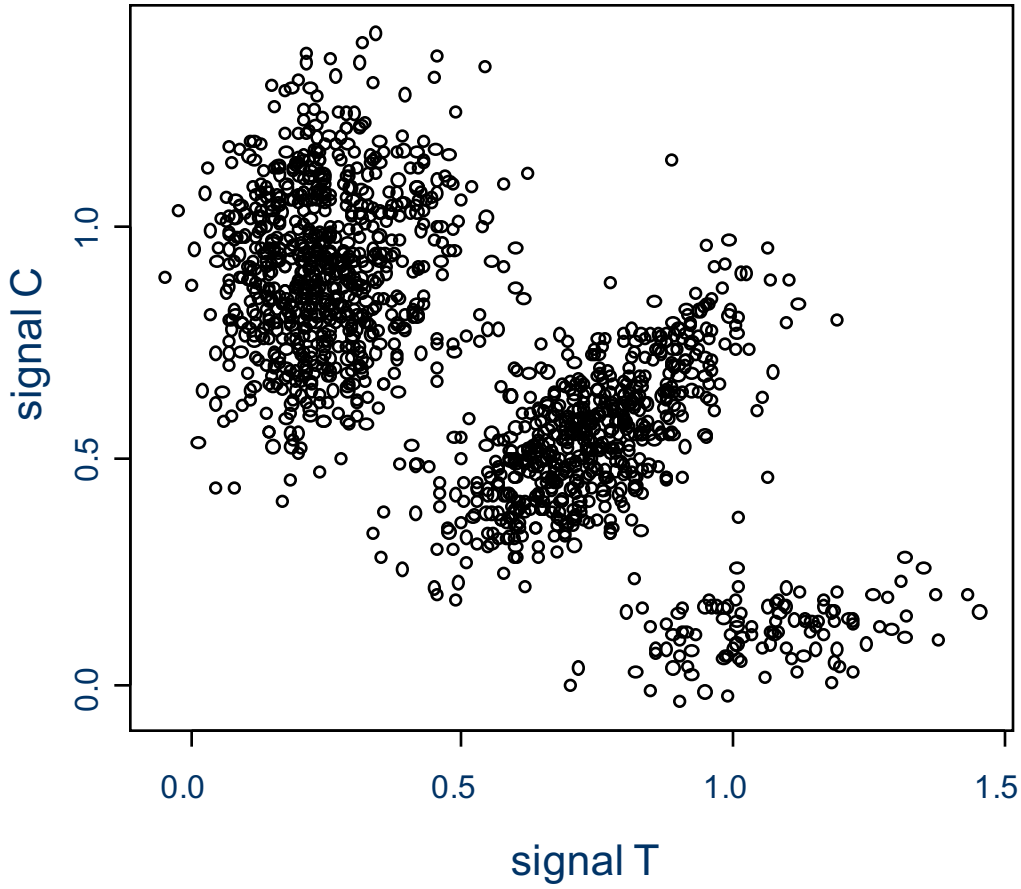
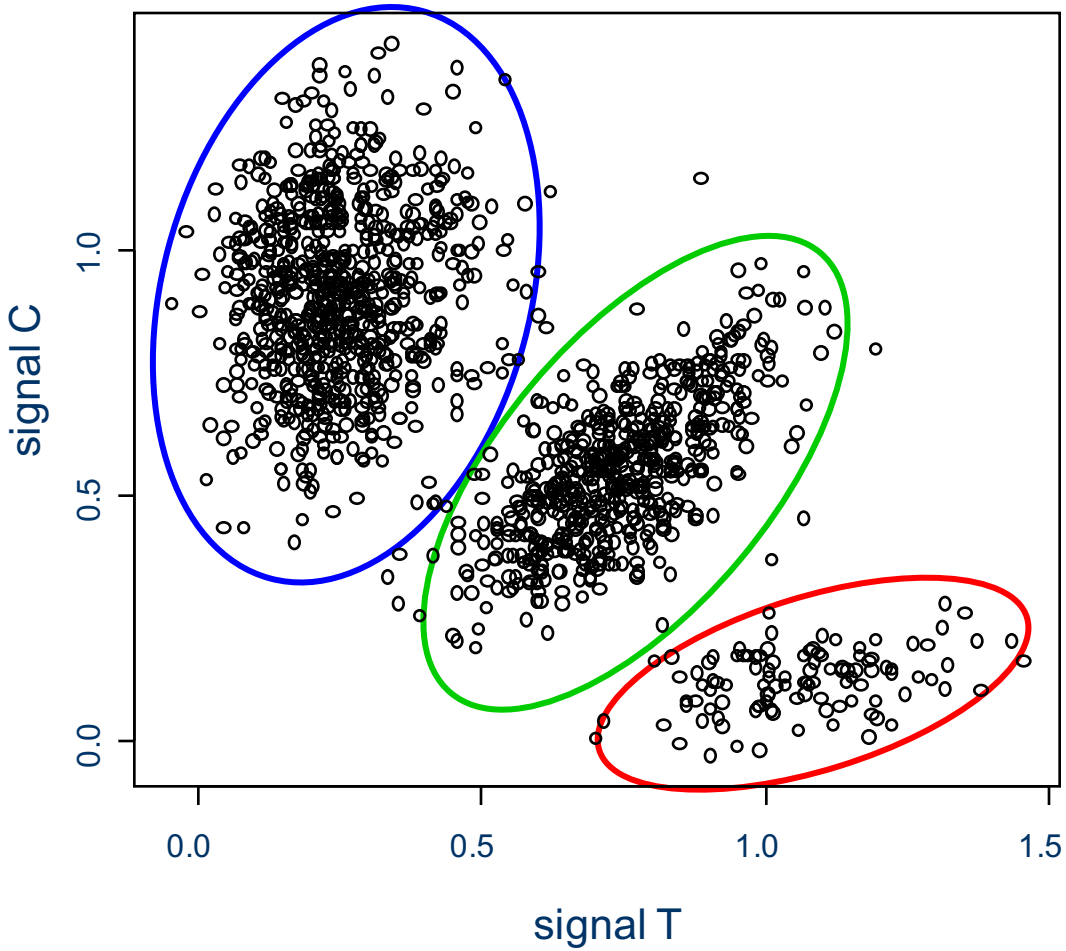
An Introduction to Cluster Analysis

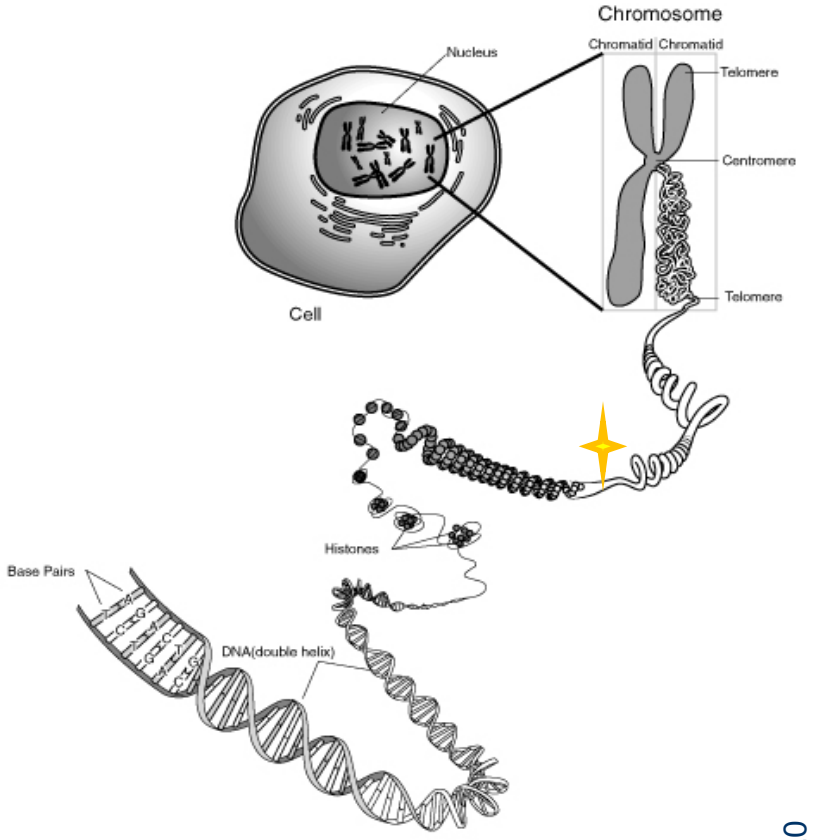
Zhaoxia Yu
Department of Statistics
Vice Chair of Undergraduate Affairs
zhaoxia@ics.uci.edu

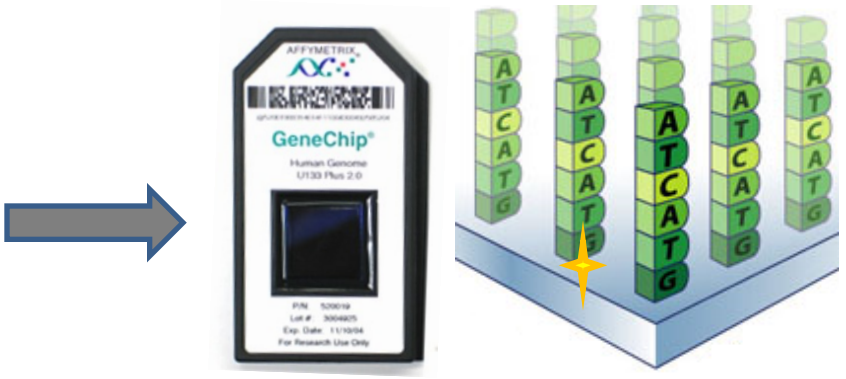
What can you say about the figure?

