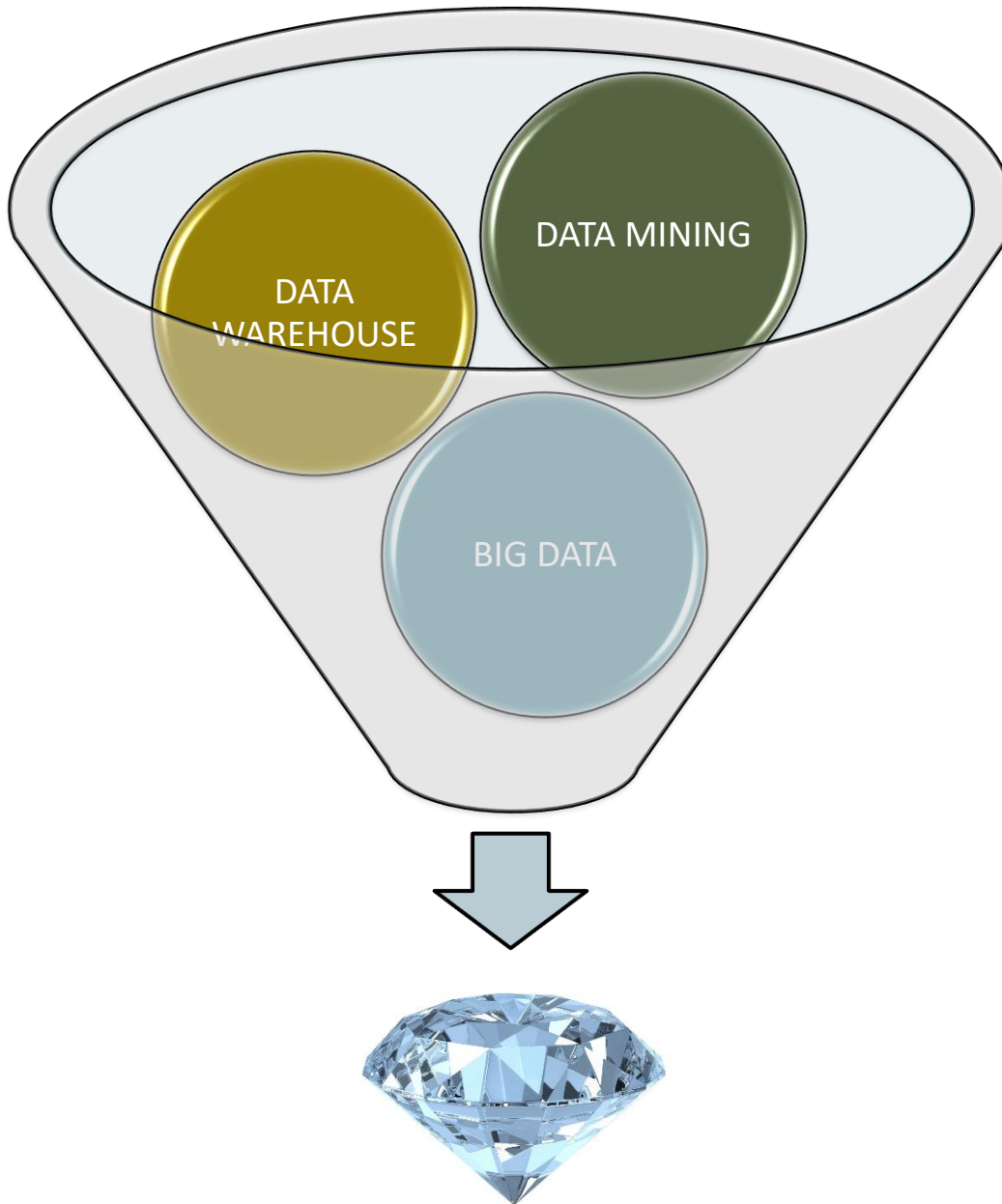




# Data Warehouse Mineração de Dados

Profa. Roberta Macêdo M. Gouveia  
robertammg@gmail.com

05/11/2015



A mina de ouro debaixo dos bits.

