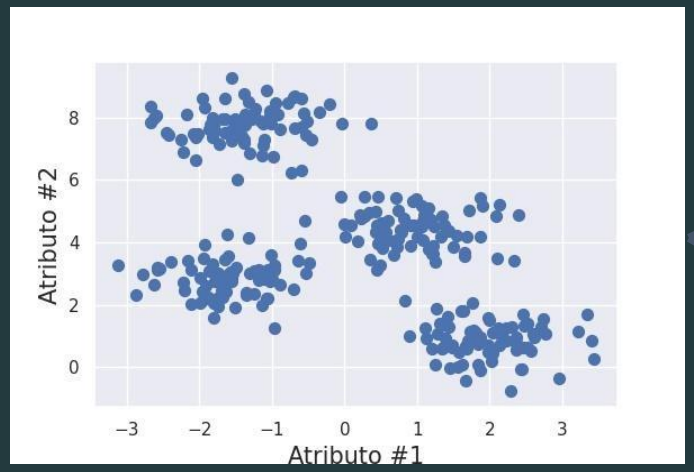
APRENDIZADO NÃO SUPERVISIONADO

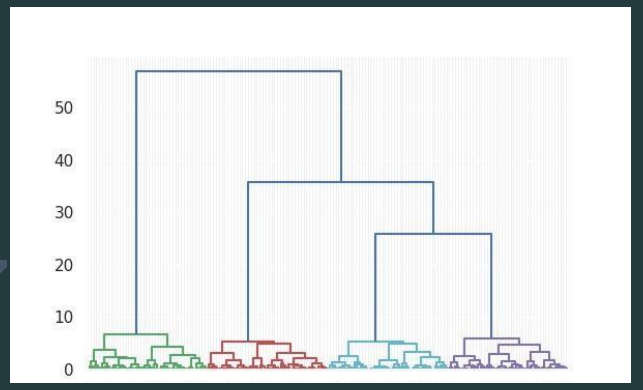
INTRODUÇÃO AOS MÉTODOS DE AGRUPAMENTO

PROFESSOR: MAURÍCIO SOBRINHO

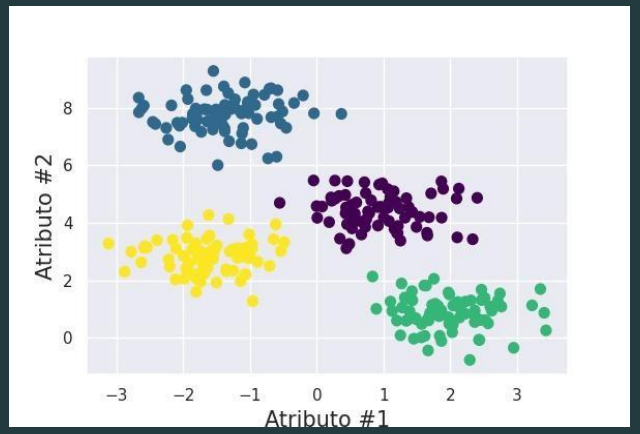
MÉTODOS PARA AGRUPAMENTO DE DADOS

- **Particionais:** organizar dados em uma partição de *k clusters*
- **Hierárquicos:** organizar dados em uma decomposição hierárquica de *clusters* e *subclusters*





