k-MEANS Clustering
and
k-Medoids Clustering

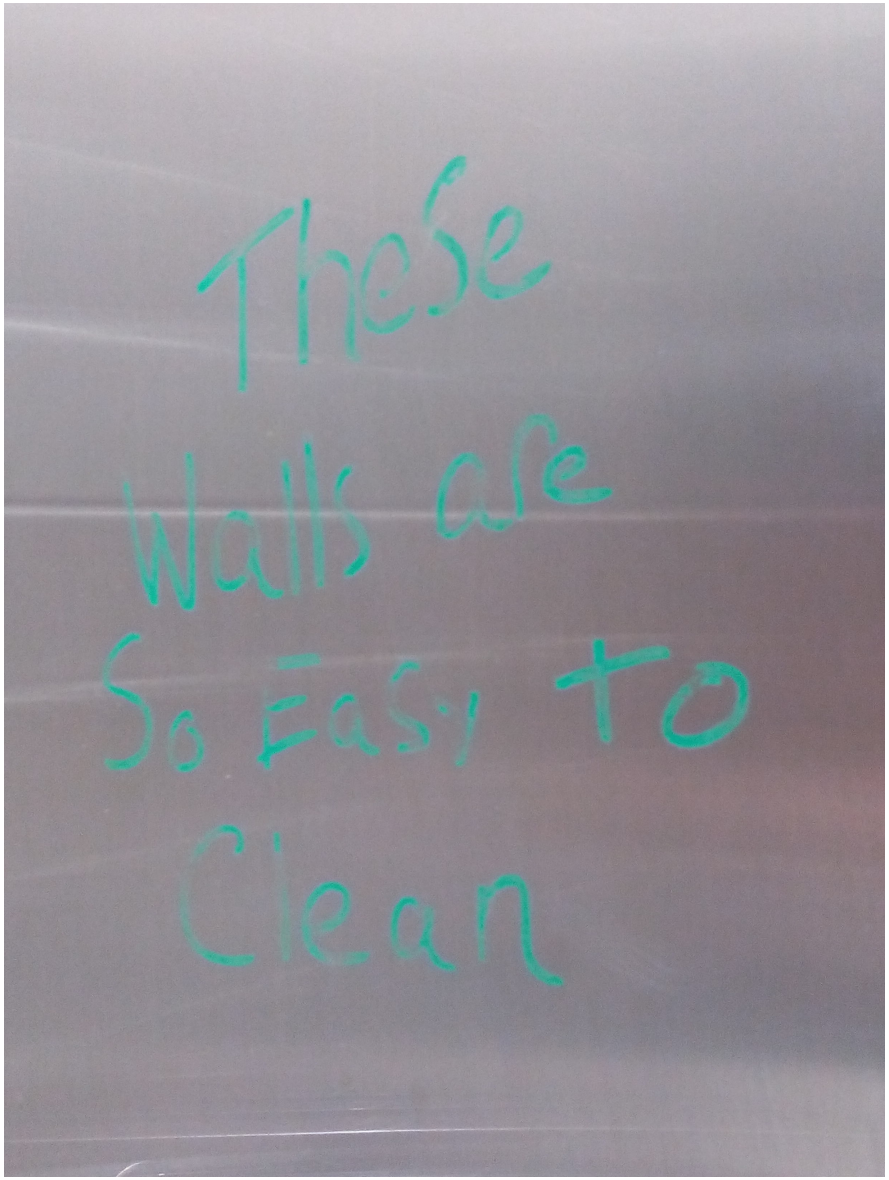
Thomas Kinsman

What is the period of rotation at infinity?

It used to move!

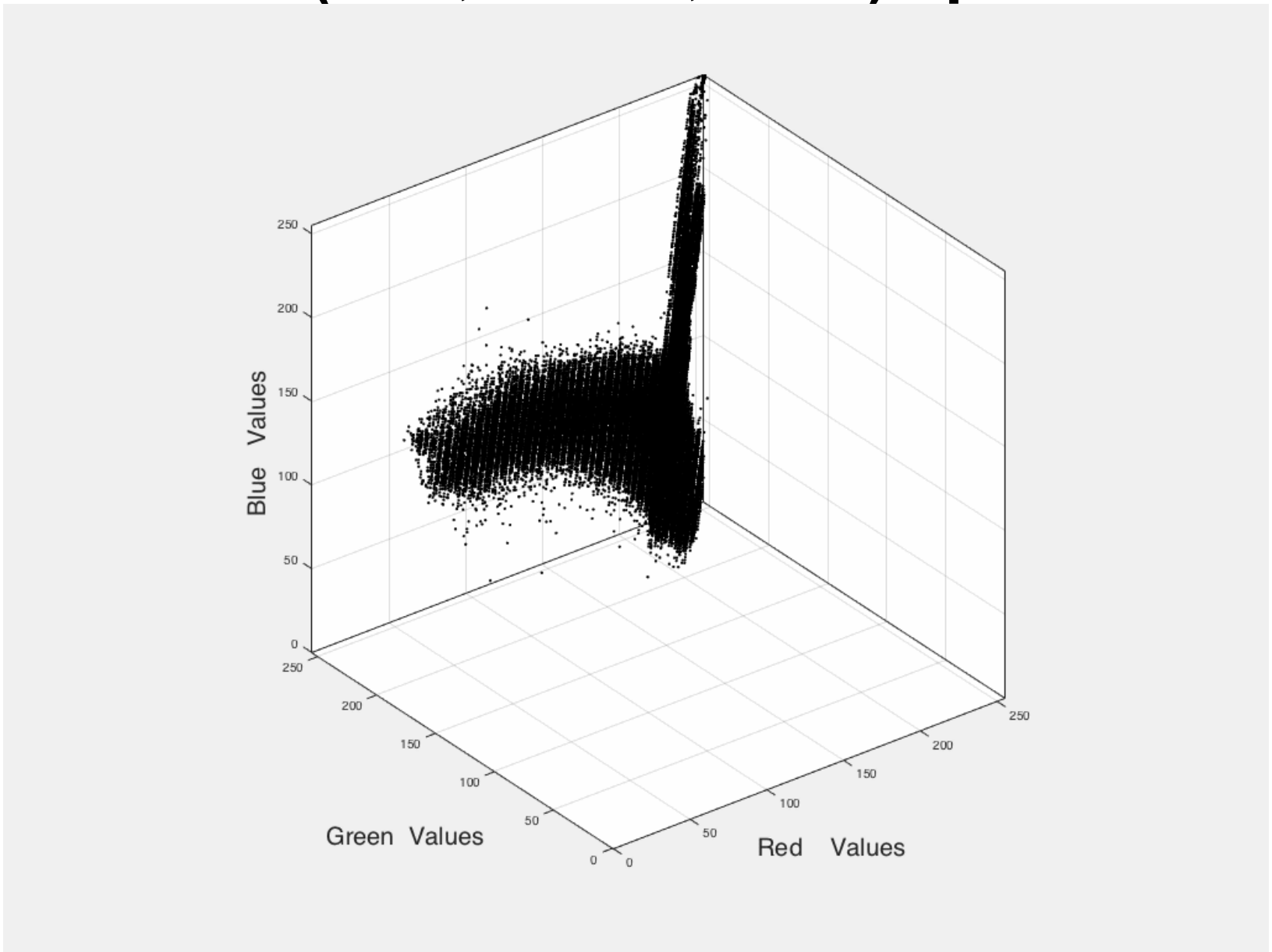
What is the rate of absorption?

What pixels are Graffiti?



These
Walls are
So Easy to
Clean

Data in (Red, Green, Blue) space



k-Means, K-means, K-Means, kMeans, c-Means, C-Means, ...

- Many different spellings for the same formula
- To divide the data into k clusters
- Also called:
  - Used called/used for “Vector Quantization”
  - Closely related to Lloyd’s algorithm
  - Used in Compression
- So, when you search for it, don’t be surprised if it has a different name.

Don’t write this down yet

Generic k-Means:

A. Select K random initial seed points as prototypes for the clusters:
B. Repeat:
    i. Assign all data points to the closest center
    ii. For each cluster formed, find the new prototype for the center
C. Stop when the prototypes do not move (Or move by less then some tolerance.)

Here's a joke:

- How many clusters does k-mean's return?

Measures of Central Tendency

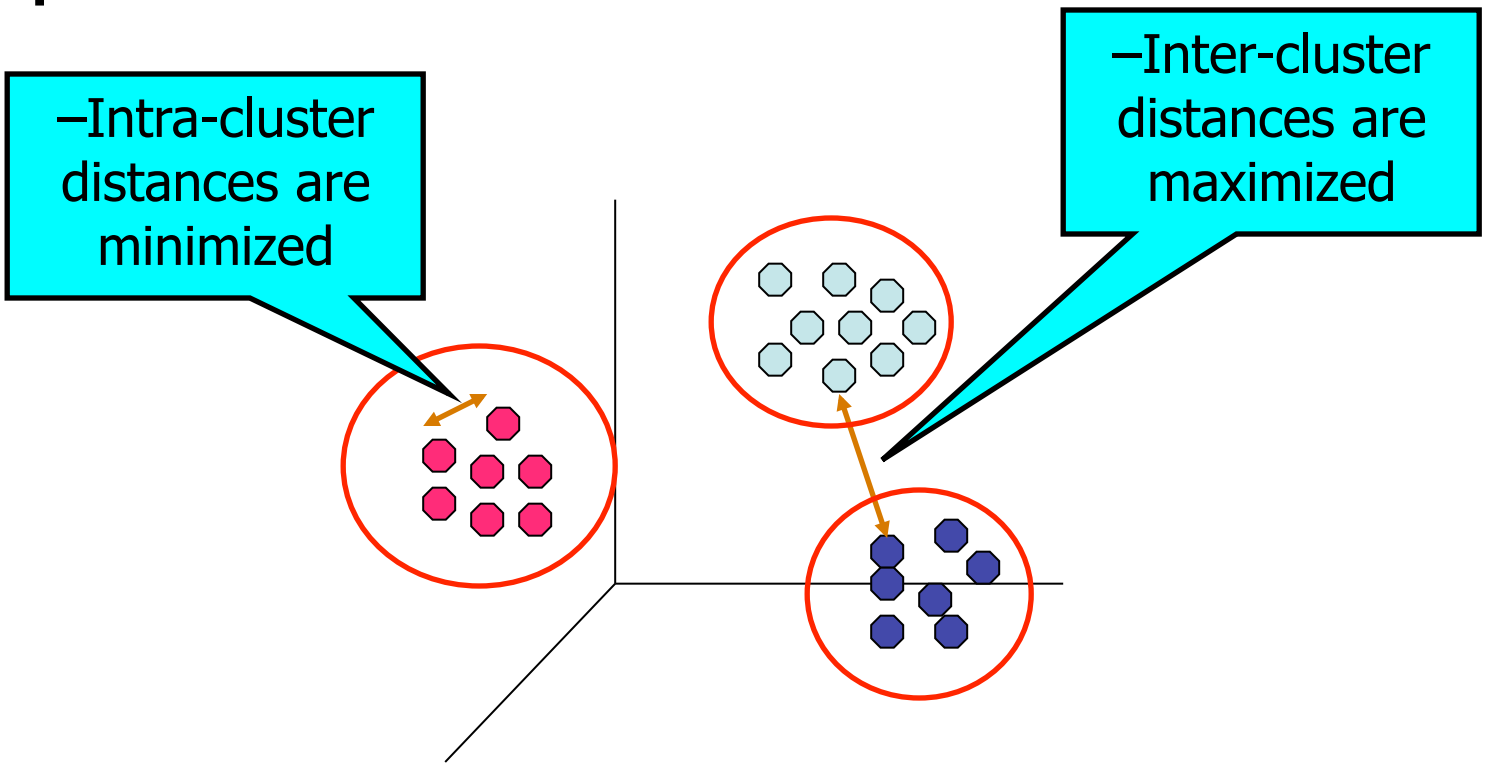
1. **Mean** – the "average"
2. **Mode** – the most common value.
3. **Median** – the central value.
4. **Centroid** -- the N-Dimensional center of mass.

