# Cluster Analysis - II

George Michailidis
Department of Statistics
The University of Michigan

March 5, 2009

## 1. Partinioning algorithms

- K-means and its variants
- Density based clustering

## K-means clustering

- Each clustering is associated with a **centroid**
- Each object in the data is assigned to the cluster with the closest centroid
- Number of clusters $K$ must be **a priori specified**

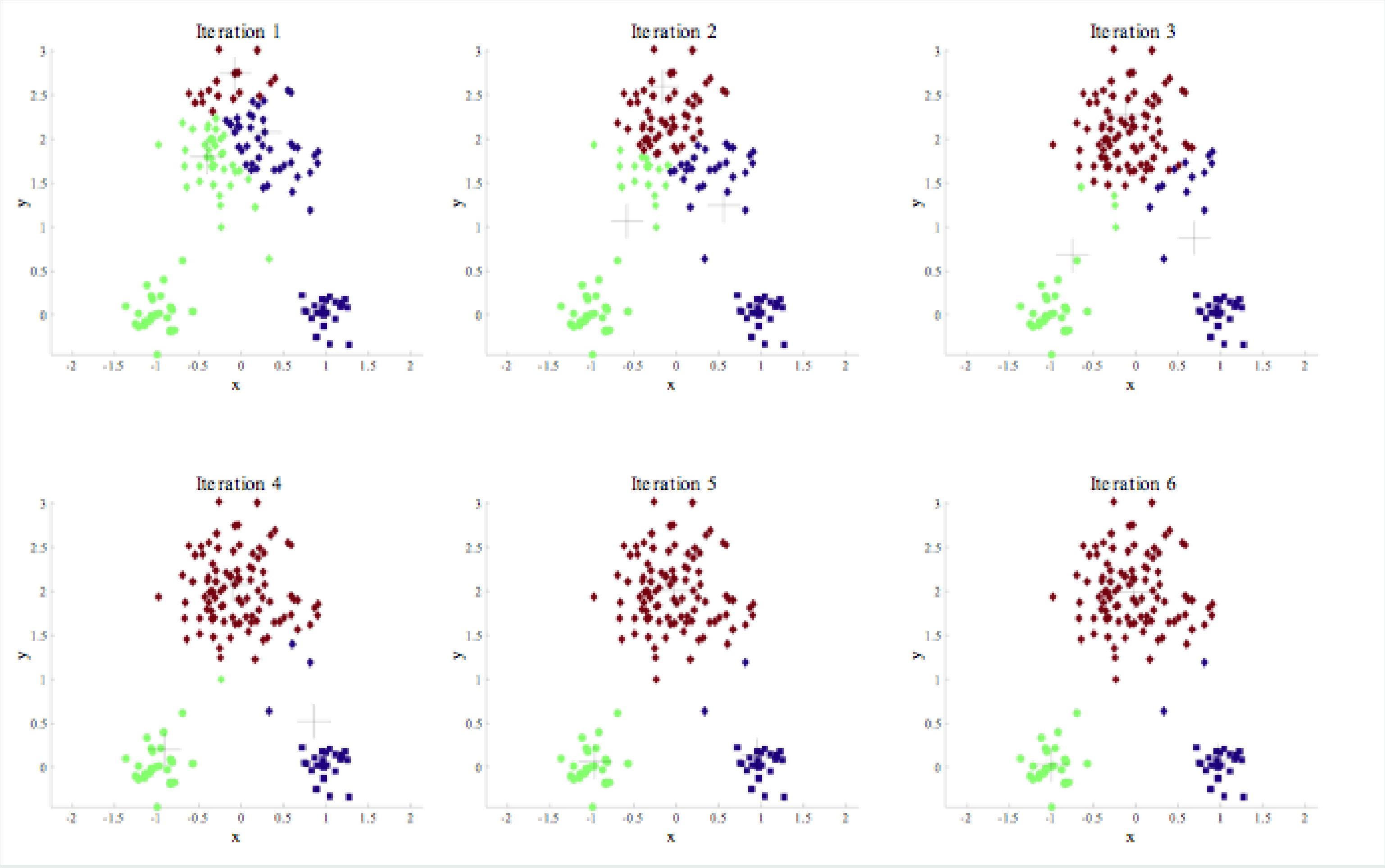
K-means algorithm:

