Agrupamento de Dados (*Clustering*)

Organização

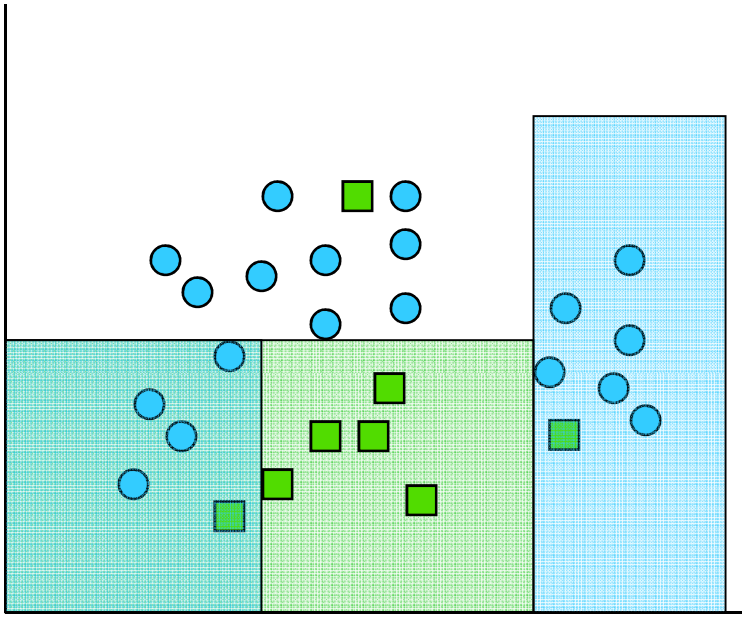
1. Introdução
2. Medidas de (Dis)similaridade
3. Métodos de Agrupamento (métodos hierárquicos, de partição)
4. Critérios numéricos para definir o número de *clusters*

1. Introdução

Aplicações para Técnicas de *Agrupamento* :

- **Marketing:** descobrir grupos de clientes e usá-los para marketing direcionado.
- **Astronomia:** encontrar grupos de estrelas e galáxias similares.
- **Estudos sobre terrremotos:** observar se epicentros estão agrupados em falhas continentais.
- **Bioinformática:** encontrar clusters de genes com expressões semelhantes.
- **Organização de textos** (*text mining*).
- Etc.

Classificação X *Clustering*



Classificação:

Aprender um método para predizer a classe de uma instância (objeto, exemplo, registro) a partir de exemplos pré-rotulados (classificados)

Clustering:

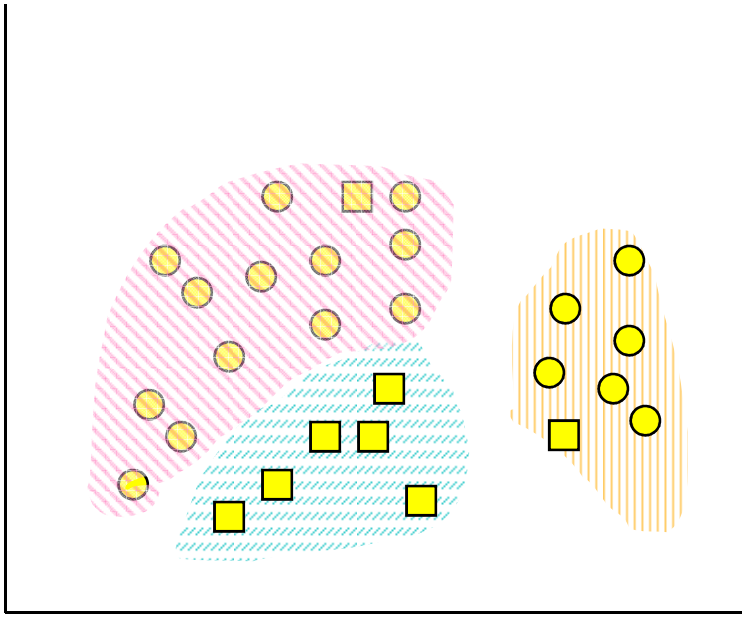
Encontrar os rótulos das classes e o número de classes diretamente a partir dos dados.

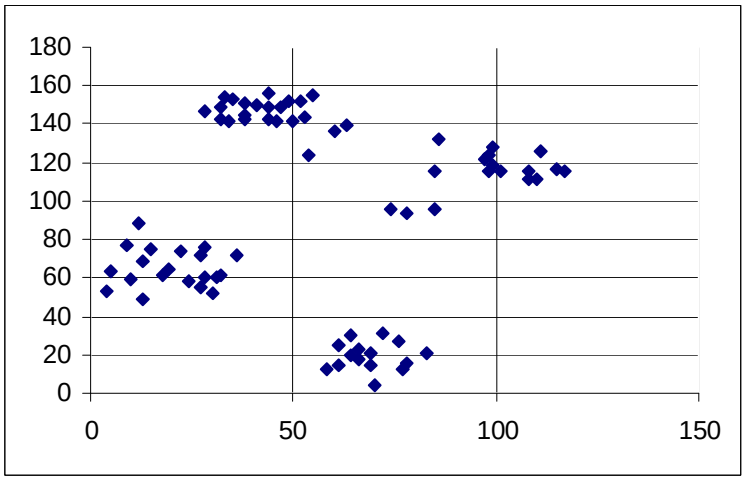
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

Agrupamento de Dados (*Clustering*)

- Aprendizado não supervisionado, segmentação;

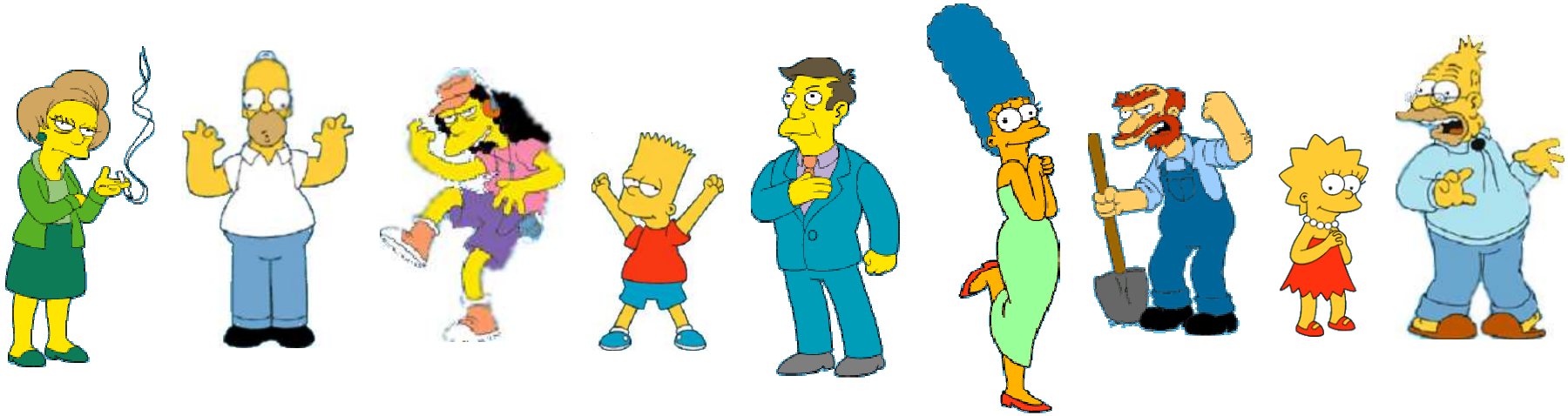
- Encontrar grupos “naturais” de objetos para um conjunto de dados não rotulados.



