Business Statistics

Clustering

Weichen Wang

Assistant Professor

Innovation and Information Management

ISLR Chapter 12.4

Beer-Diaper Syndrome

Transaction No.	Item 1	Item 2	Item 3	…Item N
100	Beer	Diaper	Chocolate	
101	Milk	Chocolate	Shampoo	
102	Beer	Wine	Vodka	
103	Beer	Cheese	Diaper	
104	Ice Cream	Diaper	Beer	
…				

Beer-Diaper Syndrome

Trans No.	Item 1	Item 2	Item 3	…	Day	Time	Customer Info.
100	Beer	Diaper	Chocolate		Fri	6:15pm	Male, 30, …
101	Milk	Chocolate	Shampoo		Sun	10:10am	Female, 25,…
102	Beer	Wine	Vodka		Sat	5:30pm	Male, 24,…
103	Beer	Cheese	Diaper		Fri	6:30pm	Male, 32,…
104	Ice Cream	Diaper	Beer		Fri	7:00pm	Male, 28,…
…							

Stock Daily Closing Price 2000-2013

Altria (formerly Philip Morris; MO)
Apple (AAPL)
Automatic Data Processing (ADP)
Corrections Corporation of America (CXW)
Equifax (EFX)
Ford (F)
General Electric (GE)
Graham Holding Companies (GHC)
Proctor and Gamble (PG)
United States Steel (X)
Yahoo! (YHOO)
Amazon (AMZN)
Archer Daniels Midland (ADM)
Bank of America (BAC)
Dow Chemicals (DOW)
ExxonMobil (XOM)
Halliburton (HAL)
Goldman Sachs (GS)
Microsoft (MSFT)
Time Warner (TWX)
Walmart (WMT)
Yum! Brands (YUM)

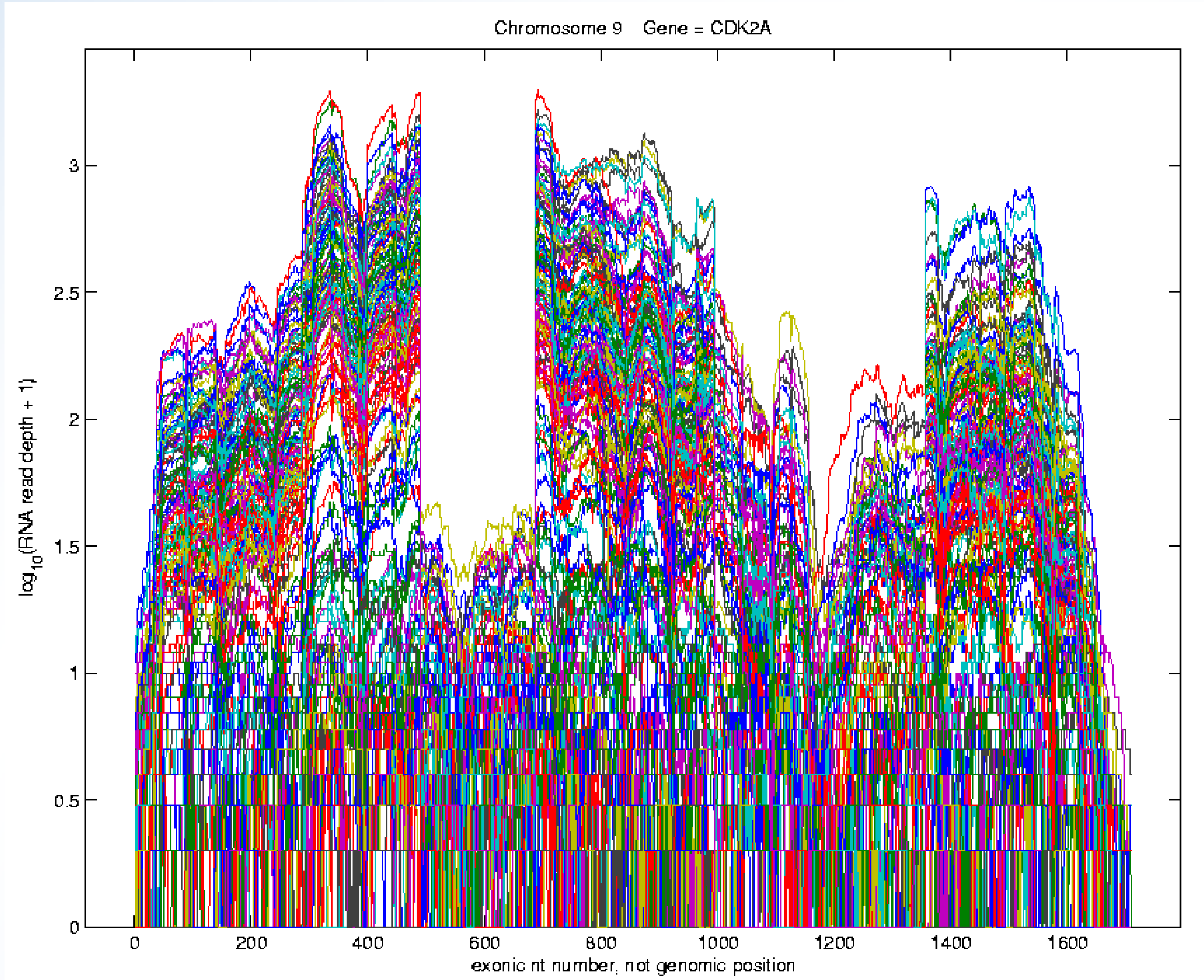
Clustering after PCA

United States Steel (X)
Dow Chemicals (DOW)
ExxonMobil (XOM)
Halliburton (HAL)
Equifax (EFX)
Ford (F)
Archer Daniels Midland (ADM)
Graham Holding Companies (GHC)
General Electric (GE)
Bank of America (BAC)

