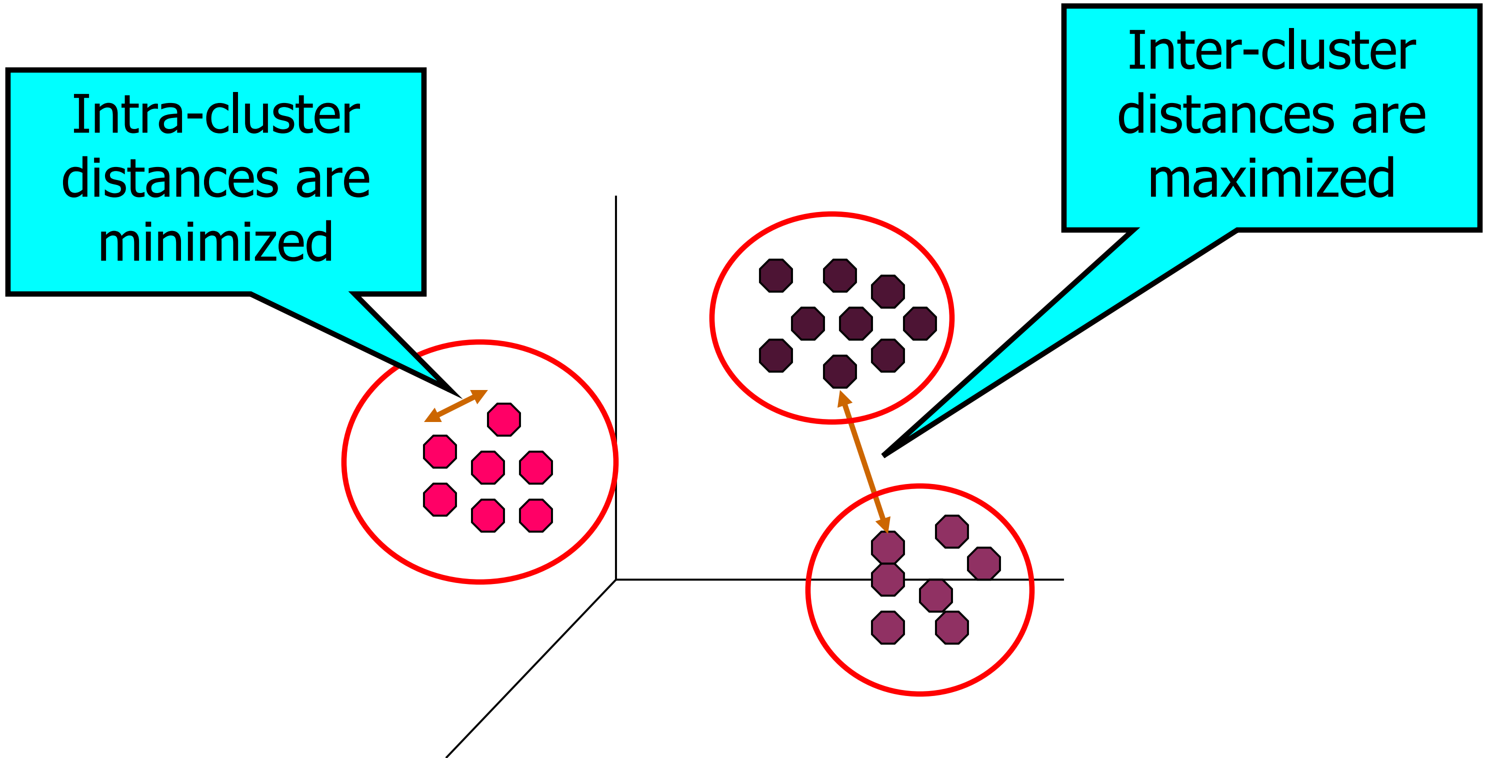
CLUSTERING

WHAT IS CLUSTER ANALYSIS?

- Given a set of objects, place them in groups such that:

  the objects in a group are similar (or related)

  different from (or unrelated to) the objects in other groups



CLUSTERING ALGORITHMS

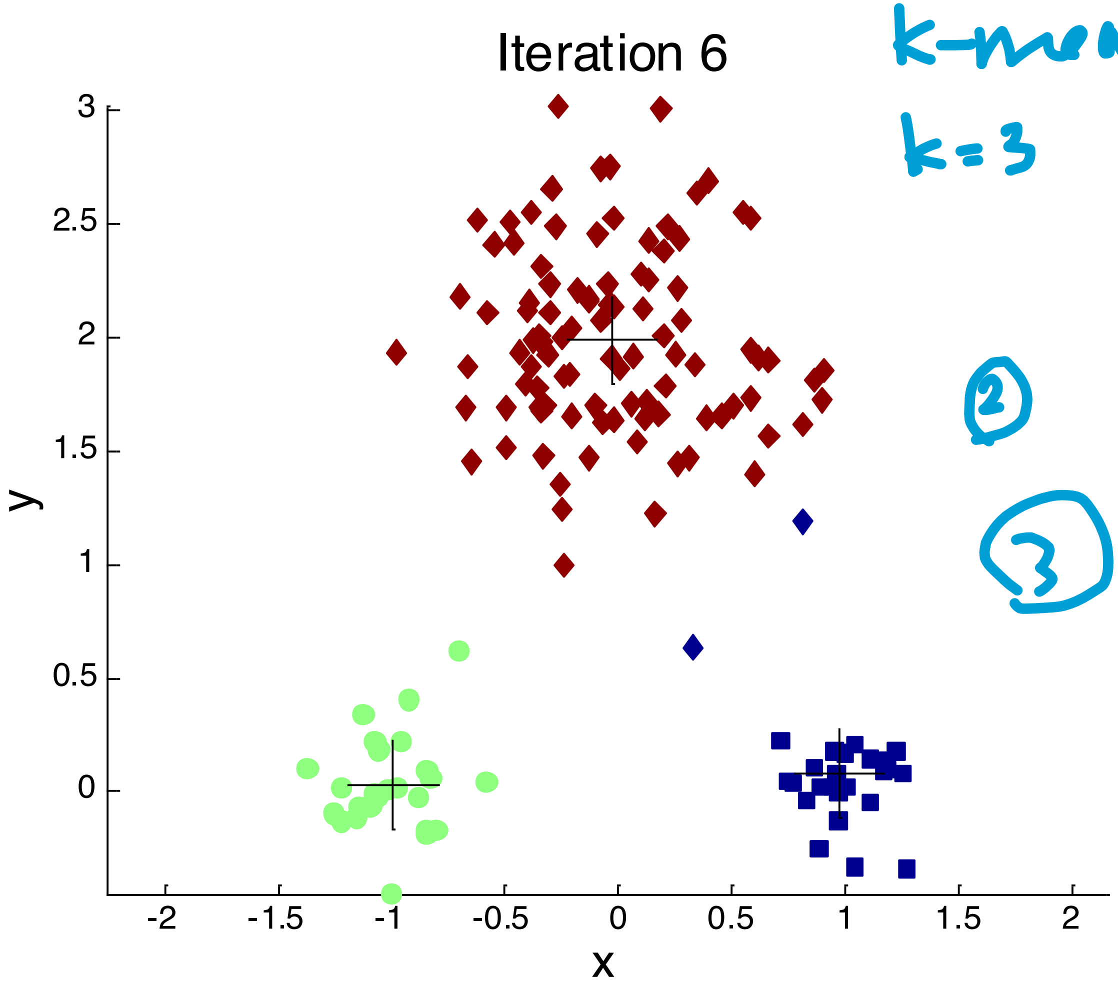
- K-means and its variants
- Hierarchical clustering
- Density-based clustering

