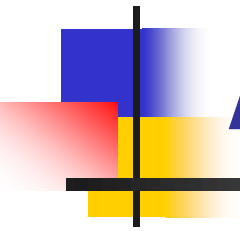
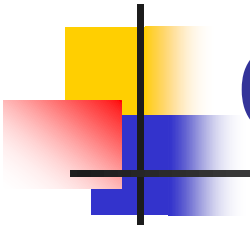


SCC5895 – Análise de Agrupamento de Dados

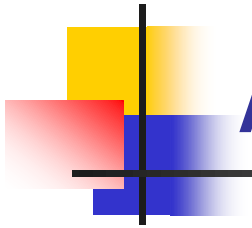
Introdução

Prof. Eduardo Raul Hruschka
PPG-CCMC / ICMC / USP



Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - Elaborados por Eduardo R. Hruschka e Ricardo J. G. B. Campello
  - de (Tan et al., 2006)
  - de E. Keogh (SBBD 2003)
  - de G. Piatetsky-Shapiro (KDNuggets)



Aula de Hoje

- Motivação
- Conceitos Básicos

Motivação

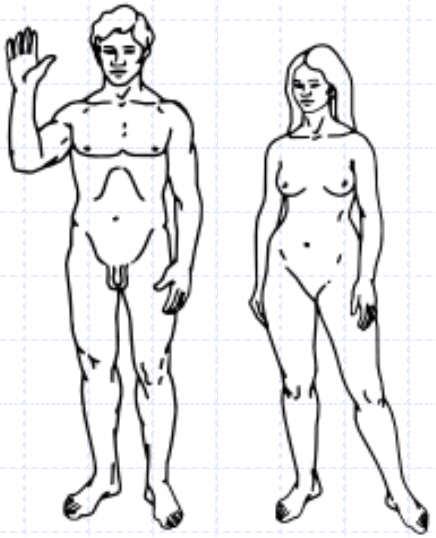
Humanos se interessam por *categorizações*

- Música: erudita, popular, religiosa, etc.

- Filmes: Animação, Aventura, Comédia, Drama, etc.

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

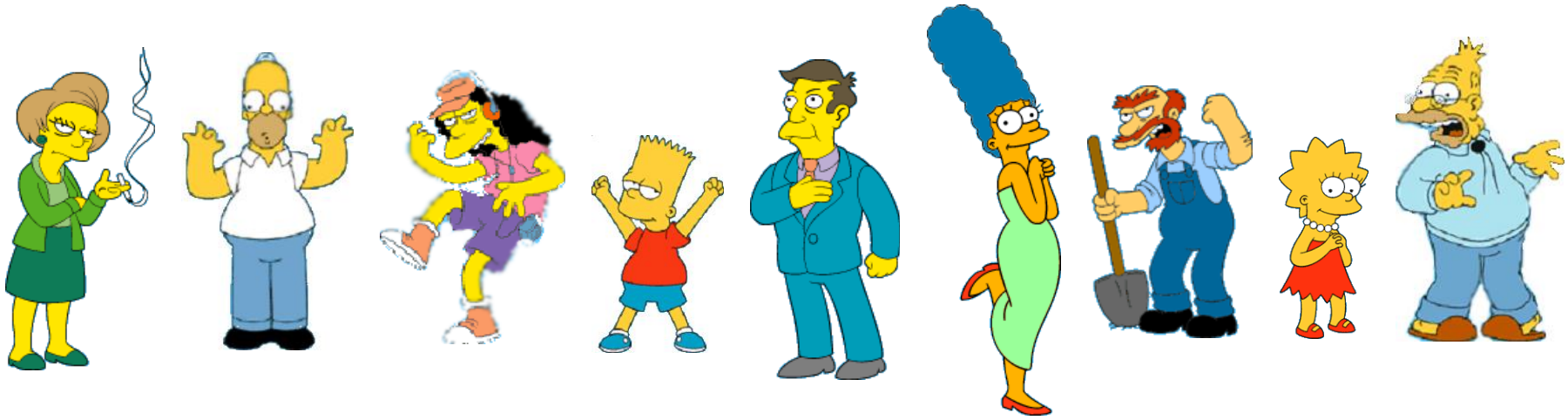
- Biologia:
  - Reino: Animalia
  - Ramo: Chordata
  - Classe: Mammalia
  - Ordem: Primatas
  - Família: *Hominidae*
  - Gênero: *Homo* (homem moderno e parentes)
  - Espécie: *Homo sapiens*



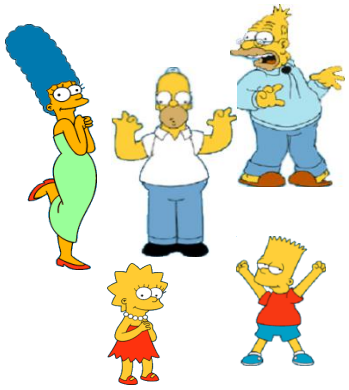
No entanto...

- Existem muitas situações nas quais não sabemos de antemão uma maneira apropriada de **agrupar** uma coleção de objetos de acordo com suas "similaridades"
  - massas de dados, possivelmente descritas por várias características (atributos) diferentes...
- Frequentemente não sabemos sequer se existe algum **agrupamento natural** dos objetos segundo um conjunto de características que descrevem esses objetos
  - que possa ser representativo de um ou mais fenômenos de interesse *por trás* dos dados em questão

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



Grupo 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 Grupo ou *Cluster* ?

- Definições subjetivas:
  - "Semelhanças entre objetos"...
    - Por ex., quais atributos devemos considerar (e como) para avaliar similaridades?

