Processo de Descoberta de Conhecimento em Base de Dados (KDD - Knowledge Discovery in Databases)

Profa. Leticia T. M. Zoby

(leticia.zoby@udf.edu.br)

O processo de KDD – Conceitos Básicos

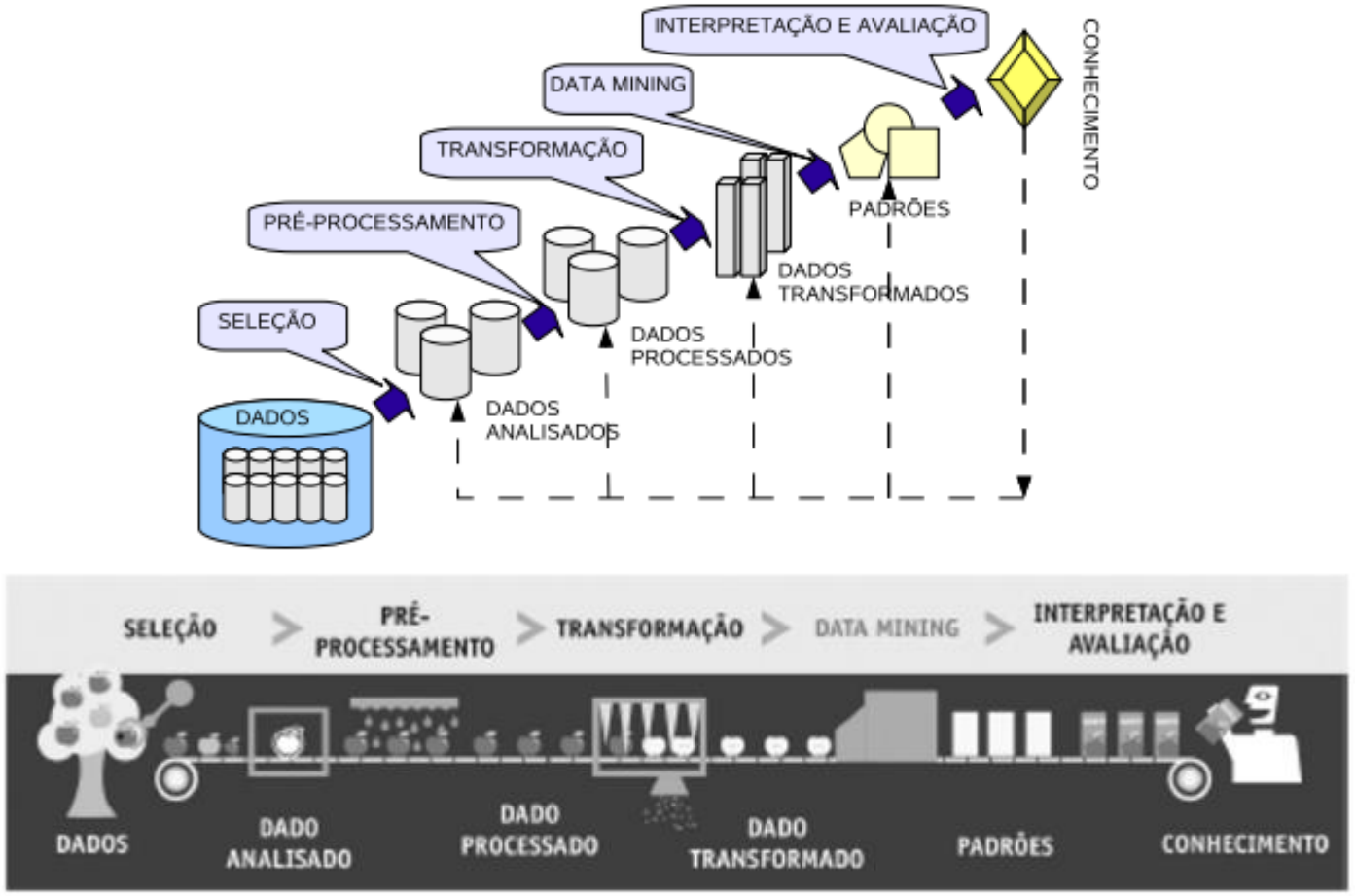
- Inviável análise dos dados armazenados de forma manual. Necessário a utilização de ferramentas para análise de grande conjuntos de dados.
  - O Processo de Descoberta de Conhecimento em Base de Dados – KDD.
- O KDD é um processo não trivial para extração de informações/ padrões válidos, novos e potencialmente compreensível em grandes conjuntos de dados.
- O KDD além de buscar conhecimento a partir de dados, também procura aplicar os algoritmos no conjunto de dados e funcionar de forma eficiente, e como a interação homem-máquina pode ser modelado e suportado.

