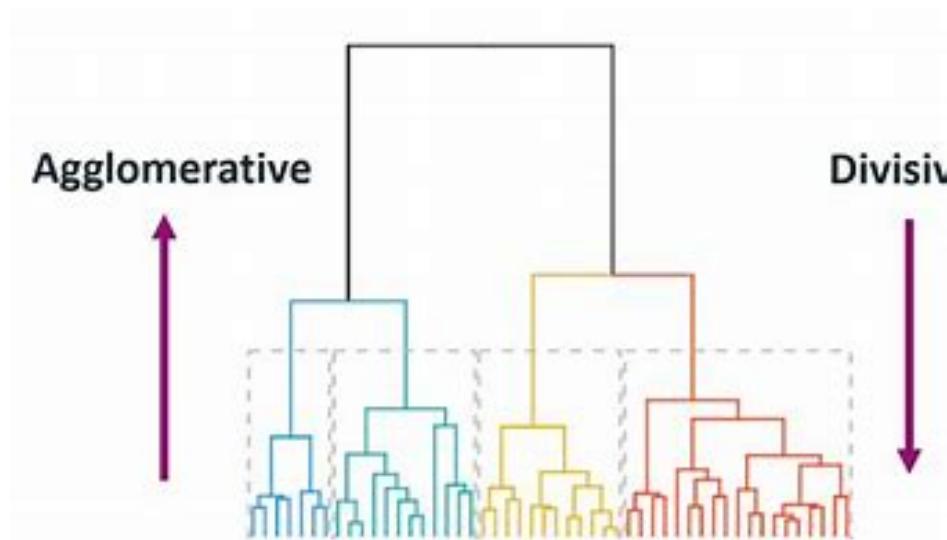


# Hierarchical Clustering

- Pendekatan klastering berbasis hirarki
- Terbagi kedalam dua kelompok:
  - Agglomerative Clustering (bottom up)
  - Divisive Clustering (top down)

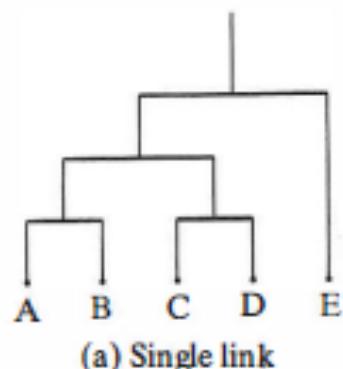


# Agglomerative Clustering

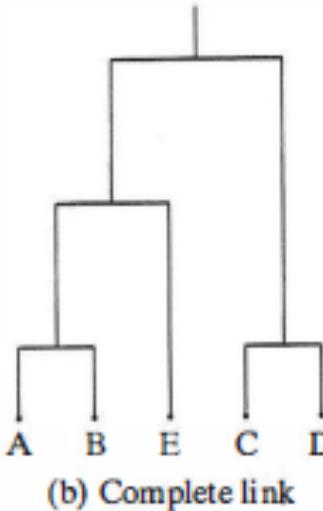
- Menggunakan distance matrix
- Menghasilkan Dendrogram

Distance Matrix

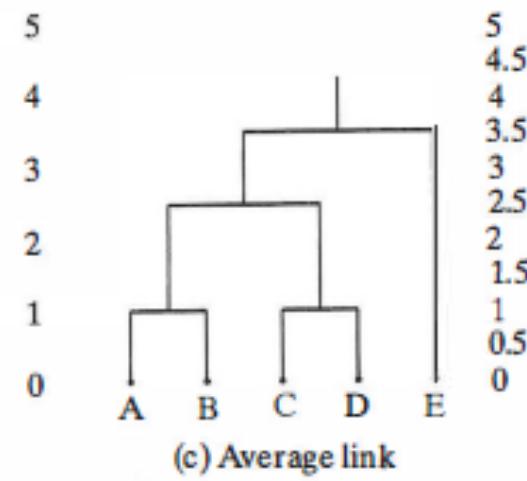
Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



(a) Single link



(b) Complete link



(c) Average link

$$d_{AB \rightarrow C} = \min(d_{AC}, d_{BC})$$

$$d_{AB \rightarrow C} = \max(d_{AC}, d_{BC})$$

$$d_{AB \rightarrow C} = \text{avg}(d_{AC}, d_{BC})$$

## Contoh

Dalam contoh ini sudah terbentuk distance matrix, kemudian kita bisa menggunakan beberapa linkage untuk membentuk dendrogram dan melakukan clustering.

