K-means clustering

RADI608: Data Mining and Machine Learning
RADI602: Data Mining and Knowledge Discovery

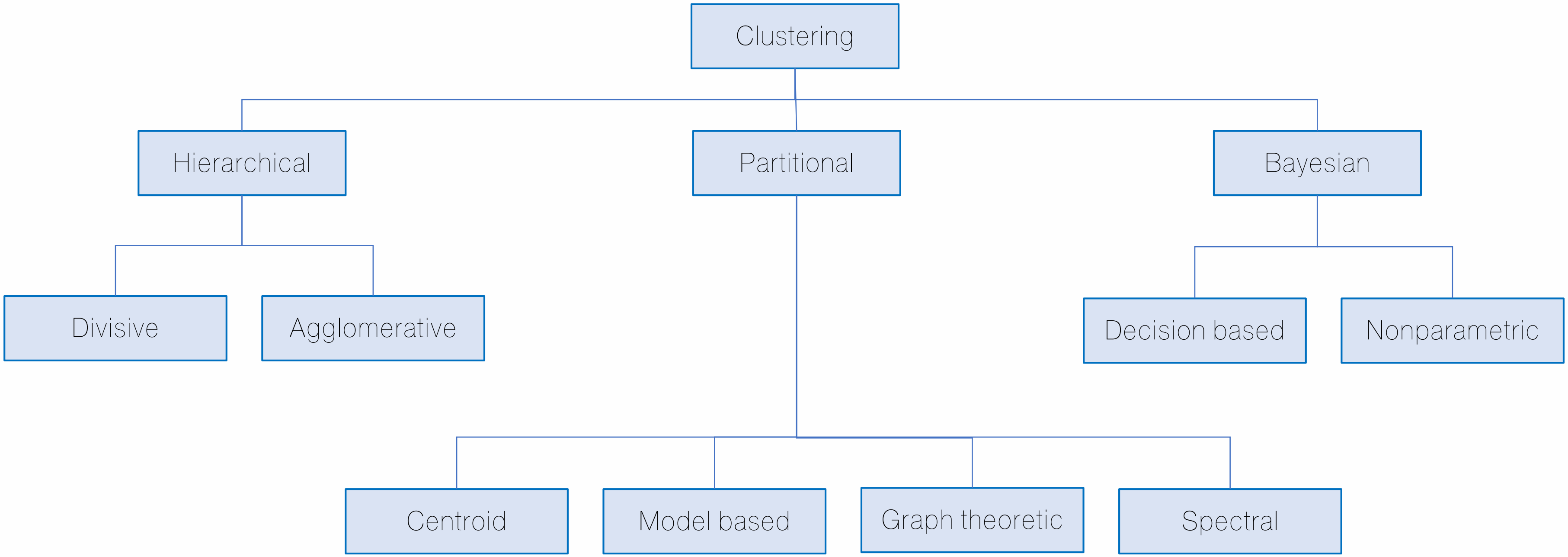
Lect. Anuchate Pattanateepapon. D.Eng

Section of Data Science for Healthcare

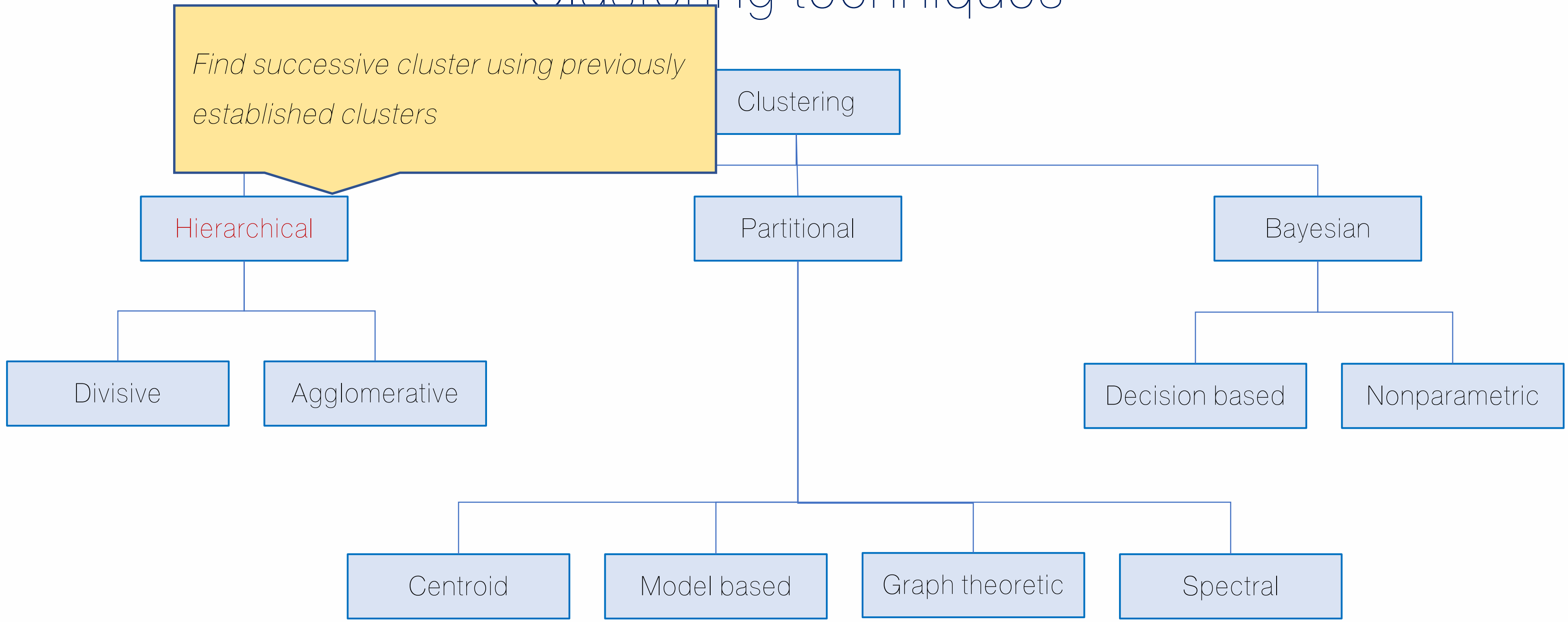
Department of Clinical Epidemiology and Biostatistics

