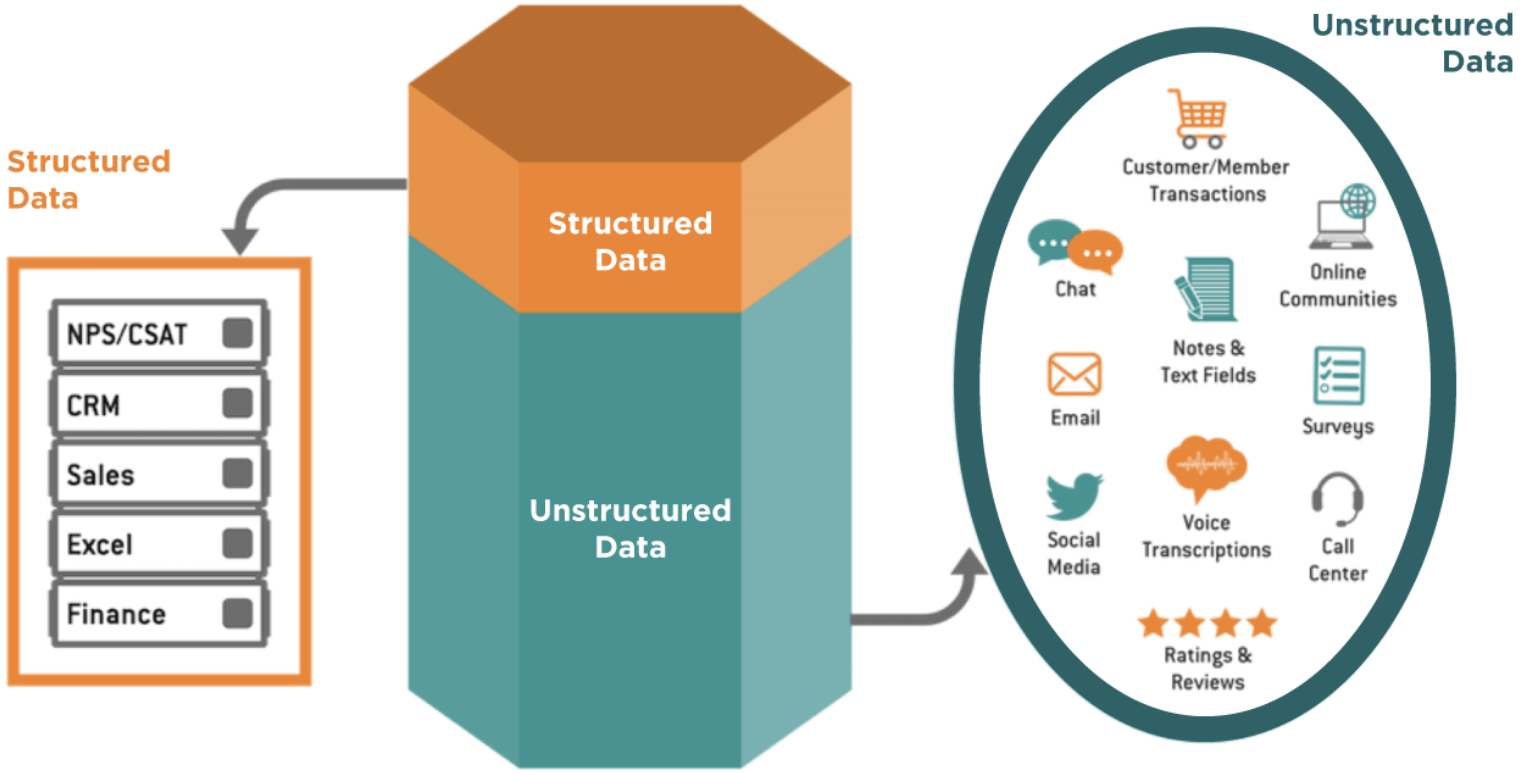
UnSupervised Learning

MUKESH KUMAR

Agenda

- Data
- Pre-requisites
- What is unsupervised learning
- K-means Clustering
- How it works?
- Elbow method
- Application of Clustering

Data


Structured Data
NPS/CSAT
CRM
Sales
Excel
Finance
Structured Data
Unstructured Data
Unstructured Data
Customer/Member Transactions
Chat
Online Communities
Notes & Text Fields
Email
Surveys
Social Media
Voice Transcriptions
Call Center
Ratings & Reviews

