LABORATÓRIO MÍDIACOM LabGen

Universidade Federal Fluminense

Algoritmos de Agrupamento

TET00343 - Tópicos Especiais em Engenharia de Telecomunicações I

Prof. Diogo Mattos

Nicollas Rodrigues (Estágio em Docência)

Departamento de Engenharia de Telecomunicações – TET/TCE/UFF

Instituto de Computação – IC/UFF

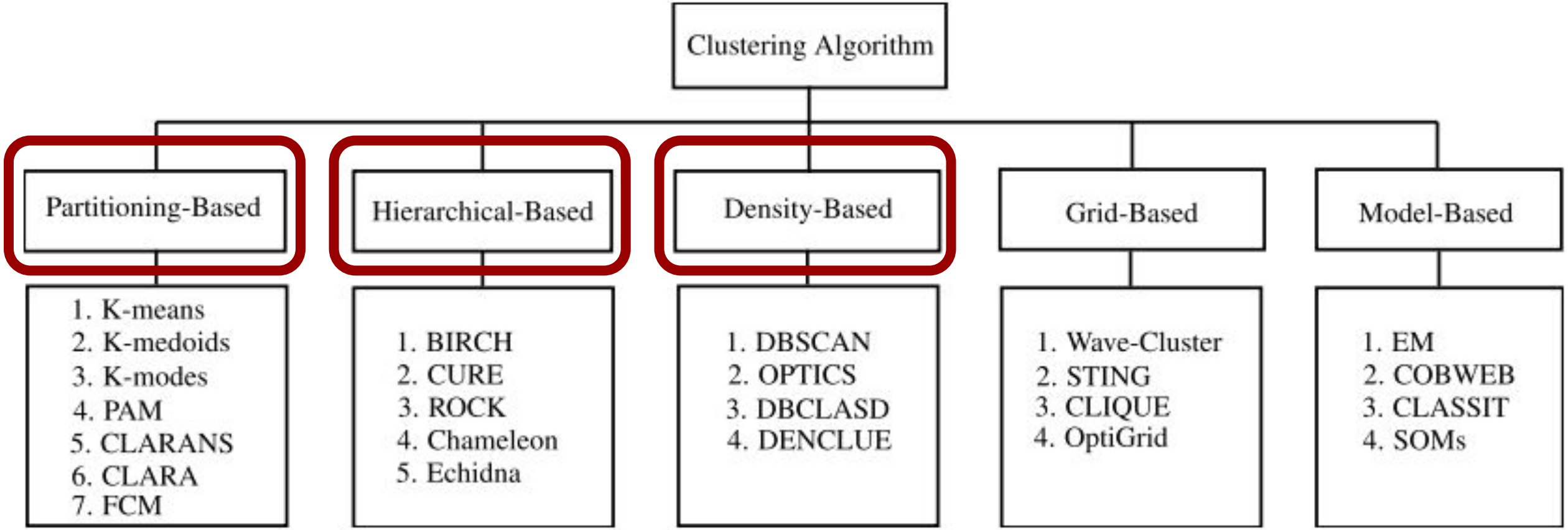
Universidade Federal Fluminense

Aprendizado não-Supervisionado

- **Principal Característica**
  - **Independe de qualquer marcação** sobre os dados (*ground truth*)

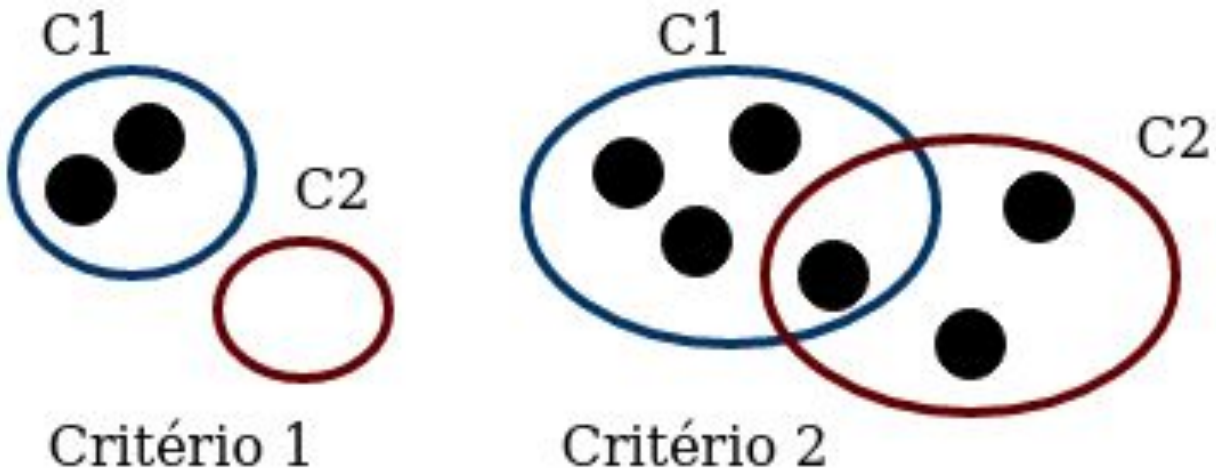
- **Algoritmos de Agrupamento** (Clusterização)
  - Principal exemplo de algoritmos não supervisionados
  - Identifica padrões entre as entradas e **agrupa aquelas similares entre si**
    - Grupos → agrupamentos (*clusters)*
  - Diversos algoritmos com diferentes…
    - lógicas operacionais
    - casos de uso
    - escalabilidades
    - desempenhos

