Mineração de Dados 2017.2

Agrupamento de dados

Thiago Ferreira Covões

(slides baseados no material do Prof. Eduardo Hruschka [erh@icmc.usp.br])

1. Motivação

Diversas ciências se baseiam na organização de objetos de acordo com suas similaridades;

- Biologia:

Reino: Animalia
Ramo: Chordata
Classe: Mammalia
Ordem: Primatas
Família: *Hominidae*
Gênero: *Homo* (homem moderno e parentes)
Espécie: *Homo sapiens*

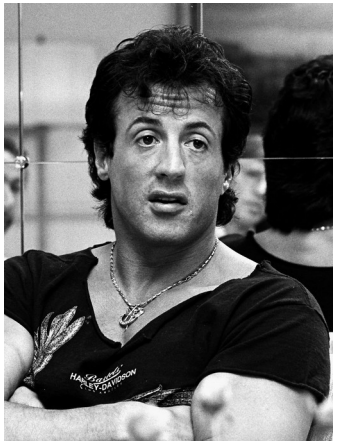
Humanos se interessam por *categorizações*:

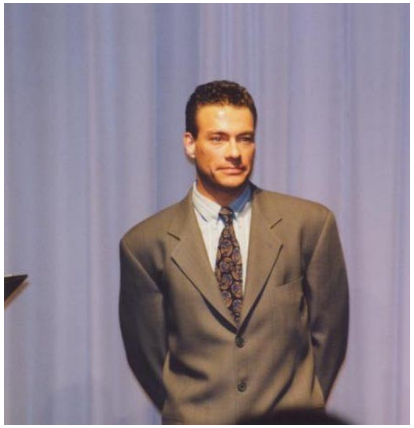
- Música: erudita, popular, religiosa, etc..

- Filmes:

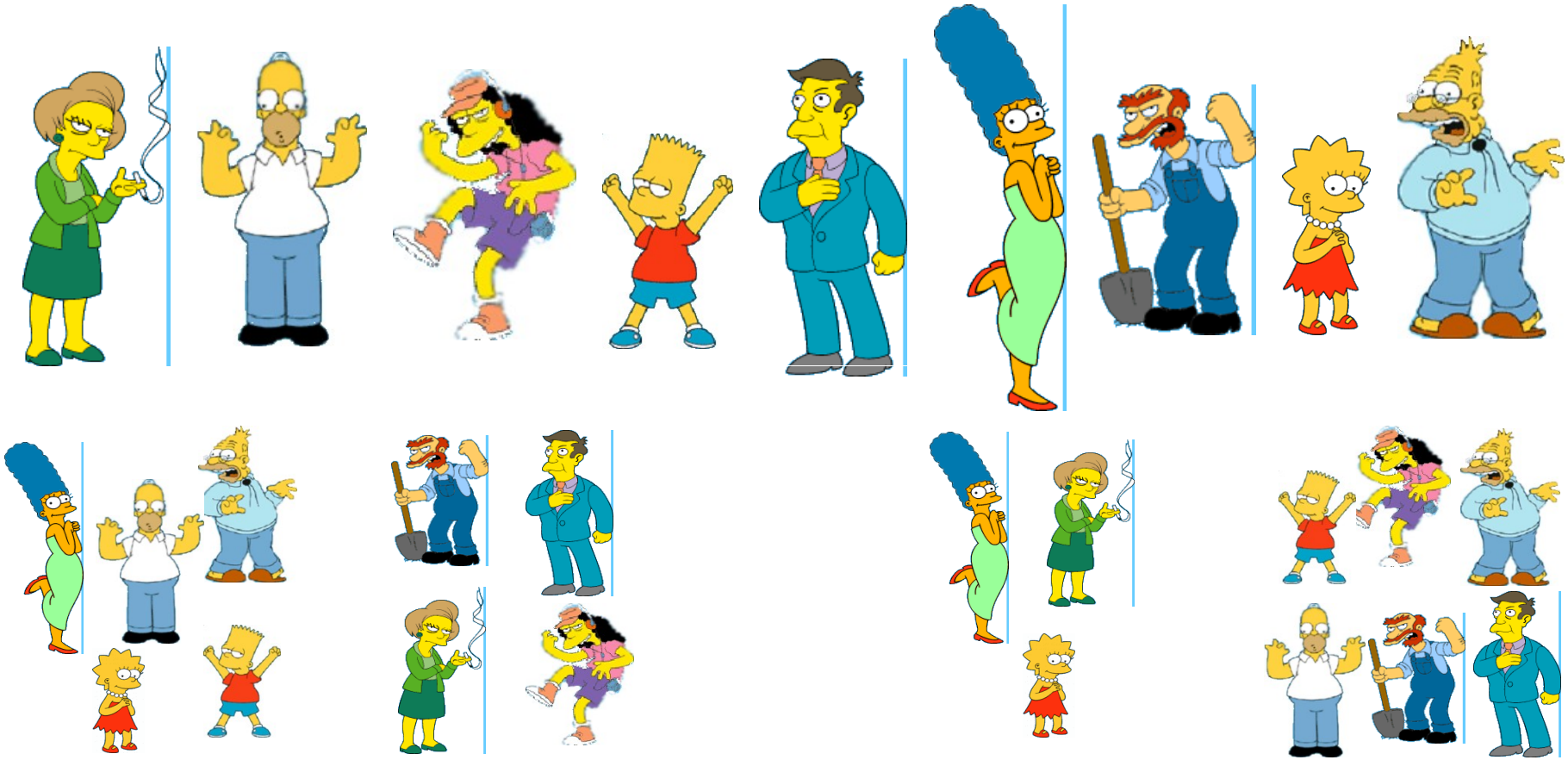
- Animação, Aventura, Comédia, Drama, Musical, etc...

