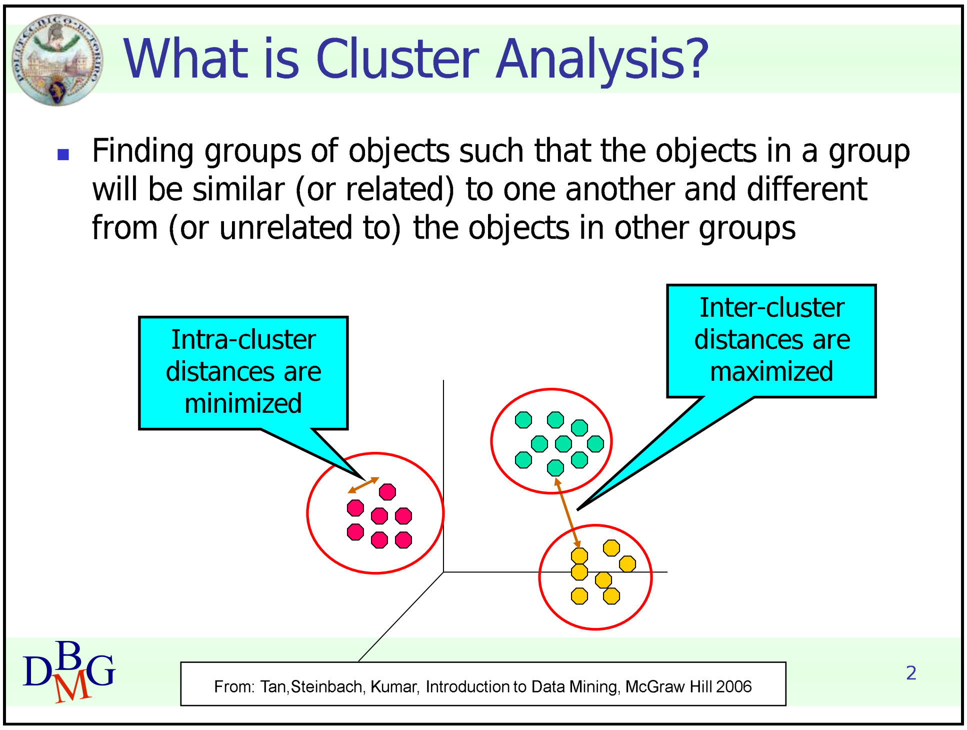
Clustering fundamentals

Elena Baralis, Tania Cerquitelli

Politecnico di Torino

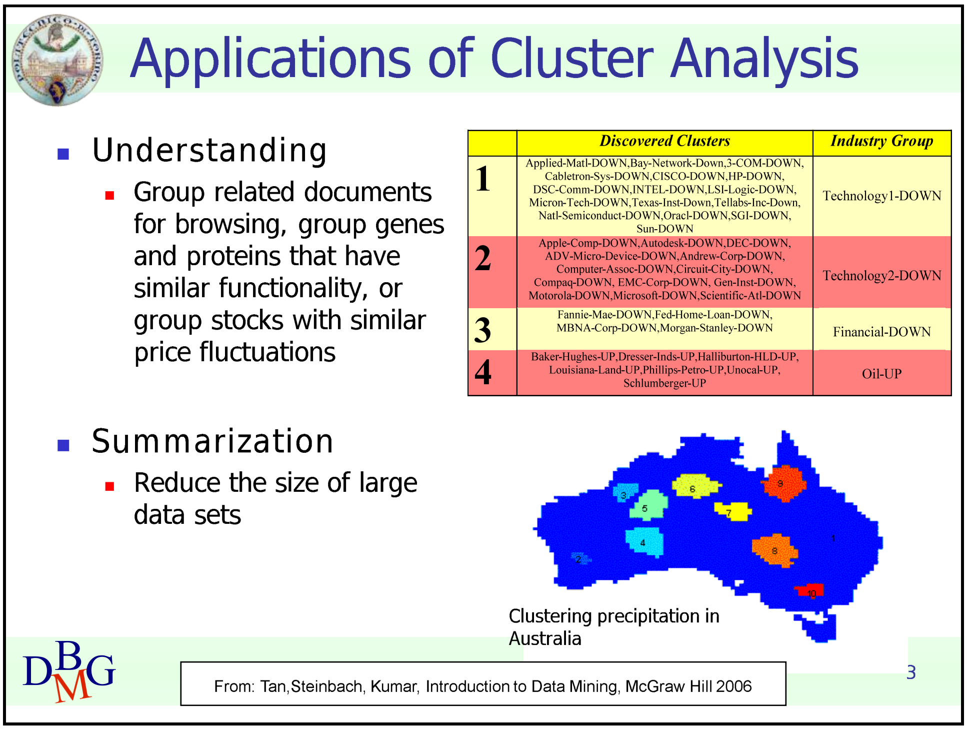


What is Cluster Analysis?

- 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

Intra-cluster distances are minimized
Inter-cluster distances are maximized

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Applications of Cluster Analysis

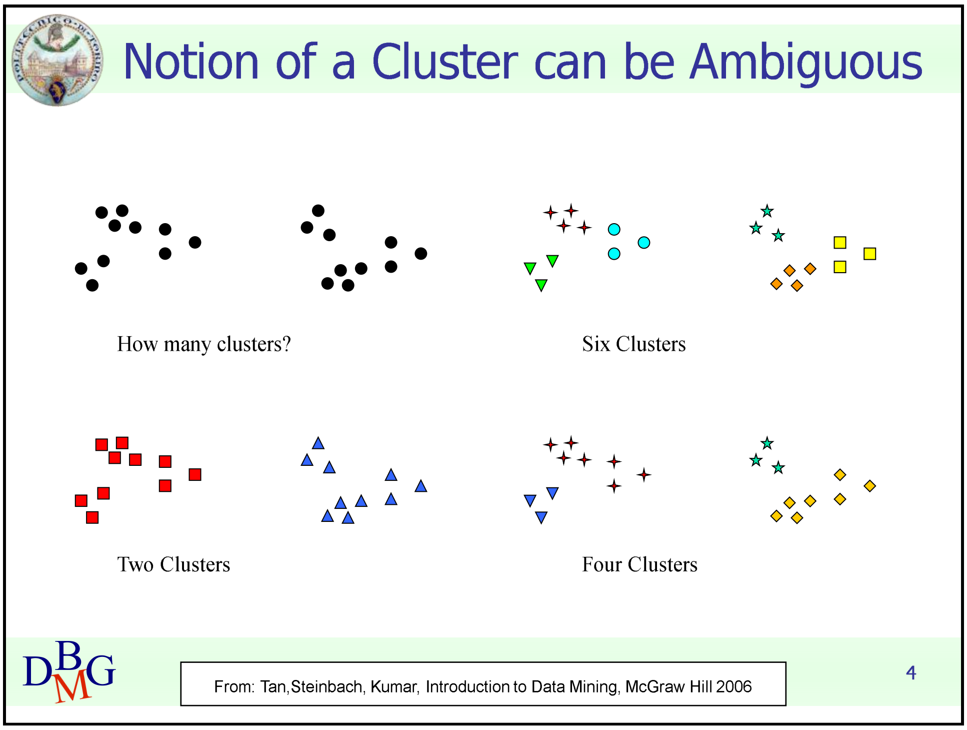
- Understanding
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations
- Summarization
  - Reduce the size of large data sets

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Clustering precipitation in Australia

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

3



Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

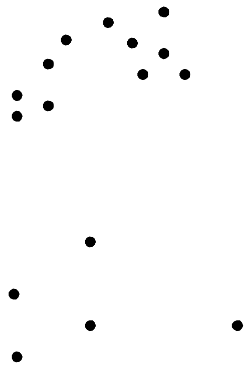
4

Types of Clusterings

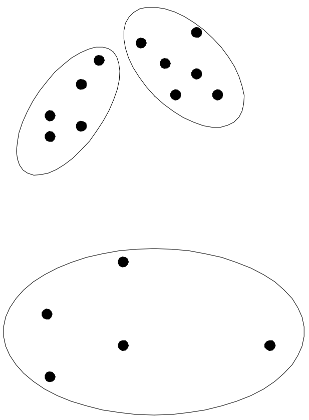
- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Partitional Clustering

