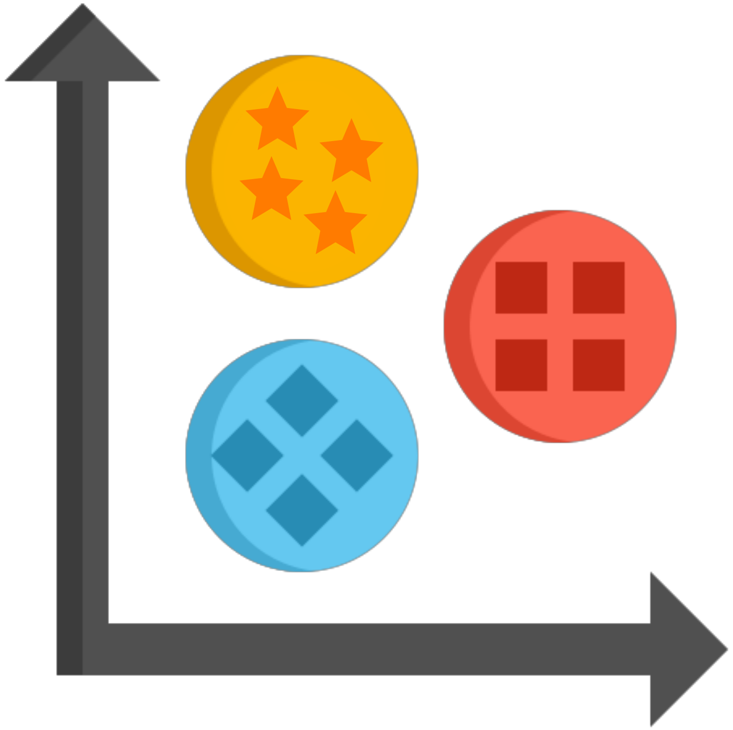
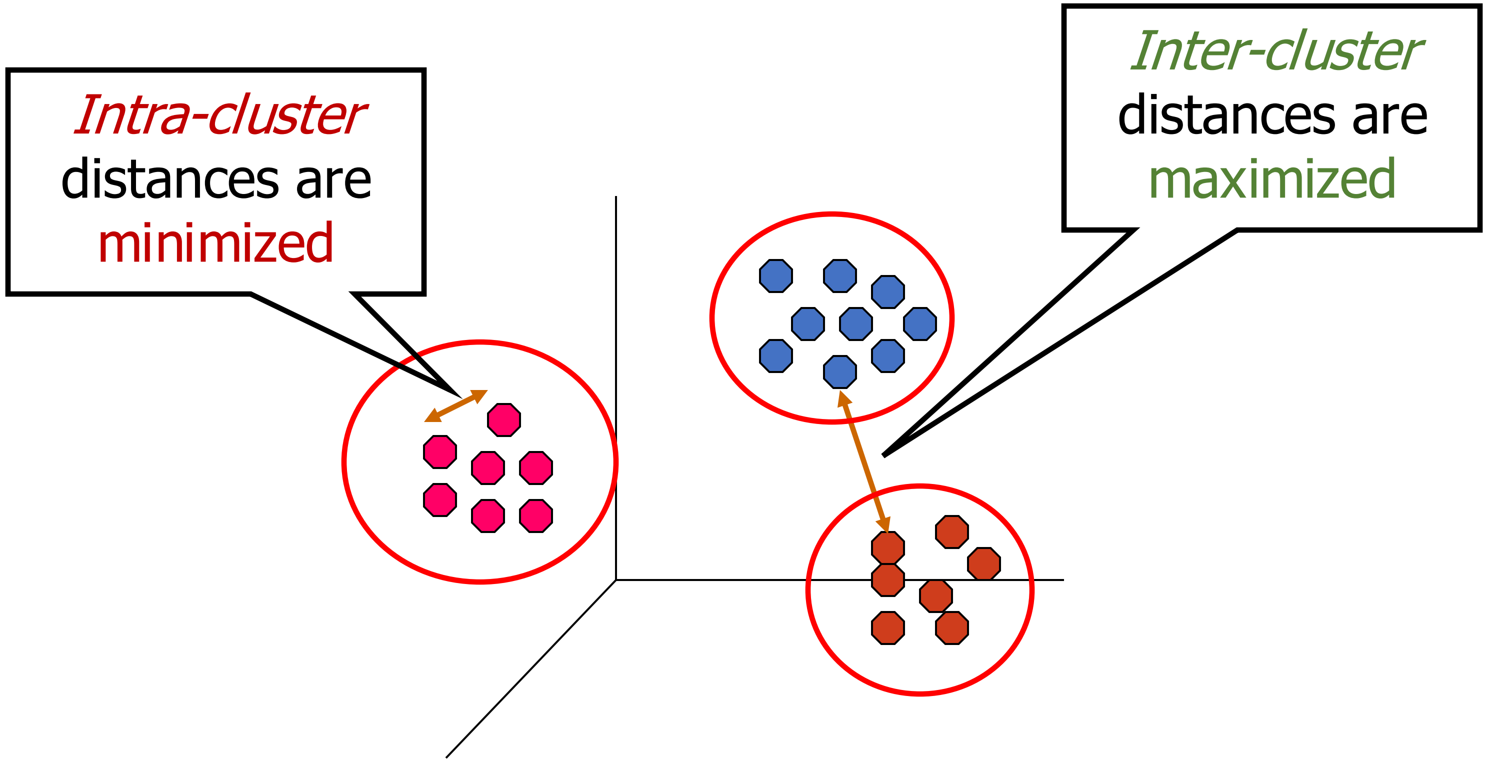
UNIVERSITÀ DEL SALENTO

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) to one another and **different** from (or unrelated to) the objects in **other groups**



Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Types of Clusterings

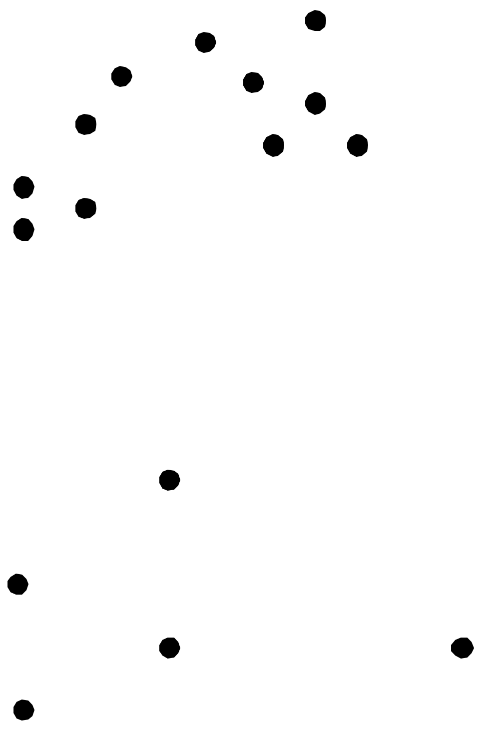
A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

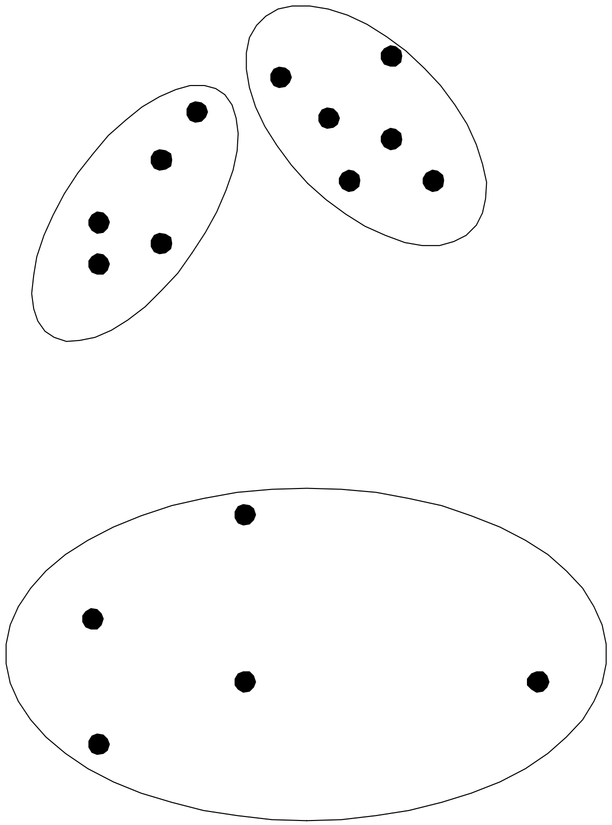
- **Partitional Clustering**: A division of data objects into non-overlapping subsets (clusters)

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

Partitional Clustering

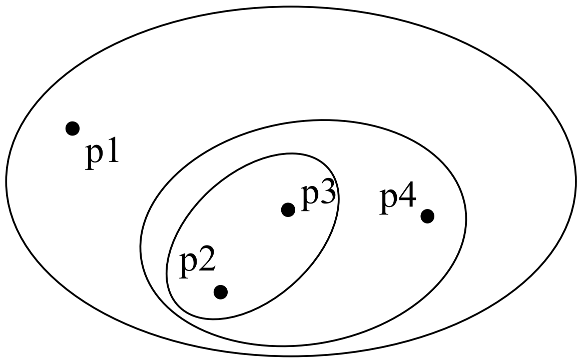


Original Points

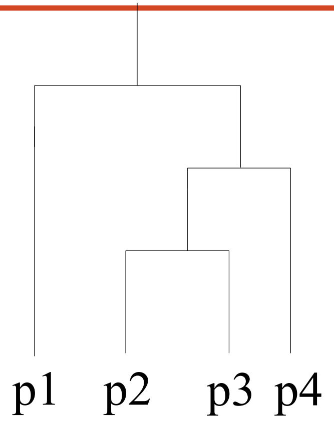


A Partitional Clustering

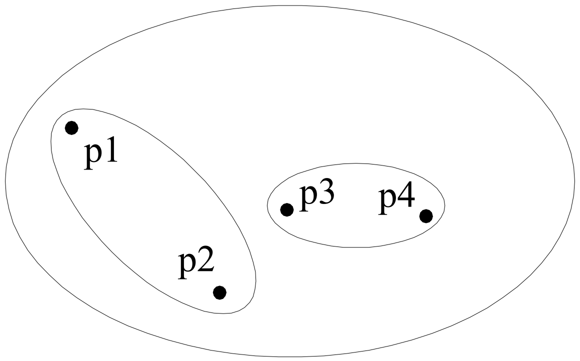
Hierarchical Clustering



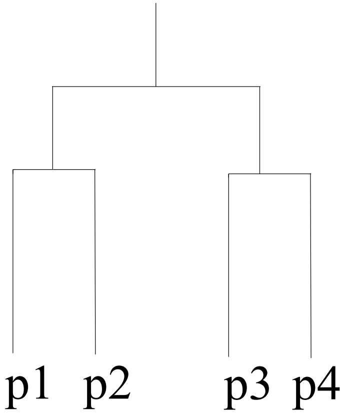
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Other Distinctions Between Sets of Clusters

Exclusive versus **non-exclusive** versus **Fuzzy**

- **Overlapping** or non-exclusive clusterings, points may belong to multiple clusters:
  - Can belong to multiple classes or could be 'border' points
- **Fuzzy** clustering (one type of non-exclusive)
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
- Probabilistic clustering has similar characteristics

