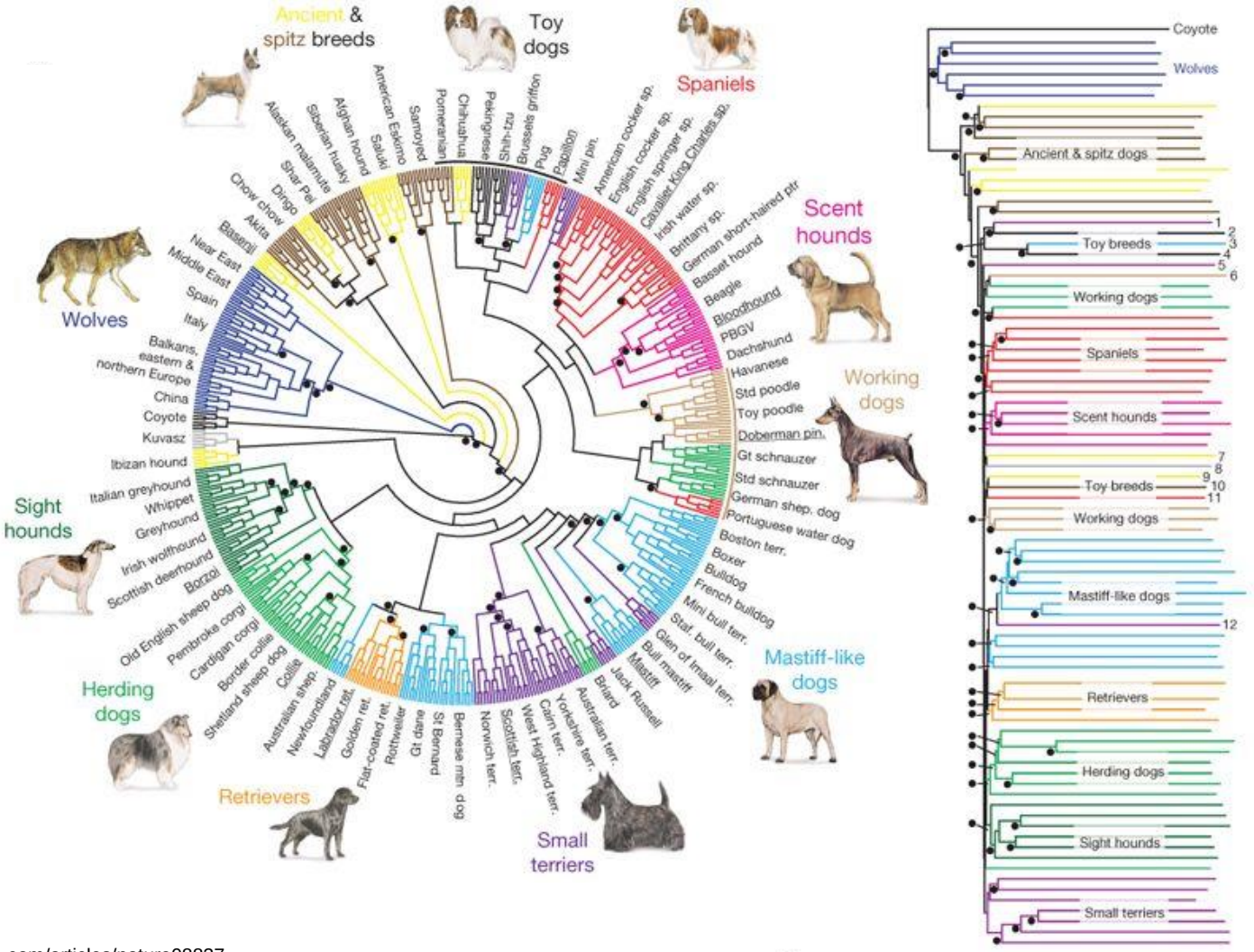
MIN601

Fundamentos de Aprendizagem de Máquina

Prof. Anderson Harayashiki Moreira

Aula 11: Clustering Hierárquico

O que é Clustering Hierárquico?

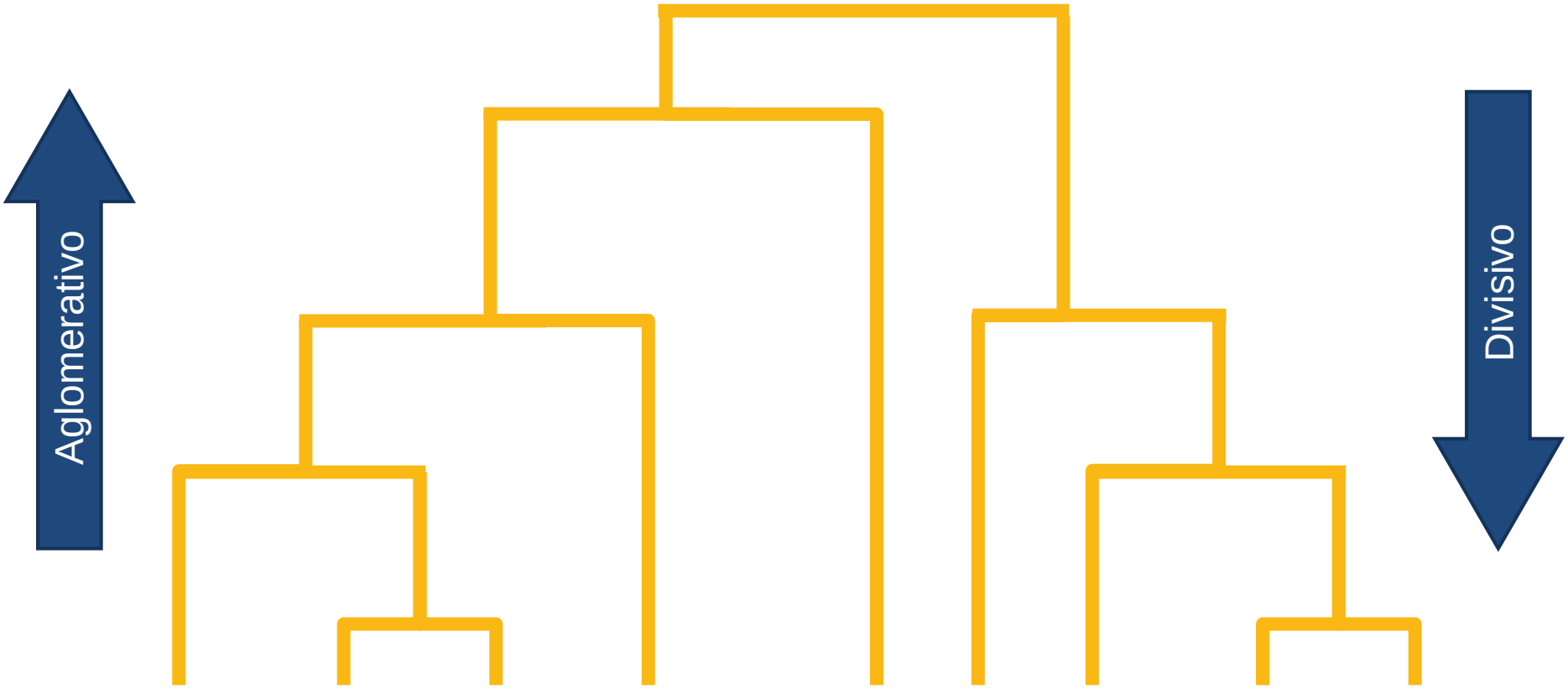


Source: https://www.nature.com/articles/nature08837

O que é Clustering Hierárquico?

