Curso 2 – CD, AM e DM

Mineração de Dados

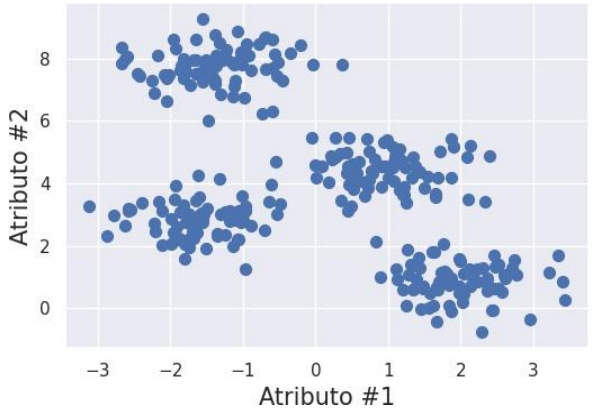
Parte 3
Extração de Padrões
Agrupamento Hierárquico

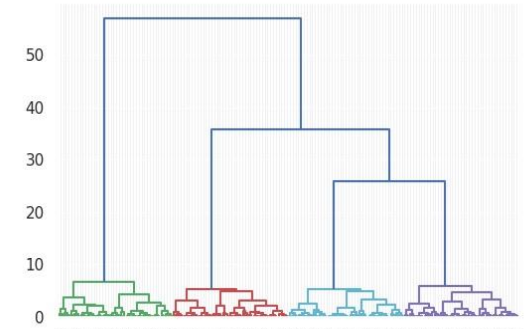
Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

Métodos para Agrupamento de Dados

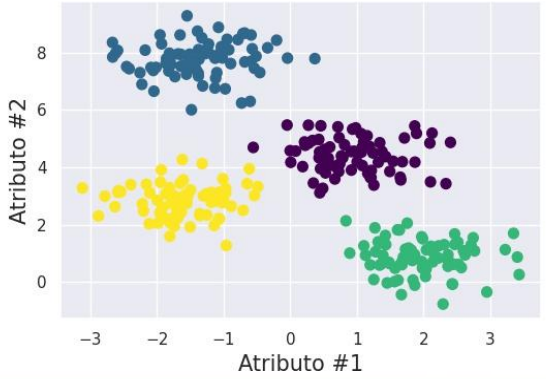
- **Particionais:** organizar dados em uma partição de *k clusters*
- **Hierárquicos:** organizar dados em uma decomposição hierárquica de *clusters* e *subclusters*

Conjunto de Dados





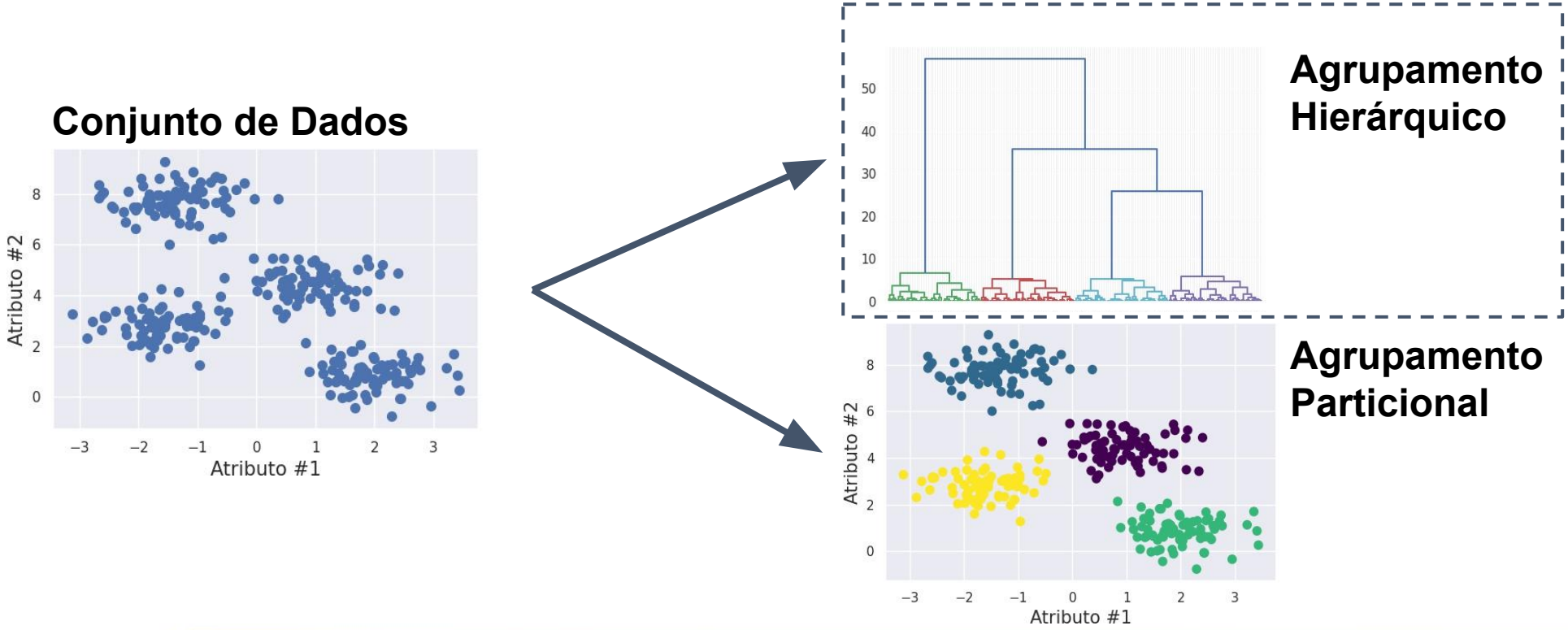
Agrupamento Hierárquico



Agrupamento Particional

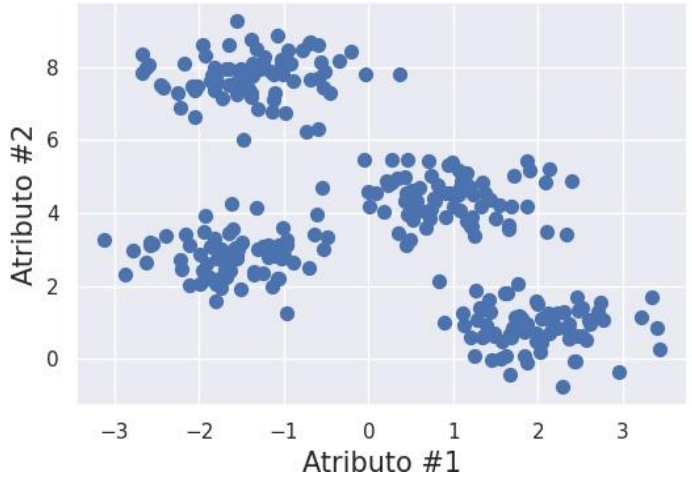
Métodos para Agrupamento de Dados

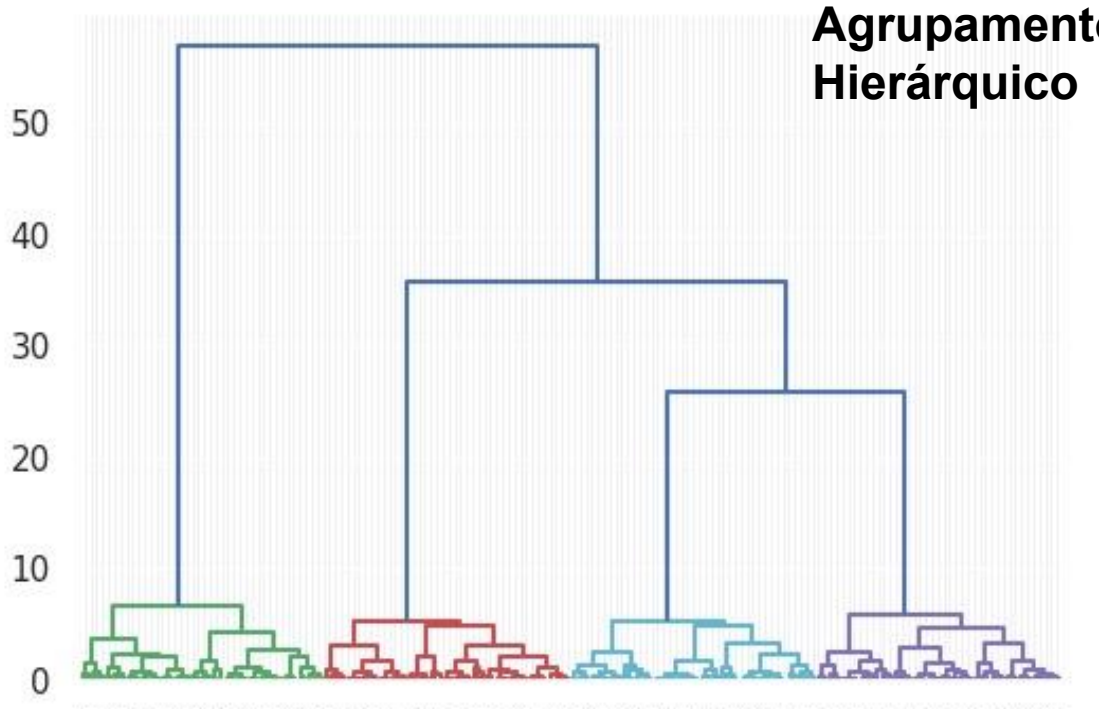
- **Particionais:** organizar dados em uma partição de *k clusters*
- **Hierárquicos:** organizar dados em uma decomposição hierárquica de *clusters* e *subclusters*