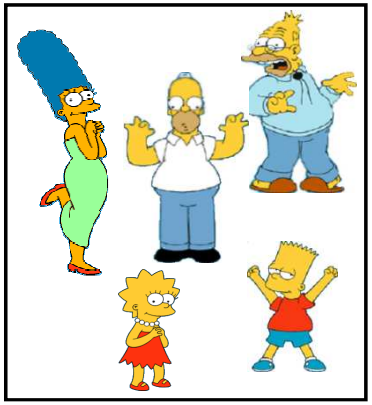


Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

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

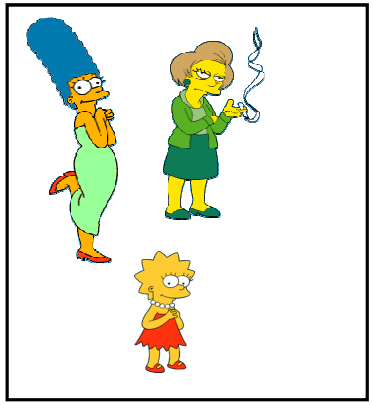


Cluster é um conceito subjetivo!



Família

Empregados da escola



Mulheres

Homens

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

O que é um cluster?

- Definições subjetivas;
- Homogeneidade (coesão interna) e Heterogeneidade (separação entre grupos);
- Diversos critérios (índices numéricos);
- Induzir, *impor* uma estrutura aos dados.

Visualizando *clusters* :

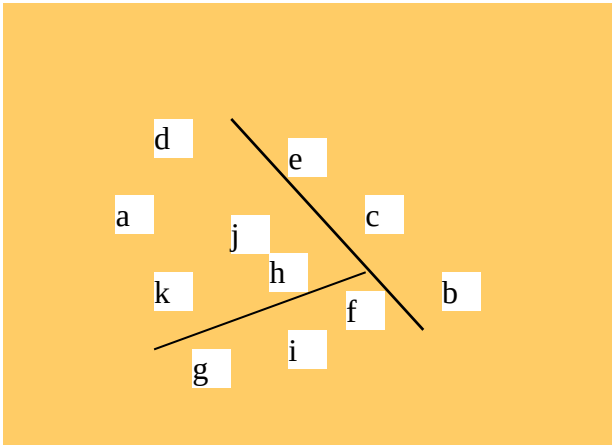
- Sistema visual humano é muito poderoso para reconhecer padrões;
- Entretanto...
  - "*Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent*." (Carl Sagan)
- Everitt et al., Cluster Analysis, Chapter 2 (Visualizing Clusters), Fourth Edition, Arnold, 2001.

Métodos para *clustering*

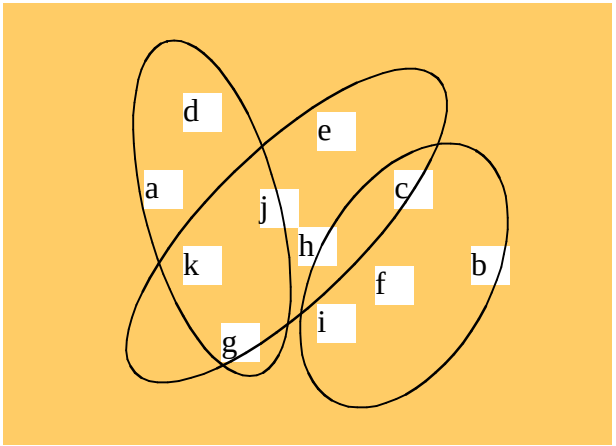
- Muitos métodos/algoritmos diferentes:
  - Para dados numéricos e/ou simbólicos
  - Determinísticos X probabilísticos
  - Para obter partições ou hierárquicos
  - Partições *rígidas* ou *sobrepostas*

Clusters com e sem sobreposição:

Sem sobreposição



Com sobreposição



Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

Avaliação de Métodos de Agrupamento:

- Inspeção *manual/visual;*
- *Benchmarking* em bases rotuladas;
- Medidas de qualidade dos *clusters:*
  - Medidas de distância.
  - Alta similaridade interna ao cluster e baixa similaridade entre clusters distintos.

Organização

1. Introdução
2. **Medidas de Similaridade**
3. Métodos de Agrupamento (métodos hierárquicos, partições)
4. Critérios numéricos para definir o número de *clusters*

2. Medidas de Similaridade

Como definir *similaridade*?

Adotaremos uma abordagem matemática.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Propriedades desejáveis para medidas de distância:

$D(A,B) = D(B,A)$ ***Simetria***

Caso contrário seria possível dizer que "João é parecido com José", mas "José não é parecido com João".

$D(A,A) = 0$ ***Auto-similaridade***

Caso contrário seria possível dizer que "João é mais parecido com José do que ele mesmo".

$D(A,B) = 0$ se $A=B$ ***Separação***

Caso contrário seria impossível "separar" objetos diferentes.

$D(A,B) \leq D(A,C) + D(B,C)$ ***Desigualdade triangular***

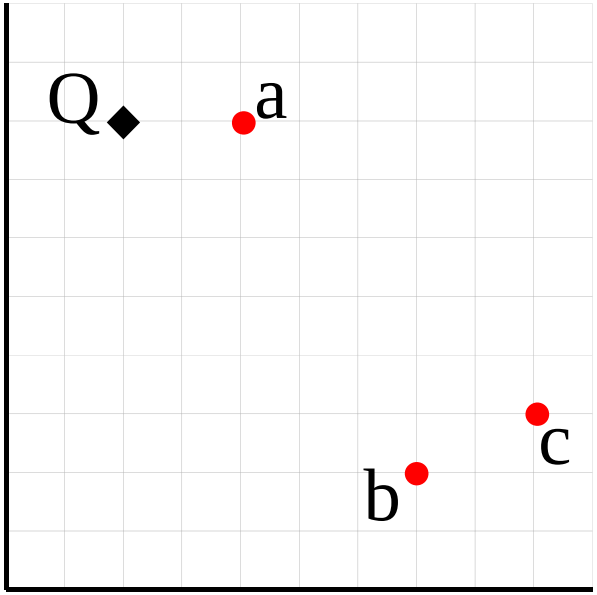
Caso contrário seria possível dizer:

A é muito parecido com C, B é muito parecido com C, mas A é muito dif. de B.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Importância da *desigualdade triangular:*

Suponhamos que se deseja encontrar o objeto mais próximo do objeto Q em uma base de dados formada por 3 objetos. Assumamos que a desigualdade triangular se aplica e que se disponha de uma tabela de distâncias entre todos os objetos da base de dados.



	a	b	c
a		6.70	7.07
b			2.30
c			

Inicialmente calculamos a distância entre Q e **a** (**2 unidades**) e entre Q e **b** (**7.81 unidades**). Neste caso, não é necessário calcular a distância entre Q e **c**, pois:

$$D(Q,\mathbf{b}) \leq D(Q,\mathbf{c}) + D(\mathbf{b},\mathbf{c})$$
$$D(Q,\mathbf{b}) - D(b,\mathbf{c}) \leq D(Q,\mathbf{c})$$
$$7.81 - 2.30 \leq D(Q,\mathbf{c})$$
$$5.51 \leq D(Q,\mathbf{c})$$

Ou seja, já se pode afirmar que **a** está mais próximo de Q do que qualquer outro objeto da base de dados.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Notação:

- Denotemos por $\mathbf{X}_{nxp}$ uma matriz formada por n linhas (correspondentes aos objetos da base de dados) e p colunas (atributos):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Sempre que não causar confusão, por simplicidade cada objeto da base de dados (linha da matriz) será denotado por um vetor $\mathbf{x}^i$. Exemplo:

$$\mathbf{x}^1 = \begin{bmatrix} x_{11} & \cdots & x_{1p} \end{bmatrix}$$

Medindo (Dis)similaridade:

- Distância Euclidiana (norma L_2) entre objetos *i* e *j* :

$$d_{ij}^{E} = \delta_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Exemplo: $\mathbf{x}^1=[2\ 3]$; $\mathbf{x}^2=[1\ 2]$; $\mathbf{x}^3=[0\ -1]$;

$d_{12}=?$ $d_{13}=?$ $d_{31}=?$

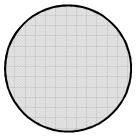
$$d_{12}^{E} = \delta_{12} = \sqrt{(2-1)^2 + (3-2)^2} = \sqrt{2}$$

Distância Euclidiana é uma medida de similaridade?