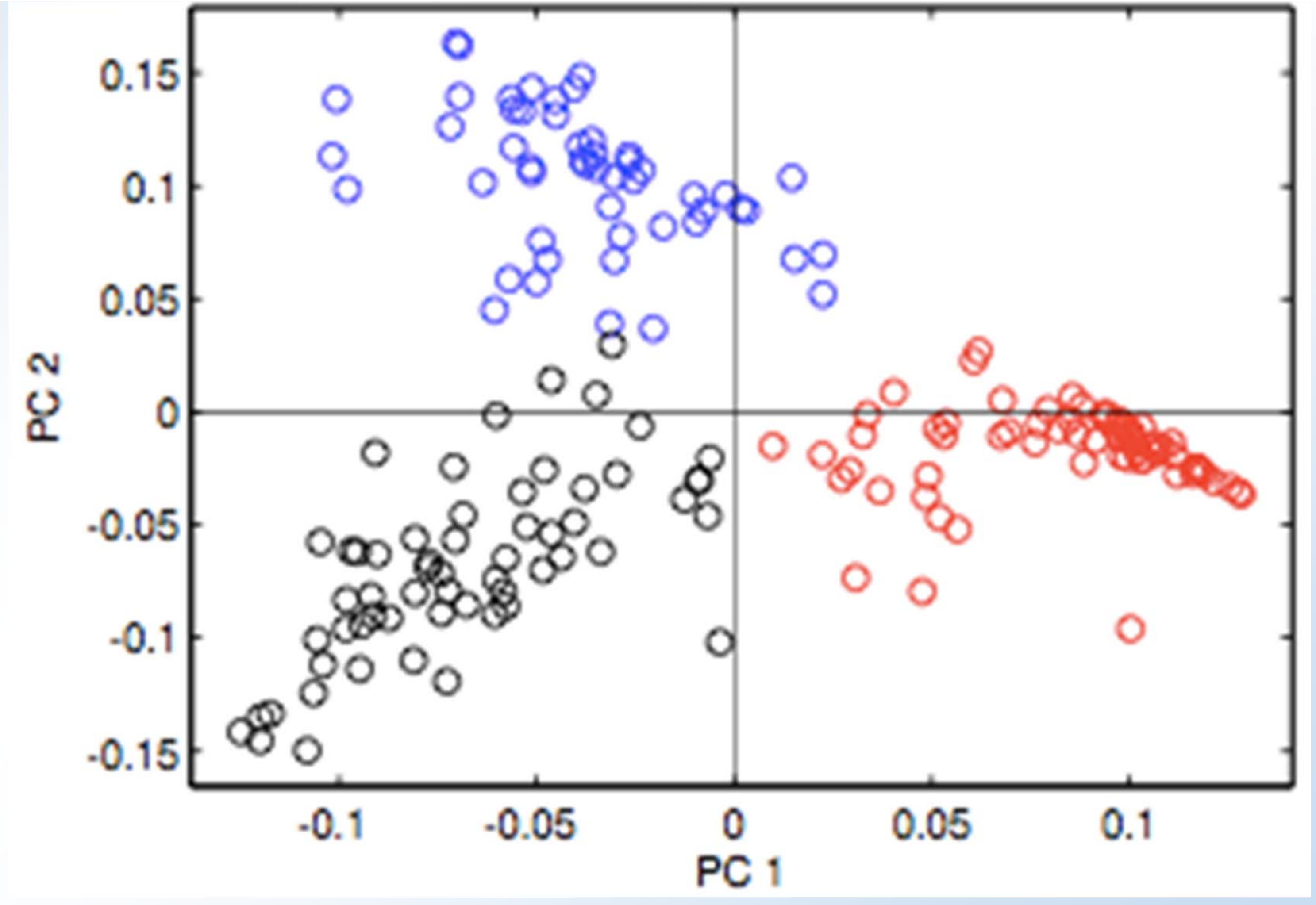
Amazon (AMZN)
Apple (AAPL)
Corrections Corporation of America (CXW)
Goldman Sachs (GS)
Microsoft (MSFT)
Time Warner (TWX)
Yahoo! (YHOO)

Altria (formerly Philip Morris; MO)
Proctor and Gamble (PG)
Walmart (WMT)
Yum! Brands (YUM)
Automatic Data Processing (ADP)

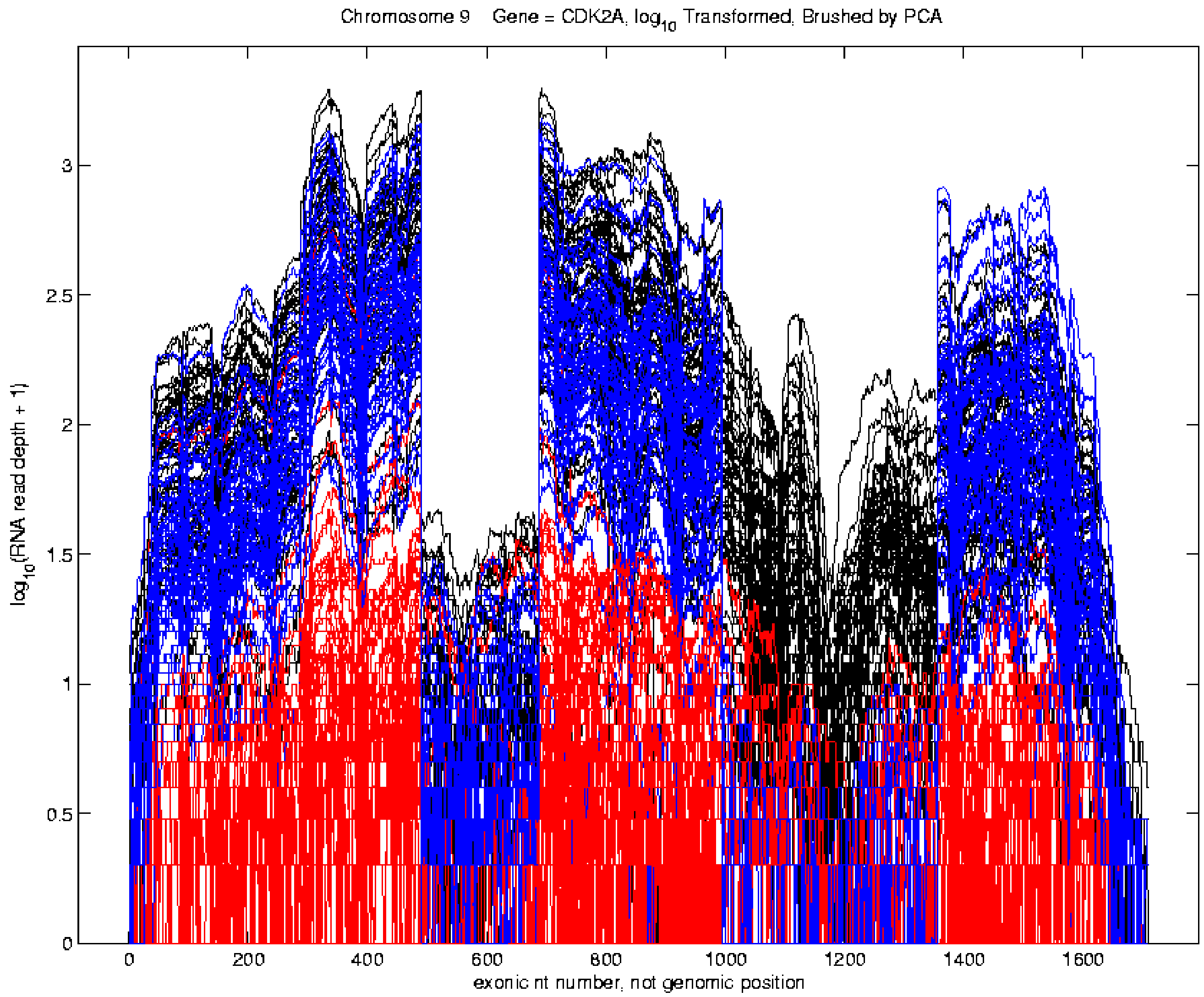
Gene CDK2A RNASeq: 180 Samples, 1700 Locations



