BIOST 546: Machine Learning for Biomedical Big Data

Ali Shojaie

Lecture 11: Clustering - Part I
Spring 2017

Recap

- Dimension reduction methods
  - PCA
  - MDS

Today's Class

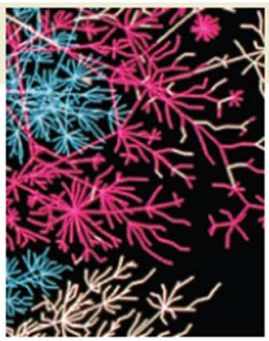
- Basics of clustering
- Hierarchical clustering

Dimension Reduction vs Cluster Analysis

- Dimension reduction methods find a low-dimensional representation of the observations that contain the good fraction of information in the original data
  - PCA tries to maximize the variance
  - MDS tries to preserve the distances among observations
- Clustering looks to find homogeneous subgroups among the observations

Cluster Analysis

Cluster analysis is one of the most-widely used techniques in analysis of omics data.



_computational BIOLOGY

PRIMER

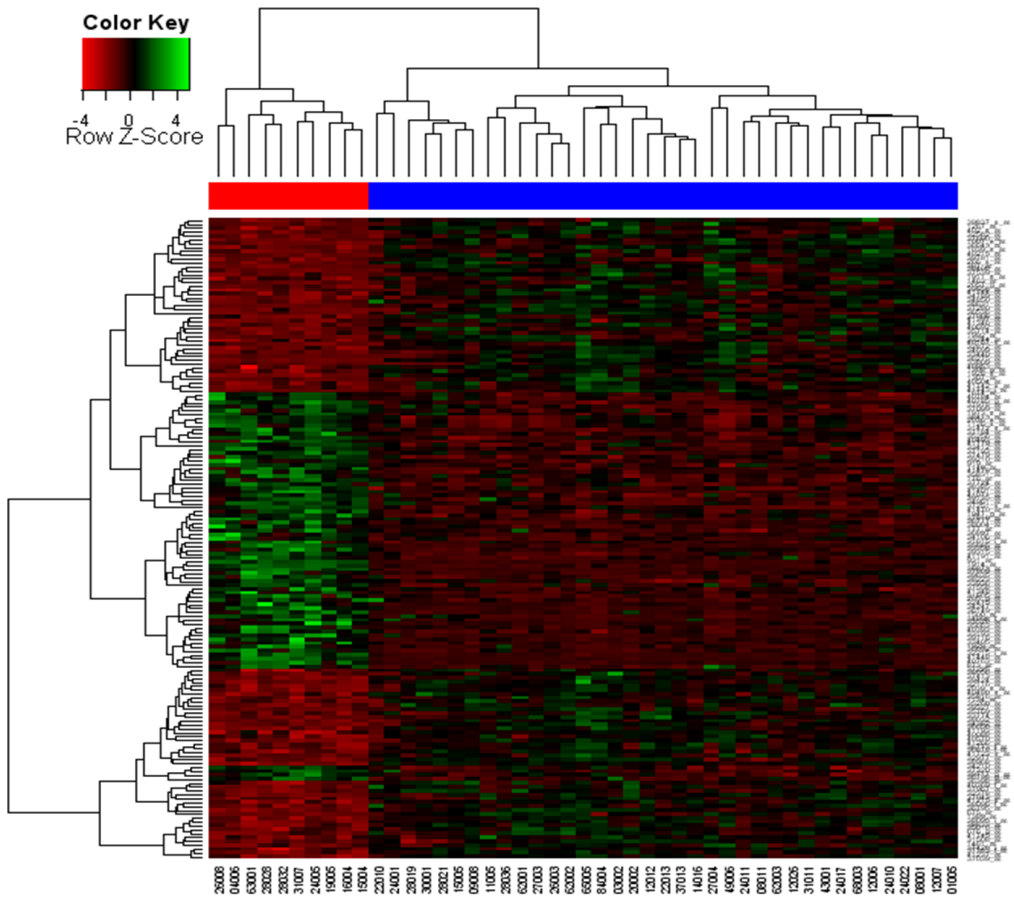
How does gene expression clustering work?

Patrik D'haeseleer

Clustering is often one of the first steps in gene expression analysis. How do clustering algorithms work, which ones should we use and what can we expect from them?

Cluster Analysis

Almost all papers on omics research, have a *heatmap*, which often includes a clustering of genes/samples.



Objective

- Grouping objects into **meaningful** subsets or **clusters**, such that objects within each cluster are more **similar** to one another than objects in other clusters.

Objective

- Grouping objects into **meaningful** subsets or **clusters**, such that objects within each cluster are more **similar** to one another than objects in other clusters.
- Need to define what a “meaningful” cluster is

Objective

- Grouping objects into **meaningful** subsets or **clusters**, such that objects within each cluster are more **similar** to one another than objects in other clusters.
- Need to define what a “meaningful” cluster is
- Need to define what we mean by “similarity”

Objective

- Grouping objects into **meaningful** subsets or **clusters**, such that objects within each cluster are more **similar** to one another than objects in other clusters.
- Need to define what a “meaningful” cluster is
- Need to define what we mean by “similarity”
- Can cluster **observations** or **features**:
  - **observations**: clustering cancer samples to find cancer sub-types
  - **features**: clustering genes based on similar functions (pathways)
  - Clustering features is similar to clustering observations (work with the transpose of the data matrix)

Defining **meaningful** clusters

How many clusters?

How many clusters?

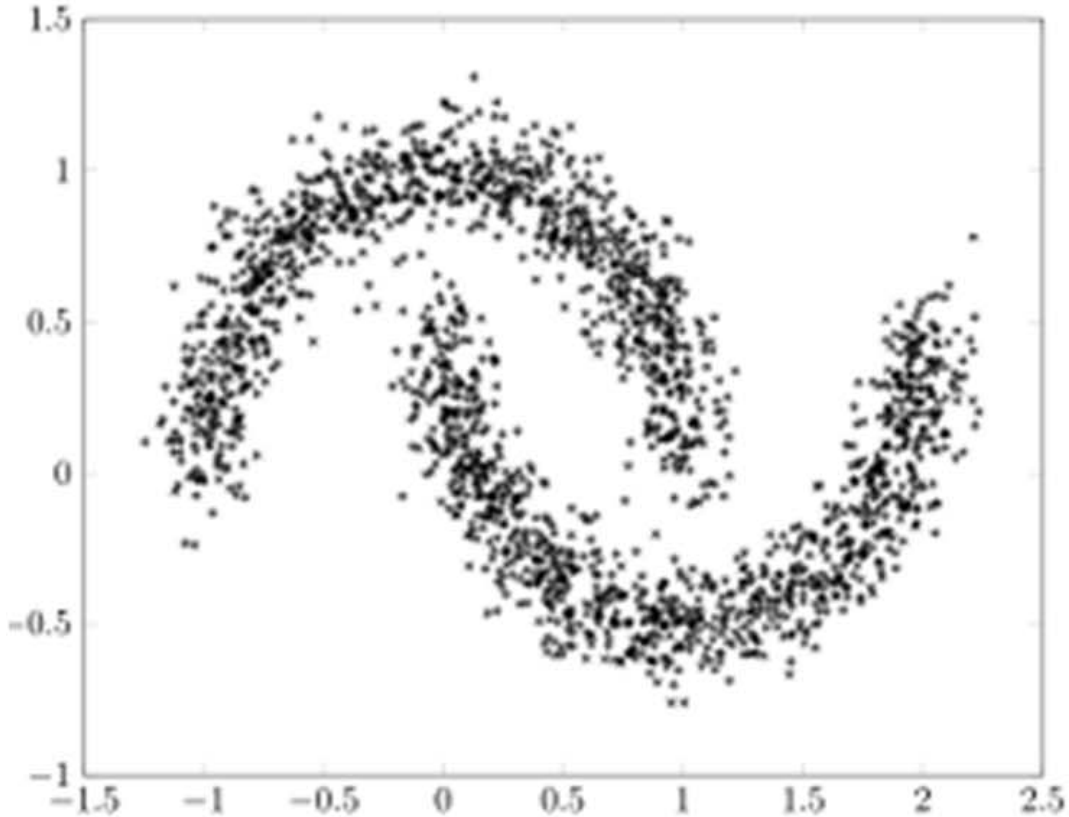
Six Clusters

Two Clusters

Four Clusters

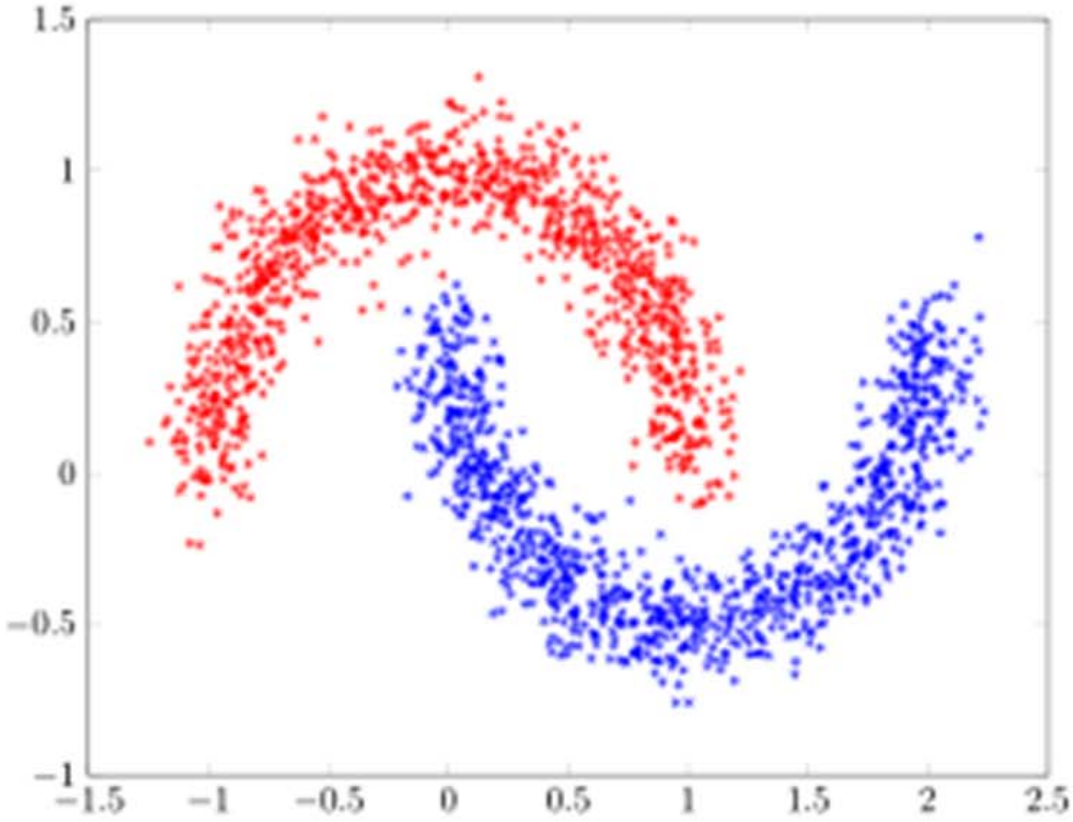
Defining **meaningful** clusters

Which points are more similar?