* Implementações mais eficientes trabalham com distâncias Euclidianas quadradas.

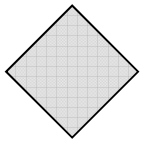
Outras medidas de distância:

$$d_{ij}^{Minkowski} = \sqrt[r]{\sum_{k=1}^{p}(x_{ik} - x_{jk})^r} \, , r \geq 1$$



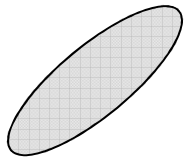
Euclidiana (r = 2)

$$d_{ij}^{M} = \delta_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$



Manhattan (L_1)

Outras medidas de distância: Mahalanobis, correlação, etc.



Mahalanobis

A escolha da medida de distância usualmente depende da aplicação. Eficiência computacional também é importante.

Atributos nominais:

- Estado civil, passatempo preferido, modelo do carro, clube favorito, país, religião, etc.
- Medidas de dissimilaridade entre objetos *i* e *j* descritos por atributos nominais *(simple matching):*

$$d_{SM}(i,j) = \sum_{k=1}^{k=p} s_k \qquad \begin{cases} (x_{ik} = x_{jk}) \Rightarrow s_k = 0; \\ (x_{ik} \neq x_{jk}) \Rightarrow s_k = 1; \end{cases}$$

Bases de dados formadas por diversos tipos de atributos:

- Método de Gower(1971) sem ponderação:

$$s_{ij} = \sum_{k=1}^{p} s_{ijk}$$

Para atributos nominais/binários:

$$\begin{cases} (x_{ik} = x_{jk}) \Rightarrow s_{ijk} = 1; \\ (x_{ik} \neq x_{jk}) \Rightarrow s_{ijk} = 0; \end{cases}$$

Para atributos numéricos:

$$s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k$$

R_k= faixa de observações do atributo k *(termo de normalização).*

Medidas de Dis(similaridade) entre clusters:

- Diversos algoritmos de *clustering* requerem avaliações de dissimilaridades entre *clusters*.
- Como medir dissimilaridade?

  Basicamente há duas abordagens possíveis:

  - Sumarizar as proximidades dos objetos de cada *cluster* (e.g. distâncias entre centróides, distância média entre objetos de clusters distintos);
  - Objetos representativos de cada *cluster (e.g.* medóides, menor dissimilaridade entre objetos de *clusters* distintos, maior distância entre objetos de *clusters* distintos).

Proximidade entre grupos para dados contínuos e nominais:

- Contínuos:

$$dist_{AB}^{E} = \delta_{AB} = \sqrt{\sum_{k=1}^{p} (\bar{x}_{Ak} - \bar{x}_{Bk})^2}$$

$$\bar{\mathbf{x}}'_A = [\bar{x}_{A1}, \ldots, \bar{x}_{Ap}] \qquad \bar{\mathbf{x}}'_B = [\bar{x}_{B1}, \ldots, \bar{x}_{Bp}]$$

- Nominais (índice de dissimilaridade de Balakrishnan e Sanghvi, 1968):

Considerando c_k valores distintos para cada variável k temos:

$$G^2 = \sum_{k=1}^{p} \sum_{l=1}^{c_k} \frac{(p_{Akl} - p_{Bkl})^2}{p_{kl}} \qquad p_{kl} = \frac{1}{2}(p_{Akl} + p_{Bkl})$$

Ponderação de atributos:

- Atribuir maior ou menor importância w_k para cada atributo k no cálculo das (dis)similaridades:
  - Conhecimento de domínio;
  - Seleção de atributos ($w_k=0$ *ou* $w_k=1$*);*
  - Peso w_k inversamente proporcional à variabilidade dos valores do atributo k*;*
    - Miligan & Cooper (1988): amplitude do intervalo é mais eficaz do que o desvio padrão. Tais procedimentos podem ser vistos como uma forma de padronização;
    - Questionável, pois não se leva em consideração o poder de discriminação da variável, que pode ser alto justamente devido à grande variabilidade dos valores totais;
    - Alternativa: estimar a variabilidade levando em conta os *clusters*. Problema: clusters são de fato desconhecidos *a priori*.

Padronização:

- Normalização linear;
- Escore z;
- Divisão pela amplitude ou pelo desvio padrão;

→ Critérios de padronização baseados na variabilidade assumem que a importância do atributo é inversamente proporcional à variabilidade de seus valores.

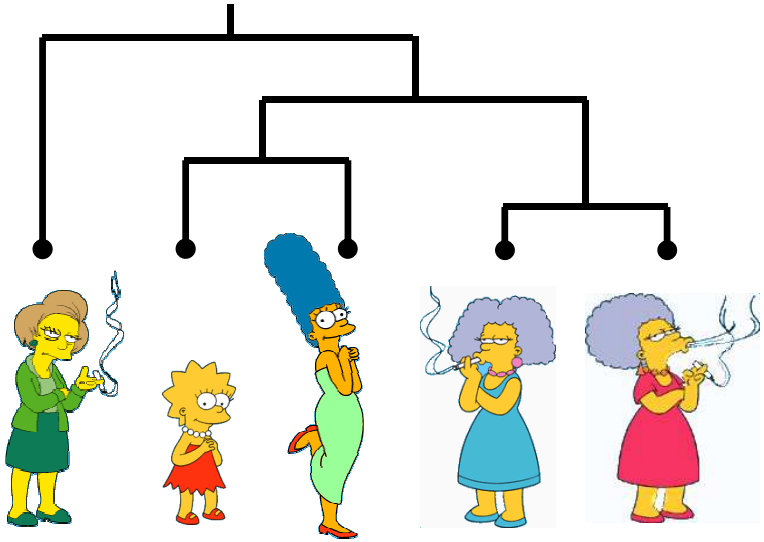
Organização

1. Introdução
2. Medidas de Similaridade
3. **Métodos de Agrupamento (métodos hierárquicos,** partições)
4. Critérios numéricos para definir o número de *clusters*

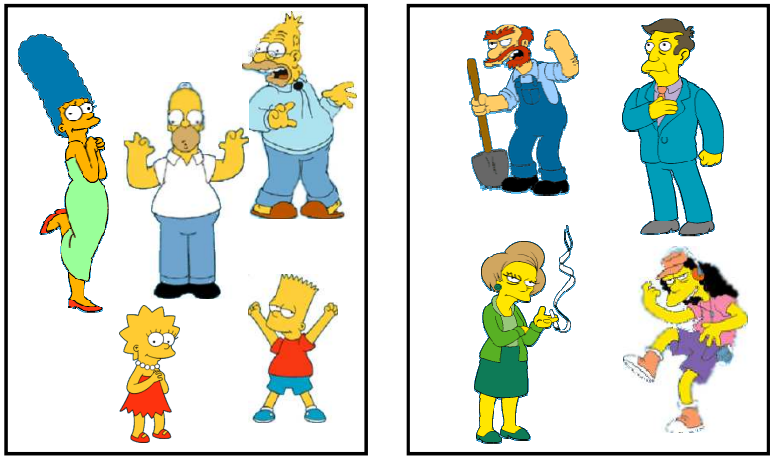
Métodos de Agrupamento

• **Métodos para particionamento:** construir várias partições e avaliá-las segundo algum critério.
• **Métodos hierárquicos:** criar uma decomposição hierárquica do conjunto de objetos usando algum critério.

