Data Mining
Cluster Analysis: Basic Concepts and Algorithms

Slides credit:

1. Introduction to Data Mining, 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

2. Mining of Massive Datasets at Stanford University

High Dimensional Data

Given a cloud of data points we want to understand its structure

The Problem of Clustering

Given a **set of points**, with a notion of **distance** between points, **group the points** into some number of ***clusters***, so that

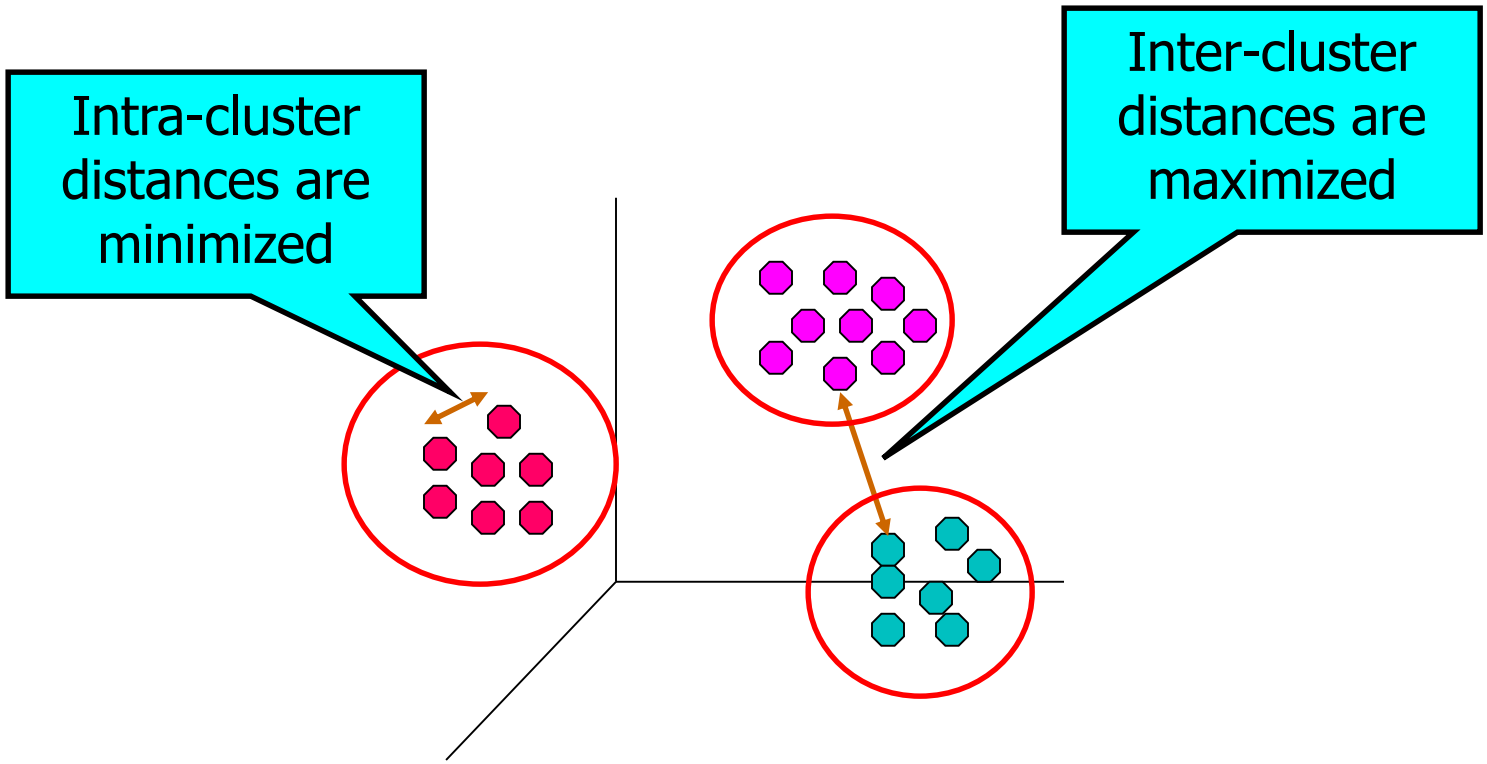
- Members of a cluster are close/similar to each other
- Members of different clusters are dissimilar

Usually:

- Points are in a high-dimensional space
- Similarity is defined using a distance measure
  - Euclidean, Cosine, Jaccard, edit distance, …

What is Cluster Analysis

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



Notion of a Cluster can be Ambiguous

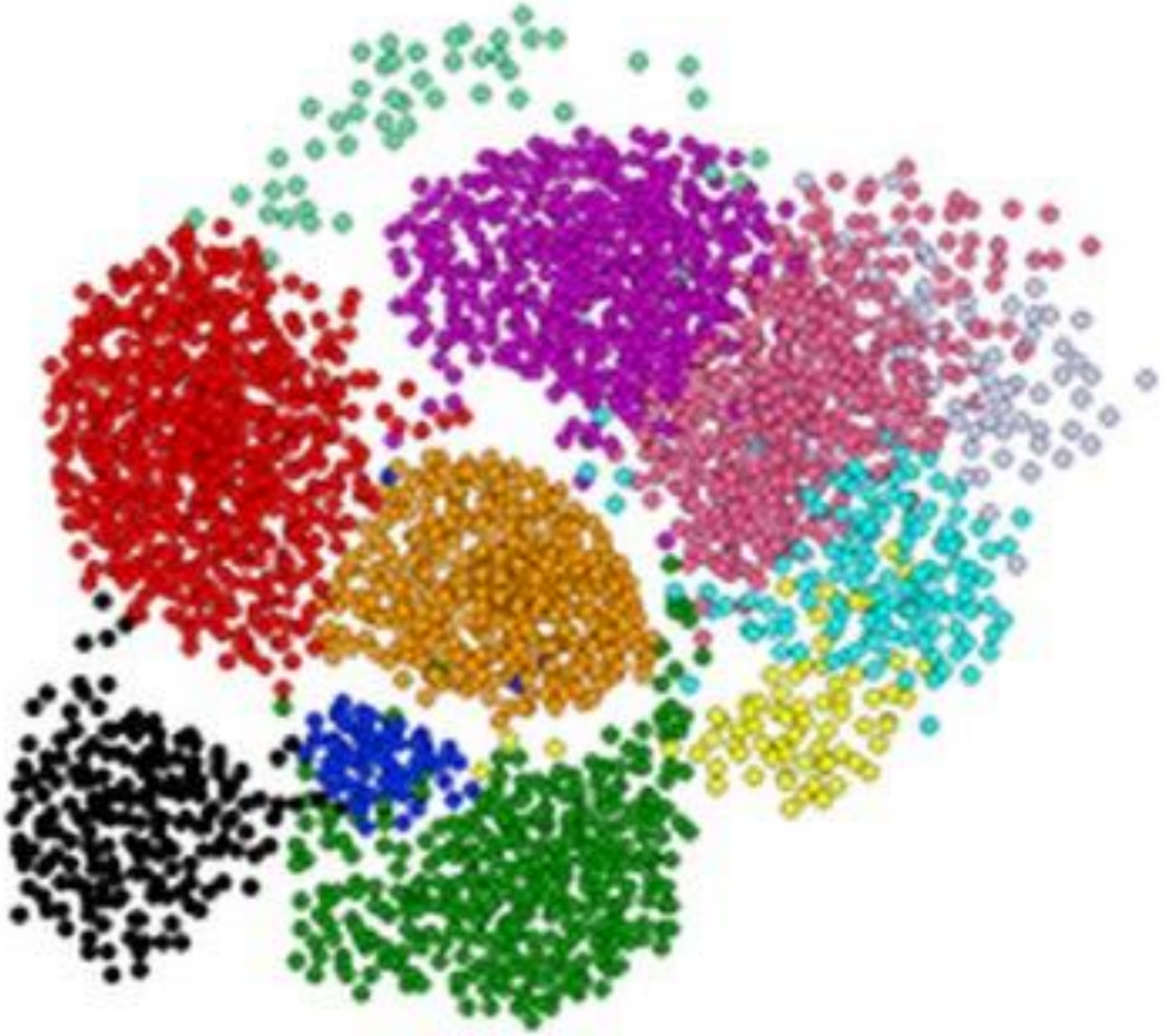
How many clusters?

Six Clusters

Two Clusters

Four Clusters

Clustering is a hard problem!



Clustering Problem: Galaxies

A catalog of 2 billion "sky objects" represents objects by their radiation in 7 dimensions (frequency bands)

Problem: Cluster into similar objects, e.g., galaxies, nearby stars, quasars, etc.

Sloan Digital Sky Survey

Clustering Problem: Music CDs

Intuitively: Music divides into categories, and customers prefer a few categories

- But what are categories really?

Represent a CD by a set of customers who bought it:

Similar CDs have similar sets of customers, and vice-versa

Clustering Problem: Music CDs

Space of all CDs:

Think of a space with one dim. for each customer

- Values in a dimension may be 0 or 1 only
- A CD is a point in this space (x_1, x_2,…, x_k), where x_i = 1 iff the i^{th} customer bought the CD

For Amazon, the dimension is tens of millions

Task: Find clusters of similar CDs

Clustering Problem: Documents

Finding topics:

Represent a document by a vector $(x_1, x_2, \ldots, x_k)$, where $x_i = 1$ iff the i^{th} word (in some order) appears in the document

Documents with similar sets of words may be about the same topic

Cosine, Jaccard, and Euclidean

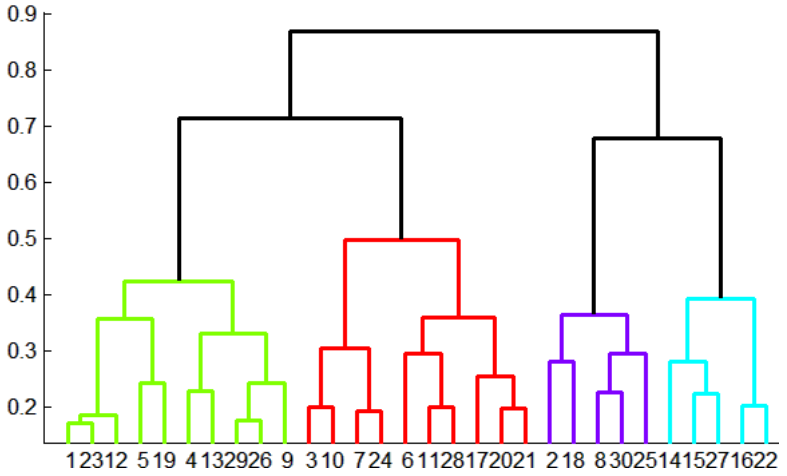
As with CDs we have a choice when we think of documents as sets of words or shingles:

- **Sets as vectors:** Measure similarity by the **cosine distance**
- **Sets as sets:** Measure similarity by the **Jaccard distance**
- **Sets as points:** Measure similarity by **Euclidean distance**

Overview: Methods of Clustering

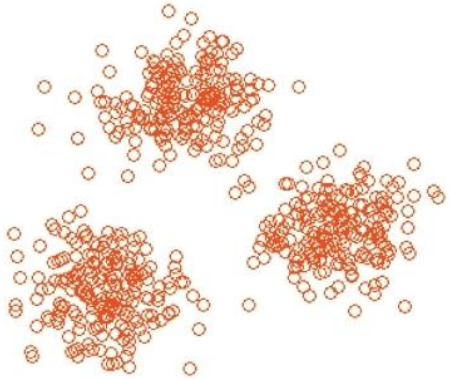
Hierarchical:

- **Agglomerative** (bottom up):
  - Initially, each point is a cluster
  - Repeatedly combine the two "nearest" clusters into one
- **Divisive** (top down):
  - Start with one cluster and recursively split it



Point assignment:

- Maintain a set of clusters
- Points belong to "nearest" cluster



Types of Clusterings

A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

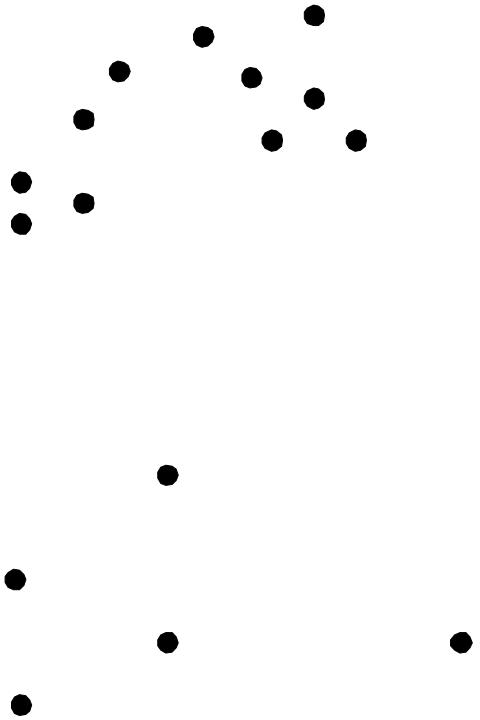
Partitional Clustering

- A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

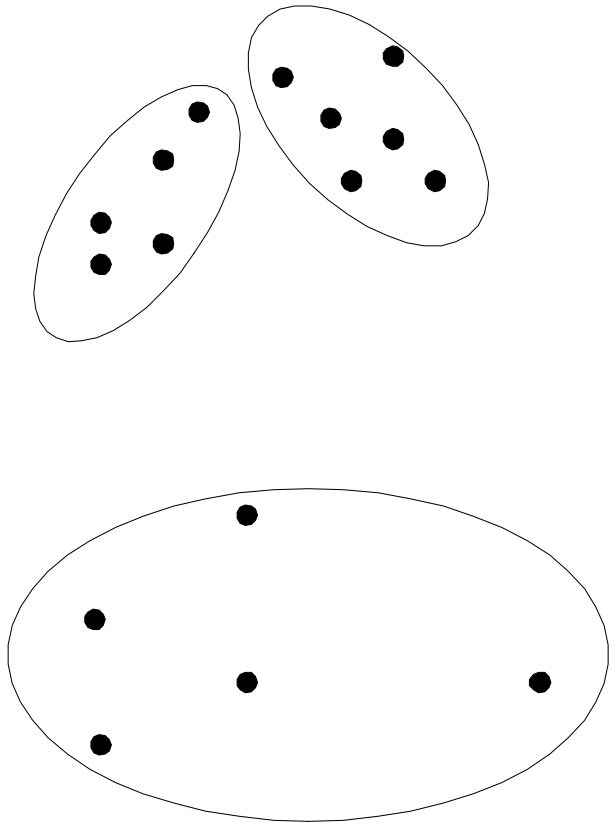
Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree

Partitional Clustering

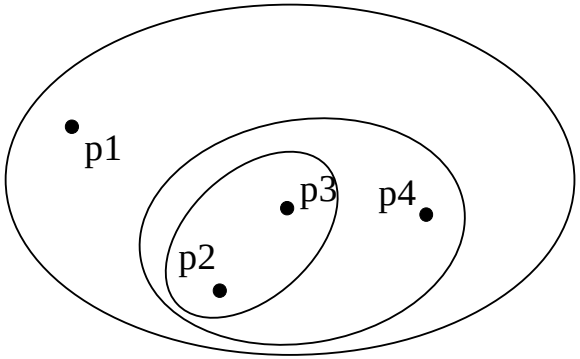


Original Points

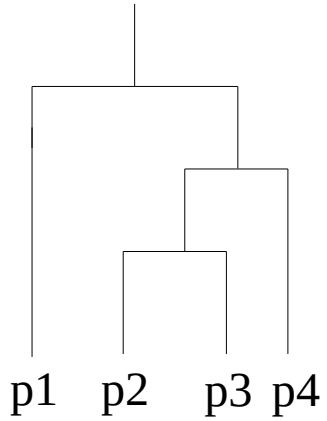


A Partitional Clustering

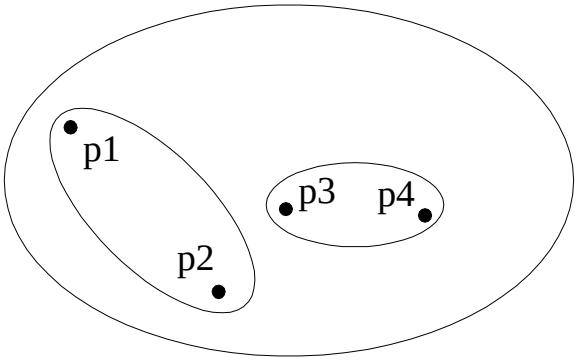
Hierarchical Clustering



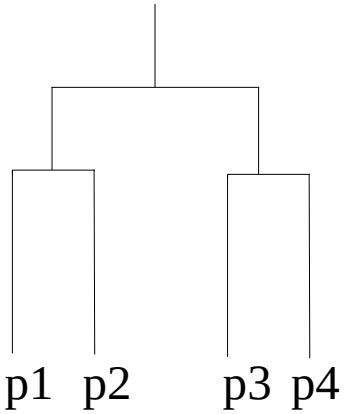
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Other Distinctions Between Sets of Clusters

Exclusive versus non-exclusive

- In non-exclusive clusterings, points may belong to multiple clusters.
- Can represent multiple classes or ‘border’ points

Fuzzy versus non-fuzzy

- 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

Heterogeneous versus homogeneous

- Clusters of widely different sizes, shapes, and densities

Objective Function for clustering

Clusters Defined by an Objective Function

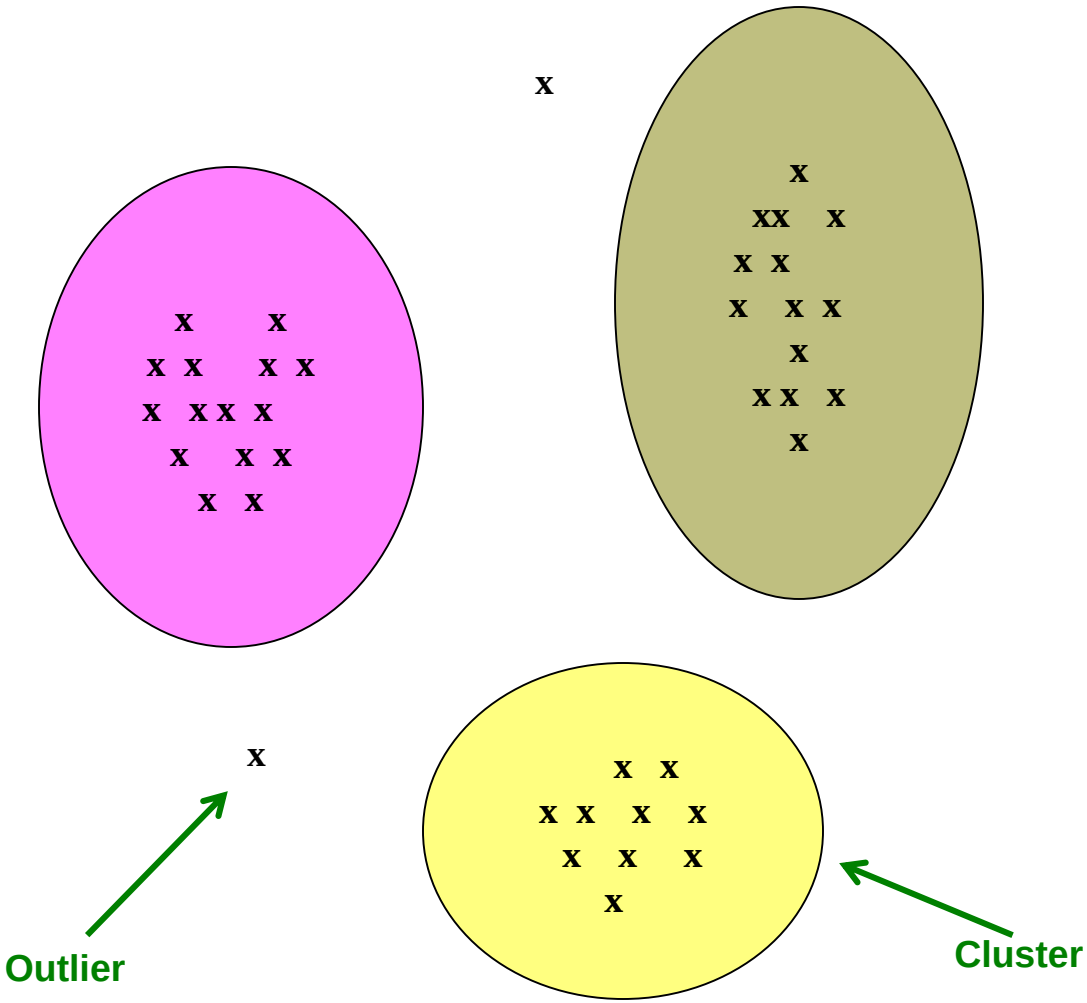
- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.
  - Hierarchical clustering algorithms typically have local objectives
  - Partitional algorithms typically have global objectives
- A variation of the global objective function approach is to fit the data to a parameterized model.
  - Parameters for the model are determined from the data.
  - Mixture models assume that the data is a ‘mixture' of a number of statistical distributions.