Agrupamento Hierárquico

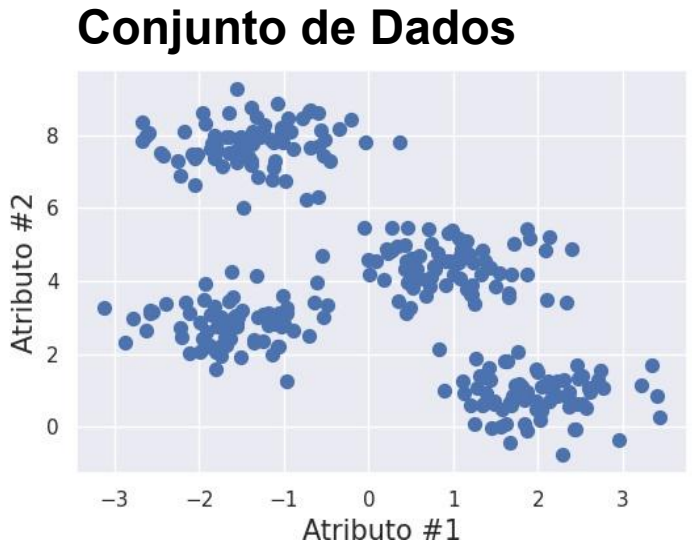
- **Dendrograma:** diagrama com a estrutura hierárquica que representa o resultado de um agrupamento.
  Sumariza a formação dos *clusters* e *subclusters*.

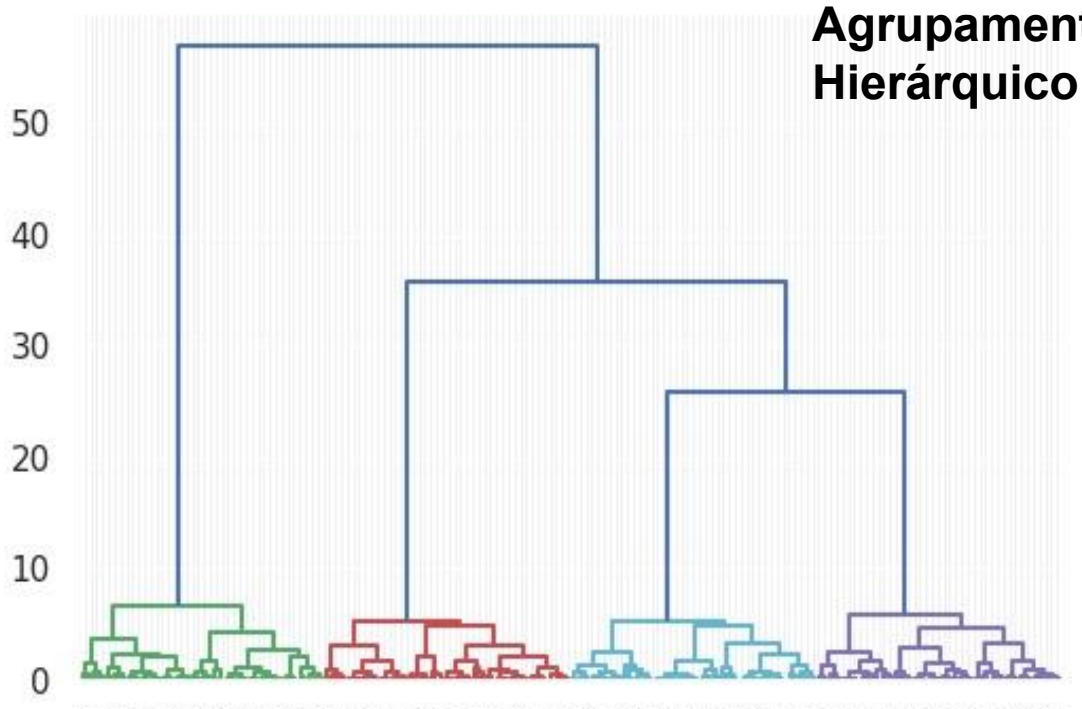




Agrupamento Hierárquico

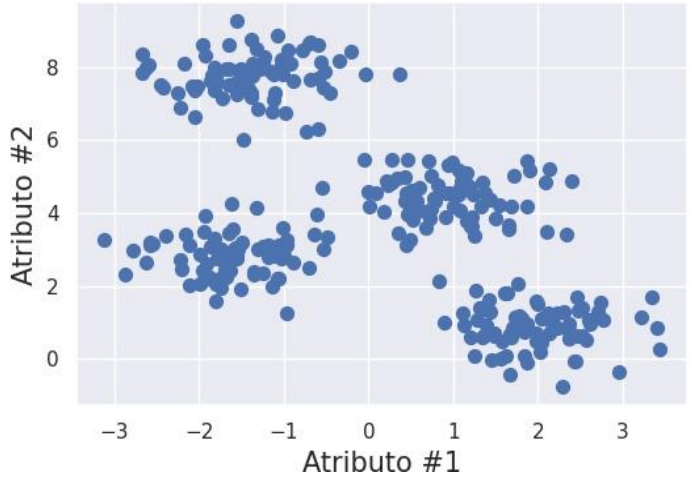
- Os objetos do conjunto de dados estão organizados no eixo x do dendrograma. A altura dos arcos indica a dissimilaridade entre objetos e grupos de objetos.

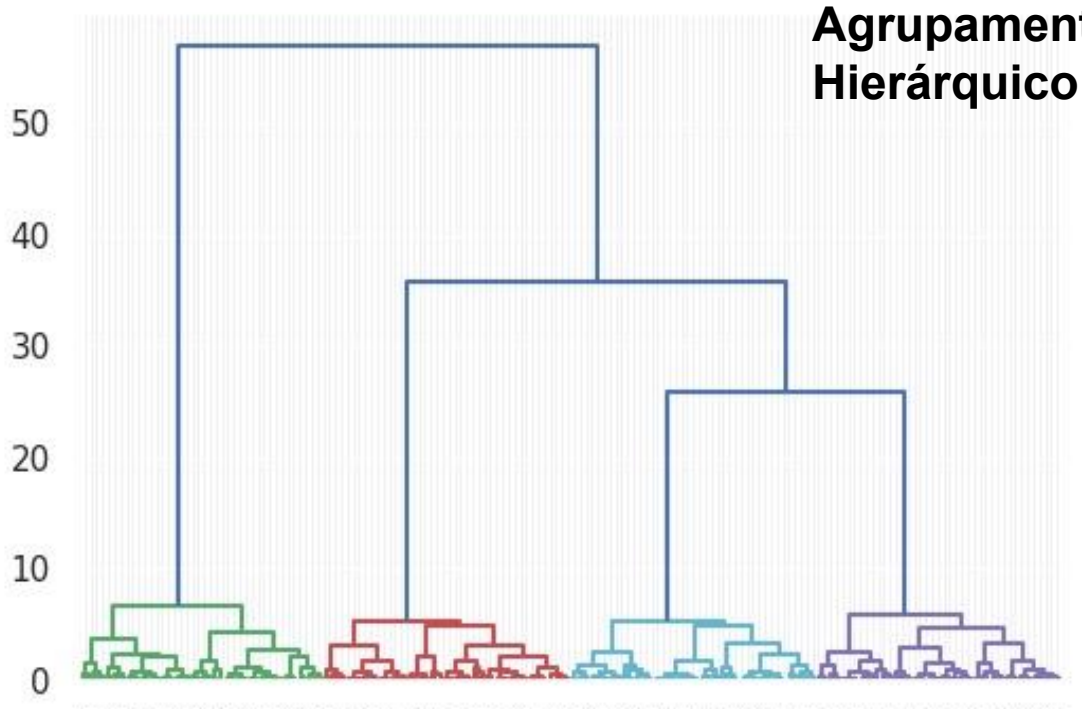




Agrupamento Hierárquico

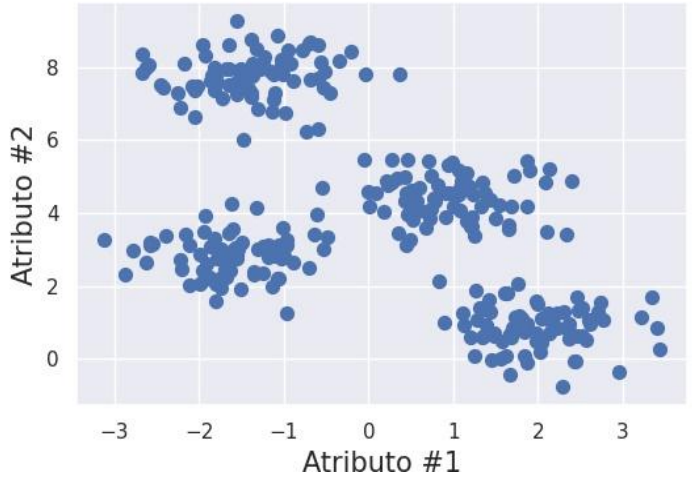
- Podemos inspecionar o dendrograma para estimar o número natural de clusters. No exemplo, há 4 subárvores bem separadas.

