# Big Data (MHI222956/MHI225101)

## 7.1 ML introduction & KNN

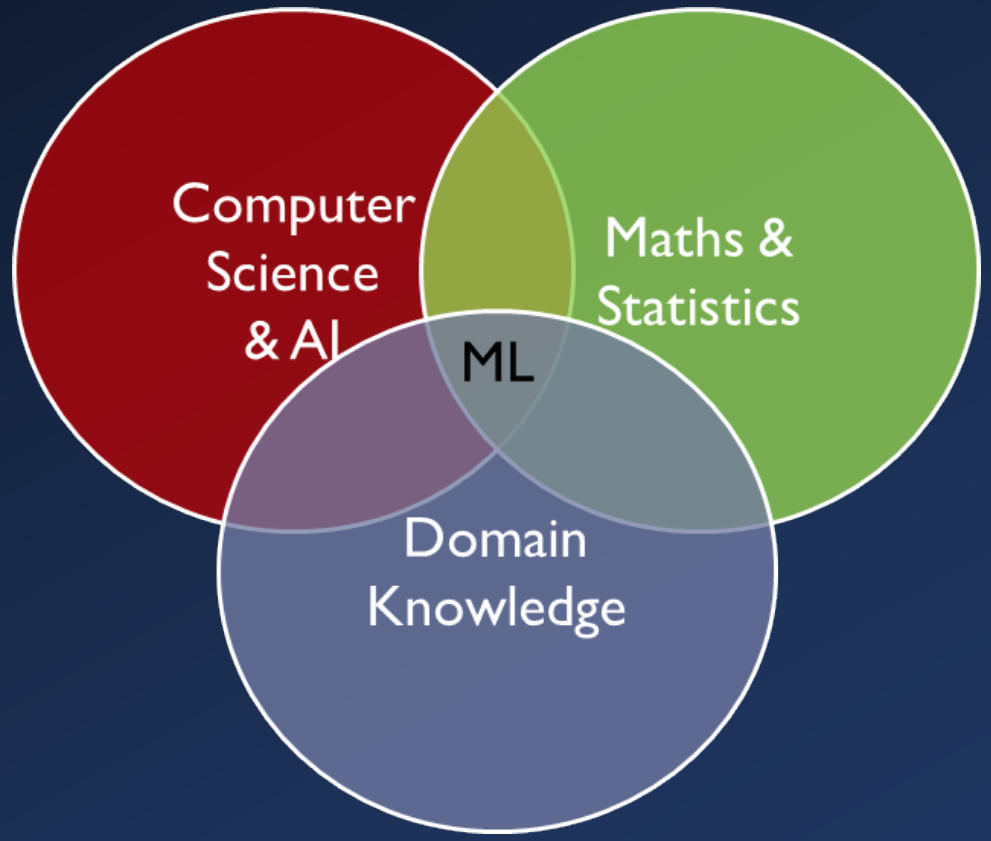
The simple linear form of data science process:

# What is Machine Learning?

- learn from data
- without being explicitly programmed
- can be used to discover hidden patterns and trends in the data
- allows for data driven decisions to be made

Machine learning algorithms and techniques are used to build models, to discover hidden patterns and trends in the data allowing for data-driven decisions to be made.

# An interdisciplinary field



# Some Terminologies

## Data Mining

Became popular around the time that the use databases became common place.

Was used to refer to activities related to finding patterns in databases and data warehouses.

## Artificial Intelligence (AI)

The theory and development of computer systems able to perform tasks normally requiring human intelligence.

Machine learning is a subset of AI.

## Data Science

A new term that is used to describe processing and analyzing data to extract meaning.

