

# **KDD E MINERAÇÃO DE DADOS**

## **O Processo de KDD: Visão Geral**

**Prof. Ronaldo R. Goldschmidt**

[ronaldo@de9.ime.eb.br](mailto:ronaldo@de9.ime.eb.br)

[rribeiro@univercidade.br](mailto:rribeiro@univercidade.br)

[geocities.yahoo.com.br/ronaldo\\_goldschmidt](http://geocities.yahoo.com.br/ronaldo_goldschmidt)

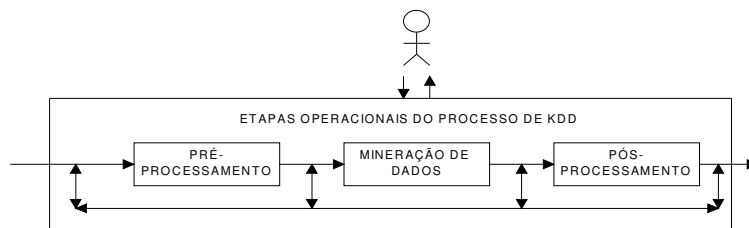
### **O PROCESSO DE KDD: VISÃO GERAL**

- **CARACTERIZAÇÃO**
- **ÁREAS DE ORIGEM**
- **GERAÇÕES DE KDD E MINERAÇÃO DE DADOS**
- **TAXONOMIA DE ATIVIDADES EM KDD**
- **ETAPAS DO PROCESSO**
- **TAREFAS DE KDD**
- **OPERAÇÕES E MÉTODOS**
- **TÉCNICAS E ALGORITMOS**
- **O PAPEL DO USUÁRIO NO PROCESSO DE KDD**
- **MACRO-OBJETIVOS DA MD**
- **ORIENTAÇÃO DA MD**
- **CONSIDERAÇÕES TÉCNICAS SOBRE PROJETOS**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **iterativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]



## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **iterativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Interação: Combinação de Ações Homem-Máquina.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Iteração: Refinamentos Sucessivos.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Padrão: Forma de Representação do Conhecimento.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Compreensão: Padrão Representado de Forma Intelegível.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Validade: Aplicação Adequada a um Contexto.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Inovação: Mudança de Ctos Anteriores p/ Ctos Descobertos.**

## O PROCESSO DE KDD: VISÃO GERAL

- **KDD – Knowledge Discovery in Databases**

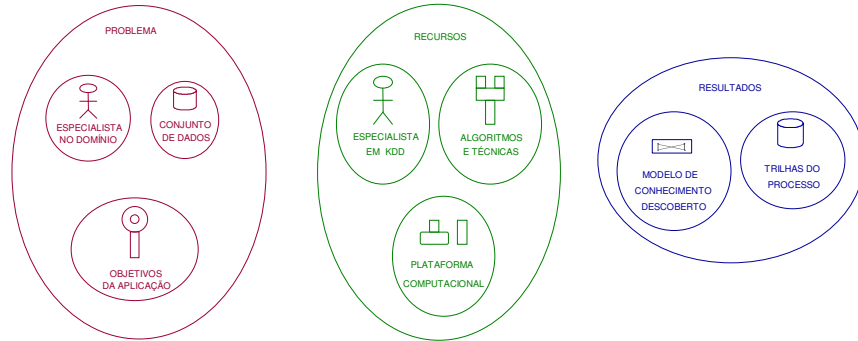
“É um **processo**, de várias etapas, não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis, válidos, novos** e potencialmente **úteis** a partir de bases de dados.” [Fayyad et al., 1996]

**Utilidade: Benefícios da Aplicação.**

## O PROCESSO DE KDD: VISÃO GERAL

### Aplicação de KDD:

- Envolve os seguintes elementos:



## 3. FORMALIZAÇÃO DO MODELO PROPOSTO

### Tipos de Profissionais em Aplicações de KDD:

- Especialista em KDD
- Especialista no Domínio da Aplicação de KDD

### Tipos de Conhecimento em Aplicações de KDD:

- Conhecimento Independente do Domínio da Aplicação
- Conhecimento Dependente do Domínio da Aplicação
- Conhecimento em KDD Aplicado ao Domínio da Aplicação

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos do Problema:



### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos do Problema: Conjunto de Dados

- Estrutura tabular bidimensional
- Contém Casos (aspecto extensional)
- Contém Características (aspecto intensional)
- Esquema é o conjunto de características
- Não necessariamente um Data Warehouse

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos do Problema: Especialista no Domínio da Aplicação

- Conhecimento prévio sobre o domínio da aplicação (background knowledge)
- Consenso quando possível
- Dispõe de meta dados sobre o conjunto de dados
- Papel importante na formulação dos objetivos
- Papel importante na avaliação de resultados

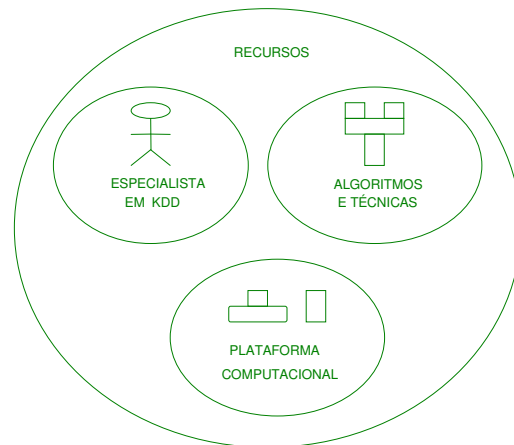
### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos do Problema: Objetivos da Aplicação

- Retratam **restrições e expectativas** acerca do modelo a ser gerado
- Em geral dependem da opinião dos especialistas no domínio da aplicação
- Nem sempre conseguem ser bem definidos no início do processo de KDD

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Recursos:



### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Recursos: Especialista em KDD

- Dispõe de conhecimento prévio sobre como realizar KDD
- Deve ter experiência neste tipo de trabalho técnico
- Interage com o especialista no domínio da aplicação
- Em geral pertence a uma equipe
- Responsável pela condução do processo de KDD

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Recursos: Algoritmos e Técnicas (Ferramentas)

- Referem-se aos **recursos de software** disponíveis para aplicação nas etapas do Processo de KDD.
- Algoritmos podem ser adaptados.
- Devem ser compatíveis com a plataforma computacional disponível.
- Uma mesma operação de KDD pode ser implementada por diversos destes recursos, de forma isolada ou conjugada.

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Recursos: Plataforma Computacional

- Referem-se aos **recursos de hardware** disponíveis para execução das Operações de KDD.
- São de grande relevância em Aplicações de KDD devido ao grande consumo de tempo em geral requerido.
- Mais memória e mais capacidade de processamento → maior dinâmica ao processo de KDD.
- Plataformas que viabilizem **computação paralela e distribuída** podem otimizar o desempenho de inúmeras Aplicações de KDD.

### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Resultados:



### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Resultados: Mod. de Conhecimento Descoberto

- Abstração de dados expressa em alguma linguagem obtida a partir da aplicação de KDD.
- Deve ser avaliado em relação ao cumprimento das expectativas formuladas nos objetivos da aplicação.
- Comparação entre modelos de conhecimento é muito comum.
- Conjugação de modelos pode ocorrer.

