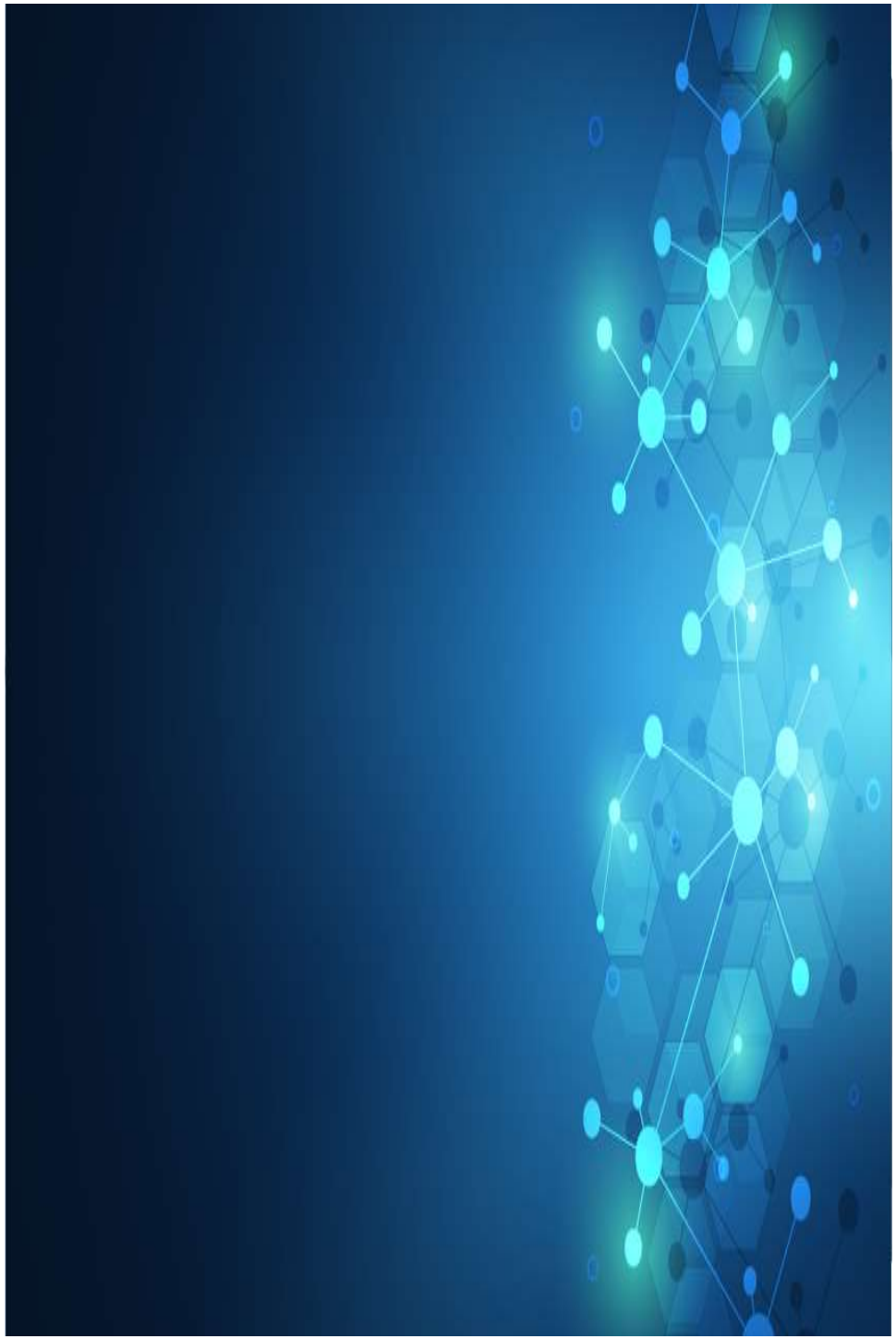


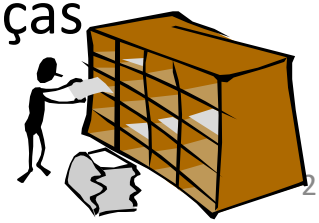
Agrupamento de Dados (*Clustering*)

Huei Diana Lee

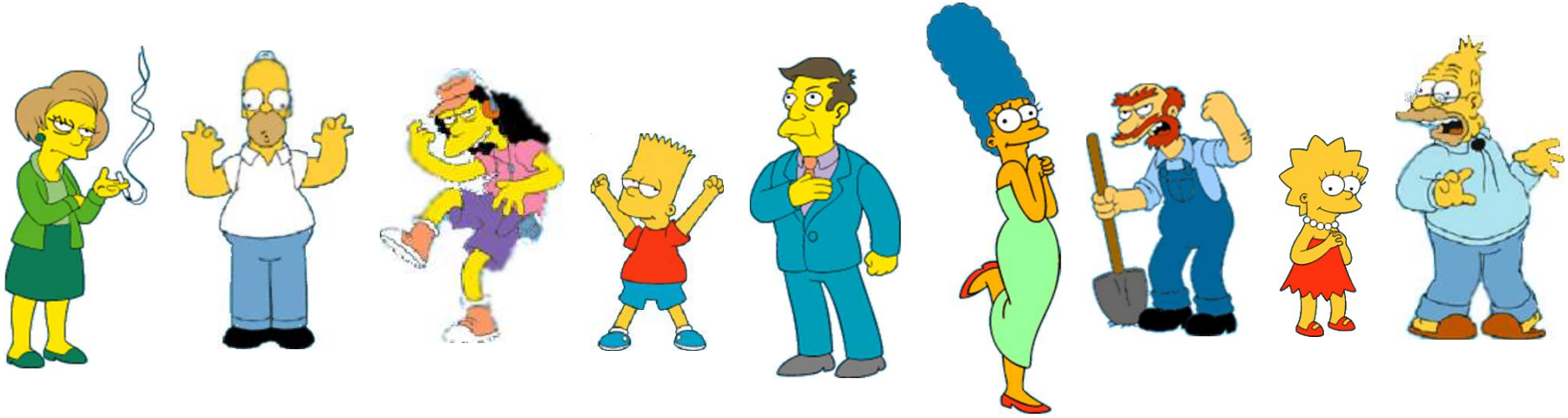
Inteligência Artificial
CECE/UNIOESTE-FOZ

Clustering

- ***Clustering*** (categorização, segmentação ou agrupamento): objetivo de agrupar objetos identificando grupos (clusters) baseados em certos atributos
- Critério de agrupamento:
  - maximizar as similaridades e
  - minimizar as diferenças mediante algum critério
- Exemplo:
  - um conjunto de novas doenças podem ser agrupadas em várias categorias baseadas nas similaridades de seus sintomas, e os sintomas comuns das doenças podem ser usados para descrever um grupo novo de doenças

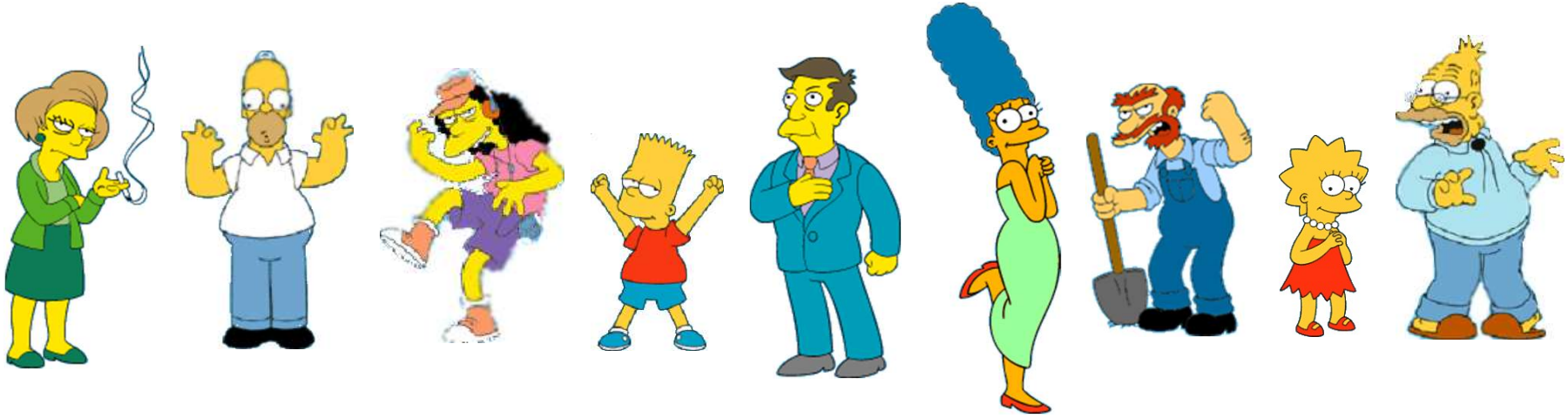


Qual é o agrupamento natural entre esses objetos?

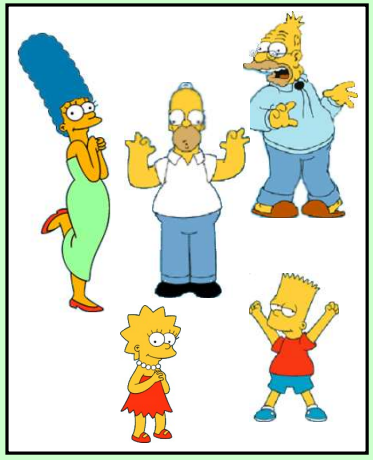


Agrupamento é subjetivo

Qual é o agrupamento natural entre esses objetos?

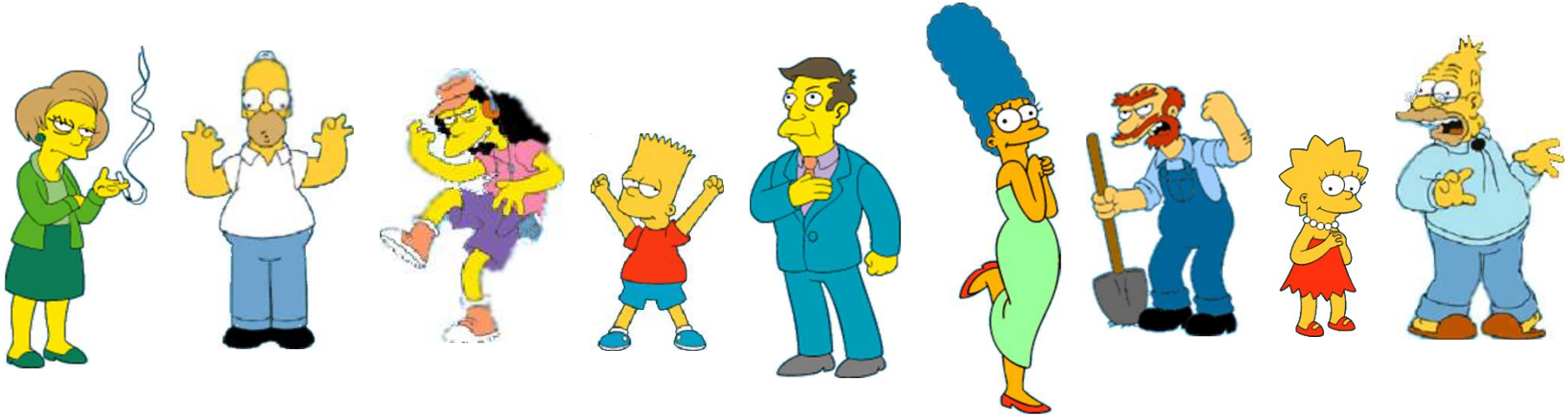


Agrupamento é subjetivo

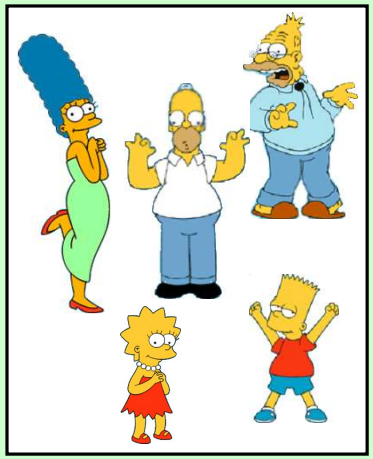


Família Simpson

Qual é o agrupamento natural entre esses objetos?



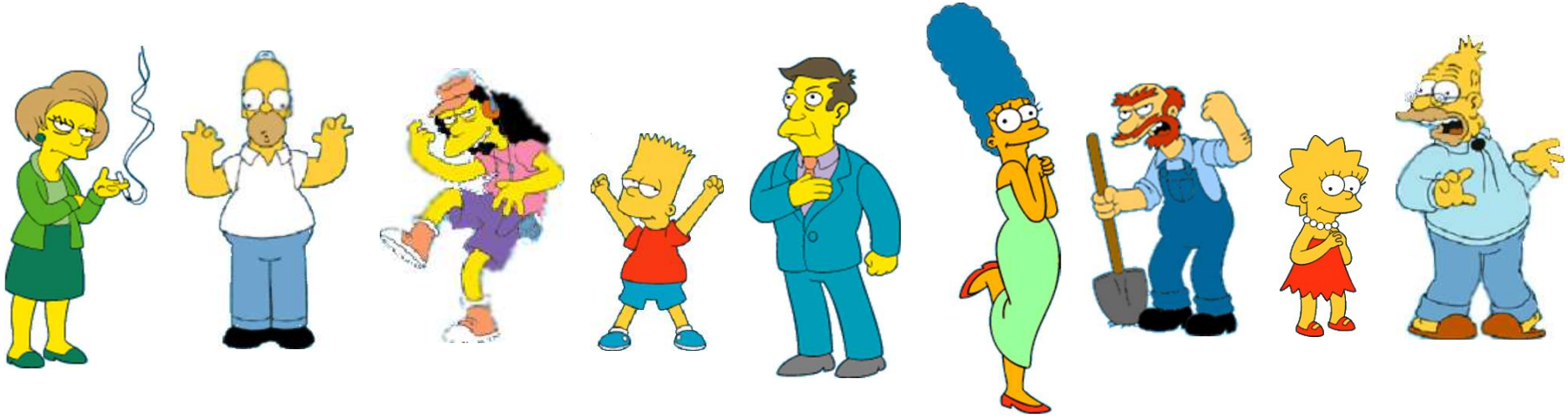
Agrupamento é subjetivo



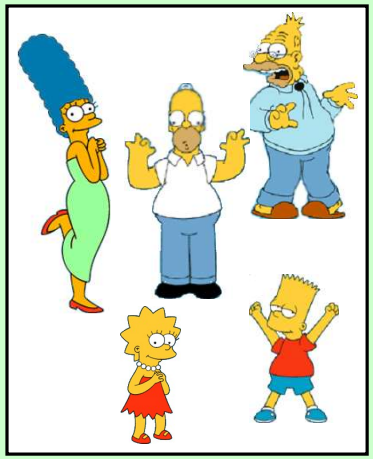
Família Simpson

Empregados da escola

Qual é o agrupamento natural entre esses objetos?

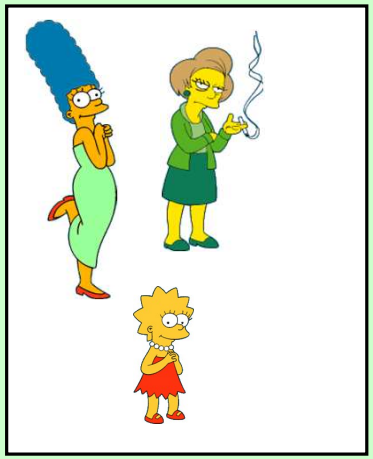


Agrupamento é subjetivo



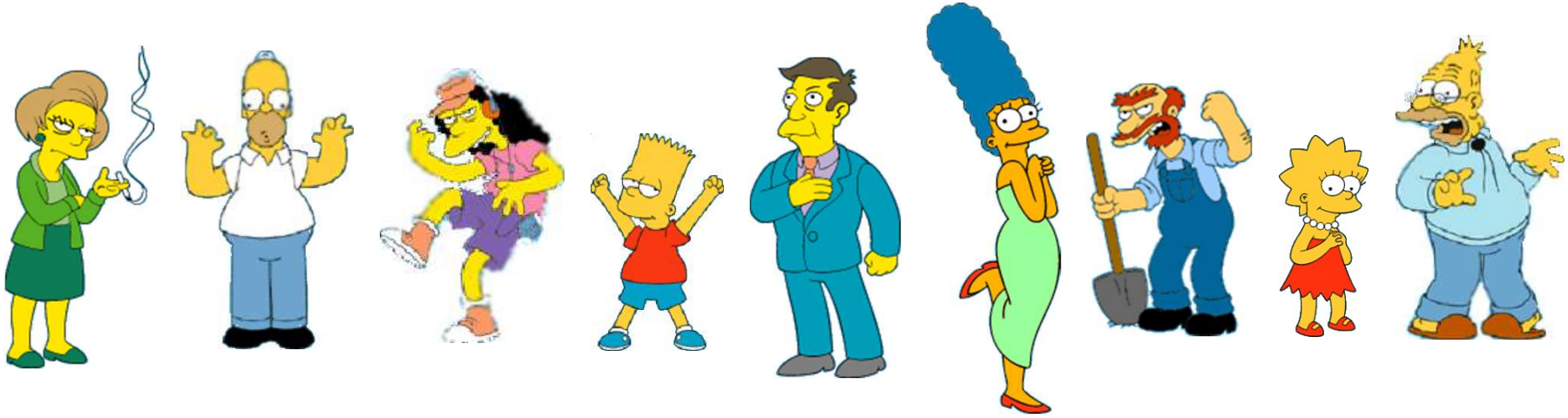
Família Simpson

Empregados da escola

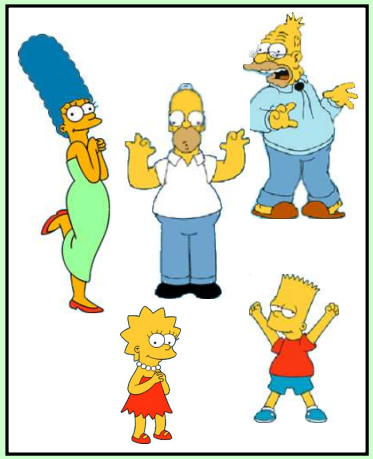


Mulheres

Qual é o agrupamento natural entre esses objetos?

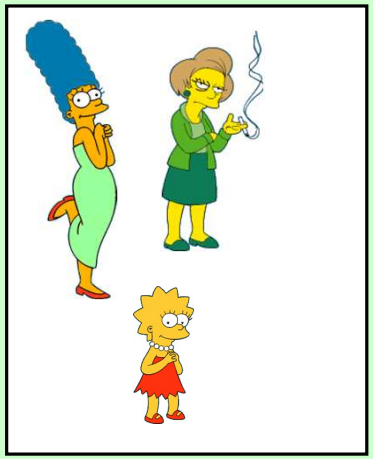


Agrupamento é subjetivo



Família Simpson

Empregados da escola



Mulheres

Homens

Propriedades Desejáveis de um Algoritmo de Agrupamento

- **Escalabilidade** (em termos de espaço e tempo)
- **Habilidade de trabalhar com diferentes tipos de dados**
- **Necessidade mínima de conhecimento** de domínio para determinar os parâmetros de entrada
- **Habilidade de lidar com ruído e *outliers***
- **Insensibilidade** relativa à ordem dos registros de entrada
- **Incorporação de restrições** especificadas pelo usuário
- **Interpretabilidade** e **usabilidade**

Clustering: O que é Similaridade?

Qualidade, caráter ou condição das coisas similares.

Dicionário Houaiss

Similaridade é difícil de definir, mas...
"*Nos sabemos quando a vemos*"

O real significado de similaridade é uma questão filosófica

http://sites.psu.edu/siowfa15/wp-content/uploads/sites/29639/2015/11/bp2-post-3-dogs.jpg

Clustering: O que é Similaridade?

