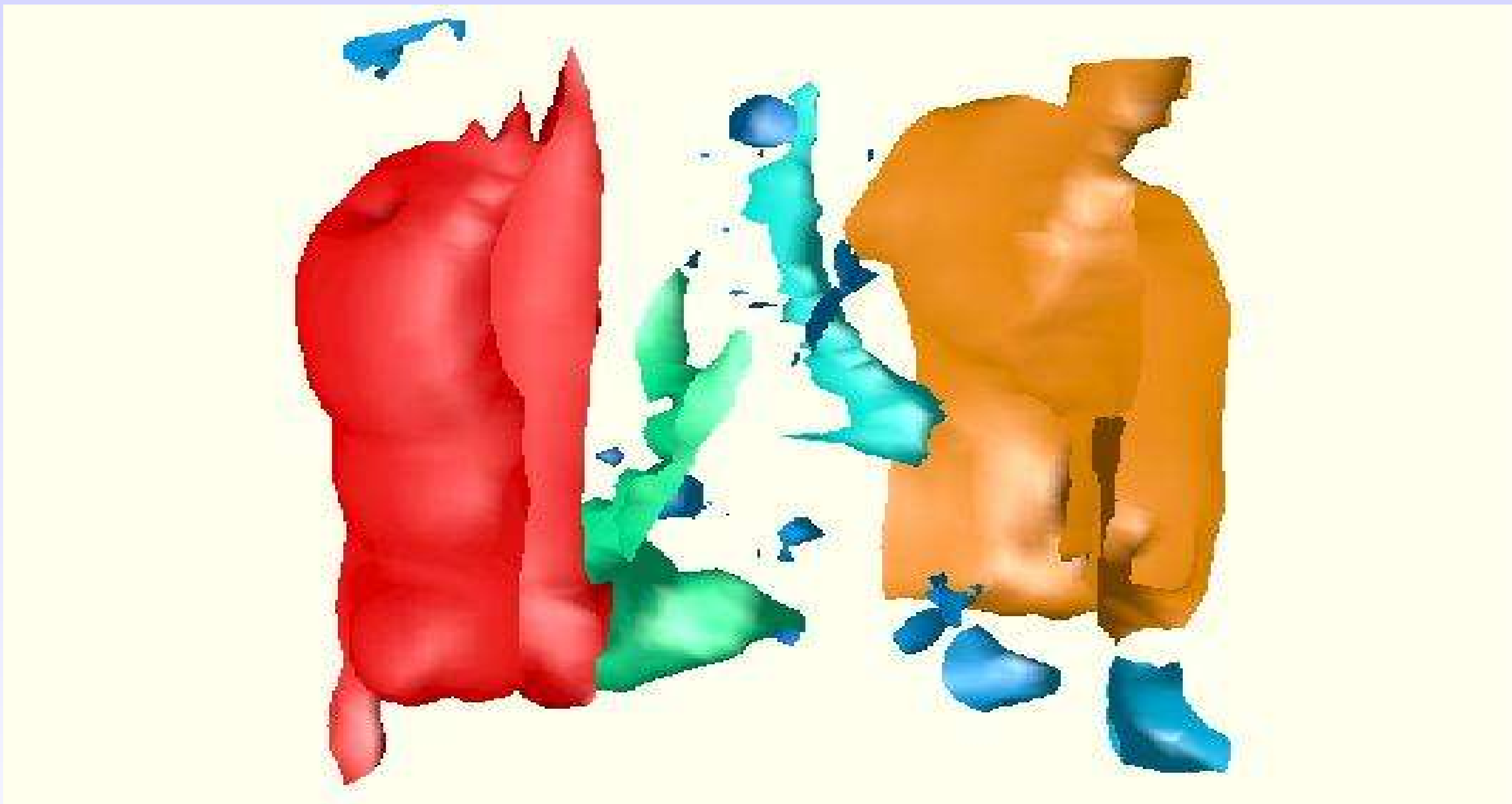
Data Clustering



Joseph Anthony C. Hermocilla

Outline of Topics

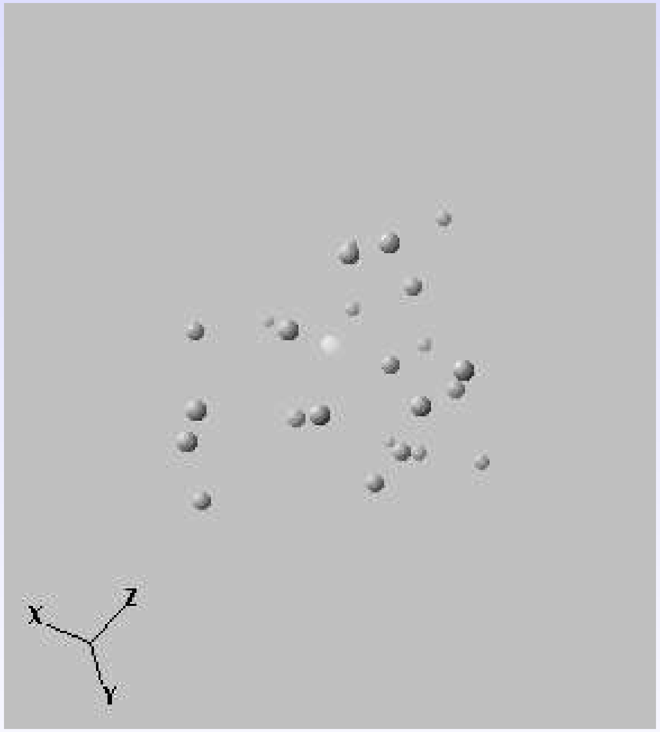
- Introduction to Data Clustering
- Definitions and Notations
- Similarity Measures
- Clustering Techniques
- Applications
- Summary

Introduction to Data Clustering

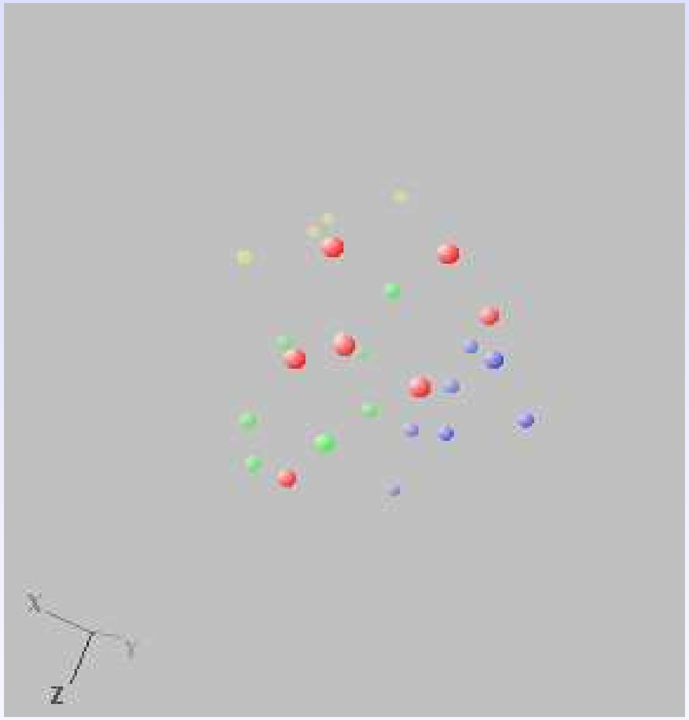
- Data analysis underlies many computing applications
- Either exploratory or confirmatory (based on the availability of approporiate models)
- Common Factor: A grouping or classification of measurements based on (1) goodness-of-fit to a postulated model or (2) natural groupings or clustering relvealed through analysis

Introduction…

- **Cluster Analysis**: The *organization* of a collection of *patterns* into *clusters* based on *similarity*



