

Tera

AULA 27

Unsupervised Learning:
Clustering

Instrutor: [Raphael Ballet](#)

Background:

- Engenheiro de Controle e Automação (IMT)
- Mestre em Sistemas Aeroespaciais e Mecatrônica (ITA)
- Lead Data Scientist - Elo7

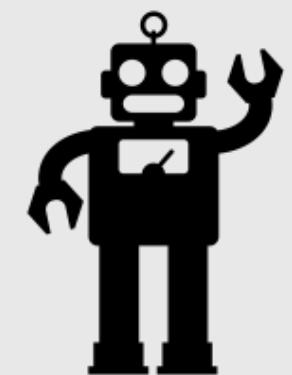
Interesses:



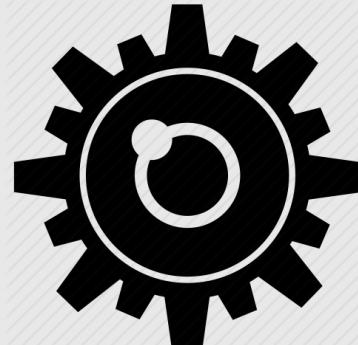
Drones



Aprendizado
de Máquina



Robótica



Visão
Computacional



Processamento de
Linguagem Natural



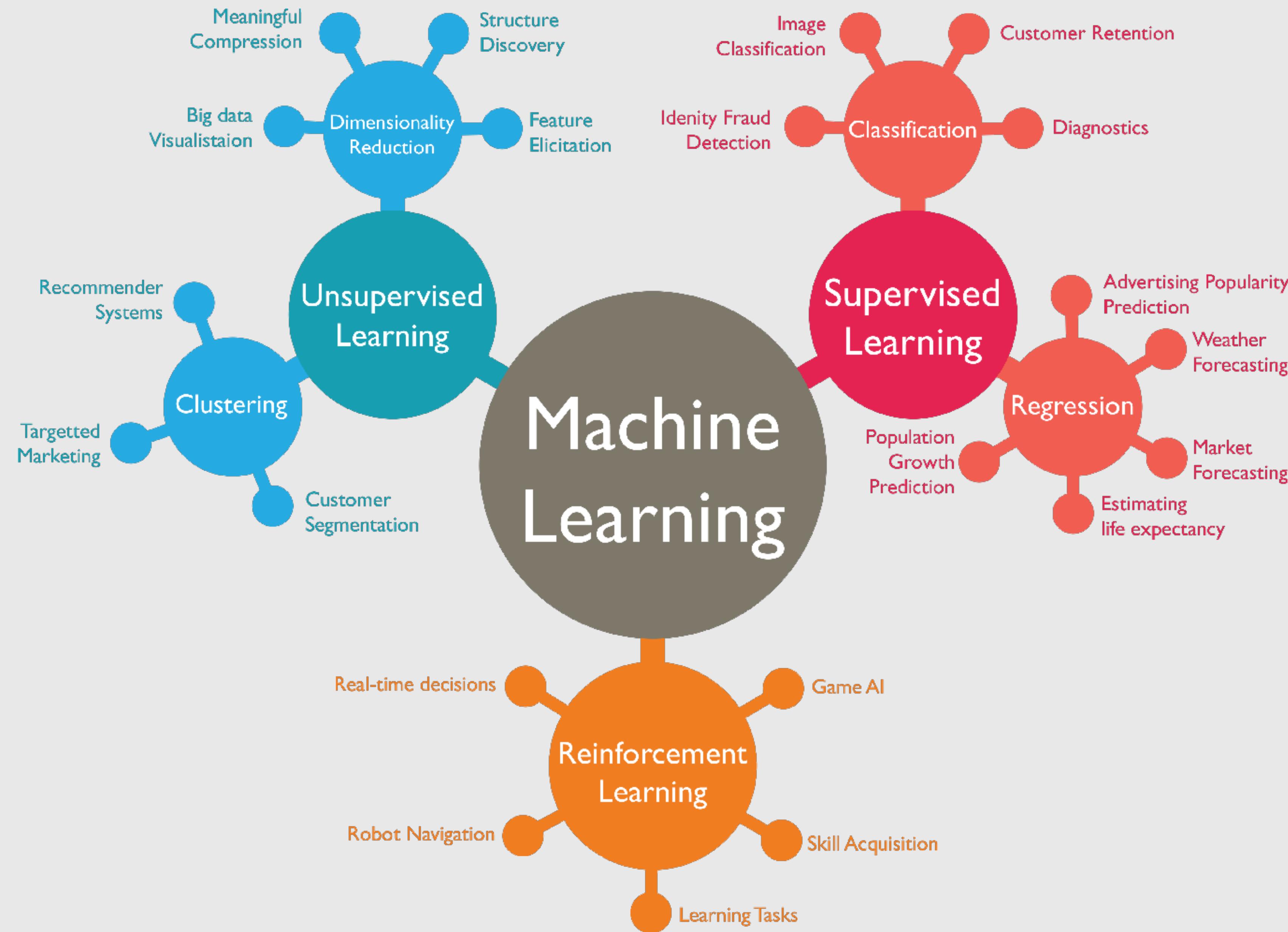
Sistemas de
recomendação

Planejamento:

1. Introdução a clustering
2. Clustering: K-Means
3. Clustering: DBSCAN
4. Clustering: Hierarchical Clustering
5. Clustering em NLP
6. Case Final

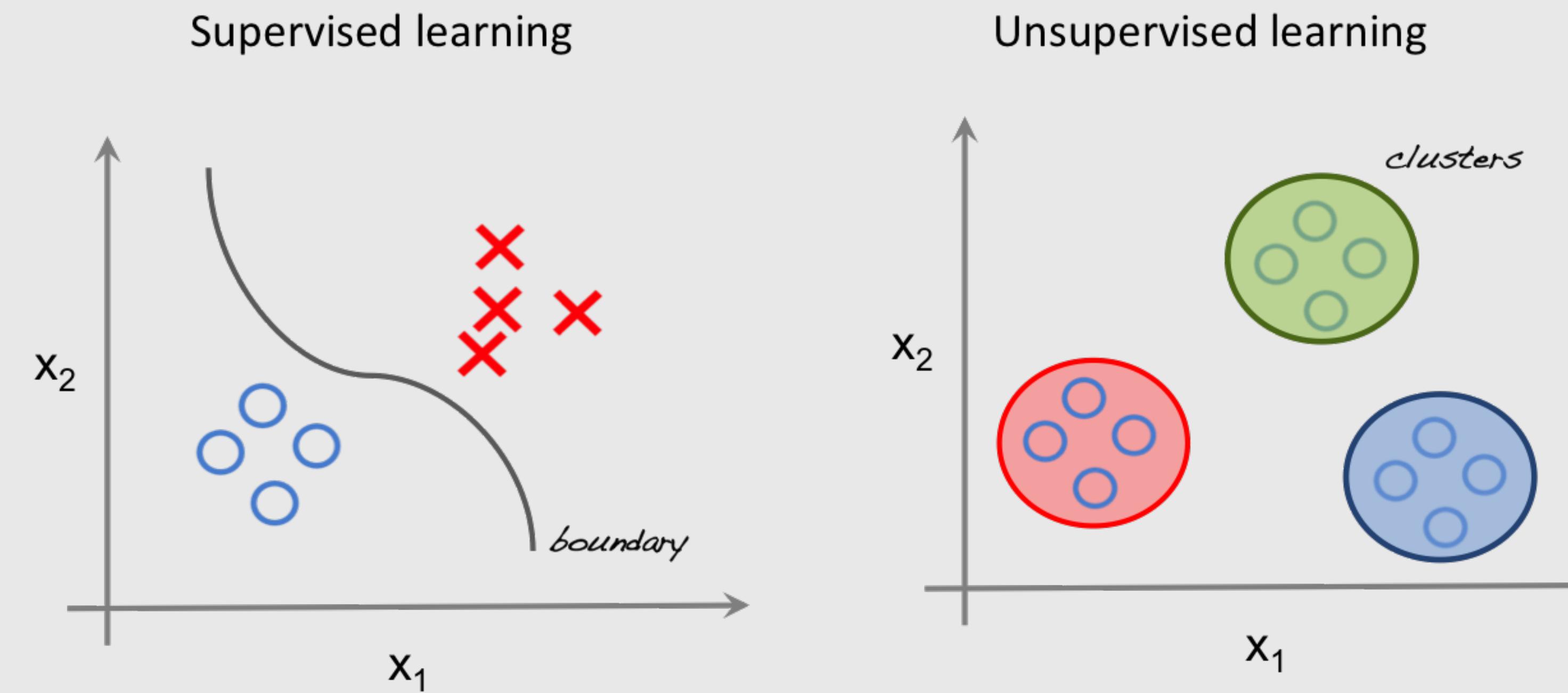
T

1. INTRODUÇÃO



1. INTRODUÇÃO:

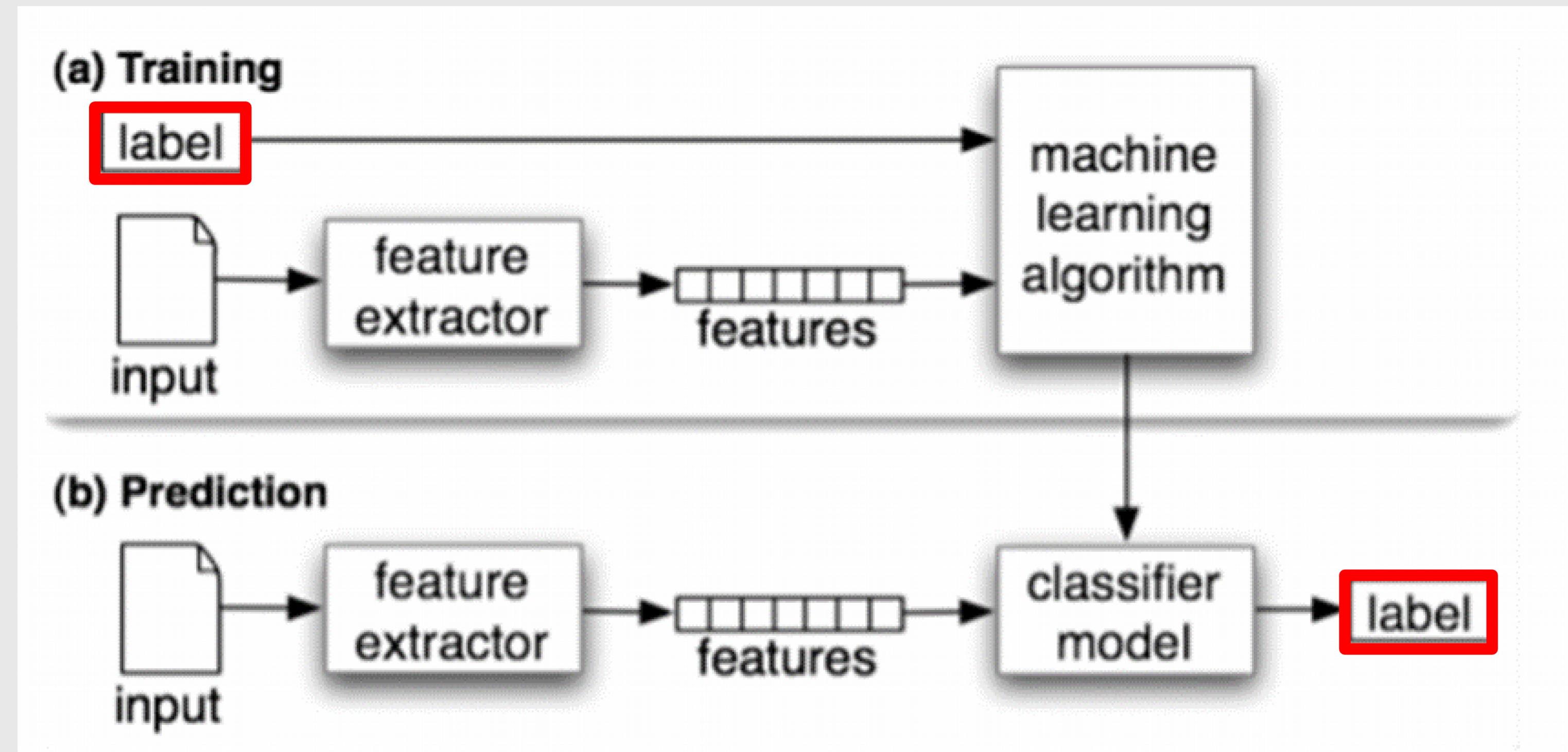
- Aprendizagem supervisionada vs.
Aprendizagem não supervisionada



I

Aprendizagem supervisionada:

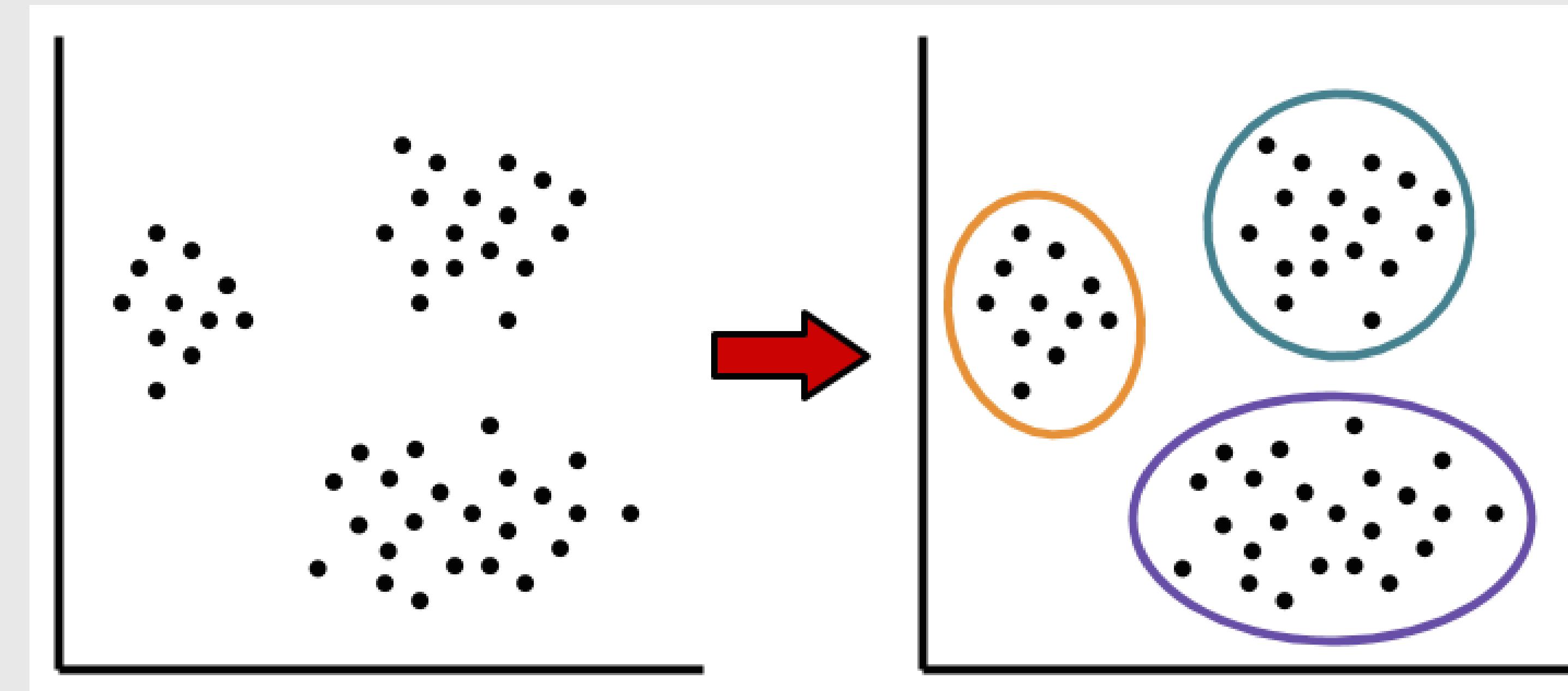
- Dados possuem **rótulo** (label)
- Treinamento com rótulo = Supervisão
- Métodos: Classificação / Regressão



T

Aprendizagem não supervisionada:

- Dados não rotulados (sem supervisão)
- Métodos: Clustering / density estimation
- Análises exploratórias
- Métricas de sucesso qualitativas e não global
- Ex: padrões de usuários, cluster de produtos / clientes etc.



INTRODUÇÃO:

- **Objetivos:**
 - Análise exploratória
 - Encontrar padrões nos dados
 - Resumir dados



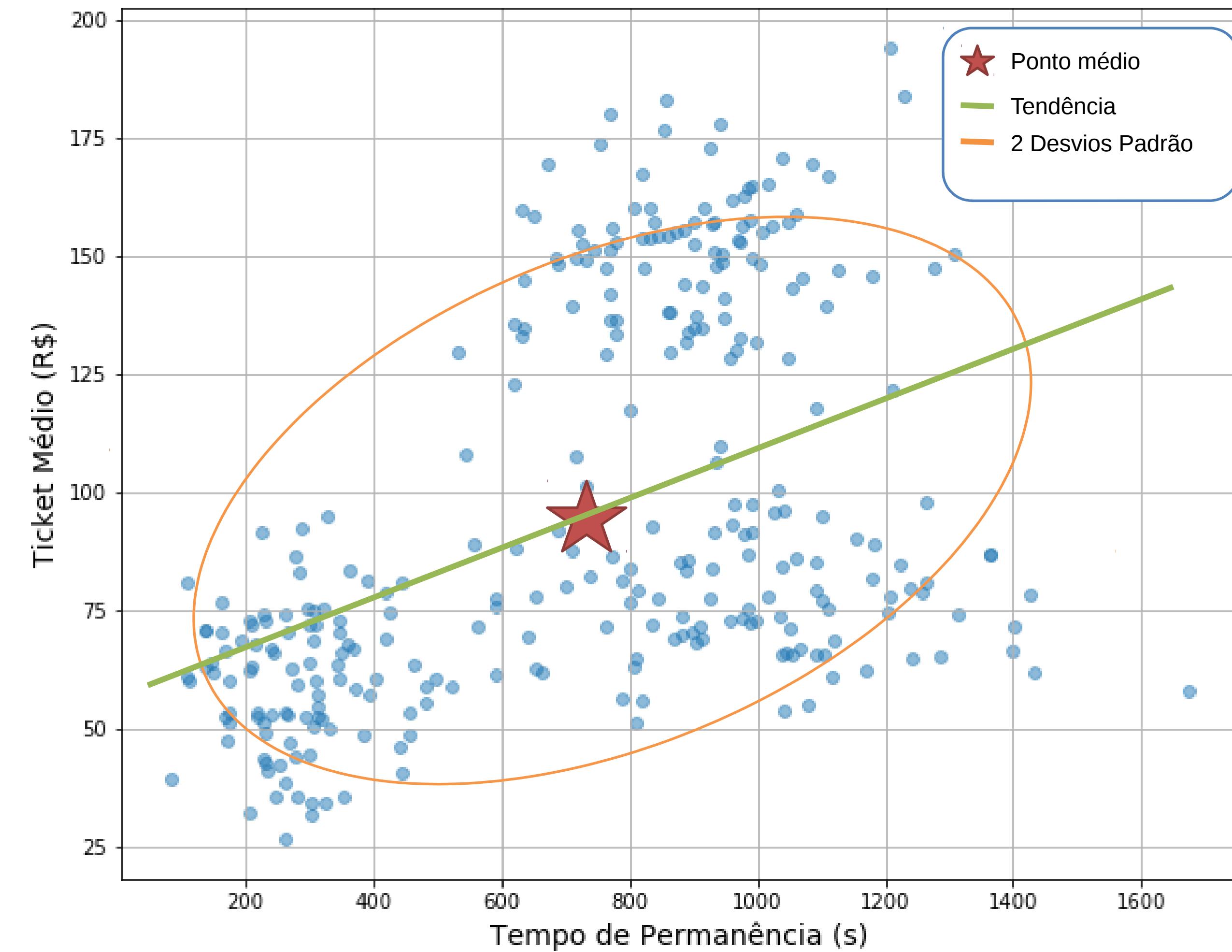
T

Exemplo:

- **Exemplo real Elo7:**
 - Análise exploratória (Notebook)

Exemplo Elo7:

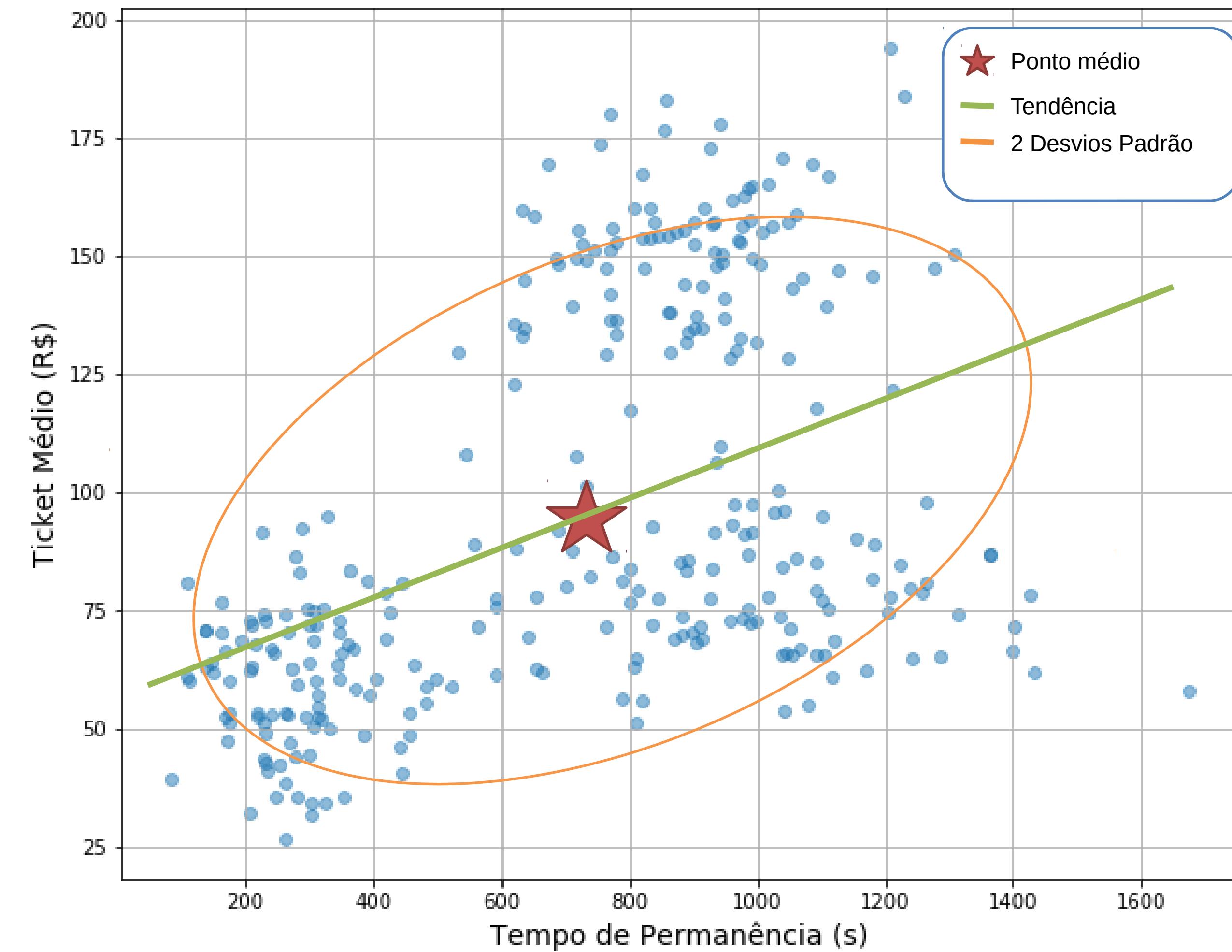
- **Análise exploratória:**
 - Medidas centrais
 - Tendência