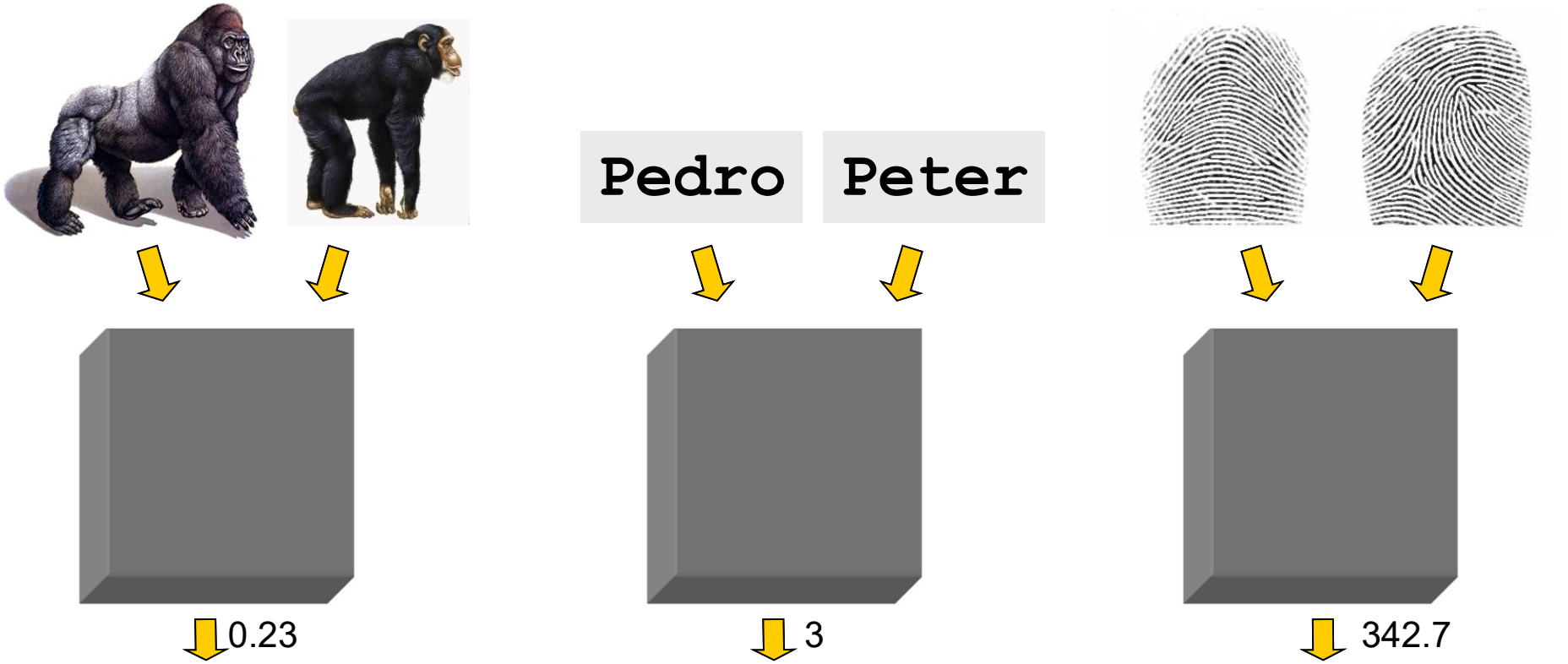
http://rockntech.com.br/cachorros-que-se-parecem/
http://dogbreedsjournal.com/best-dogs-for-kids/

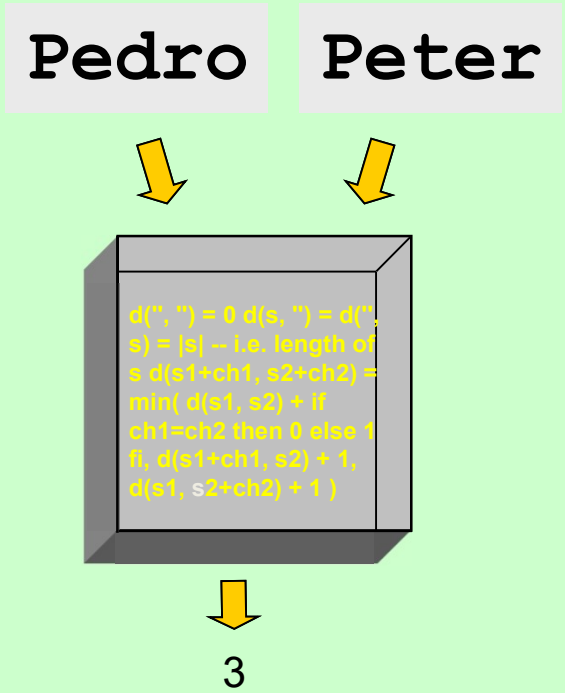
Calculando a Distância

- A **distância** é o método mais natural para dados numéricos
- Valores pequenos indicam maior similaridade
- Métricas de Distância
  - Euclideana
  - Manhattan
  - Entre outras
- Não generaliza muito bem para dados não numéricos
  - Qual a distância entre “masculino” e “feminino”?

Definindo Medidas de Distância

Definição: Sejam O_1 e O_2 dois objetos de um universo de possíveis objetos. A distância (dissimilaridade) entre O_1 e O_2 é um número real denotado por $D(O_1,O_2)$





- As caixas pretas contém alguma função de duas variáveis
- Essas funções podem ser simples ou complexas
- Em qualquer caso é natural perguntar, quais propriedades essas funções devem possuir

Quais propriedades uma medida de distância deve possuir?

• $D(A,B) = D(B,A)$	*Simetria*
• $D(A,A) = 0$	*Constância de auto-similaridade*
• $D(A,B) = 0$ sse A = B	*Positividade*
• $D(A,B) \leq D(A,C) + D(B,C)$	*Desigualdade Triangular*

Motivos das Propriedades Desejáveis de Medidas

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

Caso contrário você poderia afirmar que "Alex parece com Bob, mas Bob não parece com Alex."

$D(A,A) = 0$ Constância de Auto-simetria

Caso contrário você poderia afirmar que"Alex parece mais com Bob, do que o próprio Bob."

$D(A,B) = 0$ sse $A=B$ Positividade

Caso contrário existiriam objetos no seu mundo que são diferentes, mas você não consegue diferenciá-los.

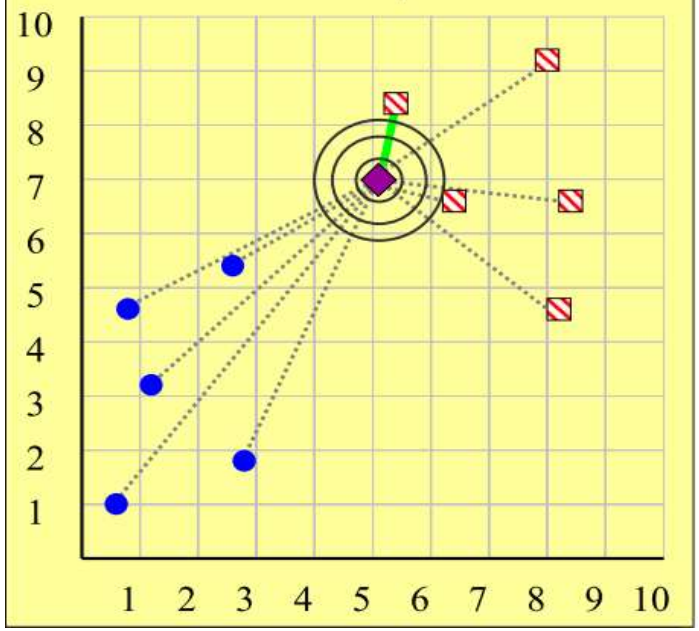
$D(A,B) \leq D(A,C) + D(B,C)$ Desigualdade Triangular

Caso contrário você poderia afirmar que "Alex é parecido com Bob, e Alex é parecido com Carl, mas Bob não se parece com Carl."

Medidas de Distância

- Norma Lp

$$L_p(x,y) = \left(\sum_{i=1}^{W} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

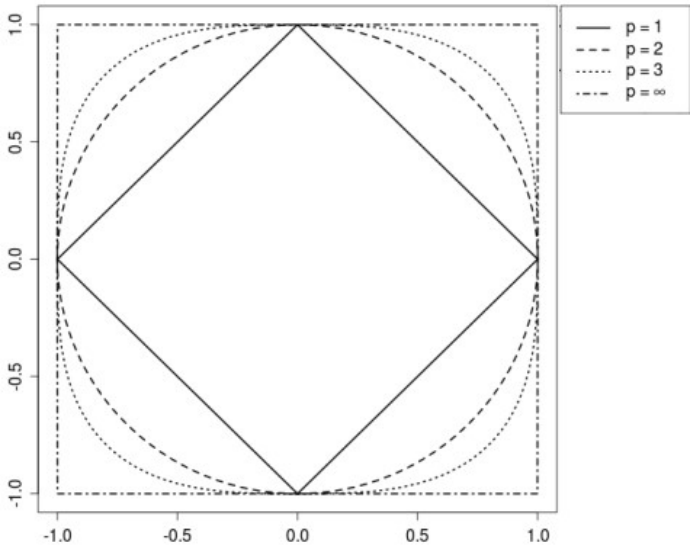


$p = 1$: Manhattan, também conhecida como *City Block* (L_1);

$p = 2$: Euclidiana (L_2);

$p = 3$: Métrica L_3 (L_3);

$p = \infty$: Chebychev, também denominada Infinita (L_∞).



Medidas de Distância

Coeficiente de Jacard

Estatística usada para comparar similaridade e diversidade entre conjuntos

$sim(t_i , t_j)$

= (número de atributos em comum)/

(número total de atributos em ambos)

= (intersecção entre t_i e t_j) / (união entre t_i e t_j)

Medidas de Distância

Coeficiente de Jacard

Coeficiente de Jacard

$$s_{ij} = \frac{p}{p+q+r}$$

onde:

p = no. de variáveis positivas para ambos

q = no. de variáveis positivas no i-ésimo objeto e negativas para j-ésimo objeto

r = no. de variáveis negativas no i-ésimo objeto e positivas no j-ésimo objeto

s = no. de variáveis negativas para ambos

$t = p+q+r+s$ = número total de variáveis

Distância de Jaccard pode ser obtida de:

$$d_{ij} = 1 - s_{ij} = 1 - \frac{p}{p+q+r} = \frac{p+q+r-p}{p+q+r} = \frac{q+r}{p+q+r}$$

Medidas de Distância

Coeficiente de Jacard - Exemplo

Fruta	Formato Esférico	Doce	Azedo	Crocante
Object A=Maçã	Yes(1)	Yes(1)	Yes(1)	Yes(1)
Object B=Banana	No(0)	Yes(1)	No(0)	No(0)

Cada objeto representado por quatro variáveis -> objeto possui quatro dimensões

Coordenadas Maçã = (1,1,1,1)
Coordenadas Banana = (0,1,0,0)

p=1, q=3, r=0 e s=0

p = no. de variáveis positivas para ambos
q = no. de variáveis positivas no i-ésimo objeto e negativas para j-ésimo objeto
r = no. de variáveis negativas no i-ésimo objeto e positivas no j-ésimo objeto

Coeficiente de Jaccard entre Maçã e Banana =1/(1+3+0)= 1/4

Distância de Jaccard entre Maçã e Banana =1-(1/4) = 3/4

Medidas de Distância

Distância de Hamming

- Em teoria da informação:

  Distância de Hamming (str1, str2)

  = Número de posições nas quais símbolos correspondentes são diferentes

  = Número mínimo de substituições necessárias para mudar uma string para a outra ou

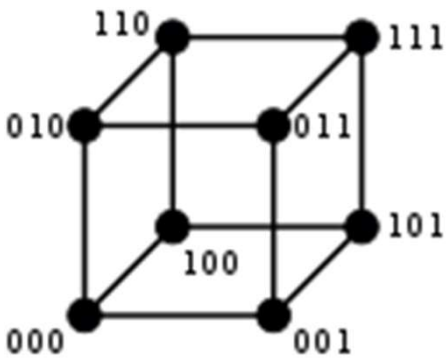
  = Número de erros que poderiam ter transformado uma string na outra

- Uma das principais aplicações: teoria de codificação

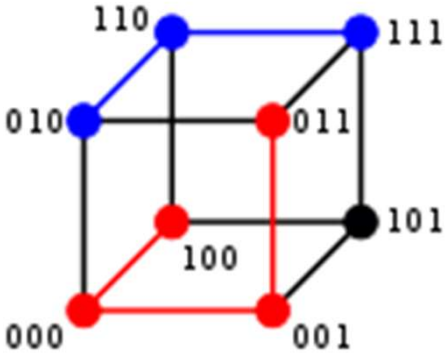
https://en.wikipedia.org/wiki/Hamming_distance

Medidas de Distância

Distância de Hamming



Cubo binário de 3-bits para encontrar a Distância de Hamming



Dois exemplos:

- 100→011 possui distância 3 (vermelho)
- 010→111 possui distância 2 (azul)

hnn("karolin“ , "kathrin“) = 3
hnn("karolin“ , "kerstin“) = 3

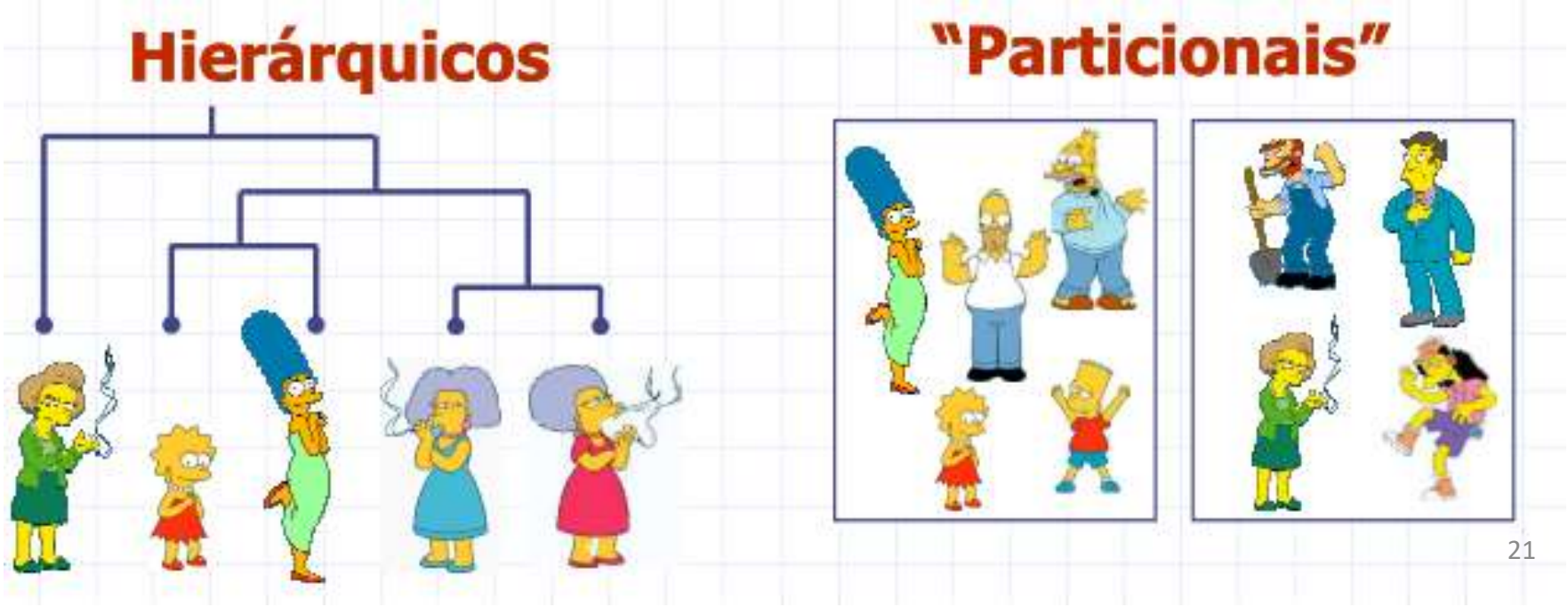
hnn(**1011101 , 1001001)** = 2

hnn(**2173896 , 2233796)** = 3

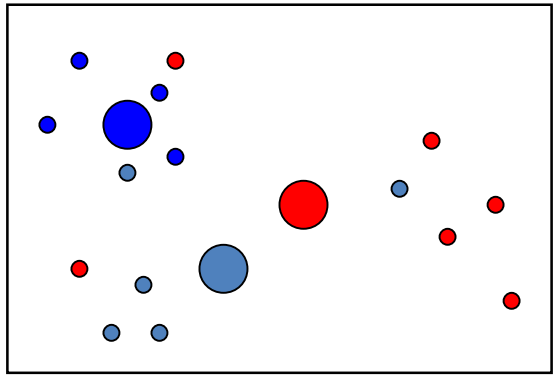
https://en.wikipedia.org/wiki/Hamming_distance

Clustering

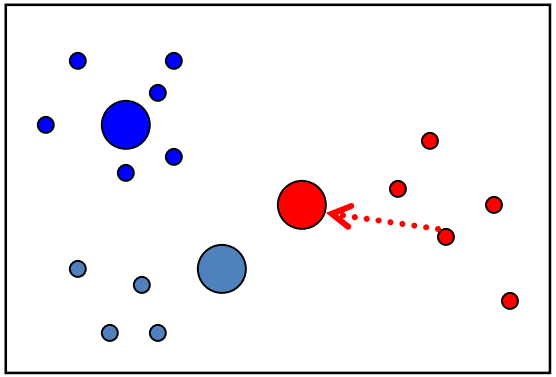
- **Estratégias de Clustering:**
  - **Particionais:** construir várias partições e avaliá-las segundo algum critério (ex.: K-means)
  - **Hierárquicos:** criar uma decomposição hierárquica do conjunto de objetos usando algum critério



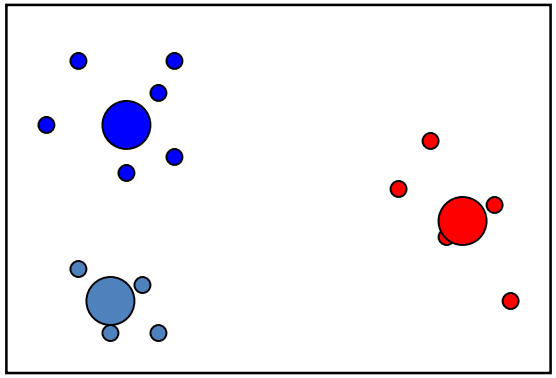
K-means: Exemplo, K = 3



Passo 1: Escolha aleatória de clusters e cálculo dos centróides (círculos maiores)

