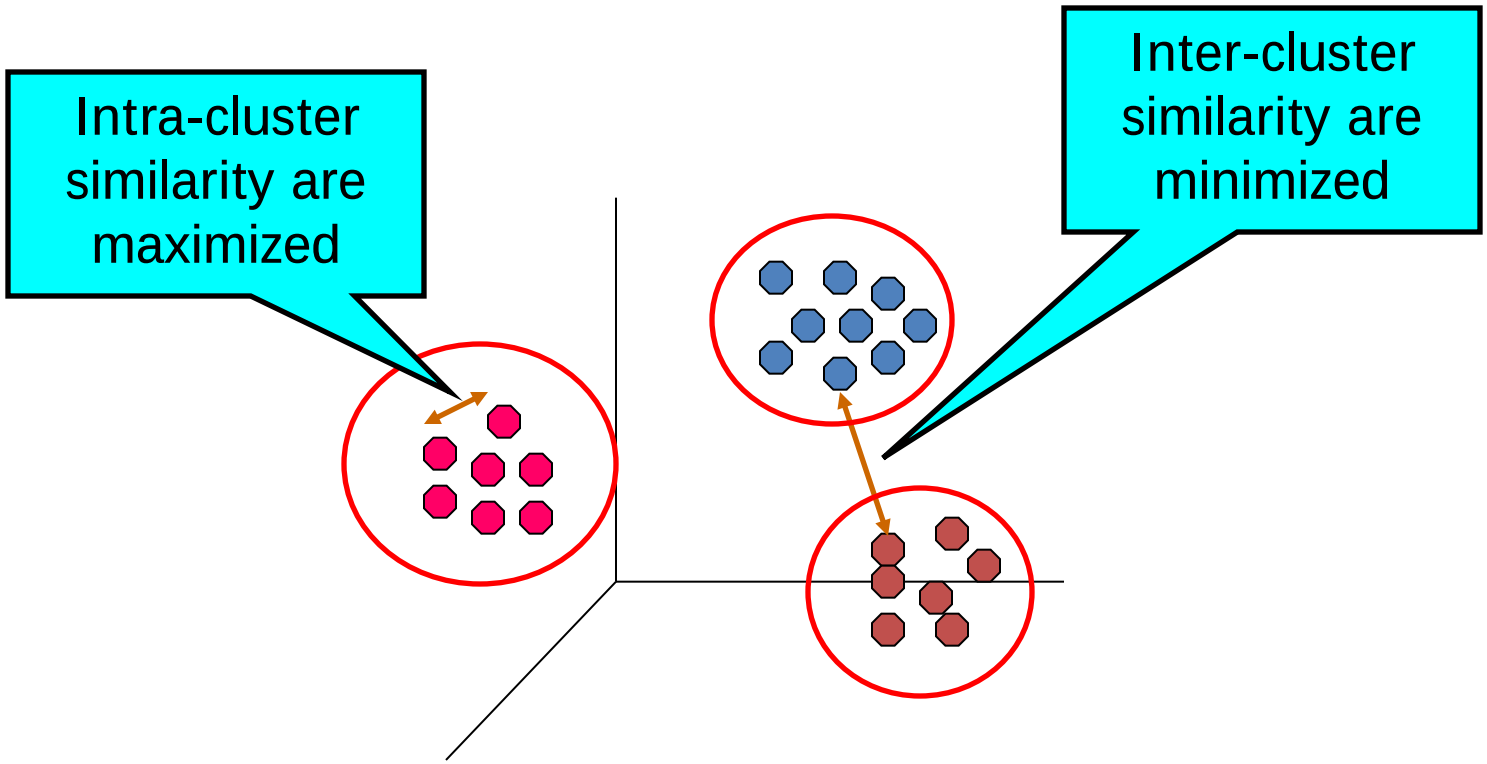
Clustering

UE 141 Spring 2013

Jing Gao
SUNY Buffalo

Definition of Clustering

- 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

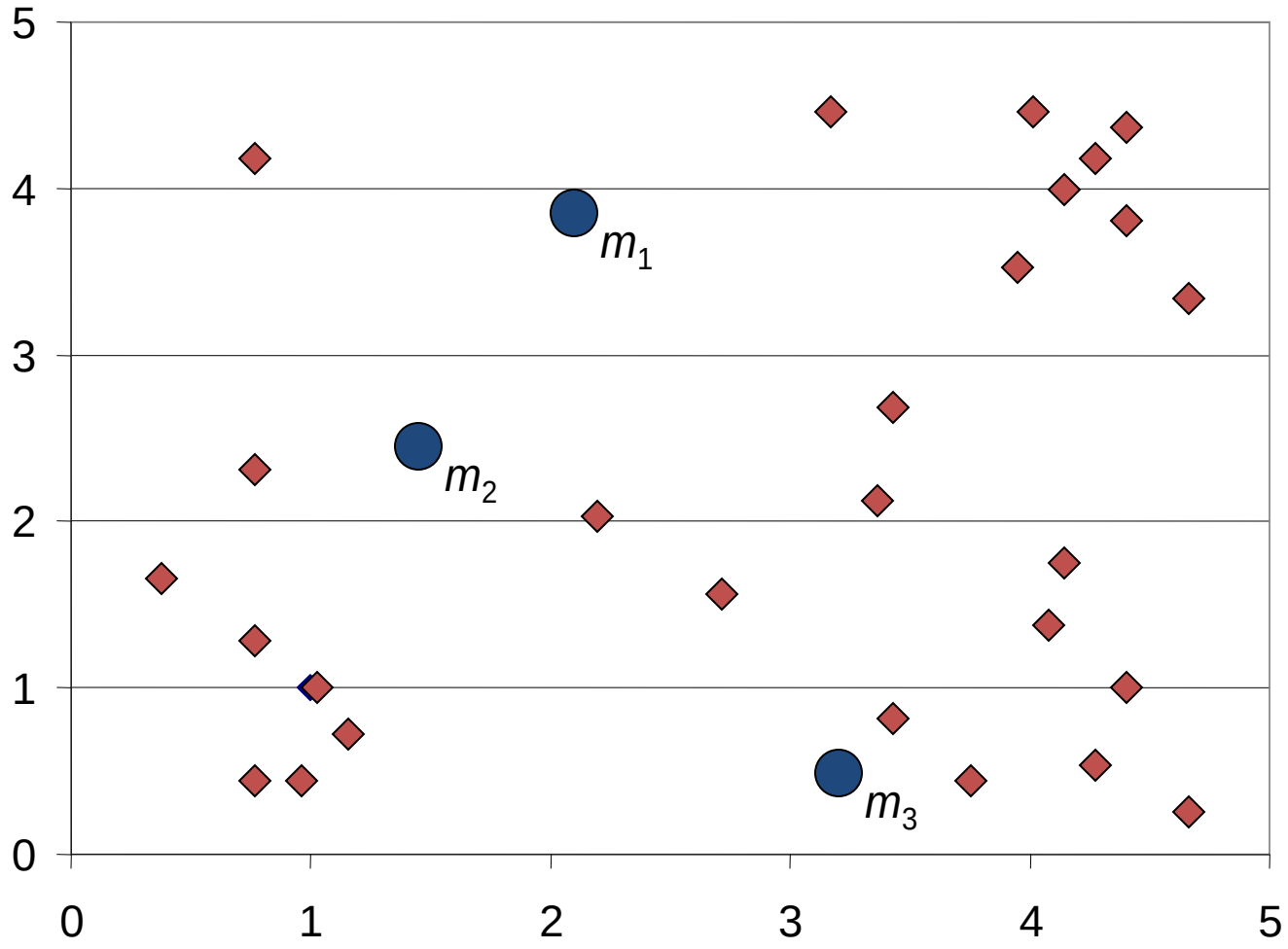


K-means

- **Partition $\{x_1,\ldots,x_n\}$ into K clusters**
  - K is predefined
- **Initialization**
  - Specify the initial cluster centers (centroids)
- **Iteration until no change**
  - For each object x_i
    - Calculate the distances between x_i and the K centroids
    - (Re)assign x_i to the cluster whose centroid is the closest to x_i
  - Update the cluster centroids based on current assignment

