CLUSTERING

KMEANS & AGGLEMERATIVE HIERARCHICAL CLUSTERING

Mr.Gangadhar Immadi

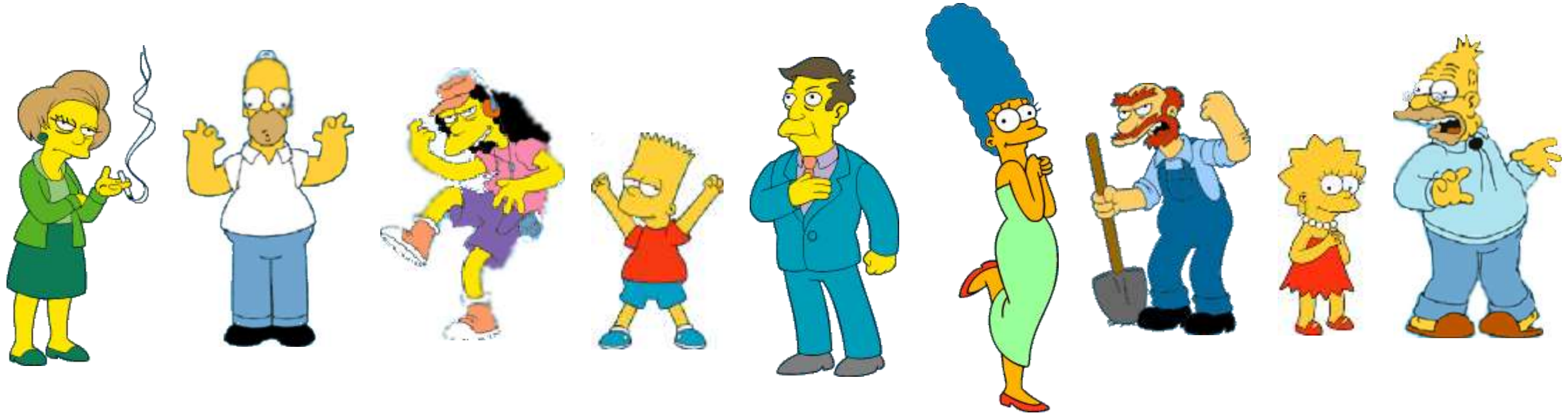
immadi.gangadhar@gmail.com

9986789040

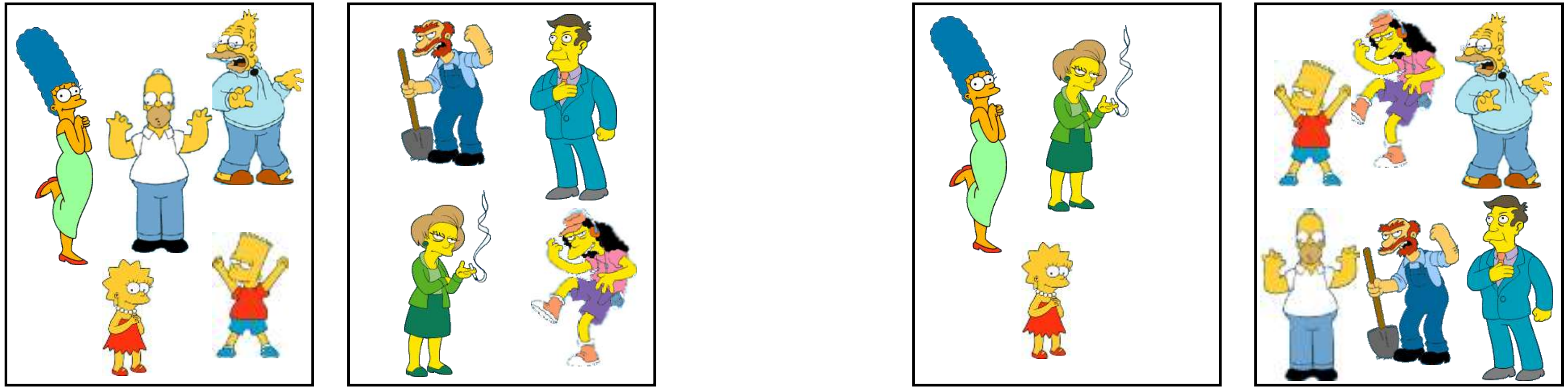
Clustering

- Unsupervised learning
- Organizing data into classes such that there is
  - high intra-class similarity
  - low inter-class similarity
- Finding the class labels and the number of classes directly from the data
- finding natural groupings among objects.

What is a natural grouping among these objects?