Defining **meaningful** clusters

Which points are more similar?



Similarity/Dissimilarity Measures

- An $n \times n$ matrix, calculated from the data matrix $\mathbf{X}$

Similarity/Dissimilarity Measures

- An $n \times n$ matrix, calculated from the data matrix $\mathbf{X}$
- Similarity measure:
  - A numerical measure $s(i,j)$ that indicates how **similar** two objects are (high $s \equiv$ high similarity). Similarity is often normalized to have a magnitude in the range $[0,1]$.
  - Properties: $s(i,j) \geq 0$ and $s(i,j) = s(j,i)$

Similarity/Dissimilarity Measures

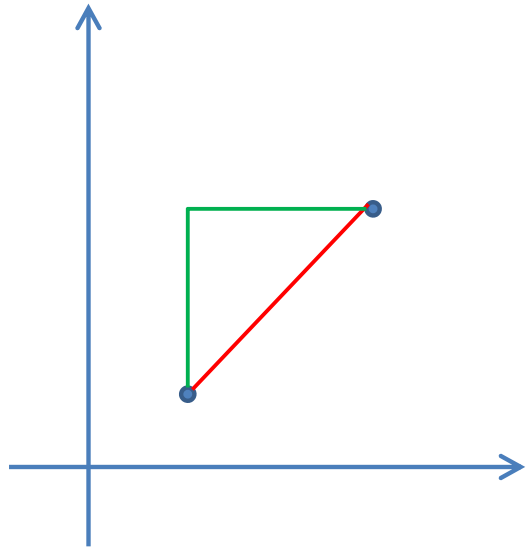
- An $n \times n$ matrix, calculated from the data matrix $\mathbf{X}$
- Similarity measure:
  - A numerical measure $s(i,j)$ that indicates how **similar** two objects are (high $s \equiv$ high similarity). Similarity is often normalized to have a magnitude in the range $[0,1]$.
  - Properties: $s(i,j) \geq 0$ and $s(i,j) = s(j,i)$
- Dissimilarity measure:
  - A numerical measure $d(i,j)$ that indicates how **different** two objects are (lower $d \equiv$ high similarity). Unlike similarity, the upper bound may vary
  - Properties: $d(i,j) \geq 0$ and $d(i,j) = d(j,i)$
  - Any **distance** naturally defines a dissimilarity measure

Common Dissimilarity Measures

- **Euclidean** distance: $d(i,j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$
- **Manhattan** distance: $d(i,j) = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$
- **Mahalanobis** distance: For p-dimensional vectors X_i and X_j,
  $d(i,j) = (X_i - X_j)^{\mathsf{T}} \Sigma^{-1} (X_i - X_j)$,
  where Σ is the covariance matrix of X_k's

Common Dissimilarity Measures

- **Euclidean** distance: $d(i,j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$
- **Manhattan** distance: $d(i,j) = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$
- **Mahalanobis** distance: For p-dimensional vectors X_i and X_j, $d(i,j) = (X_i - X_j)^{\mathsf{T}} \Sigma^{-1} (X_i - X_j)$, where Σ is the covariance matrix of X_k's



Common Similarity Measures

- **Correlation** coefficient
  - *Pearson* correlation (sample version):

$$r(X_i, X_j) = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_k (X_{ik} - \bar{X}_i)^2 \sum_k (X_{jk} - \bar{X}_j)^2}}$$

This is the "usual" correlation coefficient, and measures the **linear** association between X_i and X_j

Common Similarity Measures

- **Correlation** coefficient
  - *Pearson* correlation (sample version):

$$r(X_i, X_j) = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_k (X_{ik} - \bar{X}_i)^2 \sum_k (X_{jk} - \bar{X}_j)^2}}$$

    This is the "usual" correlation coefficient, and measures the **linear** association between X_i and X_j
  - *Spearman* correlation: Same as Pearson correlation, but applied to **ranked** observations.
    Corresponds to an **increasing monotonic trend** between X_i and X_j (more appropriate for non-Gaussian observations)
  - *Kendall's* τ: uses directly rankings among pairs of observations

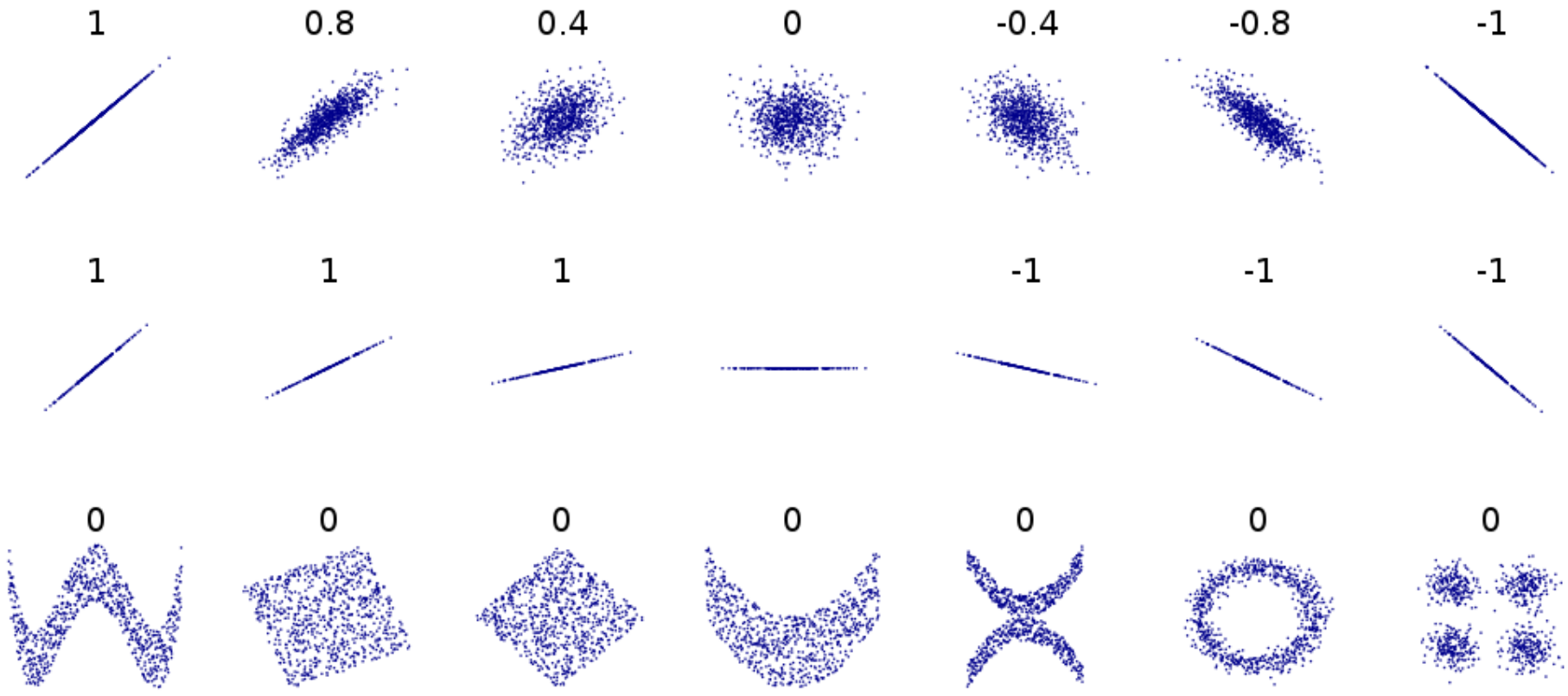
Common Similarity Measures

- **Correlation** coefficient
  - *Pearson* correlation (sample version):

$$r(X_i, X_j) = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_k (X_{ik} - \bar{X}_i)^2 \sum_k (X_{jk} - \bar{X}_j)^2}}$$

    This is the "usual" correlation coefficient, and measures the **linear** association between X_i and X_j
  - *Spearman* correlation: Same as Pearson correlation, but applied to **ranked** observations.
    Corresponds to an **increasing monotonic trend** between X_i and X_j (more appropriate for non-Gaussian observations)
  - *Kendall's* τ: uses directly rankings among pairs of observations
- **Cosine** measure: $\cos(X_i, X_j) = \frac{X_i \cdot X_j}{\|X_i\| \|X_j\|} = \frac{\sum_k X_{ik} X_{jk}}{\sqrt{\sum_k X_{ik}^2 \sum_k X_{jk}^2}}$ Measures the angle between two vectors, which determines how much they align.

Correlation Coefficient



Correlation Coefficient

Which data set has higher Pearson correlation?