Clustering on the PCA Scores

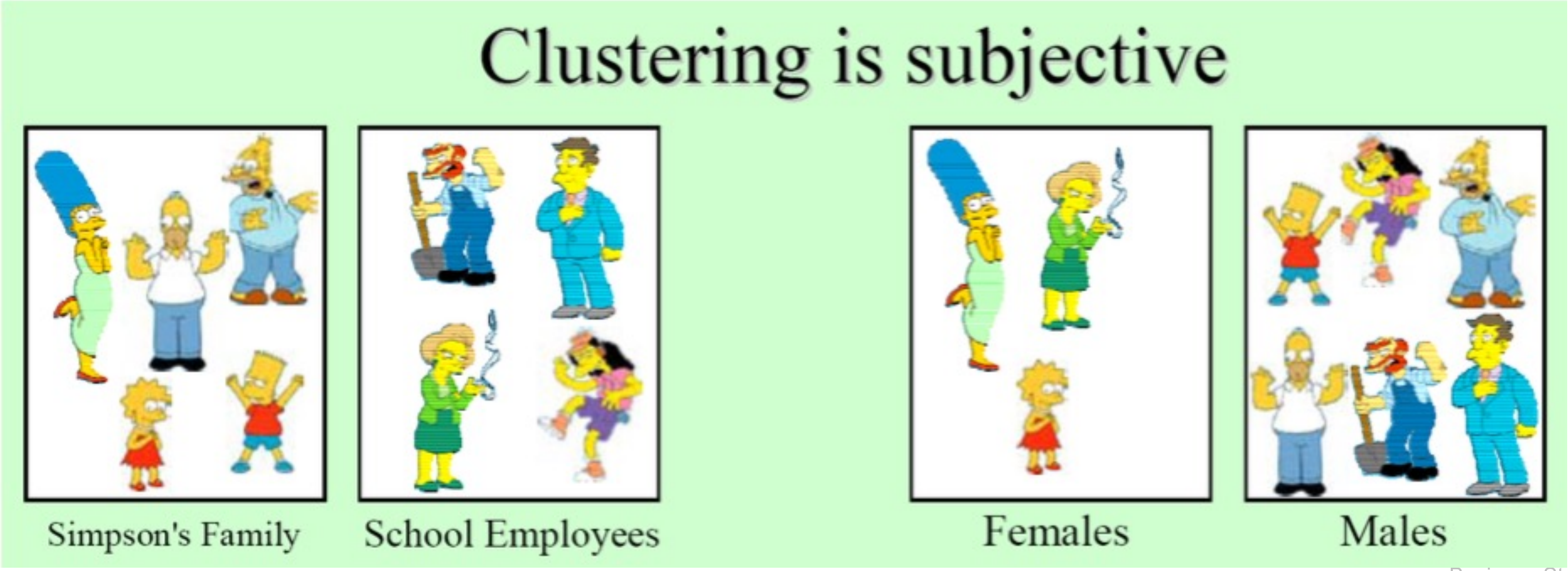


Back to the Gene Profiles



What is Clustering?

- Clustering: the process of grouping a set of objects into classes of similar objects
  - Need a similarity measure
  - High intra-class similarity
  - Low inter-class similarity



PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.

- Clustering looks for homogeneous subgroups among the observations.