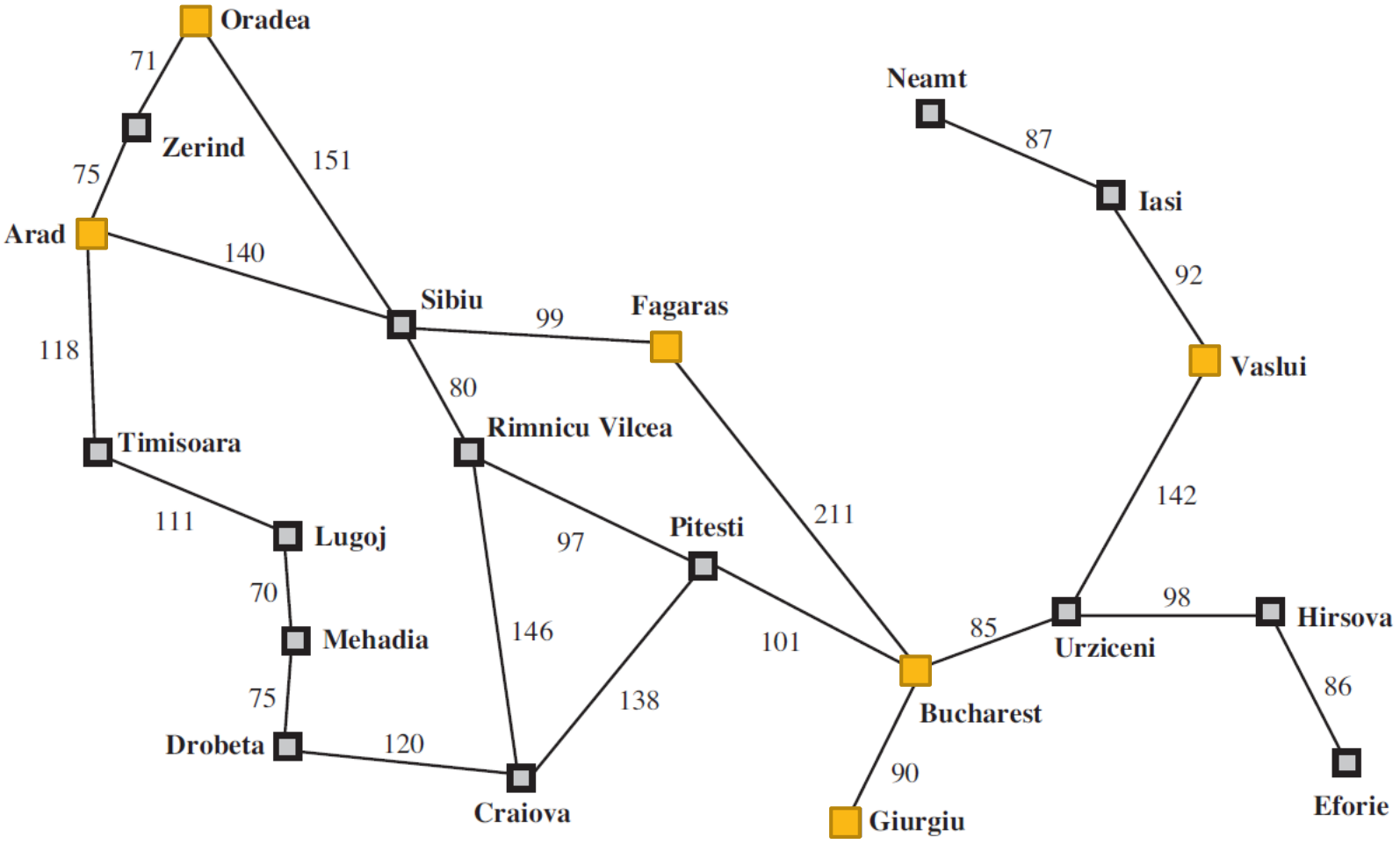
- Algoritmos de **Clustering Hierárquico** constroem uma hierarquia de clusters onde cada nó é um cluster que consiste nos clusters de seus nós-filhos.
- Existem dois tipos de Clustering Hierárquico:
  - Aglomerativo (Agglomerative)
  - Divisivo (Divisive)

O que é Clustering Hierárquico?



Clustering Aglomerativo

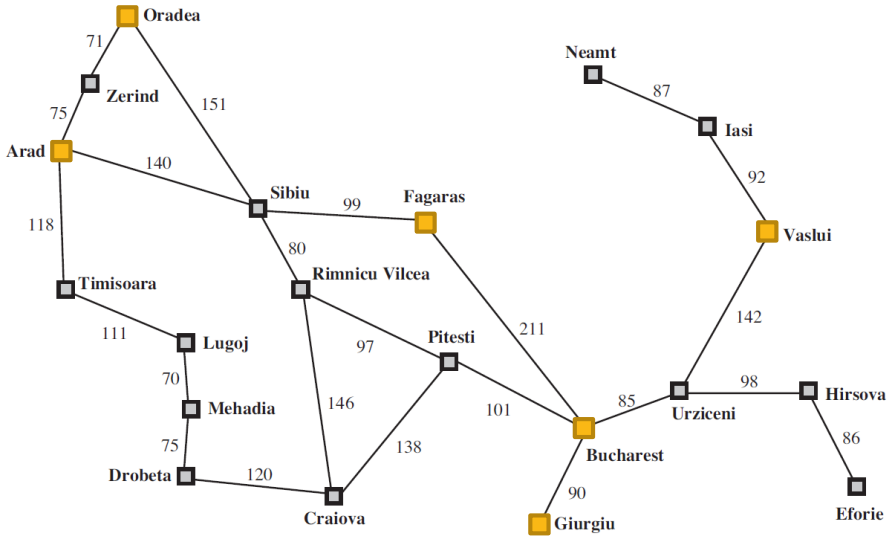
Clustering Aglomerativo



Clustering Aglomerativo

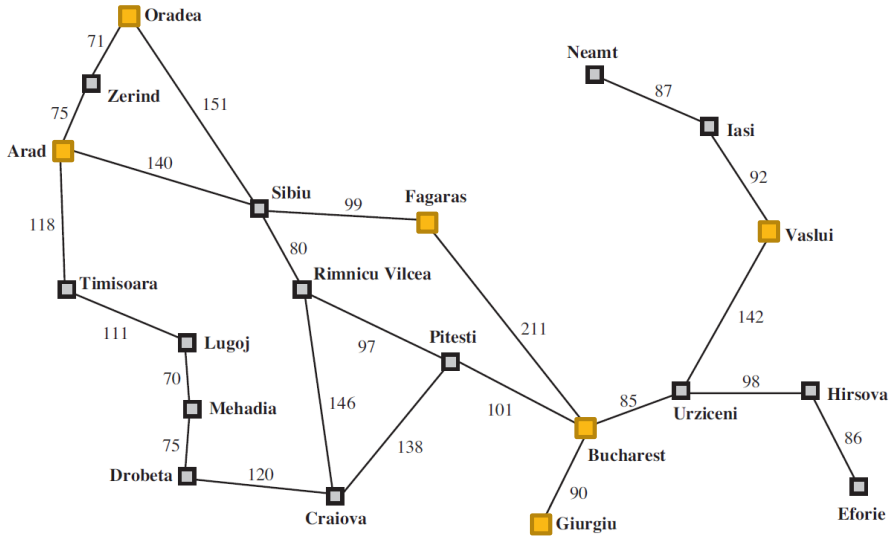
	A	B	F	G	O	V
A	0	418	239	508	146	645
B	418	0	211	90	429	227
F	239	211	0	301	250	438
G	508	90	301	0	519	317
O	146	429	250	519	0	656
V	645	227	438	317	656	0

