Aula 11 – Agrupamento Hierárquico

1001524 – Aprendizado de Máquina I
2023/1 - Turmas A, B e C
Prof. Dr. Murilo Naldi

naldi@ufscar.br

Agradecimentos

- Parte do material utilizado nesta aula foi cedido pelos professores Ricardo Campello, Eduardo Hruscka e André Carvalho e, por esse motivo, o crédito deste material é deles
- Parte do material utilizado nesta aula foi disponibilizado por M. Kumar no endereço: www-users.cs.umn.edu/~kumar/dmbook/index.php
- Baseado no material de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.comwww.kdnuggets.com
- Figuras também foram adaptadas de Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.
- Agradecimentos a Intel Software e a

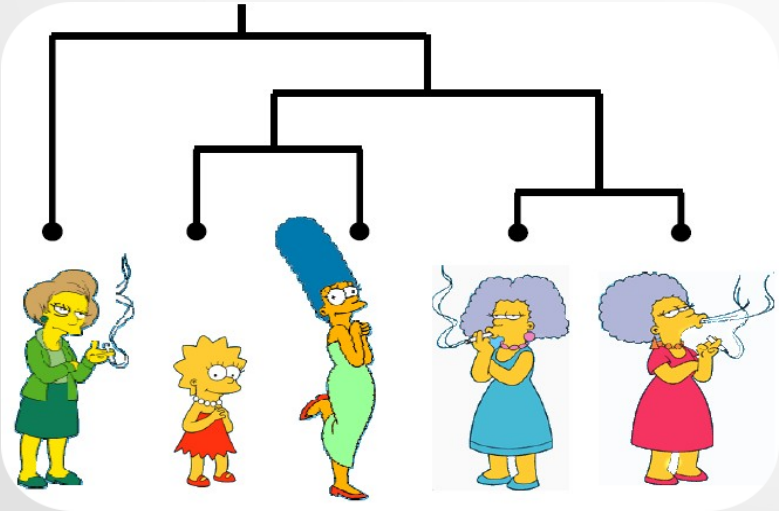
  Intel IA Academy pelo material

  disponibilizado e recursos didáticos

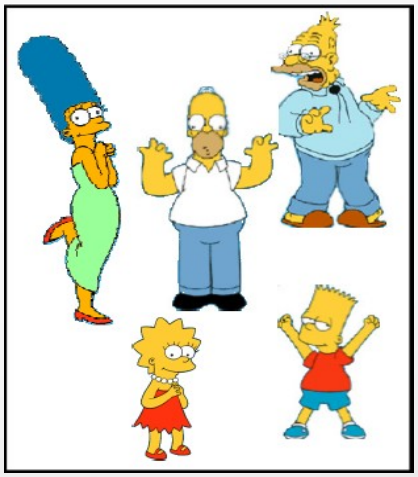
Métodos Particionais vs Hierárquicos

- Métodos Particionais:
  - constroem uma partição dos dados
- Métodos Hierárquicos:
  - constroem uma hierarquia de partições

Hierarquia



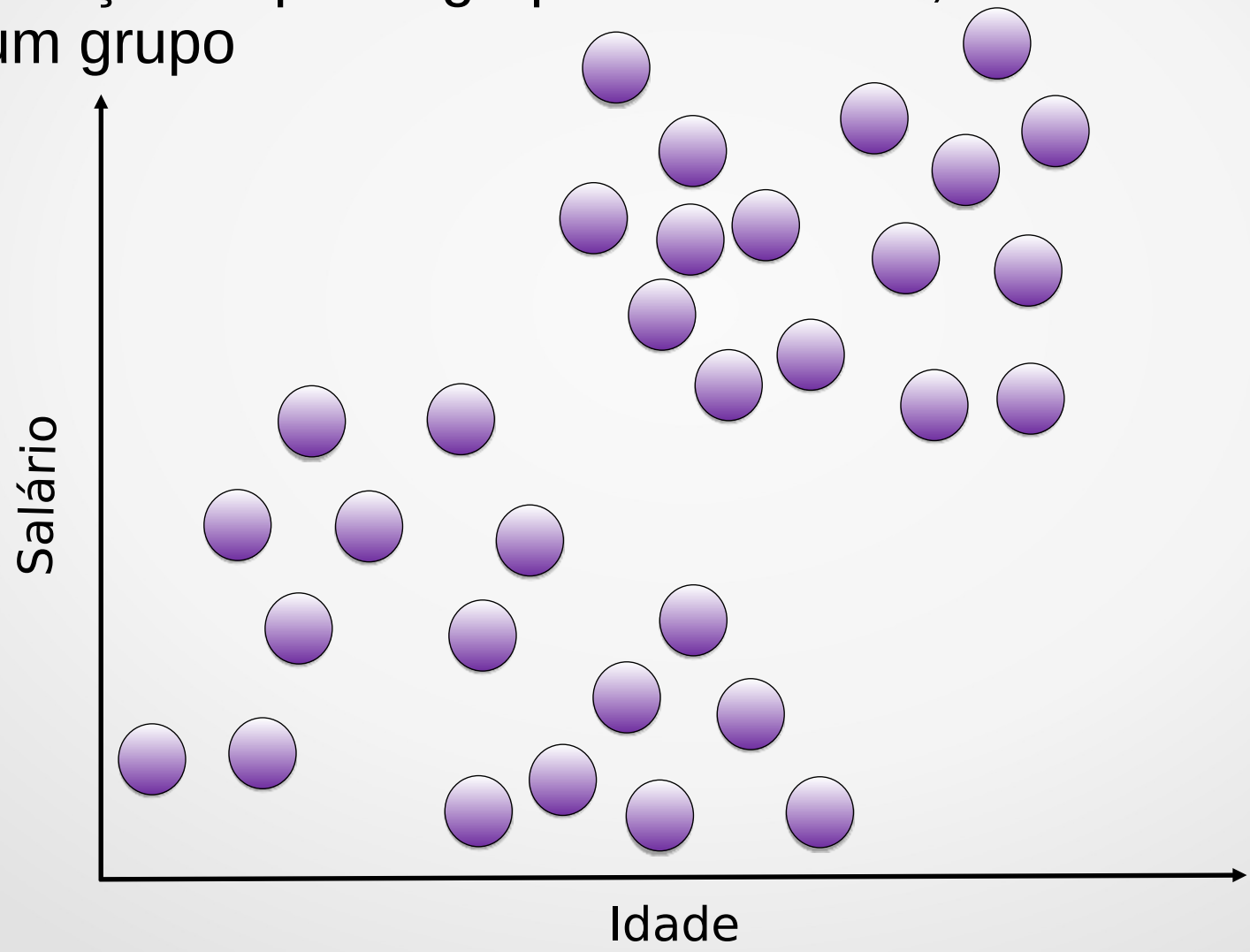
Partição



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

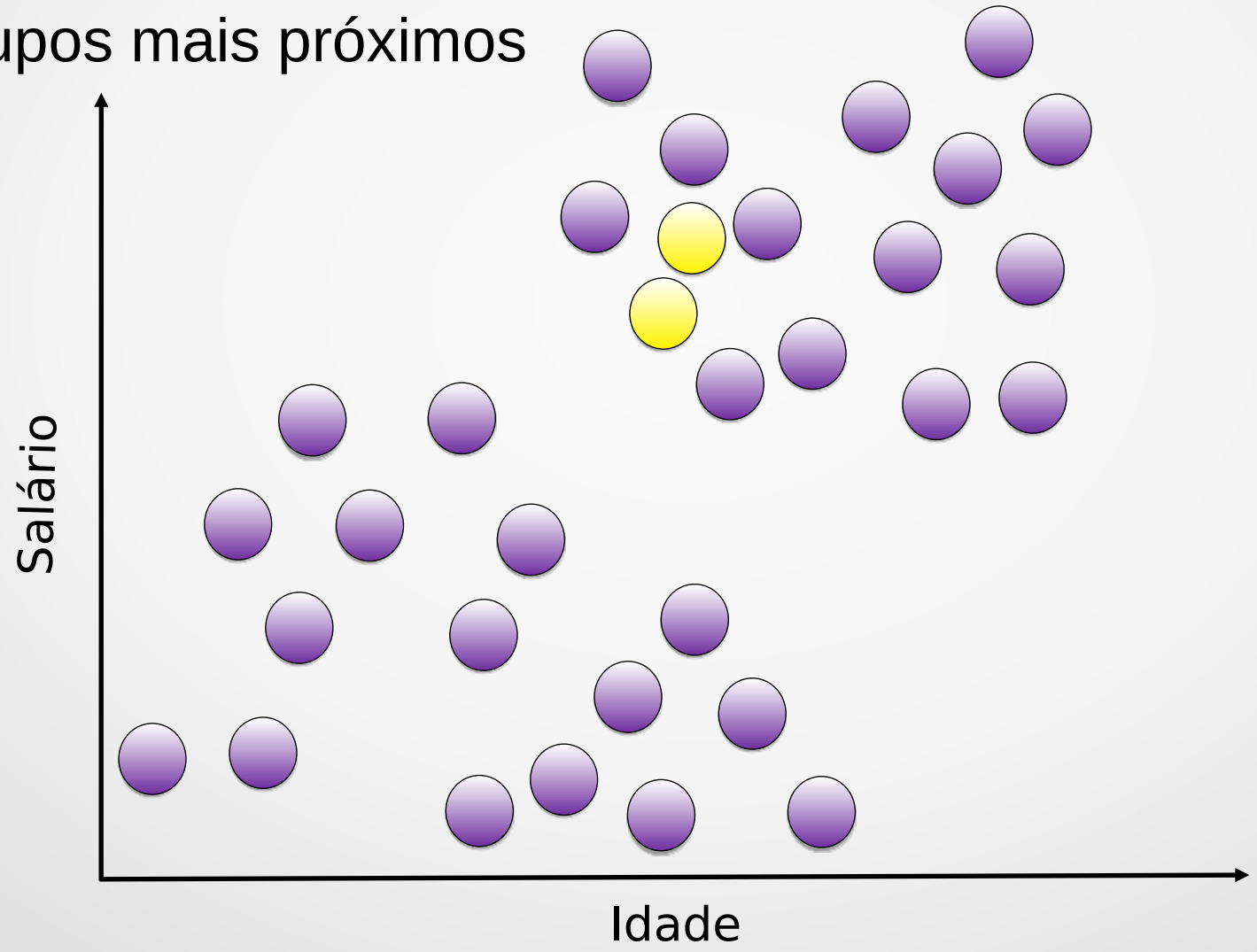
Algoritmo Aglomerativo

- Começando pelo agrupamento trivial, onde cada objeto é um grupo



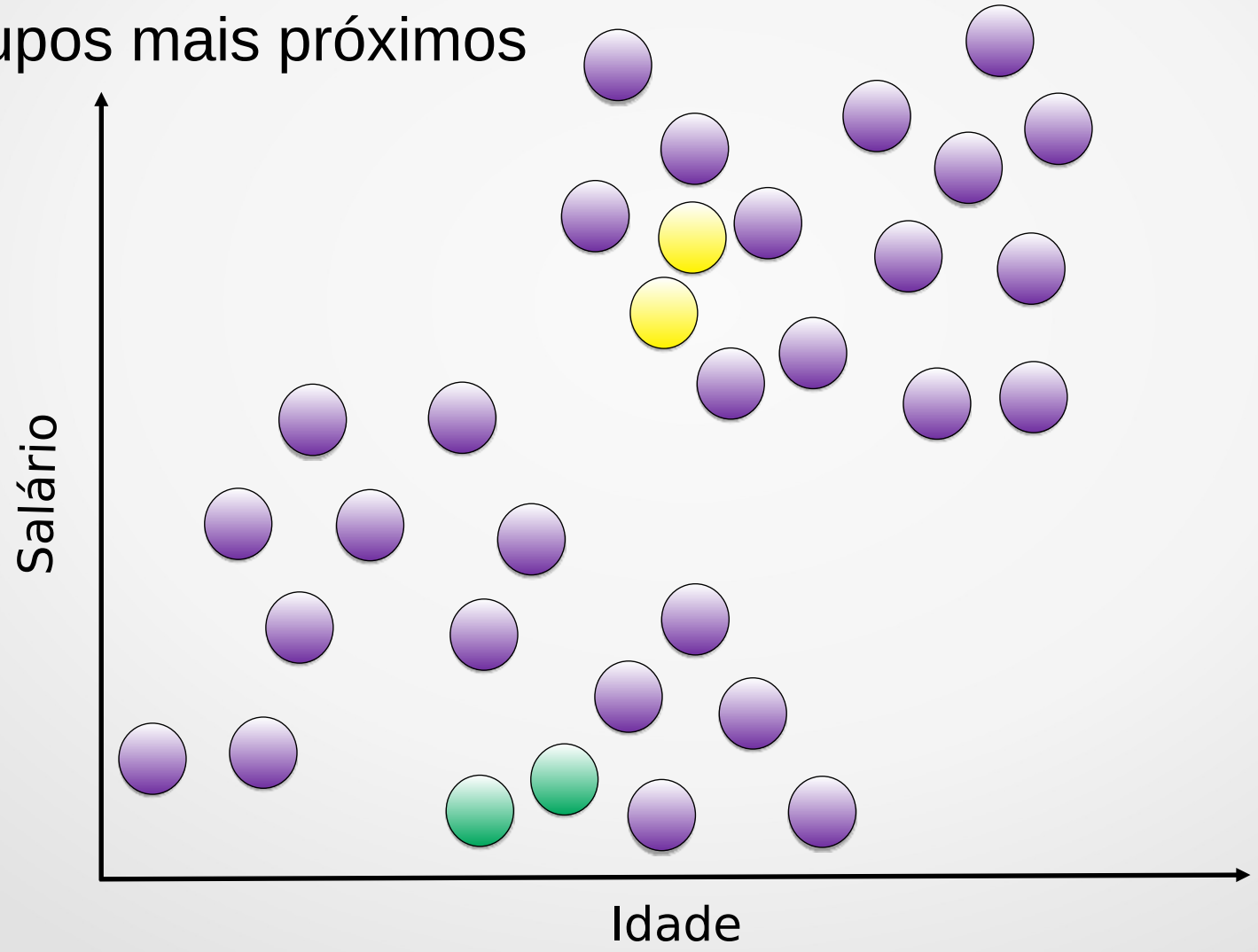
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



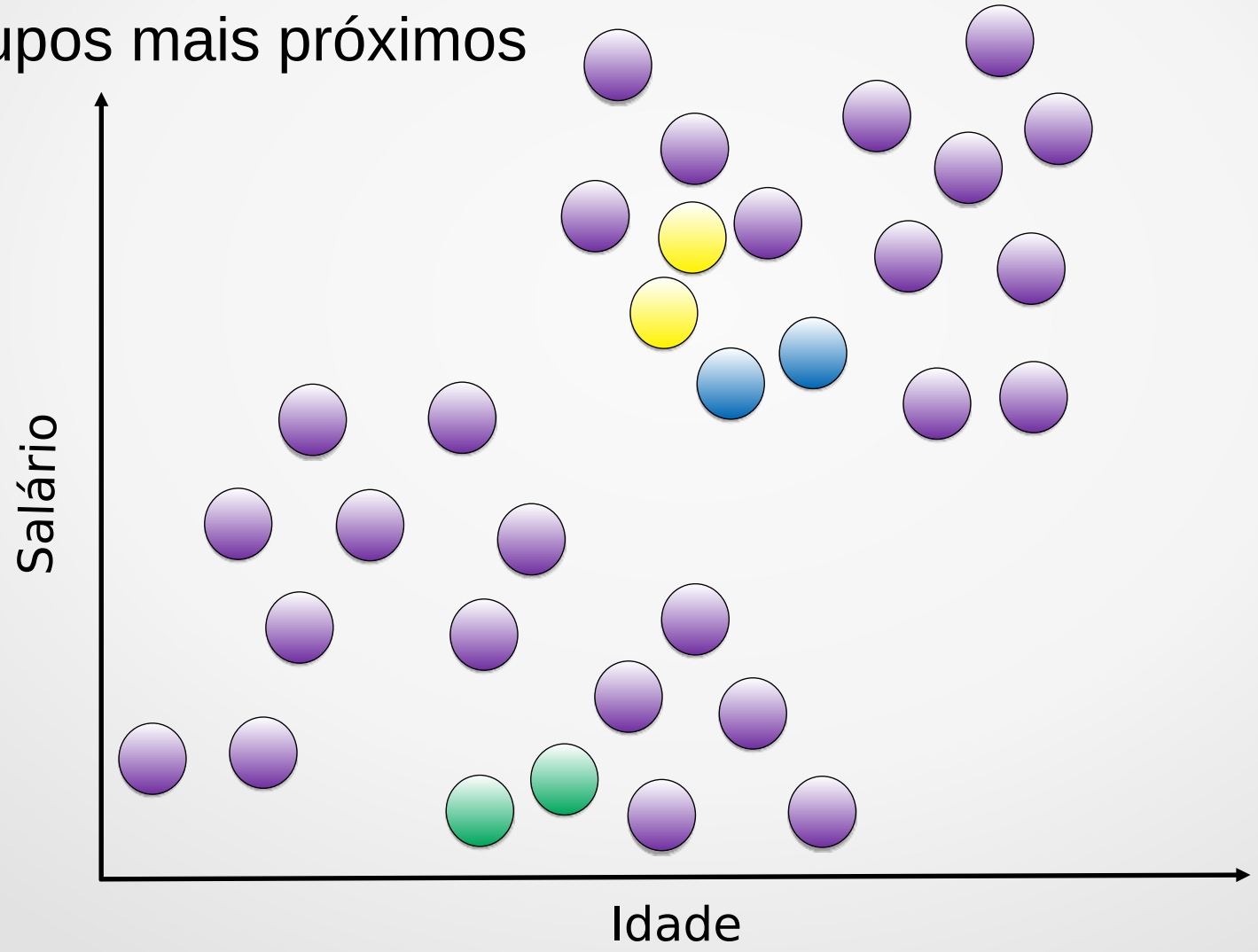
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



