Data Mining

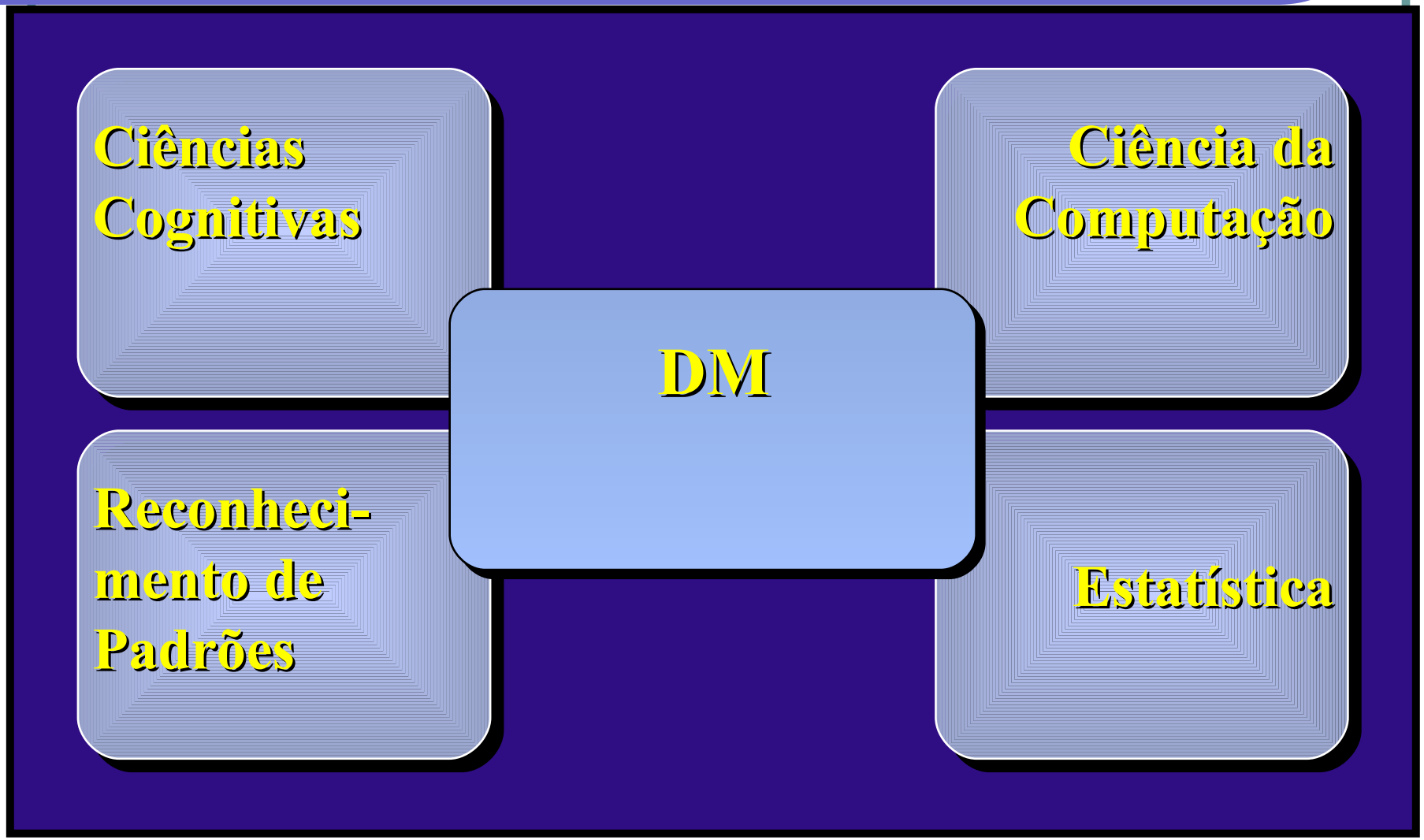
Introdução à *Data Mining*

- Processo de descoberta de conhecimento
  - *Data warehouse* e bancos de dados multidimensionais
  - Pré-processamento: seleção de atributos
  - Pós-processamento: refinamento do conjunto de regras descobertas
- Abordagens para acelerar o *data mining* em grandes bancos de dados
  - Abordagens com redução de dados
  - Abordagens sem redução de dados

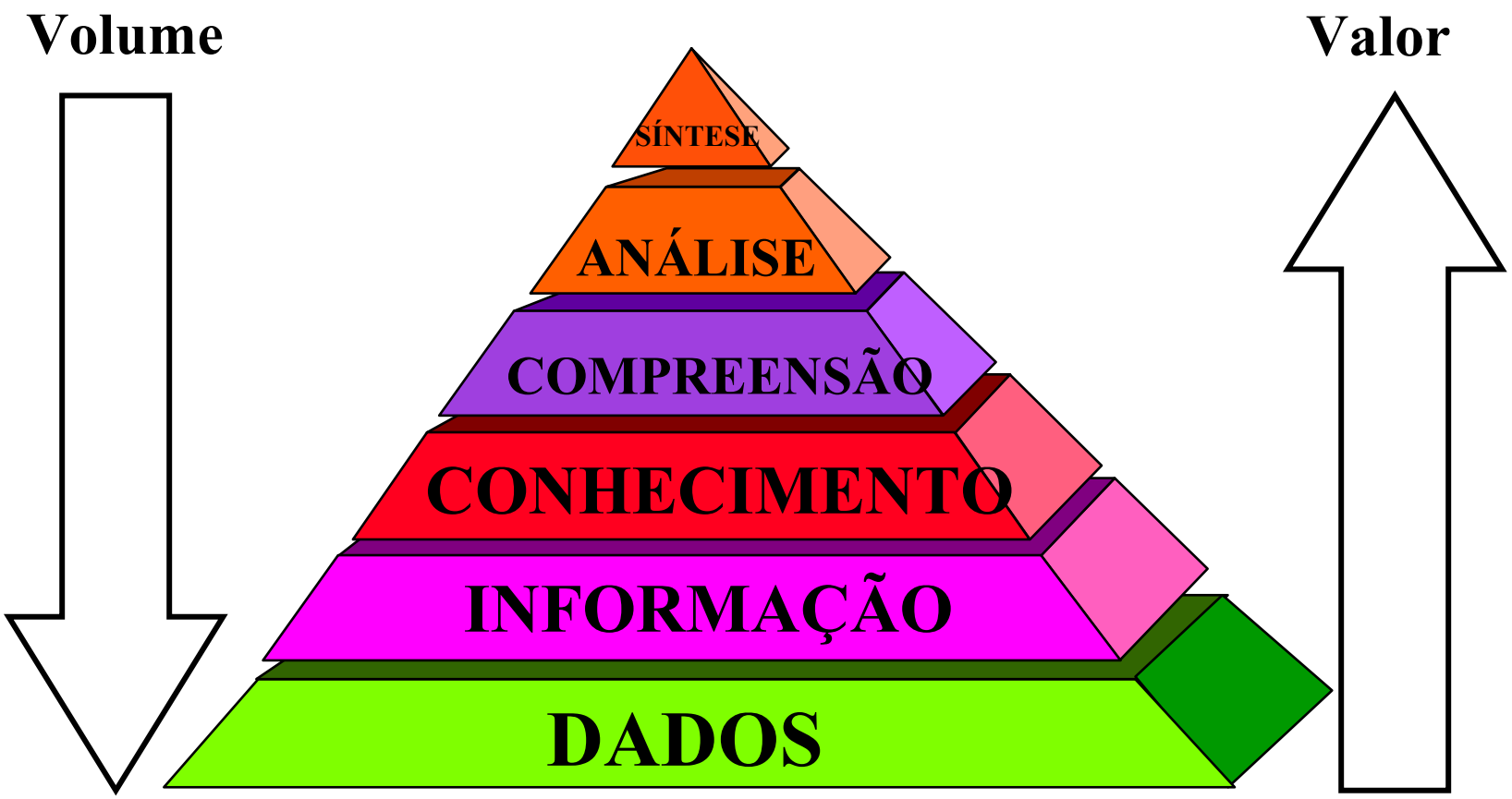
Introdução à *Data Mining*

- Tarefas desempenhadas por sistemas de DM
  - Descoberta de regras de associação
  - Classificação
  - Clustering (agrupamento)
- Métodos para DM
  - Indução de regras
  - Redes neurais
  - Aprendizado baseado em casos (vizinho mais próximo)
  - Algoritmos genéticos

DM incorpora várias técnicas de outras disciplinas



Estrutura


Volume
Valor
SÍNTESE
ANÁLISE
COMPREENSÃO
CONHECIMENTO
INFORMAÇÃO
DADOS

Exemplo: Extração de Informação

- Banco de dados sobre vendas de vários tipos de produtos com dados de clientes e produtos
- Quantos videogames do tipo XYZ foram vendidos para o cliente ABC na data *dd/mm/aaaa*?
- Aplicação: atividades do dia-a-dia da empresa (baixo nível de administração)

Exemplo: Extração de conhecimento

- Quais os clientes que têm alta probabilidade de comprar videogames?

- SE (idade < 18)
  E (profissão = “estudante”)
  ENTÃO (compra = “videogame”) (90%)

- Aplicações:
  - Marketing (mala direta direcionada)
  - Planejamento de estoque/novas filiais
  - e outras decisões estratégicas (alto nível de administração)

Associação

Descoberta de Regras de Associação

- Cada registro corresponde a uma transação de um cliente com itens assumindo valores binários (sim/não), indicando se o cliente comprou ou não o respectivo item

Descoberta de Regras de Associação

No.	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

Descoberta de Regras de Associação

Uma regra de associação é um relacionamento:

SE (X) ENTÃO (Y)

onde X e Y são conjuntos de itens

Para cada regra são atribuídos 2 fatores:

$$\text{Suporte (Sup.)} = \frac{\text{No. de registros com X e Y}}{\text{No. Total de registros}}$$

$$\text{Confiança (Conf.)} = \frac{\text{No. de registros com X e Y}}{\text{No. de registros com X}}$$

Algoritmo

Fase I: Descobrir conjuntos de itens freqüentes

Descobrir todos os conjuntos de itens com suporte maior ou igual ao mínimo suporte especificado pelo usuário

Fase II: Descobrir regras com alto fator de confiança

A partir dos conjuntos de itens freqüentes, descobrir regras de associação com fator de confiança maior ou igual ao especificado pelo usuário

Suporte de conjuntos de itens

Suporte (Sup.) = $\frac{\text{No. de registros com X e Y}}{\text{No. Total de registros}}$

- Passo 1: Calcular suporte de conjuntos com 1 item

Item	Sup
leite	**0,2**
café	**0,3**
cerveja	**0,2**
pão	**0,5**
manteiga	**0,5**
arroz	**0,2**
feijão	**0,2**

Itens freqüentes (SUP ≥ 0,3): café, pão, manteiga

Suporte de conjuntos de itens

Suporte (Sup.) = $\frac{\text{No. de registros com X e Y}}{\text{No. Total de registros}}$

- Passo 2: Calcular suporte de conjuntos com 2 itens

- Otimização: Se um item *I* não é freqüente, um conjunto com 2 itens, um dos quais é o item *I*, não pode ser freqüente. Logo, conjuntos contendo item *I* podem ser ignorados

Item	Sup
café, pão	**0,3**
café, manteiga	**0,3**
manteiga, pão	**0,4**

Conjuntos de itens freqüentes (SUP ≥ 0,3): {café, pão}, {café,manteiga}, {manteiga,pão}

Suporte de conjuntos de itens

Suporte(Sup.) = $\frac{\text{No. de registros com X e Y}}{\text{No. Total de registros}}$

- Passo 3: Calcular suporte de conjuntos com 3 itens

- Otimização: Se o conjunto de itens *{I, J}* não é freqüente, um conjunto com 3 itens incluindo os itens *{I,J}* não pode ser freqüente. Logo, conjuntos contendo itens *{I,J}* podem ser ignorados

Item	Sup
café, pão, manteiga	**0,3**

Conjuntos de itens freqüentes (SUP ≥ 0,3): {café, pão,manteiga}

Fator de confiança de regras candidatas

Confiança(Conf.) = $\frac{\text{No. de registros com X e Y}}{\text{No. de registros com X}}$

Conjunto de itens: {café, pão}

SE café ENTÃO pão. Conf = 1,0.

SE pão ENTÃO café. Conf = 0,6.

Conjunto de itens: {café, manteiga}.

SE café ENTÃO manteiga. Conf = 1,0.

SE manteiga ENTÃO café. Conf = 0,6.

Conjunto de itens: {manteiga, pão}.

SE manteiga ENTÃO pão. Conf = 0,8.

SE pão ENTÃO manteiga. Conf = 0,8.

Fator de confiança de regras candidatas

Confiança(Conf.) = $\frac{\text{No. de registros com X e Y}}{\text{No. de registros com X}}$

Conjunto de itens: {café, manteiga, pão}.

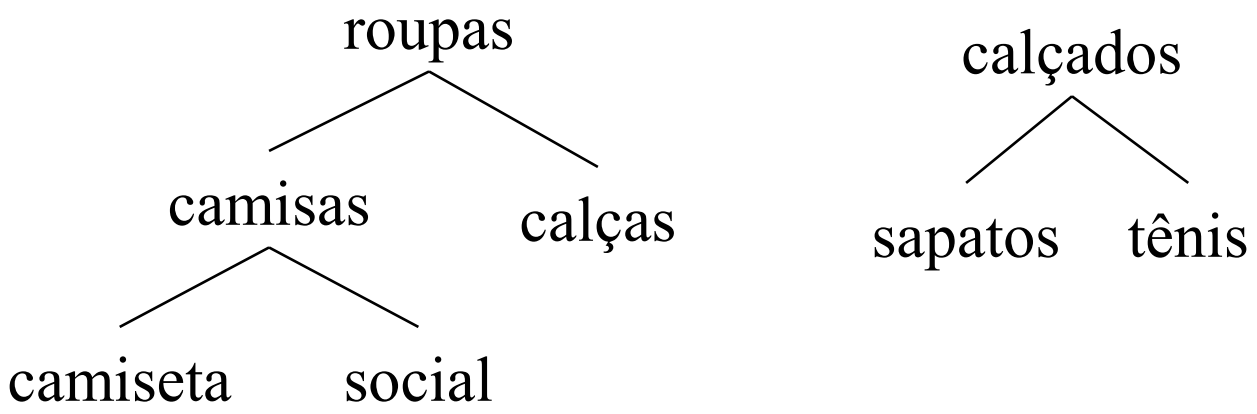
SE café, pão ENTÃO manteiga. Conf = 1,0.
SE café, manteiga ENTÃO pão. Conf = 1,0.
SE manteiga, pão ENTÃO café. Conf = 0,75.
SE café ENTÃO pão, manteiga. Conf = 1,0.
SE pão ENTÃO café, manteiga. Conf = 0,6.
SE manteiga ENTÃO café, pão . Conf = 0,6.

Finalmente, seleciona-se regras com Conf. Maior ou igual ao valor mínimo especificado pelo usuário (ex. 0,8).

Regras de Associação Descobertas
(Sup ≥ 0,3; Conf ≥ 0,8)

Conjunto de itens freqüente: {café, pão}.	Sup = 0,3.
Regra: SE (café) ENTÃO (pão).	Conf = 1.
Conjunto de itens freqüente: {café, manteiga}.	Sup = 0,3.
Regra: SE (café) ENTÃO (manteiga).	Conf = 1.
Conjunto de itens freqüente: {pão, manteiga}.	Sup = 0,4.
Regra: SE (pão) ENTÃO (manteiga).	Conf = 0,8.
Regra: SE (manteiga) ENTÃO (pão).	Conf = 0,8.
Conjunto de itens freqüente: {café, pão, manteiga}.	
Regra: SE café, pão ENTÃO manteiga.	Conf = 1,0.
Regra: SE café, manteiga ENTÃO pão.	Conf = 1,0.
Regra: SE café ENTÃO pão, manteiga.	Conf = 1,0.

Regras de Associação e Vários Níveis Hierárquicos



- Regras em níveis mais baixos da hierarquia podem não ter suporte mínimo
- Heurísticas podem ser usadas para podar regras não-interessantes

Regras de Associação com Restrições

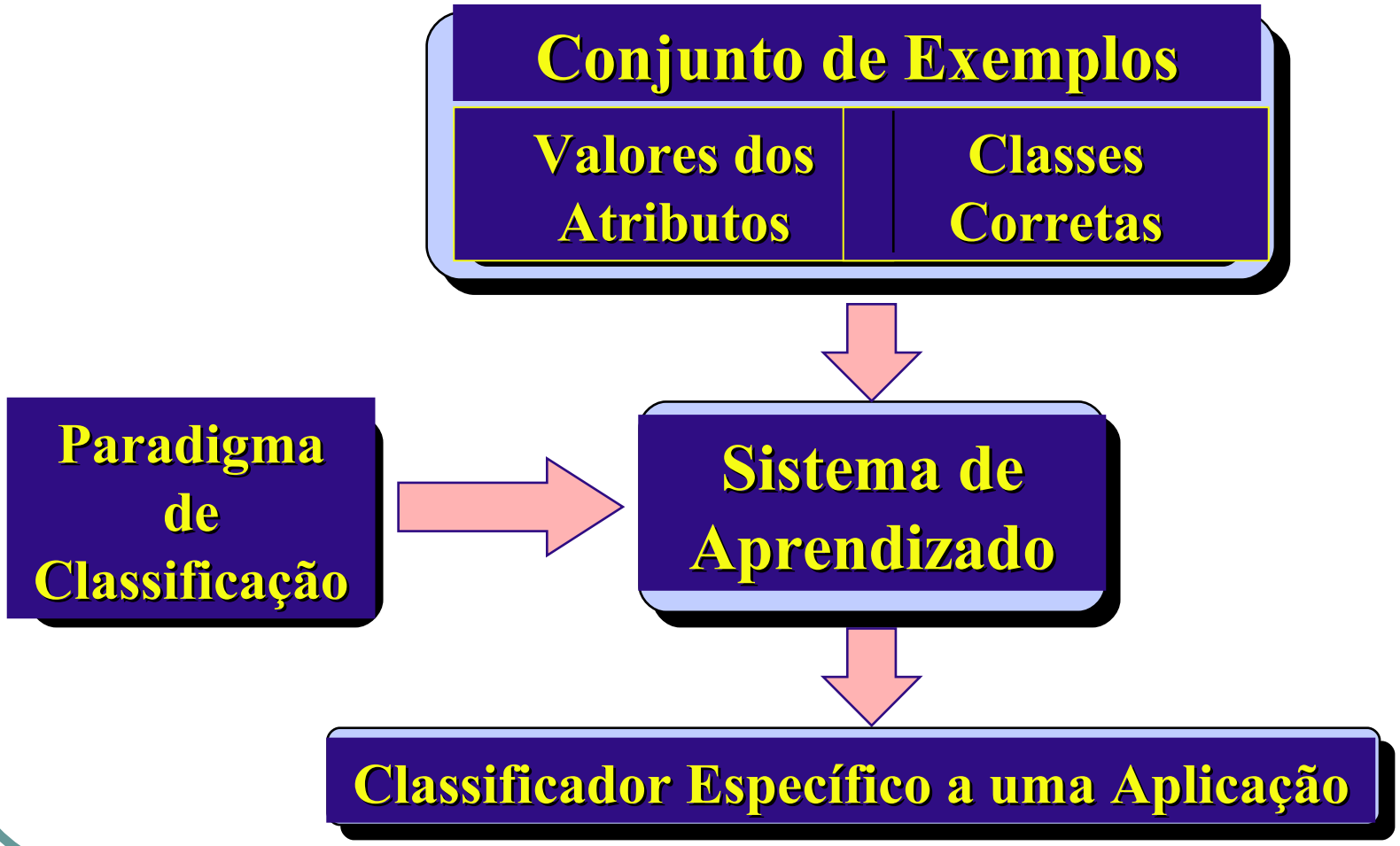
- O usuário pode estar interessado em apenas um subconjunto de associações, referente a produtos mais interessantes para ele
- Em vez de filtrar as regras descobertas, é muito mais eficiente incorporar restrições/heurísticas no algoritmo para descoberta de regras
- A preferência do usuário pode referenciar itens em vários níveis de hierarquias.

Classificação

Classificação

- Cada registro pertence a uma classe, indicada pelo valor de um atributo meta
- Cada registro consiste de:
  - um atributo meta
  - um conjunto de atributos previsores
- **Tarefa:** descobrir um relacionamento entre os atributos previsores e o atributo meta, usando registros cuja classe é conhecida
- **Objetivo:** usar o relacionamento descoberta para prever a classe (o valor do atributo meta) de um registro com classe desconhecida

Representação da Classificação



Método Indutivo

- O paradigma de aprendizado indutivo busca aprender conceitos através de instâncias destes conceitos

Método Indutivo

- O classificador utiliza os conceitos aprendidos para classificar novos exemplos.

Exemplo de Classificação

- Uma editora internacional publica o livro “Guia de Restaurantes Franceses na Inglaterra” em 3 países: Inglaterra, França e Alemanha
- A editora tem um banco de dados sobre clientes nesses 3 países, e deseja saber quais clientes são mais prováveis compradores do livro (para fins de mala direta direcionada)
  - **Atributo meta: comprar (sim/não)**
- Para coletar mais dados: enviar material de propaganda para uma amostra de clientes, registrando se cada cliente que recebeu a propaganda comprou ou não o livro

Exemplo de Classificação

Sexo	País	Idade	Comprar (meta)
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

Regras de classificação descobertas a partir dos dados ao lado

SE (País = “Alemanha”)
ENTÃO (Comprar = “NÃO”)

SE (País = “Inglaterra”)
ENTÃO (Comprar = “SIM”)

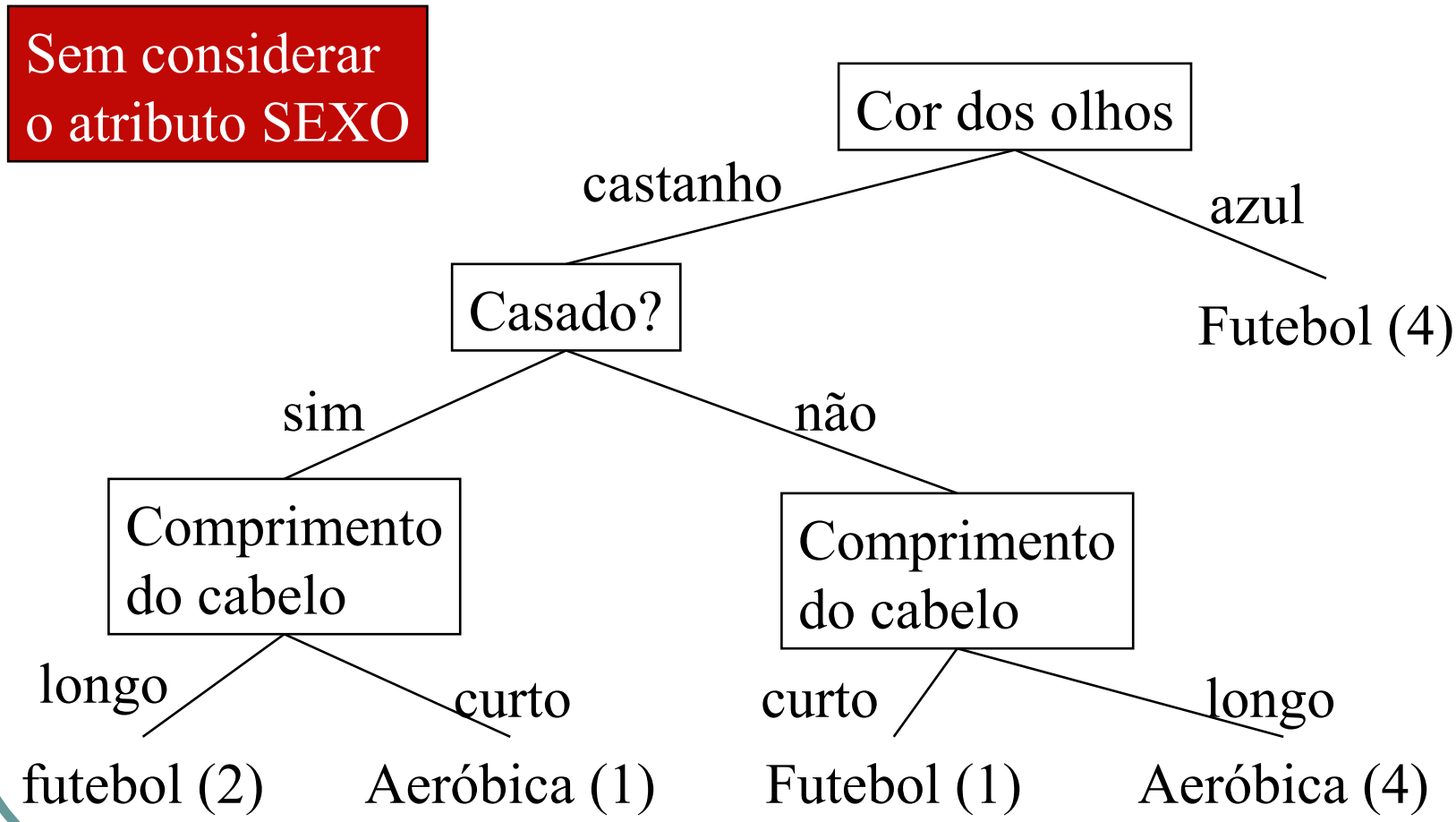
SE (País = “França”) E (Idade ≤ 25)
ENTÃO (Comprar = “SIM”)

SE (País = “França”) E (Idade > 25)
ENTÃO (Comprar = “NÃO”)

Prevendo esporte praticado por estudantes: futebol vs. aeróbica

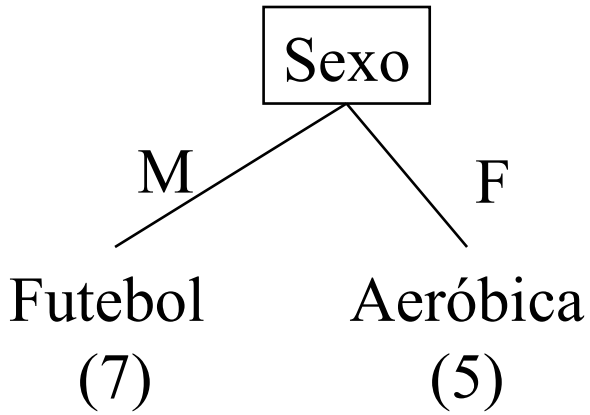
Cor dos olhos	Casado	Sexo	Comprimento do cabelo	Esporte (meta)
castanho	sim	M	longo	futebol
azul	sim	M	curto	futebol
castanho	sim	M	longo	futebol
castanho	não	F	longo	aeróbica
castanho	não	F	longo	aeróbica
azul	não	M	longo	futebol
castanho	não	F	longo	aeróbica
castanho	não	M	curto	futebol
castanho	sim	F	curto	aeróbica
castanho	não	F	longo	aeróbica
azul	não	M	longo	futebol
azul	não	M	curto	futebol

Árvore de decisão para prever esporte



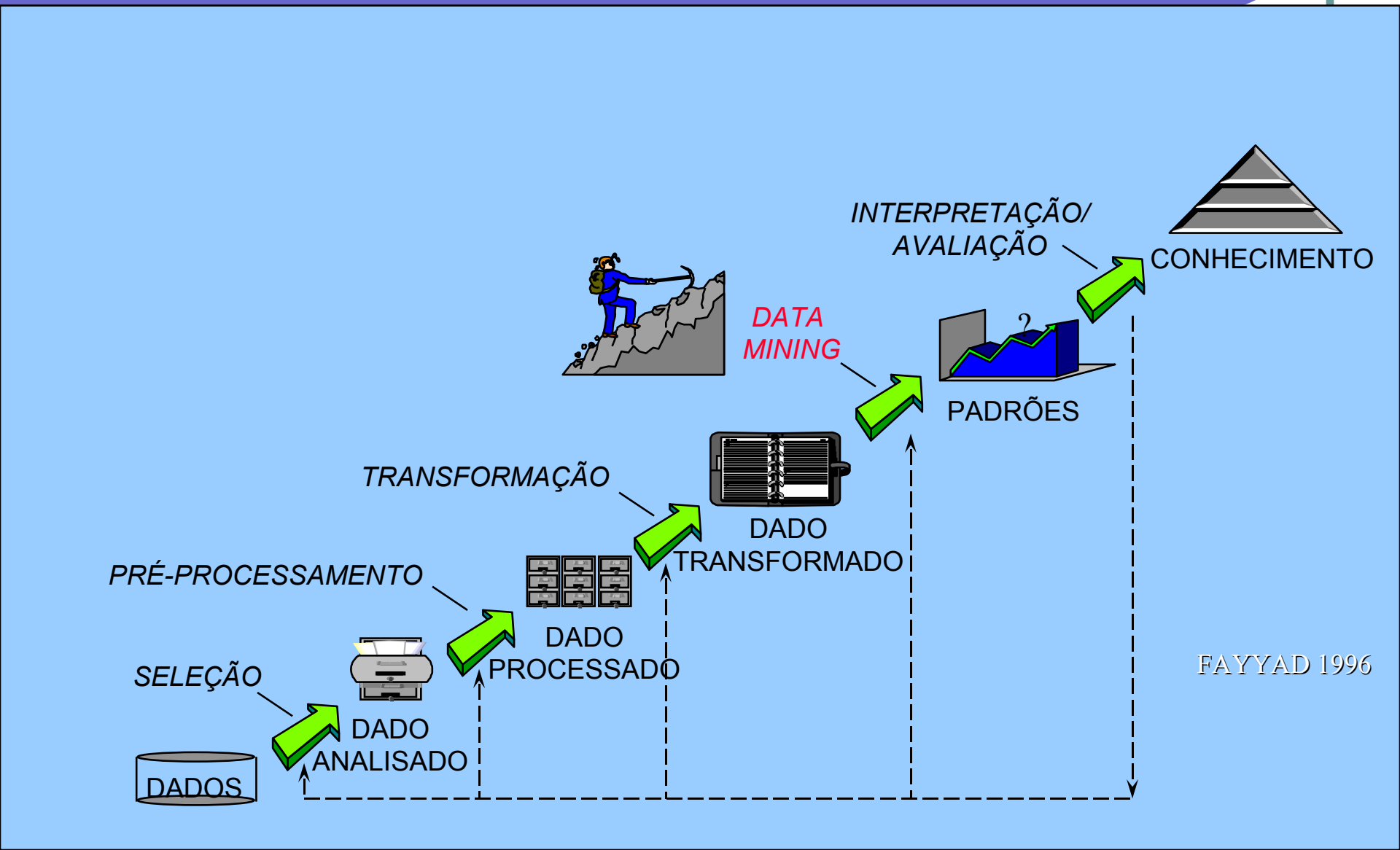
Árvore de decisão para prever esporte

Sem considerar o atributo SEXO



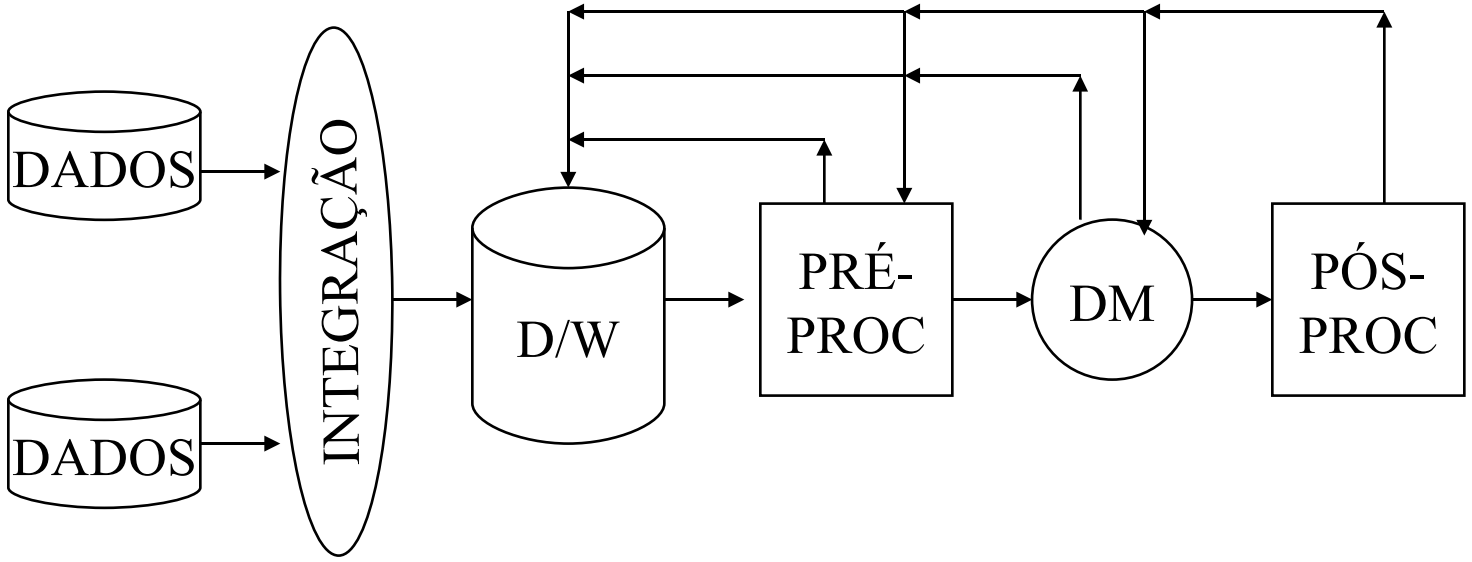
- Se há estudantes do sexo M que praticam aeróbica, mas que não foram incluídos nos dados de treinamento, essa árvore de decisão falhará na previsão para esses estudantes, enquanto a árvore anterior poderia acertar a previsão para eles
- O grau de exatidão das previsões feitas pela árvore depende da natureza do domínio e da qualidade dos dados

Processo de KDD (*Knowledge Discovery in DataBases*) - Etapas



Data Warehouse

Uma visão geral do processo de Descoberta de Conhecimento em Bancos de Dados



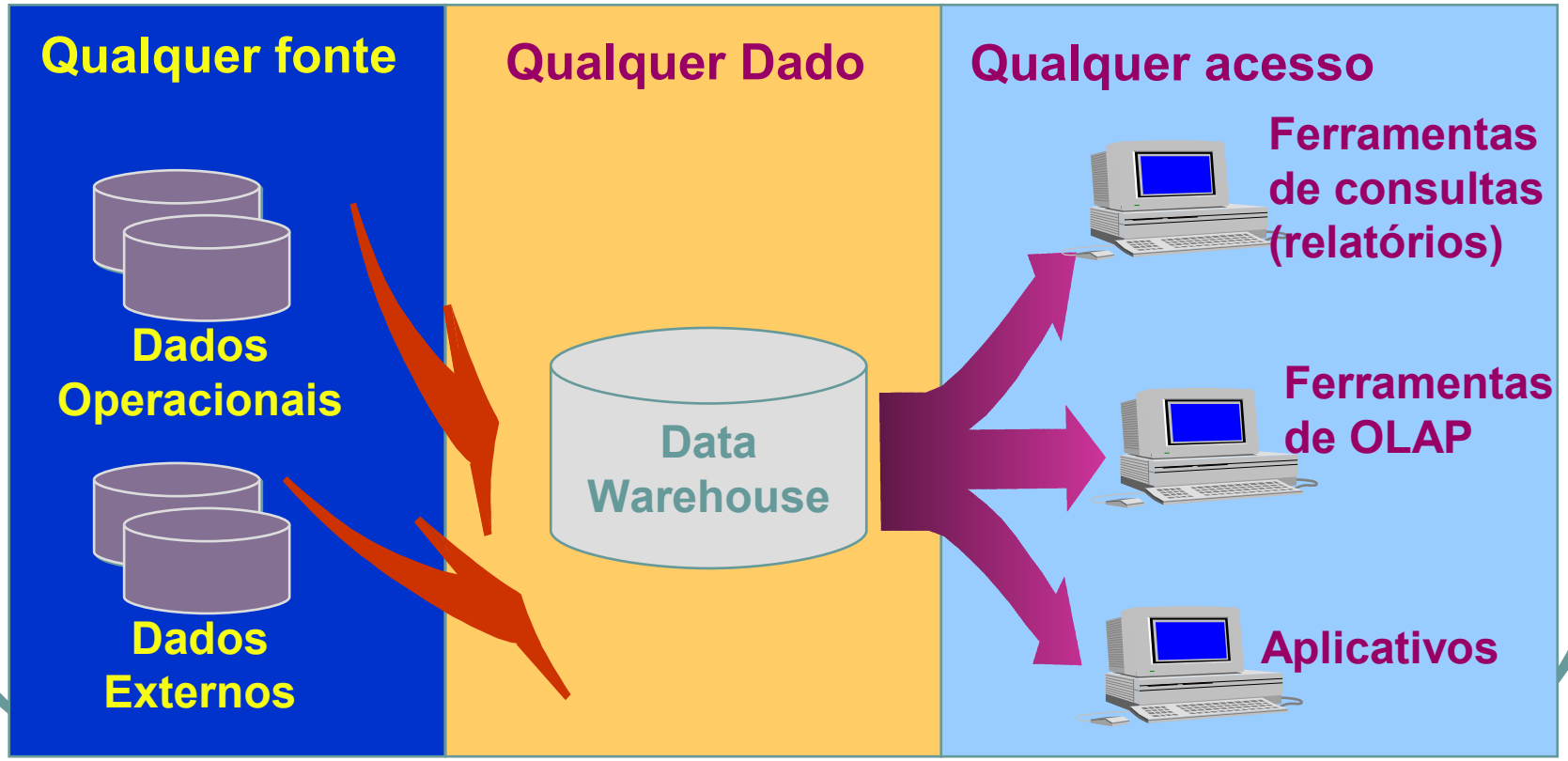
- D/W: Data Warehouse
- PRÉ-PROC: Pré-processamento para DM
- DM: Data Mining
- PÓS-PROC: Pós-processamento dos resultados de DM

Data Warehouse

- É um repositório de dados:
  - Integrados
  - Orientados para análise (alto nível de abstração)
  - Somente-leitura
  - Projetado para ser usado como suporte a sistema de apoio à decisão e sistemas de *data mining*

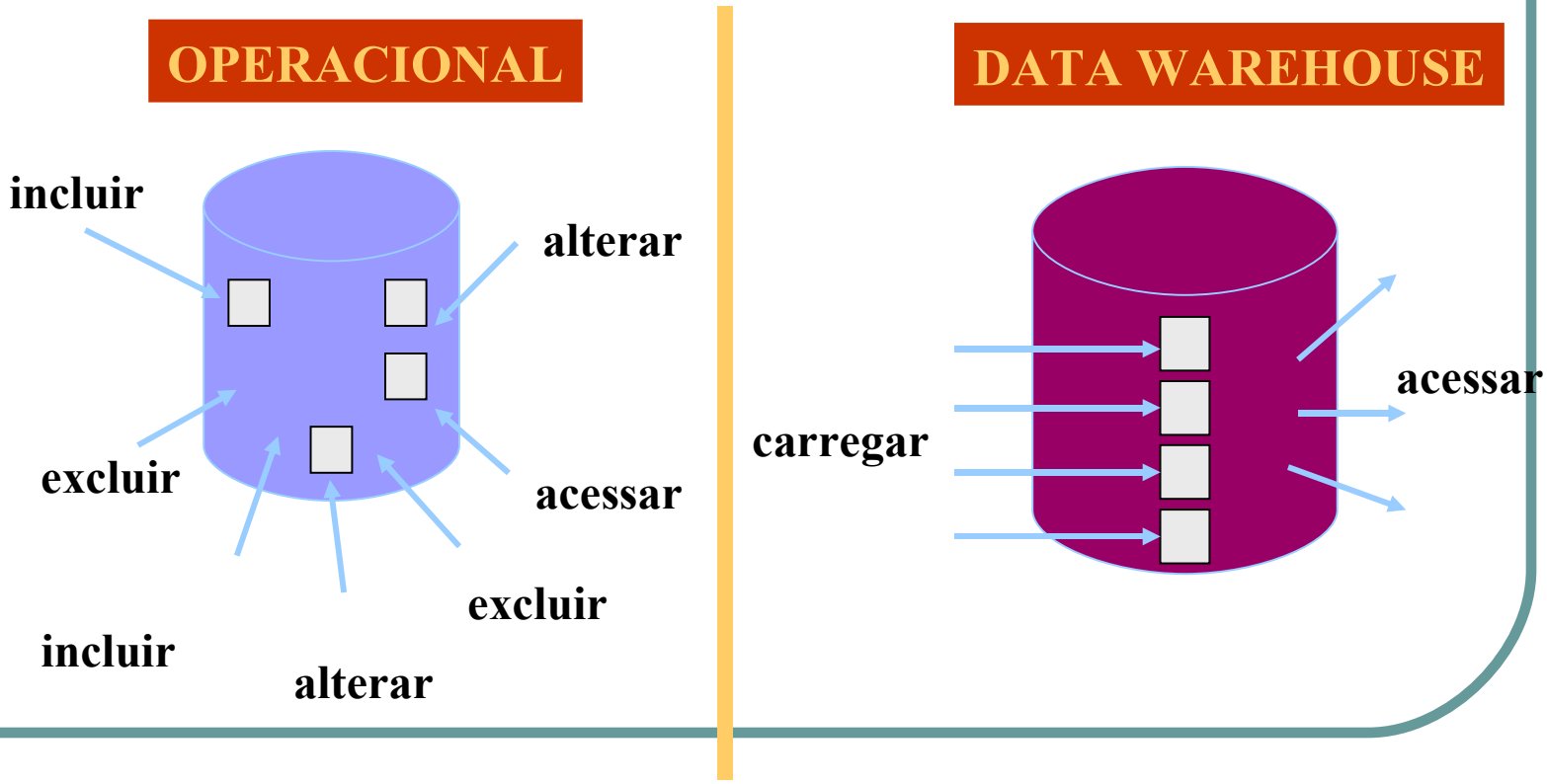
Componente de um Data Warehouse

- Data Warehouse não é o fim, ele é um meio que as empresas dispõem para analisar informações podendo utilizá-las para a melhoria dos processos atuais e futuros



Data Warehouse

Não volatilidade



Trabalho

- Baseado na tabela a seguir (utilizando classificação)

inteligência	beleza	situação financeira	classe
sim	bonito	rico	namorar_sim_namora
não	feio	pobre	namorar_não_namora
sim	feio	pobre	namorar_sim_namora
sim	feio	mediano	namorar_sim_namora
não	bonito	pobre	namorar_não_namora
não	bonito	mediano	namorar_sim_namora
não	bonito	rico	namorar_sim_namora
não	feio	rico	namorar_sim_namora

Trabalho

- Montar a árvore de decisão
- Montar as regras
- Criar a base de dados em formato para executar no *Weka*
- Mostrar a árvore de decisão gerada pelo Weka
- Verificar a confiabilidade da classificação através da matriz de confusão
- Inserir novos registros na base de dados e executar novos testes