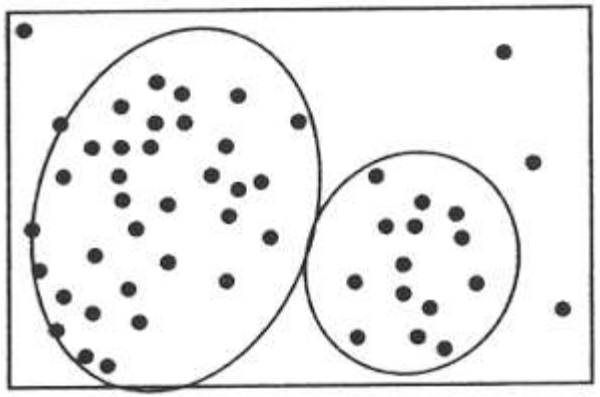
Cluster Analysis

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Natural grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes

Objective of *clustering* algorithms

- Partition the objects into groups.
  - Objects with similar categorical attribute values are placed in the same group.
  - Objects in different groups contain dissimilar categorical attribute values.



General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity

- The quality of a clustering depends on:
  - Appropriateness of method for dataset.
  - The (dis)similarity measure used
  - Its implementation.

- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Data Structures

- Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

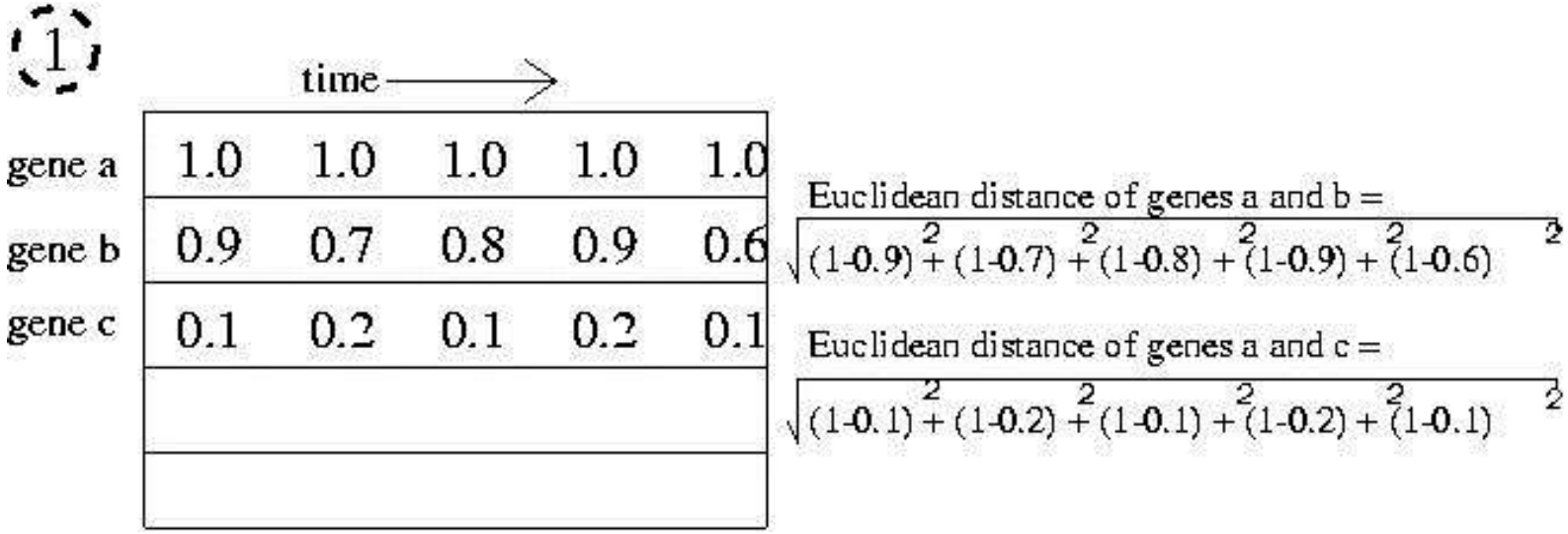
$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

