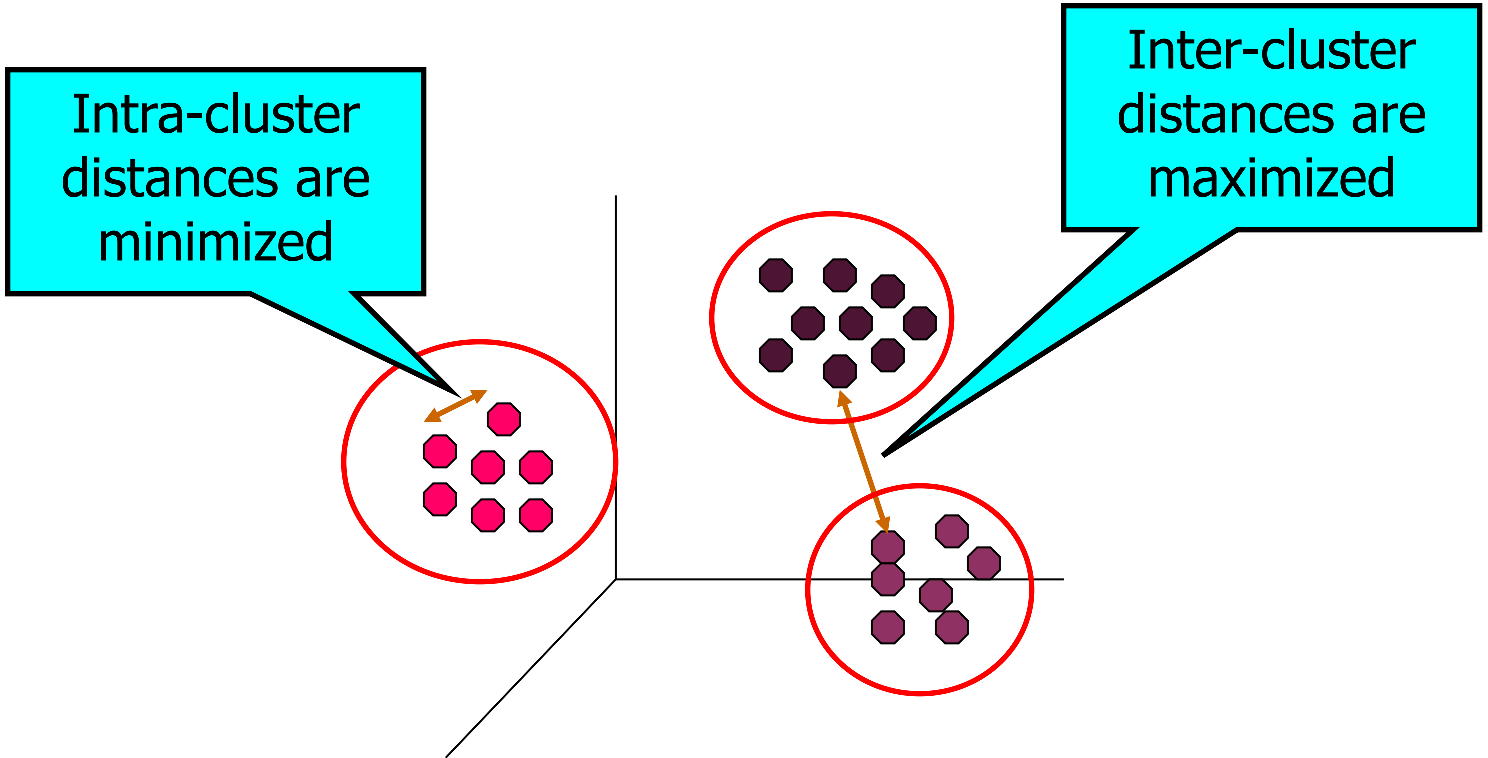
CLUSTERING

WHAT IS CLUSTER ANALYSIS?

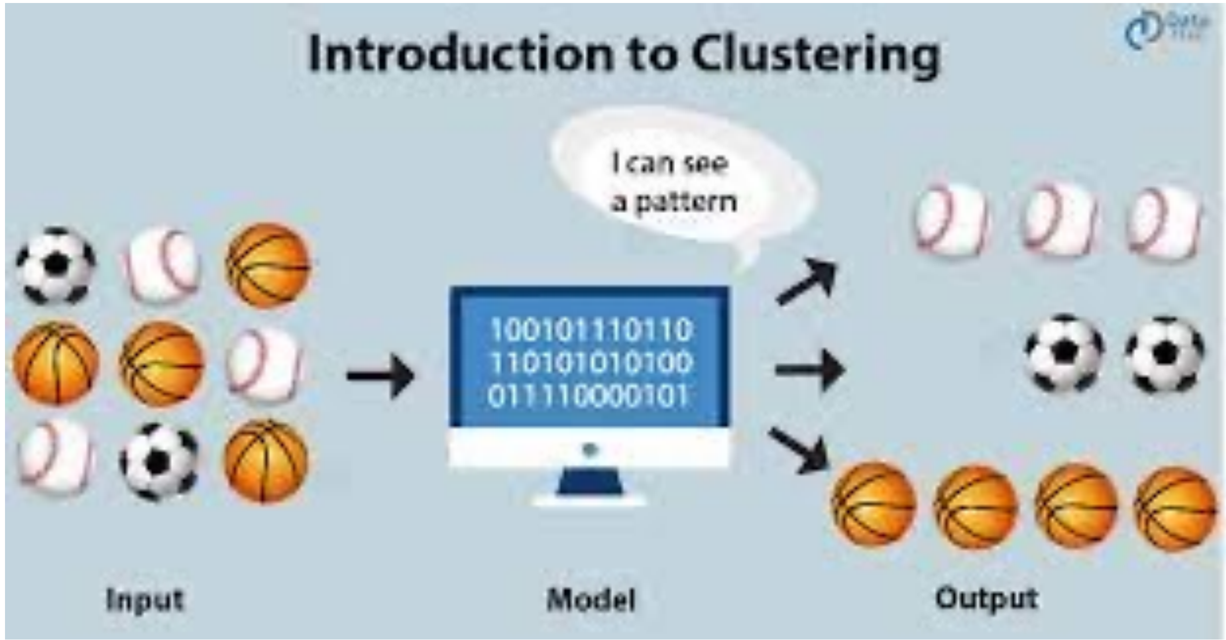
- Given a set of objects, place them in groups such that:

  the objects in a group are similar (or related)

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

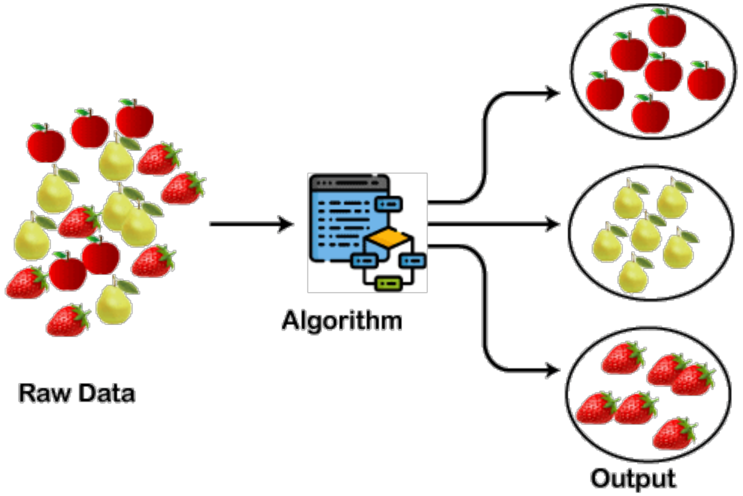


APPLICATIONS OF CLUSTER ANALYSIS



- **Understanding**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

- **Summarization**
  - Reduce the size of large data sets



NOTION OF A CLUSTER CAN BE AMBIGUOUS

How many clusters?

Six Clusters

Two Clusters

Four Clusters

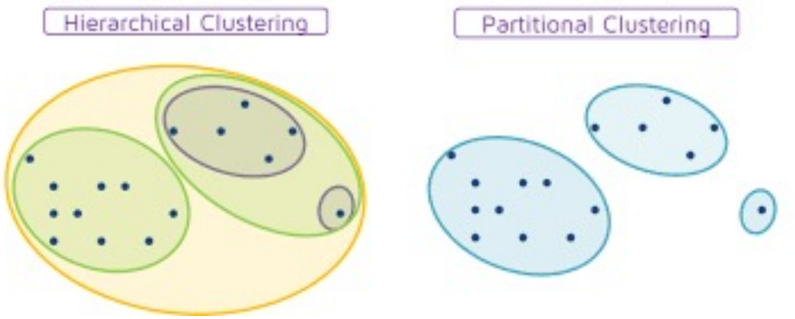
TYPES OF CLUSTERINGS

- A clustering is a set of clusters

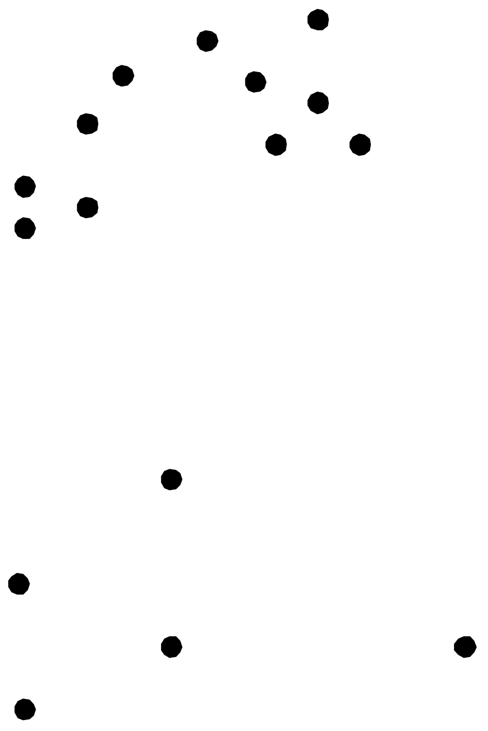
- Important distinction between hierarchical and partitional sets of clusters

  - Partitional Clustering
  - A division of data objects into non-overlapping subsets (clusters)

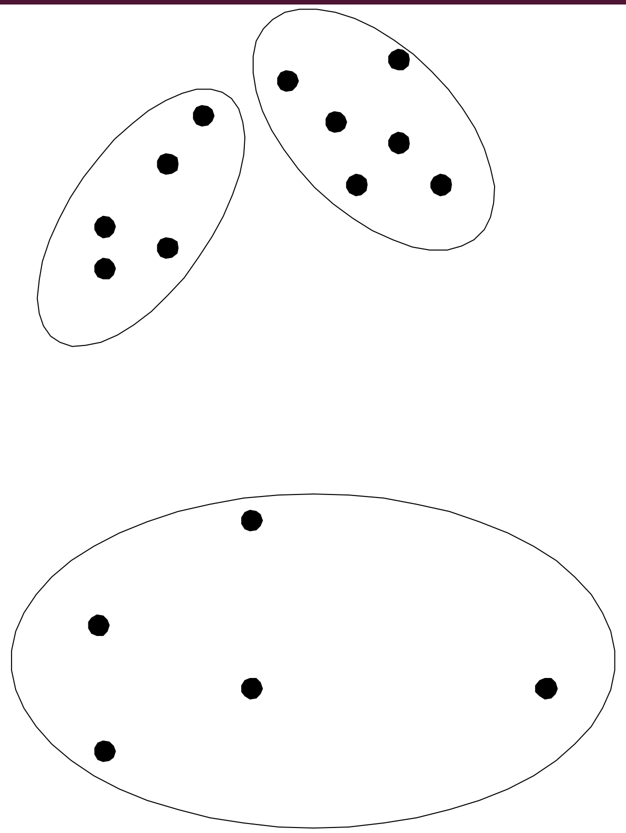
  - Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree



PARTITIONAL CLUSTERING

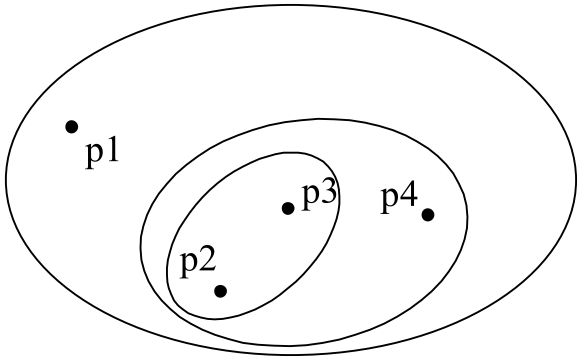


Original Points

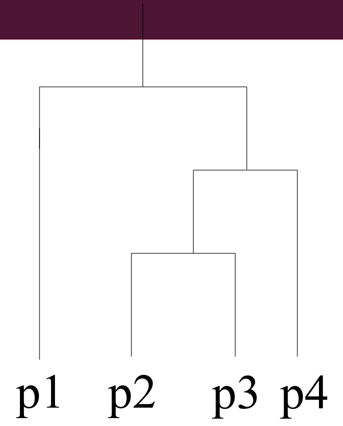


A Partitional Clustering

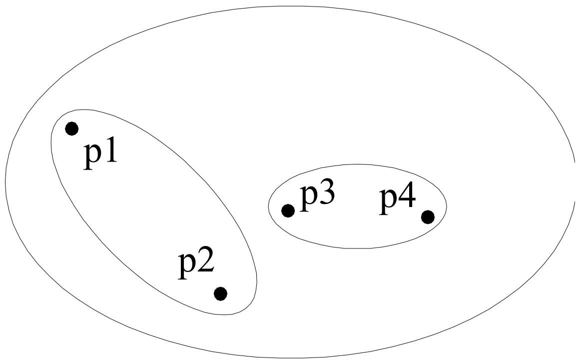
HIERARCHICAL CLUSTERING



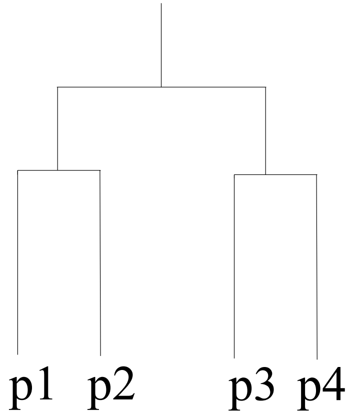
Traditional Hierarchical Clustering



Traditional Dendrogram



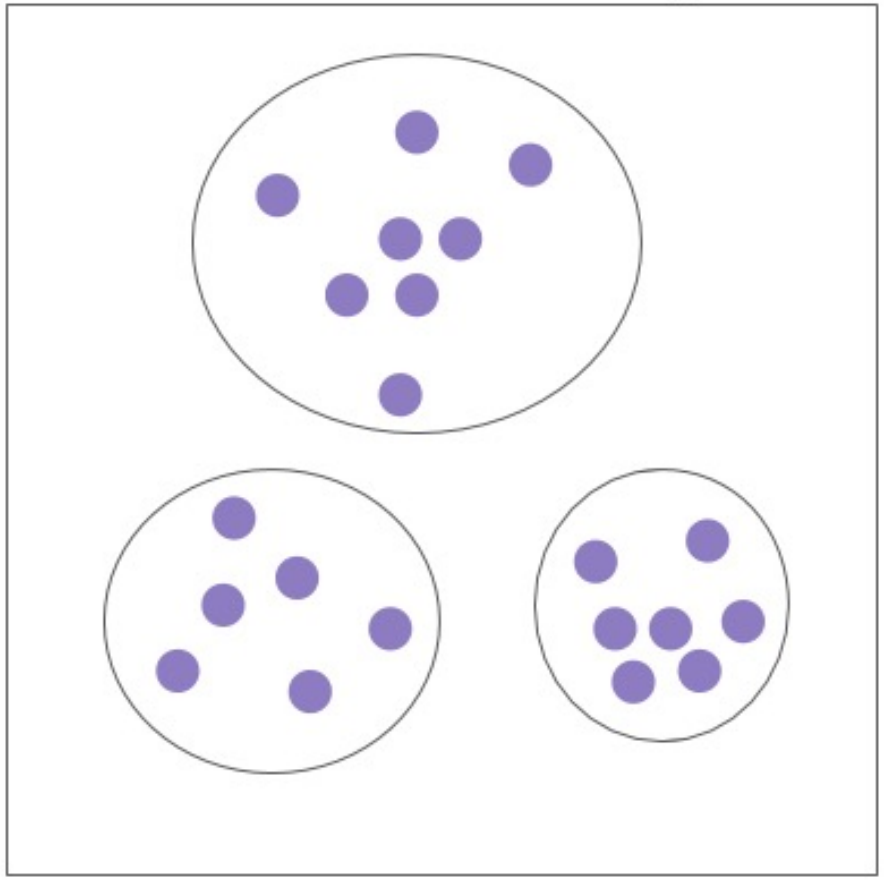
Non-traditional Hierarchical Clustering



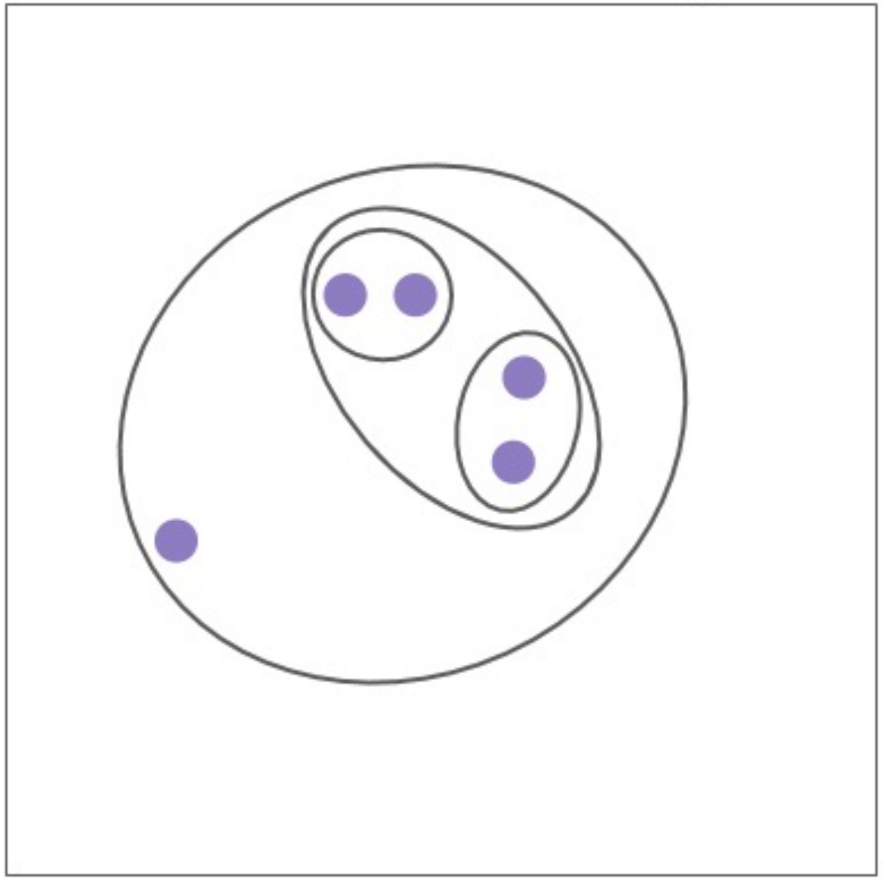
Non-traditional Dendrogram

TYPES OF CLUSTERING

Partitional Clustering



Hierarchical Clustering



OTHER DISTINCTIONS BETWEEN SETS OF CLUSTERS

- Exclusive versus non-exclusive
  - non-exclusive clustering:

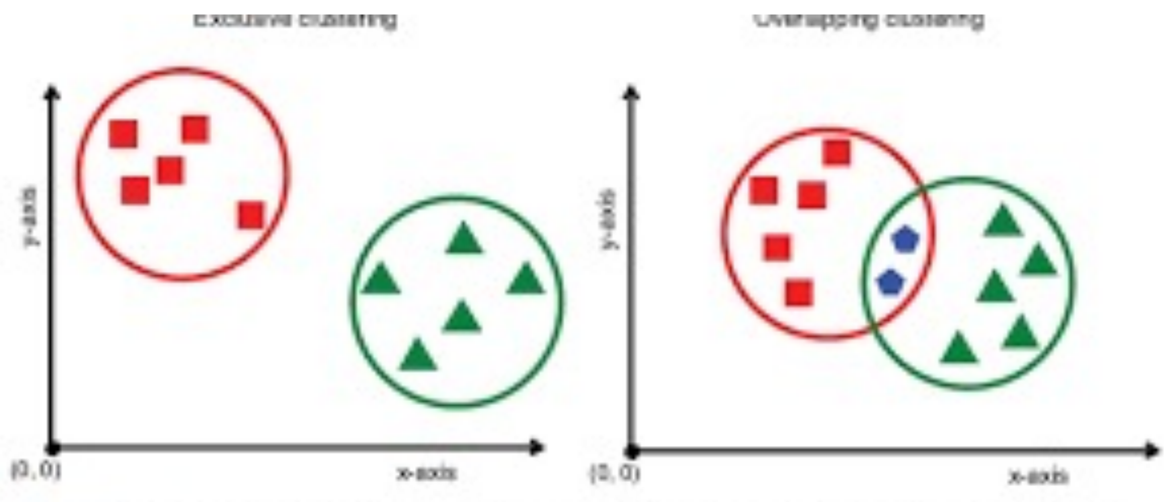
    points may belong to multiple clusters.

    an belong to multiple classes or could be 'border' points

    fuzzy clustering : a point belongs to every cluster with some weight between 0 and 1

