Extração de Conhecimento & Mineração de Dados

- Nesta apresentação é dada uma breve introdução à Extração de Conhecimento e Mineração de Dados

José Augusto Baranauskas
Departamento de Física e Matemática – FFCLRP-USP

augusto@ffclrp.usp.br
http://dfm.ffclrp.usp.br/~augusto

“Leis”, Gigantes e Monstros

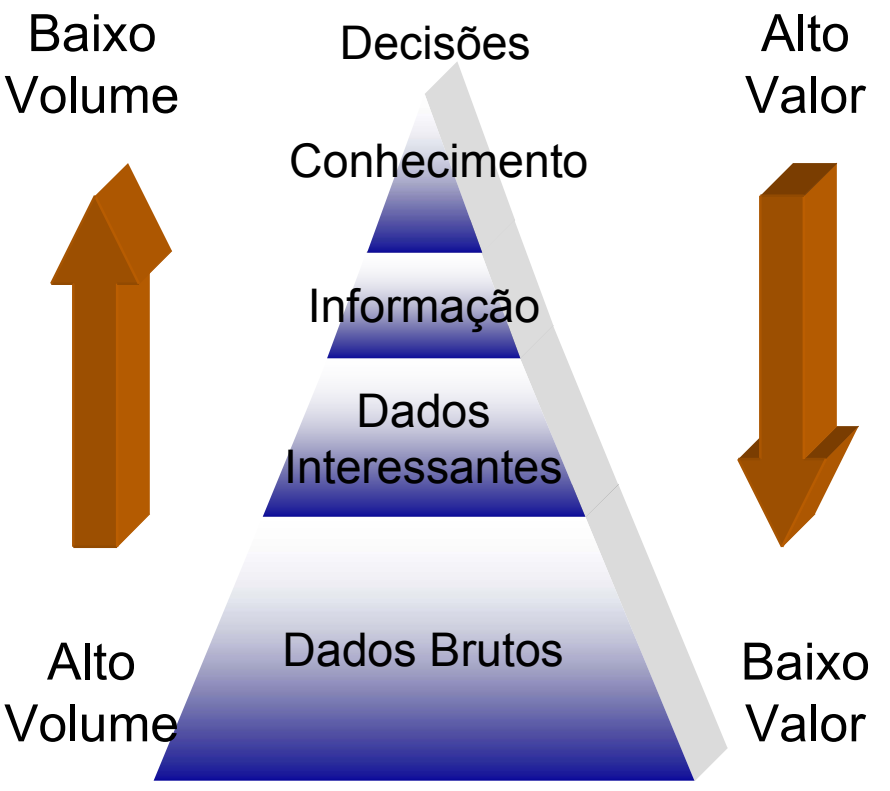
- Lei de Moore: Capacidade de processamento dobra a cada 18 meses (CPU, memória, cache)
- Capacidade de armazenamento dobra a cada 10 meses
- O que estas duas “leis” combinadas produzem?
  - Um *gap* crescente entre nossa habilidade de gerar dados e nossa habilidade de fazer uso dele

“Leis”, Gigantes e Monstros

- Biblioteca do Congresso (EUA)
  - ~10 Terabytes de texto
  - ~3 Petabytes, incluindo vídeo, áudio, etc
- Etimologia
  - Gigabyte (10^9) termo do Latim *Gigas* para **Gigante**
  - Terabye (10^{12}) termo do Grego*Teras* para **Monstro**
  - Próximos prefixos: Peta, Exa e então
    - Zeta (10^{21}): última (letra)
    - Yota (10^{24}): após...
- Em 2000, 11% de toda informação gerada pela humanidade foi gerada em 1999 apenas
- A maior parte da informação nunca vista por um ser humano

Por quê Mineração de Dados?

- Número de fontes de dados tem aumentado de modo exponencial
- Os dados têm a tendência de crescer de modo a preencher seu contêiner
  - Alta dimensão (muitos campos)
  - Muitos registros
  - Novas fontes
- Usuário final usualmente não é um estatístico



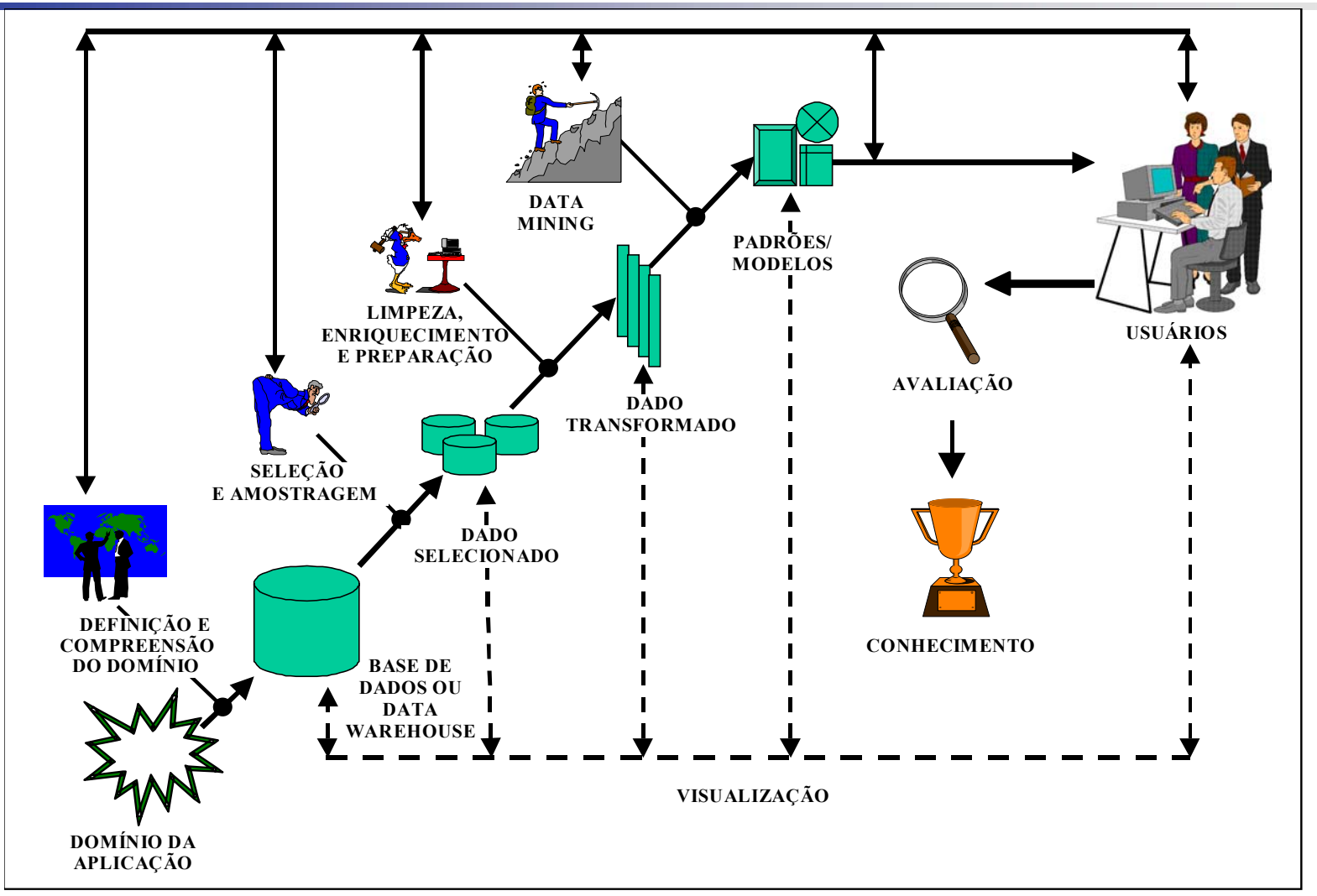
O que é Mineração de Dados?

- **Encontrar estruturas interessantes nos dados**
  - O que é estrutura? Padrões interessantes, modelos preditivos, relacionamentos ocultos
- Exemplos de tarefas abordadas em Mineração de Dados
  - Modelagem Preditiva (classificação, regressão)
  - Segmentação (Clustering)
  - Afinidade (Sumário/Resumo dos Dados)
    - Relações entre campos, associações, visualização

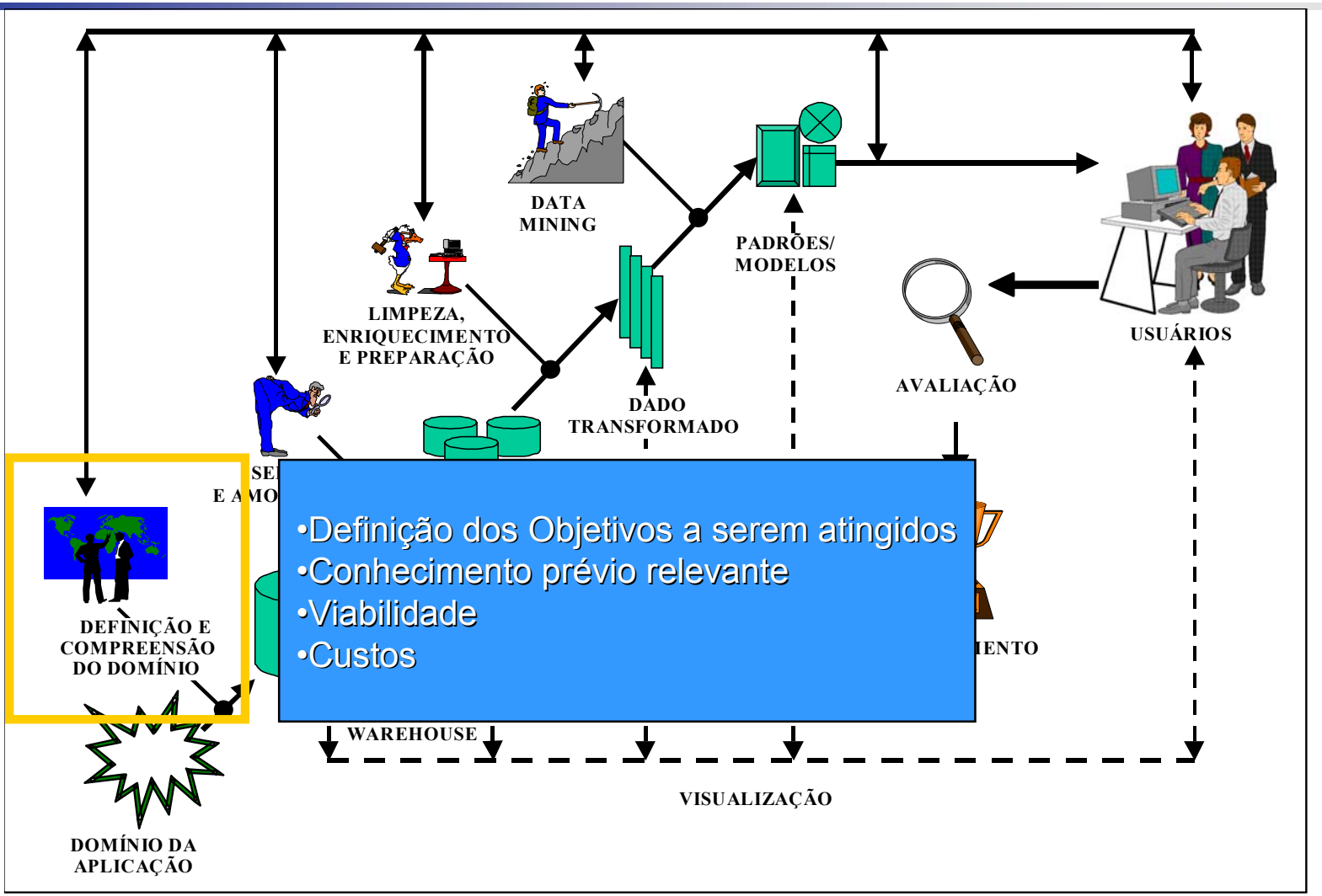
KDD & DM

- KDD – Knowledge Discovery (in Databases): Descoberta de Conhecimento
- DM – Data Mining: Mineração de Dados
- Uma área científica em rápido crescimento
- Um campo multidisciplinar:
  - Bancos da dados e data warehousing
  - Métodos de modelagem e visualização de dados
  - Aprendizado de Máquina
  - Estatística
  - Sistemas Especialistas e Aquisição de Conhecimento
- Recursos
  - Fundamentos teóricos/matemáticos
  - Aprendizado de Máquina e Inferência Lógica
  - Estatística e sistema dinâmicos
  - Sistemas gerenciadores de bases de dados

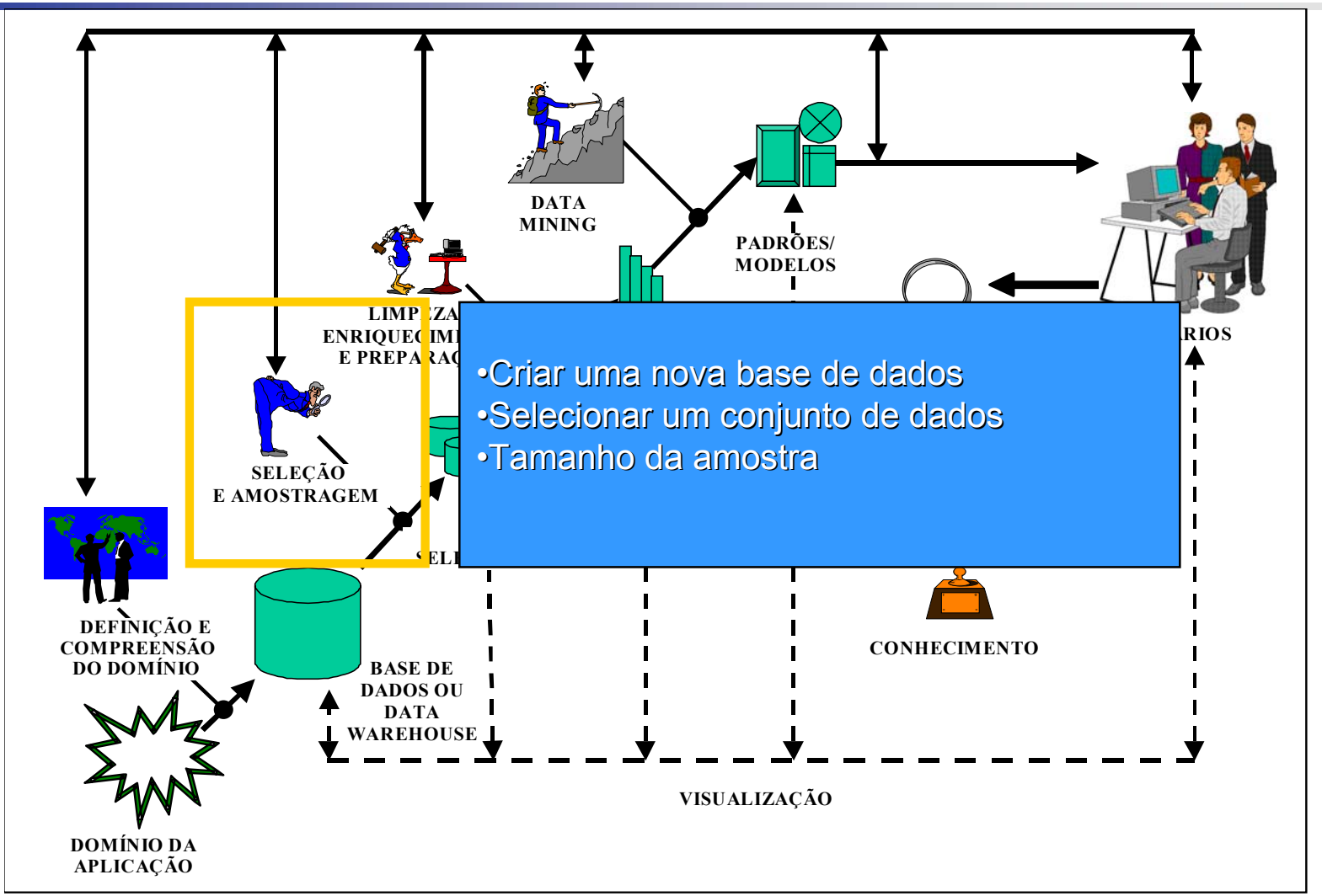
Etapas do Processo de KDD



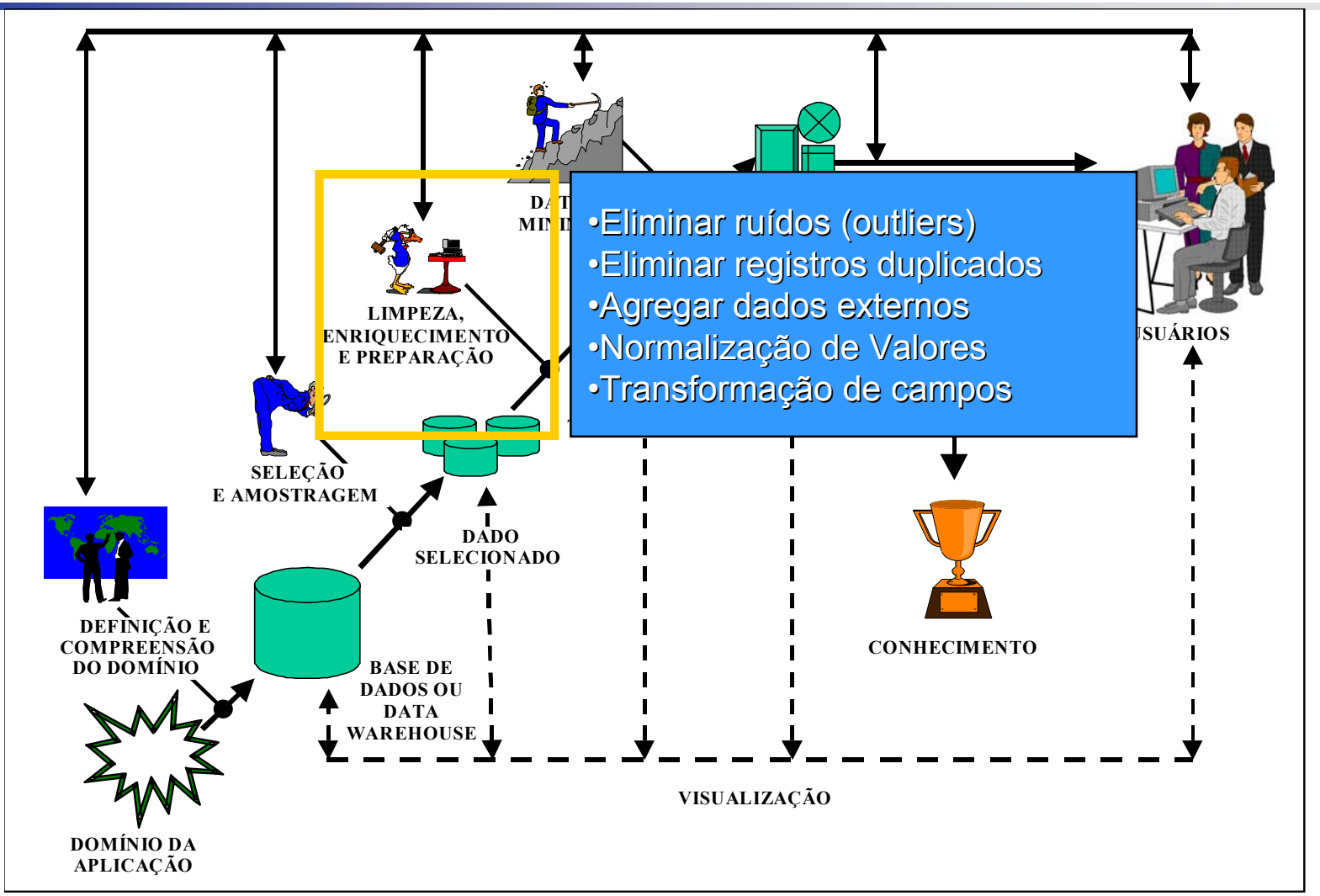
Etapas do Processo de KDD



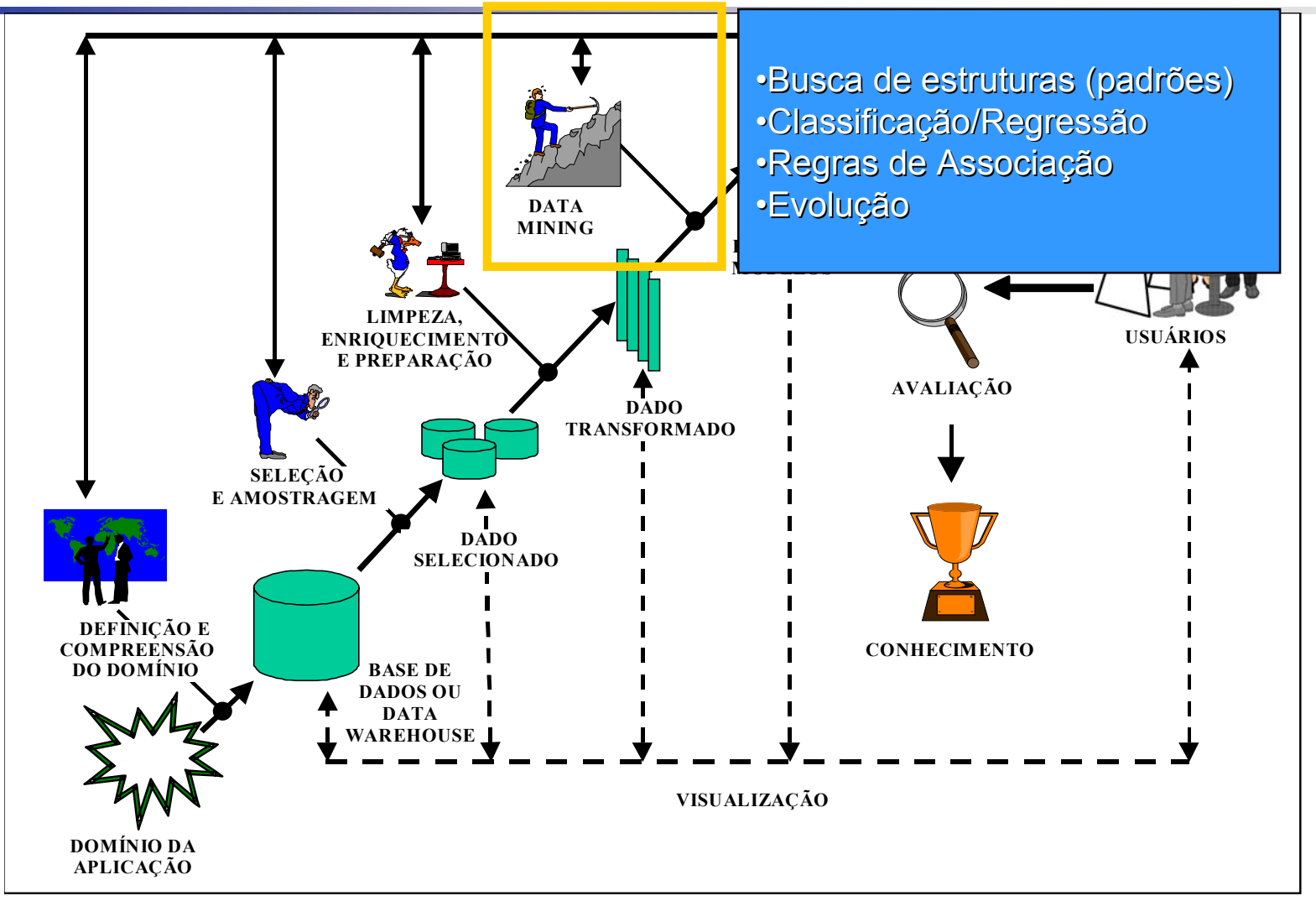
Etapas do Processo de KDD



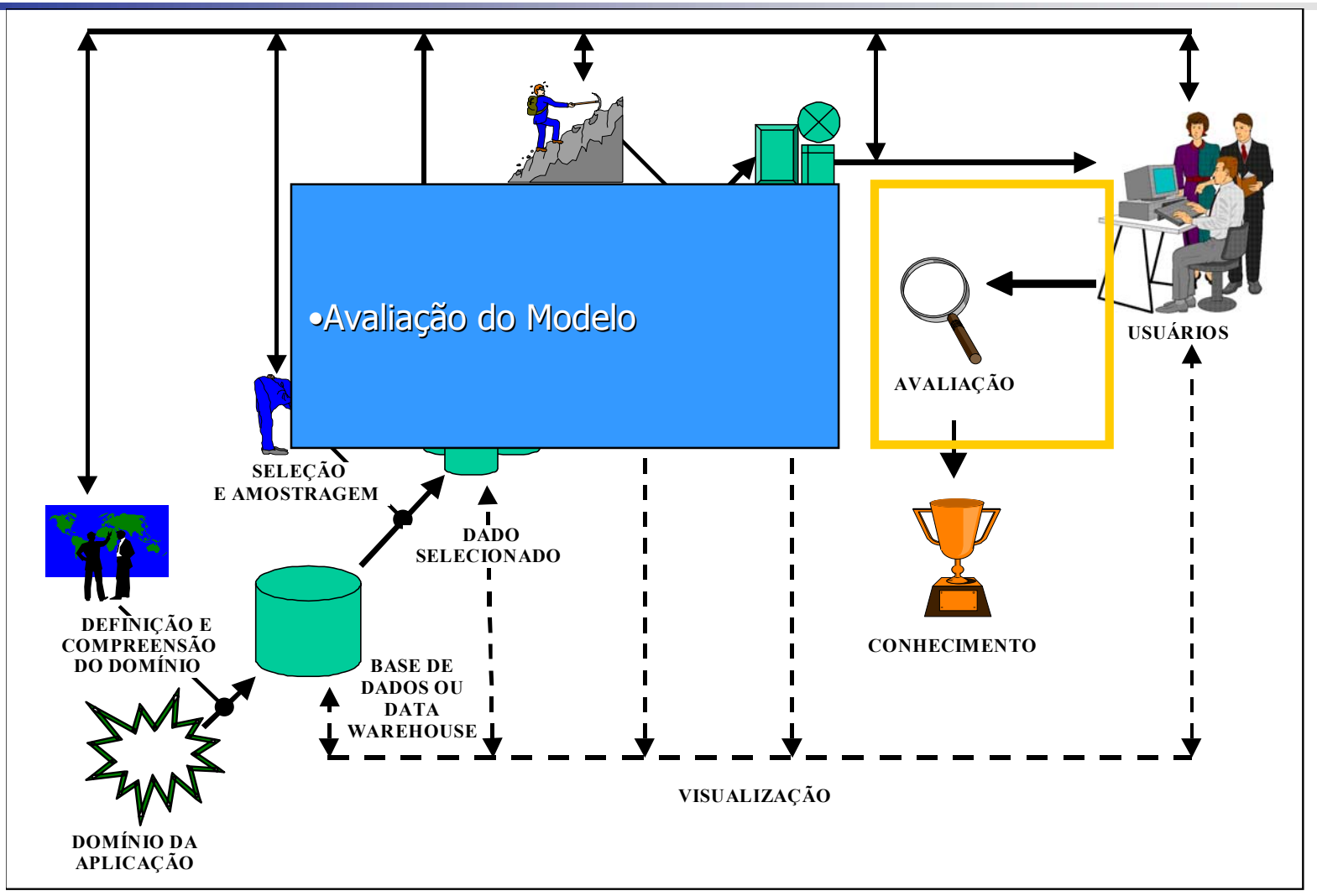
Etapas do Processo de KDD



Etapas do Processo de KDD



Etapas do Processo de KDD



KDD & DM

- KDD: O processo de selecionar e processar os dados que permitam identificar estruturas interessantes:
  - Pré-processamento
    - Preparação de dados
    - Redução de dados
  - Mineração de Dados
  - Pós-processamento ou Análise da Solução
- DM: Uma etapa no processo de KDD
  - Descoberta automática de padrões
  - Desenvolvimento de modelos preditivos e explicativos

KDD

- Resultados Possíveis:
  - Confirmação do óbvio
  - Conhecimento novo
  - Nenhum relacionamento encontrado (dados aleatórios)
- Problemas:
  - Identificação dos dados relevantes
  - Representação dos dados
  - Busca por modelos ou padrões válidos

Pré-Processamento

- Preparação
  - Especificação do Problema (Objetivos)
  - Qualidade dos Dados
  - Definição de Atributos
  - Extração e Integração
  - Transformação de Dados
  - Limpeza
  - Composição de Atributos
- Redução

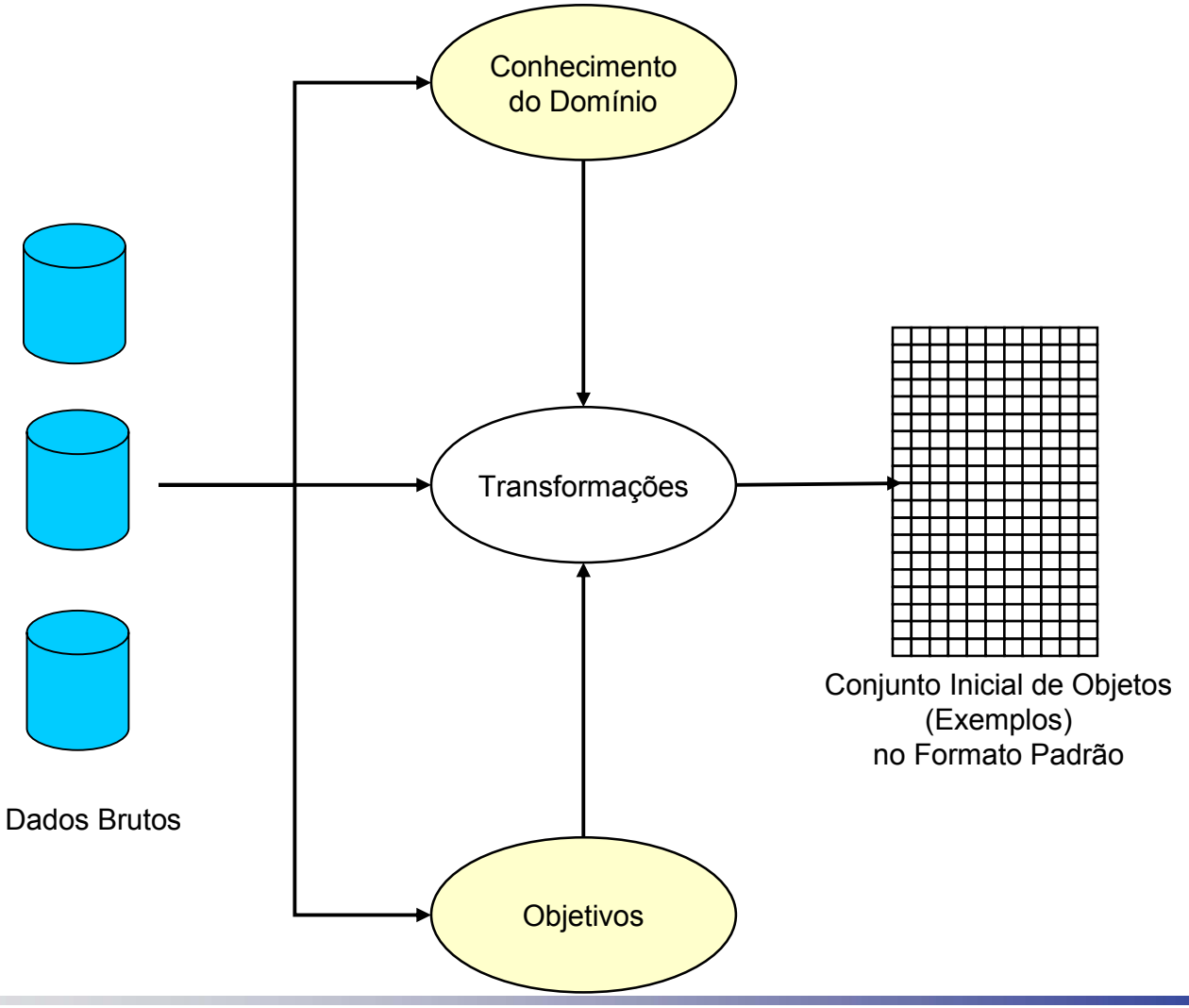
Especificação do Problema

- Solucionar o(s) problema(s) correto(s)
- Definição precisa do problema
  - Problema solucionável pela análise de dados
- Considerar tempo-de-vida da solução
  - Soluções devem se adaptar ao longo do tempo
  - Solução será utilizada uma vez e descartada
- Identificar a entidade de interesse = objeto
  - Paciente
  - Gene
- Maiores detalhes em (Pyle, 1999)

Qualidade dos Dados

- *Missing* values
- Ruído (dados incorretos, dados redundantes)
- Ramificações
  - Pode não ser possível descobrir conhecimento, porque não há padrões estatisticamente significantes nem relações que caracterizam os dados minerados
  - O conhecimento descoberto é inconsistente com o conhecimento prévio extraído
  - Os padrões descobertos são muito específicos ou muito genéricos; em todo caso, eles não são úteis
  - O conhecimento extraído pode levar à decisões incorretas
- Assegurar a qualidade dos dados pode consumir entre 20-40% de todo processo de KDD

Preparação de Dados



Definição de Atributos

- Com base nos dados brutos e no conhecimento prévio do domínio, é necessário definir quais atributos são importantes para atingir a meta do processo de KDD
- A definição dos atributos inicialmente é efetuada de forma manual, quando o especialista humano seleciona um subconjunto do total de atributos disponíveis nos dados brutos
- Como isso implica que muitas decisões de um ser humano estão envolvidas, em caso de dúvida, deve-se incluir atributos extras. Isso deve-se ao fato que os algoritmos de aprendizado têm facilidade de lidar com atributos extras, mas possuem dificuldades no processo de compor novos atributos com maior capacidade de predição.

Definição de Atributos

- Escolha dos atributos depende da tarefa de modelagem
  - Análise Preditiva
    - Atributos independentes (entrada)
    - Atributo(s) meta
  - Segmentação/Clustering
    - Atributos são escolhidos para “definir” similaridade entre objetos
  - Resumo (itemsets freqüentes, regras de associação)
    - Atributos = itens de interesse

Extração e Integração

- Os dados brutos podem se encontrar sob diferentes formas de armazenamento: arquivos, base de dados ou dataware house
- Assim, é necessário realizar a extração e integração dos dados provenientes de diferentes fontes em diferentes formatos, para o formato padrão
- No caso de dados relacionais, isso pode requerer a junção ou projeção de várias tabelas com relações de diferentes cardinalidades (um-para-muitos ou muitos-para-muitos) em uma única tabela

Construção de um Dataset

- Objeto = entidade de interesse
- Objeto = exemplo = caso = registro = linha
- Construção do dataset = coletar/calcular atributos (campos) que descrevem o objeto
  - Conhecimento específico do domínio é benéfico
  - Evitar atributos dependentes ou redundantes

Representação dos Objetos

- Cada objeto (dado) é representado usualmente por um vetor de **atributos**
- Tipos de Atributos
  - Numérico (inteiro, real)
  - Categórico (booleano, conjunto de valores)
- Por exemplo: Amostra de dados clínicos
  - Objeto: Paciente
  - Atributos:
    - Idade (atributo numérico: inteiro)
    - Peso (atributo numérico: real)
    - Sexo (atributo categórico: masculino, feminino)
    - Cor da pele (atributo categórico: branca, marrom, amarela, preta)
    - Doente? (atributo booleano: Sim, Não)

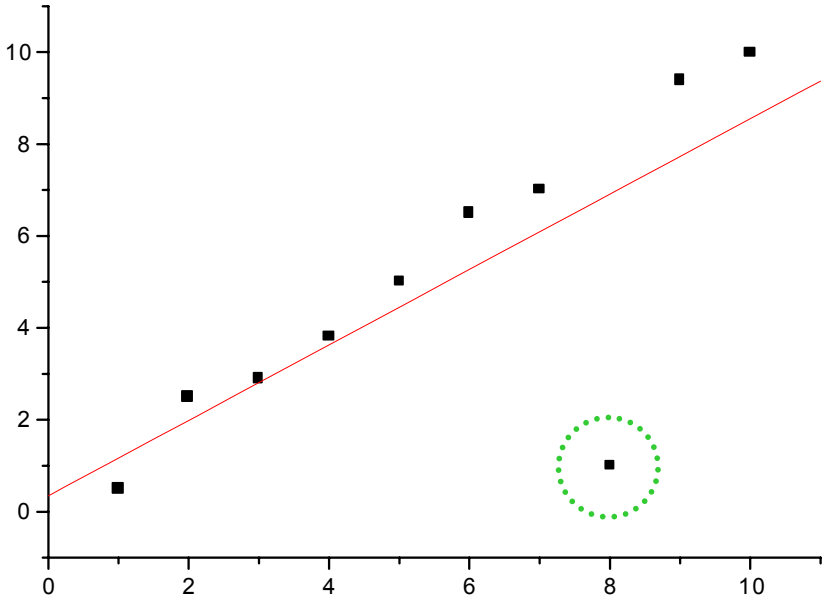
Transformação de Dados

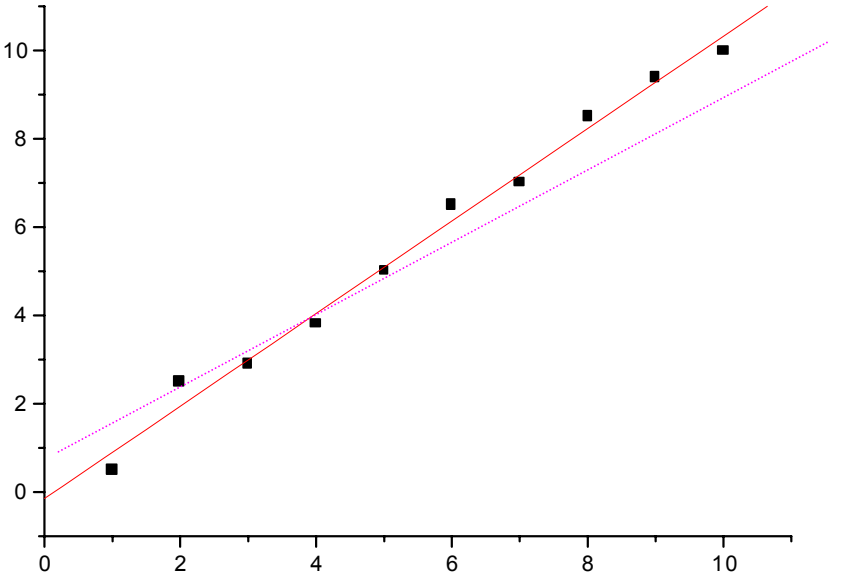
- Resumo de dados
  - dados exames individuais podem ter sido armazenados, mas um resumo diário talvez seja mais indicado para a tarefa em questão
- Transformação de tipos de dados
  - um algoritmo de aprendizado pode não ser capaz de lidar com atributos do tipo data, o que pode requerer que este atributo seja transformado no número inteiro de segundos a partir de uma determinada data inicial ou em períodos, tais como semanas, meses ou anos
- Normalização de valores
  - embora os dados no formato padrão possam ser usados por uma variedade de algoritmos, alguns deles podem requerer dados normalizados de forma a obter melhores resultados; neste caso, os dados são colocados em um intervalo específico de valores, por exemplo, entre -1 e +1

Limpeza

- De forma geral, elas podem ser divididas em dois grupos de tarefas:
  - tarefas específicas do domínio: verificação de consistência dos atributos, remover repetições indevidas
  - tarefas independentes do domínio: fornecer/definir *missing values*, remoção de ruído, tratamento de conjuntos de exemplos não balanceados, seleção de um subconjunto de atributos, construção de atributos

Limpeza





Composição de Atributos

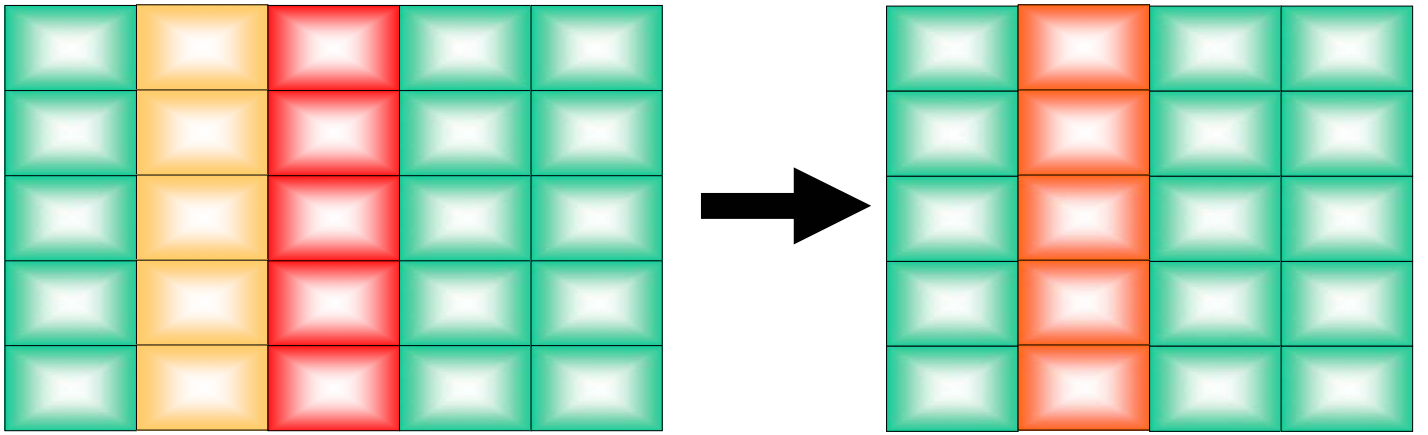
- Em alguns casos, existem transformações adicionais que podem apresentar um impacto muito grande nos resultados
- Neste sentido, a composição de atributos é um fator determinante na qualidade dos resultados, muito maior do que o próprio método de mineração adotado para produzir os resultados
- Em muitos casos, a composição de atributos é dependente do domínio da aplicação

Composição de Atributos

- Definição: processo de construção de novos atributos diretamente relevantes a partir de atributos iniciais (atributos primitivos)
- Pode ser interessante aplicar a Composição de Atributos antes da utilização de métodos de seleção de atributos (FSS), de modo que atributos possivelmente relevantes não sejam descartados

Composição de Atributos

- Combinação de atributos (AM Construtivo)



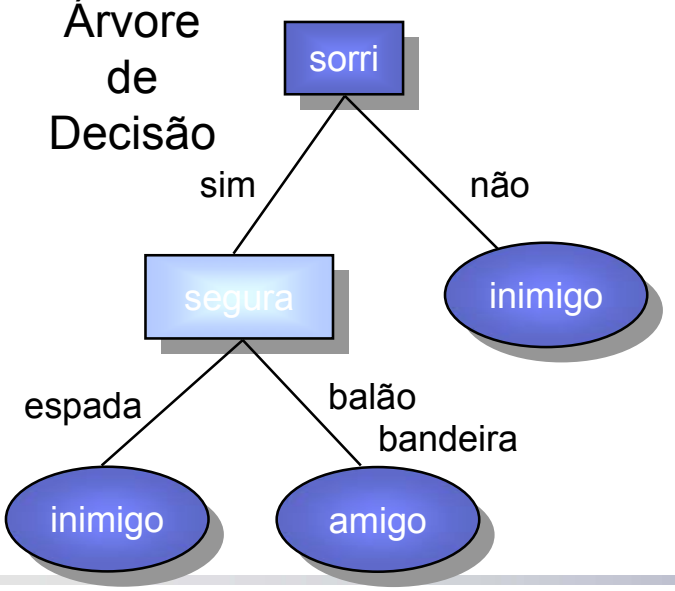
Exemplo de Robôs Amigos e Inimigos

Atributo-valor					classe
sorri	*segura*	*tem-gravata*	*cabeça*	*corpo*	
sim	balão	sim	quadrada	quadrada	amigo
sim	bandeira	sim	triangular	triangular	amigo
sim	espada	sim	redonda	triangular	inimigo
sim	espada	sim	quadrada	redonda	inimigo
não	espada	não	triangular	quadrada	inimigo
não	bandeira	não	triangular	redonda	inimigo

Exemplo de Robôs Amigos e Inimigos

Atributo-valor					classe
sorri	*segura*	*tem-gravata*	*cabeça*	*corpo*	
sim	balão	sim	quadrada	quadrada	amigo
sim	bandeira	sim	triangular	triangular	amigo
sim	espada	sim	redonda	triangular	inimigo
sim	espada	sim	quadrada	redonda	inimigo
não	espada	não	triangular	quadrada	inimigo
não	bandeira	não	triangular	redonda	inimigo

Árvore de Decisão



Regras:

Se sorri = sim *e* segura = espada
então inimigo.

Se sorri = sim *e* segura = balão ou bandeira
então amigo.

Se sorri = não
então inimigo.

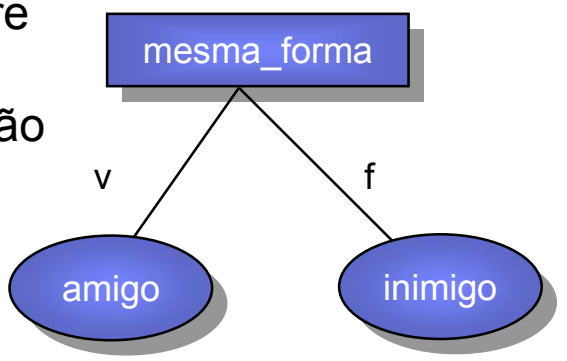
Exemplo de Robôs com o Atributo *mesma-forma*

Atributo-valor						classe
sorri	*segura*	*tem-gravata*	*cabeça*	*corpo*	*mesma_forma*	
sim	balão	sim	quadrada	quadrada	v	amigo
sim	bandeira	sim	triangular	triangular	v	amigo
sim	espada	sim	redonda	redonda	f	inimigo
sim	espada	não	quadrada	quadrada	f	inimigo
não	espada	não	triangular	triangular	f	inimigo
não	bandeira	não	redonda	redonda	f	inimigo

Exemplo de Robôs com o Atributo *mesma-forma*

Atributo-valor						classe
sorri	*segura*	*tem-gravata*	*cabeça*	*corpo*	*mesma_forma*	
sim	balão	sim	quadrada	quadrada	v	amigo
sim	bandeira	sim	triangular	triangular	v	amigo
sim	espada	sim	redonda	redonda	f	inimigo
sim	espada	não	quadrada	quadrada	f	inimigo
não	espada	não	triangular	triangular	f	inimigo
não	bandeira	não	redonda	redonda	f	inimigo

Árvore de Decisão



Regras:

Se mesma_forma = v

então amigo.

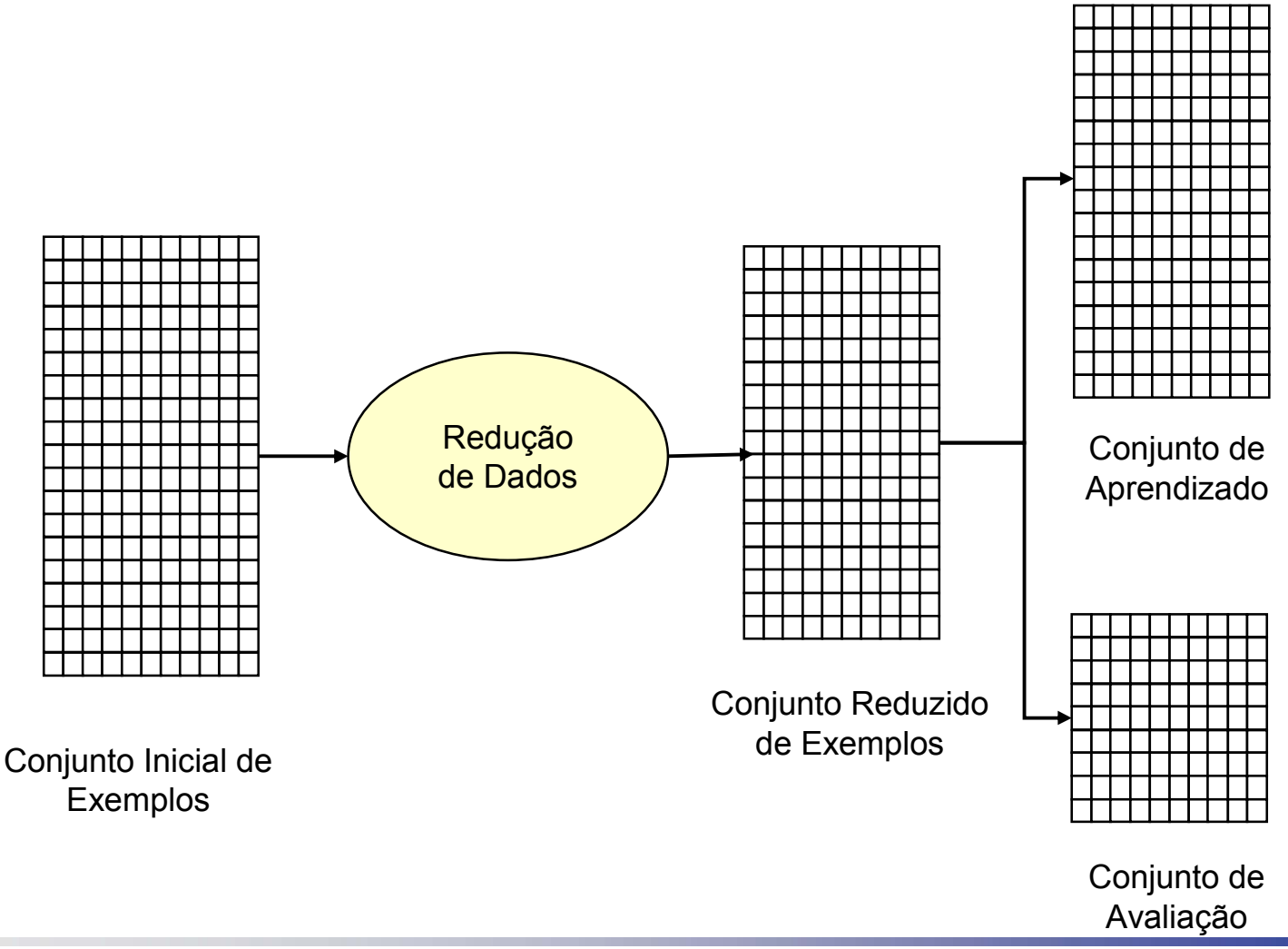
Se mesma_forma = f

então inimigo.

Redução de Dados

- Considerando a etapa de preparação de dados, é possível que uma grande quantidade de dados brutos resulte em um conjunto de exemplos, no formato padrão, de tamanho relativamente moderado
- Neste caso, é possível aplicar algoritmos de mineração diretamente
- Entretanto, para grandes conjuntos de exemplos, é bem provável que a etapa redução de dados seja necessária antes da utilização dos algoritmos de mineração

Redução de Dados

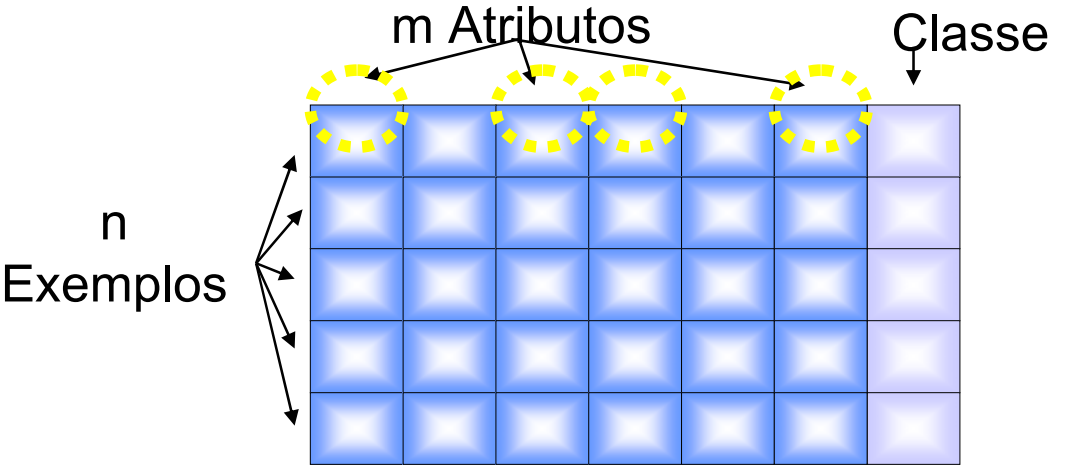


Redução de Dados

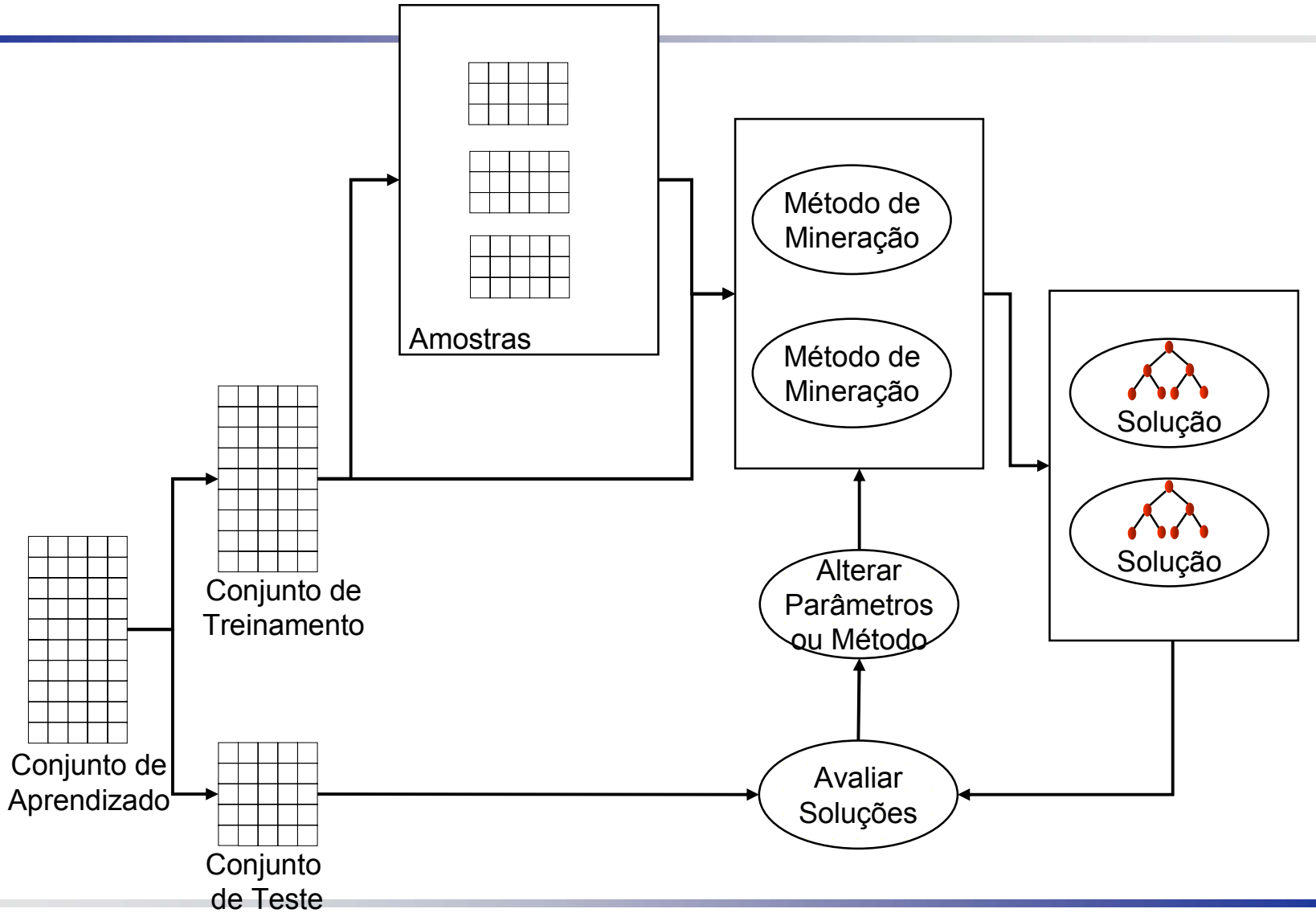
- Redução da dimensão dos dados:
  - remoção de um exemplo;
  - remoção de um atributo (maior impacto);
  - redução do número de valores de um atributo (suavizar, discretizar ou agrupar valores de um atributo)
- Estas operações tentam preservar a característica dos dados originais pela eliminação daqueles não essenciais, suavizando ou discretizando algumas características

Seleção de Atributos - FSS

- Objetivo: selecionar um subconjunto de atributos para fornecer ao indutor (Feature Subset Selection)
- Motivação:
  - Alguns indutores não trabalham bem com muitos atributos irrelevantes
  - Melhoria da precisão
  - Melhoria da compreensibilidade
- Abordagens:
  - Embutida
  - *Wrapper*
  - Filtro



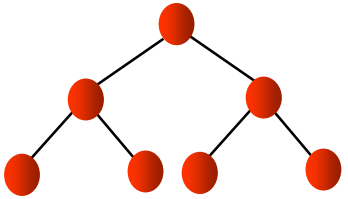
Mineração de Dados



Algoritmos de DM: Componentes

- **Modelo:** contém parâmetros que devem ser determinados a partir dos dados
  - **função** do modelo
    - Classificação/regressão
    - Segmentação (Clustering)
    - Afinidade (Sumário/Resumo dos Dados)
  - **representação** do modelo
- **Critério de preferência:** base para escolha de um modelo ou conjunto de parâmetros sobre outro
- **Algoritmo de busca:** especificação de um algoritmo para encontrar padrões particulares, a partir do modelo e critério de preferência

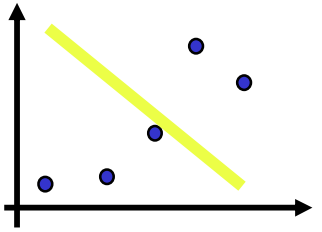
Representação do Modelo



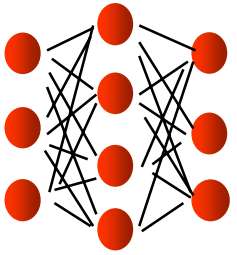
Árvores de decisão

If a = 2
then classe=bom
If b = 2 and c = quente
then classe=ruim

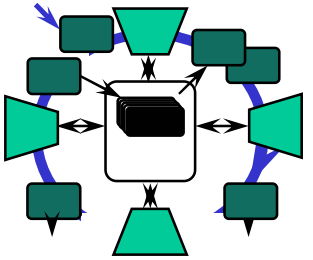
Regras



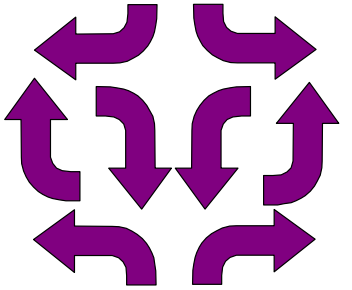
Modelos lineares



Modelos não lineares

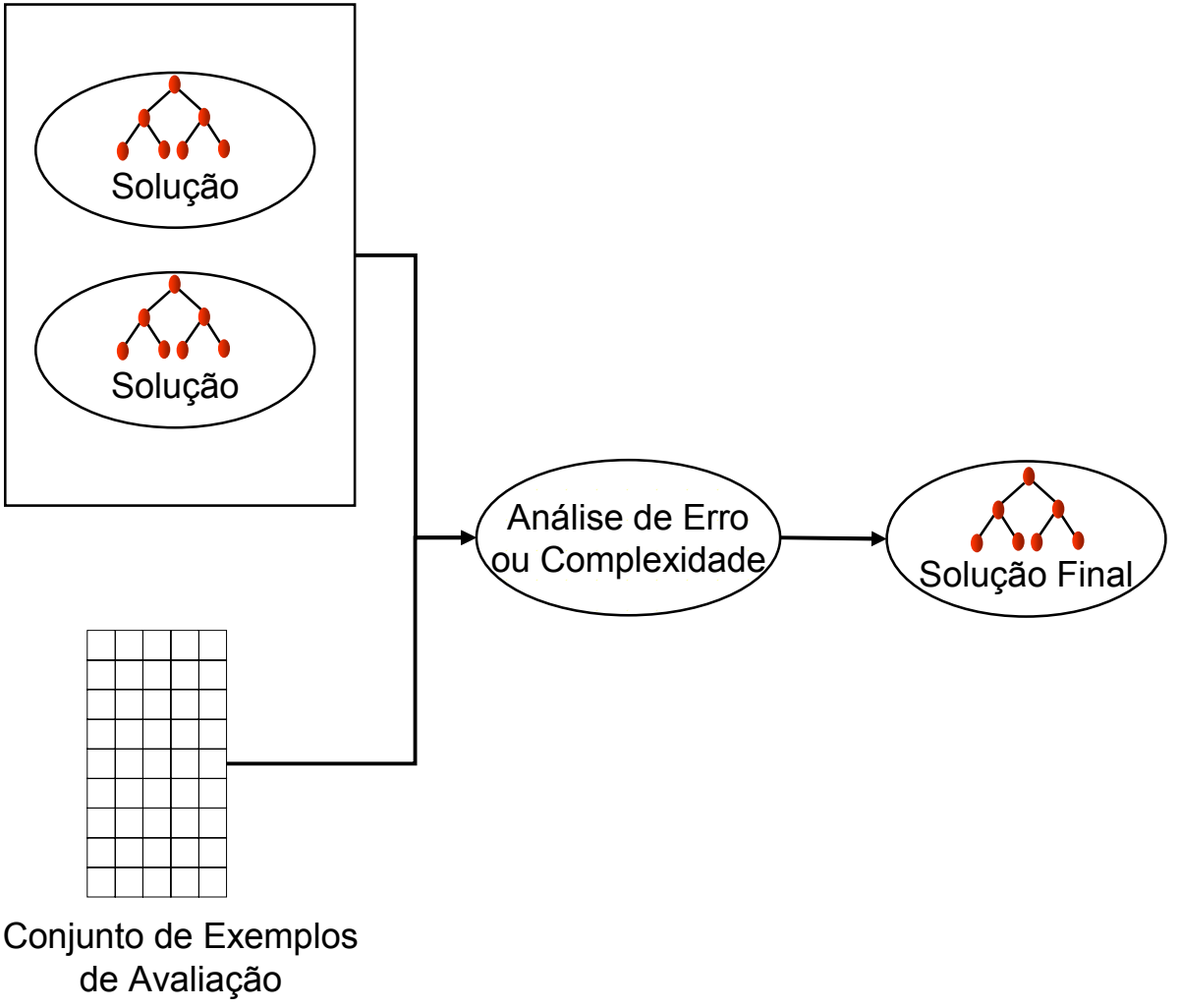


Modelos baseados em distâncias (CBR & k-NN)



Modelos relacionais

Análise da Solução



Análise da Solução

- Interpretação dos resultados: avaliação dos padrões descobertos, visualização dos padrões extraídos, remoção de padrões irrelevantes ou redundantes e tradução de padrões úteis em termos inteligíveis pelos usuários
- Uso do conhecimento extraído: incorporação do conhecimento no desempenho do sistema, tomando ações baseadas no conhecimento ou simplesmente documentando e relatando para as partes interessadas o conhecimento obtido, bem como remoção de conflitos potenciais com conhecimento previamente tido como correto (ou extraído)