

# Week 9

# Clustering

# Seokho Chi

Professor | Ph.D.

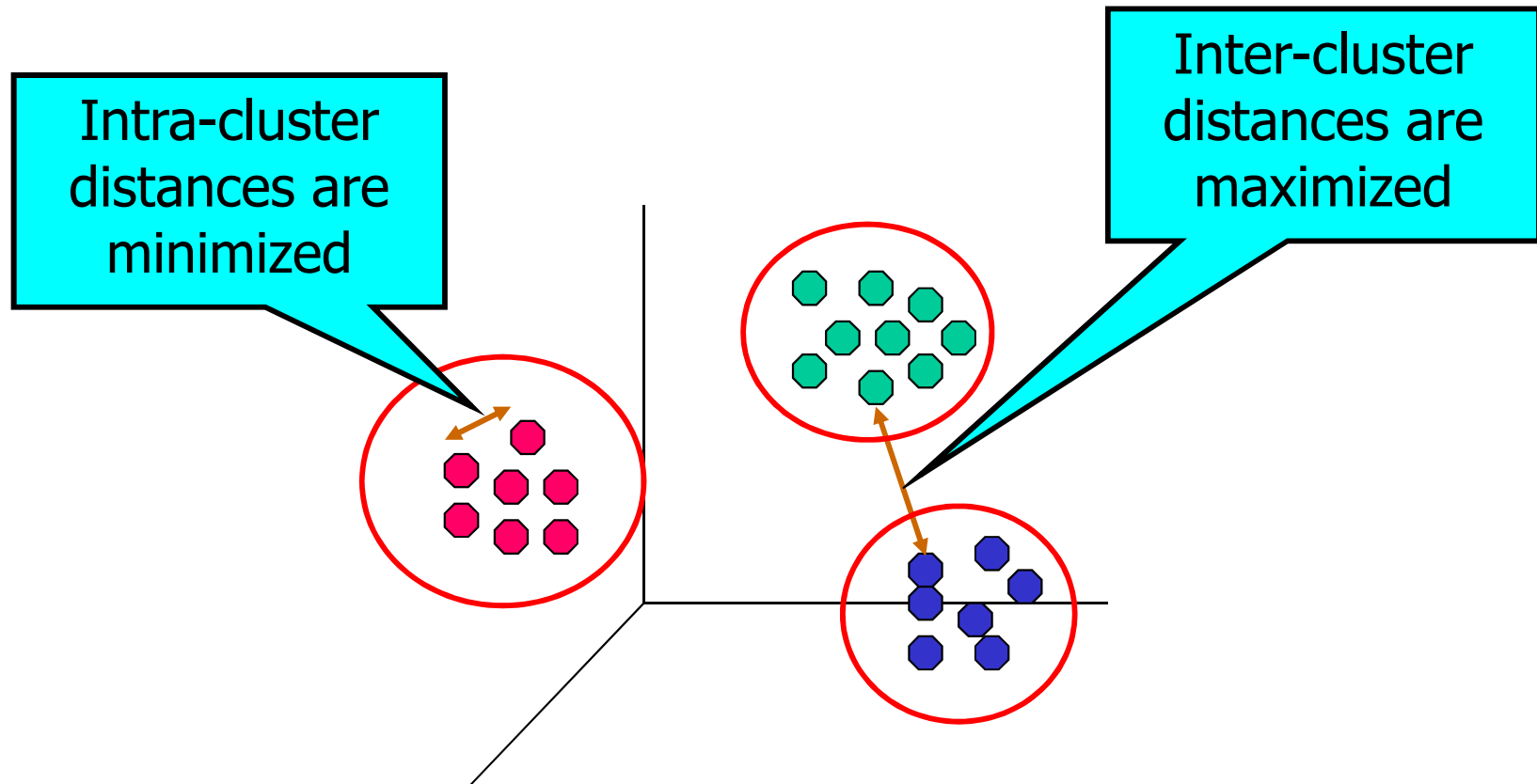
SNU Construction Innovation Lab

Source: Tan, Kumar, Steinback (2006)



# 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



*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Cluster Weblog data to discover groups of similar access patterns

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earthquake studies: Observed earthquake epicenters should be clustered along continent faults

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# What Is Good Clustering?

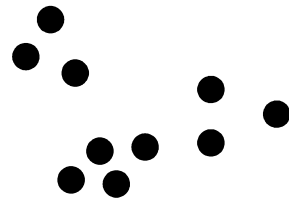
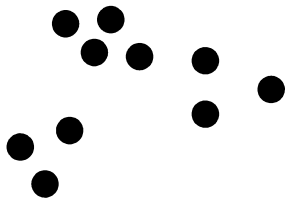
- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

*Adapted from:*

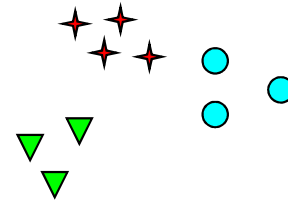
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

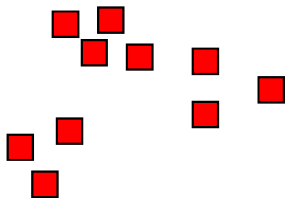
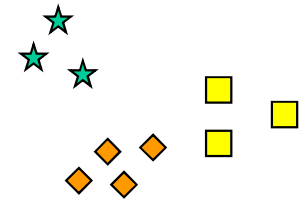
# Notion of a Cluster can be Ambiguous



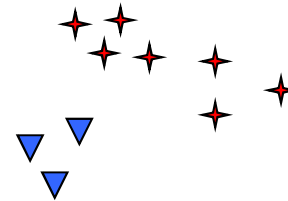
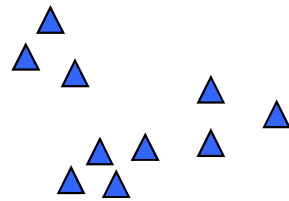
How many clusters?



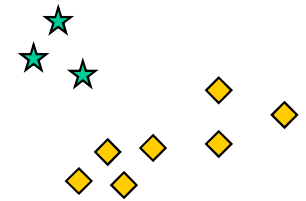
Six Clusters



Two Clusters



Four Clusters



*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Major Clustering Approaches

- Partitioning algorithms: A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchy algorithms: A set of nested clusters organized as a hierarchical tree
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# K-means Clustering

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

- 
- 1: Select  $K$  points as the initial centroids.
  - 2: **repeat**
  - 3:   Form  $K$  clusters by assigning all points to the closest centroid.
  - 4:   Recompute the centroid of each cluster.
  - 5: **until** The centroids don't change
- 

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# K-means Clustering – Details

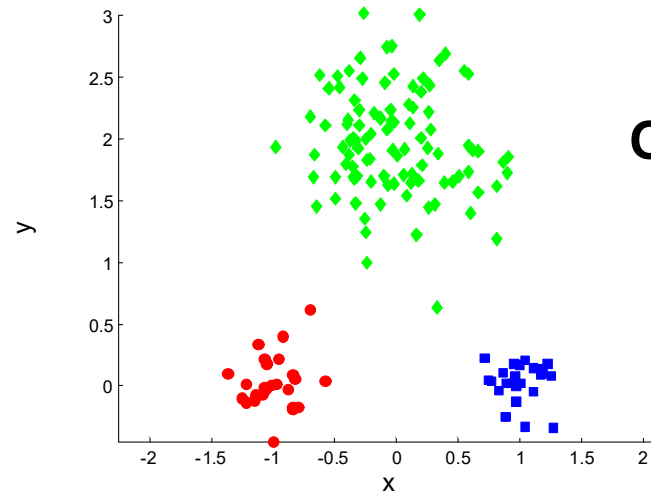
- Initial centroids are often chosen randomly.
  - Clusters produced may vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.

*Adapted from:*

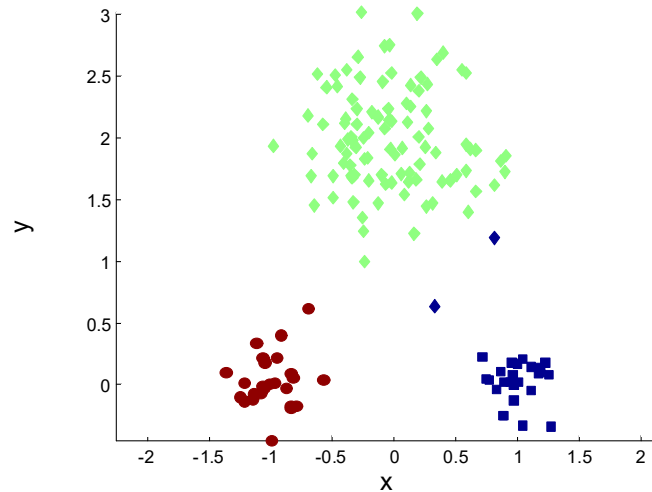
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

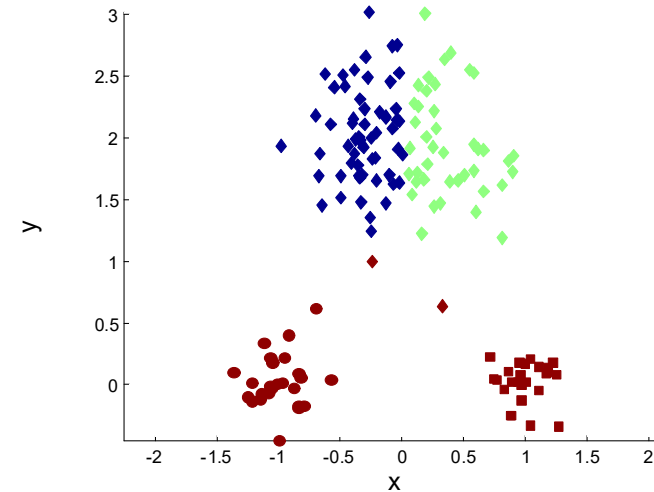
# Two different K-means Clusterings



**Original Points**



**Optimal Clustering**



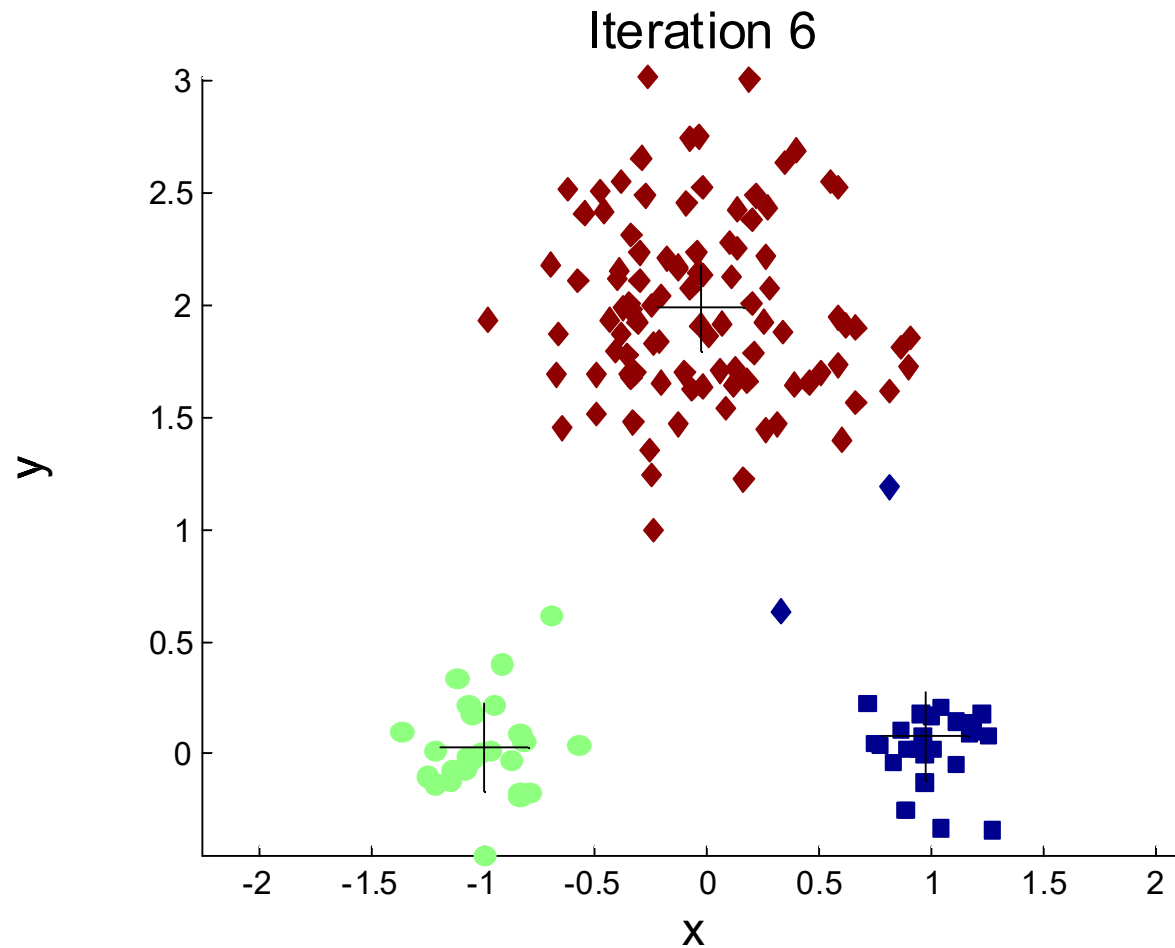
**Sub-optimal Clustering**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Importance of Choosing Initial Centroids (A)

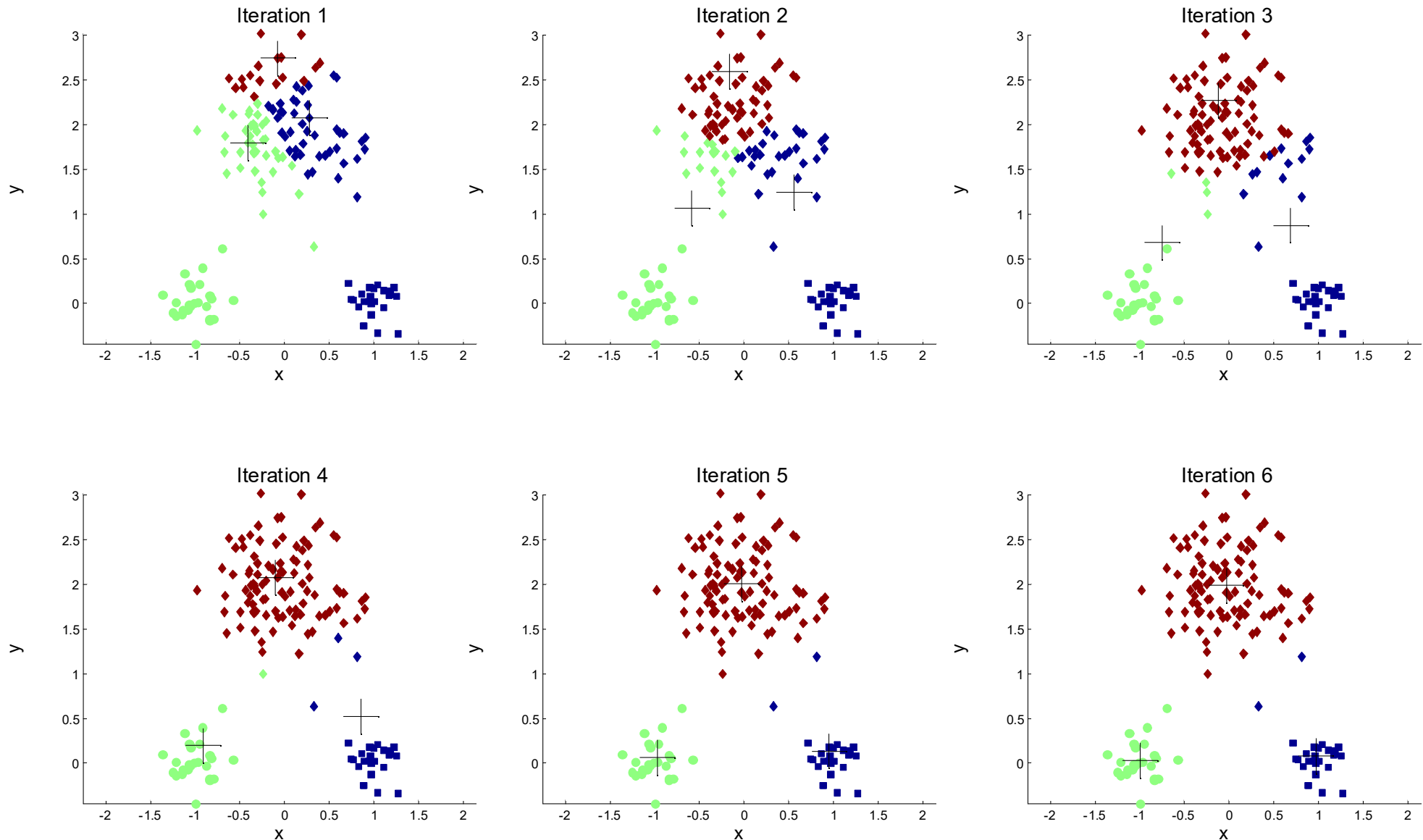


*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Importance of Choosing Initial Centroids (A)

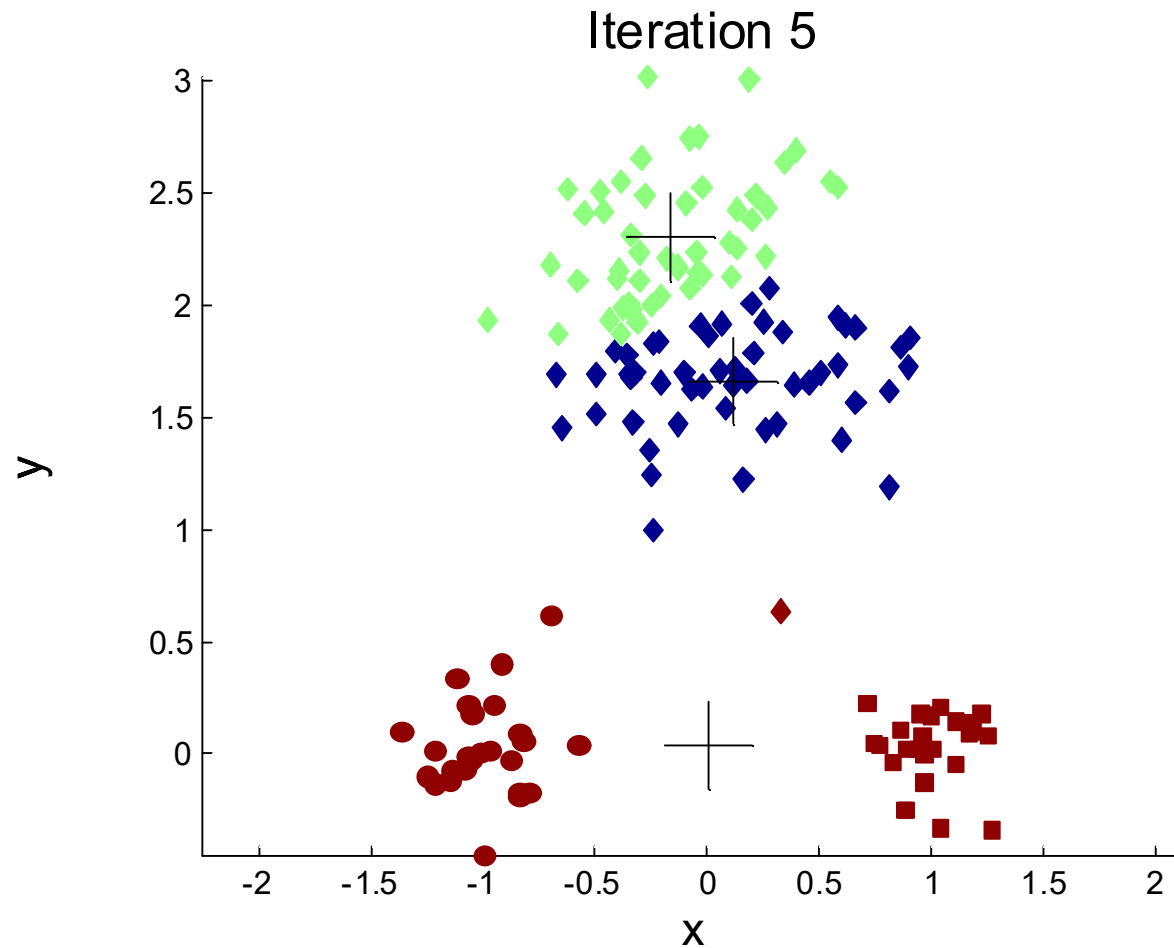


Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Importance of Choosing Initial Centroids (B)

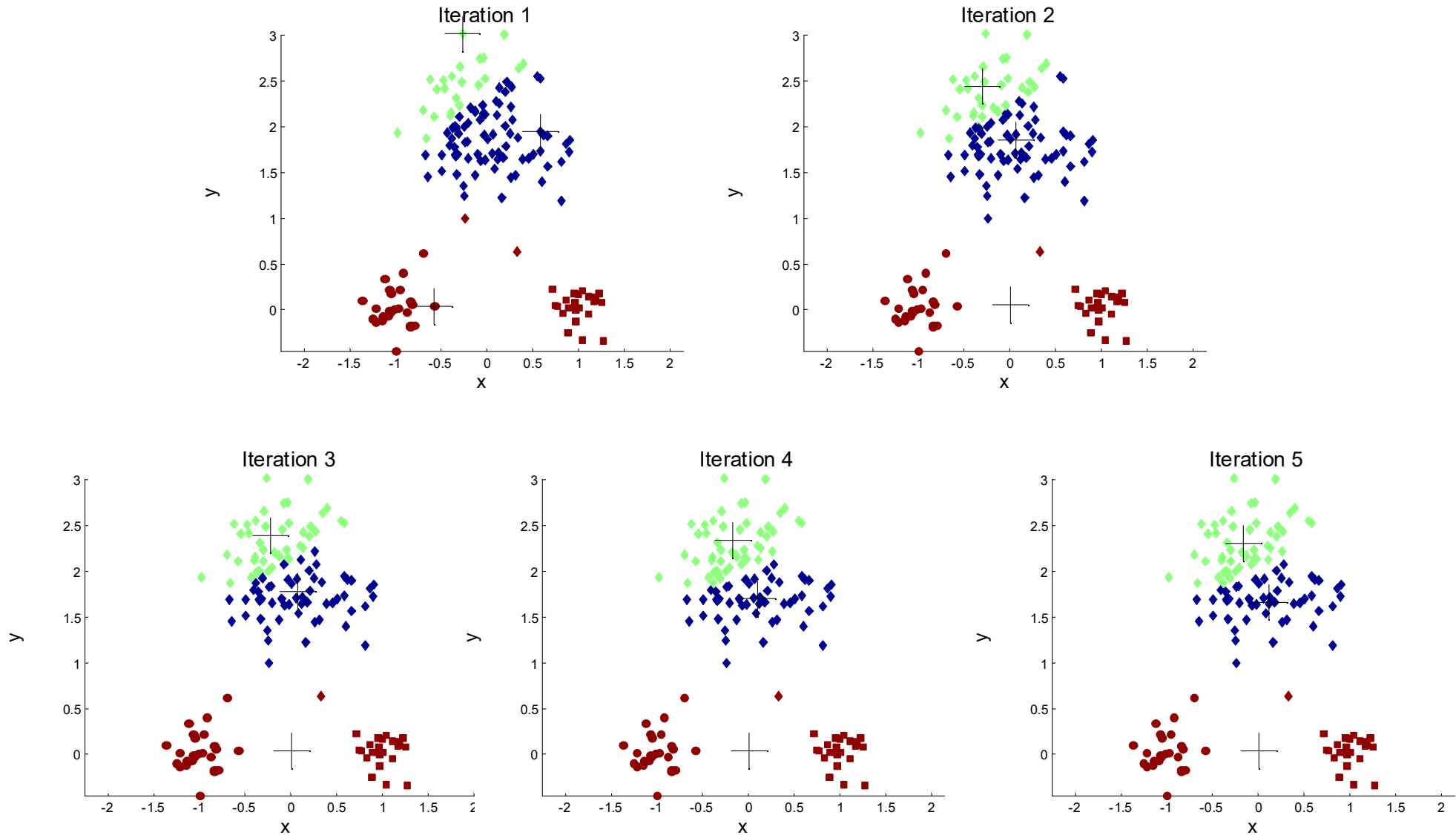


*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Importance of Choosing Initial Centroids (B)



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# 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 (center) point for cluster  $C_i$
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase  $K$ , the number of clusters

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Problems with Selecting Initial Points

- If there are  $K$  'real' clusters then the chance of selecting one centroid from each cluster is small.
  - Especially when  $K$  is large
  - Sometimes the initial centroids will readjust themselves in 'right' way, and sometimes they don't
  - Consider an example of five pairs of clusters

*Adapted from:*

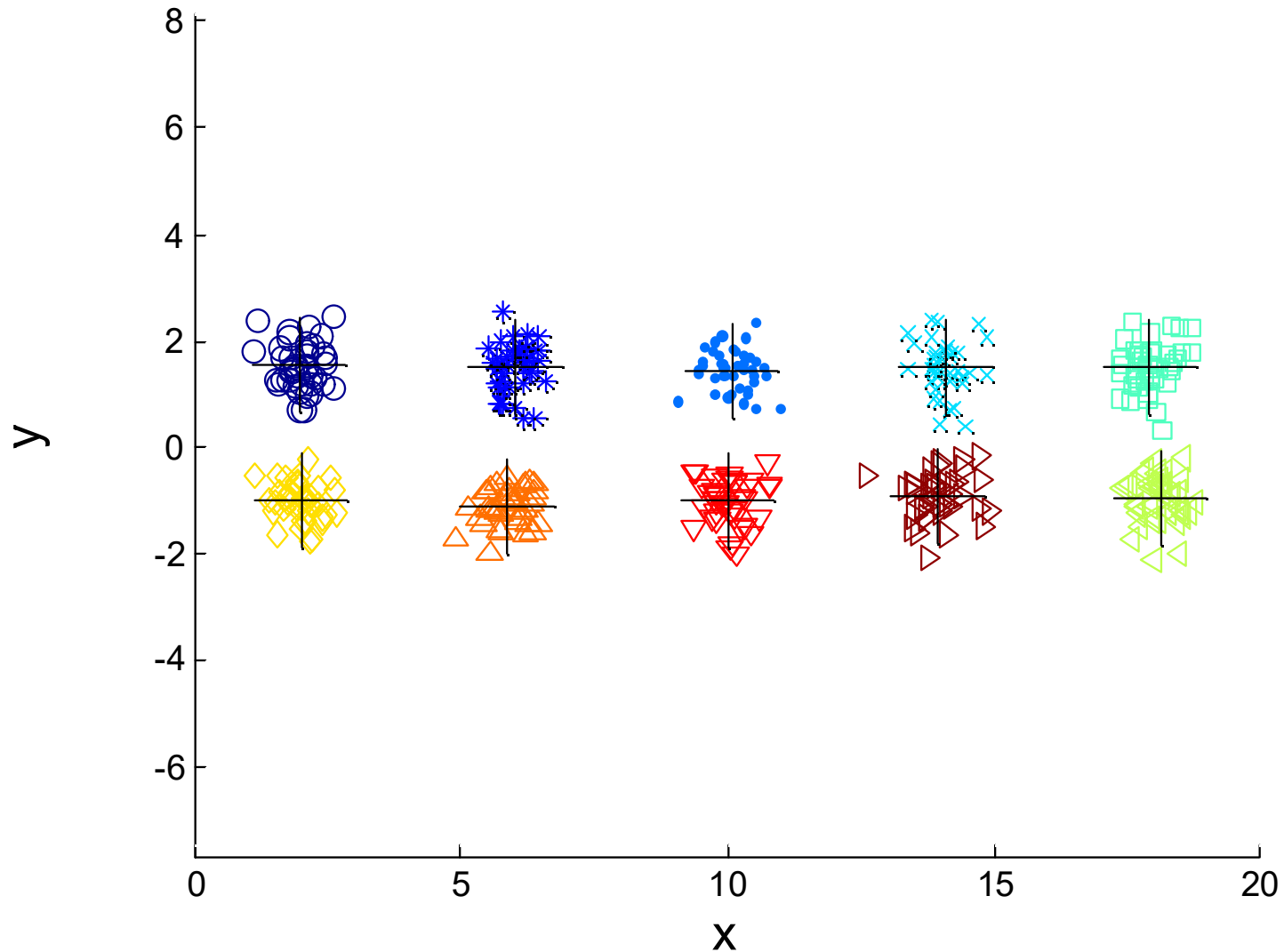
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# 10 Clusters Example (A)

