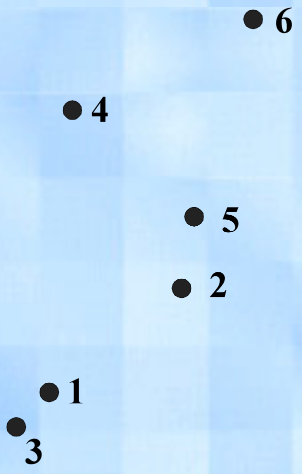
Clustering

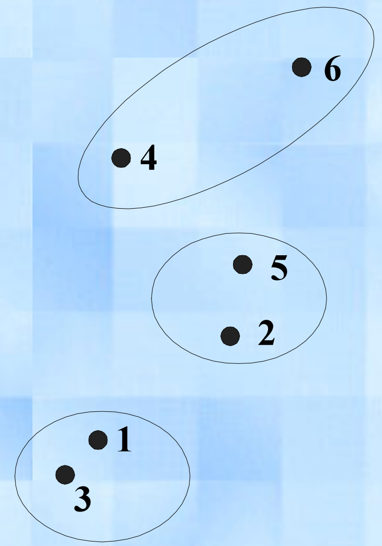
Part 2

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville

Partitional Clustering

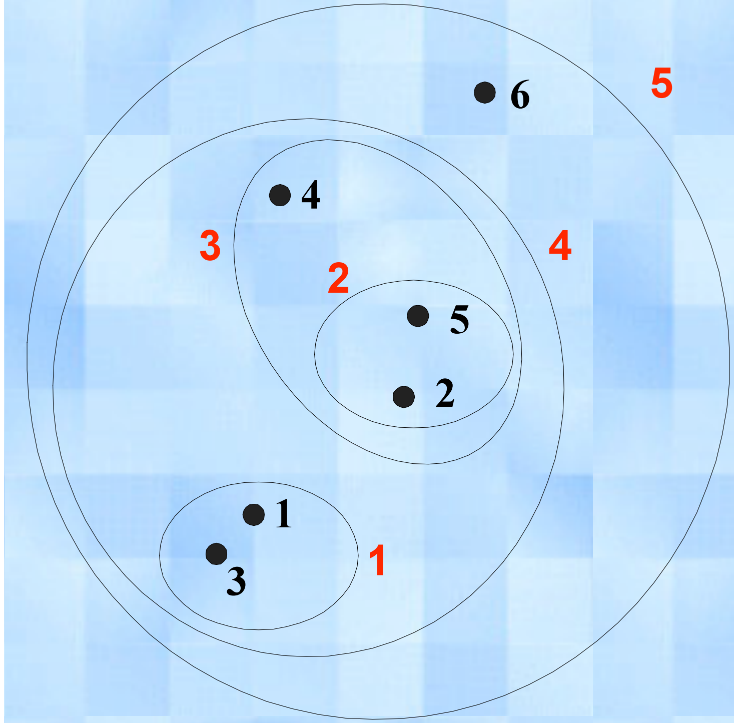


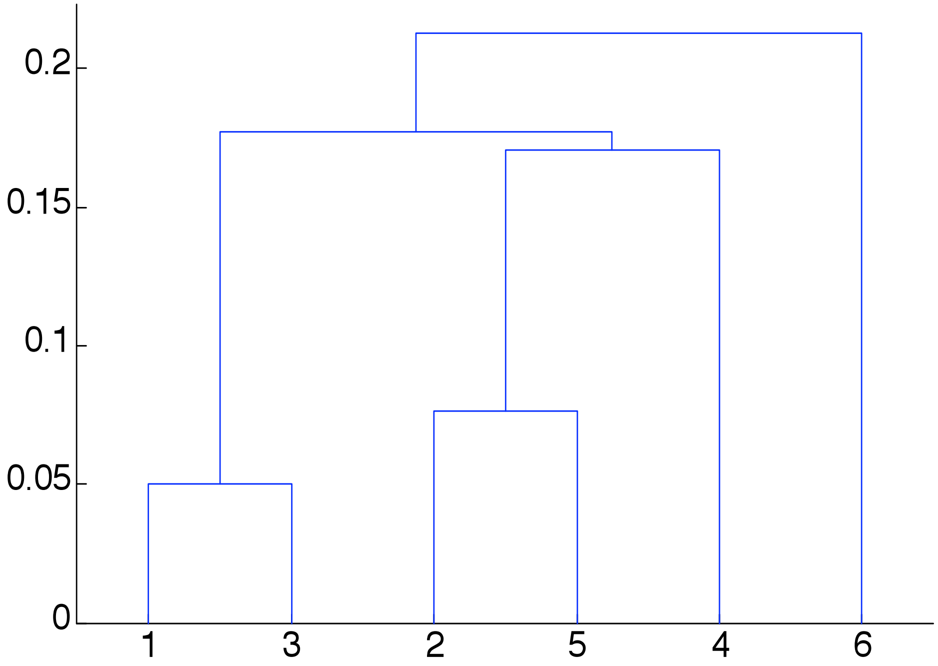
Original Points



A Partitional Clustering

Hierarchical Clustering





Traditional Hierarchical Clustering

Traditional Dendrogram

Characteristics of Clustering Algorithms

- Type of clustering the algorithm produces:
  - Partitional versus hierarchical
  - Overlapping versus non-overlapping
  - Fuzzy versus non-fuzzy
  - Complete versus partial

Characteristics of Clustering Algorithms

- Type of clusters the algorithm seeks:
  - Well-separated, center-based, density-based or contiguity-based
  - Are the clusters found in the entire space or in a subspace
  - Are the clusters relatively similar to one another, or are they of differing sizes, shapes and densities

Characteristics of Clustering Algorithms

- Type of data the algorithm can handle:
  - Some clustering algorithms need a data matrix
    - The K-means algorithm assumes that it is meaningful to take the mean (average) of a set of data objects.
    - This makes sense for data that has continuous attributes and for document data, but not for record data that has categorical attributes.
  - Some clustering algorithms start from a proximity matrix
    - Typically assume symmetry
  - Does the data have noise and outliers?
  - Is the data high dimensional?

Characteristics of Clustering Algorithms

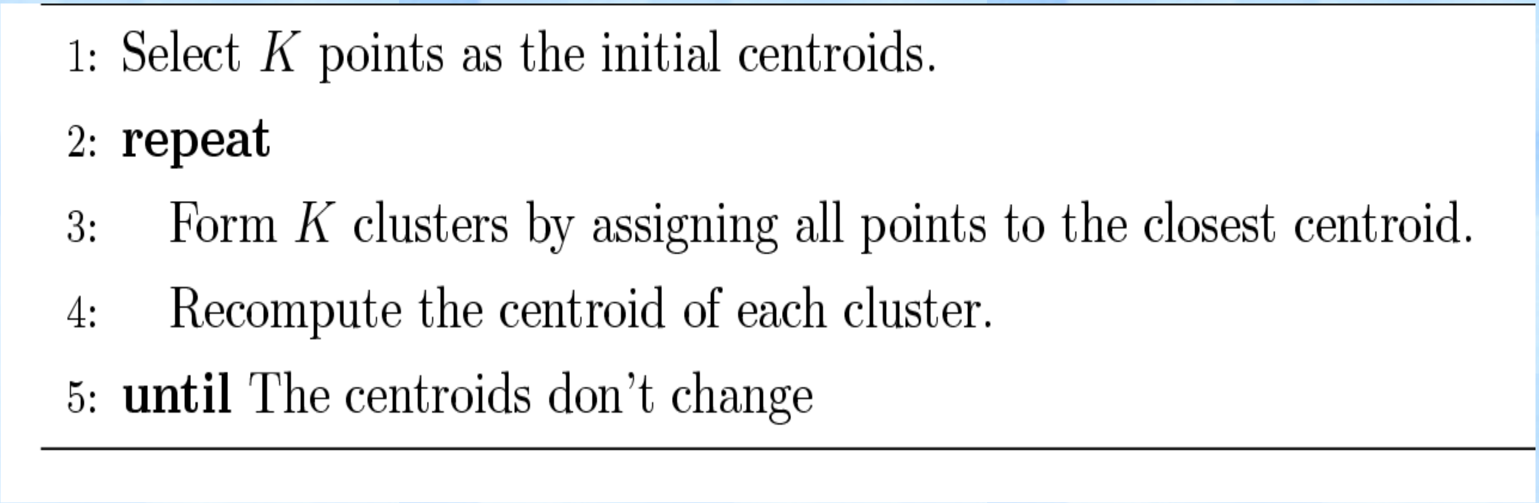
- How the algorithm operates:
  - Minimizing or maximizing a global objective function.
    - Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function. (NP Hard)
    - Can have global or local objectives.
      - Hierarchical clustering algorithms typically have local objectives
      - Partitional algorithms typically have global objectives
  - A variation of the global objective function approach is to fit the data to a parameterized model.
    - Parameters for the model are determined from the data.
    - Mixture models assume that the data is a 'mixture' of a number of statistical distributions.