Raw Data



Clustered Data

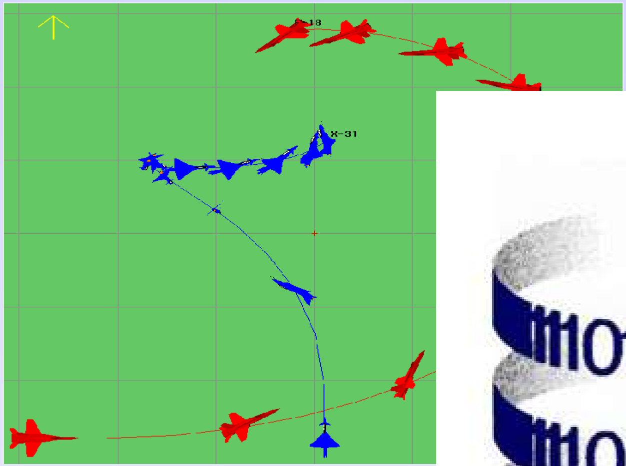
Introduction...

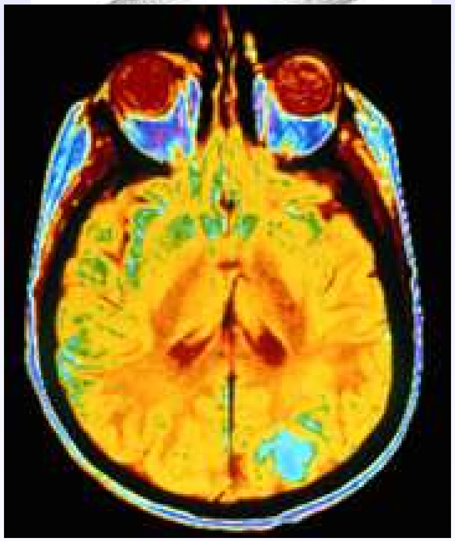
- Clustering (Unsupervised Classification) vs Discriminant Analysis (Supervised Classification)
- Discriminant Analysis: A collection of labeled(training) patterns is given; label a newly encountered, yet unrelated pattern
- Clustering: Group a given collection of unlabeled patterns into meaningful clusters. Labels are obtained solely from the data (data driven)

Training is Learning For Life

Introduction...

- Some applications
  - General Decision-making
  - Machine Learning/Recognition
  - Data Mining
  - General/Web Information Retrieval
  - Image Analysis/Medical Imaging
  - Pattern Classification
  - Genetics/Bioinformatics
  - Finance