Exemplo Elo7:

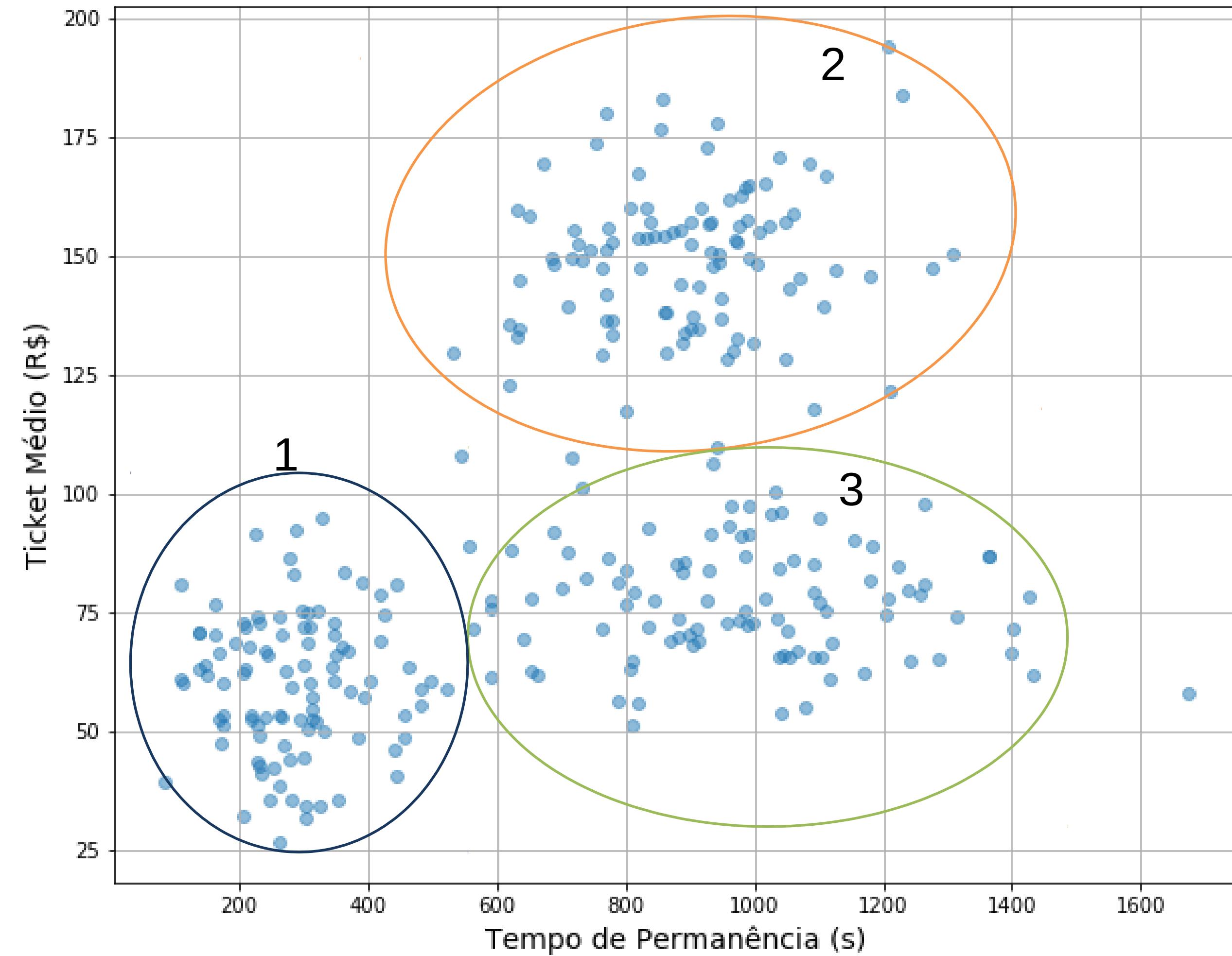
- **Análise exploratória:**
 - Medidas centrais
 - Tendência

Há algo de errado nessa análise?



Exemplo Elo7:

- Análise exploratória:
 - Medidas centrais
 - Tendência
 - Segmentação / Clustering
 - Separa (ou agrupa) fontes dos dados



Exemplo Elo7:

- **Segmentação usuários:**

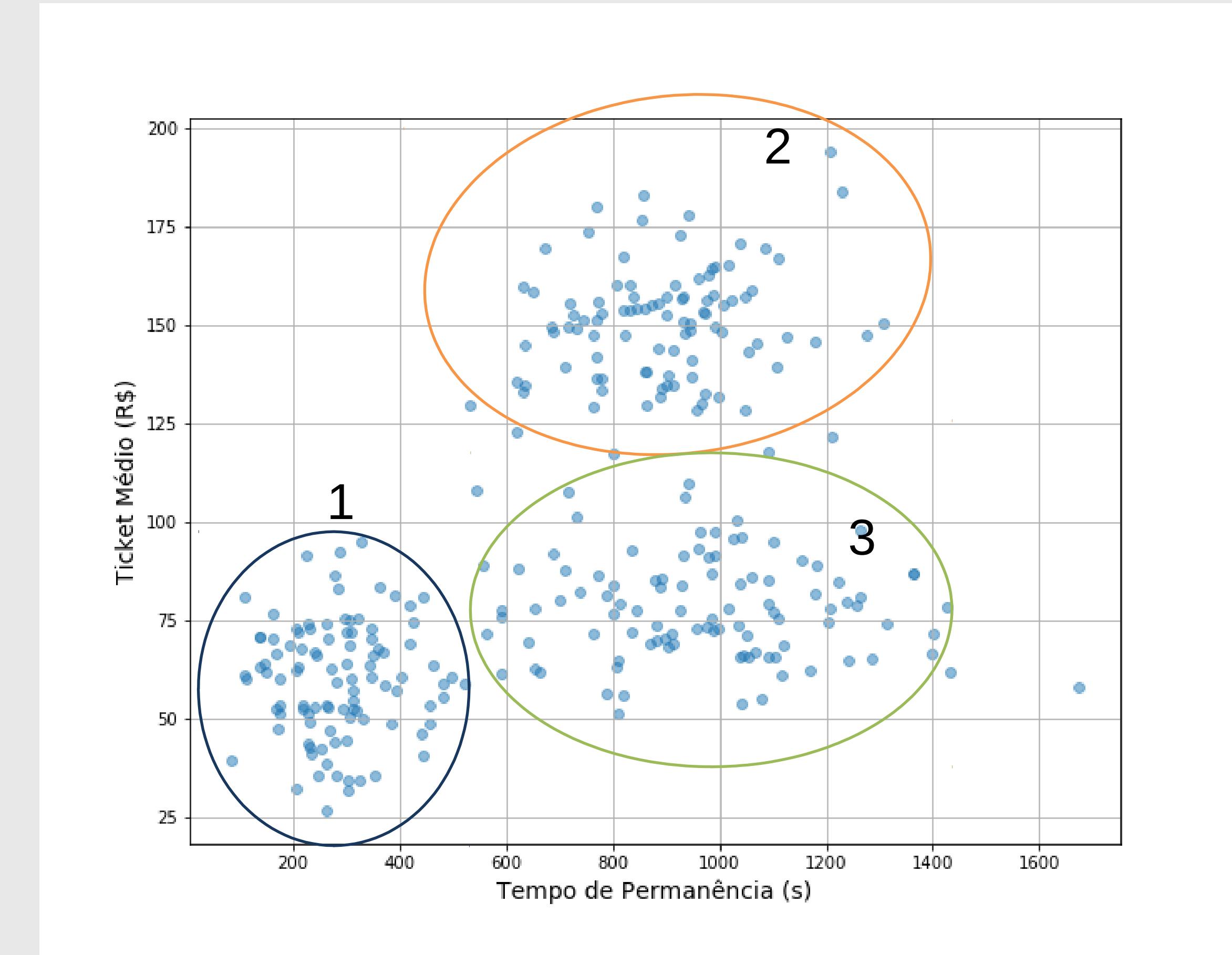
1: Compras específicas

(Ex: compras para bebês)

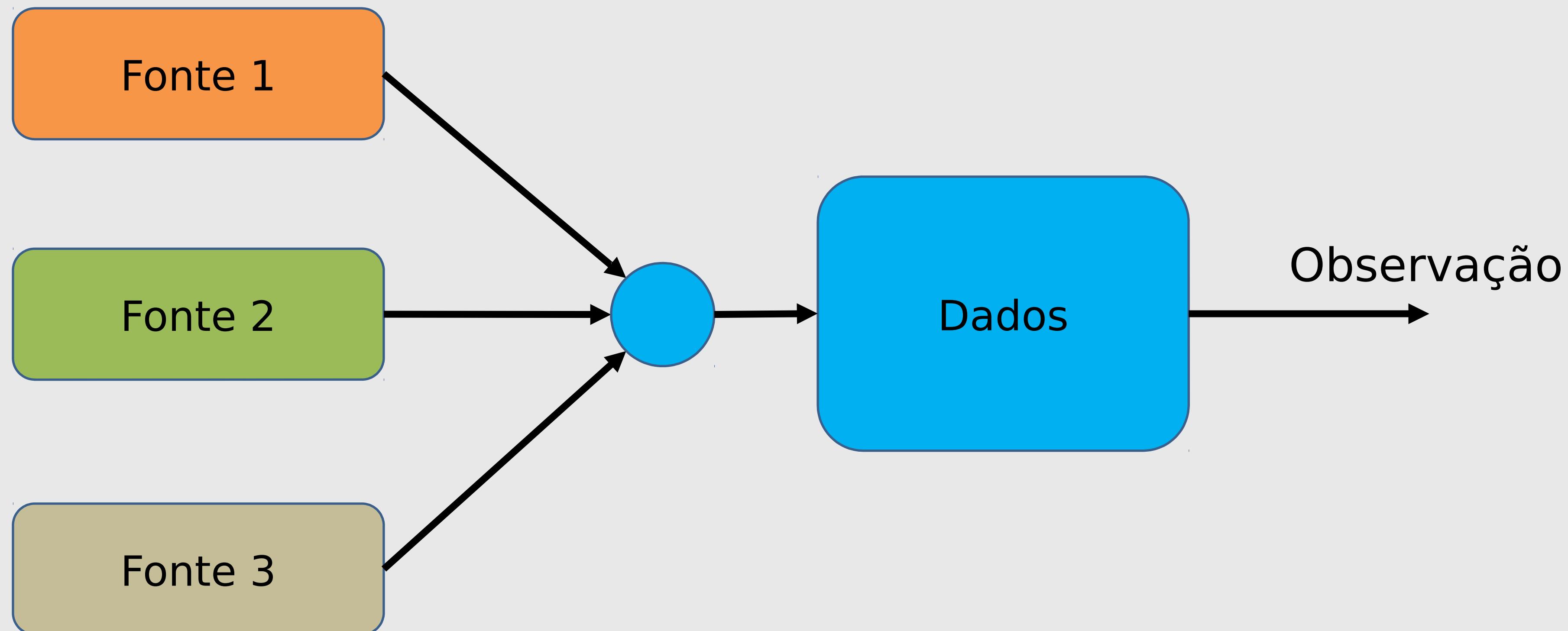
2: Compras para eventos

(Ex: preparação de casamento)

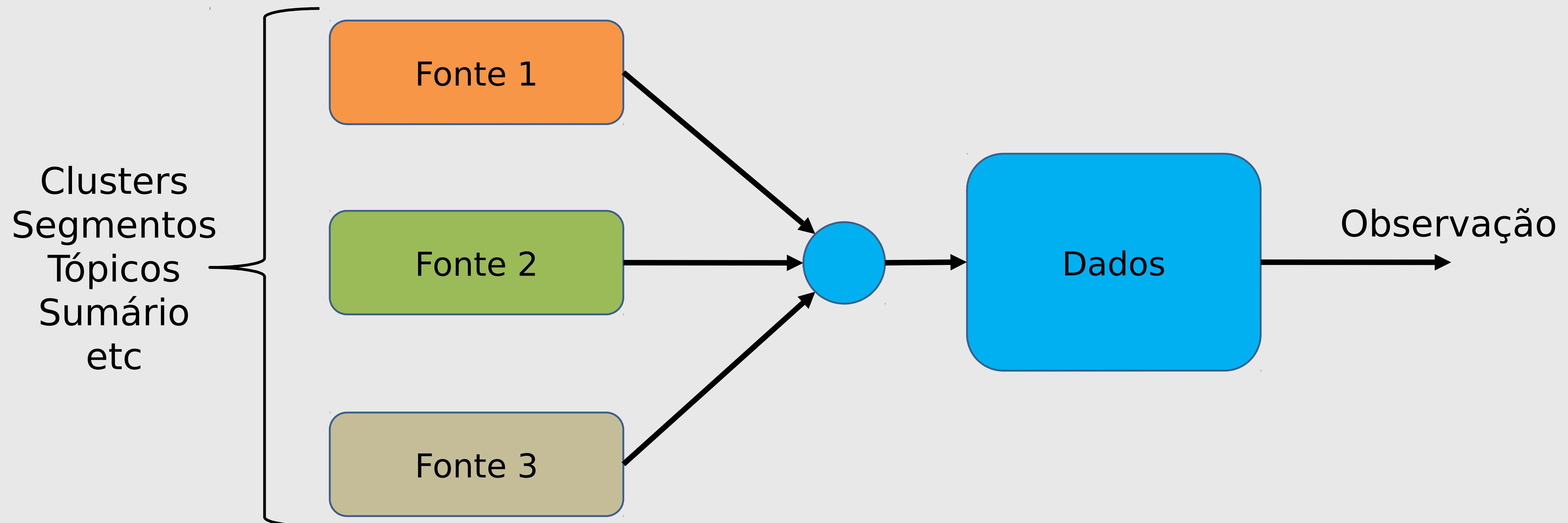
3: Navegação exploratória



Clustering

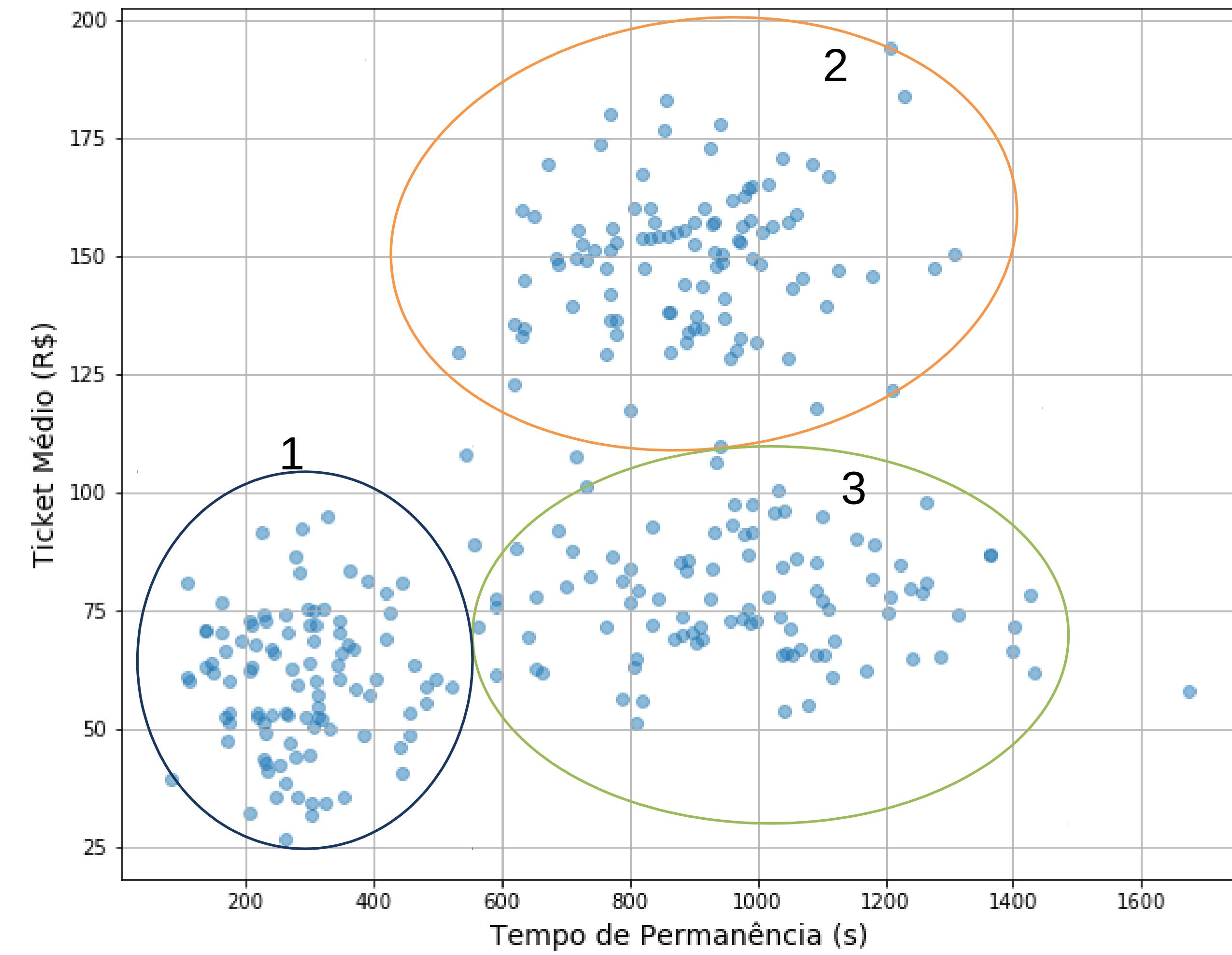


Clustering



Exemplo Elo7:

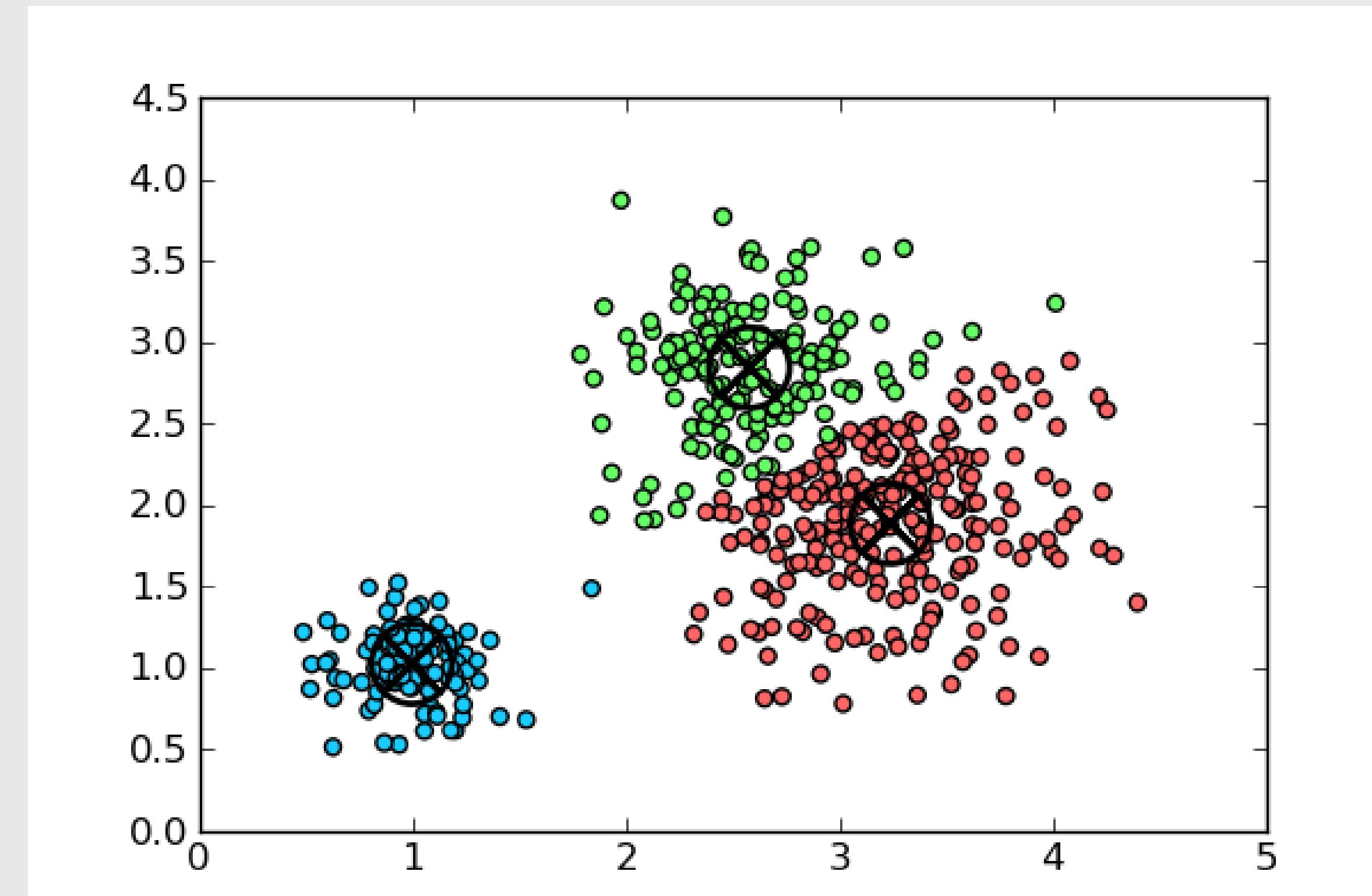
- **Clustering:**
- **Notebook**



T

2. Clustering: K-Means

- Gera clusters iterativamente
- Precisa escolher o número de clusters
- Agrupa dados pela distância do centroide



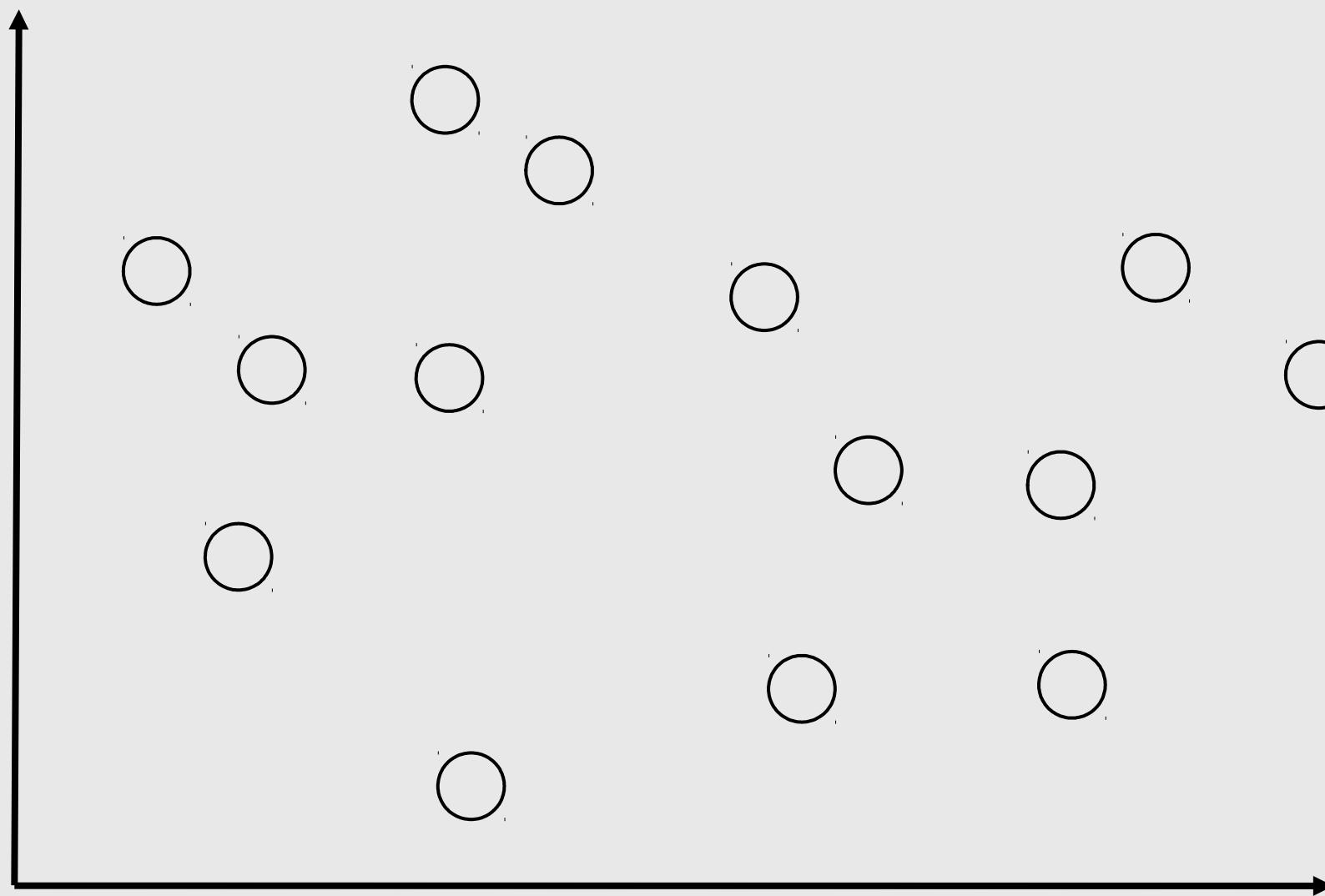
2. K-Means

- Algoritmo:
 - **Passo 1:** Defina o número de clusters
 - **Passo 2:** Escolha aleatoriamente a posição dos centroides
 - **Passo 3:** Para cada ponto, encontre o centroide mais próximo e atribua aquele cluster
 - **Passo 4:** Para cada cluster, calcule o novo centroide a partir dos pontos atribuídos anteriormente
 - **Passo 5:** Repita os passos 3-4 até não haver mais variação