Clustering is subjective



What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

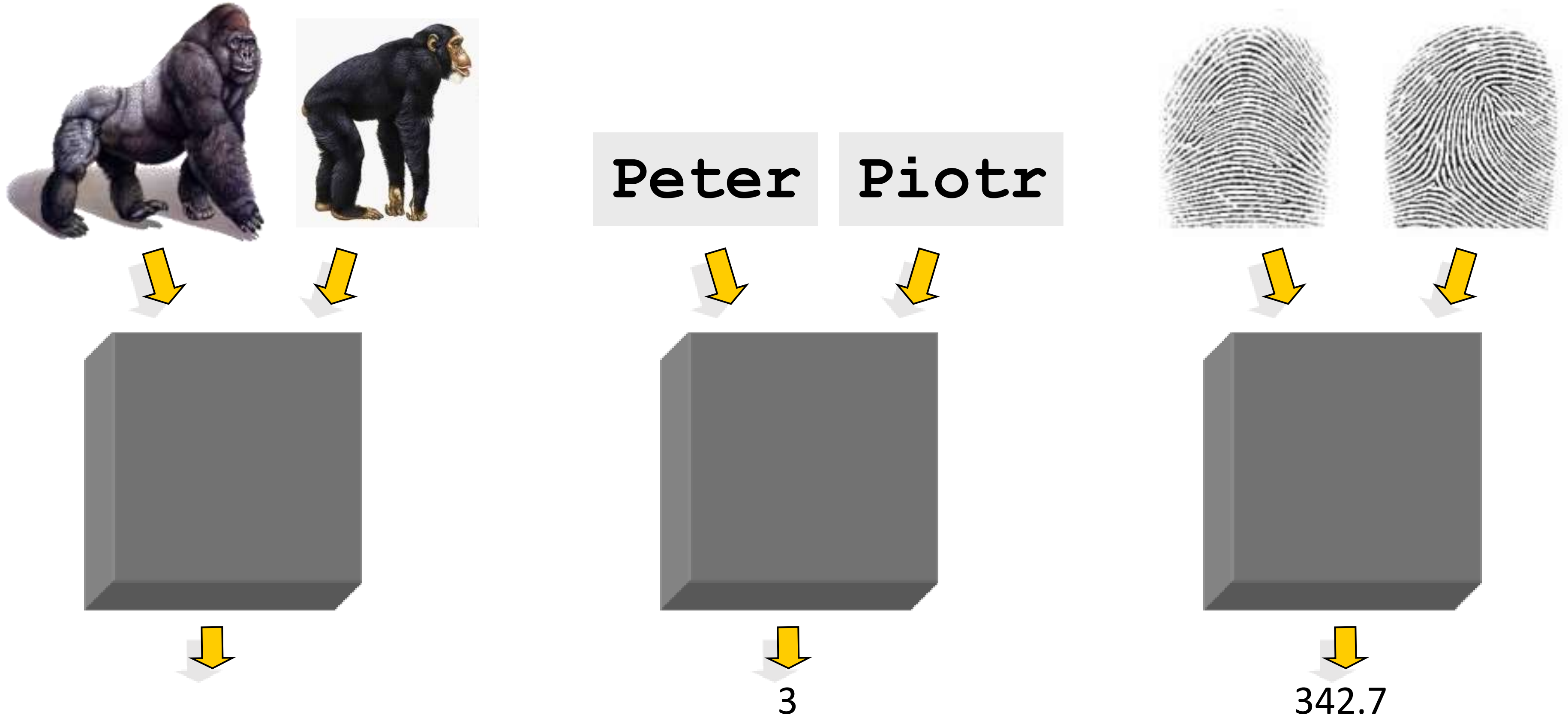
Webster's Dictionary

Similarity is hard to define, but…
"We know it when we see it"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Defining Distance Measures

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$



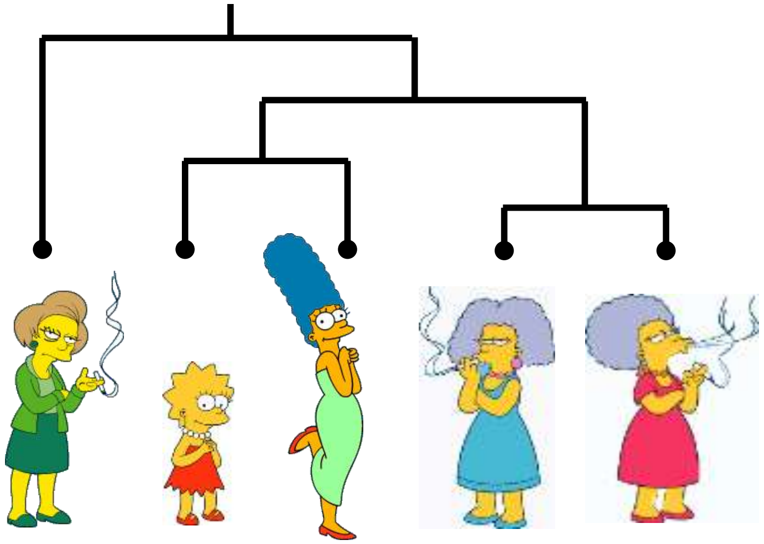
Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incorporation of user-specified constraints
- Interpretability and usability

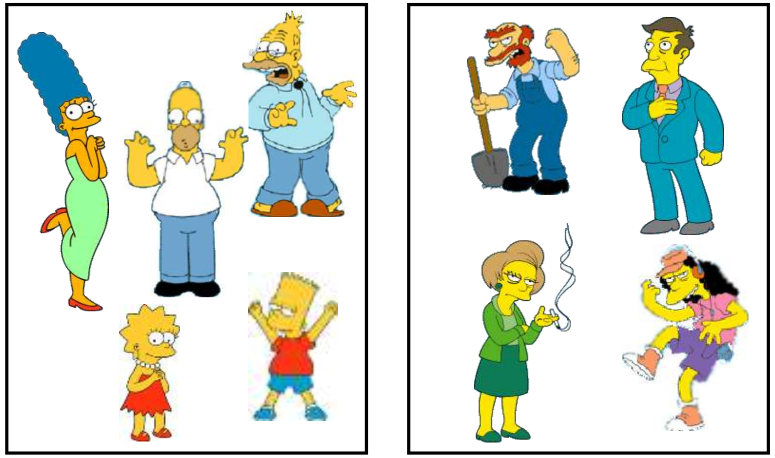
Two Types of Clustering

1. **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
2. **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical



Partitional

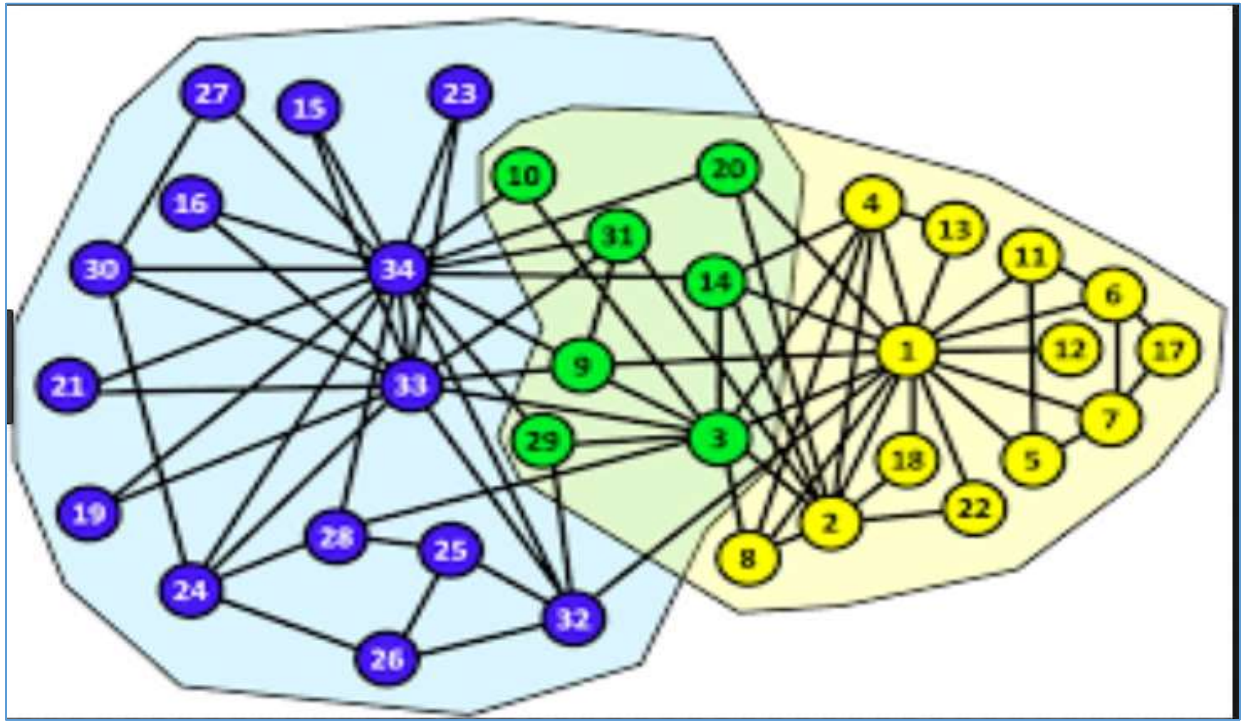


Clustering types on structure

Clustering is usually one of the first tasks performed in most analytics projects. It helps data scientists to analyze individual clusters further.

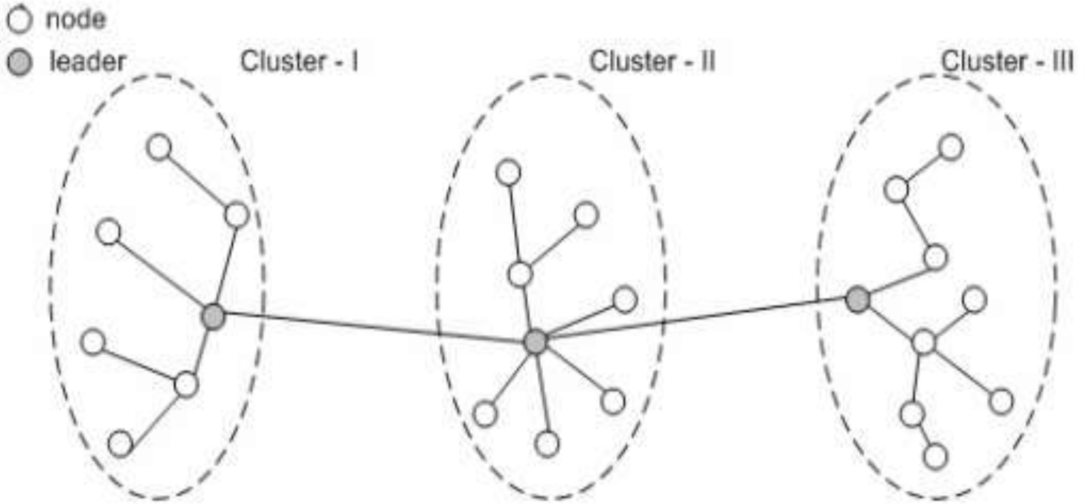
Overlapping clusters

- An observation may belong to more than one cluster



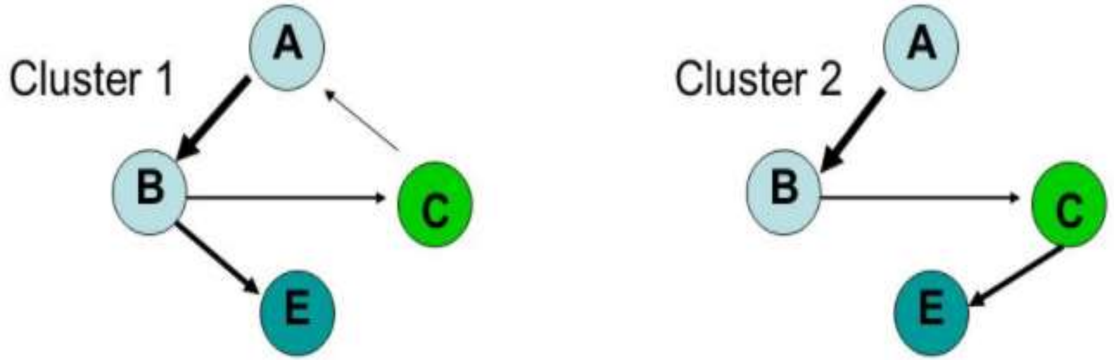
Non-overlapping clusters

Cluster in which each observation belongs to only one cluster. Non-overlapping clusters are more frequently used clustering techniques in practice.