Characteristics of Clustering Algorithms

- How the algorithm operates …
  - Map the clustering problem to a different domain and solve a related problem in that domain.
    - Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points.
    - Clustering is equivalent to breaking the graph into connected components, one for each cluster.

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.



```
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
```

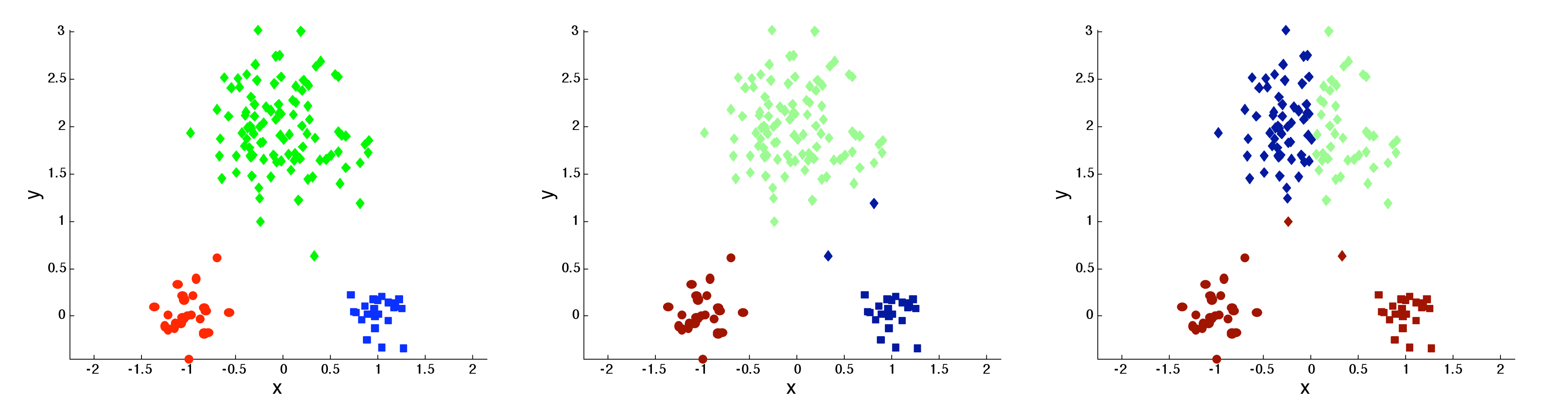
K-means Clustering – Details

- 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

Evaluating K-means Clusters

- Most common measure is the *Sum of the 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.
  - 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.