Clustering Aglomerativo



	A	B	F	G	O	V
A	0	418	239	508	146	645
B	418	0	211	90	429	227
F	239	211	0	301	250	438
G	508	90	301	0	519	317
O	146	429	250	519	0	656
V	645	227	438	317	656	0

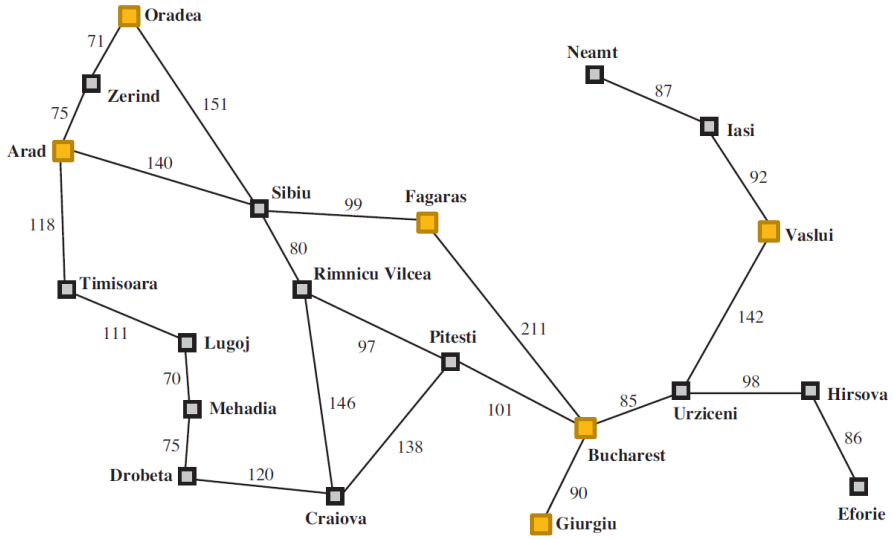
Clustering Aglomerativo



F A O V B G

	A	B	F	G	O	V
A	0	418	239	508	146	645
B	418	0	211	90	429	227
F	239	211	0	301	250	438
G	508	90	301	0	519	317
O	146	429	250	519	0	656
V	645	227	438	317	656	0

