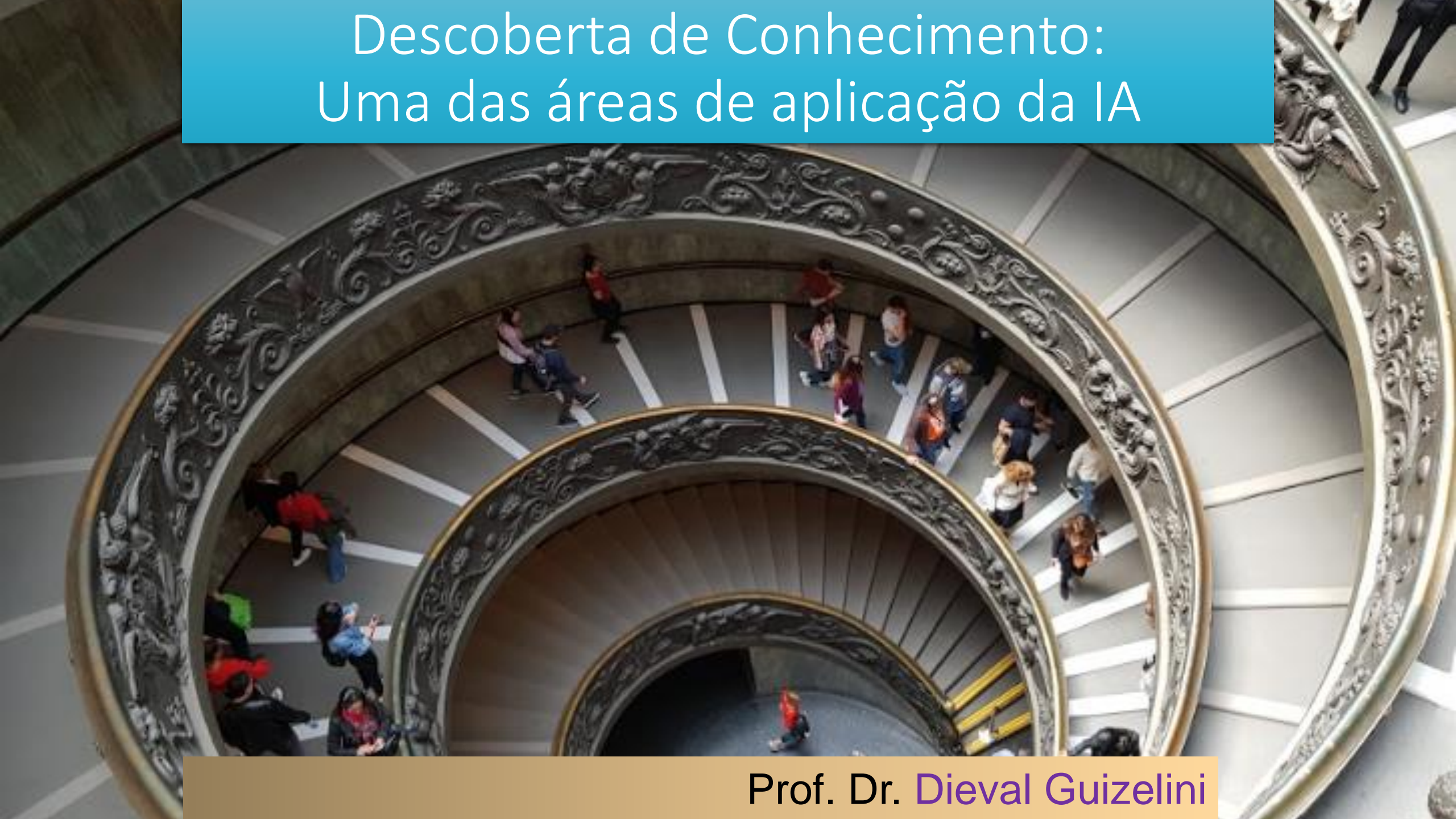


Descoberta de Conhecimento: Uma das áreas de aplicação da IA

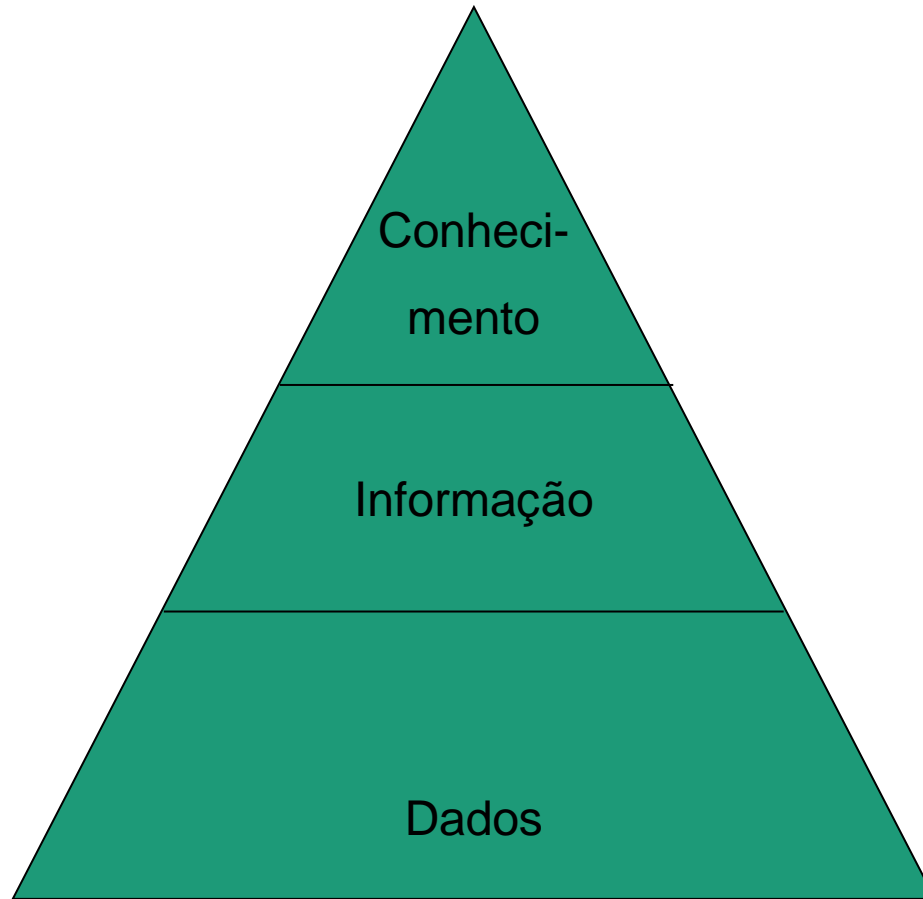


Prof. Dr. Dieval Guizelini

Conceito Geral

- Uma área de conhecimento:
Descoberta de Conhecimento em Banco de Dados
(do inglês, *Knowledge Discovery in Databases*)
 - Apesar da popularidade, a **mineração de dados** é apenas uma das etapas no processo de descoberta de conhecimento.

A hierarquia entre Dado, Informação e Conhecimento



- **Uma Regra:**
Se capacidade mensal de endividamento > 60%, então crédito = SIM
- **Um índice/indicador**
Capacidade mensal de endividamento mensal = $100 * (\text{renda mensal} - \text{despesas mensais}) / \text{renda mensal}$
- Renda mensal, despesas mensais, tickets de vendas, dados cadastrais...

Definição Segundo Fayyad

O termo KDD foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de bases de dados. Porém, uma das definições mais populares, foi proposta pelo grupo de Fayyad (1996):

“ KDD é um processo, de várias etapas, não trivial, interativo e iterativa, para identificar **padrões compreensíveis**, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados ”