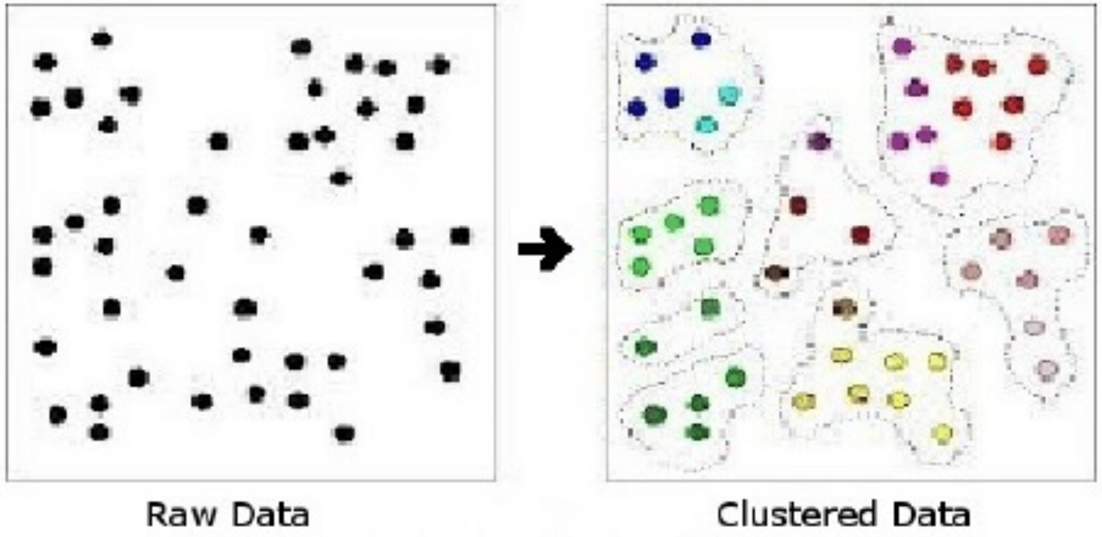
    weights sum to 1

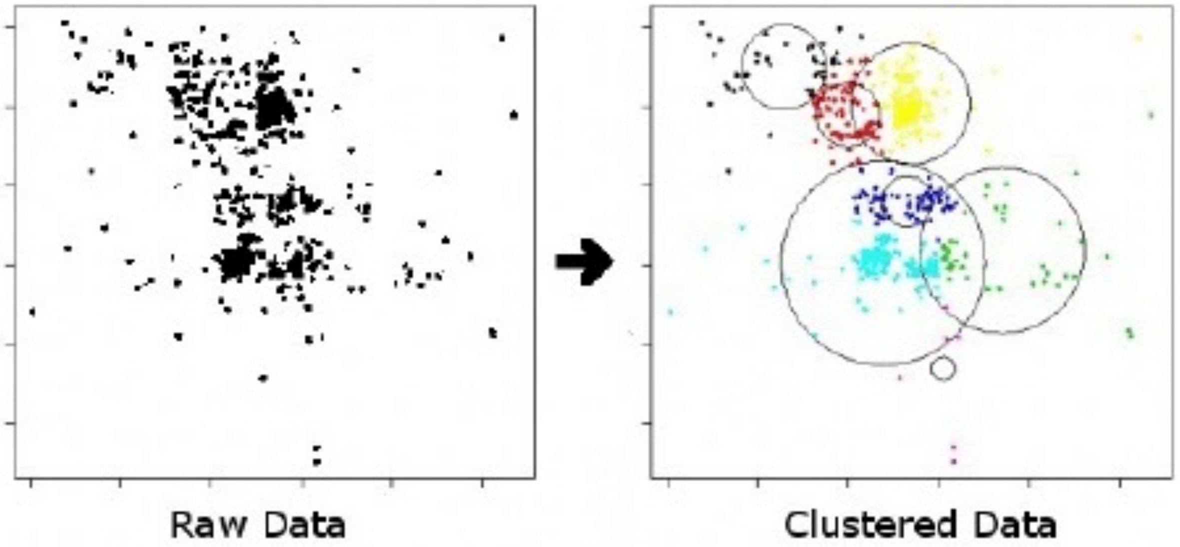
    probabilistic clustering has similar characteristics



- Partial versus complete
  - Partial: only cluster some of the data

OTHER DISTINCTIONS BETWEEN SETS OF CLUSTERS



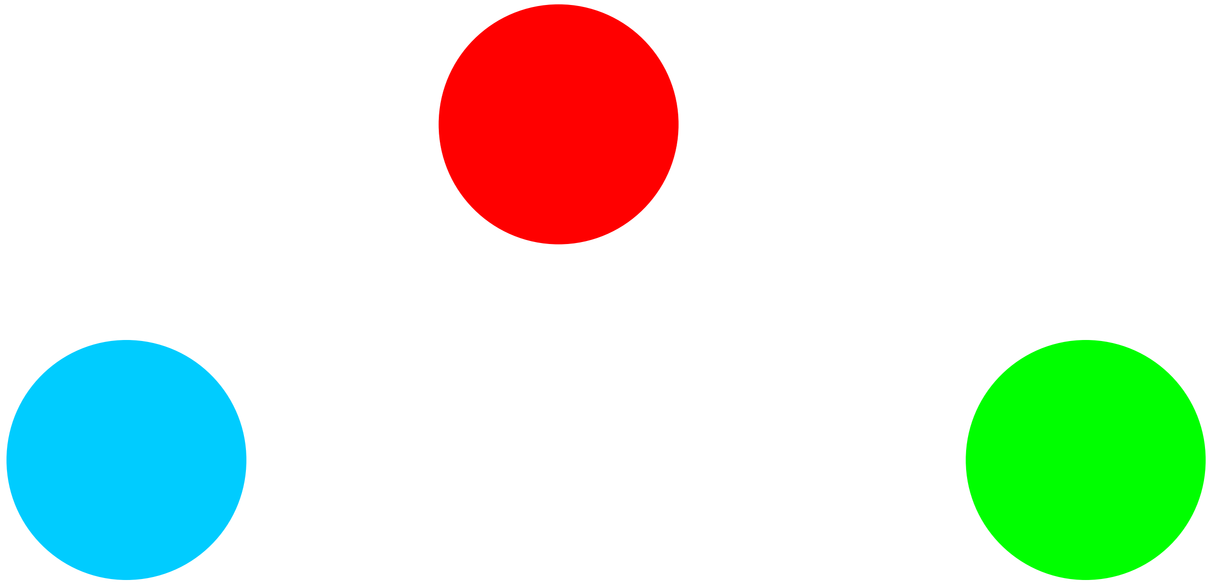


TYPES OF CLUSTERS

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

TYPES OF CLUSTERS: WELL-SEPARATED

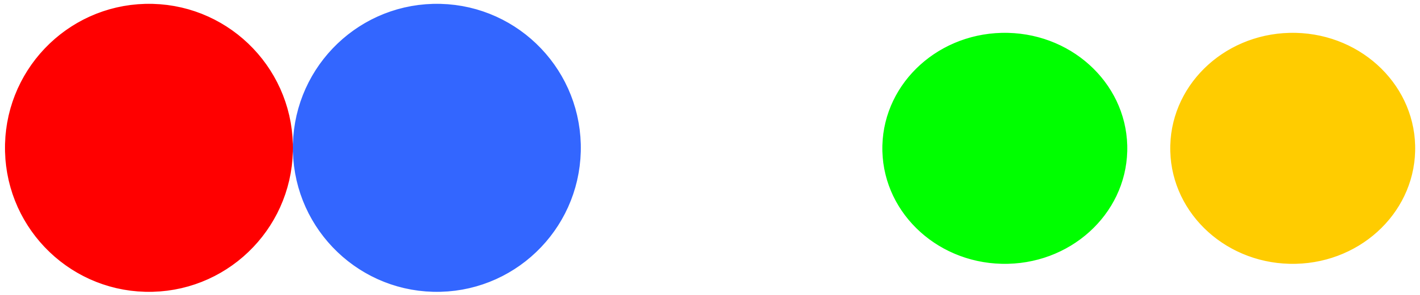
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

TYPES OF CLUSTERS: PROTOTYPE-BASED

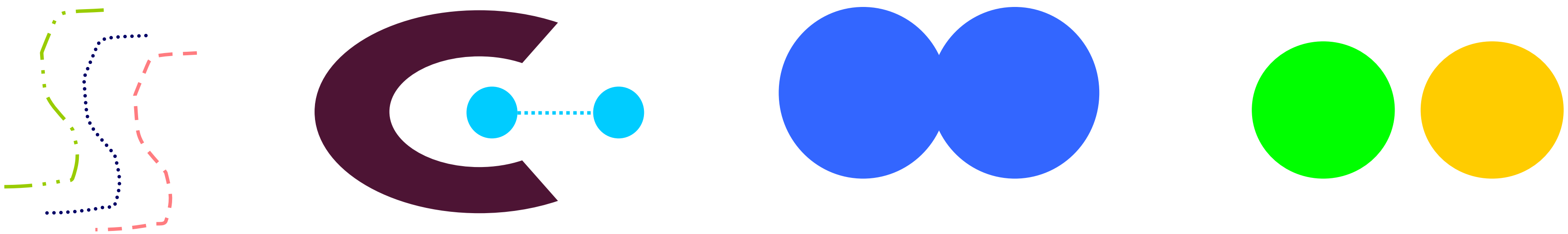
- Prototype-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster
  - centroid, the average of all the points in the cluster,
  - medoid, the most "representative" point of a cluster



4 center-based clusters

TYPES OF CLUSTERS: CONTIGUITY-BASED

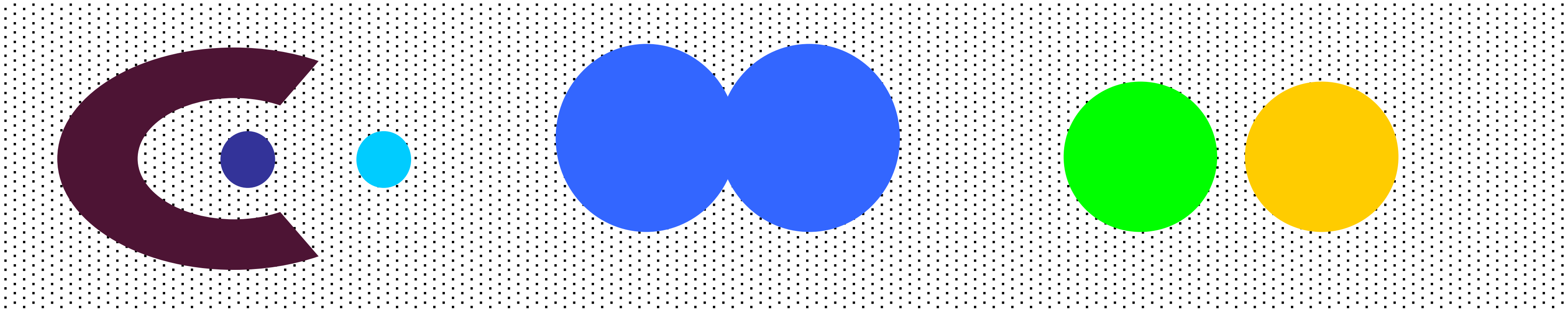
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

TYPES OF CLUSTERS: DENSITY-BASED

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

TYPES OF CLUSTERS: OBJECTIVE FUNCTION

- Clusters Defined by an Objective Function
  - Finds clusters that minimize or maximize an objective function.
  - Enumerate all possible ways of dividing the points into clusters
  - Evaluate the goodness of each potential set of clusters.
  - Objectives can be global or local
    - Hierarchical clustering algorithms typically have local objectives
    - Partitional algorithms typically have global objectives

CHARACTERISTICS OF THE INPUT DATA ARE IMPORTANT

- Type of proximity or density measure
  - Central to clustering
  - Depends on data and application
- Data characteristics that affect proximity and/or density are
  - Dimensionality
    - Sparseness
  - Attribute type
  - Special relationships in the data
    - For example, autocorrelation
  - Distribution of the data
- Noise and Outliers
  - Often interfere with the operation of the clustering algorithm
- Clusters of differing sizes, densities, and shapes

