Cluster Analysis

Dr. Zahoor Tanoli COMSATS

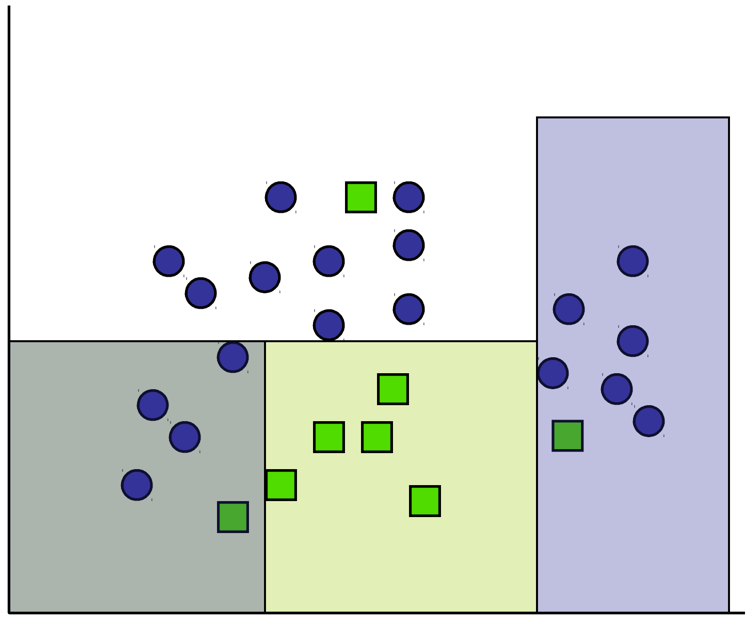
Lecture Outline

- What is Clustering
- Supervised and Unsupervised Learning
- Types of Clustering Algorithms
- Most Common Techniques
- Areas of Applications

Clustering - Definition

- Process of grouping similar items together
- Clusters should be very similar to each other but…
- Should be very different from the objects of other clusters/ other clusters
- We can say that intra-cluster similarity between objects is high and inter-cluster similarity is low
- Important human activity --- used from early childhood in distinguishing between different items such as cars and cats, animals and plants etc.

Classification vs. Clustering



Classification:

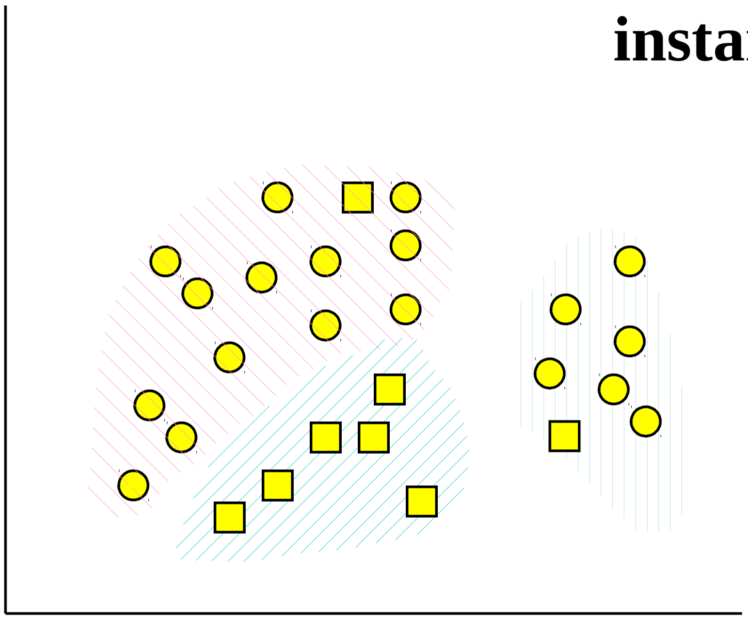
Supervised learning:

Learns a method for predicting the instance class from pre-labeled (classified) instances

Clustering

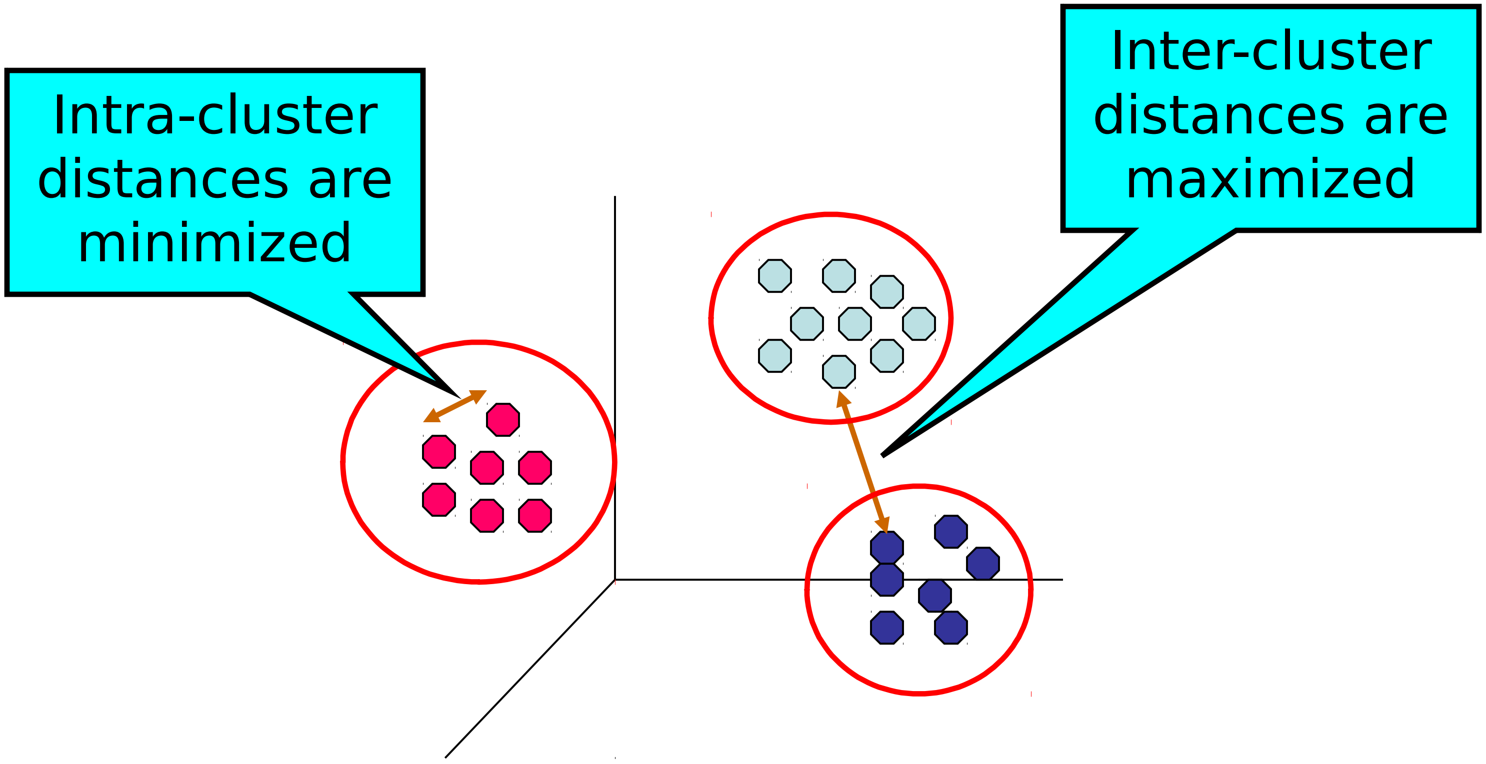
Unsupervised learning:

Finds "natural" grouping of instances given un-labeled data



What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

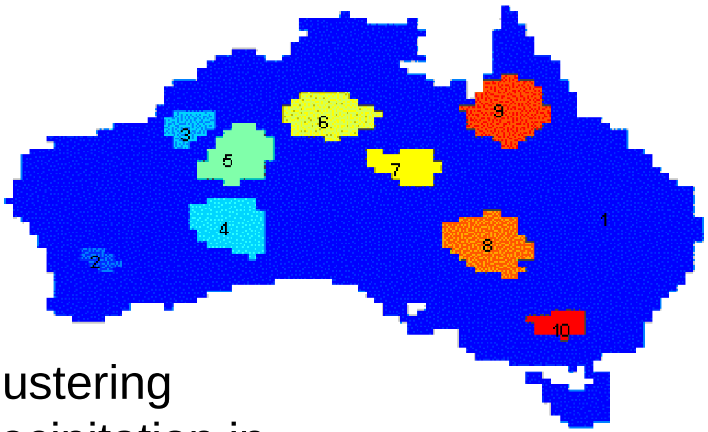


Applications of Cluster Analysis

- **Understanding**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

- **Summarization**
  - Reduce the size of large data sets



Clustering precipitation in Australia

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Types of Clustering Algorithms

- Clustering has been a popular area of research
- Several methods and techniques have been developed to determine natural grouping among the objects

Clustering Algorithms

- K-means and its variants
- Hierarchical clustering
- Density-based clustering

Simple Clustering: K-means

Works with numeric data only

1) Pick a number (K) of cluster centers (at random)
2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3) Move each cluster center to the mean of its assigned items
4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

Algorithm 1 Basic K-means Algorithm.

