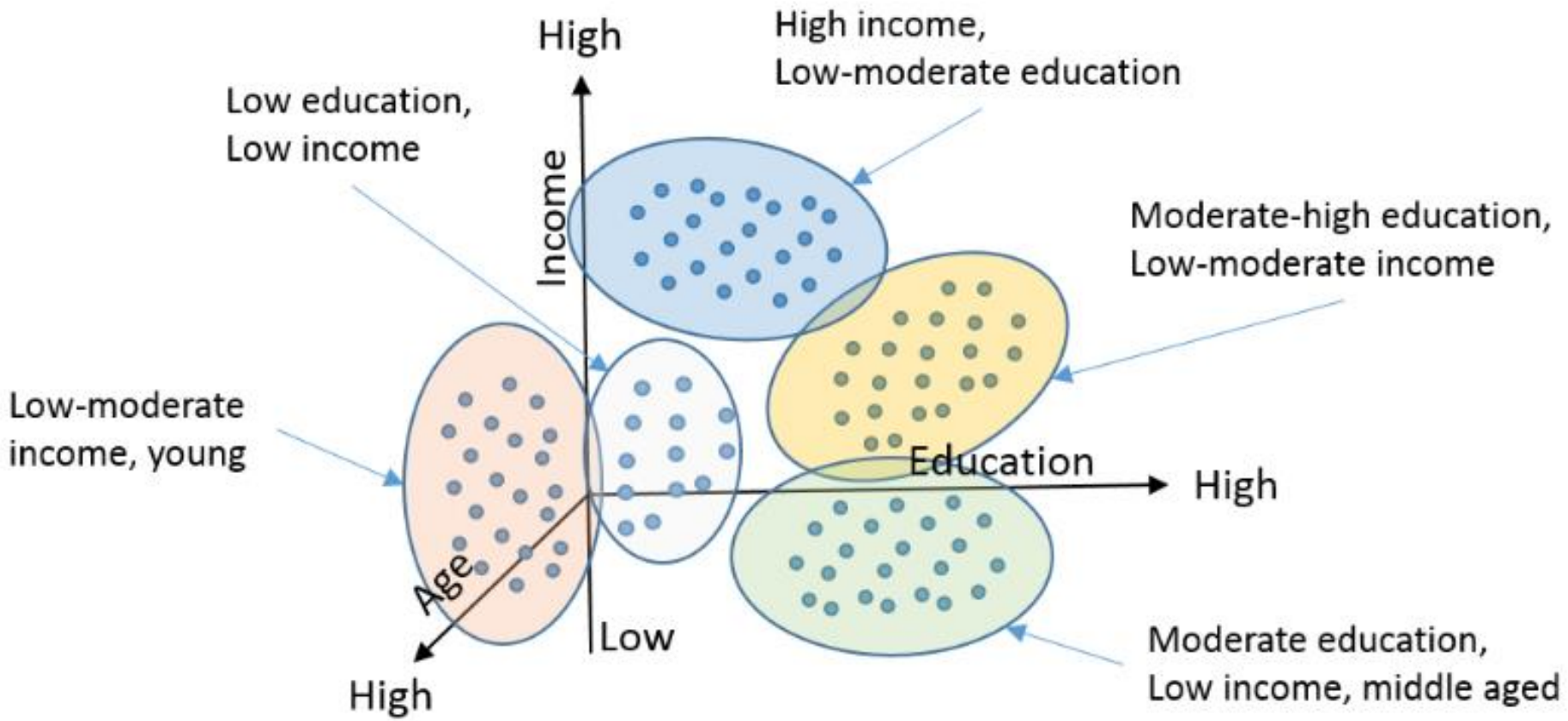
CLUSTERING

Introduction to Clustering

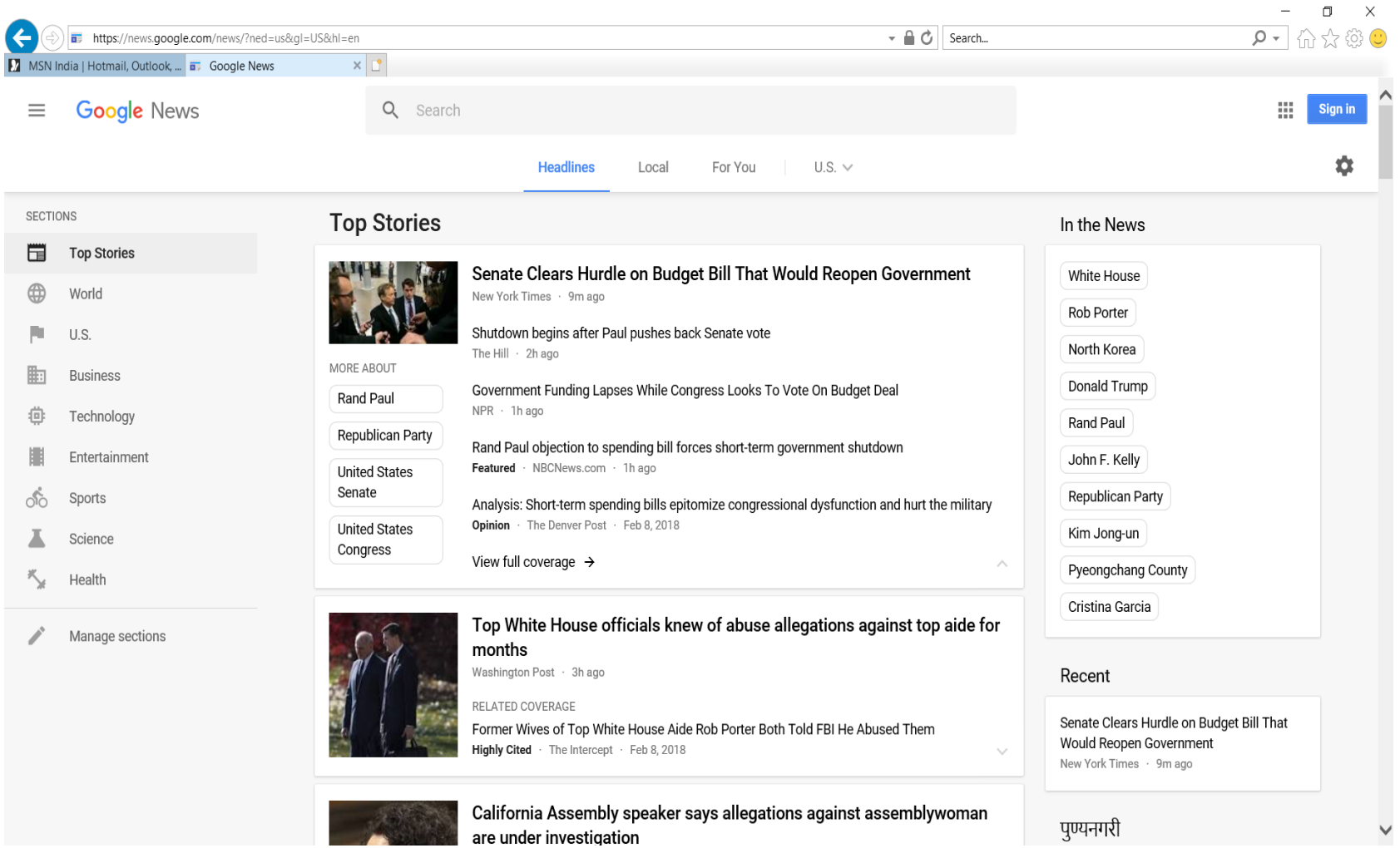
- Clustering is “the process of organizing objects into groups whose members are similar in some way”.
- A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Clustering


High
Income
Low
High
Education
High
Age
Low education,
Low income
High income,
Low-moderate education
Moderate-high education,
Low-moderate income
Low-moderate
income, young
Moderate education,
Low income, middle aged

Example

- When we search in google, we found a collection of result for a search.
- For example, if we check the google news we found a collection of news. This is an example of cluster analysis.