	A	B	C	D
A	0			
B	2.24	0		
C	2.24	1.41	0	
D	1.73	2	2	0

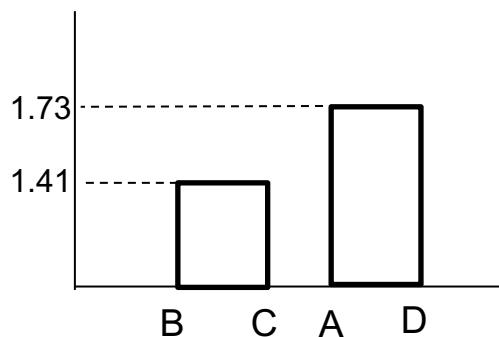
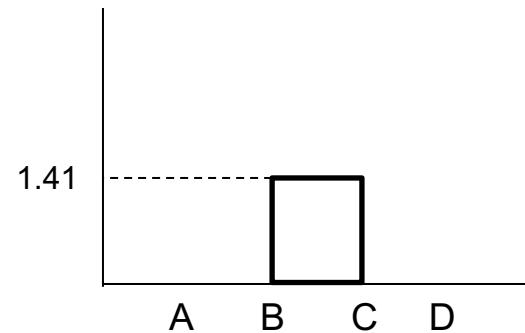
# Single Linkage

	A	B	C	D
A	0			
B	2.24	0		
C	2.24	1.41	0	
D	1.73	2	2	0

	A	BC	D
A	0		
BC	2.24	0	
D	1.73	2	0

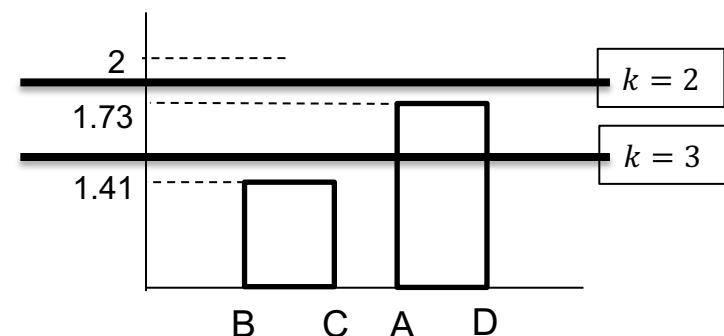
$$d_{A \rightarrow BC} = \min(d_{AB}, d_{AC}) = \min(2.24, 2.24) = 2.24$$

$$d_{D \rightarrow BC} = \min(d_{DB}, d_{DC}) = \min(2, 2) = 2$$



	AD	BC
AD	0	
BC	2	0

$$\begin{aligned} d_{BC \rightarrow AD} &= \min(d_{BCA}, d_{BCD}) \\ &= \min(2.24, 2) = 2 \end{aligned}$$



