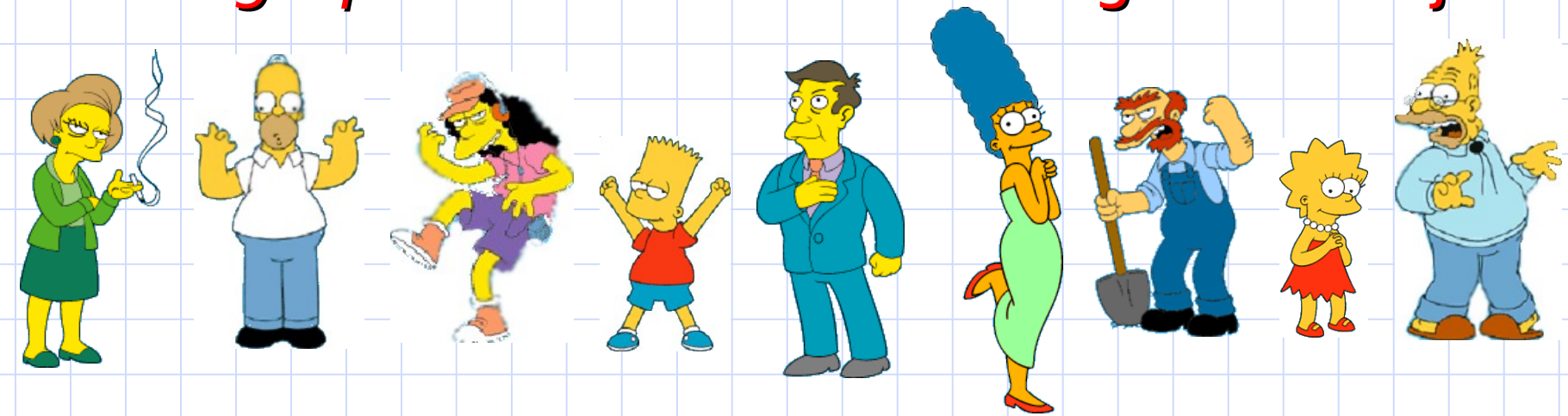
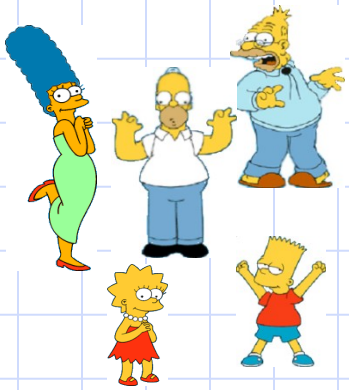
# Agrupamento de Dados

**Prof. Debora Medeiros**

*Slides adaptados de Eduardo Raul Hruschka (USP)

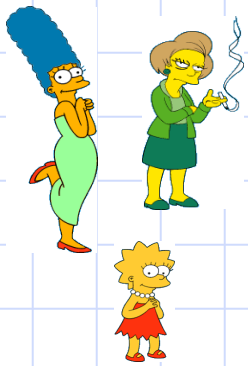
# Como *agrupar naturalmente* os seguintes objetos?



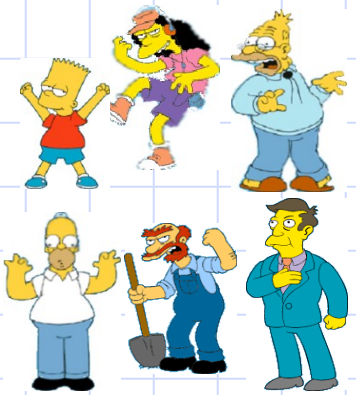


Família

Empregados



Mulheres



Homens

→ *Cluster* é um conceito subjetivo!

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

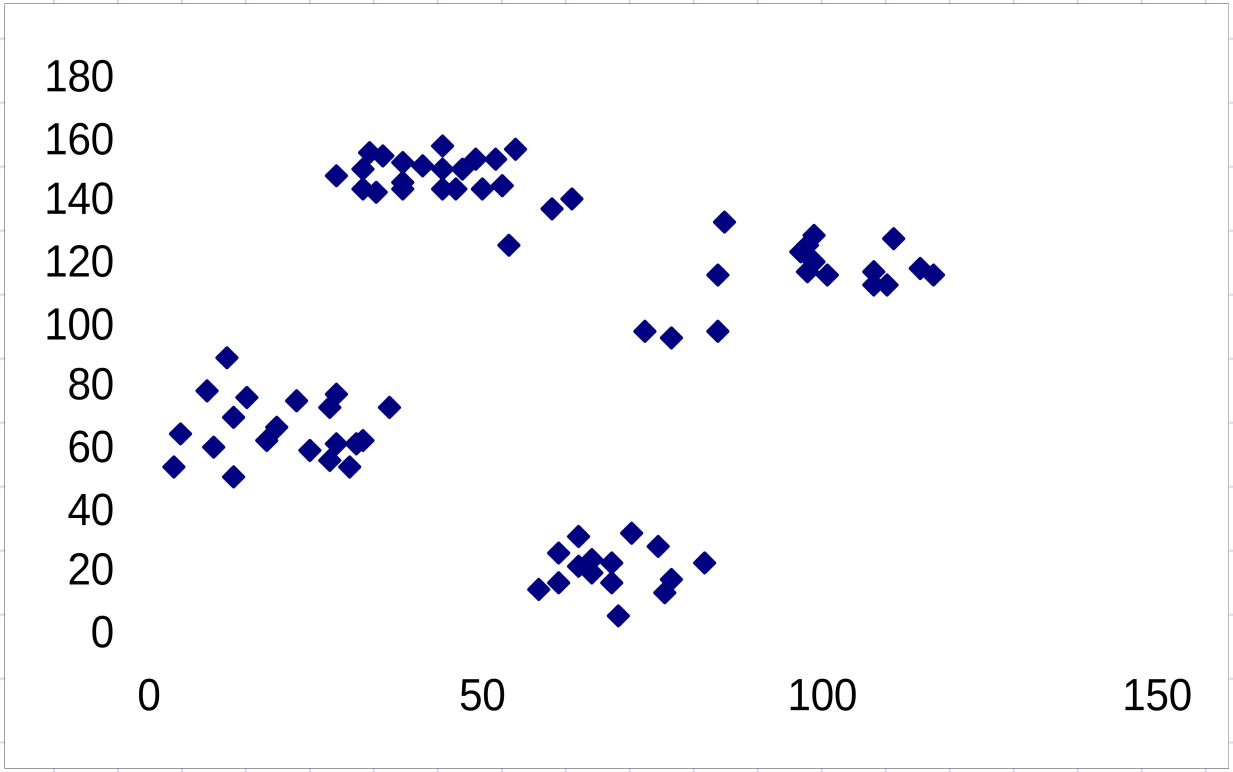
# O que é um grupo (*cluster*)?

- Definições subjetivas:
  - “Semelhanças entre objetos”.
  - Quais atributos devemos considerar para computar similaridades?



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

- Numa “abordagem matemática”, critérios numéricos usualmente consideram:
  - Homogeneidade (coesão interna);
  - Heterogeneidade (separação entre grupos);

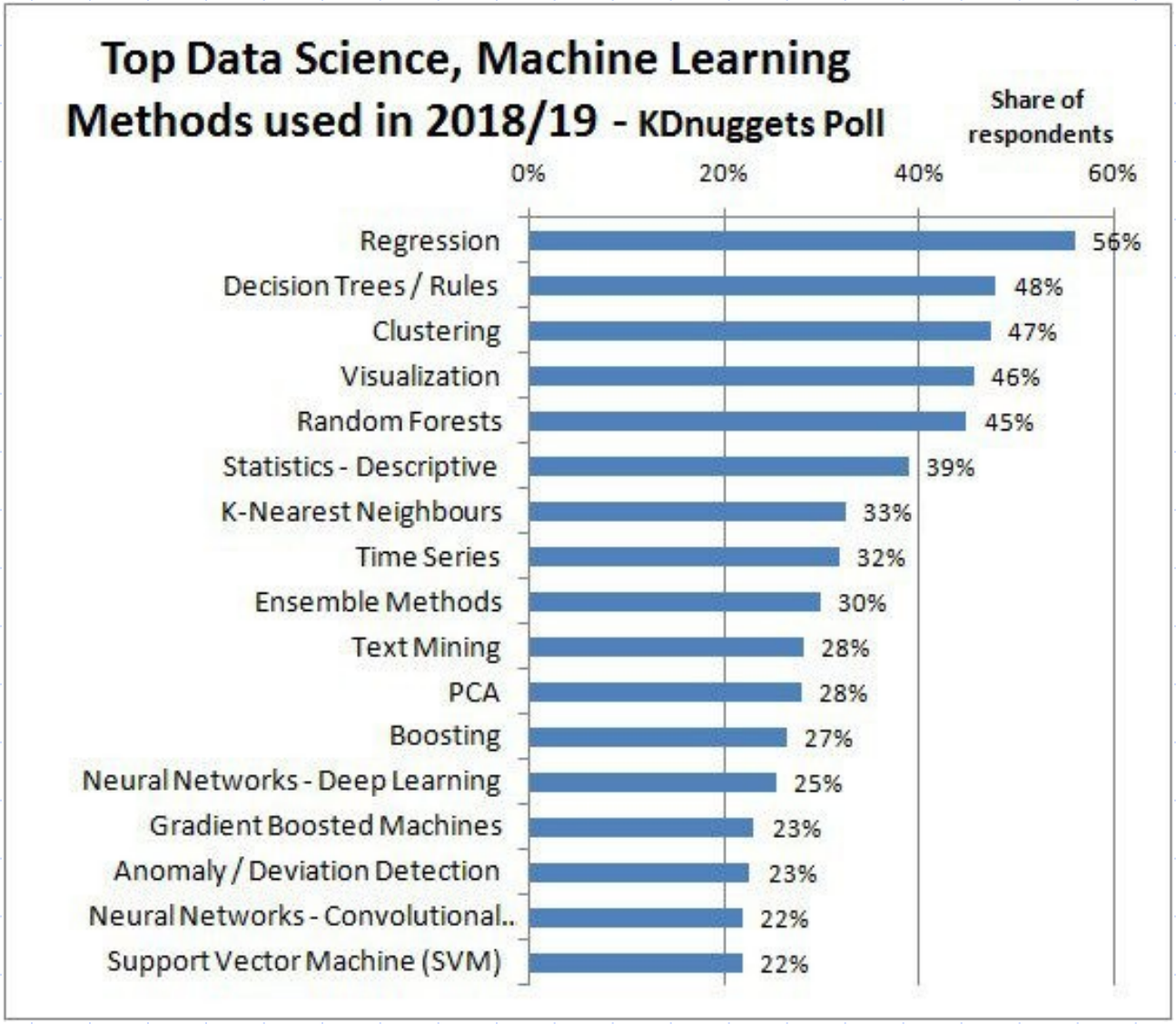



(Ruspini, 1970)

*Agrupamento de Dados* (ADs) é uma técnica importante para *Análise Exploratória de Dados* :

- Engenharia;
- Biologia;
- Psicologia;
- Medicina;
- Administração (*Marketing* , Finanças,...);
- Ciência da Computação:
  - Bioinformática;
  - Dados coletados via sensores;
  - Componentes de sistemas inteligentes;
  - Componentes de algoritmos para aprendizado de máquina, ...