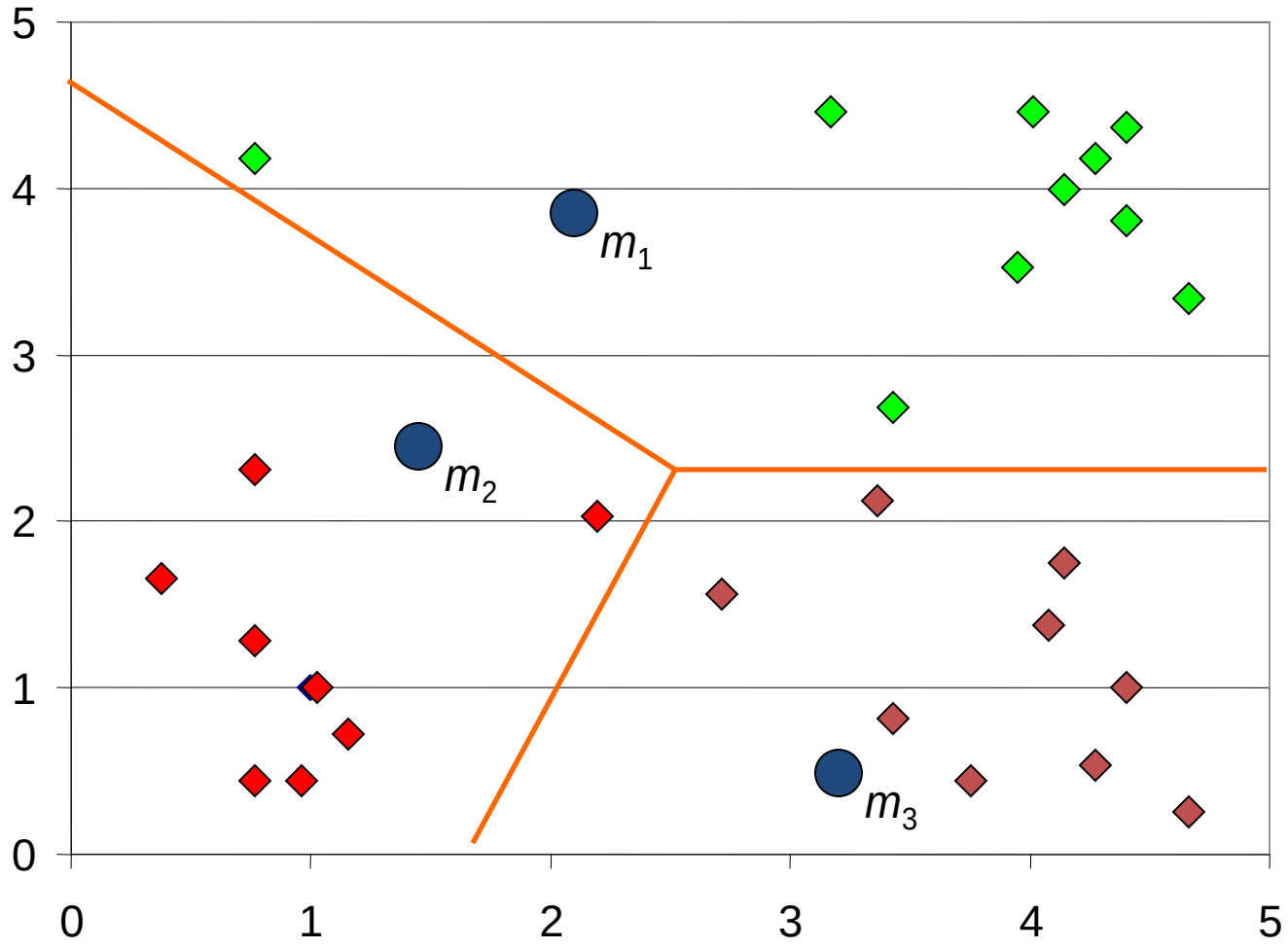
K-means: Initialization

Initialization: Determine the three cluster centers



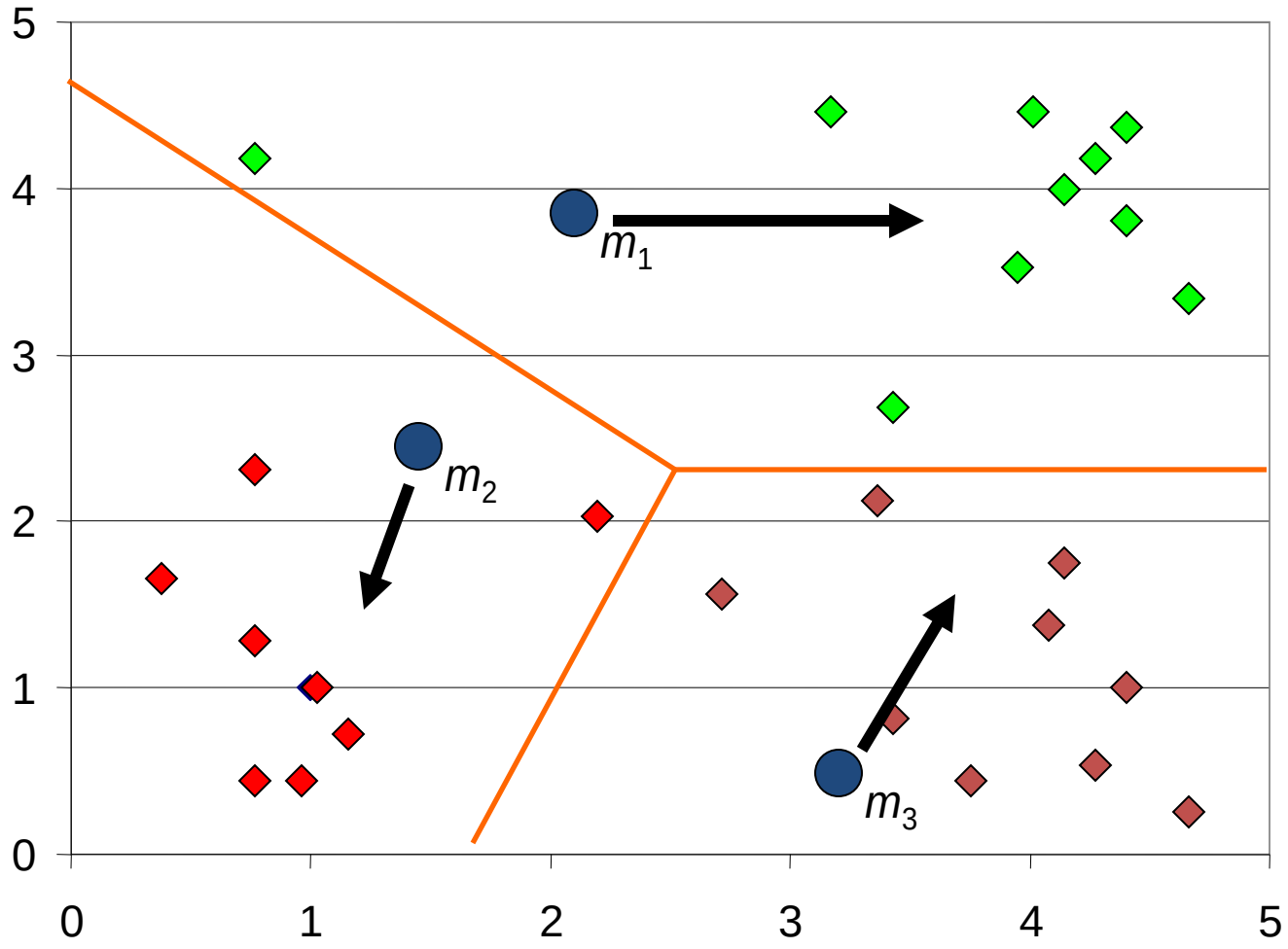
K-means Clustering: Cluster Assignment

Assign each object to the cluster which has the closet distance from the centroid to the object



