**Introduction to Machine Learning and Clustering - Summary**

**Previous Lesson Overview**

- **Topic**: Exploratory (spatial) data analysis (E(S)DA)
- **Key Points**:
    - Fundamentals and importance of E(S)DA
    - Introduction to statistical and visualization methods
    - Addressing a real-world geospatial problem with E(S)DA
    - Use of open-source Python tools

**Learning Objectives for Current Lesson**

- Explain:
    - Fundamentals of Artificial Intelligence (AI)
    - Principles of Machine Learning (ML) methods
    - Basics of clustering methods
- Apply open-source Python tools to perform geodata clustering

**Key Concepts and Definitions**

1. **Artificial Intelligence (AI)**:
    - Science of creating intelligent machines simulating human thinking and behavior.
2. **Machine Learning (ML)**:
    - Technology and algorithms enabling computers to learn from data, examples, and experience.
    - Involves finding rules consistent with the data.
3. **Types of Machine Learning**:
    - **Unsupervised Learning (Clustering)**:
        - No class labels; checks for groups/clusters in the data.
    - **Supervised Learning**:
        - Training data with known target variables; makes predictions on new data.
        - Subtypes include classification (predicting discrete labels) and regression (predicting continuous variables).

**Clustering**

- **Definition**:
    - Identifies groups of elements similar among themselves but dissimilar to other groups.
- **Importance**:
    - Facilitates the extraction of useful information by providing a high-level abstraction of data.

**Clustering Algorithms**

1. **K-means Clustering**:
    - **Process**:
        - Initialization: Arbitrary initialization of cluster centers.
        - Data Assignment: Points assigned to the closest cluster.
        - Relocation of Centers: Cluster centers moved to the mean of assigned points.
        - Iteration: Repeat until cluster centers stabilize.
    - **Limitations**:
        - Sensitive to outliers and initial setup.
        - Choosing the number of clusters (k) can be challenging.
    - **Evaluation Methods**:
        - **Silhouette Method**: Measures how similar a point is to its cluster compared to other clusters.
        - **Elbow Method**: Identifies the optimal number of clusters by plotting the sum of squared distances and finding the "elbow point."

**Practical Application Example**

- **Scenario**: Ghelgheli's Teahouse
    - **Data**: Customer purchase history, demographics, feedback, and visit times.
    - **Methods**: Data cleaning, exploratory data analysis (EDA), and clustering algorithms (K-means).
    - **Steps**:
        - Data collection and cleaning
        - EDA and data standardization
        - Choosing and applying a clustering algorithm
        - Analyzing and interpreting clusters

**Conclusion**

- **Summary**:
    - Clustering helps in understanding data by grouping similar data points.
    - Various methods and techniques can be applied for effective clustering and analysis.
    - Practical applications demonstrate the value of clustering in real-world scenarios.

**Key Questions and Answers**

1. **Exploratory (spatial) data analysis is always recommended before modeling**.
    - True
2. **For exploratory (spatial) data analysis, it is always recommended to use both statistical analysis and visualization tools**.
    - True
3. **When a geospatial dataset is clustered distributed in space, the Moran's I value is positive**.
    - True
4. **Identifying various land use categories using remote sensing images is a classification problem**.
    - True
5. **Predicting house prices based on various spatial and non-spatial factors is a regression problem**.
    - True
6. **Characterizing tick bites risk in a study area using observational and geospatial datasets can be formulated both as a regression or classification**.
    - True

UNIVERSITY OF TWENTE.

# Introduction to Machine Learning and Clustering

*Mahdi KHODADADZADEH*
*Assistant Professor*
*Faculty of Geo-Information Science and Earth Observation (ITC)*
*Department of Geo-information Processing (GIP)*
*m.khodadadzadeh@utwente.nl*

*May 2024*

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

# Previous lesson

We discussed the fundamentals and importance of E(S)DA

We introduced some statistical and visualization methods

We discussed a real-world geospatial problem which can be addressed by using E(S)DA

We introduced some open-source Python tools

# Question #1

Exploratory (spatial) data analysis is always recommended before modelling.

True

False

# Question #2

For exploratory (spatial) data analysis is always recommended to use both statistical analysis and visualization tools.

True

False

# Question #3

When a geospatial dataset is clustered distributed in space, the Moran's I value is positive.

