Introduction to clustering

clustering is the process of grouping together data objects into multiple set or clusters, so that objects with in a cluster have high similarity as compared to objects outside of it.

* Similarity is calculated or measured by distance metrics.

* The partitioning of clusters is not done by humans. It is done with help of algorithms.

* clustering is also called data segmentation because it partitions large datasets into groups according to their similarity.

* clustering is known as unsupervised learning because the class label information is not present.

Application of clustering.

* Buisness Intelligence
* Pattern Recognition
* Image Processing
* Bioinformatics
* web technology
* Text mining.

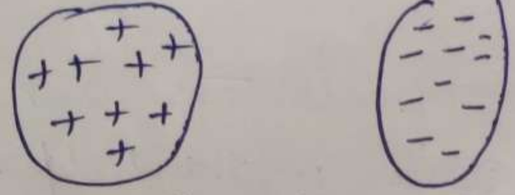
Types of clustering

clustering algorithms can be classified into two main subgroups :-

1 Hard clustering

It means each data point either belongs to cluster completely or not.

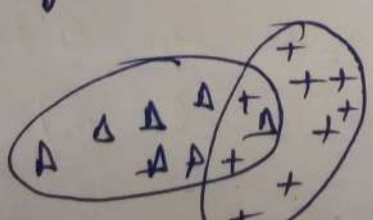
e.g k- means Clustering



'1' Data points cluster 1

"2" Data points. cluster 2

(2) Soft clustering : Here Data points / items belongs to multiple clusters.



fuzzy / c- means

depends upon prob. / membership functions.

Clustering algorithms can also be classified as follows:-

(1) Partitioning method.

(2) Hierarchical " .

(3) Density-based " .

(4) Grid-based " .

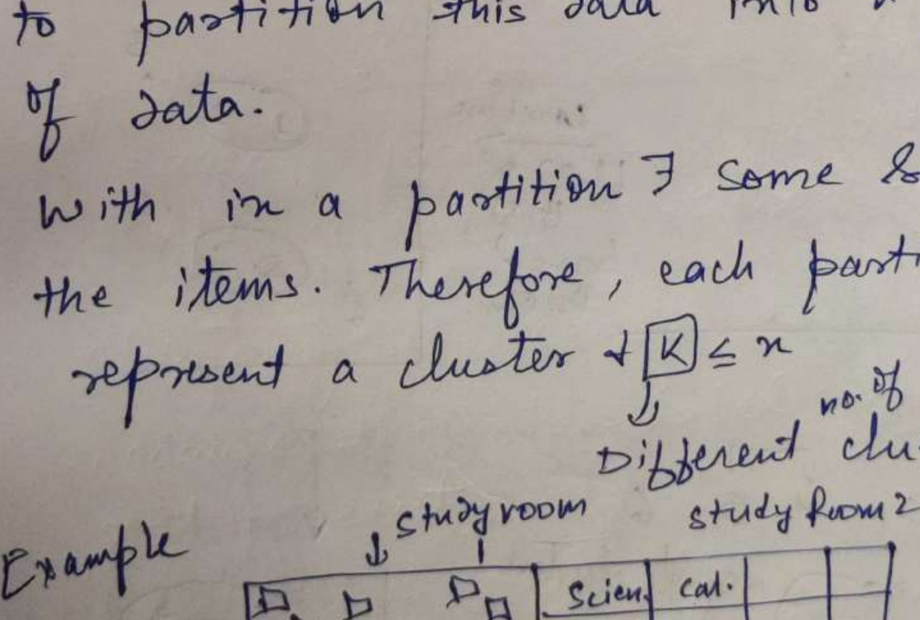
Partitioning Method

It means division, suppose we have a dataset with 'n' ~~[illegible]~~ objects + we need to partition this data into k partitions of data.

With in a partition ∃ some similarity among the items. Therefore, each partition will represent a cluster + $\boxed{K} \le n$

