Clustering

..

Supervised v/s unsupervised learning

Supervised Learning

$$Y = f(x_1, x_2, x_3, x_4, \cdots)$$

unsupervised Learning

$$\cancel{Y} = f(\underset{\displaystyle\downarrow\ (x_1, x_2, x_3, x_4, \cdots)}{x_1, x_2, x_3, x_4, \cdots})$$

Unsupervised Learning

- Since there is no **Y** (response variable)
  - There is no possibility of **prediction**
- Goal in unsupervised learning
  - Discover aspects about the data and the variables
  - Mathematically **visualize** data
  - Are there sub-groups among the variables
- Type of unsupervised learning methods
  - PCA (Principal Component Analysis)
  - Clustering

Challenges in unsupervised learning

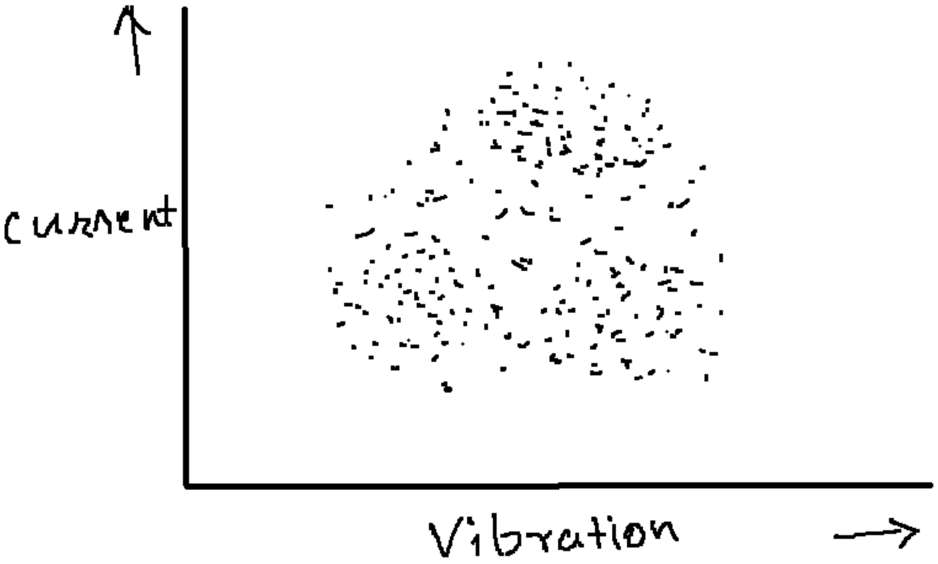
- It is a subjective process
- There is no criteria to objectively evaluate the results
  - No cross-validation methods
  - No possibility of validating results on independent data
- Consequently – unsupervised learning is performed as part of:
  - Exploratory Data Analysis
  - In order to discover **structure** in the data

Examples of unsupervised learning

Measure the following

- Vibration
- Current

During metal cutting



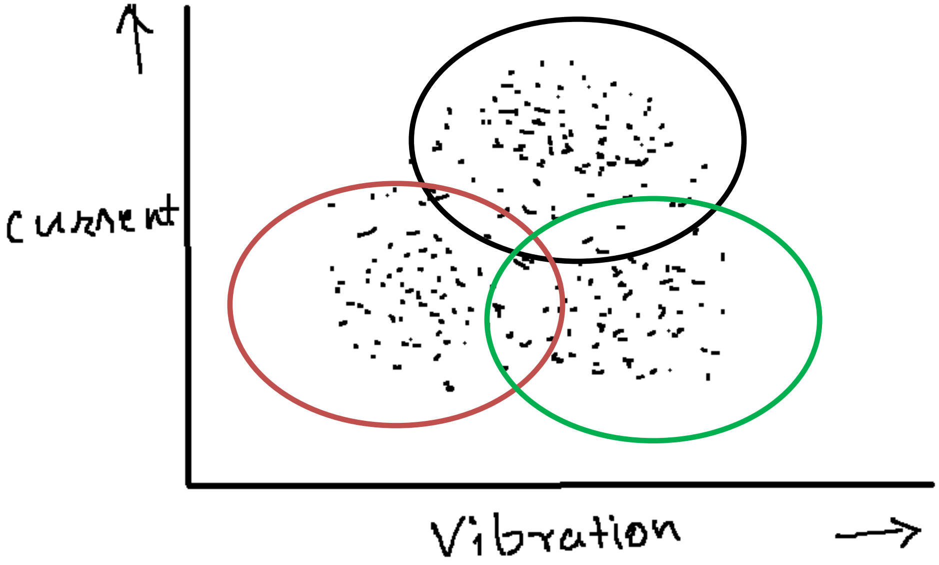
Use of unsupervised learning

Measure the following

- Vibration
- Current

During metal cutting

Is it possible to "discover" inherent groups ?



The groups observed possibly indicate a behavioural pattern of the machine tool + cutting tool system

Clustering

- Clustering: Unsupervised learning method
- Used to “discover” inherent grouping of data points
- Popular clustering methods
  - K-means clustering
    - Partition the observations into a pre-specified number of clusters
  - Hierarchical clustering
    - Number of possible clusters is an outcome of the clustering process

K-means Clustering

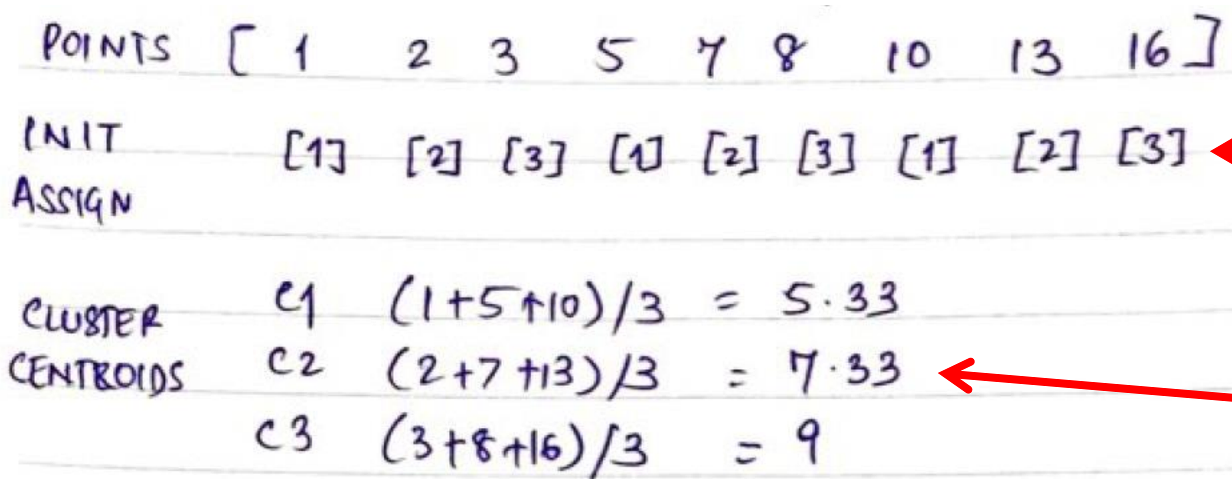
- Requires us to decide and specify, beforehand, the **desired number** of clusters
  - How do we decide?
  - Typically based on an understanding of the domain
- The method then results in assigning every point to one of these clusters
- Note:
  - Every data point will end up belonging to one and only one cluster