Faculty of Medicine Ramathibodi Hospital, Mahidol University

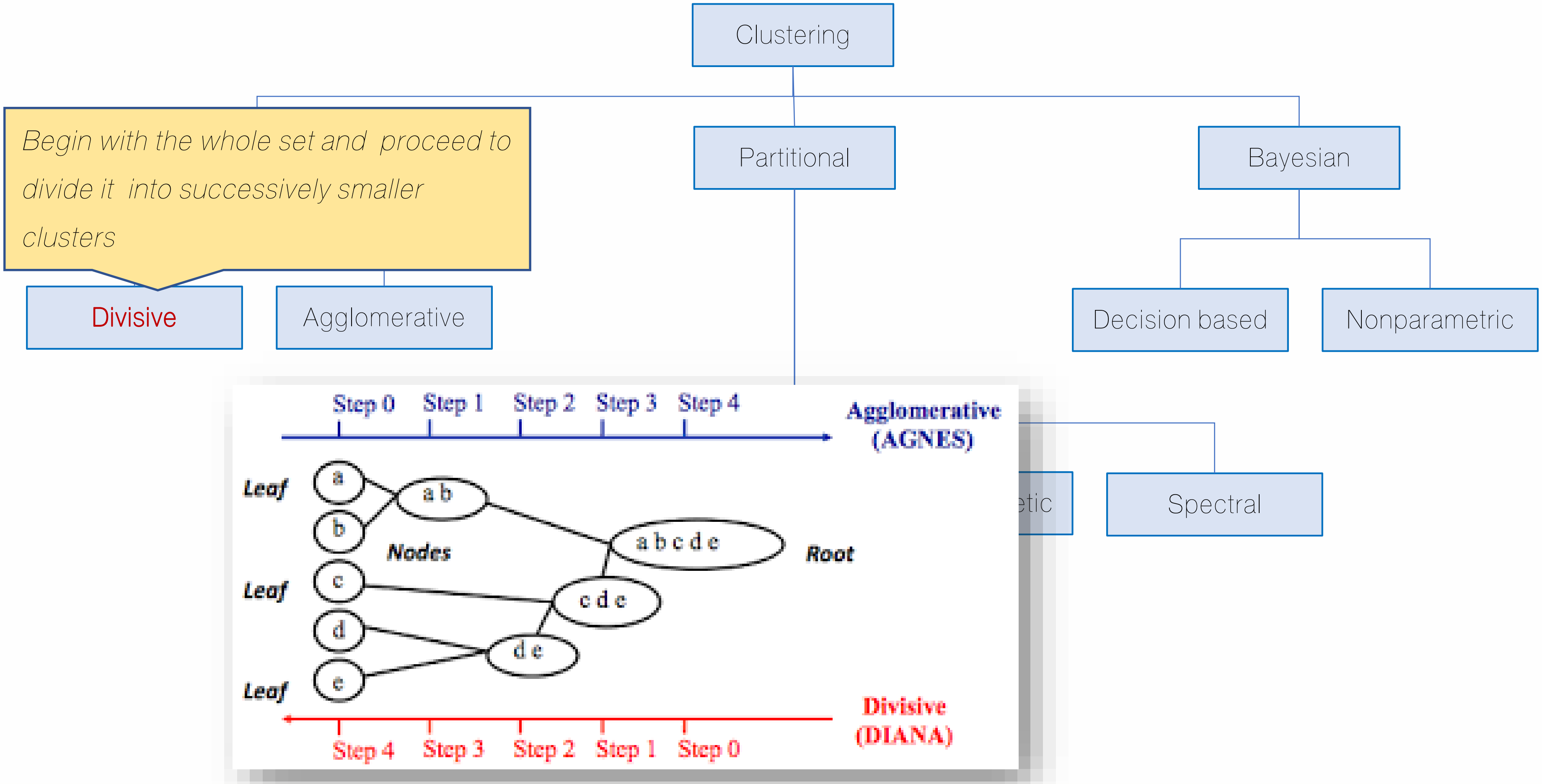
Clustering techniques



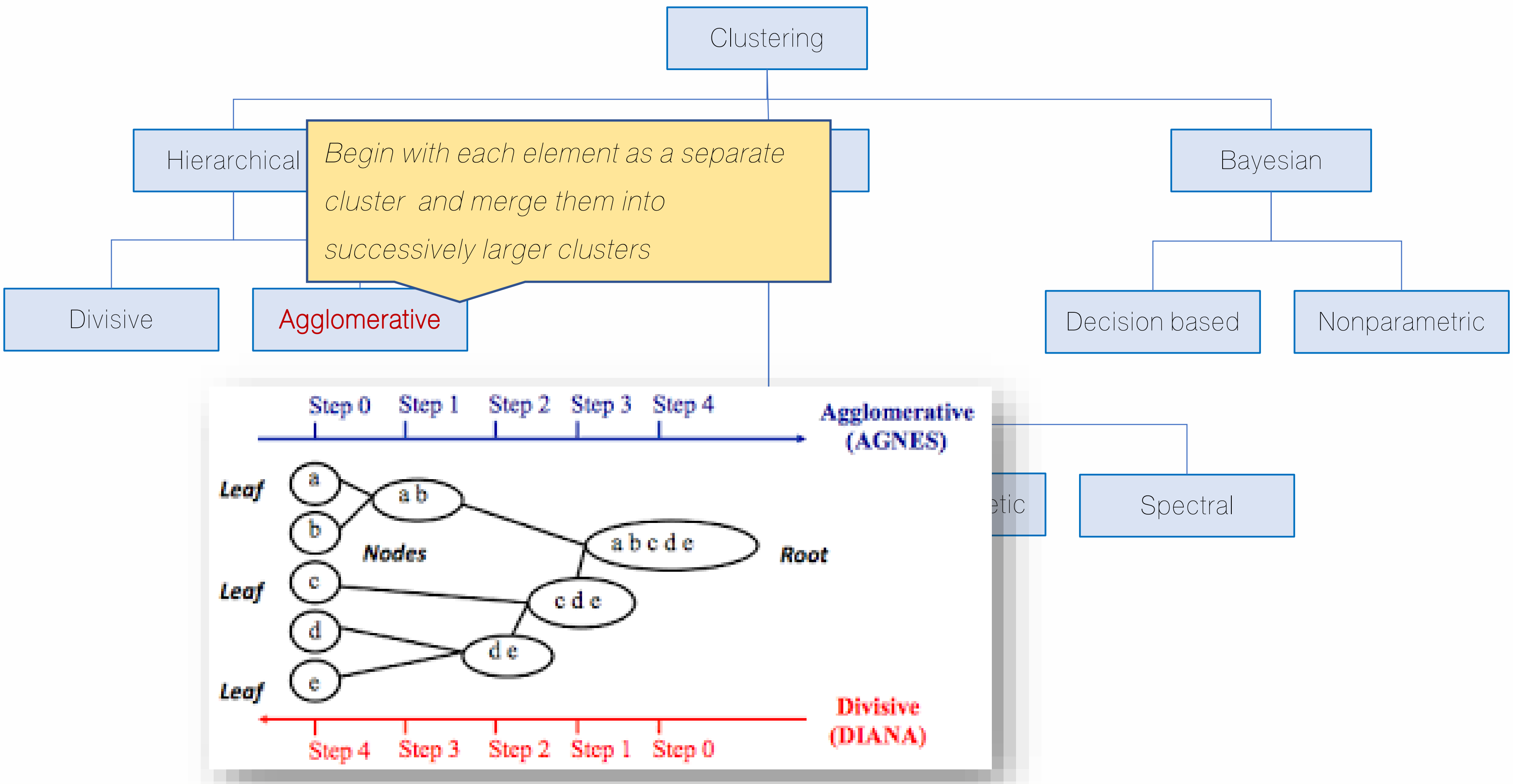
Clustering techniques



Clustering techniques

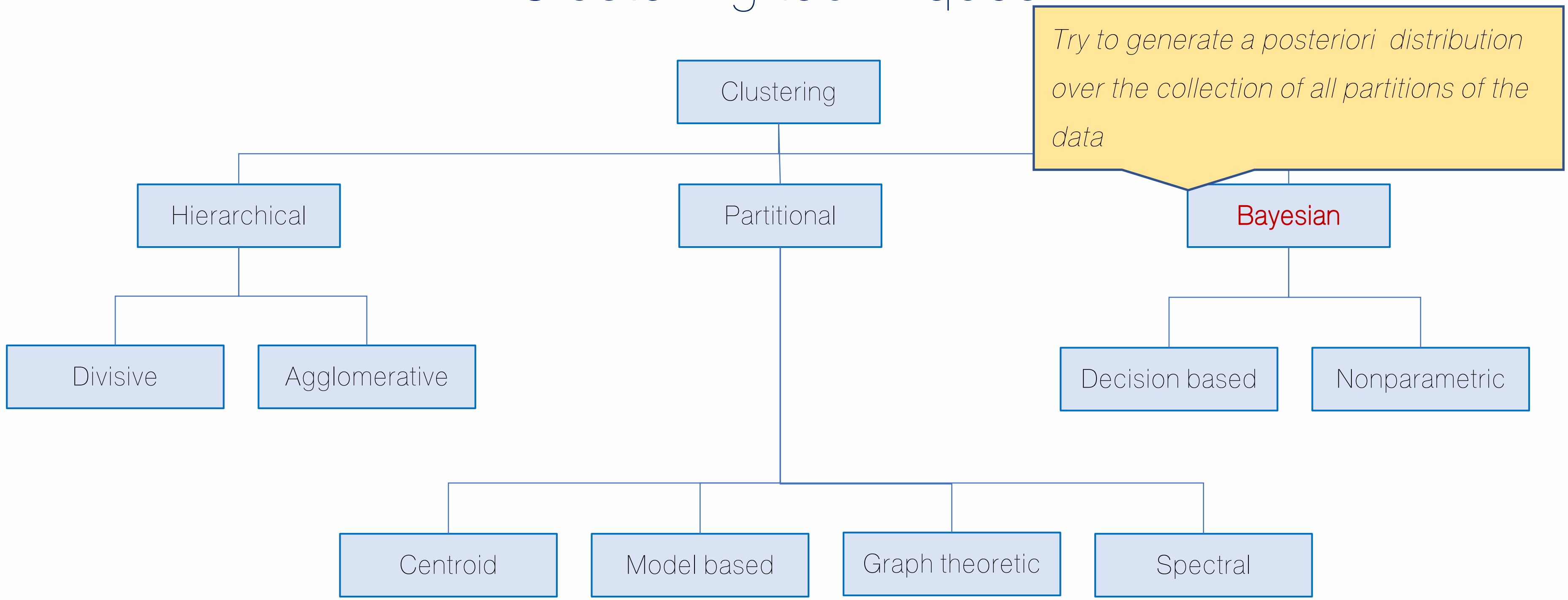


Clustering techniques

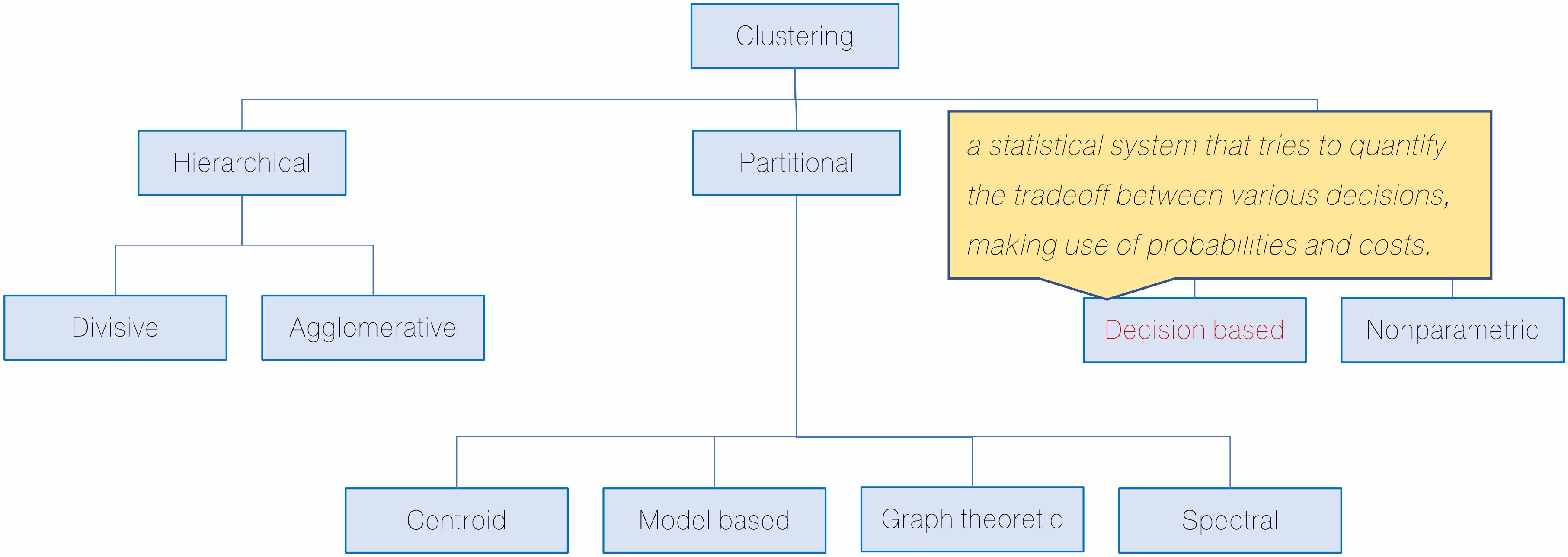


Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

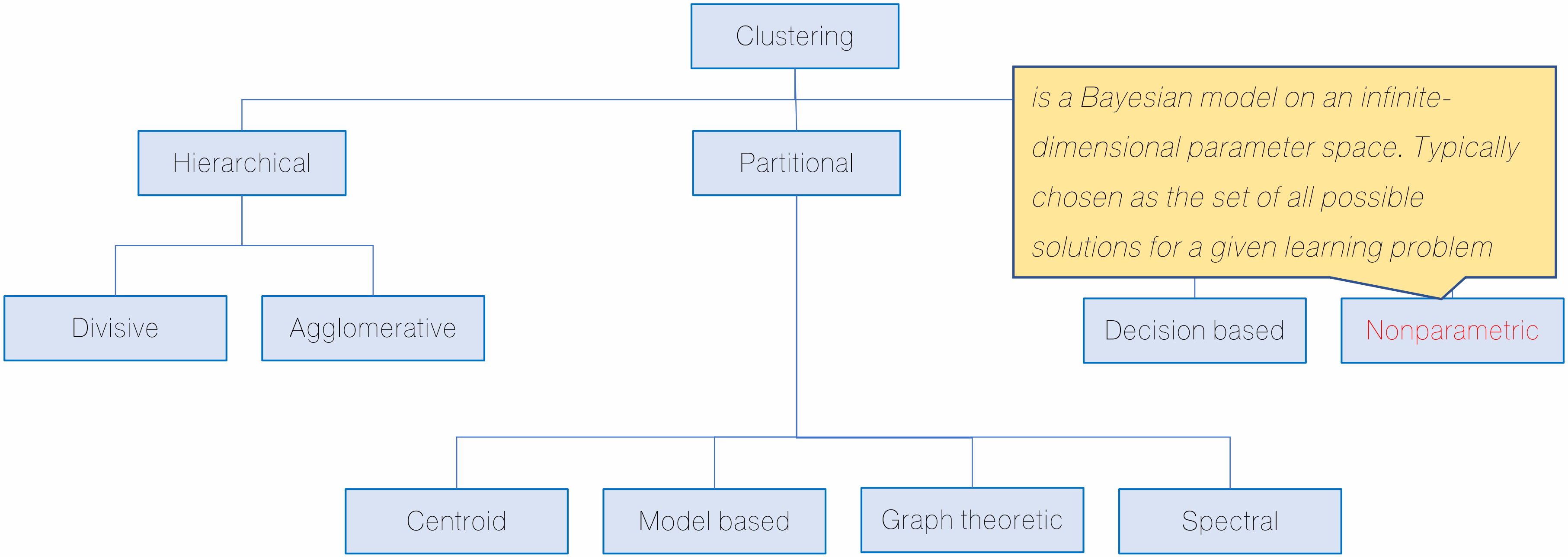
Clustering techniques



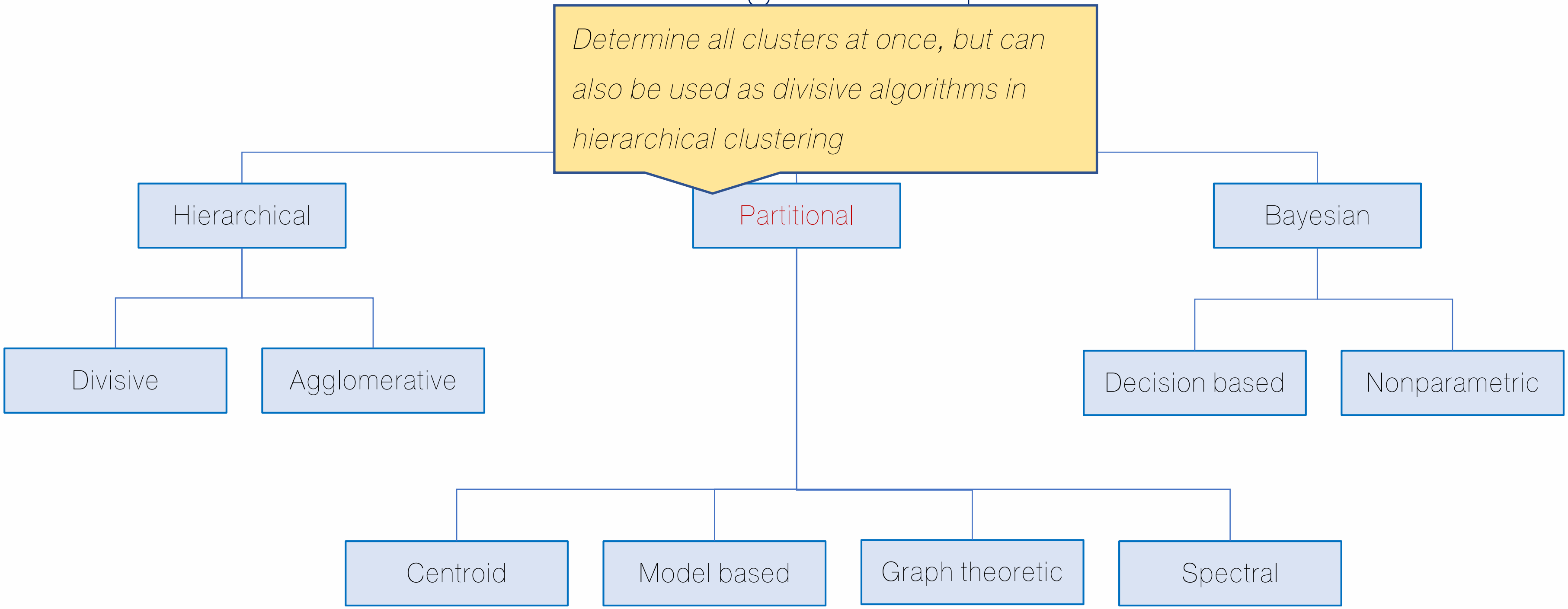
Clustering techniques



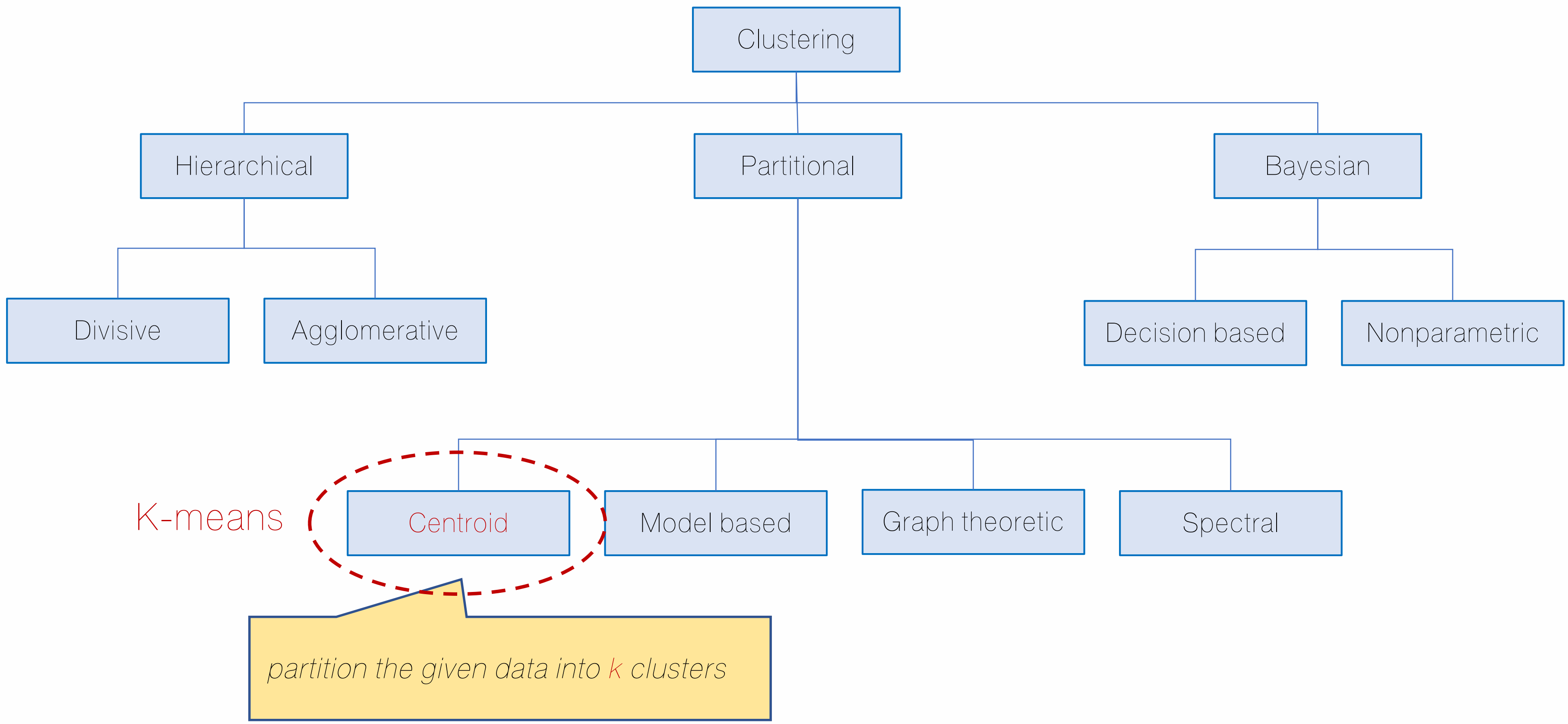
Clustering techniques



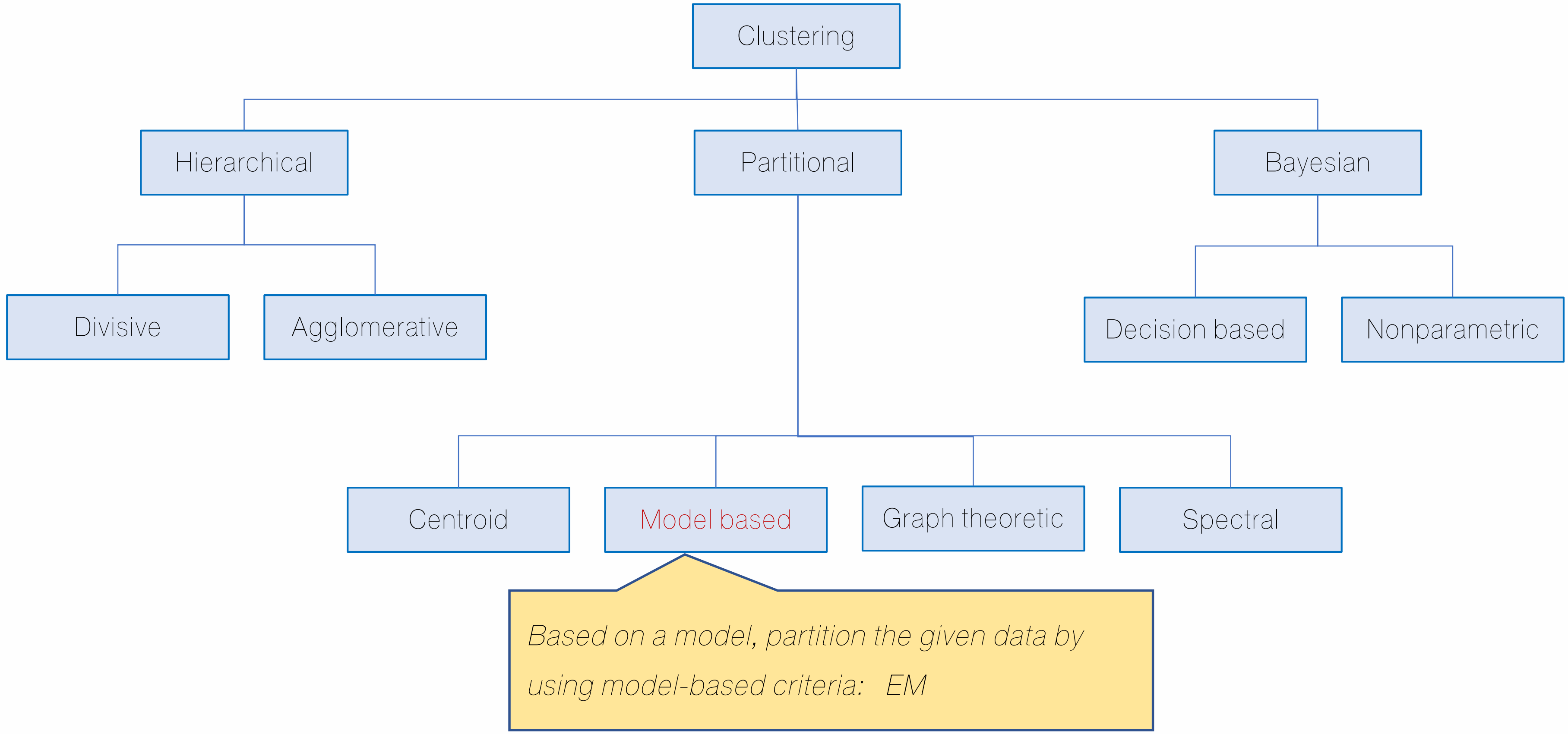
Clustering techniques



Clustering techniques

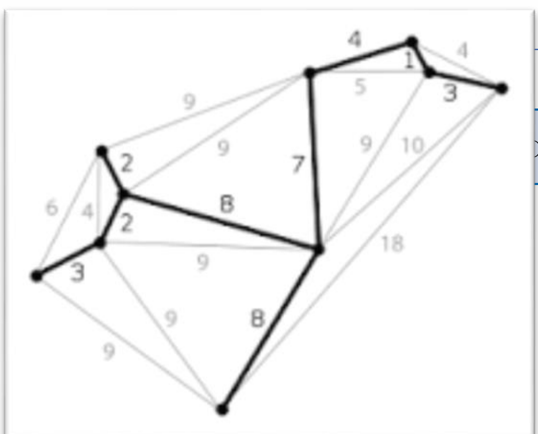


Clustering techniques



Clustering techniques

- Clustering
  - Hierarchical
    - Divisive
    - Agglomerative
  - Partitional
    - del based
    - Graph theoretic
    - Spectral
  - Bayesian
    - Decision based
    - Nonparametric