True

False


From: https://xkcd.com

# This lesson's learning objectives

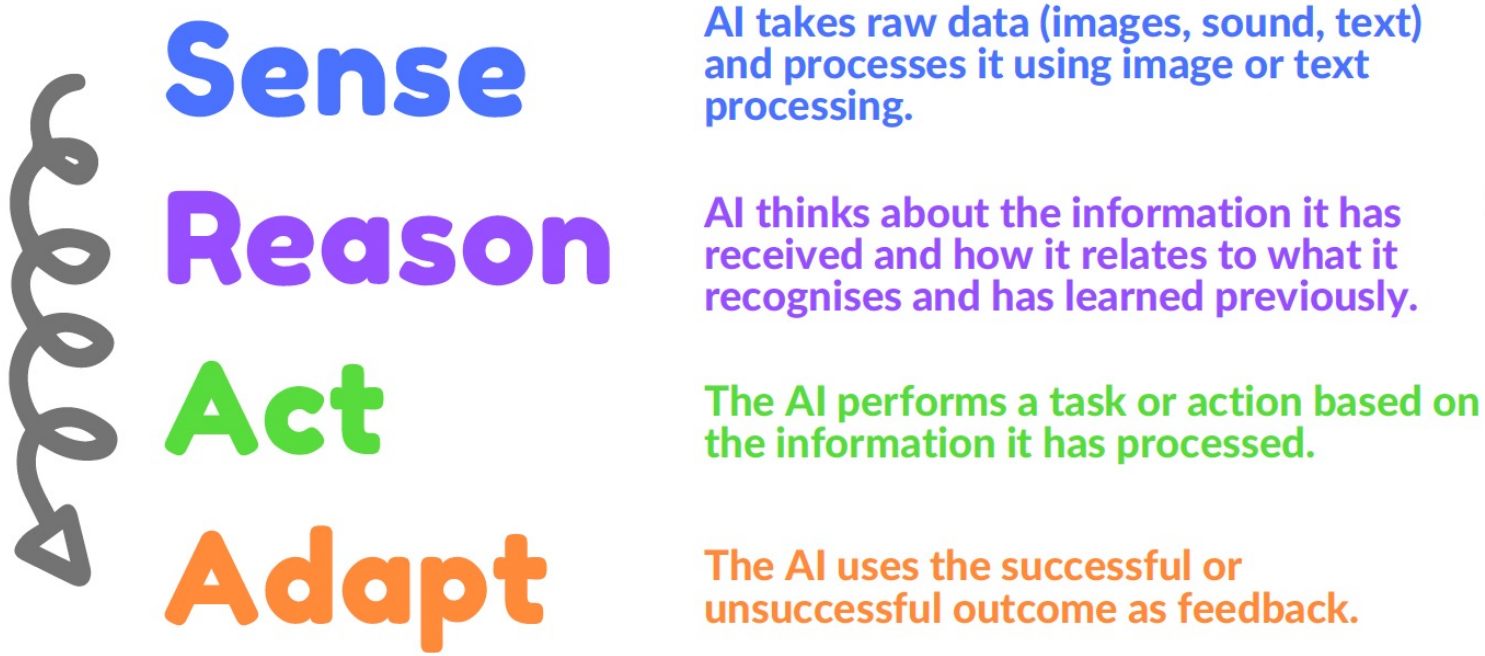
**Explain** to peers

- the fundamentals of **Artificial Intelligence**
- the principles of **Machine Learning** methods
- the basics of **clustering** methods

**Apply** some **open-source Python tools** to perform geodata clustering

# Artificial Intelligence (AI)

- AI is the **science** of **creating intelligent machines** that can **simulate human** thinking capability and behavior.



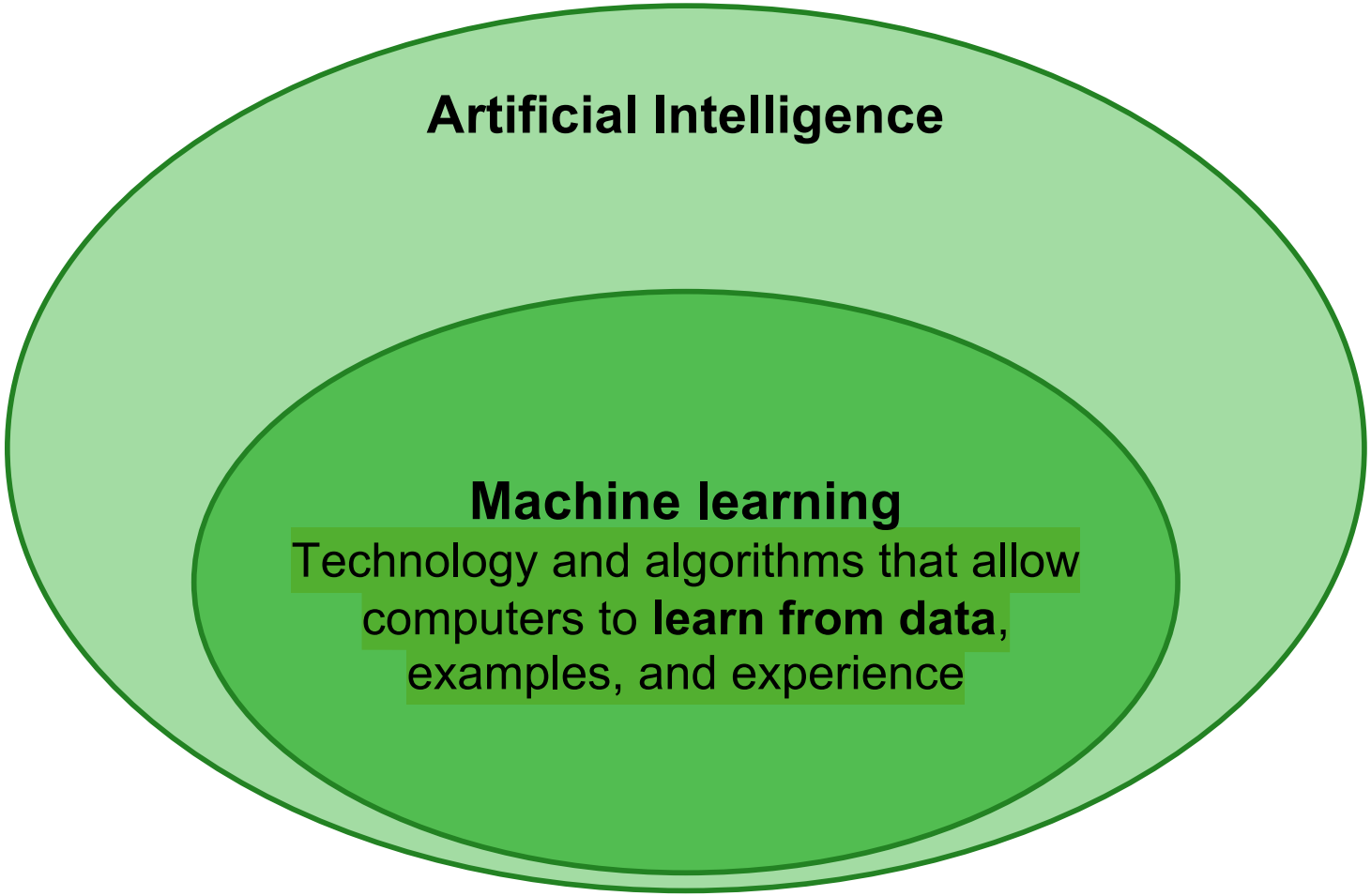

From: https://csermoocs.appspot.com/ai_secondary/unit?unit=2&lesson=4

Chat GPT



https://openai.com/blog/chatgpt

# Artificial Intelligence (AI) vs. Machine Learning (ML)



# What is Machine Learning?

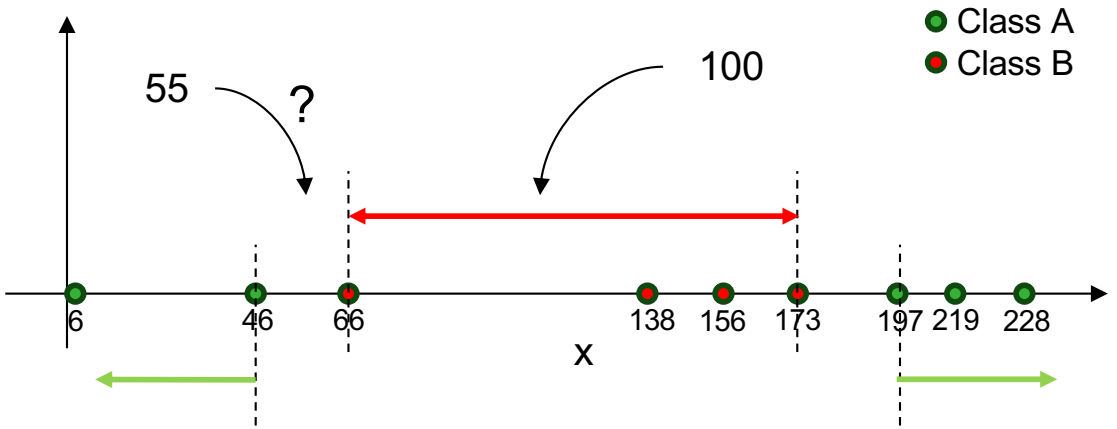
- **Machine Learning = Learning from data**

  **= Finding a rule** that is consistent with the data

Dataset

| X | Class |
|---|---|
| 228 | A |
| 66 | B |
| 138 | B |
| 219 | A |
| 156 | B |
| 46 | A |
| 197 | A |
| 6 | A |
| 173 | B |
| 100 | ? |

How can we improve the definition of the rule?