Hierárquicos



"Particionais"



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Propriedades desejáveis para um Algoritmo de Agrupamento

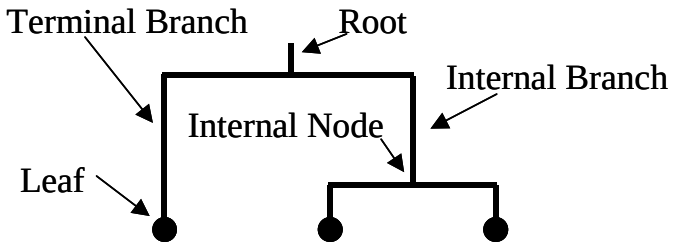
- Levando-se em conta a eficiência computacional:
  - Habilidade para lidar com diferentes tipos de dados (qualitativos e quantitativos)
  - Requerimentos mínimos em relação ao conhecimento de domínio para determinar os parâmetros de entrada;
  - Capacidade de lidar com ruído e *outliers*;
  - Insensibilidade em relação à *ordem de apresentação* dos registros de entrada (vetores de dados);
  - Incorporação de restrições definidas pelo usuário;
  - Interpretabilidade e usabilidade.

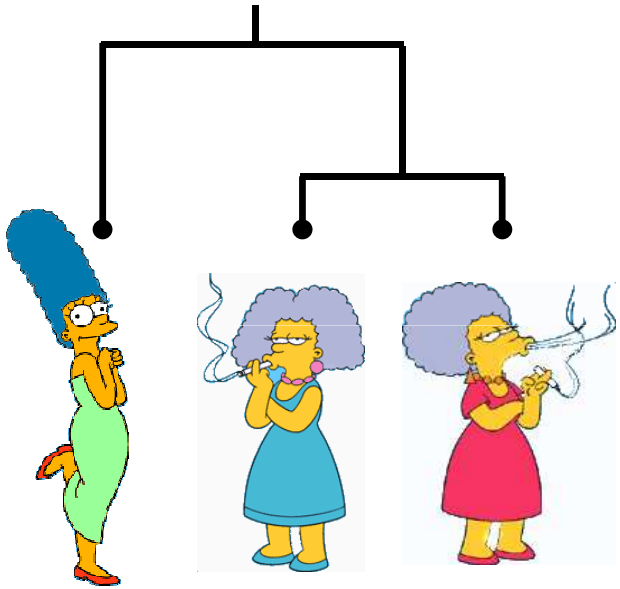
→ Vamos iniciar pelos métodos hierárquicos...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Dendrograma

É uma ferramenta útil para sumarizar medidas de (dis)similaridade:





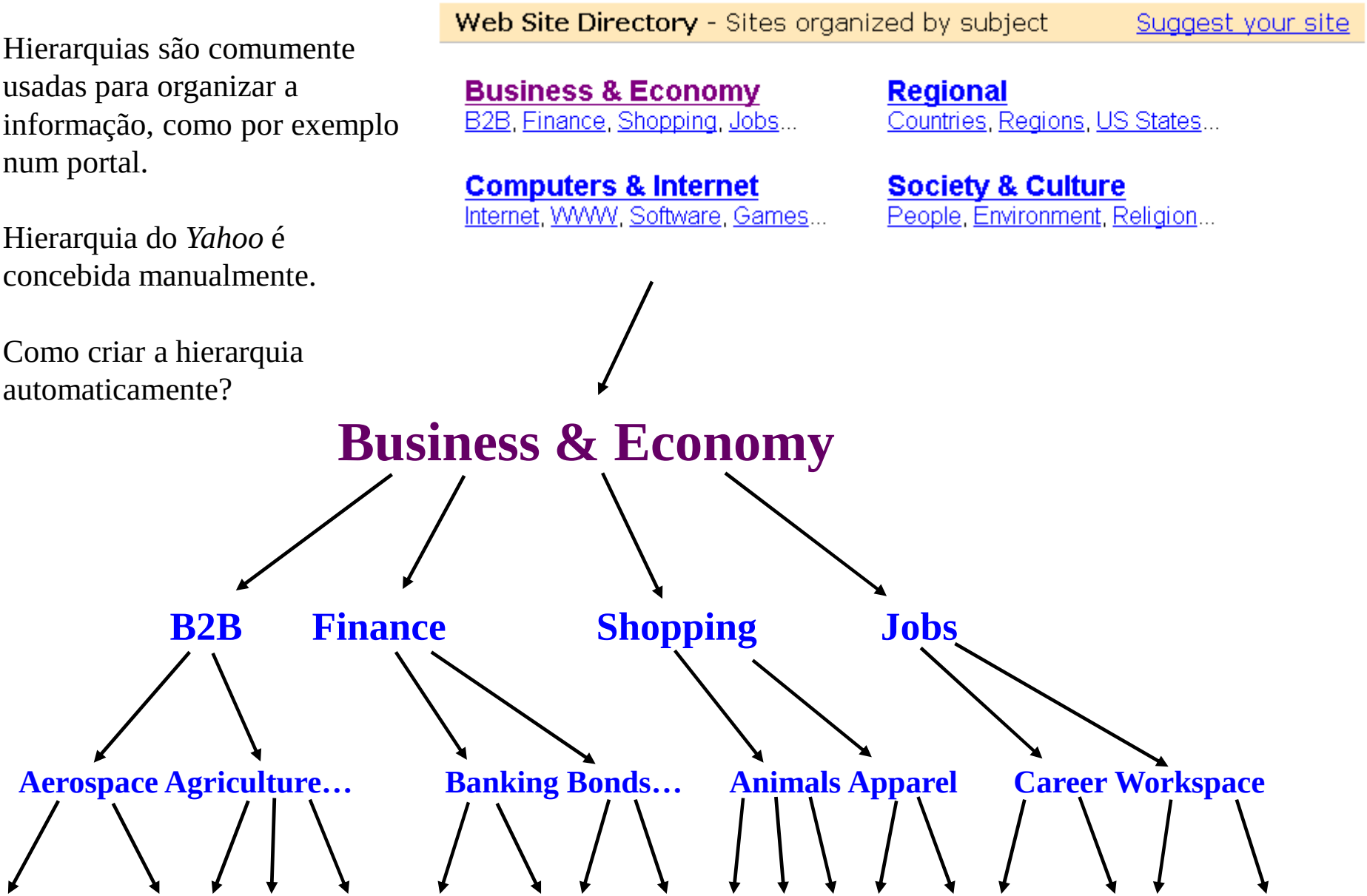
* A dissimilaridade entre dois objetos é representada como a altura do nó interno mais baixo compartilhado.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Hierarquias são comumente usadas para organizar a informação, como por exemplo num portal.