K-means Clustering

- Approach
  - Specify the number of desired clusters
  - Algorithm begins by assigning input data points randomly to these clusters
  - The algorithm then iterates:
    - Derives the centroid of every cluster
    - Finds distances of all input points from the centroids of all the clusters
    - Reassigns every data point to the "nearest" cluster
    - This process is continued until there are no more reassignments

K-means Clustering in Action

- Assume the following :
  - Input point set: [1 2 3 5 7 8 10 13 16]
    - Points along X axis …
  - Desired number of clusters : 3
- Initial cluster assignment … random



POINTS	[1	2	3	5	7	8	10	13	16]
INIT ASSIGN	[1]	[2]	[3]	[1]	[2]	[3]	[1]	[2]	[3]

CLUSTER CENTROIDS

c1 $(1+5+10)/3 = 5.33$

c2 $(2+7+13)/3 = 7.33$

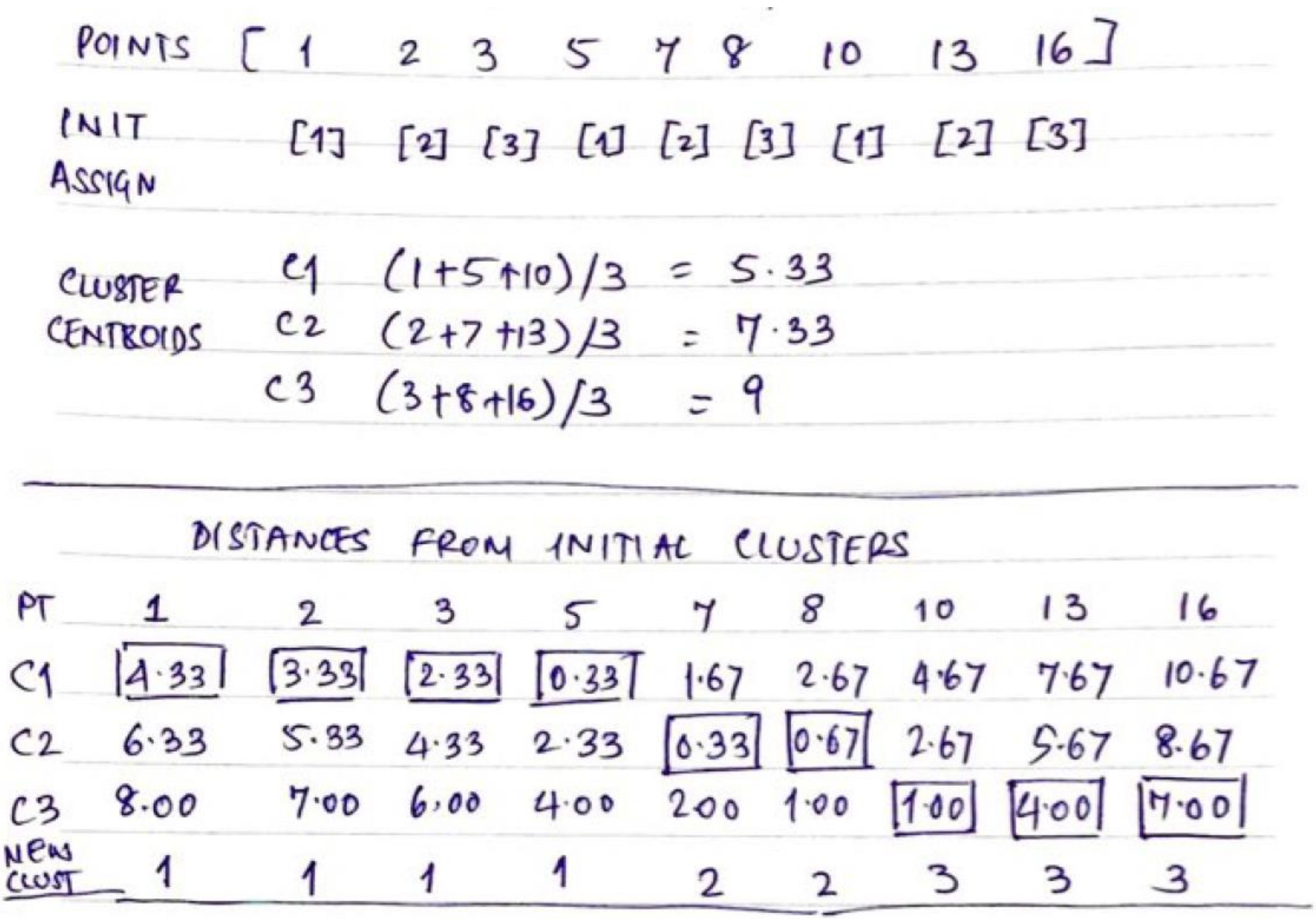
c3 $(3+8+16)/3 = 9$

Initial random assignment of points to required number of clusters

Derive the centroid of every cluster

K-means Clustering in Action

- Distance calculation & cluster re-assignment



POINTS	[1	2	3	5	7	8	10	13	16]
INIT ASSIGN	[1]	[2]	[3]	[1]	[2]	[3]	[1]	[2]	[3]