Agrupamento Hierárquico

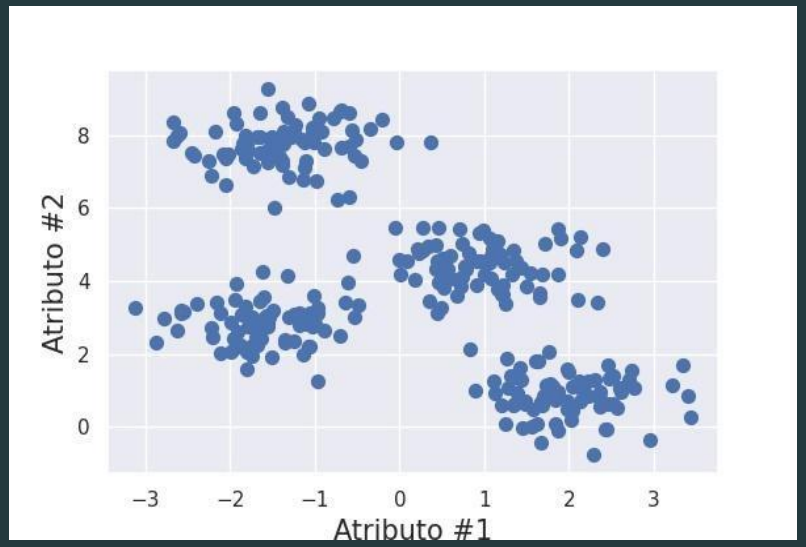


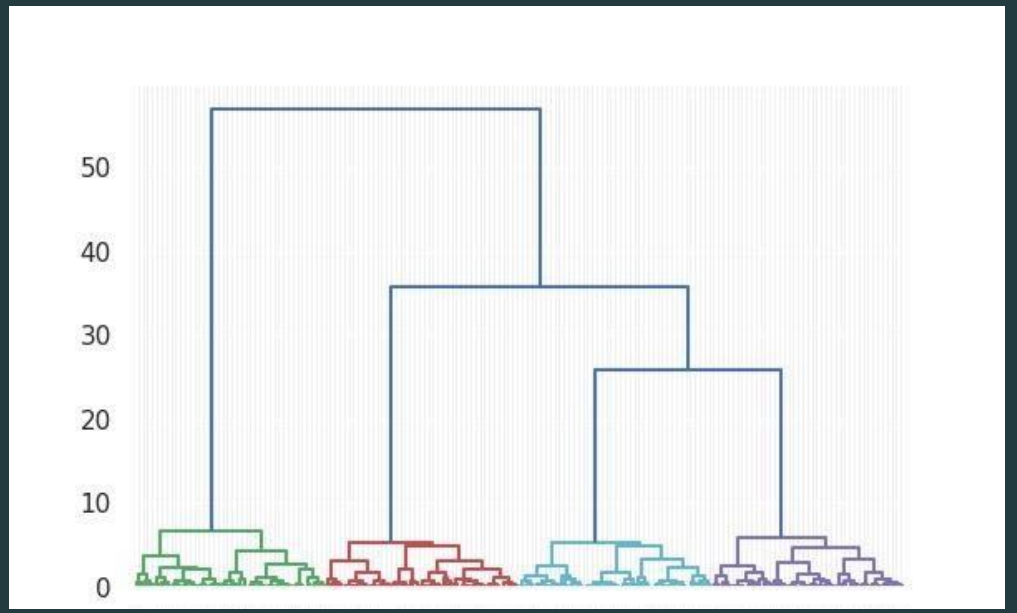
Agrupamento Particional

AGRUPAMENTO HIERÁRQUICO

- **Dendrograma:** diagrama com a estrutura hierárquica que representa o resultado de um agrupamento. Sumariza a formação dos *clusters* e *subclusters*.

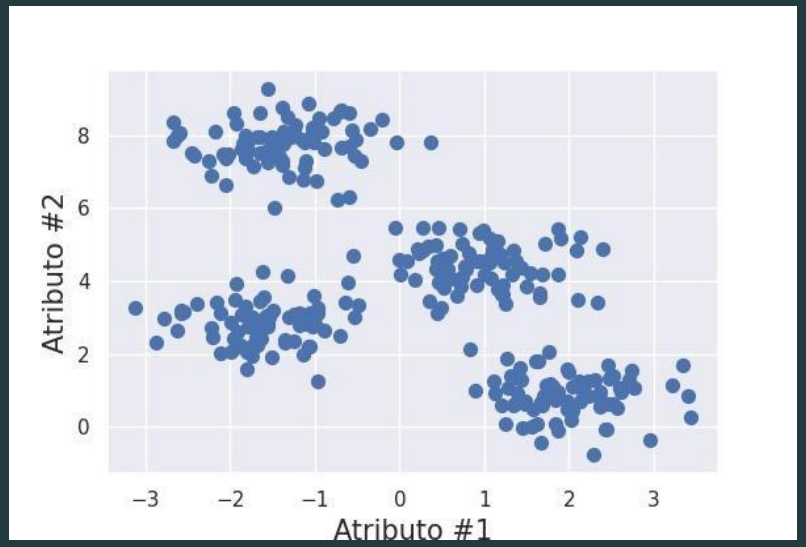
- Os objetos do conjunto de dados estão organizados no eixo x do dendrograma. A altura dos arcos indica a dissimilaridade entre objetos e grupos de objetos.

