Clodis Boscarioli
boscarioli@gmail.com

unioeste
Universidade Estadual do Oeste do Paraná

PPGTGS

PPGComp

Mineração de Dados para Tomada de Decisões: Conceitos Principais

Clodis Boscarioli

Me apresentando...

Clodis Boscarioli

unioeste

Ciência da Computação

PPGEn

PPGECEM

PPGTGS

PPGComp

Silva, L. A., Peres, S. M. e Boscarioli, C. Introdução à Mineração de Dados com Aplicações em R. Editora Elsevier, Rio de Janeiro, 2016.

Agenda

- Conceitos gerais de BI&A
- BIA& KDD
- Mineração de Dados

BI – Business Intelligence

Denomina um conjunto de metodologias e ferramentas que permitem às organizações integrar, acessar e compartilhar grandes volumes de informações, de modo a auxiliar a tomada de decisões (Gartner Group)

BA – Business Analytics

Complementa o termo BI incluindo cenários e modelos preditivos

BI e BA

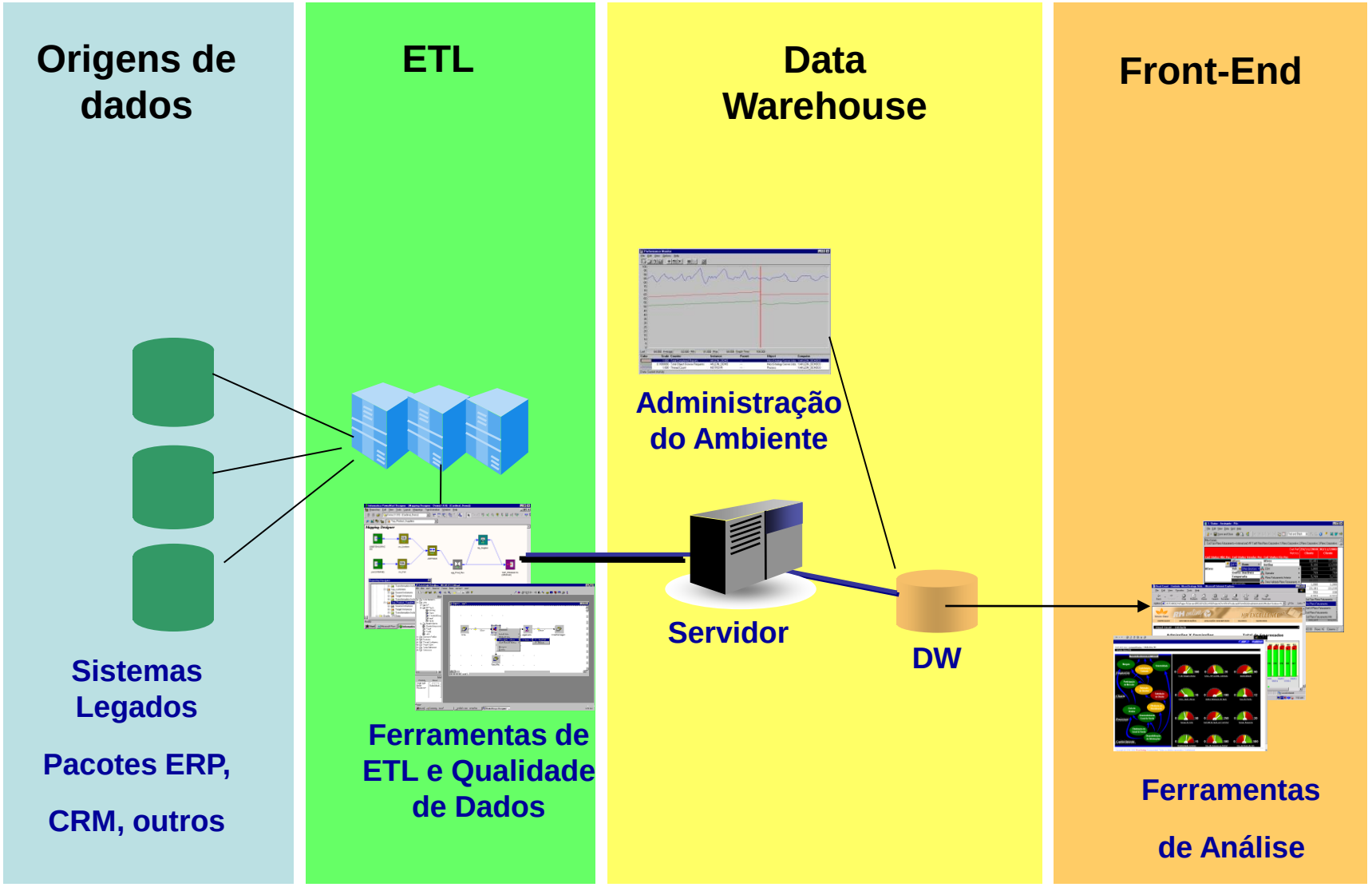
- BI: aborda o que aconteceu no passado e como aconteceu até o momento presente. Identifica tendências e padrões sem investigar ou prever o futuro.

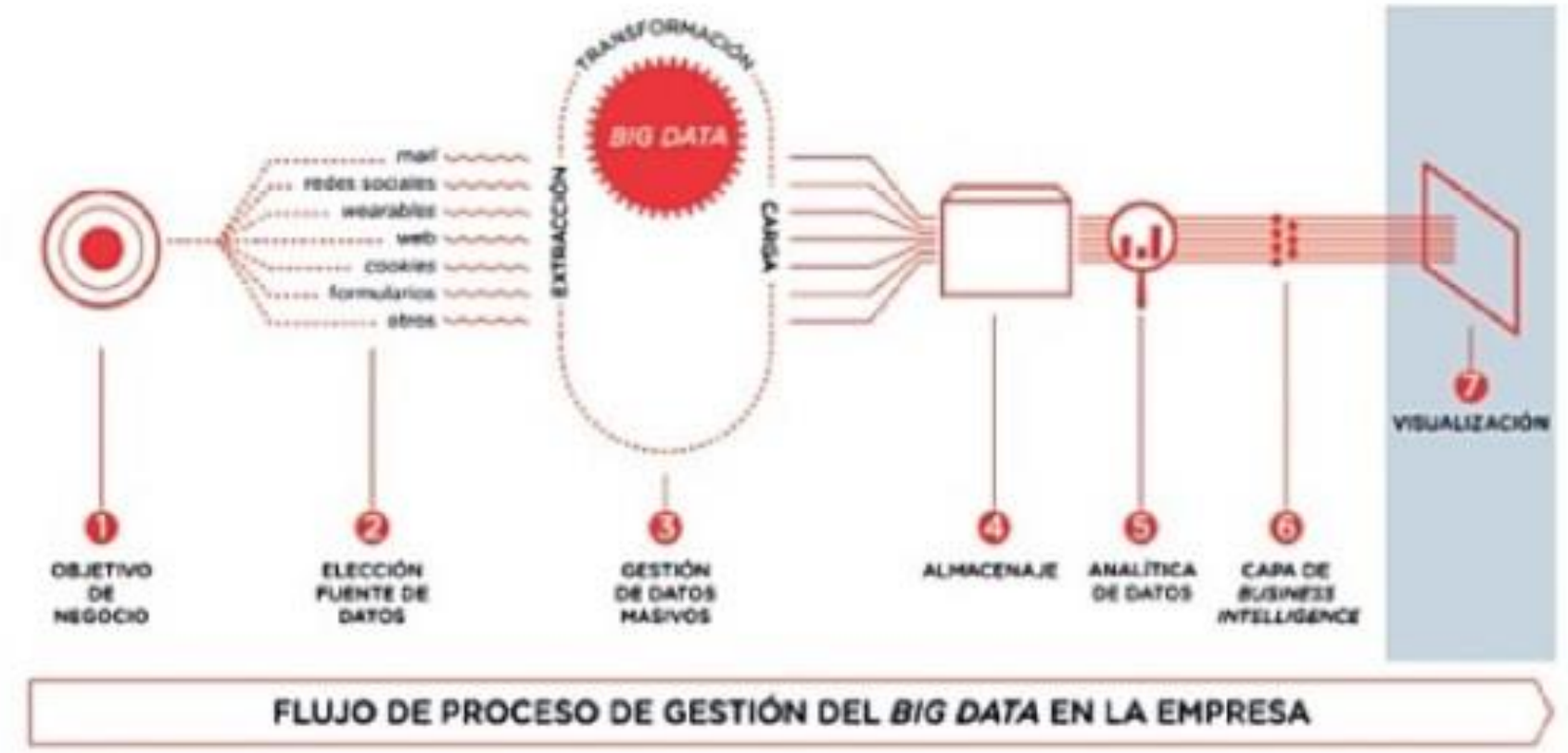
- BA - lida com o “por quê” do que aconteceu no passado. Relação de fatores e causalidades. Fazer previsões do que acontecerá no futuro.

(BI&A)

- O termo *Business Intelligence and Analytics* (BI&A) trata do desenvolvimento de tecnologias, sistemas, práticas e aplicativos para analisar dados de negócios, a fim de obter novas ideias sobre negócios e mercados, que podem ser usadas para melhorar produtos e serviços, obter melhor eficiência operacional e fomentar o relacionamento com os clientes.

Arquitetura genérica de um BI&A

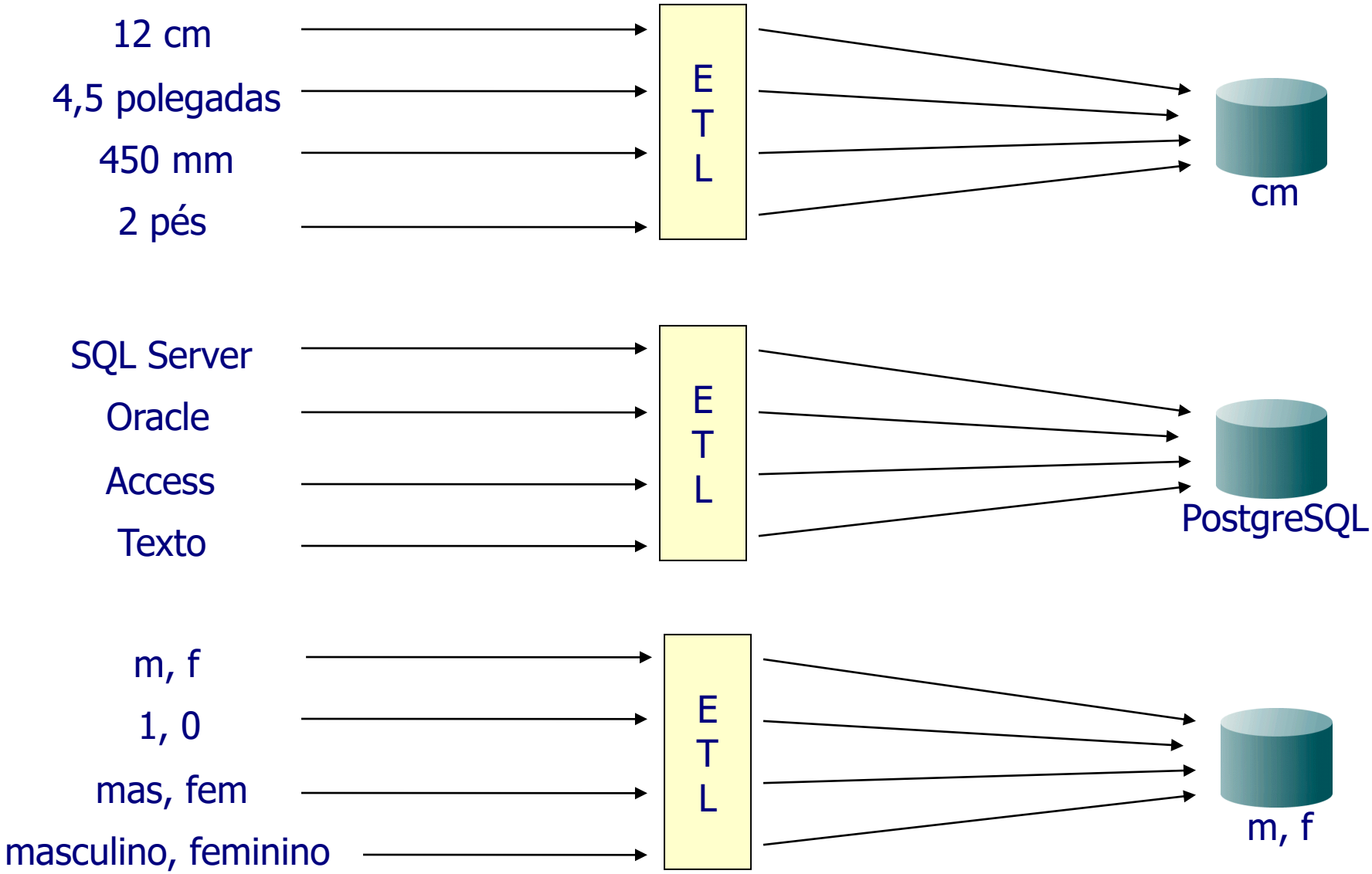




FLUJO DE PROCESO DE GESTIÓN DEL *BIG DATA* EN LA EMPRESA

Fonte: Manzano et al. (2016)

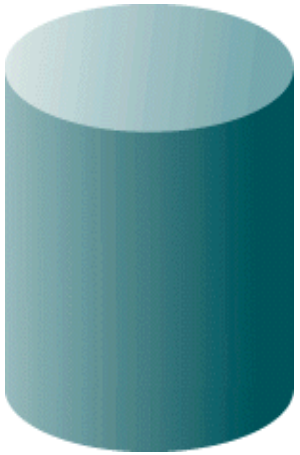
ETL – Transformação



DATA WAREHOUSE (DW) - Conceito

Armazém de Dados

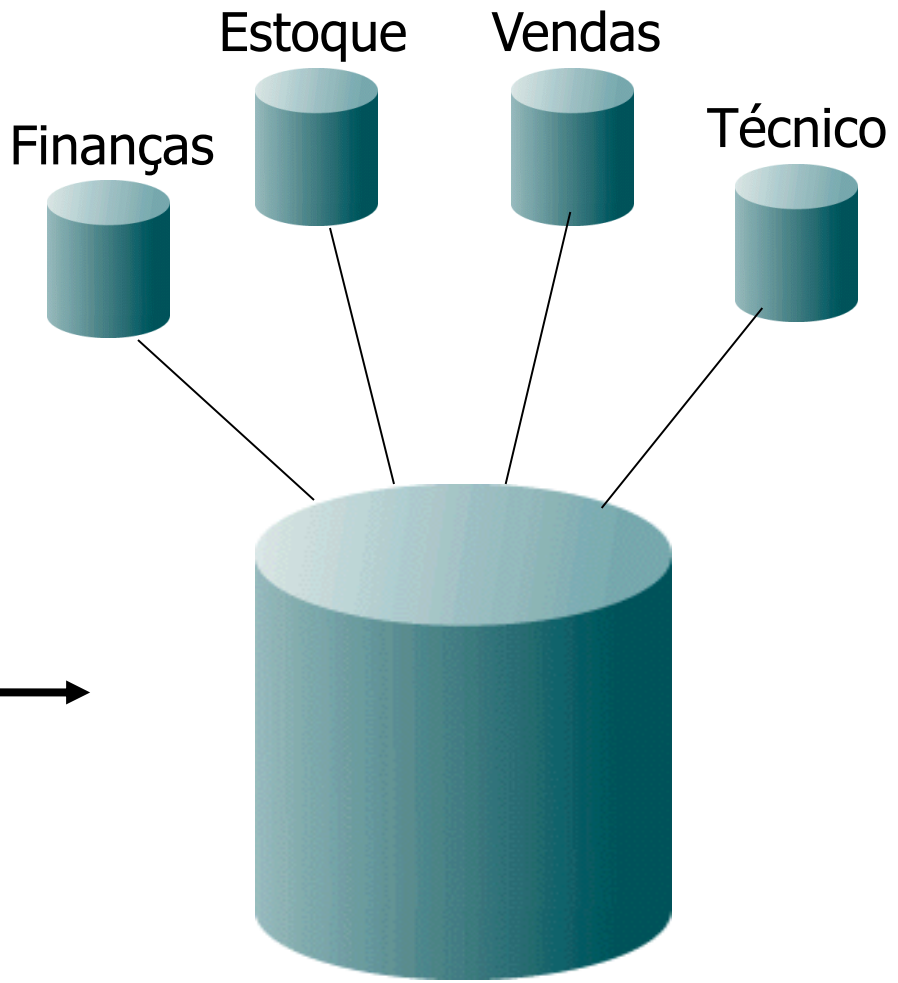
É um amplo e flexível repositório de dados, que aglutina dados de fontes heterogêneas, projetado de modo a suportar o processo de tomada de decisão.