Map Clustering Problem to a Different Problem

Map the clustering problem to a different domain and solve a related problem in that domain

- Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
- Clustering is equivalent to breaking the graph into connected components, one for each cluster.
- Want to minimize the edge weight between clusters and maximize the edge weight within clusters

Example: Clusters & Outliers



Clustering Algorithms

K-means and its variants

Hierarchical clustering

k–means Algorithm(s)

Assumes Euclidean space/distance

Start by picking ***k***, the number of clusters

Initialize clusters by picking one point per cluster

- **Example:** Pick one point at random, then ***k*-1** other points, each as far away as possible from the previous points

Populating Clusters

1) For each point, place it in the cluster whose current centroid it is nearest

2) After all points are assigned, update the locations of centroids of the ***k*** clusters

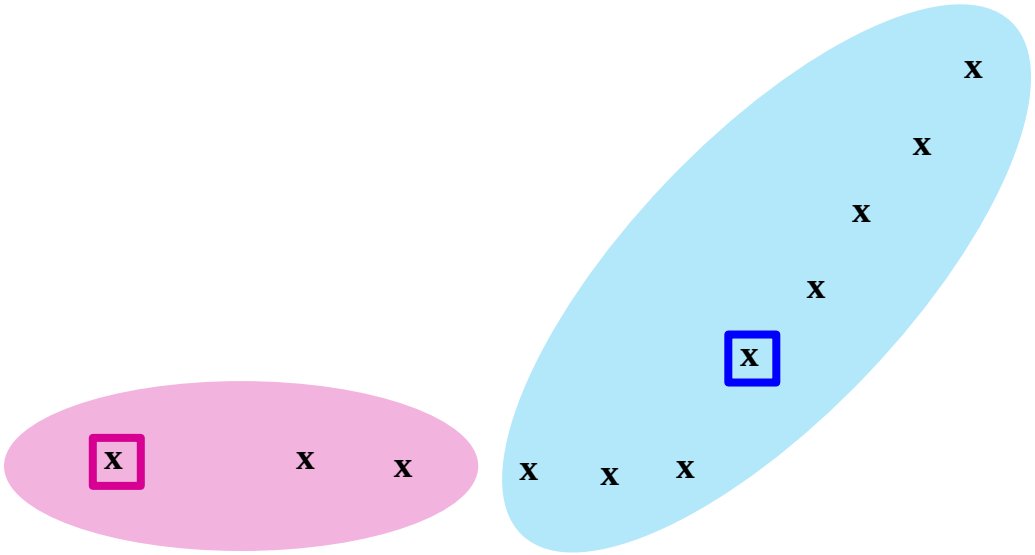
3) Reassign all points to their closest centroid

- Sometimes moves points between clusters

Repeat 2 and 3 until convergence

- **Convergence:** Points don't move between clusters and centroids stabilize

Example: Assigning Clusters

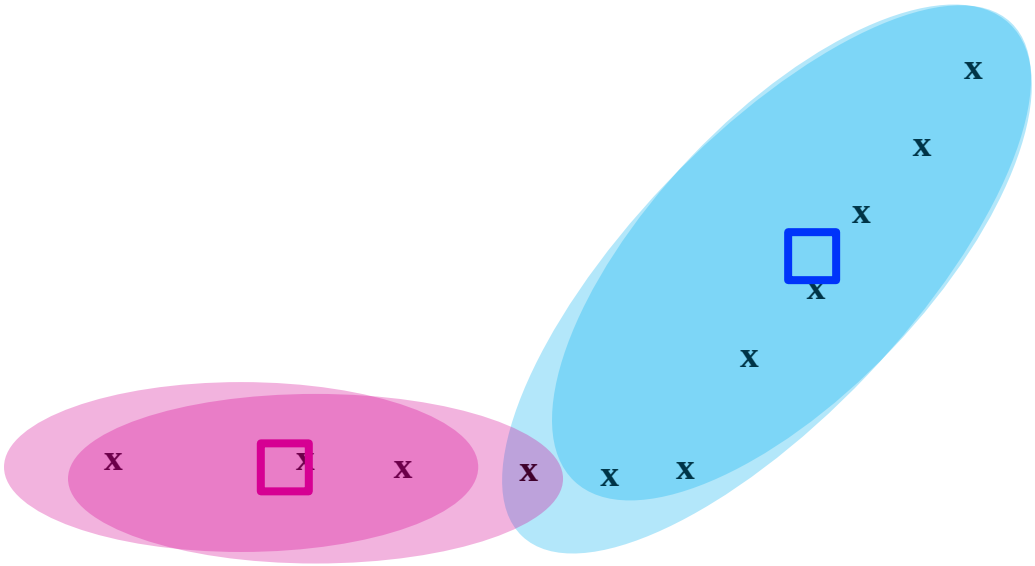


x … data point

… centroid

Clusters after round 1

Example: Assigning Clusters

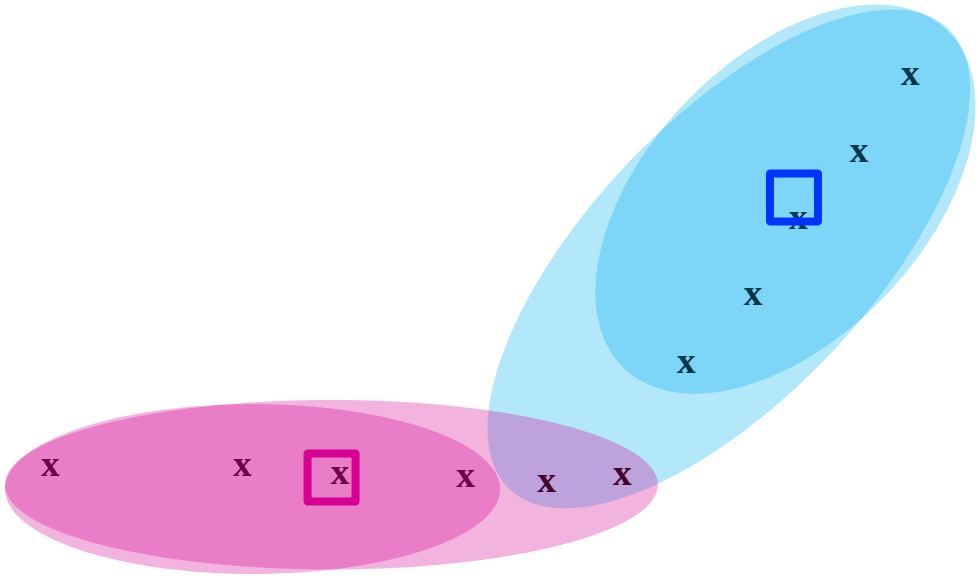


x … data point

… centroid

Clusters after round 2

Example: Assigning Clusters



x … data point
… centroid

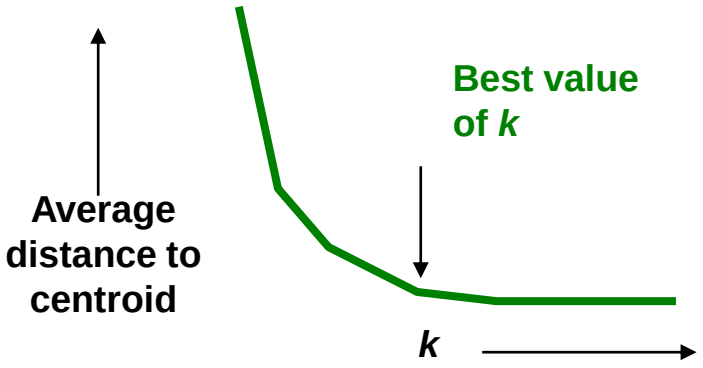
Clusters at the end

Getting the *k* right

How to select *k*?

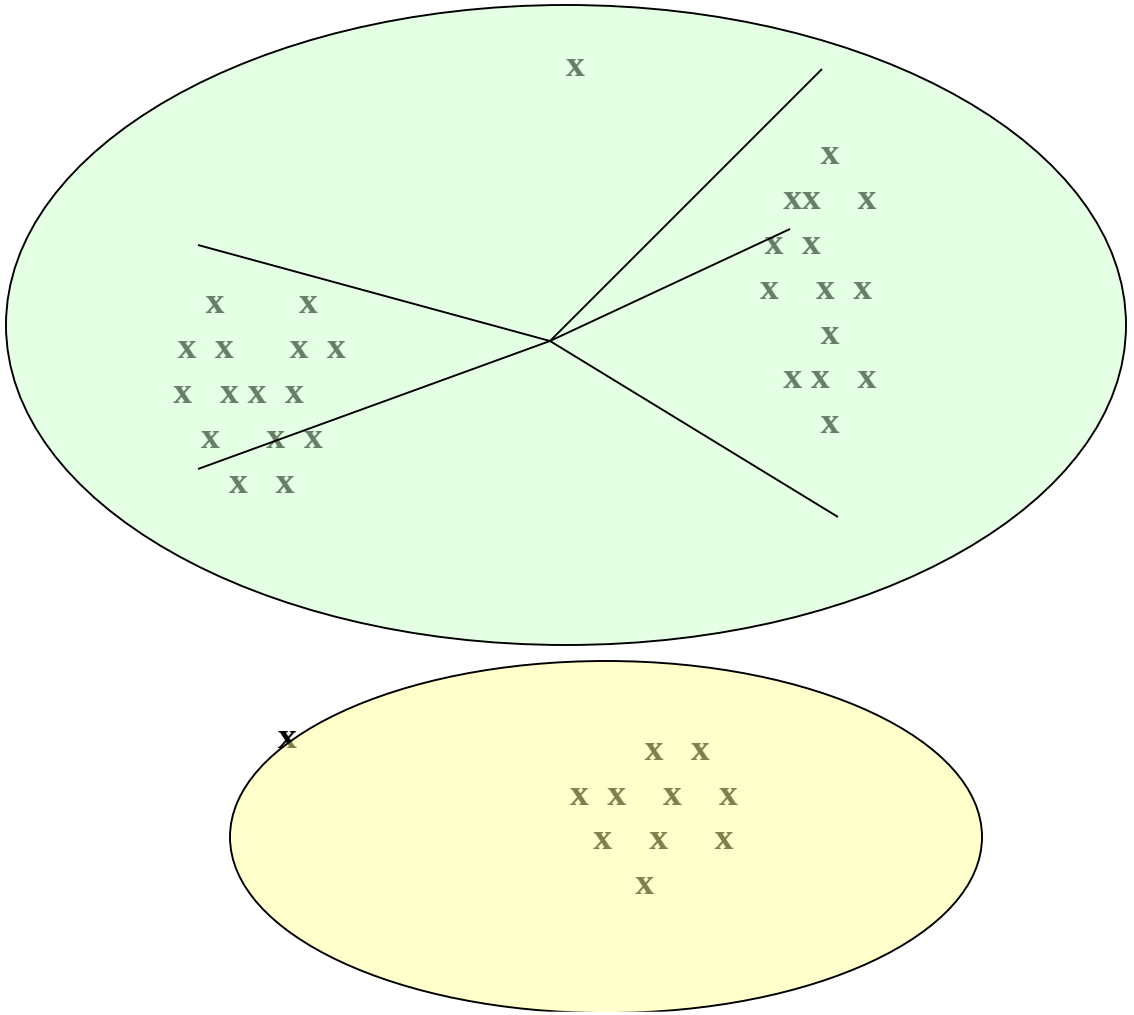
Try different **k**, looking at the change in the average distance to centroid as **k** increases

Average falls rapidly until right **k**, then changes little



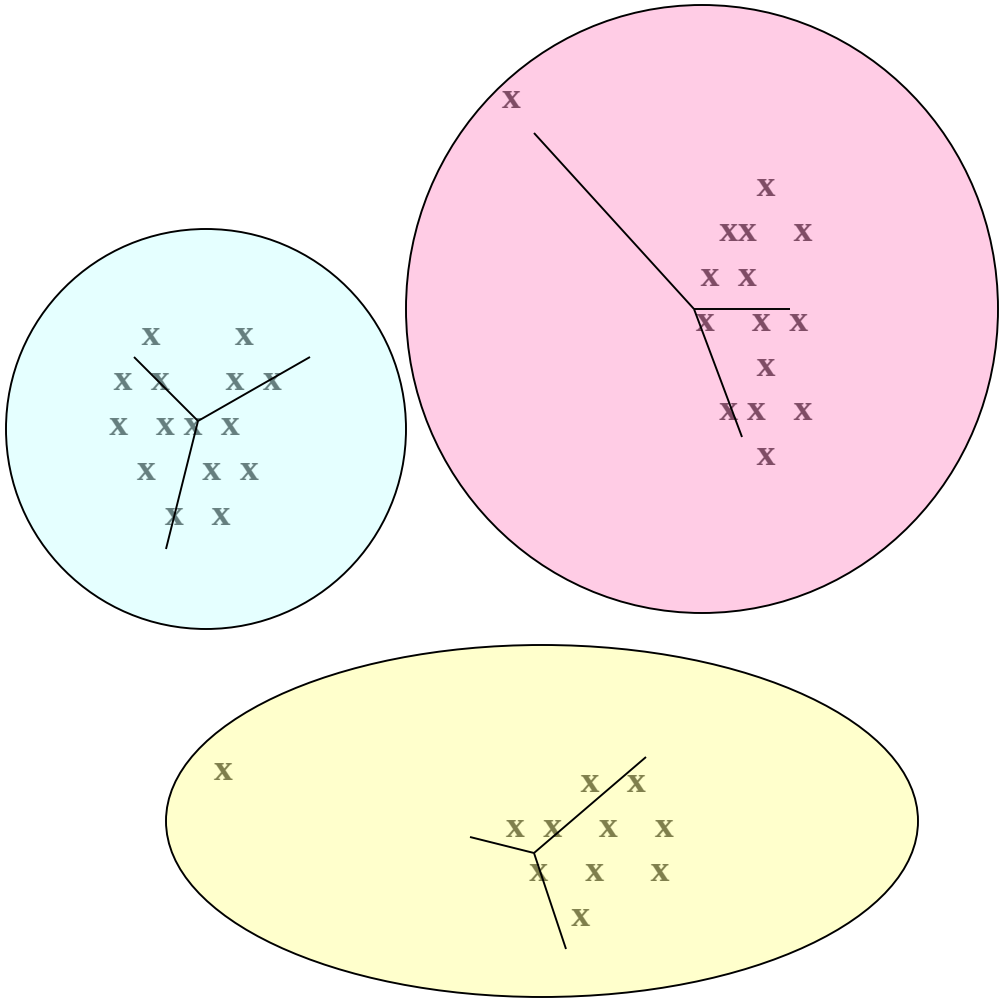
Example: Picking *k*

Too few; many long distances to centroid.



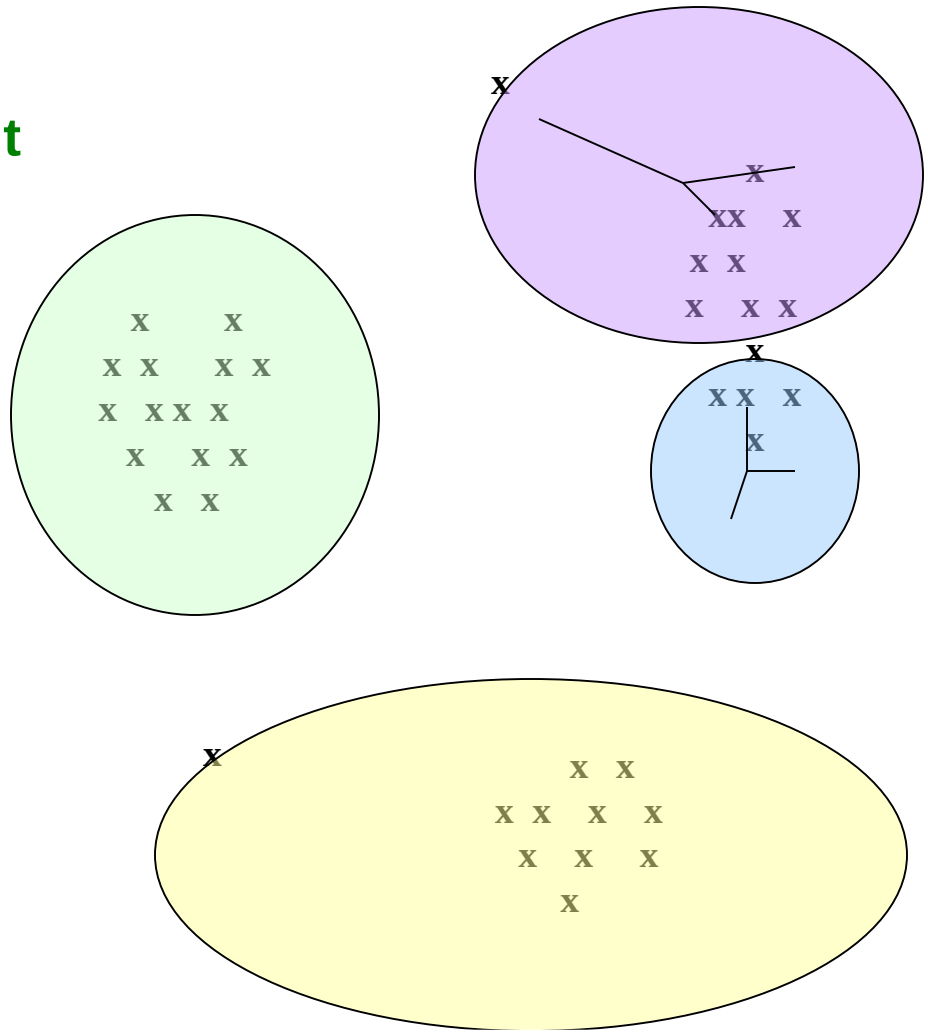
Example: Picking *k*

Just right; distances rather short.