Probabilistic clusters

An observation may belong to a cluster according to a probability distribution.



Similarity / Distance Measures - 1. Euclidean Distance

Euclidean is one of the frequently used distance measures when the data are either in interval or ratio scale.

The Eucledian distance between two *n*-dimensional observations X_1 $(x_{11}, x_{12}, \ldots, x_{1n})$ and X_2 $(x_{21}, x_{22}, \ldots, x_{2n})$ is given by

$$D(X_1, X_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \cdots + (x_{1n} - x_{2n})^2}$$

Standardized Euclidean Distance

If two features metrics and range are different which leads to skewed data.

Standardized value of the attribute = $\left(\dfrac{X_{ik} - \bar{X}_i}{\sigma_{X_i}} \right)$

Where $\bar{X}_i$ and σ_{X_i} are, respectively, the mean and standard deviation of i^{th} attribute

2. Manhattan Distance (City Block Distance)

Euclidean distance may not be appropriate while measuring distance between different locations (for example, distance between two shops in a city). In such cases, we use Manhattan distance, which is given by

$$DM(X_1, X_2) = \sum_{i=1}^{n} \left| X_{1i} - X_{2i} \right|$$

3. Minkowski Distance

Minsowski distance is the generalized distance measure between two cases in the dataset and is given by

$$\text{Minkowski } D(X_1, X_2) = \left(\sum_{i=1}^{n} \left| X_{1i} - X_{2i} \right|^{p} \right)^{1/p}$$

When p = 1, Minkowski distance is same as the Manhattan distance.

For p = 2, Minkowski distance is same as the Euclidean distance.

Clustering Algorithms

Clustering algorithms group data into finite number of mutually exclusive subsets.

Steps followed in clustering algorithms:

1. Variable selection.
2. Deciding the distance/similarity measure for measuring distance/dissimilarity between the observations.
3. Deciding the number of clusters.
4. Validation of the clusters.

1. Variable Selection

Ketchen and Shook (1996) suggest inductive, deductive, and cognitive approaches for variable selection.

- Inductive is basically an exploratory approach and starts with as many variables as possible.

- On the other hand, in deductive variable selection, suitability of the variable and theoretical basis influence selection of variables.

- Under cognitive variable selection, expert opinion plays a major role in variable selection

2. Deciding Distance/Similarity Measures

Choosing the right distance/similarity measure plays an important role in developing clusters.

3. Number of Clusters

Several approaches are available for deciding the number of clusters such as *CH* index , Hartigan statistic ,Silhouette statistic, and elbow method in which the ideal number of clusters is given by the position of elbow in an *L*-shaped curve.

Elbow : Calculate the ***Within-Cluster-Sum of Squared*** *Errors (WSS) for* ***different values of k****, and choose the k for which WSS becomes first starts to diminish.*

1. The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
2. The WSS score is the sum of these Squared Errors for all the points.
3. Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

4. Cluster Validation

The clusters created should be validated for consistency using different algorithms to ensure that the clusters represent the structures that exist in the population.

Halkidi *et al.* (2001) suggest the following measures to validate the clusters:

- **Compactness:** Closeness of each member of a cluster which can be measured through variance.

- **Separation:** Distance between different clusters.

K-Means Clustering

- K-means clustering is one of the frequently used clustering algorithms.
- It is a non-hierarchical clustering method in which the number of clusters (K) is decided *a priori*.

1. Decide on a value for k.

2. Initialize the k cluster centers (randomly, if necessary).

3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.

4. Re-estimate the k cluster centers, by assuming the memberships found above are correct.

5. If none of the N objects changed membership in the last iteration, exit. Otherwise goto 3.

Consider the following dataset and group them into 2 clusters. Use Euclidian distance

Actual Data		
Ob#	X	Y
1	1	1
2	1.5	1.5
3	1	0.5
4	0.8	1.2
5	3.3	3.1
6	2.58	3.68
7	3.5	2.8
8	3	3

Iteration - 1			
Ob#	D1	D2	Cluster #
1	0	2.82	1
2	0.70	2.12	1
3	0.5	3.2	1
4	0.28	2.84	1
5	3.11	0.31	2
6	3.11	0.79	2
7	3.08	0.53	2
8	2.82	0	2

Iteration - 2			
Ob#	D1	D2	Cluster #
1	0.9	2.99	1
2	0.61	2.29	1
3	0.55	3.37	1
4	0.31	3.0	1
5	3.02	0.2	2
6	3.03	0.74	2
7	2.99	0.53	2
8	2.74	0.17	2

Initial Centers			
	Ob#	Mean X	Mean Y
C1	1	1	1
C2	8	3	3

New Centers – after Iteration-1			
	Ob#	Mean X	Mean Y
C1	1,2,3,4	1.075	1.05
C2	5,6,7,8	3.095	3.145

Cluster the following eight points (with (x, y) representing locations) into three clusters:
A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2)

