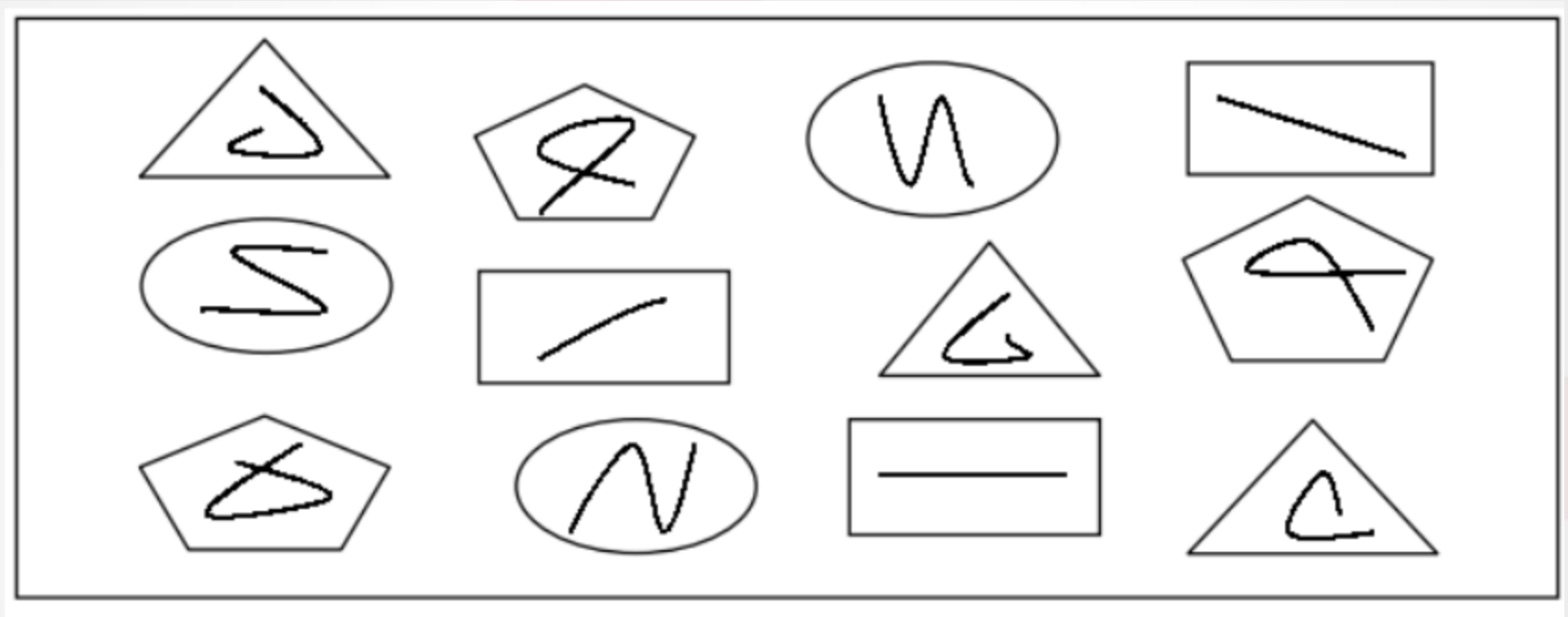