1. Select $K$ points in $p$-dimensional space as the initial centroids
2. **repeat**
3. Form $K$ clusters by assigning all object to their closest centroid
4. Recompute the centroid of each cluste
5. **until** the centroids do not change

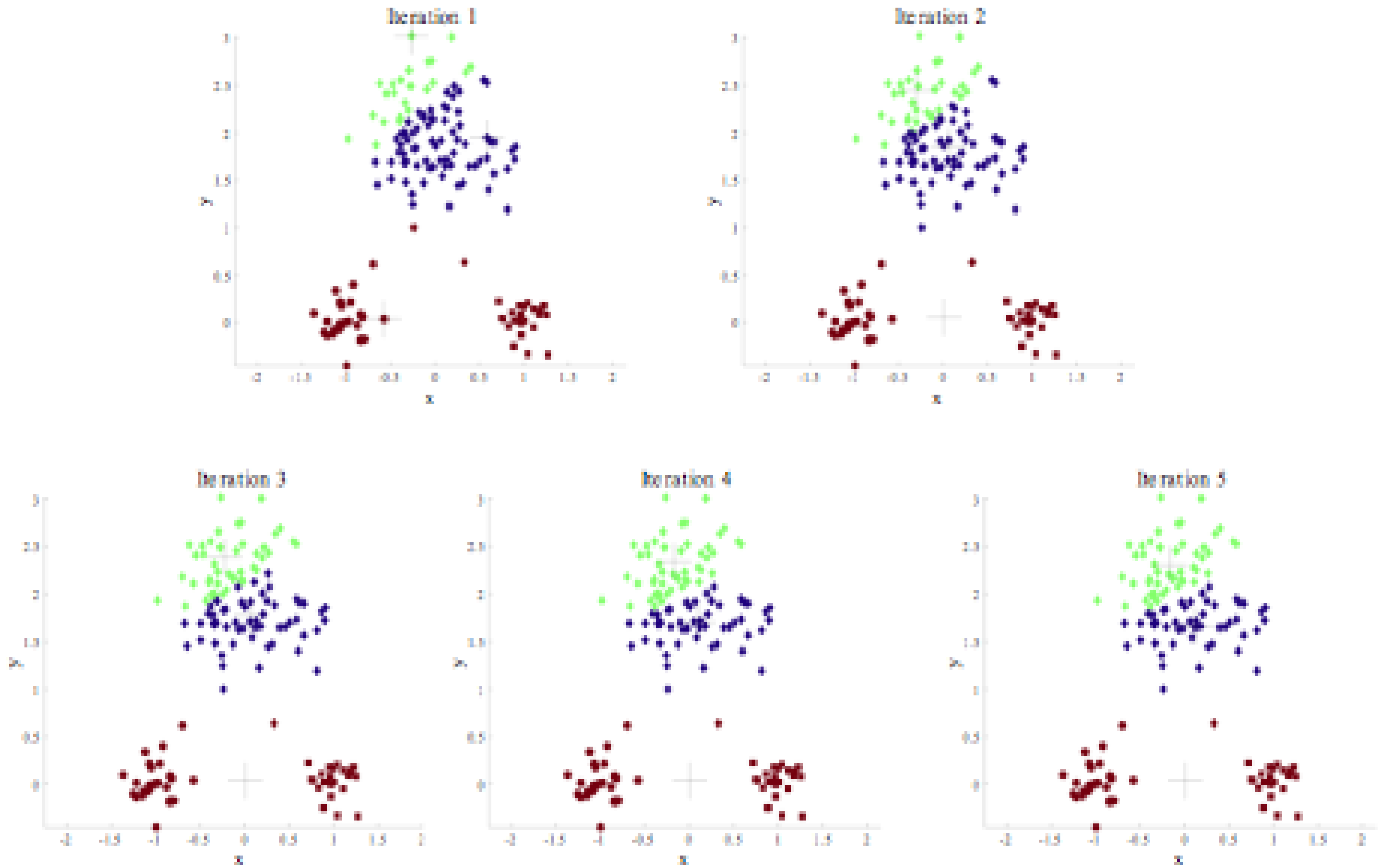
# Some important details

- Initial centroids can be chosen either at random or according to the result of a hieararchical clustering method
- The centroid is typically the multivariate mean of the objects in the cluster
- 'Closeness' is measured by Euclidean distance, or some other distance fucntion
- It can be proved that the algorithm always stops after a finite number of iterations
- Biggest imporvements occur in the first few iterations
- Computational complexity $\mathcal{O}(n * K * p)$

