Machine Learning

Aula 06

Aprendizagem não supervisionada: agrupamento

Evandro J.R. Silva

Uninassau Teresina

Sumário

Agrupamento

- O **agrupamento**, também conhecido como **clusterização** (ou *clustering*) consiste na divisão dos dados em grupos (ou *clusters*) de **características similares**, ou seja, que compartilham valores próximos entre os atributos de entrada.
- Pode ser definido como
- A forma como os grupos são divididos varia de acordo com a tarefa utilizada, podendo ser hierárquico ou particional.

Agrupamento

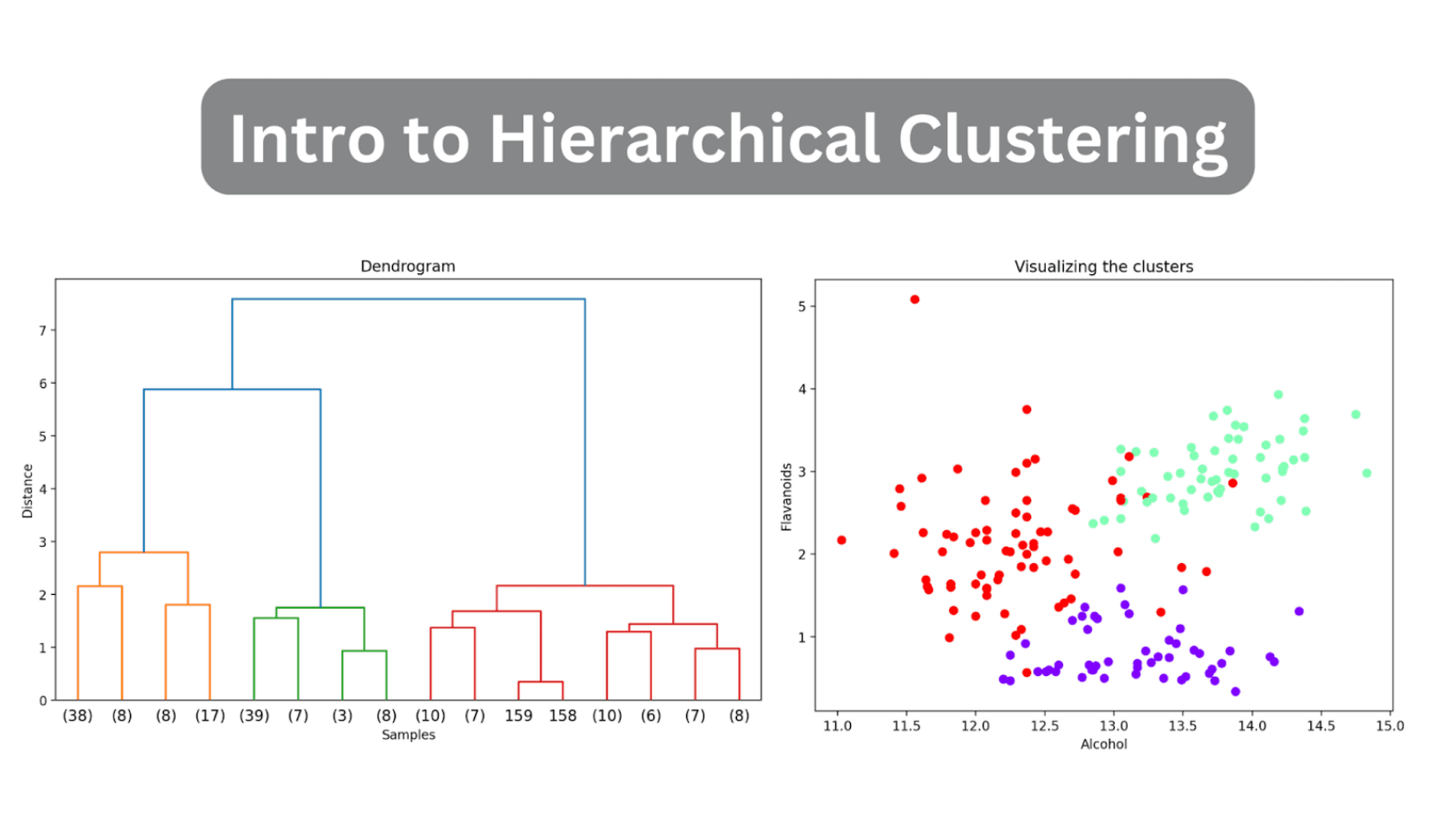


Figura 1: Agrupamento hierárquico (fonte: KDnuggets)