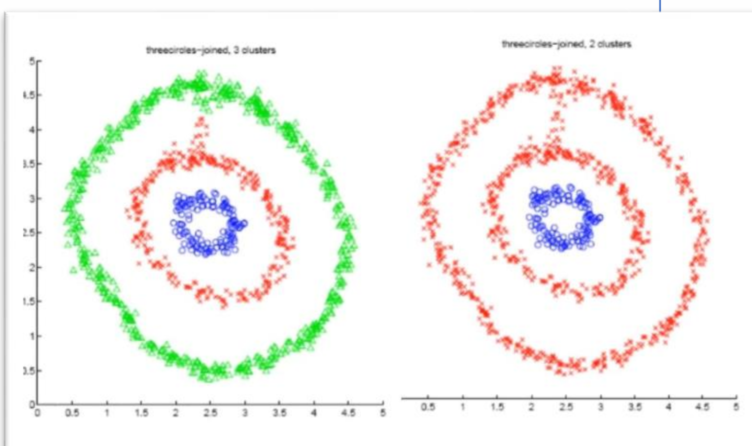


practical solutions are based on heuristics

MST: minimum spanning trees

Clustering techniques

- Clustering
  - Hierarchical
    - Divisive
    - Agglomerative
  - Partitional
    - Graph theoretic
    - Spectral
  - Bayesian
    - Decision based
    - Nonparametric



use k eigenvectors to construct k-way partitionings

K-means clustering

- K-means (MacQueen, 1967) is a partitional clustering algorithm

- Let the set of data points D be $\{x_1, x_2, \ldots, x_n\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$, $i=1,\ldots,n$ is a vector in $X \subseteq R_r$, and r is the number of dimensions

- The k-means algorithm partitions the given data into k clusters:
  - Each cluster has a cluster center, called centroid
  - k is specified by the user

K-means algorithm

The k-means algorithm works as follows:

a) select the number of clusters (k) you want to identify in your data

b) Choose k (random) data points (seeds) to be the initial centroids, cluster centers

c) Assign each data point to the closest centroid

d) Re-compute the centroids using the current cluster memberships

e) If a convergence criterion is not met, repeat steps c) and d)

K-means convergence (stopping) criterion

no (or minimum) re-assignment of data points to different clusters

or

no (or minimum) change of centroids

or

Minimum decrease in the sum of squared error (SSE)

$$\mathrm{SSE} = \sum_{j=1}^{k} \sum_{X \in G_j} d(X, c_j)^2$$

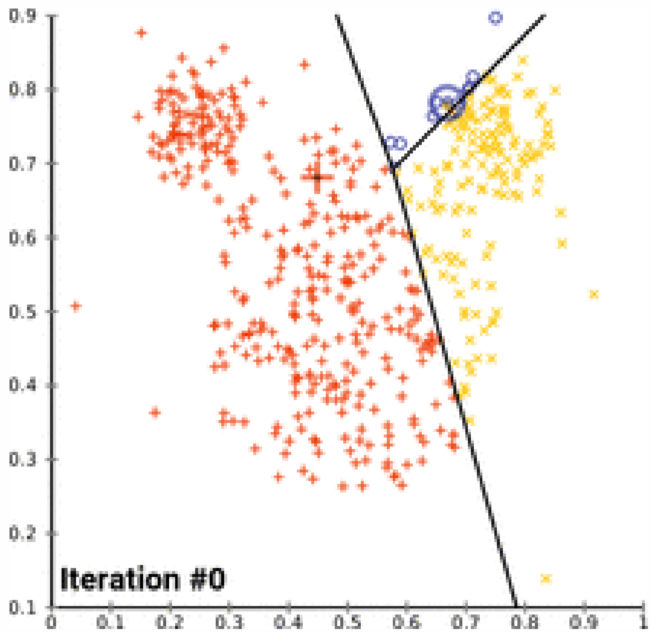
G_j is the j[th] cluster

c_j is the centroid of cluster G_j (the mean vector of all the data points in G_j)

$d(X, c_j)$ is the (Euclidean) distance between data point X and centroid c_j

K-means algorithm

Try to minimize the difference within each cluster and maximize the difference between clusters.



https://en.wikipedia.org/wiki/K-means_clustering

An example of K-means clustering

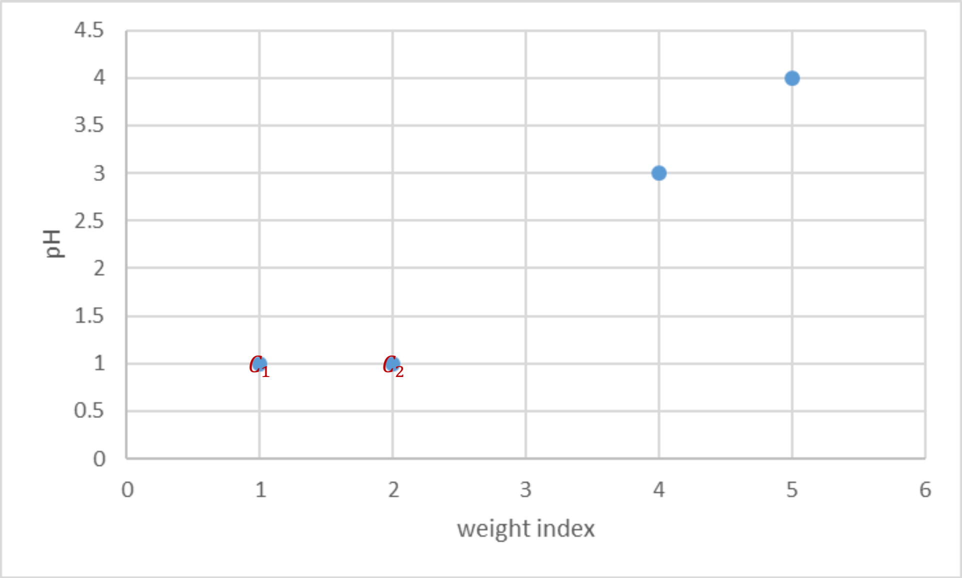
We have 4 types of medicines (samples) with 2 features (weight index and pH).

We have to group these samples into k = 2 group of medicine.

	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4

http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm

An example of K-means clustering



1. From k = 2,
2. we initialized centroid of group 1: C_1 = (1, 1) and centroid of group 2: C_2 = (2, 1)
3. Calculate the distance between cluster centroid to each object (use a Euclidean distance)

Remarks: a distance can calculate by using another methods such as, manhattan distance, etc.

	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4

An example of K-means clustering

Euclidean distance = d(a, b) = d(b, a) , $\sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$

Iteration - 0

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} C_1 = (1, 1) & group\ 1 \\ C_2 = (2, 1) & group\ 2 \end{matrix}$$

Distance between C_1 = (1, 1) and $Medicine\ A$ = (1, 1) :

= $\sqrt{(1-1)^2 + (1-1)^2}$

= 0

Distance between C_2 = (2, 1) and $Medicine\ A$ = (1, 1) :

= $\sqrt{(1-2)^2 + (1-1)^2}$

= 1

	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4

An example of K-means clustering

Euclidean distance = d(a, b) = d(b, a) , $\sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$

Iteration - 0

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{array}{ll} C_1 = (1, 1) & group\ 1 \\ C_2 = (2, 1) & group\ 2 \end{array}$$

Distance between C_1 = (1, 1) and $Medicine\ B$ = (2, 1) :

= $\sqrt{(2-1)^2 + (1-1)^2}$

= 1

Distance between C_2 = (2, 1) and $Medicine\ B$ = (2, 1) :

= $\sqrt{(2-2)^2 + (1-1)^2}$

= 0

	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4

An example of K-means clustering

Euclidean distance = d(a, b) = d(b, a) , $\sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$

Iteration - 0

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} C_1 = (1, 1) & group\ 1 \\ C_2 = (2, 1) & group\ 2 \end{matrix}$$

Distance between C_1 = (1, 1) and *Medicine* C (4, 3) :

= $\sqrt{(4-1)^2 + (3-1)^2}$

= 3.61

Distance between C_2 = (2, 1) and *Medicine C* (4, 3) :

= $\sqrt{(4-2)^2 + (3-1)^2}$

= 2.83

	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4

An example of K-means clustering

Euclidean distance = d(a, b) = d(b, a) , $\sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$

Iteration - 0

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{array}{l} C_1 = (1, 1) \quad group\ 1 \\ C_2 = (2, 1) \quad group\ 2 \end{array}$$

Distance between C_1 = (1, 1) and *medicine D* (5, 4) :