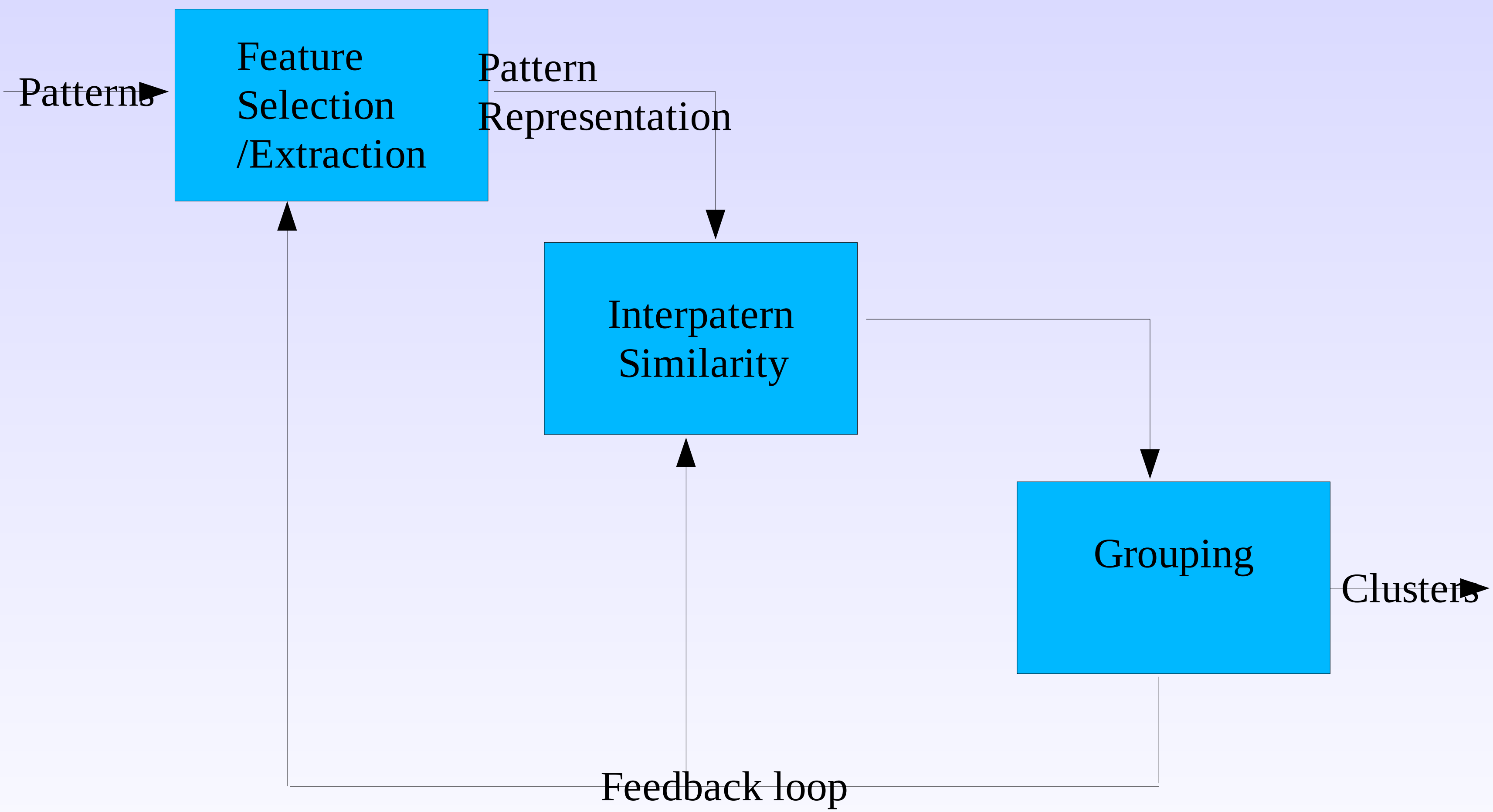




Clustering Tasks

- Pattern Representation (optionally including feature selection and/or extraction)
- Definition of a pattern proximity measure appropriate to the data domain
- Clustering or grouping
- Data abstraction
- Assessment of output

Stages in Clustering



Training is Learning For Life

Some Questions

- How should the data be normalized?
- What similarity measure is appropriate to use?
- How should domain knowledge be utilized?
- How can a very large data set be clustered efficiently?

Definitions

- A *pattern* **x** is a single data item used by the clustering algorithm. It consists of a vector of *d* measurements:
  $\mathbf{x} = (x_1, x_2, x_3, x_4, \ldots, x_d)$
- Inidividual scalar components x_i of the pattern **x** are called *features* (or *attributes*)
- *d* is the dimensionality of the pattern or of the pattern space

Definitions

- A *pattern set* is denoted by $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$. The ith pattern in X is denoted $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$. A pattern set is often referred to as $n \times d$ matrix.
- A *class* is the source of patterns whose distribution in feature space is governed by a probability density specific to the class.

Definitions

- *Hard* clustering techniques attempt to assign a *class label* l_i to each patterns $\mathbf{x}_i$, identifying its class. The set of all labels for a pattern set X is $\mathbf{L} = \{l_1, l_2, \ldots, l_n\}$, with each l_i being an element of the set $\{1, \ldots, k\}$ where k is the number of clusters.
- *Fuzzy* clustering assigns to each input pattern $\mathbf{x}_i$ a fractional degree of membership f_{ij} in each output cluster j.