# K-means in action



# What can go wrong!!



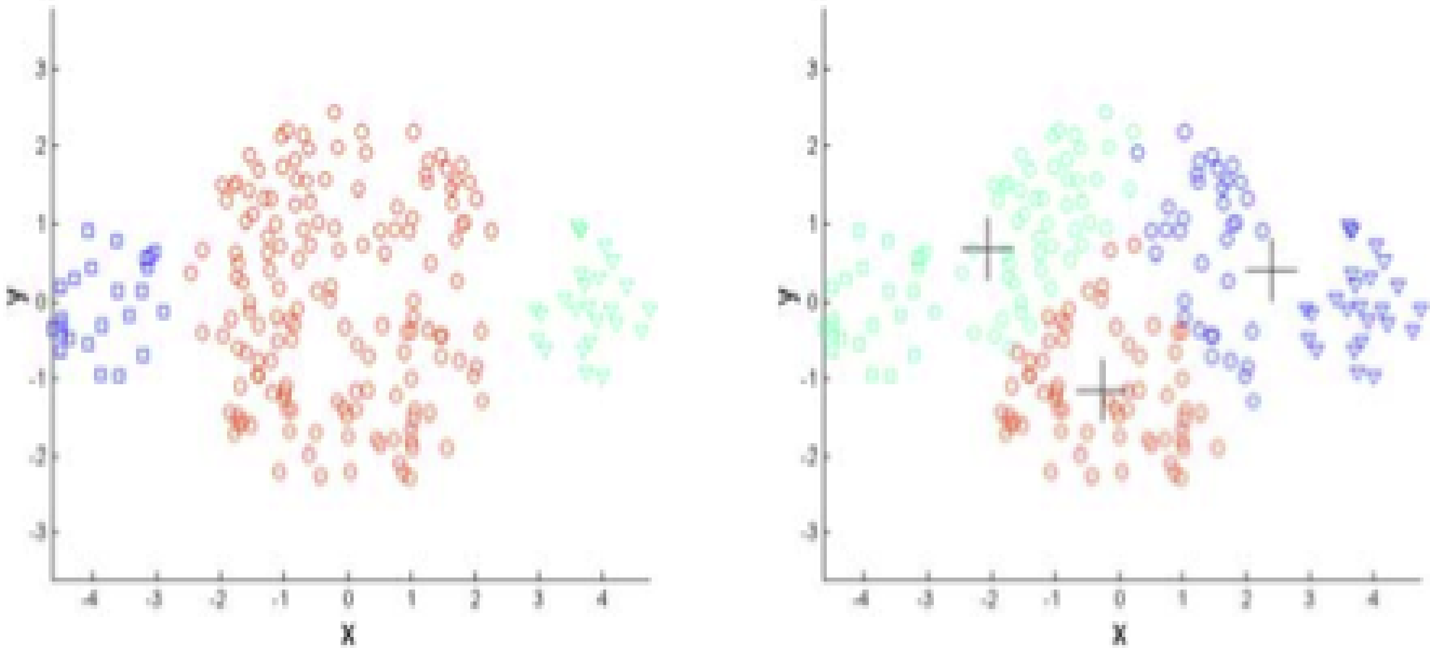
## Overcoming the initial centroids problem

- Multiple runs of the algorithm
  However, probability is not on your side!!
  For equal size 'real' clusters, probability of selecting one centroid from each cluster is $K!/K^K$
- Use the solution from some hierarchical algorithm

## 2. Some other caveats

- Algorithm may result in empty clusters
- Algorithm may result in some artificially small clusters (one idea is to eliminate outliers)
- Algorithm has a hard time with clusters of different size, density and non-spherical shape