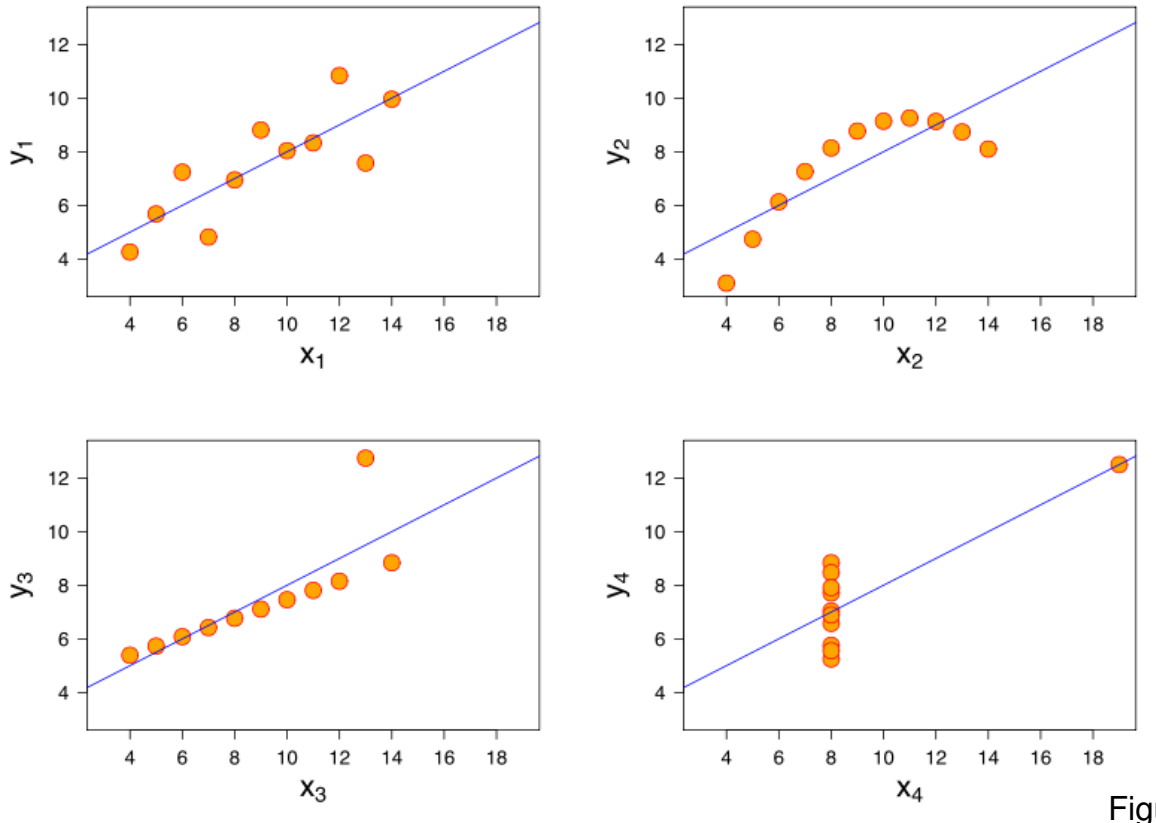


Figure courtesy of Wikipedia

Correlation Coefficient

Which data set has higher Pearson correlation?

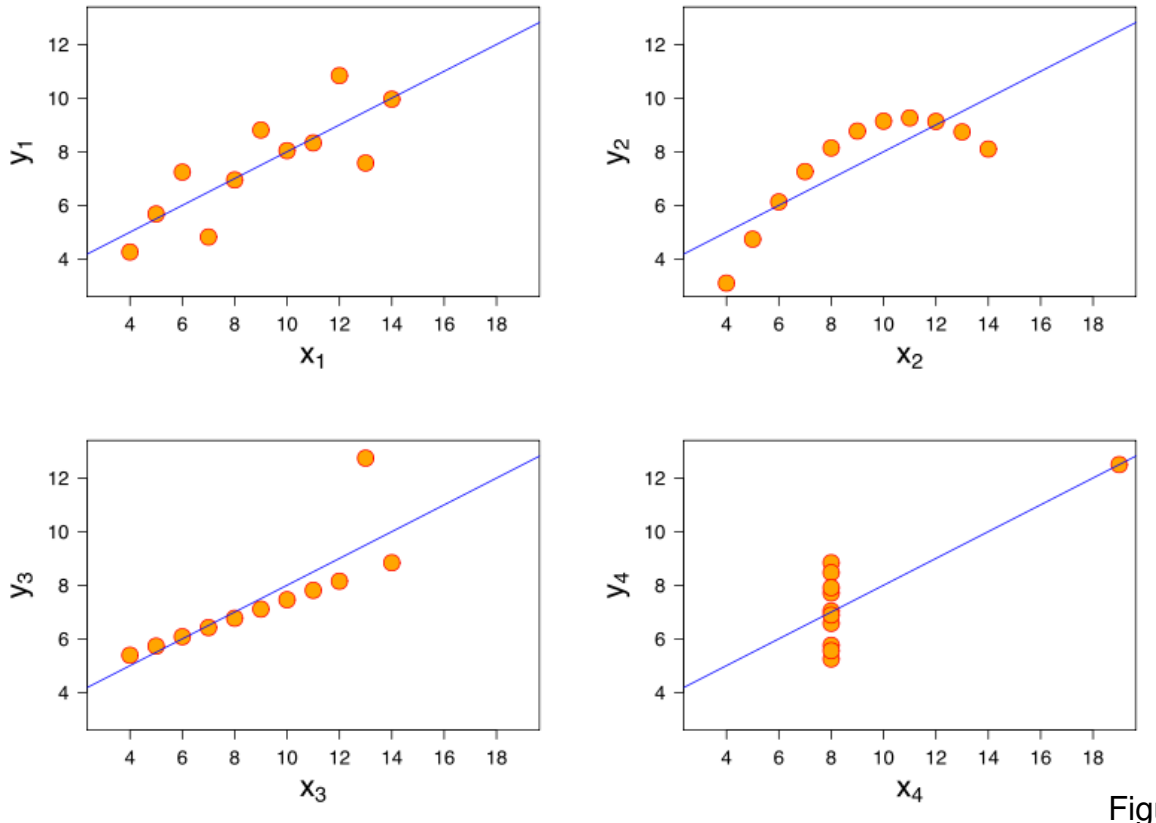


Figure courtesy of Wikipedia

In all of the above cases, $\mathbf{r = 0.816}$.

Methods of Hierarchical Clustering

Hierarchical clustering results in a sequence of solutions (nested clusters), organized in a hierarchical tree structure, called the **dendrogram**

Methods of Hierarchical Clustering

Hierarchical clustering results in a sequence of solutions (nested clusters), organized in a hierarchical tree structure, called the **dendrogram**

- **Bottom-Up** or Agglomerative: Start from n individual clusters, and group them together into using a measure of similarity.
  - Start from n individual clusters
  - At each step, merge the closest pair of clusters until all objects form a single cluster

Methods of Hierarchical Clustering

Hierarchical clustering results in a sequence of solutions (nested clusters), organized in a hierarchical tree structure, called the **dendrogram**

- **Bottom-Up** or Agglomerative: Start from n individual clusters, and group them together into using a measure of similarity.
  - Start from n individual clusters
  - At each step, merge the closest pair of clusters until all objects form a single cluster
- **Top-Down** or Divisive: Start from one cluster containing all objects, and break them down using a measure of distance
  - Start from 1 cluster
  - At each step, split the most heterogenous cluster until every cluster has only one member

Methods of Hierarchical Clustering

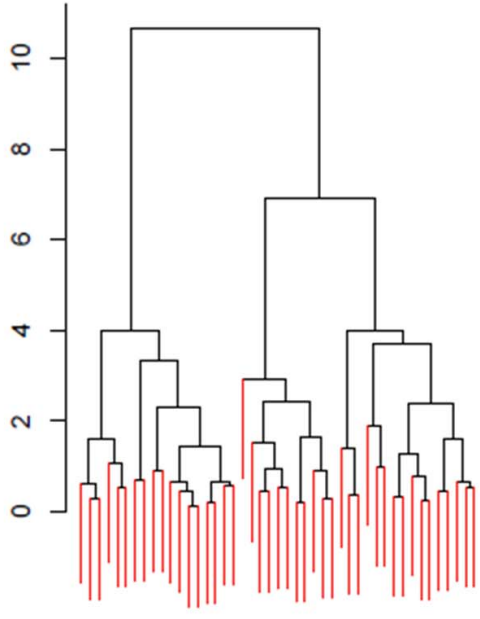
Hierarchical clustering results in a sequence of solutions (nested clusters), organized in a hierarchical tree structure, called the **dendrogram**

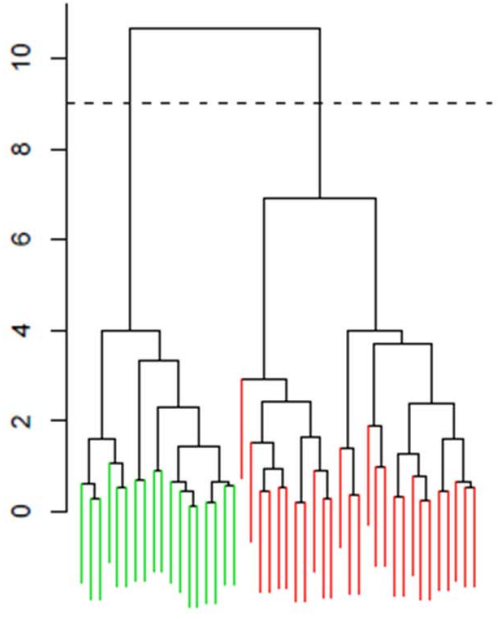
- **Bottom-Up** or Agglomerative: Start from n individual clusters, and group them together into using a measure of similarity.
  - Start from n individual clusters
  - At each step, merge the closest pair of clusters until all objects form a single cluster
- **Top-Down** or Divisive: Start from one cluster containing all objects, and break them down using a measure of distance
  - Start from 1 cluster
  - At each step, split the most heterogenous cluster until every cluster has only one member

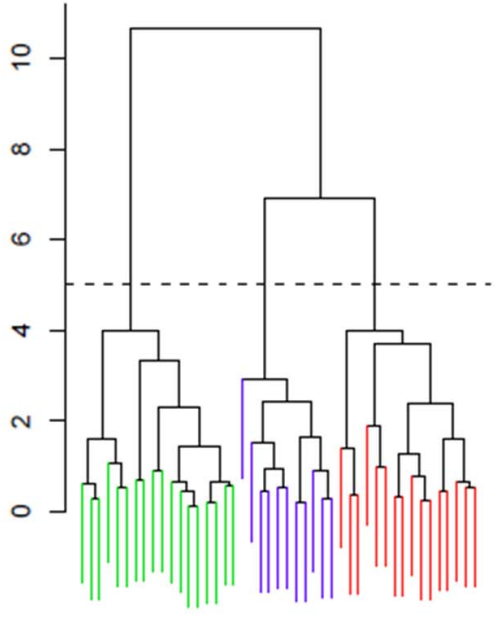
We will focus on Bottom-Up methods.

