

Descoberta de Conhecimento em Bancos de Dados

Prof. Me. Ricardo Ávila
ricardo.avila@outlook.com.br

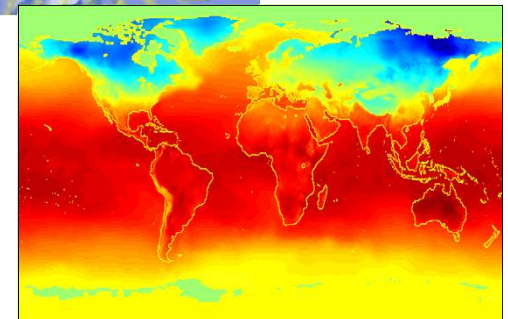
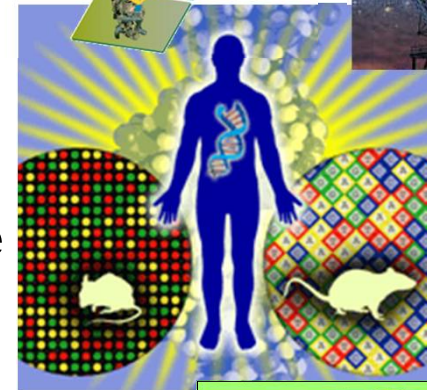
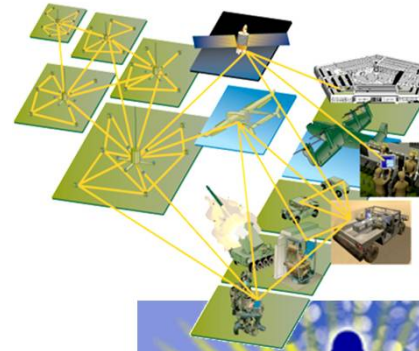
Mineração de Dados do Ponto de Vista Comercial

- Quantidades gigantescas de dados são coletados e armazenados em empresas, corporações, etc
 - Dados de comércio eletrônico,
 - Dados de navegação na internet
 - Dados de compras de clientes em grandes lojas de departamentos, supermercados,
 - Dados de transações bancárias, ou de cartão de crédito
- Computadores mais baratos e mais potentes
- Pressão da Competição



Mineração de Dados do ponto de vista científico

- Dados coletados e armazenados a velocidades enormes (GB/hora)
 - Sensores remotos em satélites
 - Telescópios
 - Microarrays gerando dados de expressões de genes
 - Simulações científicas gerando terabytes de dados.
- Técnicas tradicionais não apropriadas para analisar tais dados:
 - ruídos e grande dimensionalidade



EVOLUÇÃO DA TECNOLOGIA DE BD

❑ 1960s:

- ❑ Coleção de dados, criação do banco de dados, redes e SGBD

❑ 1970s:

- ❑ Modelo de dados relacional, implementação do DBMS relacional

❑ 1980s:

- ❑ SGBDR, modelos de dados avançados (relacional estendido, OO, dedutivo, etc.) e SGBD-ORorientado a aplicação (espacial, científica, engenharia, etc.)

❑ 1990s—2000s:

- ❑ **Data mining e data warehousing**, bancos de dados multimídia e banco de dados Web.

INTRODUÇÃO



O que é **Data Mining**?

Produzir conhecimento novo escondido em grandes bases de dados

A coleta de **dados** (transações bancárias, registros de compras, perfil de uso da internet, integração das informações de diversos sistemas, código de barras, via sensores remotos, satélites, documentos web, etc..

- tem atingido grandes proporções → acarretou problema na área do conhecimento
- novo ramo do conhecimento (KDD – *Knowledge Discovery in Databases*), o qual visa **otimizar** e **automatizar** o processo de descoberta das **tendências** e dos **padrões** contidos nos dados, potencialmente **úteis** e **interpretáveis**.

Definição

É a aplicação de técnicas **estatísticas** e de **inteligência artificial** em grandes quantidades de dados, para descobrir relações e padrões relevantes entre os dados.

É o processo de **construir modelos** baseados em um conjunto de dados, que nos permita fazer **predições**, **controlar** ou **melhorar** algum processo.

Definição

É a exploração e a análise, por meio automático ou semi-automático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativas." (Berry e Linoff, 1997).

"É o processo de reconhecimento de padrões válidos ou não, existentes nos dados armazenados em um banco de dados." [Fayyad, Piatetsky-Shapiro & Smyth, 1995]

Padrão

Um evento ou combinações de eventos numa base de dados que ocorre com mais frequência do que esperamos. Significa que sua ocorrência é significativamente diferente do que se esperaria devido ao acaso.

Padrões são guiados pelos dados e geralmente refletem os próprios dados;

- Exemplo: se “salário < T, então a pessoa não efetuou o pagamento” pode ser um padrão para uma escolha adequada de T.

Padrão

Padrões precisam ser:

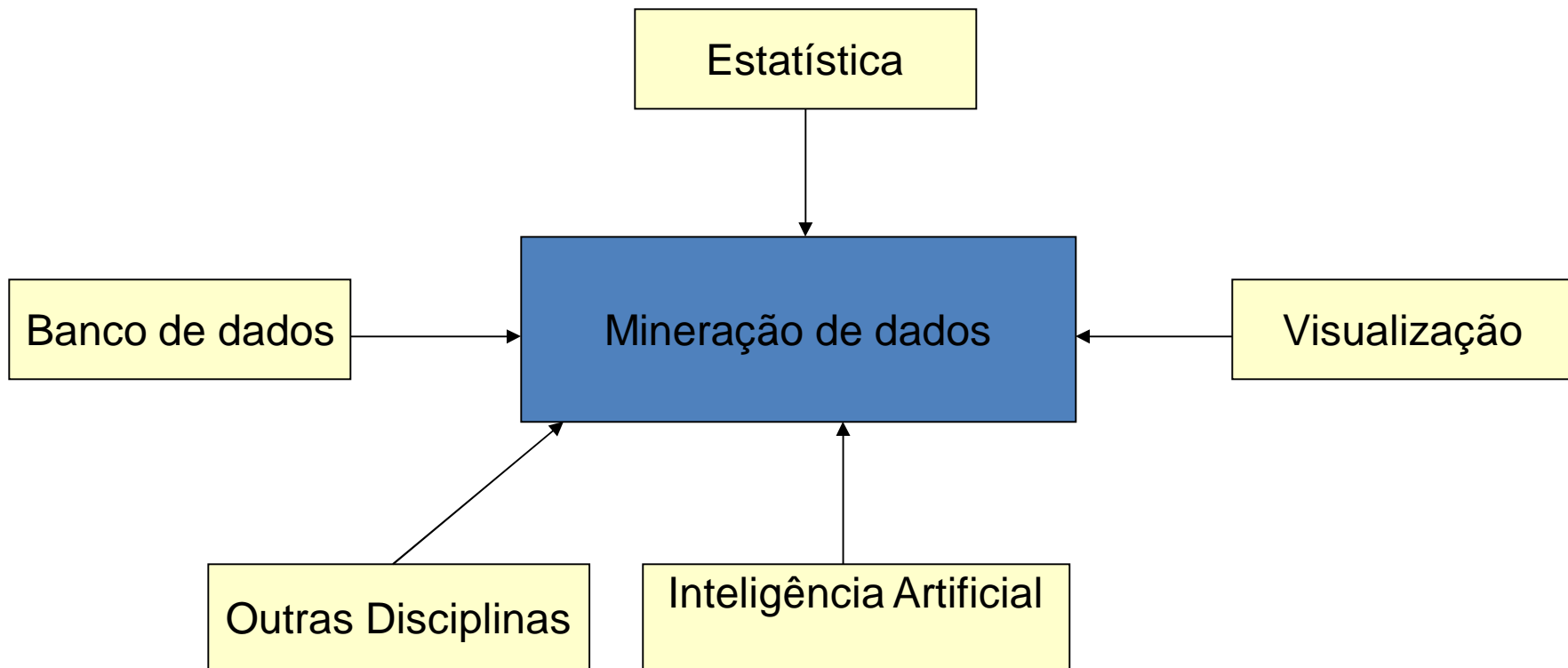
- Úteis:** Representa o grau de utilidade de um padrão, isto é, até que ponto a descoberta ajuda a responder os objetivos inerentes ao processo de KDD.
- Interpretáveis:** Um dos objetivos do KDD é gerar padrões compreensíveis para os analistas na perspectiva de um melhor entendimento dos dados.
- Válidos:** Para dados novos ou arquivo de teste com certo grau de certeza.
- Novo, desconhecido:** Especialmente no sentido de interessante, não usual.