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

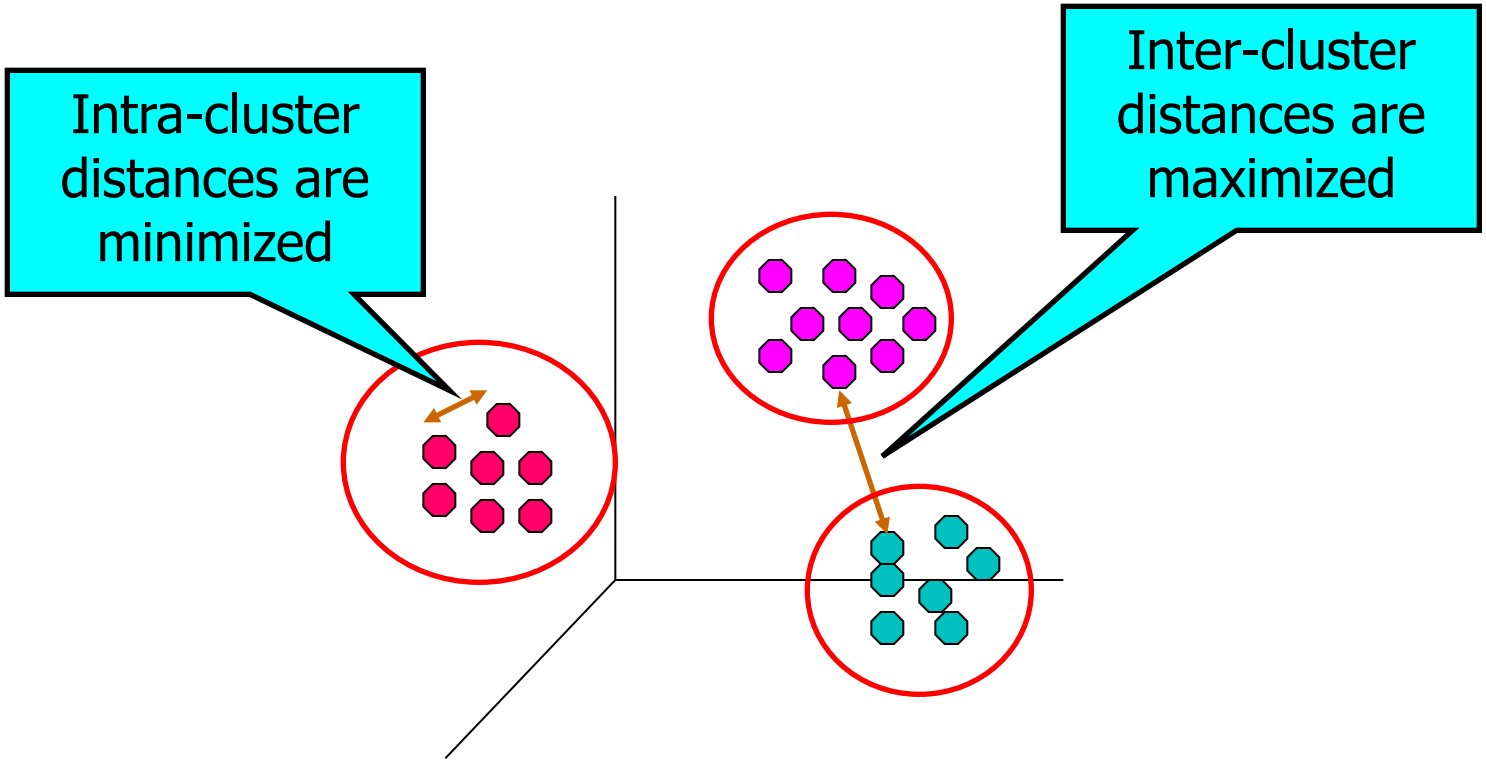
Definições Subjetivas...

- **Data Clustering:**

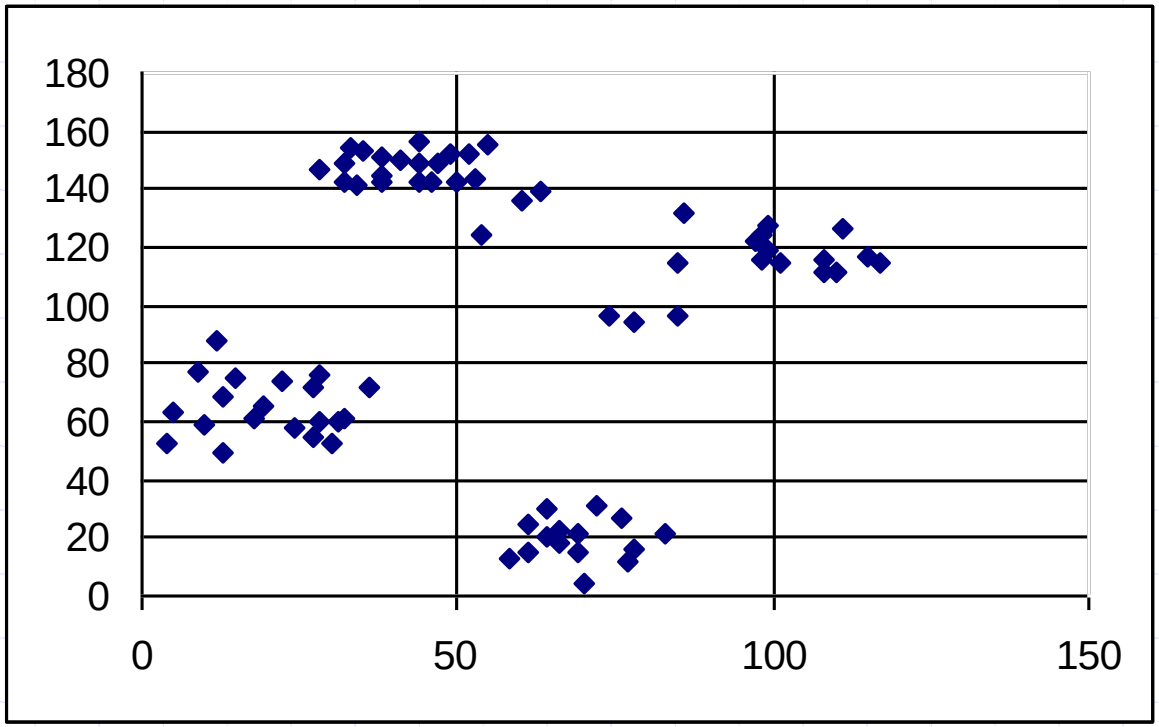
"Finding groups of objects such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups" (**Tan et al., 2006**)

"A statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics" (**Merriam-Webster Online Dictionary, 2008**)

(Uma) Visão Matemática



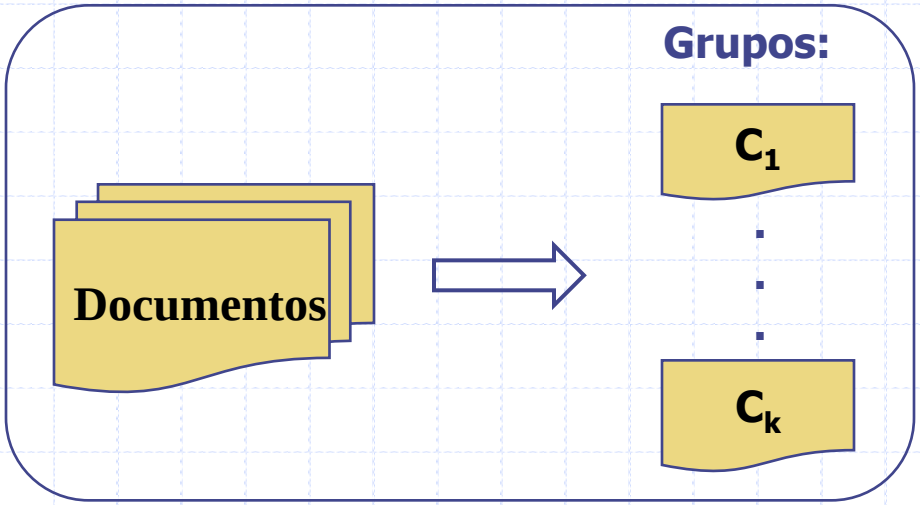
- Abordagem matemática em geral considera:
  - Homogeneidade (coesão interna)
  - Heterogeneidade (separação)
- Mesmo nesse caso, subjetividade ainda presente...