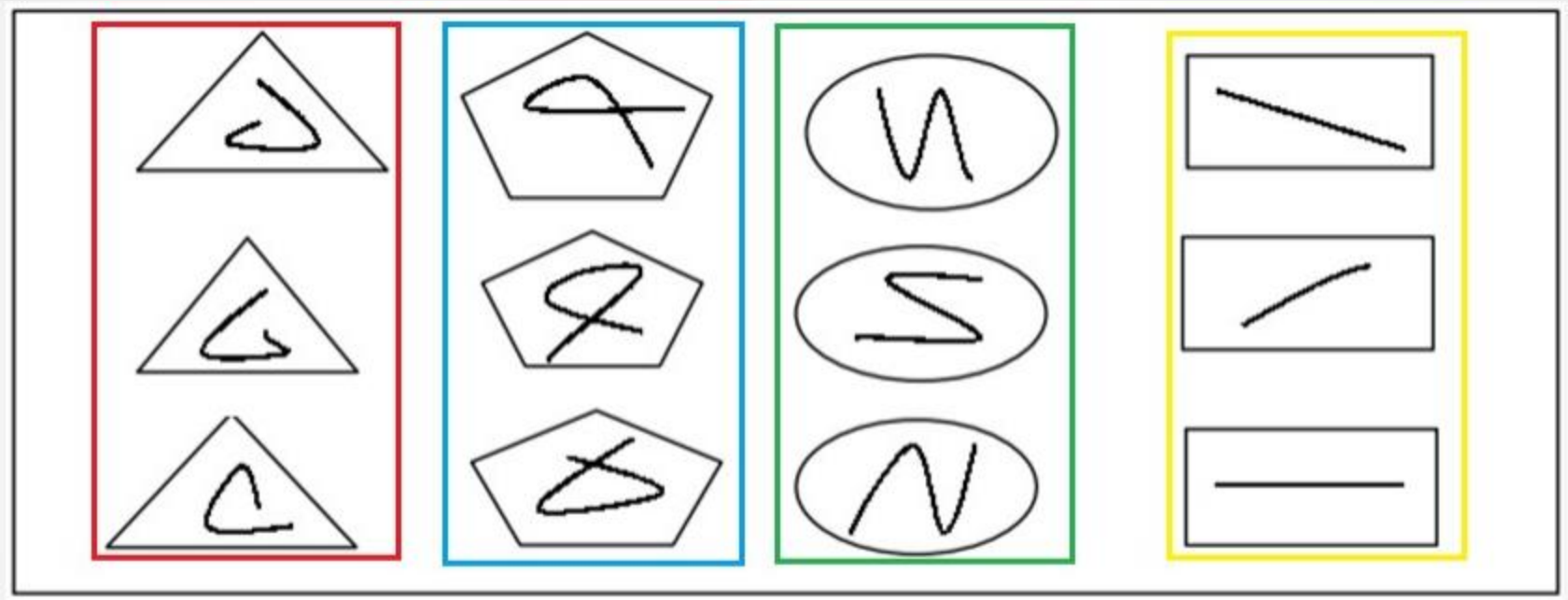
Clusterização