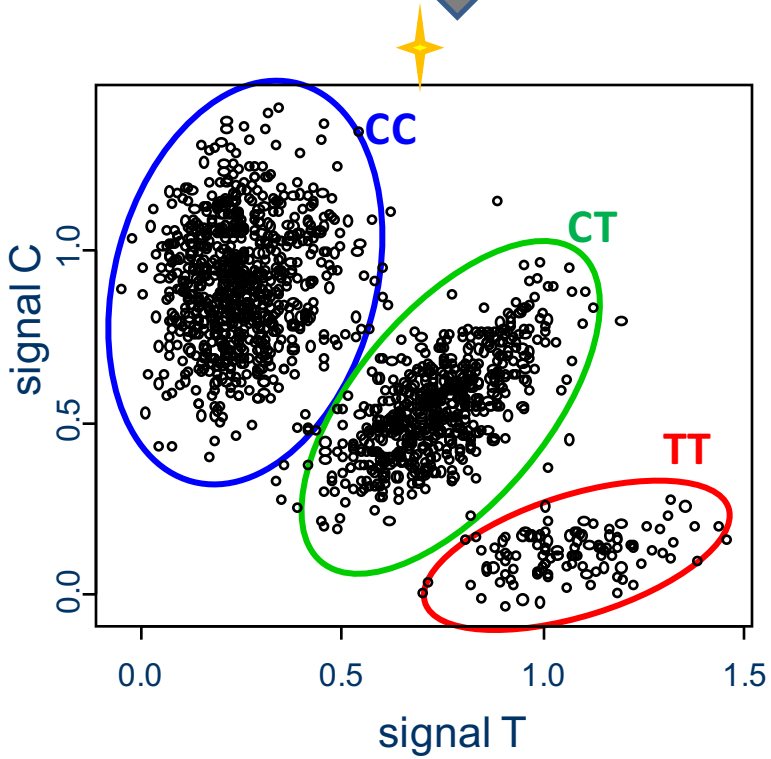


- ≈ 1500 subjects
- Two measurements per subject


signal C
0.0
0.5
1.0
0.0
0.5
1.0
1.5
signal T


Chromosome
Nucleus
Chromatid
Chromatid
Telomere
Centromere
Telomere
Cell
Histones
Base Pairs
DNA(double helix)


GeneChip


CC
CT
TT
signal C
signal T
0.0
0.5
1.0
1.5
0.0
0.5
1.0

Cluster Analysis

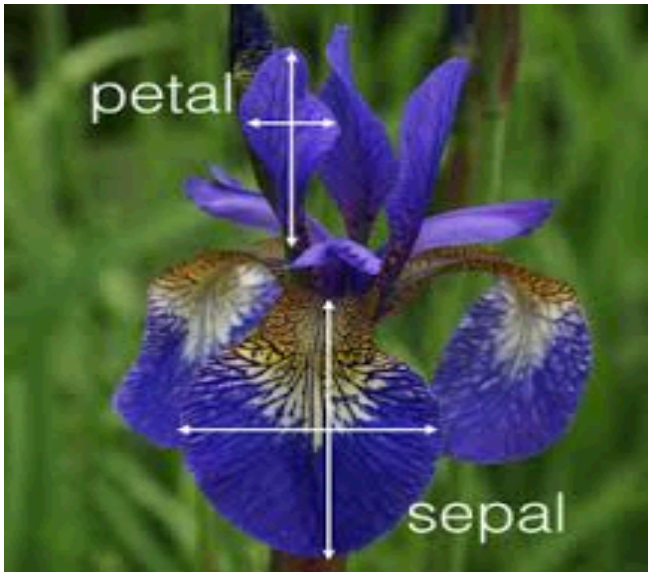
- Seeks rules to group data
  - Large between-cluster difference
  - Small within-cluster difference
- Exploratory
- Aims to understand/ learn the unknown substructure of multivariate data

Cluster Analysis vs Classification

Cluster Analysis	Classification
• Data are unlabeled	• The labels for training data are known
• The number of clusters are unknown	• The number of classes are known
• "Unsupervised" learning	• "Supervised" learning
• Goal: find unknown structures	• Goal: allocate new observations, whose labels are unknown, to one of the known classes

The Iris Data

- It was collected by F. A. Fisher
- A famous data set that has been widely used in textbooks
- Four features:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm



The Iris Data

- Three types:
  - Setosa
  - Versicolor
  - Virginica

The Iris Data

	Sepal L.	Sepal W.	Petal L.	Petal W.
[1,]	5.1	3.5	1.4	0.2
[2,]	4.9	3.0	1.4	0.2
[3,]	4.7	3.2	1.3	0.2
[4,]	4.6	3.1	1.5	0.2
[5,]	5.0	3.6	1.4	0.2
[6,]	5.4	3.9	1.7	0.4
[7,]	4.6	3.4	1.4	0.3
[8,]	5.0	3.4	1.5	0.2
[9,]	4.4	2.9	1.4	0.2
…	…	…	…	
[45,]	5.1	3.8	1.9	0.4
[46,]	4.8	3.0	1.4	0.3
[47,]	5.1	3.8	1.6	0.2
[48,]	4.6	3.2	1.4	0.2
[49,]	5.3	3.7	1.5	0.2
[50,]	5.0	3.3	1.4	0.2

Iris Setosa

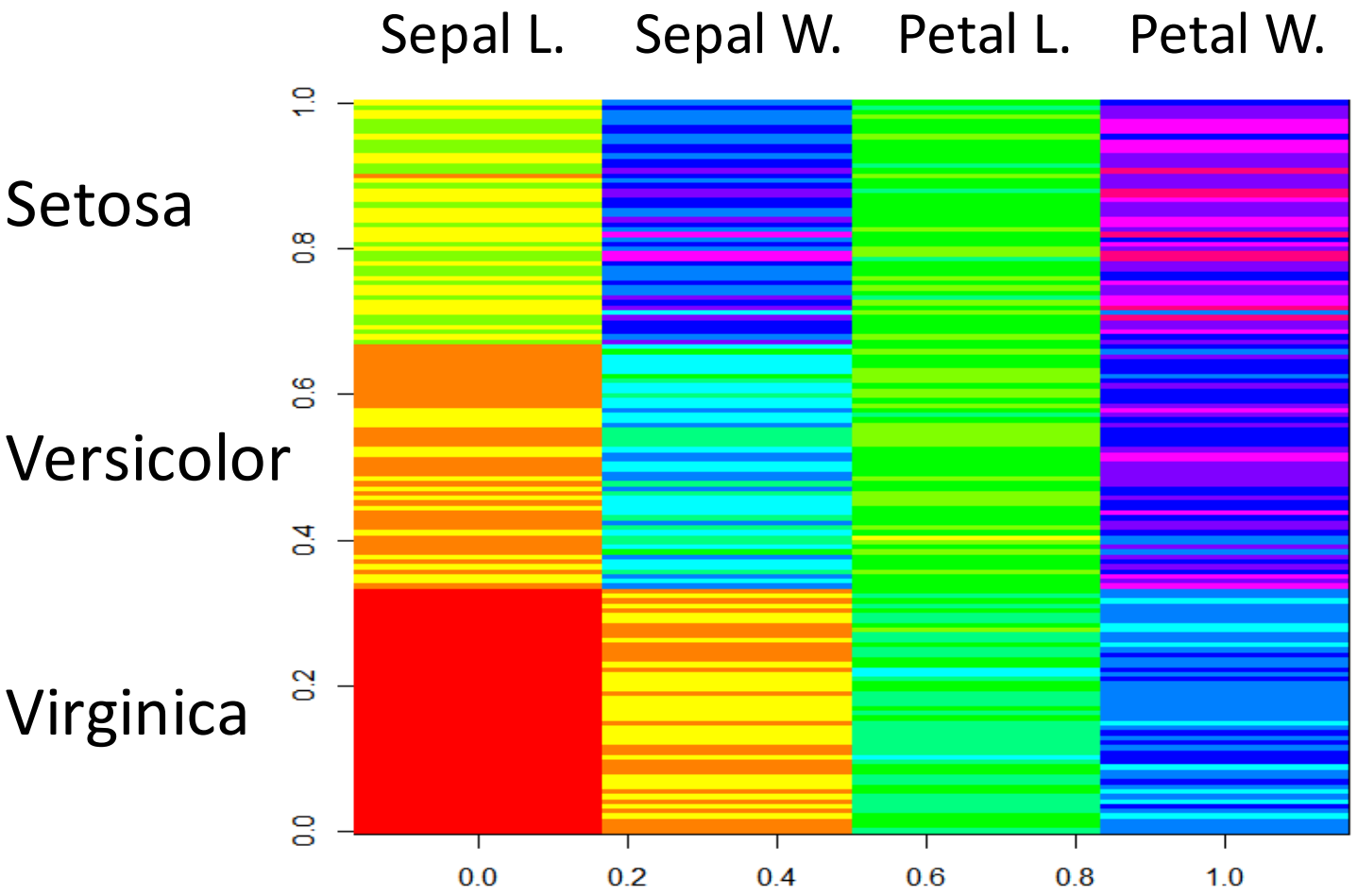
	Sepal L.	Sepal W.	Petal L.	Petal W.
[1,]	7.0	3.2	4.7	1.4
[2,]	6.4	3.2	4.5	1.5
[3,]	6.9	3.1	4.9	1.5
[4,]	5.5	2.3	4.0	1.3
[5,]	6.5	2.8	4.6	1.5
[6,]	5.7	2.8	4.5	1.3
[7,]	6.3	3.3	4.7	1.6
[8,]	4.9	2.4	3.3	1.0
[9,]	6.6	2.9	4.6	1.3
…	…	…	…	
[45,]	5.6	2.7	4.2	1.3
[46,]	5.7	3.0	4.2	1.2
[47,]	5.7	2.9	4.2	1.3
[48,]	6.2	2.9	4.3	1.3
[49,]	5.1	2.5	3.0	1.1
[50,]	5.7	2.8	4.1	1.3