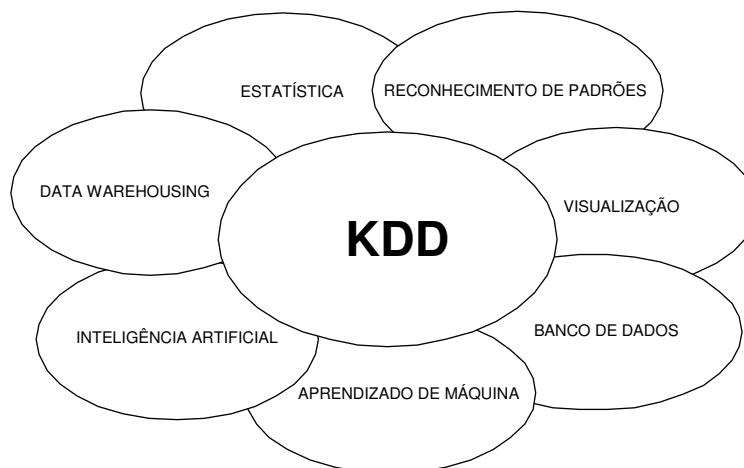
### 3. FORMALIZAÇÃO DO MODELO PROPOSTO

#### Elementos dos Resultados: Trilhas do Processo de KDD

- Estruturas de Dados que permitem armazenamento conciso de **fatos, ações e resultados intermediários** registrados ao longo do processo (históricos).
- O **conteúdo** destas estruturas pode ser utilizado como **Problema em Aplicações de KDD** cujo **objetivo** seja extrair conhecimento **sobre como realizar o Processo de KDD**.
- Podem viabilizar um processo de **aprendizado** para a **Máquina de Assistência** à Orientação do Processo de KDD.

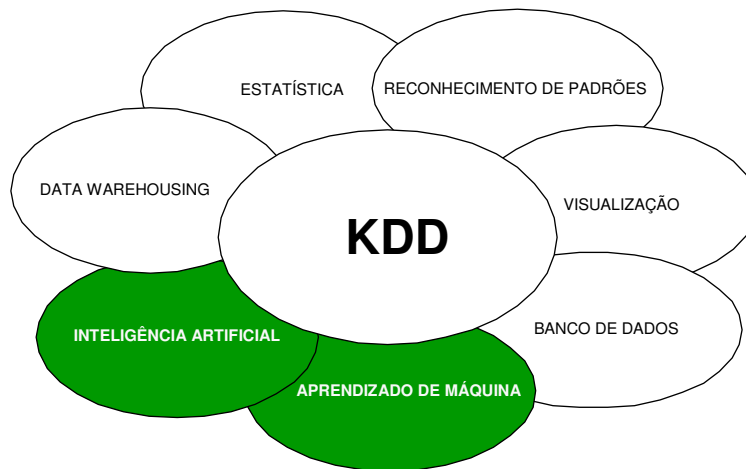
### O PROCESSO DE KDD: VISÃO GERAL

#### Áreas de Origem:



## O PROCESSO DE KDD: VISÃO GERAL

### Áreas de Origem:



## O PROCESSO DE KDD: VISÃO GERAL

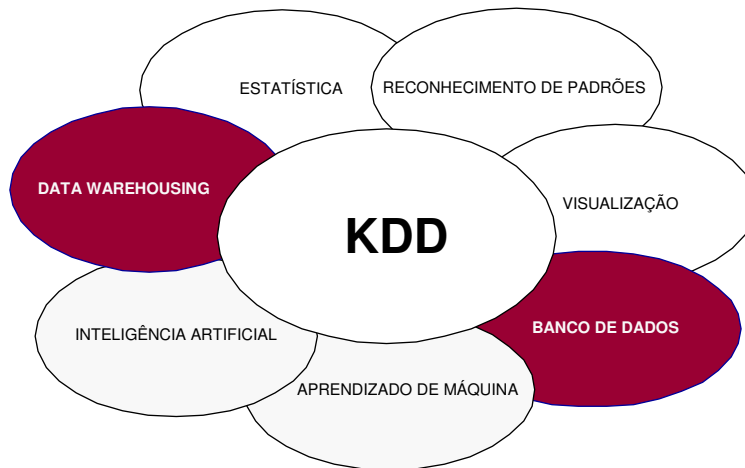
### Áreas de Origem:

#### Aprendizado de Máquina - Inteligência Artificial:

- Redes Neurais
- Algoritmos Genéticos
- Lógica Nebulosa
- Lógica Indutiva
- Árvores de Decisão

## O PROCESSO DE KDD: VISÃO GERAL

### Áreas de Origem:



## O PROCESSO DE KDD: VISÃO GERAL

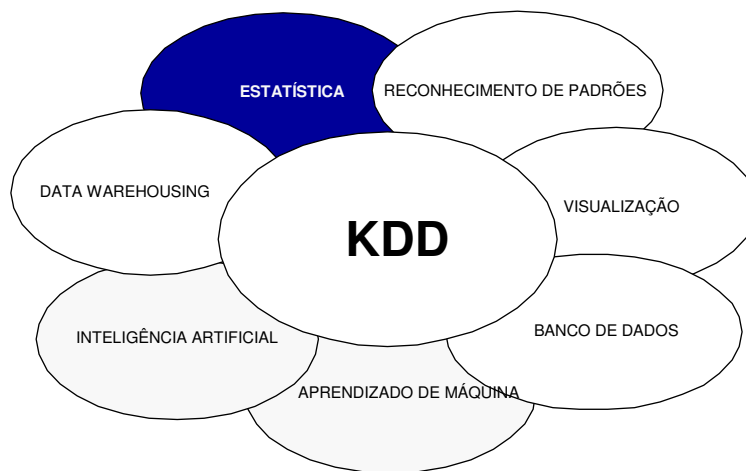
### Áreas de Origem:

#### Banco de Dados / Data Warehouses:

- **Data Warehousing**
- **SQL**
- **OLAP**
- **DMQL**
- **KMQL**

## O PROCESSO DE KDD: VISÃO GERAL

### Áreas de Origem:



## O PROCESSO DE KDD: VISÃO GERAL

### Áreas de Origem:

#### Estatística:

- **Classificadores Bayesianos**
- **Redes Bayesianos**
- **EDA - Exploratory Data Analysis**

## O PROCESSO DE KDD: VISÃO GERAL

Gerações da Mineração de Dados [Piatetsky-Shapiro, 2001]

- **1ª Geração**
  - Anos 90
  - Ferramentas de Pesquisa voltadas a uma única tarefa, sem suporte às demais etapas de KDD
  - Exemplos: c4.5, Rede Neural, Autoclass, etc...

## O PROCESSO DE KDD: VISÃO GERAL

Gerações da Mineração de Dados [Piatetsky-Shapiro, 2001]

- **2ª Geração**
  - Meados dos anos 90
  - Ferramentas chamadas “suites”: Pacote para aplicação com suporte ao pré-processamento e à visualização
  - Requerem conhecimento significativo da teoria estatística
  - Exemplos SPSS, Intelligent Miner, SAS, etc...

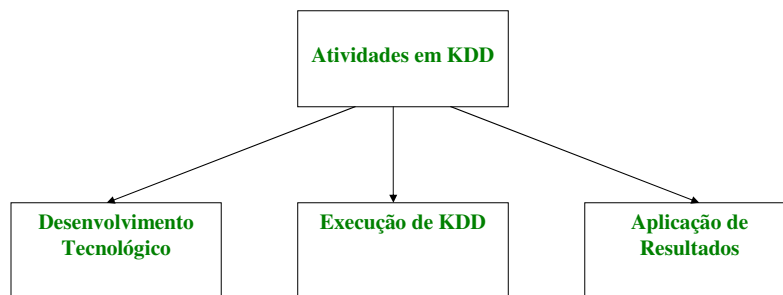
## O PROCESSO DE KDD: VISÃO GERAL

Gerações da Mineração de Dados [Piatetsky-Shapiro, 2001]

- 3ª Geração
  - Final dos anos 90
  - Soluções orientadas à resolução de problemas específicos em empresas
  - Possuem interfaces orientadas aos usuários
  - Escondem a complexidade da MD
  - Exemplos: Falcon (Detecção Fraude em Cartão)