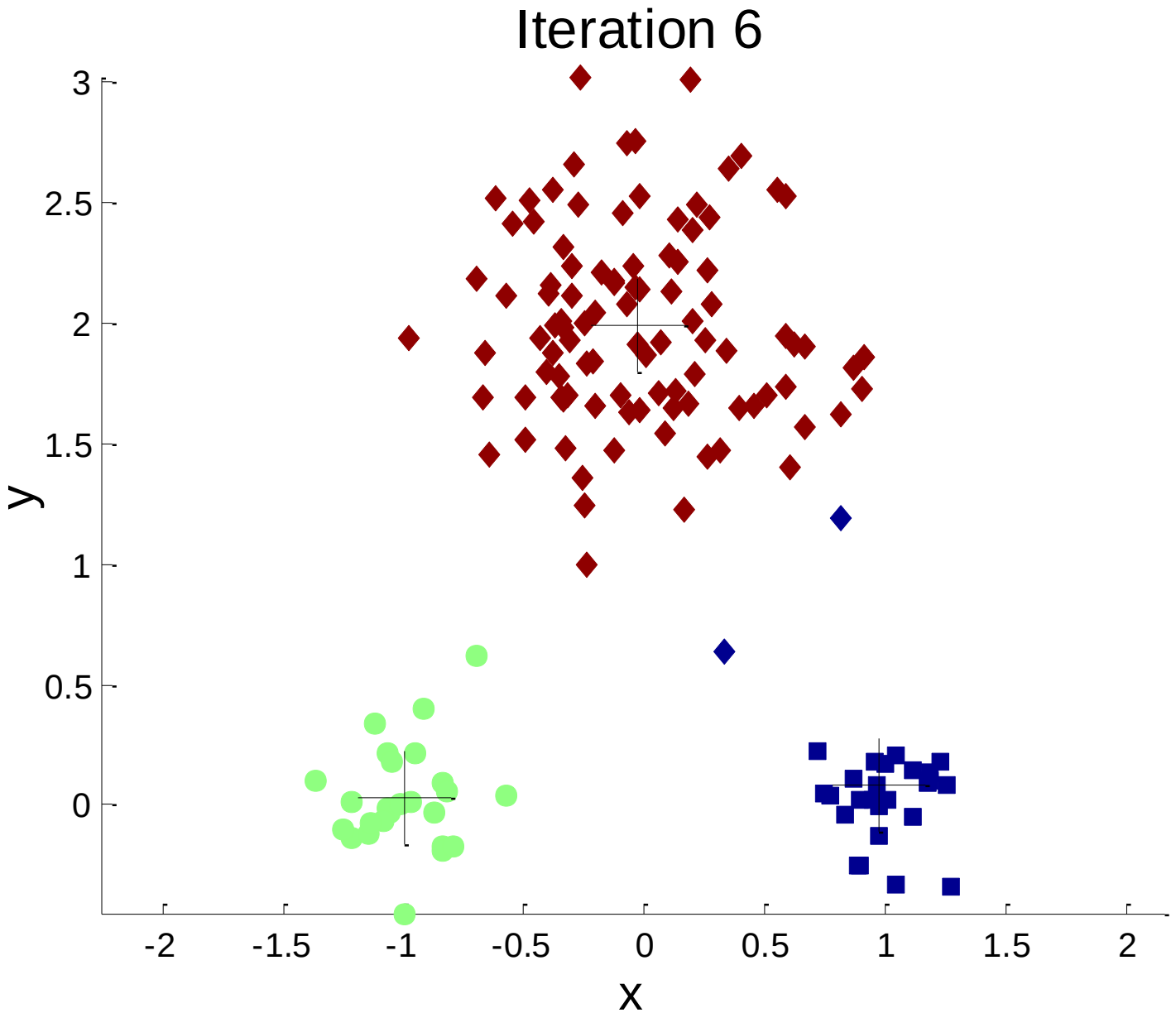


Example: Picking *k*

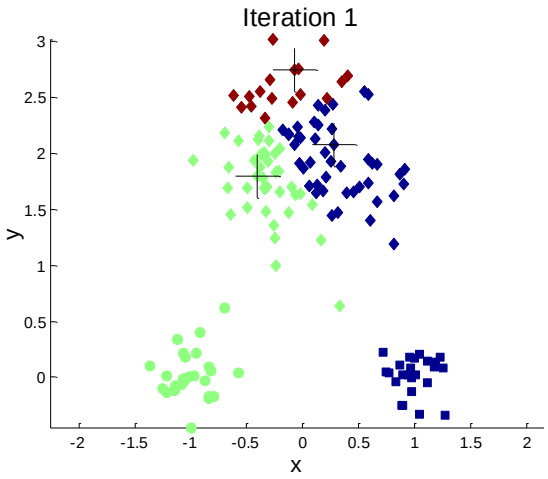
**Too many;
little improvement
in average
distance.**

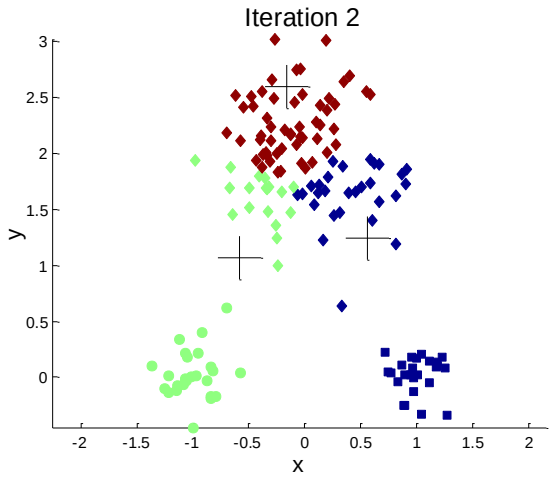


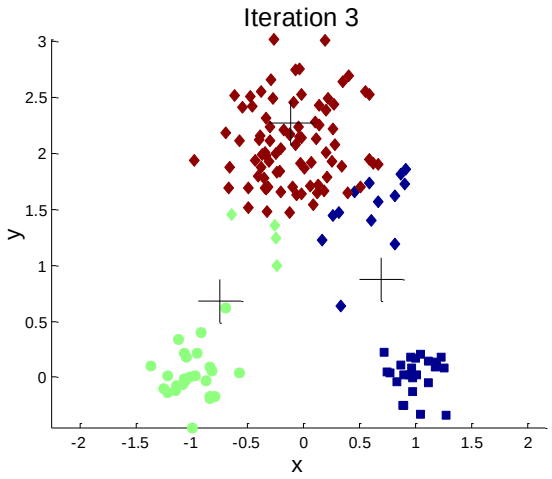
Example of K-means Clustering

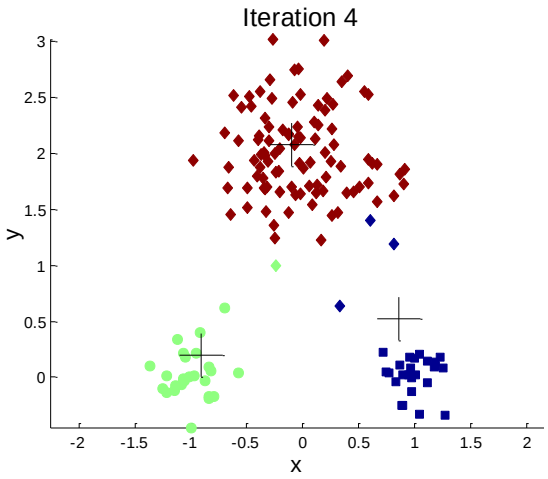


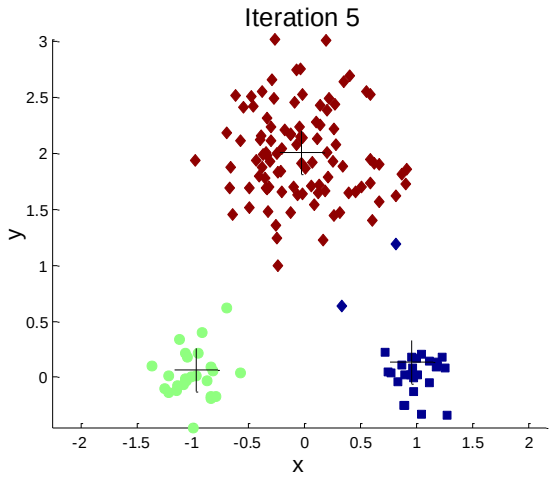
Example of K-means Clustering

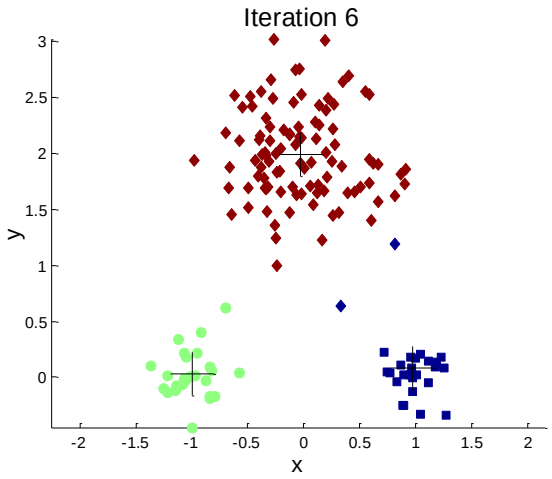












K-means Clustering – Details

Initial centroids are often chosen randomly.

- Clusters produced vary from one run to another.

The centroid is (typically) the mean of the points in the cluster.

'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.

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'

Complexity is O(n * K * I * d)

- n = number of points, K = number of clusters,
  I = number of iterations, d = number of attributes

Evaluating K-means Clusters

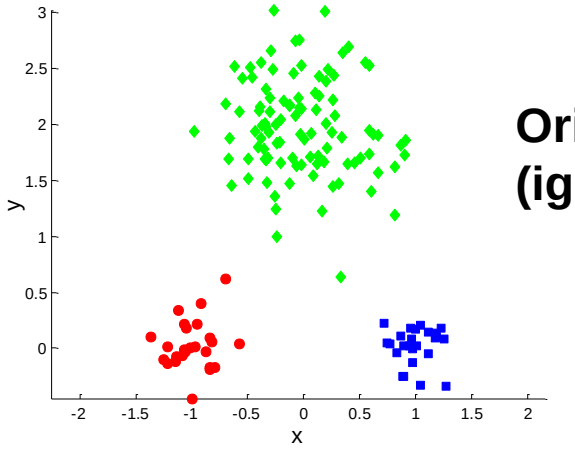
Most common measure is Sum of Squared Error (SSE)

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

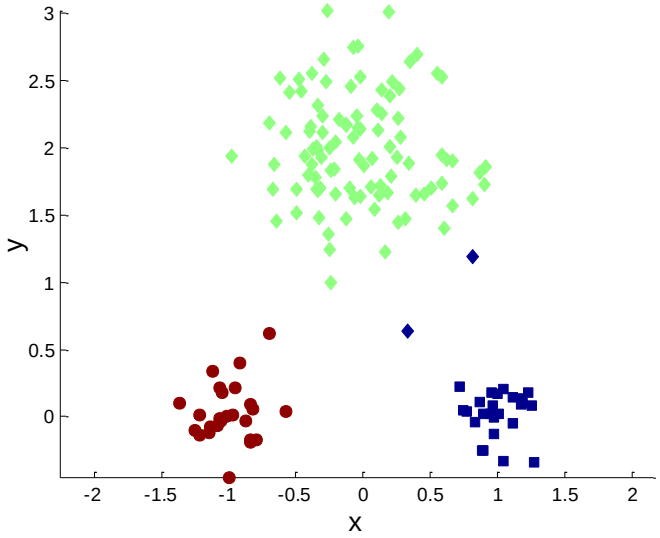
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- 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 sets of clusters, we prefer the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

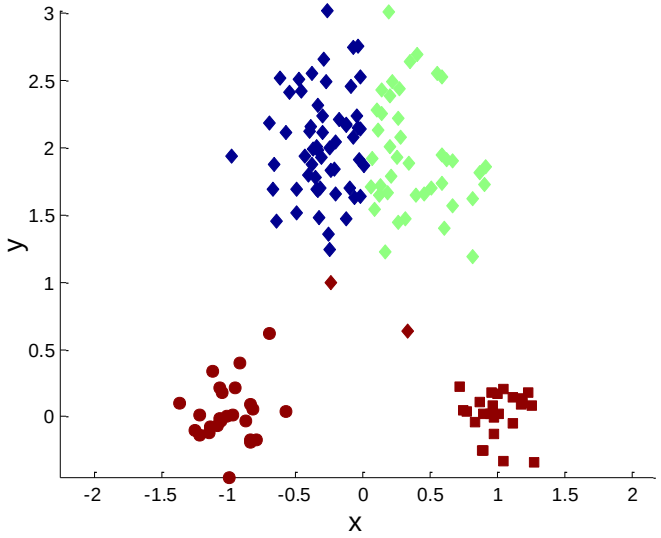
Two different K-means Clusterings



Original Points
(ignore the color)



Optimal Clustering



Sub-optimal Clustering

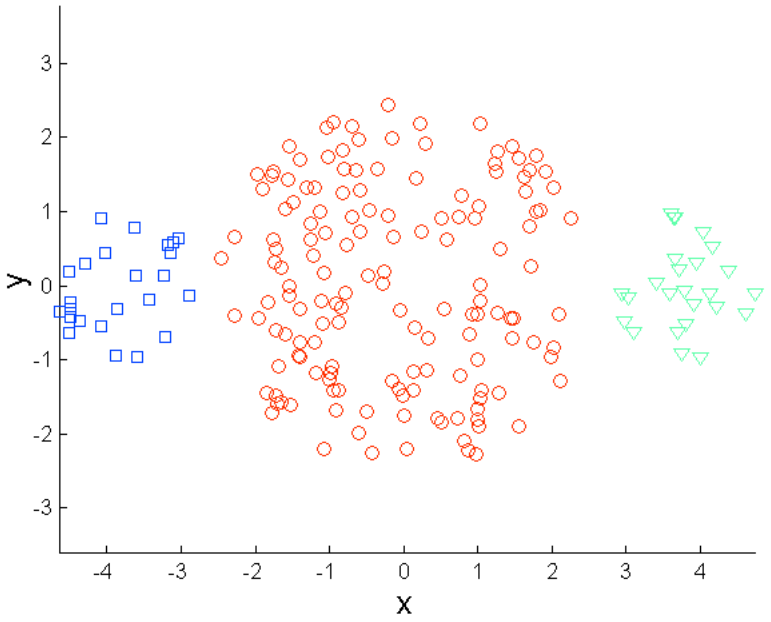
Limitations of K-means

K-means has problems when clusters are of differing

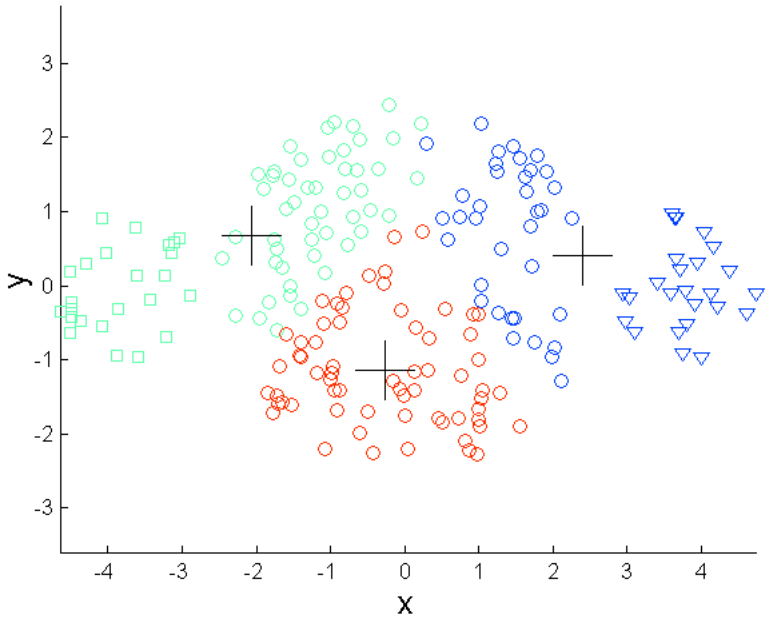
- Sizes
- Densities
- Non-globular shapes

K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

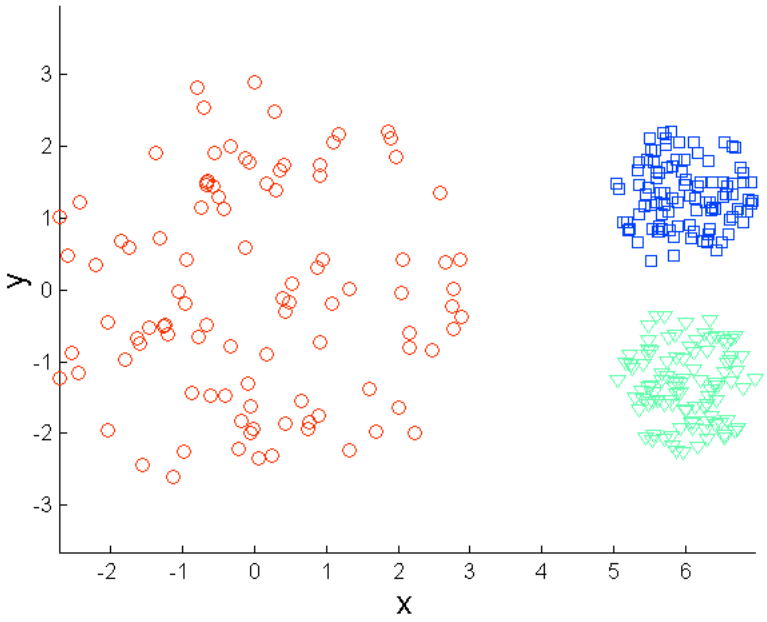


Original Points

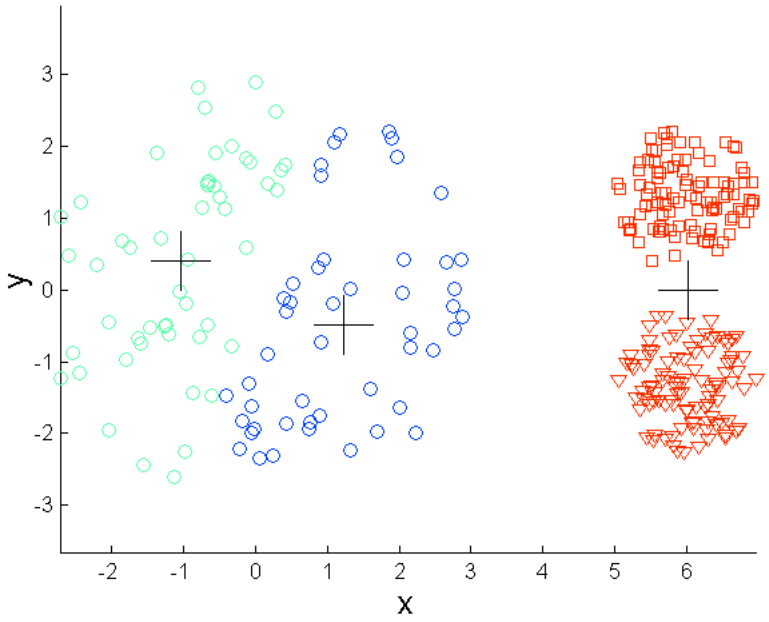


K-means (3 Clusters)