Etapas do processo de KDD

Pré-Processamento

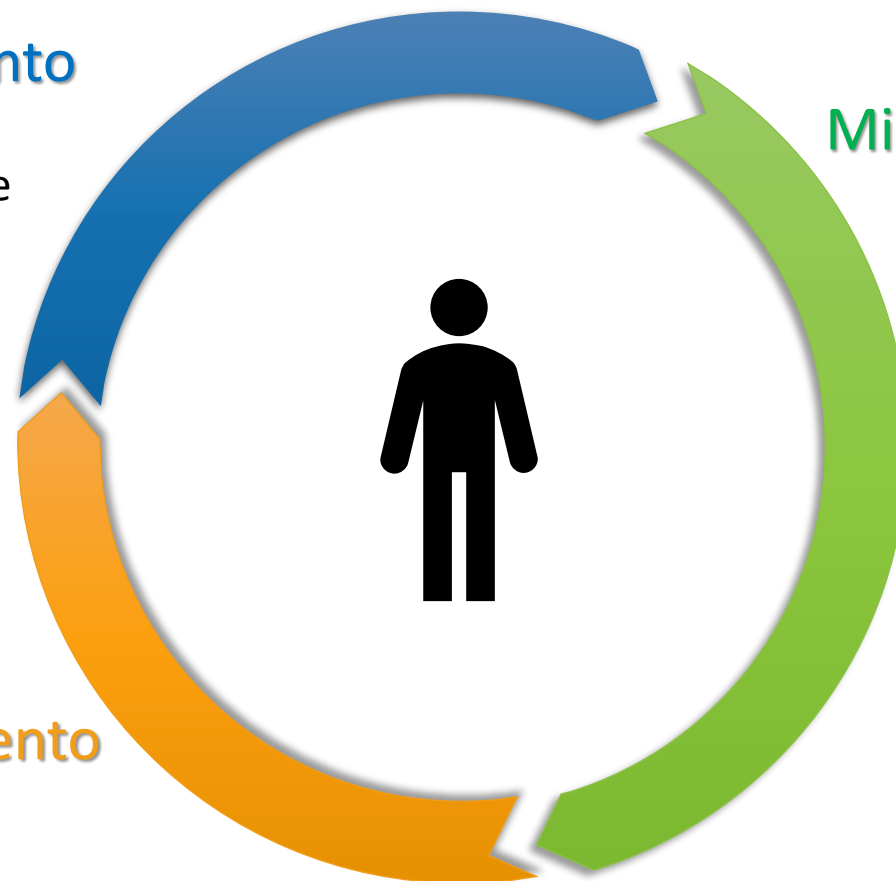
objetiva-se em obter e preparar os dados

Mineração de Dados

busca efetiva por conhecimento úteis no contexto de KDD

Pós-Processamento

tratamento do conhecimento obtido.



O problema...

- De forma geral, a complexidade do processo de KDD está na dificuldade em perceber e interpretar adequadamente inúmeros fatos observáveis durante o processo e na dificuldade em conjugar dinamicamente tais interpretações de forma a decidir quais ações devem ser realizadas em cada caso (Goldschmidt, 2003).
- Cabe ao analista humano a árdua tarefa de orientar a execução do processo de KDD.

Compreendendo o conceito

Interativo: indica a necessidade de atuação humana

Conhecimento útil: é aquele que pode ser aplicado de forma a proporcionar benefícios.

Iterativo: sugere a possibilidade de repetições

KDD é um processo, de várias etapas, não trivial, interativo e iterativa, para identificar padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados

Padrão novo: deve acrescentar novos conhecimentos aos conhecimentos existentes.