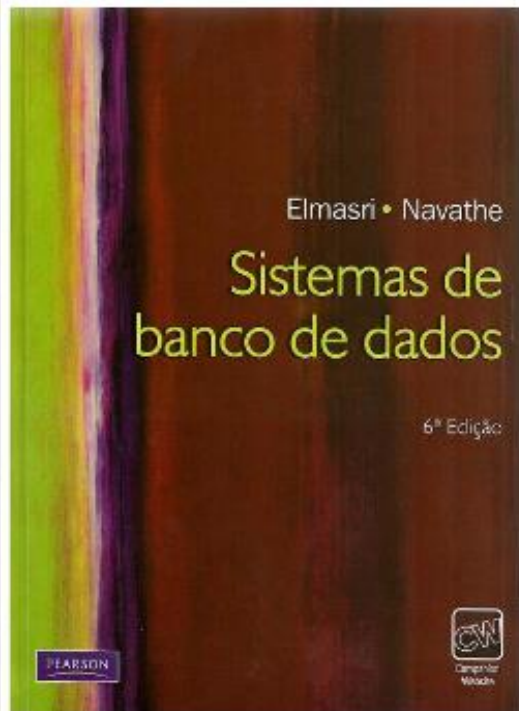
# Data Warehouse: A Memória da Empresa



# Data Mining: A Inteligência da Empresa

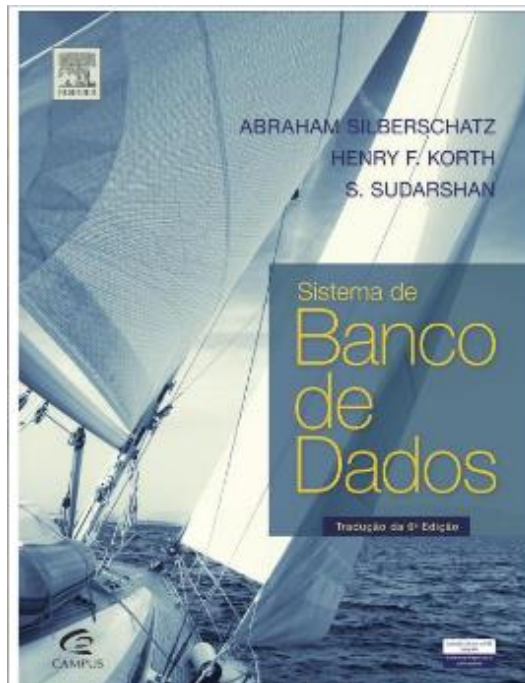


# Leituras Iniciais



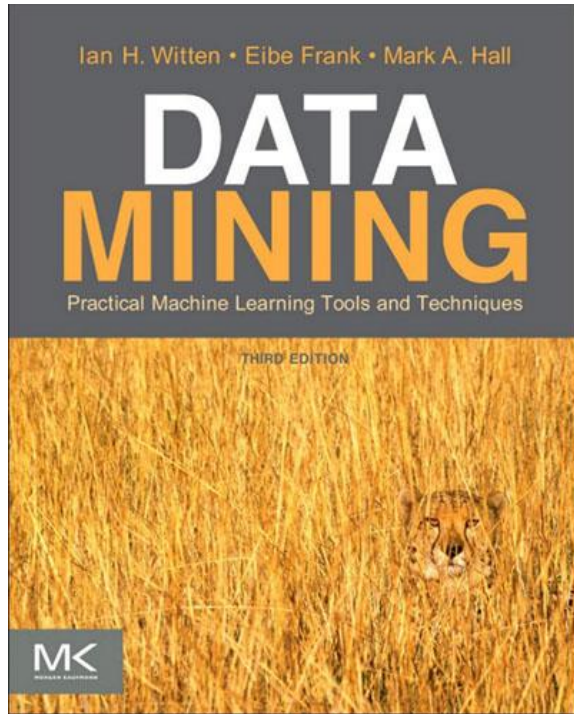
Cap. 28 - Conceitos de Mineração de Dados