Interpreting the Dendrogram

How many clusters?

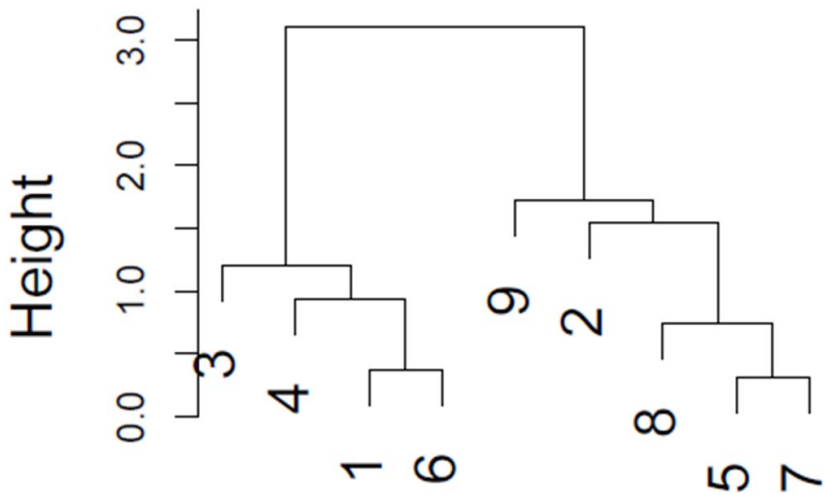






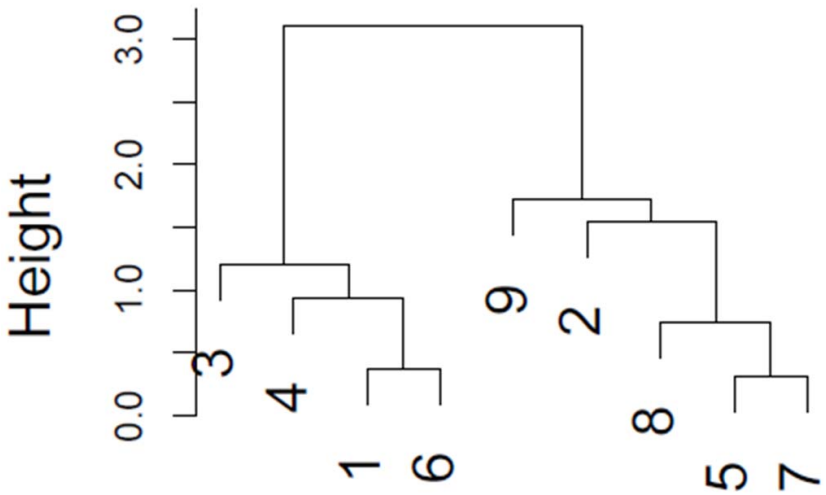
Interpreting the Dendrogram

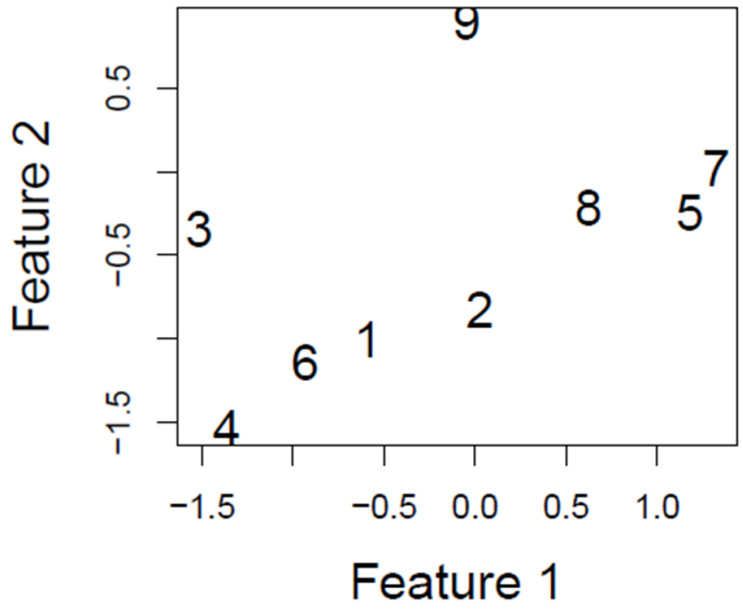
Which points are clustered together?



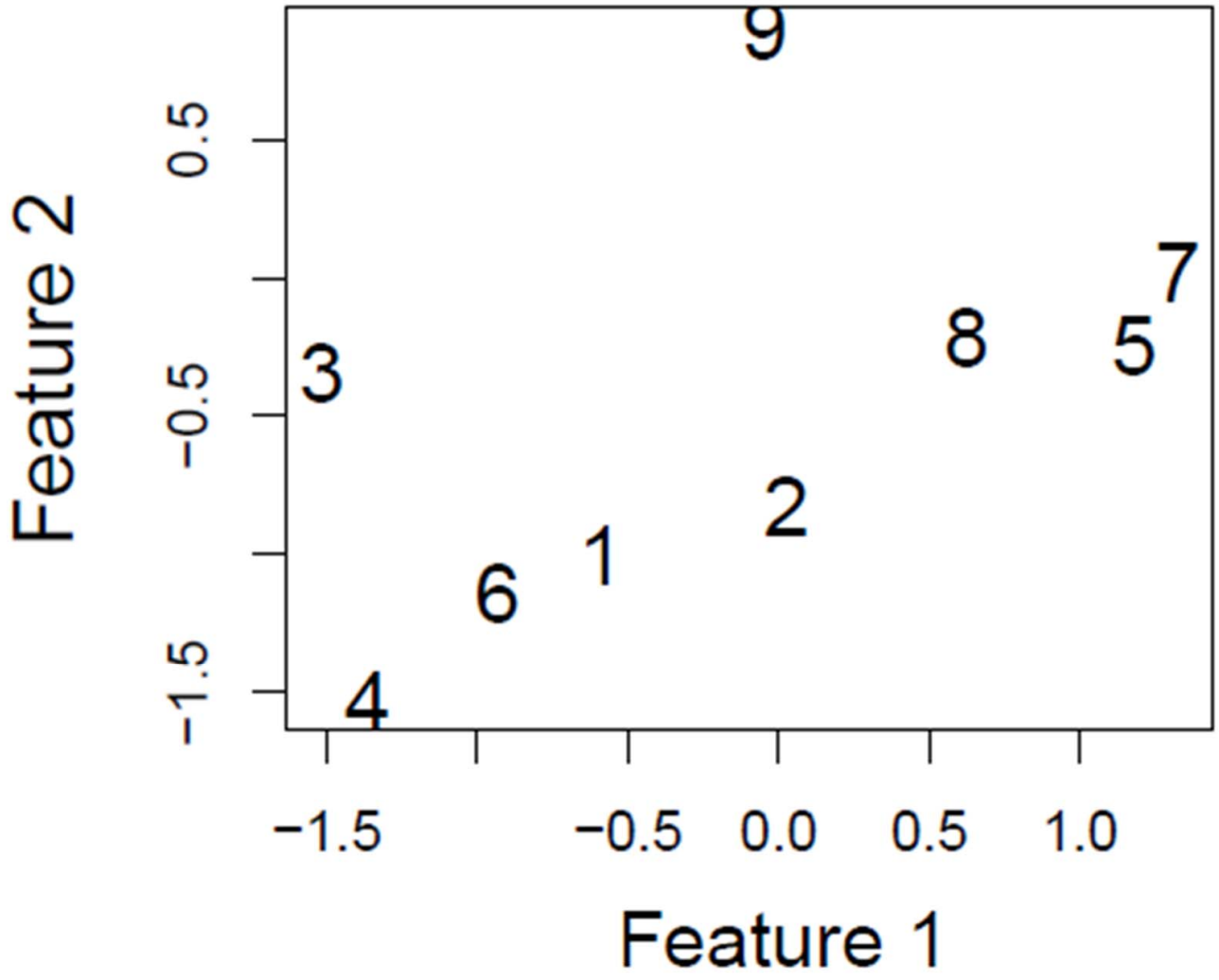
How To Read The Dendrogram?

Which points are clustered together?