Extração de conhecimento

A extração de conhecimento de bases de dados é um processo complexo e, ainda hoje, muito dependente da experiência e do trabalho do analista (formulação do problema, preparação dos dados, análises e interpretações dos resultados, avaliações).

É atribuído às máquinas a responsabilidade de manipular conjuntos de dados, procurando padrões que satisfazem os problemas apresentados, mas a interpretação depende do analista.

Data mining é uma área interdisciplinar, envolvendo banco de dados, técnicas de estatísticas e inteligência artificial como redes neurais, aprendizado de máquina, reconhecimento de padrões e visualização de dados.



Extração de conhecimento e IA

A descoberta de conhecimento surgiu como tópico de estudo na Inteligência Artificial (IA).

A IA é um ramo da ciência da computação voltada ao desenvolvimento de métodos e técnicas que permitam a construção de sistemas computacionais que exibam características associadas ao comportamento e inteligência humana (aprender, raciocinar, inferir novos conhecimentos, entender linguagens, resolver problemas, etc.).

Extração de conhecimento e IA

A descoberta de conhecimento, surgiu como tópico de estudo na IA e tem como principais objetivos:

- além da tarefa de descobrir conhecimento, descobrir formas de aquisição, armazenamento e representação deste conhecimento.

A descoberta de conhecimento, tem por função transformar dados em informação (por interpretação), derivar novas informações das existentes (por elaboração) e adquirir conhecimento novo (pelo aprendizado).

Extração de conhecimento e IA

- Os **dados** são a matéria prima e não tem caráter informativo.
- A **informação** é a seleção e a organização dos dados em um determinado contexto, para um fim determinado (é a interpretação dos dados),
- O **conhecimento** é formado por informações (na forma de fatos, regras ou heurísticas) que permitam derivar outras informações novas, por elaboração e por aprendizado.

—Exemplo:

- dt Nascimento= 02/02/2002 (dado)
- idade=10 anos é infantil (informação)
- se idade < 20 então riscoInfarto=baixo

Extração de conhecimento e IA

O conhecimento é subjetivo e depende muito do usuário, pois o que pode ser adquirido por um pode não ser por outro (por não interessar, por já ter sido adquirido ou porque o usuário não tem condições de adquiri-lo).

Por este motivo, o conhecimento deve ser adquirido de forma construtivista, onde o processo de aquisição é guiado por hipóteses, num processo interativo homem-máquina.

Extração de conhecimento e IA

Com a Internet, emails, chats, páginas WEB, etc, outras fontes de dados e informação de forma não estruturada começaram a surgir e ser armazenadas

Isso ocasionou o surgimento de um novo ramo de pesquisas na área de descoberta de conhecimento: o ramo da Descoberta de Conhecimento em Textos (KDT).

Extração de conhecimento e IA

Por outro lado, o KDT é a evolução natural da área de Recuperação de Informação (RI), já que ao invés de localizar e acessar as informações procuradas e deixar que o usuário mesmo procure o que lhe interessa, a nova área esta relacionada ao entendimento, resumo e tratamento de informações (transformando-as em conhecimento útil e aplicável).

APLICAÇÕES

SEGMENTAÇÃO DE MERCADO	IDENTIFICA AS CARACTERÍSTICAS COMUNS DE CLIENTES QUE COMPRAM OS MESMOS PRODUTOS DE UMA EMPRESA
PERDA DE CLIENTES	PREDIZ QUAIS CLIENTES PROVAVELMENTE DEIXARÃO A EMPRESA PARA UM CONCORRENTE
DETECÇÃO DE FRAUDE	IDENTIFICA QUAIS TRANSAÇÕES ESTÃO MAIS SUJEITAS A FRAUDE
MARKETING DIRETO	IDENTIFICA QUAIS PROSPECTOS DEVERIAM SER INCLUÍDOS NA MALA DIRETA PARA OBTENÇÃO DE ALTA TAXA DE RETORNO
MARKETING INTERATIVO	PREDIZ O QUE CADA INDIVÍDUO QUE ACESSA O SITE ESTÁ MAIS INTERESSADO EM VER
ANÁLISE “MARKET BASKET”	IDENTIFICA QUAIS PRODUTOS SÃO COMUMENTE COMPRADOS EM CONJUNTO
ANÁLISE DE TENDÊNCIAS	REVELA AS DIFERENÇAS ENTRE UM TÍPICO CLIENTE DE UM MÊS EM RELAÇÃO AOS MESES ANTERIORES

Mais Aplicações

- Qual o perfil do cliente que consome mais ?
- Que produtos são comprados conjuntamente ? E em sequência ?
- Meu site web tem uma boa estrutura ?
- Como as chuvas, variação de temperatura, aplicação de pesticidas afetam as colheitas ?
- Existe uma relação entre o aquecimento global e a frequência e intensidade das perturbações no ecossistema tais como secas, furacões, enchentes ?

Aplicações

O governo dos EUA se utiliza do *data mining* já há bastante tempo para identificar padrões de transferências de fundos internacionais que se parecem com lavagem de dinheiro do narcotráfico. *Data mining* usado para identificar *fraudes*.

Mineração de Dados - Por que ?

- Frequentemente existe informação “escondida” nos dados que não é evidente de ser encontrada utilizando linguagens de consulta tradicionais.
- Analistas humanos podem levar semanas para correlacionar e descobrir alguma informação útil dentro de uma grande massa de dados.
- Boa parte dos dados nunca é analisado: “cemitério” de dados.

Mineração de Dados: Por que ?

- Técnicas de Mineração podem ajudar **analistas**:
 - Entender e prever as necessidades dos clientes
 - Descobrir fraudes
 - Descobrir perfis de comportamento de clientes
- Técnicas de Mineração podem ajudar **cientistas**:
 - Classificar e segmentar dados
 - Formular hipóteses

Mineração de Dados: O que é ?

● Não

1. Fazer uma consulta no Google sobre “Data Mining ”
2. Procurar um nome numa lista telefônica
3. Fazer uma consulta SQL a um banco de dados.

● Sim

1. Agrupar documentos similares retornados pelo Google de acordo com seu contexto.
2. Descobrir se certos nomes aparecem com mais frequência em determinadas regiões da cidade (periferia, centro, bairros abastados,...)