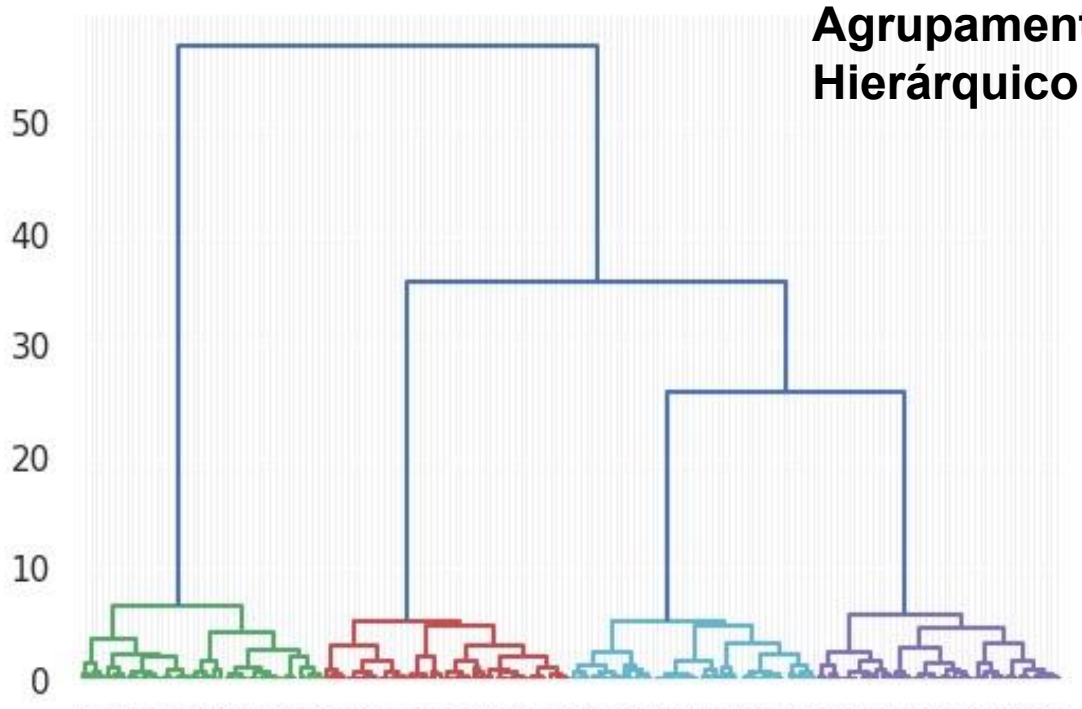




Agrupamento Hierárquico

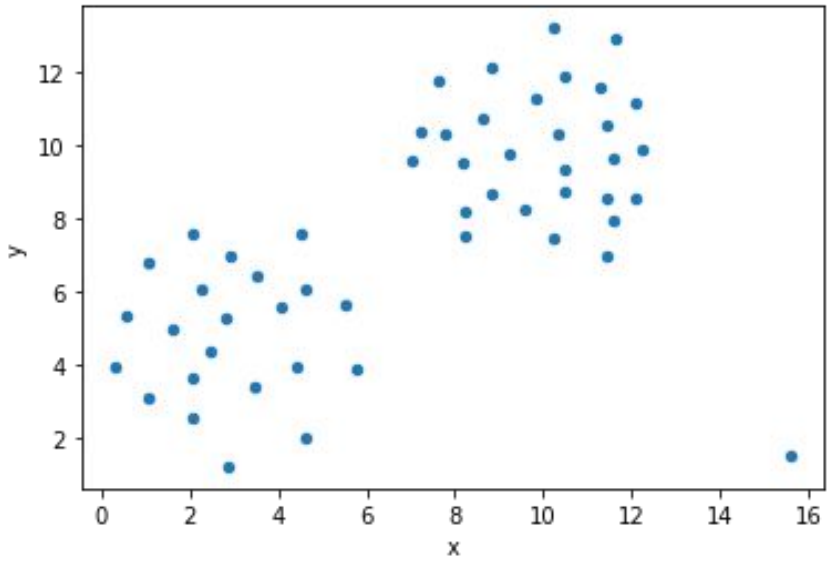
- Conceitos de homogeneidade (coesão interna) e heterogeneidade (separabilidade) dos clusters representados pela altura (eixo y) da união entre cluster.

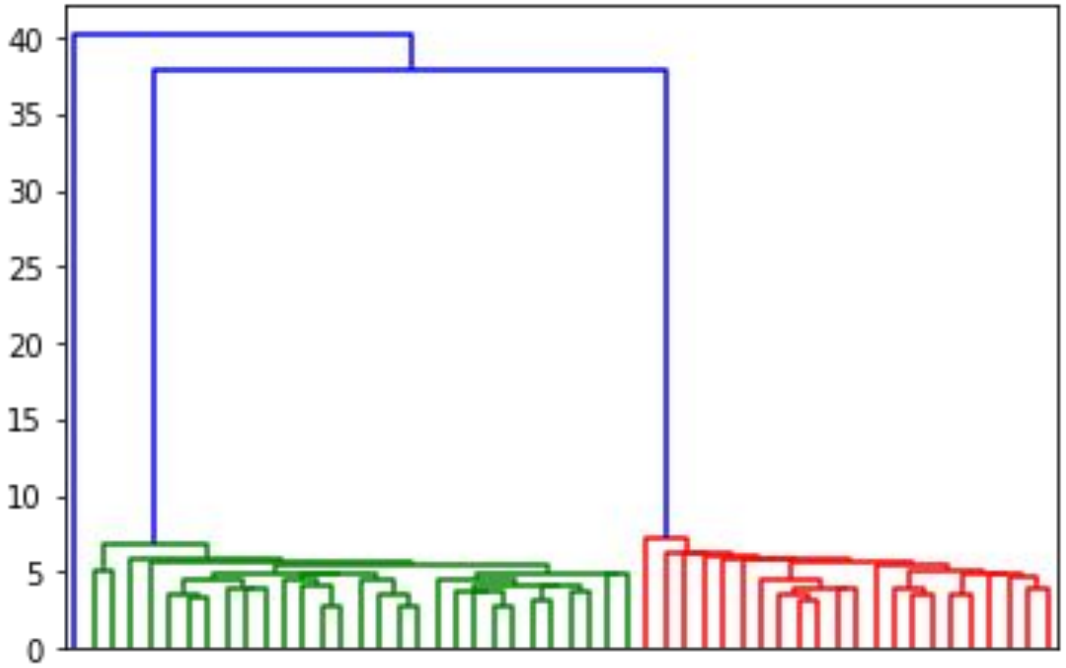




Agrupamento Hierárquico

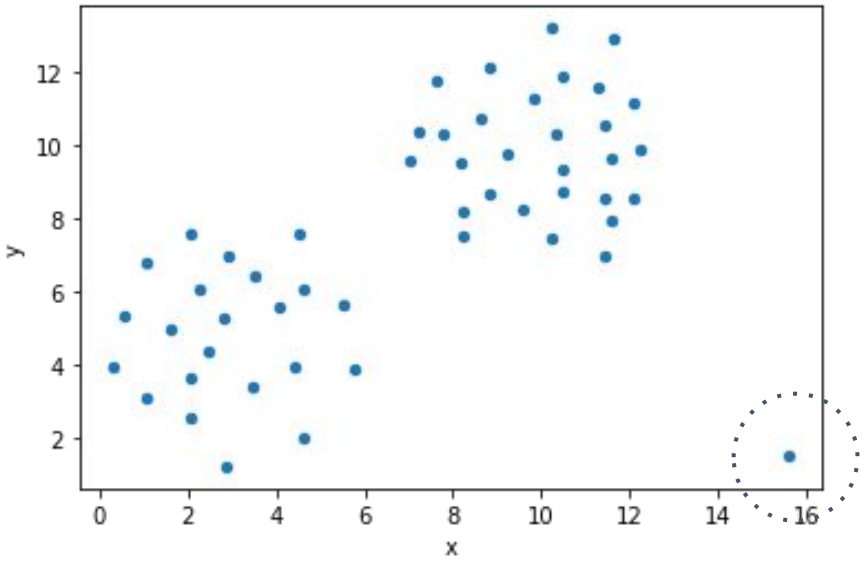
- Dendrogramas podem ser úteis para detectar *outliers*.
- Ramos isolados indicam objetos muito distante dos demais.