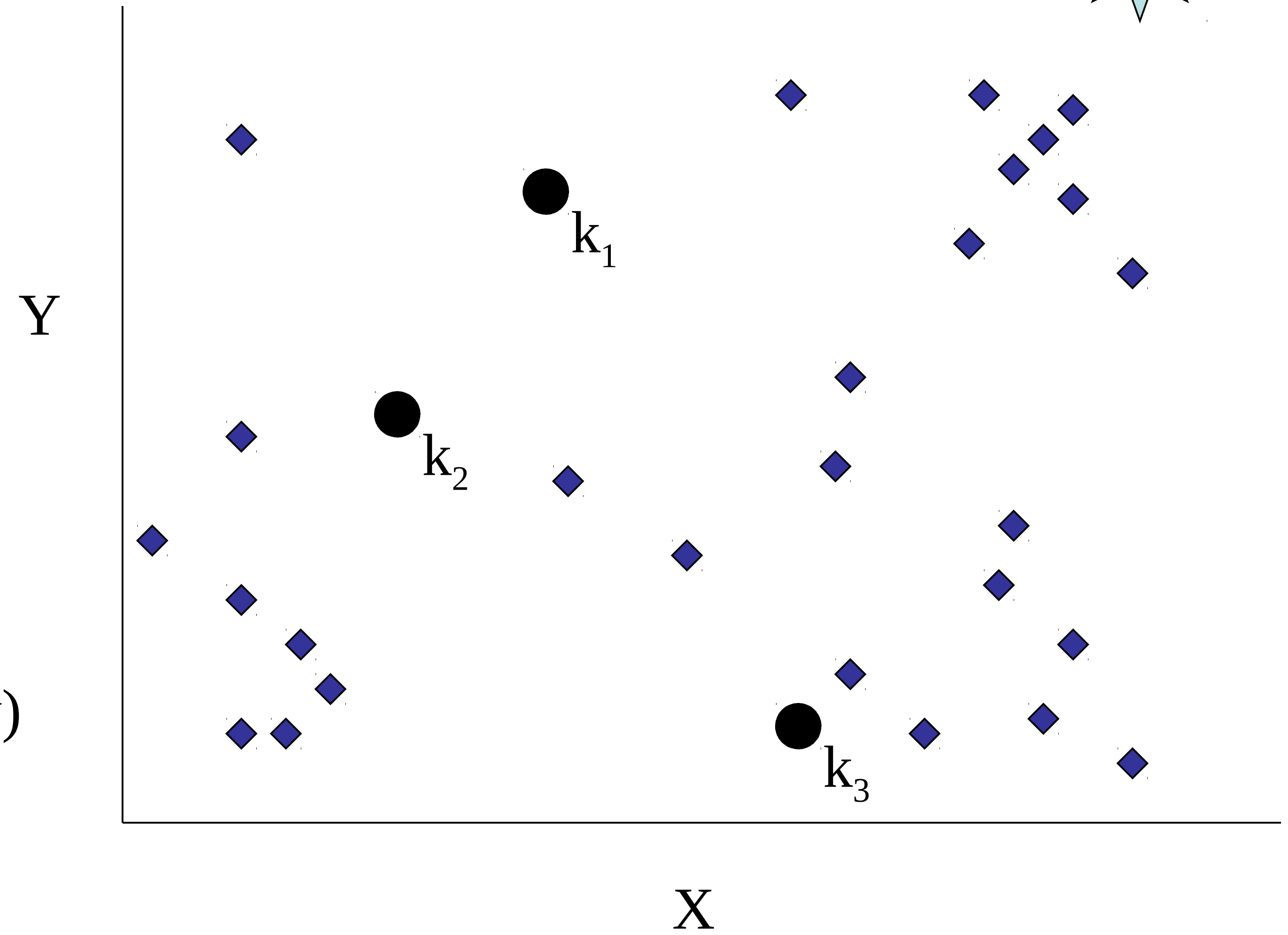
```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

K-means Clustering – Details

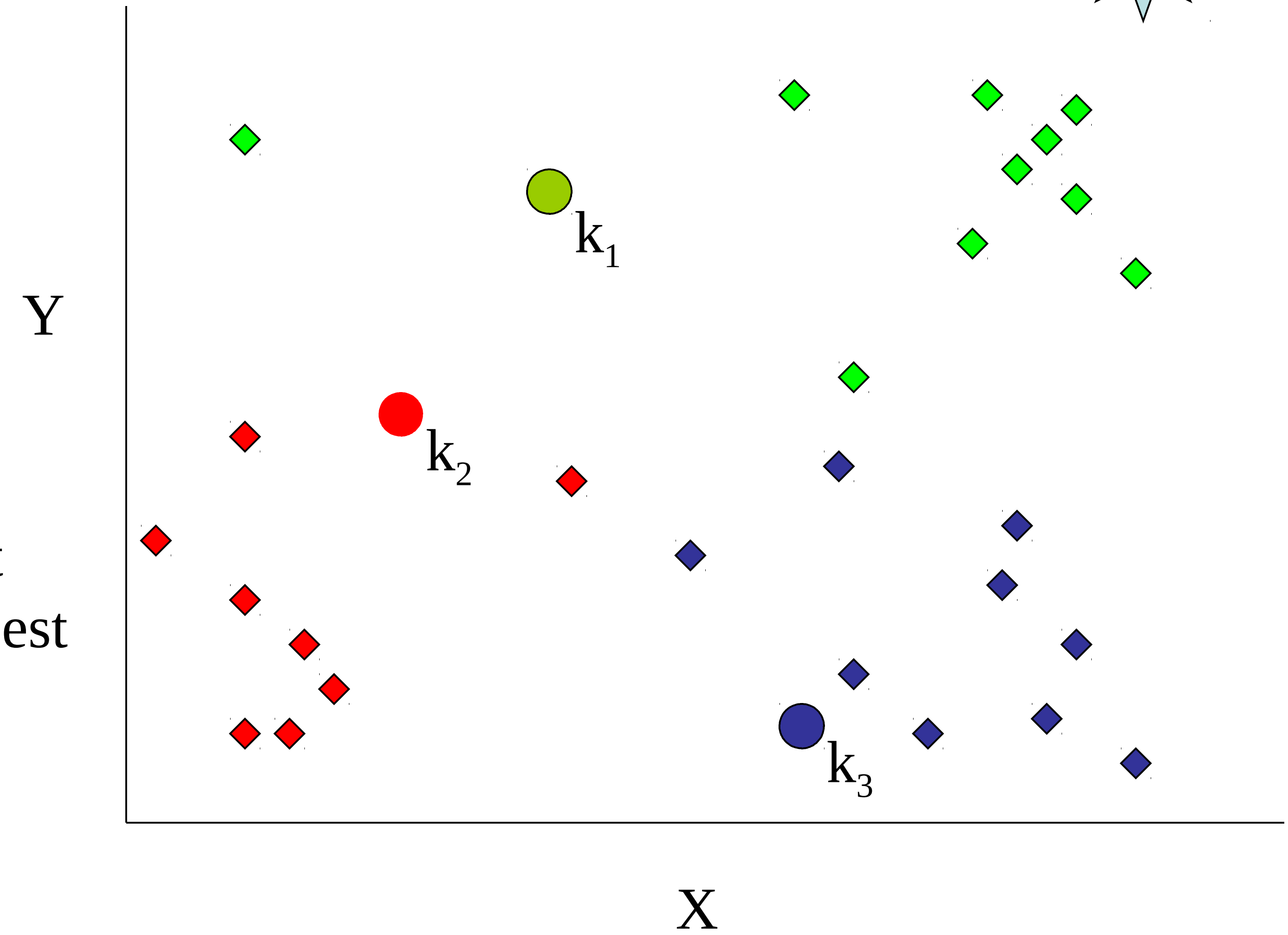
- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

K-means example, step 1

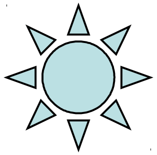
Pick 3 initial cluster centers (randomly)



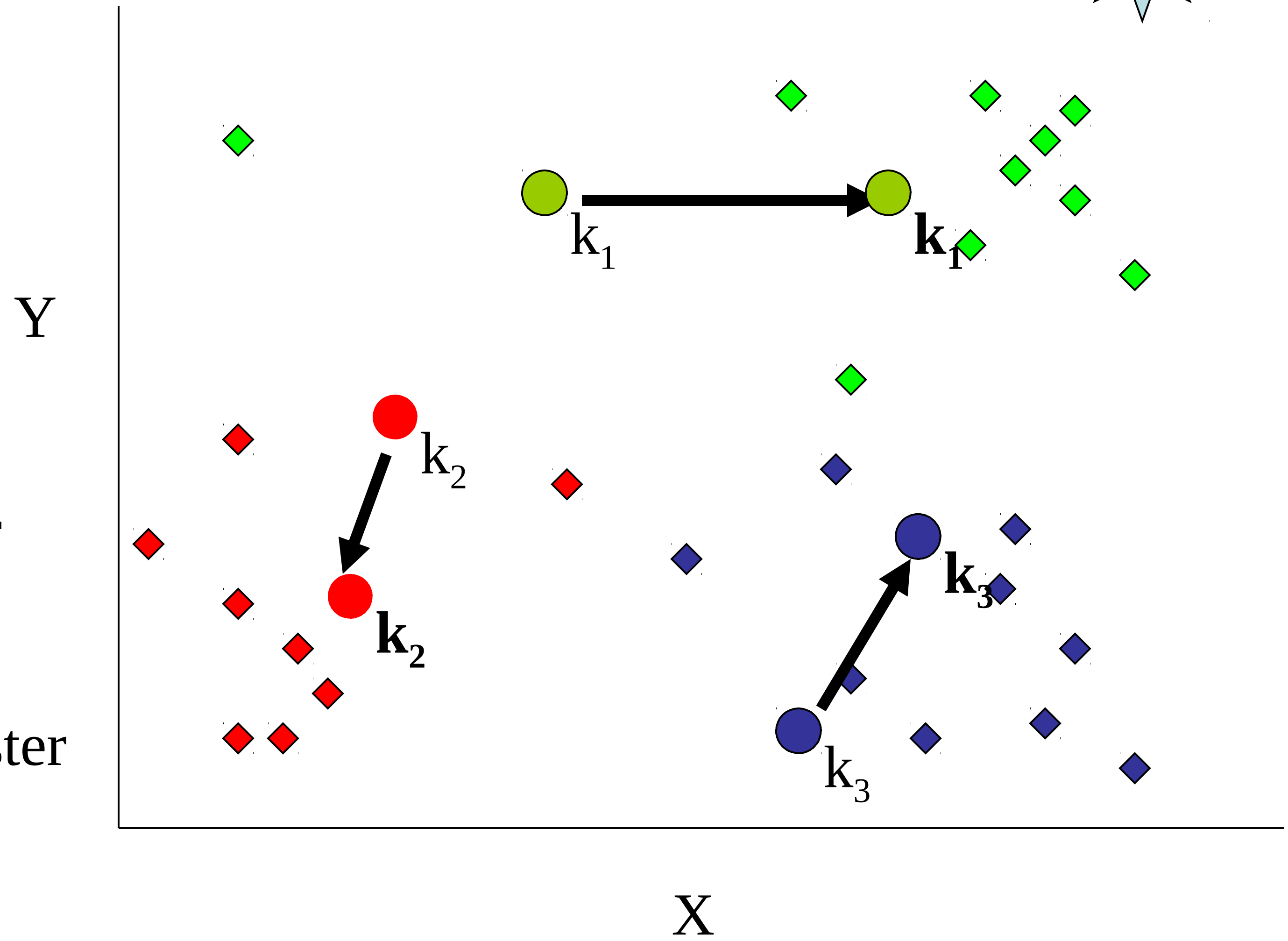
K-means example, step 2



Assign each point to the closest cluster center



K-means example, step 3

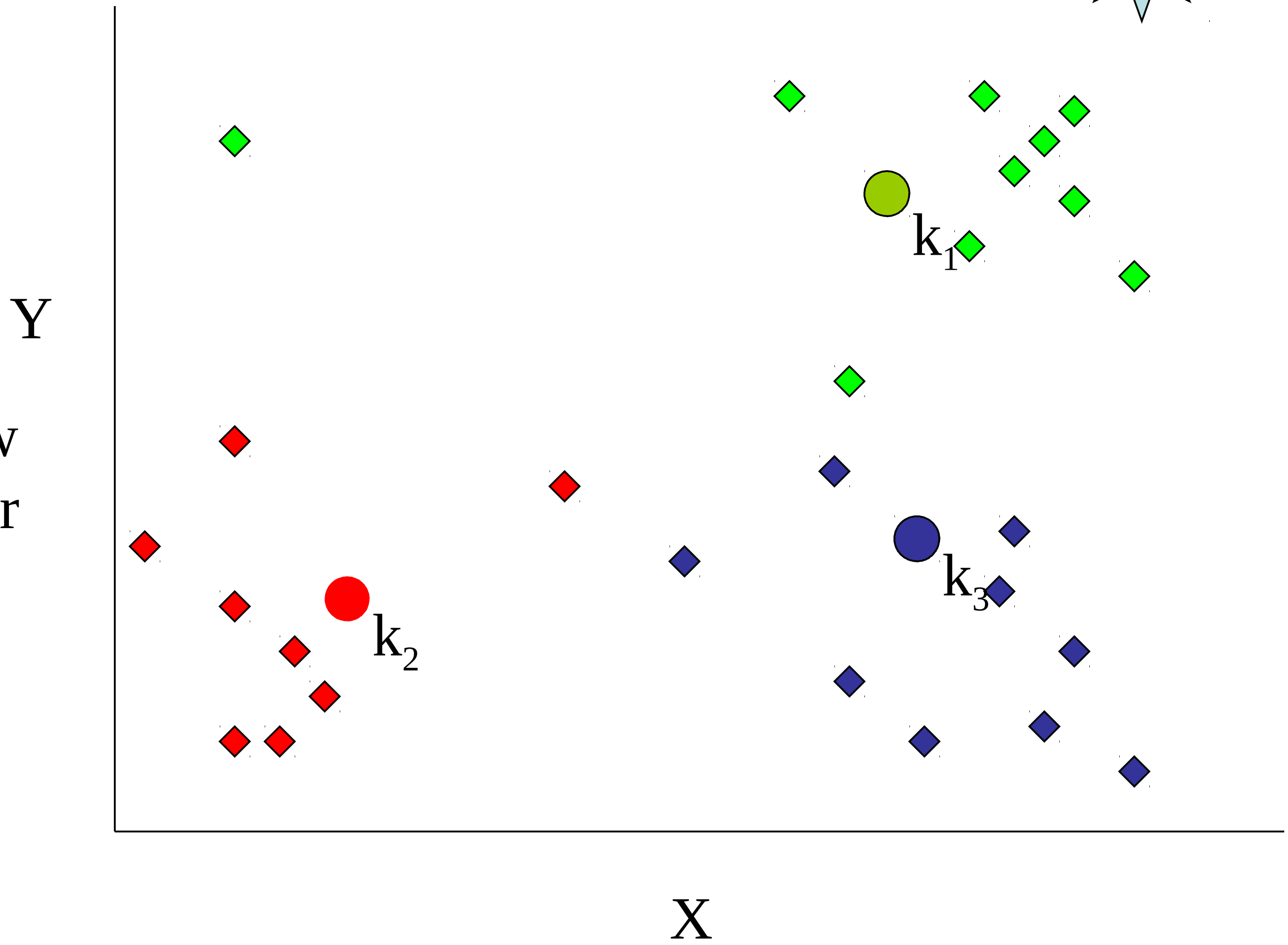


Move
each cluster
center
to the mean
of each cluster

K-means example, step 4

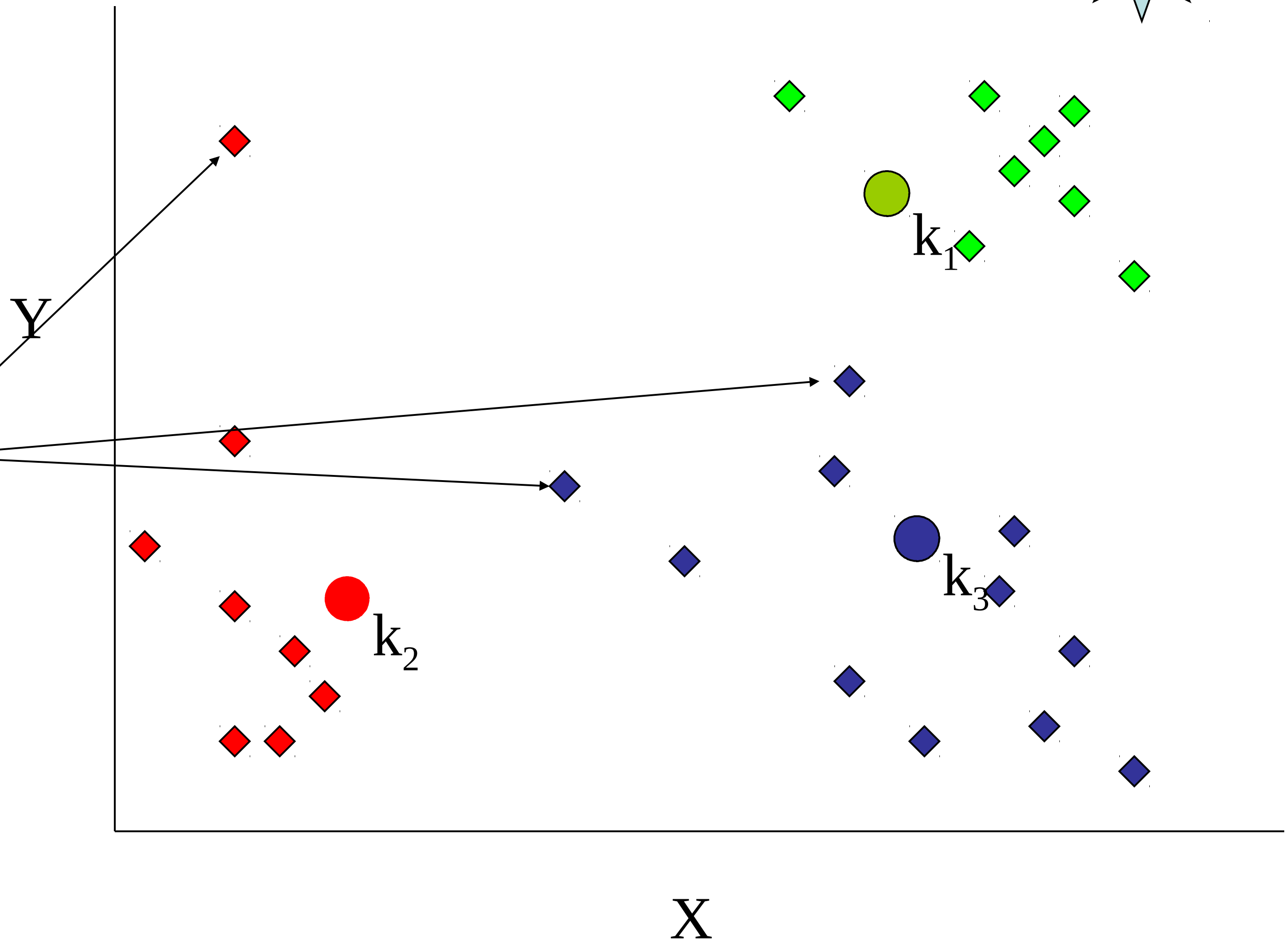
Reassign points closest to a different new cluster center

Q: Which points are reassigned?



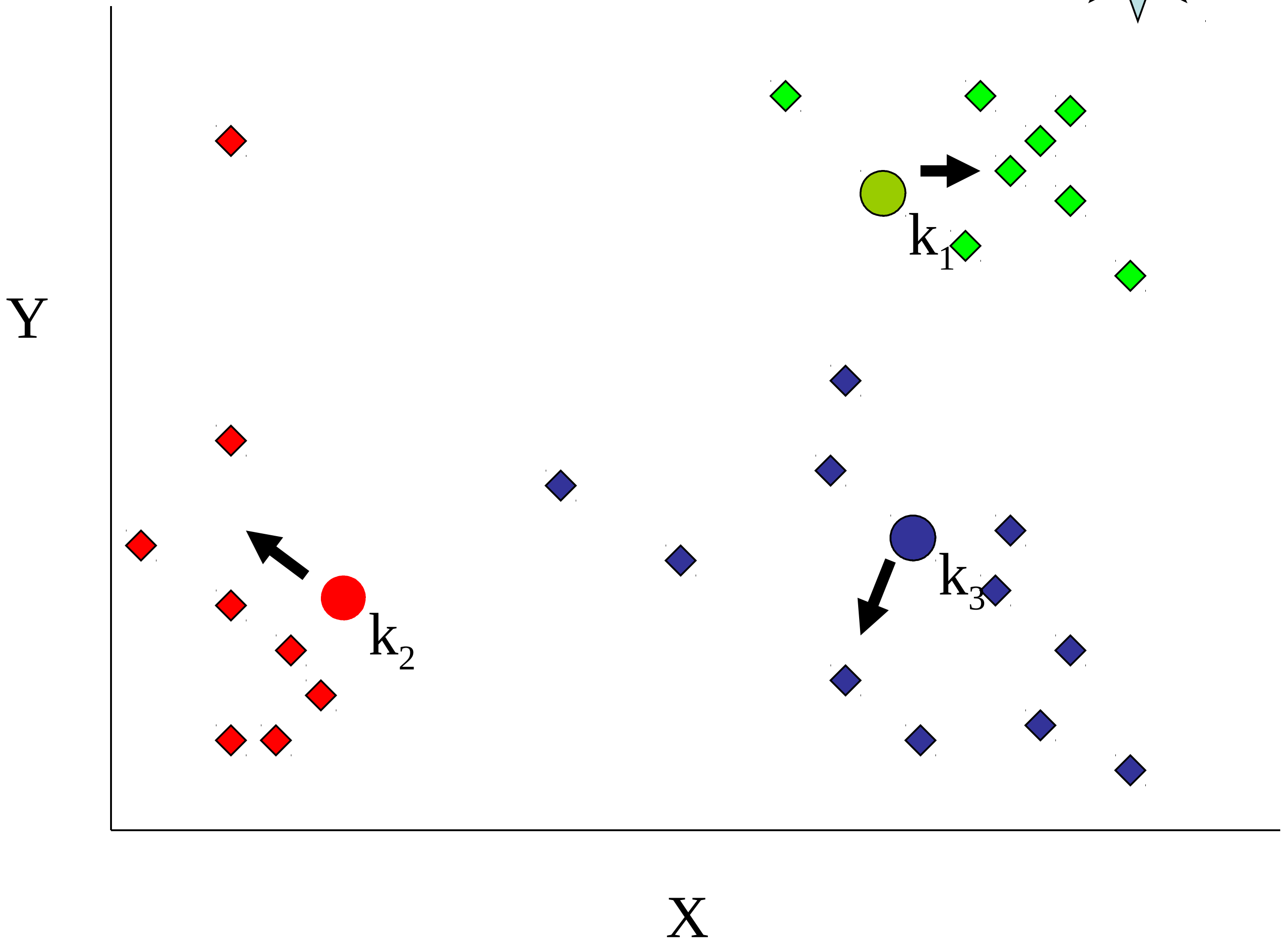
K-means example, step 4

A: three points with animation

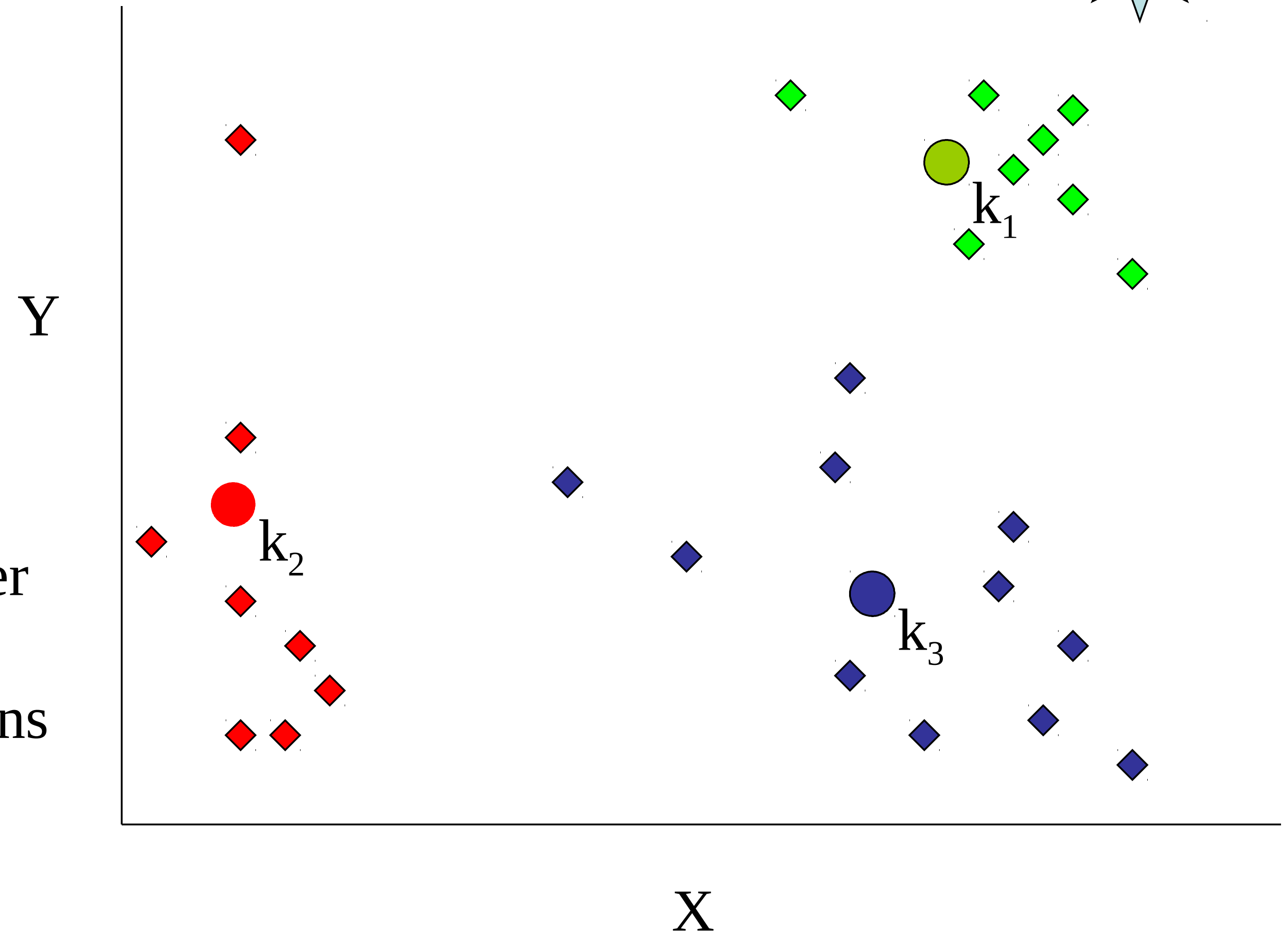


K-means example, step 4b

re-compute cluster means



K-means example, step 5



move cluster centers to cluster means

K-mean Example Run

- Suppose we have several objects (4 types of medicines) and each object have two attributes and our goal is to group these objects into K=2 group of medicine based on the two features (pH and weight index).

Object	attribute 1 (X): weight index	attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Initial value of centroids

- Suppose we use medicine A and medicine B as the first centroids.
- C1 and C2 denote the coordinate of the centroids, then $c_1 = (1,1)$ and $c_2 = (2,1)$
- Calculate the distance between cluster centroid to each object using Euclidian Distance

Continued…

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{array}$$

$$\begin{array}{cc} \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{array}{c} X \\ Y \end{array} \end{array}$$

The first row of the distance matrix corresponds to the distance of each object to the first centroid and the second row is the distance of each object to the second centroid. For example, distance from medicine C = (4, 3) to the first centroid $\mathbf{c}_1 = (1,1)$ Is $\sqrt{(4-1)^2+(3-1)^2}=3.61$ and its distance to the second centroid $\mathbf{c}_2 = (2,1)$ Is $\sqrt{(4-2)^2+(3-1)^2} = 2.83$

Objects clustering

- Assign each object based on the minimum distance
- Medicine A is assigned to group 1, medicine B to group 2, medicine C to group 2 and medicine D to group 2

$$\mathbf{G}^0 = \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \\ \begin{matrix} A & B & C & D \end{matrix} \end{array} \begin{array}{l} group-1 \\ group-2 \\ \\ \end{array}$$

