

# Do caos ao ordenado: Como a clusterização organiza e revela padrões em dados

Aline Lunkes<sup>1</sup>  
Jane Thais Soares de Oliveira<sup>2</sup>  
Honovan Paz Rocha<sup>1,2</sup>

<sup>1</sup>Programa de Pós Graduação em Modelagem Computacional e Sistemas (PPGMCS)  
Universidade Estadual de Montes Claros (Unimontes)

<sup>2</sup>Instituto de Engenharia, Ciência e Tecnologia (IECT)  
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM)

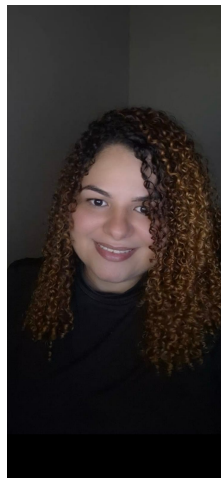
23 de fevereiro de 2025

# Quem Sou Eu?

## Jane Thais Oliveira

*Bacharel em Ciência e Tecnologia (UFVJM) Pós Graduação em Engenharia de Dados (Puc Minas)*

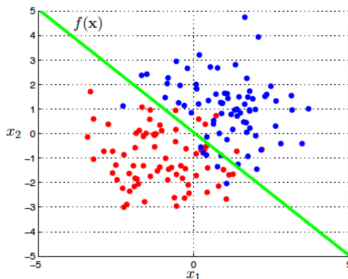
- Futura Engenheira Física pela UFVJM.
- Atua como Engenharia de Dados a +/- 4 anos.
- Atuação em projeto de pesquisa que visa aplicar e avaliar diferentes técnicas de clusterização dentro da área de Marketing.
- Projeto de pesquisa voltando implementação de técnicas de Machine Learning para classificação de dados na área de crédito.
- Interesse em Inteligência Artificial, Ciência de Dados e Clusterização
- LinkedIn: [linkedin.com/in/jane-thais-oliveira](https://www.linkedin.com/in/jane-thais-oliveira)



- 1 Introdução
  - Aprendizado Não-Supervisionado VS Supervisionado
- 2 Clustering
  - Motivação
- 3 Clustering
  - Agrupamento de Dados
  - Processo de Clusterização
  - Medidas de Similaridade e Dissimilaridade
  - Algoritmos de Agrupamento
  - Validação de Clusterização
- 4 K-means

# Supervisionado

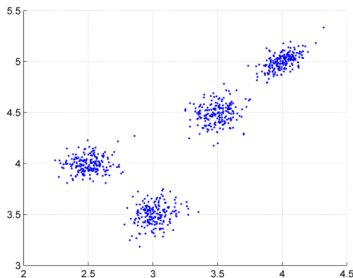
- Em Aprendizado Supervisionado, temos um conjunto de features para cada exemplo de entrada, além de um rótulo (target). O objetivo é realizar alguma predição.
- Principais Tarefas em Aprendizado Supervisionado:
  - Classificação e Regressão



Fonte - Slides de Aula - Cristiano L. Castro (UFMG)

# Não-Supervisionado

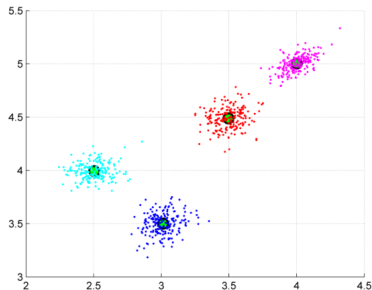
- Em aprendizado Não-Supervisionado, temos o conjunto de features para cada exemplo de entrada, mas não temos a variável rótulo (*target*). O objetivo não é previsão, mas encontrar alguma estrutura de interesse nos dados.



Fonte - Slides de Aula - Cristiano L. Castro (UFMG)

# Não-Supervisionado

- Em aprendizado Não-Supervisionado, temos o conjunto de features para cada exemplo de entrada, mas não temos a variável rótulo (*target*). O objetivo não é predição, mas encontrar alguma estrutura de interesse nos dados.



Fonte - Slides de Aula - Cristiano L. Castro (UFMG)

# Não-Supervisionado

- Principais técnicas:

- Clusterização: Agrupamento de dados semelhantes.
- Redução de Dimensionalidade (PCA): Encontrar representações compactas dos dados sem perder informações essenciais.