K-MEANS CLUSTERING

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

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

EXAMPLE OF K-MEANS CLUSTERING



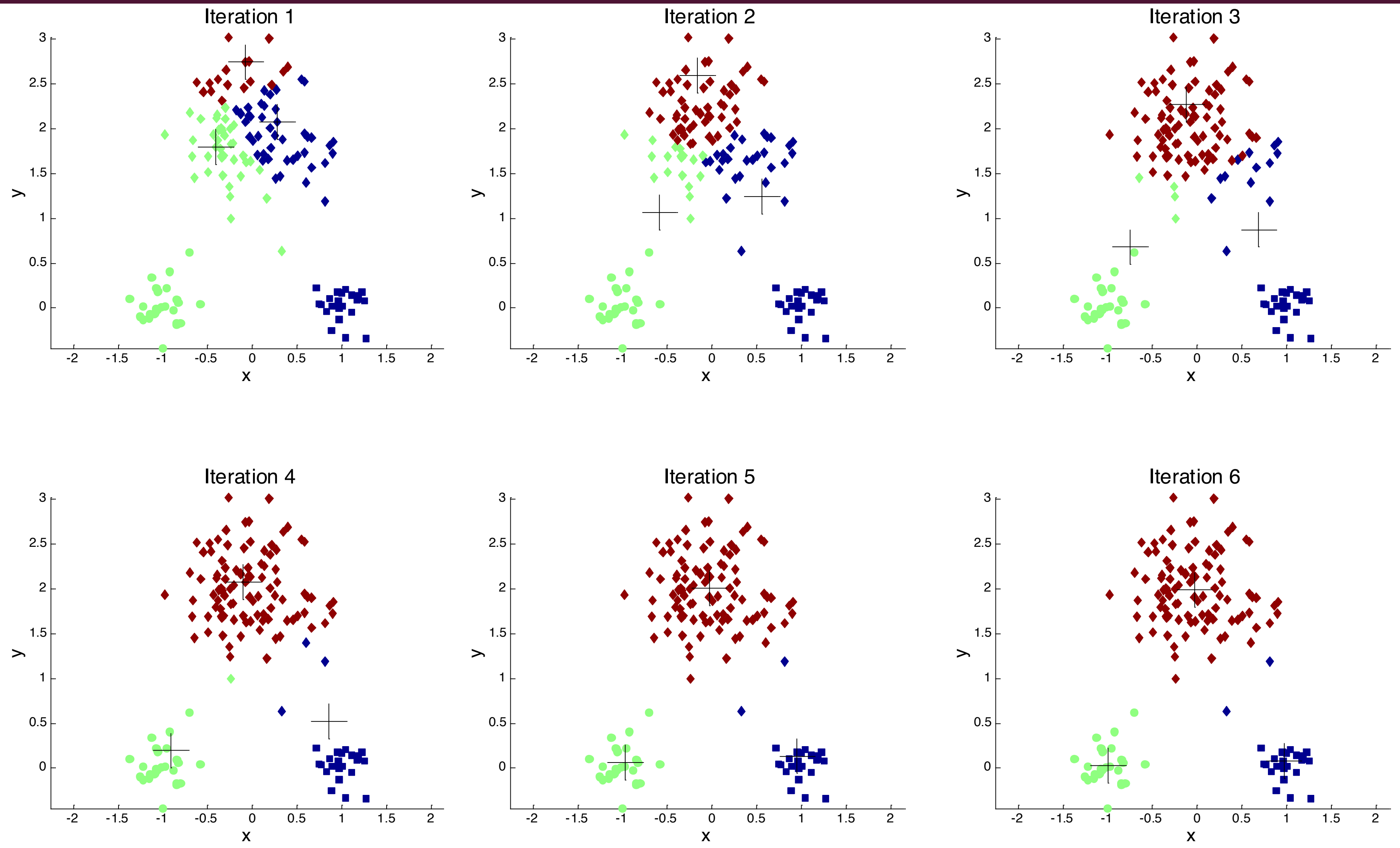
k-means:

k=3 ① randomly 3 centroids

② mind } loop

③

EXAMPLE OF K-MEANS CLUSTERING



K-MEANS CLUSTERING – DETAILS

- Simple iterative algorithm.
  - Choose initial centroids;
  - repeat {assign each point to a nearest centroid; re-compute cluster centroids}
  - until centroids stop changing.
- Initial centroids are often chosen randomly.
  - Clusters produced can vary from one run to another
- The centroid is (typically) the mean of the points in the cluster, but other definitions are possible
- K-means will converge for common proximity measures with appropriately defined centroid
- 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

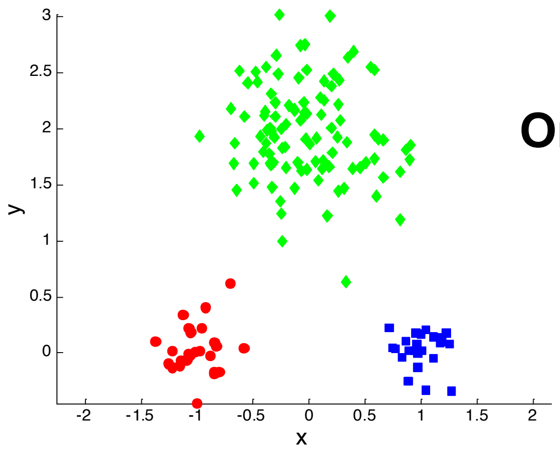
K-MEANS OBJECTIVE FUNCTION

- A common objective function (used with Euclidean distance measure) is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster center
  - To get SSE, we square these errors and sum them.

