

Mineração de dados estruturados e não-estruturados como vantagem competitiva

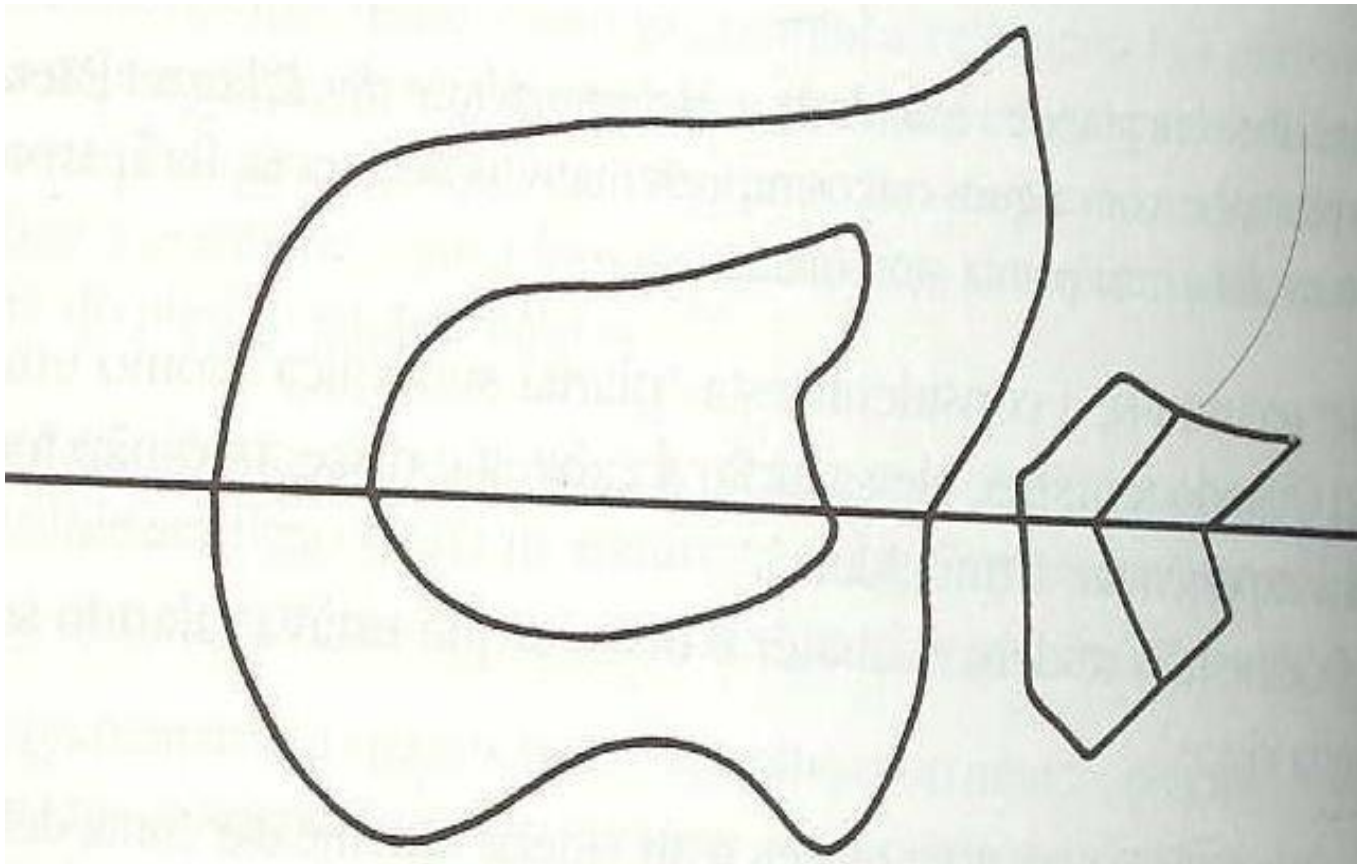
Solange O. Rezende
Departamento de Ciências de Computação
ICMC-USP, São Carlos
solange@icmc.usp.br



www.labic.icmc.usp.br

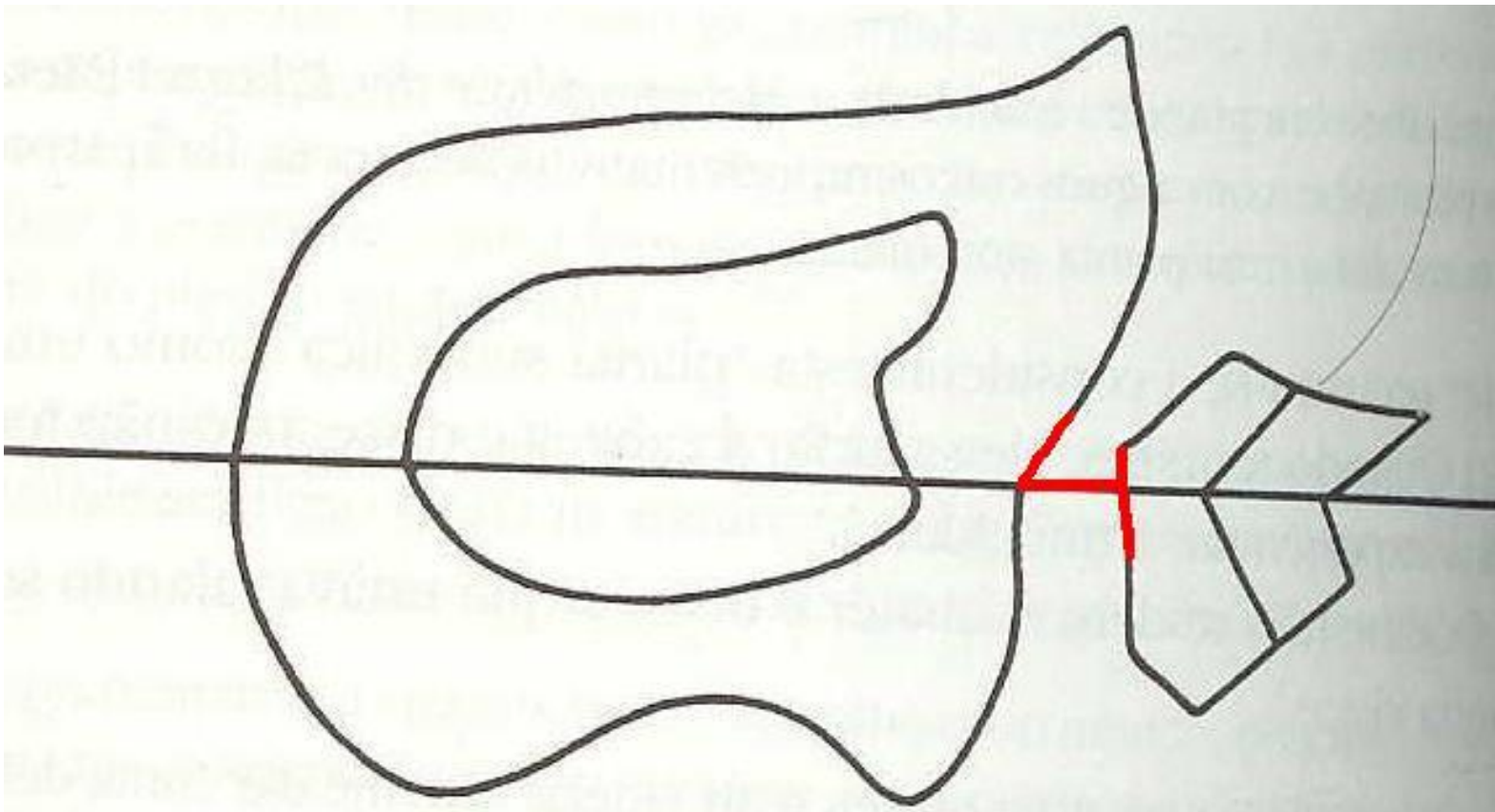
Motivação

Observe a imagem e identifique um número...



Motivação

Observe a imagem e identifique um número...



Motivação

NIKE



WAL MART



Motivação



Cassino Harrah's
(Guizzo, 2001)

16 milhões de clientes!

**Qual o perfil de cliente
proporciona maior
lucratividade?**

- Apostadores que gastam entre US\$ 100 a 500:
 - ✓ Representam 30% da clientela
 - ✓ Contribuem com 80% das receitas

- Estratégias de marketing para este “filão” mais rentável dobrou o **faturamento**

Motivação

Empresa varejista norte-americana **utiliza os dados das compras** dos clientes para criar campanhas de marketing pró ativas.



Eles conseguiram detectar um **padrão de compra de mulheres** que estavam **grávidas** e enviar a elas ofertas para gestantes.



Detalhes:

- Muitas vezes a Target sabia da gestação antes mesmo de alguns familiares. **Incluindo alguns maridos!**
- Alguns processos seguidos de revisão de estratégias



Sadia 
@SadiaOficial

[Página inicial](#)

Sobre

Fotos

Videos

[Termos de Uso](#)

Publicações

Eventos

Receitas

Curtidas

Notas

[Criar uma Página](#)

 Curtir Seguir Compartilhar ...

[Ver tudo](#)

Publicações



Sadia

31 de dezembro de 2016 -

2016 foi um ano incrível pra gente! Estamos orgulhosos de ter alcançado o primeiro lugar em sódio em 80% de nossa linha de produtos com **Varela**. Fizemos parceria com um dos maiores produtores de café do mundo, a **Oliver**. Com isso estamos trilhando um caminho de crescimento e prosperidade. Que 2017 venha com mais conquistas para todos nós! Feliz Ano Novo e venham com a **Sadia**!




 Curtir
 Comentar
 Compartilhar

👍❤️😱 166

19 compartilhamentos

Escreva um comentário...

 **Lucia Mello** Salsicha hot dog está com sabor muito diferente....conseguiu comer até crua, agora não consigo nem ver na minha frente



Dandara Oliveira Cês tão sabendo que a linguaça tá vindo 50% gordura (ou mais, depende da sorte do consumidor)? Mas que nojo, heim? Aqui em casa compramos delas há anos e ultimamente a linga tá ficando **SOMENTE GORDURA**. Isso pra vocês é Sadia? Ou mais?

Curtir · Responder · 2 - 27 de janeiro às 13:57



S Sádía Oi, Dandara! Sentimos muito que isso te agradeçemos por compartilhar com a gente sua opinião. Esse retorno dos consumidores é muito importante para nós e que possamos aprimorar não só este, mas todos os produtos e manter o controle de qualidade.



Sadia @sadia · 2 h

@Fraan Guedes A melhor pizza é nossa e o melhor elogio é o seu, Fraan! 🍕

↩ ↻ 1 ♡ 3

← Em resposta a pati

Equipe altamente
conectada com os
valores Jamie Oliver e
com os clientes

RETIROU campanha do
ar!



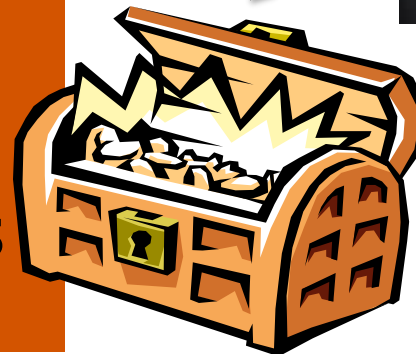
saturno @SADBUTSATURNO · 6 de fev

se a pizza de lombo da **sadia** fosse uma pessoa eu casava com ela sem pensar duas vezes

- Os **sistemas** computacionais armazenam quantidades cada vez maiores de dados.
- Esse volume de **dados** é uma valiosa **fonte de conhecimento**.
- A quantidade e complexidade dos dados **impossibilitam a exploração manual** desse conhecimento.



Necessidade de técnicas automáticas para extrair padrões dos dados armazenados.



