RV College of Engineering®

Go, change the world

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING **(IS353IA)**

Unit 5

Department Of Information Science & Engineering

Unit 5

Unsupervised Learning- Overview, What Is Cluster Analysis, Different Types of Clustering's, Different Types of Clusters

K-means-The Basic K-means Algorithm, Additional Issues, Bisecting K-means, K-means and Different Types of Clusters, Strengths and Weaknesses, K-means as an Optimization Problem

Cluster Evaluation-Overview, Unsupervised Cluster Evaluation Using Cohesion and Separation, Unsupervised Cluster Evaluation Using the Proximity Matrix, Determining the Correct Number of Clusters, Supervised Measures of Cluster Validity, Assessing the Significance of Cluster Validity Measures, Choosing a Cluster Validity Measure

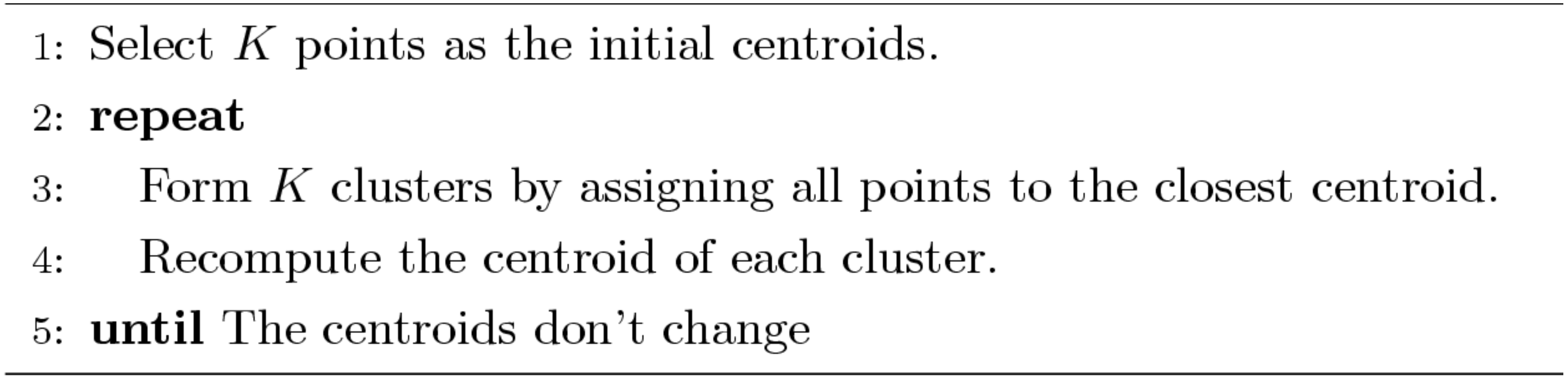