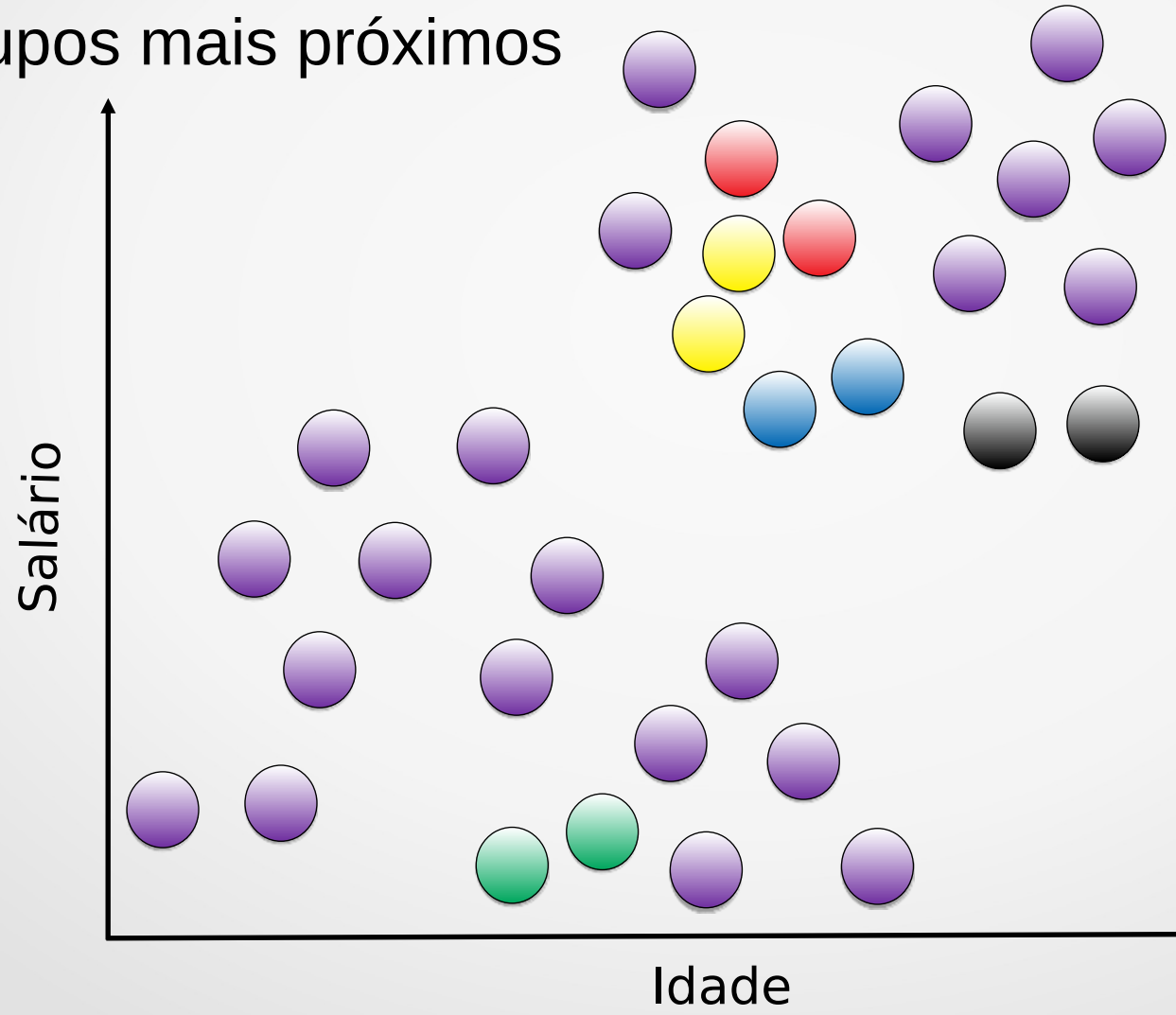
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



Algoritmo Aglomerativo

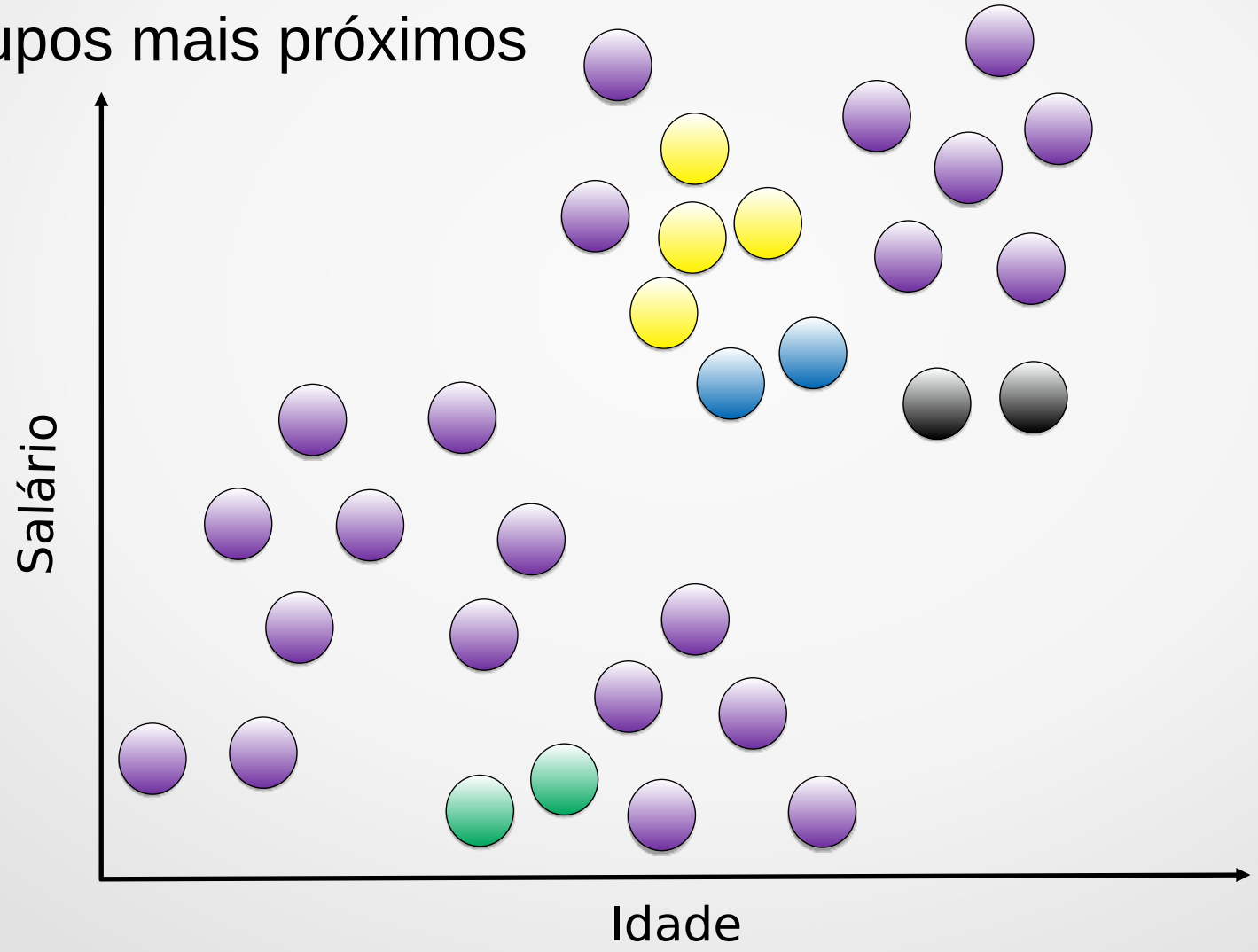
- Aglomeramos os dois grupos mais próximos



Nesse caso, houve empate na distância!

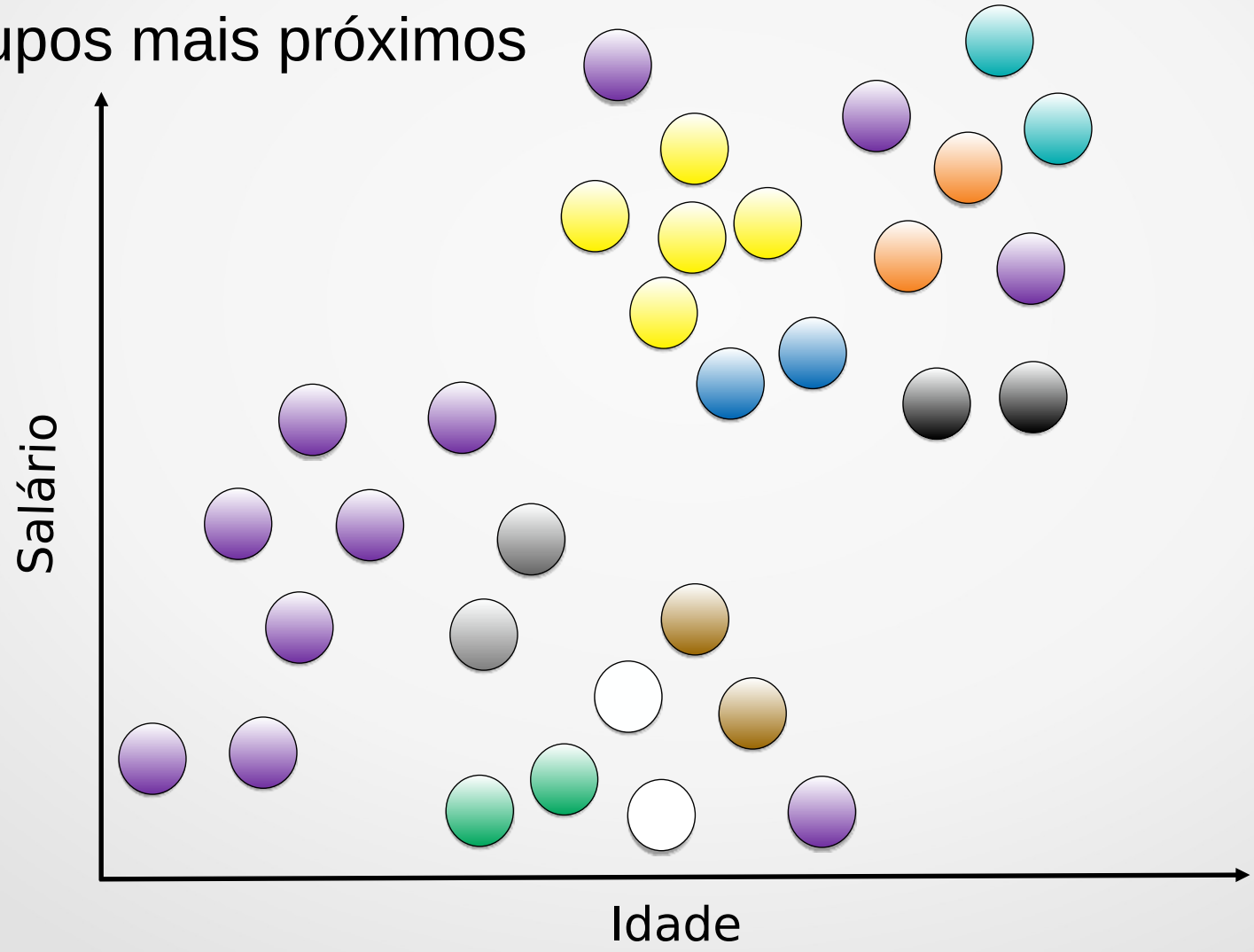
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



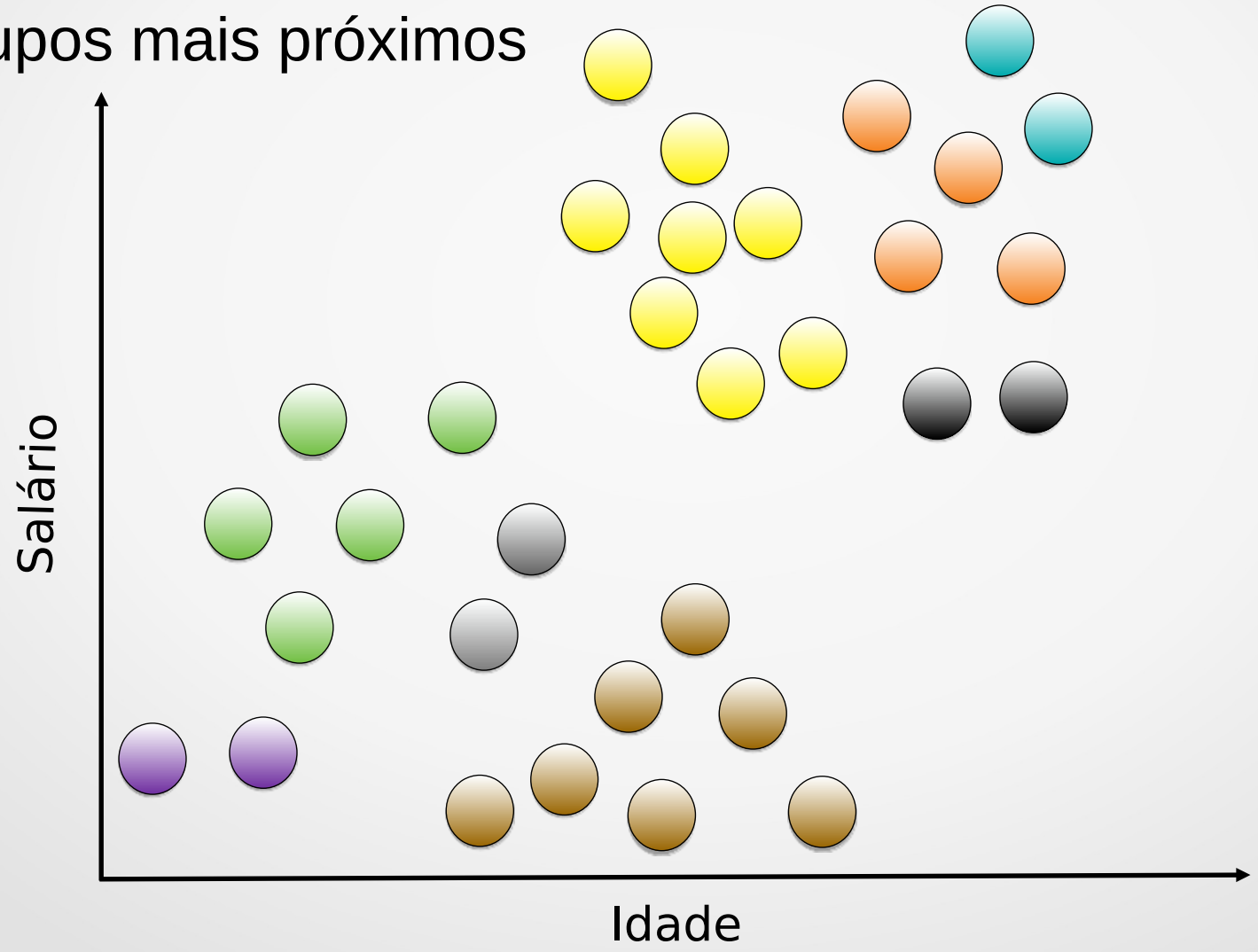
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