Contexto

Consultas - Banco de Dados e Mineração de Dados

Banco de Dados

- Encontre todos os clientes que vivem em Boa Vista
- Encontre todos os clientes que usam Mastercard
- Encontre todos os clientes que não pagaram uma parcela

Mineração de Dados

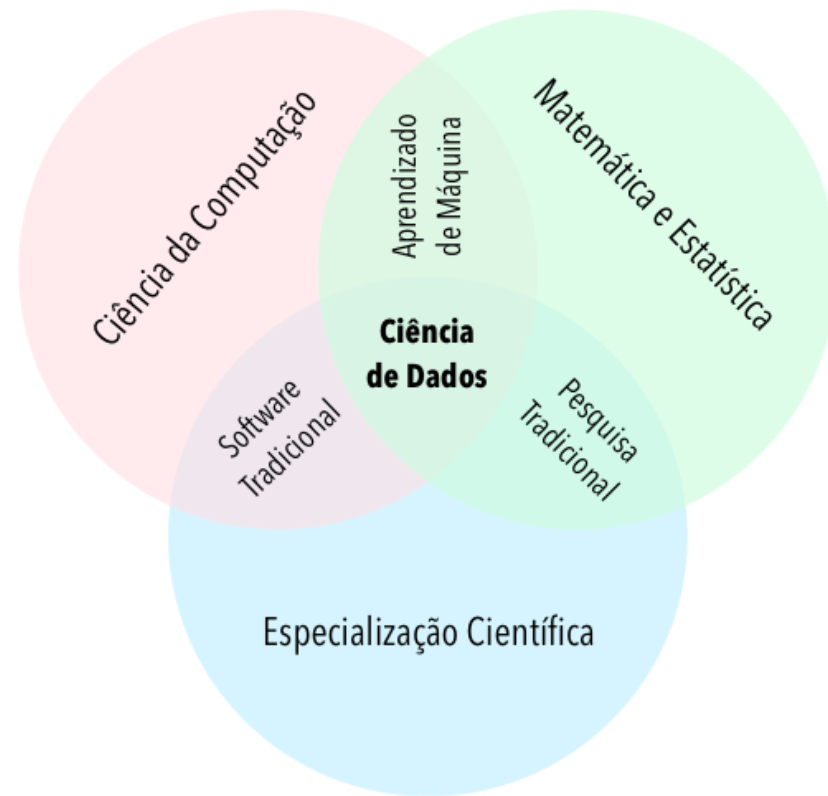
- Encontre todos os clientes que provavelmente podem não pagar uma parcela (**Classificação**)
- Agrupe todos os clientes com hábitos de consumo simples (**Agrupamento**)
- Liste todos os item que são frequentemente comprados com bicicletas (**Regras de associação**)
- Encontre qualquer cliente "incomum" (**Detecção de outliers, descoberta de anomalias**)

De dados à manipulação de conhecimento...



Aprender a partir do que temos!

Ciência de Dados é
uma área
interdisciplinar voltada
para o estudo e a
análise de dados,
estruturados ou não,
que visa a **extração**
de conhecimento ou
insights para
possíveis tomadas
de decisão



Aprender a partir do que temos!

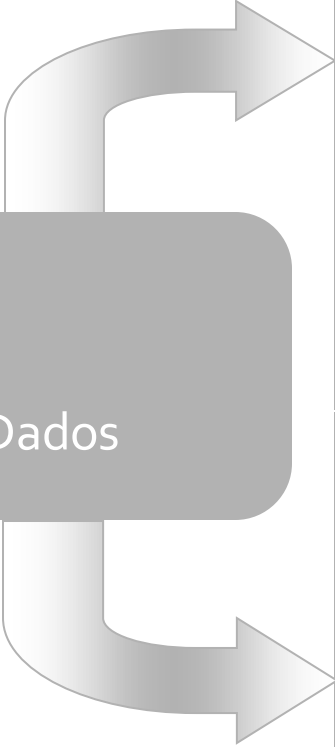
Temos muitos Dados

Mineração de Dados refere-se ao processo de **extrair conhecimento** de dados.

Auxilia o ser humano a extrair padrões **válidos, novos e potencialmente úteis** dos dados pela aplicação de diversas ferramentas e técnicas computacionais automáticas (Fayyad et al. 1996; Rezende et al., 2003).

Abordagens de MD

Metodologias e
Abordagens de
Mineração de Dados



TOP DOWN

INICIAR COM HIPÓTESES E VALIDAR AS MESMAS

AS HIPÓTESES PODEM SER CONSTITUÍDAS INICIALMENTE A PARTIR DA ABORDAGEM *BOTTOM UP* OU A PARTIR DE ALGUM CONHECIMENTO DO “MUNDO REAL”

SE A HIPÓTESE NÃO FOR SATISFEITA, REVISÁ-LA

BOTTOM UP

ANALISAR OS DADOS E EXTRAIR PADRÕES

SUPERVISIONADO:

TEM-SE ALGUMA IDÉIA DO QUE ESTÁ PROCURANDO

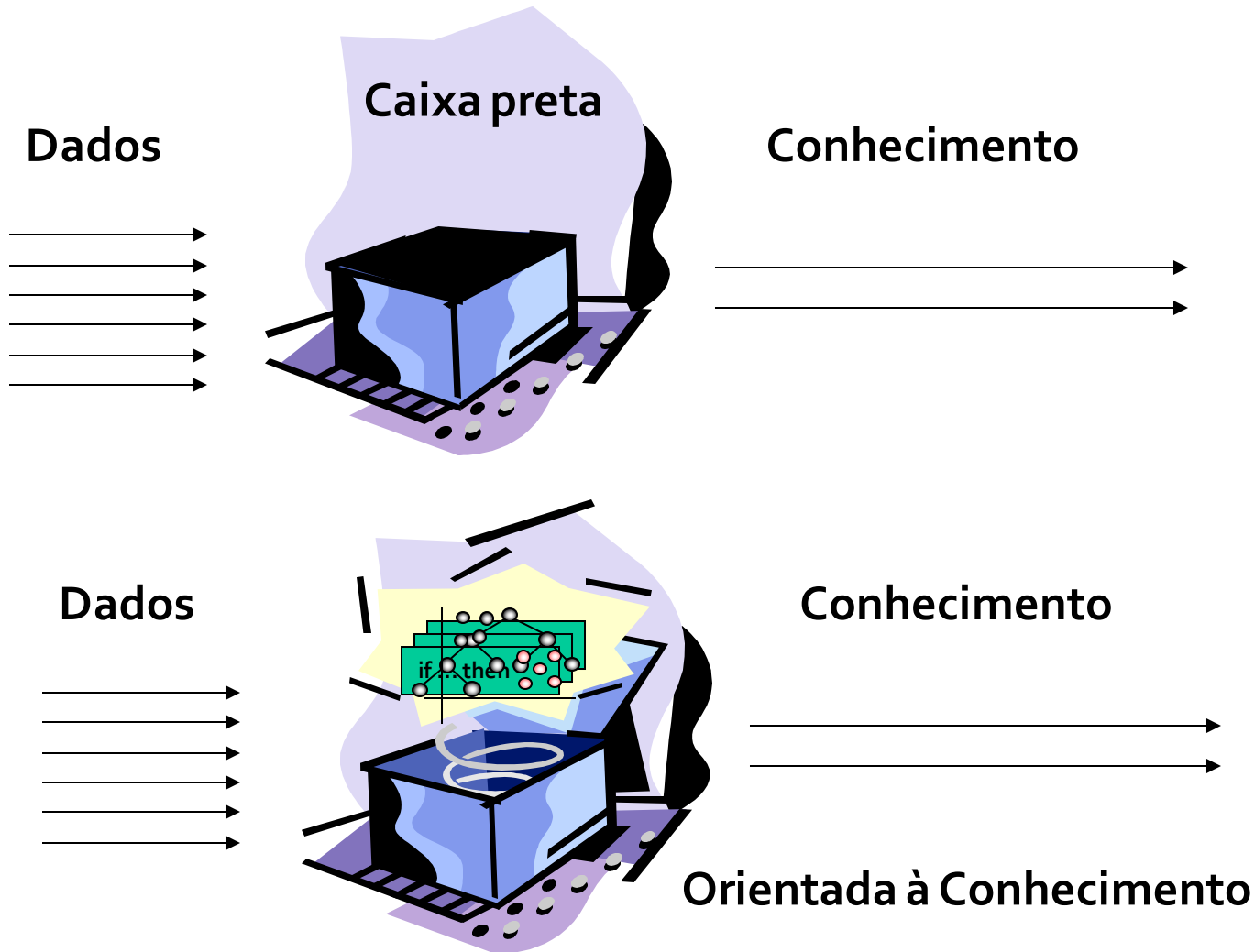
NÃO SUPERVISIONADO:

NÃO-SE TEM IDÉIA DO QUE ESTÁ PROCURANDO

(Rezende et al., 2003)

Foco no Conhecimento

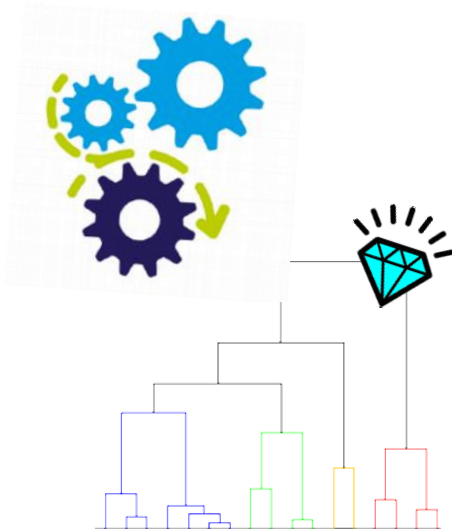
Diversos paradigmas de aprendizado



Mineração de Dados

- Termos similares
 - Exploratory data analysis
 - Data driven discovery
 - Deductive learning
 - Discovery Science
 - Knowledge Discovery

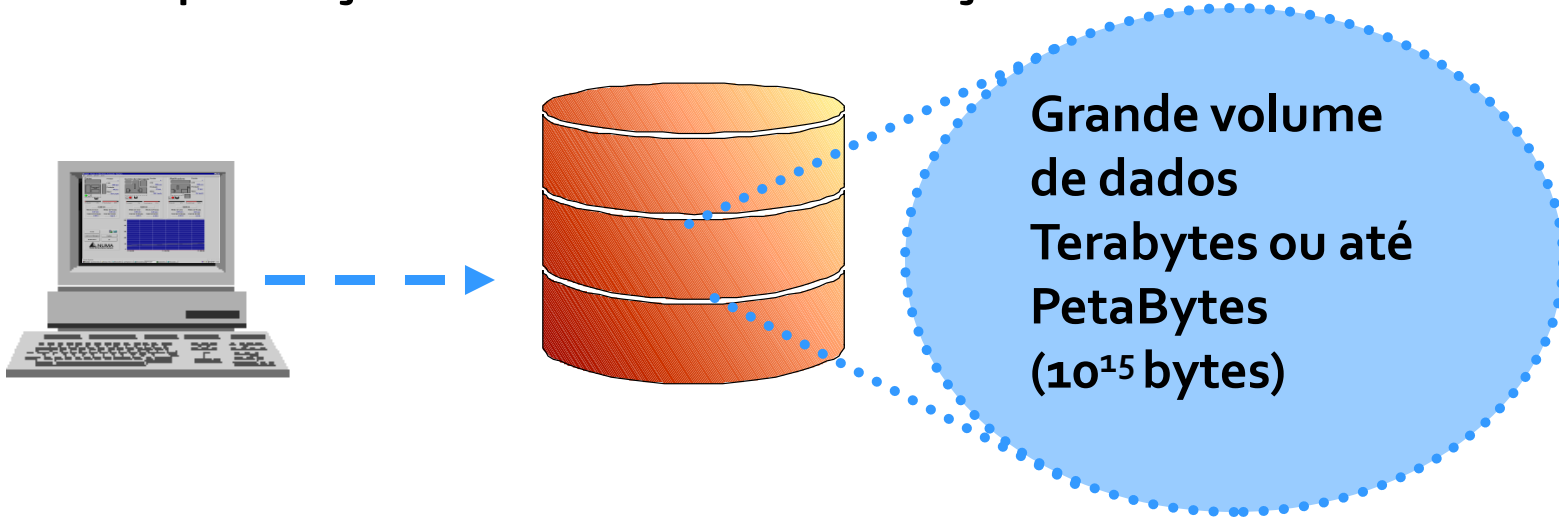
Mineração de Dados na prática



Processo de Mineração de Dados

Identificação do problema

- A exploração dos dados começa com os dados?



- A exploração normalmente começa com a identificação de uma necessidade!
- Envolve o estudo do domínio da aplicação. As decisões tomadas neste ponto guiarão os passos consecutivos e poderão ter reflexo no desempenho da aplicação!
- Auxílio do especialista de domínio é muito importante!

Processo de Mineração de Dados

Identificação do problema

- Estudo do domínio da aplicação
- Definição e identificação dos objetivos
 - Quais as principais metas do processo?
 - Quais critérios de desempenho são importantes?
 - O conhecimento extraído deve ser compreensível a seres humanos ou o modelo do tipo caixa preta é apropriado?
 - Qual deve ser a relação entre simplicidade e precisão do conhecimento extraído?
- As decisões tomadas neste ponto guiarão os passos consecutivos e poderão ter reflexo no desempenho da aplicação!