Volume Comparison

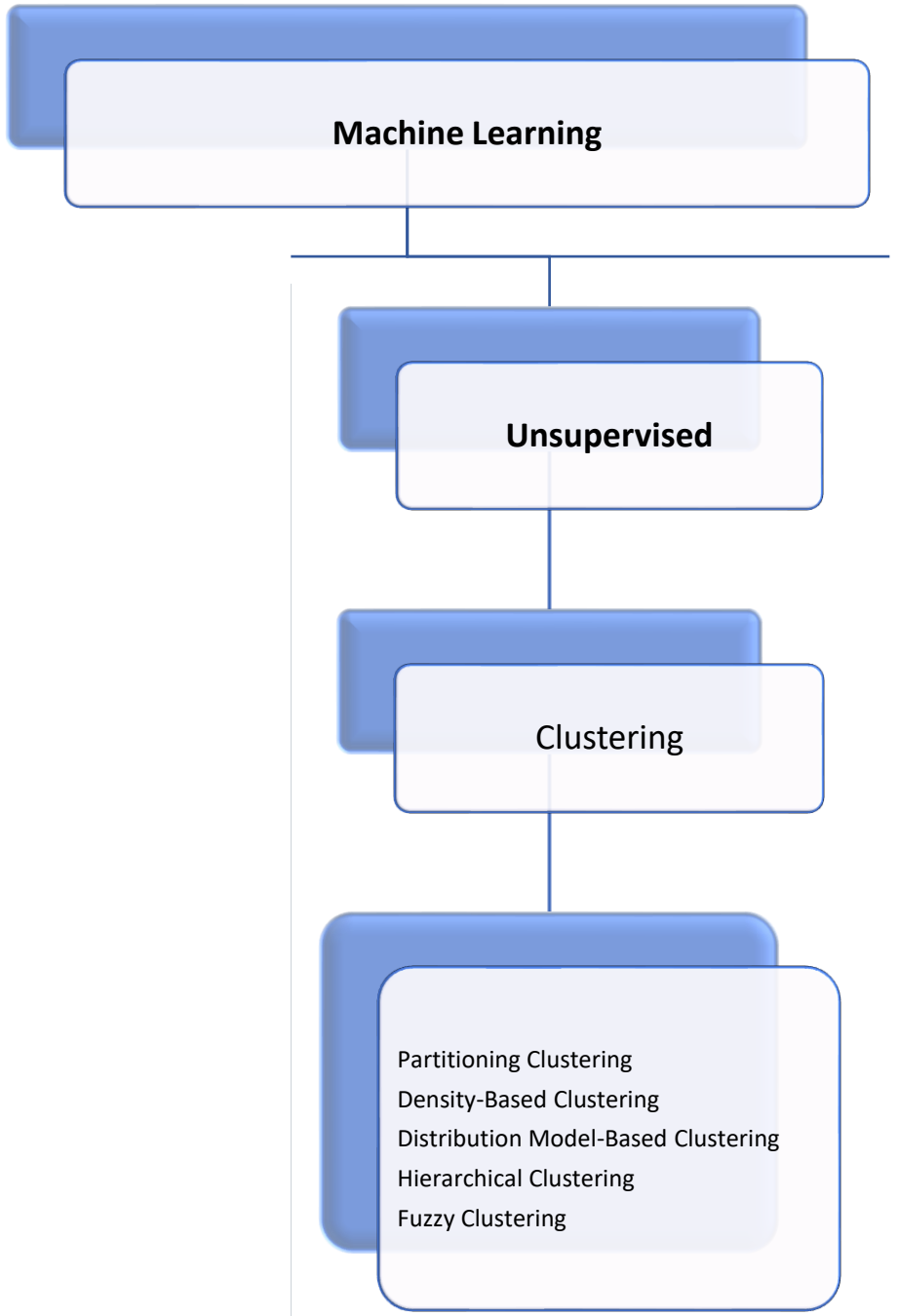
- Structured data makes up approximately 10-20% of the data generated today.
- Unstructured data constitutes the remaining 80-90%.
- Roughly 29 terabytes of data generated per second.
- over 500,000 tweets and 3.3 million Facebook posts are created every minute. This rapid growth underscores the importance of developing robust methods for managing and analyzing unstructured data effectively.

Structured Data

- **Definition:**
  - Structured data is data that is organized into a predefined format, such as rows and columns, making it easily searchable and analyzable. It follows a consistent order and can be efficiently managed in relational databases and spreadsheets.
- **Characteristics:**
  - Fixed schema: It has a defined structure and data types.
  - Searchable: Easy to query using languages like SQL.
  - Example Formats: CSV files, SQL databases, Excel sheets.
- **Examples:**
  - **Customer Information:** Names, addresses, phone numbers, and email addresses.
  - **Sales Data:** Transaction records, product IDs, quantities sold, and sales amounts.
  - **Sensor Readings:** Temperature, humidity, and pressure readings from IoT devices.