Partitional Clustering

Partitional Clustering

- Given
    - A data set of **n objects**
    - **K the number of clusters to form**
- Organize the objects into k partitions **(k<=n) where each partition** represents a cluster
- The clusters are formed to optimize an objective partitioning criterion
    - Objects within a cluster are **similar**
    - Objects of different clusters are **dissimilar**

K-means Clustering

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



```
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’ or some measure of clustering doesn’t change.
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters, I = number of iterations, d = number of attributes

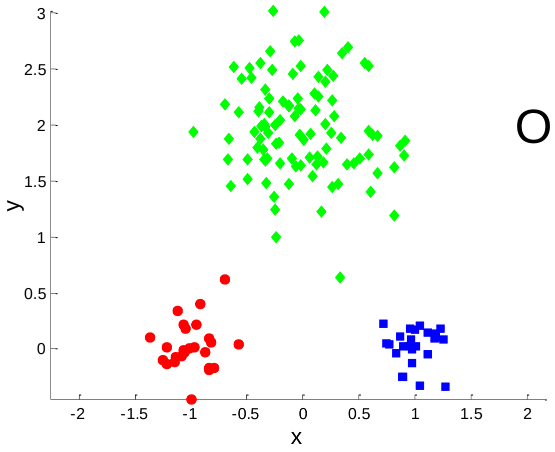
Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

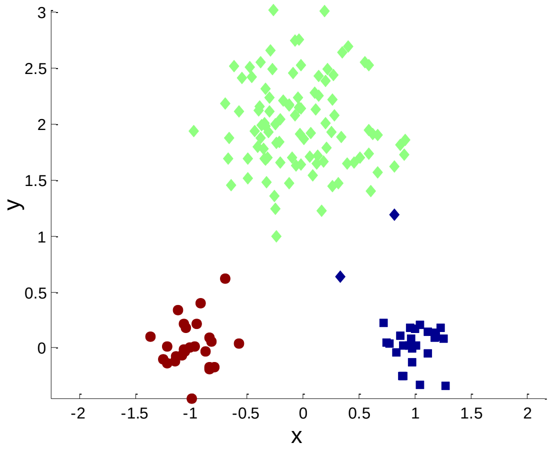