Partial versus **Complete**

- In some cases, we only want to cluster some of the data

Types of Clusters

a) Well-separated clusters

b) Prototype-based clusters

c) Contiguity-based clusters

d) Density-based clusters

e) Conceptual clusters

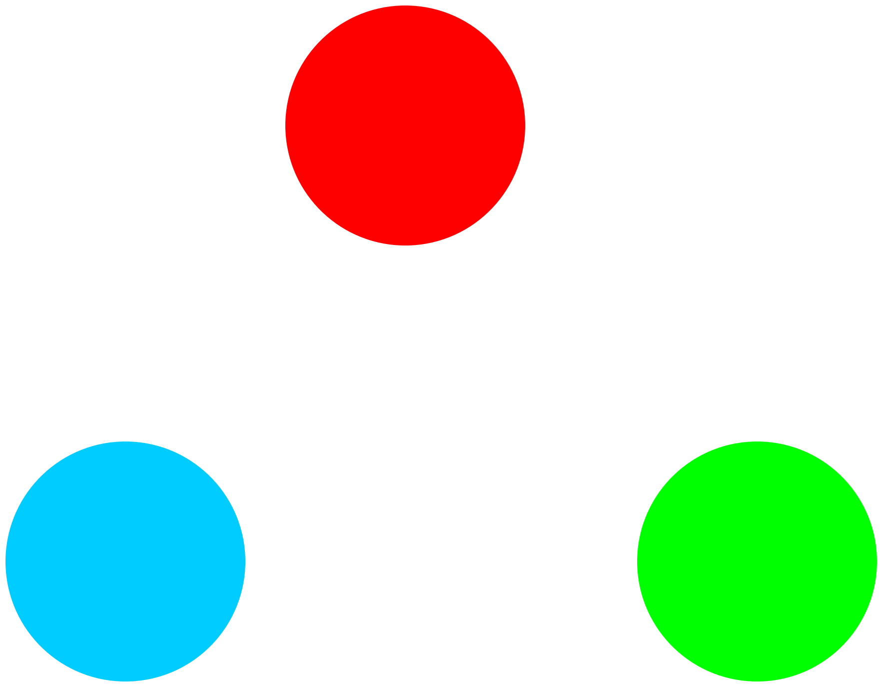
Types of Clusters: Well-Separated

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

Clusters that are quite far from each other.

The distance between any two points in different groups is larger than the distance between any two points within a group.

Well-separated clusters do not need to be **globular**, but can have any shape.

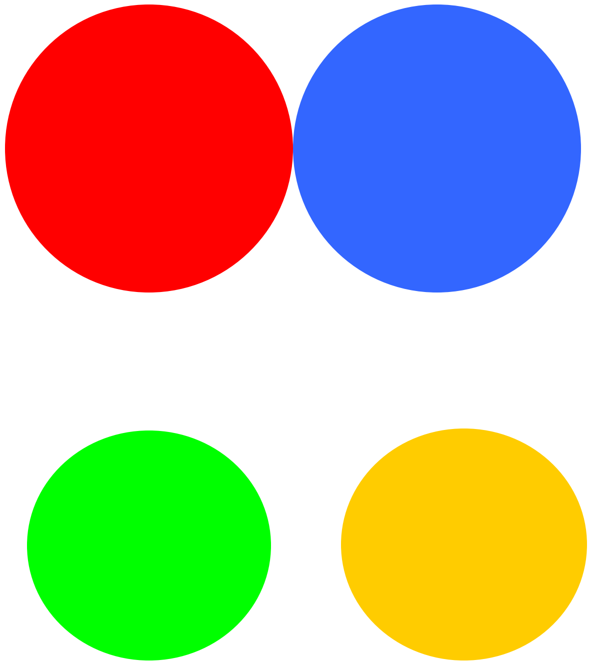


3 well-separated clusters

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

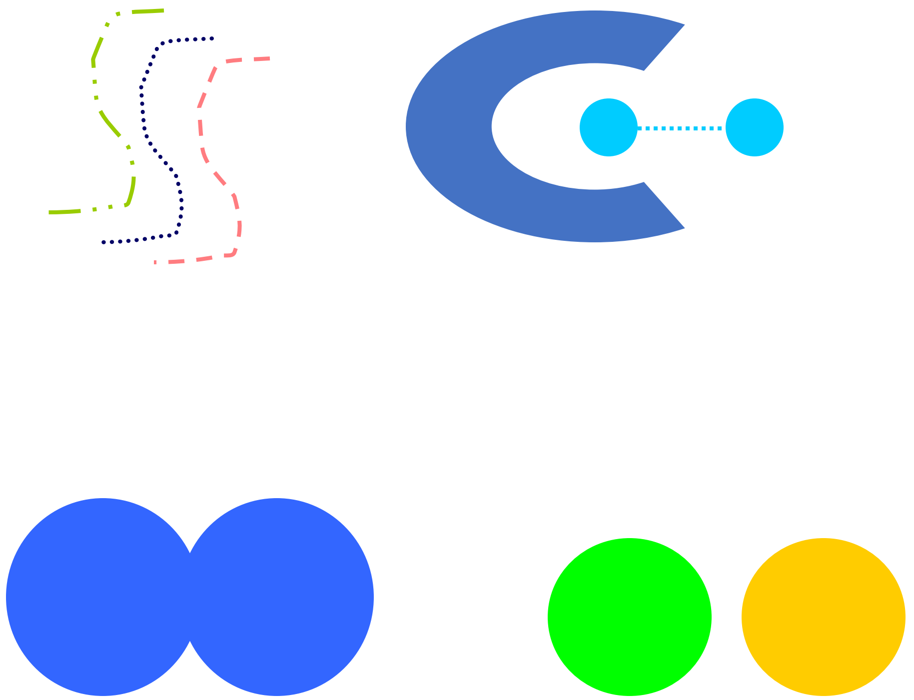
The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most "representative" point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

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

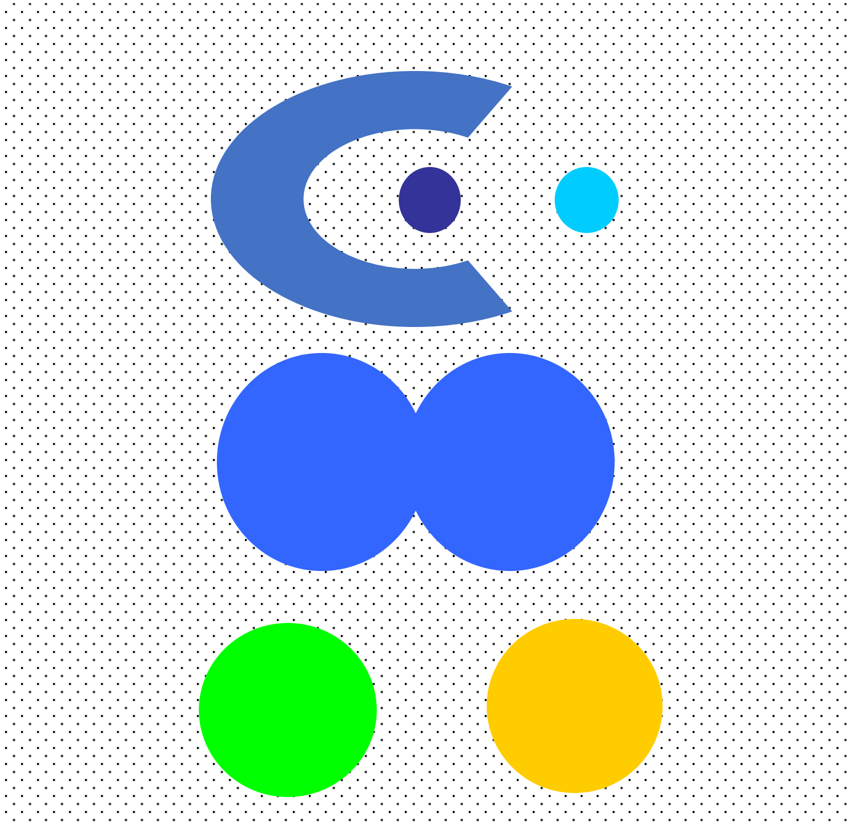


8 contiguous clusters

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.

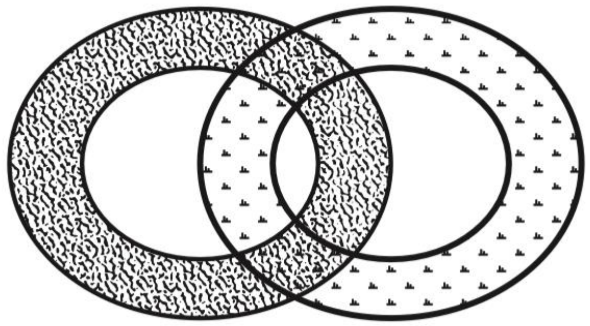
Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

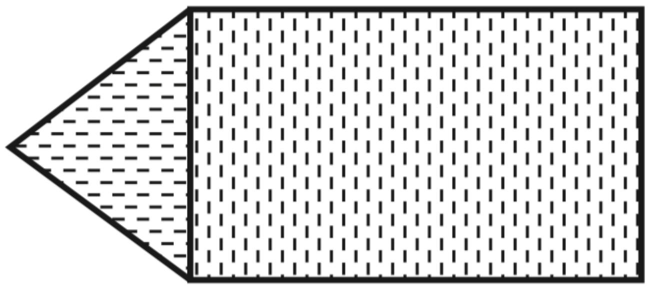
Types of Clusters: Conceptual cluster

Conceptual cluster: More generally, we can define a cluster as a set of objects that share some property.



This definition encompasses all the previous definitions of a cluster

New types of clusters --> clusters shown in figure: a triangular area is adjacent to a rectangular one



A clustering algorithm would need a very specific concept of a cluster to successfully detect these clusters.

