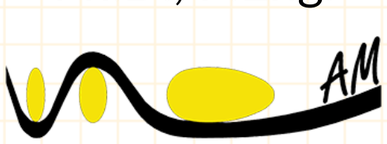
LECTURE 11
CLUSTERING

MANU 465

Ahmad Mohammadpanah

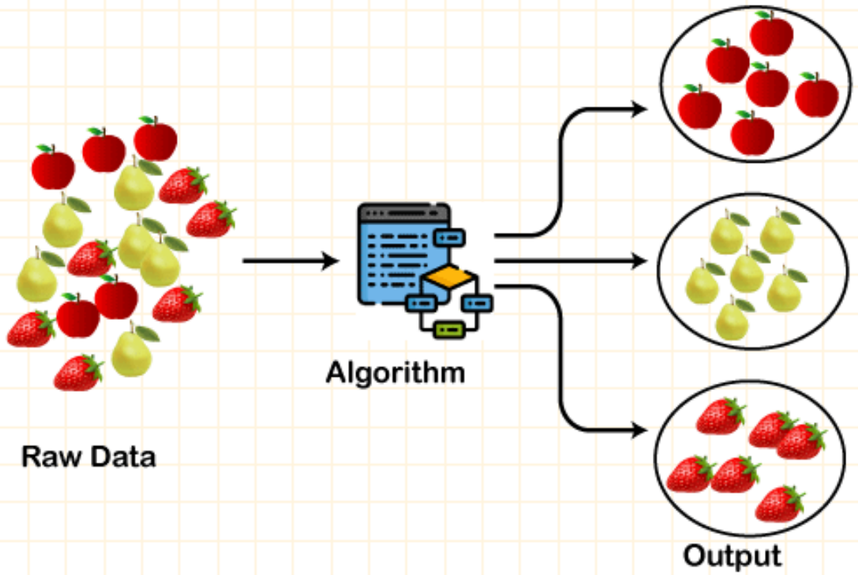
Ph.D., P.Eng.



AIntelligentManufacturing.com

Supervised vs. Unsupervised Learning:

- The main distinction between the two approaches is **the use of labeled datasets**. Supervised learning uses labeled output data, while an unsupervised learning algorithm does not.
- Unsupervised learning models, work on their own to discover the inherent structure of **unlabeled data**.
- Selection of any of these learning depends on the factors related to the structure and the use cases of the problem.



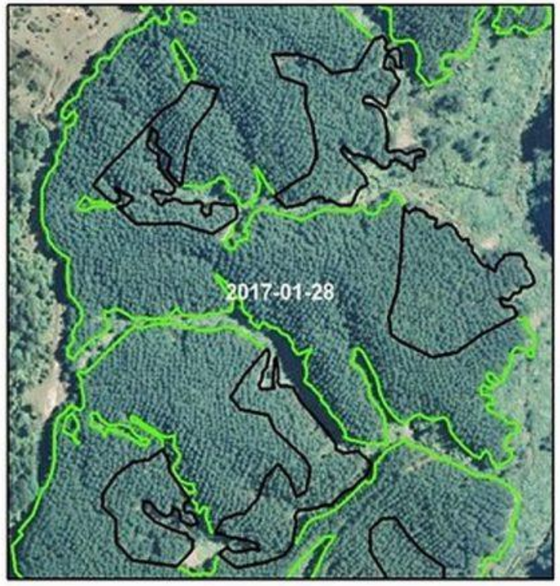
Examples of Supervised Algorithms: Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian KNN, etc.

Example of Unsupervised Algorithms: Clustering

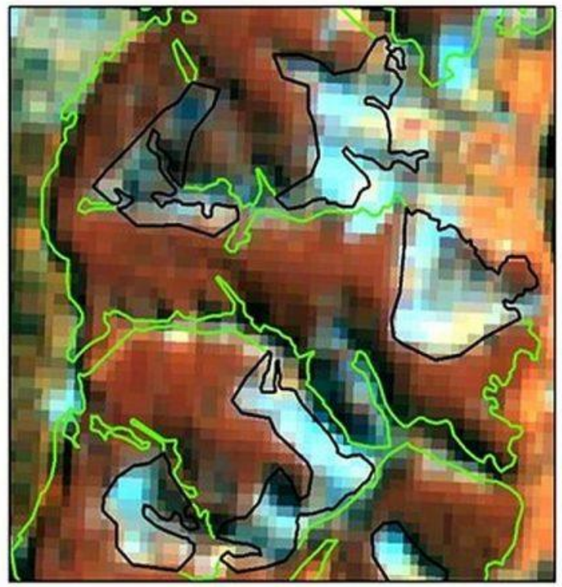
Applications:

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection.
- Analysis of satellite images

For example, you may want to measure how much total forest area there is in a region, color segmentation may be just fine.



Stocked 2017

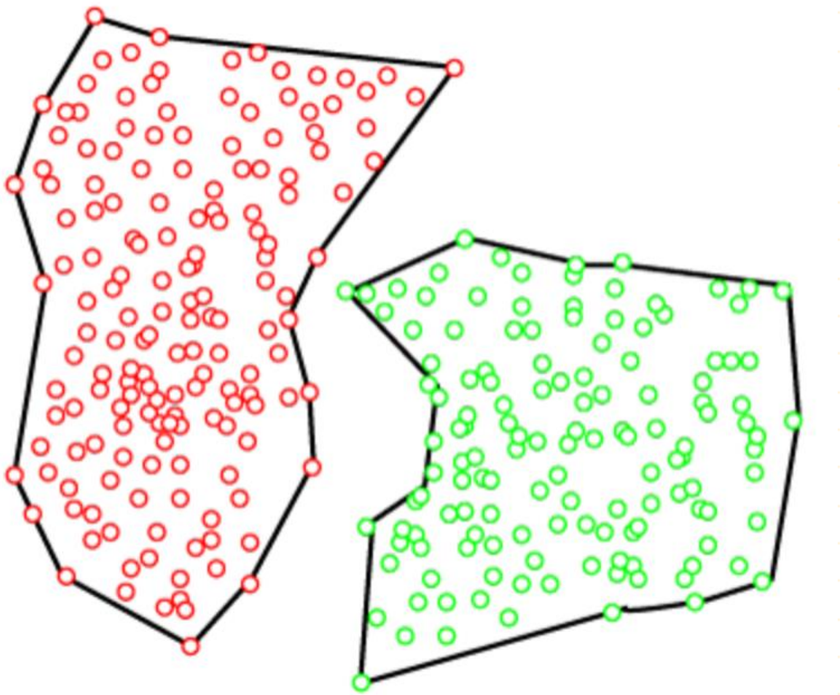


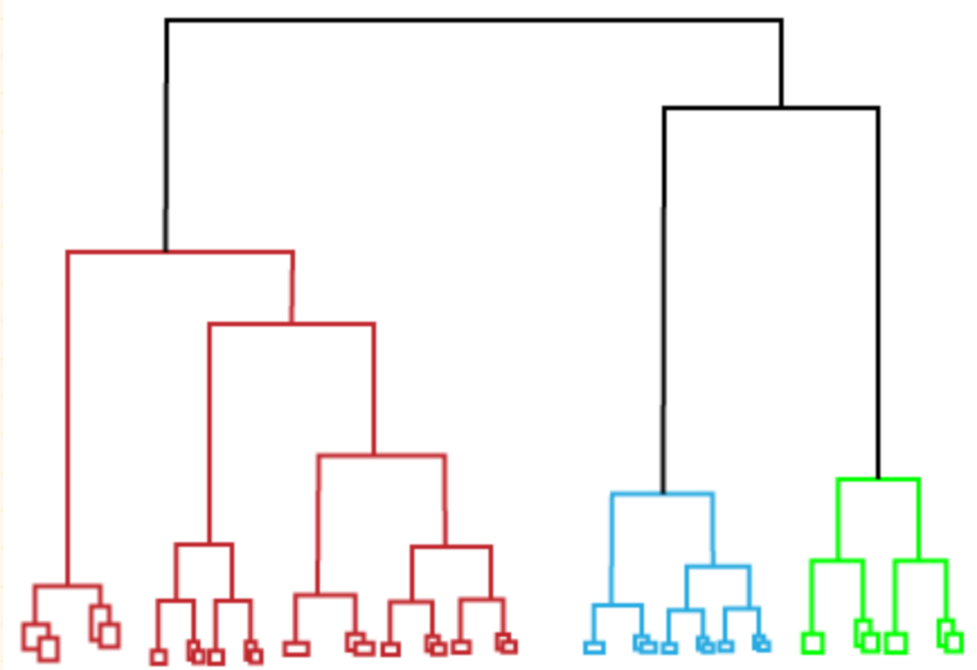
Damaged - Feb 2018 CPMS

https://foresttech.events/continuous-forest-monitoring-using-daily-satellite-imagery/