Iteration-1, *determine centroids*

- compute the new centroid of each group based on these new memberships
- . Group 1 only has one member
- Group 2 now has three members, thus the centroid is the average coordinate among the three members

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right)$$

Iteration-1, *Objects-Centroids distances*

- Compute the distance of all objects to the new centroids

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Iteration-1, *Objects clustering*

- assign each object based on the minimum distance
- Based on the new distance matrix, we move the medicine B to Group 1 while all the other objects remain

$$\mathbf{G}^1 = \begin{array}{c} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\ \begin{matrix} A & \;B & \;C & \;D \end{matrix} \end{array} \begin{matrix} group-1 \\ group-2 \\ \end{matrix}$$

Iteration 2, *determine centroids*

- calculate the new centroids coordinate based on the clustering of previous iteration
- and group 2 both has two members, thus the new centroids are

$$\mathbf{c}_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(1\tfrac{1}{2}, 1\right) \qquad \mathbf{c}_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = \left(4\tfrac{1}{2}, 3\tfrac{1}{2}\right)$$

Iteration-2, *Objects-Centroids distances*

- New distance matrix

$$\mathbf{D}^2 = \begin{array}{c} \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \end{array} \begin{array}{l} \mathbf{c}_1 = (1\tfrac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2}) \quad group-2 \\ \\ X \\ Y \end{array}$$

Iteration-2, *Objects clustering*

- Assign each object based on the minimum distance

$$\mathbf{G}^2 = \begin{array}{c} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{array}{l} group-1 \\ group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array}$$

- Grouping of last iteration and this iteration reveals that the objects does not move group anymore
- Computation of the k-mean clustering has reached its stability and no more iteration is needed

Final grouping

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Squared Error Criterion

- The $k-$Means algorithm (with a Euclidean distance measure) is equivalent to a minimisation of the squared error criterion:

$$E = \sum_{i=1}^{k} \sum_{p \,\epsilon\, C_i} |p - m_i|^2$$

  where E is the sum of square error for all objects in the database, p is the point in space corresponding to a given object and m_i is the centre of cluster C_i.