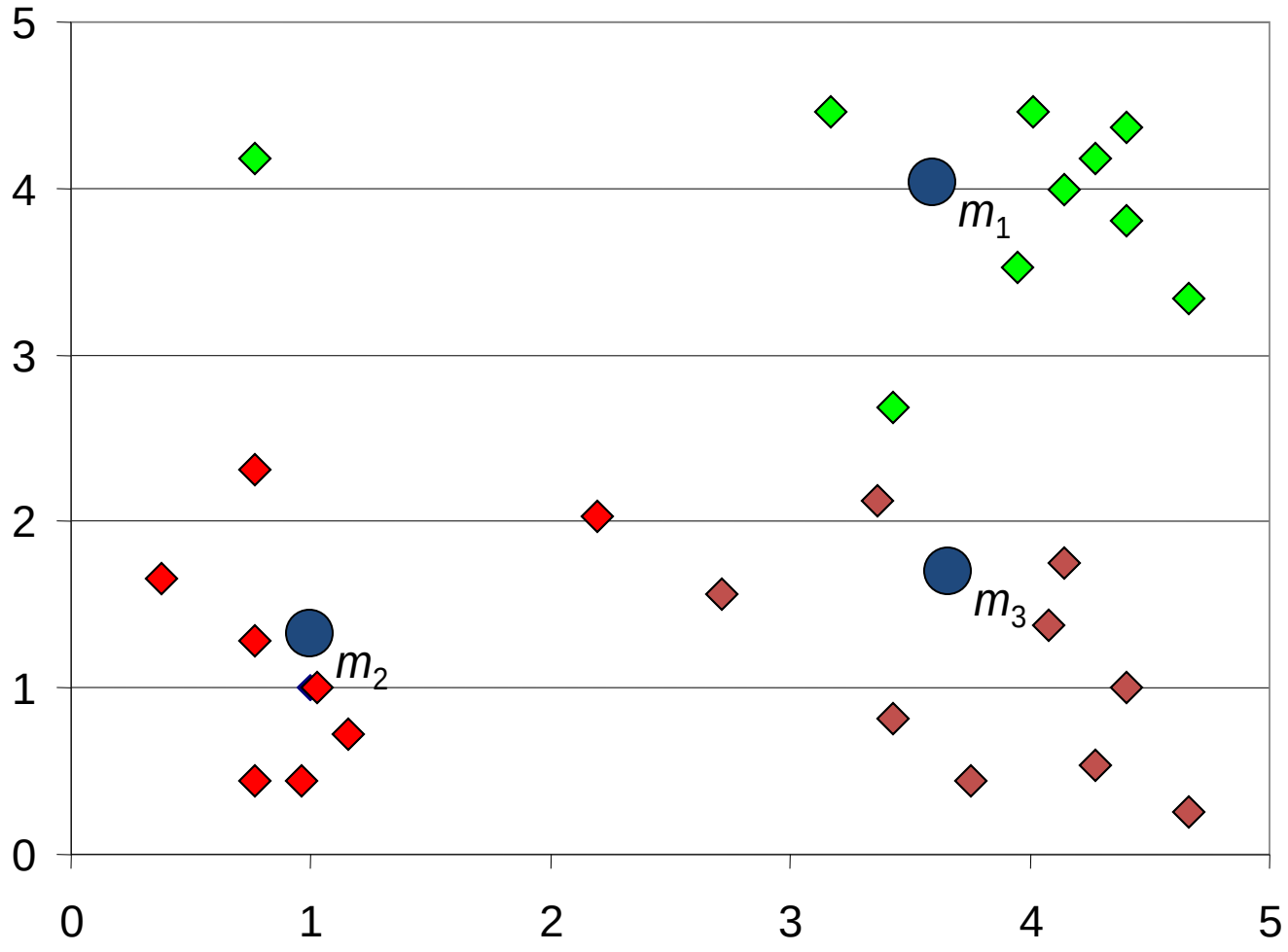
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



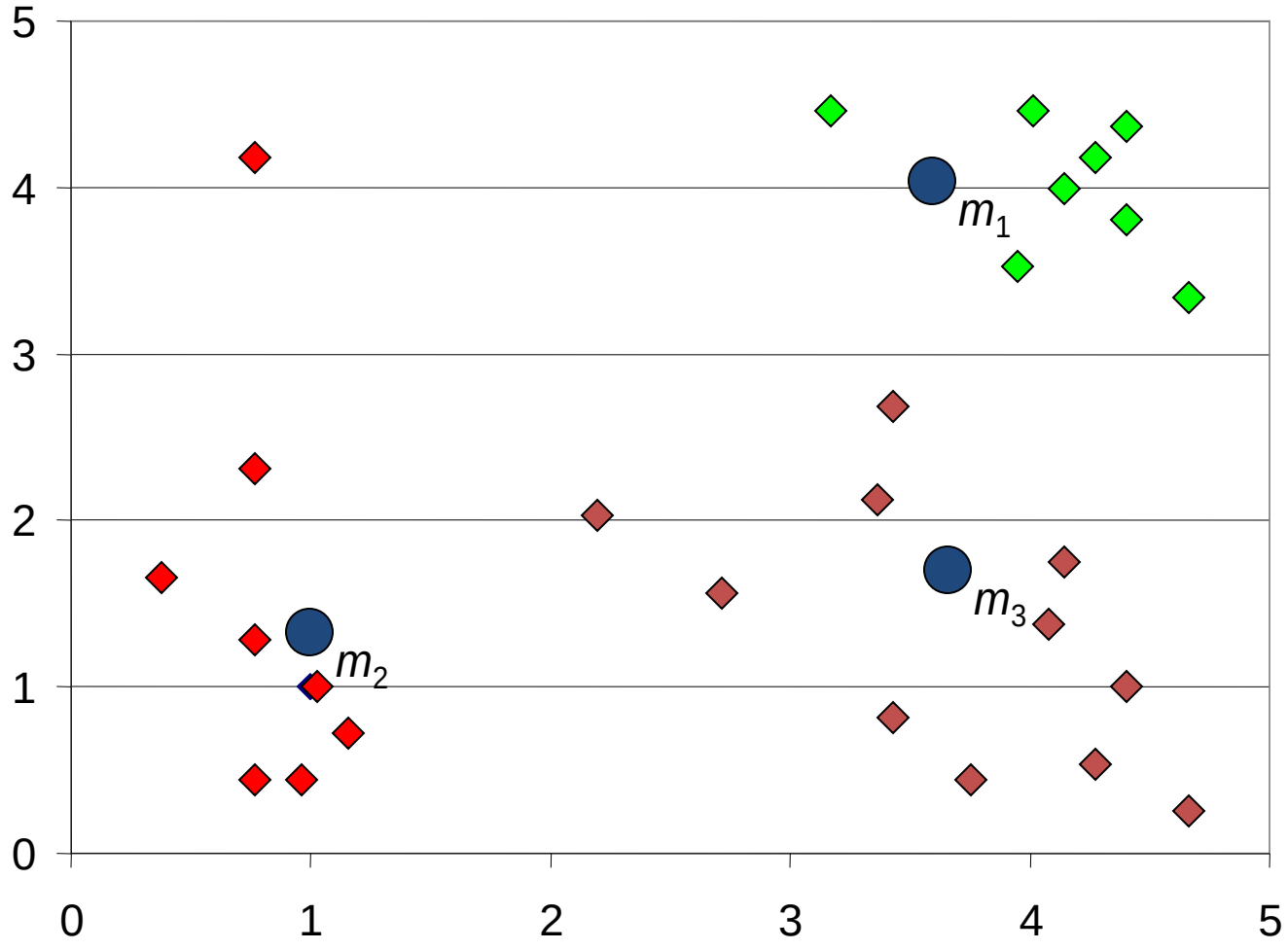
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