$k = 2 \rightarrow \{AD\} \ \{BC\}$

$k = 3 \rightarrow \{A\} \ \{D\} \ \{BC\}$

# Davies Bouldin Index

$$DBI = \frac{1}{N} \sum_{i=1}^N R_i$$

$$R_i = \max_{i \neq j} R_{i,j}$$

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

$$SSW_i = \left( \frac{1}{T_i} \sum_{j=1}^{T_i} |x_j - A_i|^p \right)^{1/p}$$

$$SSB_{i,j} = \left( \sum_{k=1}^n |A_{i,k} - A_{j,k}|^p \right)^{1/p}$$

$p=1$  for Manhattan Distance,  $p=2$  for Euclidean Distance

$n$  = the vector length of centroid

$N$  = number of the cluster

$T_i$  = size of cluster  $i$

$A_i$  = centroid of the cluster  $i$

$x$  = data points

$SSW_i$  = Sum of Square Within-Cluster  $i$

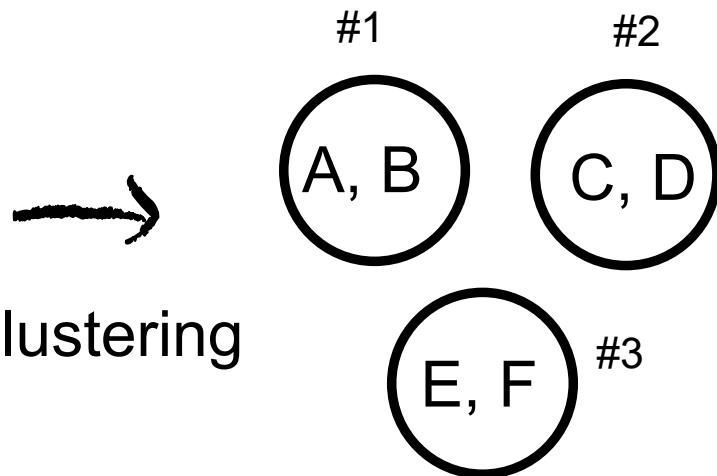
$SSB_{i,j}$  = Sum of Square Between-Cluster  $i$  and  $j$

# Davies Bouldin Index

Raw data

Data	x	y
A	1	2
B	3	1
C	5	2
D	4	5
E	7	8
F	7	10

Clustering



DBI ?

$$DBI = \frac{1}{N} \sum_{i=1}^N R_i$$

$$R_i = \max_{i \neq j} R_{i,j}$$

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

$$SSW_i = \left( \frac{1}{T_i} \sum_{j=1}^{T_i} |x_j - A_i|^p \right)^{1/p}$$

$$SSB_{i,j} = \left( \sum_{k=1}^n |A_{i,k} - A_{j,k}|^p \right)^{1/p}$$

# Hierarchical Clustering di Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn import datasets
iris = datasets.load_iris()

df=pd.DataFrame(iris['data'])
print(df.head())
```

	0	1	2	3
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

Penggunaan data iris yang terdiri dari 4 atribut: sepal length, sepal width, petal length, dan petal width.

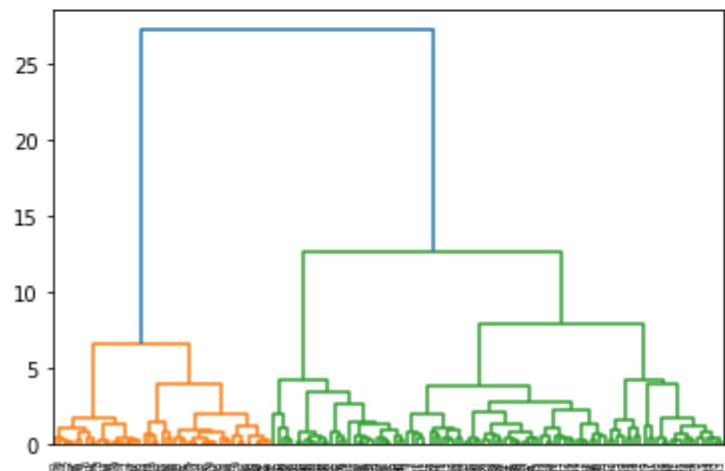


	A	B	C	D	E	
1	Sepal Length	Sepal Width	Petal Length	Petal Width	Class	
2	5.1	3.5	1.4	0.2	Iris-setosa	
3	4.9	3	1.4	0.2	Iris-setosa	
4	4.7	3.2	1.3	0.2	Iris-setosa	
5	4.6	3.1	1.5	0.2	Iris-setosa	
6	5	3.6	1.4	0.2	Iris-setosa	

```
# Import the whiten function
from scipy.cluster.vq import whiten
scaled_data = whiten(df.to_numpy())
```

```
# Import the fcluster and linkage functions
from scipy.cluster.hierarchy import fcluster, linkage
# Use the linkage() function
distance_matrix = linkage(scaled_data, method = 'ward', metric = 'euclidean')
```

```
# Import the dendrogram function
from scipy.cluster.hierarchy import dendrogram
# Create a dendrogram
dn = dendrogram(distance_matrix)
# Display the dendrogram
plt.show()
```



# Latihan Soal (Kuis)

Lakukan pengelompokan menjadi 3 cluster pada distance matrix dibawah ini.

	A	B	C	D	E
A	0				
B	2.5	0			
C	10.44	12.5	0		
D	4.12	6.4	6.48	0	
E	11.75	13.93	1.41	7.35	0