



DIGITAL TALENT SCHOLARSHIP 2019



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

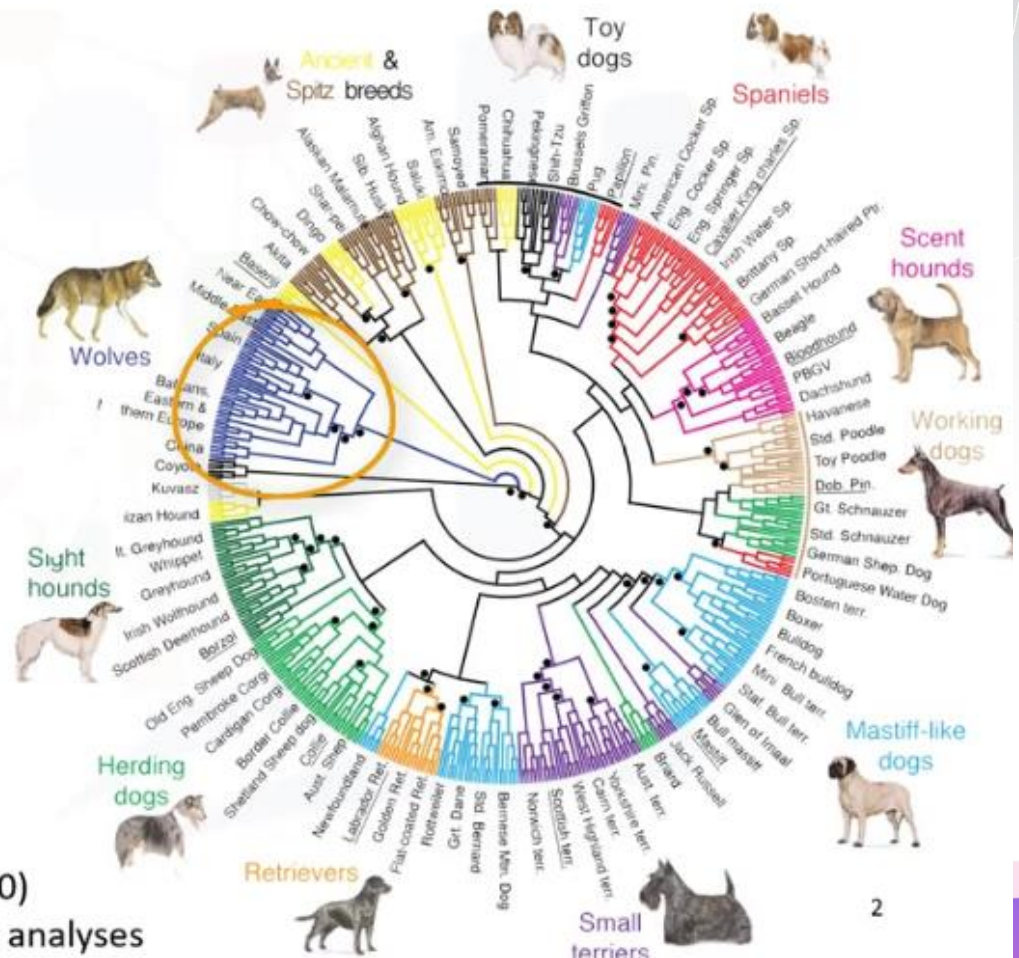
Clustering: Hierarchical

M. Ramli & M. Soleh



Hierarchical clustering

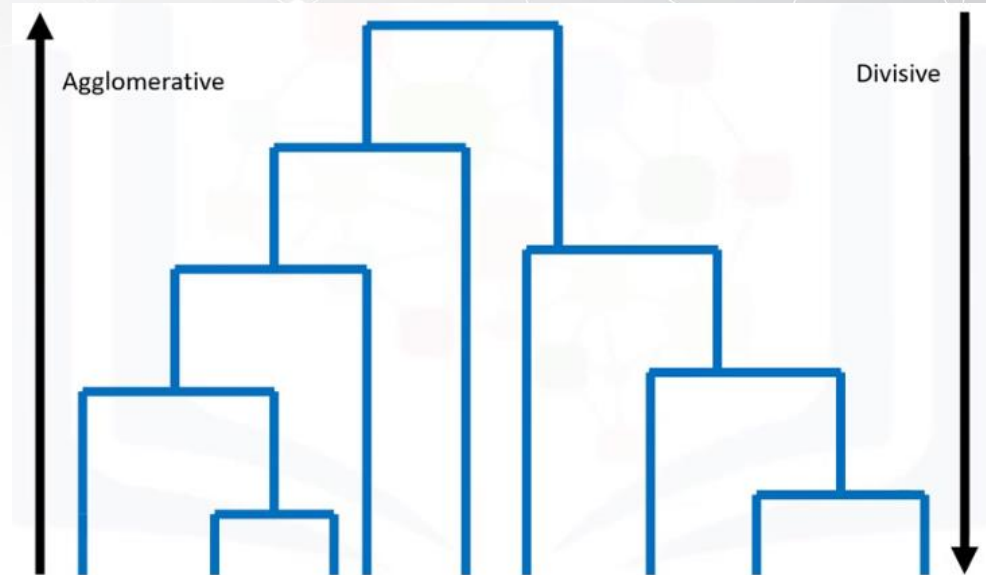
Algoritma **hierarchical clustering** membangun hierarki pada cluster, di mana setiap node adalah sebuah cluster yang terdiri dari cluster-cluster dibawahnya.