CLUSTERING ALGORITHMS

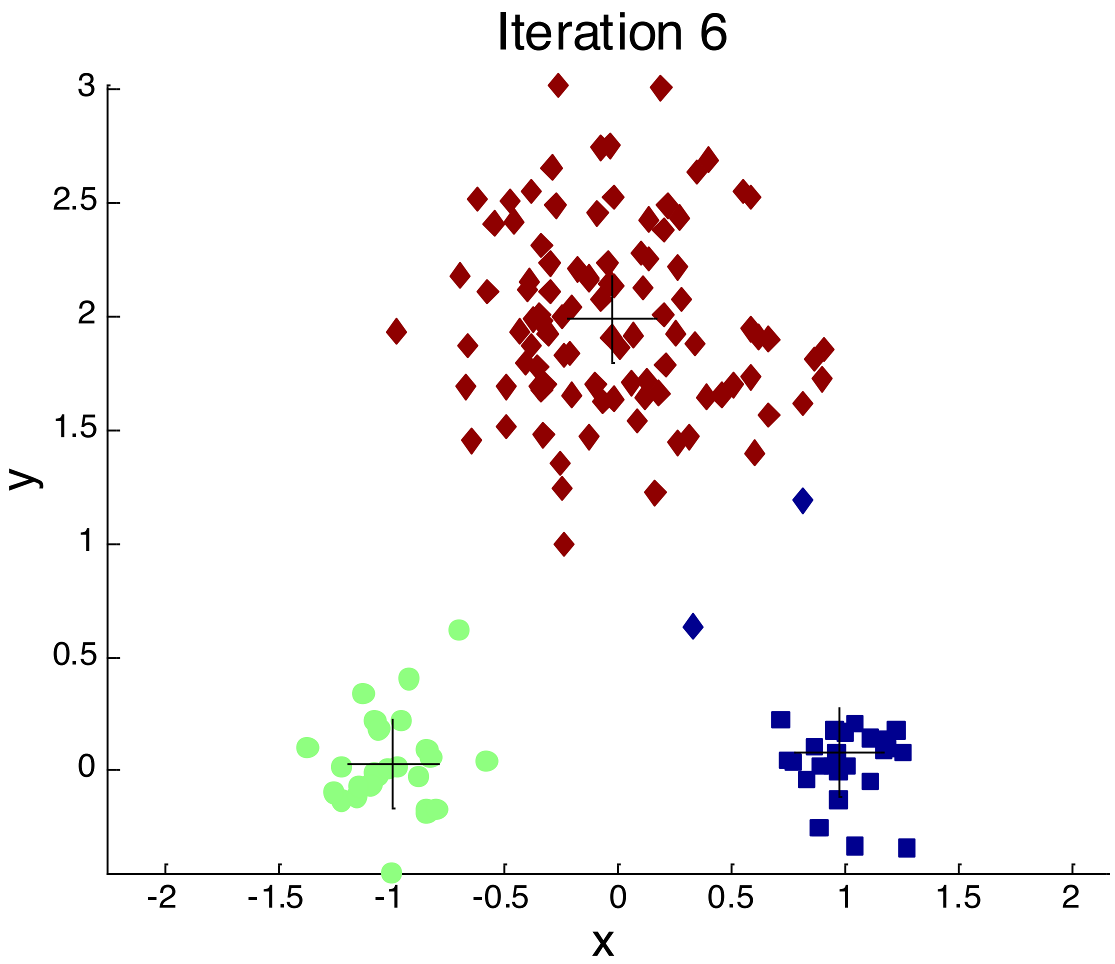
- K-means and its variants
- Hierarchical clustering
- Density-based clustering

K-MEANS CLUSTERING

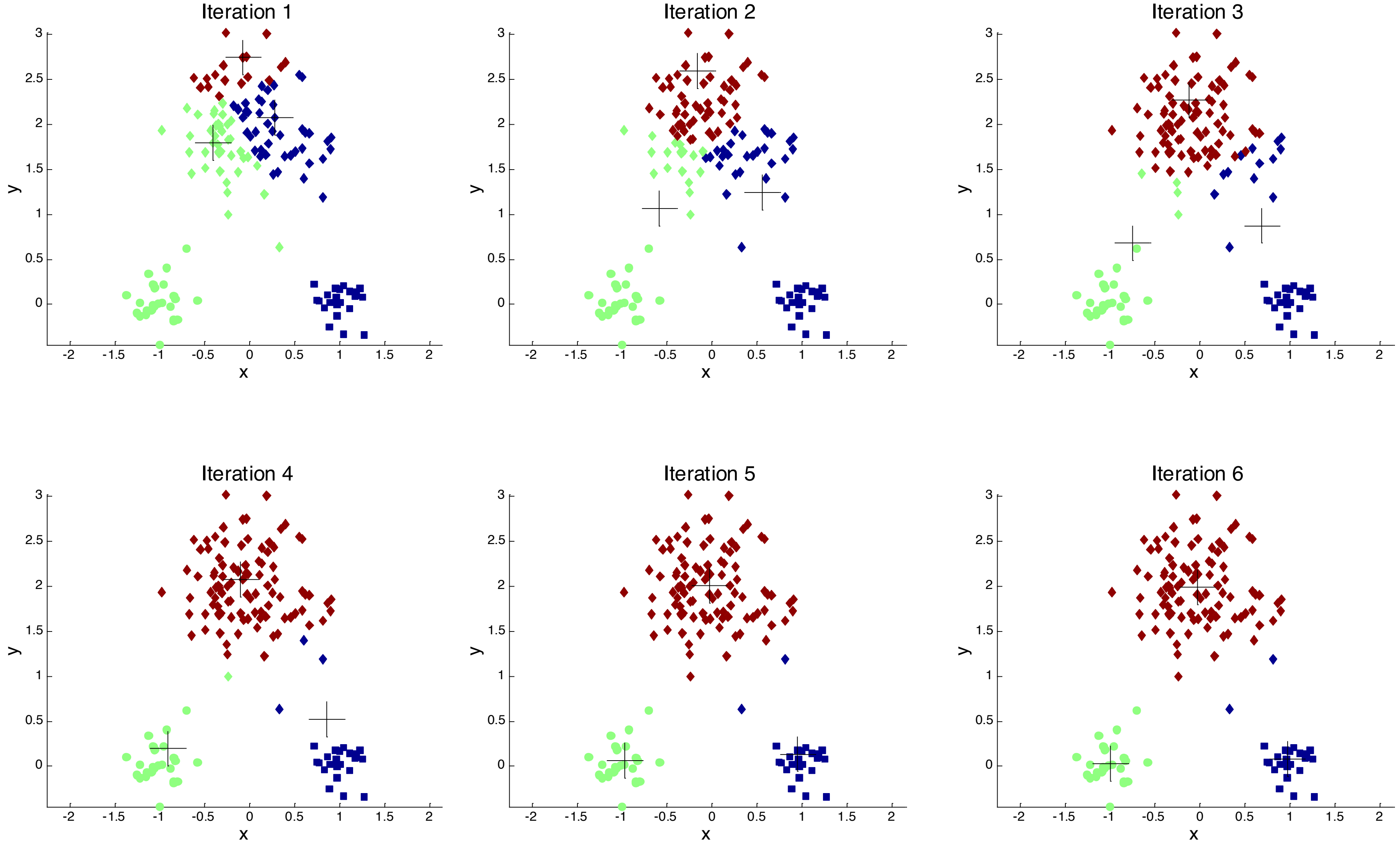
- Partitional clustering approach
- Number of clusters, K, must be specified
- Each cluster is associated with a centroid
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

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

EXAMPLE OF K-MEANS CLUSTERING



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

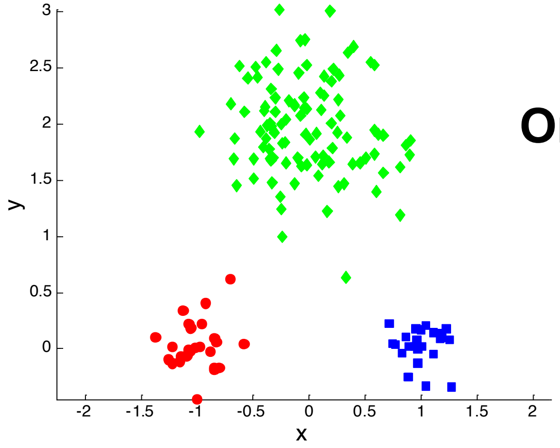
K-MEANS OBJECTIVE FUNCTION

- 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.

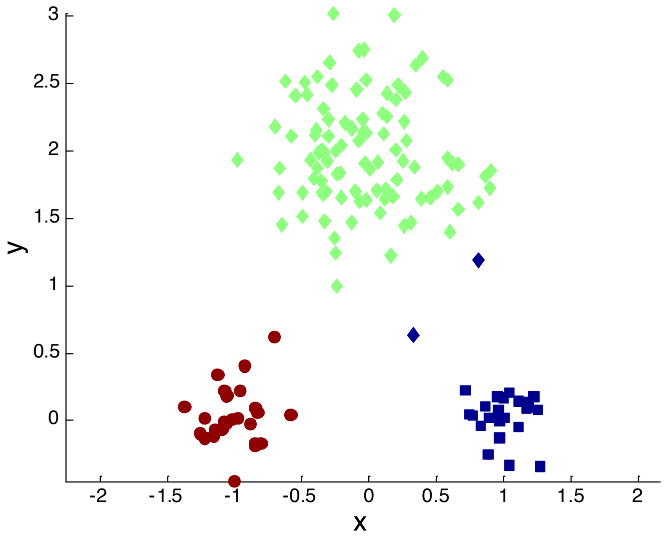
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - 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.