- **Ambiente separado**
- **Disponibilidade**
- **Integrado**
- **Retrato no tempo**
- **Orientado por assunto**
- **Fácil acesso**

DW - Organização

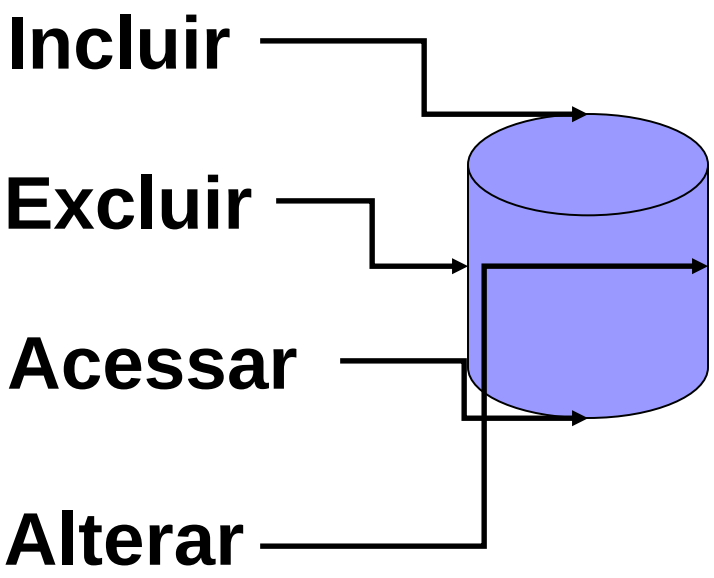
DATA MART
DW Departamental



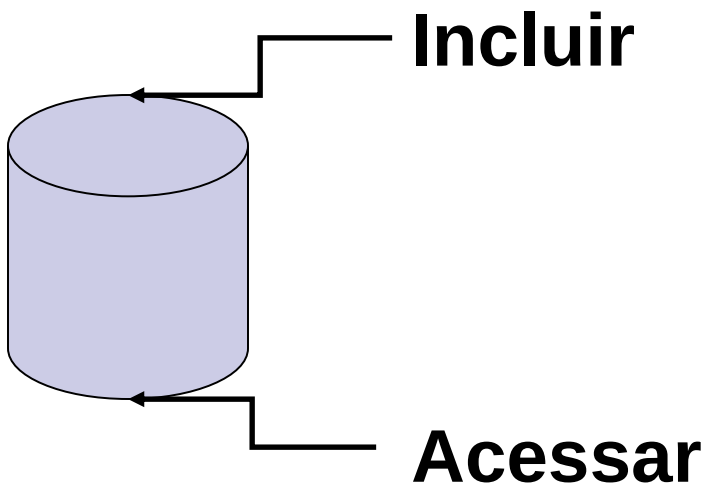
DATA WAREHOUSE
Corporativo

Conceitos Principais

Banco de dados Transacional

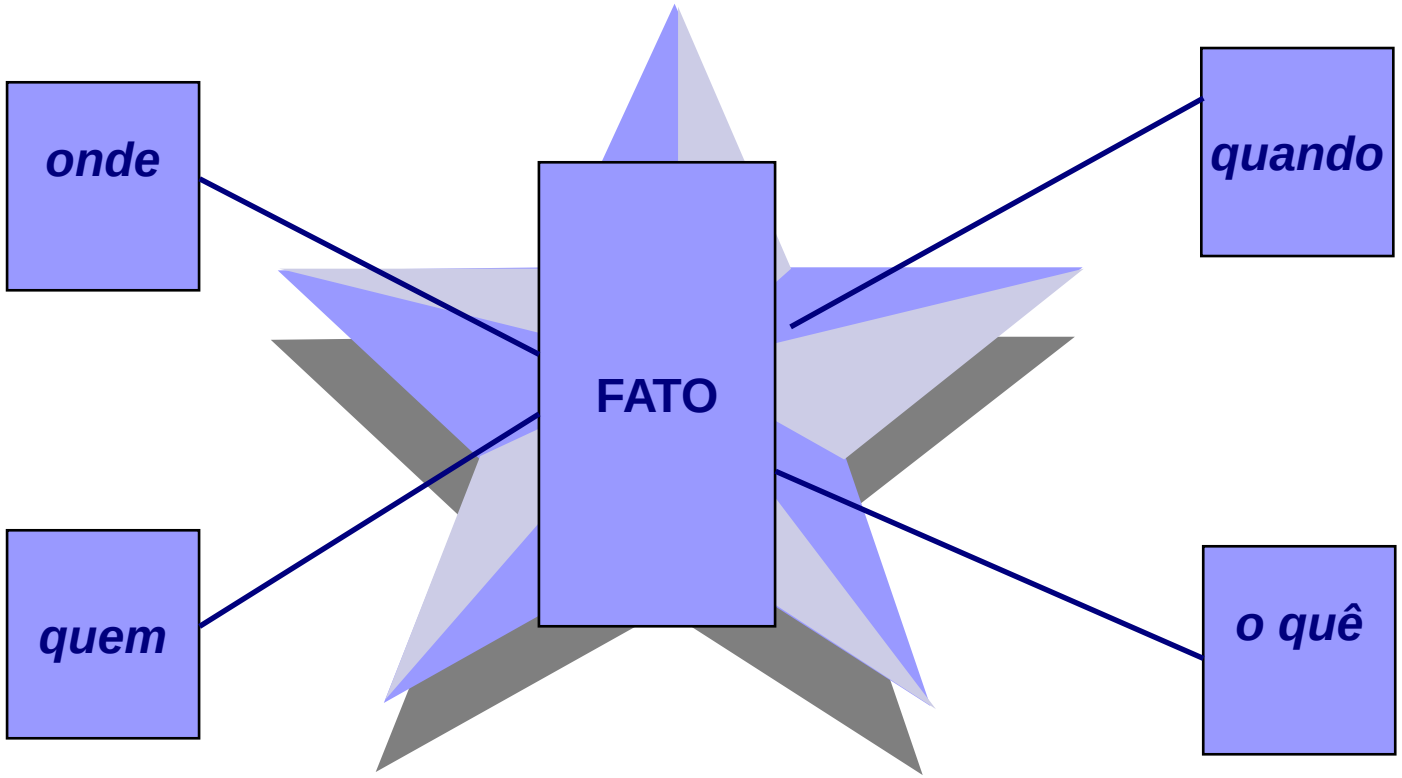


Data Warehouse



Modelo Dimensional → Esquema Estrela

- Uma tabela de fatos cercada de tabelas de dimensões



O componente *Front-End*

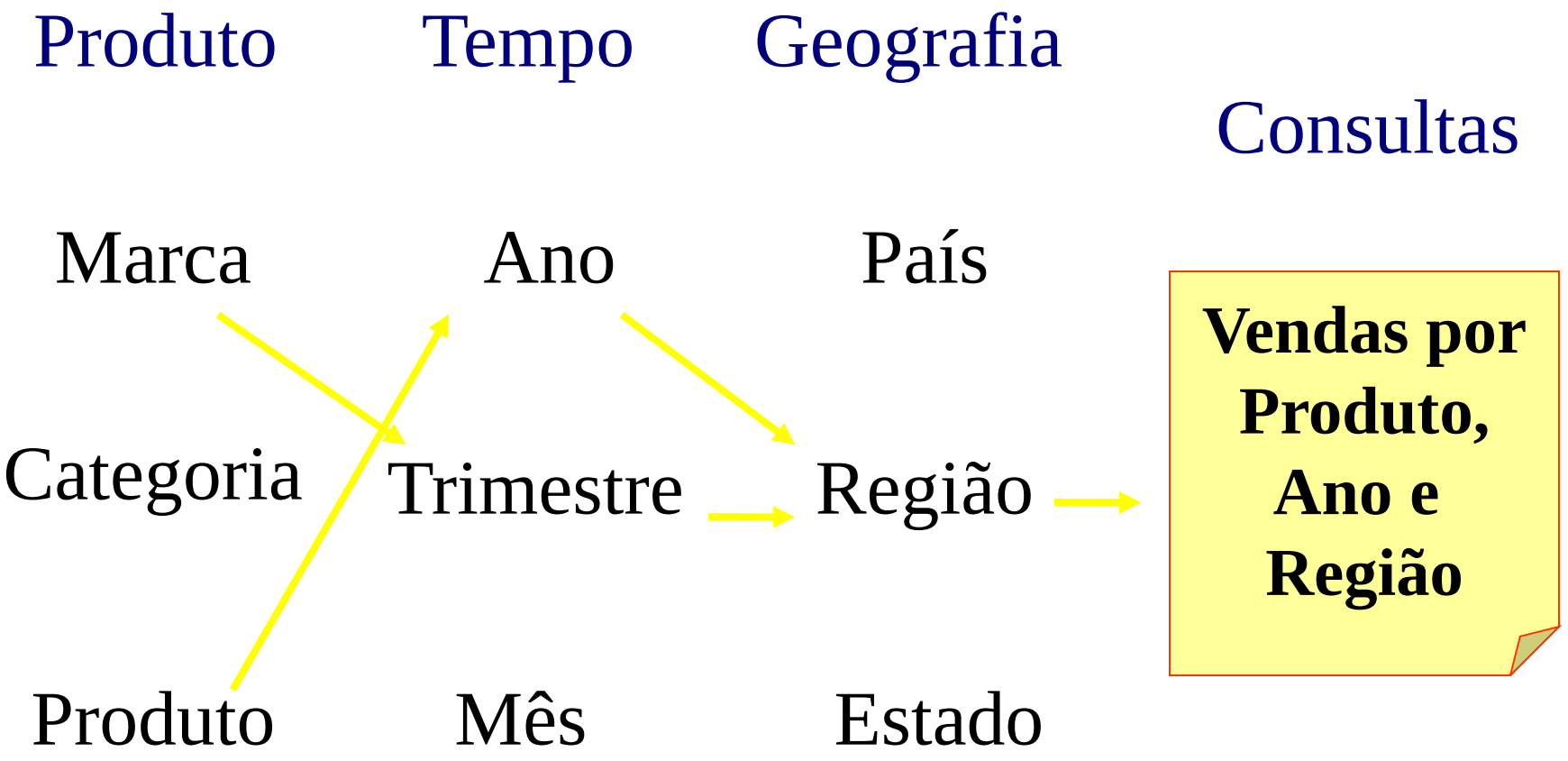
- É o que permite ao usuário final acessar os dados do DW
- Também disponibilizado em ambiente web
- Disponibiliza relatórios e um diversificada forma de análises

OLAP (*Online Analytical Processing*)

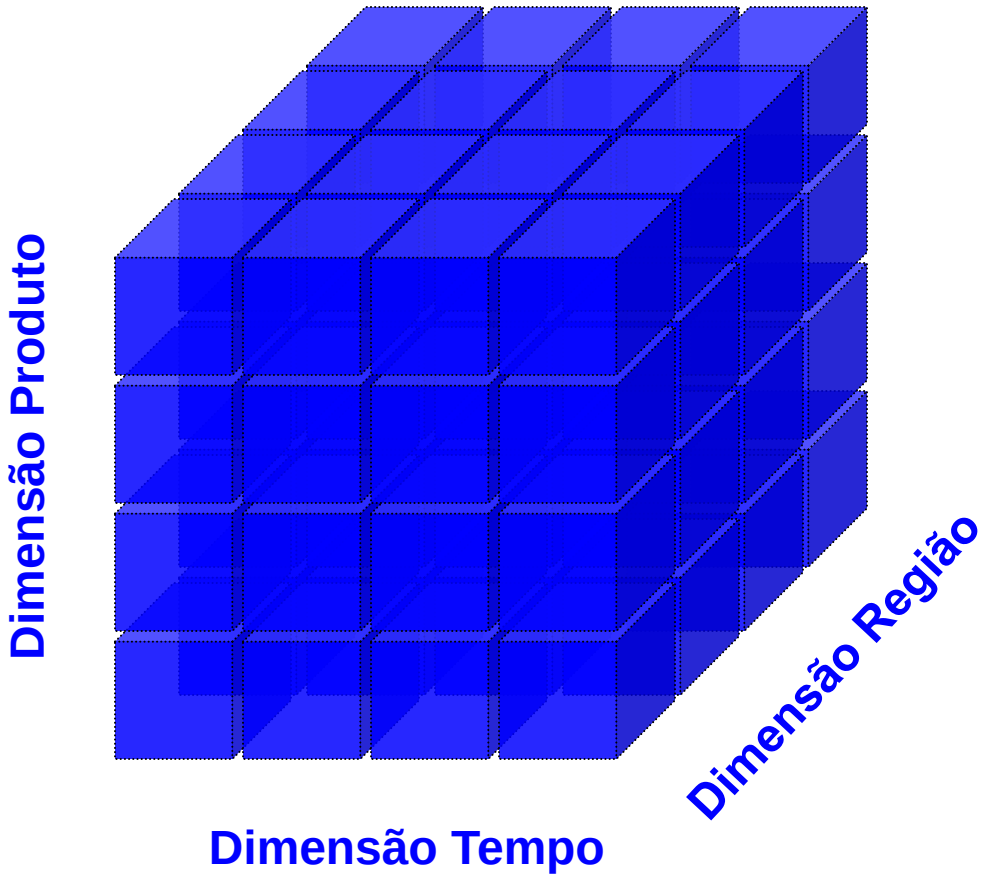
- Conjunto de processos para criação, gerência e manipulação de dados multimensionais para análise e visualização, visando maior compreensão dos dados pelos usuários finais.

- Facilidade para fazer análises, definir agregações e cruzamentos, permitindo visualizar os dados em múltiplos níveis de hierarquias e diferentes perspectivas.