Measures of Central Tendency

1. **Mean** – the "average"
2. **Mode** – the most common value.
3. **Median** – the central value.
4. **Centroid** – the N-Dimensional center of mass.
   – the COM.
5. **Medoid** – the data point closest to the COM.

Review – Ideal Clustering

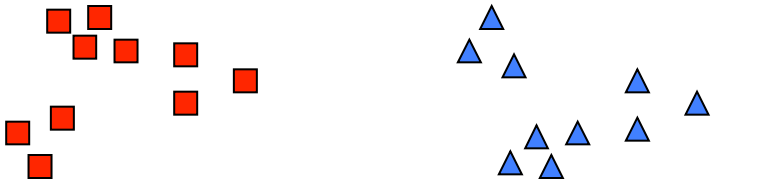
- **Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups**



The Ambiguity of Clusters

–How many clusters?

–Six Clusters



–Two Clusters

–Four Clusters

Clustering Vocabulary

- **A clustering** is a set of clusters
  It's a terrible term because "clustering" implies a process or action. Better term would be "partitioning"
- Hierarchical vs Partitional clusters:
  - Hierarchical clustering:

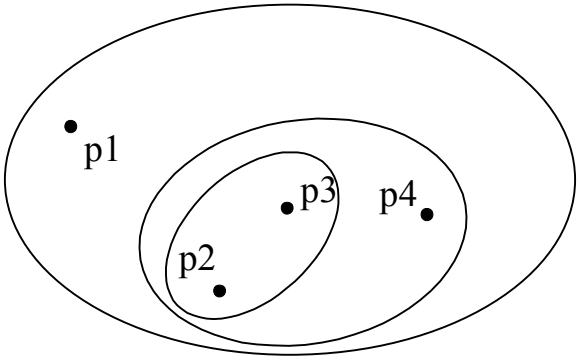
    A set of nested clusters organized as a hierarchical tree.

    Agglomerative clustering is an example.

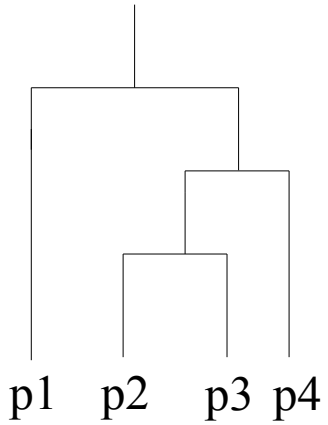
  - Partitional Clustering:

    Division of data points into non-overlapping clusters such that each data point is in exactly one cluster.

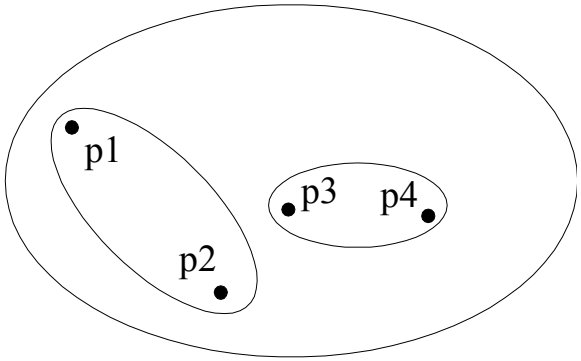
Hierarchical Clustering



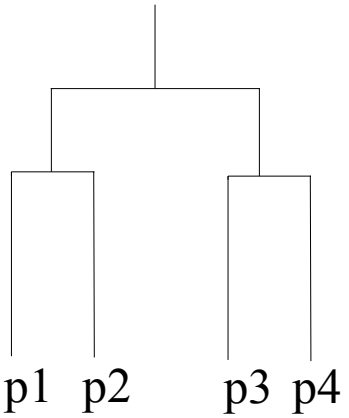
Traditional Hierarchical Clustering



Traditional Dendrogram

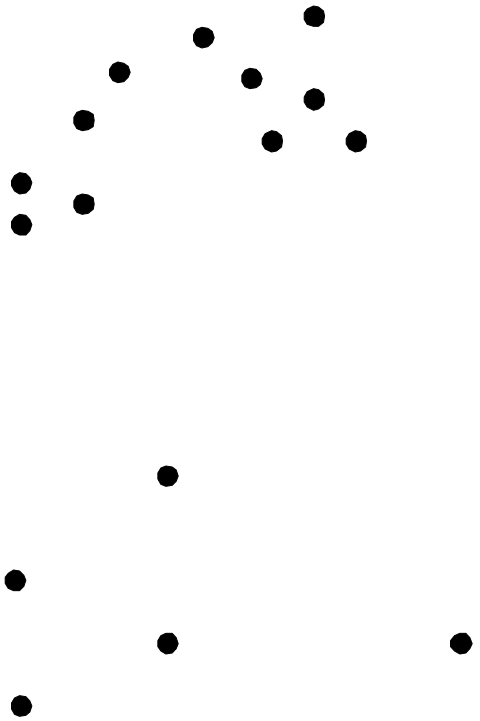


Non-traditional Hierarchical Clustering

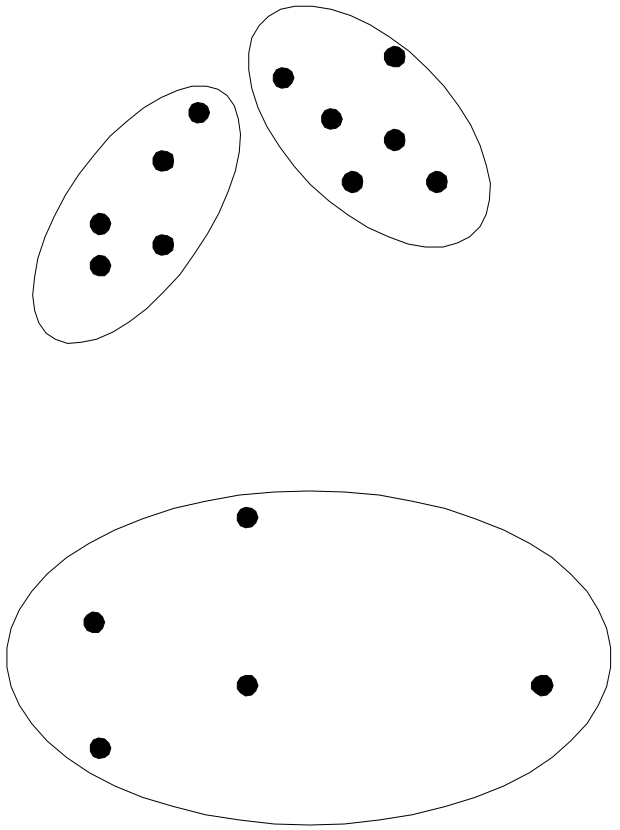


Non-traditional Dendrogram

Partitional Clustering



–Original Points



–A Partitional Clustering

Characteristics of the Data Are Important

- Type of proximity or density measure
  - This is a derived measure, but the epitome of importance
  - The relative distances between points is important
- Sparseness
  - Dictates type of similarity
  - Adds to efficiency
- Attribute type
  - Dictates type of similarity to use
- Noise and Outliers
  - Outliers always throw off the center of mass

K-means Clustering

- Partitional clustering approach
  - partitions the data points into clusters
- Each cluster is associated with a prototype (center and extent)
  - Centroid (center of mass)
  - Medoid (data point closest to center of mass)
- Data point is assigned to the cluster whose prototype is the closest to that data point
- Number of clusters, K, must be specified
- The generic algorithm is conceptually easy

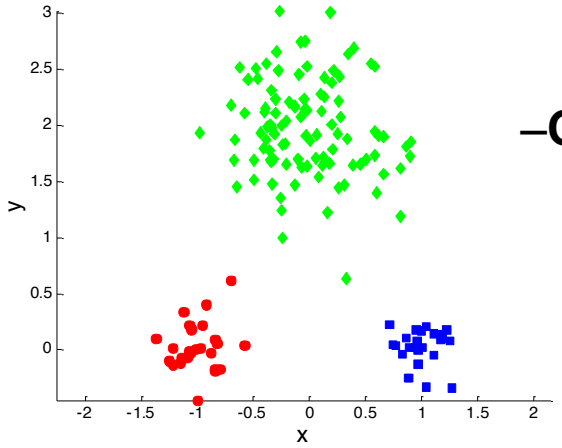
Generic k-Means

A. Select k initial seed points as prototypes for the clusters:
B. Repeat:
    i. Assign all data points to the closest center
    ii. For each cluster formed, find the new prototype for the center
C. Stop when the prototypes do not move
(Or move by less then some tolerance.)

