Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

6.873 / HST.951 Medical Decision Support
Spring 2005

Unsupervised Learning

An overview of clustering and other exploratory data analysis methods

Lucila Ohno-Machado

A few “synonyms”...

- Agminatics
- Aciniformics
- Q-analysis
- Botryology
- Systematics
- Taximetrics
- Clumping
- Morphometrics
- Nosography
- Nosology
- Numerical taxonomy
- Typology
- Clustering
- A multidimensional space needs to be reduced...

Supervised Models

	age	test1	
Case 1	0.7	-0.2	0.8
Case 2	0.6	0.5	0.4
	-0.6	0.1	0.2
	0	-0.9	0.3
	-0.4	0.4	0.2
	-0.8	0.6	0.3
	0.5	-0.7	0.4

Using these

we predict probability of diagnosis, prognosis

We are chasing PARTICULAR patterns in the data...

Evaluate against "gold standard"

Unsupervised Models

	age	test1	Cluster
Case 1	0.7	1	1
Case 2	0.6	0.5	1
	-0.6	0.1	2
	0	-0.9	3
	-0.4	0.4	2
	-0.8	0.6	2
	0.5	-0.7	3

We are chasing ANY pattern in the data...

We will need to interpret (label) the pattern

Using these

we put cases into clusters

Exploratory Data Analysis

- Goal is to flatten the dimensions of data to the spaces that we are familiar with (2-D and 3-D)
- We can "see" the data in these dimensions and extract patterns

- We are looking for clusters of data with similar characteristics overall
- Hypothesis generation versus hypothesis testing
- Fishing expedition versus confirmatory analysis

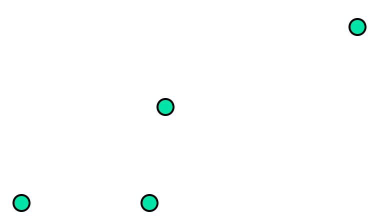
Outline

- Proximity
  - Distance Metrics
  - Similarity Measures
- Clustering
  - Hierarchical Clustering
    - Agglomerative
  - K-means
- Multidimensional Scaling

Spatial relations

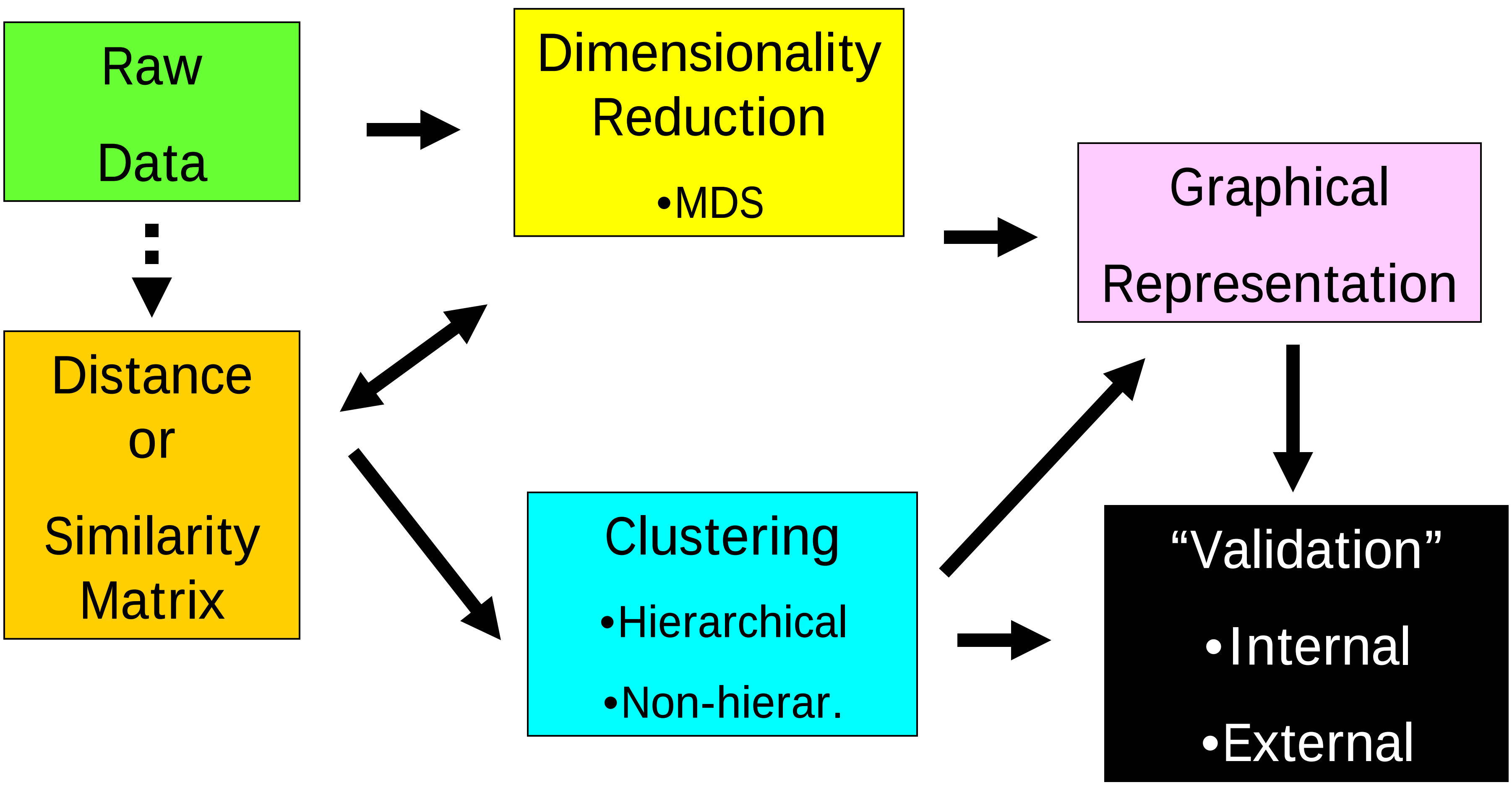
- Distance and dissimilarity
  - E.g. Euclidean distance, perceived difference
- Proximity and similarity measures
  - E.g. correlation coefficient