Nesta aula, focaremos em Clusterização, por permitir organizar grandes volumes de dados sem supervisão.

# Motivação

## Contexto Relevante

- Crescimento exponencial de dados em diversas áreas, como marketing, saúde e tecnologia.
- Necessidade de compreender **padrões ocultos** em grandes volumes de dados para otimizar estratégias e decisões.
- Métodos tradicionais de análise não capturam a complexidade e diversidade dos dados.



# Dados, dados e mais dados

Mas você sabe o que são dados?



# Dados, dados e mais dados

Mas você sabe o que são dados?



São registros observados ou medidos que podem se transformar em informação.

# Problema

## Por que se preocupar com agrupamentos de dados?

- Diariamente, o mundo gera cerca de 2,5 quintilhões de dados. E 90% dos dados disponíveis hoje foram gerados nos últimos 3 anos, segundo o IBM.
- O comportamento do Consumidor Moderno. O estudo da Epsilon (2018) aponta 80% dos consumidores preferem experiência personalizada. 90% acham atraente personalização.
- Em segurança, queremos detectar fraudes bancárias agrupando transações suspeitas. A Kaspersky, empresa especializada em cibersegurança, registrou 1,92 milhões de tentativas de golpes em pequenas e médias empresas entre outubro de 2022 e outubro de 2023.

**Lembre-se: Dados não são informação, mas podem se tornar!**

# motivação

Portanto, temos que:

- É mais fácil e barato obter dados não rotulados. Dados rotulados necessitam de intervenção humana.
  - Reconhecimento de Fala, dados de sensores, Dados clínicos para diversas tarefas, etc...
  - Extração de Características e Redução de Dimensionalidade;
- O entendimento da variância e de agrupamentos estruturais dos dados, podem ser uma ferramenta de pré-processamento extremamente importante para o aprendizado Supervisionado.

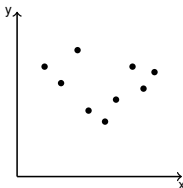
# Agrupamento de Dados

## Definição

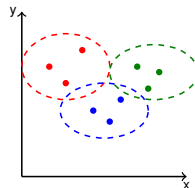
- Agrupamento (*clustering*) é a organização de dados em grupos (*clusters*) de acordo com a similaridade entre os objetos.
- Objetivo: Maximizar a similaridade dentro dos grupos e minimizar a similaridade entre grupos.

Carmichael e Julius (1968)

Antes do Cluster



Depois do Cluster



Fonte: Autor

# Definição Matemática da Clusterização

Dado um conjunto de pontos no espaço  $X = \{x_1, x_2, x_3, \dots, x_n\}$ , o objetivo da clusterização é dividir esses pontos em  $k$  grupos  $C_1, C_2, C_3, \dots, C_k$  tal que:

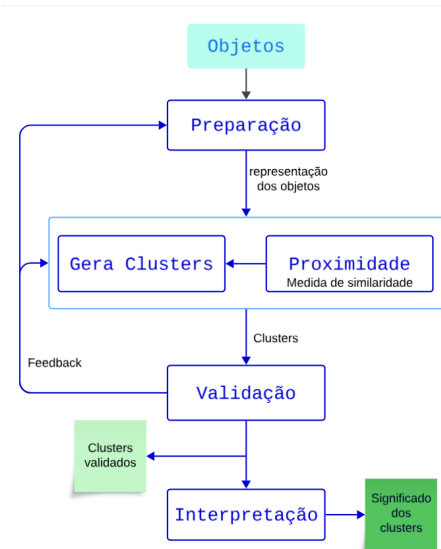
- **Máxima similaridade intra-cluster** – Os pontos dentro de um cluster devem estar próximos uns dos outros.
- **Mínima similaridade inter-cluster** – Os pontos de clusters diferentes devem estar bem separados.

# Processo de Clusterização

## Etapas principais

- **Preparação dos Dados:** Remoção de valores ausentes, tratamento de *outliers* e normalização.
- **Medida de Similaridade:** Cálculo da proximidade entre os objetos, com base em distâncias (e.g., Euclidiana, Manhattan).
- **Aplicação do Algoritmo:** Execução do método de clusterização selecionado.
- **Validação:** Avaliação da qualidade dos *clusters* gerados (e.g., Índice de Silhueta, Davies-Bouldin).

