Cluster Analysis

Intuitive Idea

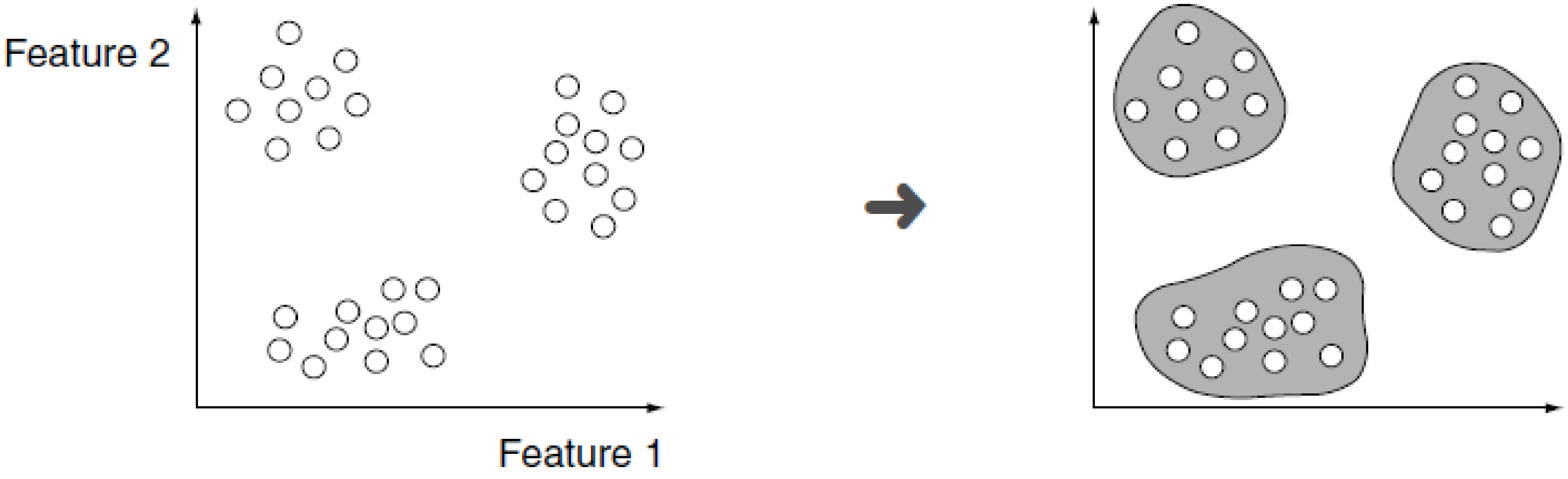
What is Cluster Analysis?

- **Cluster**: a collection of observations
  - Similar to one another within the same cluster
  - Dissimilar to the observations in other clusters
- **Cluster analysis**
  - Grouping a set of data observations into classes
- Clustering is unsupervised classification: no predefined classes—descriptive data mining.
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ denote the p-dimensional feature vectors of n objects:

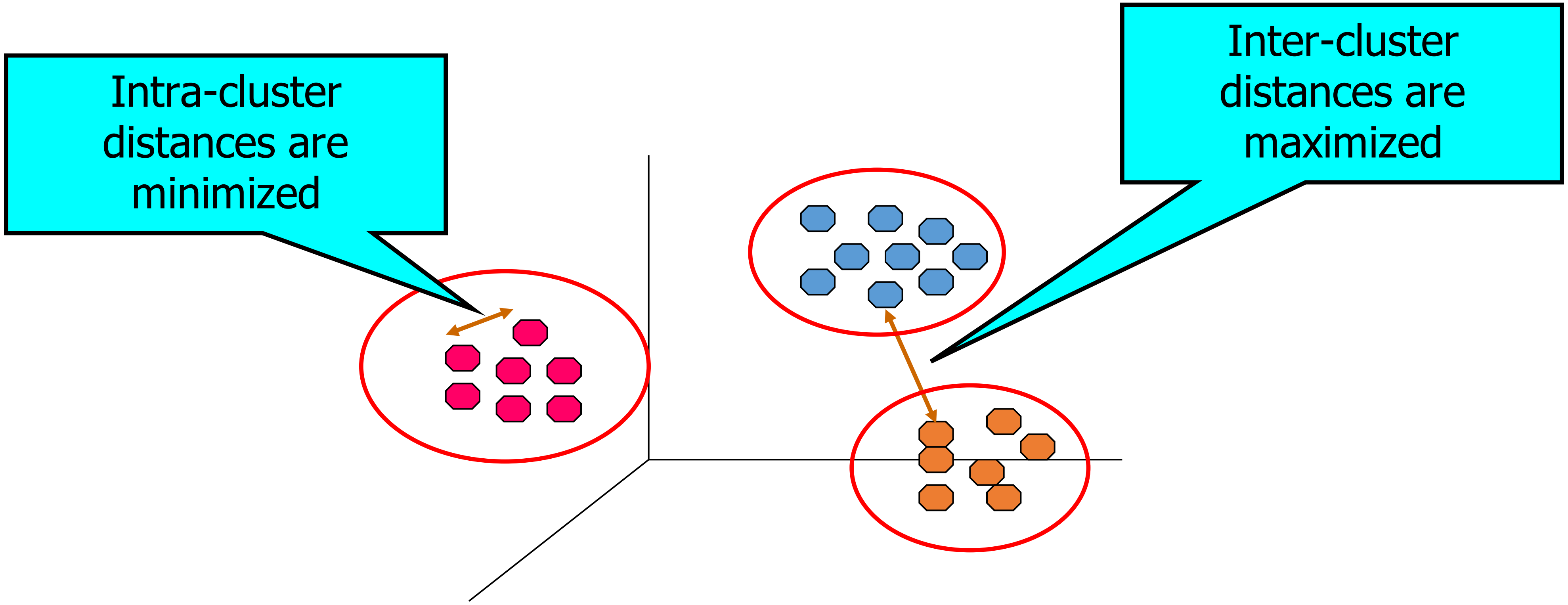
	Feature 1	Feature 2	...	Feature p	no Target concept
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	...	x_{1_p}	c_1
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	...	x_{2_p}	c_2
$\vdots$					$\vdots$
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	...	x_{n_p}	c_n

30 two-dimensional feature vectors $(n = 30,\ p = 2)$:



What is Cluster Analysis?

- Find groups (clusters) of data points such that data points in a group will be similar (or related) to one another and different from (or unrelated to) the data points in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Applications of clustering

- **Understanding**
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations

- **Summarization**
  - Reduce the size of large data sets

Specific Examples—Where Clusters Help

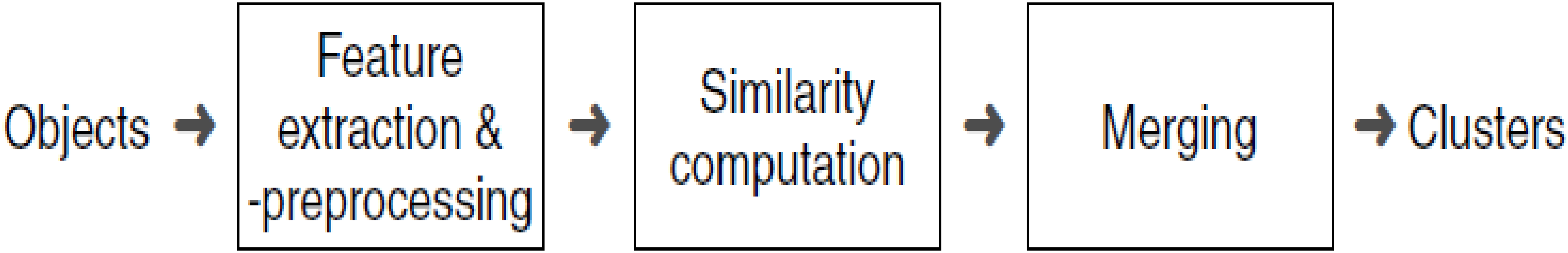
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Spatial Data Analysis: Create thematic maps by identifying areas of similar land use (by clustering feature spaces) in an earth observation dataset
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- Fraud: Identifying groups of individuals that, as a group, are very different to the other groups.
- WWW: Document classification, question categorisation, and web log data to discover similar access patterns.
- City Planning: Identify services for households according to their house type, value, and geographic location.

What Is Good Clustering?

