DIGITAL TALENT SCHOLARSHIP 2019

DIGITAL TALENT SCHOLARSHIP

digitalent.kominfo.go.id

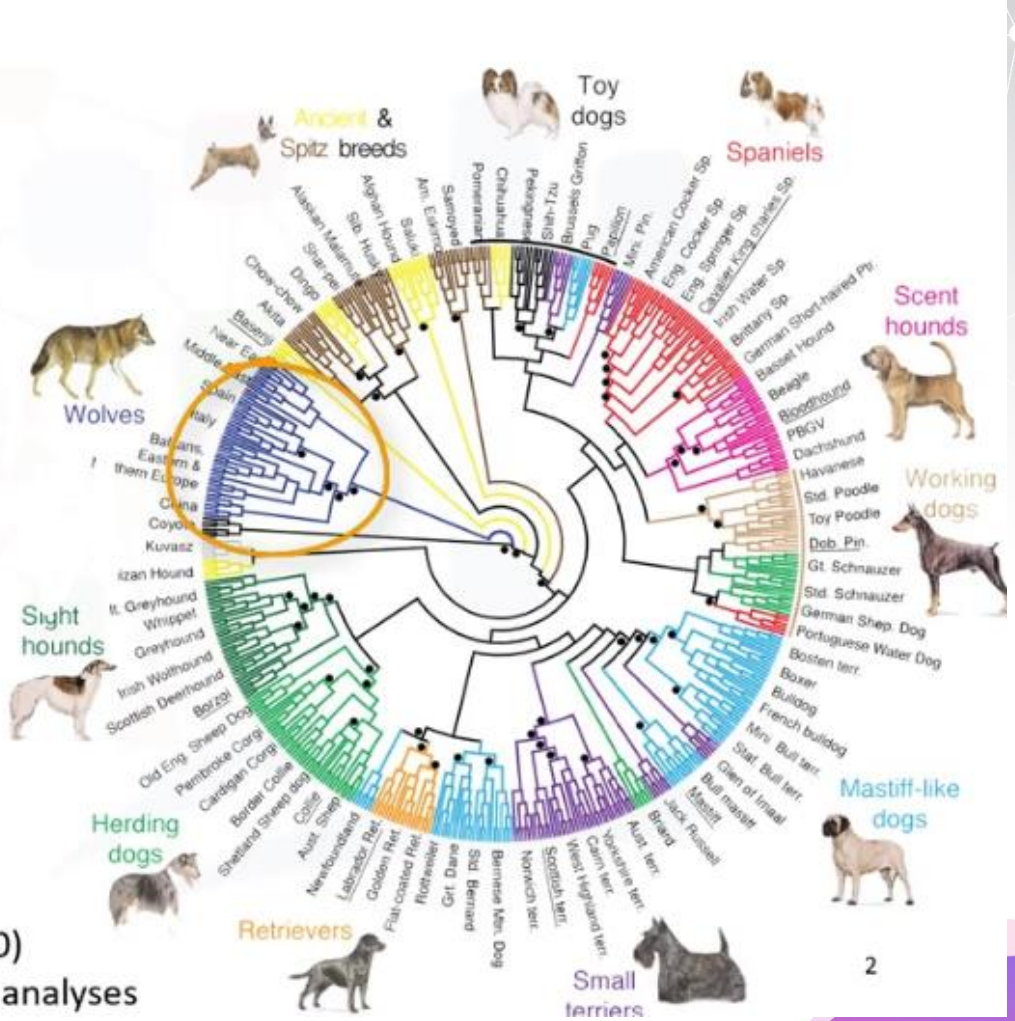
Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