Limitations of K-means: Differing Density

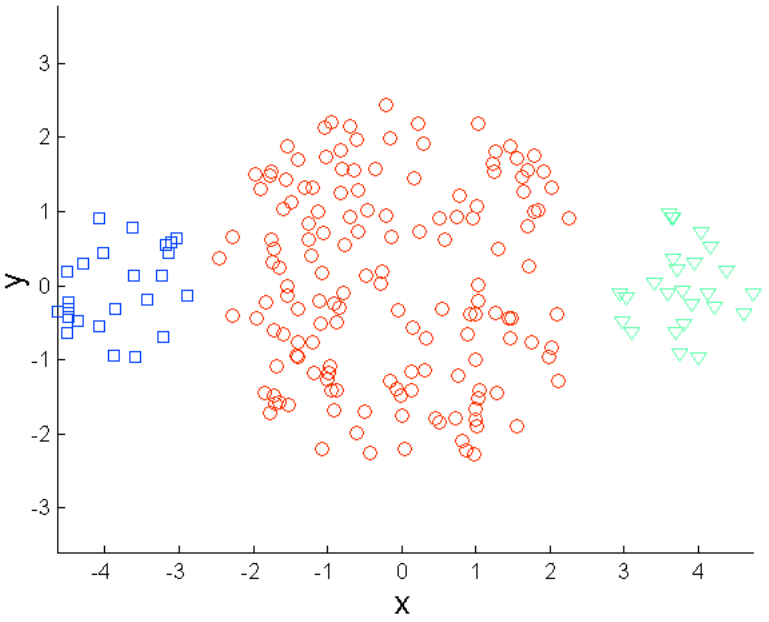


Original Points

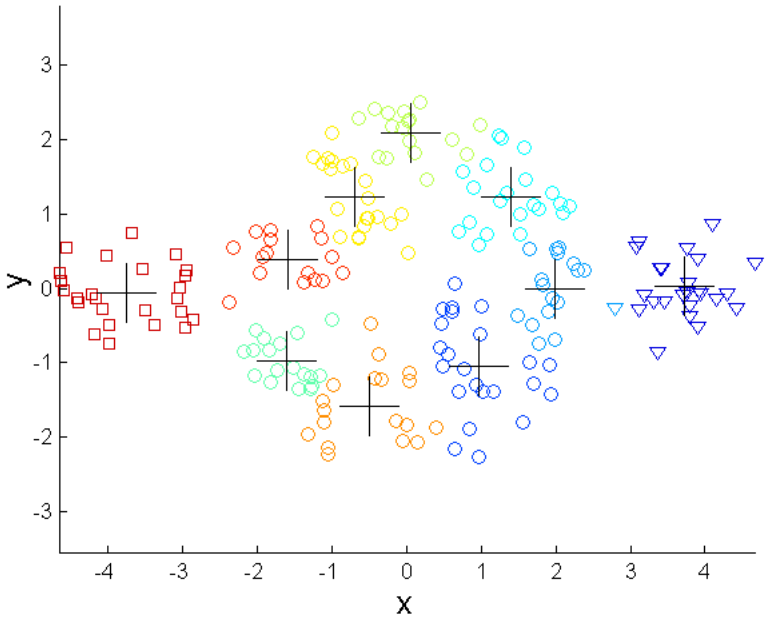


K-means (3 Clusters)

Overcoming K-means Limitations



Original Points

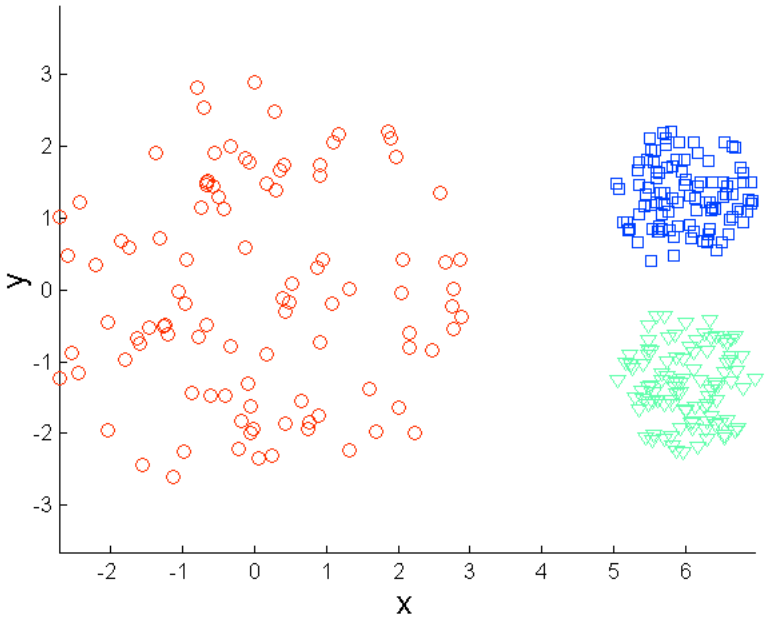


K-means Clusters

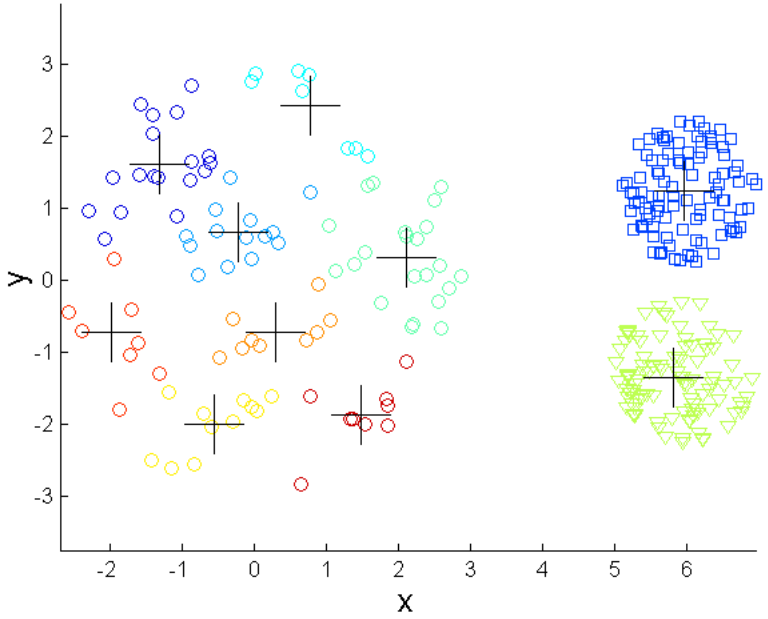
One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations

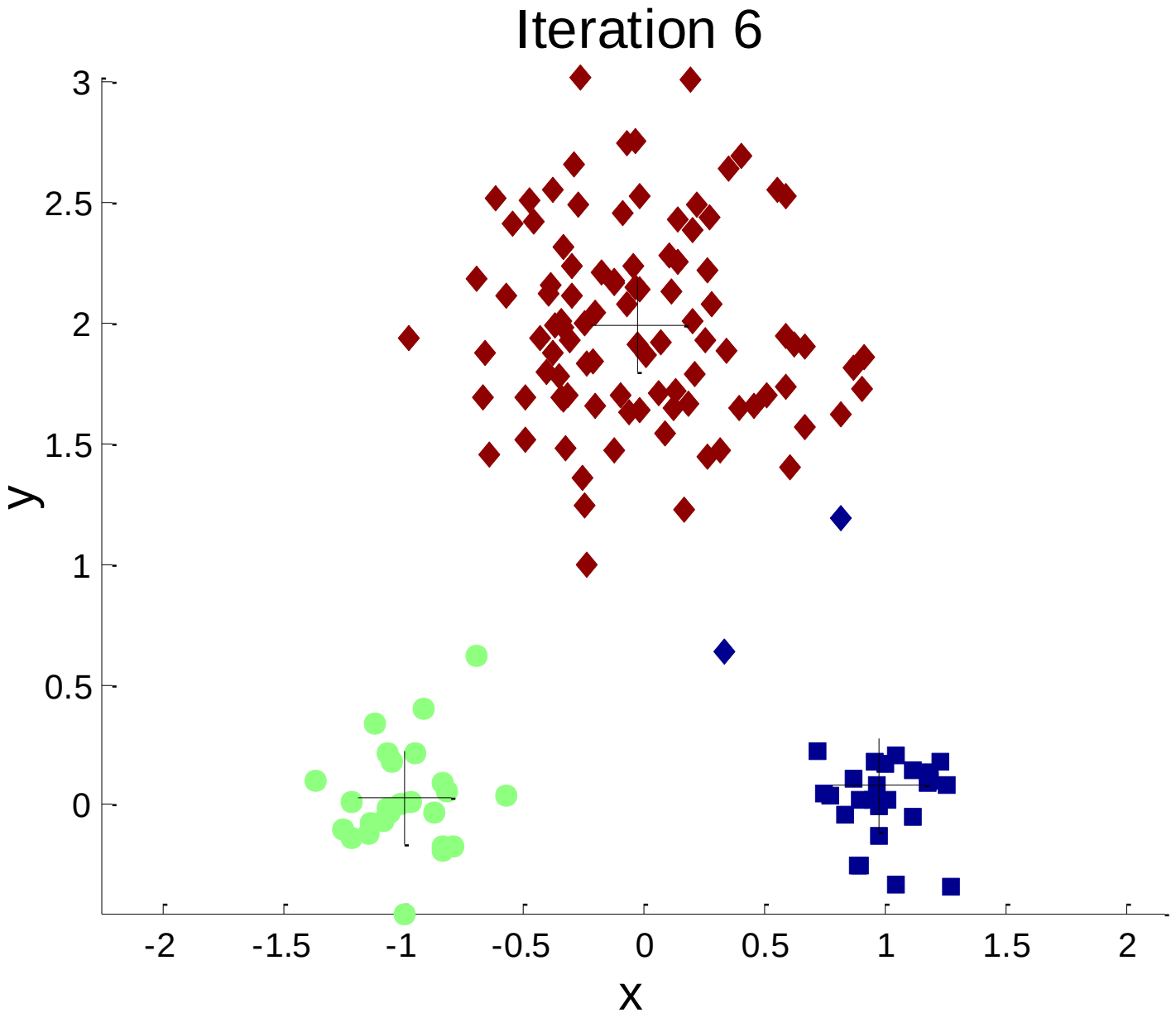


Original Points

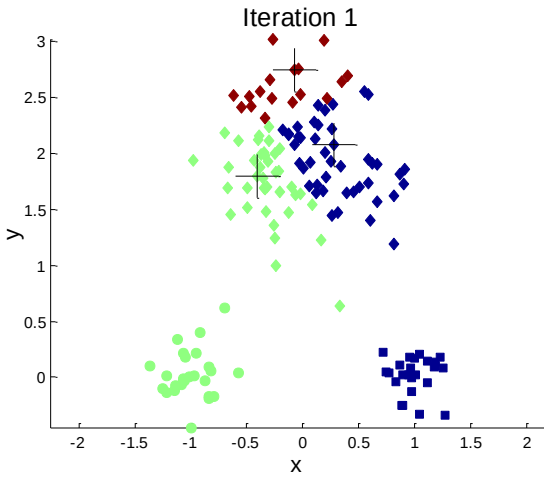


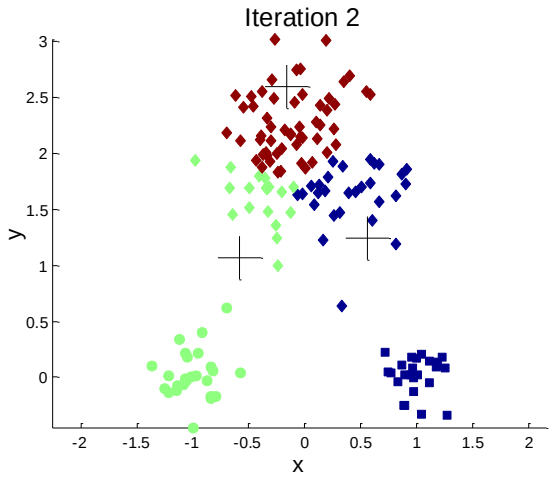
K-means Clusters

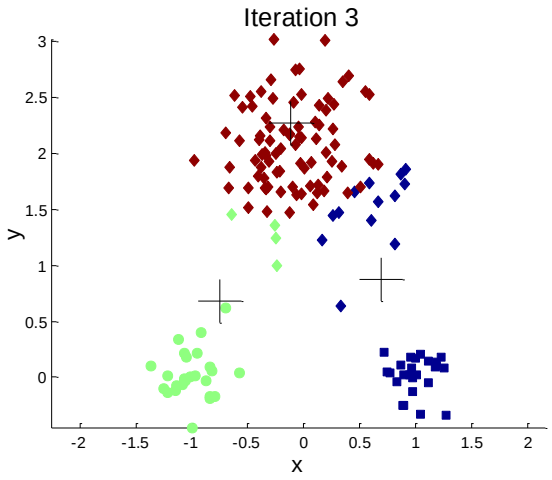
Importance of Choosing Initial Centroids

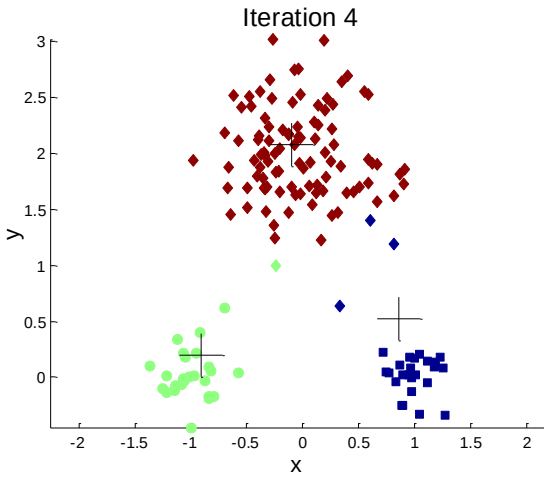


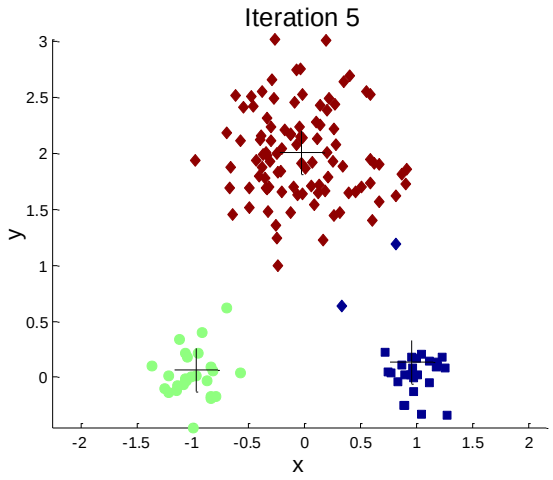
Importance of Choosing Initial Centroids