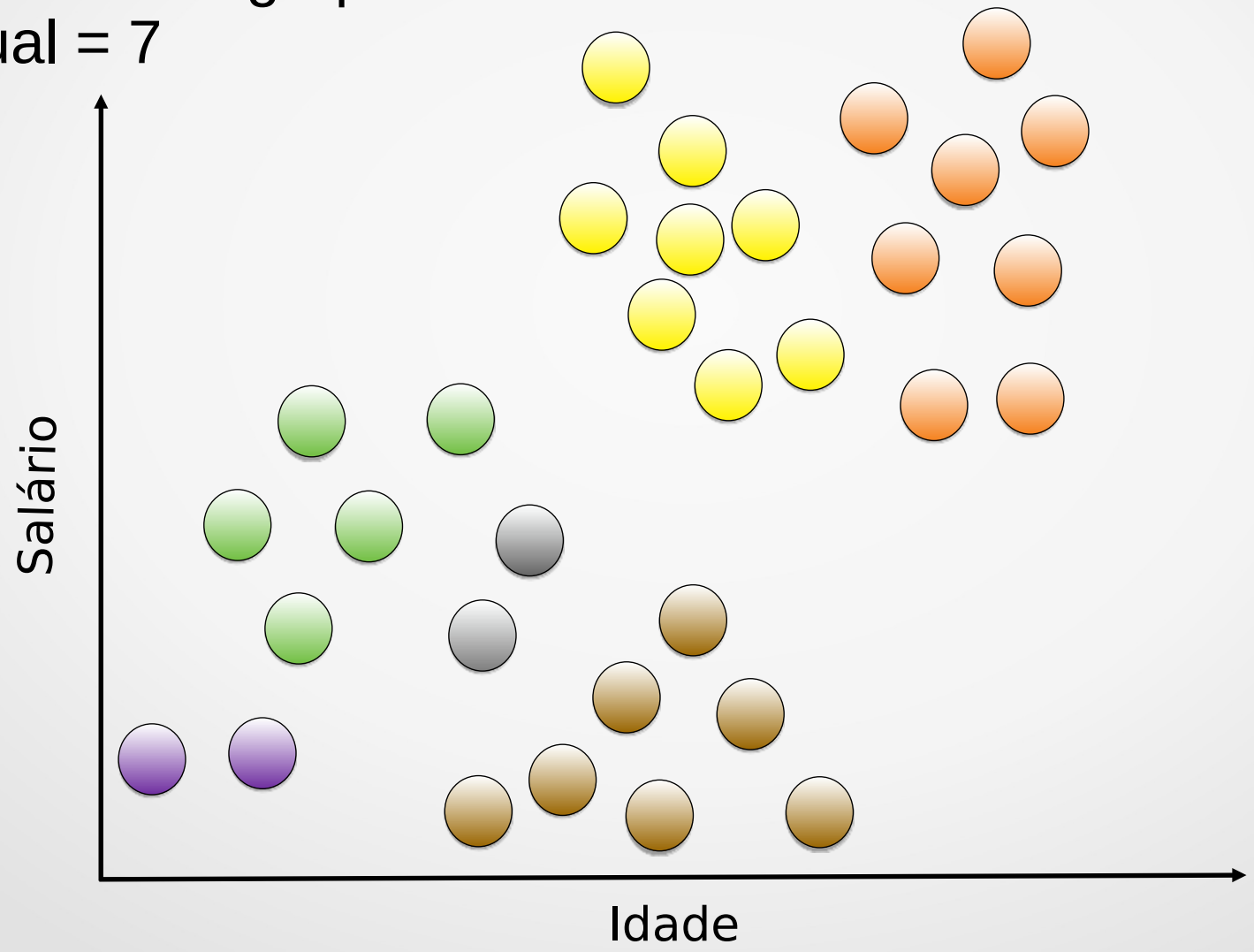
Algoritmo Aglomerativo

- Aglomeramos os dois grupos mais próximos



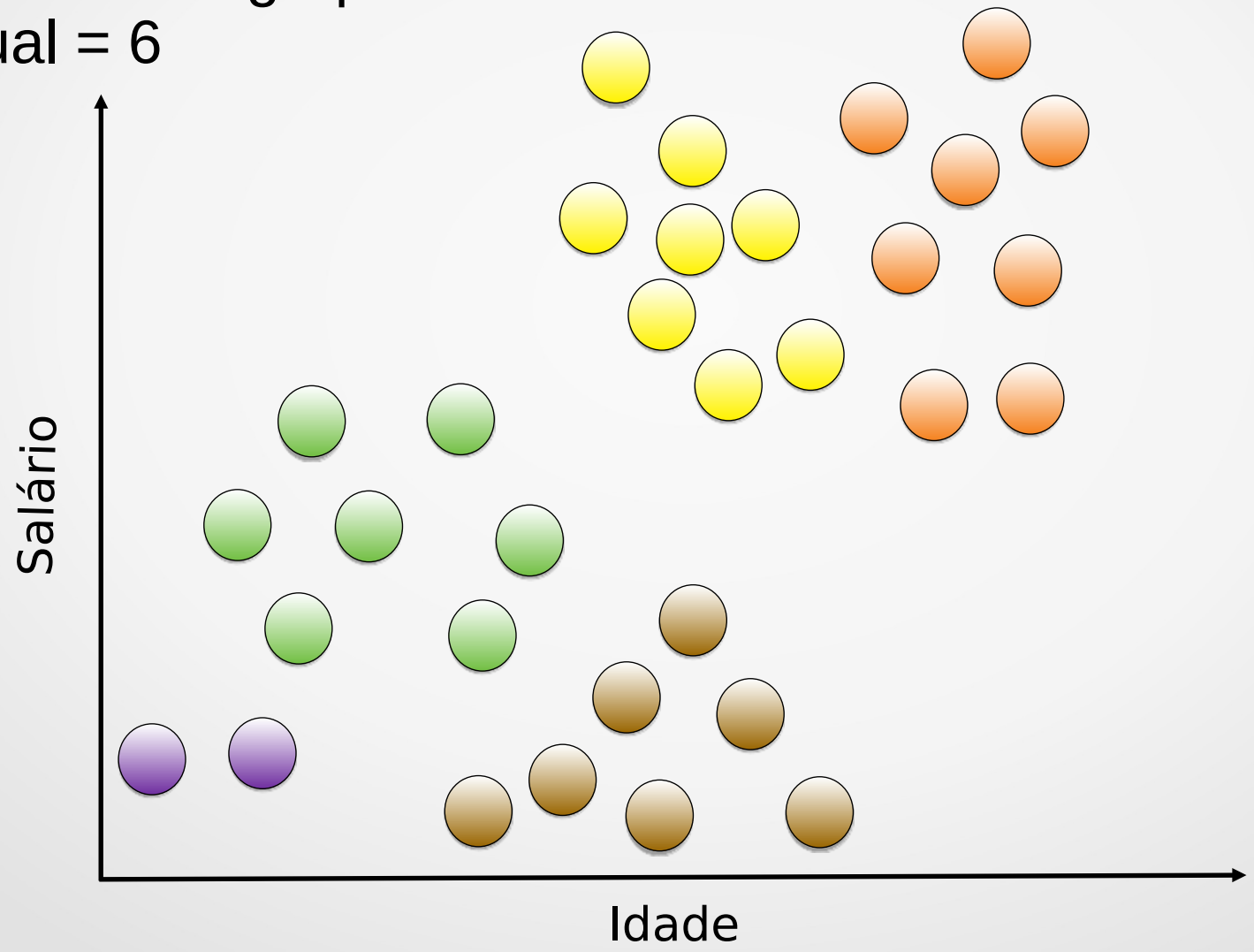
Algoritmo Aglomerativo

- Número de grupos atual = 7



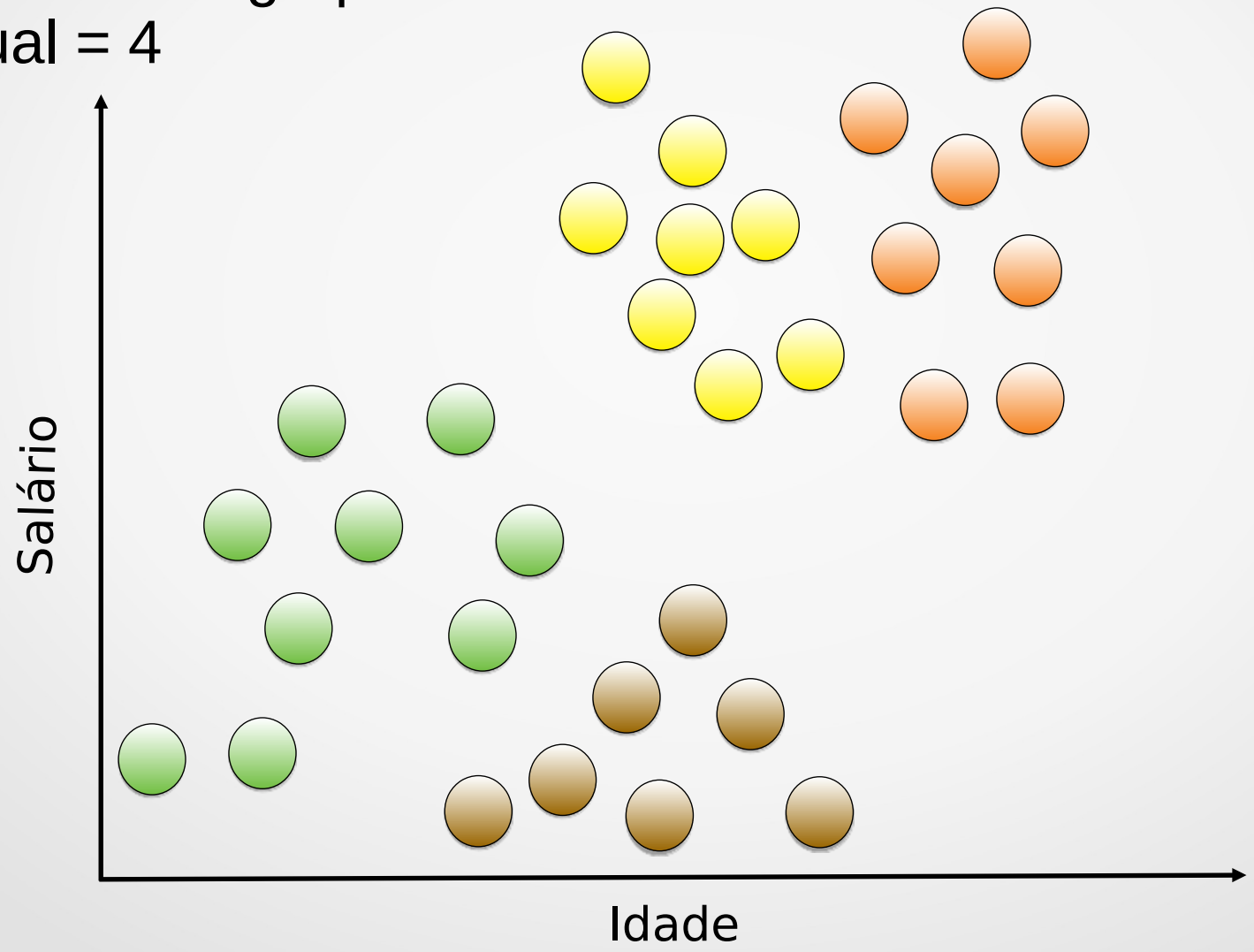
Algoritmo Aglomerativo

- Número de grupos atual = 6



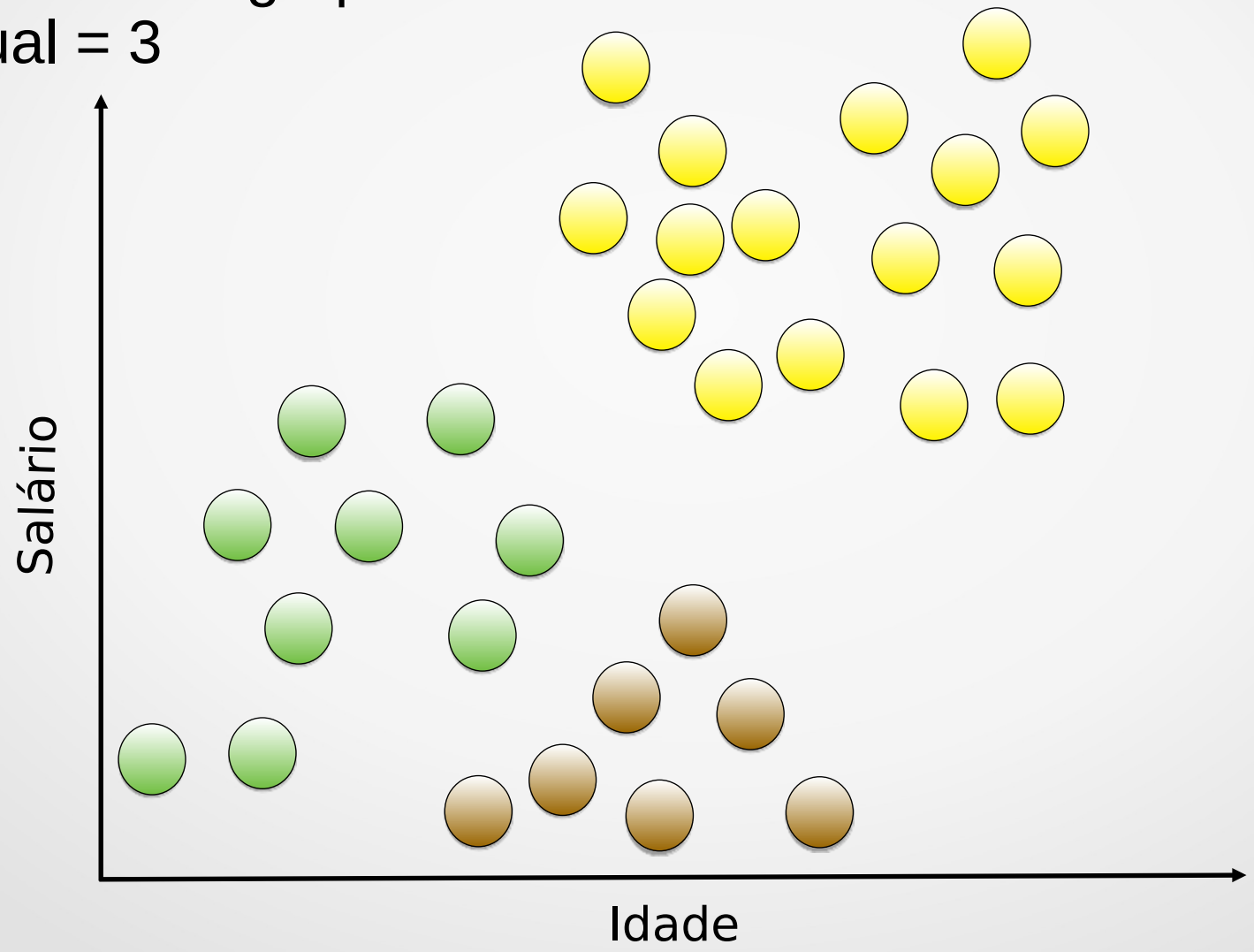
Algoritmo Aglomerativo

- Número de grupos atual = 4



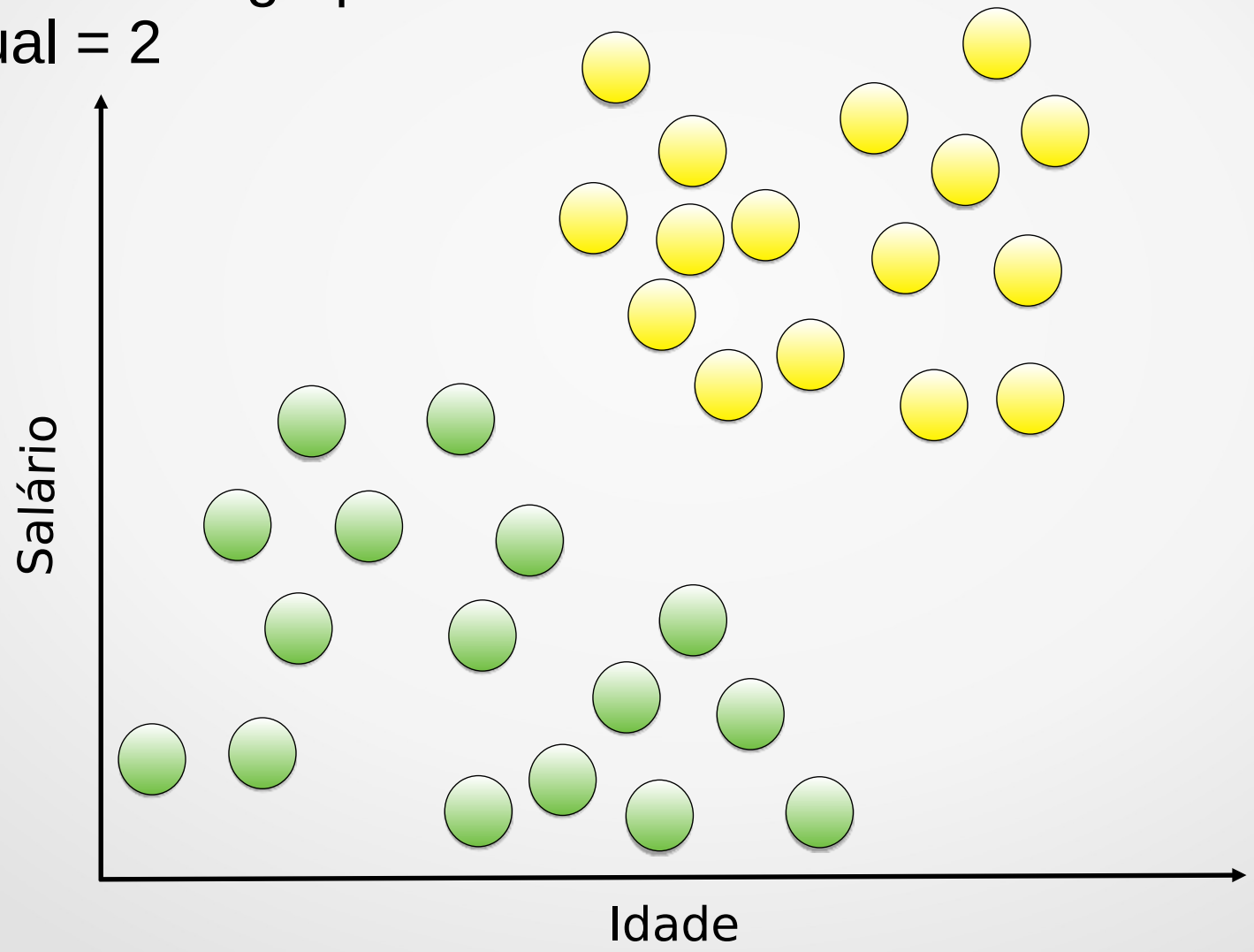
Algoritmo Aglomerativo

- Número de grupos atual = 3



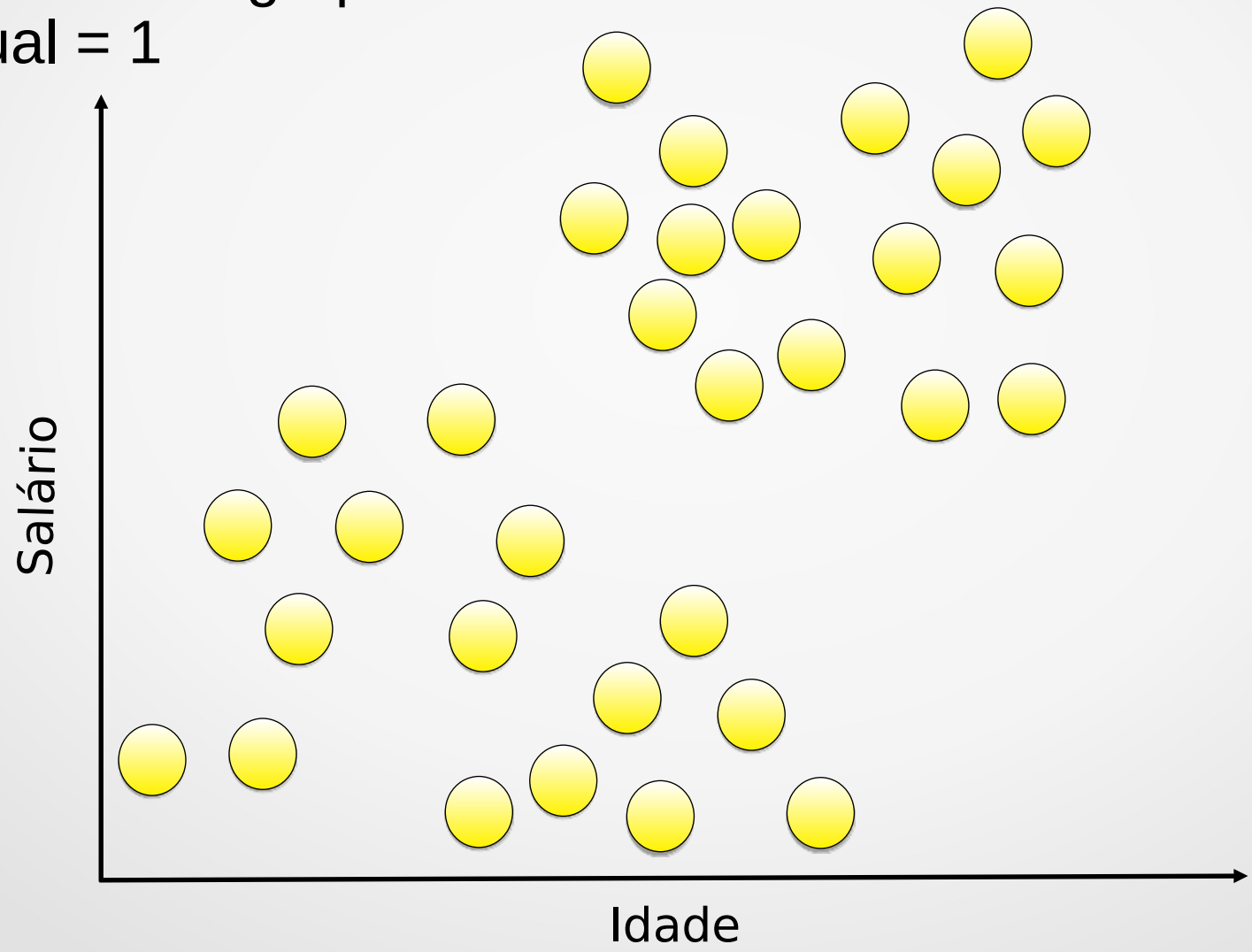
Algoritmo Aglomerativo

- Número de grupos atual = 2



Algoritmo Aglomerativo

- Número de grupos atual = 1



Definição de hierarquia

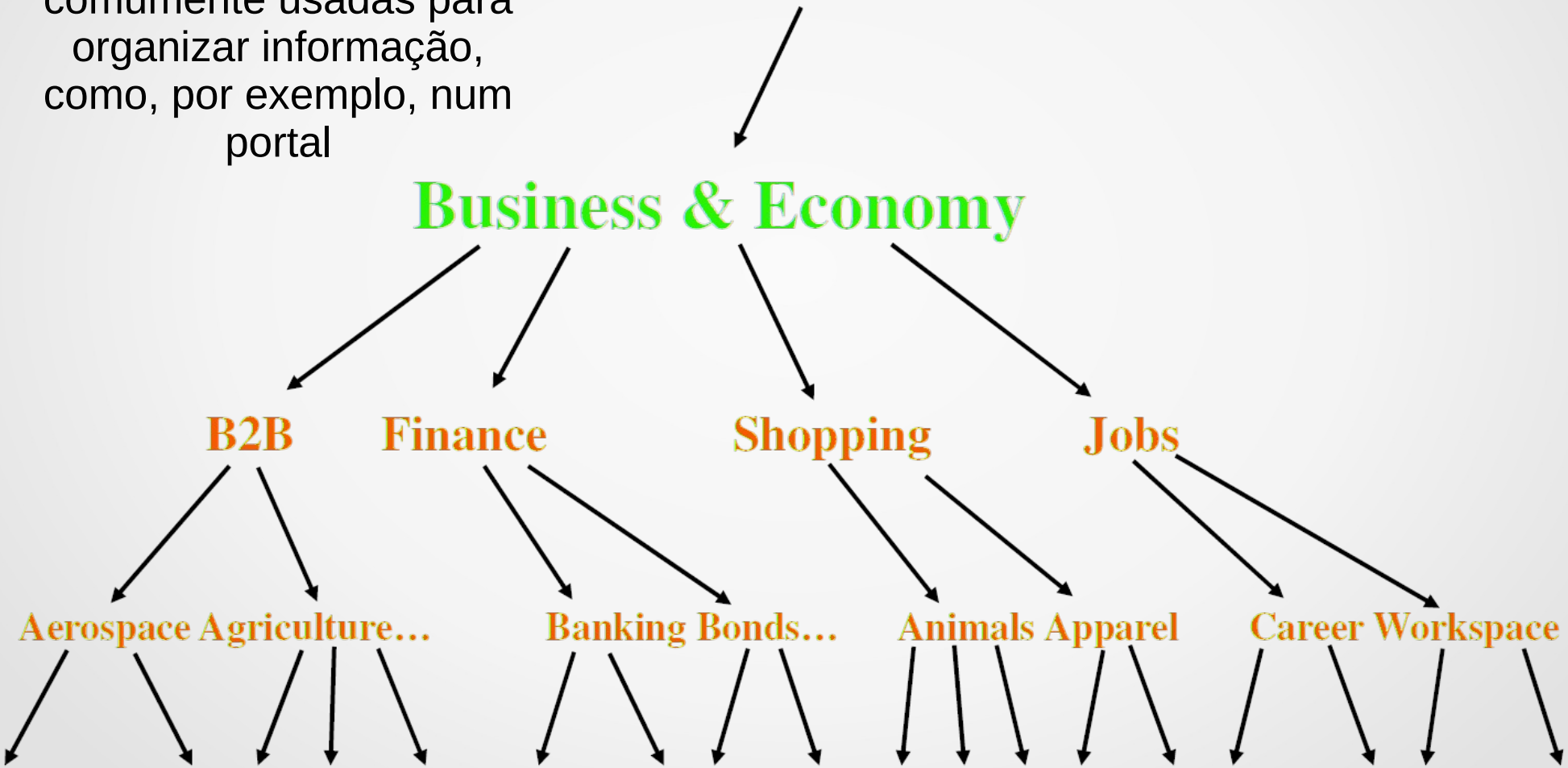
- Hierarquia (de partições de dados):
  - Sequencia de partições aninhadas
    - Uma partição P_1 está *aninhada* em P_2 se cada componente (grupo) de P_1 é um subconjunto de um componente de P_2
- Exemplo:
  - $P_1 = \{ (\mathbf{x}_1), (\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$
  - $P_2 = \{ (\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$
- Contra-Exemplo:
  - $P_2 = \{ (\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$
  - $P_3 = \{ (\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_5) \}$