Clustering

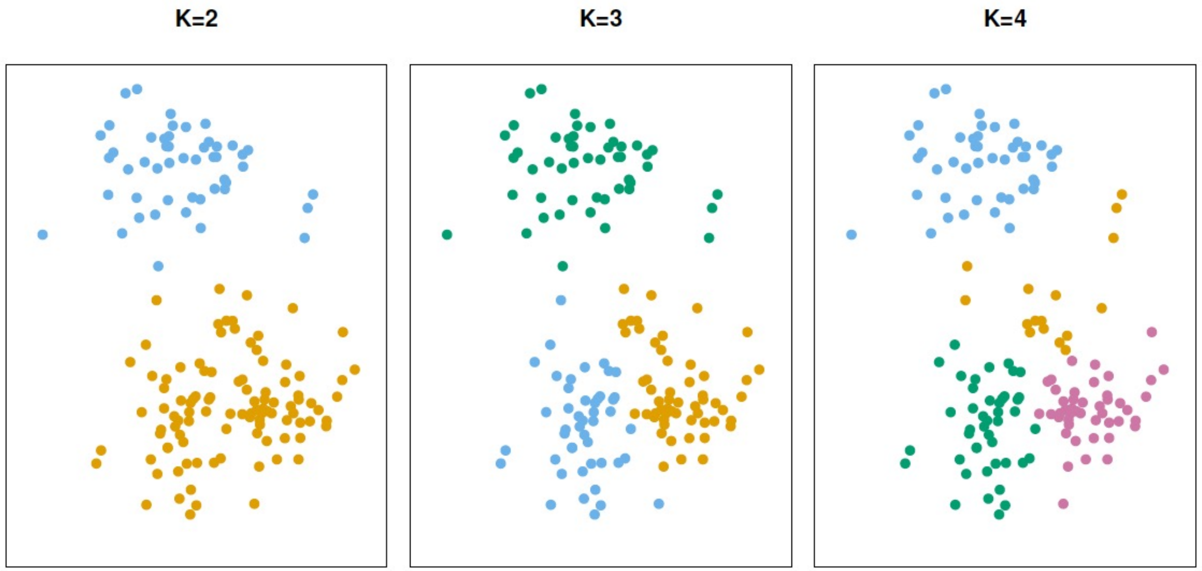
- Marketing: customer segmentation (discovery of distinct groups of customers) for target marketing
- Car insurance: identification of customer groups with high average claim cost
- Stock selection: groups of stocks that have similar trends
- Netflix recommendation: viewers with similar taste for movies
- Flatiron health: identification of patient subgroups that certain treatment works the best

Two Clustering Methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters

- In hierarchical clustering, we do not know in advance how many clusters we want. We end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n

K-means Clustering



- Simulated data with 150 observations in 2-dimensional space
- Panels show the results of applying K-means clustering with different values of K, the number of clusters
- The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm

K-means Clustering

Given K,

- Assign obs into K non-overlapping clusters
- Minimize the within-cluster variation

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

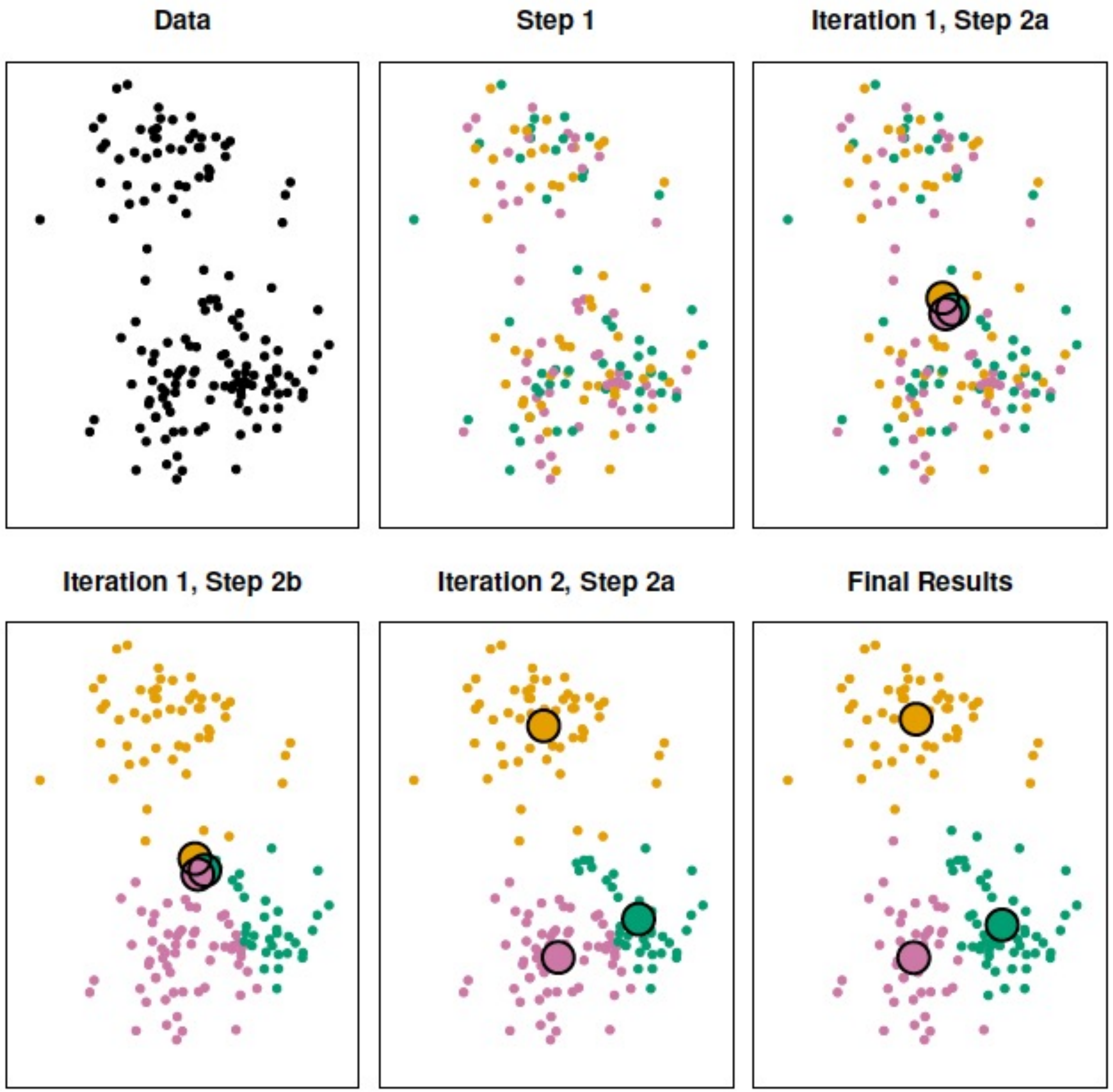
K-Means Clustering Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:
   – 2.1. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   – 2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Example



Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Details of the Previous Figure

The progress of the K-means algorithm with K=3.