O processo de KDD

- Ao se definir o objetivo para execução do KDD, ainda pode ser dividido em dois: **verificação** e **descoberta**.
  - Na verificação o sistema é limitado a confirmar hipóteses do usuário.
  - A descoberta, o sistema automaticamente encontra novos padrões, este ainda é dividido em **predição** (o sistema encontra padrões para prever o futuro de alguma entidade) e **descrição** (o sistema encontra padrões para o usuário com a representações de forma compreensível).

O processo de KDD

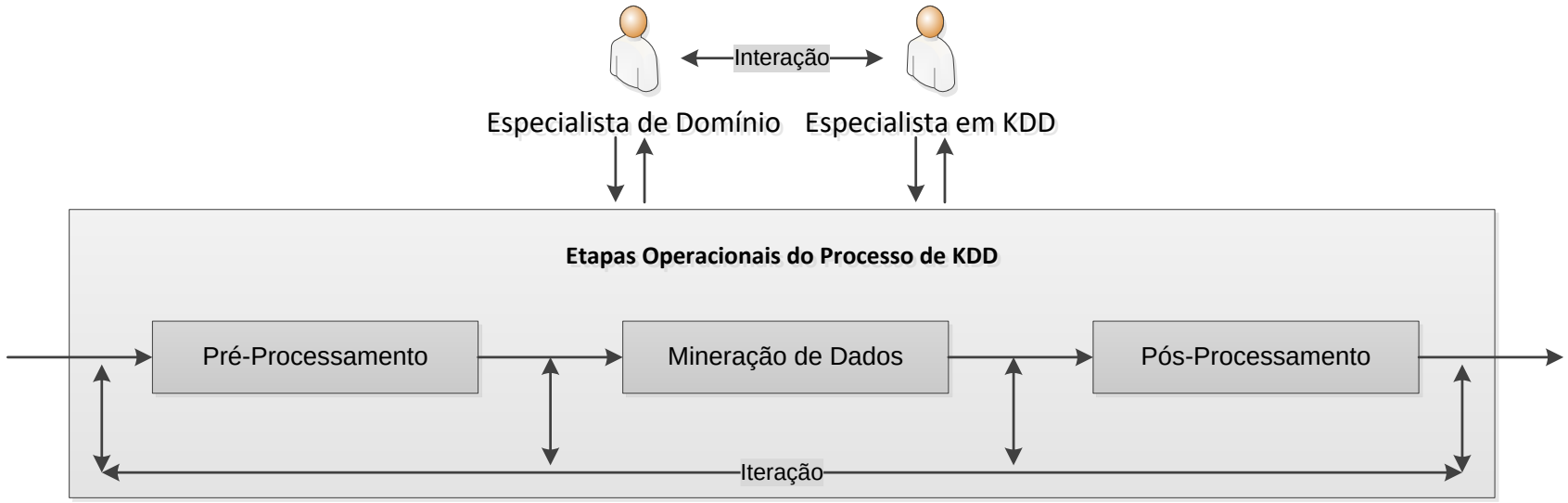
- Os passos (FAYYAD, 1996):



Processo de KDD (FAYYAD, apud GONÇALVES, 2001)