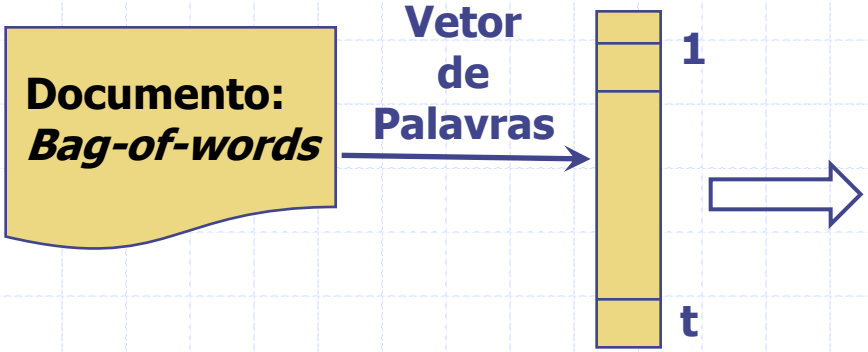


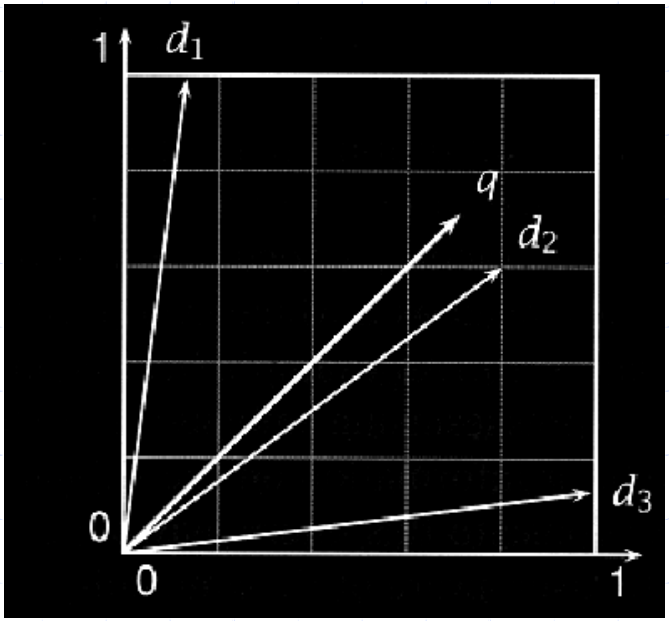
(Ruspini, 1970)

- Apesar de todas as dificuldades, a literatura sobre Análise de Agrupamento de Dados é rica e muito bem estabelecida;
  - Trabalhos importantes datam da década de 50.
  - P. ex., vide A. K. Jain, ***Data Clustering: 50 Years Beyond K-Means***, Pattern Recognition Letters, 2010
    - *"according to JSTOR [jst, 2009], data clustering first appeared in the title of a 1954 article dealing with anthropological data"*
- Há medidas de dis(similaridade) bem estudadas e fundamentadas para diversos tipos de dados e domínios de aplicação:
  - Dados Numéricos, Categóricos/Nominais, Binários, ...

- Por exemplo, em Mineração de Textos:







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

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

– **Nota**: *Data clustering is also known as Q-analysis, typology, clumping, and taxonomy, depending on the field where it is applied [Jain & Dubes, 1988]*

Exemplos de Aplicações

- **Marketing:** descobrir grupos de clientes / nichos de mercado e usá-los para marketing direcionado
- **Astronomia:** encontrar grupos de estrelas e galáxias
- **Bioinformática:** encontrar grupos de genes com expressões semelhantes
- **Mineração de Textos:** categorização de documentos
- **Vários Outros:** proc. imagens, controle, det. anomalias, ...

Kdnuggets Pool: *"Data mining/analytic methods you used frequently in the past 12 months" [203 voters]*

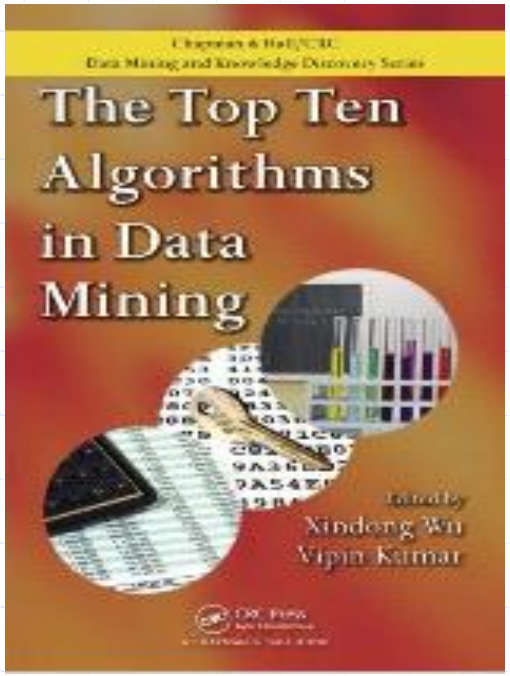
http://www.kdnuggets.com/polls/2007/data_mining_methods.htm

Decision Trees/Rules (127) 62.6%
Regression (104) 51.2%
Clustering (102) 50.2%
Statistics (descriptive) (94) 46.3%
Visualization (66) 32.5%
Association rules (53) 26.1%
Sequence/Time series analysis (35) 17.2%
Neural Nets (35) 17.2%
SVM (32) 15.8%
Bayesian (32) 15.8%
Boosting (30) 14.8%
Nearest Neighbor (26) 12.8%
Hybrid methods (24) 11.8%
Other (23) 11.3%
Genetic algorithms (23) 11.3%
Bagging (22) 10.8%

(Resultados mudaram pouco em 2011 --- http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html)

IEEE ICDM and ACM SIGKDD Poll

- 2 Algoritmos (k-means e EM) listados entre os **Top 10 Most Influential Algorithms in DM**:



- Wu, X. and Kumar, V. (Editors), *The Top Ten Algorithms in Data Mining*, CRC Press, 2009

- X. Wu et al., "Top 10 Algorithms in Data Mining", *Knowledge and Info. Systems*, vol. 14, pp. 1-37, 2008

- **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)

➢ Área de pesquisa importante e fortemente ativa.