## O PROCESSO DE KDD: VISÃO GERAL

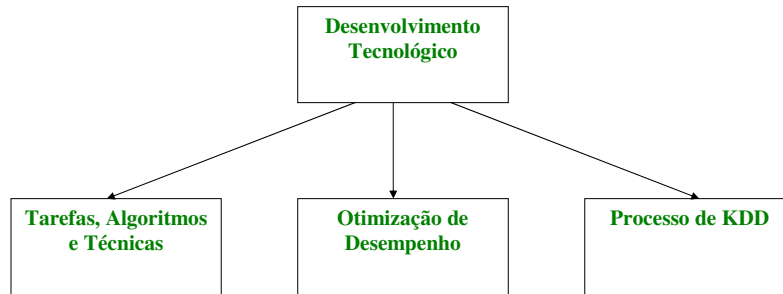
Uma Taxonomia:



[Goldschmidt et al., 2002a]

## O PROCESSO DE KDD: VISÃO GERAL

Uma Taxonomia:

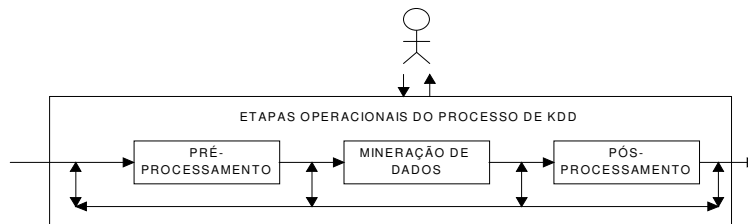


[Goldschmidt et al., 2002a]

## O PROCESSO DE KDD: VISÃO GERAL

Processo de Descoberta do Conhecimento em Bases de Dados

• Visão Pragmática [Goldschmidt et al., 2002a]:



• Operações e Métodos de KDD

## O PROCESSO DE KDD: VISÃO GERAL

Processo de Descoberta do Conhecimento em Bases de Dados

**Exemplos de Operações de KDD – Pré-Processamento:**

- **Redução de Dados: Vertical / Horizontal**
- **Limpeza: Remoção Inconsistências / Preenchimento Valores Ausentes**
- **Codificação: Categórica-Numérica / Numérica-Categórica**
- **Normalização de Dados: Linear / Máximo / Soma**
- **Partição dos Dados: Treino-Teste / K-Folders**

## O PROCESSO DE KDD: VISÃO GERAL

**Operações de Pré-Processamento: Seleção/Redução de Dados**

- **Horizontal: escolha de casos**
  - **Amostragem**
  - **Segmentação do BD**
- **Vertical: escolha de características**
  - **Atributos relevantes**
  - **Redução de dimensionalidade**

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Pré-Processamento : Limpeza

- Verificação de consistência entre informações
- Correção de erros
- Preenchimento de valores desconhecidos
- Eliminação de informações redundantes
- Eliminação de valores não pertencentes ao domínio

### Exemplo: Data de Nascimento

- Corretas nas seguradoras de vida;
- 30% a 40% em branco ou incorretas nos bancos;

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Pré-Processamento: Codificação

- Divide valores de atributos contínuos em intervalos codificados.

Ex: Renda                    [0, 1000]    → Faixa 1  
                                  [1001, 3000] → Faixa 2  
                                  [3001, 5000] → Faixa 3  
                                  etc...

- Representa valores de atributos categóricos por códigos.