Agrupamento

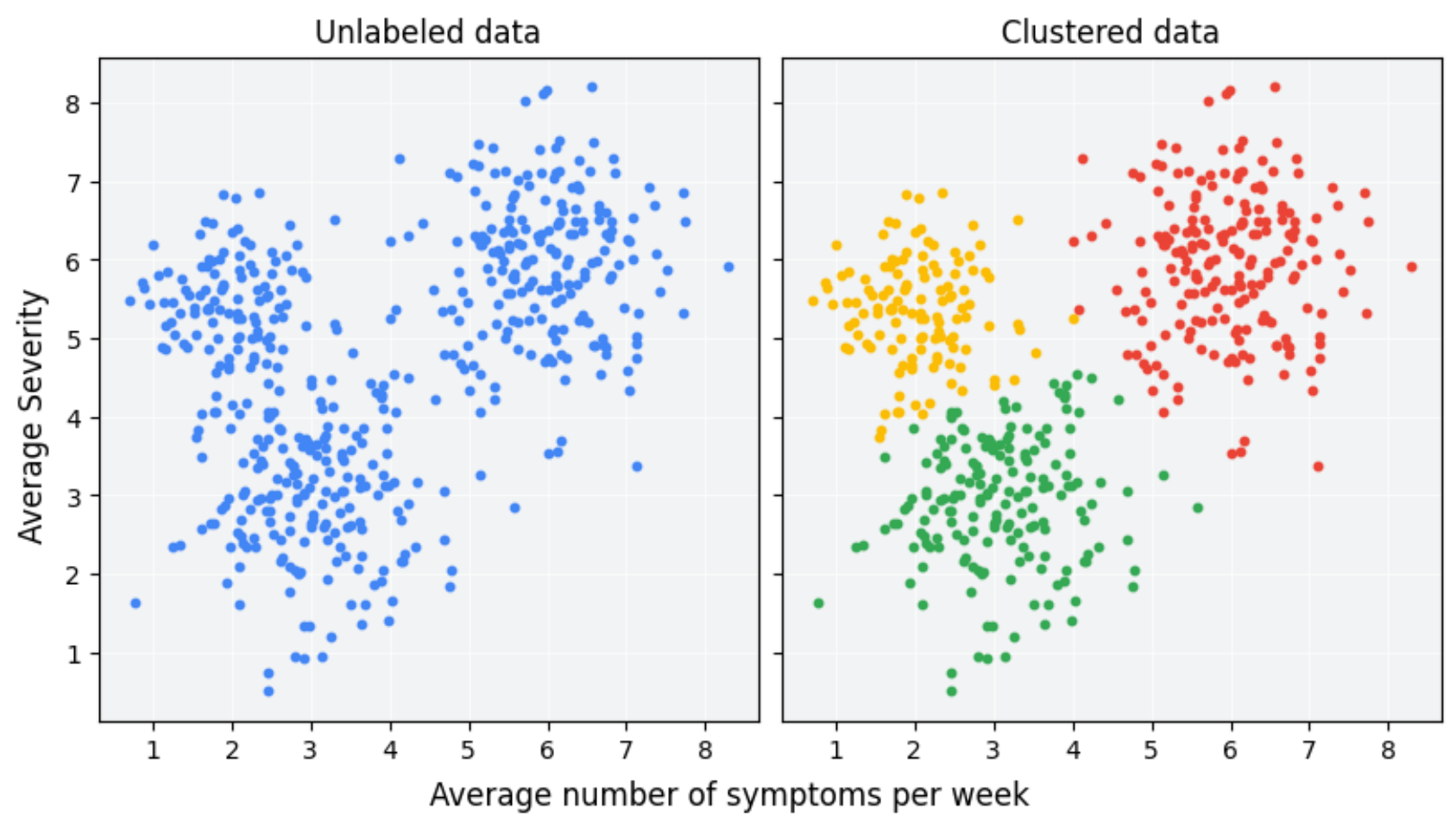
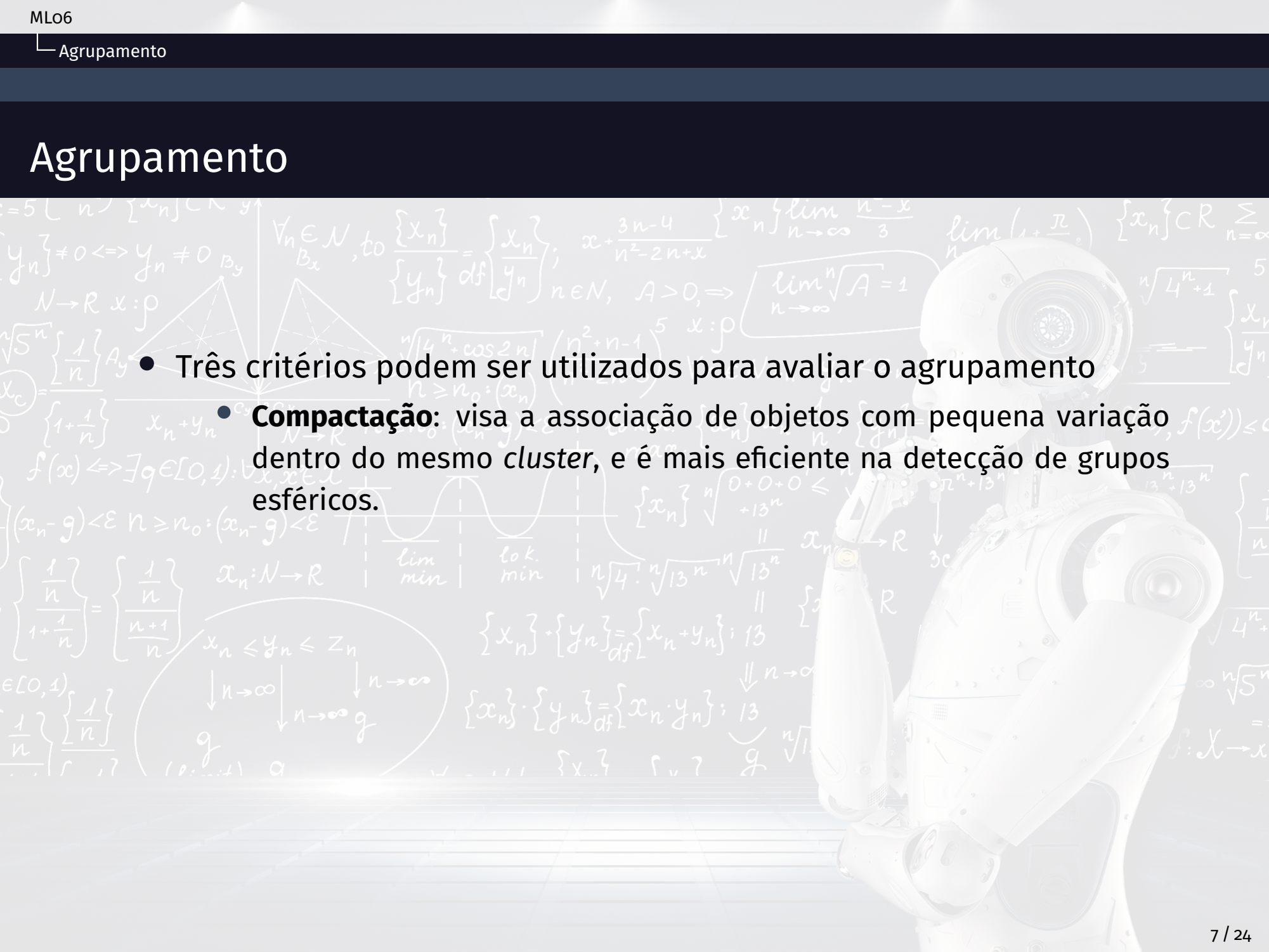