Unstructured Data

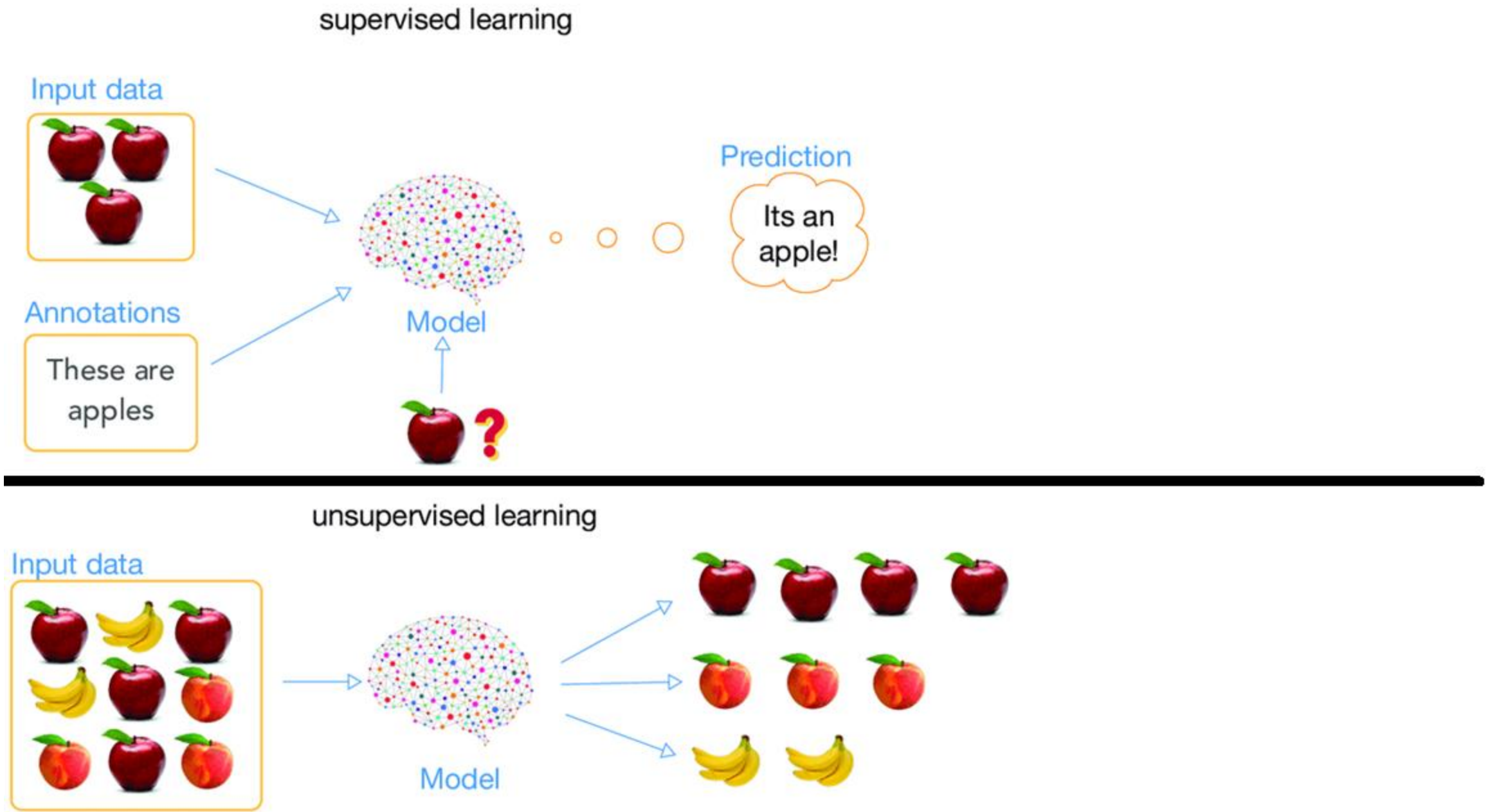
- **Definition:**
  - Unstructured data is data that does not have a predefined format or organization, making it more challenging to collect, process, and analyze.
- **Characteristics:**
  - No fixed schema: Data can be stored in various formats.
  - Difficult to search: Requires advanced techniques to query and analyze.
  - Example Formats: Text files, multimedia files, social media content.
- **Examples:**
  - **Text Documents:** Word documents, PDF files, and emails.
  - **Images:** Photos, scans, and medical images.
  - **Videos:** Recorded lectures, movies, and surveillance footage.
  - **Social Media Posts:** Tweets, Facebook updates, and Instagram photos.


