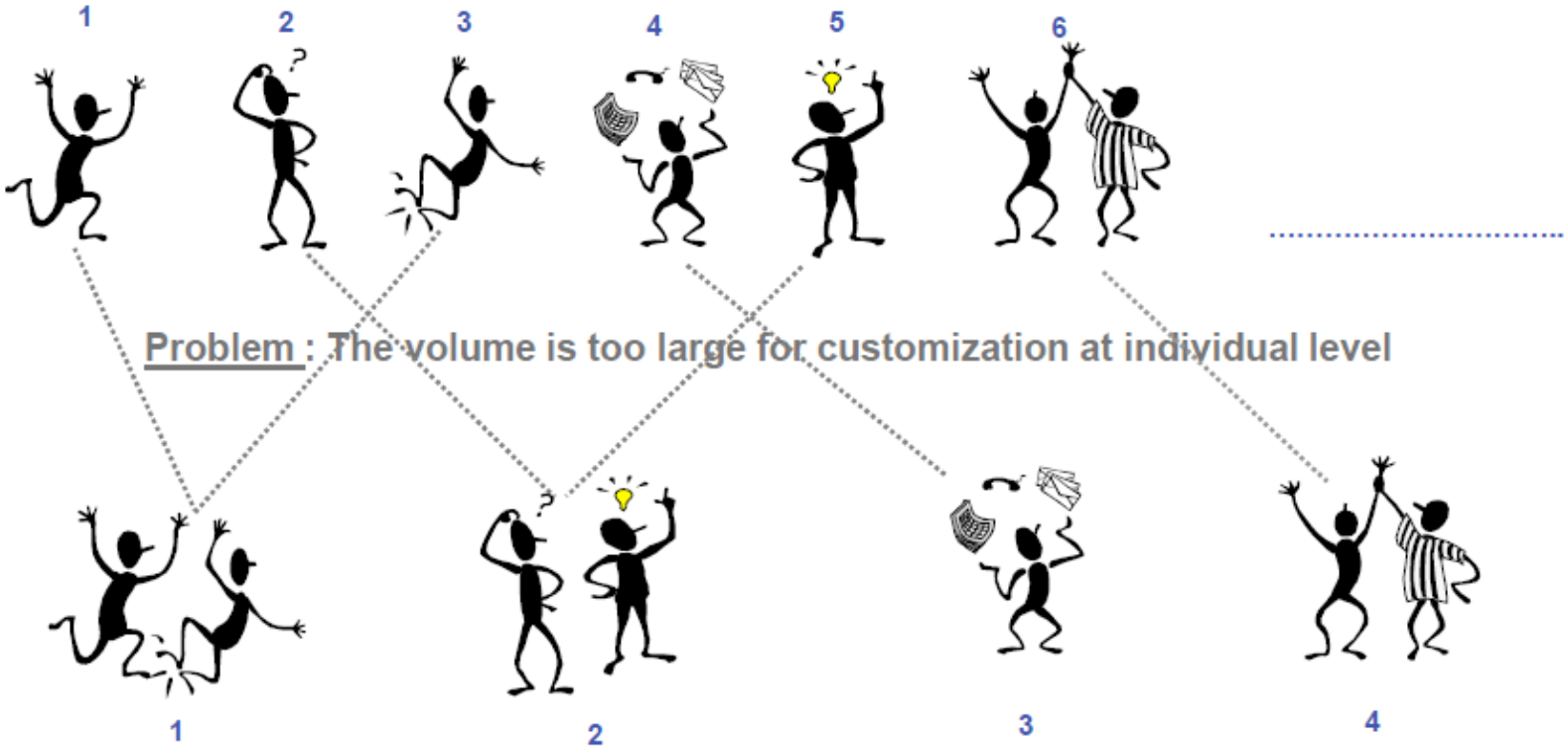
K-means Clustering

Why Segmentation?

Each individual is so different
that ideally we would want to reach out to each one of them in a different way



Solution : Identify segments where people have same characters and target each of these segments in a different way

Approach to Segmentation

Approach to Segmentation

Segmentation is of 2 types

Objective Segmentation

Clear Objective to divide population

- Response rate
- Increase in Sales
- Conversion proportion

Objective defined Analysis. To identify the desired segment within population. Then devising strategy to tap the potential within.