OBSERVANDO E APRENDENDO

Exemplo: um proprietário de uma pequena loja de vinhos conhece tudo sobre vinhos, por exemplo, o tipo de uva, a região onde a uva foi cultivada, o clima, o solo, a altitude dos parreirais, aroma, sabor, cor, o processo de fabricação. Os clientes gostam de visitar sua loja pois, também, aprendem muito sobre vinhos. Porém, só isto não basta, o proprietário precisa conhecê-los, como por exemplo, qual o tipo de vinho que o cliente gosta? Qual o poder aquisitivo? Assim, ele poderá dar um atendimento diferenciado (um a um) aos clientes. Temos, portanto, duas necessidades:

conhecimento e aprendizado

Uma pequena loja \Rightarrow poucos clientes \Rightarrow atendimento personalizado

Uma grande empresa \Rightarrow milhares de clientes \Rightarrow dificuldade em dar um atendimento dedicado

OBSERVANDO E APRENDENDO

Qual a tendência nos dias atuais?

Ter clientes leais, através de um relacionamento pessoal, ***um-para-um***, entre a empresa e o cliente.

Dentro desta tendência, as empresas desejam identificar os clientes cujos valores e necessidades sejam compatíveis com o uso prolongado de seus produtos, e nos quais é válido o risco de investir em promoções com descontos, pacotes, brindes e outras formas de criar essa relação pessoal.

Esta mudança de foco requer mudanças em toda a empresa, mas principalmente nos setores de marketing, vendas e atendimento ao cliente.



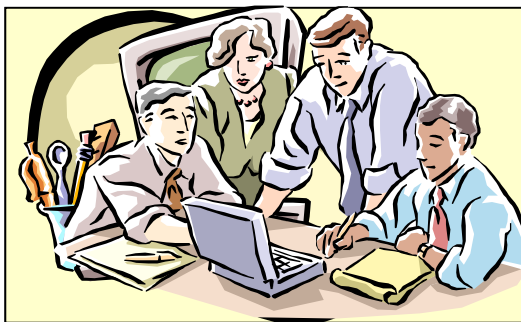
Memória e Inteligência



Na pequena empresa, o proprietário com sua inteligência e memória aprende, conhece o cliente.



Data Warehouse



Data Mining



Na grande empresa, a memória é o
data warehouse, enquanto a
inteligência é o *data mining*

Data Warehouse: a memória da empresa

Para criar relações um-para-um em uma grande empresa, o proprietário humano precisa ser substituído por uma máquina capaz de tratar grandes números, o computador. A memória do proprietário é substituída por um grande banco de dados denominado de **Data Warehouse**, enquanto a capacidade de **aprendizado** é substituída por técnicas de inteligência artificial e estatística, genericamente denominadas de **Data Mining**.

Diariamente são gerados e armazenados dados como: o número do telefone, a duração da chamada telefônica, o número do cartão de crédito, o endereço da entrega, **o produto escolhido**, renda do consumidor, escolaridade do consumidor, gasto com lazer, etc.

Certamente, só armazenar dados não significa aprender sobre o cliente.

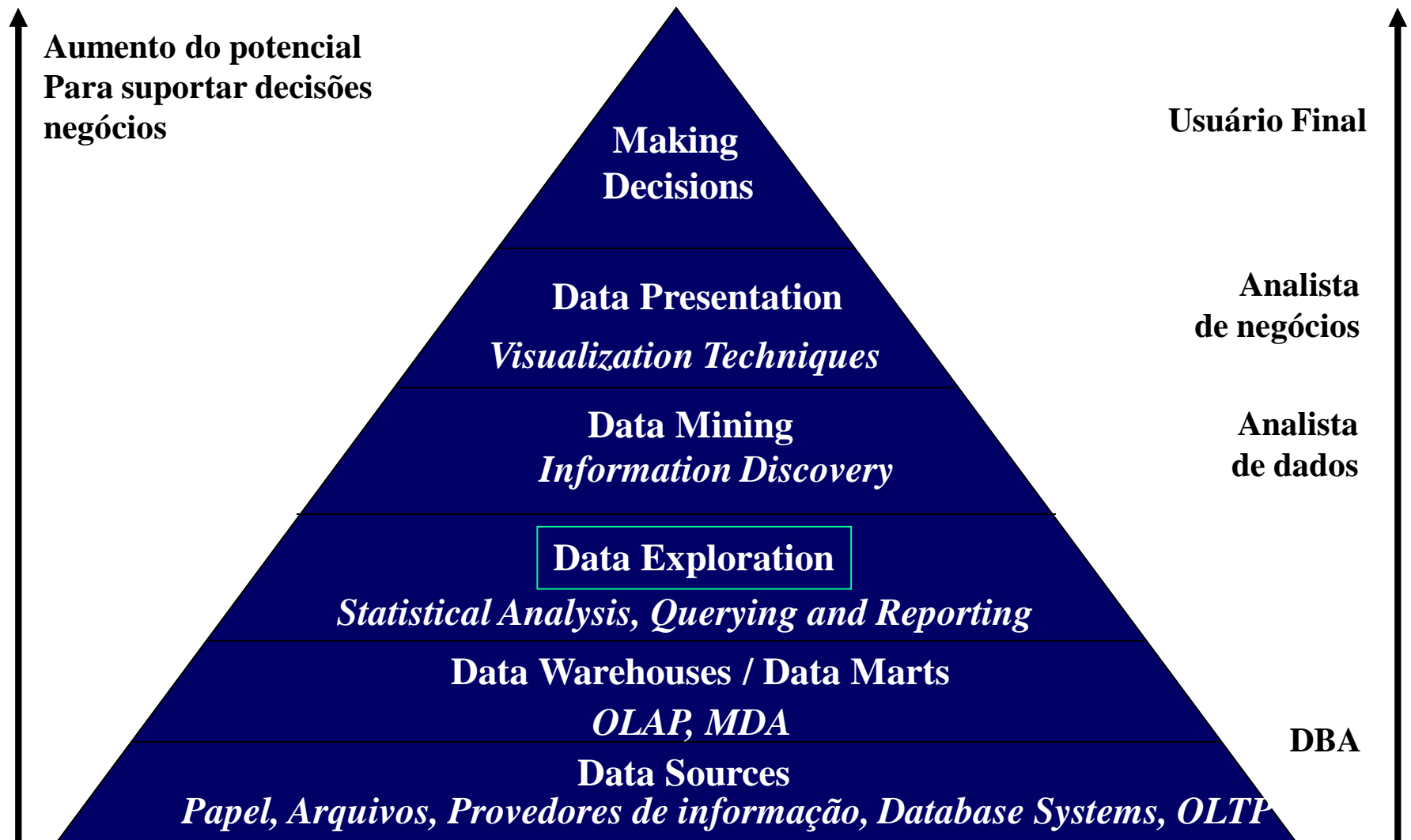
Data Mining: a inteligência da empresa

Para o aprendizado ocorrer, uma série de informações de diferentes formatos e fontes precisa ser organizada de maneira consistente na grande memória empresarial.

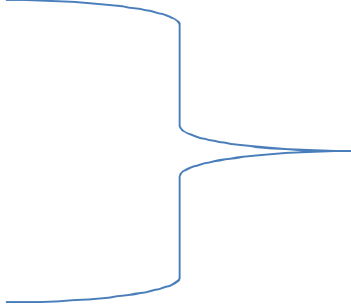
Após isto, métodos de análise estatística e inteligência artificial precisam ser aplicados sobre os dados.

