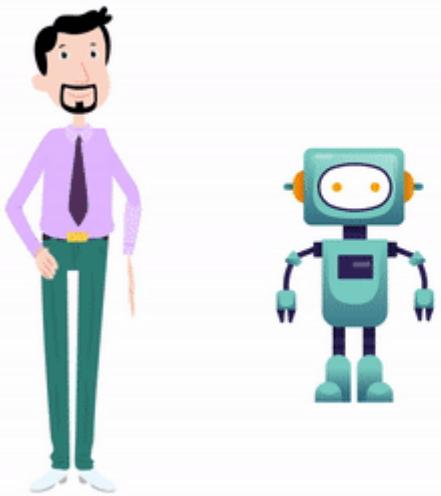
Unsupervised Learning

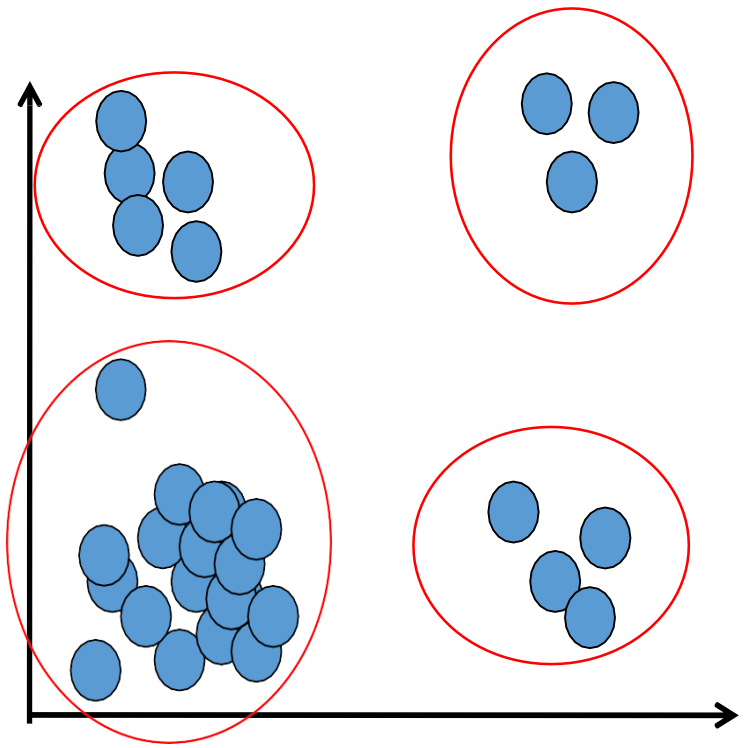


Module No. 6	Unsupervised Learning

Unsupervised Learning - Agglomerative clustering, Hierarchical clustering, k-means clustering, limitations. Introduction to Gaussian Mixture Models, Maximum Likelihood Estimation, parameter estimation for a mixture of gaussians, Expectation Maximization.

Clustering

- Clustering is a technique for identifying similarity groups in data, called clusters.



The goal of clustering is to

- Group data points that are close (or **similar**) to each other
- Identify such groupings (or clusters) in an **unsupervised** manner

- How to define similarity?
- How many iterations are needed to check cluster quality?

Clustering

Supervised learning: discover patterns in the data with known target (class) or label.

These patterns are then utilized to predict the values of the target attribute in future data instances.

Examples ?

Unsupervised learning: The data have no target attribute.

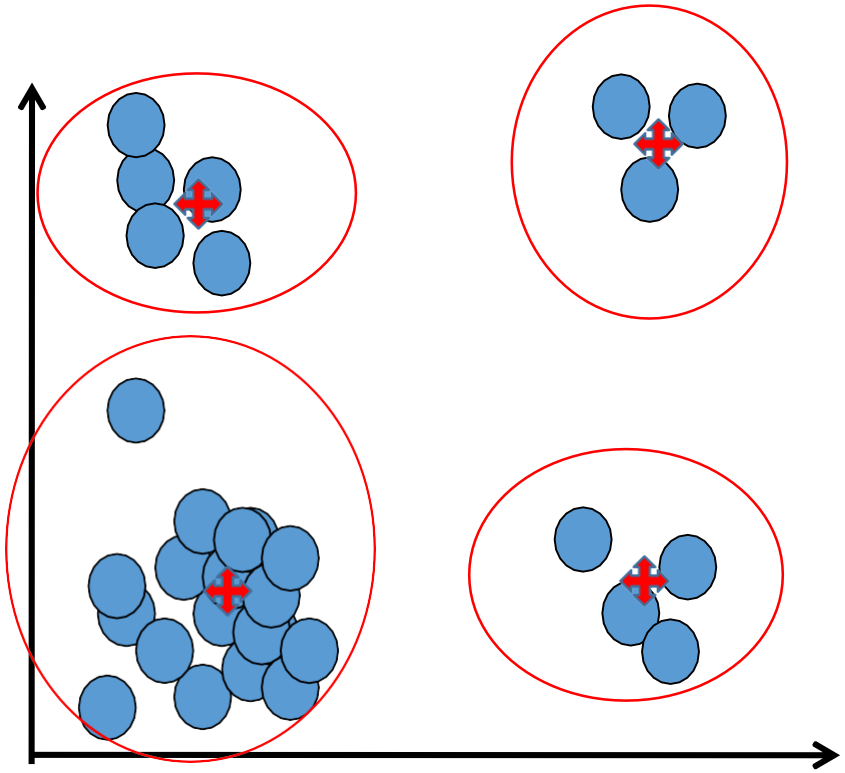
We want to explore the data to find some intrinsic structures in them.