Clustering: Hierarchial

Nama pembicara dengan gelar

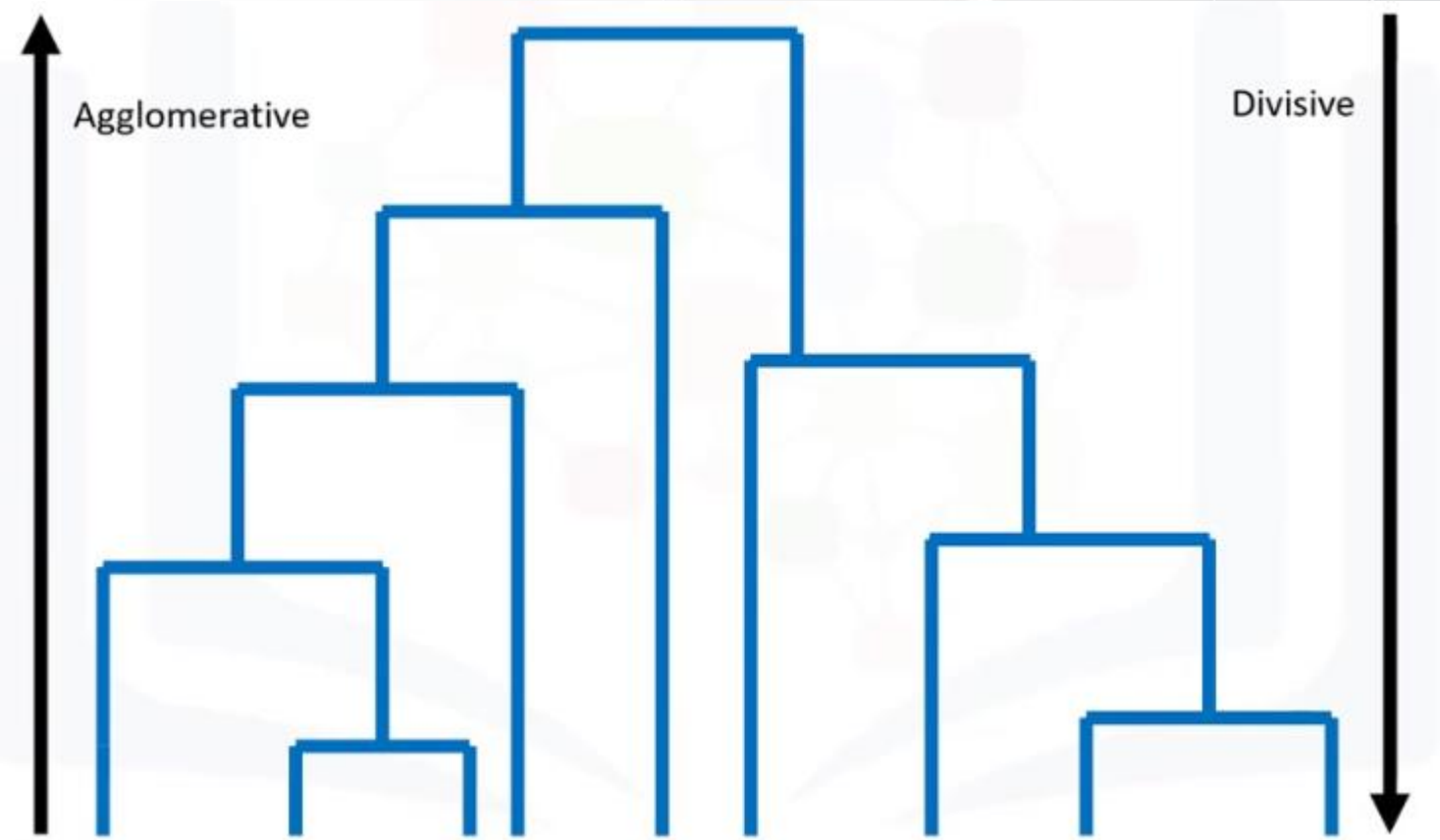
Hierarchical clustering

Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consists of the clusters of its daughter nodes.



Source: von Holdt B.M. et al. (2010)
Genome-wide SNP and haplotype analyses

Hierarchical clustering



Agglomerative Clustering

- This method build the hierarchy from the individual element by progressively merging cluster
- **dis** is a distance value between the city

j

i

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						

dis(i,j)

Agglomerative Clustering

- Find the first closest cluster, Montreal and Ottawa

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						

Agglomerative Clustering

- Then the Montreal and Ottawa are merged
- The table is constructed

	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					

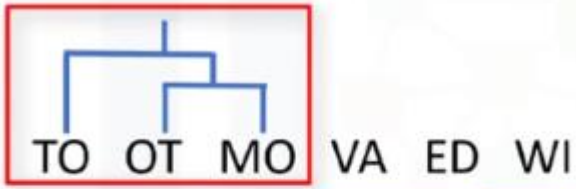
Agglomerative Clustering

- Find the closest distance from Montreal and Ottawa by calculate the distance each data to the mean of Montreal/Ottawa cluster

