



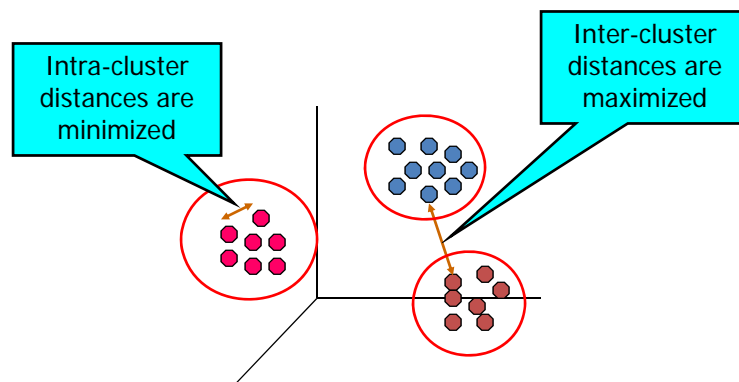
Clustering (Part I)

nanopoulos@ismll.de



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





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

Document classification

Cluster Weblog data to discover groups of similar access patterns

3



Marketing

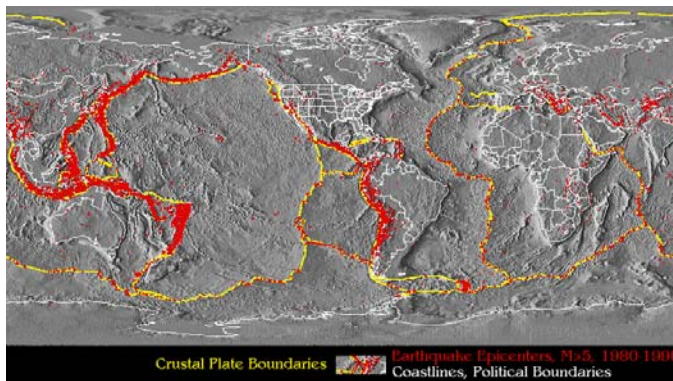


finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;

4



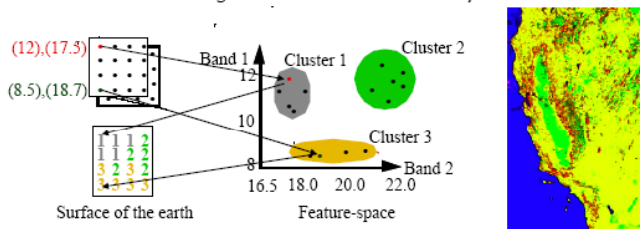
Geology



clustering observed earthquake epicenters to identify dangerous zones;

A Typical Application: Thematic Maps

- Satellite images of a region in different wavelengths
 - Each point on the surface maps to a high-dimensional feature vector $p = (x_1, \dots, x_d)$ where x_i is the recorded intensity at the surface point in band i .
 - Assumption: each different land-use reflects and emits light of different wavelengths in a characteristic way.





Application: Web Usage Mining

Determine Web User Groups

Sample content of a web log file

```
roublon.informatik.uni-muenchen.de lopa - [04/Mar/1997:01:44:50 +0100] "GET /-lopa/ HTTP/1.0" 200 1364
roublon.informatik.uni-muenchen.de lopa - [04/Mar/1997:01:45:11 +0100] "GET /-lopa/x/ HTTP/1.0" 200 712
fixer.sega.co.jp unknown - [04/Mar/1997:01:58:49 +0100] "GET /db/porada.html HTTP/1.0" 200 1229
scooter.pa-x.dec.com unknown - [04/Mar/1997:02:08:23 +0100] "GET /db/kriegel_e.html HTTP/1.0" 200 1241
```

Generation of sessions

➡ Session ::= <IP_address, user_id, [URL₁, . . . , URL_k]>

which entries form a single session?

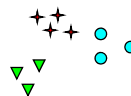
$$\text{Distance function for sessions: } d(x, y) = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}$$



Notion of a Cluster can be Ambiguous



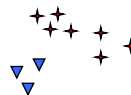
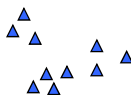
How many clusters?