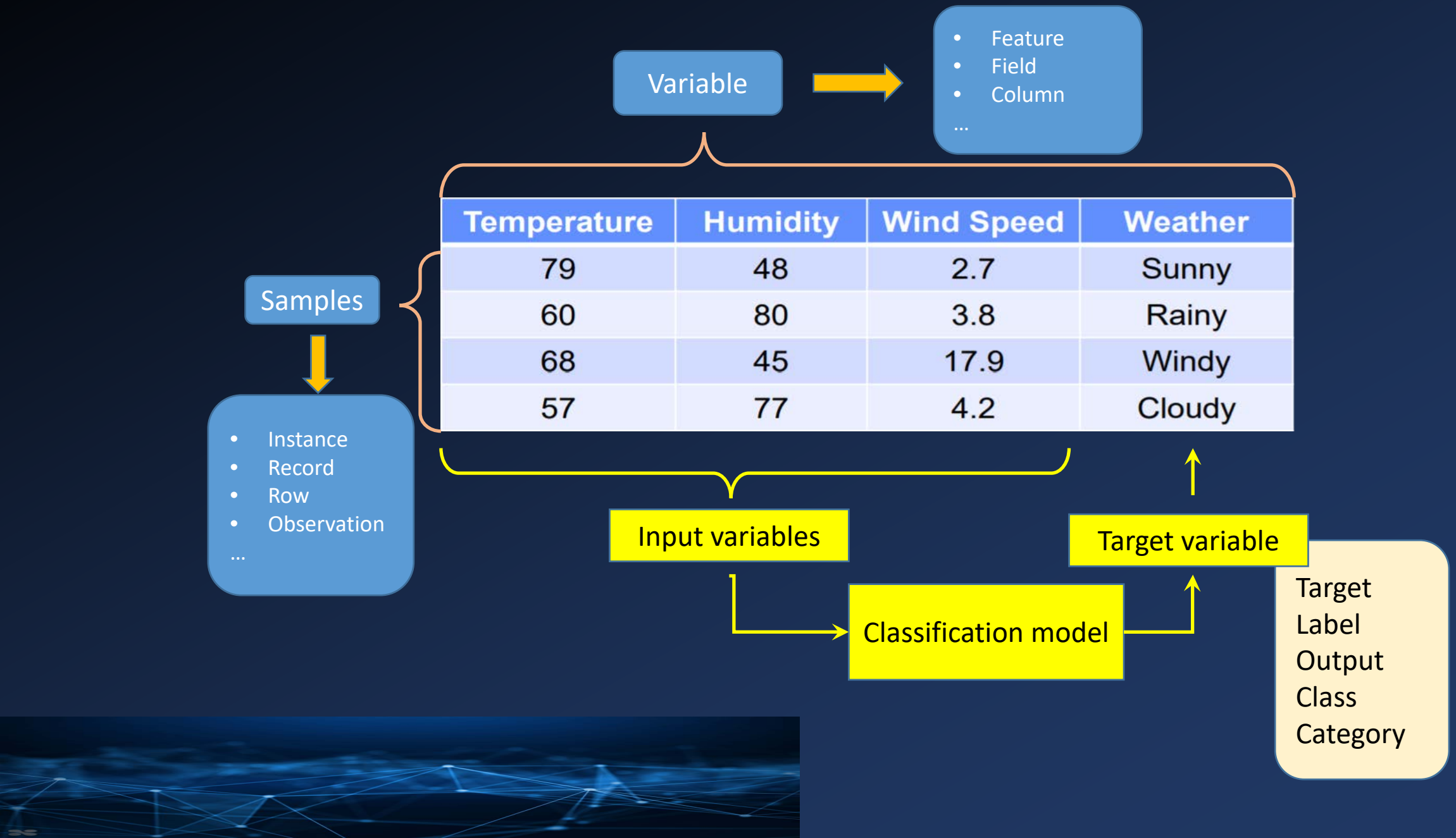
Machine learning techniques can also be used here.

Became popular at the same time that big data began appearing,

Usually refers to extracting meaning from big data and so includes approaches for collecting, storing and managing big data.