Algoritmos de Agrupamento
Subdivisões



Algoritmos de Agrupamento Baseados no Particionamento

- Classificação é dada àqueles **algoritmos cumprem simultaneamente dois critérios**
  - **Primeiro Critério** → Obrigatoriedade de ter **pelo menos uma amostra em cada agrupamento** criado
  - **Segundo Critério** → **Exclusividade de pertencimento**, ou seja, cada amostra deve pertencer a somente um agrupamento
- Exemplos Mais Comuns
  - **K-means**
  - **K-medoids**

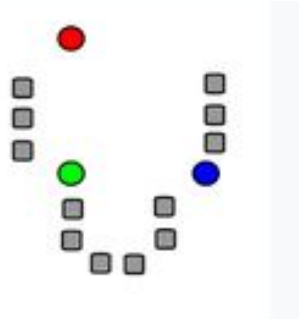


Algoritmos Baseados no Particionamento
K-means

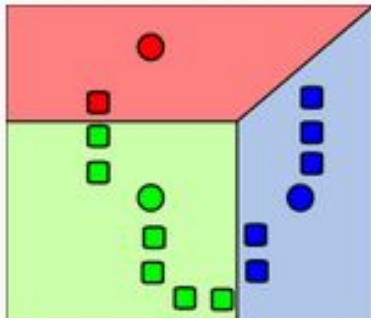
- **Heurística clássica** proposta por MacQueen (1967) e Lloyd (1957/82)
  - **Ideia Geral:**
    - Particionar dados de entrada em *k* agrupamentos pela minimização da soma dos quadrados (SSE) das distâncias em cada agrupamento

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

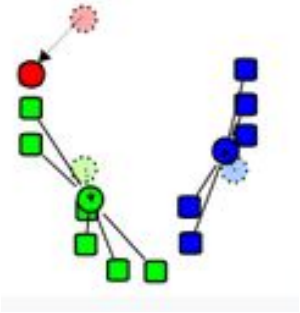
- **Lógica de Execução**
  - **Parâmetros iniciais:** *k* → arbitrariamente escolhido pelo usuário



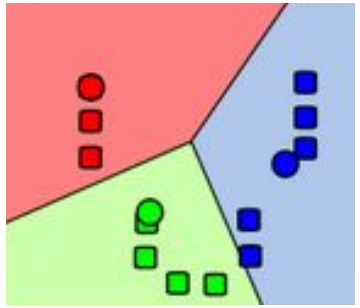
Etapa 1: Escolha aleatória dos centróides (ponto médio) de cada um dos *k* agrupamento



Etapa 2: Cálculo da distância entre cada amostra e os centróides e alocação no agrupamento cujo centróide está mais próximo



Etapa 3: A cada iteração dos centroídes são recalculados com base na média espacial



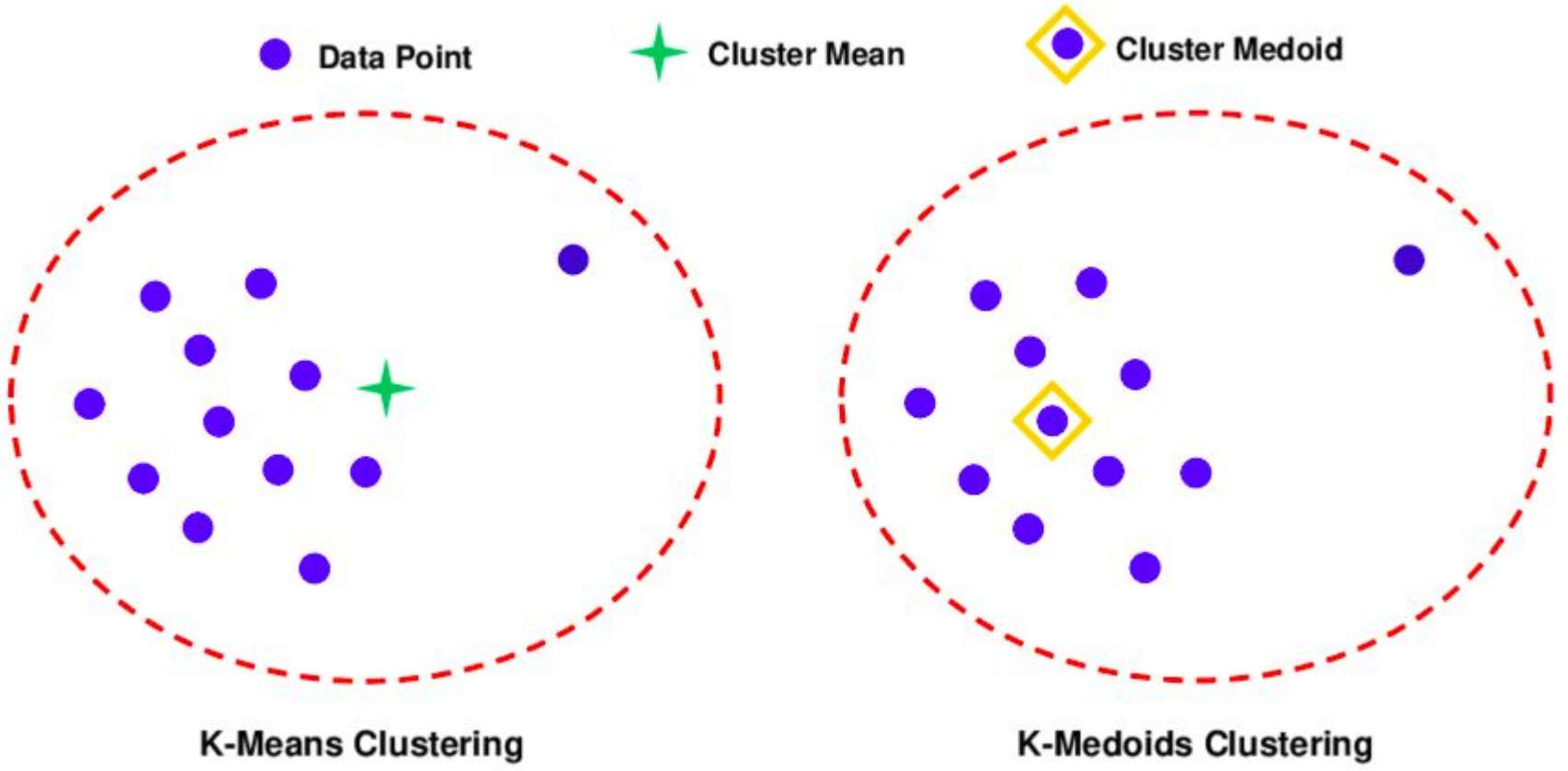
Etapa 4: O algoritmo finaliza quando cessam as alterações na alocação de amostras