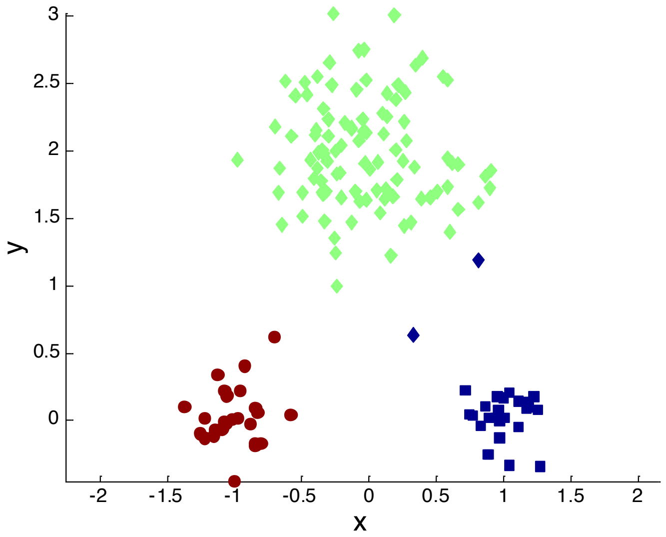
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the centroid (mean) for cluster C_i
  - SSE improves in each iteration of K-means until it reaches a local or global minima.