Rule = {A: X≥197 or X≤46} and {B: 66≤X≤173}

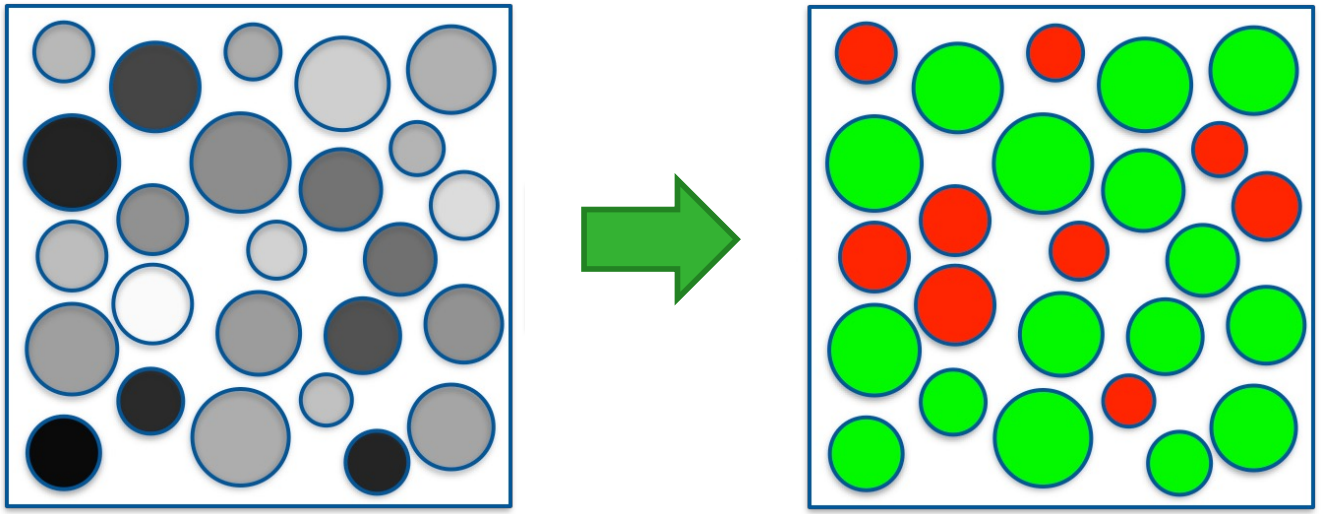
# What is Machine Learning?

- In machine learning, a feature is a measurable property of a phenomenon being observed.
- If we have multiple features, can you easily define a rule?

Dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | Class |
|---|---|---|---|---|---|---|---|
| 200 | 220 | 195 | 190 | 188 | 230 | 228 | A |
| 60 | 70 | 55 | 50 | 60 | 73 | 67 | B |
| 130 | 140 | 140 | 140 | 145 | 148 | 138 | B |
| 200 | 210 | 230 | 230 | 220 | 200 | 209 | A |
| 150 | 160 | 143 | 140 | 155 | 164 | 156 | B |
| 40 | 50 | 74 | 75 | 45 | 55 | 46 | A |
| 190 | 200 | 205 | 200 | 200 | 230 | 197 | A |
| 1 | 10 | 15 | 15 | 1 | 12 | 6 | A |
| 170 | 180 | 160 | 165 | 178 | 180 | 173 | B |

# What is Machine Learning?



Can we define a set of rules (specific procedure) that a computer follows to perform this task?

Algorithm

# When should we use Machine Learning?

- Problems where there are **no human experts**, so data cannot be labelled or categorized
- Problems where there are human experts, but it is **very hard (or impossible) to define rules**
- Problems where there are human experts and the rules can be defined, but where it is **not cost effective to implement**

- When we want to
  - ✓ **Detect** patterns/structures/themes/trends etc. in the data
  - ✓ **Make predictions** about future data and **make decisions**

# Categories of Machine Learning

## Unsupervised learning (clustering)

Class labels unknown
Checking for groups/clusters in the data

Dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|
| 200 | 220 | 195 | 190 | 188 | 230 | 228 |
| 60 | 70 | 55 | 50 | 60 | 73 | 67 |
| 130 | 140 | 140 | 140 | 145 | 148 | 138 |
| 200 | 210 | 230 | 230 | 220 | 200 | 209 |
| 150 | 160 | 143 | 140 | 155 | 164 | 156 |
| 40 | 50 | 74 | 75 | 45 | 55 | 46 |
| 190 | 200 | 205 | 200 | 200 | 230 | 197 |
| 195 | 205 | 200 | 190 | 180 | 210 | 188 |
| 170 | 180 | 160 | 165 | 178 | 180 | 173 |


| Class |
|---|
| A |
| B |
| B |
| A |
| B |
| A |
| A |
| A |
| B |

# Categories of Machine Learning

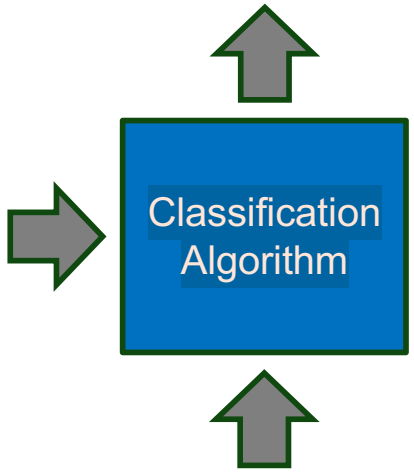
## Supervised learning

Training data with known target variable
Making predictions on unseen/new data

Discrete labels
or
Continuous

Dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | Class |
|---|---|---|---|---|---|---|---|
| 200 | 220 | 195 | 190 | 188 | 230 | 228 | A |
| 60 | 70 | 55 | 50 | 60 | 73 | 67 | B |
| 130 | 140 | 140 | 140 | 145 | 148 | 138 | B |
| 200 | 210 | 230 | 230 | 220 | 200 | 209 | A |
| 150 | 160 | 143 | 140 | 155 | 164 | 156 | B |
| 40 | 50 | 74 | 75 | 45 | 55 | 46 | A |




| | | | | | | |
|---|---|---|---|---|---|---|
| 190 | 200 | 205 | 200 | 200 | 230 | 197 |
| 195 | 205 | 200 | 190 | 180 | 210 | 188 |
| 170 | 180 | 160 | 165 | 178 | 180 | 173 |