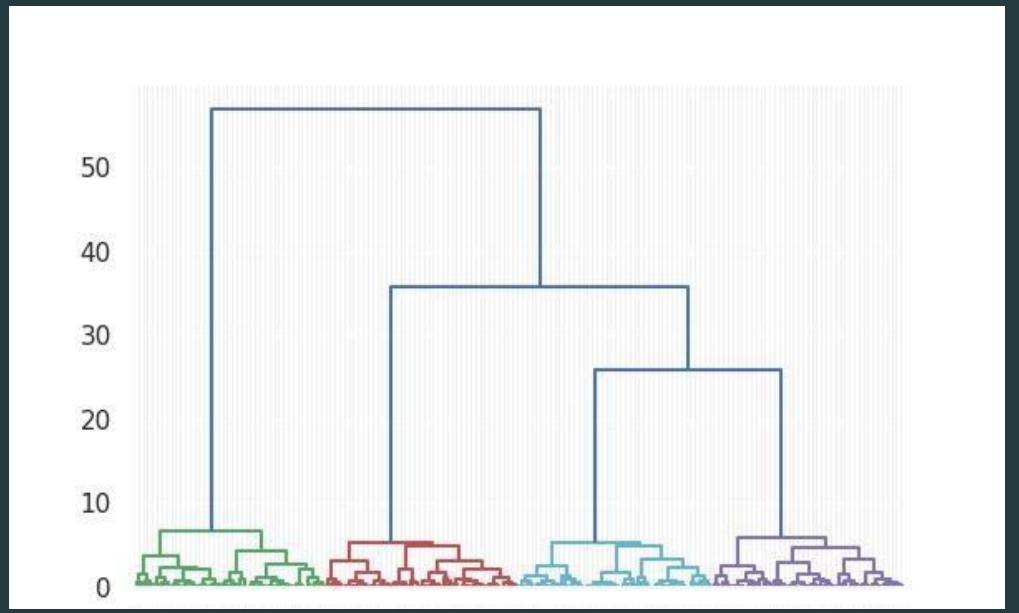




AGRUPAMENTO HIERÁRQUICO

- Podemos inspecionar o dendrograma para estimar o número natural de clusters. No exemplo, há 4 subárvores bem separadas.
- Conceitos de homogeneidade (coesão interna) e heterogeneidade (separabilidade) dos clusters representados pela altura (eixo y) da união entre cluster.





AGRUPAMENTO HIERÁRQUICO

- Dois métodos clássicos para agrupamento hierárquico

Aglomerativos:

➔ Iniciar alocando cada objeto em um *cluster*

➔ Encontrar o melhor par de *clusters* para unir

➔ Repetir até formar um único *cluster*

Divisivos:

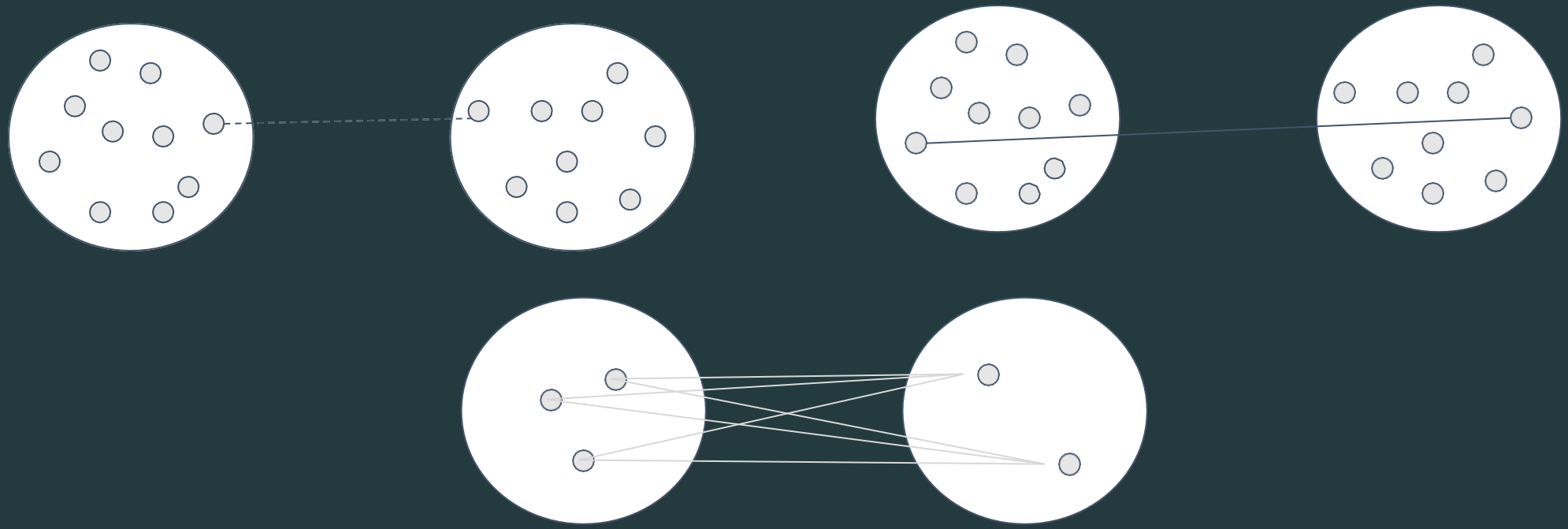
➔ Iniciar alocando todos os objetos em um único *cluster*