P(a, b) = |x2 – x1| + |y2 – y1|

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

We have only one point A1(2, 10) in Cluster-01.
cluster center remains the same. (2,10)

Center of Cluster-02
= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)
= (6, 6)

Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

Center of Cluster-01
= ((2 + 4)/2, (10 + 9)/2)
= (3, 9.5)

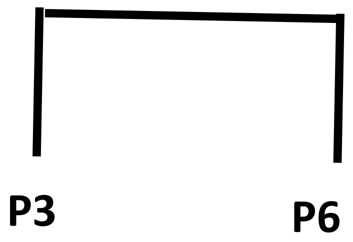
Center of Cluster-02
= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)
= (6.5, 5.25)

Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

Apply Agglomerative Hierarchical algorithm for the following dataset

Actual Data		
Ob#	X	Y
P1	0.4	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.3

Ob#	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.2	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0



- Recalculate the distance matrix

1.) MIN[DIST((P3,P6),P1] = MIN[DIST(P3,P1)(P6,P1)] = MIN(0.22,0.23) = 0.22

2.) MIN[DIST((P3,P6),P2)] = MIN[DIST((P3,P2),(P6,P2))] =MIN[0.15,0.25] = 0.15

3.) MIN[DIST((P3,P6),P4)] = MIN[DIST((P3,P4),(P6,P4))] = MIN[0.15,0.22] = 0.15

4.) MIN[DIST((P3,P6),P5)] = MIN[DIST((P3,P5),(P6,P5))] = MIN[0.28,0.39] = 0.28

Ob#	**P1**	**P2**	**P3,P6**	**P4**	**P5**
P1	0				
P2	0.23	0			
P3,P6	0.22	0.15	0		
P4	0.37	0.2	0.15	0	
P5	0.34	0.14	0.28	0.29	0