I

2. K-Means

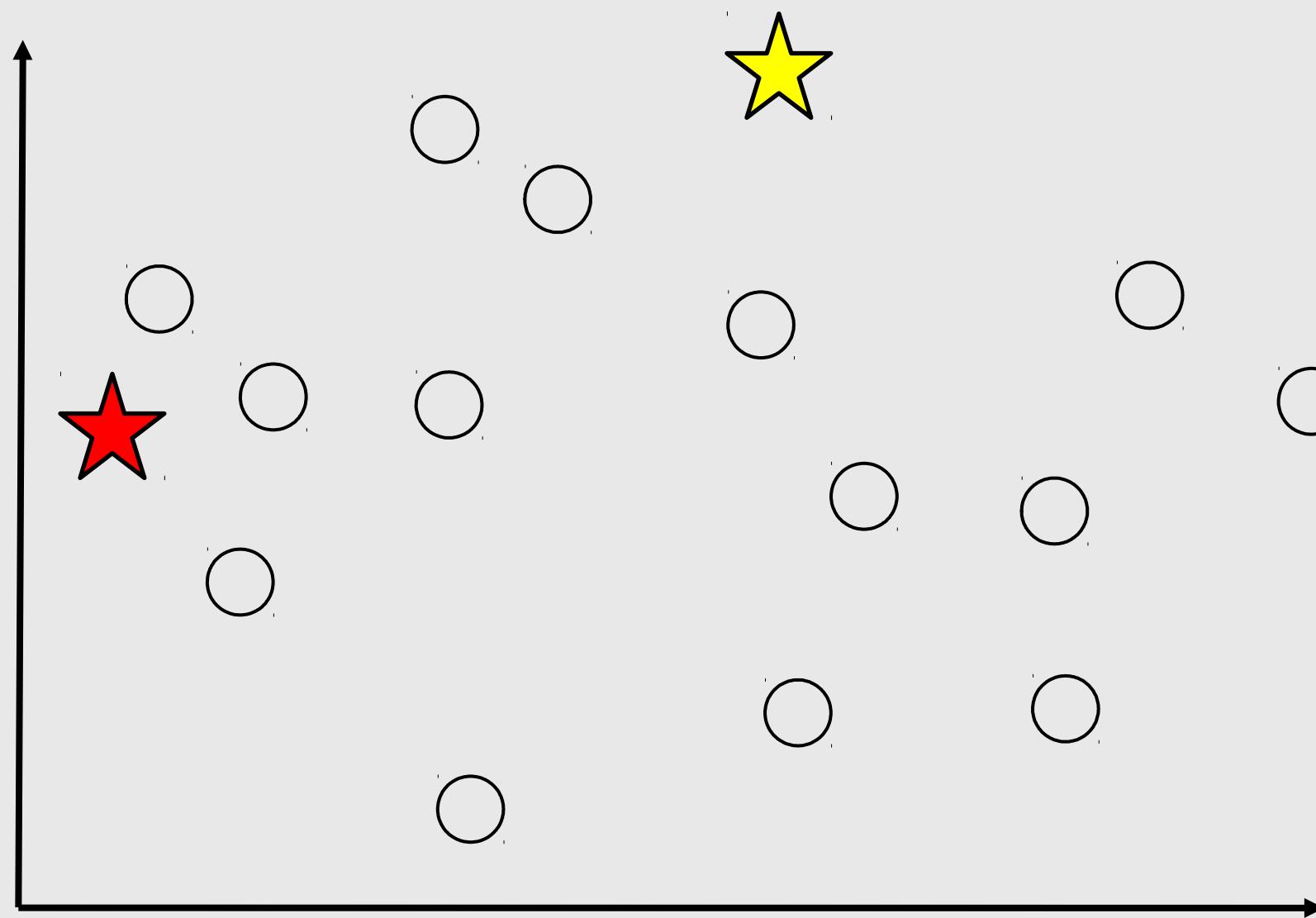
- **Passo 1:** Defina o número de clusters ($k = 2$)



I

2. K-Means

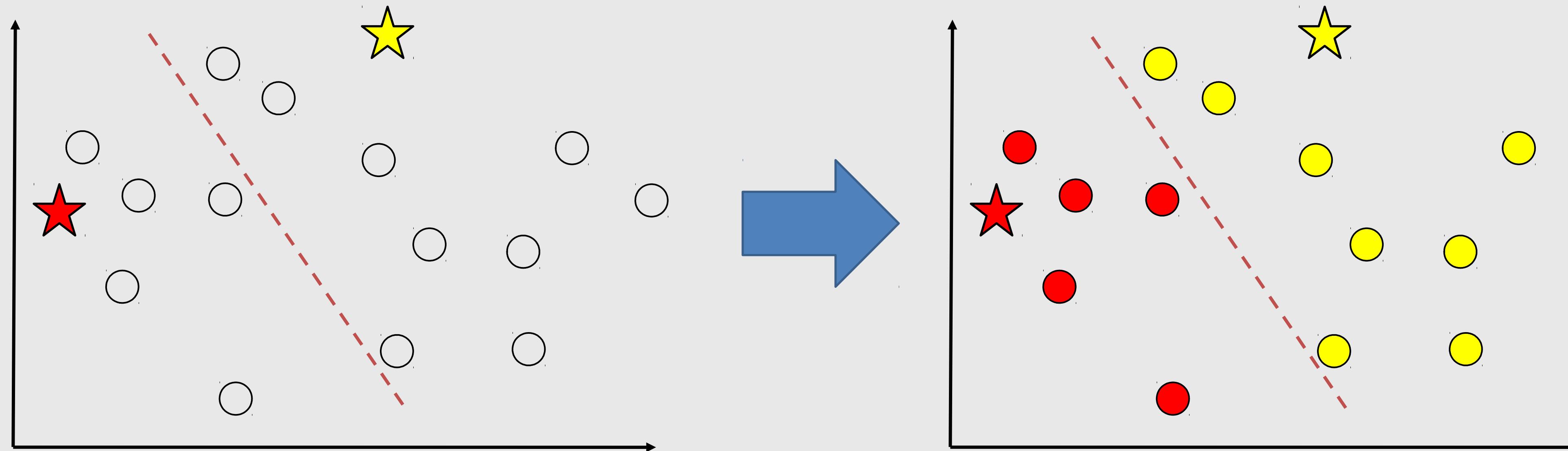
- **Passo 2:** Escolha aleatoriamente a posição dos centroides



I

2. K-Means

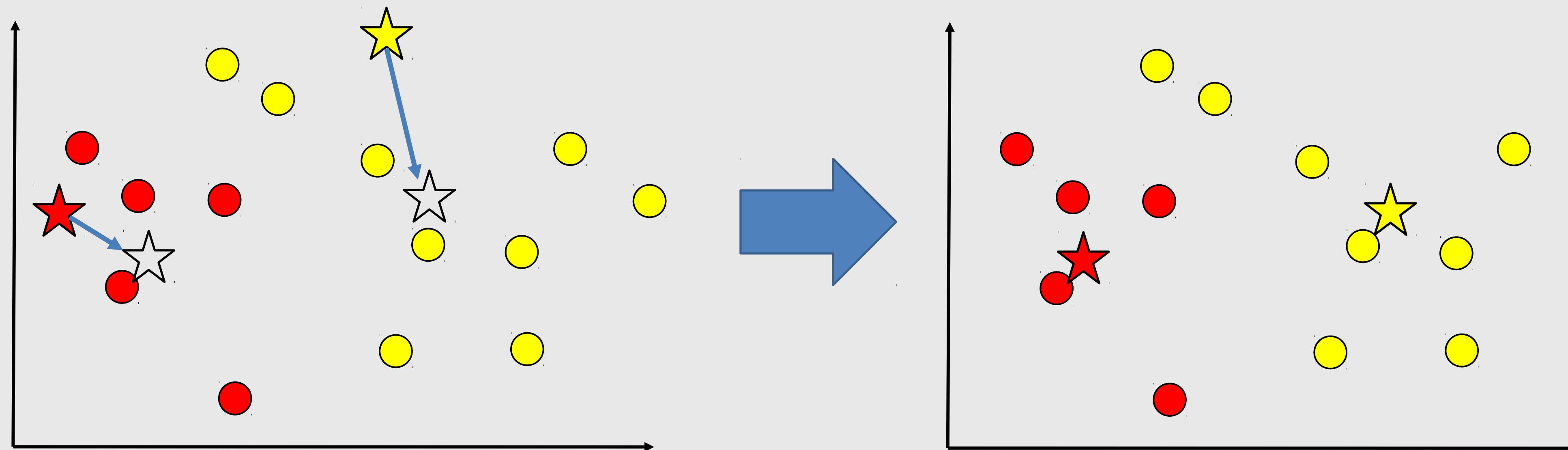
- **Passo 3:** Para cada ponto, encontre o centroide mais próximo e atribua àquele cluster



I

2. K-Means

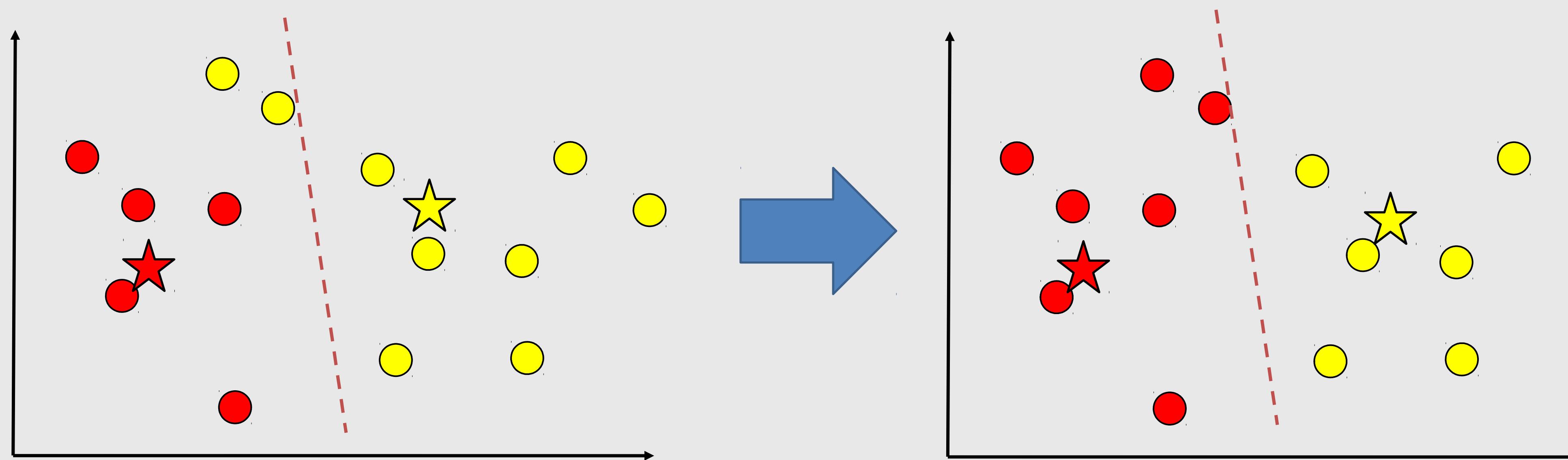
- **Passo 4:** Para cada cluster, calcule o novo centroide a partir dos pontos atribuídos anteriormente



I

2. K-Means

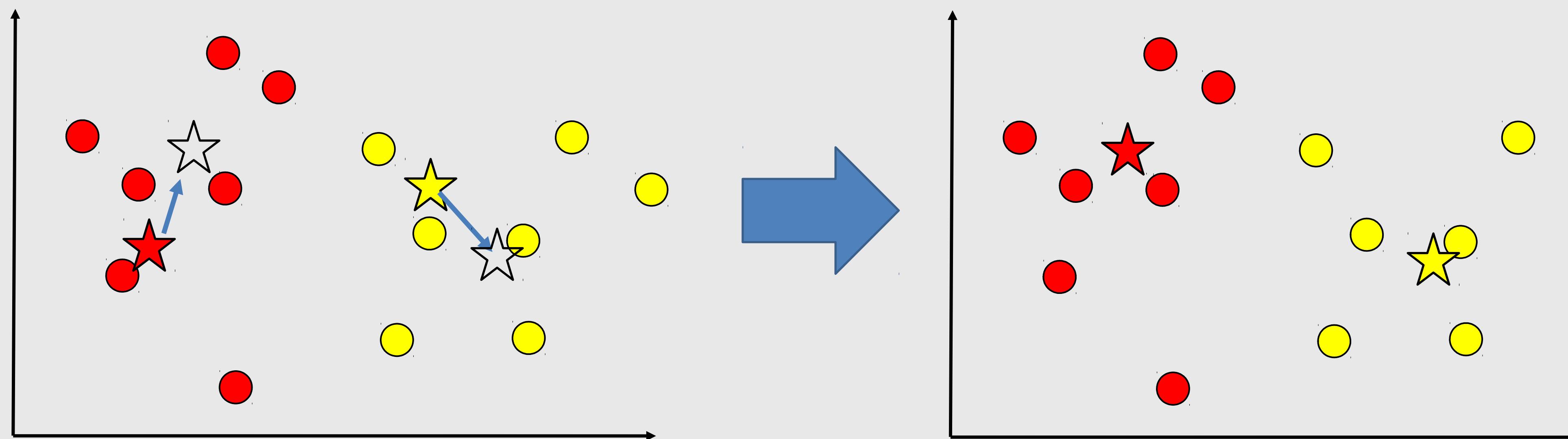
- **Passo 3:** Para cada ponto, encontre o centroide mais próximo e atribua aquele cluster



I

2. K-Means

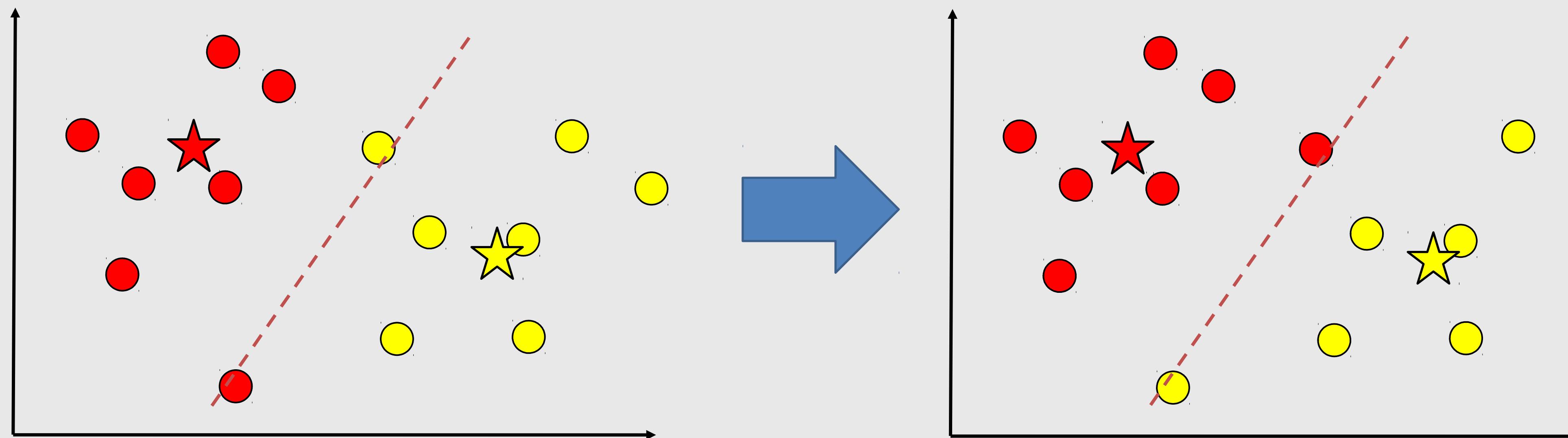
- **Passo 4:** Para cada cluster, calcule o novo centroide a partir dos pontos atribuídos anteriormente



I

2. K-Means

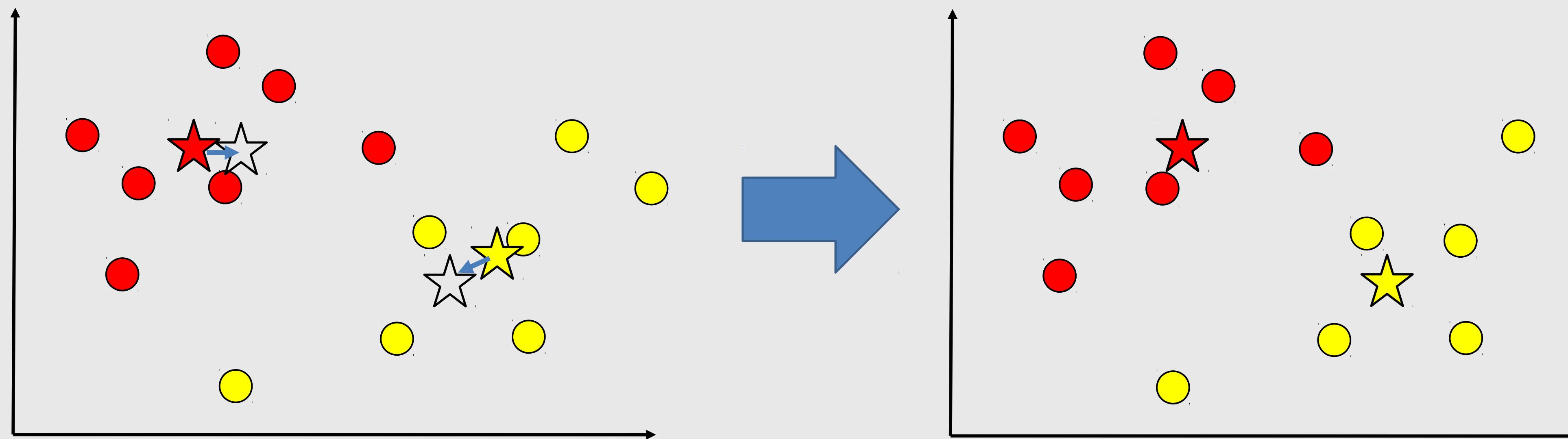
- **Passo 3:** Para cada ponto, encontre o centroide mais próximo e atribua aquele cluster



I

2. K-Means

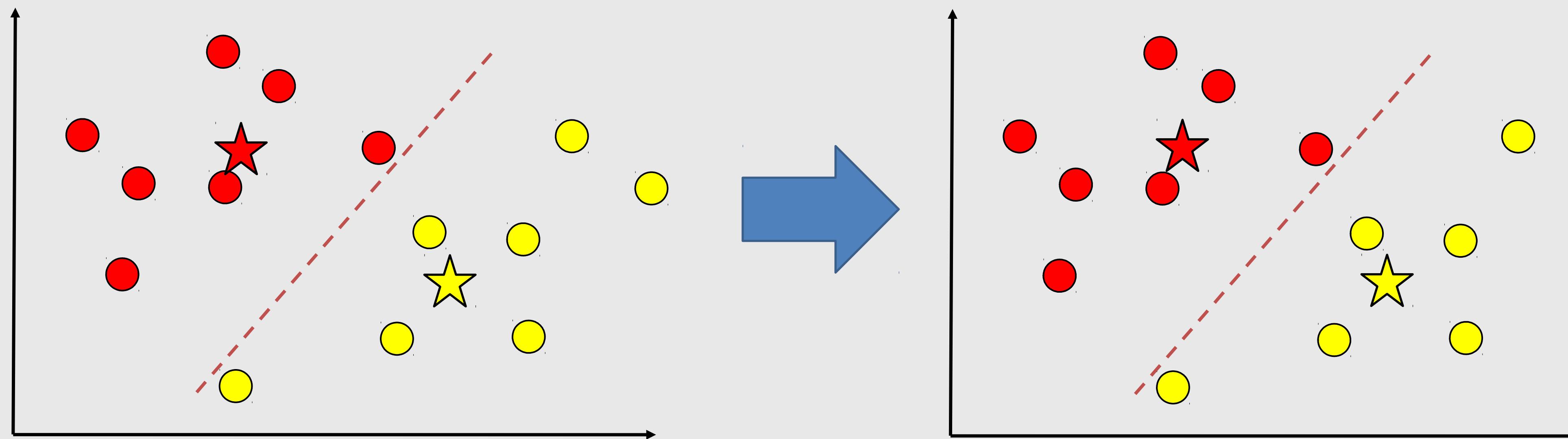
- **Passo 4:** Para cada cluster, calcule o novo centroide a partir dos pontos atribuídos anteriormente



I

2. K-Means

- **Passo 3:** Para cada ponto, encontre o centroide mais próximo e atribua aquele cluster



