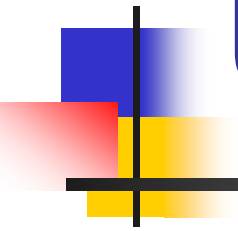


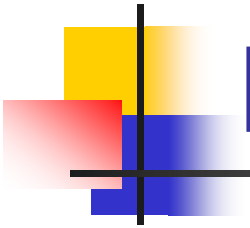
Unsupervised Learning

A little knowledge…

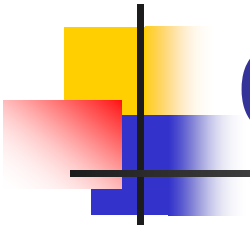
A few “synonyms”…

- Agminatics
- Aciniformics
- Q-analysis
- Botryology
- Systematics
- Taximetrics
- Clumping
- Morphometrics
- Nosography
- Nosology
- Numerical taxonomy
- Typology
- Clustering

Exploratory Data Analysis



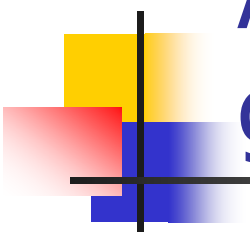
- Visualization methods with little or no numerical manipulation
- Dimensionality Reduction
  - Multidimensional Scaling
  - Principal Components Analysis, Factor Analysis
  - Self-Organizing Maps
- Cluster Analysis
  - Hierarchical
    - Agglomerative
    - Divisive
  - Non-hierarchical



Outline

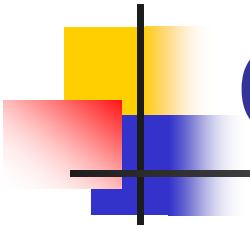
- Proximity
  - Distance Metrics
  - Similarity Measures
- Multidimensional Scaling
- Clustering
  - Hierarchical Clustering
    - Agglomerative
  - Criterion Functions for Clustering
- Graphical Representations

Algorithms, similarity measures, and graphical representations



- Most algorithms are not necessarily linked to a particular metric or similarity measure
- Also not necessarily linked to a particular graphical representation

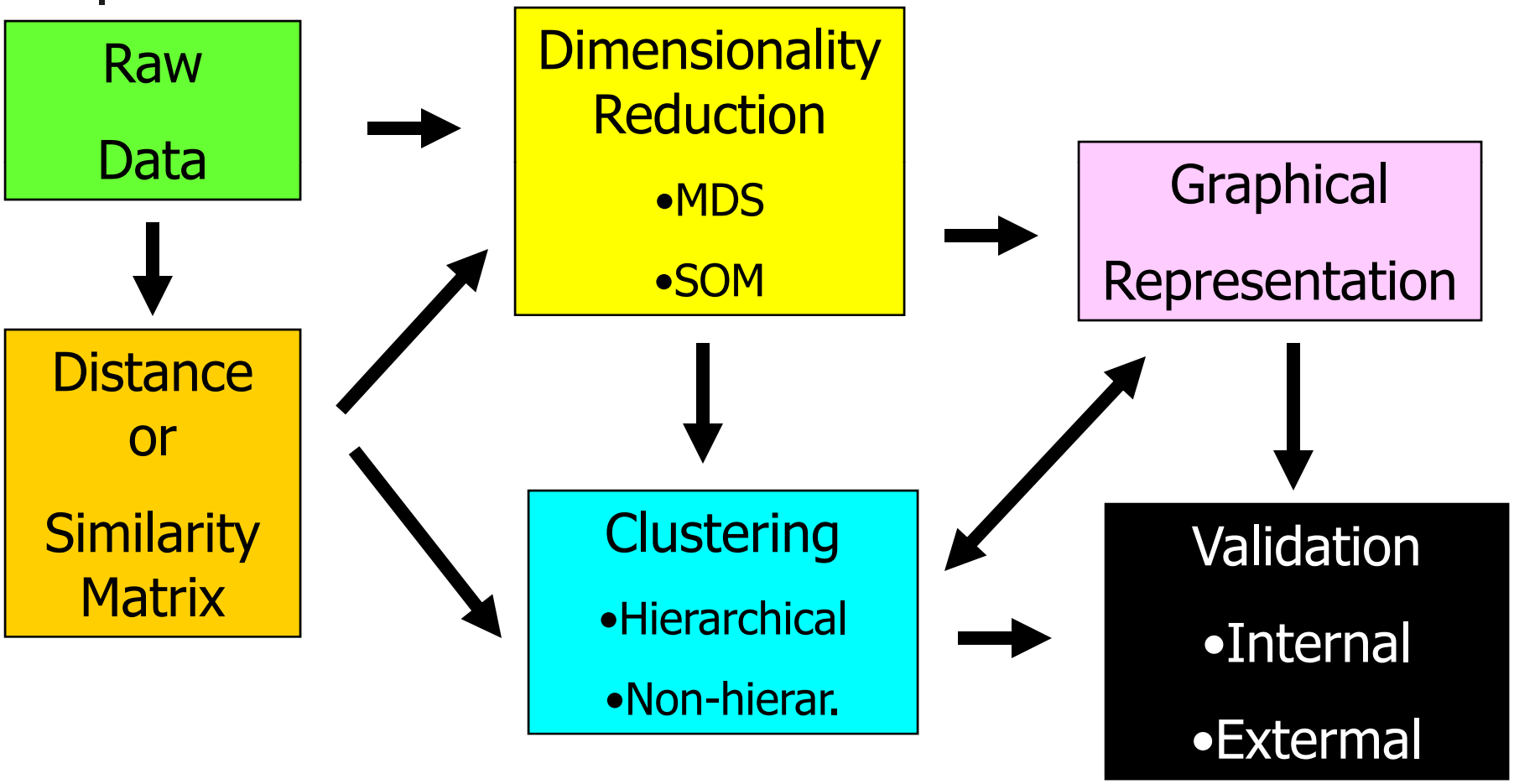
- Be aware of the fact that old algorithms are being reused under new names!