Iteration 4



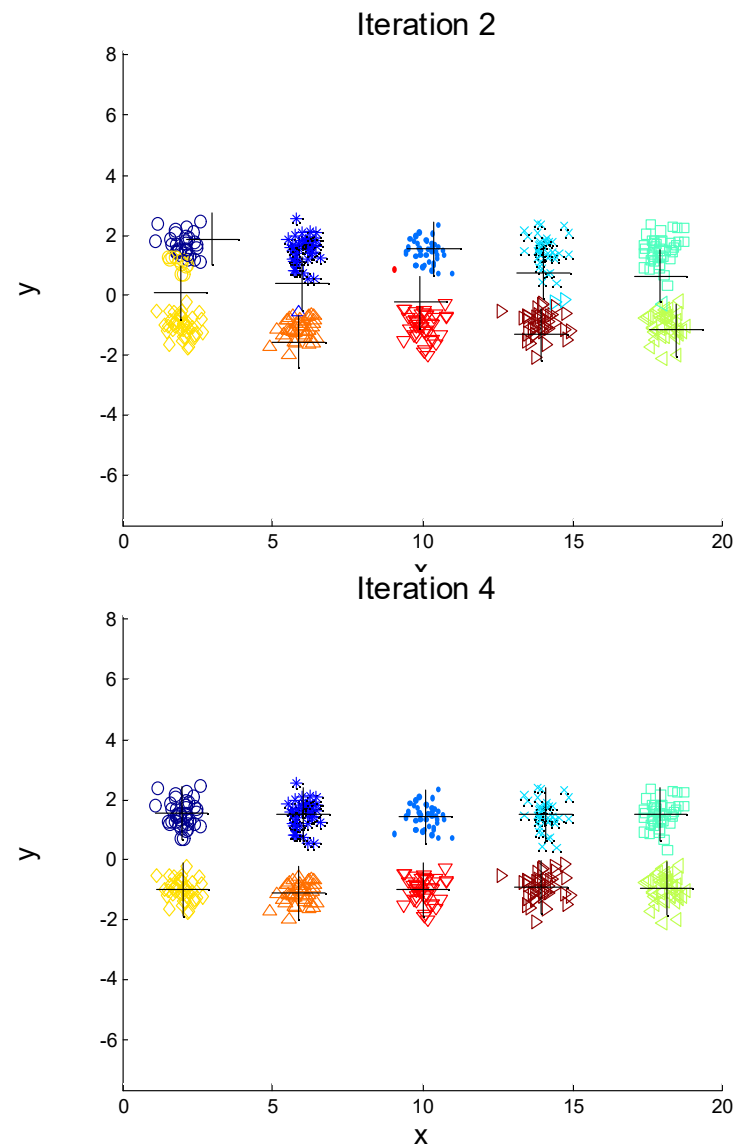
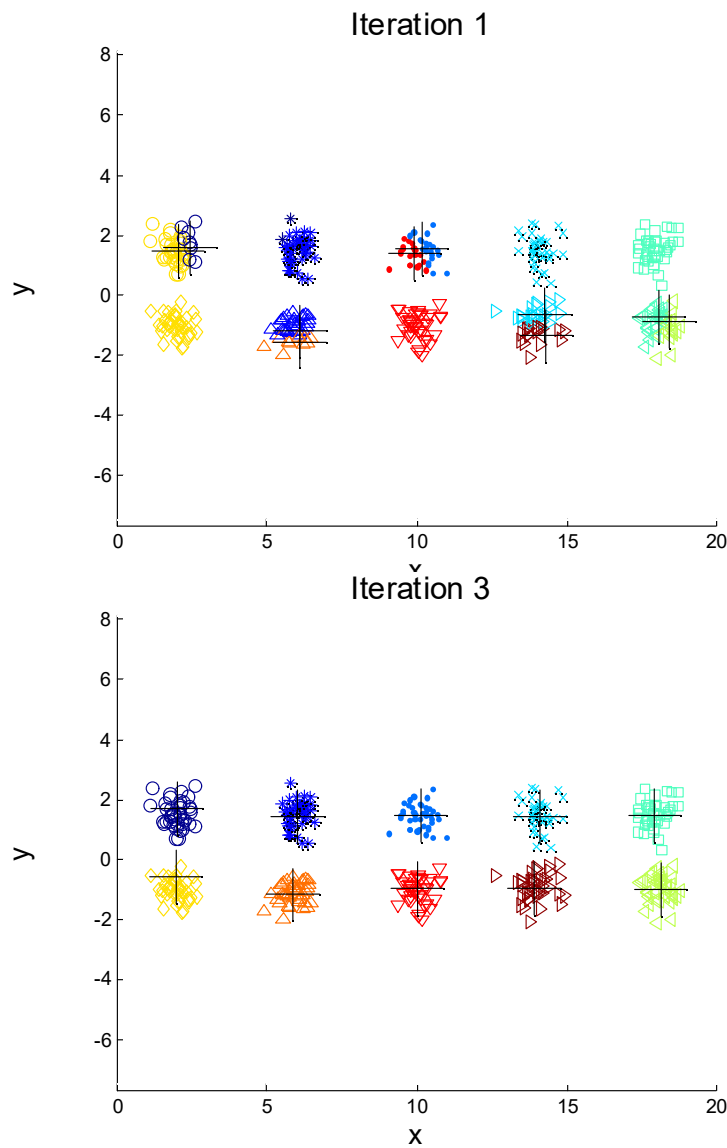
**Starting with two initial centroids in one cluster of each pair of clusters**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 10 Clusters Example (A)



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

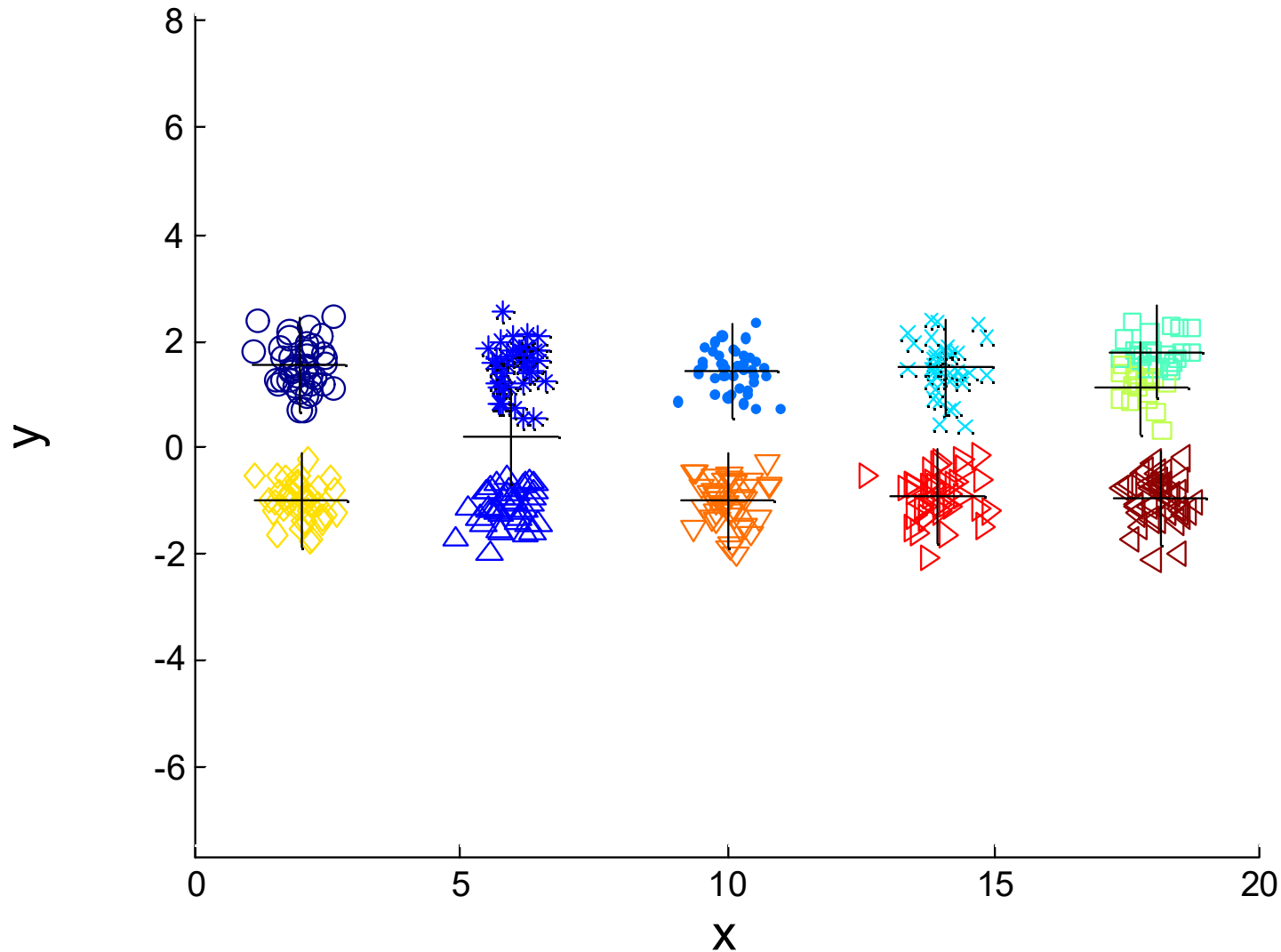
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 10 Clusters Example (B)

Iteration 4



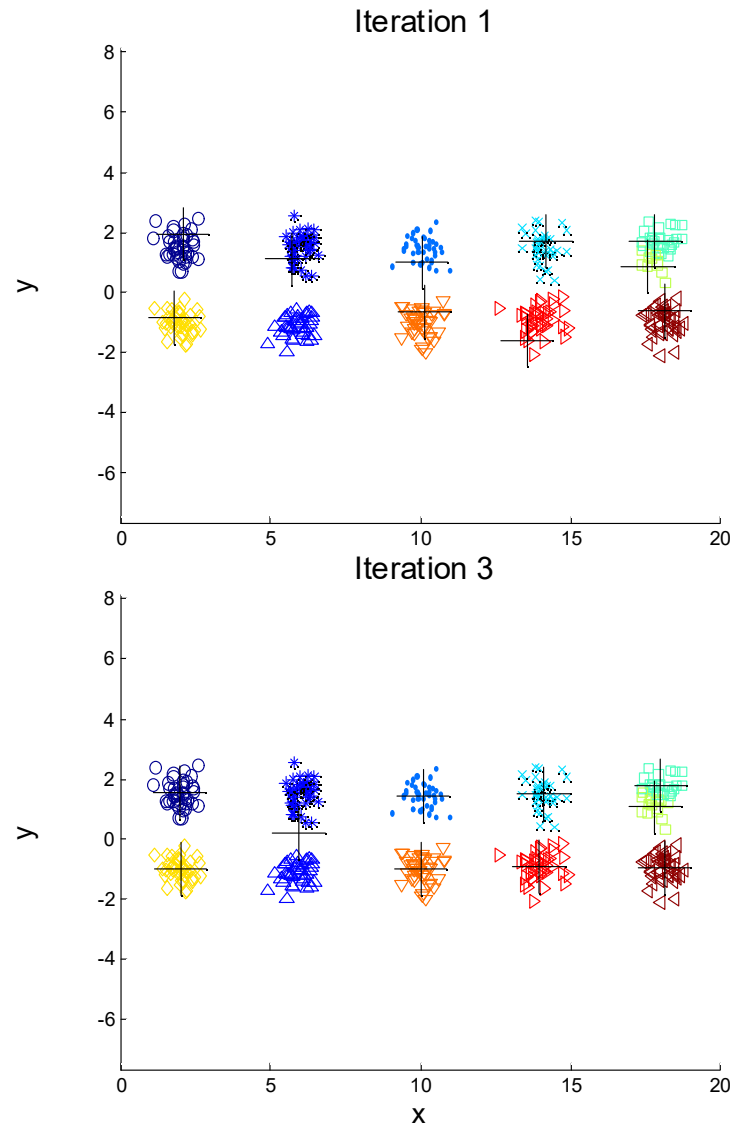
**Starting with some pairs of clusters having three initial centroids, while other have only one or two.**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 10 Clusters Example (B)



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

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 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
- Post-processing
  - Merge similar clusters after the processing
- Bisecting K-means
  - Start with one cluster, and then continuously split
  - Not as susceptible to initialization issues

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Pre-processing and Post-processing

- Pre-processing
  - Normalize the data (prevent from being dominated by big attribute values)
  - 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