Algoritmos Baseados no Particionamento
K-means

- **Vantagens:**
  - Muito popular, simples e rápido
  - Baixa complexidade O(*nkt*) da versão *standard* do algoritmo
    - *n* → número de amostras
    - k → número de agrupamentos arbitrariamente escolhidos
    - t → número de iterações até a convergência (pode ser escolhida)
- **Desvantagens:**
  - Indeterminação quanto ao número adequado de *k*
    - Determinar o valor ótimo de *k*
  - Desempenho altamente dependente das coordenadas iniciais → centróides iniciais
  - Sensibilidade à amostras anômalas e *outliers*
    - Amostras muito distantes influenciam o cálculo do centróide de cada iteração
  - Clusters exclusivamente esféricos

Algoritmos Baseados no Particionamento
K-medoids

- **Variante do *k-means***
  - **Diferença:** Uso de amostras de entrada (reais) como o centro dos agrupamentos (centróide), ao invés de pontos médios (espaciais)



Algoritmos Baseados no Particionamento
K-medoids

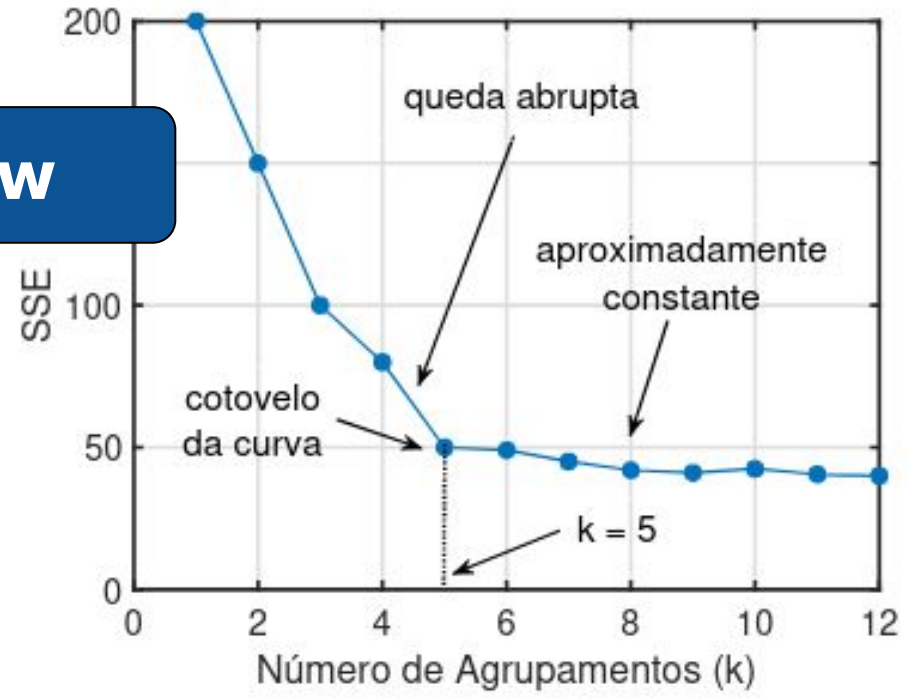
- **Similar ao k-means**
  - **Diferença:** Uso de amostras de entrada reais como o centro dos agrupamentos (centróide), ao invés de pontos médios (espacial)
- **Vantagens** (em relação ao *K-means*):
  - Maior robustez a dados ruidosos e *outliers*
  - Capacidade de lidar com alta dimensionalidade
    - Dimensionalidade → Tamanho do vetor de cada amostra ≠ Número de amostras
  - Facilidade de interpretação
    - As saídas do algoritmo são amostras reais
- **Desvantagens**
  - Indicado para pequenos conjuntos de dados
    - Quantidade limitada de amostras
  - Igualmente dependente do número de *k*