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, i.e. the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

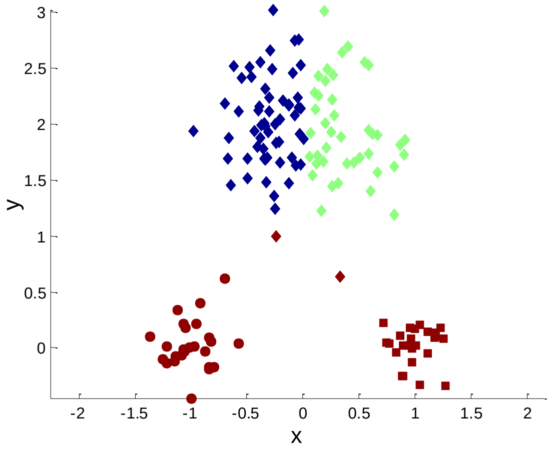
Two different K-means Clusterings



Original Points

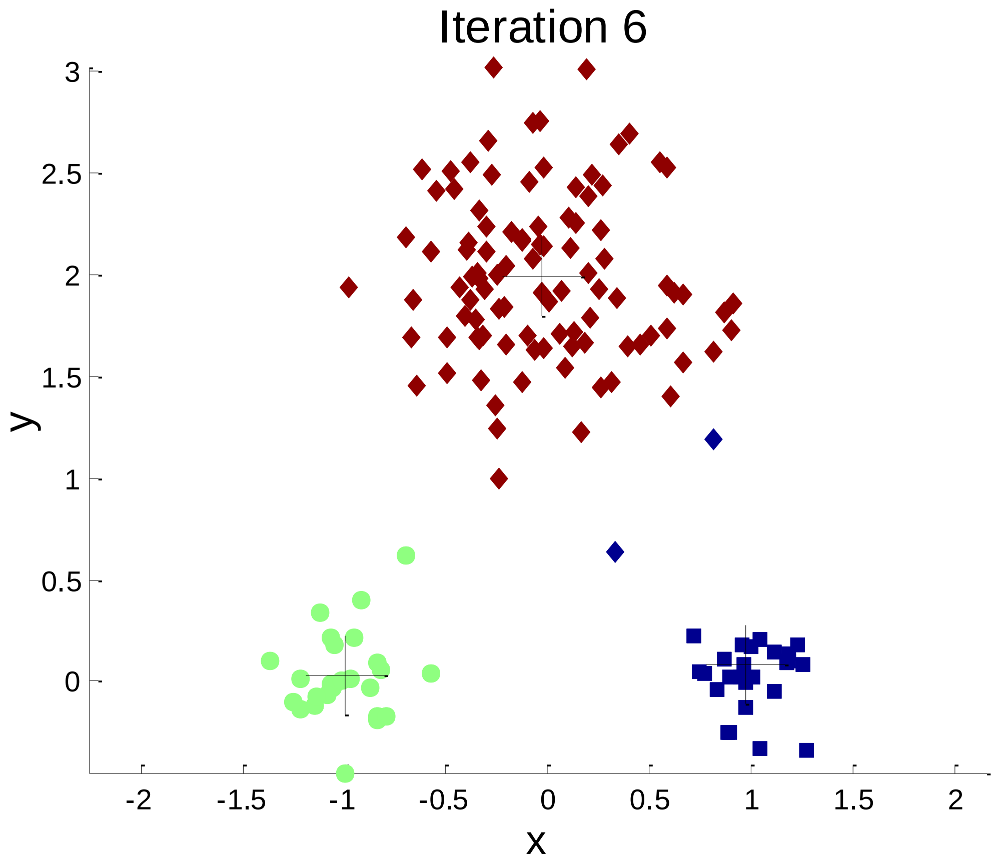


Optimal Clustering

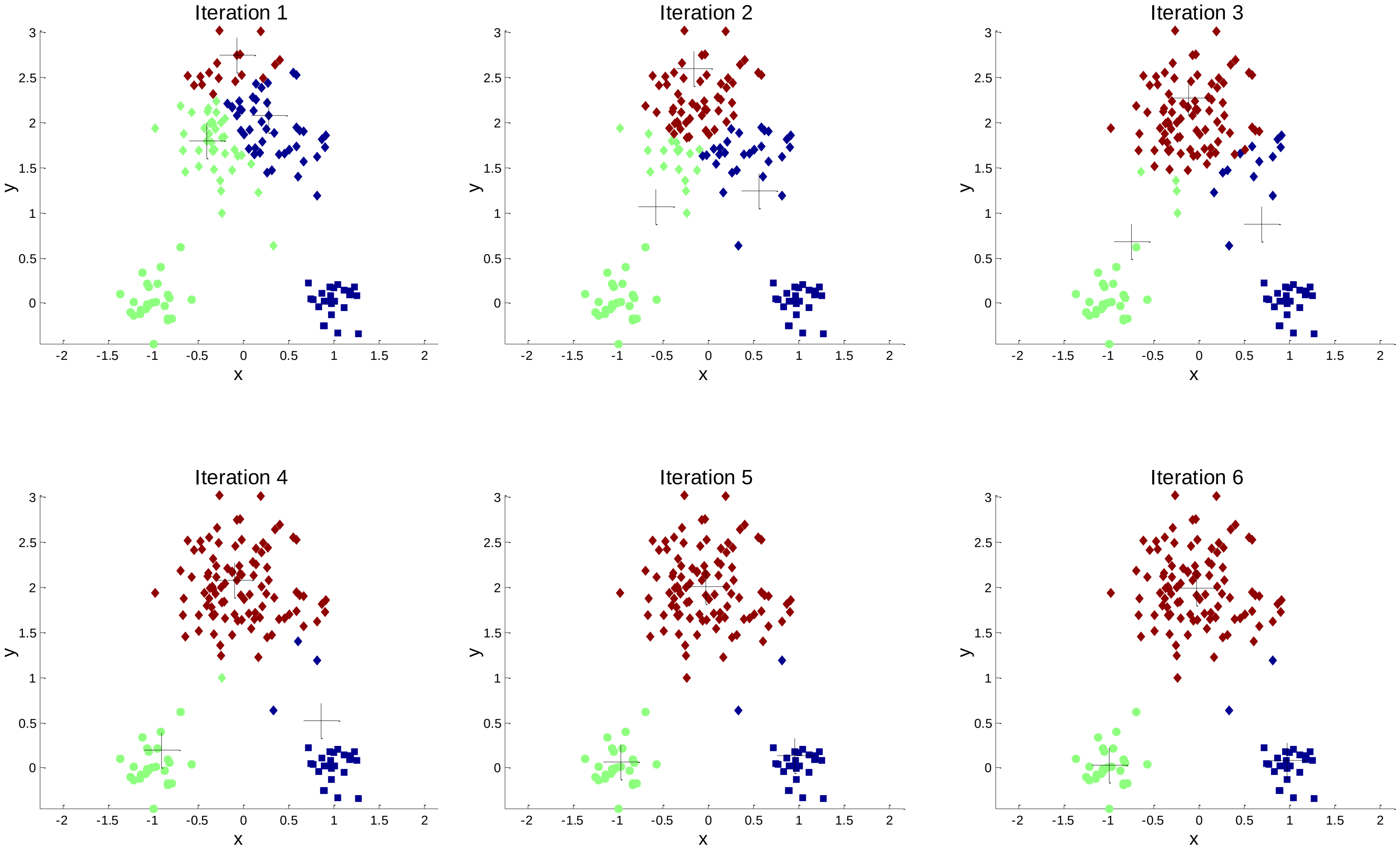


Sub-optimal Clustering

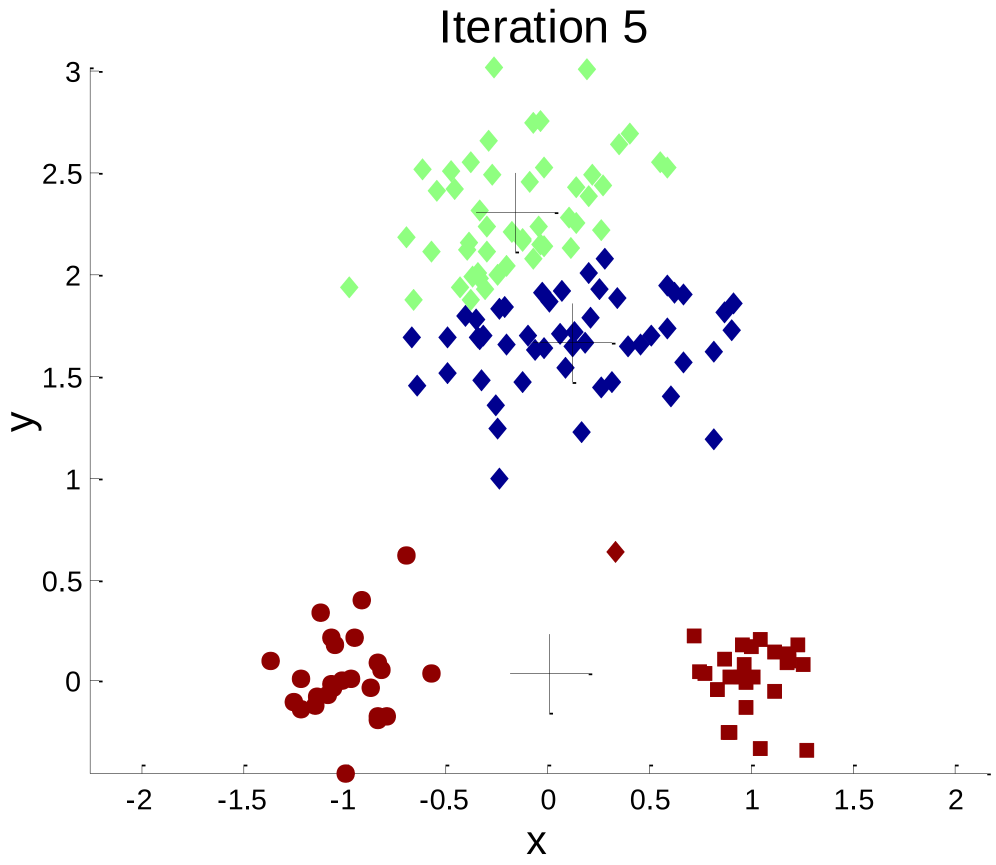
Importance of Choosing Initial Centroids (Case i)



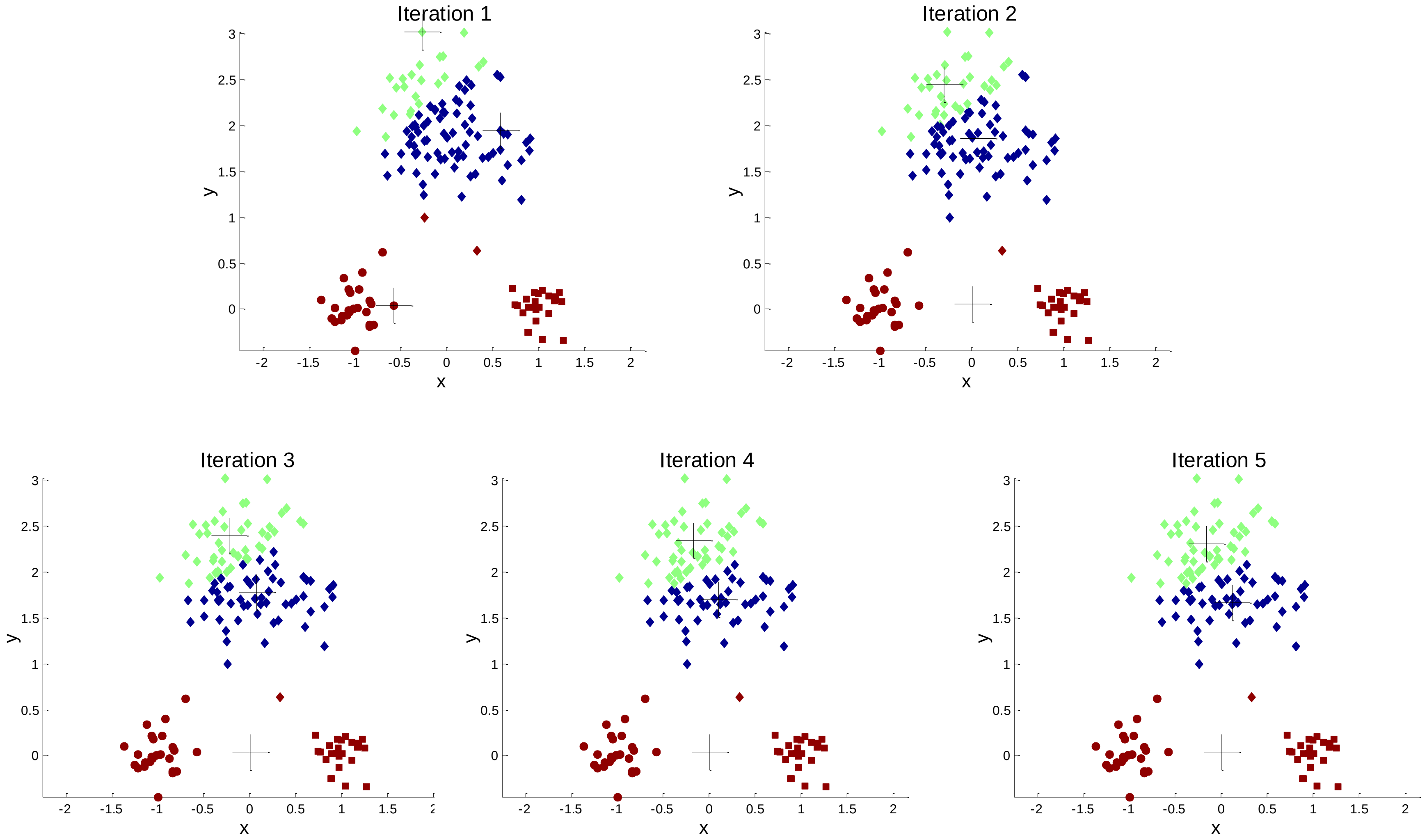
Importance of Choosing Initial Centroids (Case i)



Importance of Choosing Initial Centroids (Case ii)



Importance of Choosing Initial Centroids (Case ii)



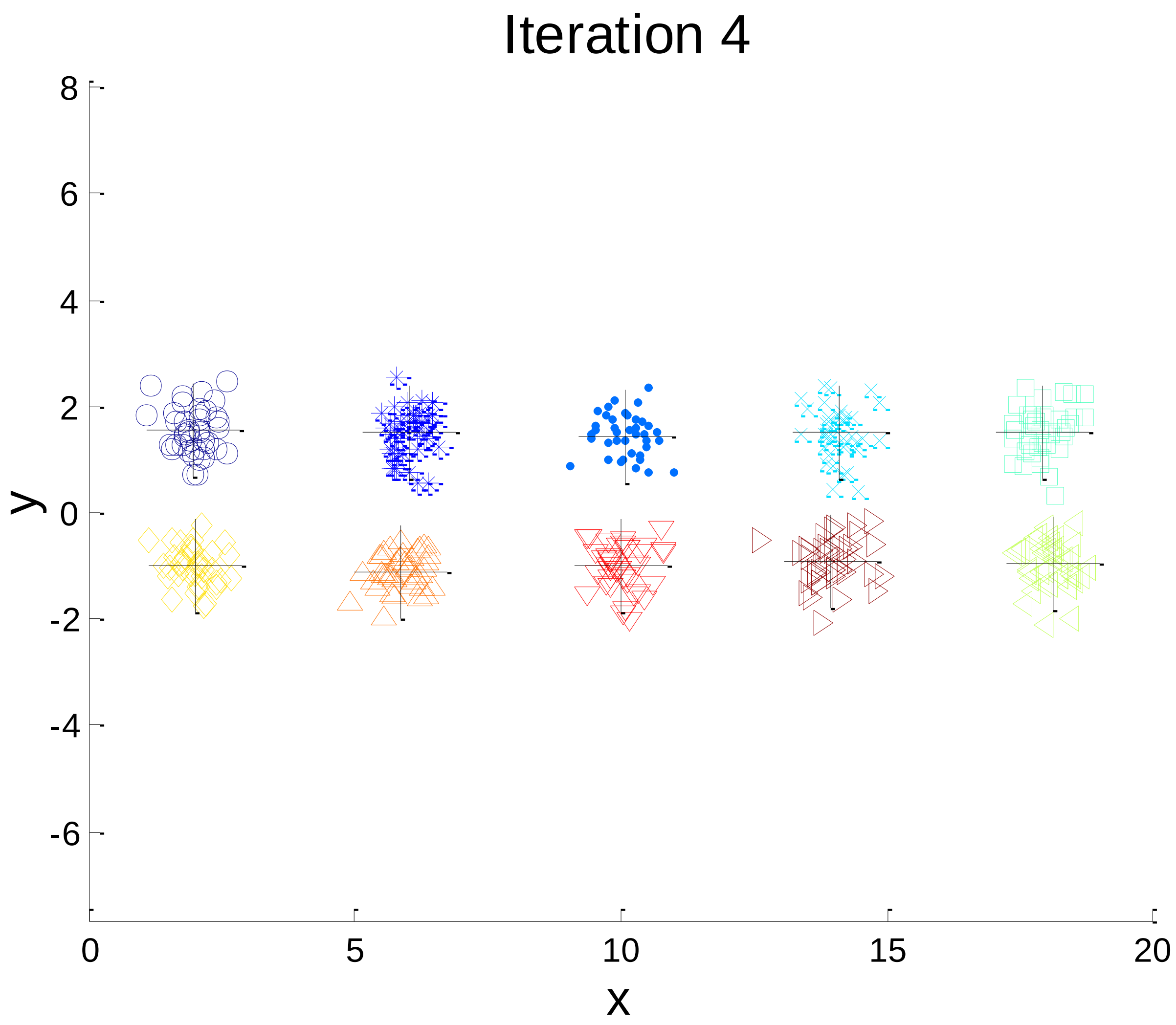
Problems with Selecting Initial Points

- If there are K ‘real’ clusters then the chance of selecting one centroid from each cluster is small.