Definitions

- A *distance measure* (a specialization of a proximity measure) is a metric (or quasi-metric) of the feature space used to quantify the similarity of patterns.

Pattern Represention

- Concerns with the representation of the pattern, basically the selection of the attributes to be included in the pattern
- User's role: gather facts and conjecture about the data
- Perform feature selection and feature extraction
- Feature Selection: Select the features to be used for the pattern
- Feature Extraction: Generate new features from existing features which will then be used for the pattern
- Pattern representation greatly affects clustering results

Pattern Representation

- Patterns are represented as multidimensional vectors where each dimension is a single feature
- Features can be qualitative, quantitative, structured
- Quantitative (can be measured on a ration scale)
  - Continuous (weight)
  - Discrete (units)
  - Interval (event duration)

Pattern Representation

- Qualitative
  - Nominal or unordered (color)
  - Ordinal (ranks)
- Structured
  - Trees
  - Symbolic objects (logical conjunctions of events)

Similarity Measures

- Similarity is fundamental to the definition of cluster
- For continous measures: *Euclidean distance,* a generalization of the Minkowski metric.

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2}$$

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^p\right)^{1/p}$$

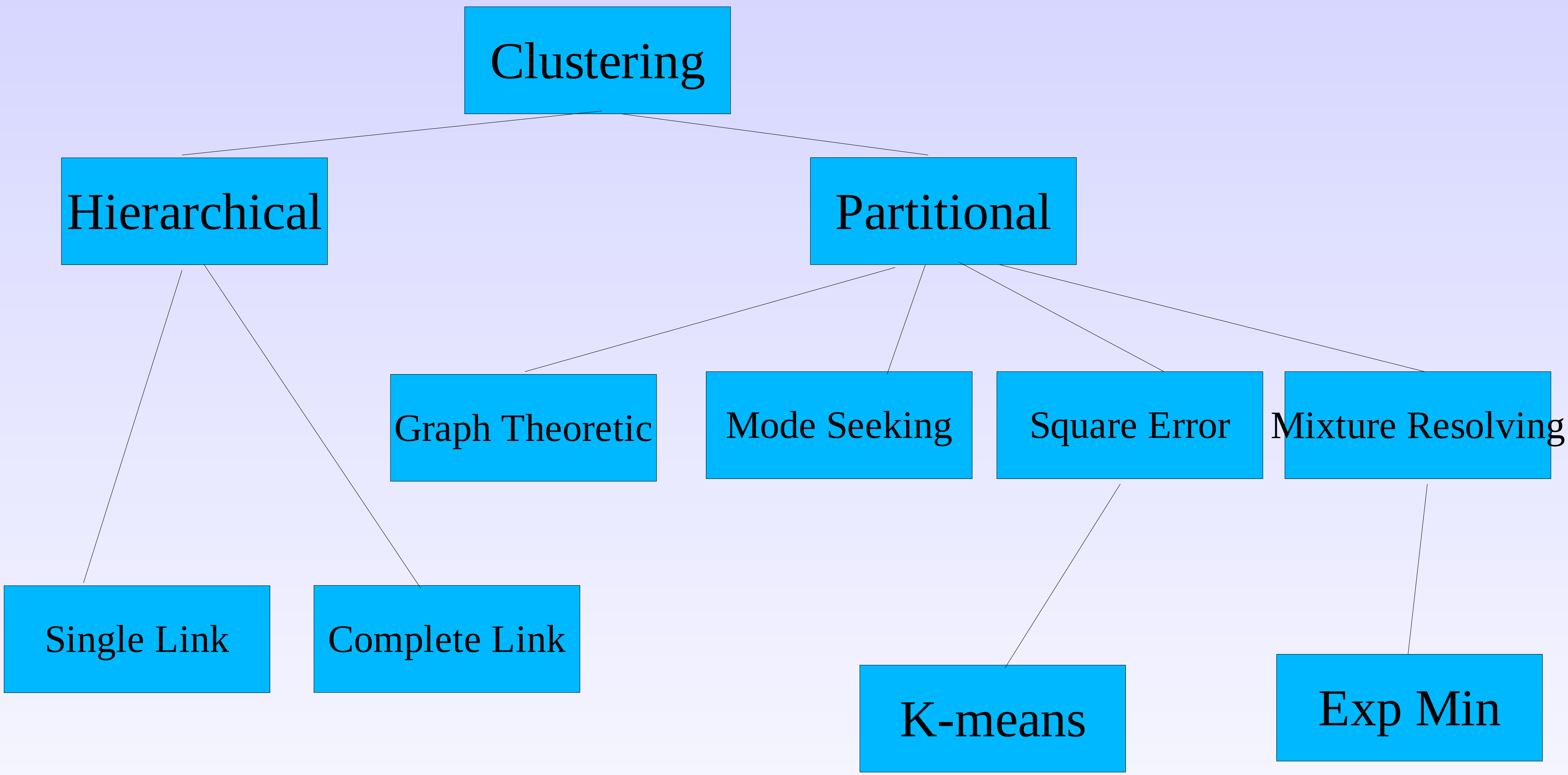
Euclidean Distance

- Works well with with compact or isolated data sets
- Problem: Largest-scaled feature dominate other features
- Solution: Normalize data to a common range or variance
- Linear correlation among features can also distort distance measures
- Some clustering algorithms work on a matrix or proximity values instead of the original pattern set. Precomputation of the n(n-1)/2 pairwise distance values for the n patterns

Heterogeneous Type Patterns

- Some of the features of a pattern may not be continuous
- Mixture of similarity measures have been proposed (Minkowski and population count)
- Patterns can also be represented using strings or trees (syntactic clustering)