Machine Learning
Unsupervised
Clustering
Partitioning Clustering
Density-Based Clustering
Distribution Model-Based Clustering
Hierarchical Clustering
Fuzzy Clustering

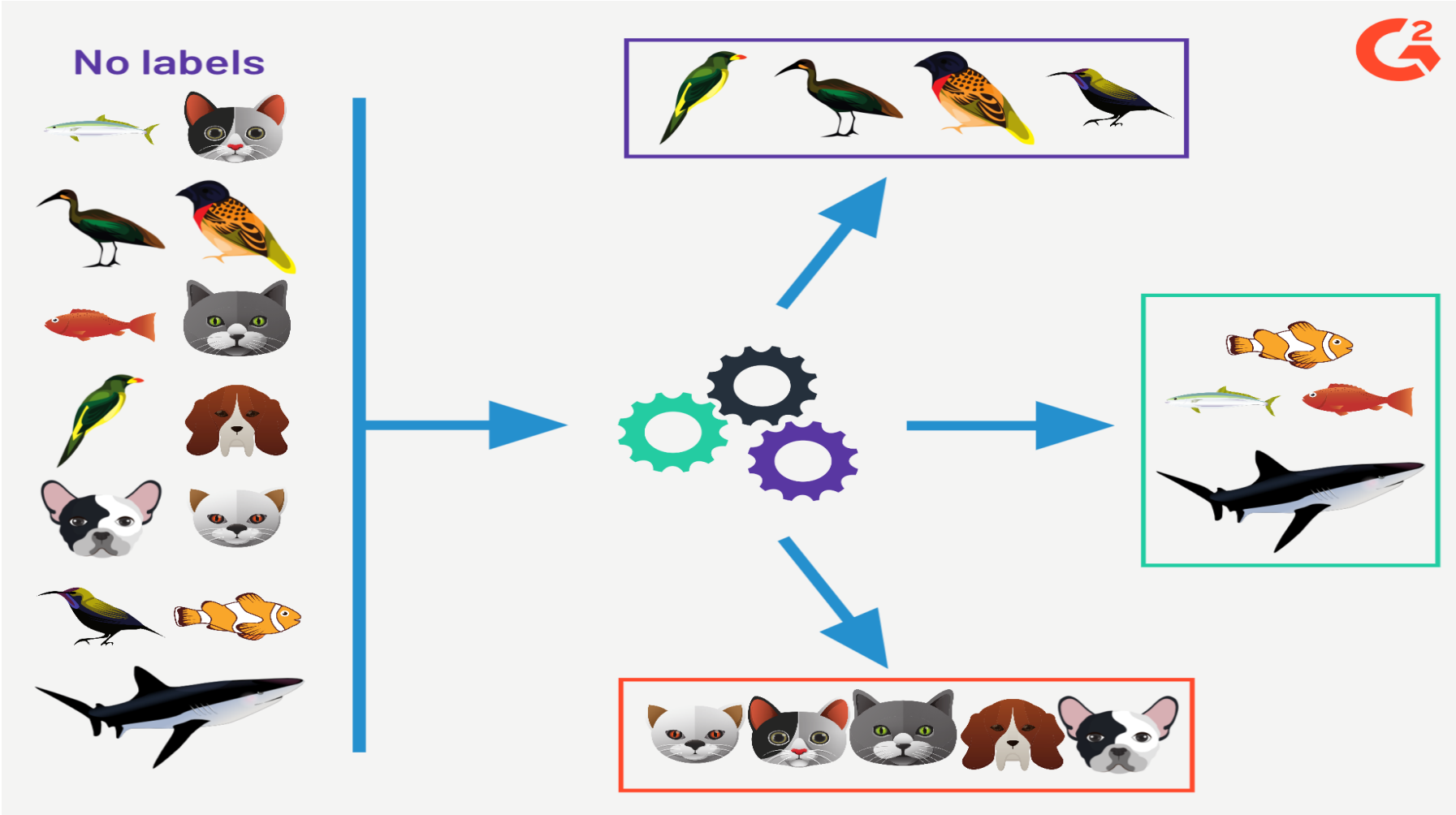
Unsupervised Learning

- Unsupervised learning, also known as unsupervised machine learning, uses machine learning (ML) algorithms to analyze and cluster unlabeled data sets.

- These algorithms discover hidden patterns or data groupings without the need for human intervention.


supervised learning
Input data
Prediction
Its an apple!
Annotations
These are apples
Model
unsupervised learning
Input data
Model