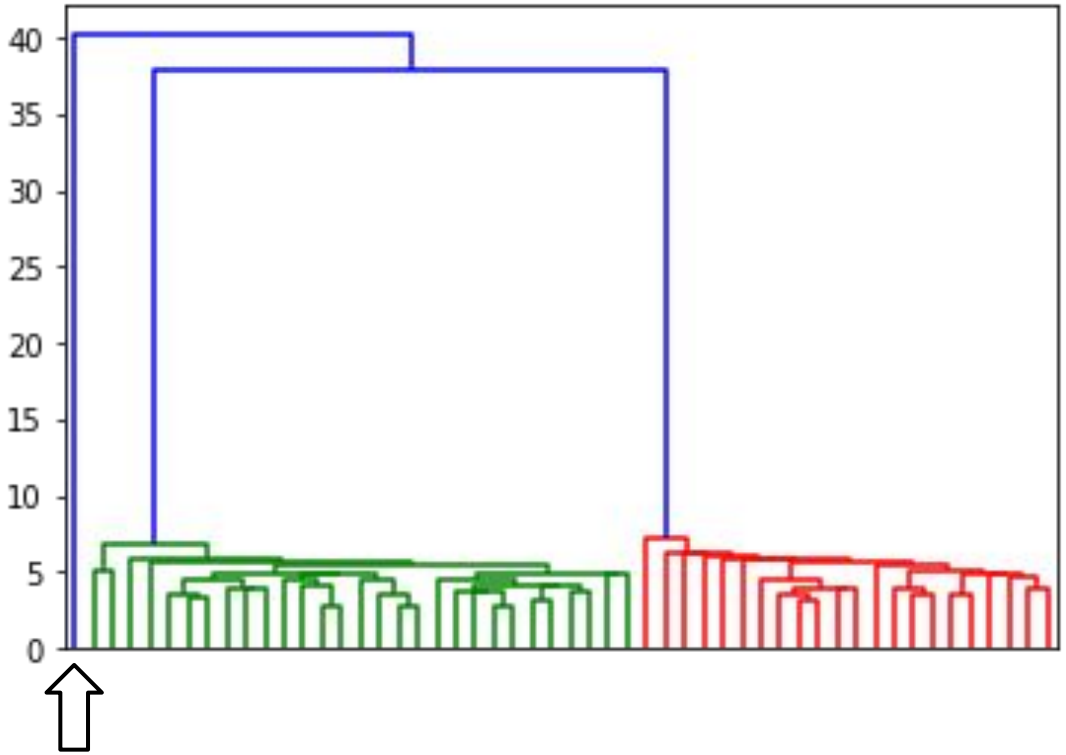




Agrupamento Hierárquico

- Dendrogramas podem ser úteis para detectar *outliers*.
- Ramos isolados indicam objetos muito distante dos demais.





Outlier

Agrupamento Hierárquico

- Dois métodos clássicos para agrupamento hierárquico

Aglomerativos:

➔ Iniciar alocando cada objeto em um *cluster*

➔ Encontrar o melhor par de *clusters* para unir

➔ Repetir até formar um único *cluster*

Divisivos:

➔ Iniciar alocando todos os objetos em um único *cluster*

➔ Dividir um *cluster* em dois *subclusters*

➔ Repetir a divisão até que cada objeto seja um *cluster*

Agrupamento Hierárquico

- Dois métodos clássicos para agrupamento hierárquico

Aglomerativos:

➔ Iniciar alocando cada objeto em um *cluster*

➔ Encontrar o melhor par de *clusters* para unir

➔ Repetir até formar um único *cluster*

Divisivos:

➔ Iniciar alocando todos os objetos em um único *cluster*

➔ Dividir um *cluster* em dois *subclusters*

➔ Repetir a divisão até que cada objeto seja um *cluster*

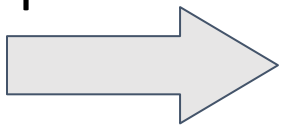
Agrupamento Hierárquico

- **Matriz de dissimilaridades:** armazena as distâncias entre cada par de objetos do conjunto de dados

Matriz atributo-valor

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Aplicar medida de proximidade



Matriz dissimilaridades

$$\mathbf{M} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

- As diagonais são dissimilaridades entre um mesmo objeto
- Por simetria, podemos usar apenas a triangular superior ou inferior da matriz de dissimilaridades

Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Fonte do exemplo:
Everitt, B. (2001). Landau S. Leese M.: Cluster Analysis, 4'th edition. Edward Arnold Publishers Ltd

Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

  Inicialmente, cada objeto é um cluster.

$$
\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array}
\begin{array}{c}
\begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\
\begin{pmatrix}
0 & & & & \\
2 & 0 & & & \\
6 & 5 & 0 & & \\
10 & 9 & 4 & 0 & \\
9 & 8 & 5 & 3 & 0
\end{pmatrix}
\end{array}
$$

1 2 3 4 5

Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

1 2 3 4 5

Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Atualizar a matriz de dissimilaridades!

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Agrupamento Hierárquico

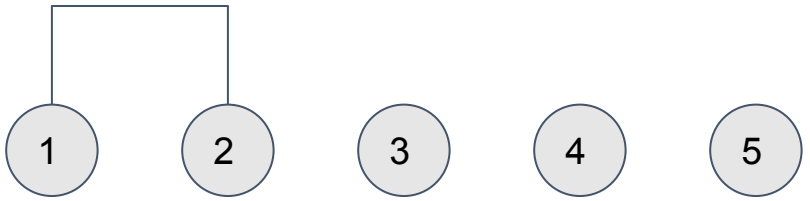
- **Exemplo de Agrupamento Hierárquico**

Considere uma matriz de dissimilaridades calculada para 5 objetos

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Atualizar a matriz de dissimilaridades!

	12	3	4	5
12	0			
3		0		
4		4	0	
5		5	3	0



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

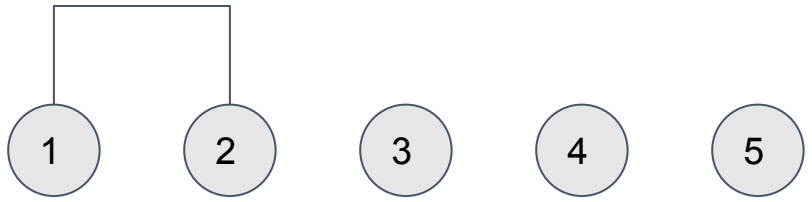
Considere uma matriz de dissimilaridades calculada para 5 objetos

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Atualizar a matriz de dissimilaridades!

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5$$
$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 9$$
$$d_{(12)5} = \min\{d_{15}, d_{25}\} = d_{25} = 8$$

	12	3	4	5
12	0			
3		0		
4		4	0	
5		5	3	0



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

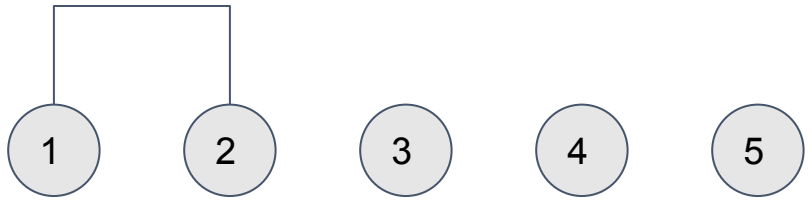
Considere uma matriz de dissimilaridades calculada para 5 objetos

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Atualizar a matriz de dissimilaridades!

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5$$
$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 9$$
$$d_{(12)5} = \min\{d_{15}, d_{25}\} = d_{25} = 8$$

	12	3	4	5
12	0			
3	5	0		
4	9	4	0	
5	8	5	3	0



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	12	**3**	**4**	**5**
12	0			
3	5	0		
4	9	4	0	
5	8	5	3	0

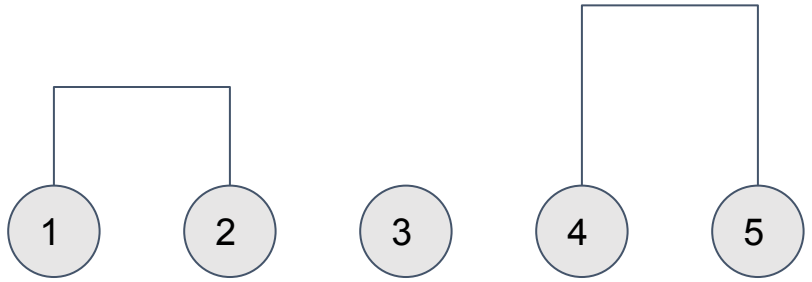
Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	12	3	4	5
12	0			
3	5	0		
4	9	4	0	
5	8	5	3	0



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

Considere uma matriz de dissimilaridades calculada para 5 objetos

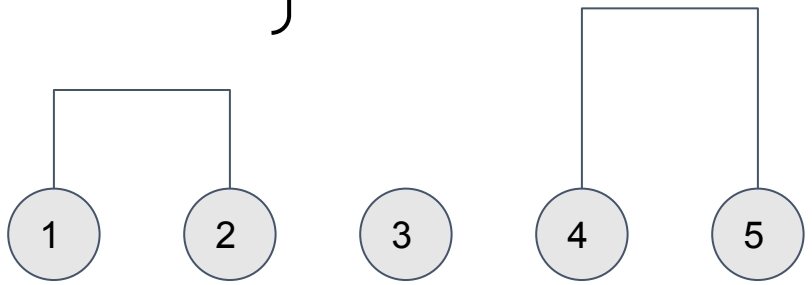
	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Atualizar a matriz de dissimilaridades!

$$d_{(12)(45)} = \min\{d_{14}, d_{15}, d_{24}, d_{25}\} = d_{25} = 8$$

$$d_{(45)3} = \min\{d_{43}, d_{53}\} = d_{43} = 4$$

	12	3	45
12	0		
3	5	0	
45			0



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

Considere uma matriz de dissimilaridades calculada para 5 objetos

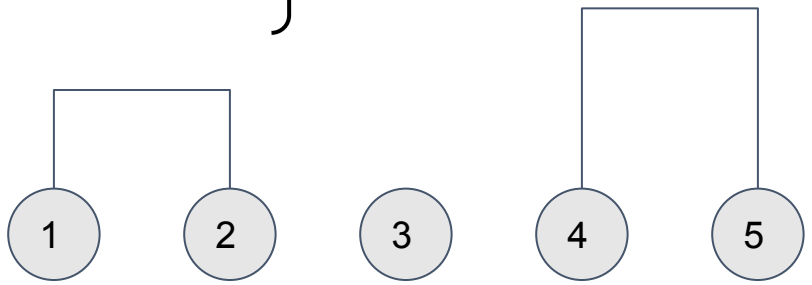
	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Atualizar a matriz de dissimilaridades!