Definição de hierarquia

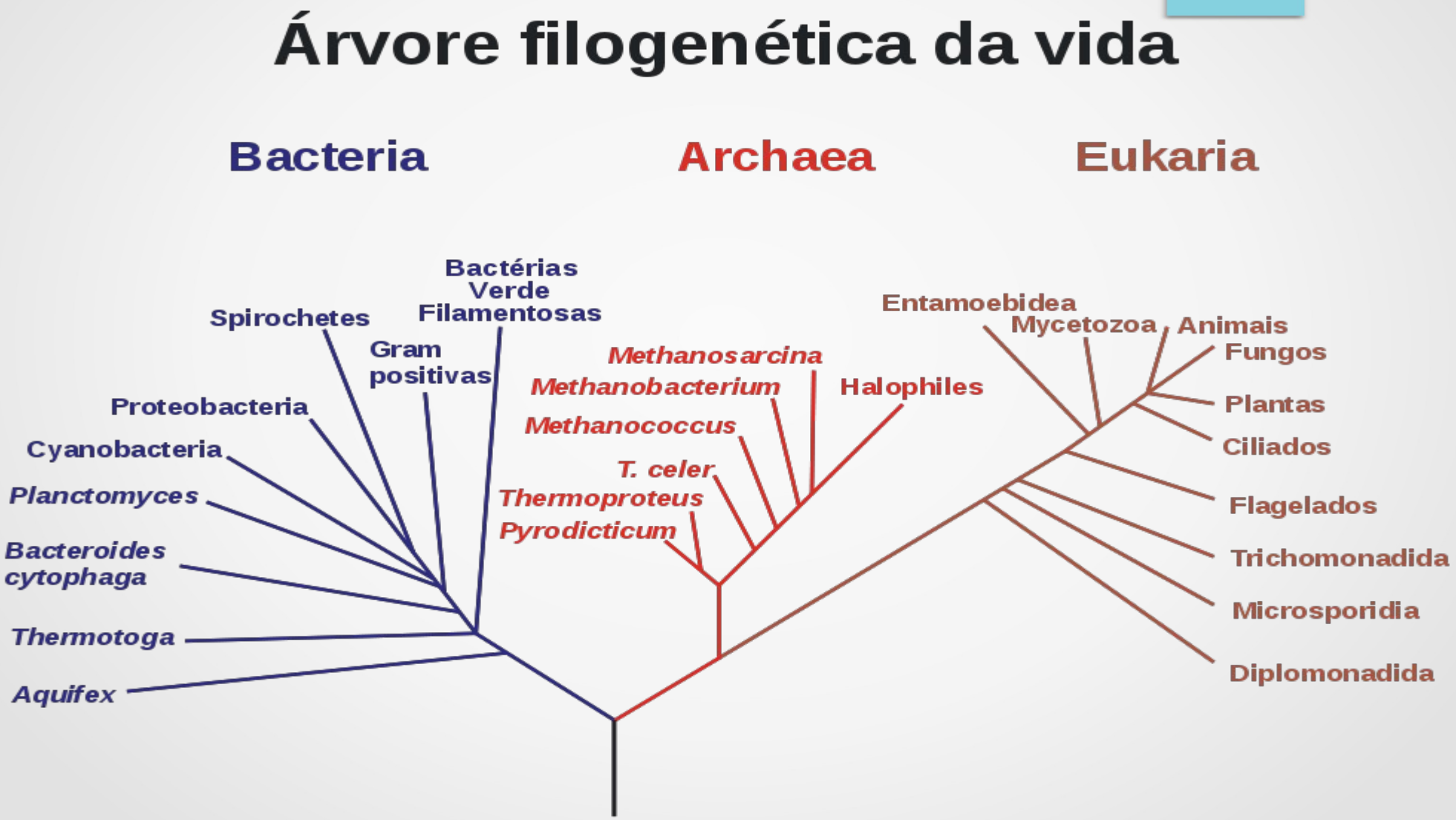
- Uma hierarquia completa:
  - Inicia ou termina com partição totalmente disjunta
    - Exemplo: P = { ($\mathbf{x}_1$), ($\mathbf{x}_2$), ($\mathbf{x}_3$), ($\mathbf{x}_4$), ($\mathbf{x}_5$), ($\mathbf{x}_6$)}
    - Também denominada “solução trivial”
  - Inicia ou termina com partição totalmente conjunta
    - Exemplo: P = { ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, $\mathbf{x}_6$) }
  - Geralmente possui N – 2 partições intermediárias

Hierarquia

Hierarquias são comumente usadas para organizar informação, como, por exemplo, num portal



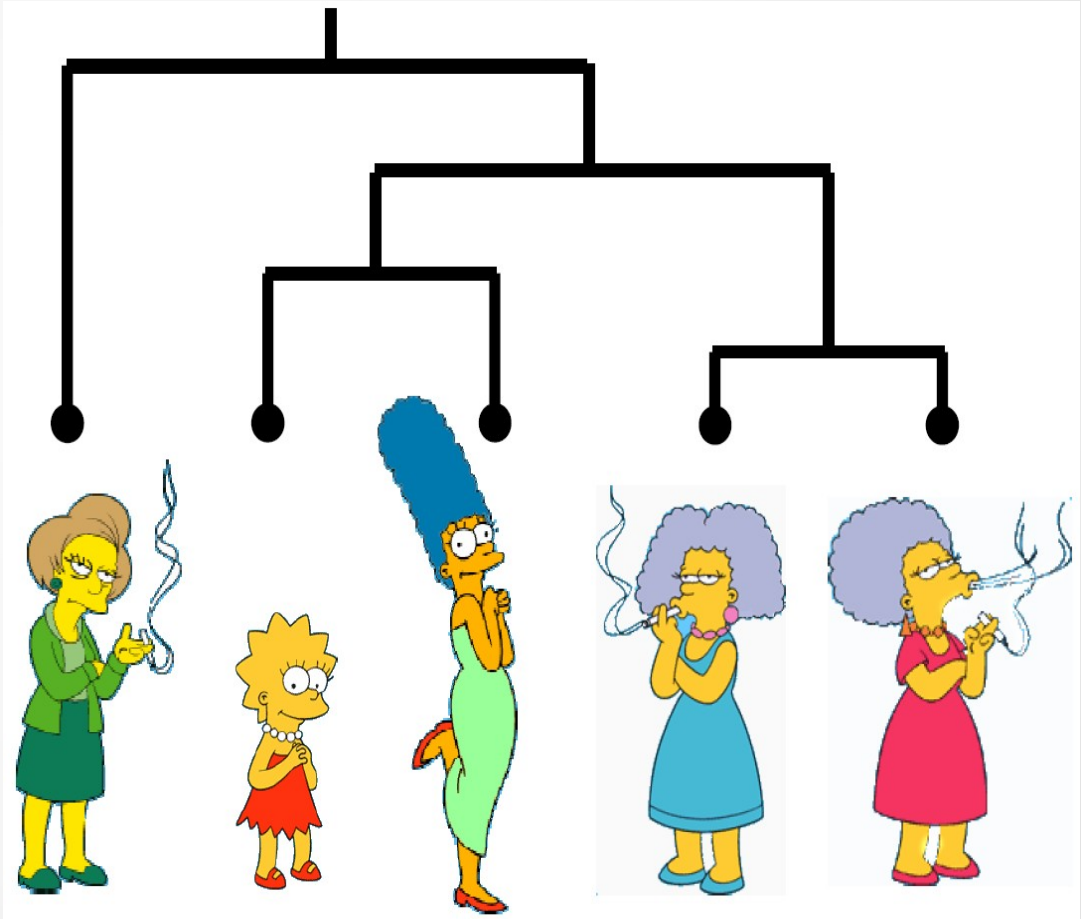
Outro exemplo de hierarquia



http://pt.wikipedia.org/wiki/%C3%81rvore_filogen%C3%A9tica

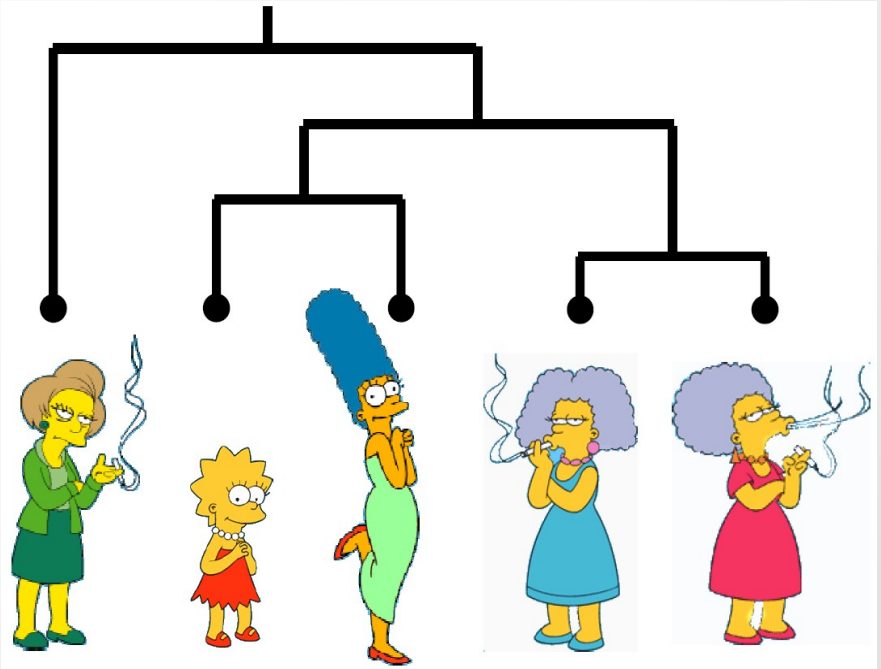
Métodos para agrupamento hierárquico

- Bottom-Up (aglomerativos):
  - Iniciar colocando cada objeto em um grupo
  - Encontrar o melhor par de grupos para unir
  - Unir o par de grupos escolhido



Métodos para agrupamento hierárquico

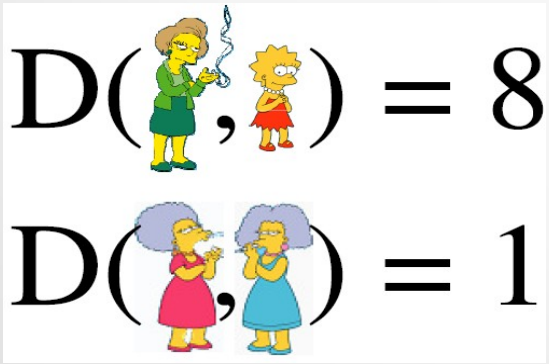
- Top-Down (divisivos):
  - Iniciar com todos objetos em um único grupo
  - Sub-dividir o grupo em dois novos grupos
  - Aplicar o algoritmo recursivamente, até que cada objeto forme um grupo por si só

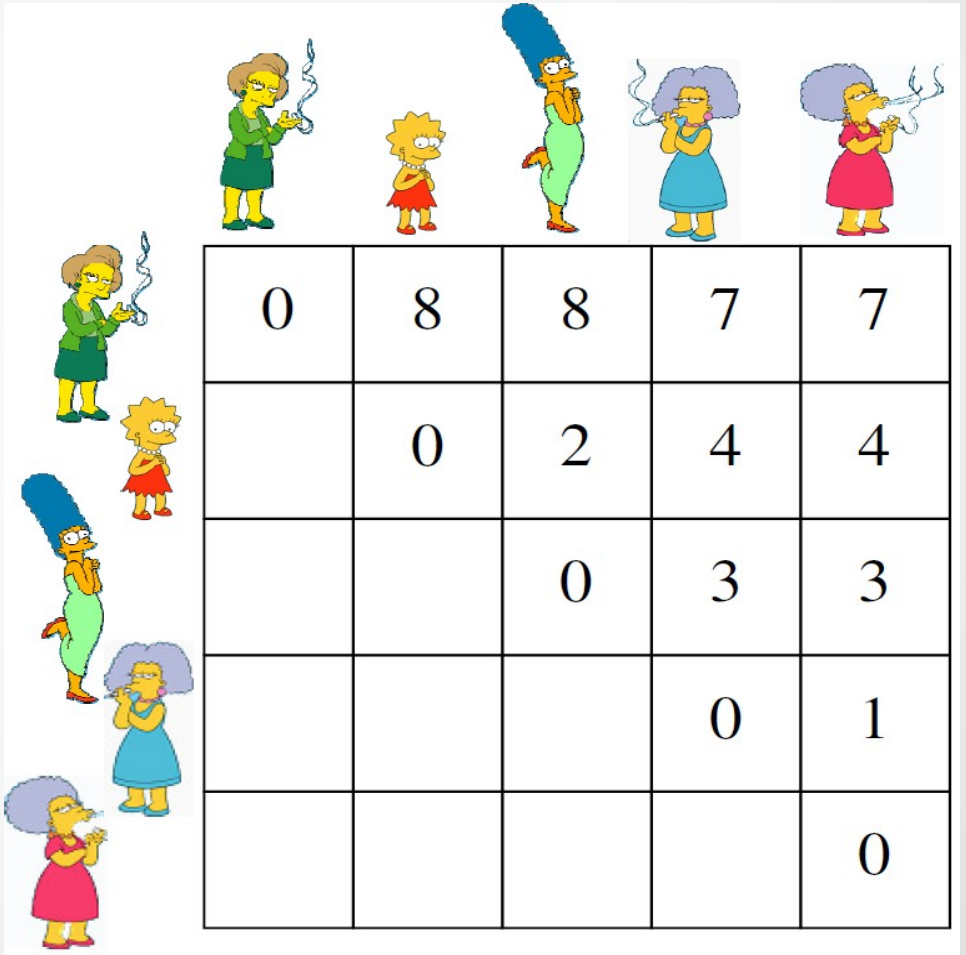


Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Algoritmo relacional

- Algoritmos hierárquicos podem operar somente sobre uma matriz de dissimilaridades: são (ou podem ser) relacionais !

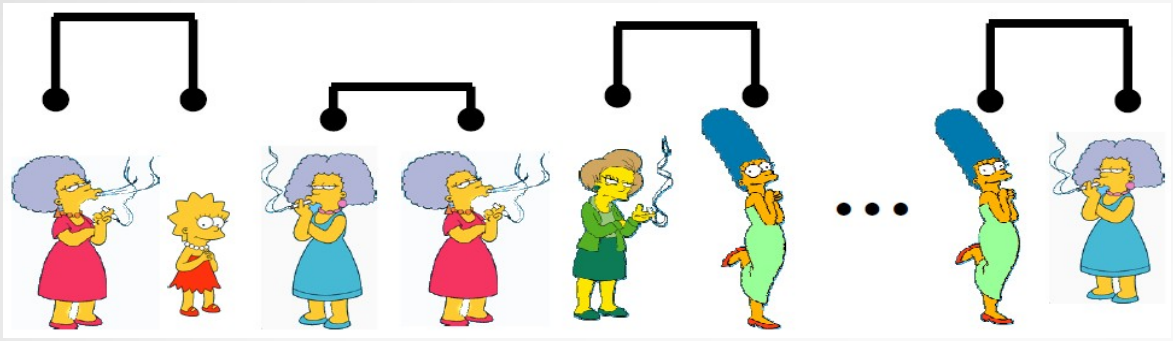




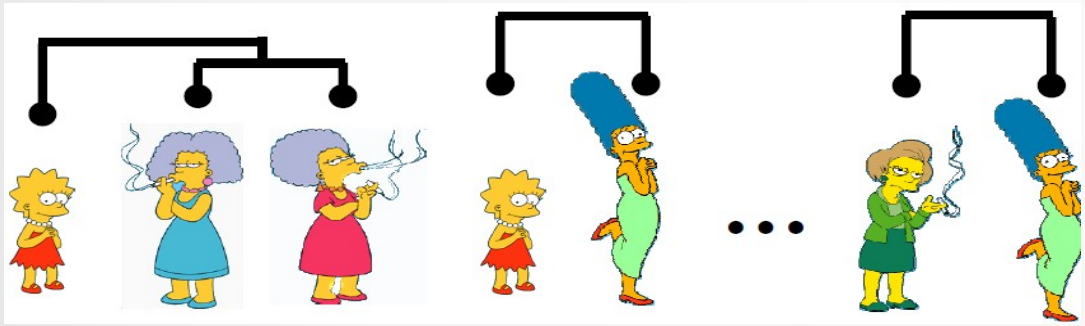
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

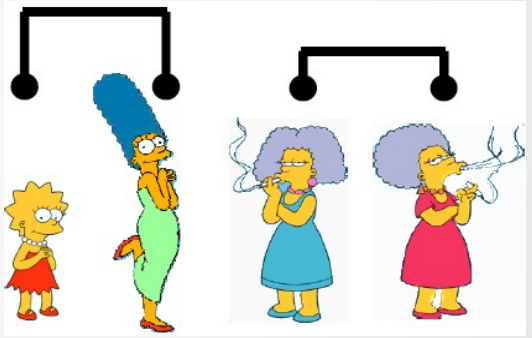
Exemplo

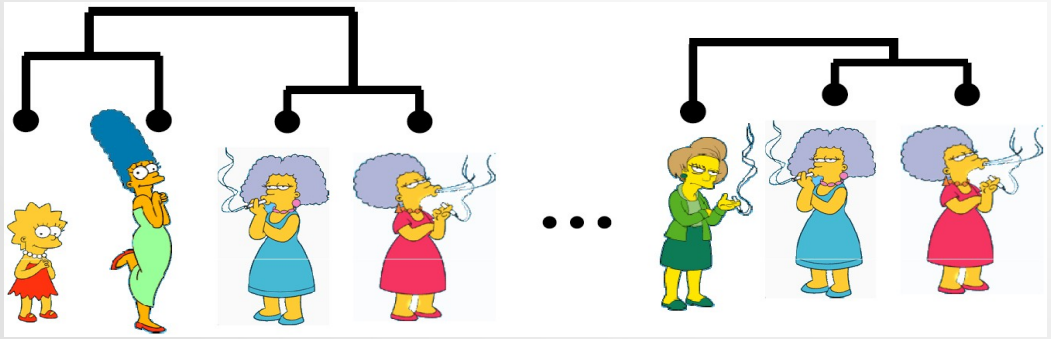


Mais similar!

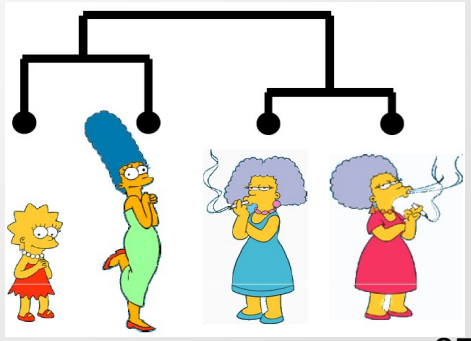


Mais similar!