Common Algorithms:

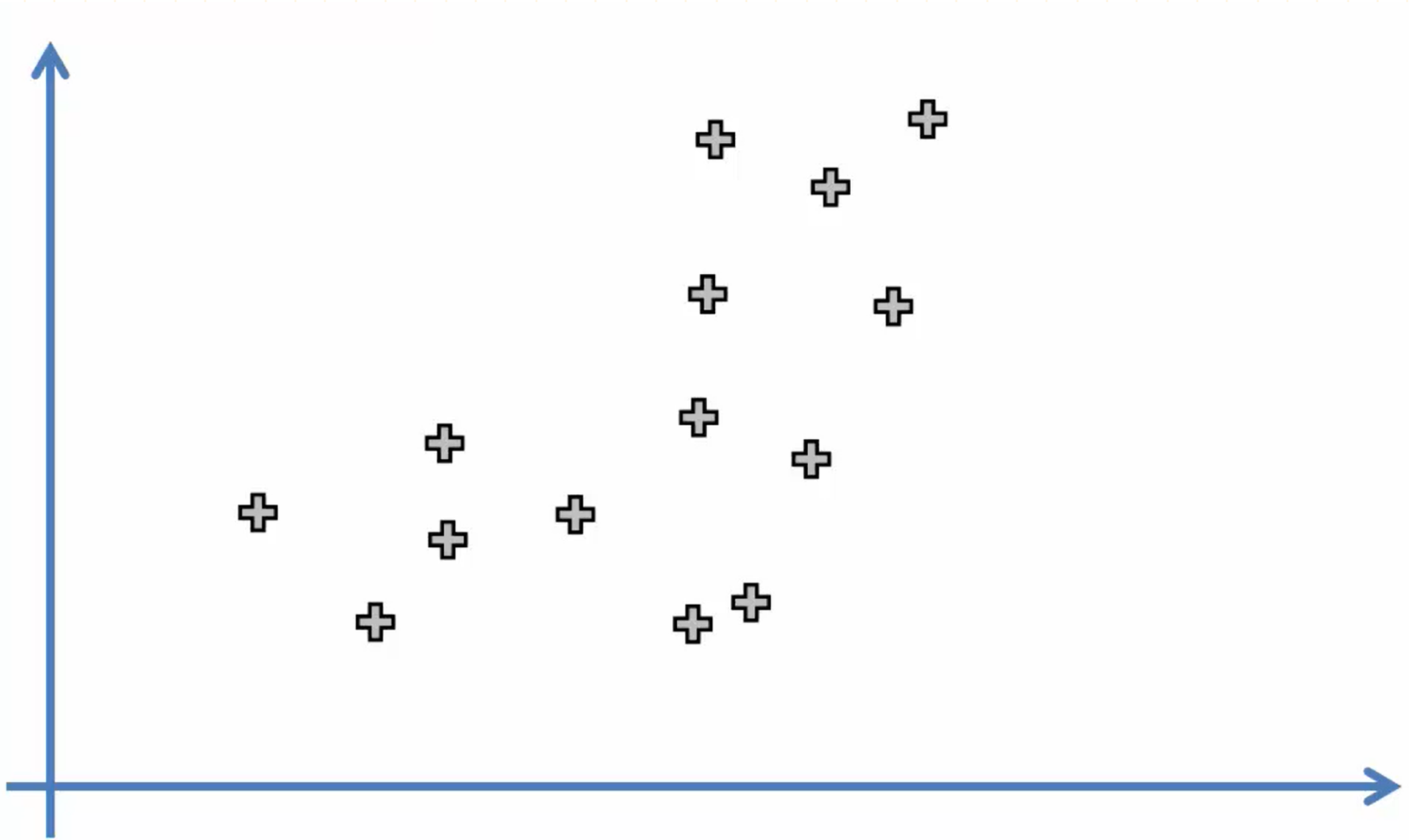
- K-Mean
- Gaussian Mixture
- Hierarchical





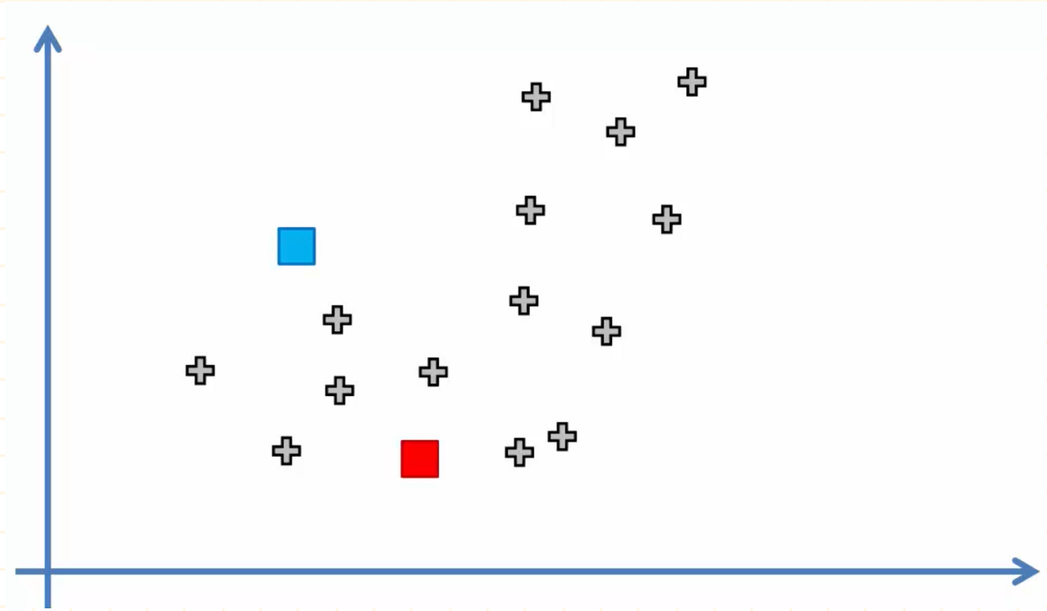
Check Canvas Examples 18, 19, 20, and 21

K-Mean Clustering

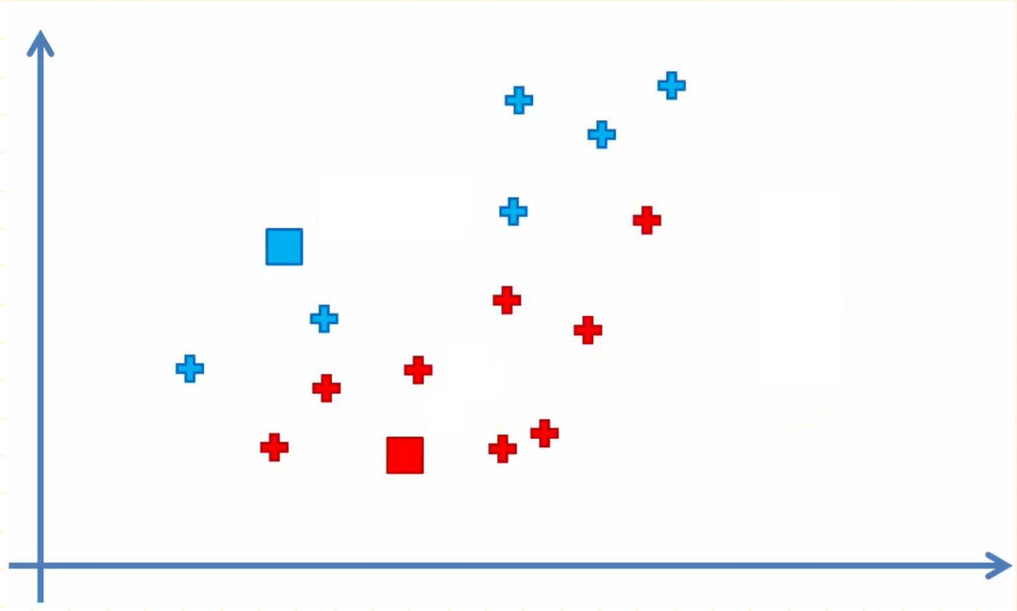


How does K-Mean Clustering work?