  - Chance is relatively small when K is large
  - If clusters are the same size, n, then

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

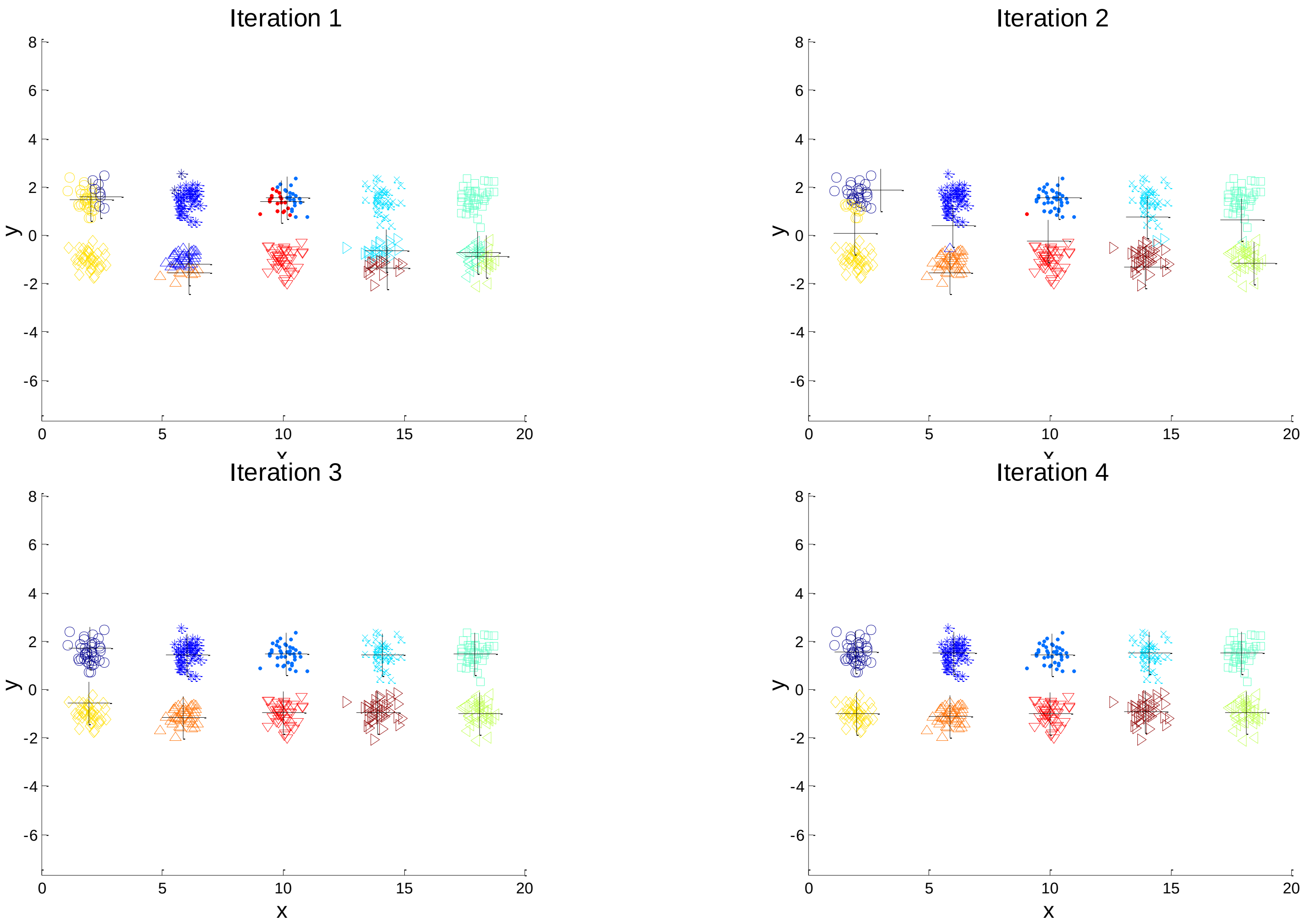
  - For example, if K = 10, then probability = $10!/10^{10}$ = 0.00036
  - Sometimes the initial centroids will readjust themselves in ‘right’ way, and sometimes they don’t
  - Consider an example of five pairs of clusters
- Initial centers from different clusters may produce good clusters

10 Clusters Example (Good Clusters)



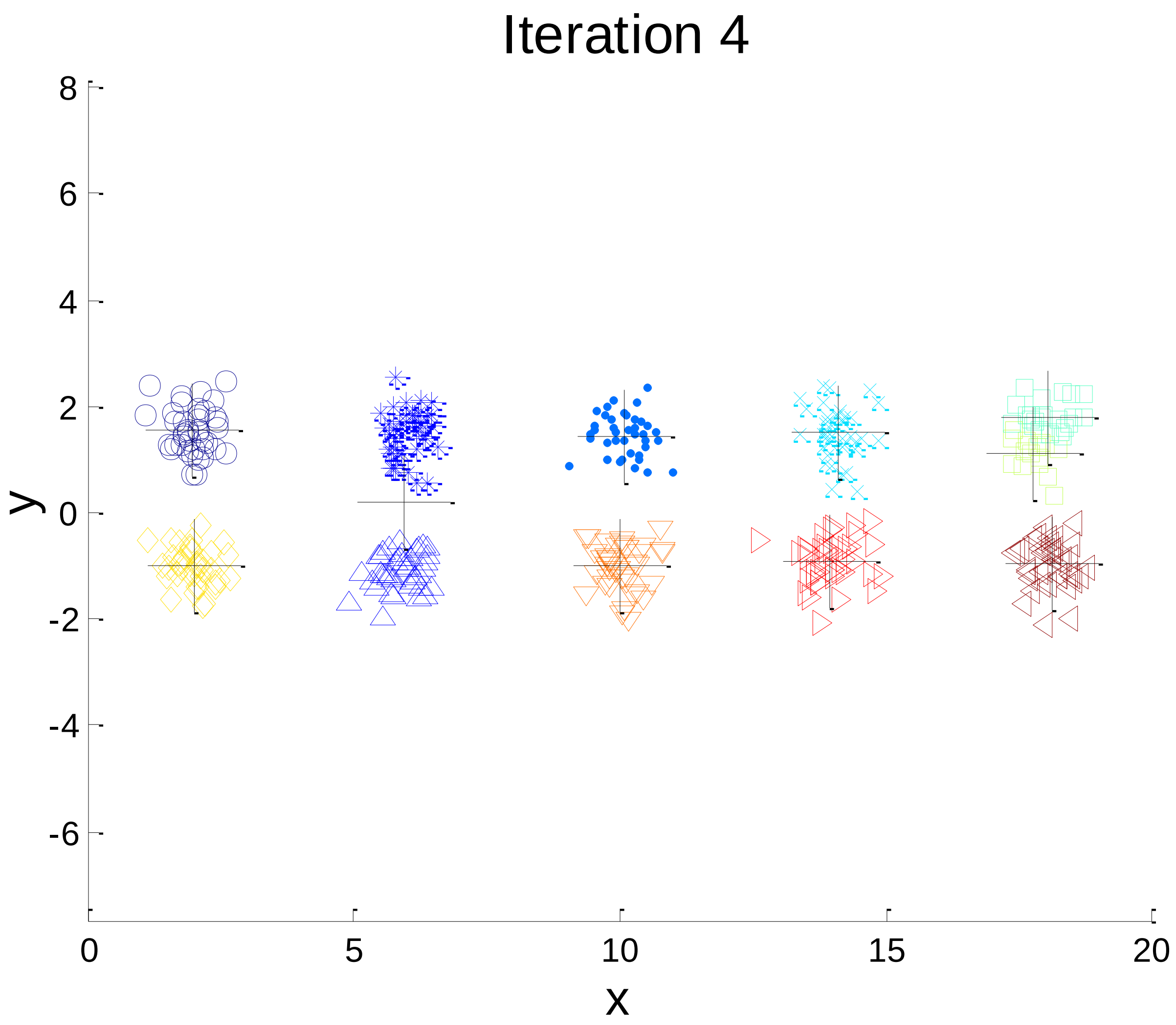
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example (Good Clusters)



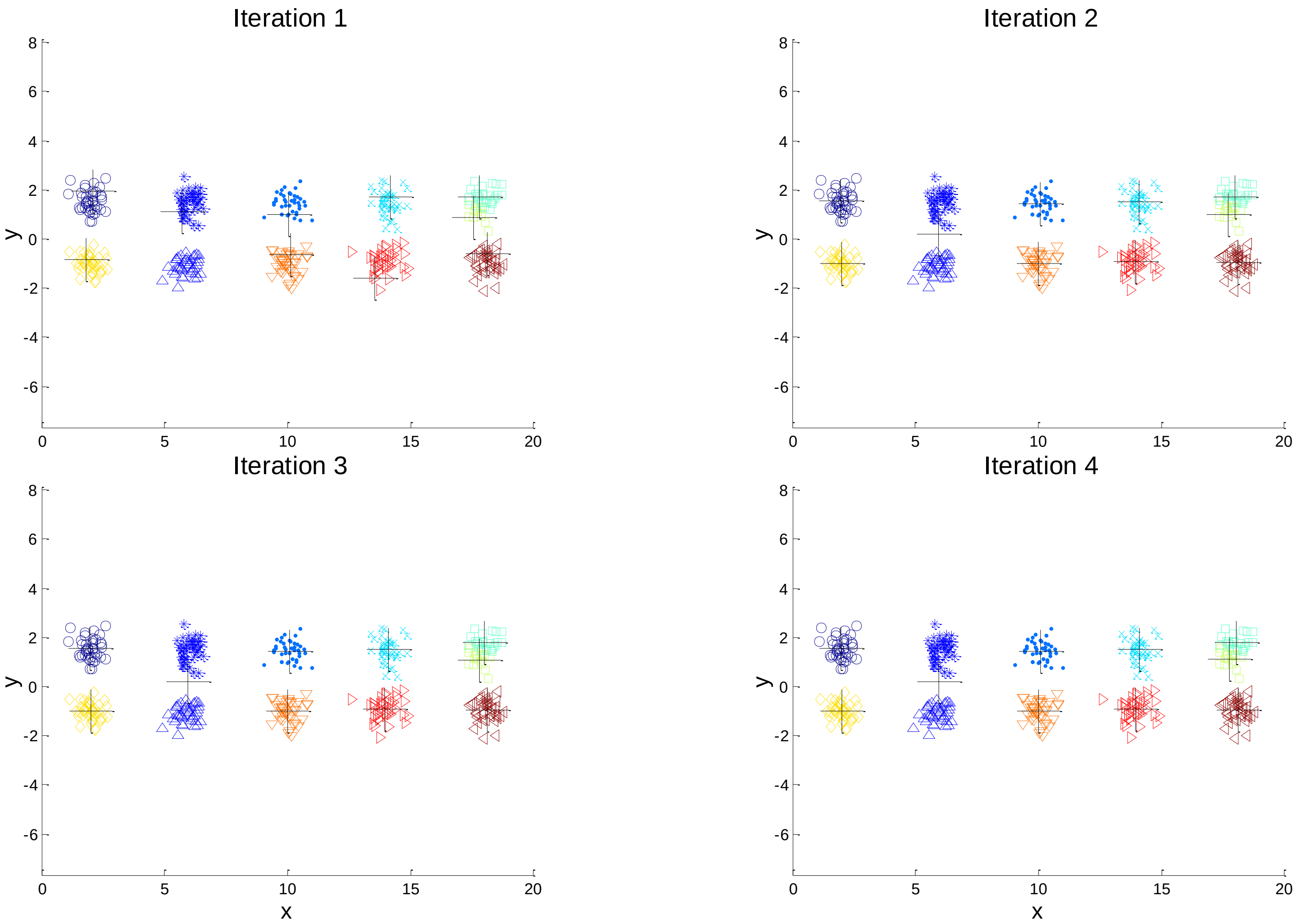
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example (Bad Clusters)



Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example (Bad Clusters)



Starting with some pairs of clusters having three initial centroids, while other have only one.