$$d_{(12)(45)} = \min\{d_{14}, d_{15}, d_{24}, d_{25}\} = d_{25} = 8$$

$$d_{(45)3} = \min\{d_{43}, d_{53}\} = d_{43} = 4$$

	12	3	45
12	0		
3	5	0	
45	8	4	0



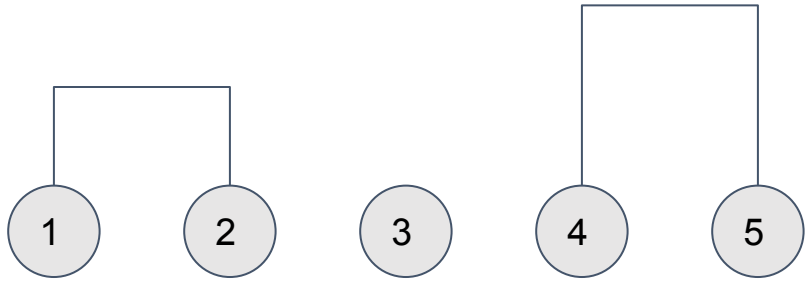
Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	12	**3**	**45**
12	0		
3	5	0	
45	8	4	0



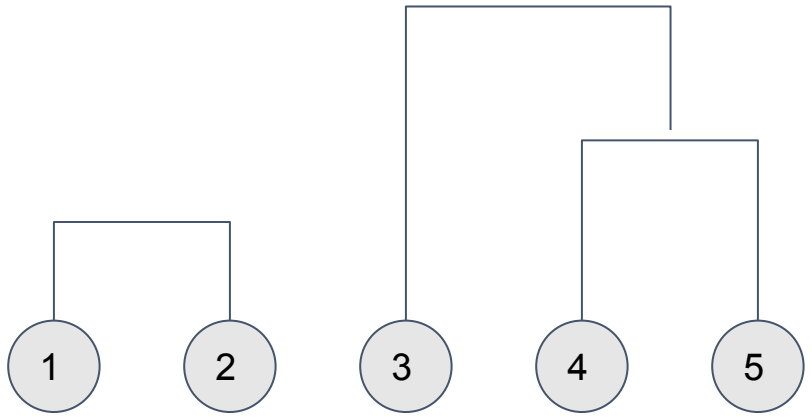
Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Qual é o melhor par de *clusters* para unir?

	12	**3**	**45**
12	0		
3	5	0	
45	8	4	0



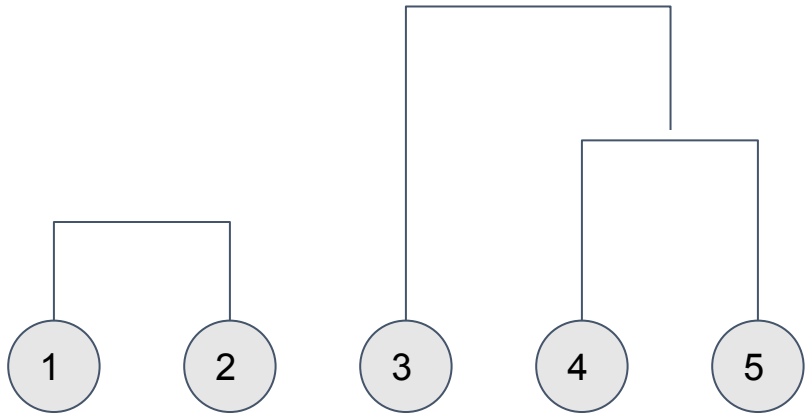
Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

Atualizar a matriz de dissimilaridades!

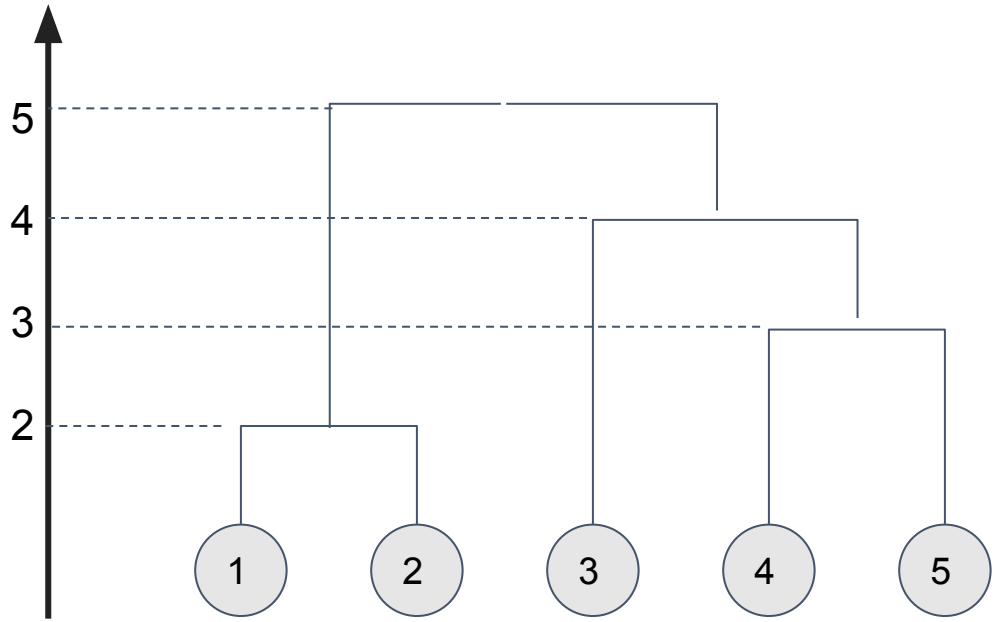
$$\begin{array}{c} \\ \mathbf{12} \\ \mathbf{345} \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{12} & \mathbf{345} \end{array} \\ \begin{bmatrix} 0 & \\ 5 & 0 \end{bmatrix} \end{array}$$



Agrupamento Hierárquico

- **Exemplo de Agrupamento Hierárquico**

  Considere uma matriz de dissimilaridades calculada para 5 objetos

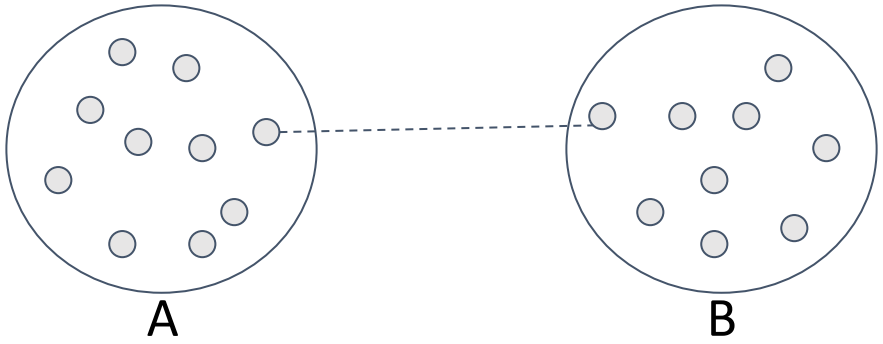


	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Agrupamento Hierárquico

- **A etapa mais importante é determinar distância entre clusters!**

  Nosso exemplo utilizou a estratégia de Vizinho mais Próximo (MIN)

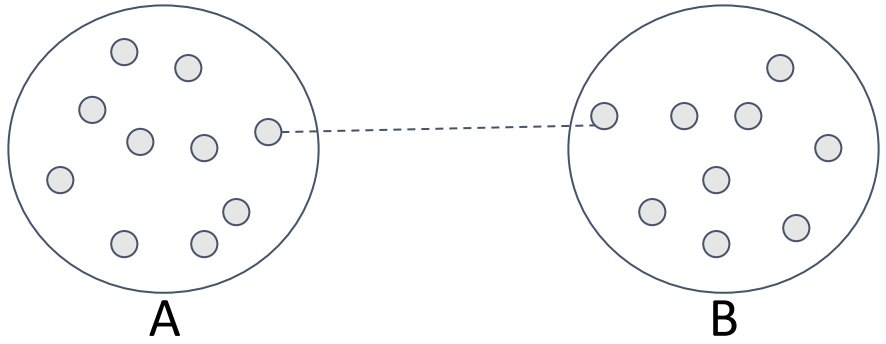


Agrupamento Hierárquico

- **A etapa mais importante é determinar distância entre clusters!**

  Nosso exemplo utilizou a estratégia de Vizinho mais Próximo (MIN)

Distância entre clusters A e B = menor distância de qualquer objeto do cluster A para qualquer objeto do cluster B