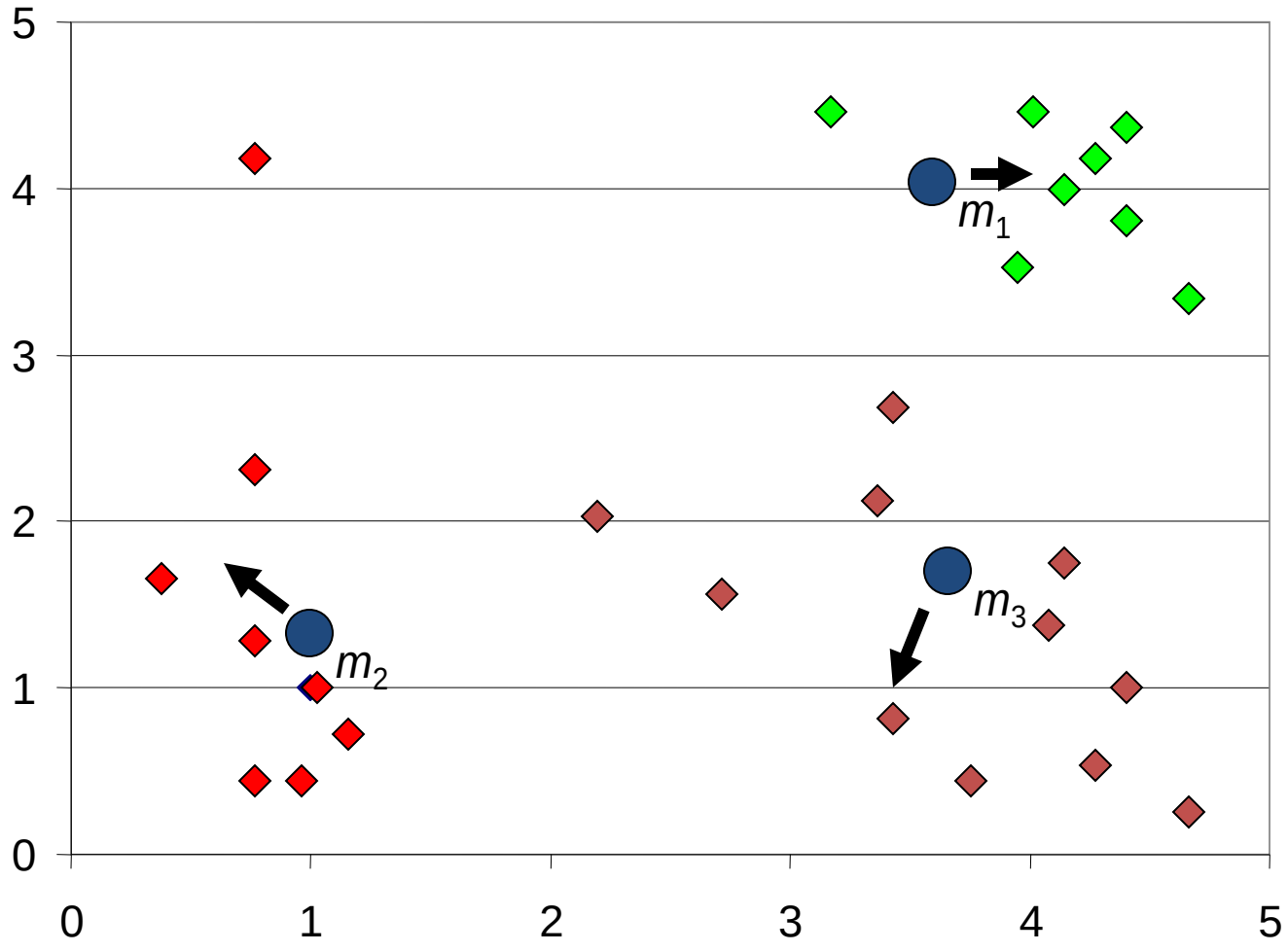
K-means Clustering: Cluster Assignment

Assign each object to the cluster which has the closet distance from the centroid to the object



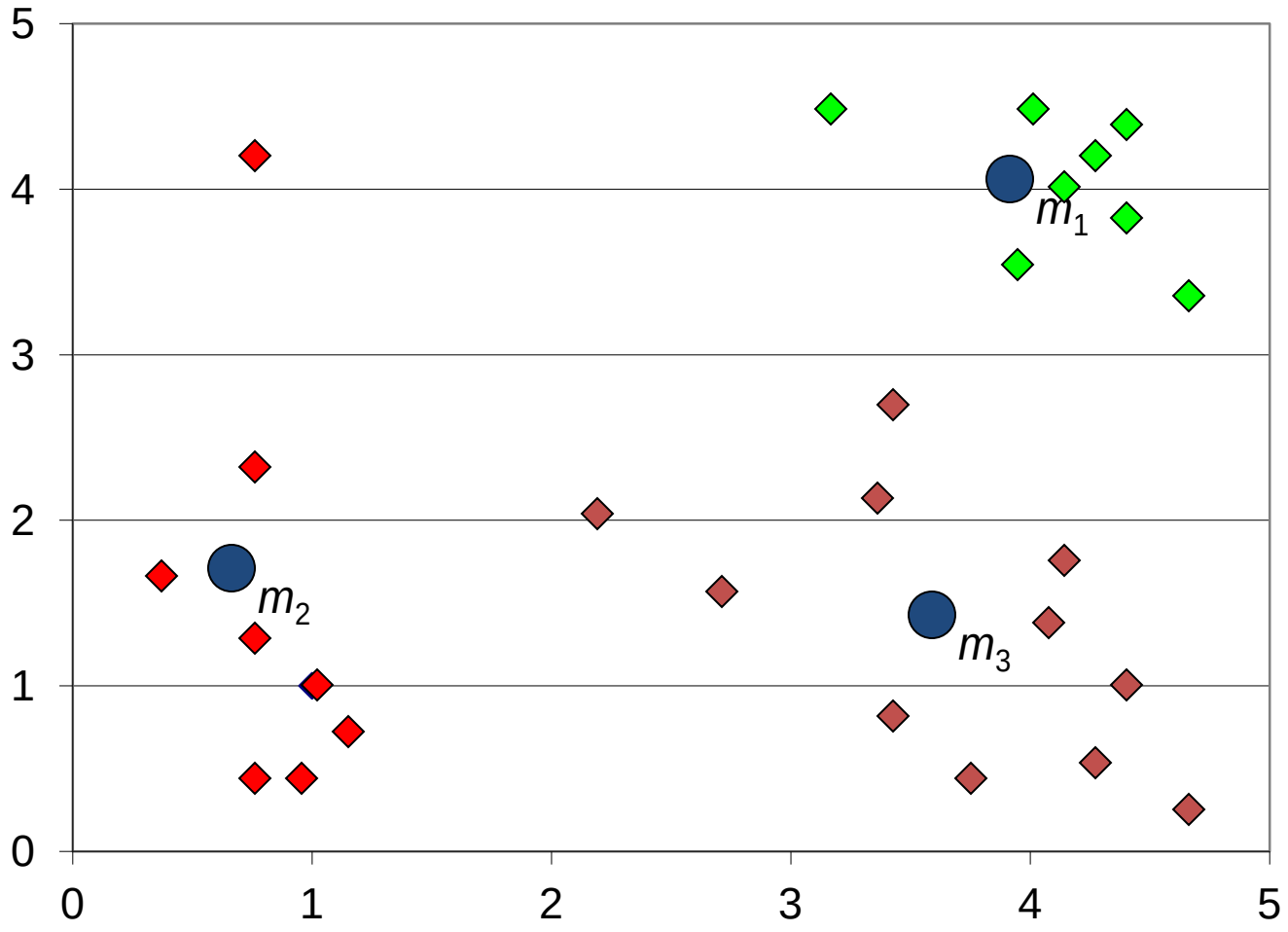
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