O processo de KDD

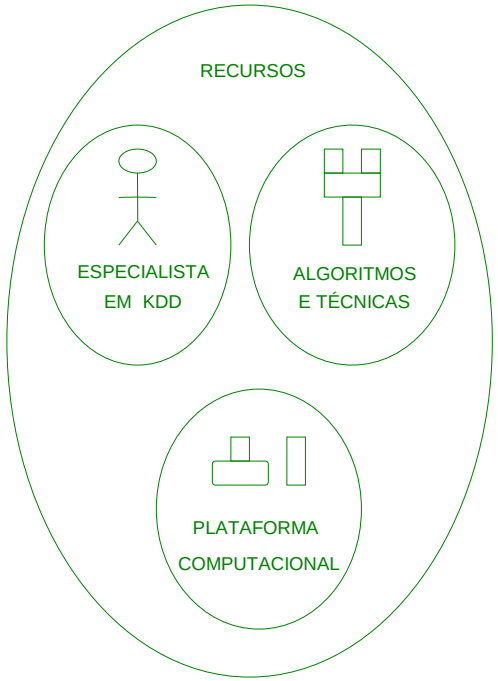
- Os passos (R. Goldschmidt; E. Passos; E. Bezerra, 2015):

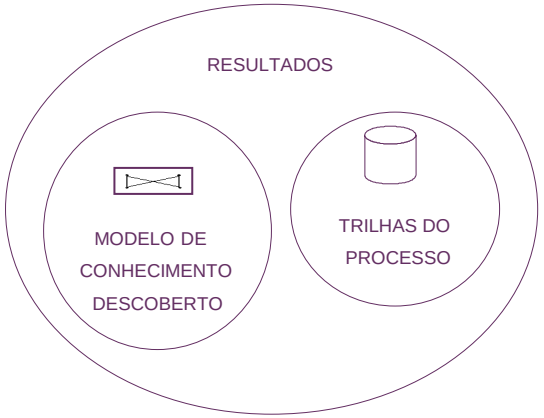


- Interação: Combinação de Ações Homem-Máquina.
- Iteração: Refinamentos Sucessivos.
- Padrão: Forma de Representação do Conhecimento.
- Compreensão: Padrão Representado de Forma Inteligível.
- Validade: Aplicação Adequada a um Contexto.
- Inovação: Mudança de Ctos Anteriores p/ Ctos Descobertos.
- Utilidade: Benefícios da Aplicação.

O processo de KDD: Visão Geral

- Aplicação de KDD
  - envolve os seguintes elementos (R. Goldschmidt; E. Passos; E. Bezerra, 2015):

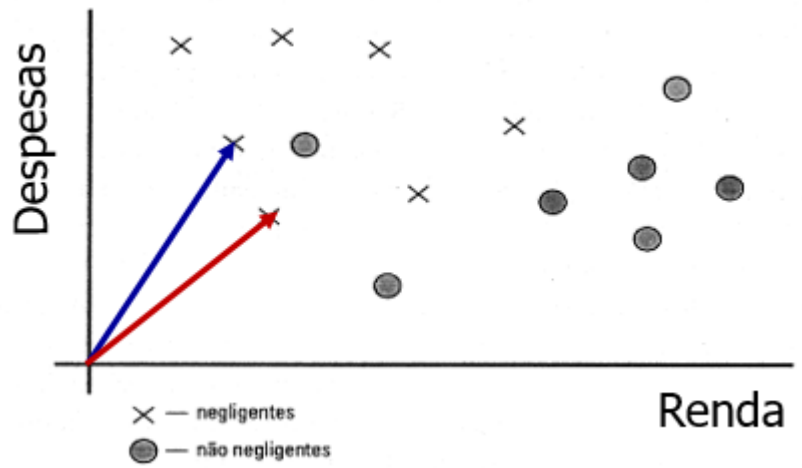




O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

- Elementos do Problema: CONJUNTO DE DADOS
  - Cada caso corresponde a um vetor em um espaço n-dimensional



Fundamentação: ***Álgebra Linear***.

Conceito de ***similaridade*** ou ***distância*** entre pontos (vetores).

Qto ***menor*** a ***distância*** entre 2 pontos, ***maior*** a ***similaridade*** entre os objetos representados.