= $\sqrt{(5-1)^2 + (4-1)^2}$

= 5

Distance between C_2 = (2, 1) and *medicine D* (5, 4) :

= $\sqrt{(5-2)^2 + (4-1)^2}$

= 4.24

An example of K-means clustering

4. Object clustering: assign each object based on the minimum distance. A assign to group 1, B, C, D assign to group 2

Iteration - 0

$$D^0 = \begin{array}{c} \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \\ \begin{matrix} A & \quad B & \quad C & \quad D \end{matrix} \end{array} \begin{matrix} C_1 = (1, 1) & group\ 1 \\ C_2 = (2, 1) & group\ 2 \end{matrix}$$

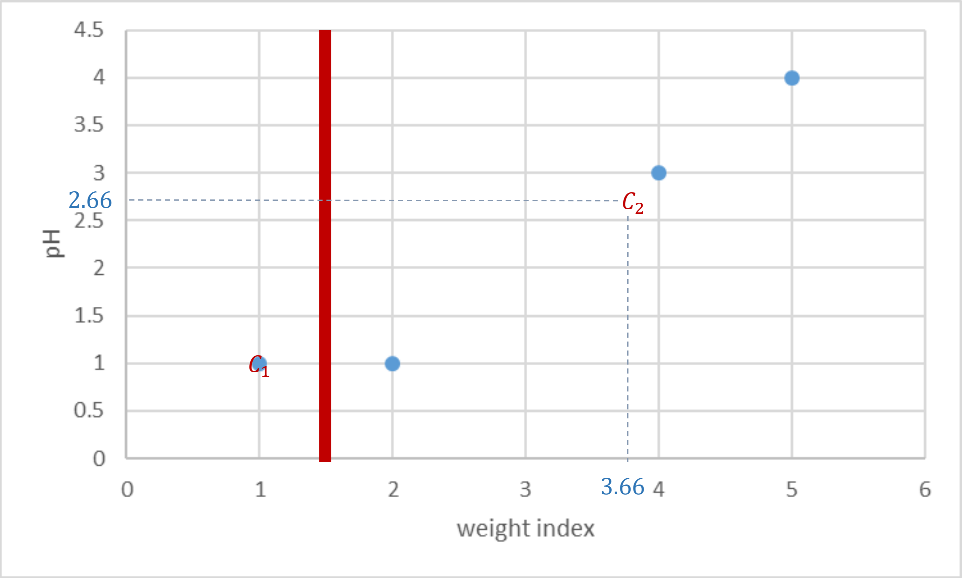
$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group\ 1 \\ group\ 2 \end{matrix}$$

5. Determine centroids for interation-1: calculate new centroid of each group

group 1 has one member, then centroid $C_1 = (1, 1)$

group 2 has three members, then $C_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right)$

An example of K-means clustering



$$C_1 = (1, 1) \text{ and } C_2 = \left(\frac{11}{3}, \frac{8}{3}\right)$$

Iteration - 1

An example of K-means clustering

Iteration - 1

$$D^1 = \begin{array}{c} \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \\ \begin{matrix} A & \quad B & \quad C & \quad D \end{matrix} \end{array} \begin{array}{ll} C_1 = (1, 1) & group\ 1 \\ C_2 = (11/3,\ 8/3) & group\ 2 \end{array}$$

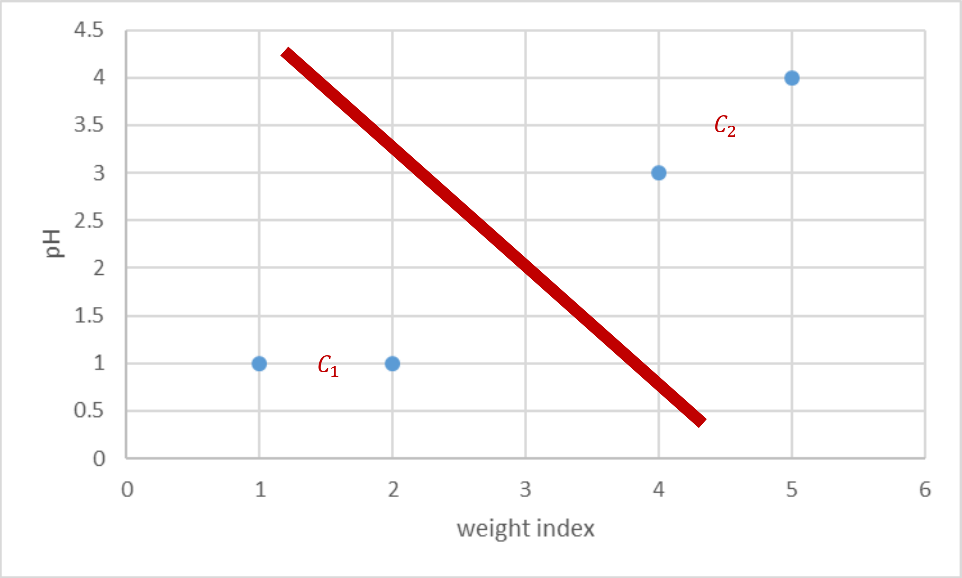
$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{array}{l} group\ 1 \\ group\ 2 \end{array}$$

new centroid of each group

group 1 has two members, then centroid $C_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\frac{1}{2}, 1)$

group 2 has two members, then $C_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\frac{1}{2}, 3\frac{1}{2})$

An example of K-means clustering



$$C_1 = (1\tfrac{1}{2}, 1) \text{ and } C_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$

Iteration - 2

An example of K-means clustering

Iteration - 2

$$D^2 = \begin{bmatrix} 0.50 & 0.50 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{array}{l} C_1 = (1\frac{1}{2}, 1) \quad group\ 1 \\ C_2 = (4\frac{1}{2}, 3\frac{1}{2}) \quad group\ 2 \end{array}$$

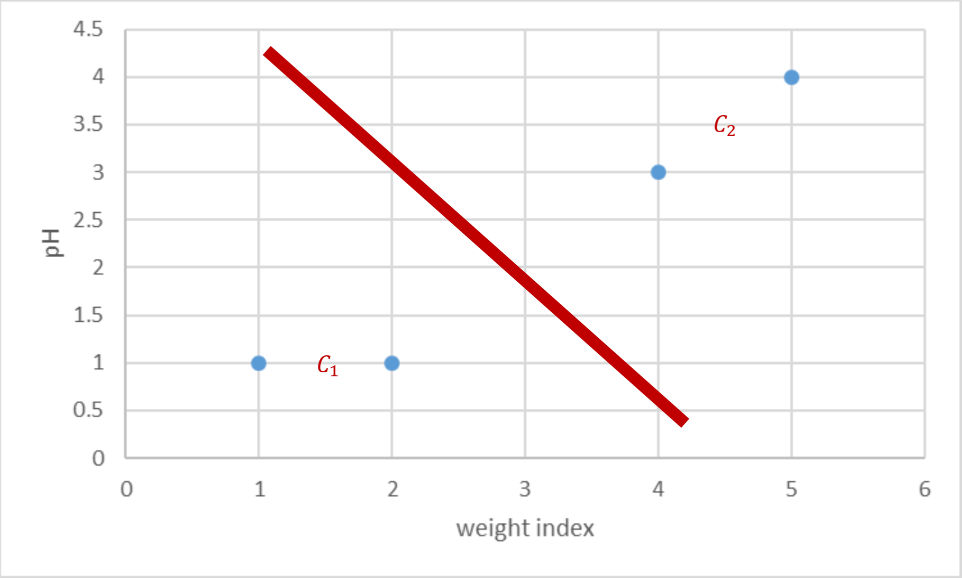
$$\begin{array}{cccc} A & B & C & D \end{array}$$

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{array}{l} group\ 1 \\ group\ 2 \end{array}$$

$$G^1 = G^2$$

The object does not move anymore, then k-mean clustering has reached its stability and no more iteration is needed.

An example of K-means clustering



	medicine A	medicine B	medicine C	medicine D
weight index	1	2	4	5
pH	1	1	3	4
Group or Class	1	1	2	2

K-means clustering in Python

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

K-means clustering in Python

```
from sklearn.cluster import KMeans
import numpy as np
X = np.array([[1, 2], [1, 4], [1, 0],
              [10, 2], [10, 4], [10, 0]])
kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
kmeans.labels_
```

Set K = 2 clusters

Set random state = 0

array([1, 1, 1, 0, 0, 0])

```
kmeans.predict([[0, 0], [12, 3]])
```

Predict unlabeled data

array([1, 0])

```
kmeans.cluster_centers_
```

array([[10., 2.], [1., 2.]])