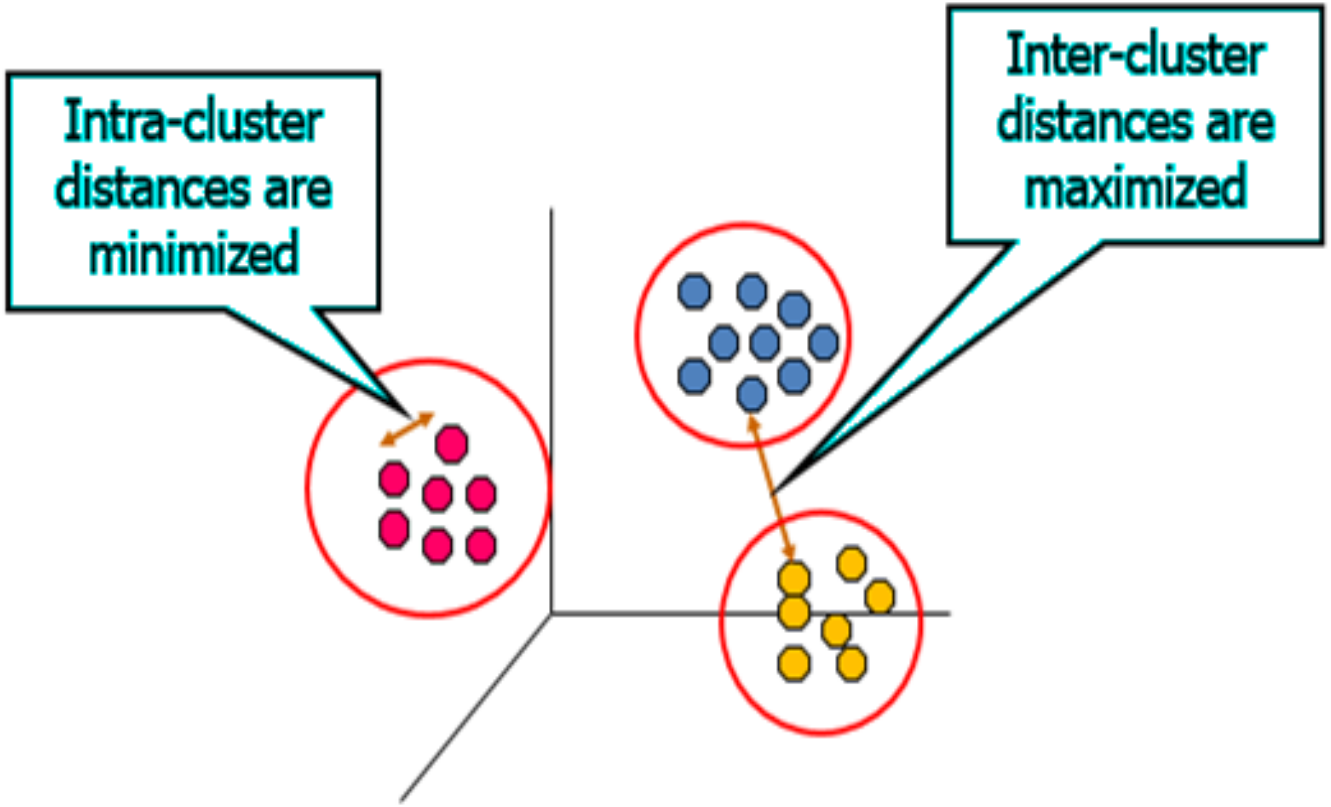


Applications of Clustering

- **Business:** Businesses collect large amounts of information about current and potential customers. Clustering can be used to segment customers into a small number of groups for additional analysis and marketing activities
- **Marketing:** Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.
- **Climate:** Understanding the Earth's climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in atmospheric pressure and ocean temperature that have a significant impact on climate.
- **Psychology and Medicine:** An illness or condition frequently has a number of variations, and cluster analysis can be used to identify these different subcategories. For example, clustering has been used to identify different types of depression. Cluster analysis can also be used to detect patterns in the spatial or temporal distribution of a disease.

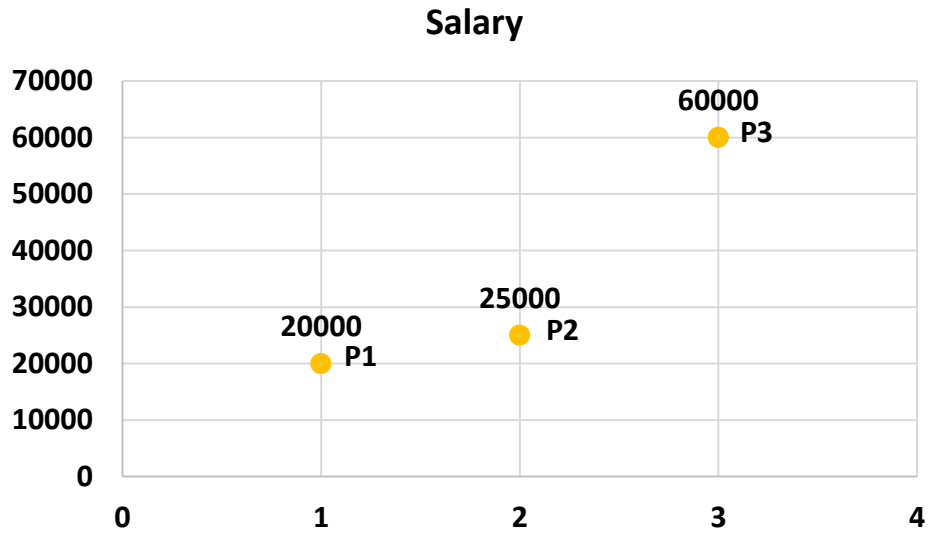
OBJECTIVE OF CLUSTER ANALYSIS

- Intra cluster distance is the sum of distances between objects in the same cluster.
- This distance should always be minimized.
- Inter cluster distance is the distance between objects in the different cluster.
- This distance should always be maximized.
- Lets have a look at an example will discuss in next slide