Conceitos Básicos

Algumas Definições (Everitt, 1974):

- *Um cluster (grupo) é um conjunto de entidades semelhantes e entidades pertencentes a diferentes clusters não são semelhantes;*

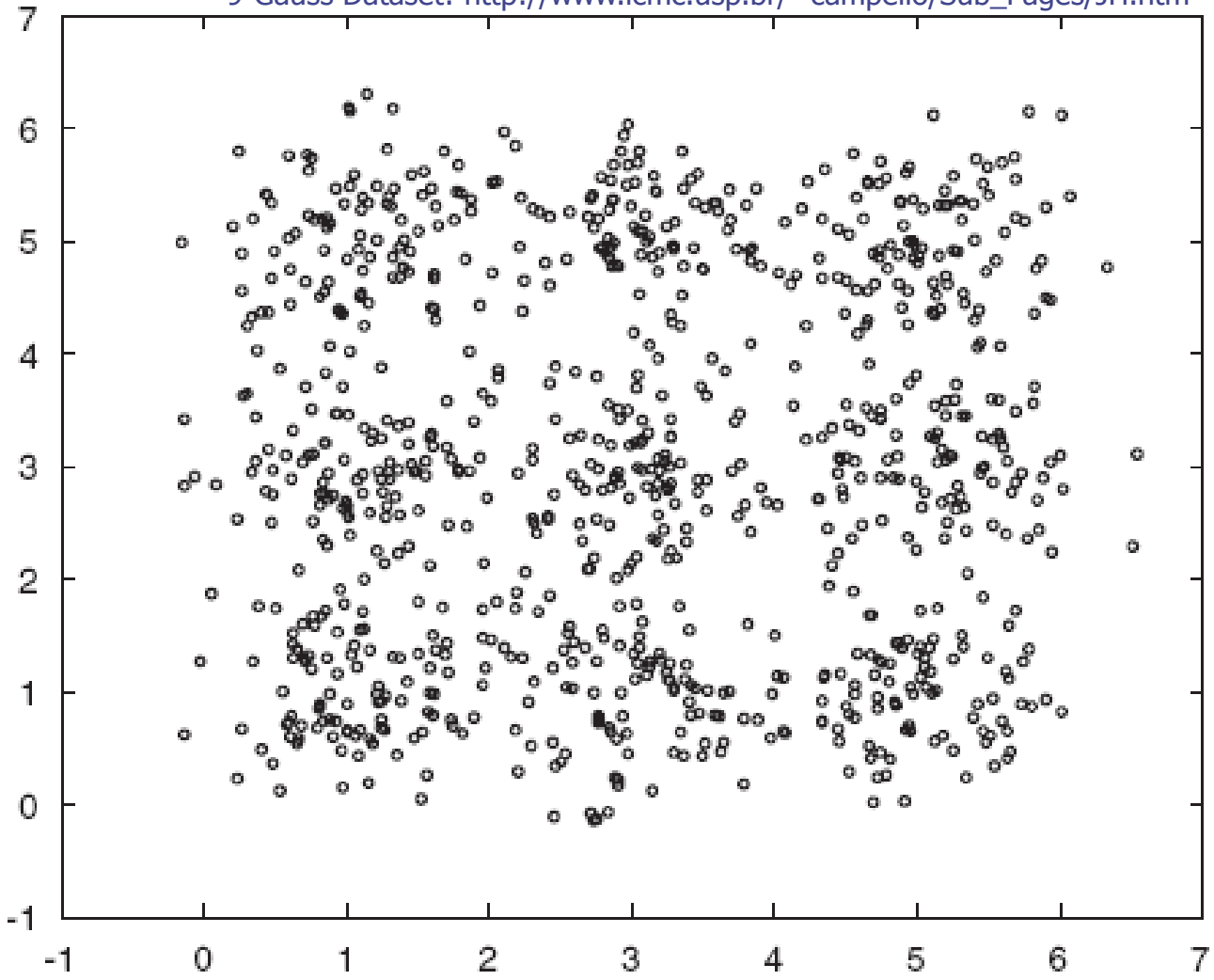
- *Um grupo é uma aglomeração de pontos no espaço tal que a distância entre quaisquer dois pontos no grupo é menor do que a distância entre qualquer ponto no grupo e qualquer ponto fora deste;*

- *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;*

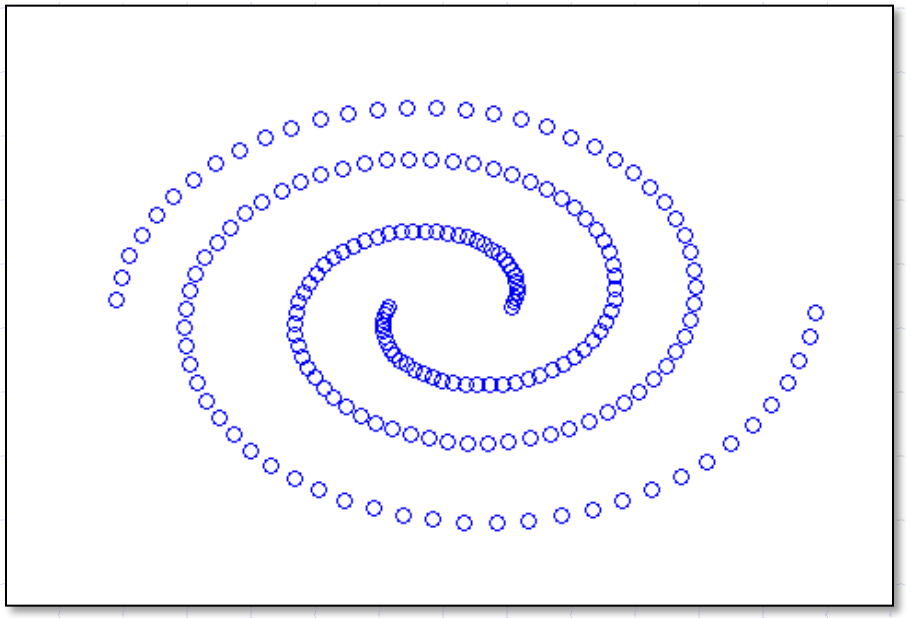
➢ **Humanos reconhecem *clusters* no plano quando os vêem, sem saber explicar exatamente porquê (Jain & Dubes, 1988) ...**

Quais são os grupos ?