Ex: Sexo                    M → 1            F → 0

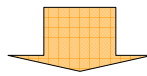
## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Pré-Processamento: Enriquecimento

Ex: Perfil do Cliente

**Atributos:**

- Renda
- Despesas
- Tipo de Residência
- Bairro de Residência



**Atributos:**

- Renda
- Despesas
- Tipo de Residência
- Bairro de Residência
- Valor Médio Imóvel

## O PROCESSO DE KDD: VISÃO GERAL

### Processo de Descoberta do Conhecimento em Bases de Dados

#### Exemplos de Operações de KDD – Mineração de Dados:

- Classificação
- Associação
- Sequências
- Previsão de Séries Temporais
- Detecção de Desvios
- Clustering

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Mineração de Dados: Classificação

#### Ex de Aplicação:

Sexo	País	Idade	Comprar
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
M	França	55	Não

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Mineração de Dados: Classificação

#### Ex de Aplicação:

#### Algumas Regras:

- Se (País = Alemanha) Então Comprar = Não
- Se (País = Inglaterra) Então Comprar = Sim
- Se (País = França e Idade  $\leq$  25) Então Comprar = Sim
- Se (País = França e Idade  $>$  25) Então Comprar = Não

## O PROCESSO DE KDD: VISÃO GERAL

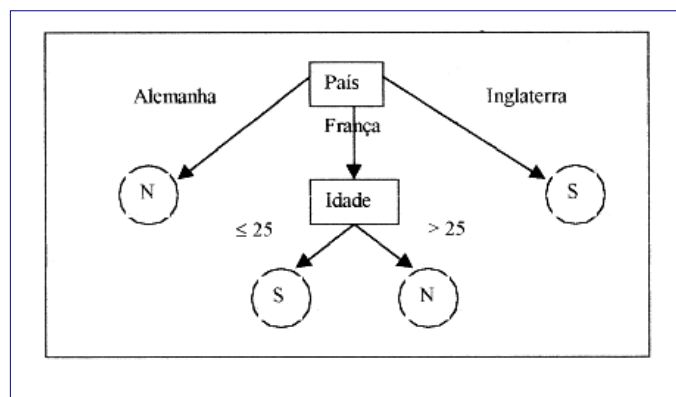
Processo de Descoberta do Conhecimento em Bases de Dados

Exemplos de Operações de KDD – Pós-Processamento:

- **Análise de Modelos**
- **Corte de Regras / Poda de Árvores (Tree Pruning)**
- **Visualização de Gráficos**
- **Organização de Resultados**
- **Avaliação do Modelo de Conhecimento Gerado**
- **Conversão de Representações**

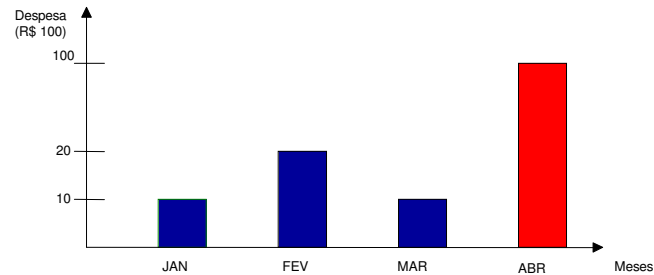
## O PROCESSO DE KDD: VISÃO GERAL

Operações de Pós-Processamento: Análise de Modelos  
Exemplo (Árvore de Decisão):



## O PROCESSO DE KDD: VISÃO GERAL

**Operações de Pós-Processamento: Visualização de Gráficos**  
**Exemplo:**



## O PROCESSO DE KDD: VISÃO GERAL

**Operações de Pós-Processamento: Organização de Resultados**  
**Exemplo:**

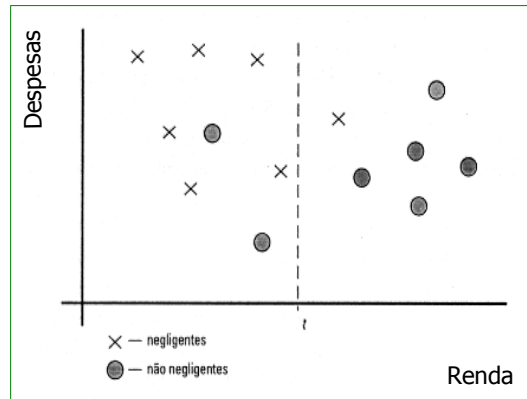
- Se (País = Alemanha) Então Comprar = Não
- Se (País = Inglaterra) Então Comprar = Sim
- Se (País = França e Idade  $\leq$  25) Então Comprar = Sim
- Se (País = França e Idade  $>$  25) Então Comprar = Não