Solutions to Initial Centroids Problem

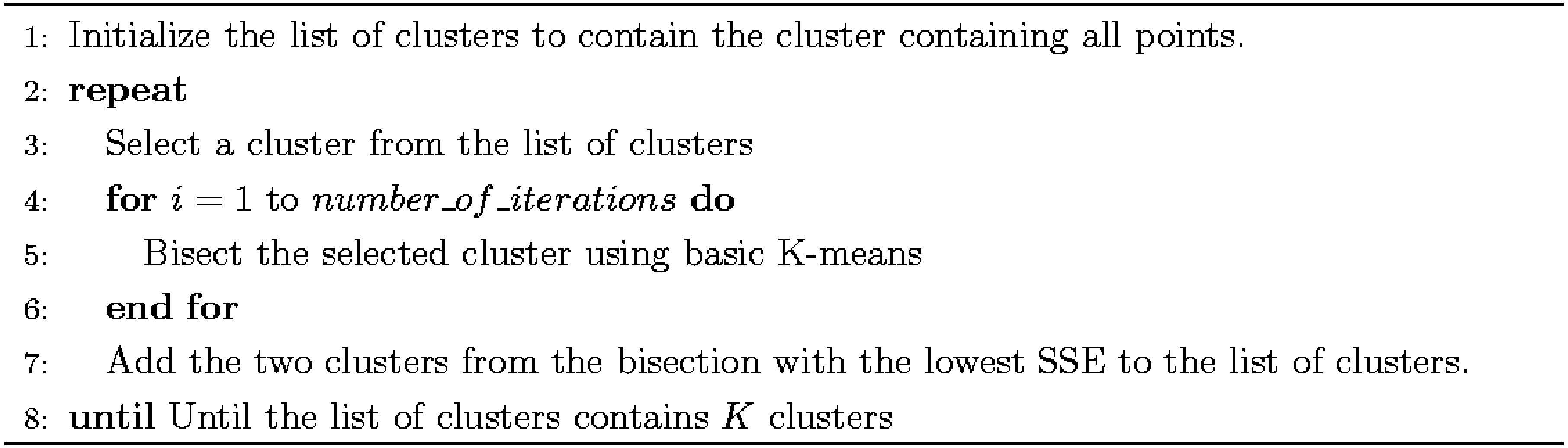
- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Post-processing
- Bisecting K-means
  - Not as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

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

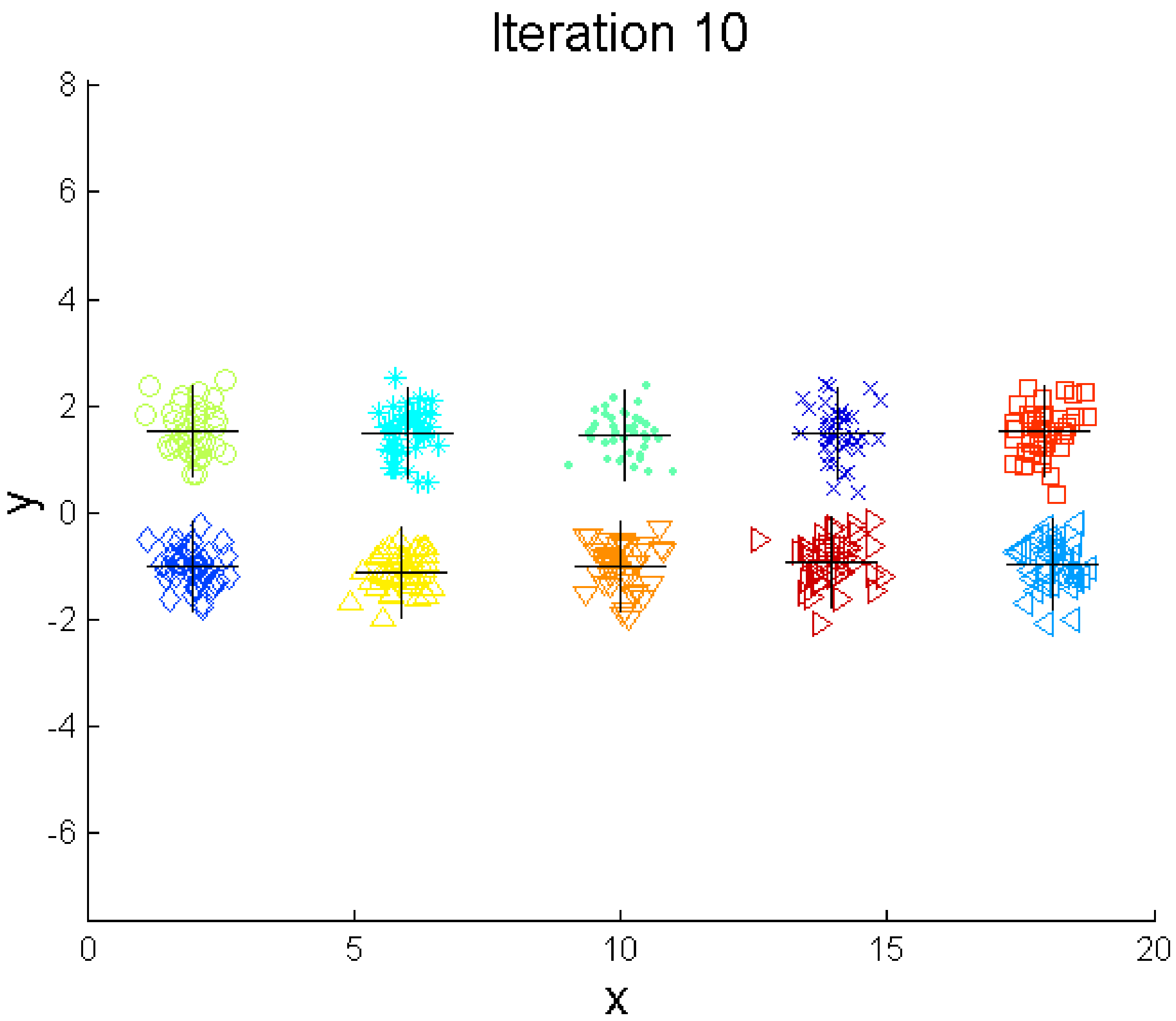
Bisecting K-means

- Bisecting K-means algorithm
  - Variant of K-means that can produce a partitional or a hierarchical clustering



```
Initialize the list of clusters to contain the cluster containing all points.
repeat
    Select a cluster from the list of clusters
    for $i = 1$ to $number\_of\_iterations$ do
        Bisect the selected cluster using basic K-means
    end for
    Add the two clusters from the bisection with the lowest SSE to the list of clusters.