Iris Versicolor

	Sepal L.	Sepal W.	Petal L.	Petal W.
[1,]	6.3	3.3	6.0	2.5
[2,]	5.8	2.7	5.1	1.9
[3,]	7.1	3.0	5.9	2.1
[4,]	6.3	2.9	5.6	1.8
[5,]	6.5	3.0	5.8	2.2
[6,]	7.6	3.0	6.6	2.1
[7,]	4.9	2.5	4.5	1.7
[8,]	7.3	2.9	6.3	1.8
[9,]	6.7	2.5	5.8	1.8
…	…	…	…	
[45,]	6.7	3.3	5.7	2.5
[46,]	6.7	3.0	5.2	2.3
[47,]	6.3	2.5	5.0	1.9
[48,]	6.5	3.0	5.2	2.0
[49,]	6.2	3.4	5.4	2.3
[50,]	5.9	3.0	5.1	1.8

Iris Virginica

The Iris Data



Clustering Methods

- Model-free:
  - Nonhierarchical clustering. K-means.
  - Hierarchical clustering. Based on similarity measures
- Model-based clustering

Model-Free Clustering
Nonhierarchical Clustering: K-Means

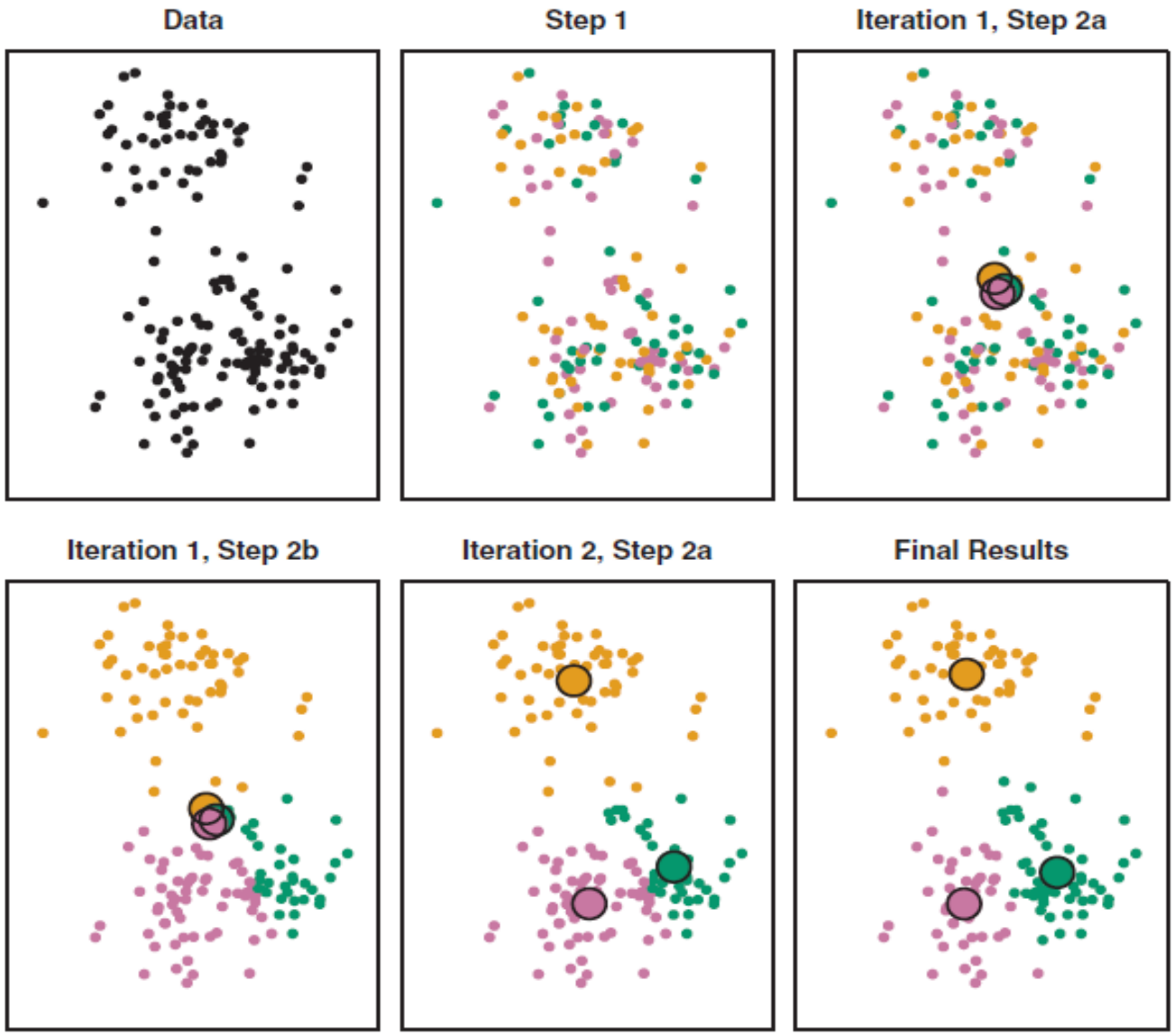
K-Means

- Assign each observation to the cluster with the nearest mean

- “Nearest” is usually defined based on Euclidean distance

K-Means: Algorithm

- Step 0: Preprocess data. Standardize data if appropriate
- Step 1: Partition the observations into *K* initial clusters.
- Step 2
  - 2.a (update step): Calculate the centroids.
  - 2.b (assignment step): Assign each observation to its nearest cluster.
- Repeat step 2 until no more changes in assignments

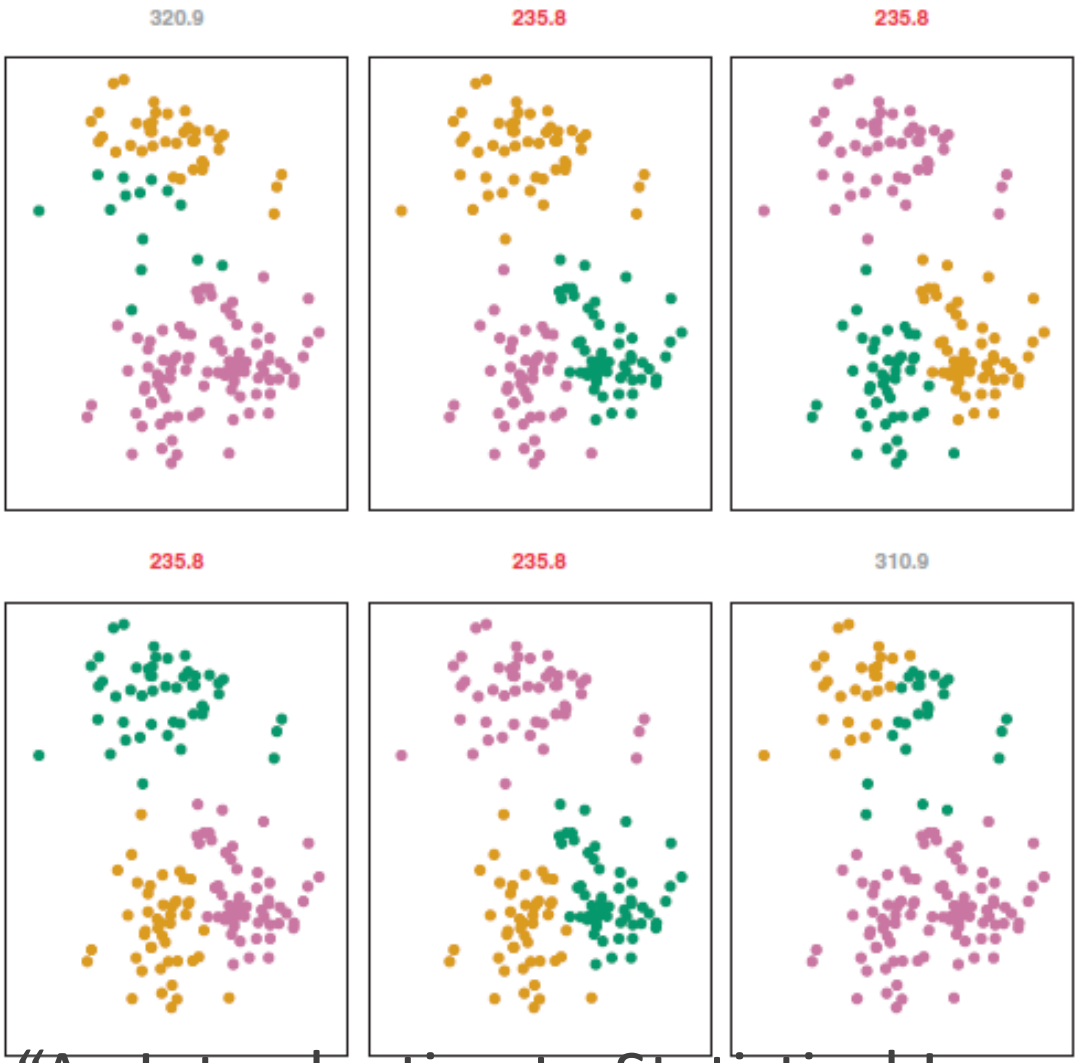


From “An Introduction to Statistical Learning”

Remarks

- Before convergence, each step is guaranteed to decrease the within-cluster sum of squares objective
- Within a finite number of steps, the algorithm might converge to a (local) minimum
- Use different and random initial values