Hierarquias e Agregados

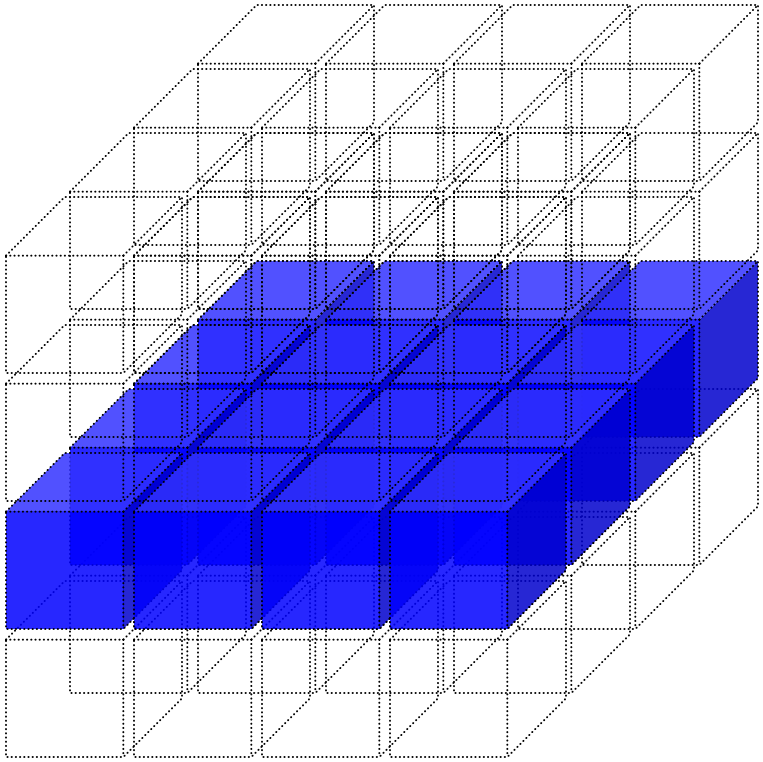


Exemplos no Cubo de Dados

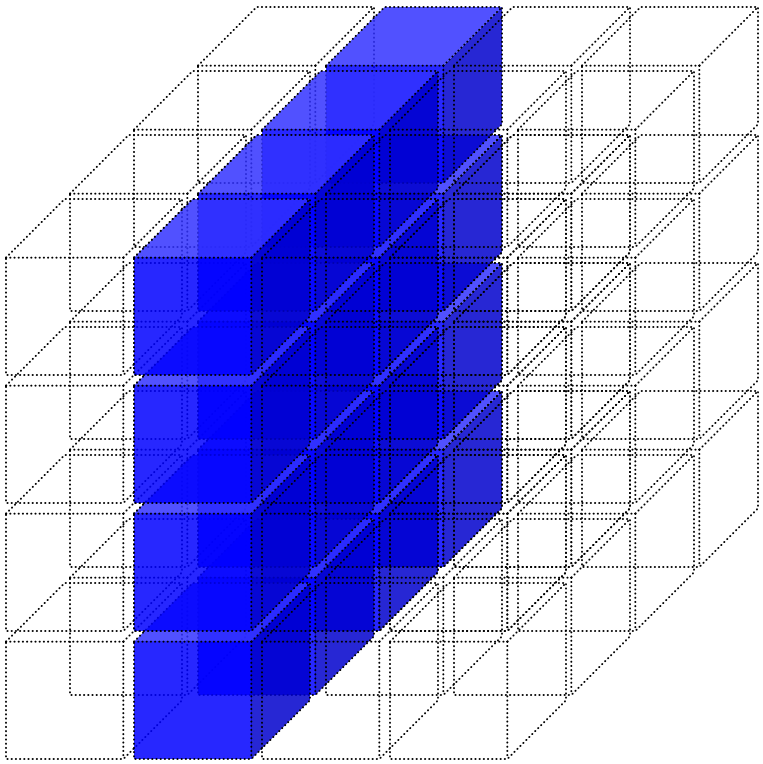


Slice and Dice

Visão Produto

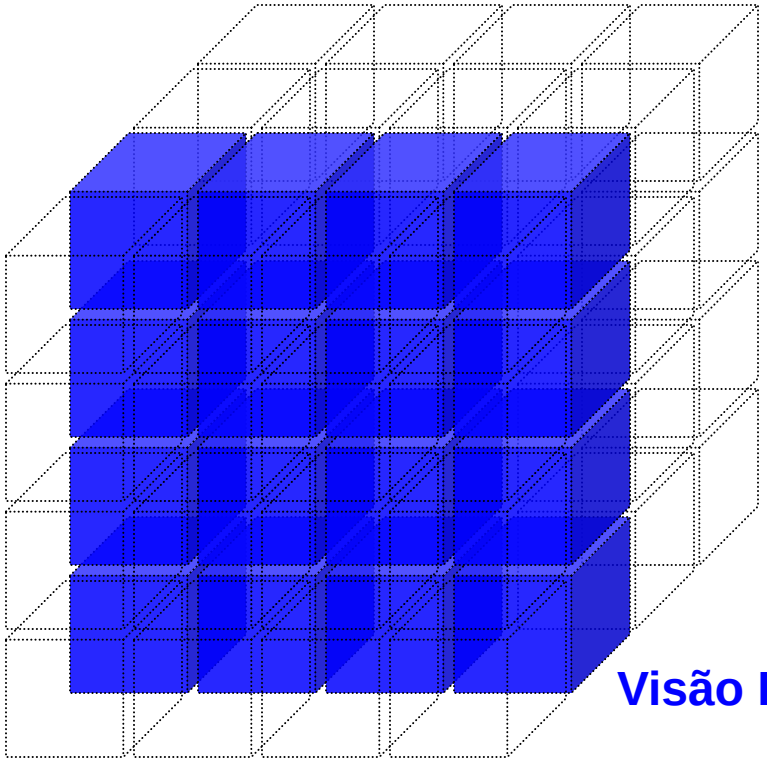


Slice and Dice



Visão Tempo

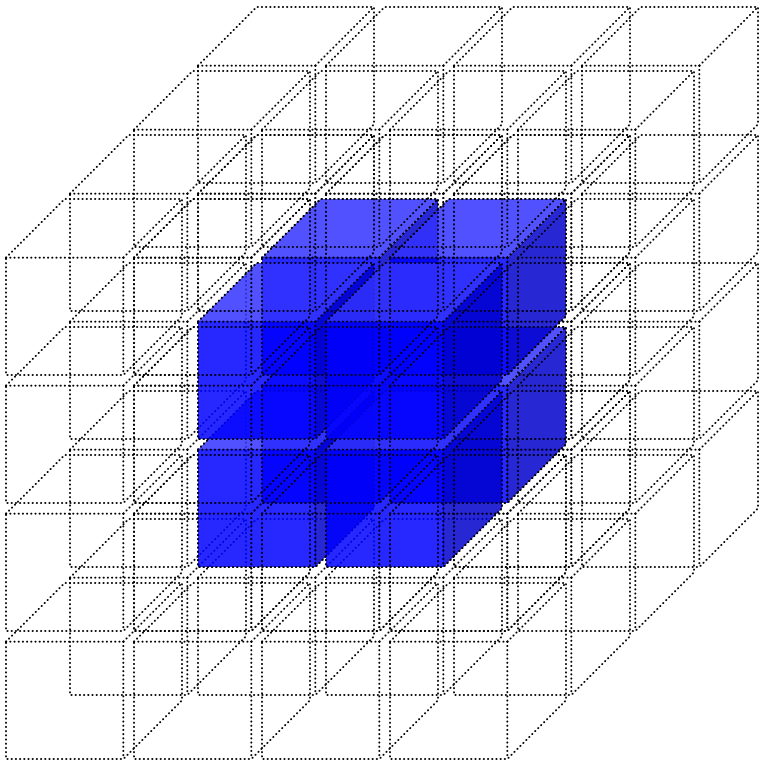
Slice and Dice

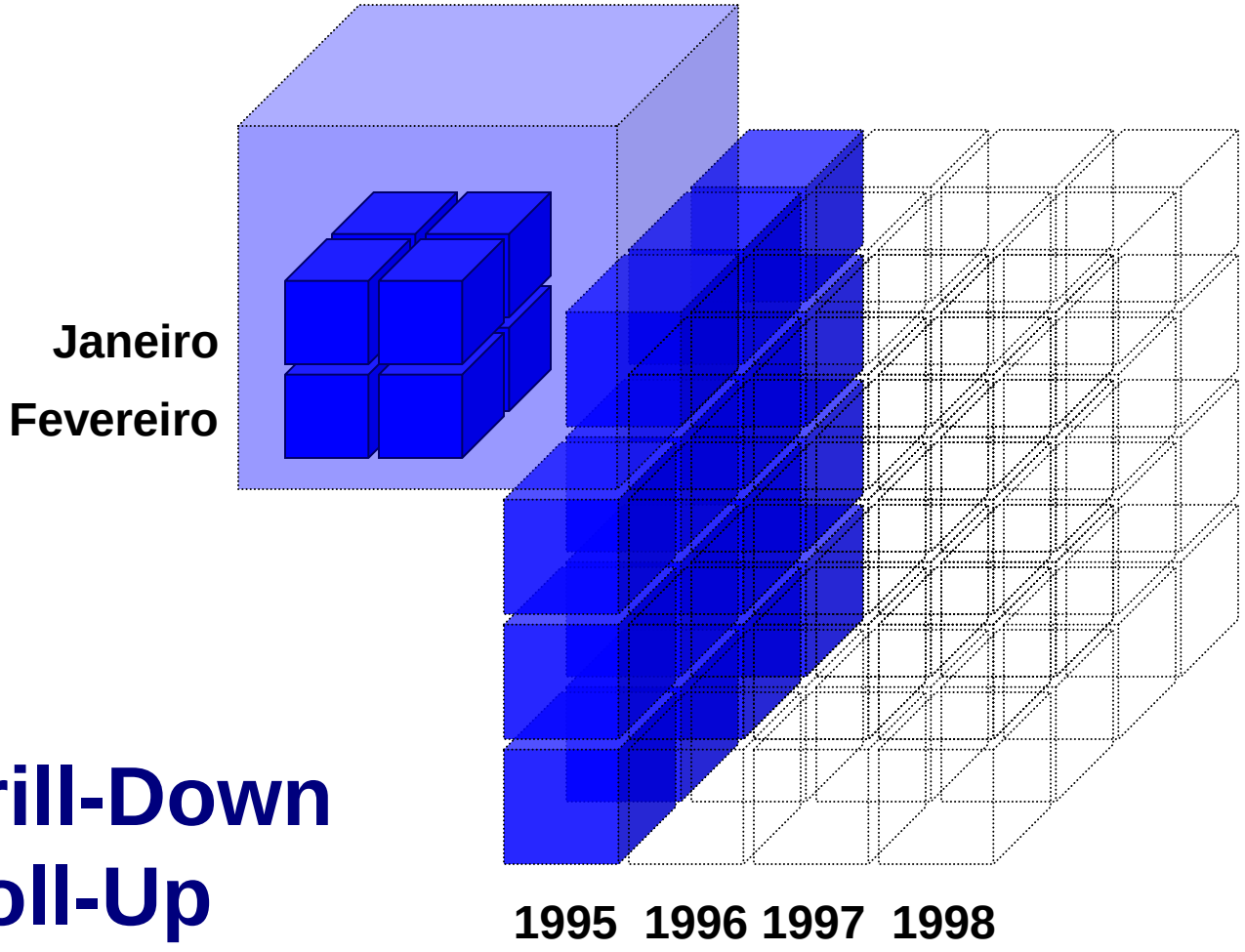


Visão Região

Slice and Dice

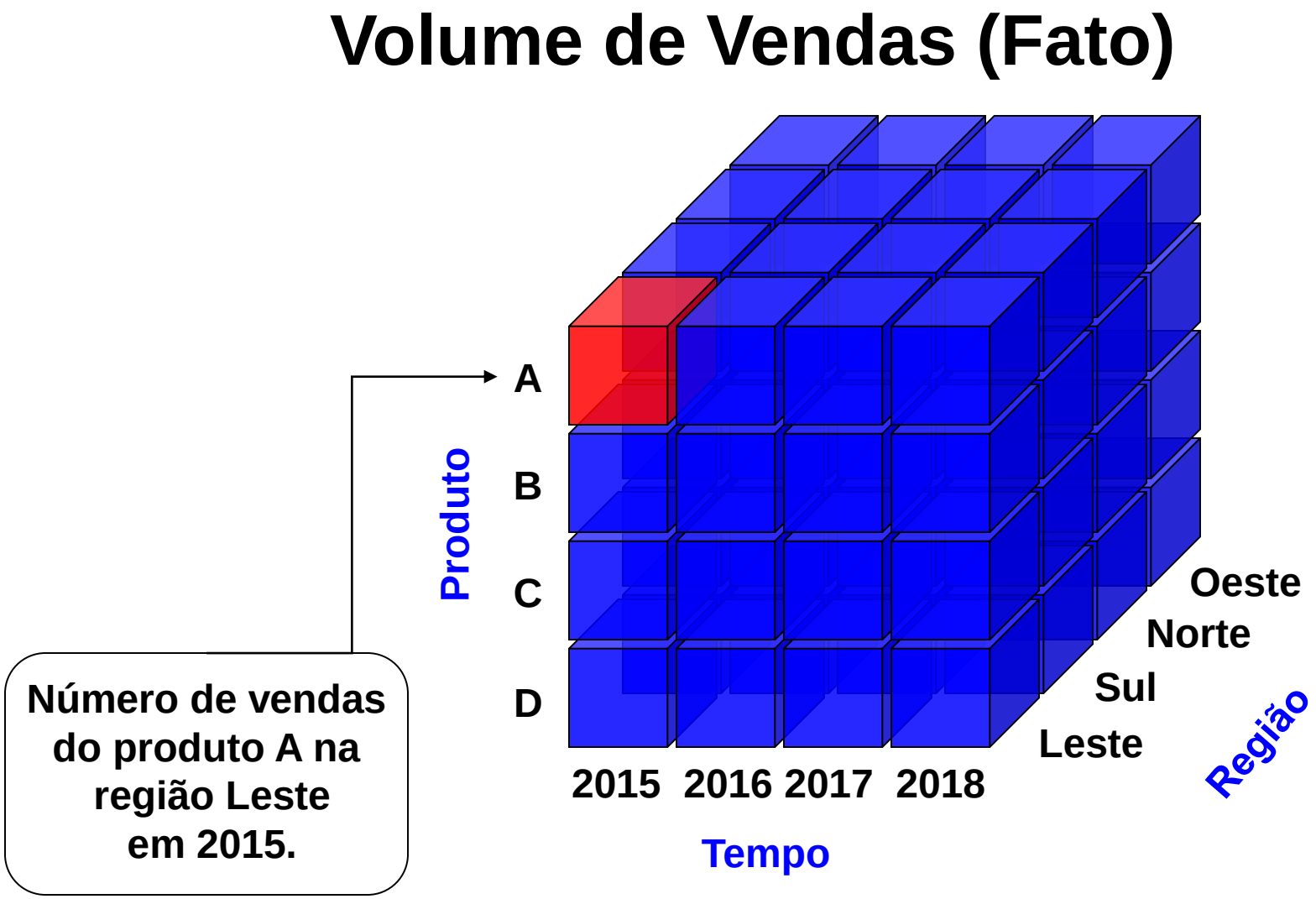
Visão ad-hoc



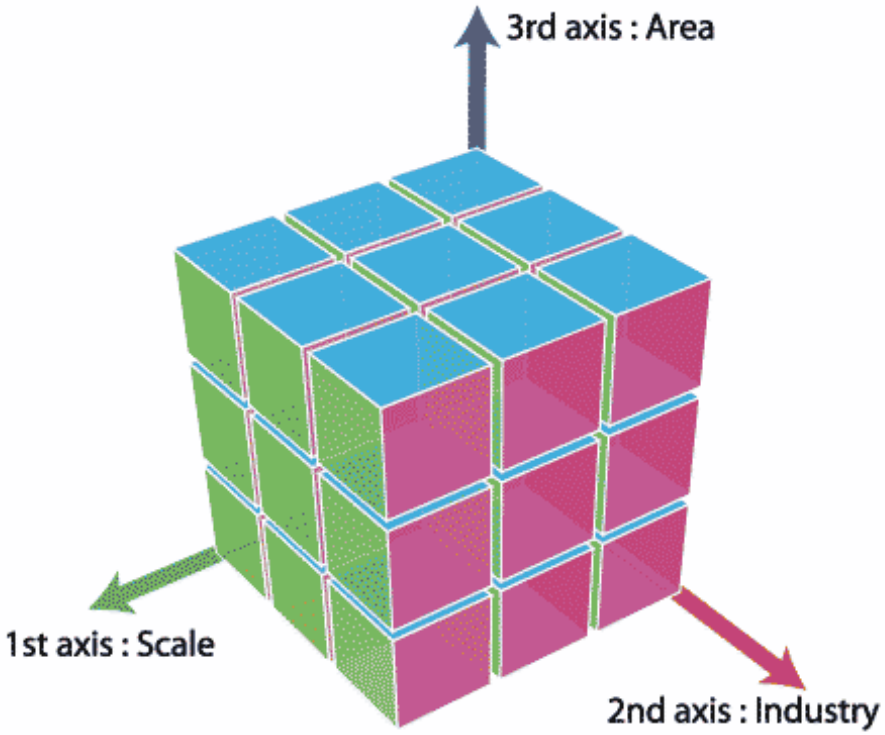


Drill-Down
Roll-Up

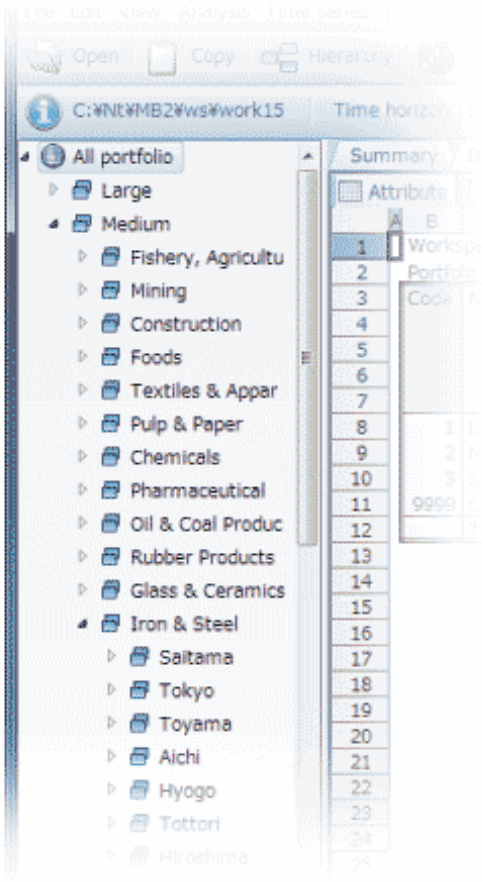
Analisando o Cubo

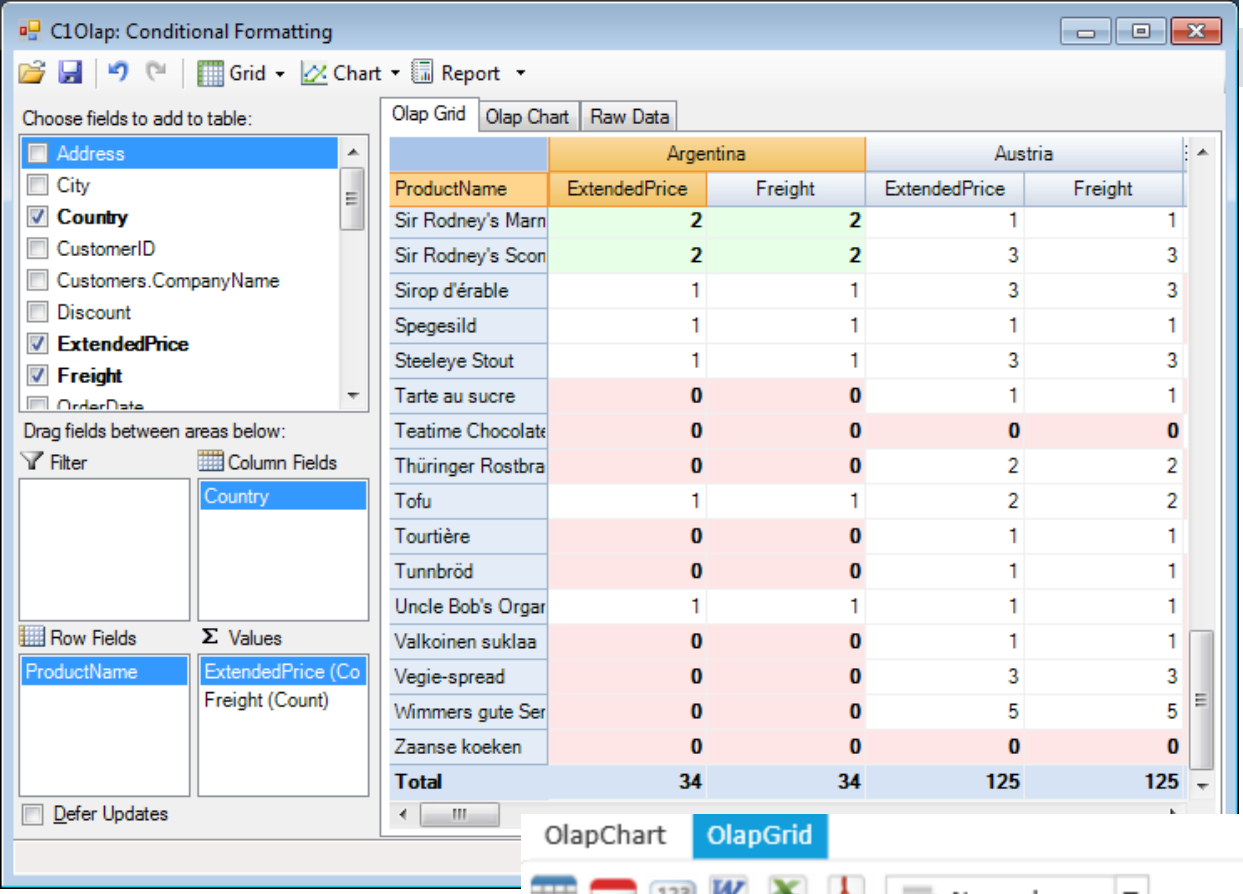


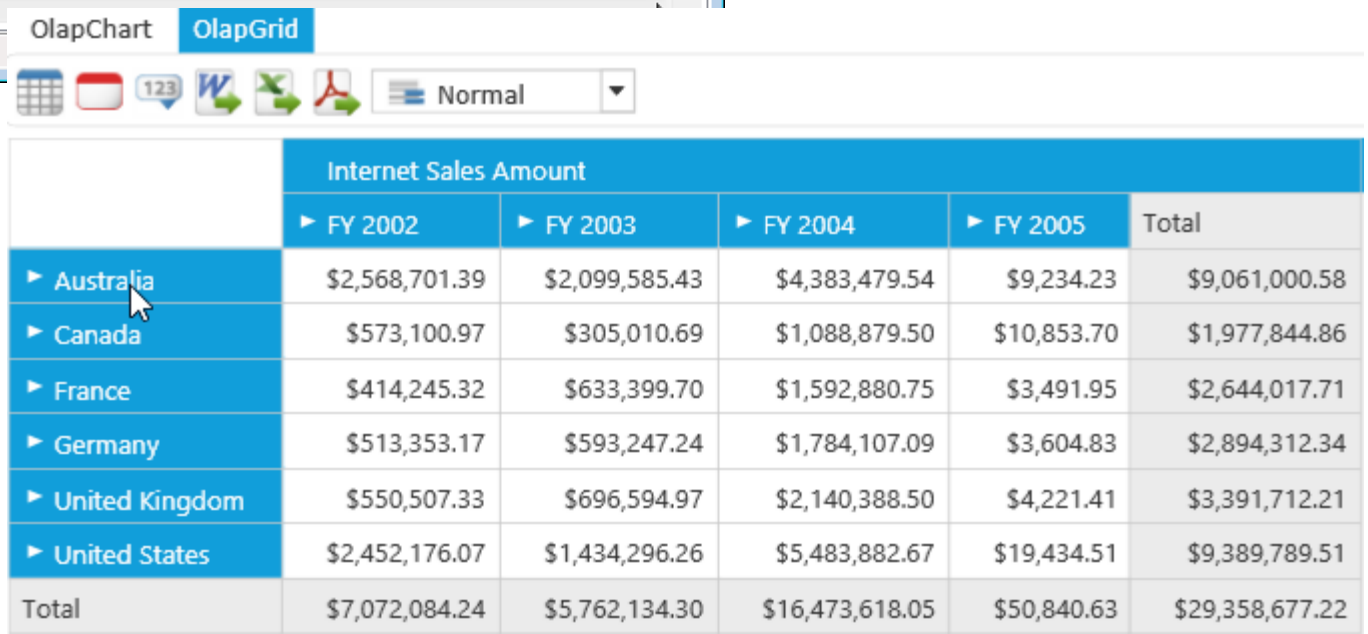
Internal expression of multi-dimensional database
(Hyper-cube structure)