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Bisecting K-means

- Bisecting K-means algorithm
  - Start with one cluster and split continuously
  - No need to determine the number of clusters
  - 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
```

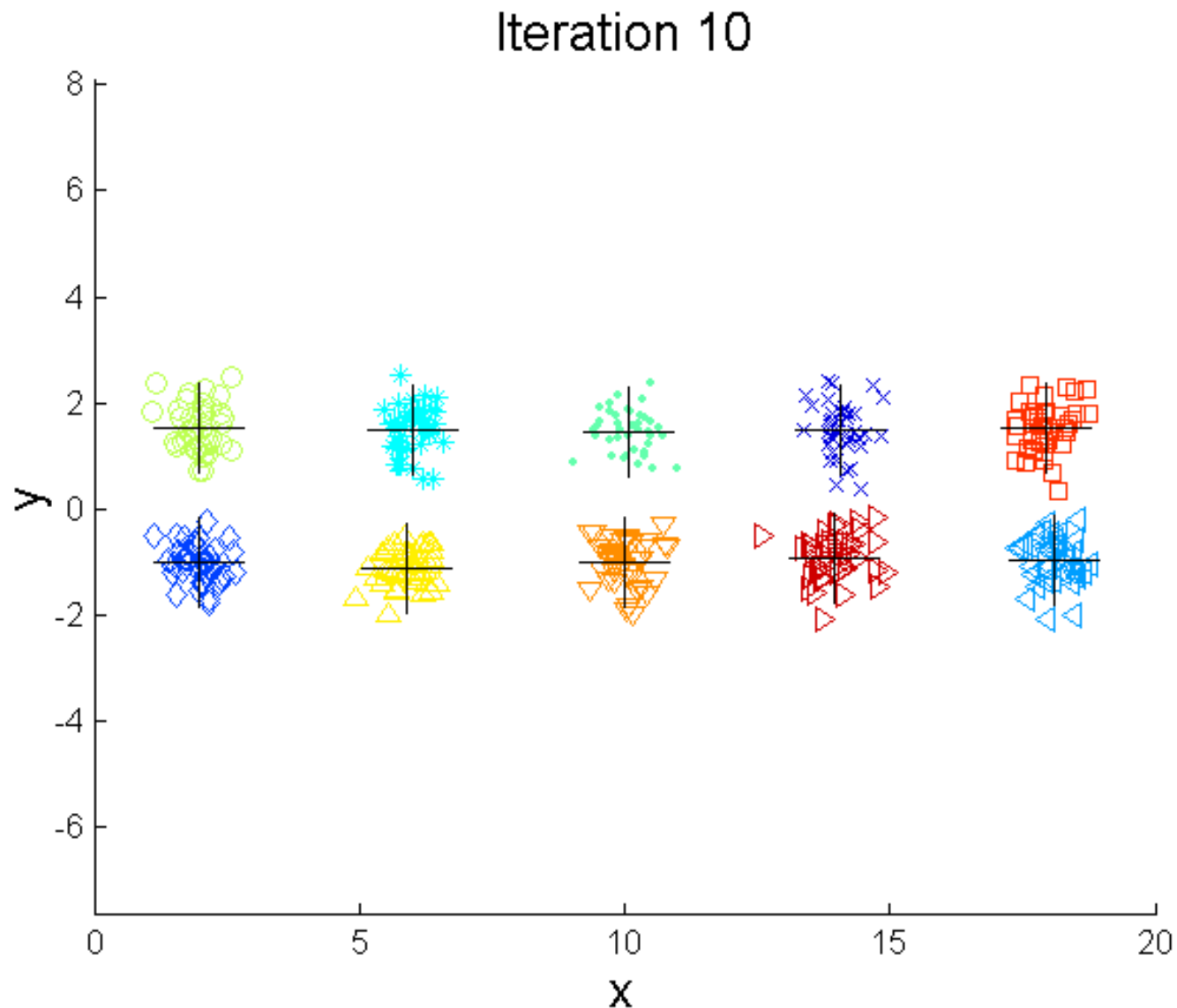
---

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Bisecting K-means Example



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques



# Limitations of K-means

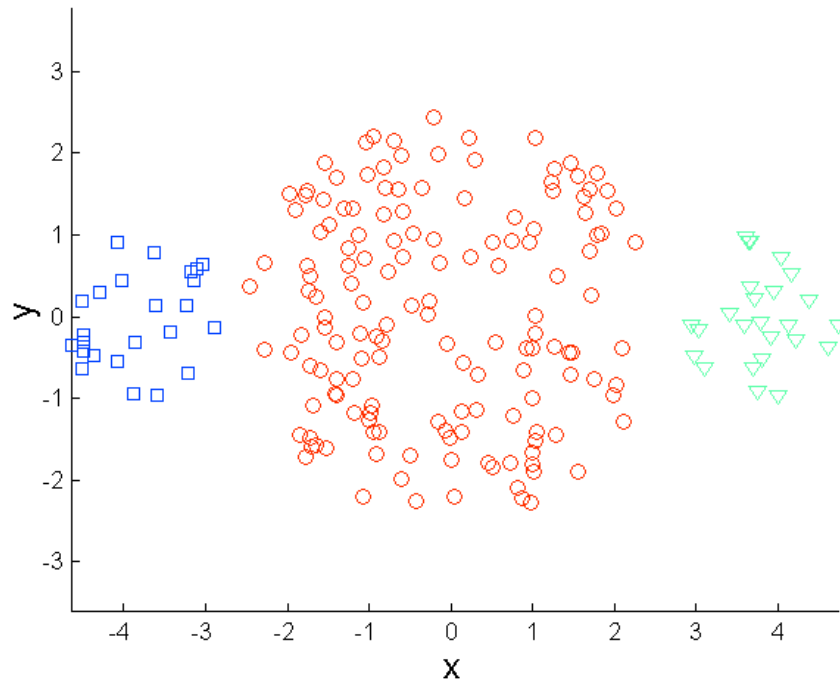
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers.

*Adapted from:*

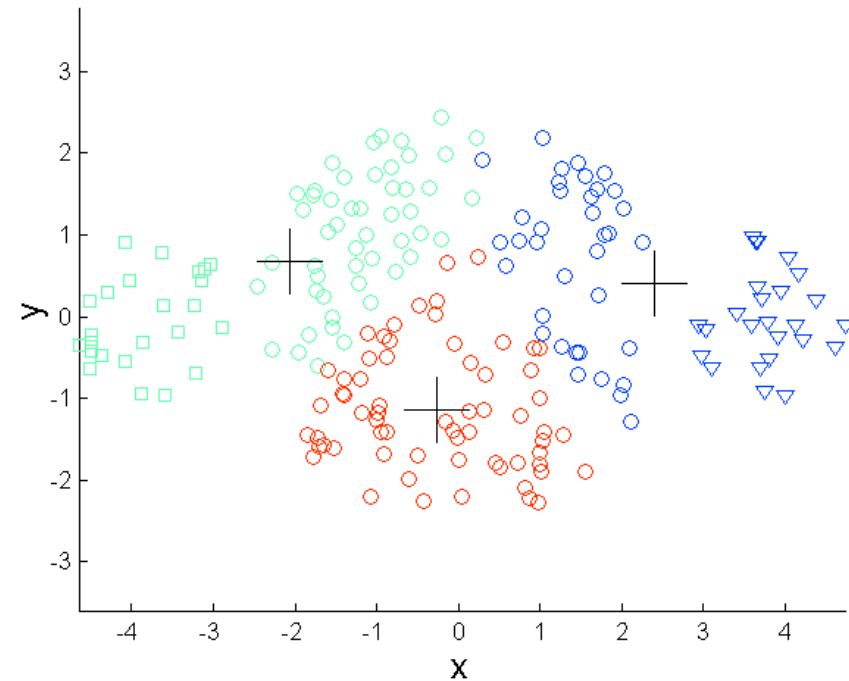
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Limitations of K-means: Differing Sizes



**Original Points**



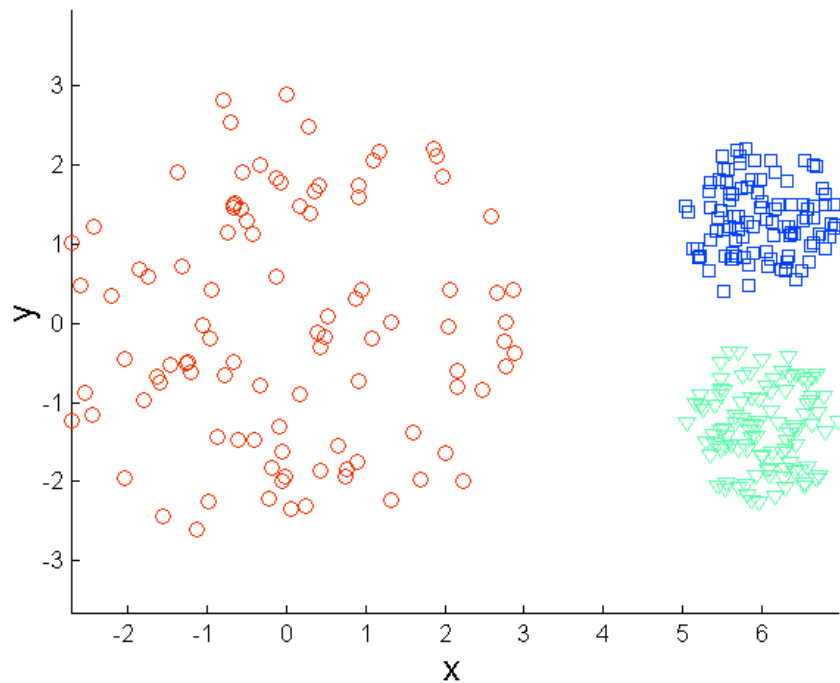
**K-means (3 Clusters)**

*Adapted from:*

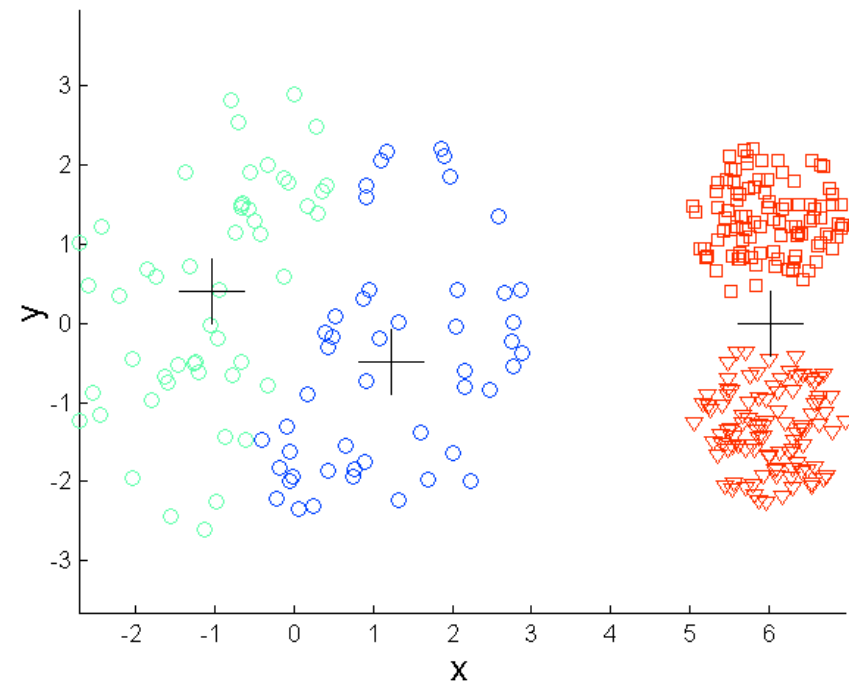
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Limitations of K-means: Differing Density



**Original Points**



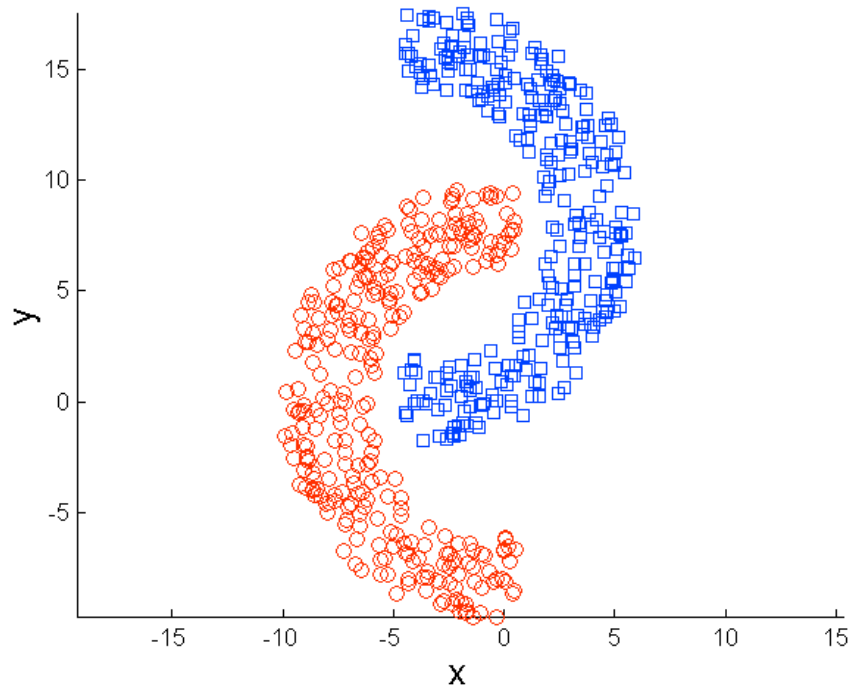
**K-means (3 Clusters)**

*Adapted from:*

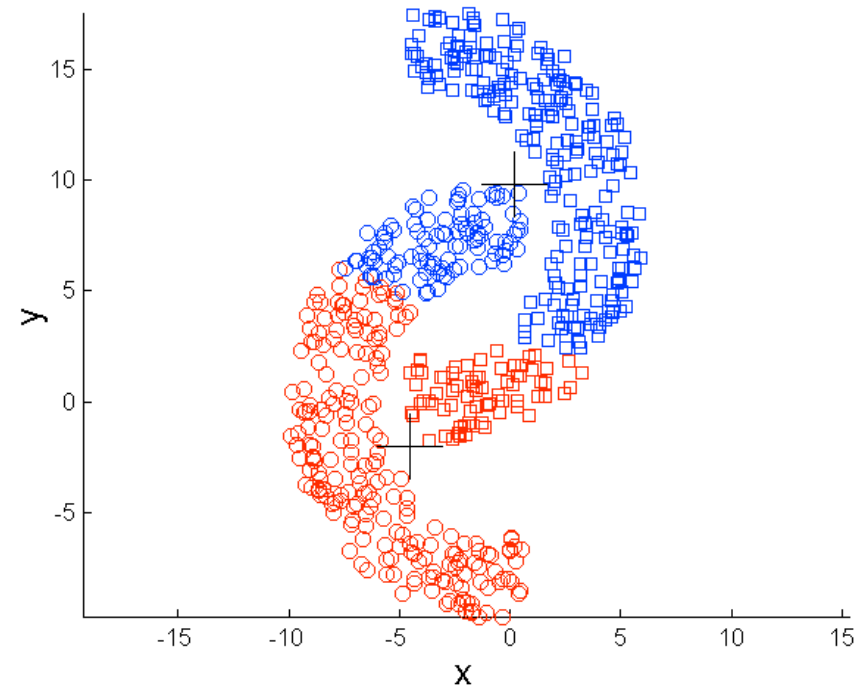
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Limitations of K-means: Non-globular Shapes



**Original Points**



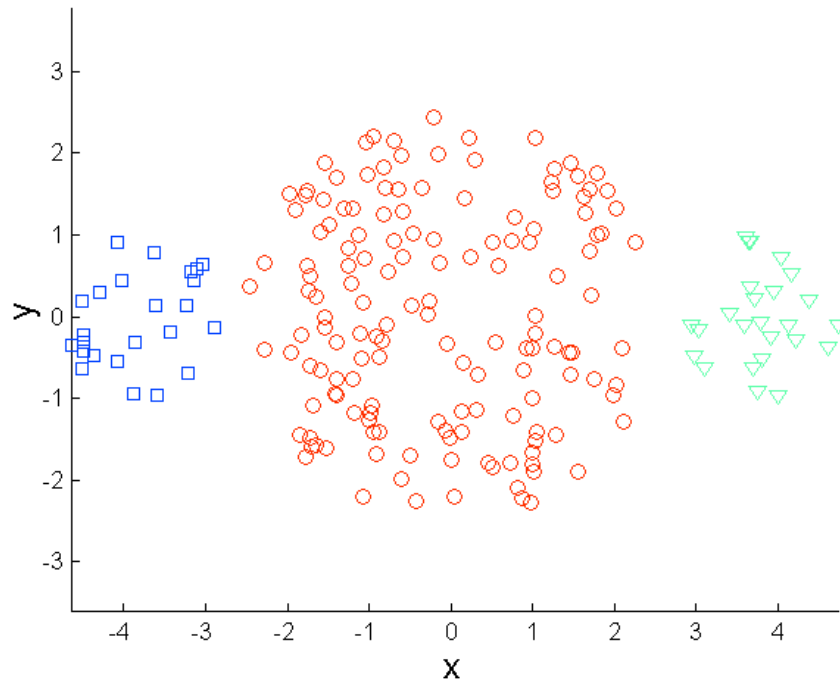
**K-means (2 Clusters)**

*Adapted from:*

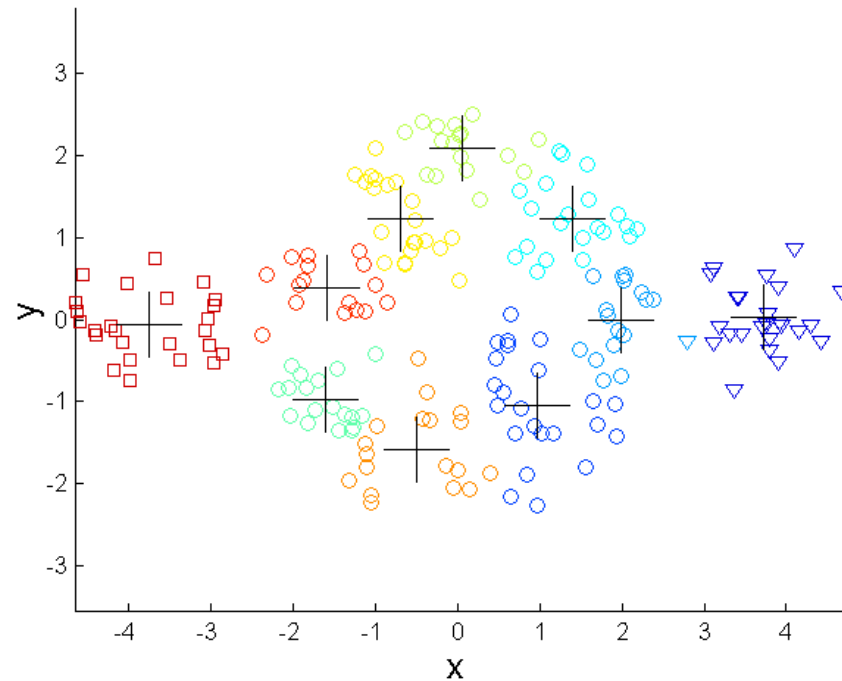
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Overcoming K-means Limitations



**Original Points**



**K-means Clusters**

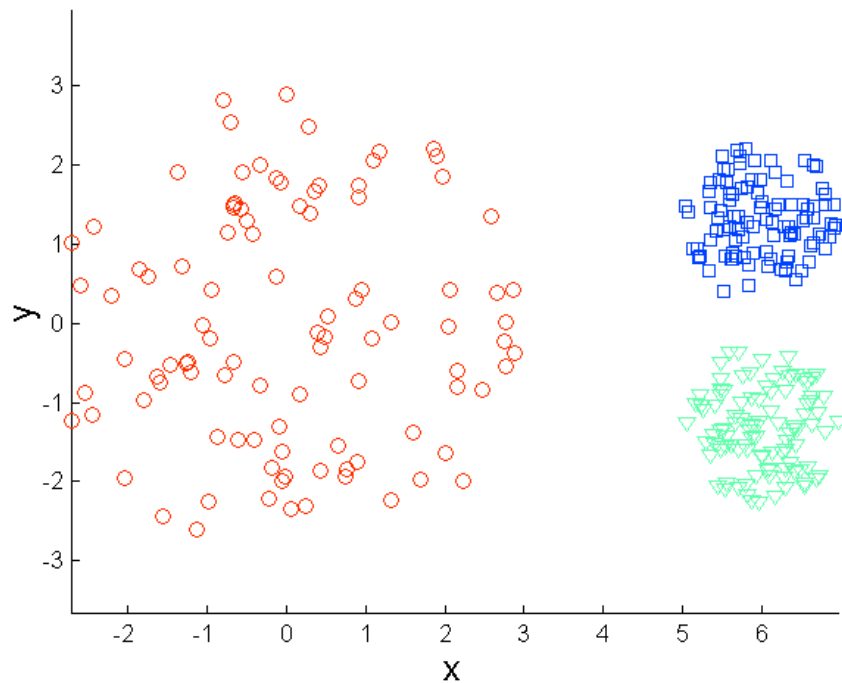
One solution is to use many clusters.  
Find parts of clusters, but need to put together.

*Adapted from:*

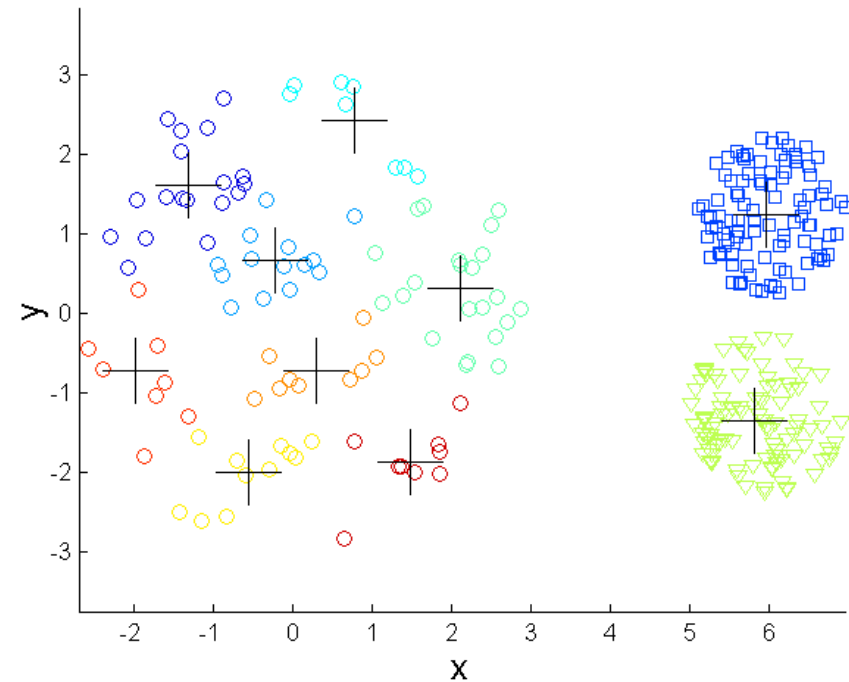
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Overcoming K-means Limitations



**Original Points**



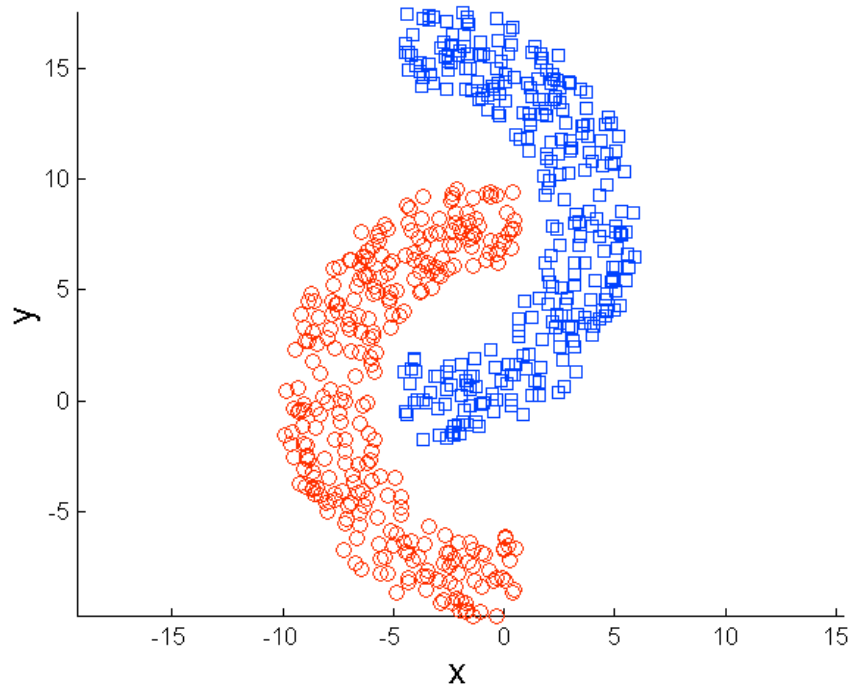
**K-means Clusters**

*Adapted from:*

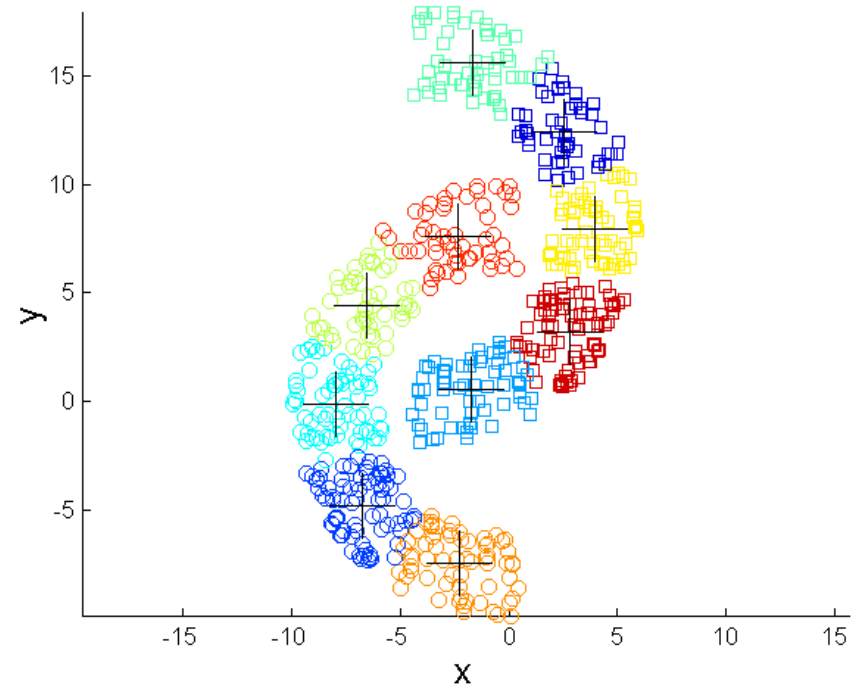
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Overcoming K-means Limitations



**Original Points**



**K-means Clusters**

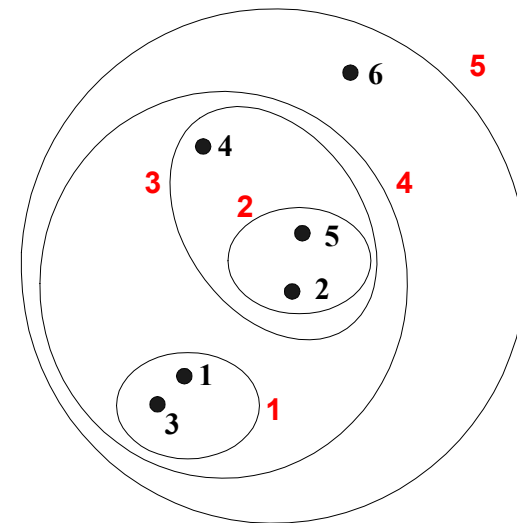
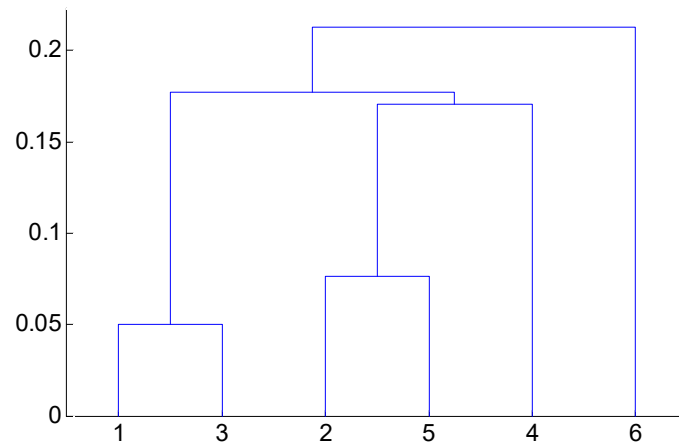
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 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



*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# 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

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 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 a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix

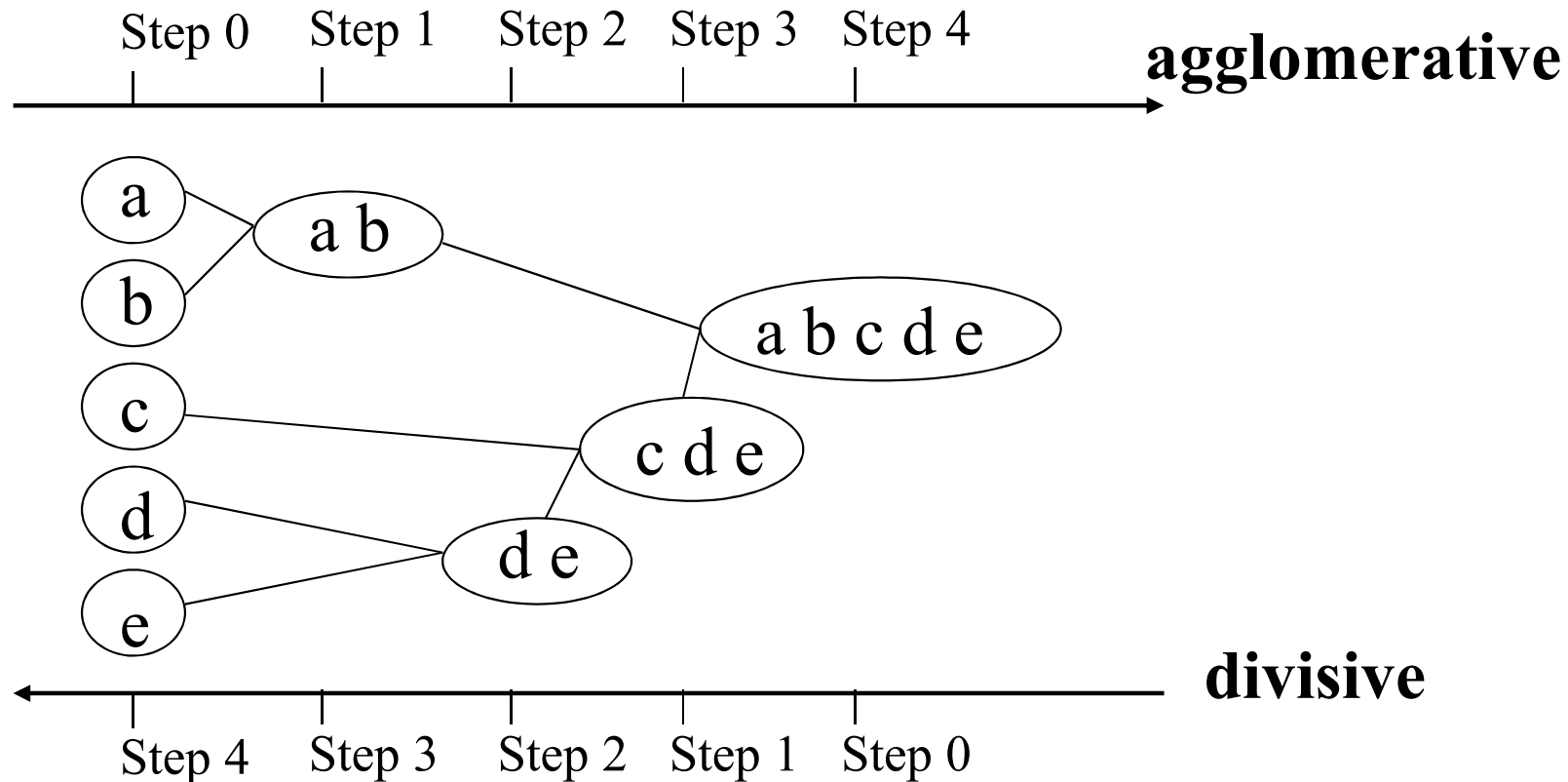
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Hierarchical Clustering

- This method does not require the number of clusters  $k$  as an input, but needs a termination condition



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Agglomerative Clustering Algorithm

- More 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

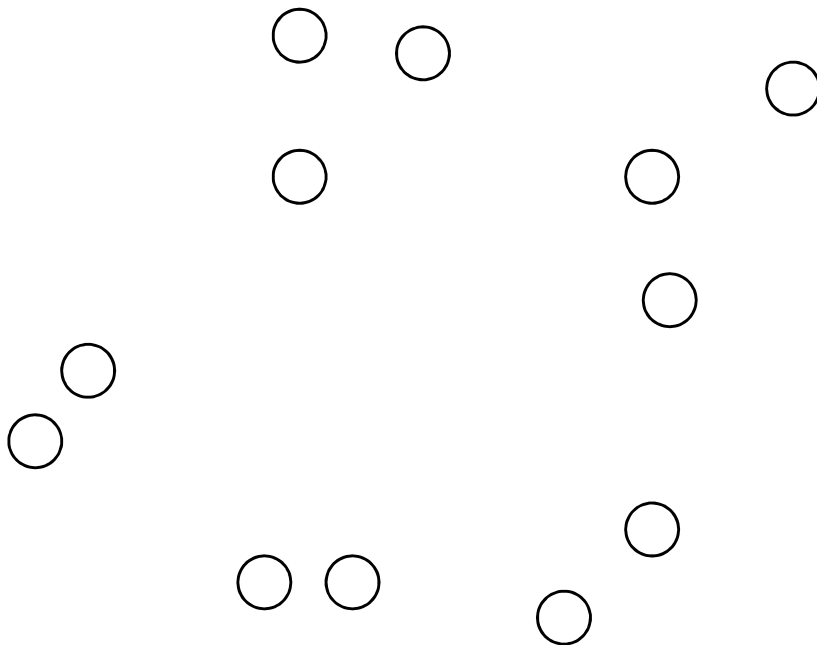
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Starting Situation

- Start with clusters of individual points and a proximity matrix



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

**Proximity Matrix**

    . . .      
 p1 p2 p3 p4 . . . p9 p10 p11 p12

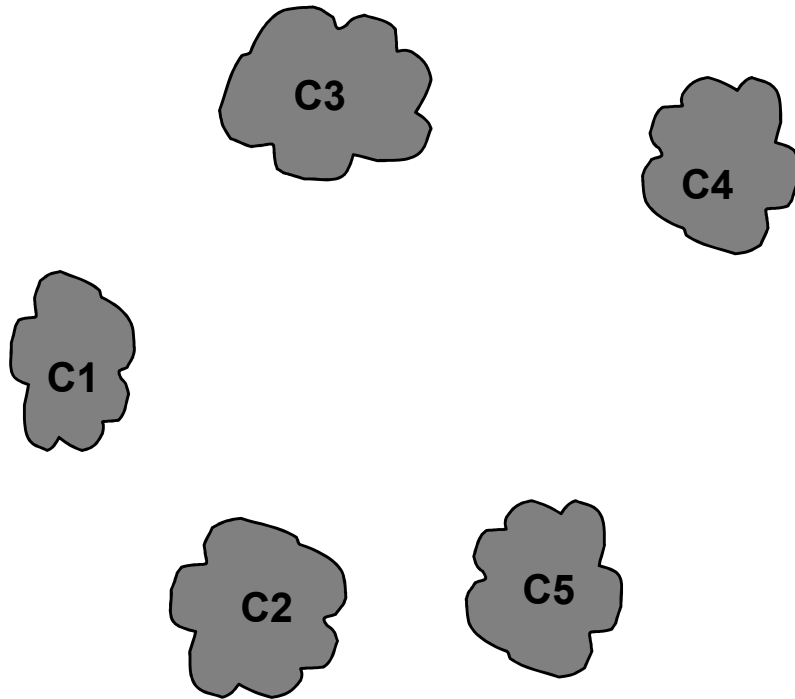
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

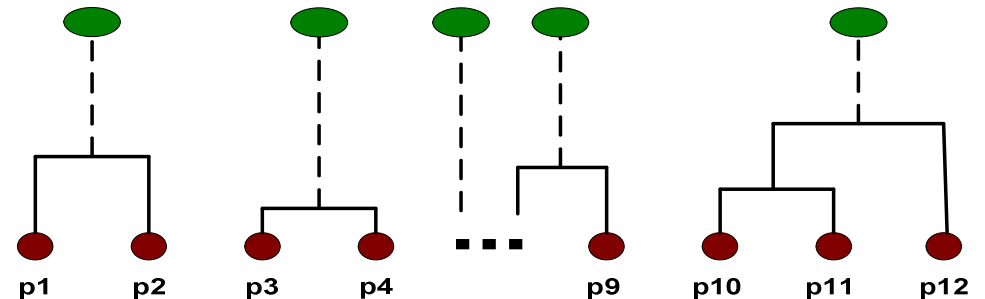
# Intermediate Situation

- After some merging steps, we have some clusters



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

**Proximity Matrix**



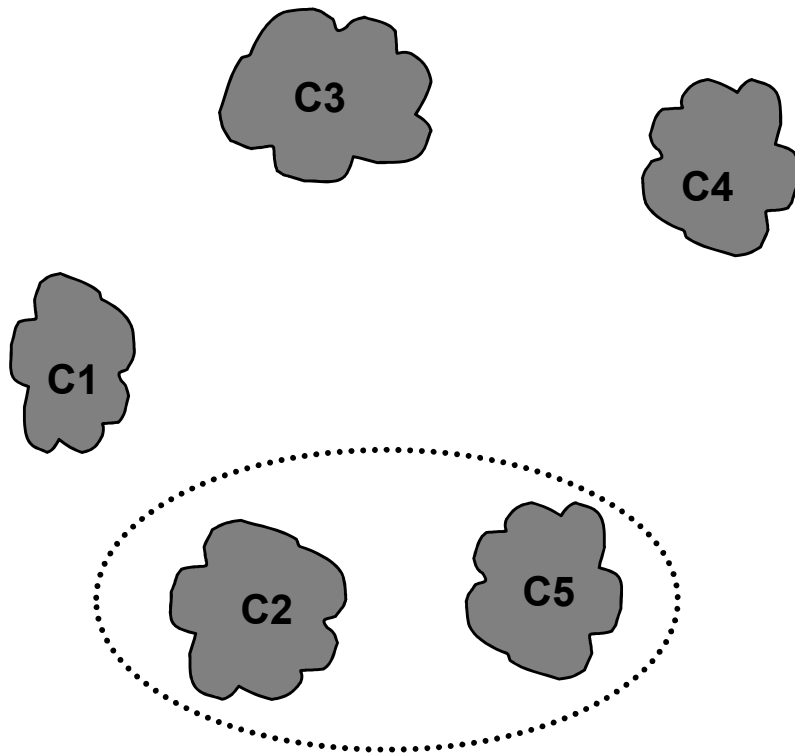
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

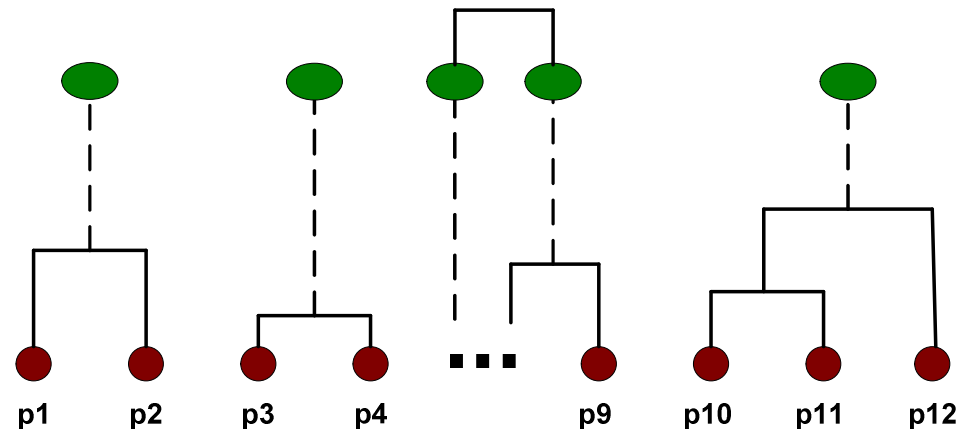
# Intermediate Situation

- 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**



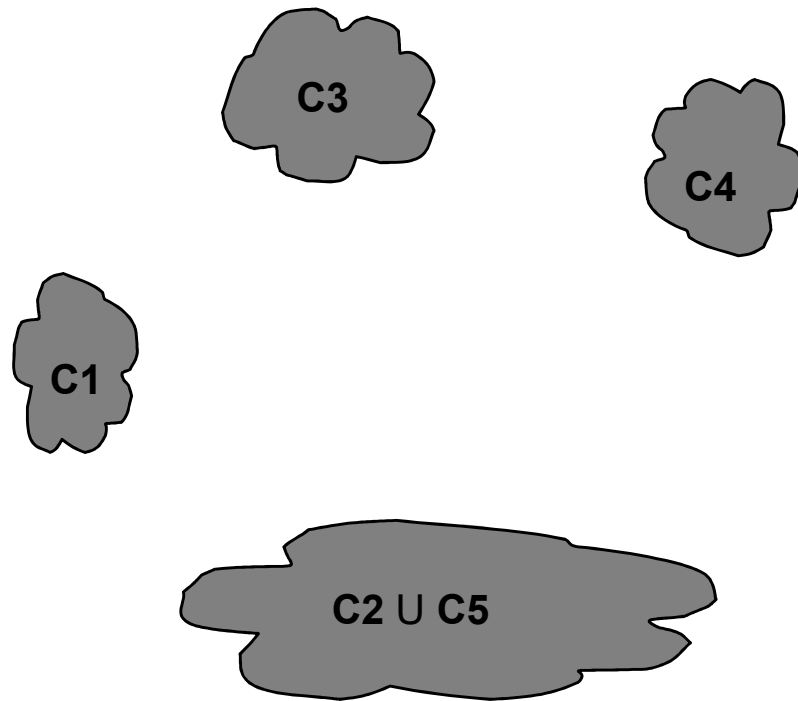
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

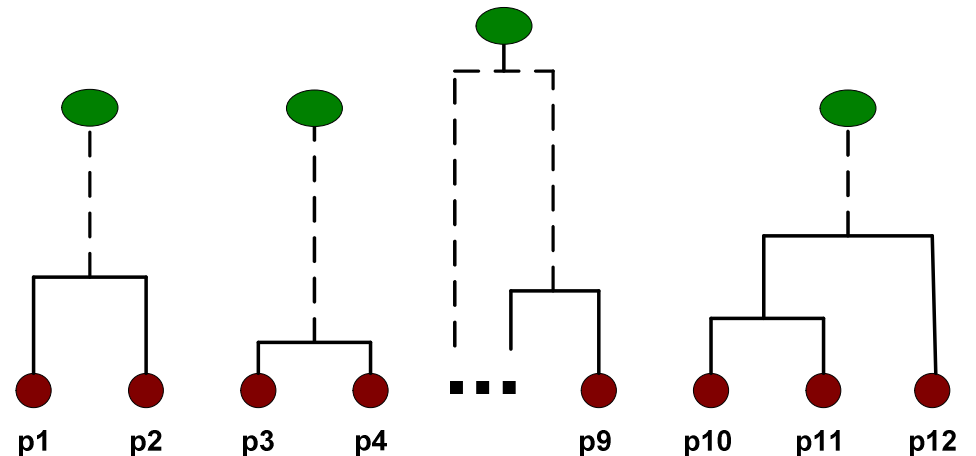
# After Merging

- The question is "How do we update the proximity matrix?"



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

**Proximity Matrix**



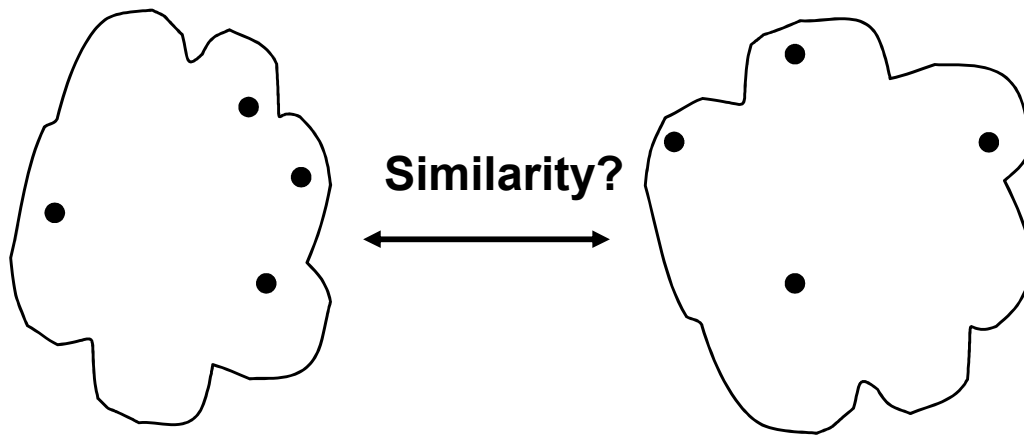
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques



# How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

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