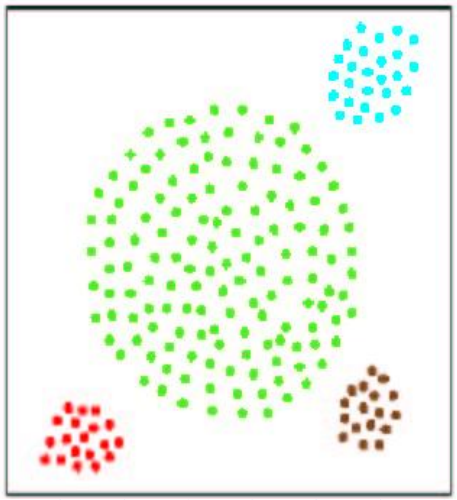
9-Gauss Dataset: http://www.icmc.usp.br/~campello/Sub_Pages/JH.htm

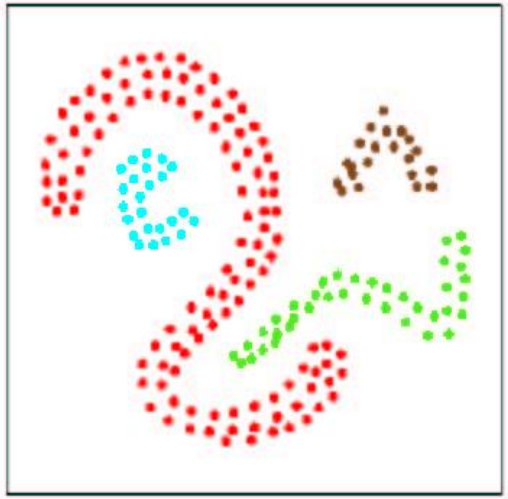


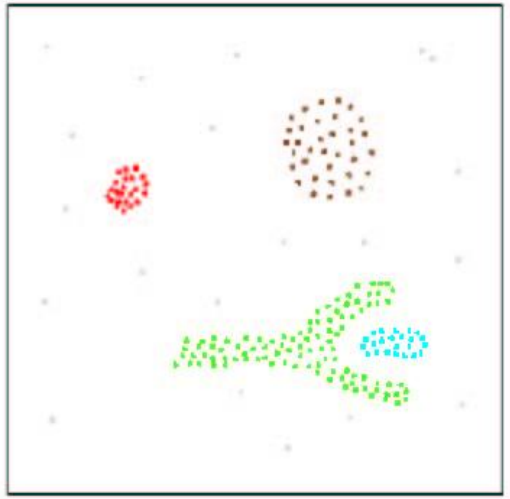
Quais são os grupos ? ...



Aplicações práticas ?







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

- Algoritmos de *clustering* induzem *clusters;*
- Os *clusters* a serem induzidos dependem de uma série de fatores, além dos dados propriamente ditos:
  - medidas de dis(similaridade), índices de avaliação, parâmetros def. pelo usuário, etc.
  - fortemente dependente do domínio / problema
  - relação c/ **bias indutivo** em aprendizado de máquina
- Perspectiva de **Aprendizado de Máquina**:
  - *projetista define o que o computador pode aprender*
  - *existem **centenas** de algoritmos...*

Abordagens de Clustering

- Muitos métodos / algoritmos diferentes:
  - Para dados numéricos e/ou simbólicos
  - Para dados **relacionais** ou **não relacionais**
  - Para obter **partições** ou **hierarquias** de partições
    - Partição: conjunto de clusters que compreendem os dados
  - Partições **mutuamente exclusivas** ou **sobrepostas**
  - **...**

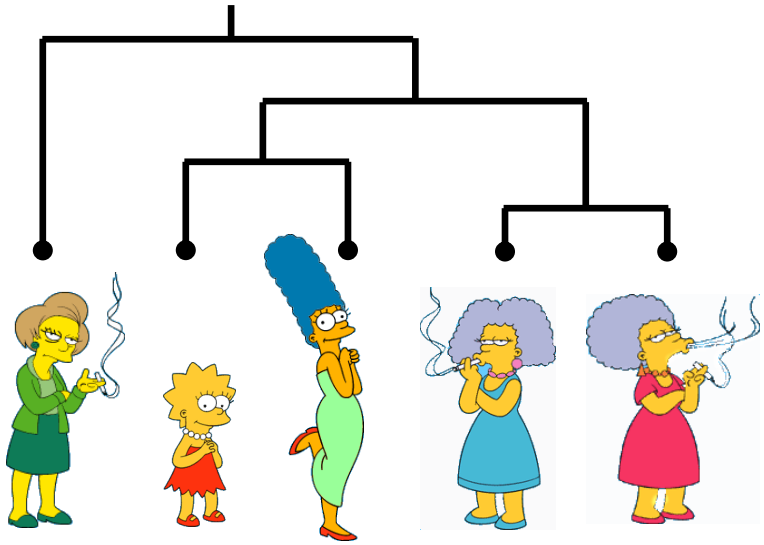
Métodos Relacionais *vs* Não Relacionais

- Métodos **Não Relacionais**:
  - Requerem os objetos a serem agrupados (dados originais)
- Métodos **Relacionais**:
  - Requerem apenas as (dis)similaridades entre os objetos
  - Vantagens:
    - Abordagem unificada para tratamento de atributos mistos
    - Dados sigilosos
    - Domínios de aplicação subjetivos (e.g. ciências sociais)
  - Desvantagem: Custo computacional em geral mais elevado