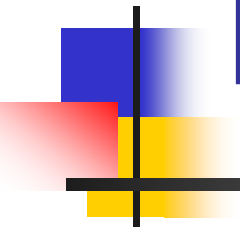


Common mistakes

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

Unsupervised Learning





Metrics

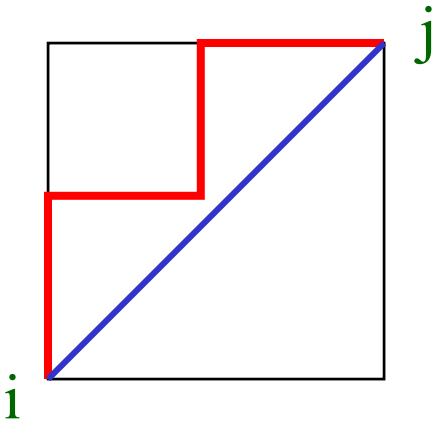
Minkowski r-metric

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

- Manhattan
  - (city-block)

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

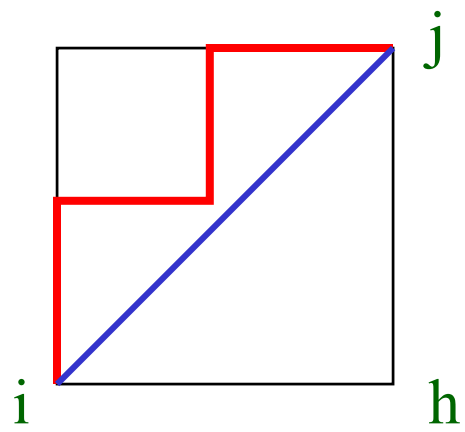
- Euclidean



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

Metric spaces

- Positivity
  Reflexivity $d_{ij} > d_{ii} = 0$
- Symmetry $d_{ij} = d_{ji}$
- Triangle
  inequality $d_{ij} \leq d_{ih} + d_{hj}$

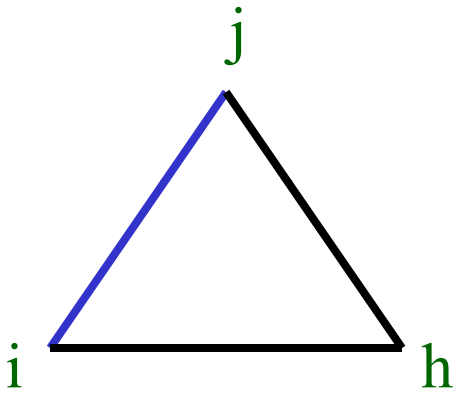


More metrics

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

  replaces

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



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

  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 = [1,0,1]

j = [0,0,1]