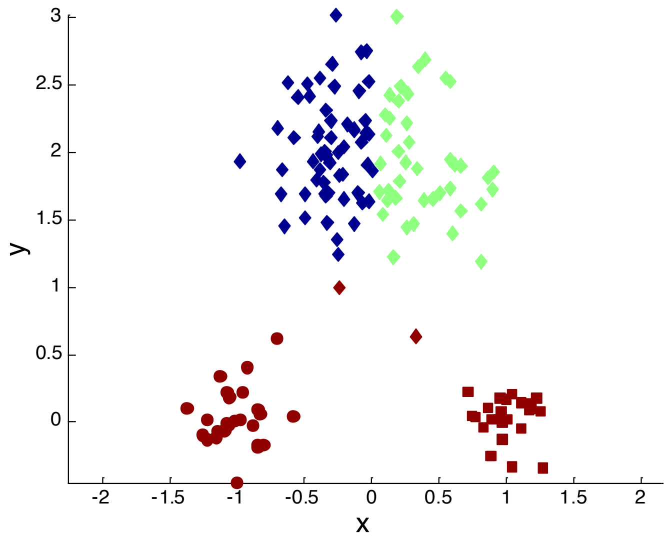
TWO DIFFERENT K-MEANS CLUSTERING



Original Points

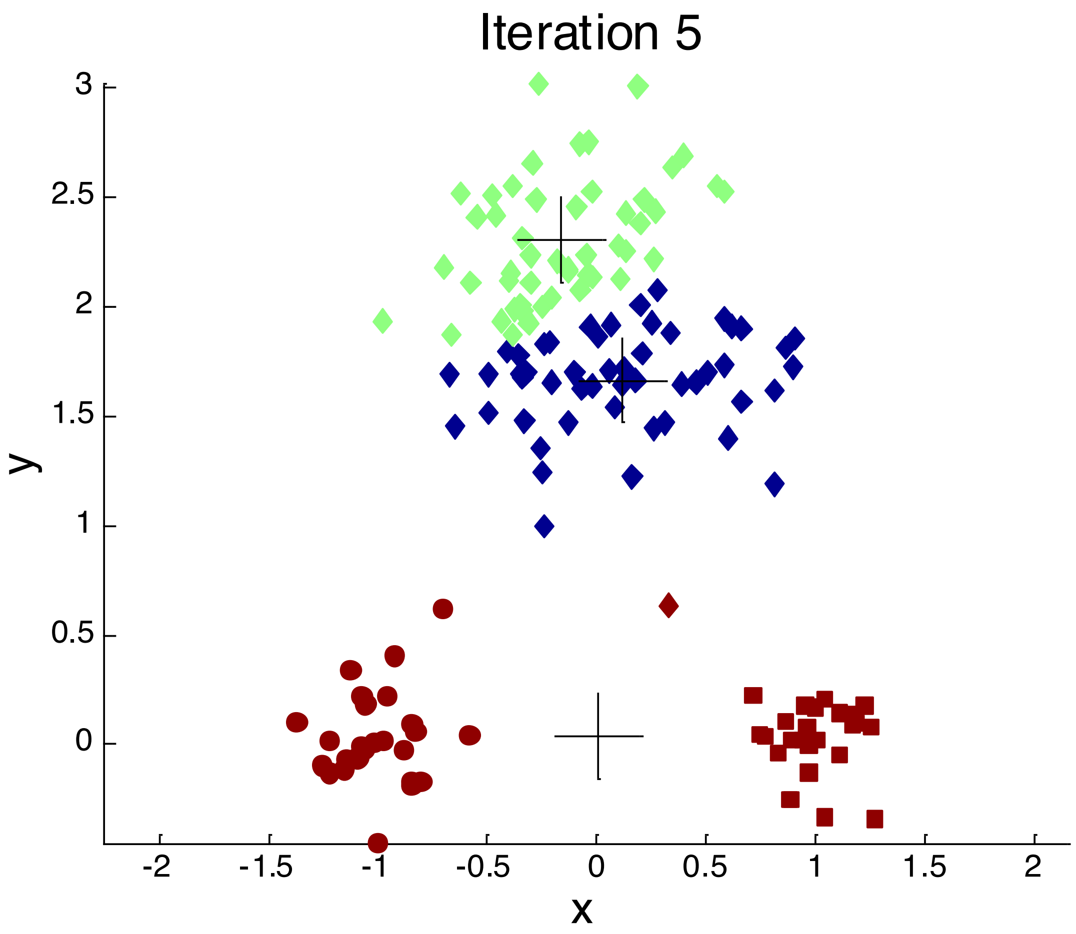


Optimal Clustering

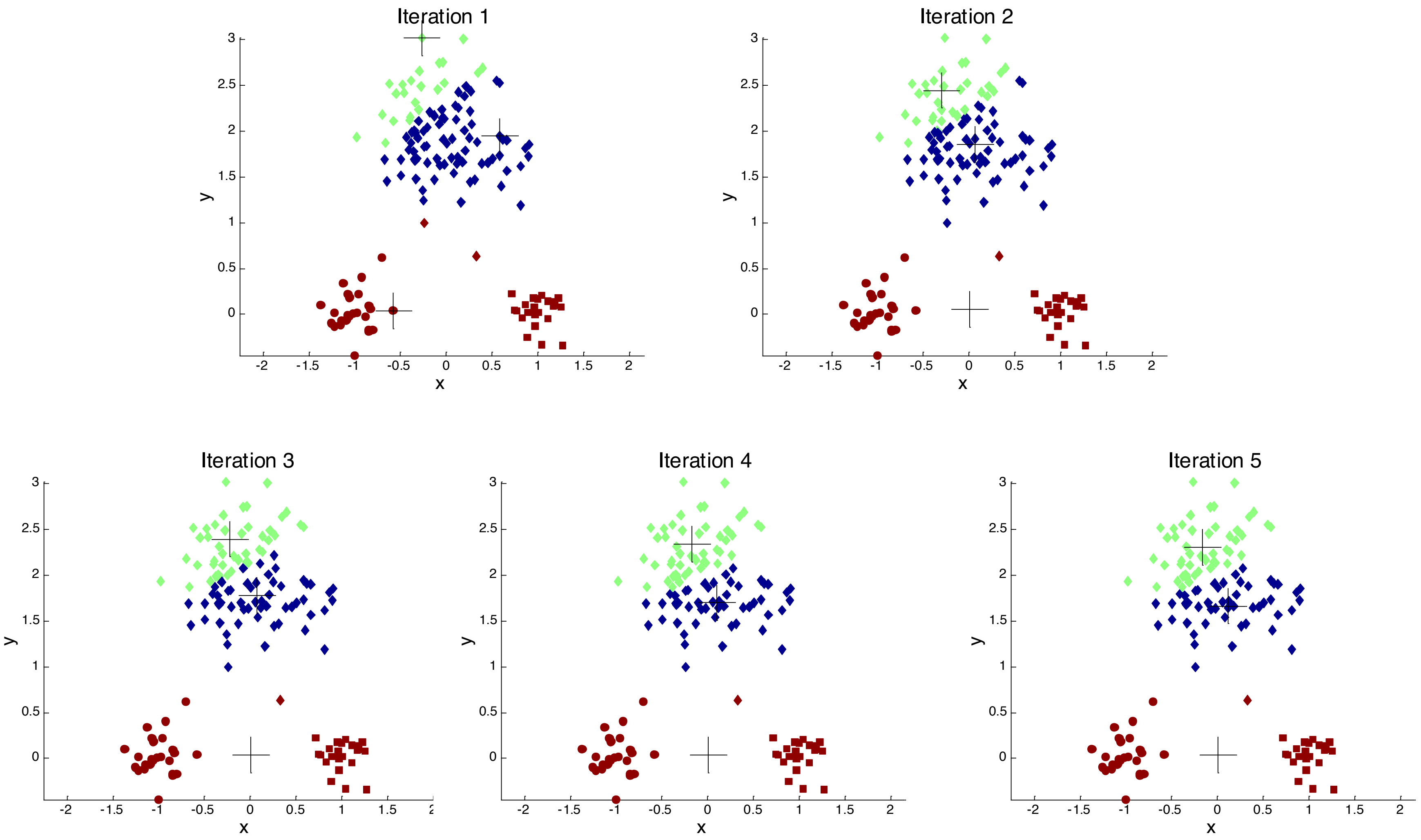


Sub-optimal Clustering

IMPORTANCE OF CHOOSING INITIAL CENTROIDS …



IMPORTANCE OF CHOOSING INITIAL CENTROIDS …



SOLUTIONS TO INITIAL CENTROIDS PROBLEM

- Multiple runs

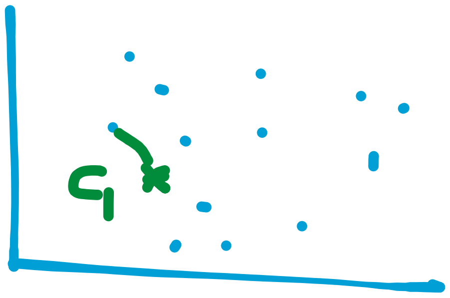
- Use some strategy to select the k initial centroids and then select among these initial centroids
  - Select most widely separated
  - K-means++ is a robust way of doing this selection
  - Use hierarchical clustering to determine initial centroids

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

K-MEANS++

- The k-means++ algorithm guarantees an approximation ratio O(log k) in expectation, where k is the number of centers

$k=3$



To select a set of initial centroids, C, perform the following

Select an initial point at random to be the first centroid C_1

For k – 1 steps For 2 steps

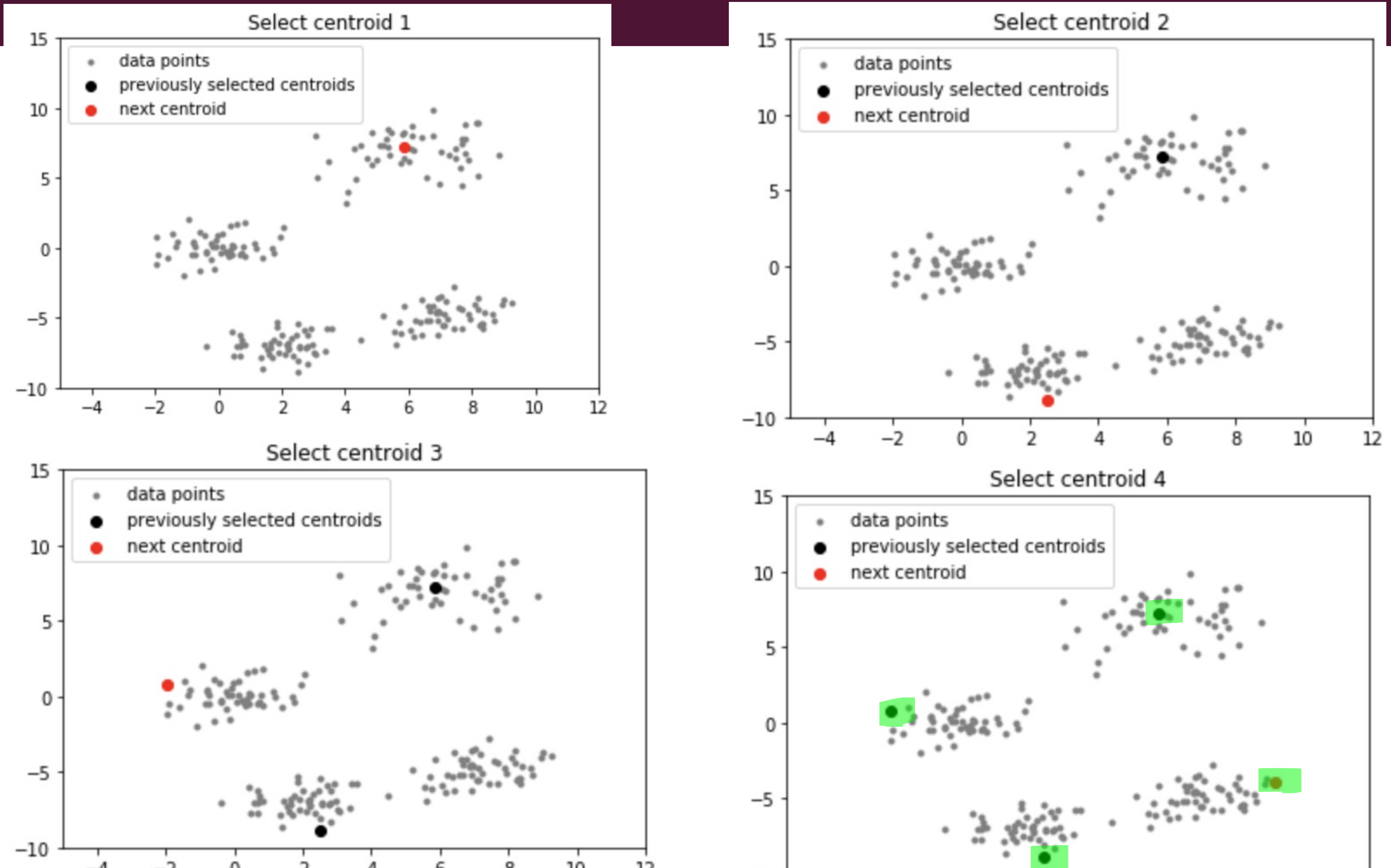
For each of the N points, x_i, $1 \leq i \leq N$, find the minimum squared distance to the currently selected centroids,