Processo de Mineração de Dados

Identificação do problema

- Auxílio do especialista de domínio pode ser necessária!
 - Estudo acerca do domínio
 - Aquisição de conhecimento inicial sobre o problema
- Deve-se, também, selecionar os dados com os quais irá trabalhar
 - Devem ser representativos ao domínio e aplicação de conhecimento
 - Podem ser oriundos de diversas fontes

Processo de Mineração de Dados

Pré-Processamento



Grande foco de
diferença entre
Mineração de Dados
estruturados e não
estruturados

Pré-processamentos

- ❑ Representação estruturada: permite ser manipulada por algoritmos de aprendizado de máquina
- ❑ Concisa: remover informação redundante
- ❑ Orientada ao domínio: considerar restrições e características do problema

Pré-processamento

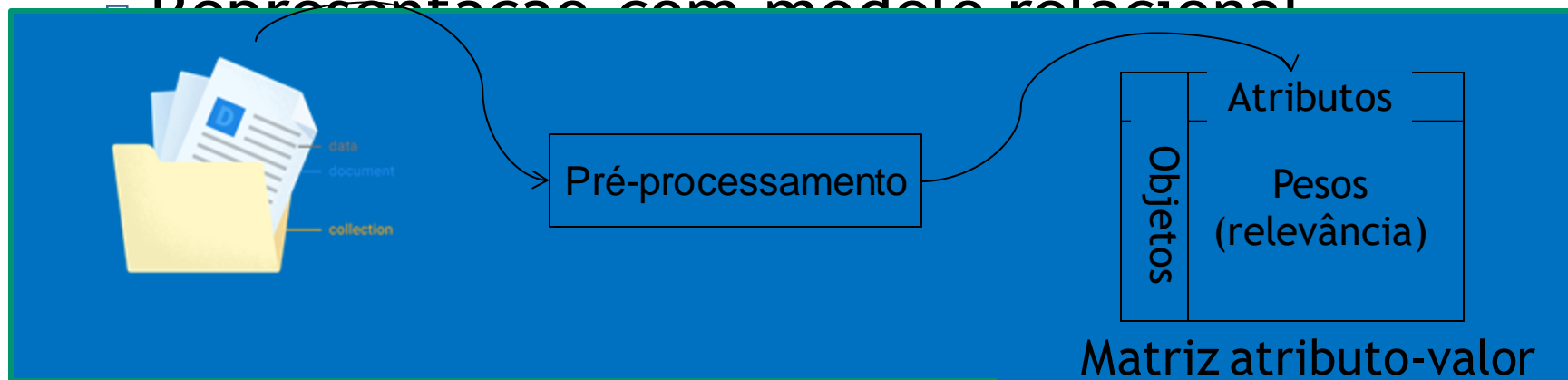
? Pré-processamentos dos textos

? Representação estruturada: permite ser manipulada por algoritmos de aprendizado de máquina

? Duas estratégias:

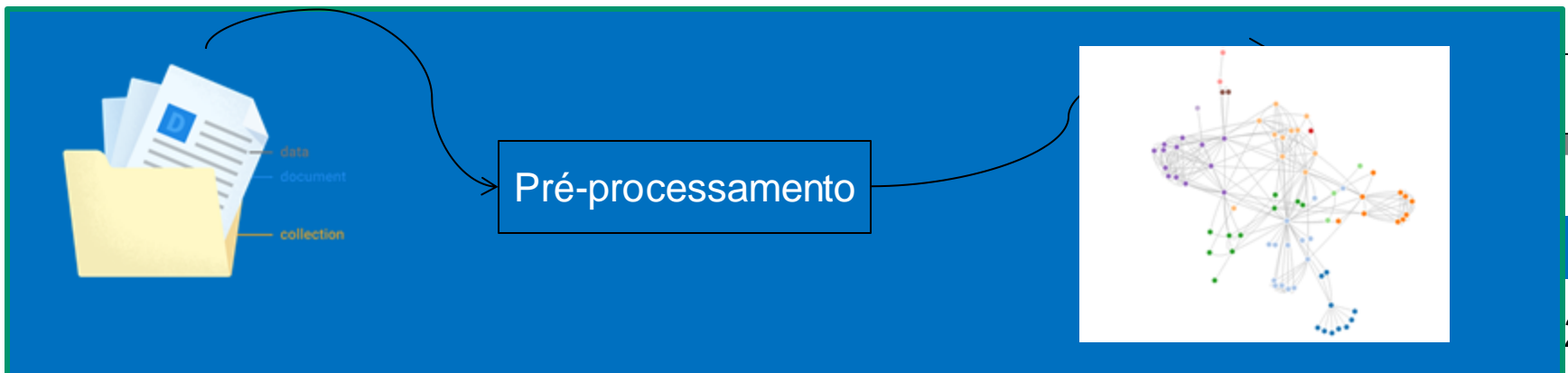
? Representação com modelo espaço-vetorial

= Representação com modelo relacional



Pré-processamento

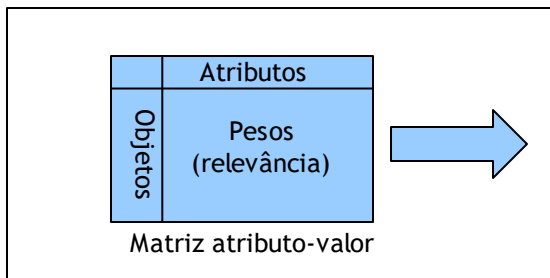
- ❑ Pré-processamentos
 - ❑ Representação estruturada: permite ser manipulada por algoritmos de aprendizado de máquina
 - ❑ Duas estratégias:
 - ❑ Representação com modelo espaço-vetorial



Pré-processamento

❓ Representação estruturada: permite ser manipulada por algoritmos de aprendizado de máquina

- ❓ Pode ser construída com técnicas estatísticas simples
- ❓ Permite o uso de diferentes algoritmos de aprendizado de máquina
- ❓ Representação que pode limitar a interpretabilidade por humanos
- ❓ Correlação entre atributos é implícita e dependente do algoritmo de aprendizado



Exemplo de modelo espaço-vetorial (*bag-of-words*)

Text	This	Is	A	Nice	Hotel	Not	All	at
This is a nice hotel	1	1	1	1	1	0	0	0
Not a nice hotel! not at all	0	0	1	1	1	2	1	1

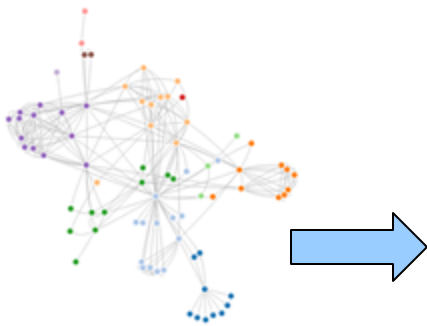
Pré-processamento

❑ Representação estruturada: permite ser manipulada por algoritmos de aprendizado de máquina

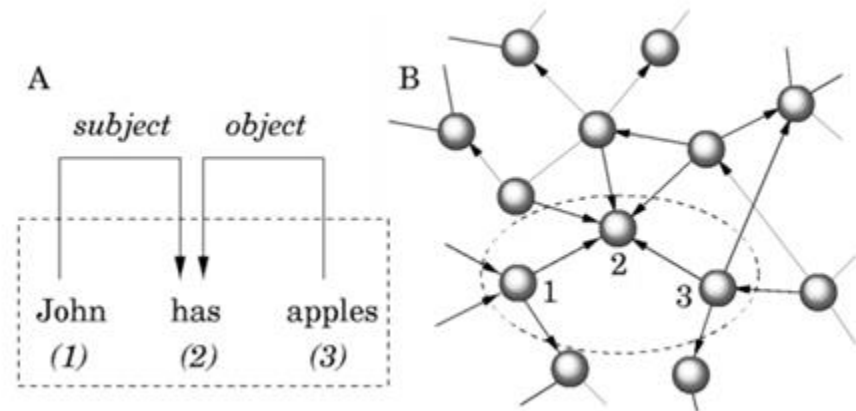
- ❑ Geralmente depende Processamento de Linguagem Natural
- ❑ Permite a construção de representações mais interpretáveis
- ❑ Nós e arestas podem representar conceitos semânticos e características do domínio
- ❑ Menor quantidade de algoritmos de aprendizado que consideram as relações semânticas

	Atributos
Objetos	Pesos (relevância)

Matriz atributo-valor



Exemplo de modelo relacional



Processo de Mineração de Dados

Pré-Processamento

Extração e integração

- Os dados podem vir de diferentes fontes
- Além disso, podem estar em diferentes formatos, como arquivos texto, arquivos Excel, banco de dados relacionais, Data Warehouse...
- É necessária a **unificação**, formando uma única fonte de dados

	X_1	X_2	...	X_m	Y
E_1	x_{11}	x_{12}	...	x_{1m}	y_1
E_2	x_{21}	x_{22}	...	x_{2m}	y_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
E_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Processo de Mineração de Dados

Pré-Processamento

- É uma das etapas que mais demanda tempo em um processo de Mineração de Dados
 - É, talvez, a que tenha menos “glamour” técnico
 - Envolve muitas atividades manuais
- Transformação dos dados para deixá-los adequados à etapa de Extração de Padrões
 - Extração e Integração
 - Transformação
 - Limpeza
 - Redução dos Dados



Processo de Mineração de Dados

Pré-Processamento

Transformação

- Adequação aos algoritmos de Extração de Padrões
 - Resumo
 - Transformação de tipo
 - Normalização de atributos contínuos
- Podem ser muito importantes em alguns domínios, como em aplicações que envolvem séries temporais como previsões no mercado financeiro

Processo de Mineração de Dados

Pré-Processamento

Limpeza

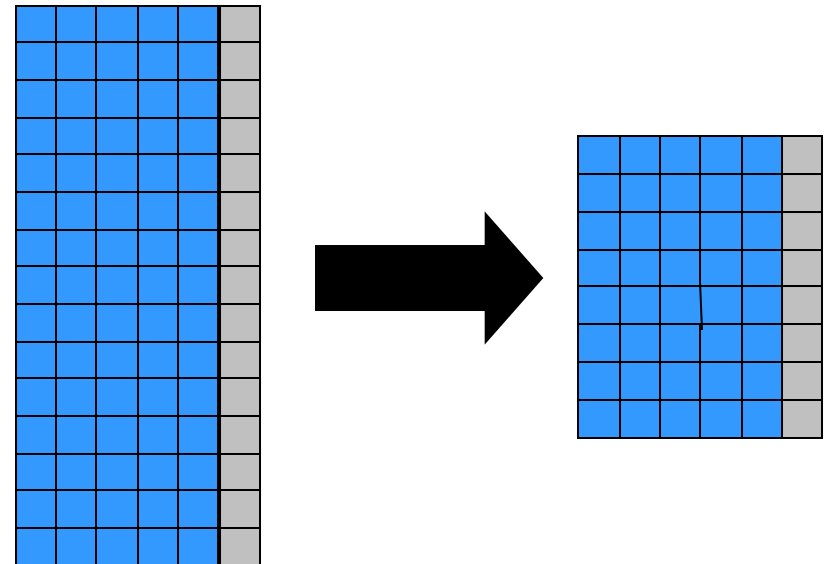
- Dados podem apresentar problemas provenientes da coleta (digitação ou leitura por sensores)
- Qualidade é muito importante
 - Utilizar o conhecimento do domínio
 - Decisão da estratégia de tratamento de atributos incompletos, remover ruídos
- Deve lidar com a completude dos dados
 - Muitas vezes, lida-se com ausência de dados
 - Exemplo: campos não preenchidos em um formulário

Processo de Mineração de Dados

Pré-Processamento

Redução de Dados

- Limitações de espaço em memória, tempo de processamento
- Atributos redundantes nos dados; atributos desnecessários
- A redução pode ser realizada de três formas:
 - **Número de exemplos**