Figura 2: Agrupamento particional (fonte: Developers - Google)

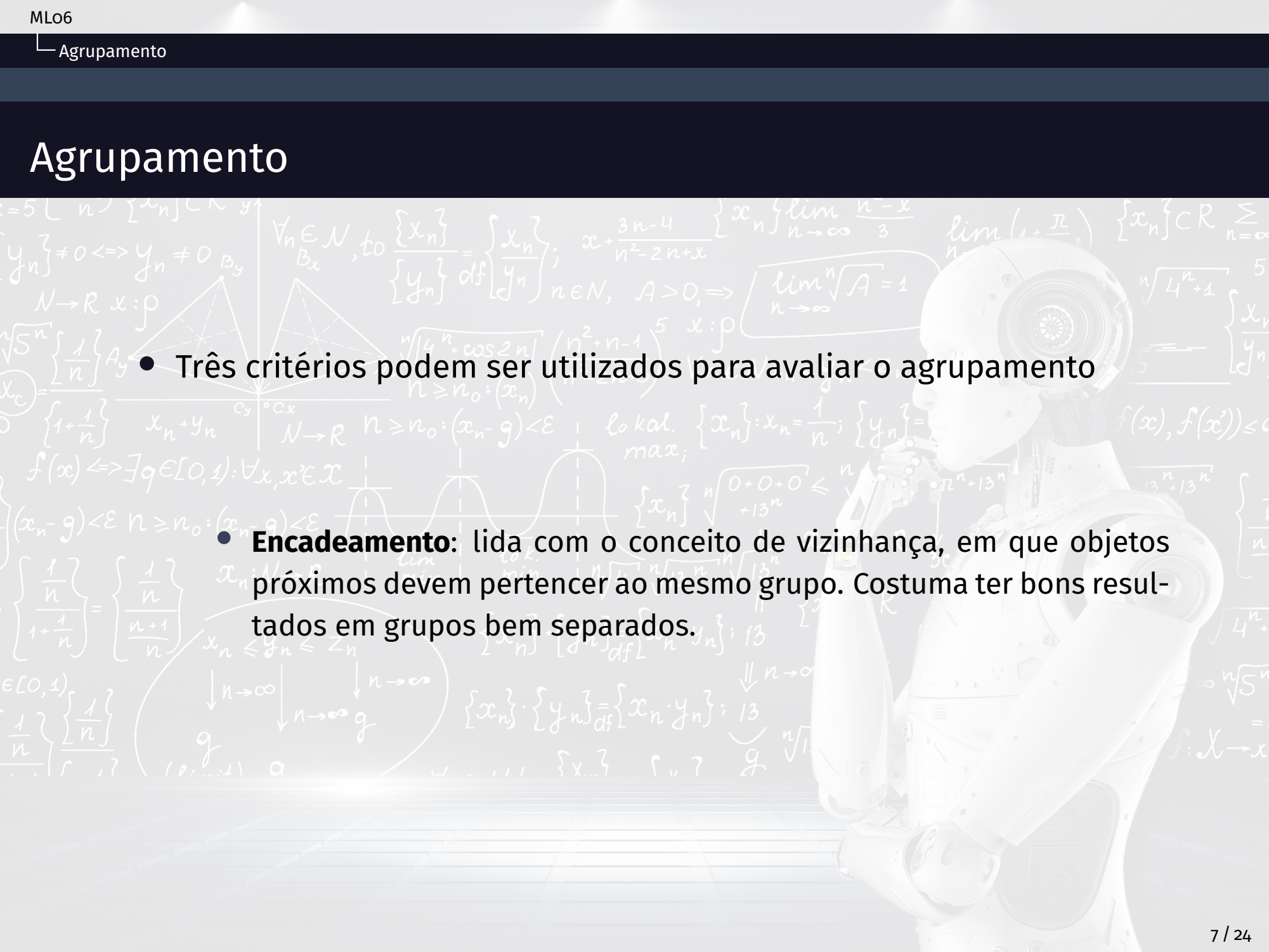
Agrupamento

- Três critérios podem ser utilizados para avaliar o agrupamento
  - **Compactação**: visa a associação de objetos com pequena variação dentro do mesmo *cluster*, e é mais eficiente na detecção de grupos esféricos.



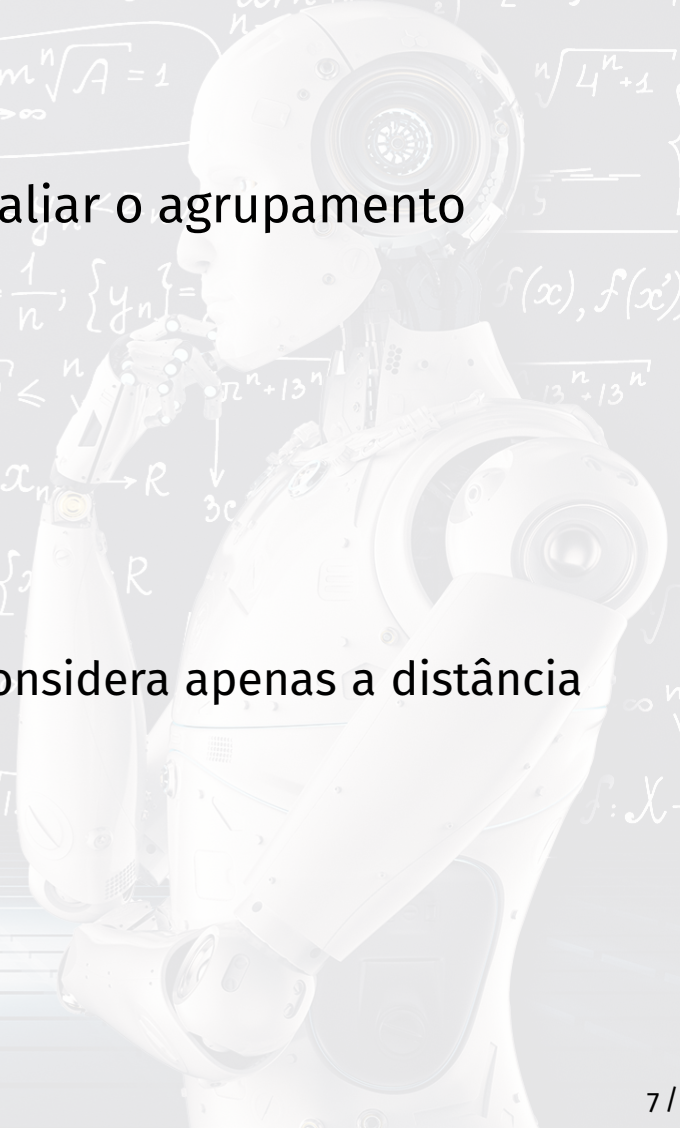
Agrupamento

- Três critérios podem ser utilizados para avaliar o agrupamento
  - **Encadeamento**: lida com o conceito de vizinhança, em que objetos próximos devem pertencer ao mesmo grupo. Costuma ter bons resultados em grupos bem separados.