Two different K-means Clusterings

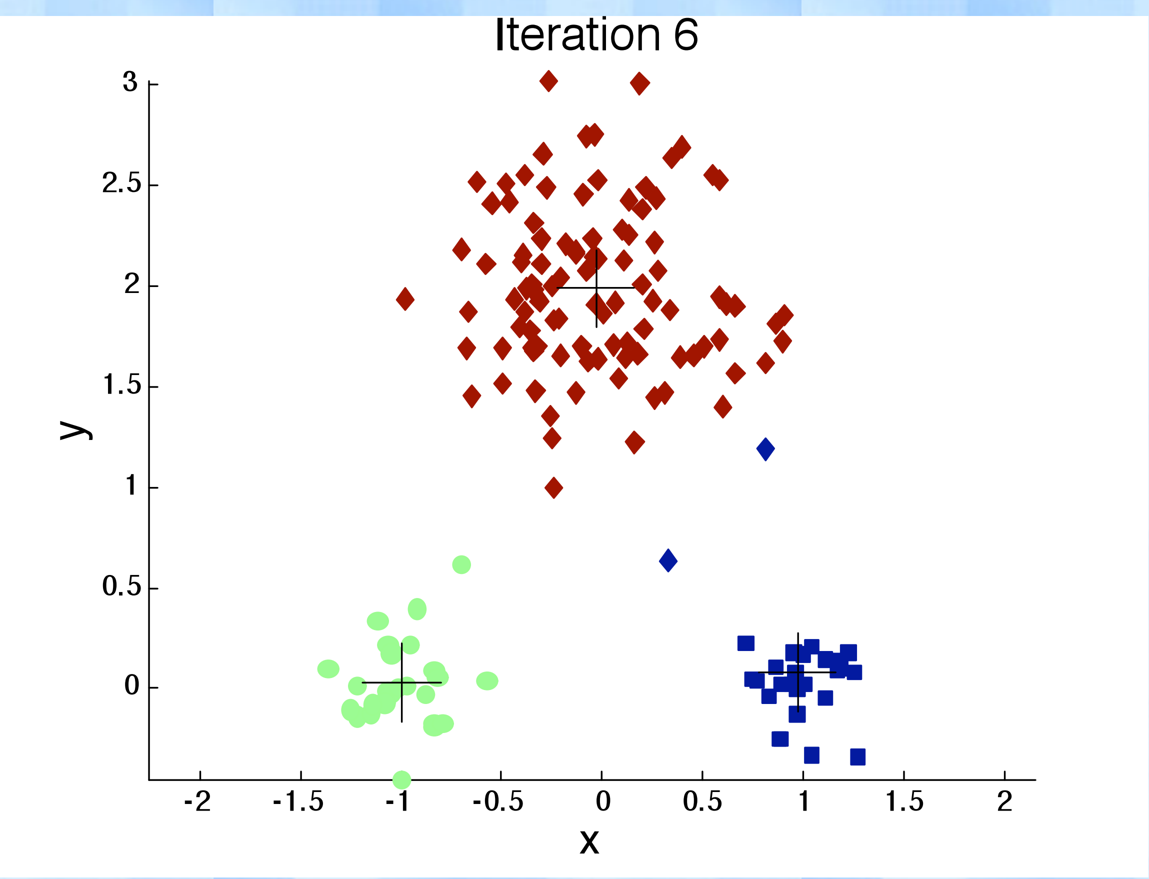


Original Points

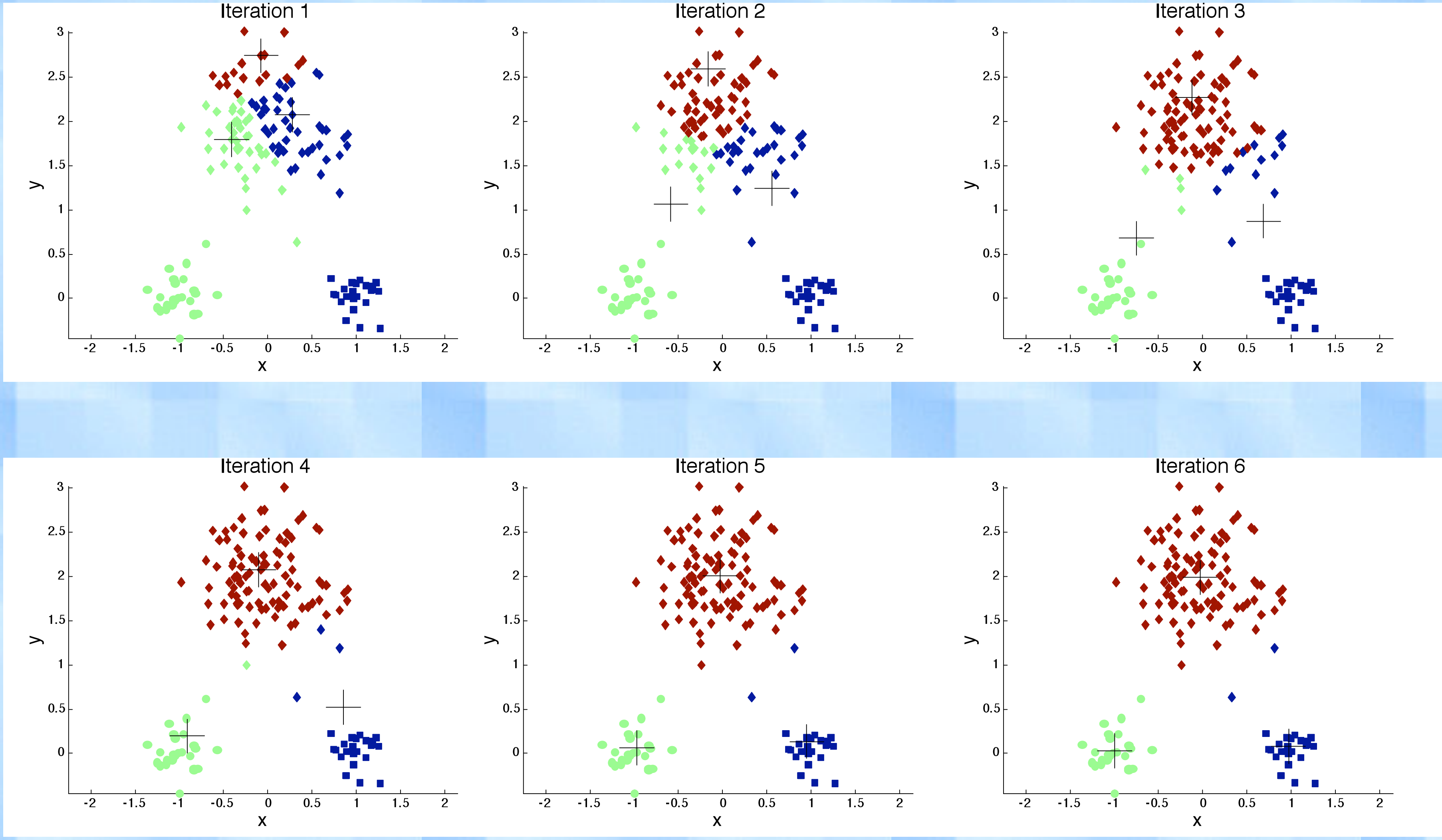
Optimal Clustering

Sub-optimal Clustering

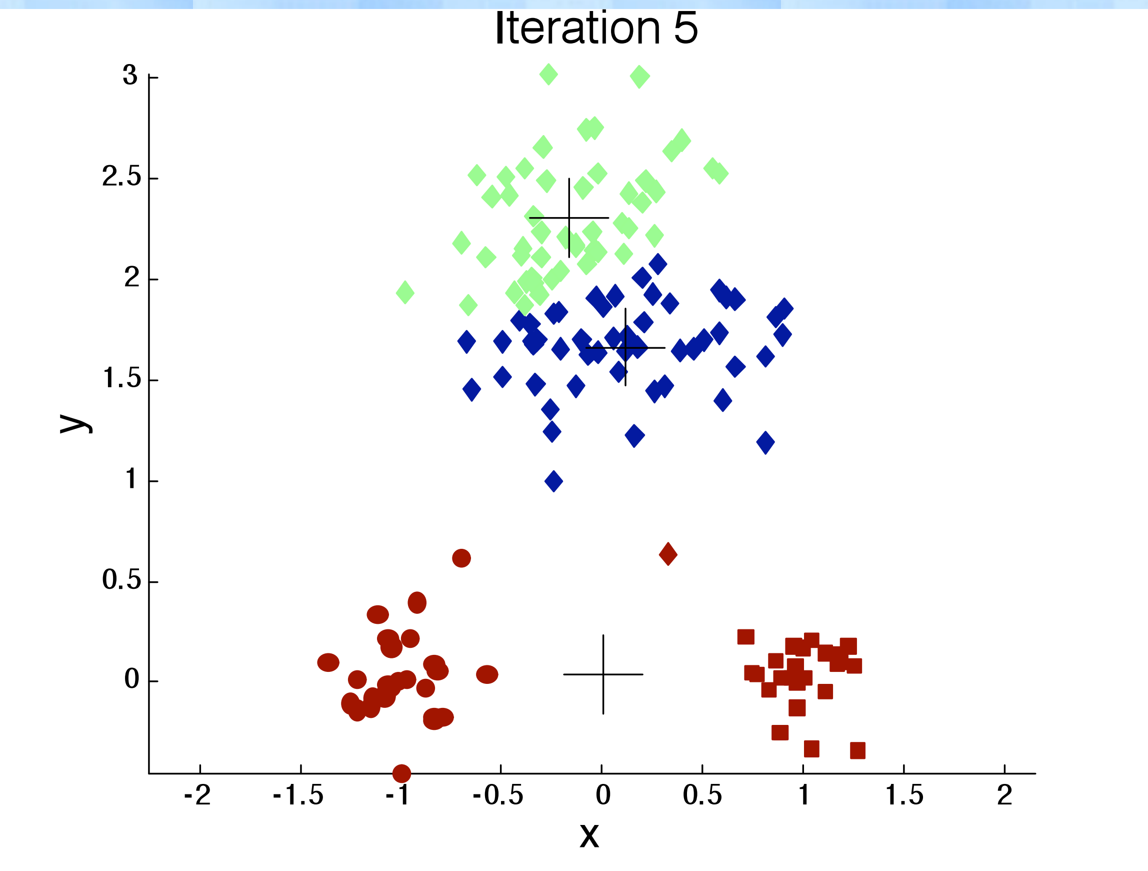
Importance of Choosing Initial Centroids



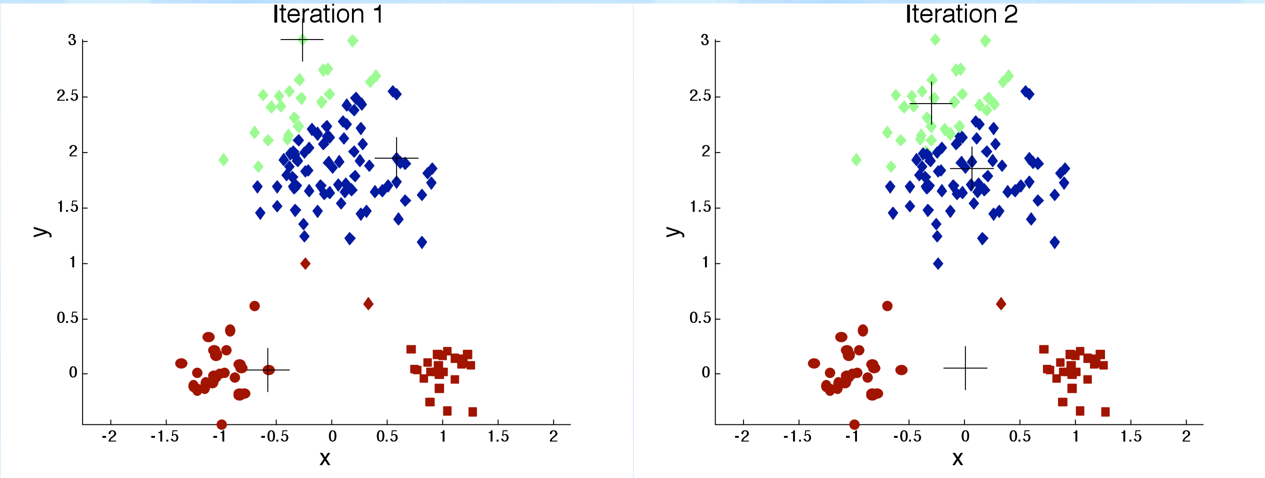
Importance of Choosing Initial Centroids