**https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html**



# 2. Conceitos Básicos

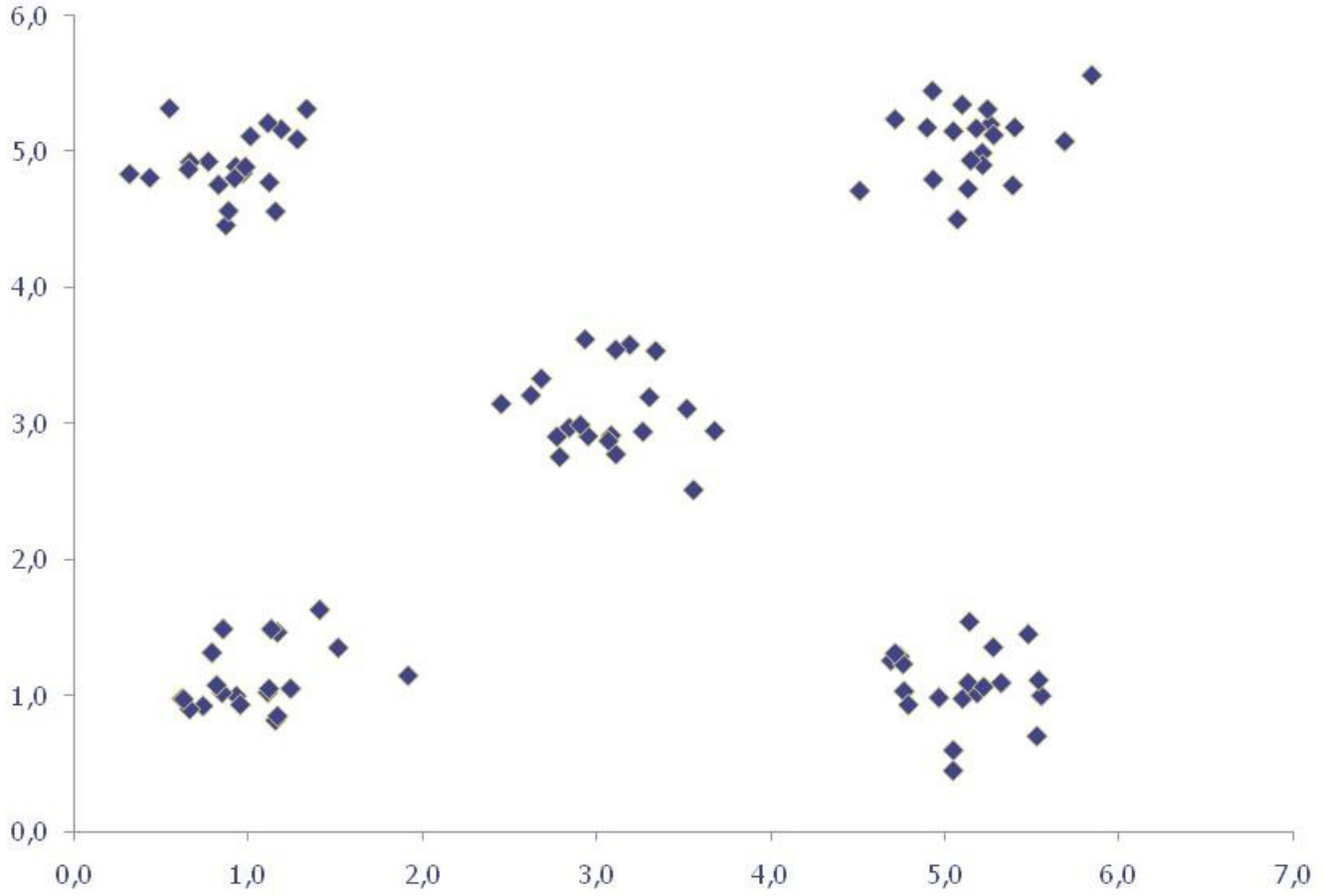
Algumas Definições (Everitt, 1974):

• *Um cluster (**grupo**) é um conjunto de **entidades semelhantes**, e entidades pertencentes a **diferentes clusters** **não são semelhantes**.*

• *Um grupo é uma aglomeração de pontos no espaço tal que a **distância** entre quaisquer **dois pontos no grupo** é **menor** do que a **distância** entre qualquer ponto no grupo e qualquer **ponto fora** deste.*

• *Grupos podem ser descritos como **regiões** conectadas de um espaço multidimensional contendo uma **densidade** de pontos relativamente **alta**, **separada** de outras tais regiões por uma região contendo uma **densidade** relativamente **baixa** de pontos.*

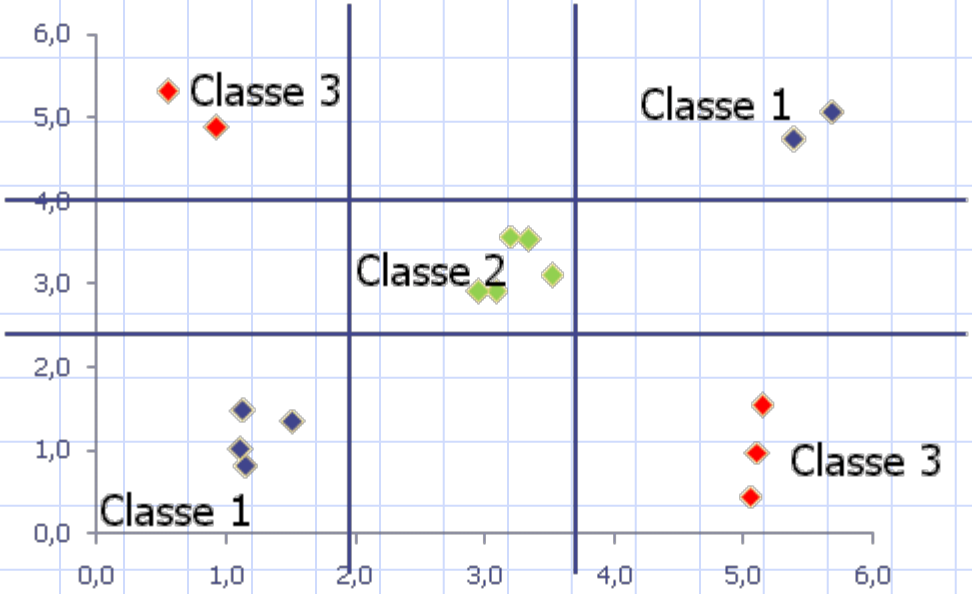
➢ **Humanos reconhecem *clusters* no plano quando os vêem, sem saber explicar exatamente porquê (Jain & Dubes, 1988) ...**




**Agrupamento:** Indução de grupos a partir da base de dados...

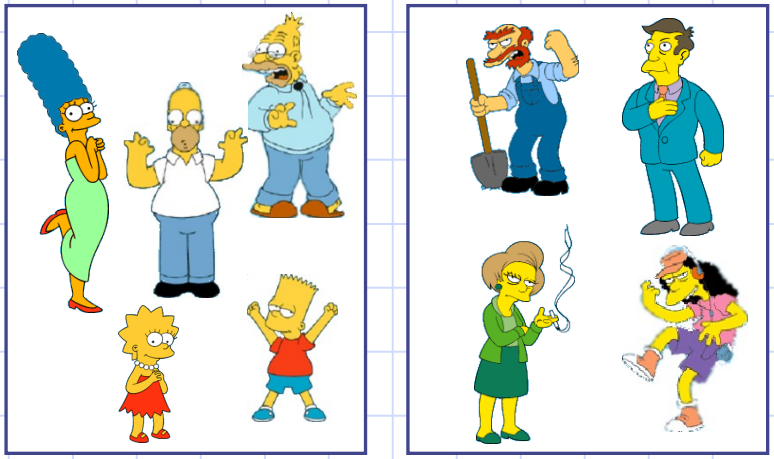
# Agrupamento X Classificação?

...

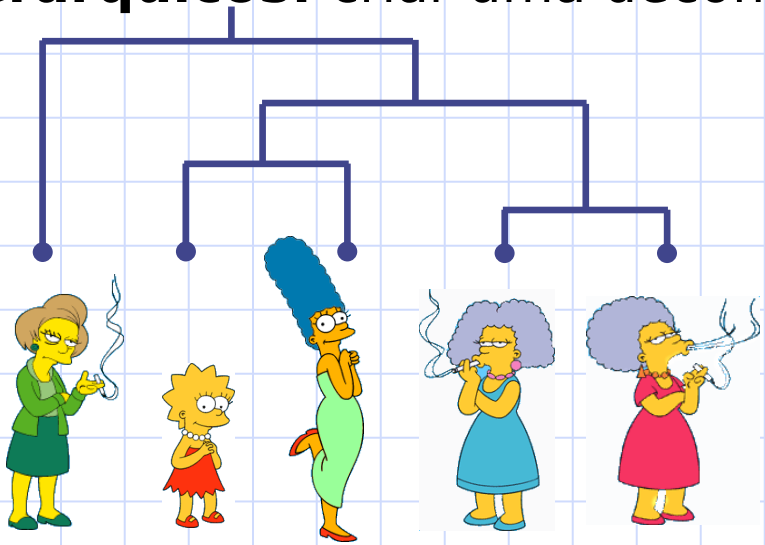



5 grupos

**Algoritmos para particionamento:** construir partições.



**Algoritmos hierárquicos:** criar uma decomposição hierárquica.

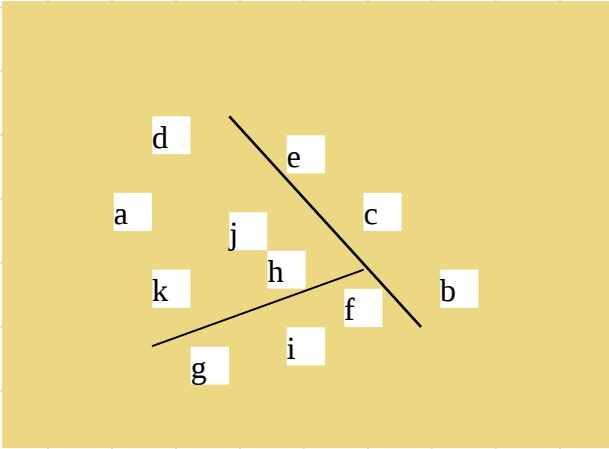




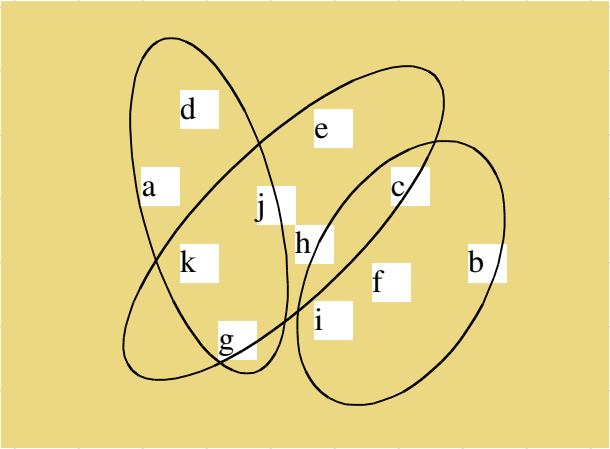
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

- **Métodos para Particionamento**:

***Rígidos***




***Sobreposição***




- Em princípio requerem a definição, a priori, do número de grupos;
- Métodos hierárquicos, por sua vez...