	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					

Agglomerative Clustering

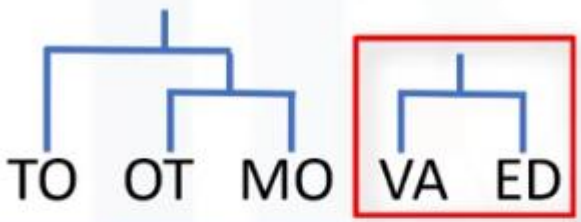
- The closest is Toronto
- Then connect Toronto to Montreal/Ottawa cluster
- This make another cluster



	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					

Agglomerative Clustering

- The Vancouver is closest to Edmonton
- Then created them as one hierarchy cluster



	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				

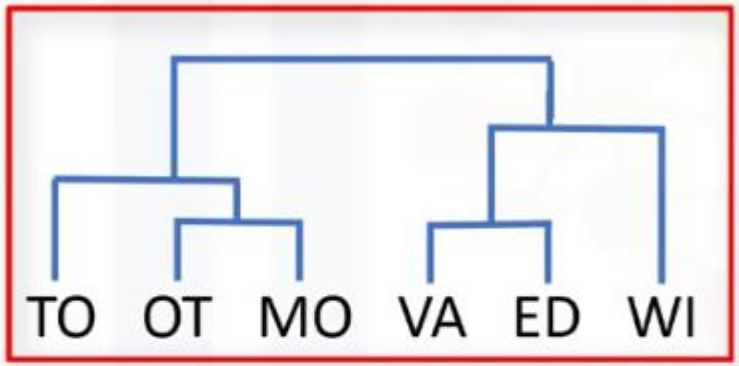
Agglomerative Clustering

- By same way agglomerative cluster build the hierarchy

	TO/OT/MO	VA/ED	WI
TO/OT/MO		2840	1676
VA/ED			1667
WI			

Hierarchical Clustering

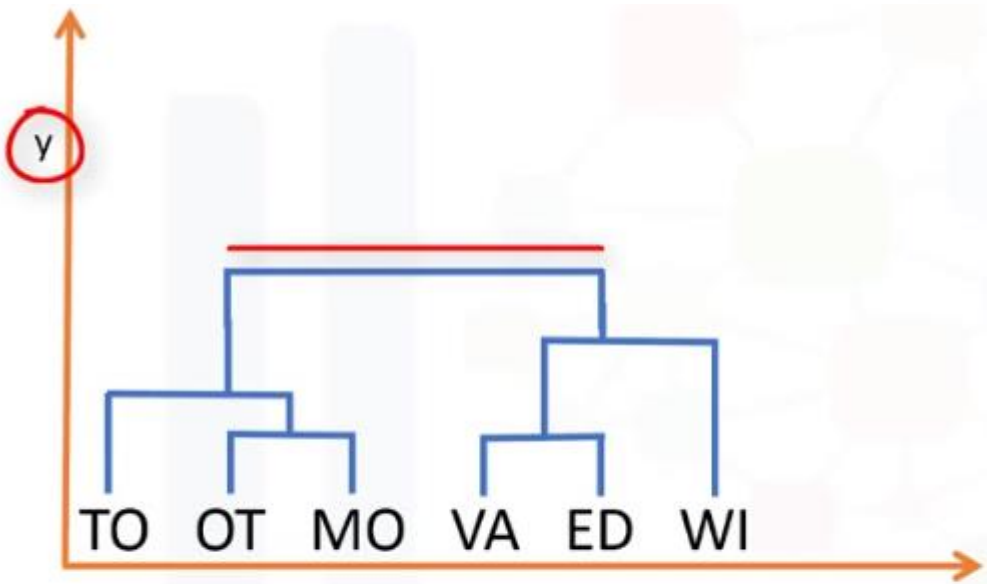
- The process is stopped when the single cluster is built
  - The cluster are totally merged
  - The tree are completed