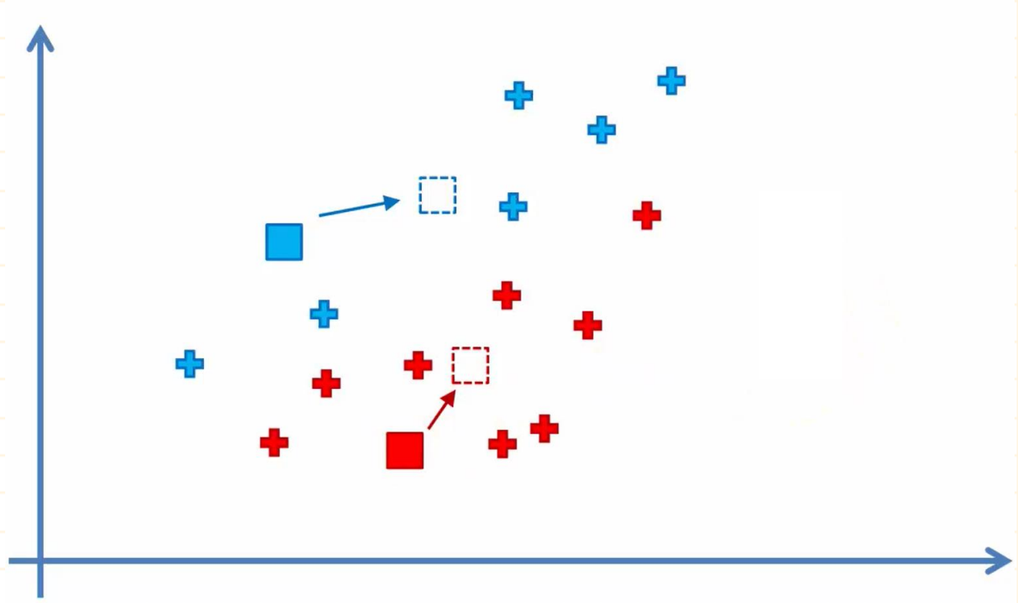
Step 1.



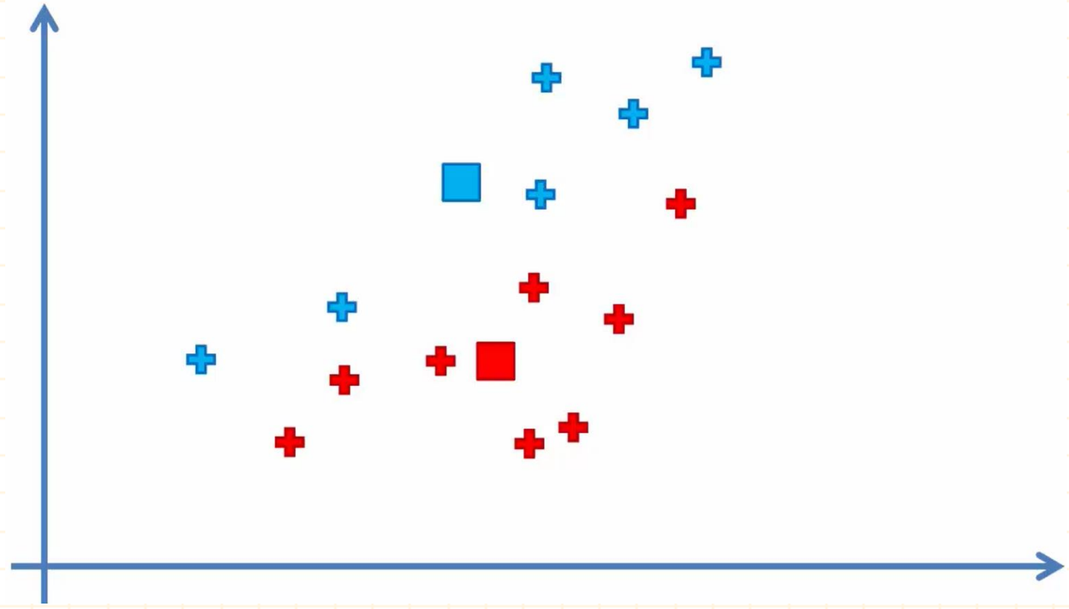
Step 2.



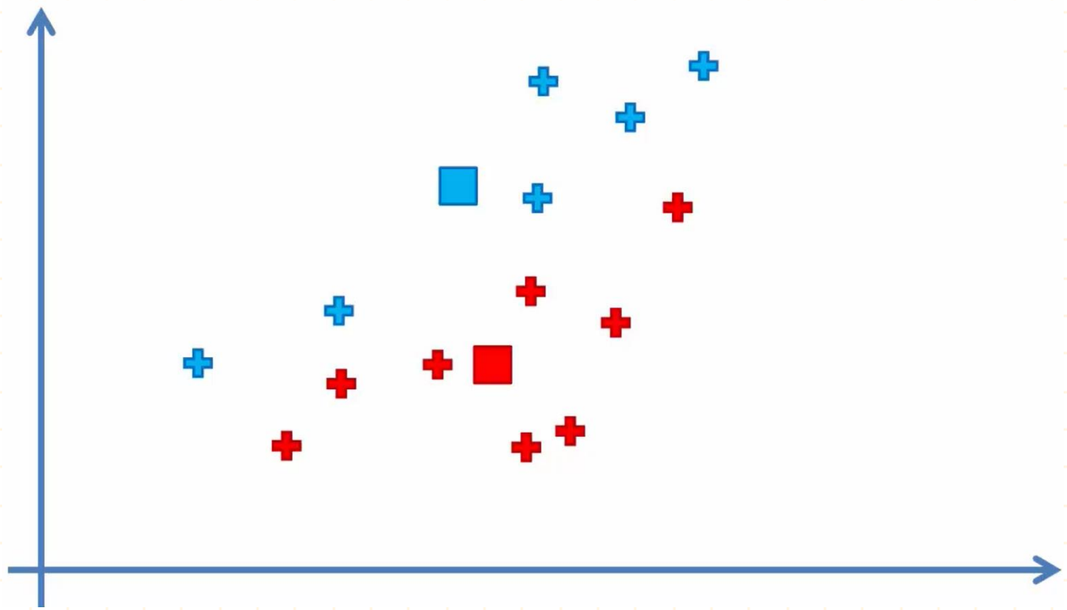
Step 3.