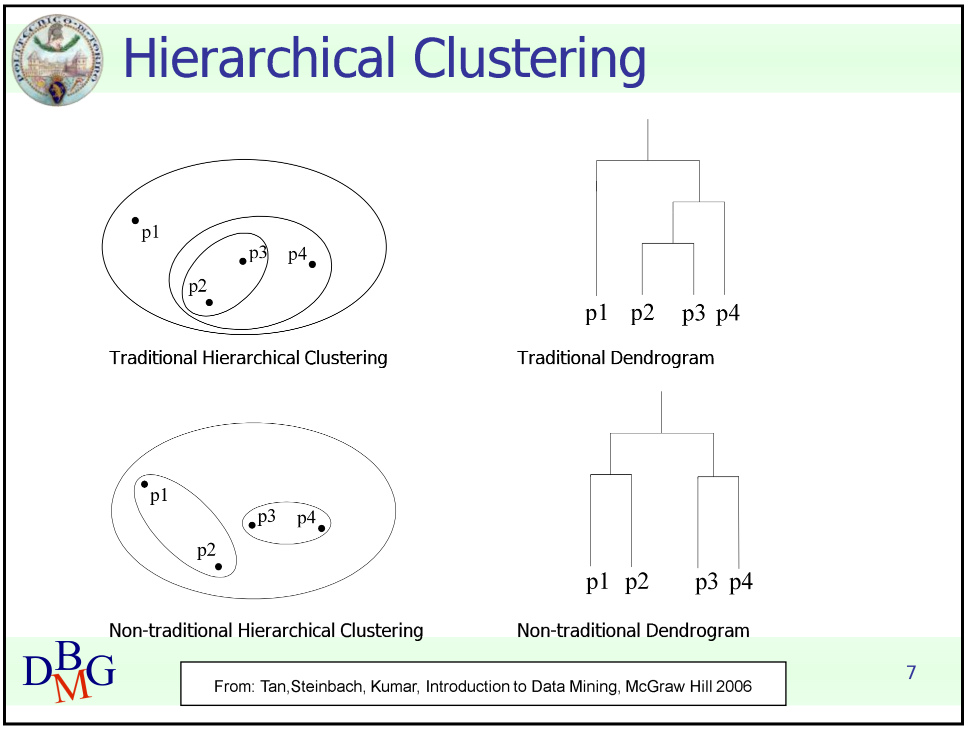


Original Points



A Partitional Clustering

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Hierarchical Clustering

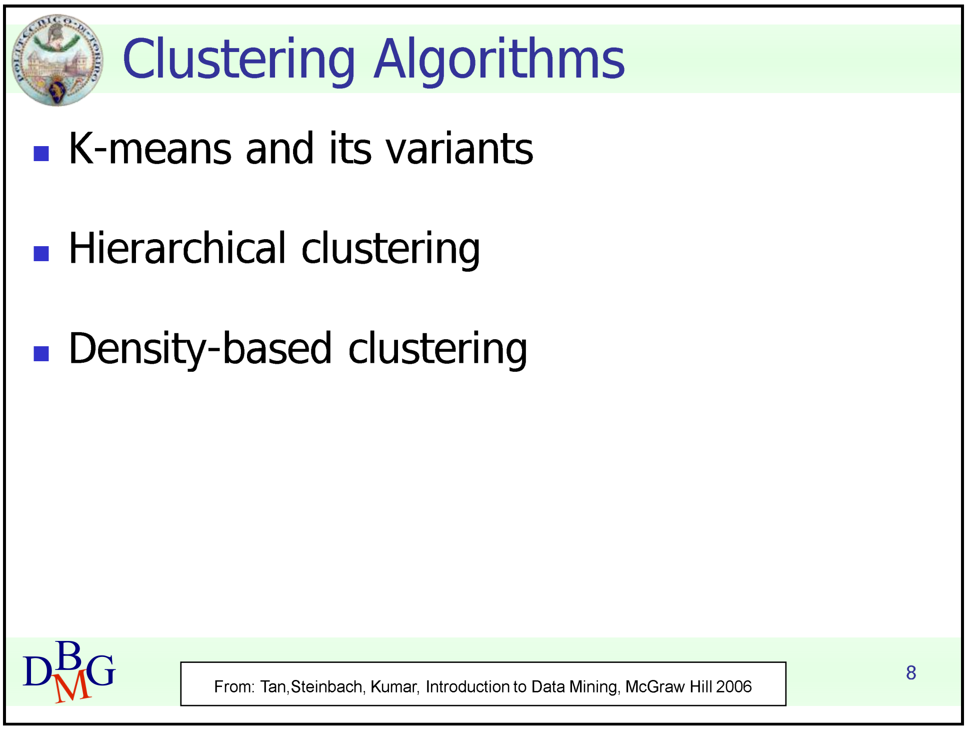
Traditional Hierarchical Clustering

Traditional Dendrogram

Non-traditional Hierarchical Clustering

Non-traditional Dendrogram

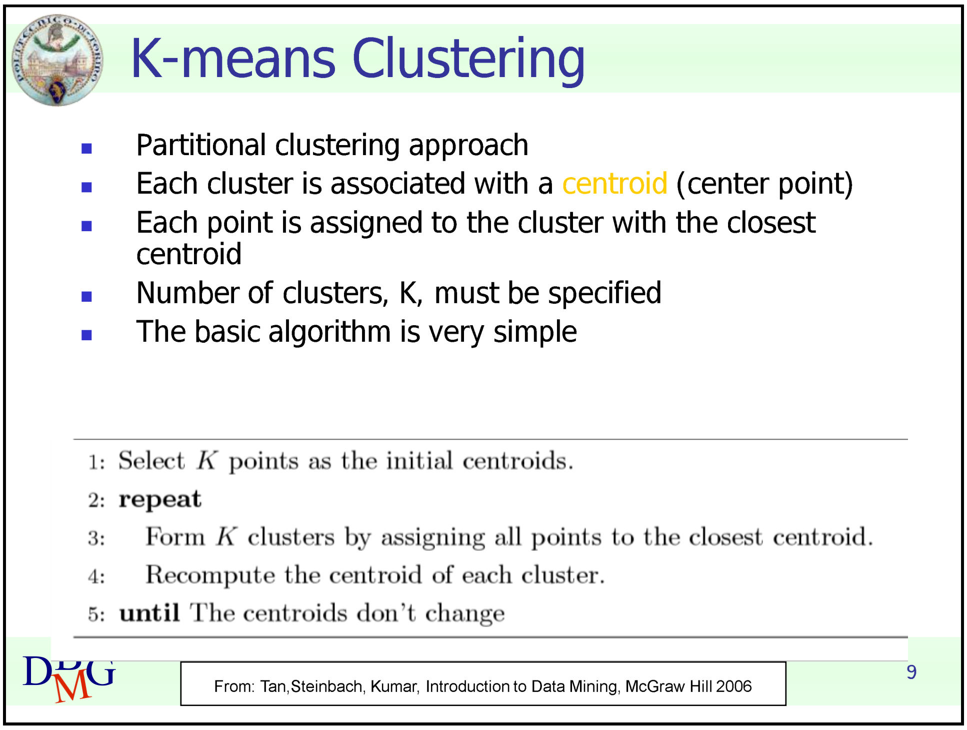
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Clustering Algorithms

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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