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		

Hierarchical Clustering

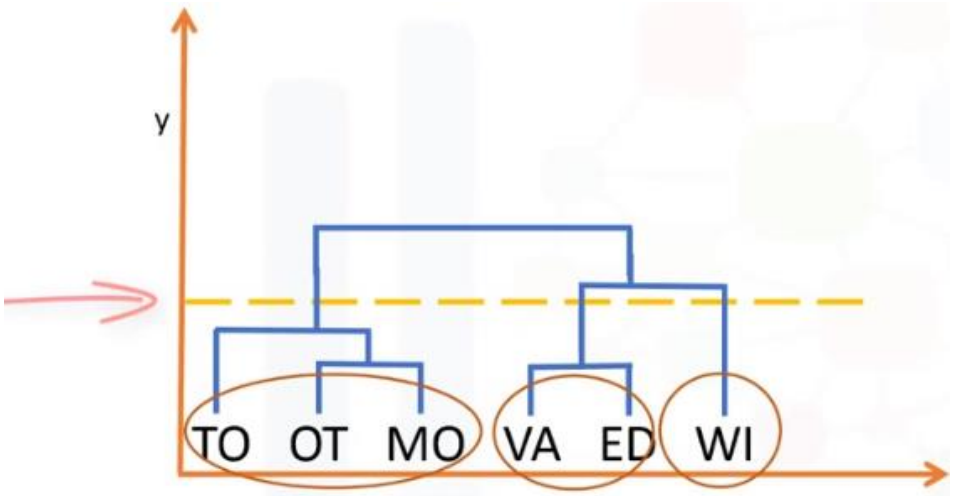
- The hierarchical clustering is described in Dendrogram
  - Each merge is represented by horizontal line
  - **y** show the similarity that two cluster that were merged
- Essentially the hierarchical clustering doesn't have number of class



Dendrogram

Hierarchical Clustering

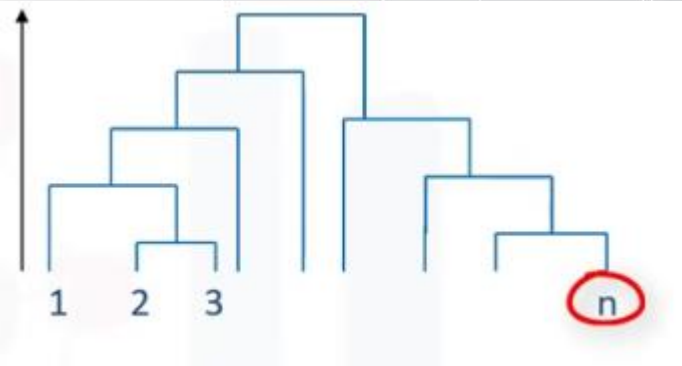
- So, how to solve the problem that require number of classes?
  - By disjoint the clusters using flag value
  - By some y value the hierarchy will be cut
- For example by cutting hierarchy by the value of similarity, we can create three cluster



Dendrogram

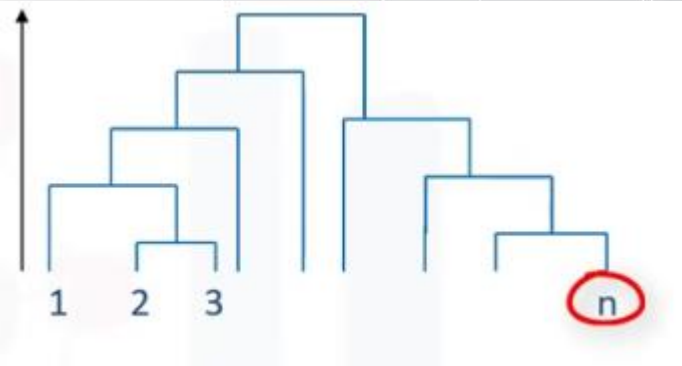
Agglomerative algorithm

1. Create n clusters, one for each data point



Agglomerative algorithm

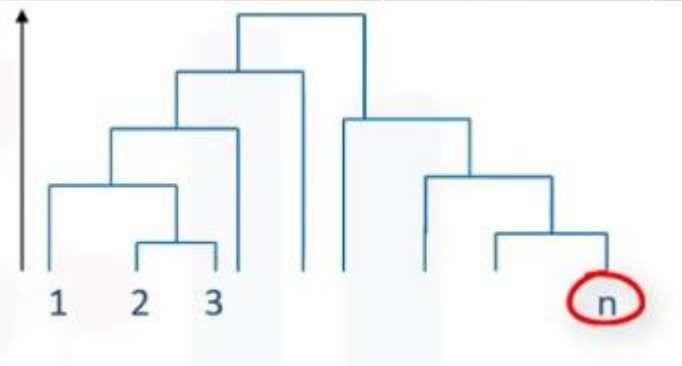
1. Create n clusters, one for each data point
2. Compute the Proximity Matrix