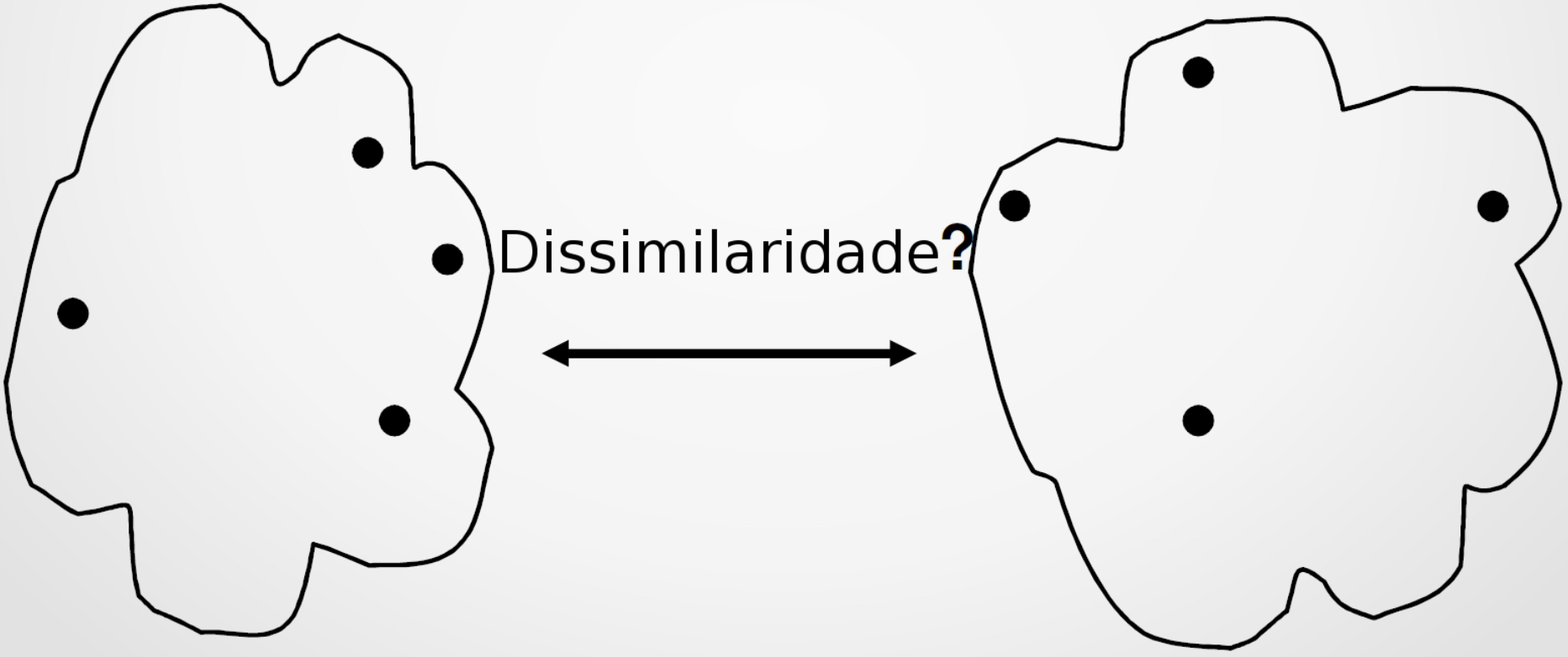
Mais similar!



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como definir dissimilaridade entre grupos?

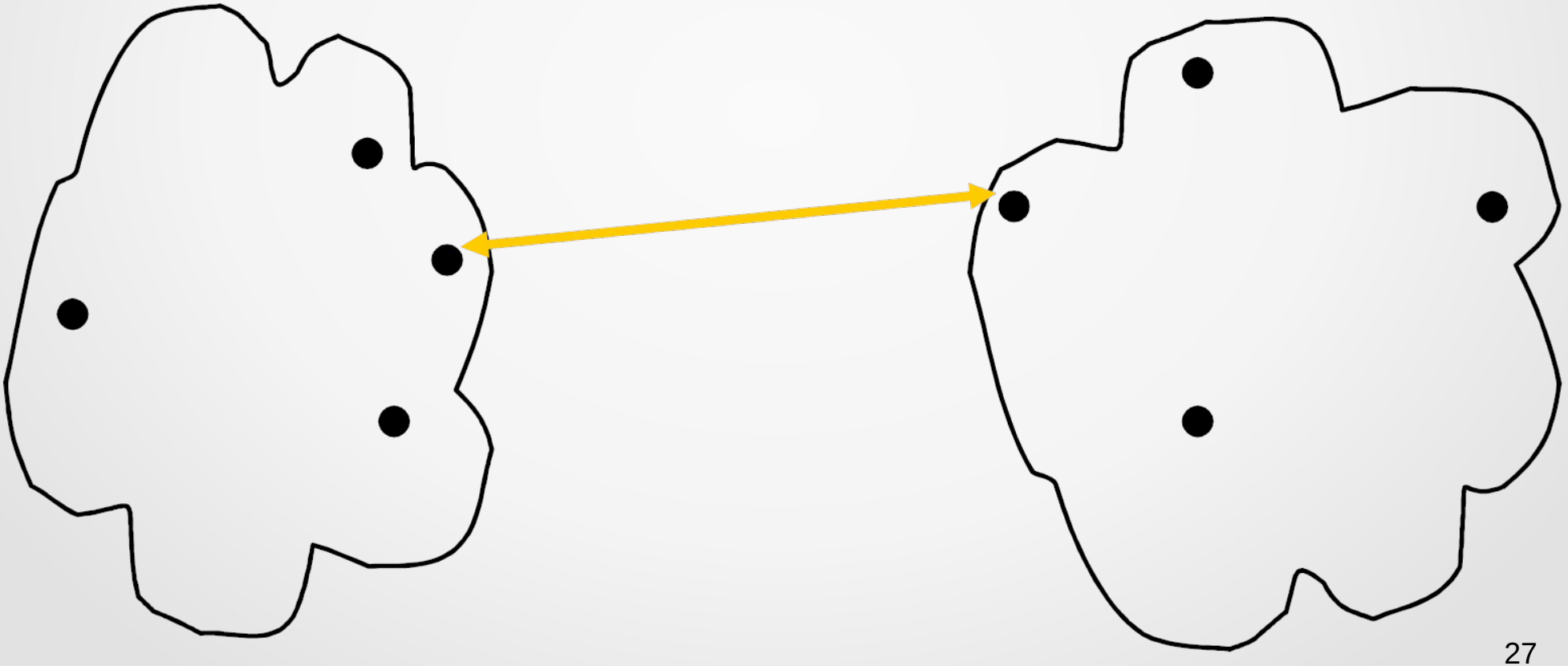
- Mínima
- Máxima
- Média
- Centroides
- Outras...



www-users.cs.umn.edu/~kumar/dmbook/index.php

Ligação Simples

- **Mínima**
- Máxima
- Média
- Centroides
- Outras...



www-users.cs.umn.edu/~kumar/dmbook/index.php

Vantagem ligação simples

- É apropriada para encontrar grupos formados por densidades

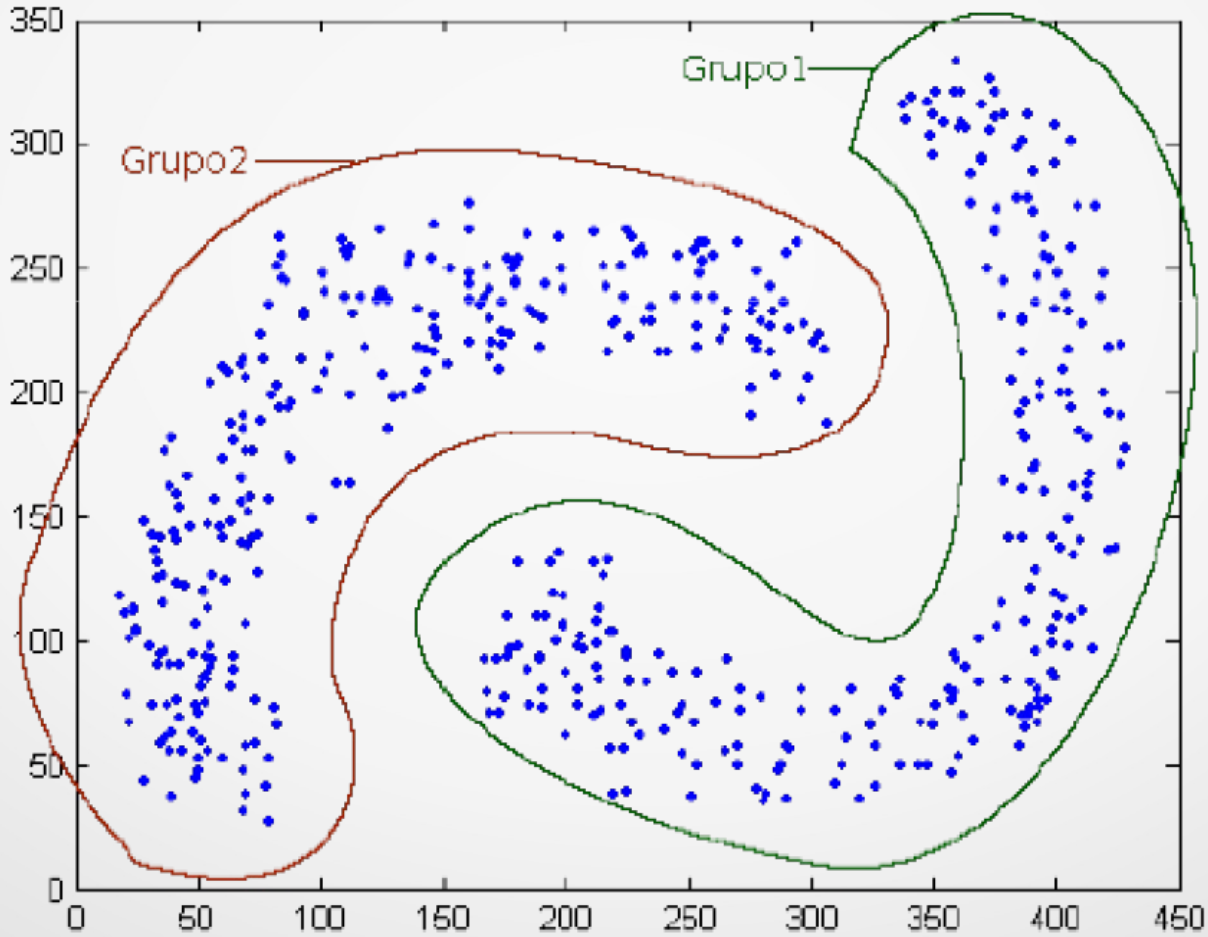
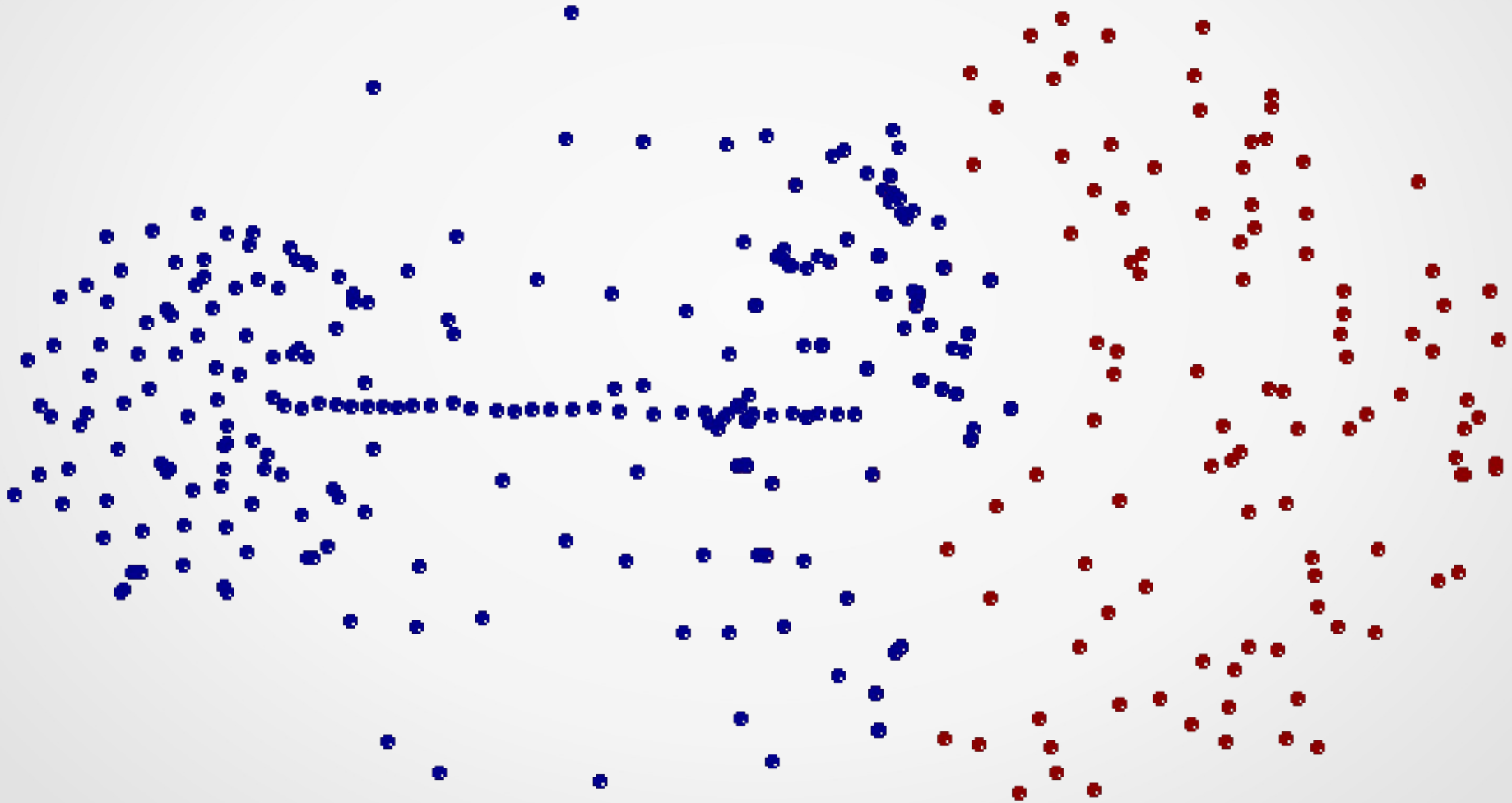


Figura por Lucas Vendramin

Desvantagem ligação simples

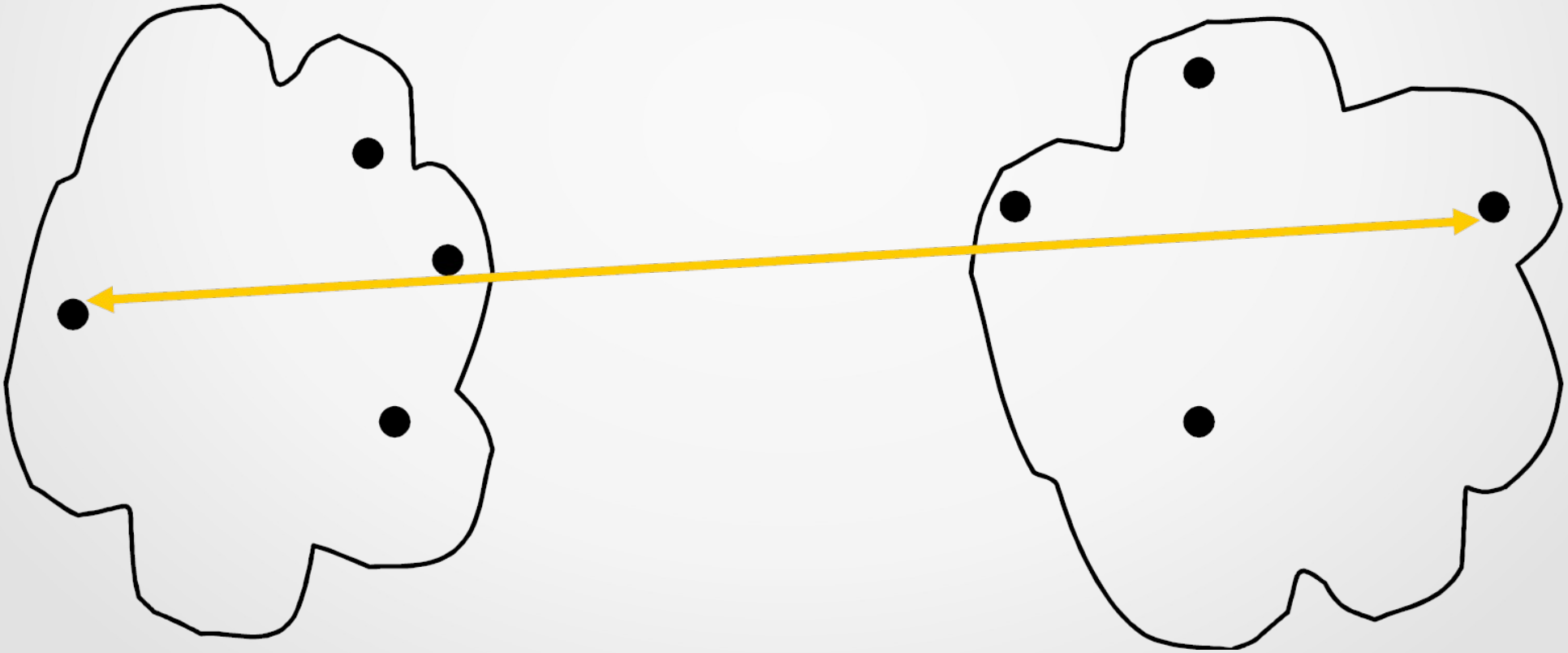
- Sensível a ruídos e *outliers*



www-users.cs.umn.edu/~kumar/dmbook/index.php

Ligação Completa

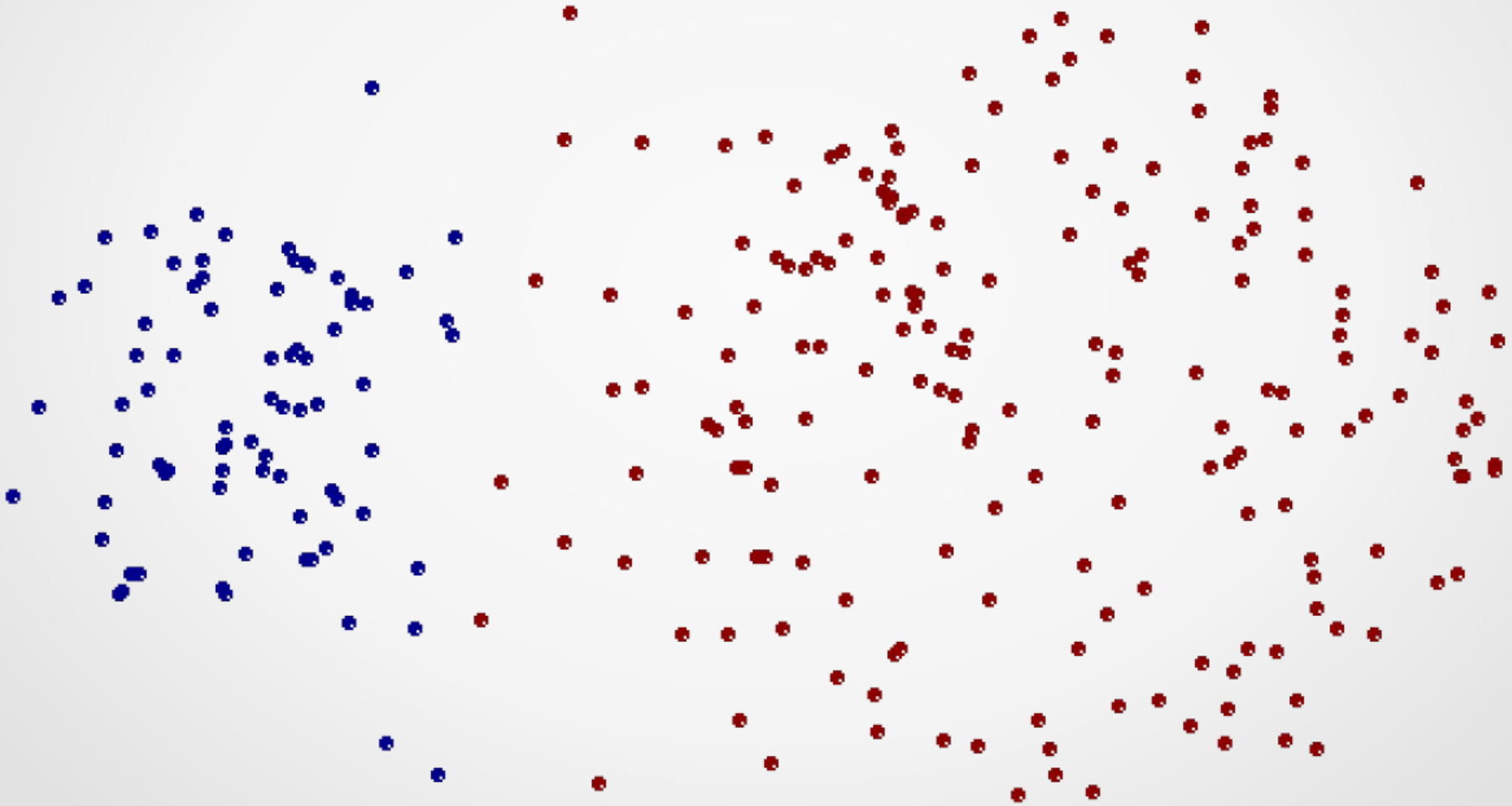
- Mínima
- **Máxima**
- Média
- Centroides
- Outras...



