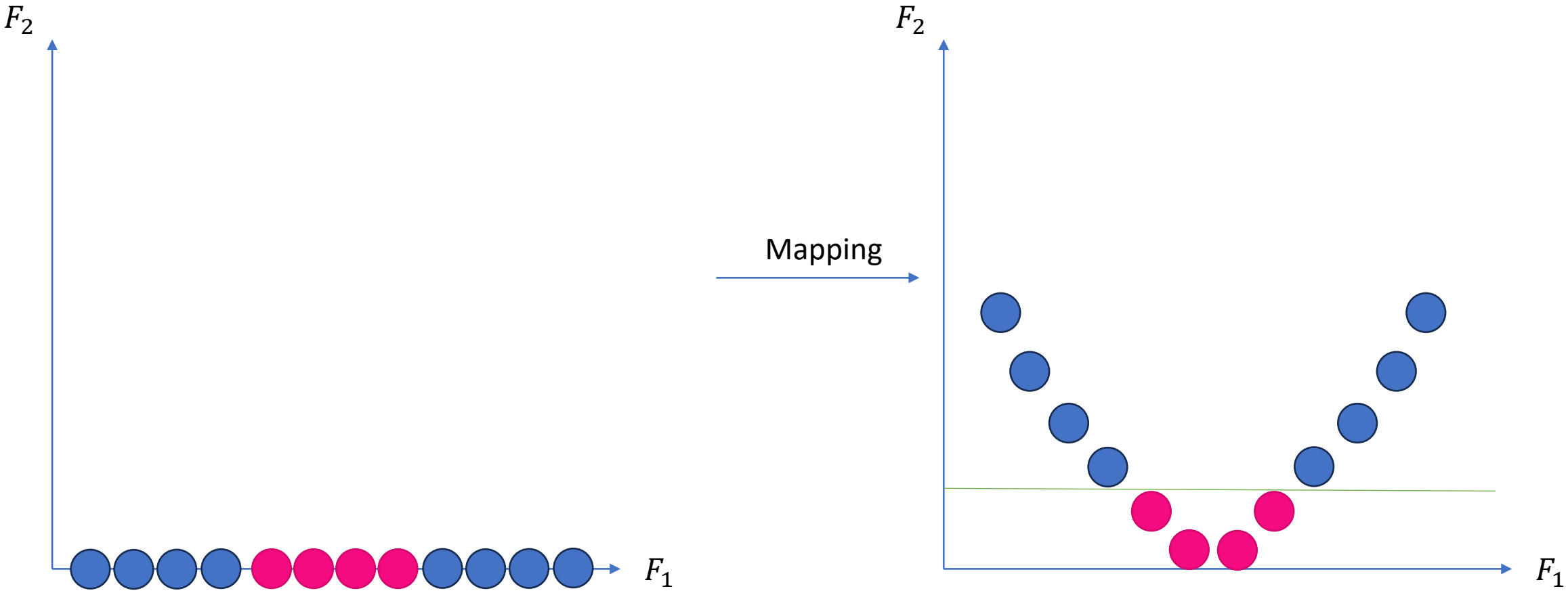
✓ Better Choice

$d_1 > d_2$

Support Vector Machine (SVM)