K-means clustering in Python

run kmean2.py

```
print(__doc__)


# Code source: Gaël Varoquaux
# Modified for documentation by Jaques Grobler
# License: BSD 3 clause

import numpy as np
import matplotlib.pyplot as plt
# Though the following import is not directly being used, it is required
# for 3D projection to work
from mpl_toolkits.mplot3d import Axes3D

from sklearn.cluster import KMeans
from sklearn import datasets

np.random.seed(5)

iris = datasets.load_iris()
X = iris.data
y = iris.target

estimators = [('k_means_iris_8', KMeans(n_clusters=8)),
              ('k_means_iris_3', KMeans(n_clusters=3)),
              ('k_means_iris_bad_init', KMeans(n_clusters=3, n_init=1,
                                               init='random'))]
```

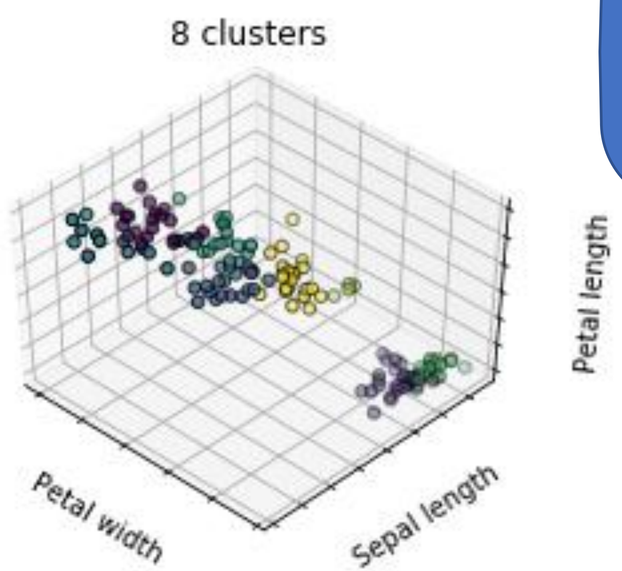
K-means clustering in Python

Set K = 8 clusters:

KMeans(n_clusters=8)

Confusion Matrix

```
[[ 0 28  0  0  0 22  0  0]
 [ 0  0 20  0  3  0  4 23]
 [22  0  0 12 15  0  0  1]
 [ 0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0]]
```



K-means clustering in Python

Set K = 3 clusters:

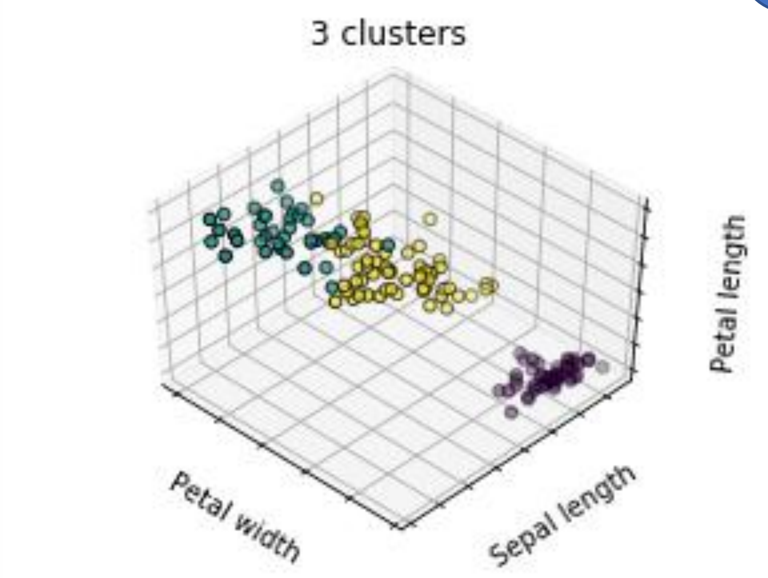
KMeans(n_clusters=3)

Confusion Matrix

[[50 0 0]

[0 2 48]

[0 36 14]]



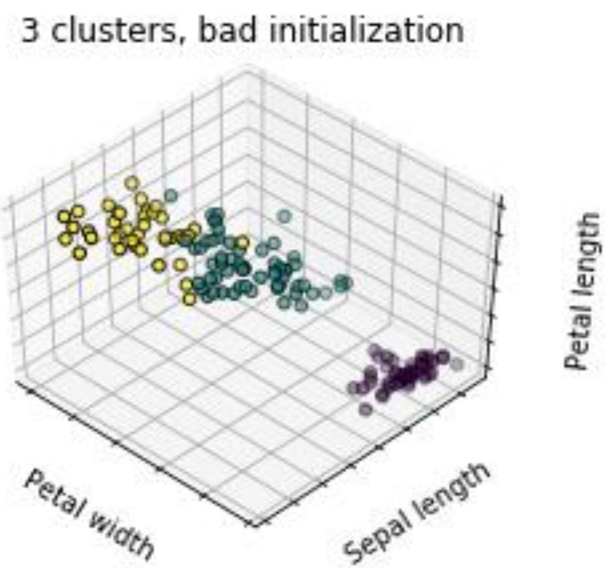
K-means clustering in Python

Set K = 3 clusters with bad initialization

KMeans(n_clusters=3, n_init=5, init='random')

A Bad initialization is on the classification process:
By setting n_init to only 5 (default is 10), the amount of times that the algorithm will be run with different centroid's initialization is reduced.

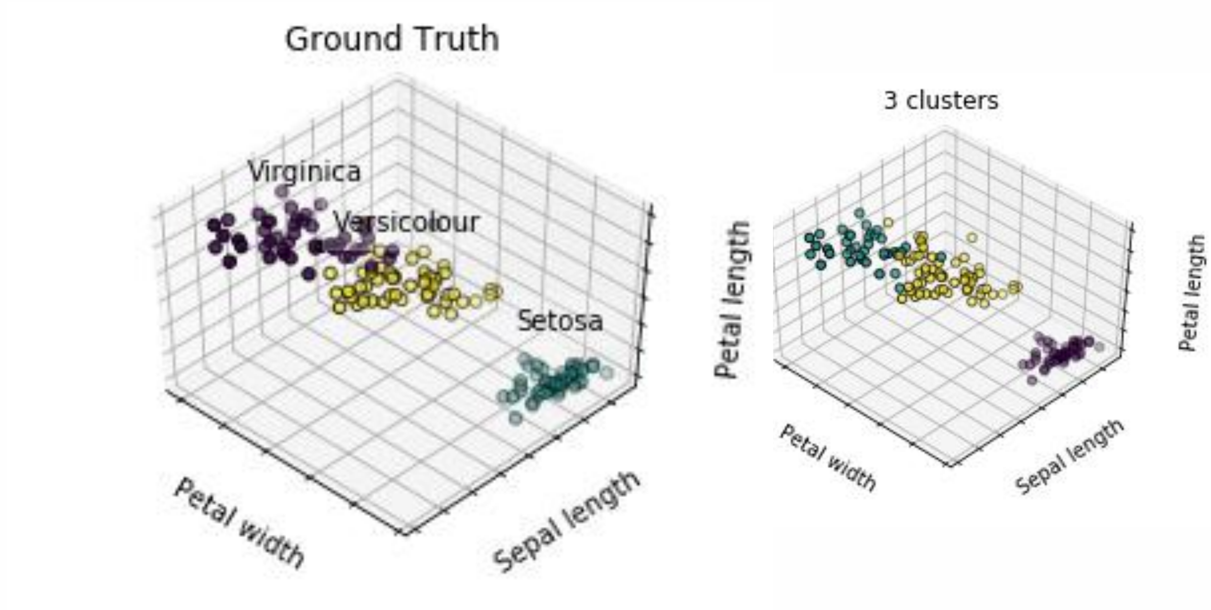
Note: Run only 5 times and select the best on



Confusion Matrix
[[0 0 50]
[2 48 0]
[36 14 0]]

K-means clustering in Python

The comparison between K-Means clustering at K = 3 and the Ground Truth (Real labeled data)



Introduction of "K" estimation techniques

a) Applied a Hierarchical Clustering

b) Applied an elbow plot

Hierarchical Clustering in Python

1. Compute distance between every pairs of point/cluster
   - Distance between point is just using the distance function
   - Compute distance between point X to cluster A may involve many choices (such as the min/max/average distance between the point X and points in the cluster A
   - Compute distance between cluster A and the other from cluster B and then pick either min/max/average of these pairs
2. Combine the two closet point/cluster into a cluster, Go back to 1) until only one big cluster remains

https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

Hierarchical Clustering in Python

```
from sklearn import datasets
from scipy.cluster.hierarchy import dendrogram, linkage
from matplotlib import pyplot as plt
iris = datasets.load_iris()
X = iris.data

linked = linkage(X, 'single')
labelList = range(150)
plt.figure(figsize=(10, 7))
dendrogram(linked,
           orientation='top',
           labels=labelList,
           distance_sort='descending',
           show_leaf_counts=True)
plt.show()
```