Six Clusters




Two Clusters



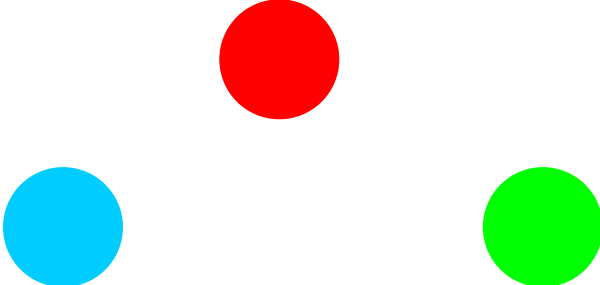
Four Clusters






Types of Clusters: Well-Separated

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



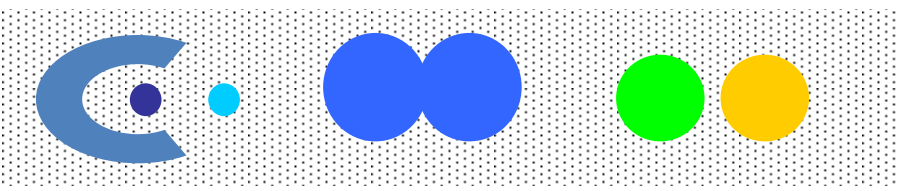
3 well-separated clusters

9



Types of Clusters: Density-Based

Density-based
 A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
 Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

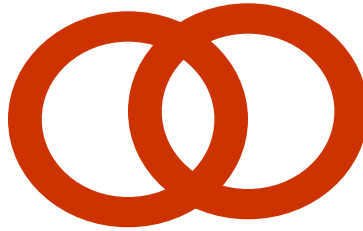
10



Types of Clusters: Conceptual Clusters

Shared Property or Conceptual Clusters

Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

11



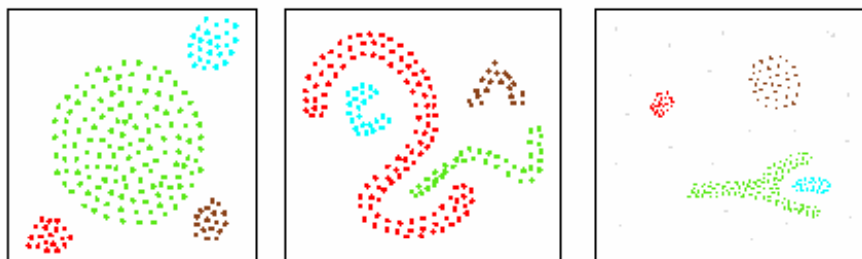
Requirements for Clustering Algorithms

- Scalability
 - Ability to deal with different types of attributes
 - Discovery of clusters with arbitrary shape
 - Minimal requirements for domain knowledge to determine input parameters
 - Able to deal with noise and outliers
 - Insensitive to order of input records
 - High dimensionality
 - Incorporation of user-specified constraints
 - Interpretability and usability
-

12



Arbitrary shapes



13



Minimal parameters

Clustering Parameters

Check columns to cluster

Column Name	Type
<input checked="" type="checkbox"/> CALORIES	INTEGER
<input checked="" type="checkbox"/> PROTEIN	INTEGER
<input checked="" type="checkbox"/> FAT	INTEGER
<input checked="" type="checkbox"/> SODIUM	INTEGER
<input checked="" type="checkbox"/> FIBER	REAL
<input checked="" type="checkbox"/> CARBO	REAL
<input checked="" type="checkbox"/> SUGARS	INTEGER
<input checked="" type="checkbox"/> POTASS	INTEGER
<input checked="" type="checkbox"/> VITAMINS	INTEGER

Uncheck All Check All

Linkage Methods

- Average Linkage (UPGMA)
- Average Group Linkage
- Complete Linkage
- Single Linkage
- Shneiderman's 1by1

Clustering Direction

- Cluster Rows
Load Similarity Matrix for Rows
- Cluster Columns
Load Similarity Matrix for Columns

Node Arrangement Methods

- Keep Right Child Redder
- Keep Right Child Small

Similarity/Distance Measure

Pearson Correlation Coefficient

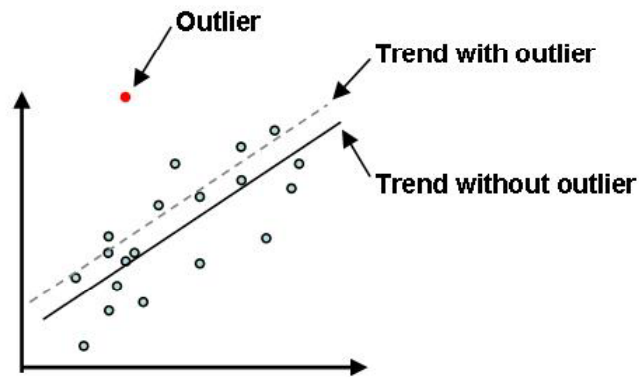
- Move center to average (Don't fix average at 0)
- Take absolute value
- Use PValues as Weights

OK Cancel

14



Noise/outliers



15



Major Clustering Approaches

Partitioning algorithms: Construct various partitions and then evaluate them by some criterion

Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion


Density-based: based on connectivity and density functions

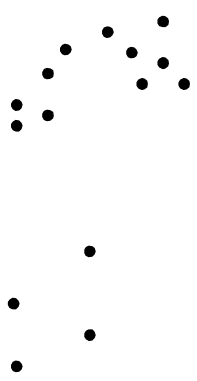
Grid-based: based on a multiple-level granularity structure