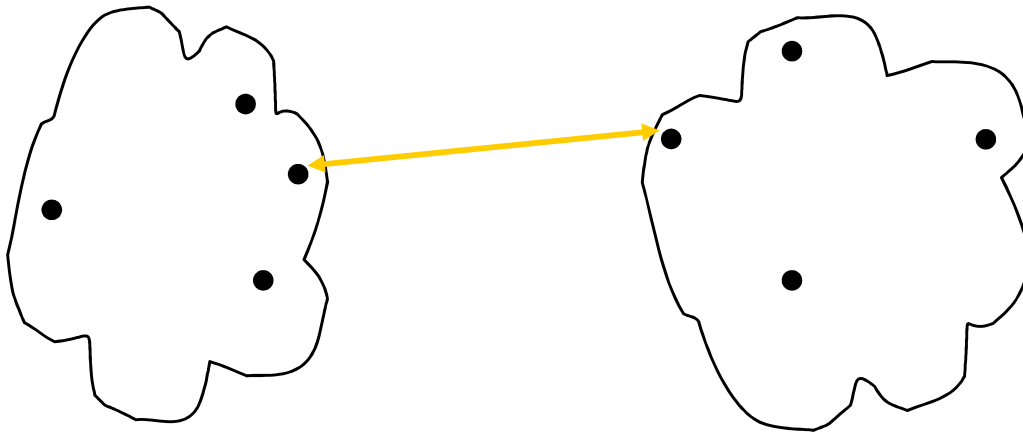
**Proximity Matrix**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# How to Define Inter-Cluster Similarity



- **MIN**
- **MAX**
- **Group Average**
- **Distance Between Centroids**
- **Other methods driven by an objective function**
  - Ward's Method uses squared error

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

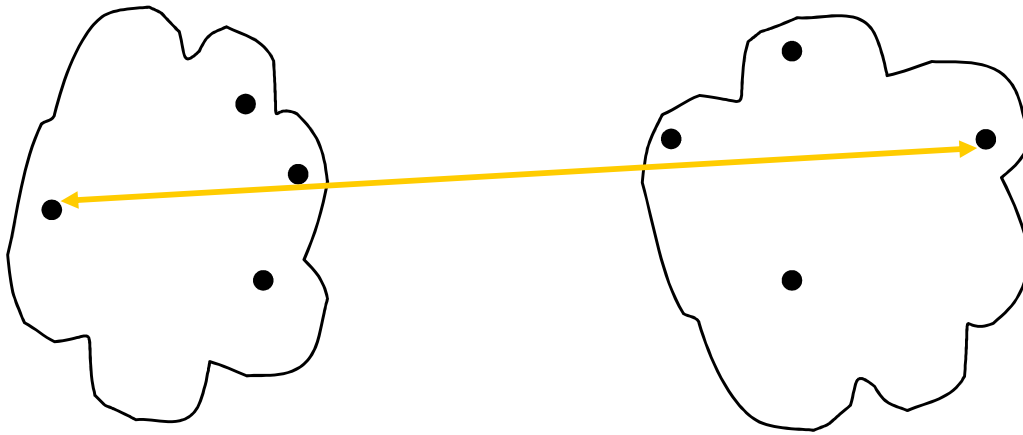
**Proximity Matrix**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

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

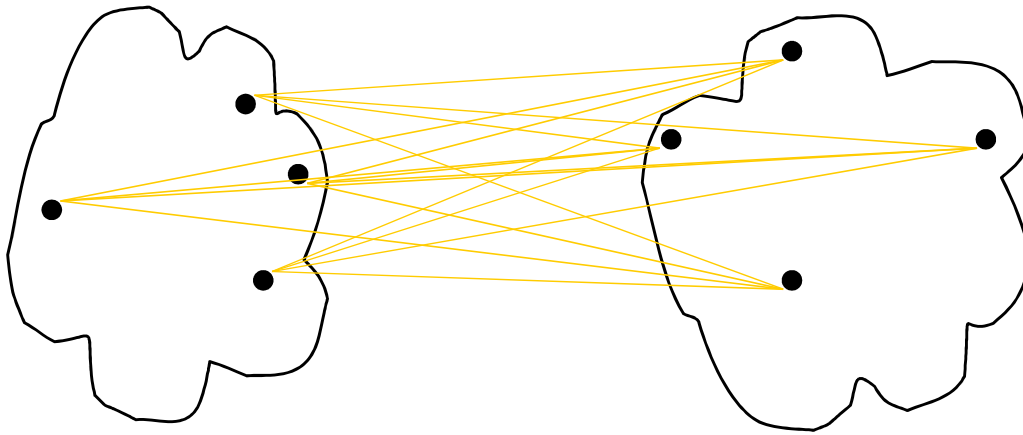
**Proximity Matrix**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# How to Define Inter-Cluster Similarity



- MIN
- MAX
- **Group Average**
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

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

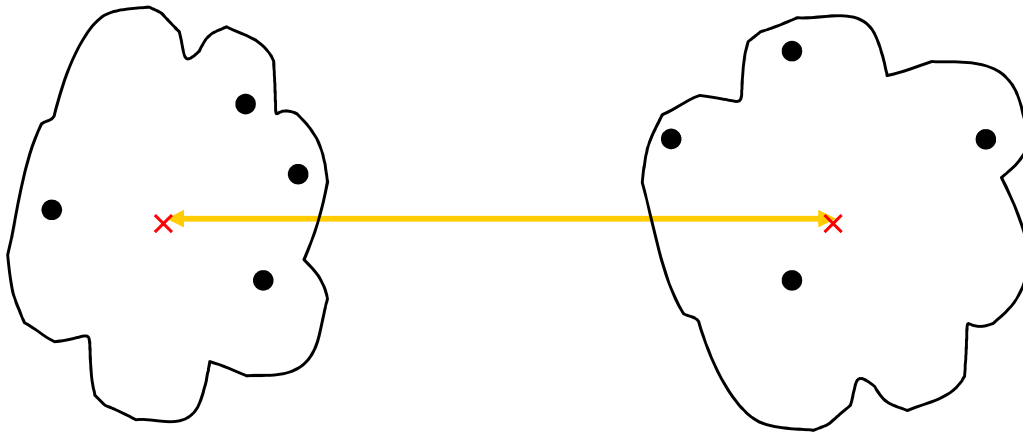
**Proximity Matrix**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

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

**Proximity Matrix**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Cluster Similarity: MIN or Single Link

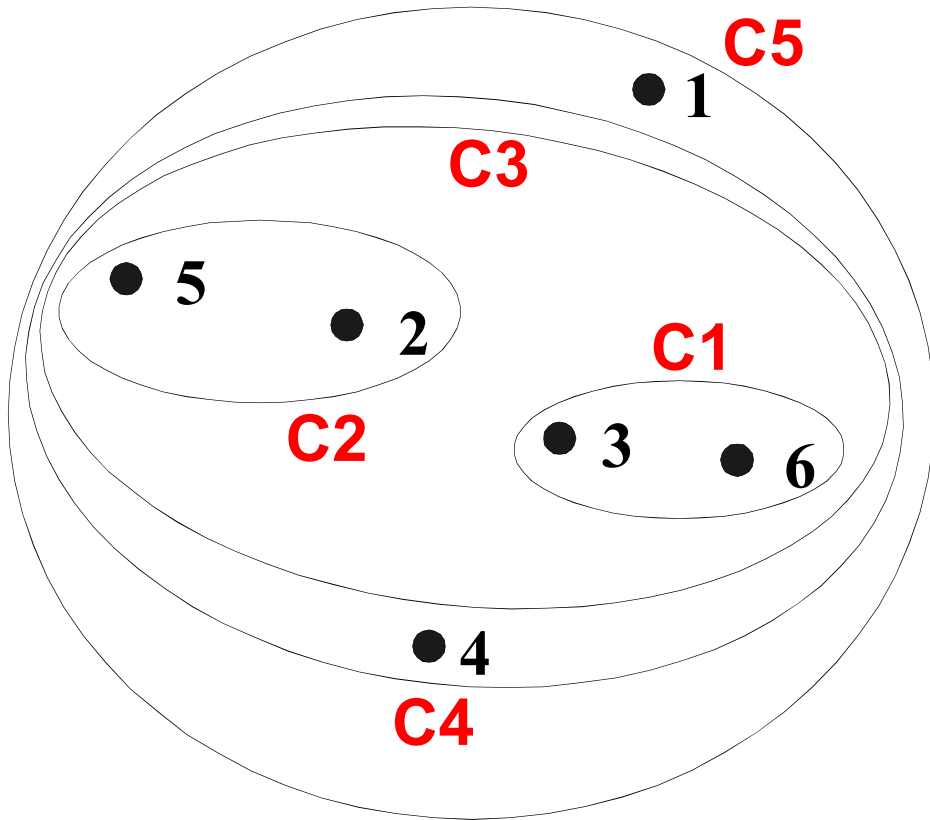
- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.

*Adapted from:*

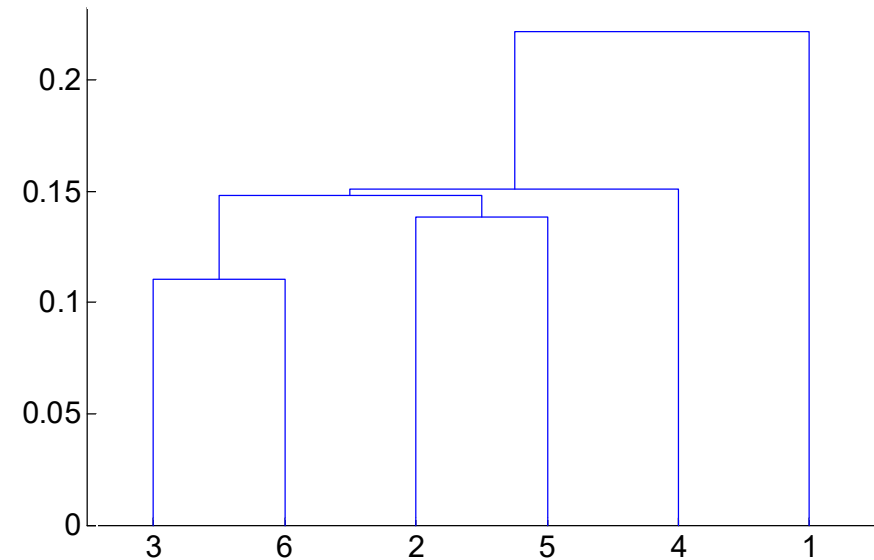
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Hierarchical Clustering: MIN



**Nested Clusters**



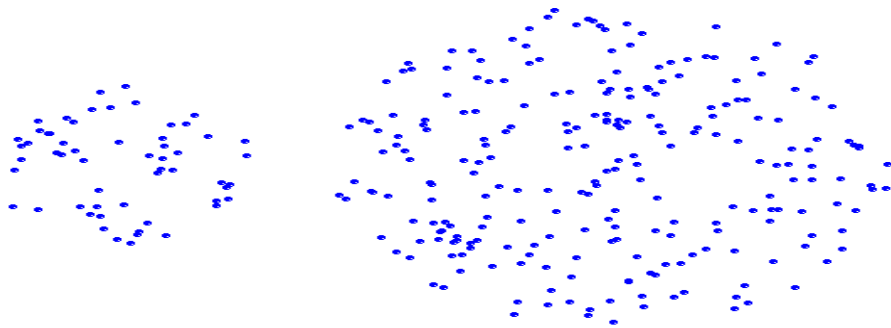
**Dendrogram**

*Adapted from:*

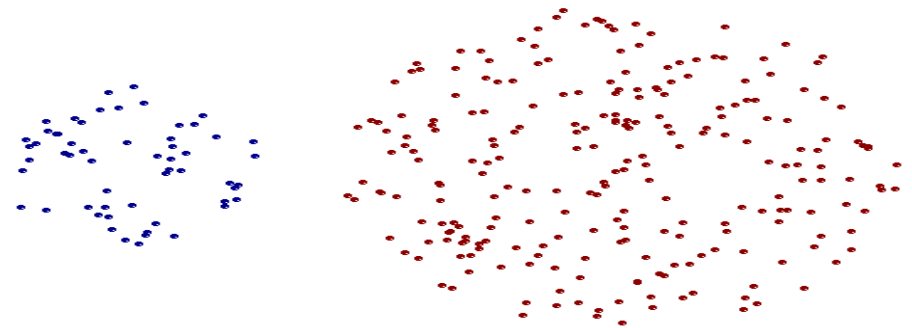
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Strength of MIN



**Original Points**



**Two Clusters**

- **Can handle non-elliptical shapes**

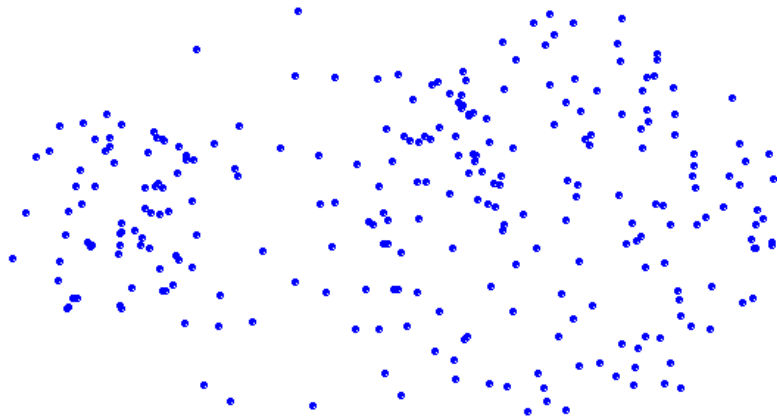
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

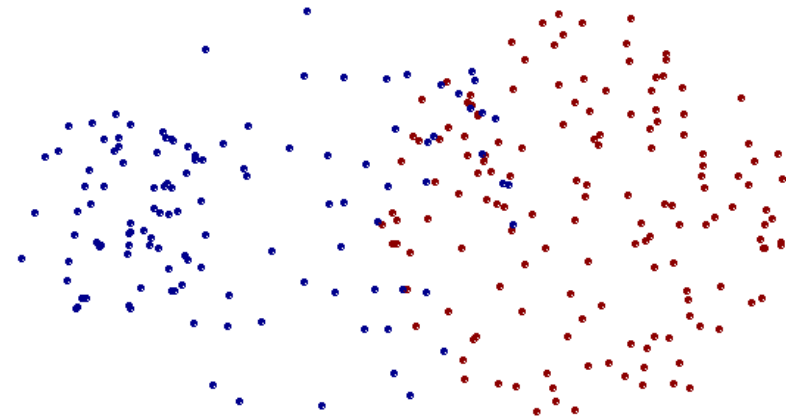
*Han, Kamber - Data Mining: Concepts and Techniques*



# Limitations of MIN



**Original Points**



**Two Clusters**

- **Sensitive to noise and outliers**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

*Adapted from:*

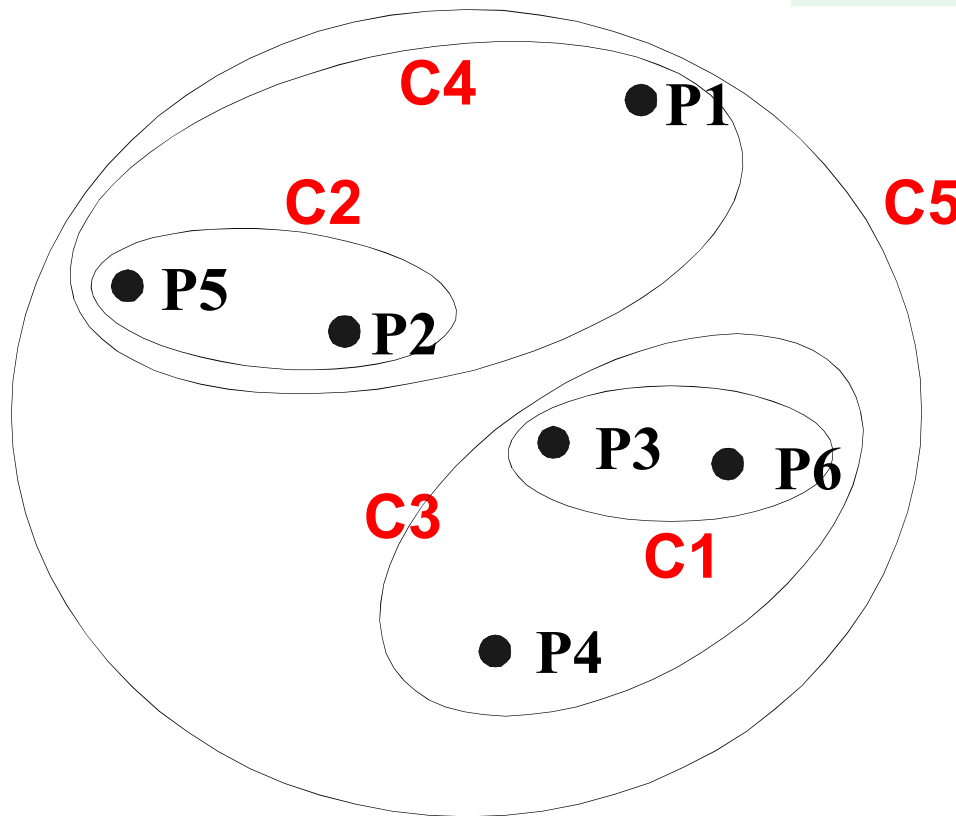
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

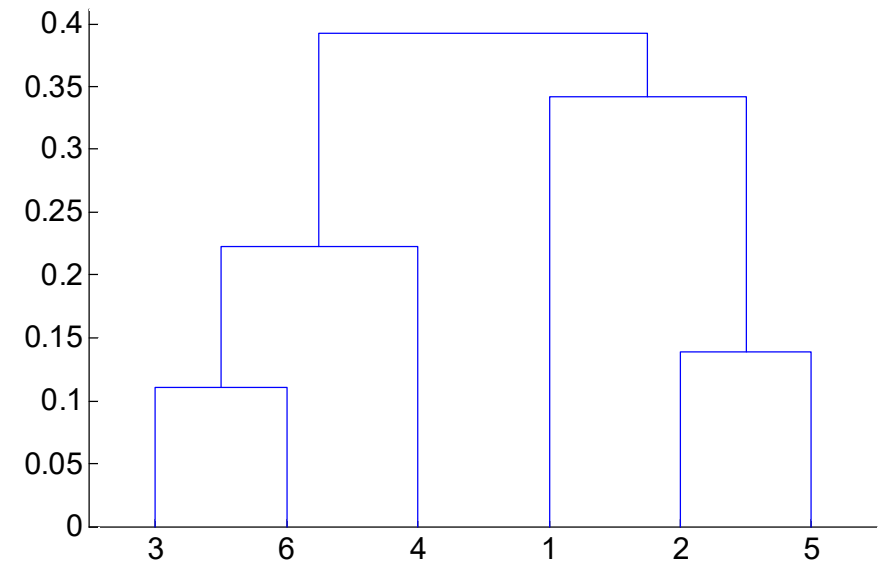
# Hierarchical Clustering: MAX

	C1 or C2	P1	P4
C1(P3, P6)	P5 & P6	P1 & P6	P4 & P6
C2(P2, P5)	P5 & P6	P1 & P5	P4 & P5

*\*MAX → then, FIND MIN among MAX*



**Nested Clusters**



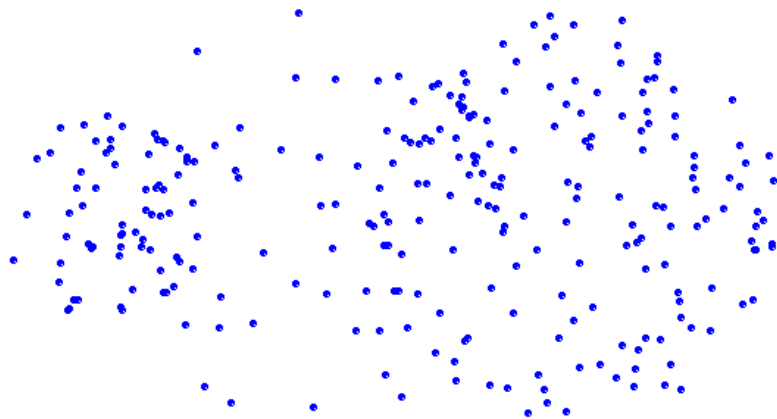
**Dendrogram**

*Adapted from:*

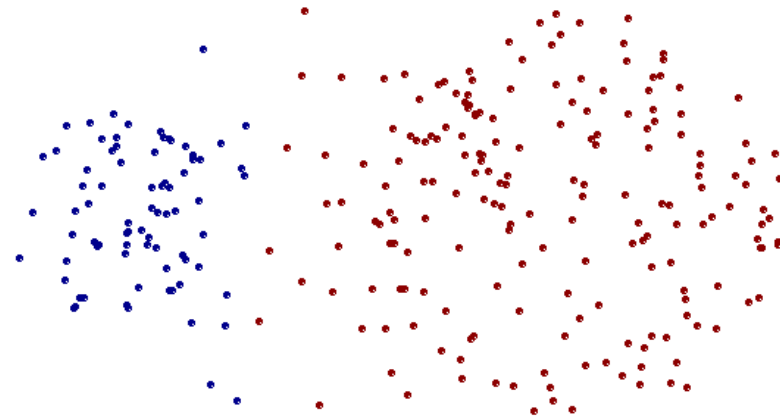
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Strength of MAX



**Original Points**



**Two Clusters**

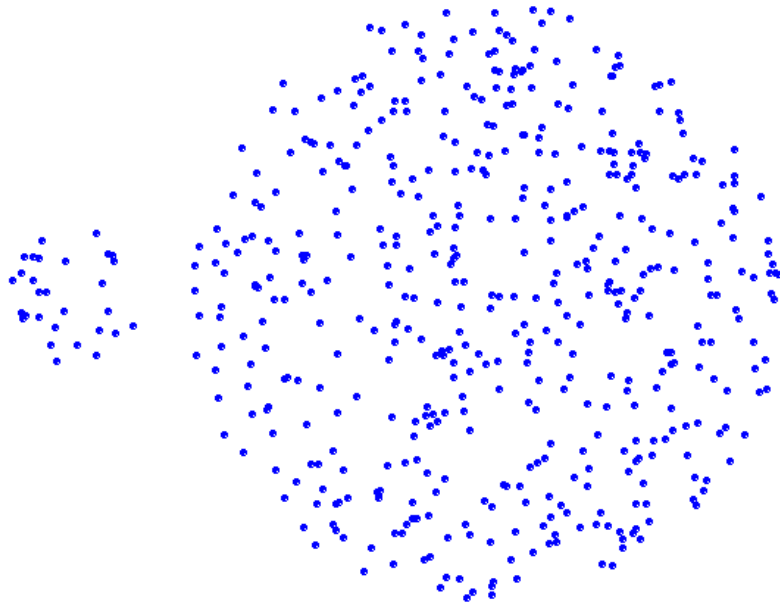
- **Less susceptible to noise and outliers**

*Adapted from:*

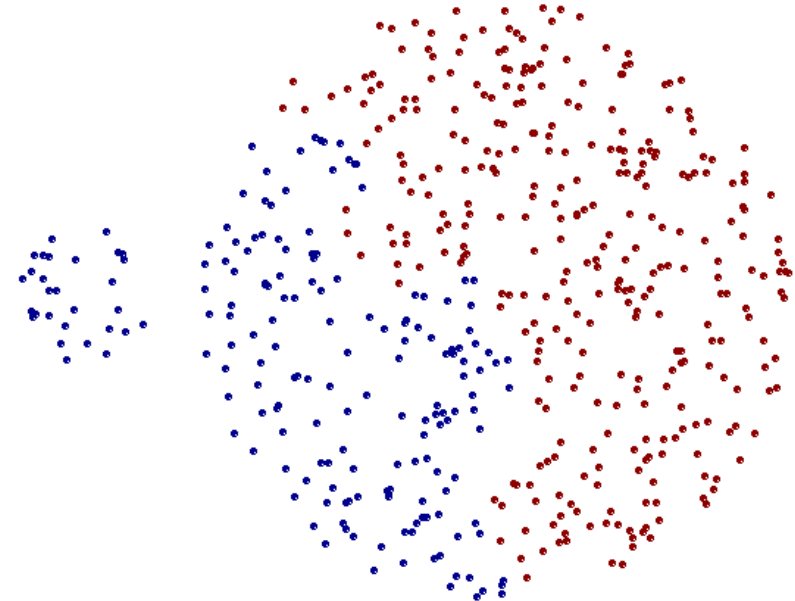
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Limitations of MAX



**Original Points**



**Two Clusters**

- **Tends to break large clusters**
- **Biased towards globular clusters**

이 슬라이드 후에 연습문제

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Cluster Similarity: Group Average

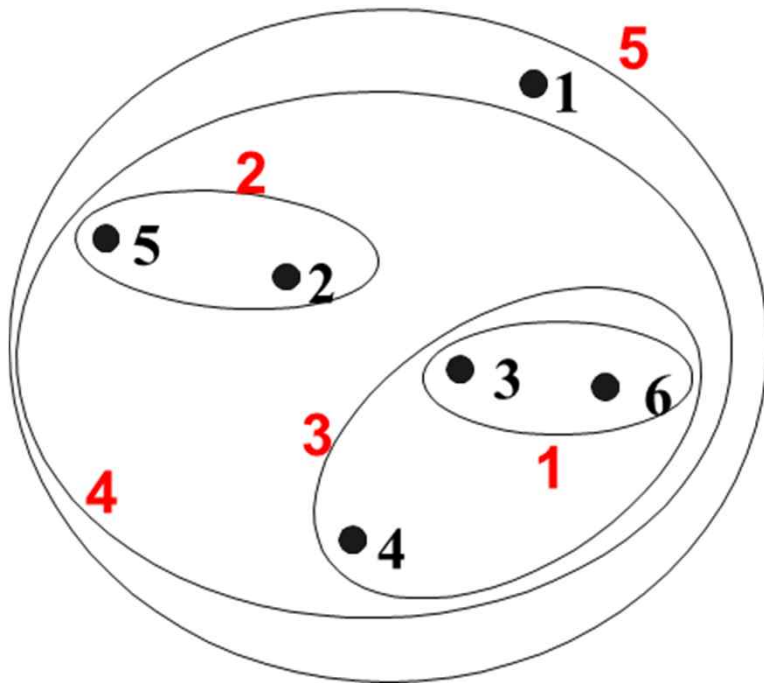
- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.
- Need to use average connectivity for scalability since total proximity favors large clusters

*Adapted from:*

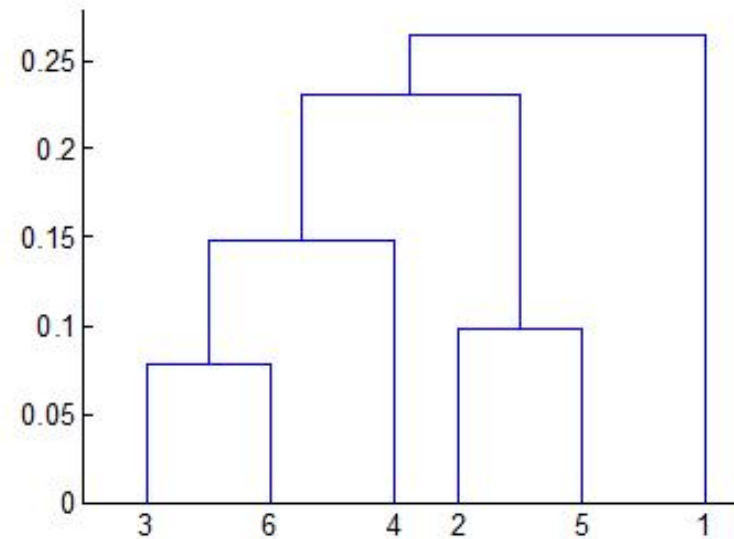
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Hierarchical Clustering: Group Average



**Nested Clusters**



**Dendrogram**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# Cluster Similarity: Ward's Method

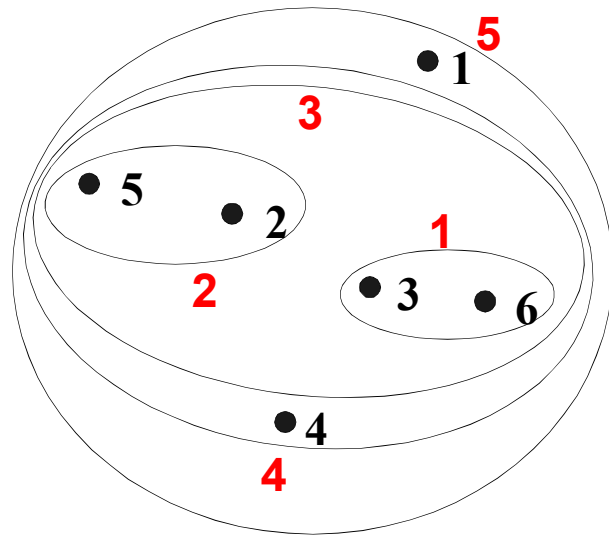
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

*Adapted from:*

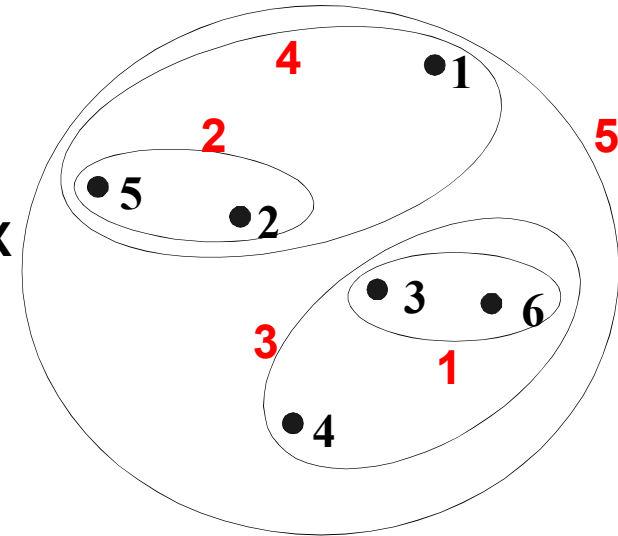
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

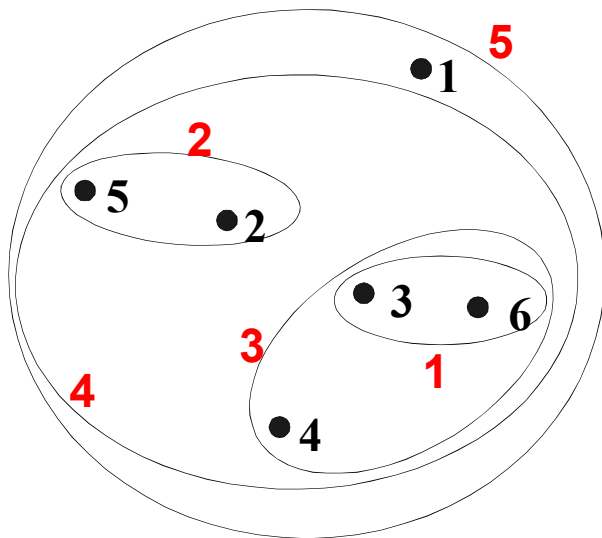
# Hierarchical Clustering: Comparison



MIN

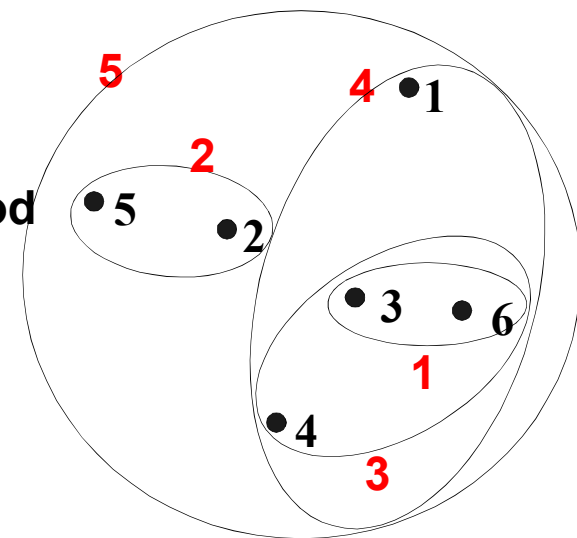


MAX



Group Average

Ward's Method



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Hierarchical Clustering: Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- Sensitivity to noise and outliers
- Difficulty handling different sized clusters and convex shapes
- Large computation time and storage (compute all the matrix first)
- No objective function: change depends on different measurement scheme