# Categories of Machine Learning

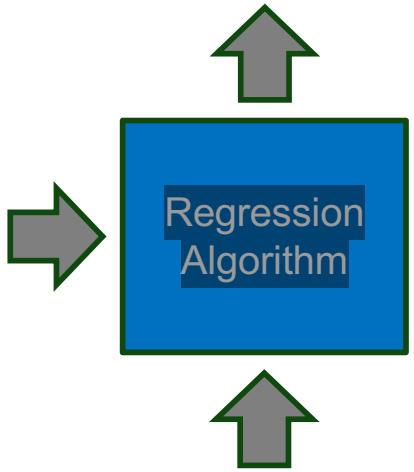
## Supervised learning

Training data with known target variable
Making predictions on unseen/new data

Discrete labels
or
Continuous

Dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | Target |
|---|---|---|---|---|---|---|---|
| 200 | 220 | 195 | 190 | 188 | 230 | 228 | 10 |
| 60 | 70 | 55 | 50 | 60 | 73 | 67 | 20 |
| 130 | 140 | 140 | 140 | 145 | 148 | 138 | 22 |
| 200 | 210 | 230 | 230 | 220 | 200 | 209 | 5 |
| 150 | 160 | 143 | 140 | 155 | 164 | 156 | 15 |
| 40 | 50 | 74 | 75 | 45 | 55 | 46 | 12 |




| Target |
|---|
| 12 |
| 18 |
| 22 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 190 | 200 | 205 | 200 | 200 | 230 | 197 |
| 195 | 205 | 200 | 190 | 180 | 210 | 188 |
| 170 | 180 | 160 | 165 | 178 | 180 | 173 |

# Question #1

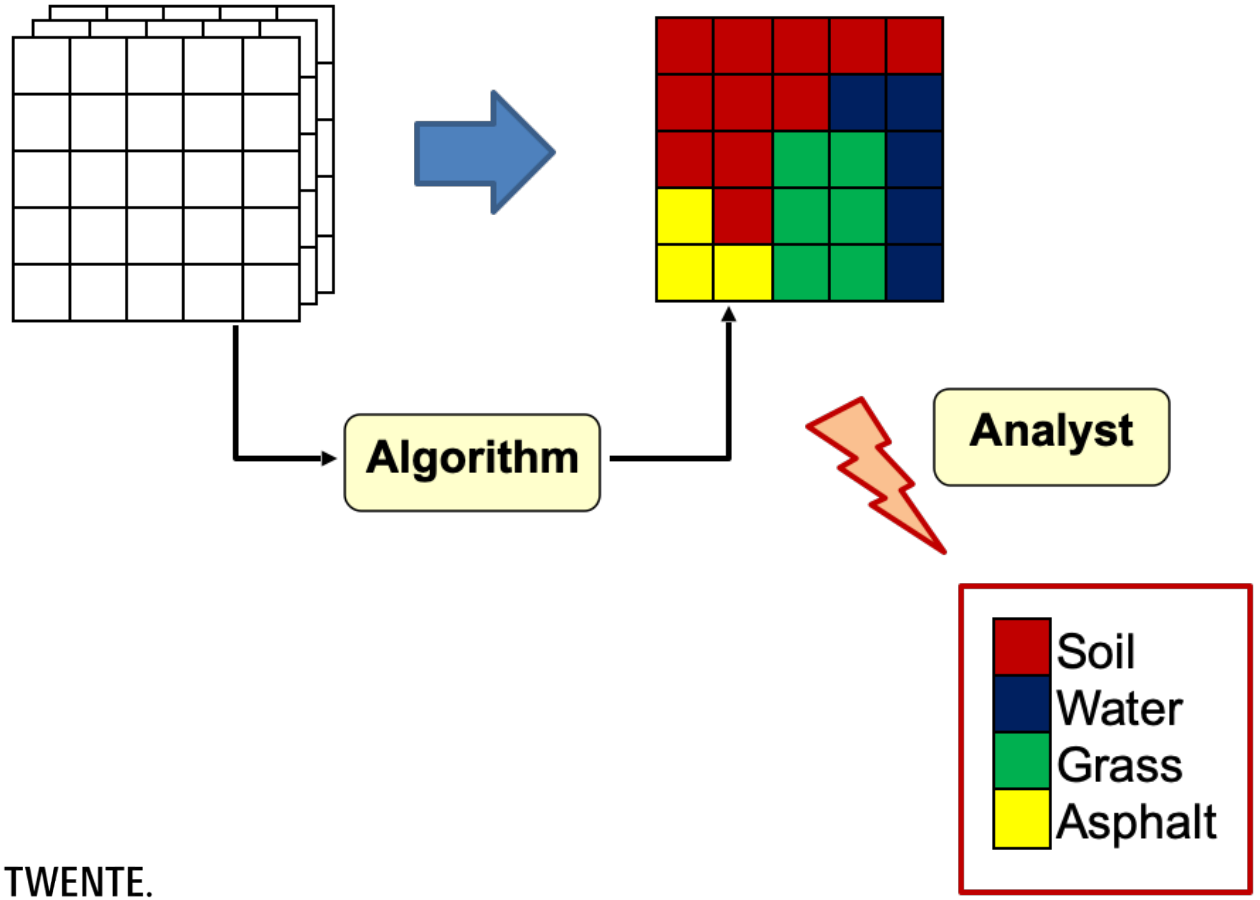
Identifying various land use categories using remote sensing images is a classification problem.