$C_1, \ldots, C_j$, $1 \leq j < k$, i.e., $\min_j d^2(C_j, x_i)$

Randomly select a new centroid by choosing a point with probability proportional to $\dfrac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$

End For

K-MEAN++



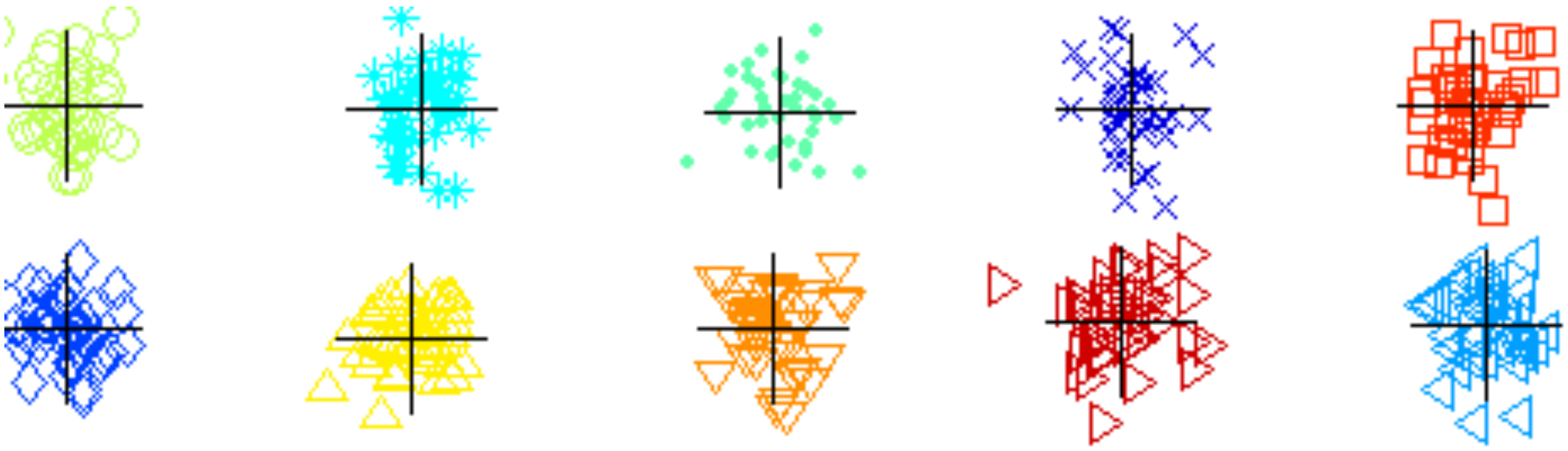
Select centroid 1
data points
previously selected centroids
next centroid
Select centroid 2
data points
previously selected centroids
next centroid
Select centroid 3
data points
previously selected centroids
next centroid
Select centroid 4
data points
previously selected centroids
next centroid