K-means Clustering: Update Cluster Centroid

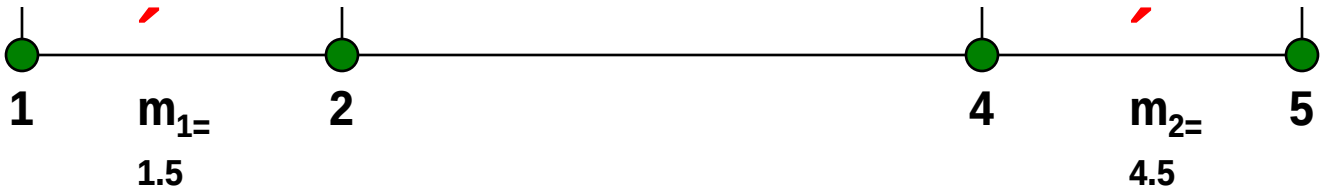
Compute cluster centroid as the center of the points in the cluster



Sum of Squared Error (SSE)

- Suppose the centroid of cluster C_j is m_j
- For each object x in C_j, compute the squared error between x and the centroid m_j
- Sum up the error of all the objects

$$SSE = \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$



$$SSE = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

How to Minimize SSE

$$\min \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$

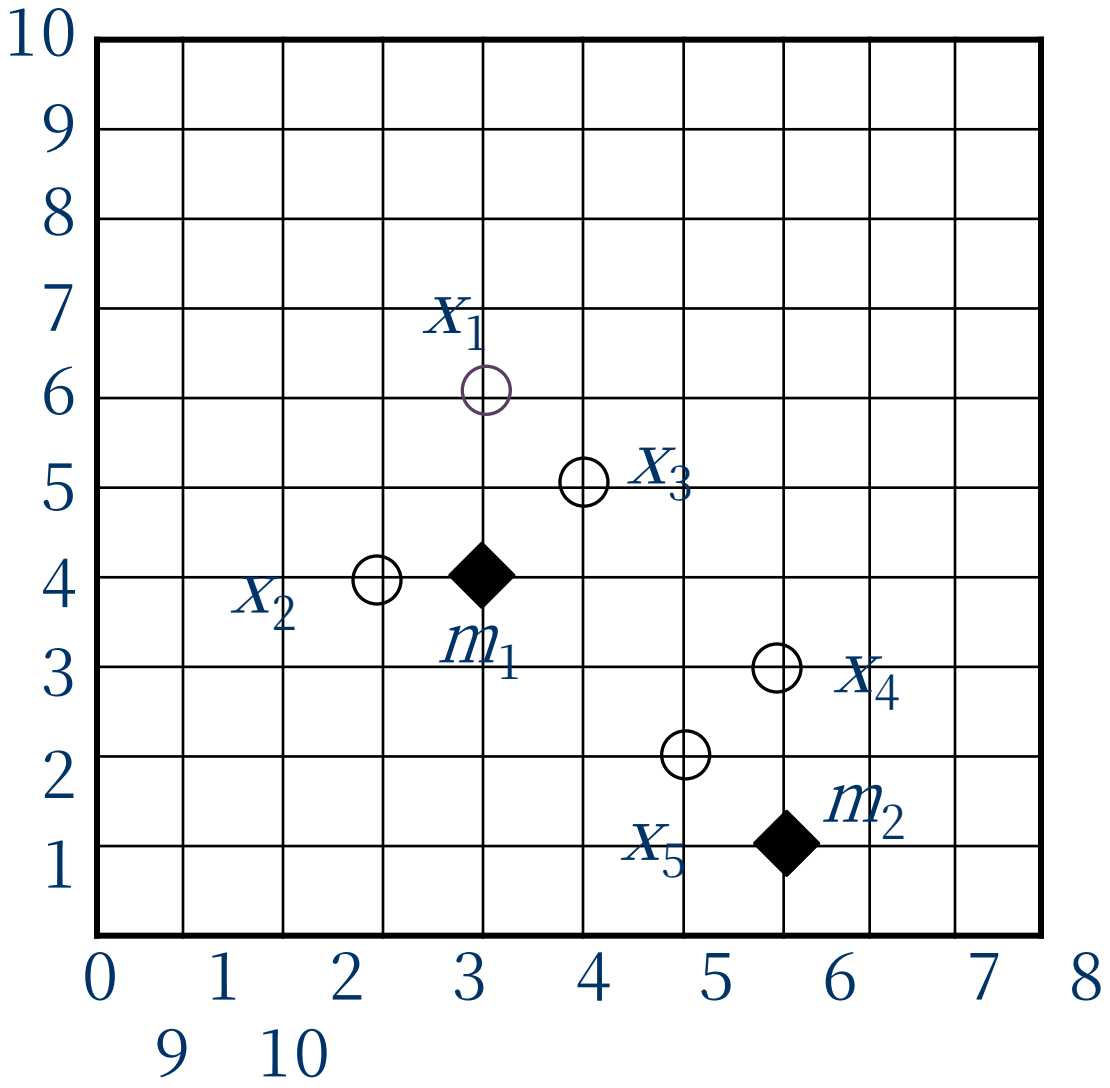
- **Two sets of variables to minimize**
  - Each object *x* belongs to which cluster? $x \in C_j$
  - What's the cluster centroid? m_j
- **Minimize the error wrt each set of variable alternatively**
  - Fix the cluster centroid—find cluster assignment that minimizes the current error
  - Fix the cluster assignment—compute the cluster centroids that minimize the current error

Cluster Assignment Step

$$\min \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$

- Cluster centroids (m_j) are known
- For each object
  - Choose C_j among all the clusters for x such that the distance between x and m_j is the minimum
  - Choose another cluster will incur a bigger error
- Minimize error on each object will minimize the SSE

Example—Cluster Assignment



Given m_1, m_2, which cluster each of the five points belongs to?

Assign points to the closet centroid—minimize SSE

$$x_1, x_2, x_3 \in C_1$$

$$x_4, x_5 \in C_2$$

$$SSE = (x_1 - m_1)^2 + (x_2 - m_1)^2 + (x_3 - m_1)^2 + (x_4 - m_2)^2 + (x_5 - m_2)^2$$