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

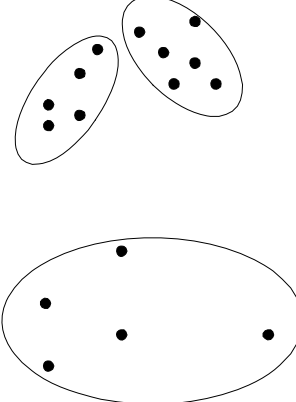
16

Partitional Clustering






Original Points



A Partitional Clustering

Partitioning Algorithms: Basic Concept



Partitioning method: Construct a partition of a database D of n objects into a set of k clusters

Given a k , find a partition of k clusters that optimizes the chosen partitioning criterion

- Global optimal: exhaustively enumerate all partitions
- Heuristic methods: *k-means* and *k-medoids* algorithms

k-means: Each cluster is represented by the center of the cluster

k-medoids or PAM (Partition around medoids): Each cluster is represented by one of the objects in the cluster

18



Enumerate all possible partitions

$S(n,k)$: number of possible k partitions with n points in total

$S(n-1,k-1)$ cases of $k-1$ partitions with $n-1$ points in total.

The n -th point creates a new partition that is combined with each of existing $S(n-1,k-1)$ such that we have k partitions with n points in total. Totally $S(n-1,k-1)$ cases.

$S(n-1,k)$ cases of k partitions with $n-1$ points in total. The n -th point can be included in each of the partitions so that we have k partitions with n points in total. Totally we have $k S(n-1,k)$ cases.

19



Enumerate all possible partitions: example

Example: A, B, C, D points. $S(4,3) = ?$

$(A, B) \quad (C)$

$(A, C) \quad (B)$

$(A) \quad (B, C)$

$S(3,2) = 3$

$(A, B) \quad (C) \quad (D)$

$(A, C) \quad (B) \quad (D)$

$(A) \quad (B, C) \quad (D)$

$S(3,2)$ cases

$(A) \quad (B) \quad (C)$

$S(3,3) = 1$

$(A, D) \quad (B) \quad (C)$

$(A) \quad (B, D) \quad (C)$

$(A) \quad (B) \quad (C, D)$

$3 \times S(3,3)$ cases

20

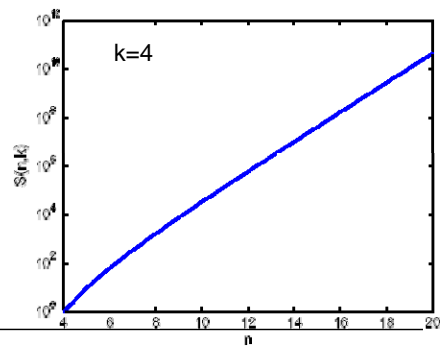


Enumerate all possible partitions

$$S(n, k) = S(n - 1, k - 1) + kS(n - 1, k)$$

$$S(n, 1) = 1, \quad S(n, n) = 1, \quad S(n, k) = 0 \text{ for } k > n$$

$$S(n, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^{k-i} \binom{k}{i} i^n$$