Different Initial Values

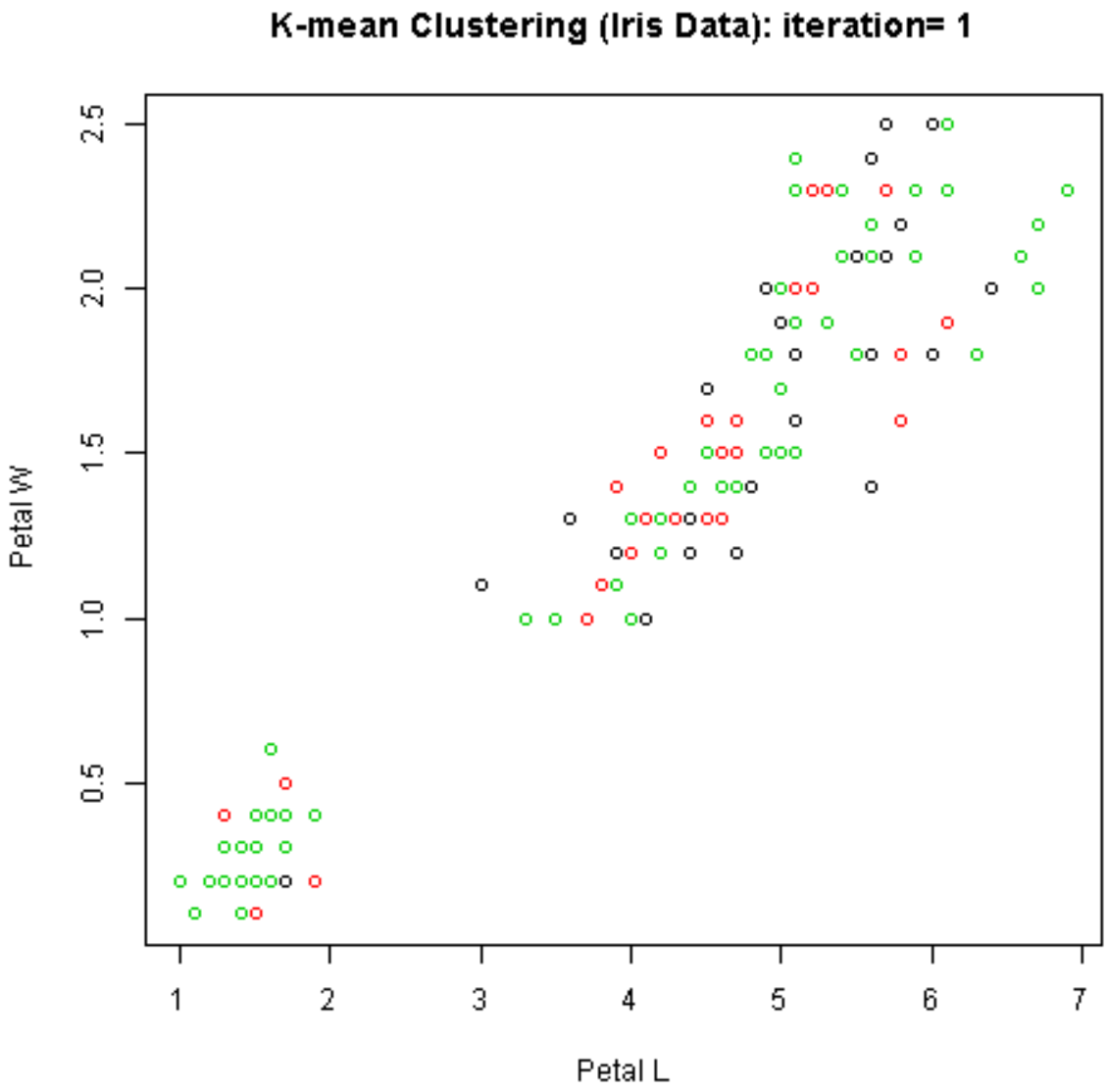


From “An Introduction to Statistical Learning”

Example: Cluster Analysis of Iris Data (Petal L & W)

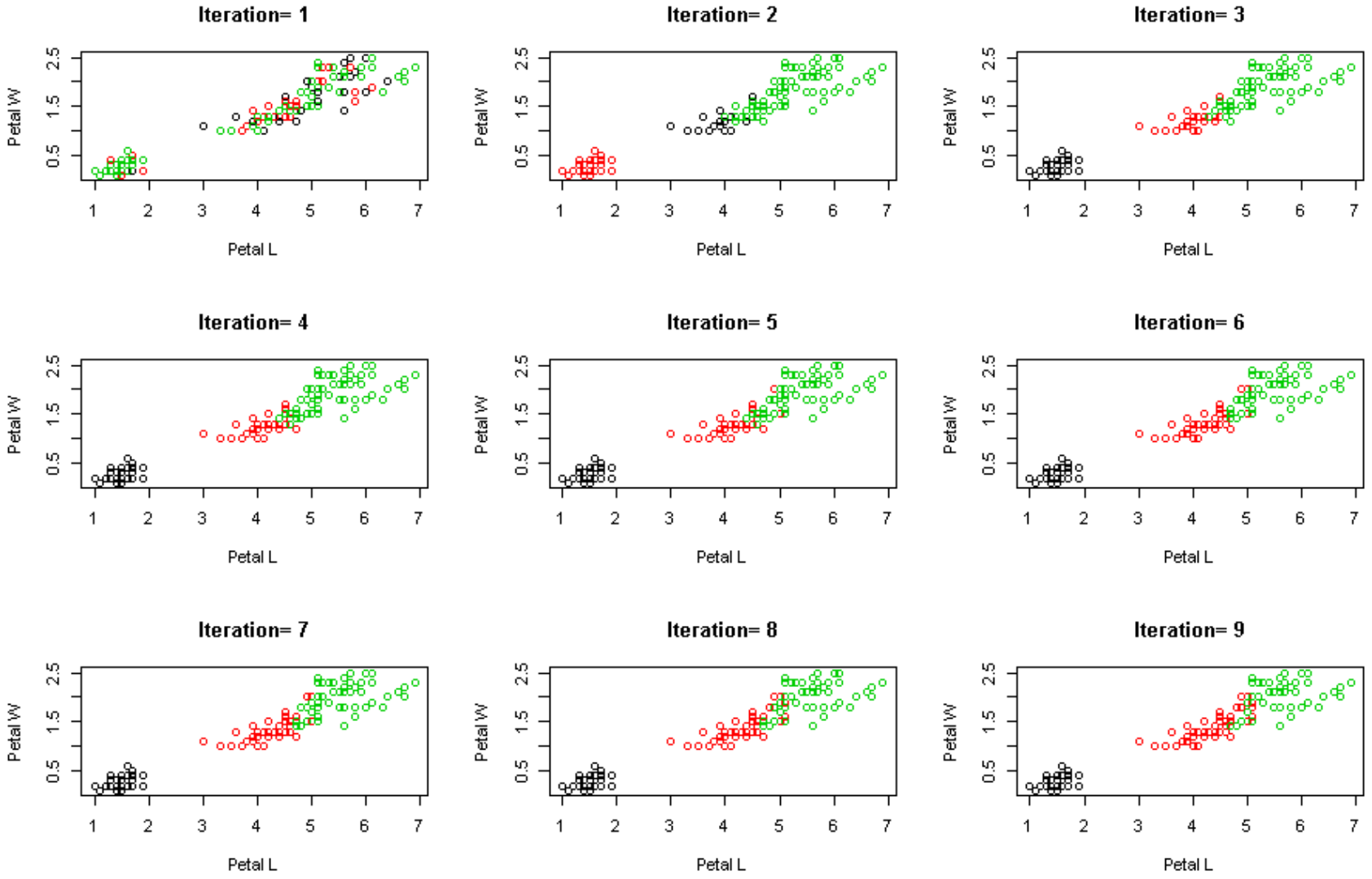
- Pretend that the iris types of the observations are unknown => cluster analysis
- As an example, and for illustration purpose, we will use petal length and width
- Choose *K*=3
- K-means

K-Mean Clustering: Iris (Petal L & W)

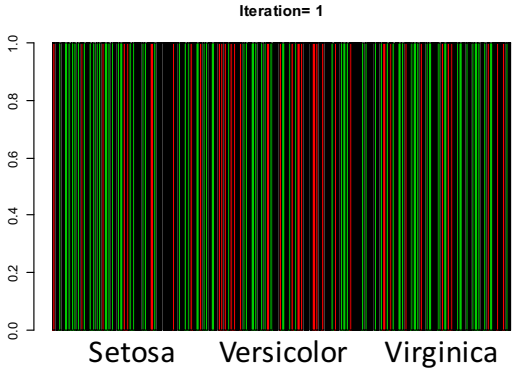


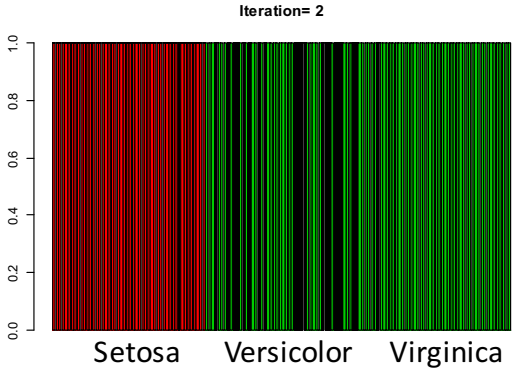
Note: the animation in the figure doesn't work appropriately on MAC.