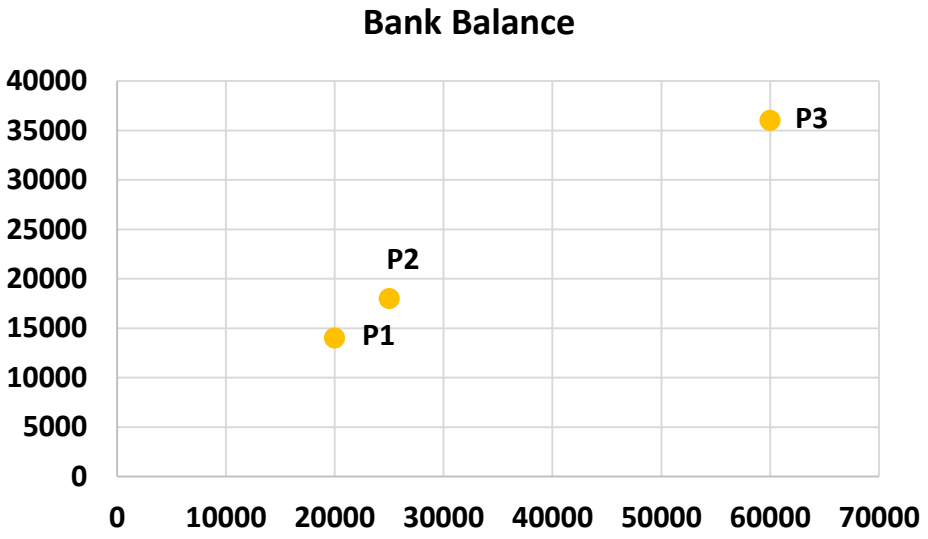
Introduction to clustering

- Lets take an example to show how the clusters are formed & distance are measured.
- Lets look at a graph, the salary of the three person P1, P2 and P3 is 20,000, 25,000 and 60,000 respectively and we have to divide them into 2 clusters.
- The difference between the salary of person P1 and P2 is 5,000 and
- The difference between the salary of person P2 and P3 is 35,000.
- The difference between the salary of person P1 and P3 is 40,000.
- The difference between the salary of person P1 and P2 is less as compared to P2 and P3 or P1 and P3
- Hence, P2 will combine with P1 to form a cluster.



Introduction to clustering

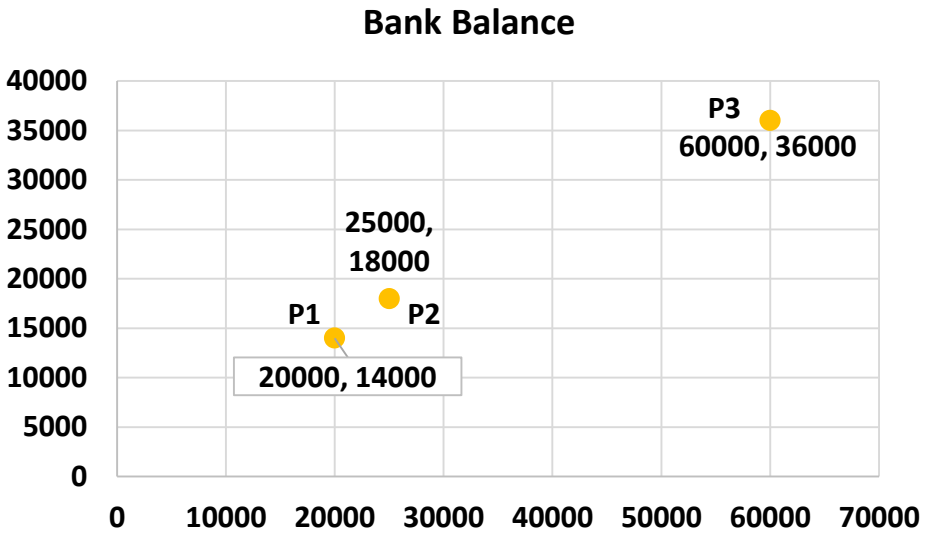
- Now lets take a 2 dimensional data, consider salary along with the bank balance.
- To find out the distance between the 2 dimensional data.
- We require some other distance measures which will help to form a cluster. The following are some of the measures
  - ✓ Euclidean distance
  - ✓ Manhattan distance
  - ✓ Mahalanobis distance
- Euclidean distance is the most popular distance measure to find out the clusters.



Salary	Bank Balance
20000	14000
25000	18000
60000	36000

Introduction to clustering

- Lets find out the distance between the points using Euclidean distance
- The coordinate of point P1(x1,y1) i.e. (20000,14000) and P2(x2,y2) i.e.(60000,36000).
- The formula for calculating the distance is simple geometric distance formula.
- $d(x,y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Lets calculate the distance between the points in the next slide.



	Salary	Bank Balance
P1	20000	14000
P2	25000	18000
P3	60000	36000

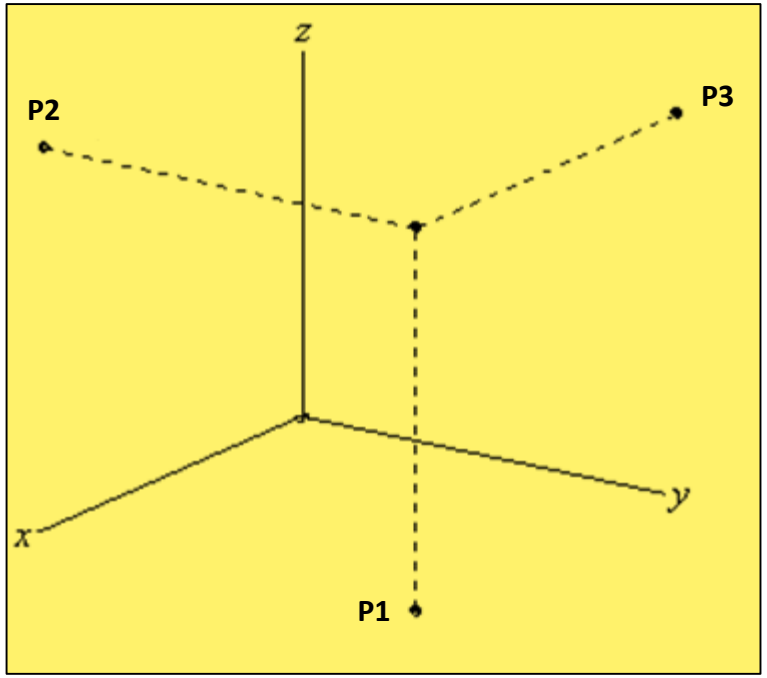
Introduction to clustering

- So, Lets measure the distance between P1, P2 and P3 points.
- The Euclidean distance between point P1 and P2 is $\sqrt{(25000-20000)^2+(18000-14000)^2}$ = 6403.12
- The Euclidean distance between point P1 and P3 is $\sqrt{(60000-20000)^2+(36000-14000)^2}$ = 45650.85
- The Euclidean distance between point P2 and P3 is $\sqrt{(60000-25000)^2+(36000-18000)^2}$ = 39357.34
- The Euclidean distance between P1 and P2 is lowest. Hence, P1 and P2 will form one cluster.