A mineração dos dados consiste em **descobrir relações entre produtos, classificar consumidores, prever vendas, localizar áreas geográficas potencialmente lucrativas para novas filiais, inferir necessidades, entre outras.**

Data Mining and BI



Etapas do Processo de KDD (DCBD) – (Fayyad 1996)

- ☐ Seleção
 - ☐ Limpeza dos Dados
 - ☐ Transformação
 - ☐ Mineração
 - ☐ Avaliação ou Pós-Processamento
 - ☐ Visualização dos Resultados
 - ☐ Conhecimento
- 
- Pré-processamento ou
Preparação dos dados

Mineração: Etapa central do processo de **Descoberta de Conhecimento**

Conhecimento

Análise do Resultado

Mineração

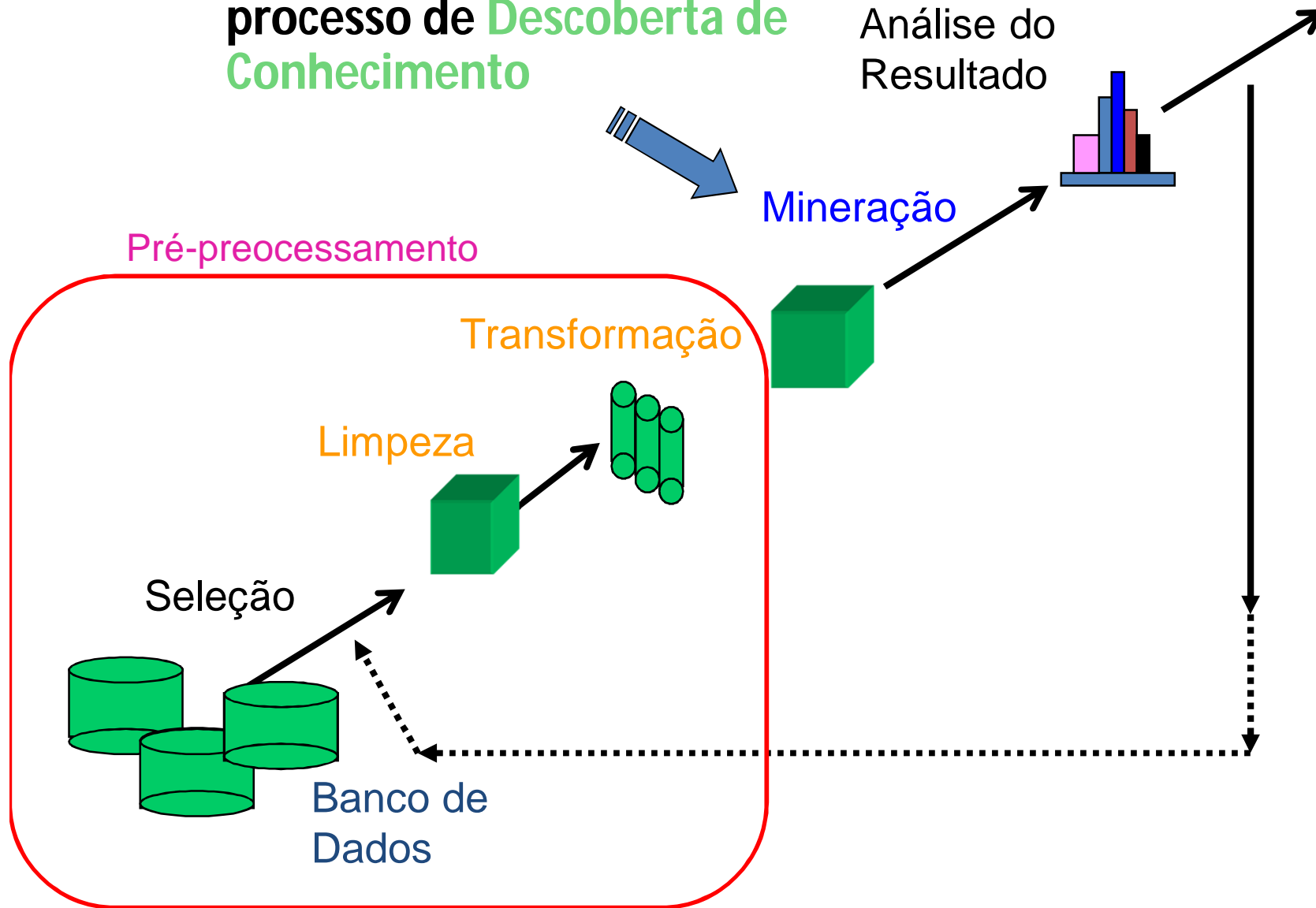
Pré-processamento

Transformação

Limpeza

Seleção

Banco de Dados



METODOLOGIA PARA DCBD

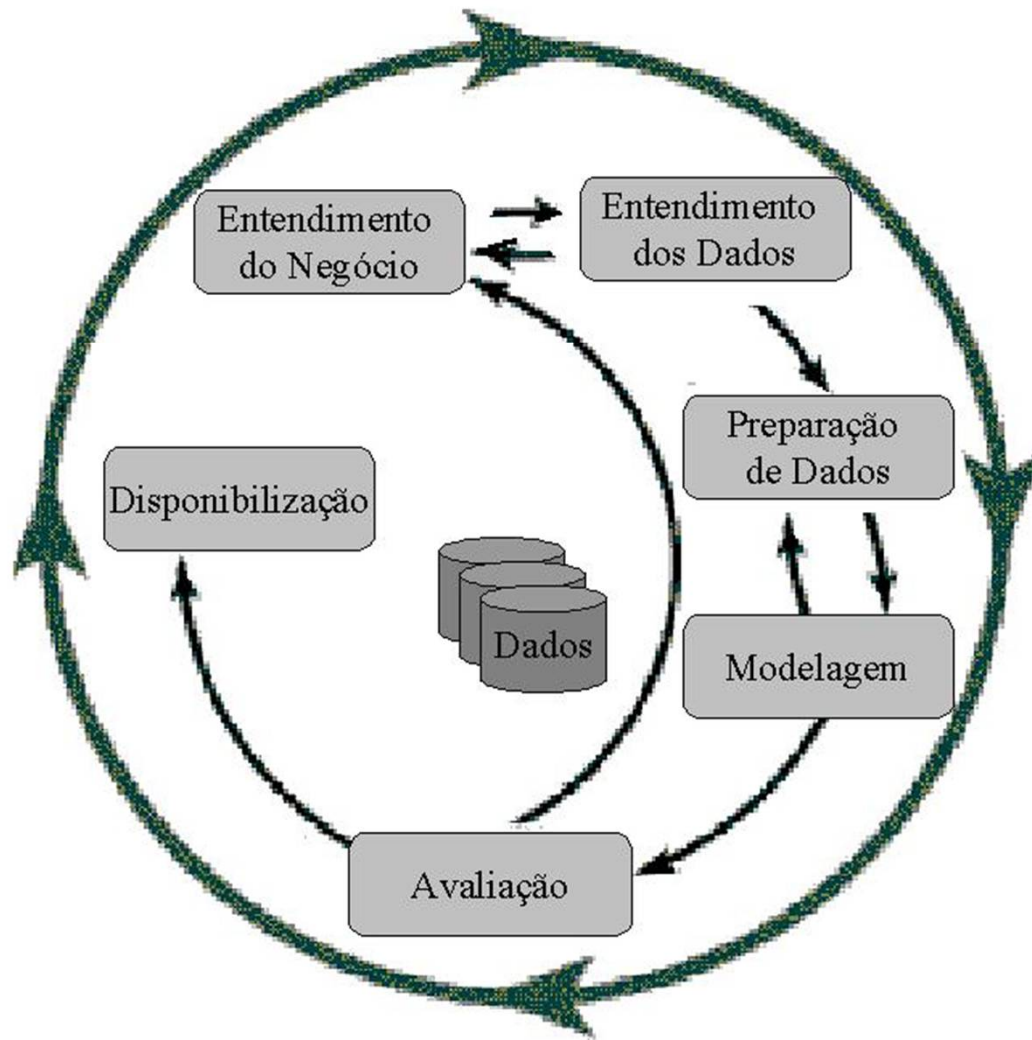
Metodologia CRISP-DM

- CRISP-DM = Cross – Industry Standard Process for Data Mining

(projeto ESPRIT com vários parceiros industriais)

- Geral - não se restringe a ferramenta ou tecnologia específica

Fases do CRISP-DM



Entendimento do Negócio (ou do problema)

Identificação dos objetivos do usuário sob o ponto de vista de DCBD e preparação de um plano inicial

- Determinar os objetivos
- Avaliar a situação: disponibilidade de recursos, limitações, etc.
- Determinar os objetivos da DC: objetivo, tipo de problema (classificação, *clustering*,...), critérios para avaliação do modelo.
- Produzir plano do projeto

Exemplo

Objetivo: uma empresa de fornecimento de água (como o a CASAN) deseja diminuir seus custos operacionais

Como: diminuindo o consumo de energia elétrica

Exemplo

Objetivo: uma empresa de fornecimento de água (como a CASAN) deseja diminuir seus custos operacionais

Como: diminuindo o consumo de energia elétrica

Predição de consumo de água:

Dados:

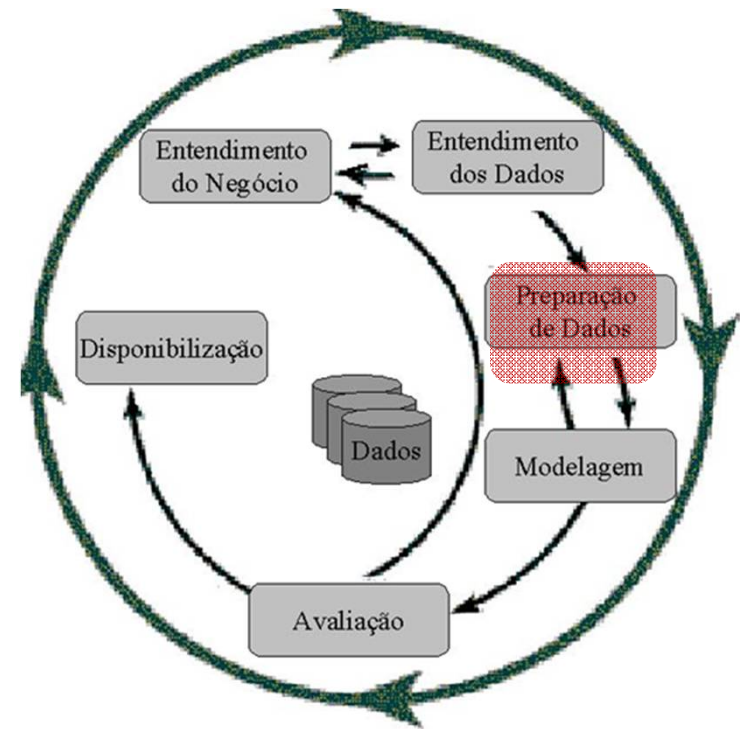
- consumo diário de água
- aspectos climáticos (temperatura, umidade do ar, ...)
- dia da semana e do mês, feriado, férias, ...