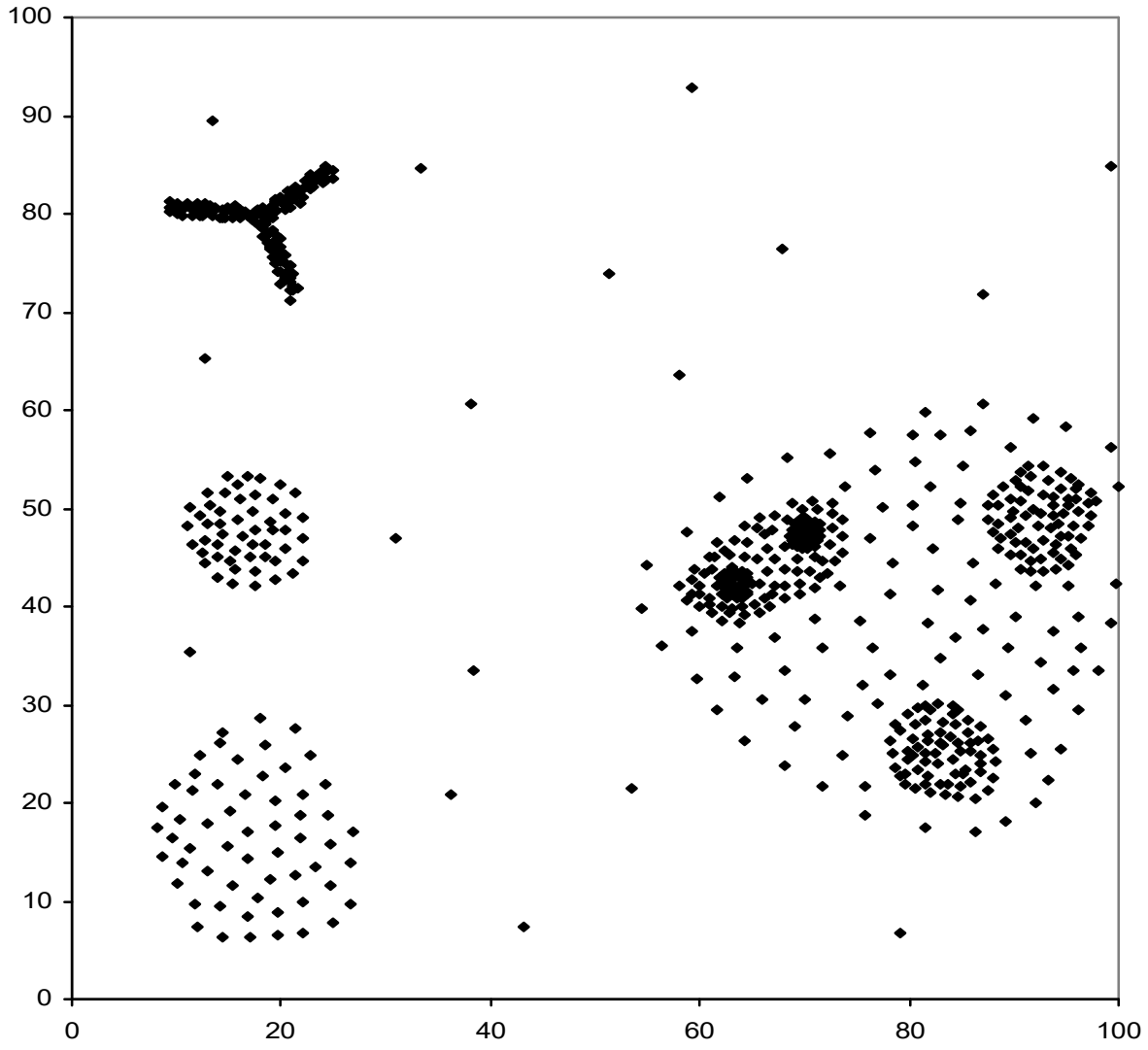


Passo 2: Atribua cada ponto ao centróide mais próximo

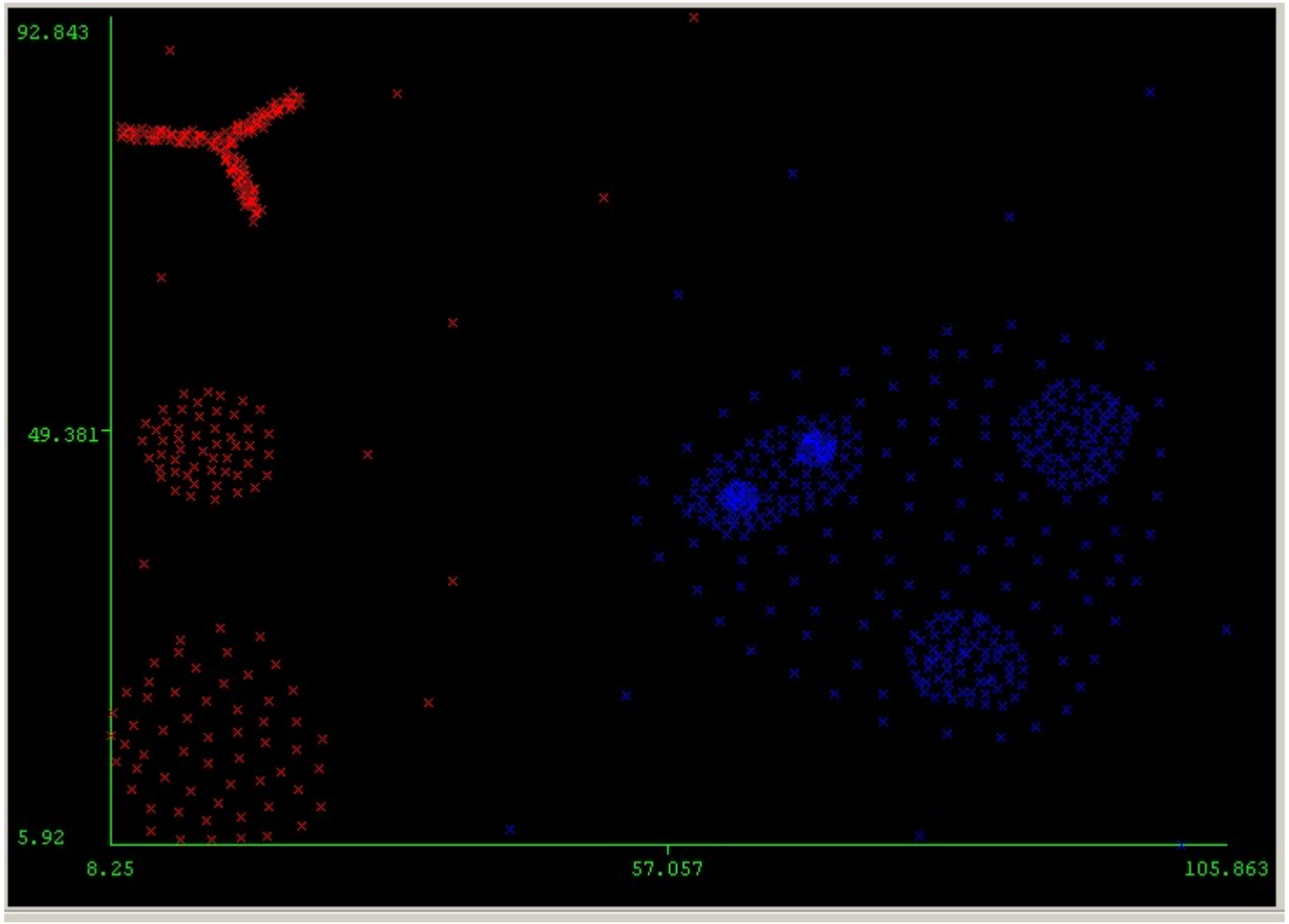


Passo 3: Recalcule centróides (neste exemplo, a solução é agora estável)