- *Top left:* The observations are shown.
- *Top center:* In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- *Top right:* In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- *Bottom left:* In Step 2(b), each observation is assigned to the nearest centroid.
- *Bottom center:* Step 2(a) is once again performed, leading to new cluster centroids.
- *Bottom right:* The results obtained after 10 iterations.

K-means are sensitive to initial assignment

Run multiple times from different random initial configurations and select the results with the minimal objective.

Simple but often not ideal

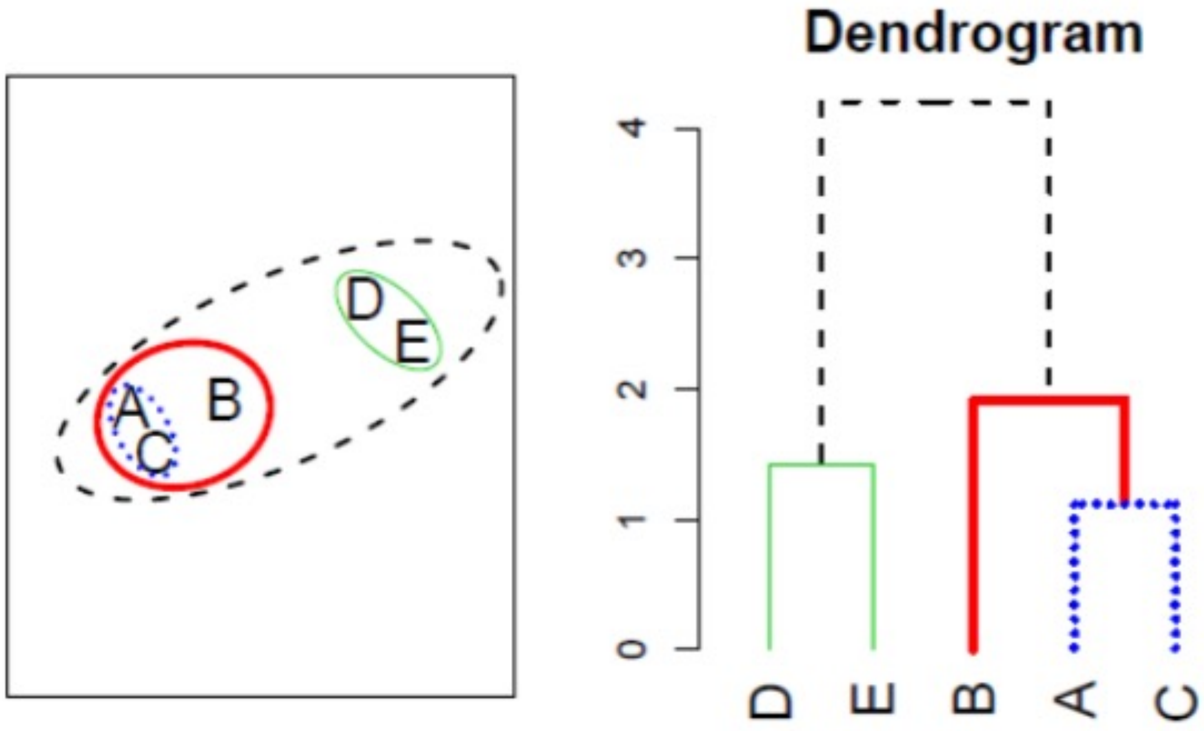
- Variable results with noisy data and outliers
- Very large or very small values can skew the centroid positions, and give poor clustering
- Only suitable for the cases where we can expect clusters to be `clumps' that are close together – e.g. terrible in the two-spirals and similar cases

Hierarchical Clustering

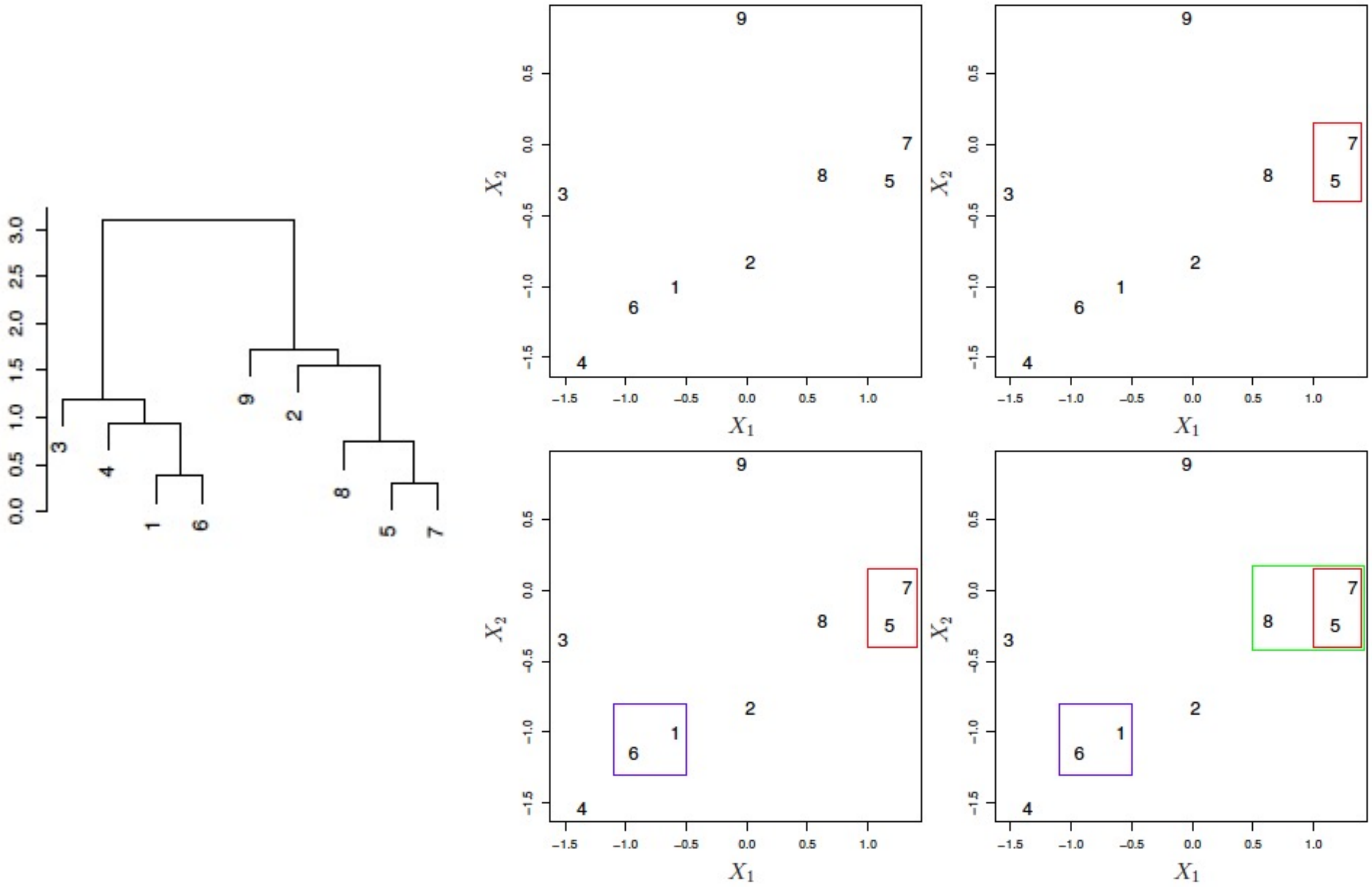
- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- Hierarchical clustering is an alternative approach which does not require a particular choice of K.
- Next, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