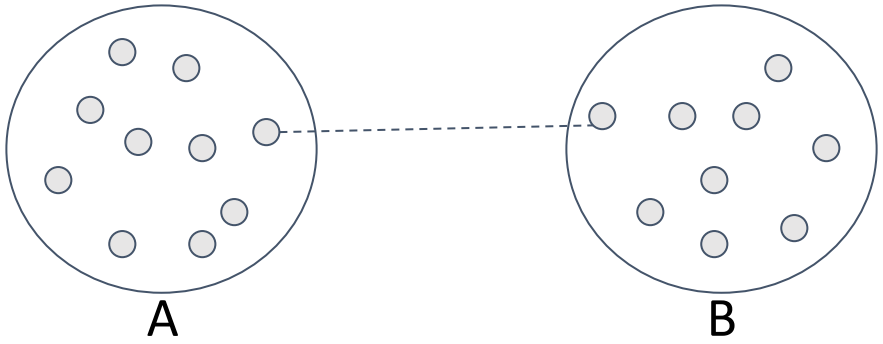
Agrupamento Hierárquico

- **A etapa mais importante é determinar distância entre clusters!**

  Nosso exemplo utilizou a estratégia de Vizinho mais Próximo (MIN)

Distância entre clusters A e B = menor distância de qualquer objeto do cluster A para qualquer objeto do cluster B

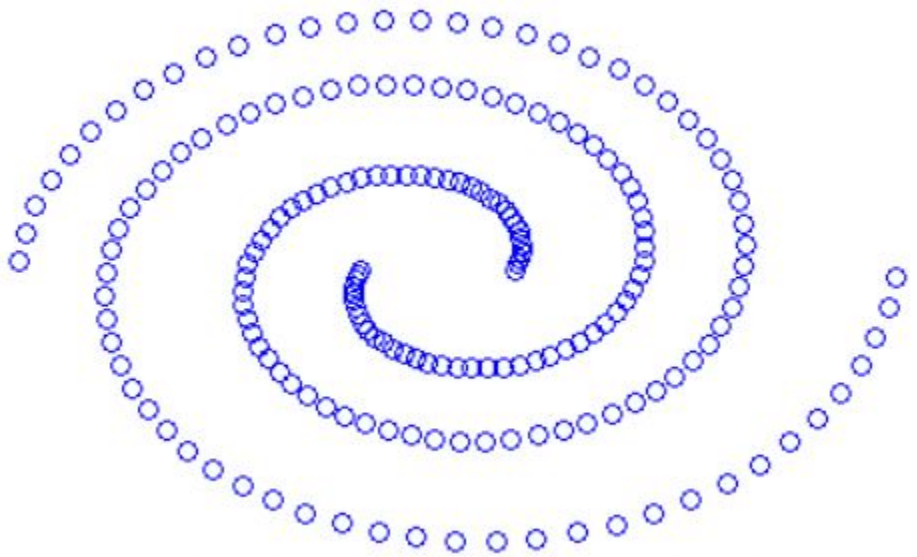
Single-Link



Agrupamento Hierárquico

- **Single-Link (MIN)**

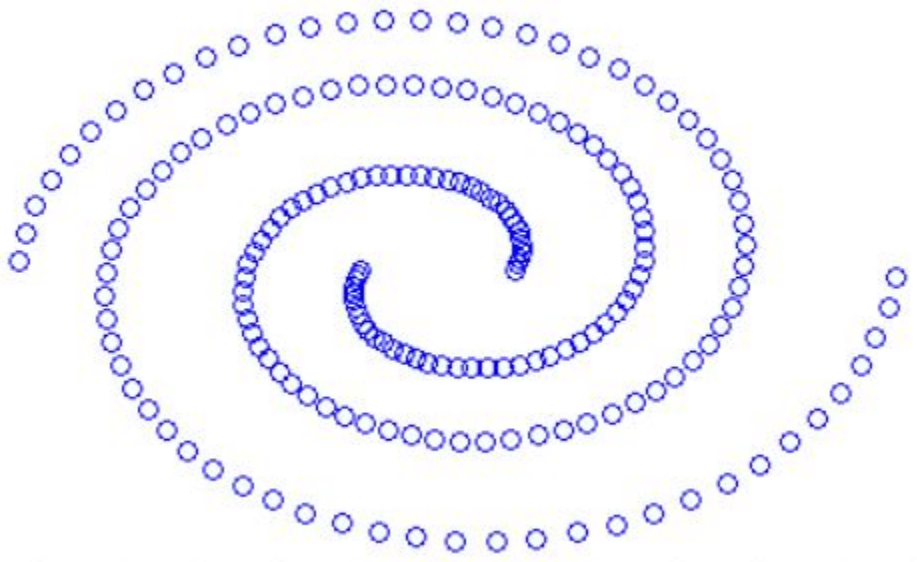
  Tende a capturar *clusters* baseado em contiguidade ou encadeamento.



Agrupamento Hierárquico

- **Single-Link (MIN)**

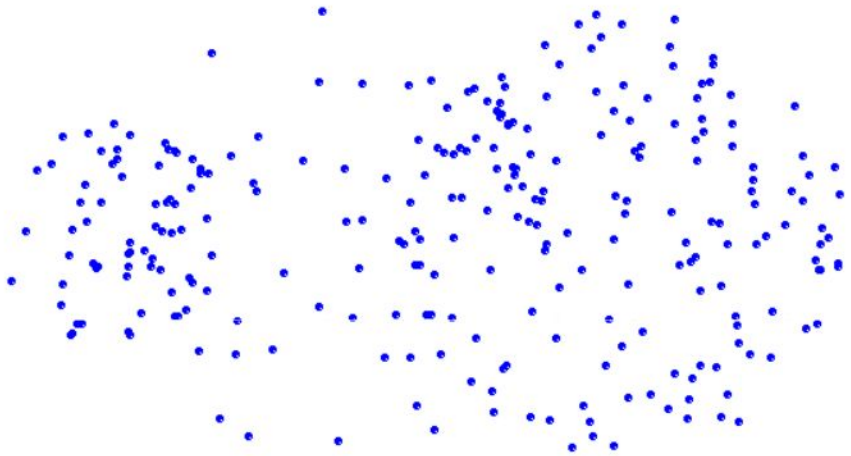
  Tende a capturar *clusters* baseado em contiguidade ou encadeamento.



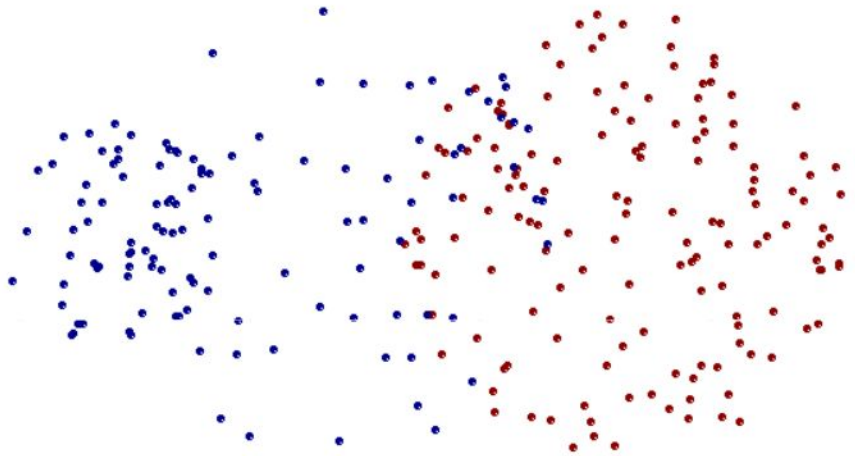
Sensível a ruídos!

Agrupamento Hierárquico

- **Single-Link (MIN)**



Pontos Originais

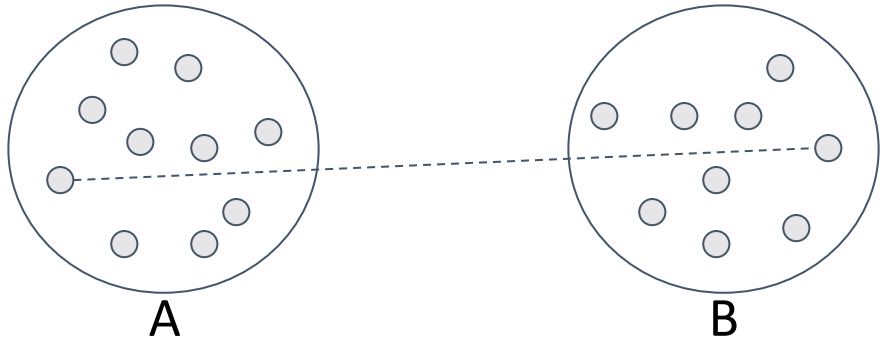


Pontos Agrupados
(dois clusters)