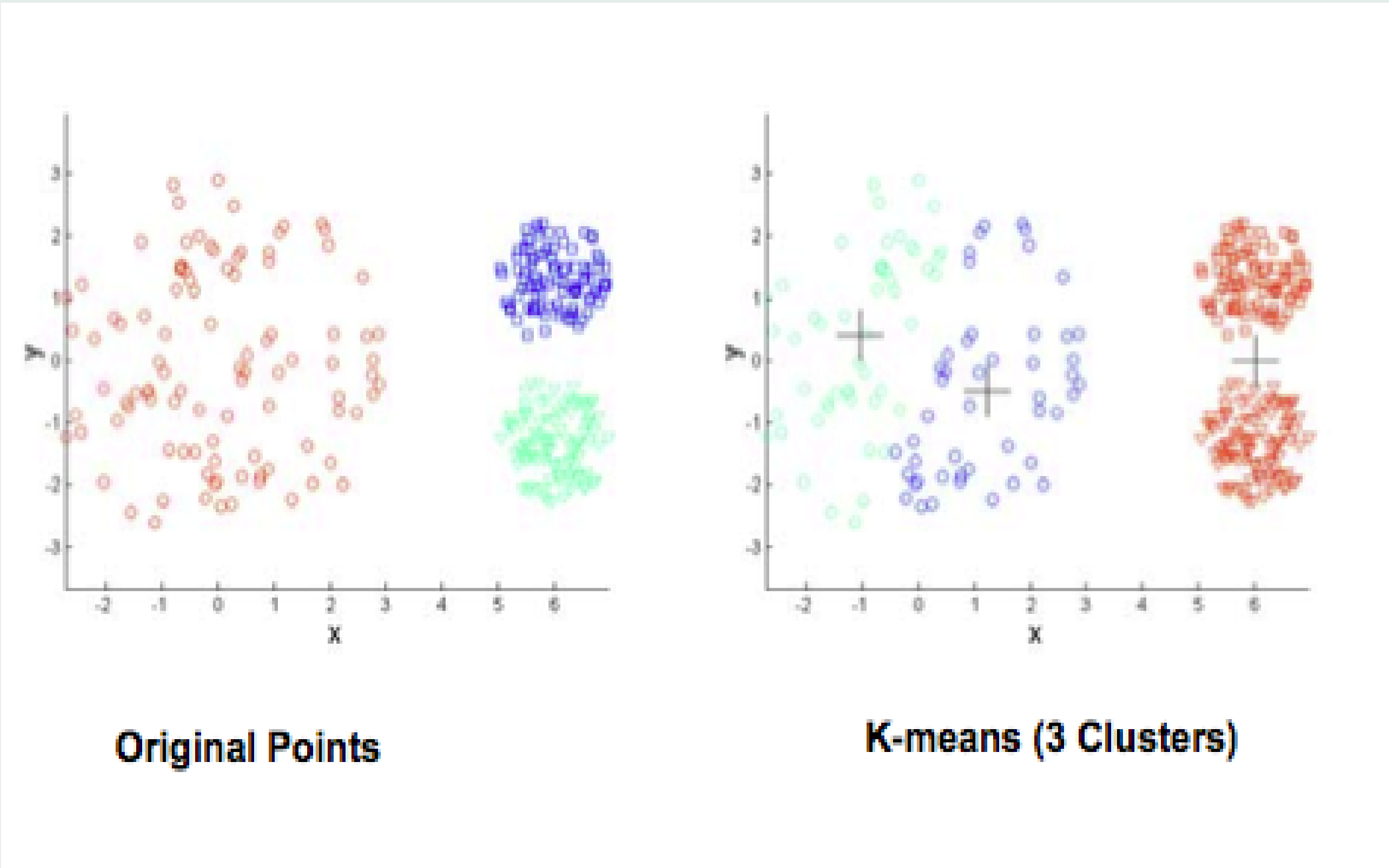
# Limitations of K-means: different sizes