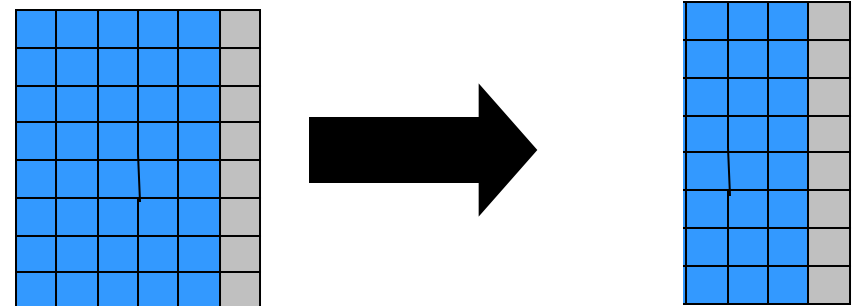


Processo de Mineração de Dados

Pré-Processamento

Redução de Dados

- Limitações de espaço em memória, tempo de processamento
- Atributos redundantes nos dados; atributos desnecessários
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - **Número de atributos**



Processo de Mineração de Dados

Pré-Processamento

Redução de Dados

- Limitações de espaço em memória, tempo de processamento
- Atributos redundantes nos dados, atributos desnecessários, ...
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - Número de atributos
 - **Número de valores**

Discretização

A se $\text{atr} < 2,5$
B se $2,5 \leq \text{atr} < 3,5$
C se $3,5 \leq \text{atr}$

atr
1
1
2
3
3
3
4
5
5
7

A
B
C

Processo de Mineração de Dados

Pré-Processamento

Redução de Dados

- Limitações de espaço em memória, tempo de processamento
- Atributos redundantes nos dados, atributos desnecessários, ...
- A redução pode ser realizada de três formas:
 - Número de exemplos
 - Número de atributos
 - **Número de valores**
 - Discretização
 - **Suavização**

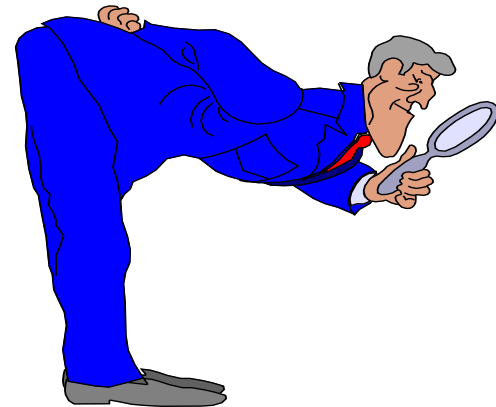
atr	
1	1
1	1
2	1
3	3
3	3
3	3
4	5
5	5
5	5
7	5

Valor
mediano

Processo de Mineração de Dados

Extração de padrões

- Etapa é direcionada ao cumprimento dos objetivos identificados na fase de identificação do problema
- Processo iterativo
 - Escolha da **atividade** e da **tarefa**
 - Escolha do **algoritmo**
 - Extração dos **padrões**



Processo de Mineração de Dados

Extração de padrões

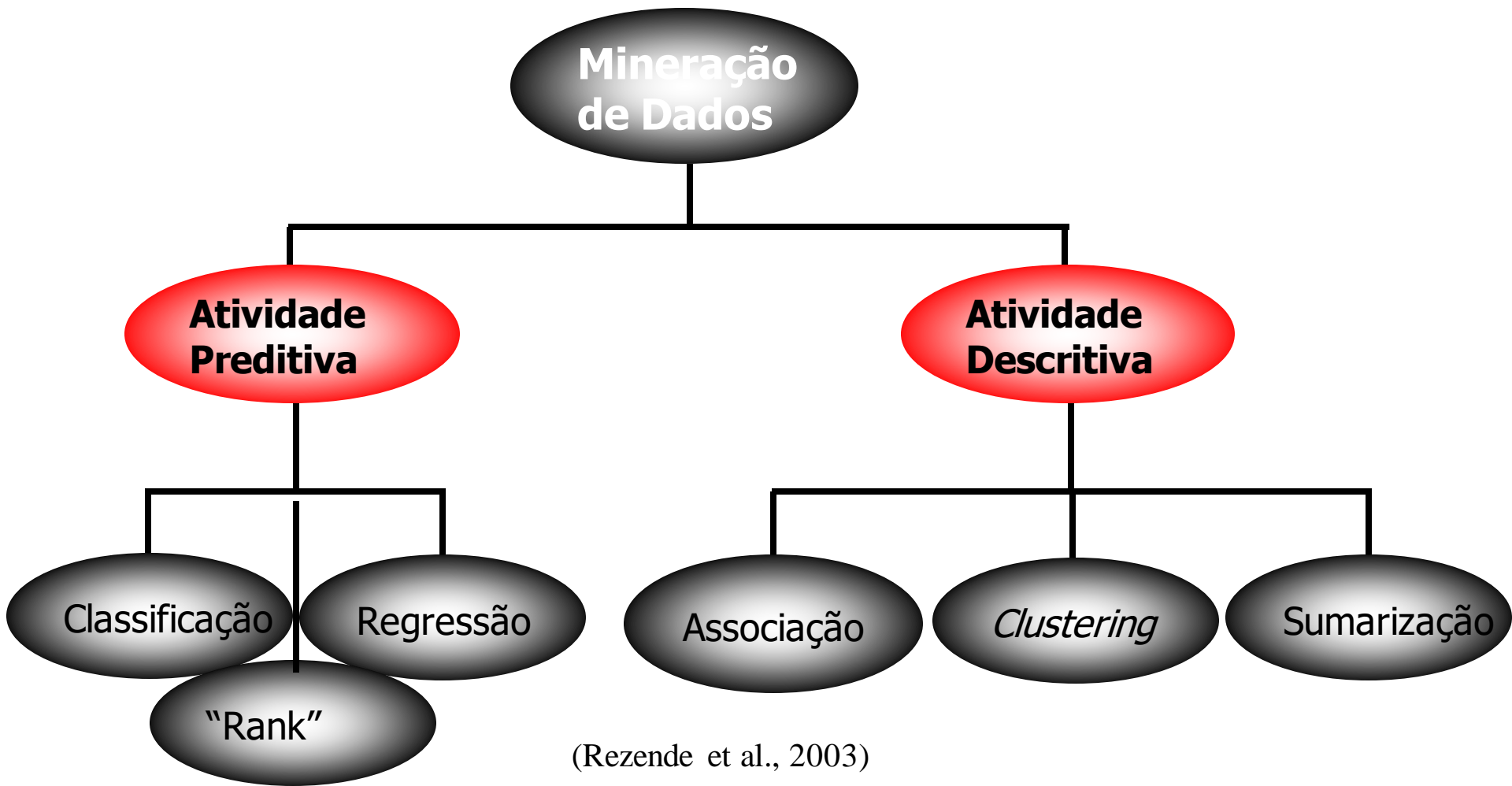
Escolha da atividade e da tarefa

- Deve ser feita de acordo com os objetivos desejáveis para a solução a ser encontrada
- Atividades podem ser agrupadas em:
 - Atividades **preditivas**: corresponde ao aprendizado **supervisionado**
 - Atividades **descritivas**: corresponde ao aprendizado **não-supervisionado**

Processo de Mineração de Dados

Extração de padrões

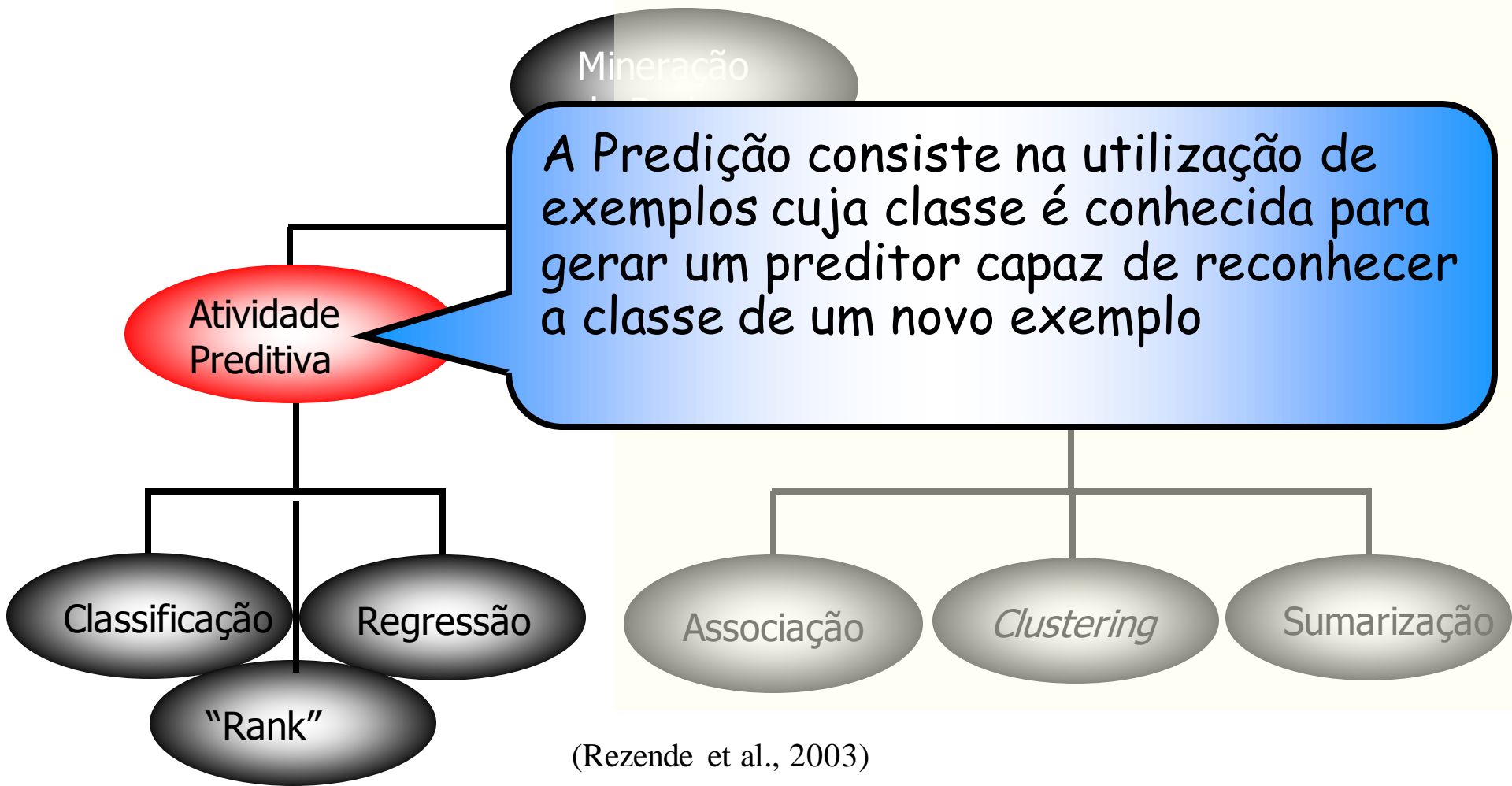
Escolha da atividade e da tarefa



Processo de Mineração de Dados

Extração de padrões

Escolha da atividade e da tarefa

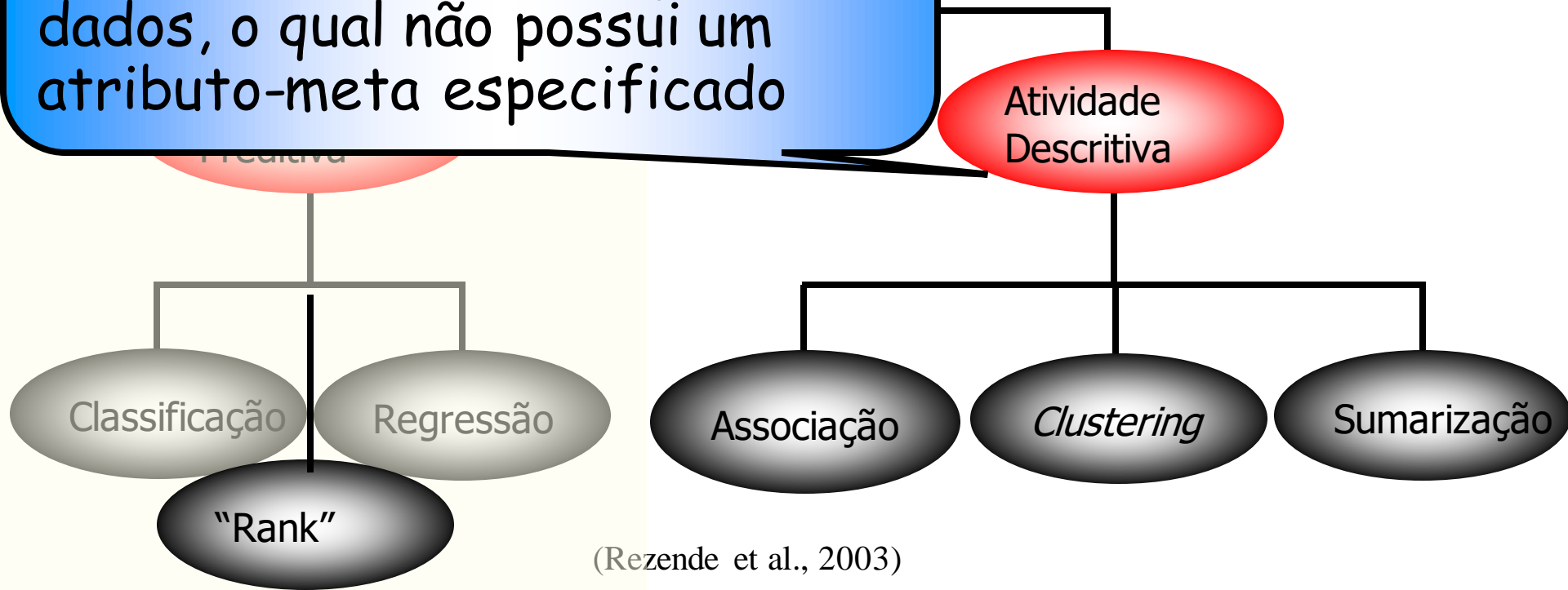


Processo de Mineração de Dados

Extração de padrões

Escolha da atividade e da tarefa

A Descrição consiste na identificação de padrões de comportamento no conjunto de dados, o qual não possui um atributo-meta especificado