- High Quality:
    - high intra-class similarity
    - low inter-class similarity
- The Quality depends on:
    - similarity measure
    - algorithm for searching

- *Depends on the opinion of the user, and the algorithm's ability to discover hidden patterns that are of interest to the user.*

Main Stages of a Cluster Analysis



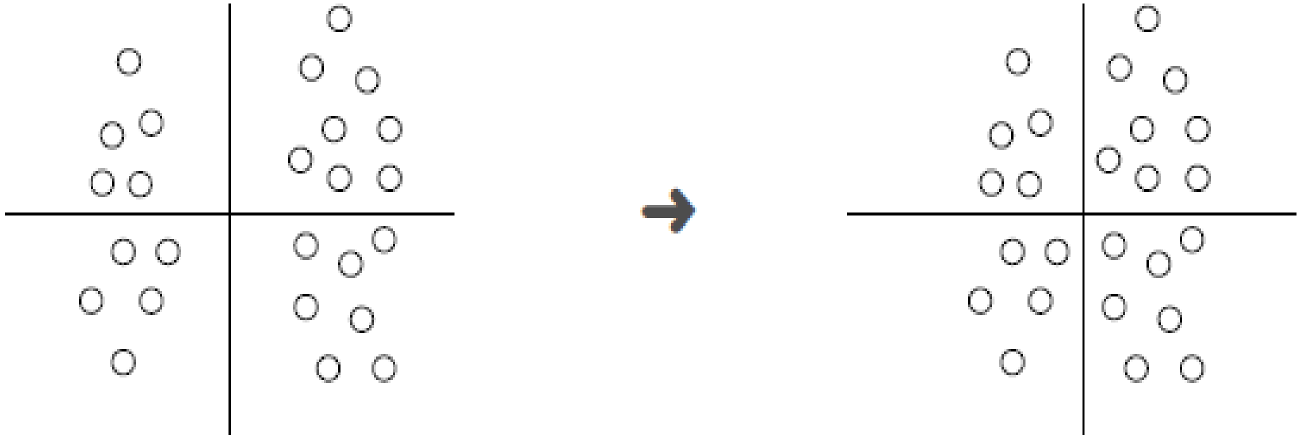
Objects → Feature extraction & -preprocessing → Similarity computation → Merging → Clusters

Feature Extraction and Preprocessing

Required are (possibly new) features of high variance. Approaches:

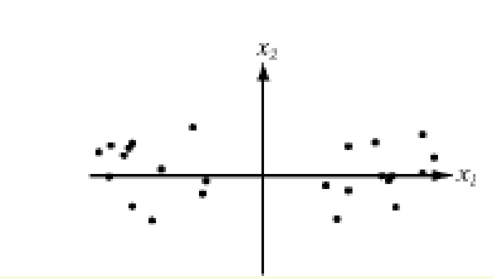
- analysis of dispersion parameters
- dimension reduction: PCA, factor analysis, MDS
- visual inspection: scatter plots, box plots

Feature standardization can dampen the structure and make things worse:

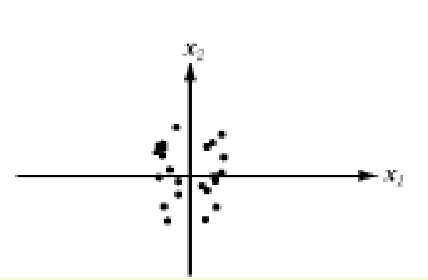


Feature Scale

- old problem: how to choose appropriate relative scale for features?
  - [length (in meters or cms?), weight(in in grams or kgs?)]
  - In supervised learning, can normalize to zero mean unit variance with no problems
  - in clustering this is more problematic, ***if variance in data is due to cluster presence, then normalizing features is not a good thing***