Agrupamento Hierárquico

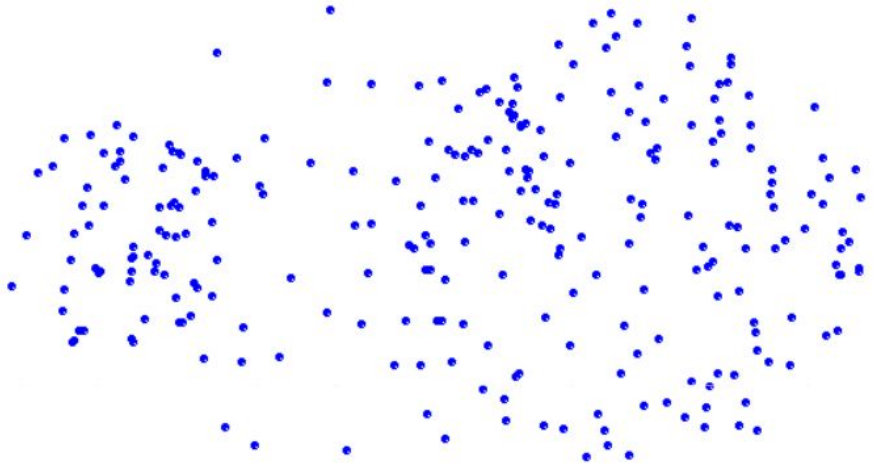
- **Complete-Link (MAX)**

Distância entre clusters A e B = maior distância de qualquer objeto do cluster A para qualquer objeto do cluster B

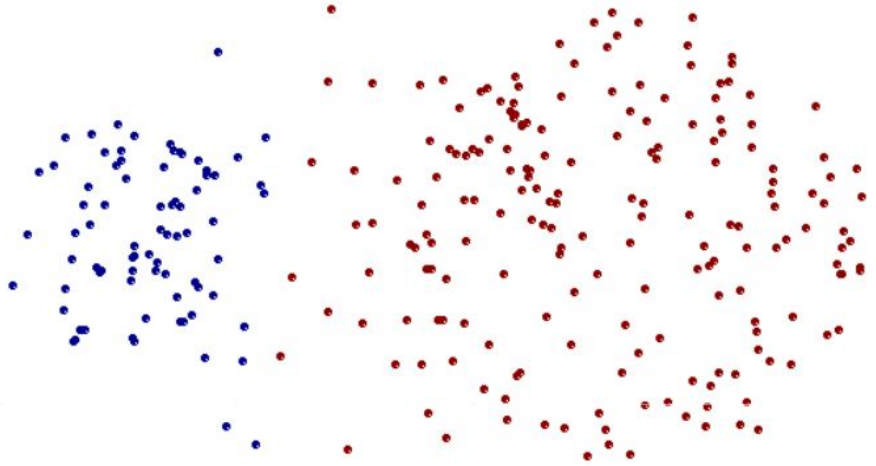


Agrupamento Hierárquico

- **Complete-Link (MAX)**



Pontos Originais

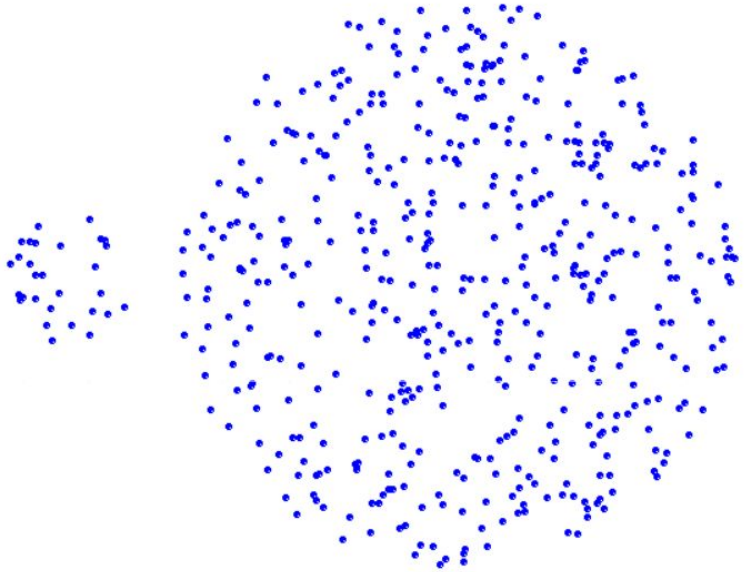


Pontos Agrupados
(dois clusters)

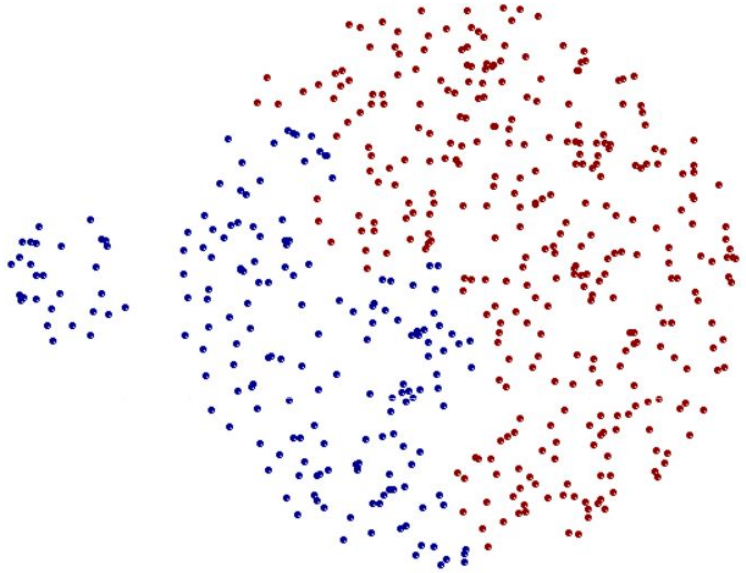
Tende a encontrar clusters com diâmetros semelhantes.
Vantagem: menos sensível a ruído.

Agrupamento Hierárquico

- **Complete-Link (MAX)**



Pontos Originais



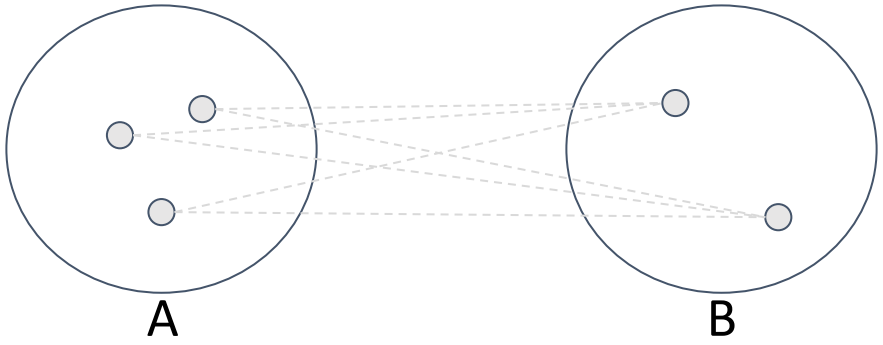
Pontos Agrupados
(dois clusters)

Tende a encontrar clusters com diâmetros semelhantes.
Desvantagem: tende a dividir grandes *clusters*.

Agrupamento Hierárquico

- **Average-Link (Média)**

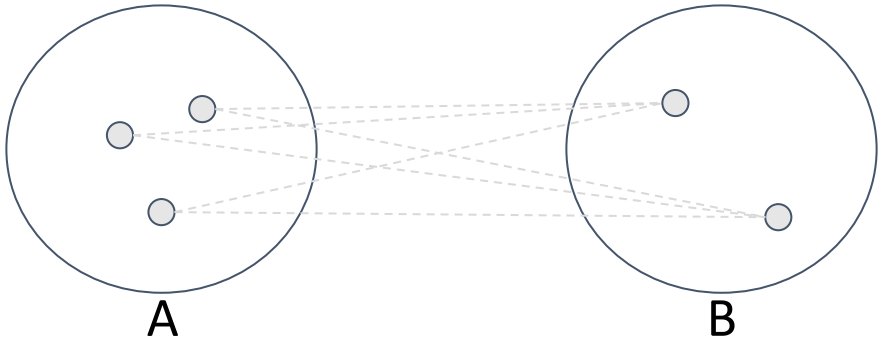
Distância entre clusters A e B = média das distâncias de todos os objetos do *cluster* A em relação aos objetos do *cluster* B



Agrupamento Hierárquico

- **Average-Link (Média)**

Distância entre clusters A e B = média das distâncias de todos os objetos do *cluster* A em relação aos objetos do *cluster* B



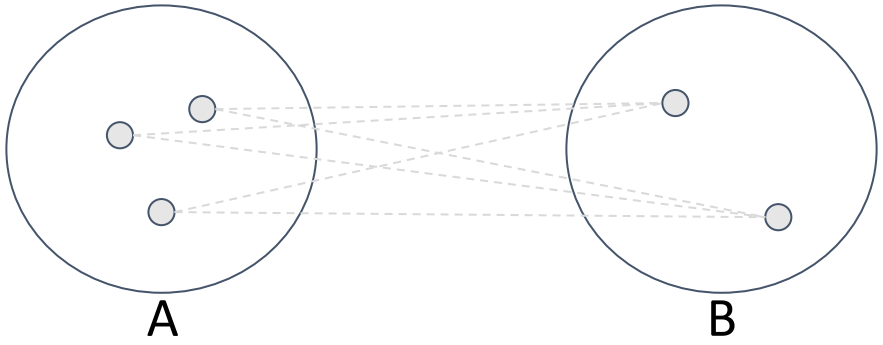
$$d(C_i, C_j) = \frac{\sum_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)}{|C_i||C_j|}$$