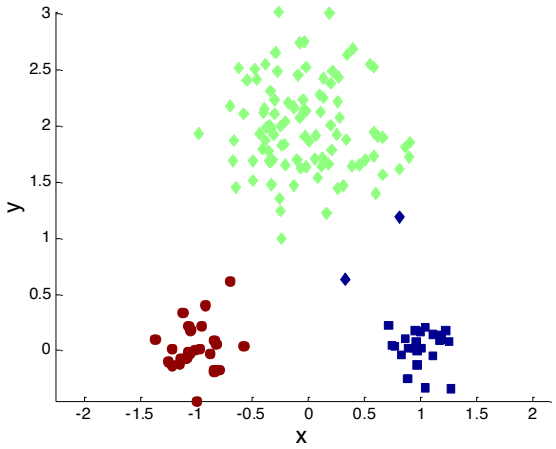
K-means Clustering – Details

- Initial centroids are chosen randomly from the data.
  - Clusters produced vary from one run to another.
- Closeness or proximity uses an appropriate distance metric
- 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’

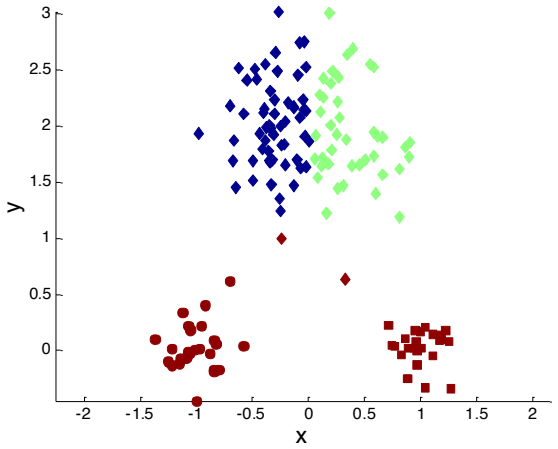
Two different K-means Clusterings – Two Runs



–Original Points



–Optimal Clustering



–Sub-optimal Clustering

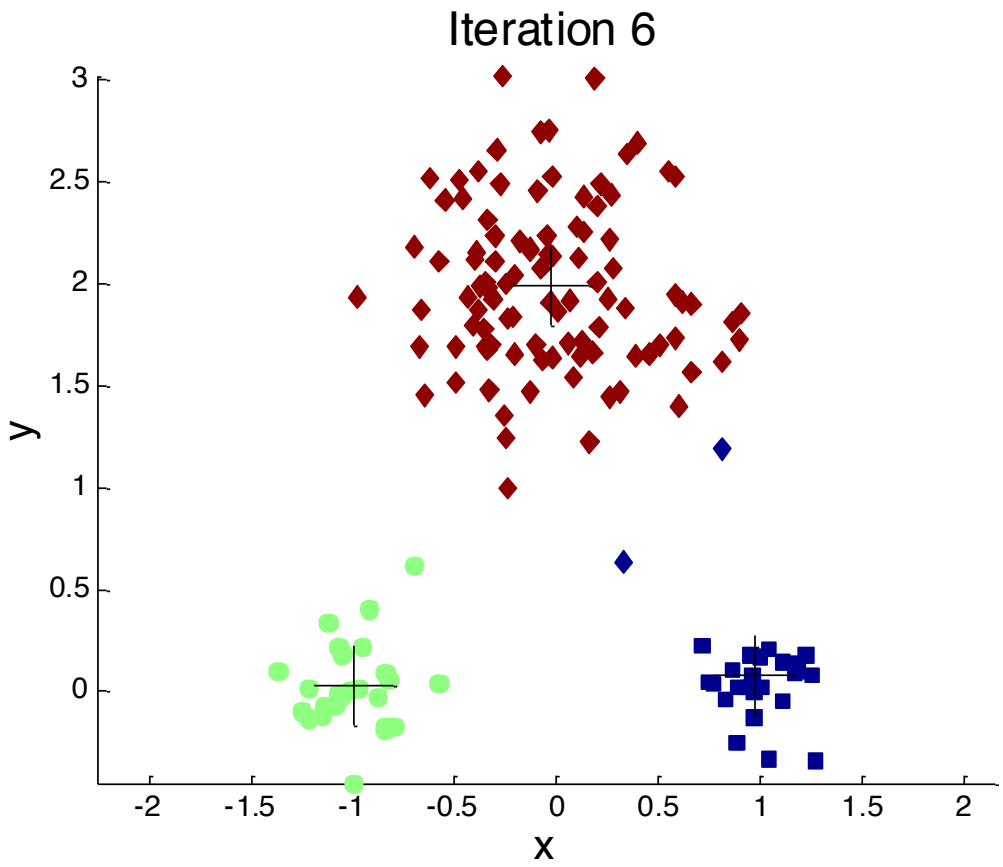
SSE - Evaluating K-means Clusters

- Most common measure is Sum of 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.

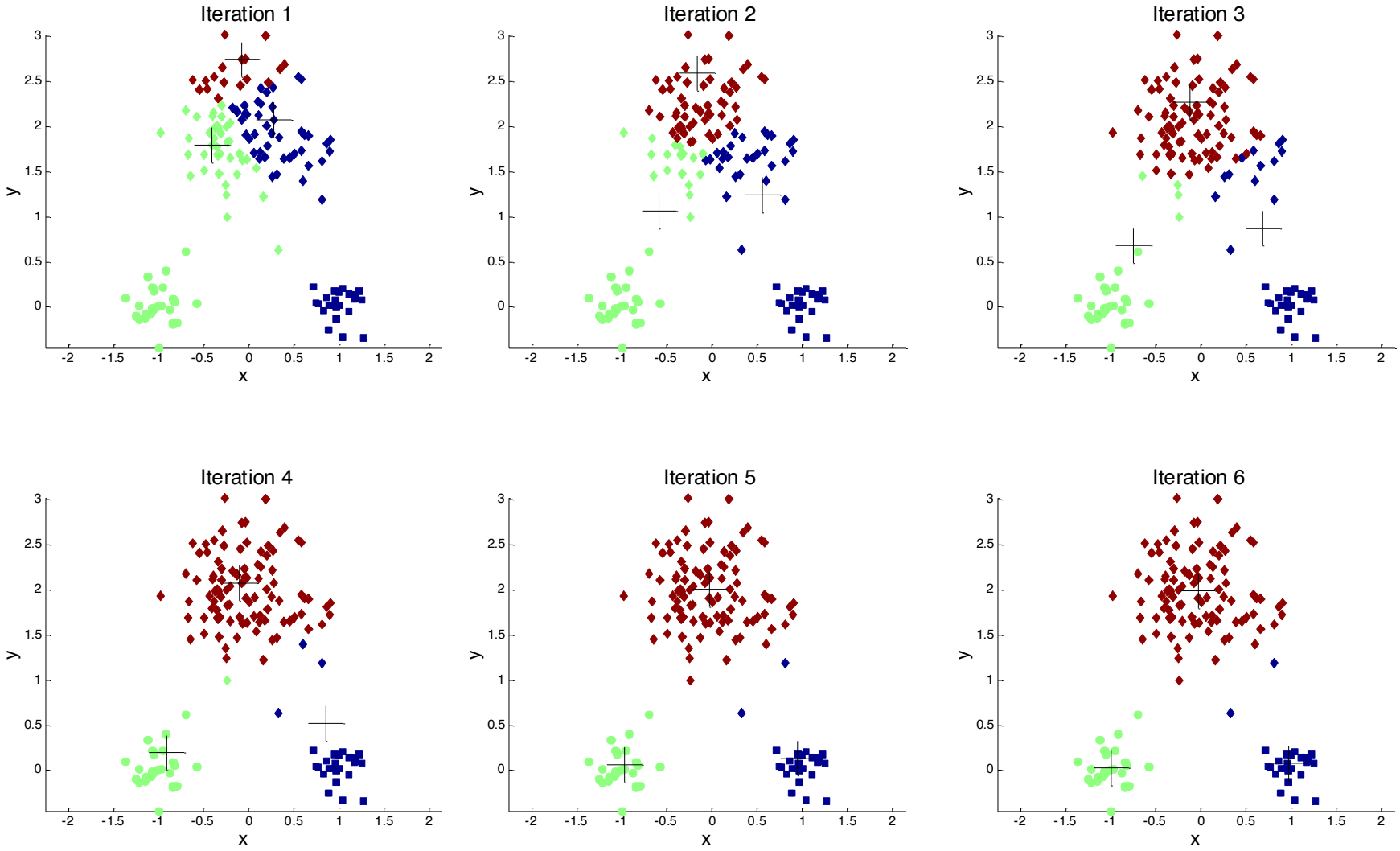
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two ***clusterings***, we can choose the one with the smallest error
  - An easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with a given K will have a lower SSE
    - A larger K can have a lower SSE
    - A good clustering with a low K can have a lower SSE than a poor clustering with higher K