# Algoritmos particionais

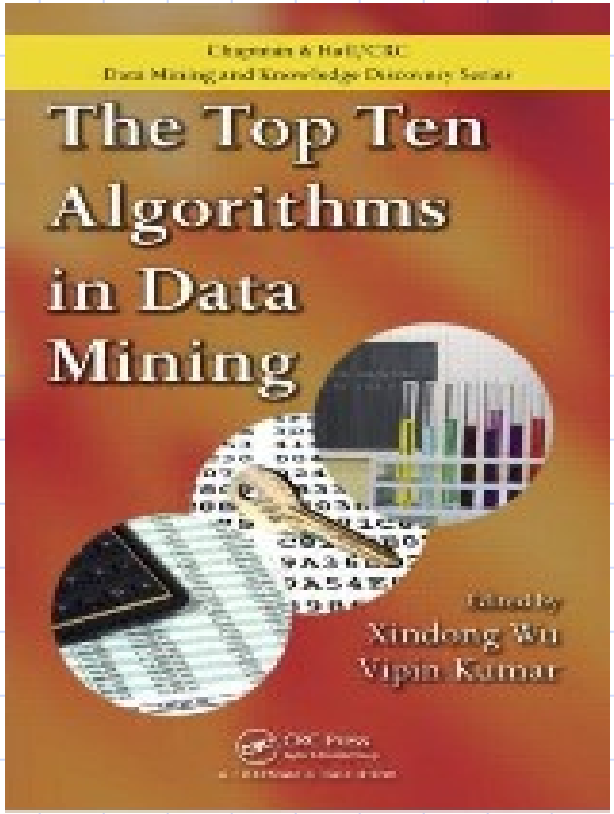
## Partições rígidas

**Abordaremos um algoritmo amplamente usados na prática:**

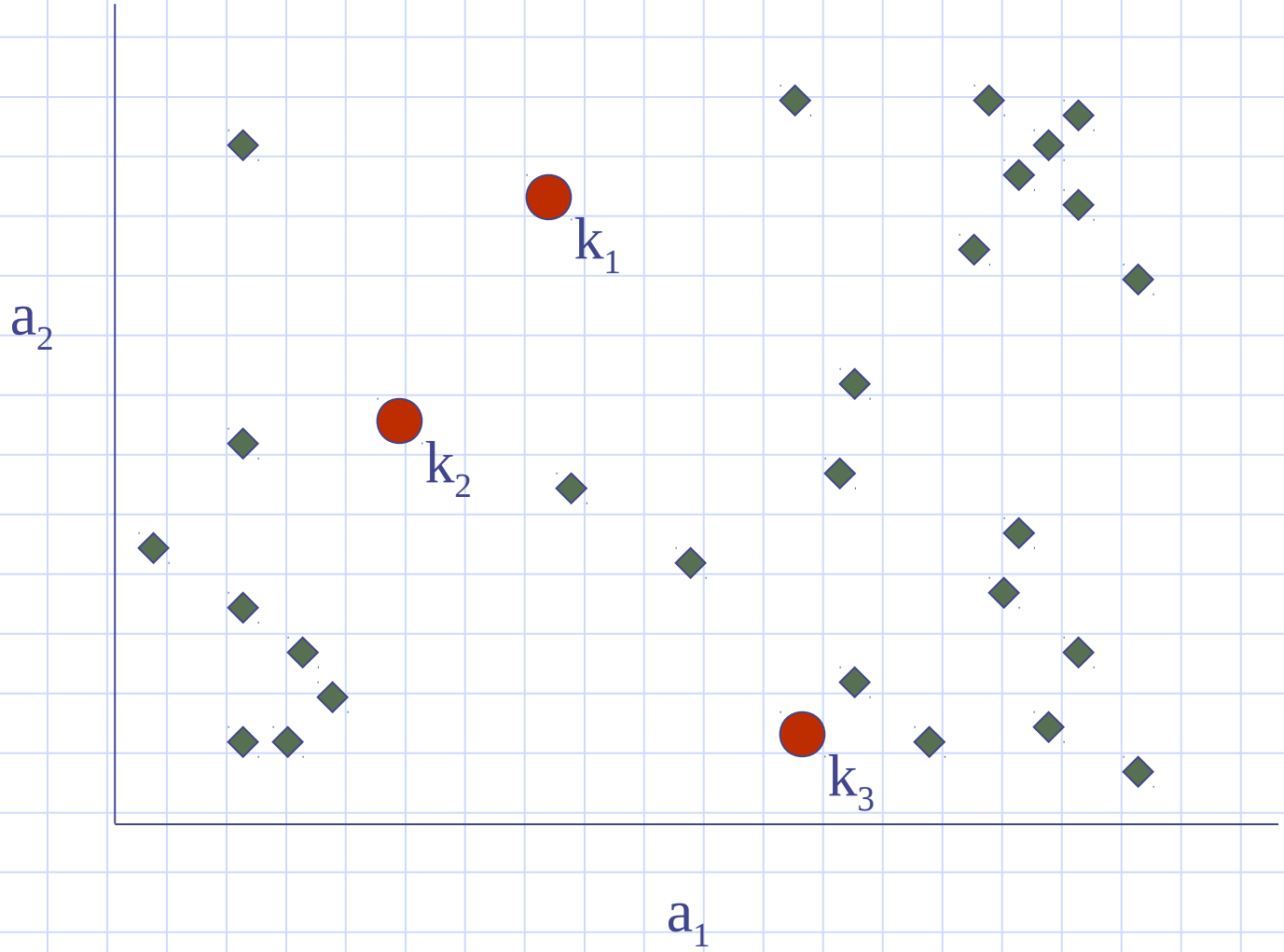
- ***k*-médias (*k-means*);**

# Algoritmo *k*-médias (MacQueen, 1967)

- Amplamente usado na prática:
  - Simplicidade;
  - Interpretabilidade;
  - Eficiência Computacional.



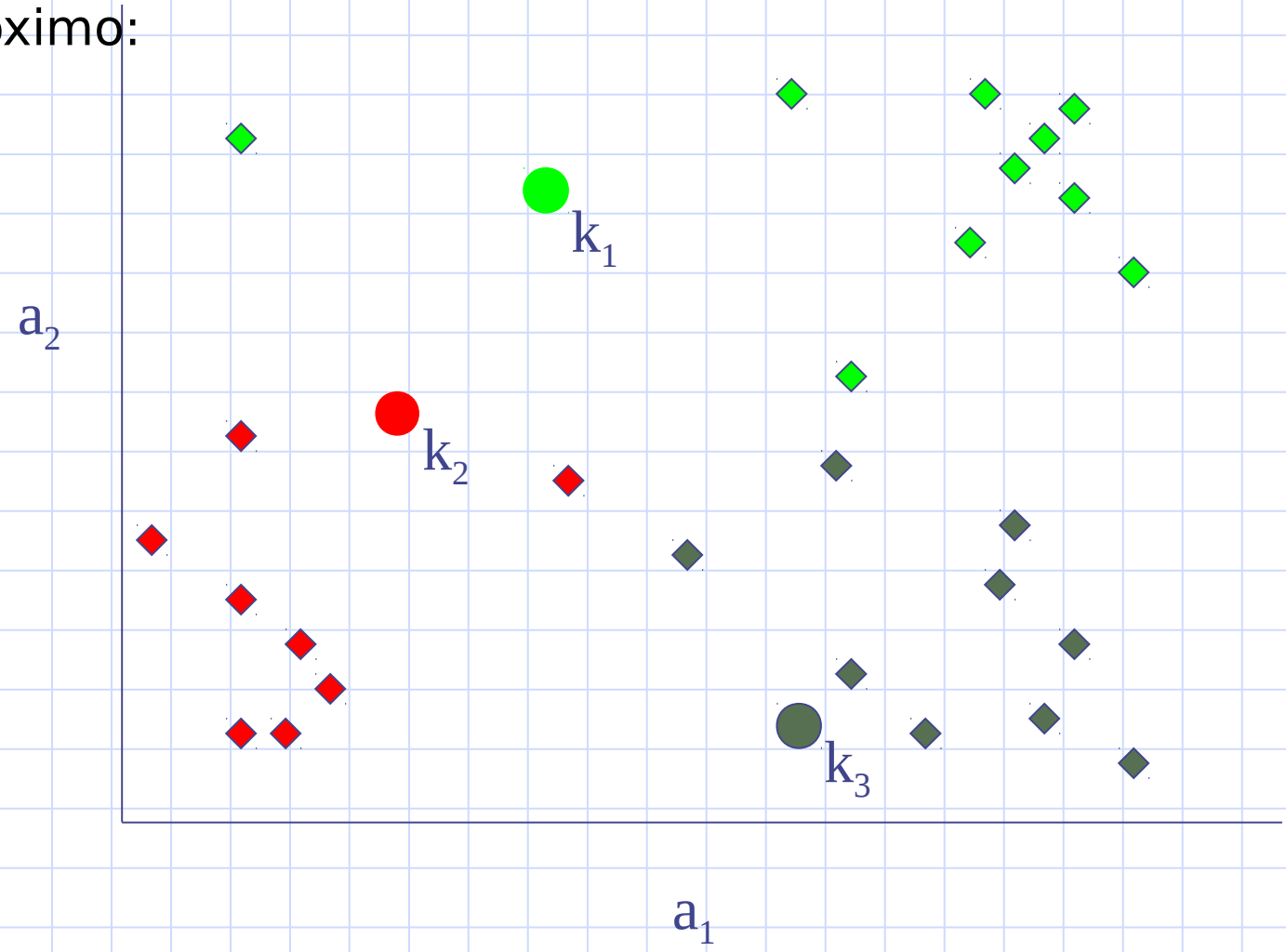
- Assumamos que queremos encontrar três *clusters* (*k* = 3) para uma base de dados bi-dimensional:



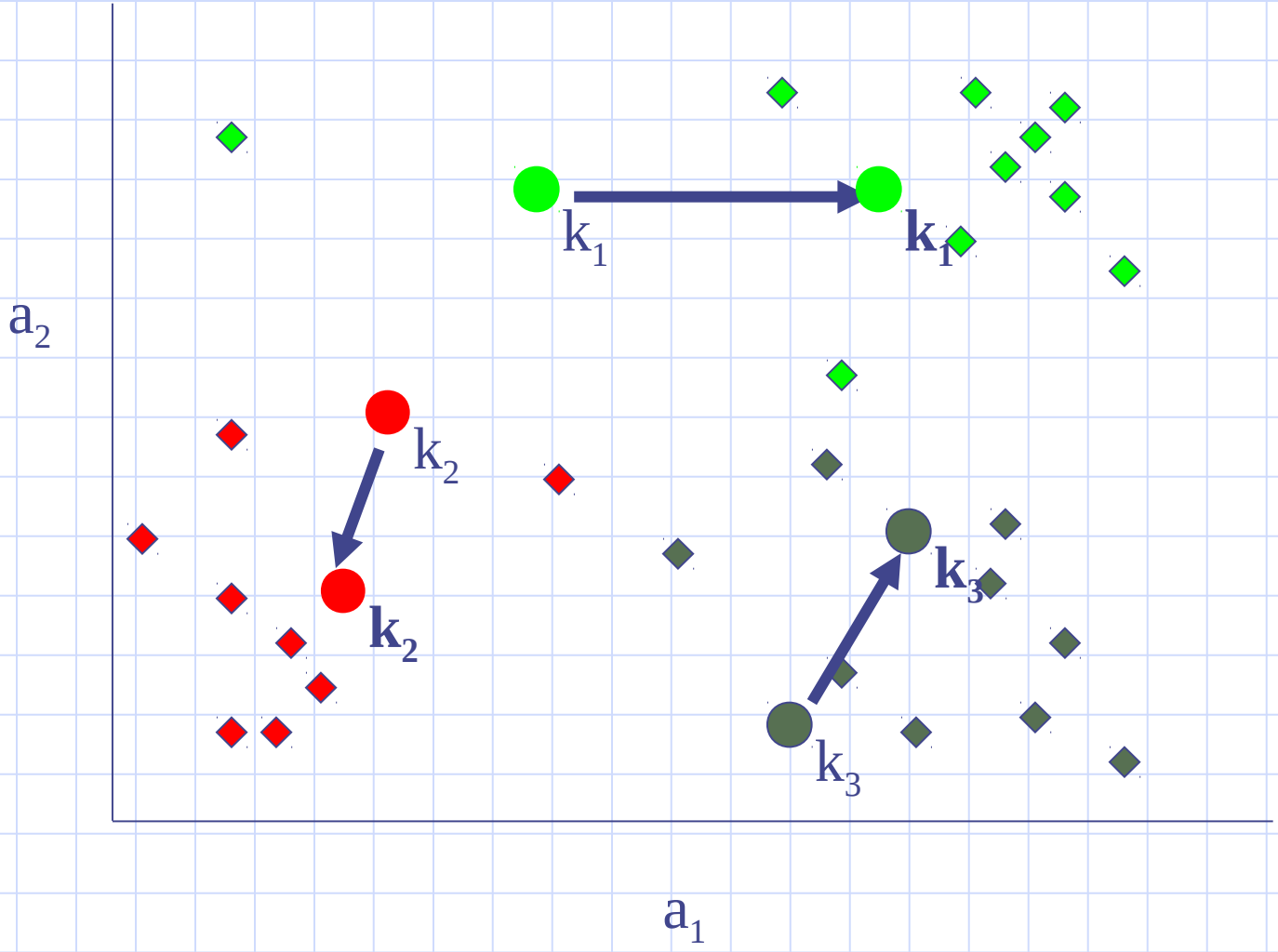



From the slides by Gregory Piatetsky-Shapiro & Gary Parker - available at www.kdnuggets.com

• Calcular dissimilaridades entre objetos e protótipos ($k_1$,$k_2$,$k_3$), encontrando grupos iniciais pela regra do vizinho mais próximo:



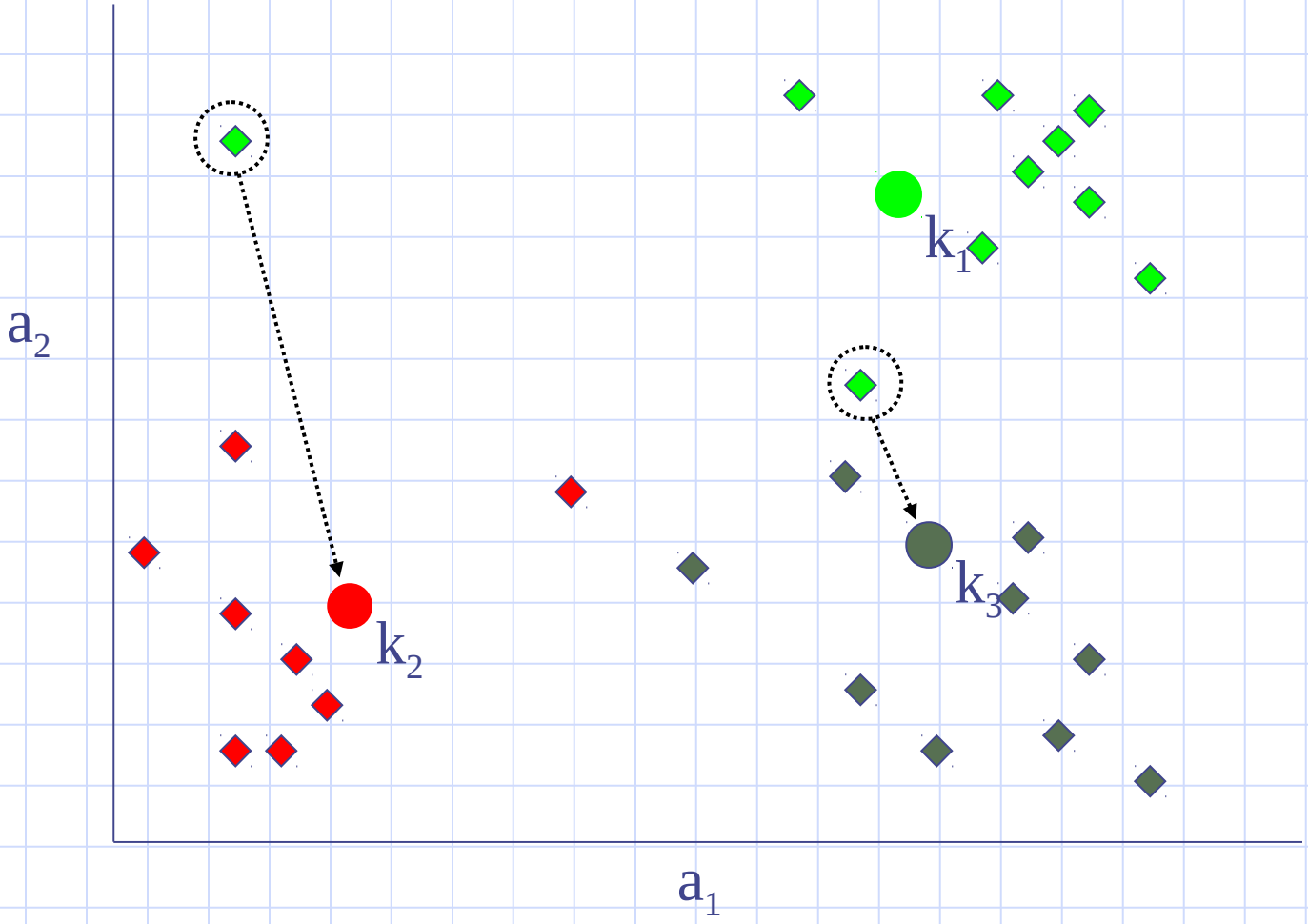
- Atualizar os protótipos (centróides) dos grupos:






From the slides by Gregory Piatetsky-Shapiro & Gary Parker - available at www.kdnuggets.com

- Calcular dissimilaridades entre objetos e centróides;
- Atualizar *clusters* (regra do vizinho mais próximo);




- Repetir até convergência/ número de iterações.

# Algoritmo básico:

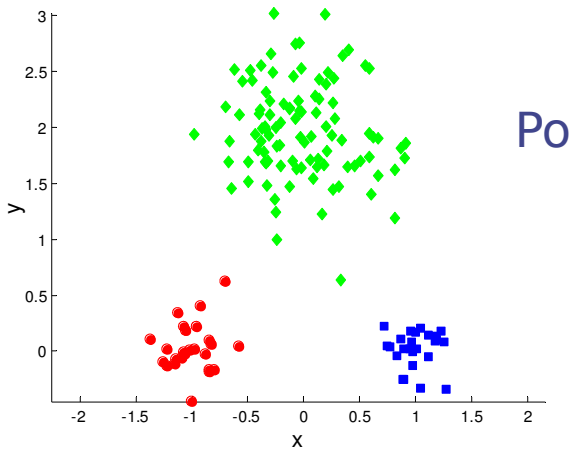
1. Selecionar *k* pontos (*centróides* iniciais);
2. Repetir até “convergir”:
   2.1 Formar *k* grupos atribuindo cada ponto ao seu centróide mais próximo;
   2.2 Re-computar o centróide de cada grupo;

# Detalhes sobre o *k*-médias:

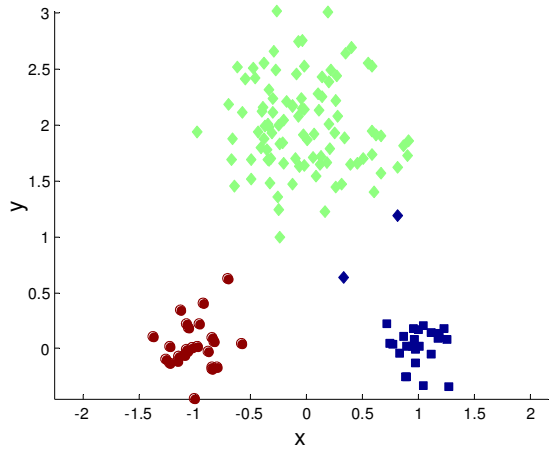
- Centróides iniciais são frequentemente escolhidos aleatoriamente.
  - *Clusters* obtidos podem variar de uma rodada para outra?
- *Proximidade* pode ser medida por meio de Distância Euclidiana, co-seno, correlação, etc.
- *k*-média converge, geralmente nas primeiras iterações;

- Vejamos alguns exemplos interessantes...

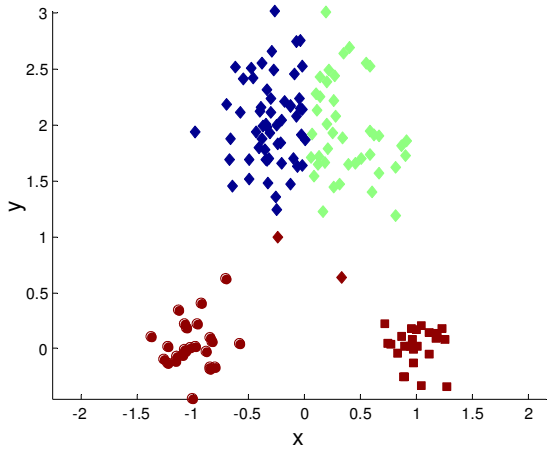
# Consideremos duas partições diferentes obtidas para $k = 3$:




Pontos originais




Partição ótima

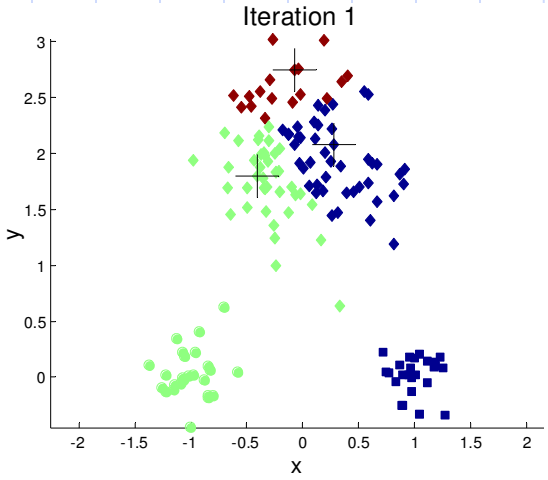


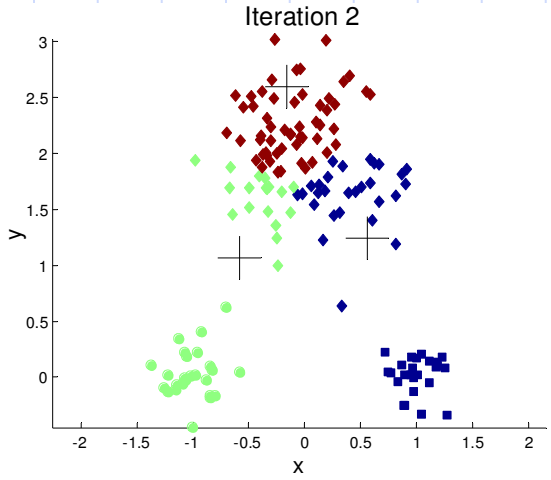

Partição Sub-ótima

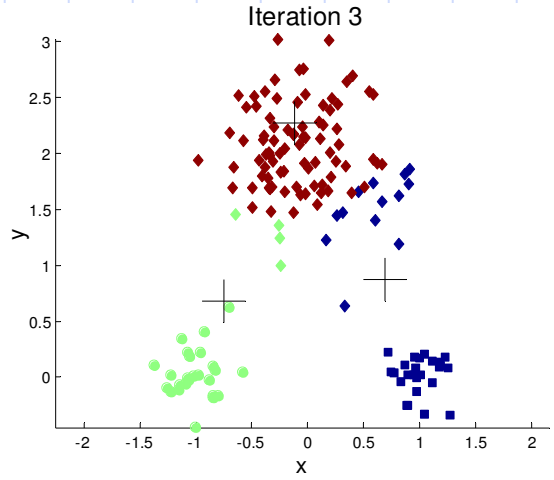


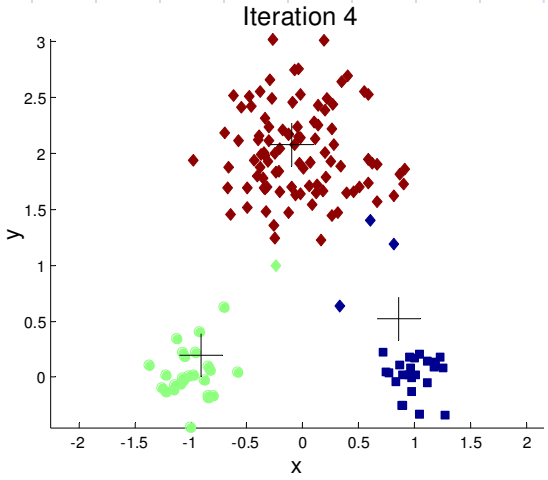
Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

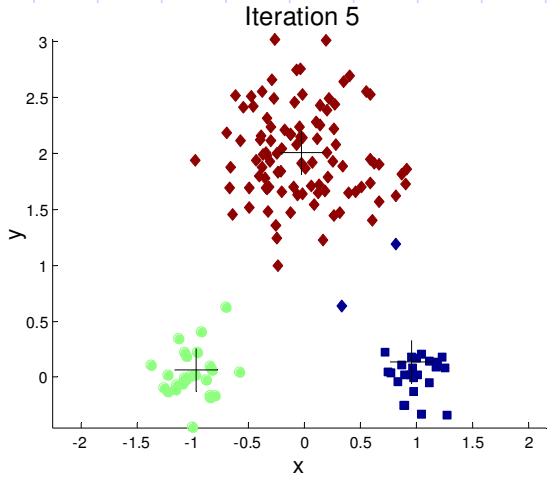
# Importância da escolha dos centróides iniciais:

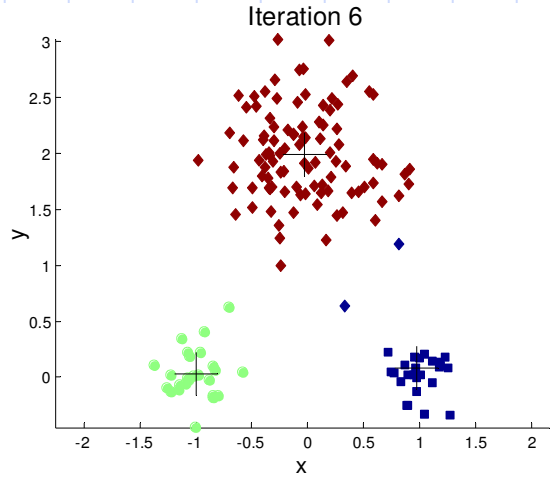


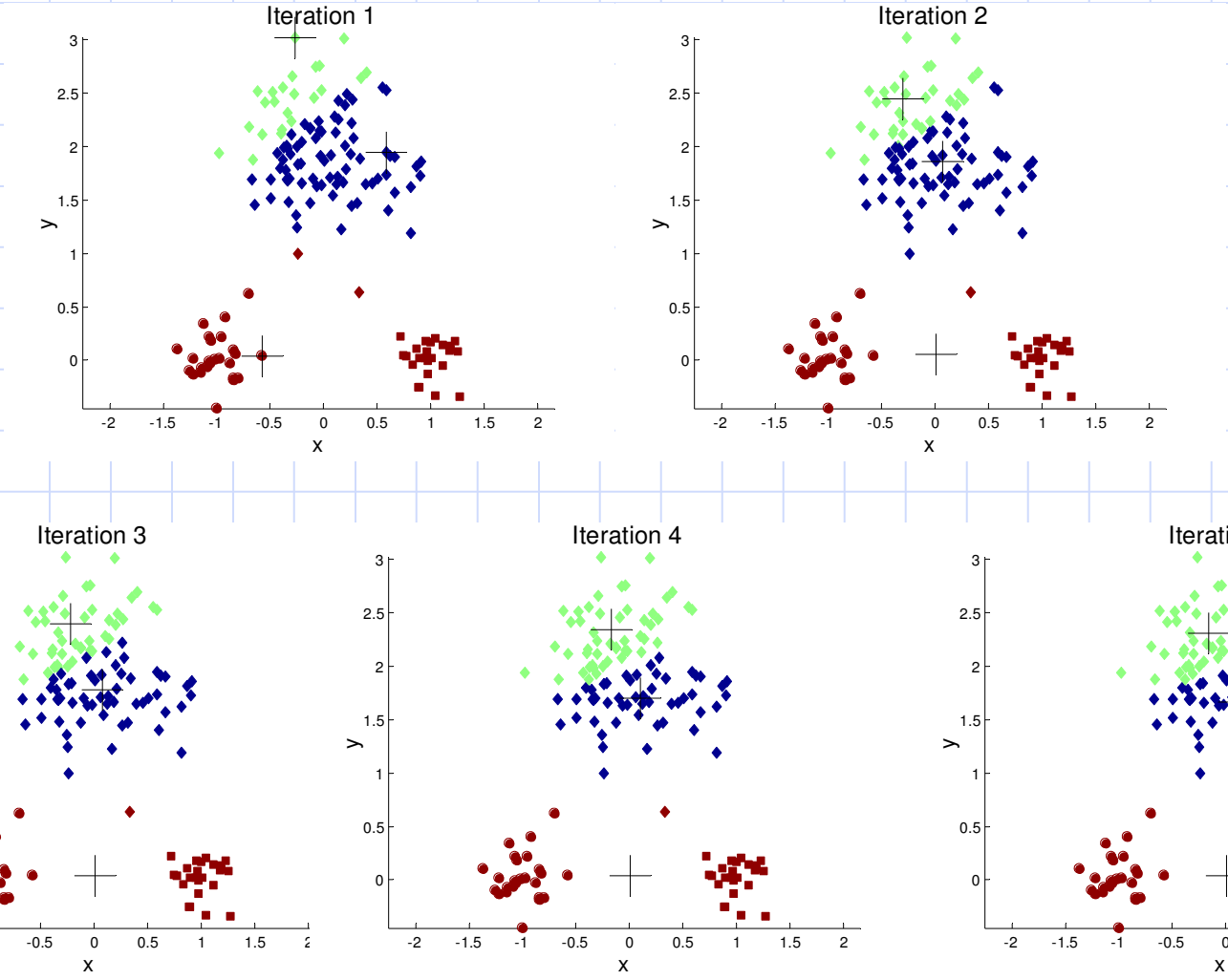






Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Importância da escolha dos centróides iniciais...




Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Soluções para inicialização?

- Múltiplas execuções:
  - Ajuda, mas pequena $P_{sucesso}$ ;
- Amostragem via métodos hierárquicos;
- Seleção “informada” de centróides distantes entre si;
- Algoritmos de busca (e.g., evolutivos);

# Avaliando os grupos obtidos:

- Soma dos erros quadráticos:

$$SEQ = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$$

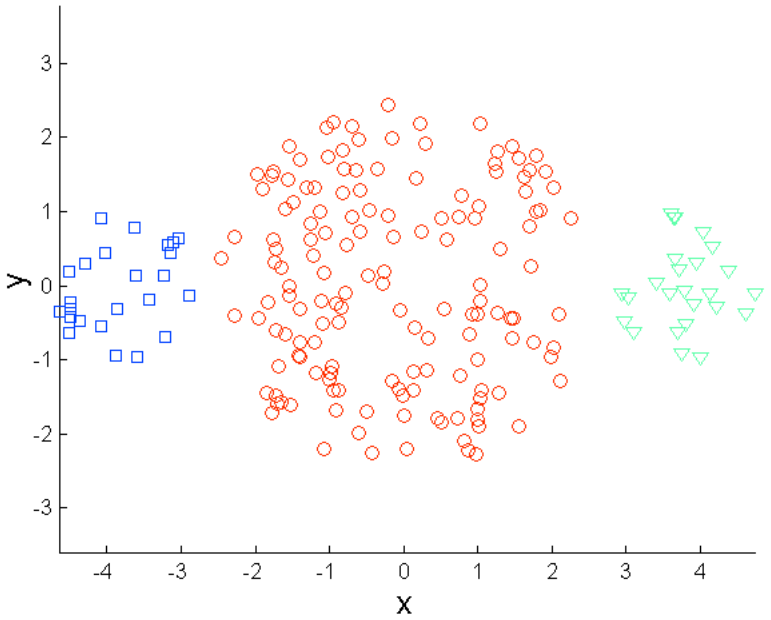
  - Dadas duas partições, escolher aquela que apresenta SEQ menor;
  - Aumento de *k* : tende a diminuir, por si só, SEQ;

# Limitações do *k*-médias:

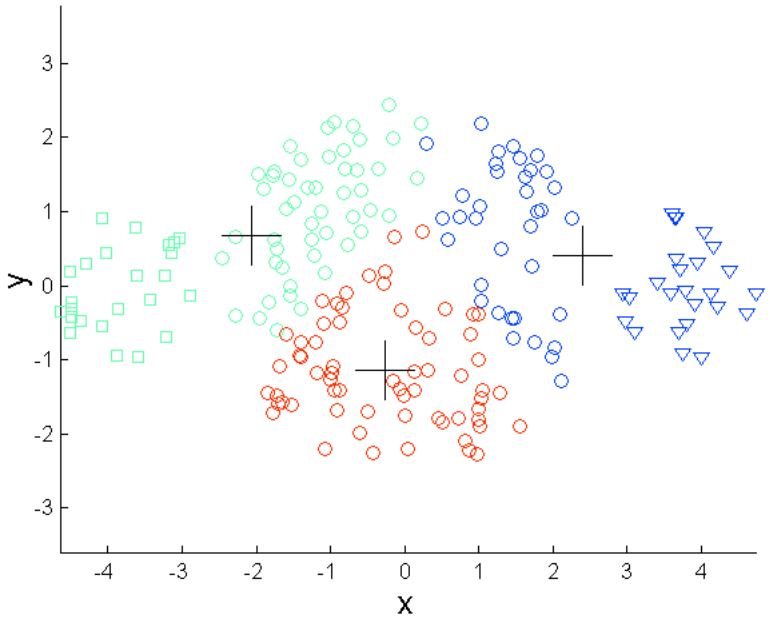
- Grupos de diferentes:
  - Tamanhos;
  - Densidades;
  - Formas não globulares.

- *Outliers*.

# Limitações do *k*-médias: grupos de tamanhos diferentes

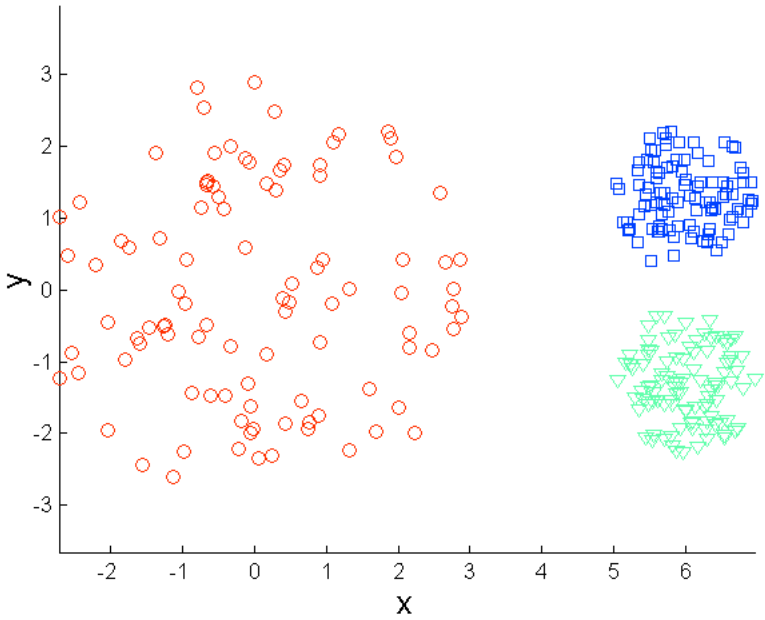



Pontos originais

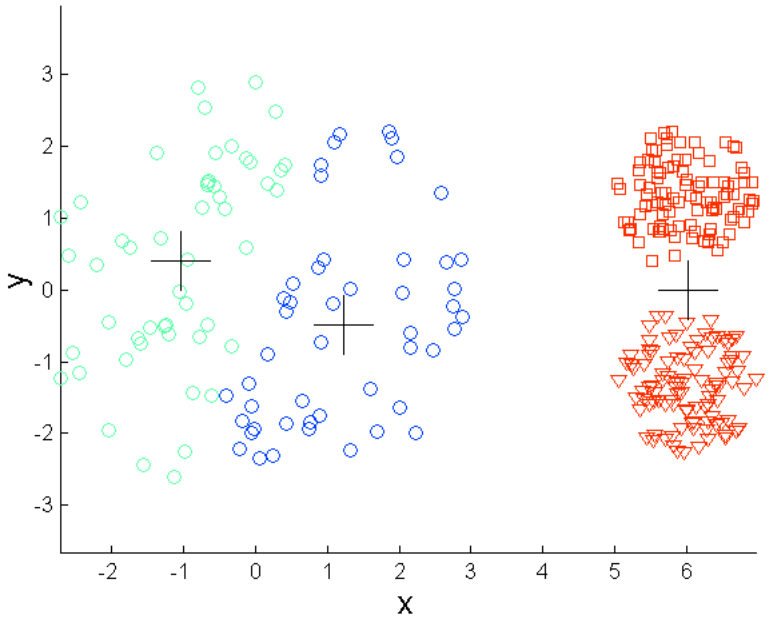



*3*-médias

# Limitações do *k*-médias: densidades diferentes




Pontos originais

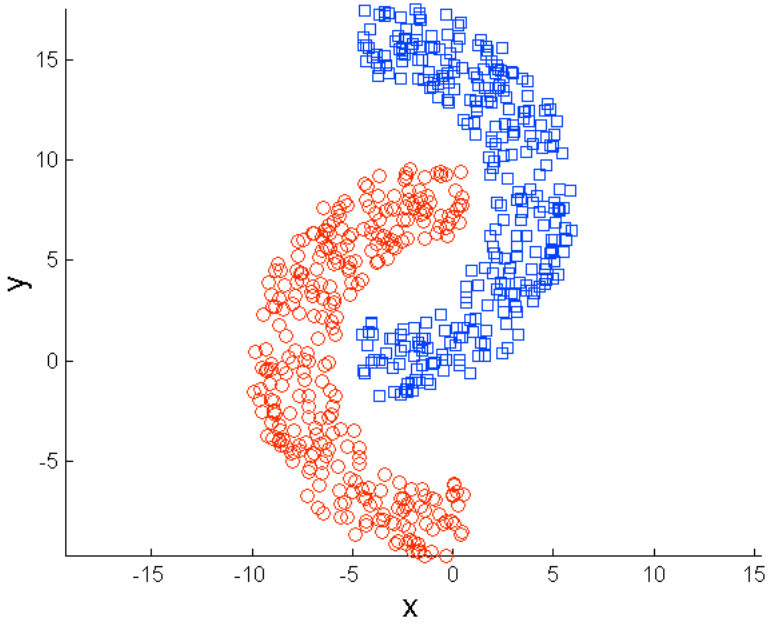



3-médias

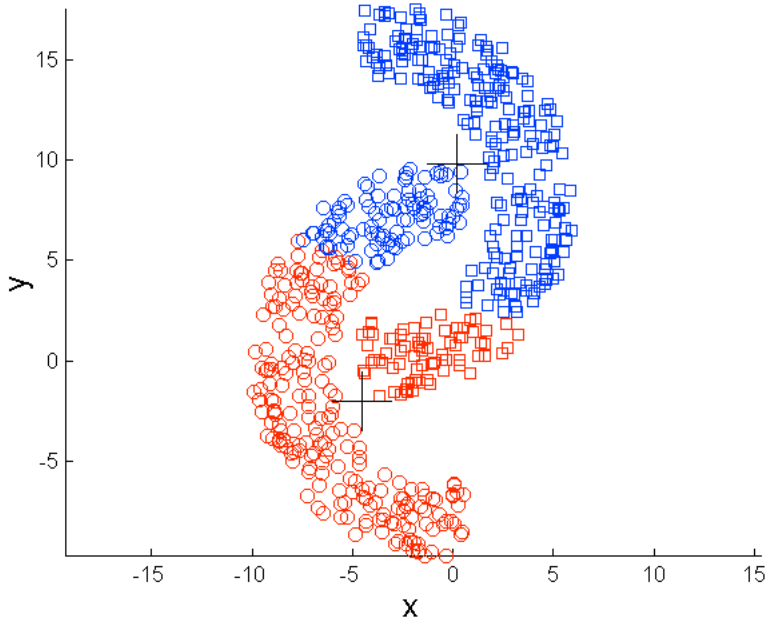


Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Limitações do *k*-médias: formas não globulares