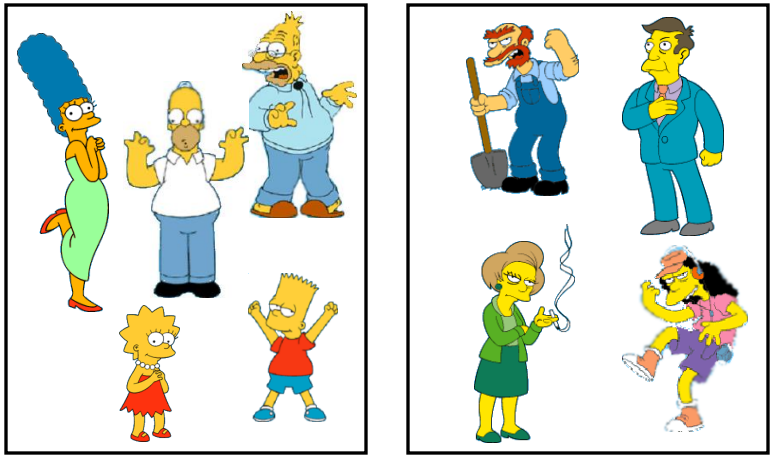
Métodos Particionais *vs* Hierárquicos

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

Hierárquicos

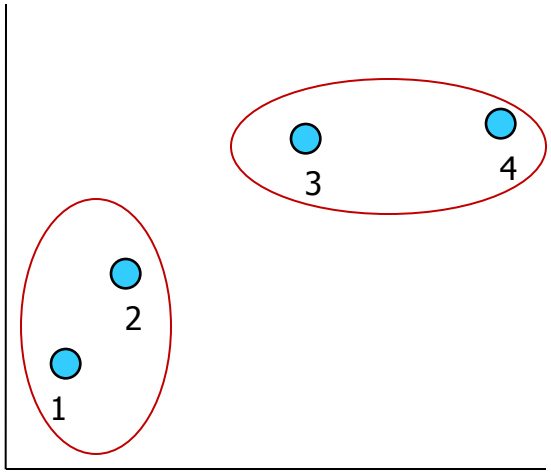


Particionais



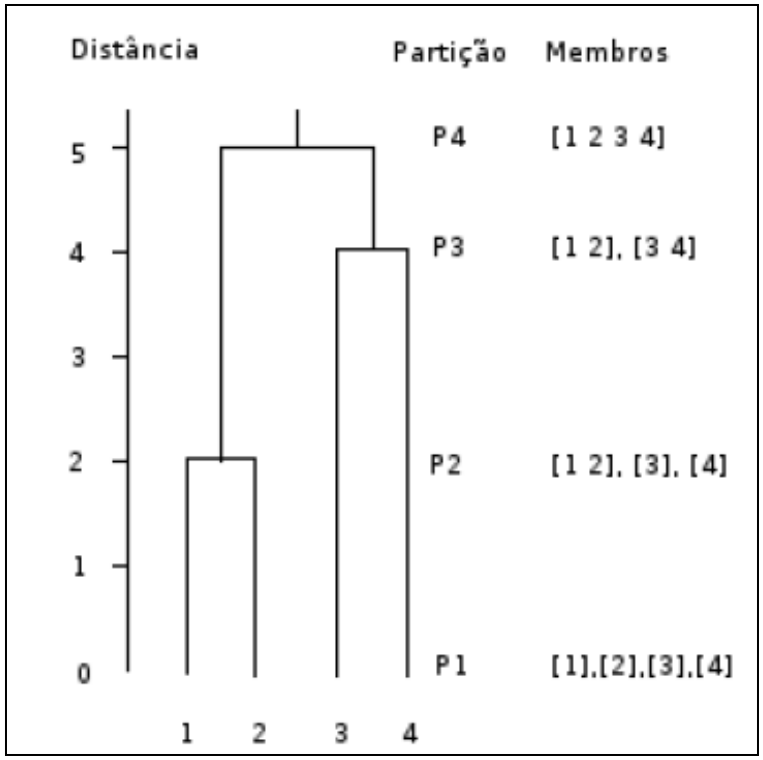
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Partições x Hierarquias



Partição

Figura por Lucas Vendramin

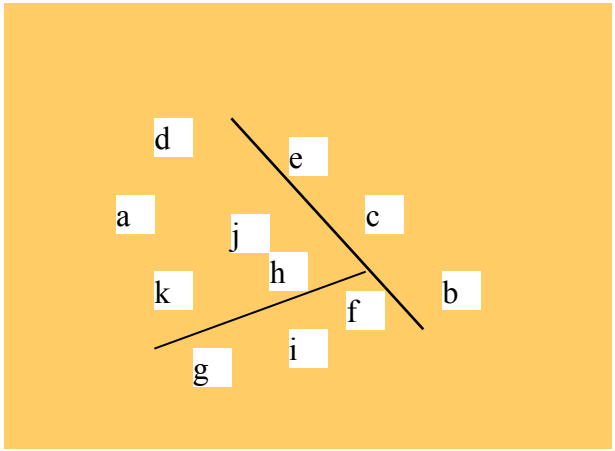


Hierarquia
(dendrograma)

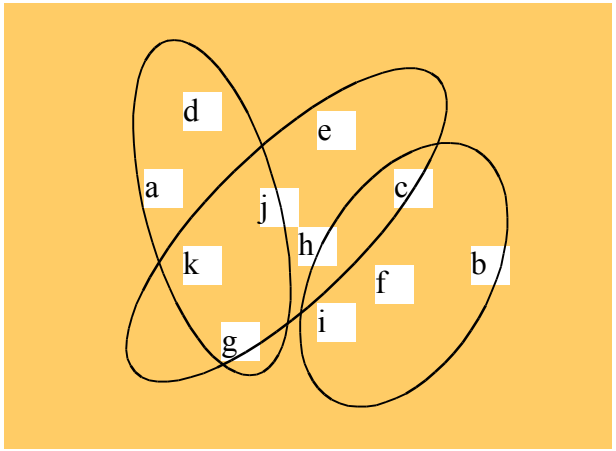
Cada um tem suas características próprias, assim como vantagens e desvantagens

Partições com e sem Sobreposição

Sem sobreposição



Com sobreposição



Cada um tem suas características próprias,
assim como vantagens e desvantagens

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

O que não é análise de agrupamento?

Classificação Supervisionada

- Disponibilidade de rótulos de classes

Segmentação Simples

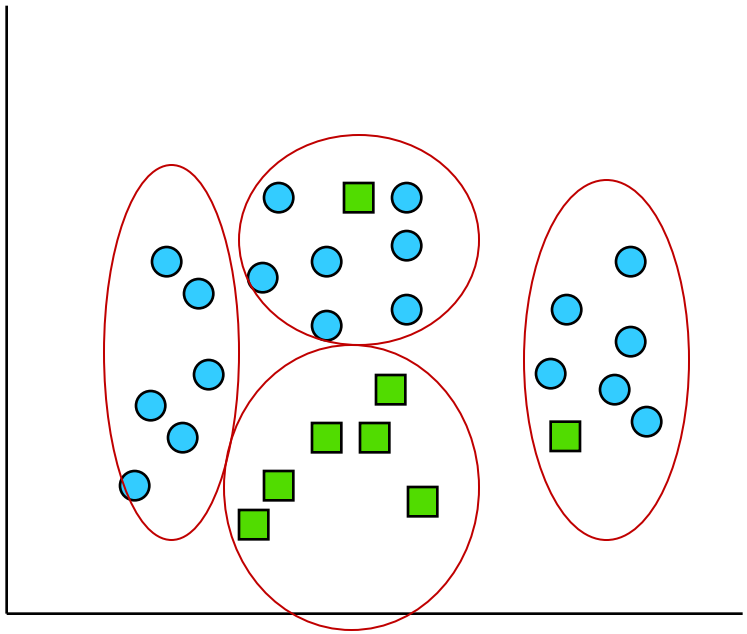
- Dividir estudantes em diferentes grupos alfabeticamente