UnSupervised Learning



Prerequisites for Understanding Clustering

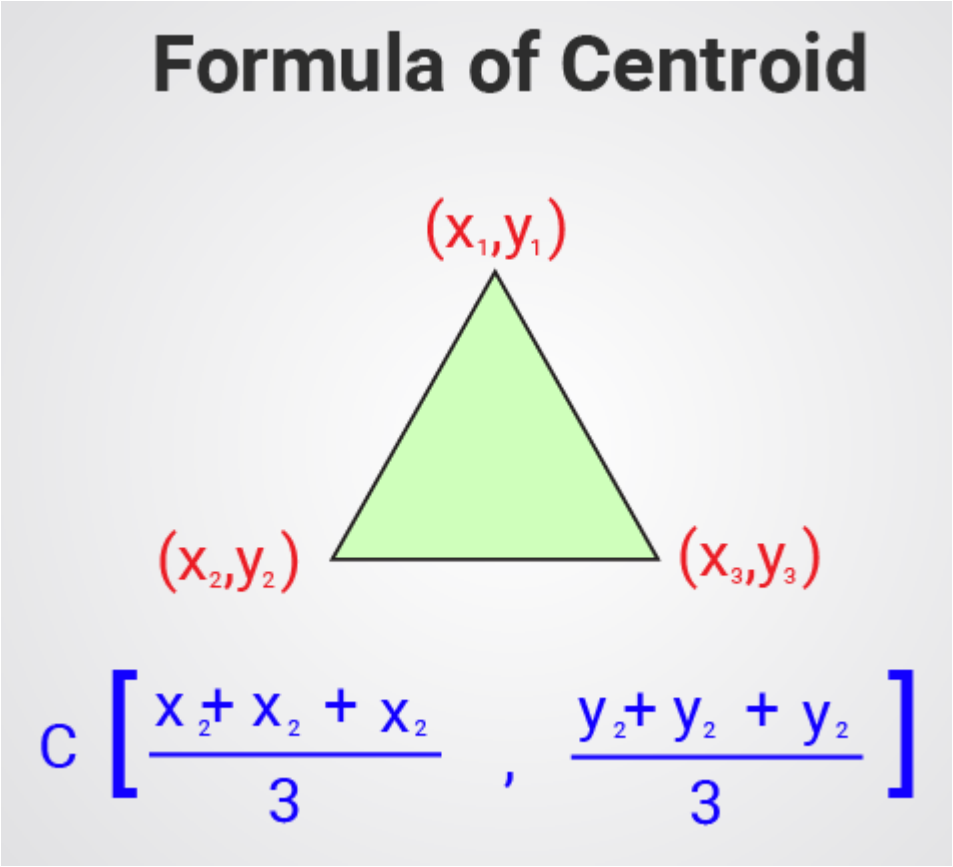
- Centroid
- Euclidian Distance

Centroid

- **Definition:**
  - A centroid is the center of a cluster in a dataset. It is a point that represents the average position of all the points in the cluster.

- In a 2-dimensional space, the centroid is the mean of the x-coordinates and the y-coordinates of all the points in the cluster.

- In higher dimensions, the centroid is calculated by taking the mean of each coordinate dimension.



Centroid

- **Calculation:**
  - For a set of points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ in 2D:

$$\text{Centroid} = \left(\frac{x_1 + x_2 + ... + x_n}{n}, \frac{y_1 + y_2 + ... + y_n}{n} \right)$$

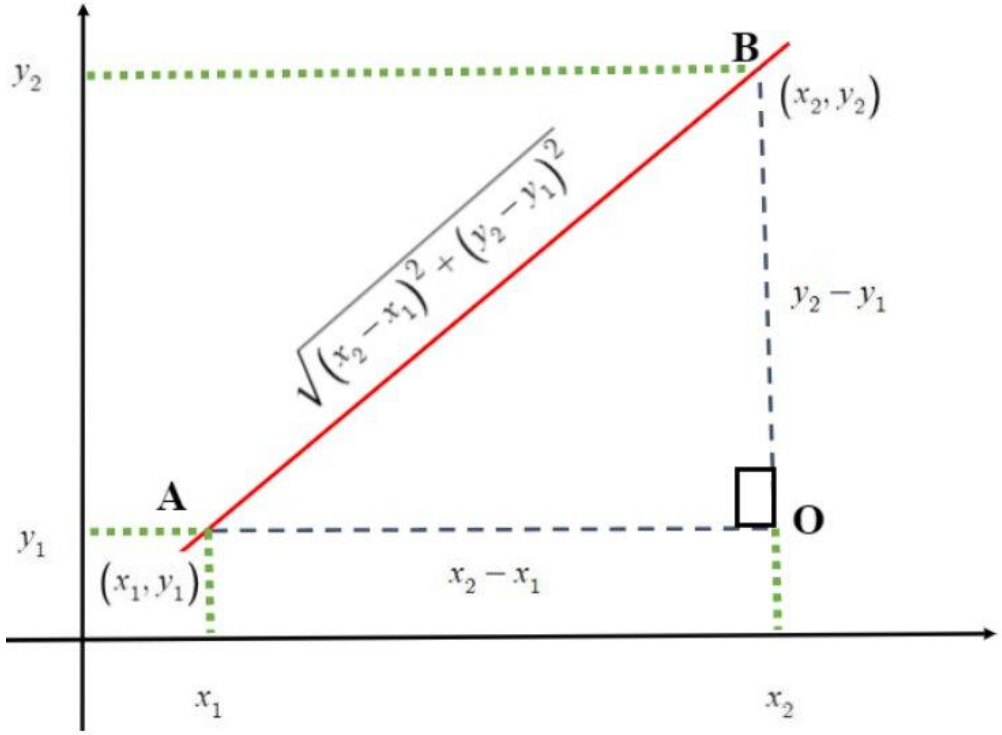
  - For points $(x_1, x_2, ..., x_m), (y_1, y_2, ..., y_m), ..., (z_1, z_2, ..., z_m)$ in m-dimensional space:

$$\text{Centroid} = \left(\frac{x_1 + x_2 + ... + x_m}{n}, \frac{y_1 + y_2 + ... + y_m}{n}, ..., \frac{z_1 + z_2 + ... + z_m}{n} \right)$$