- **Finalizado! Não há mais variação**

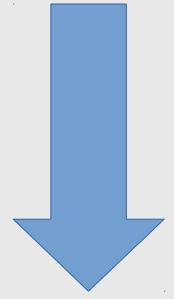
T

2. K-Means

- **Exemplo:**

2. K-Means

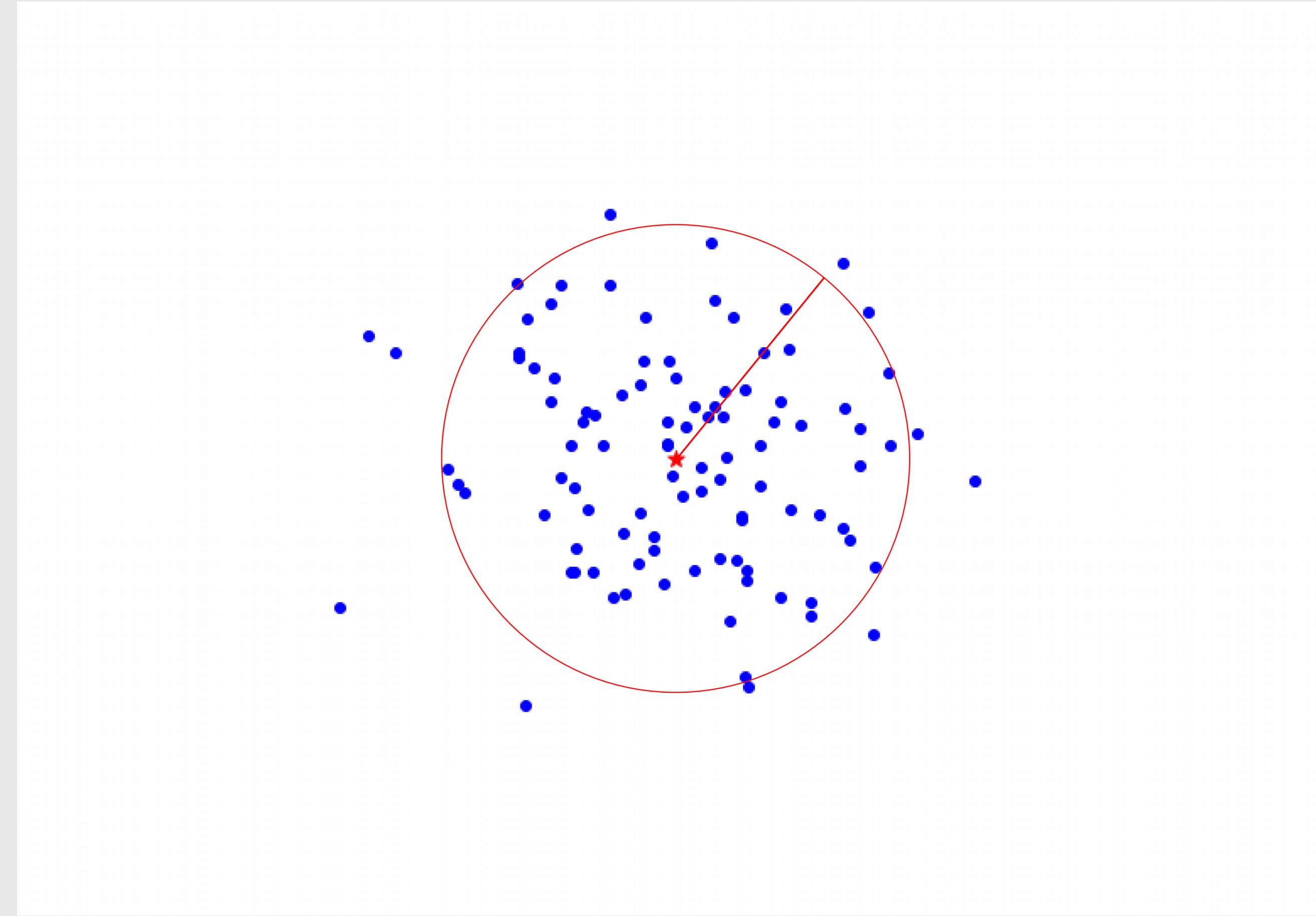
- **Avaliar um bom cluster:**
 - Dados dentro do cluster possuem perfil semelhante
 - Dados bem distribuídos nos clusters (balanceados)
 - Não muito disperso



Inércia

2. K-Means

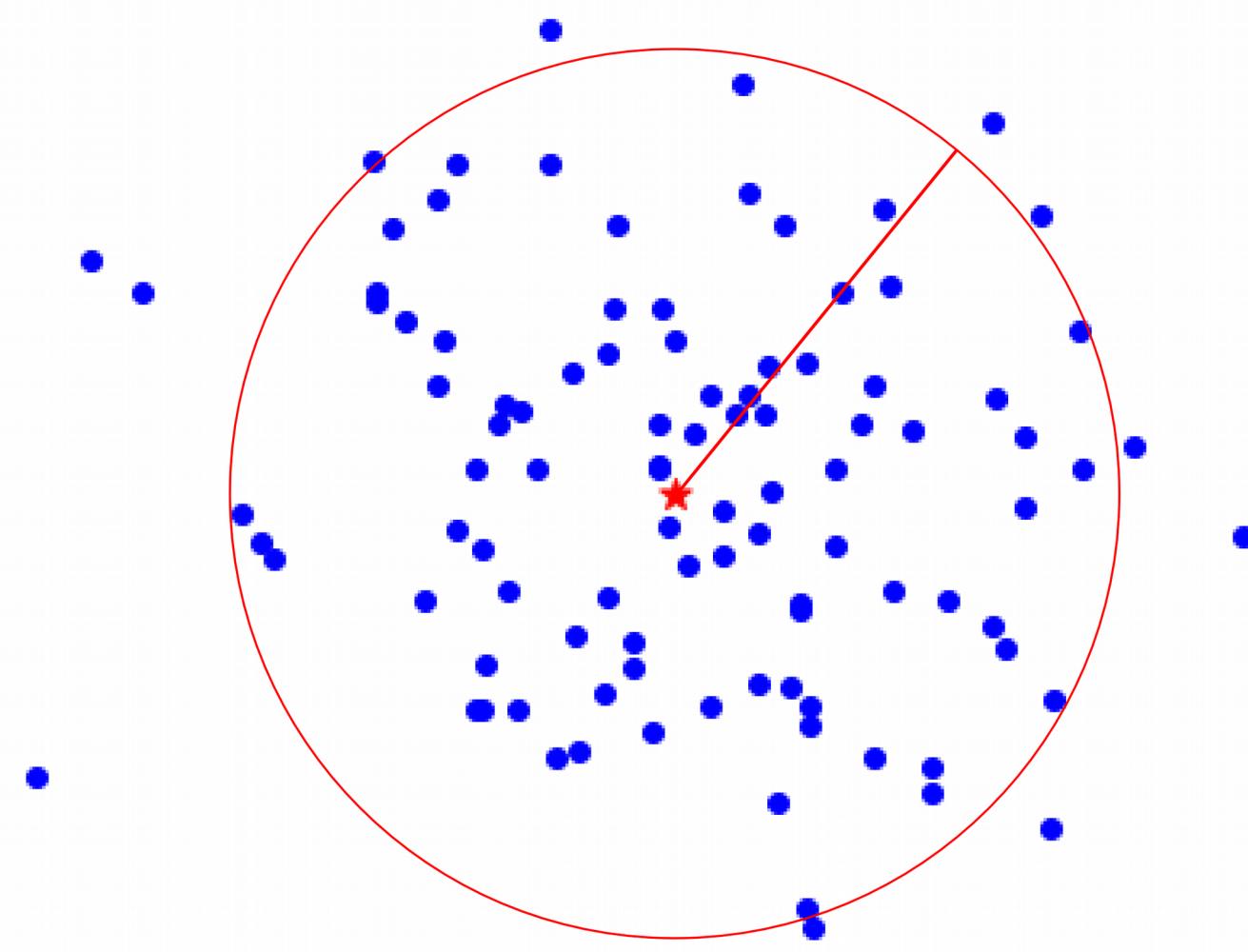
- **Inércia:**
 - Nível de dispersão dos pontos em relação ao centroide



I

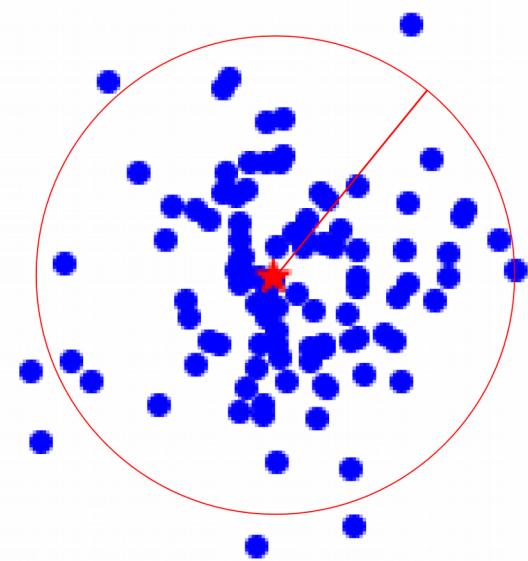
2. K-Means

- Inércia:
 - Elevada



2. K-Means

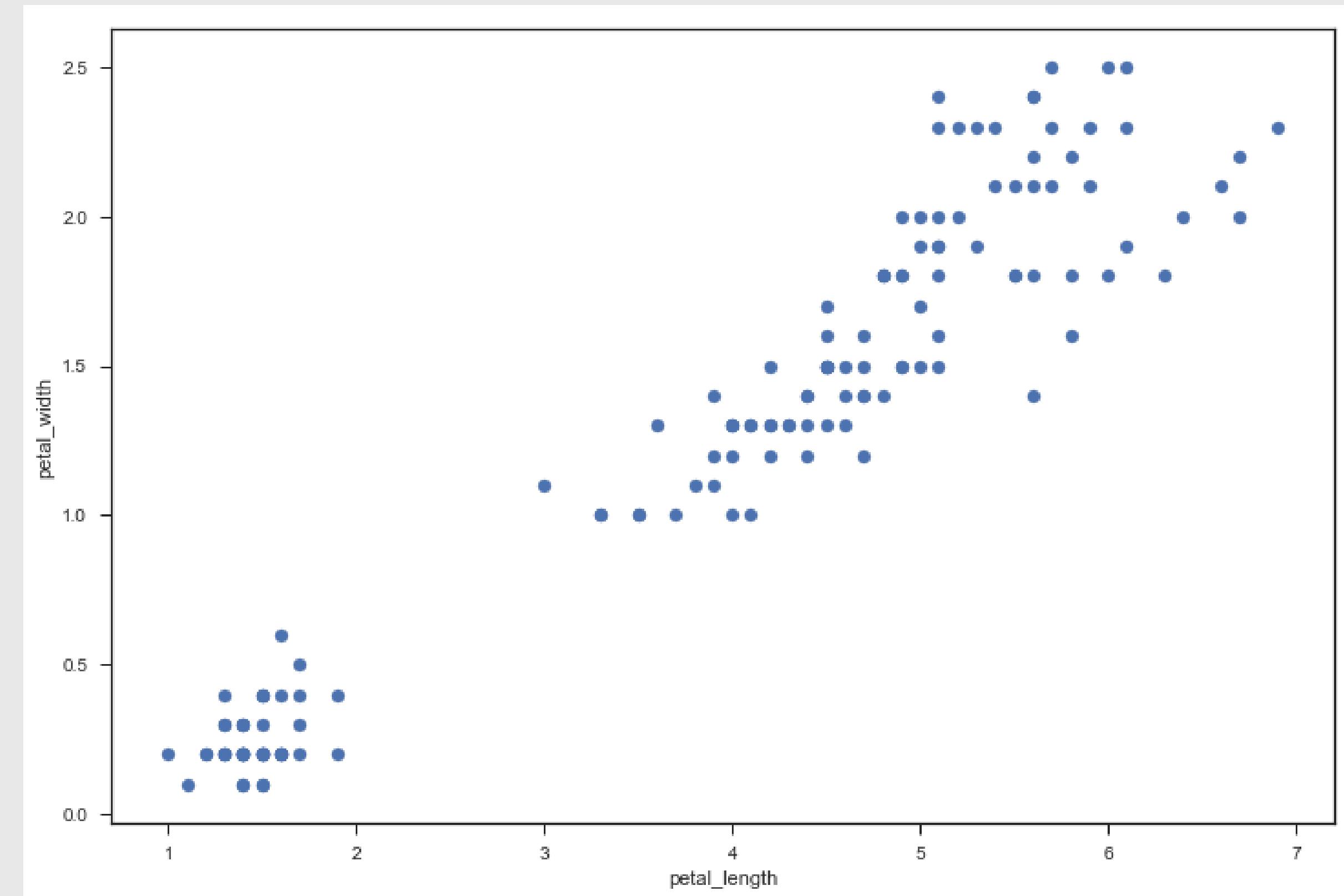
- Inércia:
 - Baixa



T

2. K-Means: Inércia

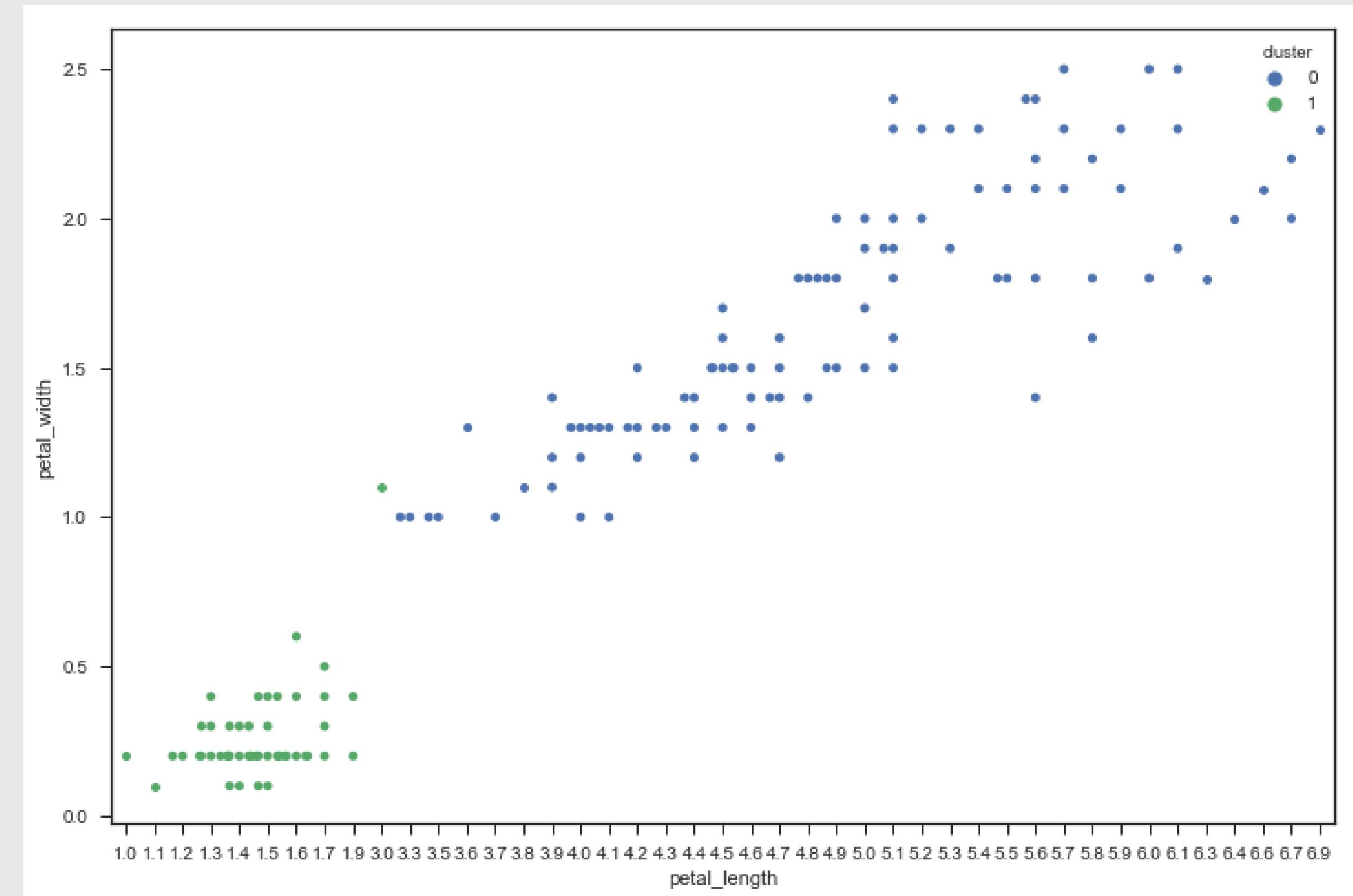
- Tem relação com o número de clusters
- Ex: k=1
- Inércia alta



T

2. K-Means: Inércia

- Tem relação com o número de clusters
- Ex: k=2
- Inércia média

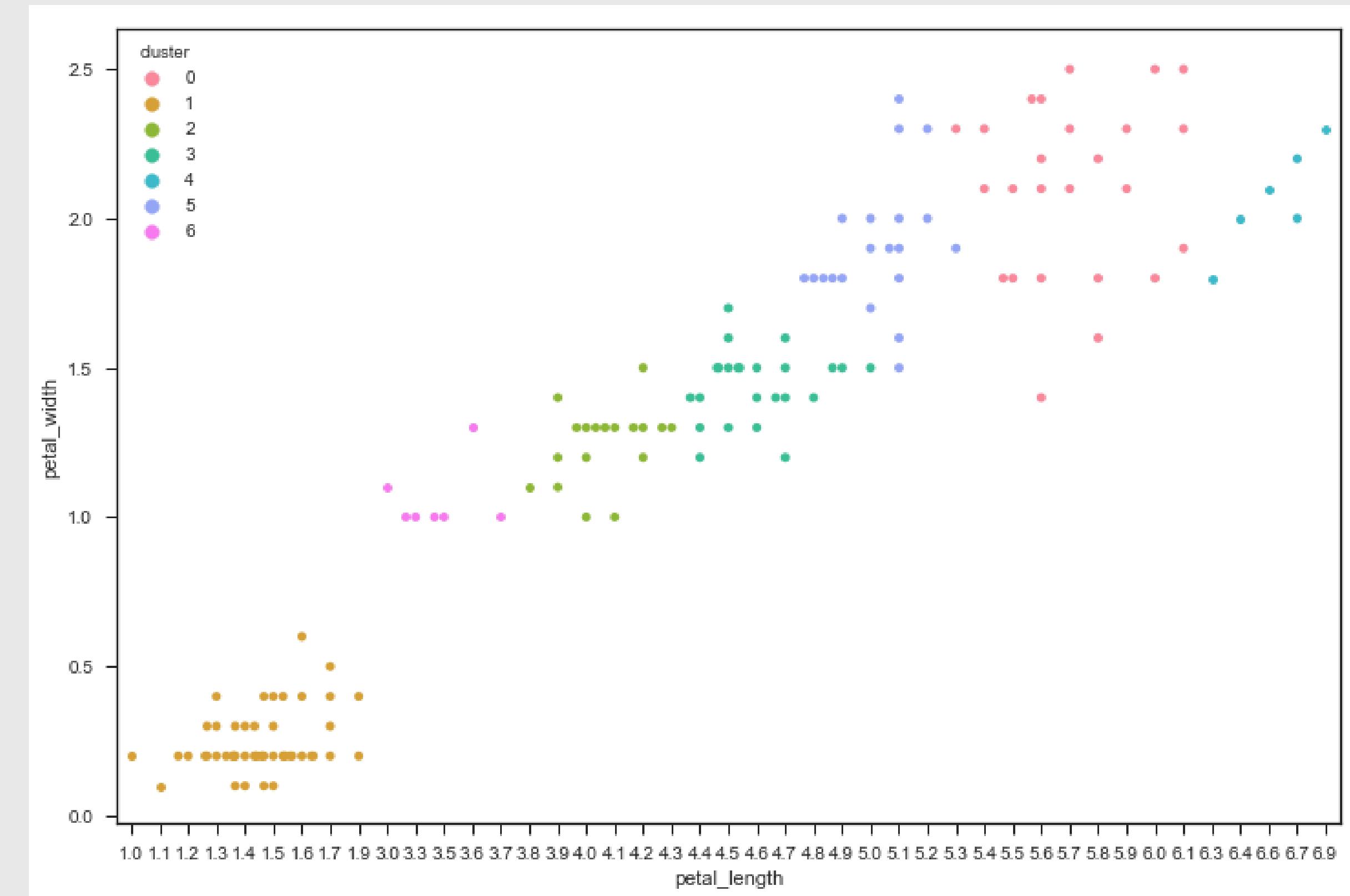


Inércia = 86

T

2. K-Means: Inércia

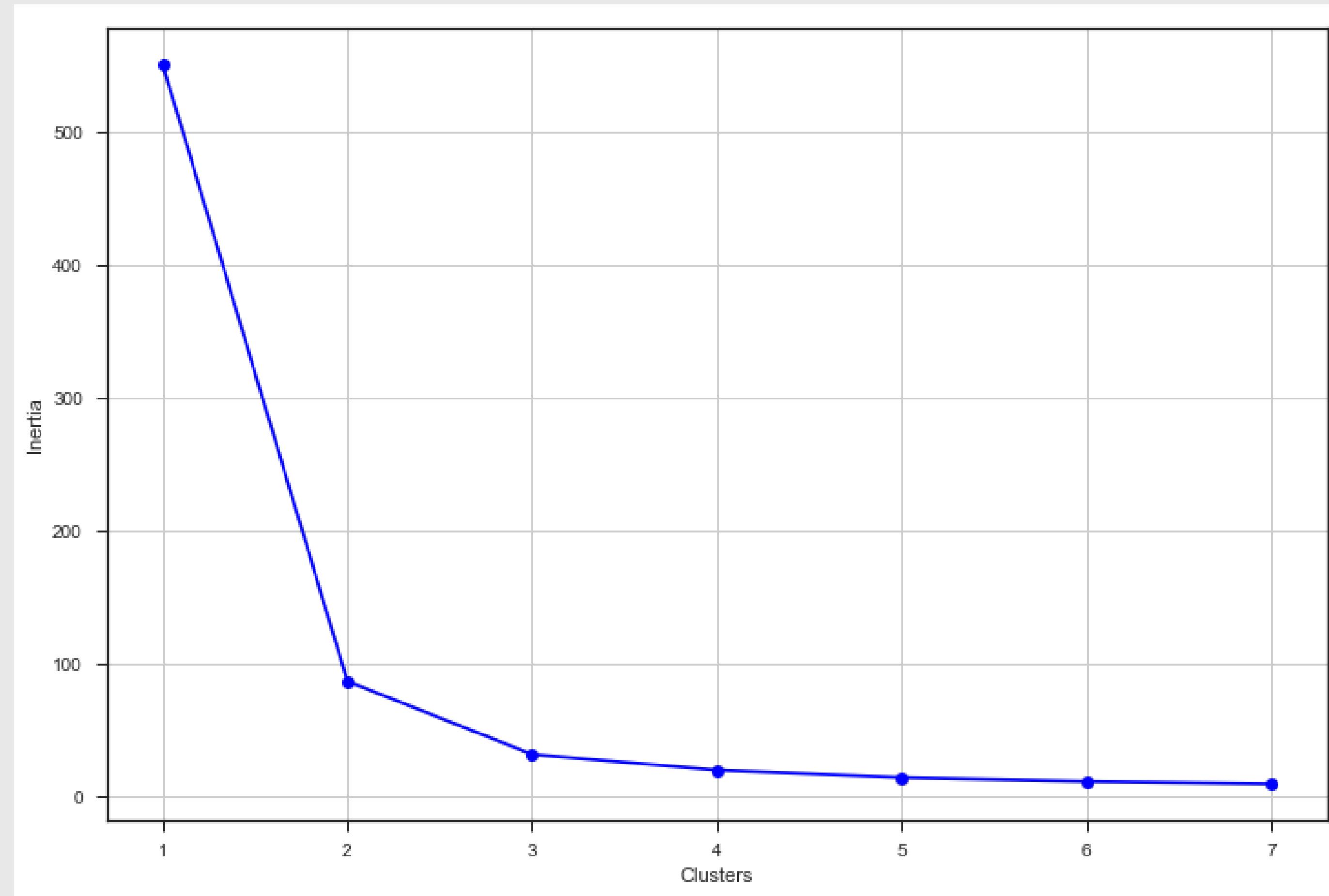
- Tem relação com o número de clusters
- Ex: k=7
- Inércia baixa