Pontos originais

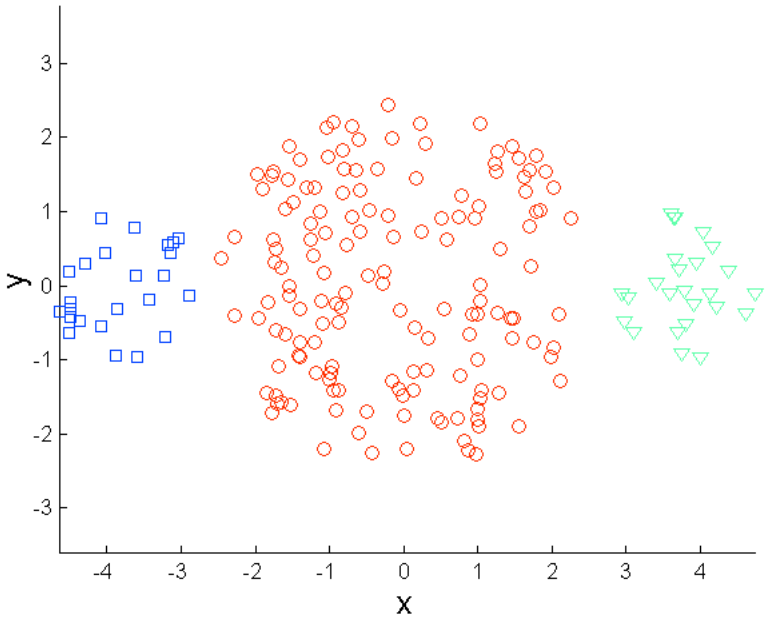



*2*-médias

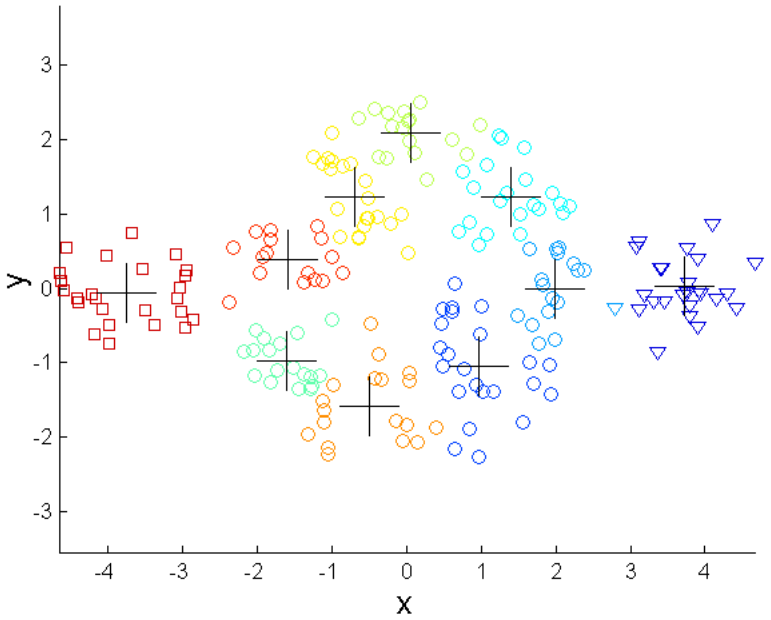


Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Superando algumas limitações do *k-médias*




Pontos originais

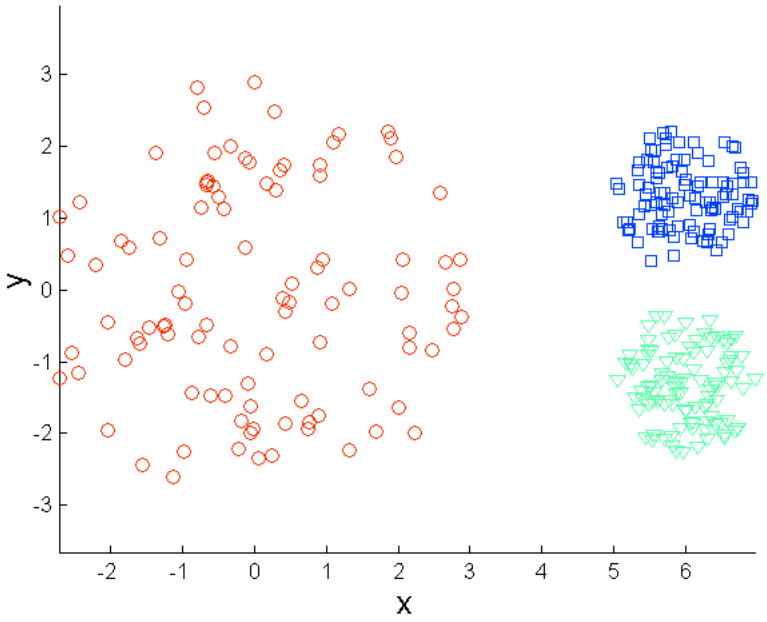



“Mais Grupos”

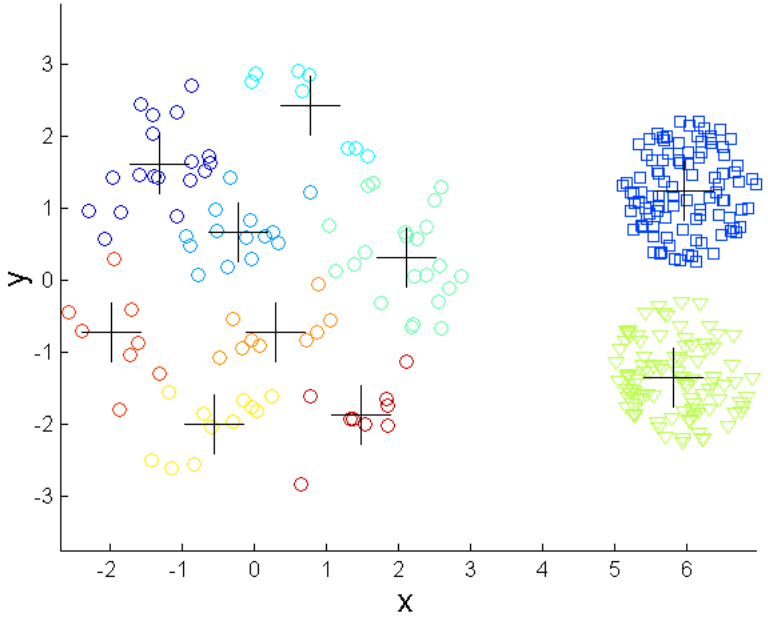


Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Superando algumas limitações do *k-médias*...




Pontos originais

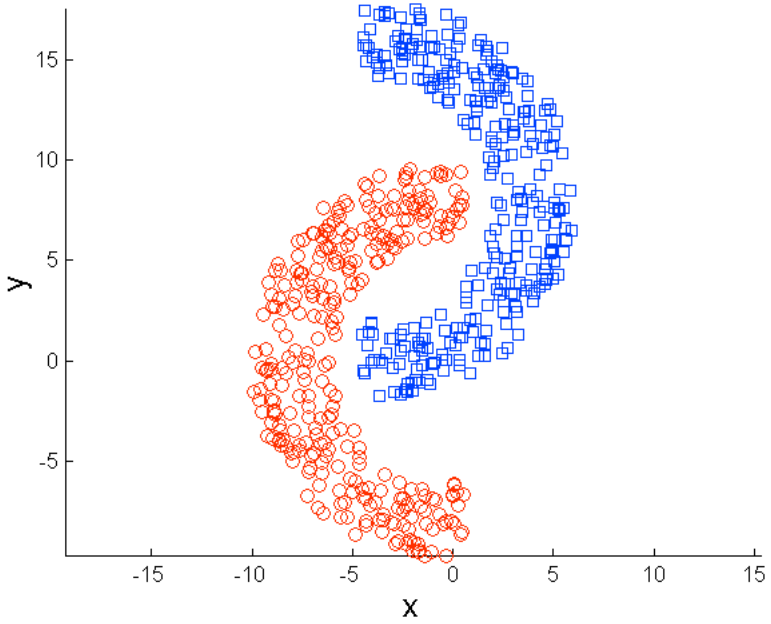


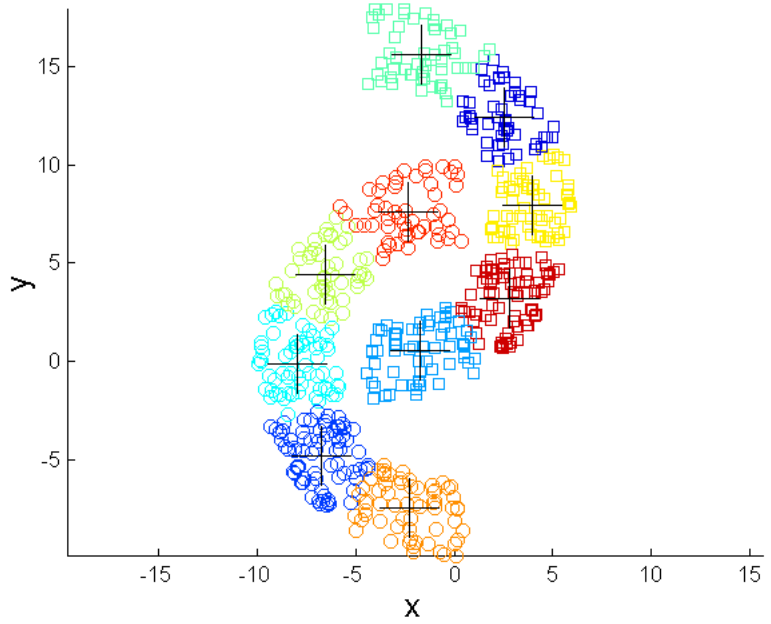

"Mais grupos"



Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Superando algumas limitações do *k-médias*...




Pontos originais

"Mais grupos"



Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Validação

# **Critérios de validade**:

Medidas numéricas para julgar diversos aspectos de validade:

- Índices Internos: medem a qualidade de uma estrutura de agrupamento independentemente de informação externa (e.g., EQ);
- Índices Relativos: comparam duas partições (ou grupos).
- Índices Externos: medem o quão bem os rótulos dos grupos representam categorias pré-estabelecidas (e.g., Rand);

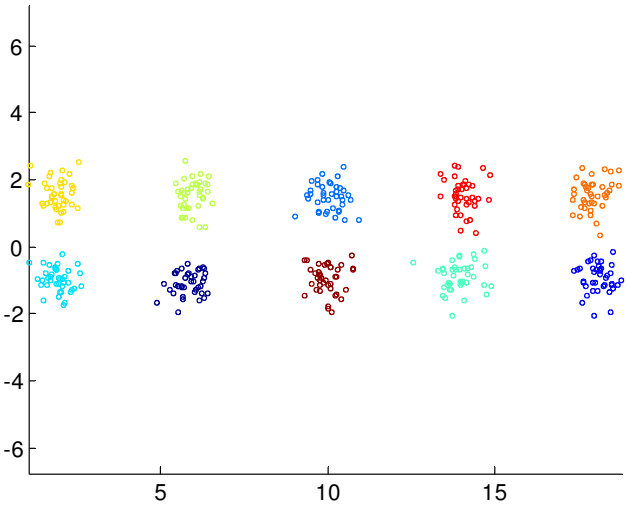
➢ Iniciaremos por um índice interno que já foi "indiretamente" estudado...