Centroid Problems

Euclidean Distance

- Euclidean distance is a measure of the straight-line distance between two points in Euclidean space. It is the most common distance metric used in clustering.



Euclidean Distance

- **Formula:**
  - For two points (x_1, y_1) and (x_2, y_2) in 2D:

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

  - For points $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ in n-dimensional space:

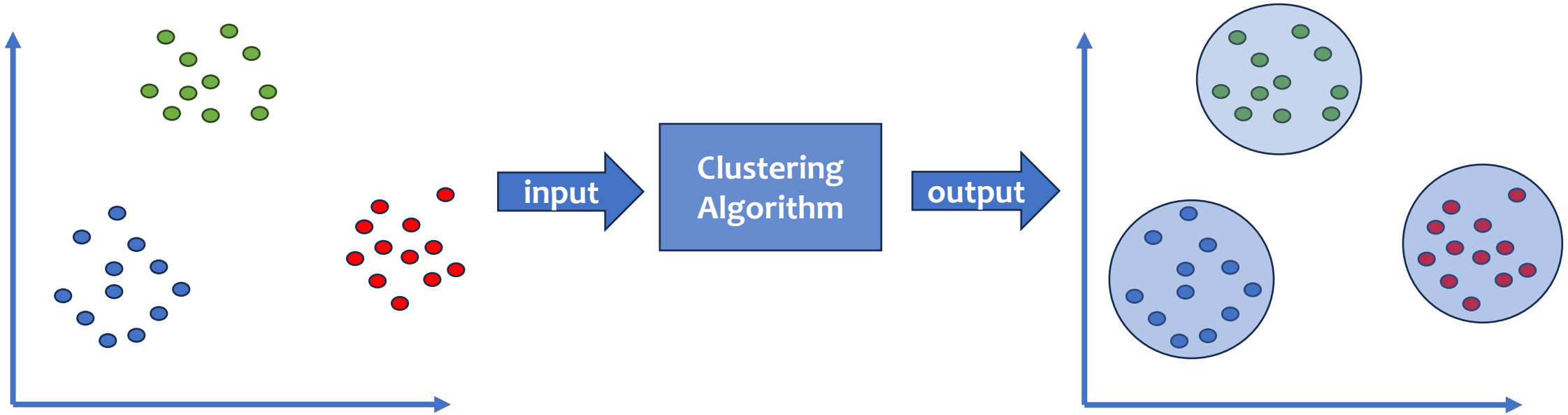
$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

Euclidean Examples:

Partitioning Clustering

K means Clustering

CLUSTERING



Clustering

- Clustering is a fundamental unsupervised learning technique used to identify groupings within a dataset

- It aims to partition data point into clusters such that point within the same cluster are more similar to each other than to those in other clusters

K means Algorithm

1. Decide how many clusters you want, i.e. choose k
2. Randomly assign a centroid to each of the k clusters
3. Calculate the distance of all observation to each of the k centroids
4. Assign observations to the closest centroid
5. Find the new location of the centroid by taking the mean of all the observations in each cluster
6. Repeat steps 3-5 until the centroids do not change position

K-means Clustering Example:

Elbow Method

- Elbow Method is a technique that we use to determine the number of centroids(k) to use in a k-means clustering algorithm.