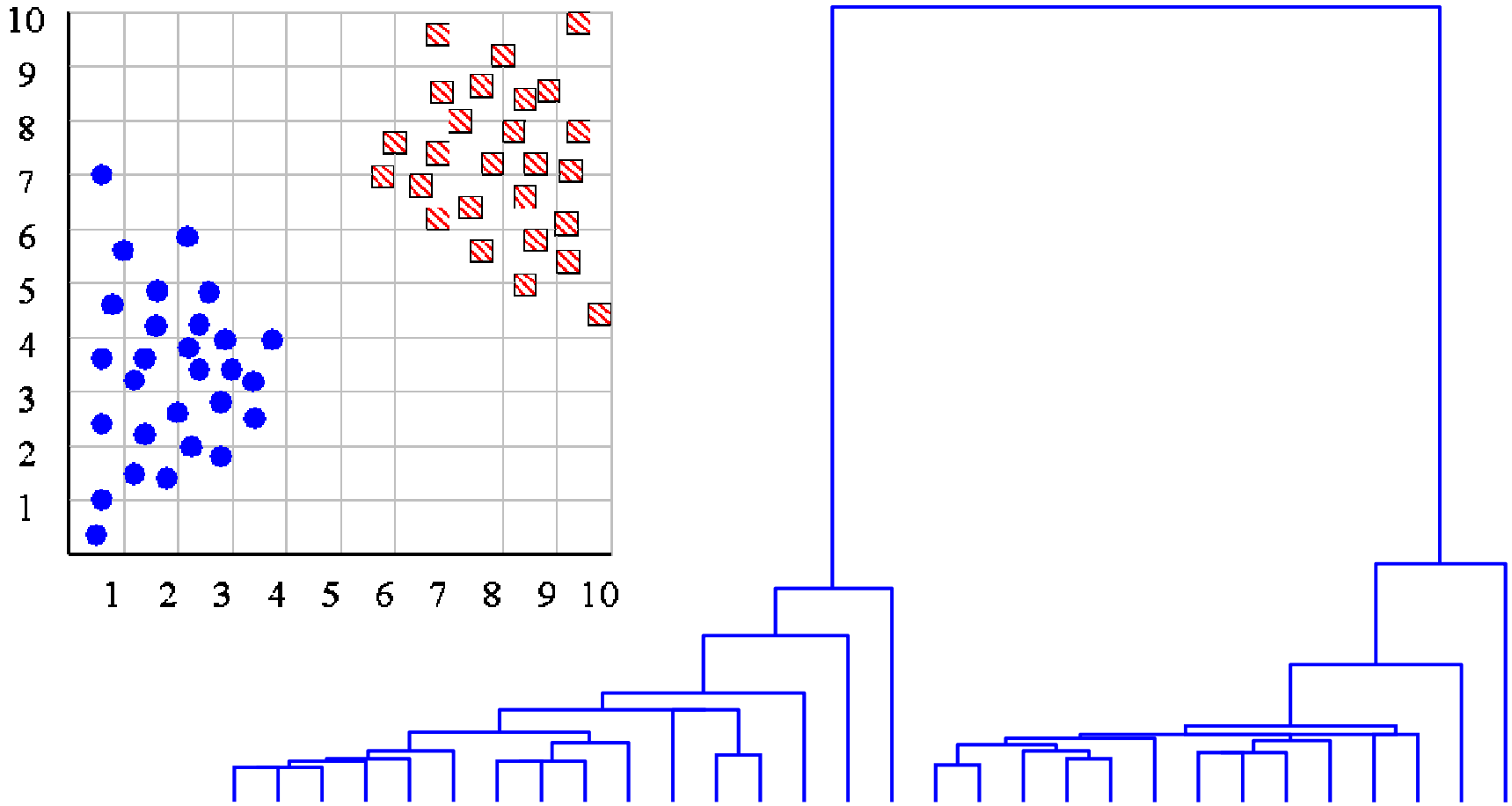
Hierarquia do *Yahoo* é concebida manualmente.

Como criar a hierarquia automaticamente?



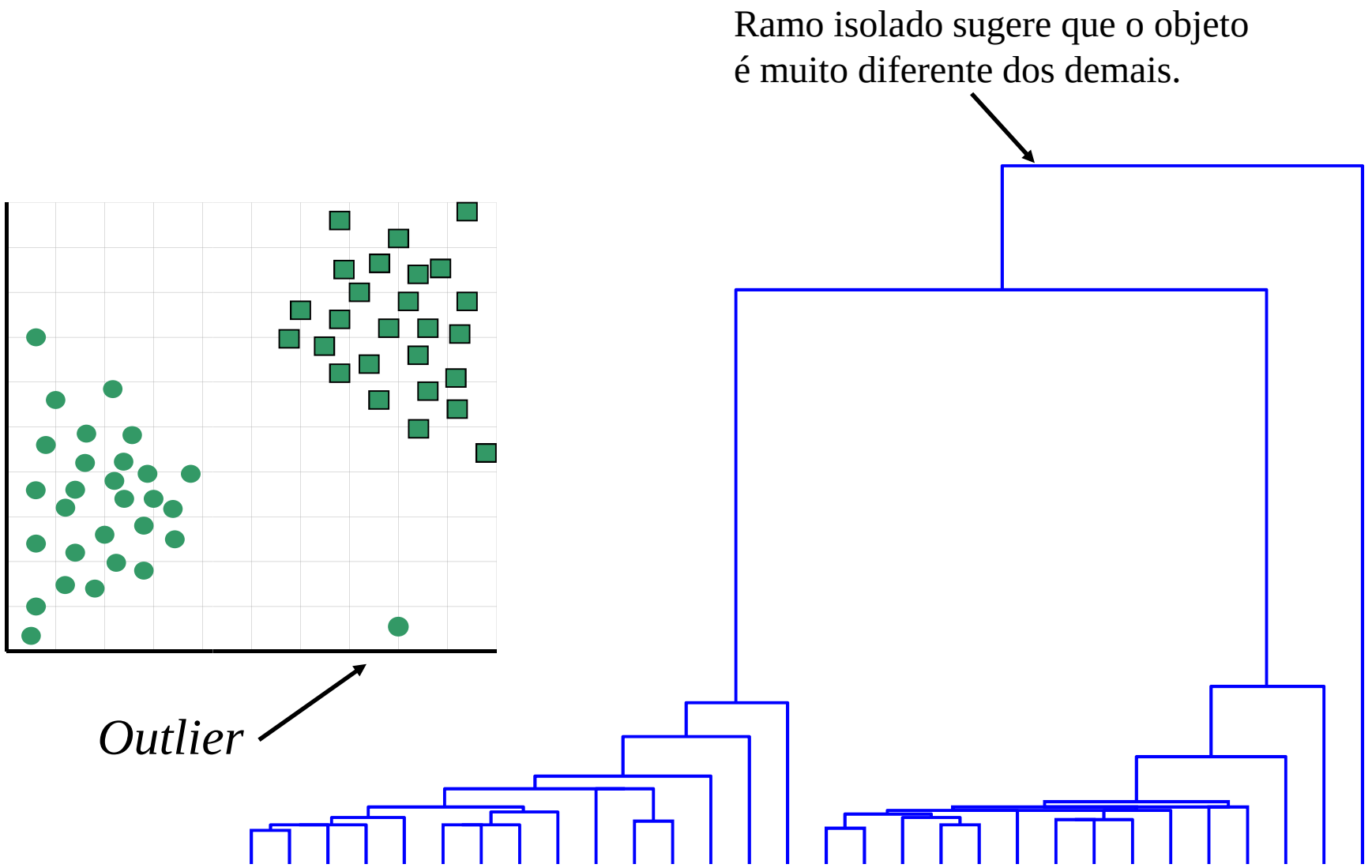
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Pode-se examinar o dendrograma para estimar o número *correto* de *clusters*. No caso abaixo, existem duas sub-árvores bem separadas, sugerindo dois grupos de dados. Infelizmente na prática as distinções não são tão simples…