www-users.cs.umn.edu/~kumar/dmbook/index.php

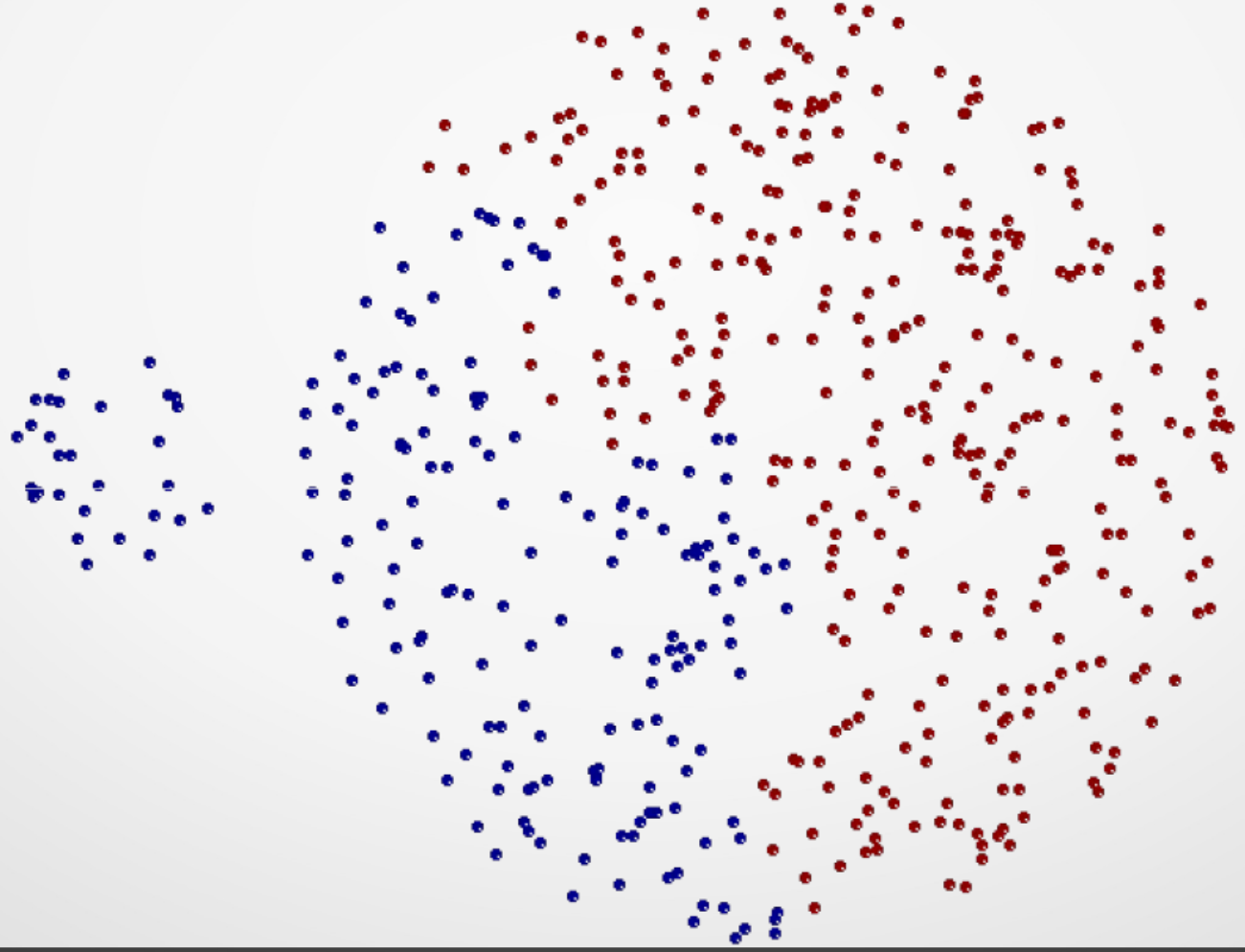
Vantagem ligação completa

- Menos sensível a ruídos e *outliers*



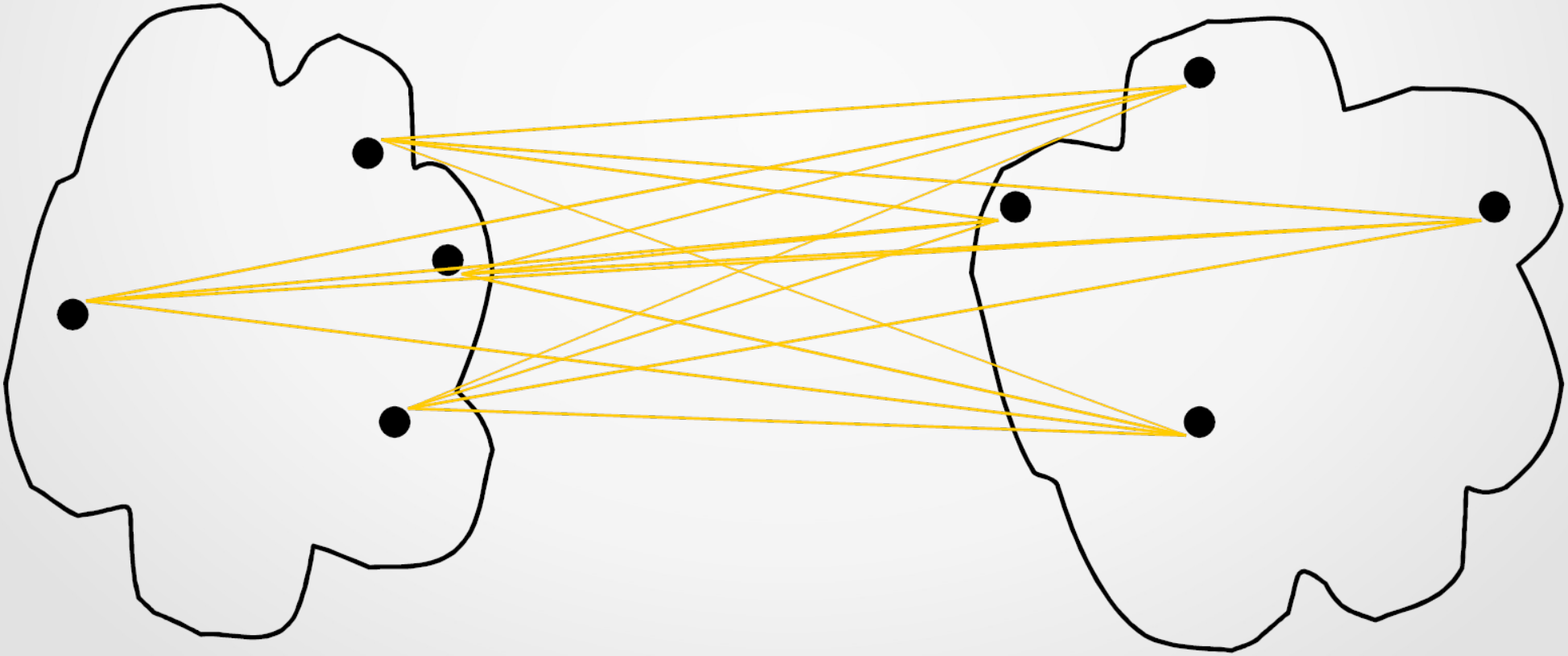
Desvantagem ligação completa

- Tende a quebrar grupos grandes
- Só funciona bem com grupos globulares



Ligação Média

- Mínima
- Máxima
- **Média**
- Centroides
- Outras...



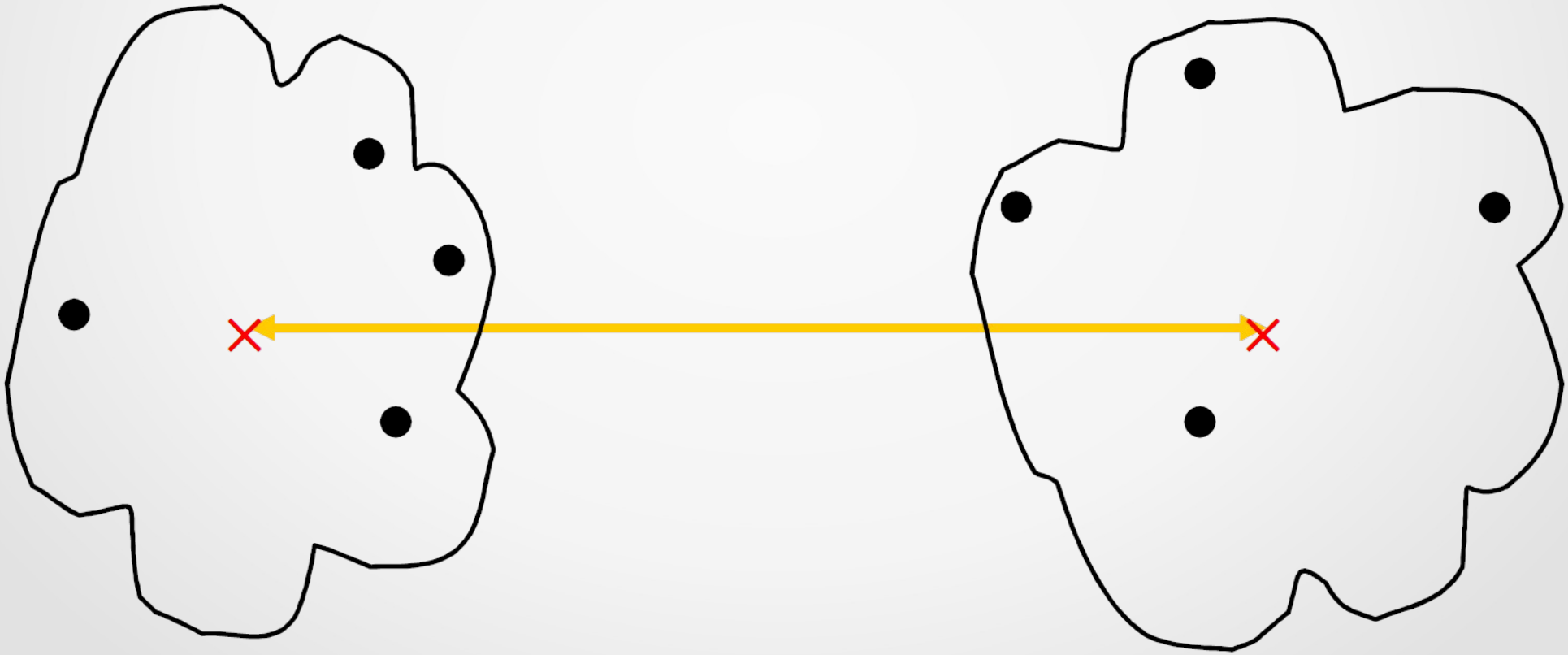
www-users.cs.umn.edu/~kumar/dmbook/index.php

Vantagens e desvantagens

- Ligação média consiste em calcular a média das distâncias entre todos os objetos dos dois grupos
- Utilizar ligação média obtêm resultados com características intermediárias entre o algoritmo de ligação simples e completa
  - Formatos globulares
  - Menos susceptível a ruídos e *outliers*

Uso de Centroides

- Mínima
- Máxima
- Média
- **Centroides**
- Outras...



www-users.cs.umn.edu/~kumar/dmbook/index.php

Vantagens e desvantagens

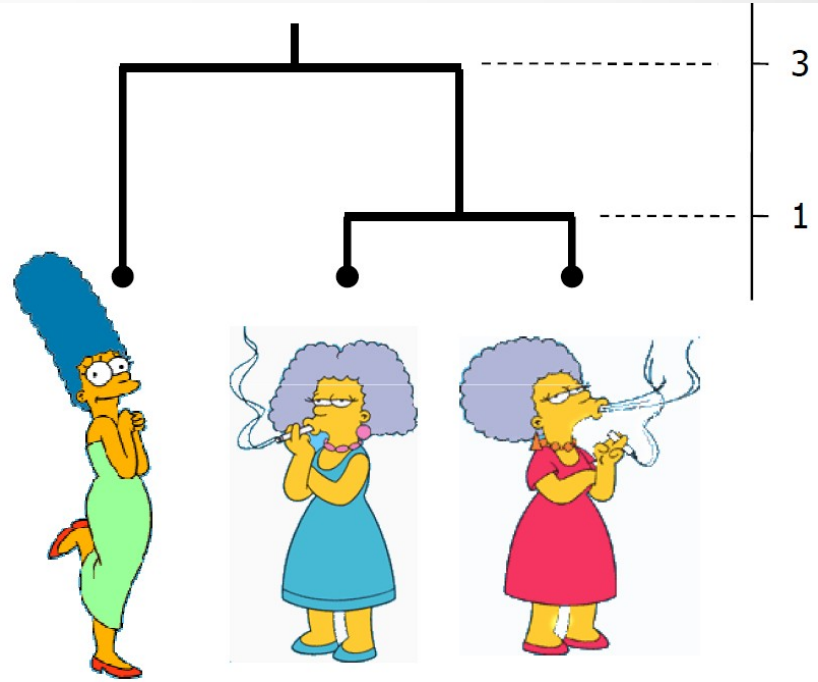
- O centroide do grupo resultante de uma união é calculado como uma média dos Centroides dos dois grupos originais ponderada pela quantidade de objetos nos grupos
- Mas perde-se informação sobre o formato dos grupos
- É limitado a distância Euclidiana e faz o algoritmo deixar de ser relacional

Outras...

- *Ward's*
  - Procura formar variância mínima entre os grupos gerados
- *Bisecting k-means*
  - Algoritmo hierárquico divisivo baseado em k-médias, sempre utilizado para dividir partições em duas

Dendrograma

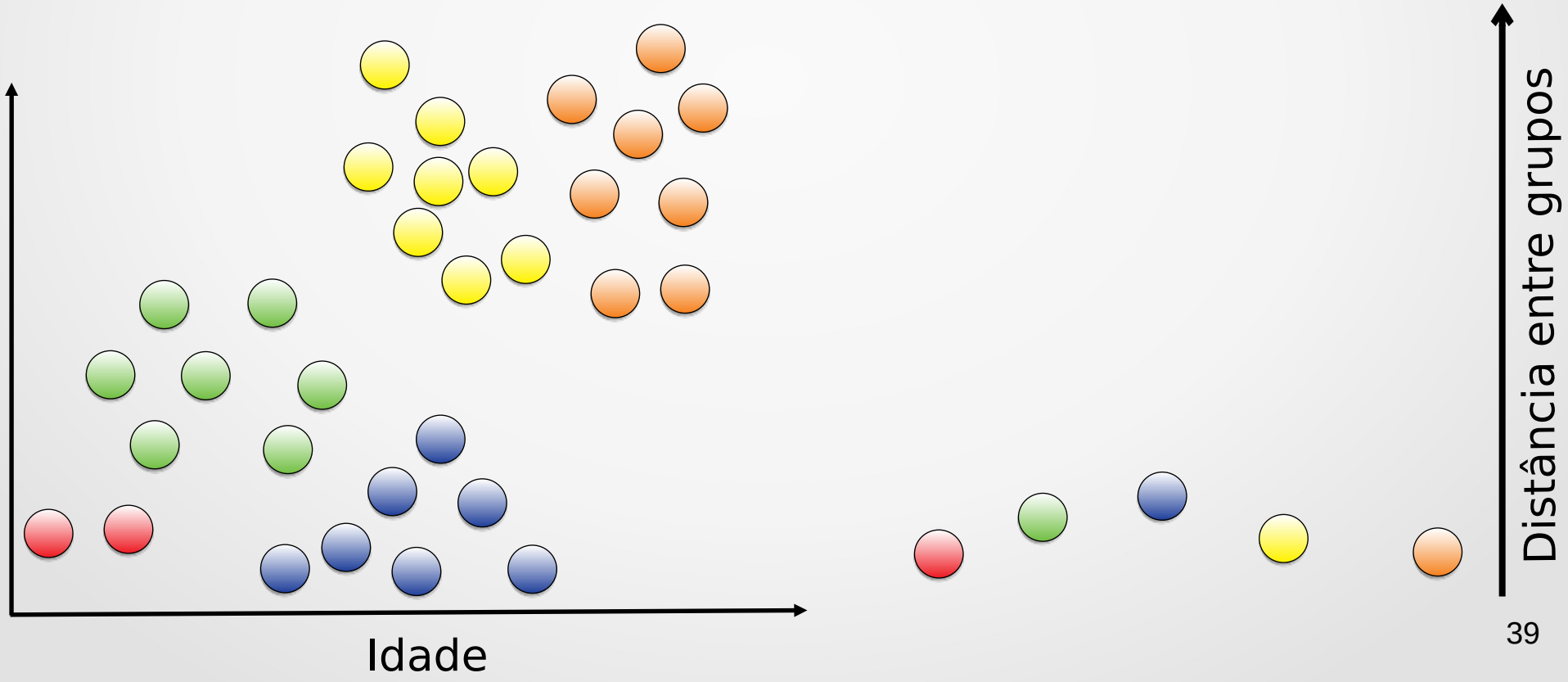
- Forma de representação tipica de um agrupamento hierárquico
- Muito importante pois possui:
  - Hierarquia
  - Dissimilaridades
- A dissimilaridade entre dois grupos é apresentada como a altura do nó interno mais baixo compartilhado