# Main Categories of ML

- Classification - to predict category value
- Regression - to predict a numeric value
- Cluster analysis - to organize similar items into groups
- Association analysis - to get a set of rules to capture associations between items or events


Variable
Feature
Field
Column
...
Temperature
Humidity
Wind Speed
Weather
79
48
2.7
Sunny
60
80
3.8
Rainy
68
45
17.9
Windy
57
77
4.2
Cloudy
Samples
Instance
Record
Row
Observation
...
Input variables
Target variable
Classification model
Target
Label
Output
Class
Category

**SUPERVISED LEARNING**

**UNSUPERVISED LEARNING**

# Supervised machine learning

- Supervised learning is where you have input variables (x) and an output variable (Y) – **labelled data** - and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

# Supervised machine learning

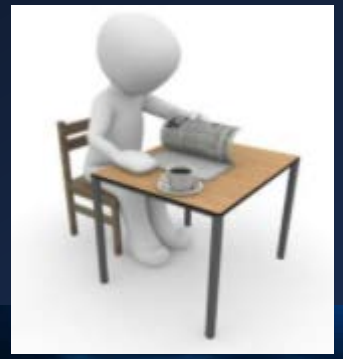
- The majority of practical machine learning uses supervised learning.

- Some popular examples of supervised machine learning algorithms are:
  - KNN
  - Decision tree
  - Linear regression.
  - SVM (Support Vector Machines)

- Supervised learning problems can be further grouped into regression and classification problems.
  - **Classification**: A classification problem is when the output variable is a category
  - **Regression**: A regression problem is when the output variable is a numeric value

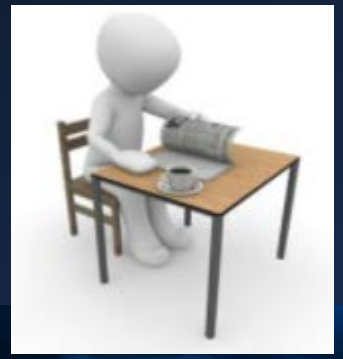
# Unsupervised machine learning

- Unsupervised learning is where you only have input data (X) and no corresponding output variables – **unlabelled data**.
    - The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.