BISECTING K-MEANS

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

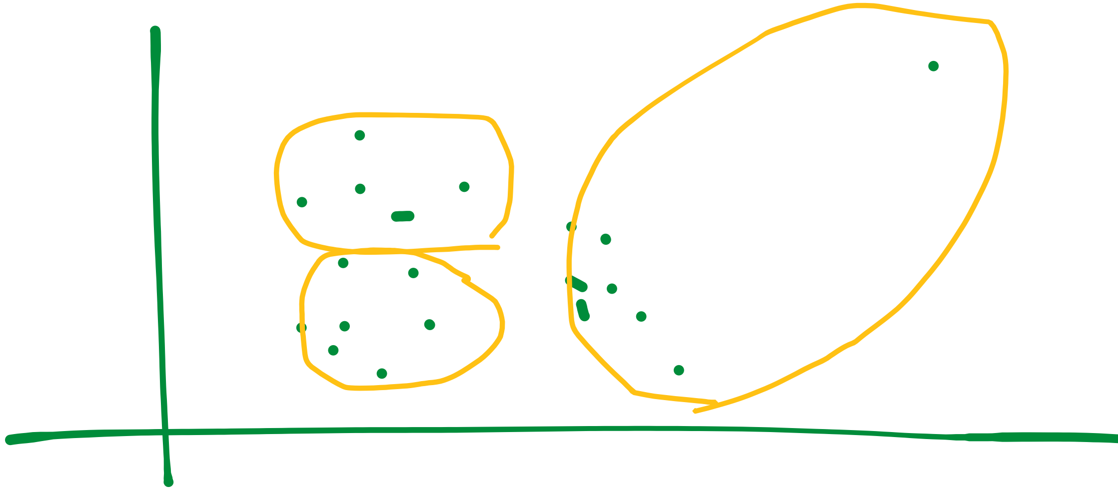
```
Initialize the list of clusters to contain the cluster containing all points.
repeat
   Select a cluster from the list of clusters
   for i = 1 to number_of_iterations do
      Bisect the selected cluster using basic K-means
   end for
   Add the two clusters from the bisection with the lowest SSE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

BISECTING K-MEANS EXAMPLE



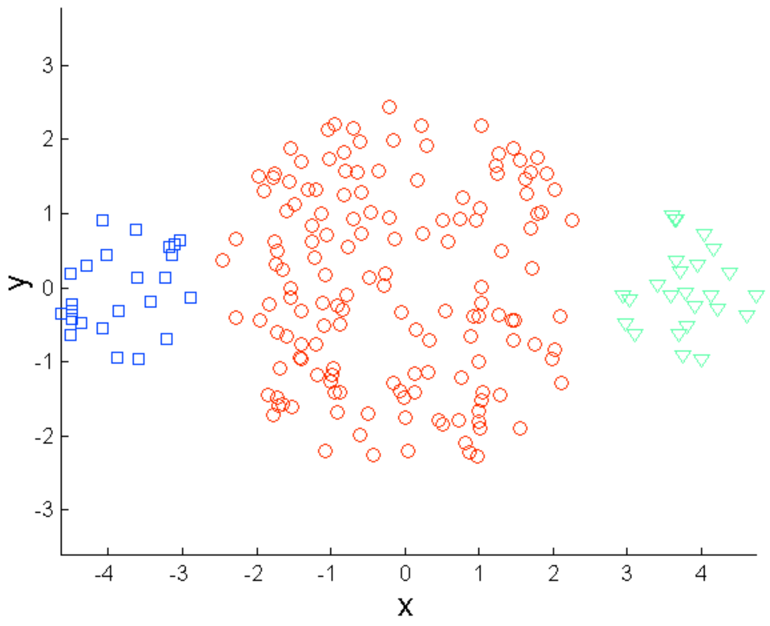
LIMITATIONS OF K-MEANS

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

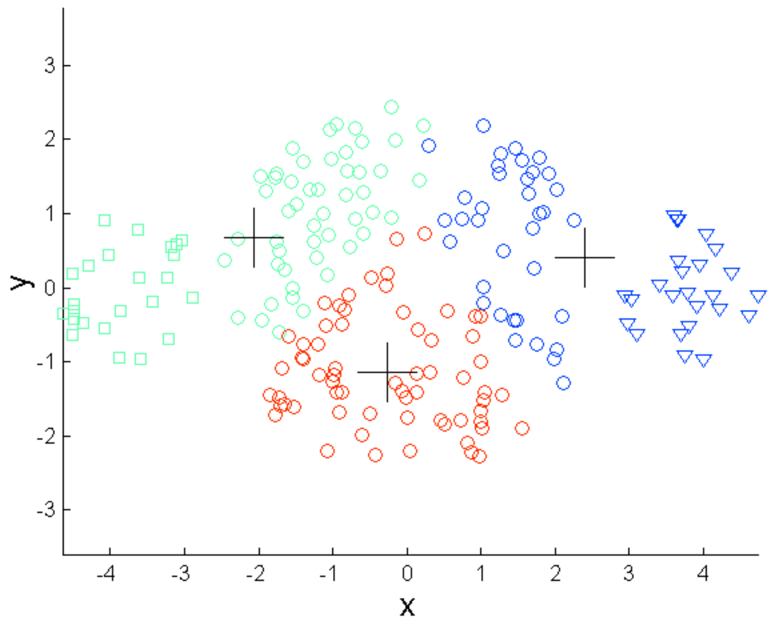


- K-means has problems when the data contains outliers.
  - One possible solution is to remove outliers before clustering

LIMITATIONS OF K-MEANS: DIFFERING SIZES



Original Points