Cap. 29 - Visão Geral de Data Warehousing e OLAP

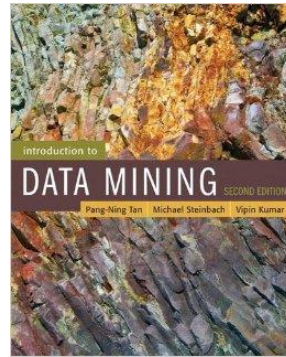


Cap. 20  
Depósito e  
Mineração de Dados

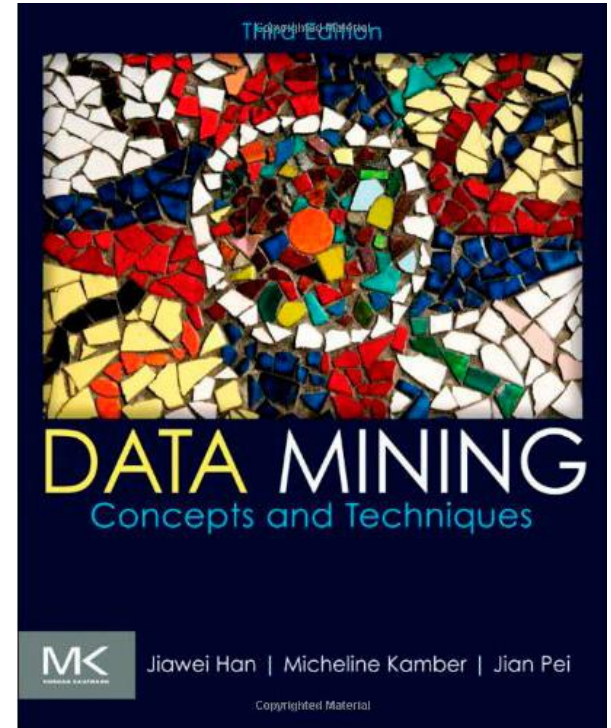
# Bibliografias Específicas



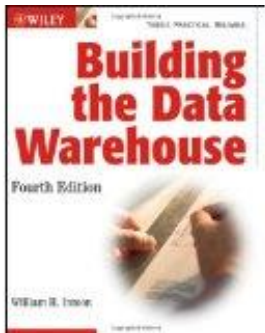
Ian H. Witten



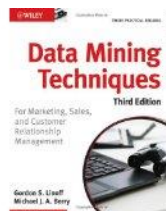
Pang-Ning Tan



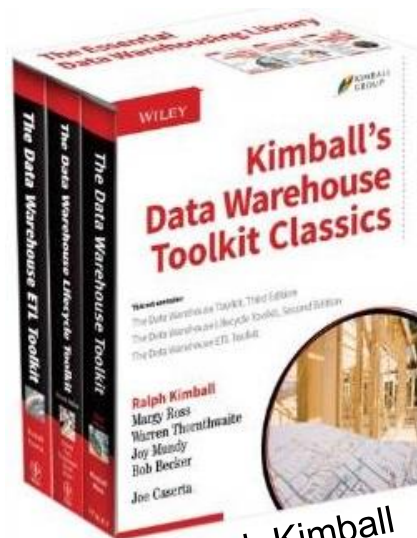
Jiawei Han; Micheline Kamber; Jian Pei



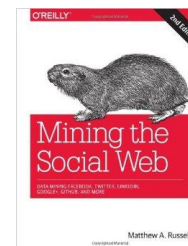
William H. Inmon



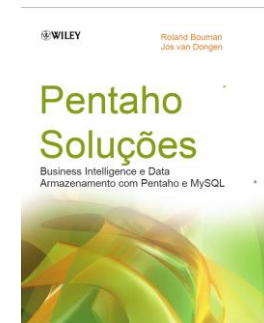
Gordon S. Linoff  
Michael J. A. Berry



Ralph Kimball



Matthew A. Russell

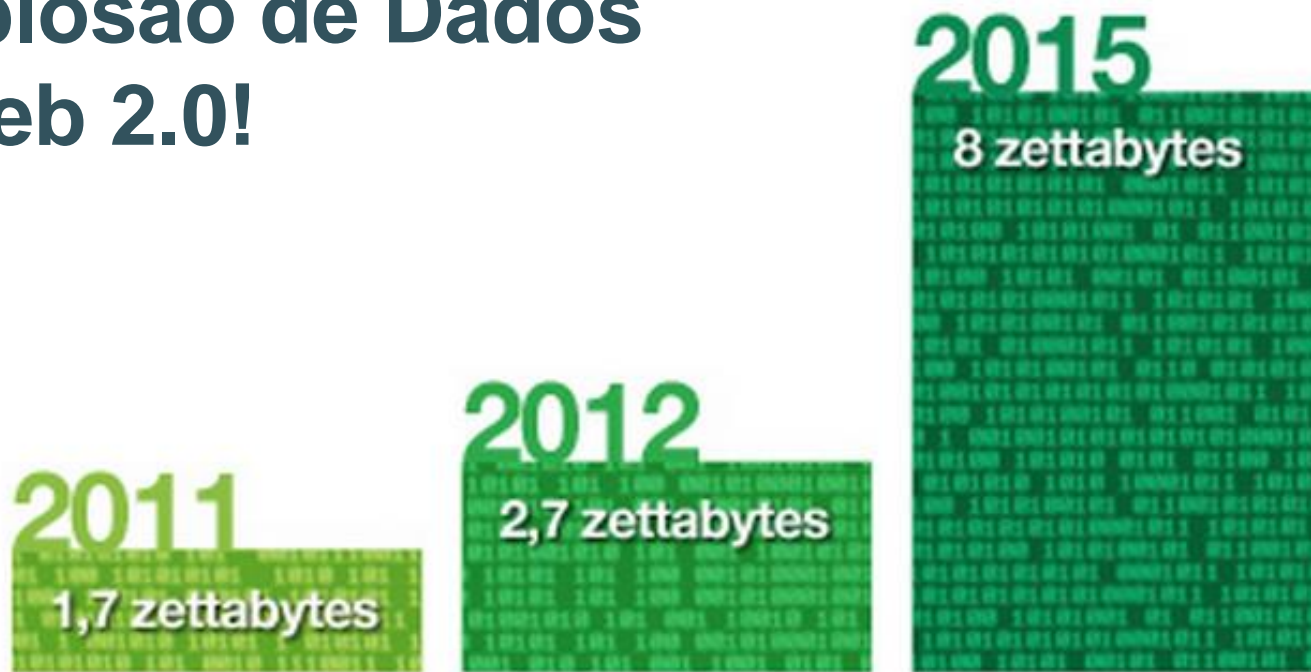


Roland Bouman



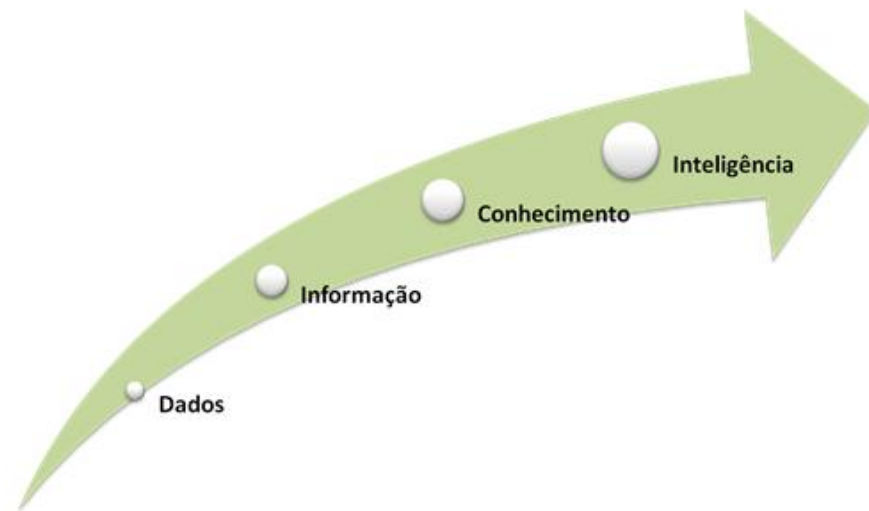
Viktor Mayer-schonberger

# A explosão de Dados na Web 2.0!



Fonte: IBM - Volume de Informação Digital - *International Data Corporation (IDC)*  
[http://www.ibm.com/midmarket/br/pt/infografico\\_bigdata.html](http://www.ibm.com/midmarket/br/pt/infografico_bigdata.html)

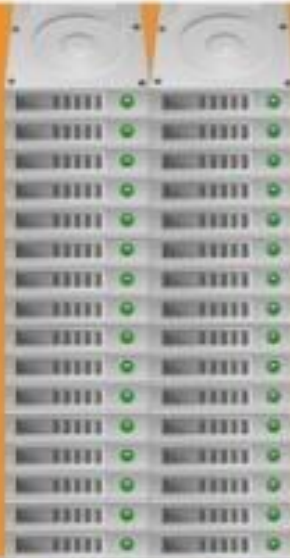
1 kilobyte	1,000,000,000,000,000,000,000
1 megabyte	1,000,000,000,000,000,000,000,000
1 gigabyte	1,000,000,000,000,000,000,000,000,000
1 terabyte	1,000,000,000,000,000,000,000,000,000,000
1 petabyte	1,000,000,000,000,000,000,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000,000,000,000,000,000,000