O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

- Elementos do Problema: ESPECIALISTA DO DOMÍNIO DA APLICAÇÃO

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

O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

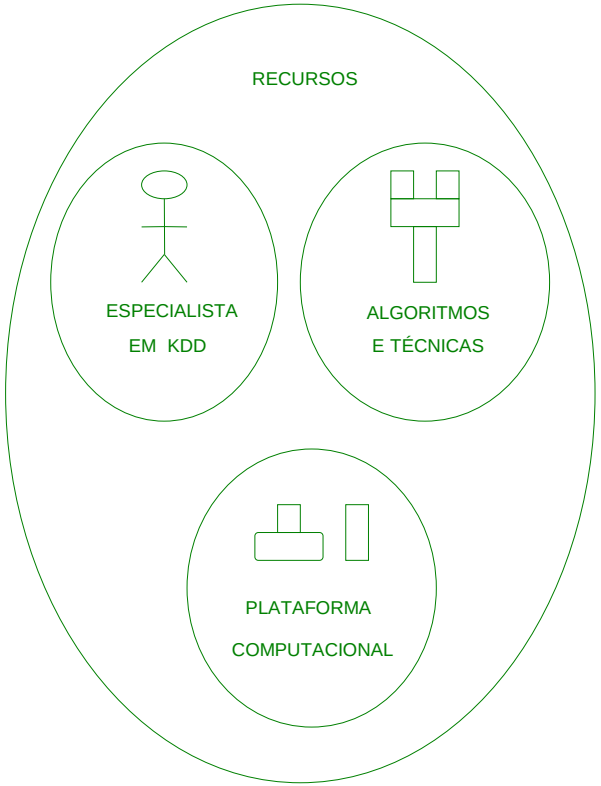
- Elementos do Problema: OBJETIVOS DA APLICAÇÃO

  - 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

O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

- Elementos dos Recursos



O processo de KDD: Visão Geral

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

O processo de KDD: Visão Geral

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.

O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

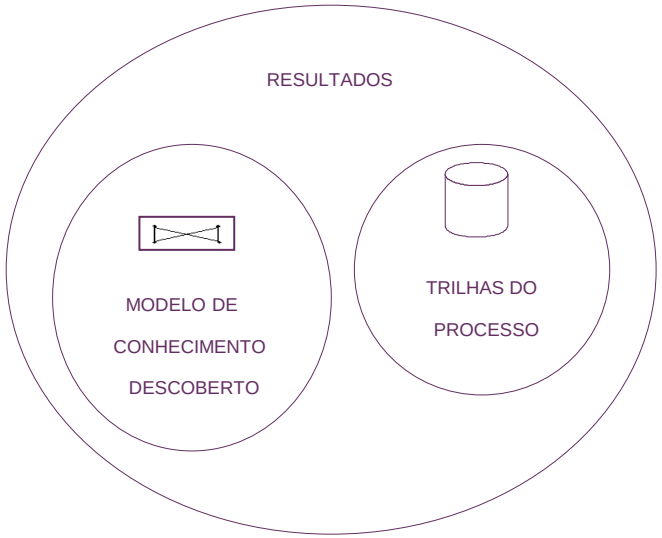
- Elementos dos Recursos: PLATAFORMA COMPUTACIONAL

  - 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.

O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

- Elementos dos Resultados



O processo de KDD: Visão Geral

FORMALIZAÇÃO DO MODELO PROPOSTO

- Elementos dos Resultados: MODELO 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.

O processo de KDD: Visão Geral

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 uma **Máquina de Assistência** à Orientação do Processo de KDD.

O processo de KDD: Visão Geral

- Áreas de Origem:

Classificadores Bayesianos
Redes Bayesianos
EDA - *Exploratory Data Analysis*

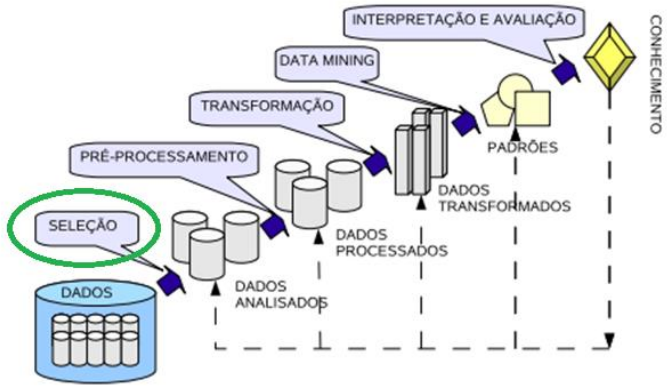
Data
Warehousing
SQL
OLAP
DMQL
KMQL
NoSQL