True

False

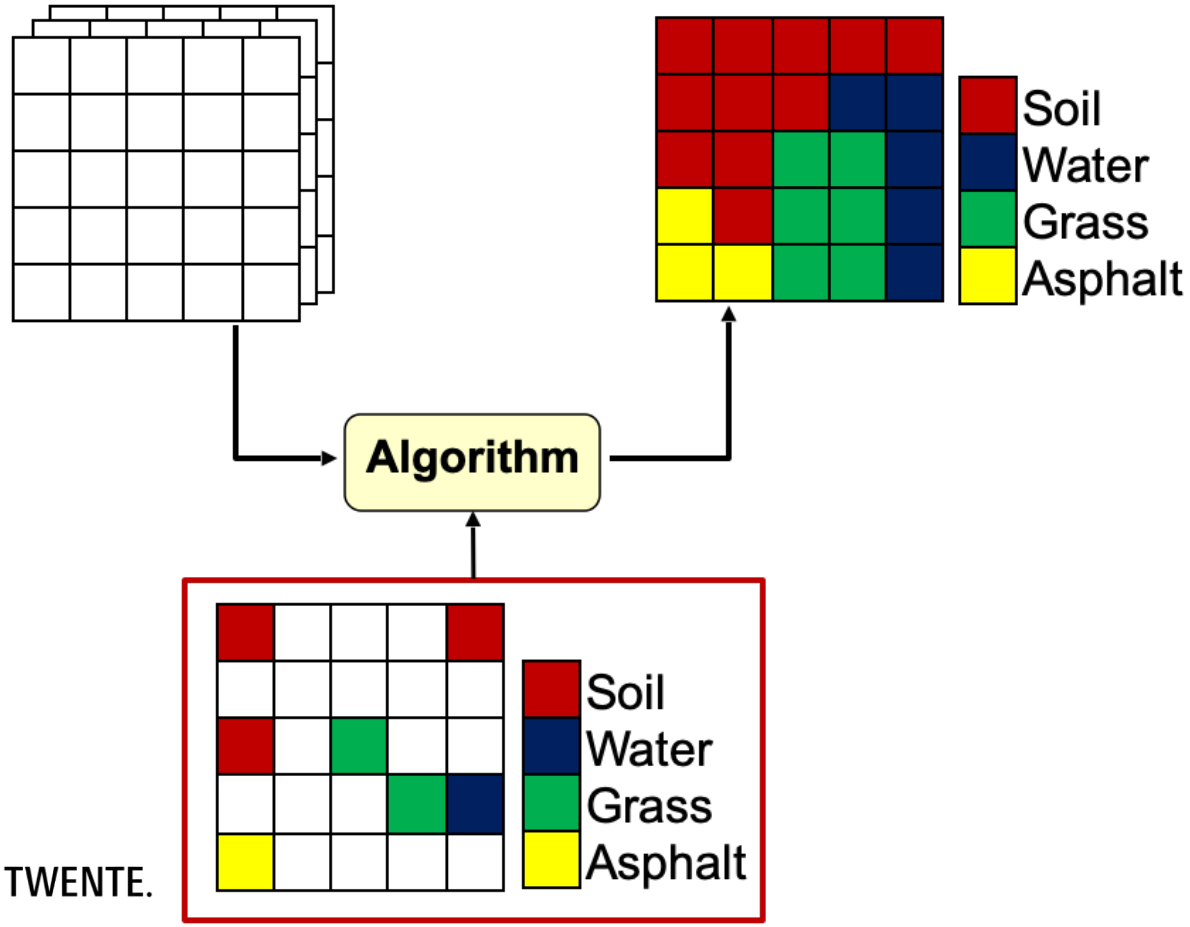
# Clustering

- Machine learning algorithm finds hidden patterns or intrinsic structures in input data



# Classification

- Machine learning algorithm trains a model on known input and output data so that it can predict future outputs (labels).



# Question #2

Predicting house prices based on various spatial and non-spatial factors is a regression problem.

True

False

Predicting house prices is a classic regression task. In this context, you'll typically use features like square footage, number of bedrooms, location, and other relevant factors to estimate the sale price of a house. The goal is to create a model that can generalize well to unseen data.

# Question #3

We would like to characterize tick bites risk in a study area using some observational and geospatial datasets. We can formulate this problem both as a regression or classification.

True

False

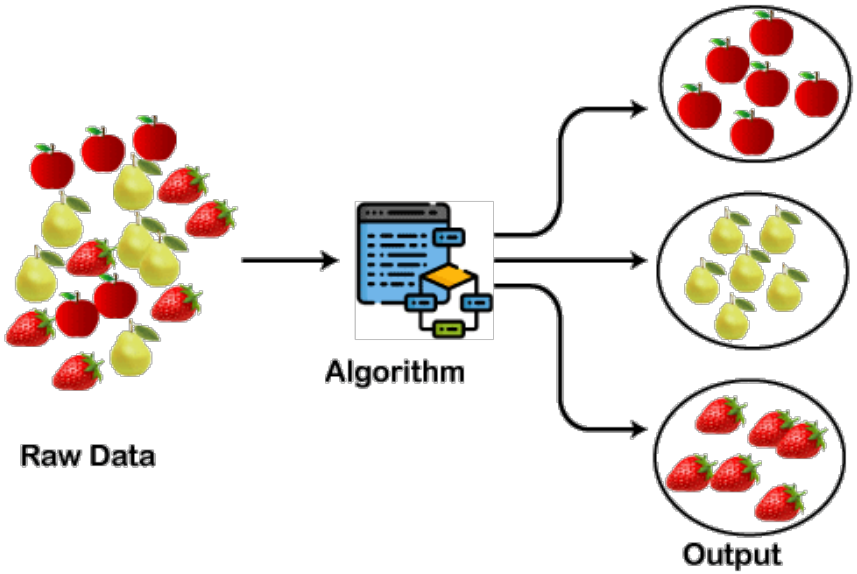
| Target variable | |
|---|---|
| Categorical | Continuous |
| 1: Very low risk<br>2: Low risk<br>3: Moderate risk<br>4: High risk<br>5: Very high risk | A probability or a score, the range could be:<br>0.0 to 1.0 |

# Unsupervised Learning

# (Clustering)

# Clustering

- *Clustering* is an important task in data mining that aims at **identifying groups of elements** that are similar among themselves but dissimilar to the elements in other groups.
- It provides a **high-level abstraction** of the data, which facilitates the **extraction of useful information**.




https://www.javatpoint.com/clustering-in-machine-learning

# Unsupervised Learning - clustering

- Contrarily to classification/regression task there is **NOT a target**
- It is suitable in problems where we have **unlabeled objects** or where the process of labelling is expensive (time / money).
- It is often done in **combination with the E(S)DA** to get a better grip on the objects at hand