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Density-Based Clustering Methods

- Clustering based on density, such as density-connected points
- Major features
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# DBSCAN

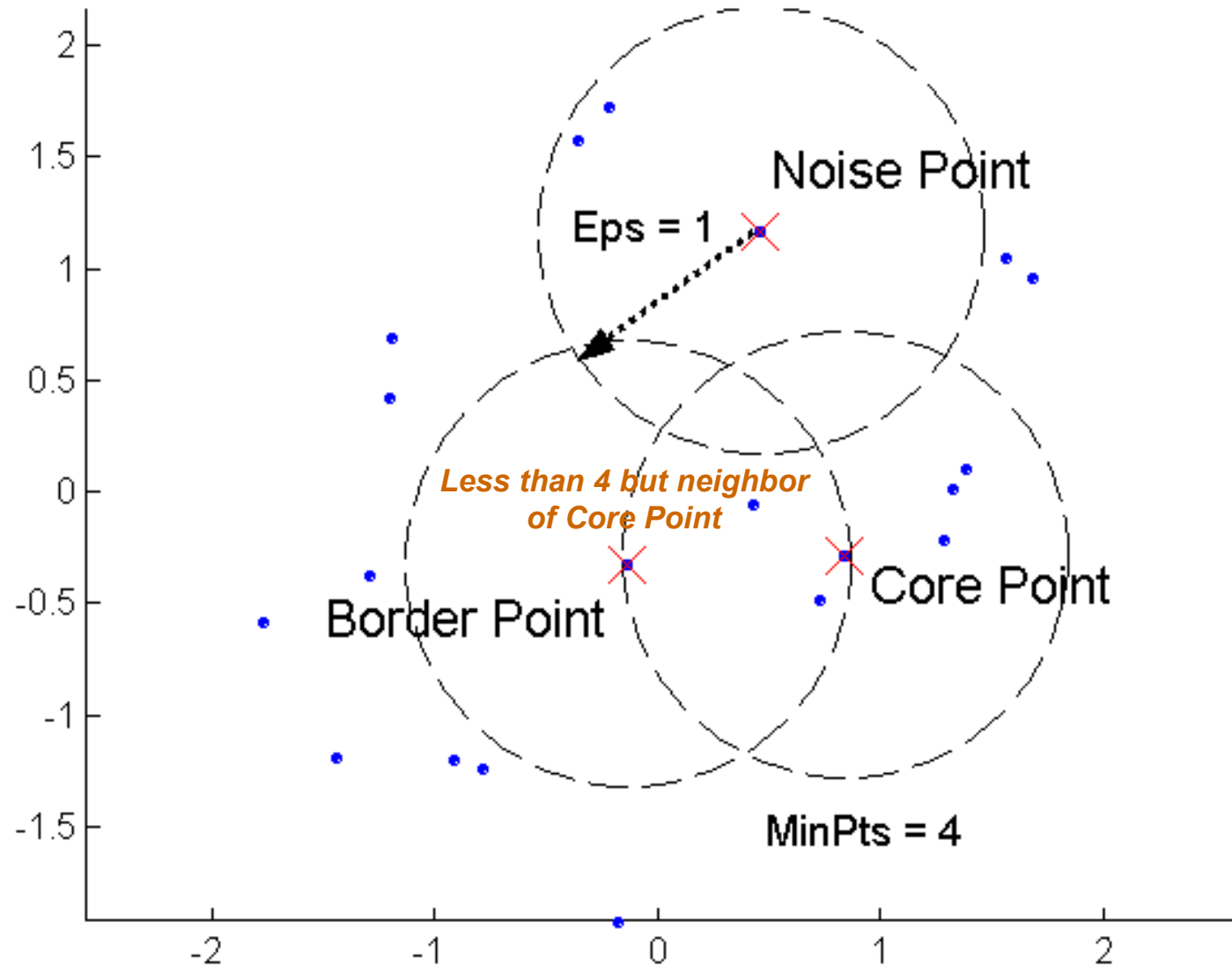
- DBSCAN is a density-based algorithm
  - Density = number of points within a specified radius (Eps, Epsilon: 임의의 매우 작은 양수)
  - A point is a **core point** if it has more than a specified number of points (MinPts) within Eps
  - These are the points that are at the interior of a cluster
  - A **border point** has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A **noise point** is any point that is not a core point or a border point.

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Core, Border, and Noise Points



Adapted from:

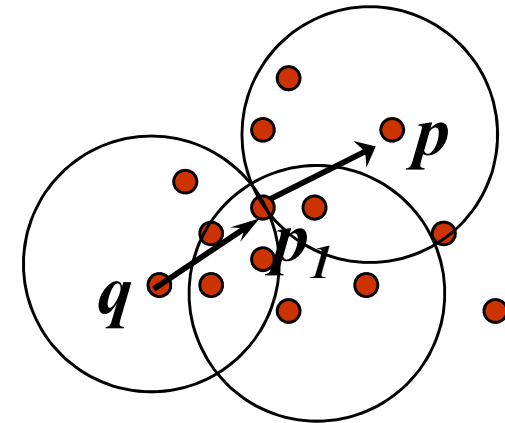
Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Density-Based Clustering Concepts

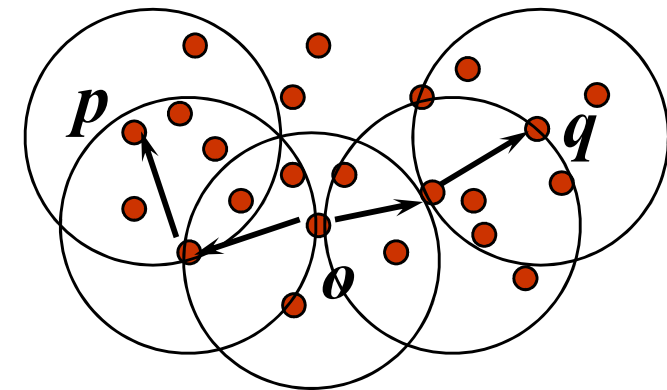
## ■ Density-reachable

- A point  $p$  is density-reachable from a point  $q$  with respect to  $Eps$ ,  $MinPts$  if there is a chain of points  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  such that  $p_{i+1}$  is directly density-reachable from  $p_i$



## ■ Density-connected

- A point  $p$  is density-connected to a point  $q$  with  $Eps$ ,  $MinPts$  if there is a point  $o$  such that both,  $p$  and  $q$  are density-reachable from  $o$  with respect to  $Eps$  and  $MinPts$ .



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# DBSCAN Algorithm

- Arbitrary select a point  $p$
- Retrieve all points density-reachable from  $p$  with respect to ***Eps*** and ***MinPts***.
- If  $p$  is a core point, a cluster is formed.
- If  $p$  is a border point, no points are density-reachable from  $p$  and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

1. ***Label all points as core, border, or noise***
2. ***Eliminate noise points***
3. ***Make each group of connected cores***
4. ***Border points: Edge***

Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

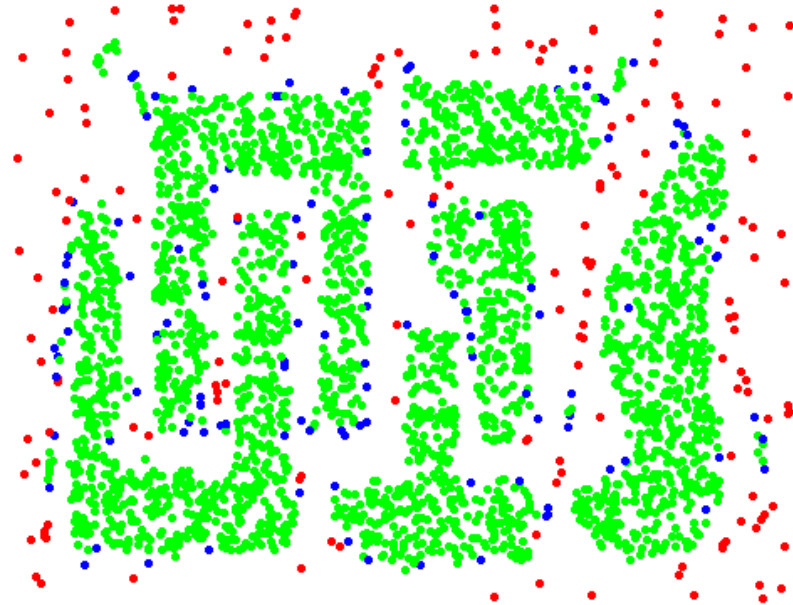
Han, Kamber - Data Mining: Concepts and Techniques



# DBSCAN: Core, Border and Noise Points



Original Points



Point types: **core**,  
**border** and **noise**

**Eps = 10, MinPts = 4**

*Adapted from:*

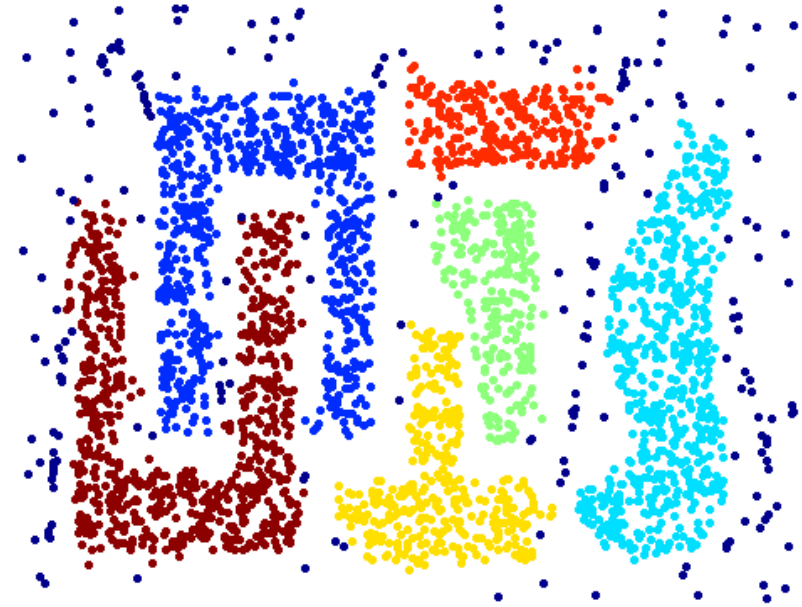
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# When DBSCAN Works Well



**Original Points**



**Clusters**

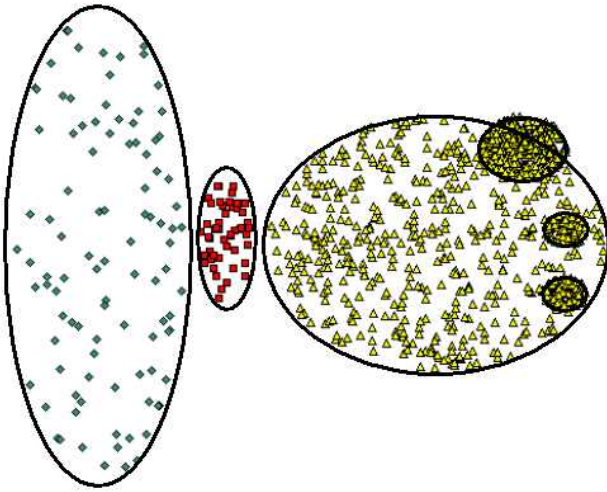
- **Resistant to Noise**
- **Can handle clusters of different shapes and sizes**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

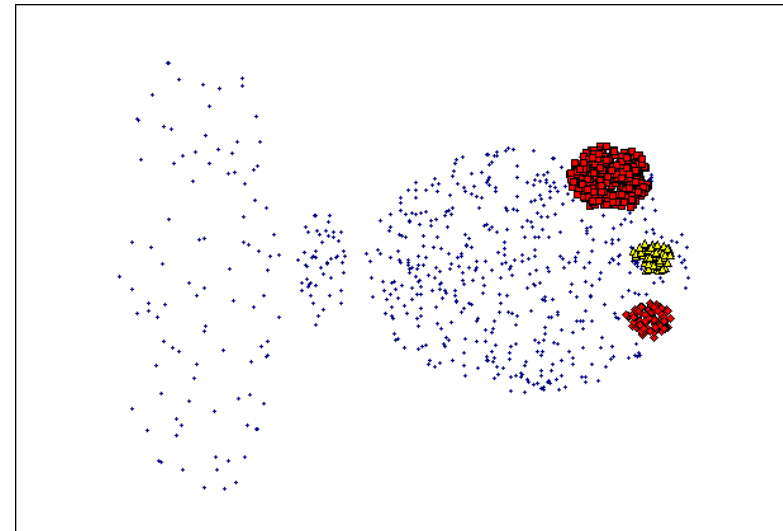
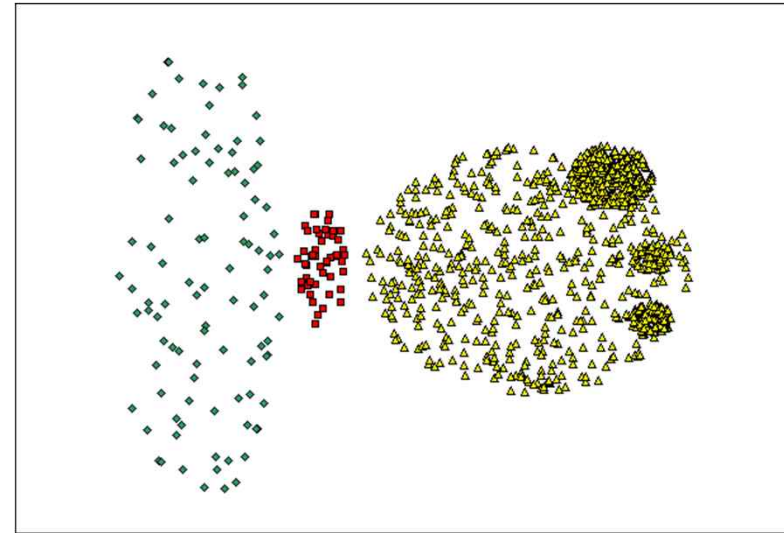
*Han, Kamber - Data Mining: Concepts and Techniques*

# When DBSCAN Does NOT Work Well



**Original Points**

- Varying densities
- Eps = ?, MinPts = ?
- High-dimensional data → Low densities



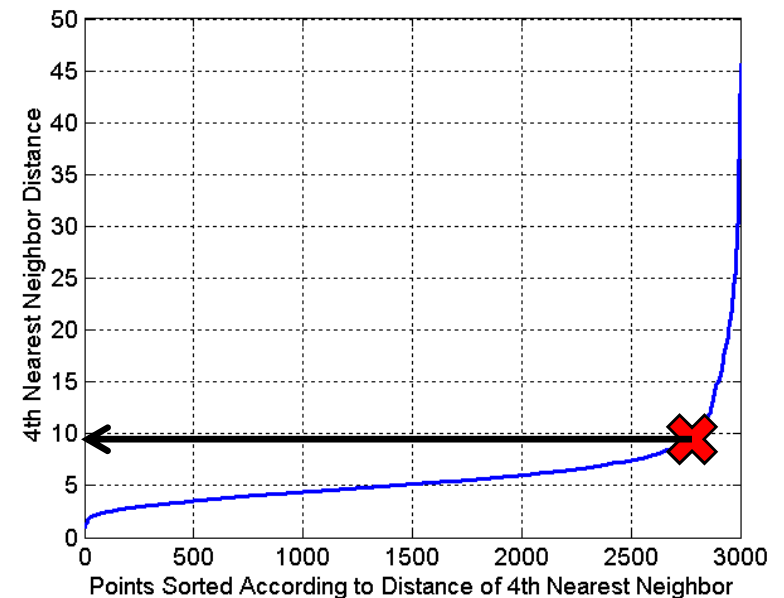
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# DBSCAN: Determining Eps and MinPts

- Idea: Look at the behavior of the distance from a point to its  $k^{th}$  nearest neighbor ( $k$ -dist)
- For points that belong to some cluster, the value of  $k$ -dist will be small since they have many neighboring points.
- For points that are not in a cluster, such as noise points, the  $k$ -dist will be relatively large.
- Thus, compute the  $k$ -dist for all the data points → sort them in increasing order → plot the sorted value → find a sharp change for Eps  
→ take the value of  $k$  as MinPts



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# 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

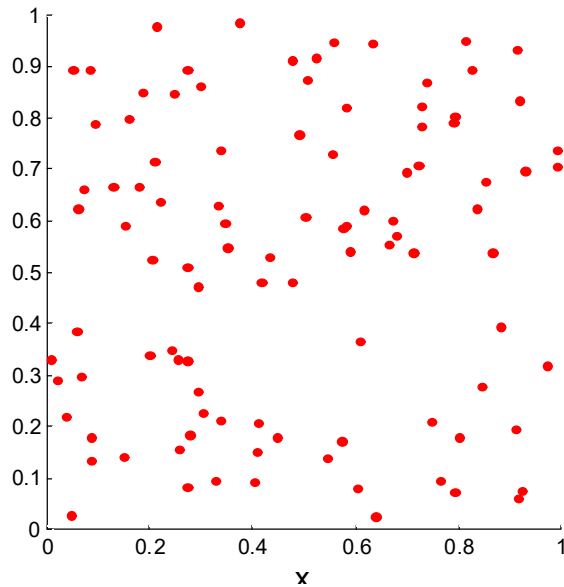
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

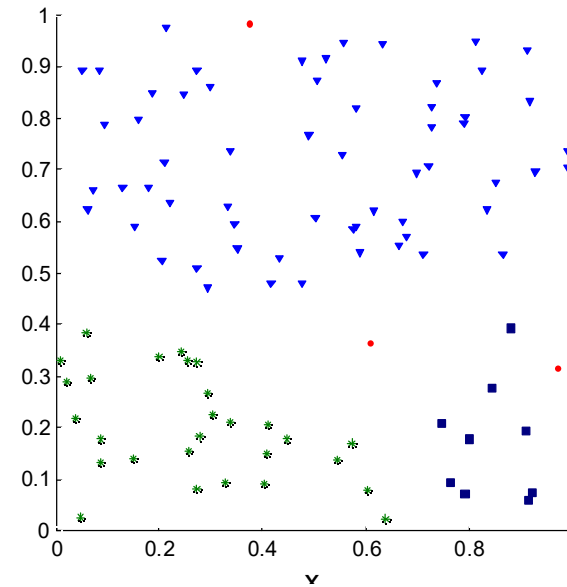
# Clusters found in Random Data

Random  
Points



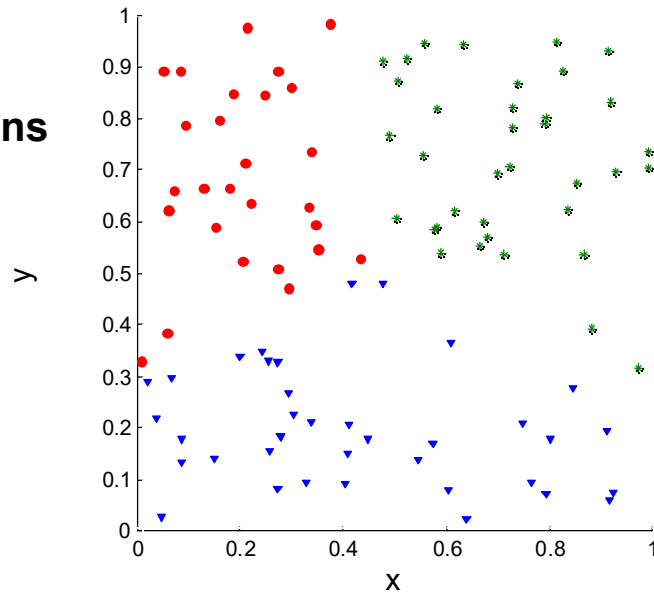
y

DBSCAN



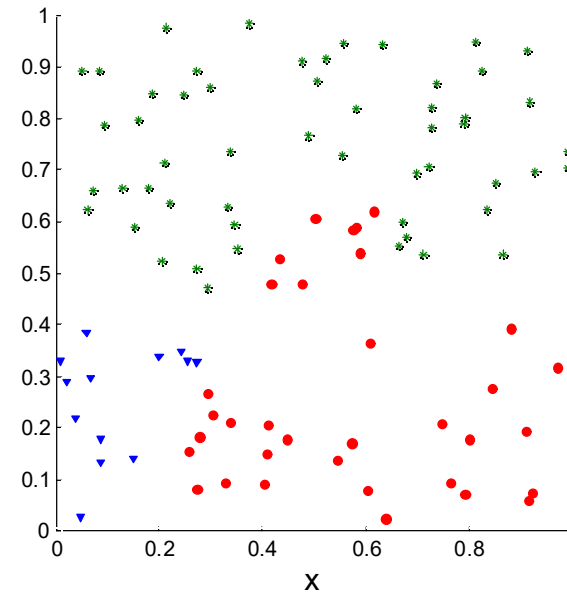
y

K-means



x

Complete  
Link



x

# 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.

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

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

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# External Measures of Cluster Validity: Entropy and Purity