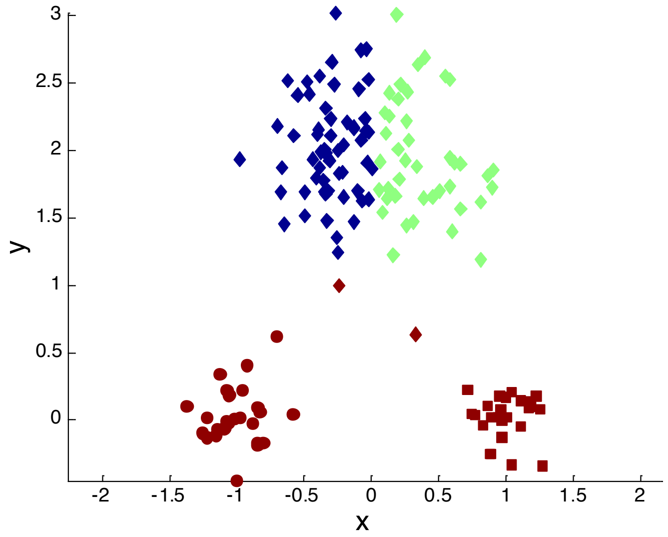
TWO DIFFERENT K-MEANS CLUSTERING



Original Points

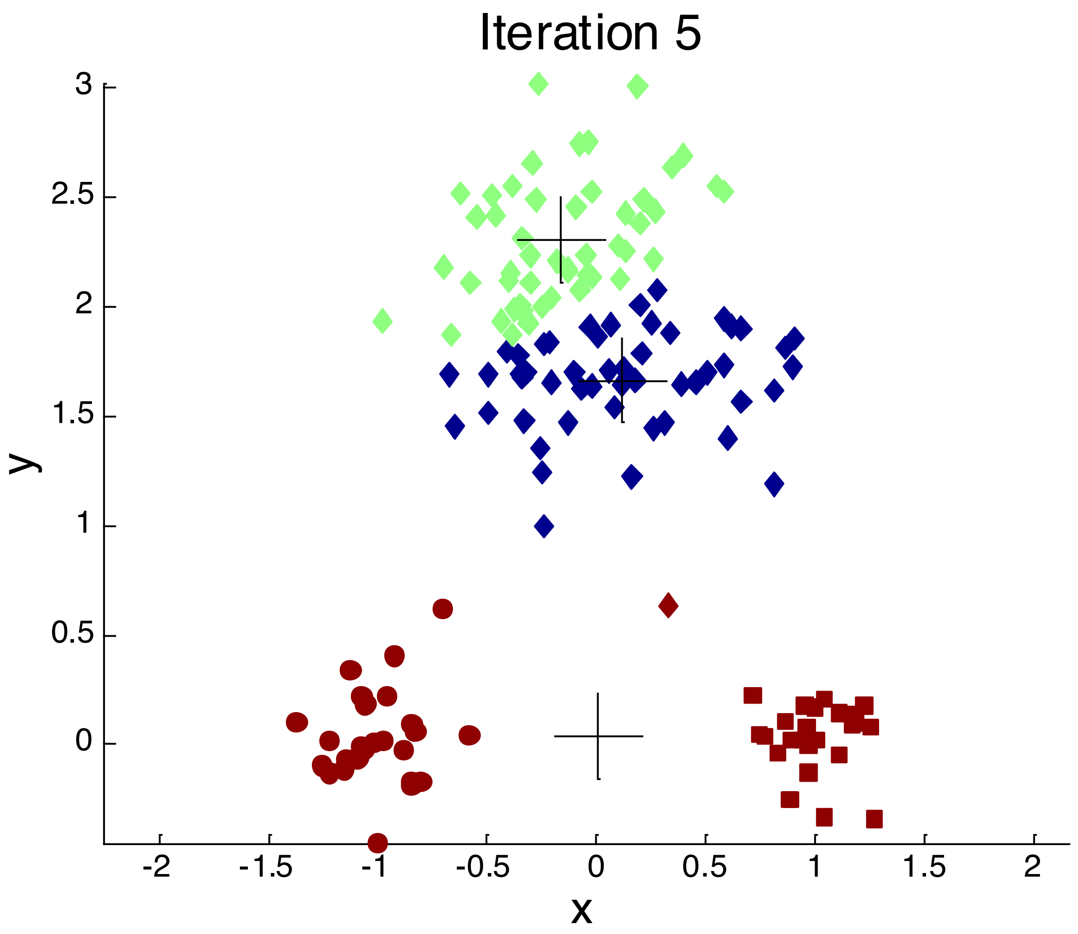


Optimal Clustering

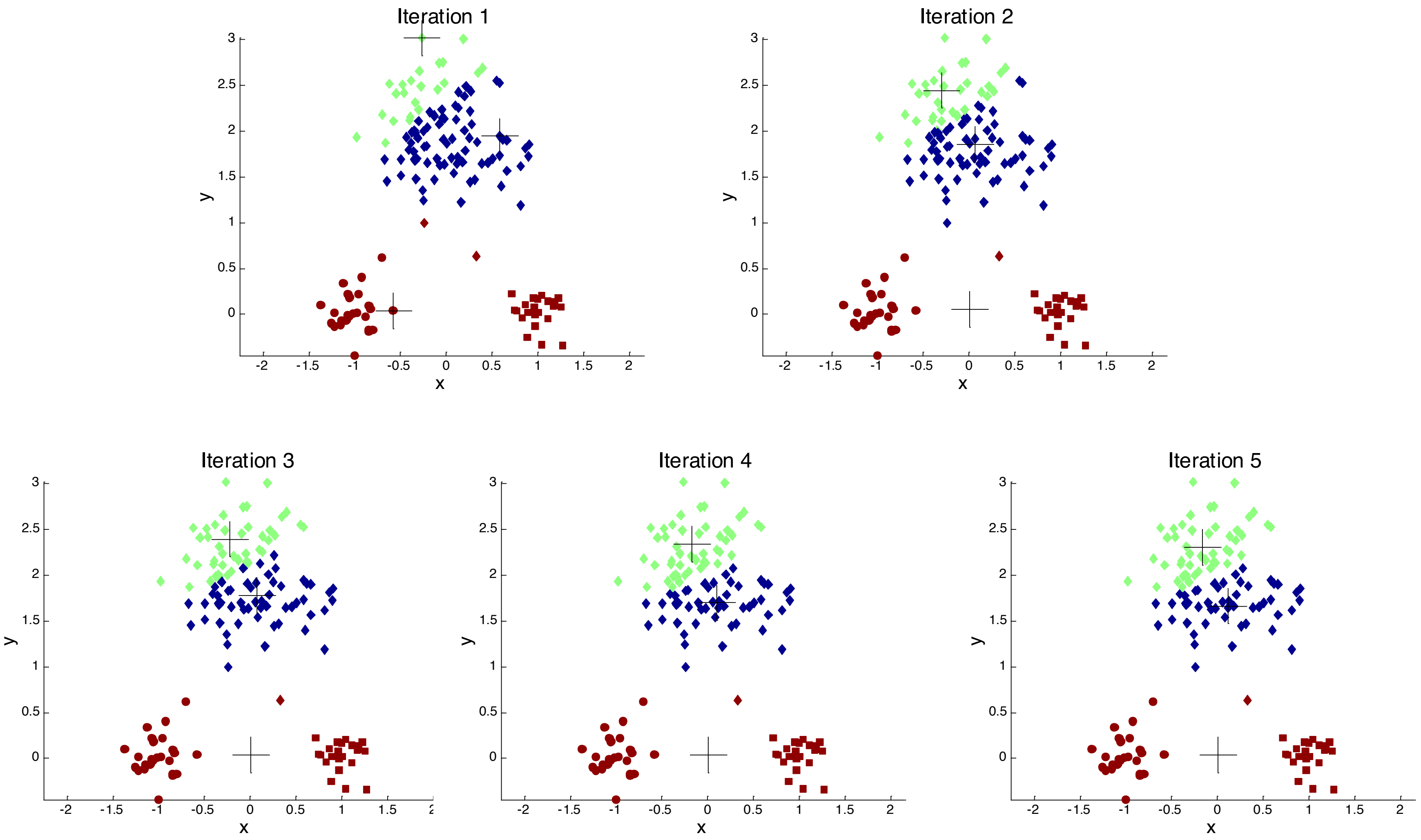


Sub-optimal Clustering

IMPORTANCE OF CHOOSING INITIAL CENTROIDS …



IMPORTANCE OF CHOOSING INITIAL CENTROIDS …



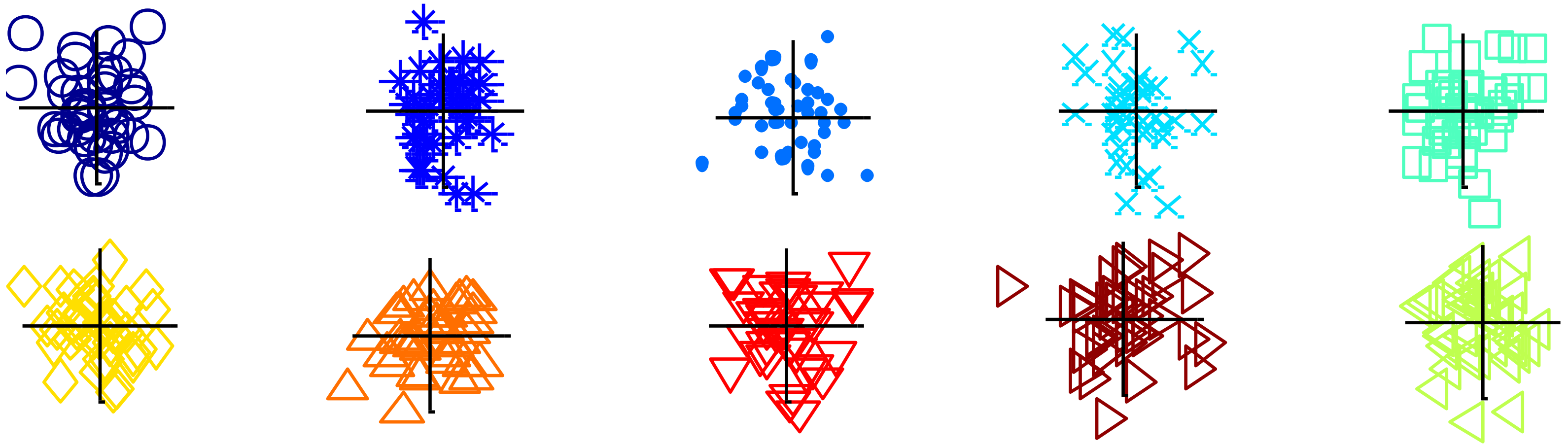
PROBLEMS WITH SELECTING INITIAL POINTS

- 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

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

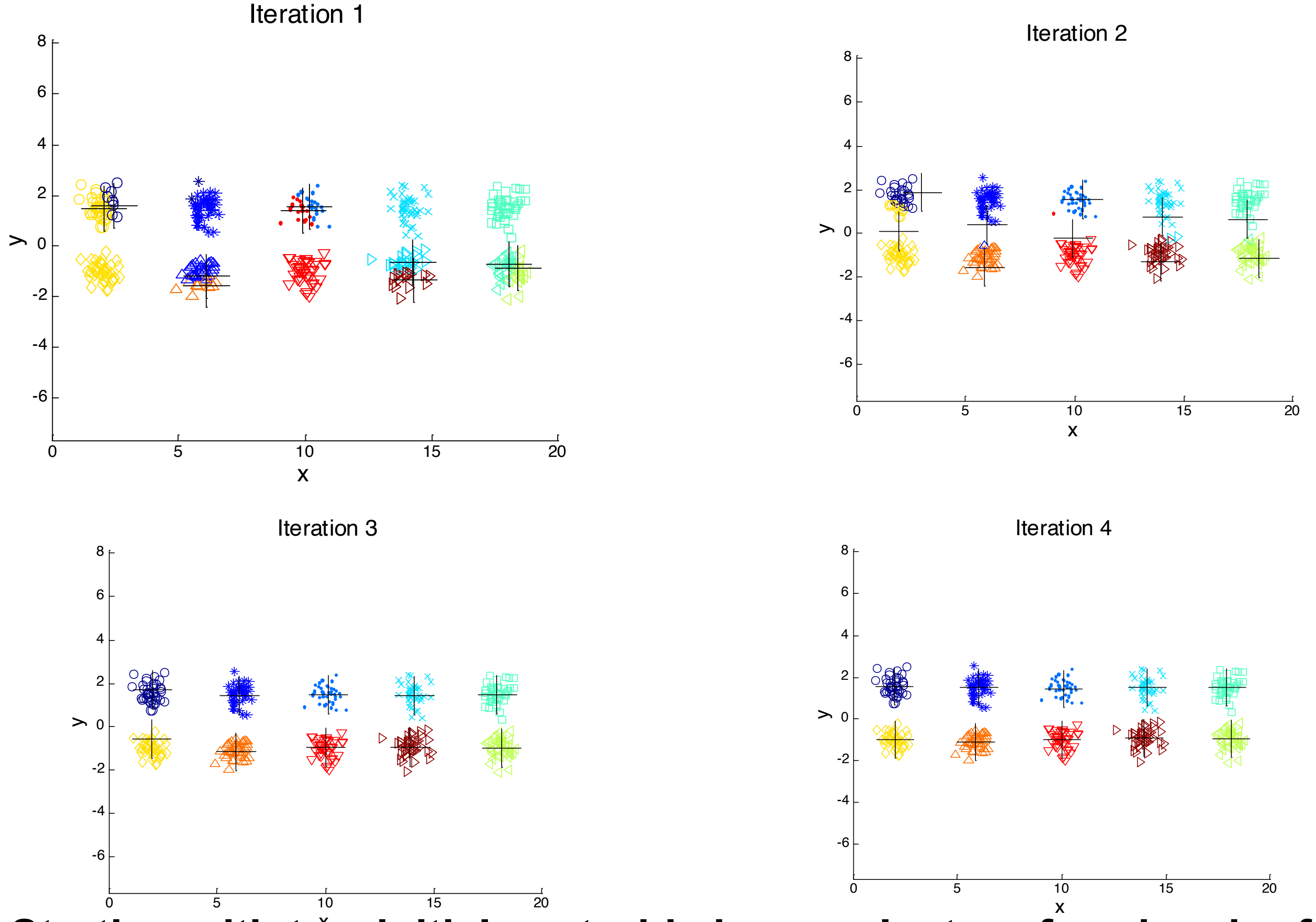
  - For example, if K = 10, then probability = $10!/10^{10}$ = 0.00036