➔ Dividir um *cluster* em dois *subclusters*

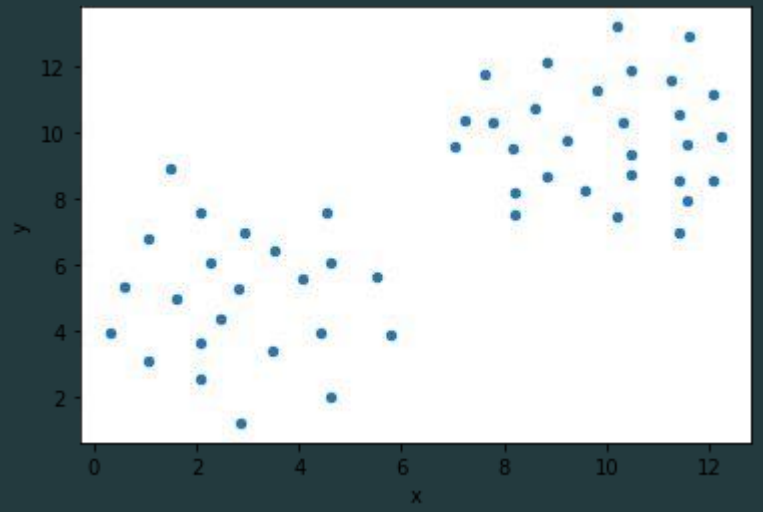
➔ Repetir a divisão até que cada objeto seja um *cluster*

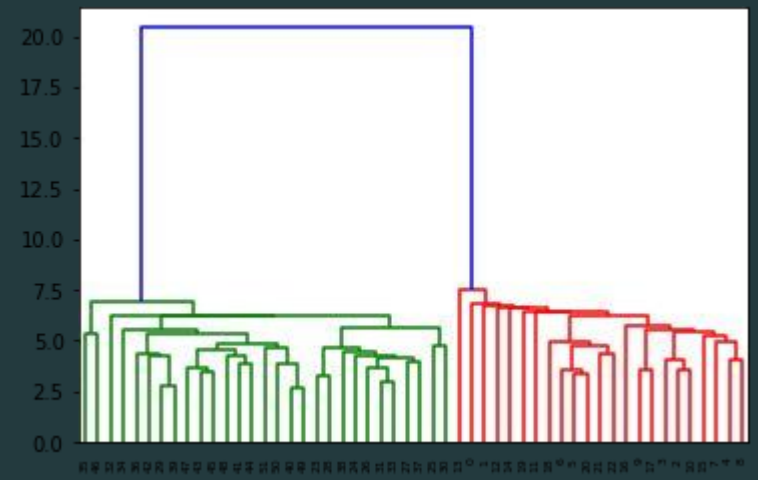
AGRUPAMENTO HIERÁRQUICO

- Single-Link (Min)
- Complete-Link (Max)
- Average-Link (Média)