Can we perform regression here ?

Examples ?

Cluster

- A cluster is represented by a single point, known as centroid (or cluster center) of the cluster.



- Centroid is computed as the mean of all data points in a cluster

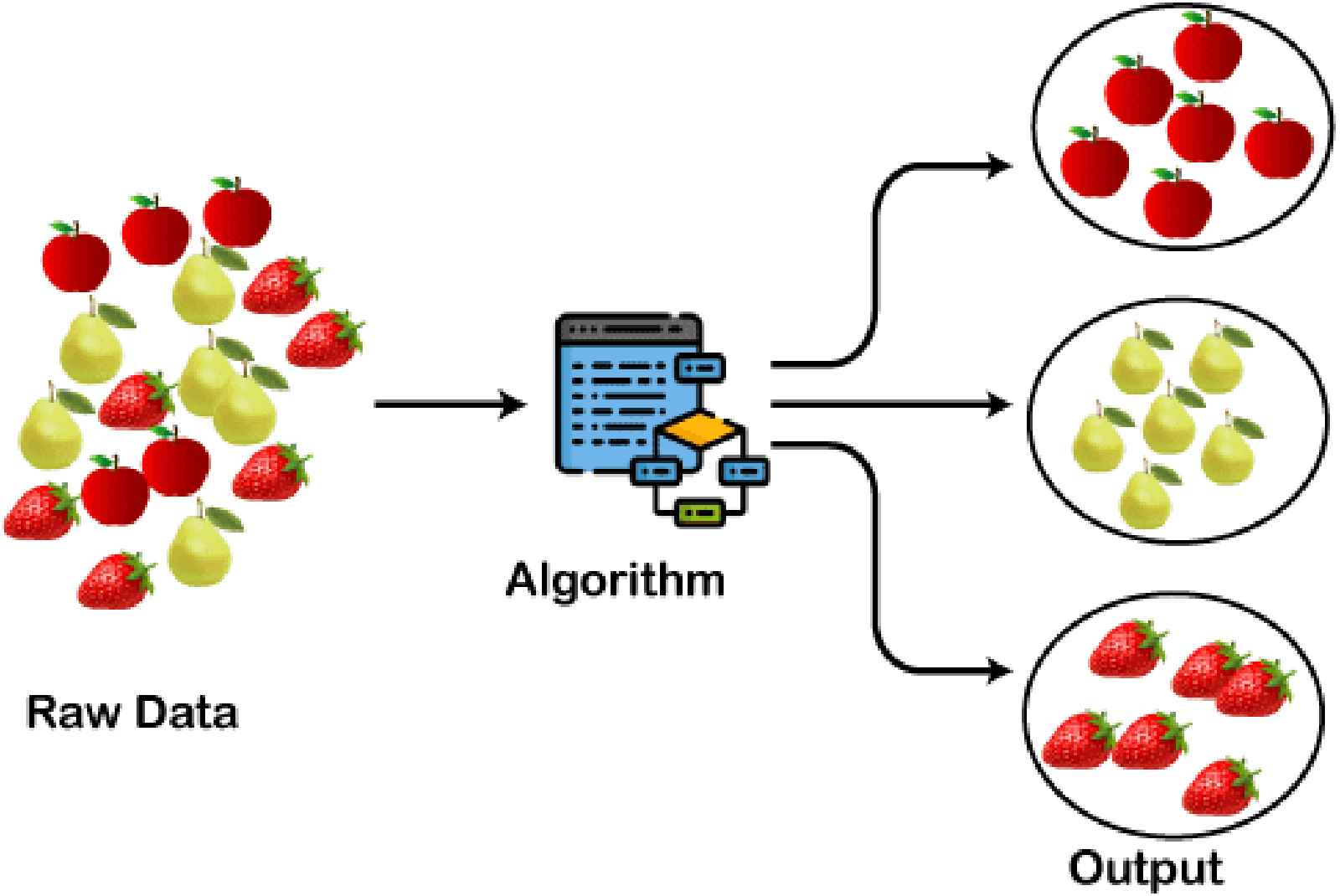
$$Cj = \sum xi$$

- Cluster boundary is decided by the farthest data point in the cluster.

Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- Earthquake studies: Observed earthquake epicenters should be clustered along continent faults.
- Image processing: Clustering parts of the image having similar RGB values, so that image is clustered into regions such as sky, greenery, road, house, etc.

Clustering



Major Clustering Approaches