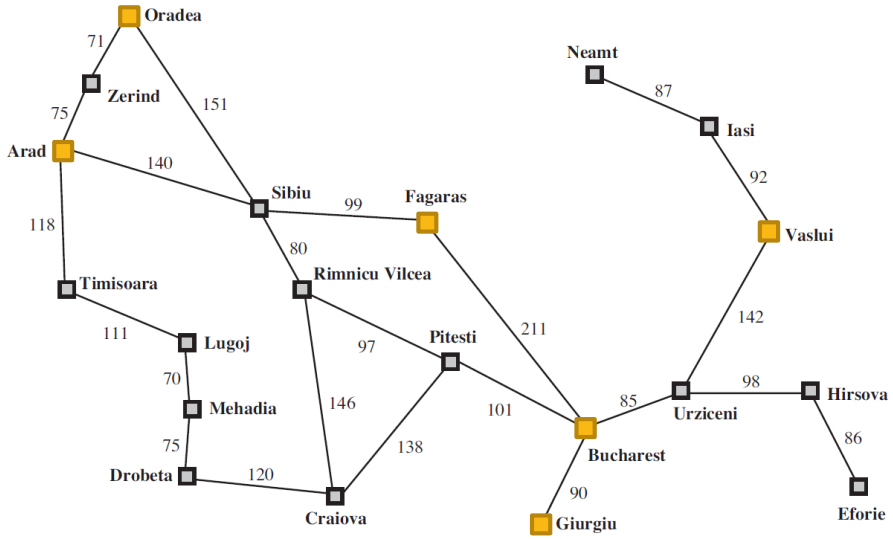
Clustering Aglomerativo



F A O V B G

	A	B	F	G	O	V
A	0	418	239	508	146	645
B		0	211	90	429	227
F			0	301	250	438
G				0	519	317
O					0	656
V						0

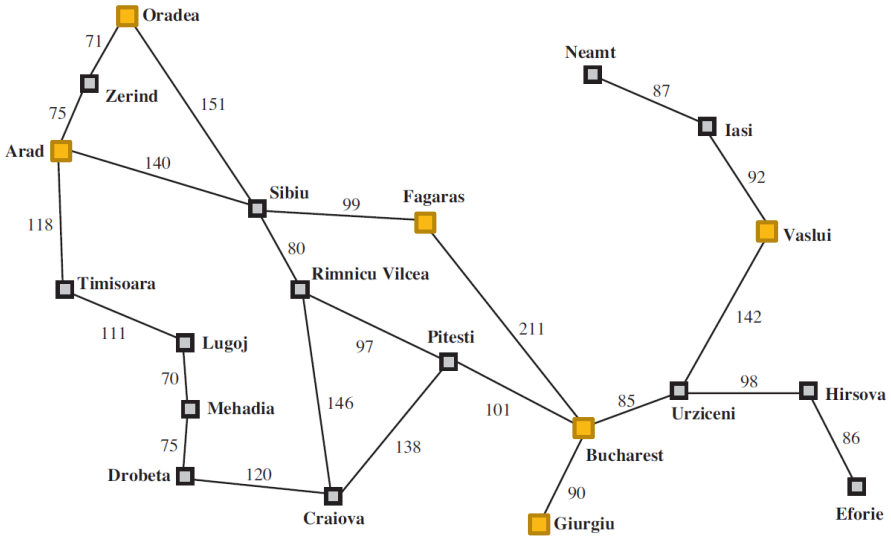
Clustering Aglomerativo



F A O V B G

	A	B	F	G	O	V
A	0	418	239	508	146	645
B		0	211	90	429	227
F			0	301	250	438
G				0	519	317
O					0	656
V						0

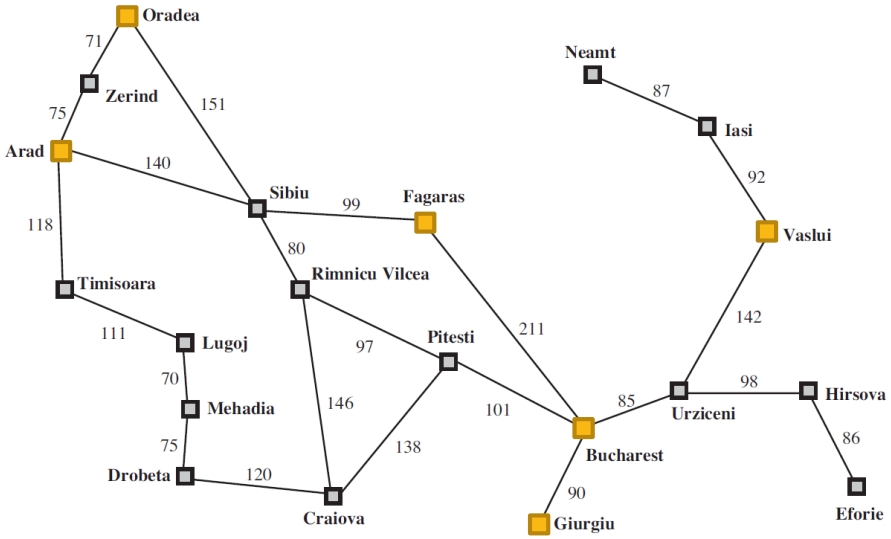
Clustering Aglomerativo



F A O V B G

	A	B	F	G	O	V
A	0	418	239	508	146	645
B		0	211	90	429	227
F			0	301	250	438
G				0	519	317
O					0	656
V						0

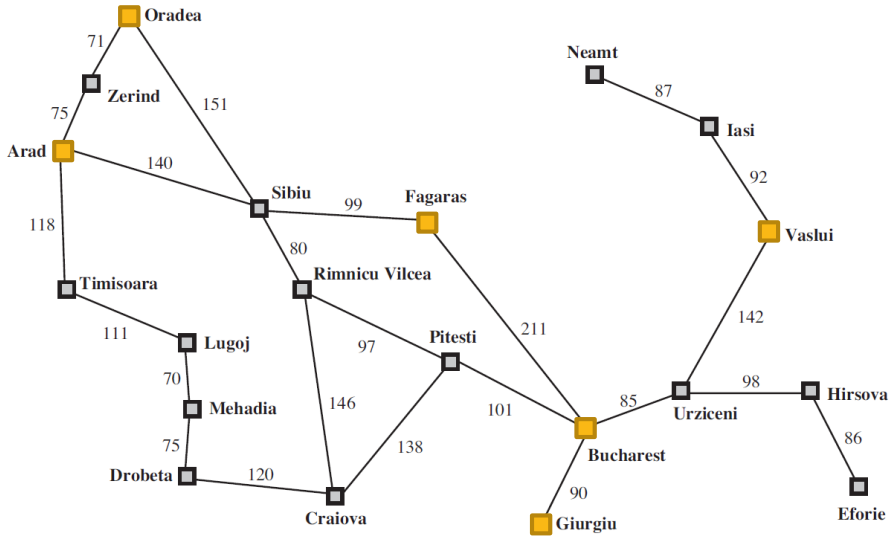
Clustering Aglomerativo



F A O V B G

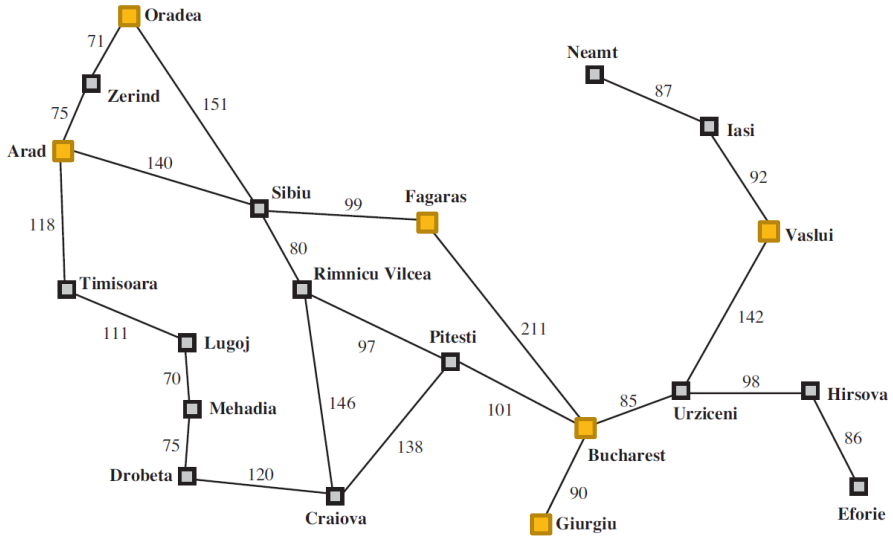
	A	F	B / G	O	V
A	0	239	463	146	645
F		0	256	250	438
B / G			0	474	272
O				0	656
V					0

Clustering Aglomerativo



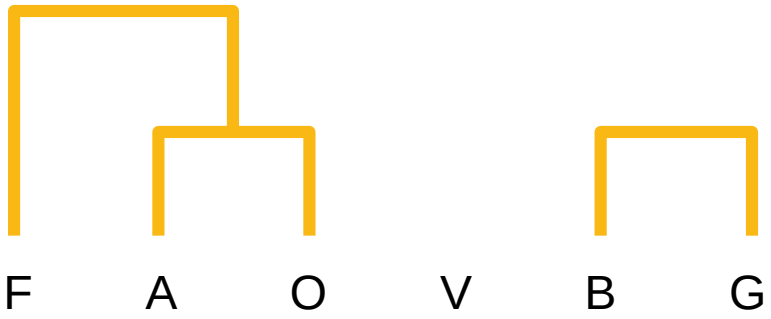
	A	F	B / G	O	V
A	0	239	463	146	645
F		0	256	250	438
B / G			0	474	272
O				0	656
V					0