(Rezende et al., 2003)

Processo de Mineração de Dados

Extração de padrões

Escolha do algoritmo

- Para efetuar a busca de padrões, podem ser utilizados algoritmos de Aprendizado de Máquina, ou outros...
- A escolha de um algoritmo é vista como um processo analítico, pois nenhum deles tem desempenho ótimo em todos os domínios de aplicação
- Um fator relacionado com a configuração dos parâmetros dos algoritmos é a complexidade da solução a ser buscada
- Vários algoritmos estão disponíveis para cada atividade
- Considerar a Representação do Conhecimento
 - Árvores de Decisão
 - Regras
 - Redes Neurais Artificiais...

Processo de Mineração de Dados

Extração de padrões

Escolha do algoritmo

- Todo algoritmo indutivo tem um bias
 - Tendência a privilegiar um tipo de hipóteses (modelos) em detrimento de outros
- Desempenho de um algoritmo varia com o domínio
 - Não existe um algoritmo que seja ótimo para todas as aplicações
- Análise experimental é **FUNDAMENTAL**

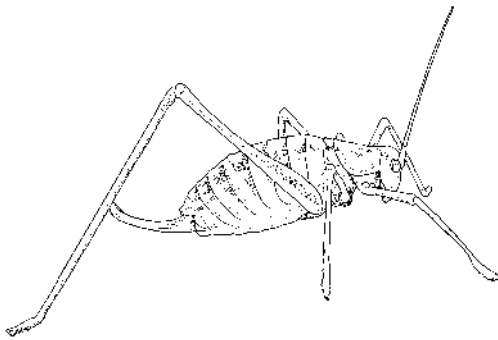
As Principais Tarefas/Métodos de Mineração de Dados

- Classificação
 - Agrupamento
 - Associações
- A maioria das demais tarefas (por exemplo, **descoberta de outliers ou detecção de anomalias**) fazem uso pesado de uma ou mais das tarefas acima.

Métodos Básicos

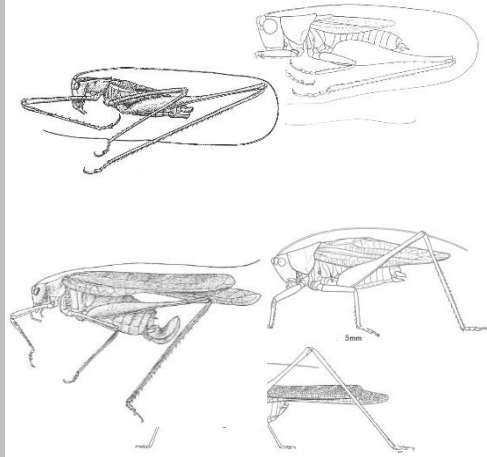
Classificação – definição informal

Dada uma coleção anotada de dados. Neste caso, cinco instâncias de **Esperanças** e cinco de **Gafanhotos**, decida a qual tipo de inseto o exemplo não rotulado pertence.

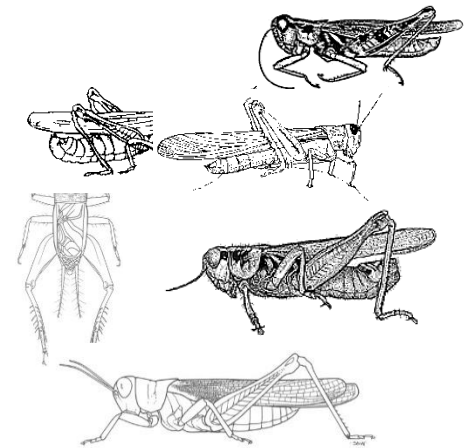


Esperança ou **Gafanhoto** ?

Esperança



Gafanhoto



Métodos Básicos

Processo de Classificação

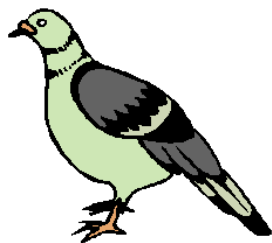
- Constituído de dois passos
 - **Aprendizado (ou treinamento)**: constrói o modelo de classificação
 - Utiliza um conjunto de dados para os quais é sabido o valor da classe
 - **Classificação**: modelo é utilizado para prever a classe de um objeto
 - Aplicado sobre objetos para os quais não se sabe o valor da classe
- Fase de aprendizado consiste no **aprendizado de uma função objetivo f** que mapeia um conjunto de atributos x a um dos rótulos de classes y (resulta na função $y=f(x)$)
 - Este mapeamento é apresentado na forma de regras de classificação, árvores de decisão ou fórmulas matemáticas



Voltaremos ao slide anterior em dois minutos. Enquanto isso vamos jogar um jogo rápido.

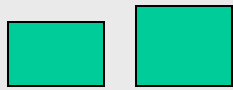
Vou mostrar a vocês alguns problemas de classificação que foram mostrados a pombos!

Vamos ver se você é tão esperto quanto um pombo



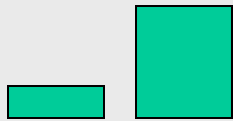
Problema do Pombo 1

Exemplos da classe A



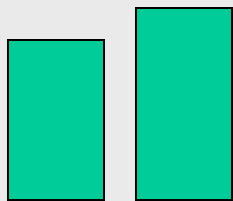
3

4



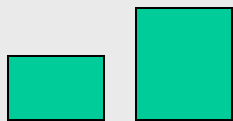
1.5

5



6

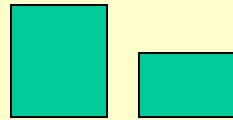
8



2.5

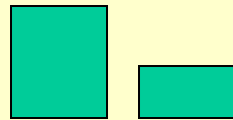
5

Exemplos da classe B



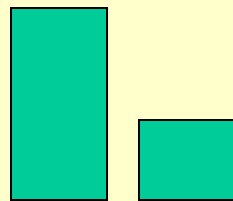
5

2.5



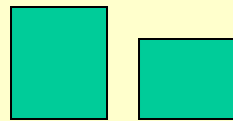
5

2



8

3

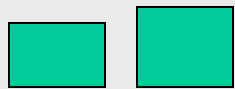


4.5

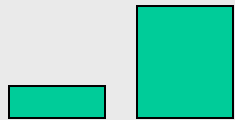
3

Problema do Pombo 1

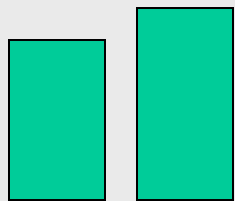
Exemplos da classe A



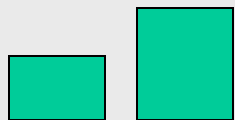
3 4



1.5 5

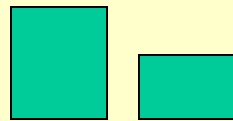


6 8

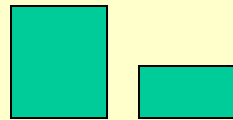


2.5 5

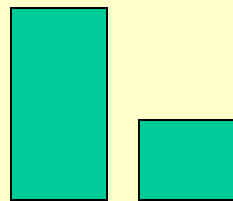
Exemplos da classe B



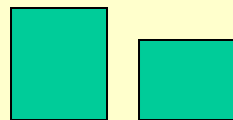
5 2.5



5 2

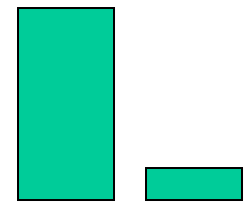
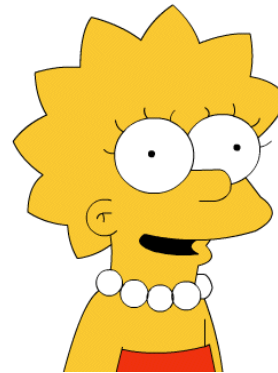


8 3



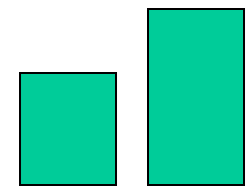
4.5 3

De qual classe é este objeto?



8 1.5

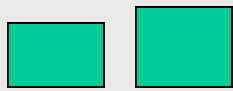
Que tal este, **A** ou **B**?



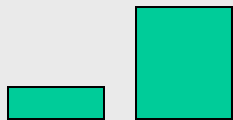
4.5 7

Problema do Pombo 1

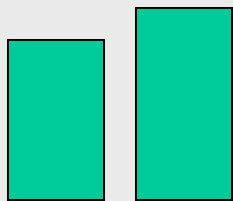
Exemplos da classe A



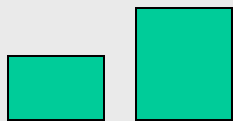
3 4



1.5 5

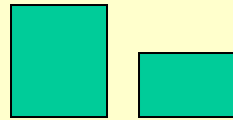


6 8

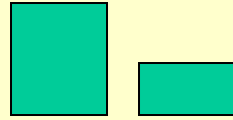


2.5 5

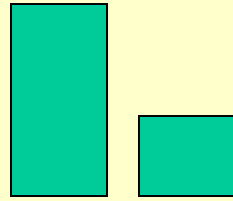
Exemplos da classe B



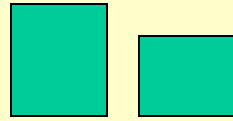
5 2.5



5 2



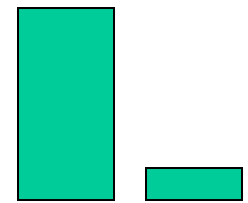
8 3



4.5 3



Este é um **B**!

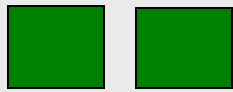


8 1.5

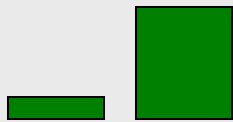
Eis a regra. Se a barra esquerda é menor que a barra direita, é um **A**, caso contrário é um **B**.

Problema do Pombo2

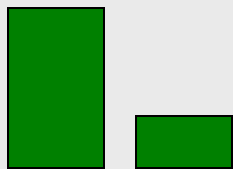
Exemplos da classe A



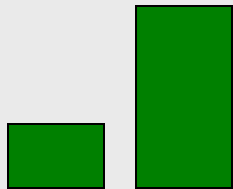
4 4



1 5

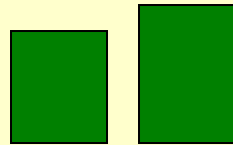


6 3

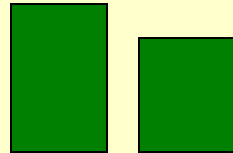


3 7

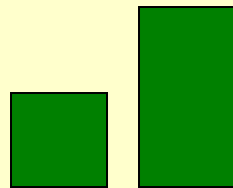
Exemplos da classe B



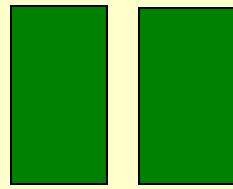
5 6



7 5

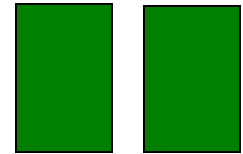


4 8



7 7

É um **B**?