Hierarchical Clustering: Bottom-Up

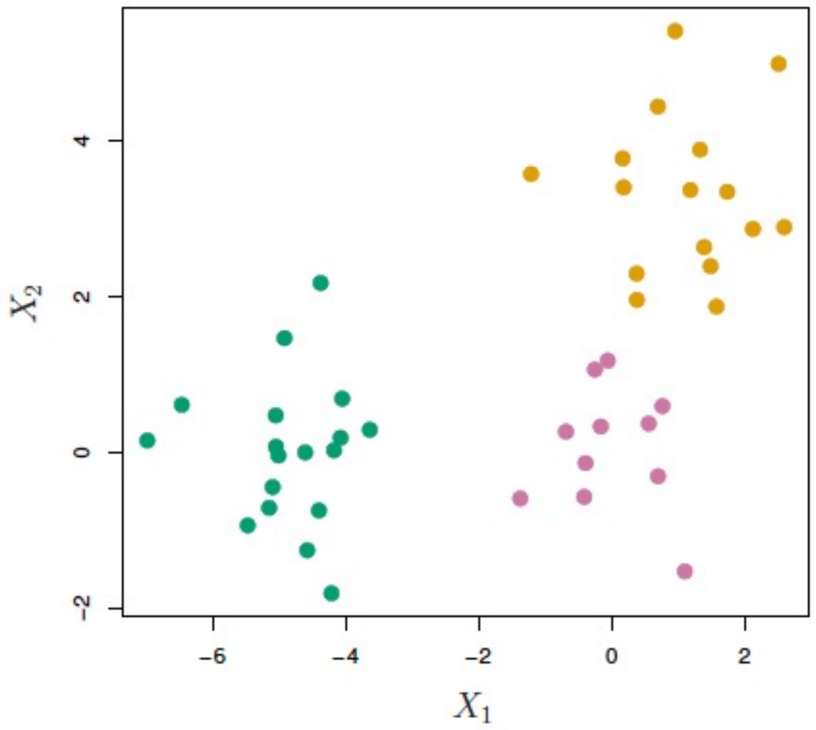
- Start with each point as its own cluster
- Identify the closest two clusters and merge them
- Repeat until all points are in a single cluster
- Axis: distance/dissimilarity between clusters