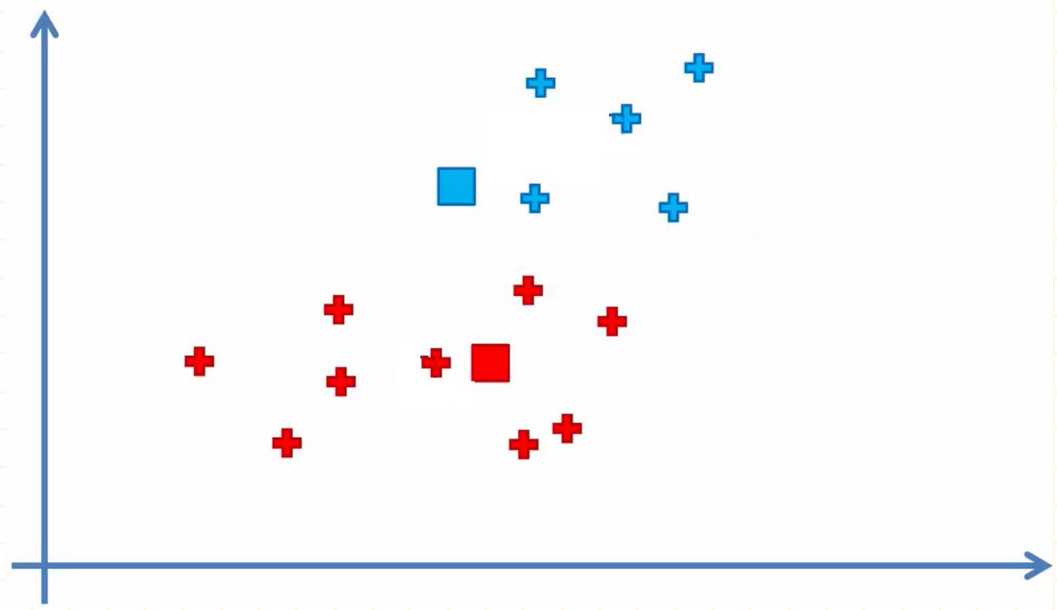
Step 4.



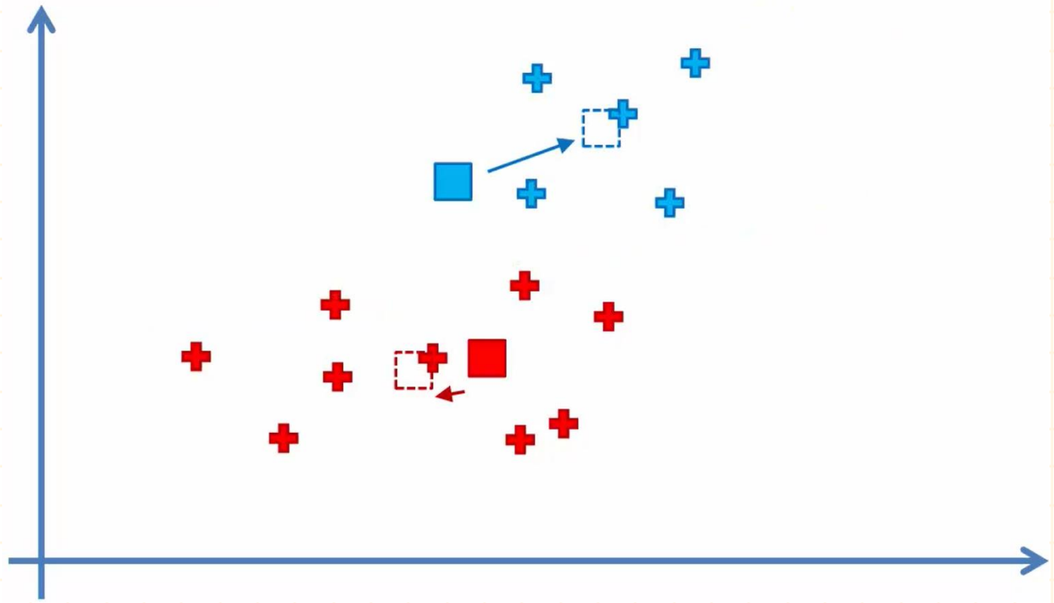
Step 5.



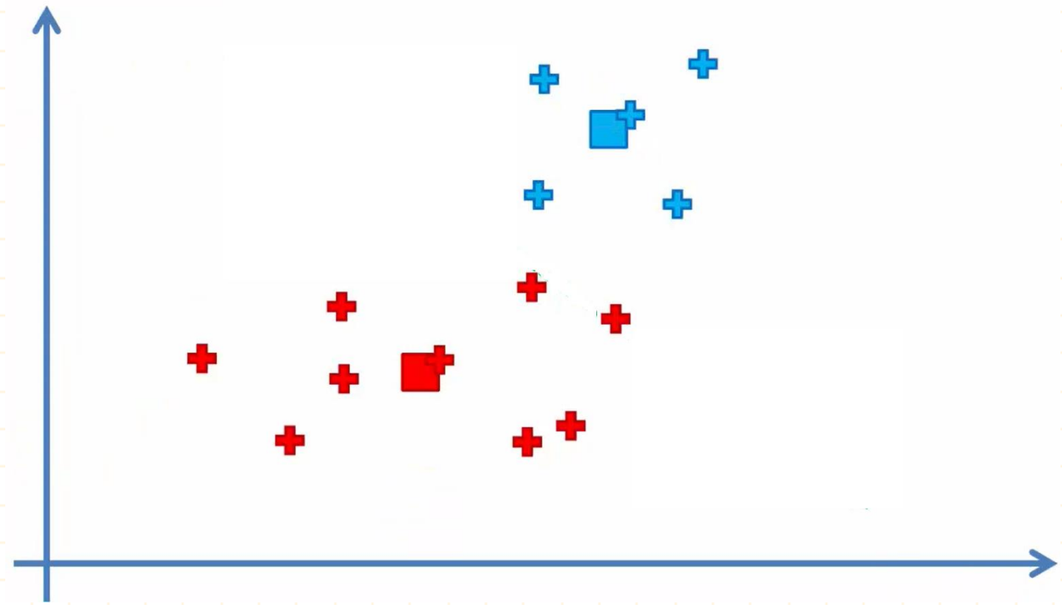
Step 6.



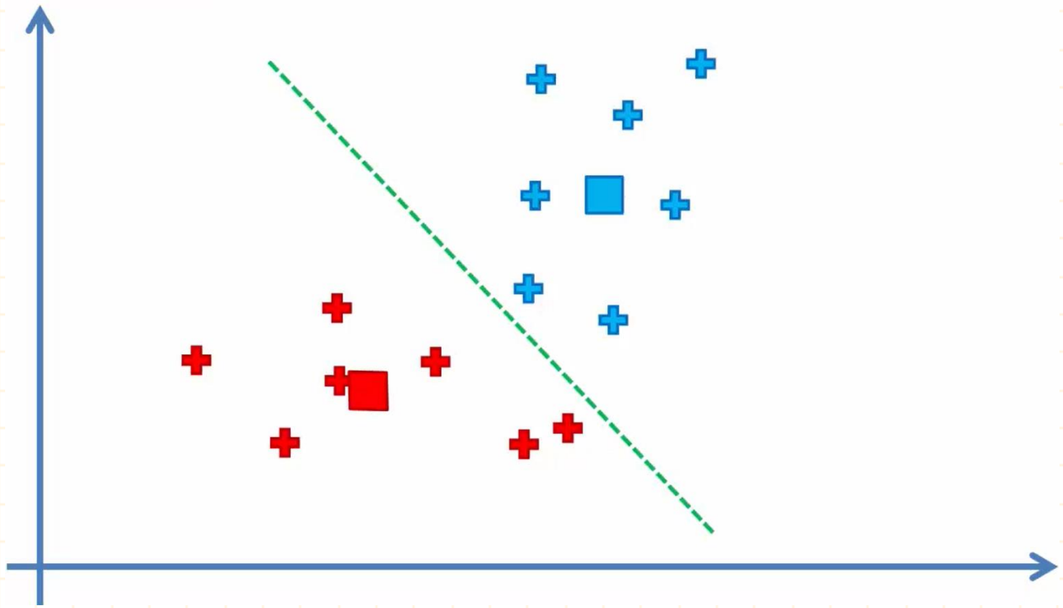
Step 7.