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 (autocorrelation)

Noise and Outliers

Clustering Algorithms

K-means

Hierarchical clustering

Density-based clustering

K-means Clustering

Partitional clustering approach

Number of clusters, **K**, <u>must be specified</u>

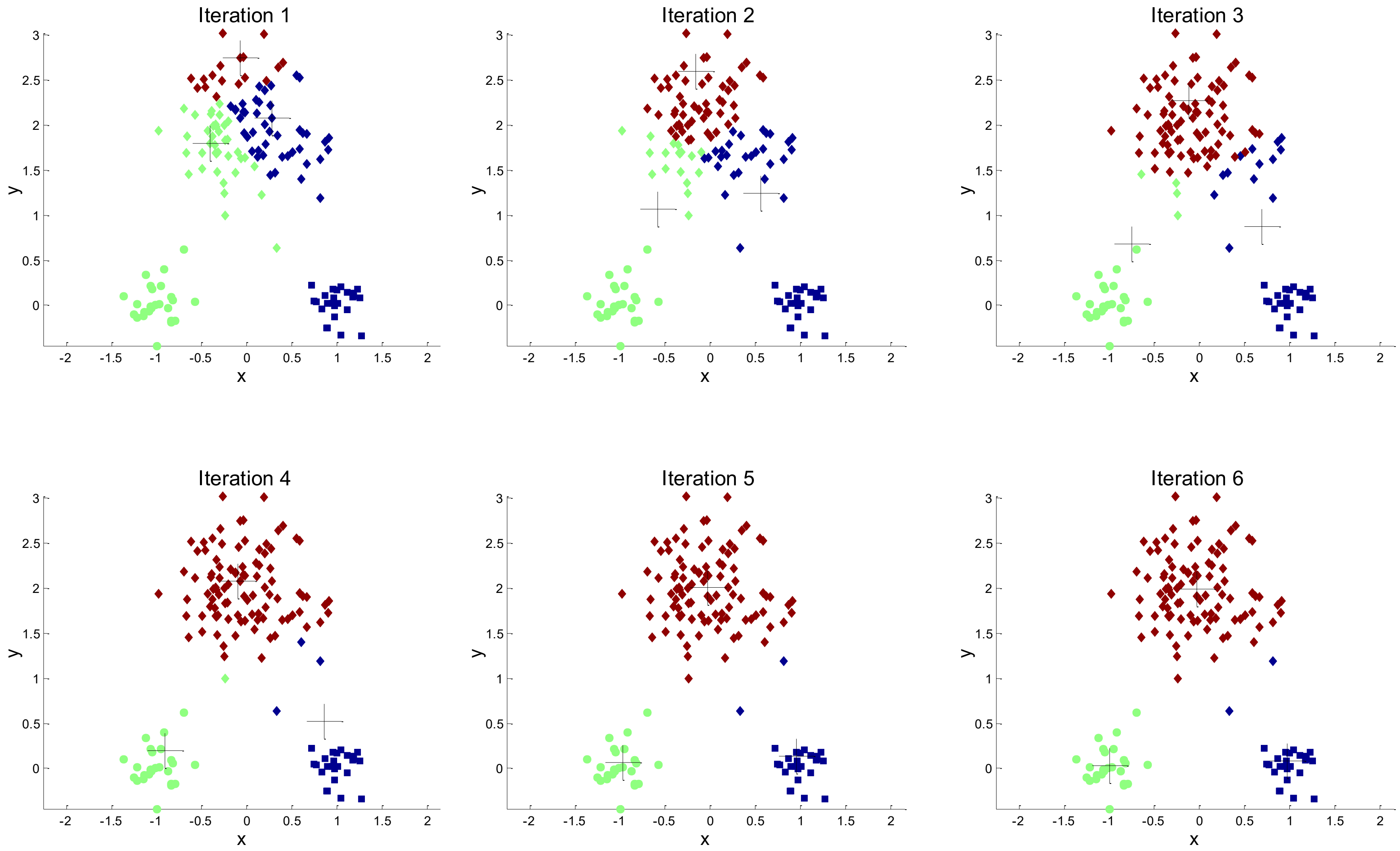
Each cluster is associated with a centroid (center point)

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



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'

K-means Objective Function

A common **objective function** (used with Euclidean distance measure) is

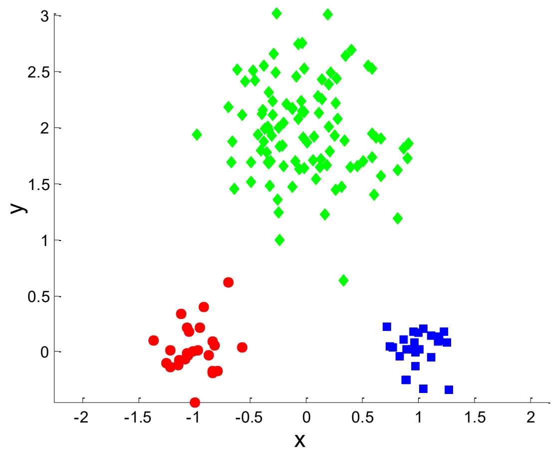
Sum of Squared Error **(SSE)**

- For each point, the error is the distance to the nearest cluster center
- To get SSE, we square these errors and sum them.

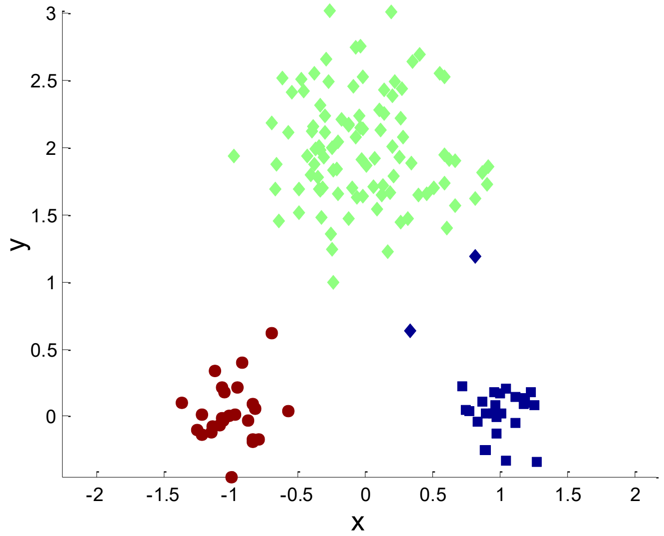
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- *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.

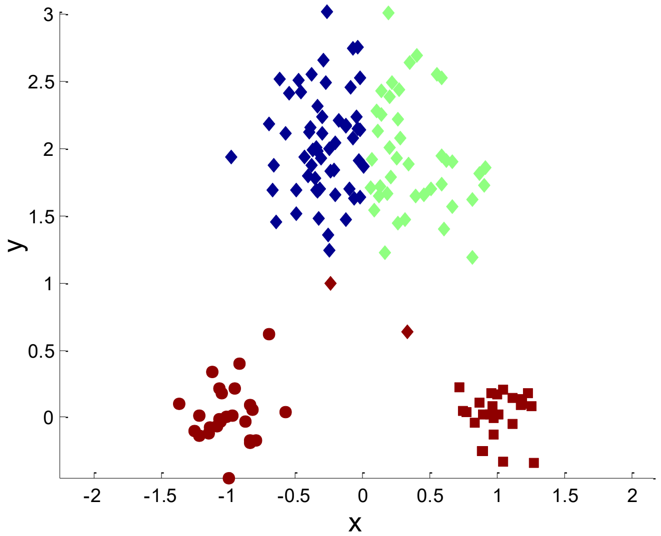
Two different K-means Clusterings



Original Points

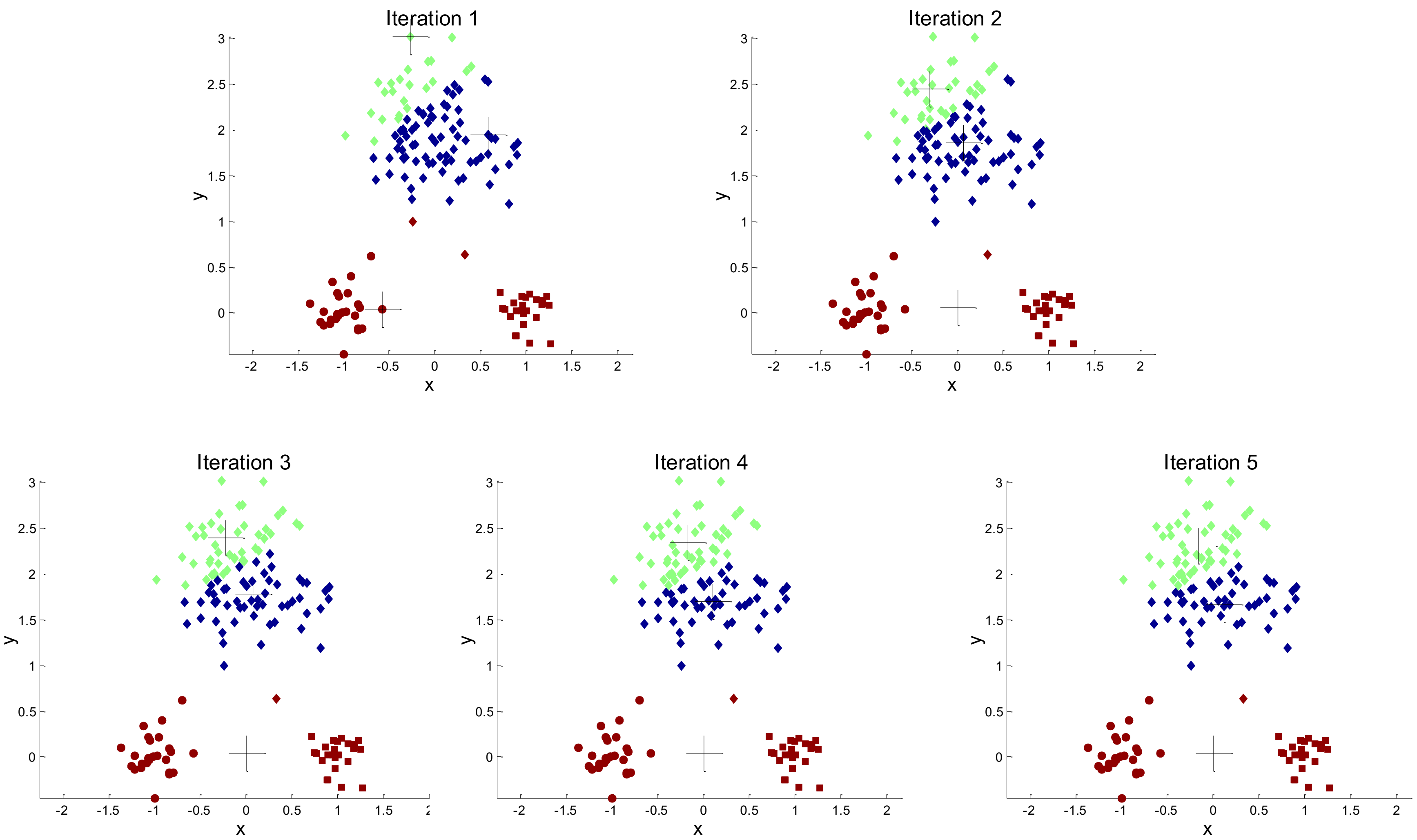


Optimal Clustering



Sub-optimal Clustering

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.