Introduction to clustering

- We have already calculated Euclidean distance in 2 dimensional plane
- Now, lets consider bank balance and age along with salary of a person.
- Now, we have a 3 dimensional plane and have to form a cluster of it.
- Here, we have to use the same procedure discussed above but with three variable.
- Suppose we have P1(x1,y1,z1) i.e. P1(20000,14000,25), P2(x2,y2,z2) i.e. P2(25000,18000,33) and P3(x3,y3,z3) i.e. (60000,36000,45) variables and have to form a cluster.
- But, to form a cluster we need to do scaling of the variables because of different units of measurement.
- Let's discuss scaling in the next slides

	Salary	Bank Balance	Age
P1	20000	14000	25
P2	25000	18000	33
P3	60000	36000	45



Data Preparation

- Before using the actual clustering algorithm we need to prepare a data for clustering.
- There are two important concepts of clustering when it comes to data preparation for clustering.
    1. Scaling
- Lets discuss scaling first
- Scaling is the concept of adjusting the values of the variables to take into account the fact that different variables are measured on very different scales.

Data Preparation - Scaling

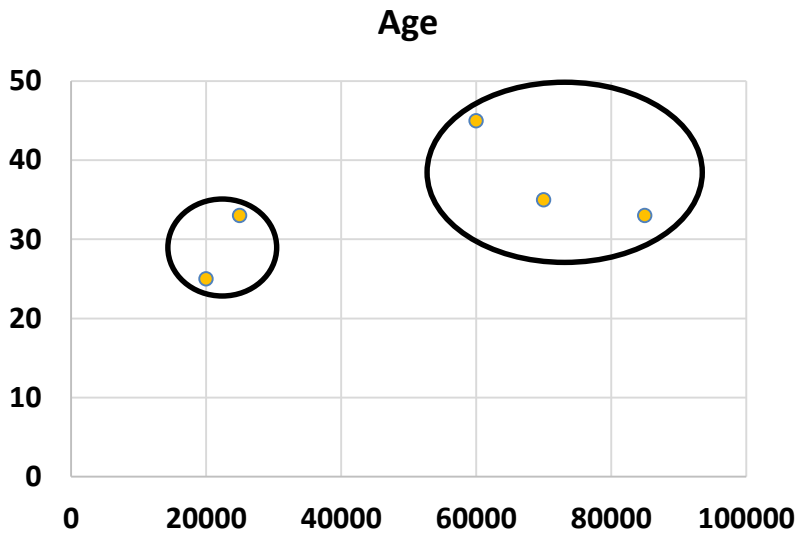
- Lets take an example discussed in previous slide where Salary and bank balance have same unit of measurement and age have different unit of measurement.
- Here if you see the scale of the age is varying between 25 - 45 and salary is varying between 20,000 – 85,000.
- The scale of measurement of age and the scale of measurement of salary, so all the difference on the sale variable gets magnified a lot more.
- But the algorithm calculates the distance and the square of the differences on both axis adds that and take square root of it.

	Salary	Bank Balance	Age
P1	20000	14000	25
P2	25000	18000	33
P3	60000	36000	45
P4	70000	48000	35
P5	85000	50000	33

Data Preparation - Scaling

- Plot a data points on graph, we see the age on the graph.
- Look at data and graph, a person P3 having a salary of 60000 is not going to be different from a Person P4 having salary 70000 But the age is going to be different.
- The actual clusters that should be formed are these.
- The clusters that are formed are not real clusters here. Because, the difference between the scales of the variables that we are measuring is very different.
- In order to do effective clustering, we need to adjust the variables to a common scale.

	Salary	Bank Balance	Age
P1	20000	14000	25
P2	25000	18000	33
P3	60000	36000	45
P4	70000	48000	35
P5	85000	50000	33



Data Preparation - Scaling

- The first regards the relative scales of the variables being measured.
- The available cluster analysis algorithms depend on the concept of measuring the distance or some other measure of similarity between the different observations we're trying to cluster.
- If one of the variables is measured on a much larger scale than the other variables, then whatever measure we use will be overly influenced by that variable.
- The traditional way of scaling variables is z-scores i.e. to subtract variable from mean, and divide by their standard deviation t

Data Preparation - Scaling

- This always have a mean of zero and variance of one.
- In case of variables that contain outliers, then this sort of standardization may be too severe, scaling down the outlying observations so that they appear to be closer to the others.
- One alternative is to use the median absolute deviation in place of the standard deviation;
- Subtract variable from median and divide by either the interquartile range or the median absolute deviation.
- For the common methods of measuring distances, centering the data by subtracting the mean or median is not really critical; it's the division by an appropriate scaling factor that's important.

Types of clustering

- These methods of clustering is used to deal with large amount of data.

There are two types of clustering.

1. Hierarchical clustering: This algorithm operates on the principle that data-point closer to base point will behave more similar compared to a data-point which is far from base point.
2. These find successive clusters using previously established clusters.
   - ✓ Agglomerative ("bottom-up"): Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.
   - ✓ Divisive ("top-down"): Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2. Partitional clustering: Partitional algorithms determine all clusters at once. They include:
   - ✓ K-means and derivatives - k means algorithm is able to handle large number of data points.

K means Clustering Example

- Lets take an example of First Cry to understand how clustering works.
- First Cry is a online shopping platform and selling bunch of product categories, to understand how clustering works.
- Let us consider only two product categories and its daily sales for a while.
- Product Categories – Baby Products and clothing.
- In this table, we can see day wise sale of baby products and clothing for 10 days.

Sales	Baby Products	Clothing
Store 1	13	20
Store 2	14	15
Store 3	10	8
Store 4	14	18
Store 5	9	6

K means Clustering

- K-means clustering is a type of unsupervised learning, which is used when we have unlabelled data.
- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable k.
- The algorithm works iteratively to assign each data point to one of k groups based on the features that are provided.
- Data points are clustered based on feature similarity.

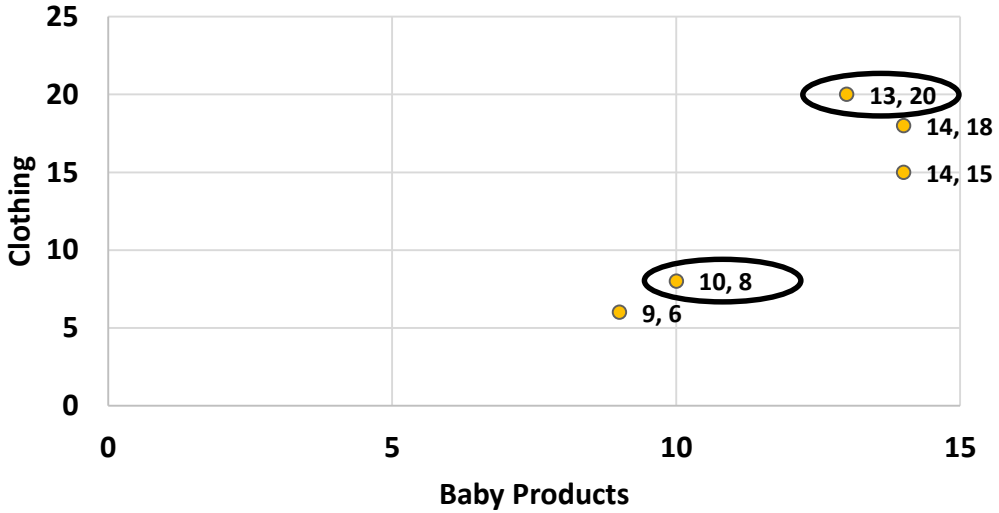
K means Clustering

- The *k-means* algorithm is implemented in 4 steps (assumes partitioning criteria is: maximize intra-cluster similarity and minimize inter-cluster similarity)
- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative reallocation technique to improve the partitioning by moving objects from one group to other.
- Initial centroids are often chosen randomly. Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance
- K-means will converge for common similarity measures.