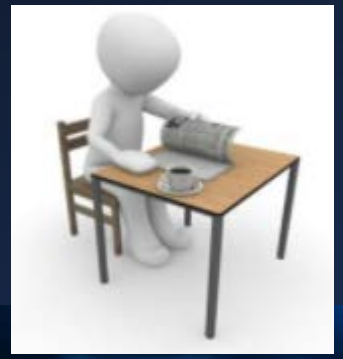
# Unsupervised machine learning

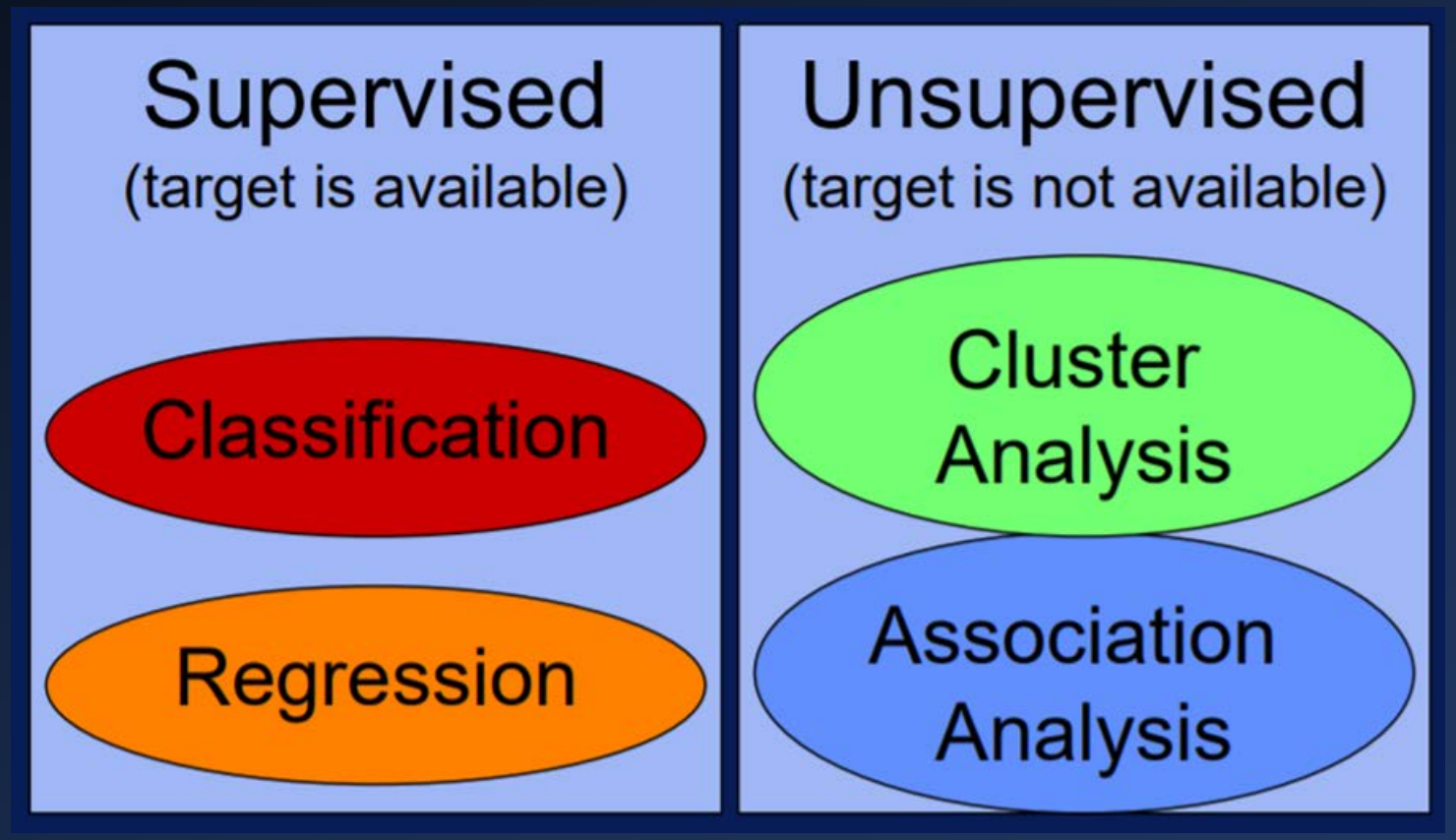
- Unlike supervised learning, there is no correct answers in unsupervised learning and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.
  - Some popular examples of unsupervised learning algorithms are:
    k-means for clustering problems.
    Apriori algorithm for association rule learning problems.



# Unsupervised machine learning

- Unsupervised learning problems can be further grouped into clustering and association problems.
  - **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.
  - **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy A also tend to buy B.




Supervised
(target is available)
Classification
Regression
Unsupervised
(target is not available)
Cluster Analysis
Association Analysis

# Semi-supervised machine learning

- Problems where you have a large amount of input data (X) and only some of the data is labelled (Y) are called semi-supervised learning problems.
- Fall somewhere in between supervised and unsupervised learning
- Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabelled data is cheap and easy to collect and store.

# Semi-supervised machine learning

- Use both labelled and unlabelled data for training:
  - You can use unsupervised learning techniques to discover and learn the structure in the input variables.
  - You can also use supervised learning techniques to make best guess predictions for the unlabelled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

- A good example is a photo archive where only some of the images are labelled, (e.g. dog, cat, person) and the majority are unlabelled.

# Reinforcement machine learning

- A learning method that interacts with its environment by producing actions and discovers errors or rewards.
- Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning.
- Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.
- This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. These algorithms are useful in the field of Robotics, Gaming etc.