Objetivo da mineração: prever o consumo de água de forma a minimizar o bombeamento em horários mais caros

Entendimento dos Dados

A partir da coleta inicial, explorar os dados, verificando suas propriedades e qualidade

- Coletar dados iniciais
- Descrever os dados
- Explorar os dados
 - analisar a descrição dos dados
 - usar técnicas de visualização
- Verificar a qualidade dos dados



O que são dados?

- Uma coleção de objetos e seus atributos
- Um atributo é uma propriedade ou característica de um objeto
 - Exemplos: a cor dos olhos de uma pessoa, a sua temperatura, etc.
 - Atributos também são conhecidos como variáveis, campos, características
- Uma coleção de atributos descreve um objeto
 - Objetos também são conhecidos como registro, ponto, caso, exemplo, entidade, instância

Atributos

Objetos



<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Atributos discretos e contínuos

- Atributo discreto
 - Tem um número finito de valores
 - Exemplos: CEP, CPF, o conjunto de palavras em um documento
 - Frequentemente representado como uma variável inteira.
- Atributo contínuo
 - Tem um número real como valor do atributo
 - Exemplos: temperatura, altura, ou peso.
 - Na prática, valores reais só podem ser medidos e representados usando um conjunto finito de dígitos.
 - Variáveis contínuas são tipicamente representadas como variáveis de ponto-flutuante.

Tipos de conjuntos de dados

- **Registros**

- Matriz de dados
- Dados de documentos
- Dados de transações

- **Grafos**

- World Wide Web
- Estruturas moleculares

- **Ordenados**

- Dados temporais e espaço-temporais
- Dados seqüenciais
- Dados de seqüências genéticas

Descrição dos dados

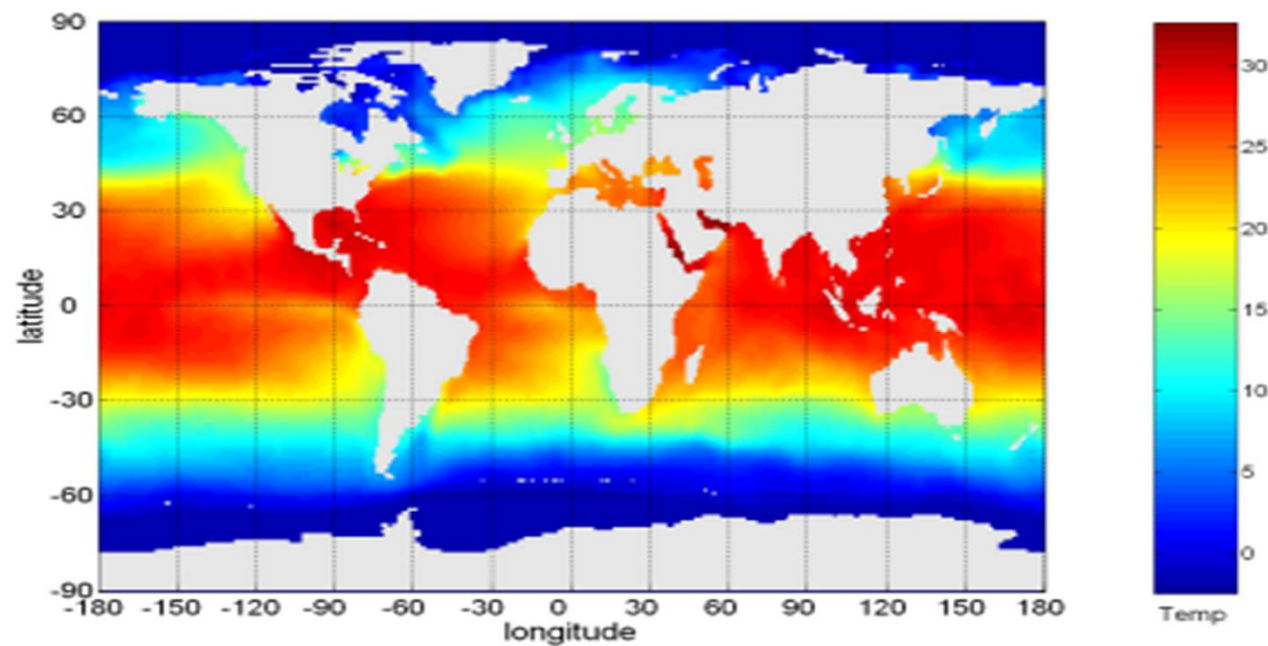
- Descrever os dados
 - Número de atributos e instâncias em cada arquivo
 - Tipos e faixas de valores dos atributos
 - Significado de cada atributo e sua importância para o objetivo
 - Estatísticas básicas para alguns atributos (média, DP, máximo, mínimo, percentil, frequência, moda, etc.)
 - Relações entre os atributos-chave