Algoritmos Baseados em Particionamento
Valor Ótimo de *k*

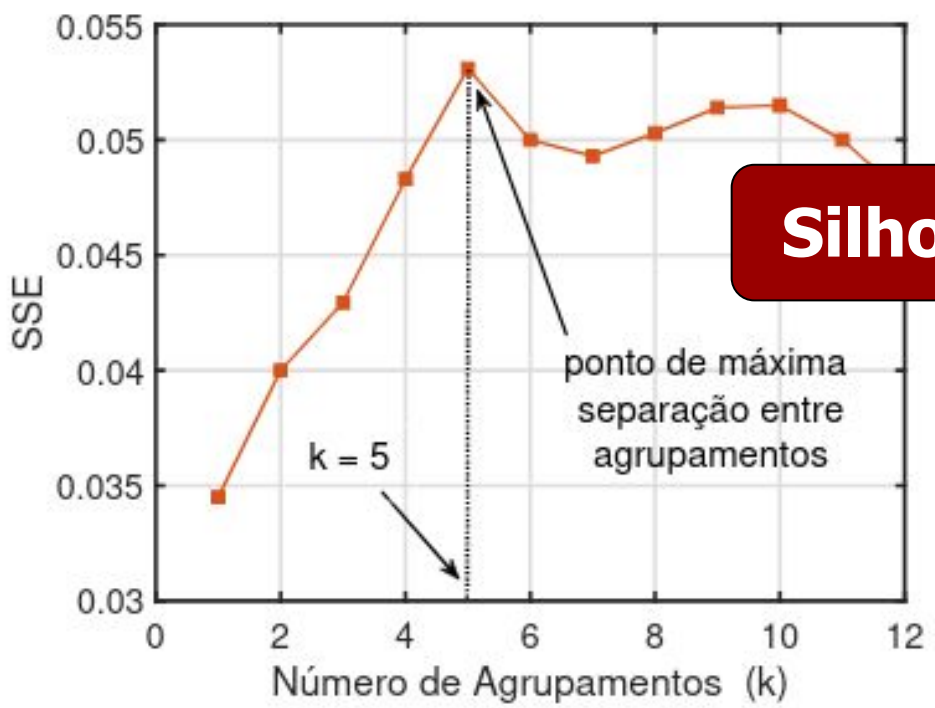
LABORATÓRIO MÍDIACOM LabGen

- **Desvantagem singular:**
  - Indeterminação quanto ao número ótimo de agrupamento *k* → impacta no desempenho
  - **Solução:** Análise gráfica prévia do dados utilizando **simultaneamente** o...
    - **Método Elbow** → mede a dispersão das amostras dentro dos clusters (compactação)
    - **Método Silhouette** → mede a qualidade dos clusters (separação)

Elbow

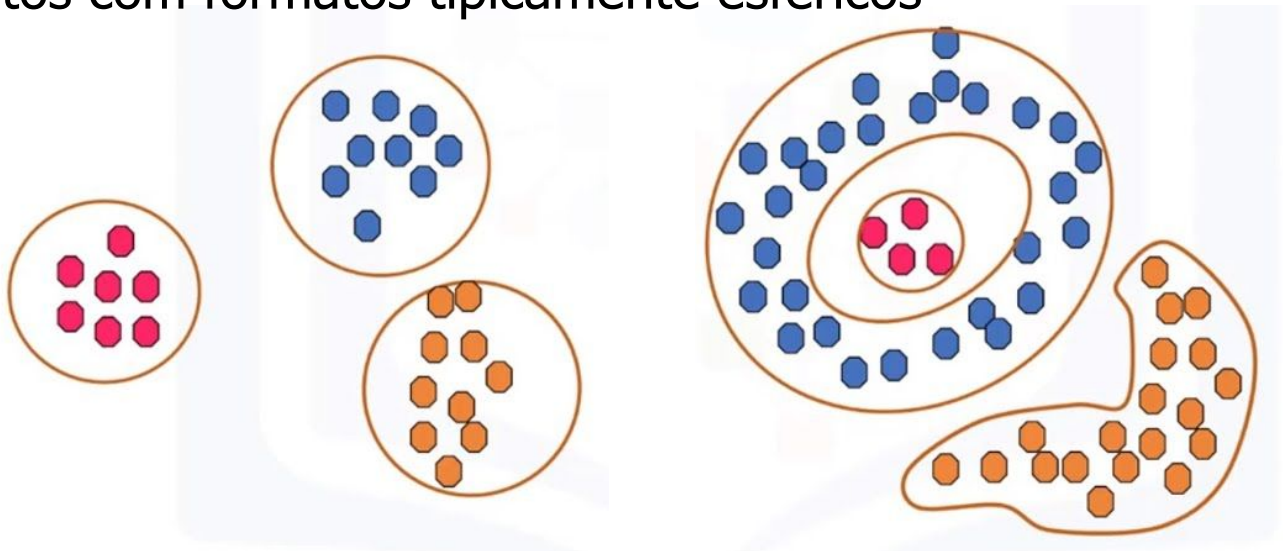


Silhouette



Algoritmos de Agrupamento Baseados em Densidade

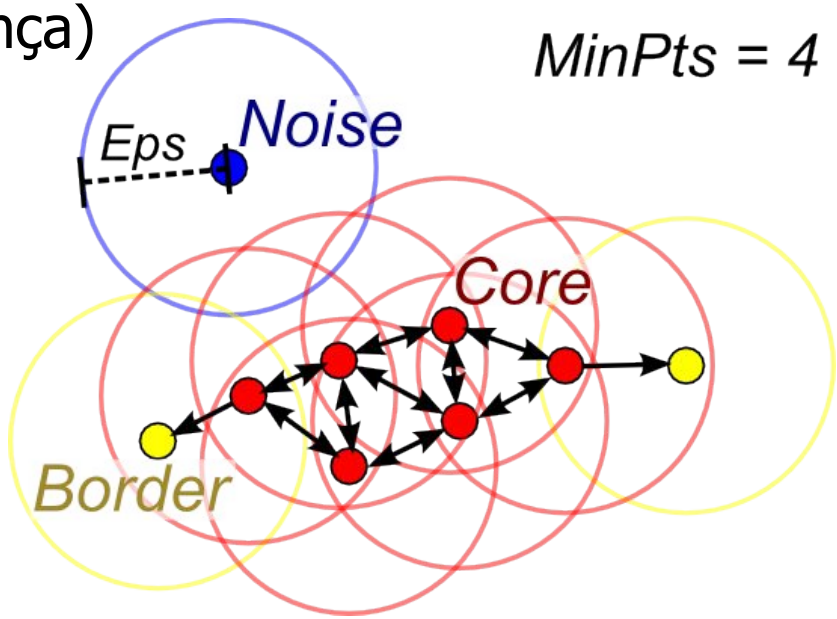
- Classificação dada aos algoritmos que **consideram a abordagem do vizinho mais próximo** (*nearest neighbour*)
  - **Agrupamento** (cluster) → componente denso conectado
    - Crescimento de um agrupamento ocorre em qualquer direção que a densidade o conduza
- **Vantagem** (sobre os algoritmos de particionamento)
  - Independe de uma escolha antecipada do número de agrupamentos
  - Possibilidade de descobrir agrupamentos com formas arbitrárias
    - k-means, k-medoids, etc → agrupamentos com formatos tipicamente esféricos
- **Desvantagem**
  - Alta complexidade
    - Tempo de convergência



