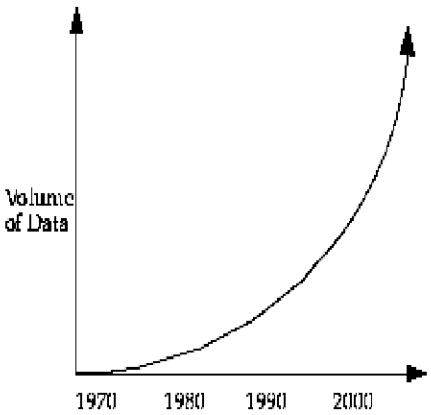
DATA MINING

Data Mining

- **Motivações**
  - Grande disponibilidade de dados armazenados eletronicamente
  - Existem informações úteis, invisíveis, nesses grandes volumes de dados
  - Aproveitar para prever um conhecimento futuro (ir além do armazenamento explícito de dados).



[01] Queens University - Belfast

Introdução

- **O que é Data mining:**
  - "Data mining (mineração de dados), é o processo de extração de conhecimento de grandes bases de dados, convencionais ou não.
  - Utiliza técnicas de inteligência artificial que procuram relações de similaridade ou discordância entre dados.
  - Seu objetivo é encontrar, automaticamente, padrões, anomalias e regras com o propósito de transformar dados, aparentemente ocultos, em informações úteis para a tomada de decisão e/ou avaliação de resultados.

Fernando S. Goulart Júnior & Robson do Nascimento Fidalgo (1998)

3

Introdução

- **Uma empresa utilizando data mining é capaz de:**
  - Criar parâmetros para entender o comportamento do consumidor;
  - Identificar afinidades entre as escolhas de produtos e serviços;
  - Prever hábitos de compras;
  - Analisar comportamentos habituais para detectar fraudes.

[09] - Dissertação - UERJ (1997)

4

Introdução

- **Data mining X Data warehouse:**
  - Data mining ⇒ extração inteligente de dados;
  - Data warehouse ⇒ repositório centralizado de dados;
  - Data mining não é uma evolução do Data warehouse;
  - Data mining não depende do Data warehouse, mas obtém-se melhores resultados quando aplicados em conjunto;
  - Cada empresa deve saber escolher qual das técnicas é importante para o seu negócio. (Uma, outra ou as duas);
  - Data Warehouse aliado a ferramentas estatísticas desempenham papel semelhante ao data mining, ***mas não descobrem novos padrões de comportamento***. (a não ser empiricamente).

5

Introdução

- **Evolução até o data mining**

Evolução	Perguntas	Tecnologia disponível	Características
Coleção de dados 1960	"Qual foi meu rendimento total nos últimos cinco anos ?"	Computadores, Fitas, discos	Retrospectiva, Dados estáticos como resposta
Acessos aos dados 1980	"Qual foi meu rendimento no Brasil no último janeiro ?"	RDBMS, SQL, ODBC	Restropectiva, dados dinâmicos a nível de registos como resposta
Data warehousing & suporte a decisão 1990	"Qual foi meu rendimento no Brasil no último janeiro? Do sul até o nordeste	Processamento analítico on-line, banco de dados multidimencionais, data warehousing	Retrospectiva, dados dinâmicos em múltiplos níveis como resposta
Data Mining Atualmente	"Porque alguns produtos são mais vendidos na região sul ?"	Algoritmos avançados, computadores multiprocessados, B.D. grandes e poderosos	Prospectivo, Informações (perspectivas) como resposta.