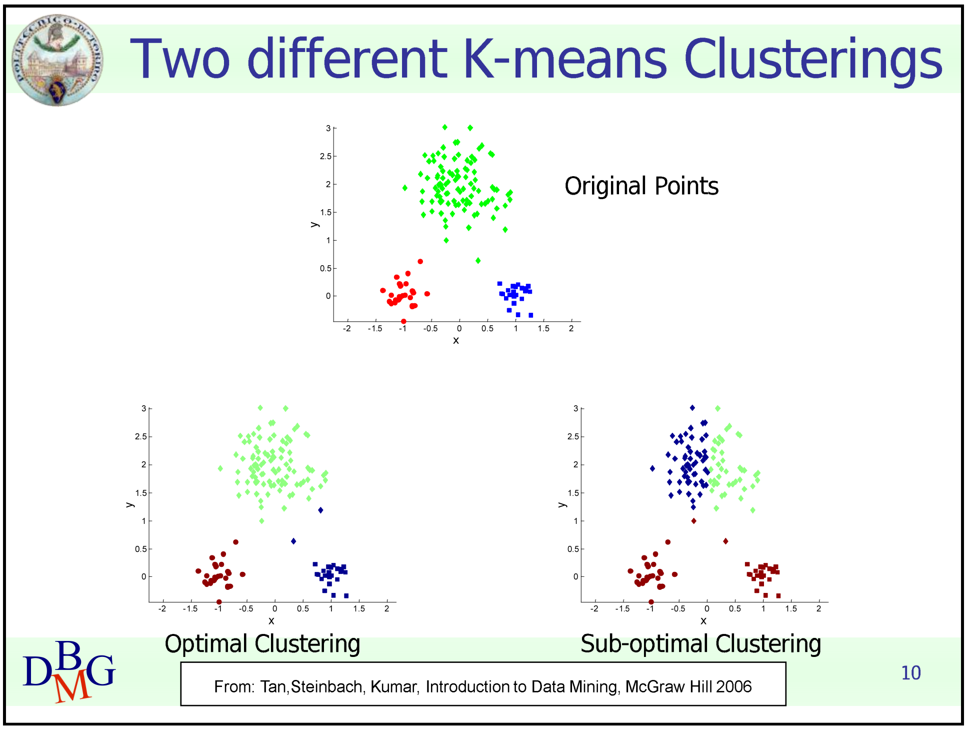


K-means Clustering

- 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

```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



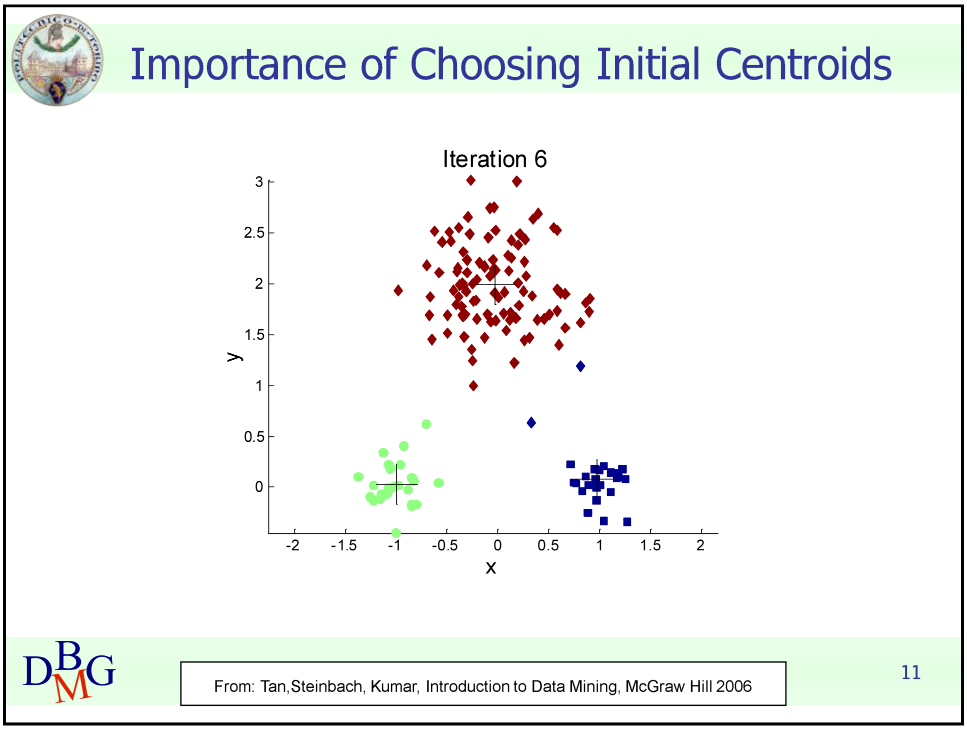
Two different K-means Clusterings

Original Points

Optimal Clustering

Sub-optimal Clustering

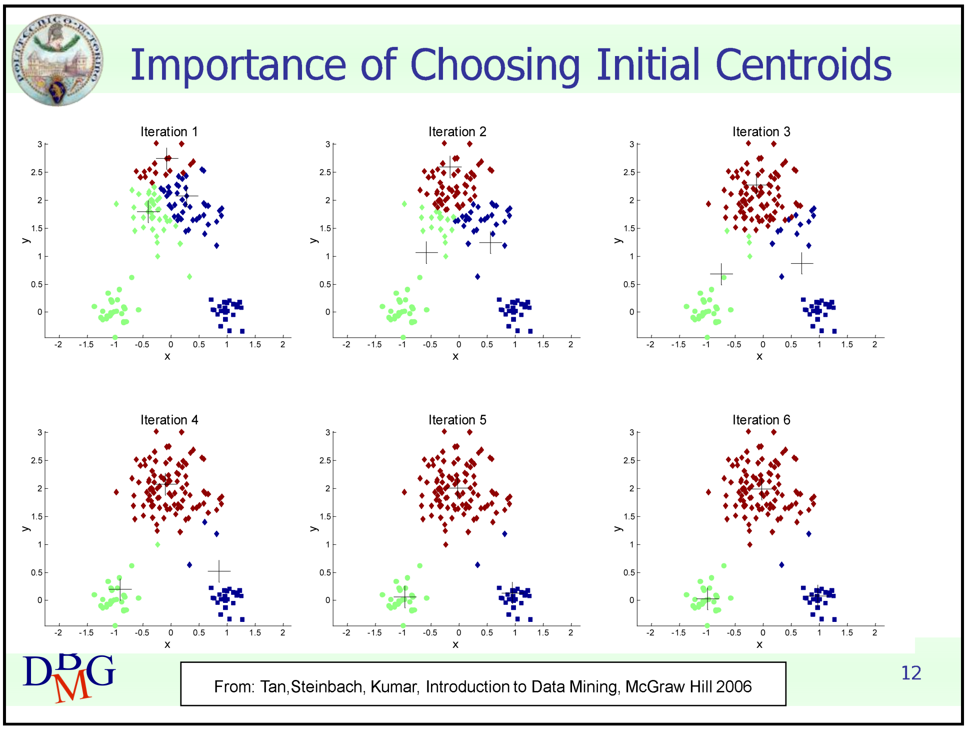
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



Importance of Choosing Initial Centroids

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

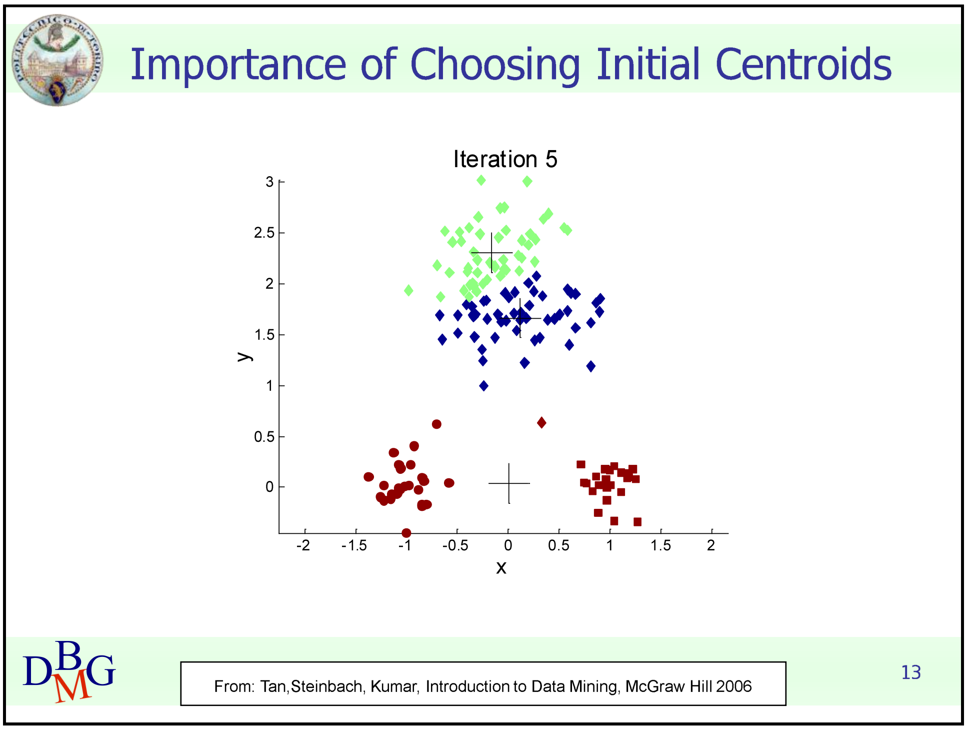
11



Importance of Choosing Initial Centroids

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

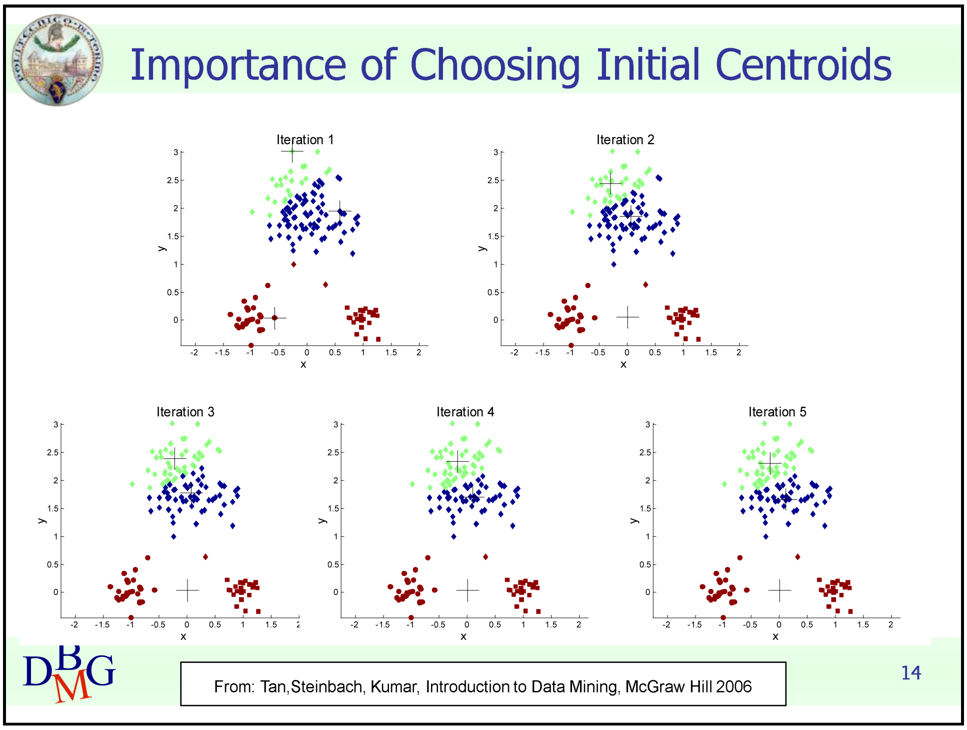
12



Importance of Choosing Initial Centroids

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

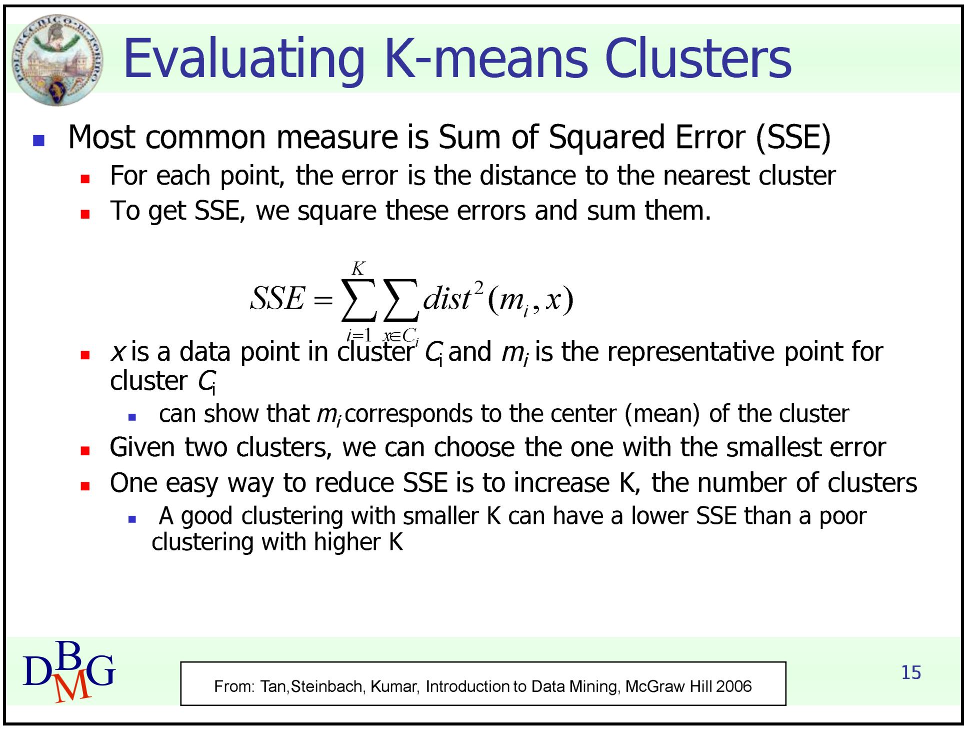
13



Importance of Choosing Initial Centroids

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