until Until the list of clusters contains $K$ clusters
```

Bisecting K-means Example

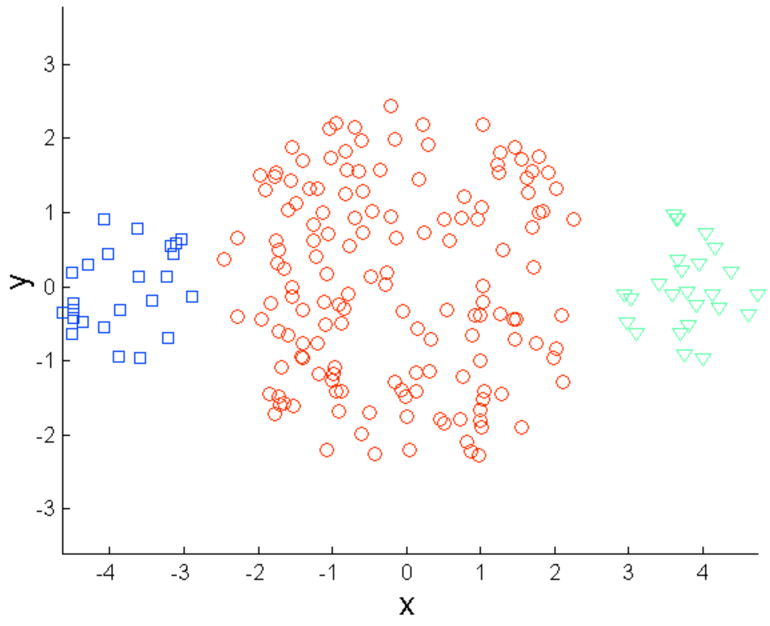


Limitations of K-means

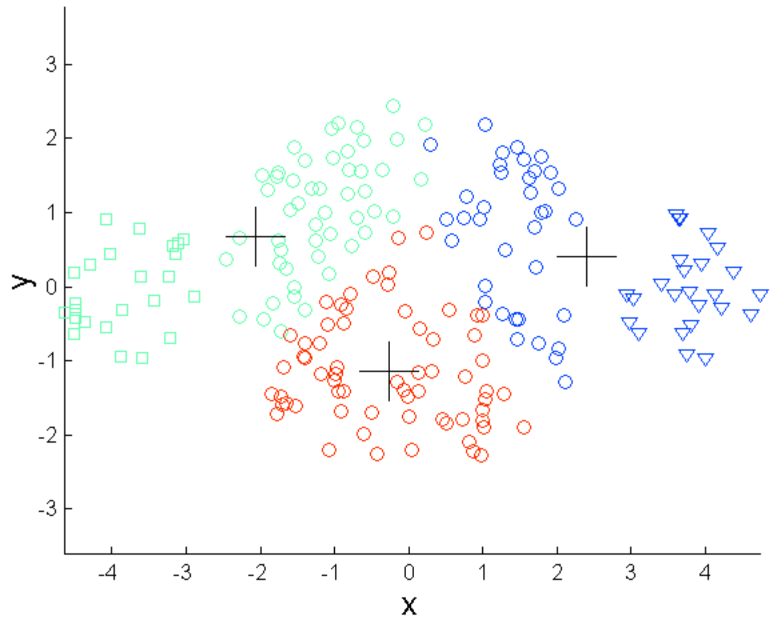
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

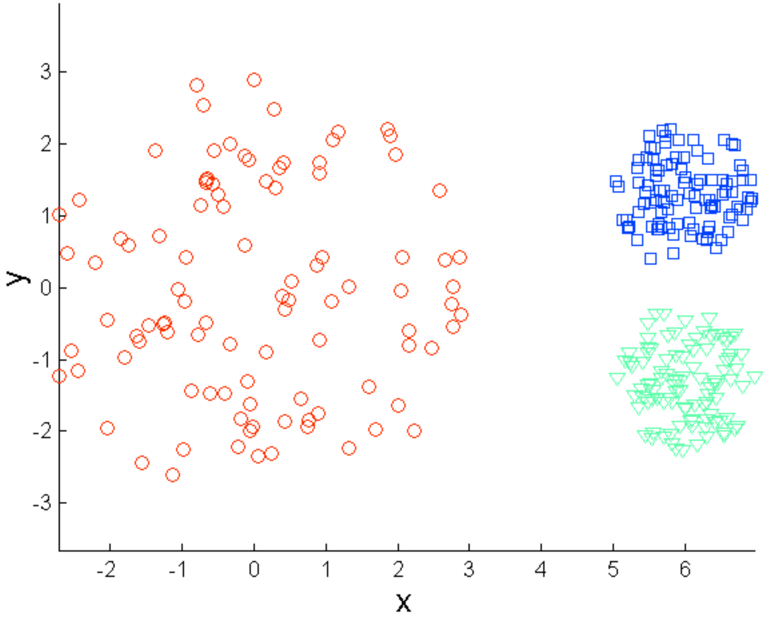


Original Points

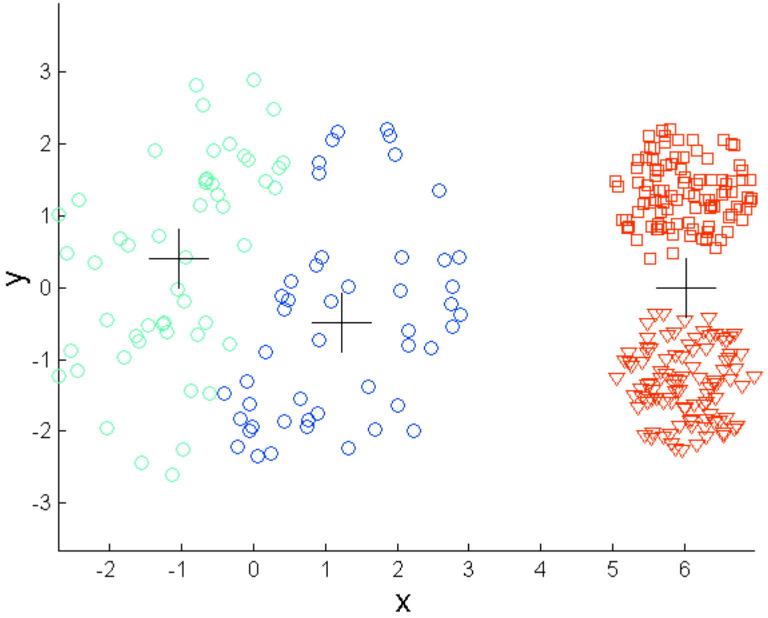


K-means (3 Clusters)

Limitations of K-means: Differing Density

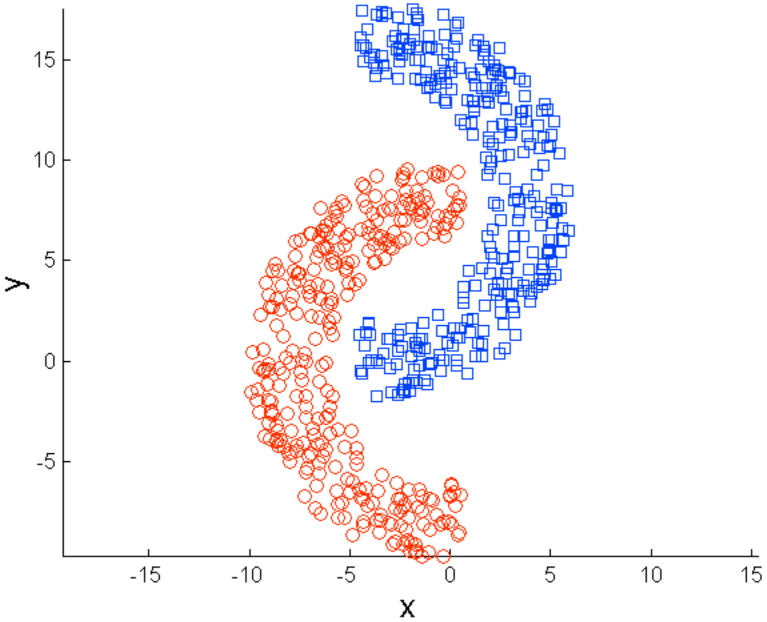


Original Points

