# Banco de Dados II

Mineração de Dados – Parte 1

# Sistemas de Informação

- Entrada: dados
- Processamento: algoritmos
- Saída: informação

# Origem dos Dados....

- Banco de Dados

# Banco de Dados

- Banco de Dados ?
  - “Uma coleção de dados relacionados” (ELMASRI & NAVATHE, 2010)
  - “ Conjunto de dados integrados que tem por objetivo atender a uma comunidade de usuários” (HEUSER, 2009)
  - “Uma coleção de dados”

    (KORTH; SILBERSCHATZ; SUDARSHAN, 2006)

# Inicialmente...

- Sistemas de Informação
  - Automatizar rotinas
- Dados
  - Relatórios básicos – exigências legais
    - Balancete, relatório de vendas, relação de clientes...

O quê mais eu posso fazer com dados armazenados nos bancos de dados?

# Conhecimento

- Difícil definir...
- “Conhecimento é um entendimento ou modelo sobre pessoas, objetos ou eventos, derivado de informações sobre eles” (GORDON; GORDON, 2006)
- Exemplo:
  - Os gestores de uma loja obtêm conhecimento sobre as preferências dos clientes (perfil) a partir das informações obtidas como resultado de dados acumulados sobre estes clientes (compras, etc.)

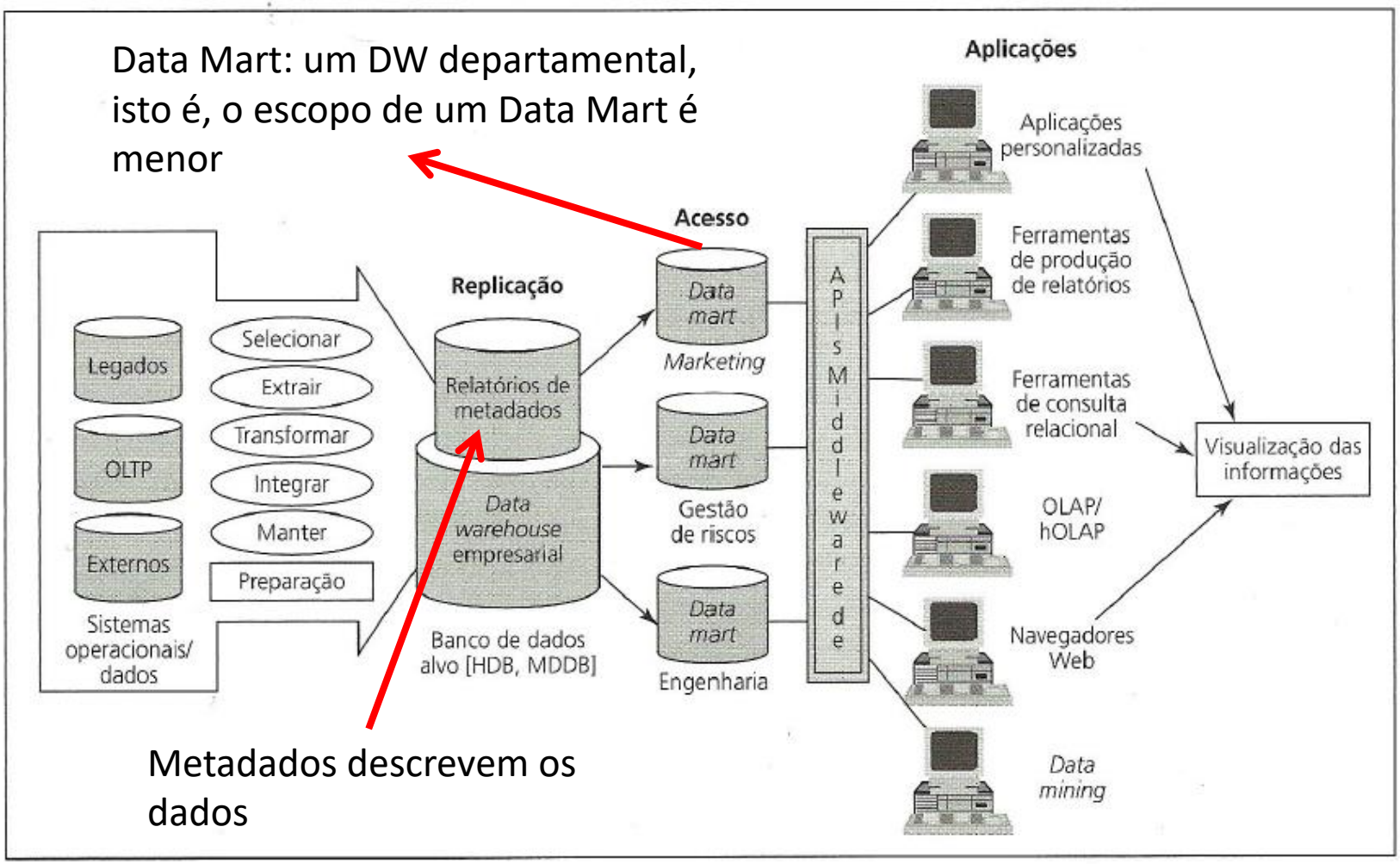
# Inteligência de Negócios - *Business Intelligence*