K means Algorithm

- Lets see how to solve k means algorithm using First Cry example discussed previously.
- Lets take a sample of 5 days to solve using k means algorithm.
- **Step 1:** Specify desired number of clusters k.
- Let us choose arbitrarily, k = 2 for these 5 data points in 2-D space.
- Initially, we considered Day 1 and Day 3.

Sales	Baby Products	Clothing
Day 1	13	20
Day 2	14	15
Day 3	10	8
Day 4	14	18
Day 5	9	6



K means example

- **Step 2:** Randomly assign each data point to a cluster:
- To assign data point to a cluster we need to calculate distance between two points and this calculated using Euclidean distance.
- $d(x,y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- where x2 and y2 are data points and x1 and y1 are Centroid.
- Lets calculate distance between Centroid1(13,20) and day 2(14,15).
- $d(x,y) = \sqrt{(14 - 13)^2 + (15 - 20)^2}$ = 5.09
- Similarly, calculate distance between Centroid 2(10,8) and Day 2(14,15).
- $d(x,y) = \sqrt{(14 - 10)^2 + (15 - 8)^2}$ = 8.1
- Similarly, we have calculated the distance between centroids and remaining Days shown in table.

	Sales	Baby Products	Clothing
Centroid 1	Day 1	13	20
	Day 2	14	15
Centroid 2	Day 3	10	8
	Day 4	14	18
	Day 5	9	6

Sales	Dist. from C1	Dist. from C2
Day 1	**0.0**	12.4
Day 2	**5.1**	8.1
Day 3	12.4	**0.0**
Day 4	**2.2**	10.8
Day 5	14.6	**2.2**

K means example

- Now consider the minimum distance between the clusters.
- In first iteration, we can see that for Day 1, Day 2 and Day 4 have the minimum distance from C1.Hence, it's assigned to Cluster 1.
- Similarly for Day 3 and Day 5 have the minimum distance from C2. Hence, it's assigned to Cluster 2.
- As we selected clusters arbitrarily, hence we need to cross check it weather the clusters are formed are correct or not.
- To cross check it we need to again do the same procedure again but with new mean.

Sales	Dist. from C1	Dist. from C2	Clusters
Day 1	**0.0**	12.4	1
Day 2	**5.1**	8.1	1
Day 3	12.4	**0.0**	2
Day 4	**2.2**	10.8	1
Day 5	14.6	**2.2**	2

K means example

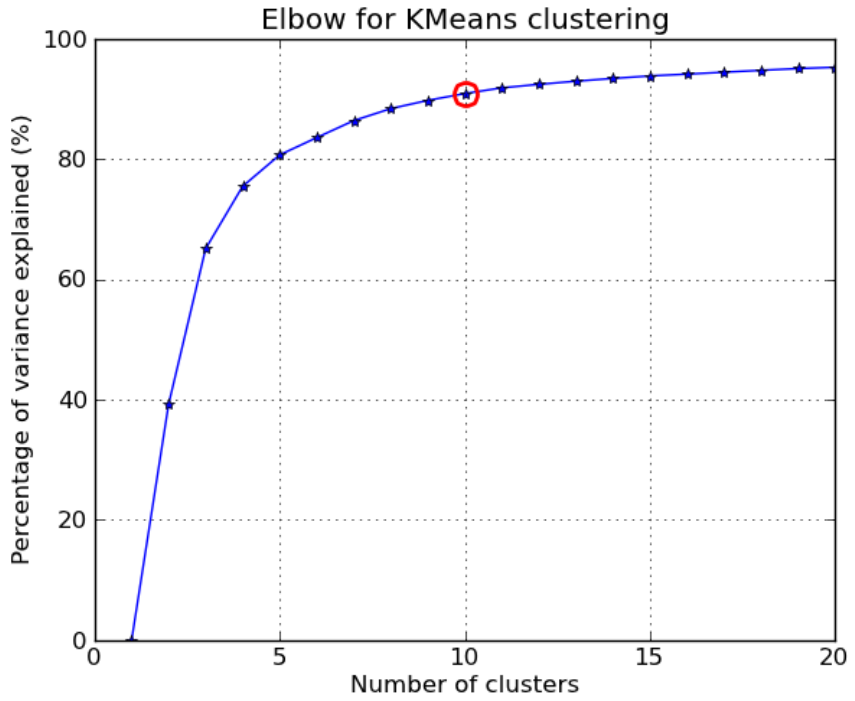
- **Step 3:** Now the new mean(Centroid) is the average of the data points in a cluster.
- Here, new centroid for cluster 1 is $\left(\frac{13+14+14}{3}, \frac{20+15+18}{3}\right)$ i.e. (13.7, 17.7).
- New centroid for cluster 2 is $\left(\frac{10+9}{2}, \frac{8+6}{2}\right)$ i.e. (9.5, 7.0).
- Now again calculate distance between individual data points and centroids using same formula.
- We will get the distance mentioned in Table.
- In the 2nd iteration we can see that the clusters have not changed.
- Hence, algorithm has converged.
- If the algorithm is not converged, then we need to iterate it again and again till the time its not converged.
- **Hence, we get 2 clusters, Cluster 1 having Day 1, Day 2 and Day 4 sale and Cluster 2 having Day 3 and Day 5 Sale.**

Sales	Baby Products	Clothing	Clusters
Day 1	13	20	1
Day 2	14	15	1
Day 3	10	8	2
Day 4	14	18	1
Day 5	9	6	2

Sales	Dist. from C1	Dist. from C2	Cluster
Day 1	2.4	13.5	1
Day 2	2.7	9.2	1
Day 3	10.3	1.1	2
Day 4	0.5	11.9	1
Day 5	12.6	1.1	2

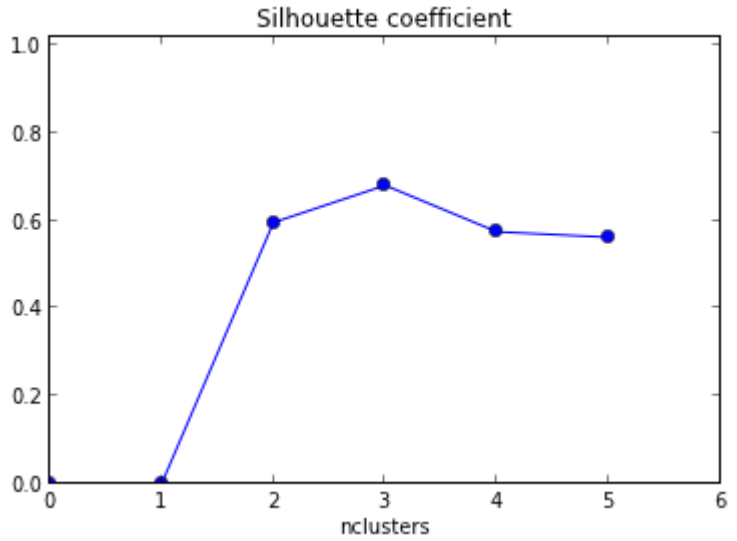
How to decide the numbers of Clusters or K means