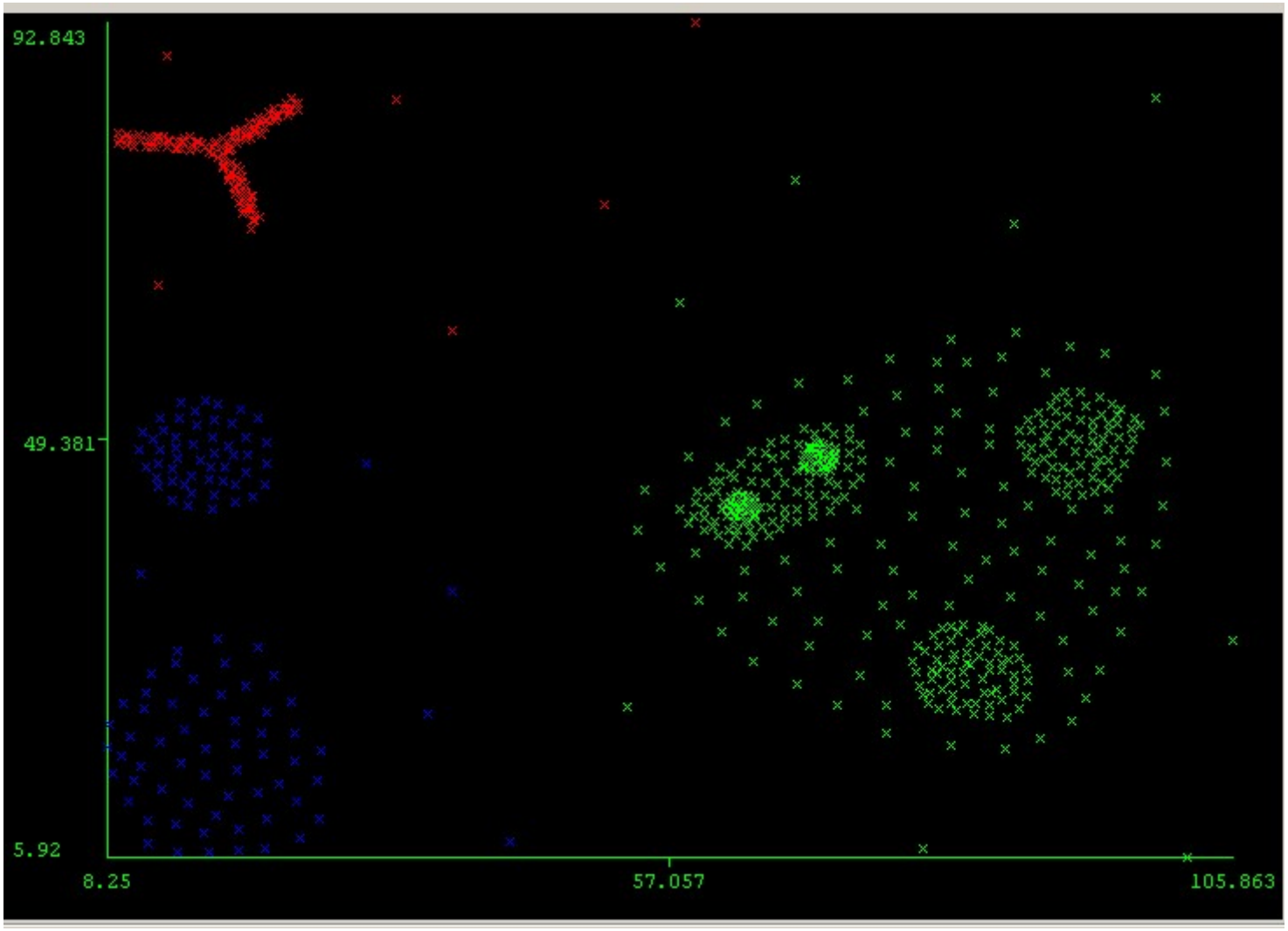
K-means: Exemplo



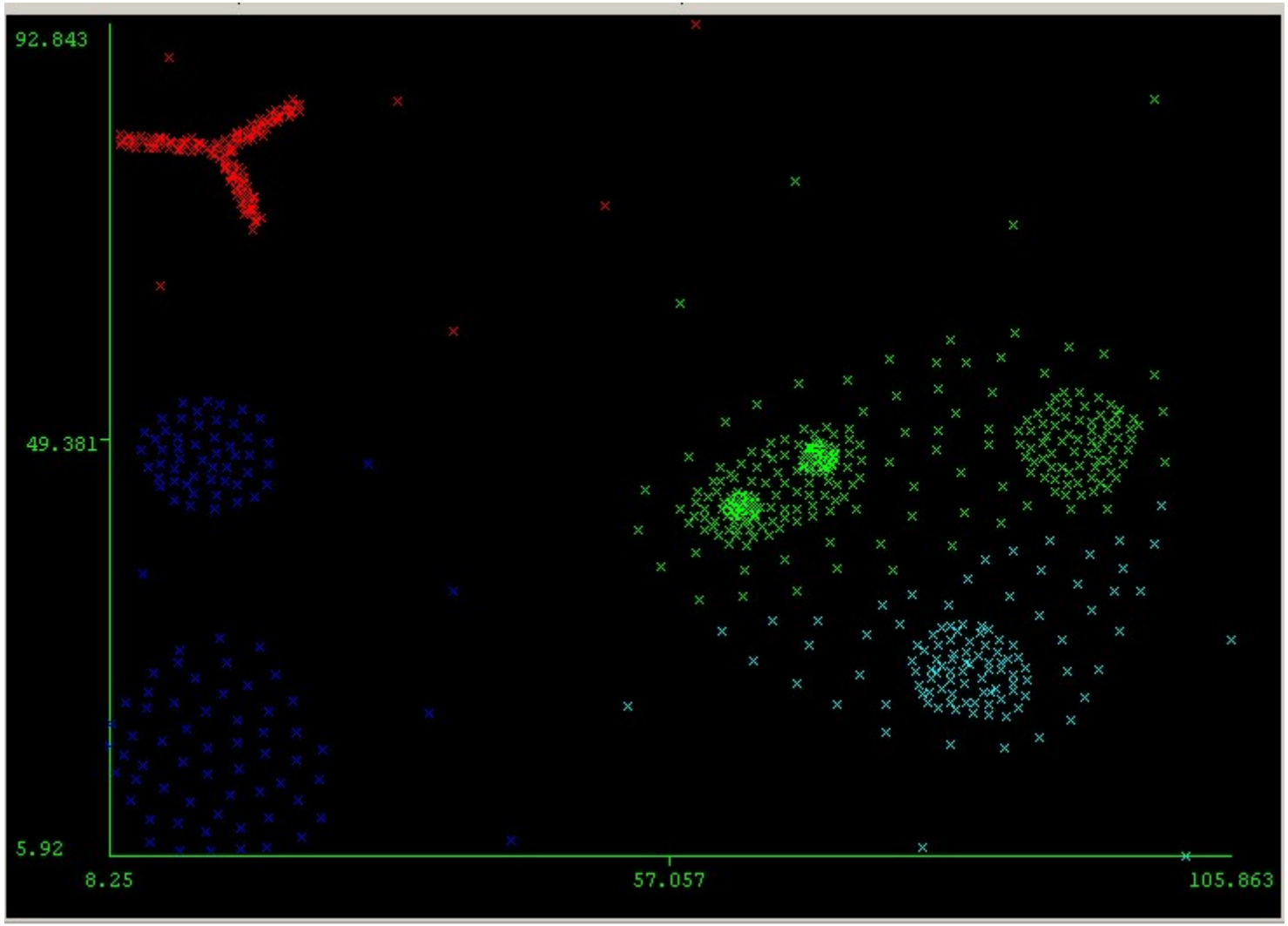
K-means: Exemplo, K=2



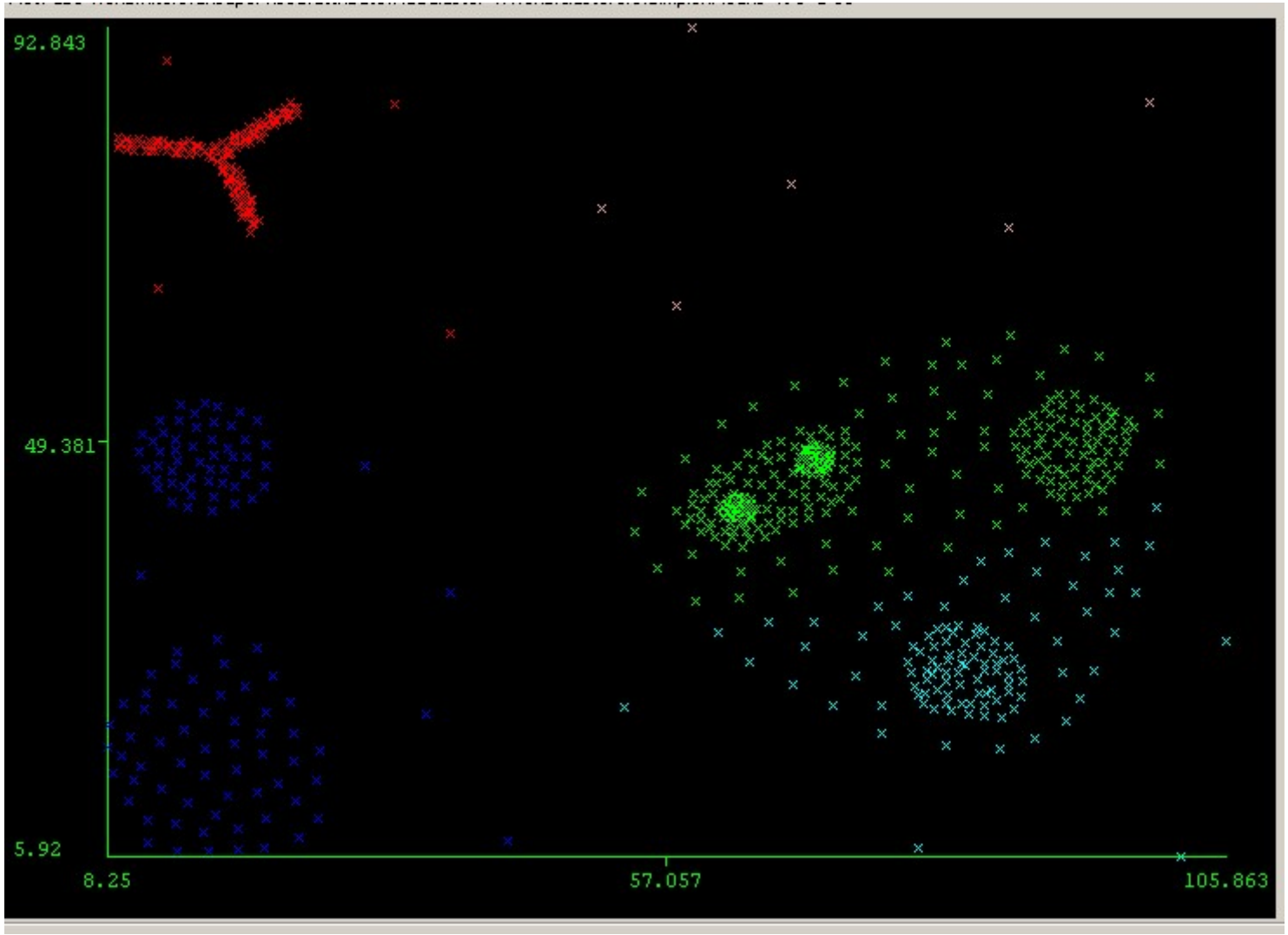
K-means: Exemplo, K=3



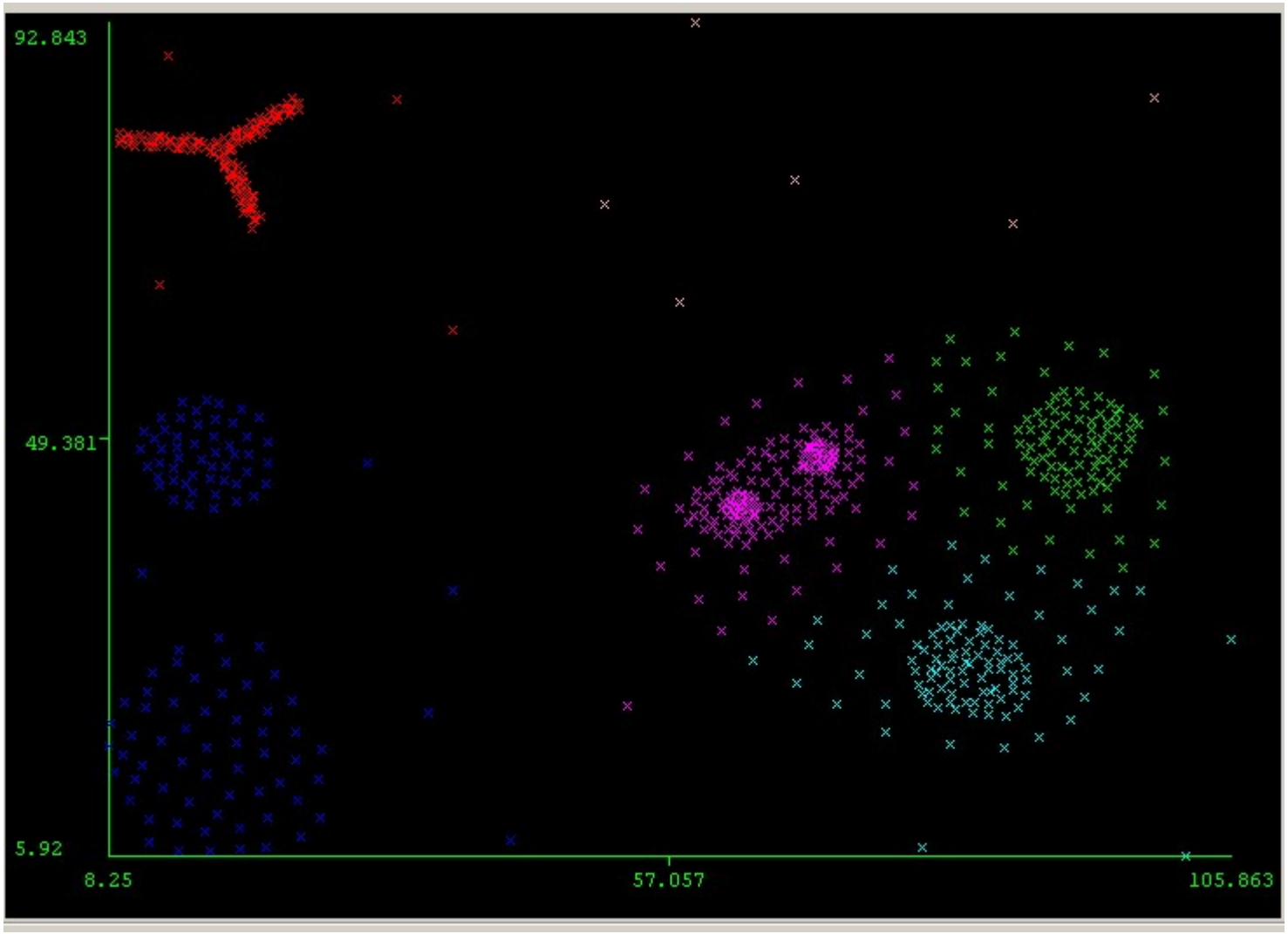
K-means: Exemplo, K=4



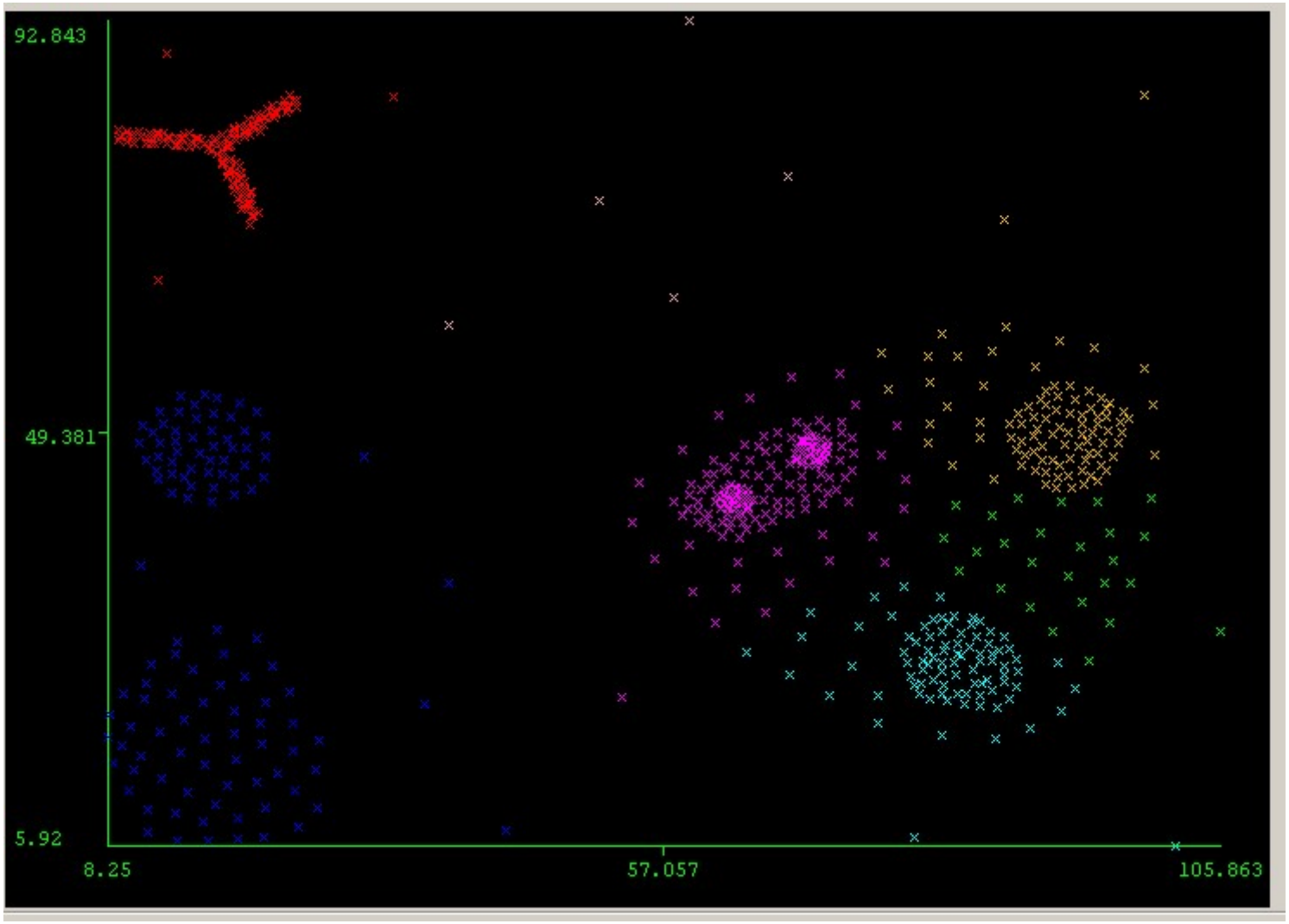
K-means: Exemplo, K=5



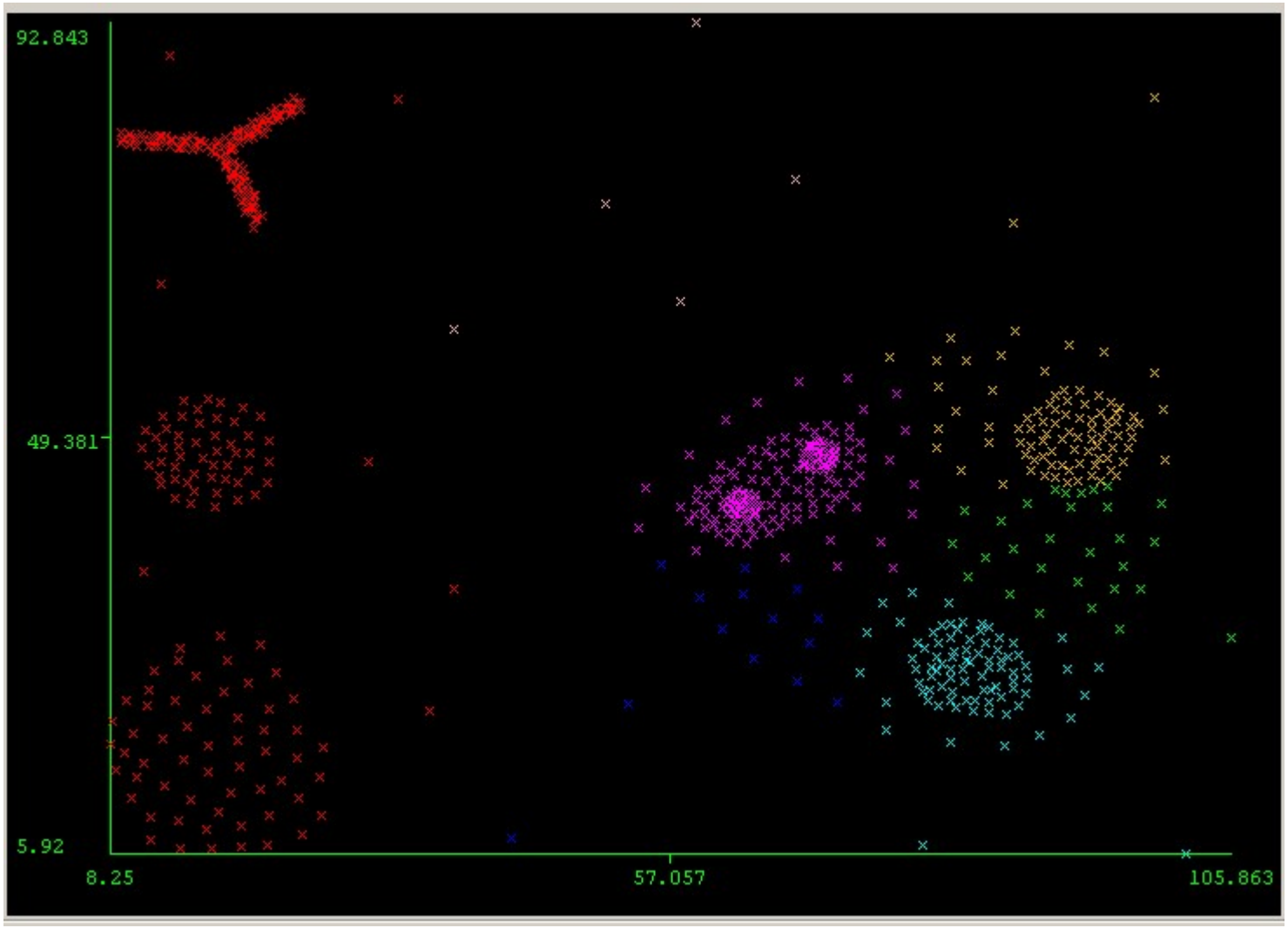
K-means: Exemplo, K=6



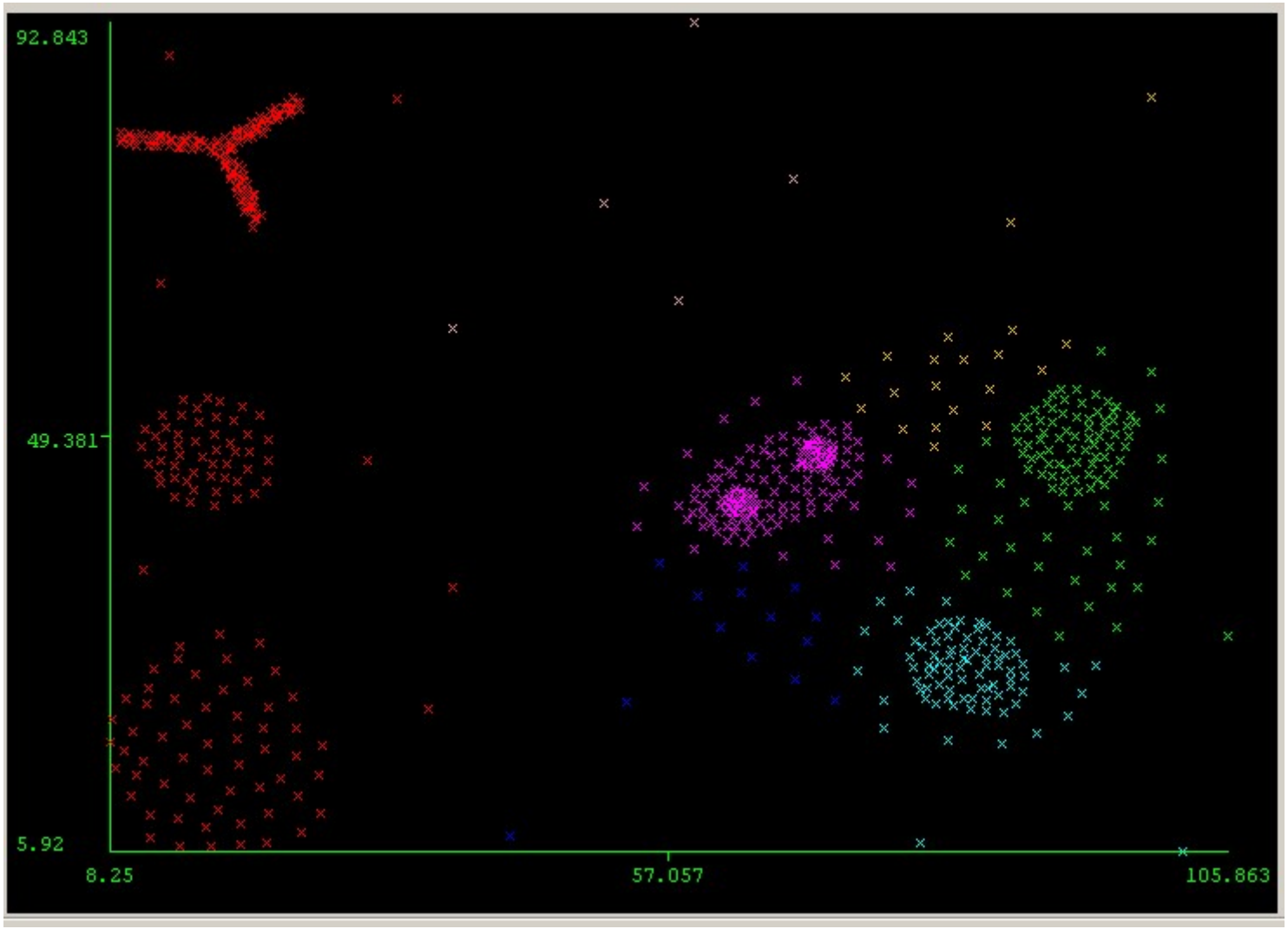
K-means: Exemplo, K=7



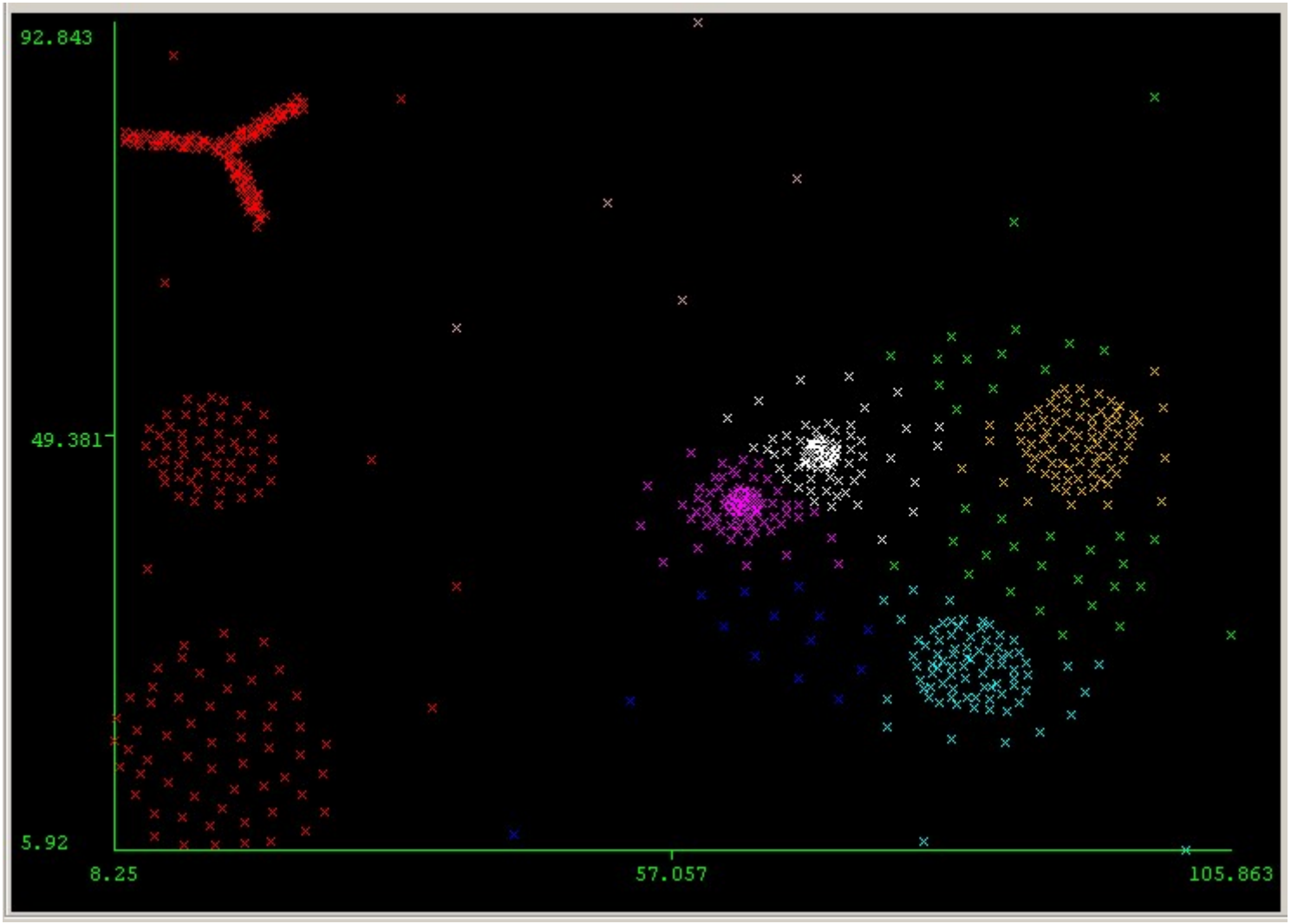
K-means: Exemplo, K=8



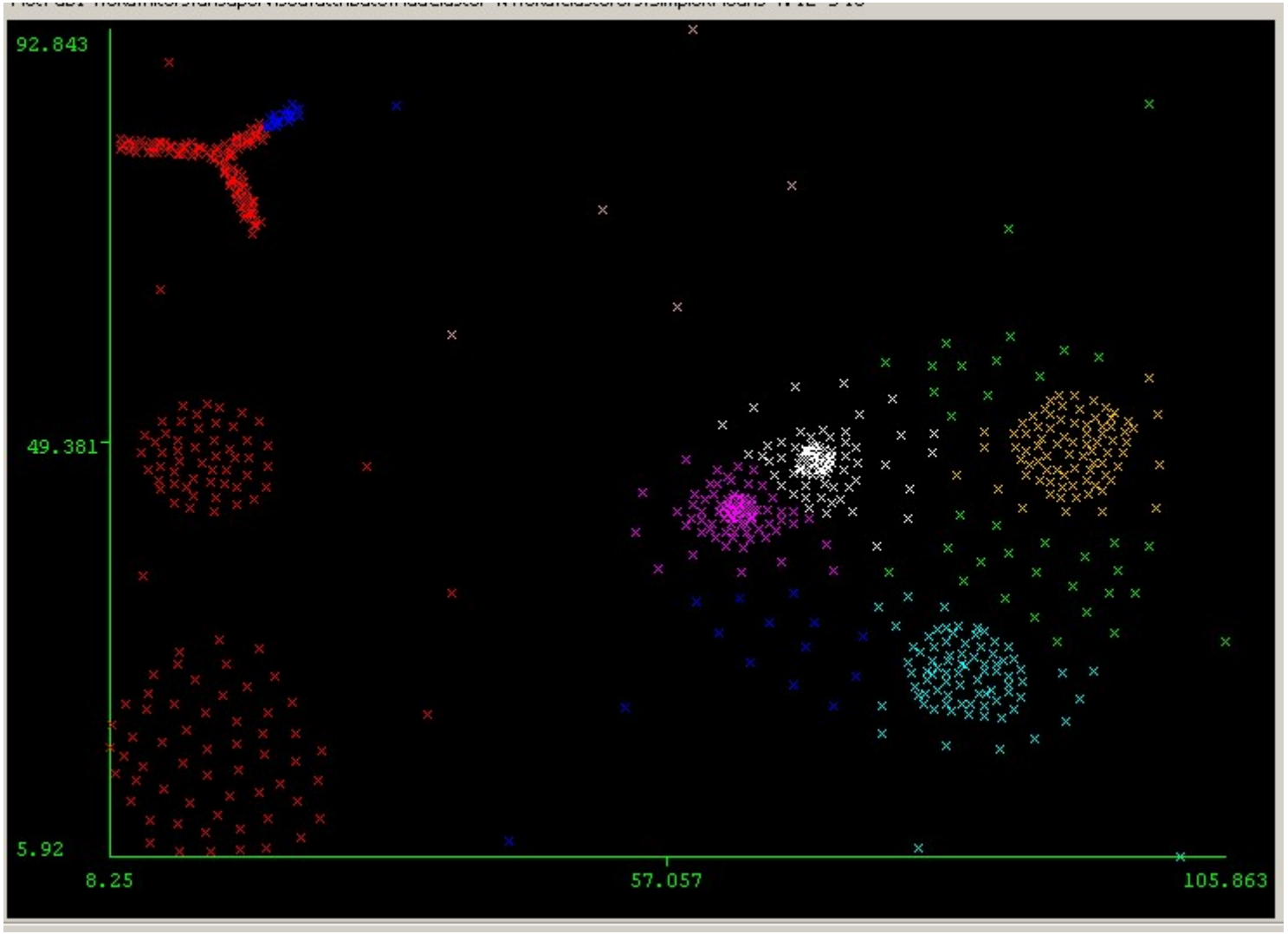
K-means: Exemplo, K=9



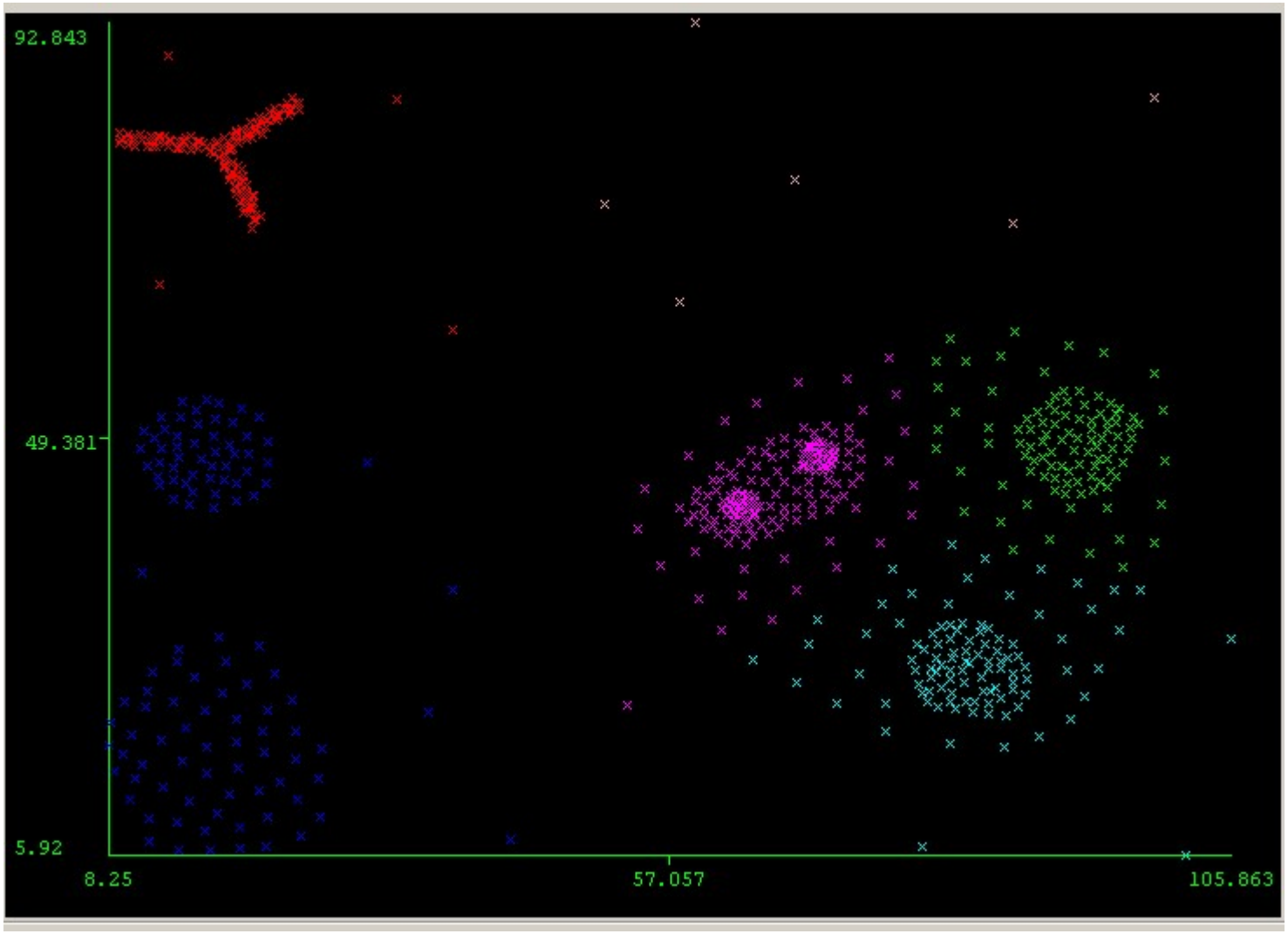
K-means: Exemplo, K=10



K-means: Exemplo, K=12

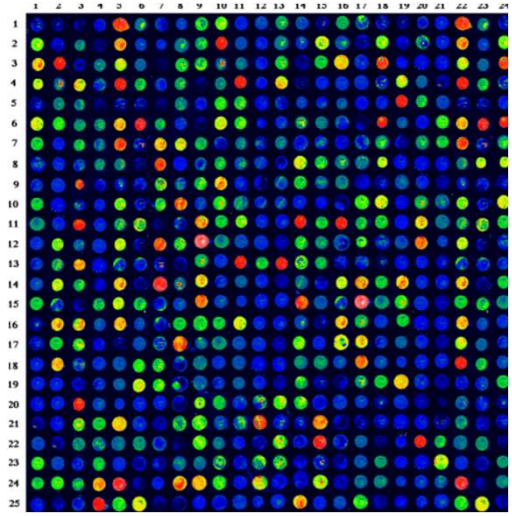


K-means: Exemplo, K=6

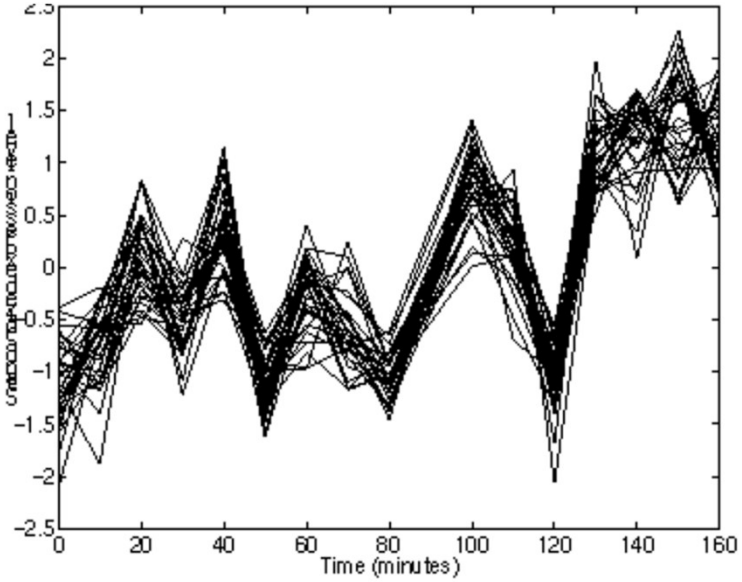


K-means
Exemplo de Aplicação

- Clustering de Genes
  - Uma série de experimentos de microarray medindo a expressão de um conjunto de genes a intervalos regulares de tempo numa célula
  - Normalização permite comparação entre microarrays
  - Produz clusters de genes que variam de forma similar ao longo do tempo
  - Hipótese: genes que variam da mesma forma podem ser/estar co-regulados



Amostra de um Array. Linhas são genes e colunas são pontos no tempo



Um cluster de genes co-regulados

Comentários sobre o Método *K-Means*

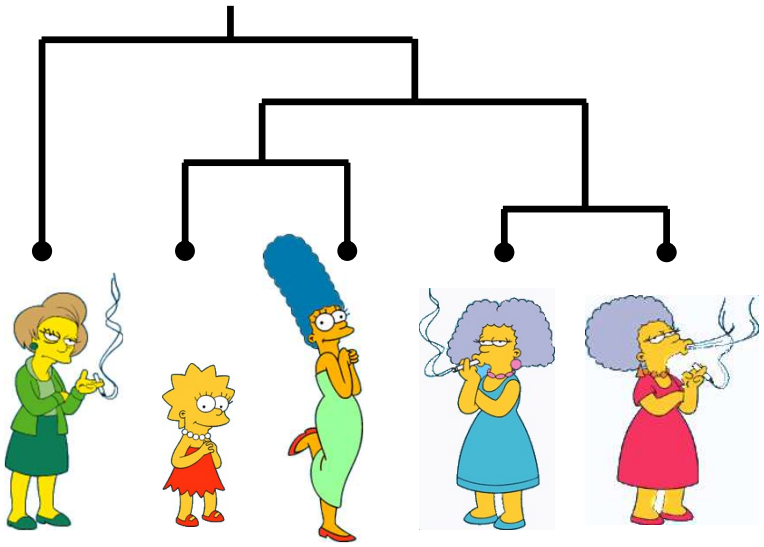
- Pontos fortes
  - *Relativamente eficiente*: *O*(*t.k.n*)

  onde *n* = o # de objetos, *k* = o # de *clusters*, e *t = o* # iterações. Normalmente, *k e t* << *n*