- Um “guarda chuva” que inclui arquiteturas, ferramentas, bancos de dados e metodologias (TURBAN et al., 2009)
- Objetivo:
  - Facilitar o processo de tomada de decisão
  - Permitir o acesso interativo aos dados, proporcionar a manipulação destes dados e fornecer aos gerentes e analistas de negócios a capacidade de realizar a análise adequada (TURBAN et al., 2009)

# *Business Intelligence*

- Principais componentes :
  - Um repositório/ banco de dados (Data Warehouse)
  - Ferramentas para visualização (OLAP)
  - Ferramentas de Descoberta de Conhecimento em Banco de Dados (Data mining, Text Mining...)

# DW - Ambiente




Fonte: TURBAN et al., 2009

# Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery Database – KDD*)

- Processo não trivial de identificação de padrões, a partir de dados válidos, novos, potencialmente úteis e compreensíveis. (FAYYAD, 1996).
- Data Mining é normalmente considerado parte do processo de DCBD.

# Processo de DCBD

1. **Definição do problema**. É necessário conhecimento do domínio
2. **Seleção dos dados**. Uma parte dos dados é selecionada.
3. **Limpeza dos dados/Pré-processamento dos dados**. Inconsistências são corrigidas (o destino por ser um DW).
4. **Transformação dos dados**. Eventualmente é reduzido o número de variáveis ou de registros a serem consideradas no processo de mineração de dados. Exemplo: discretização.
5. **Mineração dos dados/Data Mining**. Envolve escolha de algoritmos de mineração.
6. **Interpretação dos dados**. Os resultados do processo de mineração são interpretados

**Resultado: Conhecimento...(ou não...)**

# Data Mining - Mineração de Dados

- Mineração de dados utiliza técnicas e algoritmos de diferentes áreas do conhecimento:
  - Inteligência artificial (aprendizagem de máquina)
  - Banco de dados (recursos para manipular grandes bases de dados)
  - Estatística (avaliação e validação de resultados)
- Mineração de Dados => Algoritmos

# O Valor da Informação

- Informação descoberta deve ser
  - Nova
  - Inesperada
  - Válida (estatisticamente)