- *If q = 2*, *d* is Euclidean distance:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

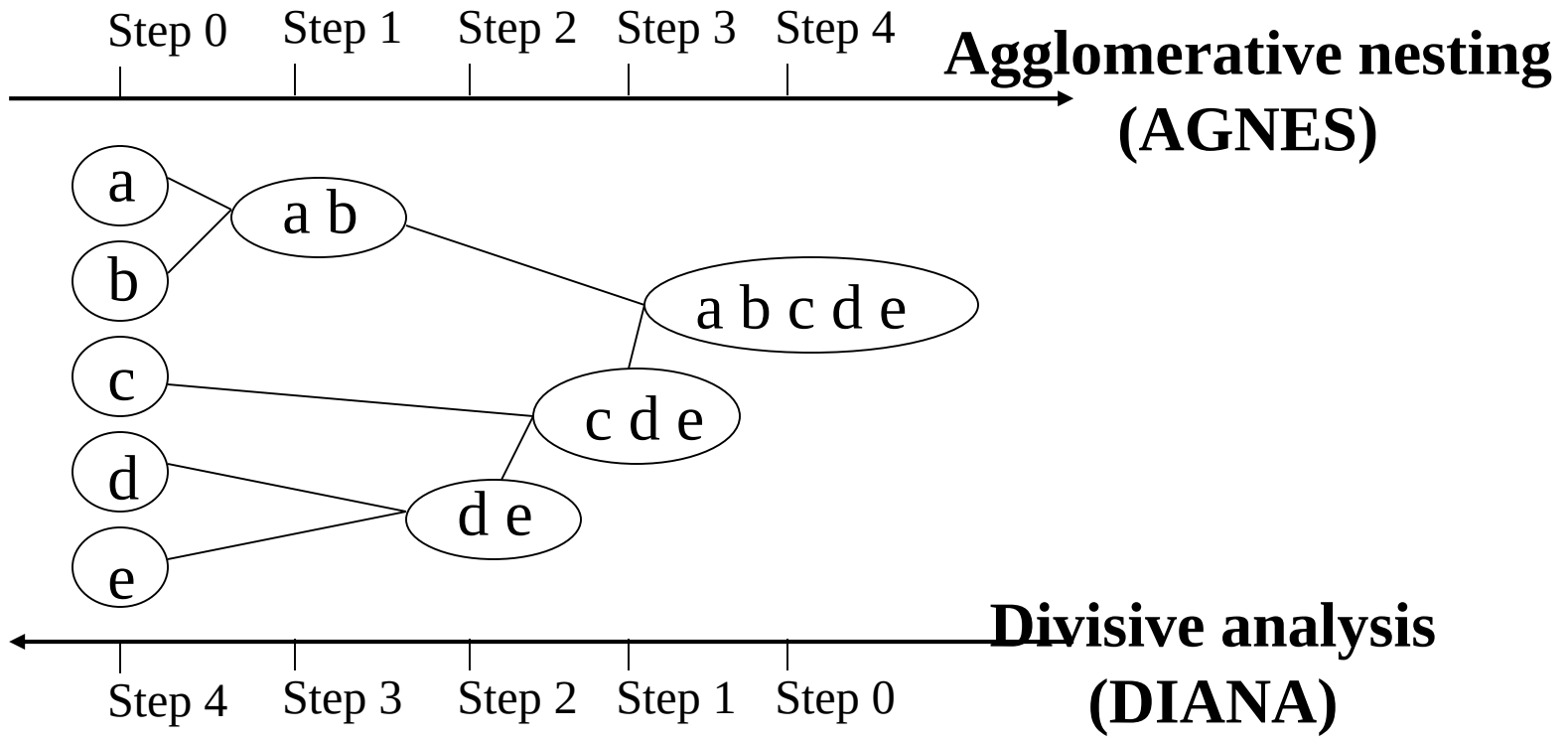
  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures.