Distance matrix



	House	Harvard	MIT	BWH
House				
Harvard	15			
MIT	18	4		
BWH	10	3	5	

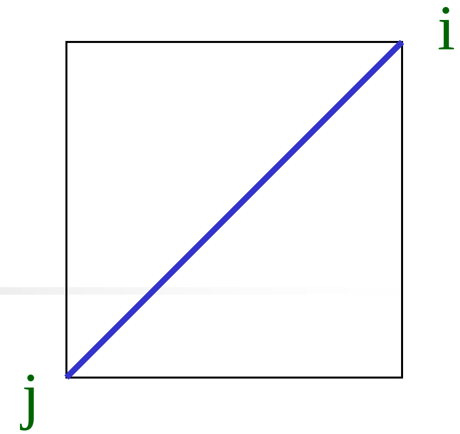
Unsupervised Learning



Algorithms, (dis)similarity measures, and graphical representations

- Most algorithms are not necessarily linked to a particular metric or (dis)similarity measure
- Also not necessarily linked to a particular graphical representation
- Cluster techniques were popular in the 50/60s (psychology experiments)
- There has been recent interest in biomedicine because of the emergence of high throughput technologies
- Old algorithms have been rediscovered and renamed

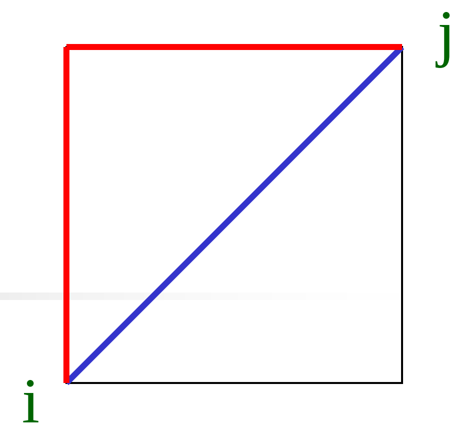
Metrics (distances)



- K dimensional data
- Euclidean

$$d_{ij} = \left\{ \sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^2 \right\}^{1/2}$$

Minkowski r-metric



- K dimensional data
- Euclidean $$d_{ij} = \left\{ \sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^2 \right\}^{1/2}$$
- Manhattan $$d_{ij} = \left\{ \sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^1 \right\}^{1/1}$$
  - (city-block)
- Generalized $$d_{ij} = \left\{ \sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^r \right\}^{1/r}$$

Metric spaces

- Positivity
  Reflexivity

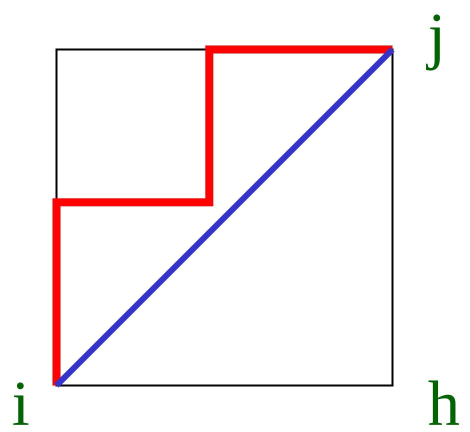
$$d_{ij} > d_{ii} = 0$$

- Symmetry

$$d_{ij} = d_{ji}$$

- Triangle
  inequality

$$d_{ij} \leq d_{ih} + d_{hj}$$

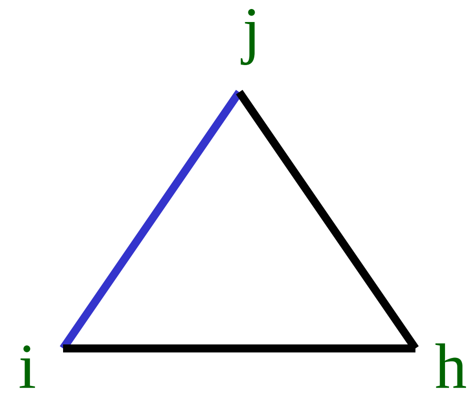


More metrics

- Ultrametric $d_{ij} \leq \max\lfloor d_{ih}, d_{hj} \rfloor$

  replaces

  $d_{ij} \leq d_{ih} + d_{hj}$



- Four-point additive condition $d_{hi} + d_{jk} \leq \max\lfloor (d_{hj} + d_{ik}), (d_{hk} + d_{ij}) \rfloor$

  replaces

  $d_{ij} \leq d_{ih} + d_{hj}$

Similarity measures

- Similarity function
  - For binary, “shared attributes”

$$s(i,j) = \frac{i^t j}{\|i\| \|j\|}$$

$$s(i,j) = \frac{1}{\sqrt{2 \times 1}}$$

i^t = [1,0,1]

j^t = [0,0,1]

Variations...