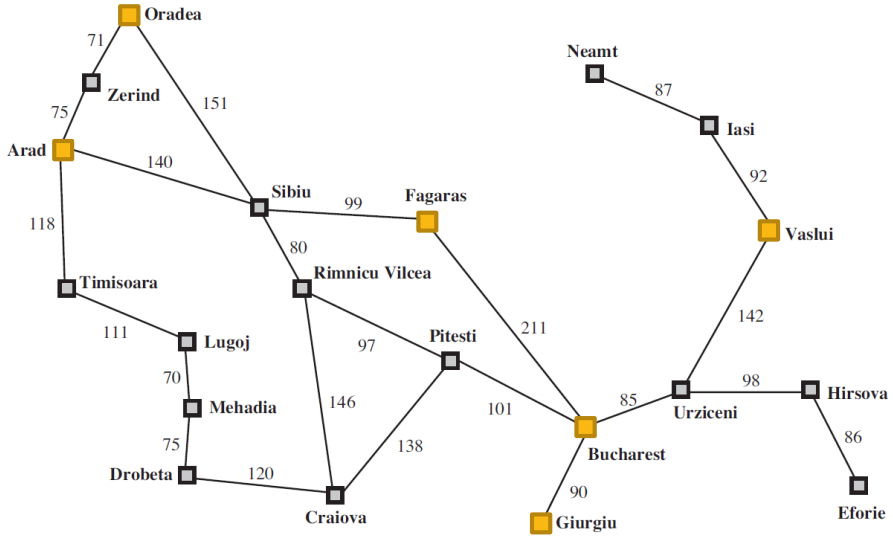
Clustering Aglomerativo



	A / O	F	B / G	V
A / O	0	166	390	572
F		0	256	438
B / G			0	272
V				0

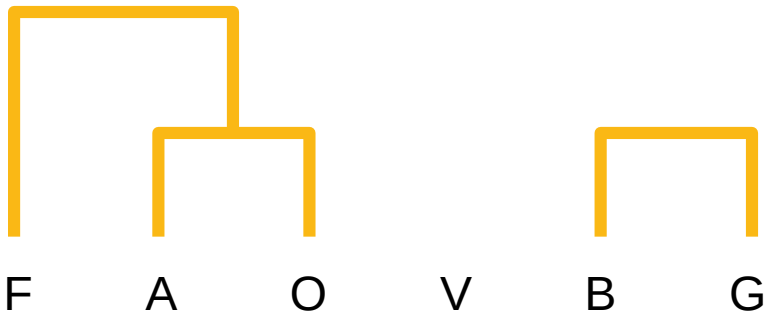
Clustering Aglomerativo

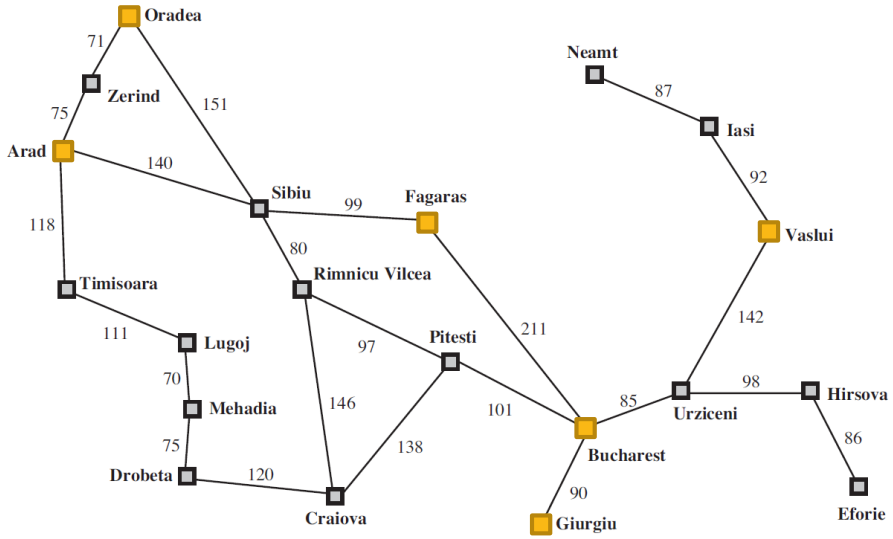




	A / O	F	B / G	V
A / O	0	166	390	572
F		0	256	438
B / G			0	272
V				0

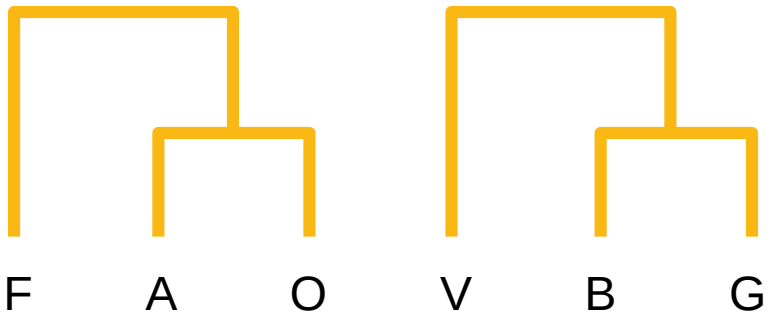
Clustering Aglomerativo

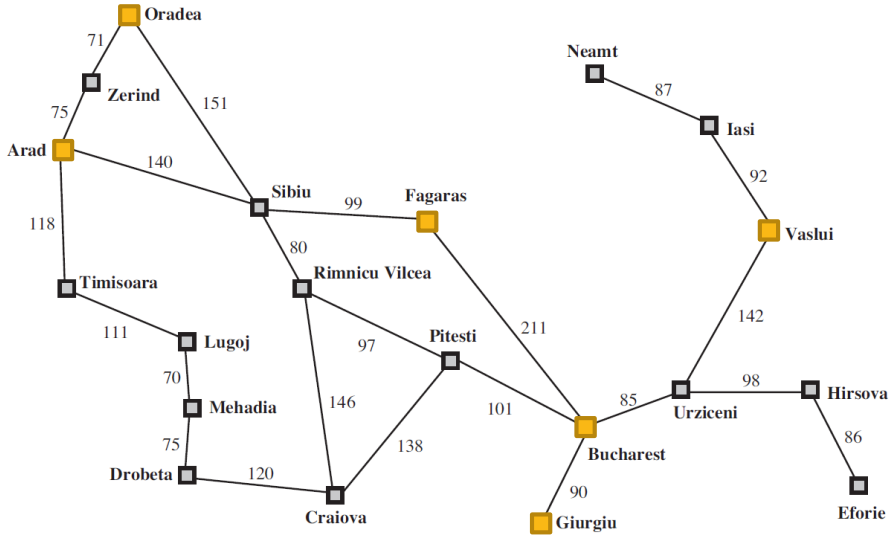




	A / O / F	B / G	V
A / O / F	0	323	505
B / G		0	272
V			0

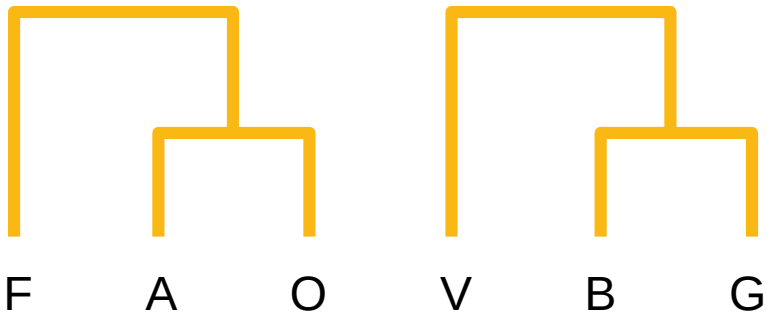
Clustering Aglomerativo

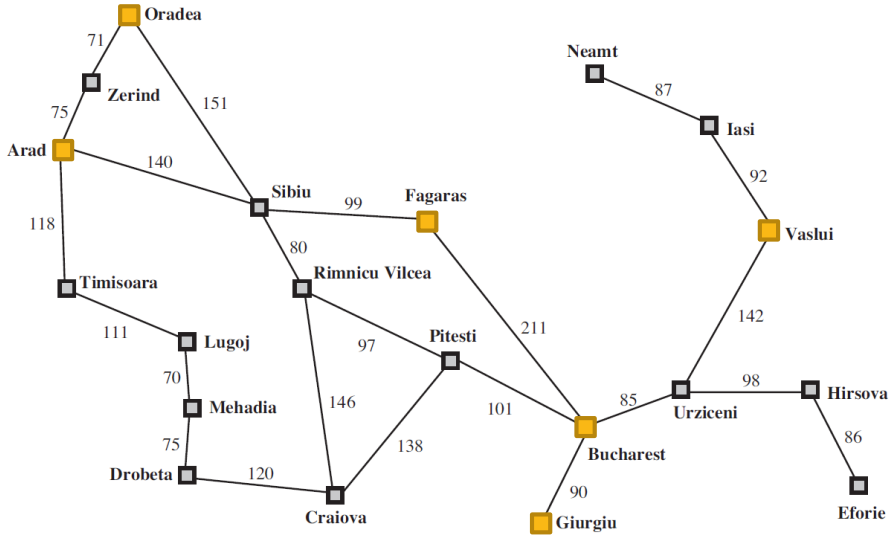




	A / O / F	B / G	V
A / O / F	0	323	505
B / G		0	272
V			0

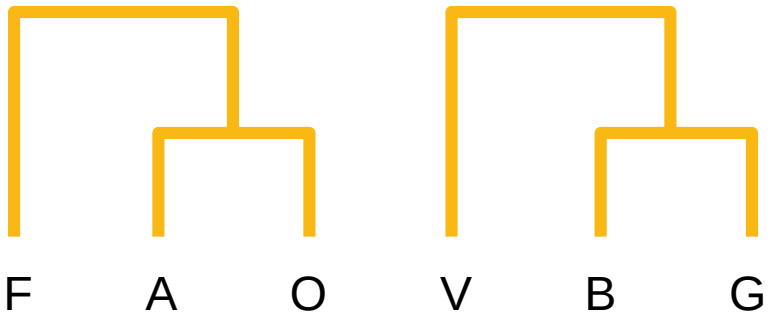
Clustering Aglomerativo

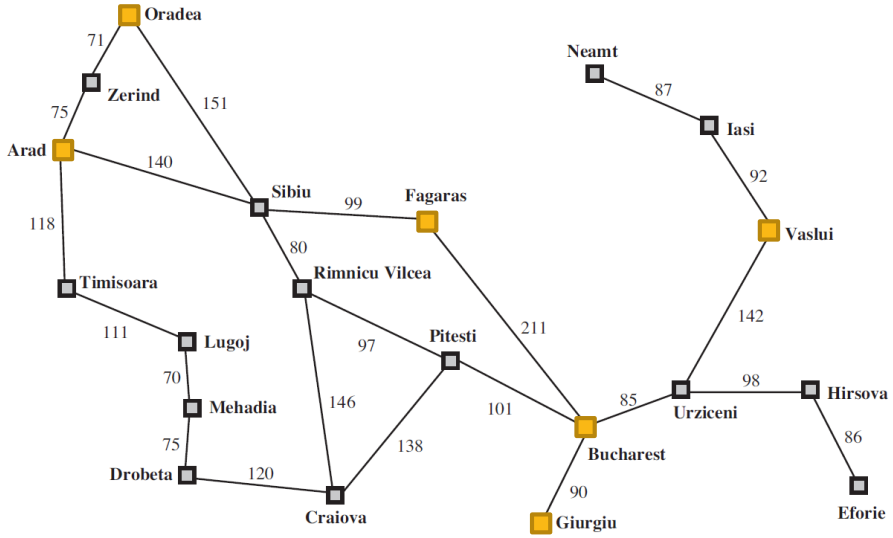




	A / O / F	B / G / V
A / O / F	0	323
B / G		0

