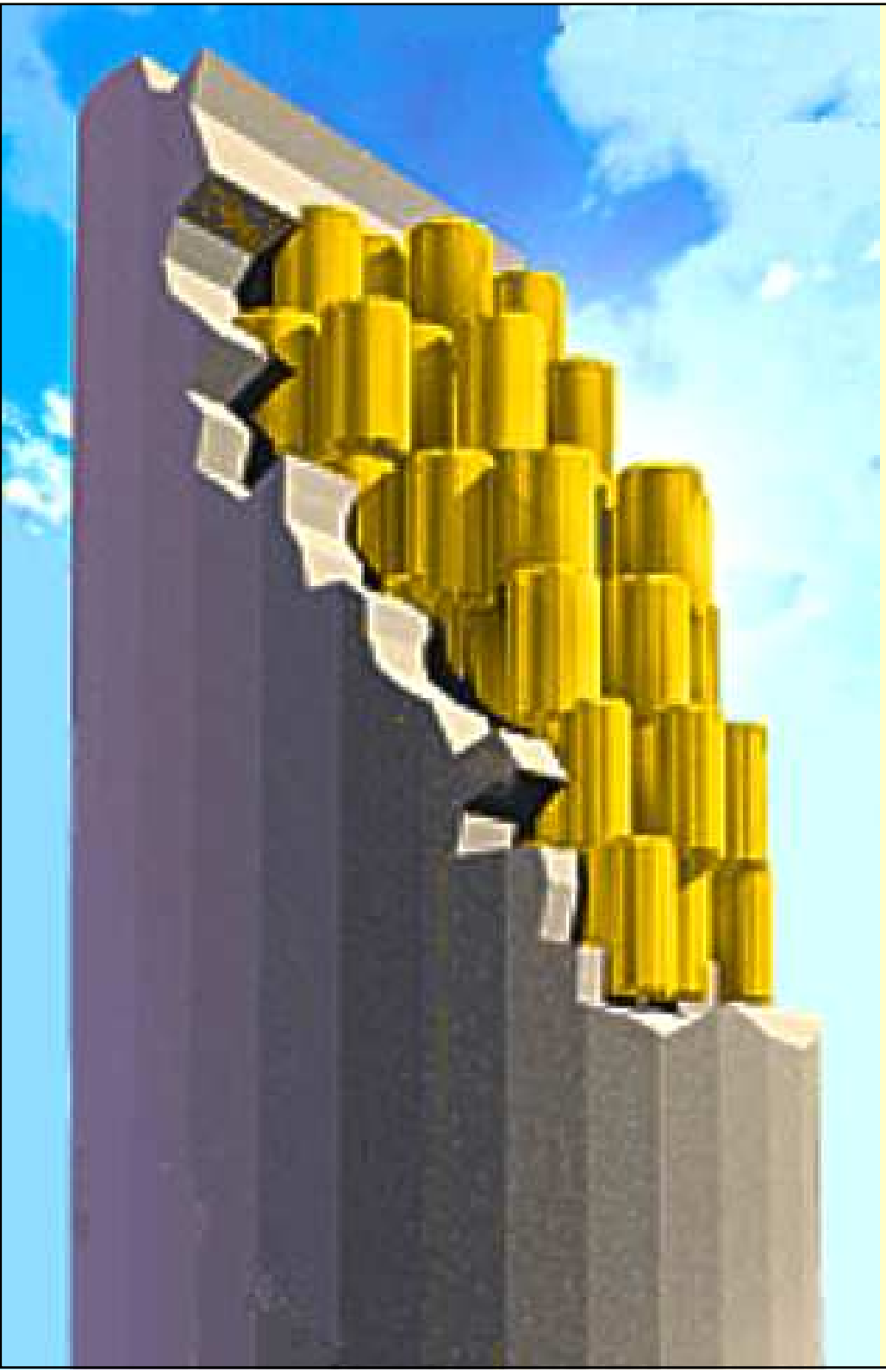


INFOIMAGEM 2002

Princípios

Essenciais do

Data Mining

Sergio Navega
Intelliwise Research and Training
http://www.intelliwise.com/snavega

Conteúdo

- A Pirâmide do Conhecimento
- O Processo de Data Mining
- DM e Outras Disciplinas
- Dedução e Indução
- A Questão dos Níveis
- Indução Orientada a Atributos
- Algumas Técnicas Importantes
  - Árvores de Decisão
  - Regras Caracterizadoras
  - Regras Associativas
  - Regras de Evolução Temporal

Uma História Interessante

Em 1974, uma caldeira de um destróier da Marinha americana explodiu.

Investigadores descobriram que a caldeira já havia sido reparada muitas vezes

Bancos de dados acumulavam dados dos reparos; informação permanecia “escondida”

Marvin Denicoff (Office of Naval Research) ficou encarregado de pesquisar o problema

O Valioso é Raro (e vice-versa)

Esta hierarquia é muito similar ao que ocorre no cérebro de organismos inteligentes