Example

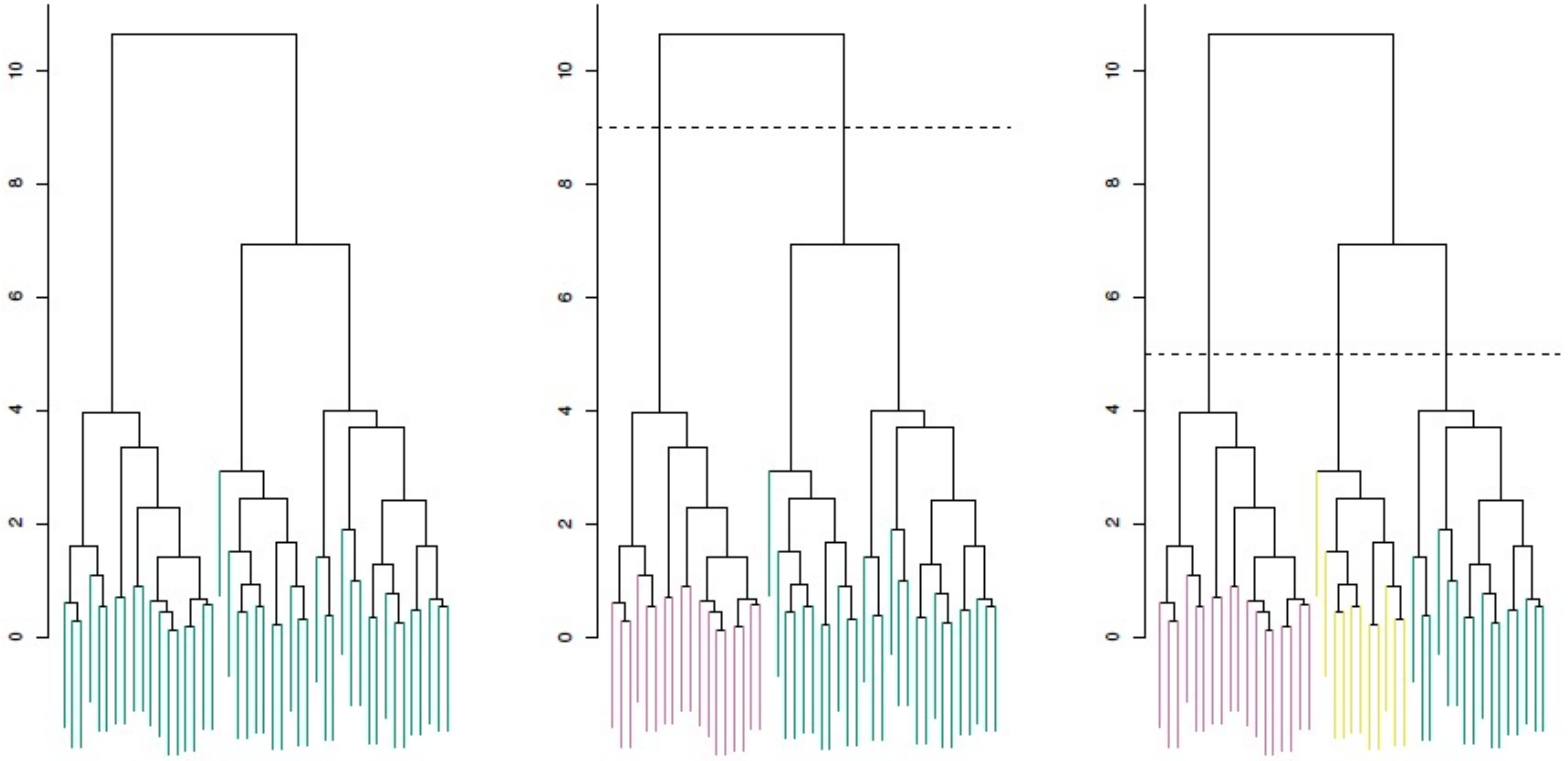


Another Example



- 45 observations in 2-dimensional space.
- Three distinct classes, shown in separate colors.
- Pretend these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

Application of Hierarchical Clustering



Details of the Previous Figure

- *Left:* Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- *Center:* The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- *Right:* The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure

Hierarchical Clustering Algorithm

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

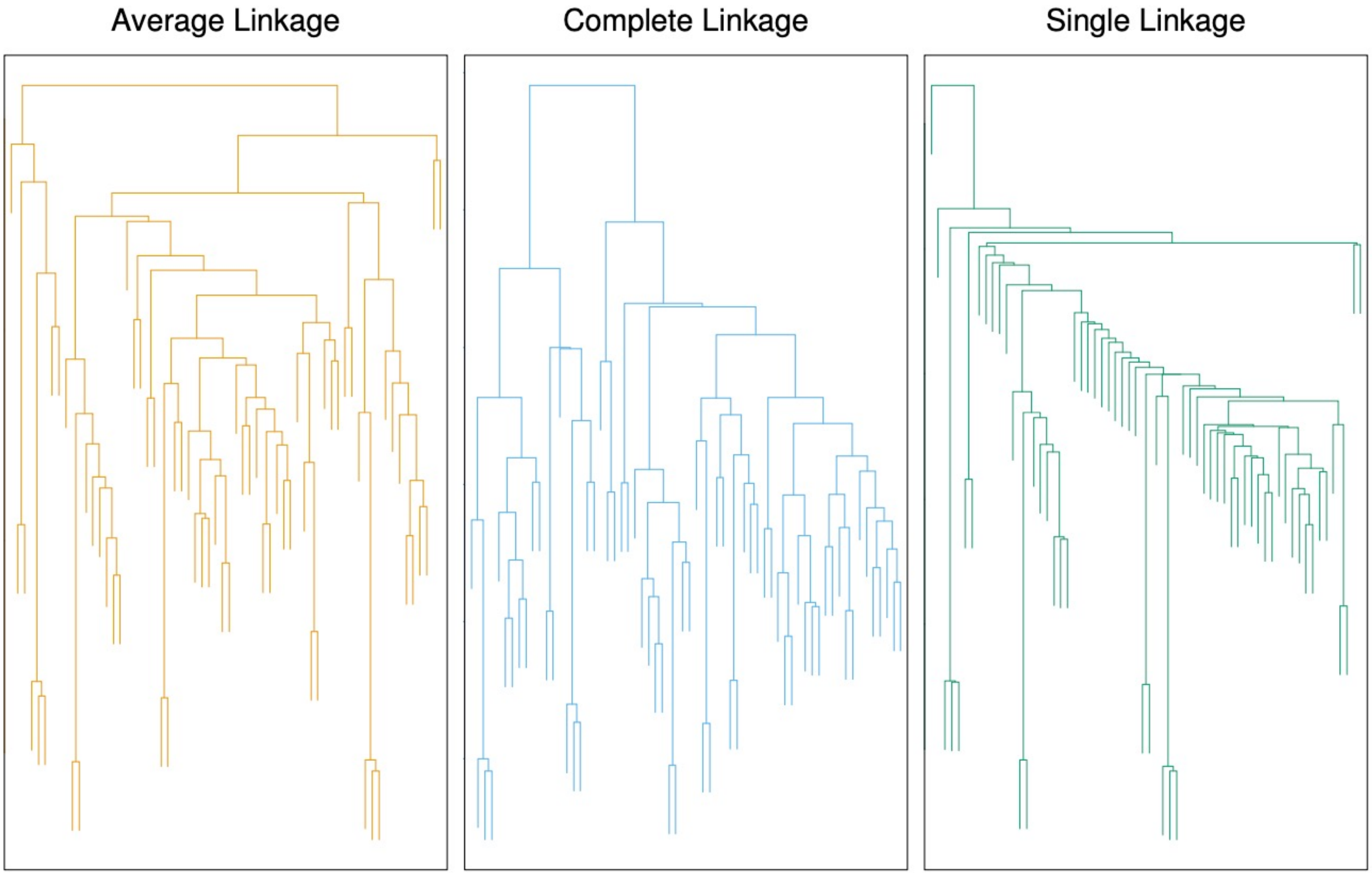
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Types of Linkage (Dissimilarity Measure)

Linkage	Description
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B.