  - Freqüentemente termina em um *ótimo local*. O *ótimo global* pode ser encontrado utilizando técnicas como: *annealing determinístico* e *algoritmos genéticos*
- Pontos fracos
  - Aplicável somente quando a *média* pode ser definida. E sobre dados categóricos?
  - É necessário especificar *k,* o *número* de *clusters*, *a priori*
  - Incapaz de lidar com ruído e *outliers*
  - Não adequado para descobrir *clusters* com formatos *não-convexos*

Agrupamento Hierárquico

Número de dendogramas com n folhas = $(2n-3)!/[(2^{(n-2)})\,(n-2)!]$

Número de Folhas	Número de Possíveis Dendogramas
2	1
3	3
4	15
5	105
...	…
10	34.459.425

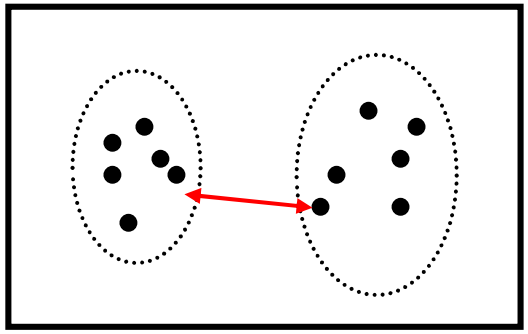


Como não podemos testar todas as possíveis árvores, nós podemos realizar uma busca heurística usando…

Bottom-Up (aglometrativo): Começando com cada item em seu próprio *cluster*, encontrar o melhor par para aglomerar em um novo *cluster*. Repetir até que todos os *clusters* tenham sido aglomerados em um único

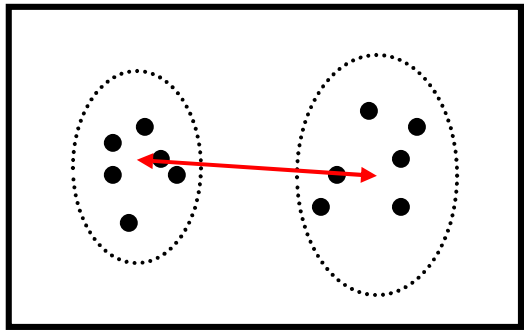
Top-Down (divisivo): Começando com todos os dados em um único *cluster*, considerar cada possível maneira de dividir o *cluster* em dois. Escolher a melhor divisão e repetir recursivamente em ambos os lados

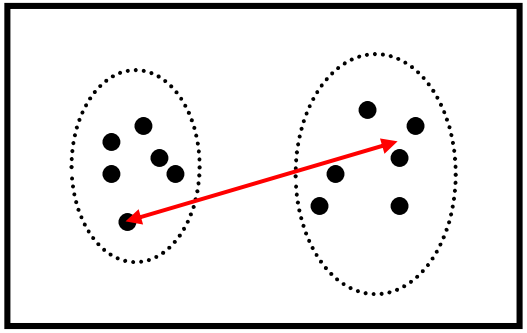
Clustering: Agrupando Clusters



Single Link: Distância entre dois clusters é a distância entre os pontos mais próximos. Também chamado “agrupamento de vizinhos”

Average Link: Distância entre clusters é a distância entre os centróides

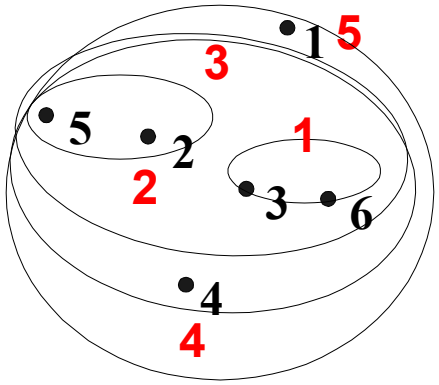




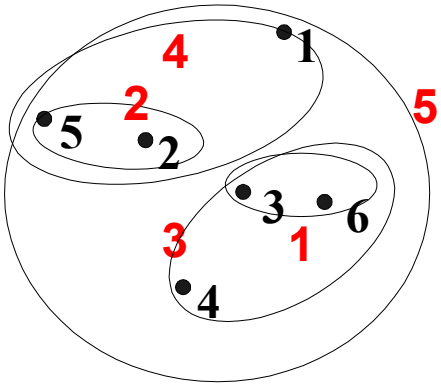
Complete Link: Distância entre clusters é a distância entre os pontos mais distantes