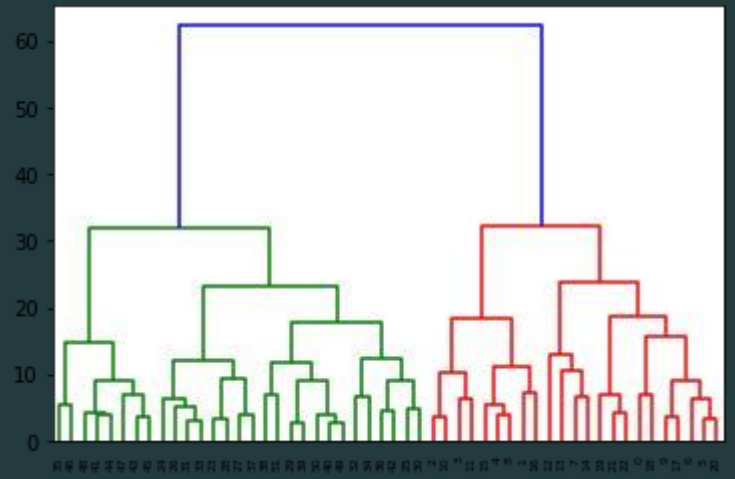


AGRUPAMENTO HIERÁRQUICO

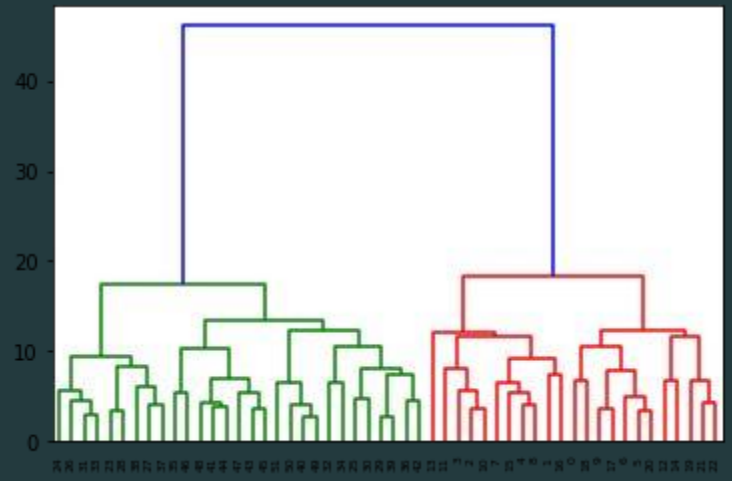




Single-Link



Complete-Link



Average-Link

AGRUPAMENTO PARTICIONAL

- Falaremos sobre métodos de agrupamento para obter partições rígidas dos dados
- **Partição rígida:** clusters não possuem sobreposição
  - Dado um conjunto de ***n*** objetos

