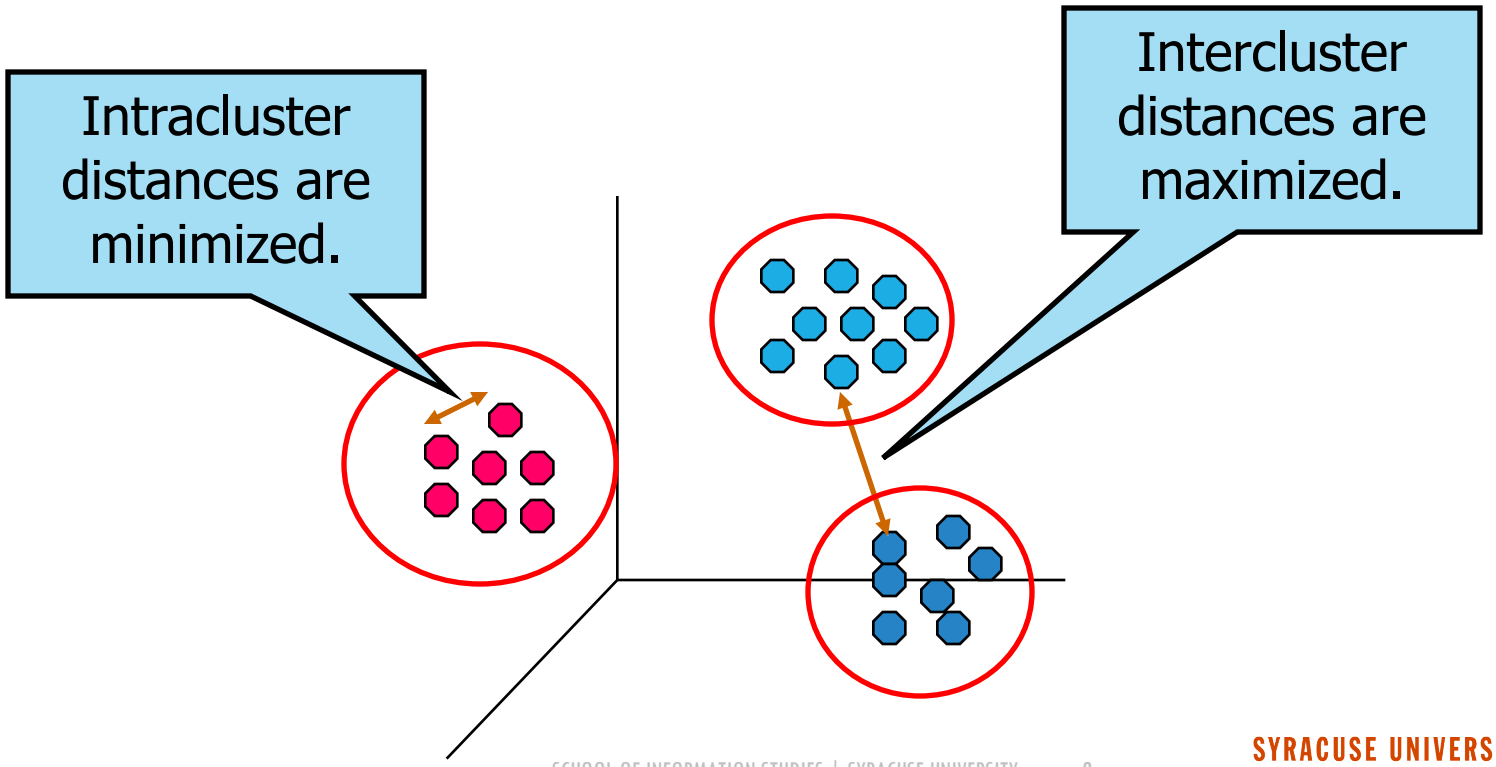
WHAT IS CLUSTERING ANALYSIS?

SYRACUSE UNIVERSITY
School of Information Studies

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



WHAT IS CLUSTER ANALYSIS?