Inércia = 9

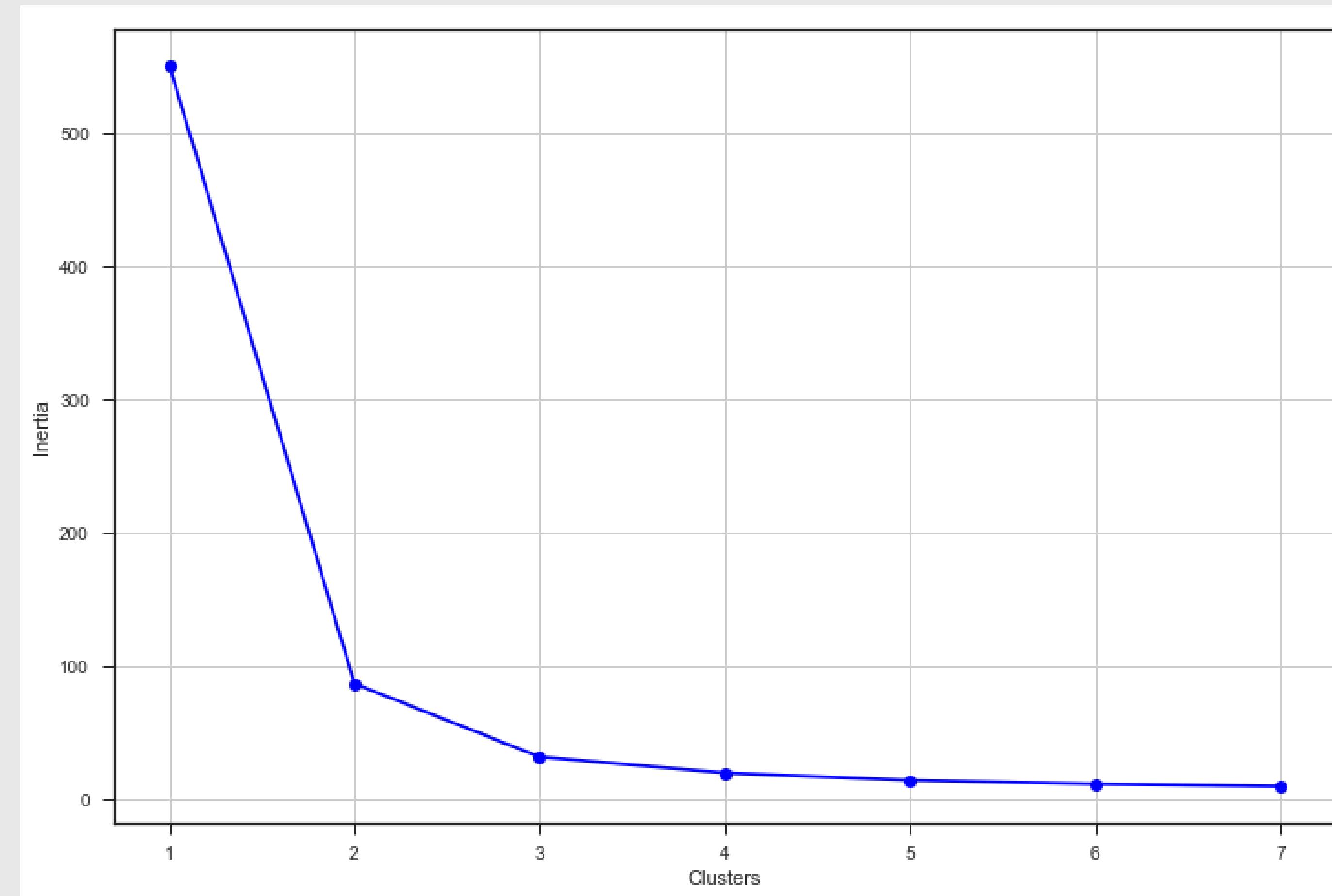
2. K-Means: Inércia

- Inércia em função do número de clusters



2. K-Means: Inércia

- Número ideal:
 - Primeiro ponto de menor angulação - 3

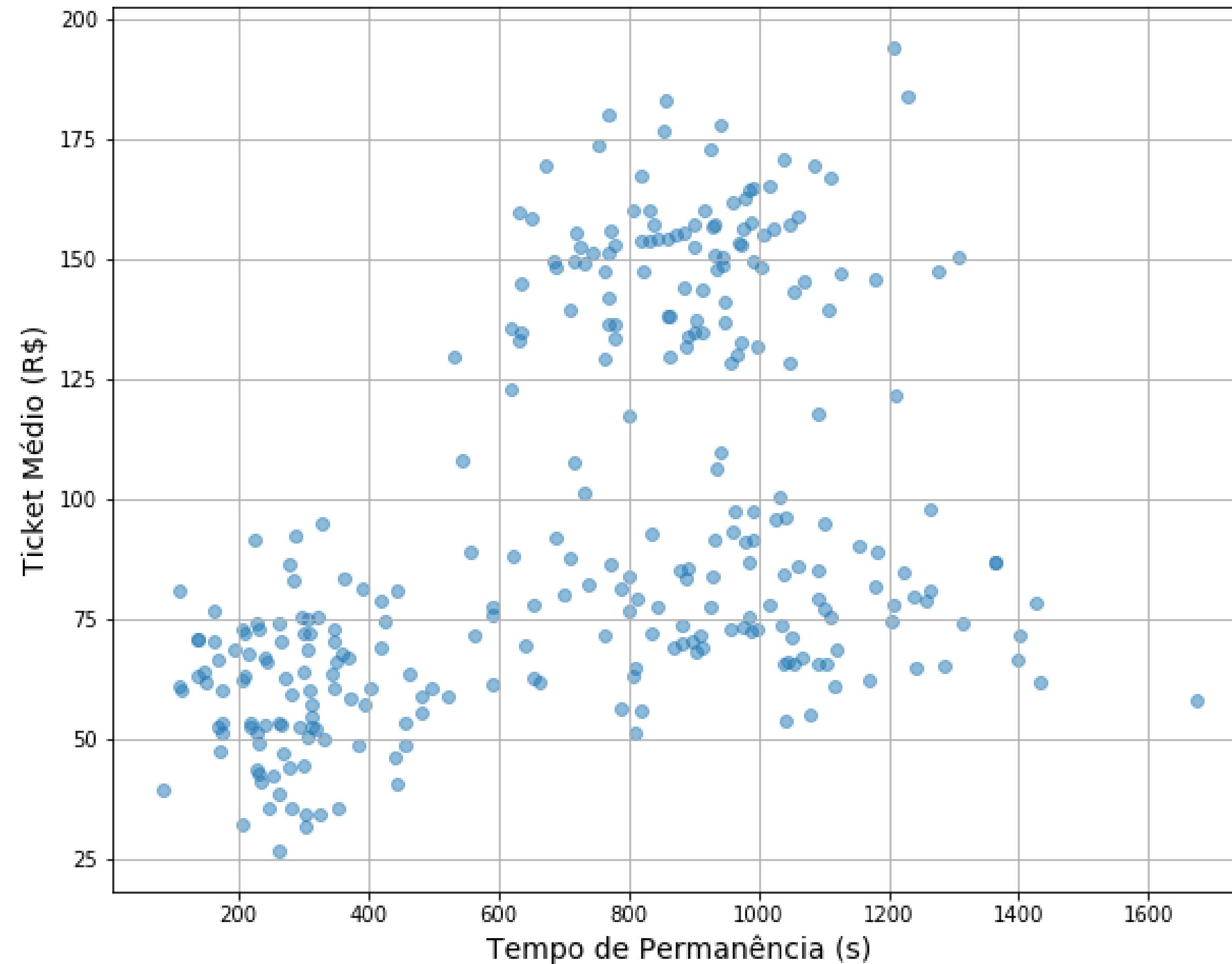


I

2. K-Means

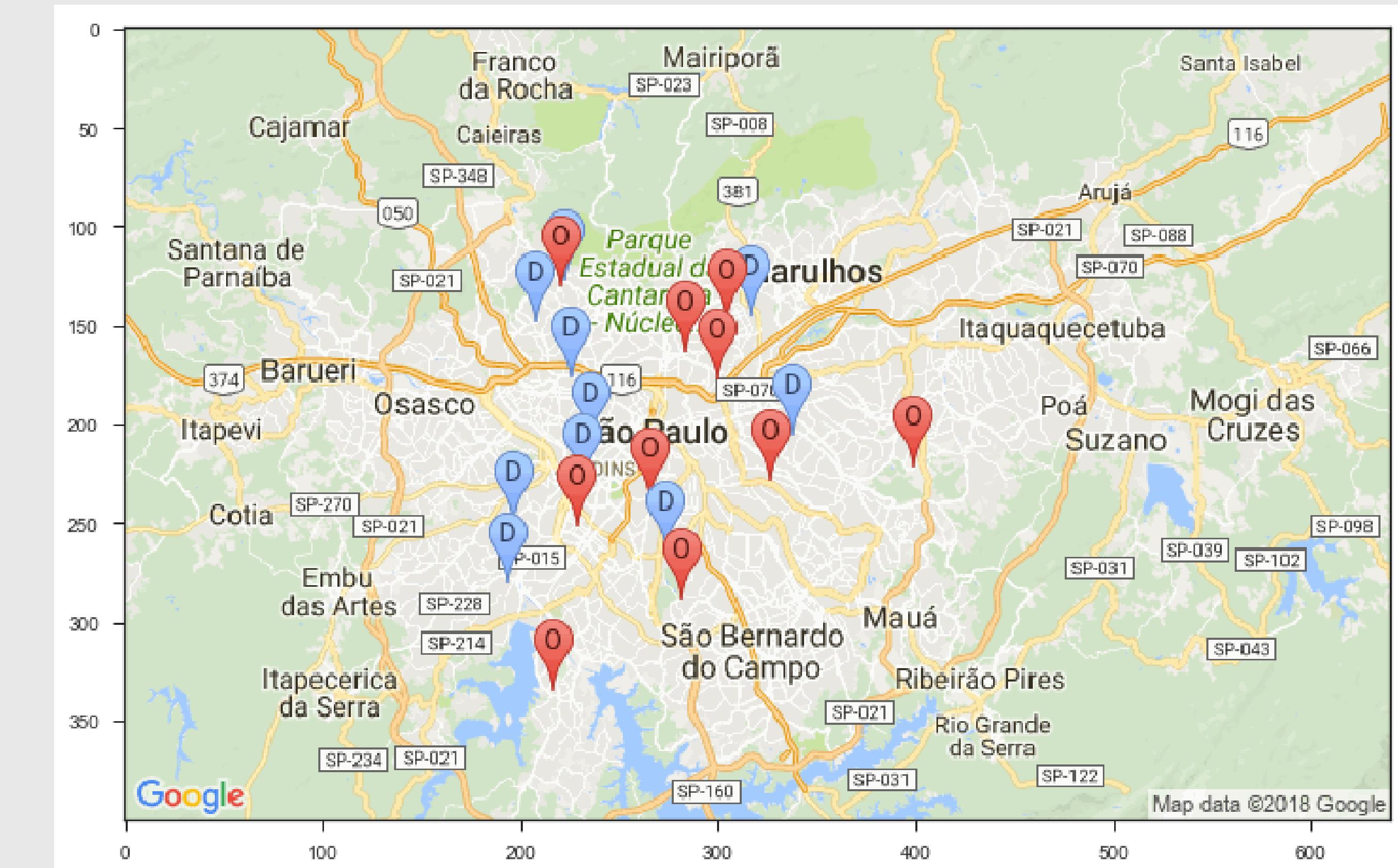
- **Case Elo7**

Notebook



Exemplo Elo7:

- Encontrar clusters de rota de frete



T

2. K-Means

- **Case**
Notebook



2. K-Means

- **Vantagens K-Means**
 - Algoritmo simples
 - Resultado intuitivo
 - Funciona bem na prática
 - Pode ser utilizado para conjunto grande de dados

2. K-Means

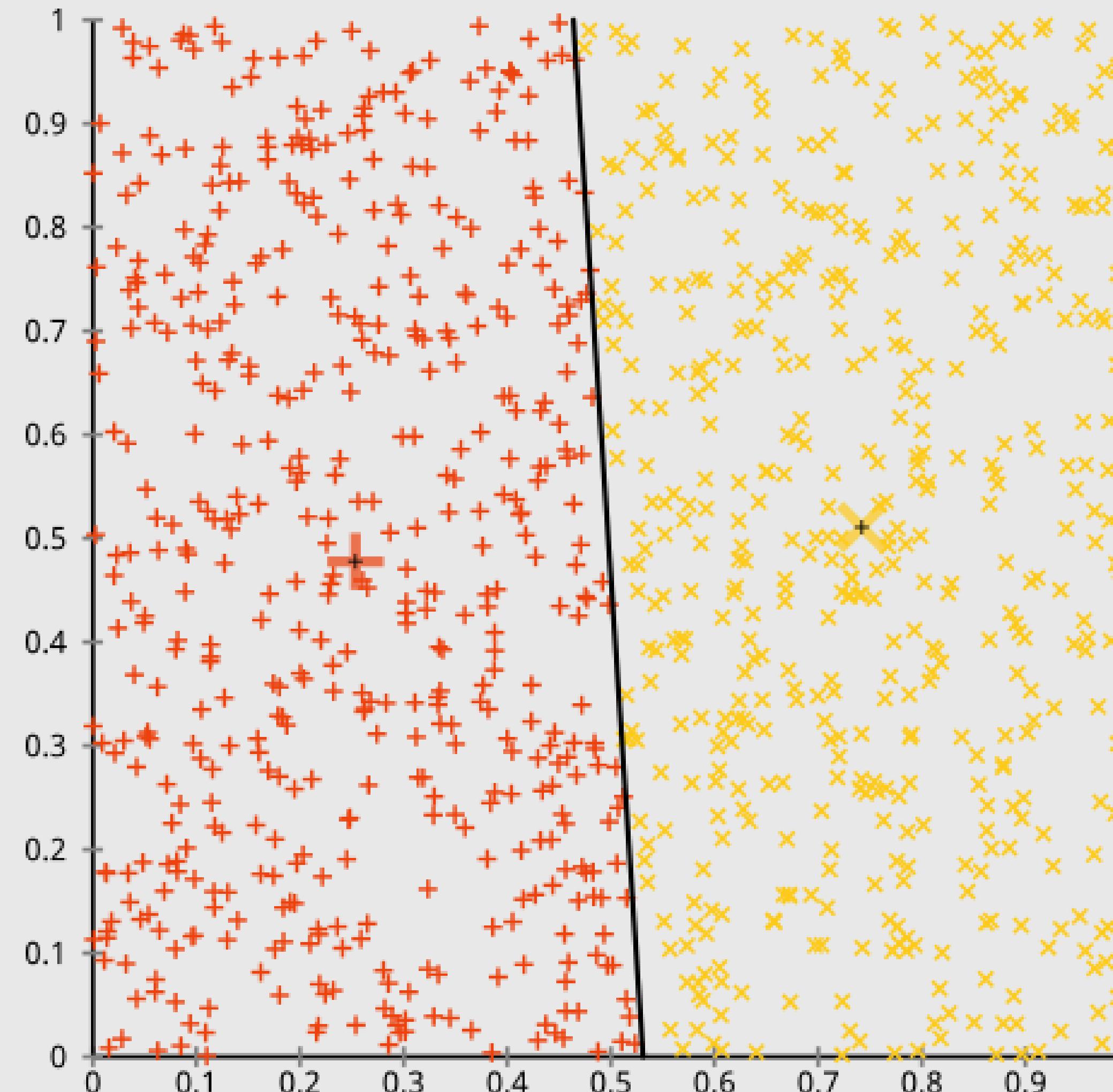
- **Desvantagens K-Means**

- Necessita escolher o número de clusters
- Pode convergir para resultados indesejados
- Resultados podem divergir a cada repetição (inicialização aleatória)
- Algoritmo “cego”: Encontra clusters até em locais onde não há
- Não há relação de hierarquia entre clusters