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

Hierarchical clustering

- **Aglomeratif Clustering** dikenal sebagai AGNES (Agglomerative Nesting). Algoritma dimulai dengan memperlakukan setiap objek sebagai sebuah cluster tunggal. Selanjutnya, pasangan cluster digabungkan secara berurutan sampai semua cluster telah bergabung menjadi satu cluster besar yang berisi semua objek. Hasilnya adalah representasi objek berbasis pohon, bernama dendrogram.
- hierarchical clustering yang memecah belah, juga dikenal sebagai DIANA (Divisive ANALysis) adalah kebalikan dari pengelompokan aglomeratif



Agglomerative Clustering

- Metode ini membangun hierarki dari elemen individual dengan secara progresif menggabungkan cluster
- **dis** adalah nilai jarak antara kota



	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

- Temukan cluster terdekat pertama, Montreal dan 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

- Kemudian Montreal dan Ottawa digabung
- Selanjutnya tabel dibangun kembali



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

Agglomerative Clustering

- Temukan jarak terdekat dari Montreal dan Ottawa dengan menghitung jarak setiap data dengan nilai rata-rata yang dimiliki Montreal/Ottawa cluster.

TO OT MO VA ED WI

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



Agglomerative Clustering

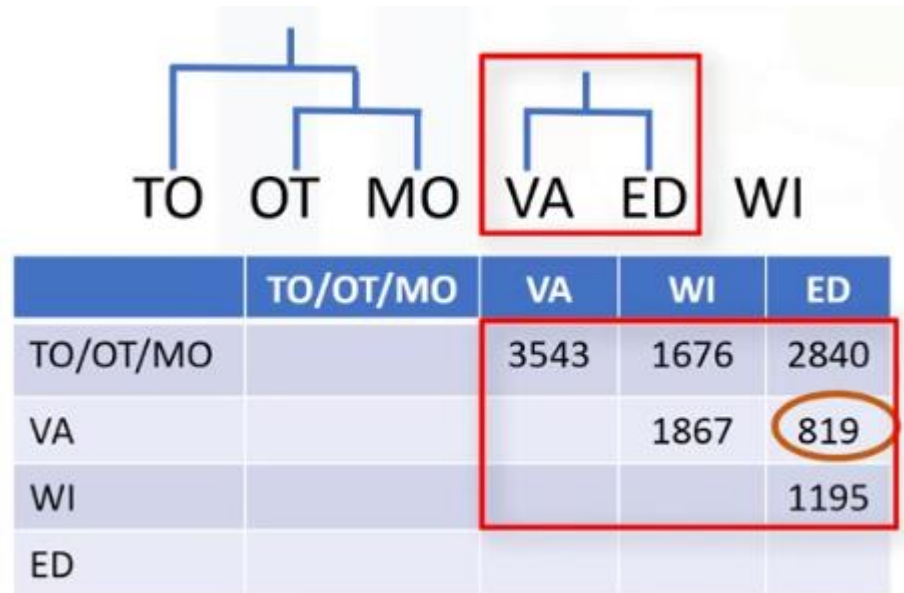
- Kota yang terdekat adalah Toronto
- Kemudian hubungkan Toronto ke Montreal / Ottawa cluster
- Hal ini membangun cluster lain