Lógica Nebulosa
Lógica Indutiva
Árvores de Decisão

O processo de KDD

- Os passos (FAYYAD, 1996):
  - Seleção
  - Pré-processamento
  - Transformação
  - Data mining (aprendizagem)
  - Interpretação e Avaliação

O processo de KDD



Seleção

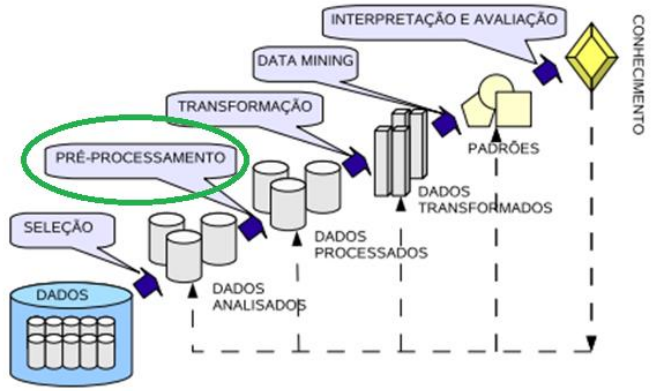
- Nesta etapa é escolhido o conjunto de dados, pertencente a um domínio, contendo todas as possíveis variáveis (também chamadas de características ou atributos) e registros (também chamados de casos ou observações) que farão parte da análise.

- Definir quais os dados que serão importantes.

- Os dados coletados podem possuir diferentes formatos (arquivo texto, banco de dados, planilhas). Cada formato de armazenamento será exigido uma ferramenta especifica para a extração desses dados. A extração destes dados pode se tornar complexa, e se realizada de forma incorreta pode significar o fracasso das etapas subsequentes.

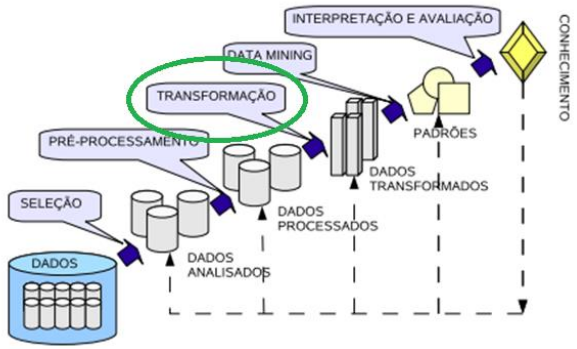
O processo de KDD

Pré-processamento (e limpeza dos dados)



- E importante no processo pois a qualidade dos dados vai determinar a eficiência dos algoritmos de mineração.
- Nesta etapa deverão ser realizadas tarefas que eliminem dados redundantes e inconsistentes, recuperem dados incompletos e avaliem possíveis dados discrepantes ao conjunto (*outliers*).
  - Dados ausentes (*missing values*)
  - Dados discrepantes (*outliers*)
  - Dados derivados
- Pode existir erros de digitação ou erros nos sistemas de captura dos dados por sensores.
- A limpeza dos dados tem por objetivo assegurar a **qualidade** dos dados selecionados.

O processo de KDD

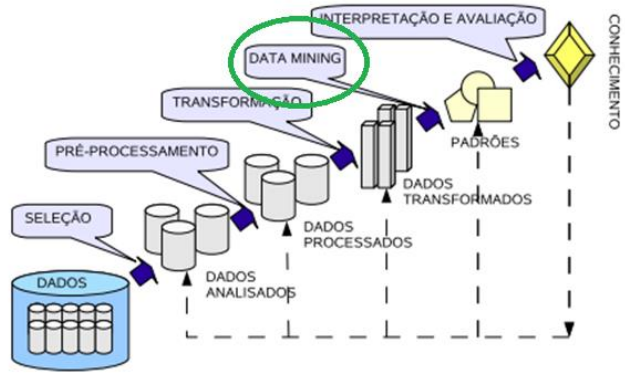


Transformação

- Após serem selecionados, pré-processados e limpos os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizado possam ser aplicados.
- Em grandes corporações e comum encontrar computadores rodando diferentes sistemas operacionais e diferentes Sistemas Gerenciadores de Bancos de Dados (SGDB). Estes dados que estão dispersos devem ser agrupados em um repositório único.
- Alguns métodos para resolução dos problemas dessa etapa:
  - Redução número de exemplos
  - Redução do número de atributos
  - Redução número dos dados

