# Mineração de Dados: visão geral

Material produzido por Marcilio Souto – DIMAp/UFRN e complementado por Alexandre Zamberlan - Unifra

# O que é Mineração de Dados?

- Processo de **identificação de padrões** válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados (Fayyad et al, 1996)

- Encontra **informações úteis embutidas em GRANDES volumes de dados**
  - Analogia com a mineração
  - Grandes volumes de dados são “peneirados” na tentativa de se encontrar alguma coisa de valor

# Exemplos

- Qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade, analisando os dados dos últimos dez anos?
- Quais são os clientes potenciais para praticar fraudes?
- Quais clientes gostariam de comprar o novo produto X?
- Que genes são determinantes para o diagnóstico de um determinado tipo de doença?

# Resumindo

A mineração de dados é o processo de extrair informações válidas antes desconhecidas, de grandes bases de dados, auxiliando em decisões cruciais (ELMASRI, 2007).

- Diferencia dados de informação;
- Faz parte do processo de descoberta de conhecimento;
- Através de suas técnicas descobre-se:
  - Padrões;
  - Informações desconhecidas;

# Descoberta de Conhecimento

- Descoberta de conhecimento ou *Knowledge Discovery in Database (KDD)* é um outro termo para o processo de Mineração de Dados

- Alguns autores consideram os termos KDD e Mineração de Dados referentes a processos distintos
  - **Mineração de Dados seria uma etapa do processo de KDD**

# Mineração de Dados - uma área multidisciplinar

- Banco de Dados
- Estatística
- Computação de alto desempenho
- **Aprendizado de Máquina**
- Visualização
- Matemática

# Mineração de Dados e Sistemas Gerenciadores de Banco de Dados

- Exemplo de um relatório de um SGBD
  - Vendas dos últimos meses para cada tipo de serviço
  - Vendas por serviço agrupadas por sexo do cliente ou idade
  - Lista dos clientes que tiveram suas apólices canceladas
  - Perguntas respondidas usando Mineração de Dados
  - Que características têm os clientes que tiveram suas apólices canceladas e como elas diferem daquelas do cleintes que as renovaram?
  - Quais clientes que possuem seguros de carro que seriam potenciais clientes para seguros de casa?

# Data Warehouse

- Data Warehouse: repositório de dados centralizado que contém dados limpos, agregados e consolidados
  - Extrai dados operacionias históricos
  - Supera inconsistências entre diferentes formatos de dados
  - Incorpora informações adicionais ou de especialistas

# On-line Analytical Processing (OLAP)

- Multi-Dimensional Data Model (Data Cube)

- Operações
  - Roll-up
  - Drill-down
  - Slice and dice
  - Rotate

# Objetivos da Mineração de Dados

- **Atividades Preditivas**: Classificação e Regressão
  - Sistemas de mineração de Dados aprendem a partir de exemplos como particionar ou classificar os dados (p. ex., gerando regras de classificação)
  - Exemplo - base de dados de clientes de um banco
    - Pergunta: Um novo cliente solicitando um empréstimo é um bom ou mau investimento?
    - Regra típica formulada:
      - Se STATUS = casado e RENDA > 2000 e PROPRIETARIO-IMÓVEL = sim então TIPO-DE-INVESTIMENTO = bom

# Objetivos da Mineração de Dados

- **Atividades Descritivas:** Associação, Clustering, Sumarização
  - Regras de Associação
    - Regras que associam um atributo de uma relação a outro
    - Abordagens orientadas a conjuntos são os meios mais eficientes para a descobertas de tais regras
  - Exemplo - base de dados de um supermercado
    - 72% de todos os registros que contêm itens A e B também contêm item C
    - A porcentagem específica de ocorrências é o fator de confiança da regra