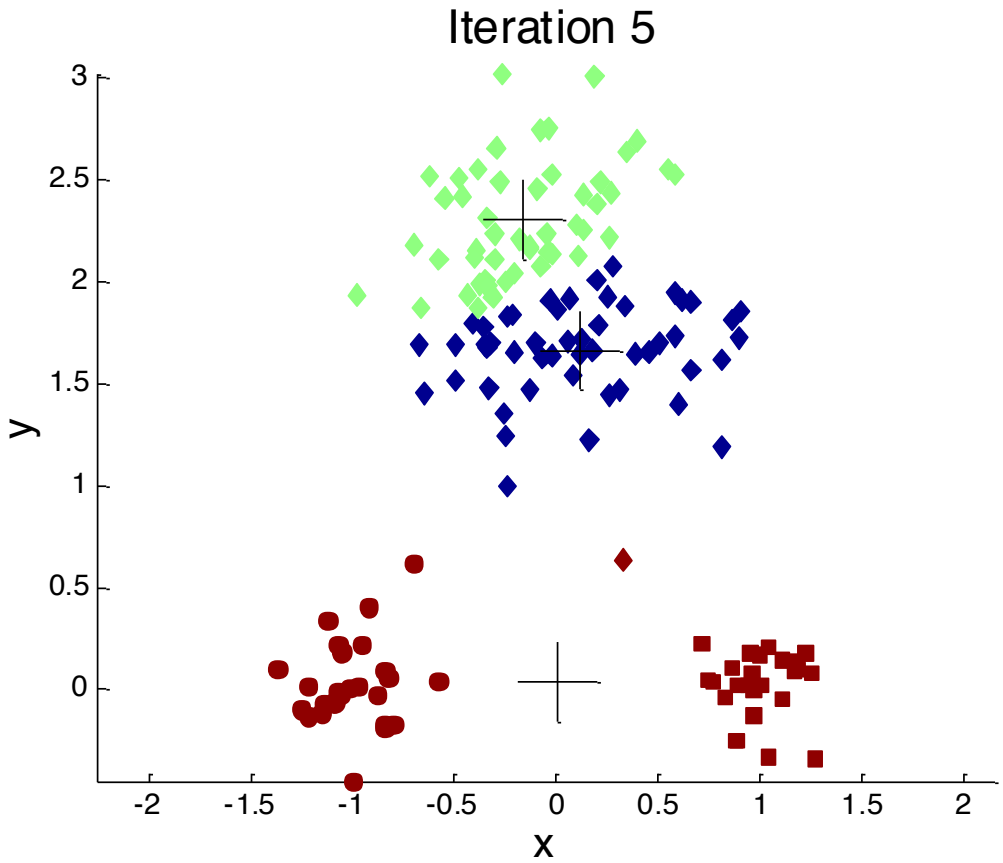
Importance of Choosing Initial Centroids



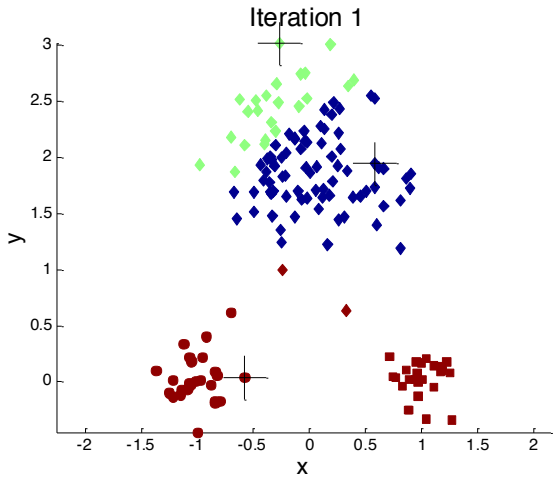
Importance of Choosing Initial Centers

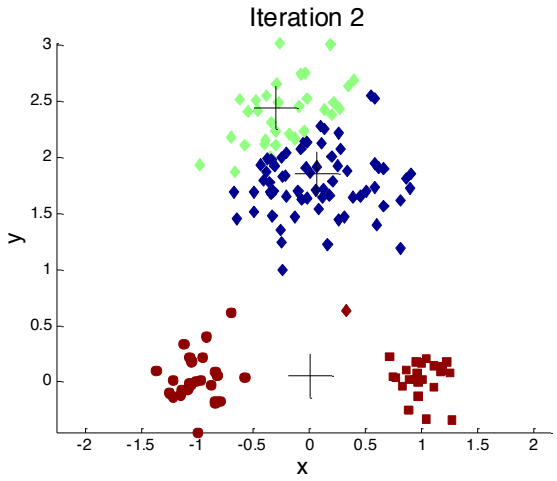


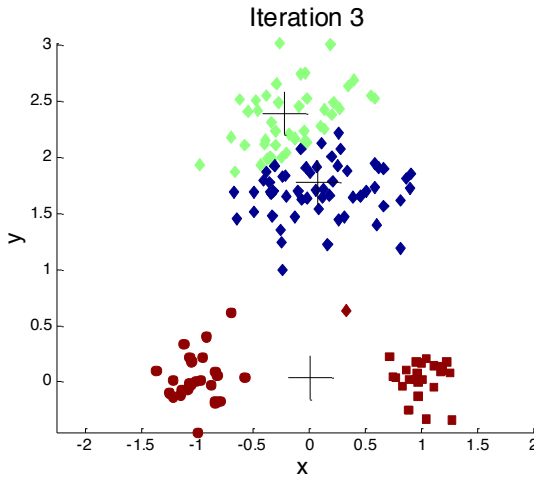
Importance of Choosing Initial Centroids …

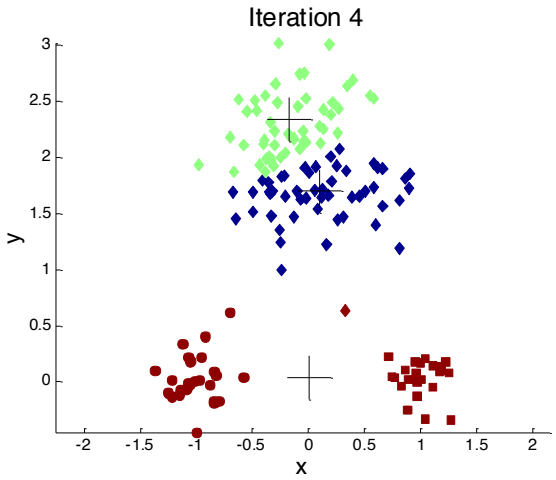


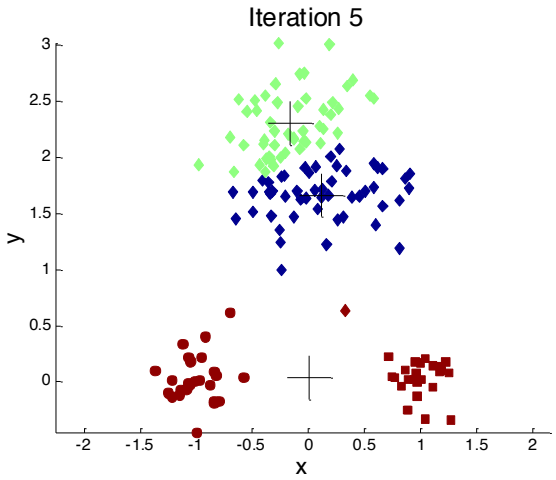
Importance of Choosing Initial Centroids …











Generic k-Means (– as before)

1. (Save room in your notes here...)

A. Select some initial seed points as prototypes for the clusters:

B. Repeat:

   i. Assign all data points to the closest center

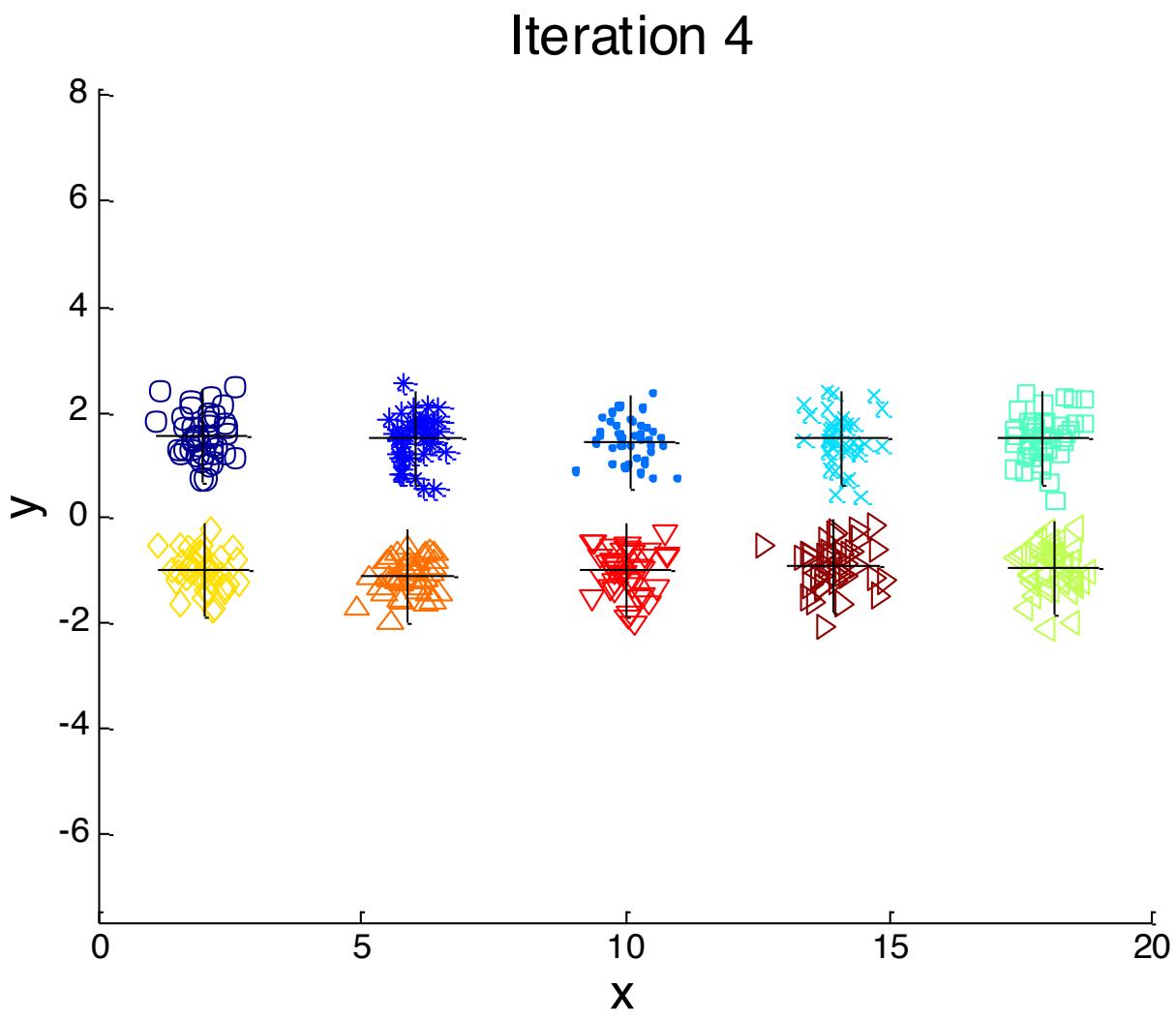
   ii. For each cluster formed, find the new prototype for the center