Exploração dos dados: Visualização

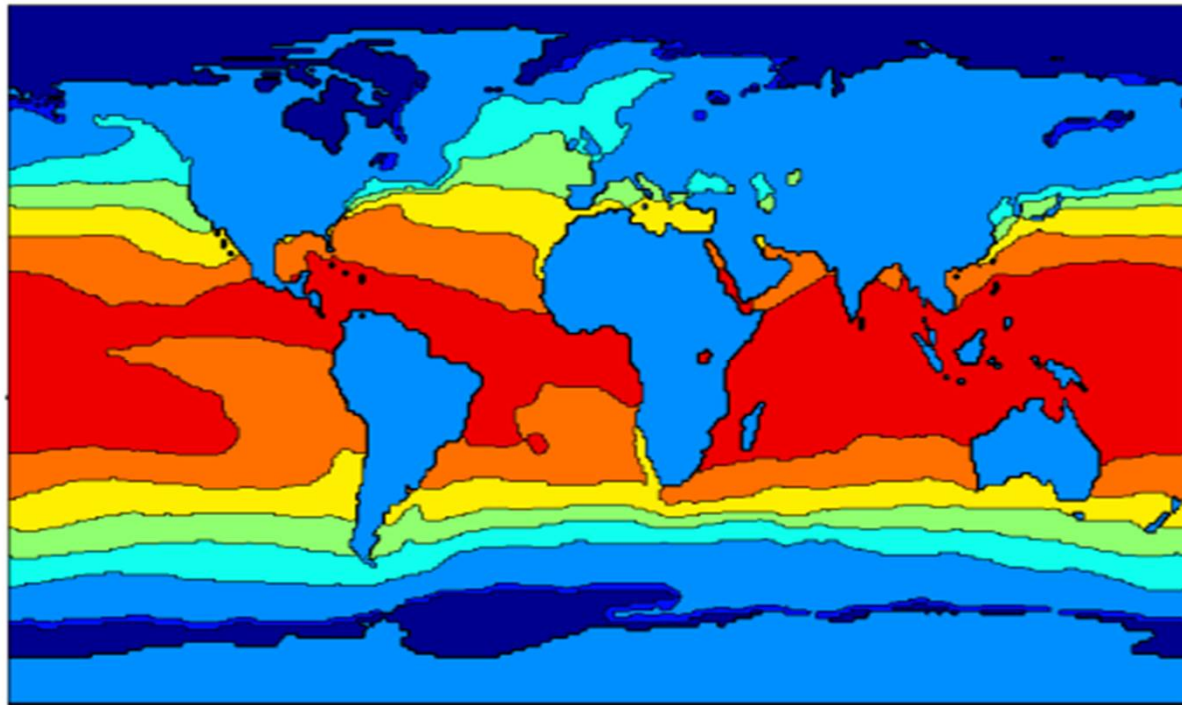
- Visualização é a conversão dos dados para um formato visual ou tabular de forma que características dos dados e da relação entre elementos dos dados possam ser percebidos visualmente
- Visualização é uma das mais poderosas técnicas para a exploração dos dados:
 - humanos têm muita habilidade de analisar grandes quantidades de informação apresentadas sob forma visual
 - podem detectar padrões gerais e tendências
 - podem detectar exceções e padrões não usuais

Exemplo: temperatura da superfície dos oceanos em julho de 1982

Dezenas de milhares de pontos estão sumarizados em uma simples figura

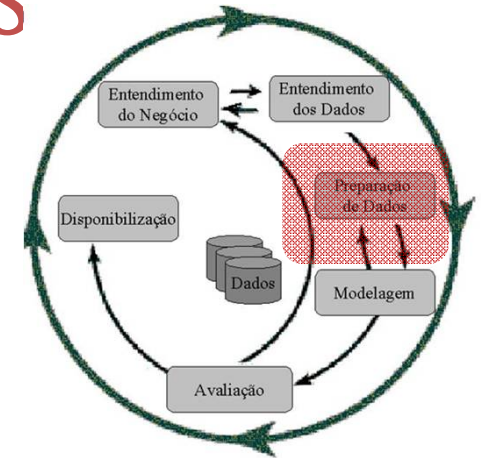


Exemplo: temperatura da superfície dos oceanos em julho de 1982



Preparação de Dados

Produção de um conjunto de dados adequado aos algoritmos de mineração



- Selecionar os dados
- Limpar os dados (dados nulos, ausentes, zerados, inválidos, etc..)
- Construir dados (criar novos atributos a partir de outros)
- Integrar dados: combinar múltiplas tabelas ou outras fontes
- Formatar dados: modificações sintáticas nos dados, sem alterar o seu significado. Ex:
 - Primeiro atributo tem que ser uma chave única
 - O arquivo tem que estar em uma ordem determinada
 - Retirar vírgulas dos campos para gerar um arquivo com atributos separados por vírgulas

Preparação de dados: seleção de dados

Seleção de atributos

- **motivos:**
 - Requisitos de tempo e espaço
 - Simplicidade do modelo gerado
 - Relevância dos atributos
 - Redundância entre atributos
 - Acurácia pode ser aumentada
- **forma:**
 - Manual
 - Por algoritmos: mais de 30 algoritmos

Preparação de dados: Limpeza dos dados

Visa garantir a qualidade dos dados

- Eliminação de dados errôneos –
- Padronização de dados: formato de datas, abreviaturas, valores de atributos (ex. sexo: M ou F, 0 ou 1, Mas e Fem, ...)
- Eliminação de duplicatas
- Tratamento de valores ausentes
 - Excluir instâncias
 - Completar valores ausentes
 - Complemento manual
 - Complemento com valor constante global: ex: “desconhecido”
 - Complementar com o valor mais provável
 - Complementar com o valor médio do atributo