Tipos de agrupamento hierárquico

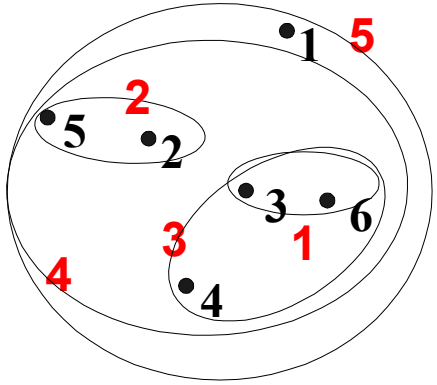
Single-link



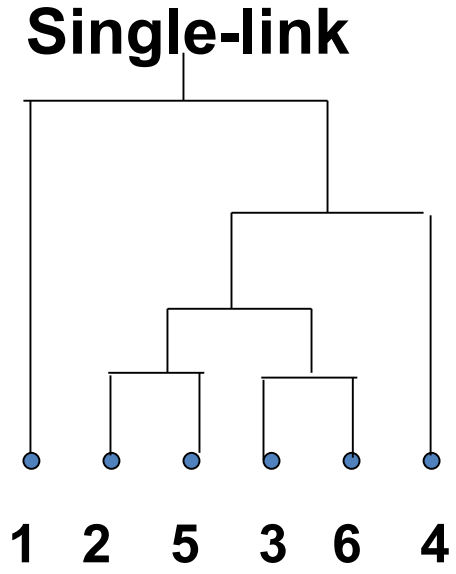
Complete-link

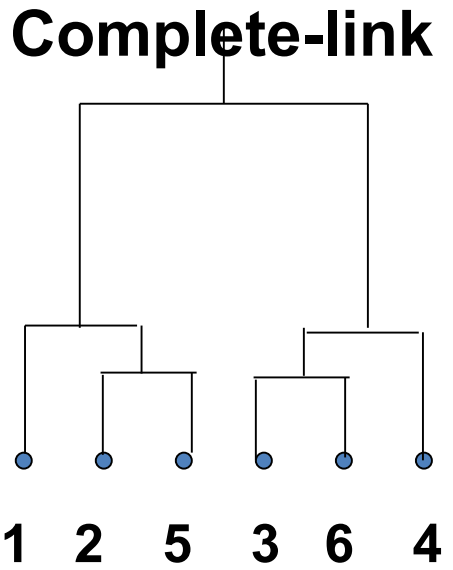


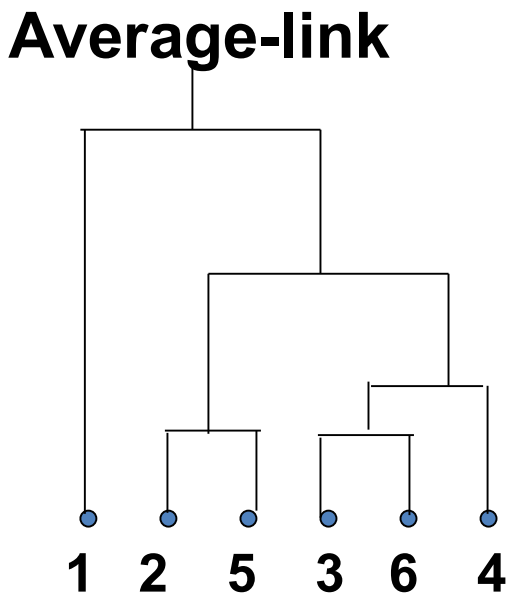
Average-link



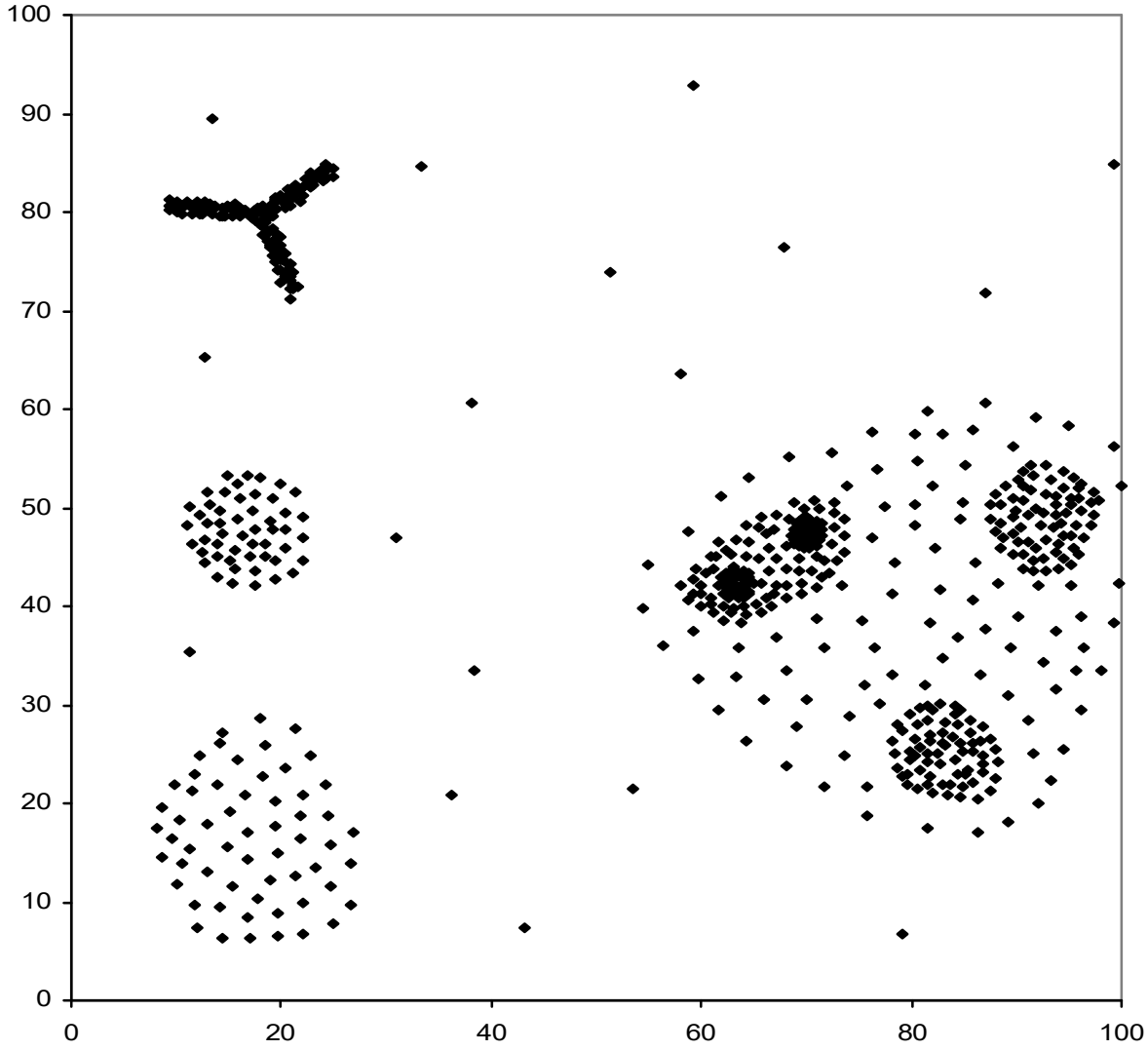
Tipos de agrupamento hierárquico



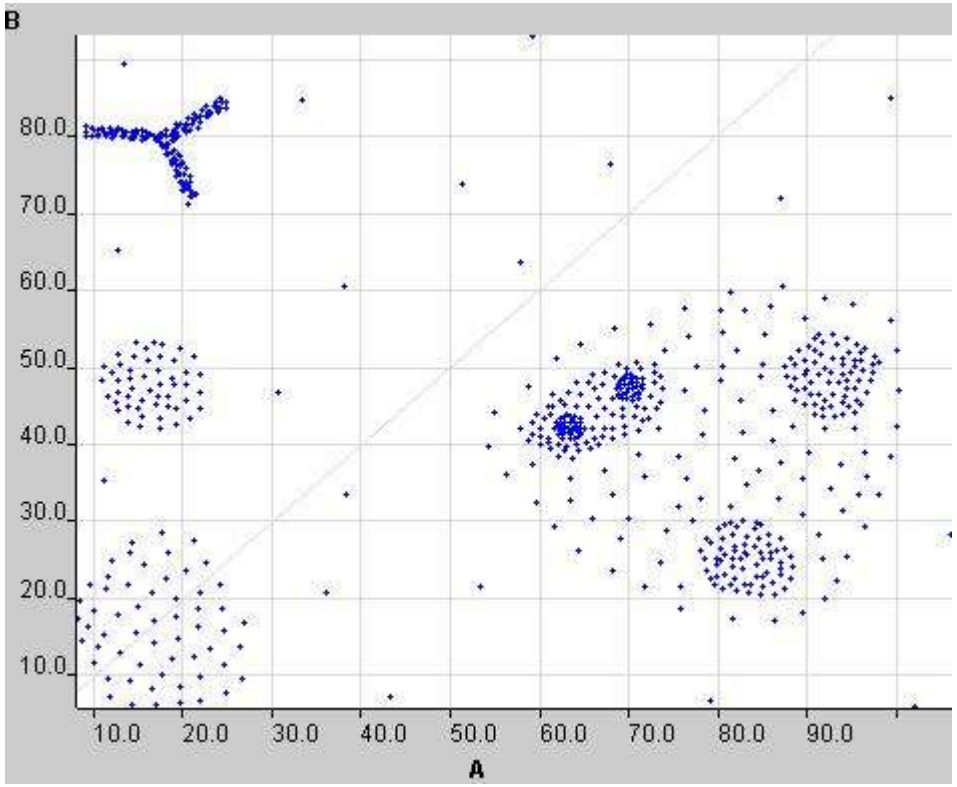


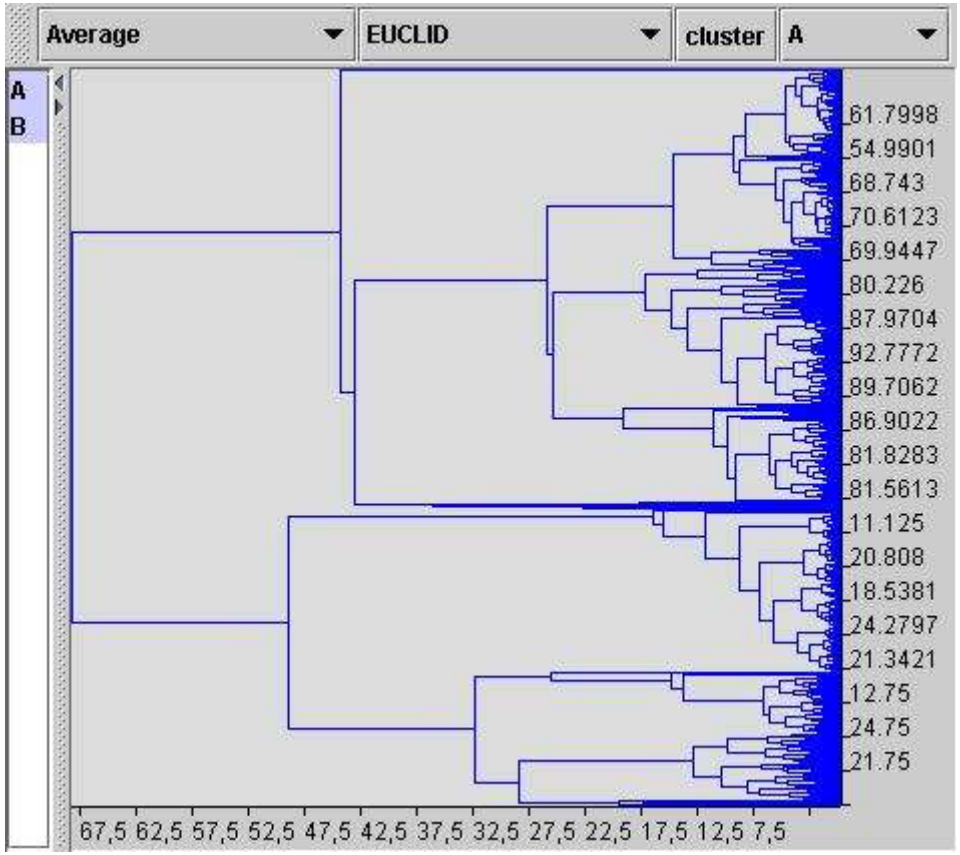


Clustering Hierárquico: Exemplo 1

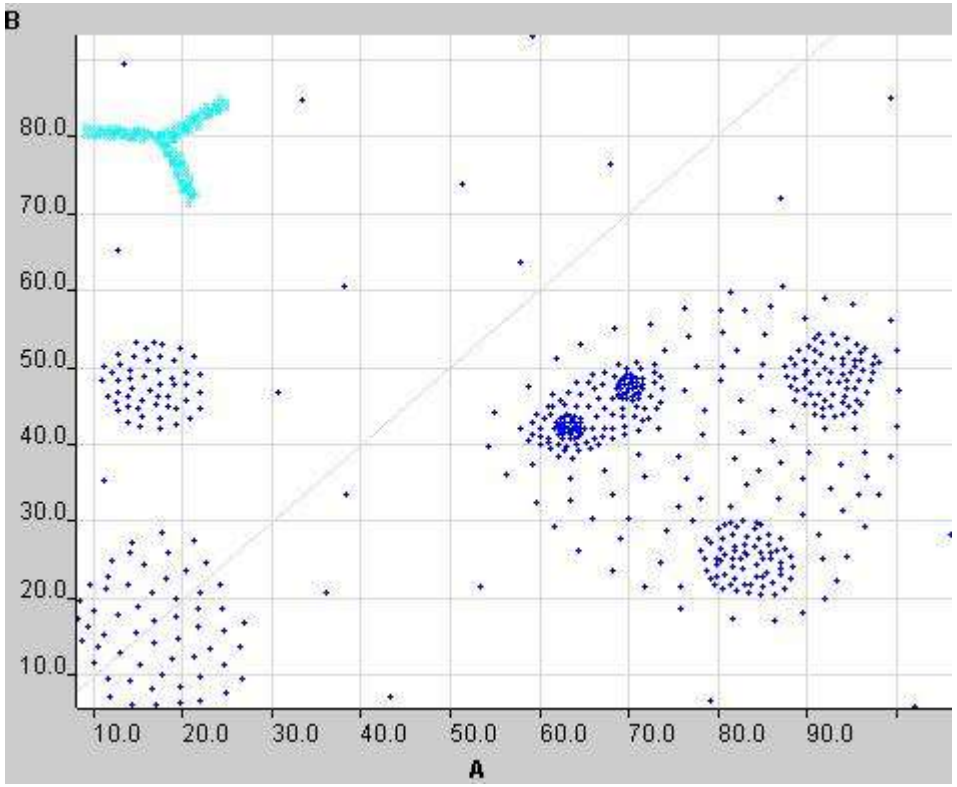


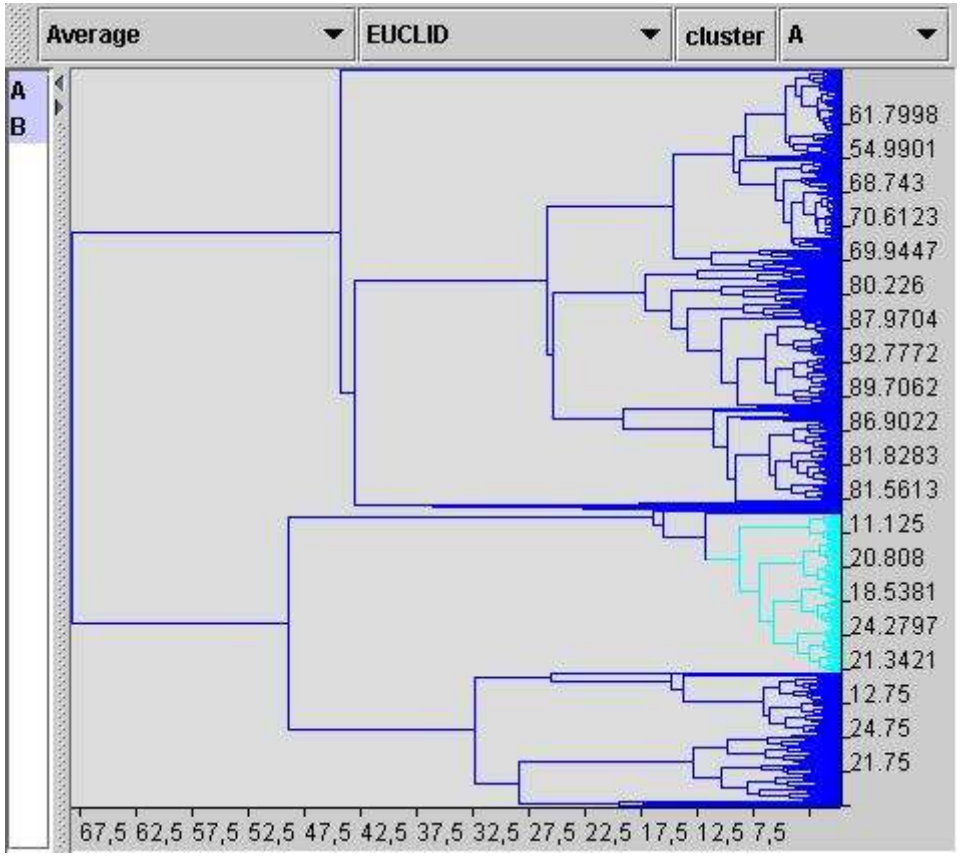
Clustering Hierárquico: Exemplo 1



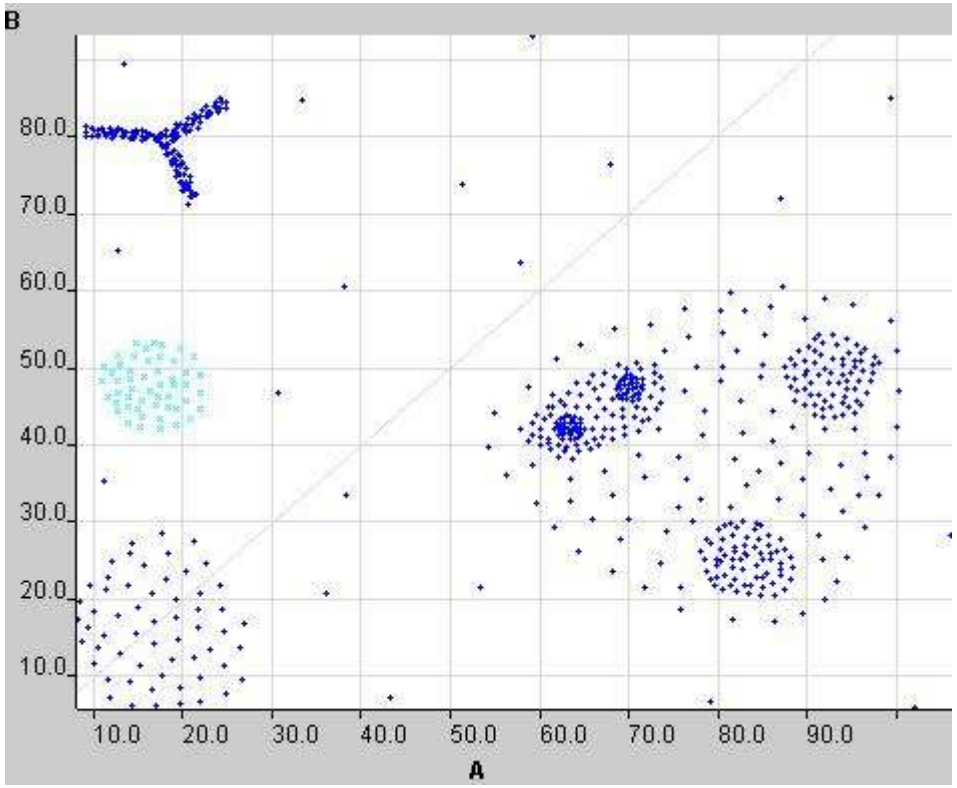


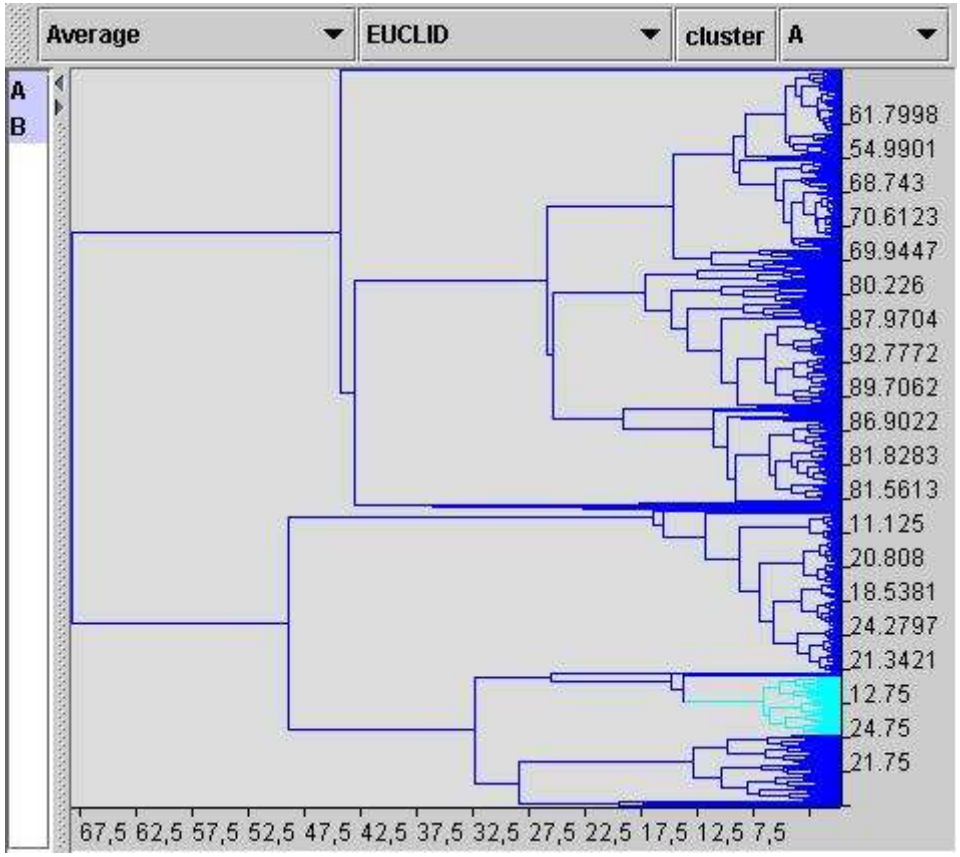
Clustering Hierárquico: Exemplo 1



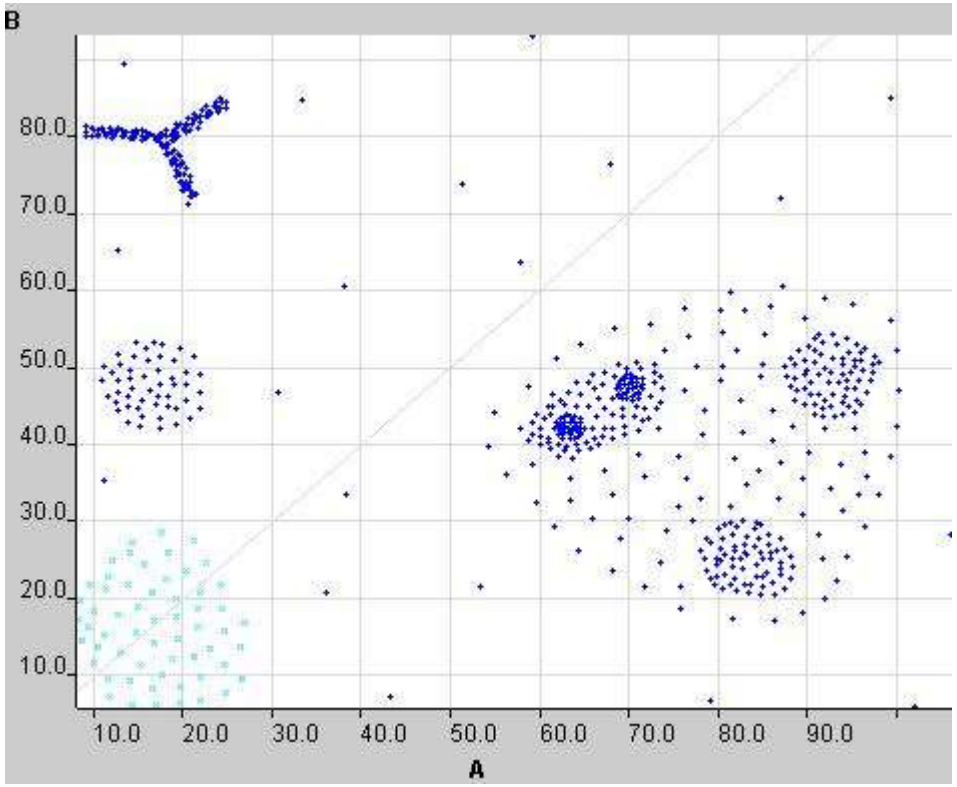


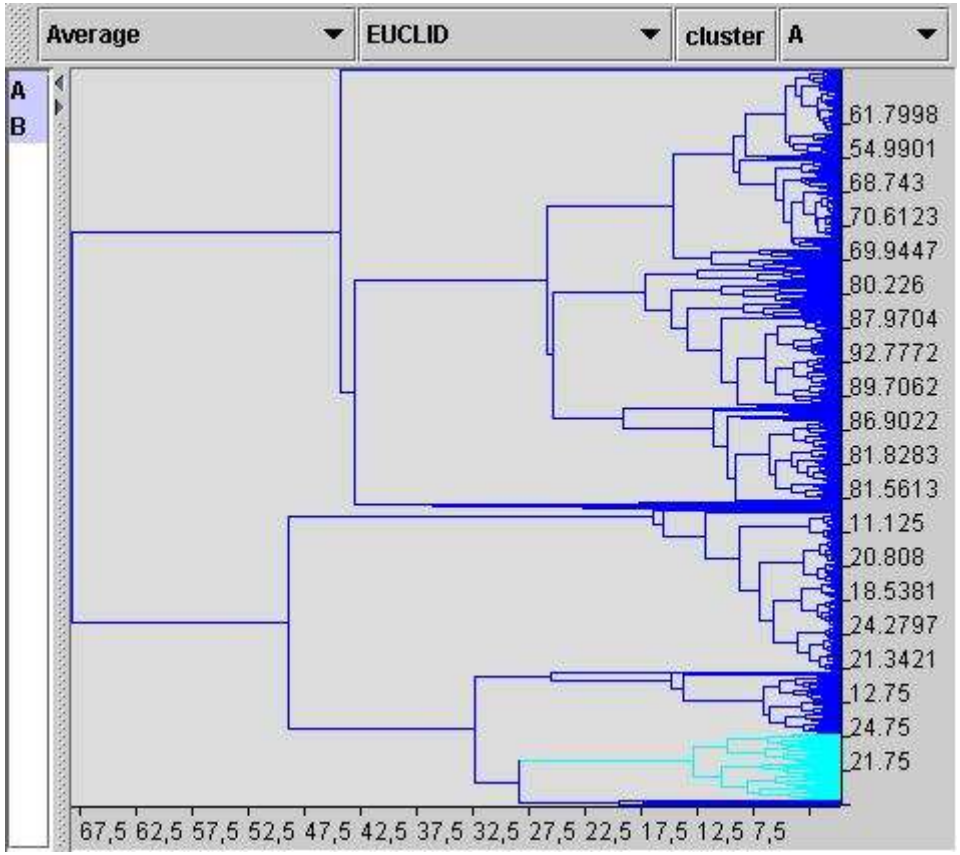
Clustering Hierárquico: Exemplo 1



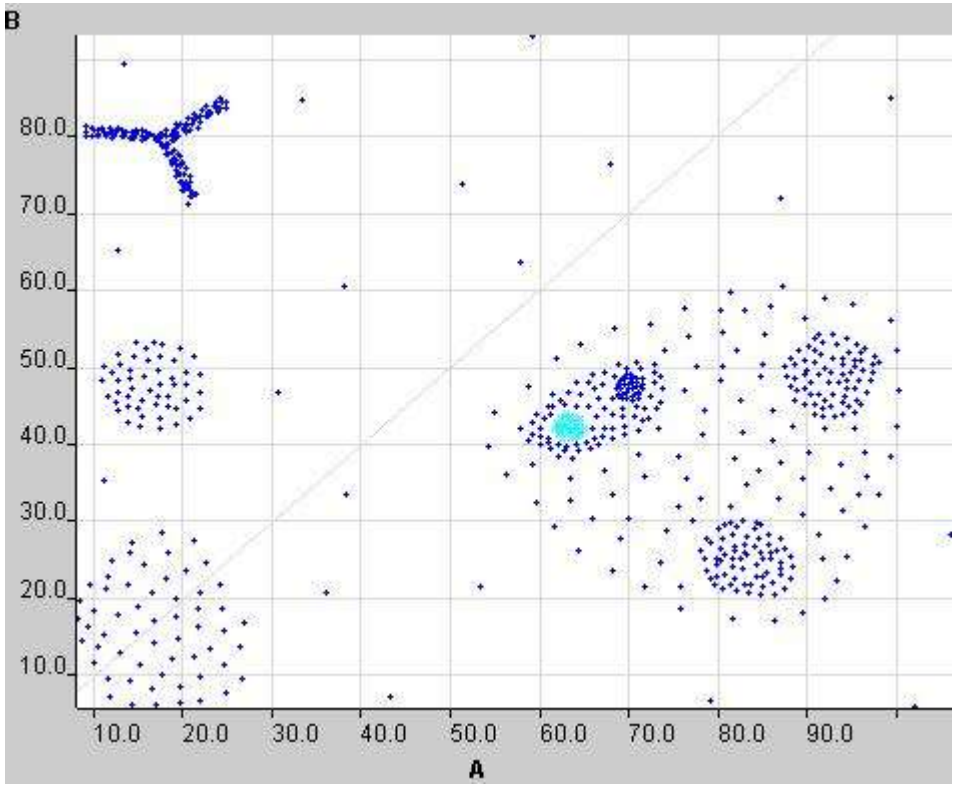


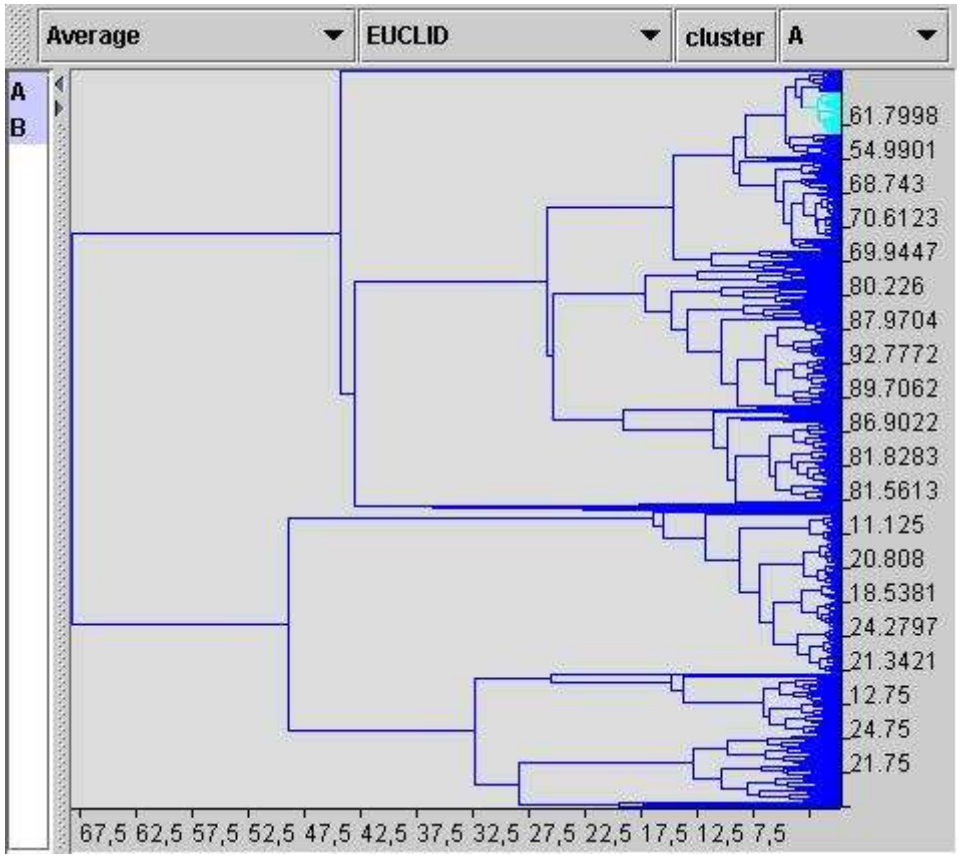
Clustering Hierárquico: Exemplo 1



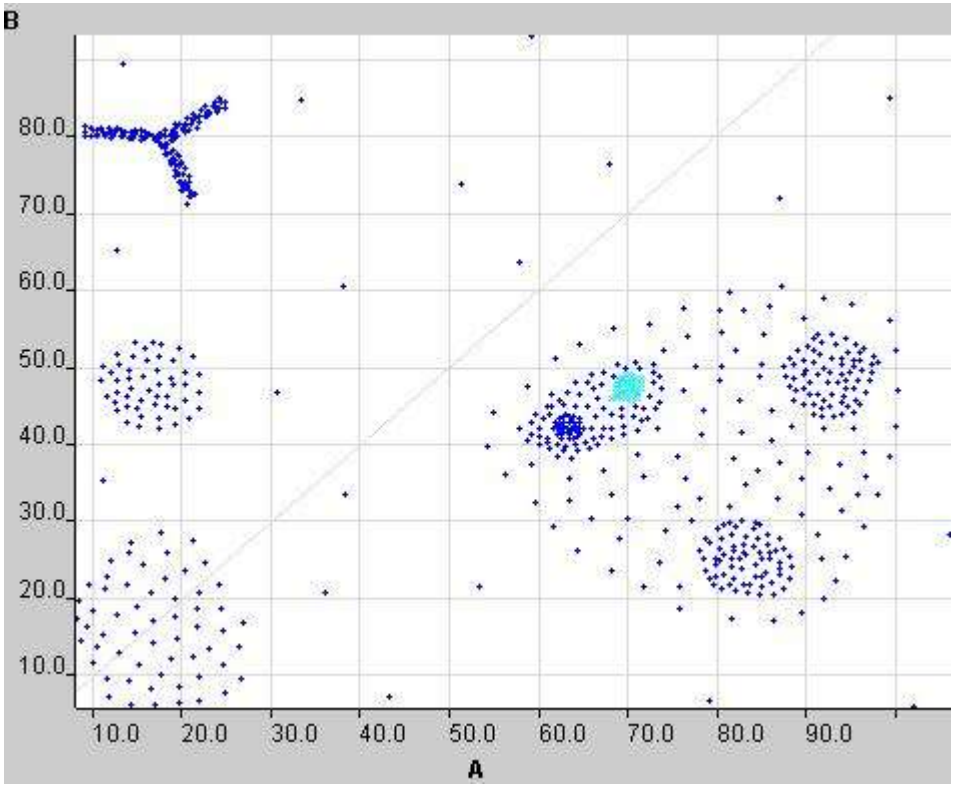


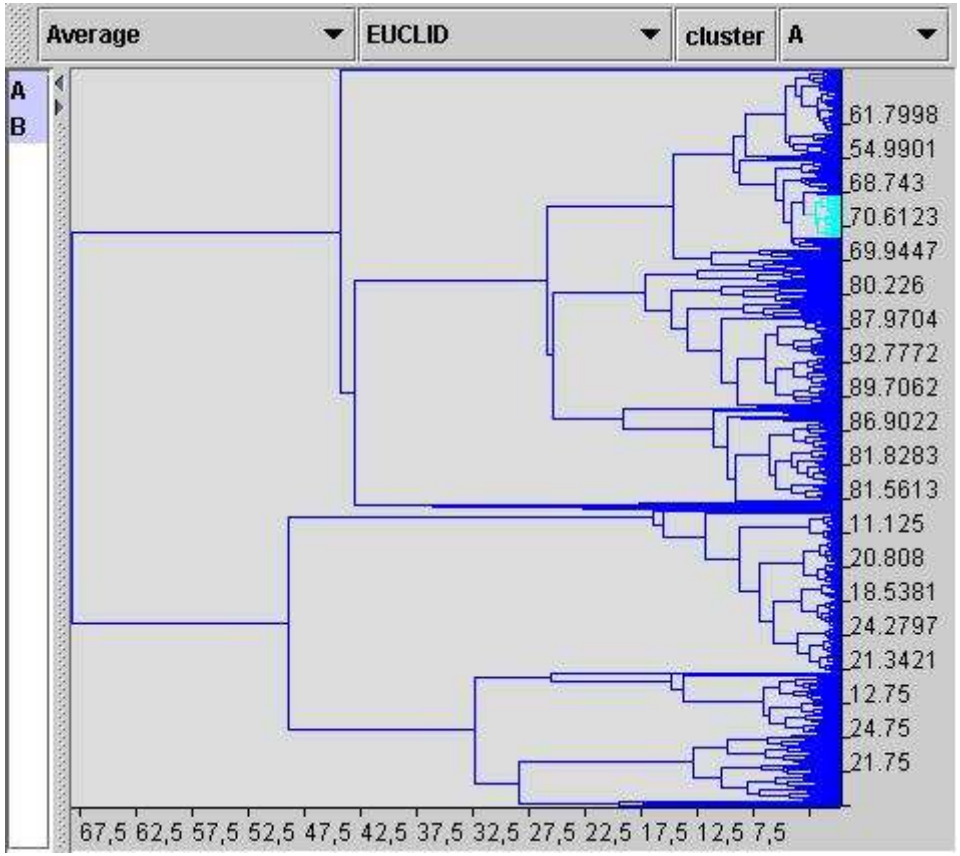
Clustering Hierárquico: Exemplo 1



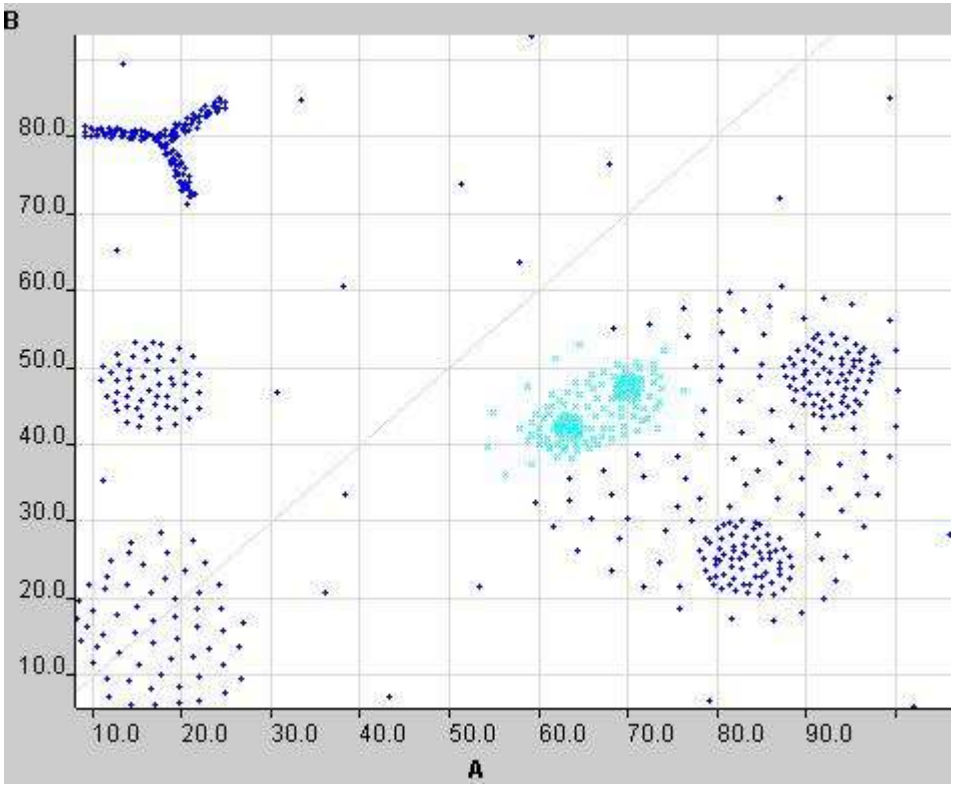


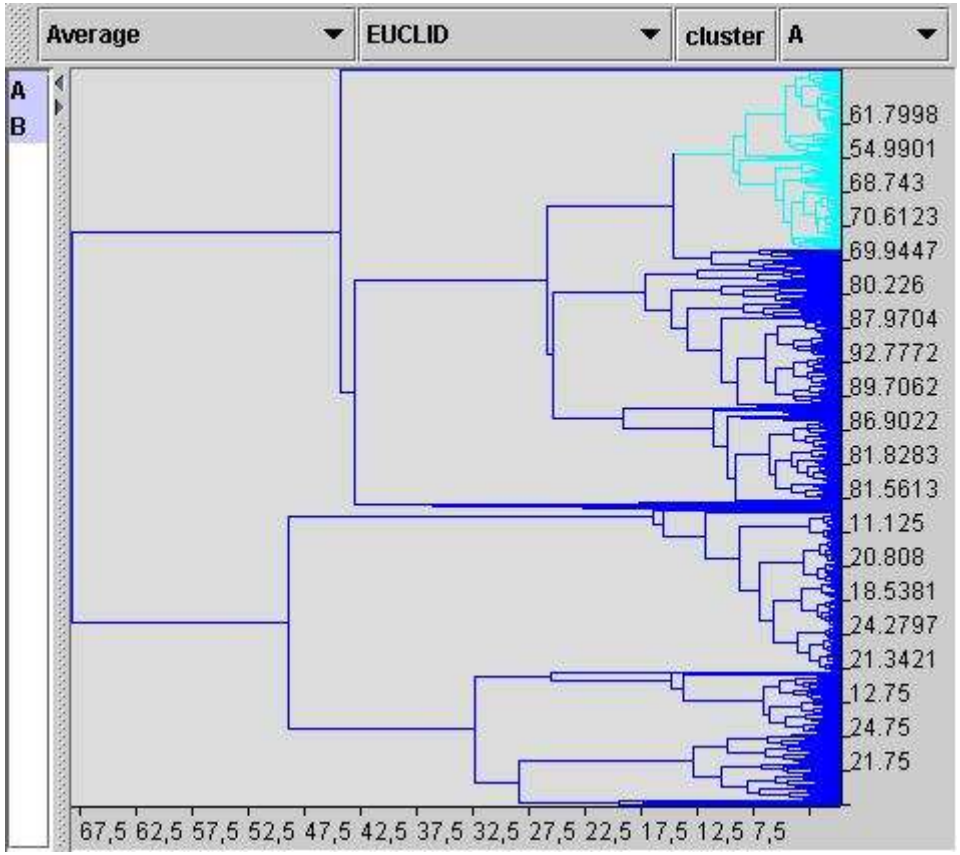
Clustering Hierárquico: Exemplo 1



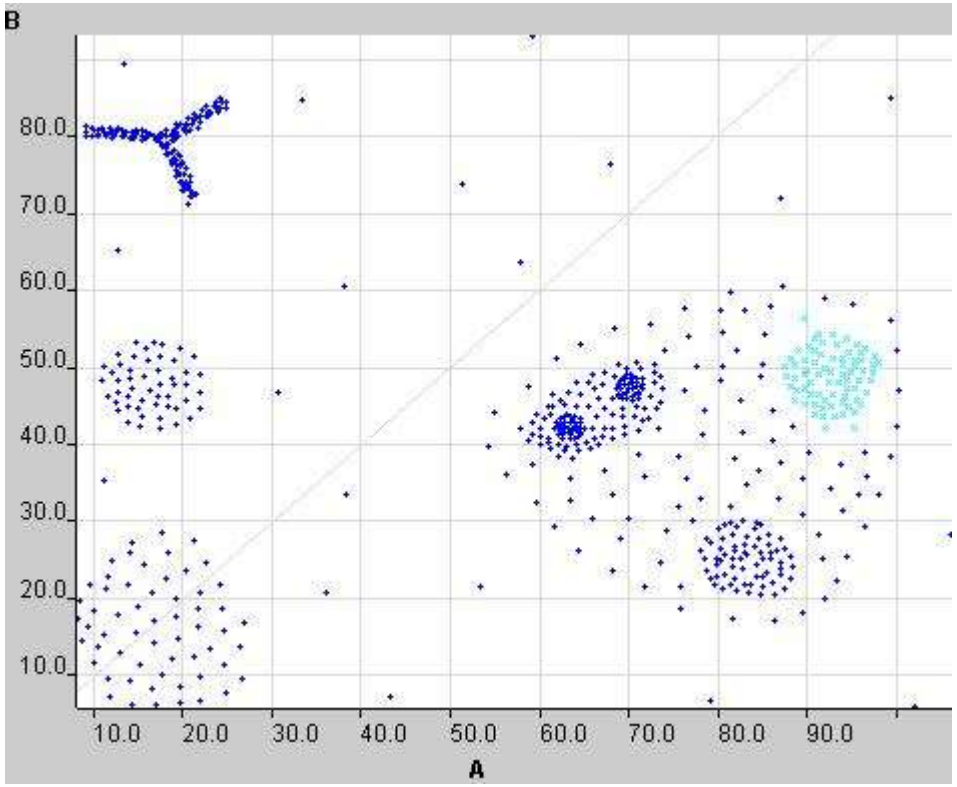


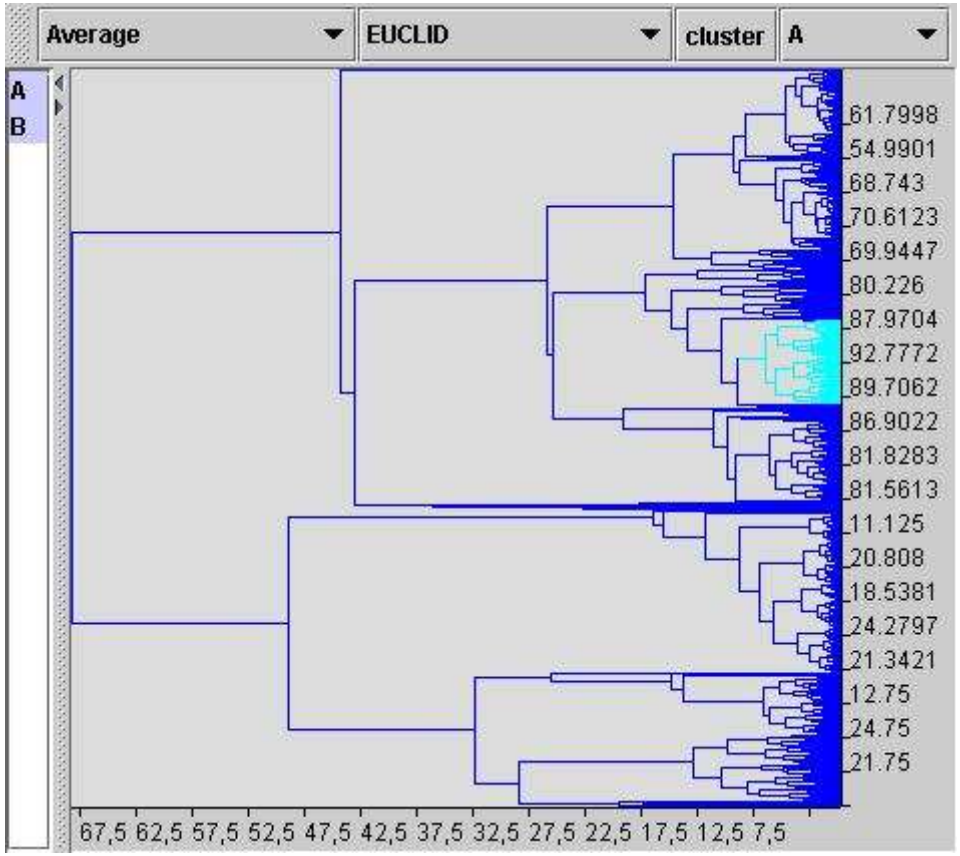
Clustering Hierárquico: Exemplo 1



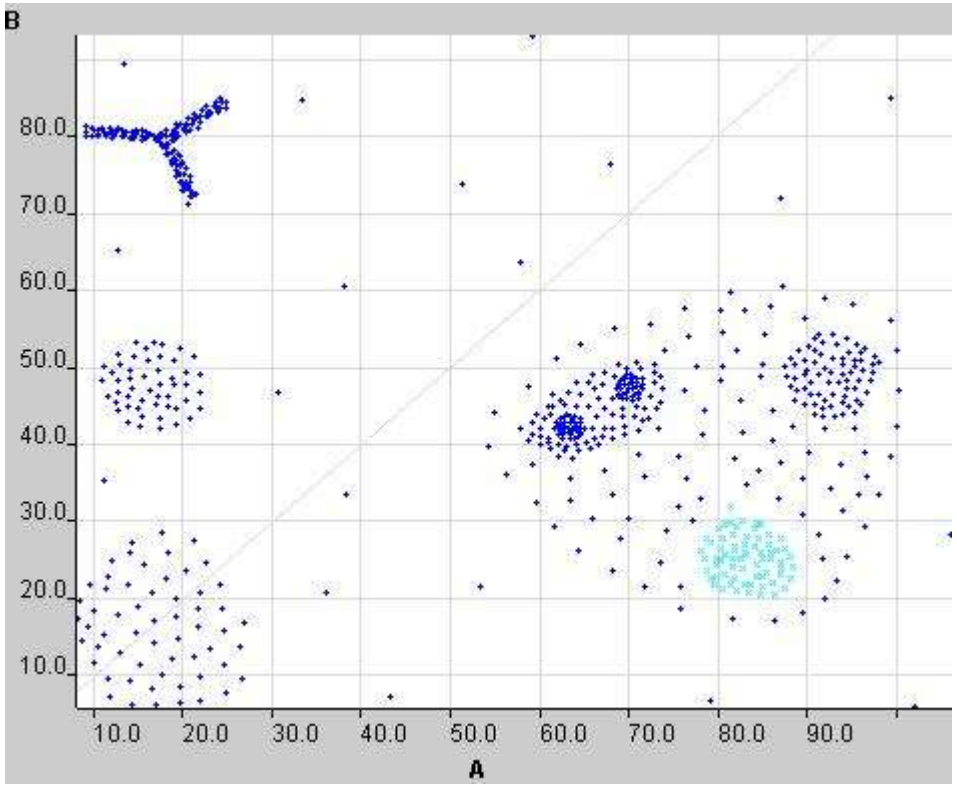


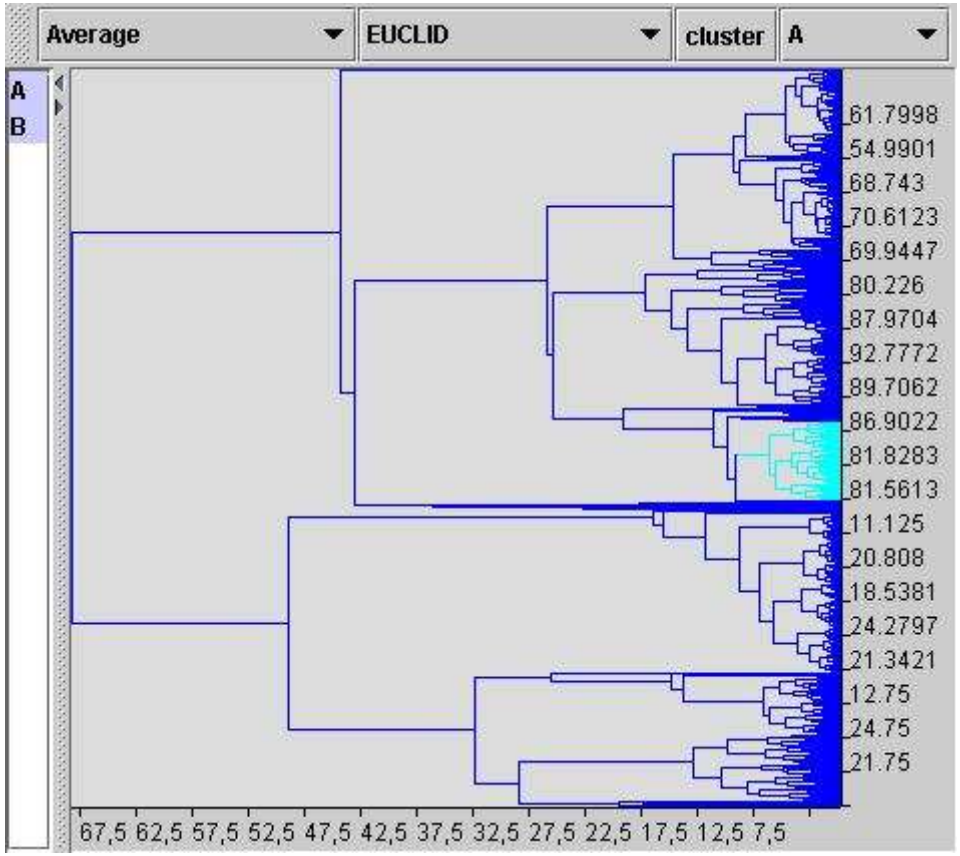
Clustering Hierárquico: Exemplo 1



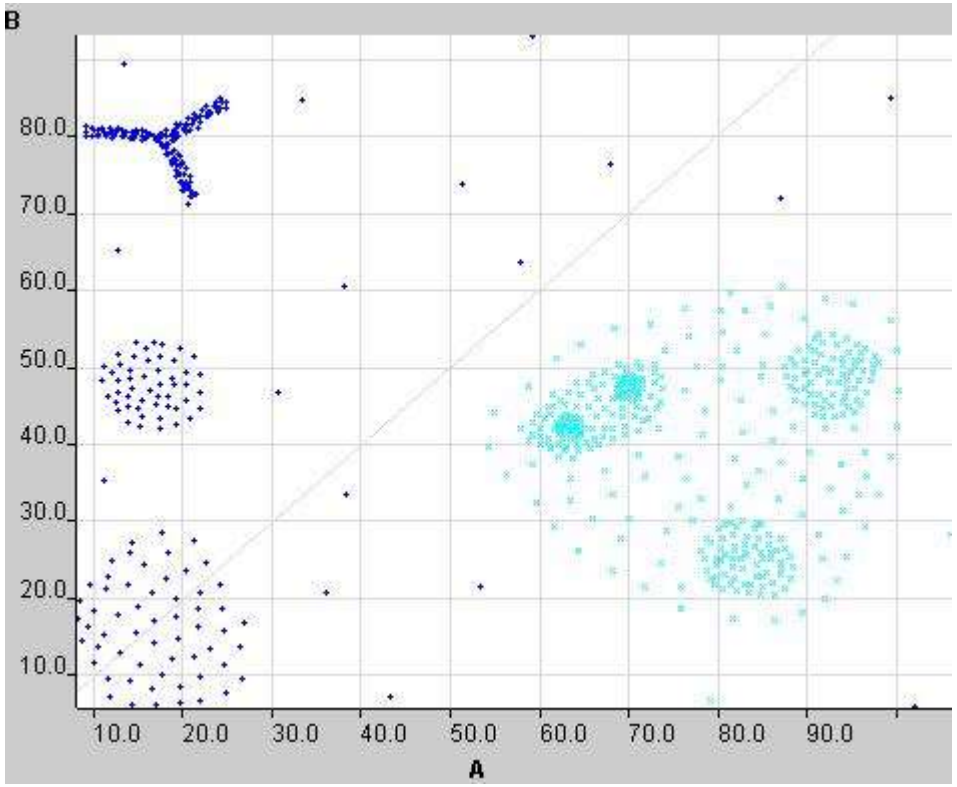


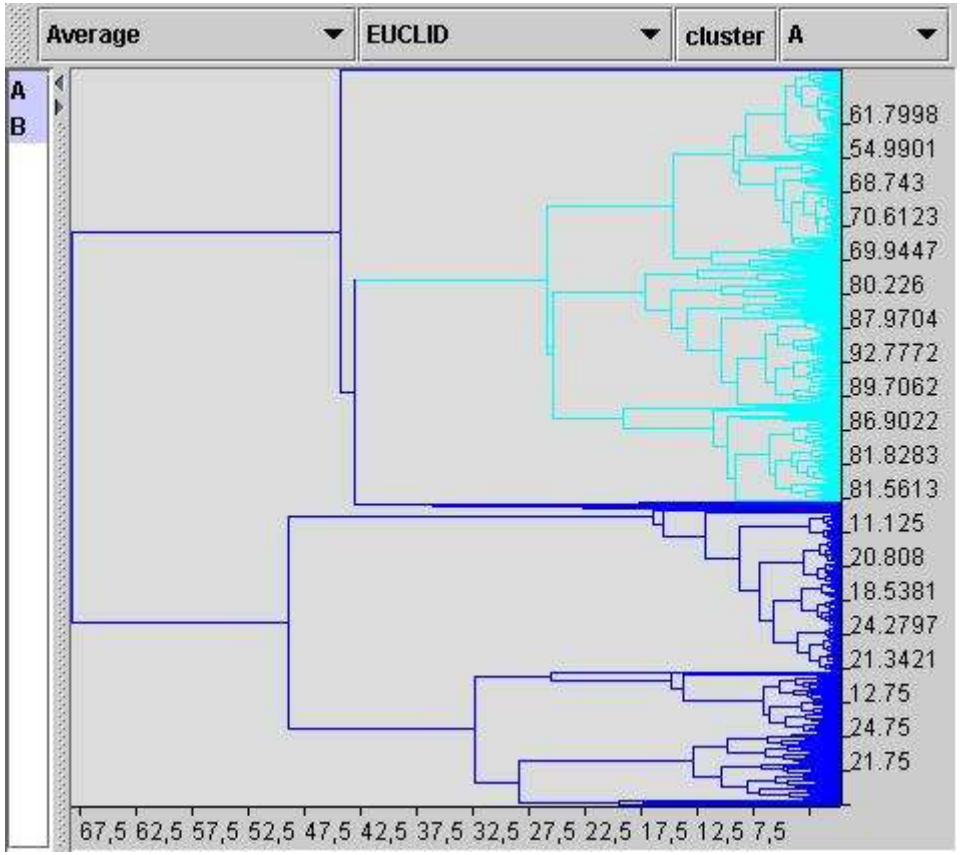
Clustering Hierárquico: Exemplo 1





Clustering Hierárquico: Exemplo 1



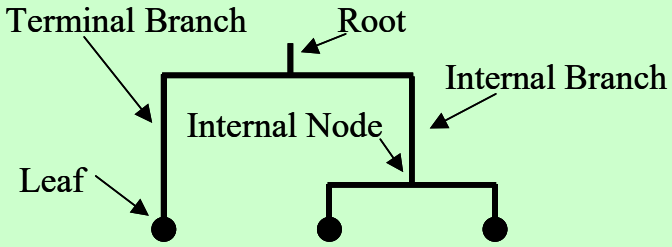


Qual tipo de agrupamento hierárquico é melhor?