Não trivial: alerta para a complexidade

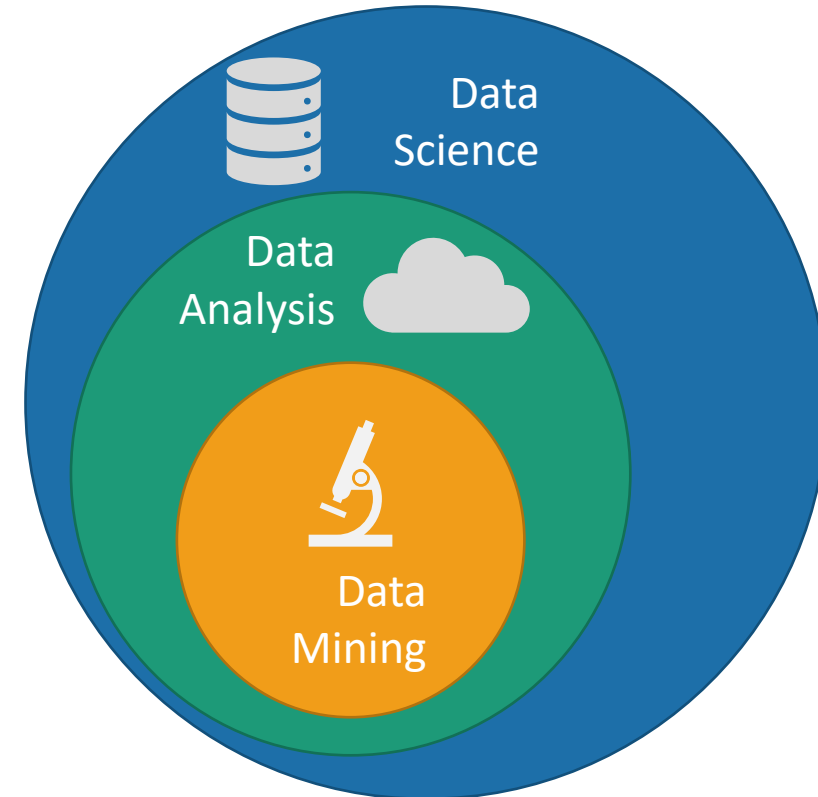
Padrão válido: indica conhecimento verdadeiro e aplicável a outros conjuntos...

Um padrão: deve ser interpretado como um conhecimento representado, segundo as normas sintáticas de alguma linguagem formal.

Historicamente multidisciplinar...

Origina-se de diversas áreas, entre elas:

- Estatística
- Inteligência Computacional
 - Aprendizado de Máquinas
 - Reconhecimento de Padrões
- Banco de Dados



As atividades podem ser organizadas em três grupos...

Desenvolvimento tecnológicos

Concepção, aprimoramento, desenvolvimento de algoritmos ferramentas, etc

Aplicação de Resultados

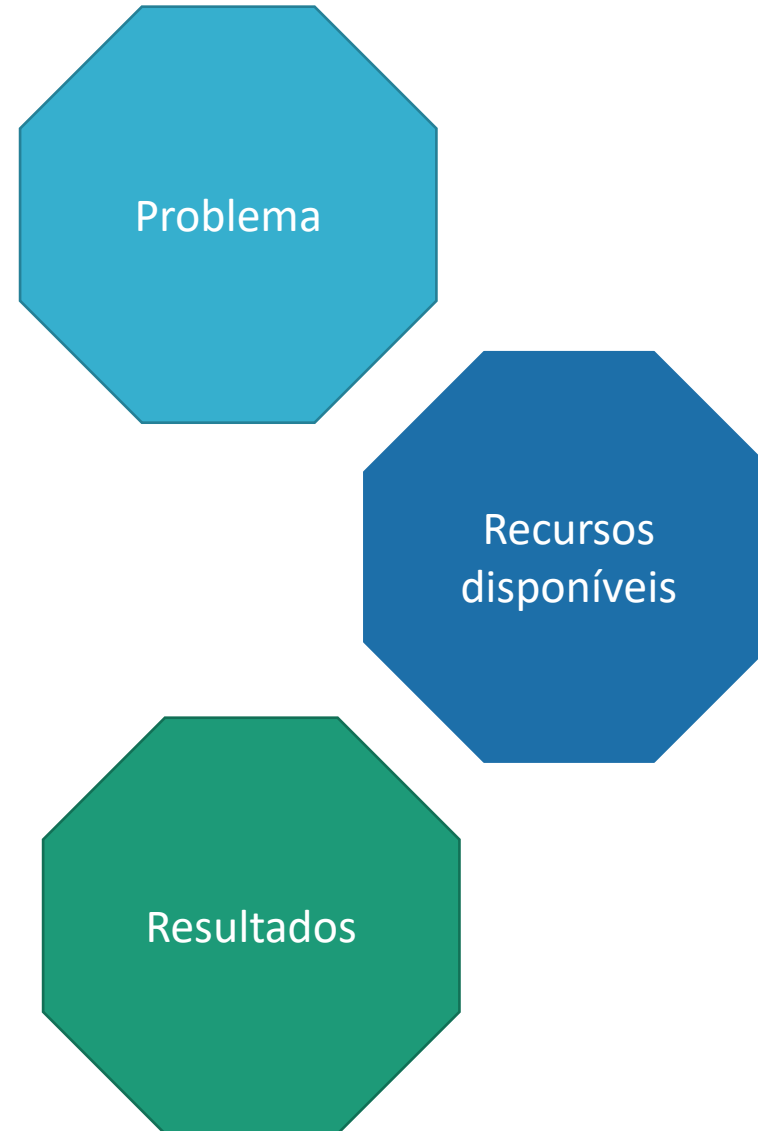
Alterações em estratégias de negócios, Novas ferramentas de busca, classificação de ativos...