- Agrupamento de atores:





Como *agrupar naturalmente* os seguintes objetos?



Família Empregados Mulheres Homens

→ *Cluster* é um conceito subjetivo!

O que é um grupo ?

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

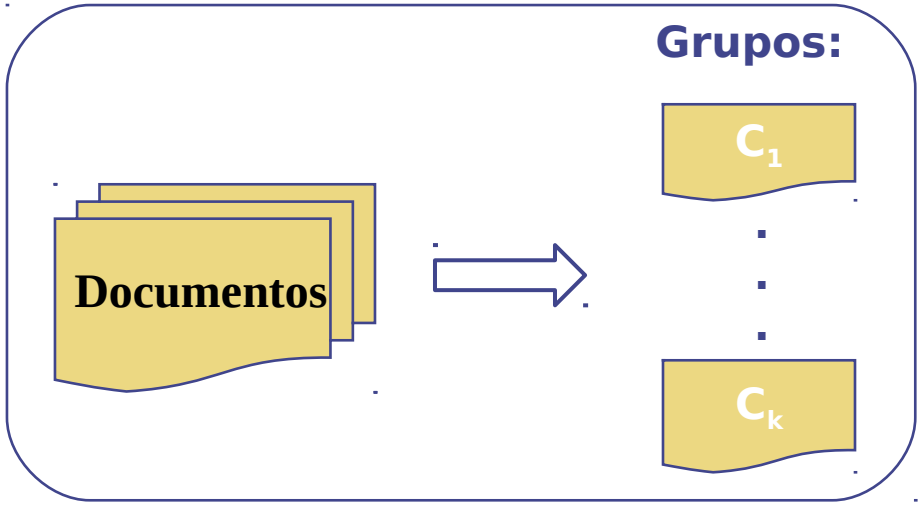
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

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

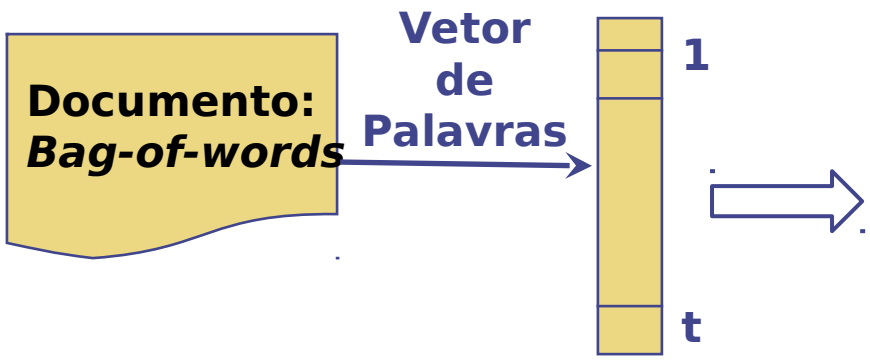
- Apesar das dificuldades apresentadas, a literatura sobre ADs é rica e bem estabelecida;