↓

CHAID Analysis

Subjective Segmentation

First level analysis to see what lies within

- Who are my customers?
- Who buys what?
- When do they buy?

Initial Analysis to Understand & Define the Population. Based on the initial understanding – Objective Based Analysis.

↓

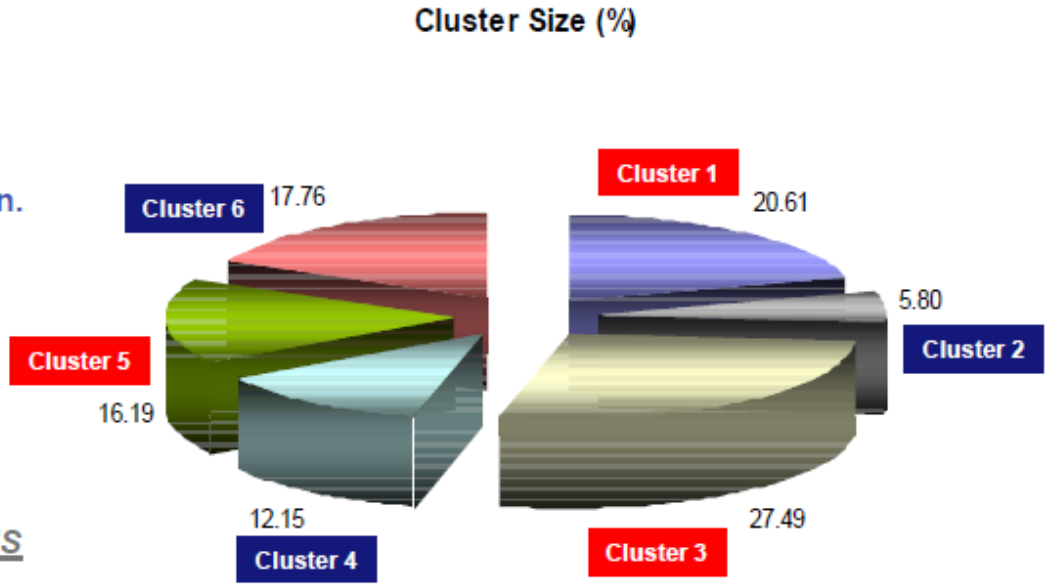
Cluster Analysis

What are clusters?

Clusters are groups within a Population.

These Groups are ***HOMOGENEOUS*** within themselves.

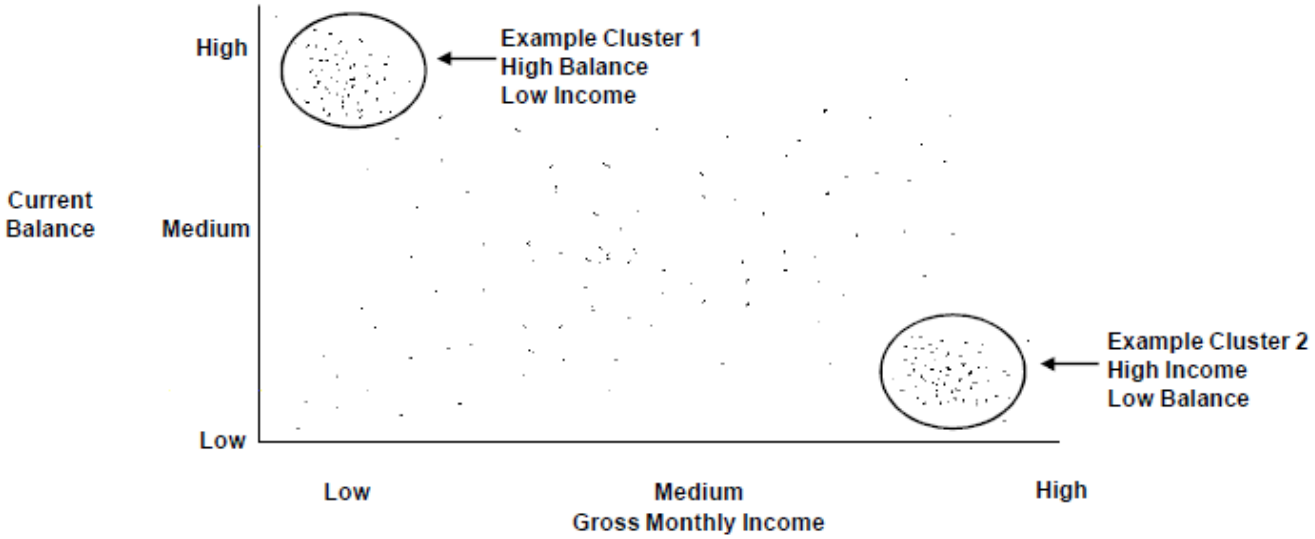
And these groups are ***HETEROGENOUS*** among each other.