- There are too many methods to determine the number of clusters but generally elbow, silhouette and gap statistic methods are used.
- The idea behind clustering methods is to define clusters such that the intra-cluster variation or total within-cluster sum of square (WSS) is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.
- **Elbow method:** The Elbow method looks at the total WSS as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

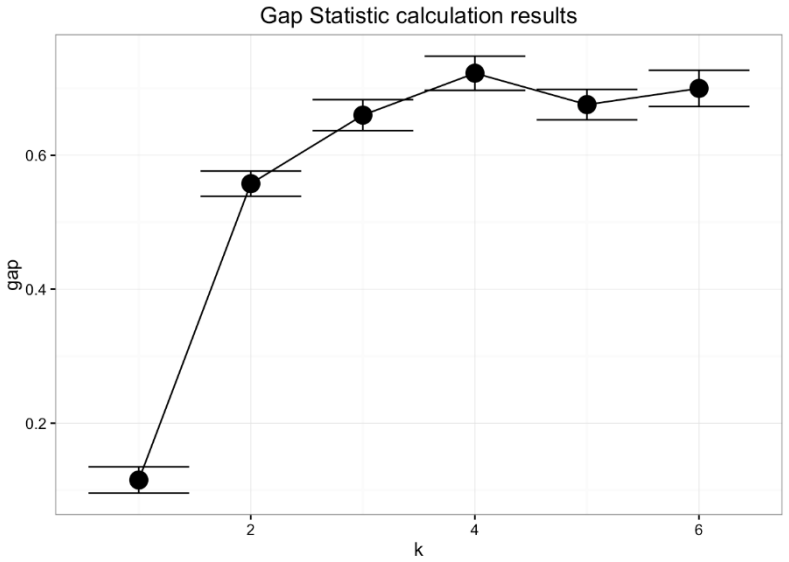


How to decide the numbers of Clusters or K means

- **Average silhouette method:** It computes the average silhouette of observations for different values of k.
- It measures the quality of a clustering i.e.it determines how well each object lies within its cluster.
- A high average silhouette width indicates a good clustering.
- The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k.

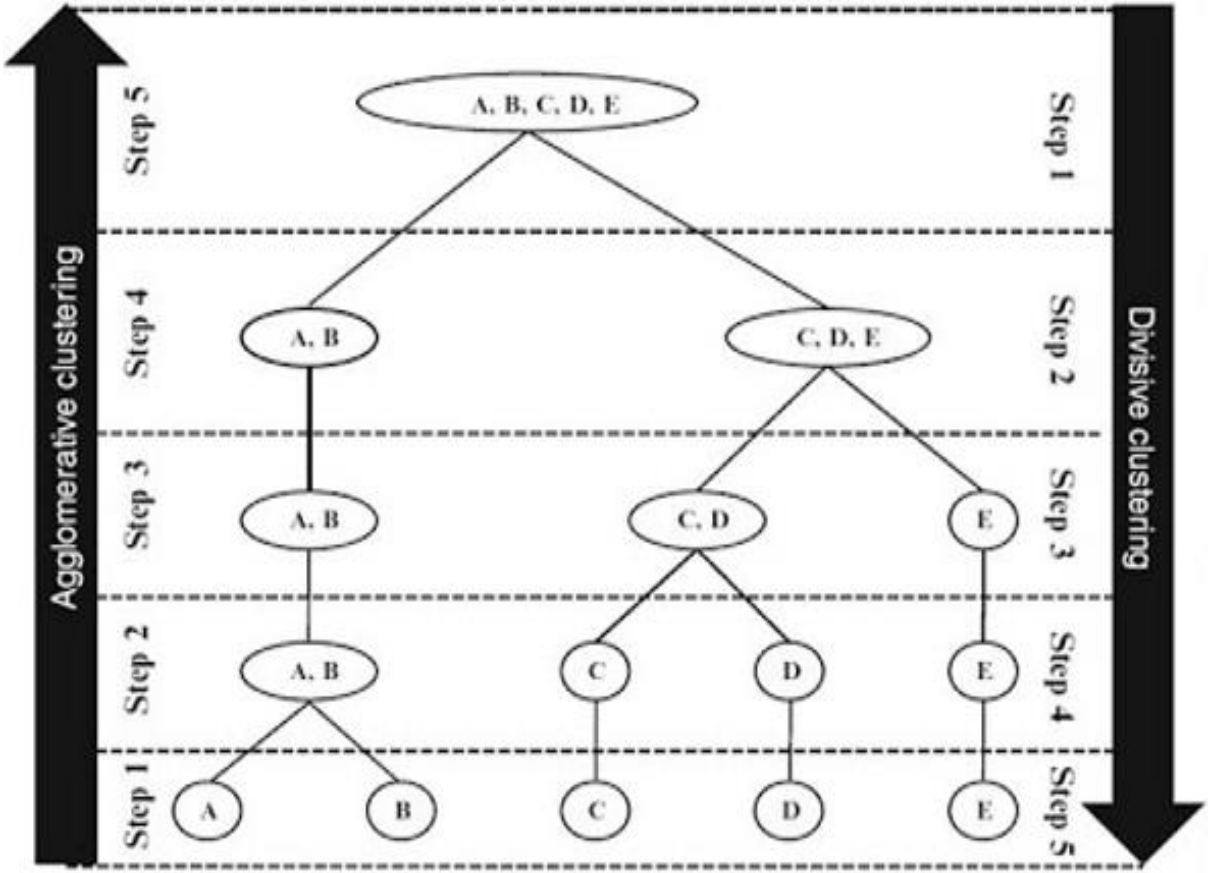


- **Gap statistic method:** The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data.
- The estimate of the optimal clusters will be the value that maximizes the gap statistic.
- This means that the clustering structure is far away from the random uniform distribution of points.



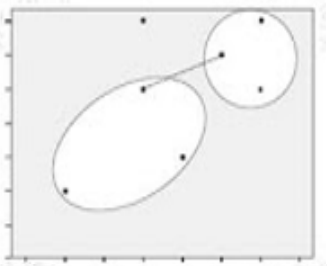
Hierarchical Clustering

- Hierarchical clustering is the another method to form clusters.
- Hierarchical clustering is an algorithm that builds hierarchy of clusters.
- Hierarchical clustering are of 2 types
  1. Agglomerative Clustering - It starts with all observations as a cluster and with each step combine observations to form one large cluster.
  2. Divisive Clustering - It starts with one large cluster and proceeds to split into smaller cluster items that are most dissimilar.
- The result of hierarchical clustering can be shown using dendrogram.