I

2. K-Means

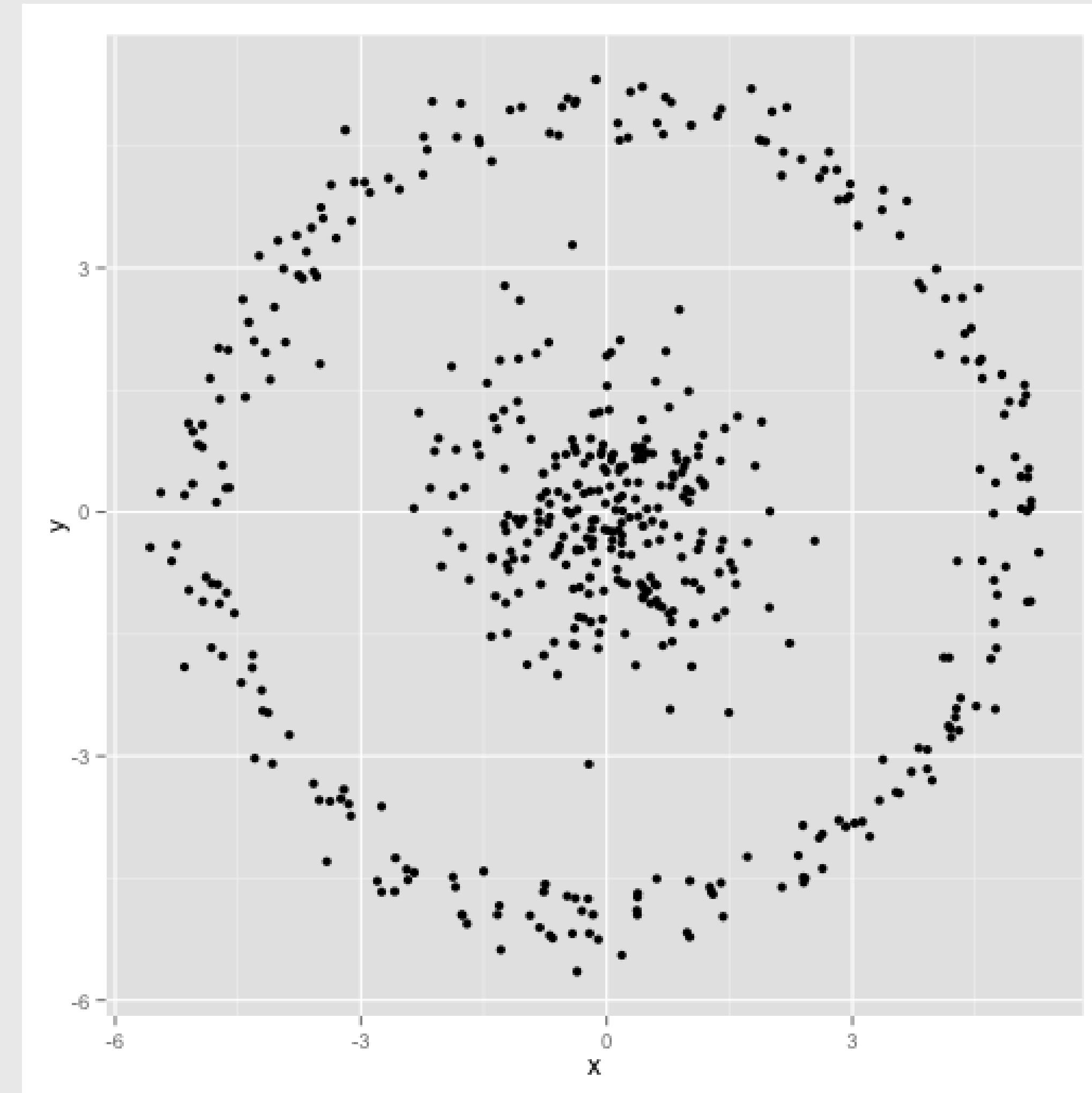
- Desvantagens K-Means



I

2. K-Means

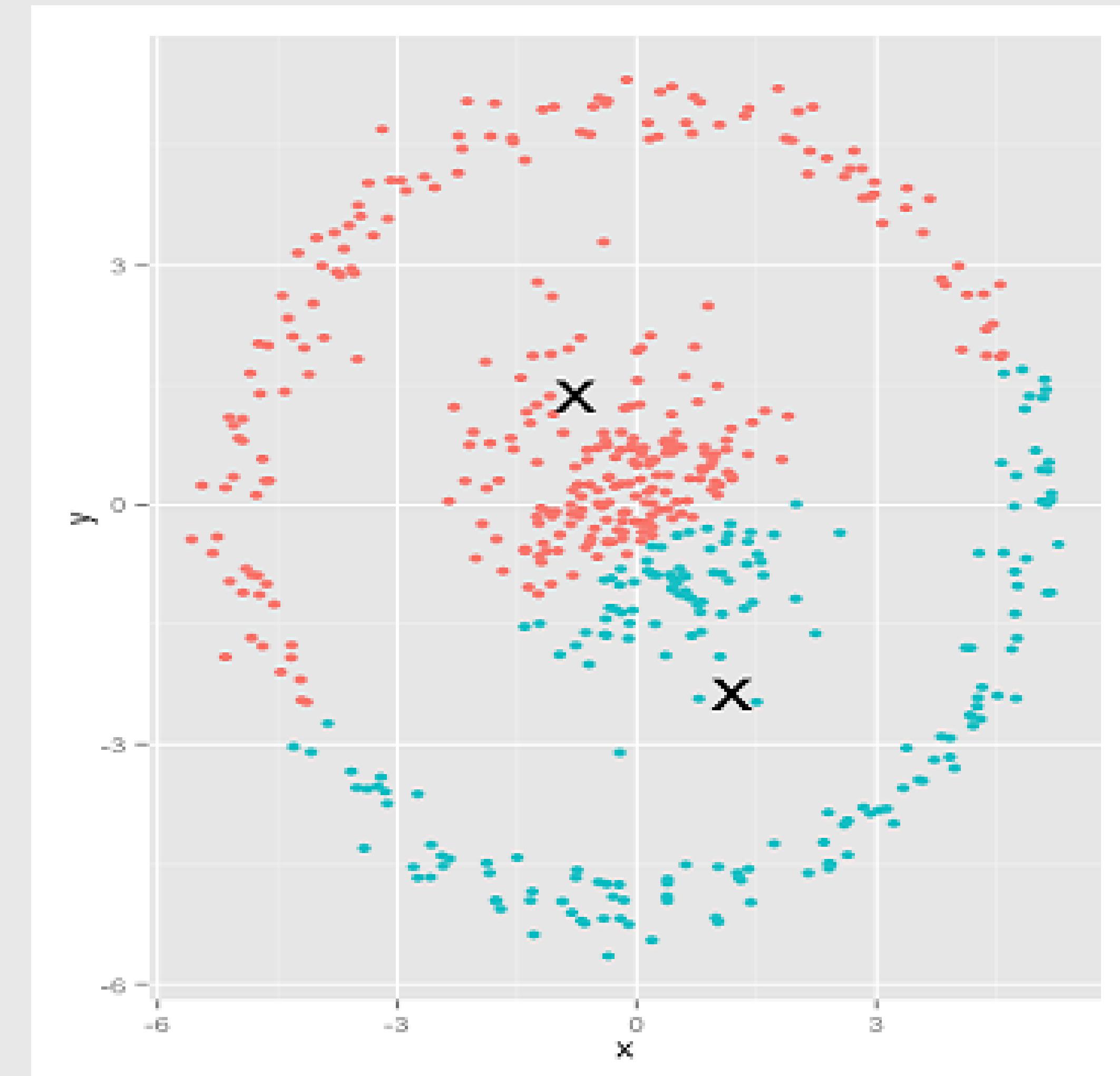
- Desvantagens K-Means



I

2. K-Means

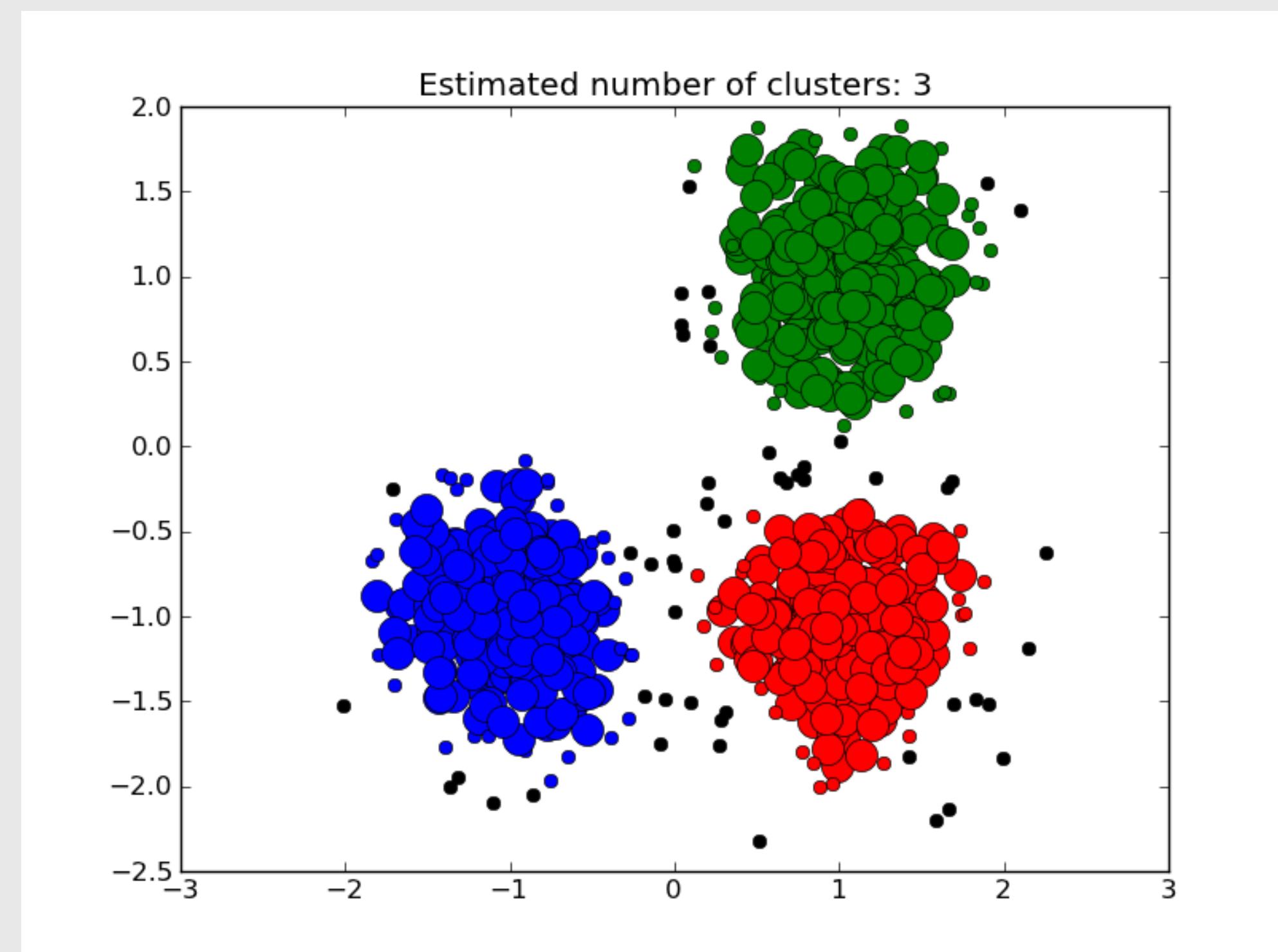
- Desvantagens K-Means



T

4. Clustering: DBSCAN

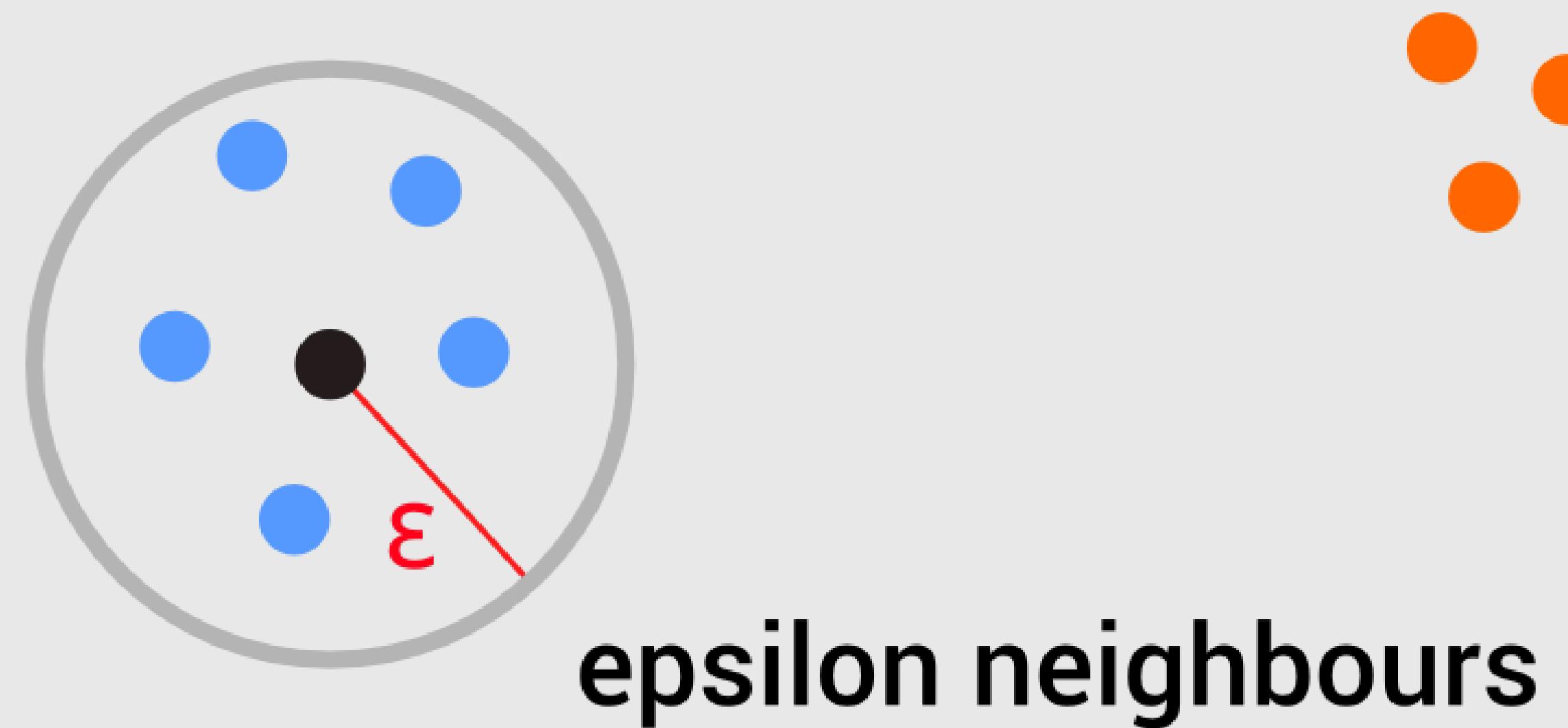
- **Density-Based Spatial Clustering Algorithm with Noise**
 - Utiliza o conceito de conectividade por densidade



I

4. Clustering: DBSCAN

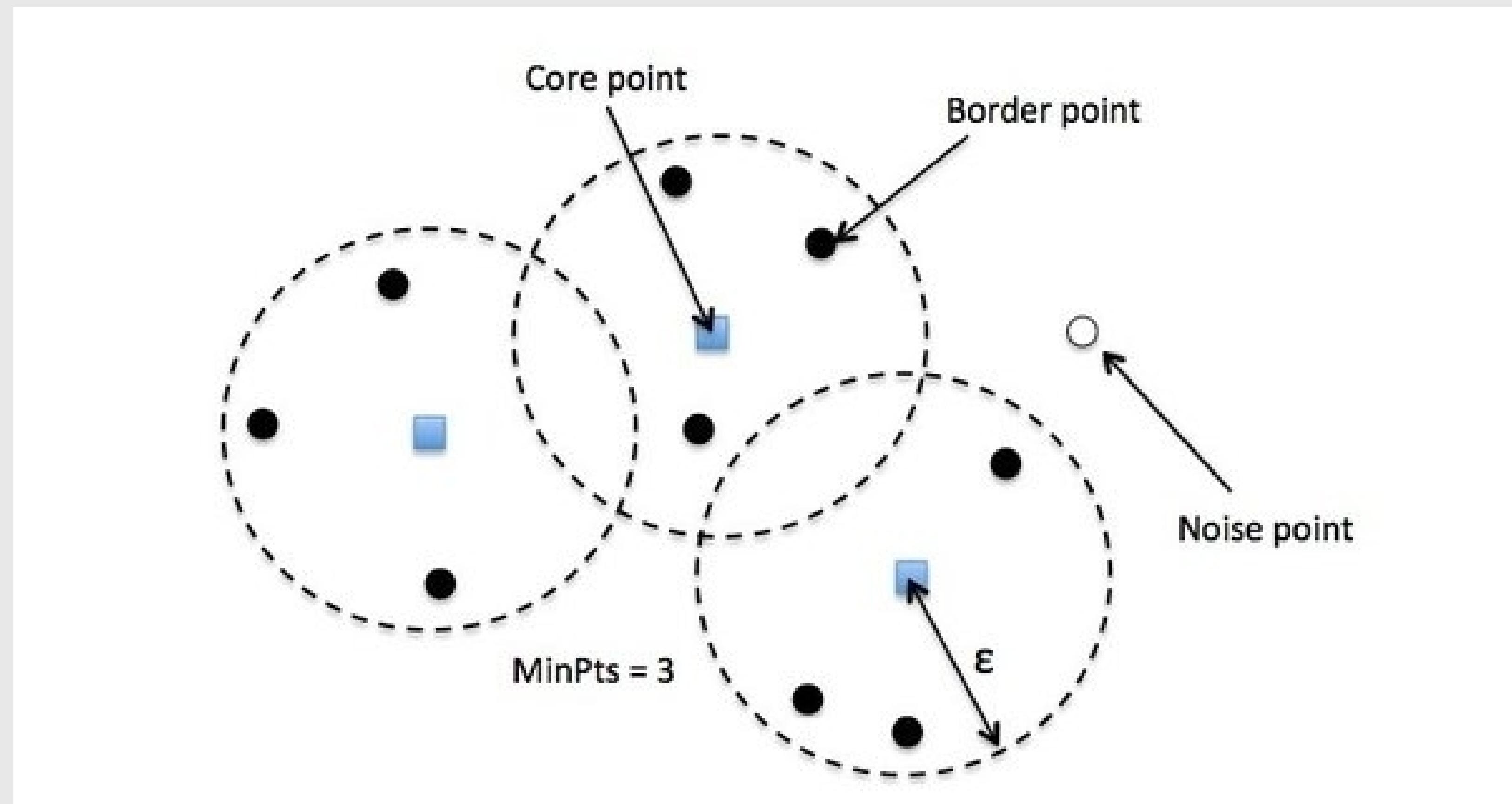
- Precisa escolher dois parâmetros: epsilon, min_points



I

4. Clustering: DBSCAN

- Conecta os pontos a partir de regiões de densidade



T

4. Clustering: DBSCAN

- Exemplos:
 - Notebook

4. Clustering: DBSCAN

- **Vantagens:**

- Não precisa escolher o número de clusters
- Consegue eliminar possíveis outliers
- Escala facilmente para datasets grandes

4. Clustering: DBSCAN

- **Desvantagens:**
 - A escolha dos parâmetros é crucial e depende de ajustes finos
 - Sofre com clusters de diferentes densidades

4. Clustering: Hierarchical Clustering

- Método iterativo
- Se baseia no princípio da conectividade das observações
- Depende da definição de distância / similaridade

I

4. Clustering: Hierarchical Clustering

- Como agrupar?



Camiseta Batman



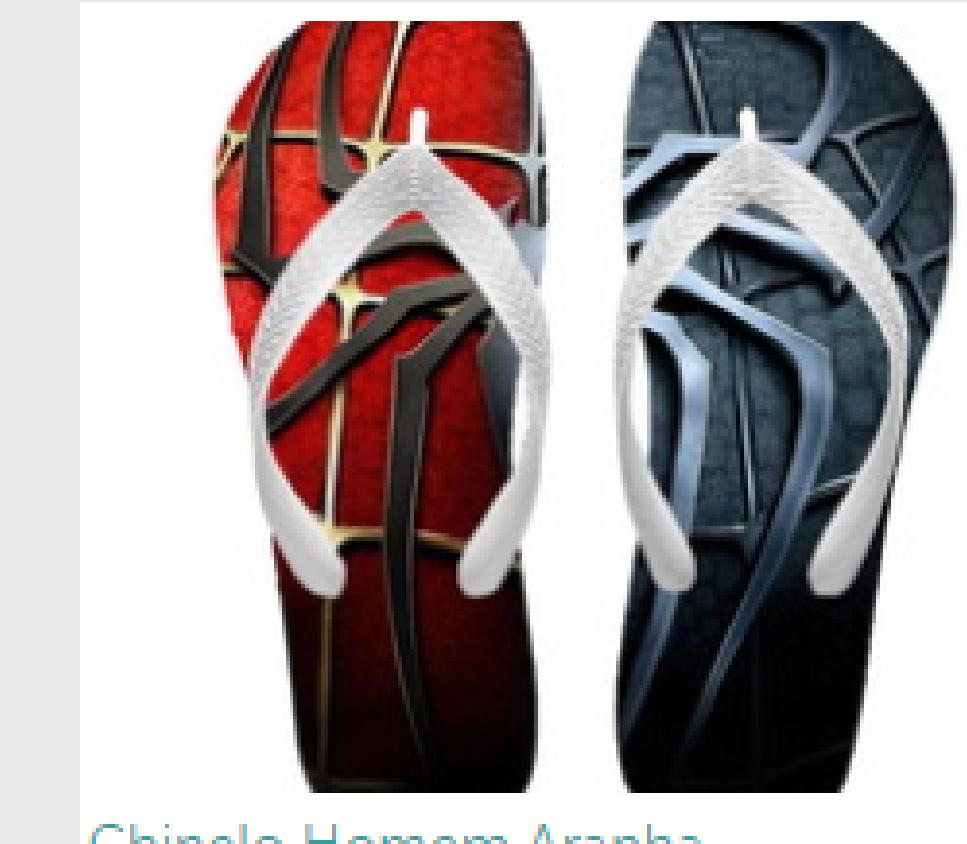
Camiseta Homem Aranha



Camiseta Senhor dos Aneis



Chinelo Batman Cute Infantil



Chinelo Homem Aranha



Chinelo Senhor dos Aneis

I

4. Clustering: Hierarchical Clustering

- Por tipo de produto?



Camiseta Batman



Camiseta Homem Aranha



Camiseta Senhor dos Aneis



Chinelo Batman Cute Infantil



Chinelo Homem Aranha



Chinelo Senhor dos Aneis

I

4. Clustering: Hierarchical Clustering

- Por tema?