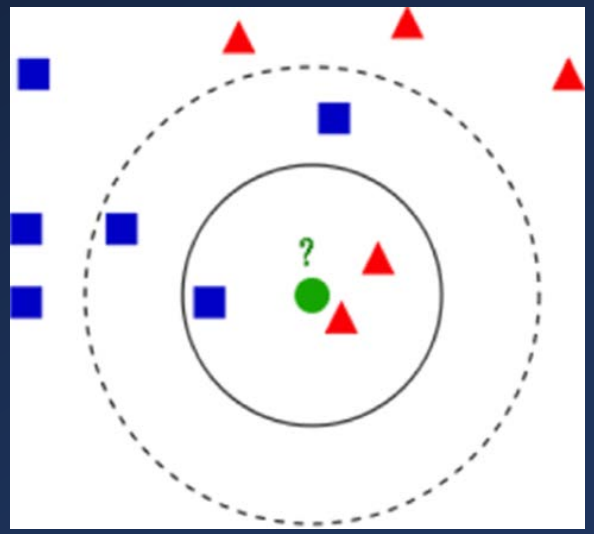
# k Nearest Neighbours

# (kNN)

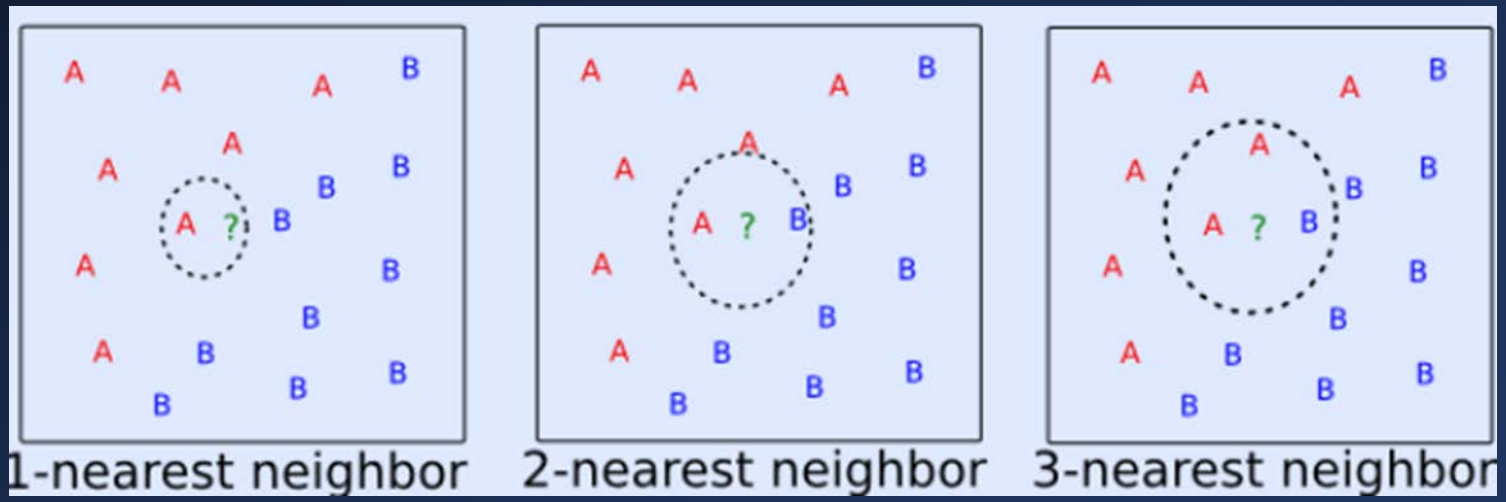
# kNN

- One of the simplest techniques to build a classification model

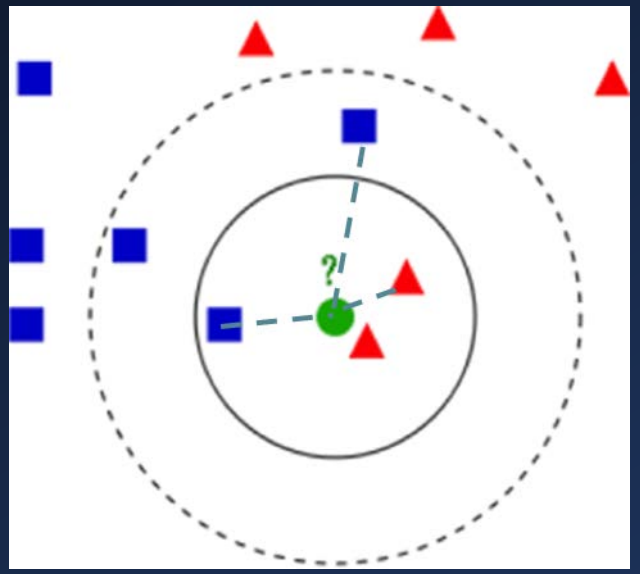
- Classify / Label a sample based on its neighbors
  - Assuming that samples with similar input values likely belong to the same class