User interface
(Tree structure)



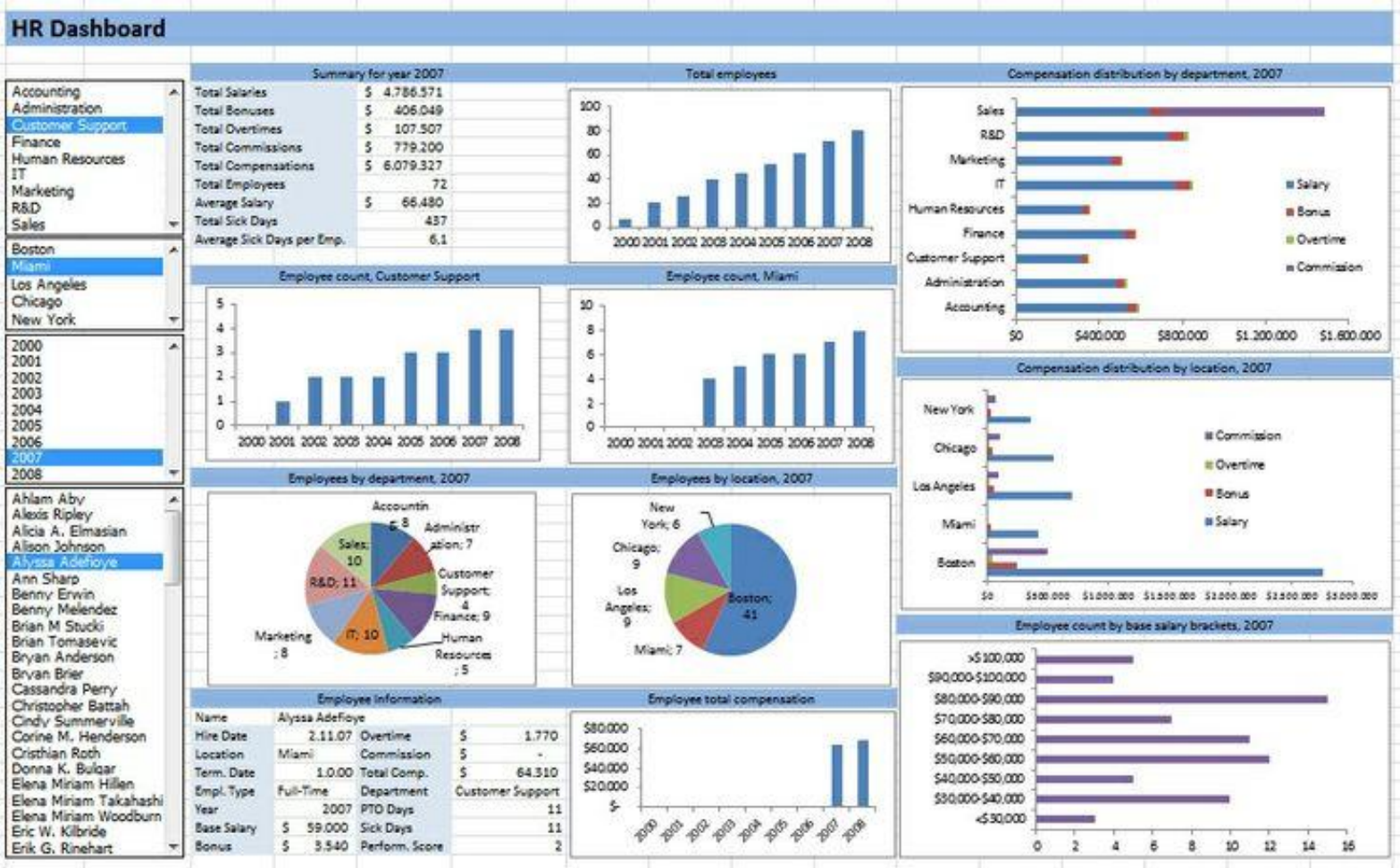

C1Olap: Conditional Formatting
Grid
Chart
Report
Choose fields to add to table:
Address
City
Country
CustomerID
Customers.CompanyName
Discount
ExtendedPrice
Freight
Drag fields between areas below:
Filter
Column Fields
Country
Row Fields
Values
ProductName
ExtendedPrice (Co
Freight (Count)
Defer Updates
Olap Grid
Olap Chart
Raw Data
Argentina
Austria
ProductName
ExtendedPrice
Freight
ExtendedPrice
Freight
Sir Rodney's Marn 2 2 1 1
Sir Rodney's Scon 2 2 3 3
Sirop d'érable 1 1 3 3
Spegesild 1 1 1 1
Steeleye Stout 1 1 3 3
Tarte au sucre 0 0 1 1
Teatime Chocolate 0 0 0 0
Thüringer Rostbra 0 0 2 2
Tofu 1 1 2 2
Tourtière 0 0 1 1
Tunnbröd 0 0 1 1
Uncle Bob's Orgar 1 1 1 1
Valkoinen suklaa 0 0 1 1
Vegie-spread 0 0 3 3
Wimmers gute Ser 0 0 5 5
Zaanse koeken 0 0 0 0
Total 34 34 125 125

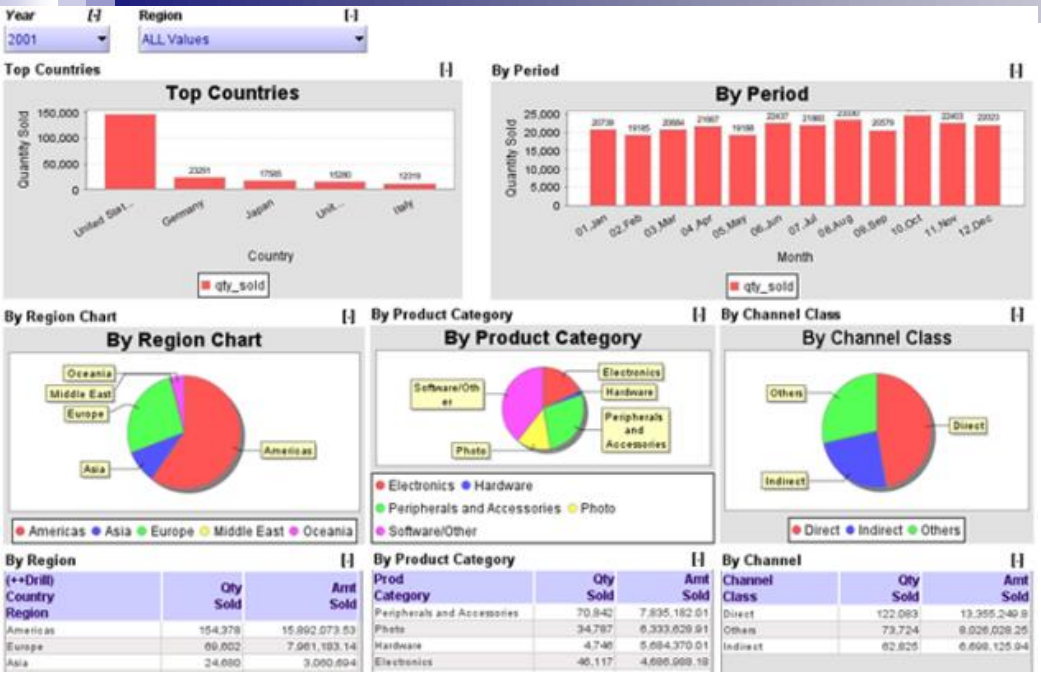

OlapChart
OlapGrid
Normal
Internet Sales Amount
FY 2002
FY 2003
FY 2004
FY 2005
Total
Australia $2,568,701.39 $2,099,585.43 $4,383,479.54 $9,234.23 $9,061,000.58
Canada $573,100.97 $305,010.69 $1,088,879.50 $10,853.70 $1,977,844.86
France $414,245.32 $633,399.70 $1,592,880.75 $3,491.95 $2,644,017.71
Germany $513,353.17 $593,247.24 $1,784,107.09 $3,604.83 $2,894,312.34
United Kingdom $550,507.33 $696,594.97 $2,140,388.50 $4,221.41 $3,391,712.21
United States $2,452,176.07 $1,434,296.26 $5,483,882.67 $19,434.51 $9,389,789.51
Total $7,072,084.24 $5,762,134.30 $16,473,618.05 $50,840.63 $29,358,677.22

Dashboard

- São relatórios de alto nível
- Extrema importância
- Abrange vários níveis do negócio

Dashboard - Exemplos




Year
2001
Region
ALL Values
Top Countries
Top Countries
Quantity Sold
150,000
100,000
50,000
0
23291
17585
15280
12319
United Stat...
Germany
Japan
Unit...
Italy
Country
qty_sold
By Period
By Period
Quantity Sold
25,000
20,000
15,000
10,000
5,000
0
01 Jan
02 Feb
03 Mar
04 Apr
05 May
06 Jun
07 Jul
08 Aug
09 Sep
10 Oct
11 Nov
12 Dec
Month
qty_sold
By Region Chart
By Region Chart
Oceania
Middle East
Europe
Americas
Asia
Americas Asia Europe Middle East Oceania
By Product Category
By Product Category
Software/Oth er
Electronics
Hardware
Peripherals and Accessories
Photo
Electronics Hardware
Peripherals and Accessories Photo
Software/Other
By Channel Class
By Channel Class
Others
Direct
Indirect
Direct Indirect Others
By Region
(++Drill) Country Region
Qty Sold
Amt Sold
Americas 154,378 15,892,073.53
Europe 69,602 7,961,183.14
Asia 24,680 3,060,694
By Product Category
Prod Category
Qty Sold
Amt Sold
Peripherals and Accessories 70,842 7,935,182.01
Photo 34,787 6,333,629.91
Hardware 4,746 5,684,370.01
Electronics 46,117 4,686,969.19
By Channel
Channel Class
Qty Sold
Amt Sold
Direct 122,083 13,355,249.9
Others 73,724 9,026,028.25
Indirect 62,825 6,698,125.94


AdminLTE
Jane Doe
Hello, Jane
Online
Search...
Dashboard
Widgets
Charts
UI Elements
Forms
Tables
Calendar
Mailbox
Examples
Dashboard
Control panel
Home > Dashboard
150
New Orders
More info
53%
Bounce Rate
More info
44
User Registrations
More info
65
Unique Visitors
More info
Server Load
Bandwidth
Transfered
Activity
FTP
Generate PDF
2006
2009
2012
CPU
Disk
RAM
Visitors
Country
Visitors
Online
Page Views
USA 209 239
India 131 958
Britain 19 417
Brazil 109 476
China 192 437
Sales
Donut
Area

BI&A → KDD

- A disponibilização dos dados não basta, sendo vital a interpretação, análise e relacionamento destes dados para que se possa desenvolver estratégias de ação.
- Para atender este novo contexto, há a Descoberta de Conhecimento em Bases de Dados - KDD – ***K****nowledge* ***D****iscovery in* ***D****atabases*

KDD - Conceito

- KDD – *Knowledge Discovery in Databases* [Fayyad]: “É um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de conjuntos de dados”.

Etapas do Processo de KDD (forma resumida)

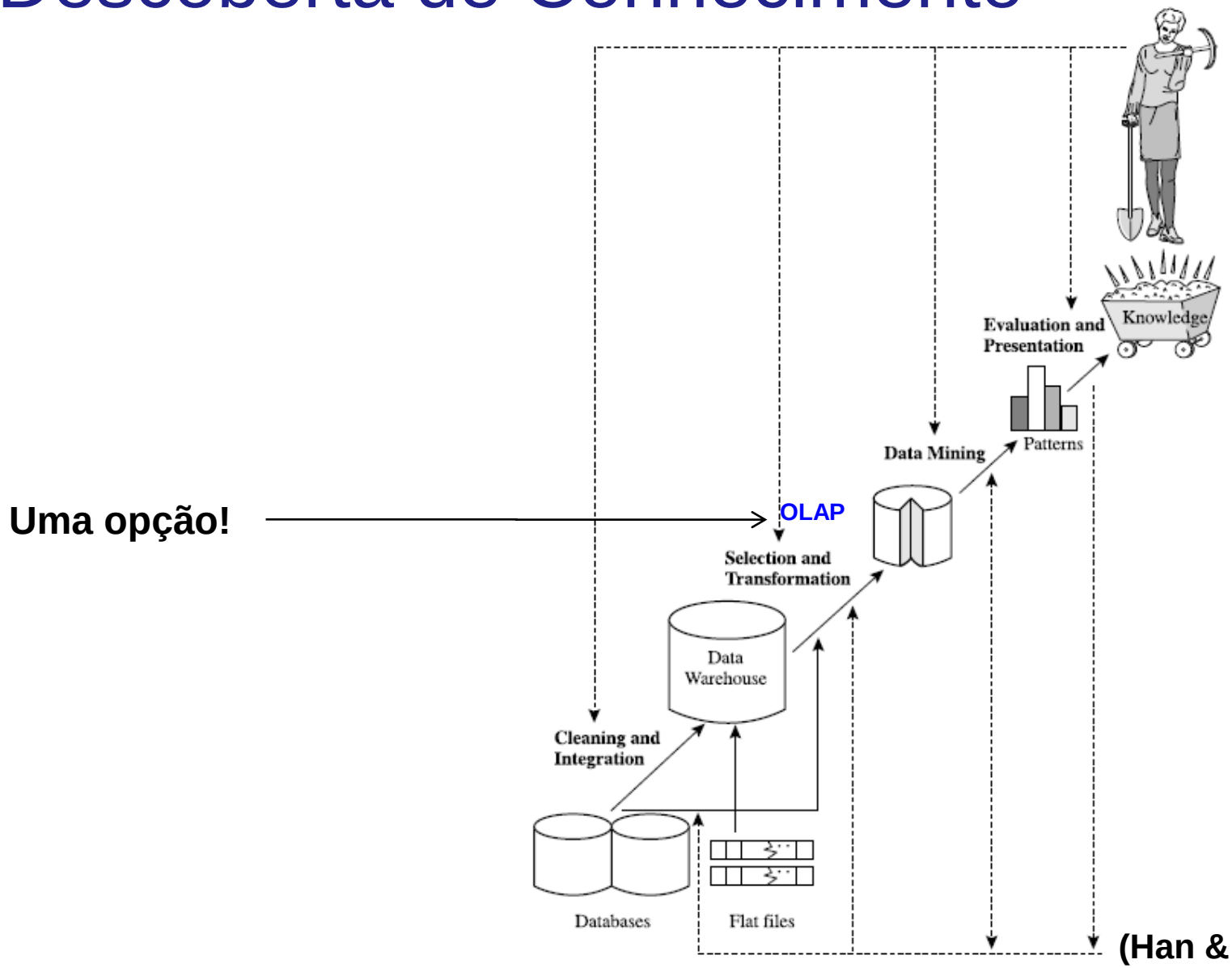
Analista humano: orienta a execução do processo

Captação
Organização
Tratamento

Busca de conhecimento útil

Tratamento do conhecimento obtido

Descoberta de Conhecimento



(Han & Kamber, 2006)

Tarefas de Mineração de Dados

- Mineração de Dados: Principal etapa do KDD: envolve a aplicação de algoritmos sobre os dados em busca de conhecimentos.

- Tarefas Preditivas
  - Classificação → incluindo Descoberta de Desvios e Previsão de Séries
  - Regressão

- Tarefas Descritivas
  - Regras de Associação → incluindo Associações Temporais
  - Agrupamentos
  - Sumarização

Aprendizagem supervisionada e não supervisionada

- **Aprendizagem supervisionada (classificação)**
  - **Supervisão: O conjunto de treinamento (observações, medições, etc.) é acompanhado dos rótulos indicativos das classes das observações;**
  - **Novos dados são classificados com base na estrutura gerada a partir do conjunto de treinamento;**

Aprendizagem supervisionada e não supervisionada

- **Aprendizagem não supervisionada (agrupamento = *clustering*)**
  - **Os rótulos das classes no conjunto de treinamento são desconhecidos;**
  - **Dado um conjunto de medidas, observações, etc. o objetivo é estabelecer a existência de classes ou grupos nos dados.**

Preparação da Entrada

- Problema: fontes diferentes de dados

  (ex.: departamento de vendas, departamento de cobrança, ...)

  - Diferenças: estilos de manter os registros, convenções, períodos de tempo, agregação dos dados, chaves primárias, erros;
  - Os dados precisam ser integrados e limpos;
  - Data warehouse.

- Denormalização não é o único problema
- Dados externos podem ser necessários
- Crítico: tipo e nível de agregação dos dados

O que é necessário fazer?
Corrigir / selecionar / transformar / limpar ...

A Tarefa de Associação

Regras de Associação

(Han & Kamber, 2006)

- Representação: considere seu universo como sendo o conjunto de produtos (itens) vendidos.
  - A existência ou ausência de cada um desses itens pode ser representada por uma variável booleana.
  - Cada compra pode ser representada por um vetor de variáveis booleanas, sendo que, de fato, nesta compra (transação) foram comprados apenas os itens valorados com *verdadeiro*.
  - Analisando esses vetores, é possível descobrir itens que frequentemente aparecem juntos (estão associados), constituindo um padrão de comportamento.
  - Esse “padrão” pode ser representado por meio de uma **regra de associação.**

antecedente → consequente

Regras de Associação

(Han & Kamber, 2006)

- Medidas de "*interessabilidade*"

computador → antivirus [suporte = 2%, confiança = 60%]

Suporte = utilidade da regra

Confiança = certeza sobre a regra

O suporte de 2% para a regra acima significa que 2% de todas as transações analisadas mostram que computadores e antivirus são comprados juntos.