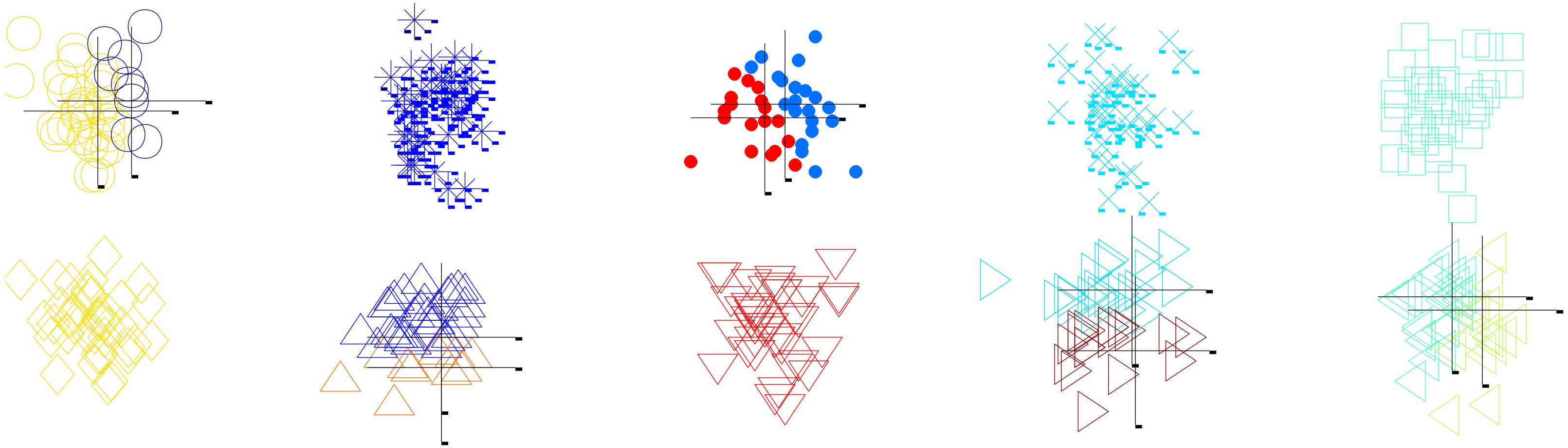
- Chance is relatively small when K is large
- If clusters are the same size, n, then

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

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

Limits in random initialization: 10 Clusters Example

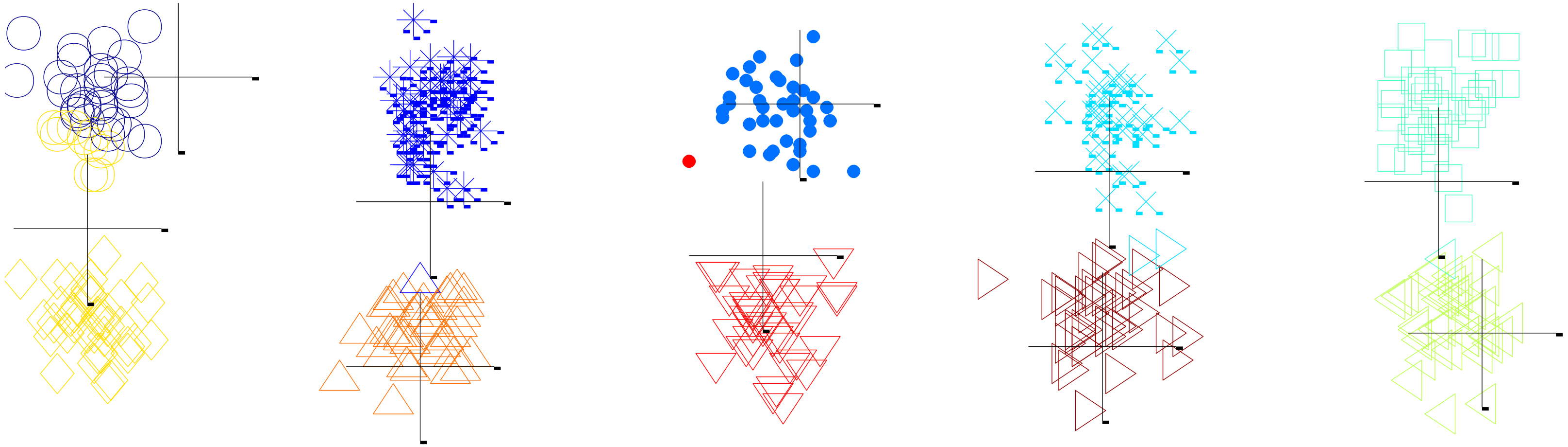
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

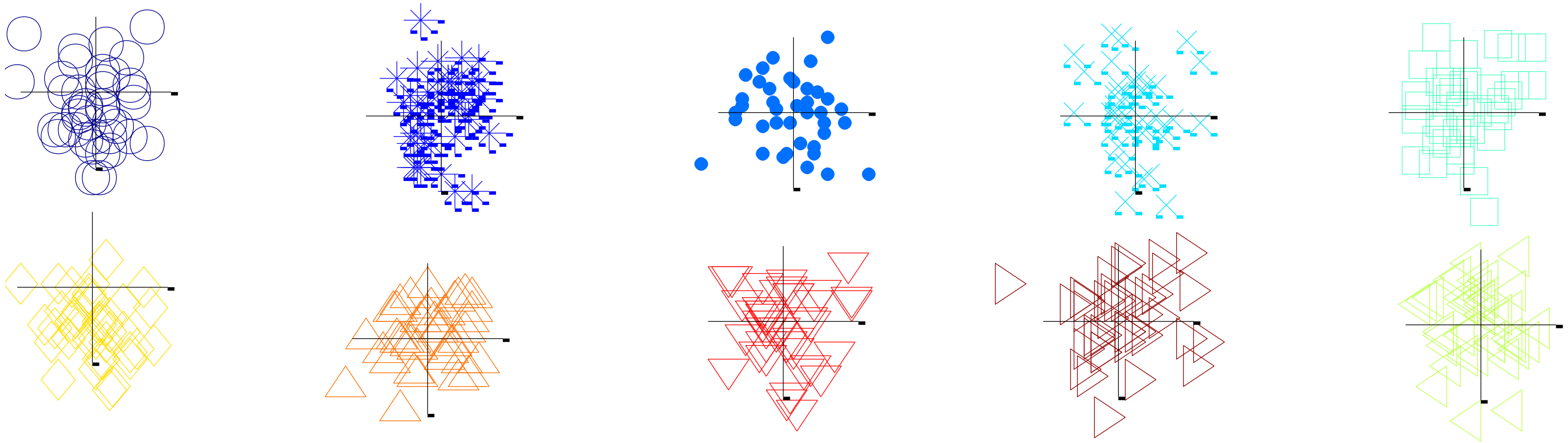
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

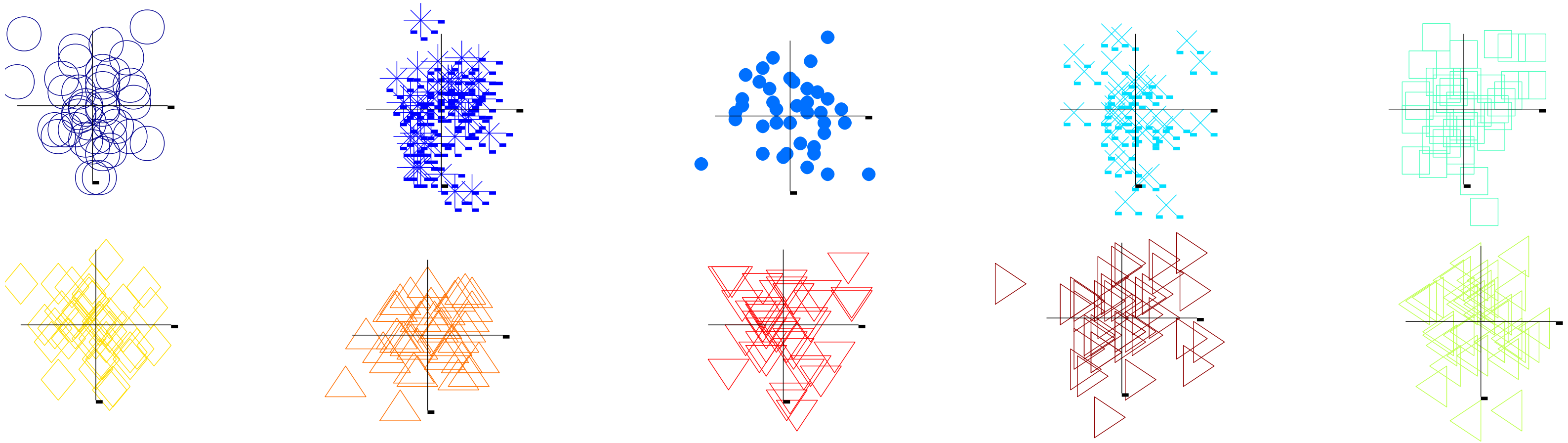
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

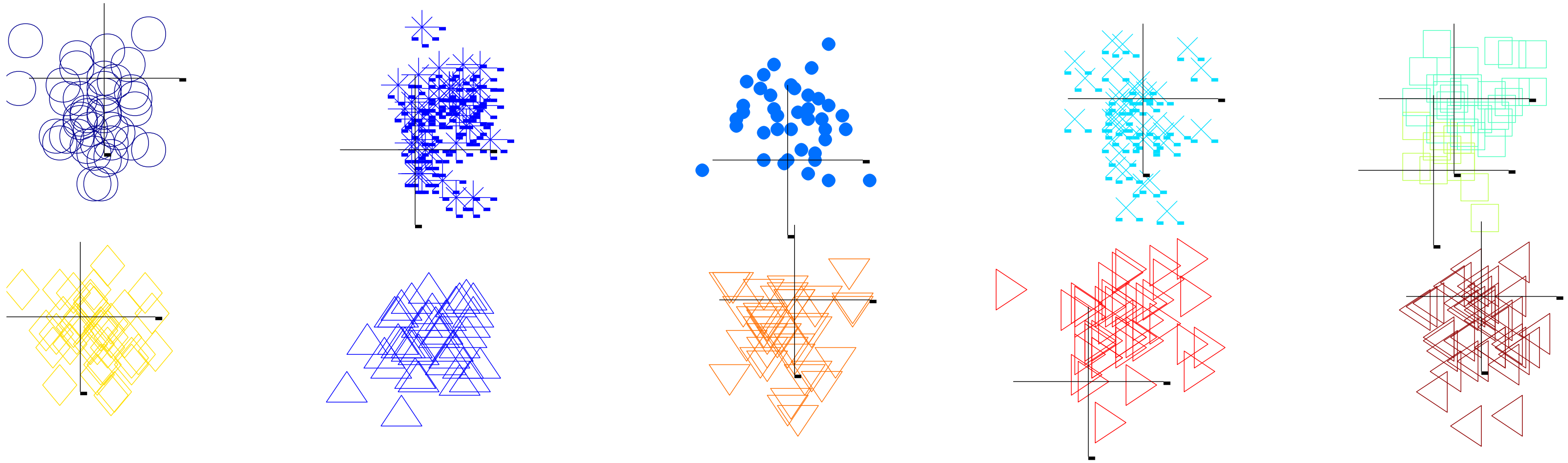
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

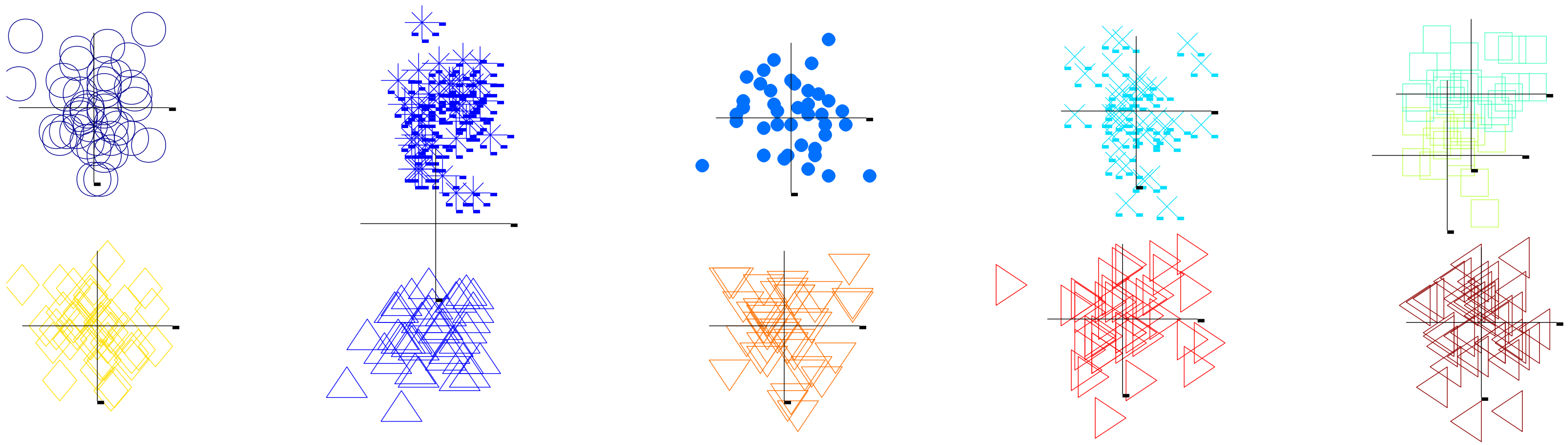
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