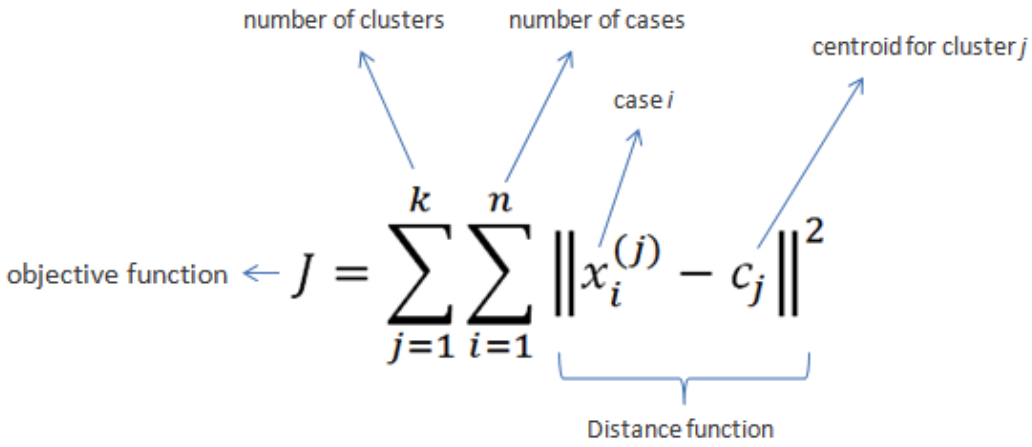
# K-means

- One of the most **popular** clustering algorithms
- It is an **iterative** algorithm that first assigns points to clusters (**k classes**) and then recomputes the centers of those clusters (means) until finding an optimum solution.

- Each point belong to one and only one cluster
- Each cluster is represented by the cluster mean (centroid)

# K-means

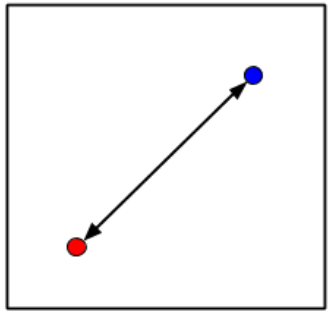
- K-means minimizes the total squared distance between each input point (xi) and its cluster center (cj)



# Common Distance Measures in Data Science

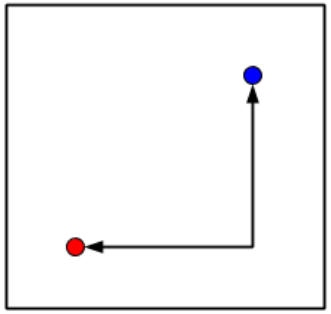
- Euclidean distance
  - The most common
- Manhattan distance
  - Approximation to Euclidean distance and cheaper to compute
  - Sum of the absolute differences of their Cartesian coordinates
- Minkowski distance
  - A generalization of both the Euclidean & Manhattan distance

Euclidean



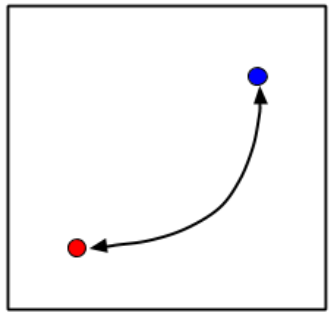
$$d(x,y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan



$$d(x,y) = \sum |x_i - y_i|$$

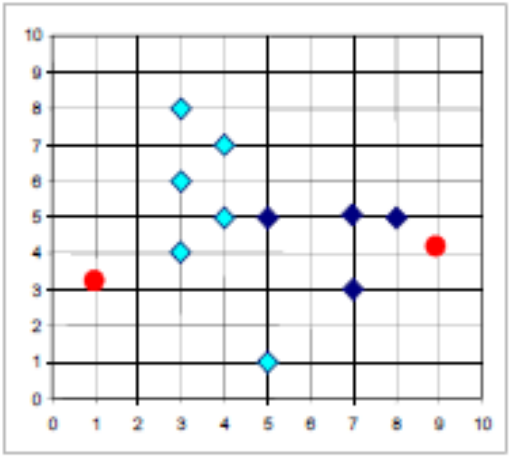
Minkowski



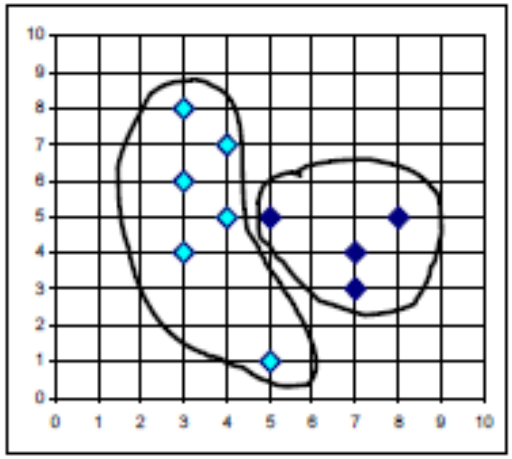
$$d(x,y) = \sqrt[h]{|x_1 - y_1|^h + |x_2 - y_2|^h + \cdots + |x_p - y_p|^h}$$

# K-means

1. **Initialization:** arbitrary initialization of the centers

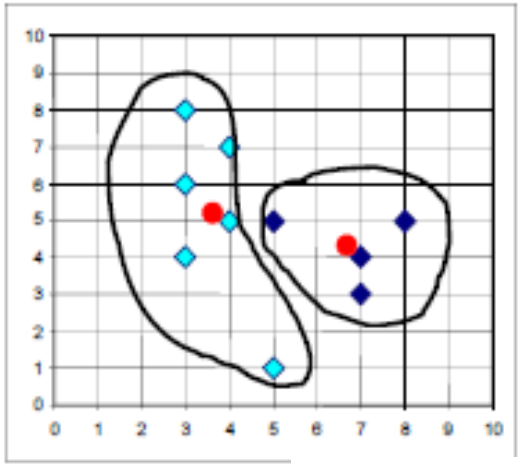



2. **Data assignment:** each point is assigned to its closest cluster (center). Ties are broken by randomly assigning the point to one of the clusters. This yields a partitioning of the data.

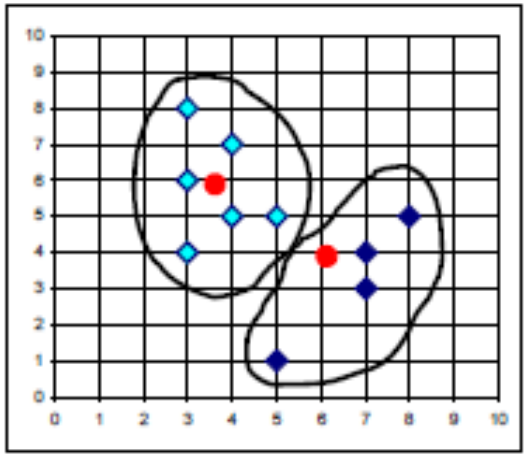


# K-means

3. **Relocating "centers":** each cluster representative is moved to the center of the points assigned to it.