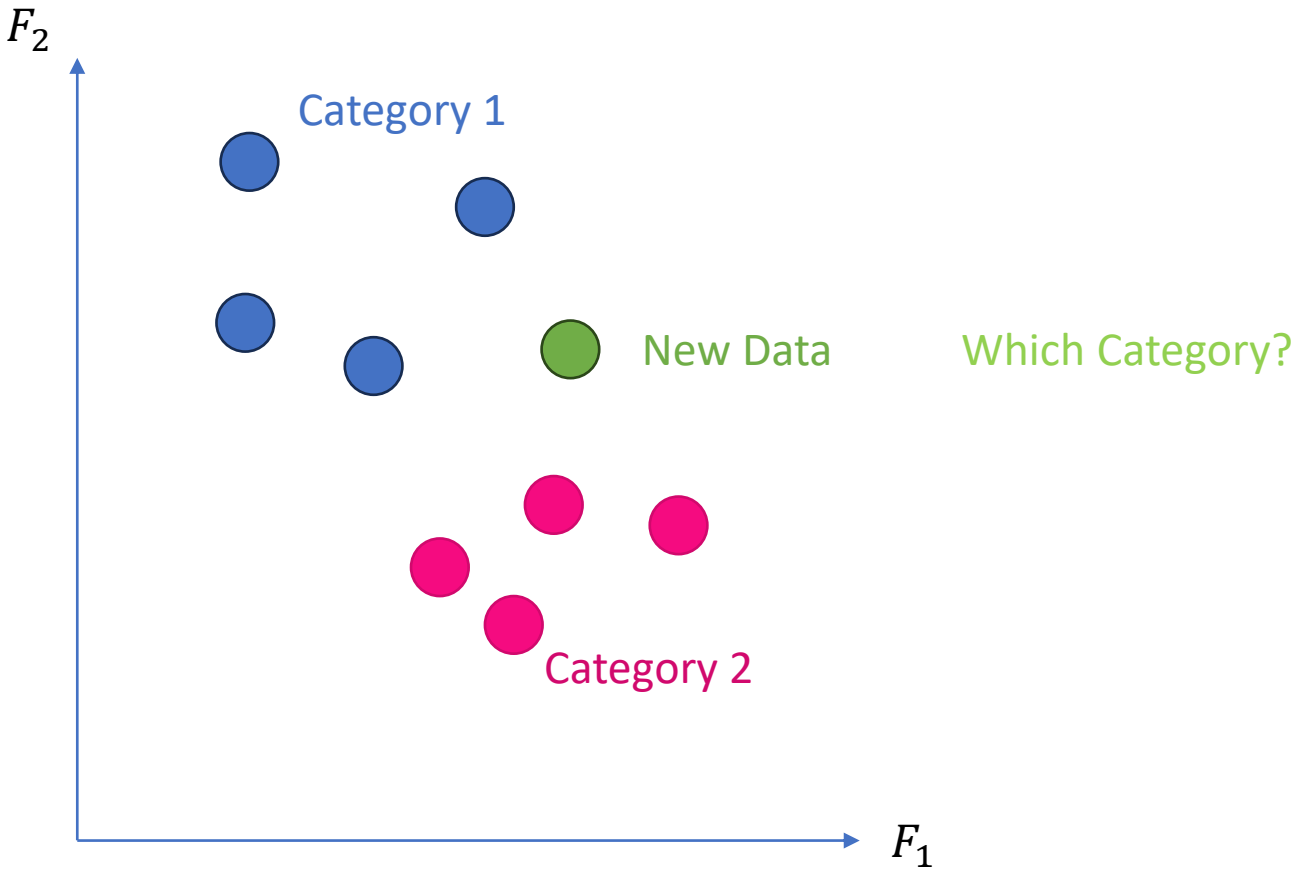
Support Vector Machine (SVM)



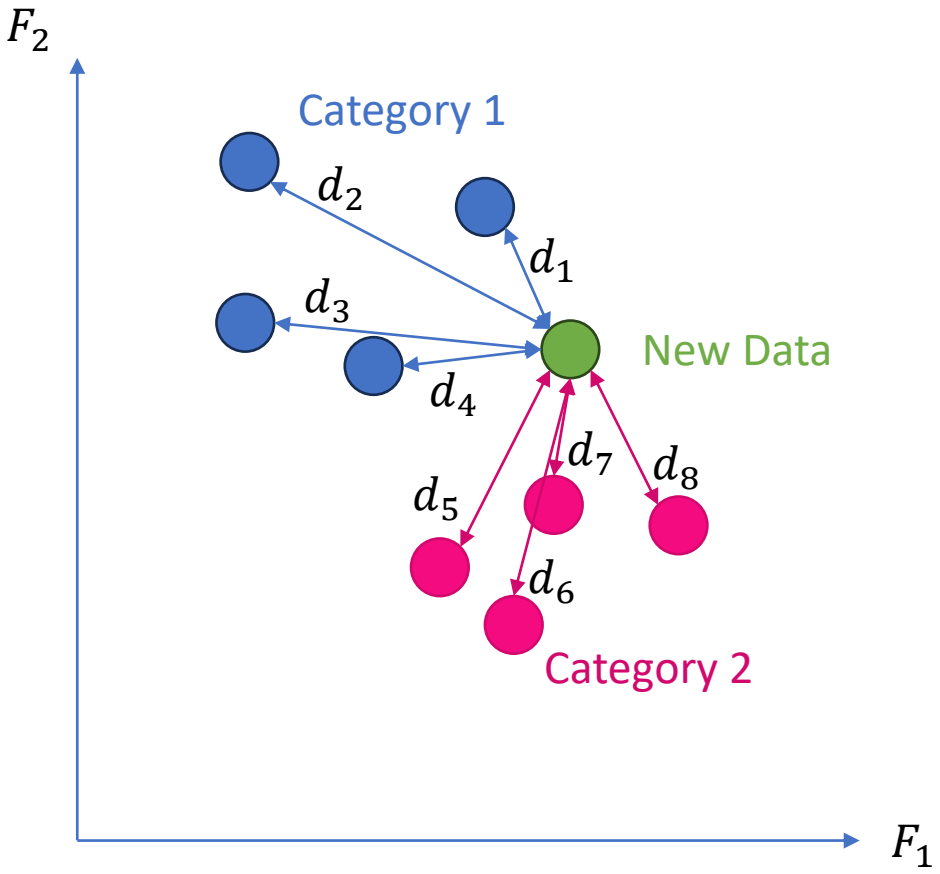
K-Nearest Neighbor (KNN)