O processo de KDD

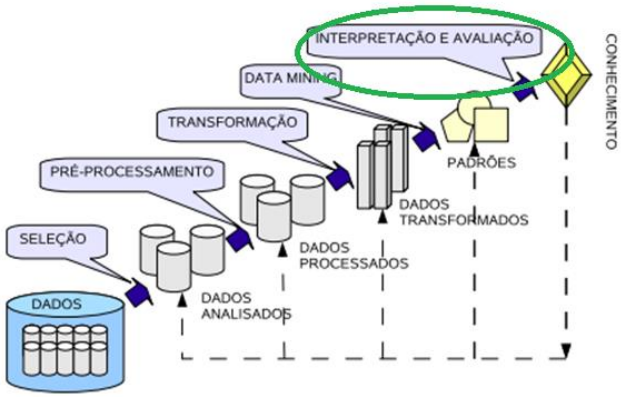
Data Mining/ Mineração de Dados



- É a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras.
- O objetivo principal do processo de *data mining* e fornecer as corporações informações que a possibilitem montar melhores estratégias de marketing, vendas e suporte, melhorando assim os seus negócios.

O processo de KDD

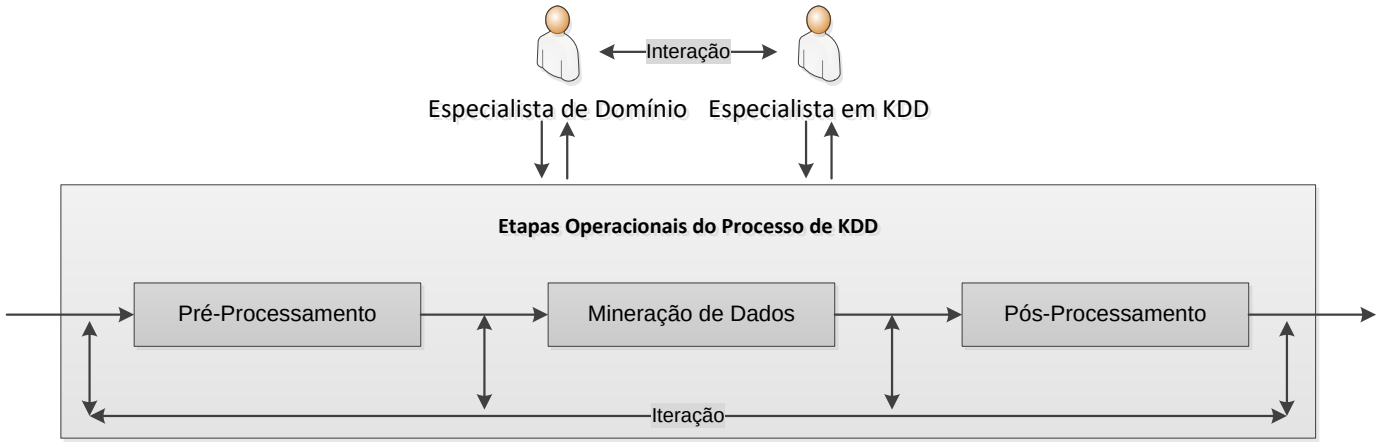
Interpretação e Avaliação



- Ser realizada em conjunto com um ou mais especialistas no assunto.
- O conhecimento adquirido através da técnica de *data mining* deve ser interpretado e avaliado para que o objetivo final seja alcançado.
- Quando o resultado não é satisfatório, o processo pode retornar a qualquer um dos estágios anteriores ou ate mesmo ser recomeçado. Duas ações mais comuns:
  - modificar o conjunto de dados inicial e/ou
  - trocar o algoritmo de data mining (ou ao menos alterar suas configurações de entrada).