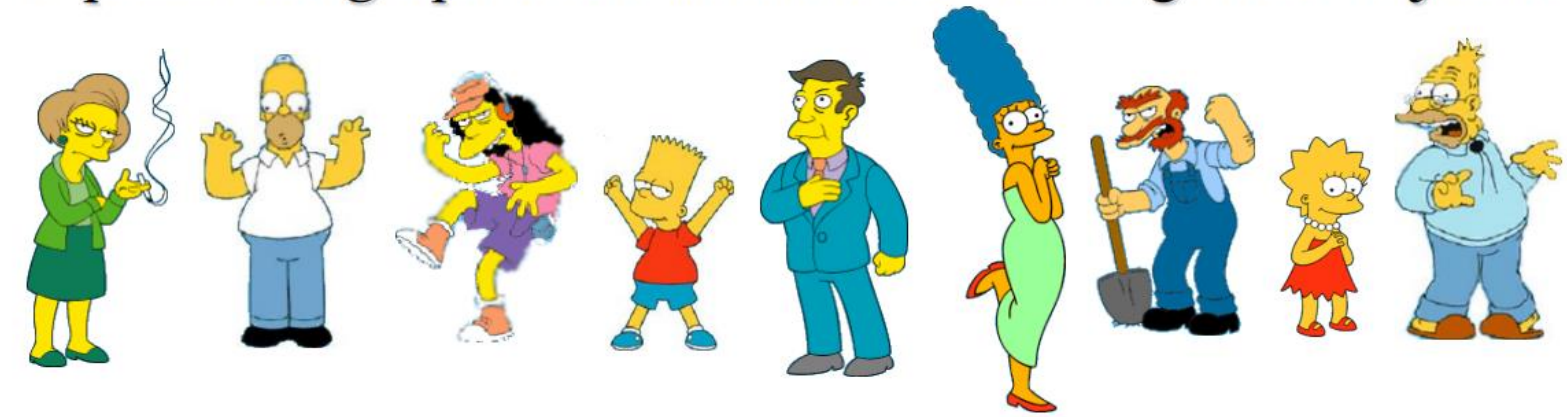


Clusterização



Clusterização

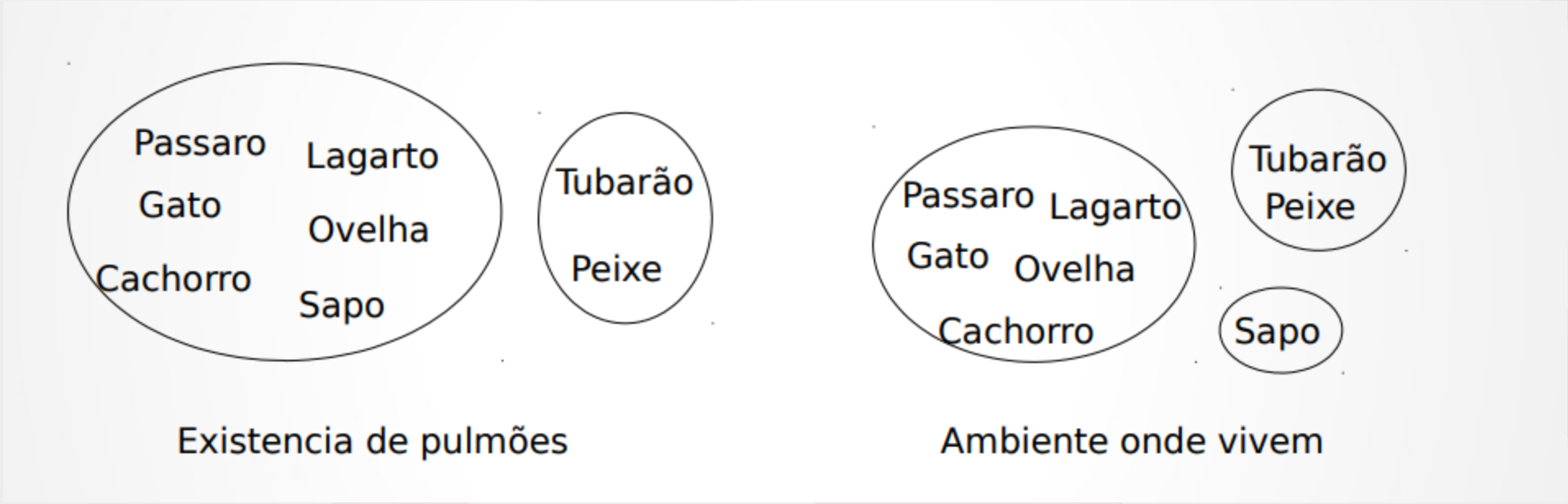
O que é um *agrupamento natural* entre os seguintes objetos?



Grupo é um conceito subjetivo:

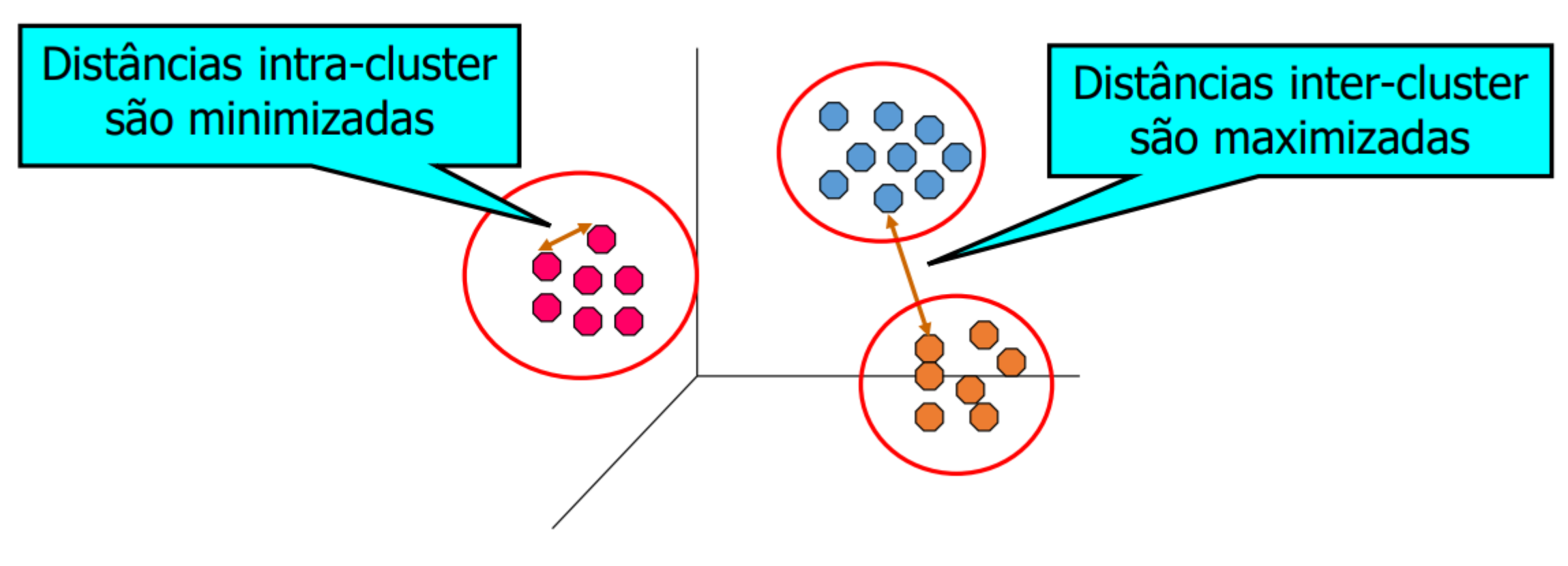
Família | Empregados da Escola | Mulheres | Homens

Clusterização



Clusterização

Clusterização



Clusterização

- As etapas do processo de aprendizagem não supervisionada são:
  - (1) Seleção de atributos
  - (2) Medida de proximidade
  - (3) Critério de agrupamento
  - (4) Algoritmo de agrupamento
  - (5) Verificação dos resultados
  - (6) Interpretação dos resultados

Clusterização

- Seleção de atributos:
  - Atributos devem ser adequadamente selecionados de forma a codificar a maior quantidade possível de informações relacionada a tarefa de interesse.
  - Os atributos devem ter também uma redundância mínima entre eles.
- Medida de Proximidade:
  - Medida para quantificar quão similar ou dissimilar são dois vetores de atributos.
  - É ideal que todos os atributos contribuam de maneira igual no cálculo da medida de proximidade.
  - Um atributo não pode ser dominante sobre o outro, ou seja, é importante normalizar os dados.

Clusterização

- Critério de Agrupamento:
  - Depende da interpretação que o especialista dá ao termo sensível com base no tipo de cluster que são esperados.
  - Por exemplo, um cluster compacto de vetores de atributos pode ser sensível de acordo com um critério enquanto outro cluster alongado, pode ser sensível de acordo com outro critério.

Clusterização

- Algoritmo de Agrupamento:
  - Tendo adotado uma medida de proximidade e um critério de agrupamento devemos escolher um algoritmo de agrupamento que revele a estrutura agrupada do conjunto de dados
- Validação dos Resultados:
  - Uma vez obtidos os resultados do algoritmo de agrupamento, devemos verificar se o resultado esta correto.
  - Isto geralmente é feito através de testes apropriados.
- Interpretação dos Resultados:
  - Em geral, os resultados do agrupamento devem ser integrados com outras evidências experimentais e análises para chegar as conclusões corretas.