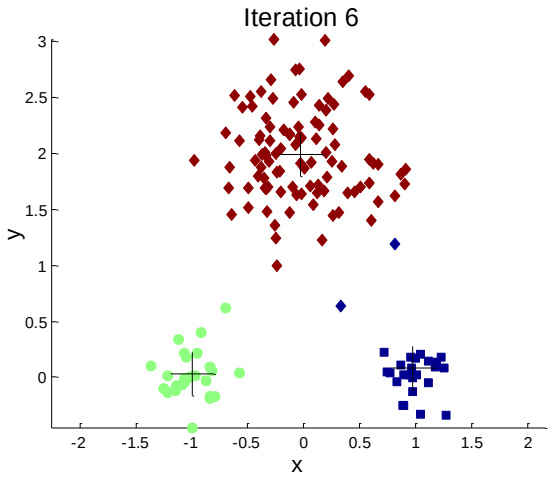




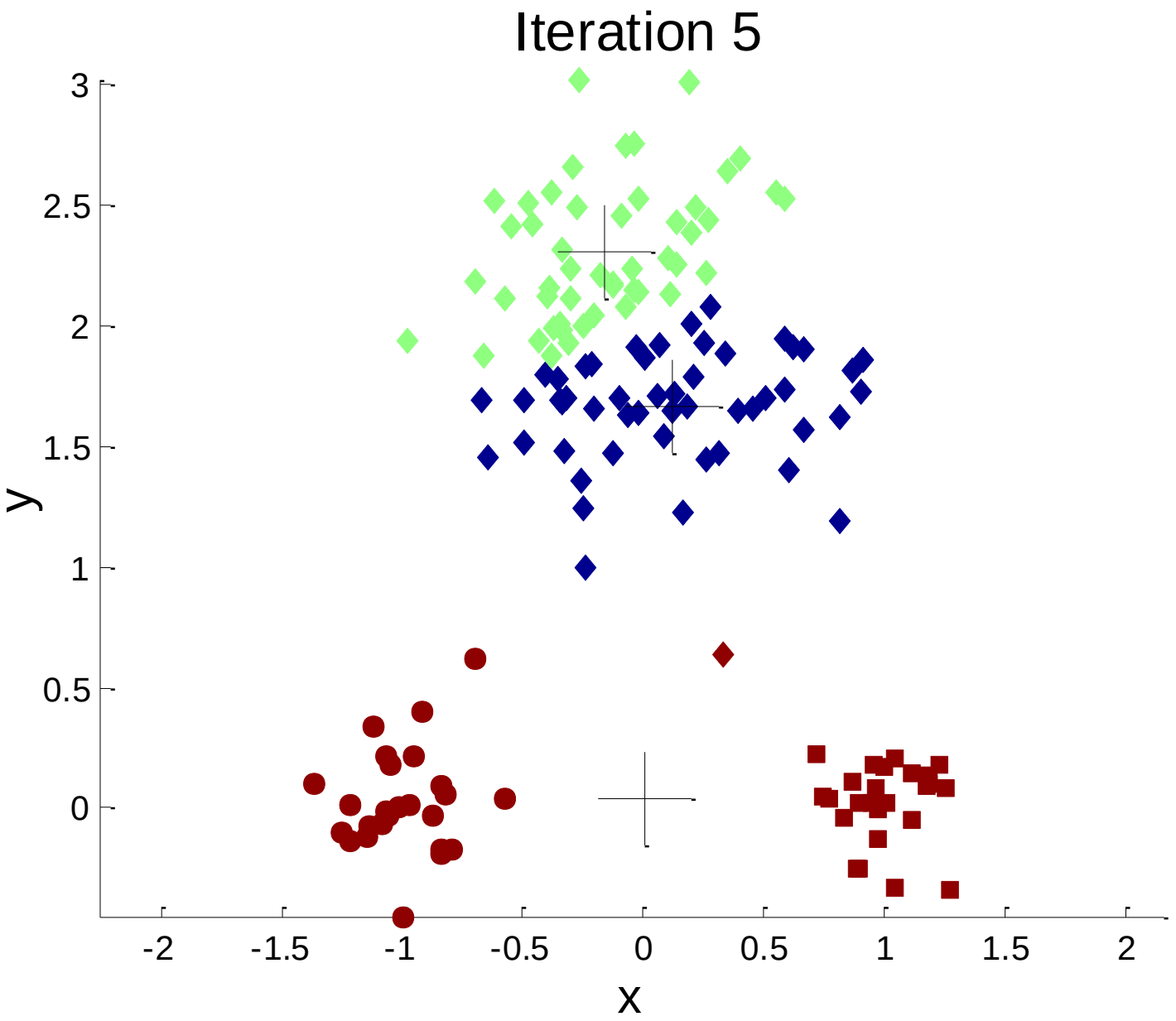




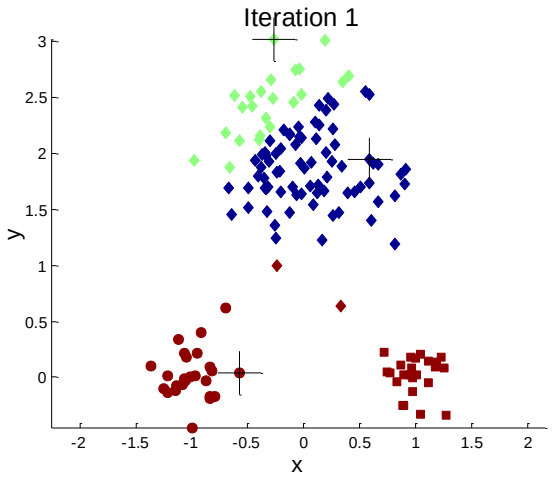


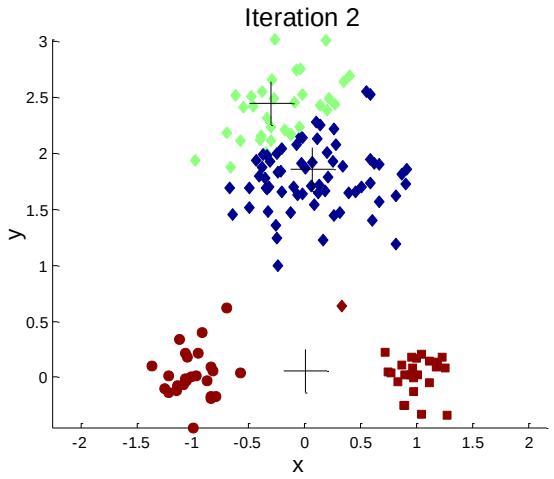


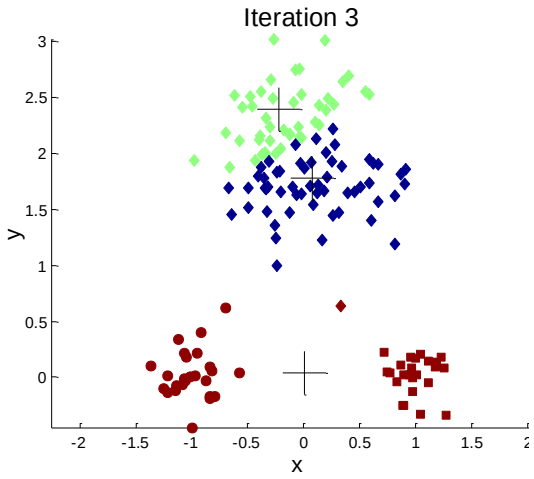
Importance of Choosing Initial Centroids ...

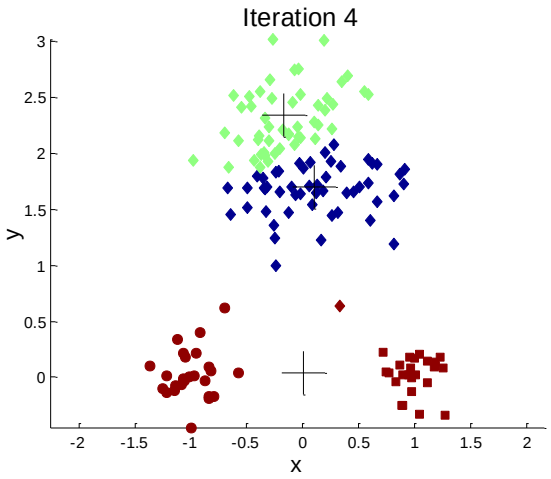


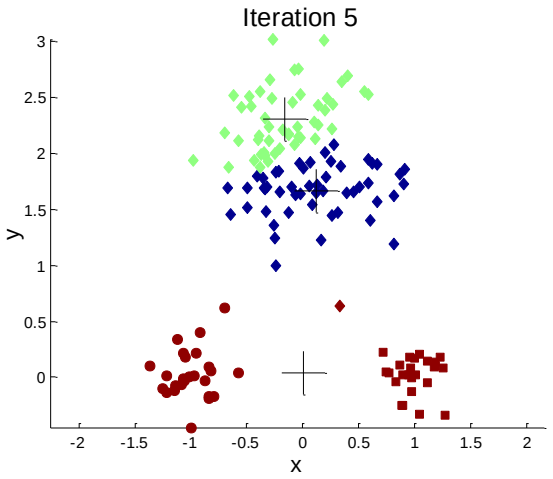
Importance of Choosing Initial Centroids ...









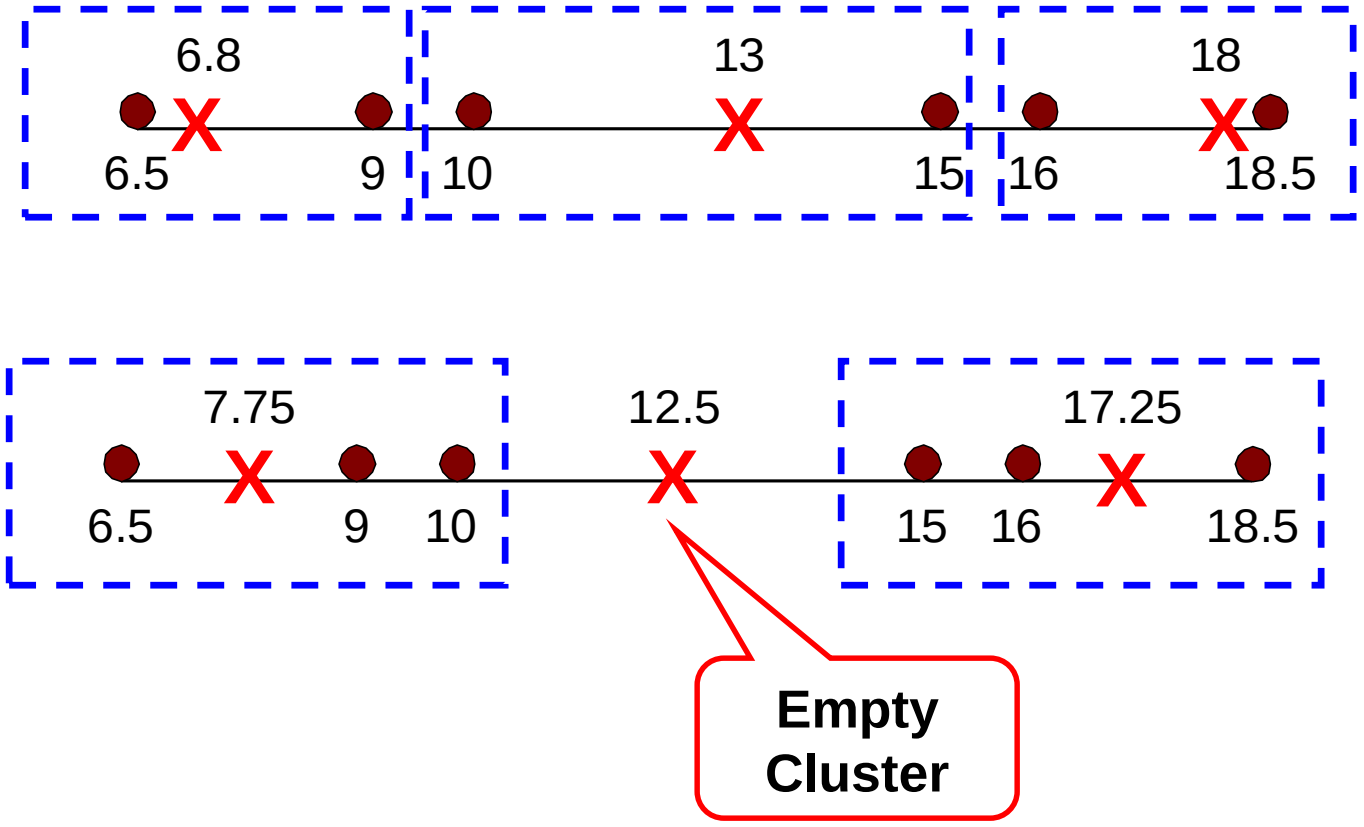


Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- 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
- Generate a larger number of clusters and then perform a hierarchical clustering
- Bisecting K-means
  - Not as susceptible to initialization issues

Empty Clusters

K-means can yield empty clusters



Handling Empty Clusters

Basic K-means algorithm can yield empty clusters

Several strategies

- Choose the point that contributes most to SSE
- Choose a point from the cluster with the highest SSE
- If there are several empty clusters, the above can be repeated several times.

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

Bisecting K-means

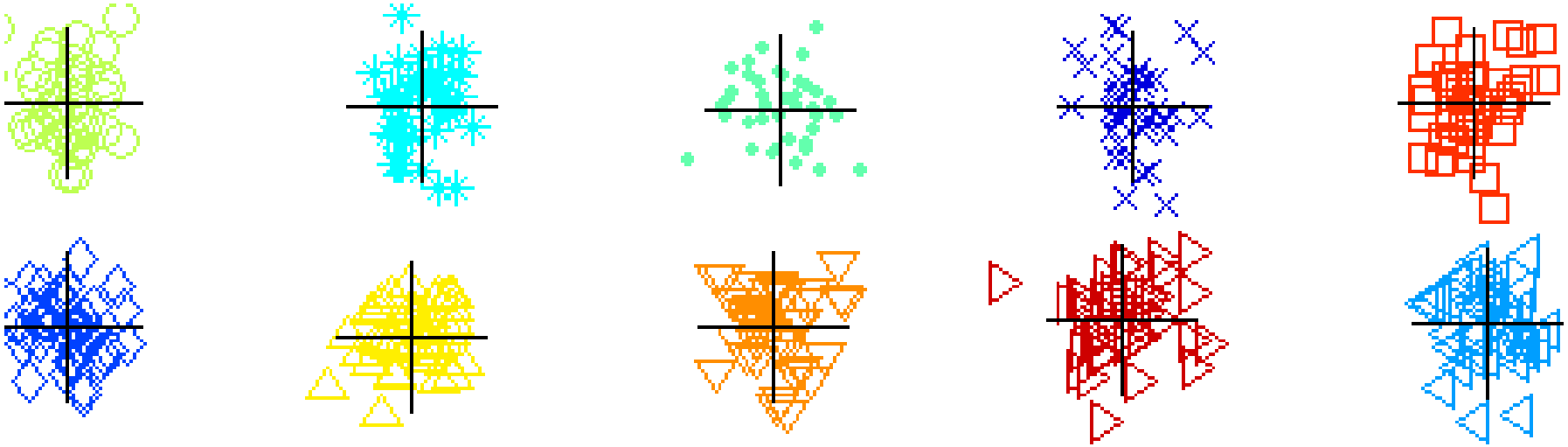
Bisecting K-means algorithm

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

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

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

Bisecting K-means Example

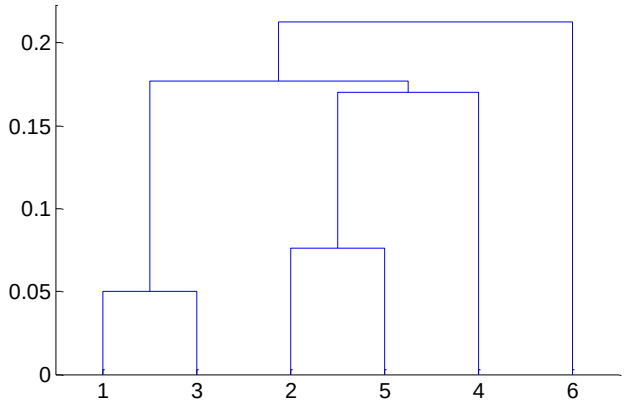


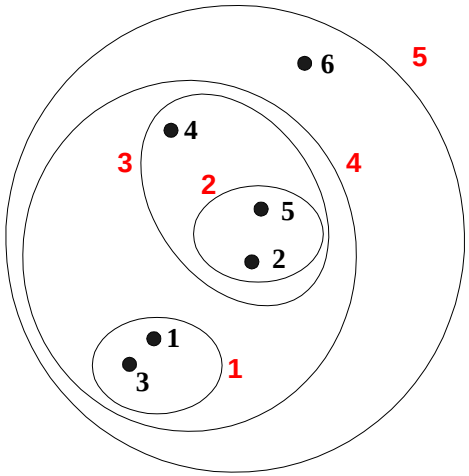
Hierarchical Clustering

Produces a set of nested clusters organized as a hierarchical tree

Can be visualized as a dendrogram

- A tree like diagram that records the sequences of merges or splits





Strengths of Hierarchical Clustering

Do not have to assume any particular number of clusters

- Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level

They may correspond to meaningful taxonomies

- Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Hierarchical Clustering

Two main types of hierarchical clustering

- Agglomerative:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive:
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains an individual point (or there are k clusters)

Traditional hierarchical algorithms use a similarity or distance matrix

- Merge or split one cluster at a time

Agglomerative Clustering Algorithm

Most popular hierarchical clustering technique

Basic algorithm is straightforward

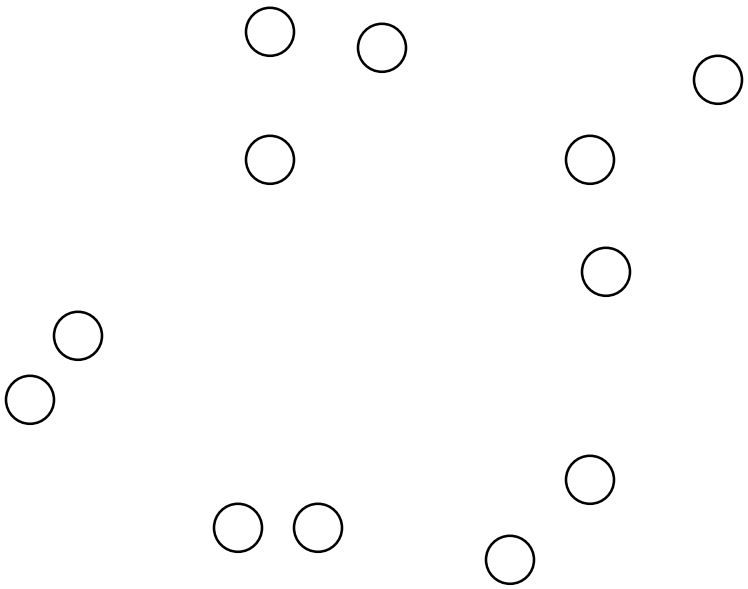
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

Starting Situation

Start with clusters of individual points and a proximity matrix

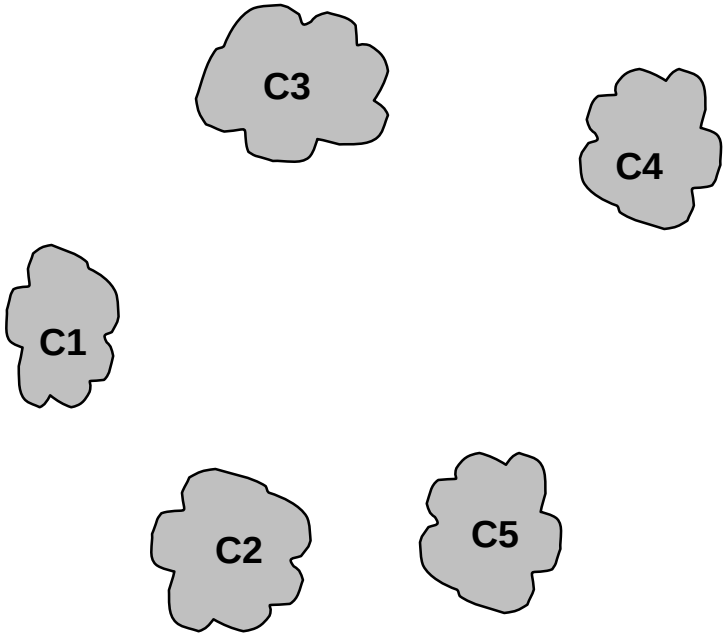


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

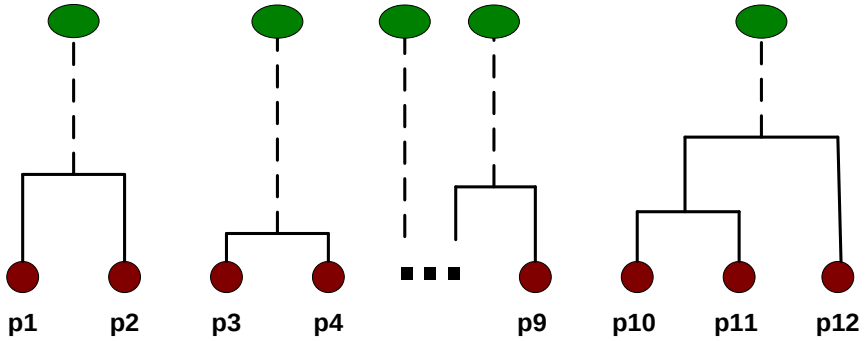
Intermediate Situation

After some merging steps, we have some clusters



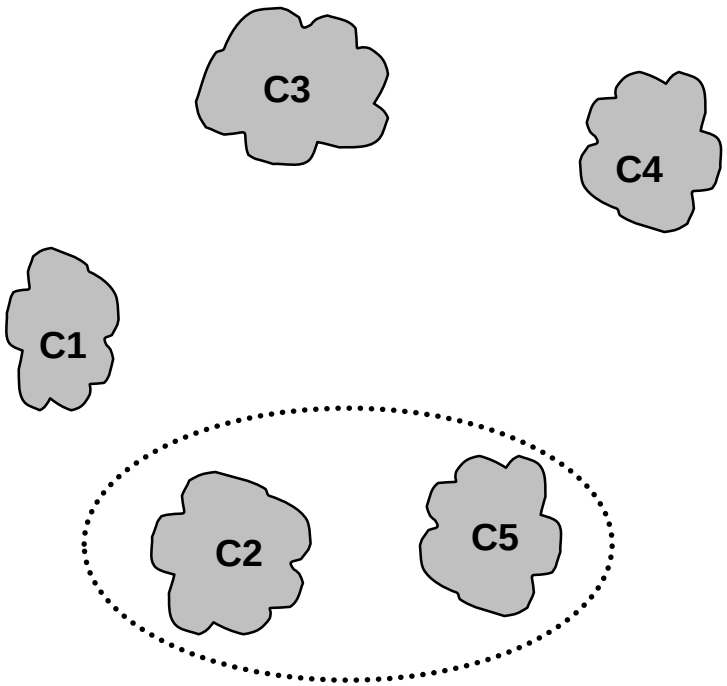
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