Agrupamento Hierárquico

- **Average-Link (Média)**

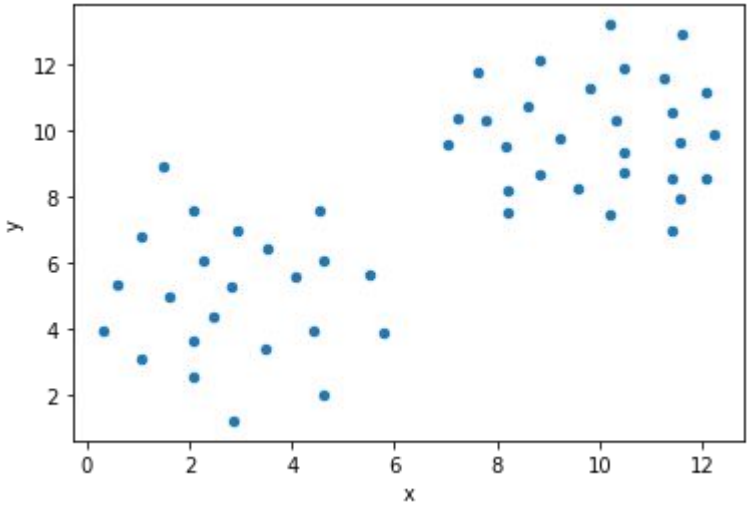
Distância entre clusters A e B = média das distâncias de todos os objetos do *cluster* A em relação aos objetos do *cluster* B

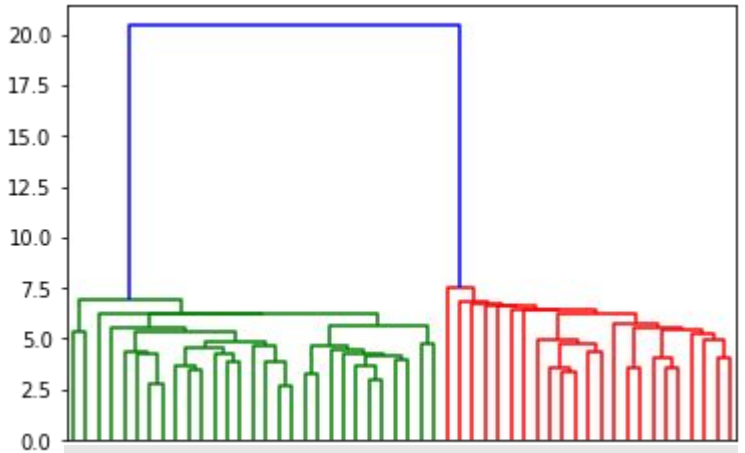
Ponto de equilíbrio entre Single-Link e Complete-Link



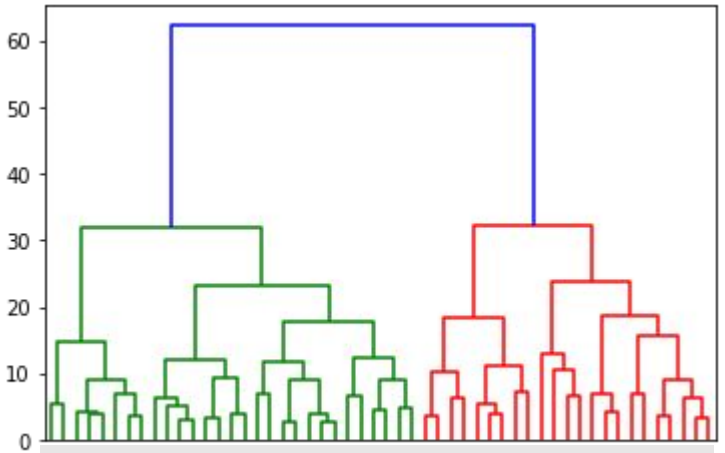
$$d(C_i, C_j) = \frac{\sum_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)}{|C_i||C_j|}$$

Agrupamento Hierárquico

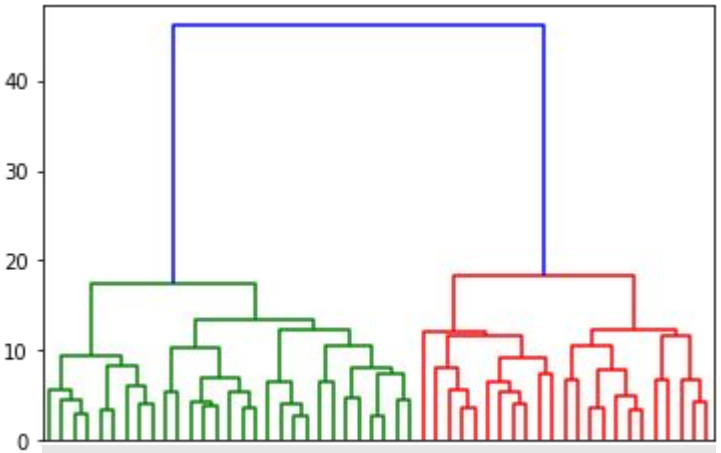




Single-Link



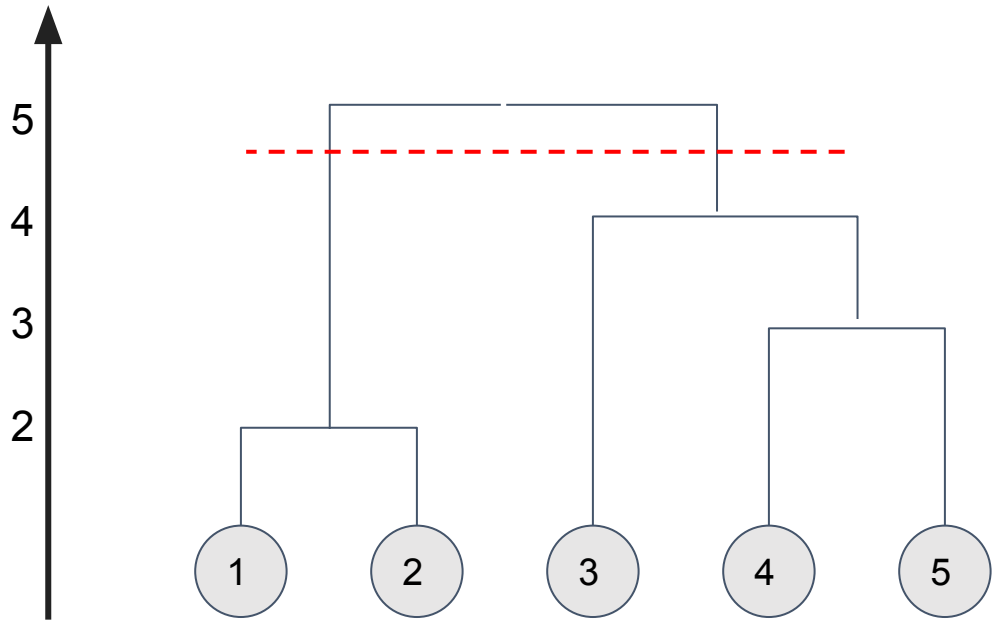
Complete-Link



Average-Link

Agrupamento Hierárquico

- Podemos extrair partições a partir de um agrupamento hierárquico
- Escolha do número de clusters *"a posteriori"*

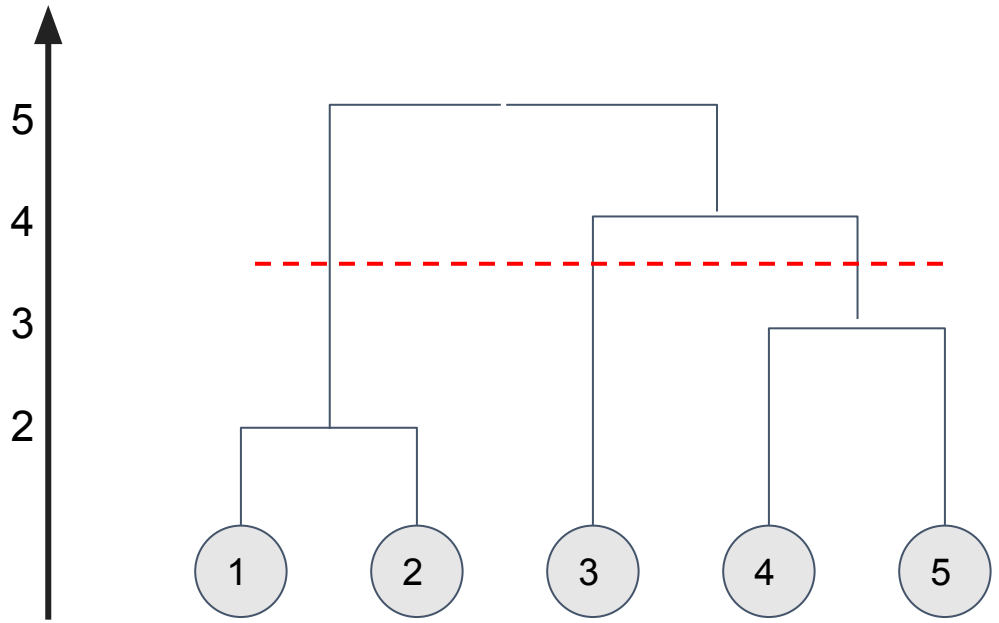


Cortes no dendrograma

Partição P = {(1,2), (3,4,5)}

Agrupamento Hierárquico

- Podemos extrair partições a partir de um agrupamento hierárquico
- Escolha do número de clusters *"a posteriori"*



Cortes no dendrograma

Partição P = {(1,2), (3), (4,5)}

Bibliografia

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.