- What is the k?
  - The value of k determines the number of nearest neighbor to consider



- Distance measure
  - to determine the “closeness”

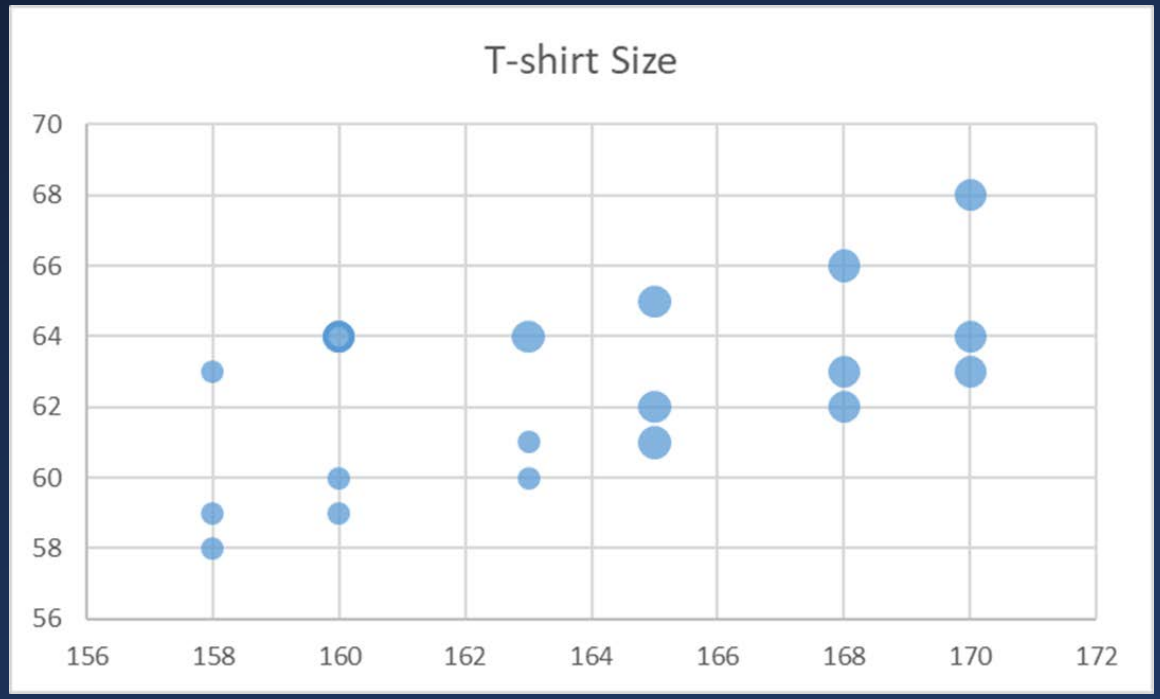



- Distance metrics
  - Euclidean Distance
  - City Block Distance (Manhattan Distance)
  - Chi square distance
  - Cosine distance