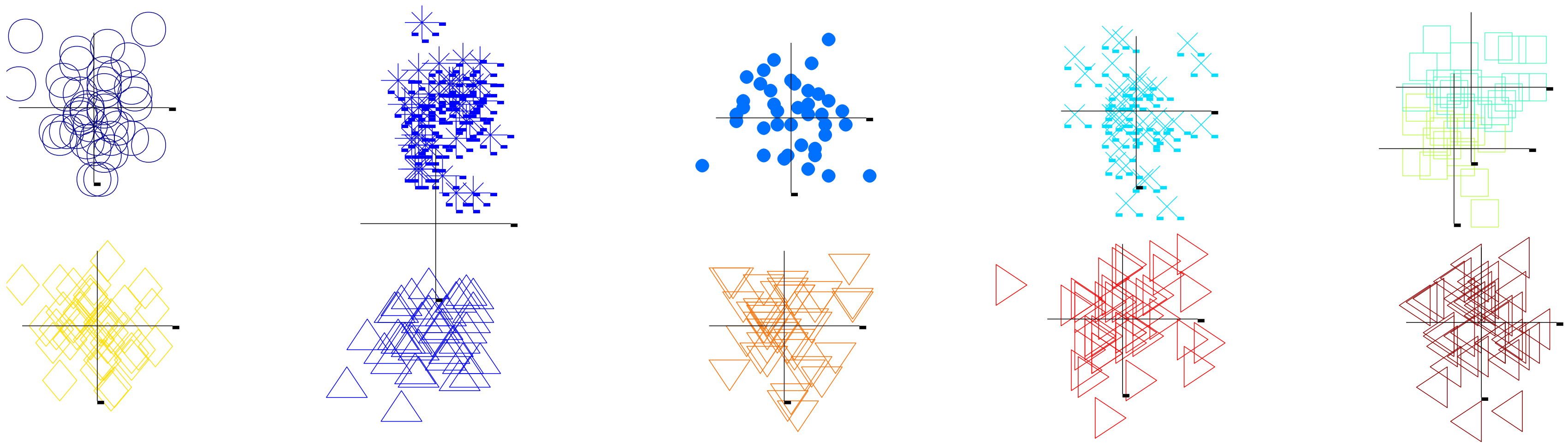
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

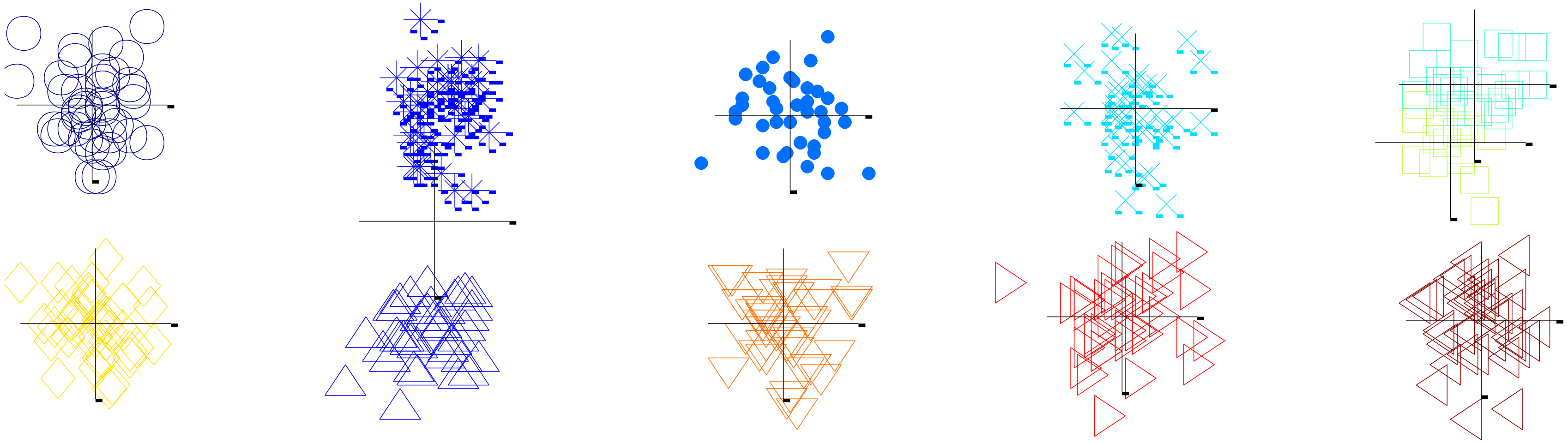
The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

Limits in random initialization: 10 Clusters Example

The data consists of 5 pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair.



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

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++

This approach can be slower than random initialization, but very consistently produces better results in terms of SSE

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

1. Select an initial point at random to be the first centroid
2. For k – 1 steps
3. 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., $\min_j d^2(C_j, x_i)$
4. Randomly select a new centroid by choosing a point with probability proportional to $\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$ is
5. End For

Bisecting K-means

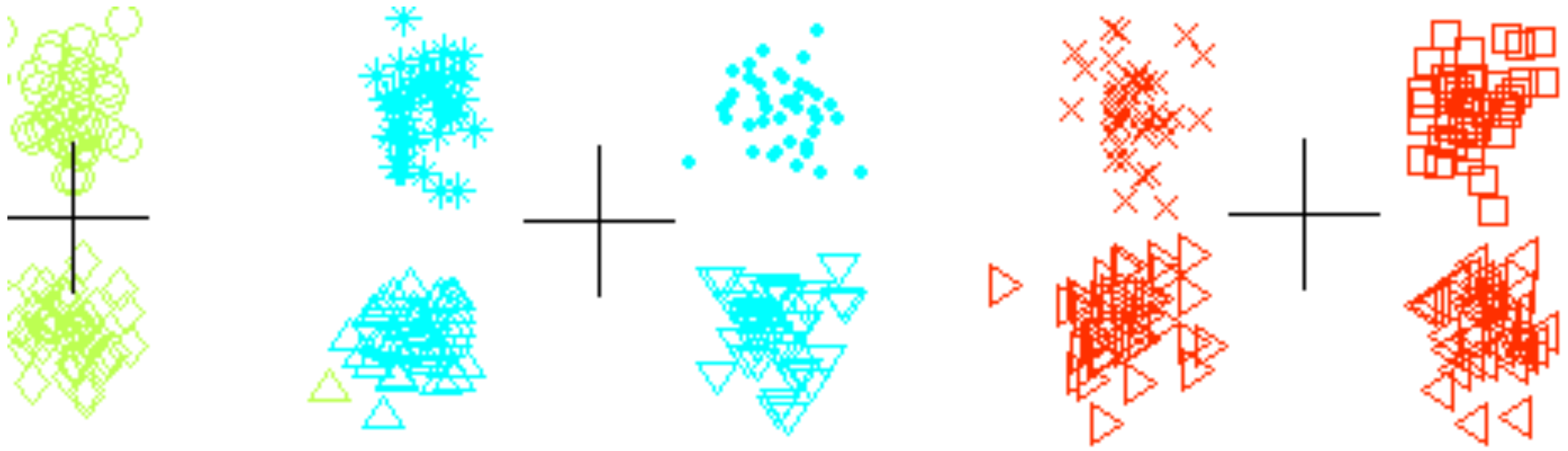
Variant of K-means that can produce a partitional or a hierarchical clustering

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

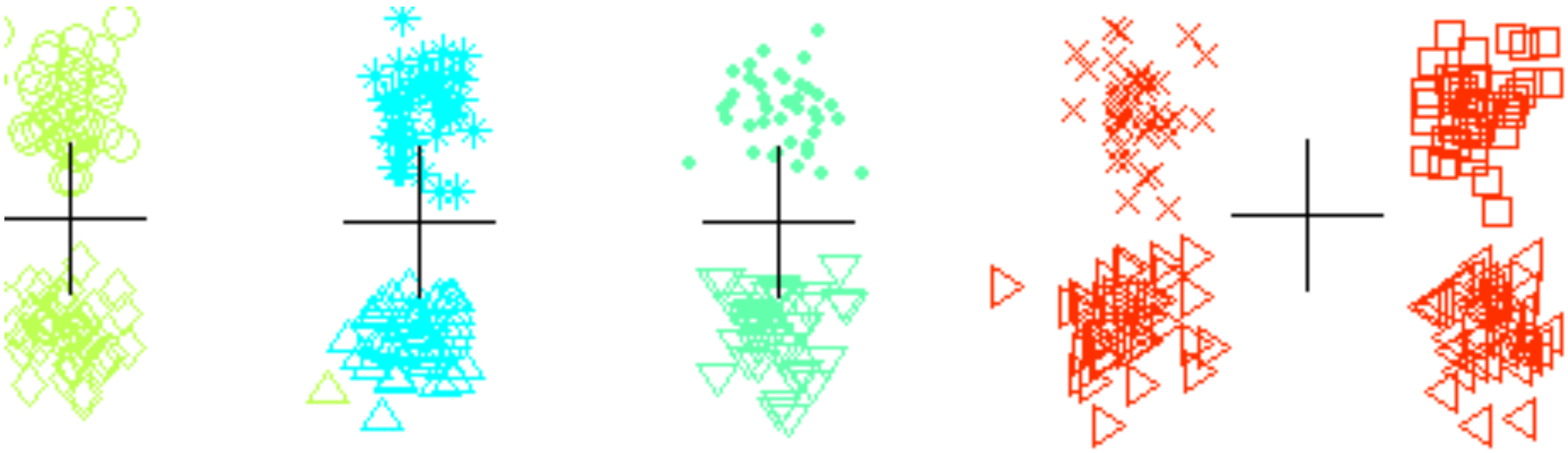
CLUTO: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