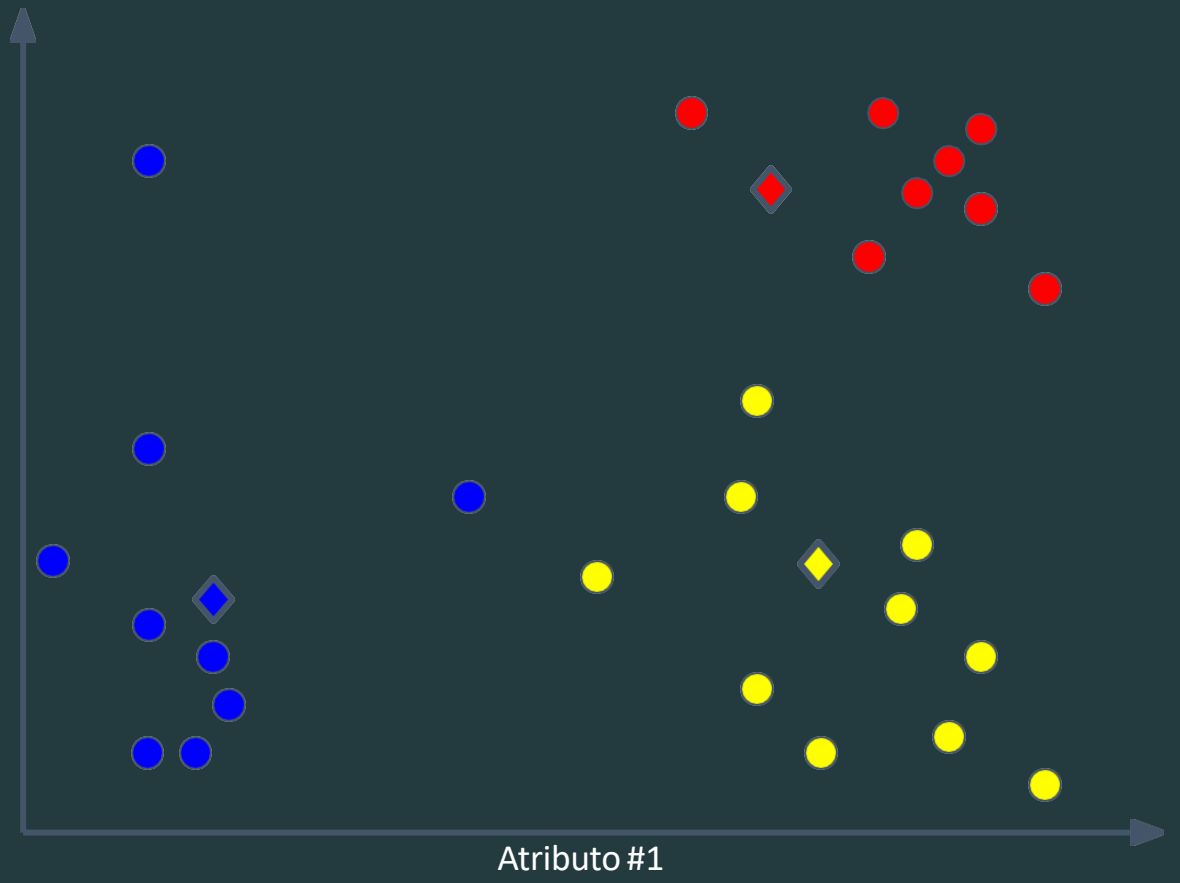
ALGORITMO *K*-MÉDIAS OU *K-MEANS*

- Amplamente usado na indústria e academia
- Características desejáveis para Mineração de Dados
  - Simplicidade
  - Interpretabilidade
  - Eficiência Computacional

ALGORITMO *K-MEANS*

Algoritmo:

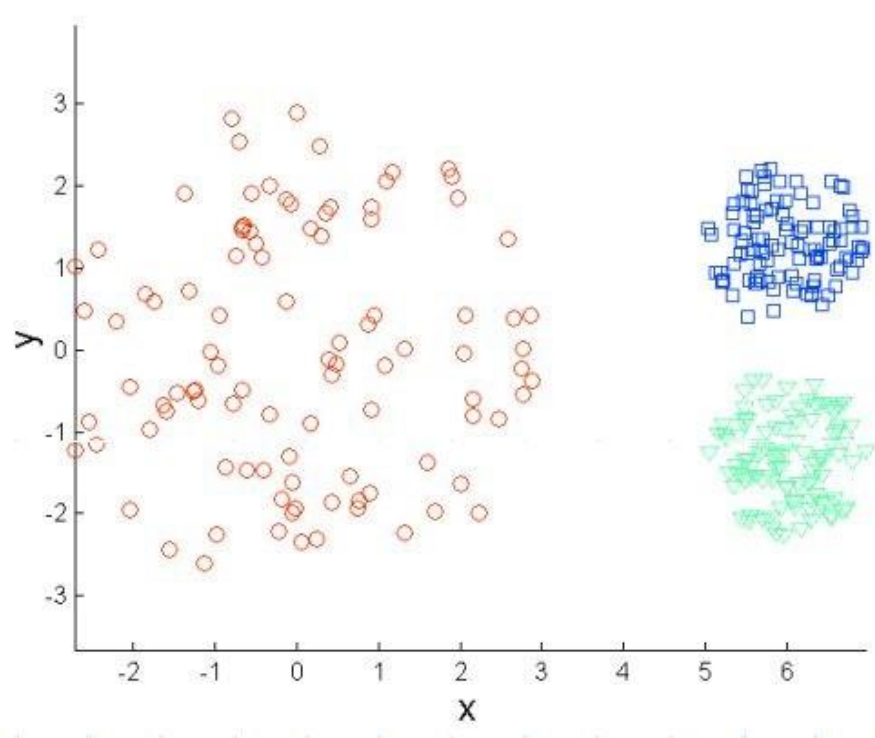
1. Selecionar k centroides iniciais
2. Repetir até convergir:
   1. Formar k clusters atribuindo cada objeto ao centroide mais próximo
   2. Atualizar o centroide de cada cluster