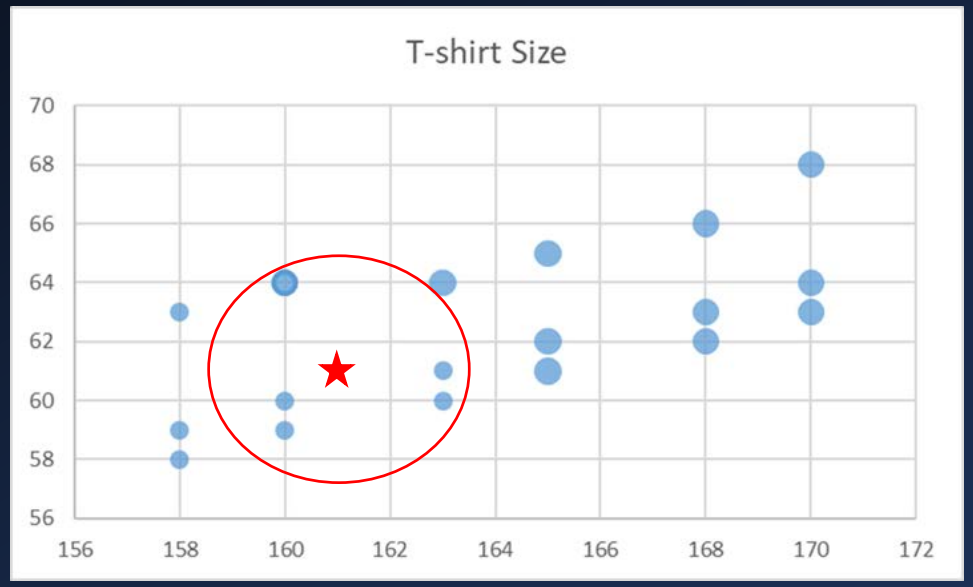
https://numerics.mathdotnet.com/Distance.html

| Height (in cms) | Weight (in kgs) | T Shirt Size |
|---|---|---|
| 158 | 58 | M |
| 158 | 59 | M |
| 158 | 63 | M |
| 160 | 59 | M |
| 160 | 60 | M |
| 163 | 60 | M |
| 163 | 61 | M |
| 160 | 64 | L |
| 163 | 64 | L |
| 165 | 61 | L |
| 165 | 62 | L |
| 165 | 65 | L |
| 168 | 62 | L |
| 168 | 63 | L |
| 168 | 66 | L |
| 170 | 63 | L |
| 170 | 64 | L |
| 170 | 68 | L |

- Suppose we have height, weight and T-shirt size of some customers:



- Now, we need to predict the T-shirt size of a new customer given only height and weight information.
- E.g., **'Tom' has height 161cm and weight 61kg.**




$$Distance = SQRT((161 - height)^2 + (61 - weight)^2)$$

| Height (in cms) | Weight (in kgs) | T Shirt Size | Distance |
|---|---|---|---|
| 158 | 58 | M | 4.24264069 |
| 158 | 59 | M | 3.60555128 |
| 158 | 63 | M | 3.60555128 |
| 160 | 59 | M | 2.23606798 |
| 160 | 60 | M | 1.41421356 |
| 163 | 60 | M | 2.23606798 |
| 163 | 61 | M | 2 |
| 160 | 64 | L | 3.16227766 |
| 163 | 64 | L | 3.60555128 |
| 165 | 61 | L | 4 |
| 165 | 62 | L | 4.12310563 |
| 165 | 65 | L | 5.65685425 |
| 168 | 62 | L | 7.07106781 |
| 168 | 63 | L | 7.28010989 |
| 168 | 66 | L | 8.60232527 |
| 170 | 63 | L | 9.21954446 |
| 170 | 64 | L | 9.48683298 |
| 170 | 68 | L | 11.4017543 |

# kNN

- No separate training phase
- Can generate complex decision boundaries allowing for complex classification decisions to be made
- Can be susceptible to noise
- Can be slow, since the distance between a new sample and all sample points in the data must be calculated in order to determine the k-Nearest Neighbors

# Summary

- What is ML
- Categories of ML
- Supervised & Unsupervised ML
- kNN