6

O Background para data mining

- **Como fazer Data Mining ?**
- **Aprendizagem computacional (Machine-learning)**
  - Automação do processo de aprendizagem, através da construção de regras baseadas em observações dos estados e transações do ambiente.
  - Examina os exemplos e seus resultados e aprende como reproduzi-los e como fazer generalizações sobre novos casos

7

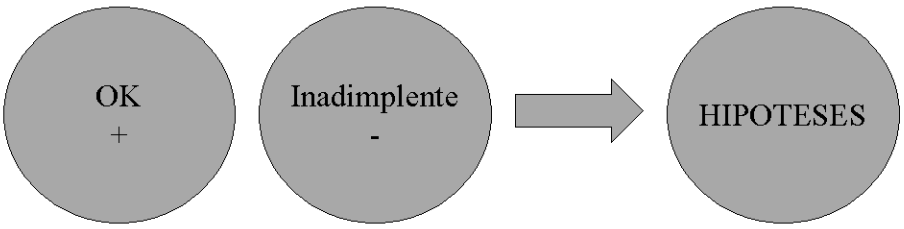
O Background para data mining

- **Aprendizagem indutiva:**
  - Faz análise nos dados para encontrar padrões
  - Agrupa objetos similares em classes
  - Formula regras
  - **Aprendizagem supervisionada (A.S.)**
    - Aprende baseando-se em exemplos ("professor" ajuda a construir um modelo def. classes e fornecendo exemplos de cada classe ⇒ formular a descrição e a forma da classe)
  - **Aprendizagem não supervisionada (A.Ñ.S.)**
    - Aprende baseando-se em observações e descobertas (não se def. classes, deve-se observar os exemplos e reconhecer os padrões por si só ⇒ uma descrição de classes para cada ambiente).

8

O Background para data mining

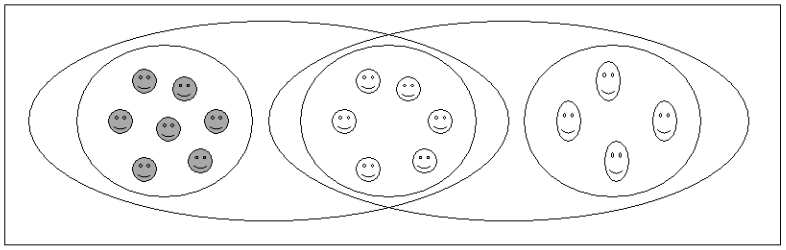
- **Modelo de verificação (Aprendizagem supervisionada)**
  - *Aprende baseando-se em exemplos pré-classificados (+/-)*
  - *Objetivo: formular descrições consistentes e gerais de classes em função de seus atributos.*



9

O Background para data mining

- **Modelo de descoberta (Aprendizagem ñ supervisionada)**
  - Aprende baseando-se em observações e descobertas;
  - Descoberta automática de informações ocultas;
  - Procura ocorrências de padrões, tendências e generalizações sobre os dados sem a intervenção do usuário;
  - Agrupar elementos similares,



10

Funções do data mining

- **Modelo de classificação :**
  - Atributos + significativos def. uma classe
  - O usuário define os atributos para cada classe **(A.S.)**
  - Aplica regras para criar modelos de ações futuras
    - Ex: Clientes com bom crédito ñ podem dever mais 10%

- **Associação:**
  - Procura registos que tenham similaridades associativas
  - Podem ser expressados por regras
    - Ex: 62% dos compradores de guaraná compram pipoca
    - 62% fator de confiança

11

Funções do data mining

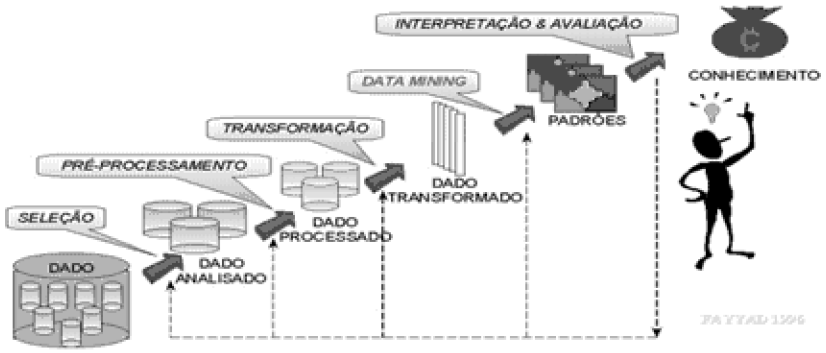
- **Padrões temporais/seqüenciais :**
  - Analisa registros num período de tempo, procurando encontrar padrões (eventos/compras) de comportamento.
  - Identificar o perfil do cliente
  - Identificar padrões que precedem outros padrões
    - Ex: Mala direta personalizada, Campanhas promocionais ..