Execução de KDD

busca efetiva por conhecimento nas bases de Dados

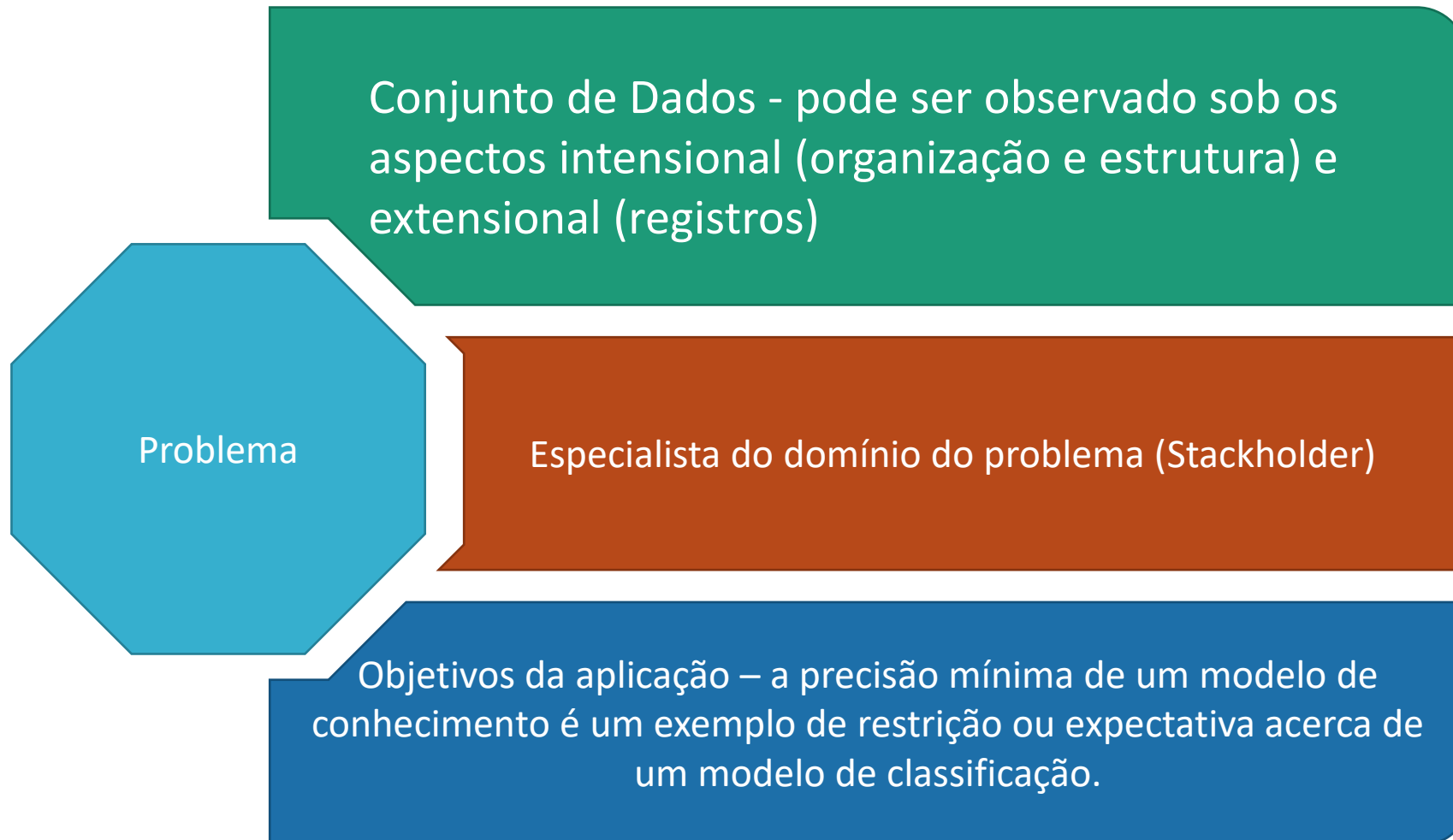
Conceitos Básicos, Caracterização do Processo