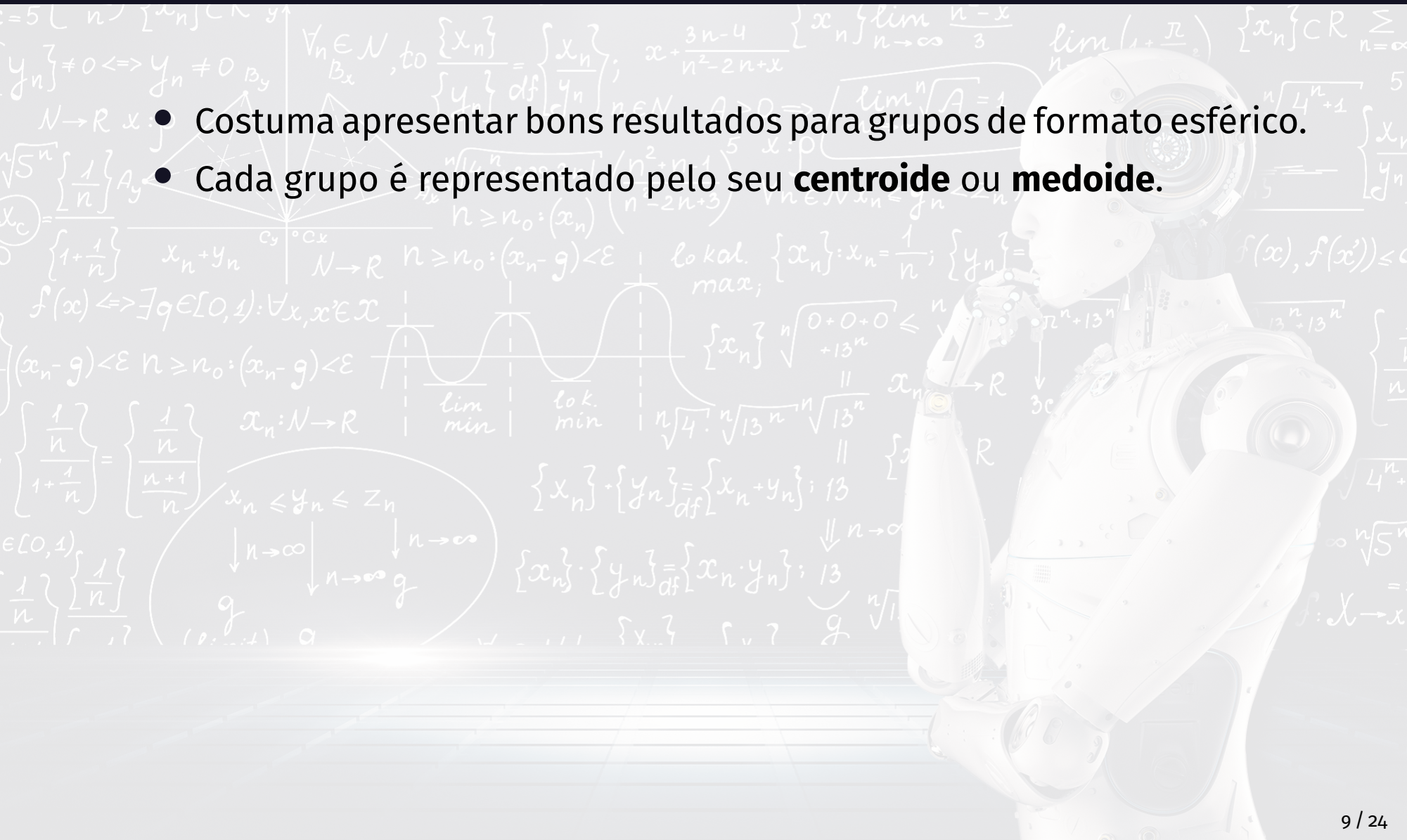
Agrupamento

- Três critérios podem ser utilizados para avaliar o agrupamento
  - **Separação espacial**: é mais genérica e considera apenas a distância entre os *clusters*.



K-means

- Costuma apresentar bons resultados para grupos de formato esférico.
- Cada grupo é representado pelo seu **centroide** ou **medoide**.



K-means

- Costuma apresentar bons resultados para grupos de formato esférico.
- Cada grupo é representado pelo seu **centroide** ou **medoide**.
  - O ponto médio do cluster.

$$\bar{x}^j = \frac{1}{n_j} \sum_{x_i \in C_j}^{n_j} x_i$$

  - onde
    - j: índice do *cluster*;
    - $\bar{x}^j$: centroide do j-ésimo *cluster*;
    - n_j: número de instâncias no *cluster* j;
    - C_j: j-ésimo *cluster*;
    - x_i: i-ésima instância.

K-means

- Costuma apresentar bons resultados para grupos de formato esférico.
- Cada grupo é representado pelo seu **centroide** ou **medoide**.
  - Neste caso o nome do algoritmo passa a ser **k-medoides**.
  - O medoide é ponto mais representativo do *cluster* e corresponde, em geral, ao ponto de menor distância para os demais membros.

$$x_{medoide} = argmin_{x_m \in C_j} \sum_{i=1}^{n_j} d(x_m, x_i)$$

  - onde
    - x_m: m-ésima instância do *cluster*;
    - d: função de distância.

K-means

- O algoritmo procura minimizar o erro (E) de acordo com a seguinte função:

$$E = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \overline{x}^{(j)})^2$$

- Onde k é a quantidade de *clusters*.

K-means

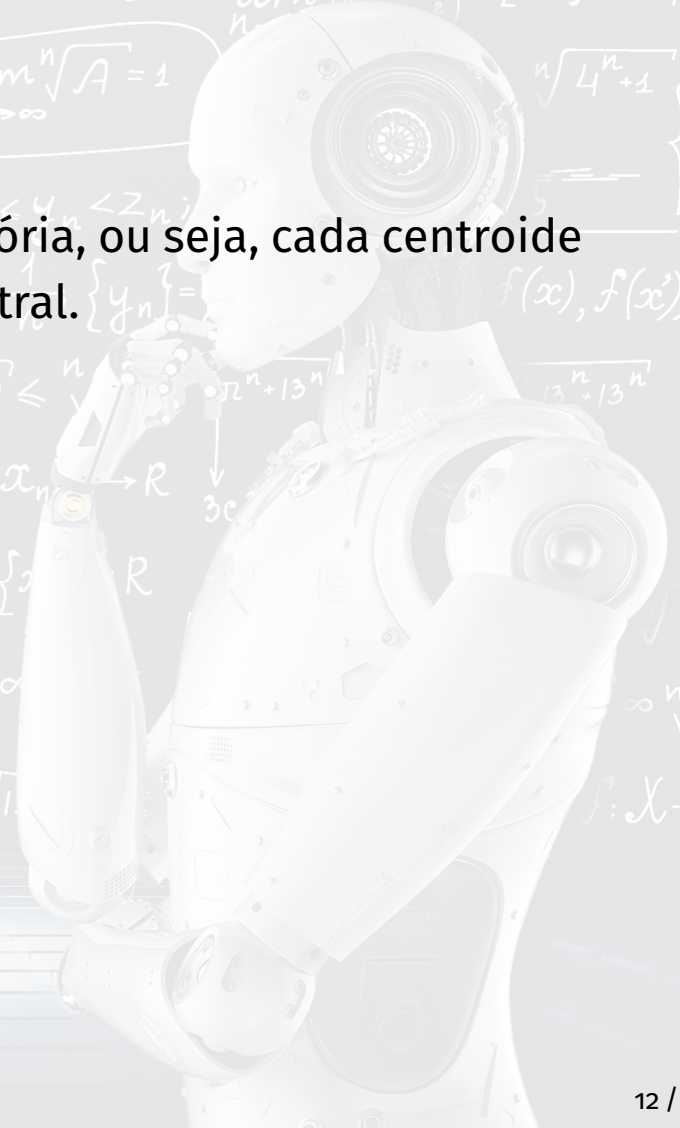
- O algoritmo procura minimizar o erro (E) de acordo com a seguinte função:

$$E = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \overline{x}^{(j)})^2$$

- Onde k é a quantidade de *clusters*.
  - O valor de k é definido pelo usuário, e pode ser qualquer valor maior ou igual a 1.
  - A definição de k pode ter base em conhecimentos prévios do usuário, ou ser um resultado alcançado a partir de *teste-e-erro*.

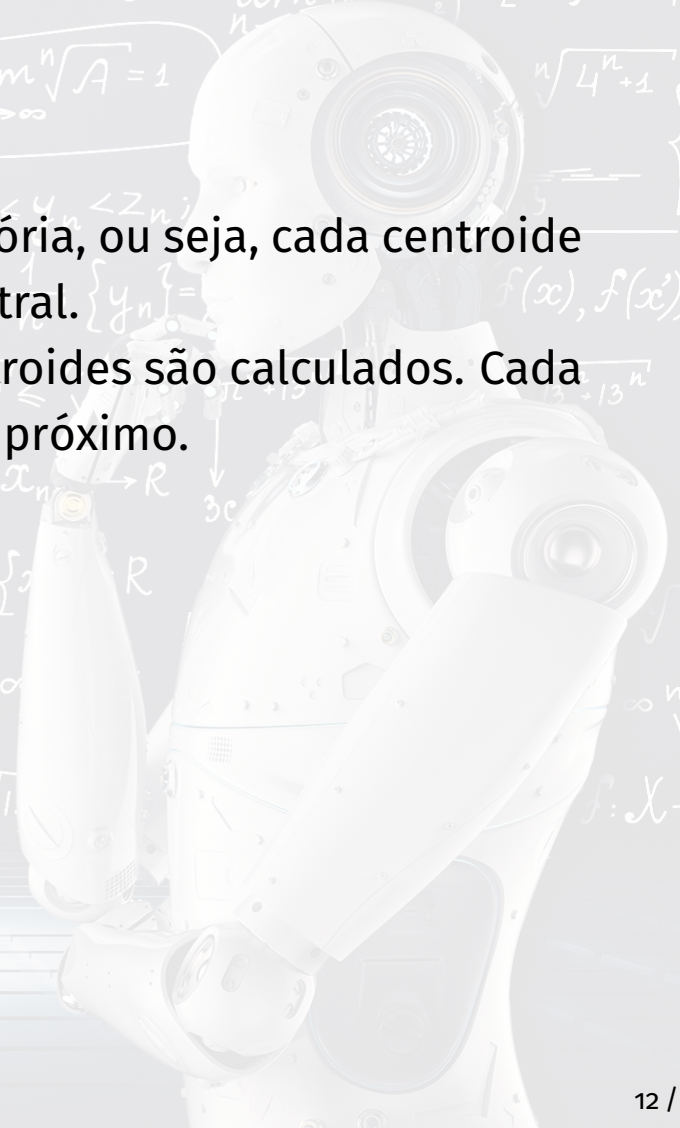
K-means

- Como funciona *na prática*
  1. *k* pontos são inicializados de forma aleatória, ou seja, cada centroide aparece em algum lugar no espaço amostral.



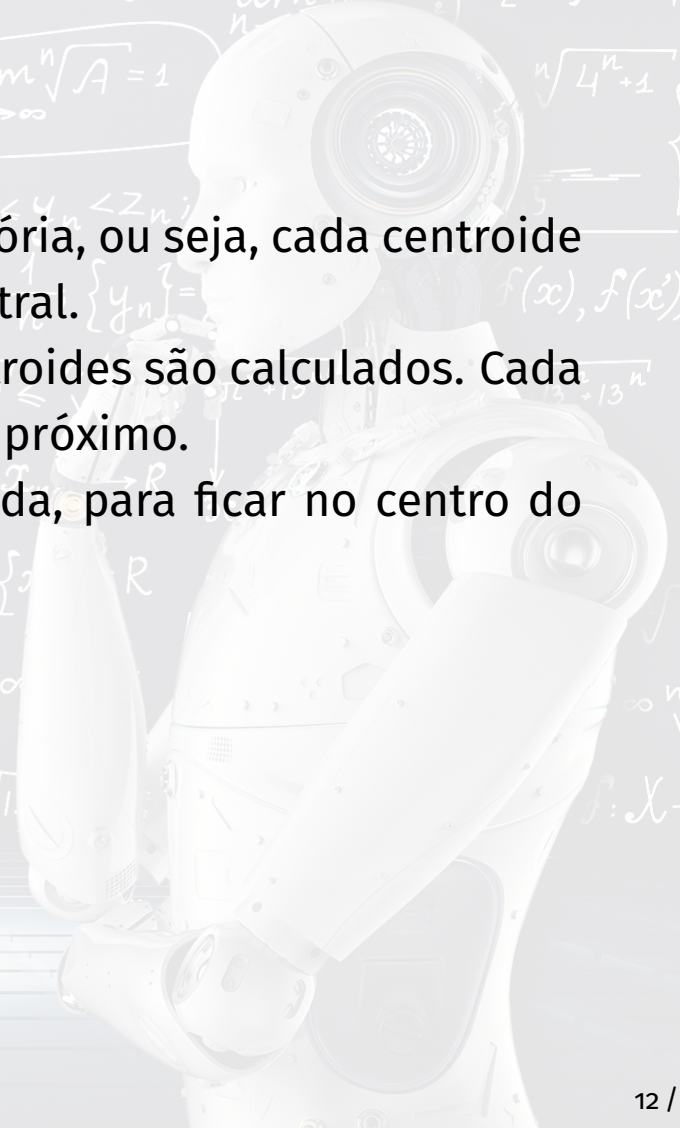
K-means

- Como funciona *na prática*
  1. *k* pontos são inicializados de forma aleatória, ou seja, cada centroide aparece em algum lugar no espaço amostral.
  2. A distância entre cada instância e os centroides são calculados. Cada instância é assinalada ao centroide mais próximo.



K-means

- Como funciona *na prática*
  1. *k* pontos são inicializados de forma aleatória, ou seja, cada centroide aparece em algum lugar no espaço amostral.
  2. A distância entre cada instância e os centroides são calculados. Cada instância é assinalada ao centroide mais próximo.
  3. A posição de cada centroide é recalculada, para ficar no centro do *cluster*.



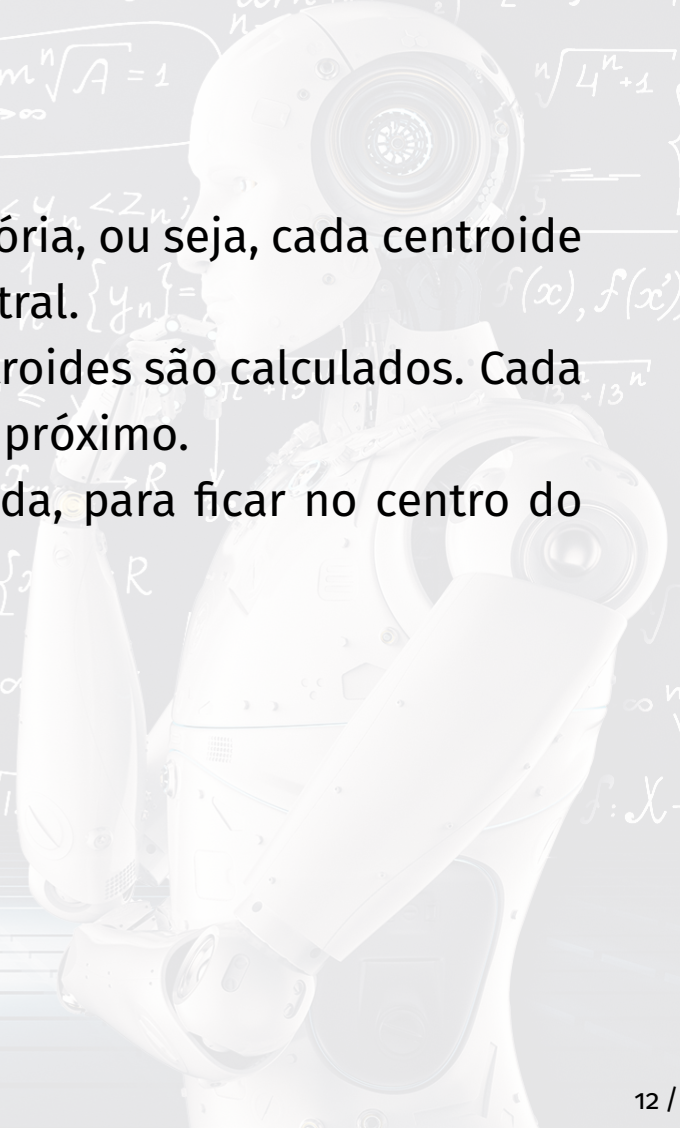
K-means

- Como funciona *na prática*
  1. *k* pontos são inicializados de forma aleatória, ou seja, cada centroide aparece em algum lugar no espaço amostral.
  2. A distância entre cada instância e os centroides são calculados. Cada instância é assinalada ao centroide mais próximo.
  3. A posição de cada centroide é recalculada, para ficar no centro do *cluster*.
  4. Os passos **2** e **3** são repetidos até que nenhuma instância seja assinalada a um *cluster* diferente.

K-means

- Como funciona *na prática*
  1. *k* pontos são inicializados de forma aleatória, ou seja, cada centroide aparece em algum lugar no espaço amostral.
  2. A distância entre cada instância e os centroides são calculados. Cada instância é assinalada ao centroide mais próximo.
  3. A posição de cada centroide é recalculada, para ficar no centro do *cluster*.

- Vídeo explicativo



Avaliação dos grupos

- Podemos calcular a **dispersão *intercluster*** e ***intracluster***.

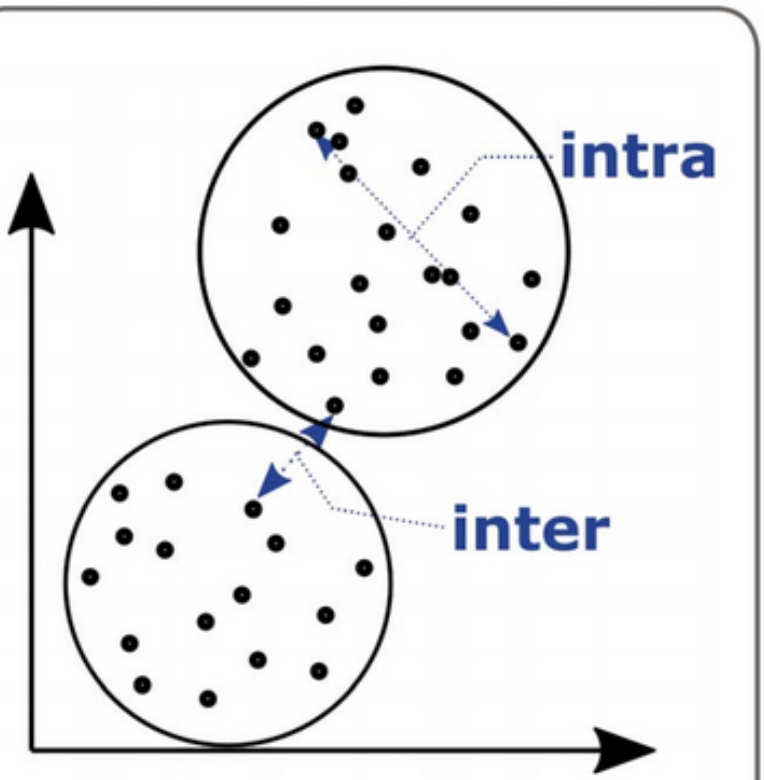
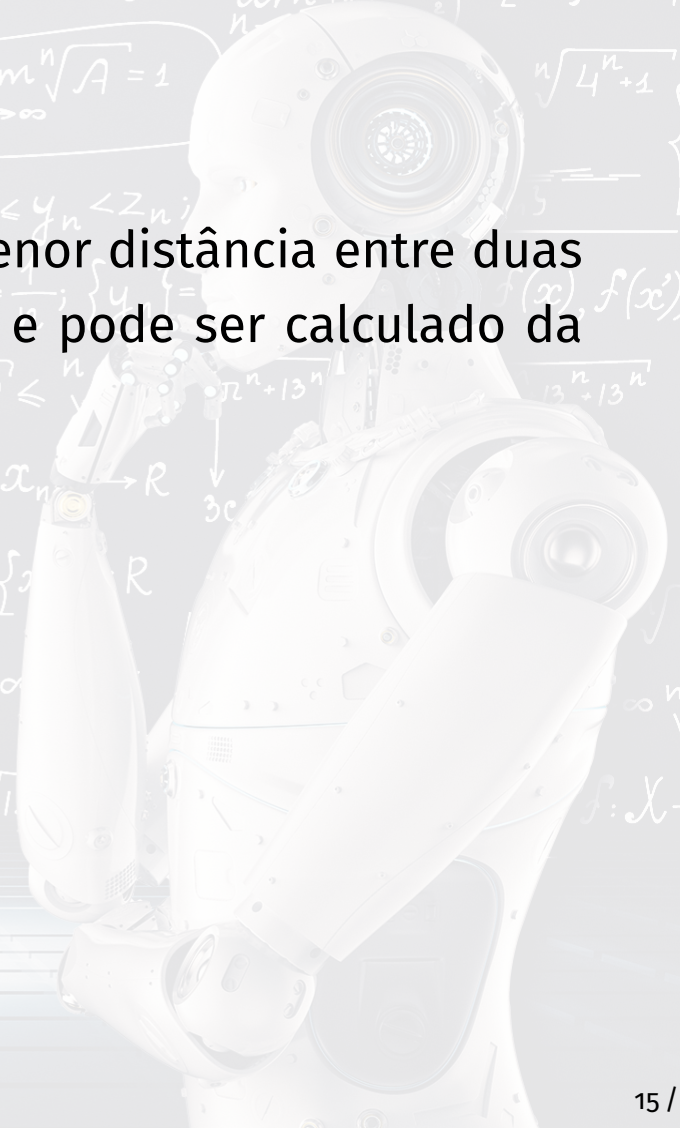


Figura 4. Tipos de dispersão: *intercluster* e *intracluster*.

Fonte: Maikon Lucian Lenz/Shutterstock.com.

Avaliação dos grupos

- A **dispersão intercluster** corresponde à menor distância entre duas instâncias pertencentes a cada um deles, e pode ser calculado da seguinte forma:
  $d(C_a, C_b) = \min d(x_{ai}, x_{bj})$, onde
  - C_a, C_b: *clusters* diferentes;
  - x_{ai}: i-ésima instâncias do *cluster* C_a;
  - x_{bj}: j-ésima instâncias do *cluster* C_b.



Avaliação dos grupos

- A medição da **dispersão intracluster** pode ser feita com o cálculo da variância:

$$var = \sqrt{\frac{1}{n_x} \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \bar{x}^{(j)})}$$

onde

- n_x: quantidade total de instâncias.

Avaliação dos grupos

- **Método *Elbow*** (ou método cotovelo)
  - Utiliza o cálculo das dispersões inter e intracluster para determinar o valor de *k*.

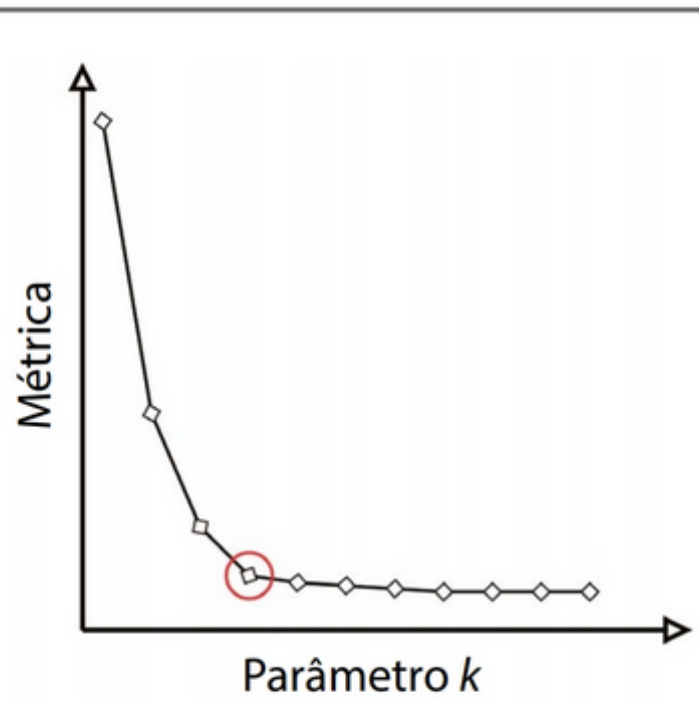


Figura 5. Método *Elbow* de seleção do parâmetro *k*.

Fonte: Adaptada de Maikon Lucian Lenz/Shutterstock.com.

Avaliação dos grupos

- **Conectividade**: mede o quanto instâncias de maior proximidade, ditas vizinhas, pertencem ao mesmo grupo:

$$conectividade = \sum_{i=1}^{n_x} \sum_{j=1}^{n_v} f(x_i, v_{ij})$$

$$f(x_i, v_{ij}) = \begin{cases} \frac{1}{j} & \text{, quando } x_i \text{ e } v_{ij} \text{ não estão no mesmo } cluster \\ 0 & \text{, quando } x_i \text{ e } v_{ij} \text{ pertencem ao mesmo } cluster \end{cases}$$

onde

- n_v: quantidade total de instâncias vizinhas à instância x_i;
- v_{ij}: j-ésima instância vizinha à instância x_i.

Avaliação dos grupos

- **Silhueta**
  - É uma métrica que avalia o quanto uma instância está adequadamente vinculada a um *cluster*.
  - Os valores variam de −1 (quando uma instância deveria estar vinculada a outro *cluster*) a 1 (quando uma instância está associada ao *cluster* ideal).

Avaliação dos grupos

- A medida da silhueta utiliza os valores
  - De distância média de uma instância para os seus parceiros de *cluster*

$$a(x_i, C_k) = \frac{1}{C_k} \sum_{\substack{x_i, x_j \in C_k \\ x_i \neq x_j}} d(x_i, x_j)$$

  - E a menor distância média para outros *clusters*

$$b(x_i) = \min_{\substack{x_i \in C_i \\ C_i \neq C_j}} a(x_i, C_j)$$

Avaliação dos grupos

$$silhueta(x_i) = \begin{cases} 1 - \dfrac{a(x_i, C_i)}{b(x_i)} & \text{, quando } a(x_i, C_i) < b(x_i) \\ 0 & \text{, quando } a(x_i, C_i) = b(x_i) \\ \dfrac{b(x_i)}{a(x_i, C_i)} - 1 & \text{, quando } a(x_i, C_i) > b(x_i) \end{cases}$$

$$silhueta(C_k) = \frac{1}{n} \sum_{x_i \in C_k}^{n} silhueta(x_i)$$

Onde n: quantidade de instâncias do *cluster*

Avaliação dos grupos

- Os critérios de avaliação mostrados não buscam por soluções “corretas”, mas buscam por soluções que encontrem estruturas apropriadas e que possam revelar outras ocultas.
- A avaliação tem por objetivo determinar se a estrutura encontrada de fato existe e se faz sentido.

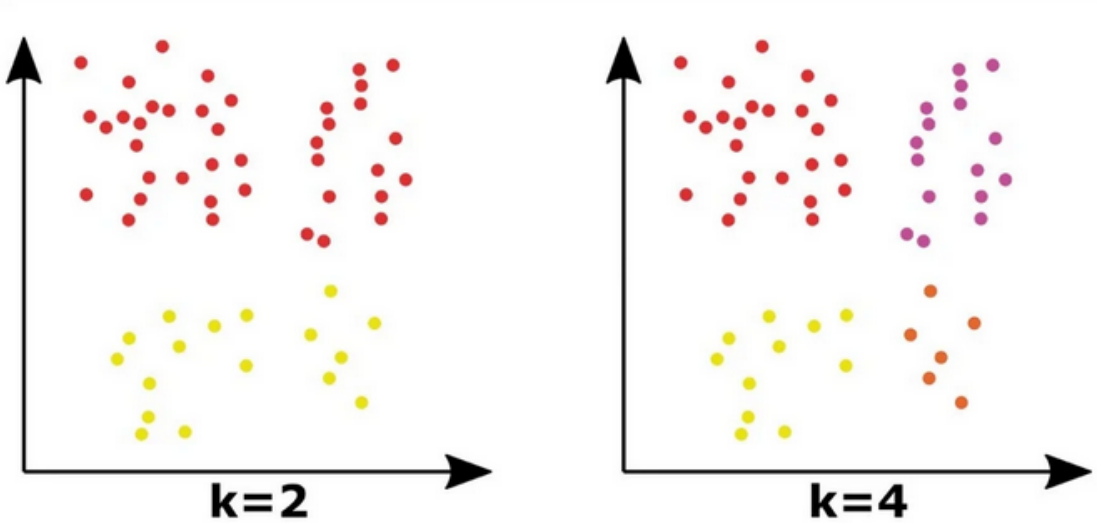


Figura 6. (a) Dois *clusters* ($k = 2$). (b) Quatro *clusters* ($k = 4$).

Fonte: Maikon Lucian Lenz/Shutterstock.com.

1 Agrupamento

2 K-means

3 Avaliação dos grupos

4 FIM

FIM