$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Agglomerative algorithm

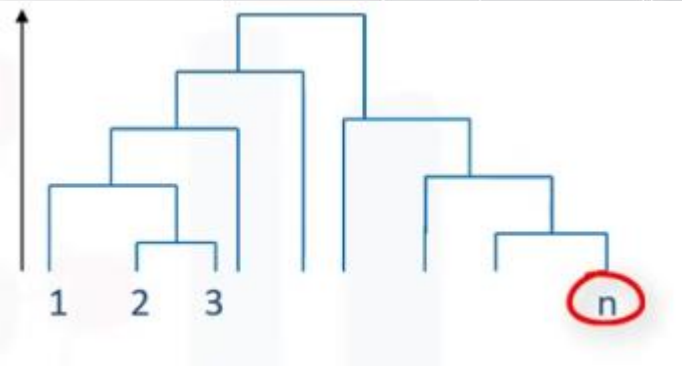
1. Create n clusters, one for each data point
2. Compute the Proximity Matrix
3. Repeat
   - Merge the two closest cluster
   - Update the proximity cluster



$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Agglomerative algorithm

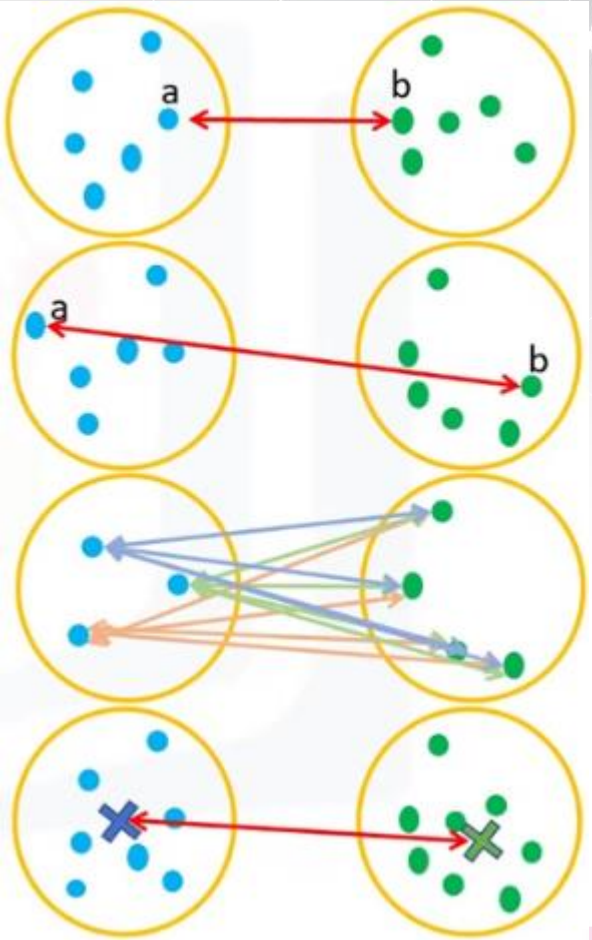
1. Create n clusters, one for each data point
2. Compute the Proximity Matrix
3. **Repeat**
   - Merge the two closest cluster
   - Update the proximity cluster
4. **Until** only a single cluster remains



$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Distance between clusters

- Single-Linkage Clustering
  - Minimum distance between clusters
- Complete-Linkage Clustering
  - Maximum distance between clusters
- Average Linkage Clustering
  - Average distance between clusters
- Centroid Linkage Clustering
  - Distance between cluster centroids



Advantages vs. disadvantages

Advantages	Disadvantages
Doesn't required number of clusters to be specified.	Can never undo any previous steps throughout the algorithm.
Easy to implement.	Generally has long runtimes.
Produces a dendrogram, which helps with understanding the data.	Sometimes difficult to identify the number of clusters by the dendrogram.

Hierarchical Clustering vs k-Means

K-means	Hierarchical Clustering
1. Much more efficient	1. Can be slow for large datasets
2. Requires the number of clusters to be specified	2. Does not require the number of clusters to run
3. Gives only one partitioning of the data based on the predefined number of clusters	3. Gives more than one partitioning depending on the resolution
4. Potentially returns different clusters each time it is run due to random initialization of centroids	4. Always generates the same clusters

IKUTI KAMI

DIGITAL TALENT SCHOLARSHIP

digitalent.kominfo
digitalent.kominfo
DTS_kominfo
Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110

digitalent.kominfo.go.id