- **Segmentação/agrupamento:**
  - Segmenta a base de dados em grupos por suas similaridade e diferenças
  - O sistema tem que descobrir por si próprio as similaridade e diferenças **(A.Ñ.S.)**
    - Ex: Clientes de bom crédito e clientes de mau crédito

12

O Processo Data Mining

- **Fases / Etapas.**
  - Seleção.
  - Pré-processamento.
  - Transformação.
  - Data mining.
  - Interpretação e Avaliação.



O Processo Data Mining

- **Seleção**
  - Selecionar ou segmentar dados de acordo com critérios definidos:
    - Ex.: Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinado.

- **Pré-processamento**
  - Estágio de limpeza dos dados, onde informações julgadas desnecessárias são removidas.
    - Ex. :O sexo de um paciente gestante
  - Reconfiguração dos dados para assegurar formatos consistentes (identificação)
    - Ex. : sexo = "F" ou "M"
      sexo = "M" ou "H"

14

O Processo Data Mining

- **Transformação**
  - Transforma-se os dados em formatos utilizáveis. Esta depende da técnica data mining usada.
    - Ex: rede neural ⇒ converter valor literal em valor numérico
  - Disponibilizar os dados de maneira usável e navegável.

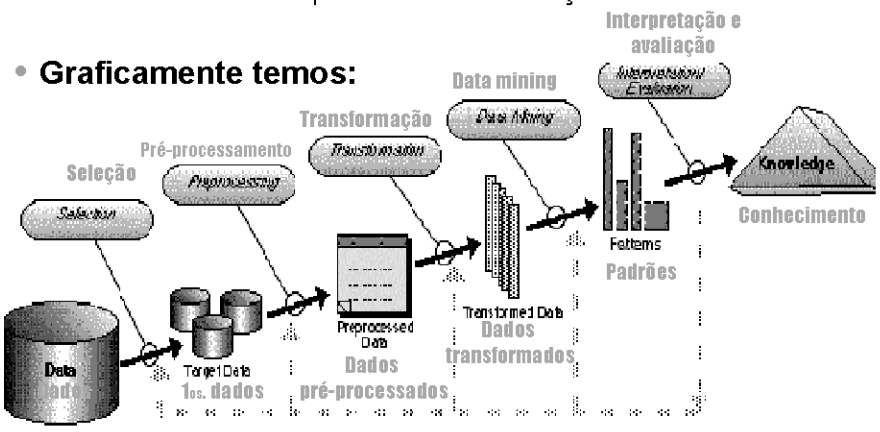
- **Data mining**
  - É a verdadeira extração dos padrões de comportamento dos dados
  - Utilizando a definição de fatos, medidas de padrões, estados e o relacionamento entre eles.

15

O Processo Data Mining

- **Interpretação e Avaliação**
  - Identificado os padrões pelo sistema, estes são interpretados em conhecimentos, os quais darão suporte a tomada de decisões humanas
    - Ex.: Tarefas de previsões e classificações