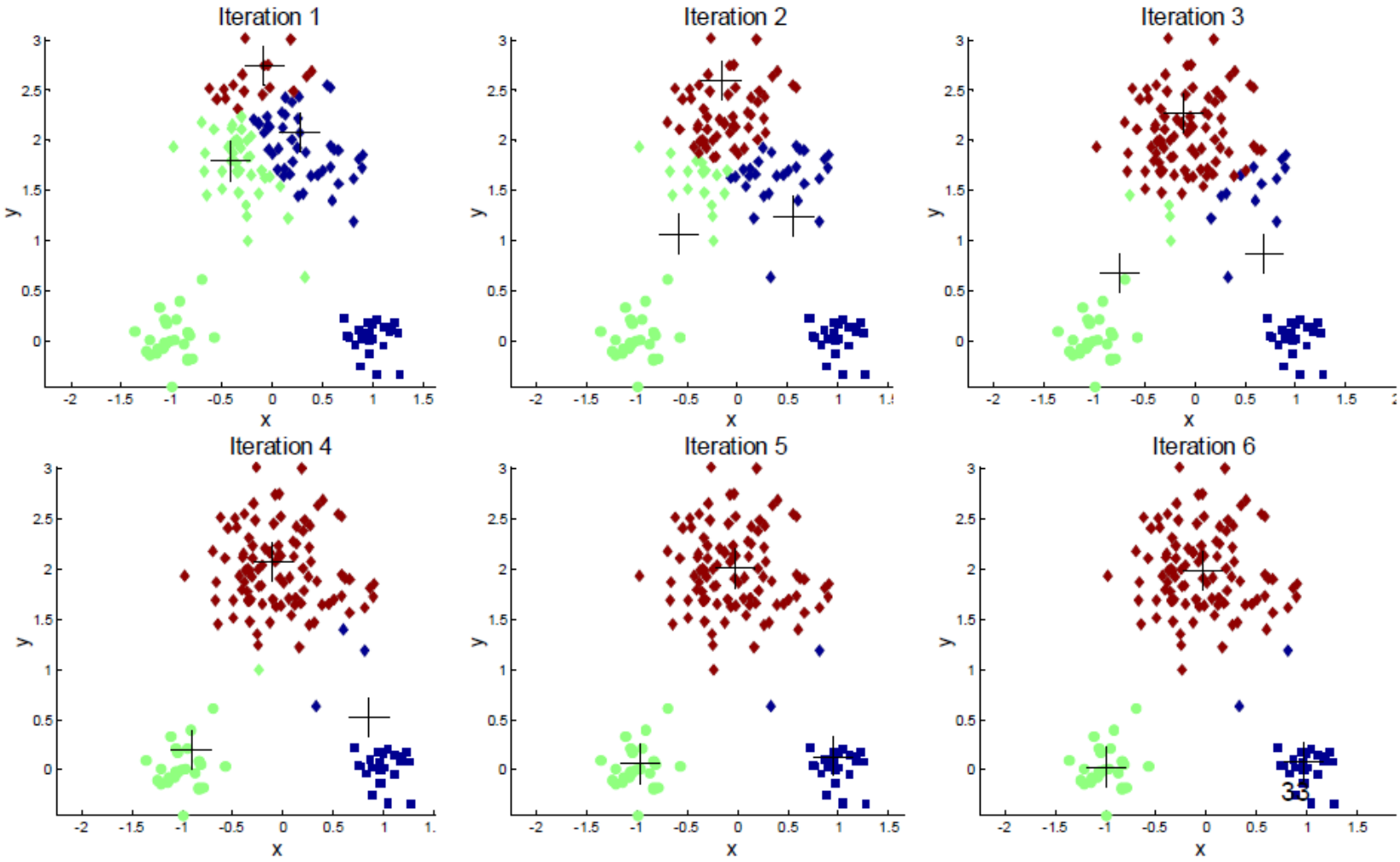



4. **Repeat 2 and 3** until the centers do not longer change.