A Pirâmide, Para o Data Mining

Uma Visão do Processo Inteiro

Limpeza

* Dinâmicas
* Incompletas
* Reduntantes
* Incoerentes
* Ruidosas
* Esparsas

DECISÕES

Visualização

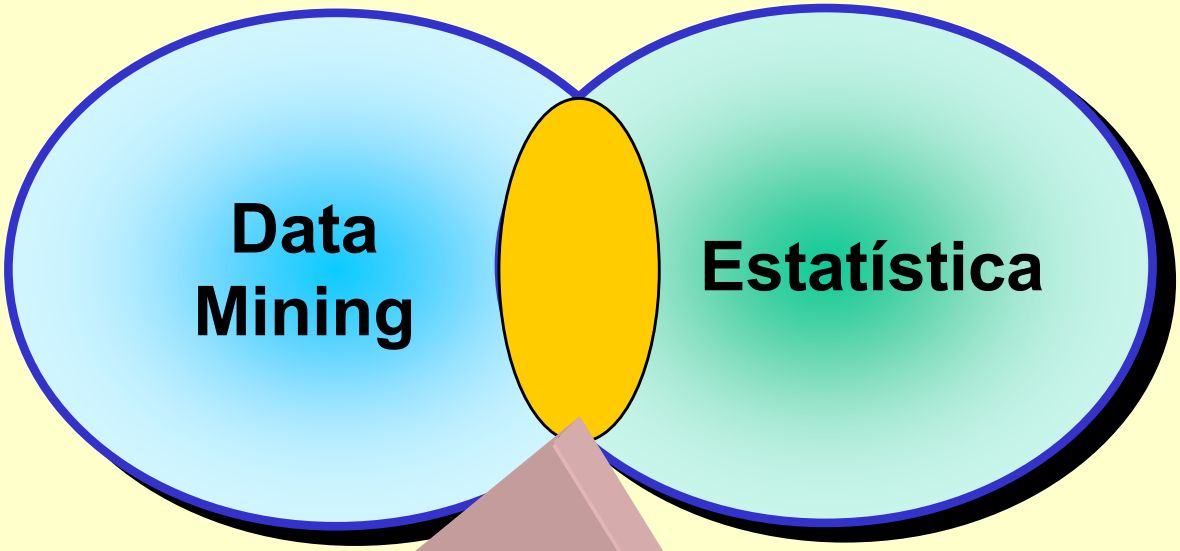
Data Mining

Data Warehouses

Bases de Dados

Fontes Diversas

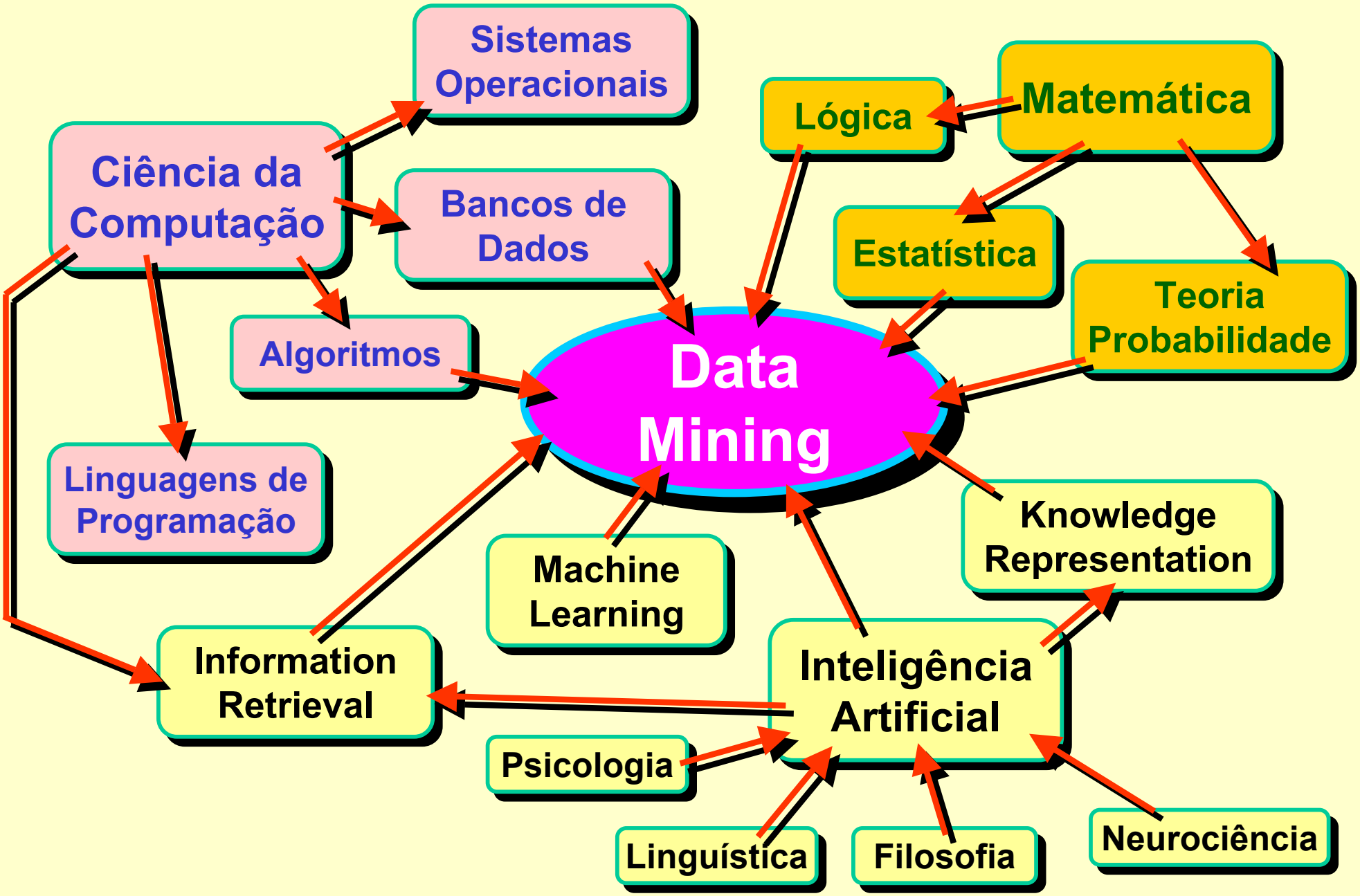
Interface Data Mining - Estatística



Assim como em estatística, em DM você tem o privilégio de nunca precisar dizer que está totalmente certo sobre uma conclusão!