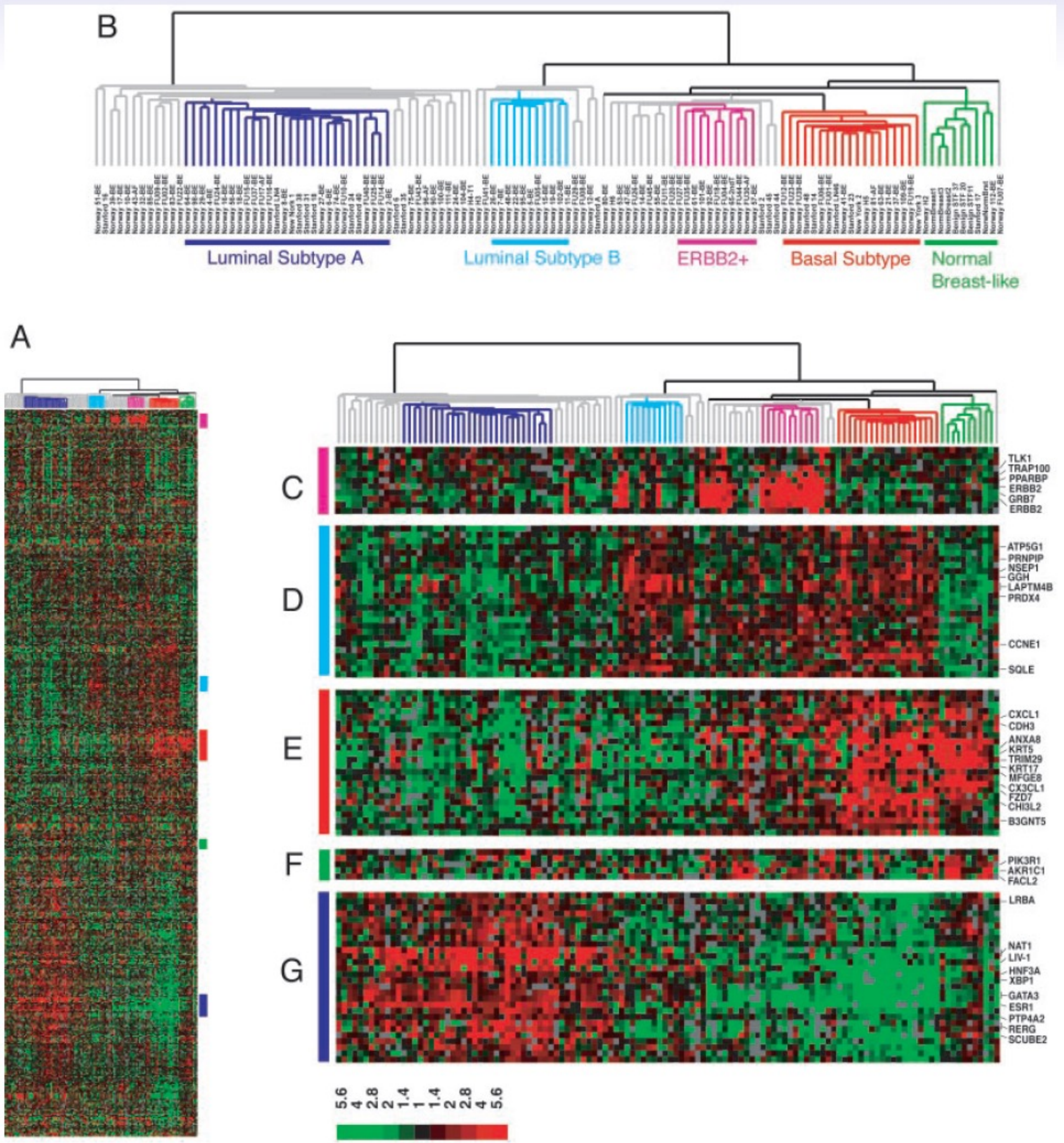
Hierarchical Agglomerative Clustering

- Is very commonly used
- Very different from *K*-means
- Provides a much richer structuring of the data
- No need to choose *K*
- But, quite sensitive to the various dissimilarity measures
  - E.g., considering clustering shoppers, Euclidean distance leads to infrequent shoppers being grouped together, while correlation distance leads to shoppers with similar preference being clustered.
- And, quite sensitive to the variable scales similar to *K*-means and PCA
  - E.g., if we do not scale to SD=1, high-frequency purchases like socks will have a much larger effect in dissimilarity of shoppers than rare purchases like computers.


Average Linkage
Complete Linkage
Single Linkage

Gene Expression Data

- "Repeated observation of breast tumor subtypes in independent gene expression data sets;" Sorlie at el, PNAS 2003
- Average linkage, correlation metric
- Clustered samples using 500 *intrinsic genes:* each woman was measured before and after chemotherapy. Intrinsic genes have smallest within/between variation.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6

Things to Consider for Clustering

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be scaled to have standard deviation one.
- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?
- In the case of K-means clustering, how many clusters should we look for in the data?

Better to try different things with subsets of the data to ensure the clustering is stable.

One possibility: Looking for "elbow" of the decay of some objective.

Summary of Unsupervised Learning

- PCA: dimension reduction, data visualization, missing data imputation, downstream regression/classification…

- Clustering: discovery of underlying data groups, bottom-up hierarchical vs K-means, dissimilarly measures…

What Have We Learnt?

- Supervised learning:
  - Regression
    - Linear / Transformation
    - Forward / Backward / Subset selection
    - Ridge / Lasso / CV
    - Poisson / GLM
    - PCR / PLS
  - Classification
    - Logistic / Multinomial
    - LDA / QDA / Naive Bayes
    - SVM
- Unsupervised learning:
  - PCA
    - Data visualization
    - Dimension reduction
    - Matrix completion
  - Clustering
    - K-means
    - Hierarchical

More Advanced Future Topics

– Nonlinearity, Splines, Local regression
– Decision tree, Random forests, Boosting
– Kernel method, SVM with kernels
– Deep learning, CNN, RNN
– Time series model, ARIMA, GARCH
– Endogeneity, Instrumental variables
– Survival Analysis
– High dimensional inference
– Casual inference