- KNN is one of the most basic yet essential classification algorithms in Machine Learning
- It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection
- KNN uses some metrics to determine the closest groups or the nearest points for a query point
  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance

K-Nearest Neighbor (KNN)

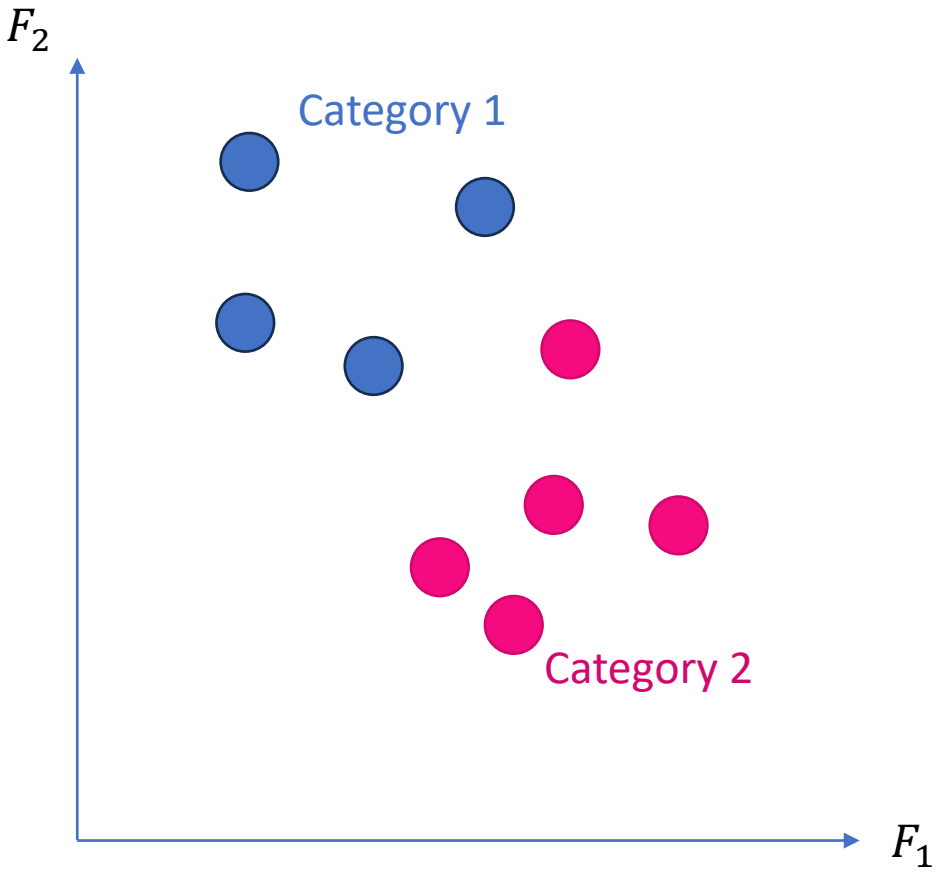


K-Nearest Neighbor (KNN)