- $|p - m_i|$ denotes the distance of a data object from the cluster centre it belongs to.
- This therefore makes the clusters as compact and as separate as possible.

Evaluating K-means Clusters

- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Bisecting K-means
  - Not as susceptible to initialization issues

Handling Empty Clusters

- Basic K-means algorithm can yield empty clusters

- Several strategies
  - Choose the point that contributes most to SSE
  - Choose a point from the cluster with the highest SSE
  - If there are several empty clusters, the above can be repeated several times.

Updating Centers Incrementally

- In the basic K-means algorithm, centroids are updated after all points are assigned to a centroid

- An alternative is to update the centroids after each assignment (incremental approach)
  - Each assignment updates zero or two centroids
  - More expensive
  - Introduces an order dependency
  - Never get an empty cluster
  - Can use “weights” to change the impact

Pre-processing and Post-processing

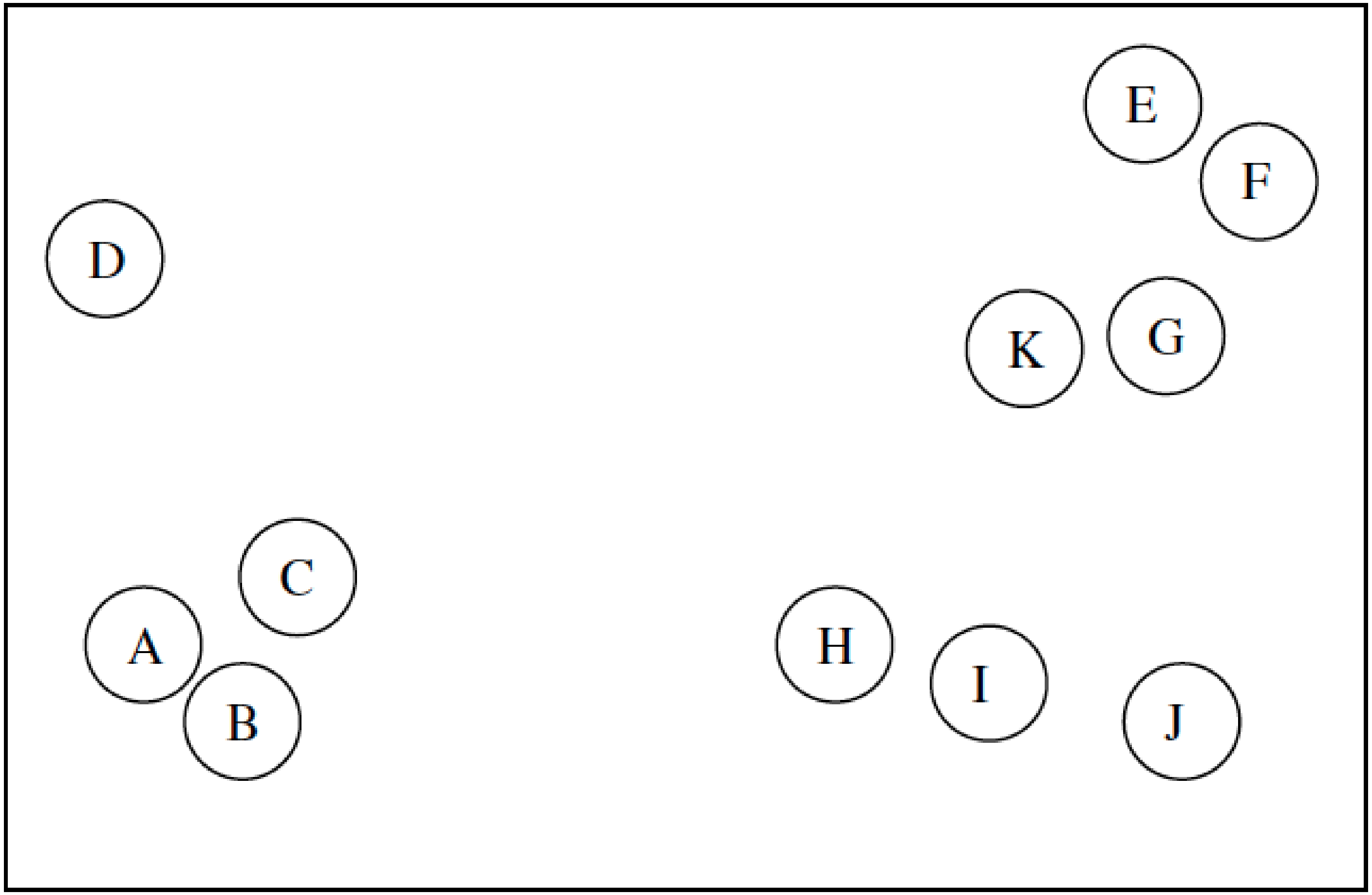
- Pre-processing
  - Normalize the data
  - Eliminate outliers

- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split ‘loose’ clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are ‘close’ and that have relatively low SSE
  - Can use these steps during the clustering process
    - ISODATA

Pros and cons of K-Means

- $k-$Means works well when the data form compact clouds, well-separated from one another.
- Method is scaleable and quite efficient for large datasets, with complexity $O(nkt)$, where k is the number of clusters, n the number of objects and t the number of iterations required. Usually $t << n$ and $k << n$.
- Can only be applied where the mean of a cluster is defined – if there are categorical (nominal or binary) data, this may not be possible.
- Users may not be able to select a good value for k.
- Algorithm is sensitive to noise and outlier points, which can influence the mean attribute values significantly.

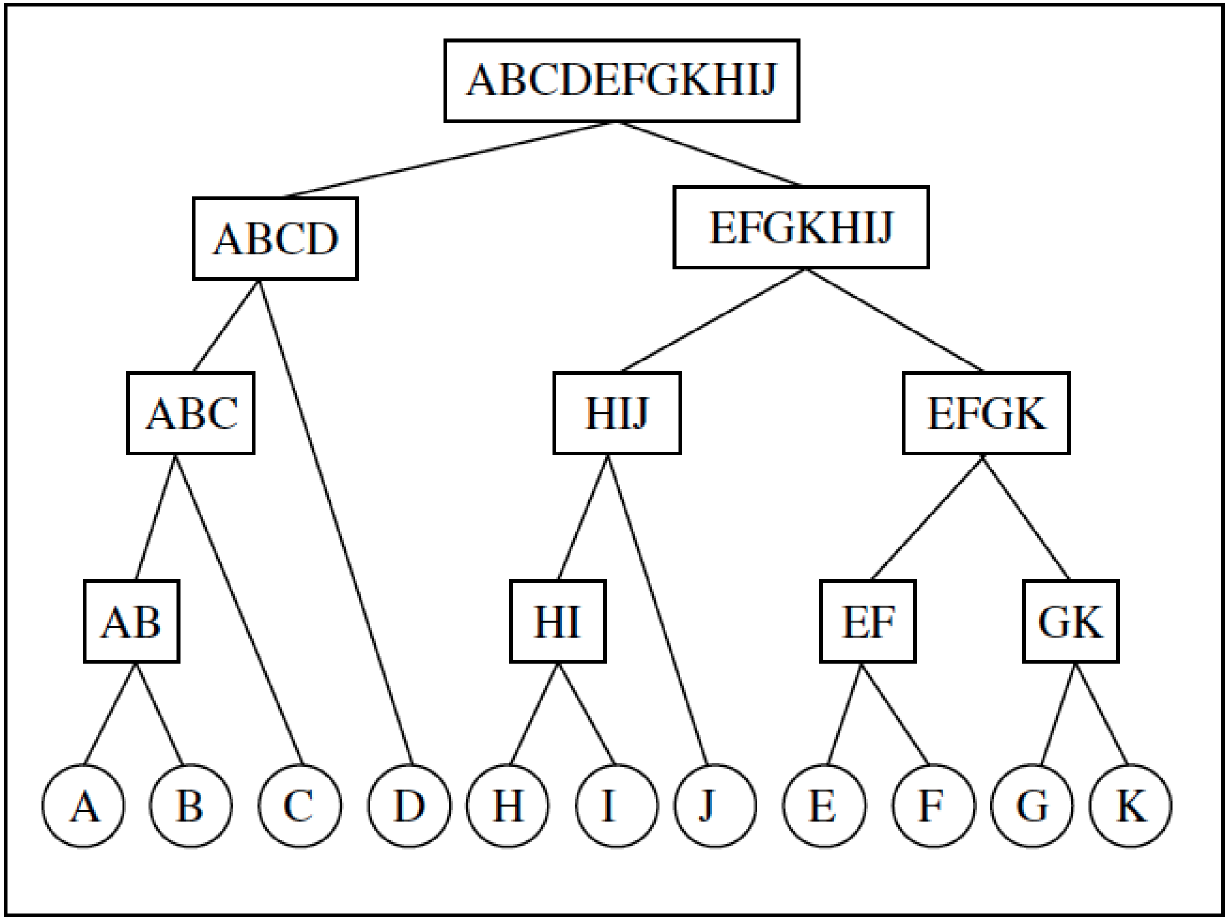
Hierarchical clustering



Hierarchical clustering

1. A and B $\rightarrow$ AB
2. AB and C $\rightarrow$ ABC
3. G and K $\rightarrow$ GK
4. E and F $\rightarrow$ EF
5. H and I $\rightarrow$ HI
6. EF and GK $\rightarrow$ EFGK
7. HI and J $\rightarrow$ HIJ
8. ABC and D $\rightarrow$ ABCD
9. EFGK and HIJ $\rightarrow$ EFGKHIJ
10. ABCD and EFGKHIJ $\rightarrow$ ABCDEFGKHIJ

Hierarchical clustering



Example run

- We have 6 objects and each have two measured values