CLUSTER CENTROIDS

c1 $(1+5+10)/3 = 5.33$

c2 $(2+7+13)/3 = 7.33$

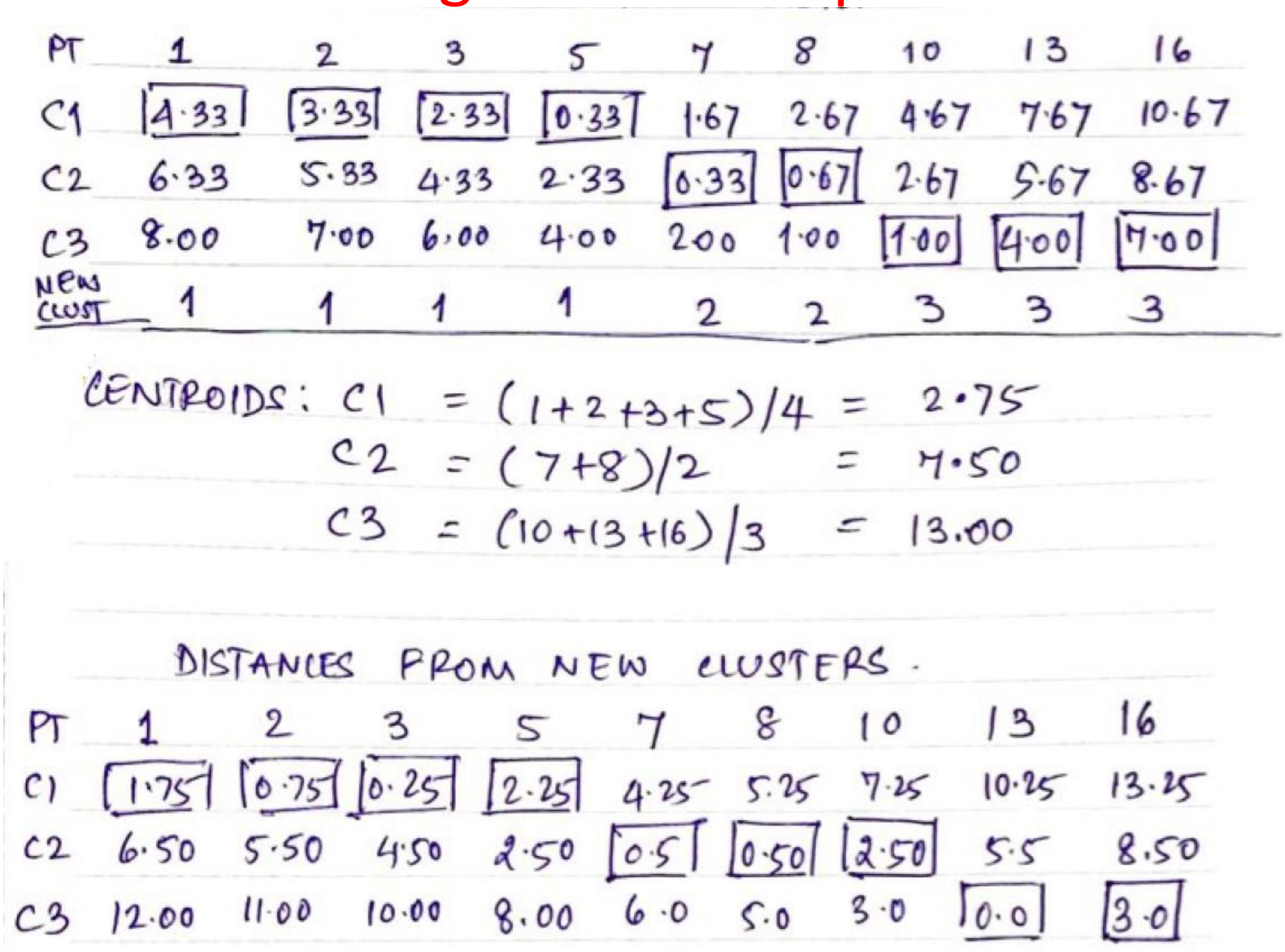
c3 $(3+8+16)/3 = 9$

DISTANCES FROM INITIAL CLUSTERS

PT	1	2	3	5	7	8	10	13	16
C1	$\boxed{4.33}$	$\boxed{3.33}$	$\boxed{2.33}$	$\boxed{0.33}$	1.67	2.67	4.67	7.67	10.67
C2	6.33	5.33	4.33	2.33	$\boxed{0.33}$	$\boxed{0.67}$	2.67	5.67	8.67
C3	8.00	7.00	6.00	4.00	200	1.00	$\boxed{1.00}$	$\boxed{4.00}$	$\boxed{7.00}$
NEW CLUST	1	1	1	1	2	2	3	3	3

K-means Clustering in Action

- Cluster re-assignment … step 2



PT	1	2	3	5	7	8	10	13	16
C1	$\boxed{4.33}$	$\boxed{3.33}$	$\boxed{2.33}$	$\boxed{0.33}$	1.67	2.67	4.67	7.67	10.67
C2	6.33	5.33	4.33	2.33	$\boxed{0.33}$	$\boxed{0.67}$	2.67	5.67	8.67
C3	8.00	7.00	6.00	4.00	200	1.00	$\boxed{1.00}$	$\boxed{4.00}$	$\boxed{7.00}$
NEW CLUST	1	1	1	1	2	2	3	3	3

CENTROIDS: C1 = (1+2+3+5)/4 = 2.75

C2 = (7+8)/2 = 7.50

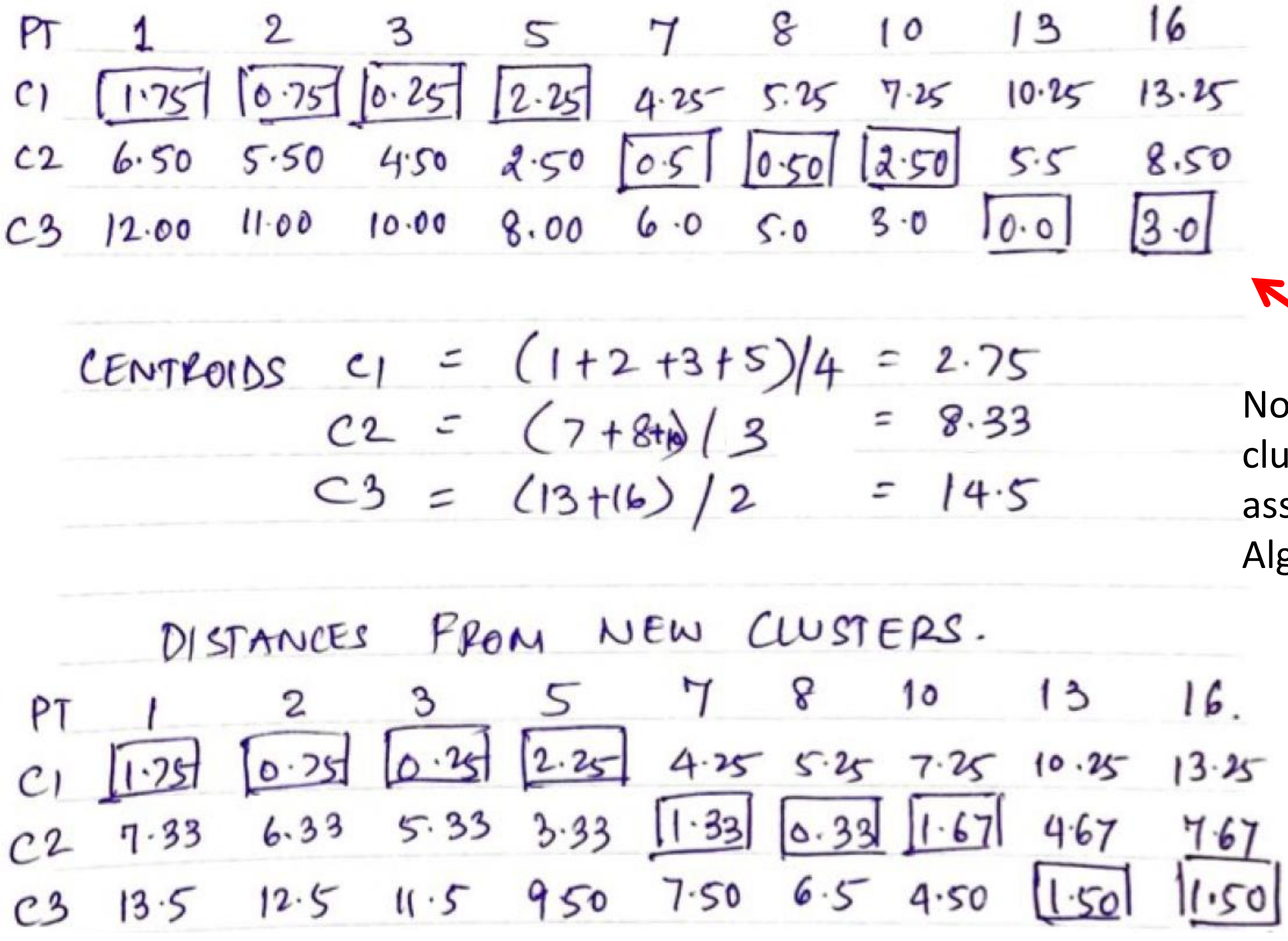
C3 = (10+13+16)/3 = 13.00

DISTANCES FROM NEW CLUSTERS.