A confiança de 60% da regra acima significa que 60% dos fregueses que compram um computador também compram um antivirus.

Regras de associação interessantes são aquelas que possuem um suporte e uma confiança mínimos (de acordo com um limite inferior pré-estabelecido).

Regras de Associação (Han & Kamber, 2006)

Processo de mineração de regras de associação:

Representações de Dados Transacionais

Base de dados transacional T_{ID}

$T_1 = \{i_1, i_3\}$
$T_2 = \{i_1, i_2, i_3\}$
$T_3 = \{i_1\}$
$T_4 = \{i_1, i_3, i_4\}$
$T_5 = \{i_3\}$
$T_6 = \{i_1, i_2\}$
$T_7 = \{i_1, i_2, i_3, i_4\}$

(a)

T_{ID}	i_1	i_2	i_3	i_4
T_1	1	0	1	0
T_2	1	1	1	0
T_3	1	0	0	0
T_4	1	0	1	1
T_5	0	0	1	0
T_6	1	1	0	0
T_7	1	1	1	1

(b)

Exemplo de base transacional genérica: (a) representação por conjuntos; (b) representação matricial

Base de dados transacional TID

$T_1 = \{i_1, i_3\}$
$T_2 = \{i_1, i_2, i_3\}$
$T_3 = \{i_1\}$
$T_4 = \{i_1, i_3, i_4\}$
$T_5 = \{i_3\}$
$T_6 = \{i_1, i_2\}$
$T_7 = \{i_1, i_2, i_3, i_4\}$
$T_8 = \{i_1, i_2, i_3, i_4\}$

4-*itemset*	Suporte
$\{i_1, i_2, i_3, i_4\}$	25%

3-*itemset*	Suporte
$\{i_1, i_2, i_3\}$	37%
$\{i_1, i_3, i_4\}$	37%
$\{i_1, i_2, i_4\}$	25%
$\{i_2, i_3, i_4\}$	25%

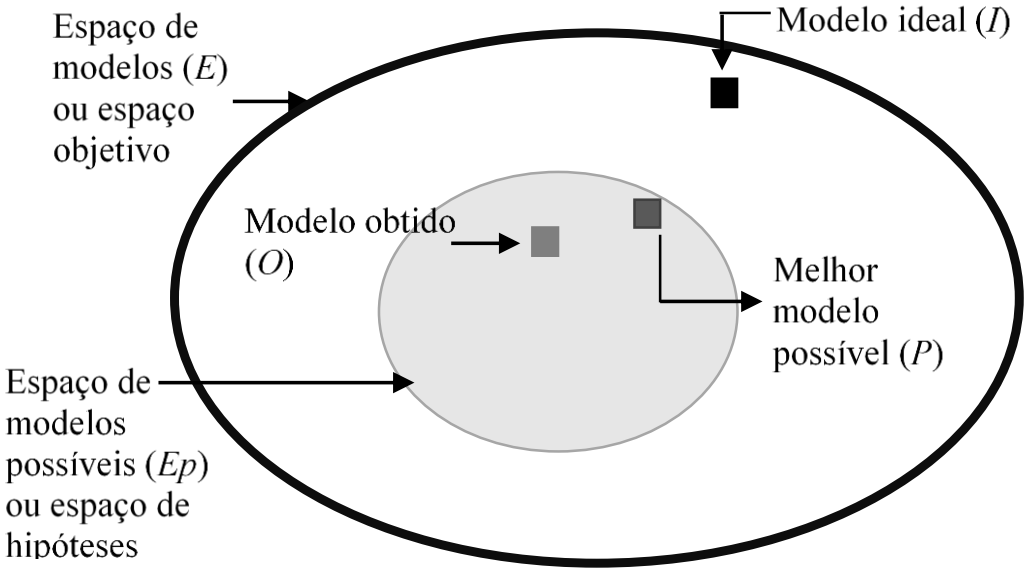
2-*itemset*	Suporte
$\{i_1, i_2\}$	50%
$\{i_1, i_3\}$	62%
$\{i_1, i_4\}$	37%
$\{i_2, i_3\}$	37%
$\{i_2, i_4\}$	25%
$\{i_3, i_4\}$	37%

1-*itemset*	Suporte
$\{i_1\}$	87%
$\{i_2\}$	50%
$\{i_3\}$	75%
$\{i_4\}$	37%

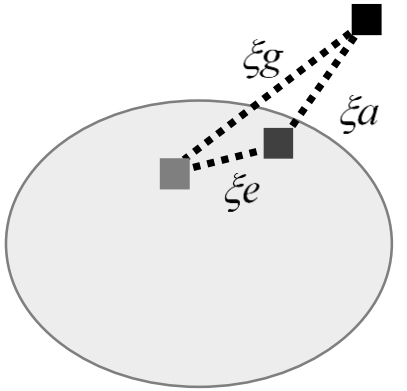
Novo contexto para a base de dados transacional; os *itemsets* e seus respectivos suportes. Note que transações diferentes podem envolver o mesmo subconjunto de itens

A tarefa de Classificação/Predição

Introdução



(a)



ξg - Erro de generalização
ξa - Erro de aproximação
ξe - Erro de estimação

(b)

Espaço de modelos (a); erros de predição (b); baseado em Lima (2004)

Coleção de dados

Podemos armazenar *características* em bases de dados

O problema de classificação agora pode ser expresso da seguinte forma:

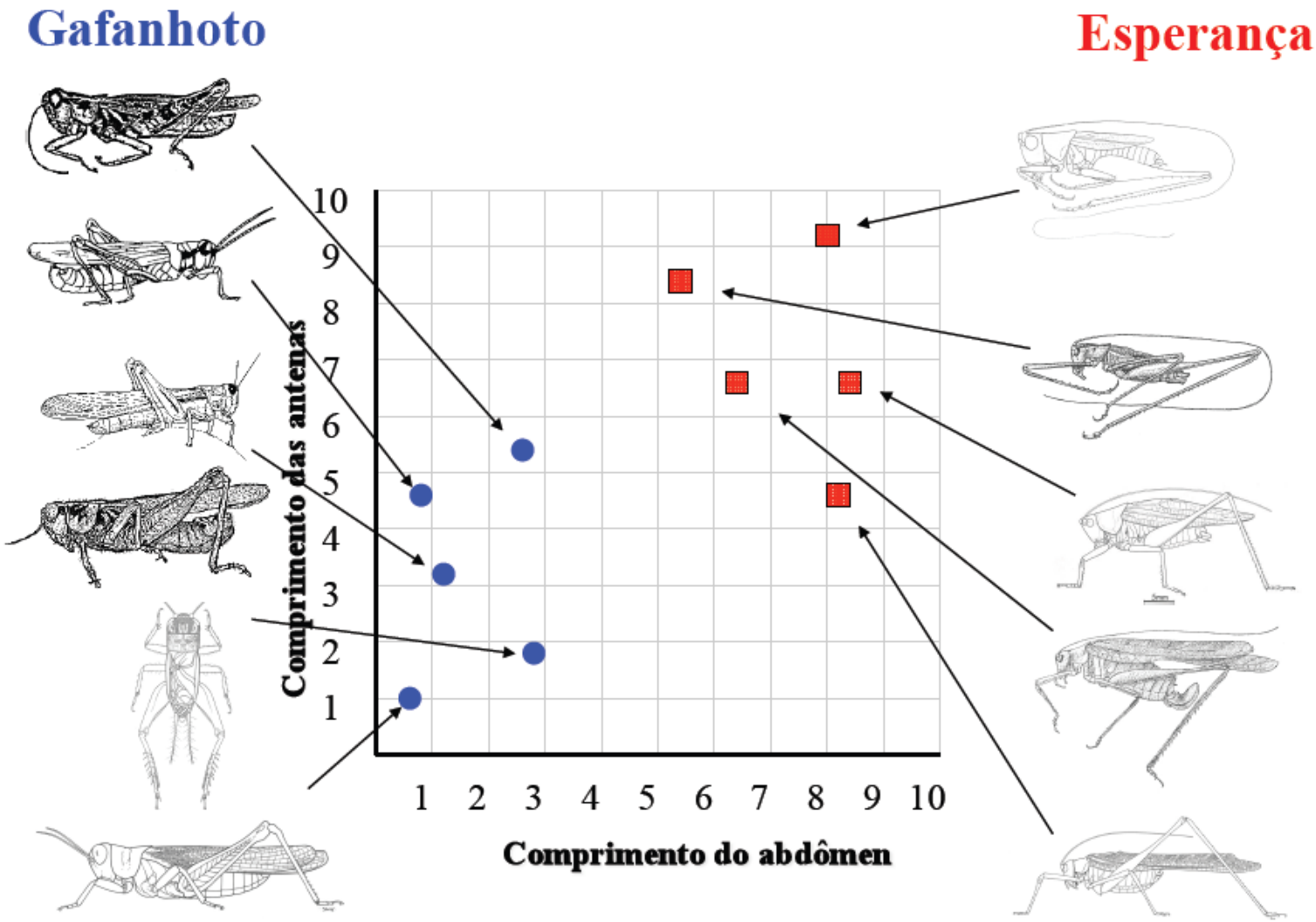
• Dada uma base de treinamento (Minha_Coleção), prediga o rótulo da classe dos exemplos ainda não vistos

Minha_Coleção

ID do inseto	Comp. do abdômen	Comp. das antenas	Classe do inseto
1	2.7	5.5	Gafanhoto
2	8.0	9.1	Esperança
3	0.9	4.7	Gafanhoto
4	1.1	3.1	Gafanhoto
5	5.4	8.5	Esperança
6	2.9	1.9	Gafanhoto
7	6.1	6.6	Esperança
8	0.5	1.0	Gafanhoto
9	8.3	6.6	Esperança
10	8.1	4.7	Esperança

Exemplo não visto =	*11*	5.1	7.0	???????

Visualização gráfica

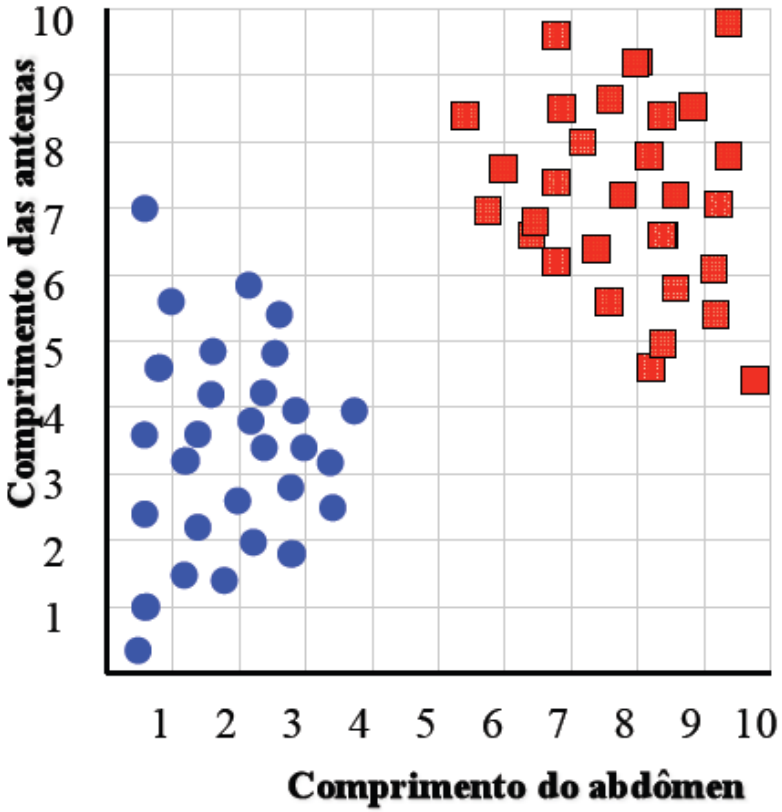


Visualização Gráfica

Gafanhoto

Esperança

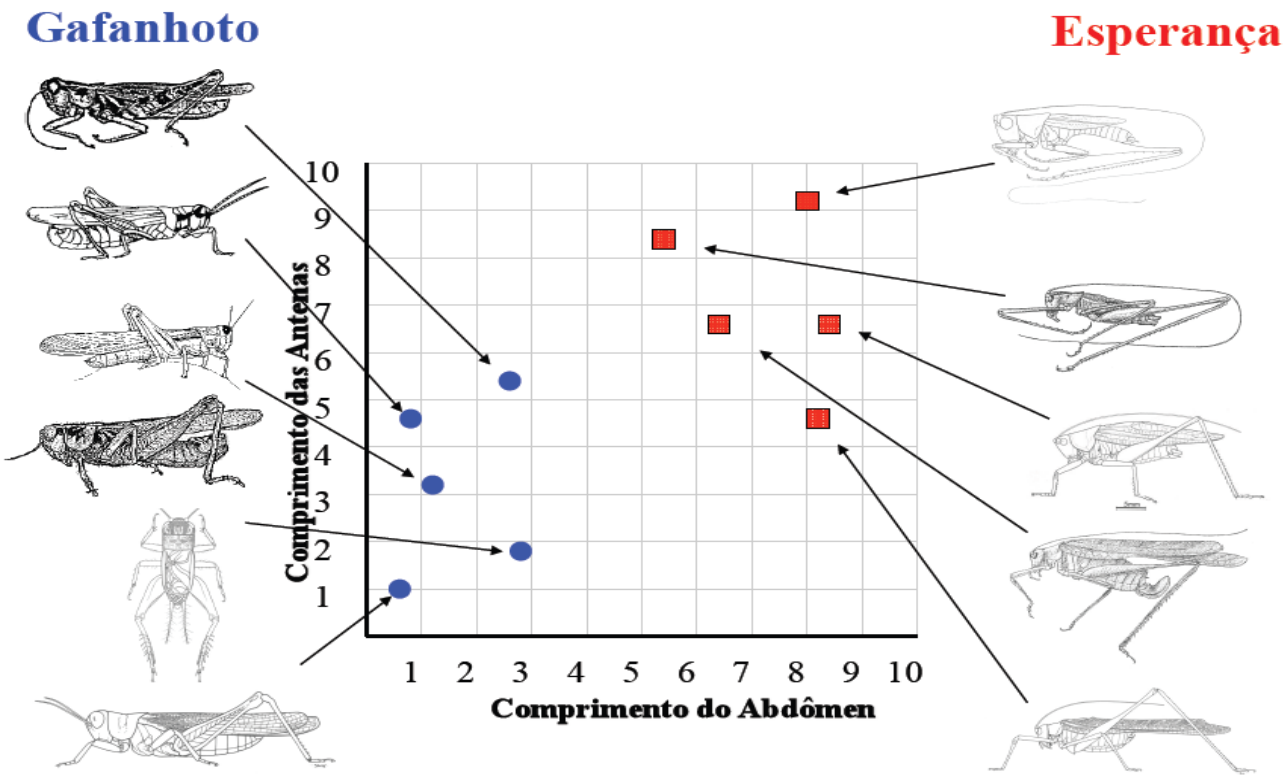
Também utilizaremos esta base de dados maior para motivação …



Cada um destes objetos de dados é chamado de…

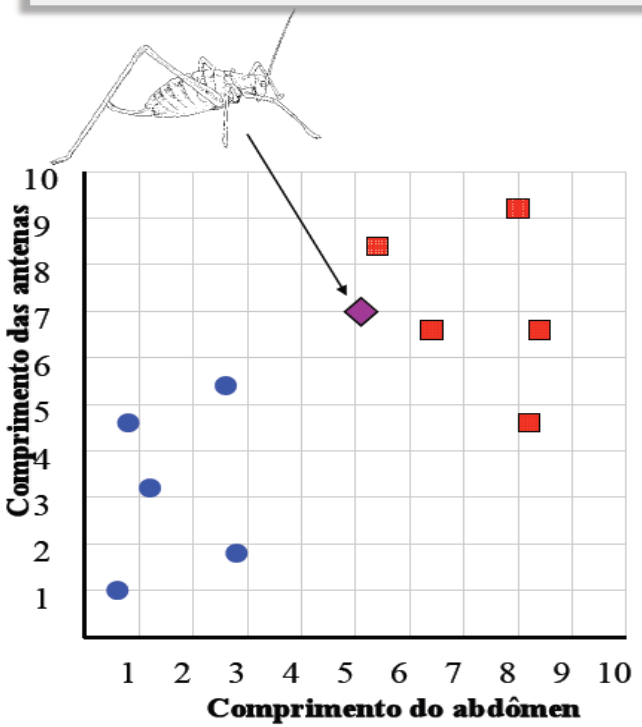
- exemplar
- exemplo (de treinamento)
- instância
- tupla

Problema Inicial



Gafanhoto ou Esperança

Exemplo não visto antes =	*11*	5.1	7.0	???????

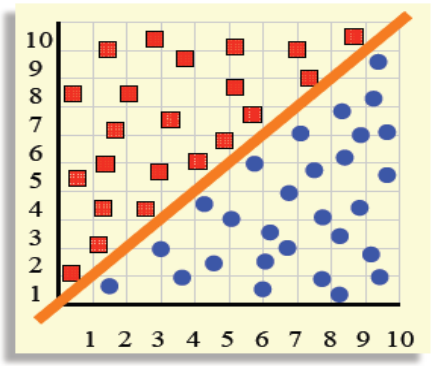


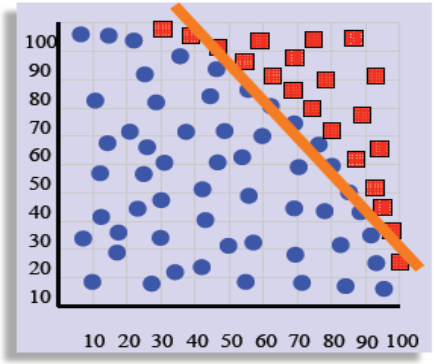
- Podemos "projetar" o **exemplo não visto antes** dentro do mesmo espaço que a base de dados.
- Acabamos de abstrair os detalhes do nosso problema particular. Será muito mais fácil conversar sobre pontos no espaço.