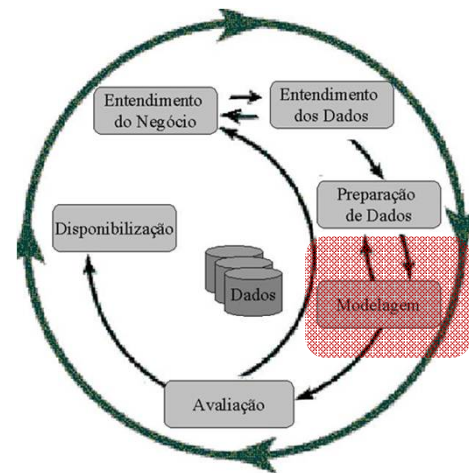
Preparação de dados: Construir dados

- Transformação de dados
 - Normalização
 - valores muito variados são transformados em um intervalo [0 – 1]
 - Transformação de valores simbólicos para numéricos
 - Ex: sexo Masculino = 0 e Feminino = 1
 - Discretização de atributos
 - transformar em faixas – ex: idade
- Criação de novos atributos.
 - Ex: área = comprimento x largura
 - Ex: idade = faixa de idades

Modelagem

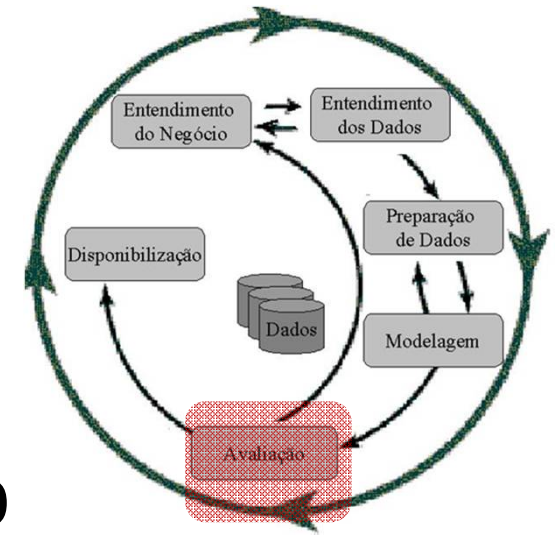
Corresponde a fase de **Mineração de Dados** utilizada por outros autores

- Selecionar a técnica de modelagem
- Gerar projeto de teste
- Construir modelo: mineração propriamente dita (aplicação do algoritmo)



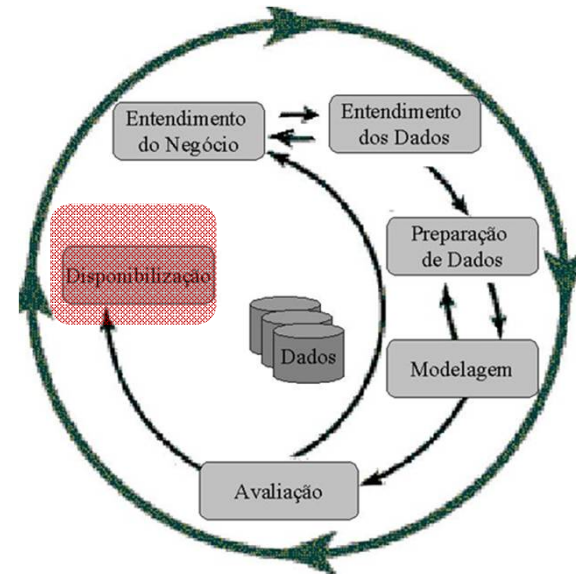
Avaliação

- Interpretar e avaliar os resultados em relação aos objetivos do usuário
- Avaliar resultados
- Revisar o processo
- Determinar próximos passos: ir para a fase final de disponibilização ou voltar para alguma etapa anterior



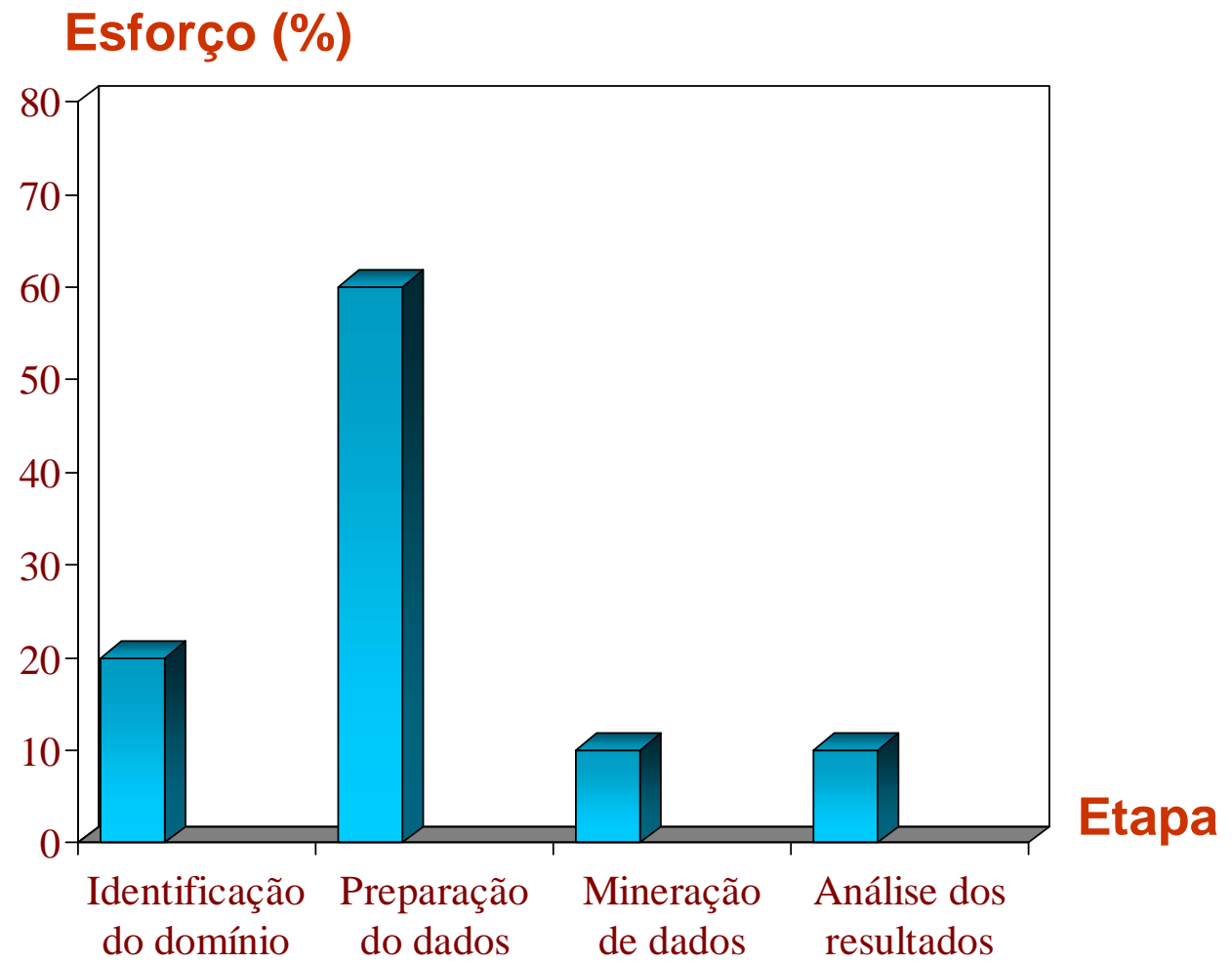
Disponibilização

- Planejar disponibilização: decidir a estratégia para a integração dos resultados obtidos no ambiente da organização
- Planejar monitoramento e manutenção:
- Produzir relatório final
- Revisar o projeto: avaliar pontos positivos e negativos do projeto, problemas e sugestões



