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
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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Company Names</th>
<th>Industry Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Applied-Matl-DOWN, Bay-Network-DOWN, 3-COM-DOWN,</td>
<td>Technology 1-DOWN</td>
</tr>
</tbody>
</table>

**Clustering precipitation in Australia**
What is not Cluster Analysis?

- **Supervised classification**
  - Have class label information

- **Simple segmentation**
  - Dividing students into different registration groups alphabetically, by last name

- **Results of a query**
  - Groupings are a result of an external specification

- **Graph partitioning**
  - Some mutual relevance and synergy, but areas are not identical
Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters
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
Partitional Clustering

Original Points

A Partitional Clustering
Hierarchical Clustering

Hierarchical Clustering

Dendrogram
Hierarchical Clustering

Dendrogram

Hierarchical Clustering

Source: http://cs.jhu.edu/~razvanm/fs-expedition/tux3.html
Hierarchical Clustering

Example: clustering of file systems from Linux Kernel 2.6.29 + tux3, based on shared external symbols (Hamming distance)

Source: http://cs.jhu.edu/~razvanm/fs-expedition/tux3.html
Other Distinctions Between Sets of Clusters

- **Exclusive versus non-exclusive**
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points

- **Fuzzy versus non-fuzzy**
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics

- **Partial versus complete**
  - In some cases, we only want to cluster some of the data

- **Heterogeneous versus homogeneous**
  - Cluster of widely different sizes, shapes, and densities
Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function
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.
Types of Clusters: Center-Based

- **Center-based**
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster.
  - The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most “representative” point of a cluster.

4 center-based clusters
Types of Clusters: Contiguity-Based

- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

8 contiguous clusters
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.
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
Clusters Defined by an Objective Function

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.
  - Hierarchical clustering algorithms typically have local objectives
  - Partitional algorithms typically have global objectives
- A variation of the global objective function approach is to fit the data to a parameterized model.
  - Parameters for the model are determined from the data.
  - Mixture models assume that the data is a 'mixture' of a number of statistical distributions.
Map the clustering problem to a different domain and solve a related problem in that domain

- Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points

- Clustering is equivalent to breaking the graph into connected components, one for each cluster.

- Want to minimize the edge weight between clusters and maximize the edge weight within clusters
Characteristics of the Input Data Are Important

- **Type of proximity or density measure**
  - This is a derived measure, but central to clustering

- **Sparseness**
  - Dictates type of similarity
  - Adds to efficiency

- **Attribute type**
  - Dictates type of similarity

- **Type of Data**
  - Dictates type of similarity
  - Other characteristics, e.g., autocorrelation

- **Dimensionality**

- **Noise and Outliers**

- **Type of Distribution**
Clustering Algorithms

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