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Distance calculation

- The proximity between object can be measured as distance matrix. Suppose we use Euclidean distance , we can compute the distance between objects using the following formula

$$d_{ij} = \left(\sum_k \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$$

- For example, distance between object A = (1, 1) and B = (1.5, 1.5) is computed as

$$d_{AB} = \left((1-1.5)^2 + (1-1.5)^2 \right)^{\frac{1}{2}} = \sqrt{\tfrac{1}{2}} = 0.7071$$

Distance matrix

- Compute all distances between objects and put the distances into a matrix form
- Distance between A and B is equal to distance between B and A
- Diagonal elements of distance matrix are zero represent distance from an object to itself

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

0.71	5.66	3.61	4.24	3.20	4.95	2.92	3.54	2.50	2.24	1.41	2.50	1.00	0.50	1.12

Joining objects to expand cluster

- We have 6 objects and we put each object into one cluster
- Goal is to group those 6 clusters such that at the end of the iterations, we will have only single cluster consists of the whole six original objects
- In each step of the iteration, we find the closest pair clusters
- The closest cluster is between cluster F and D with shortest distance of 0.5. Thus, we group cluster D and F into cluster (D, F)

Update distance matrix

- Distance between ungrouped clusters will not change from the original distance matrix

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Linking rule

- Using single linkage, specify minimum distance between original objects of the two clusters
- Using the input distance matrix, distance between cluster (D, F) and cluster A is computed as

$$d_{(D,F)\mapsto A} = \min\left(d_{DA}, d_{FA}\right) = \min\left(3.61, 3.20\right) = 3.20$$

- Distance between cluster (D, F) and cluster B is

$$d_{(D,F)\mapsto B} = \min\left(d_{DB}, d_{FB}\right) = \min\left(2.92, 2.50\right) = 2.50$$

- For C

$$d_{(D,F)\mapsto C} = \min\left(d_{DC}, d_{FC}\right) = \min\left(2.24, 2.50\right) = 2.24$$

- For E

$$d_{E\to(D,F)} = \min\left(d_{ED}, d_{EF}\right) = \min\left(1.00, 1.12\right) = 1.00$$

Updated matrix

- Looking at the lower triangular updated distance matrix, we found out that the closest distance between cluster B and cluster A is now 0.71. Thus, we group cluster A and cluster B into a single cluster name (A, B)

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00

Next update

Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0