- Fraction of *d* attributes shared

$$s(i,j) = \frac{i^t j}{d}$$

- Tanimoto coefficient

$$s(i,j) = \frac{i^t j}{i^t i + j^t j - i^t j}$$

$$s(i,j) = \frac{1}{2+1-1}$$

i^t = [1,0,1]

j^t = [0,0,1]

Popular similarity measures

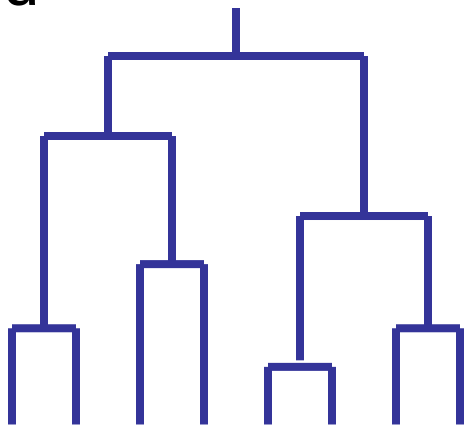
- Correlation
  - Linear
  - Rank
- Entropy-based
  - Mutual information, based on the P(i|j)
- Ad-hoc
  - Human perception

Clustering

Hierarchical Clustering

- Agglomerative Technique (average link)
  - Step 1: “Merge” 2 closest cases into a cluster
  - Step 2: Define cluster representative (e.g. , cluster means) as a “case” and remove the individual cases that compose the cluster
  - Go to step 1 until all cases are linked

- Visualization
  - Dendrogram, Tree, Venn diagram



Data Visualization

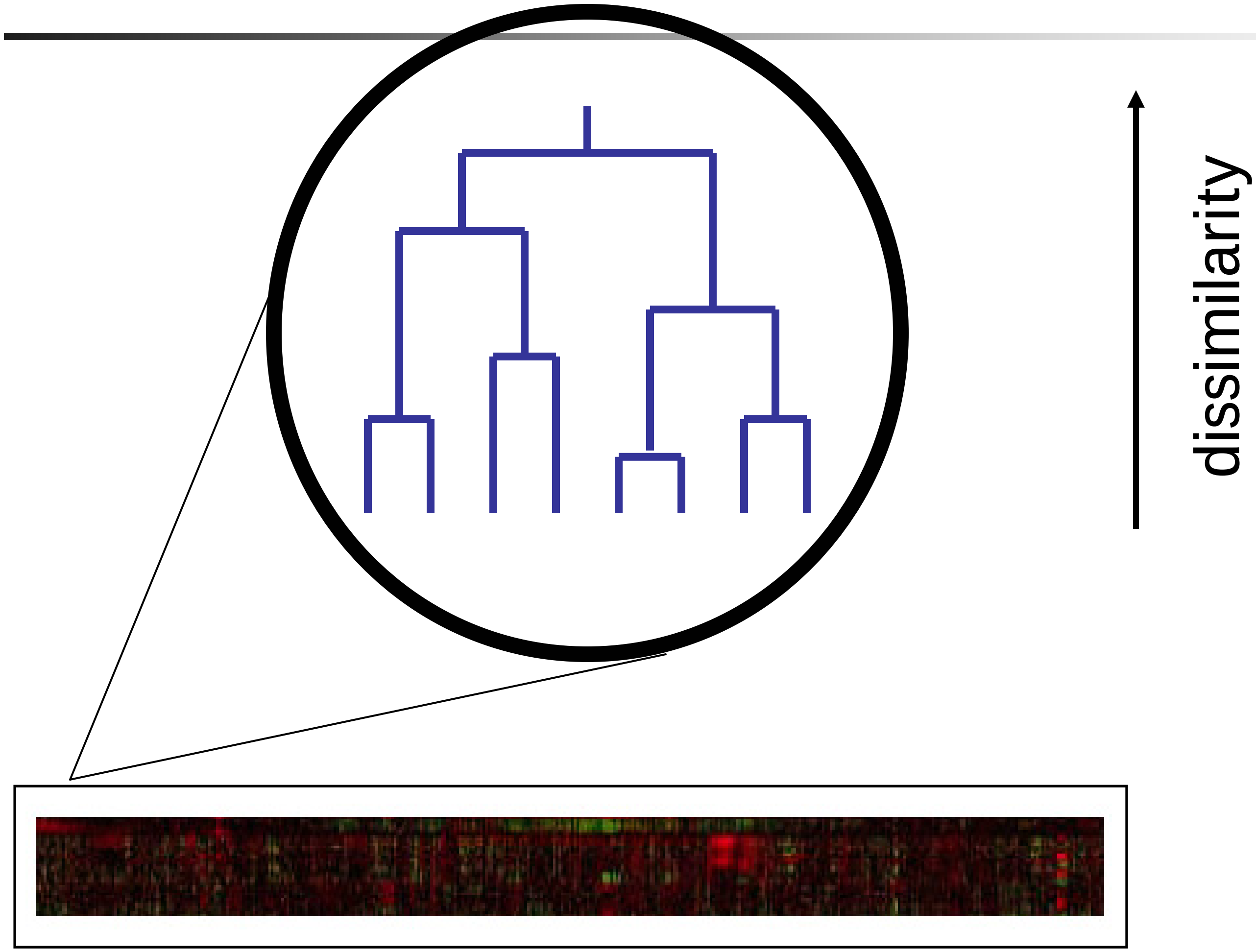
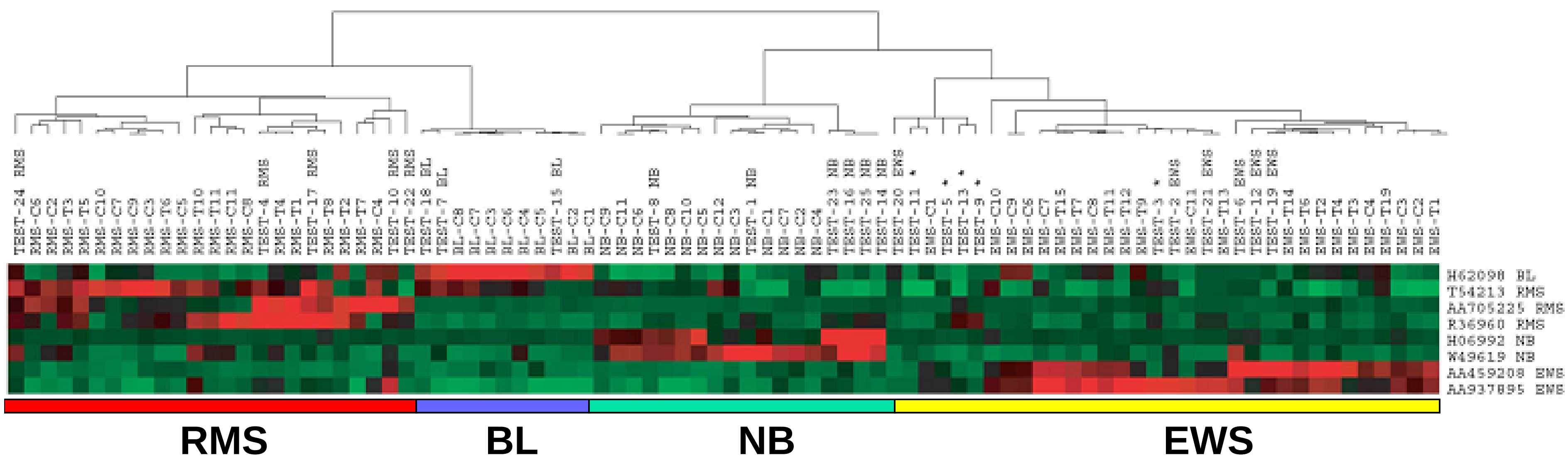