*Each iteration requires N*k comparisons. It takes long time for large datasets*

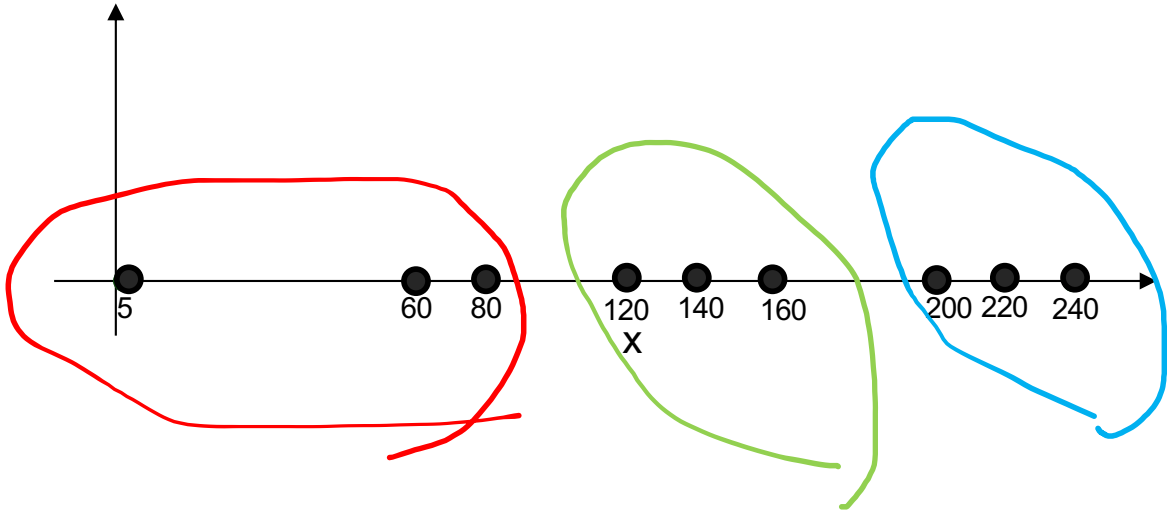
# A K-means example of 3 clusters



# K-means Limitations

Dataset

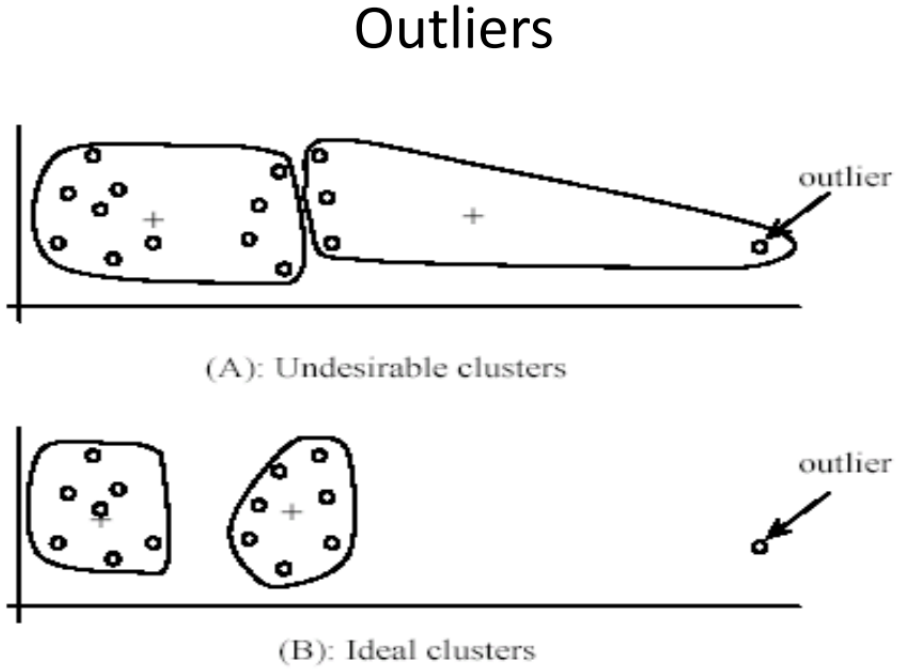
| X |
|---|
| 240 |
| 80 |
| 140 |
| 220 |
| 160 |
| 60 |
| 200 |
| 5 |
| 120 |



# K-means Limitations

The algorithm is sensitive to outliers

Clusters centers are computed using the "mean" function which is sensitive to outliers



# Kmeans limitations

It is quite sensitive to the initial set-up (location of cluster centers) which might lead to finding local minima instead of the absolute minimum.

Better to run multiple times (different initializations) or look for a robust way of initializing the algorithm.

Choosing the value of k is difficult (unless you have a priori knowledge about the “natural” number of groups present in the data).

Silhouette method

Elbow method

EDA?

- The **Silhouette method** is a technique used to evaluate the quality of clustering in unsupervised machine learning. It provides a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Here's how it works:

1. **Silhouette Score:**
    - For each data point, calculate the following:
        - **a(i):** The average distance from the data point to other points in the same cluster (intra-cluster distance).
        - **b(i):** The smallest average distance from the data point to points in a different cluster (inter-cluster distance).
    - The silhouette score for a data point i is given by:
        - **Silhouette Score(i)** = ( **b(i)−a(i)** ) / **max(a(i),b(i))**
    - The silhouette score **ranges** from -**1 to 1**:
        - A high positive value indicates that the data point is well-clustered within its own cluster.
        - A value close to 0 suggests that the data point is on the boundary between clusters.
        - A negative value indicates that the data point may be assigned to the wrong cluster.
2. **Overall Silhouette Score:**
    - Calculate the average silhouette score across all data points to assess the overall clustering quality.
    - Higher overall silhouette scores indicate better-defined clusters.
3. **Interpretation:**
    - If the silhouette score is close to 1, the clustering is effective.
    - If the score is near 0, the clusters may overlap or be poorly separated.
    - Negative scores indicate incorrect clustering.

- The **Elbow method** is a technique used to determine the optimal number of clusters (k) in a K-means clustering algorithm. Here's how it works:

    1. **K-means Clustering:**
        - K-means is an unsupervised learning algorithm that partitions data into k clusters based on similarity.
        - It assigns each data point to the nearest centroid (mean) of its cluster.
    2. **Elbow Method Steps:**
        - Run the K-means algorithm for different values of **k** (usually a range from 1 to a maximum value).
        - For each **k**, calculate the sum of squared distances (SSD) between data points and their assigned centroids.
        - Plot the SSD against the number of clusters (k).
        - Look for the "elbow point" on the plot—the point where the SSD starts to decrease at a slower rate.
    3. **Interpretation:**
        - The elbow point represents the optimal number of clusters.
        - It indicates the point where adding more clusters doesn't significantly reduce the SSD.
        - Choose **k** at the elbow point.
    4. **Example:**
        - If the plot shows a clear bend at **k=3**, it suggests that 3 clusters are appropriate for the data.

- Remember that the Elbow method is a heuristic, and sometimes the "elbow" isn't very pronounced. In such cases, consider other evaluation methods (like silhouette score) and domain knowledge.

# Team based learning

Ghelgheli decided to improve his customers' satisfaction. He heard a lot about a magic tool that could learn from data and help him tailor marketing strategies and promotions to different customer groups.

Intrigued, Ghelgheli began exploring this tool. He applied it to a set of data and discovered that his customers fell into three main categories: "Daily Commuters", "Weekend Relaxers" and "Tea Enthusiasts". The "Daily Commuters" were regular visitors who stopped by for a quick tea on their way to work. The "Weekend Relaxers" visited the teahouse mainly on weekends. The "Tea Enthusiasts" were passionate about trying different varieties and learning about tea.

After the analysis, Ghelgheli introduced a program offering discounts on morning tea purchases targeting "Daily Commuters". For the "Weekend Relaxers", he organized weekend events and introduced special weekend-only blends. To satisfy "Tea Enthusiasts", he started hosting monthly tea-tasting events and workshops. As a result, Ghelgheli's teahouse became even more popular.

- **Which data and methods do you think Ghelgheli utilized for his analysis?**
- **Can you list the potential steps that Ghelgheli took for such an analysis?**
- **Can you provide some real-life examples similar to Ghelgheli's experience?**

# Conclusion

- Data: Customer purchase history (e.g., items bought, frequency of purchases), demographics (e.g., age, occupation, location), customer feedback and reviews, and visit times and days.

- Methods: Data cleaning and preparation, EDA (descriptive statistics and visualizations), Clustering algorithms (e.g., K-means clustering) to identify distinct customer groups

- Steps: Data collection, data cleaning, EDA, data standardization (if necessary), choosing a clustering algorithm (K-means), determining the number of clusters, applying the clustering algorithm, analyzing and interpreting the resulting clusters

# Conclusion

- Please reply to the questions and write your answers on the yellow post-it

- What is the most important idea/insight you will remember from today's lesson?
- What questions do you still have?

## Conclusion

I can **Explain** to peers

- the fundamentals of **Artificial Intelligence**
- the principles of **Machine Learning** methods
- the basics of **clustering** methods

not yet ☹ very well ☺