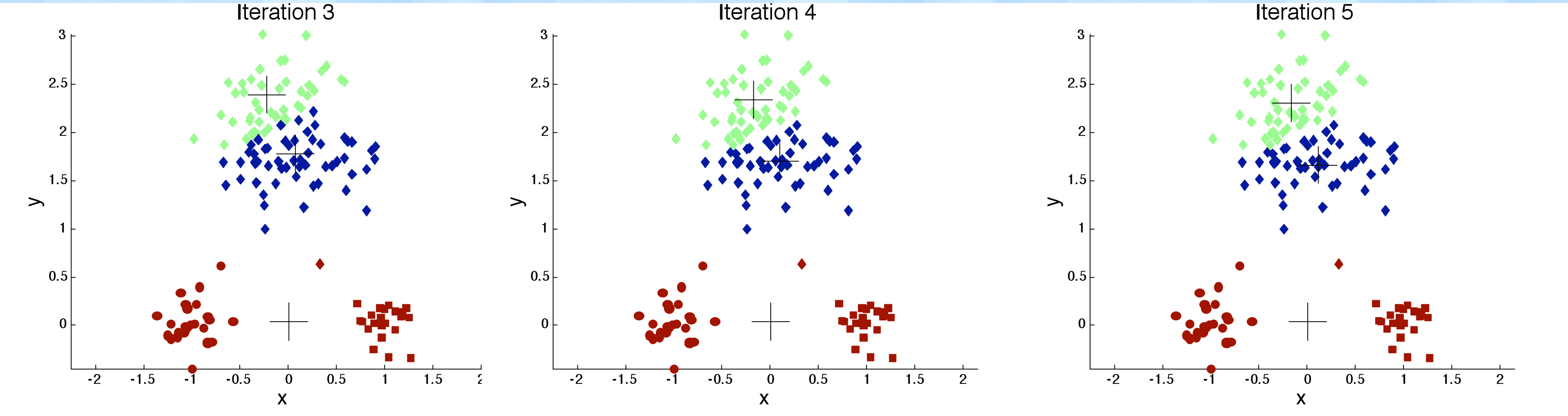


Importance of Choosing Initial Centroids …



Importance of Choosing Initial Centroids …





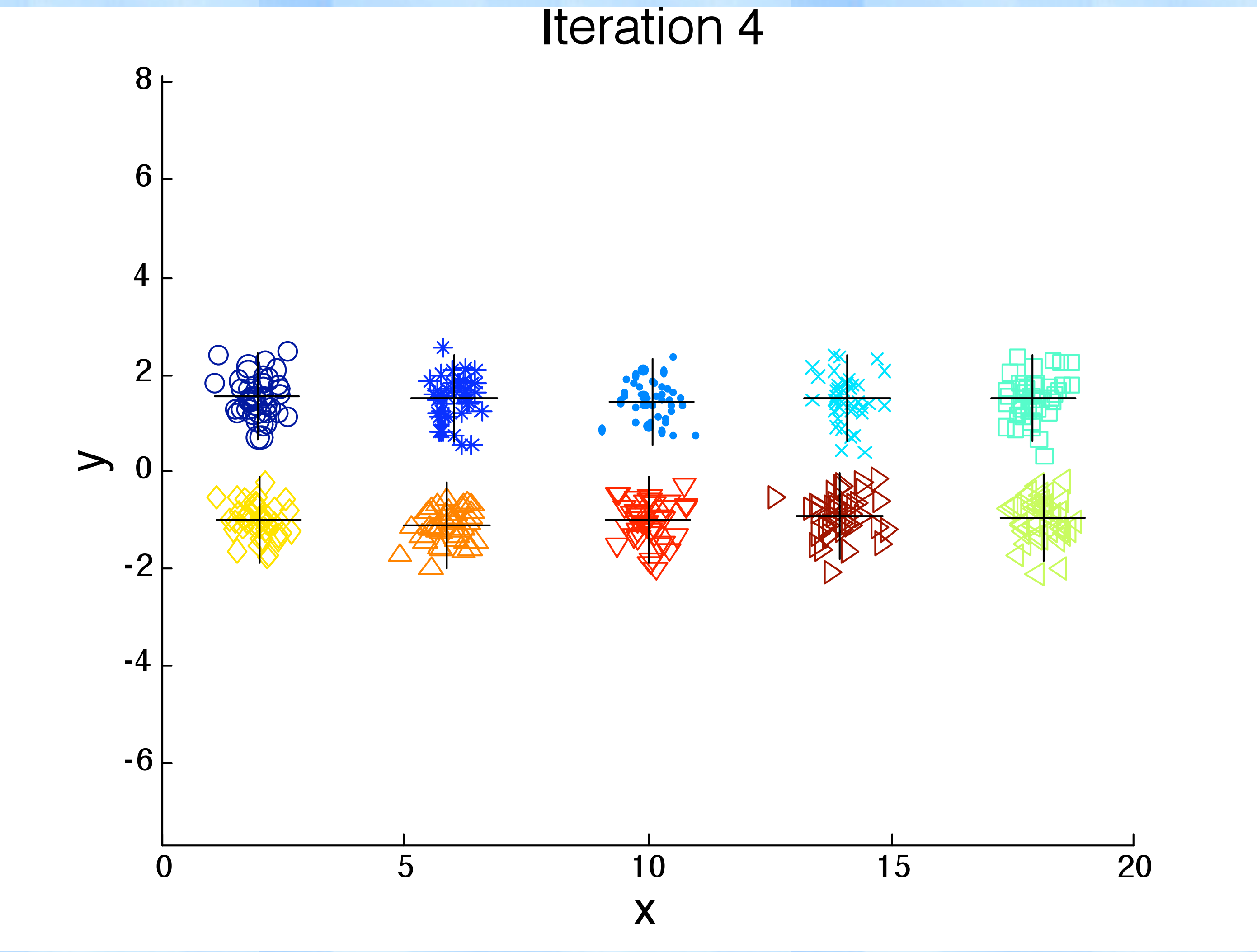
Problems with Selecting Initial Points

- 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

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

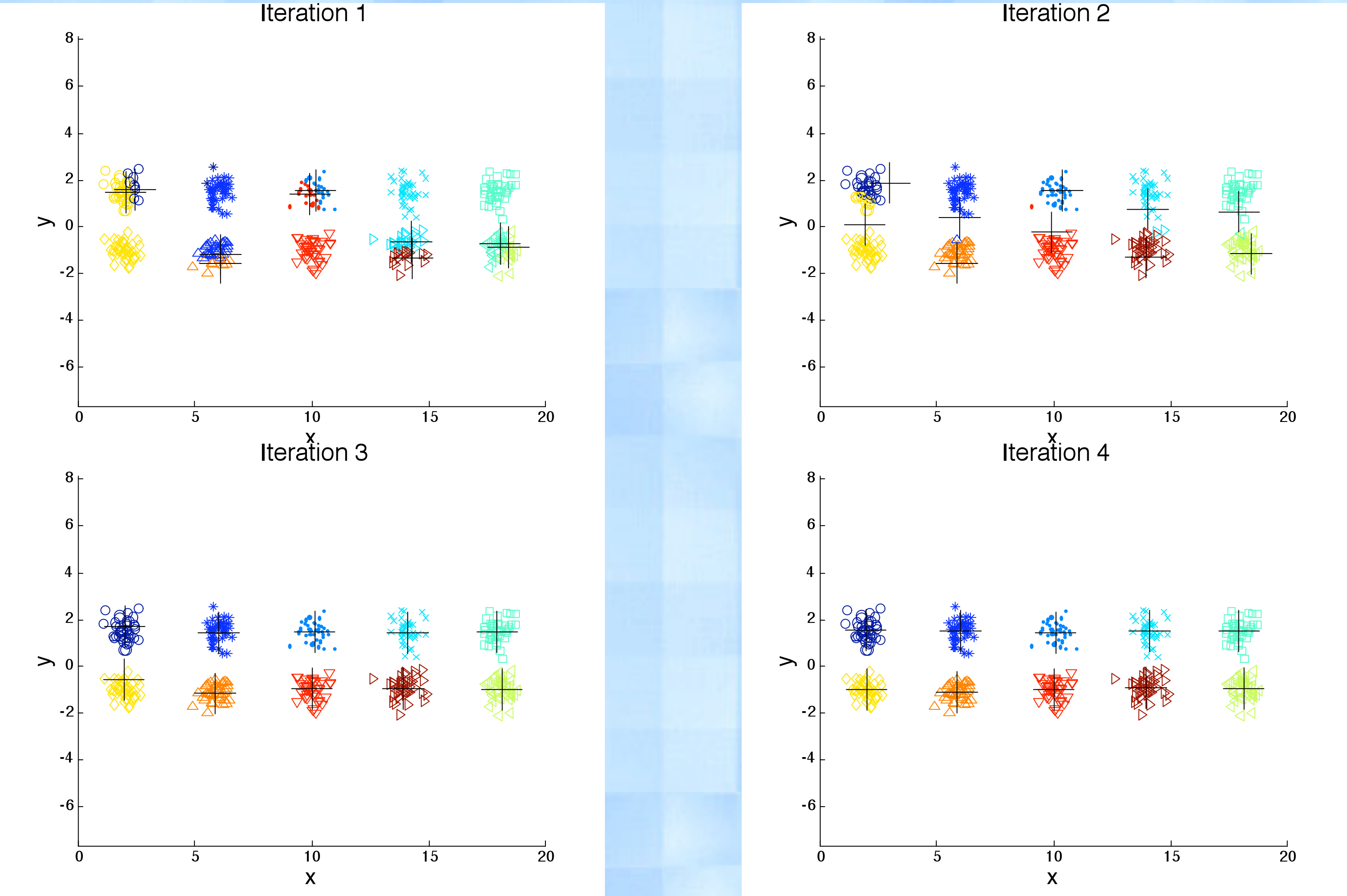
  - For example, if K = 10, then probability = 10!/1010 = 0.00036
  - Sometimes the initial centroids will readjust themselves in ‘right’ way, and sometimes they don’t
  - Consider an example of five pairs of clusters