Clusterização

- Medidas de Dissimilaridade:
  - Métrica l_p ponderada;
  - Métrica Norma l_∞ ponderada;
  - Métrica l2 ponderada (Mahalanobis);
  - Métrica lp especial (Manhattan);
  - Distância de Hamming;
- Medidas de Similaridade:
  - Produto interno (inner);
  - Medida de Tanimoto;

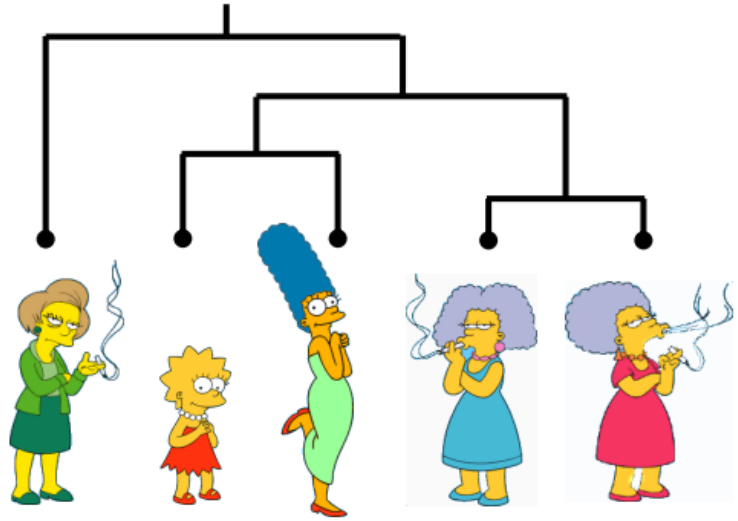
Clusterização

- Os algoritmos de agrupamento buscam identificar padrões existentes em conjuntos de dados.
- Os algoritmos de agrupamento podem ser divididos em varias categorias:
  - Métodos Hierárquicos;
  - Métodos Particionais;
  - Métodos Baseados em Densidade;
  - Métodos Baseados em Grade;
  - Métodos Baseados em Modelos;
  - Métodos Baseados em Redes Neurais;
  - Métodos Baseados em Lógica Fuzzy;
  - Métodos Baseados em Kernel;
  - Métodos Baseados em Grafos;
  - Métodos Baseados em Computação Evolucionária.

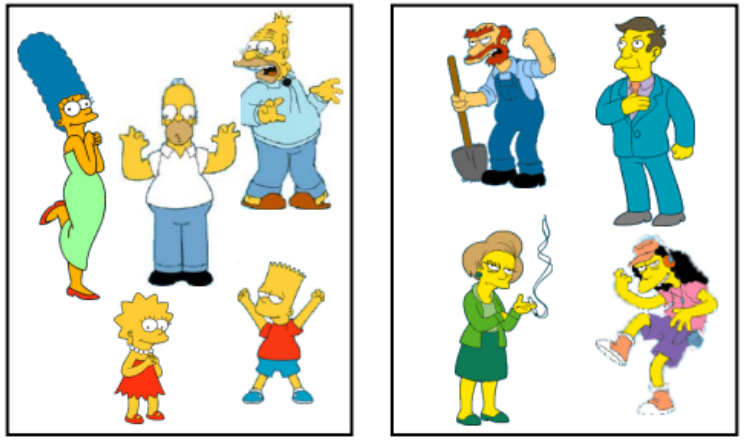
Clusterização

- **Particionais:** constroem uma partição dos dados
- **Hierárquicos:** constroem uma hierarquia de partições