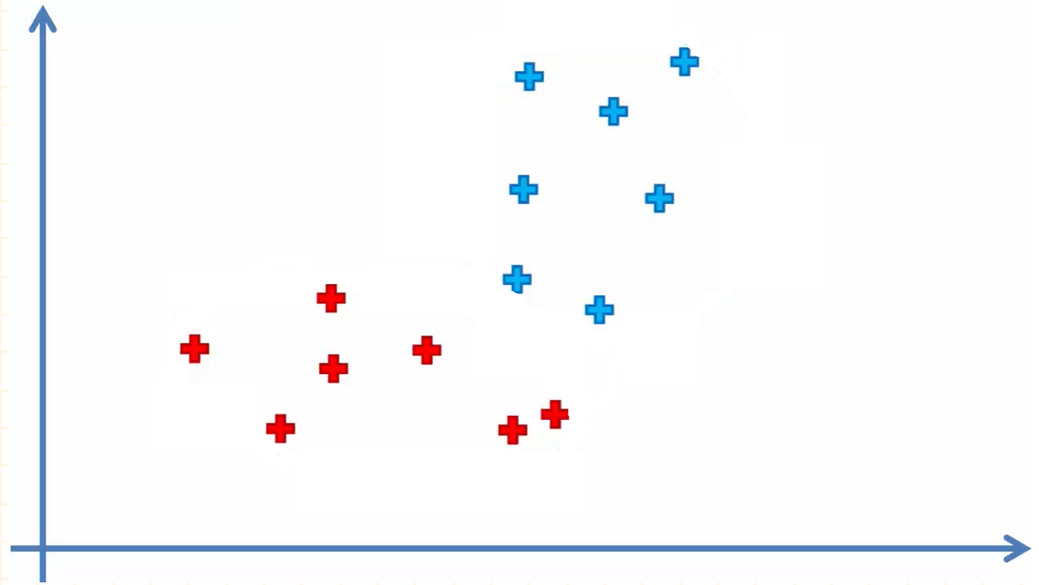
Step 8.



Step 9.



Step 10.



Example) Check Example 18 on Canvas.

Project Description

Thermoforming is usually used to manufacture relatively simple geometries from thin plastic sheets. For example, the lids of disposable coffee cups, pill blister packaging, inexpensive plastic packaging, and large items such as bathtubs and internal door liners for refrigerators. Thermoforming is also used to manufacture large components such as automotive interior panels and small boat hulls.

Watch this short video: https://www.youtube.com/watch?v=alq3RDZN4jo

Now, let's get back to ML.

This is how the coffee cup lid are produced in a factory: https://www.youtube.com/watch?app=desktop&v=YI3Cwyx1tR8

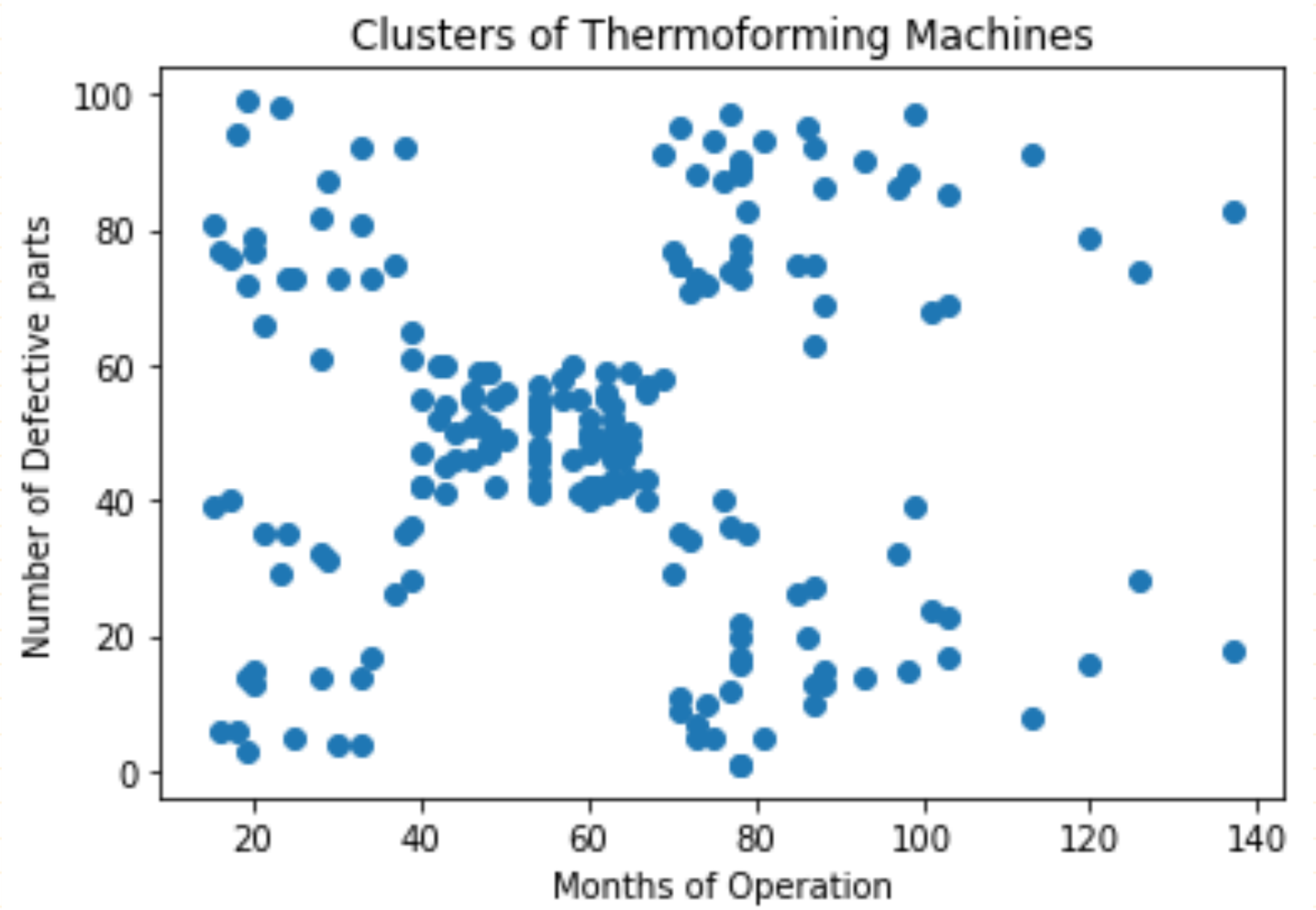
This factory has 200 thermoforming machines. Some of these machines are new and some are old from different brands. Each machine is identified with an ID number. The foreman is interested in analyzing the number of defective lids per day; to see if there is any pattern on how these machines perform.

The file 'ThermoFormingDefectiveParts.csv' summarizes the average number of defective parts per day and the age of the machine (in month) for each machine.