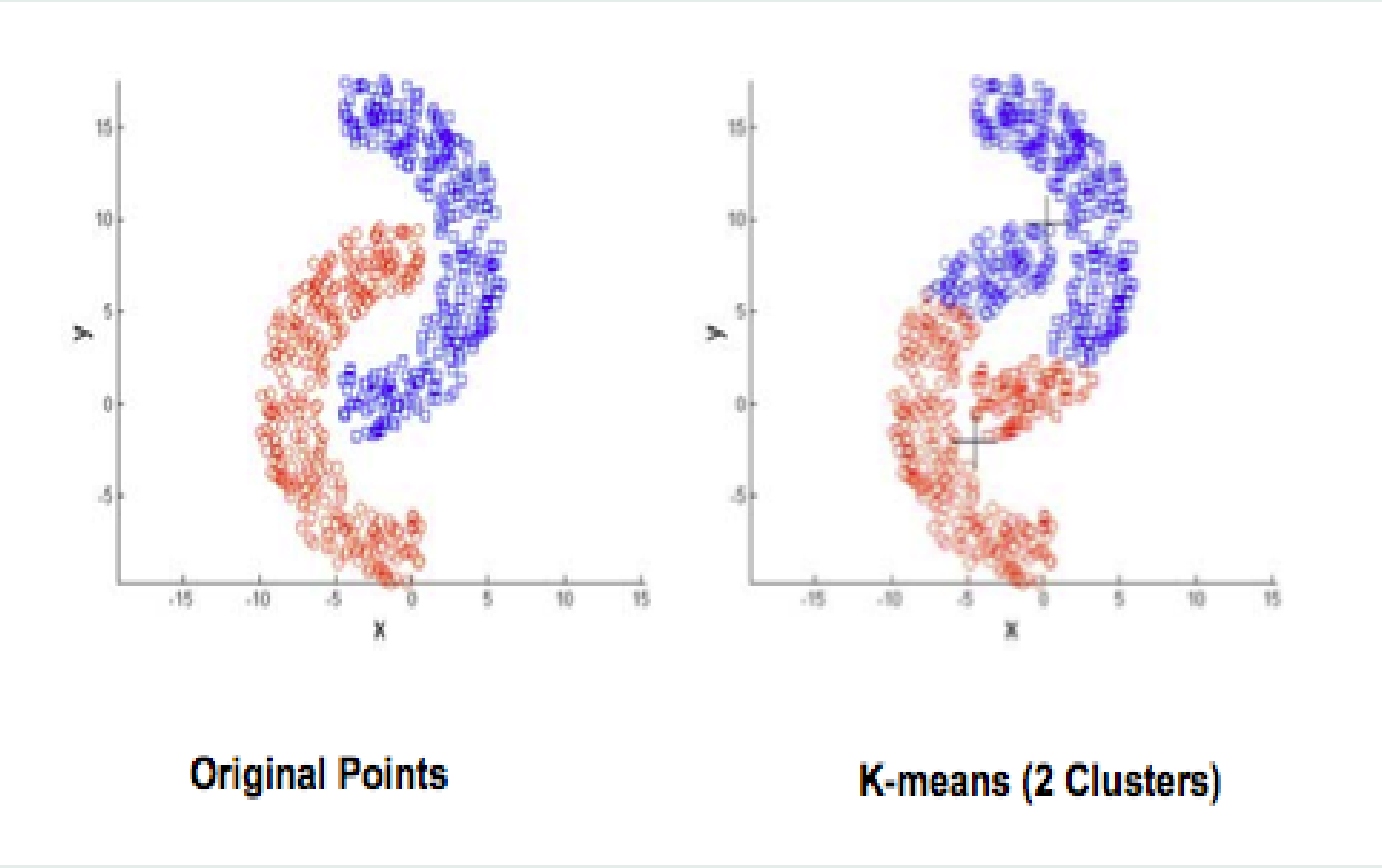
Original Points

K-means (3 Clusters)