Unsupervised learning: No predefined classes

Typical applications:

- Explore a large data set without prior knowledge about it
  - Customer segmentation, document clustering, etc.
- Classification without training data
  - Usually less accurate than supervised learning methods
- Outlier detection
  - E.g., identify plagiarism cases

REQUIREMENTS OF CLUSTERING IN DATA MINING

Scalability

Ability to explore large data set

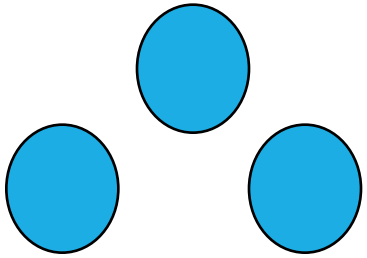
Ability to deal with different types of attributes

Nominal, ordinal, numeric

Discovery of clusters with arbitrary shape

Spherical vs. other shapes

Difficult to cluster because the two clusters are overlapped



Easy to cluster using distance-based methods

REQUIREMENTS OF CLUSTERING IN DATA MINING

Minimal requirements for domain knowledge to determine input parameters

- The number of desired clusters
- E.g., how many topics in congressional speeches

Able to deal with noise and outliers

Insensitive to order of input records

High dimensionality

- Sparse data

Interpretability and usability

TYPES OF CLUSTERINGS

A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

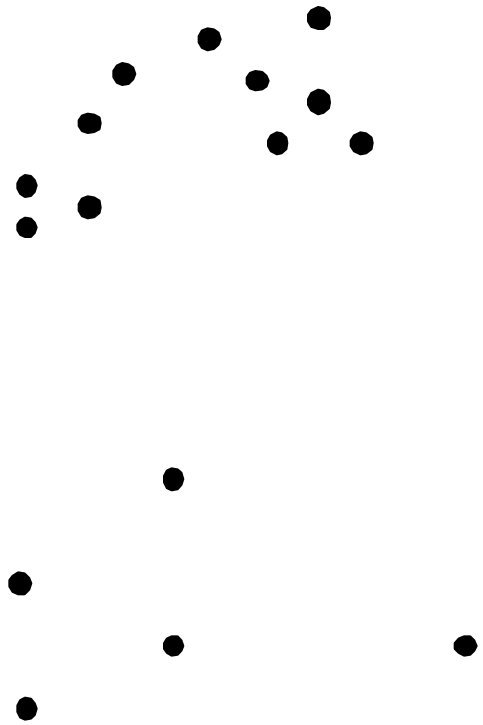
Partitional (flat) clustering:

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

Hierarchical clustering:

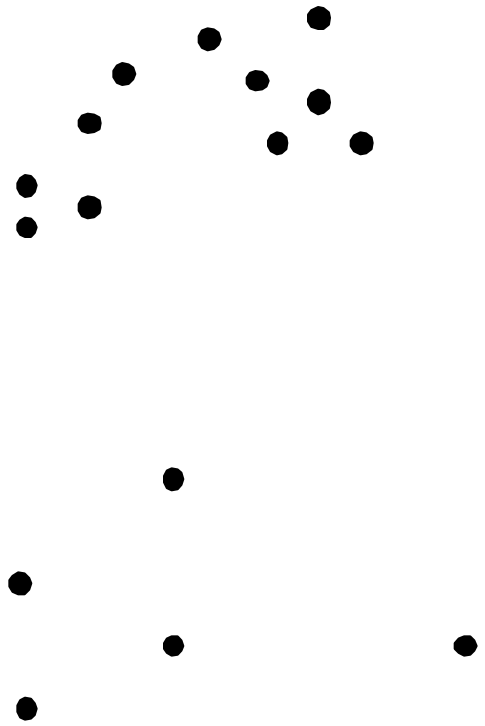
A set of nested clusters organized as a hierarchical tree