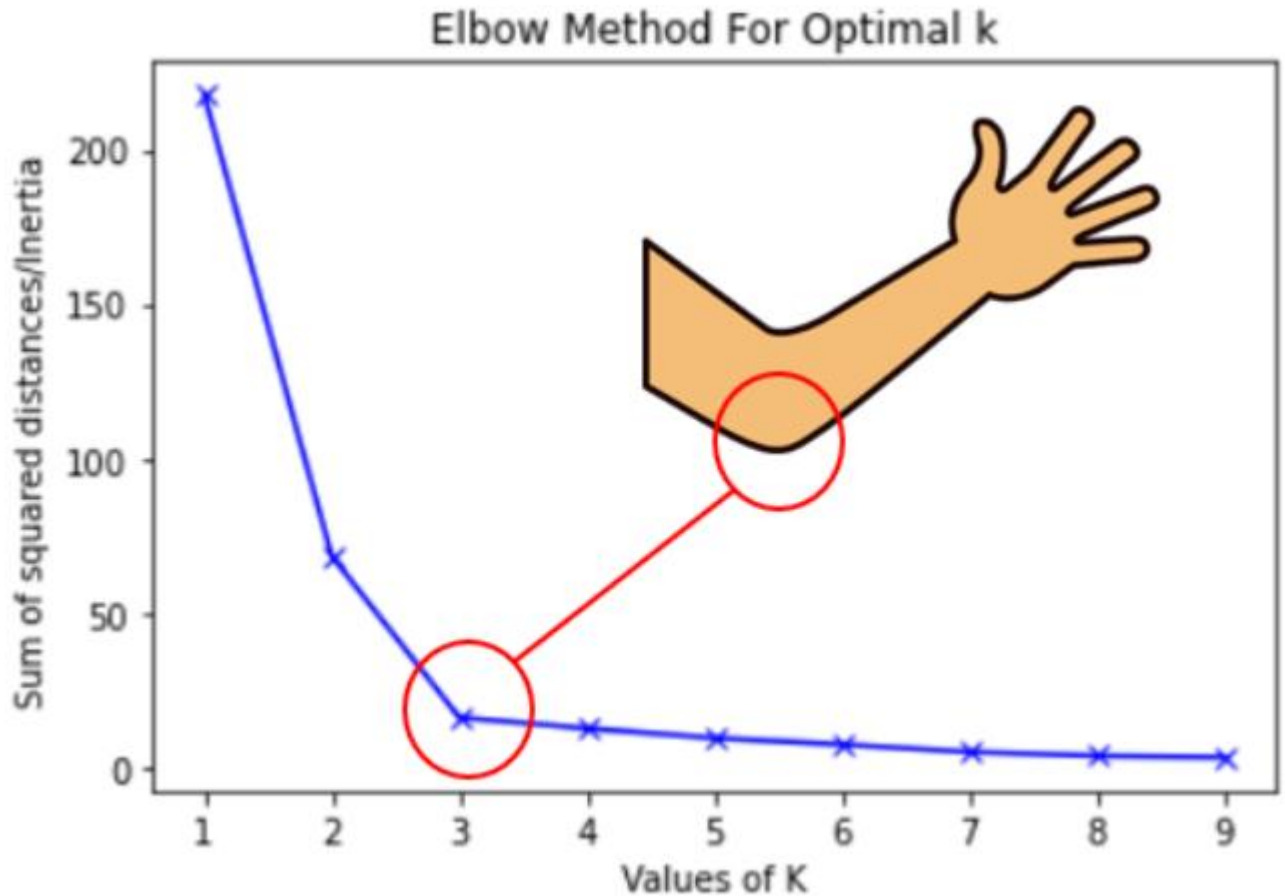
- continuously iterate for k=1 to k=n (Here n is the hyperparameter that we choose as per our requirement). For every value of k, we calculate the wit

- WCSS - It is defined as the sum of square distances between the centroids and each points.hin-cluster sum of squares (WCSS) value.

Elbow Method

- Now For determining the best number of clusters(k) we plot a graph of k versus their WCSS value. Surprisingly the graph looks like an elbow

- Also, When k=1 the WCSS has the highest value but with increasing k value WCSS value starts to decrease. We choose that value of k from where the graph starts to look like a straight line.



Line plot between K and inertia

K mean clustering visualization

- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- https://shabal.in/visuals/kmeans/1.html

Clustering Loss Function

1. Kmeans clustering partitions data into K disjoint sets or clusters where K is a pre-specified number. It can range from 1 to n where n is number of data points

2. Let the clusters be c1, c2, … , cn

3. c1 U c2 U … U cn = { 1,2,….,n} data set

4. ci n cj = 0 for all i not equal to j

Clustering Loss Function

- A good clustering is the one where the within cluster variation is minimal

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

- In words, minimize the total sum of all within cluster variations

The within cluster variation is defined as

$$\mathrm{WCSS}_k = \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mu_k\|^2$$

where:

- C_k is the set of data points in the k-th cluster.
- μ_k is the centroid (mean) of the k-th cluster.
- $\|\mathbf{x}_i - \mu_k\|^2$ is the squared Euclidean distance between a data point $\mathbf{x}_i$ and the cluster centroid μ_k.