Resultados de uma consulta (query)

- Grupos são resultado de uma especificação externa

…

Classificação X Clustering



Classificação:

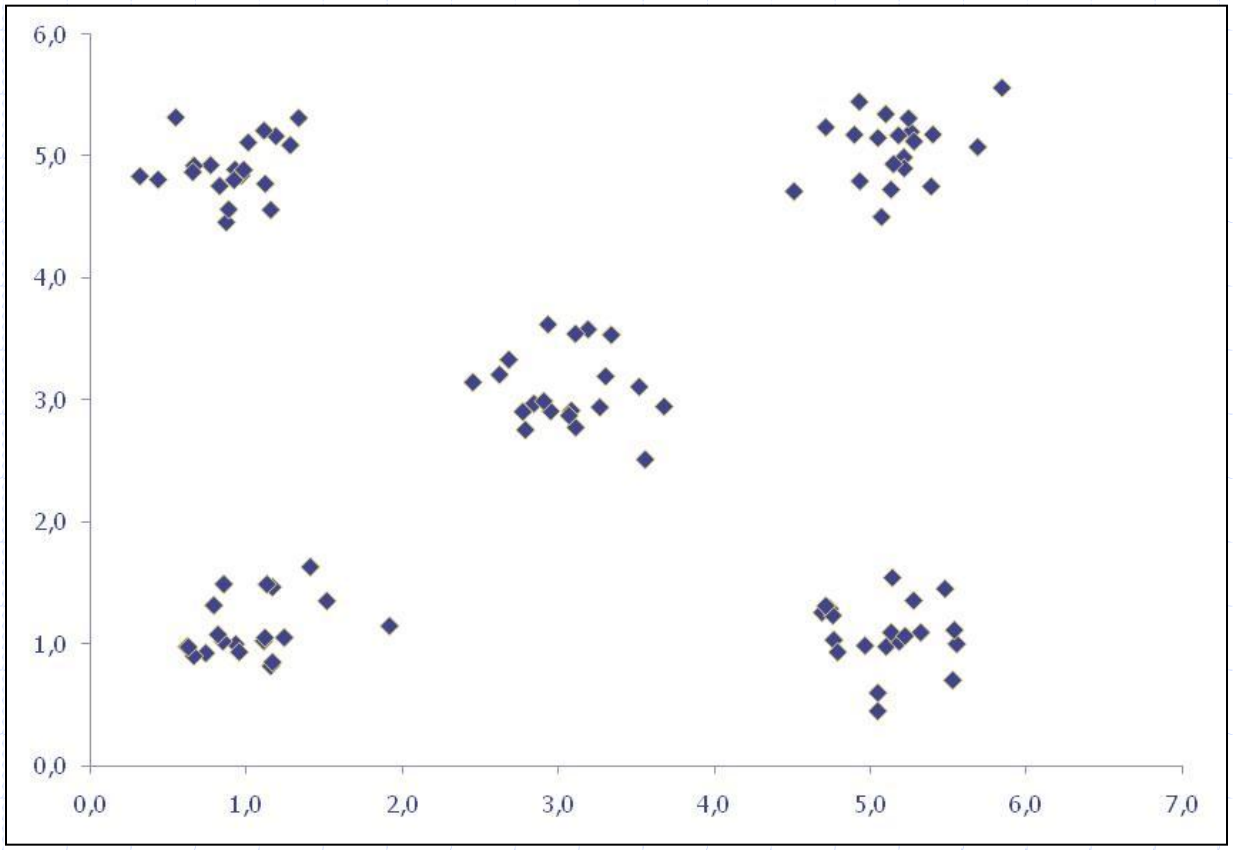
Aprender um método para predizer as categorias (classes) de padrões não vistos a partir de exemplos pré-rotulados (classificados)

Agrupamento de Dados (Clustering):

Encontrar os rótulos das categorias (grupos ou **clusters**) e possivelmente o número de categorias diretamente a partir dos dados

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

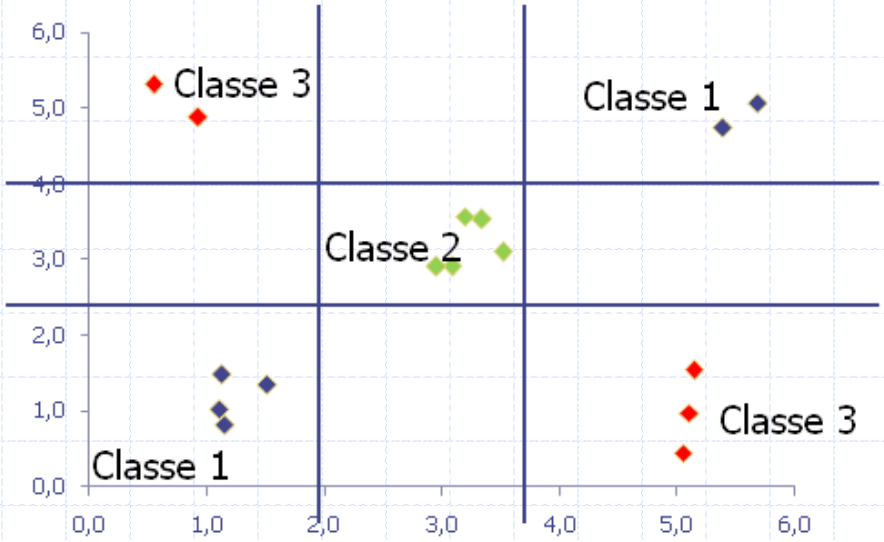
Agrupamento X Classificação?