C. Stop when the prototypes do not move
(Or move by less then some tolerance.)

Generic k-Means

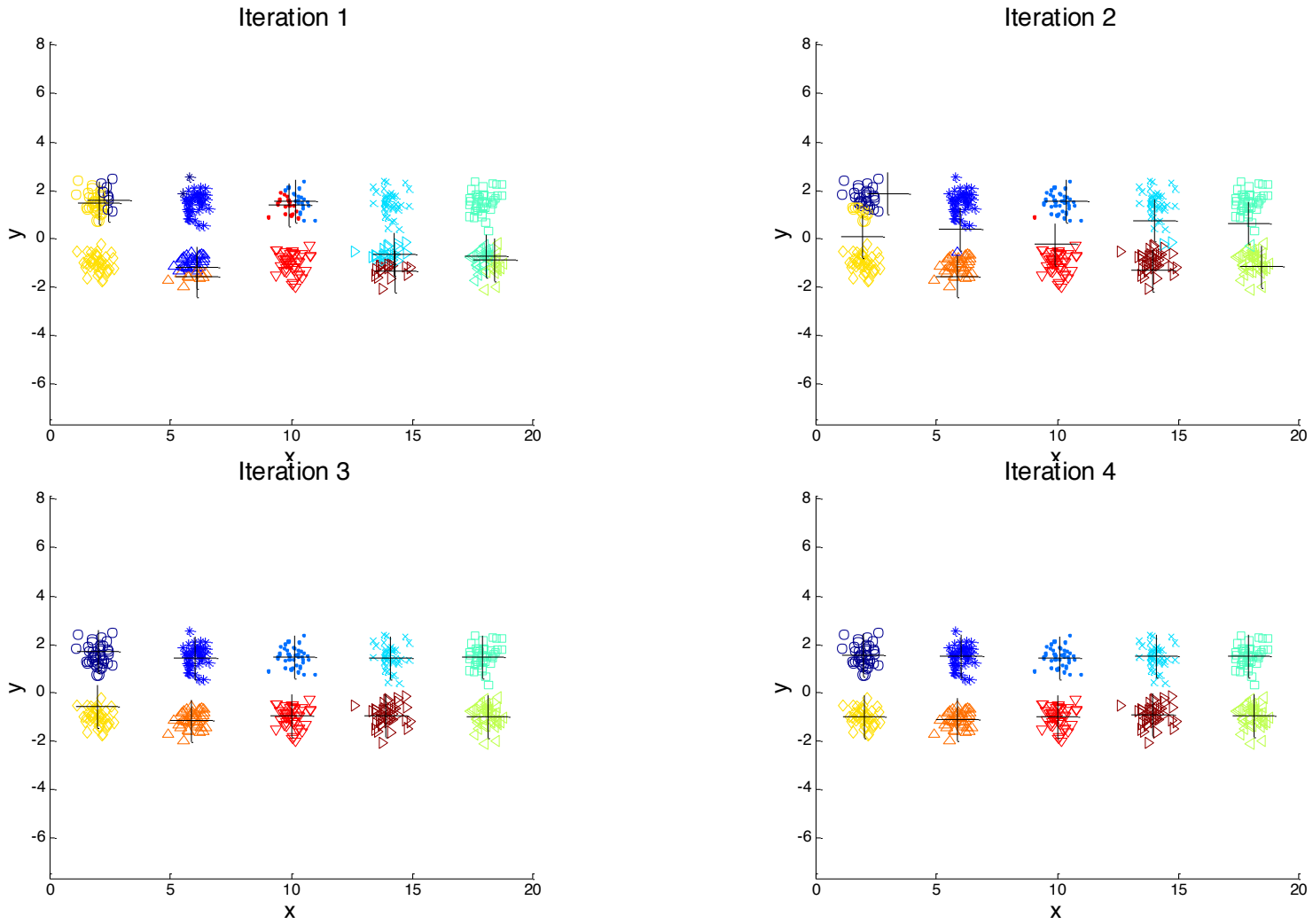
1. For some number of iterations:
   A. Select K initial seed points as prototype centers for the clusters:
   B. Repeat:
      i. Assign all data points to the closest center
      ii. For each cluster formed, find the new prototype center for the center.
      I.E. Recompute the new center of each cluster.)
   C. Stop when the prototypes do not move
   (Or move by less then some tolerance.)
   D. Evaluate the resulting clusters (using SSE) for each iteration, and save the best clustering and prototypes.
2. Use the best set of prototypes.

10 Clusters Example of Difficulty Picking Initial Cluster Centers



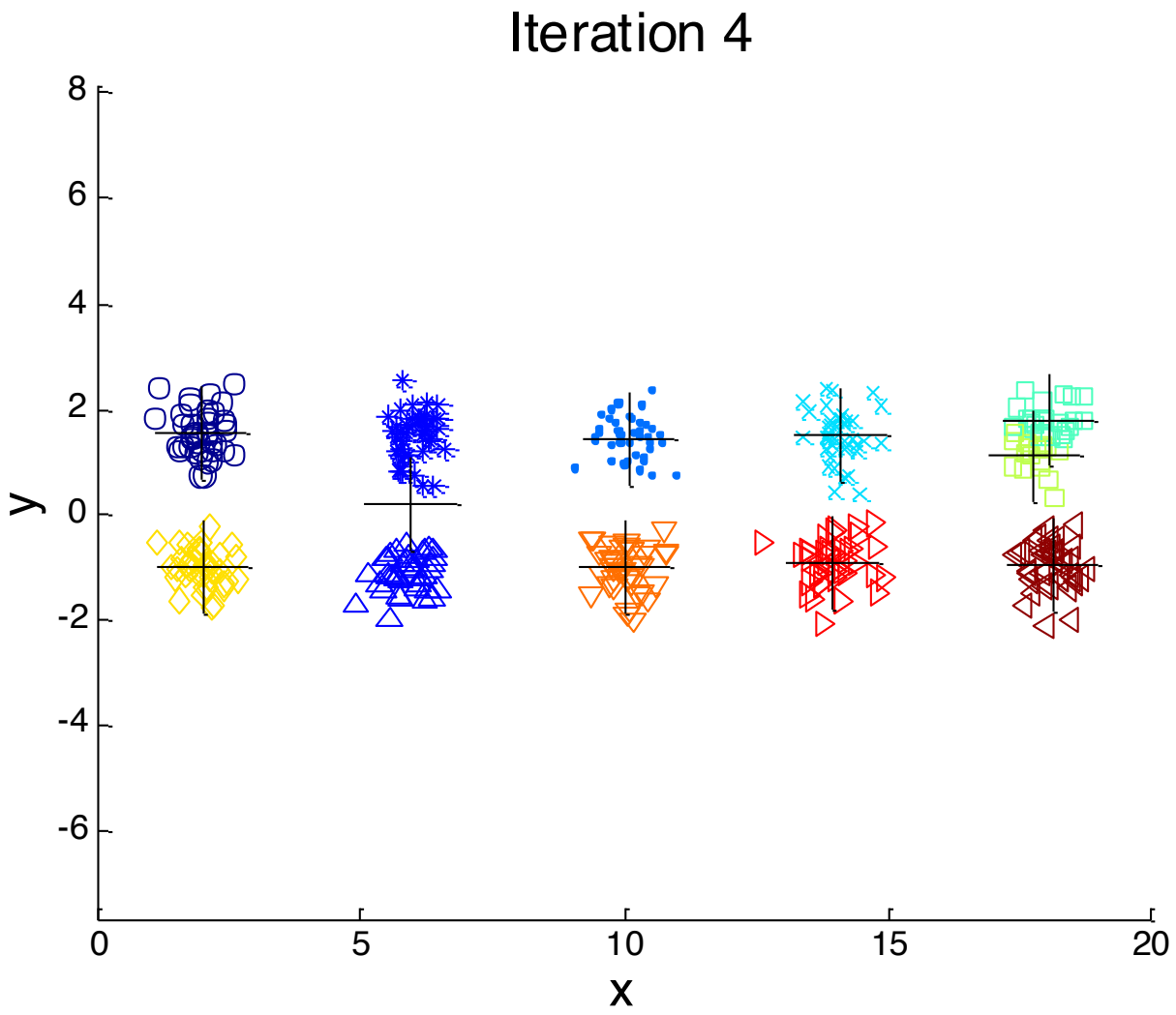
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