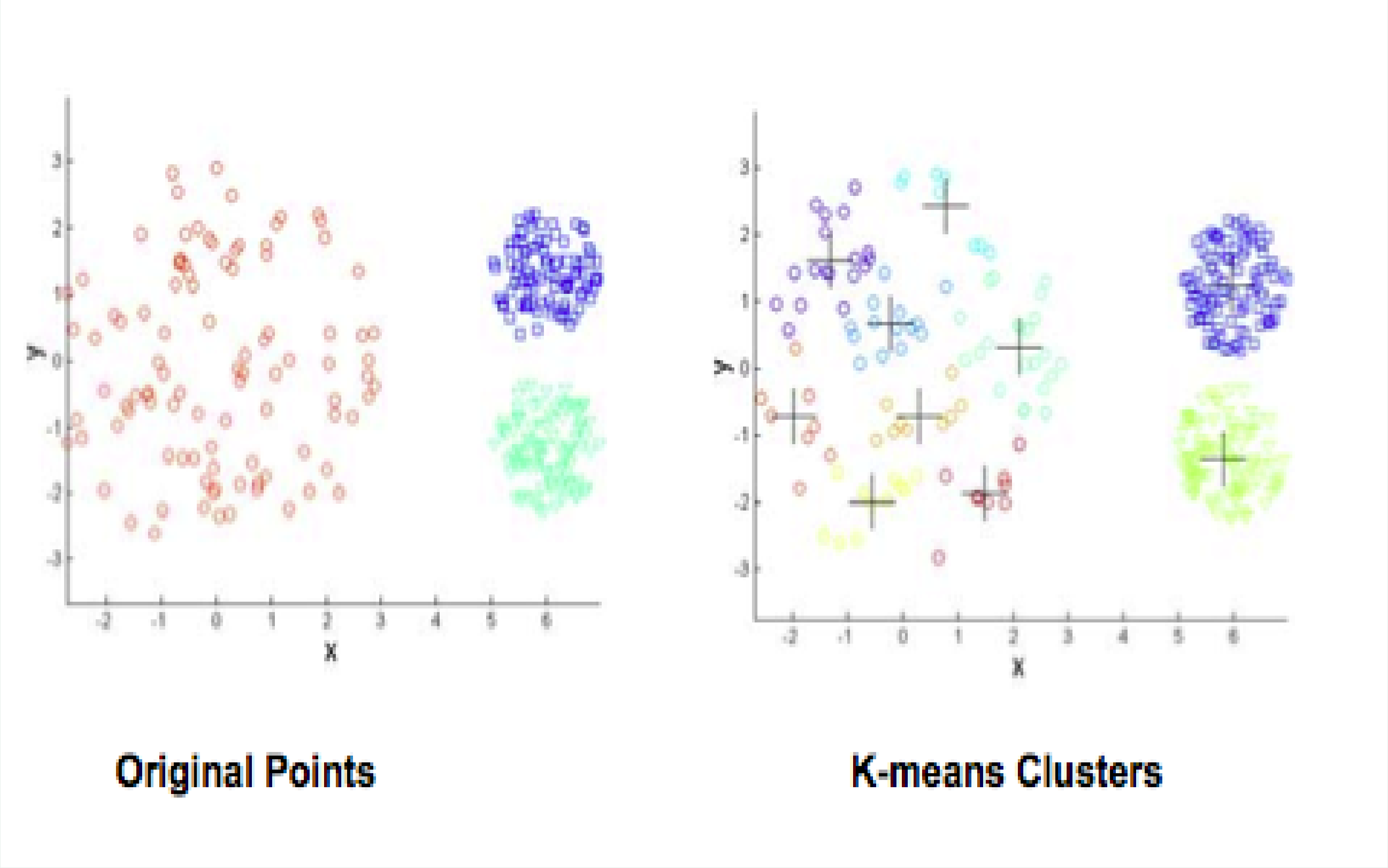
# Limitations of K-means: different densities



# Limitations of K-means: non-spherical shapes



# Potential solution: use large K and then stitch results together easier said than done!!



## 3. Evaluating the quality of a clustering solution

- Cluster homogeneity
  Within sum of squares error (WSSE)
  For each object the 'error' is the distance to its cluster centroid
  WSSE=$\sum_{k=1}^{K}\sum_{j\in C_k} d^2(m_k, x_j)$
- Given two clustering solutions, the one with smaller WSSE should be prefered
- WSSE usually decreases as $K$ increases
- Cluster separation
  Between sum of squares error (BSSE)
  For each cluster the 'error' is the distance between the cluster centroid and the 'grand mean'
  BSSE=$\sum_{k=1}^{K} d^2(m_k, m)$C_k

## The silhouette coefficient:

- Combines homogeneity and separation
- Let $a$=average distance of object $i$ to the other objects in the same cluster
- Let $b$=min(average distance of object $i$ to objects in other clusters
- $s = 1 - (a/b)$ if $a < b$; the closer to 1 the better