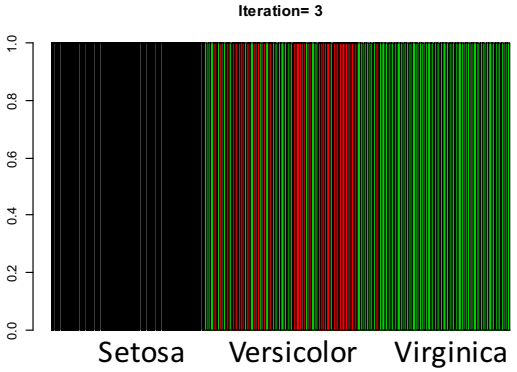
K-Mean Clustering: Iris (Petal L & W)

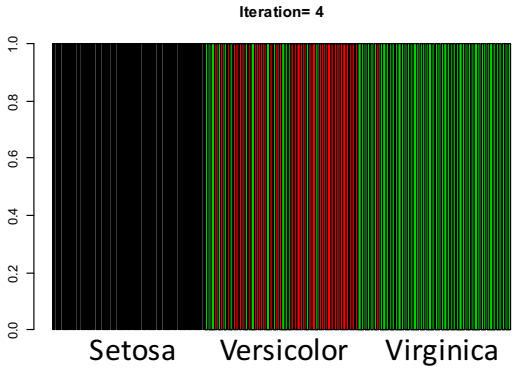


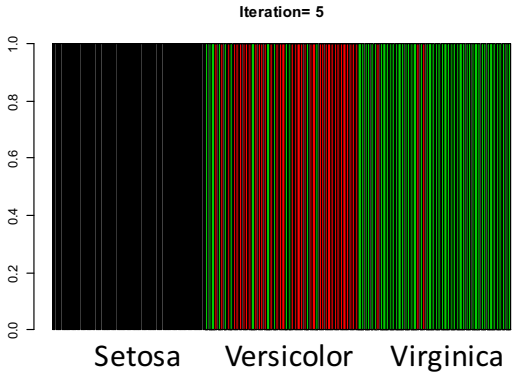
K-Mean Clustering: Iris (Petal L & W)

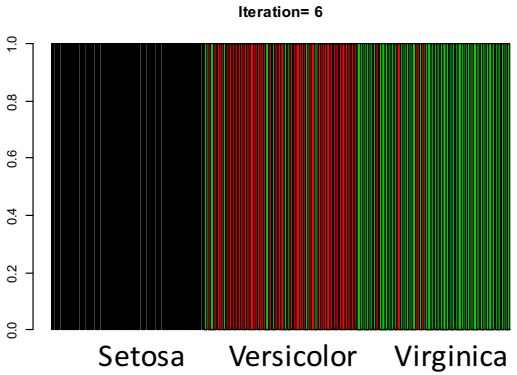


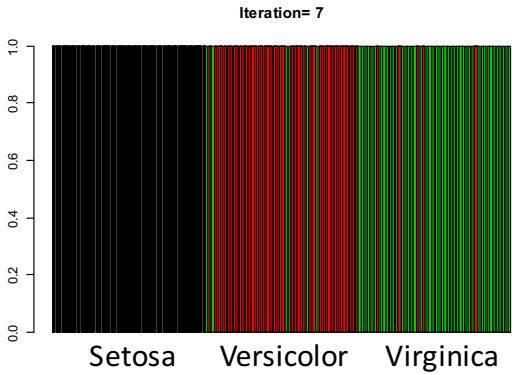


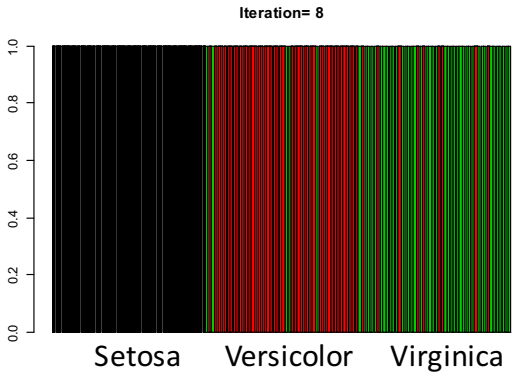


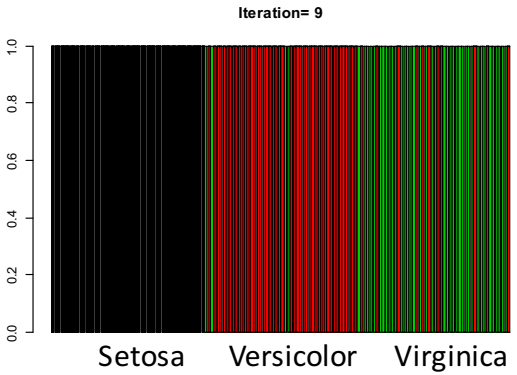




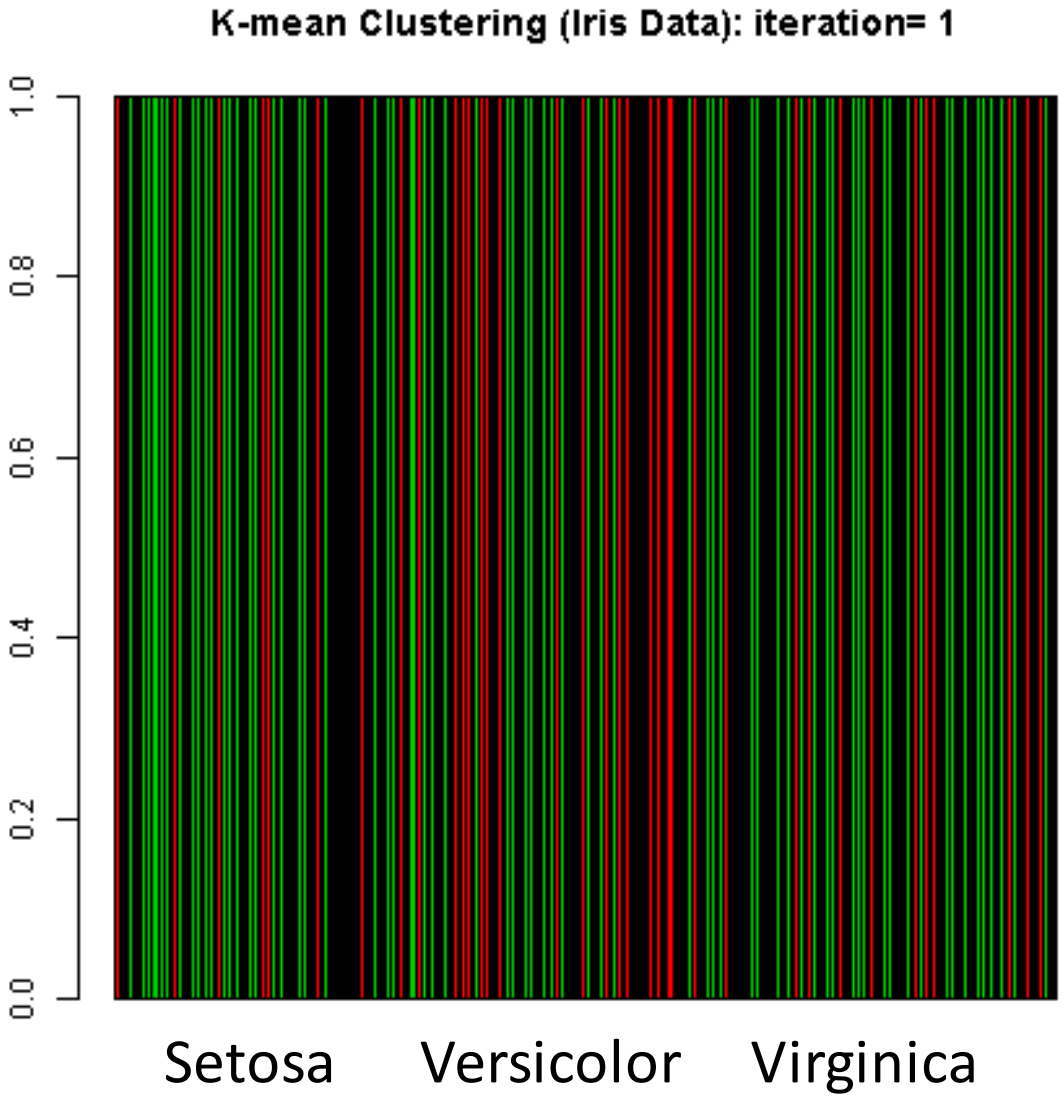








K-Mean Clustering: Iris (Petal L & W)

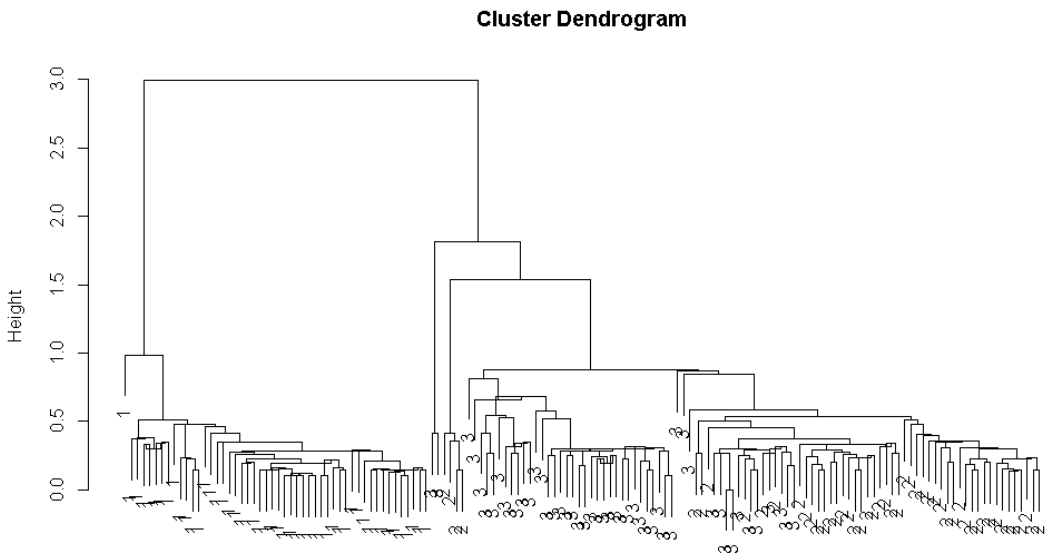


Note: the animation in the figure doesn't work appropriately on MAC.