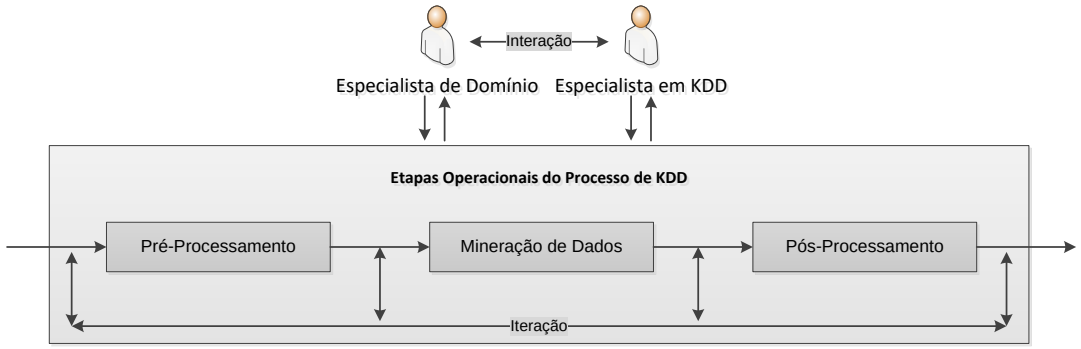
O processo de KDD

- Os passos (R. Goldschmidt; E. Passos; E. Bezerra, 2015):
  - Pré-processamento
  - Mineração de Dados /Data mining
  - Pós-processamento



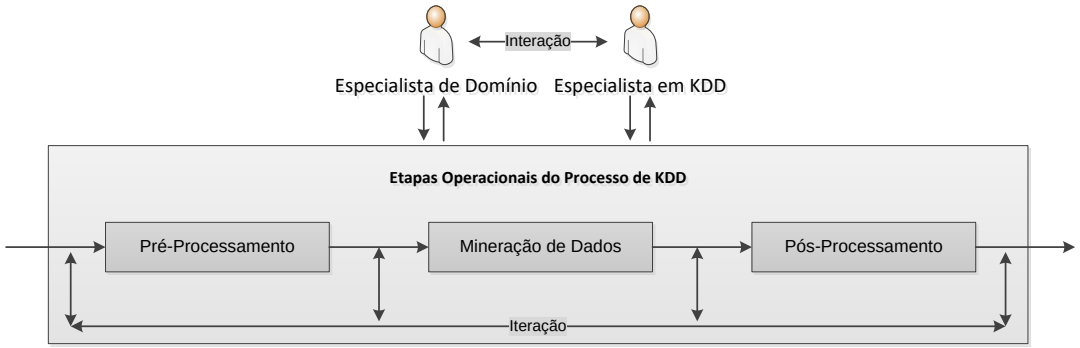
O processo de KDD

- Pré-processamento
  - Seleção de Dados
  - Limpeza dos Dados
  - Codificação dos Dados
  - Enriquecimento dos Dados



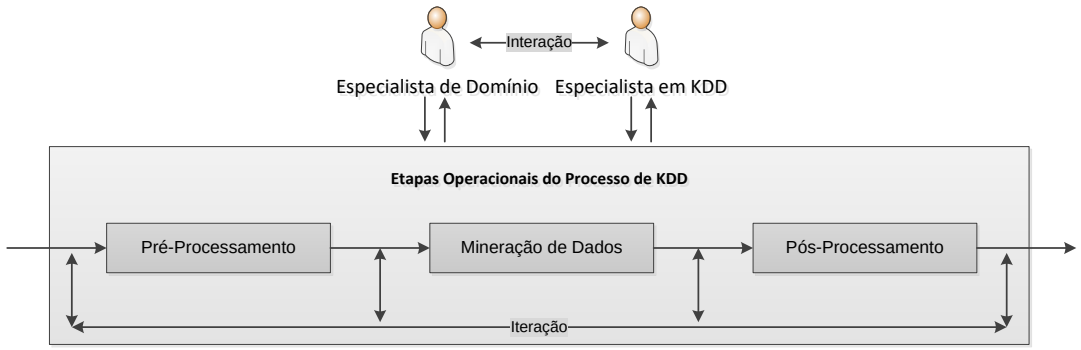
O processo de KDD

- Mineração de Dados /Data mining
  - Descoberta de Associações
  - Classificação
  - Regressão
  - Agrupamento (*Clusterização*)
  - Sumarização
  - Detecção de Desvios
  - Descobertas de Sequências