21



K-means Clustering

Partitional clustering approach

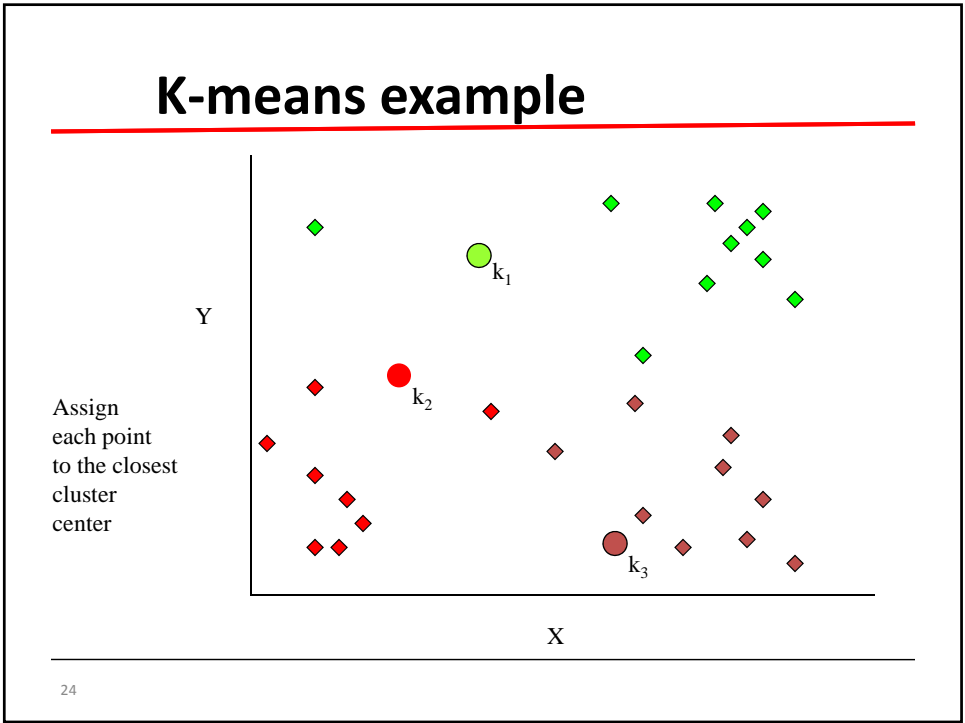
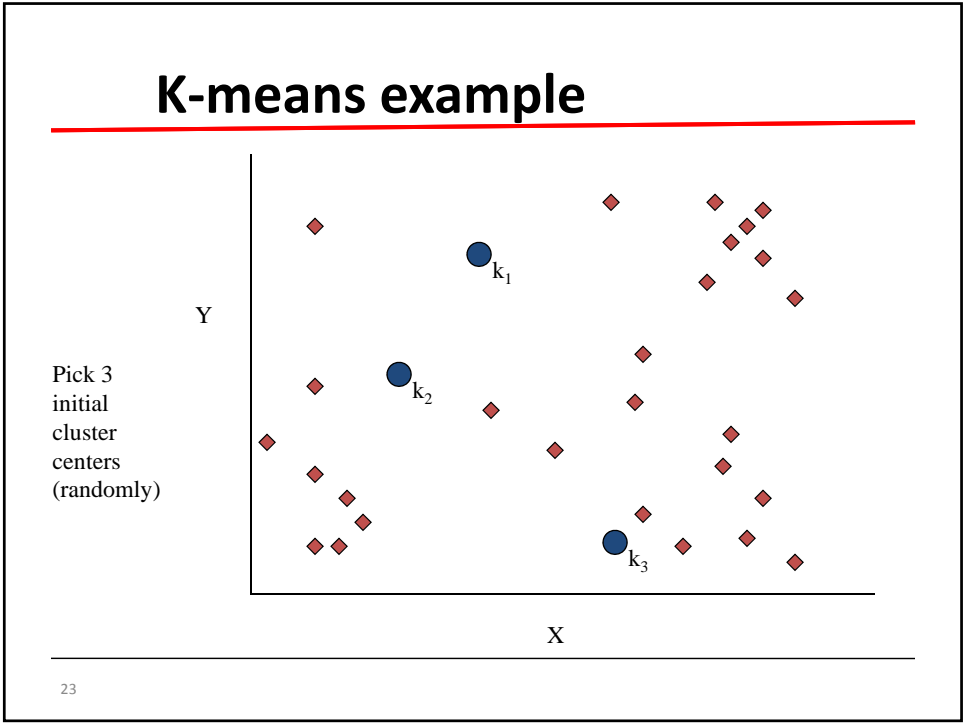
Each cluster is associated with a **centroid** (center point)