Uma aplicação de KDD é composta por três tipos de componentes:



Caracterização do Processo de KDD:

O Problema a ser submetido ao processo de KDD poder ser caracterizado por 3 elementos



Caracterização do Processo de KDD: Os recursos disponíveis

- Os recursos disponíveis para solução do problema:
 - **o especialista**: pessoa ou grupo de pessoas que possuem experiência na execução de um processo de KDD.
 - **as ferramentas de KDD**: Quais softwares, ambientes e ferramentas de análise estão disponíveis.
 - **a plataforma computacional**: Quais os recursos de hardware (memória e processamento) disponíveis.

Caracterização do Processo de KDD: Resultados

- Resultados obtidos a partir da aplicação dos recursos no problema. Compreende os modelos de conhecimento descobertos ao longo da aplicação de KDD e o histórico de ações realizadas.

Onde:

- **Modelo de conhecimento:** indica qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva algum conjunto de dados (Fayyad et al, 1996) O modelo de conhecimento com maior precisão na classificação dos clientes possui maiores chances de ser eleito como principal resultado gerado pelo processo de KDD.
- **Históricos:** como os modelos de conhecimento foram gerados, também enquadram-se como resultados do processo de KDD. São fundamentais para o controle do processo, pois permitem uma análise crítica e uma revisão das ações realizadas.

Etapa inicial: Pré-Processamento

- Compreende todas as funções relacionadas à captação, à organização e ao tratamento dos dados.
- As funções são:
 - Seleção dos dados:
 - Limpeza dos dados:
 - Codificação dos dados:
 - Enriquecimento dos dados:



Arquitetura de Dados

Etapa de Mineração de Dados

- Durante a etapa de mineração de dados é realizada a busca efetiva por conhecimento úteis no contexto da aplicação de KDD. Na mineração de dados são definidas as técnicas e os algoritmos a serem utilizados no problema.
 - Descoberta de Associação
 - Classificação
 - Regressão
 - Clusterização
 - Sumarização
 - Detecção de desvios
 - Descoberta de Sequências

Etapa de Pós-Processamento

- A etapa de pós-processamento abrange o tratamento do conhecimento obtido na mineração dos dados. Tem por objetivo facilitar a interpretação e a avaliação da utilidade do conhecimento descoberto (Fayyad et al, 1996).
- As principais funções dessa etapa são:
 - Elaboração e organização (podendo ser: simplificação, uso de gráficos e diagramas etc)
 - Conversão da forma de representação do conhecimento obtido

Por onde se começa o processo de KDD

- Em geral, todo projeto de KDD deve ser iniciado pela análise da base de dados e por entrevistas com os especialistas.
- Definição dos objetivos da pesquisa – que deverá nortear todo o processo de KDD.
- Toda aplicação de KDD deve ser classificada nas dimensões:
 - Orientação das ações a serem realizadas
 - Quanto ao macroobjetivo pretendido

Por onde se começa o processo de KDD: Quanto a orientação das ações

- A classificação da aplicação de KDD quanto à orientação das ações a serem realizadas pode ser:
 - **Validação de hipóteses postuladas:** o especialista da área em que deseja realizar o processo de KDD apresenta alguma hipótese que deve ser comprovada ou refutada, mediante análise dos dados
 - **Descoberta de Conhecimento:** enquadra-se a busca efetiva por conhecimentos a partir da abstração dos dados existentes.