- Objetivo comum de sumarizar quantidades de dados
- Identificar estruturas e relações interessantes em conjuntos de dados (data sets)
- Construir (ou auxiliar) no desenvolvimento de preditores baseados nos dados coletados

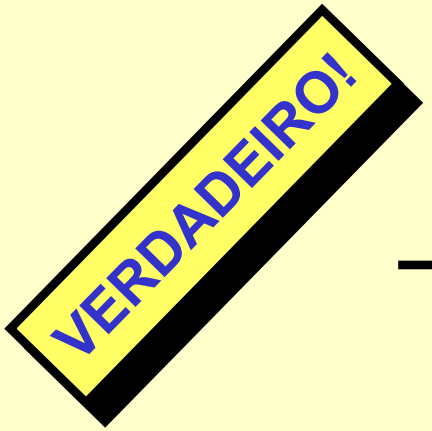
DM Está no Meio de Uma Teia de Disciplinas



Filosofia: Os Dois Lados da Questão

Dedução	Indução
Garante a verdade	Garante a consistência
Conclusões são certeiras	Conclusões são prováveis
Aumento do número de proposições	Redução no número de proposições
Racionalismo	**Empirismo**

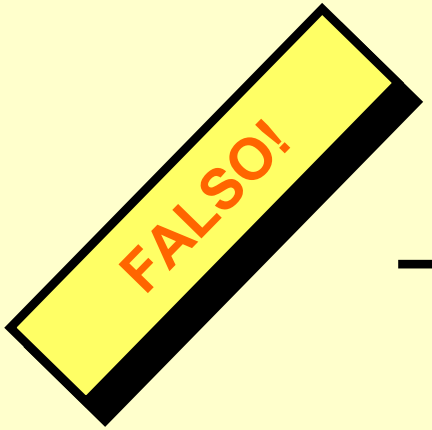
Dedução Prova a Conclusão Pelas Premissas



Todas as Baleias são Mamíferos
Todos os Mamíferos têm Pulmões

Portanto, Todas as Baleias têm Pulmões

Mamíferos bípedes, onívoros

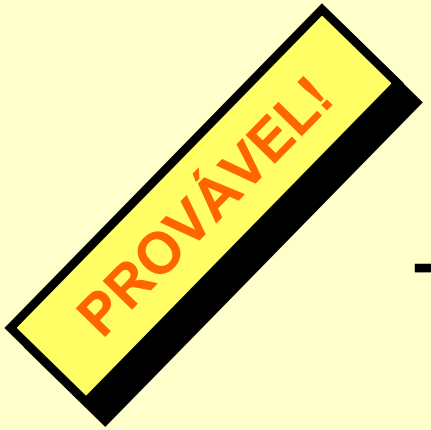


Todos os Homens São Mapeludos
Todos os Ursos São Mapeludos

Portanto, Todos os Homens São Ursos

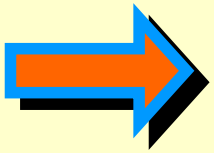
Indução é Parecida (mas nem tanto)

Indução Forte



Meu cão é amigável
O cão de meu vizinho é amigável

Portanto, a maioria dos cães são amigáveis



Algumas instâncias são usadas para "justificar" uma regra genérica;
a indução tem **caráter probabilístico**

Nem Sempre Induções São Boas

Indução Fraca

Esta pessoa conhece o Sergio
Esta outra pessoa conhece o Sergio

Portanto, todas as pessoas deste prédio conhecem o Sergio

Mesmo com essas desvantagens, a indução (e alguns outros raciocínios fracos) são a única forma de **gerar conhecimento novo**

Conceituando o Data Mining

- Bancos de dados corporativos são volumosos e potencialmente cheios de informações valiosas. Técnicas tradicionais de B.D. são inadequadas para gerar novos padrões (são dedutivas).
- A extração de informações preditivas “escondidas” em grandes bancos de dados
- Extração de padrões interessantes de grandes volumes de dados brutos

“Data Mining é o processo não trivial de identificação de padrões em dados que sejam válidos, novos, potencialmente úteis e ultimamente compreensíveis”
Usama Fayyad (1995)

“Nós estamos nos afogando em dados, mas morrendo de fome de conhecimento”
Dr. Jaiwei Han, Simon Fraser University

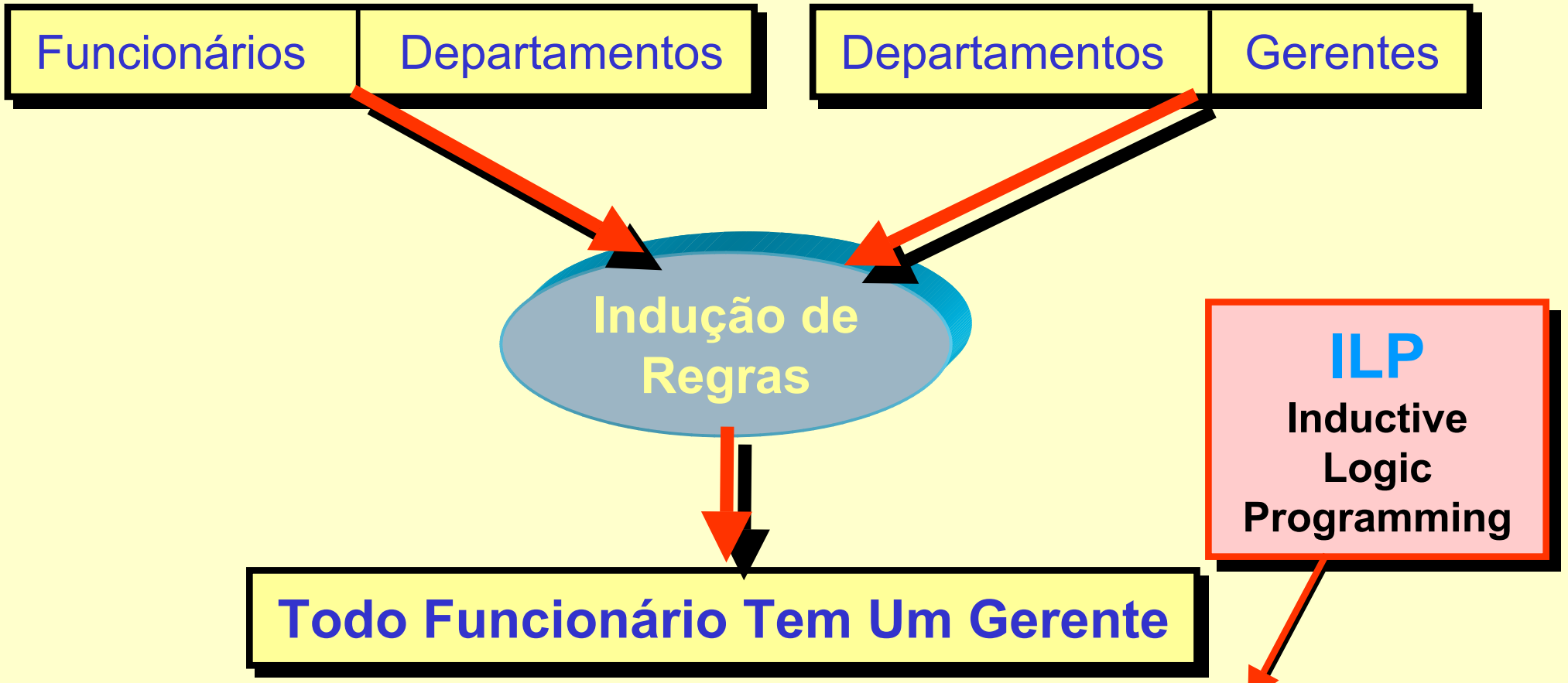
Manipulação de B.D. é Dedução

Dedução: Inferências comprovadamente corretas

Dedução é naturalmente suportada pelos B.D.

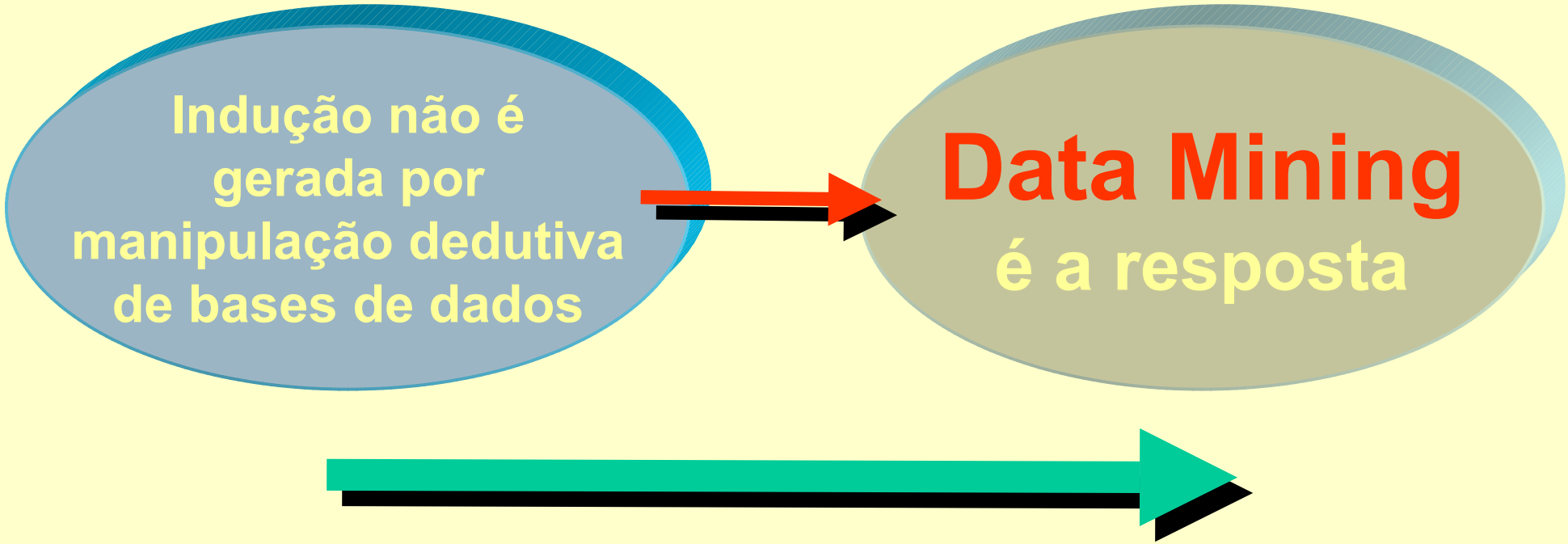
Data Mining é Indução

Usando a mesma B.D., uma indução produz:



$$\forall x, \exists y \therefore Funcionário(x) \Rightarrow Gerente(y, x)$$

Novas Técnicas São Necessárias



Perda de informação, compactação

Para fazer indução é necessário
desprezar algumas características;
Data Mining precisa "perder" alguns dados
Este é um "medo" que precisamos perder!

A Questão dos Níveis

Em Bancos de Dados, as informações estão estruturadas em níveis primitivos

Conhecimento é expresso em níveis mais elevados

Filial

Divisão Operacional

Gerência Informática

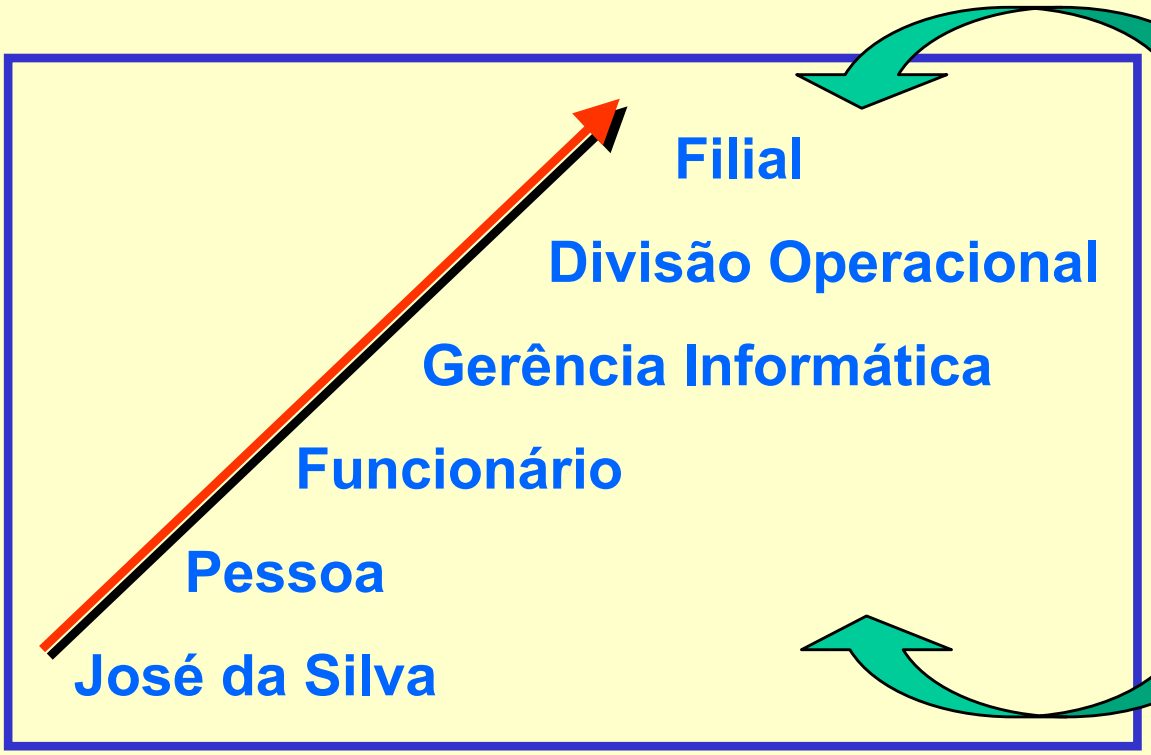
Funcionário

Pessoa

José da Silva

Mineração deve se preocupar com o nível de análise

A Questão dos Níveis



Nível Alto Demais
Regras são óbvias, revelam o senso comum

Nível Baixo Demais
Regras descobertas são muito fracas e suscetíveis a "ruído"

- Mineração em múltiplos níveis, provê diversas visões em múltiplos estágios de abstração
- Interação providencia (em tempo real) foco em áreas mais interessantes, aprofundando o processo de mineração de forma seletiva e controlada

A Indução Orientada a Atributos

- Compressão de dados onde valores de atributos são trocados por conceitos generalizados, de hierarquias superiores
- Hierarquias são normalmente fornecidas por um especialista no domínio ou gerados através de sugestões de outros métodos (ID3, C4.5)
- Pode-se remover atributos que têm grande número de valores distintos e que não possuam hierarquia superior (Ex: chaves de acesso a B.D.)
- Pode-se generalizar atributos que tenham como ser categorizados sob um mesmo nome

Exemplos de IOA

Um banco de dados gigantesco esconde relações valiosas

Estado	Cidade	População	Orç.Saúde
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....

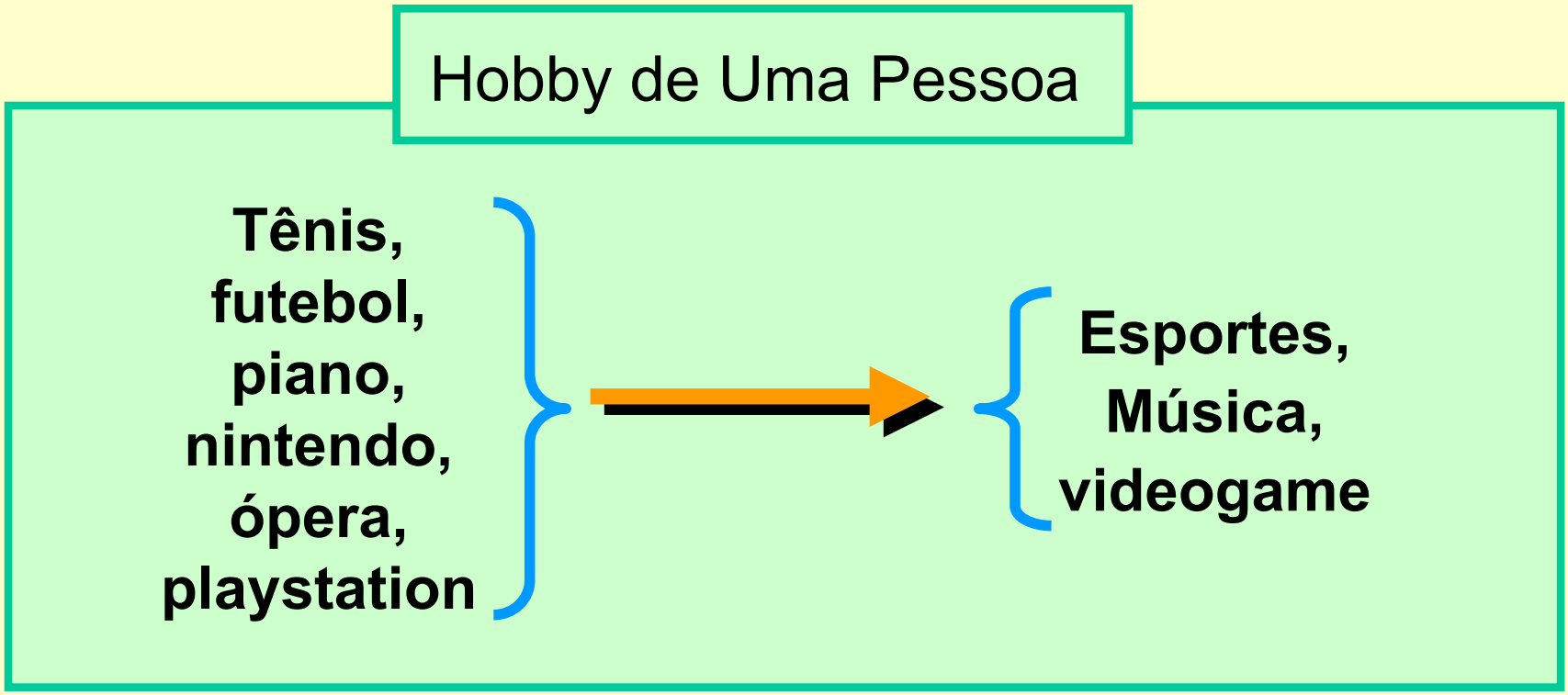
Estado → Norte, Nordeste, Sudeste, Sul, etc.

População → Pequena, Média, Grande

Gastos → 0-5%, 5-15%, 15-20% arrecadação

Generalização de atributos reduz (comprime) o tamanho do banco de dados

Exemplos de IOA



Redução de atributos promove uma generalização que pode favorecer o aparecimento de certos padrões

Três Técnicas Importantes

Árvores de Decisão

ID3 (Quinlan 1986)
C4.5 (Quinlan 1992)

Descoberta de Regras

Piatetsky-Shapiro 1991

Regras Associativas

Agrawal 1993

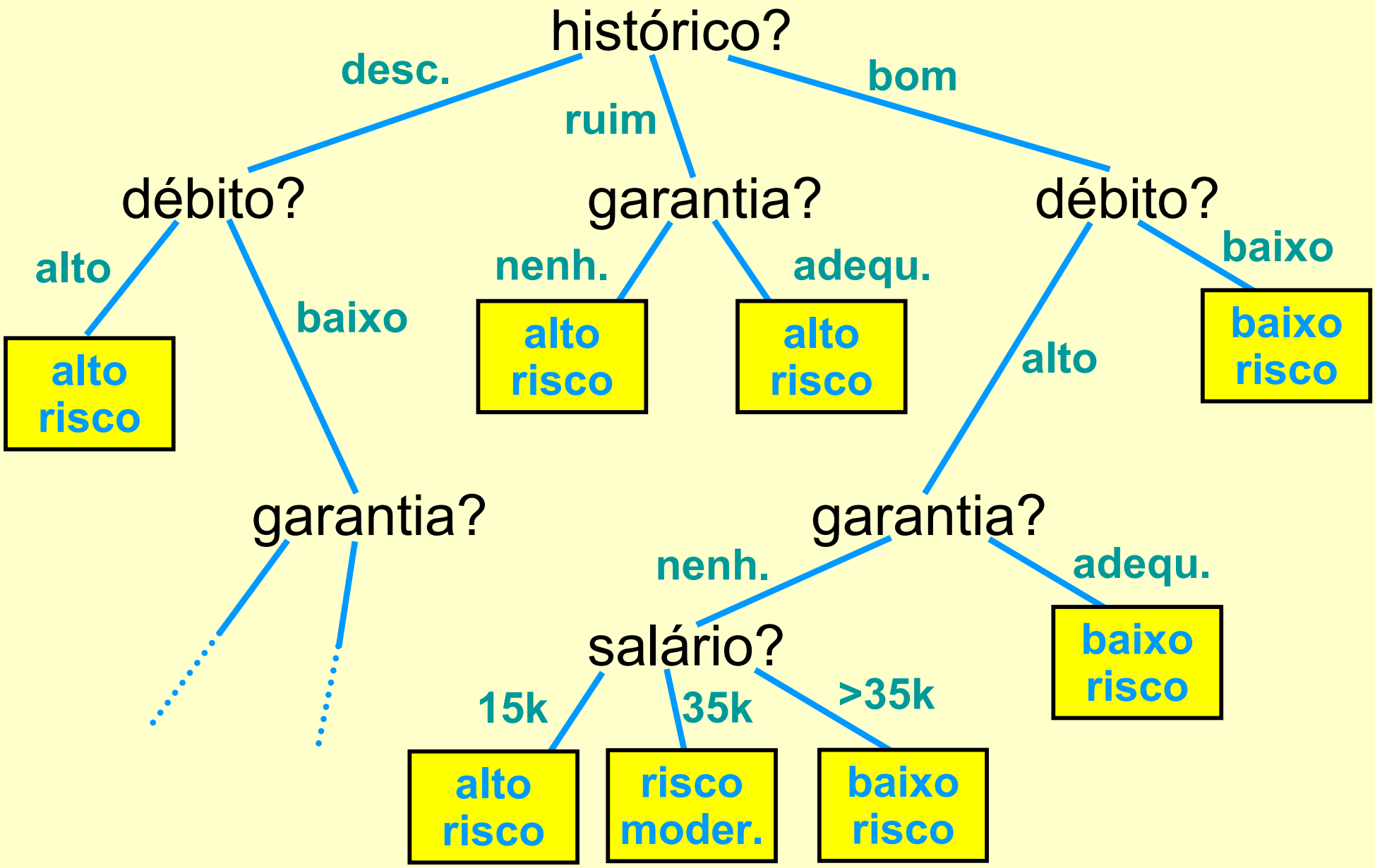
Existe atualmente um grande número de técnicas, várias delas sendo publicadas no momento em que falamos

Indução de Árvores de Decisão

Método ID3 já é obsoleto, mas é importante em termos teóricos

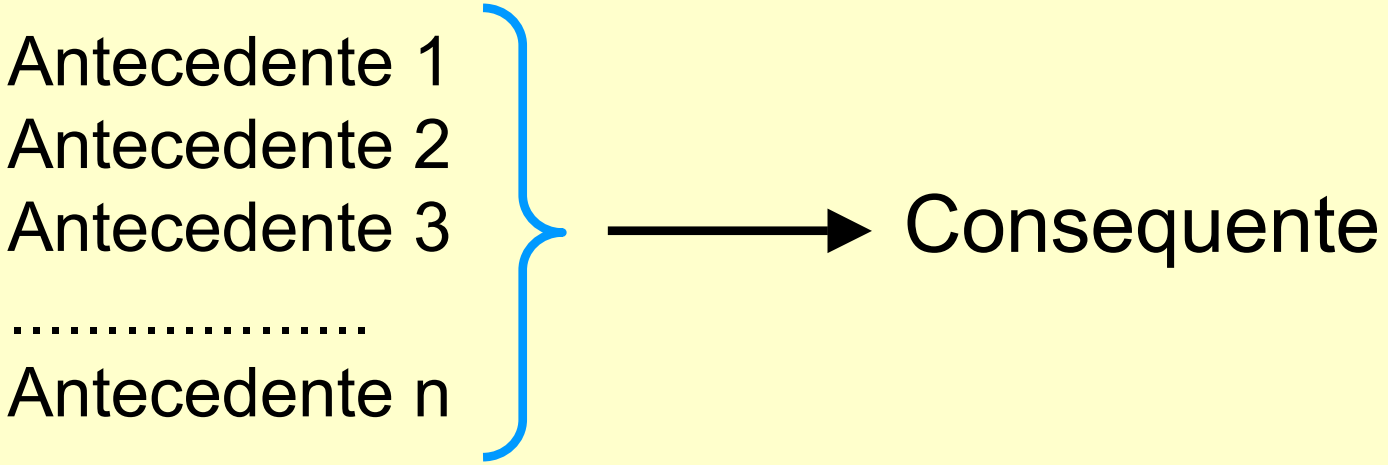
Caso	Risco	Histórico	Débito	Garantia	Salário
1	alto	ruim	alto	nenhuma	15k
2	alto	desc.	alto	nenhuma	35k
3	moder.	desc.	baixo	nenhuma	35k
4	alto	desc.	baixo	nenhuma	15k
5	baixo	desc.	baixo	nenhuma	>35k
6	baixo	desc.	baixo	adeq.	>35k
7	alto	ruim	baixo	nenhuma	15k
8	moder.	ruim	baixo	adeq.	>35k
9	baixo	bom	baixo	nenhuma	>35k
10	baixo	bom	alto	adeq.	>35k
11	alto	bom	alto	nenhuma	15k
12	moder.	bom	alto	nenhuma	35k

A Árvore Obtida Pelo ID3



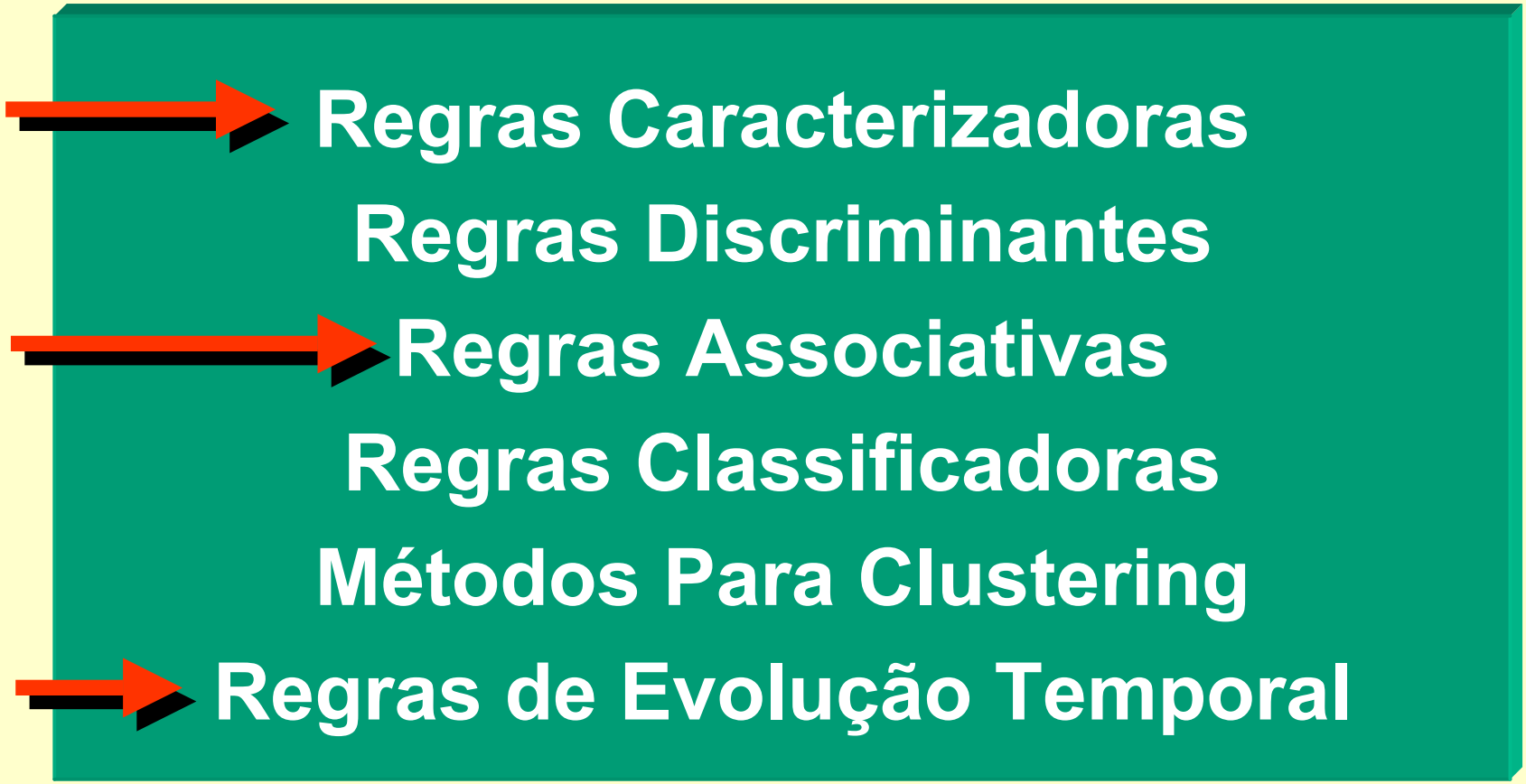
O Que é Uma Regra?

Expressões que relacionam condições iniciais a conclusões. Em argumentação, as condições iniciais são chamadas de premissas e as conclusões são chamadas de alegação. Em nosso caso, vamos usar esta nomenclatura:



Se (emprestimo > 5000)
e (salário < 1500) então {recusar crédito}

Técnicas Mais Comuns



Além dessas técnicas existem várias outras que não veremos aqui (algoritmos genéticos, métodos bayesianos, support vector machine, análise de discriminantes, regressão linear, etc.)