10 CLUSTERS EXAMPLE



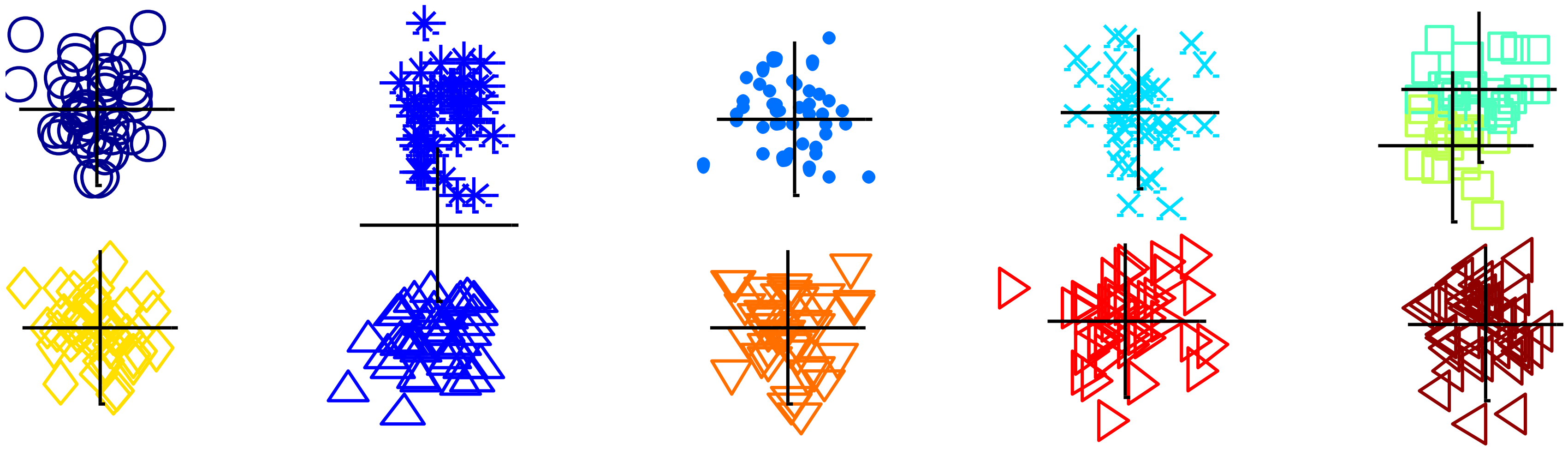
Starting with two initial centroids in one cluster of each pair of clusters

10 CLUSTERS EXAMPLE



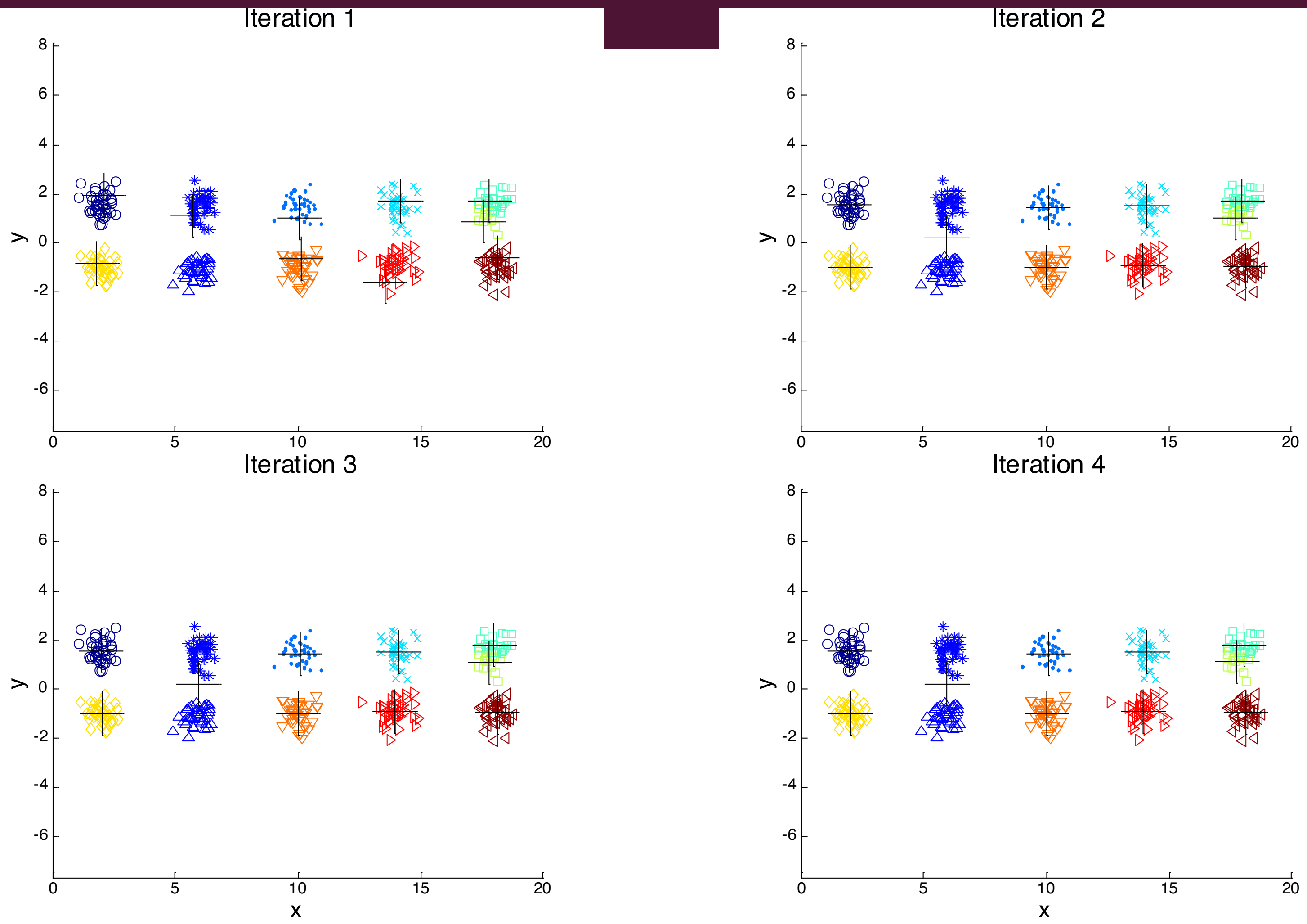
Starting with two initial centroids in one cluster of each pair of clusters

10 CLUSTERS EXAMPLE



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

10 CLUSTERS EXAMPLE



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

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

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

Select an initial point at random to be the first centroid

For k – 1 steps

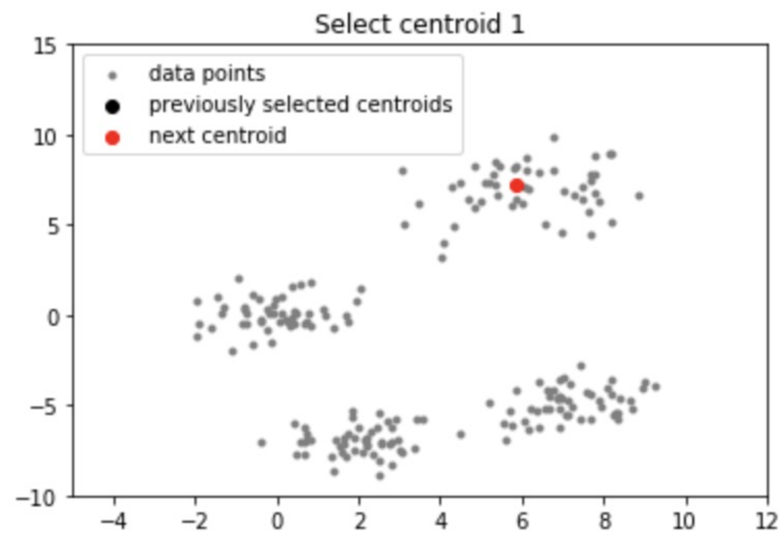
For each of the N points, x_i, $1 \le i \le N$, find the minimum squared distance to the currently selected centroids,

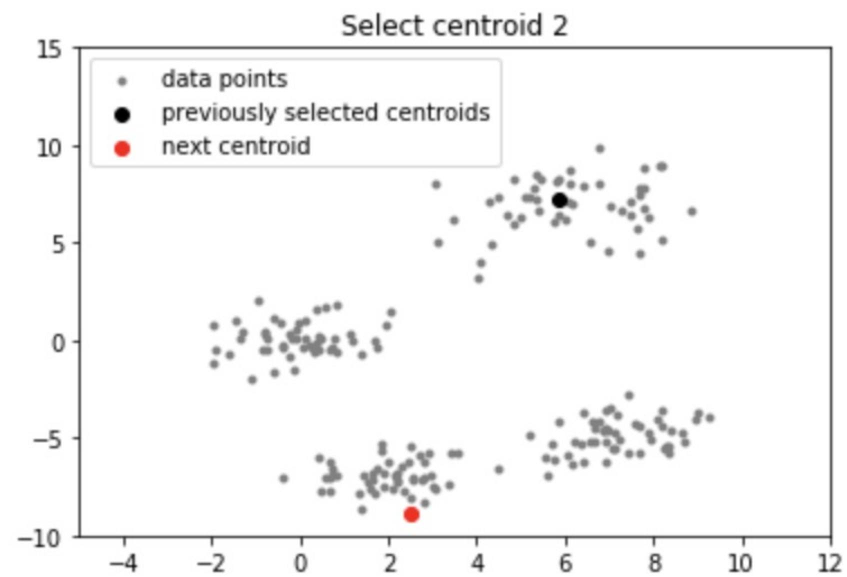
$C_1, \ldots, C_j$, $1 \le j < k$, i.e., $\min_j d^2(C_j, x_i)$

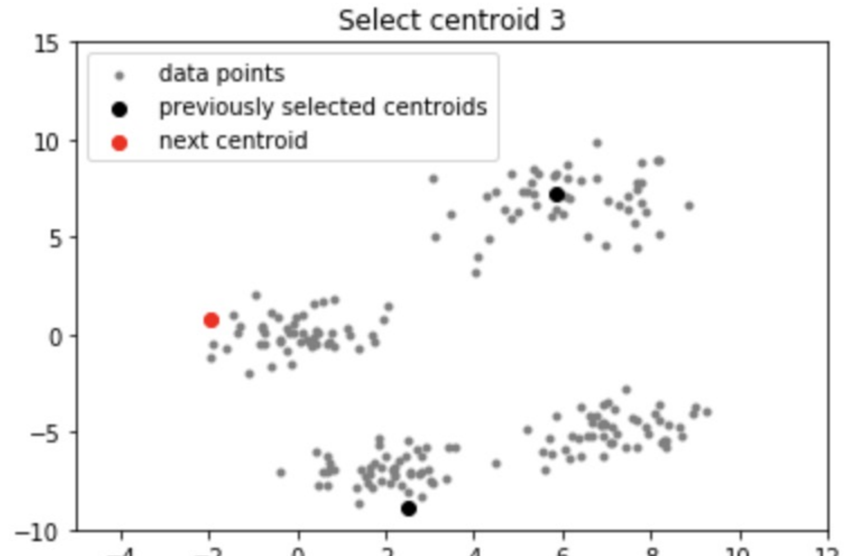
Randomly select a new centroid by choosing a point with probability proportional to $\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$