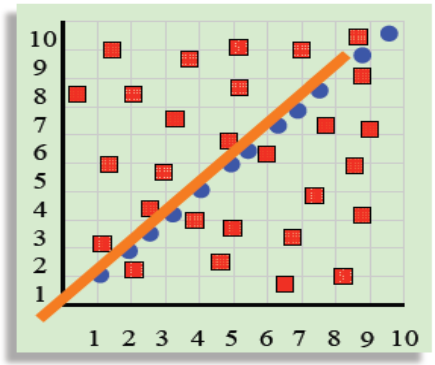
■ Esperança
● Gafanhoto

Classificadores Lineares

1) Perfeito
2) Inútil
3) Muito bom







Problemas que podem ser resolvido por um classificador linear são chamados de **linearmente separáveis**.

Distribuição espacial de atributos

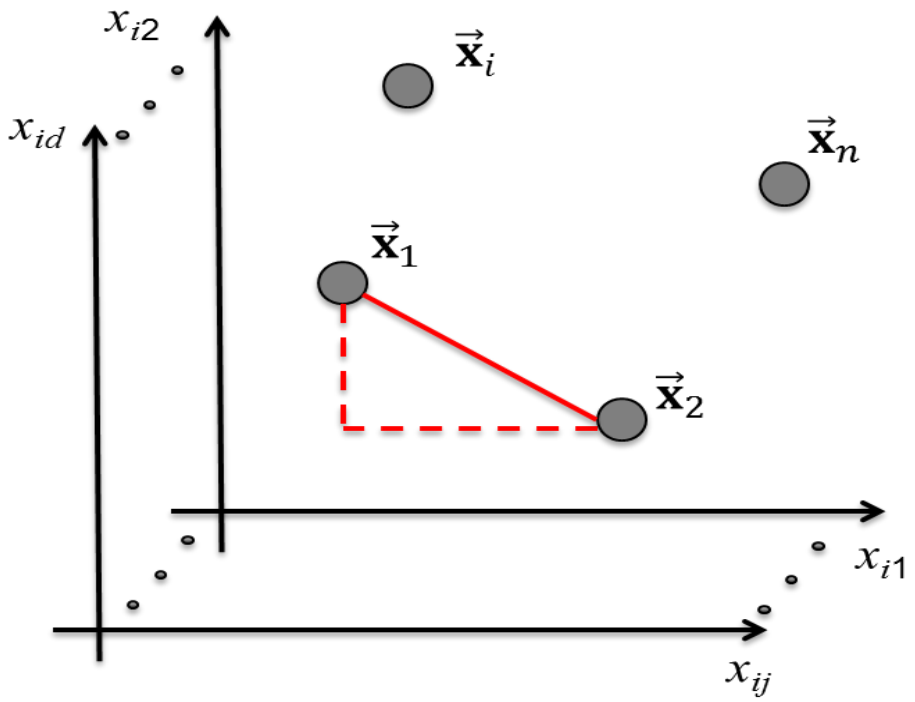
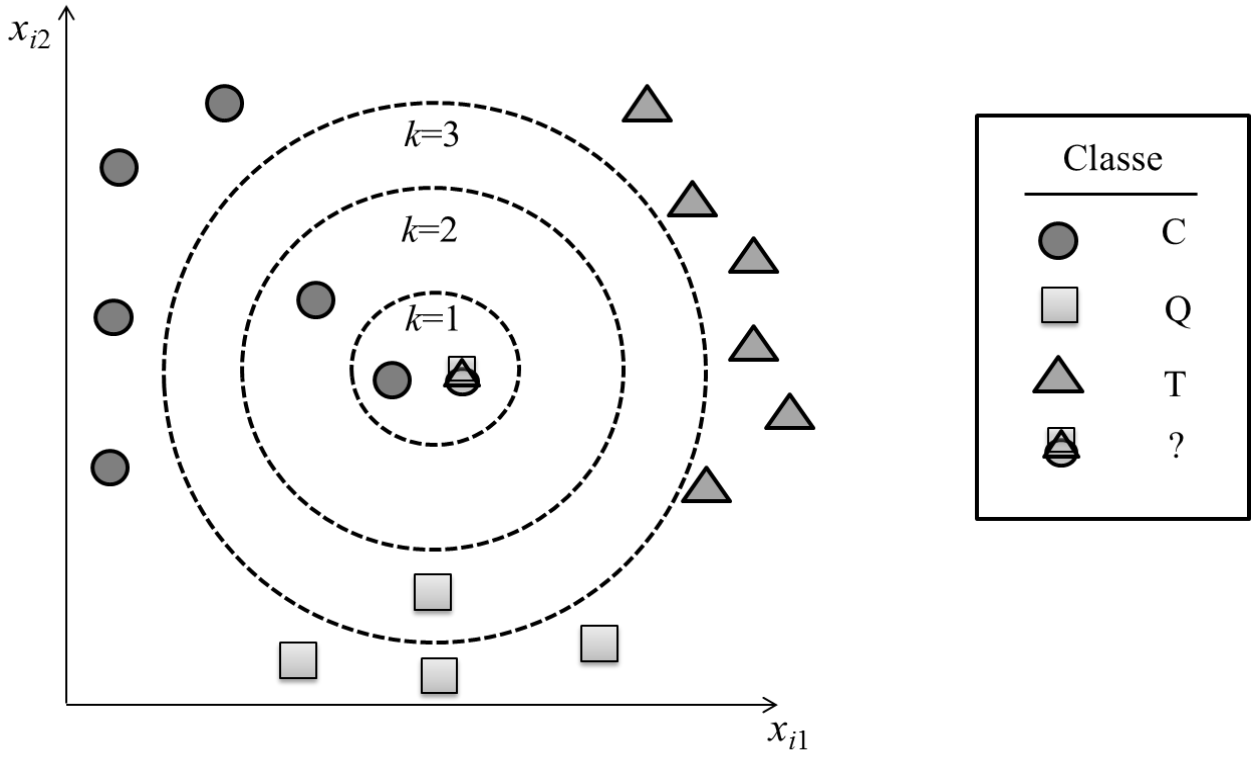


Gráfico de dispersão como uma abstração dos exemplares distribuídos no espaço dos atributos

Exemplo algoritmo kNN



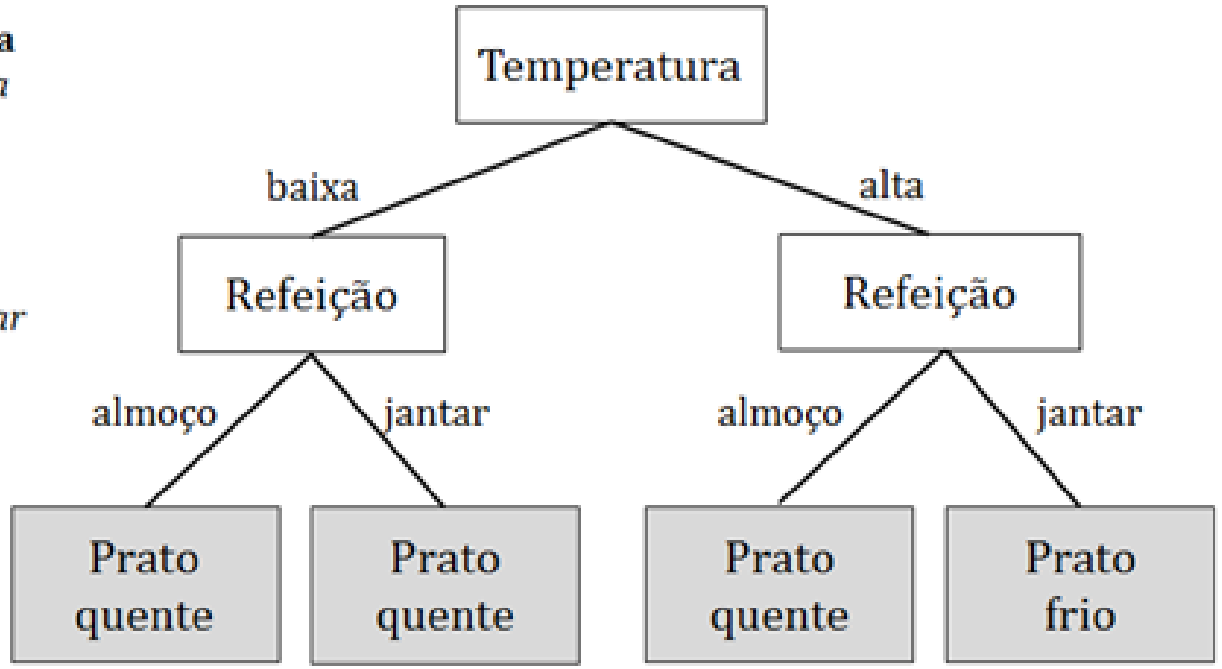
Exemplo ilustrativo para o *k-NN*

Árvore de Decisão

Atributo descritivo = **temperatura**
Valores de temperatura: *alta, baixa*

Atributo descritivo = **refeição**
Valores de refeição: *almoço, jantar*

Atributo de rótulo: **tipo de prato**
Valores de rótulo:
Prato quente, prato frio



Exemplo de um classificador hipotético, na forma de uma árvore de decisão

Treinamento e Teste

- O desempenho de um classificador pode ser medido por meio da taxa de erro:
  - A taxa de erro é a proporção de erros obtidos sobre um conjunto completo de instâncias.

O classificador prediz a classe de cada instância; se ela é correta, é contada como um "sucesso"; se não, é contada como um "erro".

  - O que interessa é o desempenho do classificador mediante "novos" dados, e não sobre os dados velhos (usados no processo de treinamento).

Treinamento, validação e teste

- Frequentemente é útil dividir o conjunto de dados disponíveis em três partes, para três diferentes propósitos:
  - Conjunto de treinamento: usado por um ou mais métodos de aprendizado para construir o classificador.
  - Conjunto de validação: usado para otimizar os parâmetros do classificador, ou para selecionar um em particular.
  - Conjunto de teste: usado para calcular a taxa de erro final do modelo já otimizado.

Uma vez que a taxa de erro foi determinada, os dados de testes podem se juntar aos dados de treinamento para produzir um novo classificador para o uso real. Não há problema nisso quando usado apenas como uma forma de maximizar o classificador que será usado na prática. O que é importante é que a taxa de erro não seja calculada com base nesse último classificador gerado. Além disso, o mesmo pode ser feito com os dados de validação. (Witten & Frank, 2005)

Análise de desempenho

Meu classificador apresentou 25% de taxa de erro (75% de taxa de acerto). Mas o que isso realmente significa? Quanto posso confiar nesta medida?

É util determinar a taxa de sucesso com relação a um intervalo de confiança.

Seja S a contagem de respostas corretas obtidas nos testes do classificador e N o número de testes realizados, então:

- **Se S = 750 e N = 1000, a taxa de sucesso é por volta de 75%. Se considerarmos 80% de confiança na medida, a taxa de sucesso fica entre 73.2% e 76.7%.**

- **Se S = 75 e N = 100, a taxa de sucesso é por volta de 75%. Se considerarmos 80% de confiança na medida, a taxa de sucesso fica entre 69.1% e 80.1%.**

? →

Matriz de Confusão

- Oferece uma medida da eficácia do modelo de classificação,mostrando o número de classificações corretas *versus o* número de classificação prevista para cada classe.

Classe	C_1 Prevista	C_2 Prevista	⋯	C_k Prevista
C_1 Real	$M(C_1,C_1)$	$M(C_1,C_2)$	⋯	$M(C_1,C_k)$
C_2 Real	$M(C_2,C_1)$	$M(C_2,C_2)$	⋱	$M(C_2,C_k)$
⋮	⋮	⋮	⋯	⋮
C_k Real	$M(C_k,C_1)$	$M(C_k,C_2)$	⋯	$M(C_k,C_k)$

$$M(C_i,C_j) = \sum_{\{\forall(x,y)\in T: y=C_i\}} \left\| h(x) = C_j \right\|$$

Matriz de Confusão para duas classes

Classe	prevista C_+	prevista C_-	Taxa de erro da classe	Taxa de erro total
real C_+	T_p	F_n	$F_n / (T_p + F_n)$	$(F_p + F_n) / n$
real C_-	F_p	T_n	$F_p / (F_p + T_n)$	

TP = True Positive (verdadeiro positivo)

FN = False Negative(falso negativo)

FP = False Positive (falso positivo)

TN = True Negative (verdadeiro negativo)

n = (TP+FN+FP+TN)

Matriz de Confusão para duas classes

- Outras métricas derivadas da tabela anterior:

C_+ Predictive Value = T_p / (T_p + F_p)
C_- Predictive Value = T_n / (T_n + F_n)
True C_+ Rate ou Sesitivity y ou Recall = T_p / (T_p + F_n)
True C_- Rate ou Specifity = T_n / (F_p + T_n)
Precision = (T_p + T_n) / n

Avaliação do classificador

- Para estimar o erro verdadeiro de um classificador, a amostra para teste deve ser aleatoriamente escolhida
- Amostras não devem ser pré-selecionadas de nenhuma maneira
- Para problemas reais, tem-se uma amostra de uma única população, de tamanho *n,* e a tarefa é estimar o erro verdadeiro para essa população

Métodos para estimar o erro verdadeiro de um classificador

- Resubstitution
- Random
- Holdout
- r-fold cross-validation
- r-fold stratified cross-validation
- Leave-one-out
- Bootstrap

Holdout **(Witten & Frank, 2005)**

- Estratégia para teste de classificador que reserva um certo montante de dados para treino e o restante para teste (podendo ainda usar parte para validação).
- Comumente esta estratégia uma 1/3 dos dados dados para teste e o restante para treinamento, escolhido randomicamente.
- É interessante assegurar que a amostragem randômica seja feita de tal maneira que garanta que cada classe é apropriadamente representada tanto no conjunto de treinamento quanto no conjunto de teste. Este procedimento é chamado de *estratificação* (*holdout estratificado*).
- Também é útil, para amenizar tendências, repetir todo o processo de treino e teste várias vezes com diferentes amostragens randômicas (holdout repetitivo/iterativo).

Cross Validation

(Witten & Frank, 2005)

- Trata-se de uma estratégia para lidar com um montante de dados limitado.

- Nesta estratégia decide-se um numero fixos de folds, ou partições dos dados. Supondo que sejam usados três folds (3-fold cross validation):
  - o conjunto de dado é dividido em três partições de tamanhos aproximadamente iguais e, de maneira rotativa, cada uma delas é usada para teste enquanto as duas restantes são usadas para treinamento.
  - ou seja: use 2/3 para treinamento e 1/3 para teste e repita o processo três vezes, tal que, no fim, cada instância tenha sido usadas exatamente uma vez para teste.
  - se a estratificação é adotada, então o procedimento se chama 3-fold cross validation estratificado (aconselhável).
  - o padrão é executar o 10-fold cross validation, 10 vezes.
  - o erro final do classificador é a média dos erros obtidos em cada iteração da estratégia cross-validation

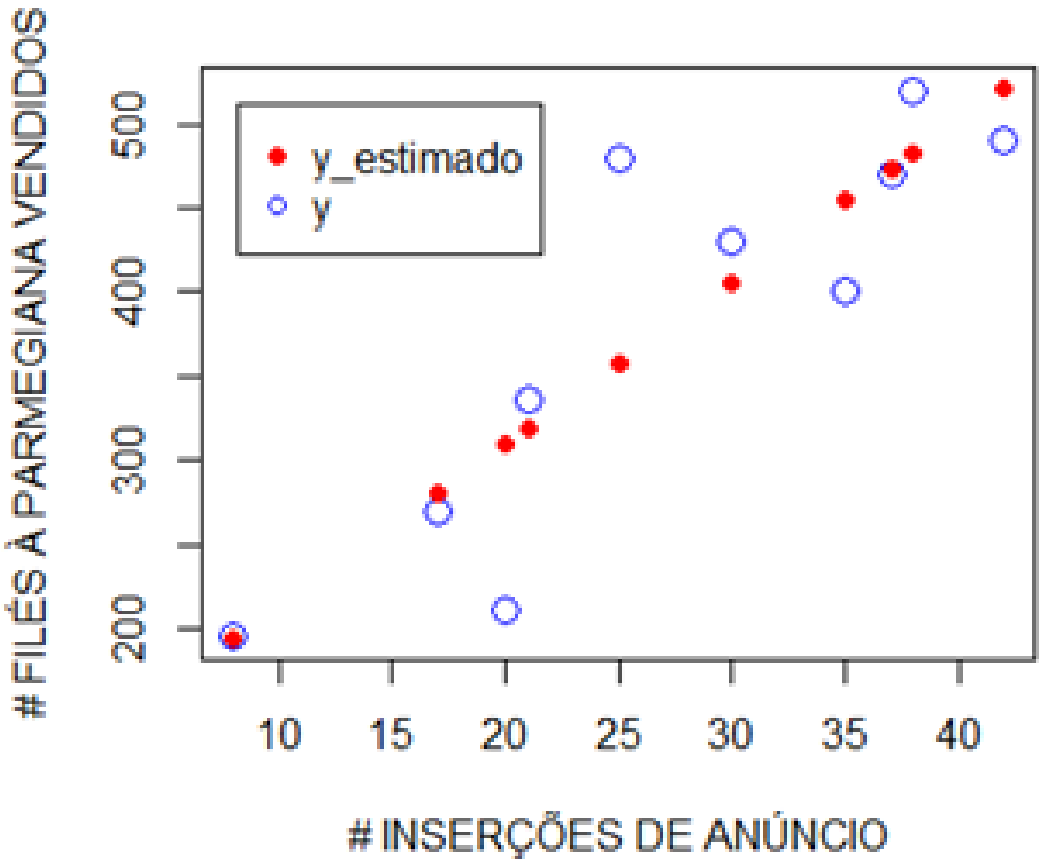
Leave-one-out **(Witten & Frank, 2005)**

- Leave-one-out cross-validation é um n-fold cross-validation, onde n é o número de instâncias no conjunto de dados.
- A avaliação é sobre a corretude de classificação da instância em teste – um ou zero para sucesso ou falha, respectivamente.
- Os resultados de todas as n avaliações, uma para cada instância do conjunto de dados, são analisados via média, e tal média representa o erro final estimado.