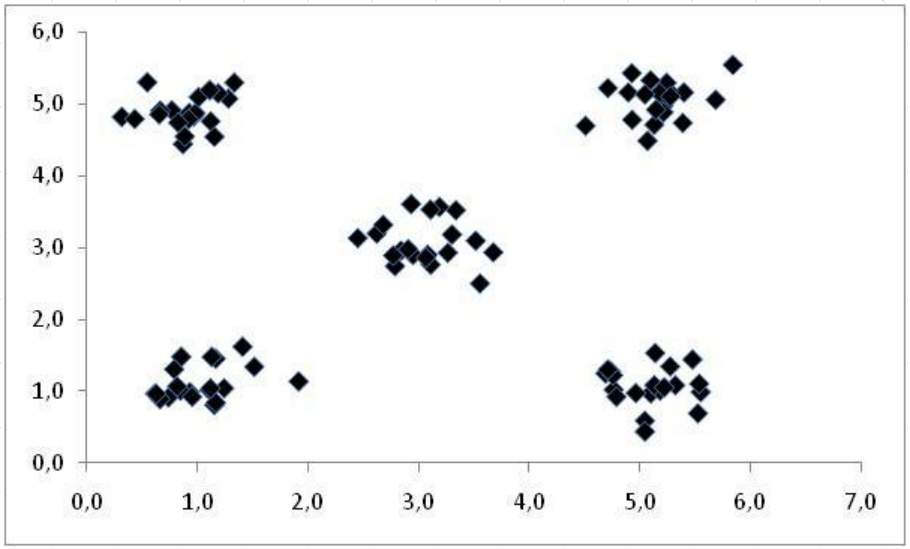


Agrupamento:
Indução de grupos a partir da base de dados...

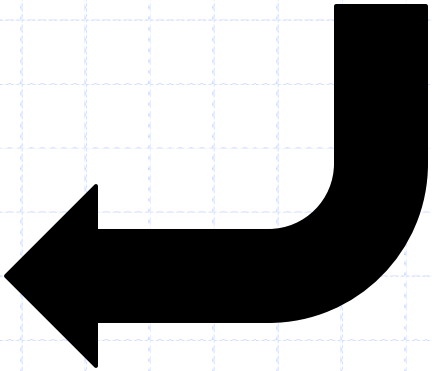
- Grupos obtidos serão então cuidadosamente estudados

Agrupamento X Classificação? ...



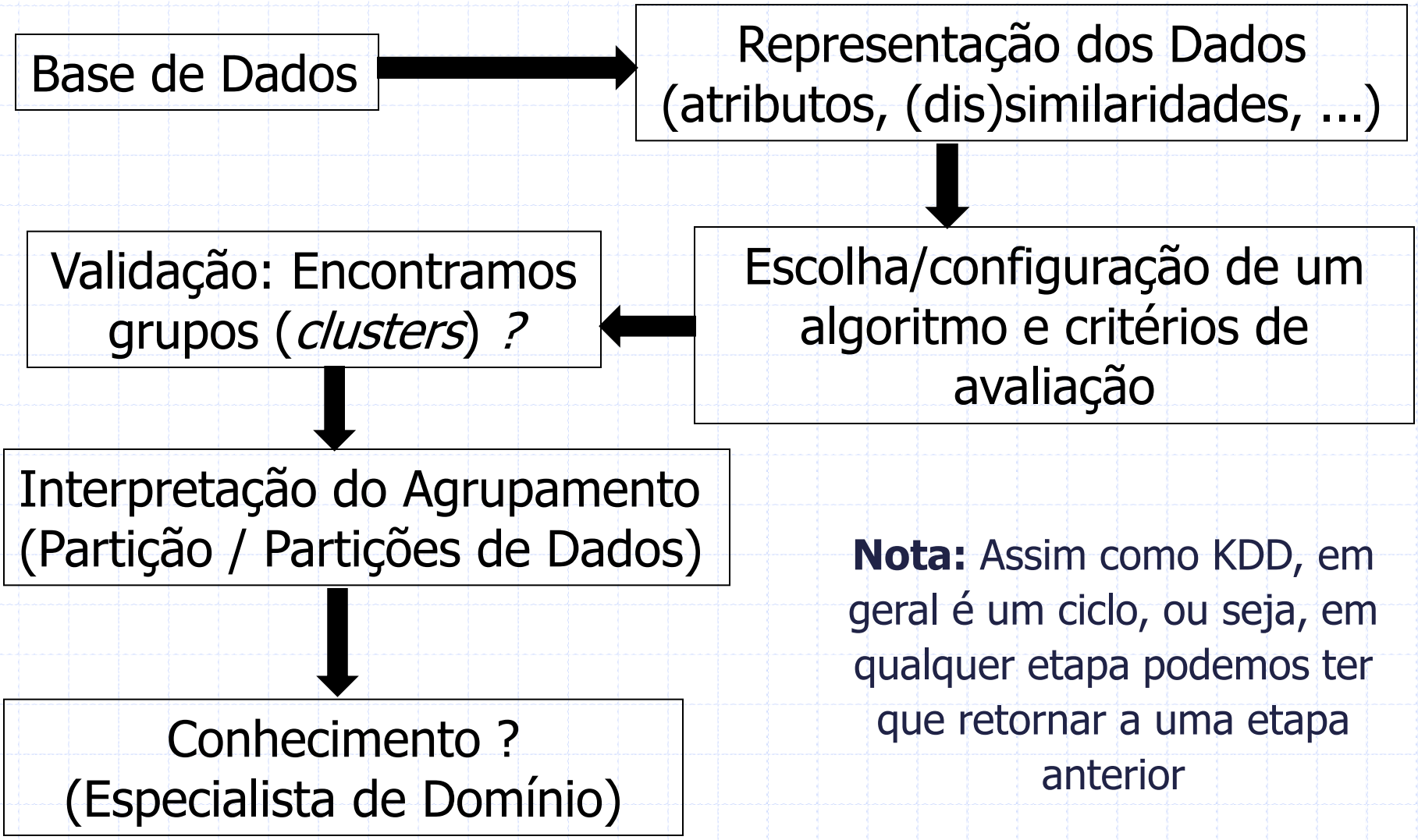


Base de treinamento com dados rotulados: classificador (modelo)

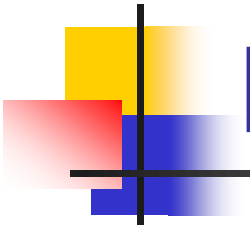


Rotular dados de teste em função do modelo obtido

Processo Básico – "Análise de Agrupamento"



Nota: Assim como KDD, em geral é um ciclo, ou seja, em qualquer etapa podemos ter que retornar a uma etapa anterior



Referências

- Jain, A. K. and Dubes, R. C., *Algorithms for Clustering Data*, Prentice Hall, 1988
- Everitt, B. S., Landau, S., and Leese, M., *Cluster Analysis*, Arnold, 4th Edition, 2001
- Gan, G., Ma, C., and Wu, J., *Data Clustering: Theory, Algorithms and Applications*, ASA SIAM, 2007
- Xu, R., Wunsch, D., *Clustering*, IEEE Press, 2009
- Tan, P.-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006
- Wu, X. and Kumar, V. (Editors), *The Top Ten Algorithms in Data Mining*, CRC Press, 2009