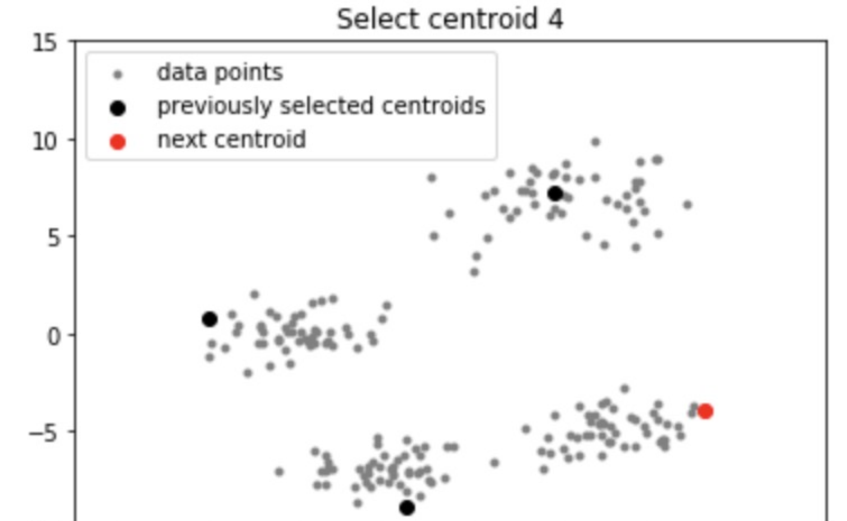
End For

K-MEAN++


Select centroid 1
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12


Select centroid 2
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12


Select centroid 3
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10

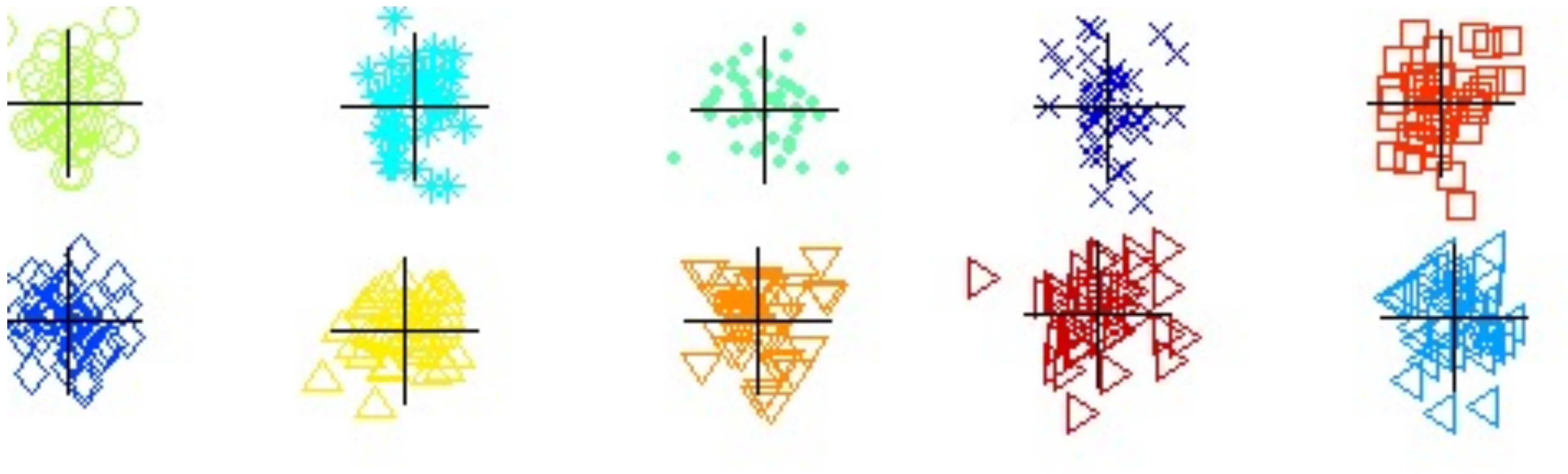

Select centroid 4
data points
previously selected centroids
next centroid
15
10
5
0
−5

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

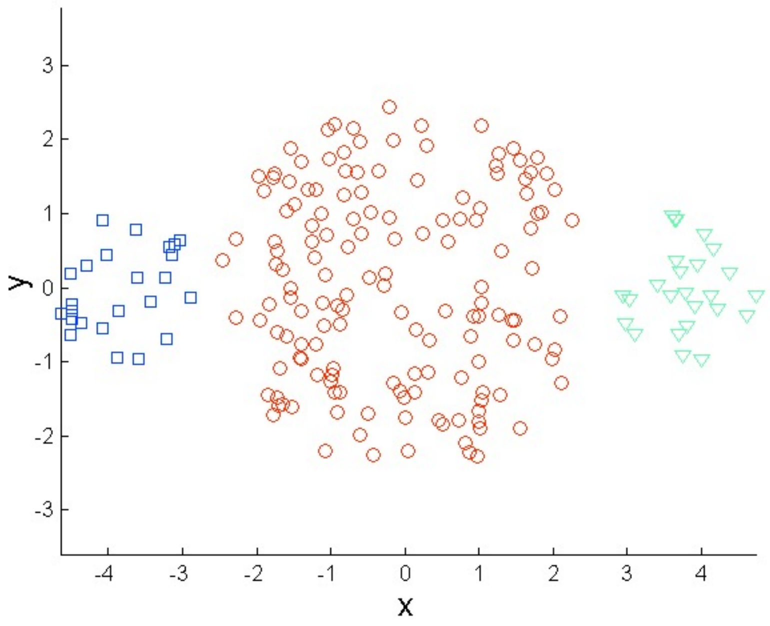


LIMITATIONS OF K-MEANS

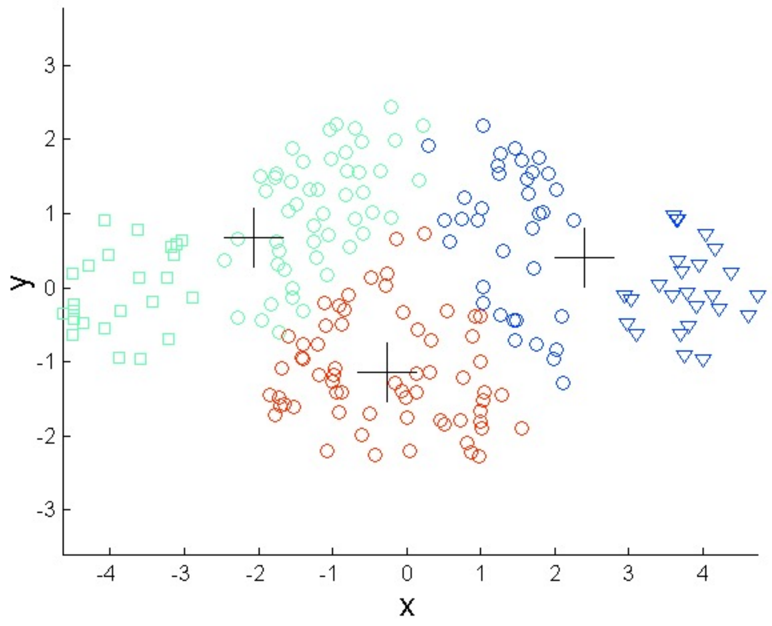
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

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

LIMITATIONS OF K-MEANS: DIFFERING SIZES

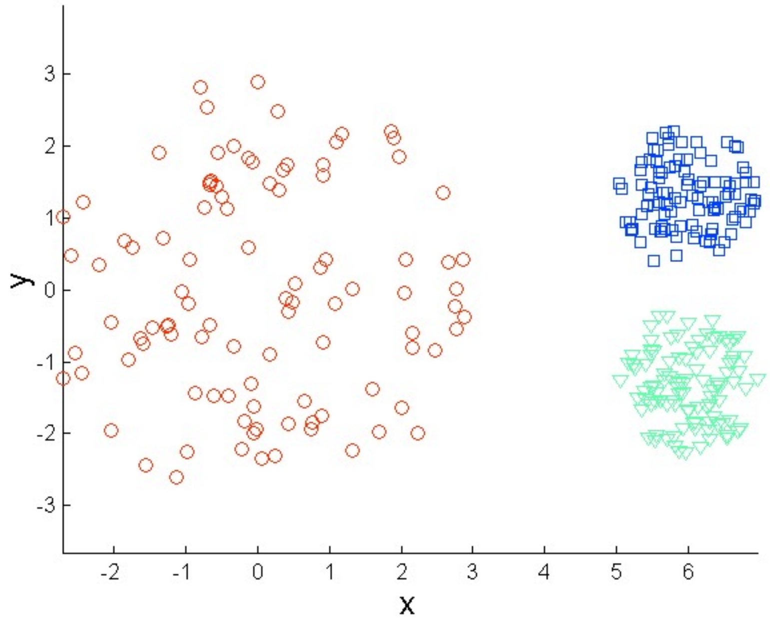


Original Points

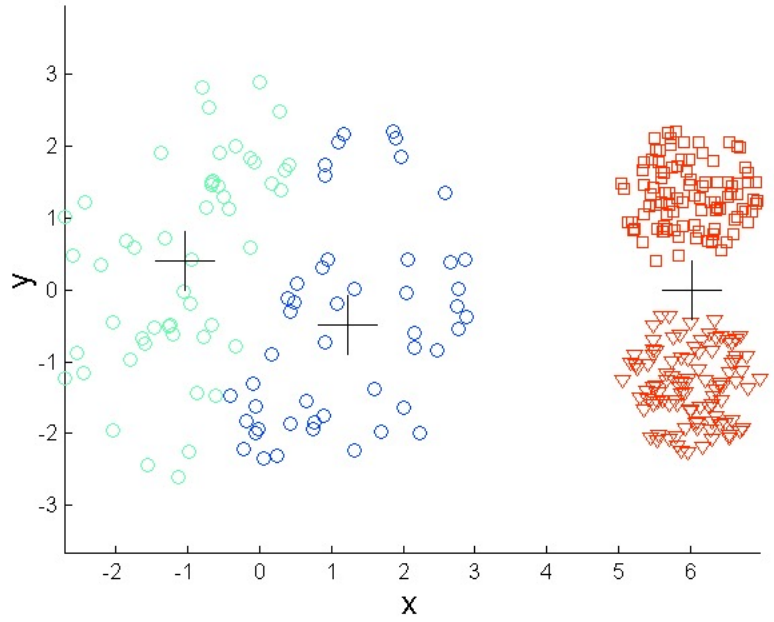


K-means (3 Clusters)