Por onde se começa o processo de KDD: Quanto aos macroobjetivos

- A classificação da aplicação de KDD quanto ao macroobjetivo pode ser:
 - **Predição:** é a busca por um modelo de conhecimento que permite a partir de um histórico de casos anteriores, prever os valores de determinados atributos em novas situações.
 - **Descrição:** busca-se um modelo que descreva, de forma compreensível pelo homem, o conhecimento existente em um conjunto de dados.

Operações e métodos de KDD

- A expressão “operações de KDD” se refere a qualquer função das etapas operacionais de KDD. Portanto uma operação é a especificação lógica de uma função. Exemplo: seleção, limpeza, codificação, classificação, descoberta de associação etc.
- As operações da etapa de mineração são, normalmente, denominadas tarefas de KDD ou tarefas de mineração de dados.
- Os métodos de KDD são implementações específicas das operações de KDD. Um método de KDD corresponde a um algoritmo particular. Exemplo: apriori, Análise de Componentes Principais (PCA), Corte de Regras etc.

Técnicas de KDD

- A expressão técnicas de KDD se refere a qualquer teoria que possa fundamentar a implementação de um método de KDD.
Exemplos de métodos da IA:
 - A Teoria de Redes Neurais subsidiou o desenvolvimento de Máquinas de Suporte Virtual (SVM) (Haykin, 1999), que por sua vez, são aplicáveis em tarefas de classificação;
 - O Método Rule Evolver (Lopes et al, 1999) foi concebido a partir da Teoria de Algoritmos Genéticos (Davis, 1999) e pode ser aplicado em operações de Classificação e Sumarização (Fayyad et al, 1996);
 - O Método NFHB-Class (Gonçalves, 2001; Souza, 1999) baseia-se na combinação de princípios da Teoria Lógica Nebulosa e das Redes Neurais e pode ser aplicado em operações de classificação.