PT	1	2	3	5	7	8	10	13	16
C1	$\boxed{1.75}$	$\boxed{0.75}$	$\boxed{0.25}$	$\boxed{2.25}$	4.25	5.25	7.25	10.25	13.25
C2	6.50	5.50	4.50	2.50	$\boxed{0.5}$	$\boxed{0.50}$	$\boxed{2.50}$	5.5	8.50
C3	12.00	11.00	10.00	8.00	6.0	5.0	3.0	$\boxed{0.0}$	$\boxed{3.0}$

K-means Clustering in Action

- Cluster re-assignment … Step 3



PT	1	2	3	5	7	8	10	13	16
C1	$\boxed{1.75}$	$\boxed{0.75}$	$\boxed{0.25}$	$\boxed{2.25}$	4.25	5.25	7.25	10.25	13.25
C2	6.50	5.50	4.50	2.50	$\boxed{0.5}$	$\boxed{0.50}$	$\boxed{2.50}$	5.5	8.50
C3	12.00	11.00	10.00	8.00	6.0	5.0	3.0	$\boxed{0.0}$	$\boxed{3.0}$

CENTROIDS C1 = (1 + 2 + 3 + 5)/4 = 2.75

C2 = (7 + 8 + 10) / 3 = 8.33

C3 = (13 + 16) / 2 = 14.5

DISTANCES FROM NEW CLUSTERS.

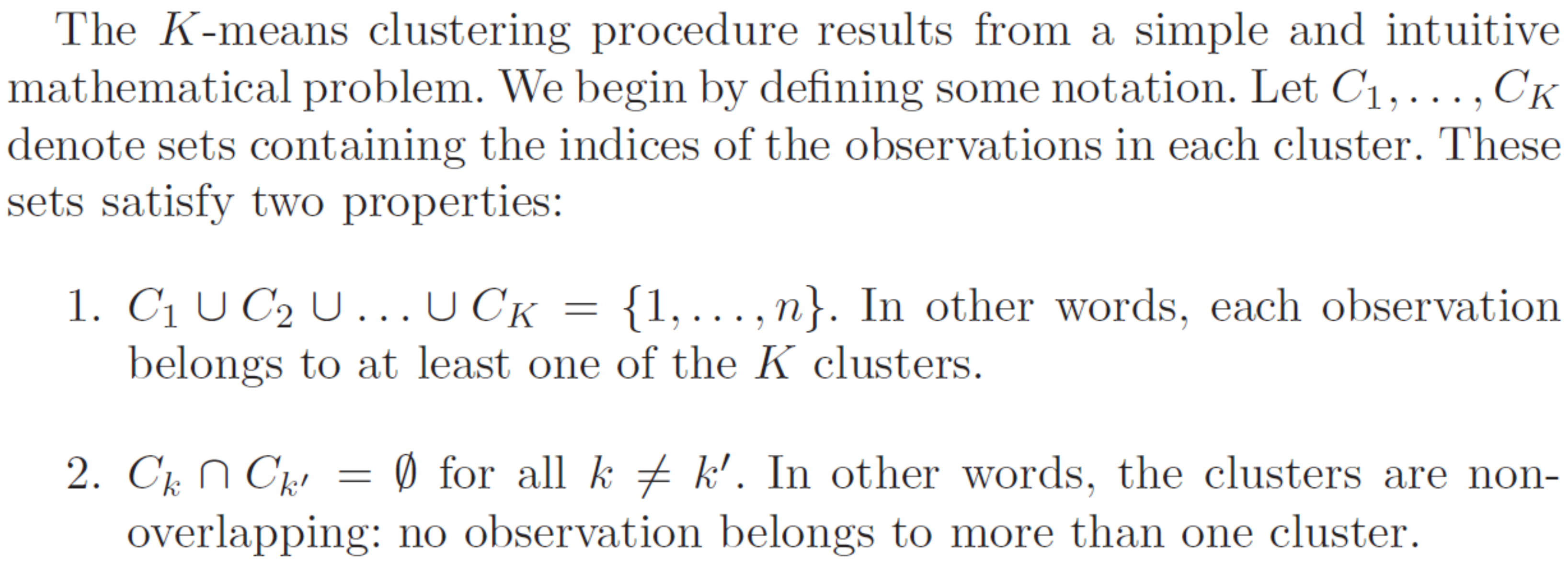
PT	1	2	3	5	7	8	10	13	16.
C1	$\boxed{1.75}$	$\boxed{0.75}$	$\boxed{0.25}$	$\boxed{2.25}$	4.25	5.25	7.25	10.25	13.25
C2	7.33	6.33	5.33	3.33	$\boxed{1.33}$	$\boxed{0.33}$	$\boxed{1.67}$	4.67	7.67
C3	13.5	12.5	11.5	9.50	7.50	6.5	4.50	$\boxed{1.50}$	$\boxed{1.50}$

No change in clusters assignments. Algorithm stops.

K-means Clustering: Final Clusters

- Input points
  - [1 2 3 5 7 8 10 13 16]
- Clusters
  - [1 2 3 5]   [7 8 10] [13 16]
- Note:
  - In K-means clustering, the initial cluster assignments are random
  - Therefore, the resulting clusters could represent only a local minima
  - **Given a data set, it is therefore important to run K-means clustering algorithm multiple times – each starting from a distinct initial random cluster assignment**

Formally …



The K-means clustering procedure results from a simple and intuitive mathematical problem. We begin by defining some notation. Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

Good clustering is one for which the **within cluster** variation is as less as possible

Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Formally …

Using Euclidean distance, the total within cluster variation can be expressed as:

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

$|C_k|$ denotes the number of observations in the kth cluster

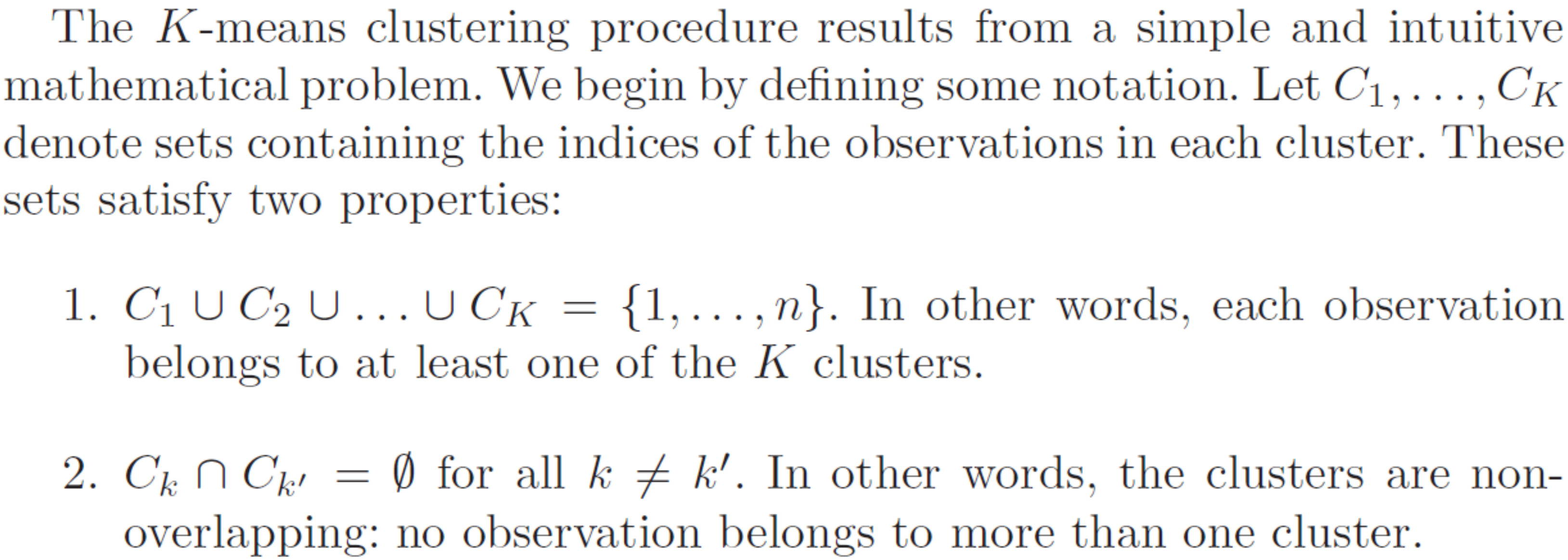
The clustering problem therefore reduces to:

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

The k-means algorithm is based on this problem formulation $\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$

Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Formally …



The K-means clustering procedure results from a simple and intuitive mathematical problem. We begin by defining some notation. Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Hierarchical Clustering

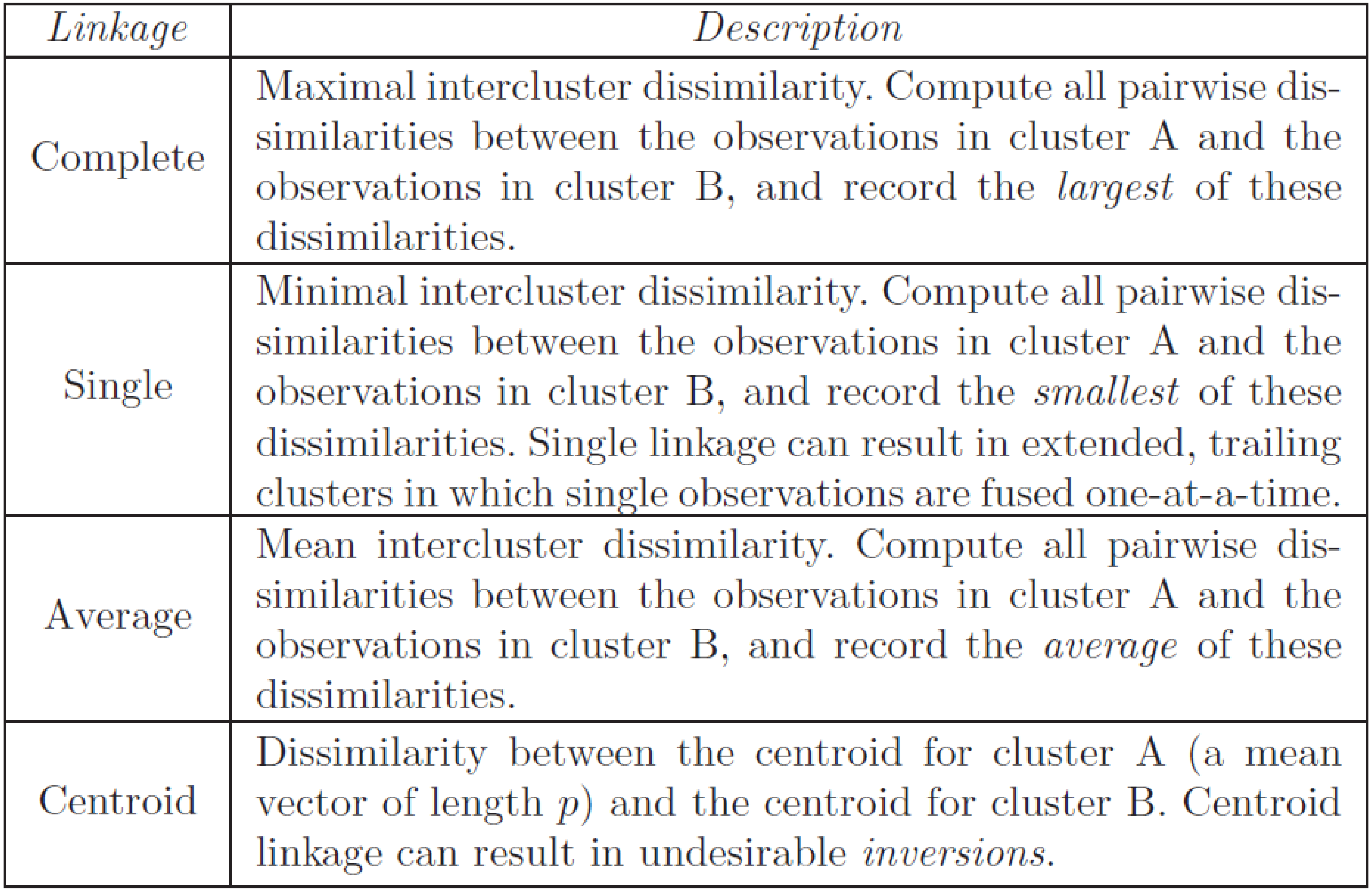
- Hierarchical clustering does not require specify the number of clusters
- It results in a hierarchical, tree-like, representation of the observations
  - Known as a ***Dendrogram***
  - *(Based on **bottom-up** or **agglomerative** clustering)*

Hierarchical Clustering: Method

- Given a point set, the Hierarchical Clustering algorithm proceeds as follows:
  - Every data point is assigned to its own unique cluster
  - The algorithm then iterates as follows:
    - A distance metric (Euclidean – by default) is used to find out pair-wise distance between all the clusters (**see next slide**)
    - A pair of clusters are found such that the distance between them is the smallest of all pair-wise distances
    - The two clusters are merged
    - The process is repeated till only one cluster remains
  - The above process results in a Dendrogram which, when cut at a particular level, results in a specific number of clusters
    - The required number of clusters can thus be obtained

Hierarchical Clustering: Distance between Clusters

- Distance between clusters is evaluated using one of the following "linkages": of these, **Average and Complete are used the most**



Linkage	*Description*
Complete	Maximal intercluster 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 intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster 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. Centroid linkage can result in undesirable *inversions*.

Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Hierarchical Clustering : Example

- Input point set: [1 2 3 5 7 8 10 13 16]
- Initial cluster: [1] [2] [3] [5] [7] [8] [10] [13] [16]
- Inter-cluster distance: Using “Complete Linkage”

	1	2	3	5	7	8	10	13	16
1	-	1	2	4	6	7	9	12	15
2	-	-	1	3	5	6	8	11	14
3	-	-	-	2	3	5	7	10	13
5	-	-	-	-	2	3	5	8	11
7	-	-	-	-	-	1	3	6	9
8	-	-	-	-	-	-	2	5	8
10	-	-	-	-	-	-	-	3	6
13	-	-	-	-	-	-	-	-	3
16	-	-	-	-	-	-	-	-	-

Hierarchical Clustering: Example

- Step 1

	1	2	3	5	7	8	10	13	16
1	-	**1**	2	4	6	7	9	12	15
2	-	-	1	3	5	6	8	11	14
3	-	-	-	2	3	5	7	10	13
5	-	-	-	-	2	3	5	8	11
7	-	-	-	-	-	1	3	6	9
8	-	-	-	-	-	-	2	5	8
10	-	-	-	-	-	-	-	3	6
13	-	-	-	-	-	-	-	-	3
16	-	-	-	-	-	-	-	-	-

Hierarchical Clustering: Example

- Step 2

	[1 2]	3	5	7	8	10	13	16
[1 2]	-	2	4	6	7	9	12	15
3	-	-	2	4	5	7	10	13
5	-	-	-	2	3	5	8	11
7	-	-	-	-	1	3	6	9
8	-	-	-	-	-	2	5	8
10	-	-	-	-	-	-	3	6
13	-	-	-	-	-	-	-	3
16	-	-	-	-	-	-	-	-