LIMITATIONS OF K-MEANS: DIFFERING DENSITY

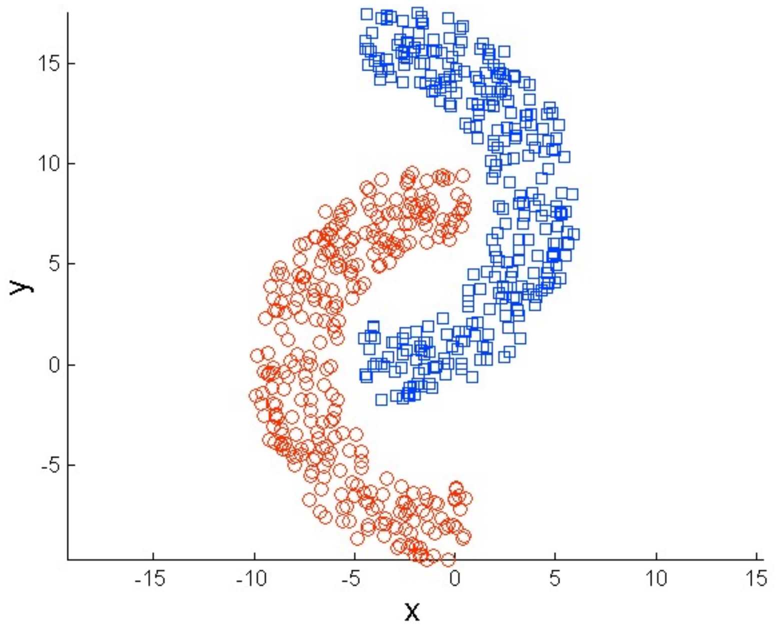


Original Points

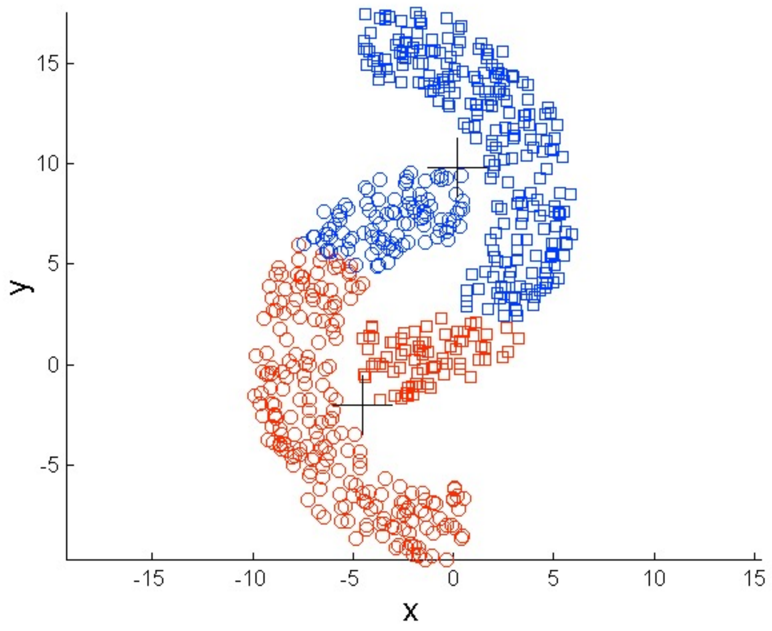


K-means (3 Clusters)

LIMITATIONS OF K-MEANS: NON-GLOBULAR SHAPES

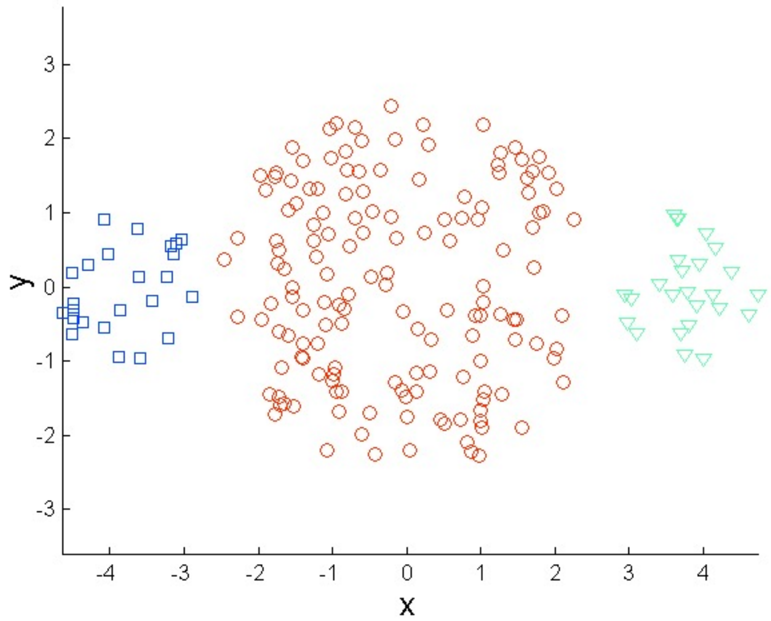


Original Points

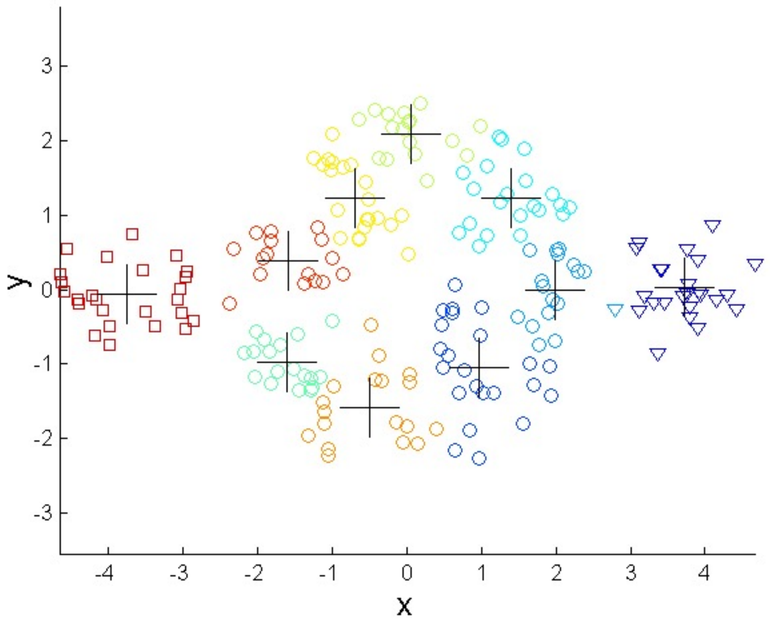


K-means (2 Clusters)

OVERCOMING K-MEANS LIMITATIONS



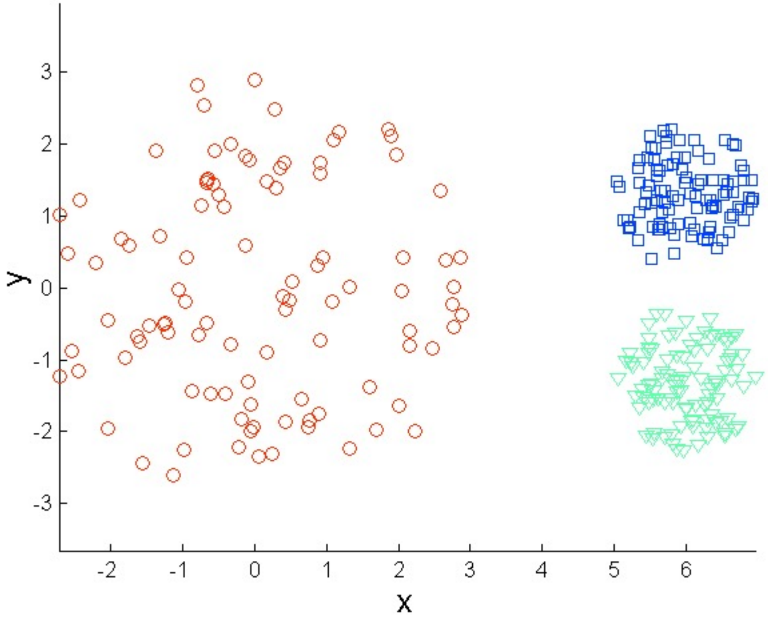
Original Points



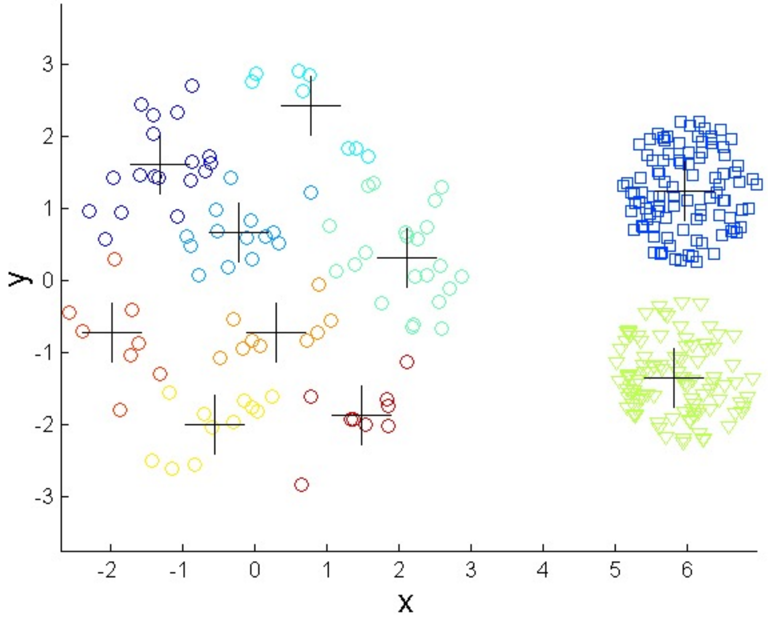
K-means Clusters

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



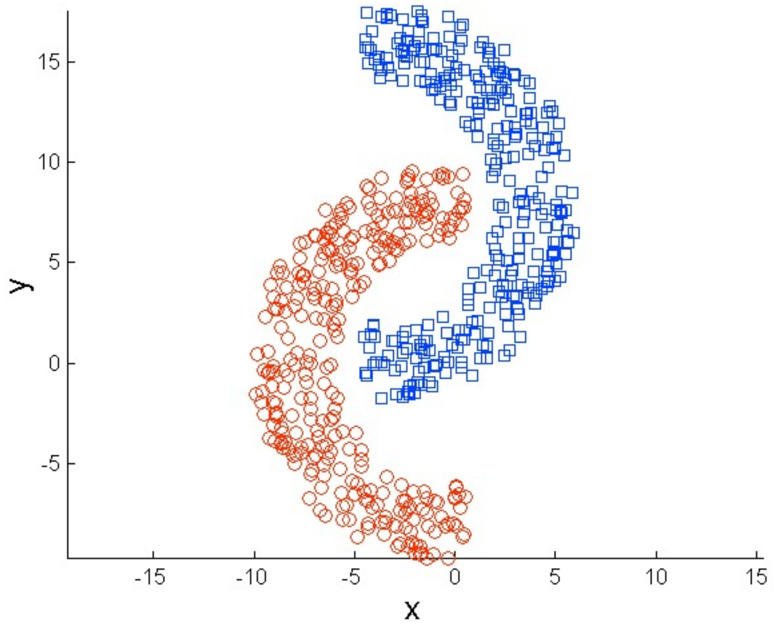
Original Points



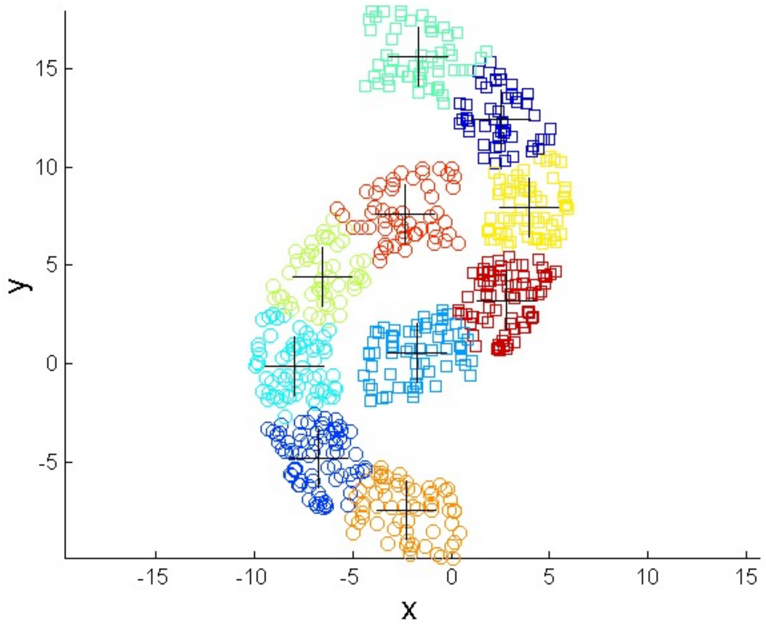
K-means Clusters

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.