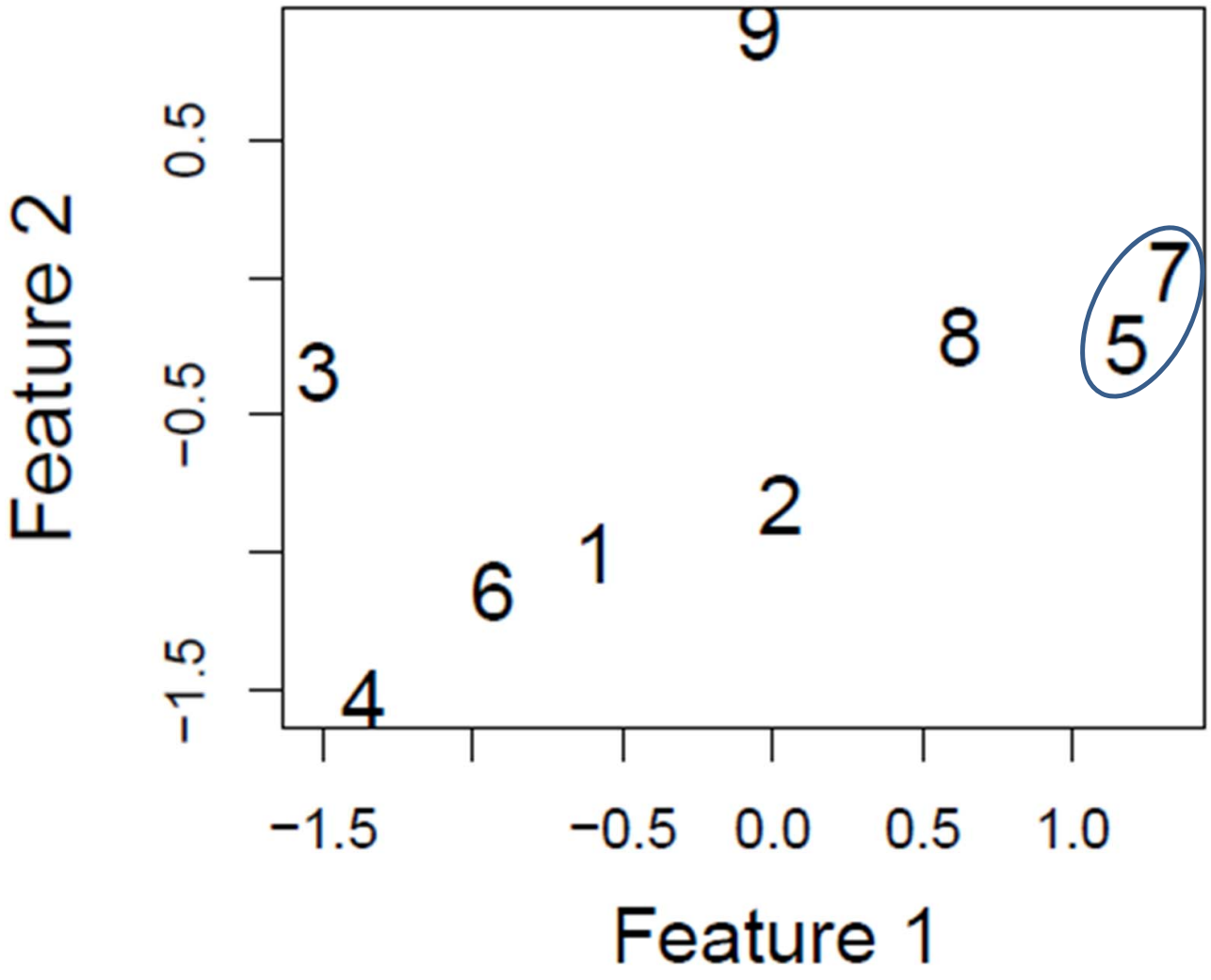




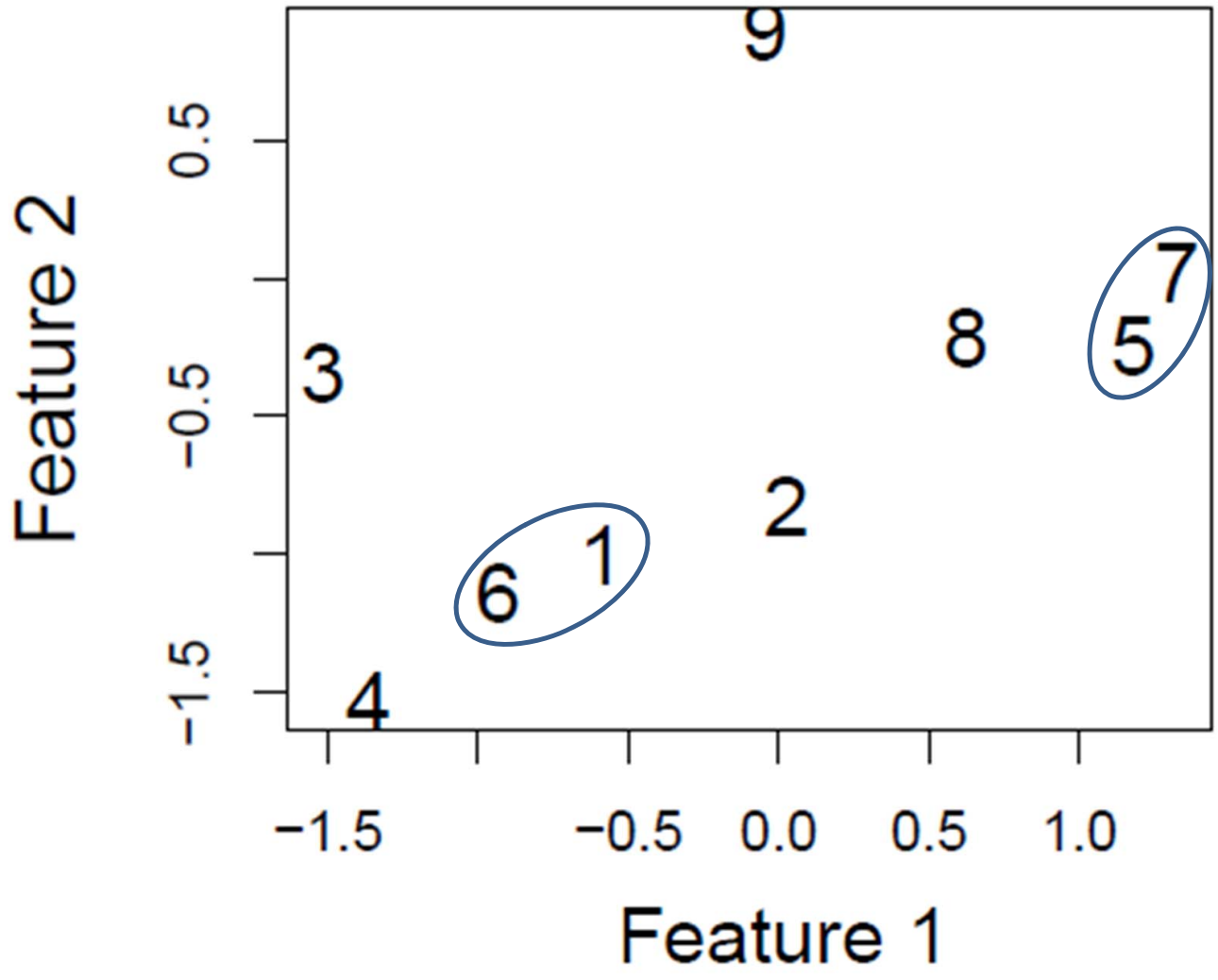
A Closer Look



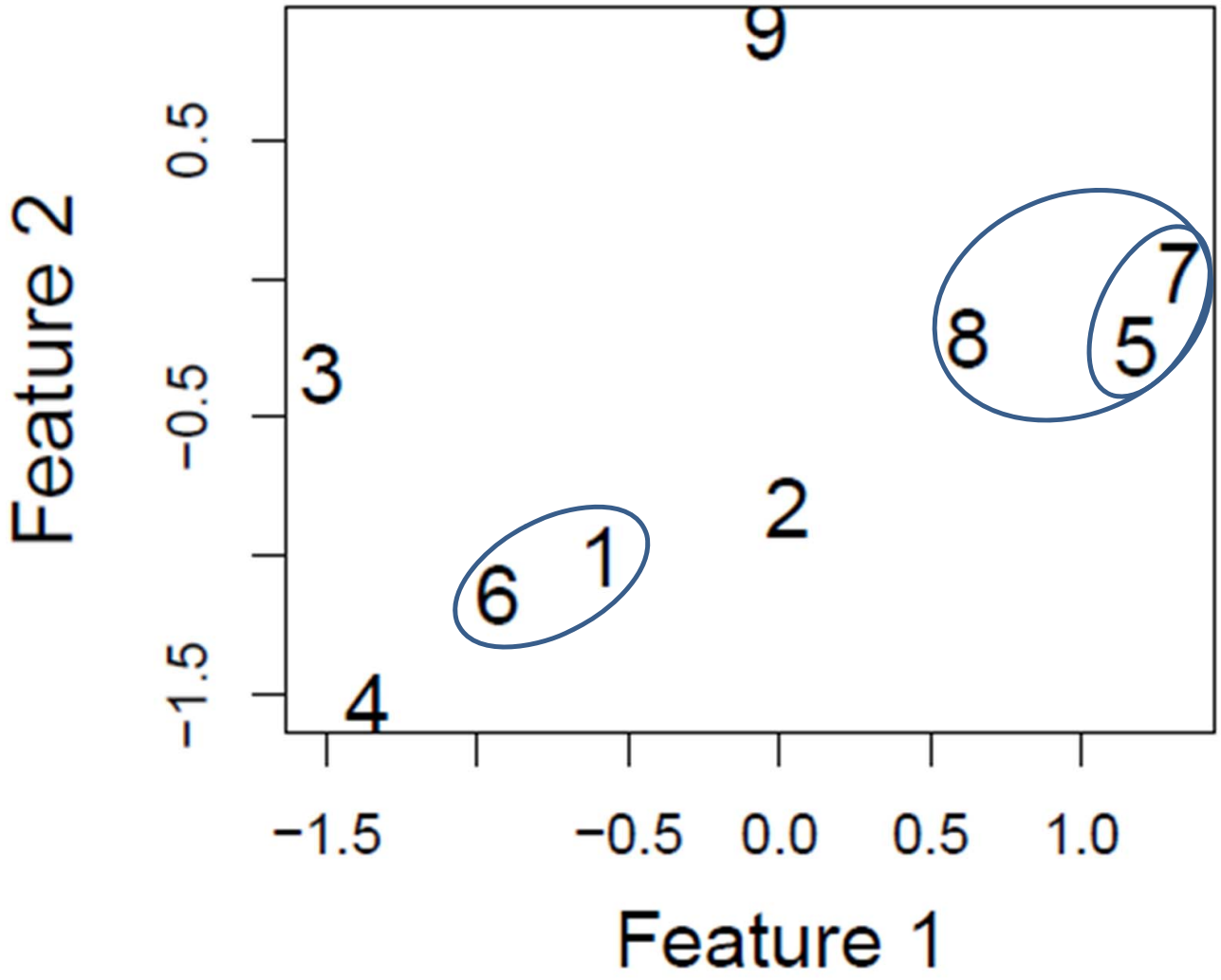
A Closer Look



A Closer Look



A Closer Look



Interpreting the Dendrogram

Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation

Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation
- As we move up the tree, some leaves begin to fuse into branches: these are observations that are similar to each other

Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation
- As we move up the tree, some leaves begin to fuse into branches: these are observations that are similar to each other
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other

Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation
- As we move up the tree, some leaves begin to fuse into branches: these are observations that are similar to each other
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other
- Observations that fuse later (near the top of the tree) can be quite different