PARTITIONAL CLUSTERING

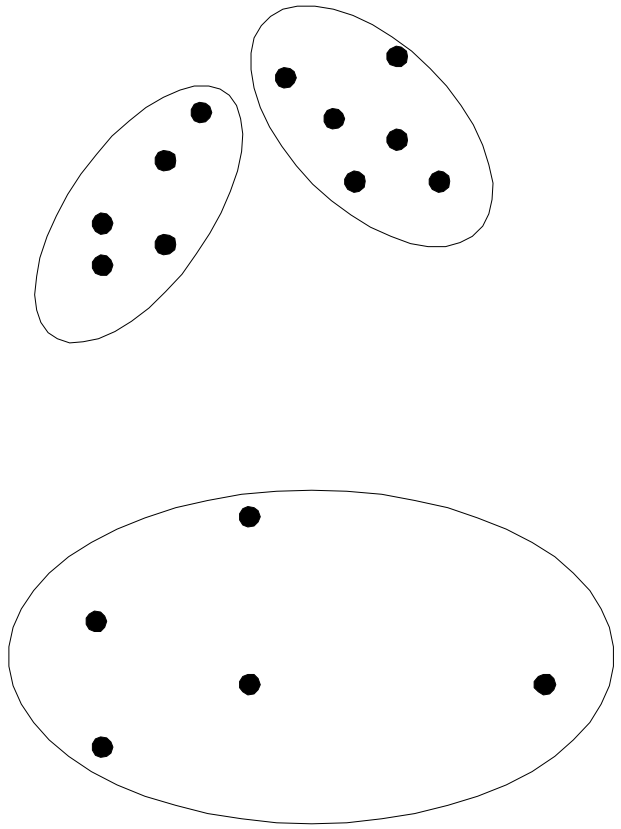


Original Points

PARTITIONAL CLUSTERING

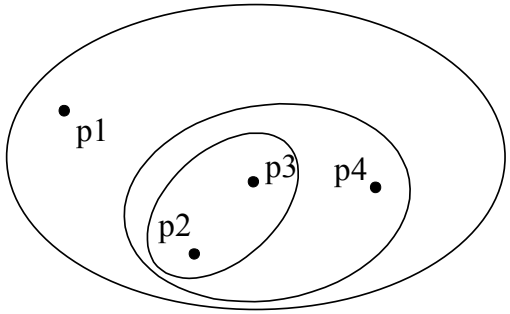


Original Points

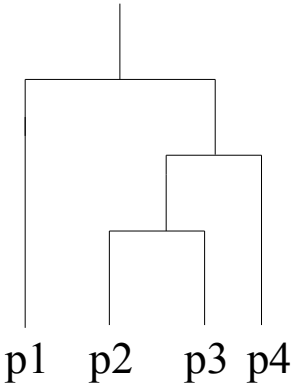


A Partitional Clustering

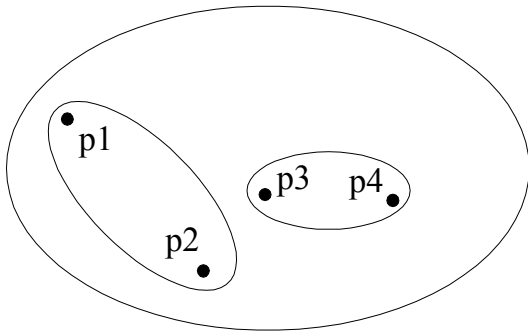
HIERARCHICAL CLUSTERING



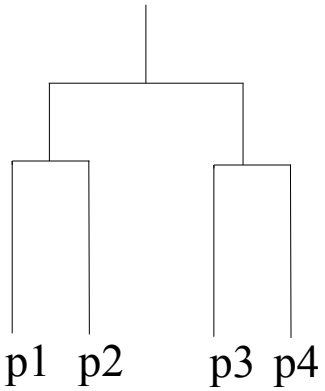
Traditional Hierarchical Clustering



Traditional Dendrogram



Nontraditional Hierarchical Clustering



Nontraditional Dendrogram

MAJOR CLUSTERING APPROACHES

Partitioning approach:

Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.

Typical methods: k-means, k-medoids, CLARANS, EM

Hierarchical approach:

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

Typical methods: DIANA, AGNES, BIRCH, ROCK, CHAMELEON