- Thus the optimization problem becomes

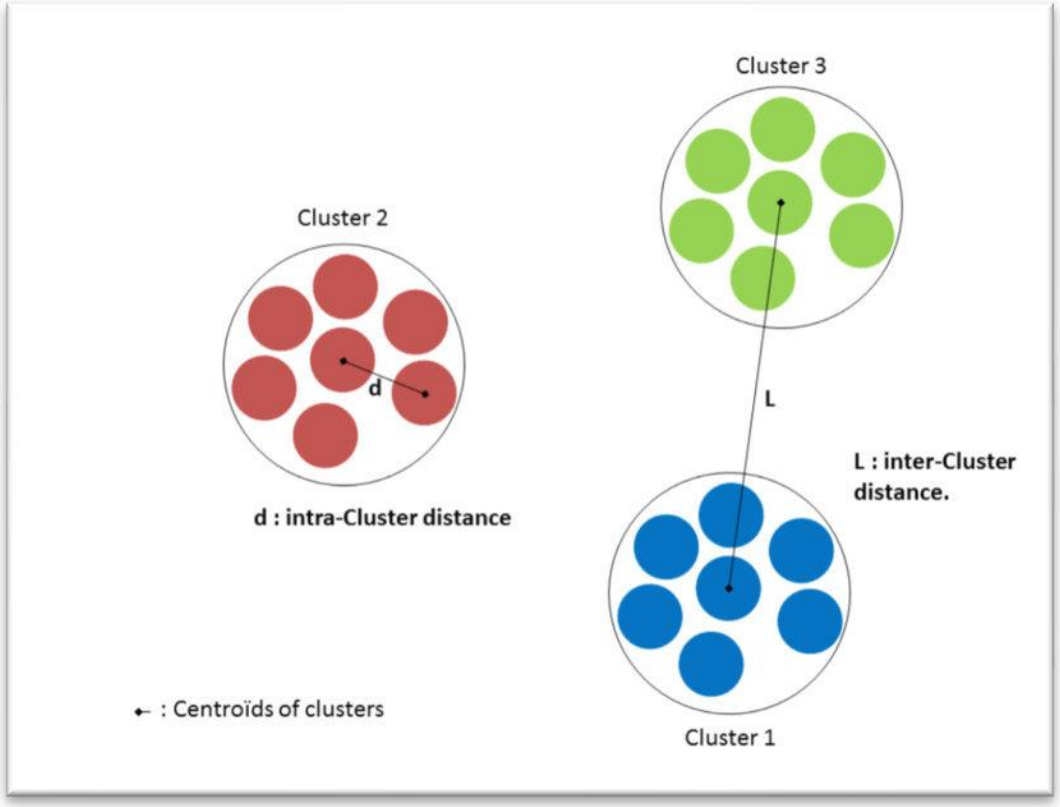
The total within-cluster variance is the sum of the within-cluster variances for all clusters:

$$\mathrm{WCSS} = \sum_{k=1}^{K} \mathrm{WCSS}_k$$

Good Clustering

- Clustering is based on two important metrics:
  - Inter-Cluster distance
  - Intra-cluster distance.

- In ideal conditions for clustering to perform well the inter-cluster distance should be very high and intra-cluster distance should be very low.



K-means Advantages

- Simple and easy to implement.
- Scalable and efficient for large datasets.
- Easy to interpret and visualize.
- Flexible and applicable to various domains.

K-means disadvantages

- Requires the number of clusters (K) to be specified.
- Sensitive to initial cluster centers.
- Assumes spherical clusters of similar sizes.
- Sensitive to outliers and noise.
- Relies on Euclidean distance, which may not be suitable for all data types.
- As the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. Reduce dimensionality either by using **PCA** on the feature data, or by using “spectral clustering” to modify the clustering algorithm as explained below.

K-means Application

1. Customer Segmentation

Marketing: Segmenting customers based on purchasing behavior, demographics, or browsing patterns to tailor marketing strategies, promotions, and product recommendations.

Retail: Grouping customers based on purchase history to identify target audiences for specific products or services.

2. Image Compression

Image Processing: Reducing the number of colors in an image by clustering similar colors and representing them with the cluster centroid. This reduces the image file size while maintaining visual quality.

3. Document Clustering

Text Mining: Grouping similar documents or text data, such as news articles, research papers, or customer reviews, based on their content. This helps in organizing large collections of documents and improving information retrieval.

K-means Application

4. Anomaly Detection

Network Security: Identifying unusual patterns or anomalies in network traffic that could indicate potential security threats or intrusions.

Fraud Detection: Detecting fraudulent transactions or activities by identifying patterns that deviate significantly from typical behavior.

5. Image Segmentation

Medical Imaging: Segmenting different regions in medical images (e.g., MRI or CT scans) for diagnostic purposes.

6. Market Basket Analysis

Retail Analytics: Grouping products that are frequently purchased together to optimize store layouts, inventory management, and cross-selling strategies.

K-means Application

7. Recommender Systems

Content Recommendation: Grouping similar users or items to provide personalized recommendations in e-commerce, streaming services, or online platforms.

8. Financial Analysis

Portfolio Management: Grouping assets with similar performance characteristics to optimize portfolio diversification.

Risk Assessment: Identifying clusters of high-risk or low-risk investments based on historical data.