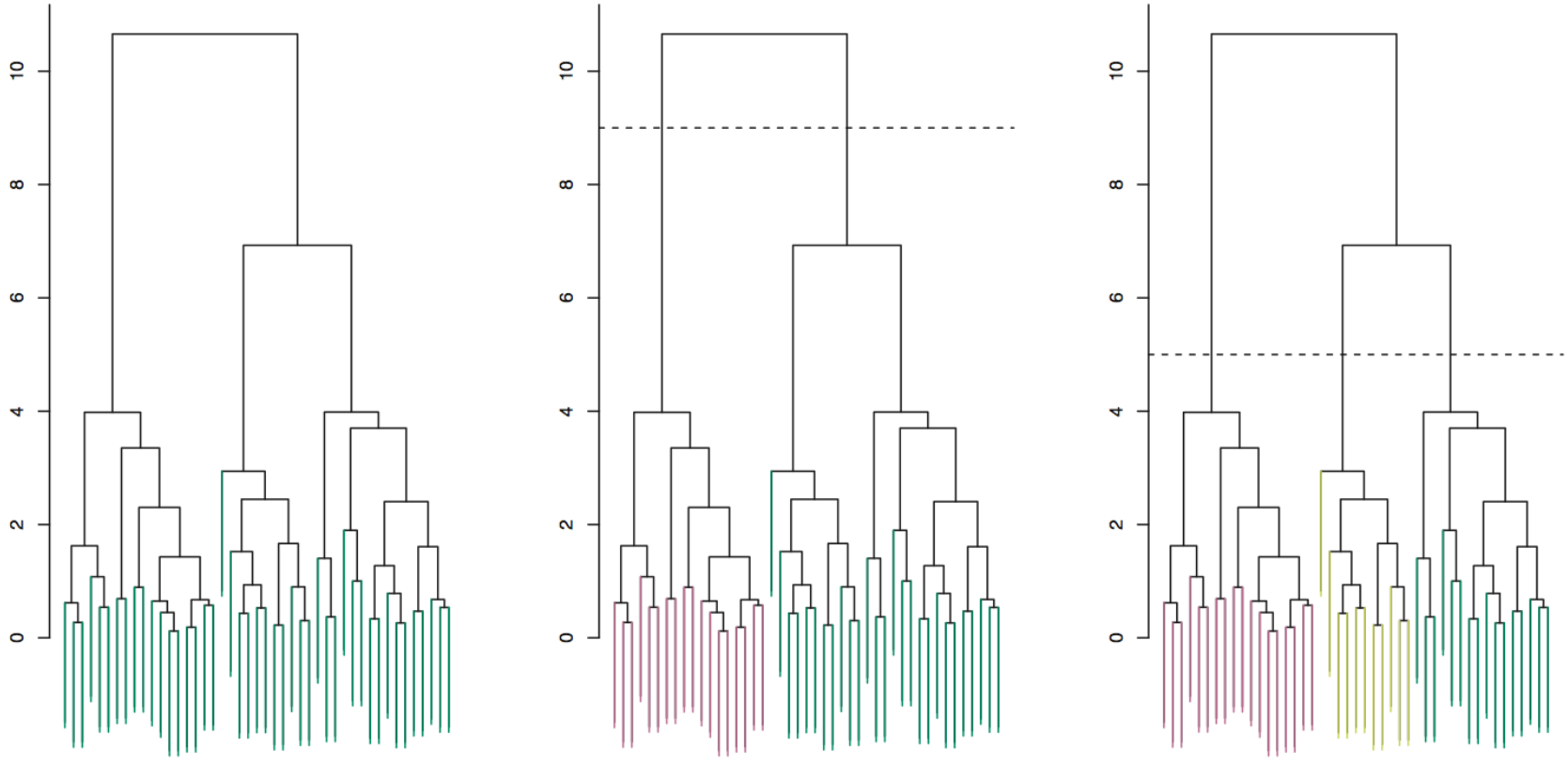
Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation
- As we move up the tree, some leaves begin to fuse into branches: these are observations that are similar to each other
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other
- Observations that fuse later (near the top of the tree) can be quite different
- The height of the point in the tree where branches containing two observations are first fused, as measured on the *vertical axis*, indicates how different they are from each other

Interpreting the Dendrogram

- At the bottom of the tree, there is a leaf for each observation
- As we move up the tree, some leaves begin to fuse into branches: these are observations that are similar to each other
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other
- Observations that fuse later (near the top of the tree) can be quite different
- The height of the point in the tree where branches containing two observations are first fused, as measured on the *vertical axis*, indicates how different they are from each other
- However, we cannot draw conclusions about the similarity of two observations based on their proximity along the *horizontal axis*

Interpreting the Dendrogram



Measures of Inter-Cluster similarity

- Single linkage (min)
- Complete linkage (max)
- Average linkage
- Distance between centroids
- Ward's method

Single Linkage

- At each step, the minimum distance between points in two clusters is used to determine which two clusters should be merged

Single Linkage

- At each step, the minimum distance between points in two clusters is used to determine which two clusters should be merged
- Can handle **diverse shapes**

Single Linkage

- At each step, the minimum distance between points in two clusters is used to determine which two clusters should be merged
- Can handle **diverse shapes**
- **Very sensitive to outliers/noise**

Single Linkage

- At each step, the minimum distance between points in two clusters is used to determine which two clusters should be merged
- Can handle **diverse shapes**
- **Very sensitive to outliers/noise**
- In practice, often results in **unbalance clusters**; may break down the clusters by “chaining” their members

Single Linkage

- At each step, the minimum distance between points in two clusters is used to determine which two clusters should be merged
- Can handle **diverse shapes**
- **Very sensitive to outliers/noise**
- In practice, often results in **unbalance clusters**; may break down the clusters by “chaining” their members
- Can result in extended, trailing clusters in which observations are fused one-at-a-time

Complete Linkage

- At each step, the maximum distance between points in two clusters is used to determine which two clusters should be merged

Complete Linkage

- At each step, the maximum distance between points in two clusters is used to determine which two clusters should be merged
- Often gives **comparable cluster sizes**

Complete Linkage

- At each step, the maximum distance between points in two clusters is used to determine which two clusters should be merged
- Often gives **comparable cluster sizes**
- Less sensitive to outliers

Complete Linkage

- At each step, the maximum distance between points in two clusters is used to determine which two clusters should be merged
- Often gives **comparable cluster sizes**
- Less sensitive to outliers
- Works better with **spherical distributions**

Average Linkage

- At each step, the average distance between points in two clusters is used to determine which two clusters should be merged

Average Linkage

- At each step, the average distance between points in two clusters is used to determine which two clusters should be merged
- A compromise between single and complete linkage

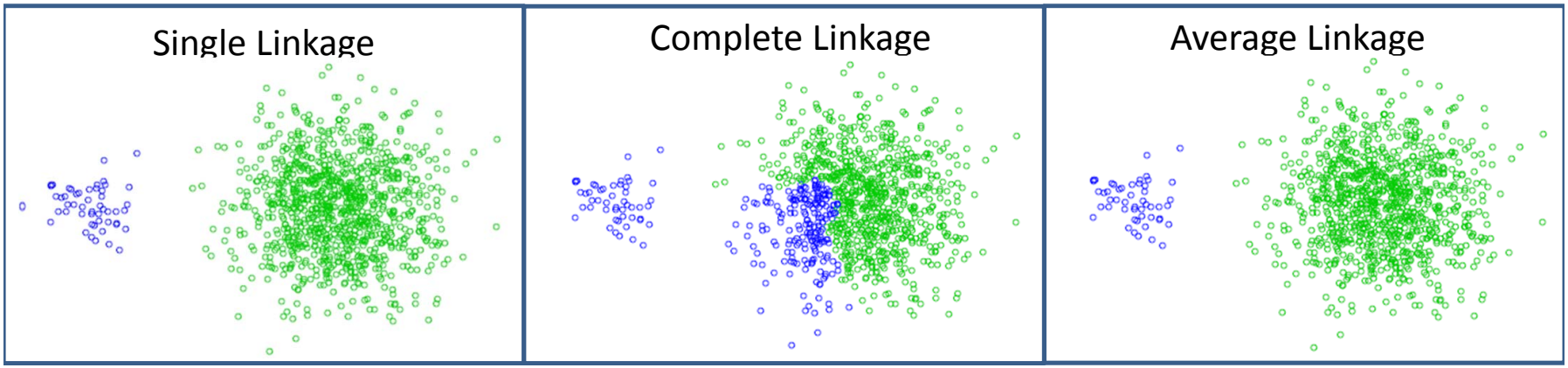
Average Linkage

- At each step, the average distance between points in two clusters is used to determine which two clusters should be merged
- A compromise between single and complete linkage
- Less sensitive to outliers

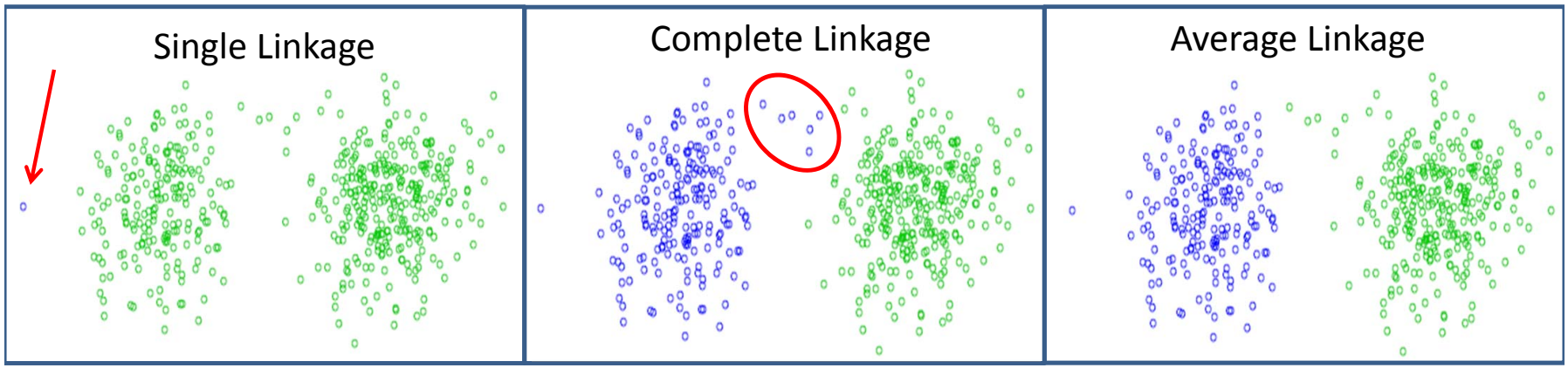
Average Linkage

- At each step, the average distance between points in two clusters is used to determine which two clusters should be merged
- A compromise between single and complete linkage
- Less sensitive to outliers
- Works better with **spherical distributions**