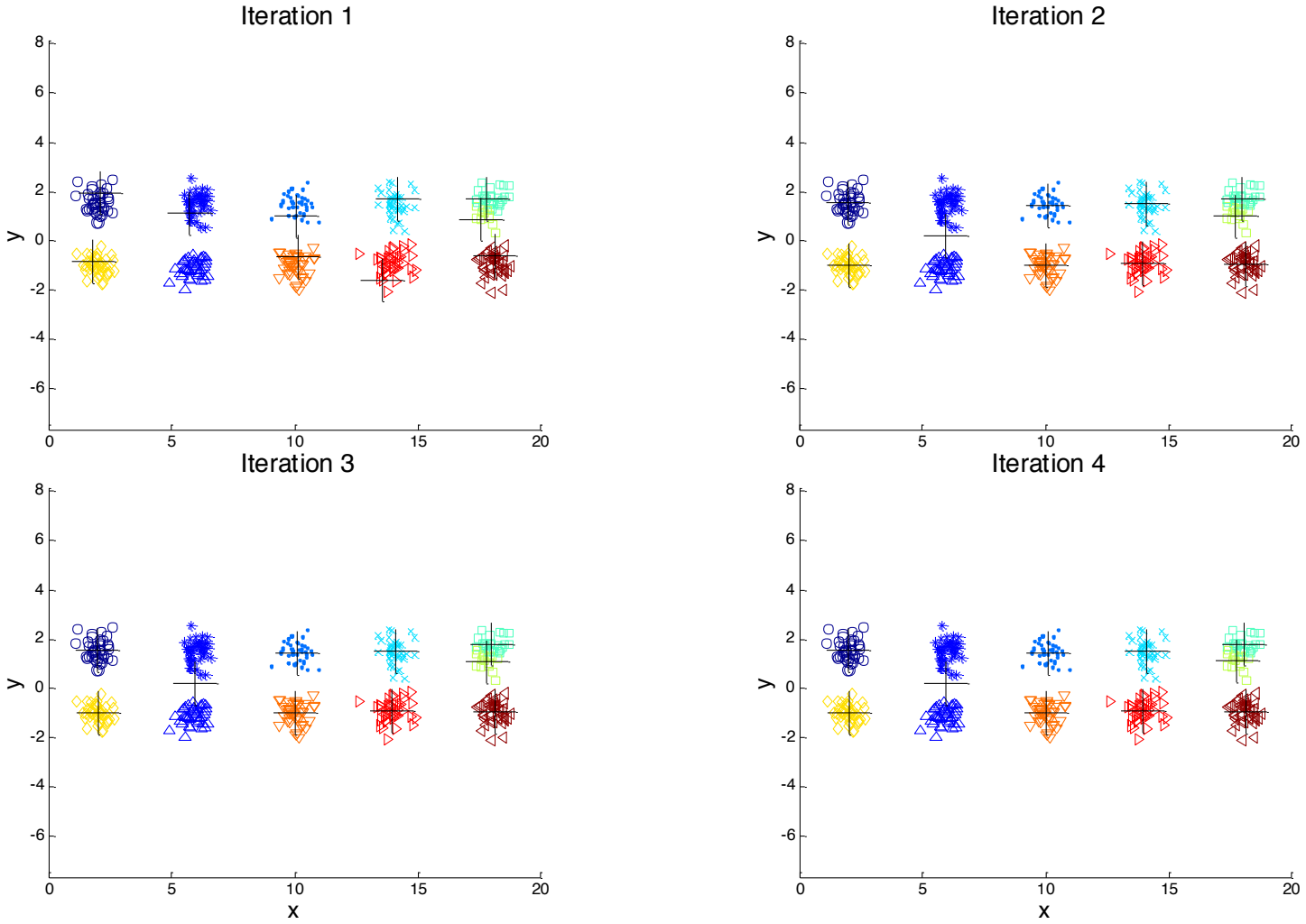
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



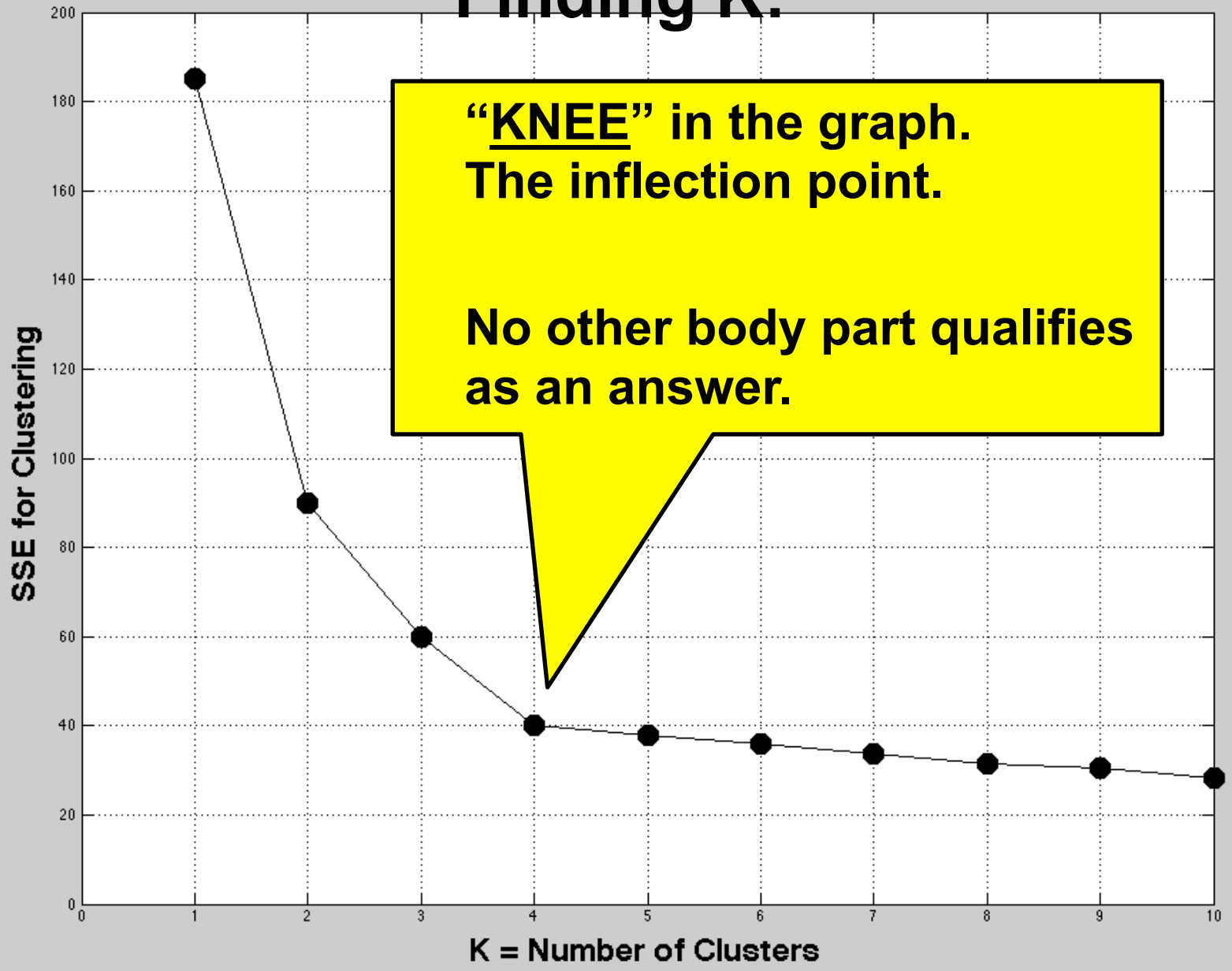
Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example



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

Finding K:


"KNEE" in the graph.
The inflection point.
No other body part qualifies as an answer.
SSE for Clustering
K = Number of Clusters
0
20
40
60
80
100
120
140
160
180
200
0
1
2
3
4
5
6
7
8
9
10

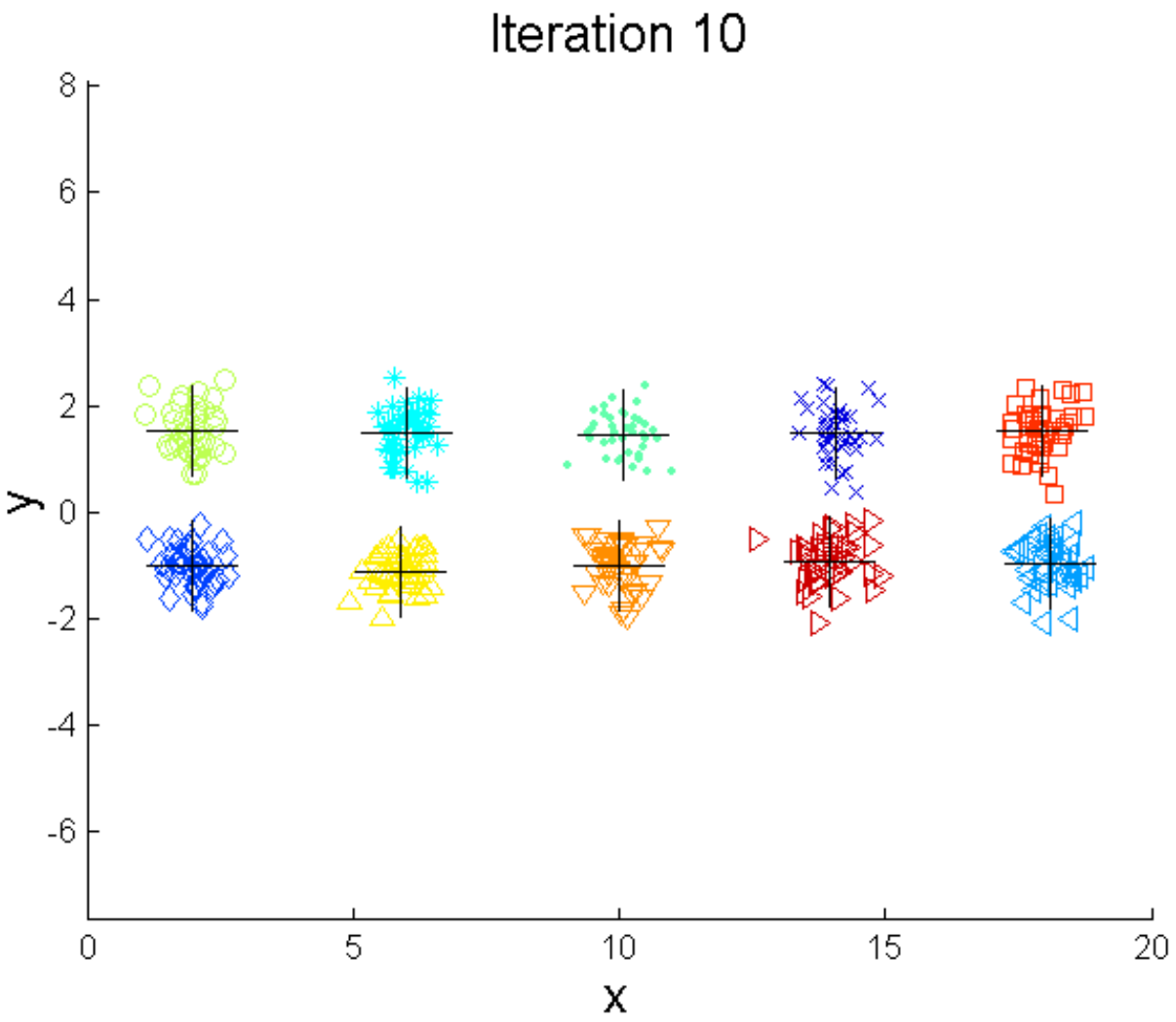
Solutions to Picking the best Initial Centroids Problem

- Multiple runs
  - Helps, but combinatorics is not on your side for many dimensions
- 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
- Postprocessing
- Bisecting K-means
  - Not as susceptible

Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - 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 is a variant