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

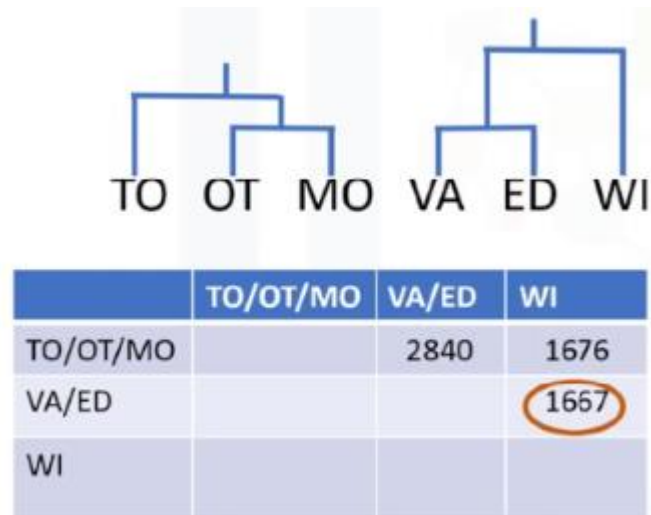
Agglomerative Clustering

- Vancouver paling dekat dengan Edmonton
- Kemudian 2 kota tersebut dibuat menjadi satu cluster hierarki



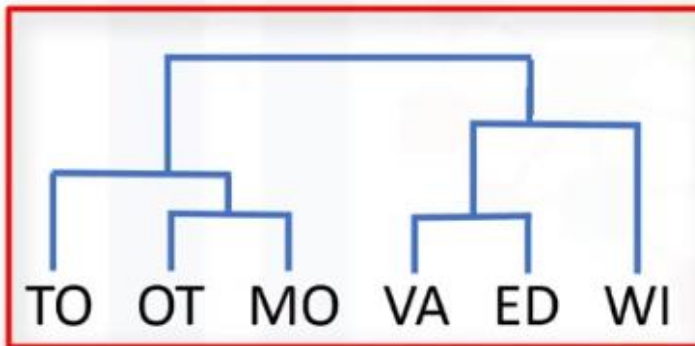
Agglomerative Clustering

- Dengan cara yang sama Agglomerative Clustering membangun hierarki



Hierarchical Clustering

- Proses dihentikan ketika satu cluster dibangun
 - Semua cluster digabung
 - Pada akhirnya pohon terbentuk