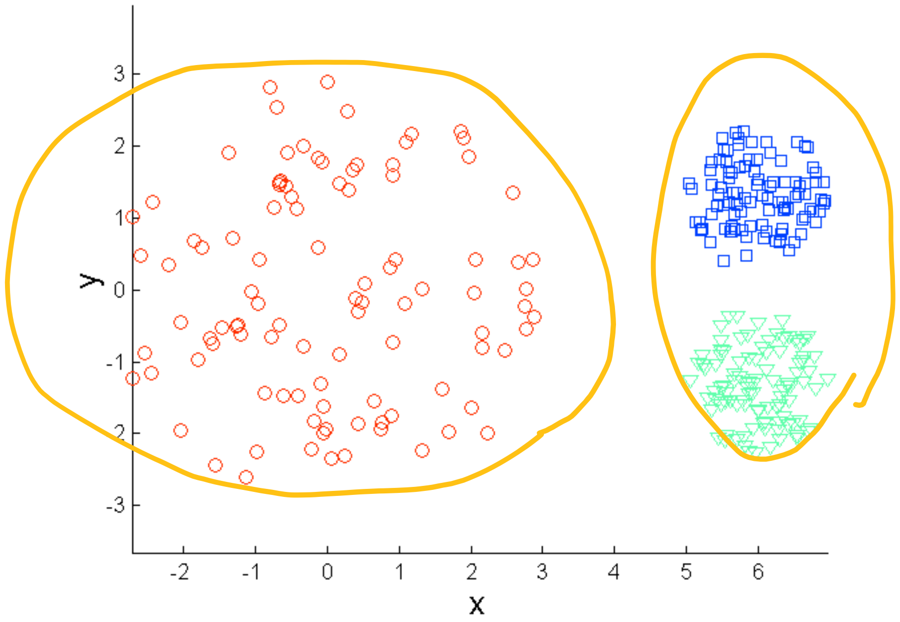


K-means (3 Clusters)

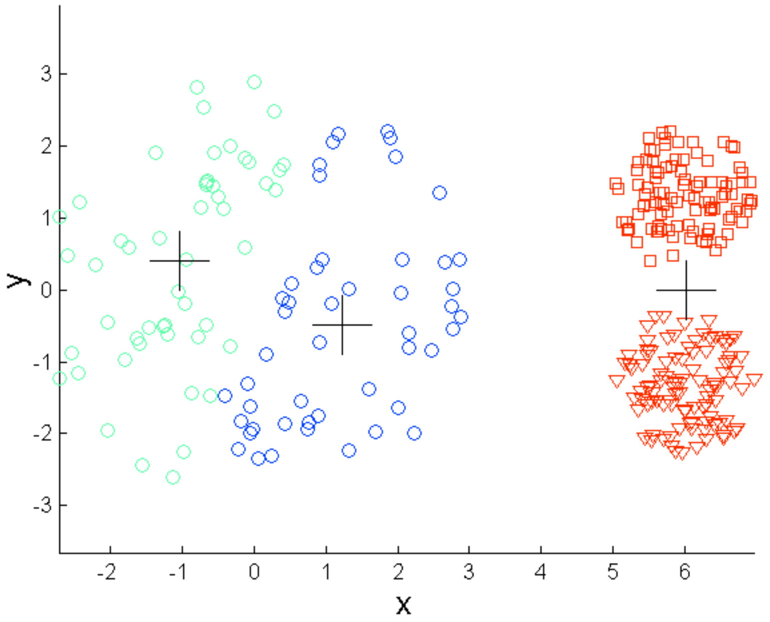
DATA MINING PIPELINE

- 1. Data pre-processing for the entire dataset:
  - Remove outlines from the entire dataset
  - Remove noises from the entire dataset
  - Re-calculate the features (derived features)
- 2. split the entire data into training, validation, and testing (k-folder cross validation)
- 3. model selection based on different measures

LIMITATIONS OF K-MEANS: DIFFERING DENSITY

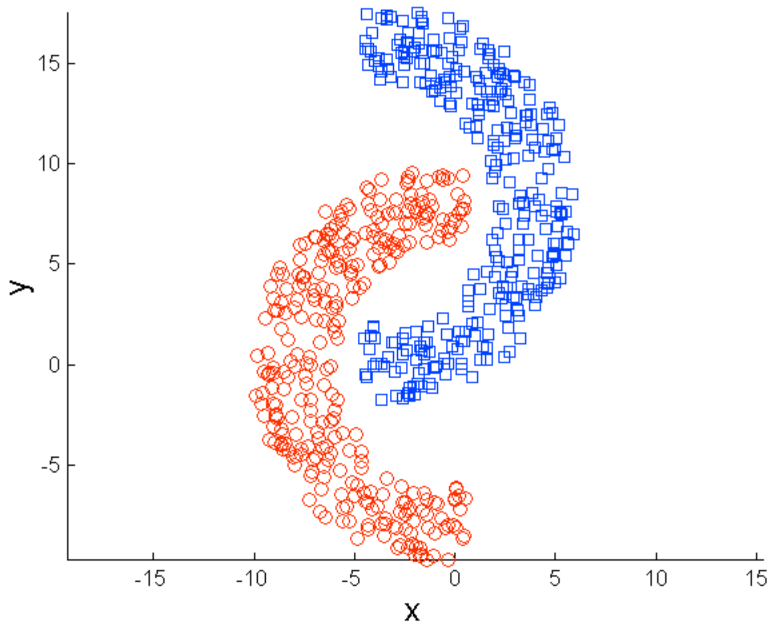


Original Points

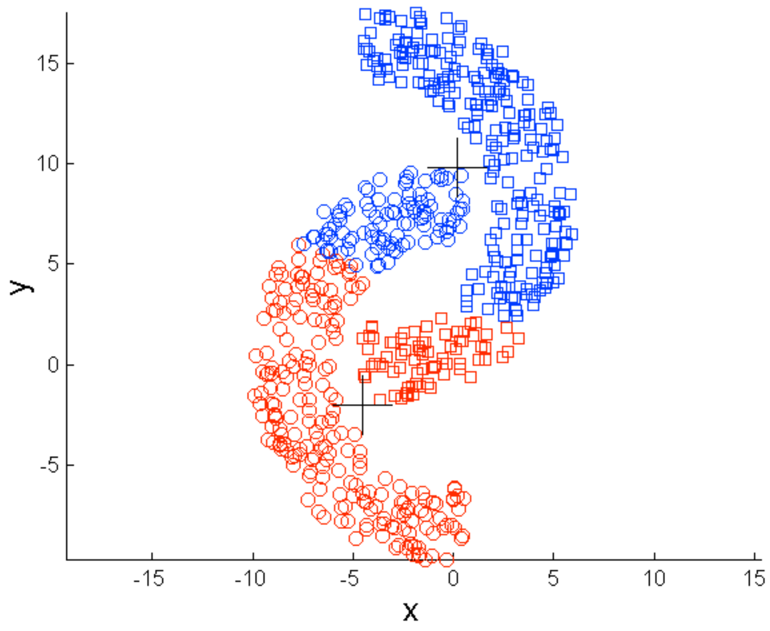


K-means (3 Clusters)

LIMITATIONS OF K-MEANS: NON-GLOBULAR SHAPES

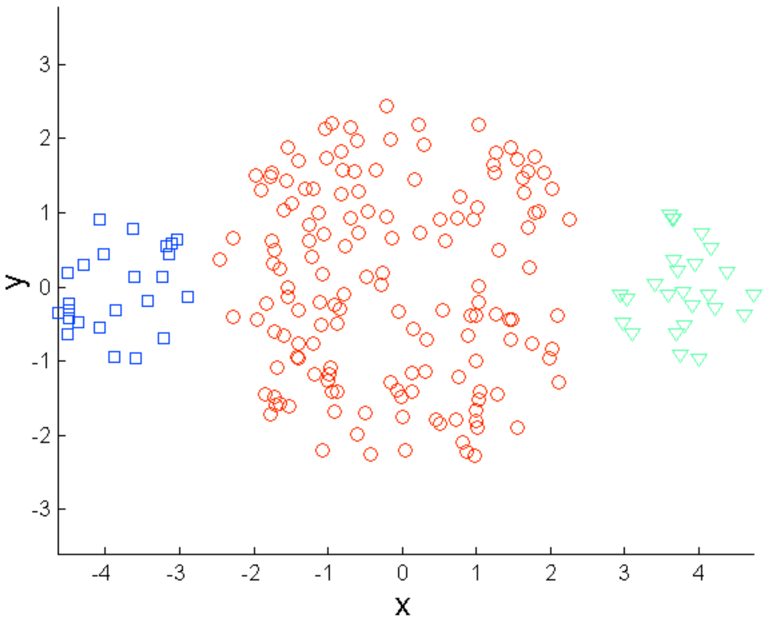


Original Points

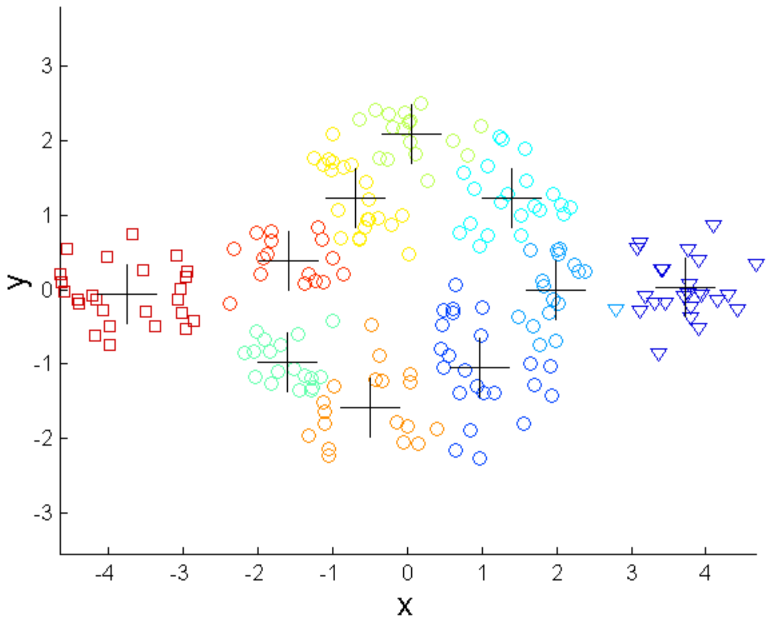


K-means (2 Clusters)

OVERCOMING K-MEANS LIMITATIONS



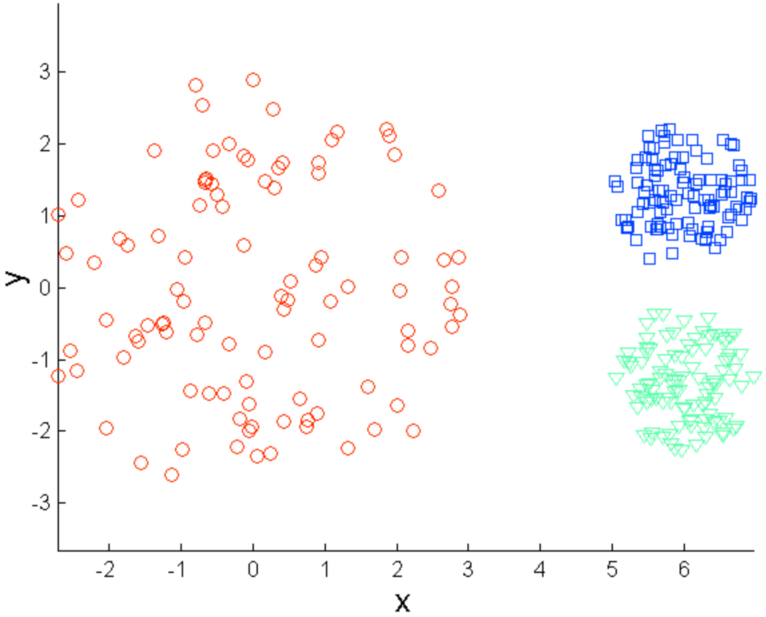
Original Points