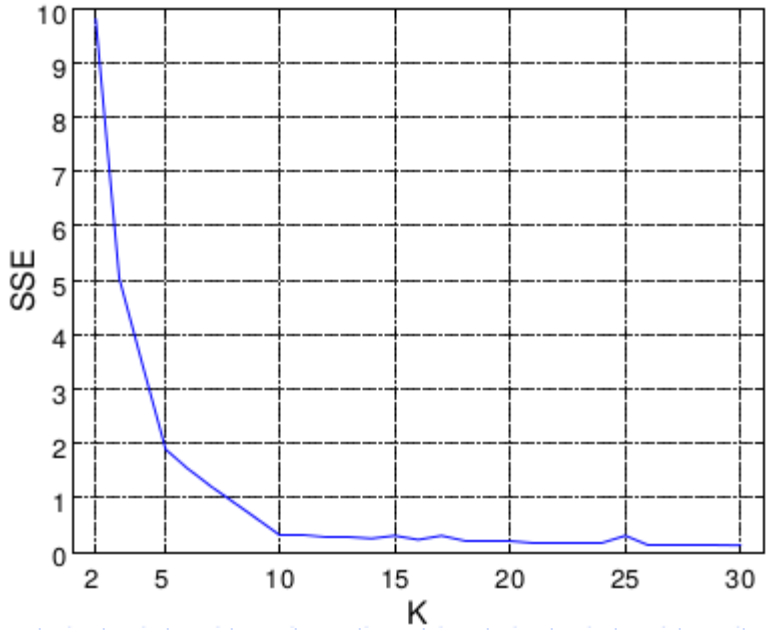


Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Erro Quadrático (EQ):

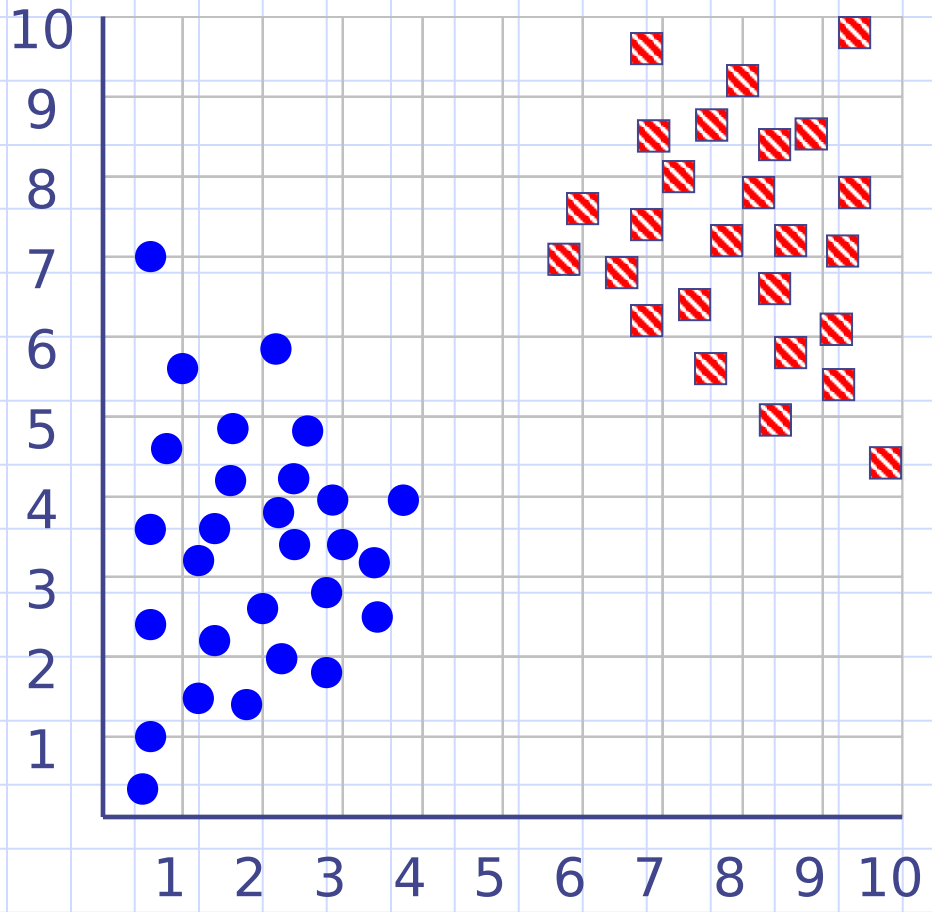
- Usado para medir a qualidade da estrutura de grupos obtida sem se considerar informações externas;
- Comparar partições (mesmo *k*) e grupos (EQ médio) ;
- Pode ser usado para estimar o número de grupo (com cuidado!):