Homogeneous segments making it possible to group people of similar characteristics.

Heterogeneous among themselves making it possible to differentiate segments within population.

Example of Clusters

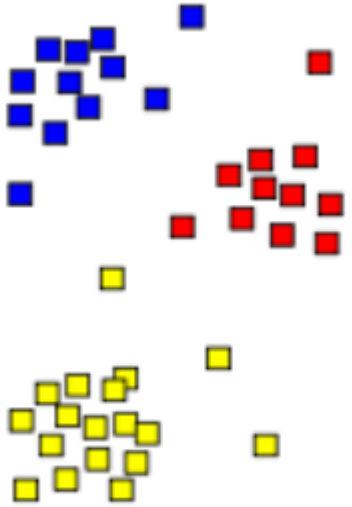


Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance. The objects in Cluster 1 have similar characteristics (High Income and Low balance), on the other hand the objects in Cluster 2 have the same characteristic (High Balance and Low Income).

But there are much differences between an object in Cluster 1 and an object in Cluster 2.

What is Clustering?

- Clustering is a 'UnSupervised Method' of statistical learning as there is no target variable to predict
- Here, the objective is to segment the data into similar groups instead of prediction
- We can cluster data into 'similar groups' and then build a predictive model for each group



Methodology of Variable Standardization

Professional School
ISO 9001 : 2000 Organisation

Why do we need 'Standardization' ?

Since the units of measurement are different for different variables, standardization is a must.

E.g.: - Consider two variables, Age and Income.
The unit of Age is 'Year' and the unit of Income is say 'Rs'.
Hence they are not comparable.
In that case there won't be an unit of measurement for the distance between two clusters.

Generally we standardize by making the mean = 0 and variance = 1 thus deunitizing the variables and bringing them on a common platform to analyze.

Post all the data treatment steps – "Cluster Development Process" is commenced upon.

Post Cluster Development – "Cluster Validation" is done on the validation sample to establish that the cluster solution is not Sample dependent.

Types of Clustering Methods

There are 2 ways in which Cluster solutions could be built up.

Hierarchical Clustering

Each observation is considered as an individual cluster. Distance from each observation to all others is calculated & the nearest observations are clubbed to form clusters. Intensive distance calculations required thus making it difficult to implement.

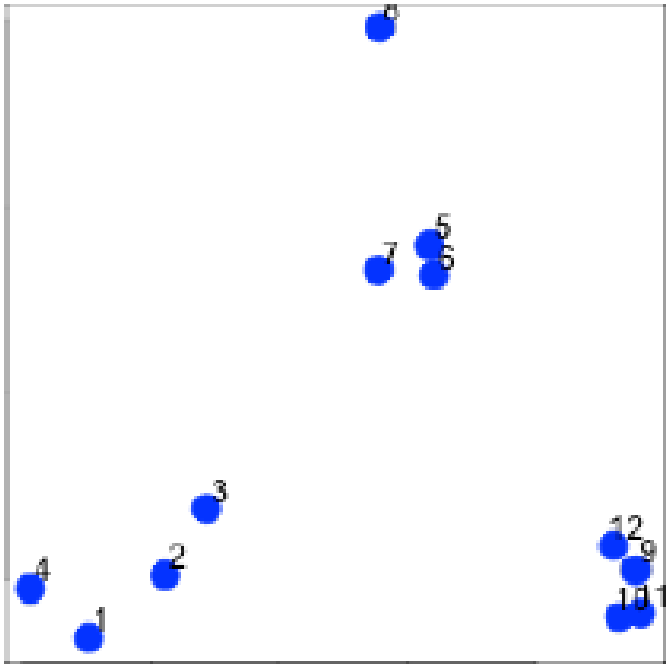
K-Means Clustering

K distinct observations are randomly selected at the highest distance from each other. Each observation is considered one by one & clubbed to the nearest Cluster. If two clusters come significantly close to each other, they are merged to each other to form a new cluster.