Algoritmos Baseados em Densidade
DBSCAN

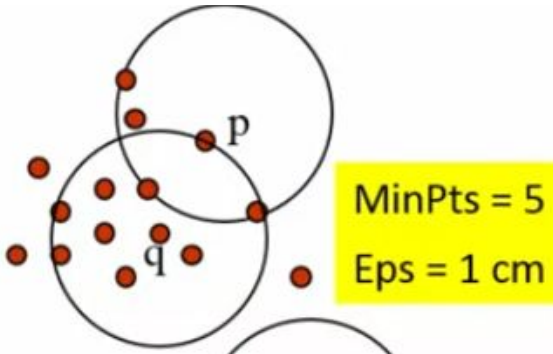
- **DBSCAN** (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído)
  - Introduzido por Ester et. al, 1996
  - Objetivo: Encontrar regiões que...
    - Satisfaçam uma **densidade de pontos mínima** estabelecida + sejam **separadas por regiões de menor densidade**
    - Parâmetros iniciais
      - **ε** → raio de observação (tamanho da vizinhança)
      - **minPts** → número de vizinhos
  - Classificação das amostras (pontos):

Ponto Central (*Core*) → amostra com minPts vizinhos dentro da vizinhança de raio ε (incluindo a própria amostra)
Ponto de Borda (*Border*) → amostra que não satisfaz o minPts, porém é vizinha de um ponto central
Ponto Ruído (*Noise*) → amostra que não satisfaz o minPts e não é vizinha de um ponto central

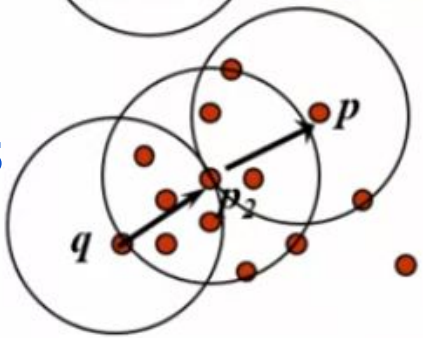


Algoritmo DBSCAN
Conceitos

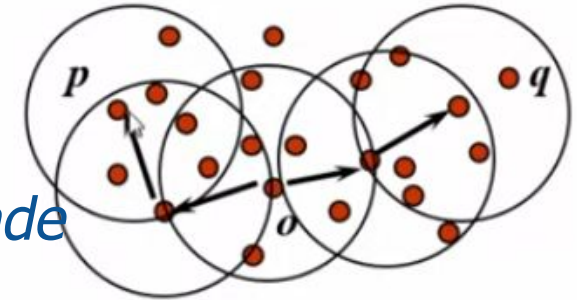
- **Alcançável diretamente por densidade** (*directly density-reachable*)**:**
  - Um ponto ***p*** é alcançável diretamente pelo ponto ***q*** se...
    - *o* ponto ***q*** está dentro da vizinhança ε do ponto central ***p***
  - **Obs:** Pontos só podem ser alcançados diretamente a partir dos pontos centrais



- **Alcançável por densidade** (*density-reachable)*:
  - Um ponto ***p*** é alcançável por densidade pelo ponto ***q*** se...
    - houver um caminho de pontos que sejam diretamente alcançáveis entre si até o ponto *p*
  - Obs: O ponto inicial ***q*** e todos os pontos no caminho devem ser pontos centrais, exceto o ponto ***q***

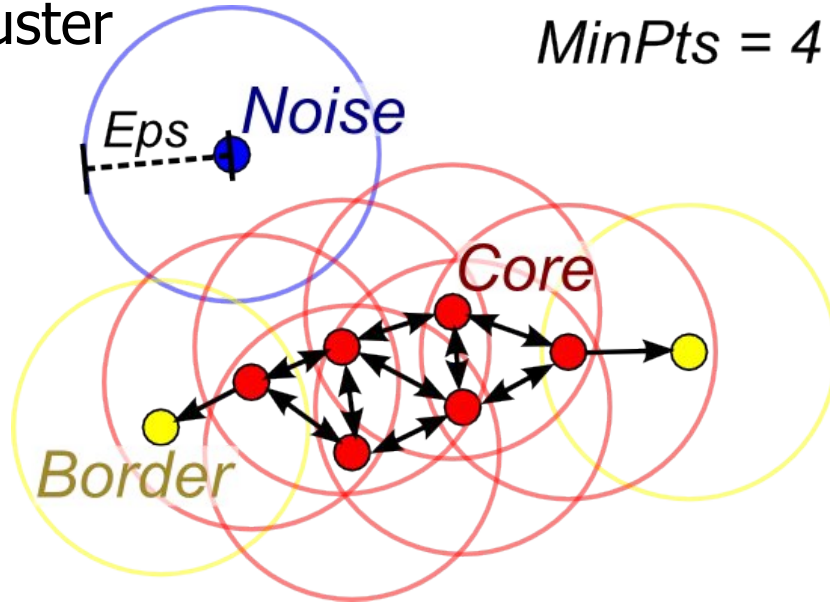


- **Conectado por densidade** (*density-connected*):
  - Um ponto ***p*** é conectado por densidade pelo ponto ***q*** se...
    - houver um ponto ***o*** *tal que* ***p*** *e* ***q*** *sejam alcançáveis por densidade a partir do ponto o*
  - **Obs:** *p* e *q* são pontos de borda