Camiseta Batman



Camiseta Homem Aranha



Camiseta Senhor dos Aneis



Chinelo Batman Cute Infantil



Chinelo Homem Aranha

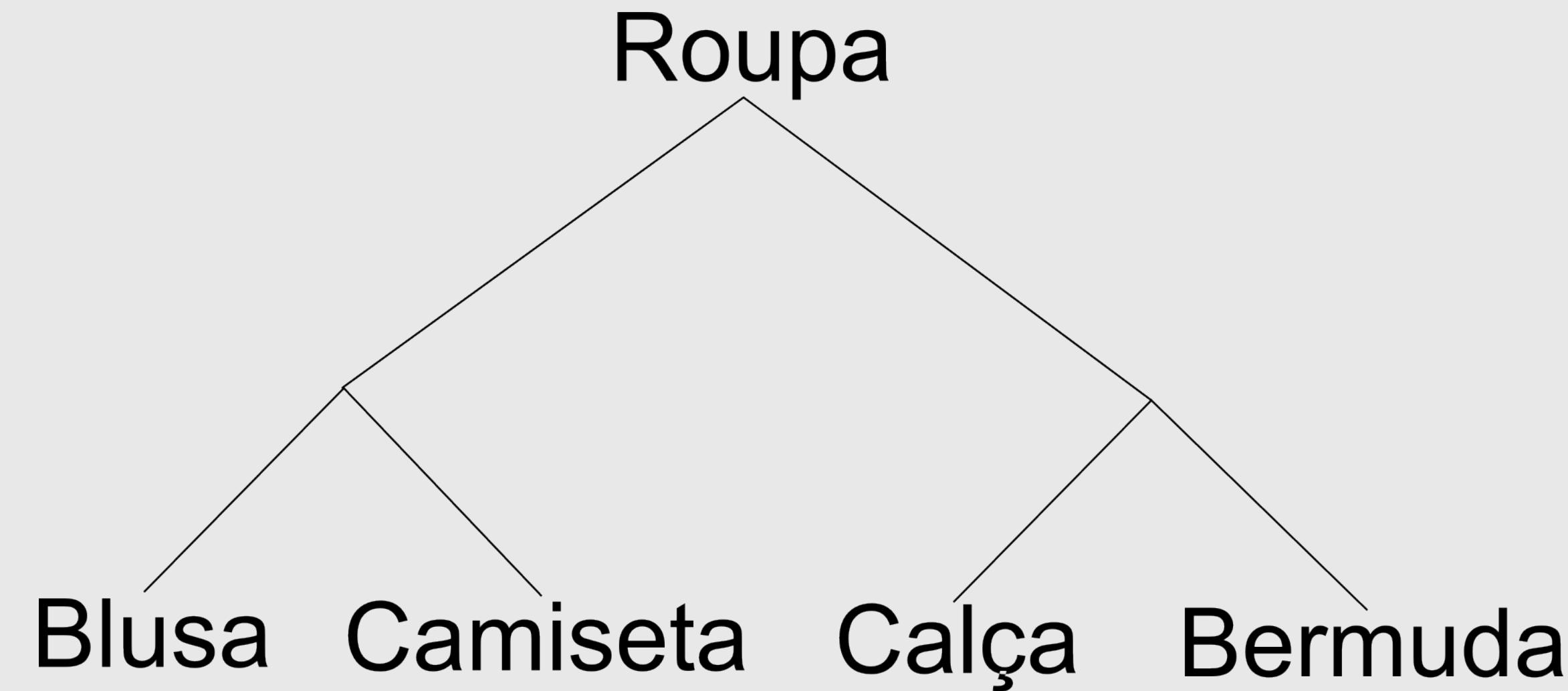


Chinelo Senhor dos Aneis

T

4. Clustering: Hierarchical Clustering

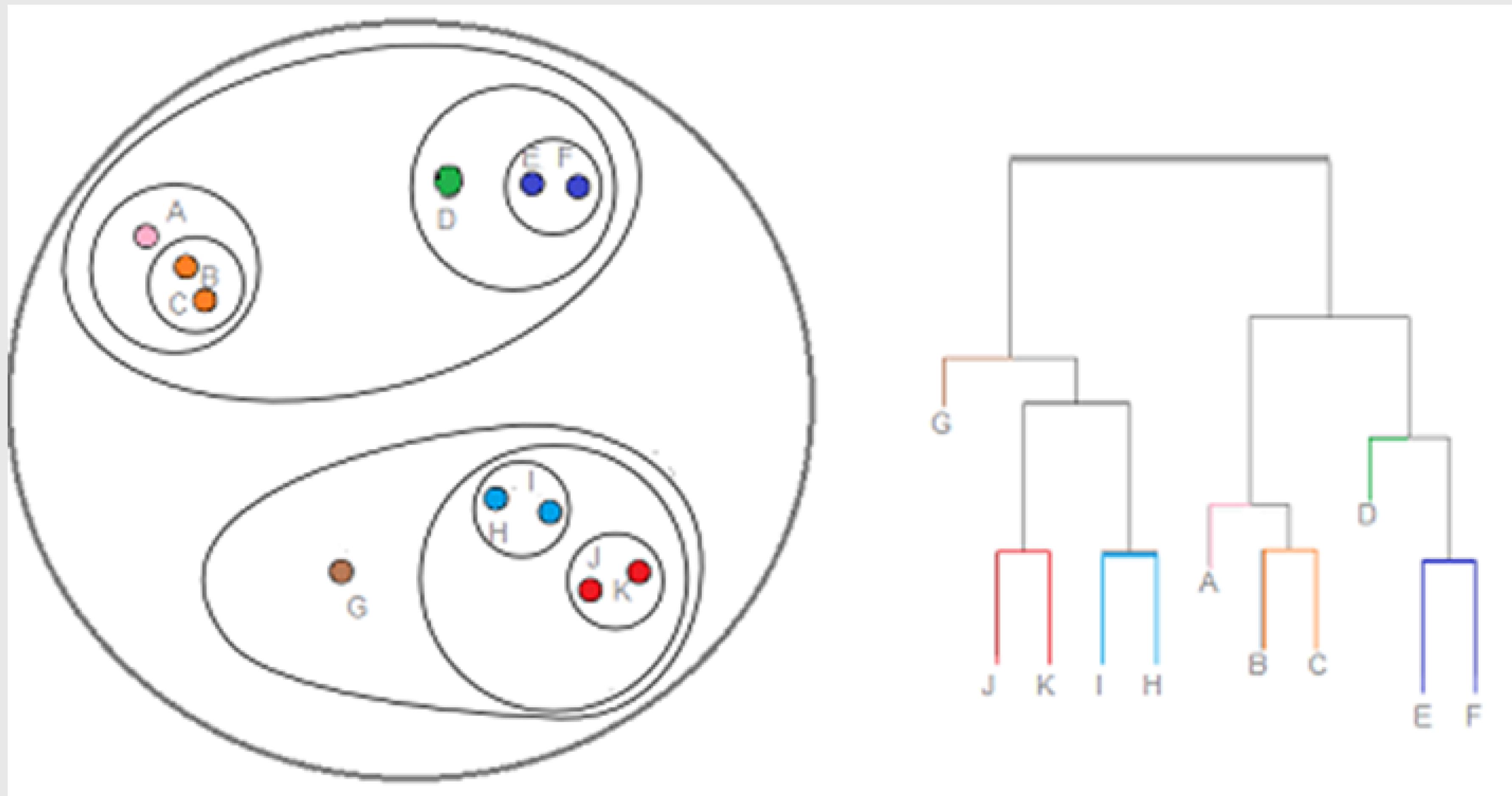
- Existe hierarquia



I

4. Clustering: Hierarchical Clustering

- Visualizar histórico de divisão: **Dendrograma**



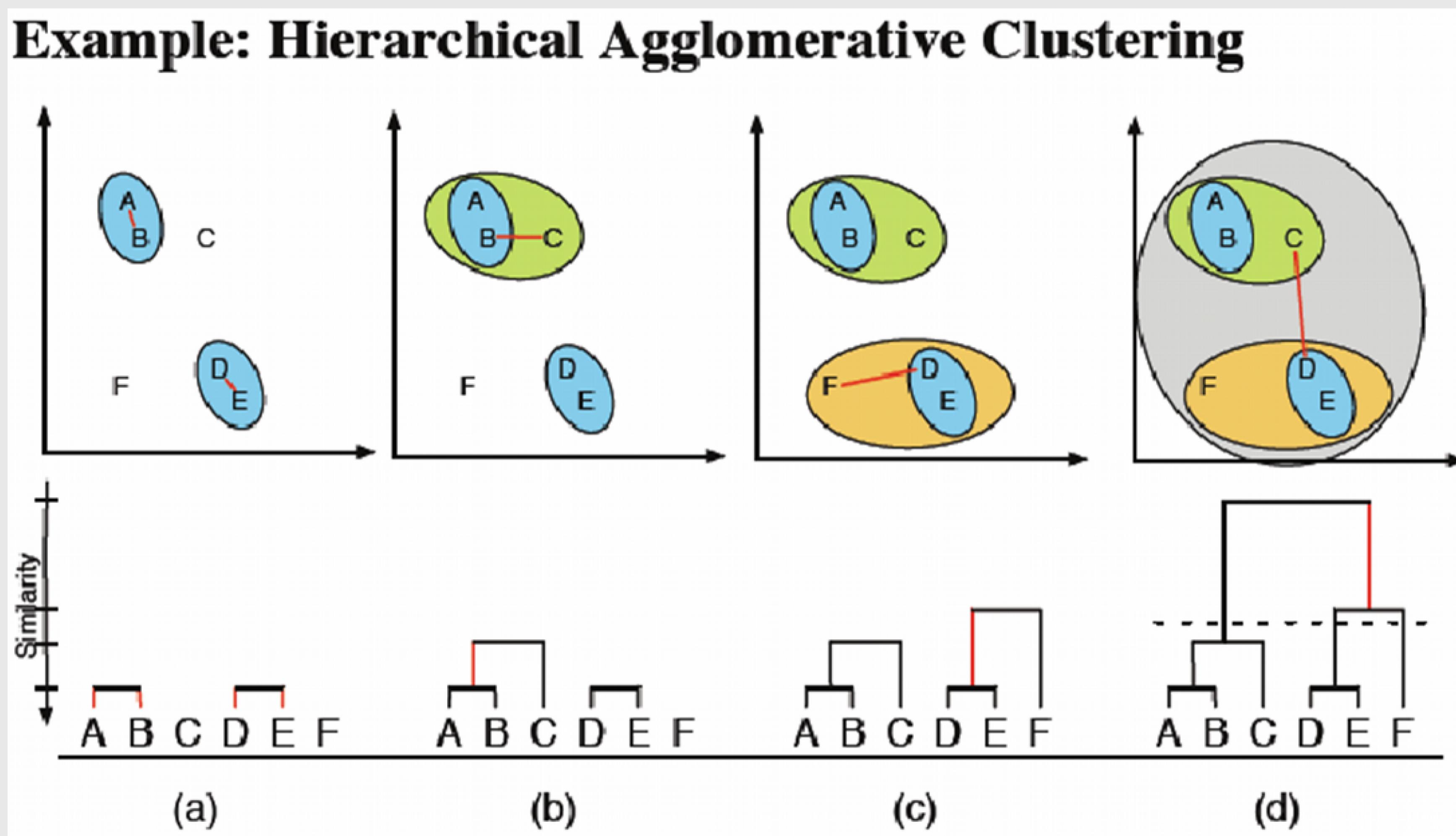
4. Clustering: Hierarchical Clustering

- 2 métodos principais:
 - Aglomerativo (Mais utilizado)
 - Por divisão

T

4. Clustering: Hierarchical Clustering

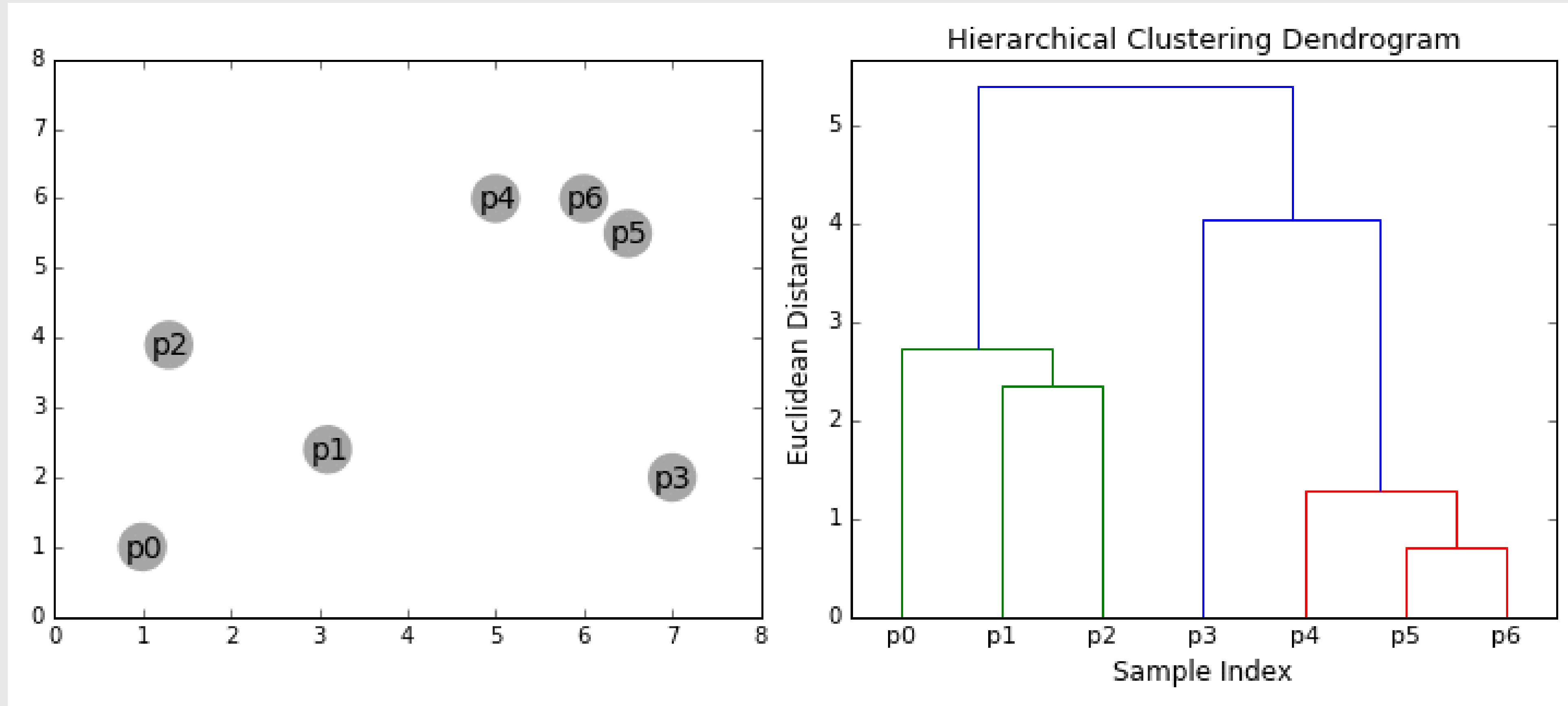
- Método aglomerativo:



I

4. Clustering: Hierarchical Clustering

- Método aglomerativo:



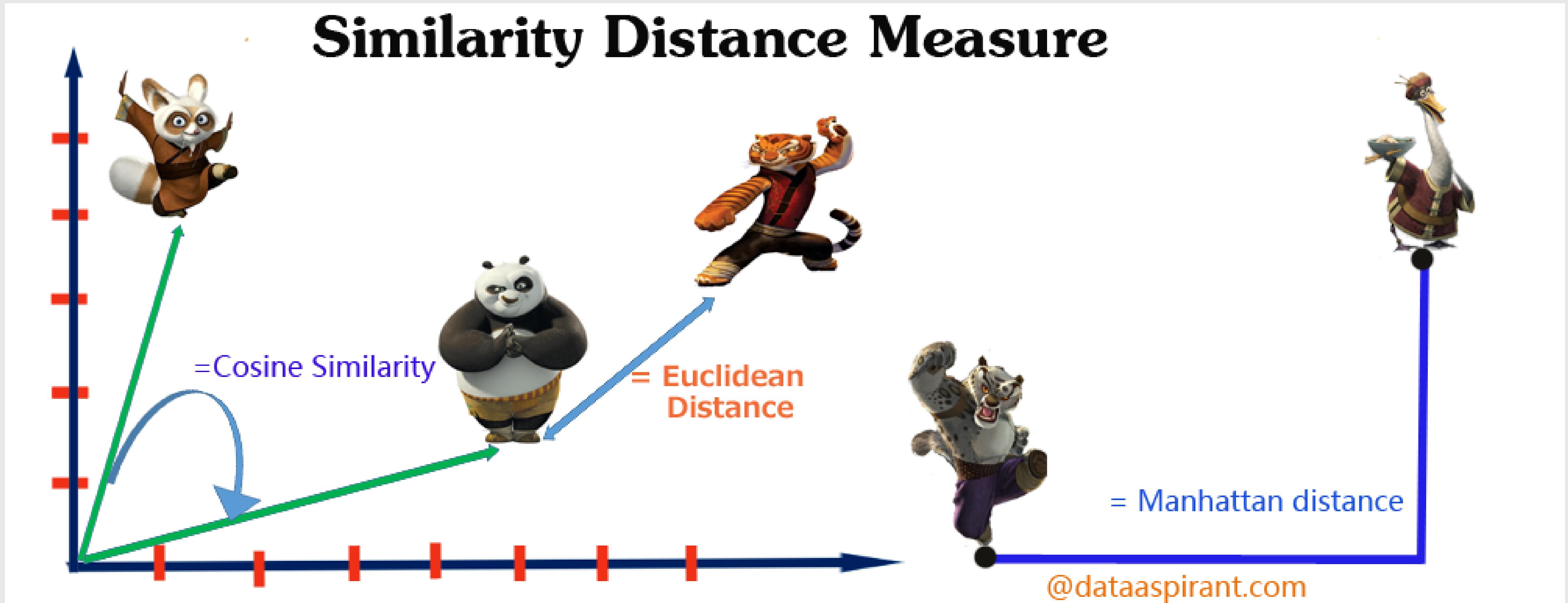
4. Clustering: Hierarchical Clustering

- Como agrupar clusters?
 - Distância (afinidade) entre clusters:
 - Euclidiana
 - Manhattan
 - Cosseno

T

4. Clustering: Hierarchical Clustering

- Tipos de distância:



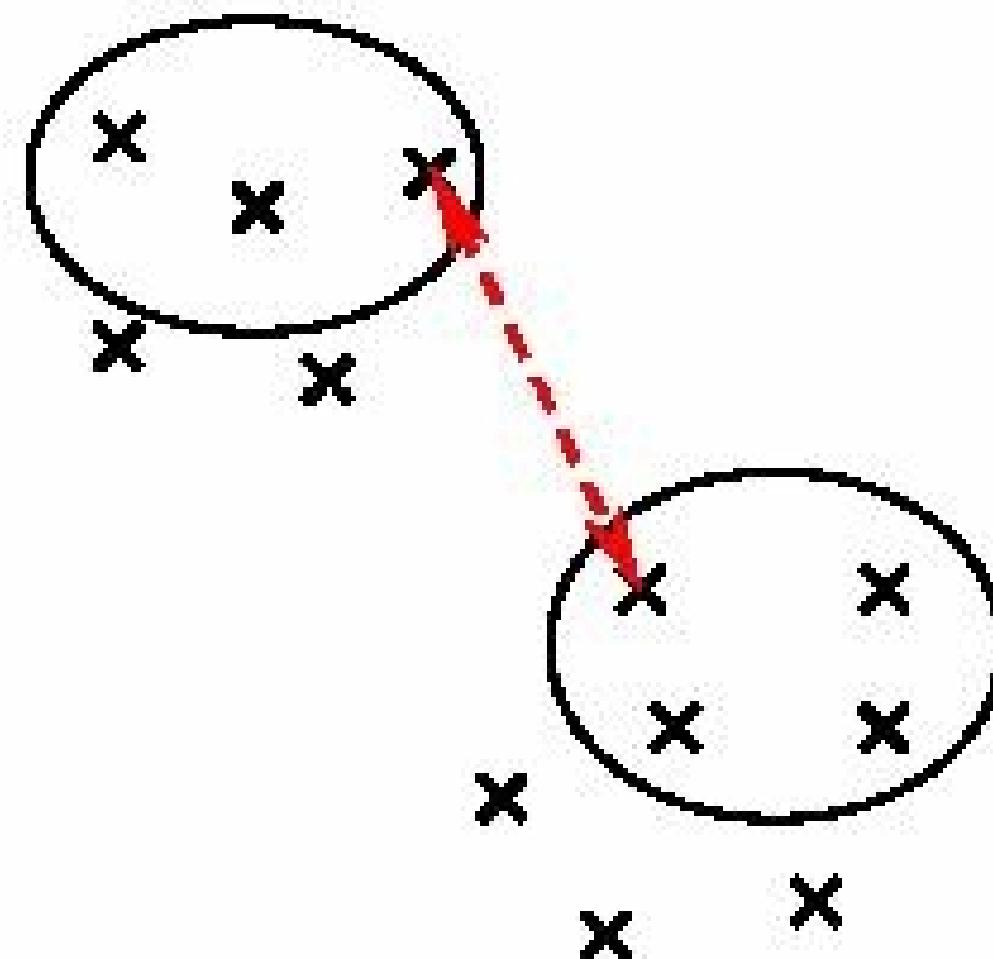
4. Clustering: Hierarchical Clustering

- Como agrupar clusters?
 - Tipos de ligação:
 - **Ward**: Mínima variância
 - **Completa**: Máxima distância
 - **Média**: Distância média
 - **Simples**: Menor distância

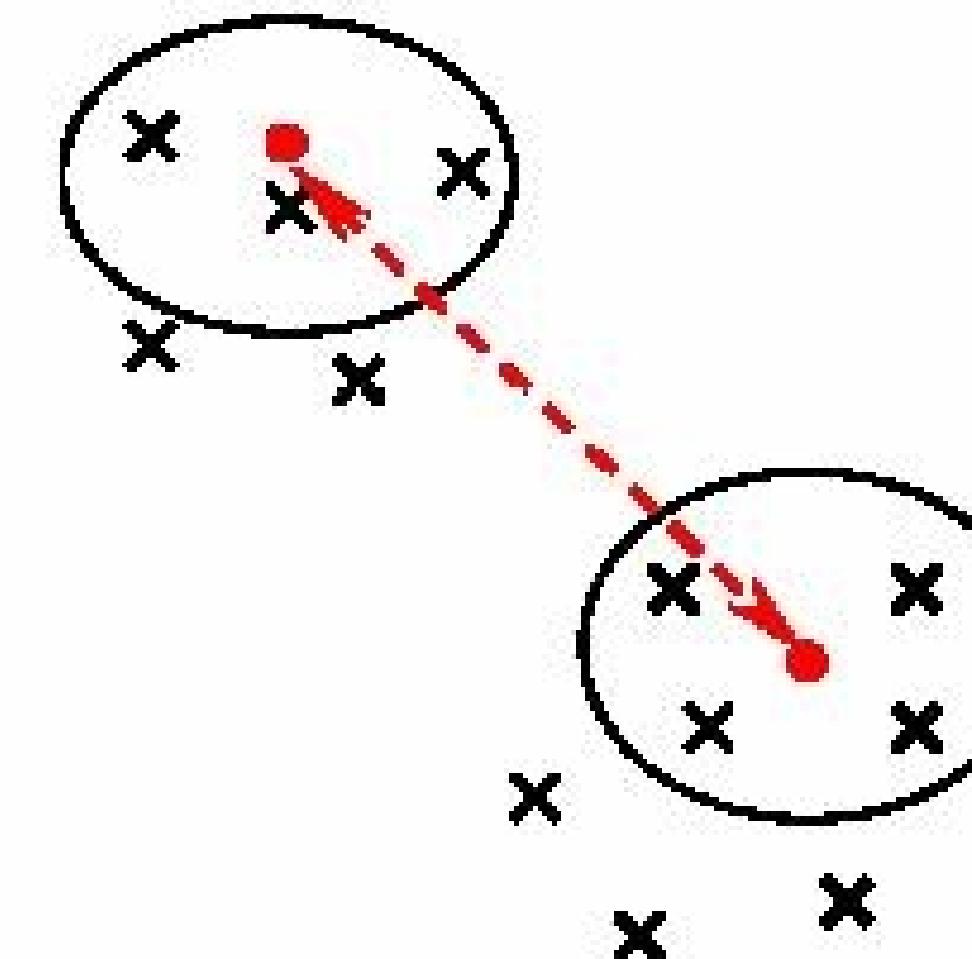
4. Clustering: Hierarchical Clustering

- Tipos de ligação:

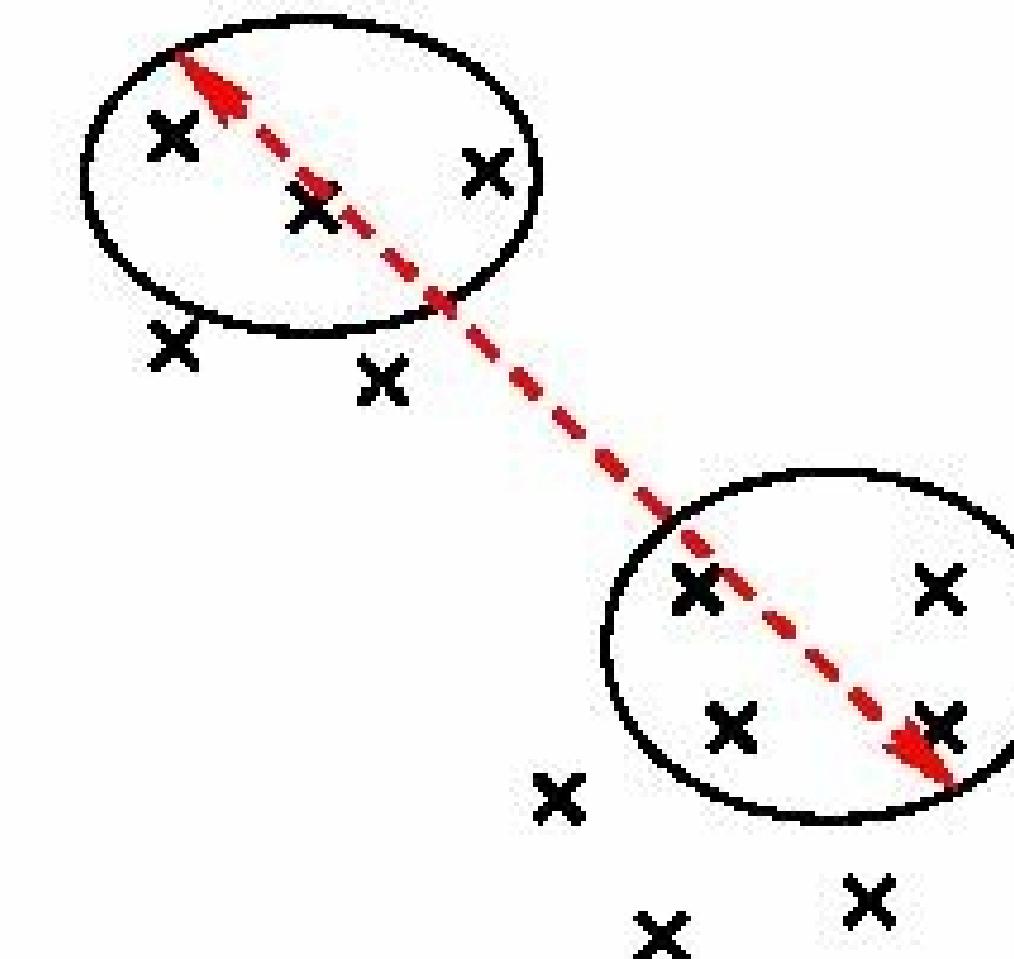
- Simple linkage



- Average linkage



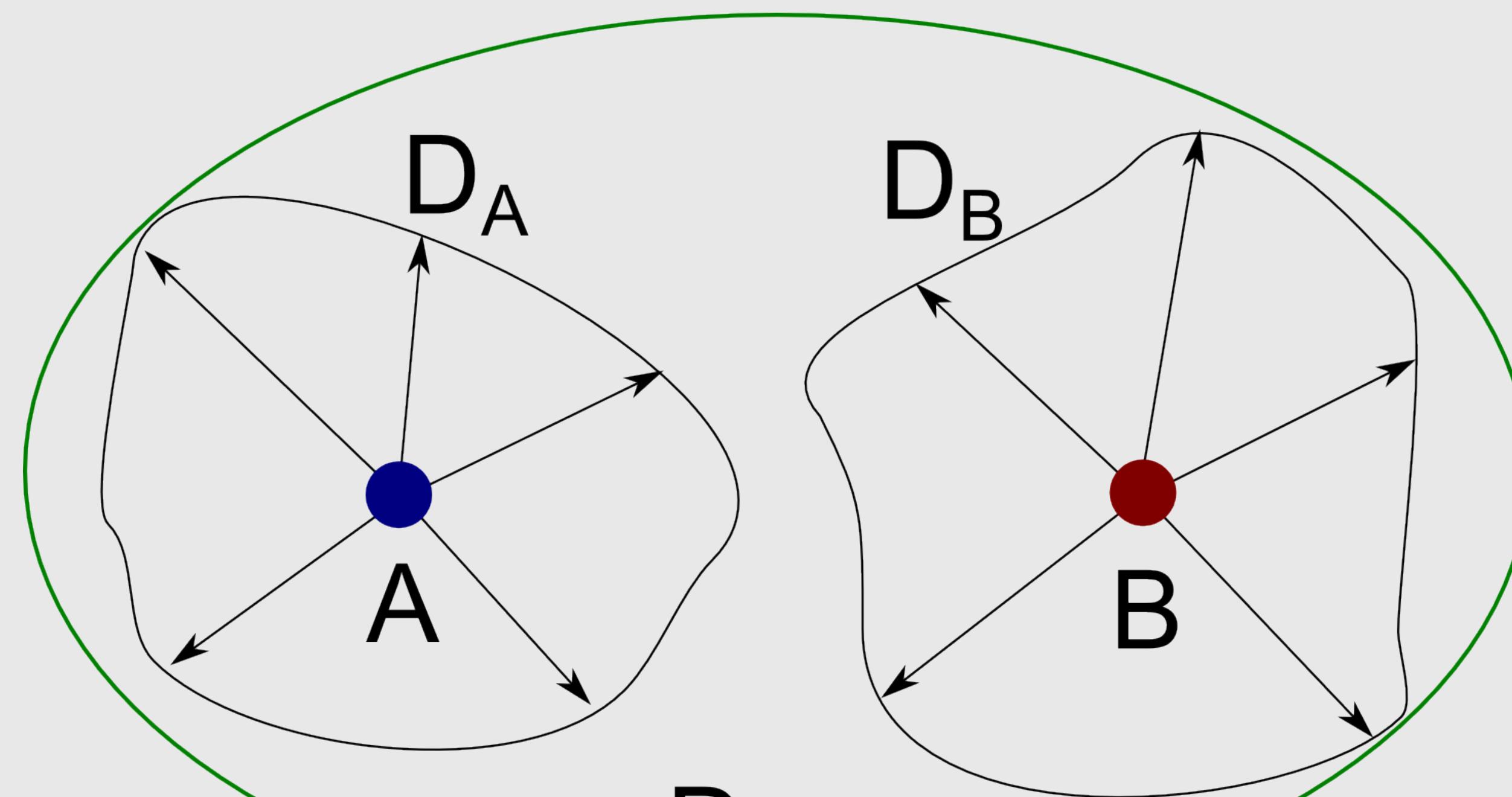
- Complete linkage



T

4. Clustering: Hierarchical Clustering

- Tipos de ligação: **Ward**

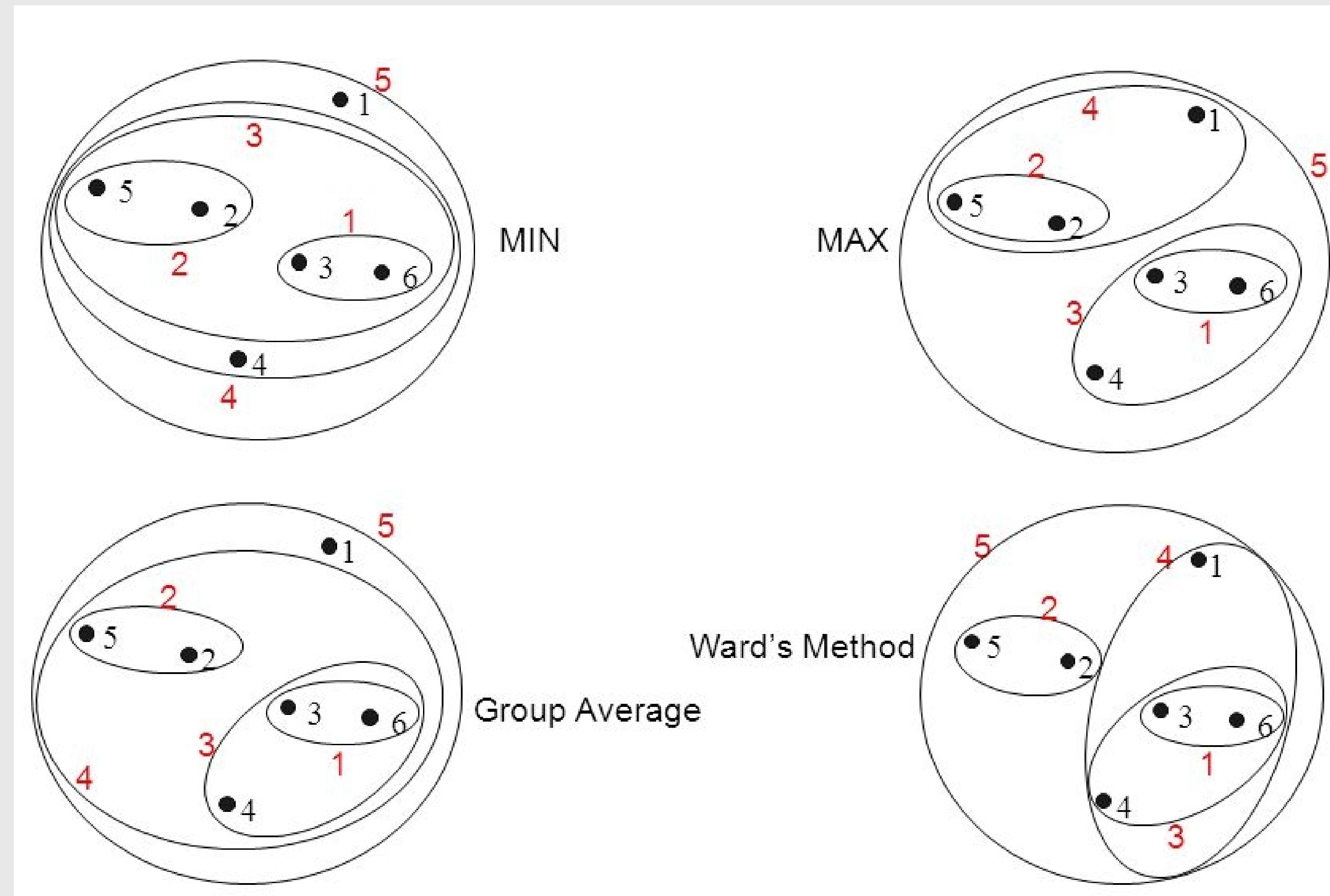


$$\text{Ward} = D_A + D_B - D_{AB}$$

I

4. Clustering: Hierarchical Clustering

- Tipos de ligação: Resultados distintos!