14



Evaluating K-means Clusters

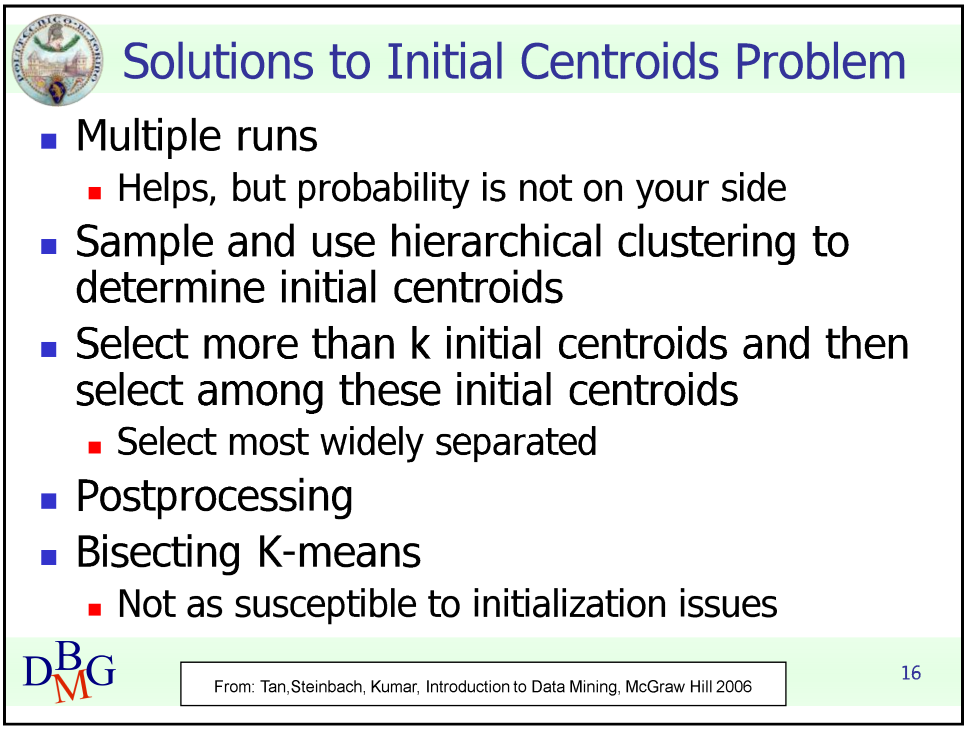
- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

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

  - *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 clusters, we can choose 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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

15

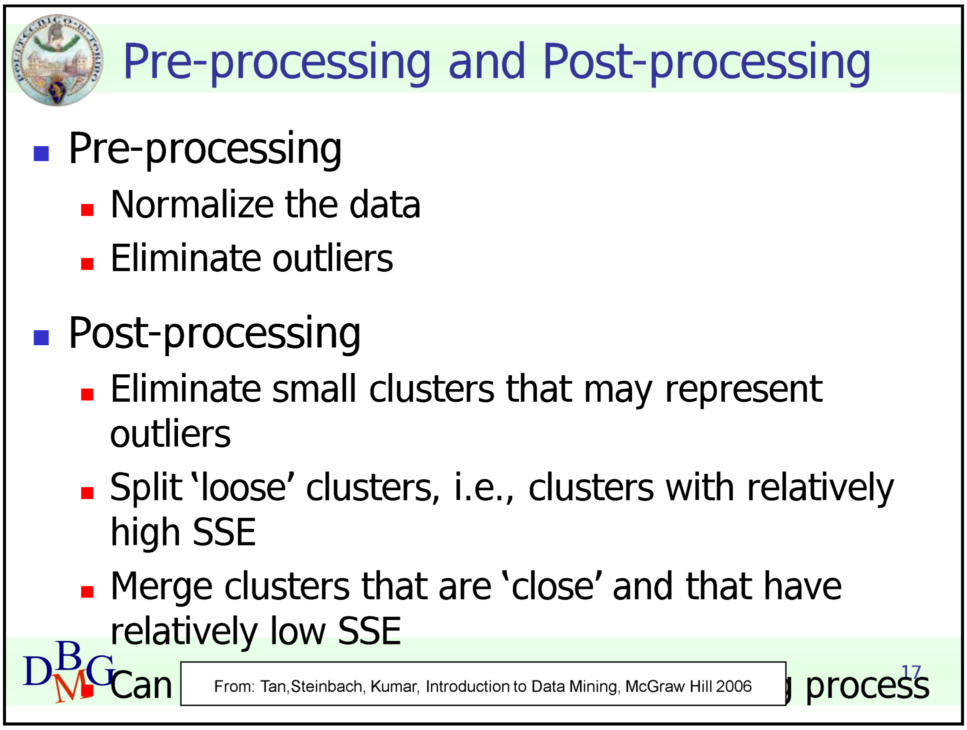


Solutions to Initial Centroids Problem

- 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
- Bisecting K-means
  - Not as susceptible to initialization issues

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

16



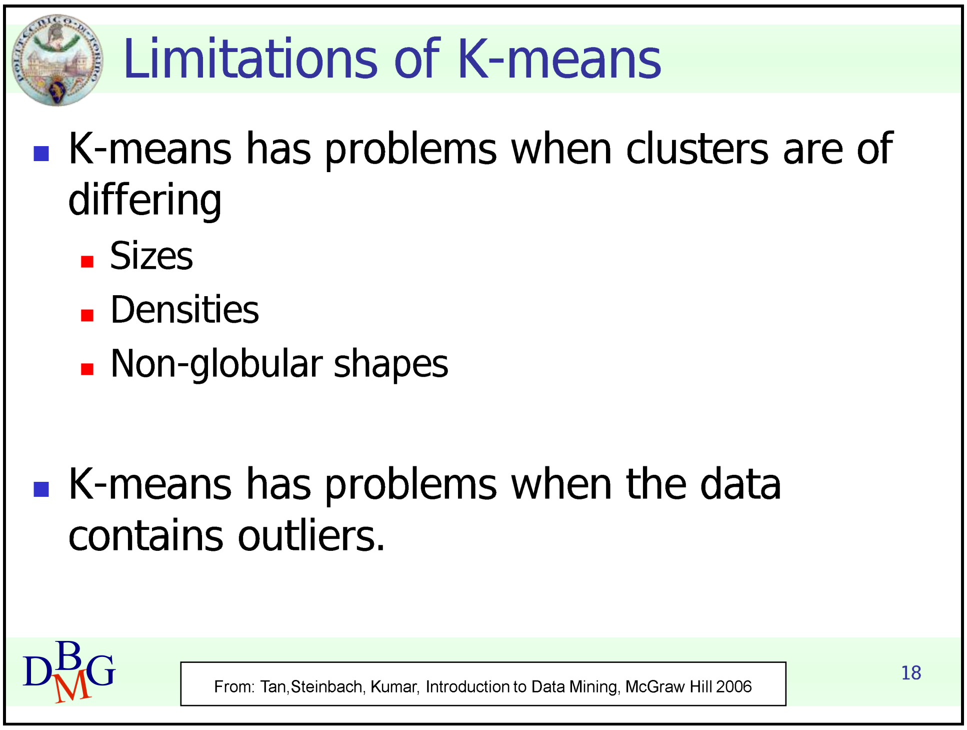
Pre-processing and Post-processing

- 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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

process

17

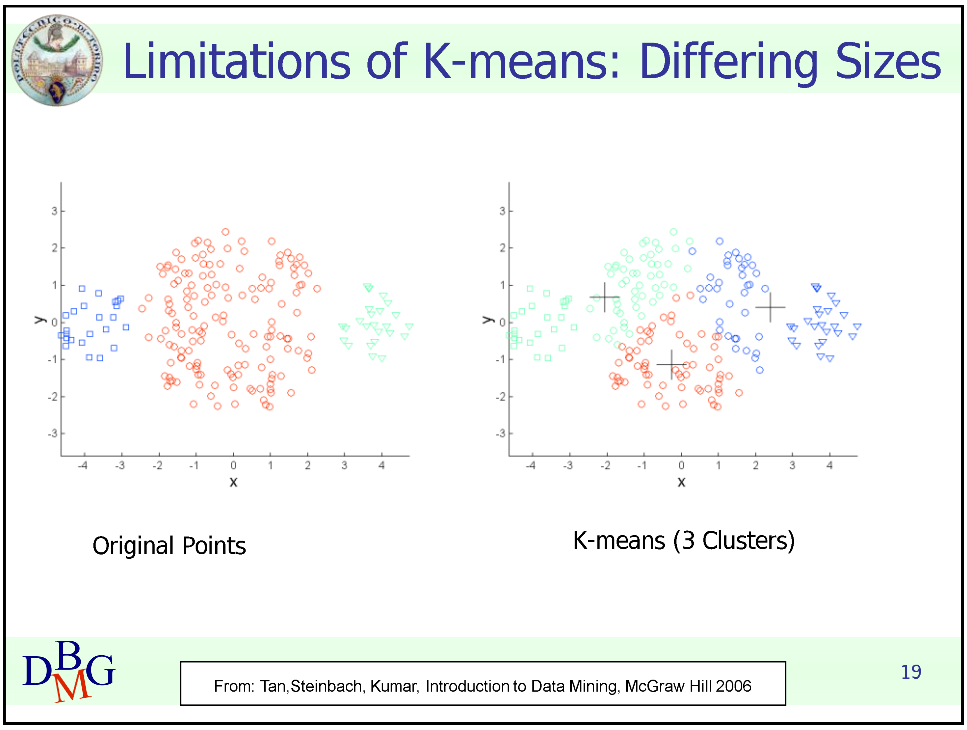


Limitations of K-means