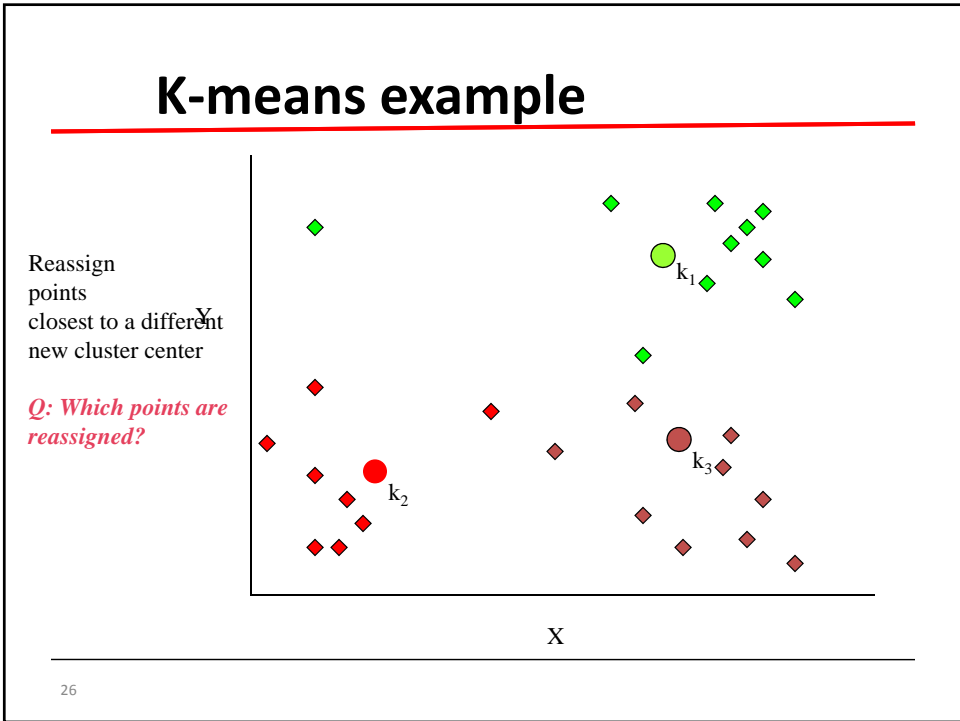
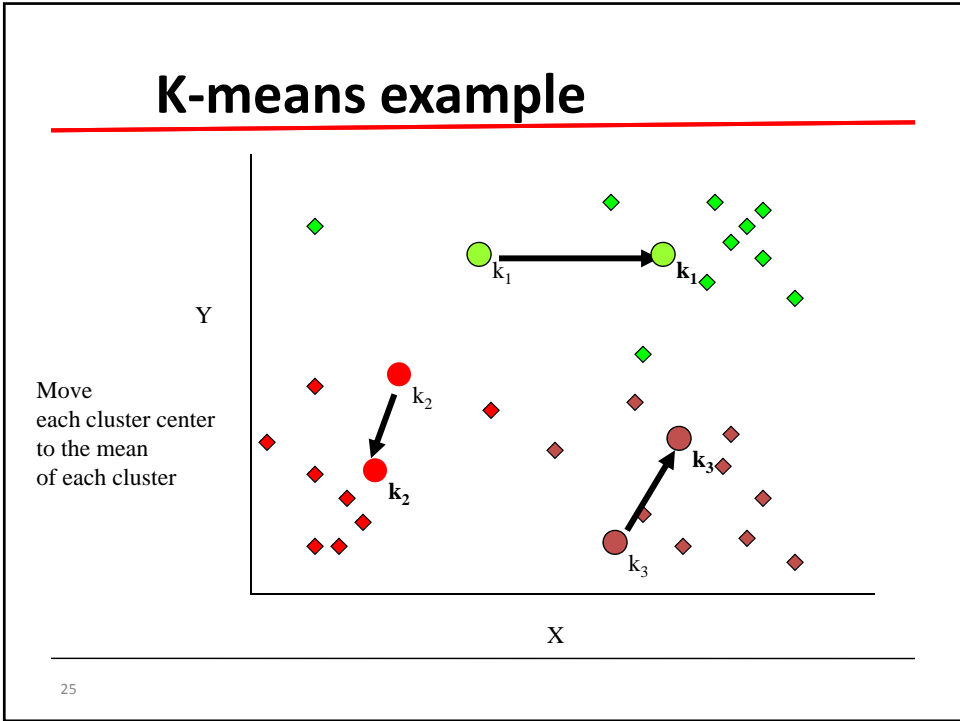
Each point is assigned to the cluster with the closest centroid