(FACELI, 2006)



# Medidas de Similaridade e Dissimilaridade

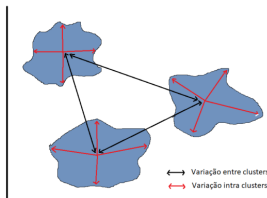
## Definição

- Similaridade: Mede o grau de semelhança entre dois objetos.
- Dissimilaridade: Mede o grau de diferença entre dois objetos.

(CASTRO; FERRARI, 2017)

## Medidas mais comuns

- **Distância Euclidiana:**  $d(x_i, x_j) = \sqrt{\sum_{l=1}^m (x_{il} - x_{jl})^2}$
- **Distância de Manhattan:**  $d(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}|$



Fonte: Adaptada de Lauretto (2017)



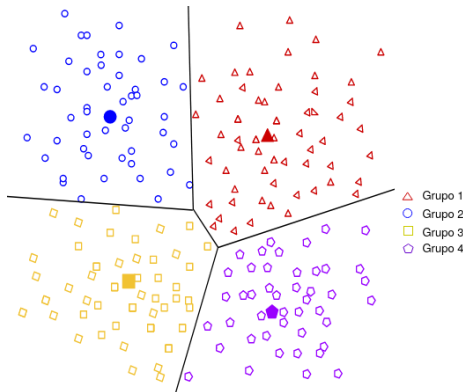
# Algoritmos de Agrupamento

## Classificação dos Algoritmos

- **Particionais:** Dividem os dados diretamente em  $k$  *clusters* (e.g., K-Means, K-Medoids).
- **Hierárquicos:** Formam uma estrutura em árvore (*dendrograma*) (e.g., Ward's Method).
- **Baseados em Densidade:** Identificam *clusters* como regiões densas no espaço (e.g., DBSCAN).

# Particionais ou Centroides

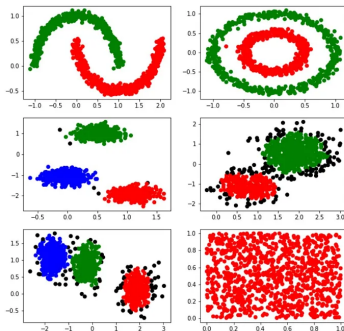
Os algoritmos de agrupamento baseados em centroide são eficientes, mas sensíveis a condições iniciais e valores discrepantes. Entre eles, o k-means é o mais usado. Ela exige que os usuários definam o número de centroides,  $k$ , e funciona bem com clusters de tamanho aproximadamente igual.



Fonte: Livro Introdução ao Machine Learning do grupo DataAt

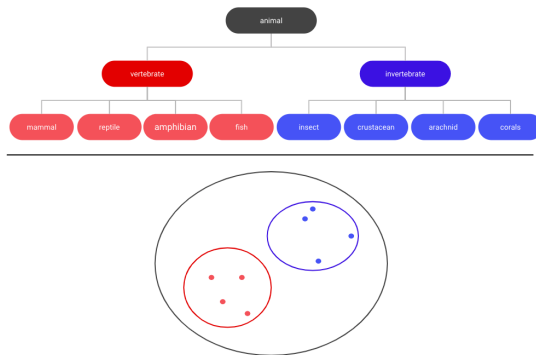
# Densidade

Este agrupamento identifica regiões alta densidade de pontos, formando clusters sem a necessidade de definir um número fixo previamente. Esse método permite detectar clusters de formatos arbitrários e separar pontos discrepantes, que não são atribuídos a nenhum grupo. No entanto, sofre ao lidar com clusters de densidades variadas e conjuntos de dados de alta dimensionalidade.



# Hierarquico

O clustering hierárquico cria uma árvore de clusters. Sendo adequado para dados hierárquicos, como taxonomias. É possível escolher qualquer número de clusters cortando a árvore no nível certo.



# Complexidade da Clusterização

## Desafios e Limitações

- **Explosão Combinatória:** A busca pela solução ótima torna-se inviável em grandes conjuntos devido à enorme quantidade de combinações possíveis.
- **Densidade e Distribuição:** Variações de densidade e distribuições complexas dificultam a aplicação de certos algoritmos.
- **Dimensionalidade:** Muitas dimensões prejudicam medidas de similaridade (*maldição da dimensionalidade*).
- **Tipo de Dados:** Diferentes tipos de variáveis exigem abordagens específicas, aumentando a complexidade.