Bisecting K-means Example

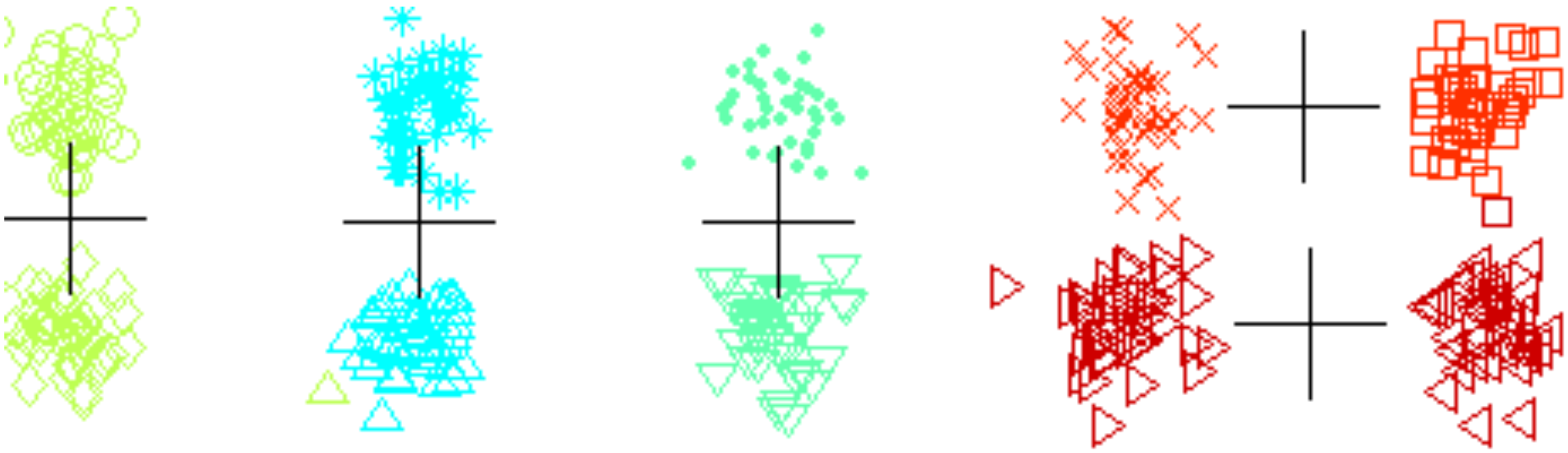
Bisecting K-means Example



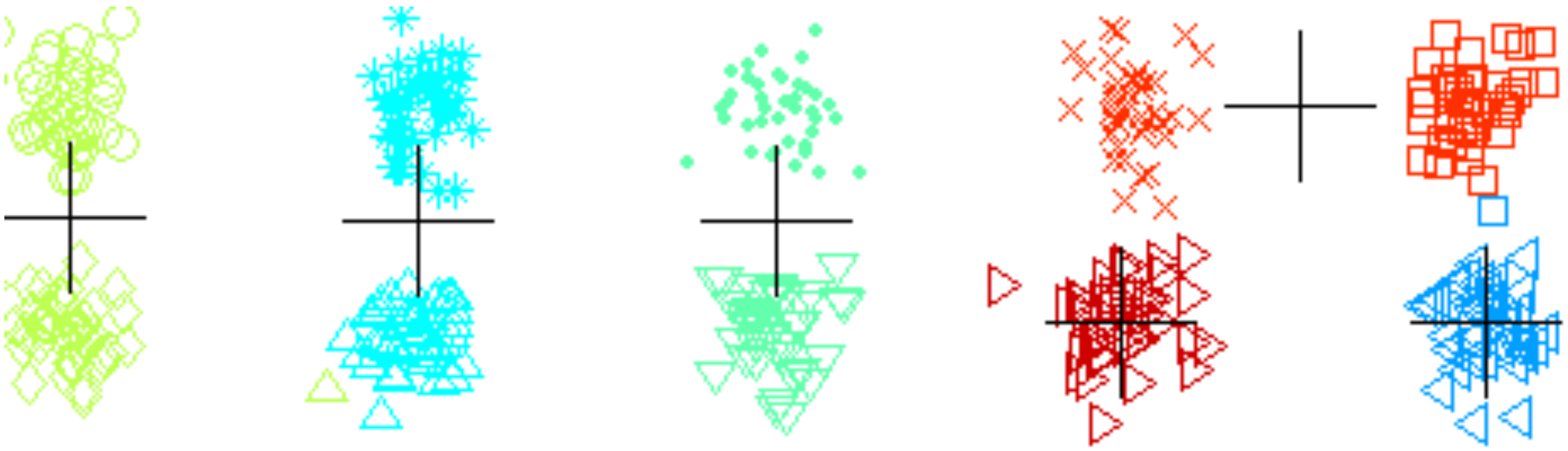
Bisecting K-means Example



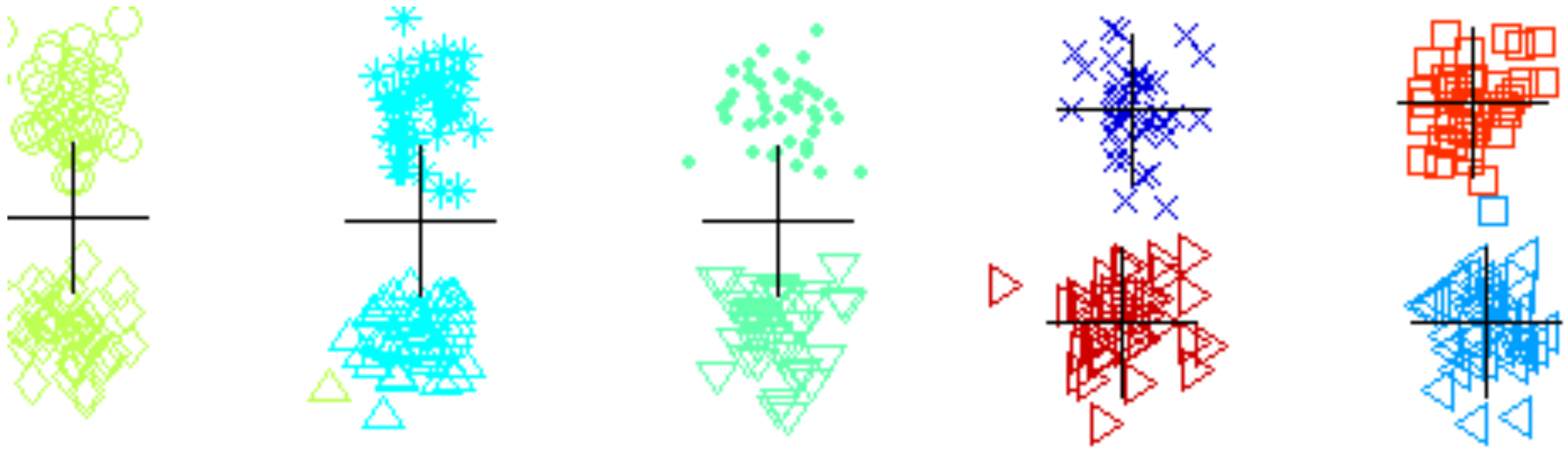
Bisecting K-means Example



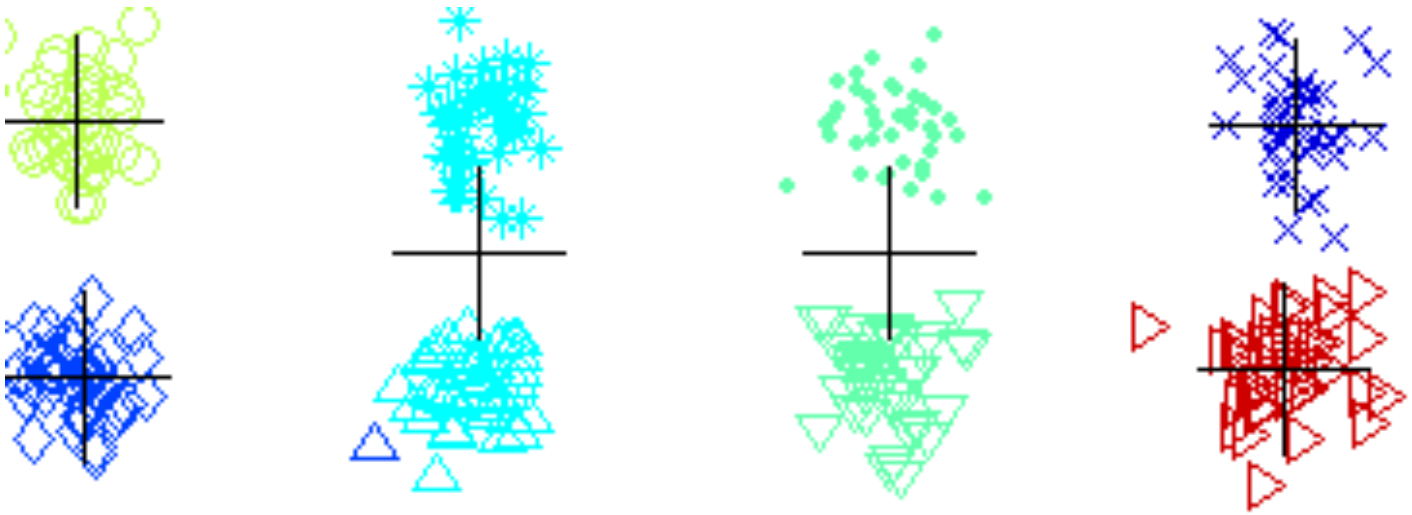
Bisecting K-means Example



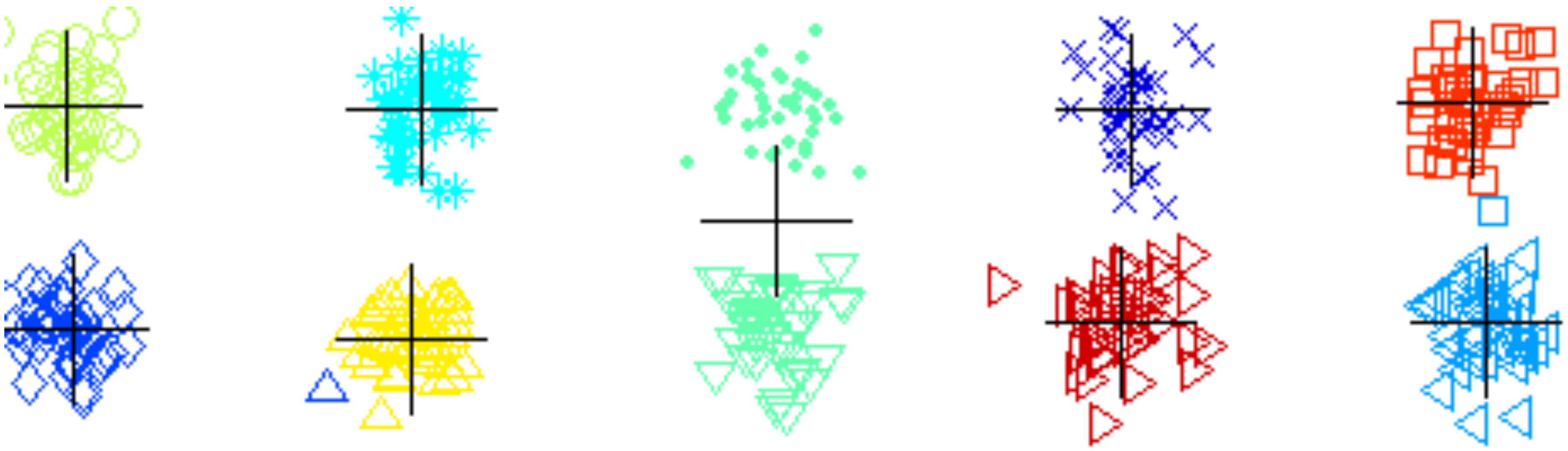
Bisecting K-means Example



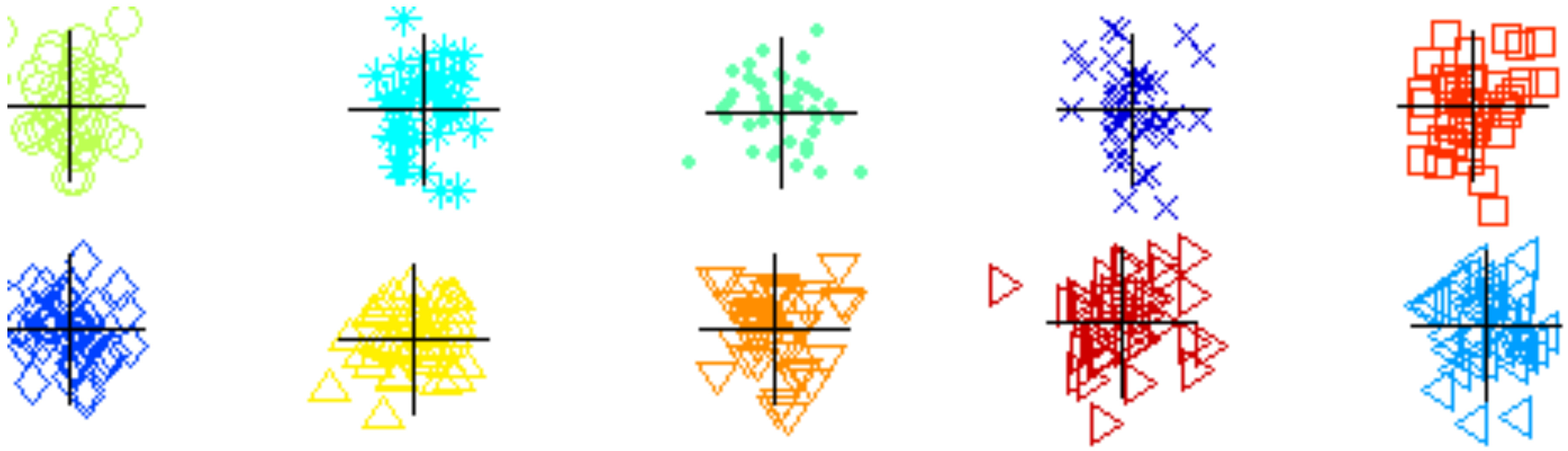
Bisecting K-means Example



Bisecting K-means Example



Bisecting K-means Example