Hierarchical Clustering

- Step 3

	[1 2]	3	5	[7 8]	10	13	16
[1 2]	-	2	4	7	9	12	15
3	-	-	2	5	7	10	13
5	-	-	-	3	5	8	11
[7 8]	-	-	-	-	3	6	9
10	-	-	-	-	-	3	6
13	-	-	-	-	-	-	3
16	-	-	-	-	-	-	-

Hierarchical Clustering

- Step 3

	[1 2 3]	5	[7 8]	10	13	16
[1 2 3]	-	4	7	9	12	15
5	-	-	3	5	8	11
[7 8]	-	-	-	3	6	9
10	-	-	-	-	3	6
13	-	-	-	-	-	3
16	-	-	-	-	-	-

Hierarchical Clustering

- Step 4

	[1 2 3]	[5 7 8]	10	13	16
[1 2 3]	-	7	9	12	15
[5 7 8]	-	-	5	8	11
10	-	-	-	3	6
13	-	-	-	-	3
16	-	-	-	-	-

- Step 5

	[1 2 3]	[5 7 8]	[10 13]	16
[1 2 3]	-	7	12	15
[5 7 8]	-	-	8	11
[10 13]	-	-	-	6
16	-	-	-	-

Hierarchical Clustering

- Step 6

	[1 2 3]	[5 7 8]	[10 13]	16
[1 2 3]	-	7	12	15
[5 7 8]	-	-	8	11
[10 13]	-	-	-	6
16	-	-	-	-

- Step 7

	[1 2 3]	[5 7 8]	[10 13 16]
[1 2 3]	-	7	15
[5 7 8]	-	-	11
[10 13 16]	-	-	-

Hierarchical Clustering

- Step 8

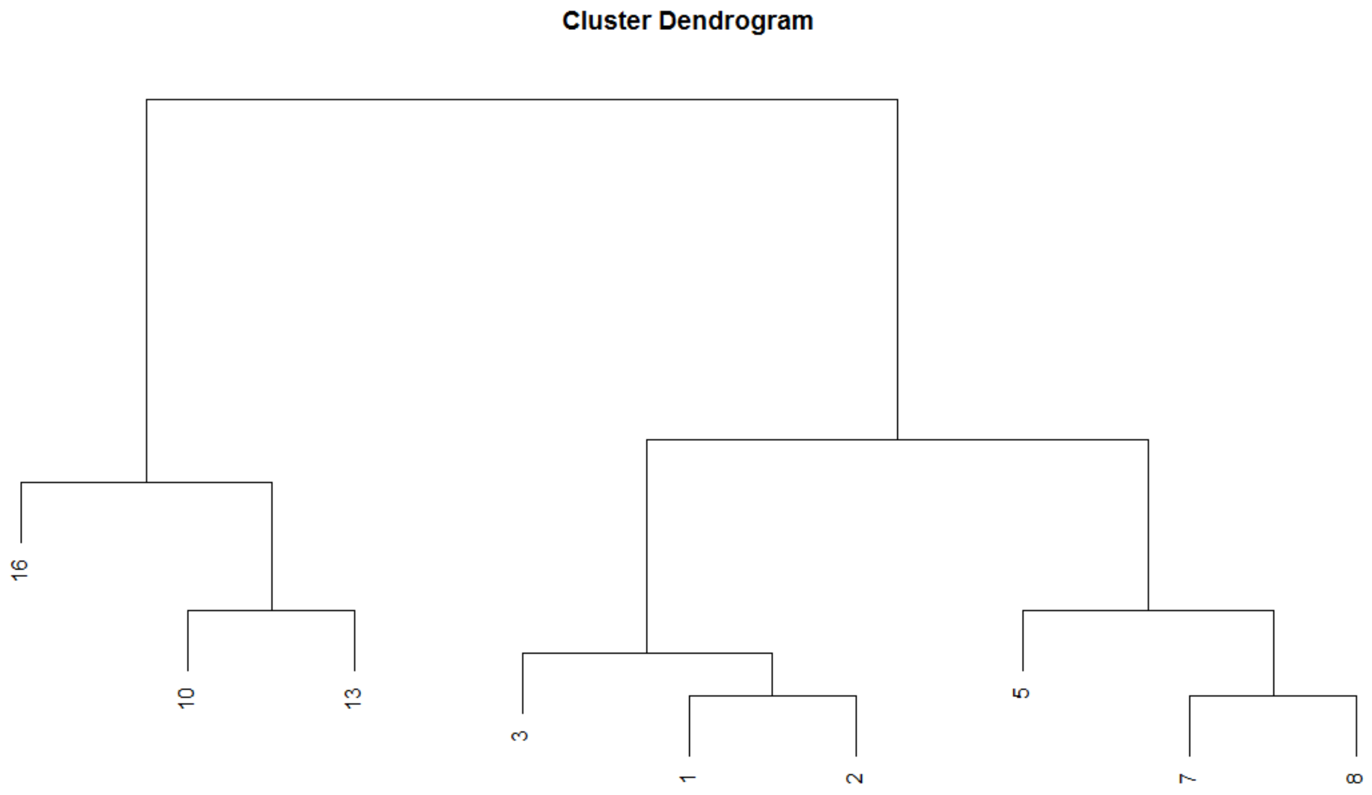
	[1 2 3 5 7 8]	[10 13 16]
[1 2 3 5 7 8]	-	15
[10 13 16]	-	-