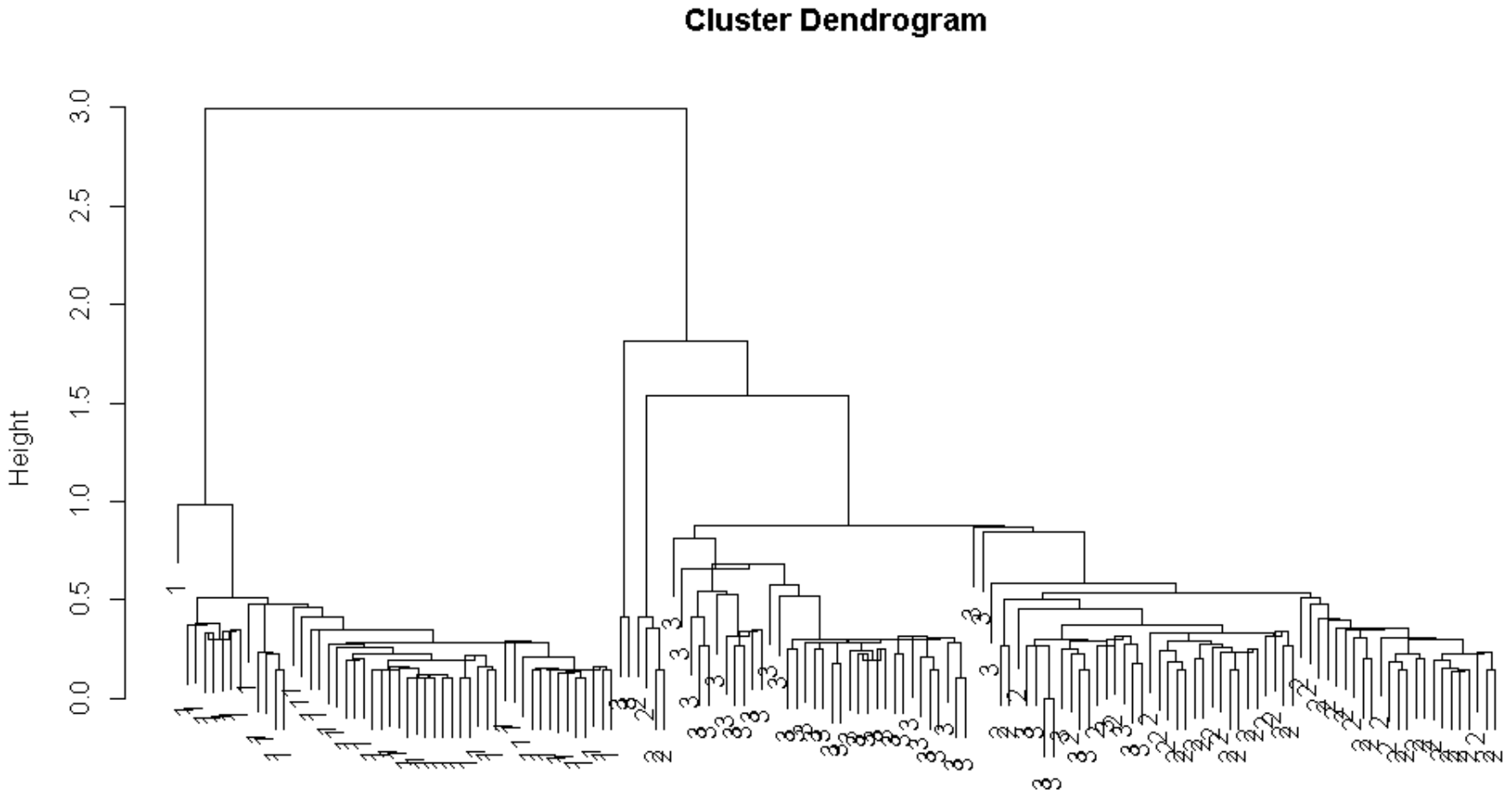
Model-Free Clustering: Hierarchical Clustering

Hierarchical Clustering

- The number of clusters is not required
- Gives a tree-based representation of observations - dendrogram



- Each leaf represents an observation
- Leaves similar with each other are fused to branches
- Leaves/branches similar with each other are fused to branches
- …

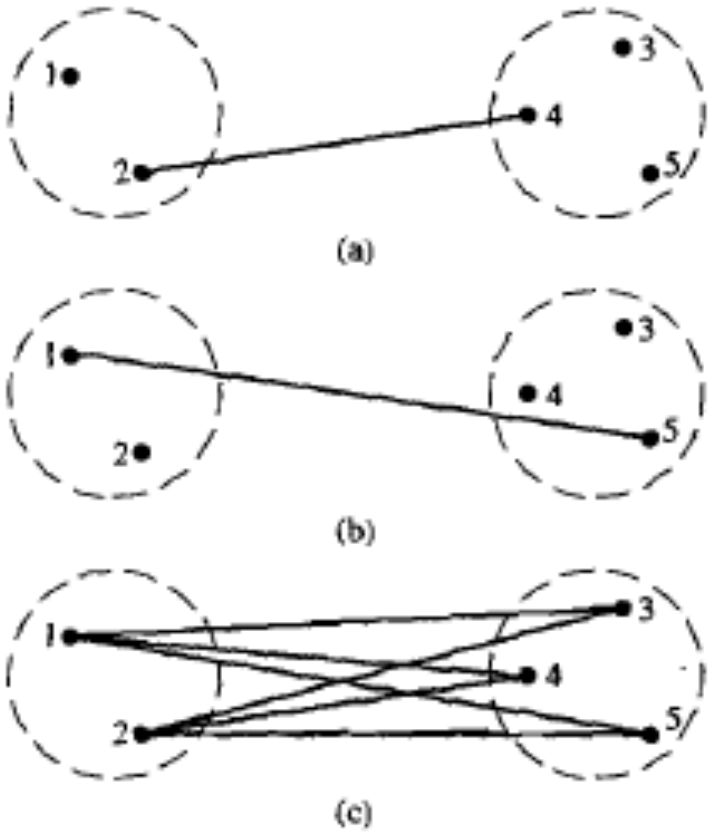


Setosa Virginica Versicolor

Hierarchical Clustering

- To grow a tree, we need to define dissimilarities (distances) between leaves/branches
  - Two leaves: easy. One can use a dissimilarity measure
  - A leaf and a branch: there are different options
  - Two branches: similar to “a leaf and a branch”, there are different options

Distance between Two Branches/Clusters



Cluster distance	
d_{24}	Single linkage
d_{15}	Complete linkage
$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$	Average linkage

Many other options!

Model-Based Clustering

Model-Based Clustering: Mixture Model

- Consider a random variable X.
- We say it follows a mixture of *K* distributions if its distribution can be represented using *K* distributions:

$$f_{Mix}(\mathbf{x}) = \sum_{k=1}^{K} p_k \, f_k(\mathbf{x})$$

- The weights p_k, k=1,…,K are nonnegative numbers and they add up to 1

Cluster Analysis Based on Mixture Model

- I present a frequentist version
  - Choose an appropriate model. E.g., A Gaussian mixture model with K=2 clusters
  - Write down the likelihood function
  - Find the maximum likelihood estimate of the parameters
  - Calculate the Pr(cluster k| observation x_i) for i=1,…,n, k=1,2

The Maximum Likelihood Estimate (MLE) of the Parameters

- An easy-to-implement algorithm to find the MLEs is called the Expectation and Maximization (EM) algorithm
- Initialize parameters
- E step: calculate "conditional" expectation.
  - "conditional" means conditional on current estimate of the parameters
  - This step involves calculating prob(cluster k|obs I, current estimate of para), k=1,…,K, i=1,…,n
  - This step is similar to the assignment step in an K-means algorithm

The Maximum Likelihood Estimate (MLE) of the Parameters

- The M step: find the set of values that maximize the conditional expectation calculated in the E step. This step updates the parameter values

- Repeat the E and M steps until convergence

EM vs K-Mean

EM	K-Mean
• Step 1: initialization	• Step 1: initialization
• E: Calculate conditional probabilities	• Step 2a: guess cluster membership
• M step: Find optimal values for parameters	• Step 2b: find cluster centers
• Repeat the E and M steps until convergence	• Repeat 2a-2b until convergence
• Allows clusters to have different shapes	