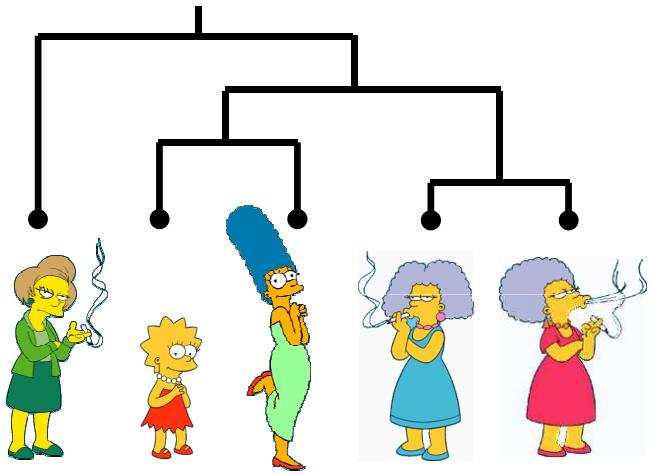
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Pode-se usar o dendrograma para detectar *outliers:*



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Métodos *clássicos* para Agrupamento Hierárquico

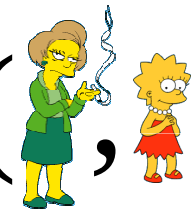


Bottom-Up (aglomerativos):
- iniciar colocando cada objeto em um *cluster*;
- encontrar o melhor par de *clusters* para uni-los;
- Repetir até que todos os *clusters* sejam reunidos em um só *cluster*.

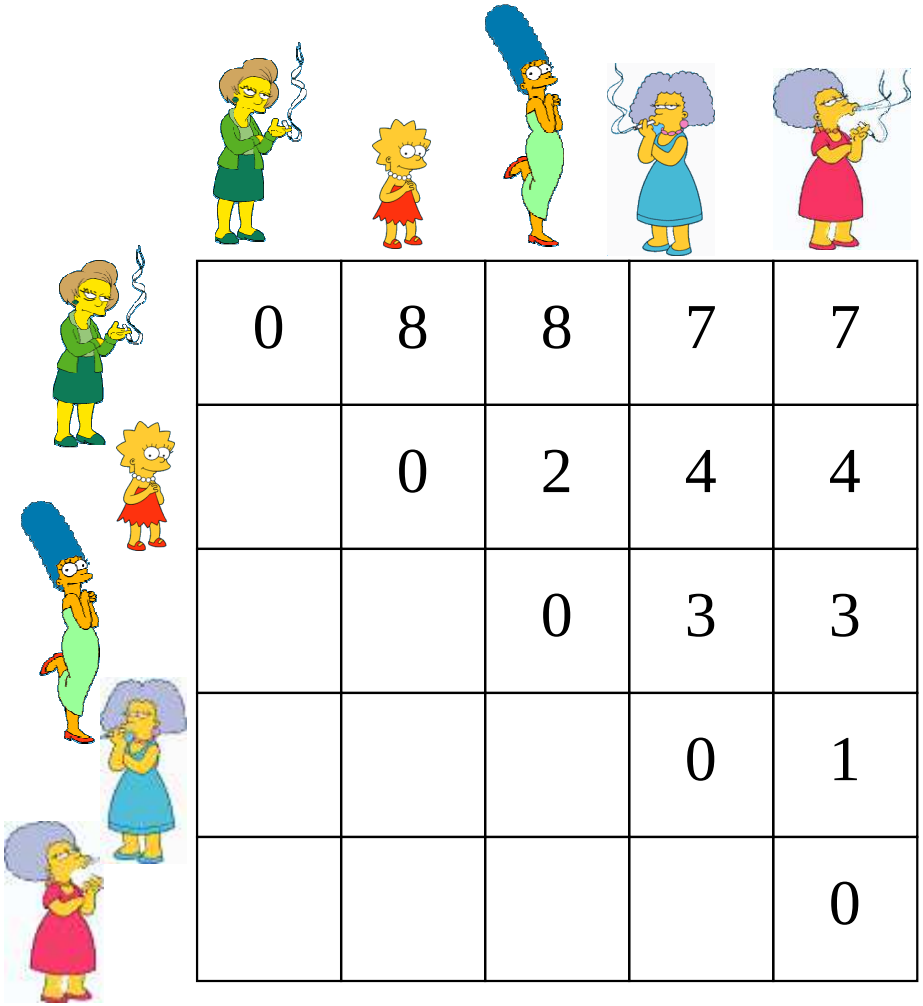
Top-Down (divisivos):
- Iniciar colocando todos os objetos em um único *cluster*;
- Considerar modos possíveis de dividir o *cluster* em dois;
- Escolher a melhor divisão e recursivamente operar em ambos os lados até que cada objeto forme um *cluster*.

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Inicialmente calculamos uma matriz de distâncias. Esta contém as distâncias entre cada par de objetos da base de dados:

D() = 8

D() = 1

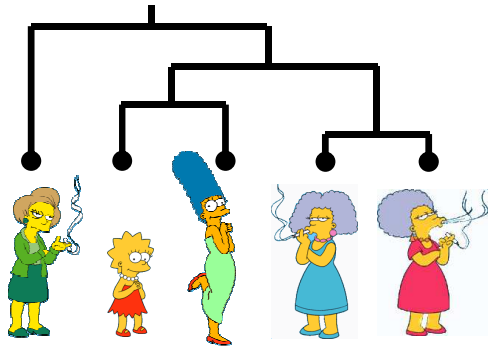


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

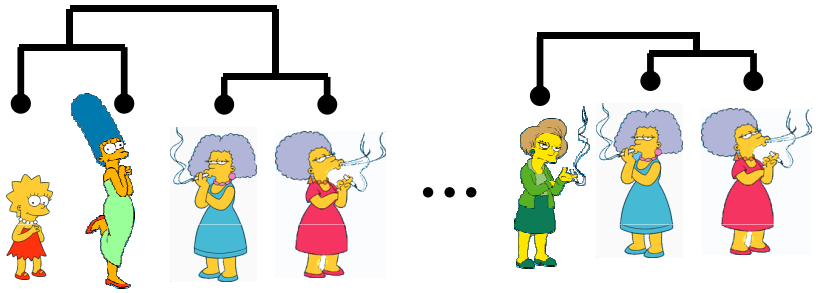
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Bottom-Up (aglomerativo):

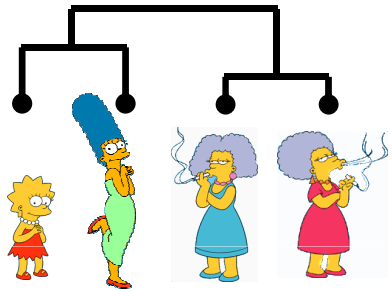
Iniciando com cada objeto em seu próprio cluster, encontrar o melhor par de *clusters* para unir em um novo *cluster.* Repetir até que todos os *clusters* sejam fundidos em um único *cluster*.