- **Graficamente temos:**



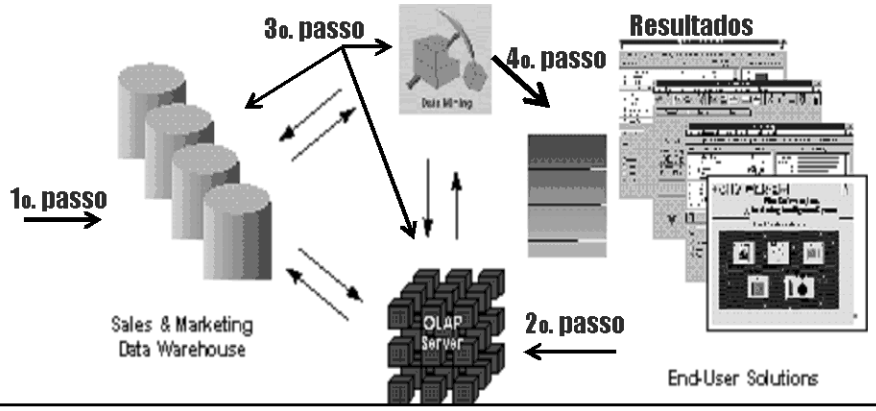
Uma arquitetura data mining

1º ⇒Data warehouse com dados pertinentes ao negócio;

2º ⇒Servidor OLAP permitindo ao usuário analisar os dados do D.W. de forma mais produtiva ;

3º ⇒ Integrar o D.W. e o OLAP com o D.M. (permitir decisões oper.)

4º ⇒ Modelos de metadados definidos pelo data mining



Técnicas

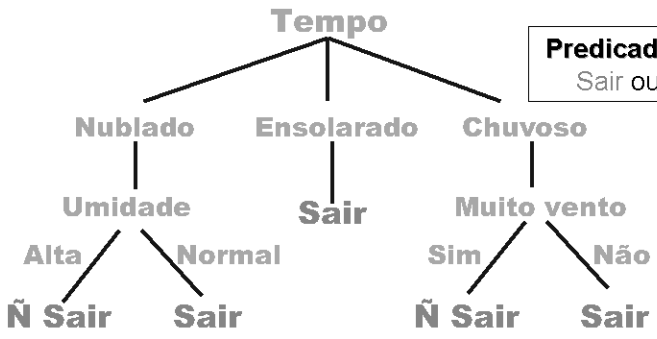
- **Indução**
  - Regras indutivas (rule induction)
    - Regra indutiva é o processo de olhar uma série de dados e, a partir dela, gerar padrões.
    - Pode-se trabalhar com dados numéricos ou não
    - Pelo fato de explorar uma série de dados, o sistema indutivo cria hipóteses que conduzem a padrões
      - Ex.: Analisada as idades de profissionais foi gerada a seguinte regra:
        - Se profissão = atleta
          ENTÃO idade < 30

Regras cobertas ⇒ comportamentos estáveis
Regras inexatas ⇒ margem de precisão "fixada" **(%)**

18

Técnicas

- **Árvores de decisão: (ID3) - 1º Exemplo**
  - Representações simples do conhecimento
  - Utilização de regras condicionais
  - A partir de um conjunto de valores decide SIM ou NÃO
  - Mais rápida e mais compreensível que redes neurais
  - Exemplo: Sair ou não de acordo com o tempo



Predicado objetivo:
Sair ou Não Sair

19

Técnicas

- **Árvores de decisão: (ID3) - 2º Exemplo**

Classificação de um indivíduo com risco de ter ou não crédito

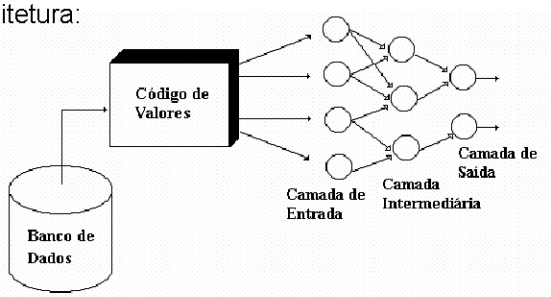
Predicado objetivo:
Crédito ou Não

Nesta árvore de decisões, regras são induzidas nos padrões dos dados e cria-se uma hierarquia de indicações "se-então".