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

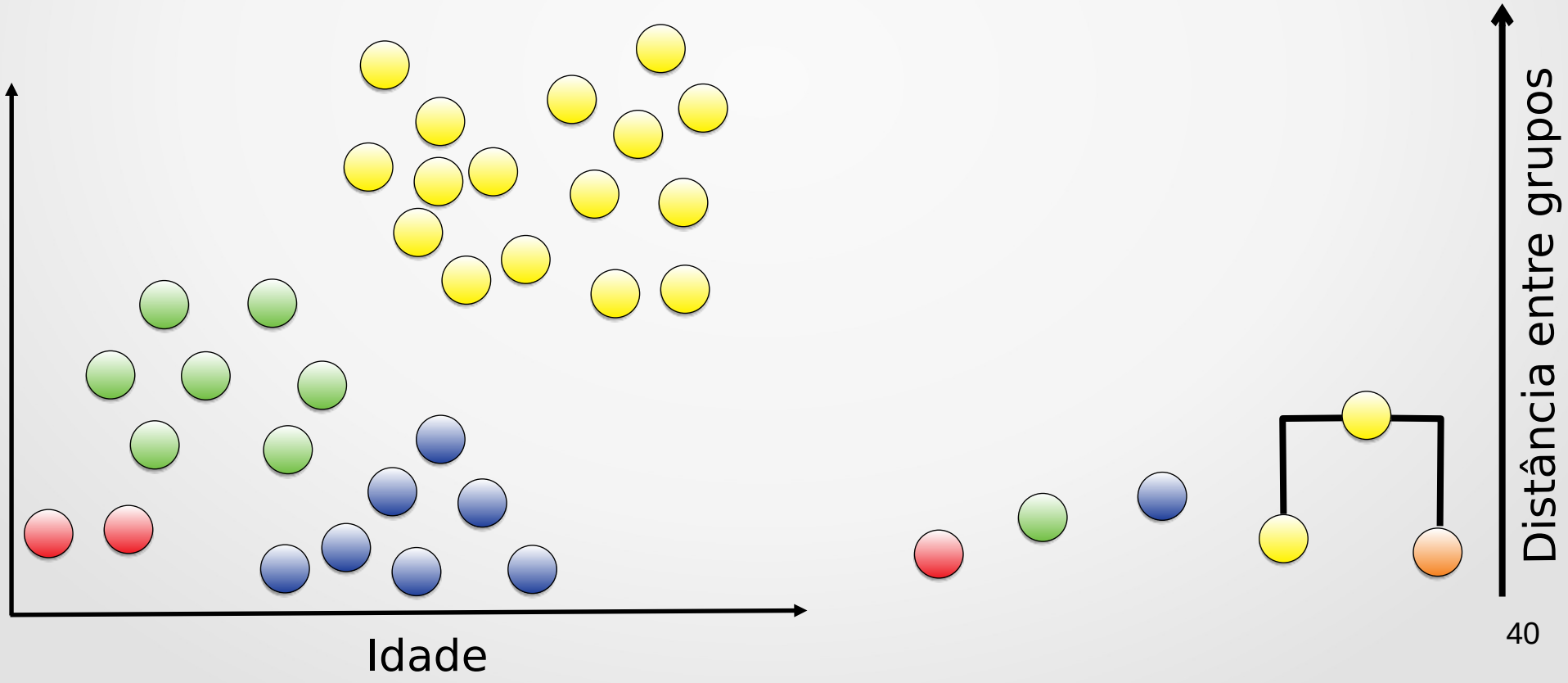
Exemplo de dendrograma

- Dendrograma deve refletir a junção dos grupos e distância onde ela ocorreu
  - Exemplo de dendrograma parcial (5+ grupos)



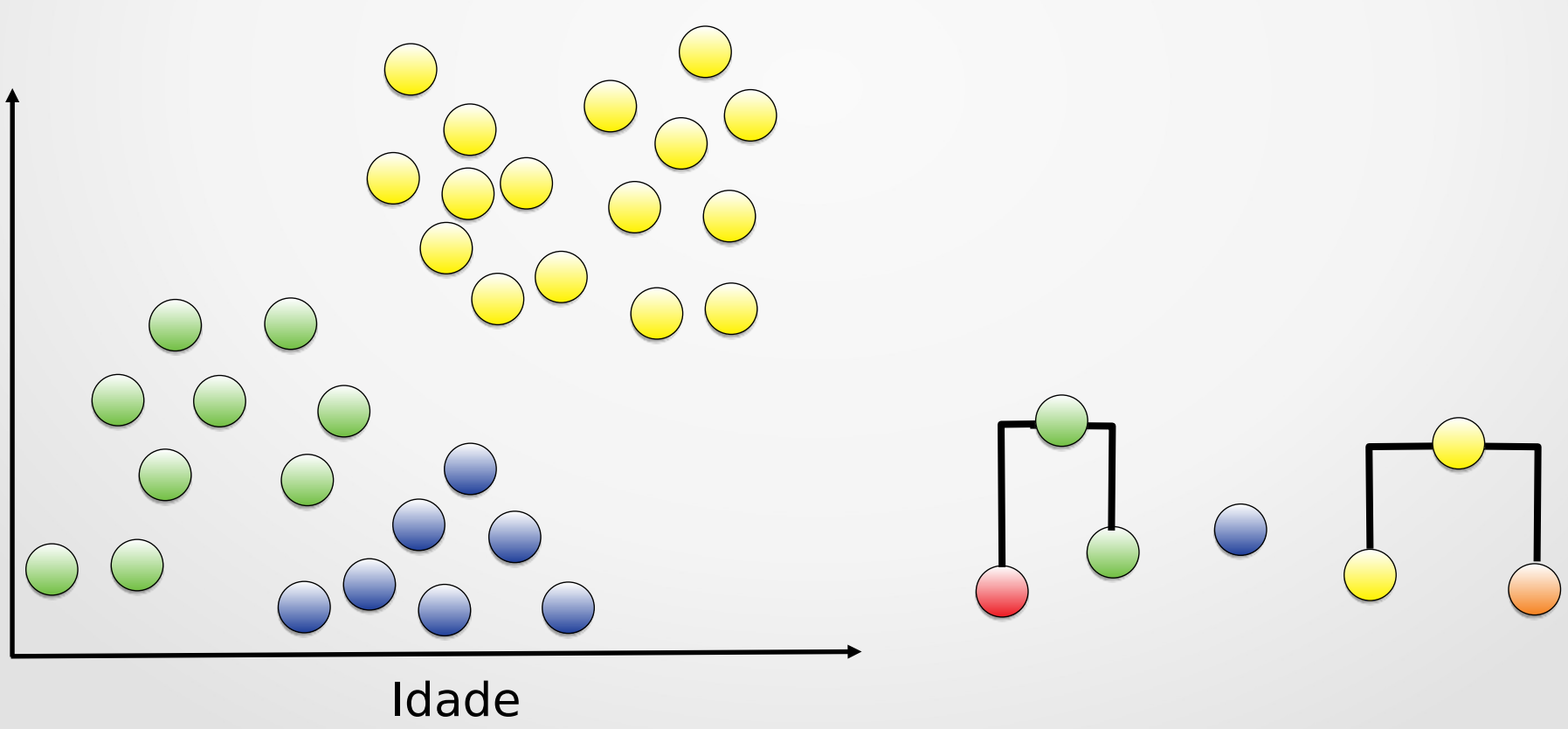
Exemplo de dendrograma

- Junção dos grupos mais próximos, segundo a medida de dissimilaridade escolhida



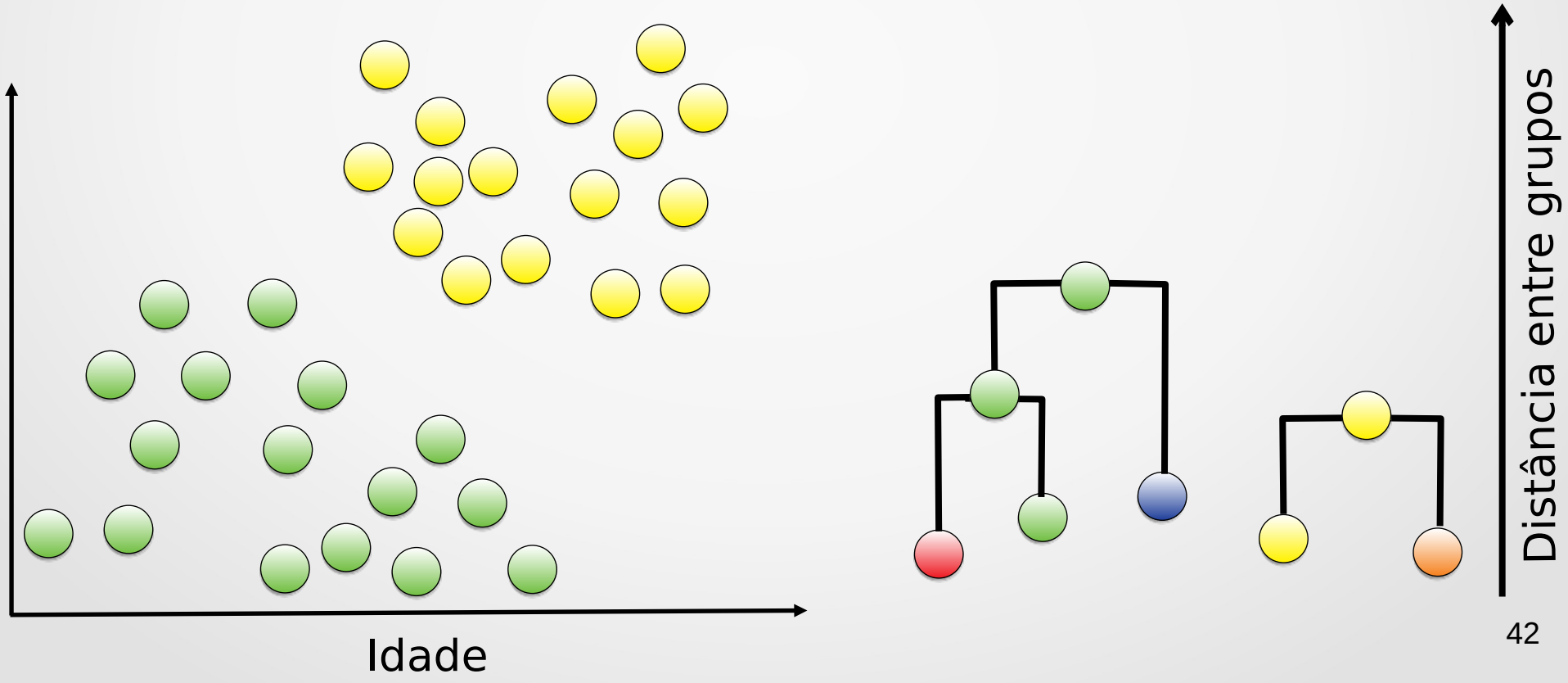
Exemplo de dendrograma

- Junção dos grupos mais próximos, segundo a medida de dissimilaridade escolhida



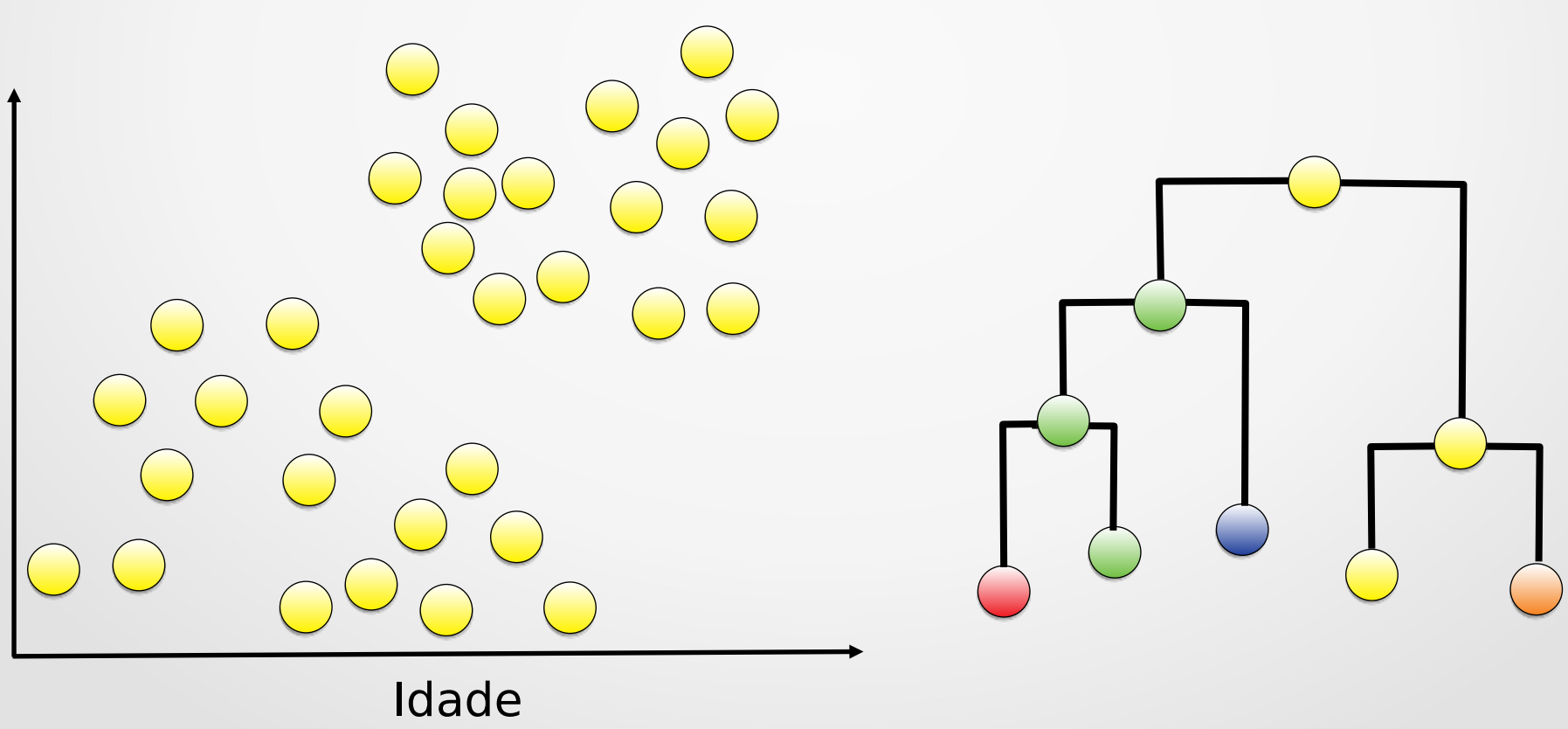
Exemplo de dendrograma

- Junção dos grupos mais próximos, segundo a medida de dissimilaridade escolhida



Exemplo de dendrograma

- Junção dos grupos mais próximos, segundo a medida de dissimilaridade escolhida

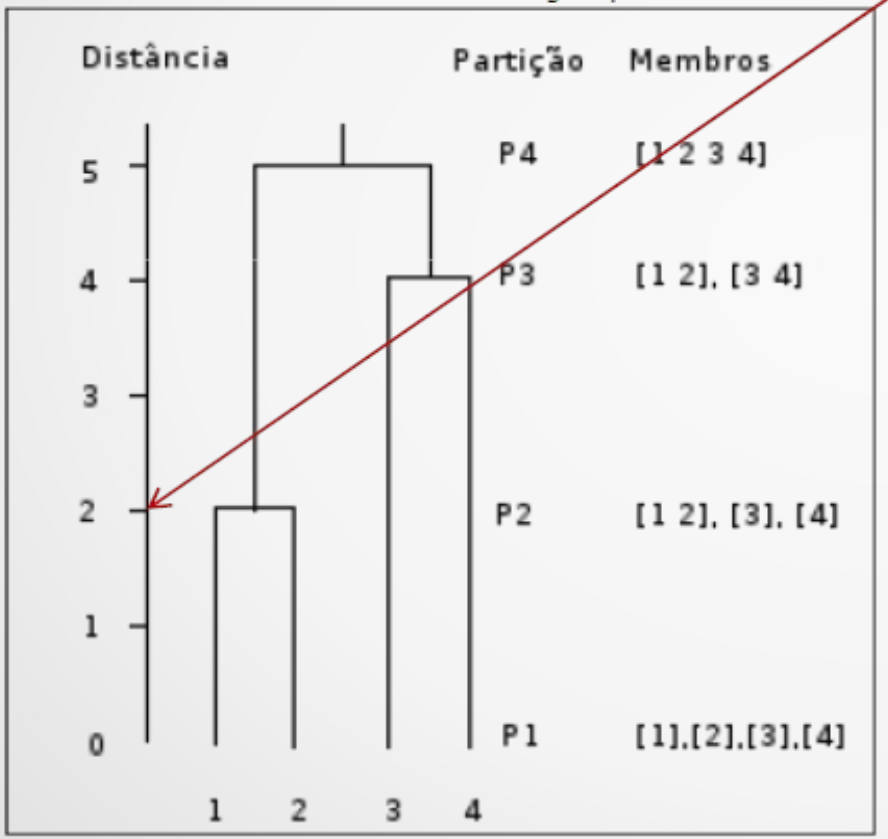


Exemplo Dendrograma

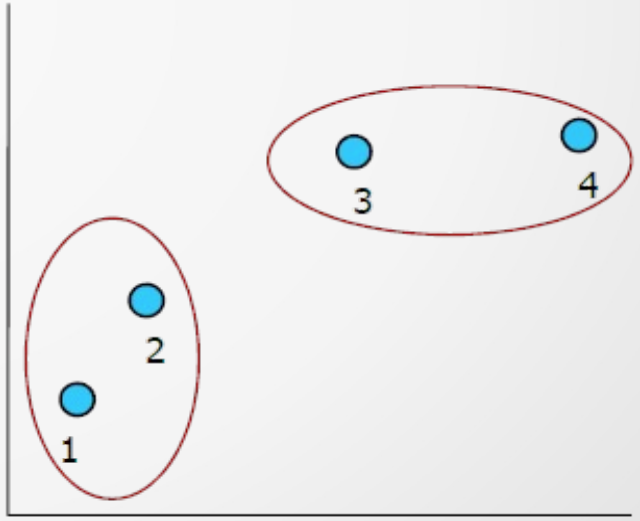
- Pode ser derivado diretamente da matriz de distâncias

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 7 & 13 \\ 2 & 0 & 5 & 10 \\ 7 & 5 & 0 & 4 \\ 13 & 10 & 4 & 0 \end{bmatrix}$$