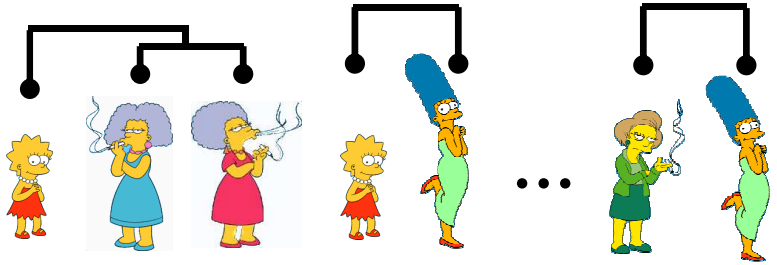
Considerar todas as uniões possíveis …



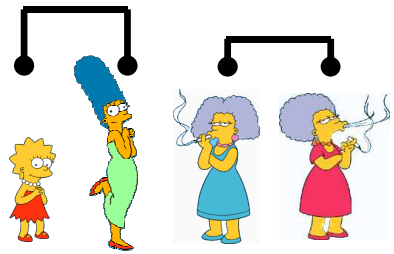
Escolher a melhor



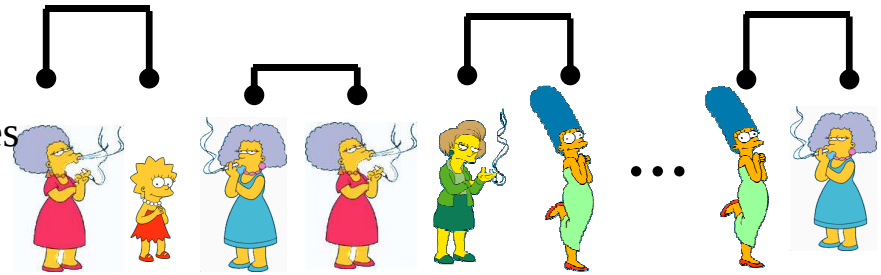
Considerar todas as uniões possíveis …



Escolher a melhor



Considerar todas as uniões possíveis …



Escolher a melhor

Como medir dis(similaridades) entre clusters?

- Vizinho mais próximo - *Single linkage* ou *nearest-neighbour (Florek – 1951, Sneath -1957)*:
  - Distância entre *clusters* é dada pela menor distância entre dois objetos (um de cada *cluster*). No método *divisivo*, seleciona-se inicialmente os objetos mais distantes entre si, agrupando-se os demais em função destas duas sementes iniciais;
  - Exemplo de método aglomerativo (Everitt et al., Cluster Analysis, 2001):

- Consideremos a seguinte matriz de distâncias iniciais (MD_1) entre 5 objetos {1,2,3,4,5}. Qual par de objetos será escolhido para formar o primeiro *cluster*?

$$MD_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ \boxed{2} & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix} \end{array}$$

- A menor distância entre objetos é $d_{12}=d_{21}=2$, indicando que estes dois objetos serão unidos em um *cluster.* Na sequência, calculamos as distâncias:

$d_{(12)3}=min\{d_{13},d_{23}\}=d_{23}=5;$

$d_{(12)4}=min\{d_{14},d_{24}\}=d_{24}=9;$

$d_{(12)5}=min\{d_{15},d_{25}\}=d_{25}=8;$

- Desta forma, obtém-se uma nova matriz de distâncias (MD_2), que será usada na próxima etapa do agrupamento hierárquico:

$$MD_2 = \begin{array}{c} 12 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & \boxed{3} & 0 \end{bmatrix}$$

- Qual o novo *cluster* a ser formado?
- Unindo os objetos 5 e 4 obtemos três clusters: {1,2}, {4,5}, {3}.
- Na sequência, calculamos:

  $d_{(12)3}=min\{d_{13},d_{23}\}=d_{23}=5$ *(na verdade já calculado)*

  $d_{(12)(45)}=min\{d_{14},d_{15},d_{24},d_{25}\}=d_{25}=8$

  $d_{(45)3}=min\{d_{43},d_{53}\}=d_{43}=4$ e obtém-se a seguinte matriz de distâncias:

$$MD_3 = \begin{array}{c} 12 \\ 3 \\ 45 \end{array} \begin{bmatrix} 0 & & \\ 5 & 0 & \\ 8 & \boxed{4} & 0 \end{bmatrix}$$

* Unir *cluster* {3} com {4,5};

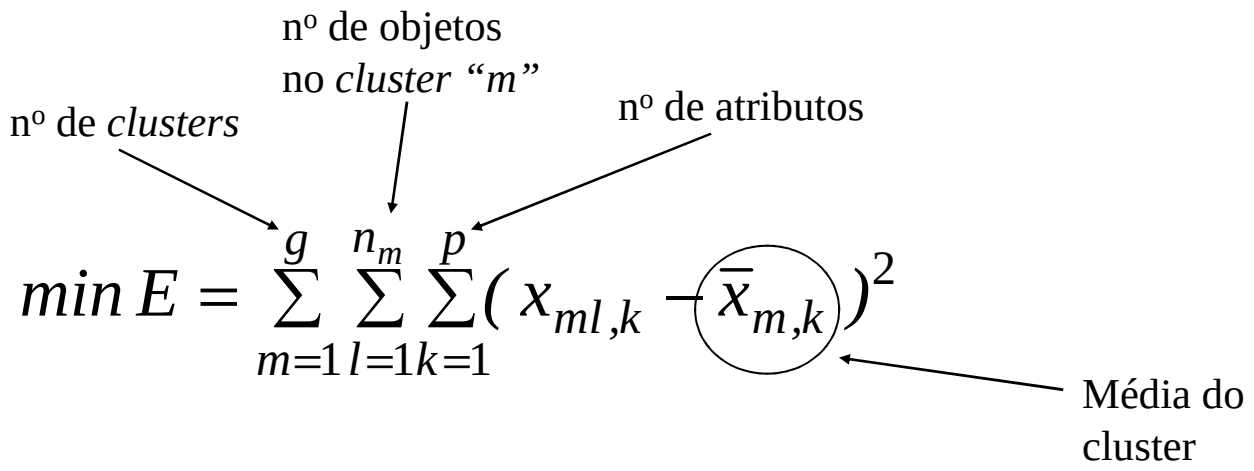
* Finalmente, unir todos os *clusters* em um único *cluster*.

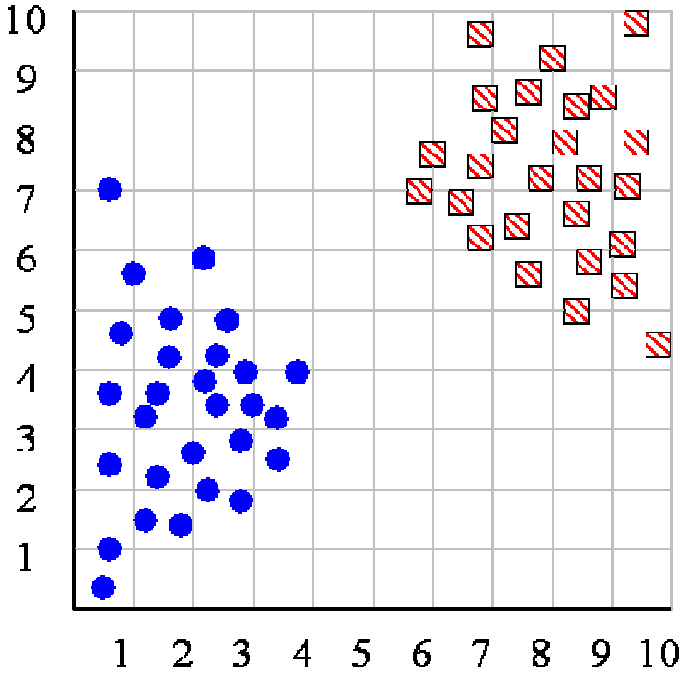
Outras alternativas de métodos aglomerativos:

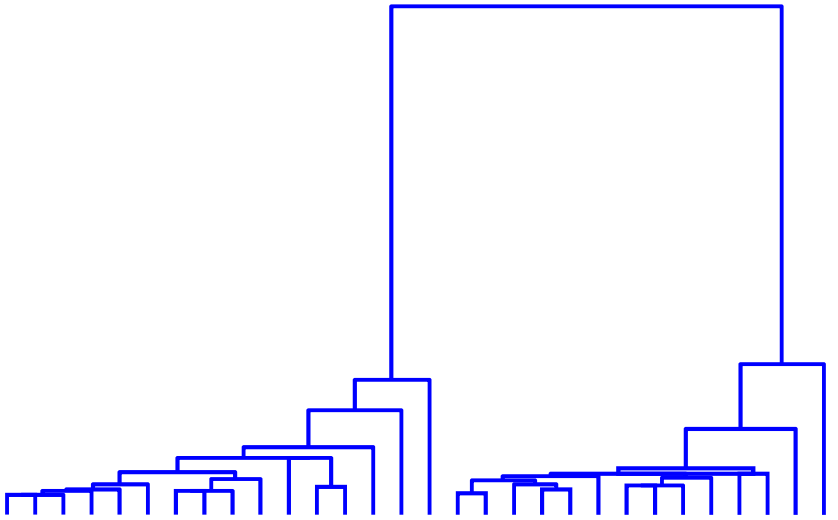
a) Vizinho mais distante (*complete linkage* ou *furthest neighbour*): distância entre *clusters é* medida em função da maior distância entre objetos;

b) Distâncias médias (*average linkage*): distância entre *clusters* é a distância média entre todos os pares de objetos de *clusters* distintos (pode-se também usar os centróides);

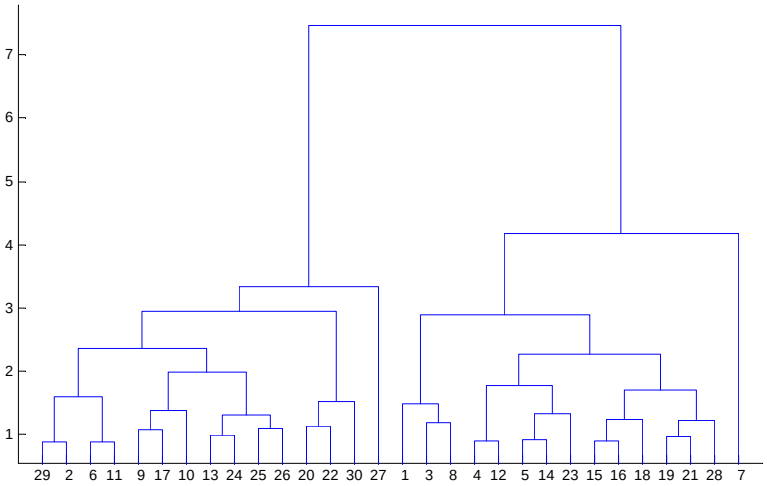
c) Método de Ward (1963):



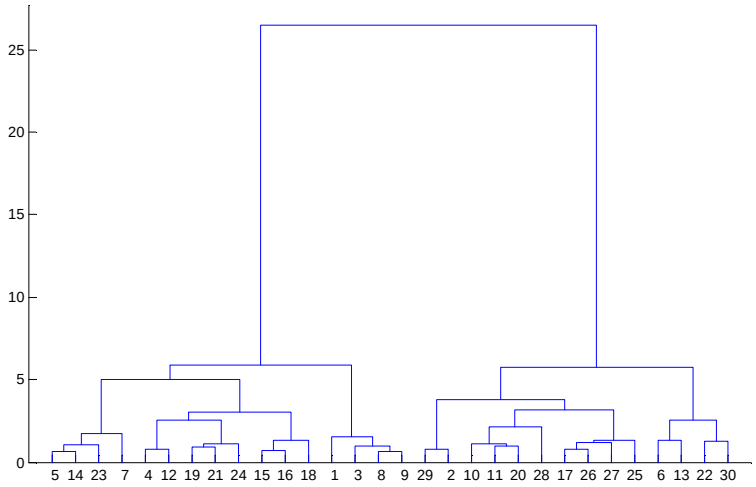




Vinculação simples *(Single linkage)*



Vinculação Média (*Average linkage)*



Vinculação de Ward

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Métodos *Divisivos*:

- Iniciam com um único *cluster,* que é então dividido em dois novos *clusters* ;
- Em cada etapa, cada *cluster* é dividido em dois novos *clusters* ;
- número de modos para dividir *n* objetos em dois *clusters* é $(2^{n-1} – 1)$. Por exemplo, para n=50 existem $5.63x10^{14}$ maneiras de se obter dois clusters!
- Em geral são menos usados do que os métodos aglomerativos.
- Existem versões razoavelmente eficientes, baseados no conceito de entropia, para bases de dados formadas por atributos binários (*monothetic divisive methods).*

- Heurística de MacNaughton-Smith et al. (1964):
  - Para um dado *cluster*, escolher o objeto mais distante dos demais. Este formará o *novo cluster.*
  - Para cada objeto, calculam-se as distâncias médias relativas ao cluster principal e ao *novo cluster,* fazendo com que o objeto mais próximo do *novo cluster* e mais distante do cluster principal faça parte do *novo cluster.*
- Exemplo de aplicação (Everitt et al., Cluster Analysis, 2001):

$$MD = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{bmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{bmatrix}$$

Como encontrar o objeto mais distante dos demais?

- Para este exemplo, objeto "1" é o mais distante *(cluster* A*);*
- Demais objetos formam o *cluster principal* (B);
- *Clusters* obtidos: A={1} e B={2,3,4,5,6,7};
- Consideremos que D_A e D_B são as distâncias médias dos objetos de B em relação aos clusters A e B respectivamente.

Mais próximos de A do que de B

Objetos B	D_A	D_B	D_B-D_A
2	10	25	15,0
3	7	23,4	16,4
4	30	14,8	-15,2
5	29	16,4	-12,6
6	38	19,0	-19,0
7	42	22,2	-19,8

Objeto escolhido para mudar de cluster

Desta forma, obtemos os clusters {1,3} e {2,4,5,6,7}.

Repetindo-se o processo temos:

Objetos B	D_A	D_B	D_B-D_A
2	8,5	29,5	12,0
4	25,5	13,2	-12,3
5	25,5	15,0	-10,5
6	34,5	16,0	-18,5
7	39,0	18,7	-20,3

Mudar para A

Novos clusters: {1,3,2} e {4,5,6,7}.

Próximo passo: todos (D_B-D_A) negativos;

Pode-se continuar o processo em cada *cluster* separadamente...

Como saber se realmente existe um agrupamento (*clustering*)?

- Admissibilidade de Mirkin (1996):
  - Existe um *agrupamento* se todas as distâncias *internas (within-cluster distances)* aos *clusters* são menores do que todas as distâncias *externas (between-cluster distances);*
  - *Compactação e Separação;*
  - Mas não esquecer que existem várias maneiras de se calcular tais distâncias!

Sumário dos Métodos Hierárquicos:

• Não necessitam especificar o número de clusters *a priori*, mas o usuário acaba tendo que escolhê-lo;

• Sofrem do defeito de que não se pode reparar o que foi feito num passo anterior;

• Escalabilidade é um problema: $O(n^2)$, n = *número de objetos*;

• Algoritmo de busca heurístico: ótimos locais são um problema;

• Interpretação dos resultados é intuitiva, mas em geral muito subjetiva. Existem também critérios numéricos para definir o número de *clusters (e.g. Everitt et al., Cluster Analysis, 2001*). Estudaremos alguns destes critérios mais adiante.

• Por ora vamos nos concentrar numa outra grande *classe* de métodos de agrupamento: métodos para obter partições.

Organização

1. Introdução
2. Medidas de Similaridade
3. Métodos de Agrupamento (métodos hierárquicos, **de partição**)
4. Critérios numéricos para definir o número de *clusters*