Limitations of K-means

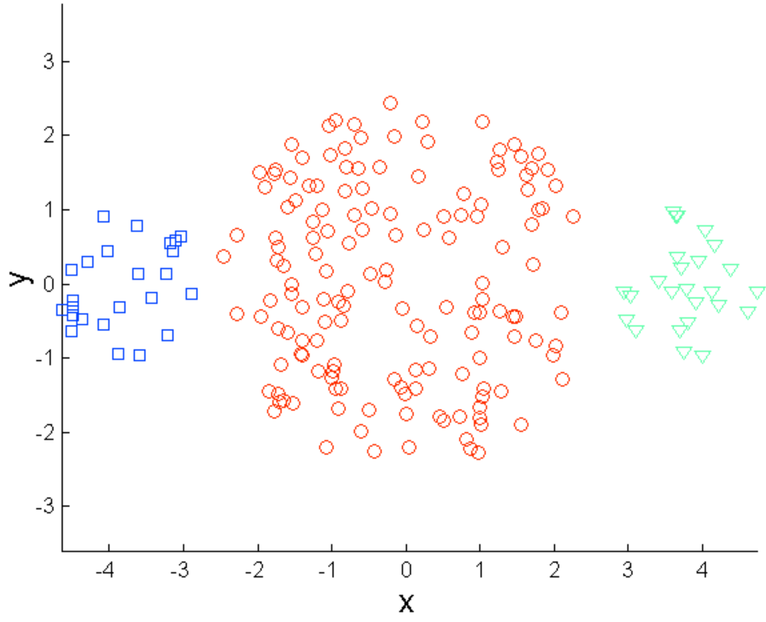
K-means has problems when clusters are of differing

- Sizes
- Densities
- Non-globular shapes

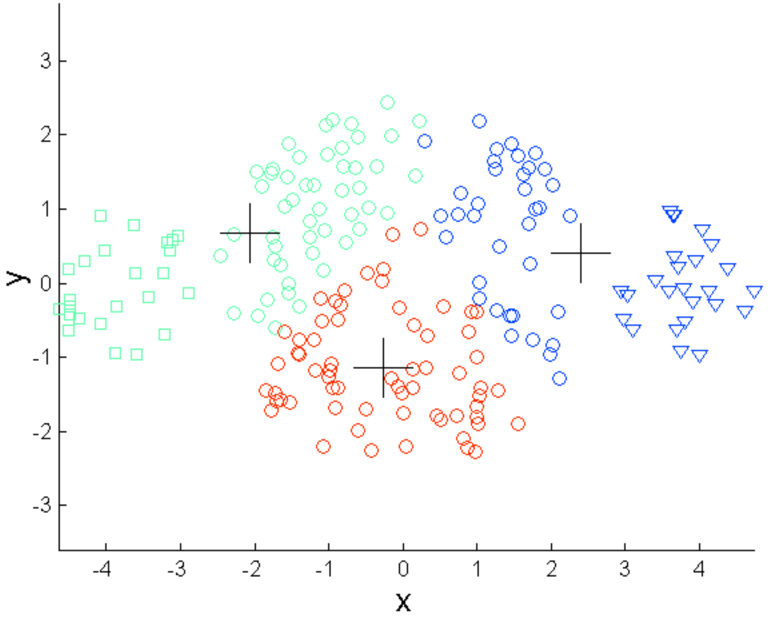
K-means has problems when the data contains outliers.

- One possible solution is to remove outliers before clustering

Limitations of K-means: Differing Sizes

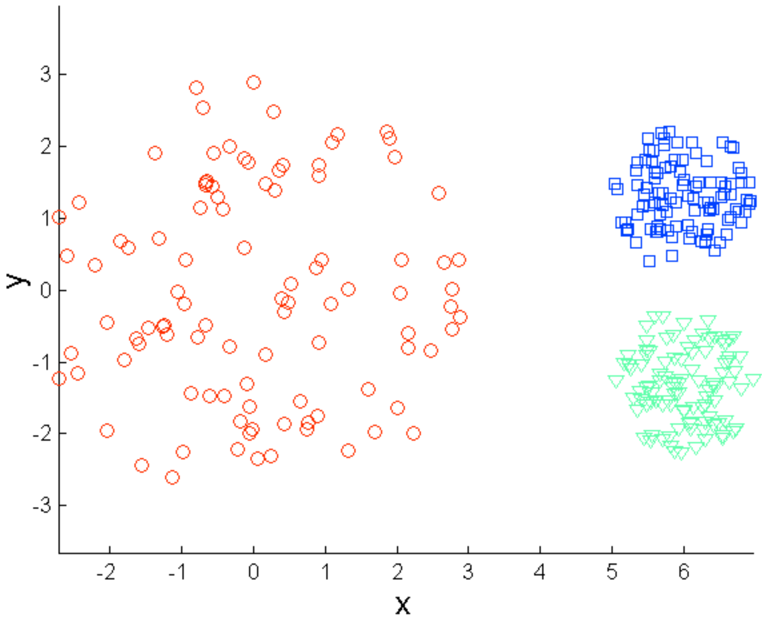


Original Points

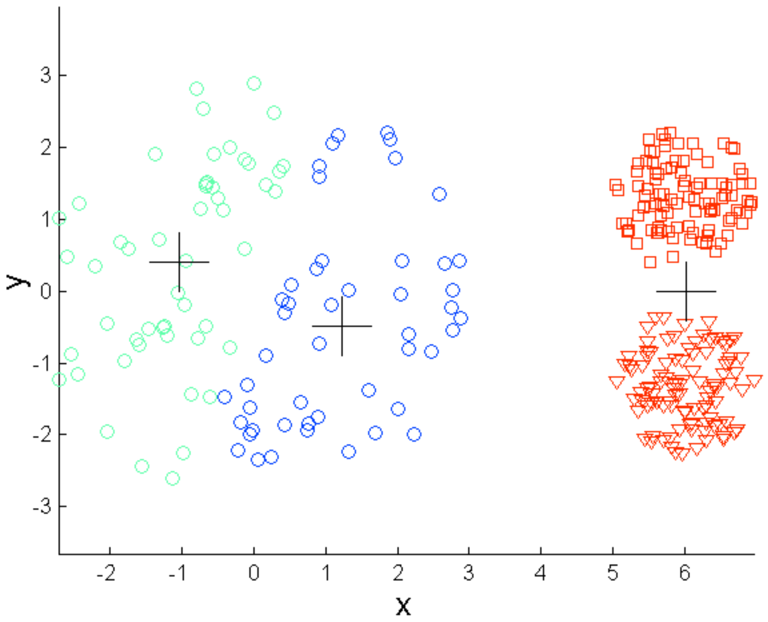


K-means (3 Clusters)