- Motivações:
  - o maior número possível de dados é usado para treinamento em cada caso, o que aumenta as chance do classificador alcançar acuidade.
  - o procedimento é determinístico.

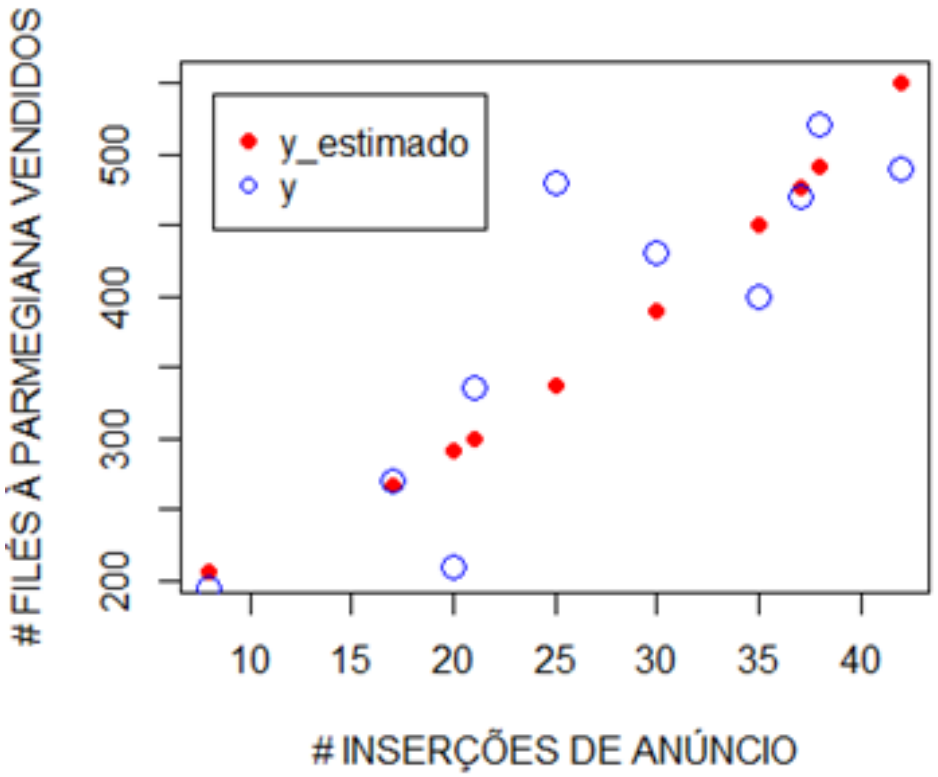
- Indicado para conjunto de dados pequenos.
- Não é possível aplicar qualquer procedimento de estratificação.

Exemplo de predição linear



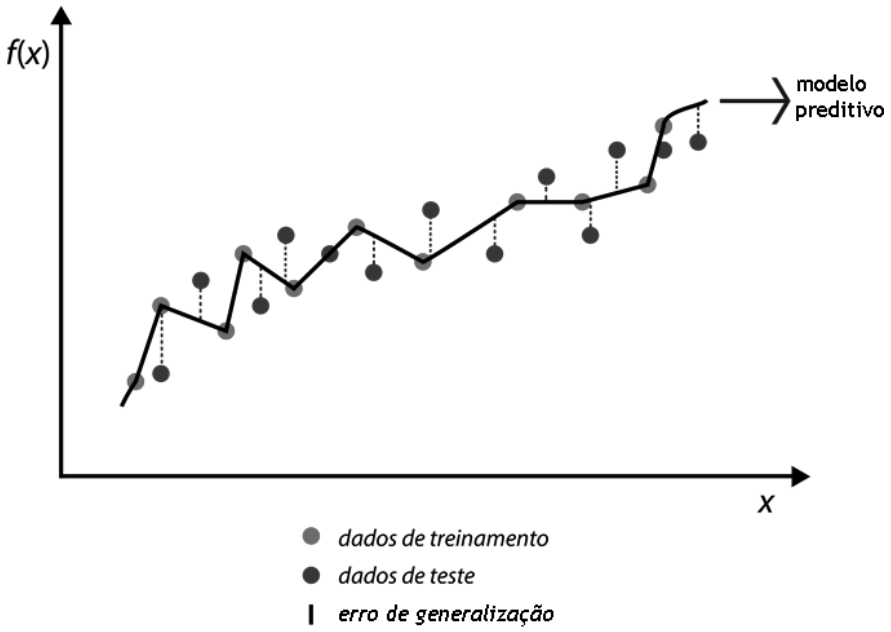
Representação gráfica para o modelo de regressão linear obtido para o conjunto PLANEJAMENTO DE PROPAGANDA

Exemplo de predição não linear

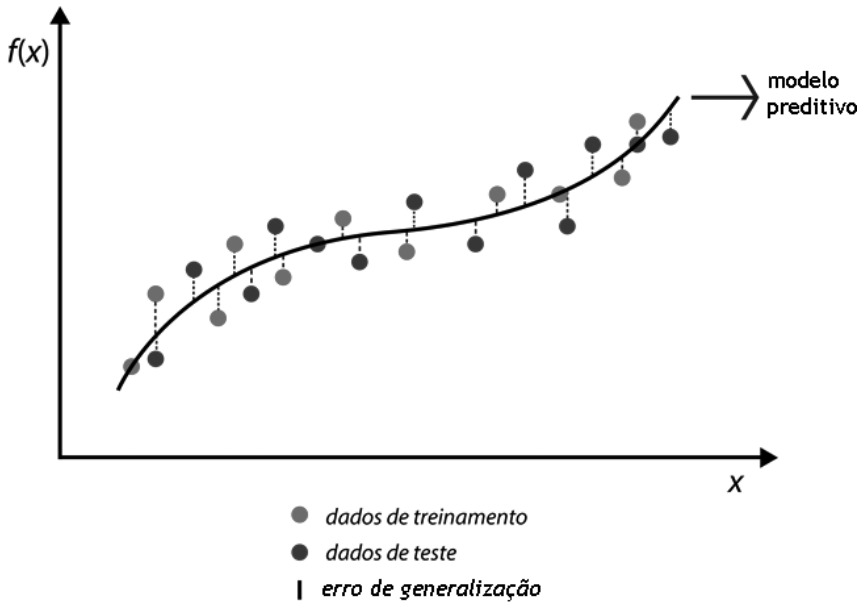


Representação gráfica para o modelo de regressão não linear obtido para o conjunto PLANEJAMENTO DE PROPAGANDA

Avaliação de modelos preditivos



(a)



(b)

Exemplos de modelo preditivo: (a) com sobreajuste; (b) sem sobreajuste

A tarefa de Agrupamento

Agrupamento (clustering)

Tarefa descritiva que agrupa exemplos (objetos) de acordo com suas características

- Objetivo: agrupar objetos em clusters (agrupamentos) de modo que objetos pertencentes a um mesmo cluster são mais similares entre si de acordo com alguma medida de similaridade pré-definida, enquanto que objetos pertences a clusters diferentes têm uma similaridade menor
- Consumo de um carro em função de suas características
- Valor de um imóvel em função de suas características e do bairro

Agrupamento (clustering)

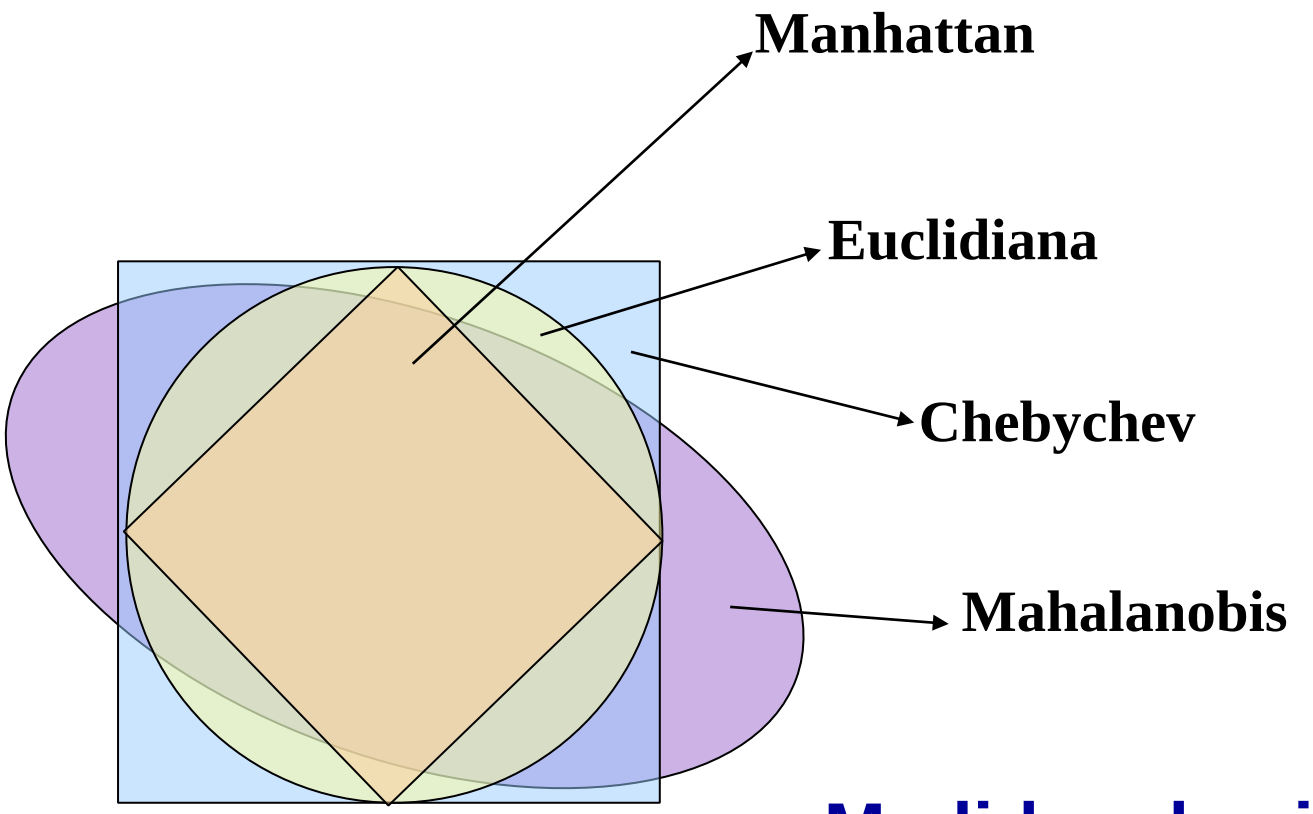
(Han & Kamber, 2006)

- A tarefa de agrupamento consiste em agrupar um conjunto de objetos físicos ou abstratos em grupos de objetos similares.

- Um grupo é uma coleção de objetos que são similares uns aos outros, dentro de um grupo, e dissimilares aos objetos de outros grupos.
  - pode ser considerada uma forma de compressão

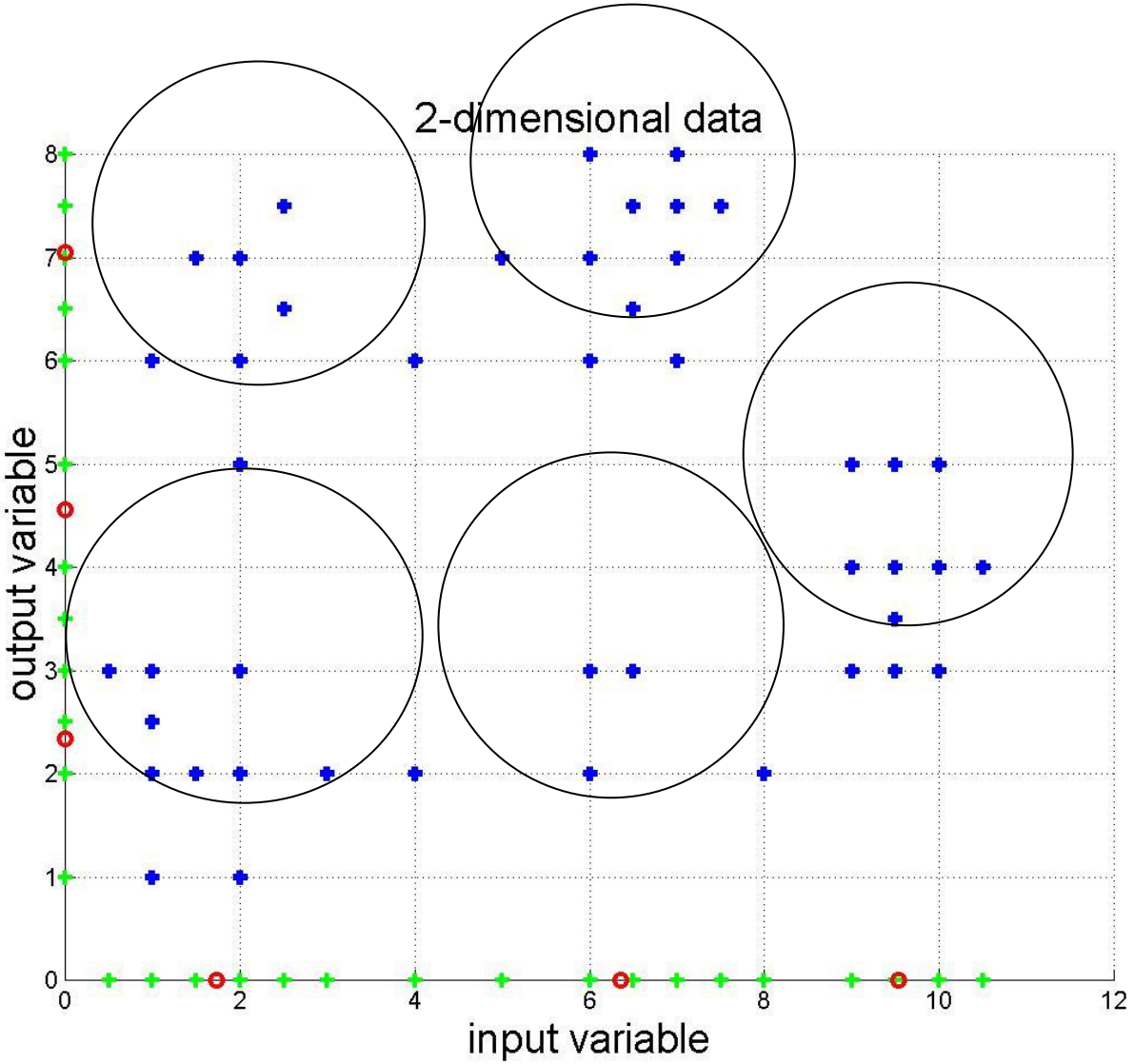
- O modelo de agrupamento não é construído com base em dados rotulados. Apenas a informação de similaridade entre os dados é usada.
  - após o modelo de agrupamento ser construídos, um processo de rotulação dos grupos formados pode ser útil.

Formatos de clusters



Medidas de similaridade

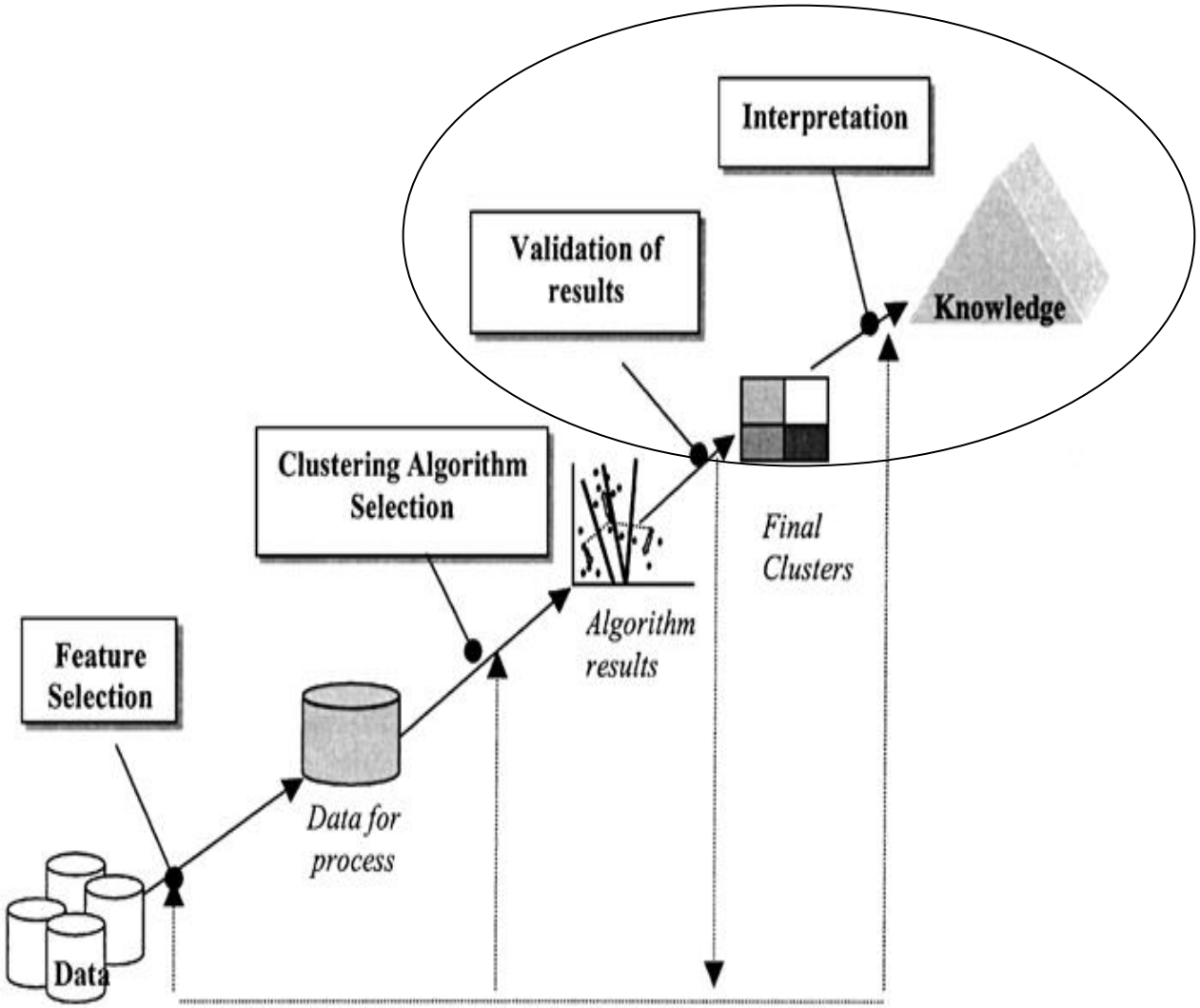
Formatos de clusters



Processo de Agrupamento

1. seleção de exemplos e seleção ou construção de atributos:
   - Seleciona atributos relevantes, ou
   - Constrói atributos representativos
2. Similaridade entre exemplos
   - Seleciona a medida de similaridade a ser utilizada, que deve ser adequada ao domínio
3. Agrupamento
   - Aplicação do algoritmo de agrupamento

Processo de Agrupamento



Categorização dos Métodos de agrupamento

(Han & Kamber, 2006)

- Métodos de particionamento
- Métodos hierárquicos
- Métodos baseados em densidade
- Métodos baseados em gride
- Métodos baseados em modelos
- Agrupamento de dados de alta-dimensão
- Agrupamento baseado em restrições

Métodos de Particionamento

(Han & Kamber, 2006)

- Os clusters são encontrados por meio da otimização de critérios tais como a função de dissimilaridade baseada em distância.
- Para alcançar a otimalidade global, é necessário enumerar, exaustivamente, todas as possíveis partições.
- Porém, dado o custo de tal procedimento, métodos heurísticos são usados:
  - algoritmo **k-means**: onde cada cluster é representado pelo valor médio dos objetos no cluster
  - algoritmo **k-medóides**: onde cada cluster é representado por um dos objetos localizado próximo ao centro do cluster.