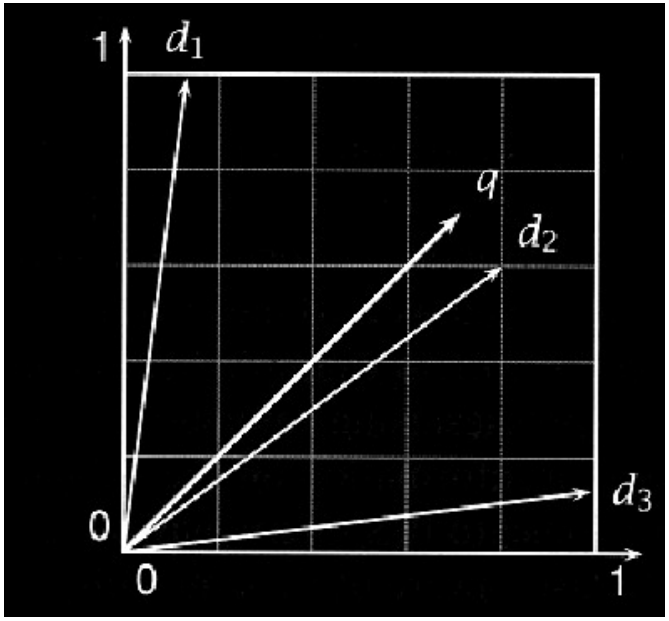
- Há medidas de dis(similaridade) bem estudadas e fundamentadas para diversos tipos de dados (e também para diversos domínios de aplicação):
  - Dados Numéricos;
  - Dados Categóricos / Nominais;
  - Dados Binários;
  - Etc.

- Mineração de Textos:



Como?





Aplicações?

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, ...

Data science/machine learning methods you used in the past 12 months: [732 voters] **(https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html)**

Regression 60%
Clustering 55%
Decision Trees/Rules 51%
K-NN 39%
…

- Gan et al. (2007) reportam uma vasta literatura sobre agrupamento de dados, que inclui:

- 13 *surveys*;
- 10 livros (de 1963 em diante);
- 76 periódicos que publicam artigos sobre ADs;
- 45 conferências que publicam sobre ADs;

- Xu & Wunsch (2009): *Web of Science* revela mais de 12.000 artigos usando o termo *cluster analysis* no título, ou nas palavras chaves, ou no resumo (oriundos de mais de 3.000 *journals* diferentes).

- Gan, G., Ma, C., Wu, J., **Data Clustering: Theory, Algorithms, and Applications**, SIAM Series on Statistics and Applied Probability, 2007.
- Xu, R., Wunsch, D., **Clustering**, IEEE Press, 2009.

ADs tem por objetivo principal organizar dados:

- Particões;
- Hierarquias.

- Aprendizado não supervisionado:
  - Inteligência Artificial;
  - Aprendizado de Máquina;
  - Reconhecimento de Padrões.

2. Conceitos Básicos

Algumas Definições (Everitt, 1974):