10 Clusters Example



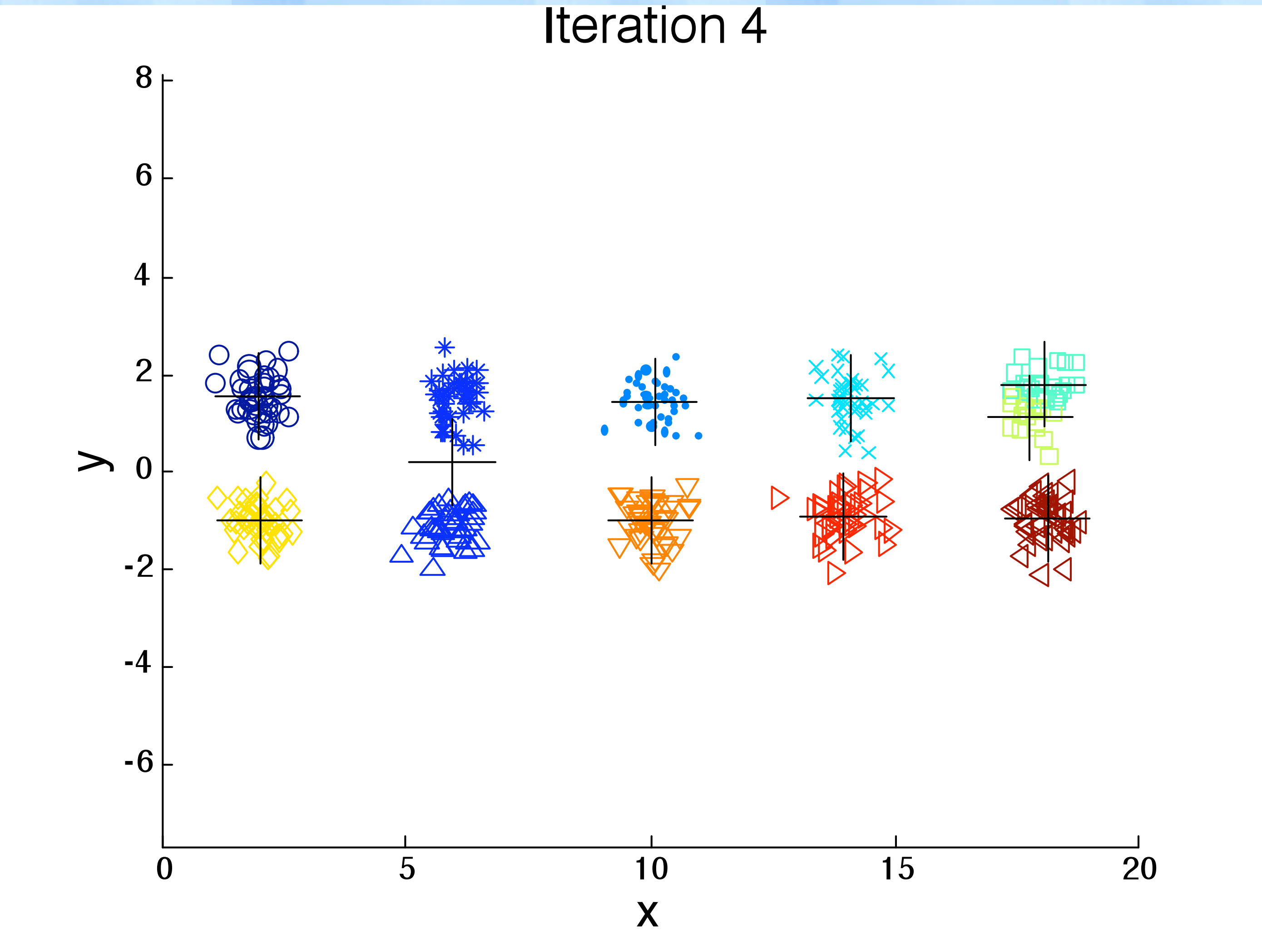
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



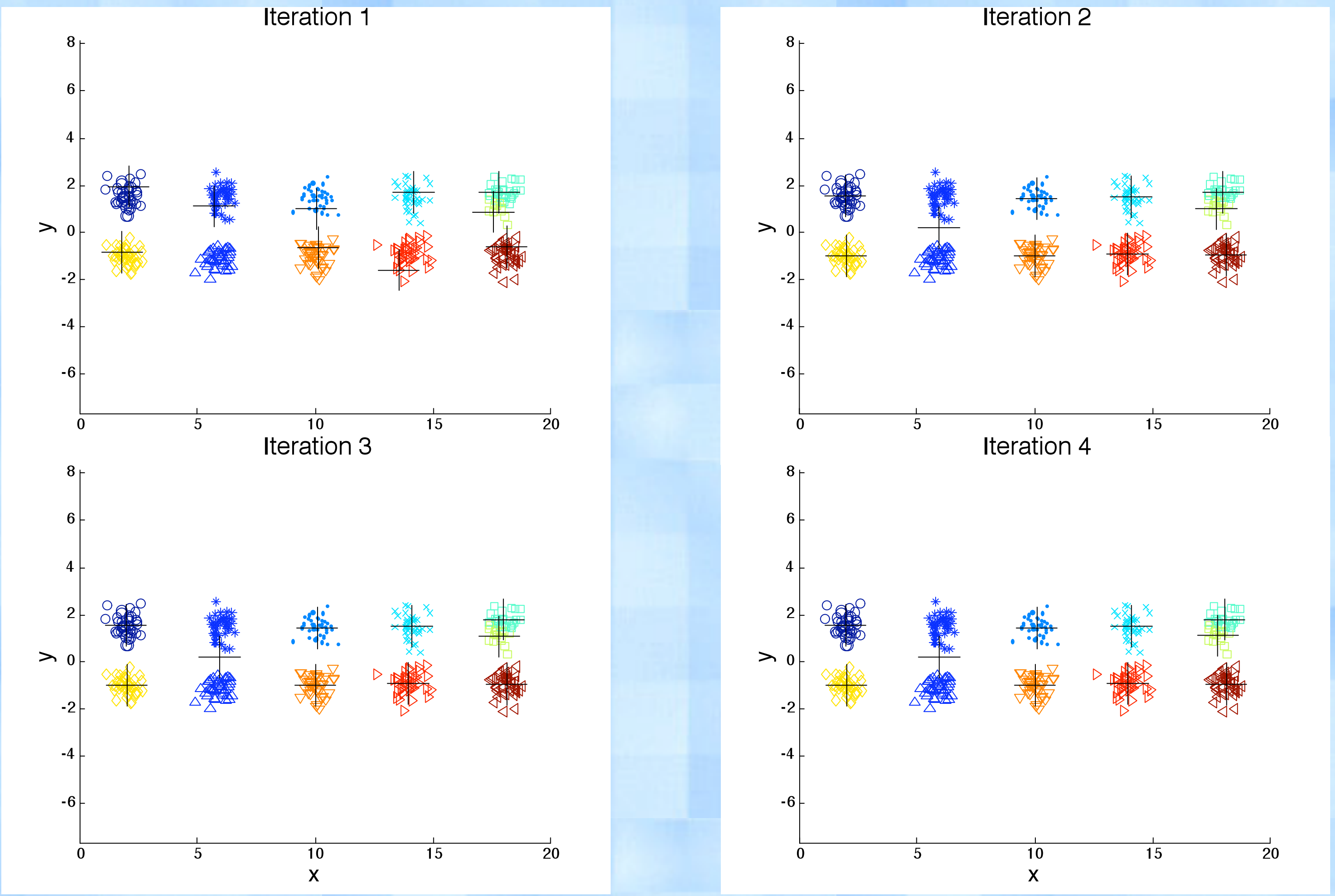
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



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

10 Clusters Example



Iteration 1

Iteration 2

Iteration 3

Iteration 4

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
- Bisecting K-means
  - Not as susceptible to initialization issues
- 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