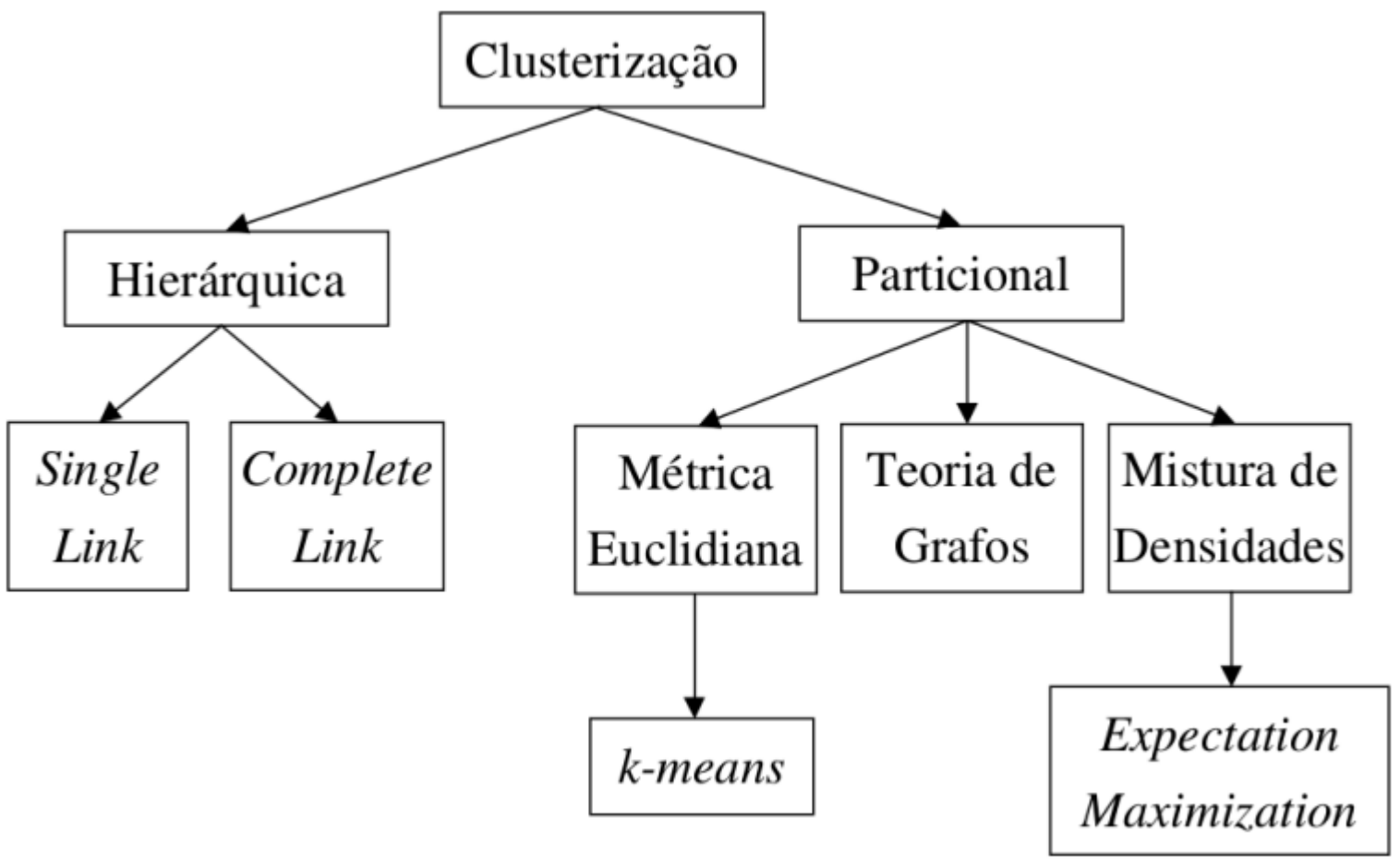
Hierárquicos



Particionais



Clusterização



Clusterização

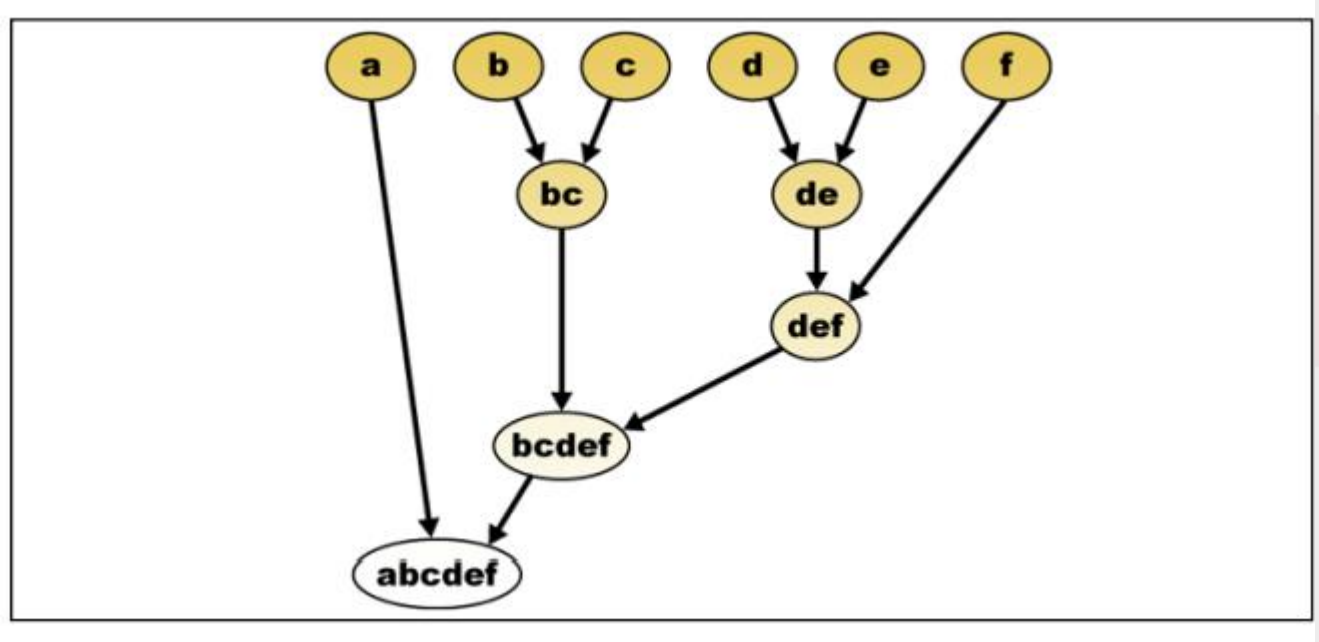
- Algoritmos de Particionamento:
  - Os algoritmos particionais dividem a base de dados em k-grupos, onde o número k é dado pelo usuário.
  - São algoritmos diretos e rápidos.
  - Geralmente, todos os vetores de características são apresentados ao algoritmo uma ou várias vezes.
  - O resultado final geralmente depende da ordem de apresentação dos vetores de características.

Clusterização

- Algoritmos Hierárquicos:
  - Algoritmos de clusterização baseados no método hierárquico (HC) organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos.
  - Os resultados de um algoritmo HC são normalmente mostrados como uma árvore binária ou dendograma, que é uma árvore que iterativamente divide a base de dados em subconjuntos menores.
  - A raiz do dendograma representa o conjunto de dados inteiro e os nós folhas representam os indivíduos.