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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

18

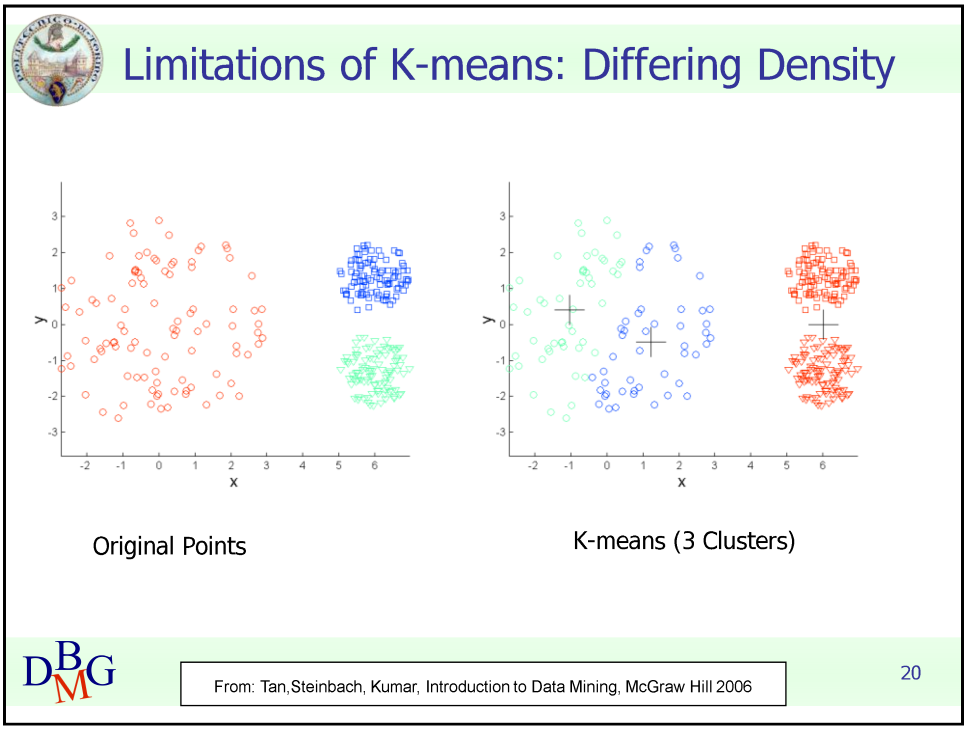


Limitations of K-means: Differing Sizes

Original Points

K-means (3 Clusters)

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

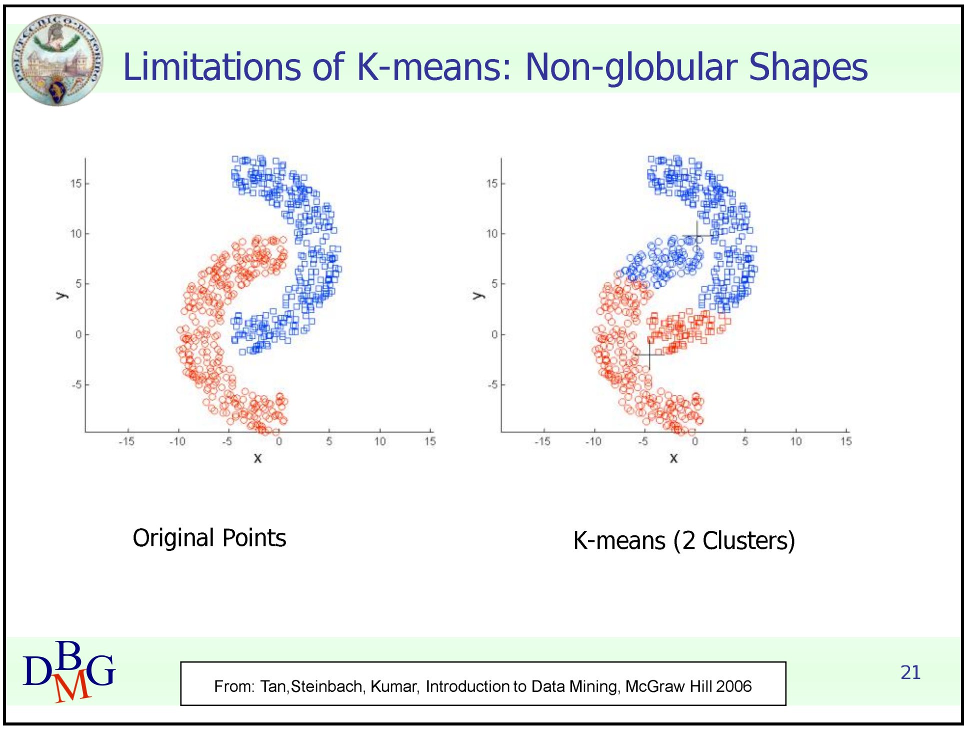


Limitations of K-means: Differing Density

Original Points

K-means (3 Clusters)

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



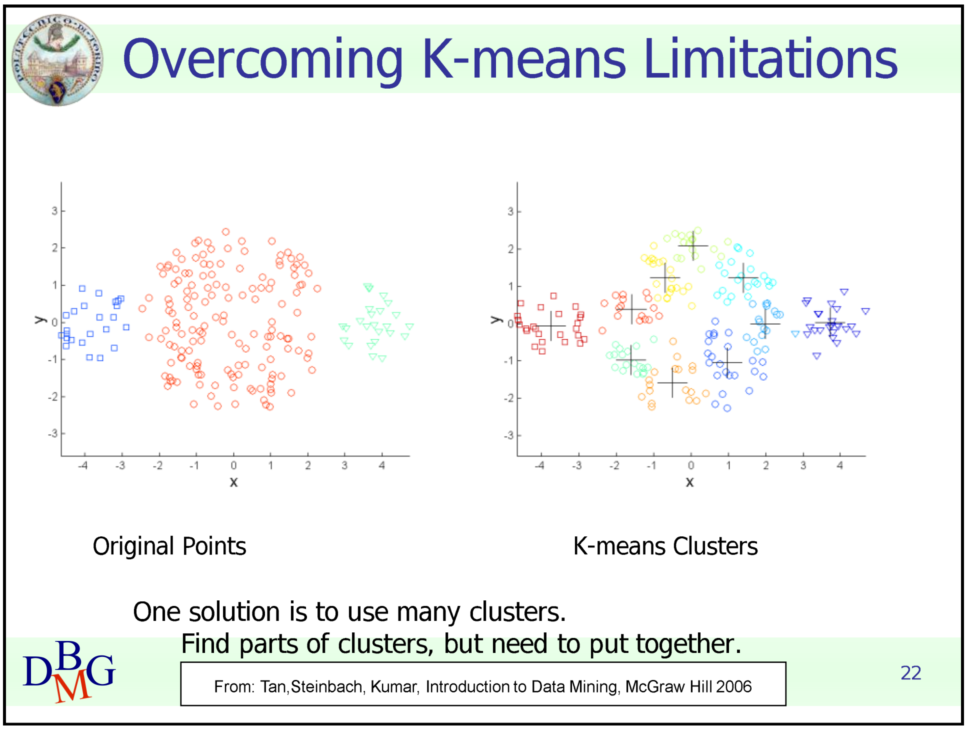
Limitations of K-means: Non-globular Shapes

Original Points

K-means (2 Clusters)

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

21



Overcoming K-means Limitations

Original Points

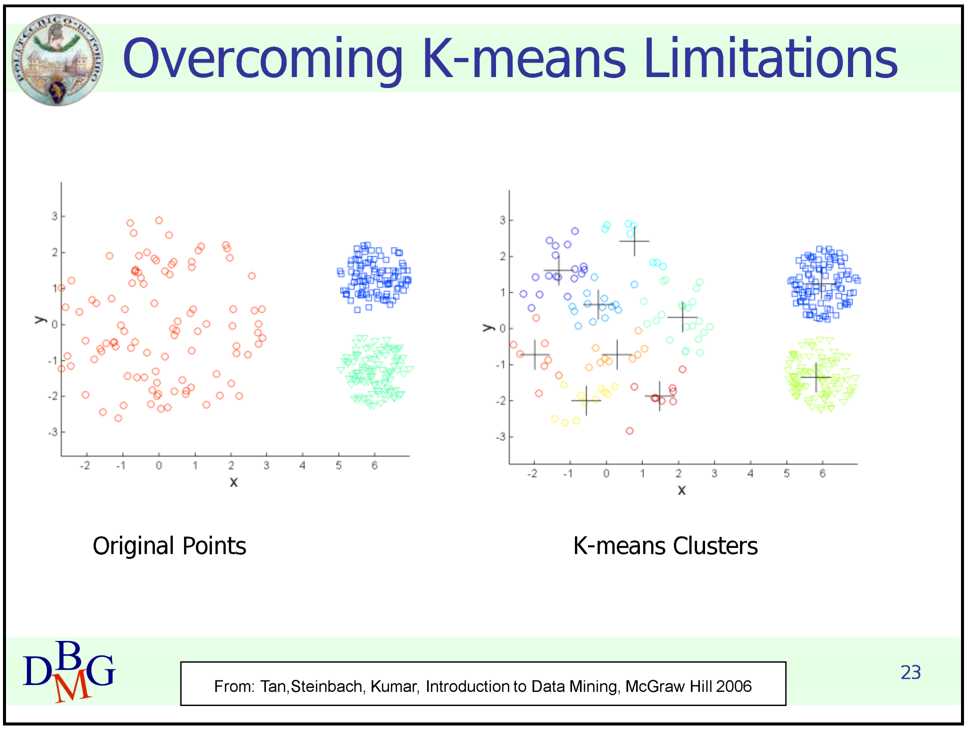
K-means Clusters

One solution is to use many clusters.

Find parts of clusters, but need to put together.

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

22



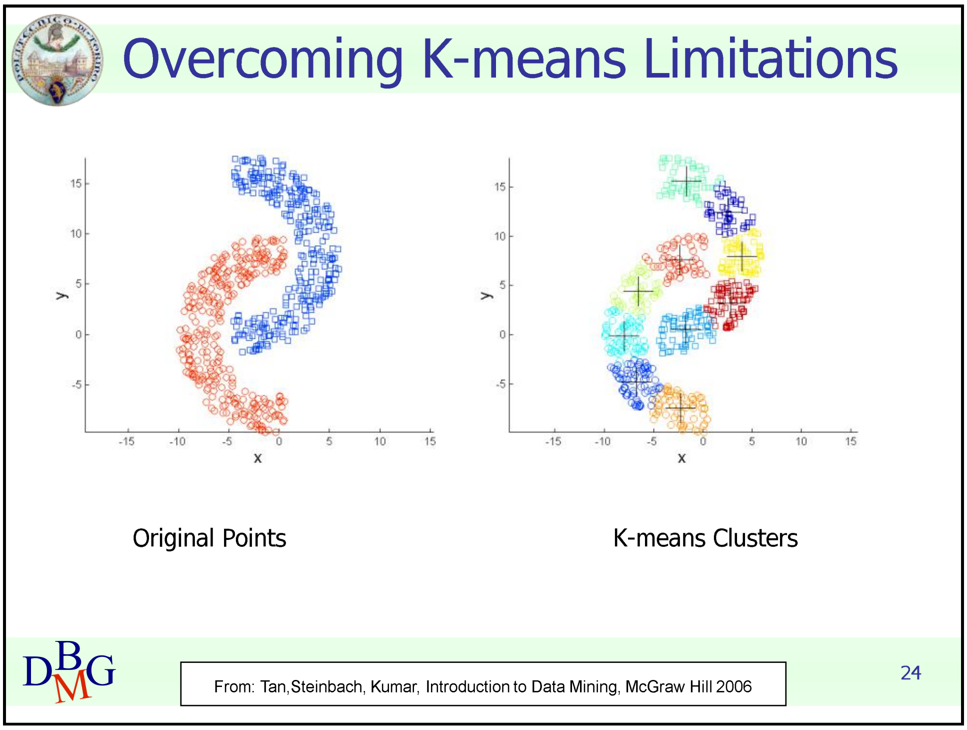