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

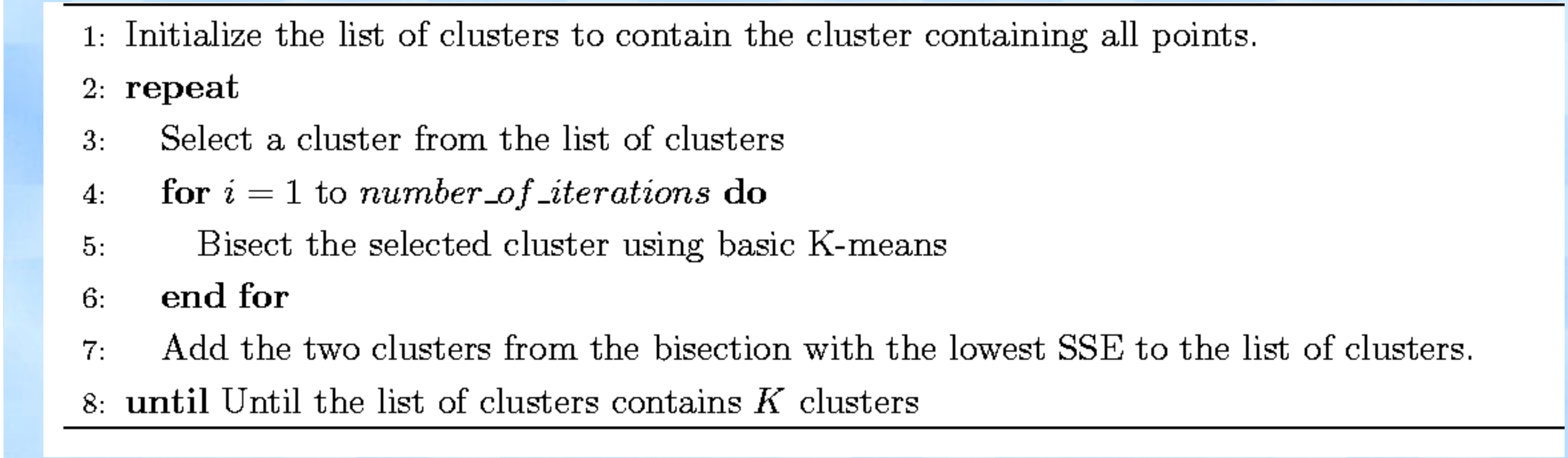
- In the basic K-means algorithm, centroids are updated after all points are assigned to a centroid.
- An alternative is to update the centroids after each assignment.
  - May need to update two centroids
  - More expensive
  - Introduces an order dependency
  - Never get an empty cluster
  - Can use “weights” to change the impact

Pre-processing and Post-processing

- Pre-processing
  - Normalize data so distance computations are fast.
  - 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