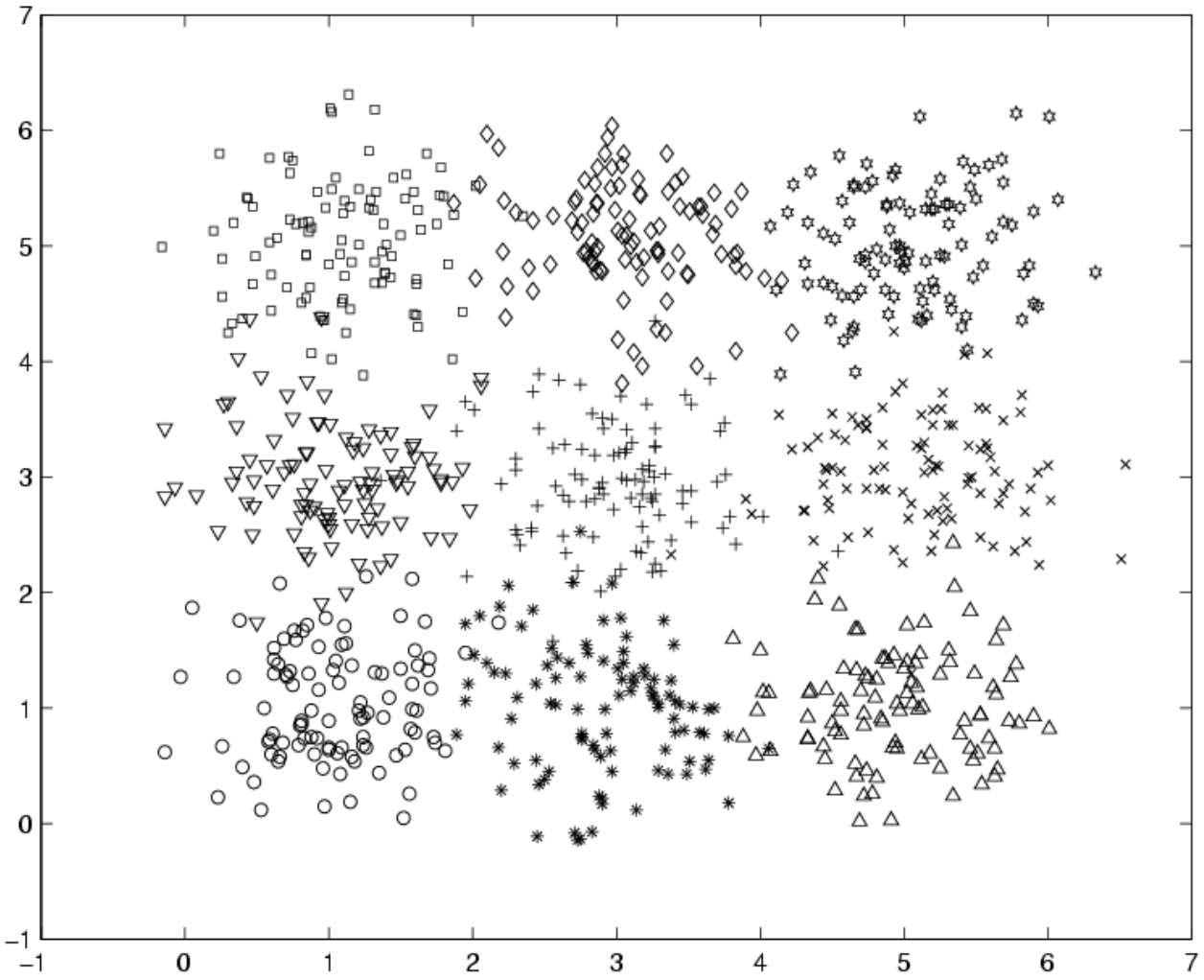
• *Um cluster (grupo) é um conjunto de entidades semelhantes, e entidades pertencentes a diferentes clusters não são semelhantes*.

• *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*.

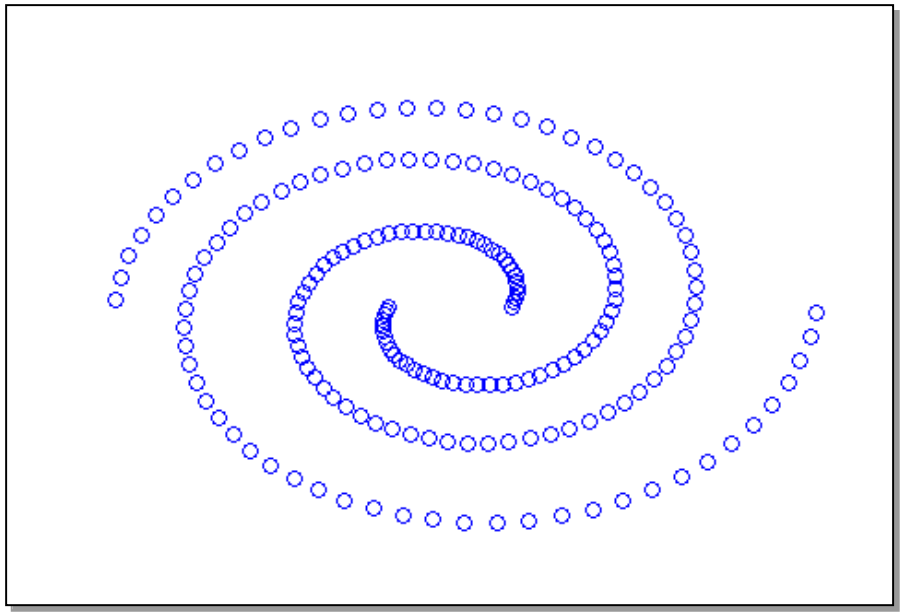
• *Grupos podem ser descritos como regiões conectadas de um espaço multidimensional contendo uma densidade de pontos relativamente alta, separada de outras tais regiões por uma região contendo uma densidade relativamente baixa de pontos*.