(JAIN; DUBES, 1988), (LIU et al., 1968), (NALDI, 2011) e etc.

# Validação de Clusterização

## Objetivo

- Garantir que os *clusters* gerados sejam significativos e representativos.

## Métricas de Validação

- **Índices Internos:** Avaliam a coesão e separação (e.g., Índice de Silhueta).
- **Índices Externos:** Comparação com uma partição de referência (e.g., Índice de Rand).
- **Índices Relativos:** Comparação entre diferentes configurações de parâmetros.

(CAMPELLO, 2007), (ROUSSEEUW, 1987), (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004) e etc.

# Métricas de Validação

| Métrica  | Descrição  |
|--|--|
| Índice de Silhueta (ROUSSEUW, 1987)                        | Mede a coesão interna e separação entre clusters.                                |
| <i>Índice Davies-Bouldin (DBI) (DAVIES; BOULDIN, 1979)</i> | Avalia a separação e compactação dos agrupamentos.                               |
| Índice PBM (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004)          | Busca maximizar as distâncias inter-grupo e minimizar as distâncias intra-grupo. |
| Índice de Dunn (BEZDEK; PAL, 1998)                         | Avalia a separação e compactação dos agrupamentos.                               |
| VCR (CALIŃSKI; HARABASZ, 1974)                             | valoriza a coesão interna dos grupos e a separação externa entre grupos.         |
| Índice de Rand (CAMPELLO, 2007)                            | Mede a similaridade entre duas partições.  |

# Algoritmos reconhecidos na literatura

- **K-Means:** Simplicidade e eficiência para dados numéricos.
- **K-Medoids:** Robusto contra *outliers*.
- **DBSCAN:** Baseado em densidade, ideal para *clusters* de formas arbitrárias.
- **Ward's Method:** Minimiza variância em uma estrutura hierárquica.
- **Expectation Maximization (EM):** Ajusta modelos probabilísticos para identificar *clusters*.

## Racional para Escolha

- Buscar diversidade de abordagens para atender diferentes tipos de dados e formatos de *clusters*.
- Testar vantagens e limitações de cada técnica no contexto a ser implementado.



# K-means ou k-médias

O K-Médias é um algoritmo de agrupamento amplamente utilizado para particionar um conjunto de dados em  $k$  grupos, onde  $k$  é o número de *clusters* desejado e deve ser definido previamente.

Ele organiza os dados de forma que os objetos de um mesmo grupo sejam mais similares entre si do que em relação aos objetos de outros grupos. A similaridade dentro de cada grupo é avaliada pelo cálculo do centroide, que representa a média dos objetos pertencentes ao grupo.

A função-objetivo do k-means é descrita pela Equação:

$$\min_{c_1, \dots, c_k} \text{obj} = \sum_{i=1}^k \sum_{x_j \in C_i} d(x_j, m_i) \quad (1)$$

onde,  $m_i$  é o centróide do grupo  $C_i$  e  $d(x_j, m_i)$  é a distancia euclidiana entre um ponto  $x_j$  e o centróide  $m_i$ .

# K-means

A principal meta do algoritmo é minimizar uma função de custo, que calcula a soma das distâncias quadráticas entre cada objeto e o centroide do seu respectivo grupo.

## Características:

- Algoritmo simples e eficiente, entretanto, o K-Médias não garante encontrar a melhor solução global, pois sua performance pode depender da escolha inicial dos centroides.
- A aplicação exige cuidado na definição do valor de  $k$  e na escolha da inicialização para garantir resultados consistentes e significativos.

# K-means - Algoritmo

---



---

```

Input:  $K$ (Número de clusters),  $\{x_{train}^{(i)}\} \ i = 1, 2, \dots, m$ 
1 begin
2   Inicializar aleatoriamente  $k$  centróides  $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ 
   repeat
     // minimize  $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$ 
       com  $\mu_1, \dots, \mu_K$  fixos
3     for  $i = 1$  to  $m$  do
4        $c^{(i)} =$  índice( $1, \dots, K$ ) do centróide mais próximo de  $x^{(i)}$ 
5     end

     // minimize  $J(\dots)$  com relação a  $\mu_1, \dots, \mu_K$ 
6     for  $k = 1$  to  $K$  do
7        $\mu_k =$  média dos pontos atribuídos ao cluster  $k$ 
8     end
9   until Convergência;
10 end

Result: Ótimos de  $[c^{(i)} \{i = 1, 2, \dots, m\}, \mu_k \{k = 1, 2, \dots, K\}]$ 