Variations…

- Fraction of d attributes shared

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

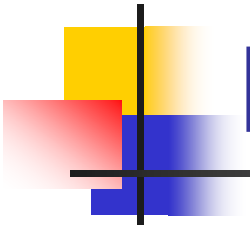
- Tanimoto coefficient

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

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

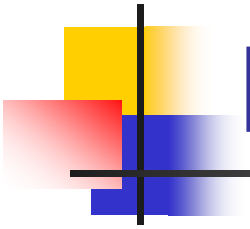
i = [1,0,1]

j = [0,0,1]

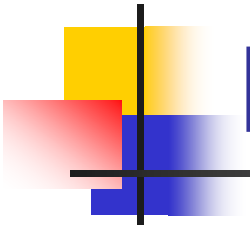


More variations…

- Correlation
  - Linear
  - Rank
- Entropy-based
  - Mutual information
- Ad-hoc
  - Neural networks



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

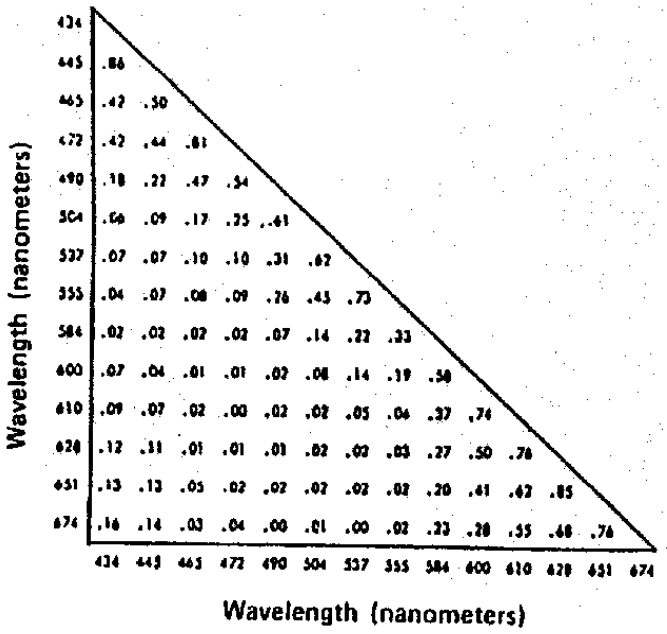
Metric and non-metric MDS

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

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

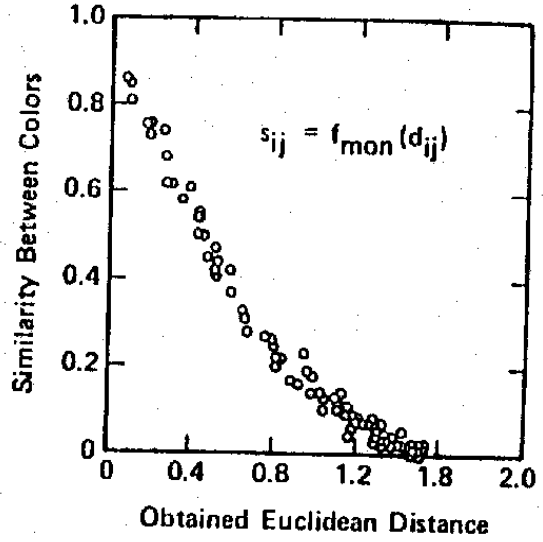
Similarity Data

Judged similarities between 14 spectral colors varying in wavelength from 434 to 674 nanometers (from Ekman, 1954).



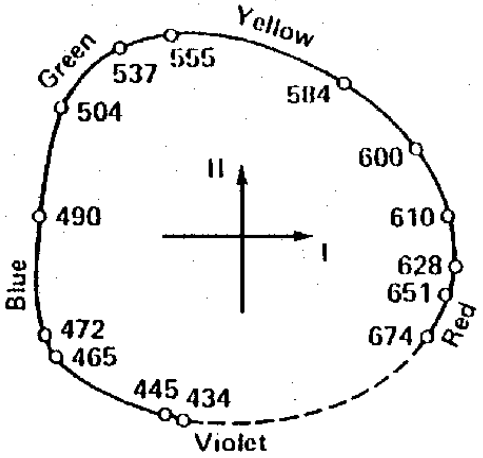
Relation of Data to Spatial Representation

Obtained relation between Ekman's original similarity data for the 14 colors and the Euclidean distances in Shepard's spatial solution.