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

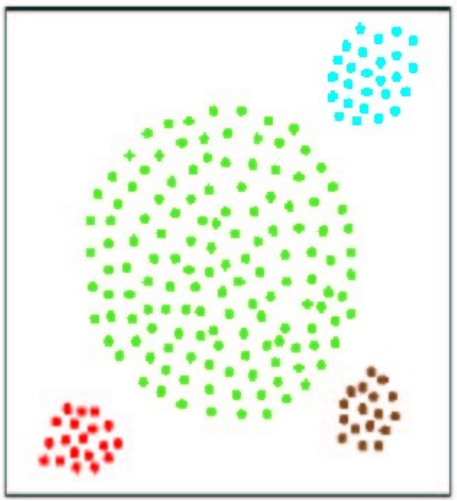
Quais são os grupos ?

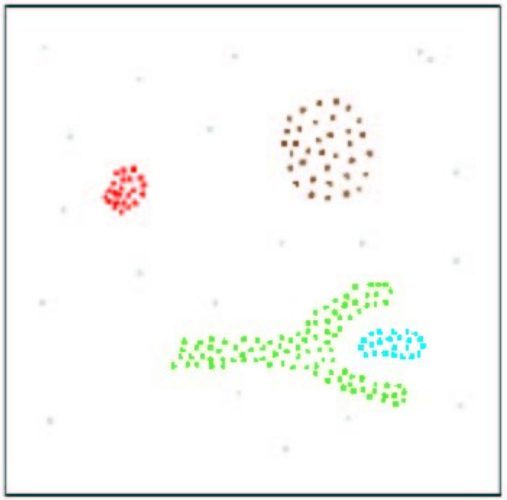


Quais são os grupos ? ...



Aplicações práticas?





I. Zhang, Density-Based Clustering Algorithms, *cs.hku.hk/~dbgroup/seminar/*

- Algoritmos para agrupamento de dados induzem *clusters*;

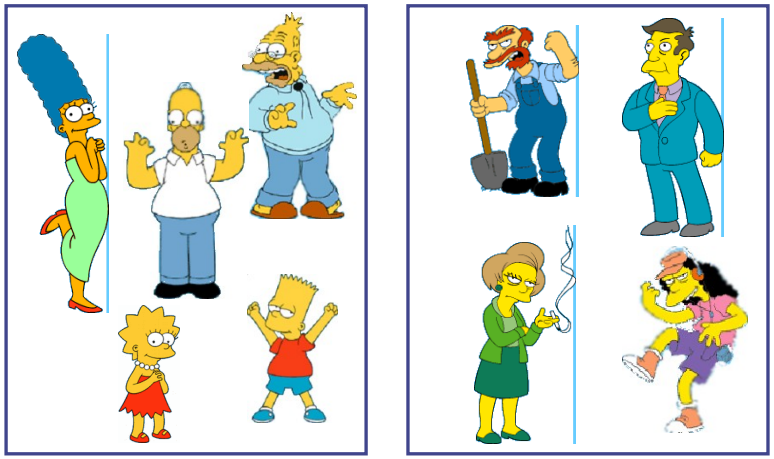
- Conceito semelhante ao de tendência (*bias*) indutiva estudado em aprendizado de máquina;

- Medidas de dis(similaridade), índices de validade relativos, parâmetros definidos pelo usuário, etc. (dependente do domínio/problema)

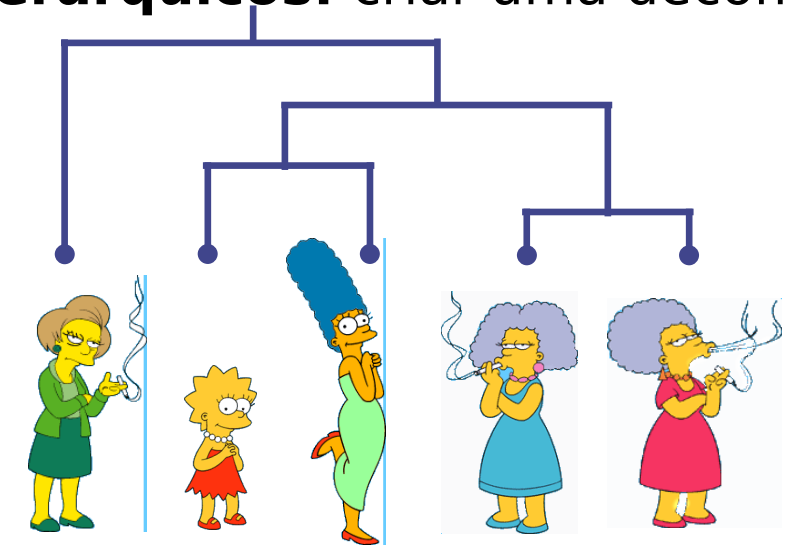
- Sob o ponto de vista de AM: *projetista define o que o computador pode aprender.*