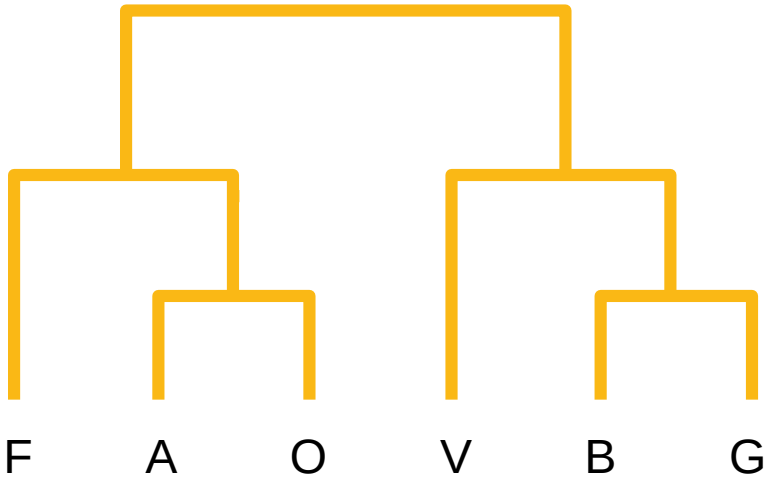
Clustering Aglomerativo

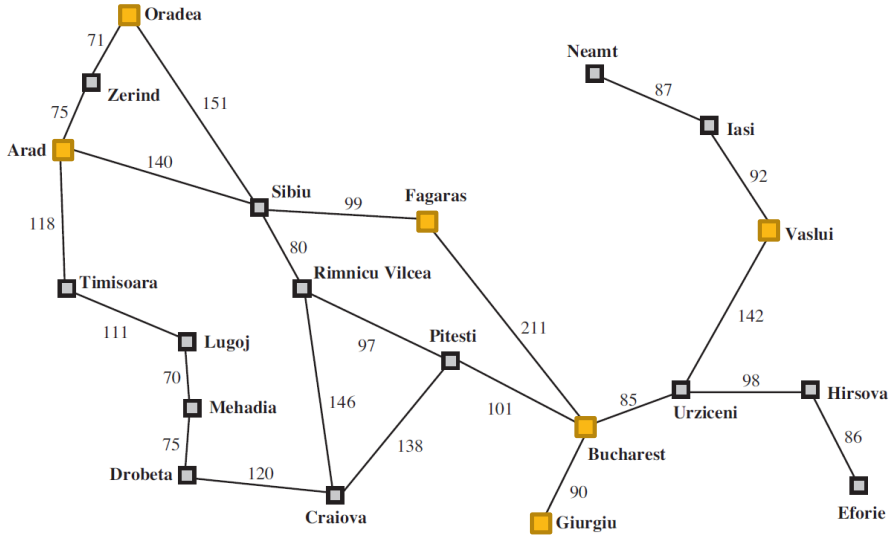




	A / O / F	B / G / V
A / O / F	0	323
B / G		0

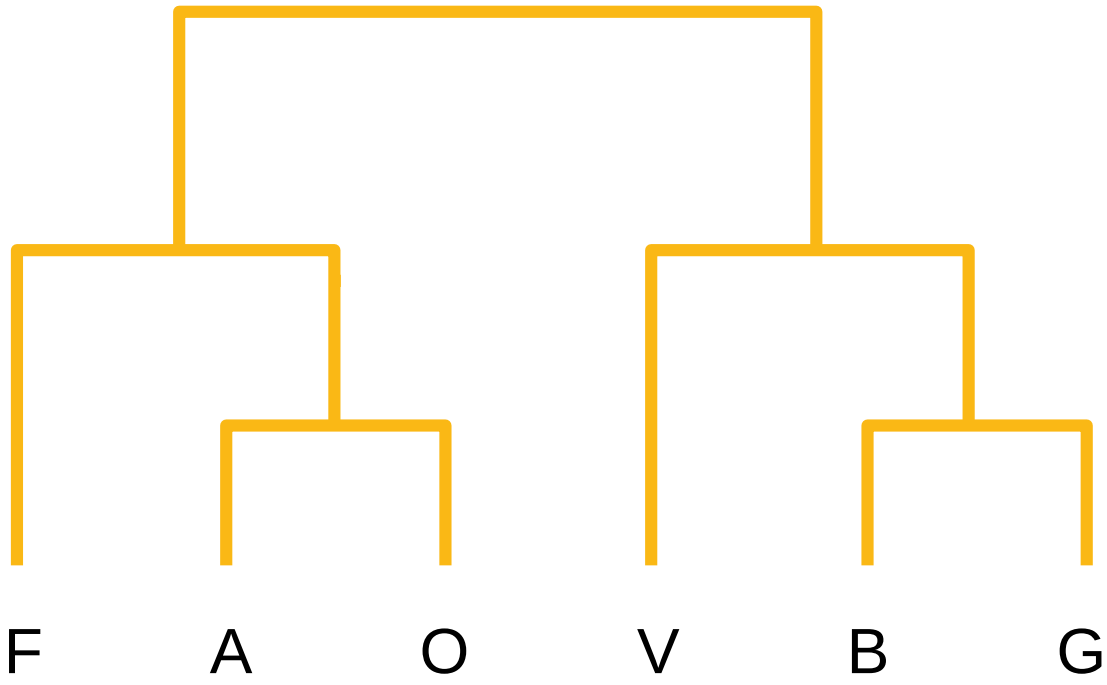
Clustering Aglomerativo



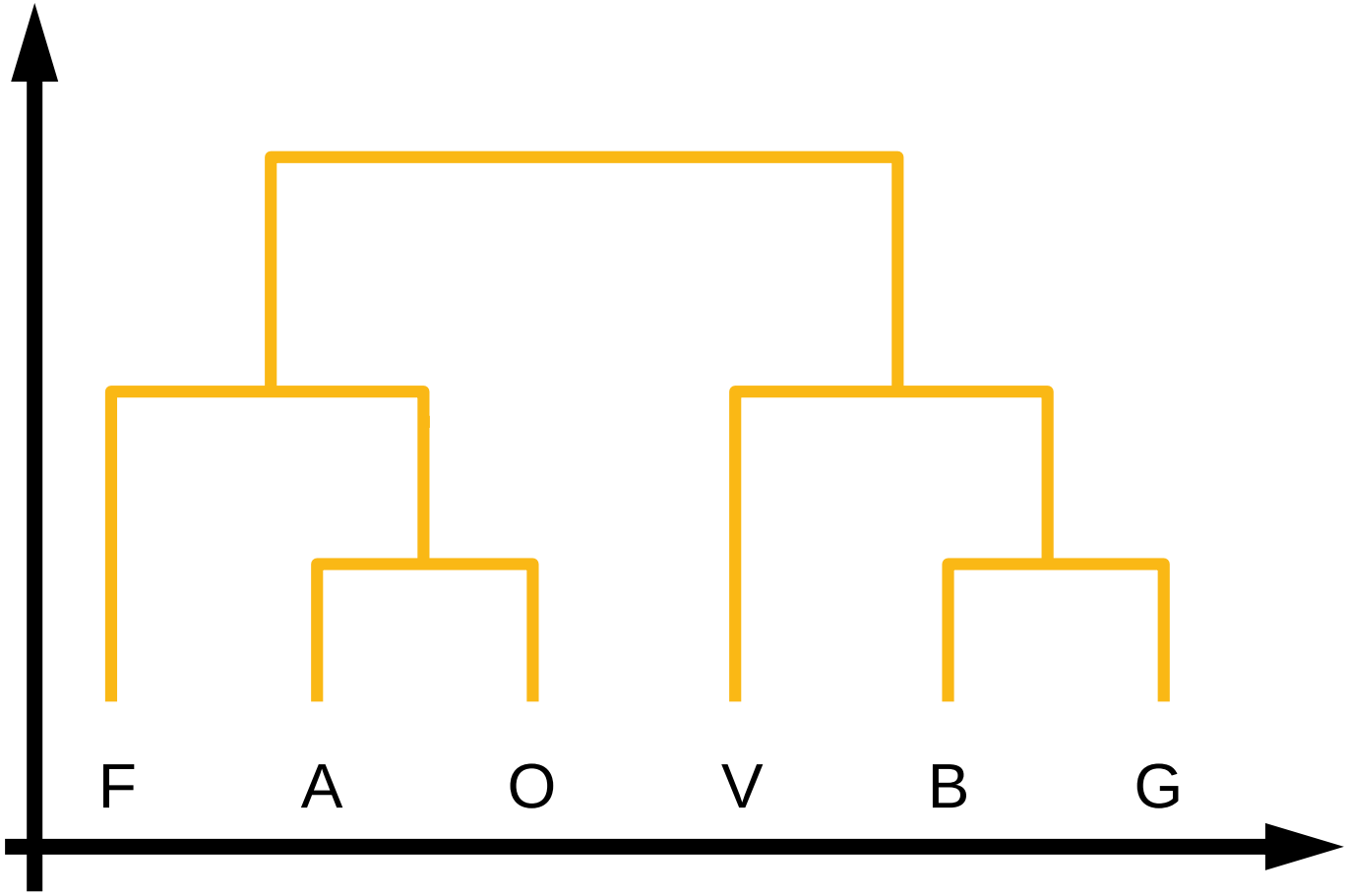


	A / O / F	B / G / V
A / O / F	0	323
B / G		0

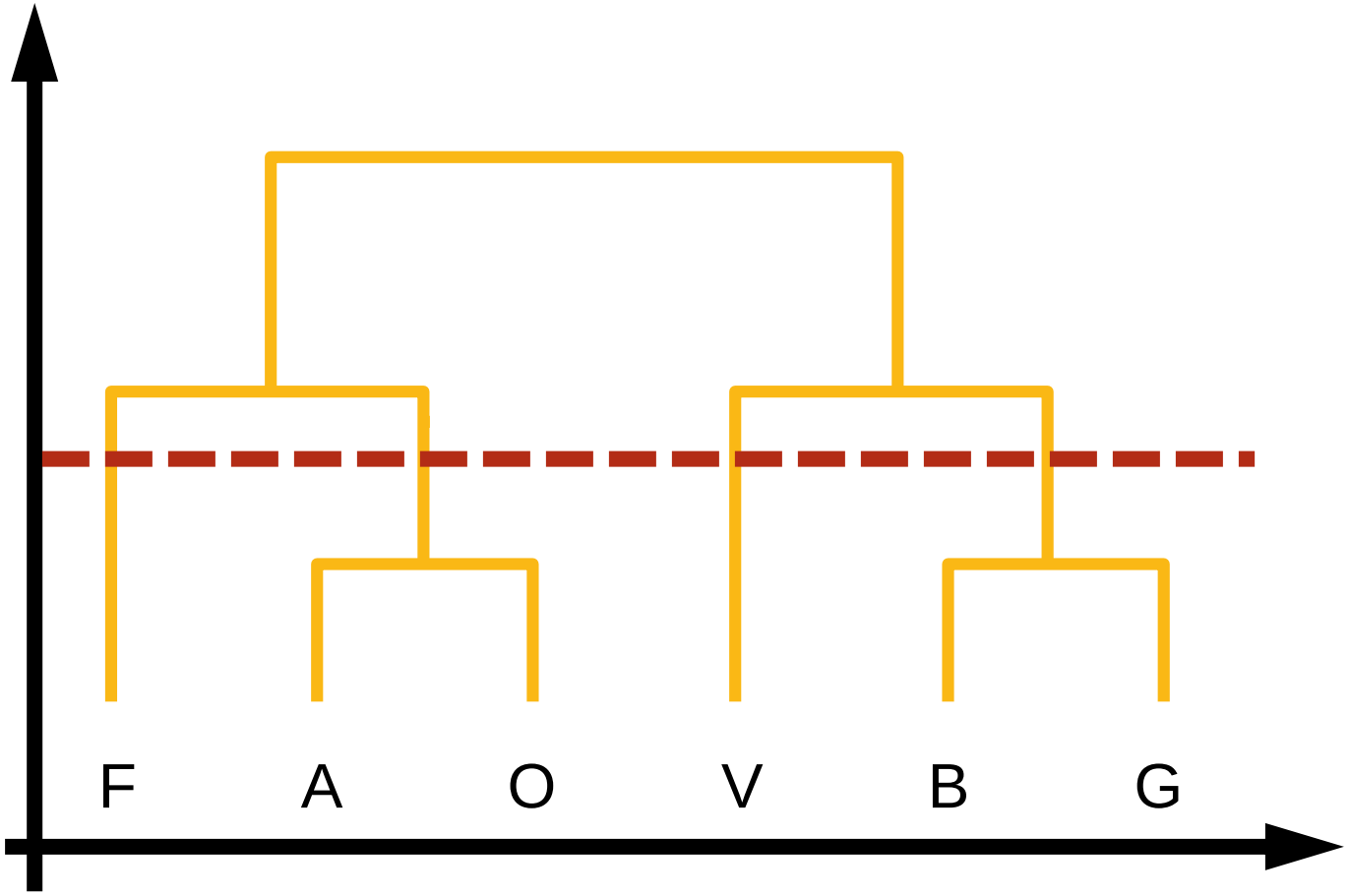
Dendograma



Dendograma



Dendograma

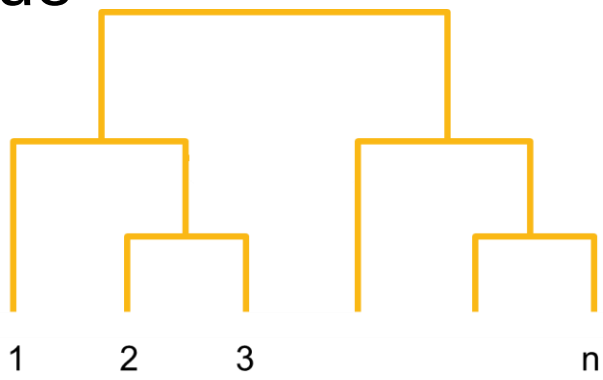


Algoritmo do Clustering Aglomerativo

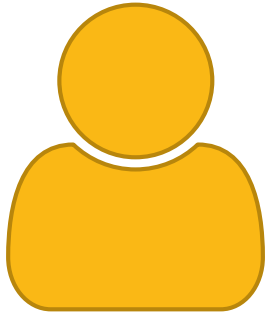
1. Criar n clusters, um para cada dado
2. Calcular a Matriz de Proximidade

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

3. **Repetir**
    1. Combinar os dois clusters mais próximos
    2. Atualizar a Matriz de Proximidade
4. **Até** sobrar um único cluster



Como calcular a distância?



Idade 50