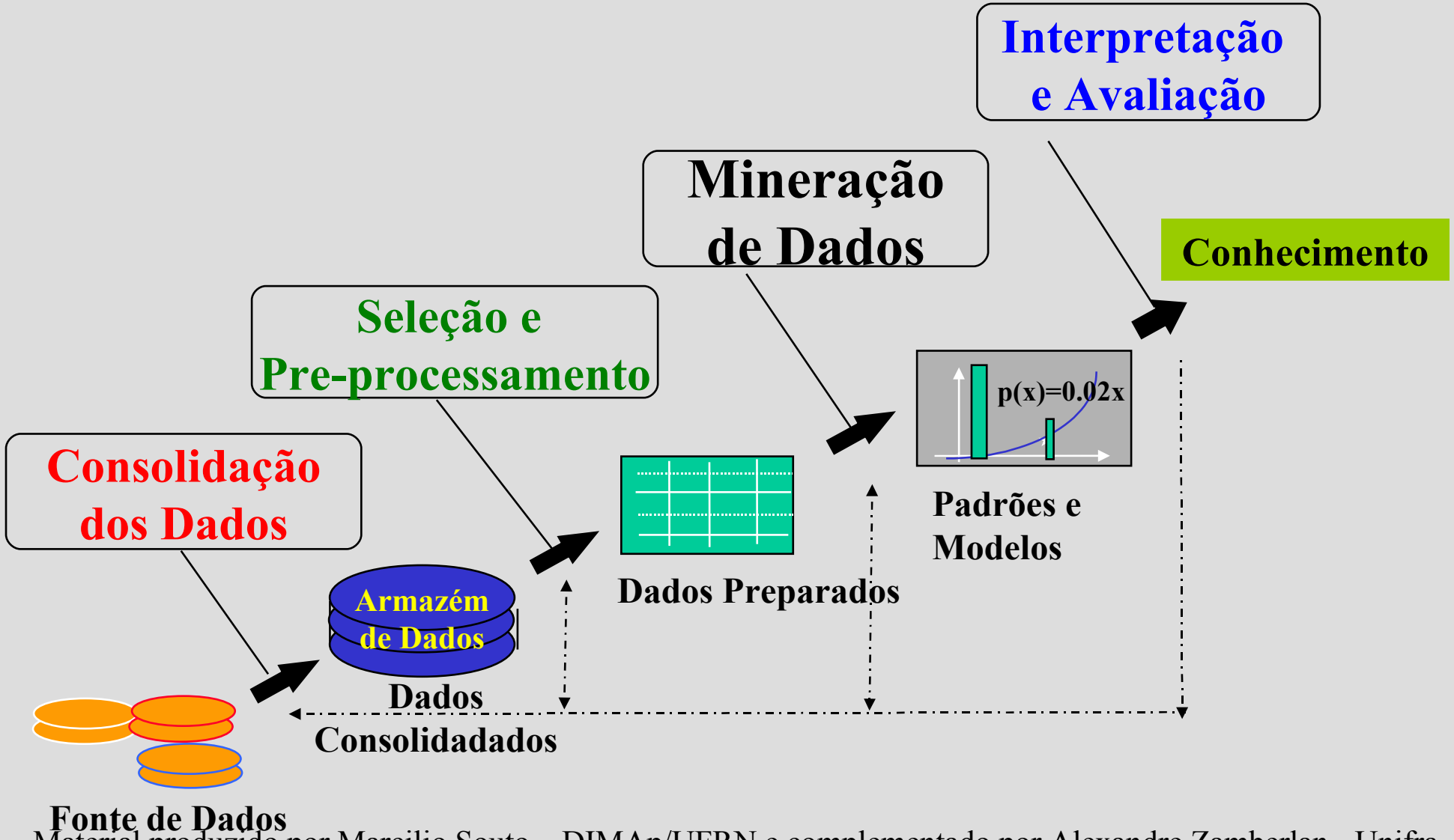
# Principais técnicas

- Regras de Associação (*Association Rules*): objetivo é encontrar qualquer relação existente entre os valores das variáveis;
- Agrupamento (*Clustering*): analisam os dados para encontrar grupos de itens que são semelhantes;
- Padrões seqüenciais (*Sequential Patterns*): é a investigação de seqüências de ações ou eventos;
- Padrões com séries temporais (*Time-Series Data*): esses padrões podem ser encontrados em posições de uma série temporal de dados, que é uma seqüência de dados capturada a intervalos regulares.

# SGBD, OLAP e Mineração de Dados

| Área | SGBD | OLAP | MD |
|---|---|---|---|
| Tarefa | Extração de dados detalhados e sumários | Sumários tendências e previsões | Descoberta de conhecimento de padrões embutidos e insights |
| Tipo de resultado | Informação | Análise | Insights e previsões |
| Método | Dedução (faça a pergunta, verifique os dados) | Modelagem de dados multi-dimensionais, agregação, estatísticas | Indução (construa o modelo, aplique-o a novos dados, obtenha o resultado) |
| Exemplo de pergunta | Quem comprou passagens internacionais nos últimos 3 anos? | Qual a renda média dos compradores de passagens internacionais por região? | Quem comprará uma passagem internacional nos últimos 6 meses e a razão? |

# Fases da Mineração de Dados



# Fases da Mineração de Dados

- Identificação do Problema
  - Quais são as principais metas do processo?
  - Quais critérios de desempenho são importantes?
  - O conhecimento extraído deve ser compreensível a seres humanos ou um modelo tipo caixa-preta é apropriado?
  - Qual a deve ser a relação entre simplicidade e precisão do conhecimento extraído?
  - Pré-processamento
  - Extração e Integração
  - Limpeza
  - Transformação
  - Seleção e Redução
- **Criação de um modelo - Aprendizado de Máquina**
  - Escolha da tarefa - classificação, regressão, associação, clustering, …
  - Escolha do(s) algoritmo(s)
  - Aplicação do(s) algoritmo(s)
- Teste do modelo
- Interpretação e avaliação

# Técnicas de Aprendizado de Máquina

- k-NN
- Naive Bayesian Learning
- Árvores de Decisão
- Regras
- Redes Neurais Artificias
- Support Vector Machines
- Ensembles
- Regras de Associação
- k-means
- Métodos de agrupamento hierárquico

# Aplicações de Mineração de Dados

- Atribuição de crédito
- Predição no mercado financeiro
- Diagnóstico de falhas em linhas de produção
- Descobertas médicas
- Detecção de fraudes
- Análise de tendências de compra
- Marketing direcionado
- ....

# Bibliografia


- Rezende, S. O. *et al.* (2003). Mineração de Dados. In Rezende, S. O. (org.) *Sistemas Inteligentes: Fundamentos e Aplicações,* Capítulo 12, pp. 307-333. Editora Manole Ltda.
- Witten, I. H. and Frank, E. (1999). *Data Mining: practical machine learning tools and techniques with Java implementations*. Chapter 1 - What's it all about?, pp. 1-36.