Overcoming K-means Limitations

Original Points

K-means Clusters

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

23



Overcoming K-means Limitations

Original Points

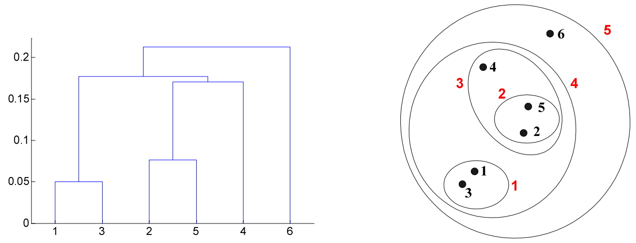
K-means Clusters

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

24

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



DBMG

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

25

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 dendogram at the proper level
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

DBMG

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

26

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
  - Merge or split one cluster at a time

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

27

Agglomerative Clustering Algorithm

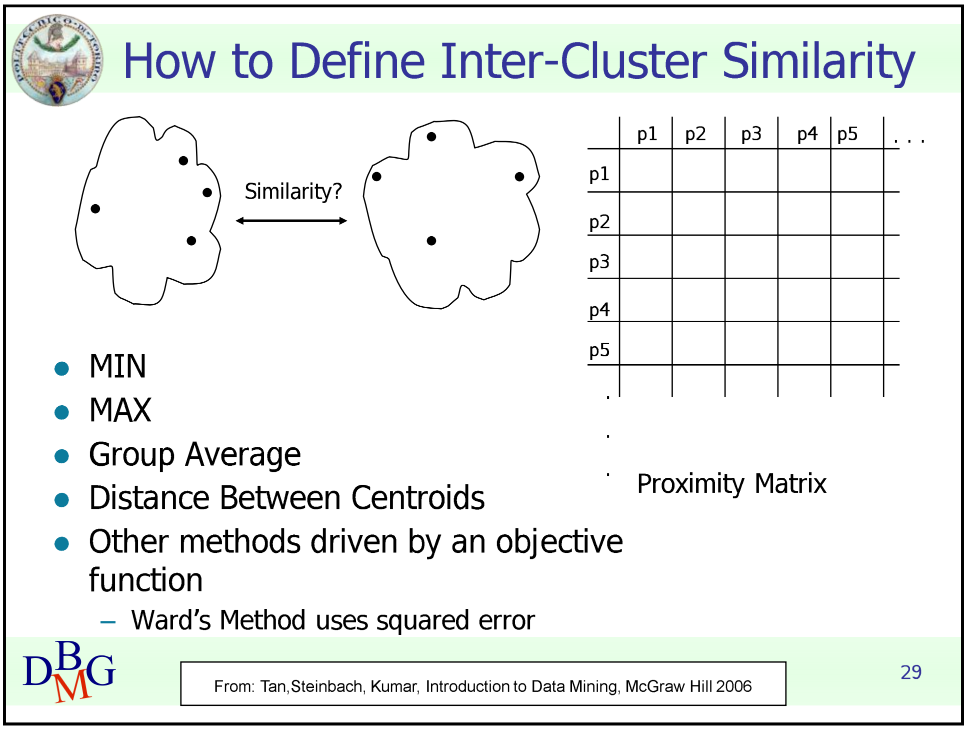
- More popular hierarchical clustering technique
- Basic algorithm is straightforward