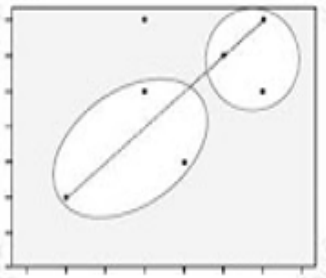


Hierarchical Clustering

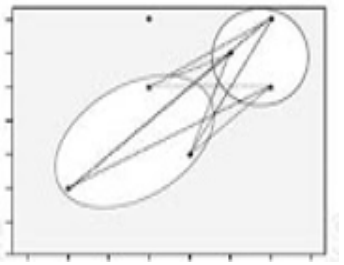
- **Single-link Method:** The distance between two clusters is the minimum of the distances between all pairs of patterns drawn one from each cluster.



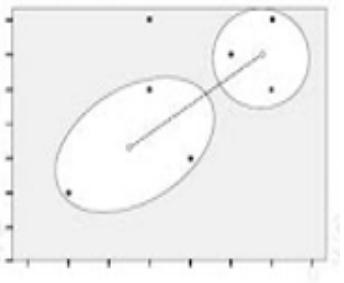
- **Complete-link Method:** The distance between two clusters is the maximum of all pair wise distances between pairs of patterns drawn one from each cluster.



- **Average-link Method:** The distance between two clusters is the average of all pair wise distances between pairs of patterns drawn one from each cluster (which is the same as the distance between the means in the vector space case – easier to calculate).

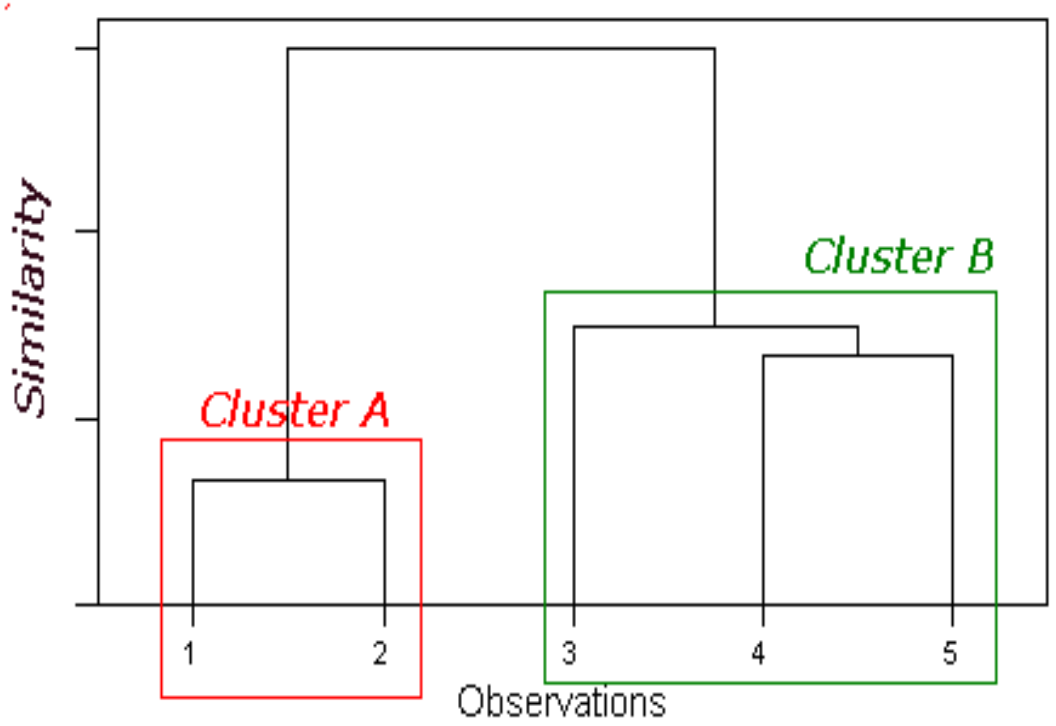


- **Centroid Method:** The geometric center is computed. The distance between two clusters equals the distance between two centroids.



Dendrogram

- Observations are on x-axis.
- The y-axis is a measure of closeness of either individual data points or clusters.
- Longer the line indicates the clusters are clearly apart from each other
- This helps in determining the number of optimal clusters
- The red and green line indicates the number of clusters



Hierarchical Clustering Example

- Lets take an example of hierarchical clustering using single linkage method.
- To solve a problem, first we need to find out the distance matrix using Euclidean distance formula.
- Here, we have calculated a distance between 1 store to other stores.
- The upper diagonal and the lower diagonal elements of the matrix are same hence, here, we have considered only lower diagonal values.

Sales	Baby Products	Clothing
Store 1 (S1)	13	20
Store 2 (S2)	14	15
Store 3 (S3)	10	8
Store 4 (S4)	14	18
Store 5 (S5)	8	10

Distance Matrix					
	Store 1 (S1)	Store 2 (S2)	Store 3 (S3)	Store 4 (S4)	Store 5 (S5)
Store 1 (S1)	0				
Store 2 (S2)	5.10	0			
Store 3 (S3)	12.37	8.06	0		
Store 4 (S4)	2.24	3.00	10.77	0	
Store 5 (S5)	11.18	7.81	2.83	10	0

Hierarchical Clustering Example

- For hierarchical clustering using single linkage method we need to consider a lowest distance in the distance matrix.
- Here, 2.24 is the lowest distance between Store 1 to Store 4.
- Hence, Store 1 and Store 4 form 1 cluster.

Distance Matrix					
	Store 1 (S1)	Store 2 (S2)	Store 3 (S3)	Store 4 (S4)	Store 5 (S5)
Store 1 (S1)	0				
Store 2 (S2)	5.10	0			
Store 3 (S3)	12.37	8.06	0		
Store 4 (S4)	2.24	3.00	10.77	0	
Store 5 (S5)	11.18	7.81	2.83	10	0

Hierarchical Clustering Example

- Once the cluster is formed, update the distance matrix with new cluster (S1, S4)
- Now calculate the minimum distance between (S1,S4) to S2,S3,S5.
- For Example, min(dist(S1,S4),S2) means calculate the min distance between (S1,S2) and (S4,S2))
- The distance between (S1,S2) is 5.10 and distance between (S4,S2) is 3.
- The minimum distance between S4 and S2 is 3 as compared to 5.10 of (S1,S2)
- Hence, the distance between (S1,S4) and S2 is 3
- Similarly, we have calculated distance between (S1,S4) and S3, S5.

Update matrix with (S1,S4)			
Minimum distance between cluster and Points	Calculation of minimum distance	Distance between points	Minimum distance
min(dist(S1,S4),S2)	d((S1, S2), (S4,S2))	(5.10,3)	3
min(dist(S1,S4),S3)	d((S1,S3),(S4,S3))	(12.37,10.77)	10.77
min(dist(S1,S4),S5)	d(S1,S5),(S4,S5))	(11.18, 10)	10

Hierarchical Clustering Example

- Update the distance matrix based on calculation in the previous slide (marked in green)
- Now, again repeat the same procedure,
- Select the lowest distance
- The lowest distance is 2.83 between (S3,S5). So, (S3,S5) form 1 cluster
- Now, again update matrix with cluster (S3,S5) and find out distance between (S3,S5) and (S1,S4), S2,S3.