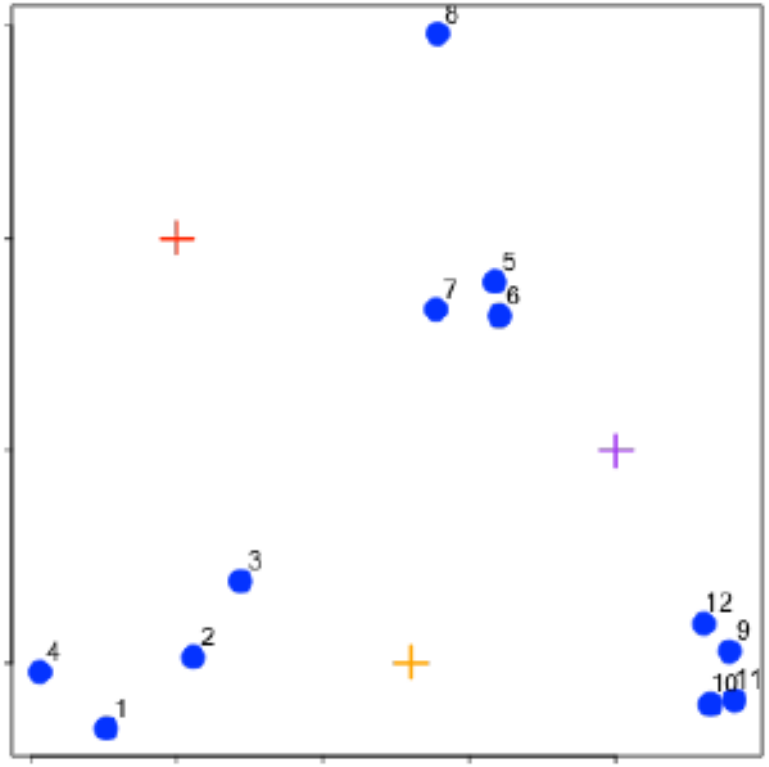
Hierarchical Clustering is not suitable for large datasets as the multitude of calculations involved would be impossibly huge. Thus K-Means clustering is the most used method of clustering.

K-means clustering

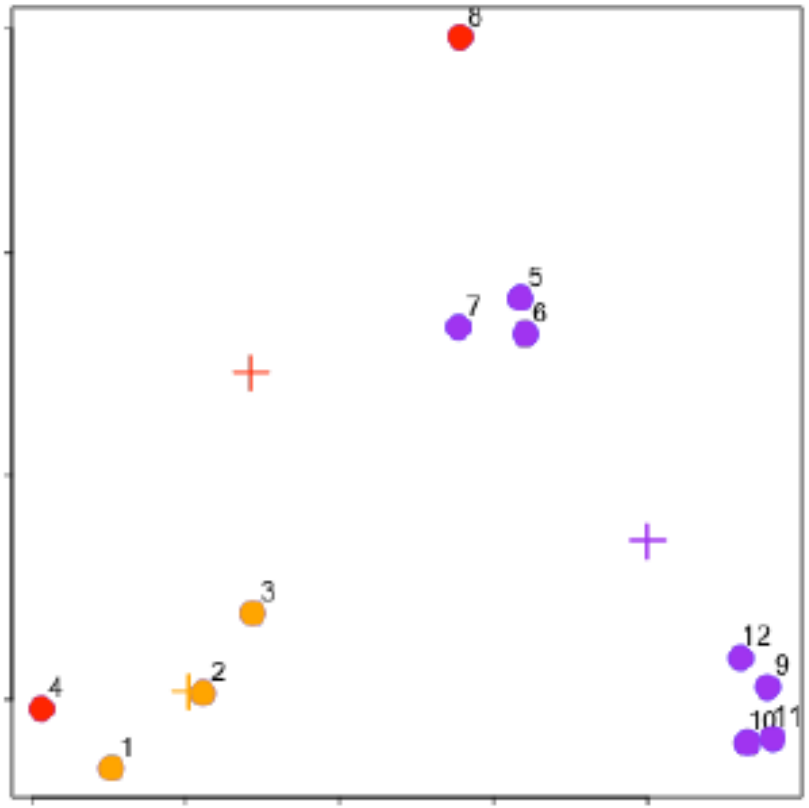
- A partitioning approach
- Fix a number of clusters
- Get "centroids" of each cluster
- Assign things to closest centroid
- Recalculate centroids