```
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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

28



How to Define Inter-Cluster Similarity

Similarity?

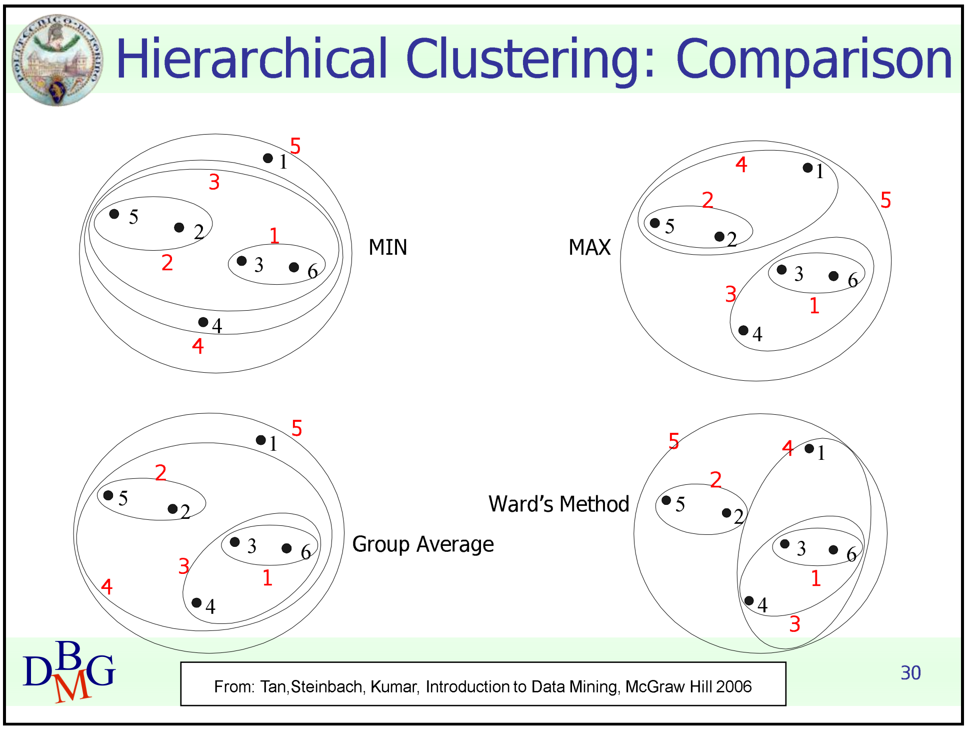
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

29



Hierarchical Clustering: Comparison

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

30

DBSCAN

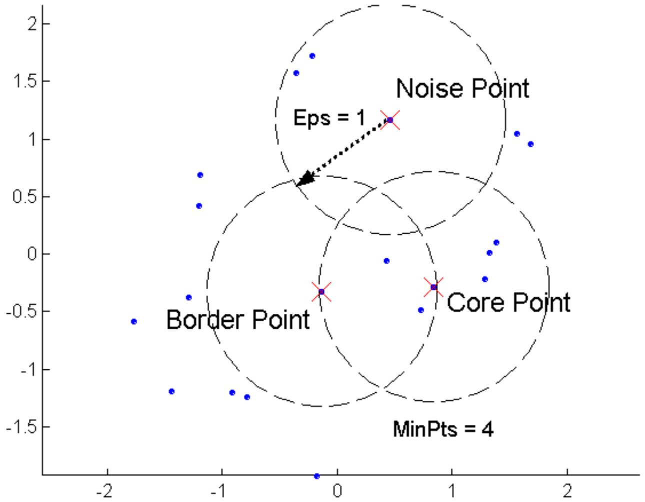
- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts 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.