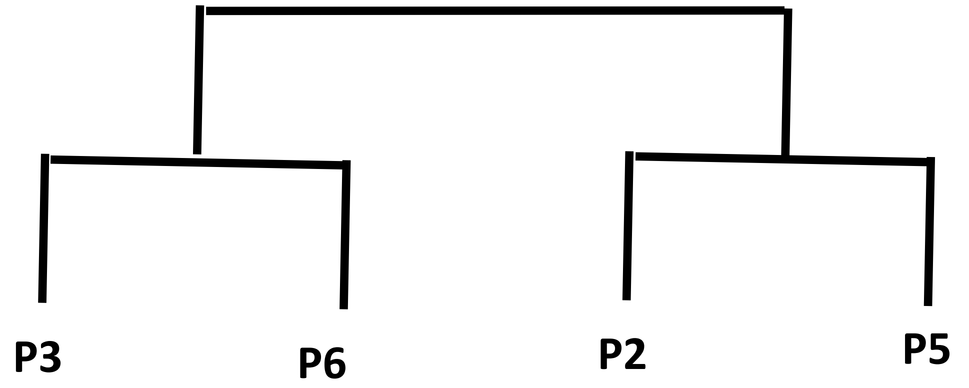
- Recalculate the distance matrix

1.) MIN[DIST((P2,P5),P1] = MIN[DIST(P2,P1)(P2,P5)] = MIN(0.23,0.34) = 0.23

2.) MIN[DIST((P2,P5),(P3,P6))] = MIN[DIST((P2,(P3,P6)),(P5,(P3,P6)))]
=MIN[0.15,0.28] = 0.15

3.) MIN[DIST((P2,P5),P4)] = MIN[DIST((P2,P4),(P5,P4))] = MIN[0.2,0.29] = 0.2

Ob#	**P1**	**P2,P5**	**P3,P6**	**P4**
P1	0			
P2,P5	0.23	0		
P3,P6	0.22	0.15	0	
P4	0.37	0.2	0.15	0

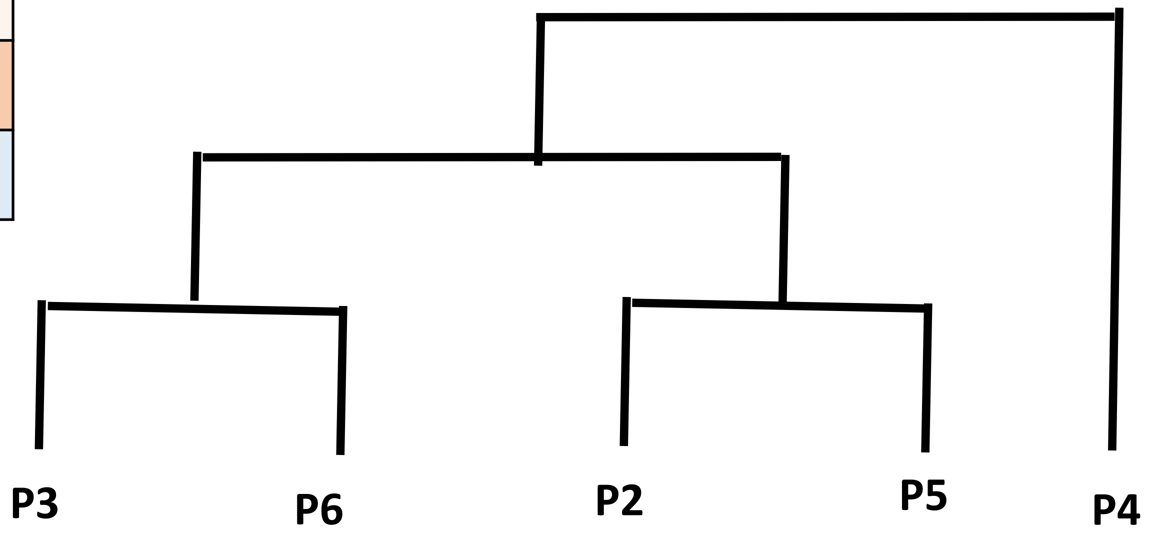


- Recalculate the distance matrix

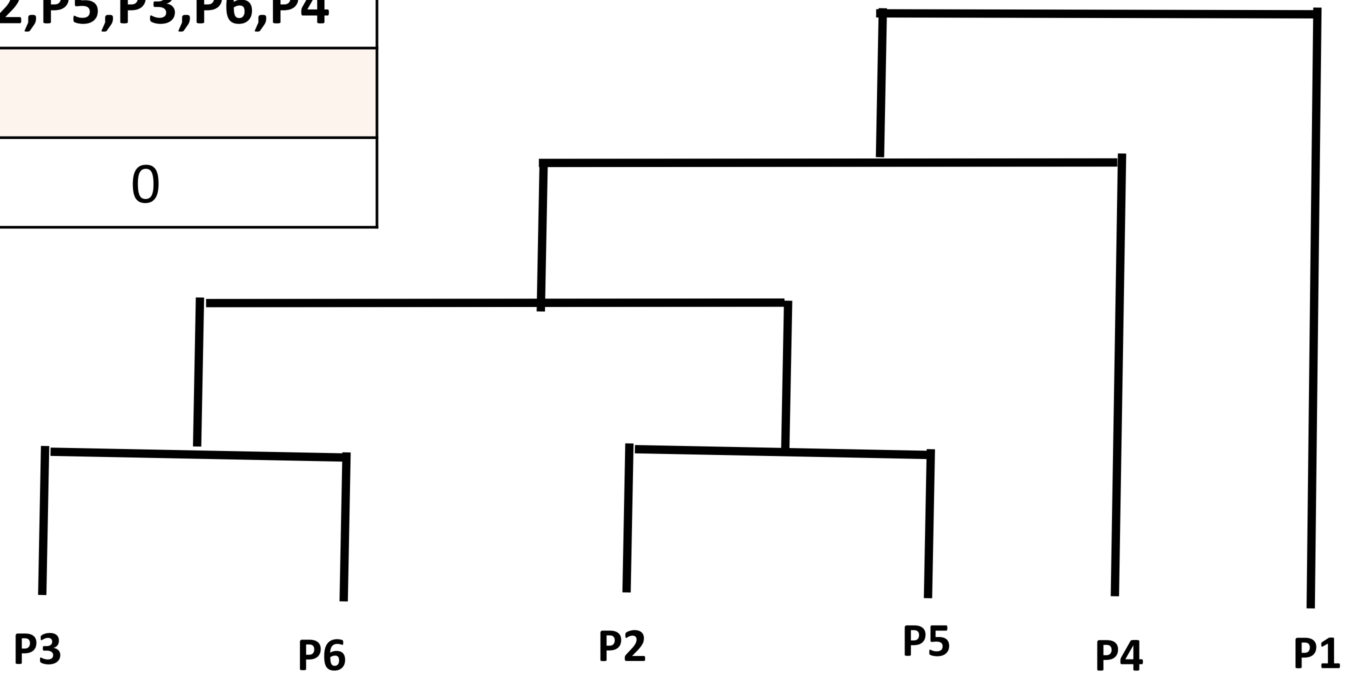
1.) MIN[DIST({(P2,P5),(P3,P6)},P1] = MIN[DIST({(P2,P5),(P1)}, {(P3,P6),(P1)}] = MIN(0.23,0.22) = 0.22

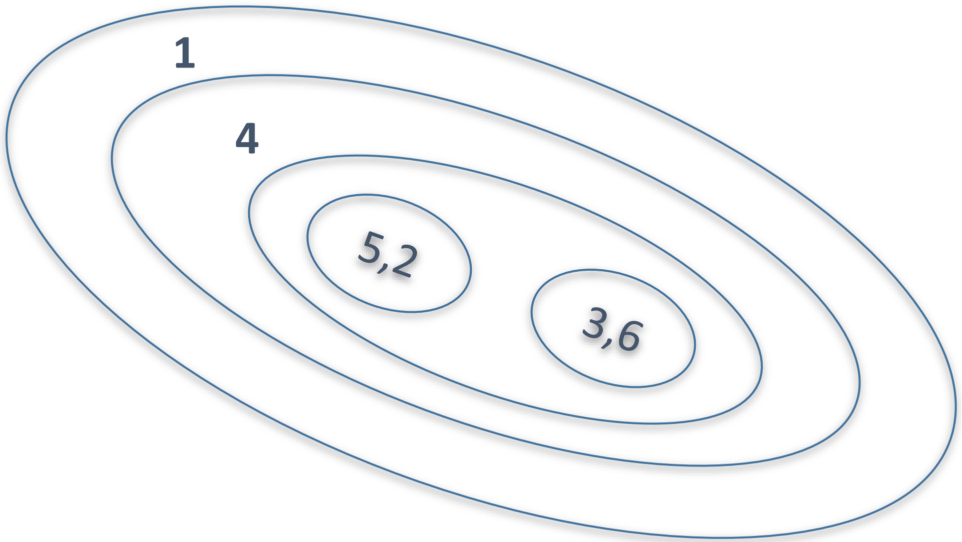
2.) MIN[DIST({(P2,P5),(P3,P6)},P4] = MIN[DIST({(P2,P5),(P4)}, {(P3,P6),(P4)}] = MIN(0.2,0.15) = 0.15