- Step 9: Final

	[1 2 3 5 7 8 10 13 16]
[1 2 3 5 7 8 10 13 16]	-

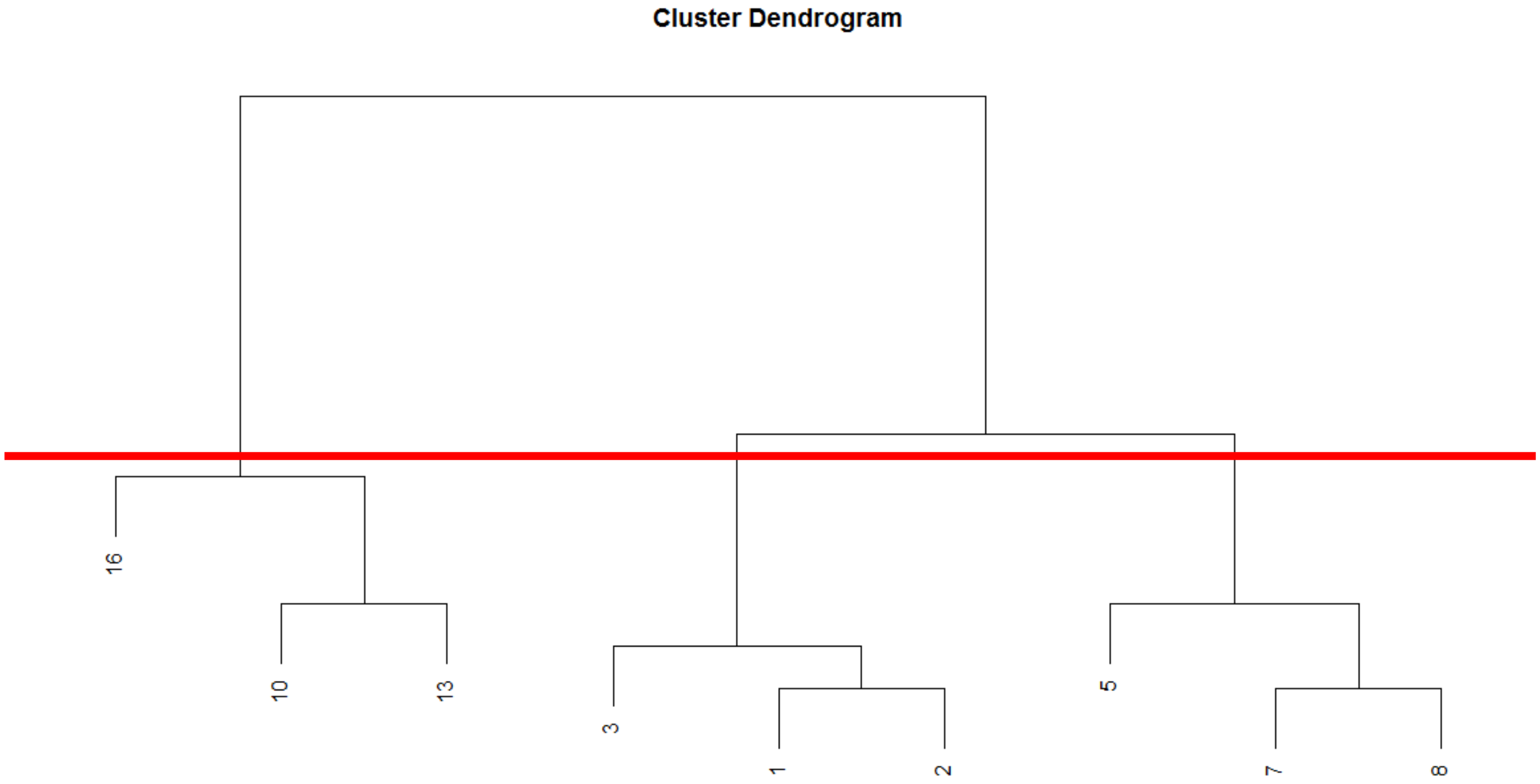
Hierarchical Clustering

- Resulting Dendrogram



Hierarchical Clustering

- Cutting the dendrogram to get clusters

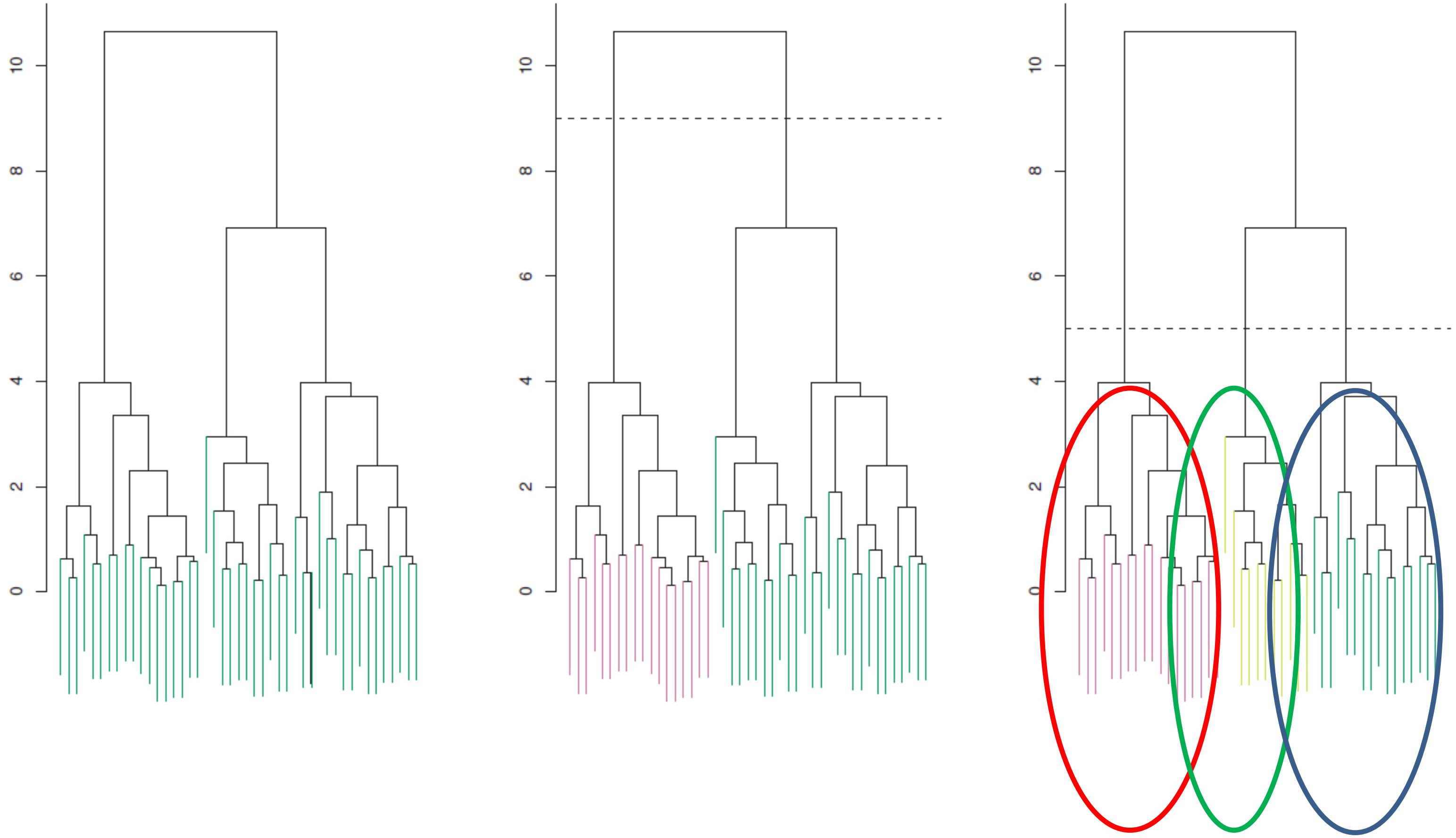


- Resulting clusters: [1 2 3] [5 7 8] [10 13 16]