20

Técnicas

- **Redes Neurais:**
  - É uma abordagem computacional que envolve desenvolvimento de estruturas matemáticas com a habilidade de aprender. (modelo do sistema nervoso para aprender)
  - Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios/nós), que possuem entrada, saída e processamento.
  - São organizados em camadas que aprendem pela modificação da conexão.
  - Arquitetura:



21

Técnicas

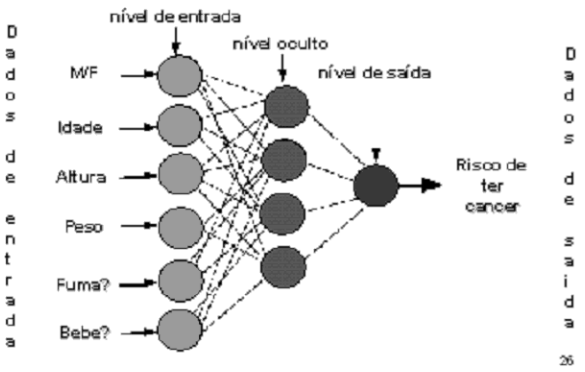
- **Redes Neurais:**
  - Para construir um modelo neural, nós primeiramente "adestramos" a rede em um dataset de treinamento e então usamos a rede já treinada para fazer predições.
  - Problemas:
    - Não retorna informação a priori
    - Não pode ser treinada em uma grande base de dados
    - Entrada não pode ser dados alfa-numéricos (mapear para numérico)
    - Nenhuma explanação dos dados é fornecida (caixa preta)

22

Técnicas

- **Redes Neurais:**
  - Exemplo prático: risco de câncer

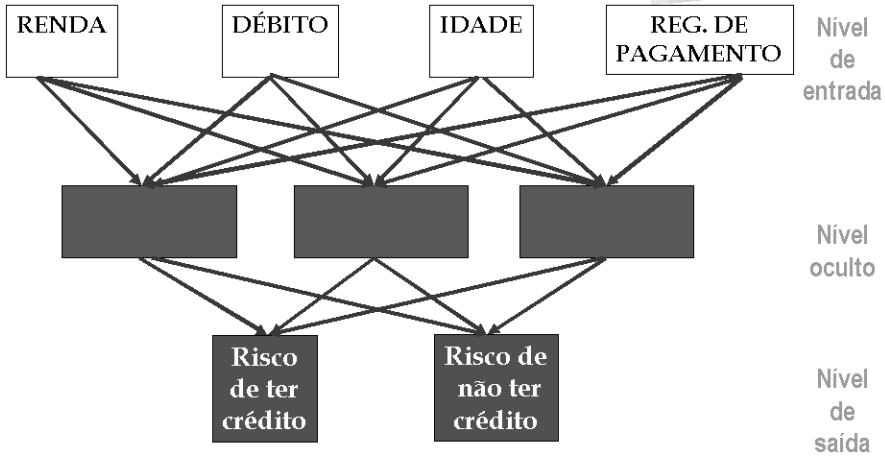
– Exemplo prático: risco de câncer



23

Técnicas

- **Redes Neurais: - 2º Exemplo**



As redes neurais usam seus dados de entrada.
Atribui pesos nas conexões entre os atributos (neurônios).
E obtém um resultado (risco de ter ou não crédito) no nível de saída.

Exemplos

- **Áreas de aplicações potenciais:**
  - Vendas e Marketing
    - Identificar padrões de comportamento de consumidores
    - Associar comportamentos à características demográficas de consumidores
    - Campanhas de marketing direto (mailing campaigns)
    - Identificar consumidores "leais"
  - Bancos
    - Identificar características de correntistas
    - Mercado Financeiro ($$$)
  - Médica
    - Comportamento de pacientes
    - Identificar terapias de sucessos para diferentes tratamentos
    - Comportamento de usuários de planos de saúde

25

Introdução

- **Exemplo (1) - Fraldas e cervejas**
  - O que as cervejas tem a ver com as fraldas ?
  - homens casados, entre 25 e 30 anos;
  - compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa;
  - Wal-Mart otimizou às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas;
  - Resultado: o consumo cresceu 30% .