DBMG

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

31

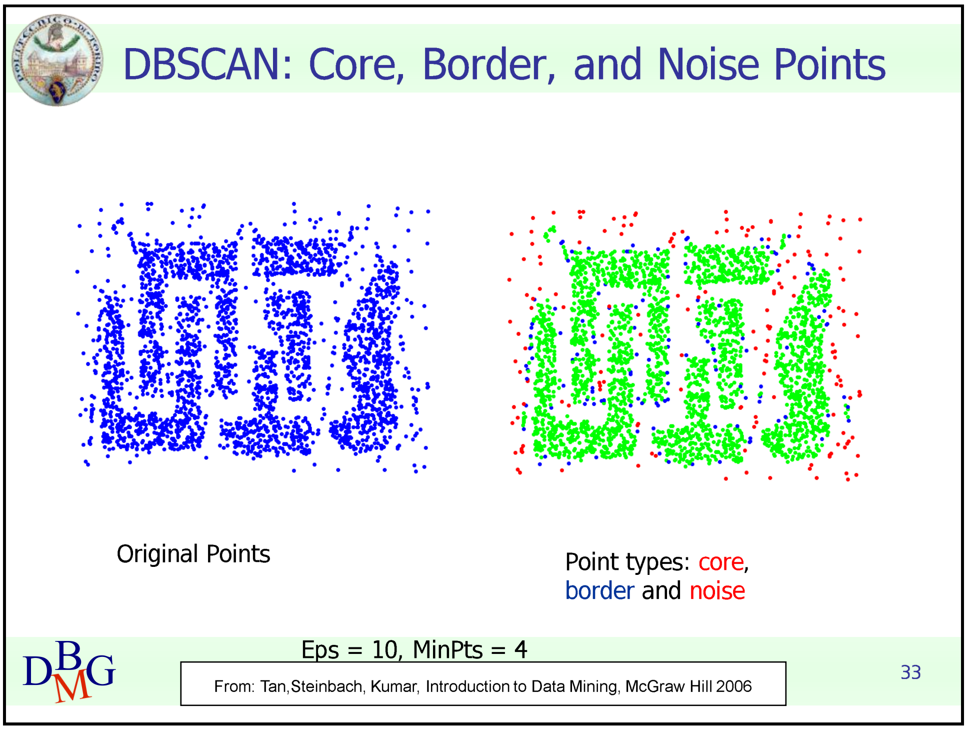
DBSCAN: Core, Border, and Noise Points



DBMG

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

32



DBSCAN: Core, Border, and Noise Points

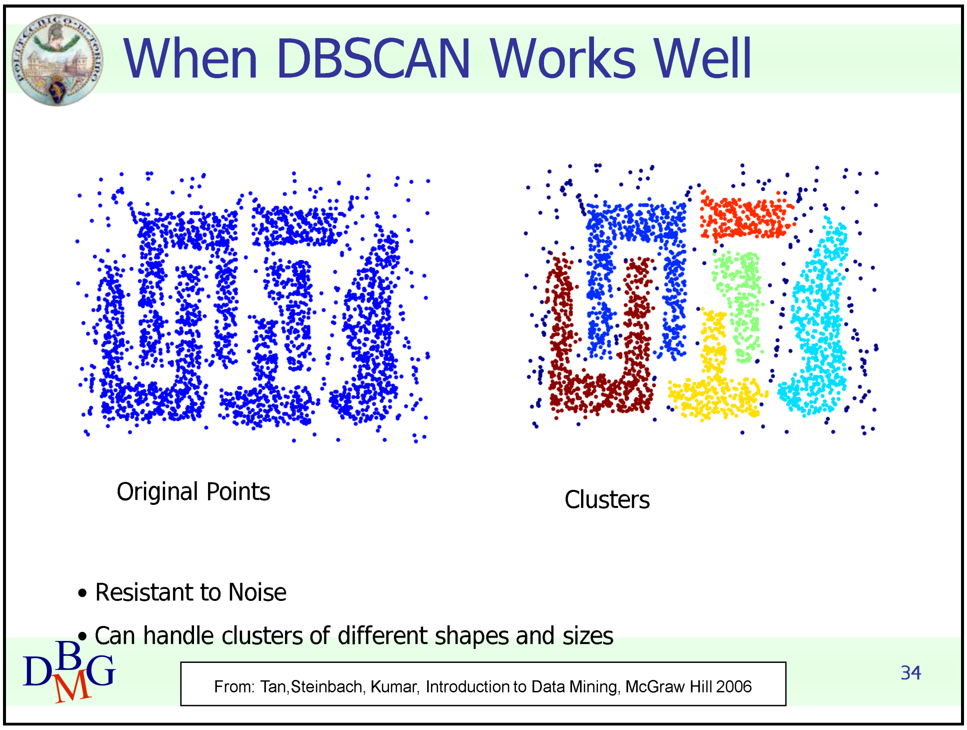
Original Points

Point types: core, border and noise

Eps = 10, MinPts = 4

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

33



When DBSCAN Works Well

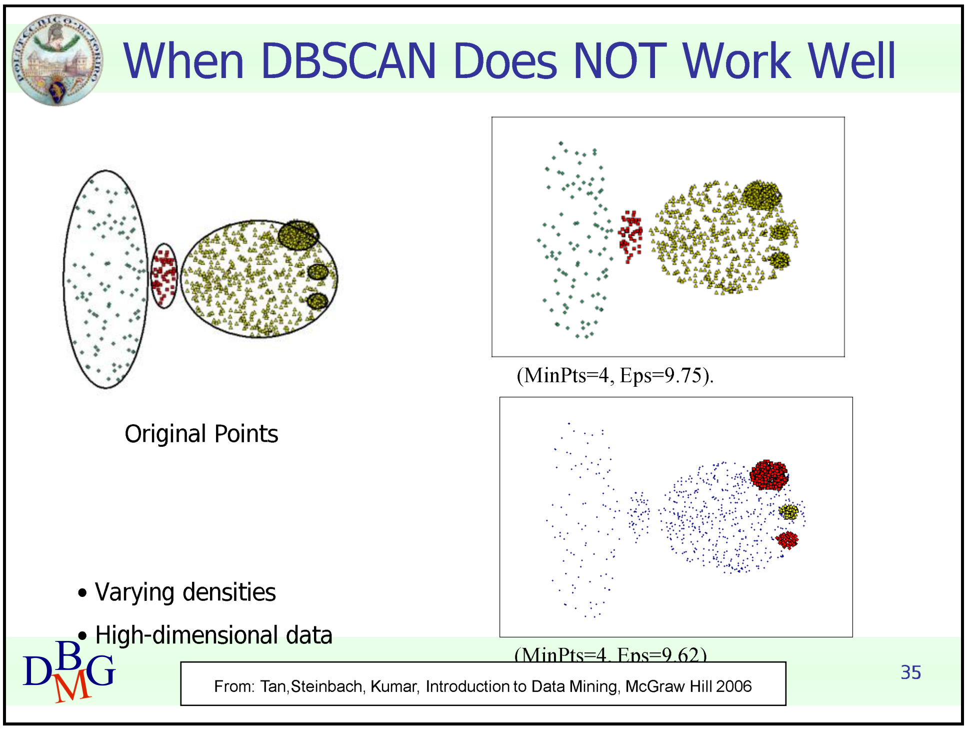
Original Points

Clusters

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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

34



When DBSCAN Does NOT Work Well

Original Points

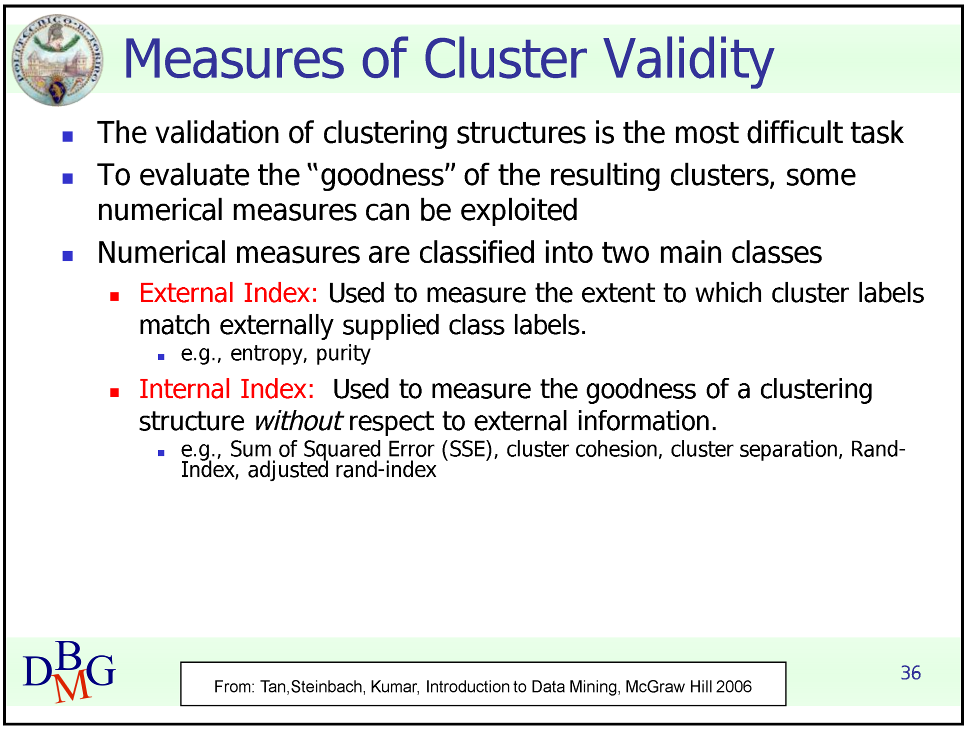
(MinPts=4, Eps=9.75).

(MinPts=4, Eps=9.62)

- Varying densities
- High-dimensional data

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

35



Measures of Cluster Validity

- The validation of clustering structures is the most difficult task
- To evaluate the "goodness" of the resulting clusters, some numerical measures can be exploited
- Numerical measures are classified into two main classes
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - e.g., entropy, purity
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - e.g., Sum of Squared Error (SSE), cluster cohesion, cluster separation, Rand-Index, adjusted rand-index

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

36

External Measures of Cluster Validity: Entropy and Purity

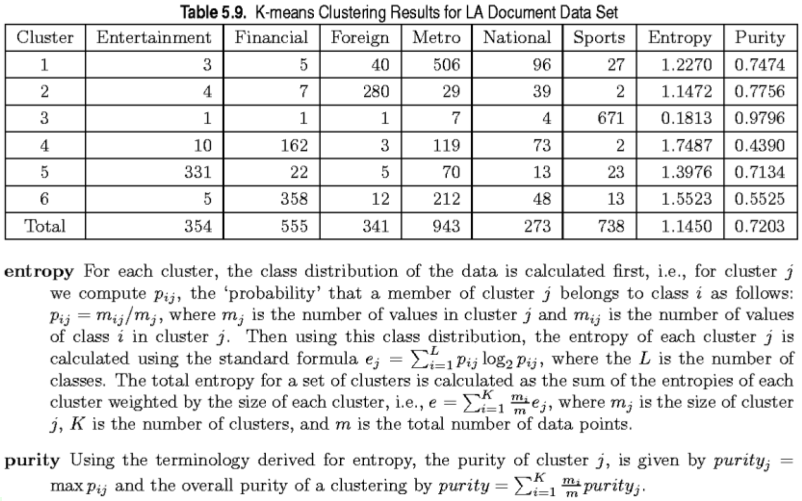


Table 5.9. K-means Clustering Results for LA Document Data Set

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

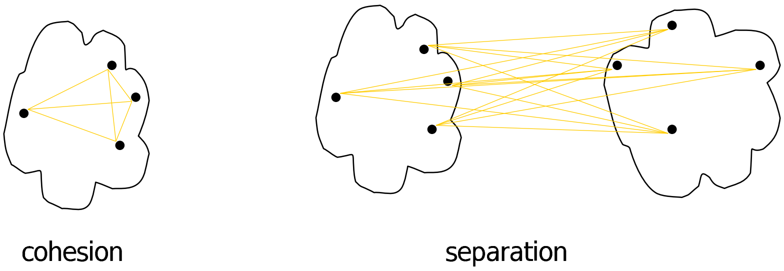
entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values 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., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, 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 $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

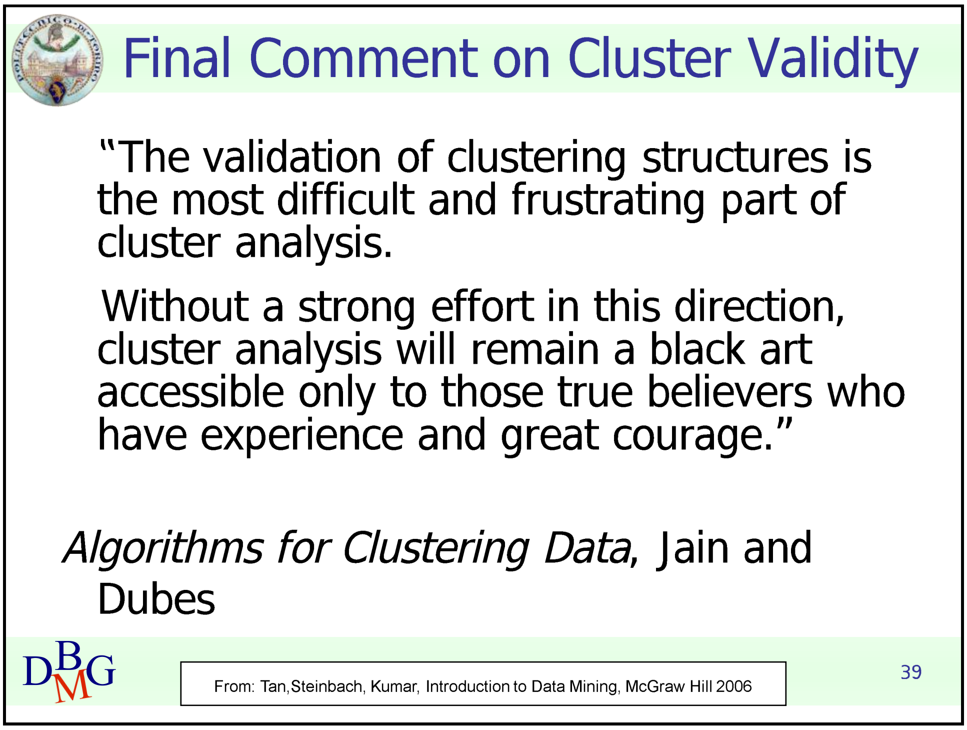
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

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.



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006



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

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006