O processo de KDD

- Pós-processamento
  - Principais funções:
    - Elaboração e organização do conhecimento obtido
      - Ex: simplicação de gráficos, diagramas, ou relatórios demostrativos
    - Conversão da forma de representação do conhecimento obtido

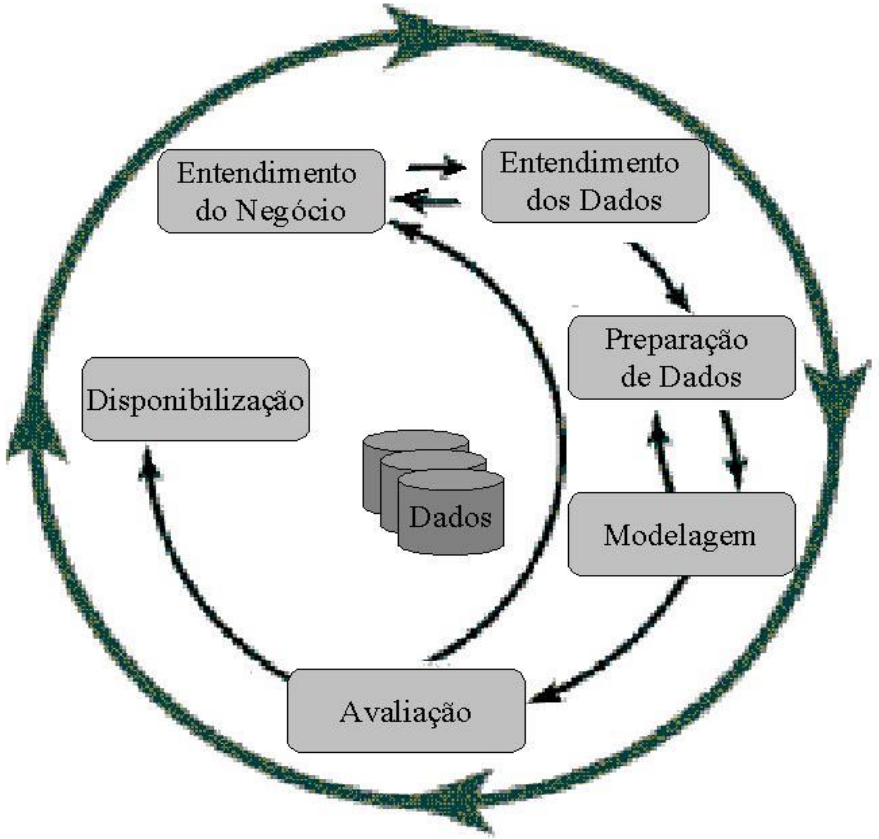


O processo de KDD

- O processo de KDD é interativo (pois o usuário pode intervir e controlar o curso das atividades) e iterativo (por ser uma sequência finita de operações em que o resultado de cada uma e dependente dos resultados das que a precedem).

Metodologia CRISP-DM

- CRISP-DM = *CRoss – Industry Standard Process for Data Mining* (projeto europeu ESPRIT com vários parceiros industriais)
- Geral - não se restringe a ferramenta ou tecnologia específica



Metodologia CRISP-DM

- Ciclo de vida de um projeto de mineração de dados:

Entendimento do Negócio	Foco no **entendimento do negócio** que visa obter conhecimento sobre os objetivos do negócio e seus requisitos.
Seleção dos Dados	Consiste no **entendimento dos dados**, que visa à familiarização com o banco de dados pelo grupo de projeto, utilizando-se de conjuntos de dados "modelo".
Limpeza dos Dados	Fase de **preparação de dados**, que consiste na preparação dos dados buscando a limpeza, a transformação, a integração e a formatação dos dados da etapa anterior.
Modelagem dos Dados	Fase que consiste na modelagem dos dados, a qual visa a aplicação de técnicas de modelagem sobre o conjunto de dados preparado na etapa anterior. Técnicas são baseadas em conceitos de: – Aprendizagem de máquina; – Reconhecimento de padrões; – Estatística.
Avaliação do processo	Visa garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas.
Execução	Esta fase consiste na definição das fases de implantação do projeto de Mineração de Dados.

Referências...