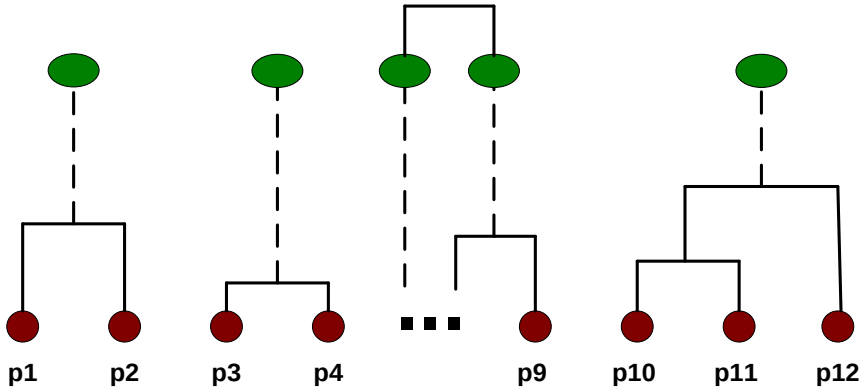
Intermediate Situation

We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



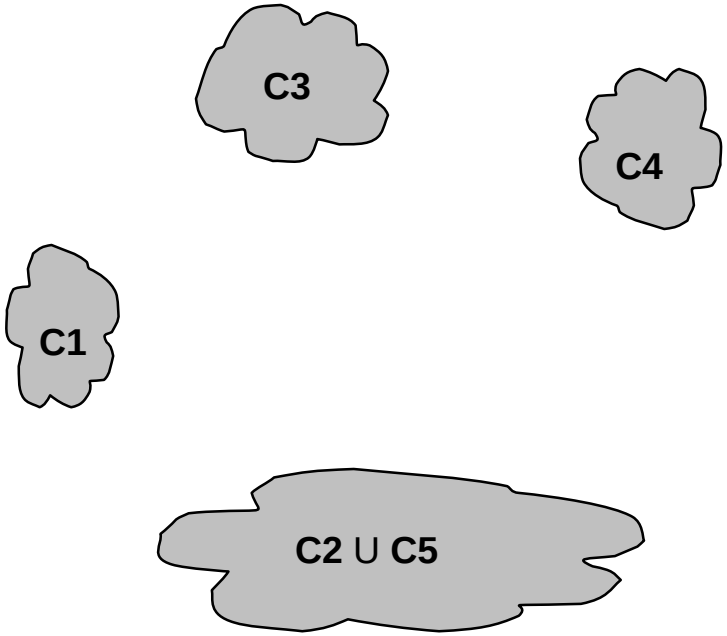
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



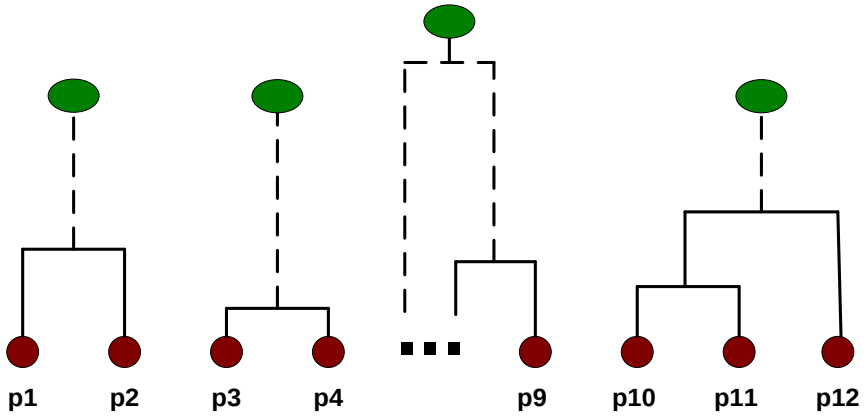
After Merging

The question is “How do we update the proximity matrix?”

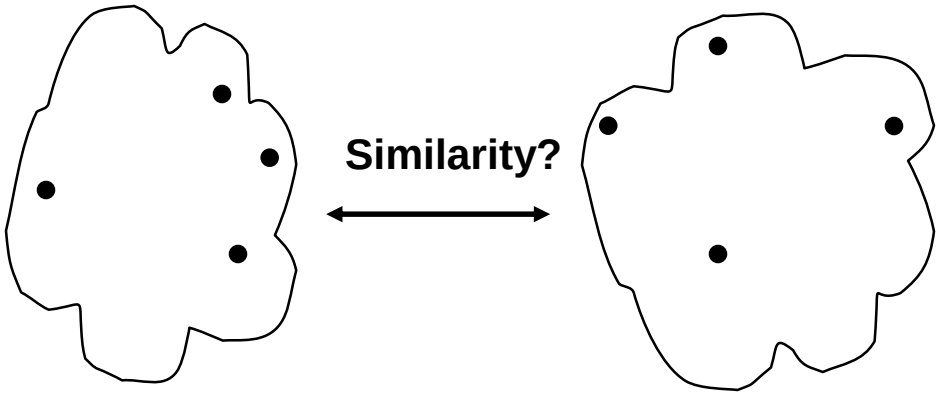


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Distance



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

MIN

MAX

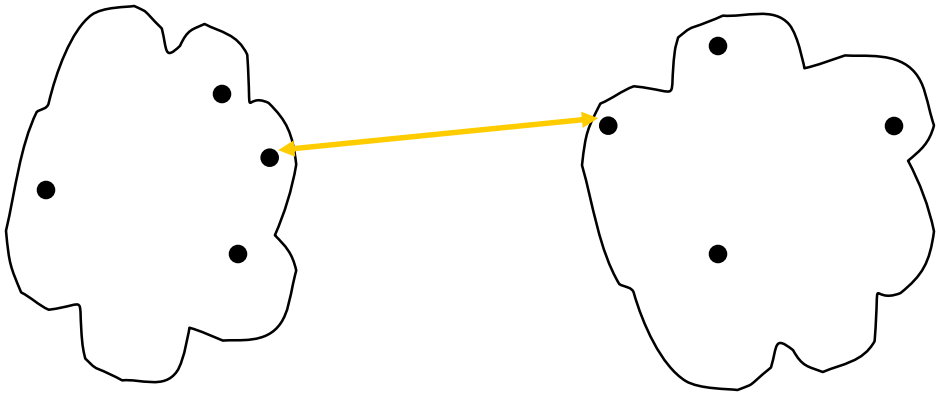
Group Average

Distance Between Centroids

Other methods driven by an objective function

- Ward's Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

MIN

MAX

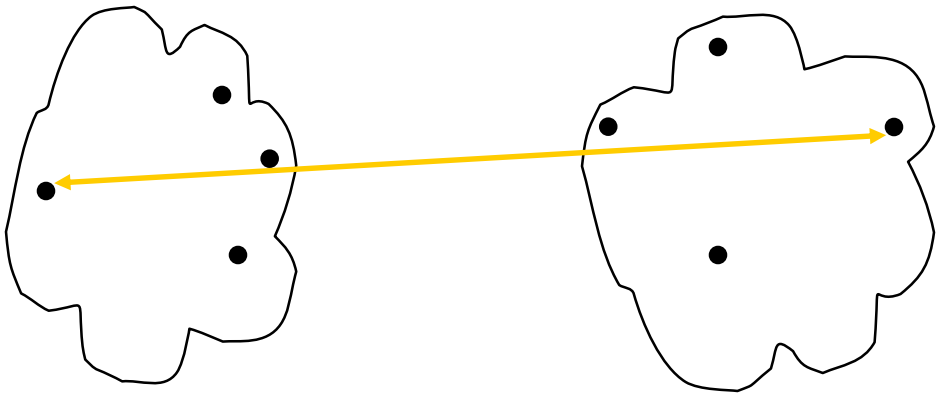
Group Average

Distance Between Centroids

Other methods driven by an objective function

- Ward's Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

MIN

MAX

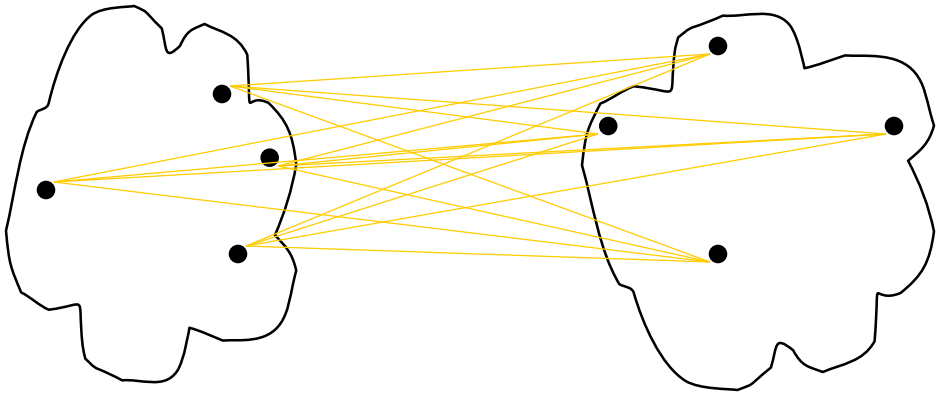
Group Average

Distance Between Centroids

Other methods driven by an objective function

- Ward's Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

MIN

MAX

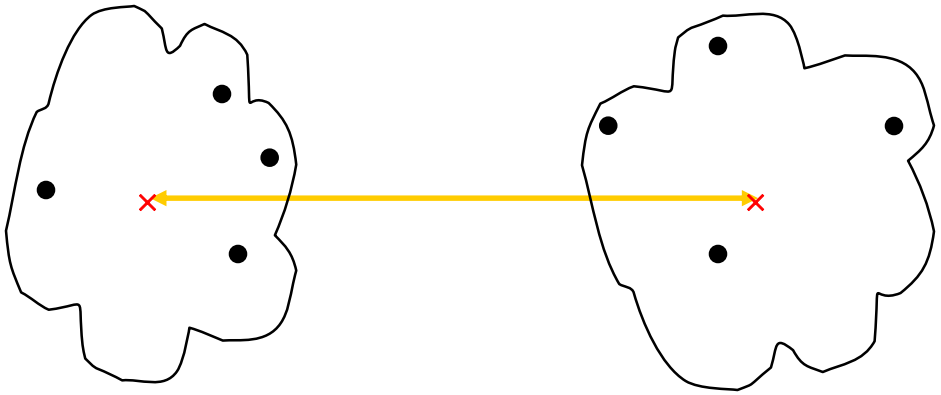
Group Average

Distance Between Centroids

Other methods driven by an objective function

- Ward's Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

MIN

MAX

Group Average

Distance Between Centroids

Other methods driven by an objective function

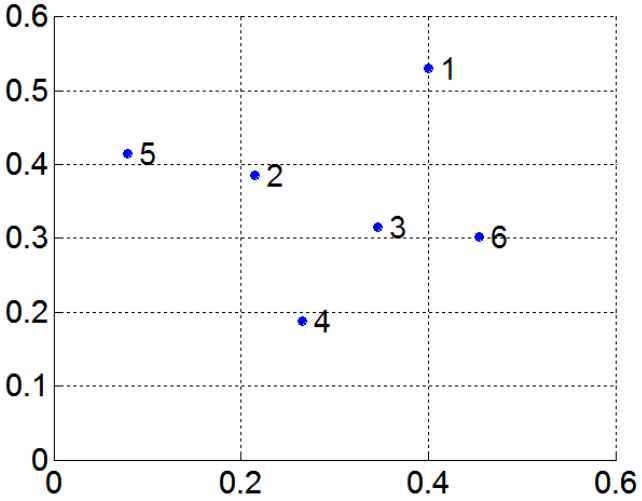
- Ward's Method uses squared error

MIN or Single Link

Proximity of two clusters is based on the two closest points in the different clusters

- Determined by one pair of points, i.e., by one link in the proximity graph

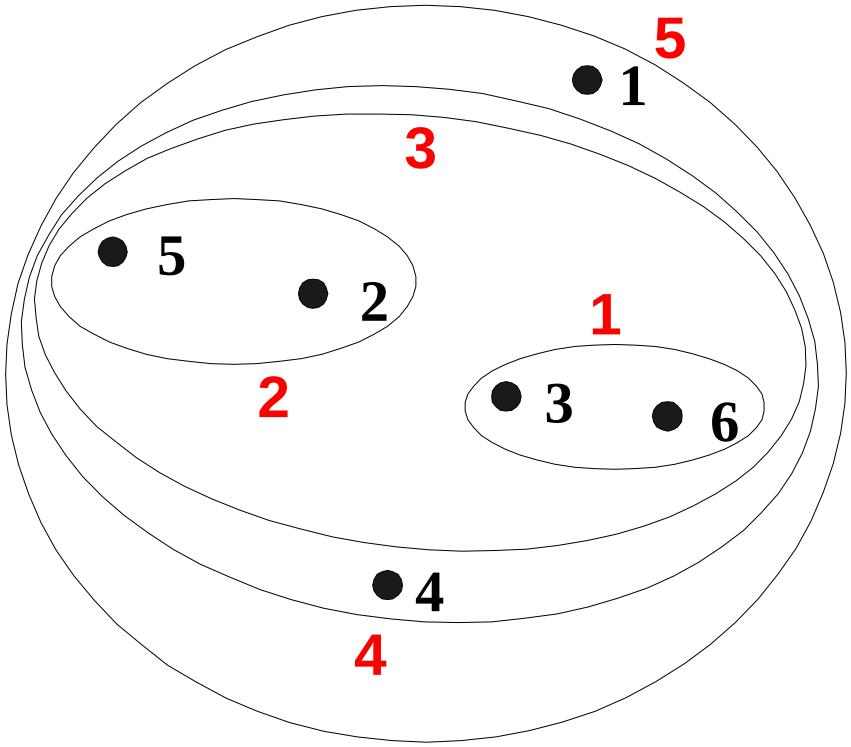
Example:



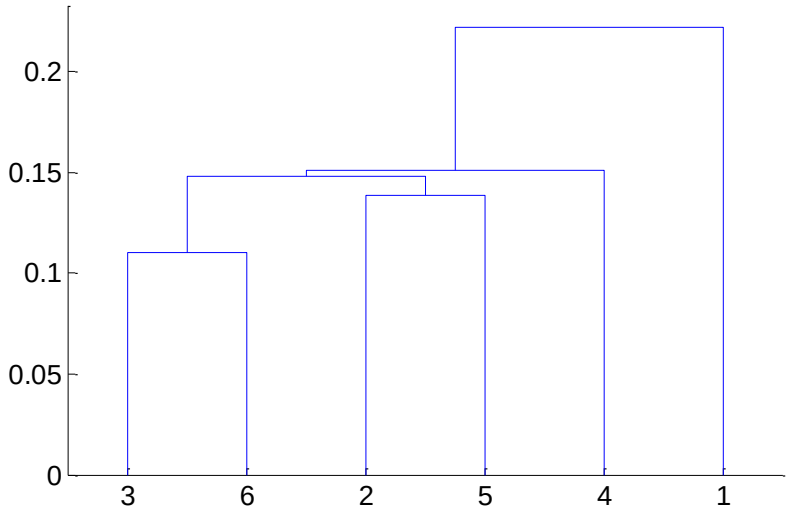
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: MIN