Cluster Centroid Computation Step

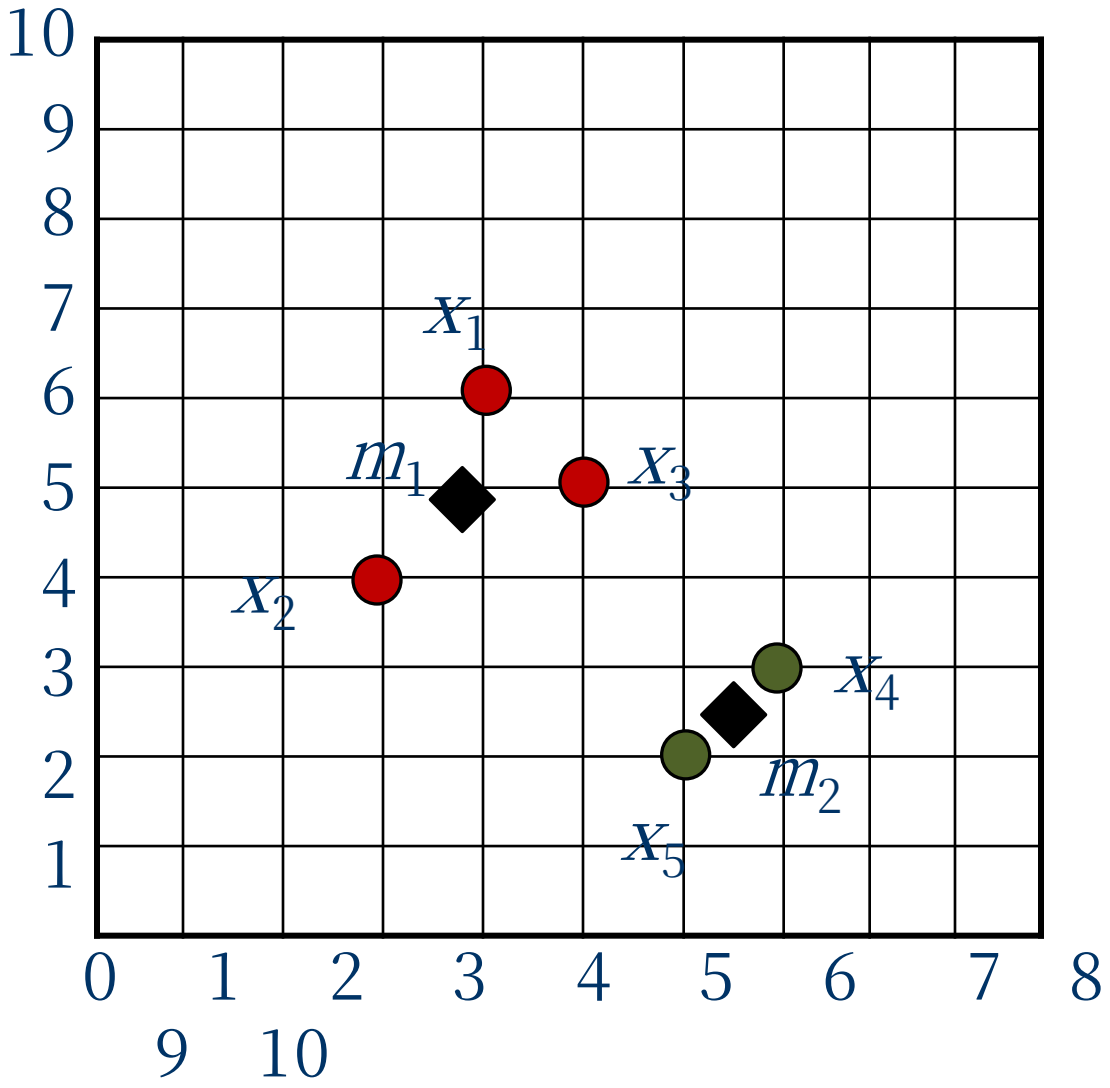
$$\min \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$

- For each cluster
  - Choose cluster centroid m_j as the center of the points

$$m_j = \frac{\sum_{x \in C_j} x}{|C_j|}$$

- Minimize error on each cluster will minimize the SSE

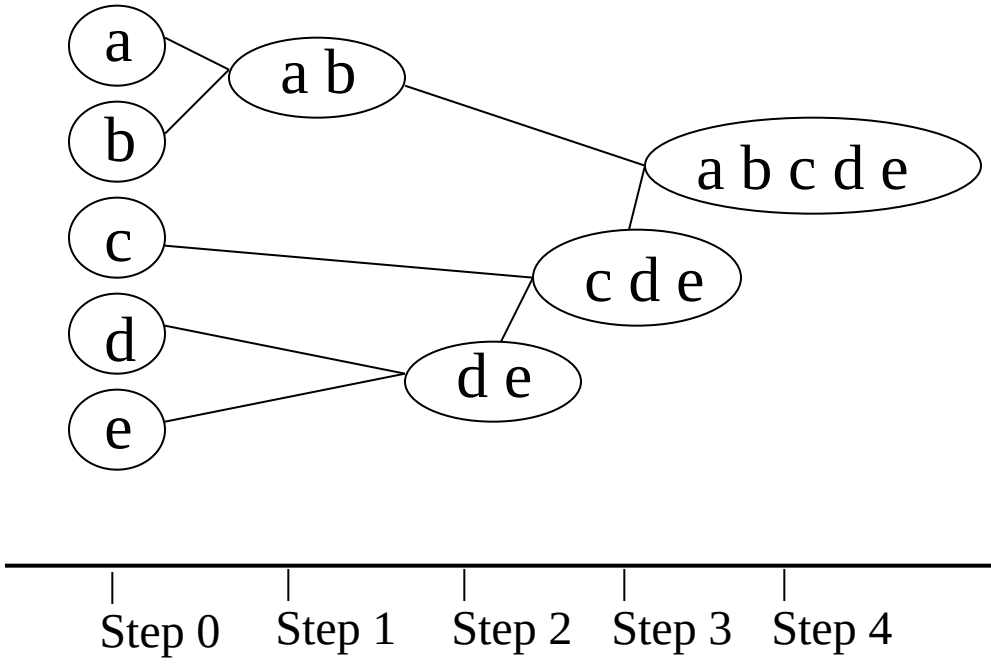
Example—Cluster Centroid Computation



Given the cluster assignment, compute the centers of the two clusters

Hierarchical Clustering

- **Agglomerative approach**



Initialization:

Each object is a cluster

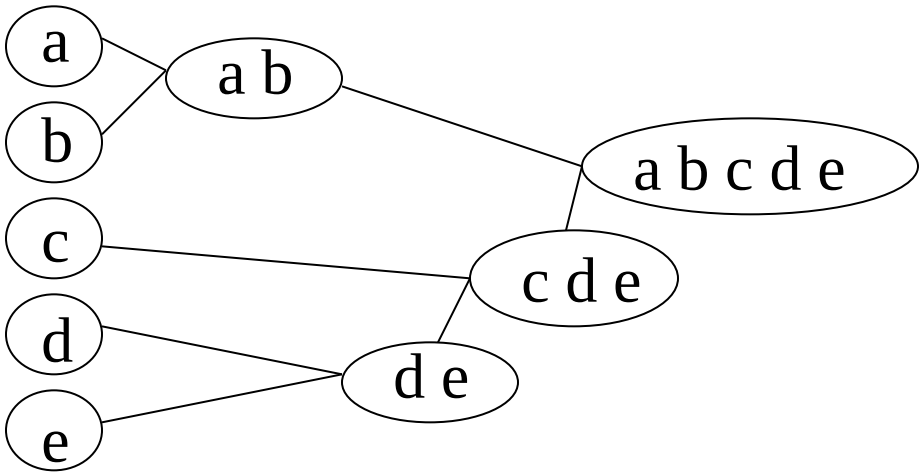
Iteration:

Merge two clusters which are most similar to each other;

Until all objects are merged into a single cluster

Hierarchical Clustering

- **Divisive Approaches**



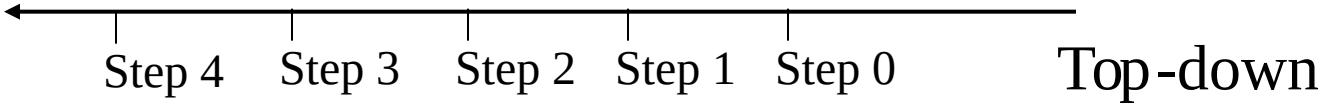
Initialization:

All objects stay in one cluster

Iteration:

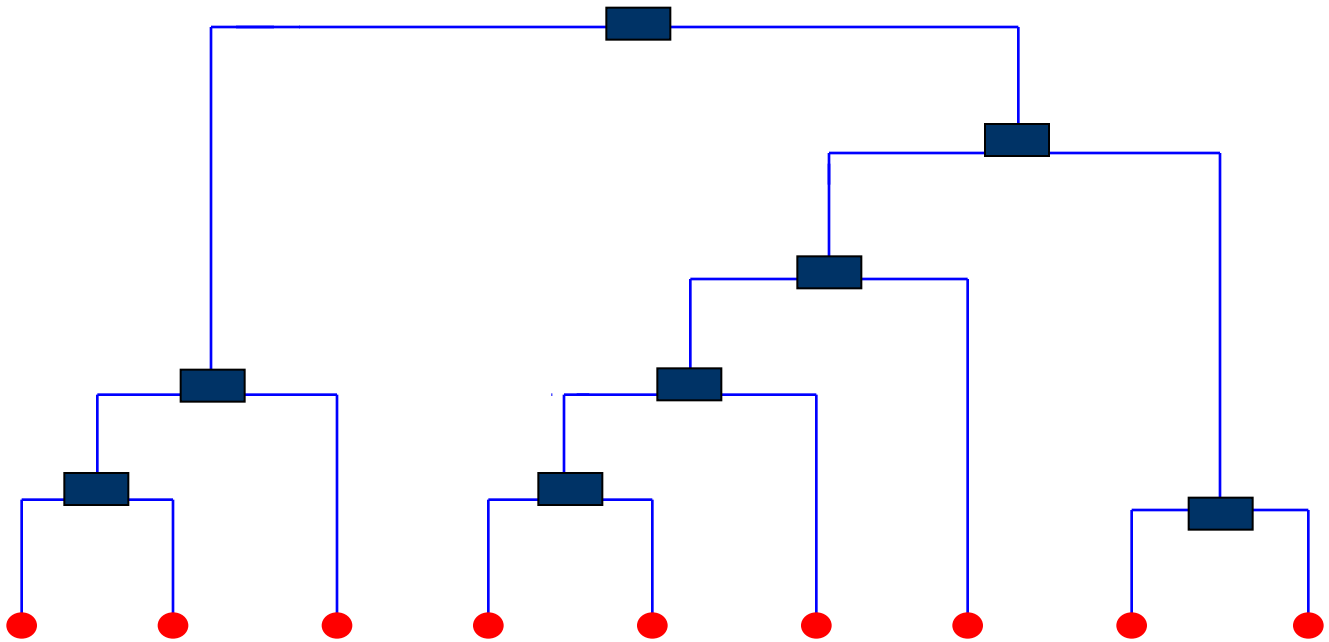
Select a cluster and split it into two sub clusters

Until each leaf cluster contains only one object



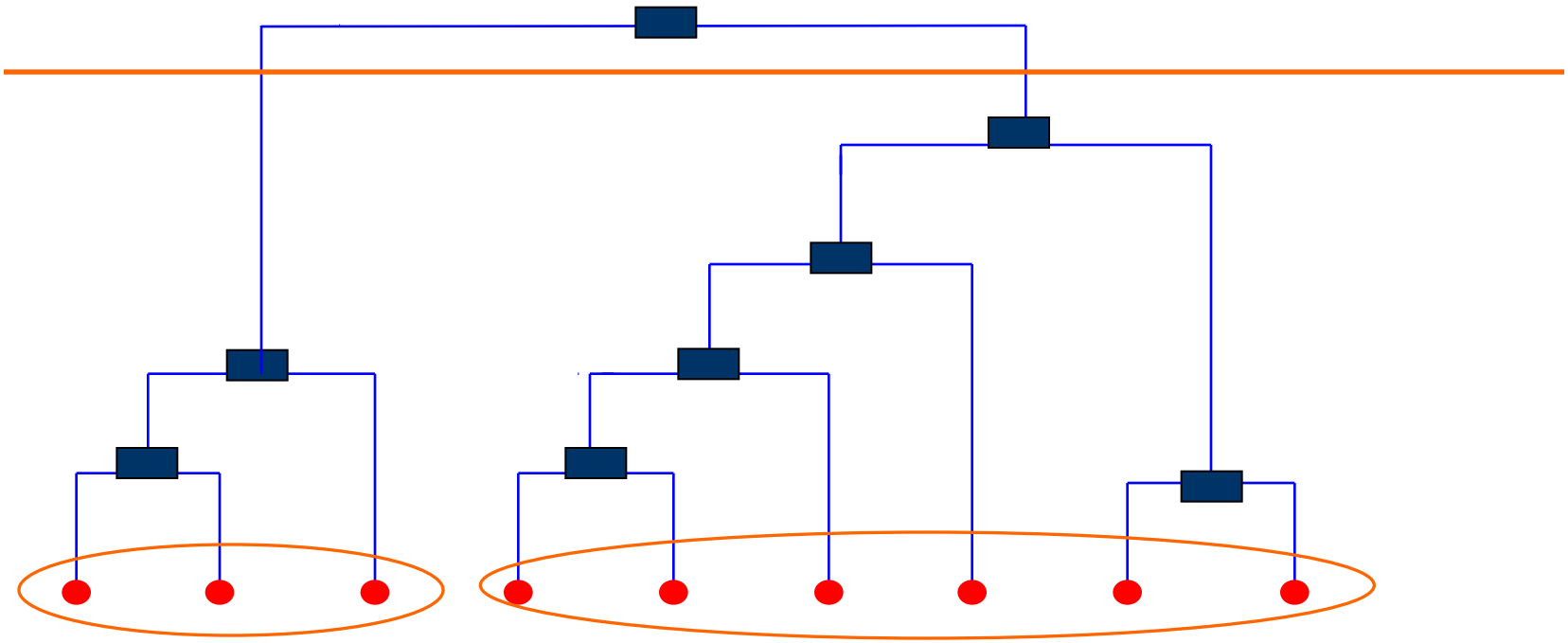
Dendrogram

- A tree that shows how clusters are merged/split hierarchically
- Each node on the tree is a cluster; each leaf node is a singleton cluster



Dendrogram

- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster

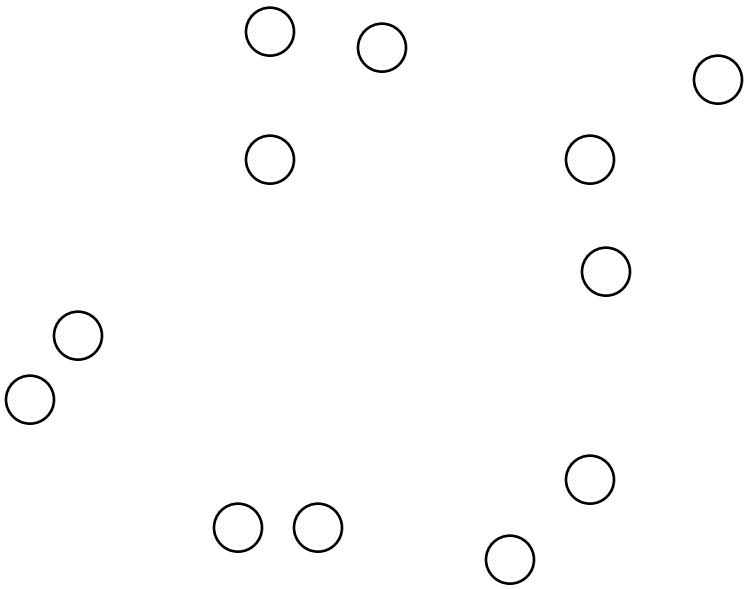


Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the distance matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the distance matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the distance between two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

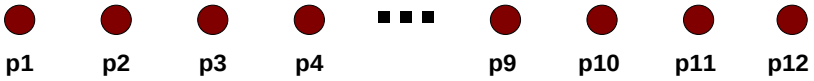
Starting Situation

- Start with clusters of individual points and a distance matrix



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Distance Matrix

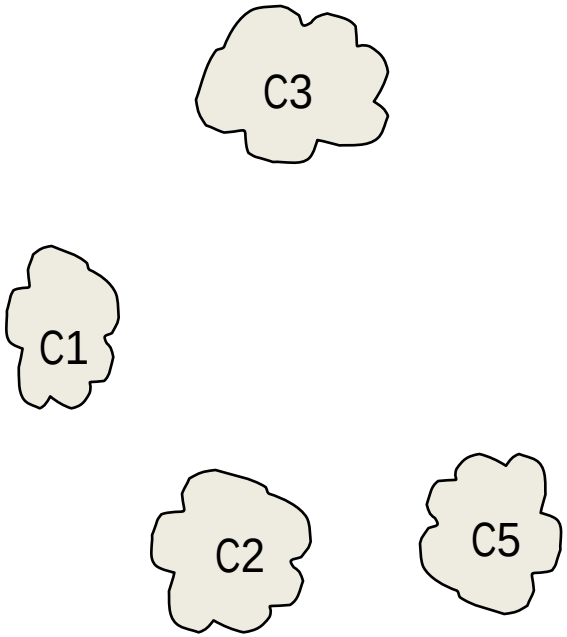


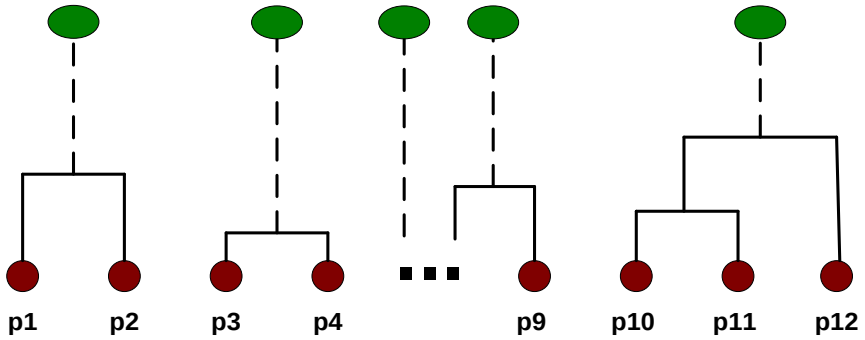
Intermediate Situation

- After some merging steps, we have some clusters
- Choose two clusters that has the smallest distance (largest similarity) to merge