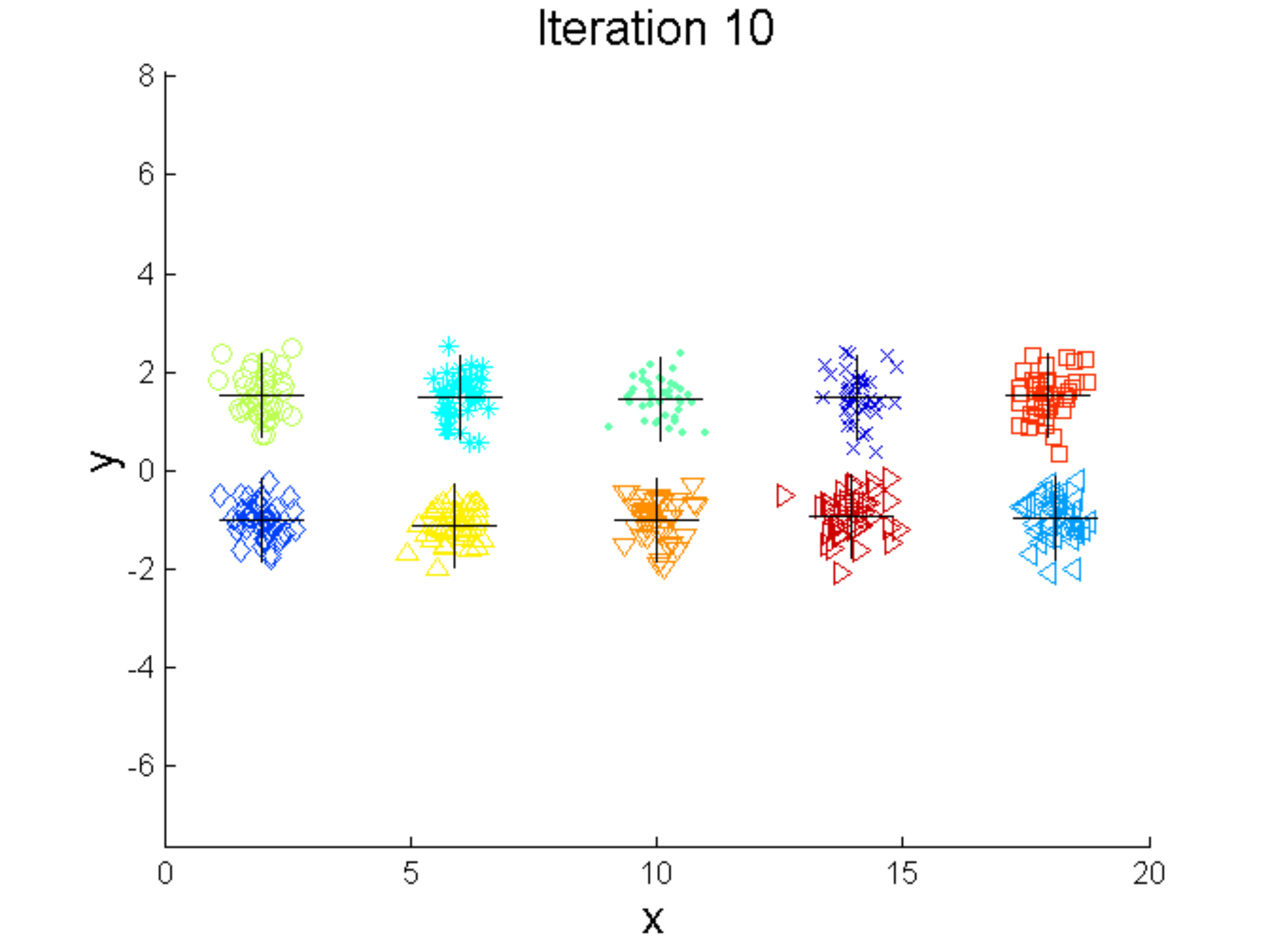
Bisecting K-means

- 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
```

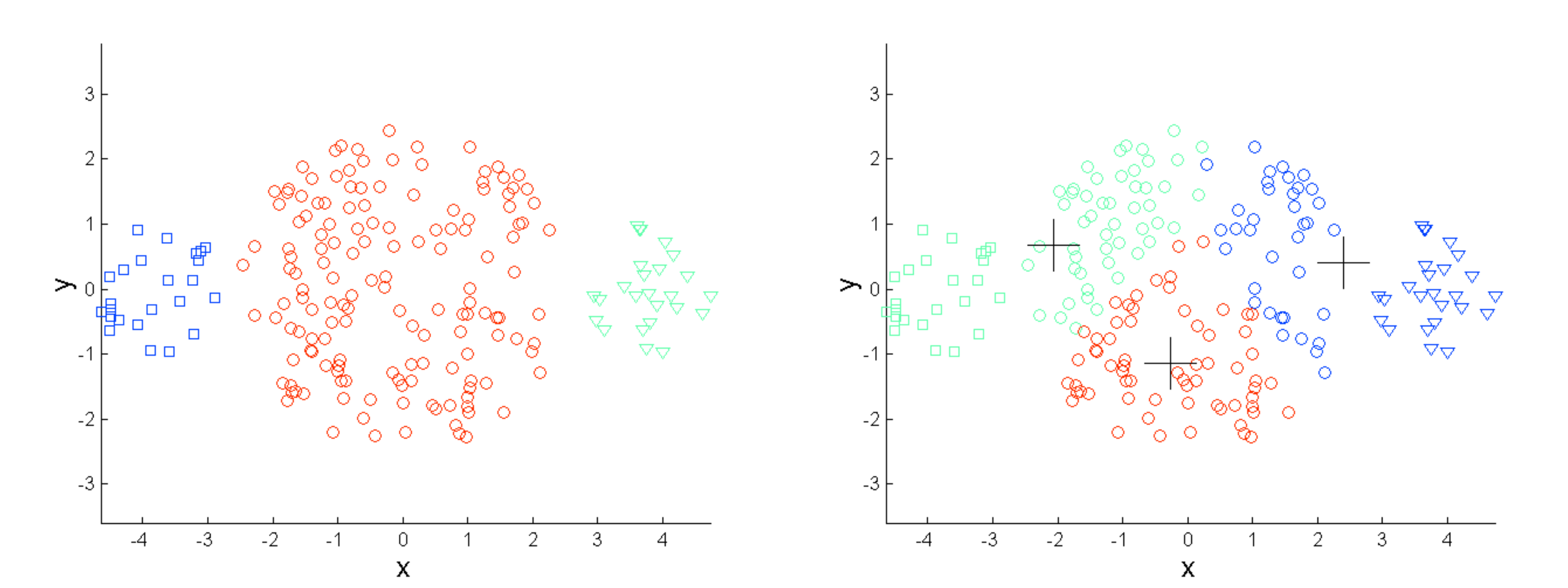
Bisecting K-means Example



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

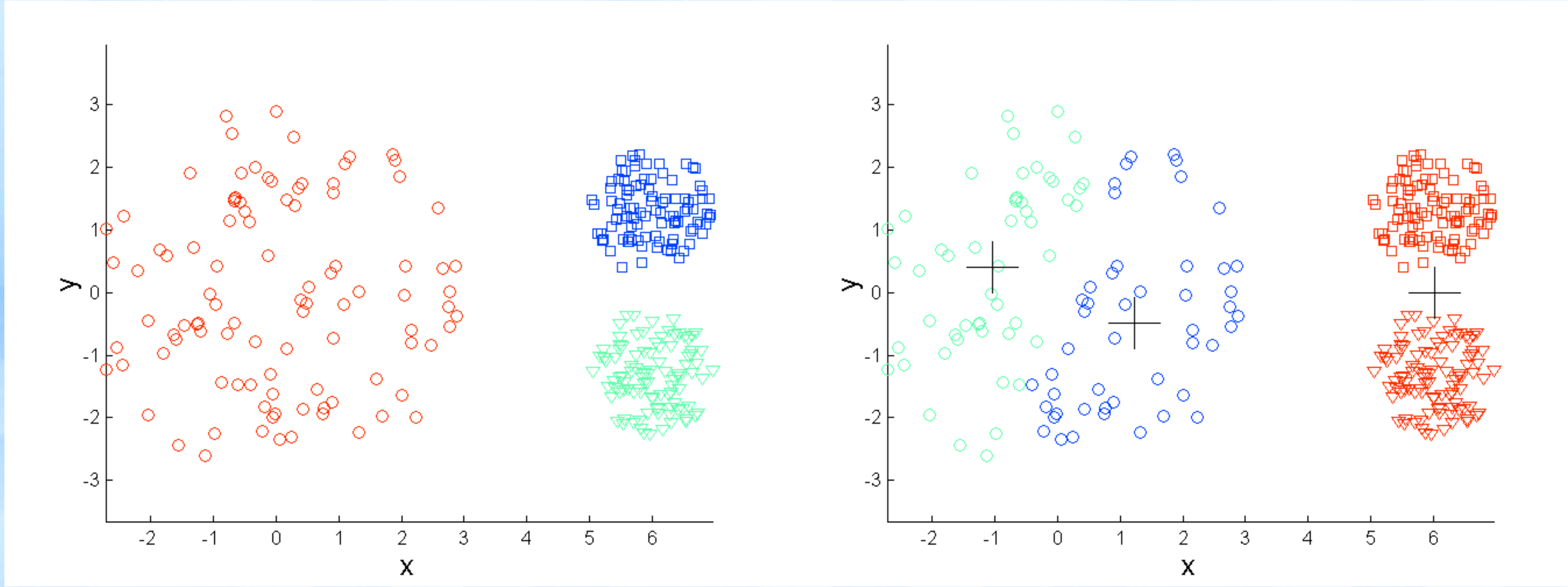
Limitations of K-means: Differing Sizes



Original Points

K-means Clusters

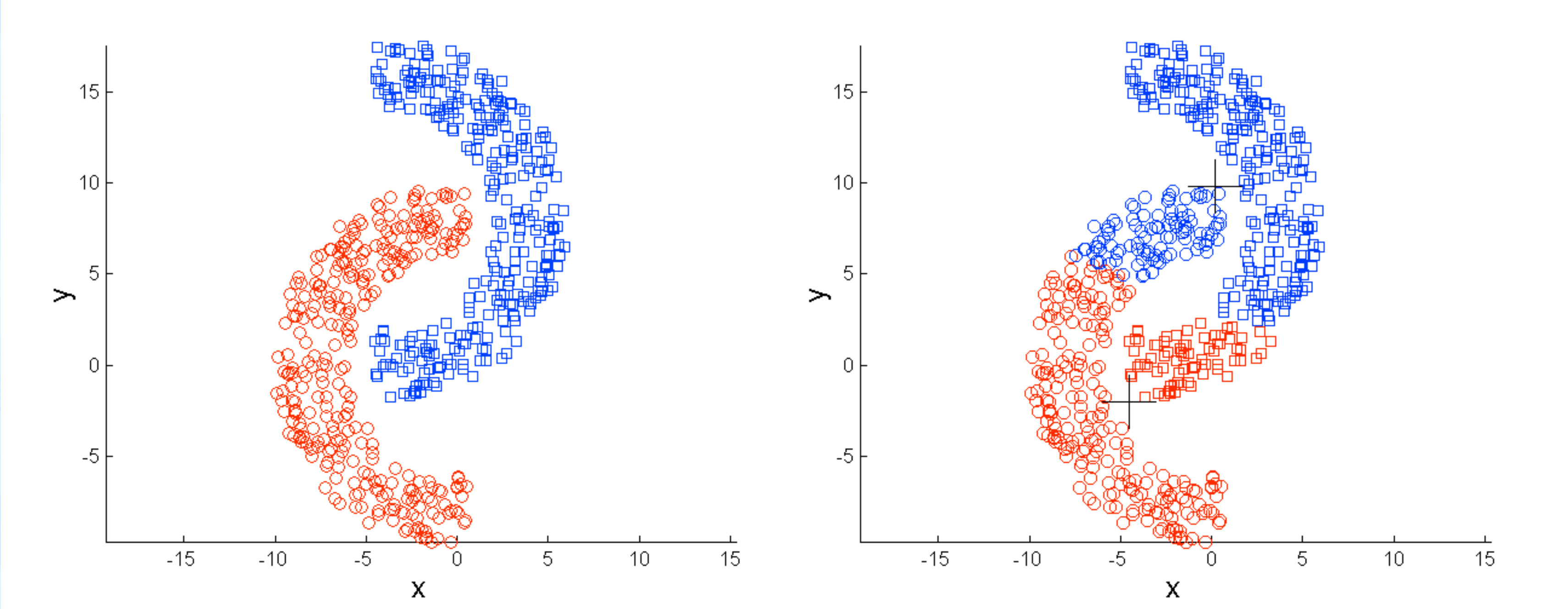
Limitations of K-means: Differing Density



Original Points

K-means Clusters

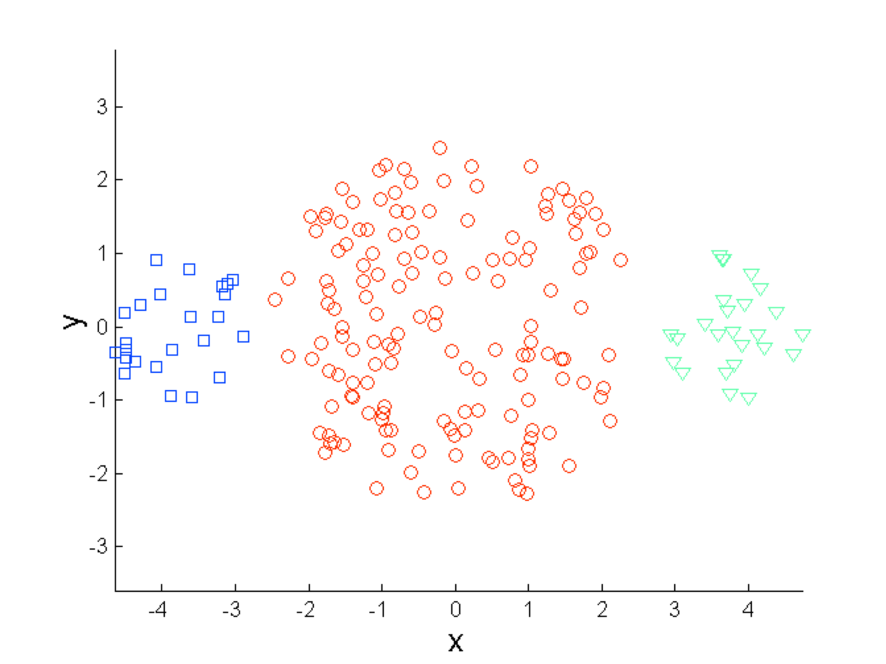