- *Existem centenas de algoritmos...*

Algoritmos para particionamento: construir partições.

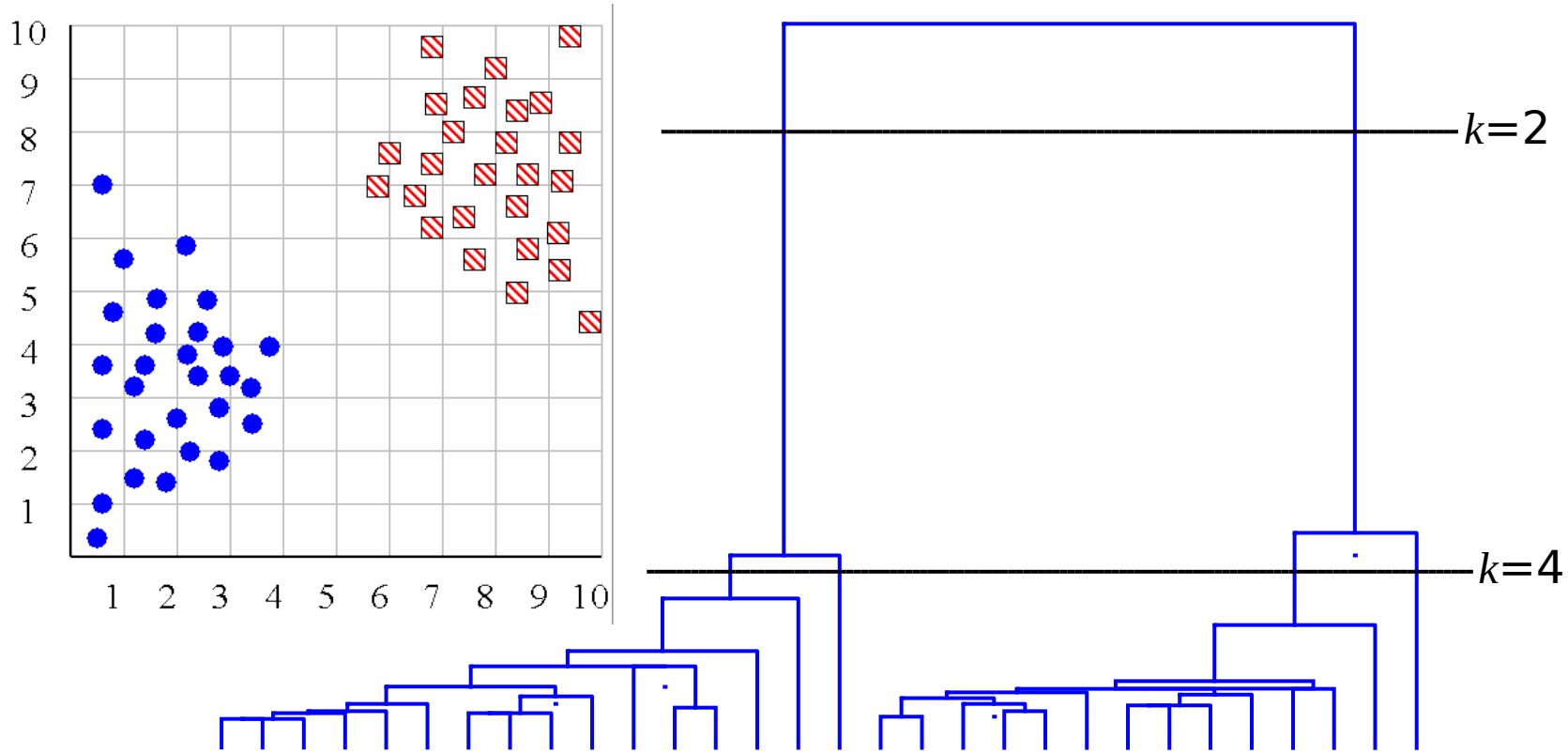


Algoritmos hierárquicos: criar uma decomposição hierárquica.