Interpreting the Dendrogram

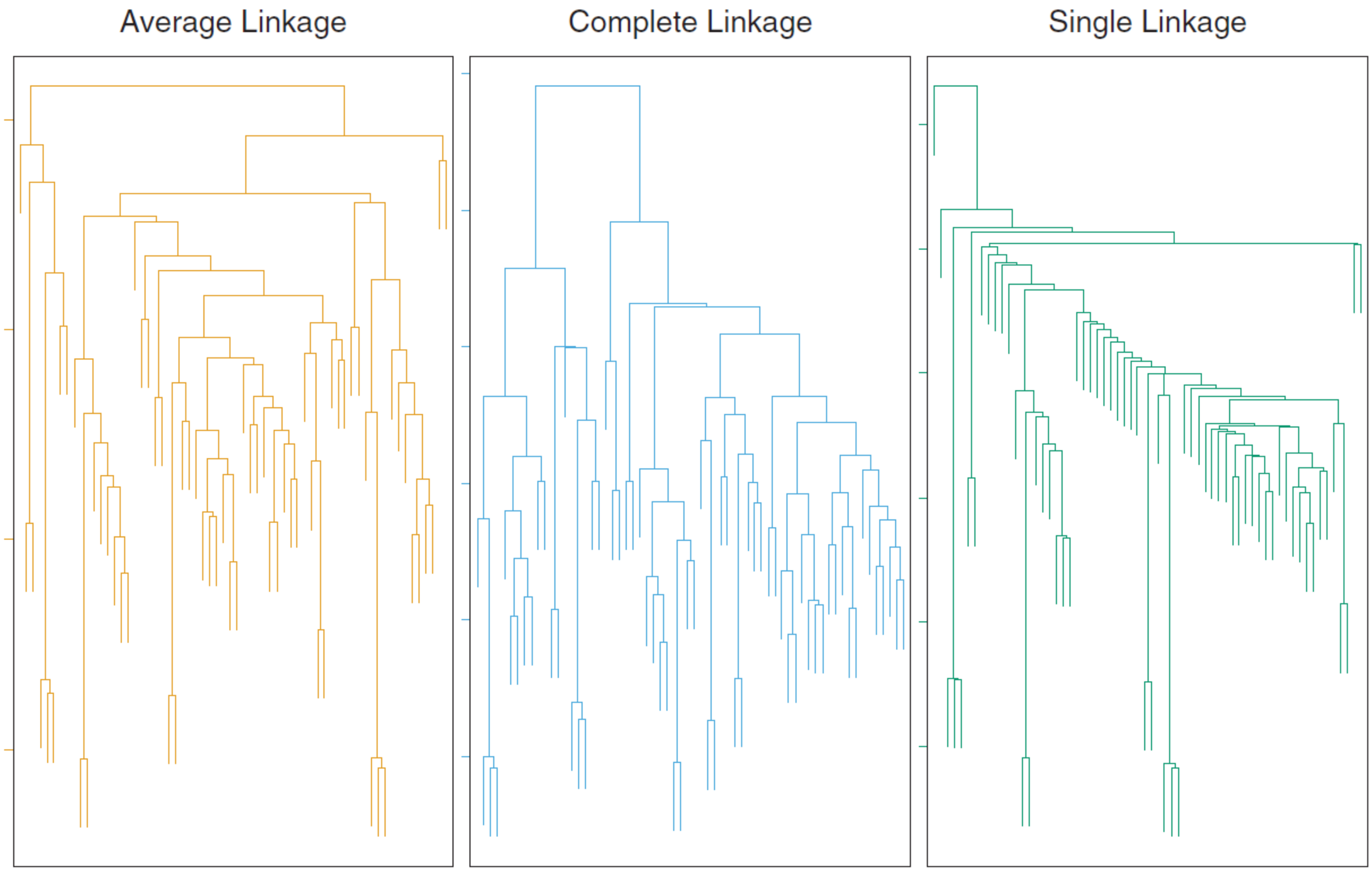
- No conclusions can be drawn
  - Based on the proximity of branches along the horizontal axis
- Observations that fuse lower in the tree
  - Close to each other
- Observations that fuse higher in the tree
  - Are dissimilar
- Height of the fusion
  - Indicate how different the branches / leaves are
- Identifying distinct number of groups
  - Based on the leg lengths

Interpreting the Dendrogram



Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Comparison between linkages

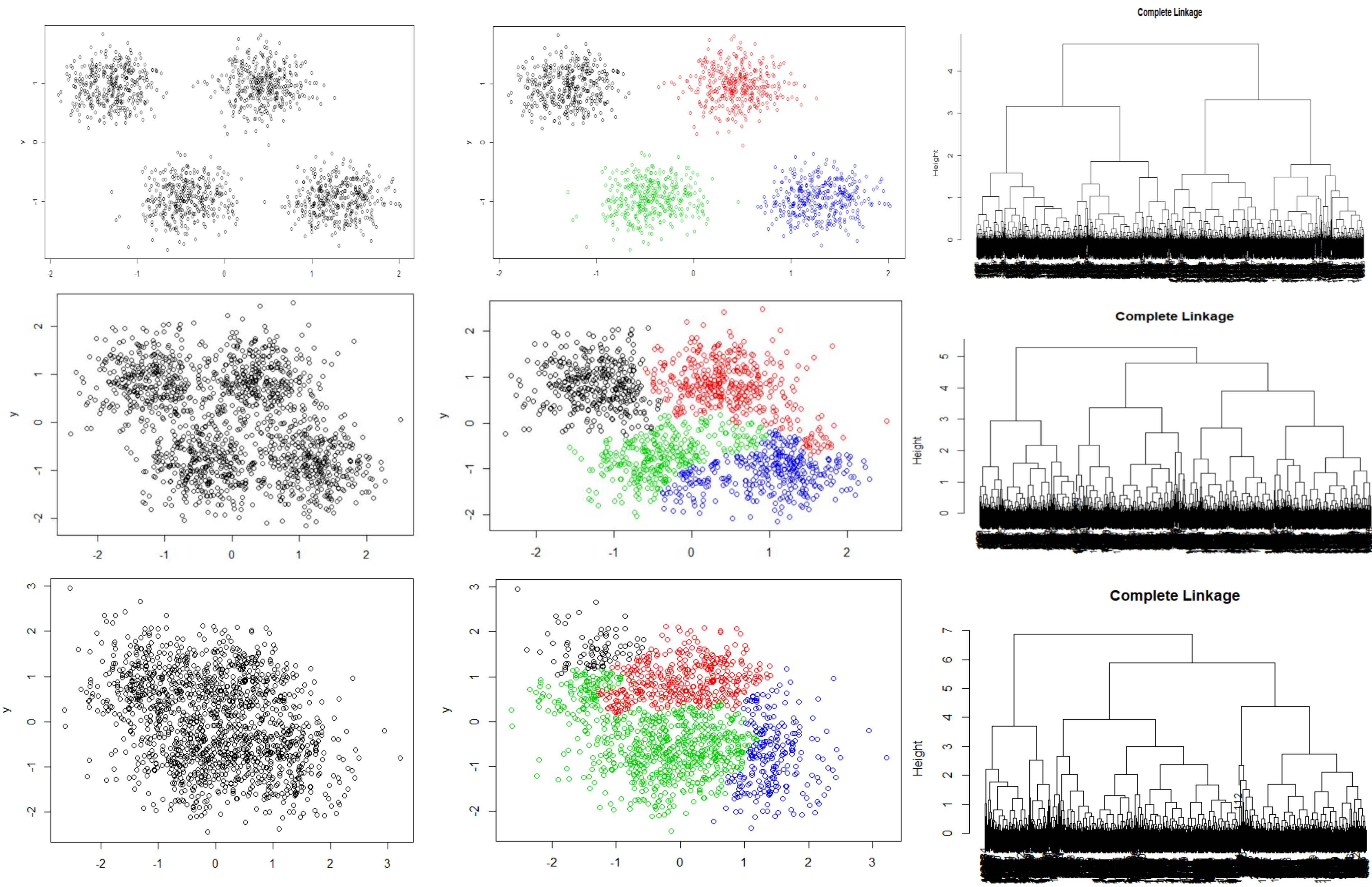


Source: *Introduction to Statistical Learning; Auth: James, Witten, Hastie, Tibshirani; Springer*

Points to be considered in clustering

- Which dissimilarity measure (linkage)?
- How many clusters?
- Presence of outliers
  - Since every point is forced into a cluster
  - This may lead to distortions
- Observations at different scales
  - How to handle such observations?
  - Should they be centered and scaled?

Clustering performance


Complete Linkage
Height
y
x
Complete Linkage
Complete Linkage