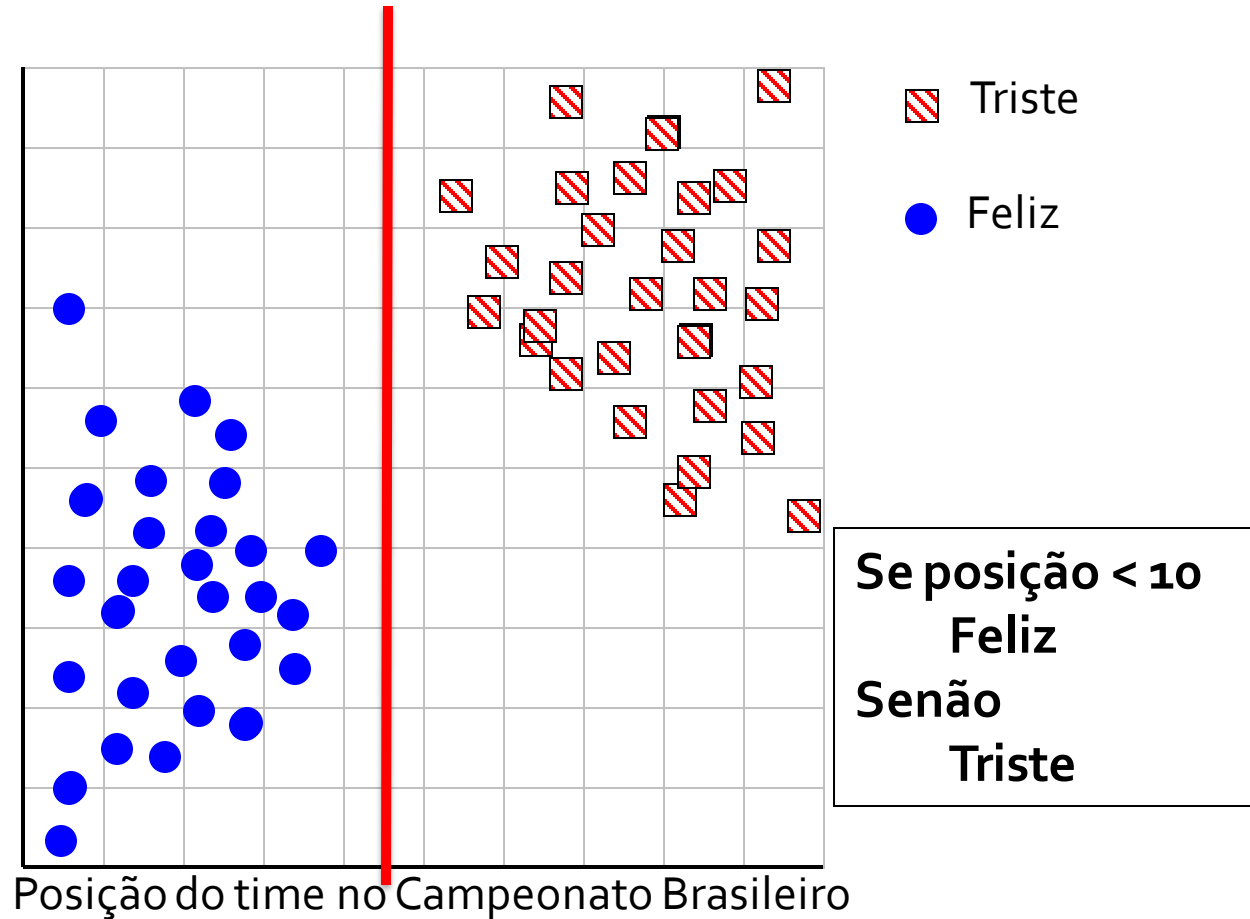


6 6

A regra é a seguinte, se o quadrado da soma das duas barras é menor ou igual a 100, é um **A**.

Caso contrário é um **B**.

Métodos Básicos Aprendizado e Classificação



Métodos Básicos

Aprendizado e Classificação

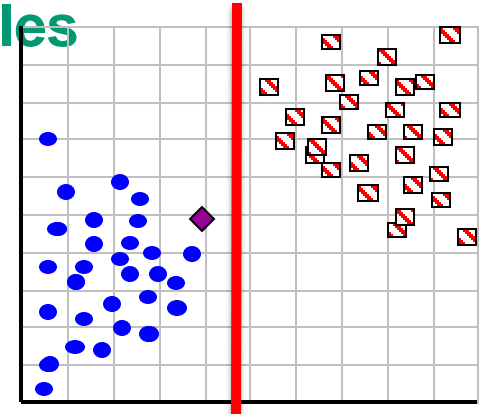
- Em quase todos os problemas de **classificação**, há uma interpretação geométrica
- Os problemas anteriores eram extremamente simples
 - Era possível aprender um modelo perfeito utilizando apenas uma linha
 - Chamamos de problema **Linearmente Separável**
 - Para classificarmos uma nova instância ainda não vista, basta usar um **Classificador Linear Simples**

Se instância ainda não vista está antes da linha

Classe Feliz

Senão

Classe Triste

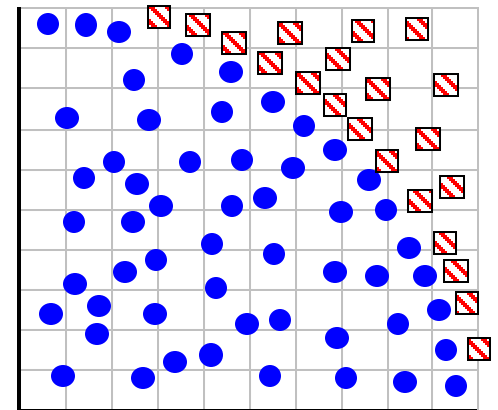
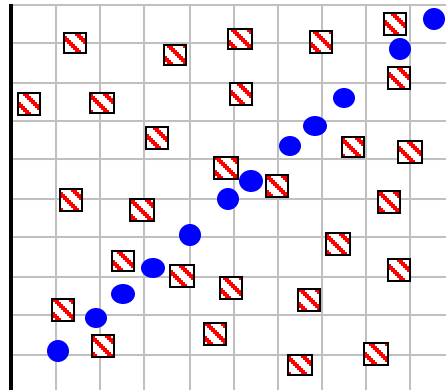
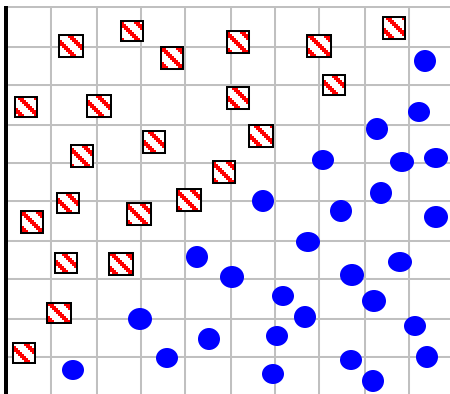


Posição do time no Campeonato Brasileiro

Métodos Básicos

Aprendizado e Classificação

- Mas nem todos os problemas são linearmente separáveis...Na verdade, quase nenhum problema real é linearmente separável!



Principais Algoritmos

☐ **K Nearest Neighbours (kNN)**

Hastie, T. and Tibshirani, R., 1996. Discriminant Adaptive Nearest, Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616. DOI= <http://dx.doi.org/10.1109/34.506411>

☐ **Naive Bayes**

Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.

☐ **C4.5**

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.

☐ **SVM**

Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.

☐ **Redes Neurais MLP (Multi-Layer Perceptron)**

Werbos, P.J. (1975). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences

Métodos Básicos

Associação

- Identifica itens que tem grande probabilidade de ocorrerem juntos em uma mesma transação da base de dados.
- Comumente utilizada em análise de cestas de mercado.



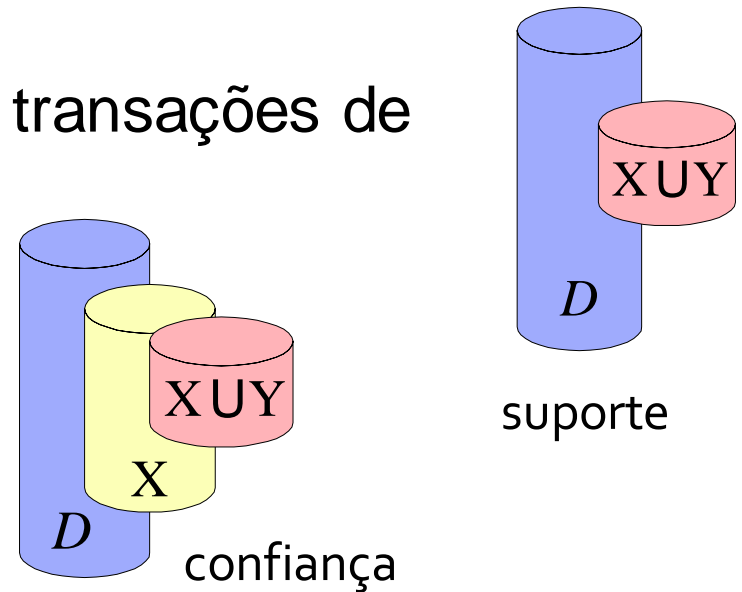
Exemplo de uma regra de associação: *fralda* → *cerveja*, indicando que o cliente que compra *fralda*, tende a comprar *cerveja*.

Métodos Básicos

Associação – Medidas interesse

Suporte de $X \rightarrow Y$: Porcentagem das transações de D que contém $X \cup Y$

Confiança de $X \rightarrow Y$:
$$\frac{\text{Suporte}(X \cup Y)}{\text{Suporte}(X)}$$



- Itemset X é frequente se $\text{Suporte}(X) \geq \text{minsup}$
- Uma **regra** $X \rightarrow Y$ é forte se $\text{Suporte}(X \rightarrow Y) \geq \text{minsup}$ e $\text{Confiança}(X \rightarrow Y) \geq \text{minconf}$

Principais Algoritmos

❑ **Apriori**

Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In Proc. of the 20th Int'l Conference on Very Large Databases (VLDB '94), Santiago, Chile, September 1994.

<http://citeseer.comp.nus.edu.sg/agrawal94fast.html>

❑ **FP-Growth** - que usa uma árvore de padrões frequentes (FP-Tree)

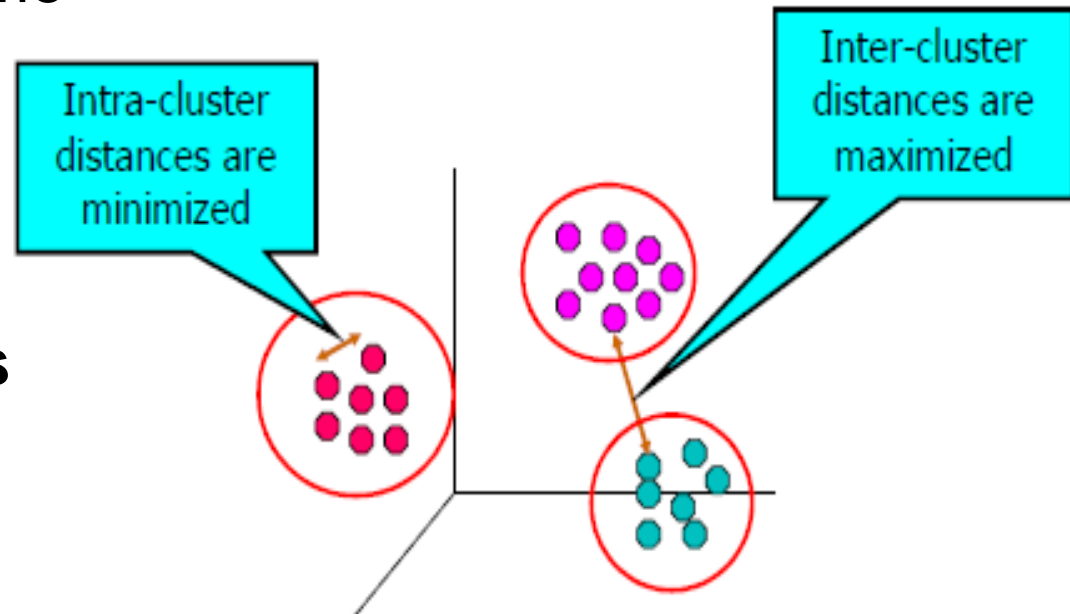
Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (Dallas, Texas, United States, May 15 - 18, 2000). SIGMOD '00. ACM Press, New York, NY, 1-12.