Bisecting K-means Example



Divisive – Bisecting K-means Algorithm

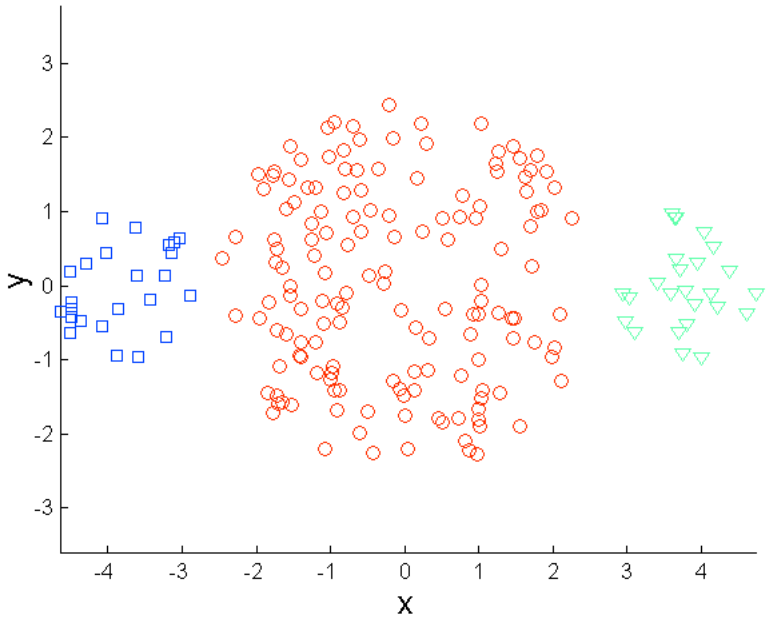
Variant of K-means that can produce a partitional or a hierarchical clustering

1. Initialize the list of clusters to contain all data points (all records)
2. repeat:
   A. Select a cluster from the list of clusters
   B. for counter=1 to number_of_iterations
      bisect the selected cluster using the basic k-Means into two clusters
   C. add the two new clusters from the bisection with the lowest SSE to the list of clusters
3. until – the list of clusters contains K clusters

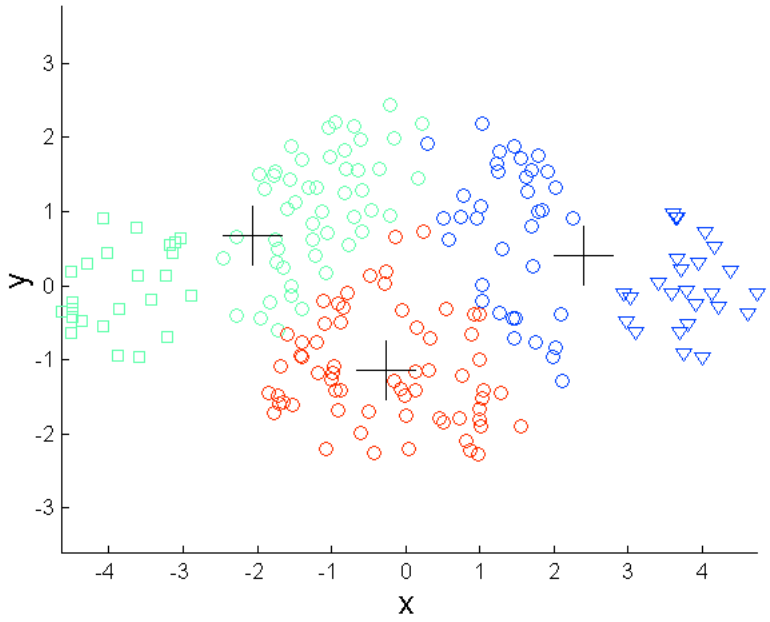
Limitations of K-means

- Has problems when clusters are of differing:
  - Sizes
  - Densities
  - Non-globular shapes (non-convex regions)
- Has problems when the data contains outliers
- YOU HAVE TO CHOOSE K!

Limitations of K-means: Differing Sizes

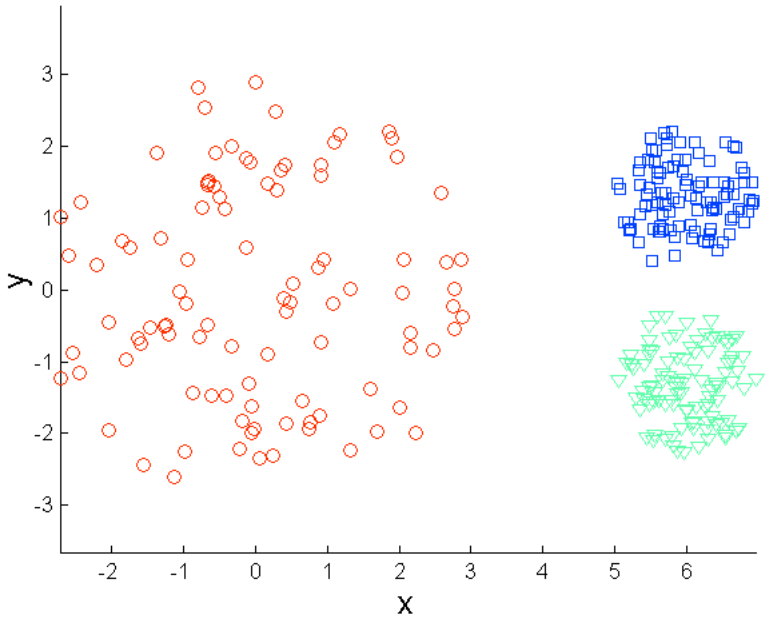


–Original Points

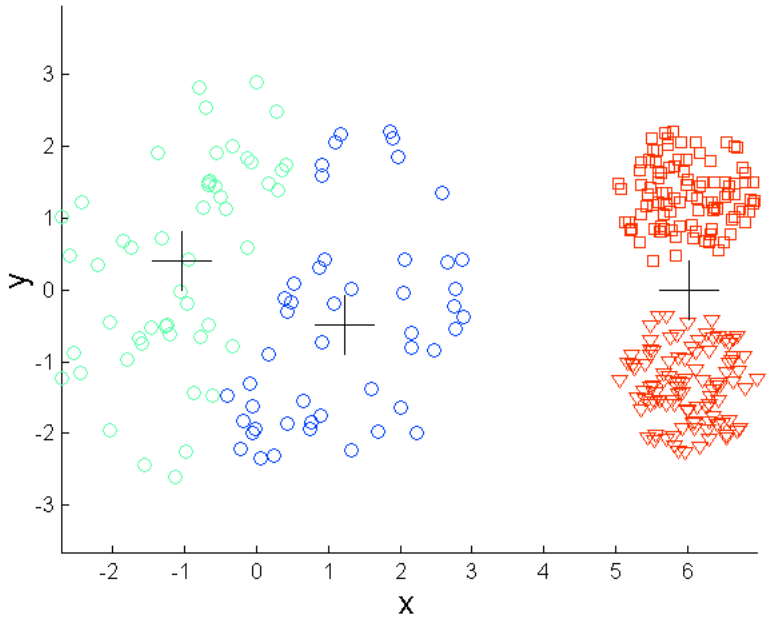


–K-means (3 Clusters)