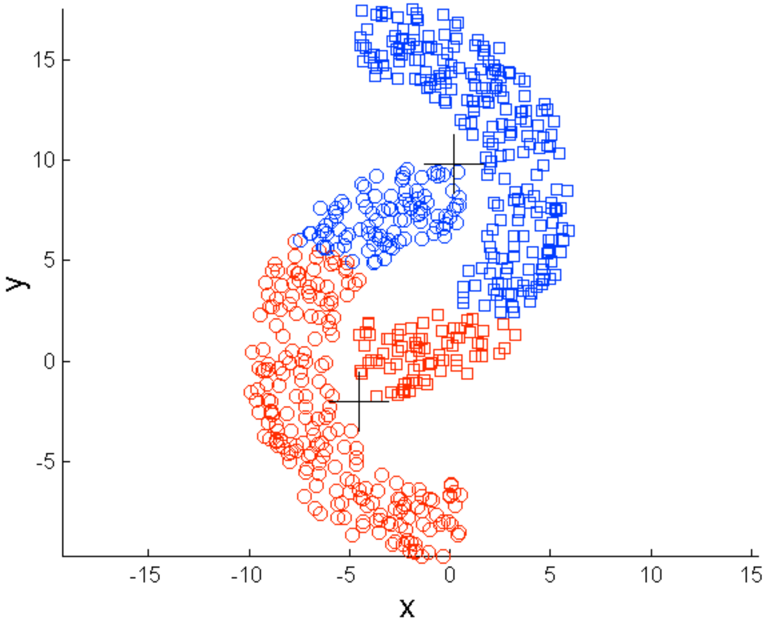


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

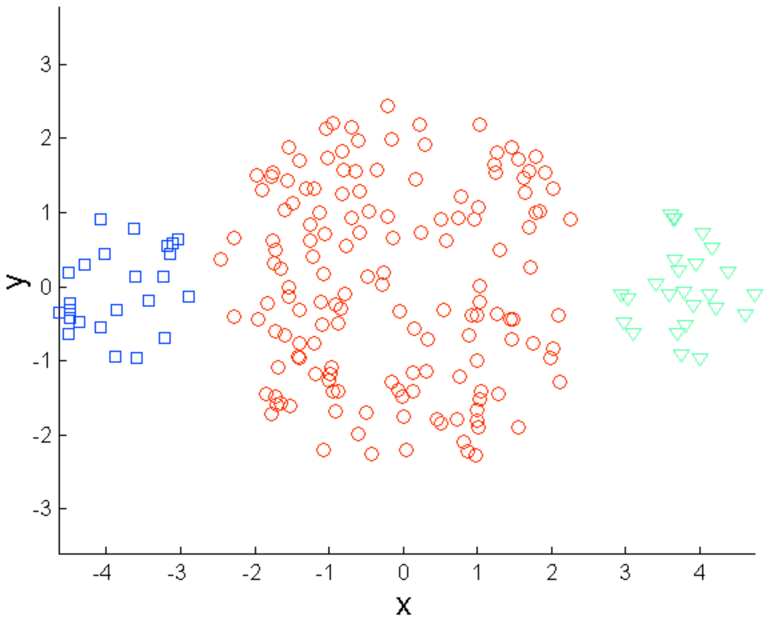


Original Points

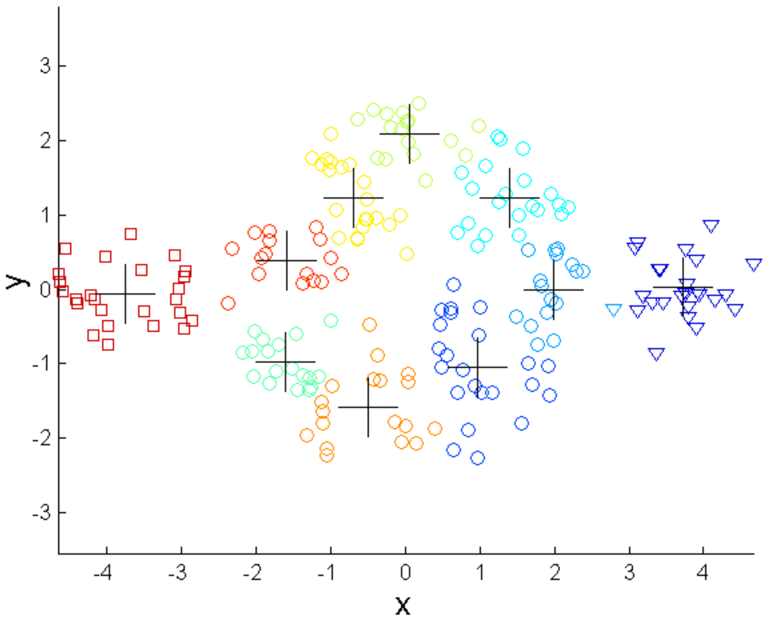


K-means (2 Clusters)

Overcoming K-means Limitations



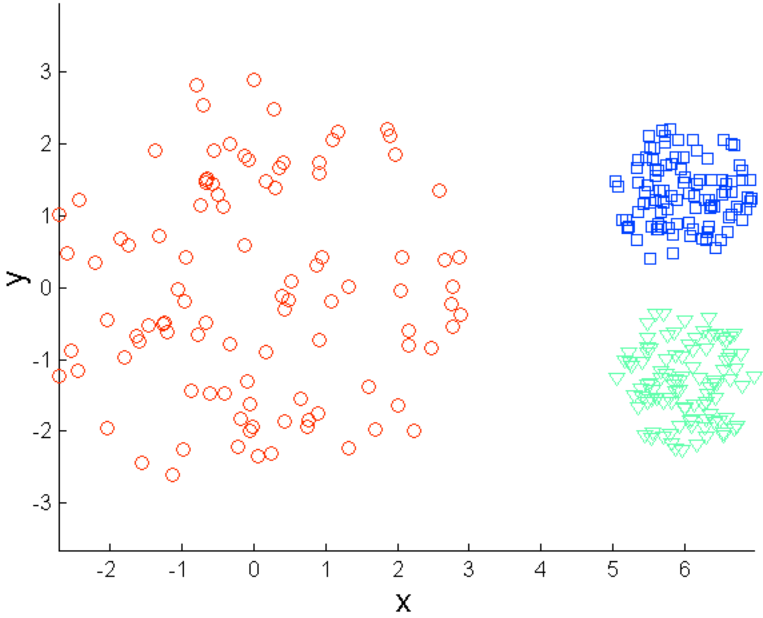
Original Points



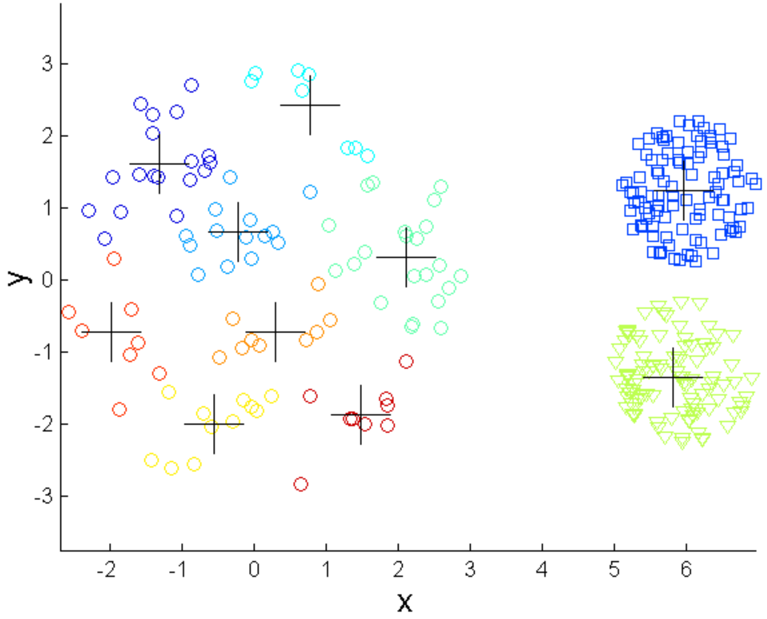
K-means Clusters

- One solution is to use many clusters.
  - Find parts of clusters, but need to put together.