K-means e K-medóides são os mais comuns dentro da categoria de particionamento.

Trabalham bem em bases de dados pequenas e médias e encontram clusters de formas esféricas.

K-means

(Han & Kamber, 2006)

- Técnica baseada em centróide.

Objetivo: maximizar a similaridade intracluster e minimizar a similaridade intercluster.

- A similaridade do cluster é medida em relação ao valor médio dos objetos no cluster (centróide ou centro de gravidade).
- Critério do erro quadrado:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

onde E é a soma do erro quadrado (distância) para todos os objetos no conjunto de dados; p é o ponto em um espaço de representação do objeto; mi é a média do clusters Ci (tanto p quanto m são multidimensionais).

K-means - Algoritmo

(Han & Kamber, 2006)

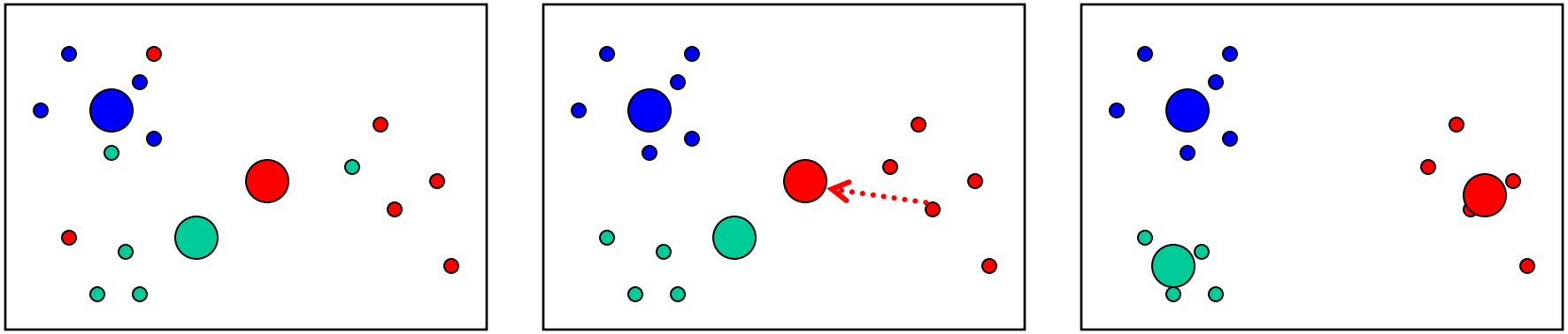
```
Input:
     k: o número de clusters;
     D: o conjunto de dados com n objetos;

Output:
     um conjunto de k clusters (os vetores protótipos de cada
clusters);

Method:
     (1)escolha k objetos de D, arbitrariamente, para representar
             os centros dos clusters (partições) iniciais;

     (2)repeat

     (3)   (re)associe cada objeto para o cluster que tem seu
                  centro mais similar ao objeto;
     (4)   atualize as médias dos clusters, i.e., calcule o valor
                  médio dos objetos para cada cluster;

     (5)until "nenhuma mudança ocorrer" (ou outro critério de
             parada);
```

K-means: Executando…

- Dado um conjunto de pontos numéricos no espaço D-dimensional e um inteiro K;
- O algoritmo gera K (ou menos) clusters da seguinte maneira:

```
Escolha K clusters aleatoriamente
Calcule o centróide para cada cluster
Repita
   Atribua cada ponto ao centróide mais próximo
   Recalcule o centróide para cada cluster
Até estabilidade
```

Ex.: K = 3



K-means: Problemas

- Os clusters finais não representam uma otimização global mas apenas local e clusters diferentes podem surgir a partir da diferença na escolha inicial aleatória dos centróides (1ª Figura);
- O parâmetro K deve ser escolhido antecipadamente, ou vários valores devem ser tentados, até encontrar o "melhor";

Os dados devem ser numéricos e devem ser comparados através da distância Euclideana.

K-means: Problemas

- **O algoritmo trabalha melhor com dados que contêm clusters esféricos; clusters com outra geometria podem não ser encontrados;**

- **O algoritmo é sensível a *outliers* (pontos que não pertencem a nenhum cluster). Esses pontos podem distorcer a posição do centróide e deteriorar o cluster;**

Métodos Hierárquicos

(Han & Kamber, 2006)

- Um método hierárquico cria uma decomposição hierárquica de um conjunto de dados. São classificados em:
  - **aglomerativos (*bottom-up*):** inicia com cada objeto formando um grupo separado e sucessivamente junta os objetos ou grupos que estão mais próximos um do outro, até que apenas um grupos seja formado ou alguma condição de parada seja alcançada.
  - **divisivos (*top-down*):** inicia com todos os objetos no mesmo grupo e a cada iteração, os divide em grupos menores, até que cada objeto esteja em um grupo ou alguma condição de parada seja alcançada.

Clustering Hierárquico: Algoritmo

- Cria uma árvore na qual os objetos são as folhas e os nós internos revelam a estrutura de similaridade dos pontos
  - A árvore é freqüentemente chamada de “dendograma”

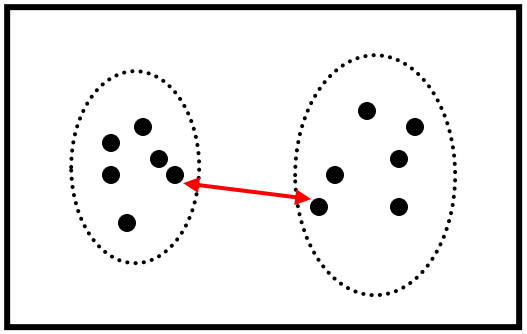
- O algoritmo pode ser resumido da seguinte maneira:

```
Coloque todos os pontos em seus próprios
clusters
Enquanto há mais de um cluster Faça
  Agrupe o par de clusters mais próximos
```

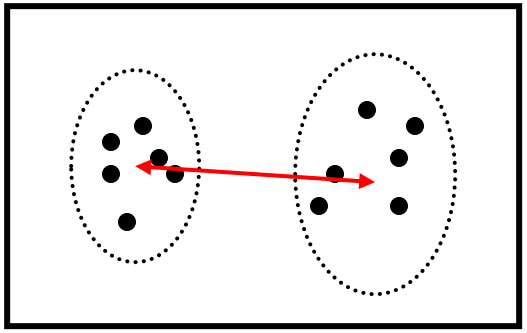
- O comportamento do algoritmo depende de como o “par de clusters mais próximo” é definido.

Clustering Hierárquico: Agrupando Clusters

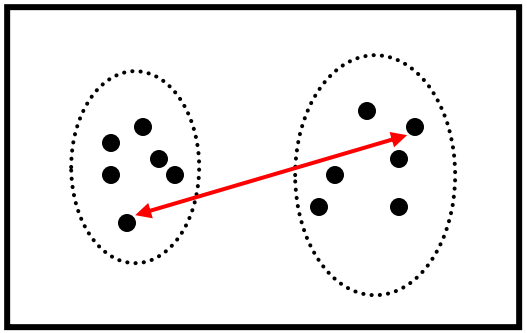
Single Link: Distância entre dois clusters é a distância entre os pontos mais próximos. Também chamado "agrupamento de vizinhos".



Average Link: Distância entre clusters é a distância entre os centróides.



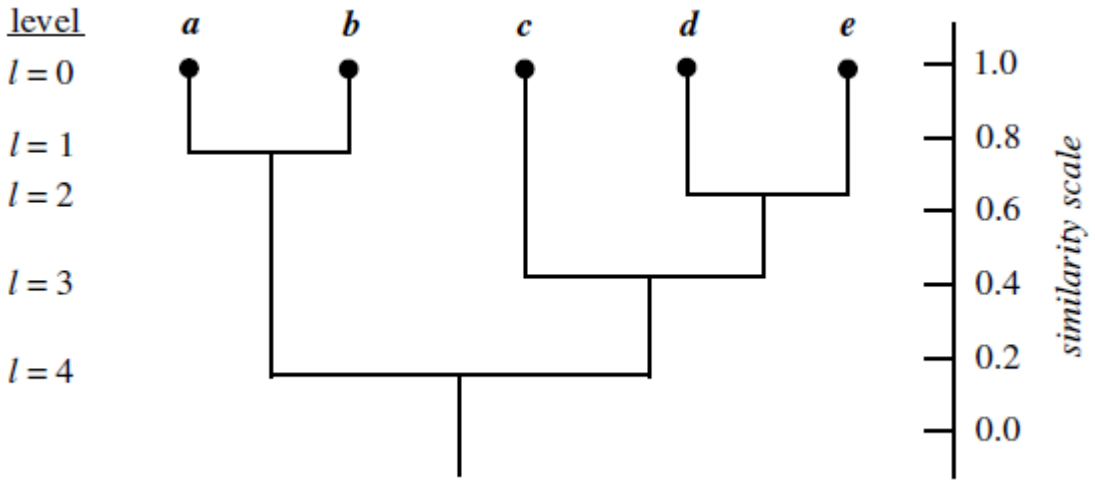
Complete Link: Distância entre clusters é a distância entre os pontos mais distantes.



Métodos Hierárquicos

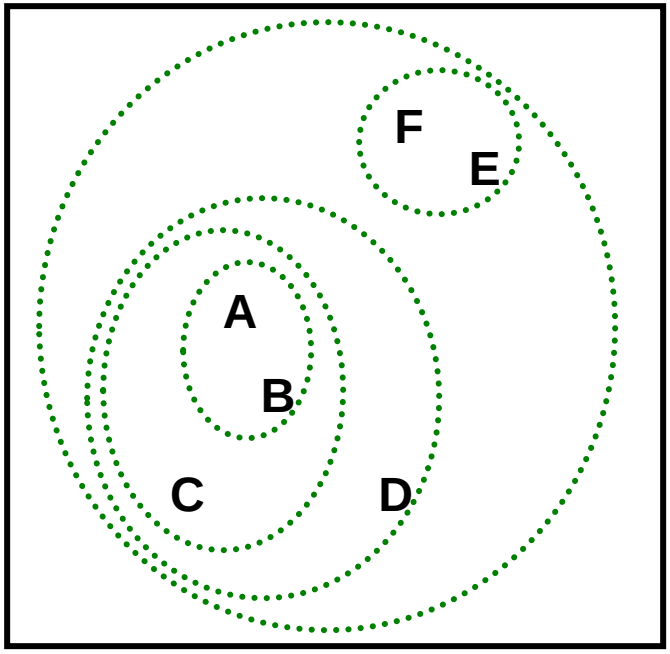
(Han & Kamber, 2006)

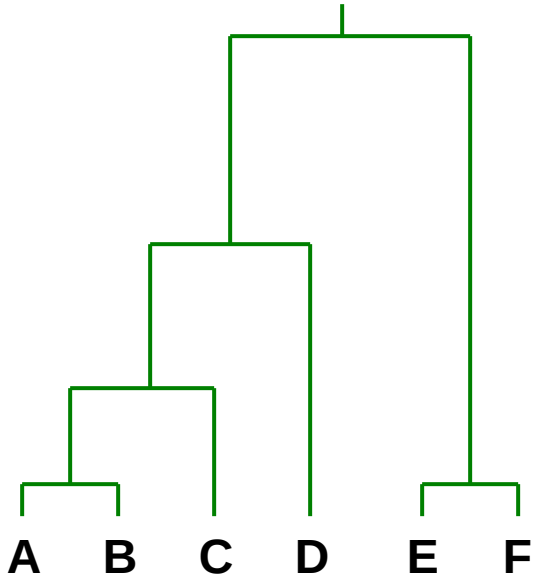
- Um dendograma é frequentemente usado para representar um agrupamento hierárquico.



Clustering Hierárquico: Exemplo

Este exemplo ilustra single-link clustering no espaço Euclideano com 6 pontos





Soluções (Há muitas)

- Pacotes em R, MatLab, Python,...
- Ferramentas livres como Weka, RapidMiner,...
- Ferramentas proprietárias como Clementine, Microsoft Power BI, Tableau,...

Pra fechar...

- Objetivei dar uma visão geral (introdutória) de Análises de dados
- Um campo aberto para pesquisas, estudos de casos e desenvolvimento de soluções
- Não há como falar em Sistemas de Informação para tomada de decisões sem falar de Análise de Dados (diferencial)

Referências

- Introdução a Banco de Dados (Apostila, Cap. 10). Prof. João Eduardo Ferreira (IME/USP)
- Notas de aula da Prof. Maria Luiza M.Campos (DCC/IM/UFRJ)
- Notas de aula do Prof. Edgard Jamhour (PPGIA/PUCPR)
- Eric Thomsen. OLAP – Construindo Sistemas de Informações Mutidimensionais. Editora Campus. Rio de Janeiro, 2002.
- Ralph Kimball. Data Warehouse Toolkit. Editora Makron Books. São Paulo, 1998.
- Laudon & Laudon. Gerenciamento de Sistemas de Informação. 3ª Edição. Editora LTC. Rio de Janeiro, 2001.
- Sistemas de Banco de Dados. (Cap. 28) Ramez Elmarsri e Sham Navathe. 4ª Edição. Ed. Pearson, 2005.
- REMENYL, D. MONEY, A. SHERWOOD-SMITH, M. The effective measurement and management of IT costs and benefits. Oxford: Butterworth-Heinemann, 2000.

Referências

- Tecnologia da Informação para Gestão. Turban, E; McLean, E.; Wetherbe, J.; Bookman; 2002.

- Corporate Information Systems Management : Text and Cases. Fourth Edition By Applegage, Lynda M. / McFarlan, F. Warren / McKenney, James L.; 1996.

- Information Systems Management in Practice. Fourth Edition; Sprague, Ralph H. / McNurlin, Barbara; Prentice-Hall; 1999.

- Management Information Systems: Organization and Technology in the Networked Enterprise. Laudon, K., Laudon, J.; Prentice Hall; 2000.

- Enterprise Architecture Planning: Developing a Blueprint for Data, Applications and Technology By Spewak, Steven; 1993.

- Software Assessment, Benchmarks, and Best Practices;. Capers Jones; Addison Wesley; 2000.

- Redes de Valor. David Bovet; Joseph Martha; Negócios Editora; 2000.

- Big Data. Eje estratégico en la industria audiovisual. Eva Patricia Fernández Manzano (Organizador). Editorial UDC; 2016.

Referências

- Mariano, V. L.; Boscarioli, C. Utilização de BI na Análise de Desempenho em uma Empresa do Agronegócio. In: Anais do SBSI 2012, São Paulo, SP, 2012.

- Zanardi, F.; Oyamada, M. S.; Boscarioli, C. Geração de Painéis Gerenciais a partir de Data Marts: Um Relato de Experiência . In: Anais do CONTECSI 2012, São Paulo, SP, 2012.

- Voltolini, R.; Boscarioli, C. Projeto, Implantação e Avaliação de uma Solução OLAP para Empresas de Venda de Telefonia Móvel Empresarial. Artigo de Conclusão de Curso. Pós-graduação em Tecnologias de Business Intelligence, Unioeste, Campus de Cascavel, 2011.

- HEHN, Herman F. e SILVA, Eloah C. A. M. P. Managerware: como extrair valor dos investimentos em Sistemas de Informação. São Paulo: Atlas, 2006.

- LAGO, S. M. S. Notas de aula. Unioeste, campus de Cascavel, 2011.

- LAUDON, Kenneth C. e LAUDON, Jane P. Sistemas de informação gerenciais. 9ª edição, São Paulo: Pearson, 2011.

- Gartner Group. 2014 - Magic Quadrant for Business Intelligence and Analytics Platforms. https://www.gartner.com/doc/2668318

Venha pesquisar comigo...

unioeste
Cascavel

PPGECEM

Programa de Pós-graduação em Educação em Ciências e Educação Matemática

Mestrado e Doutorado

www.unioeste.br/ppgecem

PPGComp

Programa de Pós-graduação em Ciência da Computação

Mestrado

www.inf.unioeste.br/pos

Na continuação...

17/06 - Palestra 2: **Modelos Preditivos de Vendas para Recomendação de Estoque**

Rodrigo Pereira Fontes

18/06 - Palestra 3: **Business *Intelligence and Analytics* na Gestão da Construção Civil**

Anderson Brunheira Lopes

Obrigado!

Clodis Boscarioli

boscarioli@gmail.com clodisboscarioli