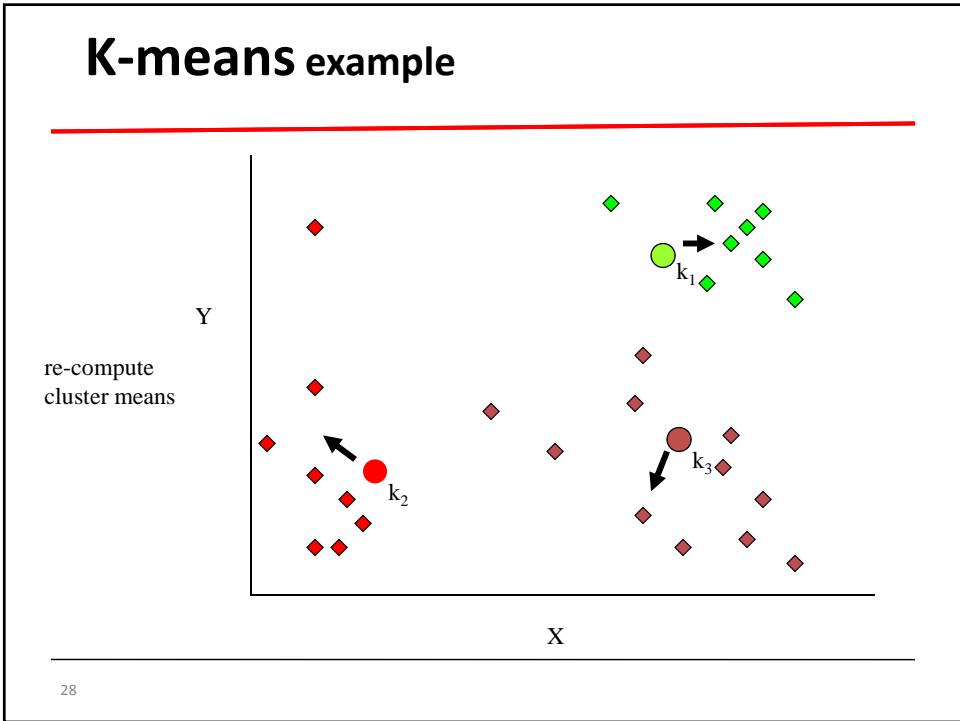
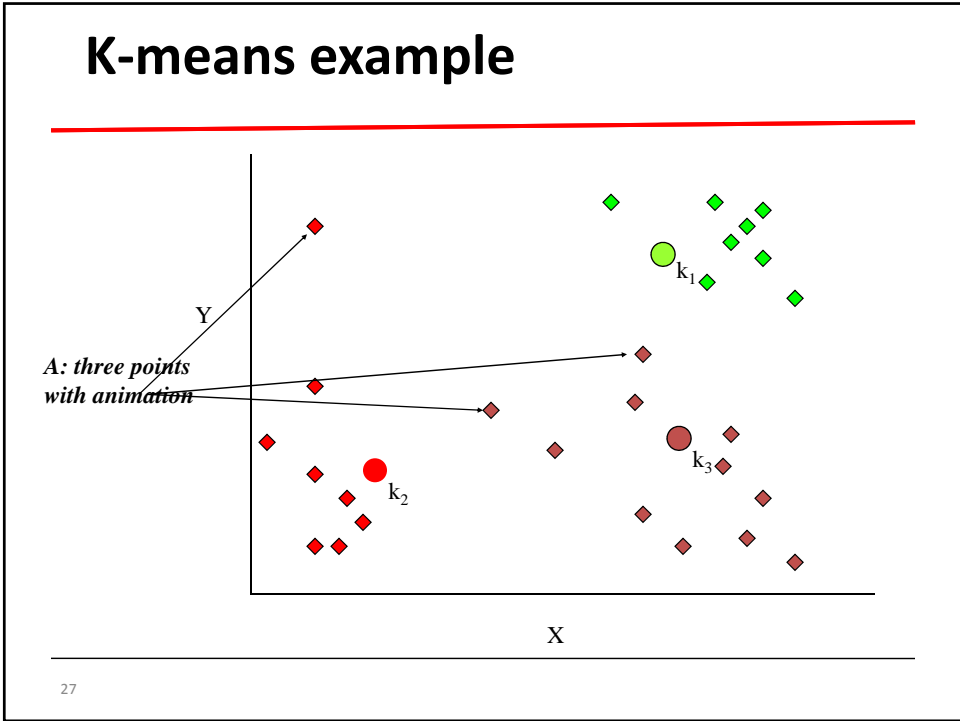
Number of clusters, K , must be specified