Clustering of genomic data sets



Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition

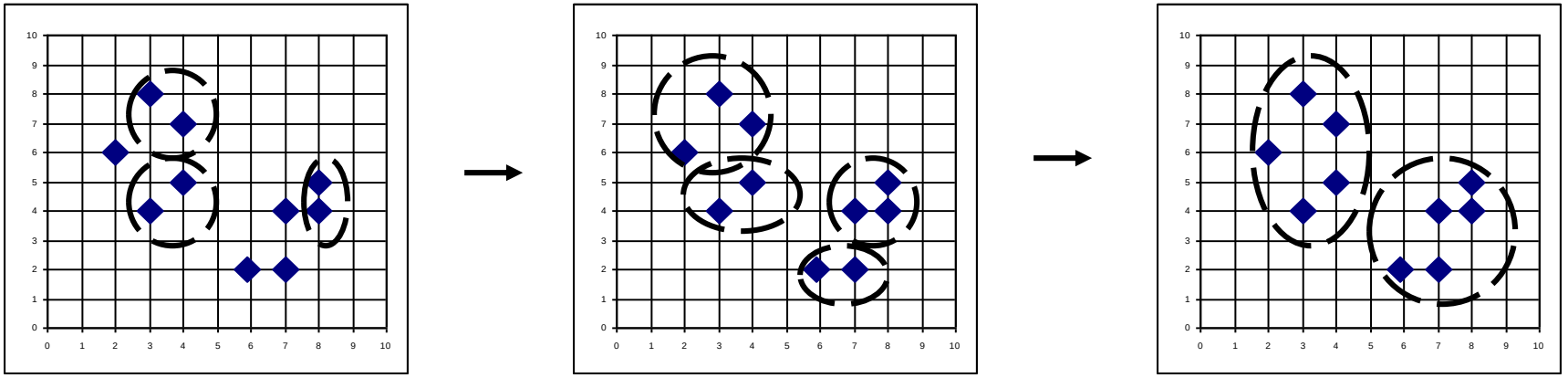


Distances between clusters

- **Average.** Average linkage averages all distances between pairs of objects in different clusters to decide how far apart they are.
- **Centroid.** Centroid linkage uses the average value of all objects in a cluster (the cluster centroid) as the reference point for distances to other objects or clusters.
- **Complete.** Complete linkage uses the most distant pair of objects in two clusters to compute between-cluster distances.
- **Median.** Median linkage uses the median distances between pairs of objects in different clusters to decide how far apart they are.
- **Single.** Single linkage defines the distance between two objects or clusters as the distance between the two closest members of those clusters.
- **Ward.** Ward's method averages all distances between pairs of objects in different clusters, with adjustments for covariances, to decide how far apart the clusters are.
- **Weighted.** Weighted average linkage uses a weighted average distance between pairs of objects in different clusters to decide how far apart they are. The weights used are proportional to the size of the cluster.