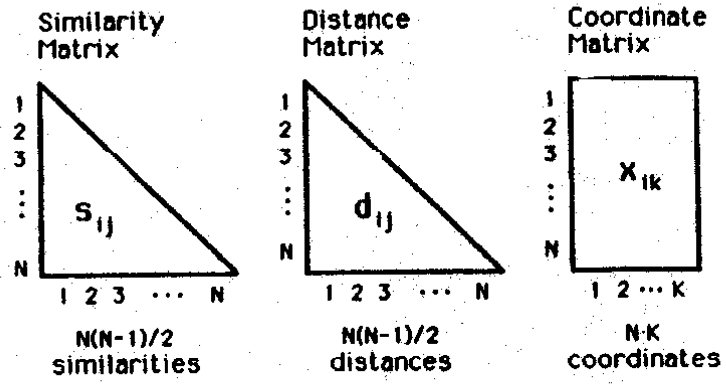


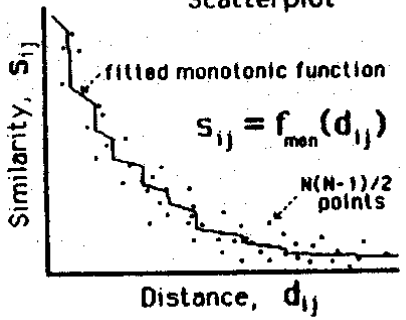
Spatial Representation

Two-dimensional spatial solution for the 14 colors obtained by Shepard (1962) on the basis of Ekman's (1954) similarity data.



Multidimensional Scaling Schema



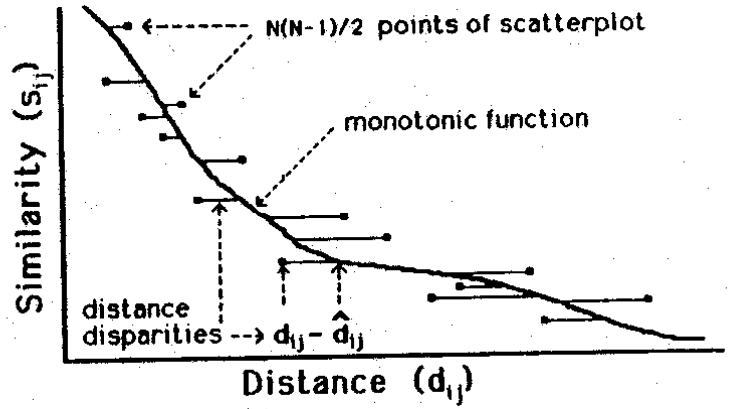


$$d_{ij} = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2}$$

x_{ik} is the coordinate for object i on dimension k

$i = 1, 2, \cdots N \qquad k = 1, 2, \cdots K$

Quantifying Departure From Monotonicity



Departure from the monotonic function is defined in terms of a (normalized) sum of the squared distance disparities:

$$\text{Stress} = \sqrt{\frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

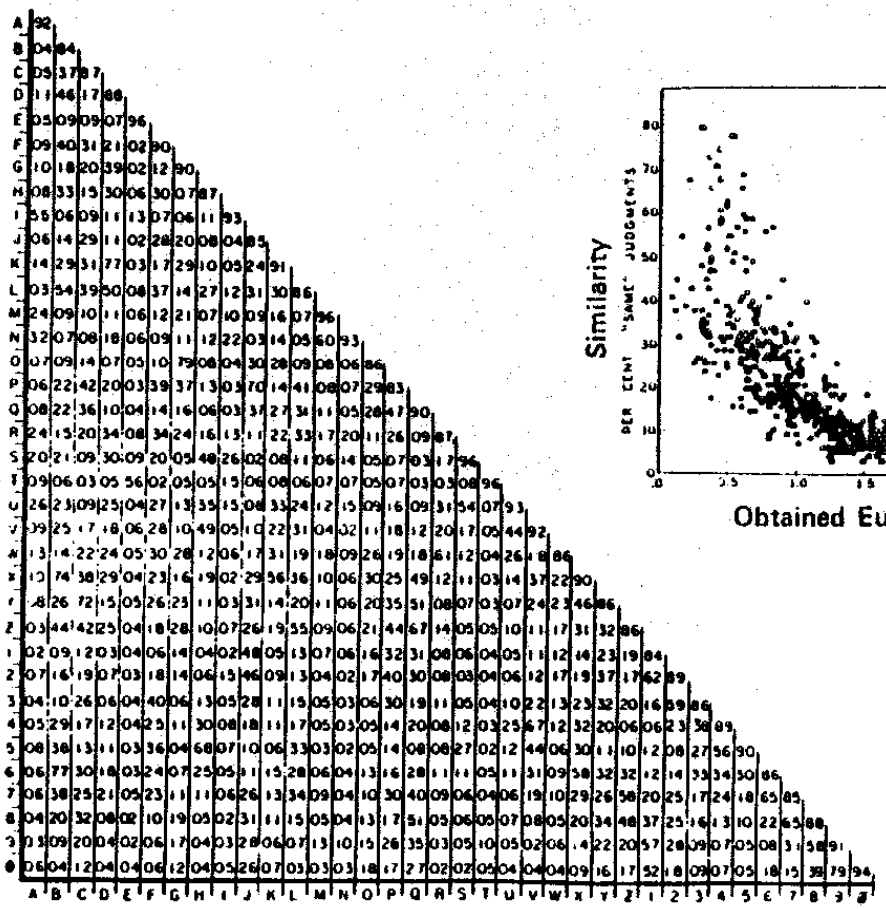
1. The square root simply compensates for the squares (of the distances and distance disparities) and thus renders stress distance-like; literally, stress is the distance from perfect monotonicity.

2. The normalization (division by the sum of squared distances) makes stress independent of the overall size of the configuration, so that stress cannot be reduced to zero merely by shrinking all distances.

3. Because stress is based on distance disparities only, it is invariant under any monotonic (order-preserving) transformations of the similarity data.

Similarity Data

Percent "same" judgments for all pairs of successively presented aural signals of the International Morse Code (from Rothkopf, 1957).

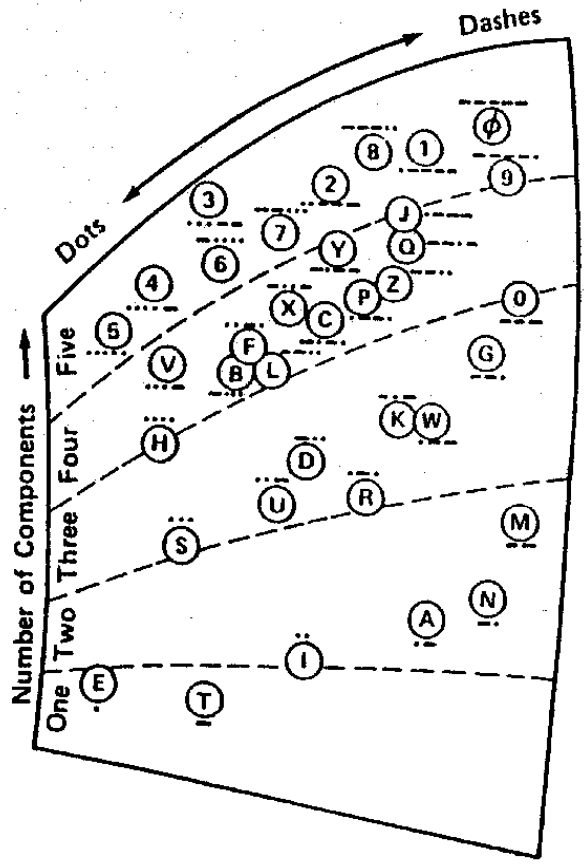


Relation of Data to Spatial Representation

Obtained relation between Rothkopf's original similarity data for the 36 Morse Code signals and the Euclidean distances in Shepard's spatial solution.

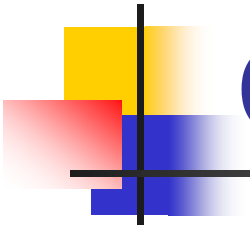
Spatial Representation

Two-dimensional spatial solution for the 36 Morse Code signals obtained by Shepard (1963) on the basis of Rothkopf's (1957) similarity data.