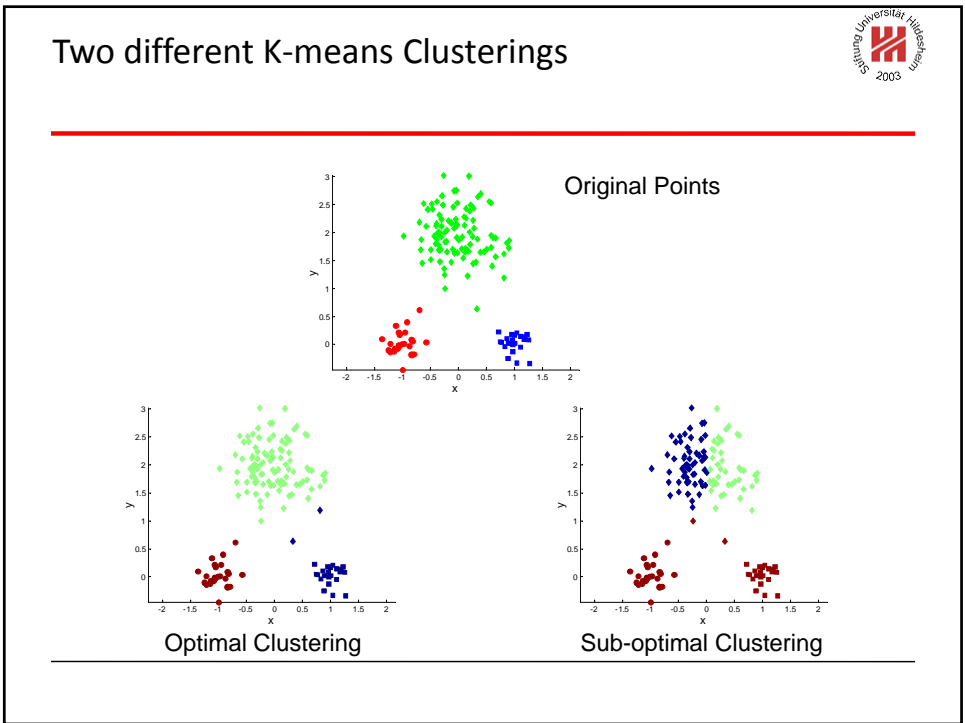
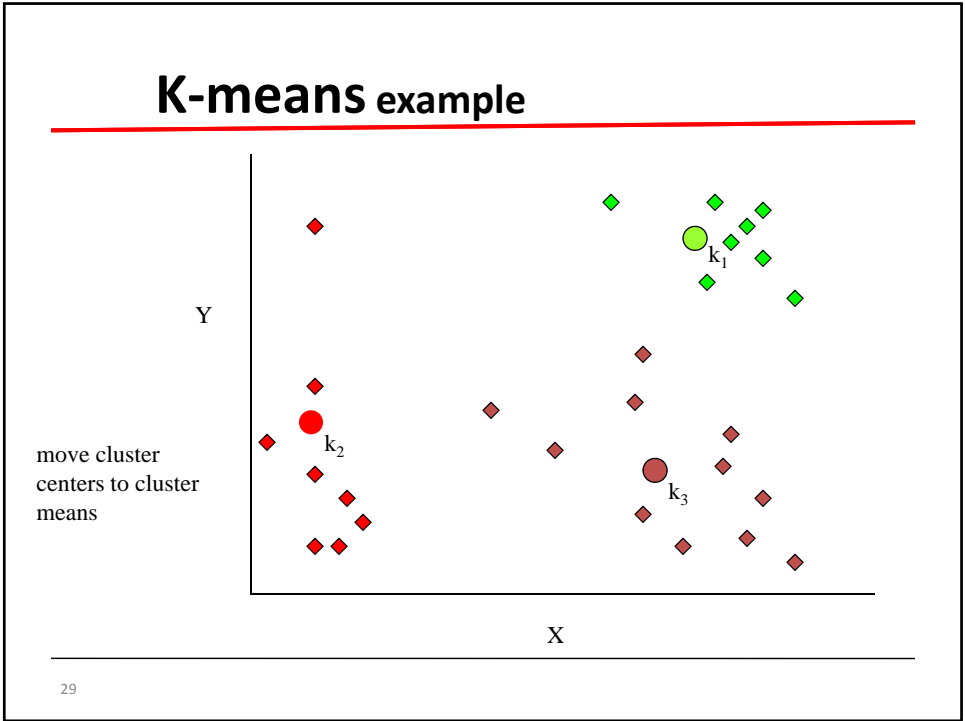
The basic algorithm is very simple