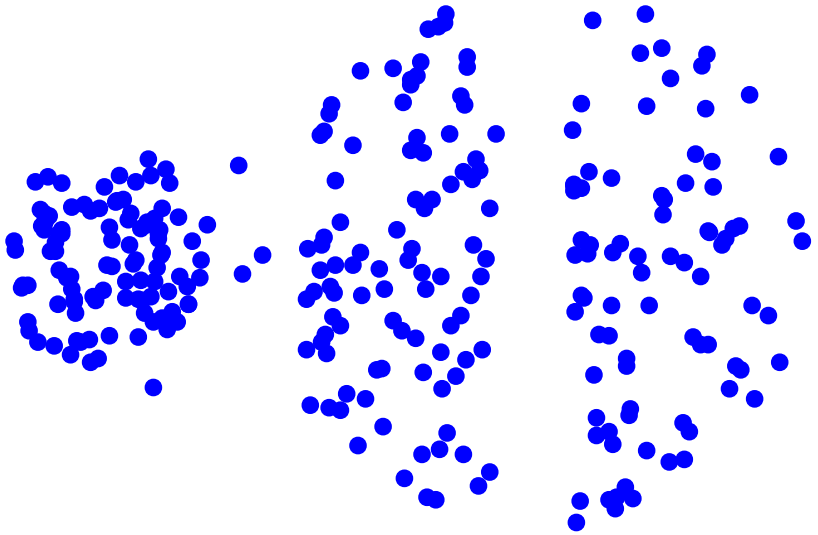


Nested Clusters

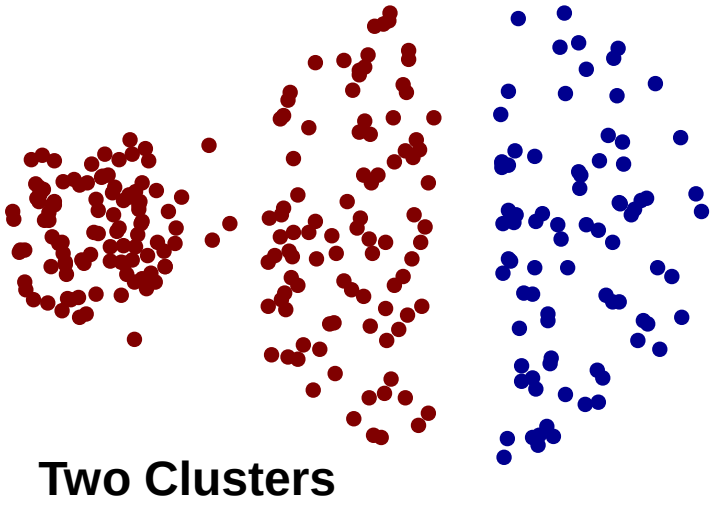


Dendrogram

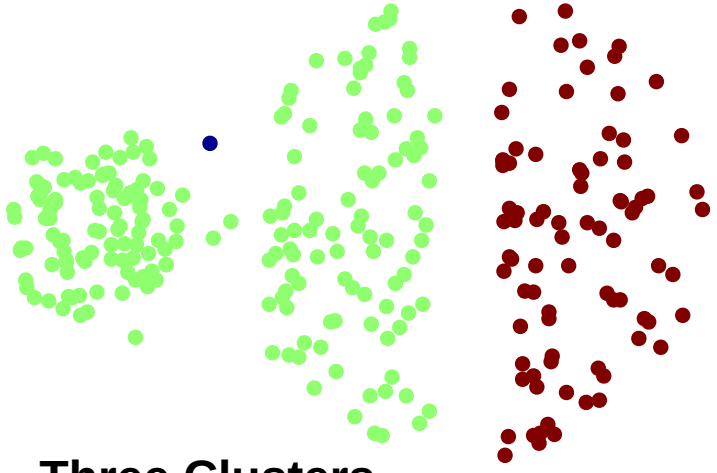
Limitations of MIN



Original Points



Two Clusters



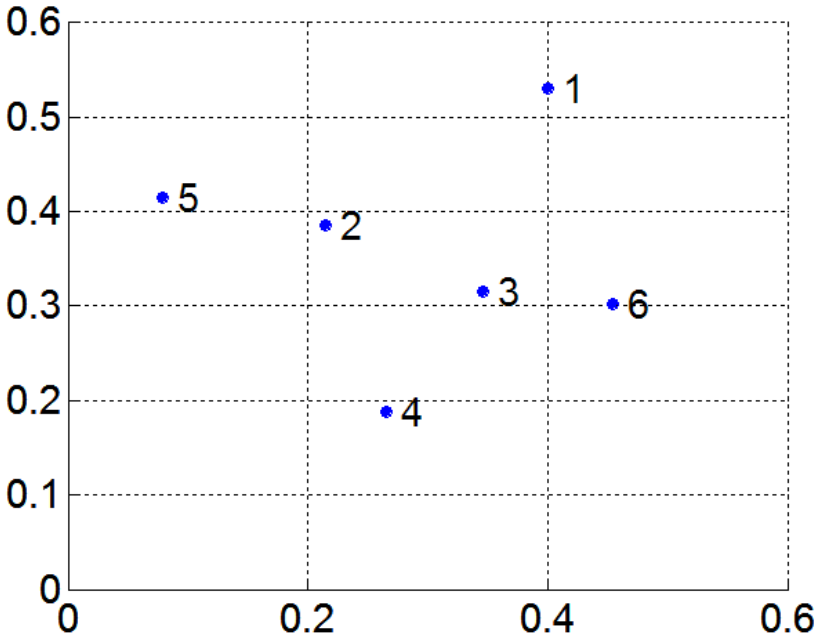
Three Clusters

- Sensitive to noise and outliers

MAX or Complete Linkage

Proximity of two clusters is based on the two most distant points in the different clusters

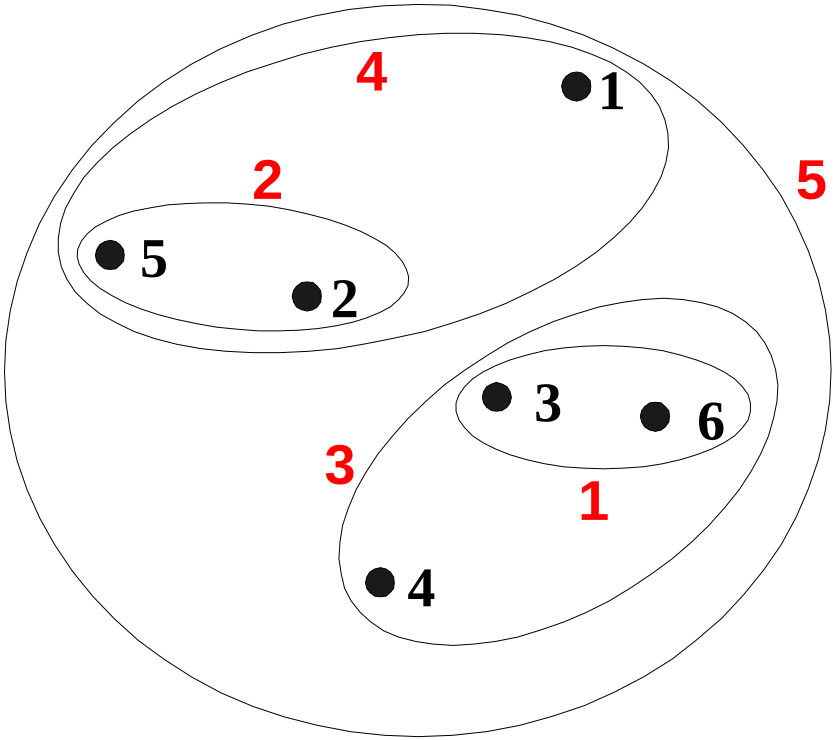
- Determined by all pairs of points in the two clusters



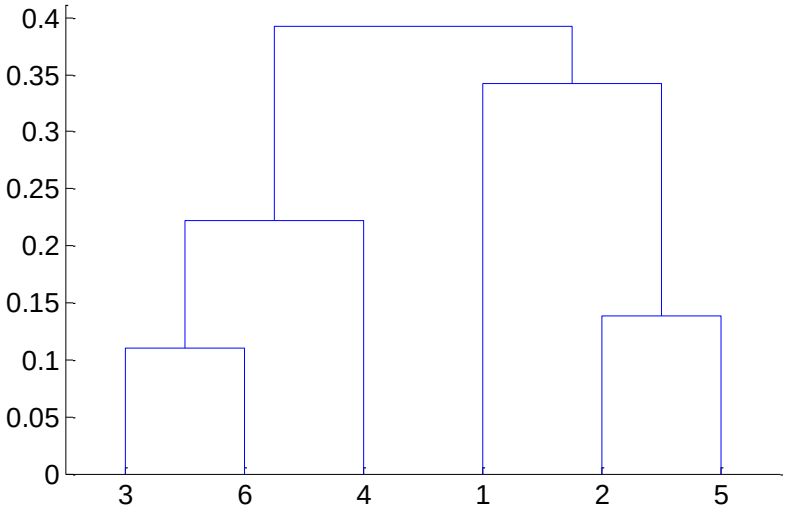
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: MAX

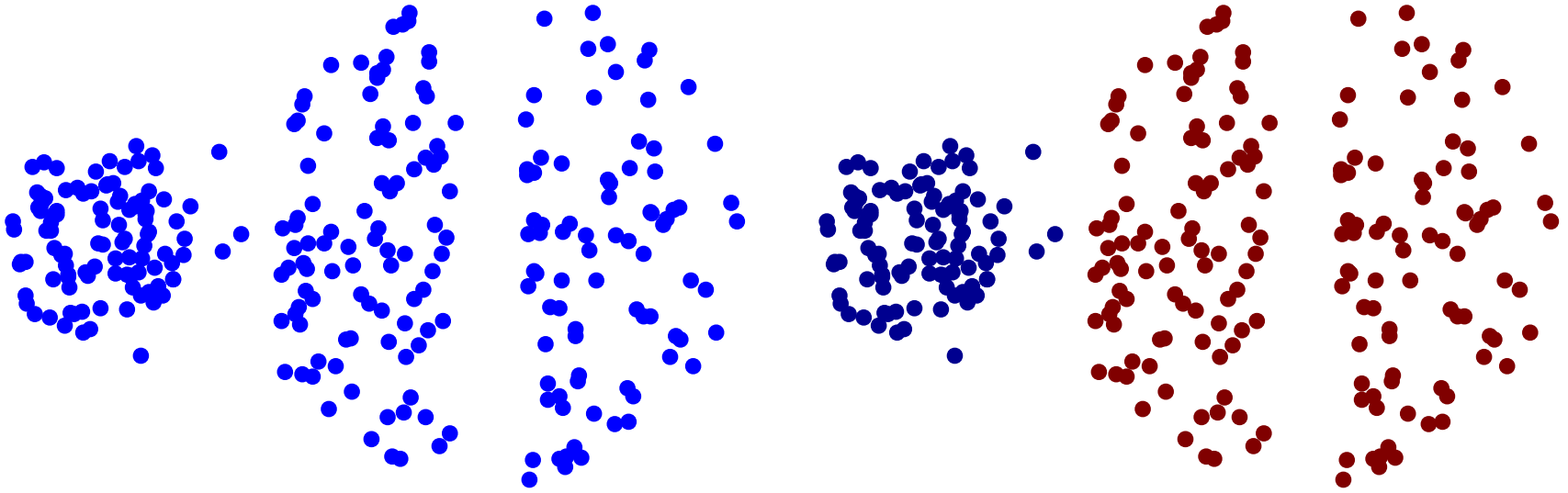


Nested Clusters



Dendrogram

Strength of MAX

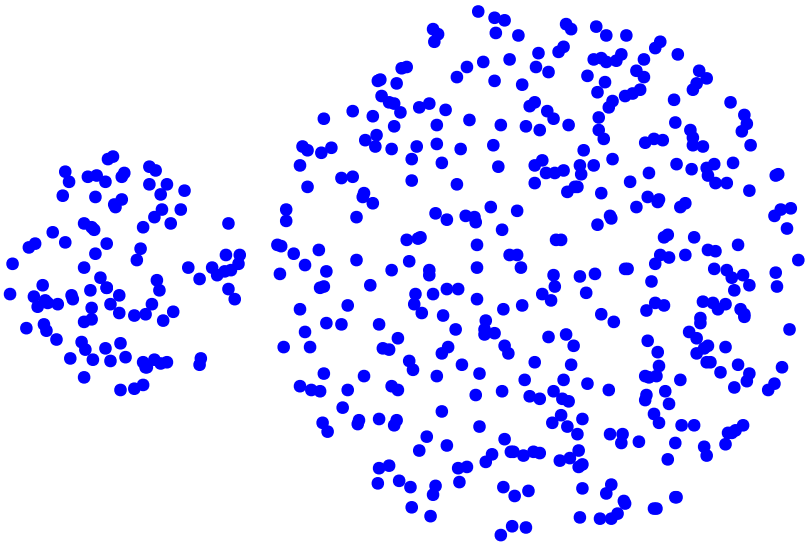


Original Points

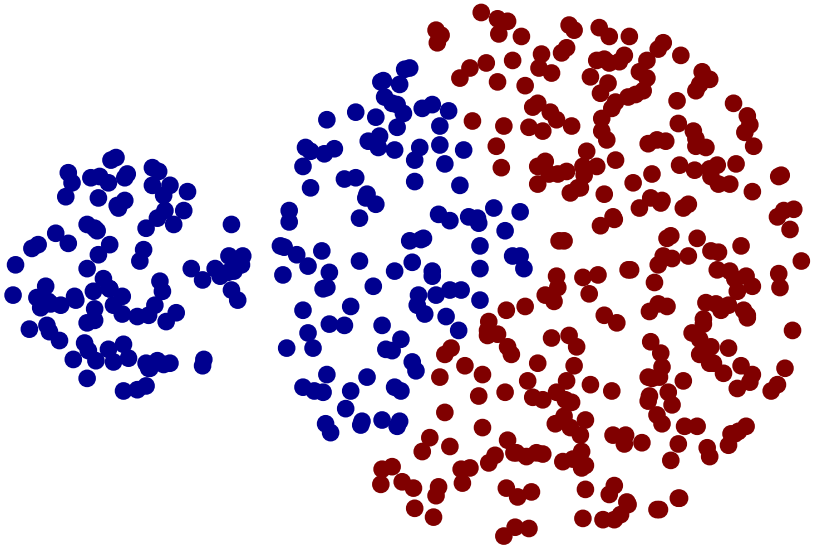
Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX



Original Points



Two Clusters

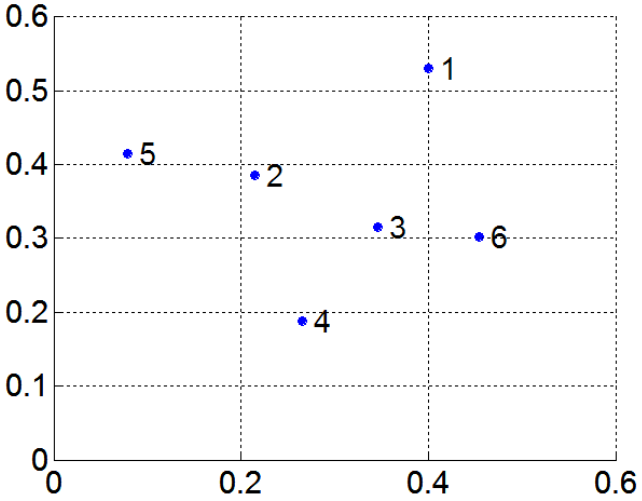
- Tends to break large clusters
- Biased towards globular clusters

Group Average

Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

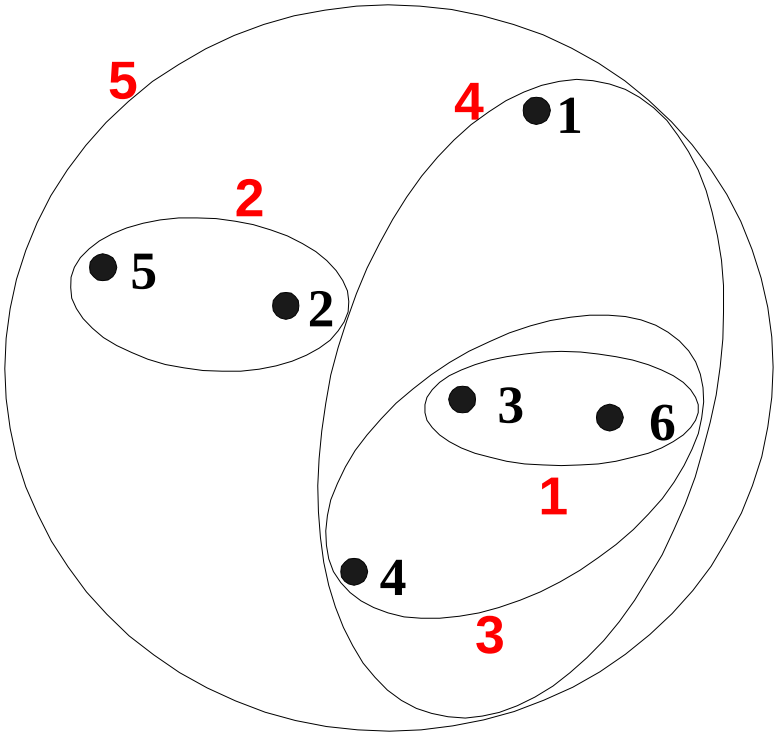
Need to use average connectivity for scalability since total proximity favors large clusters



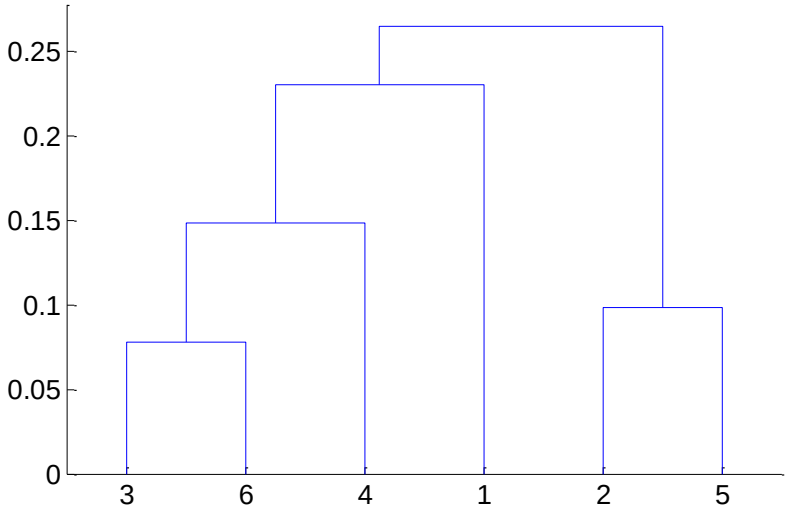
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: Group Average



Nested Clusters



Dendrogram

Hierarchical Clustering: Group Average

Compromise between Single and Complete Link

Strengths

- Less susceptible to noise and outliers

Limitations

- Biased towards globular clusters

Cluster Similarity: Ward's Method

Similarity of two clusters is based on the increase in squared error when two clusters are merged

- Similar to group average if distance between points is distance squared

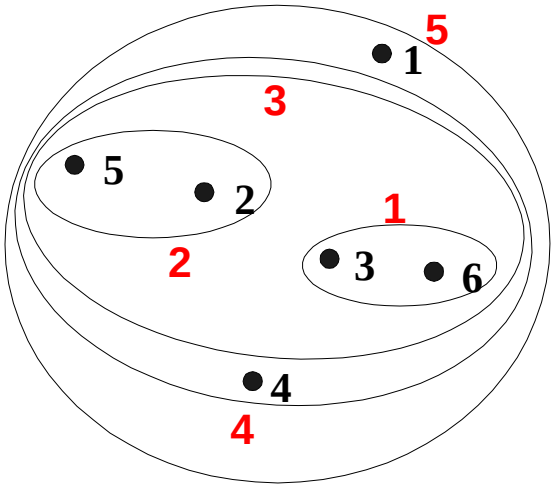
Less susceptible to noise and outliers

Biased towards globular clusters

Hierarchical analogue of K-means

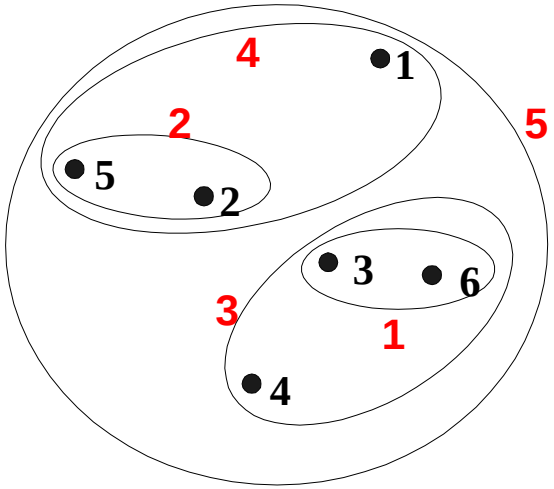
- Can be used to initialize K-means

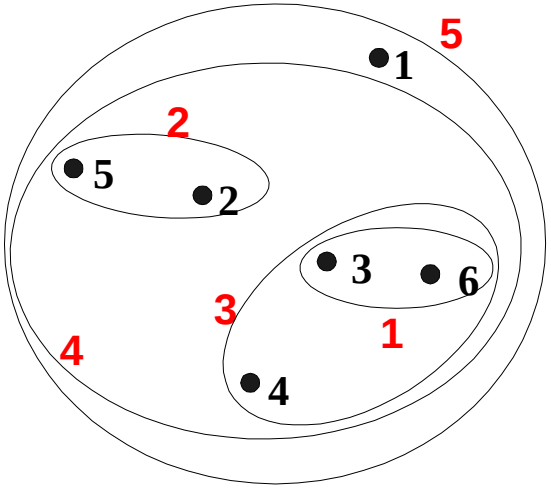
Hierarchical Clustering: Comparison



MIN

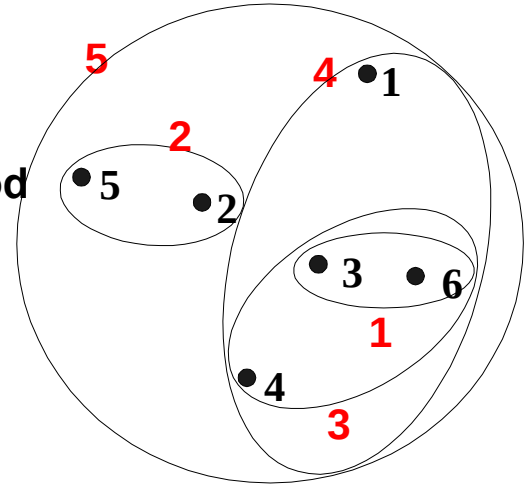
MAX





Group Average

Ward's Method



Hierarchical Clustering: Time and Space requirements

$O(N^2)$ space since it uses the proximity matrix.