As Regras Caracterizadoras

- Regras que caracterizam um conceito satisfeito por todos (ou pela maioria) dos exemplos. Também conhecidas como Regras de Sumarização.

- Aqui usa-se a IOA de forma a sugerir quais são os atributos que caracterizam uma determinada coleção de dados.

  Exemplos:
  Sintomas de uma doença específica podem ser sumarizados por uma regra caracterizadora

- Características típicas dos estudantes de MBA que decidiram pelo curso logo após terminarem graduação

Regras Caracterizadoras: Análise de Crédito

tempo t1

Cliente: **10015**
- Anos cliente: **6**
- Empréstimo:**$2800**
- Salário:**$4800**
- Possui casa:**Sim**
- Contas em atraso:**2**
- Num. Pagtos. Atraso:**3**
- Cliente rentável:**?**

tempo t2

Cliente: **10015**
- Anos cliente: **6**
- Empréstimo:**$4200**
- Salário:**?**
- Possui casa:**Sim**
- Contas em atraso:**2**
- Num. Pagtos. Atraso:**4**
- Cliente rentável:**?**

tempo t3

Cliente: **10015**
- Anos cliente: **7**
- Empréstimo:**$6720**
- Salário:**?**
- Possui casa:**Sim**
- Contas em atraso:**3**
- Num. Pagtos. Atraso:**6**
- Cliente rentável:**Não**

Se [contas em atraso] > 2 e [num.pagtos.atraso] > 1 Então
Se [cliente rentável] = “Não” Então {Recusar Crédito}

Se [contas em atraso] = 0 e ([salário] > 3000 ou
([anos cliente] > 3) Então
{Aceitar Solicitação de Crédito}

Regras Associativas

Dado um B.D. qualquer, descobrir quais as associações entre itens de forma que a presença de um item em um registro implica na presença de outro(s) item(s) no mesmo registro

Exemplo:
Para a maioria das pessoas que adquiriram pão e leite conjuntamente, também foi adquirido manteiga

{pão, leite} → manteiga

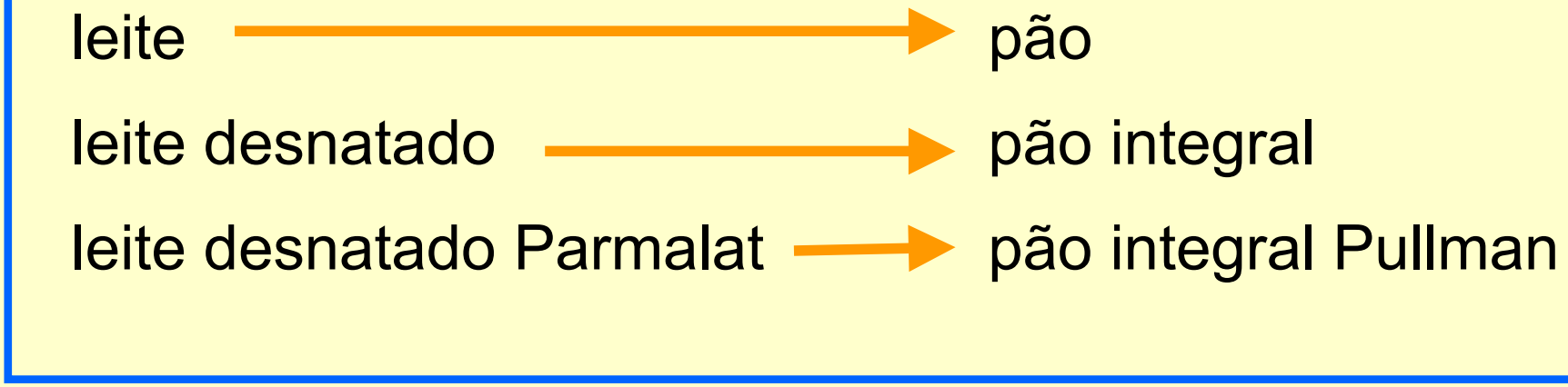
Características das Regras Associativas

- Em geral, despreza quantidades, só interessando a presença ou ausência de determinada associação. Por isso, é um processo booleano por natureza (Sim/Não)

Jiawei Han propôs algoritmos que desenvolvem associações sobre "faixas" de quantidades

Regras Associativas Multi-Níveis

Dr. Han também propôs descobrir regras em níveis conceituais específicos, como as MLAR (Multiple-Level Association Rules)



Market Basket Analysis

Esta é uma das mais interessantes aplicações das Regras Associativas, de fundamental importância para marketing

A) Achar todas as regras que tenham “diet coke” como consequentes. Irá auxiliar no planejamento de lojas para vender melhor esse produto

B) Achar todas as regras com “iogurte” como antecedente. Irá auxiliar a determinar o impacto nas vendas, caso esse produto seja retirado das prateleiras

C) Achar todas as regras com “salsicha” no antecedente e “mostarda” no consequente. Auxilia na obtenção de melhores regras para determinar que produtos devem ser vendidos em conjunto com salsichas para aumentar as vendas de mostarda

Regras de Evolução Temporal

- Procuram acompanhar a evolução no tempo de um conjunto de dados, tentando obter padrões

  Exemplo:
  Comprou um PC com CD-ROM hoje, poderá comprar um DVD-ROM em seis meses

- Comprou impressora hoje, precisará de novos suprimentos em dois meses

- Quem comprou videocassete hoje tem 3 vezes mais probabilidade de adquirir uma camcorder 7 meses após a compra

- Na Evolução Temporal interessa-nos dados que sofrem variações constantemente. Anti-exemplo: departamento de um funcionário

Usa as Outras Técnicas Como Base

- As Regras de Evolução Temporal são construídas através das técnicas anteriores (caracterização, classificação, associação, clustering)
- Achar as principais características das empresas cujas ações em bolsas de valores tiveram crescimento de 20%
- Achar as empresas que tenham ações subindo em conjunto (regras associativas)

Perguntas?

Sergio Navega
snavega@attglobal.net

Artigo sobre DM e cópia destes slides:
`http://www.intelliwise.com/reports`