- 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



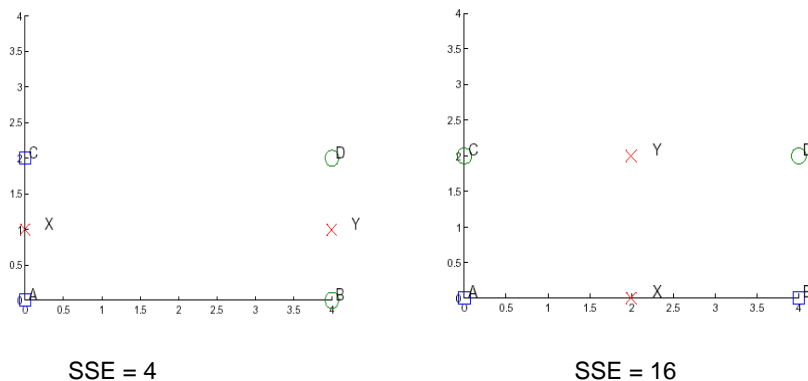








Local minima

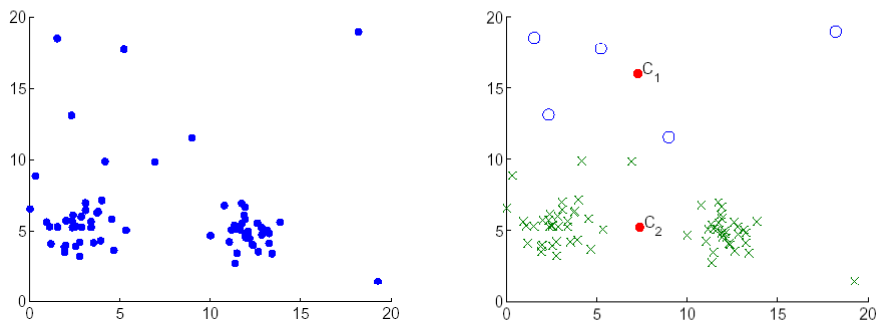


Repeat with different random initial centers

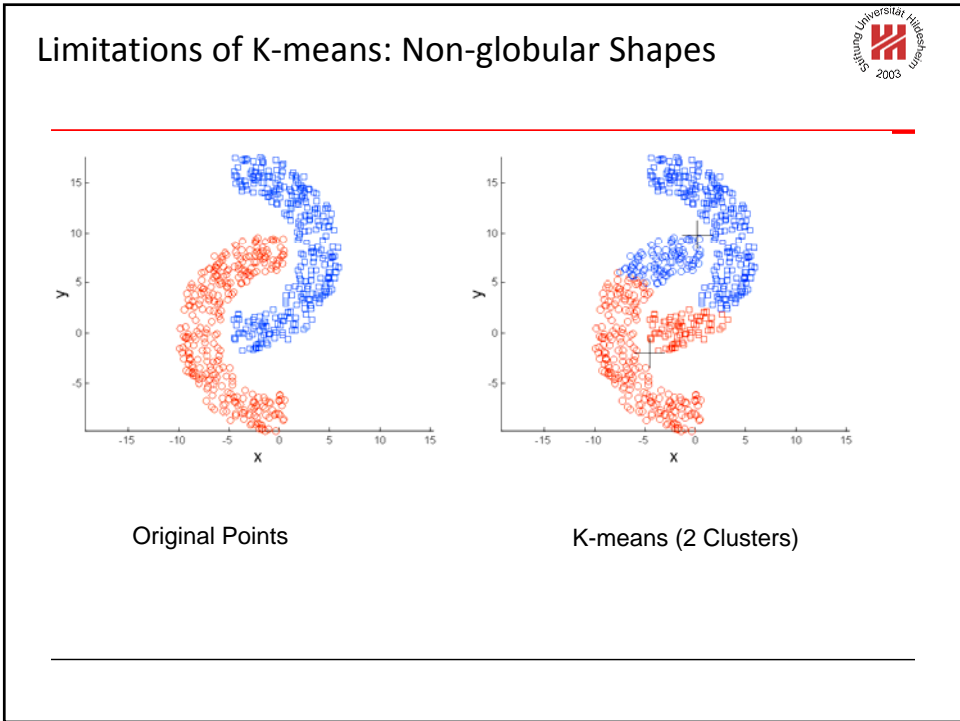
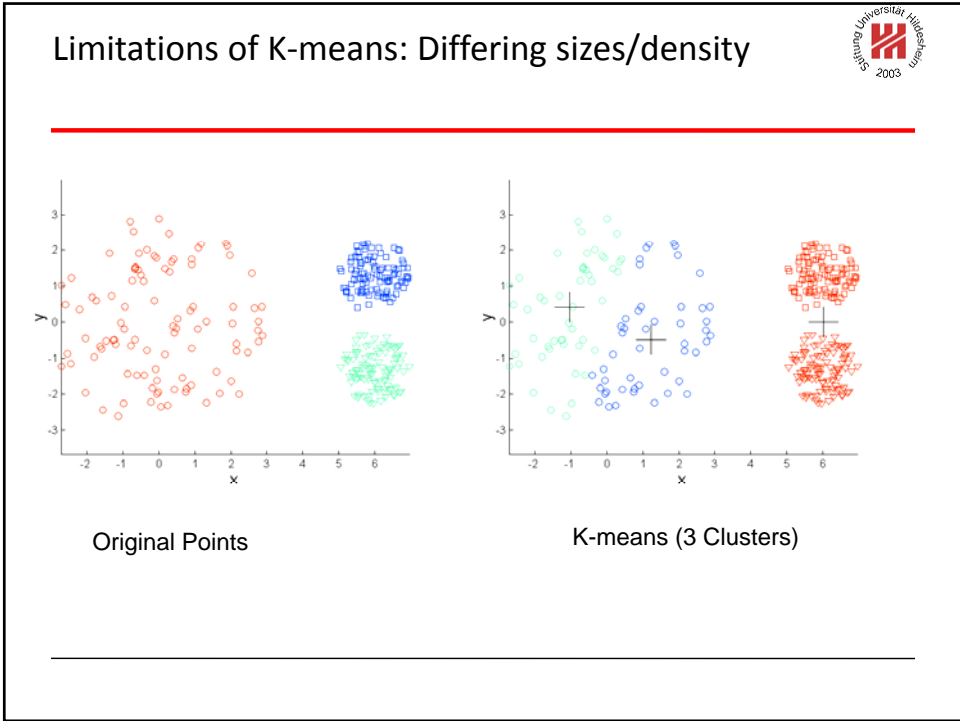
31



Sensitivity to noise/outliers



32



K-means characteristics

Pros

Simple

Fast ($O(nd)$)

Cons

Selection of k

Local minima

Sensitive to noise,
sizes, shapes