**Importância desta operação para lidar com grandes volumes de resultados → Meta Mineração de Dados**

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Pós-Processamento: Avaliação do Modelo

#### Exemplo:



#### Modelo de Conhecimento:

Se renda > R\$ t

Então Crédito = SIM

#### Avaliação do Modelo:

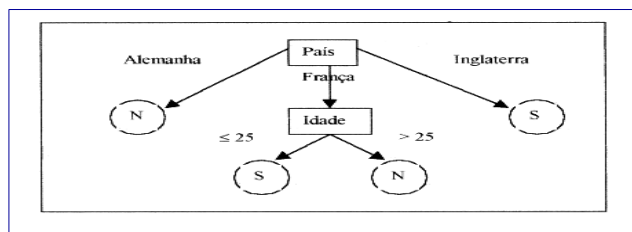
Interpretável

Precisão:  $11/14 = 78,6\%$

## O PROCESSO DE KDD: VISÃO GERAL

### Operações de Pós-Processamento: Conversão de Representações

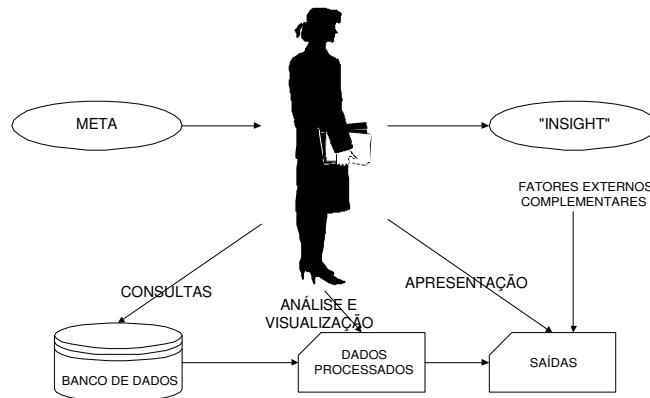
#### Exemplo:



- Se (País = Alemanha) Então Comprar = Não
- Se (País = Inglaterra) Então Comprar = Sim
- Se (País = França e Idade  $\leq 25$ ) Então Comprar = Sim
- Se (País = França e Idade  $> 25$ ) Então Comprar = Não

## O PROCESSO DE KDD: VISÃO GERAL

### A importância do usuário no processo KDD



## O PROCESSO DE KDD: VISÃO GERAL

### Macro-Objetivos da Mineração de Dados [Zaki, 2002]:

- **Predição: Histórico x Novas Situações**
- **Descrição: Modelo Descritivo do Conhecimento**

### Orientação das Tarefas de Mineração de Dados [Zaki, 2002]:

- **Para Verificação: Hipótese Postulada x Validação**
- **Para Descoberta: Extração de novos conhecimentos**

## O PROCESSO DE KDD: VISÃO GERAL

### Tarefas de Mineração de Dados:

- Associação
- Descoberta de Sequências
- Classificação / Regressão
- Clusterização
- Detecção de Desvios
- Sumarização / Descrição
- Mineração de Textos
- Mineração na Web
- Dentre outras ...

## O PROCESSO DE KDD: VISÃO GERAL

### Técnicas de Mineração de Dados:

- Tradicionais
- Específicas
- Híbridas

## O PROCESSO DE KDD: VISÃO GERAL

### Técnicas de MD Tradicionais

- Baseadas em tecnologias consagradas fora do contexto da MD
- Exemplos: Redes Neurais, Algoritmos Genéticos, Árvores de Decisão, Estatística, SQL, etc...