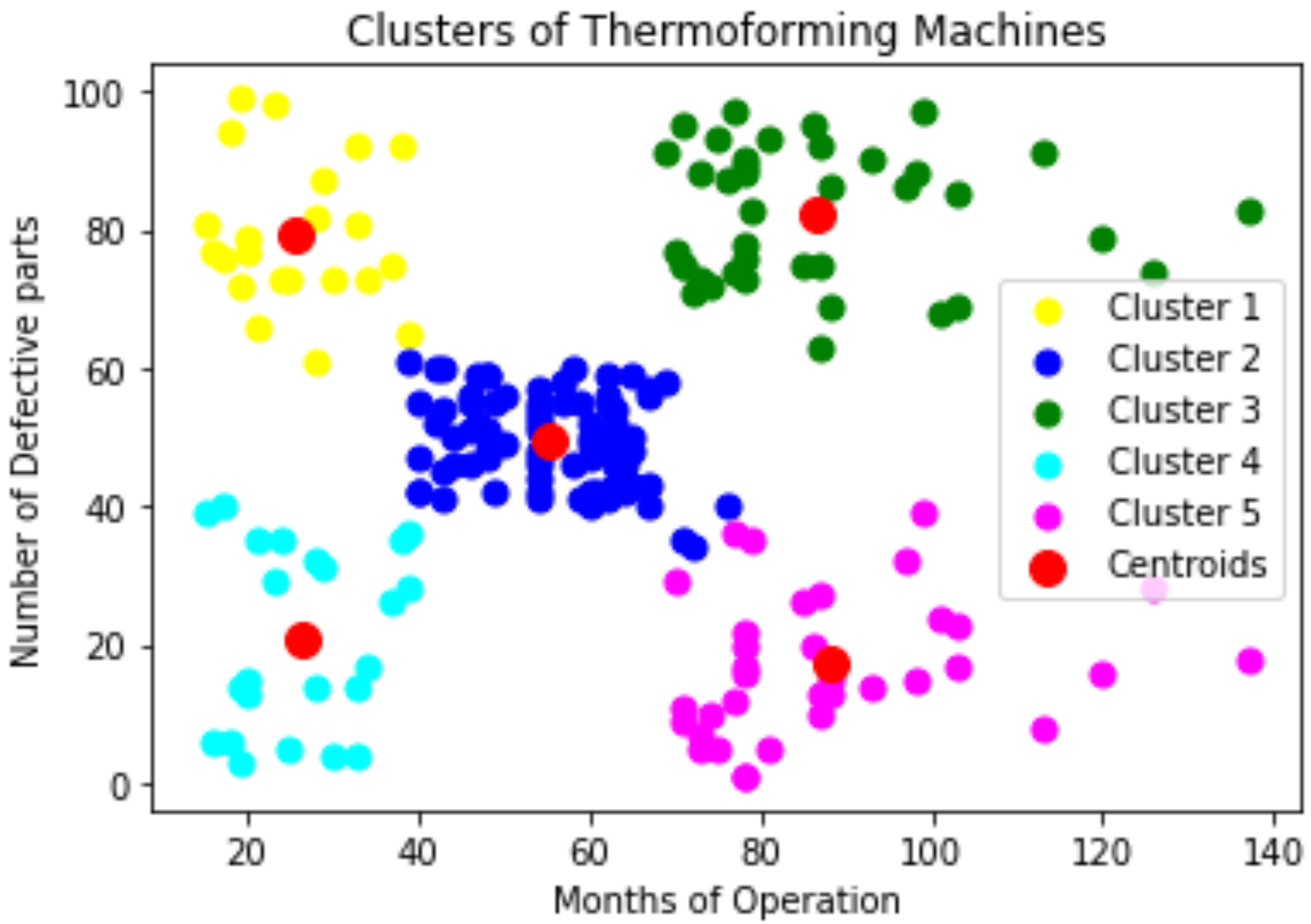
In order to start the investigation, the foreman wants to see if he can group these machines to few groups. Try to see if you could cluster these machines to an optimum number of clusters.

Check Example 18 Code.

Raw Data:



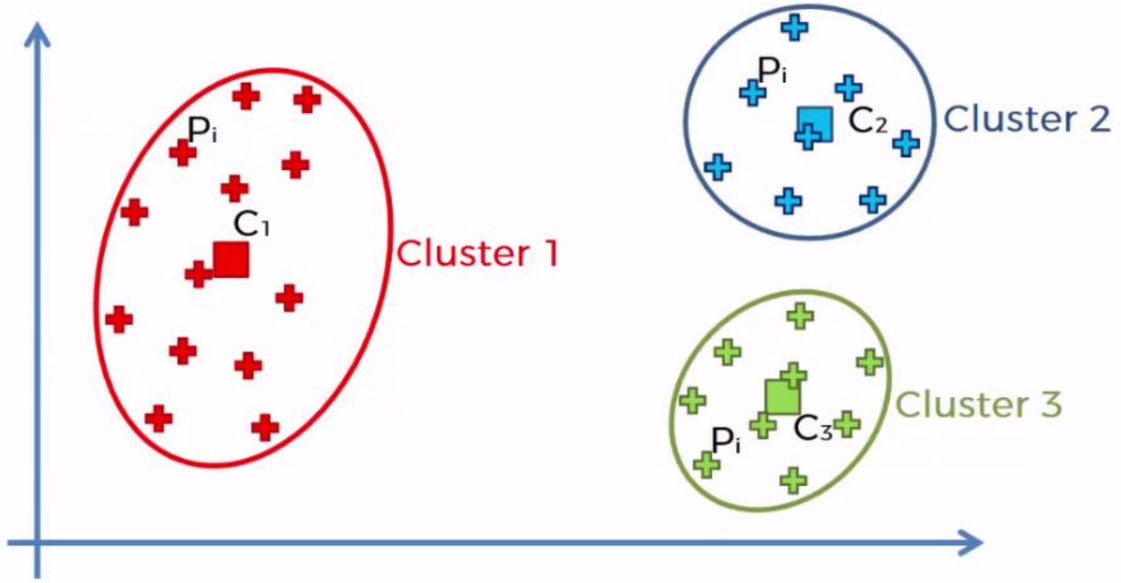
Solved by Python



Here comes, the $1,000,000 Question! What is the "Optimum Number" of Clusters?

Method 1. The Elbow

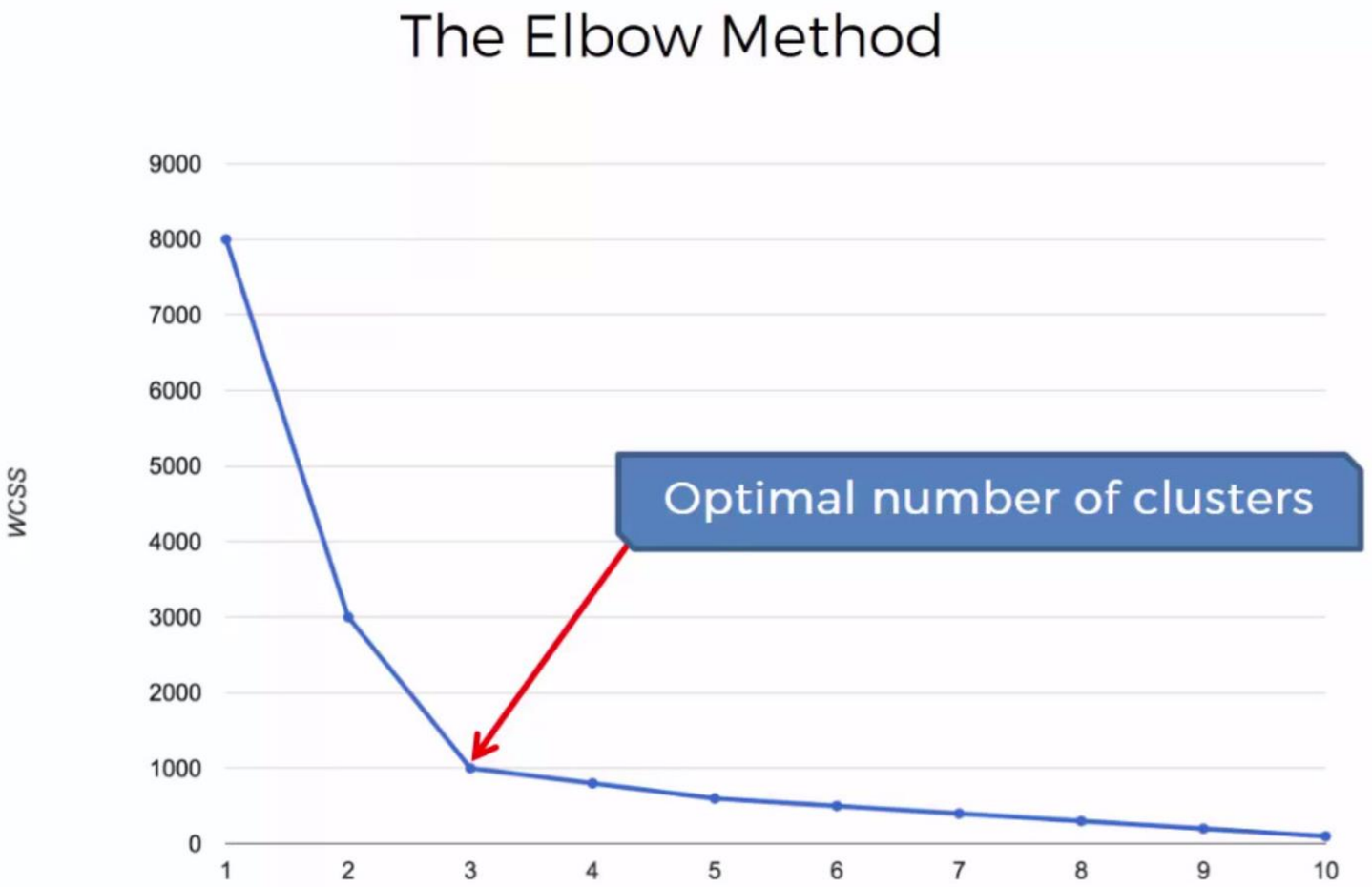
In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster.