Overcoming K-means Limitations



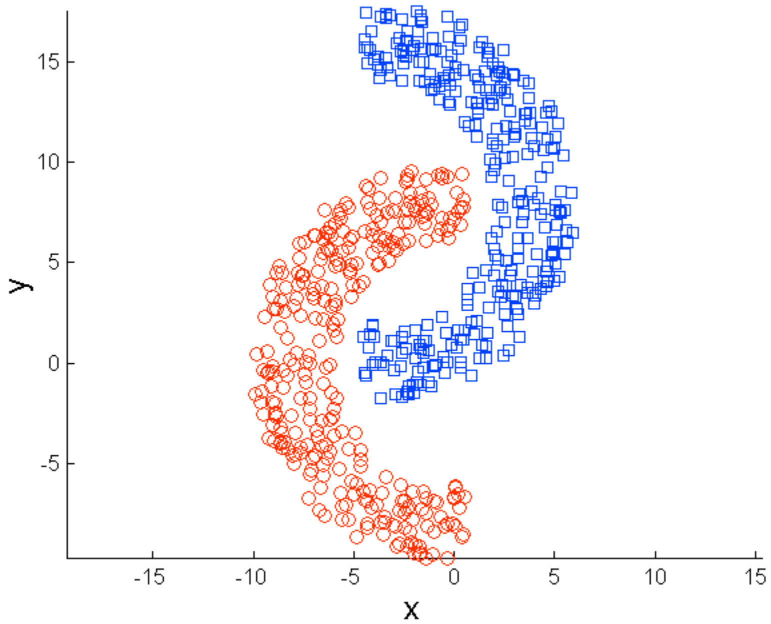
Original Points



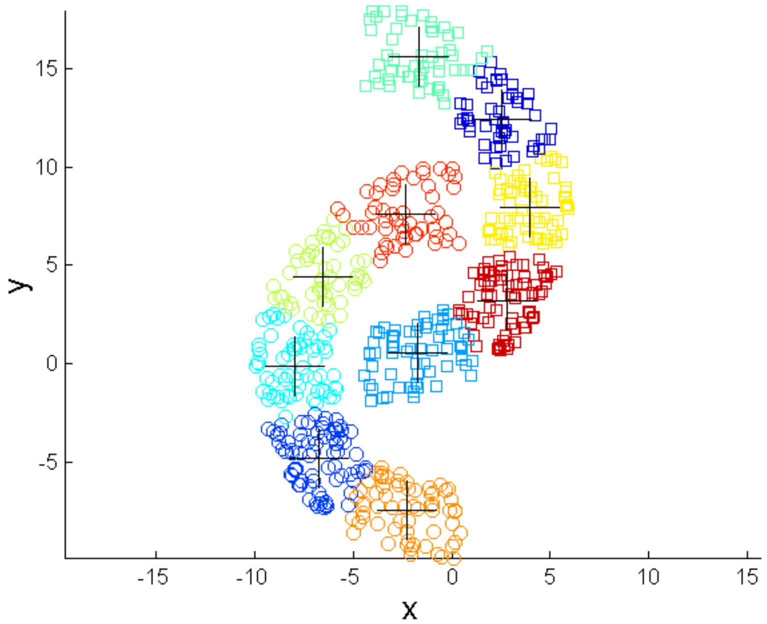
K-means Clusters

- One solution is to use many clusters.
  - Find parts of clusters, but need to put together.

Overcoming K-means Limitations



Original Points

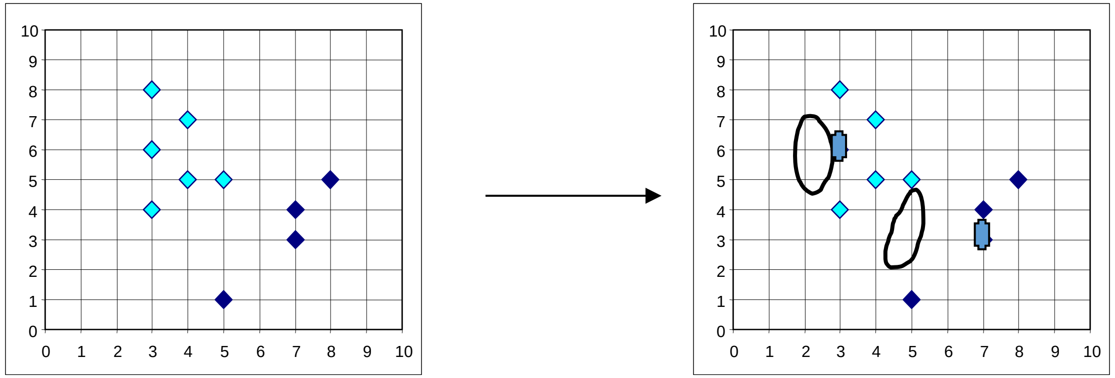


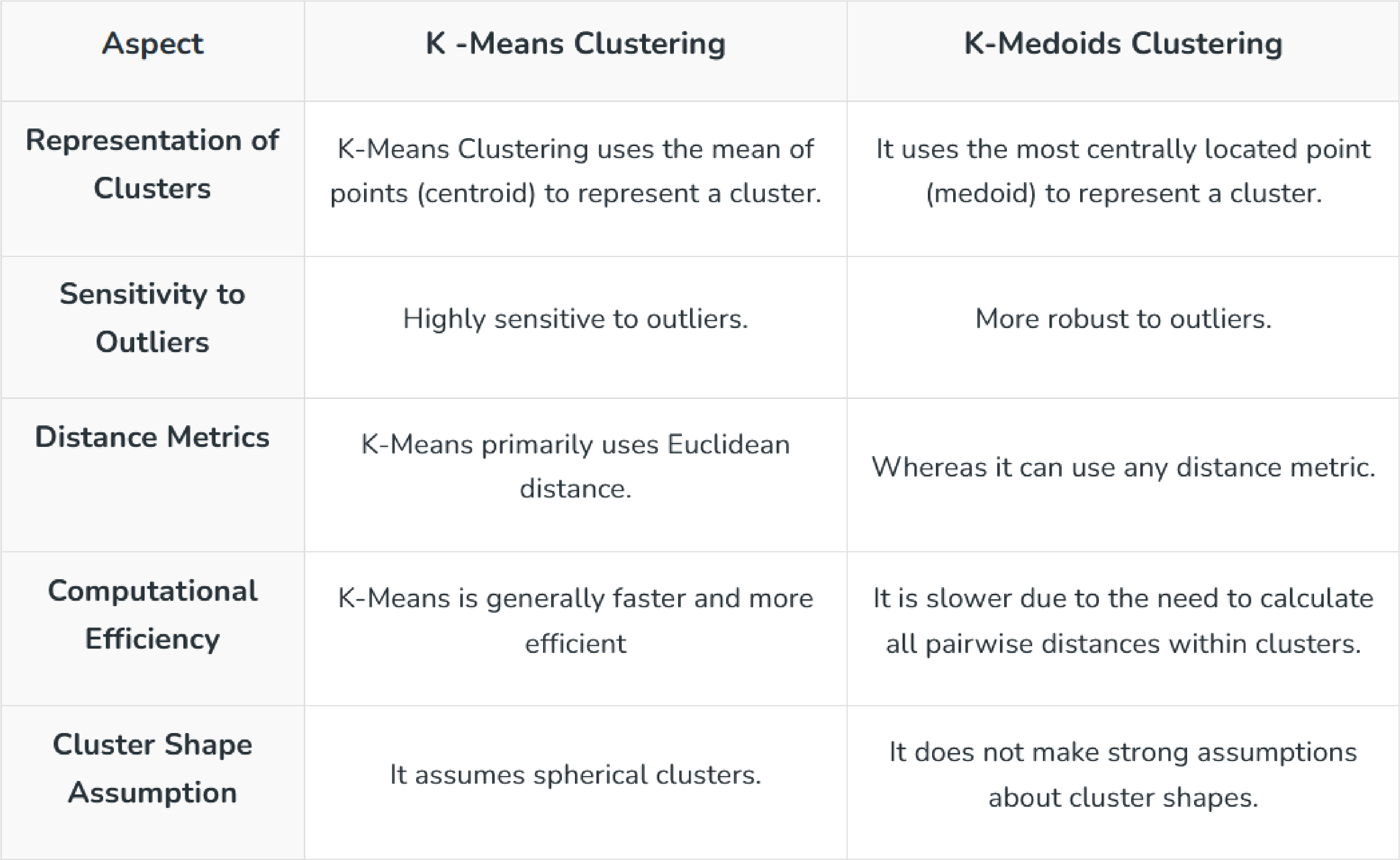
K-means Clusters

- One solution is to use many clusters.
  – Find parts of clusters, but need to put together.

Limitations of K-means: Outlier Problem

- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- <u>Solution</u>: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.





Aspect	K -Means Clustering	K-Medoids Clustering
Representation of Clusters	K-Means Clustering uses the mean of points (centroid) to represent a cluster.	It uses the most centrally located point (medoid) to represent a cluster.
Sensitivity to Outliers	Highly sensitive to outliers.	More robust to outliers.
Distance Metrics	K-Means primarily uses Euclidean distance.	Whereas it can use any distance metric.
Computational Efficiency	K-Means is generally faster and more efficient	It is slower due to the need to calculate all pairwise distances within clusters.
Cluster Shape Assumption	It assumes spherical clusters.	It does not make strong assumptions about cluster shapes.