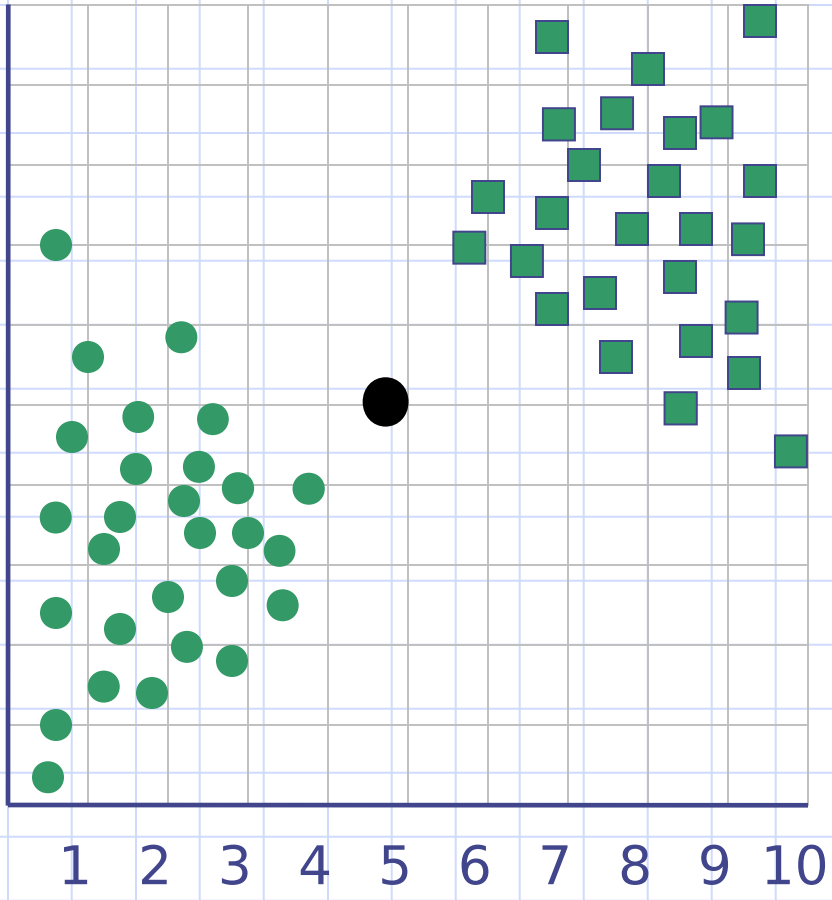

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

# Estimando o número de grupos:



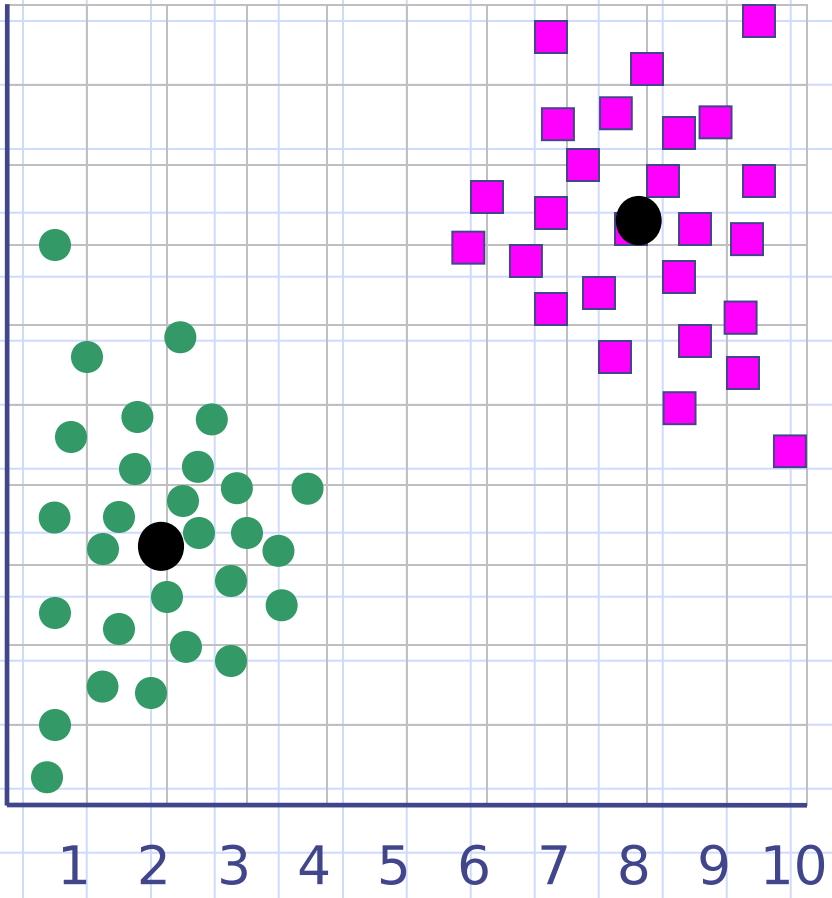

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

# Para $k = 1$, EQ = 873.

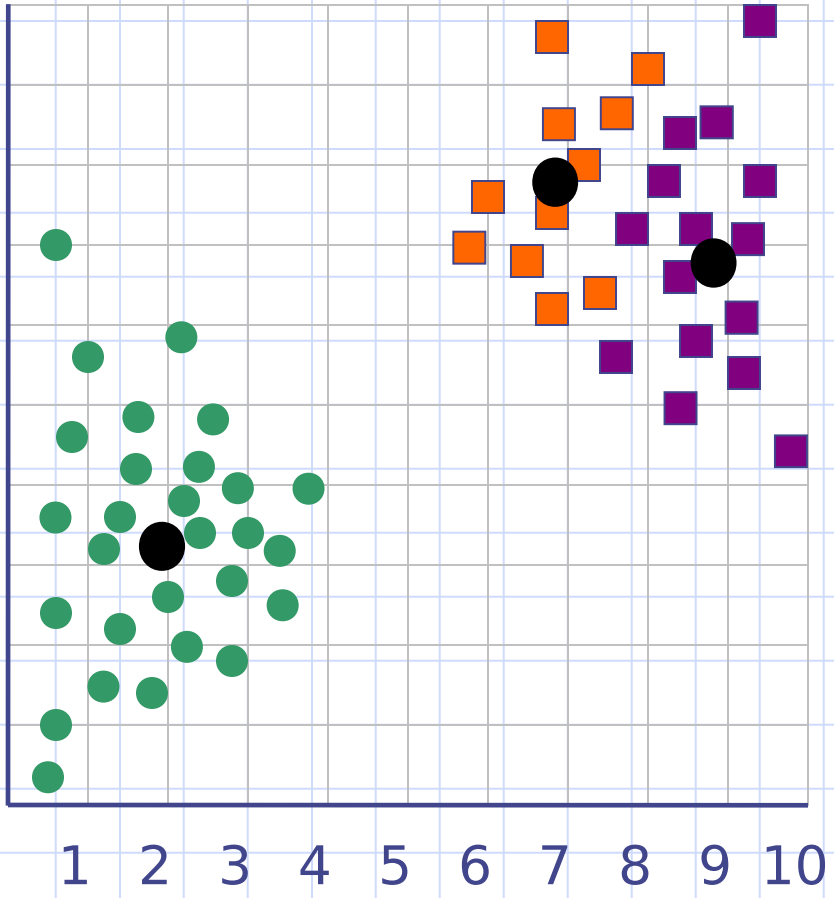



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

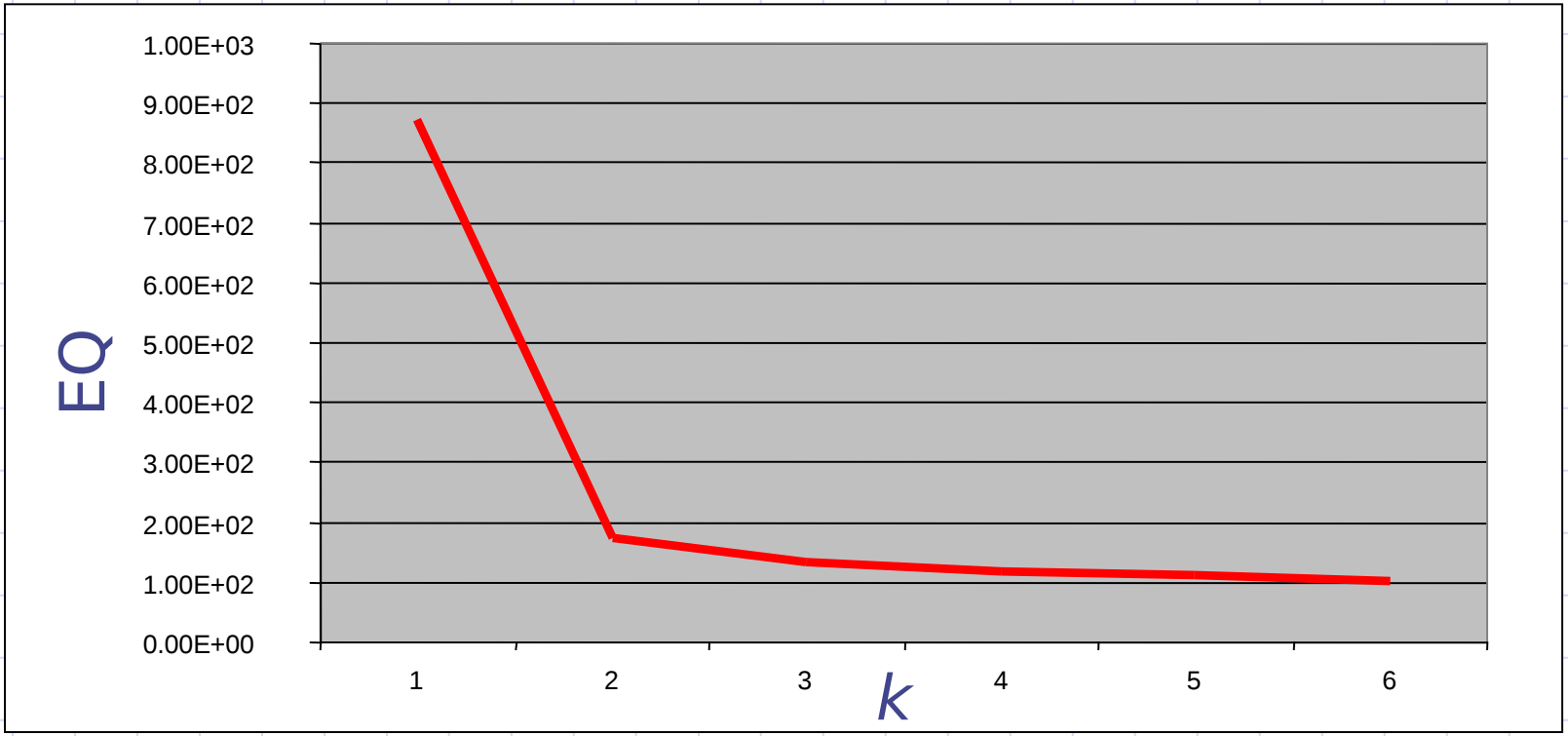
# Para $k = 2$, EQ = 173.



# Para $k$ = 3, EQ = 133.



# Plotar EQ para $k$ =1,...6 , tentando identificar um joelho:

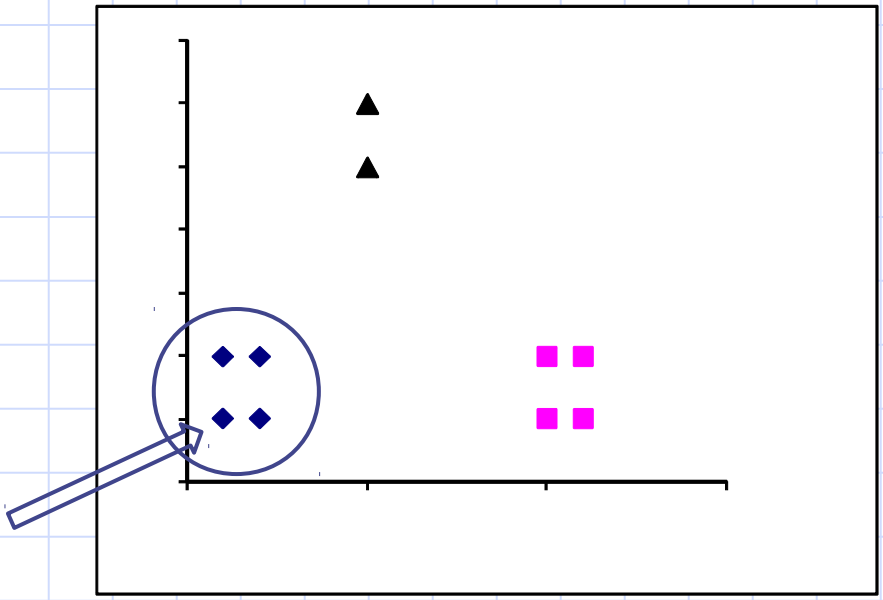


# Índices Relativos:

- Normalmente usados para comparar partições diferentes e para estimar o número de grupos de dados;

# Silhueta (Kaufman & Rousseeuw, 1990):

| *i* | *x* | *y* | *cluster* |
|---|---|---|---|
| 1 | 1 | 1 | $\mathbf{C}_1$ |
| 2 | 1 | 2 | $\mathbf{C}_1$ |
| 3 | 2 | 1 | $\mathbf{C}_1$ |
| 4 | 2 | 2 | $\mathbf{C}_1$ |
| 5 | 10 | 1 | $\mathbf{C}_2$ |
| 6 | 10 | 2 | $\mathbf{C}_2$ |
| 7 | 11 | 1 | $\mathbf{C}_2$ |
| 8 | 11 | 2 | $\mathbf{C}_2$ |
| 9 | 5 | 5 | $\mathbf{C}_3$ |
| 10 | 5 | 6 | $\mathbf{C}_3$ |

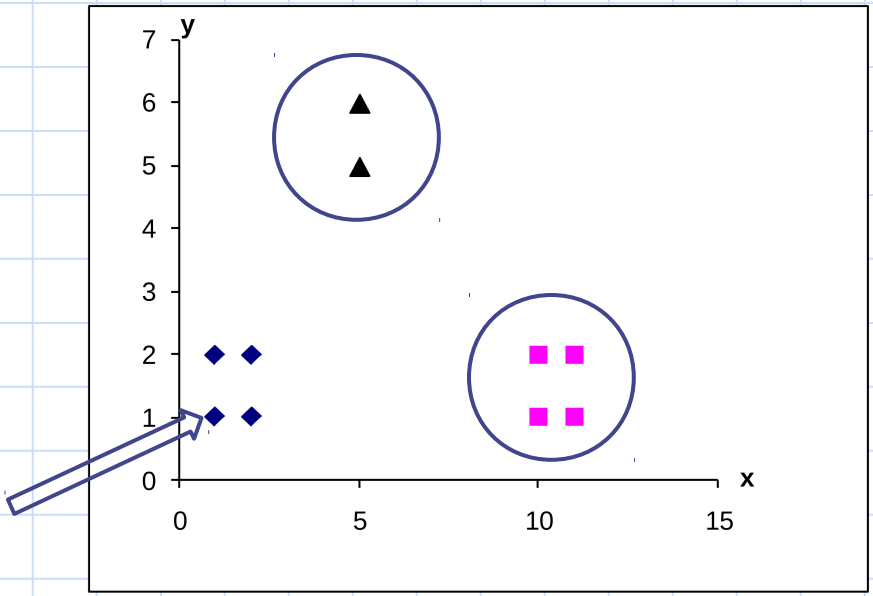


- Quão bem "classificado" está o *i*-ésimo objeto (e.g., *i*=1)?

1)Dissimilaridade em relação aos objetos do seu grupo:

- ***a*(i):** dissimilaridade média de *i* aos objetos do seu grupo:
  - $(d_{12} + d_{13} + d_{14})/3$.

| *i* | *x* | *y* | *cluster* |
|---|---|---|---|
| 1 | 1 | 1 | $\mathbf{C_1}$ |
| 2 | 1 | 2 | $\mathbf{C_1}$ |
| 3 | 2 | 1 | $\mathbf{C_1}$ |
| 4 | 2 | 2 | $\mathbf{C_1}$ |
| 5 | 10 | 1 | $\mathbf{C_2}$ |
| 6 | 10 | 2 | $\mathbf{C_2}$ |
| 7 | 11 | 1 | $\mathbf{C_2}$ |
| 8 | 11 | 2 | $\mathbf{C_2}$ |
| 9 | 5 | 5 | $\mathbf{C_3}$ |
| 10 | 5 | 6 | $\mathbf{C_3}$ |




2) O que se pode dizer em relação à sua dissimilaridade relativa aos objetos do grupo vizinho?

***b*(i):** $\min\{ (d_{15} + d_{16} + d_{17} + d_{18})/4 \ ; \ (d_{19} + d_{110})/2 \}$.

**Silhueta** (Kaufman & Rousseeuw, 1990):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$S_M = \frac{\sum_{i=1}^{N} s(i)}{N}$$

***a*(i):** dissimilaridade média de *i* aos outros objetos de seu grupo.

***b*(i):** dissimilaridade média de *i* em relação aos objetos de seu grupo vizinho.

→ **Aumentar *k* tende a diminuir *a*(*i*), mas também tende a diminuir *b*(*i*) .**

**Existem dezenas de índices reportados na literatura;**

Maior parte dos índices é baseada em conceitos de compactação e separabilidade...

# Índices Externos:

- Comparar uma partição obtida com uma partição de referência;
- Aplicações:
  - Experimentos controlados;
  - Prática?
- Consideremos:

- Um par de objetos $(\mathbf{x}_i, \mathbf{x}_j)$ de $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$;

- Uma partição dos dados $\mathbf{C}=\{\mathbf{C}_1,\mathbf{C}_2,...,\mathbf{C}_k\}$;

- Uma partição de referência $\mathbf{P}=\{\mathbf{P}_1,\mathbf{P}_2,...,\mathbf{P}_p\}$:

Há 4 casos possíveis para atribuir $(\mathbf{x}_i, \mathbf{x}_j)$:

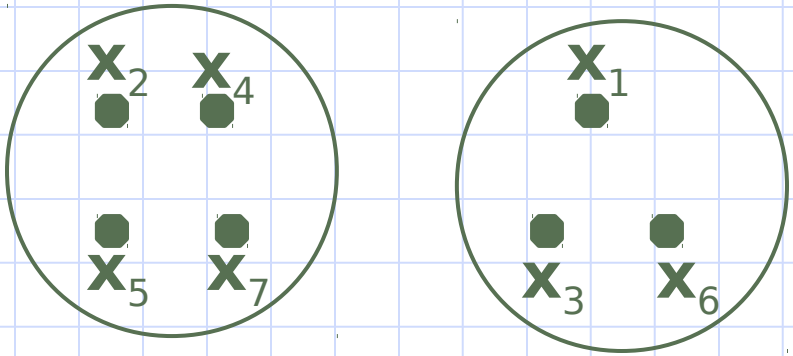
**Caso 1**: $\mathbf{x}_i$ e $\mathbf{x}_j$ estão no mesmo grupo em **C** e na mesma categoria em **P**; (a)

**Caso 2**: $\mathbf{x}_i$ e $\mathbf{x}_j$ estão no mesmo grupo em **C** e em categorias diferentes em **P**; (b)

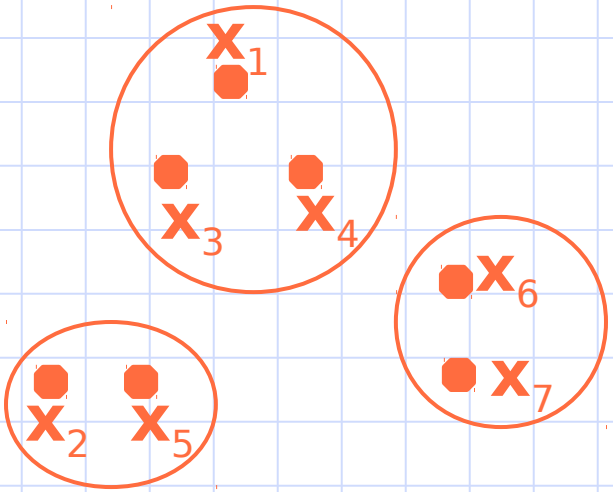
**Caso 3**: $\mathbf{x}_i$ e $\mathbf{x}_j$ estão em grupos diferentes em **C** e na mesma categoria em **P**; (c)

**Caso 4**: $\mathbf{x}_i$ e $\mathbf{x}_j$ estão em grupos diferentes em **C** e em categorias diferentes em **P**; (d)

# Exemplo (Xu & Wunch, 2009):




Partição de Referência – **P**.




*Clusters* obtidos – **C**.

$a$ = 2 (13,25)

$b$ = 3 (14,34,67)

$c$ = 7 (16,24,27,36,45,47,57)

$d$ = 9 (12,15,17,23,26,35,37,46,56)

Nº de pares:
$a+b+c+d = n(n-1)/2 = 21$.

# Alguns índices muito usados:

Jaccard (1908): $J = \frac{a}{a+b+c}$

Principais referências usadas para preparar essa aula:

- Xu, R., Wunsch, D., **Clustering**, IEEE Press, 2009.
  - Capítulo 4.

- Tan, Steinbach & Kumar, **Introduction to Data Mining**, Pearson, 2006.
  - Capítulo 8, pp. 496-515.

- Jain, A. K., Dubes, R. C., **Algorithms for Clustering Data**, Prentice Hall, 1988.
  - Capítulo 3, pp. 89-142.

- Bishop, C. M., **Pattern Recognition and Machine Learning**, 2006.
  - Capítulo 9, pp. 423-439.