- Valor da Informação => Impacto nas decisões

# Tarefas de Mineração de Dados

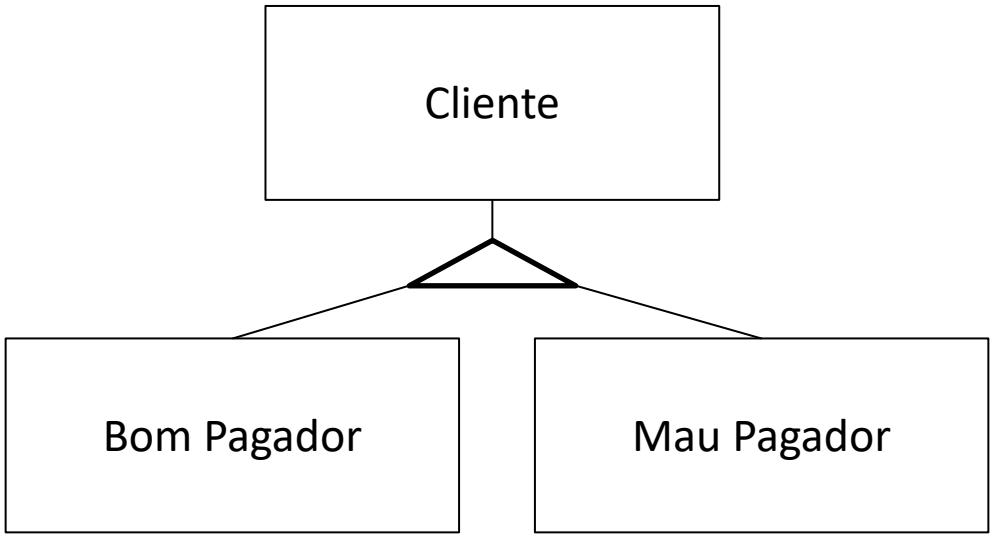
- Indica o tipo de problema que será resolvido
  - Classificação
  - Agrupamento (*Clustering*)
  - Associação
  - Regressão

# Classificação

- É construído um modelo que possa ser aplicado a dados não classificados de forma associar estes dados a classes pré-definidas

# Classificação

- Existe uma taxonomia, um conjunto de classes onde eu desejo classificar os registros



# Classificação

- Outros exemplos
  - Classificar pedidos de crédito (conceder ou não)
  - A partir do perfil do cliente (dados) prever se ele pode realizar determinada compra
  - Classificar documentos (*text mining*)

# Regressão (Estimativa)

- Usada para definir um valor a alguma variável contínua
  - Estimar número de filhos
  - Estimar valor de venda
  - Estimar nota que será atribuída a um objeto (produto)

# Associação

- No caso da classe de uma tarefa de mineração não ser determinada como na classificação uma possibilidade é o uso de algoritmos de associação
  - Algoritmo de associação
  - Itens que ocorrem juntos
  - Exemplo (clássico!!!): Fraldas -> Cerveja
    - Quem compra fraldas compra cerveja...

# Agrupamento ou *Clustering*

- Agrupar registros de uma determinada base de dados levando em conta características específicas
  - Agrupar clientes com hábitos de compra similares
  - Agrupar usuários Web com comportamento similar

# Ferramentas de Data Mining

- Wizrule (demo):
  - http://www.wizsoft.com/
- Weka (free):
  http://www.cs.waikato.ac.nz/ml/weka/
- The R Project for Statistical Computing
  - http://www.r-project.org/
- Python – inúmeras bibliotecas, ferramentas, recursos…

# Análise de Dados - Python

- Notebook
- Jupyter
- Anaconda
- Pandas

# Análise de Dados - Python

- Notebook.  Neste contexto…
  - É um documento virtual que contém código e textos explicativos (comentários do código e dos resultados).
  - Permite a execução do código e ver resultados da execução
  - Não é preciso executar todo Código do documento

# Análise de Dados - Python

- *Jupyter Notebook*
  - interface gráfica que permite a edição de *notebooks* em um navegador Web. O nome Jupyter é um acrônimo criado a partir das linguagens de programação que inicialmente foram aceitas: **Julia**, **Pyt**hon e **R**.
  - Jupyter Notebooks são armazenados como documentos JSON (extensão ".ipynb")
  - Documentos possuem 3 tipos de célua:
    - célula de código;
    - célula de Markdown possui texto formatado
      - https://pt.wikipedia.org/wiki/Markdown
      - https://docs.pipz.com/central-de-ajuda/learning-center/guia-basico-de-markdown#open
    - célula bruta.

# Análise de Dados - Python

- *Jupyter Notebook*
  - https://jupyter.org/
- *Google Colab*

  https://research.google.com/colaboratory/intl/pt-BR/faq.html