Classificação das Técnicas de KDD

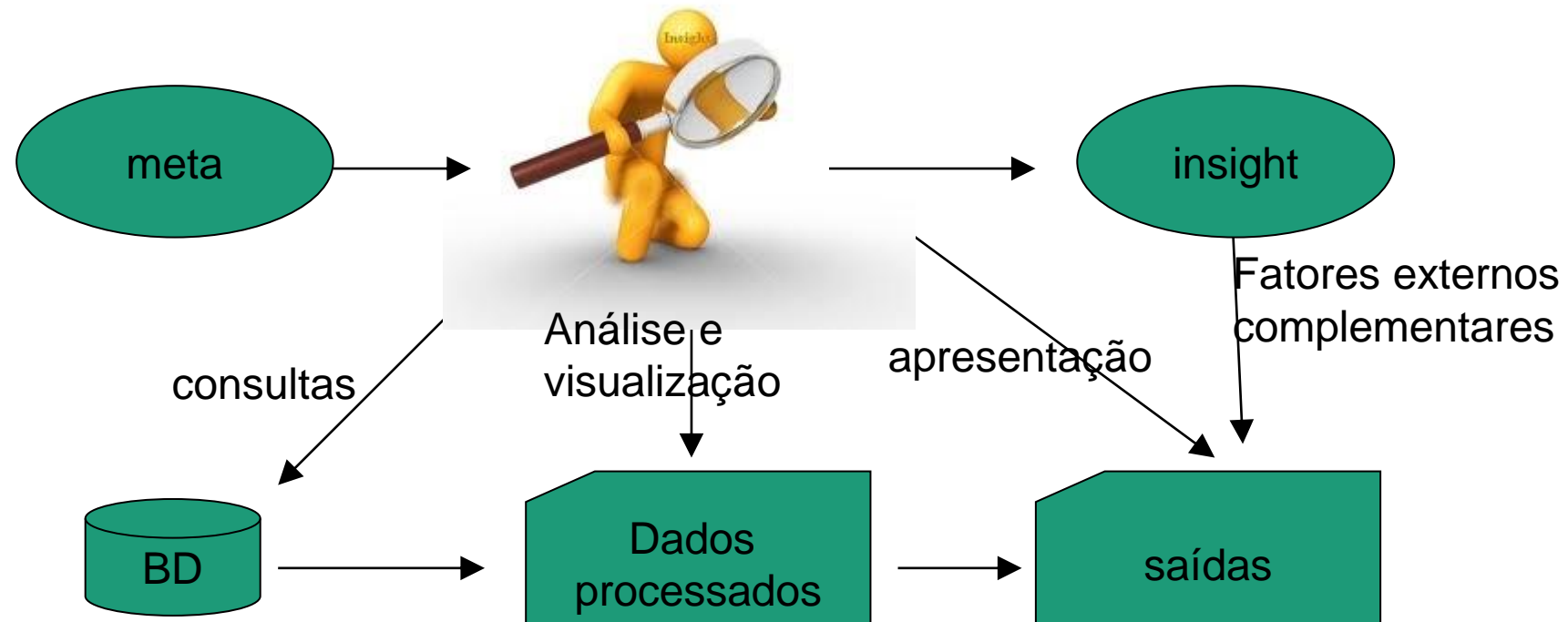
- São diversos tipos de técnicas e de algoritmos para mineração, elas podem ser divididas em:
- **Técnicas Tradicionais:** são tecnologias que existem independente do contexto de MD, ex:
 - Redes Neurais: Uma Rede Neural Artificial (RNA) é uma técnica computacional que constrói um modelo matemático inspirado em um sistema neural biológico simplificado, com capacidade de **aprendizado**, **generalização**, **associação** e **abstração**.
 - Lógica Nebulosa (Fuzzy): É uma técnica que permite a construção de sistemas que lidam com informações **imprecisas** ou **subjetivas**.
 - Algoritmos Genéticos: (AG): são modelos de otimização, inspirados na evolução natural e na genética, aplicados a problemas complexos de otimização e busca. As AGs pode ser descritas em três componentes: variáveis, restrições e objetivo. São aplicados em problemas complexos com **muitas variáveis e restrições** ou com **grandes espaços de buscas**.
 - Estatística: Fornece diversos modelos e técnicas tradicionais para **análise** e **interpretação de dados**. Ex: Redes Bayesianas, Análise Discriminante, Análise Exploratórios de dados etc.

Classificação das Técnicas de KDD (2/2)

- **Técnicas Específicas:** são técnicas desenvolvidas especificamente para a aplicação em tarefas de KDD. Ex: apriori (técnica desenvolvida para descoberta de associação).
- **Técnicas Híbridas:** São técnicas combinadas de forma a gerar sistemas híbridos.

O Papel do Usuário no Processo de KDD

- A participação do usuário é fundamental para o processo de KDD, os exemplos e problemas observados anteriormente podem ser ampliados ou reduzidos em consequência do conhecimento especializado e da participação dos usuários.



- No entanto, a formação do especialista em KDD constitui-se em uma tarefa árdua, longa e exaustiva, pois requer não somente uma fundamentação teórica sobre a área, mas também a participação destes em inúmeras experiências práticas reais.

Bibliografia

- GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. ***Data mining: um guia prático***. Rio de Janeiro: Elsevier, 2005.
- FAYYAD, U. M.; PLATETSKY-SHAPIRO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery: An Overview*. **Knowledge Discovery and Data Mining**, Menlo Park: AAAI Press, 1996
- FAYYAD, U. M.; PLATETSKY-SHAPIRO, G.; SMYTH, P. The *KDD Process for Extracting Useful Knowledge from Volumes of Data*. **Communications of the ACM**, v. 39, 1996
- HAYKIN, S. *Neural Networks: a Comprehensive Foundation*. Prentice Hall, 1999.