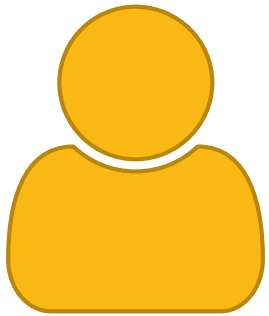
Idade 46

$$Dist(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

$$Dist(x_1, x_2) = \sqrt{(50 - 46)^2} = 4$$

Como calcular a distância?

Idade	50
Salário	166



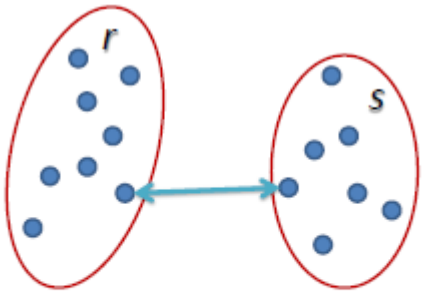
Idade	46
Salário	86

$$Dist(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

$$Dist(x_1, x_2) = \sqrt{(50 - 46)^2 + (166 - 86)^2} = 80{,}1$$

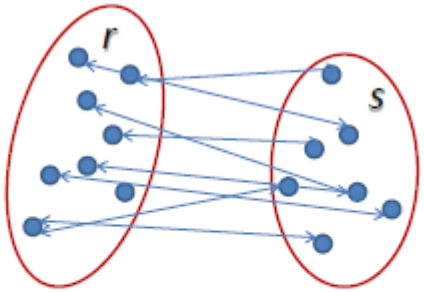
Como calcular a distância entre os clusters?

Single-Linkage Clustering



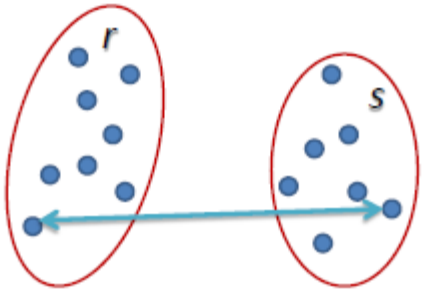
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete-Linkage Clustering



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Average-Linkage Clustering



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Prós e Contras

Vantagens

- Não precisa especificar a quantidade de clusters
- Fácil de Implementar
- Gera o dendograma que ajuda a entender os dados

Desvantagens

- Não pode voltar para um passo anterior do algoritmo
- Geralmente demora para ser executado
- Algumas vezes é difícil identificar o número de clusters no dendograma

Clustering Hierárquico vs K-Means

K-Means

- Mais eficiente
- Necessita especificar a quantidade de clusters
- Fornece apenas um particionamento dos dados baseado no número de clusters
- Potencialmente retorna clusters diferentes toda vez que se roda o algoritmo, devido a aleatoriedade da inicialização dos centroides

Desvantagens

- Pode ser lento para grandes conjuntos de dados
- Não necessita especificar a quantidade de clusters
- Fornece mais de um particionamento dos dados dependendo da resolução
- Sempre gera os mesmos clusters