T

4. Clustering: Hierarchical Clustering

- Exemplo: notebook

4. Hierarchical Clustering

- **Vantagens**
 - Algoritmo simples
 - Resultado fácil de ser explicado
 - Não precisa definir o número de clusters
 - Permite analisar as relações entre clusters (hierarquia)
 - Resposta determinística

4. Hierarchical Clustering

- **Desvantagens**
 - Inviável com datasets grandes
 - Pode convergir para decisões indesejadas
 - Difícil definir melhor região de corte
 - Depende da escolha de parâmetros de distância e ligação

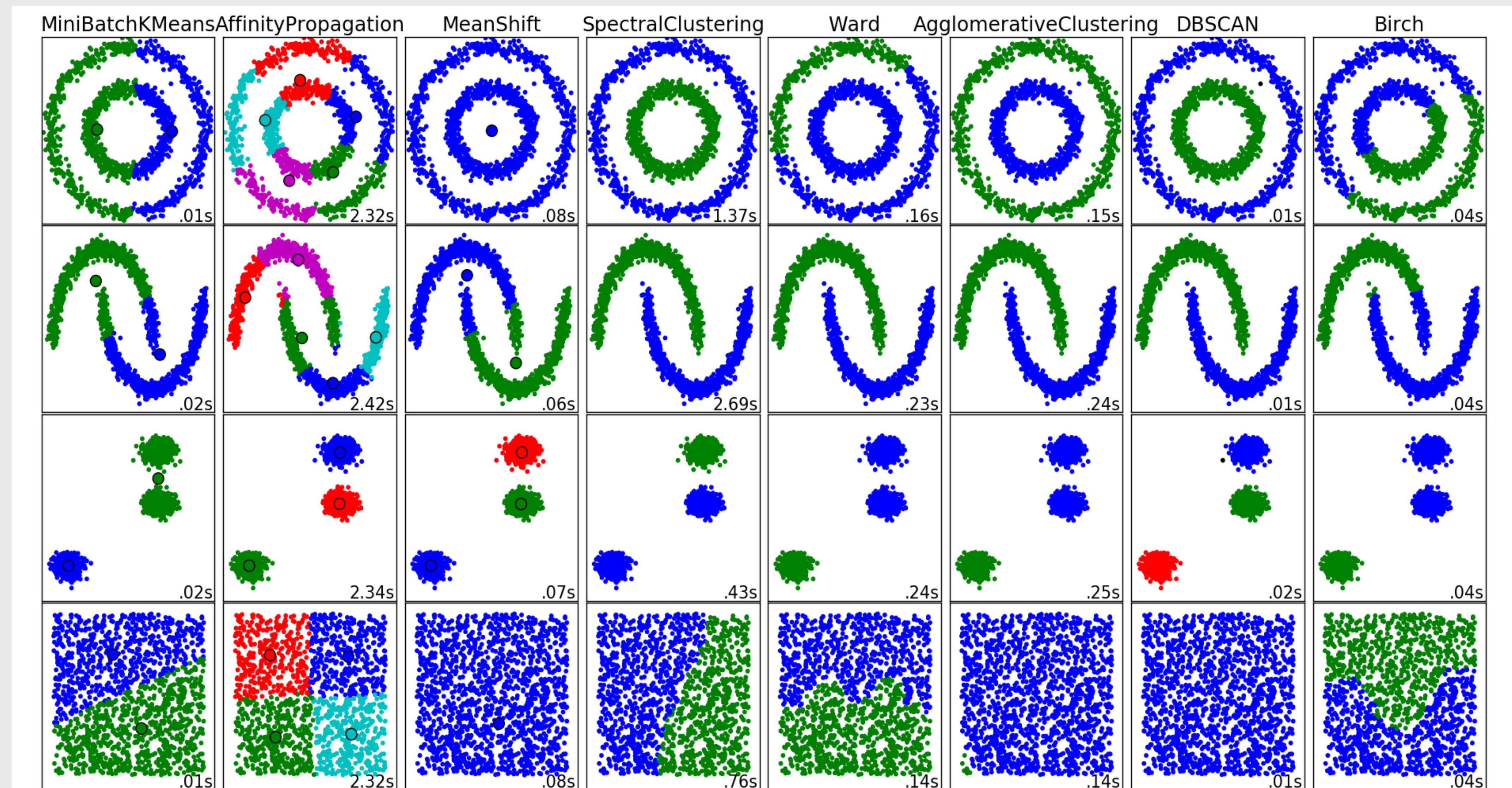
T

Clustering

- Outros algoritmos:
 - **Expectation-Maximization (EM)**
 - **Birch**
 - **Spectral Clustering**
 - etc

T

Clustering



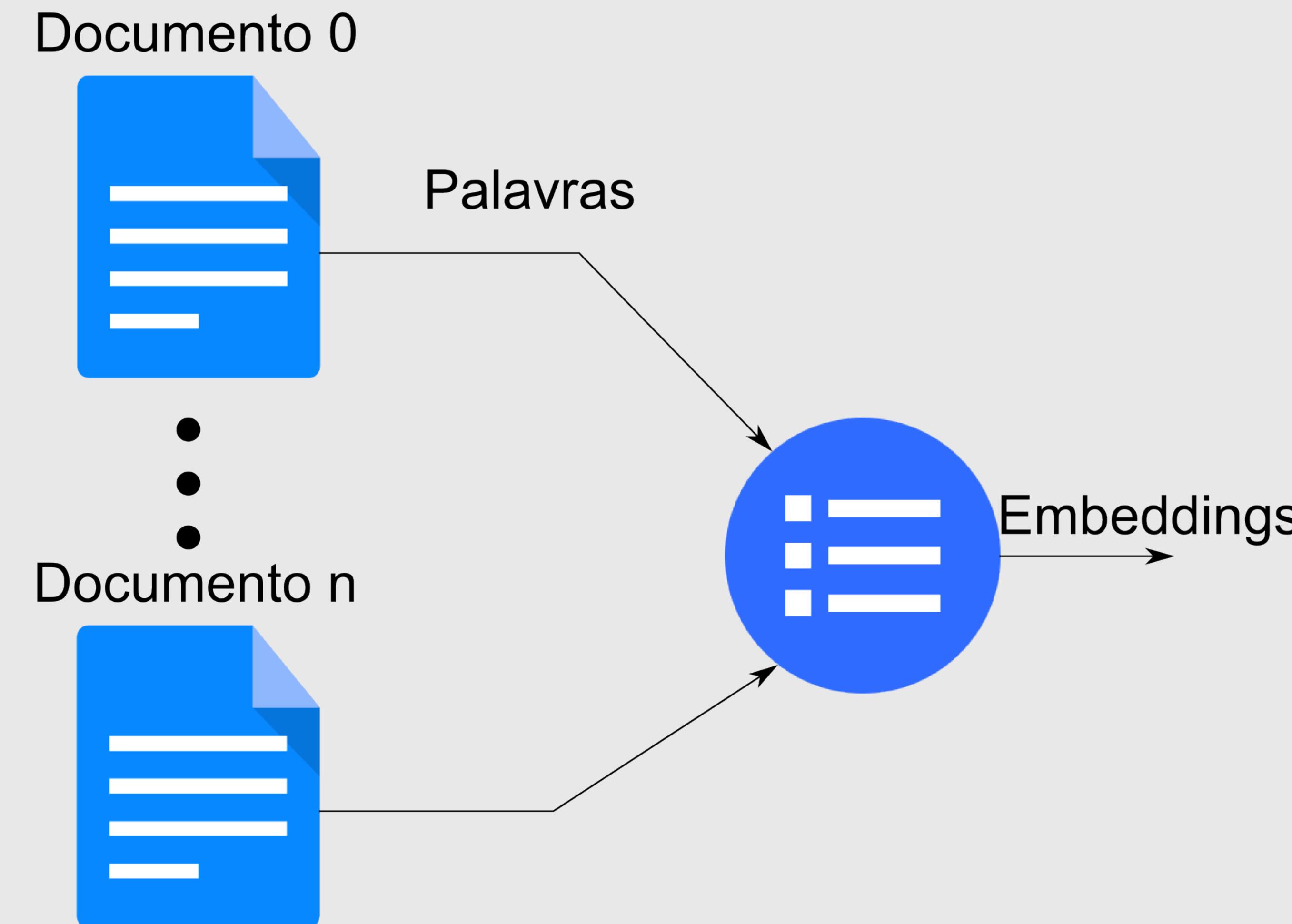
5. Clustering NLP

- Objetivos:
 - Encontrar relações entre documentos
 - Relações entre palavras
 - Documentos semelhantes possuem conjuntos de palavras semelhantes

T

5. Clustering NLP

- Vetor de atributos de um texto:



T

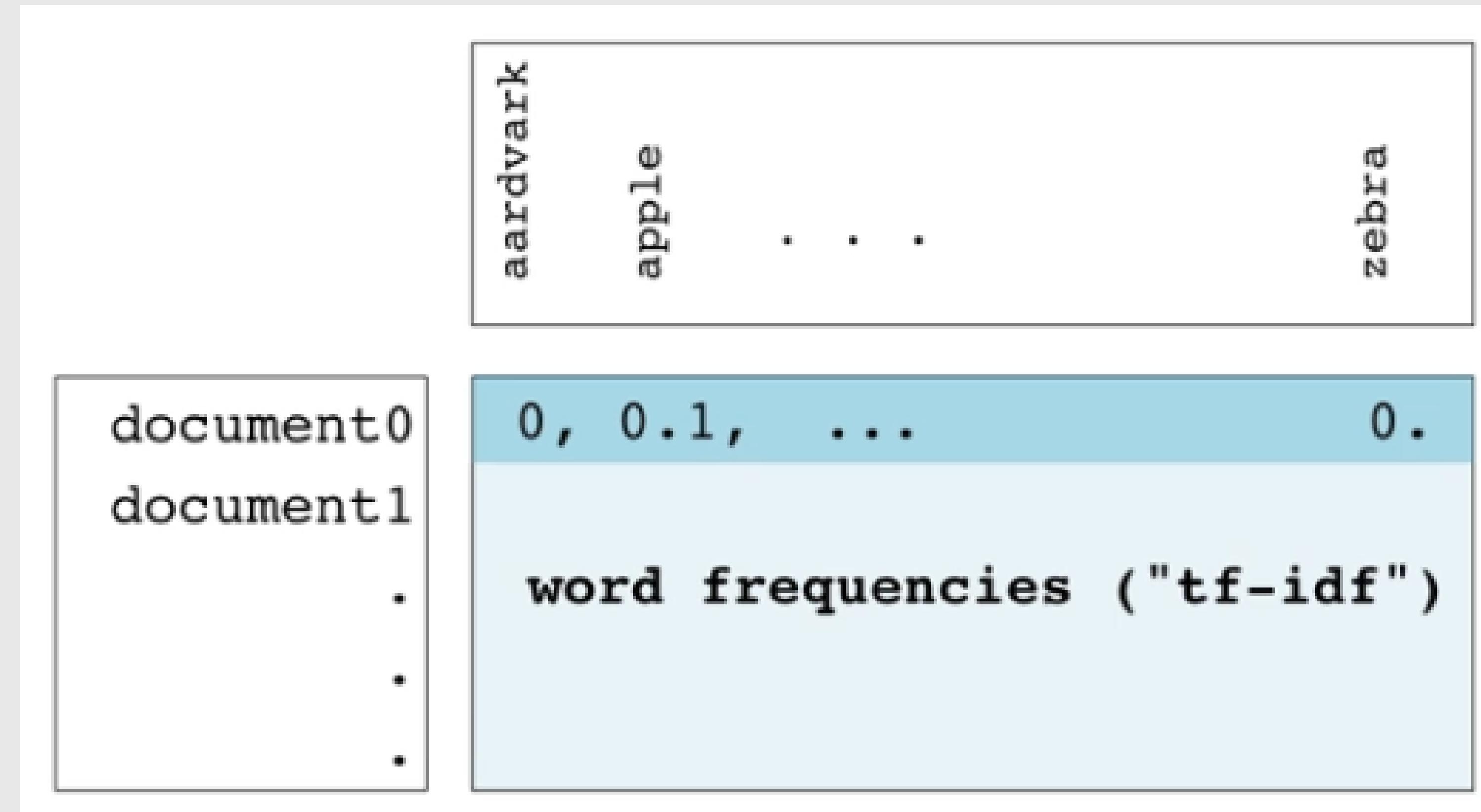
5. Clustering NLP

- Vetor de atributos de um texto (embedding) - X:
 - Bag-of-Words
 - TF-IDF
 - Word2Vec
 - Glove
 - ...

I

5. Clustering NLP

- Vetor de atributos de um texto:



T

5. Clustering NLP

- Exemplo: Clustering momento de compras Elo7

I

Case Elo7 - Subcategorias

- Navegação categorias Elo7

[elo7](#)

Produtos Buscar produtos

Cadastrar Entrar

Categorias

- Acessórios
- Aniversário e Festas
- Bebê
- Bijuterias
- Bolsas e Carteiras
- Casa
- Casamento
- Convites
- Decoração
- Doces
- Eco
- Infantil
- Jogos e Brinquedos
- Jóias
- Lembrancinhas
- Papel e Cia
- Pets
- Religiosos
- Roupas
- Saúde e Beleza
- Técnicas de Artesanato
- Materiais para artesanato

bazar do elo7

Produtos incríveis escolhidos pelas influenciadoras

CONFIRA

@isabelamarques @brugalliano @apartamento_203 @historiasdecasa

Saída Maternidade Bata Flor Goli...
Tchukinhos
R\$ 99,90

Convite de Casamento - 240g
Conviteria Royale
R\$ 1,50 R\$ 1,90

Livro do bebê elefantinho
Mundo de Jozi
R\$ 186,19

Pingente acrílico pezinho cristal
Vânia Campos Lembrancinhas
R\$ 0,34 R\$ 0,40

Bandana Pet Limonada Geladinh...
Estilo Peludo
R\$ 21,90 R\$ 24,90

Lugar Americano com Renda
Mimi ateliê
R\$ 12,90

Kit 3 Vasos Concreto Grecchi c/s...
SANTO CONCRETO
R\$ 76,70 R\$ 80,70

Vaso de Cimento Catedral
Deckler Garten
R\$ 44,90

I

Case Elo7 - Subcategorias

- Navegação categorias Elo7

N1

Categorias

- Acessórios
- Aniversário e Festas
- Bebê
- Bijuterias
- Bolsas e Carteiras
- Casa
- Casamento
- Convites
- Decoração
- Doces
- Eco
- Infantil
- Jogos e Brinquedos
- Jóias
- Lembrancinhas
- Papel e Cia
- Pets
- Religiosos
- Roupas
- Saúde e Beleza
- Técnicas de Artesanato

bazar do elo7

Produtos incríveis escolhidos pelas influenciadoras

CONFIRA

Saída Maternidade Bata Flor Goli...
Tchukinhos
R\$ 99,90

Convite de Casamento - 240g
Conviteria Royale
R\$ 1,50 R\$ 1,90

Livro do bebê elefantinho
Mundo de Jozi
R\$ 186,19

Pingente acrílico pezinho cristal
Vânia Campos Lembrancinhas
R\$ 0,34 R\$ 0,40

Bandana Pet Limonada Geladinh...
Estilo Peludo
R\$ 21,90 R\$ 24,90

Lugar Americano com Renda
Mimi ateliê
R\$ 12,90

Kit 3 Vasos Concreto Grecchi c/s...
SANTO CONCRETO
R\$ 76,70 R\$ 80,70

Vaso de Cimento Catedral
Deckler Garten
R\$ 44,90

I Case Elo7 - Subcategorias

- Navegação categorias Elo7

The image shows two screenshots of the Elo7 website illustrating the navigation process.

Left Screenshot (N1): The main categories page. A blue arrow points to the 'Casamento' button in the sidebar, which is highlighted with a green border. The sidebar also lists other categories like Acessórios, Aniversário e Festas, Bebê, Bijuterias, etc. Below the sidebar, there's a banner for 'bazar do elo7' featuring influencer @isabel and several product cards: 'Saída Maternidade Bata Flor Gol...' (R\$ 99,90), 'Convite de Casar Conviteria Royale' (R\$ 1,50), 'Bandana Pet Limonada Geladinh...' (R\$ 21,90), and 'Lugar Americano Mimi atelié' (R\$ 12,90).

Right Screenshot (N2): The 'Casamento' subcategory page. A blue arrow points to the 'Casamento' category in the sidebar, which is highlighted with a red border. The sidebar lists various sub-categories under 'Casamento'. The main content area shows search filters for 'Preço' (R\$) and 'Cidade', followed by product cards: 'Noivinhos Infantil' by IRENE ARTES (R\$ 43,90), 'Almofada rústica pa...' (partially visible), 'SAVE THE DATE CASAMENTO E NOI...' by ELEN DESIGN (R\$ 25,00 FRETE GRÁTIS), and 'Tiara mini flores Ma...' (partially visible). A banner at the bottom says 'O casamento vai sair!'.

I Case Elo7 - Subcategorias

- Árvore de categorias do Elo7:
 - Limitada a dois níveis (N1 e N2)
 - Usabilidade baixa para o comprador e vendedor
 - Consequência: produtos mal categorizados (ciclo vicioso)

I Case Elo7 - Subcategorias

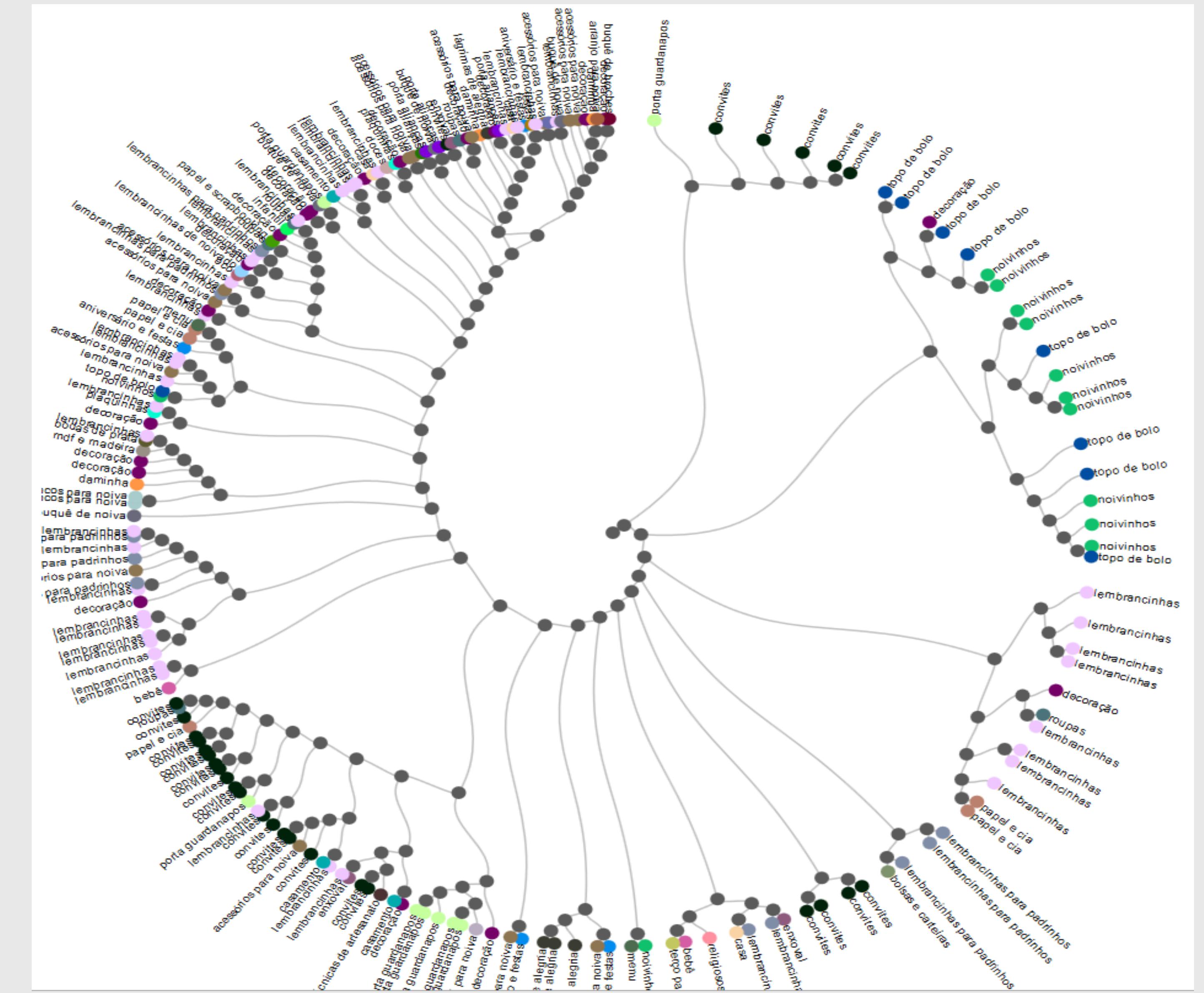
- Desejo do Elo7:
 1. Aumentar granularidade das categorias (mais níveis)
 2. Automatizar o processo de “subcategorização” dos produtos
 3. Subcategorias devem ter hierarquia



Clustering

T Case Elo7 - Subcategorias

- Dendrograma dos produtos de casamento



T

OBRIGADO!