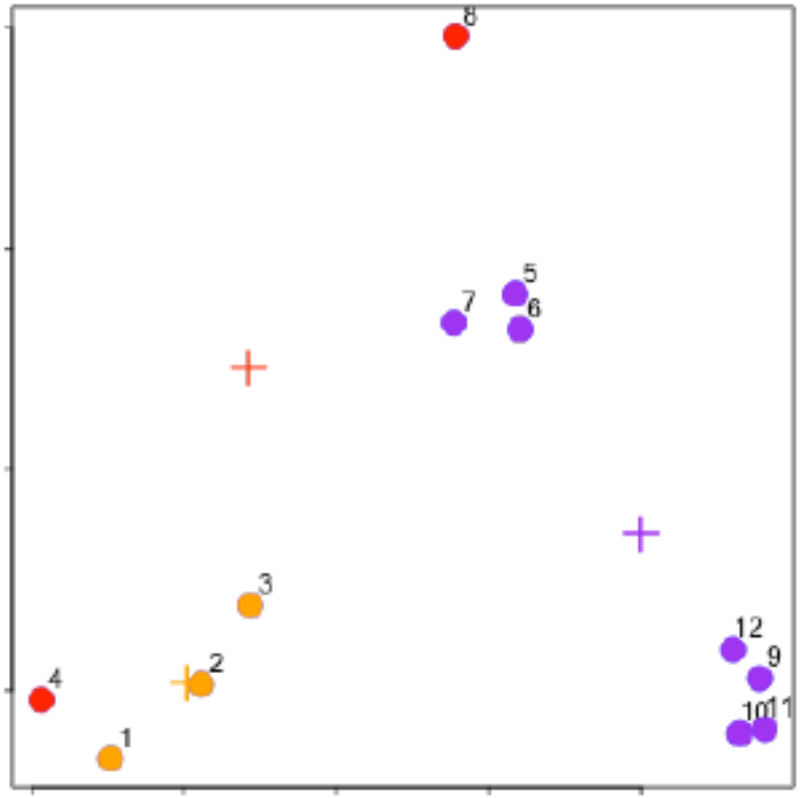
K-Means Clustering-Starting Centroids



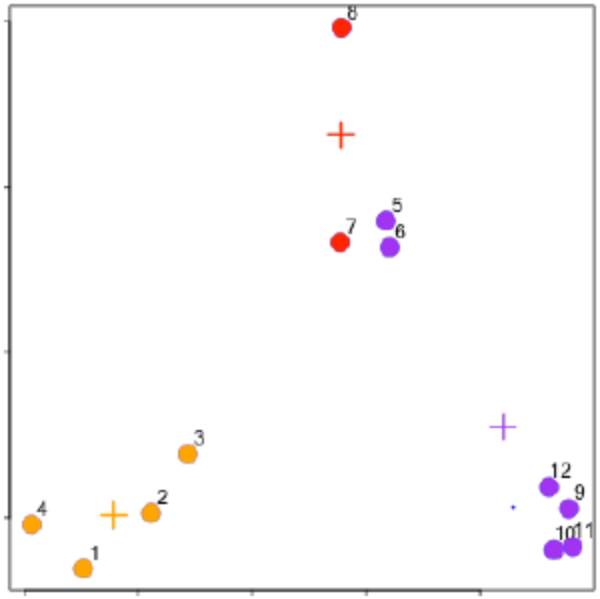
K-means clustering-Assign to closest centroids



K-means Clustering: Recalculate Centroids



K-means Clustering: Update Centroids



Thank You