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera

When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.



Method 2. Silhouette Coefficient

Silhouette Coefficient or Silhouette Score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

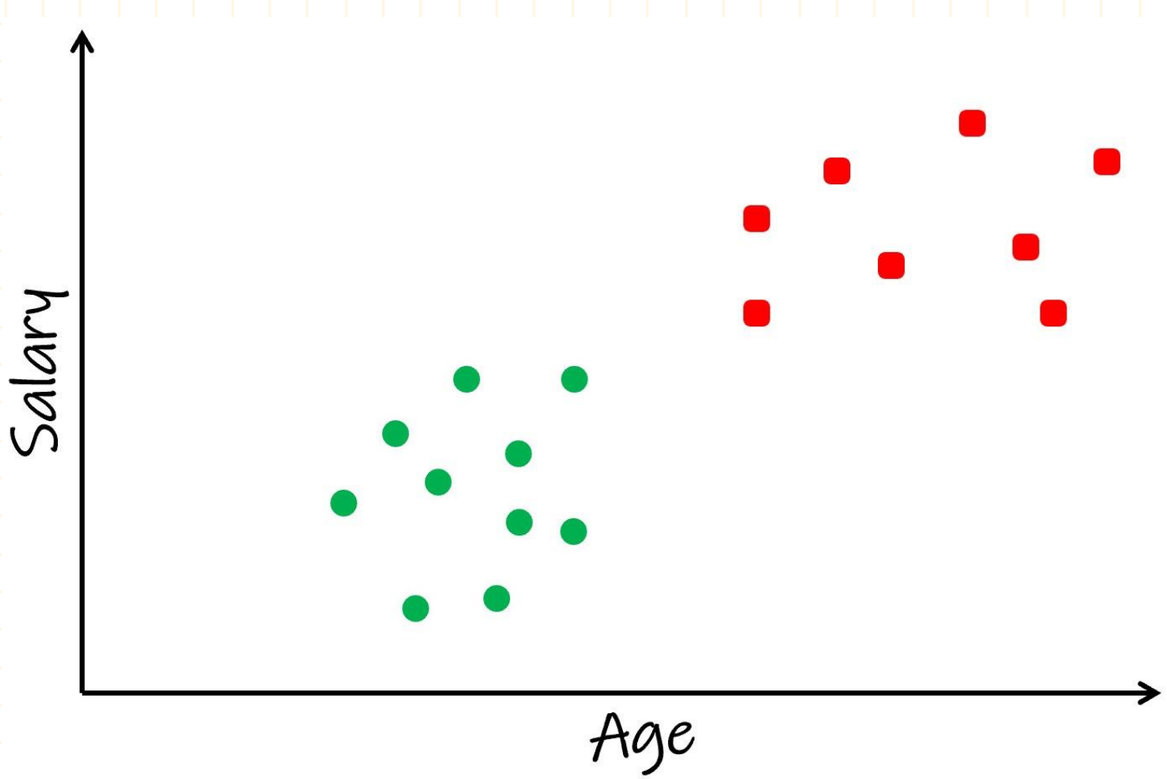
1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

$$\text{silhouette score} = \frac{b-a}{\max(a,b)}$$ (a= average distance between each point within a cluster. b= the average distance between all clusters.)

Source: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

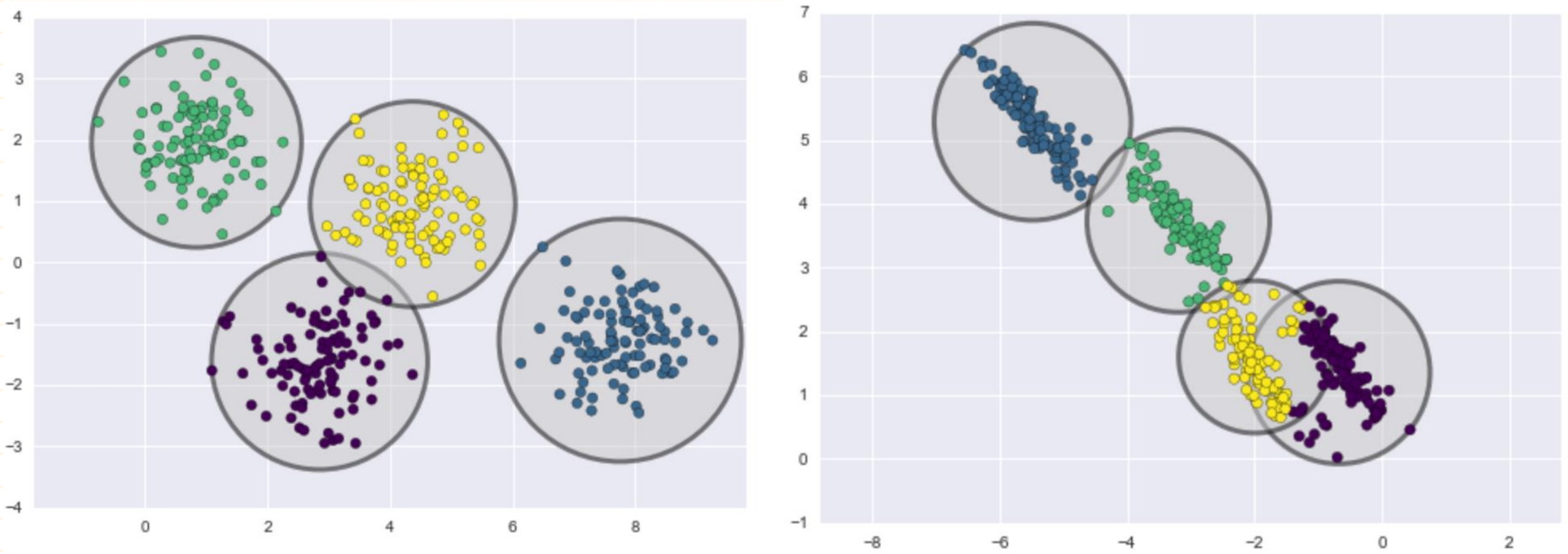


Gaussian Mixture Models for Clustering

Gaussian mixture models can be used to cluster unlabeled data in much the same way as k-means.

Advantages to using Gaussian mixture models over k-means:

1. The first difference between k-means and Gaussian Mixture Models is that K-Mean performs hard classification whereas the GMM performs soft classification. In other words, k-means tells us what data point belong to which cluster but won't provide us with the probabilities that a given data point belongs to each of the possible clusters.
2. Second, k-means does not account for variance.



Source: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

How does GMM work?

- If you are interested in detailed math and to learn “How the Algorithm Works”: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e
- Here with Animations: https://towardsdatascience.com/how-to-code-gaussian-mixture-models-from-scratch-in-python-9e7975df5252
- For in depth analysis: https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html

Example) Check Example 19 on Canvas. *ThermoformingGaussianMixture on Canvas*

Anomaly Detection

Gaussian Mixtures can be used for anomaly detection: instances located in low-density regions can be considered anomalies. You must define what density threshold you want to use. For example, in a manufacturing company that tries to detect defective products, the ratio of defective products is usually well-known. Say it is equal to 5%, then you can set the density threshold to be the value that results in having 5% of the instances located in areas below that threshold density.