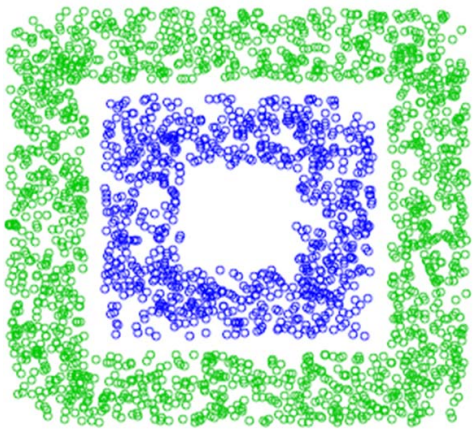
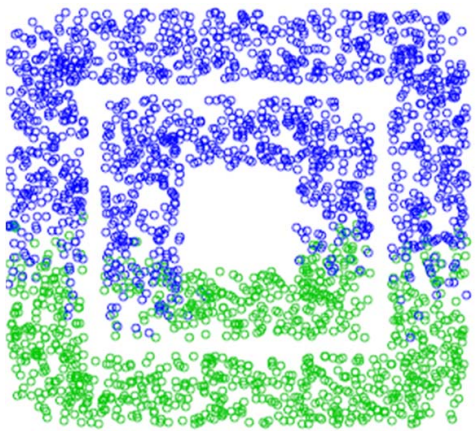
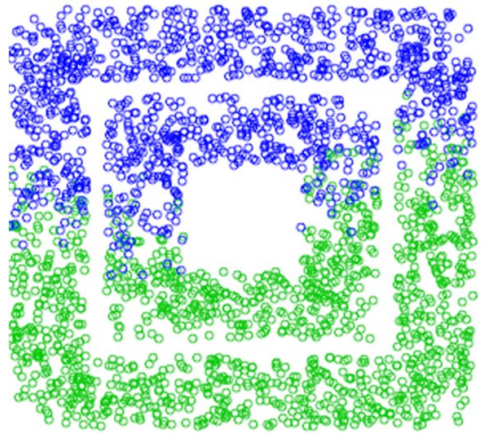
Unbalanced Clusters



Outliers



Non-Spherical Distributions

Single Linkage	Complete Linkage	Average Linkage
		

Example

- Clustering analysis on NCI60 data
- The dataset includes gene expression data for 6830 genes from 64 cancer samples
- Data can be downloaded form http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Missing values have been imputed
- Cluster the samples using different hierarchical clustering methods

Example, contd.

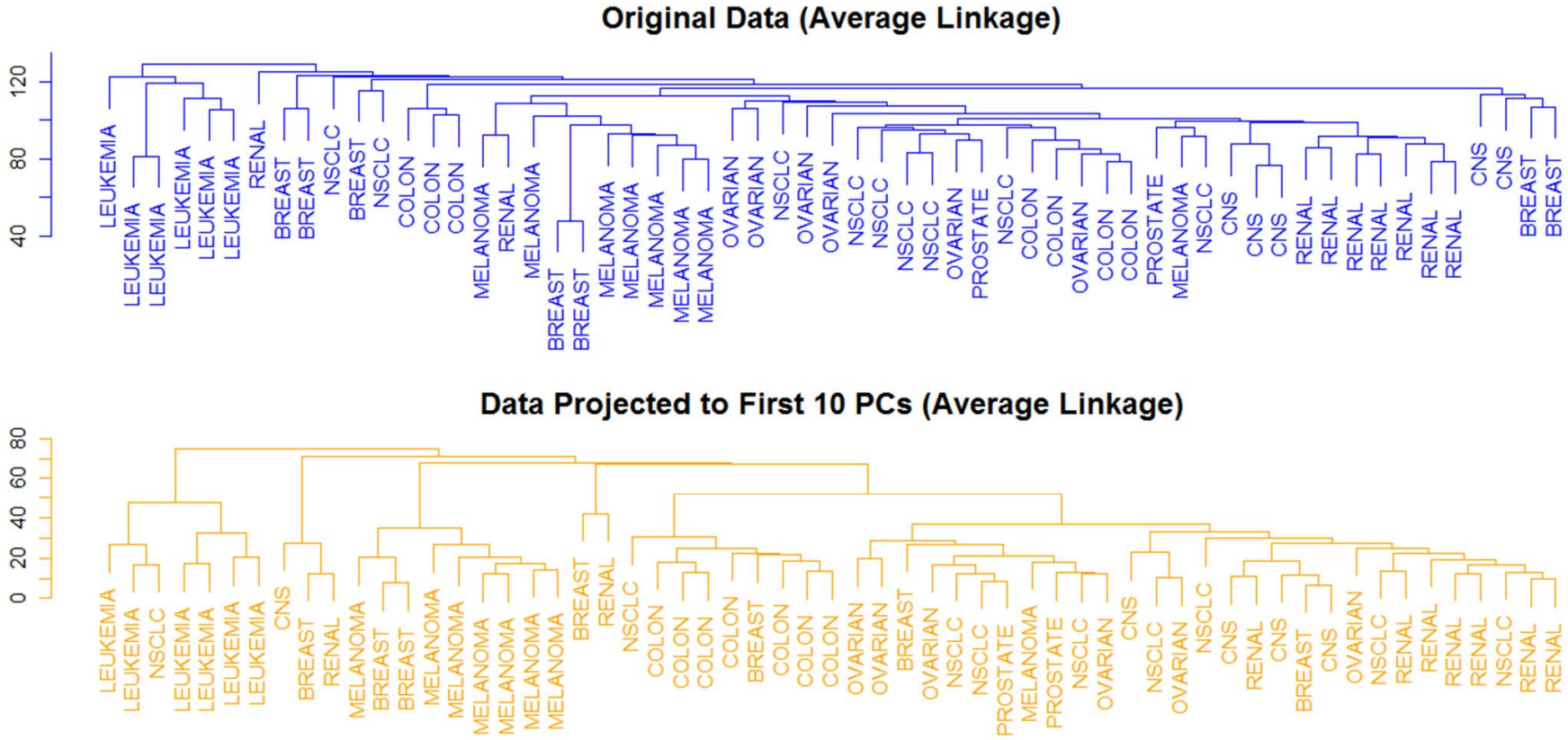
```
## Analysis NCI60 data
mydata <- read.table('nci.data')
dim(mydata)
mydata <- t(mydata)
mydata <- scale(mydata, center=F, scale=T)

labs <- read.table('nci.info', skip=14)
labs <- as.vector(as.matrix(labs))
table(labs)

## Define the distance
mydist <- dist(mydata)

## Plot the results
par(mfrow=c(1,3))
plot(hclust(mydist), labels=labs, col="green", main="Complete Linkage",
    hang=0.2, xlab="", sub="", ylab="", cex.main=1.5, cex.lab=0.6)
plot(hclust(mydist, method="average"), labels=labs, col="orange", hang=0.2
main="Average Linkage", xlab="", sub="", ylab="", cex.main=1.5, cex.lab=0.
plot(hclust(mydist, method="single"), labels=labs, col="blue", hang=0.2,
main="Single Linkage", xlab="", sub="", ylab="", cex.main=1.5, cex.lab=0.8
```

Hierarchical Clustering for NCI60 Data

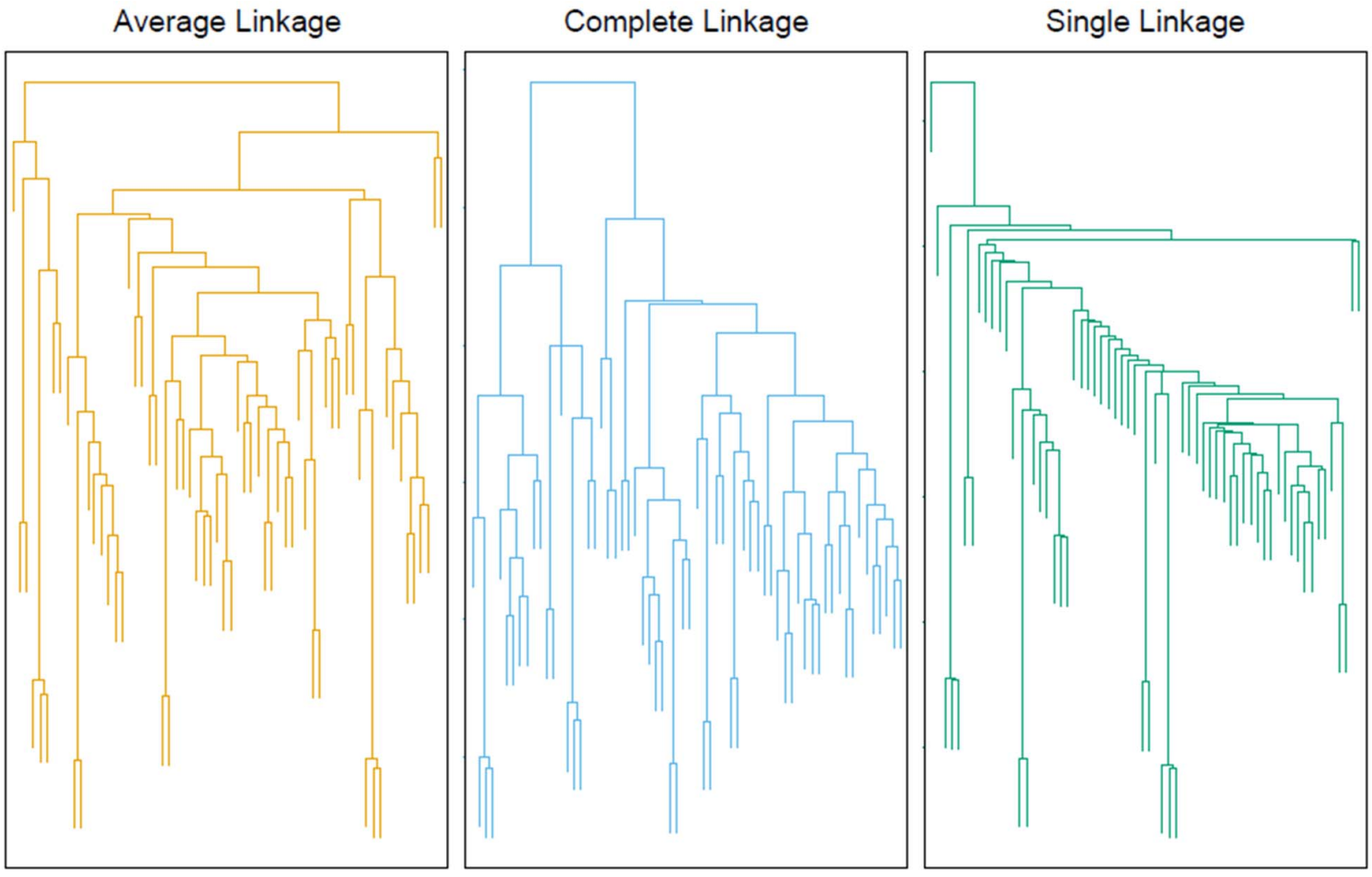


We will talk more about the second approach in the next lecture...