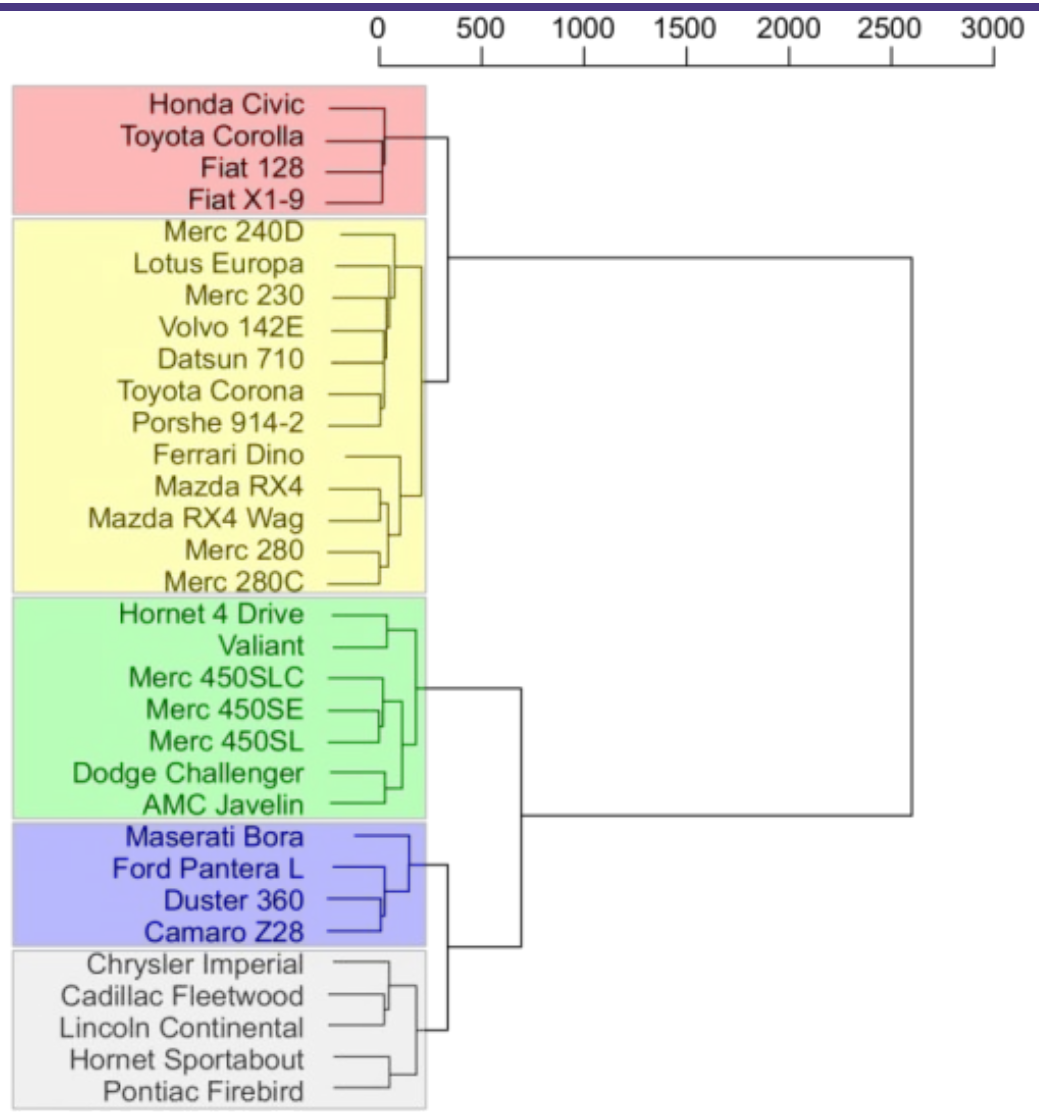
Algoritmos Baseados em Densidade
DBSCAN

- Funcionamento:
  - Encontra os **pontos na vizinhança ε de cada ponto p_i**
    - Identifica os pontos centrais → satisfazendo o **ε** e **minPts**
  - Seleciona um ponto central *p* para **formar um agrupamento (*cluster*) contendo todos os pontos alcançáveis por densidade** a partir *p*
    - inclui pontos de borda e outros pontos centrais
  - Continua o processo **selecionando outros pontos centrais ainda não processados**
    - **pontos centrais ainda não alocados** em algum cluster
  - Havendo pontos não alocados a nenhum cluster → ruído
- **Obs:**
  - Cada cluster contém pelo menos um ponto central



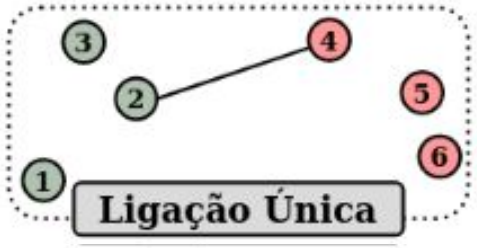
Algoritmos Hierárquicos

- Classificação dada aqueles algoritmos **que criam agrupamentos** e calculam uma **representação hierárquica dos dados de entrada**
- Representação hierárquica → **dendrograma**
  - Tipo particular de árvore binária
    - nós-folhas expressam dados individuais
- Método de construção
  - **Aglomerativo (*bottom-up*)**
  - **Divisivo (top-down)**



