



Data Science & Business Analytics

# Machine Learning Models

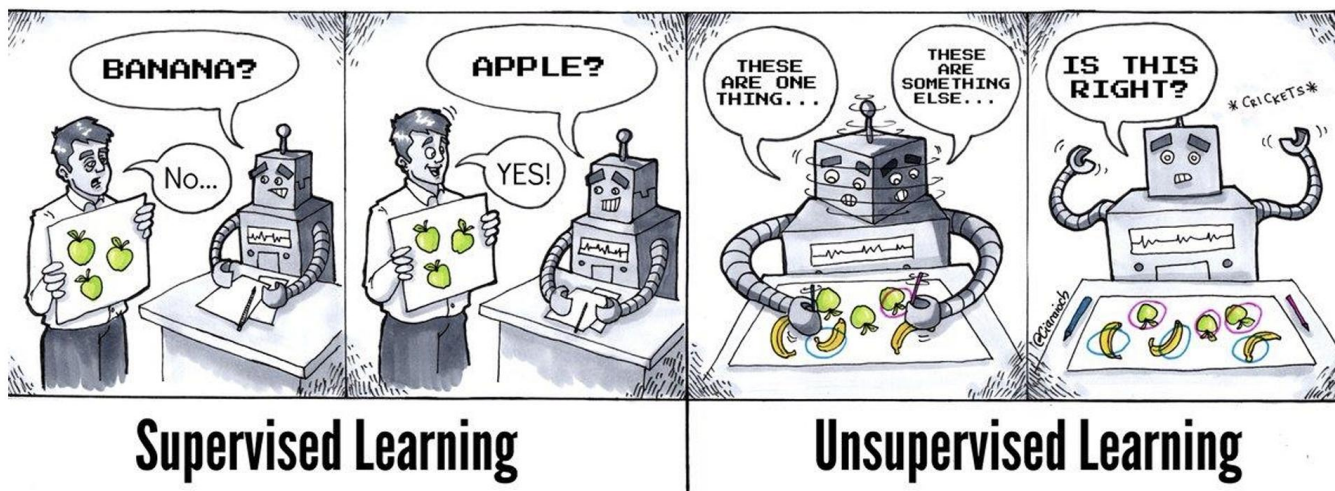
David Issá

davidribeiro.issa@gmail.com

# 1. Unsupervised Learning

# 1. Unsupervised Learning – Definição

Corresponde a um conjunto de técnicas de machine learning, cujo objetivo é o de **aprender padrões a partir de dados não rotulados (unlabeled)**.



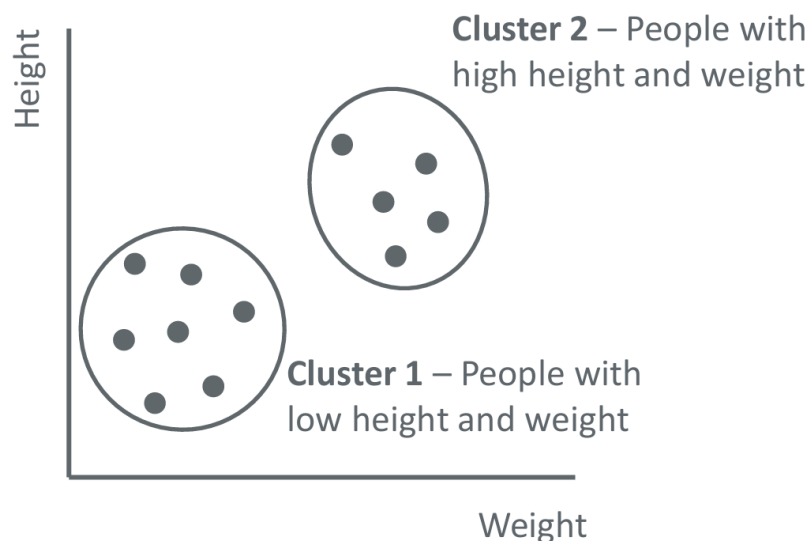
- Conjunto de dados de treino rotulado;
- Aprende a relação entre as variáveis independentes e a variável alvo;
- Utilizado para classificar novas observações.

- Os dados não têm etiquetas;
- Encontra clusters (grupos) óptimos;
- Não utilizado com novas observações.

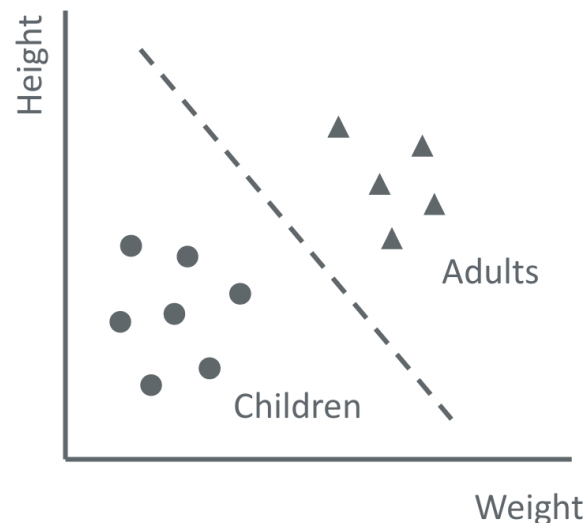
# 1. Unsupervised Learning – Clustering

Por vezes, existe alguma confusão relativamente à distinção entre algoritmos de clustering (unsupervised) e algoritmos classificação (supervised).

## Clustering



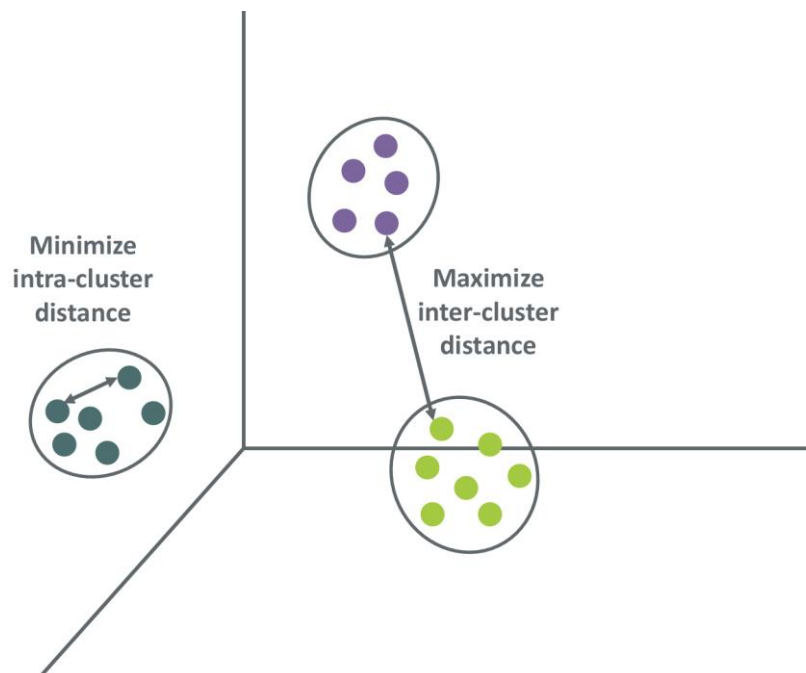
## Classification



## 2. Clustering

## 2. Clustering – Definição

Tarefa de **agrupar observações com base na sua semelhança**. Parte do princípio de que os pontos semelhantes estão relacionados e, por conseguinte, podem ser considerados um grupo (cluster).



## 2. Clustering – Definição

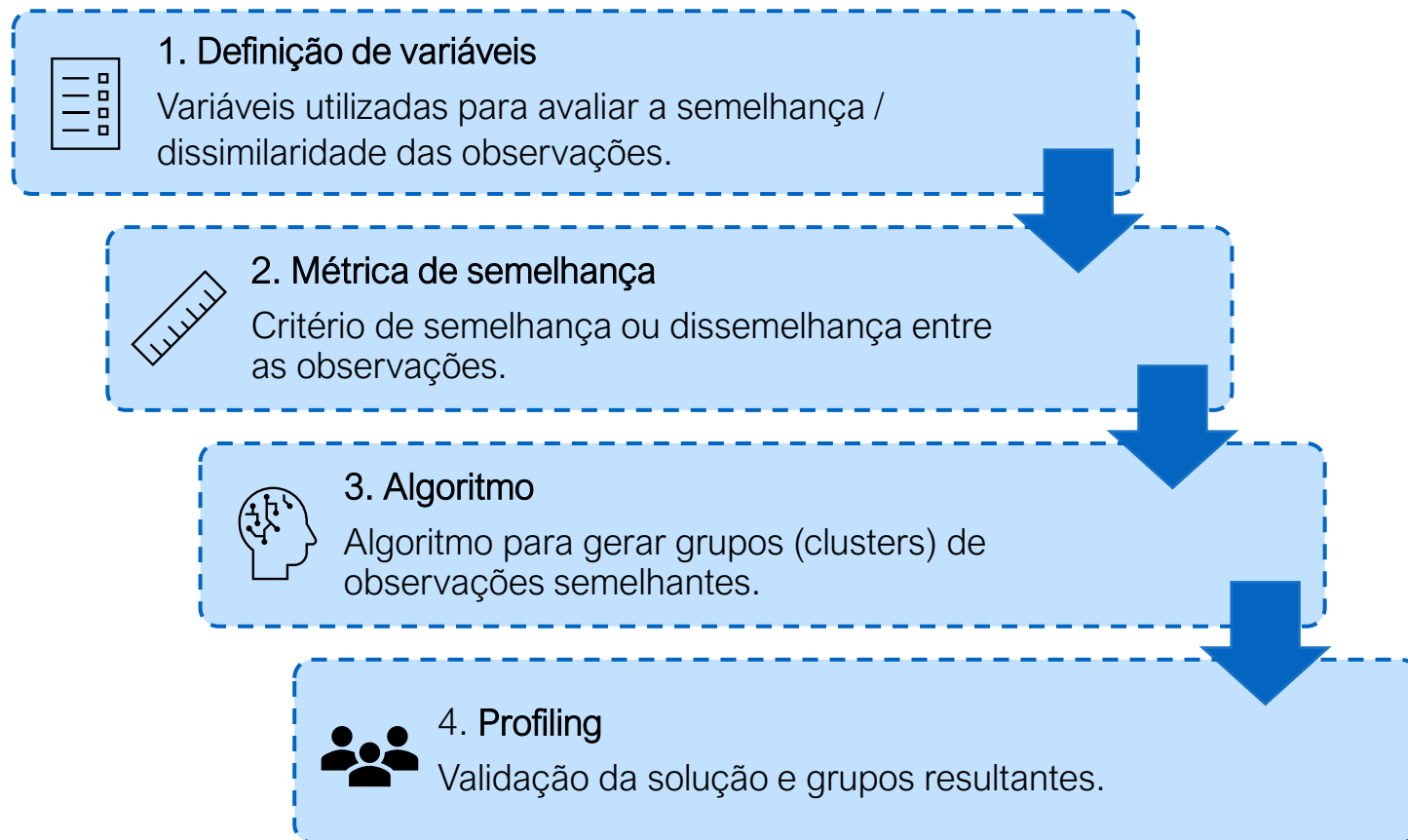
É bastante utilizado na análise exploratória de dados:

- Permite resumir grandes conjuntos de dados: trade-off entre detalhe e compreensão dos dados.
- Permite caracterizar entidades numa base de dados (clientes, produtos, etc.).

Algumas das principais aplicações das análises de clustering são:

- **Business/Marketing:** segmentação de clientes;
- **Saúde:** gestão de doenças;
- **Planeamento urbano:** planeamento de transportes.

## 2. Clustering – Principais Etapas





## 2. Clustering – Principais Etapas



### 1. Definição de variáveis

○ **tipo de problema determina as variáveis a escolher.**

- A inclusão de variáveis discriminantes é decisiva;
- A qualidade de qualquer análise de clusters é altamente condicionada pelas variáveis utilizadas,

A **escolha das variáveis deve ter um contexto teórico de suporte:**

- Este processo é efectuado a partir de um conjunto de variáveis que sabemos serem bons discriminantes para o problema em causa.

## 2. Clustering – Principais Etapas



### 2. Métrica de semelhança

Função que **recebe duas observações e devolve uma pontuação de similaridade/dissimilaridade**:

- A escolha correta é fundamental para obter bons clusters;
- A escolha depende do tipo de dados e do problema.
- Os dados são categóricos ou numéricos? A magnitude é importante? Os dados são altamente dimensionais?

## 2. Clustering – Principais Etapas



### 2. Métrica de semelhança

Função que **recebe duas observações e devolve uma pontuação de similaridade/dissimilaridade**:

- A escolha correta é fundamental para obter bons clusters;
- A escolha depende do tipo de dados e do problema.
- Os dados são categóricos ou numéricos? A magnitude é importante? Os dados são altamente dimensionais?

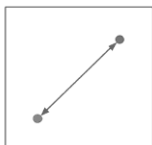
O tipo **mais comum de medidas são as medidas geométricas**:

- Distância entre os pontos de dados  $X_i$  e  $X_j$  com  $\nu$  dimensões

## 2. Clustering – Principais Etapas



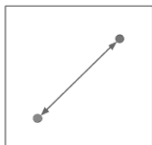
### 2. Métrica de semelhança (entre observações)



#### Distância Euclidiana (a mais usada)

- A distância entre dois elementos (i, j) é a raiz quadrada da soma dos quadrados das diferenças entre os valores de i e j para todas as variáveis ( $v = 1, 2, \dots, p$ )

$$d_{ij} = \sqrt{\sum_{v=1}^p (X_{iv} - X_{jv})^2}$$



#### Distância Euclidiana Ponderada

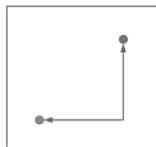
- Se a cada variável for atribuído um peso de acordo com a sua importância para a análise, a distância euclidiana ponderada assume a seguinte forma:

$$d_{ij} = \sqrt{\sum_{v=1}^p W_v (X_{iv} - X_{jv})^2}$$

## 2. Clustering – Principais Etapas

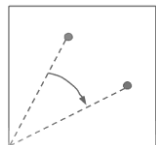


### 2. Métrica de semelhança (entre observações)



#### Distância Manhattan

- Melhor para aplicações de elevada dimensão
- Sendo a soma das diferenças absolutas, trata todas as dimensões de forma igual, impedindo que grandes diferenças numa dimensão enviem a métrica.



#### Cosine Similarity

- Mais usada para medir a semelhança de documentos em aplicações de NLP.
- Mede a semelhança entre dois vectores através do cosseno do ângulo entre os vectores.

## 2. Clustering – Principais Etapas



### 2. Métrica de semelhança (entre observações)

#### Distância Hamming

- Utilizado para comparar variáveis categóricas. É o número de posições de bits em que os dois bits são diferentes:

Id	Gender	Student	Nationality	code
u_1	Male (1)	Yes (1)	Indian (1)	111
u_2	Female (0)	No (0)	Spain (2)	002
u_3	Male (1)	No (1)	Australia (3)	113

1. Difference between u\_1 and u\_2 is 3

2. Difference between u\_2 and u\_3 is 3

3. Difference between u\_1 and u\_3 is 1

LTDR: A escolha da métrica de distância depende, em última análise, das características dos dados e do algoritmo de clustering que está a ser utilizado.

## 2. Clustering – Principais Etapas



### 2. Métrica de semelhança (entre variáveis)

#### Coeficiente de correlação Pearson

- Medir o grau de associação linear entre duas variáveis;
- Mede o grau em que duas variáveis se movem uma em relação à outra (variáveis numéricas).

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

#### Coeficiente de correlação Spearman

- Determina a força da relação entre duas variáveis (em que  $d$  é a diferença entre as classificações emparelhadas e  $n$  é o número de observações);
- Usado em variáveis numéricas e categóricas.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

## 2. Clustering – Principais Etapas



### 2. Métrica de semelhança (entre variáveis)

ID	$x$	$y$	Rank $x$	Rank $y$	$d_i^2$
1	87	241	5	6	1
2	46	160	1	1	0
3	72	210	4	4	0
4	64	195	3	3	0
5	123	285	7	7	0
6	97	233	6	5	1
7	50	174	2	2	0
Avg	77	212	NA	NA	NA

#### Coeficiente de correlação Pearson

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2} \sqrt{(y_i - \bar{y})^2}} = 0.985$$

#### Coeficiente de correlação Spearman

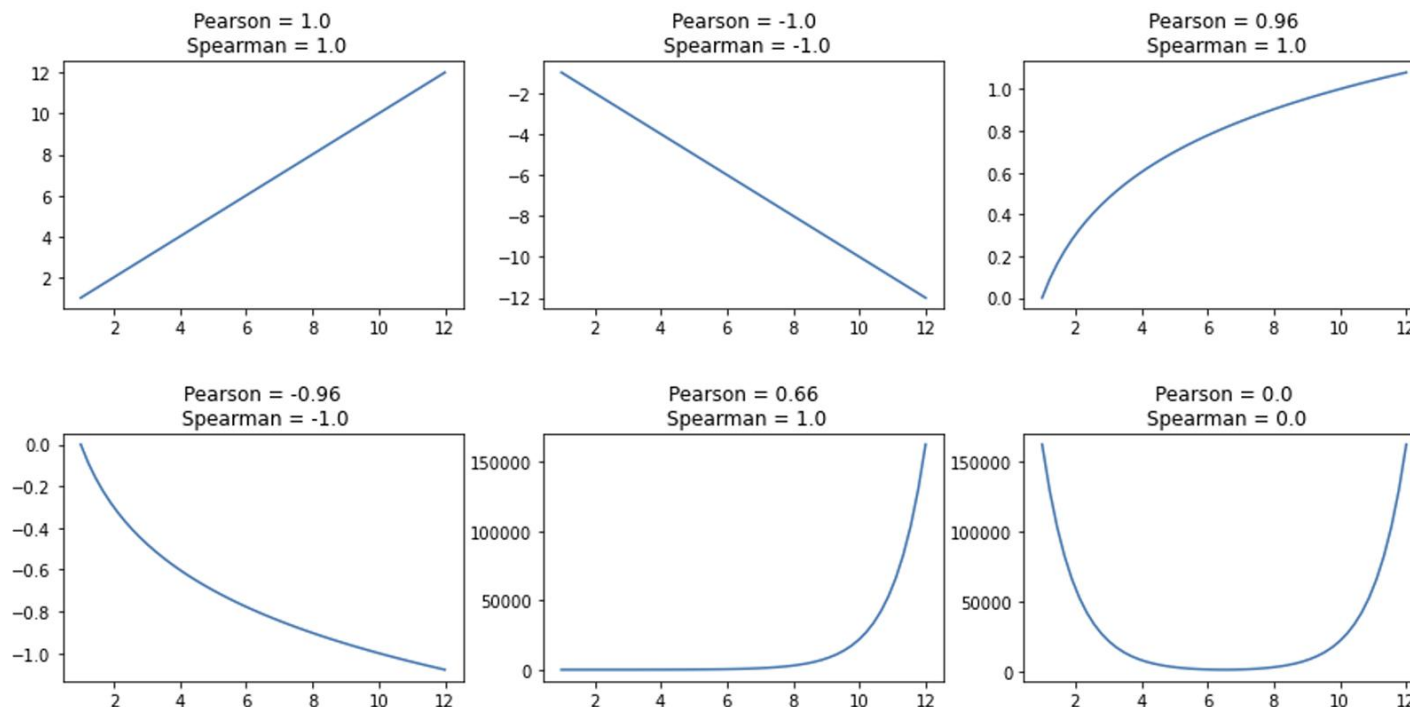
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 0.964$$



## 2. Clustering – Principais Etapas



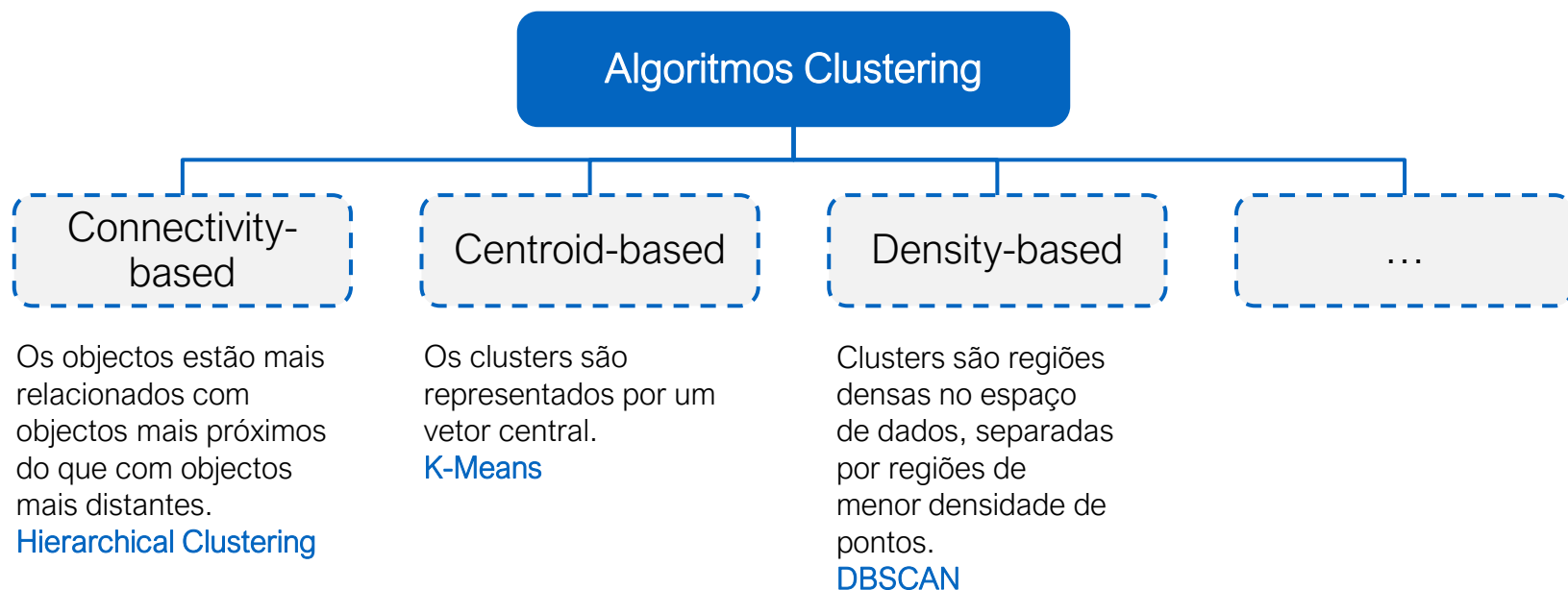
### 2. Métrica de semelhança (entre variáveis)



## 2. Clustering – Principais Etapas



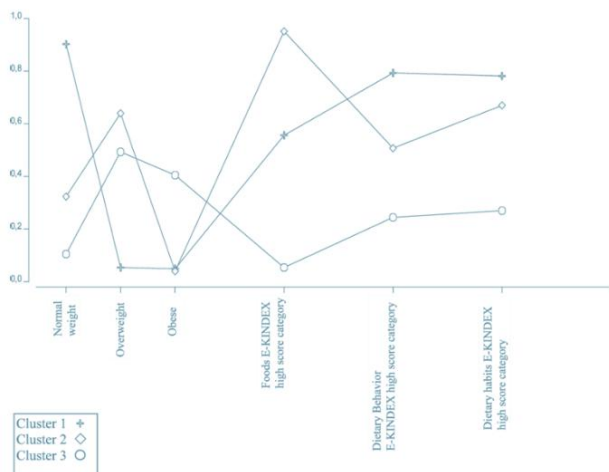
### 3. Algoritmo



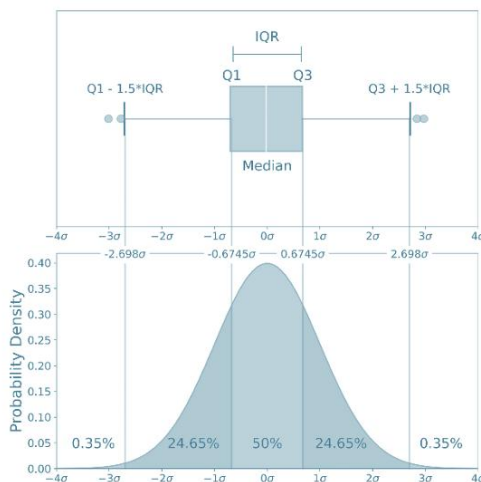
## 2. Clustering – Principais Etapas

### 4. Profiling

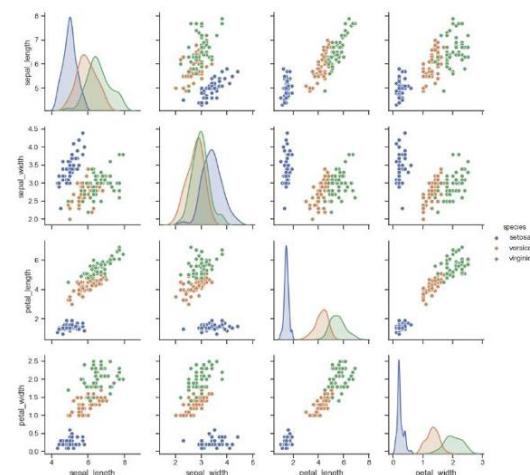
O objetivo principal é compreender o que se distingue em cada cluster.



Comparar médias  
para cada variável



Comparar distribuições  
dos clusters



Comparar correlações  
entre variáveis

## 2. Clustering – Principais Etapas

### 4. Profiling

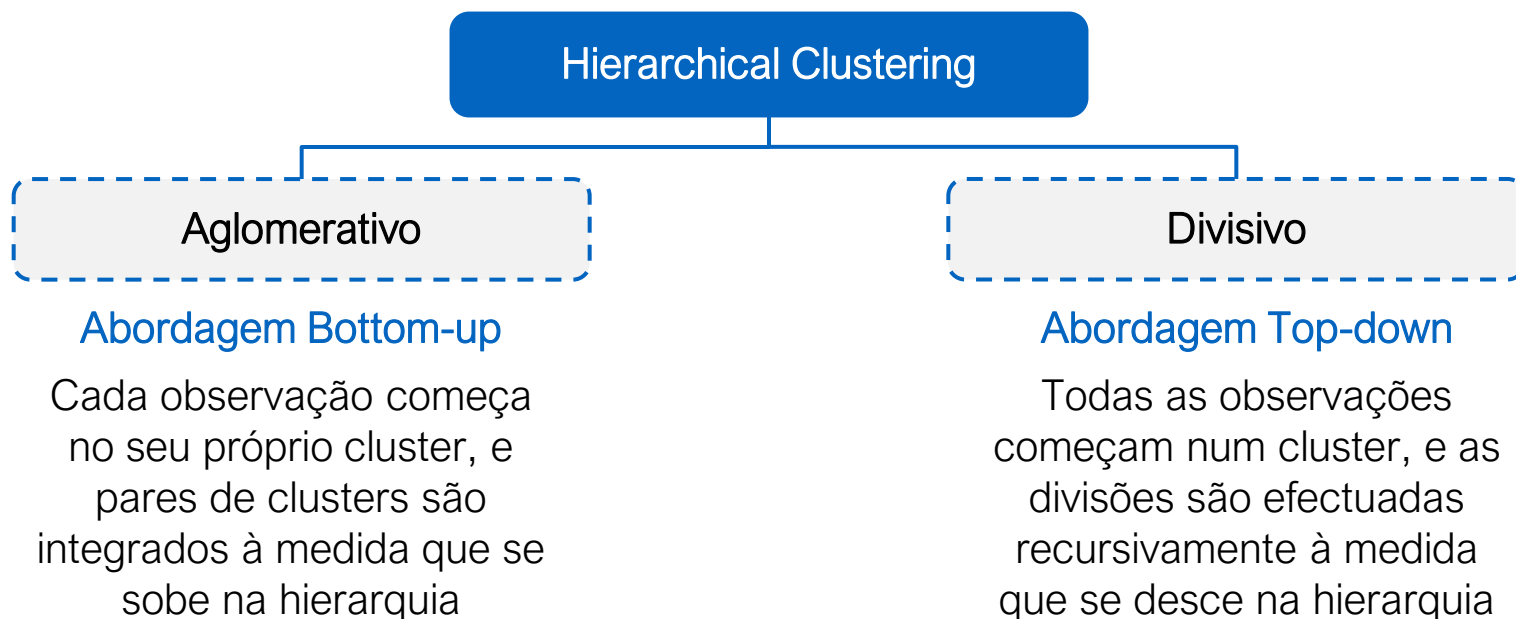
O objetivo principal é compreender o que se distingue em cada cluster.

Como interpretar os resultados?

- **Obter summary statistics para cada variável**, para o grupo de dados que compõe cada cluster;
- Analisar também em **termos de variáveis não utilizadas no clustering**;
- **Etiquetar o cluster** (por exemplo: clientes que compram tudo exceto roupa, clientes com churn, ...).

# 3. Hierarchical Clustering

### 3. Hierarchical Clustering



# 3. Hierarchical Clustering - Aglomerativo

## Passo 1

Definir uma função de distância: como calcular a distância entre 2 observações? (slides 10-14)

## Passo 2

Escolher um critério de ligação: como calcular a distância entre 2 clusters?

## Passo 3

Calcular hierarquia de clusters:

- Começar com cada observação no seu próprio cluster;
- Em cada iteração, integrar os dois clusters mais próximos;
- Parar quando todas as observações pertencerem a um único cluster.

## Passo 4

Escolher o número de clusters e obter a atribuição dos clusters.

# 3. Hierarchical Clustering - Aglomerativo

## Passo 1

Definir uma função de distância: como calcular a distância entre 2 observações? (slides 10-14)

## Passo 2

Escolher um critério de ligação: como calcular a similaridade entre 2 clusters?

## Passo 3

Calcular hierarquia de clusters:

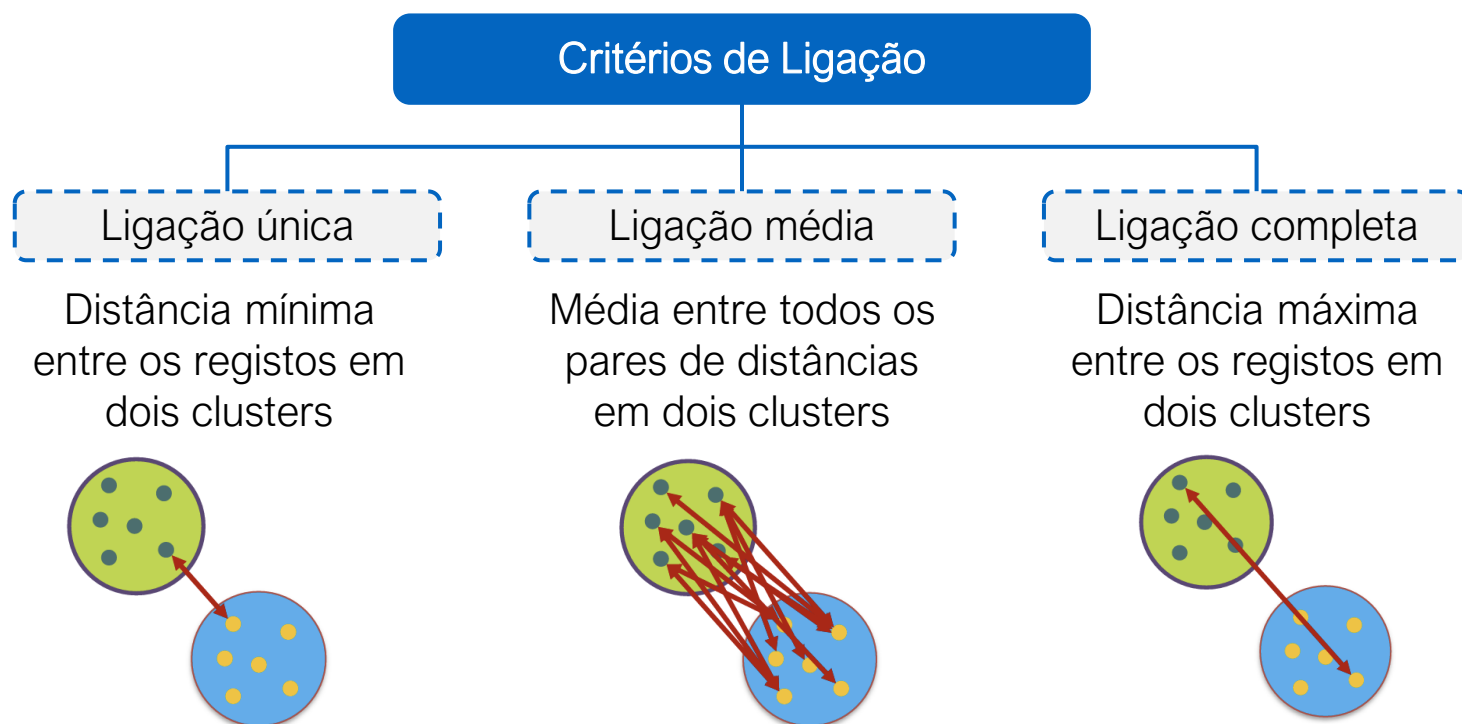
- Começar com cada observação no seu próprio cluster;
- Em cada iteração, integrar os dois clusters mais próximos;
- Parar quando todas as observações pertencerem a um único cluster.

## Passo 4

Escolher o número de clusters e obter a atribuição dos clusters.



### 3. Hierarchical Clustering - Aglomerativo



Quanto mais elevado o resultado da medida, menos similares (mais dissimilares) são os clusters.

### 3. Hierarchical Clustering - Aglomerativo

Método Ward (variância mínima): tem em conta a “perda de informação” que ocorre quando os registos são agrupados. Minimiza a variância total dentro do agrupamento.

- Quando cada cluster tem um registo, não há perda de informação.
- Quando os registos são agrupados, a informação sobre um registo individual é substituída pela informação do agrupamento a que pertence.
- Assim, o novo cluster corresponde à integração dos 2 clusters cuja combinação resulta na menor perda de informação (ou seja, menor aumento na variância do cluster).

Para medir a variância, é **usada a Sum of Squared Errors (SSE)**:

- Valores  $x_1$  (2, 6, 5, 6, 2, 2, 2, 2, 0, 0, 0)
- Média de  $x_1 = 2.5$
- $SSE = (2 - 2.5)^2 + (6 - 2.5)^2 + (5 - 2.5)^2 + \dots + (0 - 2.5)^2 = 50.5$

# 3. Hierarchical Clustering - Aglomerativo

## Passo 1

Definir uma função de distância: como calcular a distância entre 2 observações? (slides 10-14)

## Passo 2

Escolher um critério de ligação: como calcular a distância entre 2 clusters?

## Passo 3

Calcular hierarquia de clusters:

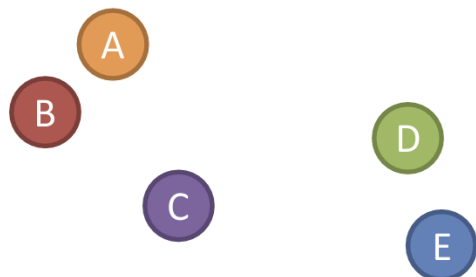
- Começar com cada observação no seu próprio cluster;
- Em cada iteração, integrar os dois clusters mais próximos;
- Parar quando todas as observações pertencerem a um único cluster.

## Passo 4

Escolher o número de clusters e obter a atribuição dos clusters.

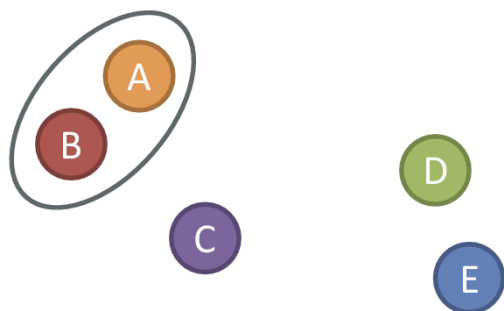
### 3. Hierarchical Clustering - Aglomerativo

Passo 1: Começar com  $n$  clusters, onde  $n$  corresponde ao número de observações.



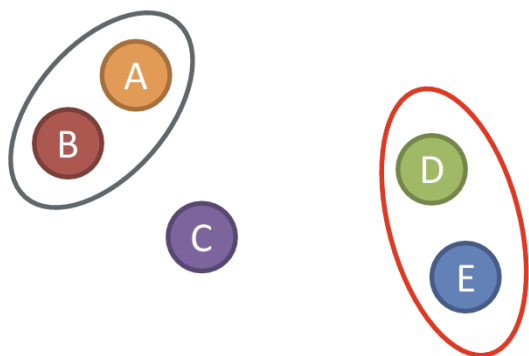
### 3. Hierarchical Clustering - Aglomerativo

Passo 2: Calcular dissimilaridade entre clusters (compostos por 1 observação cada) e integrar num novo cluster as 2 mais similares.



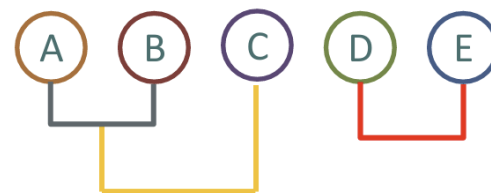
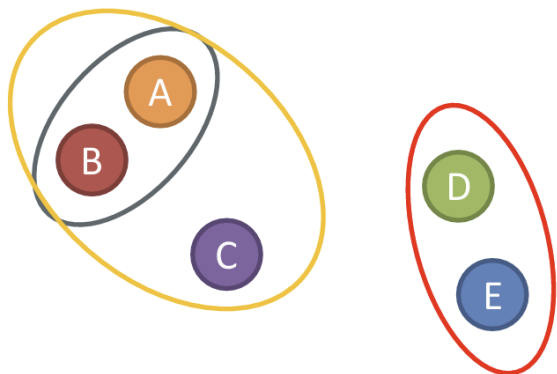
### 3. Hierarchical Clustering - Aglomerativo

Passo 3: Em cada iteração seguinte, os 2 clusters mais próximos (segundo o critério de dissimilaridade escolhido) são combinados.

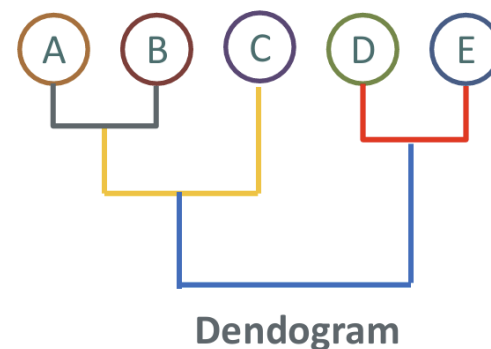
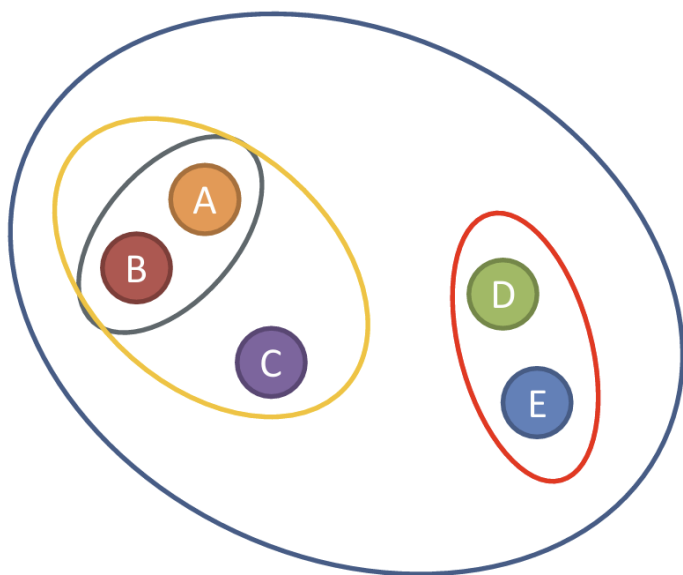


### 3. Hierarchical Clustering - Aglomerativo

Passo 3: Em cada iteração seguinte, os 2 clusters mais próximos (segundo o critério de dissimilaridade escolhido) são combinados.



### 3. Hierarchical Clustering - Aglomerativo





# 3. Hierarchical Clustering - Aglomerativo

## Passo 1

Definir uma função de distância: como calcular a distância entre 2 observações? (slides 10-14)

## Passo 2

Escolher um critério de ligação: como calcular a distância entre 2 clusters?

## Passo 3

Calcular hierarquia de clusters:

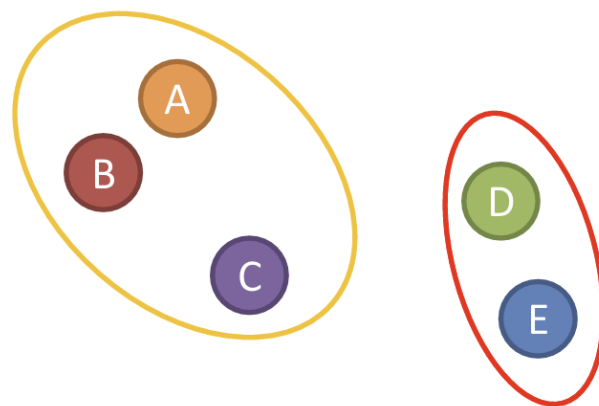
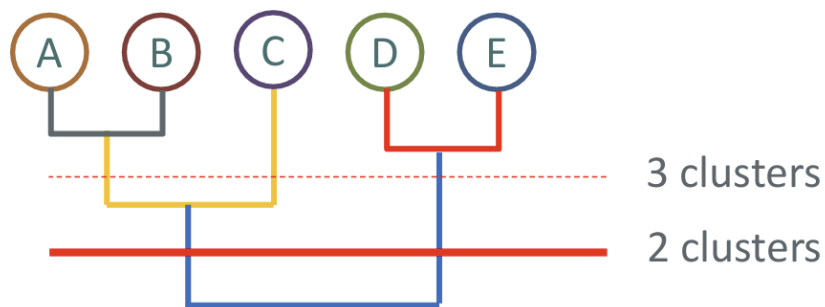
- Começar com cada observação no seu próprio cluster;
- Em cada iteração, integrar os dois clusters mais próximos;
- Parar quando todas as observações pertencerem a um único cluster.

## Passo 4

Escolher o número de clusters e obter a atribuição dos clusters.


### 3. Hierarchical Clustering - Aglomerativo


**Determinar o número de clusters:** as observações são afectadas aos clusters desenhando uma linha horizontal através do dendrograma. As observações que se juntam acima da linha formam os clusters.





### 3. Hierarchical Clustering – Vantagens e Desvantagens

O hierarchical clustering é um algoritmo “guloso”. Por conseguinte:

 É mais simples de compreender e interpretar que os algoritmos particionais.

 Uma vez efectuada uma interação (fusão ou divisão), não pode voltar atrás → Pode conduzir a soluções não óptimas.

 Computacionalmente intensivo para grandes volumes de dados → requer o cálculo de distâncias entre todos os pares de pontos de dados.

 Não pode ser utilizado para “prever” novas observações.

# 4. K-Means

## 4. K-Means

Método de clustering que tem por **objetivo dividir  $n$  observações em  $k$  clusters** em que cada observação **pertence ao cluster com a média das respetivas variáveis mais próxima**.

Dado um conjunto de dados com  $n$  objectos, o K-means **constrói  $k$  partições, em que cada partição representa um cluster ( $k \leq n$ )**.

Os dados são divididos em  $k$  clusters, satisfazendo as seguintes condições:

- Cada cluster contém pelo menos um objeto.
- Cada objeto pertence apenas a um cluster.

## 4. K-Means

O algoritmo executa os seguintes passos iterativamente:

1. Inicialização das seeds (normalmente aleatoriamente);
2. Cada observação é associada à seed mais próxima;
3. Calcula os centroides dos clusters formados;
4. Executa novamente o passo 2;
5. Processo termina quando os centroides deixam de mudar (ou as mudanças são mínimas, abaixo do threshold pré definido).

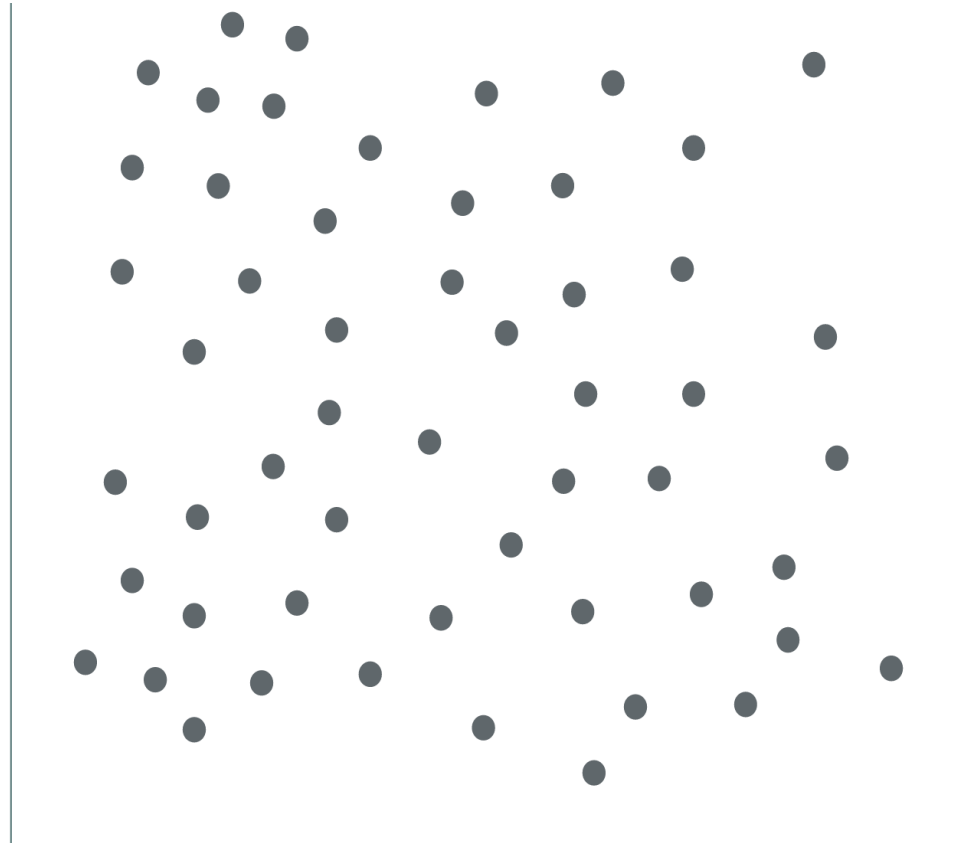
Este processo iterativo garante que a distância intra cluster é minimizada (Within Cluster Sum of Squares), onde  $x$  corresponde às variáveis e  $c$  aos centroides:

$$WCSS = \sum_{j=1}^K \sum_{i=1}^n distance(x_i, c_j)^2$$

## 4. K-Means

Os dados:

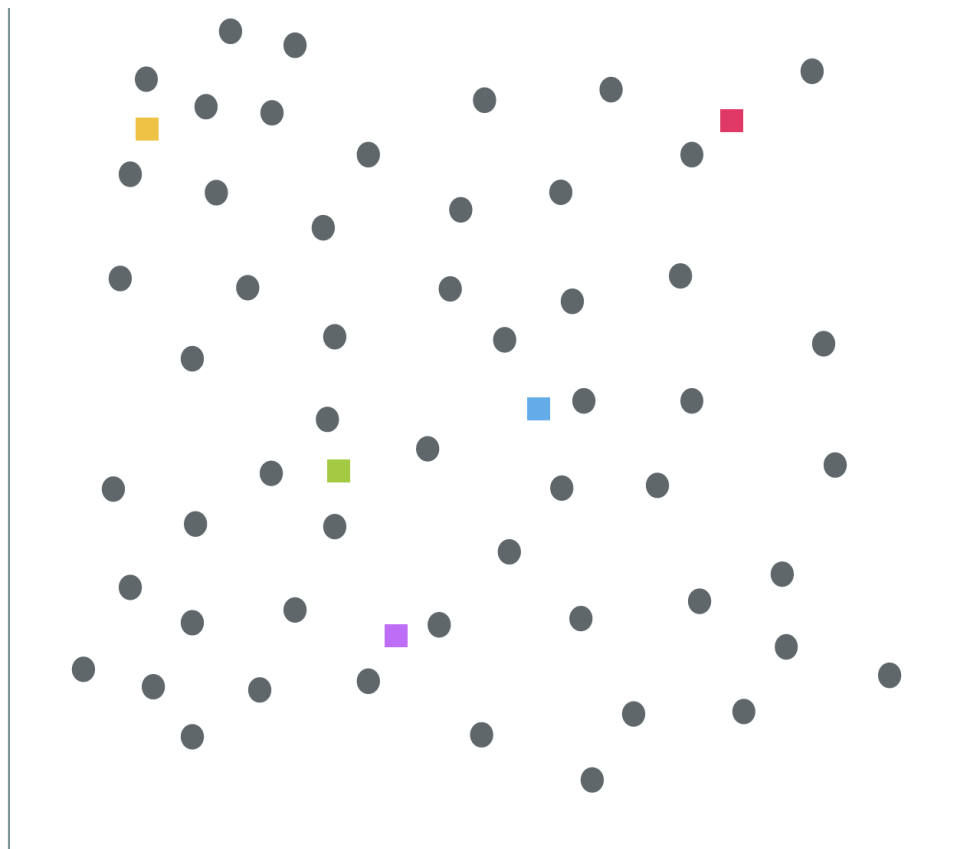
- 2 variáveis
- Agrupar em 5 clusters!



## 4. K-Means

### Inicialização

- Definir as seeds iniciais (normalmente aleatório)

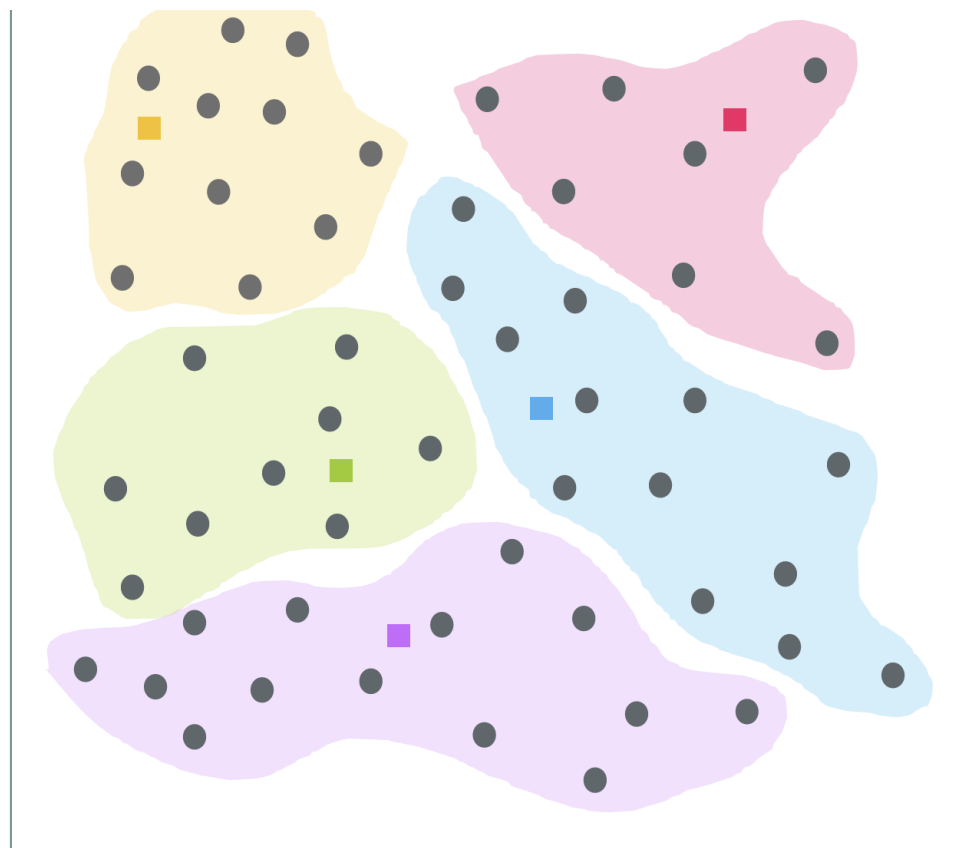




## 4. K-Means

### Iteração 1 – 1º passo

- Definir a seed mais próxima para cada observação

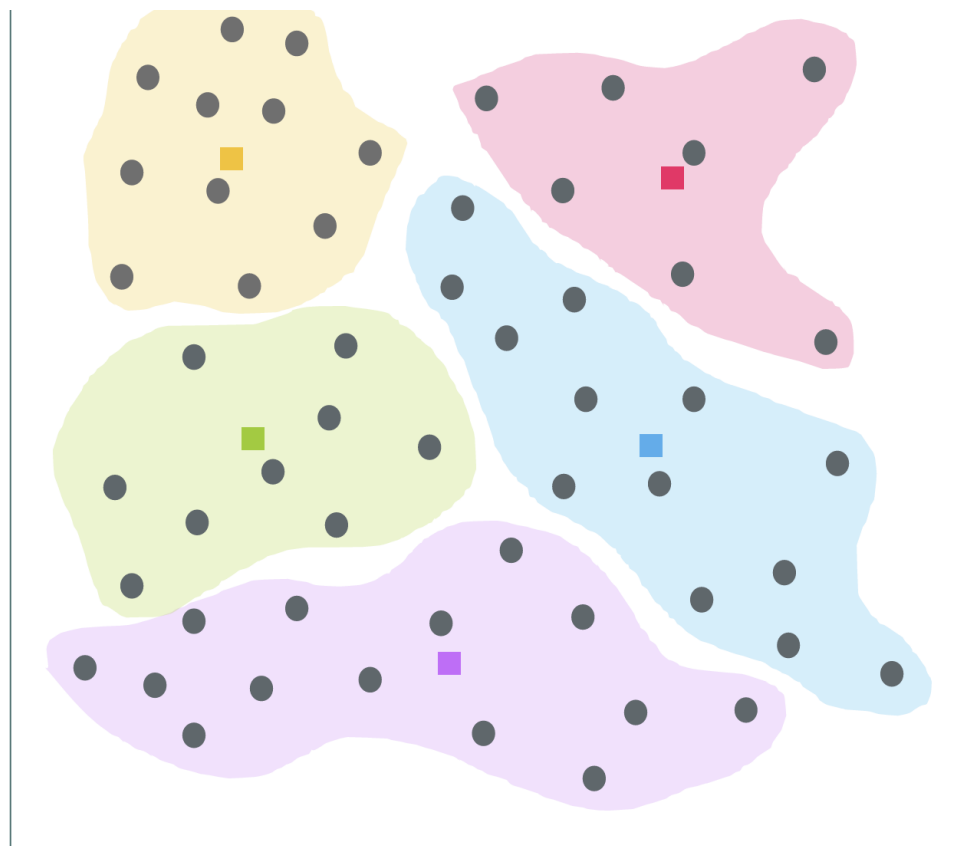


## 4. K-Means

### Iteração 1 – 2º passo

- Recalcular a seed de modo a que fique na nuvem de pontos, representando o seu centro (designado por centroide)

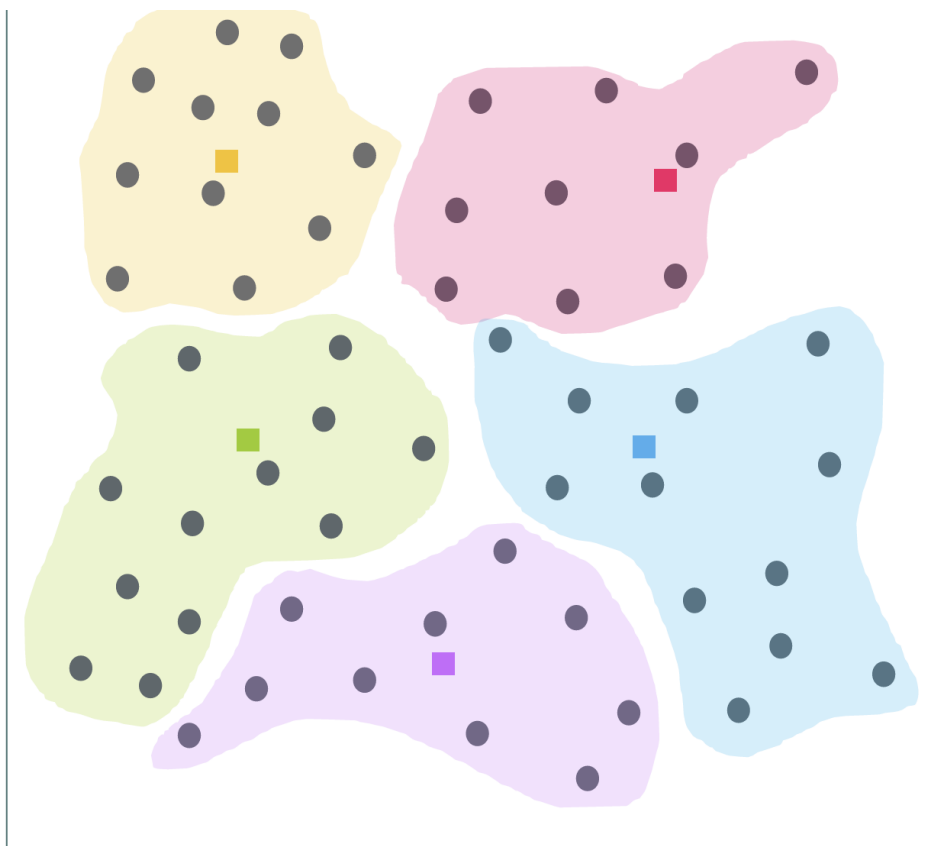
E REPETIR ITERAÇÃO....



## 4. K-Means

### Iteração 2 – 1º passo

- Definir a seed mais próxima para cada observação

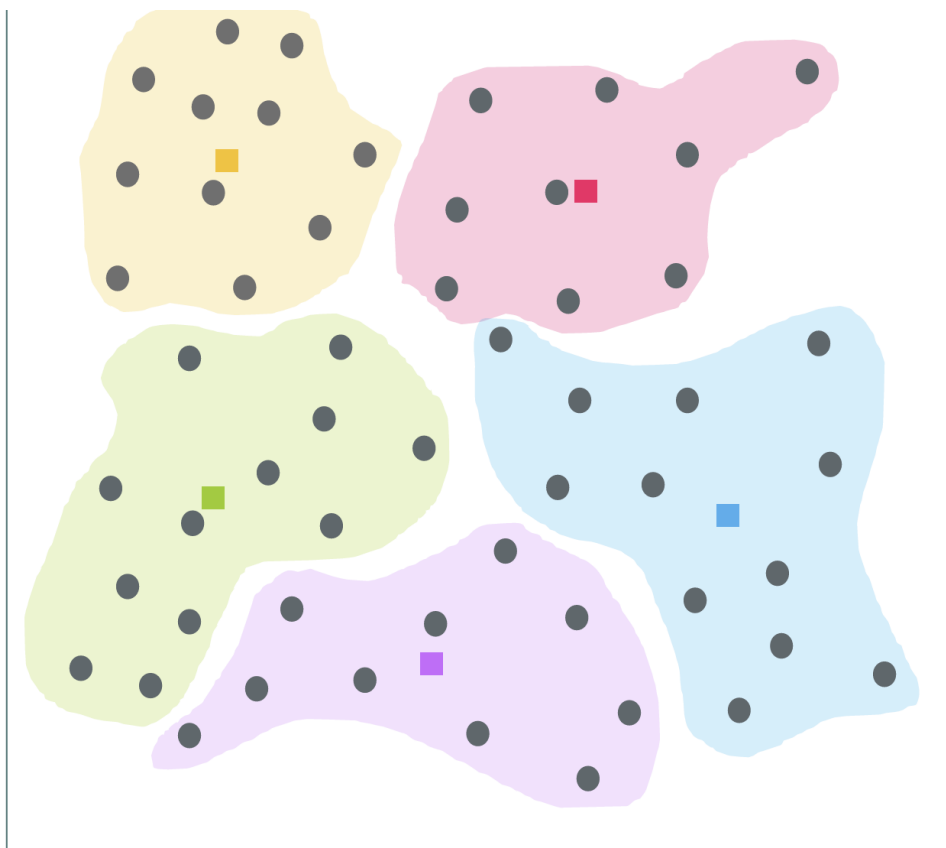


## 4. K-Means

Iteração 2 – 2º passo

- Recalcular o centroid

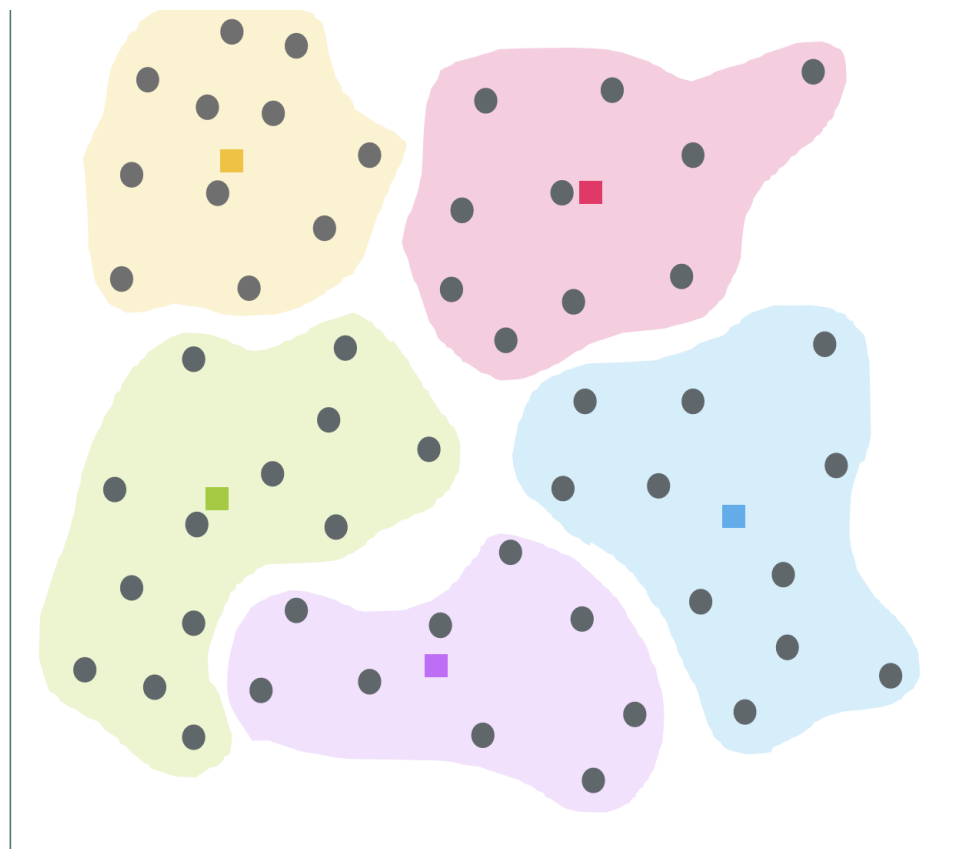
E REPETIR ITERAÇÃO....



## 4. K-Means

### Iteração 3 – 1º passo

- Definir a seed mais próxima para cada observação

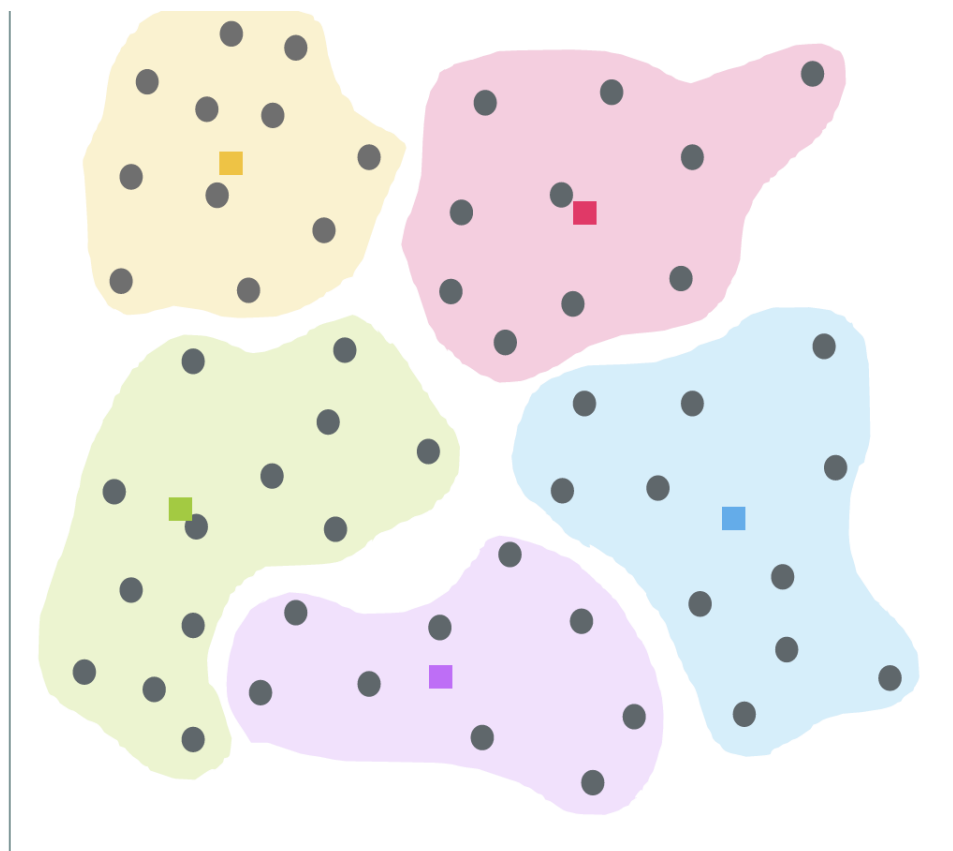


## 4. K-Means

Iteração 3 – 2º passo

- Recalcular o centroid

E REPETIR ITERAÇÃO....

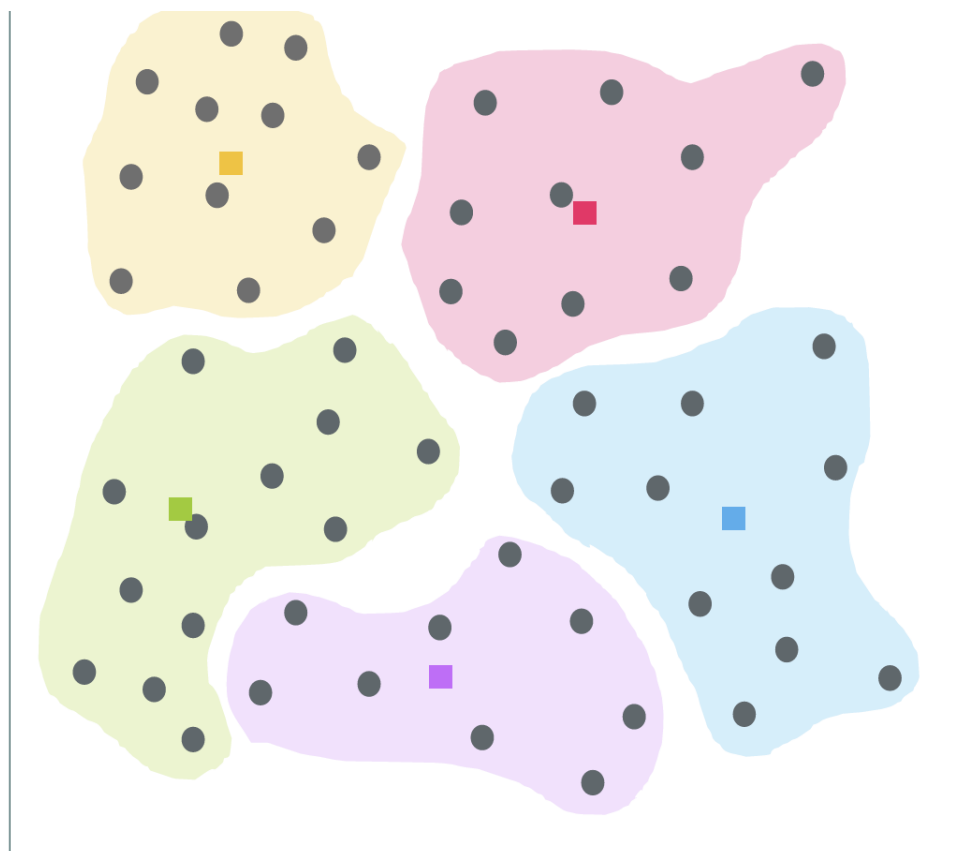


## 4. K-Means

### Iteração 4 – 1º passo

- Definir a seed mais próxima para cada observação

SEM MAIS ALTERAÇÕES...  
Solução final!

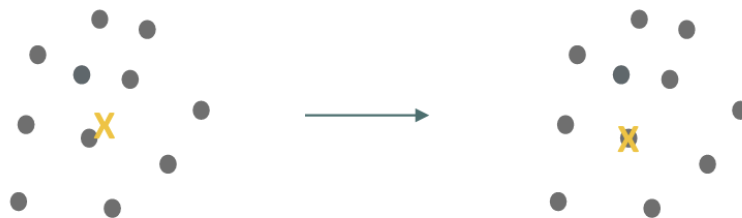


## 4. K-Means – Variantes

### K-Medoids

- Variante do K-Means que altera a forma como o centróide de cada cluster é definido;
- Cada cluster é representado por uma das observações pertencentes ao cluster.

Pode ser usado de modo alternativo ao K-Means devido à sua robustez em relação a outliers, maior interpretabilidade e permite a aplicação de mais funções de distância.



### K-Mode

- Variante do K-Means para dados categóricos, que utiliza a moda em vez da média
- Utiliza uma medida de dissimilaridade simples – Hamming distance.



# 4.K-Means – Vantagens e Desvantagens



Simples de implementar, perceber e interpretar.



Fácil adaptabilidade a novas observações.



Rápida implementação.



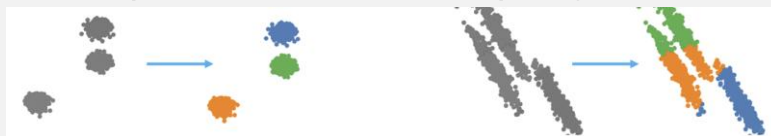
O número de seeds: é necessário definir o número de clusters a criar à priori.



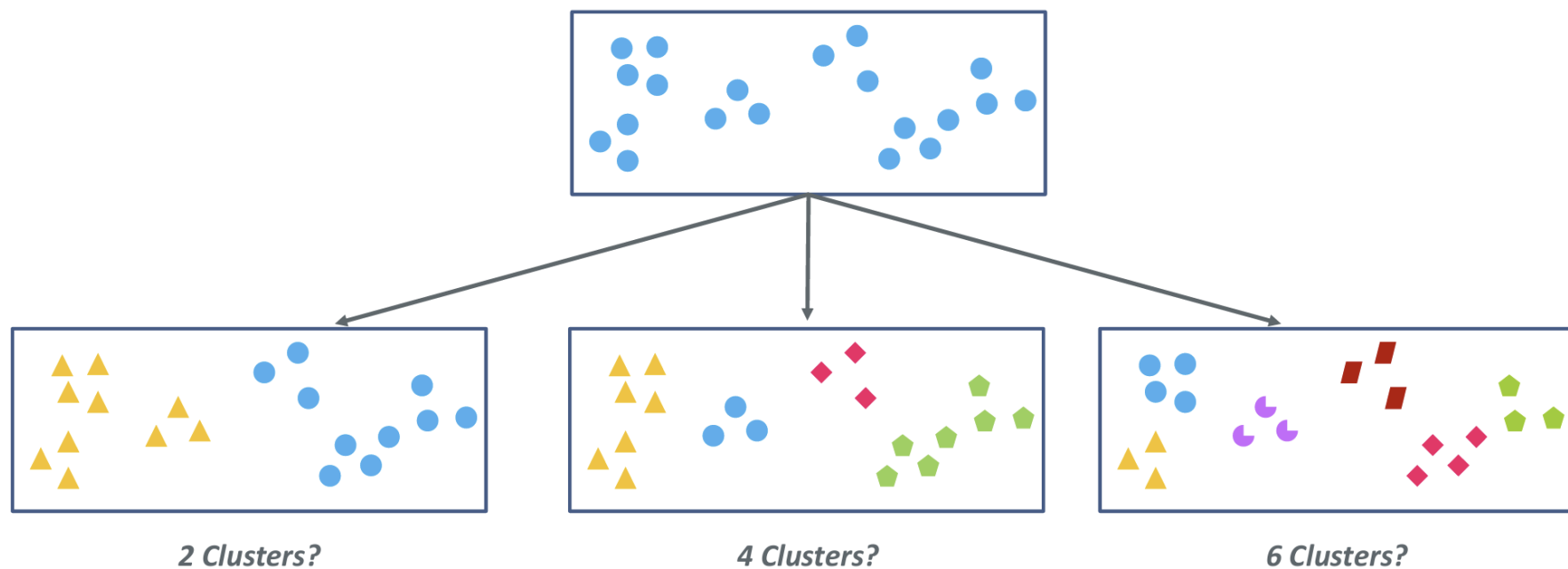
A inicialização: sensível às posições iniciais das seeds, bem como à existência de outliers.



A “forma” dos dados: os métodos de partição funcionam bem com clusters de formas esféricas. Para dados com formas complexas, os métodos de partição não são a melhor escolha.



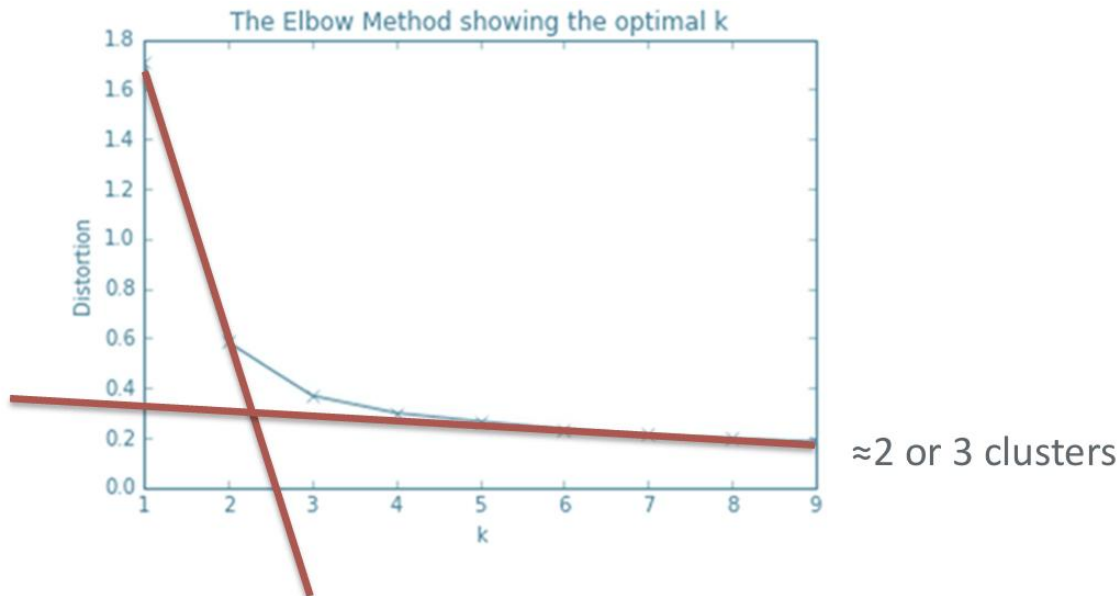
## 4. K-Means – Quantos clusters?



## 4. K-Means – Quantos clusters?

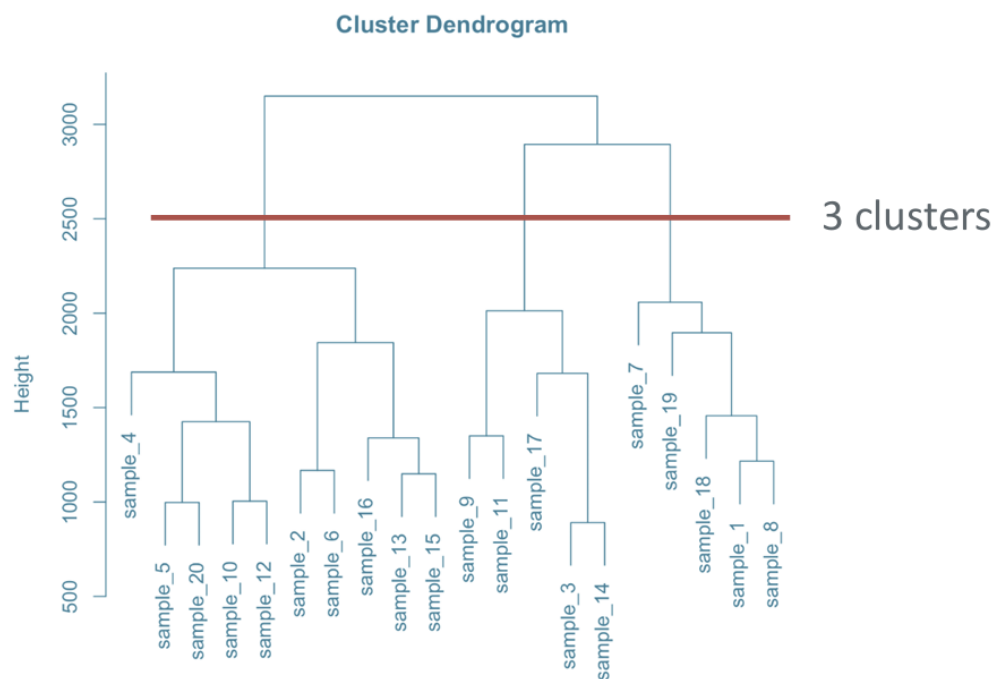
Opção 1: Produzir várias soluções de clusters com diferentes  $k$ , e escolher a melhor solução (Elbow method).

Ou seja, o número de clusters  $k$  para os quais a distorção (WCSS) é menor, mantendo o  $k$  a um nível razoável e interpretável.



## 4. K-Means – Quantos clusters?

Opção 2: Utilizar um método hierárquico para escolher o número de clusters com base no dendrograma



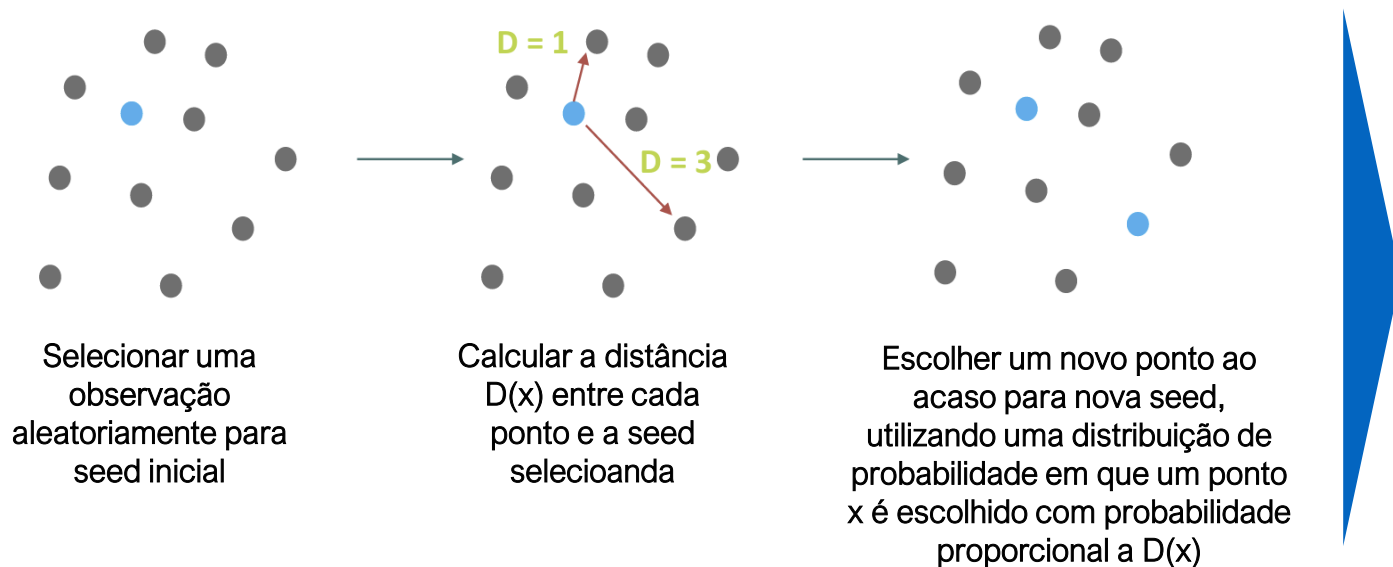
## 4. K-Means – Problema de Inicialização

Um dos problemas do K-means é a sua sensibilidade às posições iniciais das seeds.

[Uma solução possível: K-Means ++](#)

Variante do K-Means com uma nova inicialização das seeds.

- Objetivo: Abordar a sensibilidade do modelo à inicialização;
- Intuição: A dispersão das seeds iniciais conduz a boas seeds finais.



Repetir passos 2 e 3 até  $k$  seeds terem sido selecionadas.

Depois, prosseguir normalmente para definição de clusters tal como no K-Means.

# Obrigado!