Figure by MIT OCW.

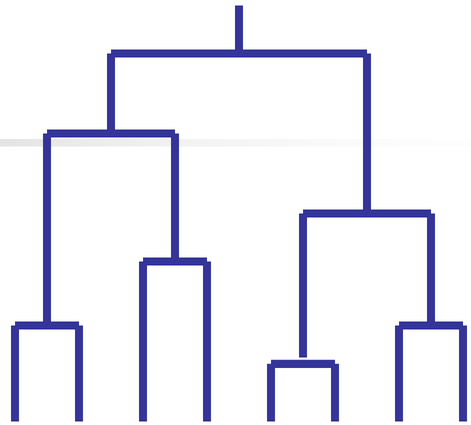
Hierarchical Clustering on Small Round Blood Cell Tumours



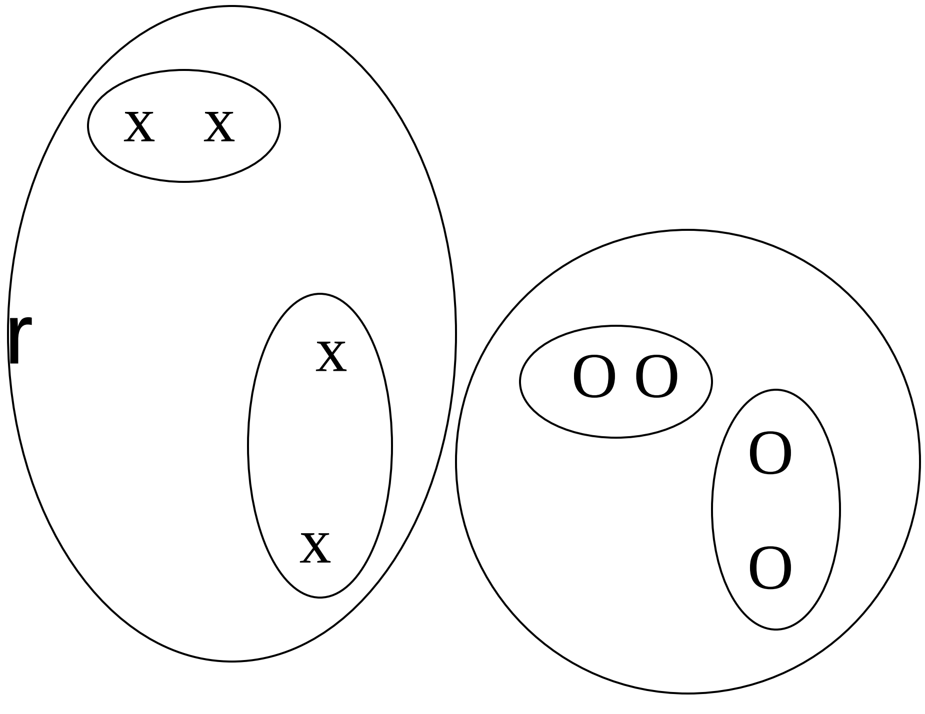
Linkages

- Average-linkage: proximity to the mean (centroid)
- Single-linkage: proximity to the closest element in another cluster
- Complete-linkage: proximity to the most distant element

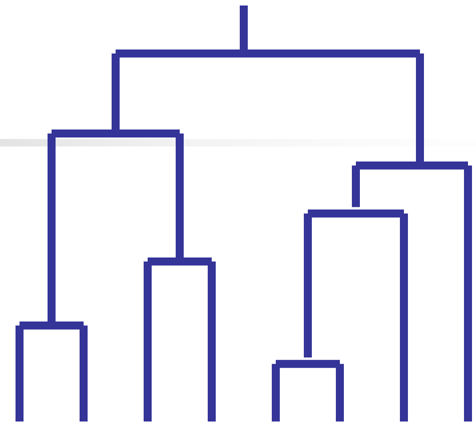
Mean Linkage



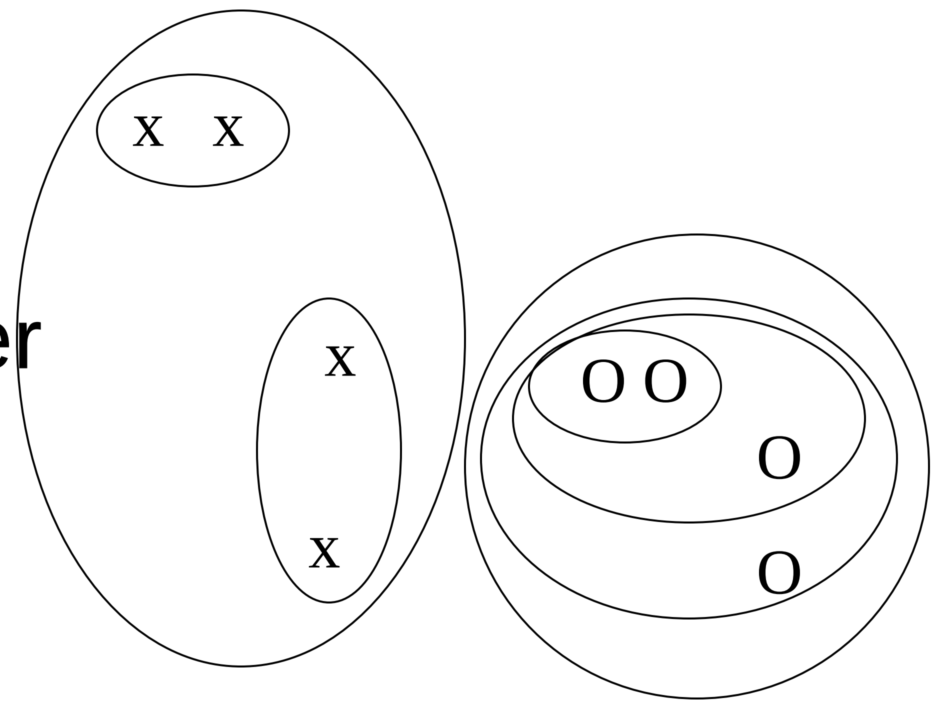
- Assign case according to proximity to the mean (centroid) of another cluster