```

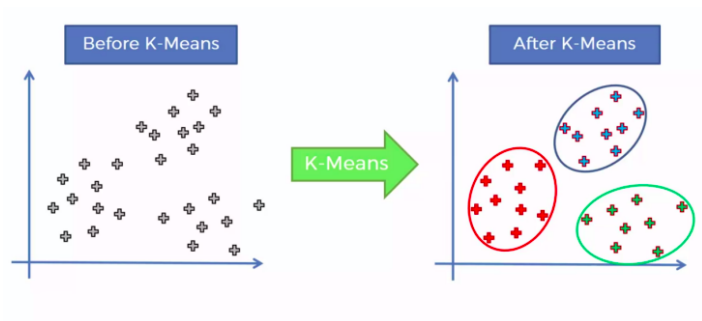
---

# K-means

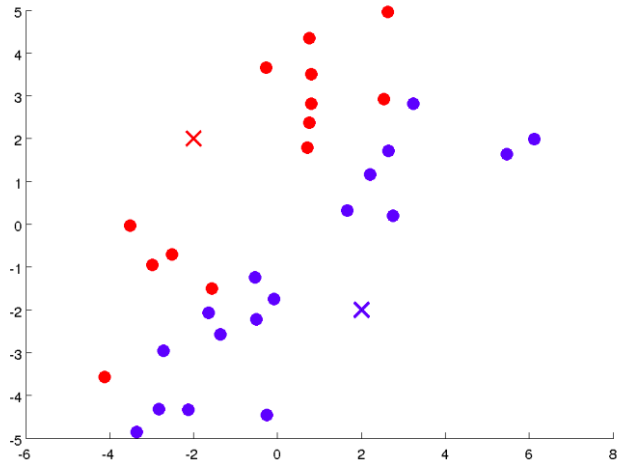
O funcionamento do K-Médias ocorre de forma iterativa, seguindo estas etapas principais:

- 1 **Inicialização:** São selecionados aleatoriamente  $k$  centroides iniciais, que podem ser objetos da base de dados ou pontos calculados.
- 2 **Atribuição de grupos:** Cada objeto da base é associado ao *cluster* cujo centroide esteja mais próximo, com base em uma medida de distância, como a Euclidiana.
- 3 **Recalcular os centroides:** Para cada *cluster*, calcula-se um novo centroide, baseado na média dos objetos atribuídos ao grupo.
- 4 **Convergência:** As etapas de atribuição e recalculação são repetidas até que os centroides não se movam mais ou até atingir o número máximo de iterações.