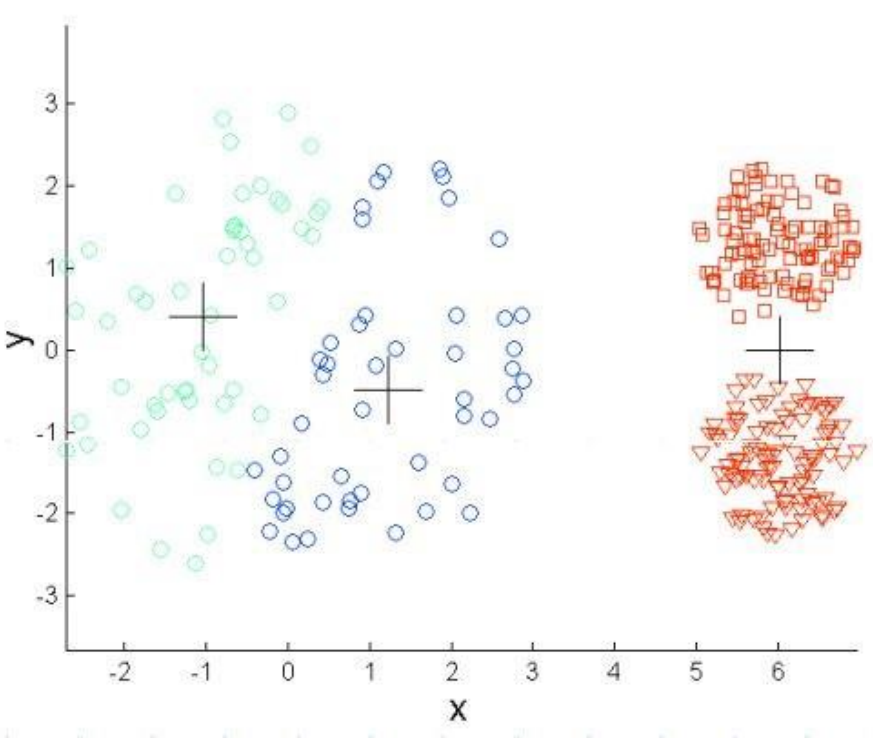
Critérios de convergência: (1) poucas mudanças nos clusters/centroides; (2) número máximo de iterações.