DOI= <http://doi.acm.org/10.1145/342009.335372>

Métodos Básicos

Agrupamento

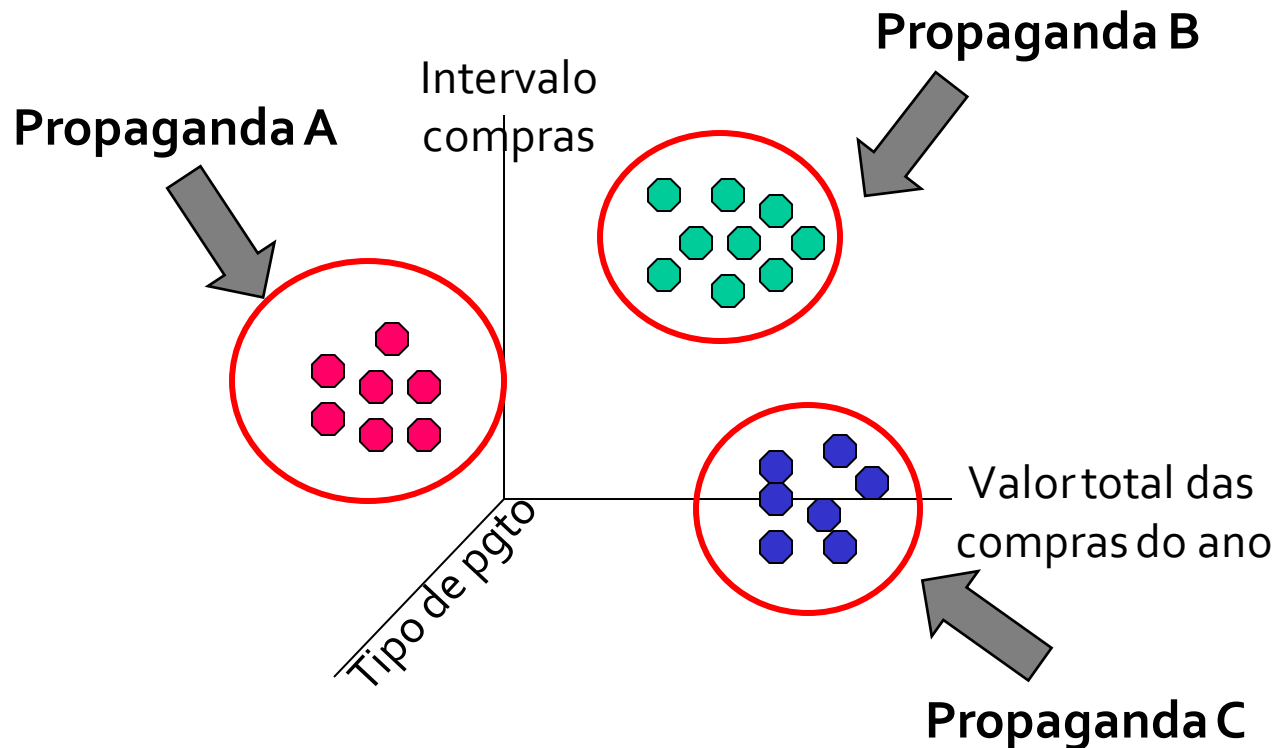
- É uma tarefa na qual pesquisam-se os dados com a finalidade de identificar grupos (cluster) de registros similares baseado em valores próximos de seus atributos
 - Elementos de um mesmo *cluster* são **mais** similares
 - Elementos de *clusters* diferentes são **menos** similares



Métodos Básicos

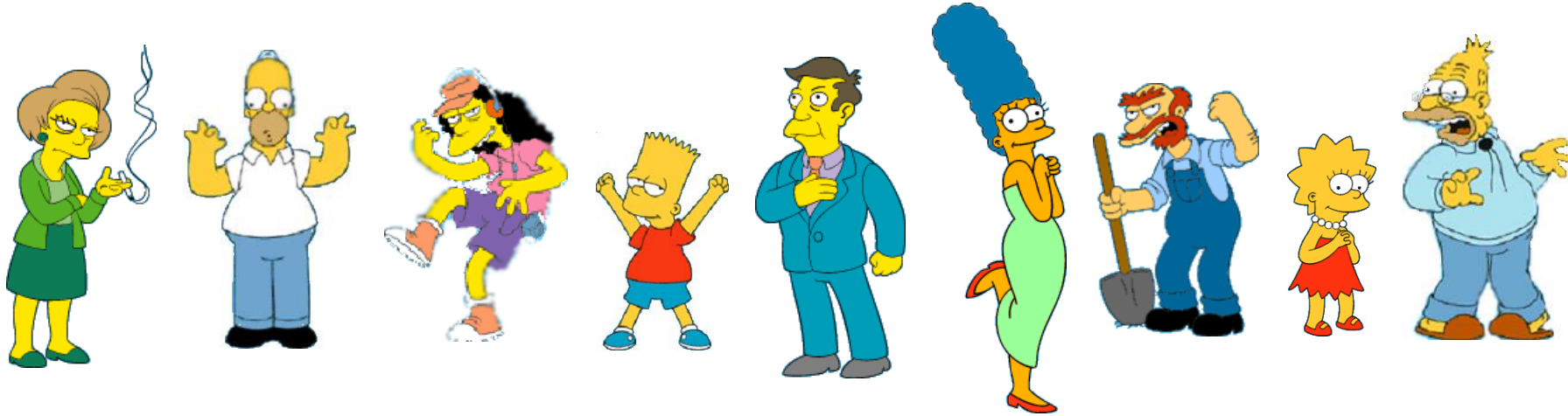
Agrupamento uso comum

Marketing: descobrir grupos de clientes / nichos de mercado e usá-los para marketing direcionado

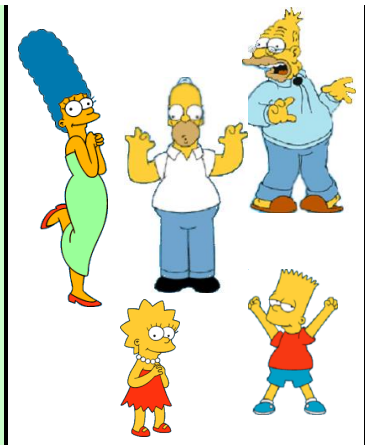


Métodos Básicos

Como agrupar estes objetos?



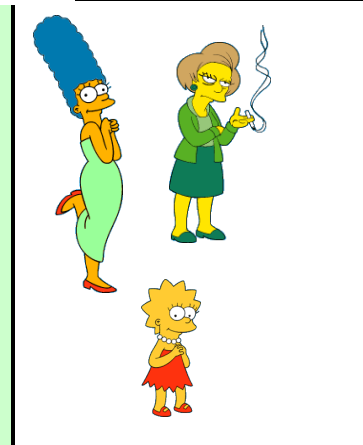
o agrupamento é subjetivo



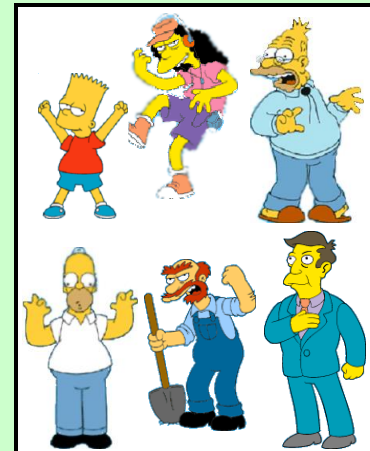
Os Simpsons



Empregados da Escola



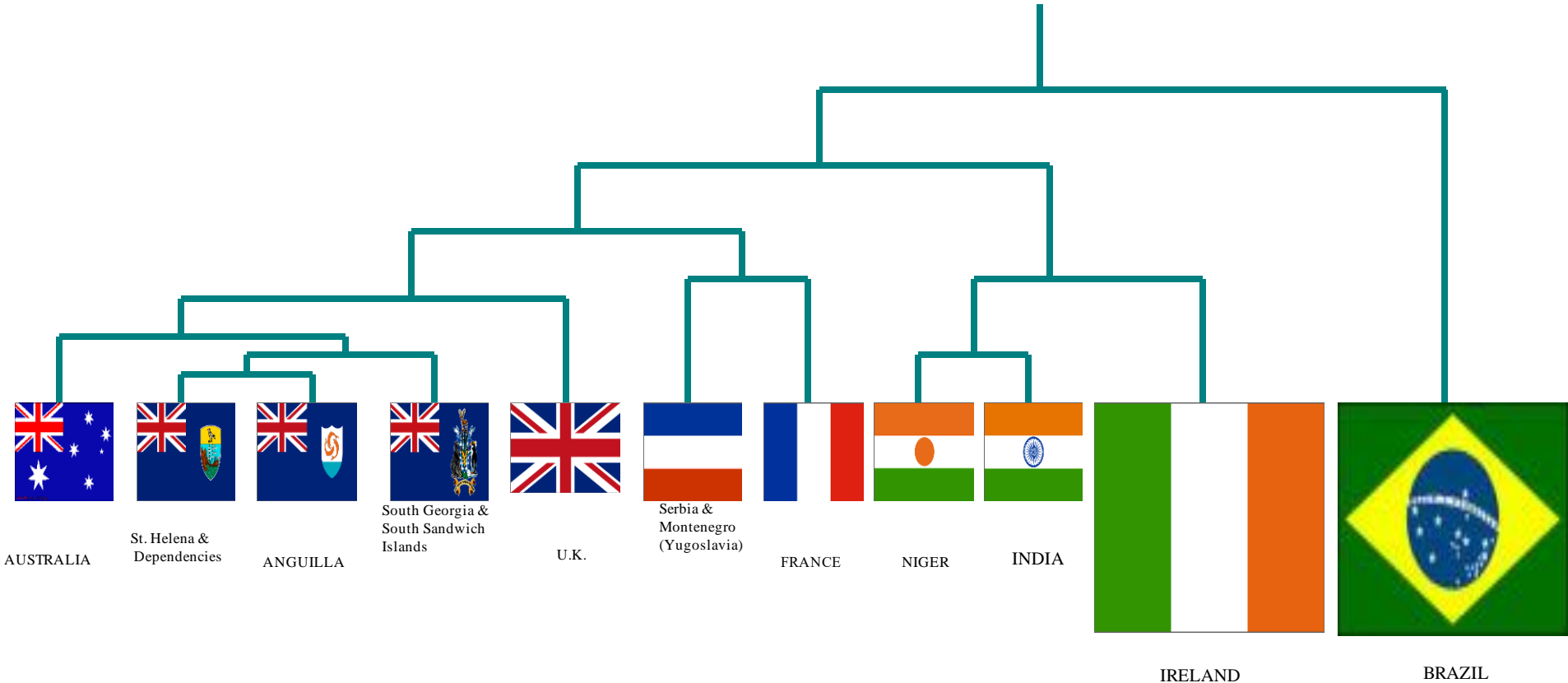
Mulheres



Homens

Métodos Básicos

Agrupamento hierárquico... como avaliar?



Como identificar/calcular as distâncias entre os objetos?

Métodos Básicos

O que é similaridade?

A qualidade, caráter ou condição das coisas similares.

(Dicionário Houaiss)



Similaridade é difícil de definir, mas...
Reconhecemos quando a vemos!

O real significado de similaridade é uma questão filosófica.

Nós vamos utilizar uma abordagem mais pragmática.

Métodos Básicos

Algumas Aplicações Agrupamento

- Marketing: descobrir grupos de clientes / nichos de mercado e usá-los para marketing direcionado
- Astronomia: encontrar grupos de estrelas e galáxias
- Estudos sobre terremotos: observar se epicentros estão agrupados em falhas continentais
- Bioinformática: encontrar grupos de genes com expressões semelhantes
- Mineração de Textos: organização de documentos
- Vários outros: proc. imagens, controle, determinar anomalias, ...

Principais Algoritmos

❑ K-Means

MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281-297.

❑ Hierarchical Clustering

Murtagh, Fionn. "A survey of recent advances in hierarchical clustering algorithms." The Computer Journal 26.4 (1983): 354-359.

❑ DBSCAN

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

❑ EM

McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.