Stress and goodness-of-fit

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



Clustering

Non-Hierarchical: Distance threshold

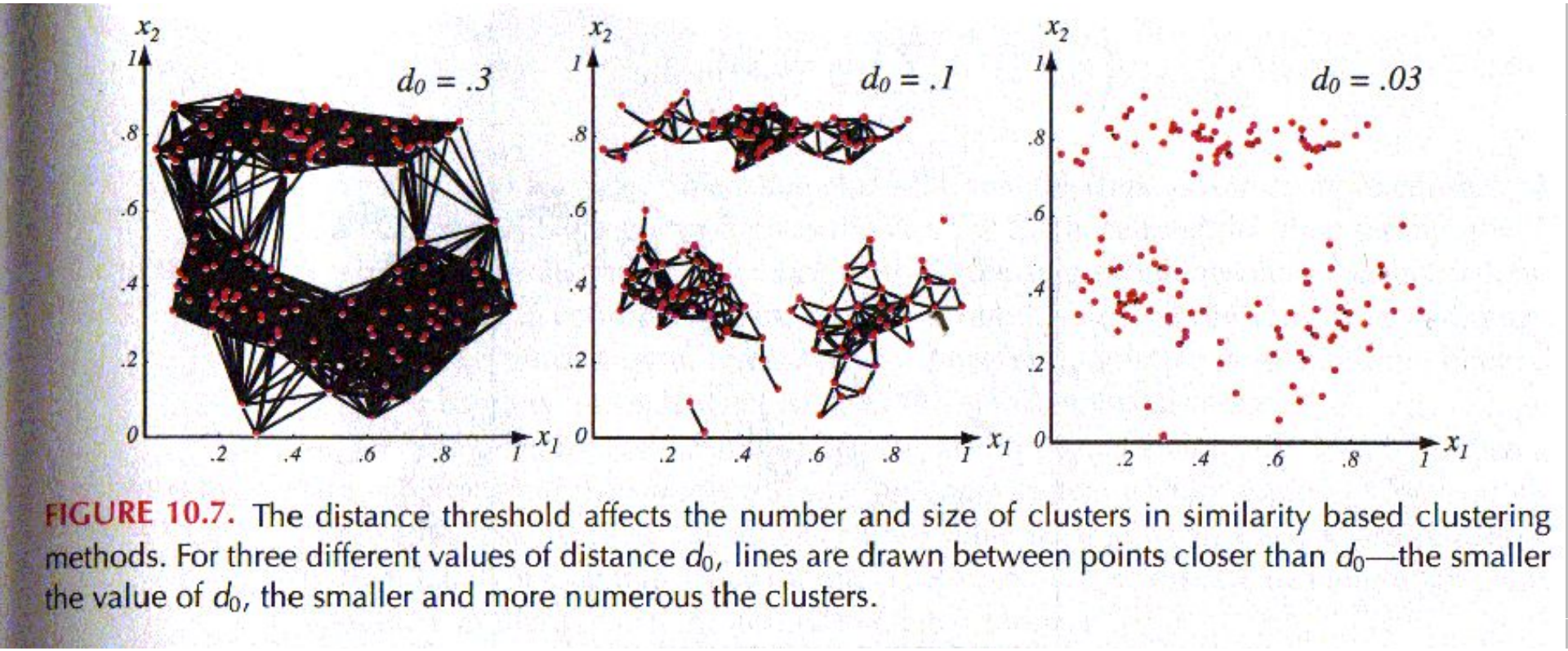
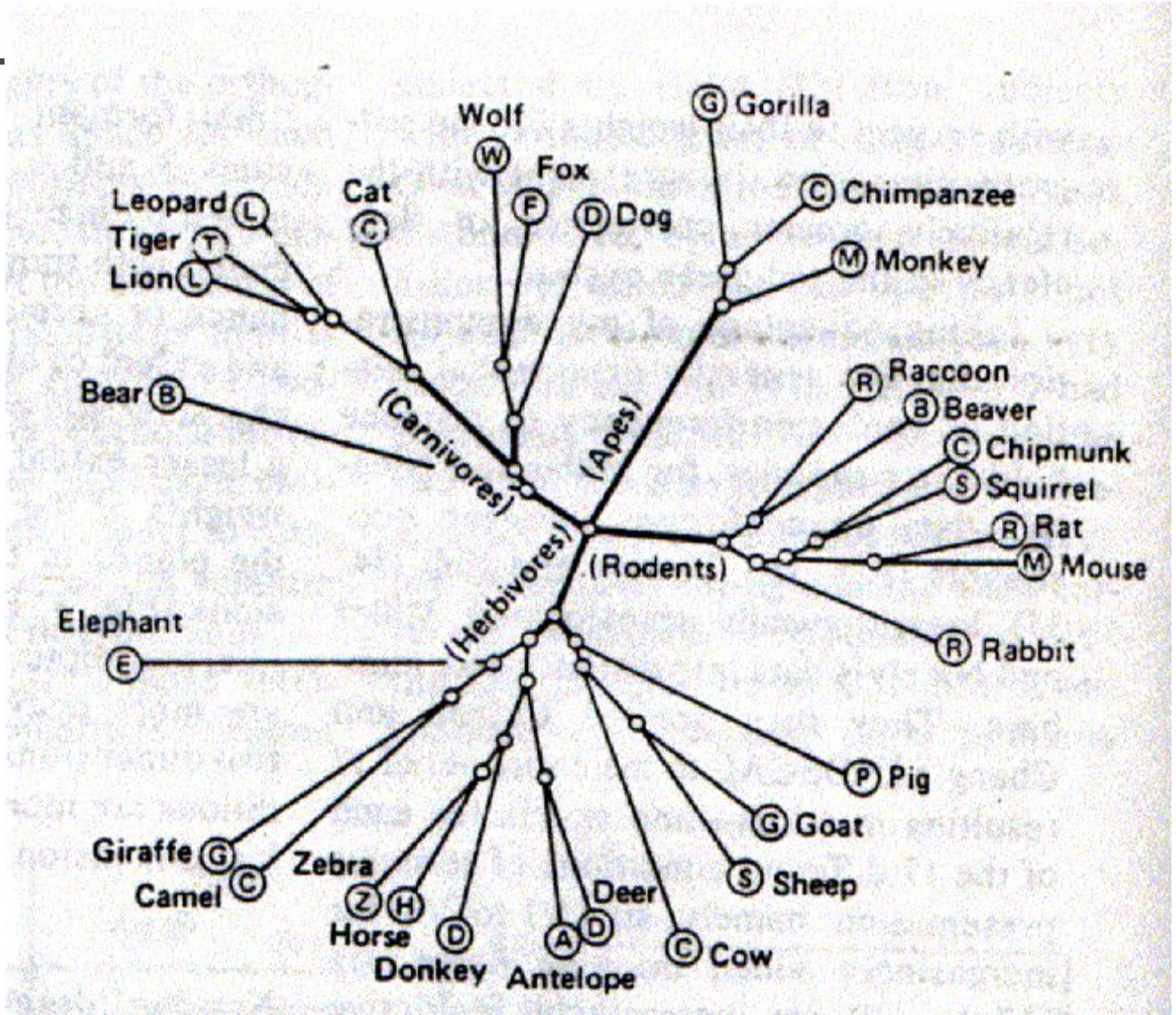
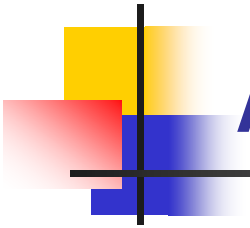


FIGURE 10.7. The distance threshold affects the number and size of clusters in similarity based clustering methods. For three different values of distance d_0, lines are drawn between points closer than d_0—the smaller the value of d_0, the smaller and more numerous the clusters.

Duda et al., "Pattern Classification"

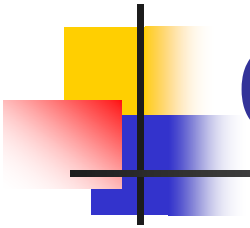
Hierarchical





Additive Trees

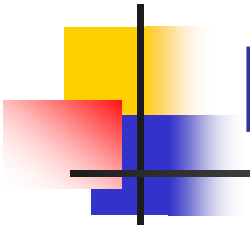
- Commonly the minimum spanning tree
- Nearest neighbor approach to hierarchical clustering
- Single-linkage



Other linkages

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

  [only that does not require all distances]



Hierarchical Clustering

- Agglomerative Technique
  - Successive “fusing” cases
  - Respect (or not) definitions of intra- and /or inter-group proximity
- Visualization
  - Dendrogram, Tree, Venn diagram

Graphical Representations

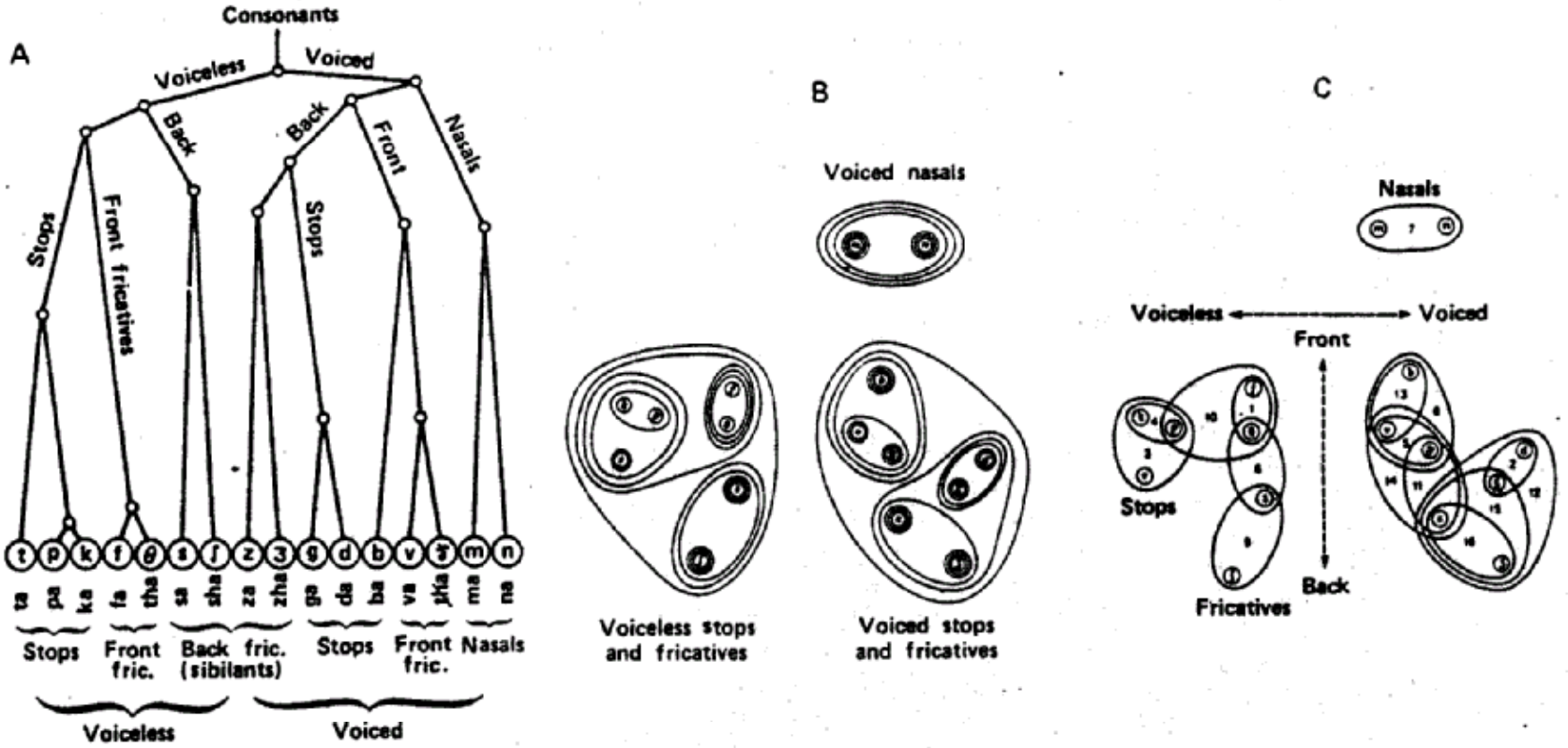


Fig. 6. Alternative clustering analyses of Miller and Nicely's (25) data on confusions of 16 consonants. (A) Hierarchical tree obtained by using Johnson's (50) nonmetric "diameter" (or "complete-link") method. (B) The same hierarchical clustering displayed as embedded in the two-dimensional scaling solution of Fig. 3A. [From Shepard (26)] (C) Nonhierarchical clustering obtained by ADCLUS analysis of the same data, embedded in the same two-dimensional scaling solution. The Arabic numerals indicate the ranks of the clusters by estimated weights. [From Shepard and Arabie (52)]

Data Visualization