Distance Matrix				
	S1,S4	S2	S3	S5
S1,S4	0			
S2	3	0		
S3	10.77	8.06	0	
S5	10	7.81	2.83	0

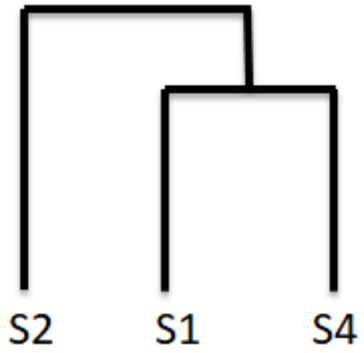
Update matrix with (S3,S5)			
Minimum distance between cluster and Points	Calculation of minimum distance	Distance between points	Minimum distance
min(dist(S3,S5),(S1,S4))	d((S3,(S1,S4)),(S5,(S1,S4))	(10.77,10)	10
min(dist(S3,S5),S2)	d((S3,S2),(S5,S2))	(8.06,7.81)	7.81

Hierarchical Clustering Example

- Update distance matrix with new distance and cluster.
- Again choose the lowest distance and form a cluster and find out the distance between points.

Distance Matrix			
	S1,S4	S2	S3,S5
S1,S4	0		
S2	3	0	
S3,S5	10	7.81	0

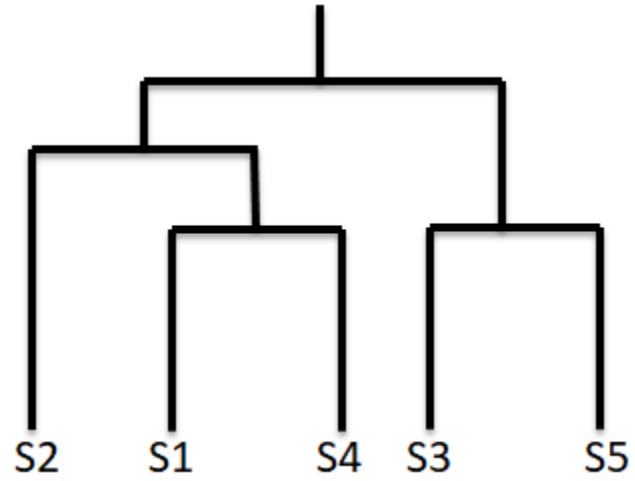
Update matrix with ((S1,S4),S2)			
Minimum distance between cluster and Points	Calculation of minimum distance	Distance between points	Minimum distance
min((S1,S4),S2),(S3, S5))	d(((S1,S4),(S3, S5)), (S2,(S3, S5)))	(10,7.81)	7.81



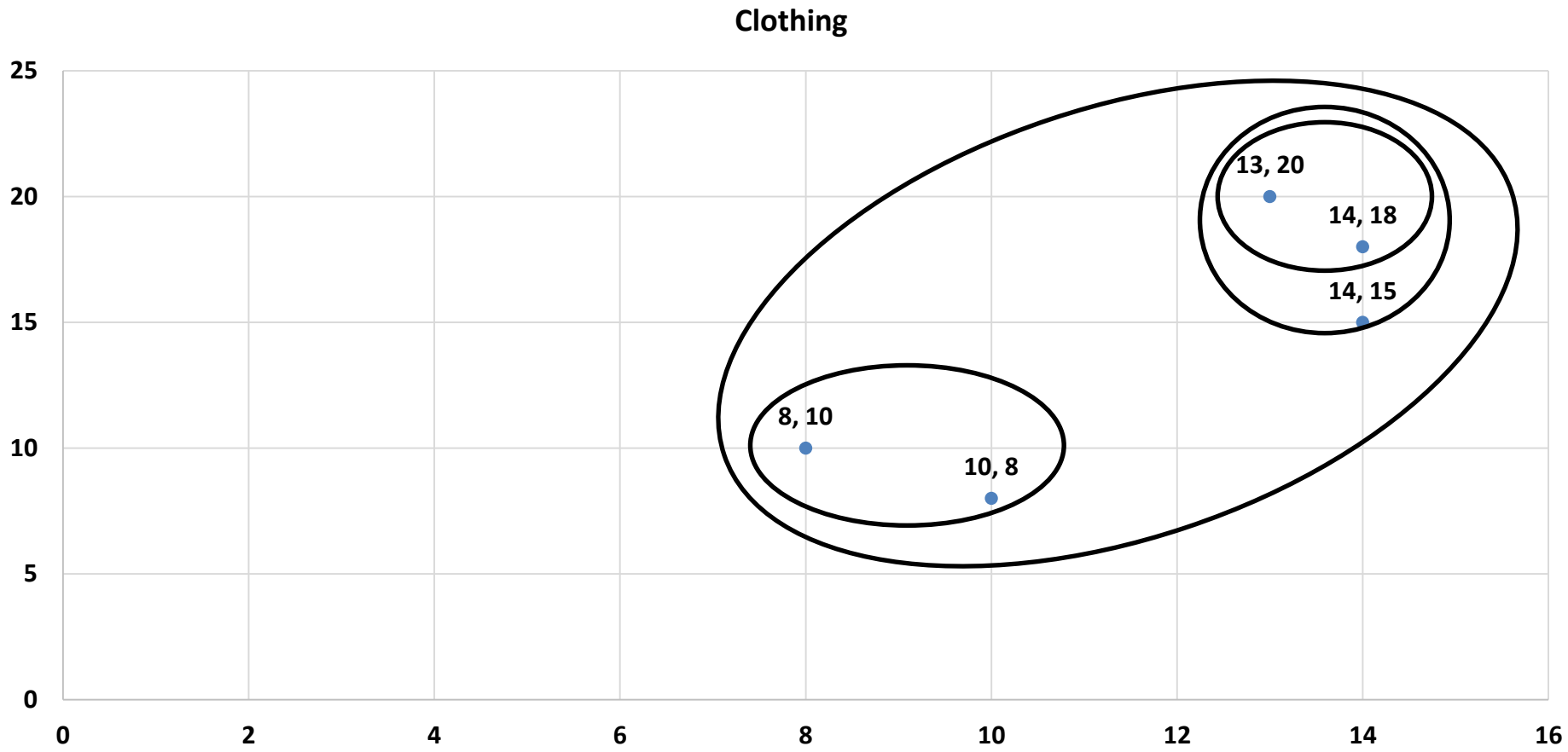
Hierarchical Clustering Example

- The updated new distance matrix is
- The new clusters are
  - ✓ Cluster 1 – Store 1 and Store 4
  - ✓ Cluster 2 – Store 3 and Store 5
  - ✓ Cluster 3 – Store 1, Store 4 and Store 2
  - ✓ Cluster 4 – Store 1, Store 4, Store 2, Store 3 and Store 5

Distance Matrix		
	(S1,S4),S2	S3,S5
(S1,S4),S2	0	
S3, S5	7.81	0



Hierarchical clustering Example



Difference between k means and Hierarchical clustering

K means Clustering	Hierarchical Clustering
K means clustering can handle big data well. Because time complexity of k means is linear i.e. O(n)	Hierarchical clustering can't handle big data well Because time complexity of hierarchical clustering is quadratic i.e. O(n2)
In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.	In Hierarchical clustering, results are reproducible
K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into.	But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.
K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering is not found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D)

Thank You.