ALGORITMO *K-MEANS*

- **Limitações do *k-Means:***
  - Clusters de densidades muito diferentes



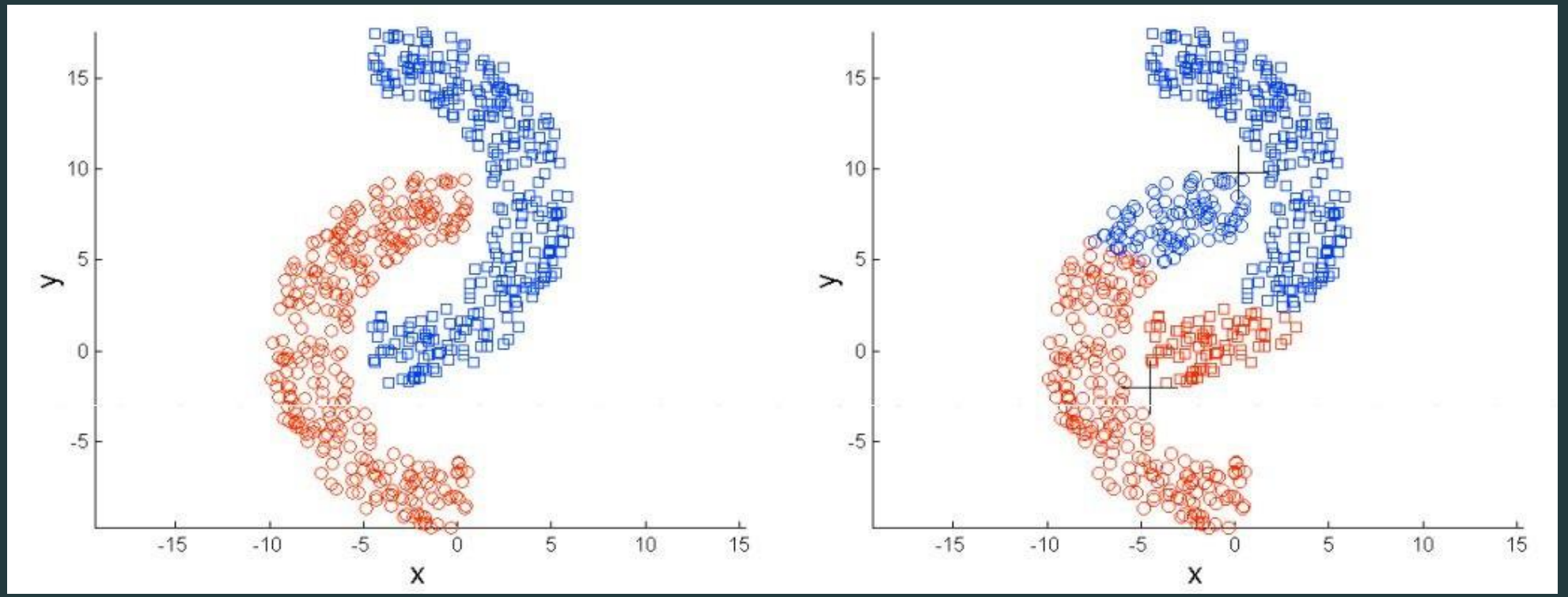
Clusters esperados



Clusters obtido

ALGORITMO *K-MEANS*

- **Limitações do *k-Means:***
  - Clusters de formatos não globulares

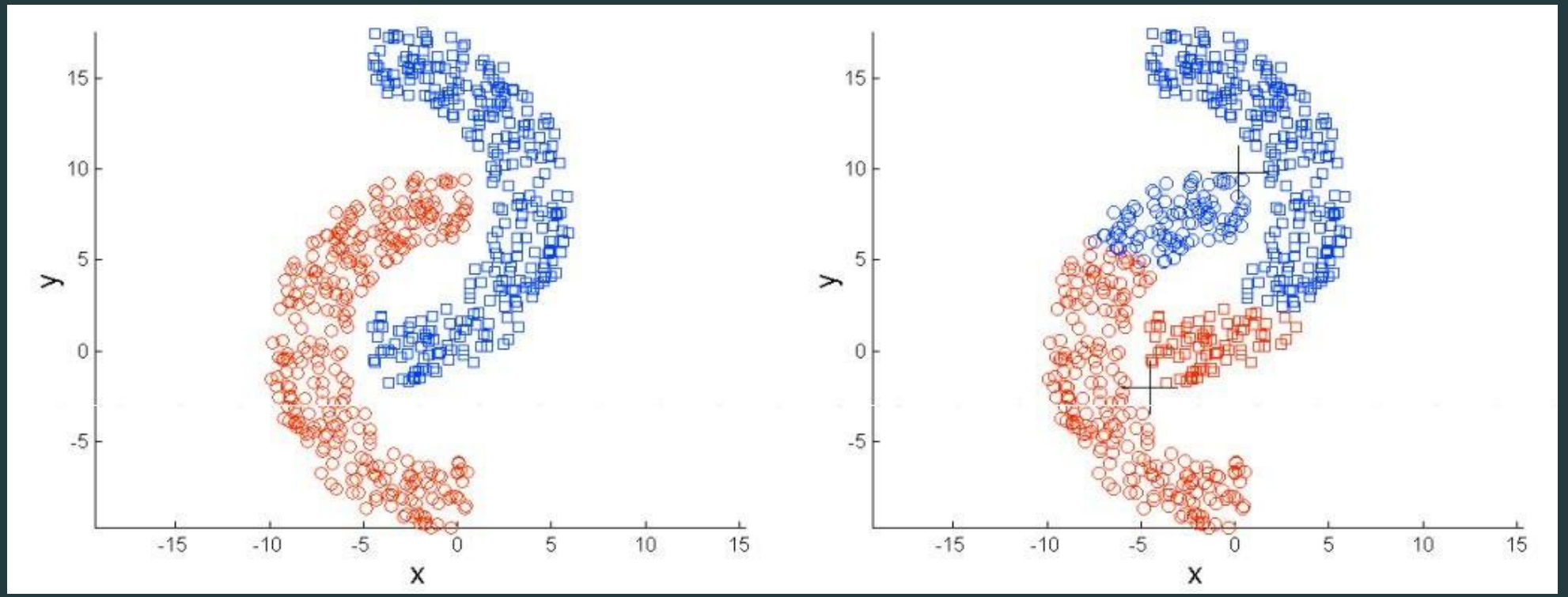


Clusters esperados

Clusters obtido

ALGORITMO *K-MEANS*

- **Limitações do *k-Means:***
  - Clusters de formatos não globulares



Clusters esperados

Clusters obtido

BIBLIOGRAFIA

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.