Ob#	**P1**	**P2,P5,P3,P6**	**P4**
P1	0		
P2,P5,P3,P6	0.22	0	
P4	0.37	0.15	0

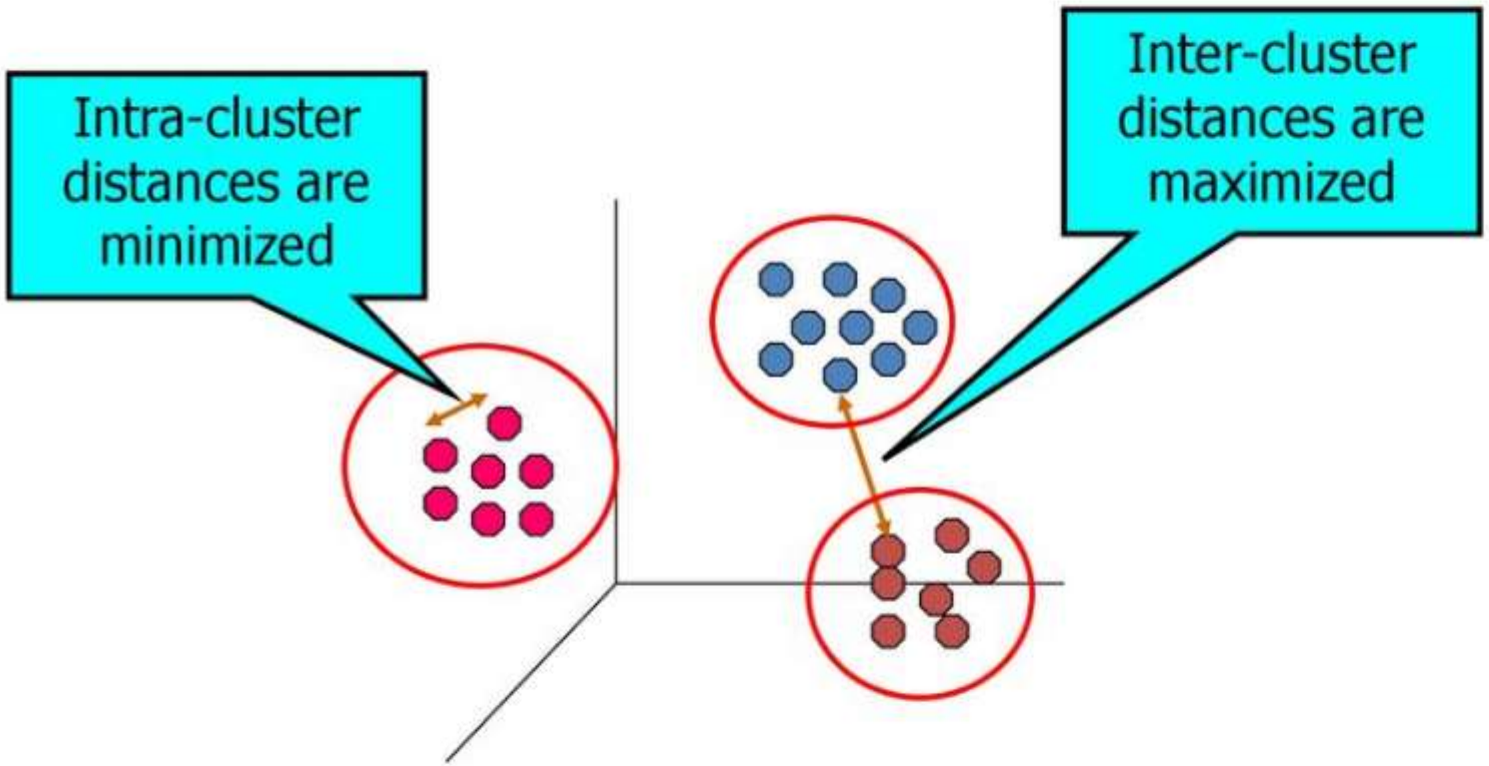


Ob#	P1	P2,P5,P3,P6,P4
P1	0	
P2,P5,P3,P6,P4	0.22	0

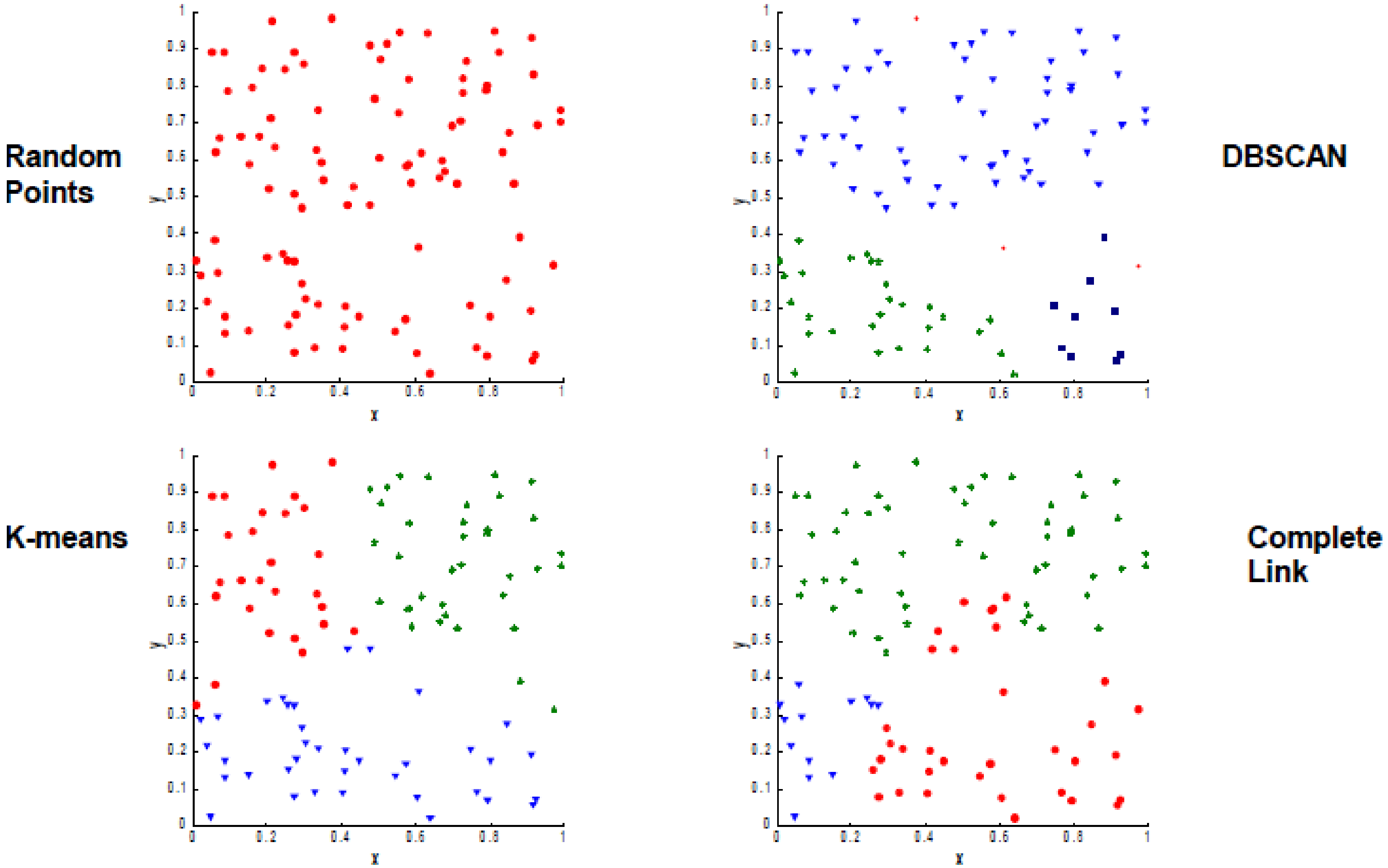




Performance Measure - Clustering



Clustering in Some random data

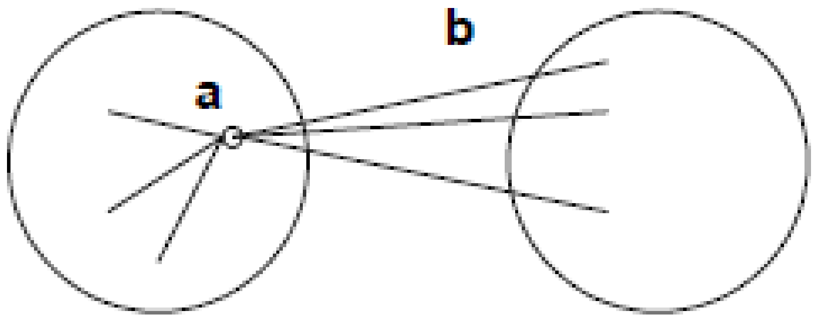


Different Aspects of Cluster Validation

- Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
- Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
- Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
  - Use only the data
- Comparing the results of two different sets of cluster analyses to determine which is better.
- Determining the 'correct' number of clusters.
- For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Internal Measures : Silhouettes coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, *I*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by
  - s = 1 – a/b if a < b, (or s = b/a - 1 if a >= b, not the usual case)
  - Typically between 0 and 1.
  - The closer to 1 the better.