before normalization



after normalization

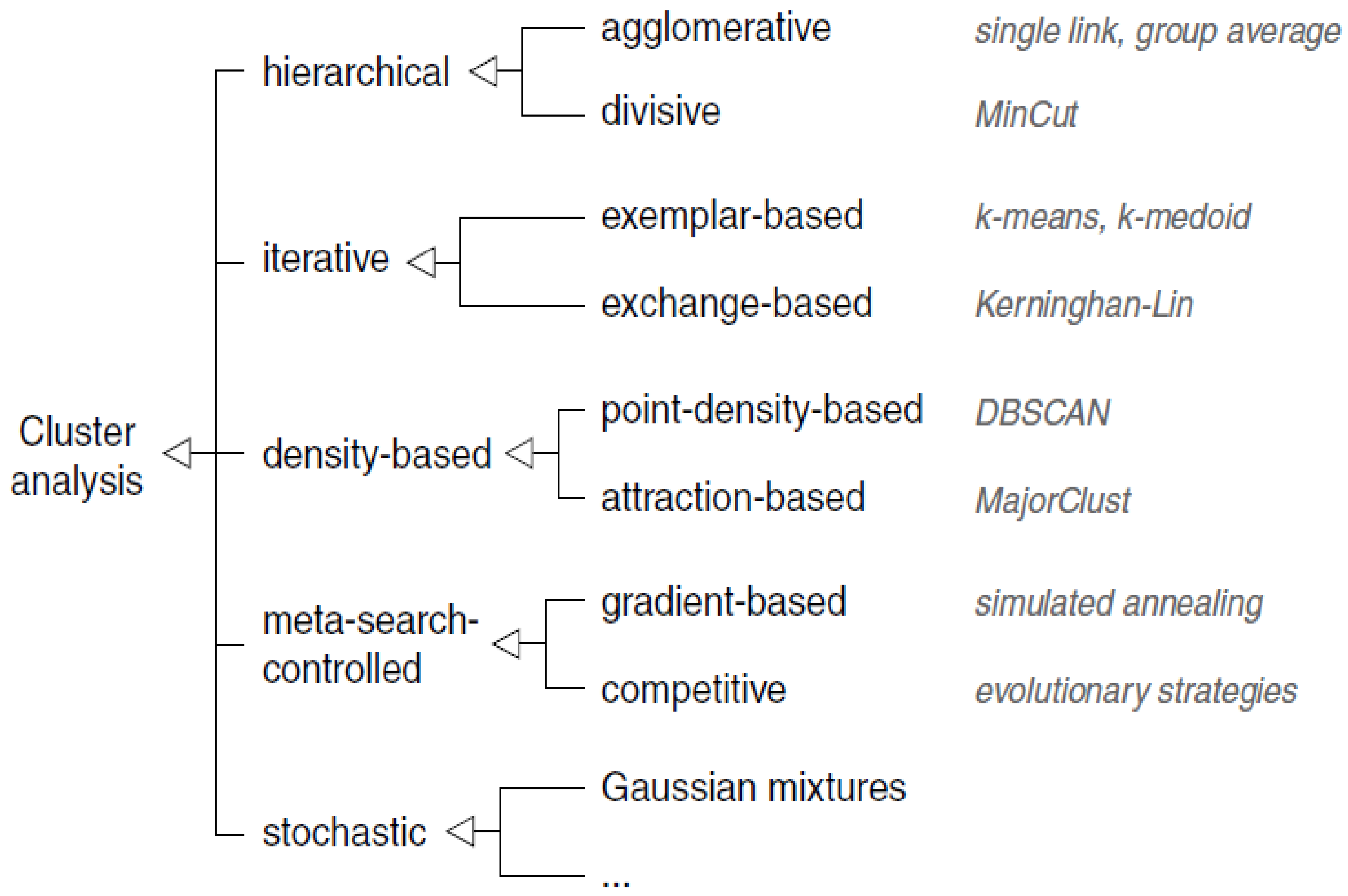
Computation of Distances or Similarities

	Feature 1	Feature 2	. . .	Feature p
$\mathbf{x}_1$	x_{1_1}	x_{1_2}	. . .	x_{1_p}
$\mathbf{x}_2$	x_{2_1}	x_{2_2}	. . .	x_{2_p}
$\vdots$				
$\mathbf{x}_n$	x_{n_1}	x_{n_2}	. . .	x_{n_p}

→

	$\mathbf{x}_1$	$\mathbf{x}_2$	. . .	$\mathbf{x}_n$
$\mathbf{x}_1$	0	$d(\mathbf{x}_1, \mathbf{x}_2)$	. . .	$d(\mathbf{x}_1, \mathbf{x}_n)$
$\mathbf{x}_2$	-	0	. . .	$d(\mathbf{x}_2, \mathbf{x}_n)$
$\vdots$				
$\mathbf{x}_n$	-	-	. . .	0

Merging Principles



Cluster analysis

hierarchical
- agglomerative — *single link, group average*
- divisive — *MinCut*

iterative
- exemplar-based — *k-means, k-medoid*
- exchange-based — *Kerninghan-Lin*

density-based
- point-density-based — *DBSCAN*
- attraction-based — *MajorClust*

meta-search-controlled
- gradient-based — *simulated annealing*
- competitive — *evolutionary strategies*

stochastic
- Gaussian mixtures
- ...