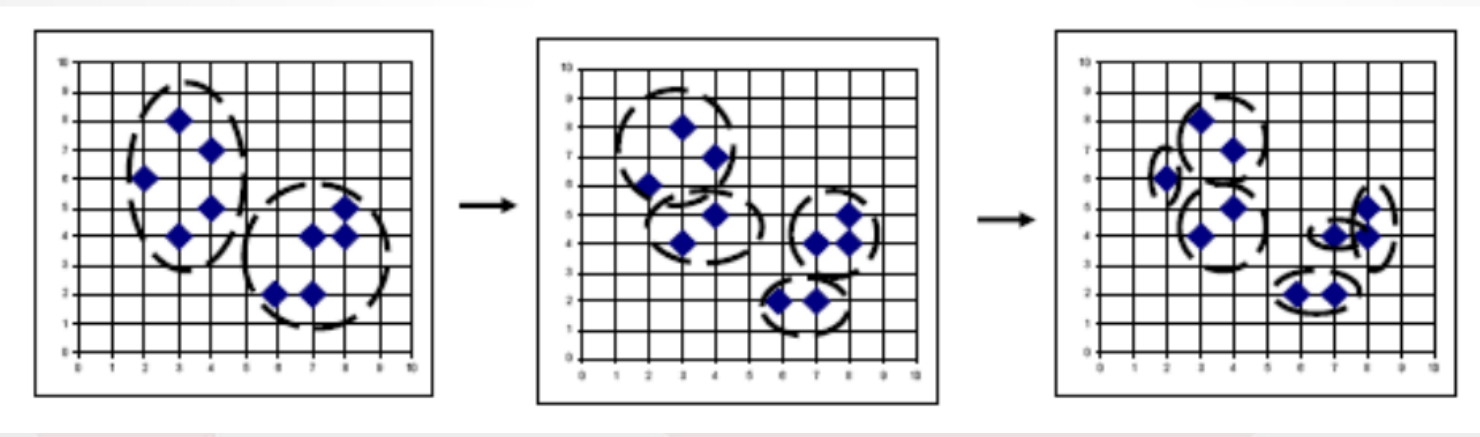
Clusterização

- Algoritmos Hierárquicos:



Clusterização

- Algoritmos Hierárquicos:
  - Aglomerativos (bottom-up):
    - Produzem uma sequência de agrupamentos com um número decrescente de clusters a cada passo.
    - Os agrupamentos produzidos em cada passo resultam da fusão de dois clusters em um.
  - Divisivos (top-down):
    - Atuam na direção oposta, isto é, eles produzem uma sequência de agrupamentos com um número crescente de clusters a cada passo.
    - Os agrupamentos produzidos em cada passo resultam da partição de um único cluster em dois.



Clusterização

- Algoritmos Baseados em Densidade:
  - Clusters são definidos como regiões densas, separadas por regiões menos densas que representam os ruídos
- Algoritmos Baseados em Grade:
  - Usam uma estrutura de dados em grade de multiresolução.
- Algoritmos Baseados em Modelo:
  - Usam um modelo de referência para cada cluster.
- Algoritmos Baseados em Lógica Fuzzy:
  - Os métodos de clusterização baseados em Lógica Fuzzy são métodos não ‘hard’, que permitem associar um indivíduo a todos os clusters usando uma função de pertinência.

Clusterização

- Algoritmos Baseados em RNA:
  - Têm suas raízes no método de clusterização ART (Teoria Ressonante Adaptativa) ou nos Mapas Auto Organizáveis de Kohonen.
- Algoritmos Baseados em Kernel:
  - Usam do espaço de características para permitir uma separação não-linear no espaço de entrada.
  - São capazes de produzir uma separação não linear entre os hiperespaços dos clusters, ao contrário dos algoritmos tradicionais que produzem por partes fronteiras lineares entre os dados

Clusterização

- Algoritmos Baseados em Grafos:
  - Buscam representar um conjunto de dados em um grafo, onde cada vértice representa um elemento do conjunto de dados e a existência de uma aresta conectando dois vértices é feita com base na proximidade entre os dois dados.
- Algoritmos Baseados em Computação Evolucionária:
  - Compreende um conjunto de técnicas de busca e otimização baseados em mecanismos da evolução biológica, tais como reprodução, mutação, recombinação e seleção natural e estão sendo utilizados amplamente pela comunidade de inteligência artificial para obter modelos de inteligência computacional.