Example : Consider the following clusters and compute Silhouettes coefficient for {1,0} of c1 with C2 and C3

- C1 = { {1,0}, { 1, 1} }
- C2 = { { 1, 2 } , { 2, 3 }, { 2 , 2 }, { 1, 2} }
- C3 = { { 3, 1 }, { 3, 3 }, { 2, 1 } }

- Cosider {1,0} in C1,

a = $\sqrt{(1-1)^2 + (0-1)^2} \;=\; 1$

For c2 rom {1,0} of c1

{1,0}-> {1,2} = $\sqrt{(1-1)^2 + (0-2)^2}$ = 2

{1,0}->{2,3} = $\sqrt{(1-2)^2 + (0-3)^2}$ = 3.16

{1,0}->{2,2} = $\sqrt{(1-2)^2 + (0-2)^2}$ = 2.24

{1,0}-> {1,2} = $\sqrt{(1-1)^2 + (0-2)^2}$ = 2

Average distance of point {1,0} in cluster 1 to all points in C2 = (2+3.16+2.24+2) / 4 = 2.325

- Similarly for cluster -3

{1,0}->{3,1} = $\sqrt{(1-3)^2+(0-1)^2}$ = 2.24

{1,0}->{3,3} = $\sqrt{(1-3)^2+(0-3)^2}$ = 3.61

{1,0}->{2,1} = $\sqrt{(1-2)^2+(0-1)^2}$ = 2.24

Average distance of point {1,0} of C1 to all points in cluster 3 =

(2.24+3.61+2.24)/3 = 2.7

b = min(2.325,2.7) = 2.325

Hence, Silhouettes coefficient of C1 = S1 = 1-(a/b) = 1-(1/2.325) = 0.569

Note : Clusters with greatest Silhouettes coefficient value is the best as per evaluation