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

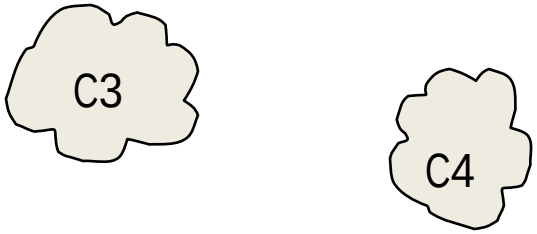
Distance Matrix

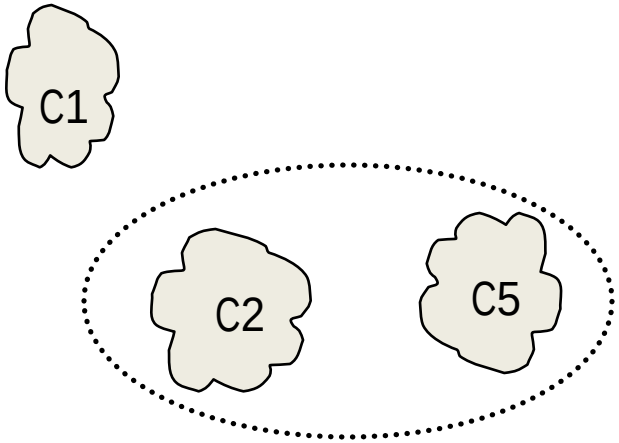




Intermediate Situation

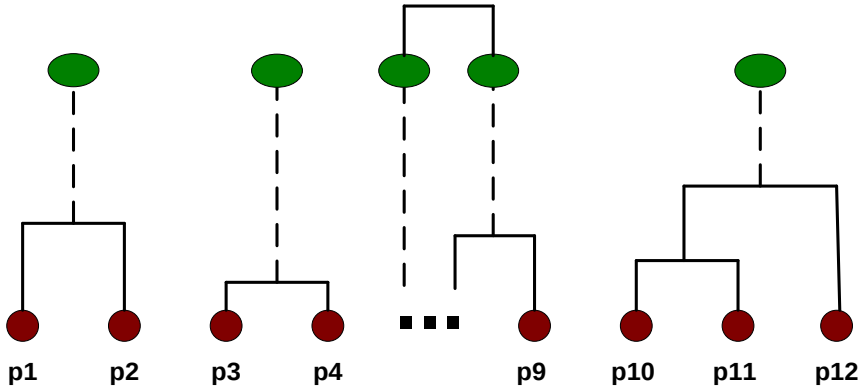
- We want to merge the two closest clusters (C2 and C5) and update the distance matrix.





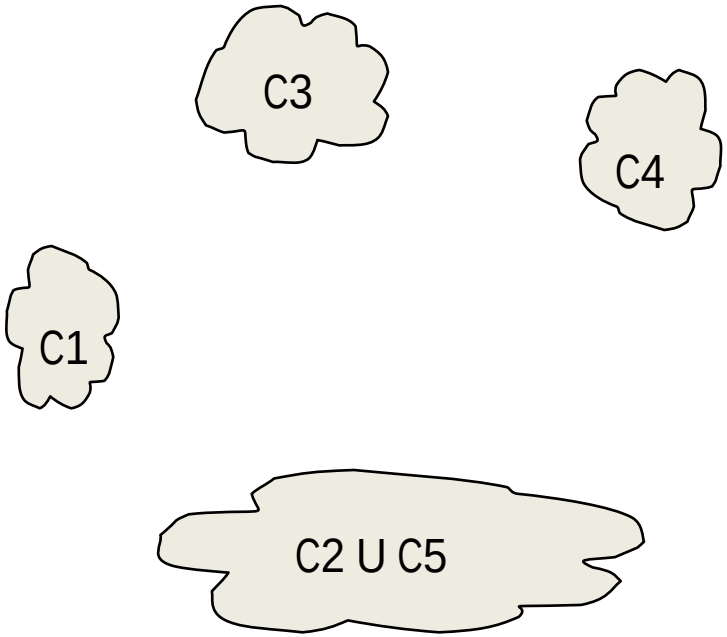
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance Matrix



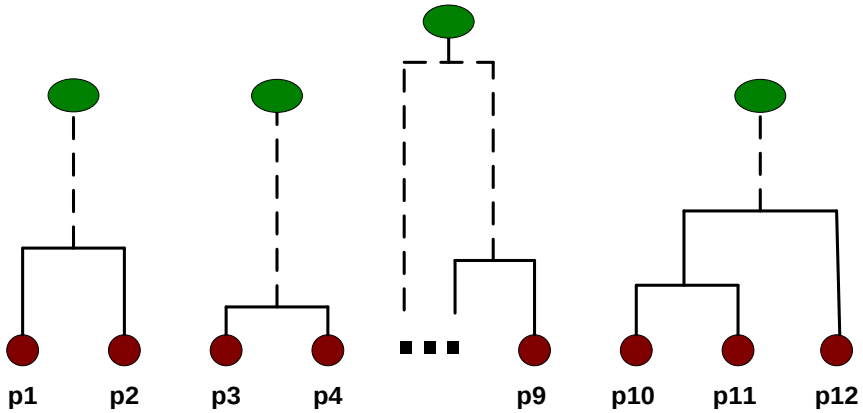
After Merging

- The question is “How do we update the distance matrix?”

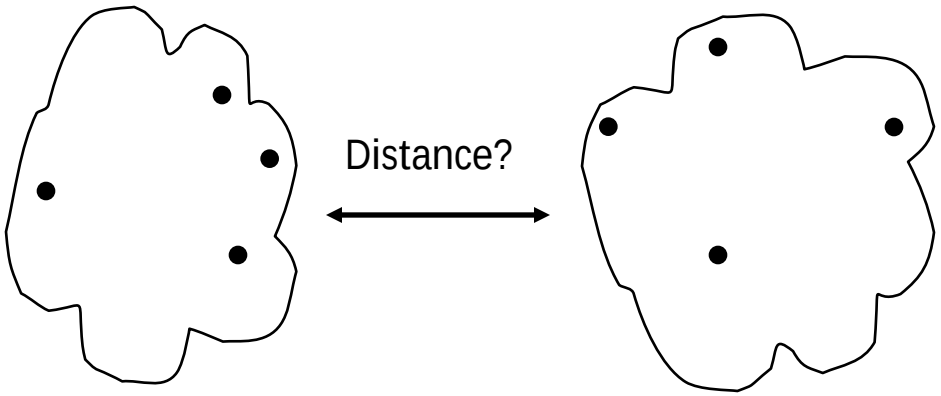


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Distance Matrix



How to Define Inter-Cluster Distance



- MIN
- MAX
- Group Average
- Distance Between Centroids
- ……

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Distance Matrix

Question

- **Talk about big data**
  - What's the definition of big data?
  - What applications generate big data?
  - What are the challenges?
  - What are the technologies for mining big data?