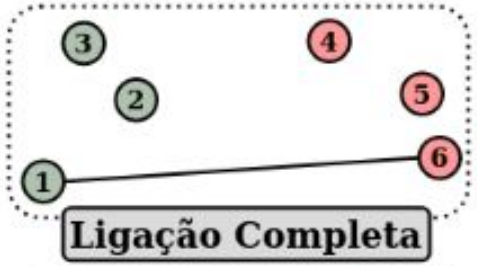
Algoritmos Hierárquicos
Método Algomerativo

- **Aglomerativo (*bottom-up*)**
  - Inicia considerando **cada amostra = agrupamento unitário**
  - Mescla recursivamente duas ou mais amostras em um novo agrupamento
    - Seguindo uma **função de ligação** (*linkage*) escolhida



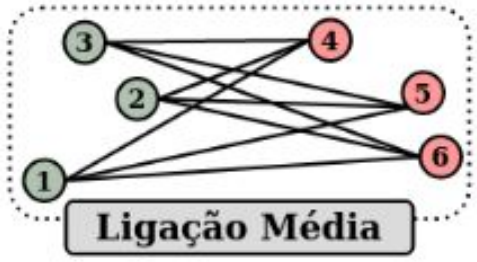
Ligação Única *(Single-linkage)* → estabelece a união considerando a distância entre as amostras mais próximos de cada agrupamento

$$D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$



Ligação Completa *(Complete-linkage)* → estabelece a união considerando a distância das amostras mais distantes entre si de cada agrupamento

$$D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

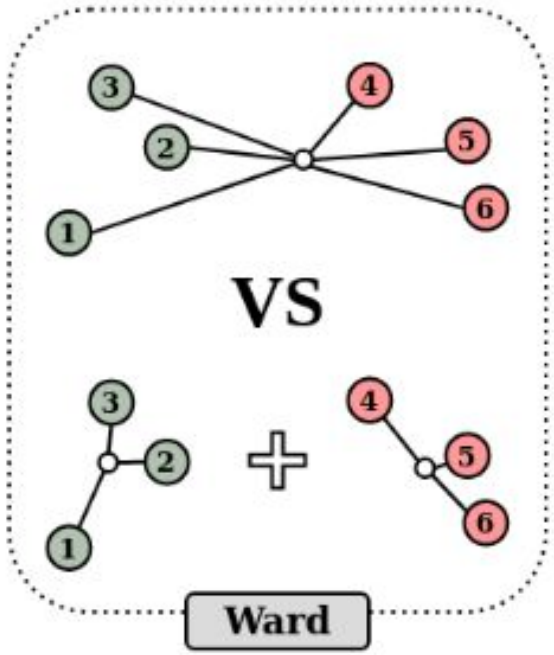


Ligação Completa *(*Average linkage*)* → estabelece a união considerando a média das distâncias de todas as amostras de um agrupamento em relação a todas as amostras d outro agrupamento.

$$D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$$

Algoritmos Hierárquicos
Método Algomerativo

- **Aglomerativo (*bottom-up*)**
  - Inicia considerando **cada amostra = agrupamento unitário**
  - Mescla recursivamente duas ou mais amostras em um novo agrupamento
    - Seguindo uma **função de ligação** (*linkage*) escolhida



Ward Linkage → considera a distância euclidiana na descoberta do par de agrupamentos que minimizam o aumento na variância total interna após a união

$$TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$$

Considerando $C_1 \rightarrow \sigma_1^2$ $C_2 \rightarrow \sigma_2^2$

Ao mesclar dois clusters em um terceiro $C_3 = C_1 + C_2$

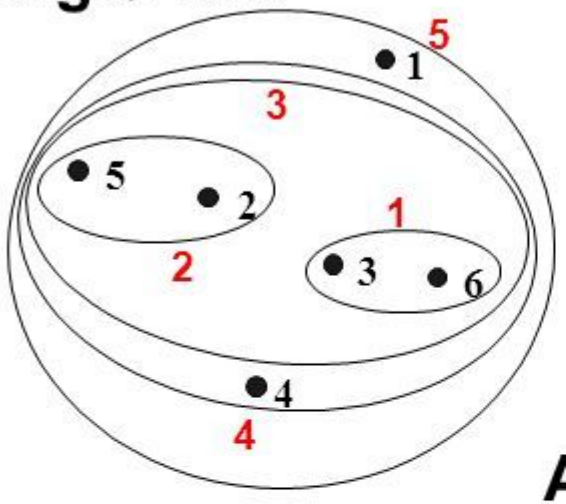
$C_3 \rightarrow \sigma_3^2 > \sigma_2^2$ e $\sigma_3^2 > \sigma_1^2$

"O quanto vai aumentar?"

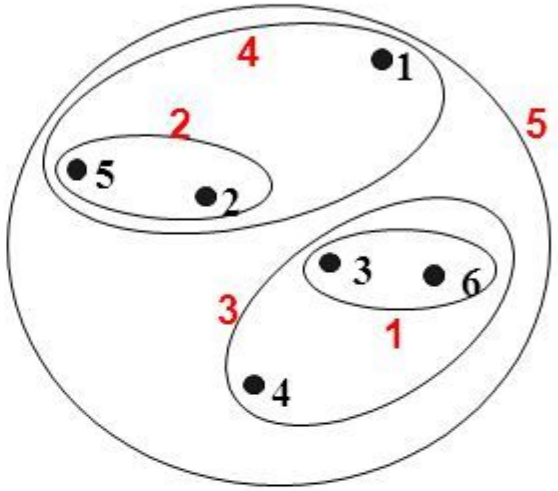
Algoritmos Hierárquicos
Função de Ligação x Impacto nos Resultados