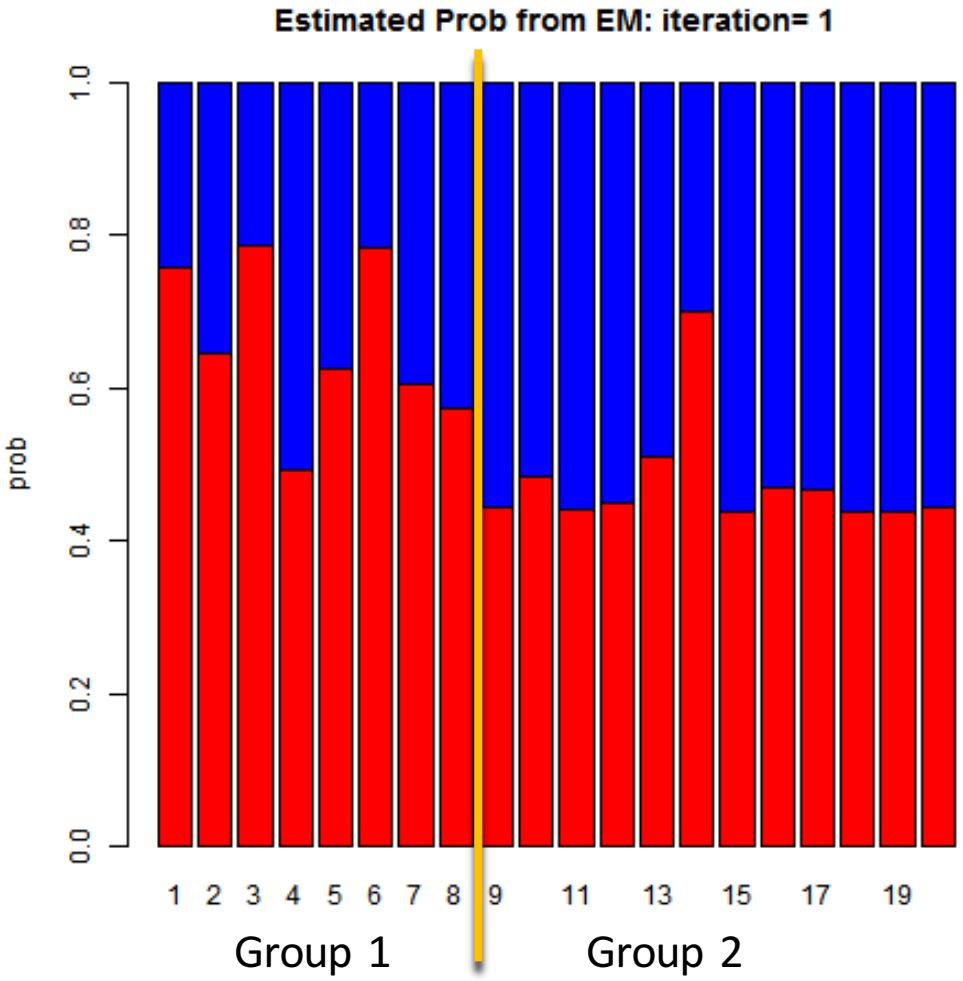
Example: Gaussian Mixture Model

- Observed data (simulated from two normal distributions)
  - 0.37 1.18 0.16 2.60 1.33 0.18 1.49 1.74 3.58 2.69
    4.51 3.39 2.38 0.79 4.12 2.96 2.98 3.94 3.82 3.59

- Assuming K=2

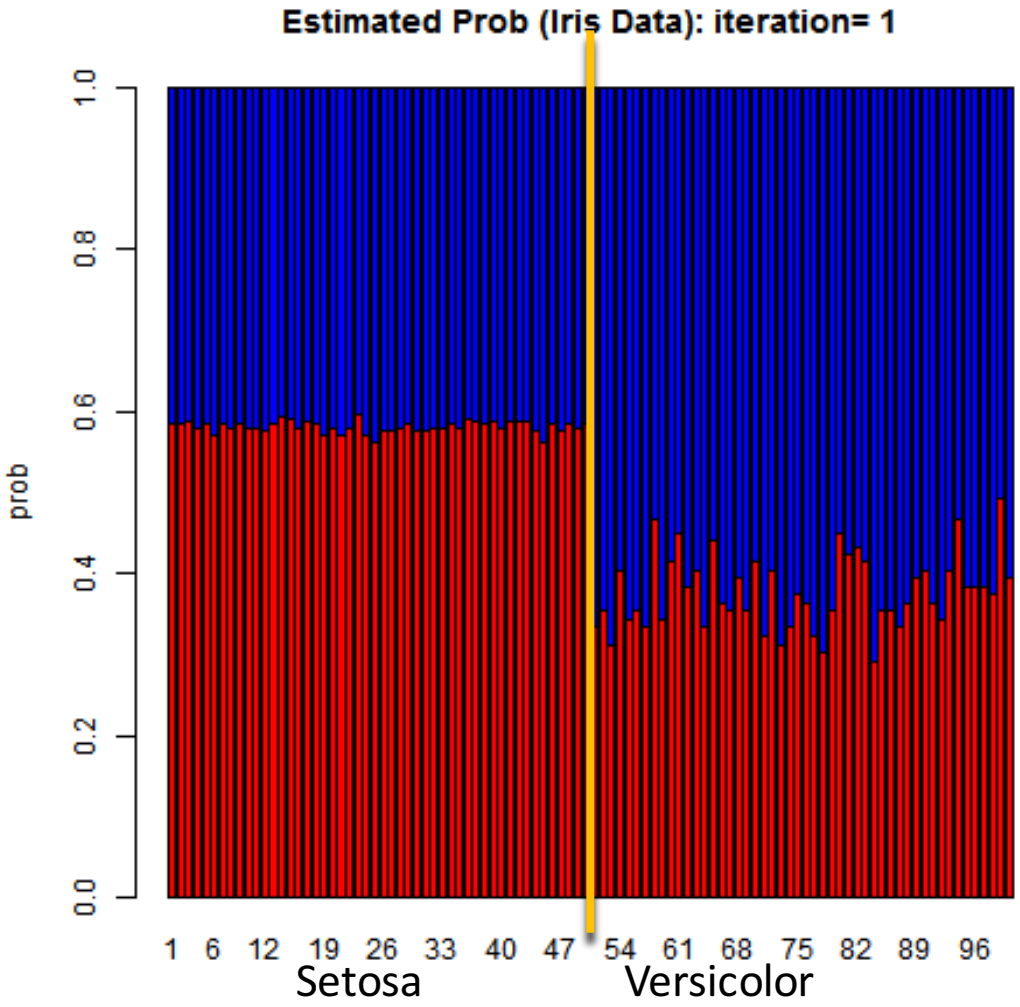
- Parameters: μ_1, μ_0, σ_1, σ_0, p

Example: simulated data



Note: the animation in the figure doesn't work appropriately on MAC.

Example: Cluster Analysis of Iris Data Using Petal Length



Note: the animation in the figure doesn't work appropriately on MAC.

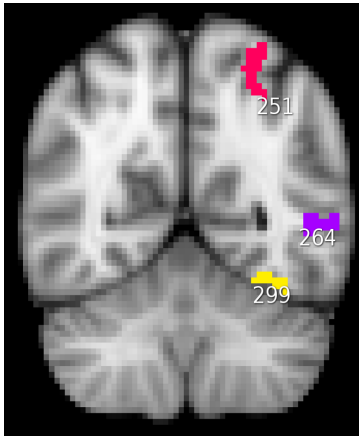
R Package: MCLUST

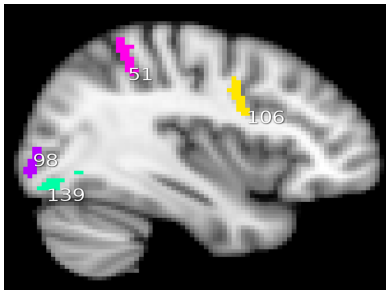
- Developed by Adrian Raftery and colleagues
- Gaussian mixture model
- EM
- Clustering, classification, density estimation
- Please try it out!