Please see Example 20 on Canvas, CNC Machine Anomaly Detection

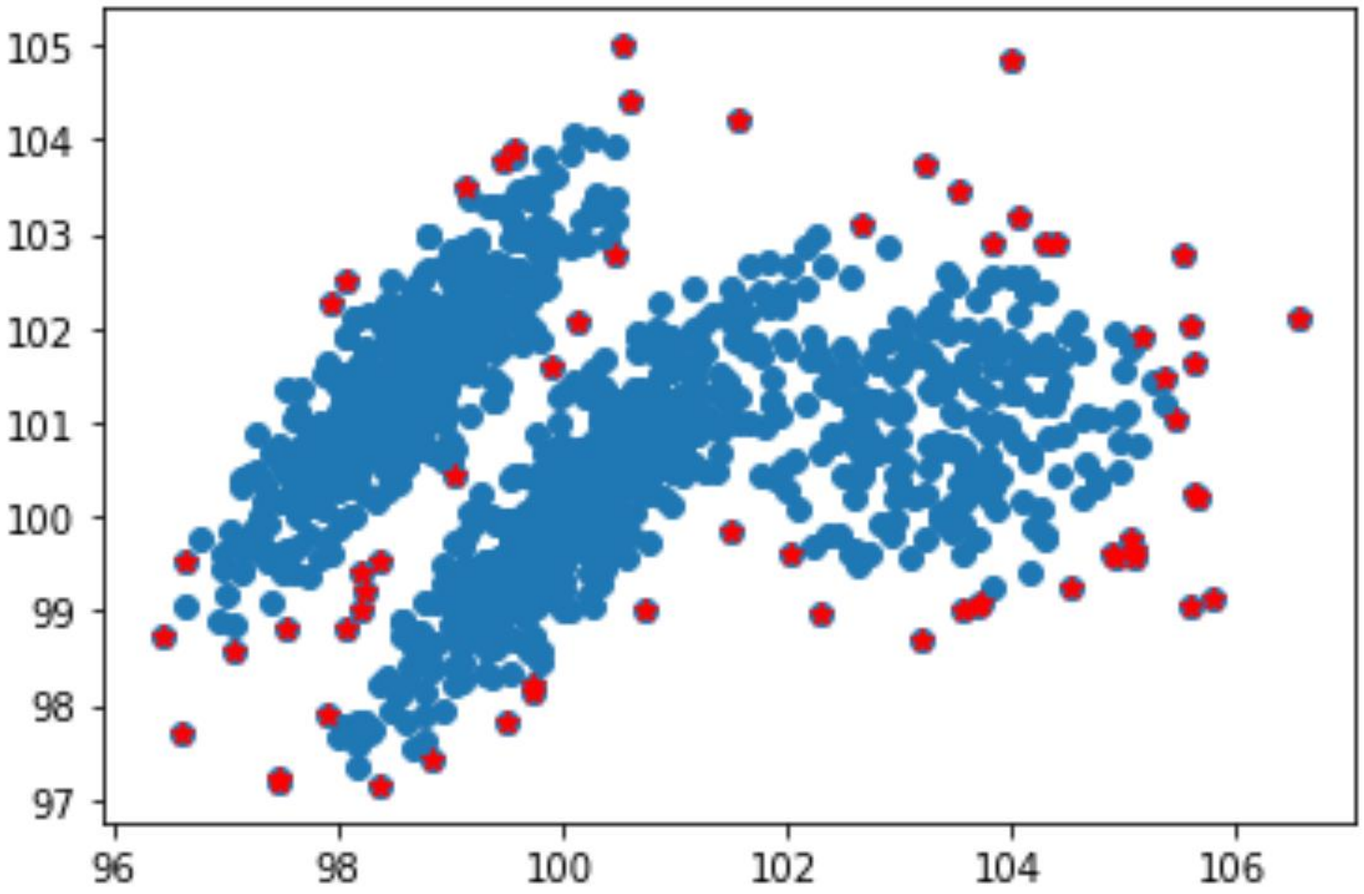


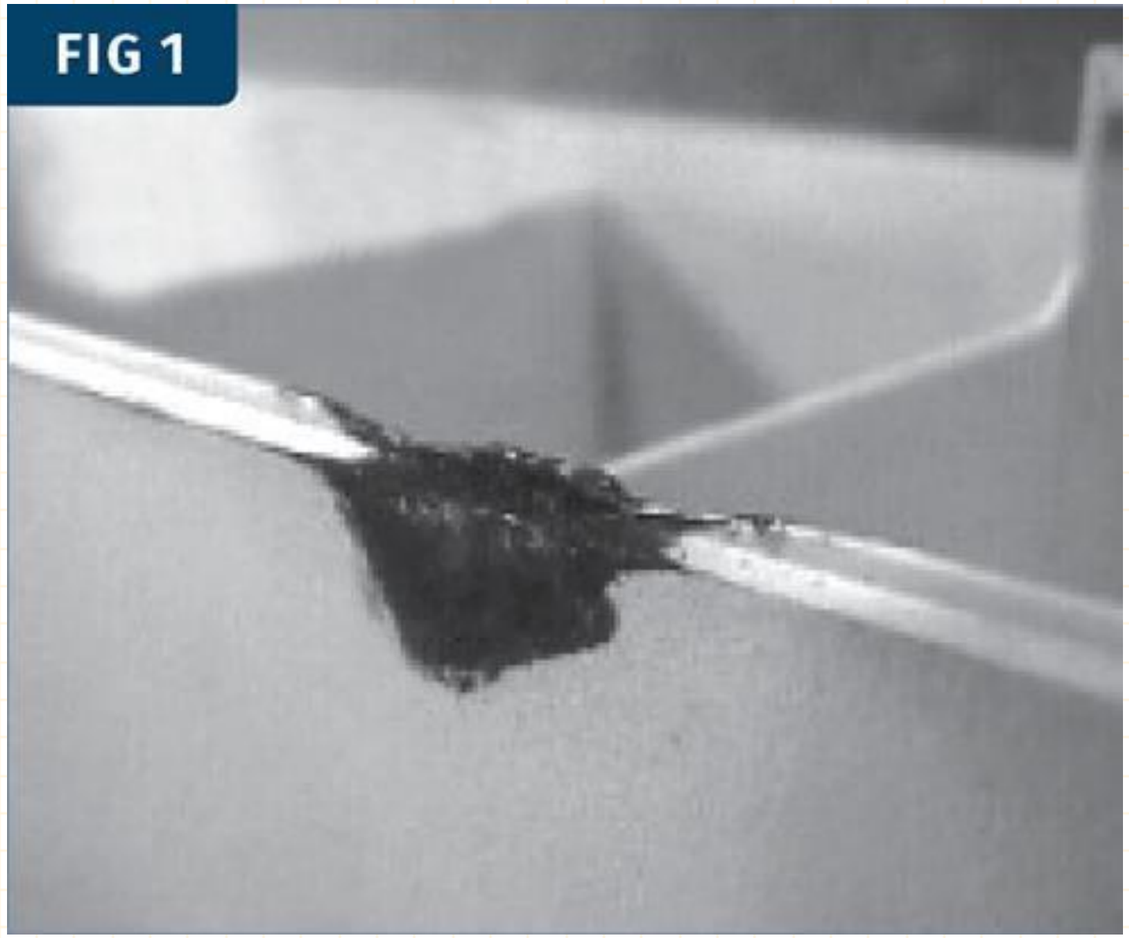
Image Segmentation by Clustering

Image segmentation is the task of partitioning an image into multiple segments. In this process, all the pixels that are part of the same object type get assigned to the same segments.

For example, in a self-driving car's vision system, all pixels that are part of a pedestrian's image are assigned to the "Pedestrian" segment, etc.

Another example, in the analysis of satellite images, for example, you may want to measure how much total forest area there is in a region, color segmentation may be just fine.

In Example 21 (on Canvas), we are interested in detecting the burning marks on parts manufactured by injection molding.



Other Topics (Optional to Study):

If interested in Image Edge Detection:

https://www.kdnuggets.com/2019/08/introduction-image-segmentation-k-means-clustering.html