❑ BIRCH

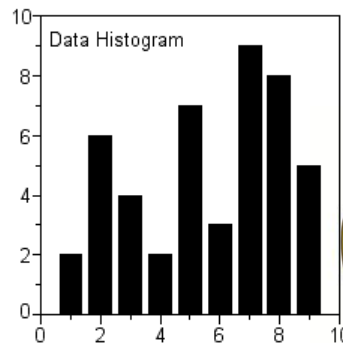
Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an eficiente data clustering method for very large databases. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (Montreal, Quebec, Canada, June 04 - 06, 1996). J. Widom, Ed. SIGMOD '96. ACM Press, New York, NY, 103-114. DOI= <http://doi.acm.org/10.1145/233269.233324>

Processo de Mineração de Dados

Pós-Processamento

Avaliação do conhecimento extraído

- O conhecimento extraído representa o conhecimento do especialista?
- De que maneira o conhecimento do especialista difere do conhecimento extraído?
- Em que parte o conhecimento do especialista está correto?



Processo de Mineração de Dados

Pós-Processamento

Avaliação do conhecimento

- Pode-se ter uma quantidade enorme de padrões que podem não ser importantes, relevantes ou interessantes aos usuários
- Não é muito interessante fornecer uma quantidade grande de padrões ao usuário para ser avaliado
- Desenvolver técnicas de apoio para fornecer padrões mais interessantes
- Envolve técnicas de visualização dos padrões

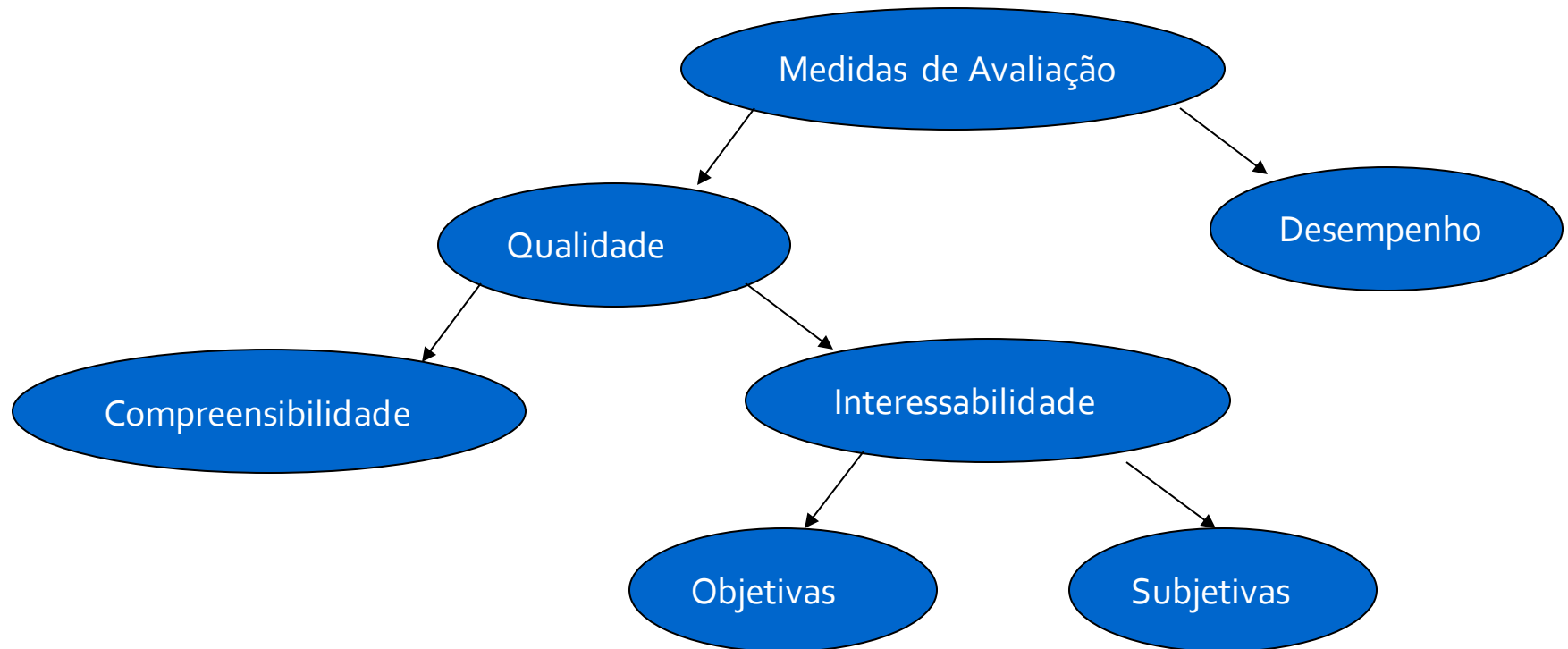


Processo de Mineração de Dados

Pós-Processamento

Medidas de avaliação

Existem diversas medidas para auxiliar o usuário no entendimento e na utilização do conhecimento adquirido



Processo de Mineração de Dados

Pós-Processamento

O processo de extração de conhecimento não termina após a descoberta dos padrões

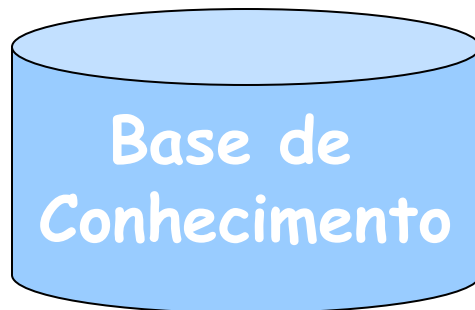
- Análise dos resultados pode indicar falhas nos dados, no pré-processamento dos dados ou mesmo na escolha dos algoritmos
- Caso falhas sejam detectadas, deve-se retornar no processo e refazê-lo

Processo de Mineração de Dados

Utilização do conhecimento

Após ter percorrido todas as etapas do processo com sucesso, o conhecimento encontra-se apto a ser utilizado pelo usuário

- Garante-se, com o correto desenvolvimento das etapas anteriores, que o **conhecimento** é válido e útil, podendo ser **aplicado no apoio à tomada de decisão**
- Formação de uma base de conhecimento, que é incorporada a um **Sistema Inteligente**



Sistema Inteligente

Processo de Mineração de Dados

Utilização do conhecimento

No final do processo de Mineração de Dados, é interessante que todo o **conhecimento adquirido seja disponibilizado** em um ambiente adequado para facilitar a sua exploração, interpretação e utilização

Algumas Ferramentas

- Weka – classificação, regressão, regras de associação e clustering além de outras funcionalidades
<http://www.cs.waikato.ac.nz/ml/weka/>
- Apriori – regras de associação
<http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html>
- Cubist – regressão e See5 – classificação
<http://www.rulequest.com/cubist-info.html> e [.../see5-info.html](http://www.rulequest.com/see5-info.html)
- Orange – uma ferramenta para visualização e análise de dados tanto para iniciantes quanto para especialistas
<http://orange.biolab.si/download/>
- R - Ferramenta open-source contendo pacotes para análise estatística e para mineração de dados

-
- KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source ([GPLv3](#)) Java software tool
 - O SAS Viya Data Mining and Machine Learning é um ambiente escalável, aberto, de análises avançadas
 - KNIME - Konstanz Information Miner

<https://www.knime.org/>

- **Torch - Topic Hierarchies**

<http://sites.labic.icmc.usp.br/torch/>

- Site LABIC Software and Application Tools

labic.icmc.usp.br

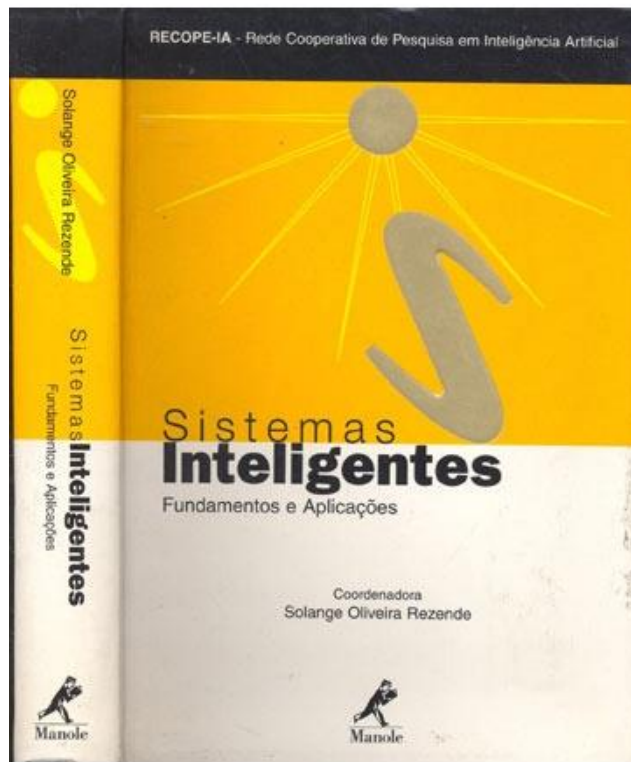
Text Categorization Tool API (implementada em Java) - para representações usando Redes

Word2Vec (implementado em Python) - para representações usando sequencia de sentenças

Referências Gerais

- Fayyad, U. M. ; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Mineração de Dados*, MIT Press, 1996.
- Witten, I. H.; Frank, E. *Mineração de Dados: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.
<http://www.cs.waikato.ac.nz/~ml/weka/book.html>
- Pyle, D. *Data Preparation for Mineração de Dados*, Morgan Kaufmann Publishers, 1999.
- Thuraisingham, B. *Mineração de Dados: Technologies, Techniques, and Trends*, CLR Press LLC, 1999.
- Rezende, S. O; *Sistemas Inteligentes: Fundamentos e Aplicações*; Ed Manole 2003.
- Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier, 2011.
- Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to Data Mining*. Addison-Wesley, 2005.
- Faceli, K,; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. L. F.; *Inteligência Artificial: Uma abordagem de aprendizado de máquina*, Editora LTC, 2012.

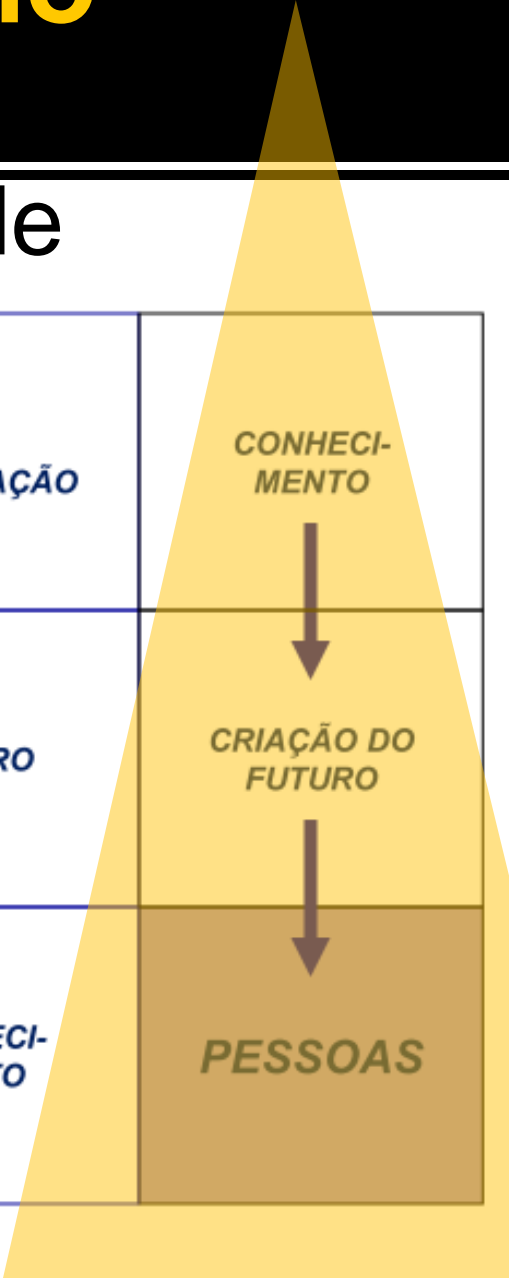
Referências Gerais



Mineração de Dados como Vantagem Competitiva

Evolução da Sociedade

SOCIEDADE	<i>AGRÍCOLA</i>	<i>INDUSTRIAL</i>	<i>DA INFORMAÇÃO</i>	<i>CONHECI- MENTO</i>
FONTE DE CRESCIMENTO	<i>PASSADO</i>	<i>PRESENTE</i>	<i>FUTURO</i>	<i>CRIAÇÃO DO FUTURO</i>
RECURSO ESTRATÉGICO (PODER)	<i>TERRA</i>	<i>CAPITAL</i>	<i>CONHECI- MENTO</i>	<i>PESSOAS</i>



Não há limites para a MD...



FIM
@Solange Rezende