26

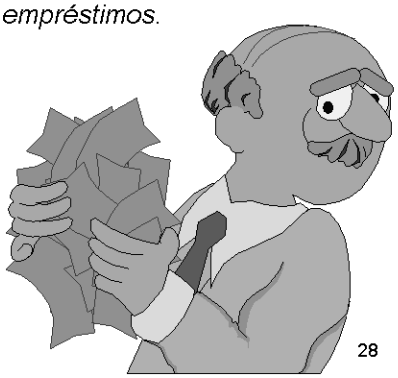
Exemplos

- **Exemplo (2) - Lojas Brasileiras (Info 03/98)**
  - Aplicou 1 milhão de dólares em técnicas de data mining
  - Reduziu de 51000 produtos para 14000 produtos oferecidos em suas lojas.
  - Exemplo de anomalias detectadas:
    - Roupas de inverno e guarda chuvas encalhadas no nordeste
    - Batedeiras 110v a venda em SC onde a corrente é 220v

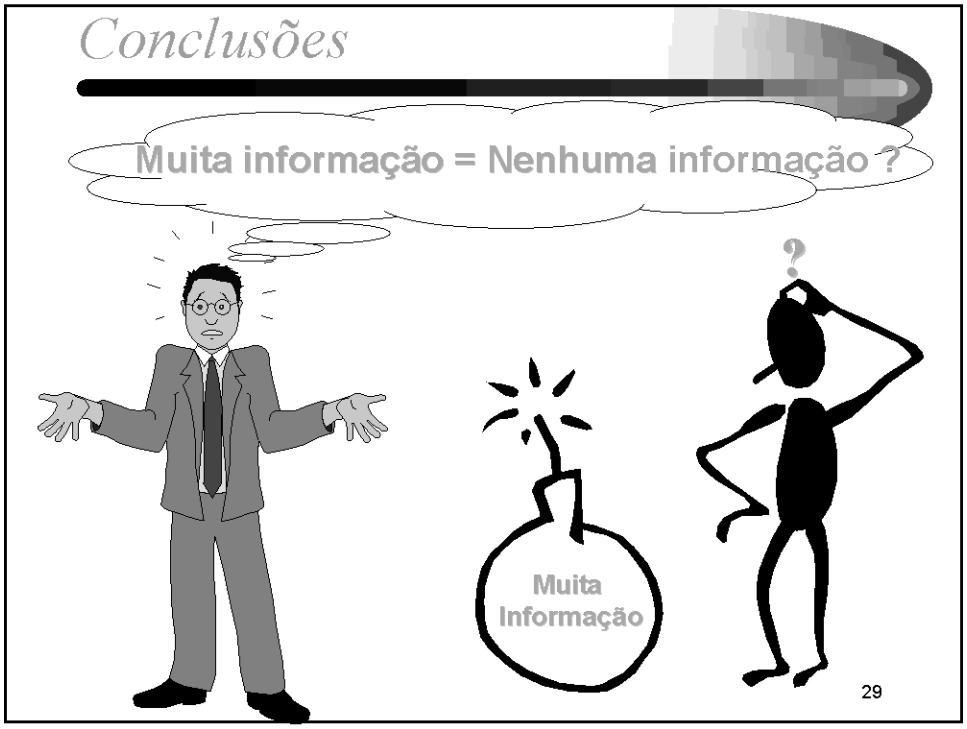
27

Exemplos

- **Exemplo (3) - Bank of America (Info 03/98)**
  - Selecionou entre seus 36 milhões de clientes
    - *Aqueles com menor risco de dar calotes*
    - *Tinham filhos com idades entre 18 e 21 anos*
    - *Resultado em três anos o banco lucrou 30 milhões de dólares com a carteira de empréstimos.*



28


Conclusões
Muita informação = Nenhuma informação ?
Muita
Informação
?
29