Exercise

- The analysis here is performed using **Euclidean distance**. *Repeat the analysis using Spearman correlations, and compare the results*.
  Hint: A good starting point is: `as.dist(1-cor(t(x)))`

Which Linkage Function?



Properties of Hierarchical Clustering

- Advantages: Gives a family of possible solutions; computationally fast

Properties of Hierarchical Clustering

- Advantages: Gives a family of possible solutions; computationally fast
- Disadvantages: No optimization criterion; final solution chosen by the data analyst; different merging (splitting) criteria give different solutions

Bi-Clustering

- Clustering variables and samples *simultaneously*

Bi-Clustering

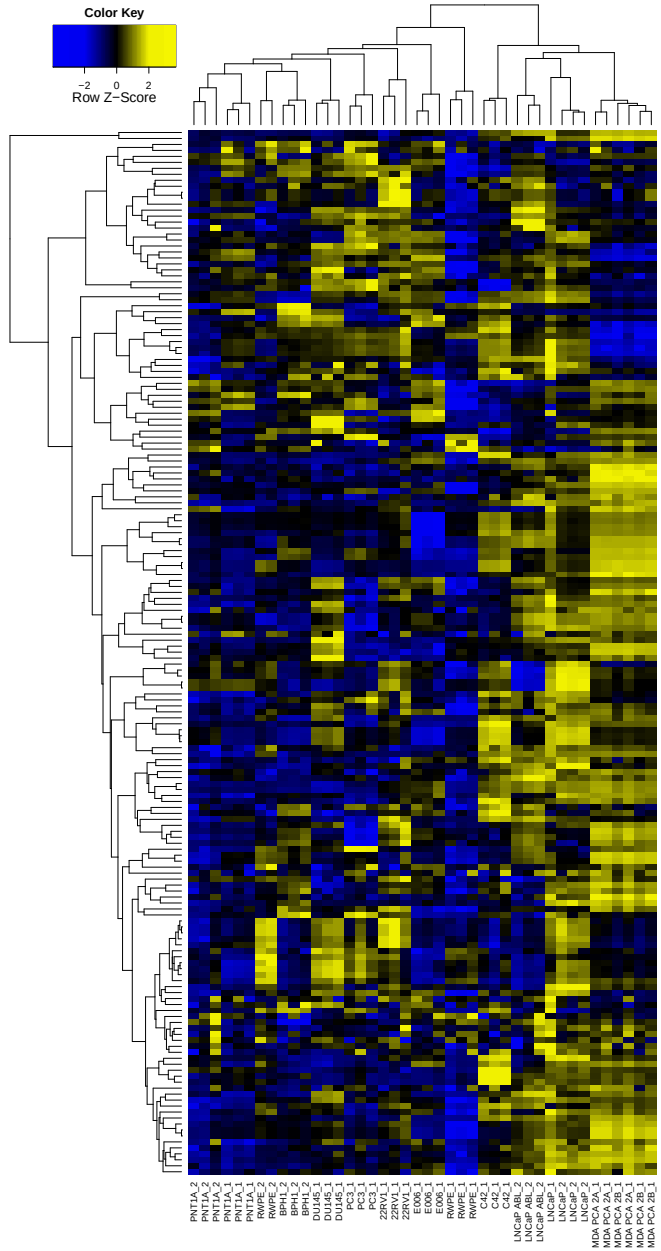
- Clustering variables and samples *simultaneously*
- Useful for e.g. finding subgroups of genes with similar activity in subclasses of cancer patients

Bi-Clustering

- Clustering variables and samples *simultaneously*
- Useful for e.g. finding subgroups of genes with similar activity in subclasses of cancer patients
- *Heatmap* of metabolite abundances in cancer cell lines

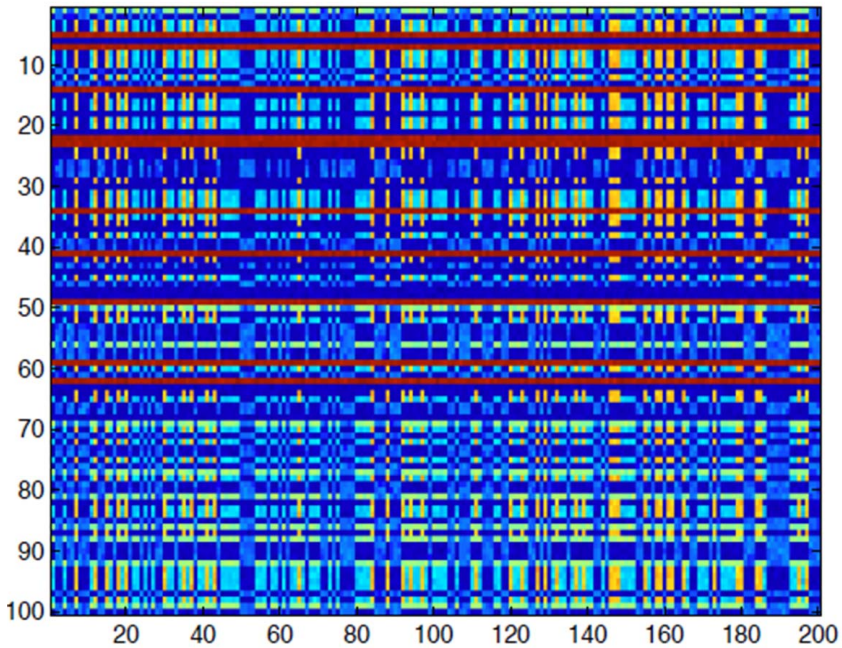
Bi-Clustering

- Clustering variables and samples *simultaneously*
- Useful for e.g. finding subgroups of genes with similar activity in subclasses of cancer patients
- *Heatmap* of metabolite abundances in cancer cell lines



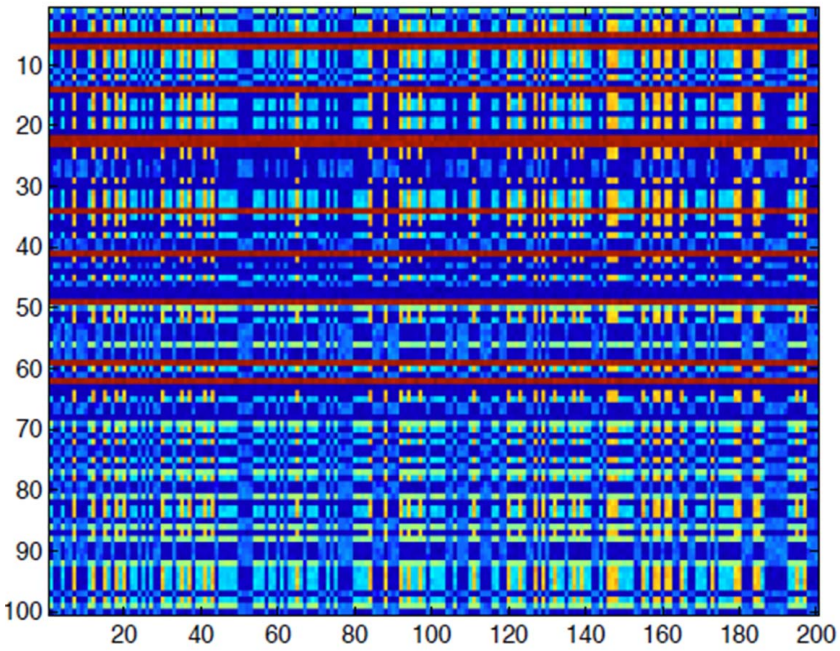
Bi-Clustering: Main Idea

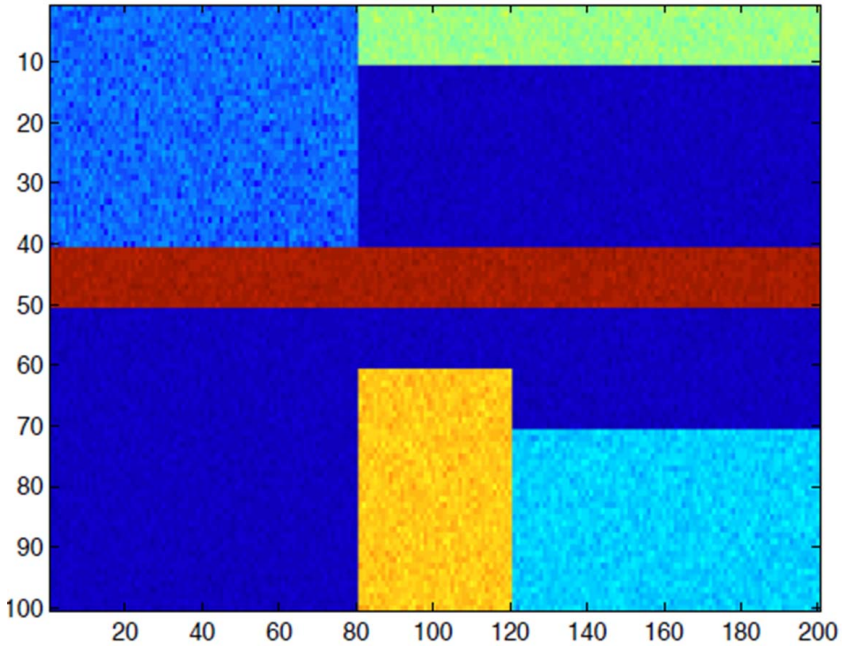
Find a rearrangement of rows and columns that gives meaningful partitions



Bi-Clustering: Main Idea

Find a rearrangement of rows and columns that gives meaningful partitions





Next Time

- Other clustering methods
  - K-means clustering
  - Model-based clustering
  - Spectral clustering (briefly)