**Table 5.9.** K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

**entropy** For each cluster, the class distribution of the data is calculated first, i.e., for cluster  $j$  we compute  $p_{ij}$ , the ‘probability’ that a member of cluster  $j$  belongs to class  $i$  as follows:  $p_{ij} = m_{ij}/m_j$ , where  $m_j$  is the number of values in cluster  $j$  and  $m_{ij}$  is the number of values of class  $i$  in cluster  $j$ . Then using this class distribution, the entropy of each cluster  $j$  is calculated using the standard formula  $e_j = \sum_{i=1}^L p_{ij} \log_2 p_{ij}$ , where the  $L$  is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e.,  $e = \sum_{j=1}^K \frac{m_j}{m} e_j$ , where  $m_j$  is the size of cluster  $j$ ,  $K$  is the number of clusters, and  $m$  is the total number of data points.

**purity** Using the terminology derived for entropy, the purity of cluster  $j$ , is given by  $purity_j = \max_i p_{ij}$  and the overall purity of a clustering by  $purity = \sum_{j=1}^K \frac{m_j}{m} purity_j$ .

Adapted from:

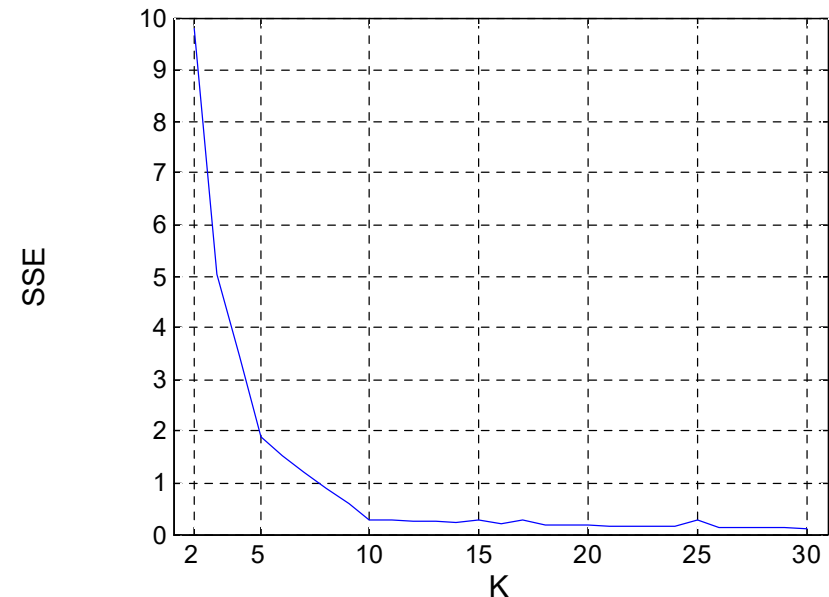
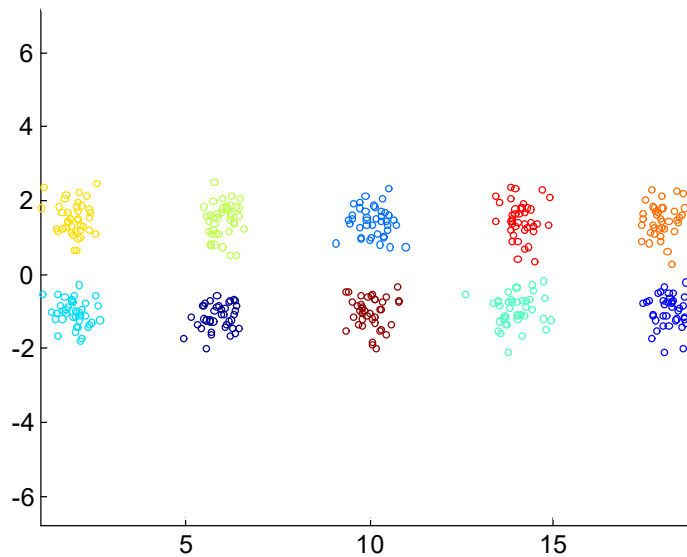
Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

**Small Entropy and Large Purity → Good**

# 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 clusters (average SSE).
- Can also be used to estimate the number of clusters



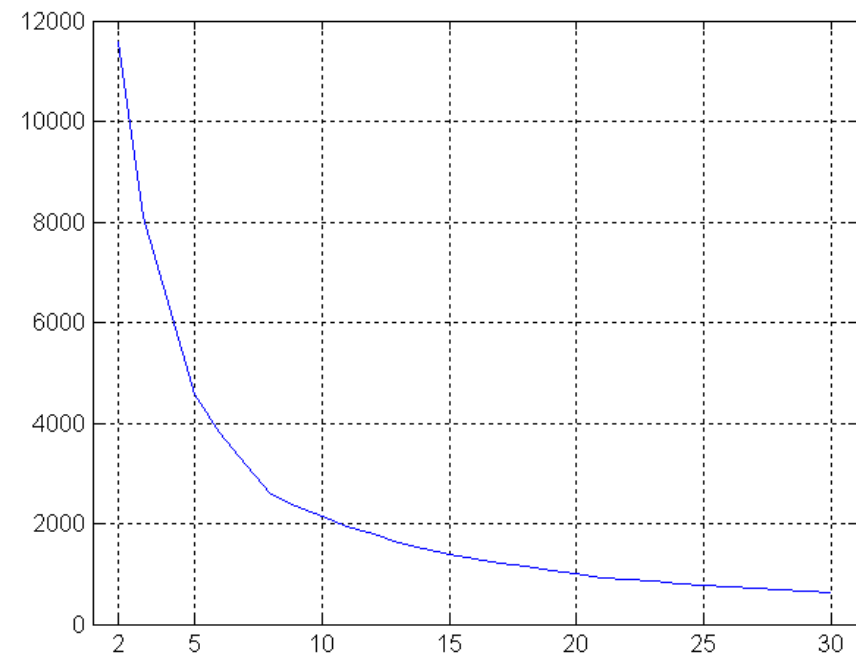
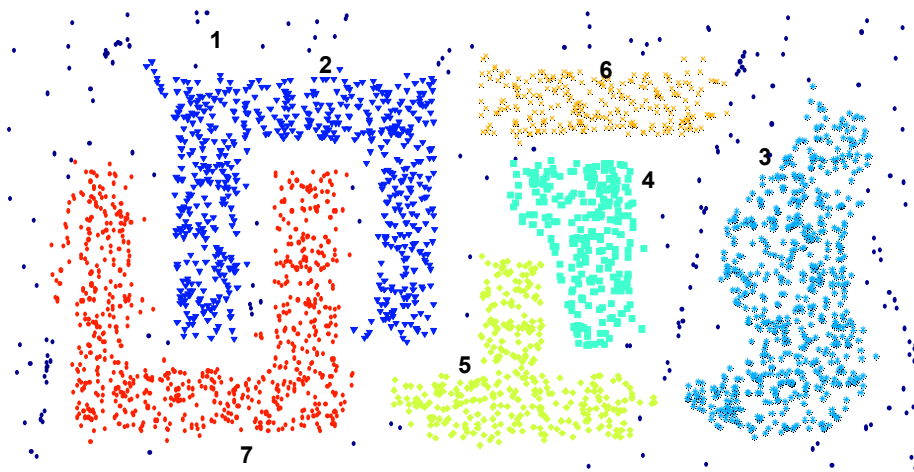
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Internal Measures: SSE

- SSE curve for a more complicated data set



**SSE of clusters found using K-means**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Internal Measures: Cohesion and Separation

- **Cluster Cohesion**: Measures how closely related are objects in a cluster
  - Example: SSE
- **Cluster Separation**: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

*클러스터 중심과 해당 클러스터의 각 데이터 간의 거리*

- Separation is measured by the between cluster sum of squares

$$BSS = \sum |C_i| (m - m_i)^2$$

*m: center point*

» Where  $|C_i|$  is the size of cluster i

**Small SSE and Large BSS → Good**

Adapted from:

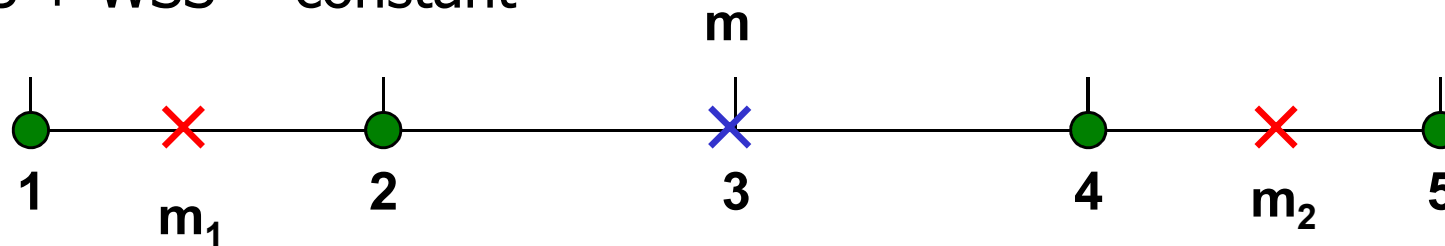
Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

*전체 데이터의 중심과 각 클러스터 중심 간의 거리*

# Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



**K=1 cluster:**

$$WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$BSS = 4 \times (3 - 3)^2 = 0$$

$$Total = 10 + 0 = 10$$

**K=2 clusters:**

$$WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

$$BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9$$

$$Total = 1 + 9 = 10$$

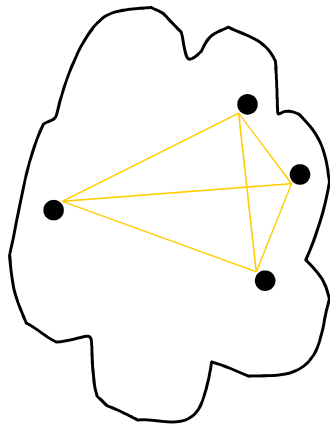
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

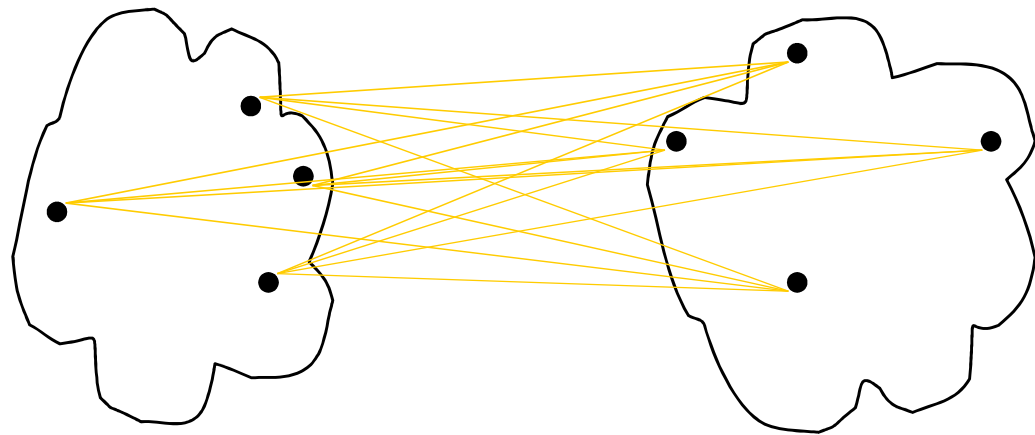
*Han, Kamber - Data Mining: Concepts and Techniques*

# Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



separation

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

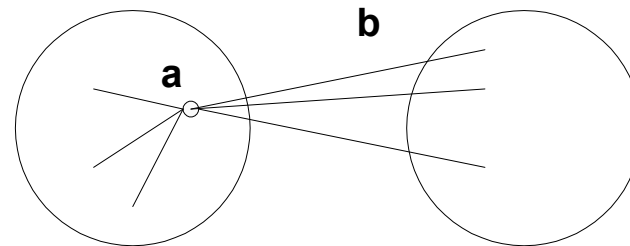
*Han, Kamber - Data Mining: Concepts and Techniques*

# Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters
- For an individual point,  $i$ 
  - Calculate  $a$  = average distance of  $i$  to the points in its cluster
  - Calculate  $b$  = min (average distance of  $i$  to points in another cluster)
  - The silhouette coefficient for a point is then given by

$$s = 1 - a/b \quad \text{if } a < b, \quad (\text{or } s = b/a - 1 \quad \text{if } a \geq b, \text{ not the usual case})$$

- Typically between 0 and 1.
- The closer to 1 the better.



*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Measuring Cluster Validity Via Correlation

- Two matrices:
  - Proximity (similarity) Matrix
  - “Incidence” Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
- High correlation indicates that points that belong to the same cluster are close to each other.

*Adapted from:*

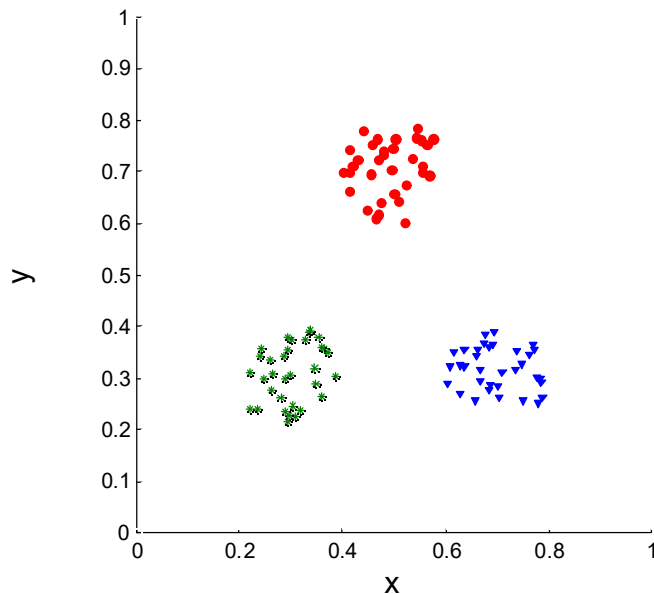
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

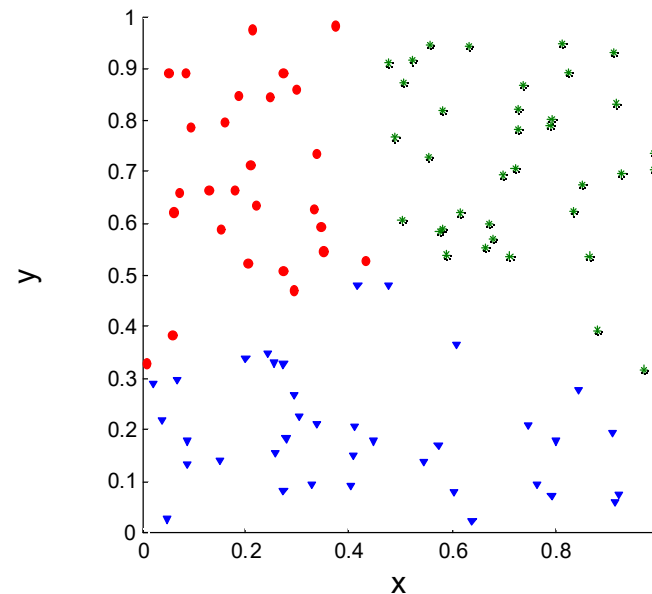


# Measuring Cluster Validity Via Correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



**Corr = -0.9235**



**Corr = -0.5810**

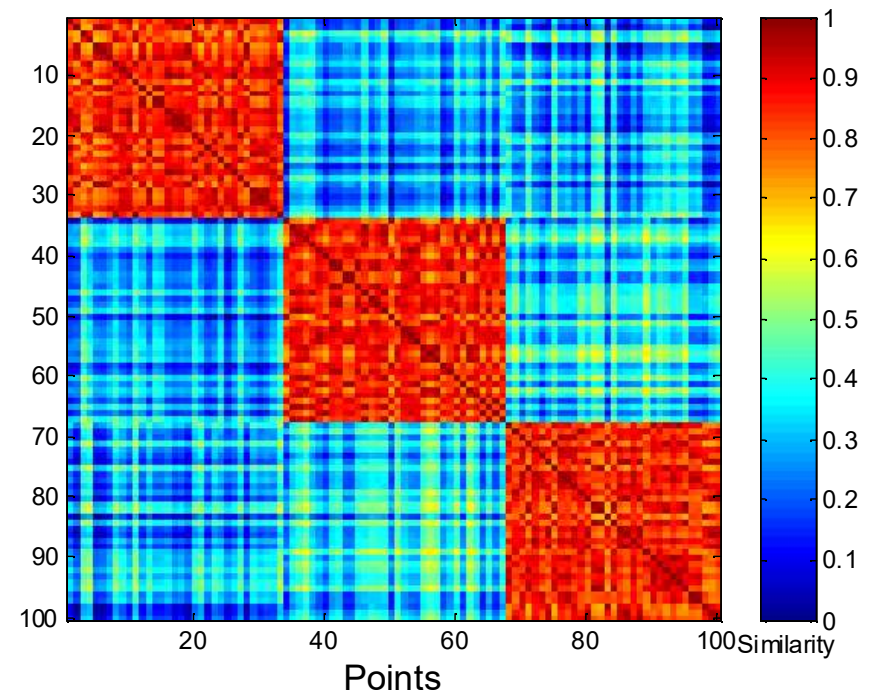
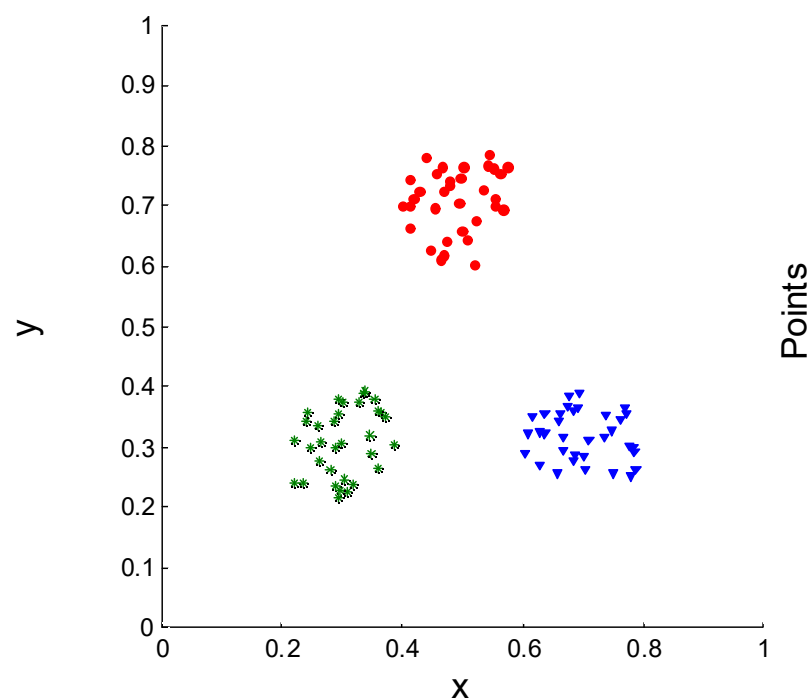