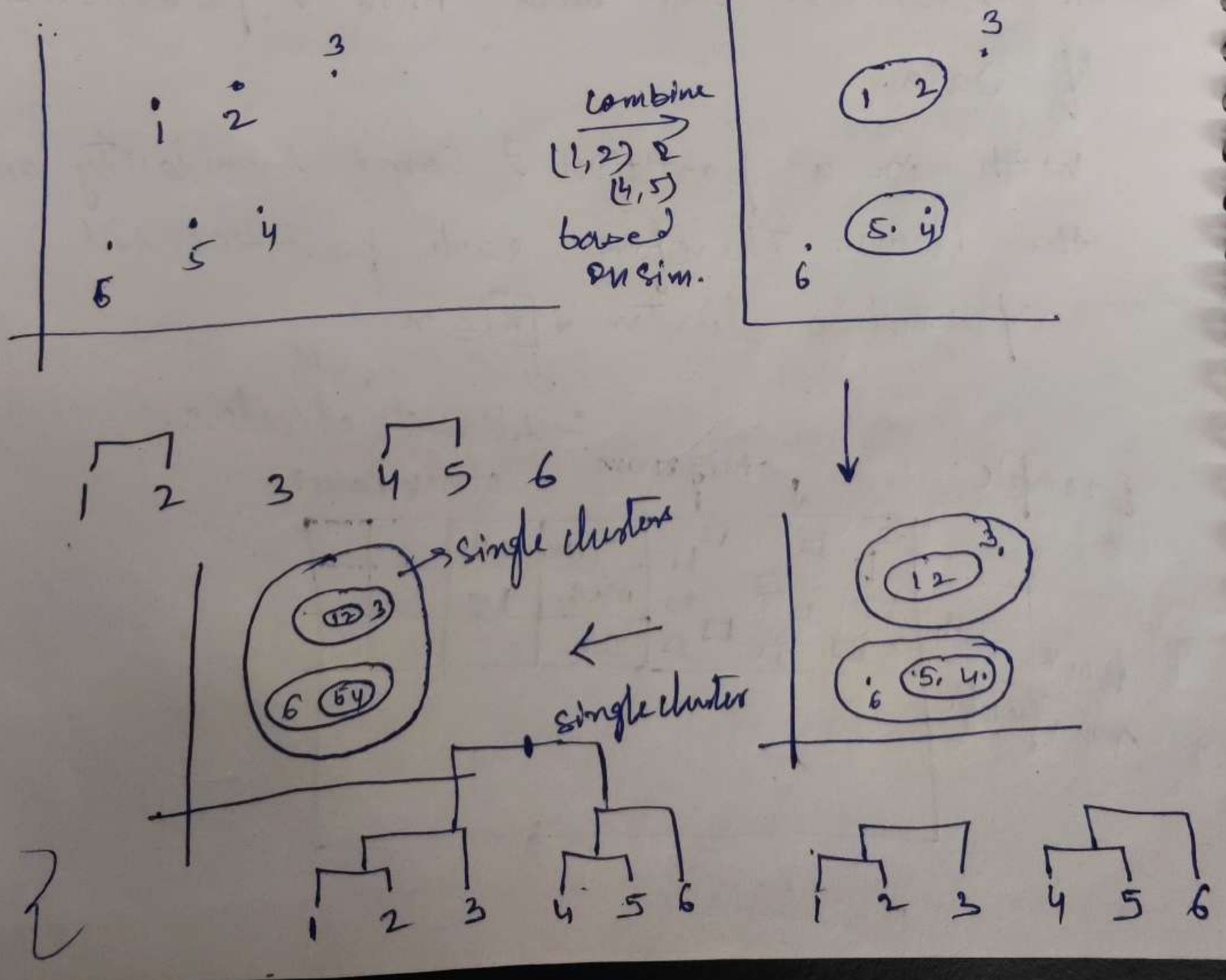
↓ Different no. of clusters.

Example



Hierarchical clustering

Suppose you have Six data points
A B C D E F
1, 2, 3, 4, 5, 6 → Types of data



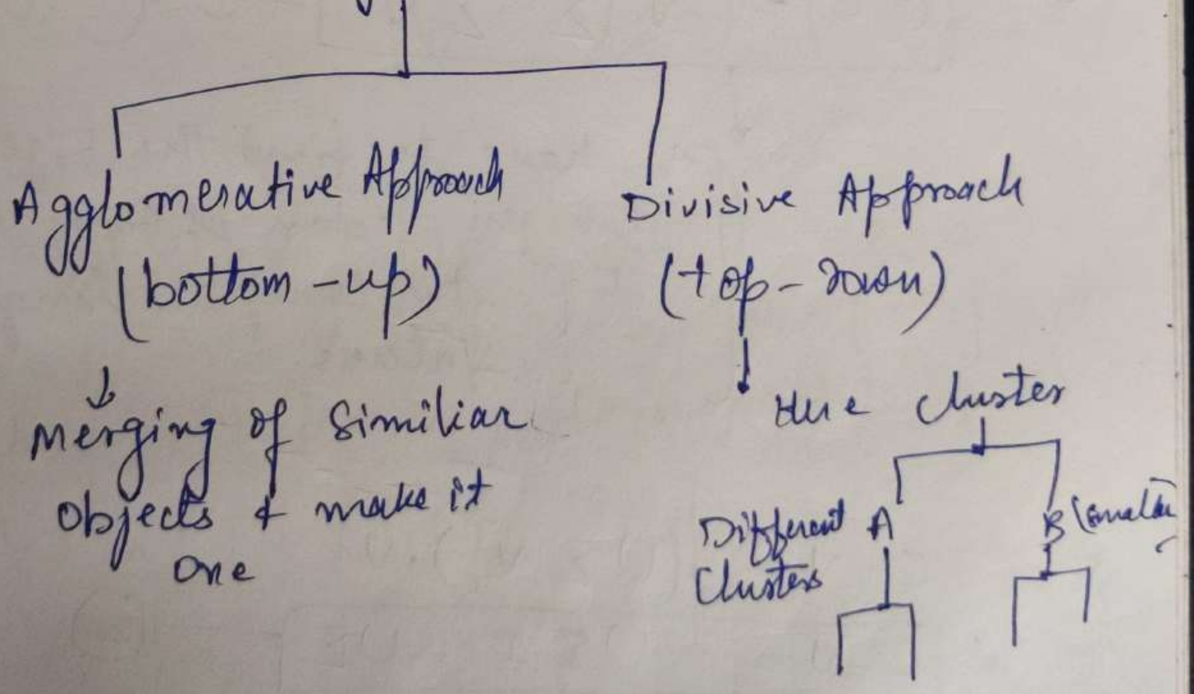
Diesel filter

Formal definition

Hierarchical clustering is an alternative approach to partitioning clustering for identifying groups in a dataset.

Main advantage :- It does not require pre-specify amount / no. of clusters to be generated. The result of Hierarchical clustering is a tree-based representation of objects which is known as dendogram.

Also, these observations can be sub-divided into groups by cutting the dendogram at a desired similarity level.


Agglomerative Approach
(bottom-up)
Merging of similiar objects & make it one
Divisive Approach
(top-down)
Hire cluster
Different Clusters
A
B (smaller)