Limitations of K-means: Differing Density

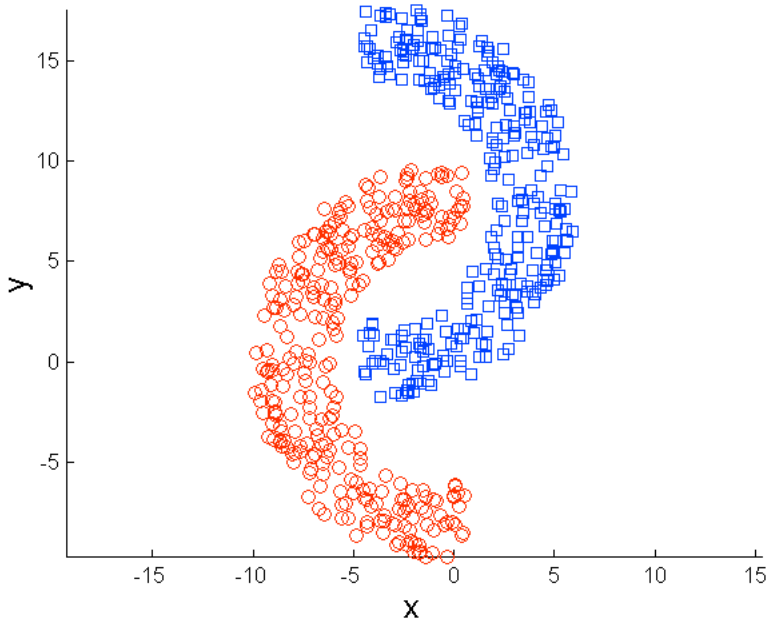


–Original Points

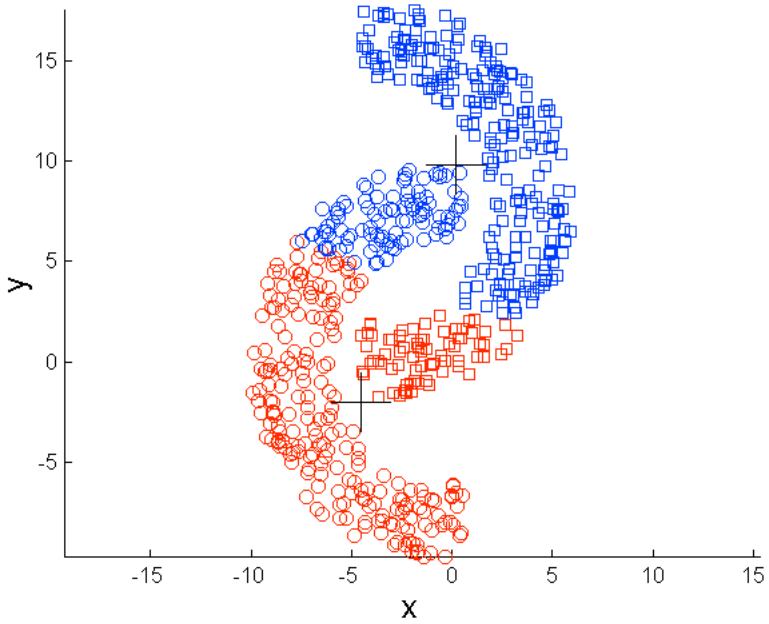


–K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

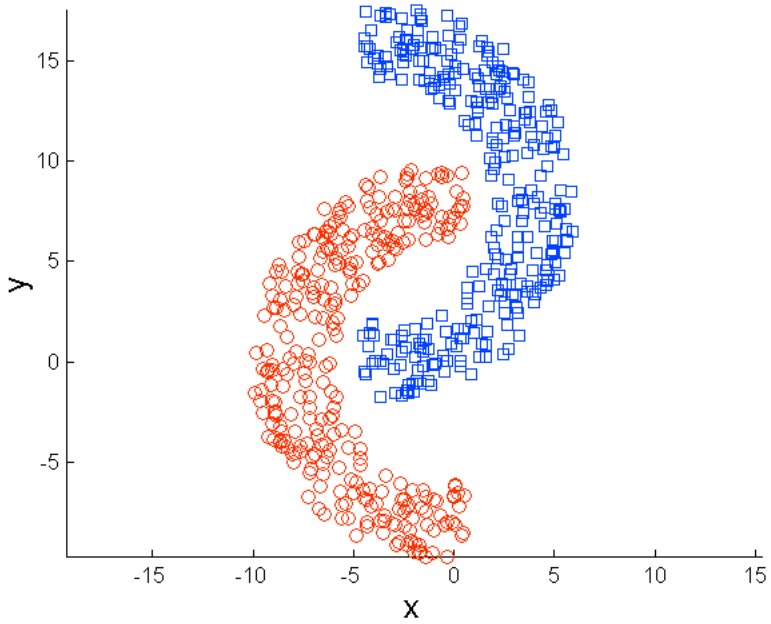


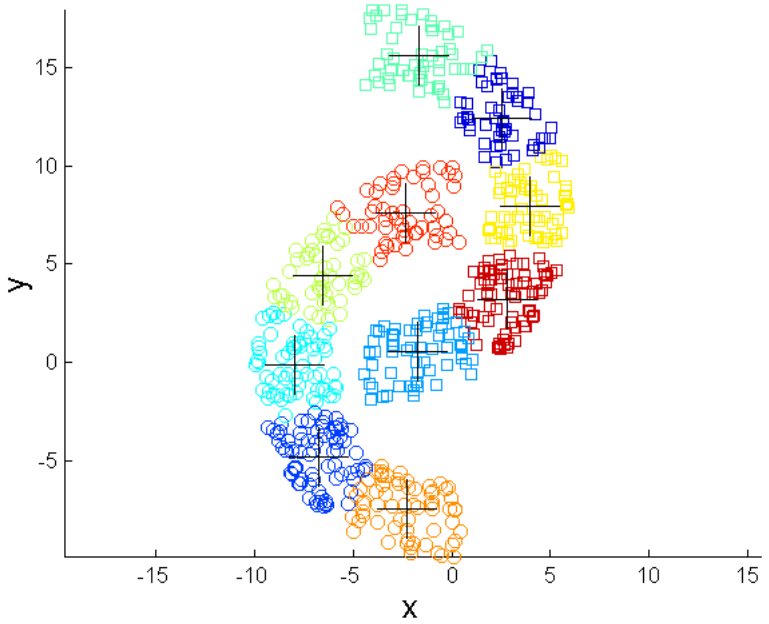
–Original Points



–K-means
(2 Non-Convex Clusters)

Overcoming K-means Limitations

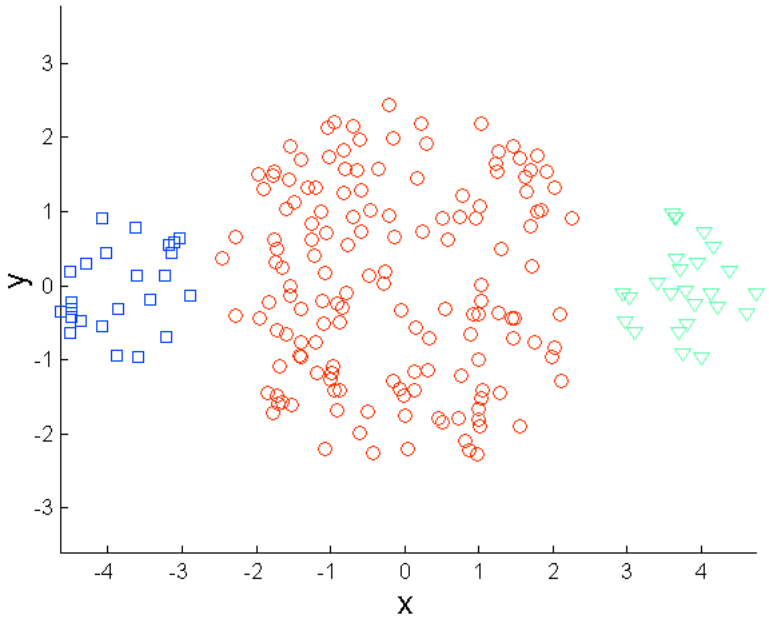


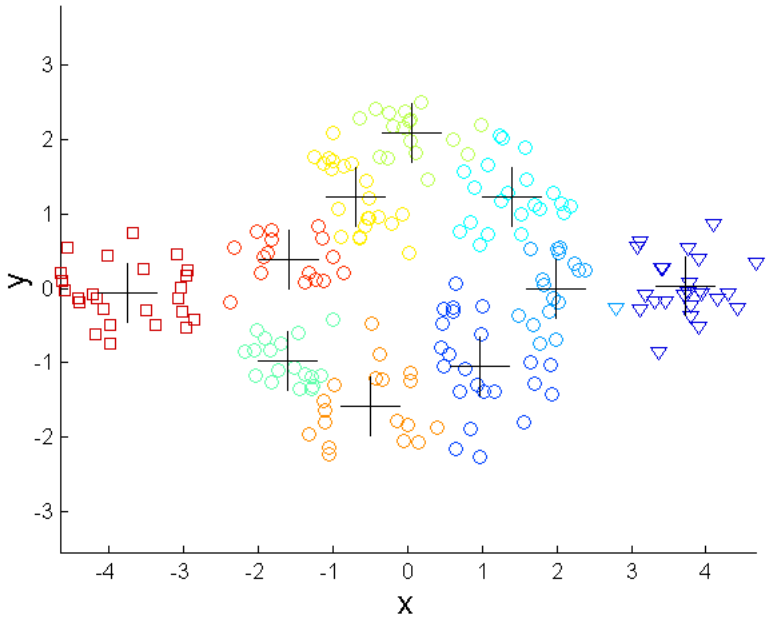


–Original Points

K-means Clusters

Overcoming K-means Limitations





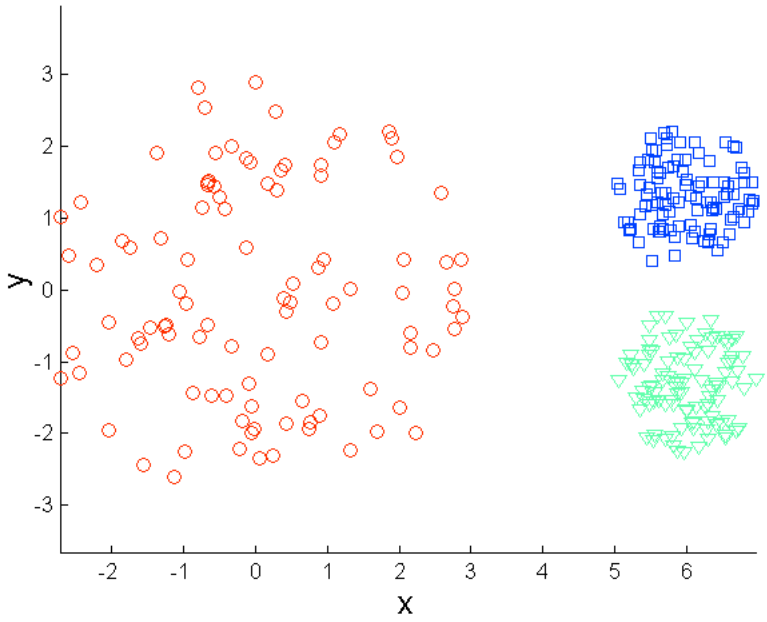
–Original Points

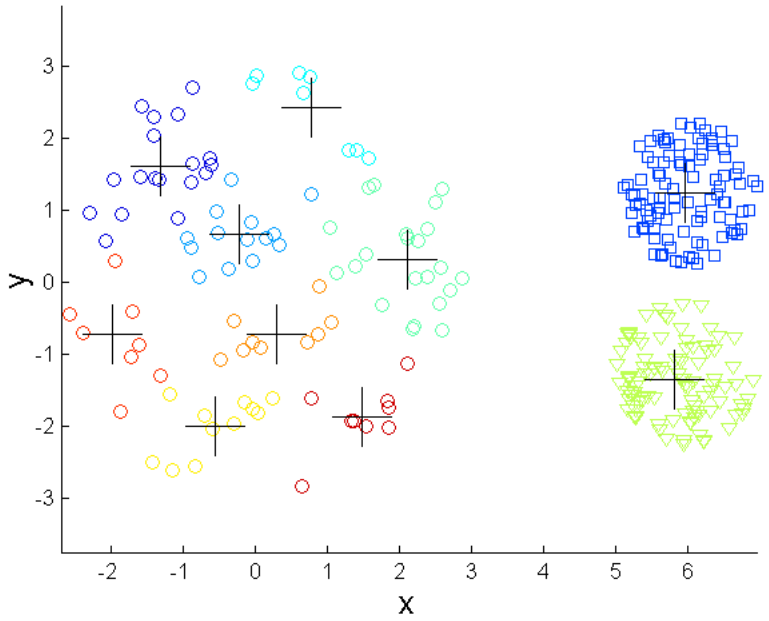
K-means Clusters

–One solution is to use many clusters.

–Find parts of clusters, but need to put together.

Overcoming K-means Limitations





–Original Points

K-means Clusters

Questions about K-means

- Why cluster or segment at all?
- What decisions do you need to make when using K-means? (distance metric is one.)
- What shortcomings are there to K-Means?
- How are these shortcomings avoided or worked around?

END