Limitations of K-means: Non-globular Shapes

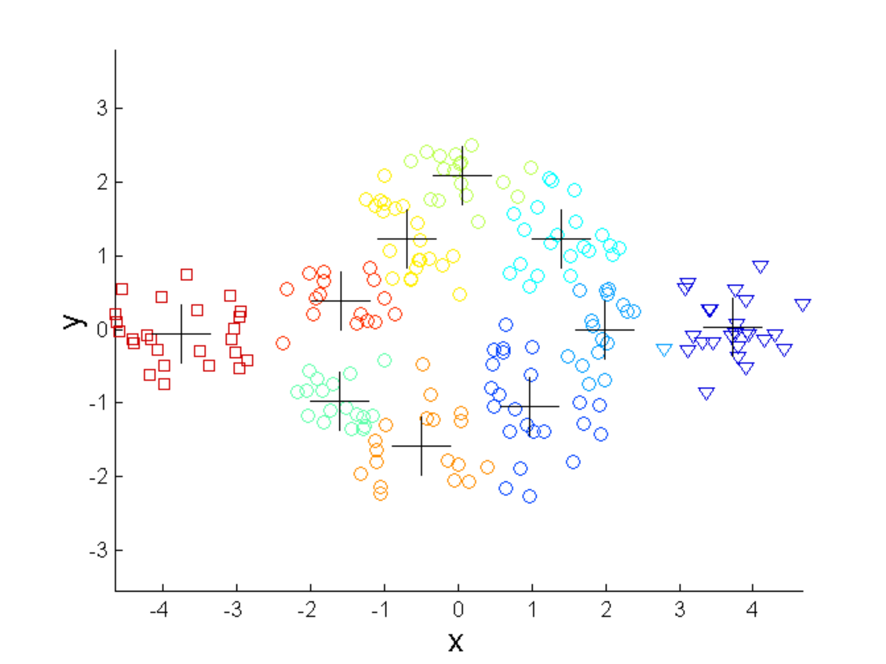


Original Points

K-means Clusters

Overcoming K-means Limitations

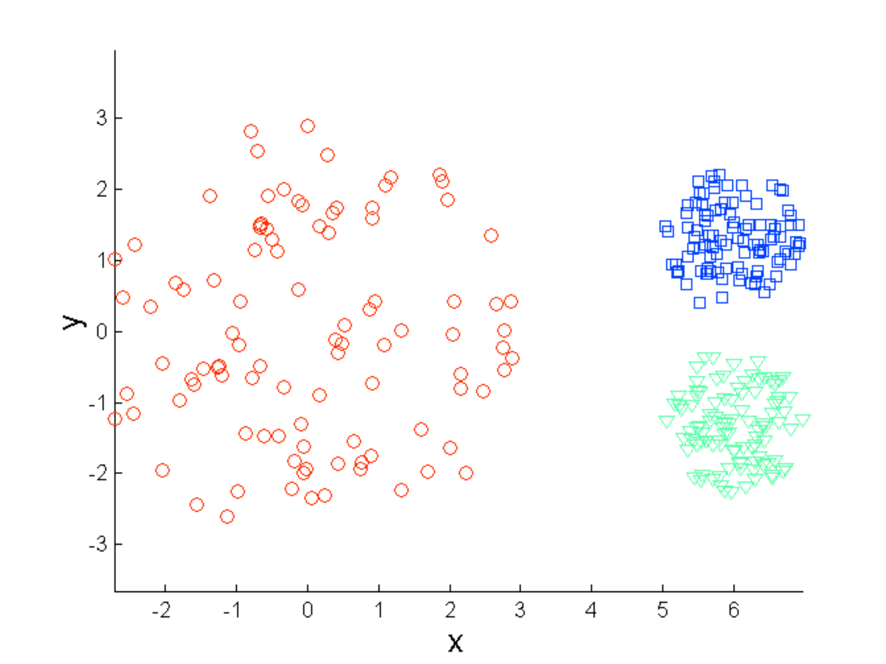




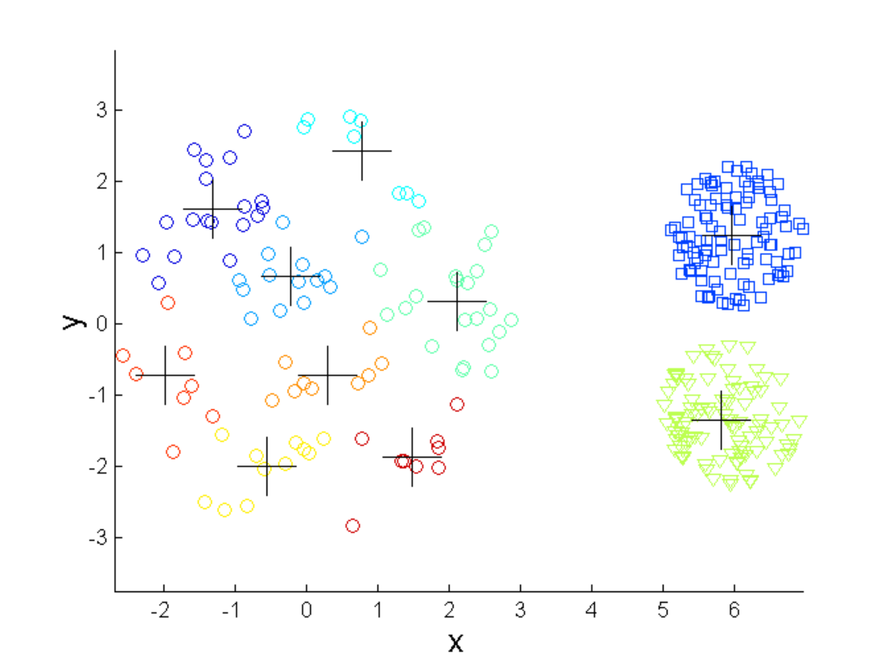
Original Points

K-means Clusters

Overcoming K-means Limitations

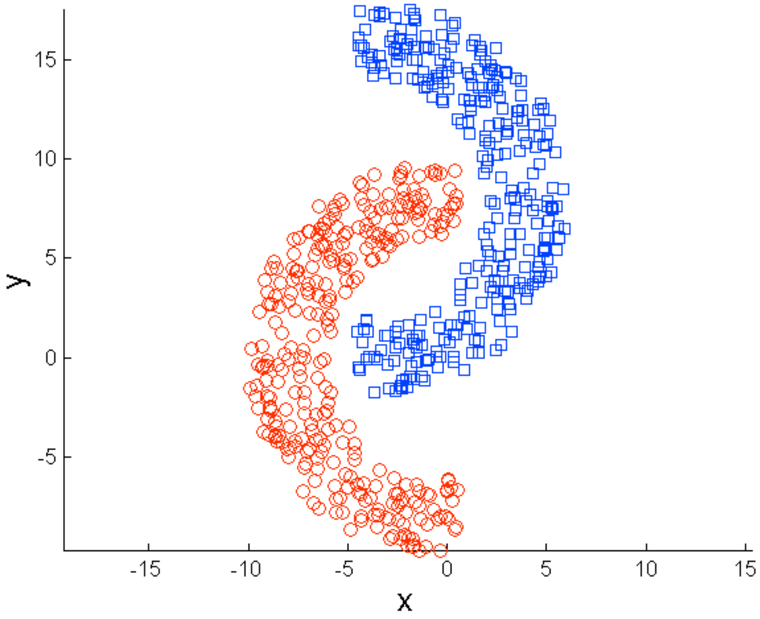


Original Points

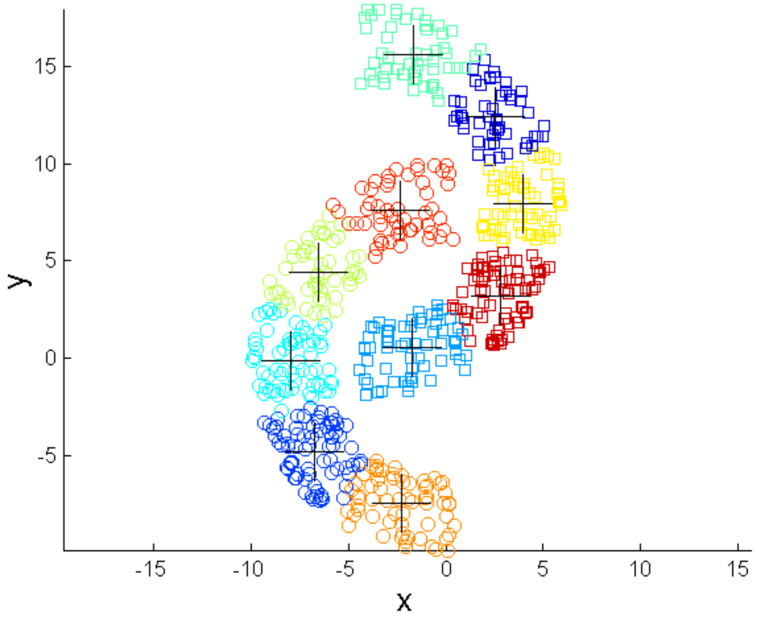


K-means Clusters

Overcoming K-means Limitations



Original Points



K-means Clusters