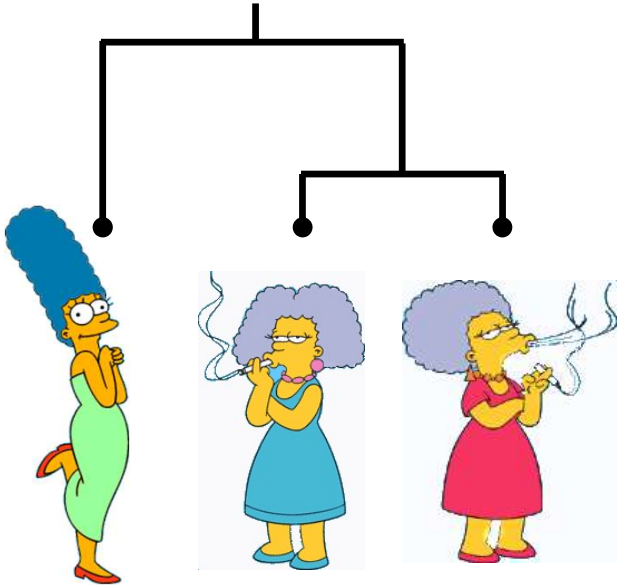
- Cada tipo tem vantagens e desvantagens
- Single-link
  - Pode encontrar grupos de formato irregular
  - Sensível para outliers
- Complete-link e Average-link
  - Robustos para outliers
  - Tendem a particionar grupos grandes
  - Preferem grupos esféricos

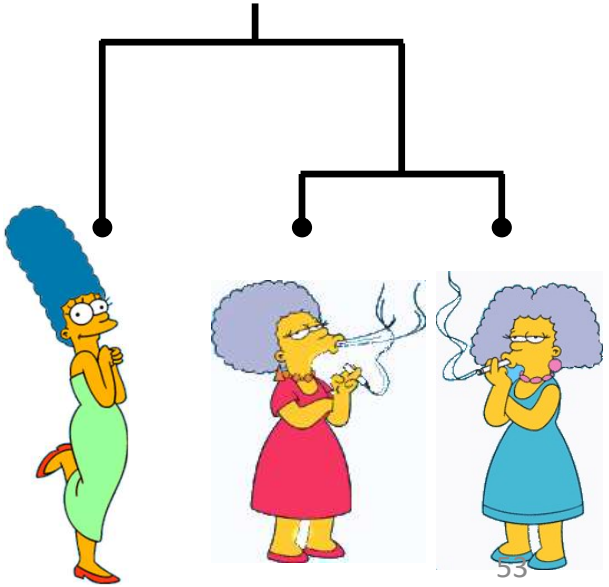
Uma Ferramenta Útil para Resumir as Similaridades

Dendograma

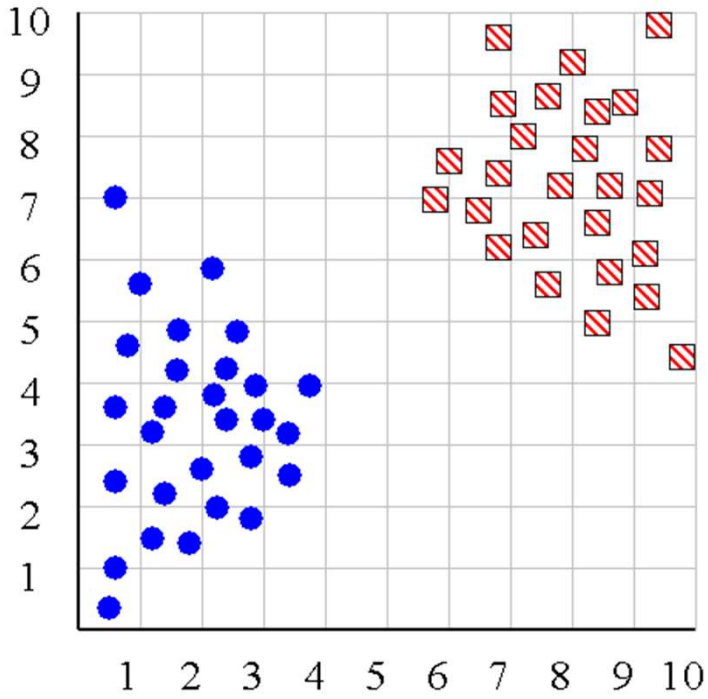


A similaridade entre dois objetos em um dendograma é representada pela altura do nó interno mais baixo que eles compartilham



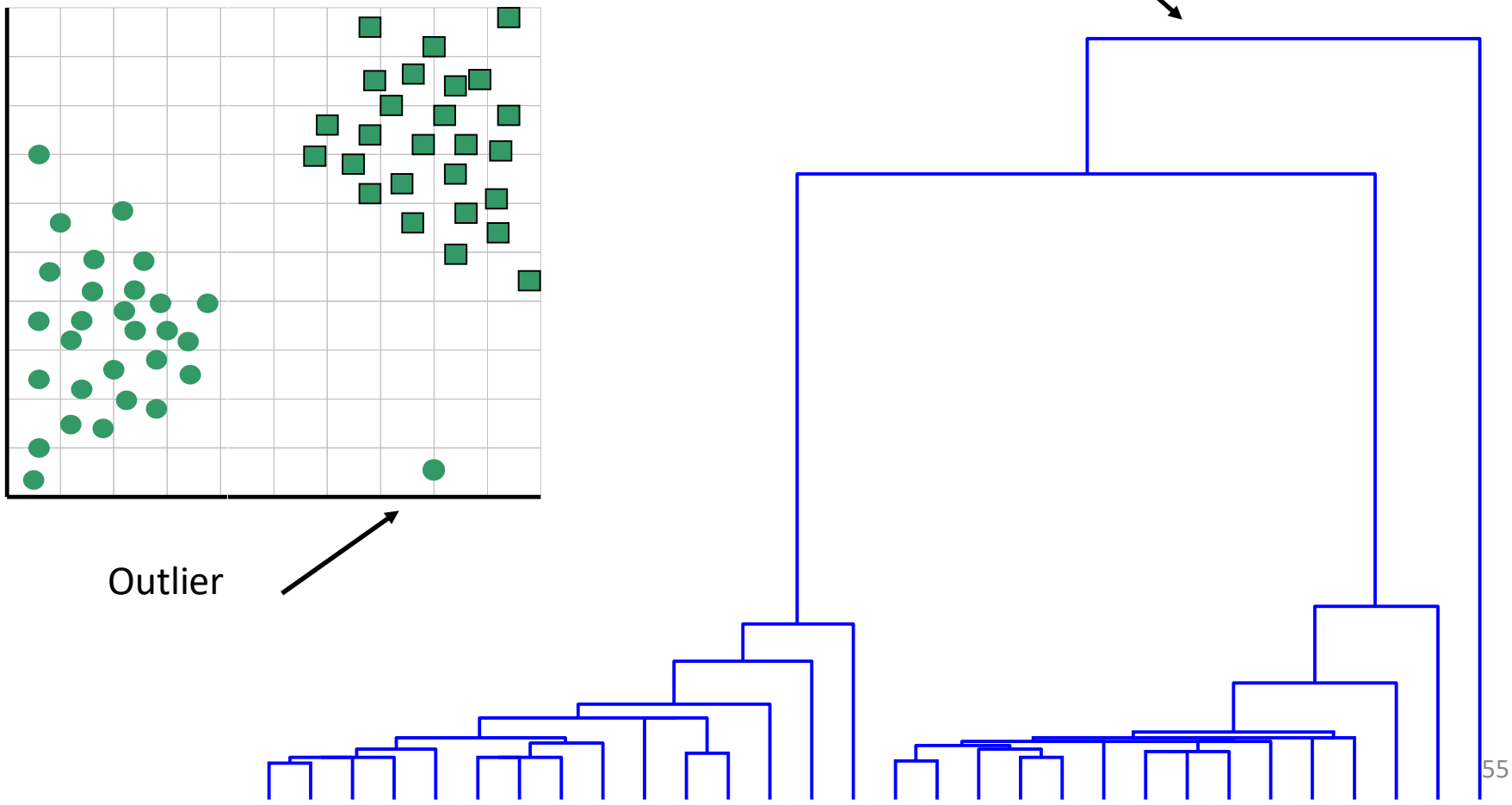


Um dendograma pode ser usado para determinar o número “correto” de agrupamentos. Por exemplo, a existência de duas árvores bem separadas é um forte indicativo de dois *clusters*.



Um possível uso de dendogramas é a detecção de *outliers*

Um ramo único e isolado sugere um dado que é muito diferente de todos os demais



Considerações sobre Clustering Hierárquico

- Não existe a necessidade de especificar o número de *clusters a priori*
- A natureza hierárquica é facilmente mapeada pela intuição humana para alguns domínios
- Eles não escalam bem: a complexidade de tempo é pelo menos $O(n^2)$, na qual n é o número de objetos
- Como qualquer algoritmo de busca heurística, mínimos locais são um problema
- A interpretação dos resultados é (muito) subjetiva

Outros métodos de agrupamento

- K-medoids: variação do K-means que usa mediana ao invés da média
- EM – agrupamento baseado em probabilidades
- SOM – mapas auto-organizáveis
- …

Avaliação de Clusters

- Avaliação Tradicional:

$$\text{Qualidade do Cluster} = \frac{\text{Distância Inter} - \text{Cluster}}{\text{Distância Intra} - \text{Clusters}}$$

- Avaliação para Clusters Hierárquicos
  - Poucos clusters
    - Cobertura grande → boa generalidade
  - Descrição de clusters grandes
    - Mais atributos → maior poder de inferência
  - Mínima (nenhuma) sobreposição (intersecção) entre clusters
    - Clusters mais distintos → conceitos melhor definidos

Desafios em Clustering

- Cálculo de Similaridade
  - Resultados dos algoritmos dependem inteiramente da métrica de similaridade utilizada
  - Os sistemas de clustering fornecem pouco auxílio em como escolher a similidade adequada aos objetos sendo estudados
  - Calcular a correta similaridade de dados de diferentes tipos pode ser difícil
  - Similaridade é muito dependente da representação dos dados:
    - Normalizar?
    - Representar um dado numericamente, categoricamente, entre outros?
- Seleção de Parâmetros
  - Algoritmos atuais requerem muitos parâmetros arbitrários, que devem ser especificados pelo usuário

Exemplos de aplicações

- Reconhecimento de padrões
- Análise de dados espaciais
  - Criação de mapas temáticos em GIS por agrupamento de espaços de características
  - Detecção de clusters espaciais e sua explanação em *data mining*
- Processamento de imagens
- Pesquisas de mercado
- WWW:
  - Classificação de documentos
  - Agrupamento de dados de weblogs para descobrir padrões similares de acesso

- Alguns slides foram baseados em apresentações de:
  - Profa. Huei Diana Lee
  - Prof. José Augusto Baranauskas
  - Prof. E. Keogh
  - Profa. Bianca Zadrozny
  - Prof. S. A. Demurjian
  - Prof. G. Piatetsky-Shapiro