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Using Similarity Matrix for Cluster Validation

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



*Similarity matrix sorted by cluster label*

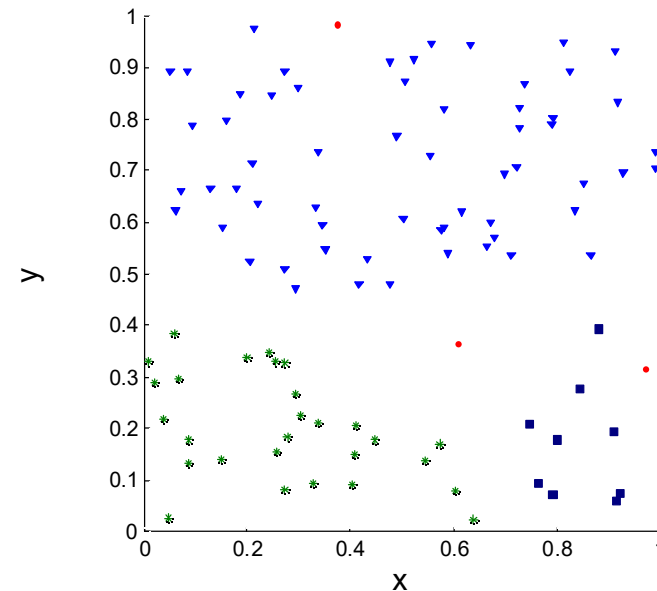
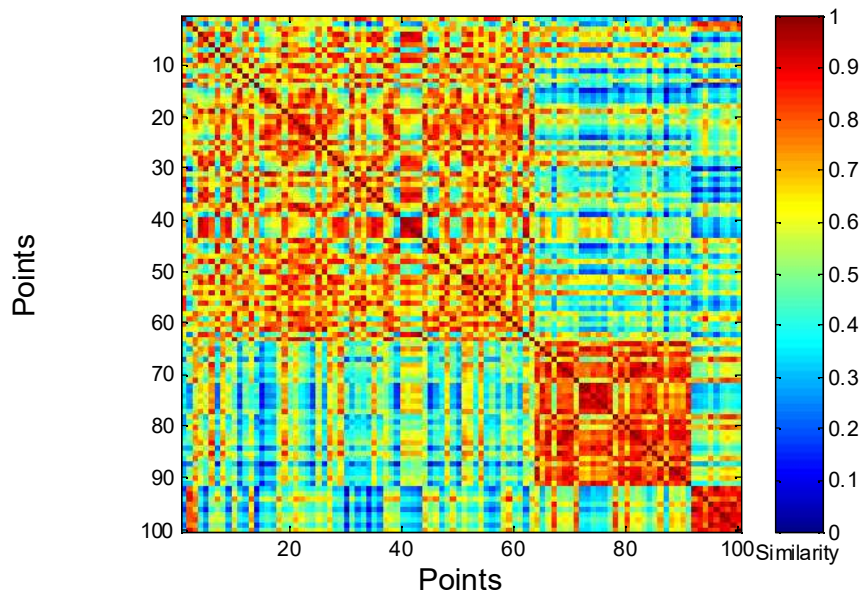
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp



**DBSCAN**

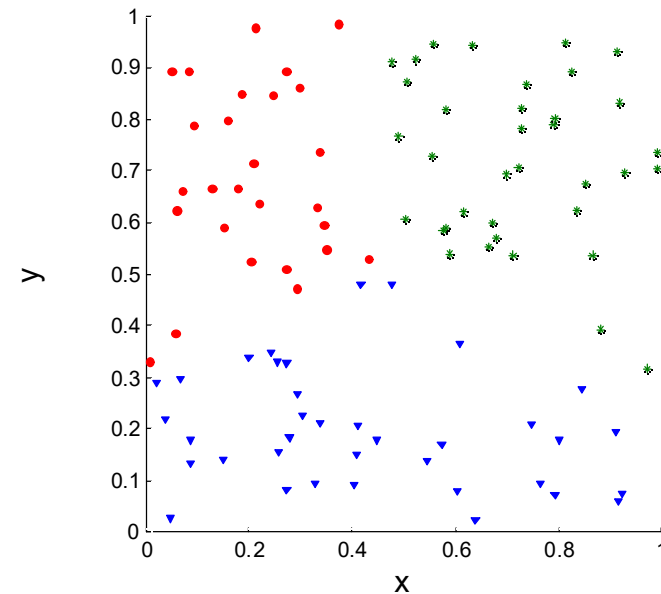
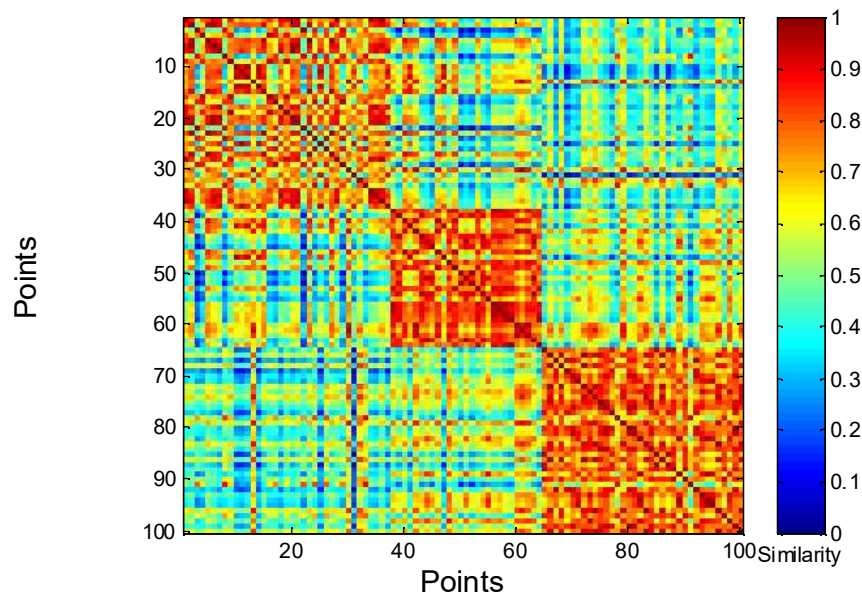
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp



**K-means**

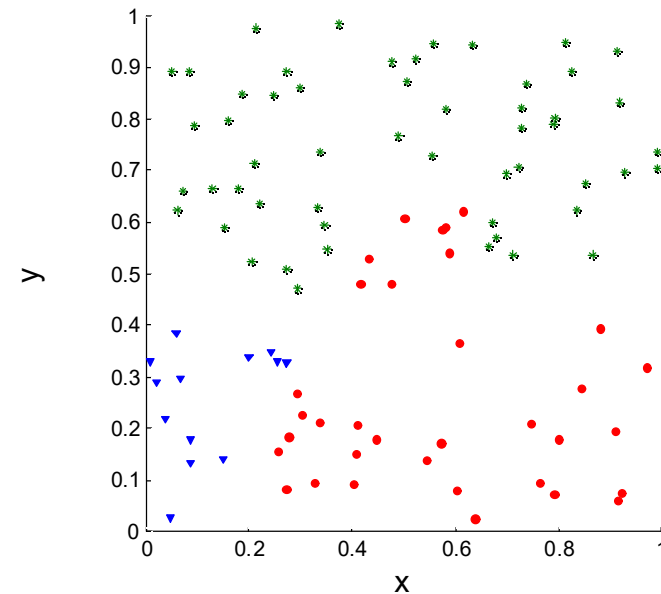
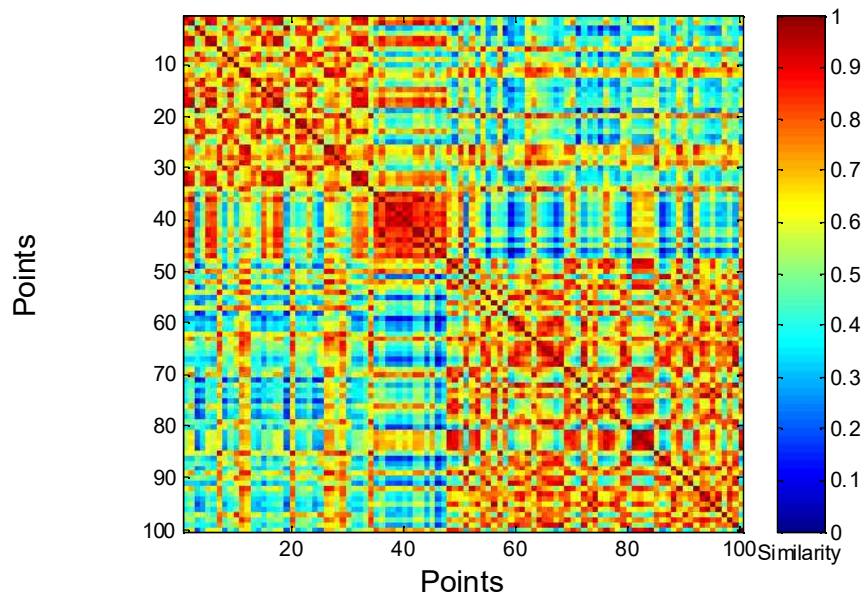
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp



**Complete Link**

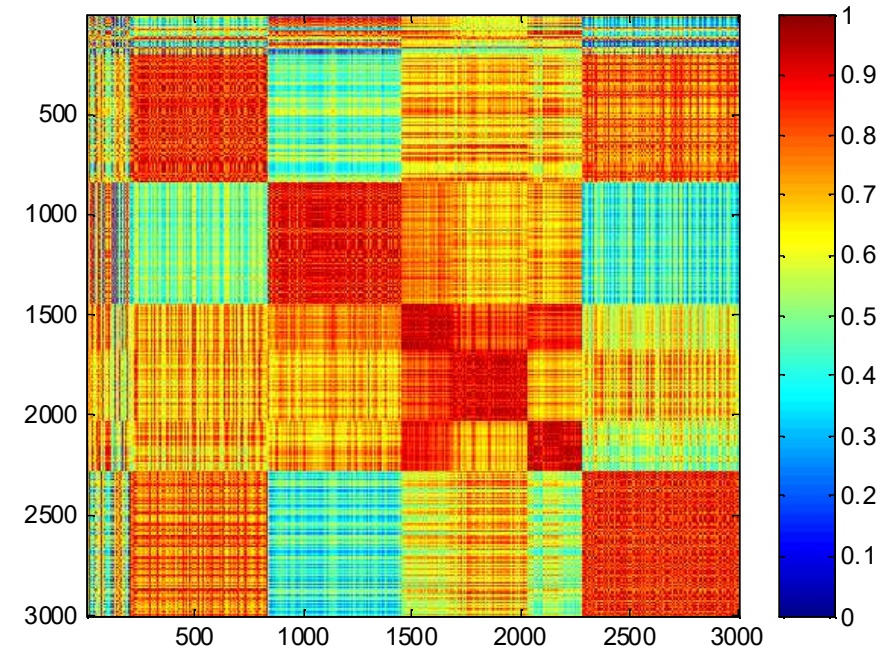
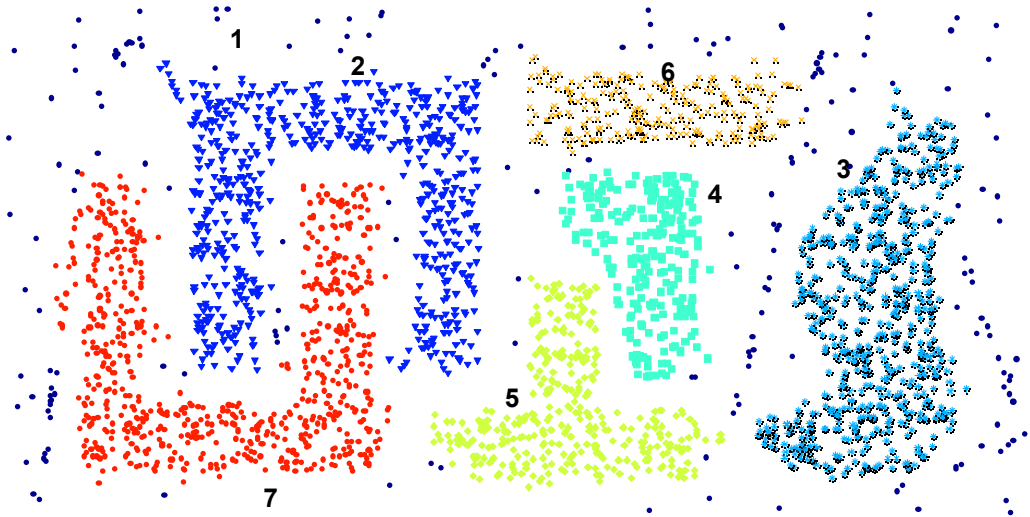
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# Using Similarity Matrix for Cluster Validation



**DBSCAN**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# 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

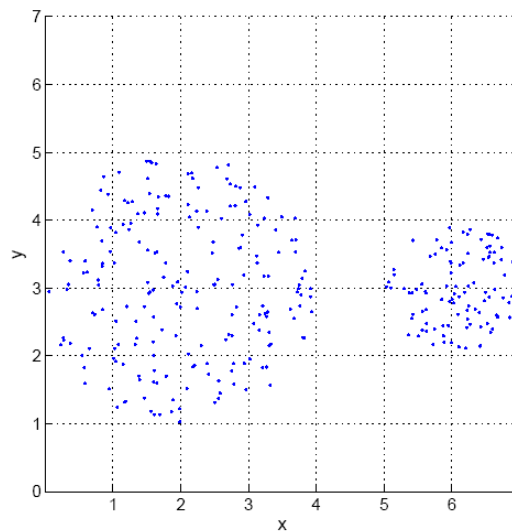
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Grid-Based Clustering

- Efficient way to organize a set of data in low dimensions
- Algorithm:
  1. Define a set of grid cells
  2. Assign objects to the appropriate cells and compute the density of each cell
  3. Eliminate cells having a density below a specified threshold
  4. Form clusters from contiguous (adjacent) groups of dense cells



**Figure 7.13.** Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

**Table 7.6.** Point counts for each grid cell.

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# Graph-Based Clustering

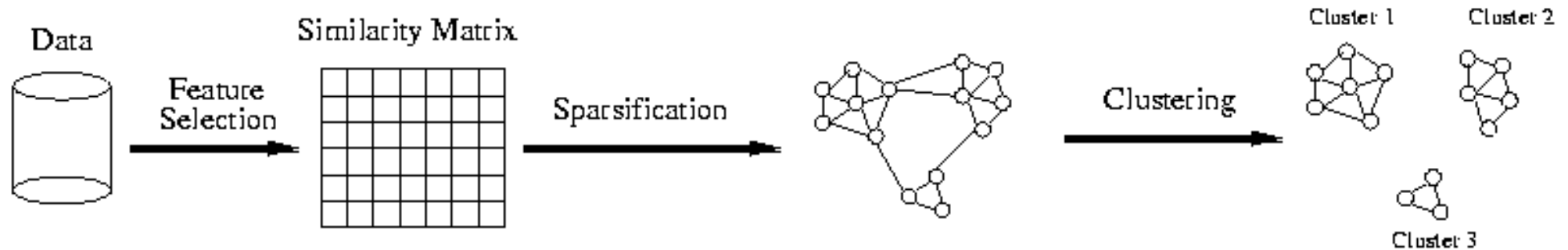
- Graph-Based clustering uses the proximity graph
  - Start with the proximity matrix
  - Consider each point as a node in a graph
  - Each edge between two nodes has a weight which is the proximity between the two points
  - Initially the proximity graph is fully connected
- In the simplest case, clusters are connected components in the graph.

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Sparsification in the Clustering Process



*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Graph-Based Clustering: Sparsification

- The amount of data that needs to be processed is drastically reduced
  - Setting many of low similarity values to 0 (break the link)
  - Break all links that have a similarity below a specified threshold or keep only links to the k nearest neighbors of point
  - Sparsification can eliminate more than 99% of the entries in a proximity matrix
  - The amount of time required to cluster the data is drastically reduced
  - The size of the problems that can be handled is increased

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Graph-Based Clustering: Sparsification

- Clustering may work better
  - Sparsification techniques keep the connections to the most similar (nearest) neighbors of a point while breaking the connections to less similar points.
  - The nearest neighbors of a point tend to belong to the same class as the point itself.
  - This reduces the impact of noise and outliers and sharpens the distinction between clusters.
- Sparsification facilitates the use of graph partitioning algorithms (or algorithms based on graph partitioning algorithms).
  - Chameleon and Hypergraph-based Clustering

*Adapted from:*

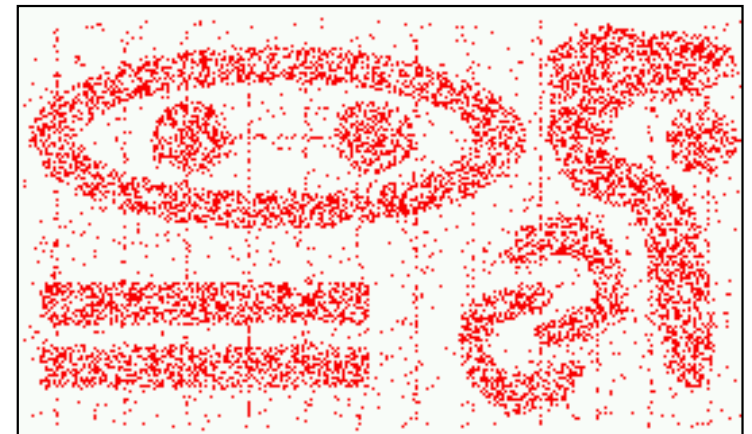
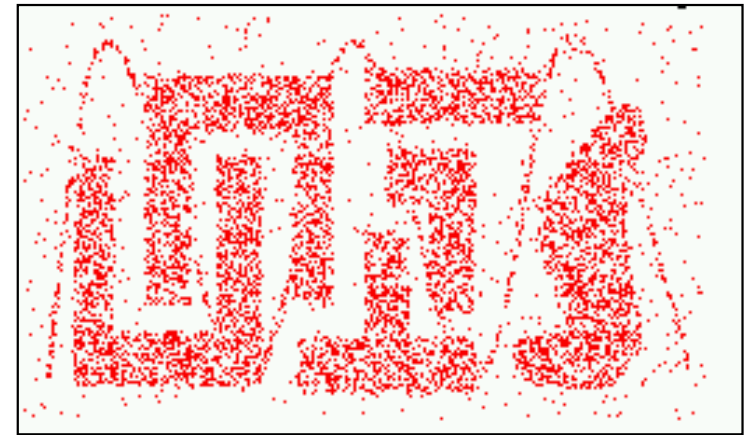
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Characteristics of Spatial Data Sets

- **Clusters are defined as densely populated regions of the space**
- **Clusters have arbitrary shapes, orientation, and non-uniform sizes**
- **Difference in densities across clusters and variation in density within clusters**
- **Existence of special artifacts (*streaks*) and noise**

**The clustering algorithm must address the above characteristics and also require minimal supervision.**



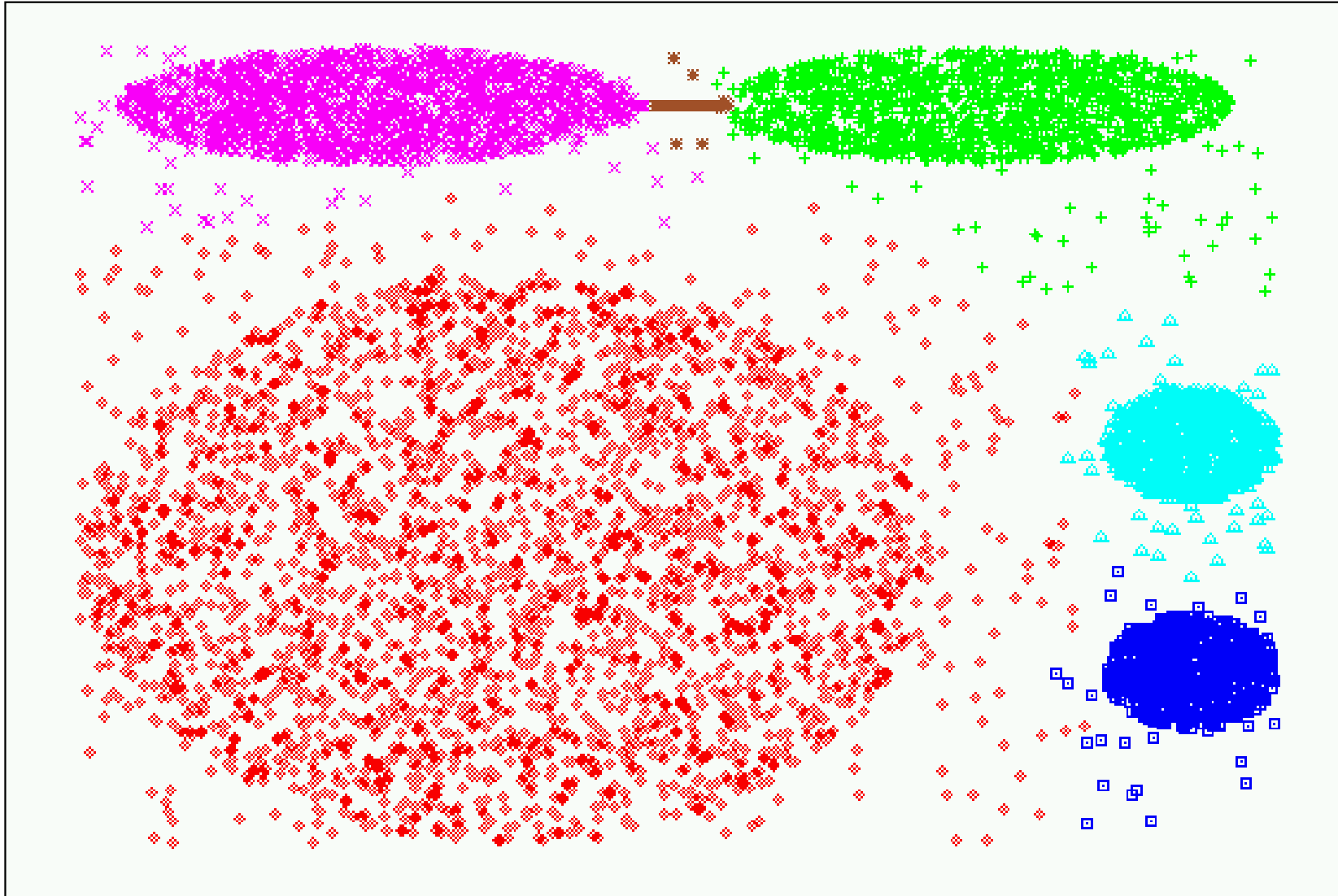
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Experimental Results: CHAMELEON

*\*An agglomerative clustering algorithm with dynamic modeling using an efficient graph partitioning algorithm (considers density, shapes, closeness, interconnectivity, etc.)*

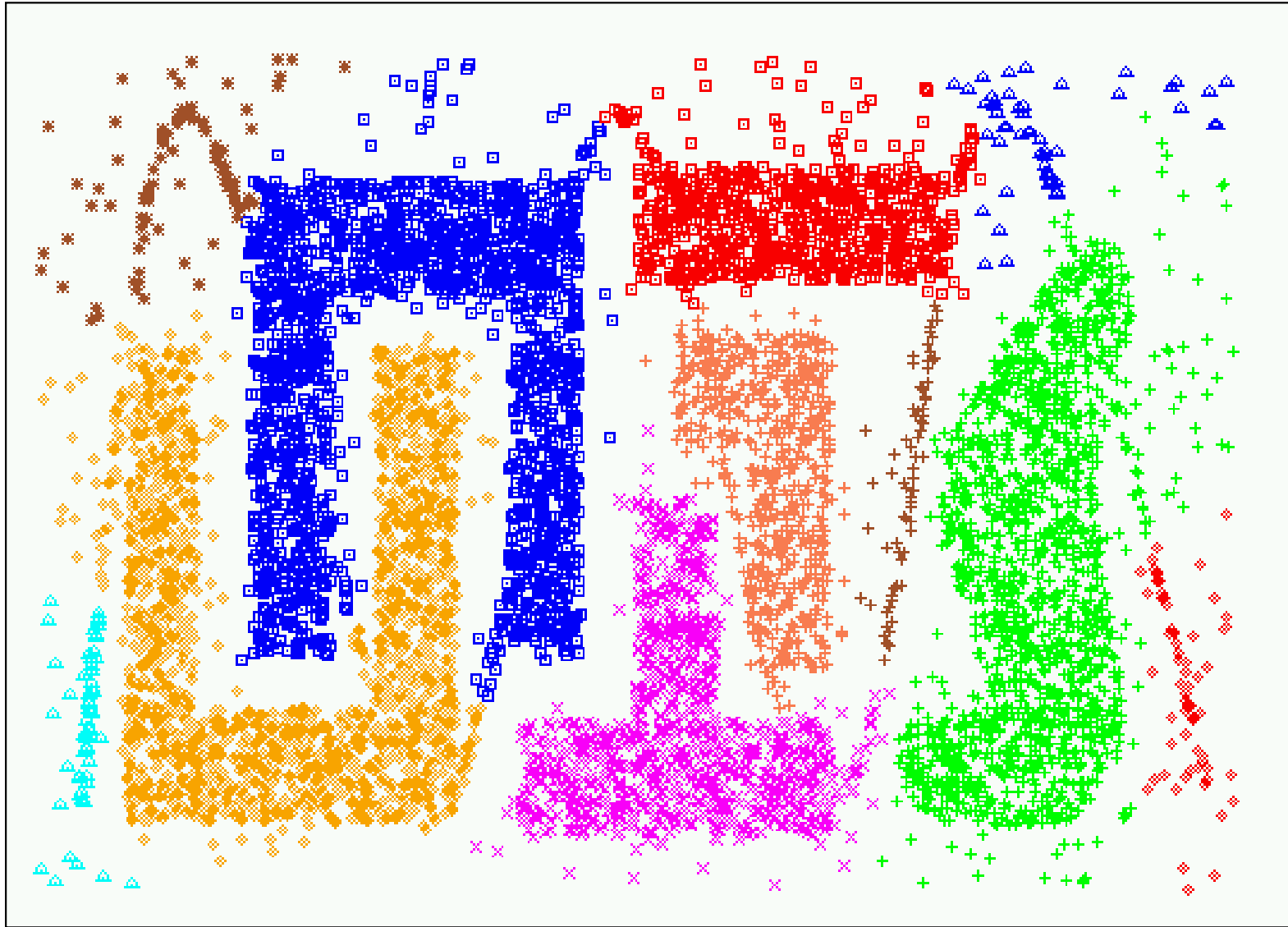


Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# Experimental Results: CHAMELEON



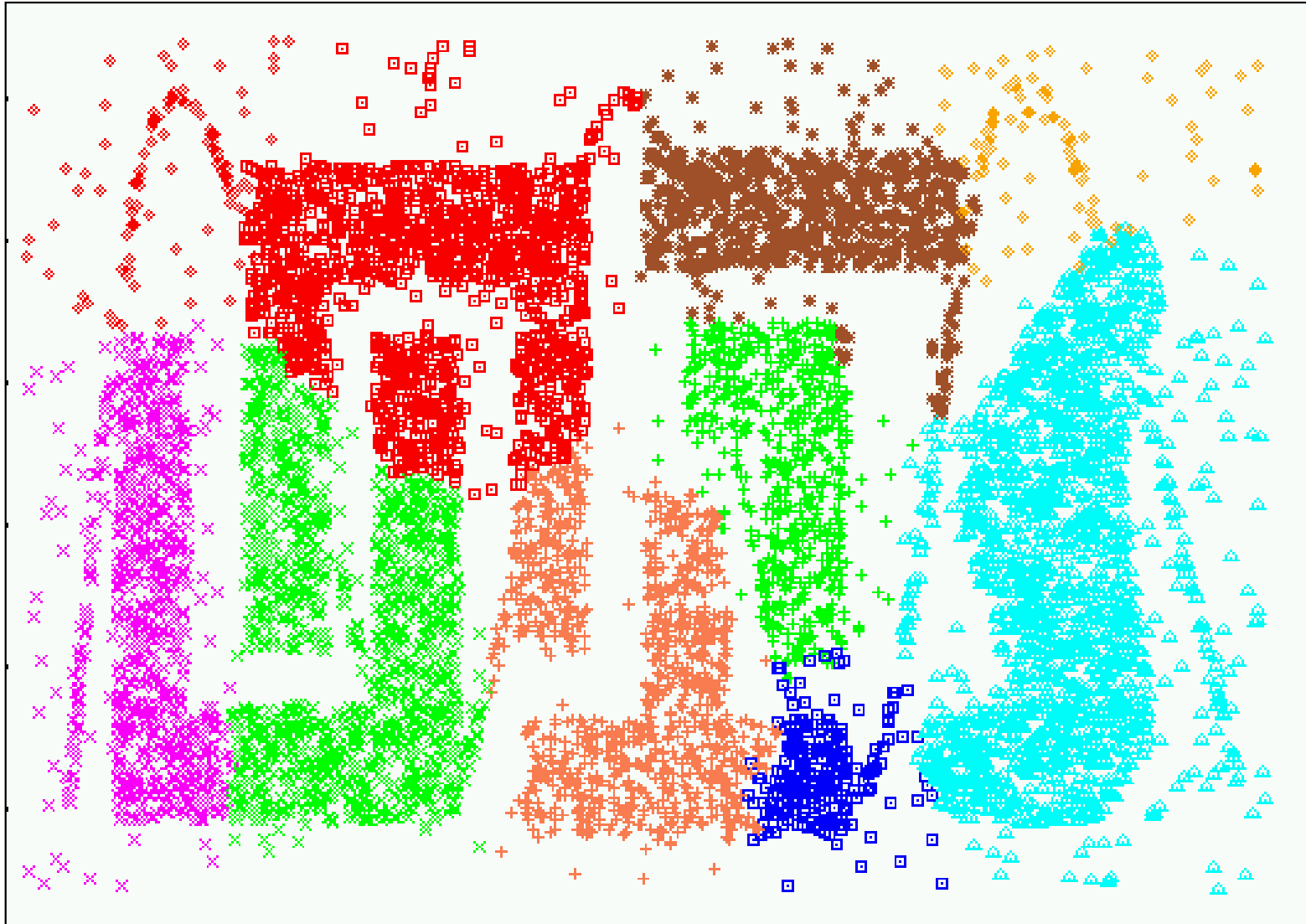
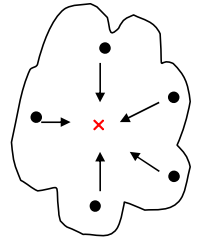
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Experimental Results: CURE ( *10 clusters* )

*\*Clustering Using REpresentatives: (1) equally partitions, (2) uses a constant number of points to represent a cluster which capture the geometry and shape of the cluster and (3) then shrinks them toward the center of the cluster by a factor, alpha*



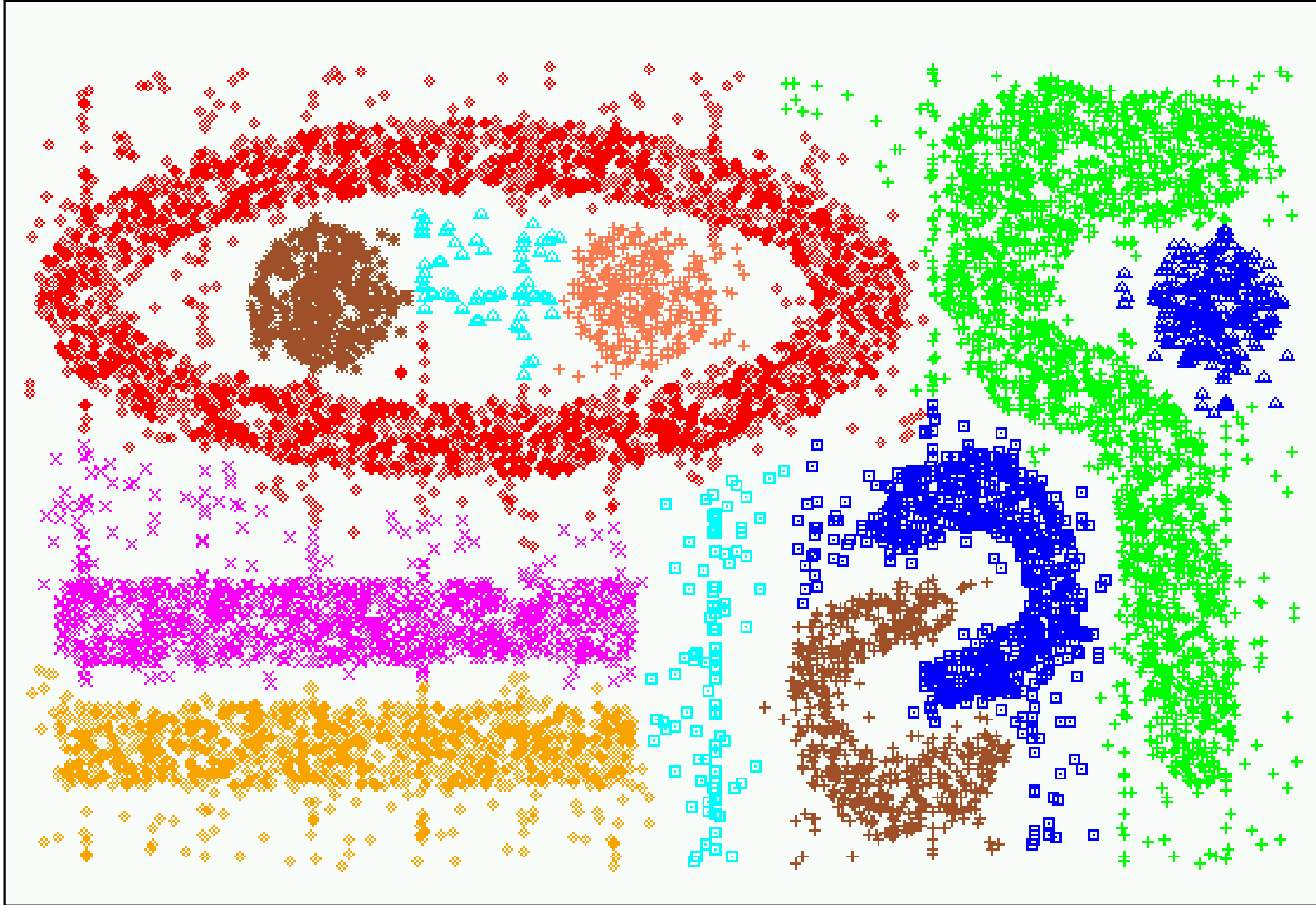
Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques



# Experimental Results: CHAMELEON

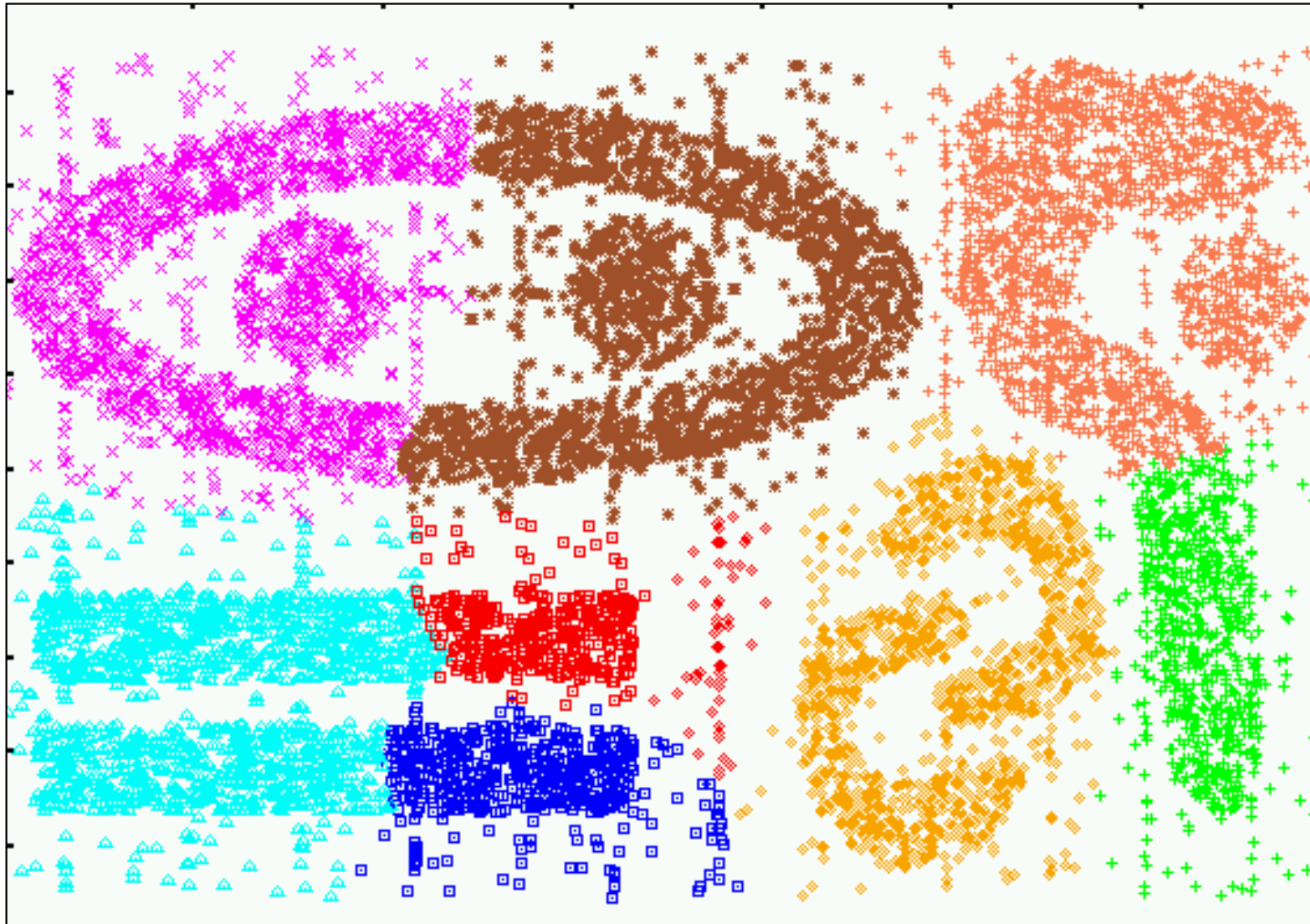


*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# Experimental Results: CURE (*9 clusters*)



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

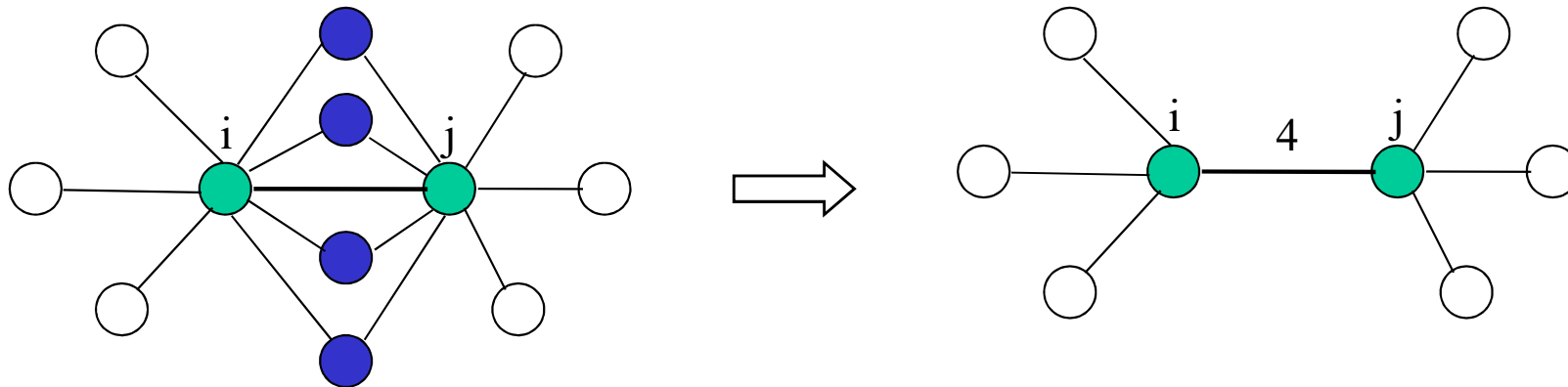
Han, Kamber - Data Mining: Concepts and Techniques

# Shared Nearest Neighbor (SNN) Similarity

If two points are similar to many of the same points, then they are similar to one another, even if a direct measurement of similarity does not indicate this.

→ Similarity is “confirmed” by common near neighbors

**SNN graph**: the weight of an edge is the number of shared neighbors between vertices given that the vertices are connected



Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

Han, Kamber - Data Mining: Concepts and Techniques

# SNN Clustering Algorithm

## 1. Compute the similarity matrix

This corresponds to a similarity graph with data points for nodes and edges whose weights are the similarities between data points

## 2. Sparsify the similarity matrix by keeping only the $k$ most similar neighbors

This corresponds to only keeping the  $k$  strongest links of the similarity graph

## 3. Construct the shared nearest neighbor graph from the sparsified similarity matrix.

At this point, we could apply a similarity threshold and find the connected components to obtain the clusters (Jarvis-Patrick algorithm)

## 4. Find the SNN density of each Point.

Using a user specified parameters,  $Eps$ , find the number points that have an SNN similarity of  $Eps$  or greater to each point. This is the SNN density of the point

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

(Note that steps 4-8 are DBSCAN)

# SNN Clustering Algorithm ...

## 5. Find the core points

Using a user specified parameter, *MinPts*, find the core points, i.e., all points that have an SNN density greater than *MinPts*

## 6. Form clusters from the core points

If two core points are within a radius, *Eps*, of each other they are place in the same cluster

## 7. Discard all noise points

All non-core points that are not within a radius of *Eps* of a core point are discarded

## 8. Assign all non-noise, non-core points to clusters

This can be done by assigning such points to the nearest core point

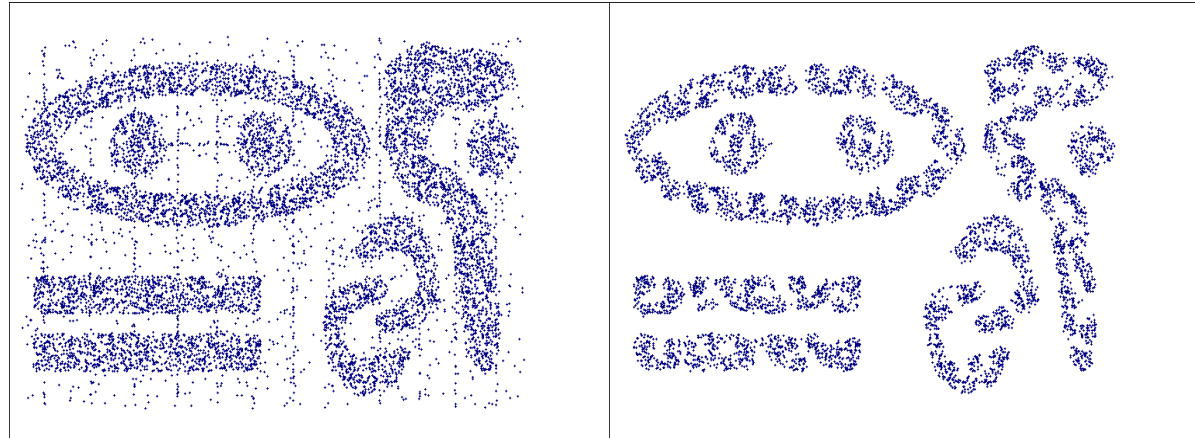
(Note that steps 4-8 are DBSCAN)

Adapted from:

Tan, Steinbach, Kumar - Introduction to Data Mining

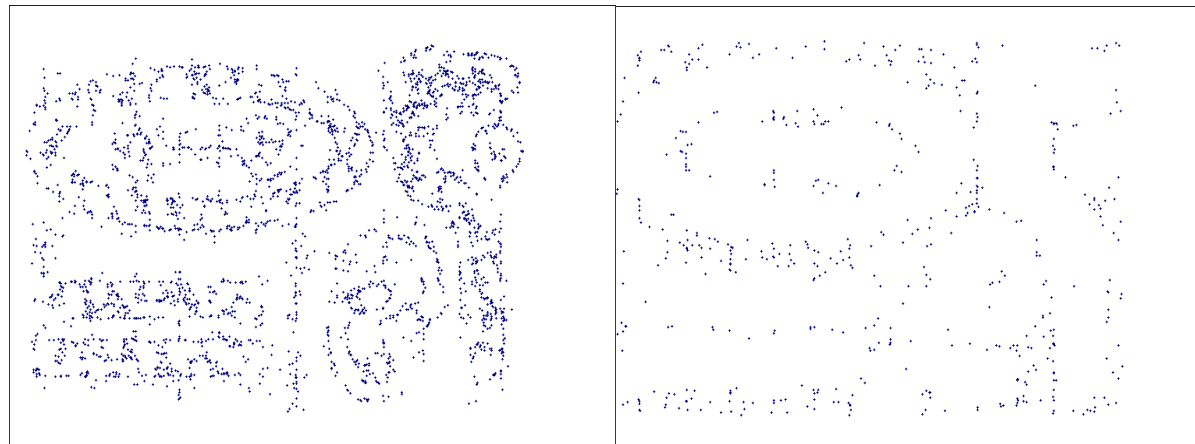
Han, Kamber - Data Mining: Concepts and Techniques

# SNN Density



**a) All Points**

**b) High SNN Density → more core**  
(High connectivity in the SNN graph)



**c) Medium SNN Density**

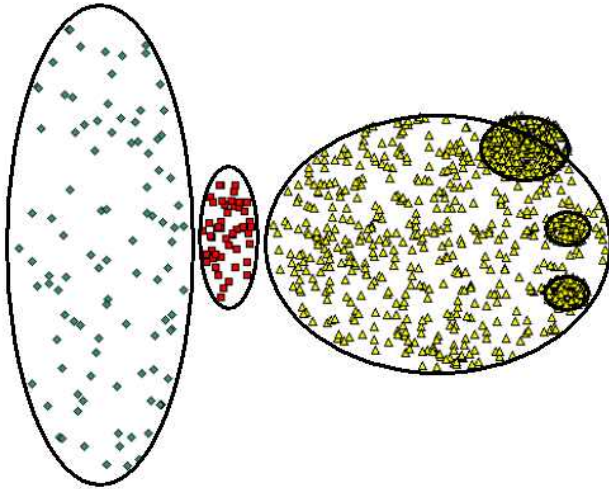
**d) Low SNN Density → more noise / outlier**

*Adapted from:*

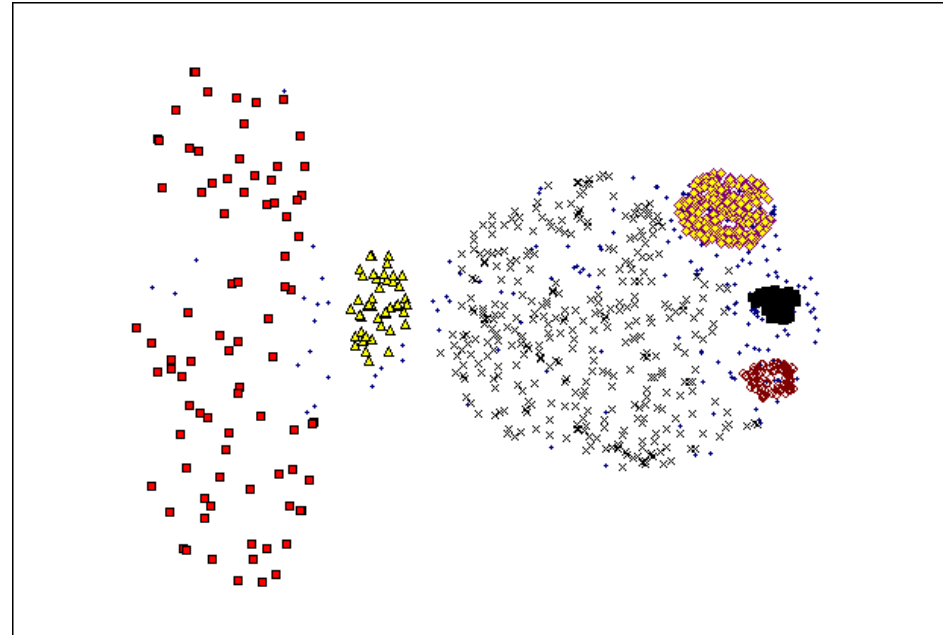
*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*

# SNN Clustering Can Handle Differing Densities



**Original Points**



**SNN Clustering**

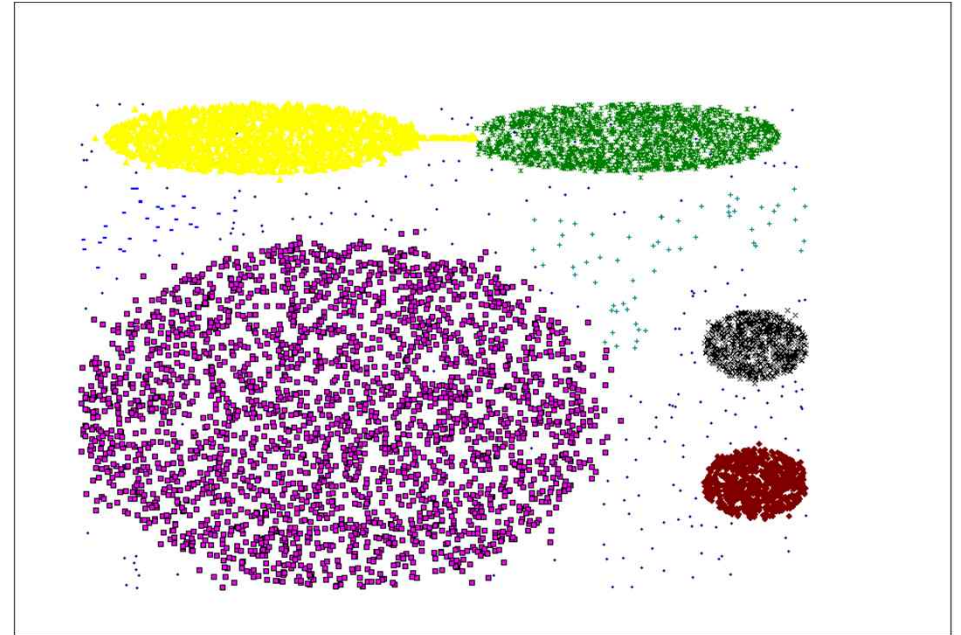
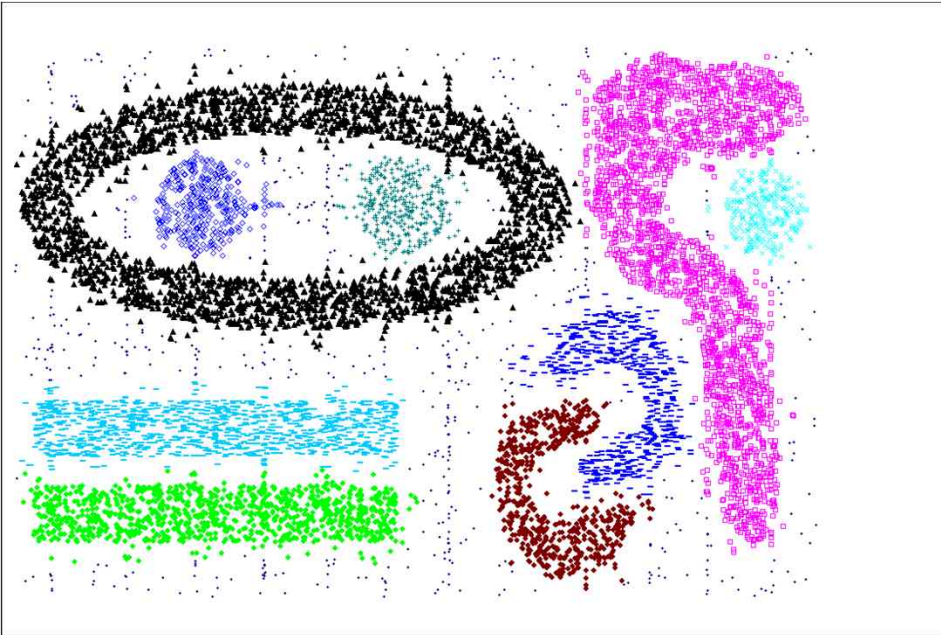
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*



# SNN Clustering Can Handle Other Difficult Situations



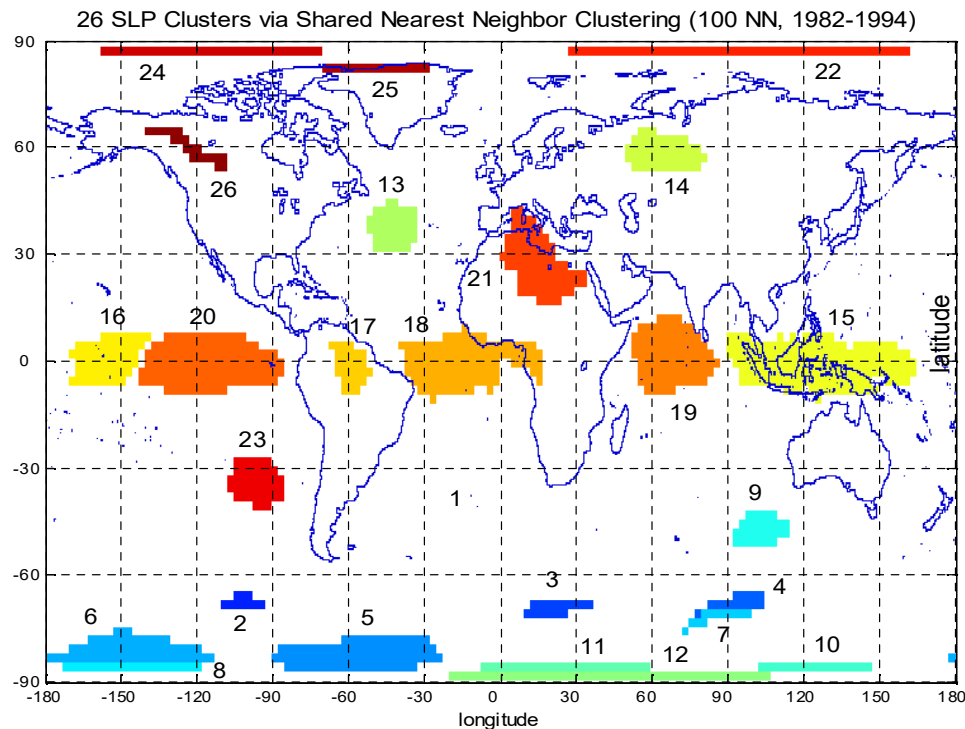
*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

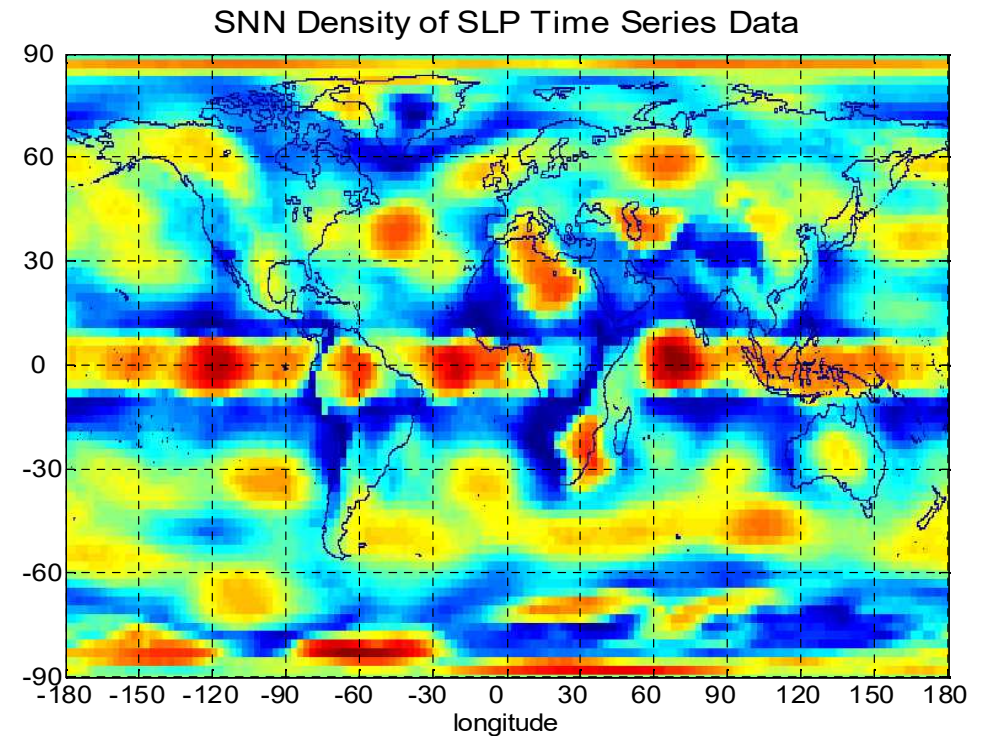
*Han, Kamber - Data Mining: Concepts and Techniques*



# Finding Clusters of Time Series In Spatio-Temporal Data



**SNN Clusters of Pressure**



**SNN Density of Pressure**

*Adapted from:*

*Tan, Steinbach, Kumar - Introduction to Data Mining*

*Han, Kamber - Data Mining: Concepts and Techniques*