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

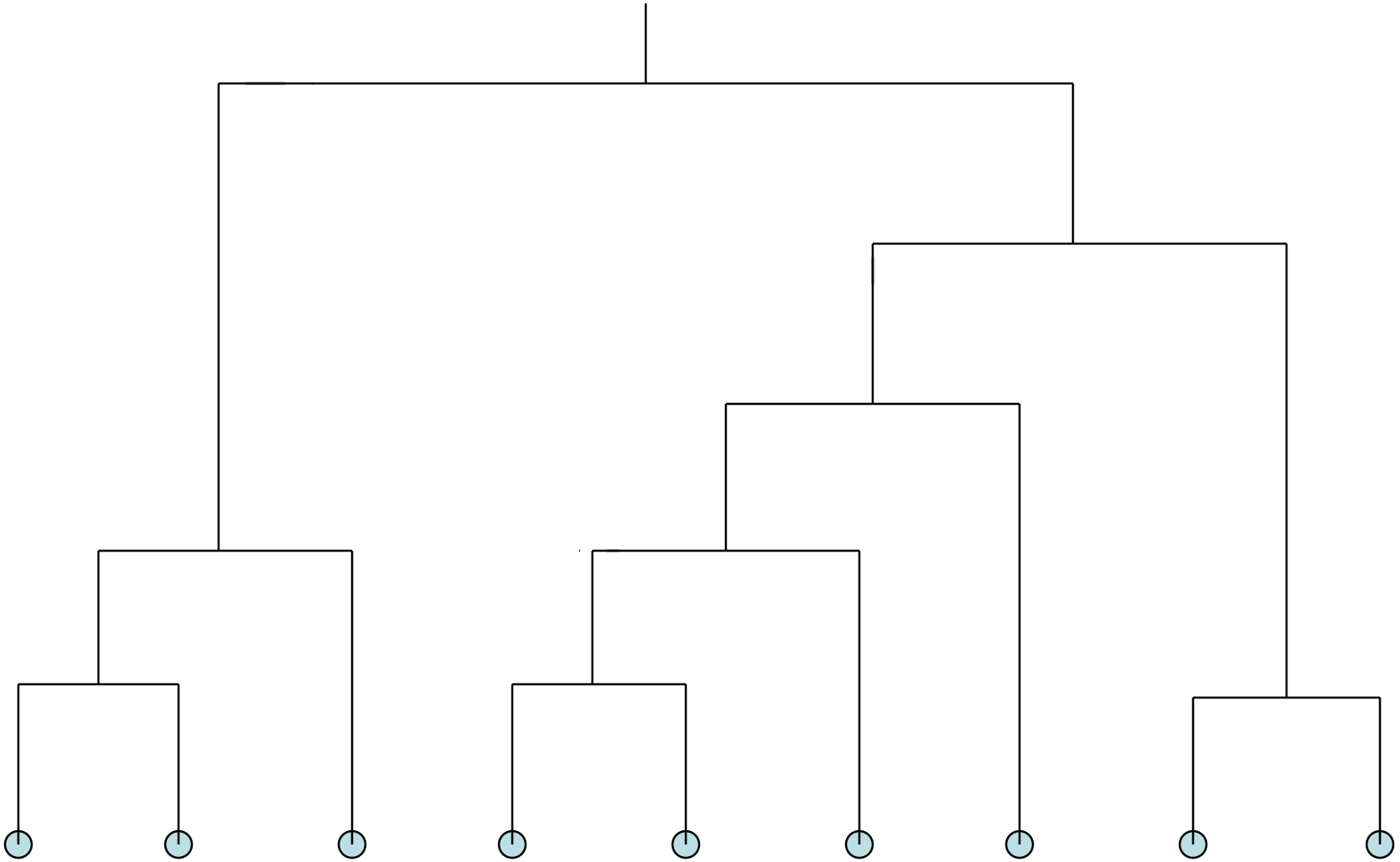


A *Dendrogram* Shows How the Clusters are Merged Hierarchically

Decompose data objects into several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

Dendrogram



```
import pandas as pd
df= pd.read_csv("E:/MY DOCUMENTS/Desktop/Python/testdata.csv")
y = df
df.drop(['SL','Companies'], axis=1,inplace=True)
print("Dimension of the data set is : ", df.shape)

#Perform Clustering
from sklearn.cluster import AgglomerativeClustering
agg=AgglomerativeClustering(n_clusters=5)
ypred=agg.fit_predict(df)
x=agg.labels_
print(x)

#Creating dendrogram
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, ward
result=ward(df)
dendrogram(result)
plt.title("DENDROGRAM")
plt.xlabel('Observations')
plt.ylabel('Distances')
plt.show()
```