$$\min(d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8) = d_7$$

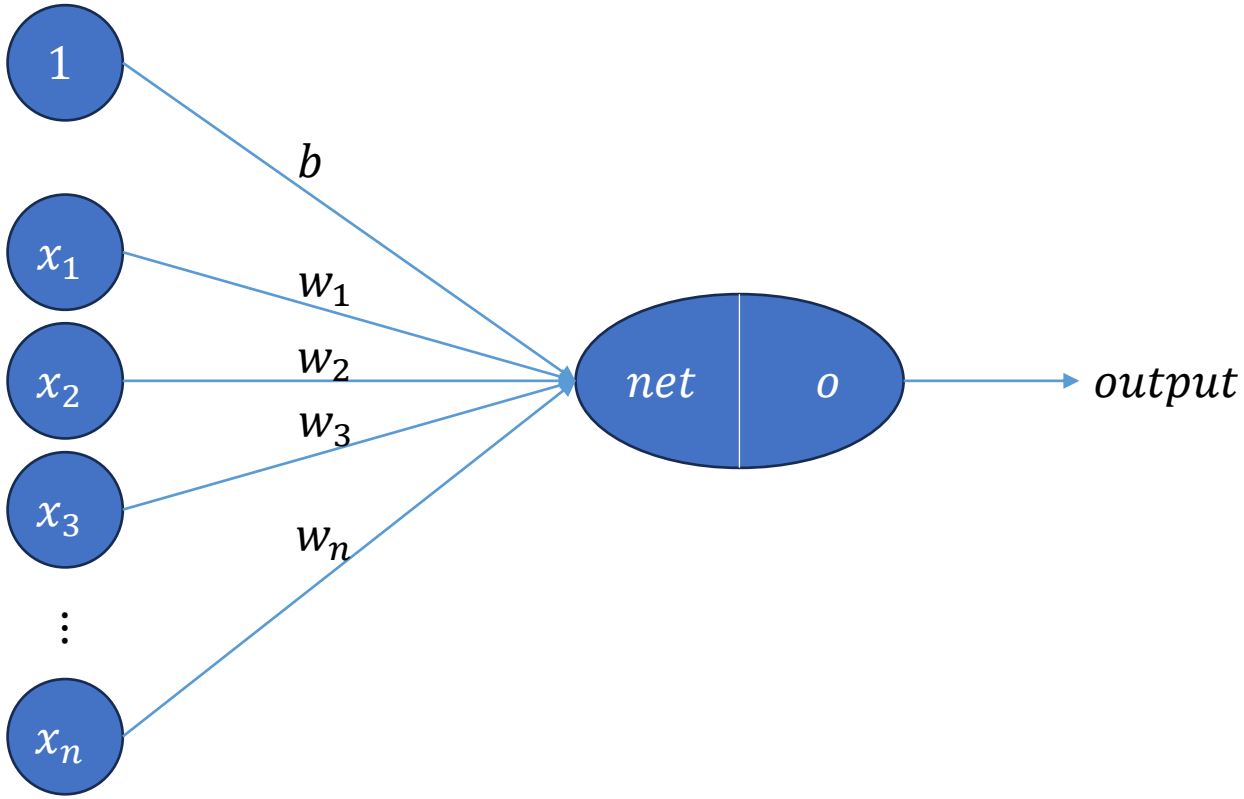
K-Nearest Neighbor (KNN)



Artificial Neural Network (ANN)

- ANN contains neurons that are arranged in a series of layers that together constitute the whole ANN in a system
- A layer can have only a dozen units or millions of units as this depends on how the complex neural networks will be required to learn the hidden patterns in the dataset
- ANN has
  - Input layer
    - The input layer receives data from the outside world which the neural network needs to analyze or learn about
  - Hidden layer(s)
    - The data passes through one or multiple hidden layers that transform the input into data that is valuable for the output layer
  - Output layer
    - The output layer provides an output in the form of a response of the ANN to input data provided

Artificial Neural Network (ANN)



$$net = w * x + b$$

$$o = activation\_function(net)$$

- w and b are trainable parameters
- x is the input