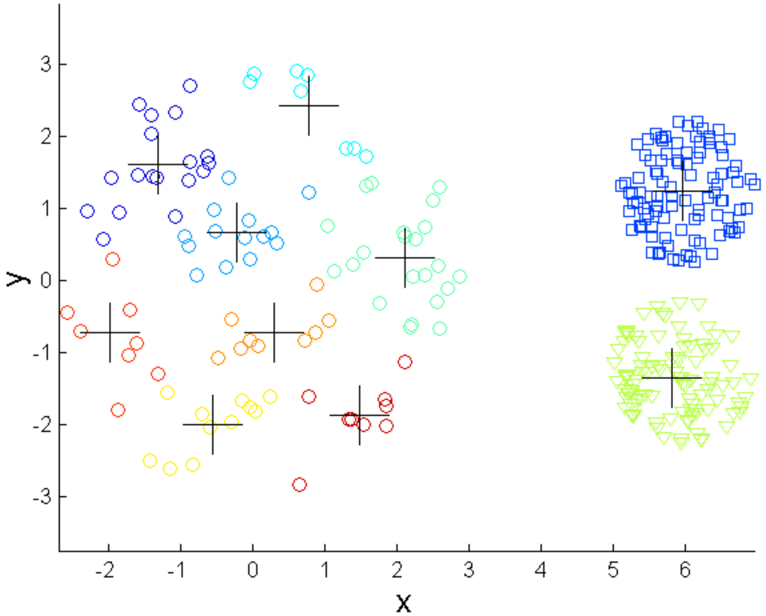
K-means Clusters

One solution is to find a large number of clusters such that each of them represents a part of a natural cluster. But these small clusters need to be put together in a post-processing step.

OVERCOMING K-MEANS LIMITATIONS



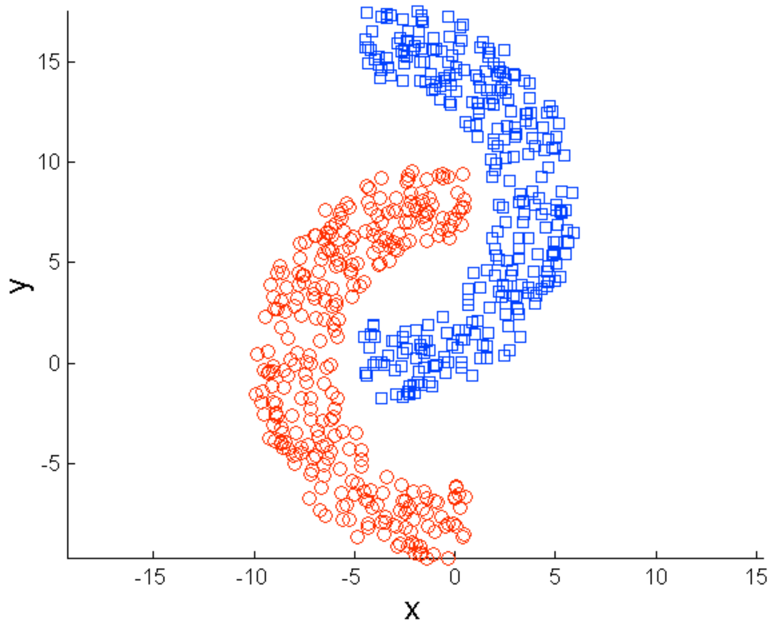
Original Points



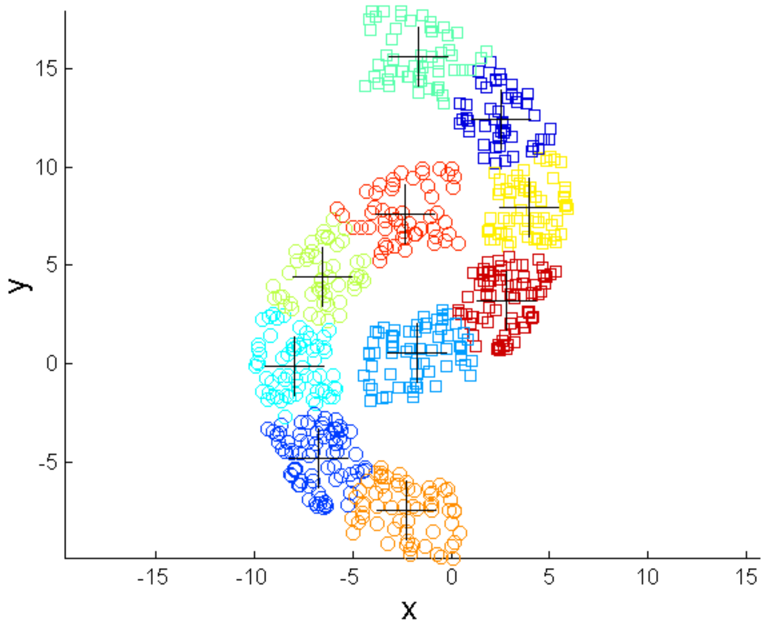
K-means Clusters

One solution is to find a large number of clusters such that each of them represents a part of a natural cluster. But these small clusters need to be put together in a post-processing step.

OVERCOMING K-MEANS LIMITATIONS



Original Points



K-means Clusters

One solution is to find a large number of clusters such that each of them represents a part of a natural cluster. But these small clusters need to be put together in a post-processing step.