- N is the number of points.

$O(N^3)$ time in many cases

- There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
- Complexity can be reduced to $O(N^2 \log(N))$ time with some cleverness

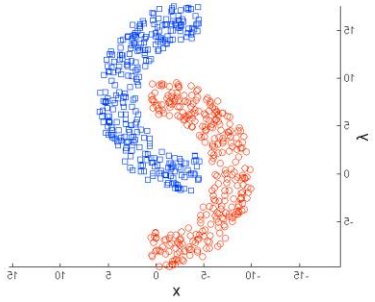
Hierarchical Clustering: Problems and Limitations

Once a decision is made to combine two clusters, it cannot be undone

No global objective function is directly minimized

Different schemes have problems with one or more of the following:

- Sensitivity to noise and outliers
- Difficulty handling clusters of different sizes and non-globular shapes
- Breaking large clusters



Cluster Validity

For supervised classification we have a variety of measures to evaluate how good our model is

- Accuracy, precision, recall

For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?

But “clusters are in the eye of the beholder”!

Then why do we want to evaluate them?

- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two sets of clusters
- To compare two clusters

Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.

   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Measures of Cluster Validity

Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.

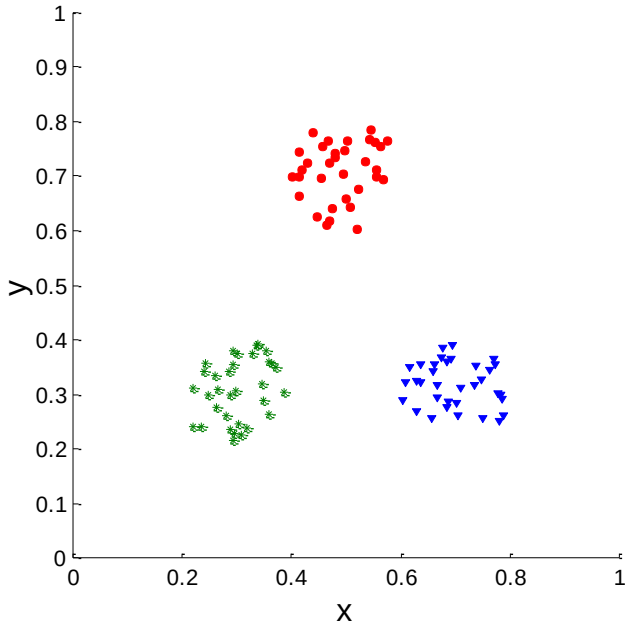
- External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
  - Entropy
- Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
  - Sum of Squared Error (SSE)
- Relative Index: Used to compare two different clusterings or clusters.
  - Often an external or internal index is used for this function, e.g., SSE or entropy

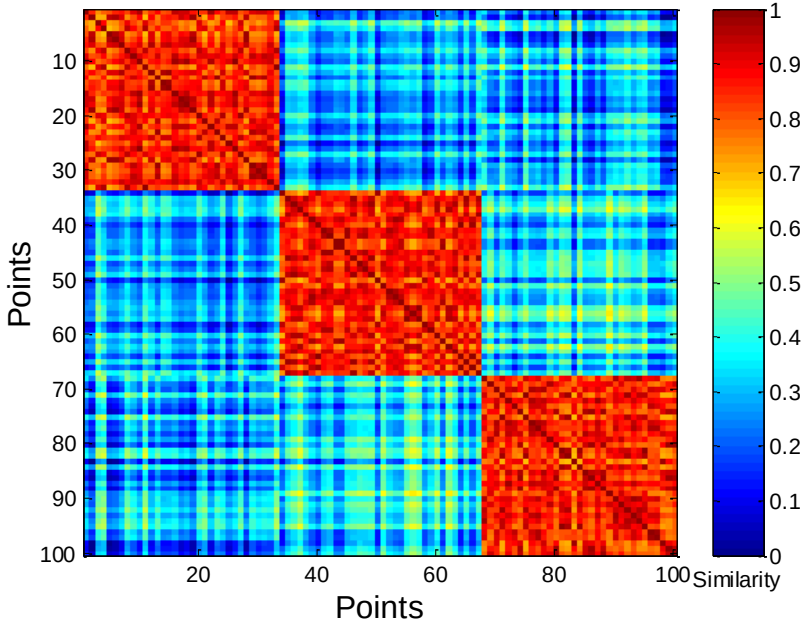
Sometimes these are referred to as criteria instead of indices

- However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Using Similarity Matrix for Cluster Validation

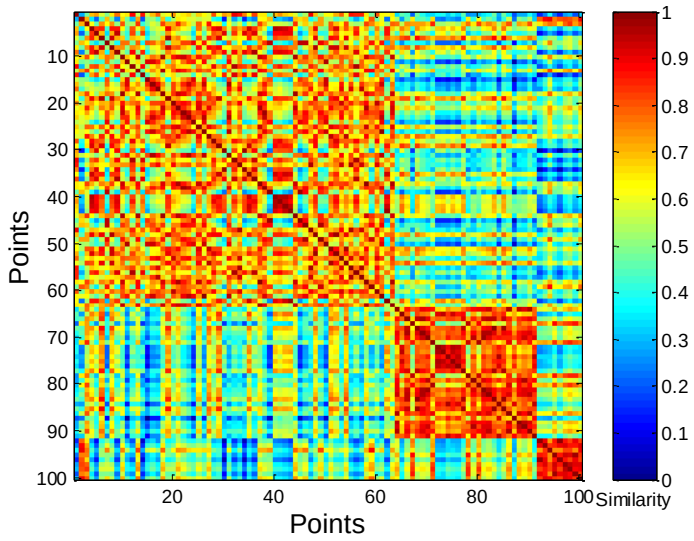
Order the similarity matrix with respect to cluster labels and inspect visually.

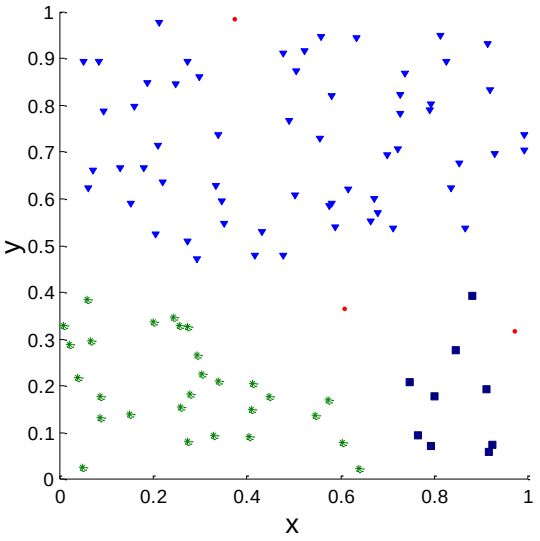




Using Similarity Matrix for Cluster Validation

Clusters in random data are not so crisp

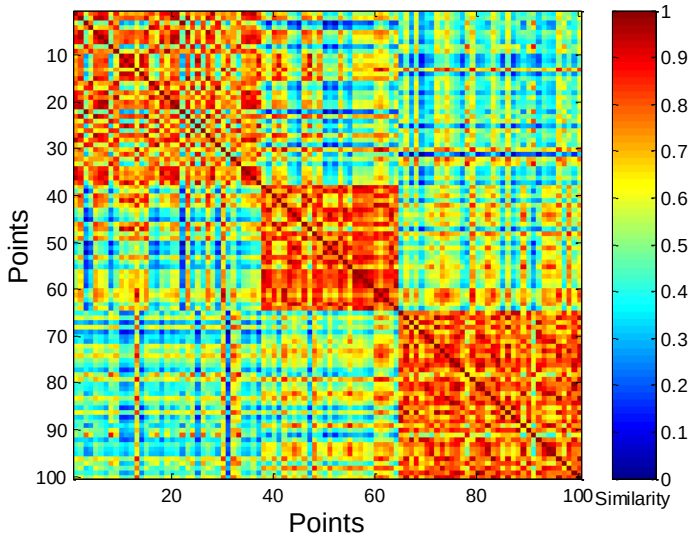


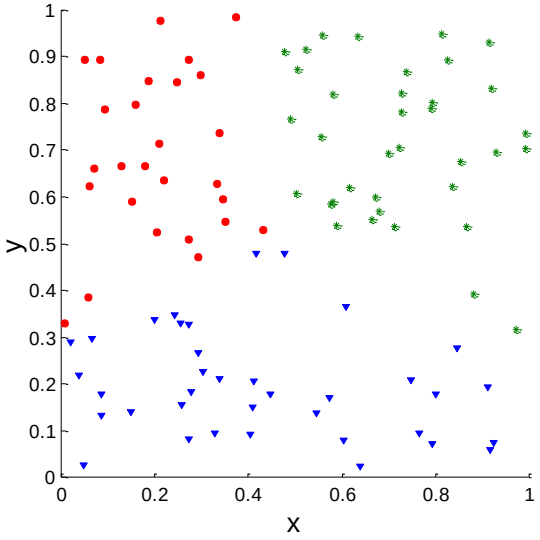


DBSCAN

Using Similarity Matrix for Cluster Validation

Clusters in random data are not so crisp

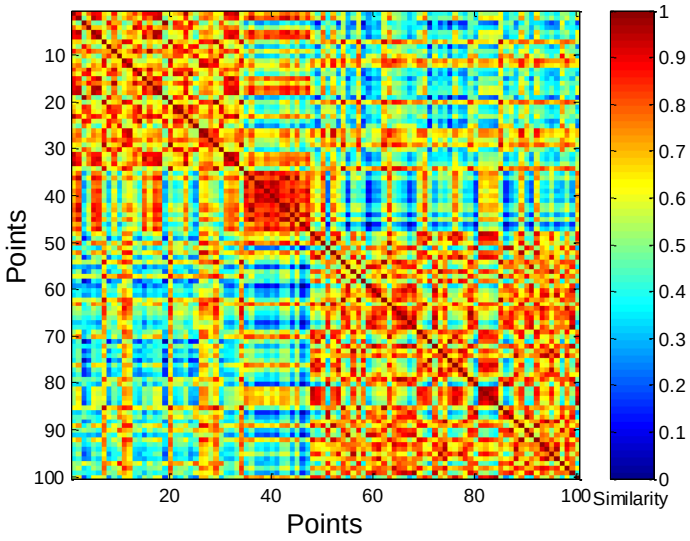


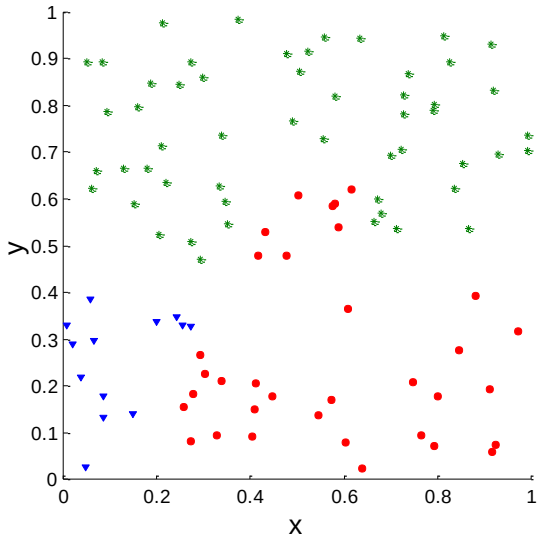


K-means

Using Similarity Matrix for Cluster Validation

Clusters in random data are not so crisp





Complete Link

Internal Measures: SSE

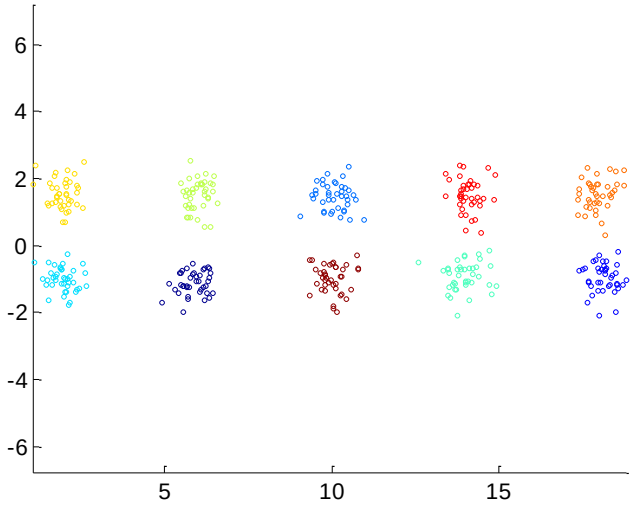
Clusters in more complicated figures aren't well separated

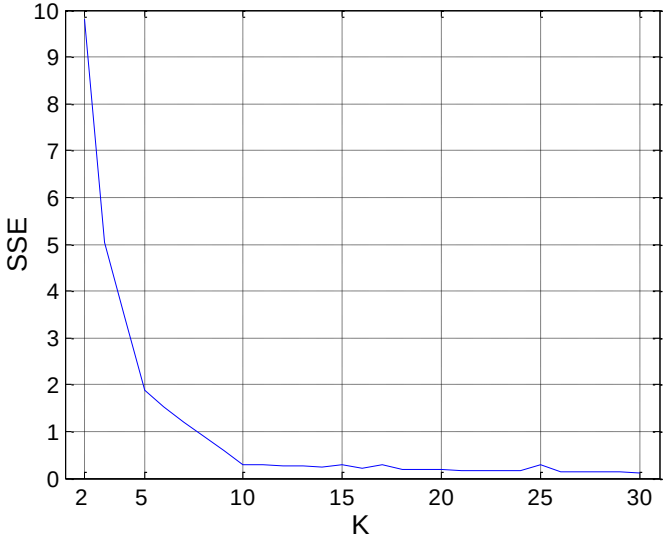
Internal Index: Used to measure the goodness of a clustering structure without respect to external information

- SSE

SSE is good for comparing two clusterings or two clusters (average SSE).

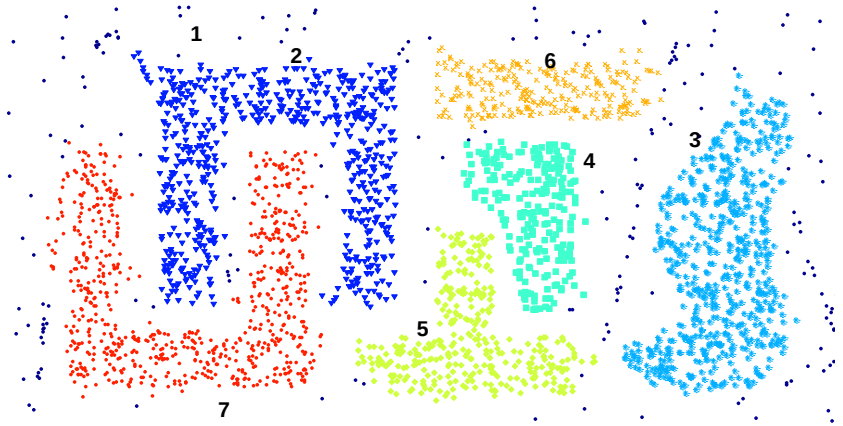
Can also be used to estimate the number of clusters

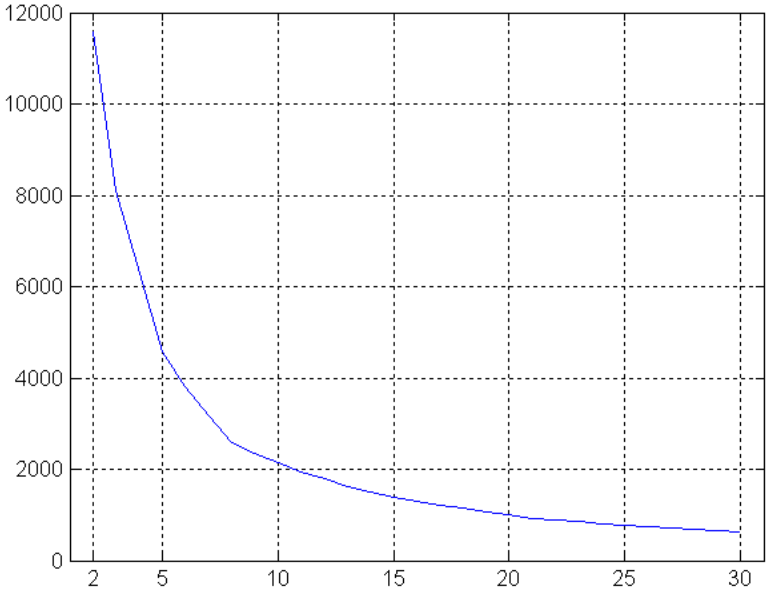




Internal Measures: SSE

SSE curve for a more complicated data set





SSE of clusters found using K-means

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes