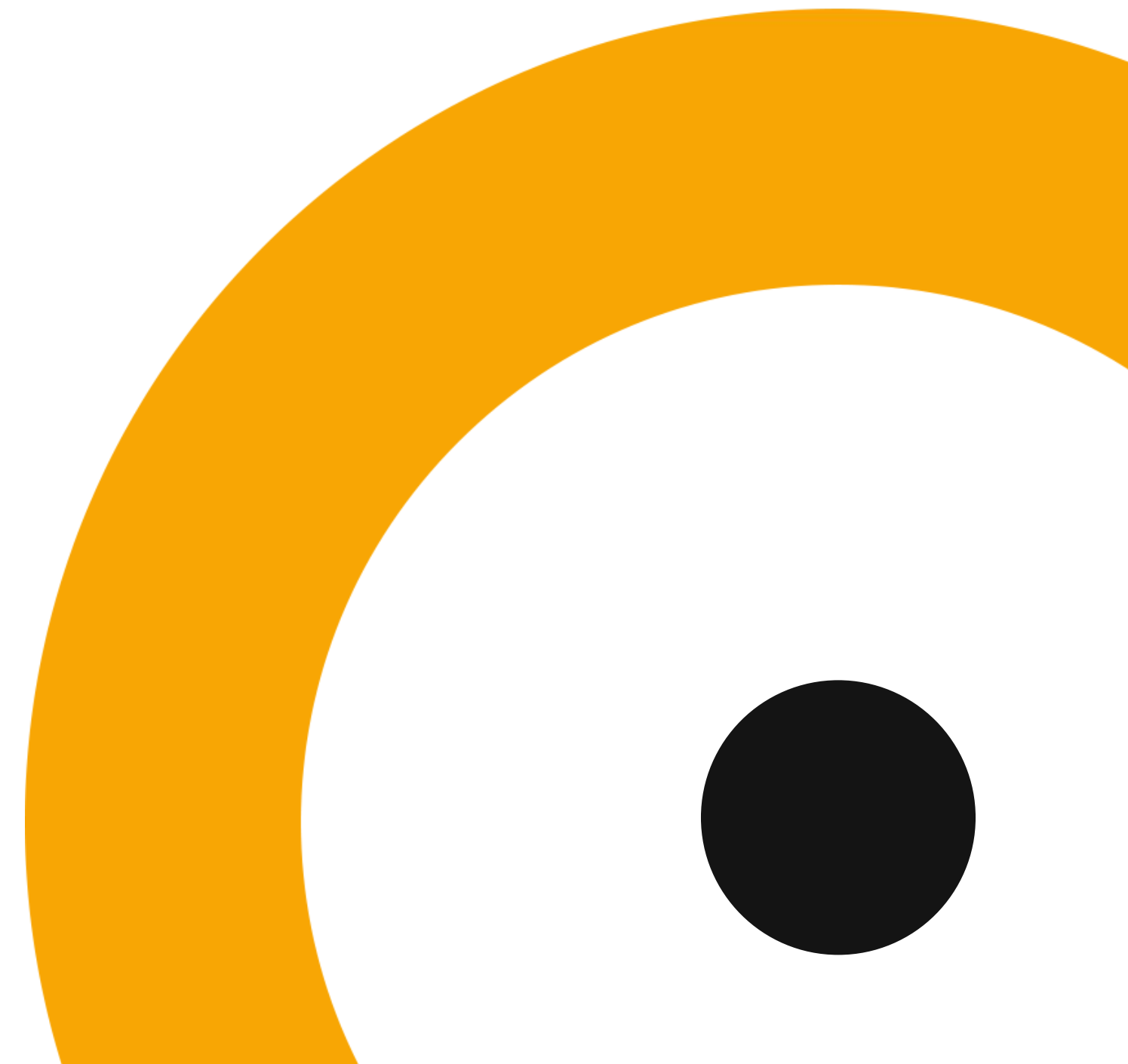


Mestrado em Data Science

# Séries Temporais

Clustering de Séries Temporais





3

# Clustering de Séries Temporales





# Clustering

Clustering é uma tarefa não supervisionada

**Objetivo é agrupar um conjunto de dados em diferentes grupos**

- Os elementos de cada grupo são semelhantes entre si, mas diferentes dos elementos dos restantes grupos
- Por vezes, é uma tarefa inserida no processo de exploração de dados de forma a descobrir a sua estrutura
  - Séries temporais com dinâmicas semelhantes
- **São usadas medidas de semelhança ou distância para calcular a proximidades entre duas observações**



# Clustering de Séries Temporais

**Cada observação representa uma série temporal**

Exemplos:

- Agrupar um conjunto de stocks (e.g. representadas por séries temporais do seu preço) para descobrir stock semelhantes
- Agrupar produtos de retalho com números de vendas semelhantes
- Entre muitos outros



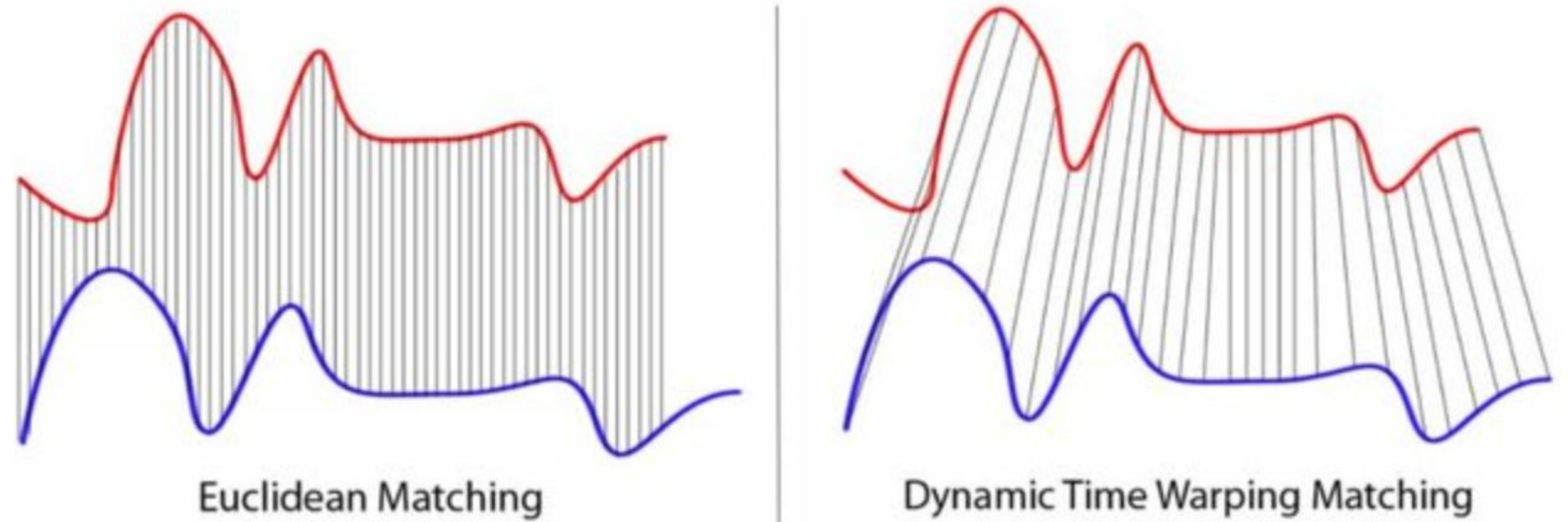


# Clustering de Séries Temporais

**Cada observação representa uma série temporal**

Desafios:

- Cada série temporal pode ser longa
  - Conjunto de dados de grande dimensão
- Tamanho das diferentes séries pode ser irregular
- Desfasamento entre padrões semelhantes





# Abordagens Principais

## Baseado nos dados brutos (*Raw-data-based*)

- O input principal são as séries temporais sem qualquer processamento
- É utilizada uma medida de semelhança apropriada para séries temporais

## Baseado em variáveis explicativas (*Feature-based*)

- As séries temporais são sumarizadas num vector de baixa dimensão
- Depois, um algoritmo de clustering convencional é aplicado

## Baseado em modelos (*Model-based*)

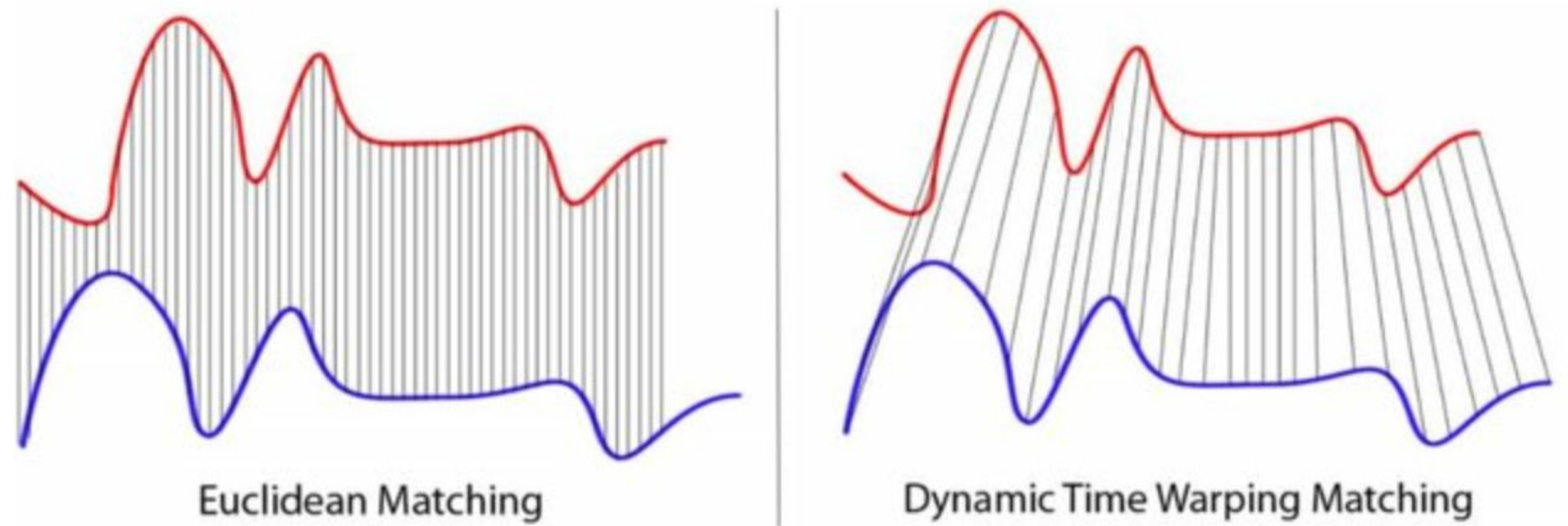
- As séries temporais são modeladas com um dado algoritmo, e representadas pelos respetivos parâmetros do modelo obtido
- Depois, um algoritmo de clustering convencional é aplicado



# Medidas de Distância

Servem para quantificar a proximidade entre observações

- Euclidiana
- Manhattan
- Correlação
- *Dynamic Time Warping*
- Entre outras...





# Medidas de Distância

Servem para quantificar a proximidade entre observações

- Euclidiana  $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- Manhattan  $d(x, y) = \sum_{i=1}^n |x_i - y_i|$





# Avaliação

## Como avaliar a qualidade dos agrupamentos?

- Medidas que quantificam a quão bem os grupos estão separados entre si (e compactos internamente)

Medidas tipicamente utilizadas:

- Coeficiente de silueta (*Silhouette*)
- Davies–Bouldin index
- Soma dos erros quadrados (Within-Cluster-Sum of Squared Errors)

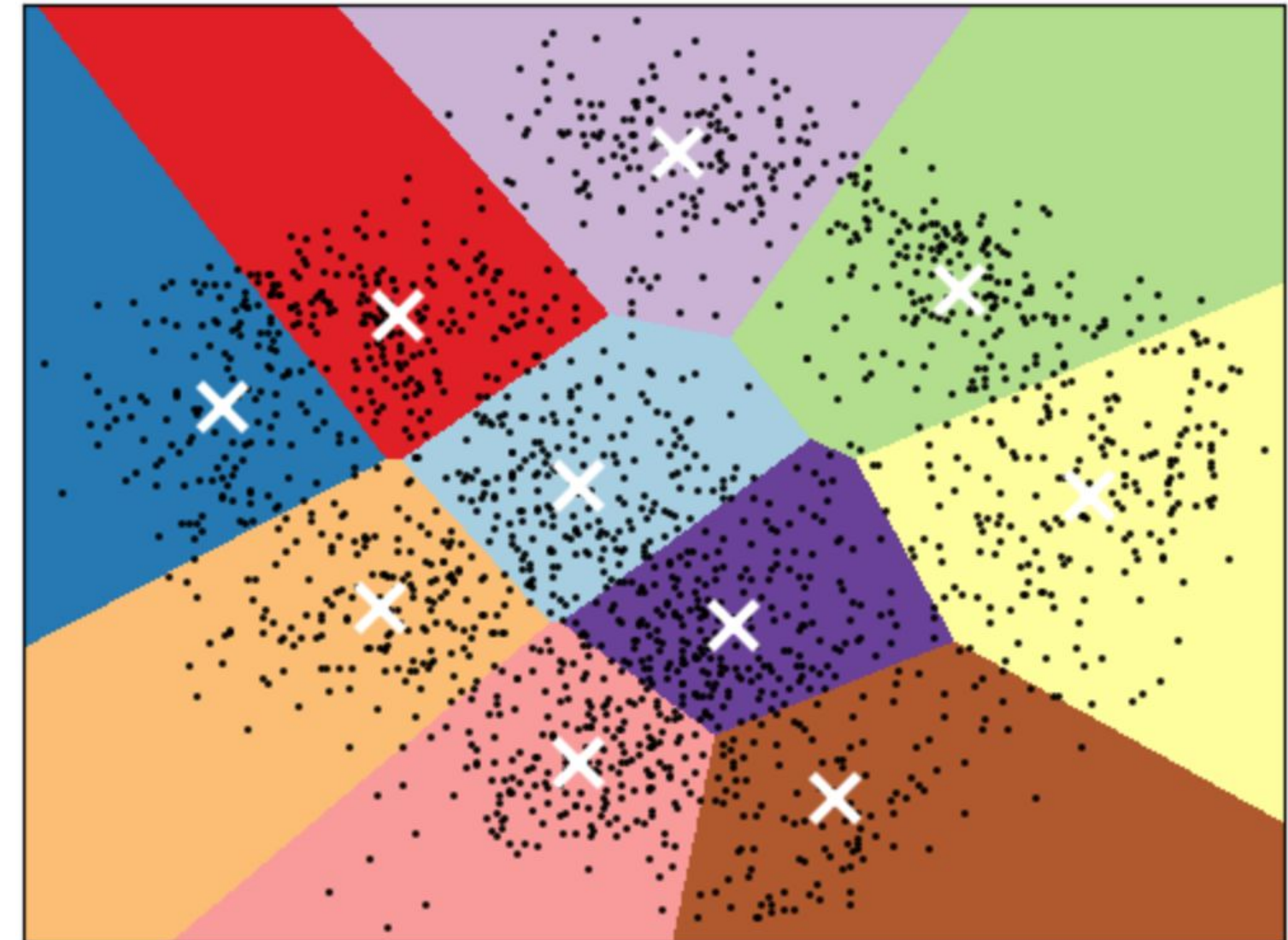


# Métodos de Clustering

## K-Means

- Método popular para clustering

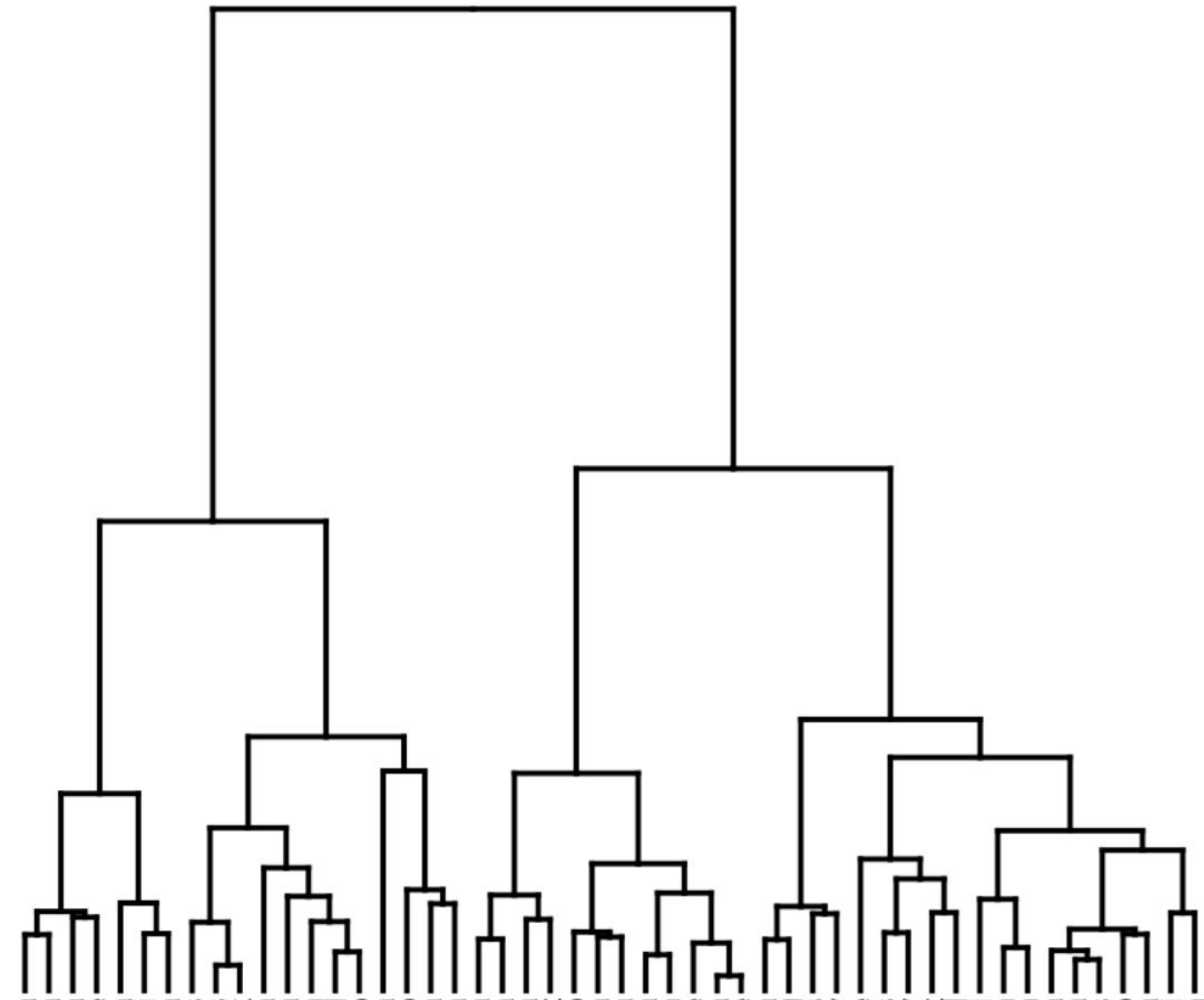
1. Primeiro, seleciona  $k$  observações iniciais como centróides
2. Repetir até convergência:
  - a. Formar  $k$  grupos ao atribuir observações ao centróide mais próximo
  - b. Re-calcular centróides (ponto médio) de cada grupo





# Métodos de Clustering

- K-Medoids
  - Parecido com K-means, mas os centróides representam observações concretas
- Clustering hierárquico
  - aglomerativo
  - divisivo





# Como Escolher o Número de Grupos?

## Método do Cotovelo

- Escolher o número de grupos no ponto de inflexão da curva do erro (cotovelo)