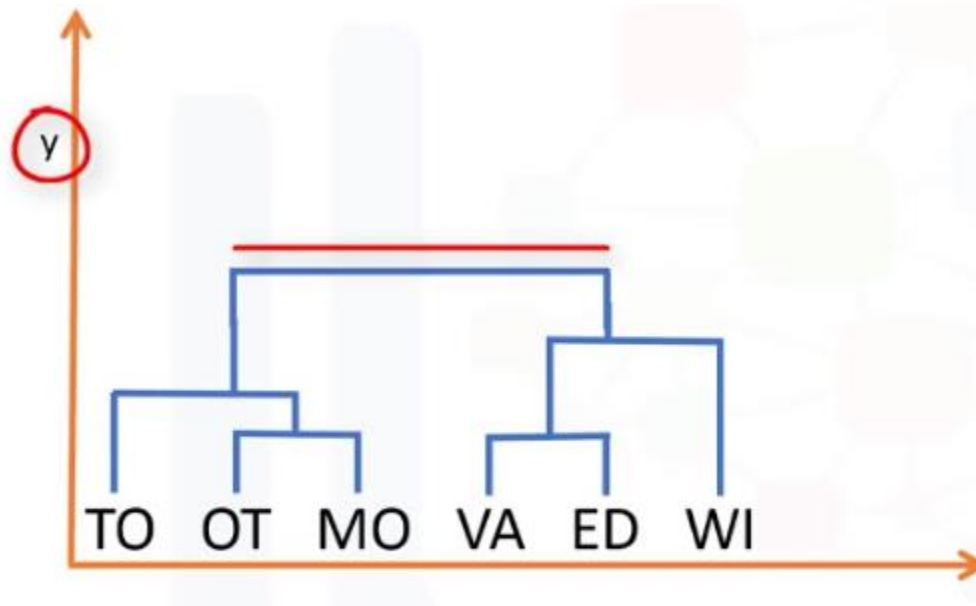


	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		



Hierarchical Clustering

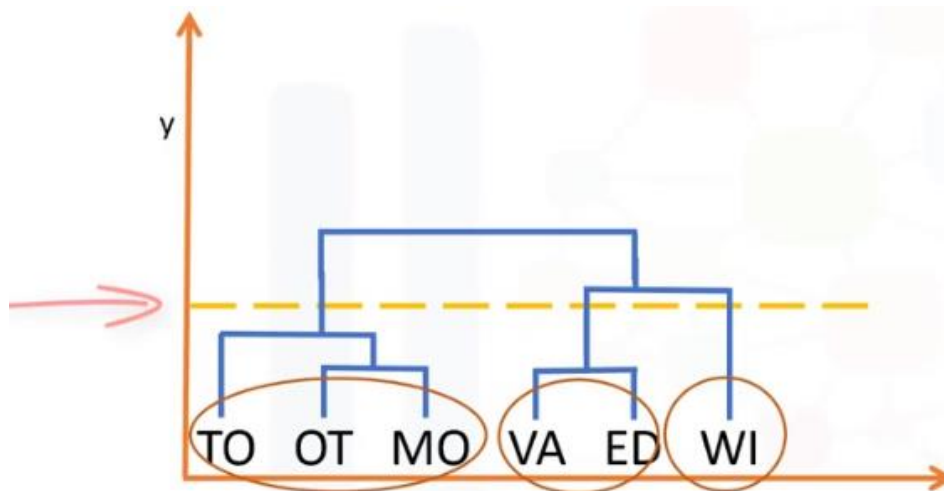
- Hierarchical clustering dijelaskan dalam Dendrogram
 - Setiap penggabungan diwakili oleh garis horizontal
 - y menunjukkan kesamaan bahwa dua cluster yang digabung
- Pada dasarnya hierarchical clustering tidak memiliki jumlah kelas



Dendrogram

Hierarchical Clustering

- Lantas, bagaimana cara mengatasi masalah yang membutuhkan sejumlah kelas?
 - Dengan memisahkan cluster dengan menggunakan nilai flag
 - Dengan sebuah nilai y hierarki dapat dipotong
- Misalnya dengan memotong hierarki berdasarkan nilai similaritas, sehingga kita dapat membuat tiga kluster



Dendrogram

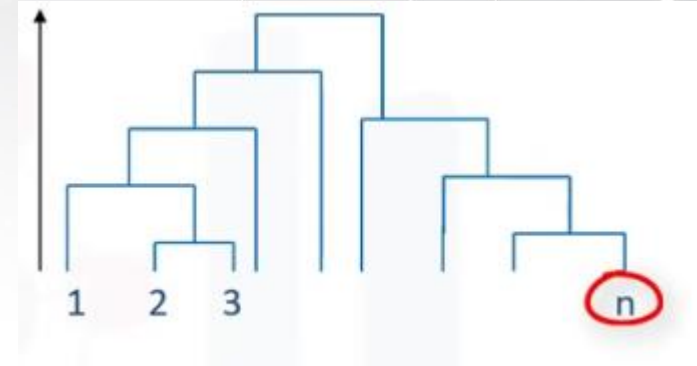


archical Clustering



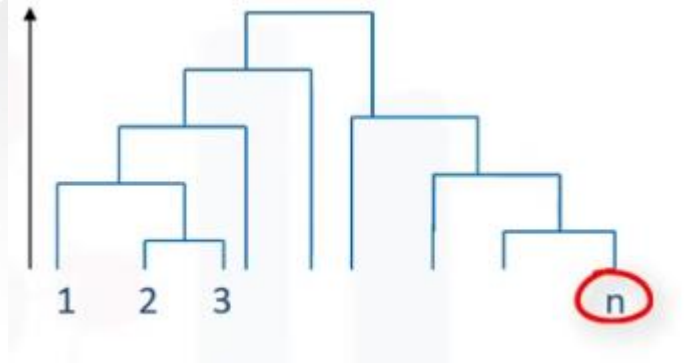
Algoritma Agglomerative

1. Beri label cluster n untuk setiap data



Algoritma Agglomerative

1. Buat n cluster, satu untuk setiap titik data
2. Hitung Proximity Matrix

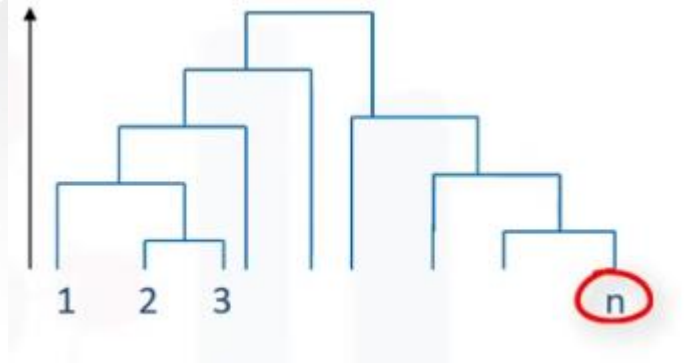


$$\begin{bmatrix}
 0 & & & & \\
 d(2,1) & 0 & & & \\
 d(3,1) & d(3,2) & 0 & & \\
 \vdots & \vdots & \vdots & & \\
 d(n,1) & d(n,2) & \dots & \dots & 0
 \end{bmatrix}$$