Single Linkage

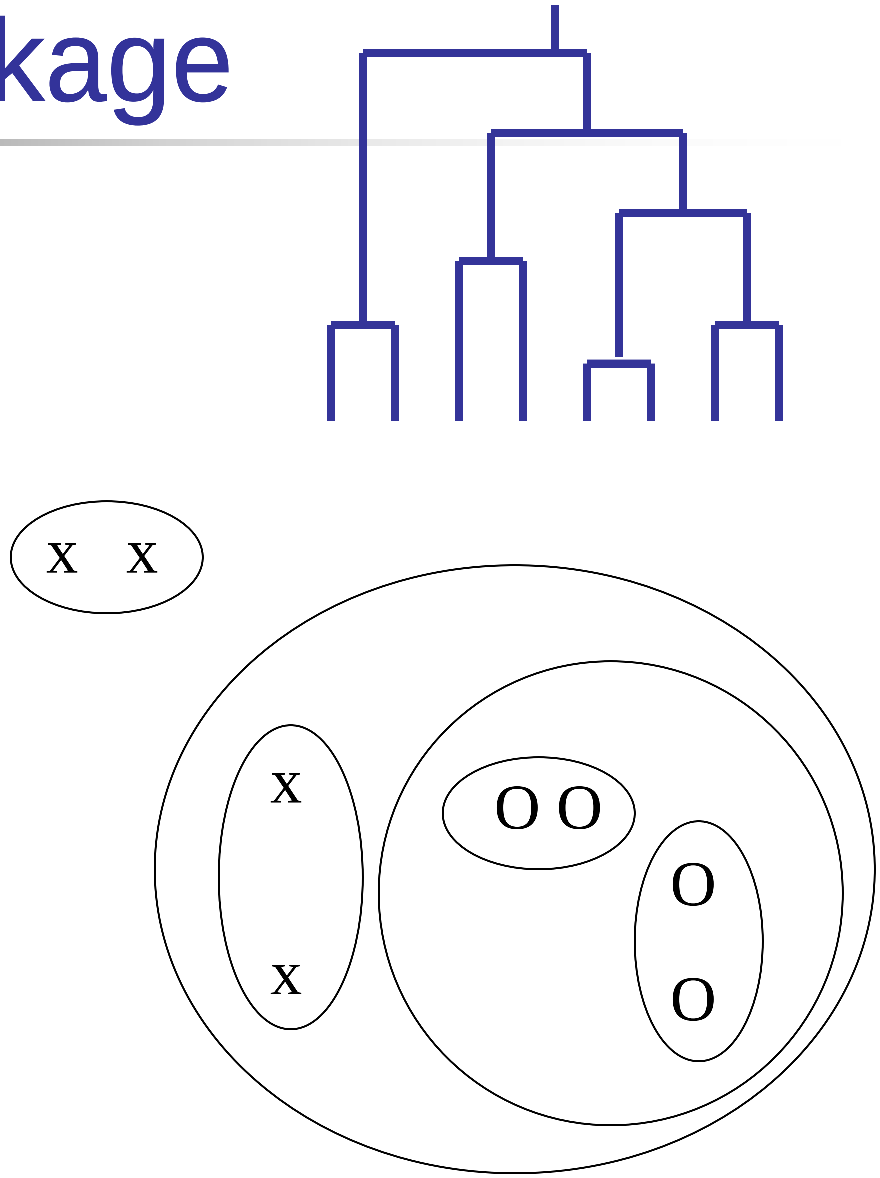


- Assign case according to proximity to the closest element in another cluster



Complete Linkage

- Assign case according to proximity to the most distant element



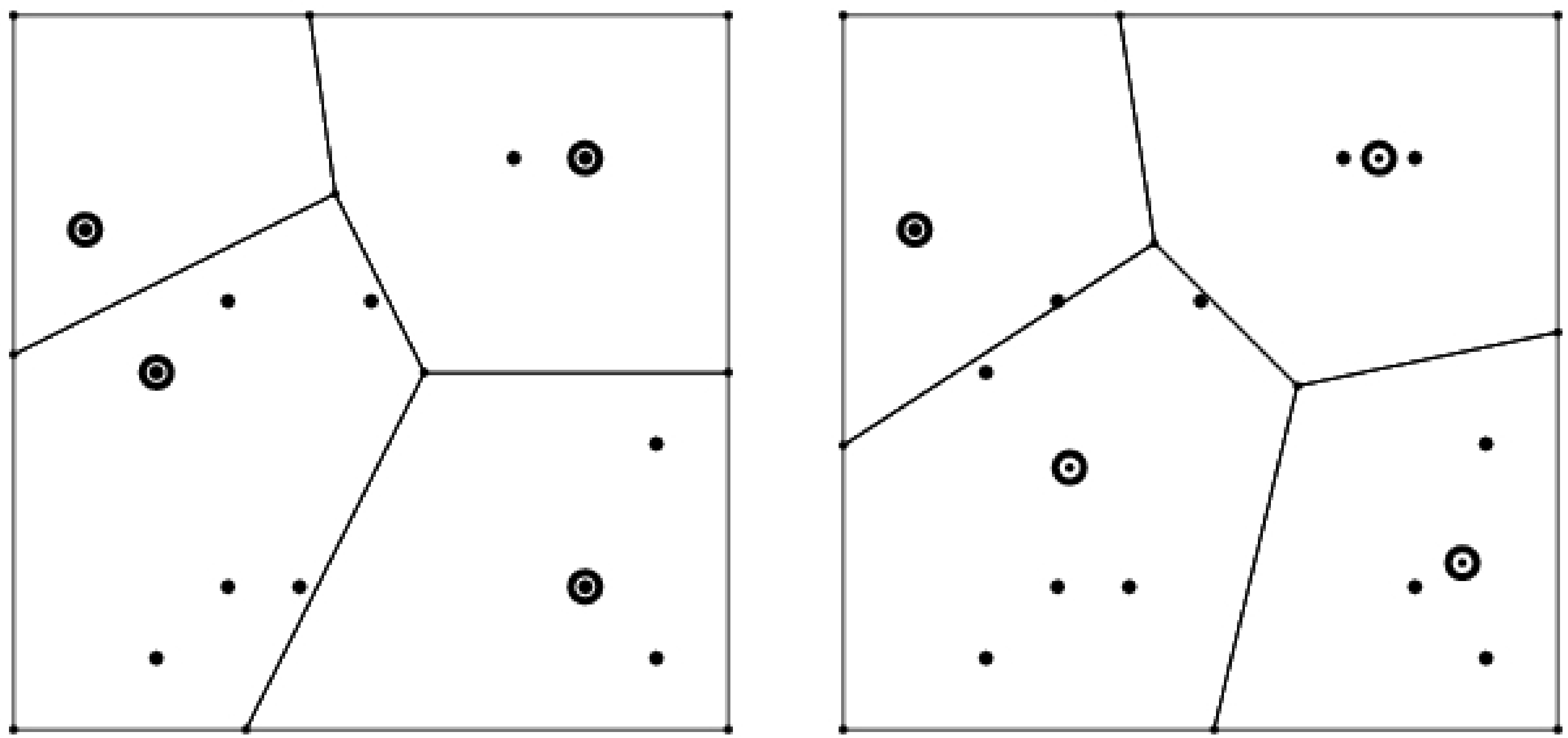
Additive Trees

- Commonly the minimum spanning tree
- Nearest neighbor approach to hierarchical clustering

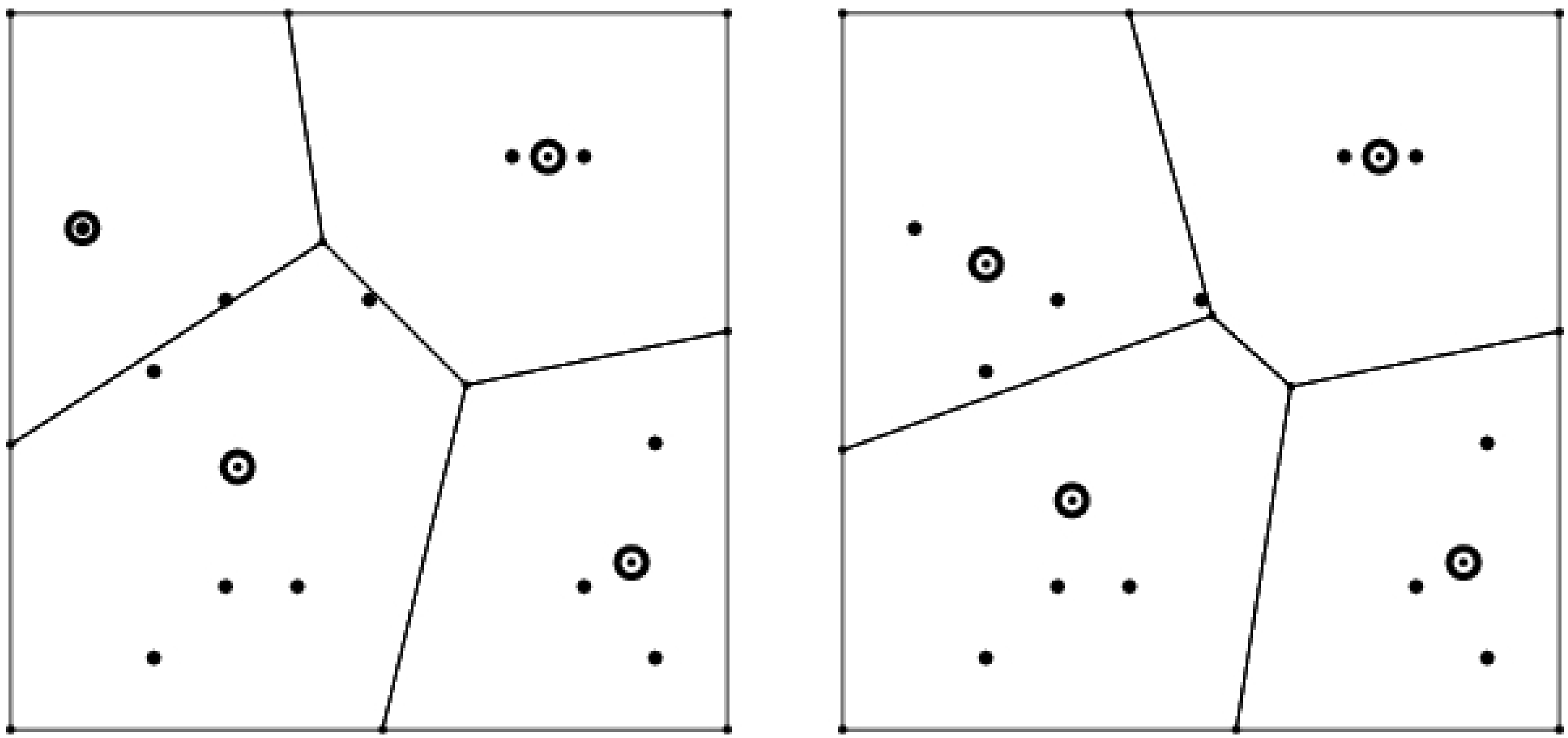
k-means clustering (Lloyd's algorithm)

1. Select k (number of clusters)
2. Select k initial cluster centers $c_1, \ldots, c_k$
3. Iterate until convergence: For each i,
    1. Determine data vectors $v_{i1}, \ldots, v_{in}$ closest to c_i (i.e., partition space)
    2. Update c_i as $c_i = 1/n\,(v_{i1} + \ldots + v_{in})$

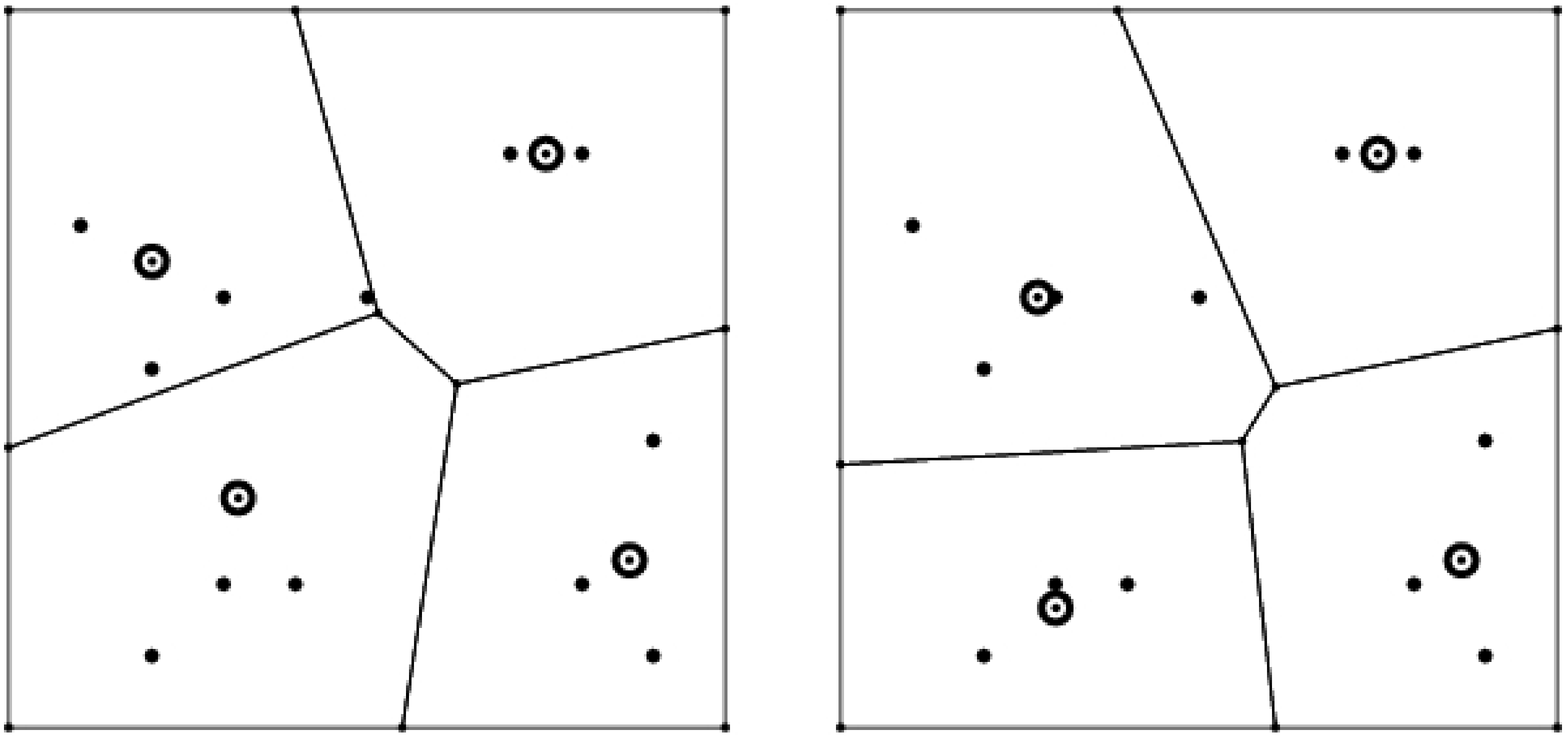
k-means clustering example



k-means clustering example



k-means clustering example



Common mistakes

- Refer to dendrograms as meaning “hierarchical clustering” in general
- Misinterpretation of tree-like graphical representations
- Ill definition of clustering criterion
  - Declare a clustering algorithm as “best”
- Expect classification model from clusters
- Expect robust results with little/poor data

Dimensionality Reduction

Multidimensional Scaling

- Geometrical models
- Uncover structure or pattern in observed proximity matrix
- Objective is to determine both dimensionality *d* and the position of points in the *d*-dimensional space

Classic Multidimensional Scaling

- Also known as principal coordinates analysis (because it is principal components analysis) ☺
- From distances, find coordinates
- Constrain origin to centroid of data

Metric and non-metric MDS

- Metric (Torgerson 1952)
- Non-metric (Shepard 1961)
  - Estimates nonlinear form of the monotonic function

$$s_{ij} = f_{mon}(d_{ij})$$

Stress and goodness-of-fit

Stress	Goodness of fit
20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

Figures removed due to copyright reasons.

Please see:

Khan, J., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nat Med* 7, no. 6 (Jun 2001): 673-9.

Visualization

- Clustering is often good for visualization, but it is generally not very useful to separate data into pre-defined categories
- But there are counterexamples...

Figures removed due to copyright reasons.

Please see:

Khan, J., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nat Med* 7, no. 6 (Jun 2001): 673-9.

Figures removed due to copyright reasons.

Please see:

Khan, J., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nat Med* 7, no. 6 (Jun 2001): 673-9.