LABORATÓRIO MÍDIACOM
LabGen

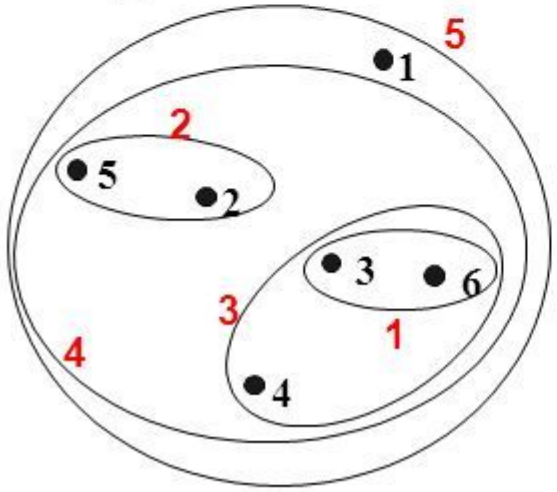
Single-link



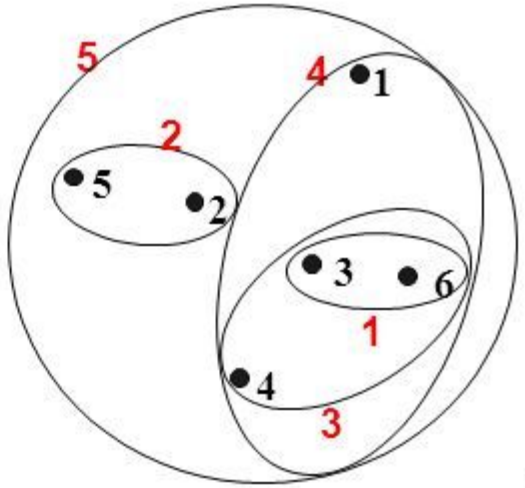
Complete-link



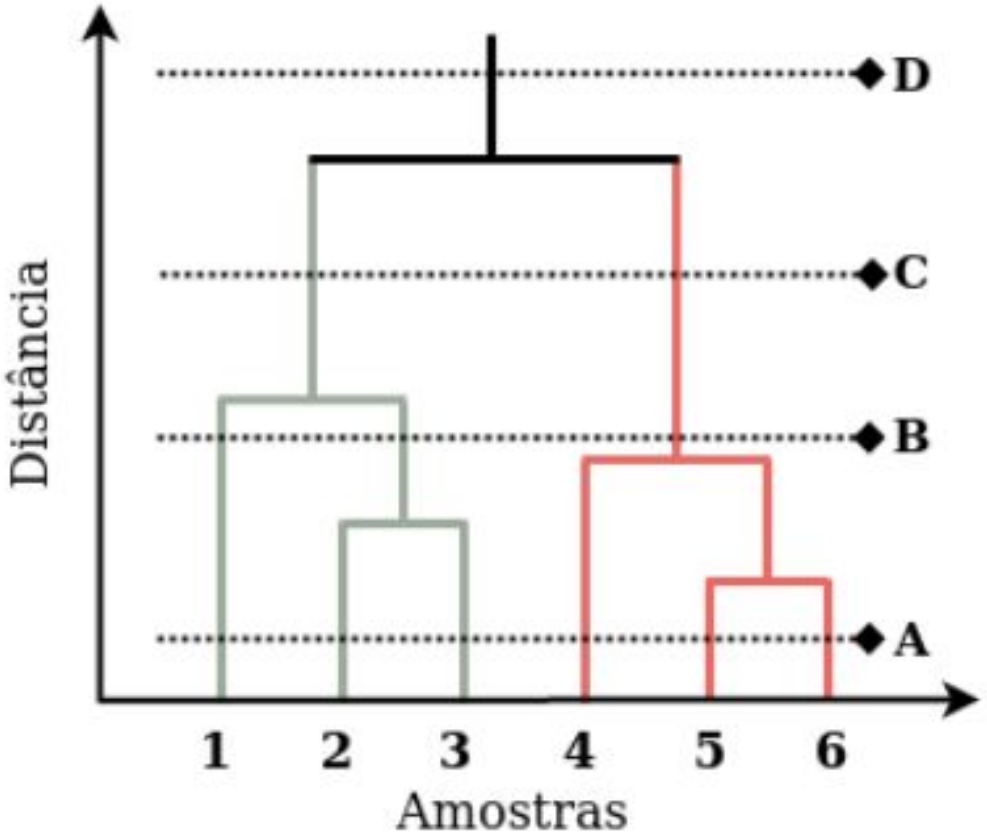
Average-link



Centroid distance



Algoritmos Hierárquicos
Método Algomerativo



Retas A-D **identificam diferentes momentos** do processo de agrupamento

- Em **A** → **6 agrupamentos unitários**
  - cada um contendo **UMA** amostra
- Em **B** → **3 agrupamentos:**
  - Agrupamento unitário (**amostra 1**)
  - Agrupamento das **amostras 2 e 3**
  - Agrupamento formado pelas **amostras 4, 5 e 6**
- Em **C** → **par de agrupamentos**
  - Agrupamento contendo as **amostras 1, 2 e 3**
  - Agrupamento das **amostras 4, 5 e 6**
- Em **D** → **alcançamos um único agrupamento superpopuloso**
  - contendo todas as amostras iniciais

Algoritmos Hierárquicos
Método Divisivo

- **Divisivo (*top-down*)**
  - Começa com um **cluster super populoso** → raiz da árvore
    - contendo todas as amostras
  - A cada iteração, um **ramo-pai** é dividido em dois subconjuntos menores, os **ramos-filhos**.
  - O processo **termina quando um critério de parada é atingido** → o número k de agrupamentos