Algoritma Agglomerative

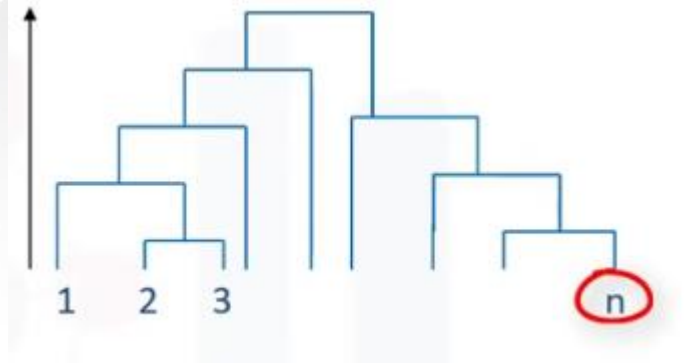
1. Buat n cluster, satu untuk setiap titik data
2. Hitung Proximity Matrix
3. **Ulangi**
 - a. Gabungkan dua kluster terdekat
 - b. Perbarui nilai proximity cluster



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

Agglomerative algorithm

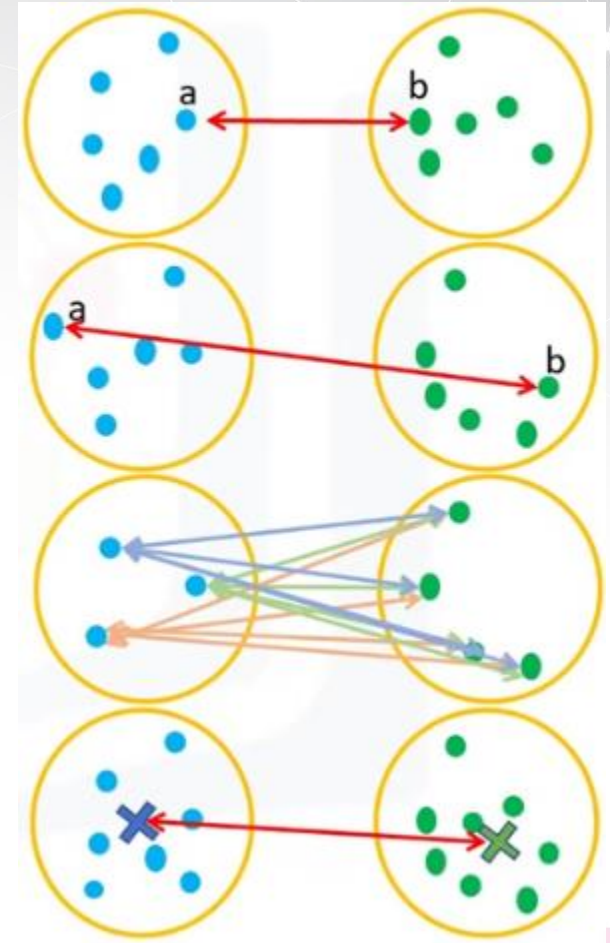
1. Buat n cluster, satu untuk setiap titik data
2. Hitung Proximity Matrix
3. **Ulangi**
 - a. Gabungkan dua kluster terdekat
 - b. Perbarui nilai proximity cluster
4. **Sampai** hanya satu cluster yang tersisa



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

Jarak antar cluster

- **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
Tidak membutuhkan nilai inisial untuk jumlah cluster	Tidak akan pernah bisa mengulangi step sebelumnya melalui algoritma yang ada
Mudah untuk diimplementasikan	Secara umum membutuhkan waktu proses yang lama
Menghasilkan Dendogram, yang dapat membantu dalam memahami data	Satu waktu dapat terjadi kesulitan untuk mengidentifikasi jumlah cluster berdasarkan Dendogram.

Hierarchical Clustering vs k-Means

K-means	Hierarchical Clustering
Lebih efisien	Dapat lebih lama untuk data yang cukup besar
Membutuhkan inisialisasi jumlah cluster	Tidak membutuhkan inisialisasi jumlah cluster
Hanya memberikan satu solusi cluster	Memberikan lebih dari satu solusi cluster tergantung pada solusi yang diinginkan
Memiliki potensial adanya perubahan cluster, jika inisial centroid diberikan secara random	Selalu menghasilkan satu jenis cluster

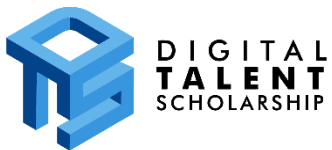






Bagian 2

Praktikum Lab

ML0101EN-Clus-Hierarchical-Cars-py-v1.ipynb

IKUTI KAMI



-  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