## O PROCESSO DE KDD: VISÃO GERAL

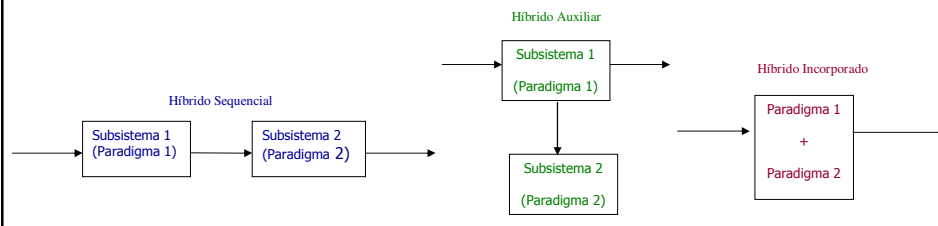
### Técnicas de MD Específicas

- Desenvolvidas especificamente para aplicação em Mineração de Dados.
- Exemplos: Apriori, GSP, ParMaxEclat, ParMaxClique, DMQL, etc..

## O PROCESSO DE KDD: VISÃO GERAL

### Técnicas de MD Híbridas

- Utilizam combinações entre as técnicas tradicionais e as técnicas específicas
- Exemplos: Apriori em PL/SQL
- Formas de associação entre duas técnicas para a construção de sistemas híbridos (Souza, 1999):



## O PROCESSO DE KDD: VISÃO GERAL

### Considerações Técnicas quanto à Realização de MD - Algumas Diretrizes:

- Disponibilidade de dados suficientes
- Suporte a grandes volumes de dados
- Verificação da relevância dos atributos
- Busca por baixo nível de ruído
- Utilização de conhecimento prévio

## O PROCESSO DE KDD: VISÃO GERAL

Considerações Técnicas quanto à Realização de MD -  
Algumas Diretrizes:

- **Suporte a vários recursos de aprendizado (aprendizado híbrido)**
- **Buscar integração com DSS - Decision Support Systems**
- **Utilização de plataformas com arquitetura expansível**
- **Suporte a Bancos de Dados Heterogêneos**

## O PROCESSO DE KDD: VISÃO GERAL

Considerações Técnicas quanto à Realização de MD -  
Algumas Diretrizes:

- **Buscar estabelecer Data Warehouses**
- **Disponibilidades de recursos para limpeza de dados**
- **Facilidades de codificação dinâmica de atributos**

## O PROCESSO DE KDD: VISÃO GERAL

Considerações Técnicas quanto à Realização de MD - Check List Inicial:

- **Fazer um Levantamento do Hardware e Software existente.**
- **Fazer uma lista de necessidades.**
  - Qual o propósito do KDD?
  - Quais são os critérios de sucesso do KDD?
  - Como será mensurado esse sucesso?
  - Bancos de Dados, Redes, Aplicações, Servidores, etc.
- **Avaliar a qualidade dos dados disponíveis.**
  - Para que propósito foi coletado?

## O PROCESSO DE KDD: VISÃO GERAL

Considerações Técnicas quanto à Realização de MD - Check List Inicial:

- **Fazer um inventário dos Banco de Dados disponíveis.**
  - Internamente e Externamente
- **Verificar a existência de um Data Warehouse.**
  - Que tipo de dados estão disponíveis
  - Podemos verificar os detalhes dos dados operacionais?
- **Formular o conhecimento que a organização necessita.**

## O PROCESSO DE KDD: VISÃO GERAL

### Considerações Técnicas quanto à Realização de MD - Check List Inicial:

- **Identificar os grupos de engenheiros de conhecimento ou os grupos de decisão que aplicarão os resultados.**
  - Que tipo de decisões precisam ser tomadas?
  - Quais padrões são úteis?
- **Analisar se o conhecimento encontrado é realmente útil para a organização.**
- **Listar os Processos e as Transformações que serão aplicados aos BD's antes que esses possam ser utilizados no KDD.**

## O PROCESSO DE KDD: VISÃO GERAL

### KDD: Um Resumo dos Principais Aspectos Abordados