Descoberta de Conhecimento em Bases de Dados

DESAFIOS:



Descoberta de Conhecimento em Bases de Dados

- O grande esforço da preparação dos dados se deve não apenas ao conhecimento dos dados, seleção e limpeza, mas a transformação
- Como a IA e BD são áreas distintas e os algoritmos de DM foram criados por pesquisadores da área de IA, a maioria deles não trabalha com BD, mas sim com arquivos texto ou uma tabela
 - Então entra a parte de desnormalizar o BD para criar o tabelão, com os dados selecionados, limpos e transformados para minerar

Outros tipos de dados

- Texto (ex: categorização de textos; “exame” de e-mails, ...)
- internet
 - conteúdo
 - estrutura
 - uso
- imagens
- seqüências de genes
- séries temporais
- dados de trajetórias
-

Estes dados requerem algoritmos mais sofisticados de mineração de dados

Tarefas de Mineração

Tarefa  descobrir um certo **tipo de padrão**

- ☐ Regras de Associação
- ☐ Análise de Sequências
- ☐ Classificação
- ☐ Agrupamento
- ☐ Outliers

Tarefas de Mineração de Dados

- **Tarefas Preditivas**

- prever o valor de um determinado atributo baseado nos valores de outros atributos

Classificação – Predição

- **Tarefas Descritivas**

- Derivar « **padrões** » : correlações, tendências, anomalias, agrupamentos dentro de uma grande massa de dados.

**Regras de Associação – Padrões Sequenciais –
Agrupamentos - Outliers**

TAREFAS, TÉCNICAS E ALGORITMOS

Aprendizado	Tarefa	Técnica
Não Supervisionado	Associação	Regras de Associação Padrões Sequenciais
	Agrupamento	Clustering
Supervisionado	Classificação	Regras de Indução Árvores de Decisão MBR – <i>Memory Based Reasoning</i> Redes Neurais
	Análise de Desvios	Árvores de Decisão Redes Neurais

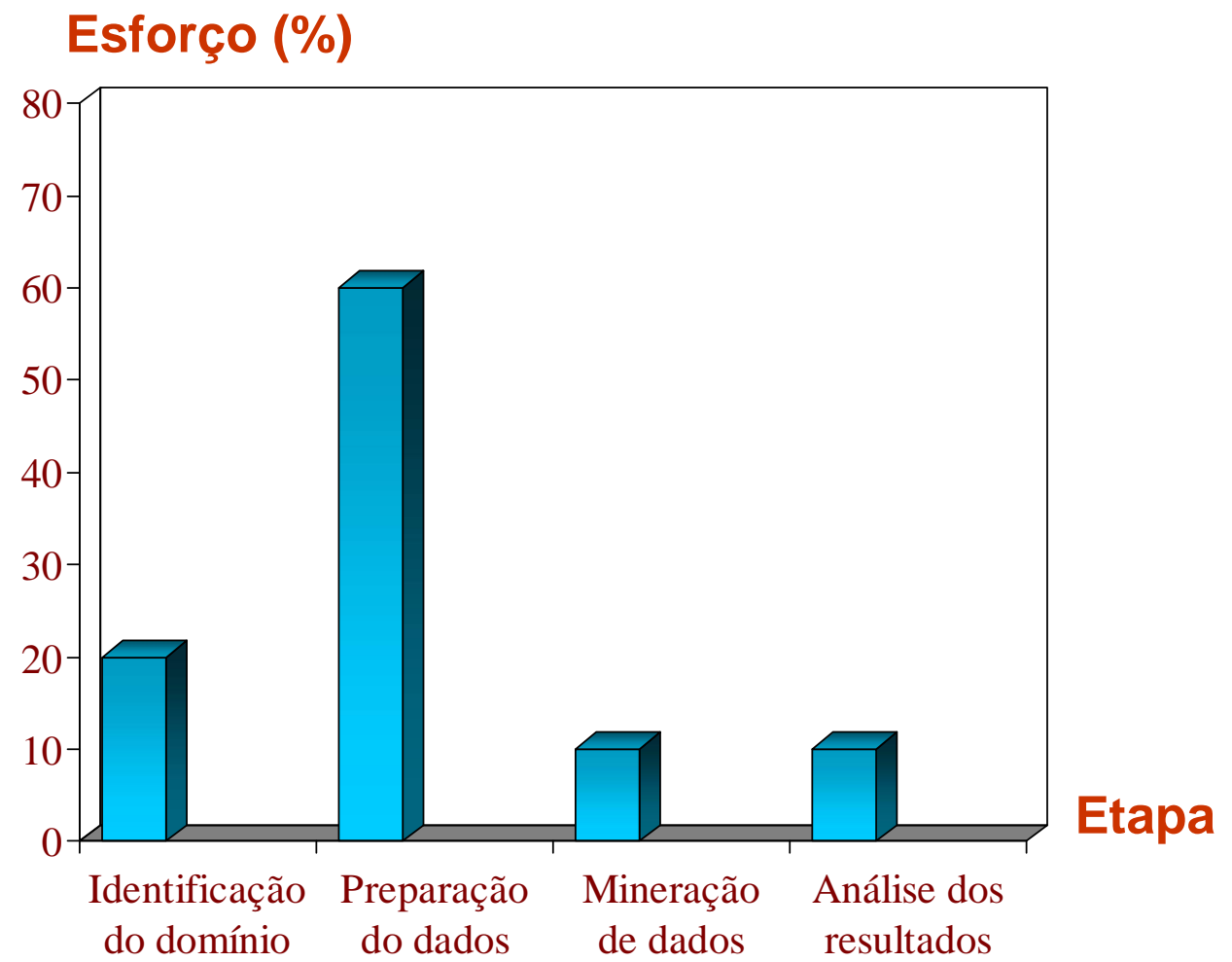
Como a Mineração é Conhecida?

- ❑ Mineração de Dados / Data Mining
- ❑ Descoberta de Conhecimento em bancos de dados (KDD)
- ❑ KDD = Knowledge Discovery in Databases

Sistemas de Mineração

- ☐ Intelligent Miner (IBM)
- ☐ DBMiner
- ☐ Enterprise Miner
- ☐ Clementine
- ☐ MineSet
- ☐ Genamics Expressions
- ☐ **Rapid Miner**
- ☐ **Weka**

Descoberta de Conhecimento em Bases de Dados



Seminários

1. Descoberta de conhecimento em Texto
 2. Data mining na Web 2.0
 3. Mineração de opinião (sentiment analysis, opinion mining)
 4. Data mining em séries temporais
 5. Data mining em grafos
 6. Data mining em trajetórias de objetos móveis
 7. Data mining em redes sociais – facebook, foursquare
 8. Data mining em redes sociais – twitter, flickr
 9. Data mining em imagens
 10. Data mining em dados geográficos
 11. Data mining em vídeos
-
- Grupos de 3 ou 4 alunos
 - Focar nos aspectos específicos do tipo de dado utilizado: pré-processamento, algoritmos utilizados etc.
 - Deverá ser entregue um relatório de 2 a 5 páginas sobre o assunto, com referências bibliográficas.