Limitations of K-means: Differing Density

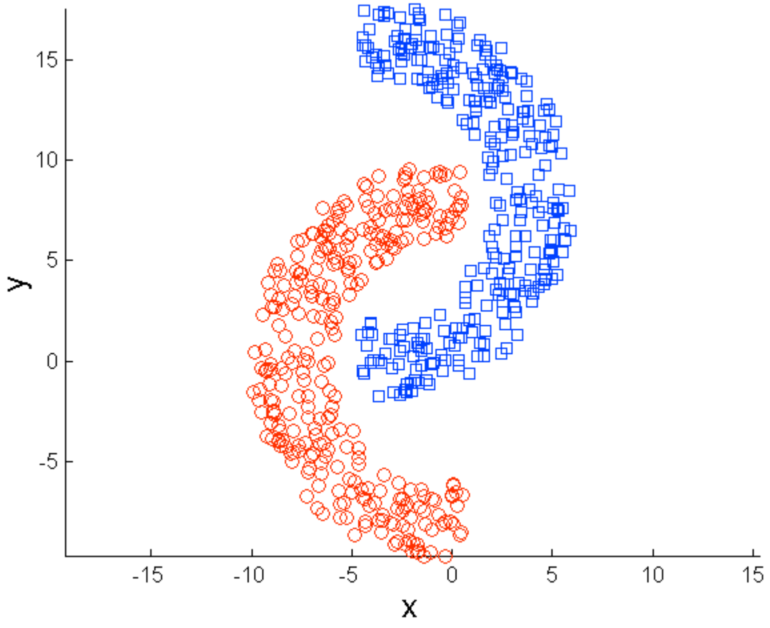


Original Points

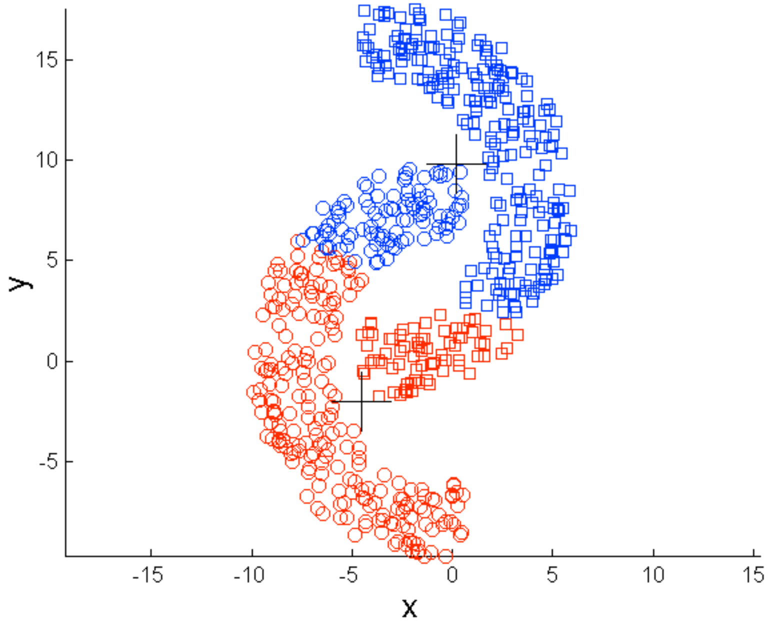


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points

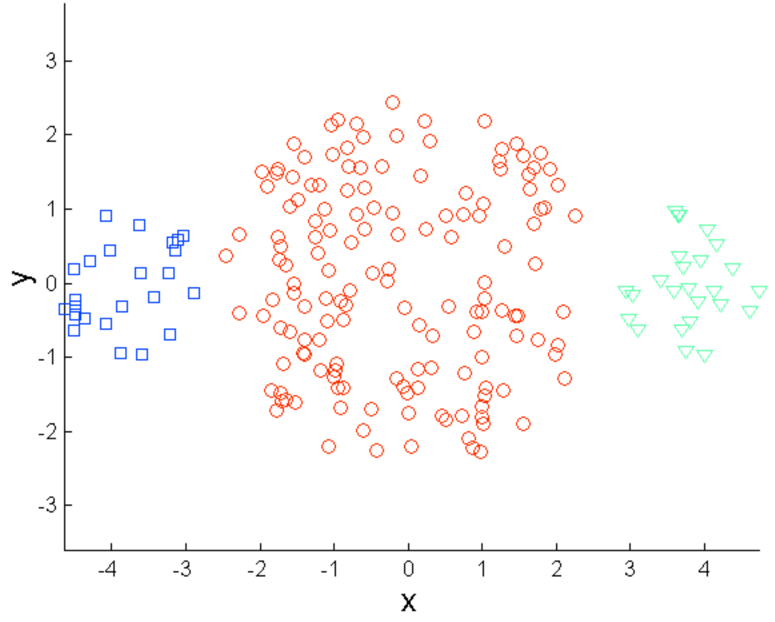


K-means (2 Clusters)

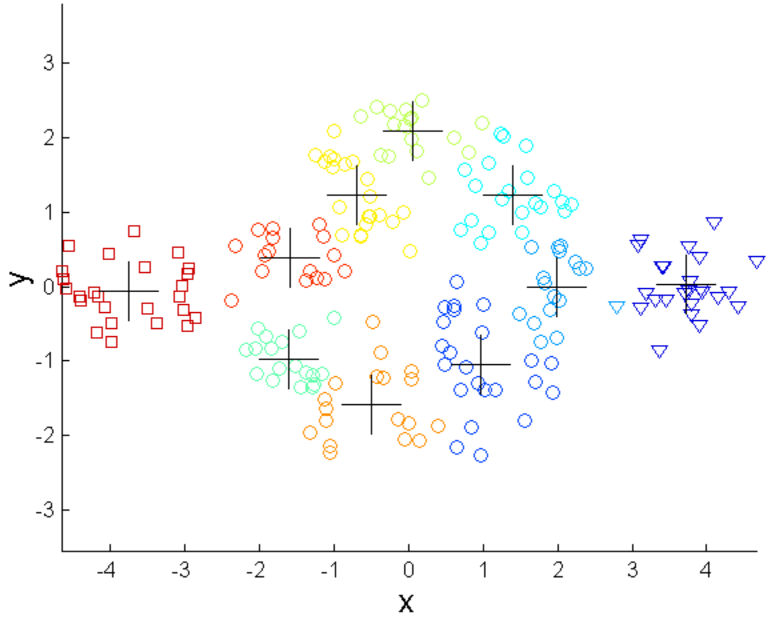
Overcoming K-means Limitations

One solution is to find a **large number of clusters** such that each of them represents a part of a natural cluster.

Small clusters need to be put together in a **post-processing** step.



Original Points

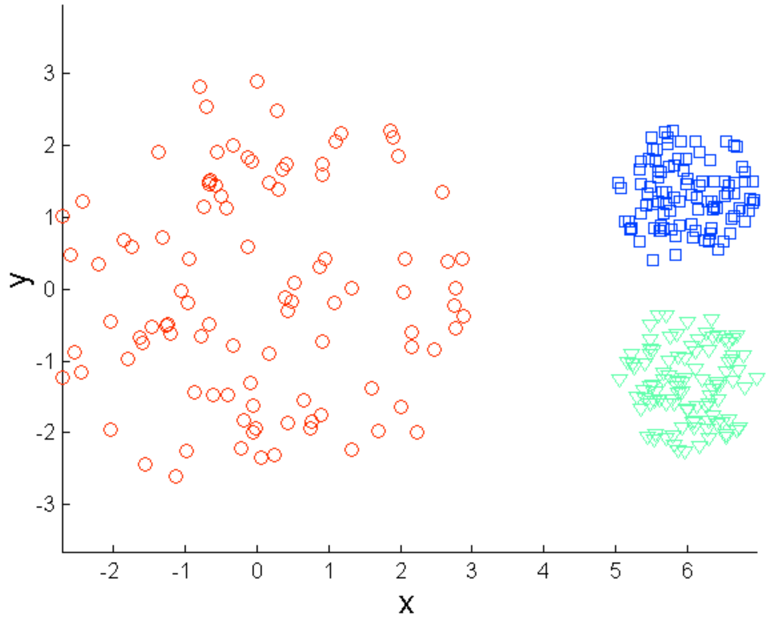


K-means Clusters

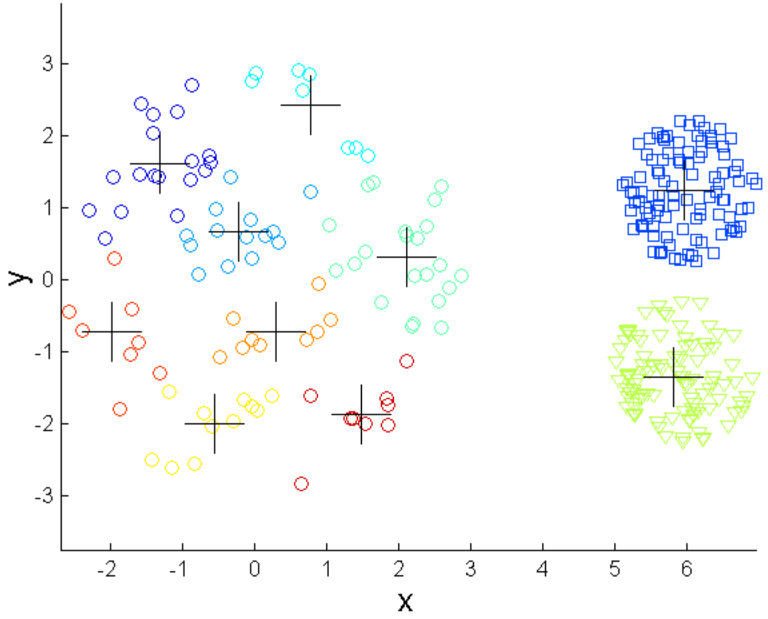
Overcoming K-means Limitations

One solution is to find a **large number of clusters** such that each of them represents a part of a natural cluster.

Small clusters need to be put together in a **post-processing** step.



Original Points

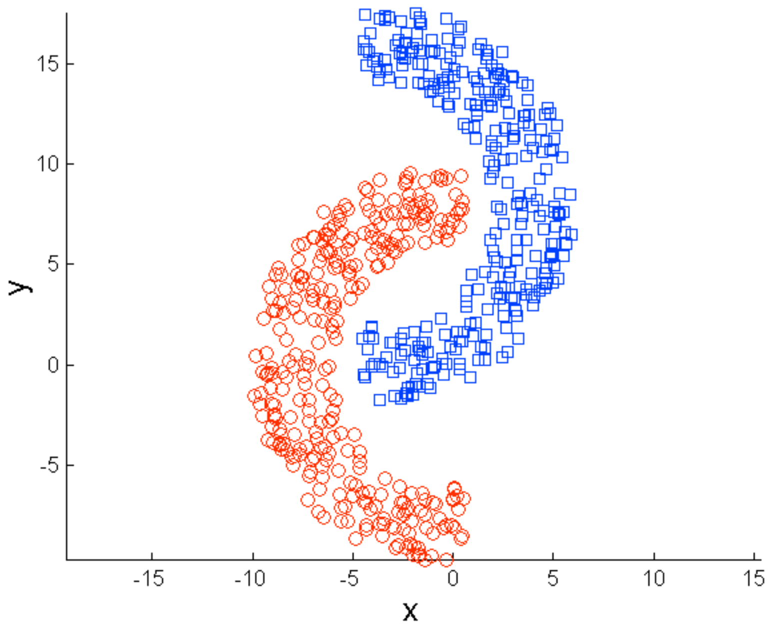


K-means Clusters

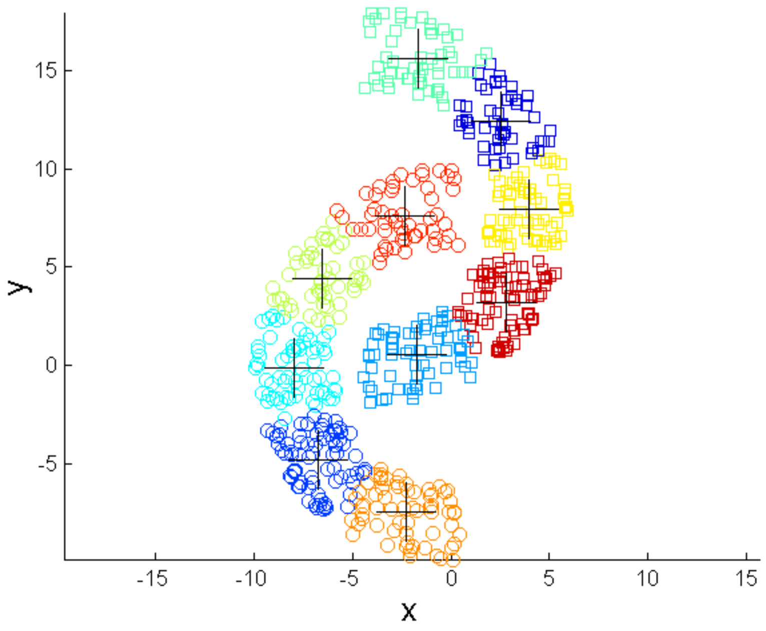
Overcoming K-means Limitations

One solution is to find a **large number of clusters** such that each of them represents a part of a natural cluster.

Small clusters need to be put together in a **post-processing** step.



Original Points



K-means Clusters