Single assign to cluster (one centroid)

labeling all 150 points

Figure size 10 * 7

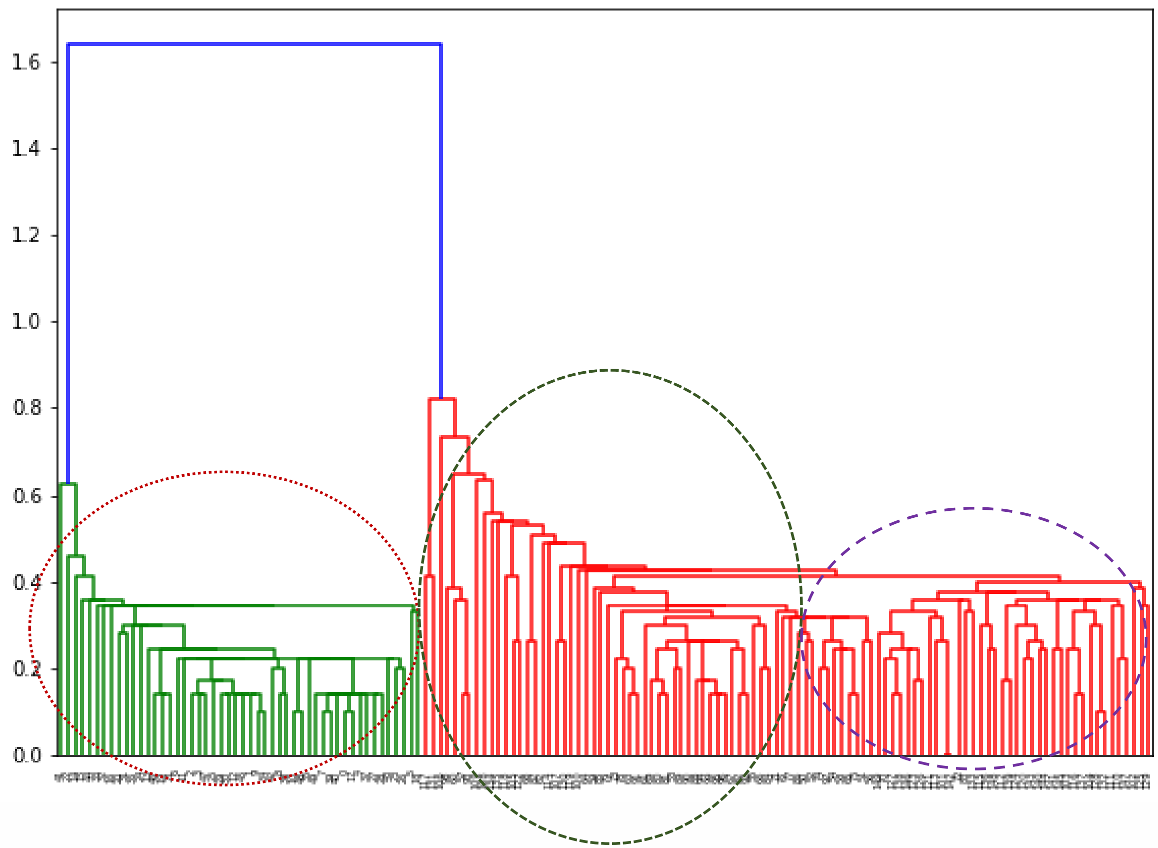
Plots the root at the top, and plot descendent links going downwards

The child with the maximum distance between its direct descendents is plotted first

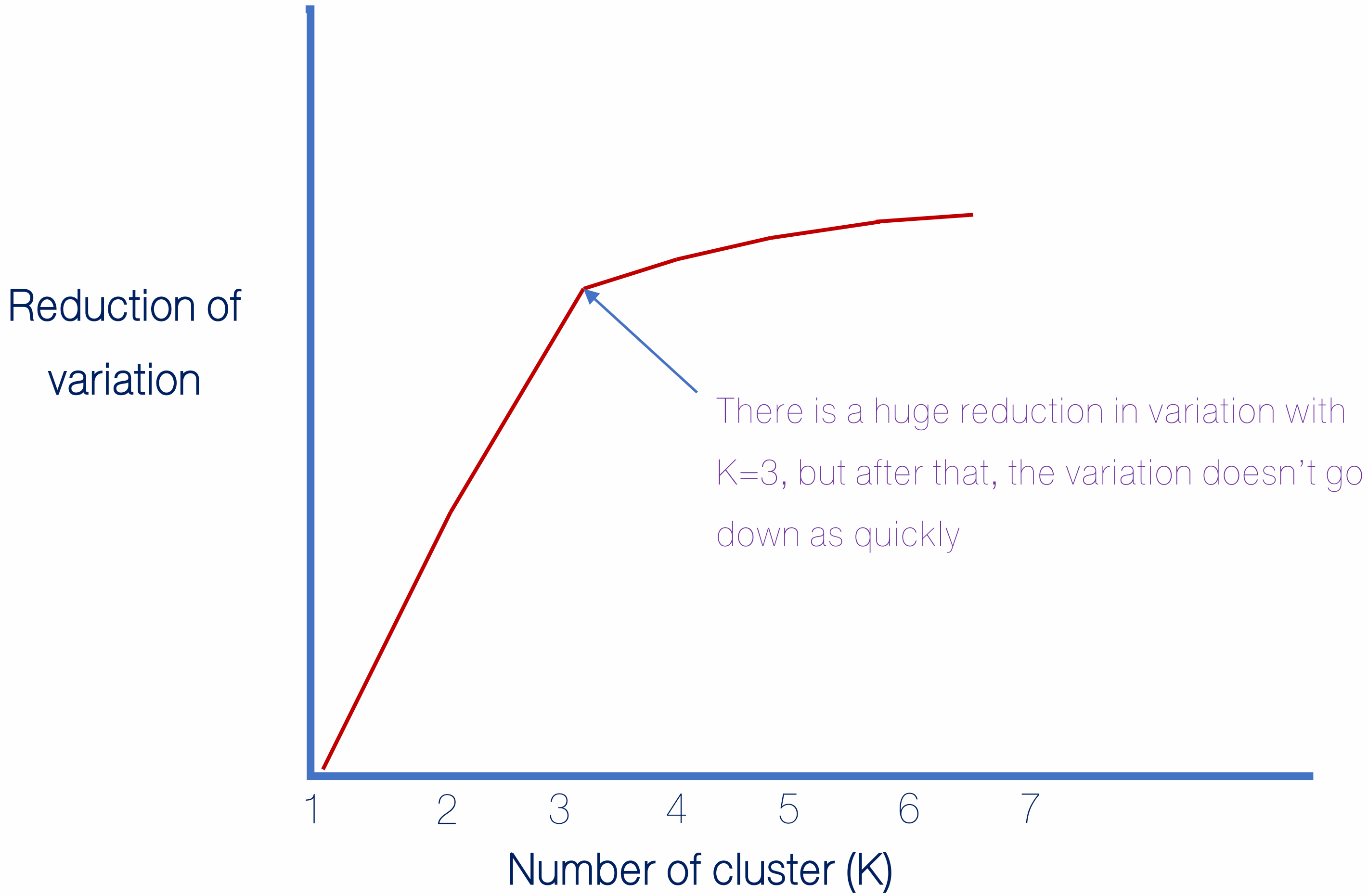
leaf nodes representing k>1 original observation are labeled with the number of observations they contain in parentheses.

Estimate a value of K by using a hierarchical clustering

Three big hierarchies



Pick up "K" by finding the elbow in the plot



Assignment:

Due date: November 21, 2022 (10 points)

Find (by manual and show calculation steps) the appropriate centroids by using a K-means clustering with Euclidean distance (K = 2)

Samples	S1	S2	S3	S4	S5	S6
Feature #1	1	2	1	5	4	5
Feature #2	1	1	3	2	3	4

Samples	STU#1	STU#2	STU#3	STU#4	STU#5	STU#6
C1	(1,1)	(2,1)	(5,2)	(5,2)	(4,3)	(2,1)
C2	(1,3)	(1,3)	(4,3)	(5,4)	(4,4)	(5,2)