Figura por Lucas Vendramin

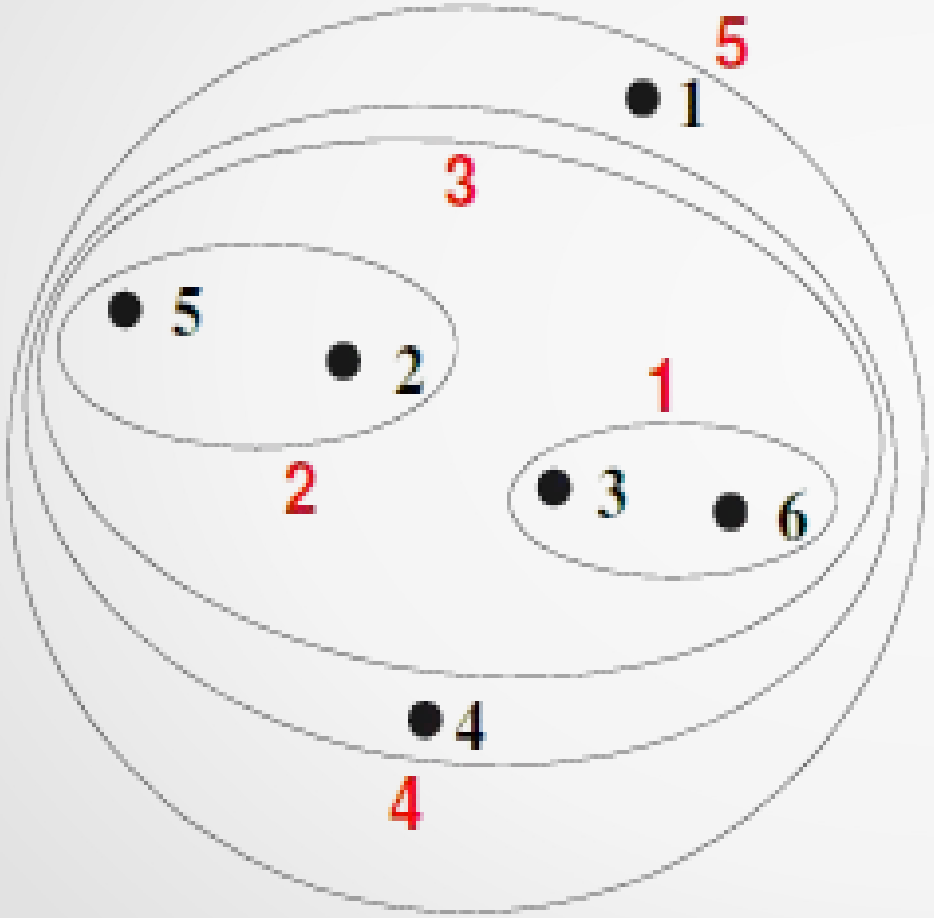


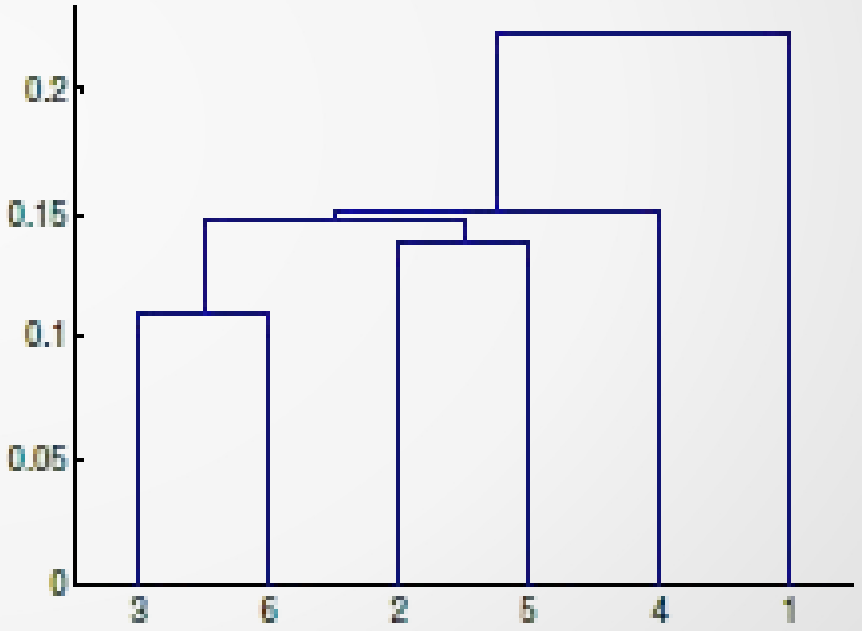
Dendrograma



uma das partições aninhadas

Outro exemplo



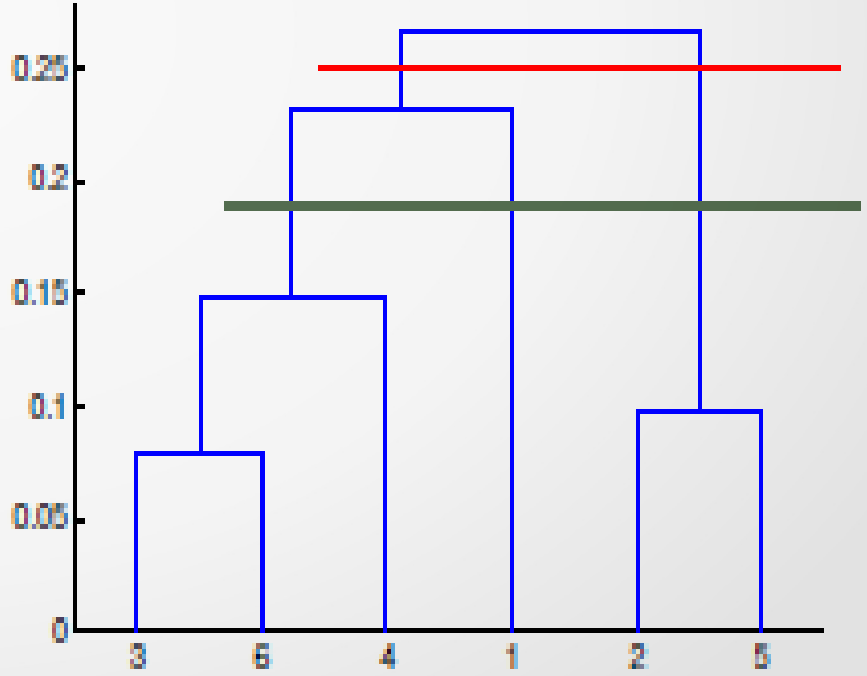


Como obter partição?

- Partições são obtidas via cortes no dendrograma
  - cortes horizontais
  - no. de grupos da partição = no. de interseções
- Exemplos:

$P_1 = \{ (\mathbf{x}_1), (\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$

$P_2 = \{ (\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$

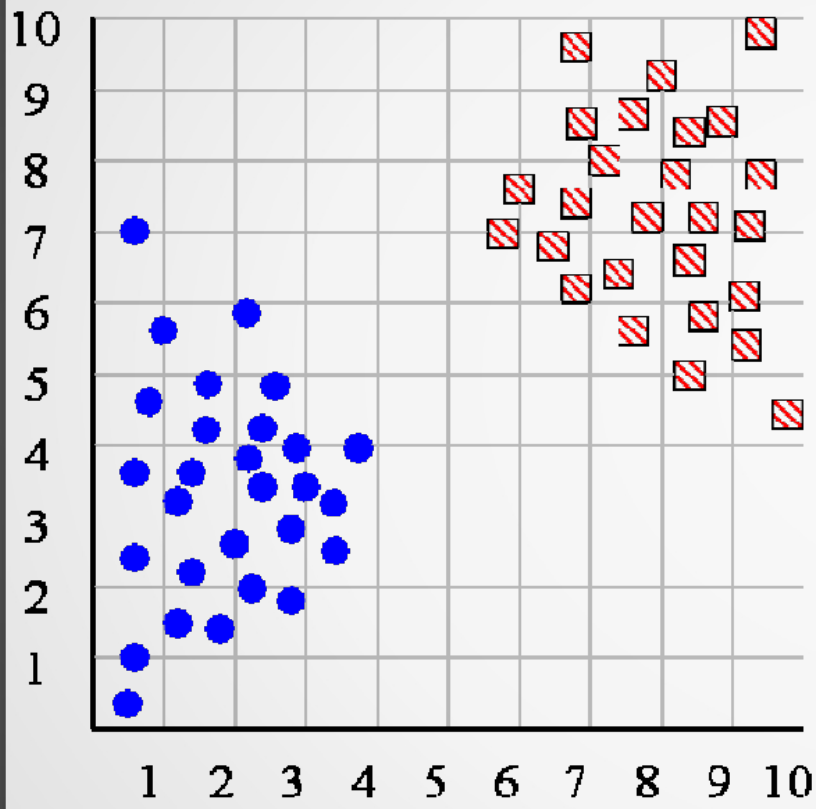


Como fazer o Dendrograma?

- Gerar o gráfico automaticamente demanda ordenar de forma apropriada os objetos no eixo horizontal
- Algoritmo Recursivo Simples:
  - Iniciar com o topo da hierarquia (grupo único)
  - Dividir o eixo horizontal em 2 subintervalos e colocar em cada um
  - Os objetos de cada um dos 2 grupos que derivam do grupo único
  - Executar recursivamente o passo anterior para cada subintervalo

Dendrograma e análise de grupos

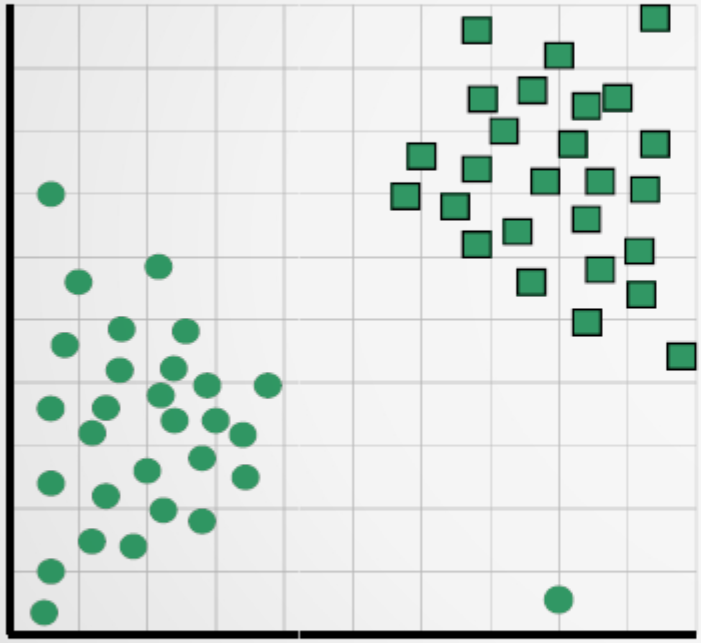
- Pode indicar o número natural de grupos



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Dendrograma e *outliers*

- Também usado para tentar detectar *outliers*



Outlier

Ramo isolado sugere que o objeto é muito diferente dos demais.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

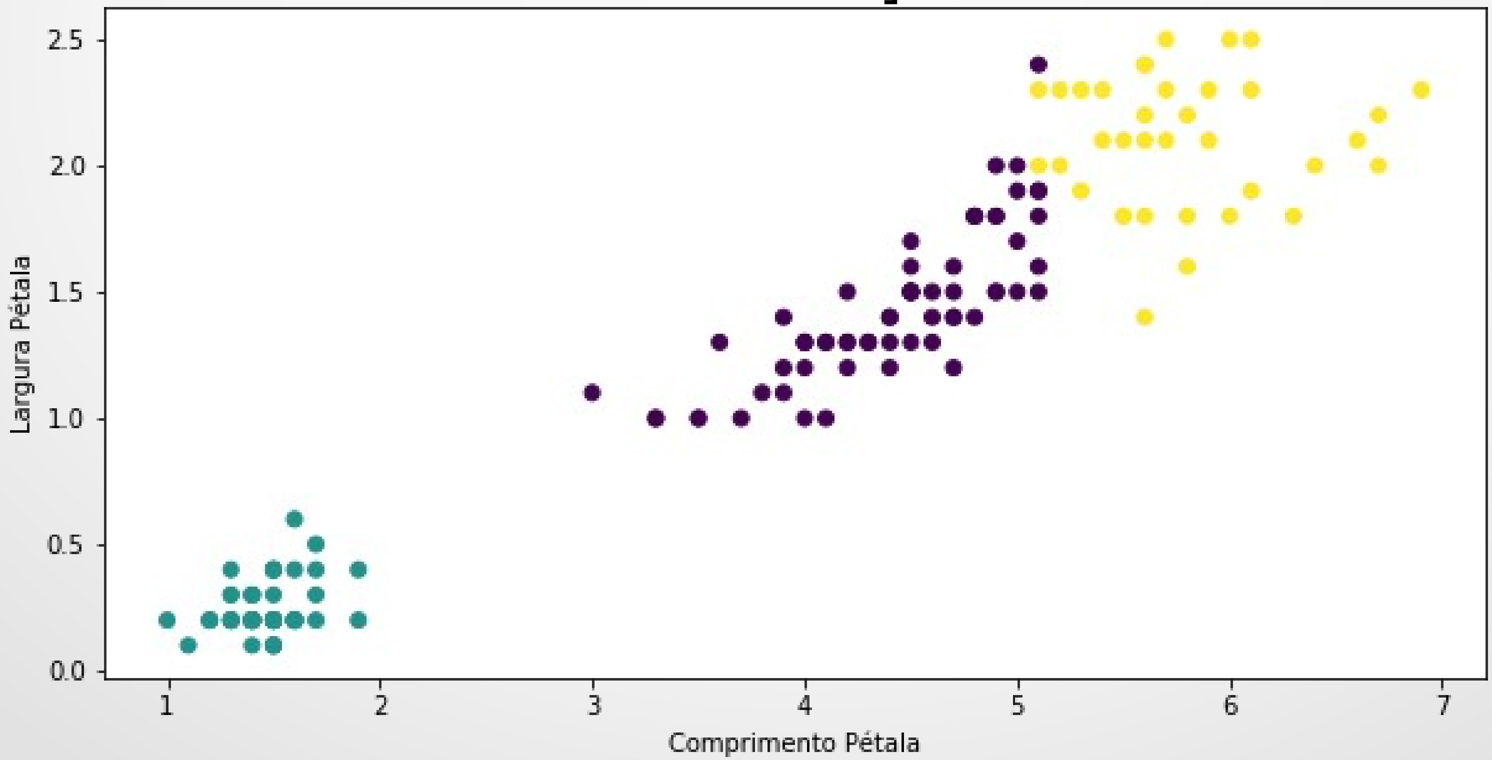
Aplicando aglomerativo sobre Iris

- Código

```
import pandas as pd
#Importa agrupamento aglomerativo do skleran
from sklearn.cluster import AgglomerativeClustering
#importa pyplot
from matplotlib import pyplot as plt
plt.rcParams['figure.figsize'] = [10, 5]
# Localização do arquivo
filepath = 'data/Iris_Data.csv'
# Importando os dados
data = pd.read_csv(filepath)
#Cria uma instância de classe agrupamento aglomerativo com
distância average
agg = AgglomerativeClustering(n_clusters=3,affinity='euclidean',
linkage='average')
#aplica agrupamento aglomerativo sobre os dados, desconsiderando
classes
agg  = agg.fit(data.iloc[:,:4])
#imprime rótulos do agrupamento aglomerativo sobre os dados
print(agg.labels_)
# Rótulo dos grupos usando comprimento e largura da pétala
fig = plt.figure()
plt.scatter(data.petal_length, data.petal_width, c=agg.labels_)
plt.xlabel('Comprimento Pétala')
plt.ylabel('Largura Pétala')
```

Resultado

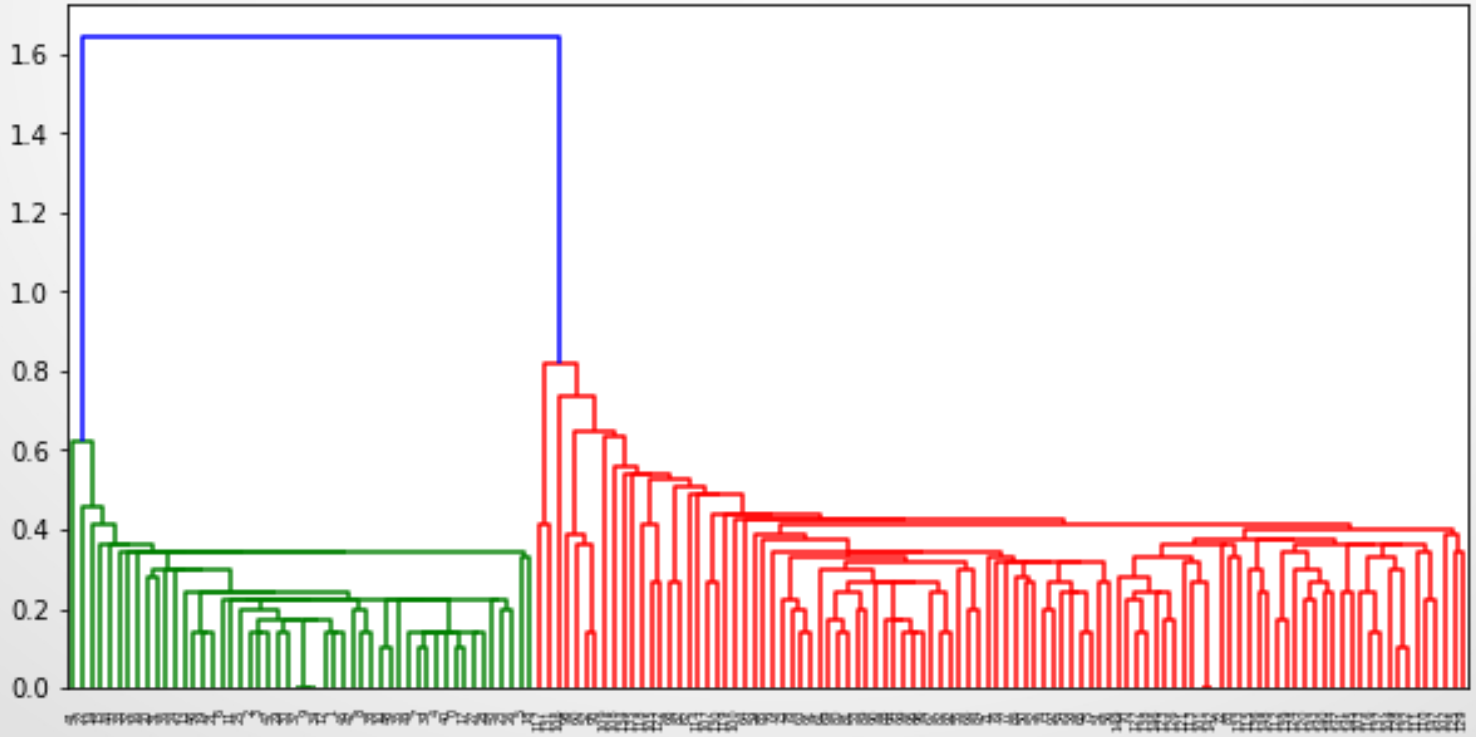
- [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 2 0 2 2 2 2 0 2 2 2 2 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2
2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2 2 0]



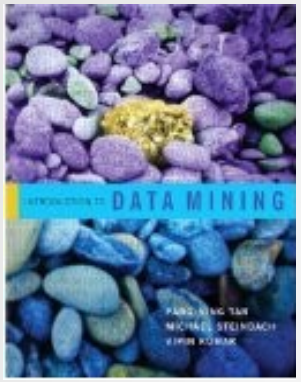
Exemplo com dendrograma

- Código

```
from scipy.cluster.hierarchy import dendrogram, linkage
#aplica agrupamento aglomerativo de ligação simples
Z = linkage(data.iloc[:,:4])
#desenvolve o dendrograma
dendrogram(Z)
```



Bibliografia

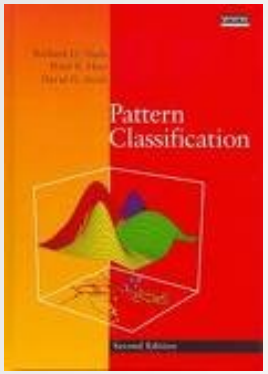


V. TAN, STEINBACH, M., KUMAR, P. Introdução ao Data Mining (Mineração de Dados). Edição 1. Ciência Moderna 2009. ISBN 9788573937619.

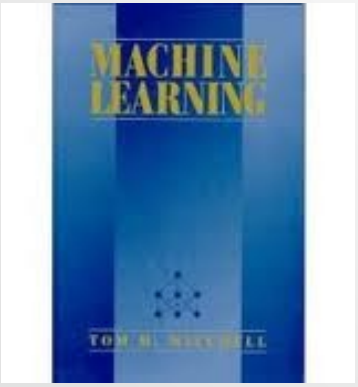
Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Katti Faceli, Ana Carolina Lorena, João Gama, André C. P. L. F. de Carvalho. Grupo Gen 2011

Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993

Referencias



Duda, R.O., Hart, P. E. and Stork, D. G. Pattern Classification (2nd Edition). Wiley-Interscience



MITCHELL, T. Machine Learning, McGraw Hill, 1997.

Referências

- Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
- Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 2005.
- Tan, P.-N., Steinbach, M., and Kumar, V., Introduction to Data Mining, Addison-Wesley, 2006
- Wu, X. and Kumar, V., The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC, 2009
- D. Steinley, K-Means Clustering: A Half-Century Synthesis, British J. of Mathematical and Stat. Psychology, V. 59, 2006