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

Métodos Hierárquicos:



- Conjunto de “partições aninhadas”;
- Número de grupos ?

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

Independentemente do método (particional ou hierárquico) o principal objetivo do agrupamento de dados é:

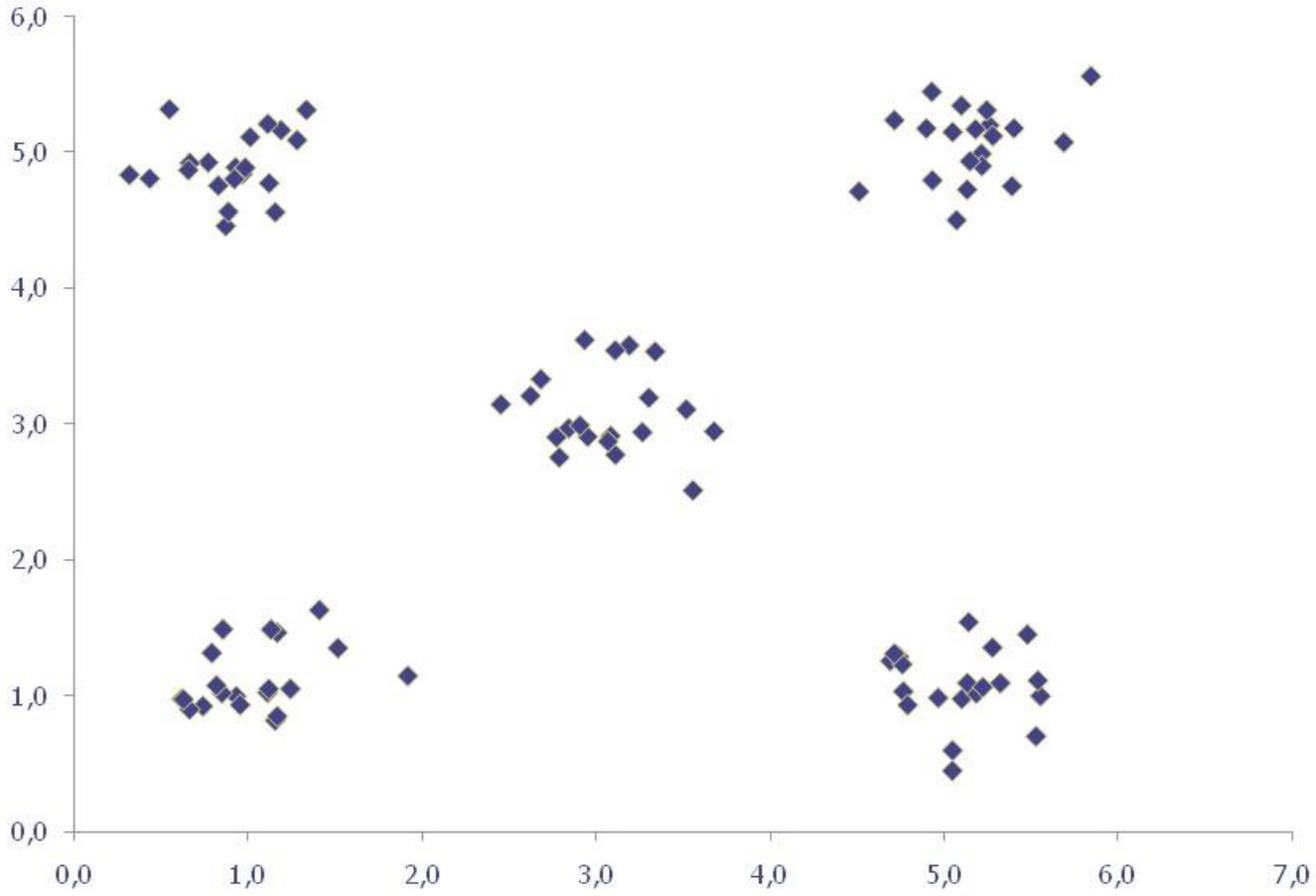
- Maximizar a homogeneidade interna ao *cluster* e a heterogeneidade entre diferentes *clusters*.