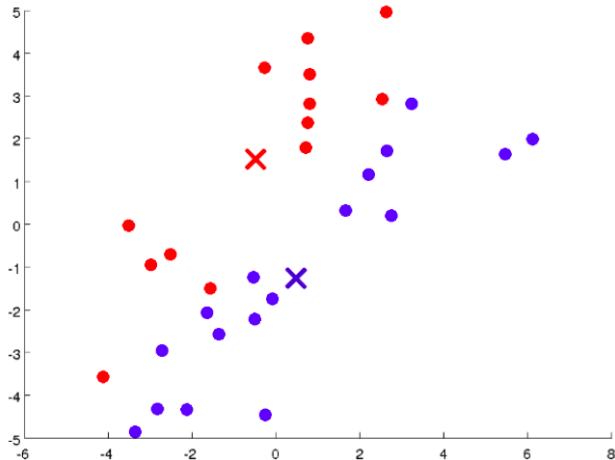
# K-means



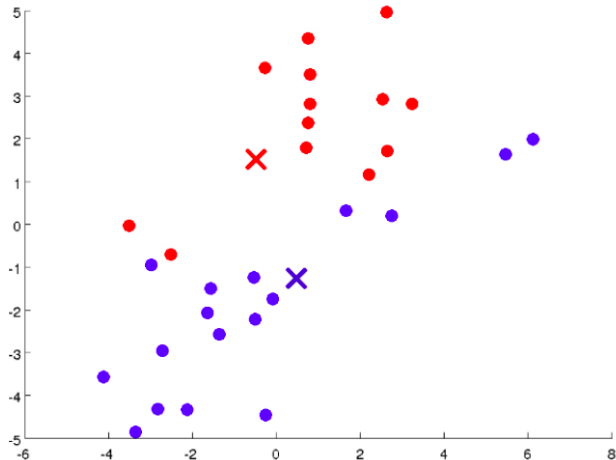
# K-means



# K-means

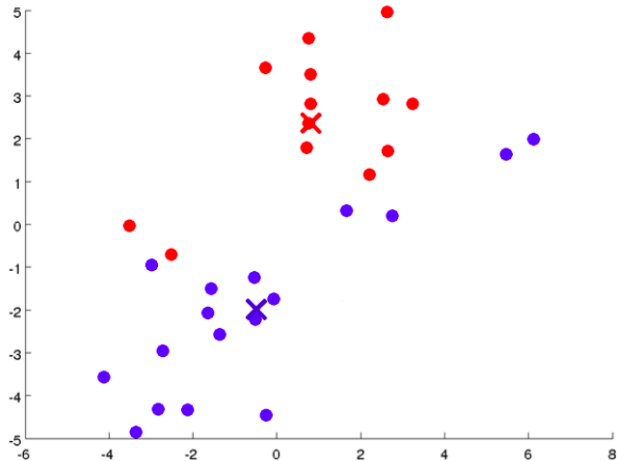


# K-means

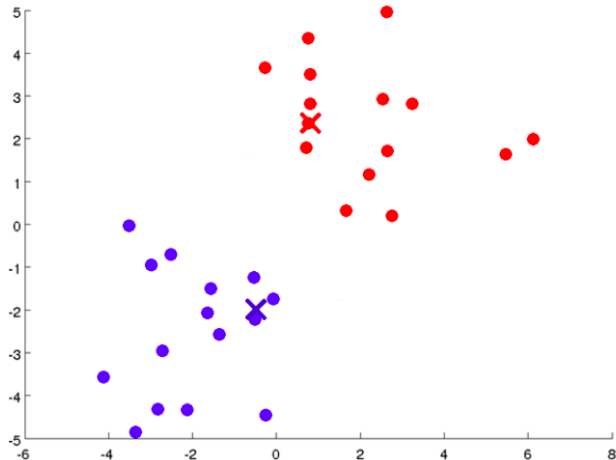




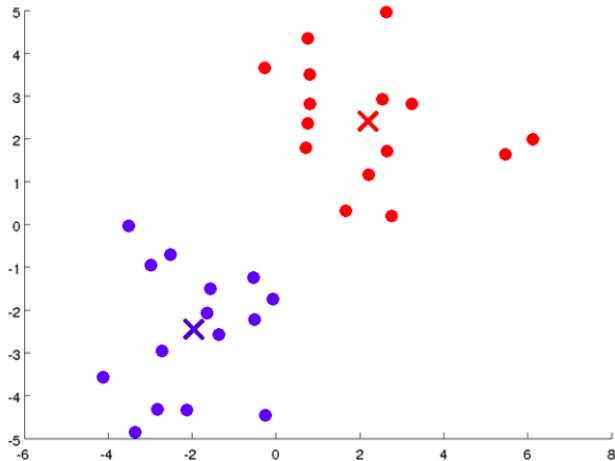
# K-means



# K-means

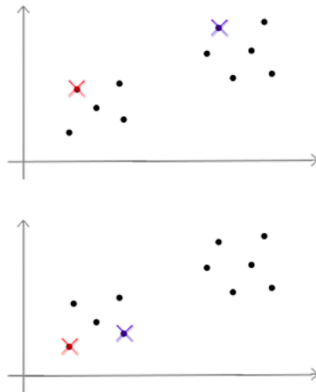


# K-means

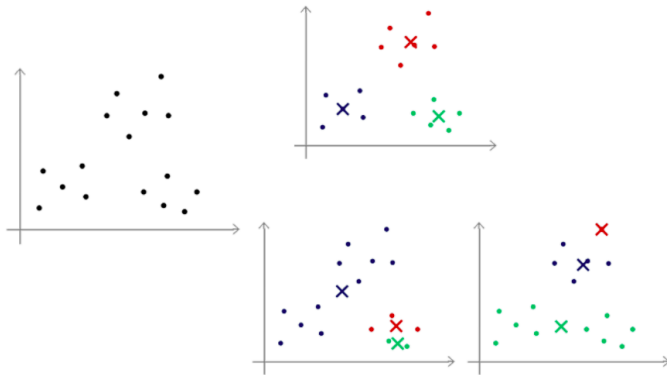


# Inicialização aleatória

- ▶ Escolha  $K < m$
- ▶ Busque aleatoriamente  $K$  exemplos de treinamento
- ▶ Atribua  $\mu_1, \dots, \mu_K$  igual aos  $K$  exemplos.