Clustering Analysis For Multidimensional Data

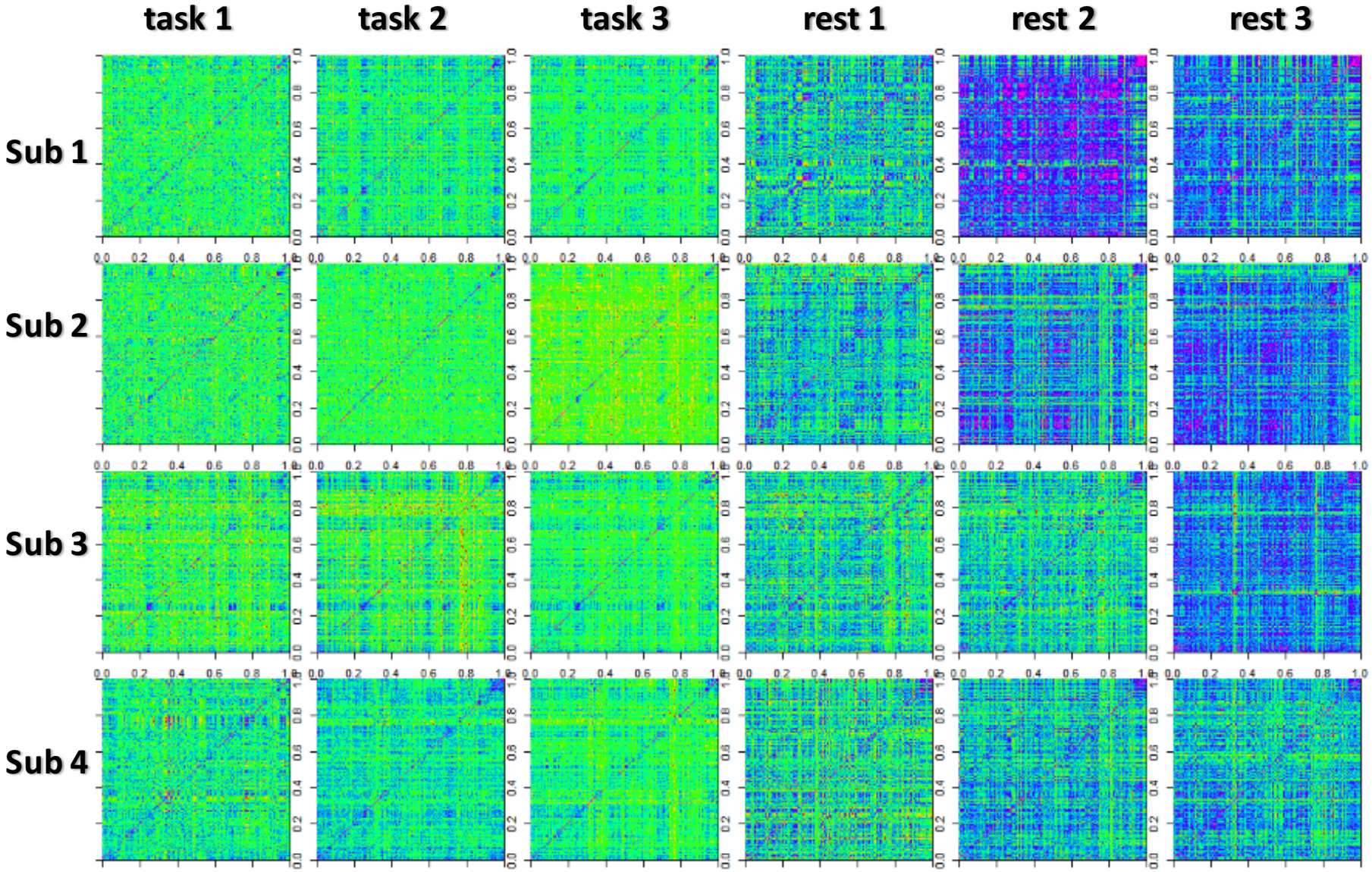
Multidimensional Data

- Human faces, images
- 3D objects
- Text documents
- Brain imaging

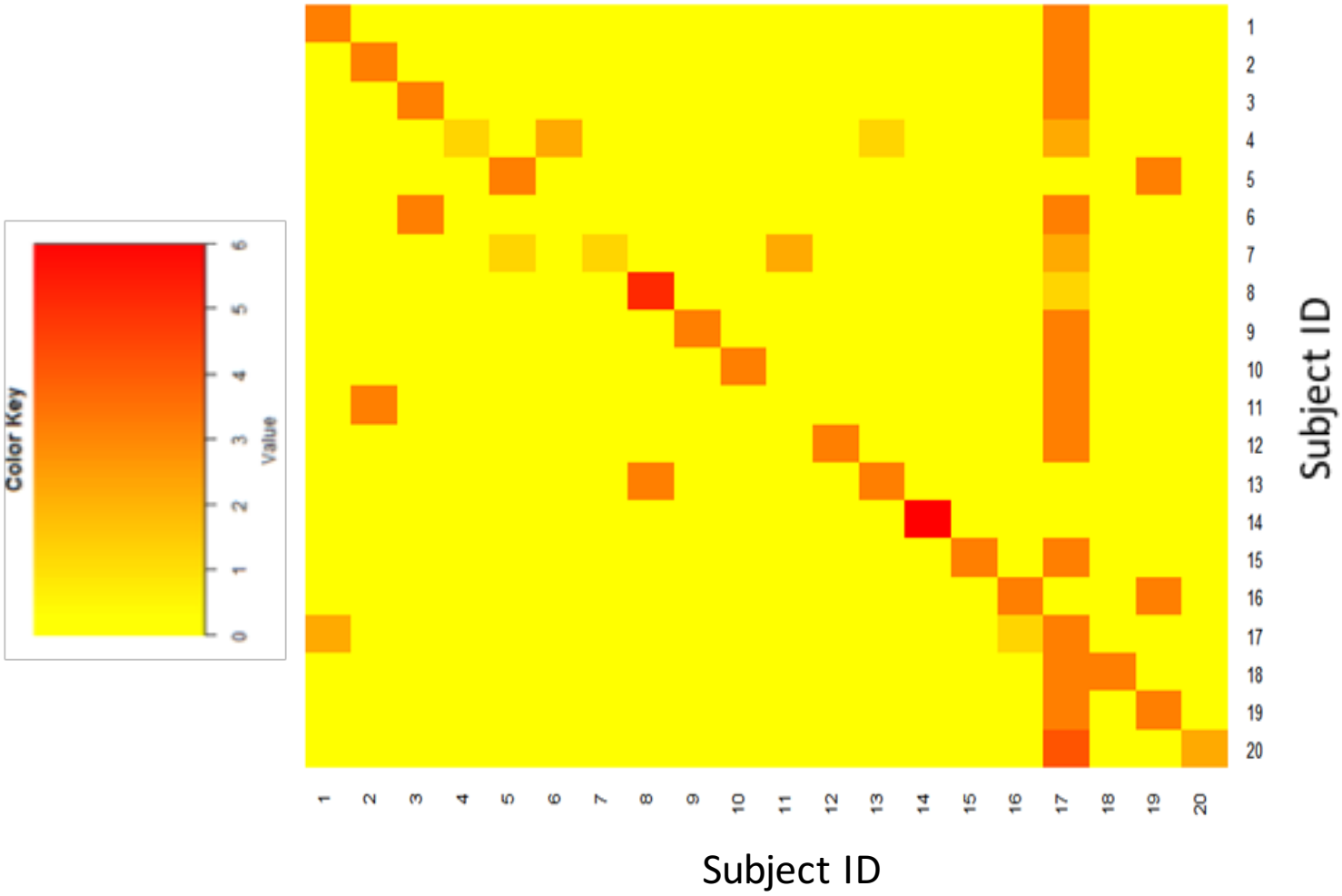


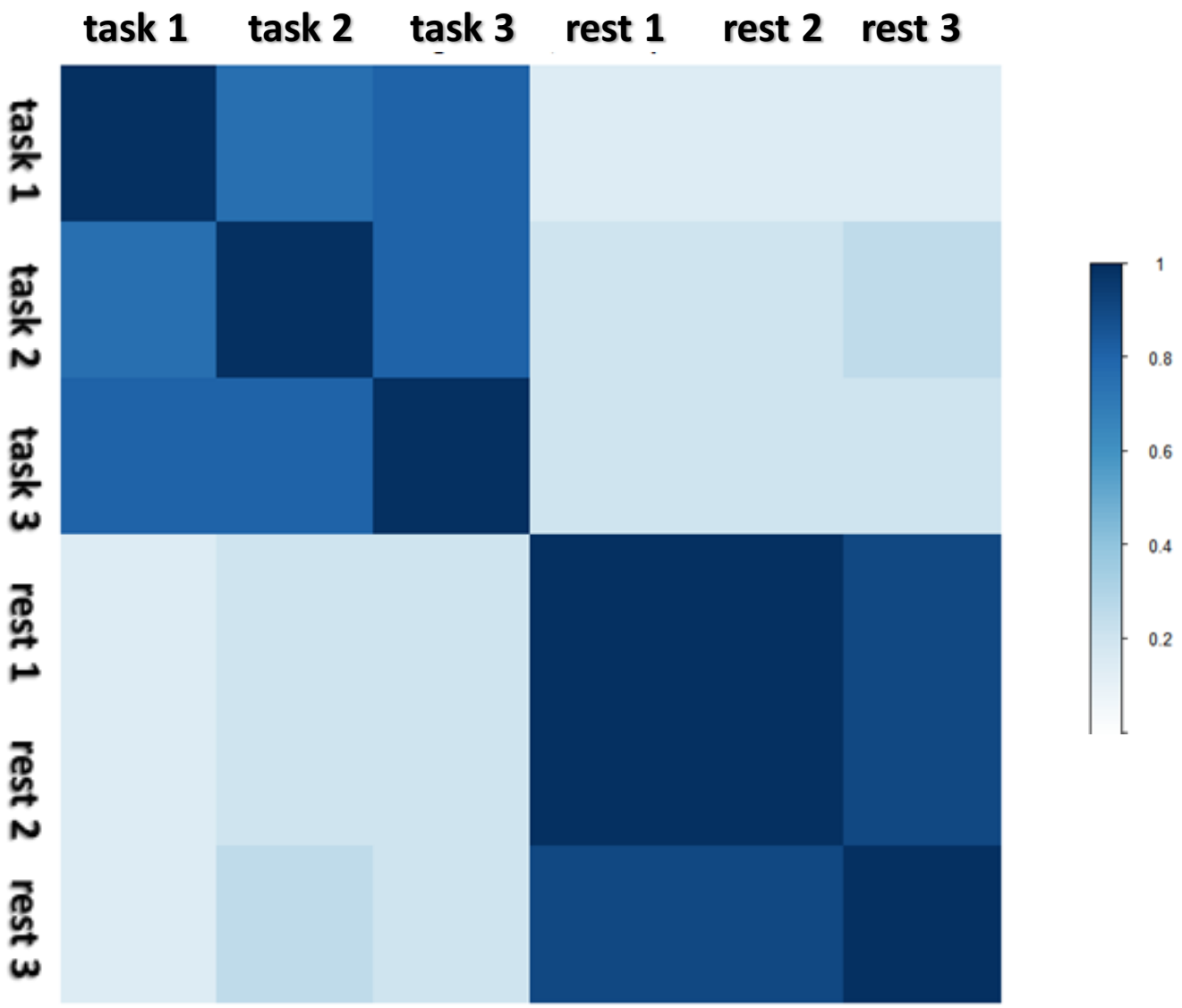


Whole Brain Connectivity



Brain Connectivity vs Fingerprint




task 1
task 2
task 3
rest 1
rest 2
rest 3
task 1
task 2
task 3
rest 1
rest 2
rest 3
1
0.8
0.6
0.4
0.2

Some Technical Details