- Using the input distance matrix (size 6 by 6), distance between cluster C and cluster (D, F) is computed as

$$d_{C\to(A,B)} = \min\left(d_{CA}, d_{CB}\right) = \min\left(5.66, 4.95\right) = 4.95$$

- Distance between cluster (D, F) and cluster (A, B) is the minimum distance between all objects involves in the two clusters $d_{(D,F)\mapsto(A,B)} = \min\left(d_{DA}, d_{DB}, d_{FA}, d_{FB}\right) = \min\left(3.61, 2.92, 3.20, 2.50\right) = 2.50$
- Similarly, distance between cluster E and (A, B) is

$$d_{E\to(A,B)} = \min\left(d_{EA}, d_{EB}\right) = \min\left(4.24, 3.54\right) = 3.54$$

Updated matrix

Min Distance (Single Linkage)

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	**4.95**	0	2.24	1.41
(D, F)	**2.50**	2.24	0	1.00
E	**3.54**	1.41	1.00	0

- Observing the lower triangular of the updated distance matrix, we can see that the closest distance between clusters happens between cluster E and (D, F) at distance 1.00. Thus, we cluster them together into cluster ((D, F), E)

Next matrix

Min Distance (Single Linkage)

Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00

- Distance between cluster ((D, F), E) and cluster (A, B) is calculated as

$$d_{((D,F),E)\mapsto(A,B)} = \min\left(d_{DA}, d_{DB}, d_{FA}, d_{FB}, d_{EA}, d_{EB}\right) = \min\left(3.61, 2.92, 3.20, 2.50, 4.24, 3.54\right) = 2.50$$

- Distance between cluster ((D, F), E) and cluster C yields the minimum distance of 1.41. This distance is computed as $d_{((D,F),E)\mapsto C} = \min\left(d_{DC}, d_{FC}, d_{EC}\right) = \min\left(2.24, 2.50, 1.41\right) = 1.41$
- Merge cluster ((D, F), E) and cluster C into a new cluster name (((D, F), E), C)

Updated matrix

Min Distance (Single Linkage)

Dist	(A,B)	(D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

- The minimum distance of 2.5 is the result of the following computation

$$d_{(((D,F),E),C)\rightarrow(A,B)} = \min\left(d_{DA}, d_{DB}, d_{FA}, d_{FB}, d_{EA}, d_{EB}, d_{CA}, d_{CB}\right)$$

$$d_{(((D,F),E),C)\rightarrow(A,B)} = \min\left(3.61, 2.92, 3.20, 2.50, 4.24, 3.54, 5.66, 4.95\right) = 2.50$$

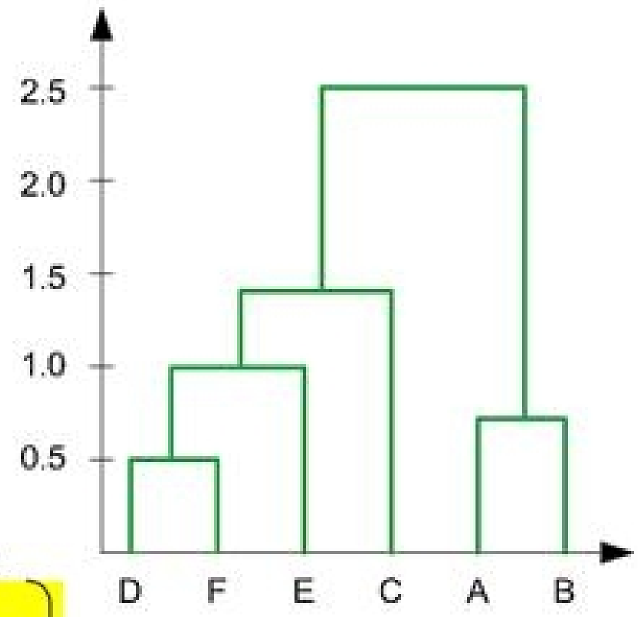
- Merge the remaining two clusters, we will get only single cluster contain the whole 6 objects. Thus, our computation is finished.

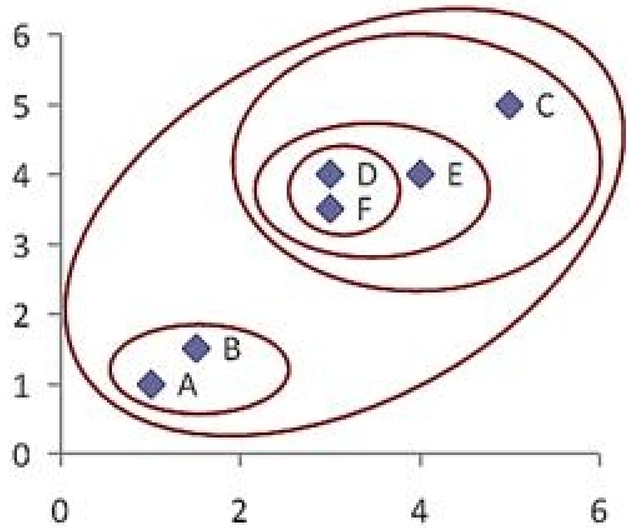
Summarized result

- In the beginning we have 6 clusters: A, B, C, D, E and F
- We merge cluster D and F into cluster (D, F) at distance **0.50**
- We merge cluster A and cluster B into (A, B) at distance **0.71**