→ Objetos que pertencem ao mesmo *cluster* devem ser mais semelhantes entre si do que em relação a objetos de outros clusters;

→ Medidas de (Dis)similaridade possuem polarizações (biases):

- vantagens/desvantagens dependentes do domínio: Distância Euclidiana, Correlação de Pearson, Coseno, etc.

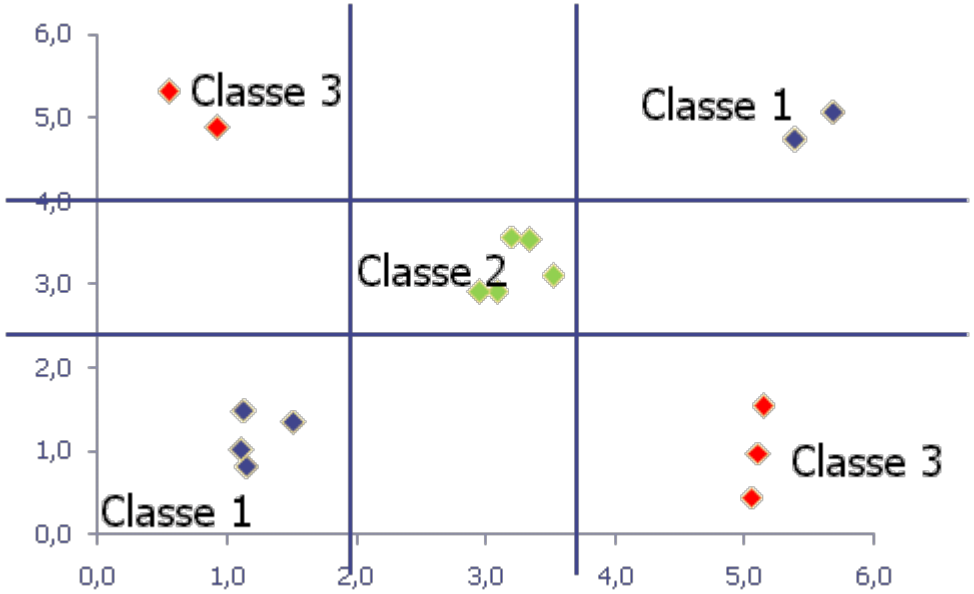
Agrupamento X Classificação?

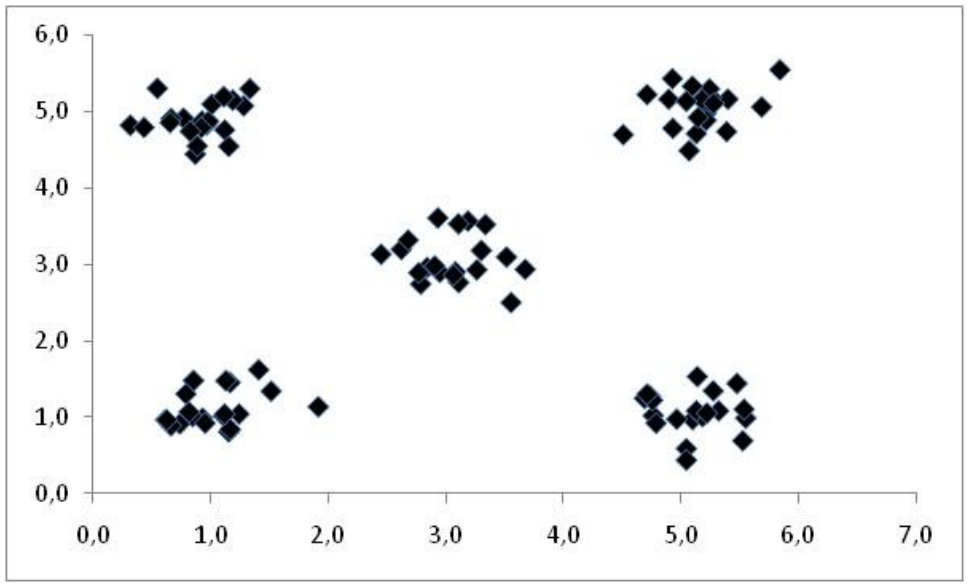


Agrupamento: Indução de grupos a partir da base de dados...

- Grupos obtidos serão então cuidadosamente estudados.

Agrupamento X Classificação?





Base de treinamento com dados rotulados:

- classificador (modelo)

Rotular dados de teste em função do modelo obtido.