- Statistical approaches have been proven to be superior to syntactic approches

Taxonomy of Techniques



Training is Learning For Life

Clustering Techniques

- Agglomerative vs Divisive
- Monothethic vs Polythetic
- Hard vs Fuzzy
- Deterministic vs Stochastic
- Incremental vs Non-incremental

K-means

(1)Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set

(2)Assign each pattern to the closest cluster

(3)Recompute the cluster centers using the current cluster memberships

(4)If a convergence criterion is not met, goto step 2. Typical convergence criteria are: no reassignment of patterns to new cluster centers or minimal increase in squared error

Representation of Clusters

- Data abstraction of clustering results
- Using the centroid or a set of distant points in the cluster
- Use nodes in a classification tree
- Use conjunctive logical expressions

Advanced Techniques

- Artificial Neural Networks (Kohonen Networks, Self Organizing Maps (SOM))
- Fuzzy Clustering
- Evolutionarty Algorithms(Evolutionary Programming (EP), Genetic Algorithms (GA), Evolutionary Strategies (ES)
- Simulated Annealing Techniques

Comparison of Techniques

- The clustering problem as an optimization problem
- Evolutionary techniques are globalized search techniques
- ANNs and GAs are parallel in nature (can take advantage of hardware configuration)
- Evolutionary techniques are population based: search using more than one solution at a time
- Evolutionary techniques are generally weak (no explicit use of domain knowldge)

Comparison of Techniques

- K-means and Kohonen networks have been applied to large data sets
- Obtaining suitable learning and control parameters is difficult for evolutionary approaches and ANN with high execution times
- Combining domain knowledge provides improvement

Complexity

• K-means	$O(nkl)$	$O(k)$
• ISODATA	$O(nkl)$	$O(k)$
• SSP	$O(nkl)$	$O(n)$
• Single line	$O(n^2 \log n)$	$O(n^2)$
• Complete Line	$O(n^2 \log n)$	$O(n^2)$

References

- Aguila, J. *SCLUST*, Special Problem. April 2003.
- *Data Clustering: A Review*. ACM Computing Surveys, Vol. 31, No. 3, Septermber 1999.