- We merge cluster E and (D, F) into ((D, F), E) at distance **1.00**
- We merge cluster ((D, F), E) and C into (((D, F), E), C) at distance **1.41**
- We merge cluster (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance **2.50**
- The last cluster contain all the objects, thus conclude the computation
- Using this information, we can now draw the final results of a dendogram

Dendogram and hierarchy

- The hierarchy is given as (((D, F), E),C), (A,B). We can also plot the clustering hierarchy into XY space





	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Hierarchical clustering

	a	b	c	d	e	f
a	0	12	6	3	25	4
b	12	0	19	8	14	15
c	6	19	0	12	5	18
d	3	8	12	0	11	9
e	25	14	5	11	0	7
f	4	15	18	9	7	0

Hierarchical clustering

	ad	b	c	e	f
ad	0	?	?	?	?
b	?	0	19	14	15
c	?	19	0	5	18
e	?	14	5	0	7
f	?	15	18	7	0

Hierarchical clustering

	ad	b	c	e	f
ad	0	8	6	11	4
b	8	0	19	14	15
c	6	19	0	5	18
e	11	14	5	0	7
f	4	15	18	7	0

Hierarchical clustering

	adf	b	c	e
adf	0	8	6	7
b	8	0	19	14
c	6	19	0	5
e	7	14	5	0

Hierarchical clustering

	adf	b	ce
adf	0	8	6
b	8	0	14
ce	6	14	0

Hierarchical clustering

	$adfce$	b
$adfce$	0	8
b	8	0

Research Problems

– **Effective and Efficient methods of Clustering**
– **Scalability**
– **Handling different types of data**
– **Handling complex multidimensional data**
– **Complex shapes of clusters**
– **Subspace Clustering**
– **Cluster overlapping etc.**

Examples of Clustering Applications

- **Marketing:** discover customer groups and use them for targeted marketing and re-organization
- **Astronomy:** find groups of similar stars and galaxies
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults
- **Genomics:** finding groups of gene with similar expressions
- …

Clustering Summary

- unsupervised
- many approaches
  - K-means – simple, sometimes useful
    - K-medoids is less sensitive to outliers
  - Hierarchical clustering – works for symbolic attributes
  - Can be used to fill in missing values

Questions

?

Quiz-2

One Error at most 2.5 points Two Errors 0 to 1 points

Assume the following dataset is given: (2,2), (4,4), (5,5), (6,6), (7,7), (9,9), (0,6), (6,0). K-Means is used with k=3 to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compute distances between centroids and objects in the dataset. Moreover, K-Means' initial clusters C1, C2, and C3 are as follows:

C1: {(2,2), (4,4), (6,6)}

C2: {(0,6), (6,0)}

C3: {(5,5), (7, 7), (9,9)}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids? [5]

d((x1,x2),(x1',x2'))= |x1-x1'| + |x2-x2'|