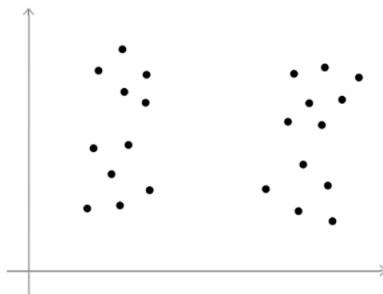
# Ótimo local



# Como escolher o melhor valor de K?

- Não há uma forma ideal. Geralmente isso é feito manualmente;
- Normalmente há ambiguidade na definição da quantidade de clusters;

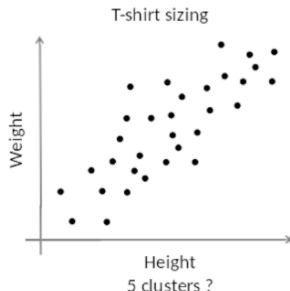
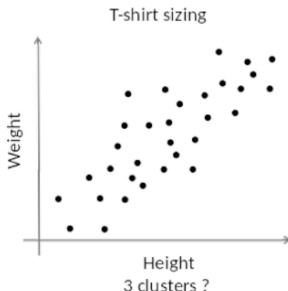
Neste conjunto de dados existem 2 ou 4 clusters?



# A escolha pode depender do problema

- O problema a ser resolvido pode conter uma métrica de avaliação capaz de avaliar o clustering?
- É importante pensar pelo ponto de vista da aplicação.
- Dimensionando camisas: 3 clusters para diminuir os custos ou 5 clusters para aumentar a satisfação do cliente?

Exemplo: Segmentação de mercado:



# Métricas de Validação de Clusterização muito utilizadas com o K-means

Algumas métricas ajudam a validar a qualidade do agrupamento:

- **Silhouette Score**

- Mede a separação entre clusters:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

- $a(i)$ : Média da distância do ponto aos outros do mesmo cluster.
- $b(i)$ : Distância mínima do ponto a um cluster vizinho.

- **Índice de Davies-Bouldin**

- Mede a compactação e separação dos clusters.
- Quanto menor, melhor.

- **Método do Cotovelo**

- Avalia a soma dos erros dentro dos clusters para diferentes valores de  $K$ .
- O ponto onde a redução do erro desacelera sugere um bom valor de  $K$ .



# K-means

Bora brincar agora?? Partiu Orange


# Obrigado!


Jane Thais Soares de Oliveira  
Email: [jane.oliveira@ufvjm.edu.br](mailto:jane.oliveira@ufvjm.edu.br)





**UFVJM**


**Universidade Federal dos Vales do  
Jequitinhonha e Mucuri**


 CAMPELLO, R. J. A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, Elsevier, v. 28, n. 7, p. 833–841, 2007.


 CARMICHAEL, J.; JULIUS, R. Finding natural clusters. *Systematic Biology*, Society of Systematic Zoology, v. 17, n. 2, p. 144–150, 1968.


 CASTRO, L. N. D.; FERRARI, D. G. *Introdução à mineração de dados*. [S.l.]: Saraiva Educação SA, 2017.

 FACELI, K. *Um framework para análise de agrupamento baseado na combinação multi-objetivo de algoritmos de agrupamento*. Tese (Doutorado) — Universidade de São Paulo, 2006.


 JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988.

 LAURETTO, M. *Análise de Agrupamentos (Clusters)*. 2017. Data de Acesso: 02 de Dezembro de 2024. Disponível em: <<https://www.each.usp.br/lauretto/cursoR2017/04-AnaliseCluster.pdf>>.

 LIU, C. L. et al. *Introduction to combinatorial mathematics*. [S.l.]: McGraw-Hill New York, 1968. v. 181.

 NALDI, M. C. *Técnicas de combinação para agrupamento centralizado e distribuído de dados*. Tese (Doutorado) — Universidade de São Paulo, 2011.

 PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters. *Pattern recognition*, Elsevier, v. 37, n. 3, p. 487–501, 2004.

 ROUSSEAU, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.