Atualmente, cerca de 15 petabytes de dados estruturados e não estruturados são gerados todos os dias. Entre eles, destacam-se vídeos, comentários em redes sociais, conteúdos de blogs e dispositivos móveis

**1ZB =**

1 bilhão de HDs  
iguais ao de um desktop



**1ZB =**

75 bilhões  
de iPads 16GB



**0,5ZB =**  
toda a Internet em 2009



**42ZB =**

Todas as palavras ditas pela  
humanidade, em toda a sua história,  
se digitalizadas.



Fonte: <http://www.monetate.com/infographic/the-retailers-guide-to-big-data/#axzz2HaZVK816>

## DESAFIO:





# Descoberta de Conhecimento em Bancos de Dados

## **Processo KDD - *Knowledge Discovery in Databases***

“É o processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis, de uma fonte de dados.”

(Usama Fayyad et al. 1996)

## **Data Warehouse**

*“A copy of transaction data specifically structured for query and analysis.”*

(Ralph Kimball, 2013)

## **Data Mining**

*“DM is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic.”*

(Ian H. Witten et al. 2011)

# DM e DW fazem parte do processo de descoberta de conhecimento em Bancos de Dados (KDD)

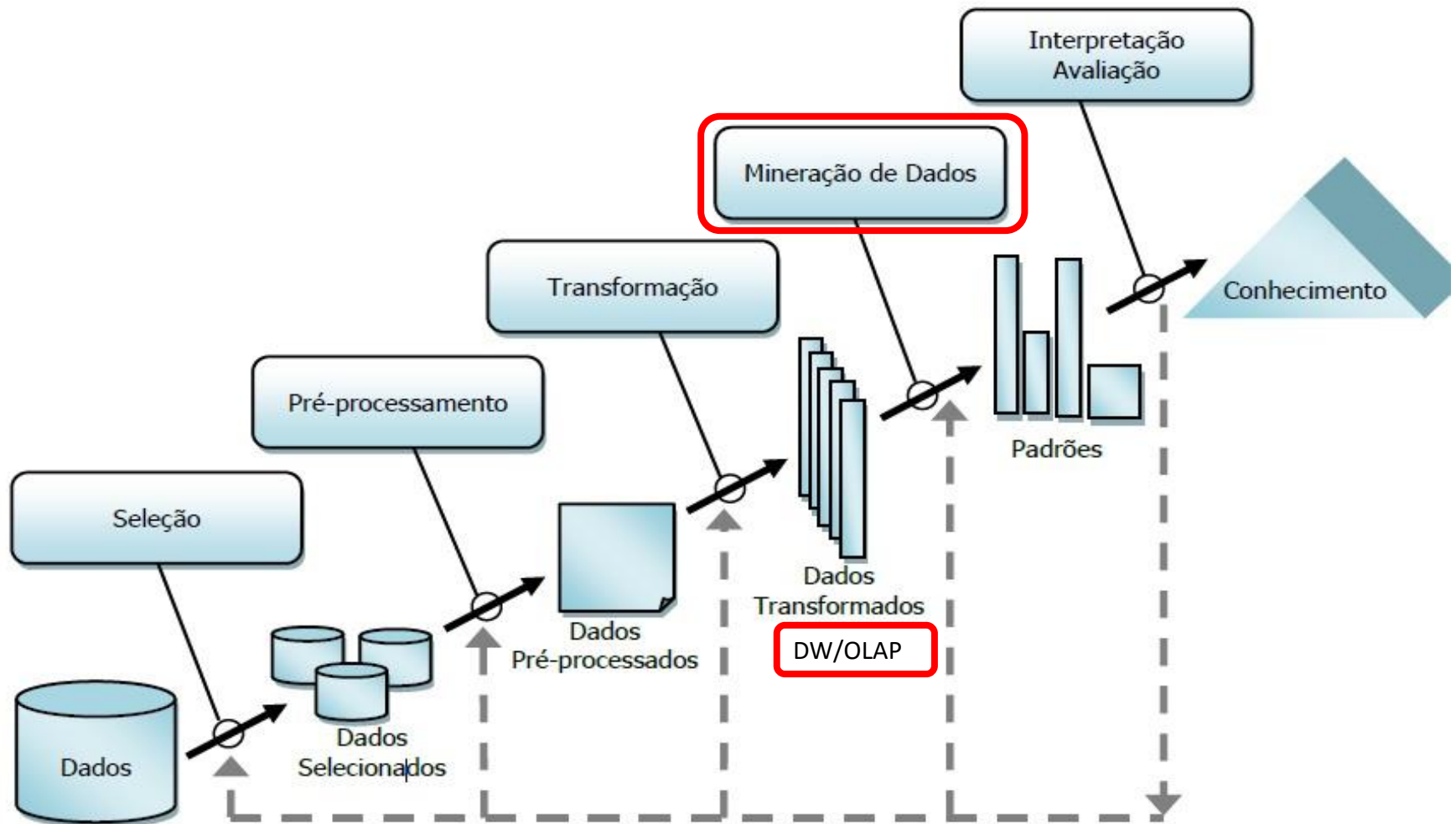
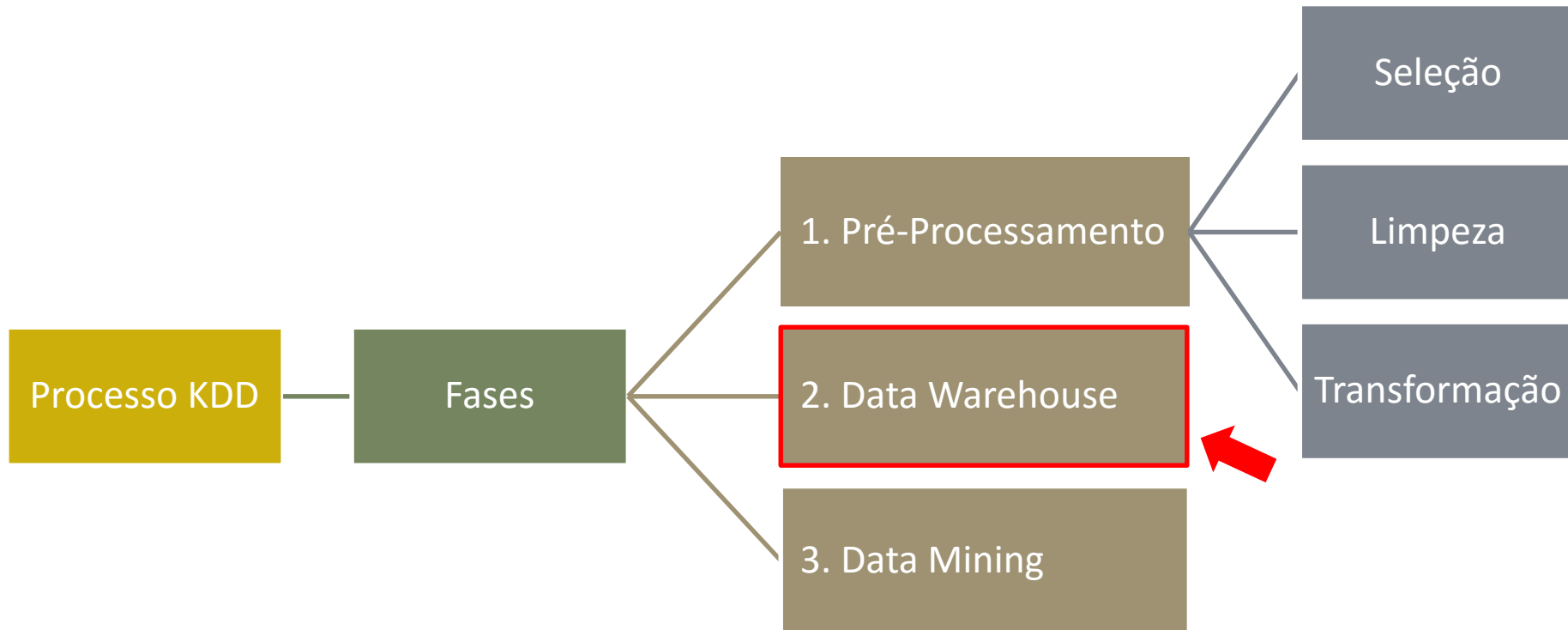
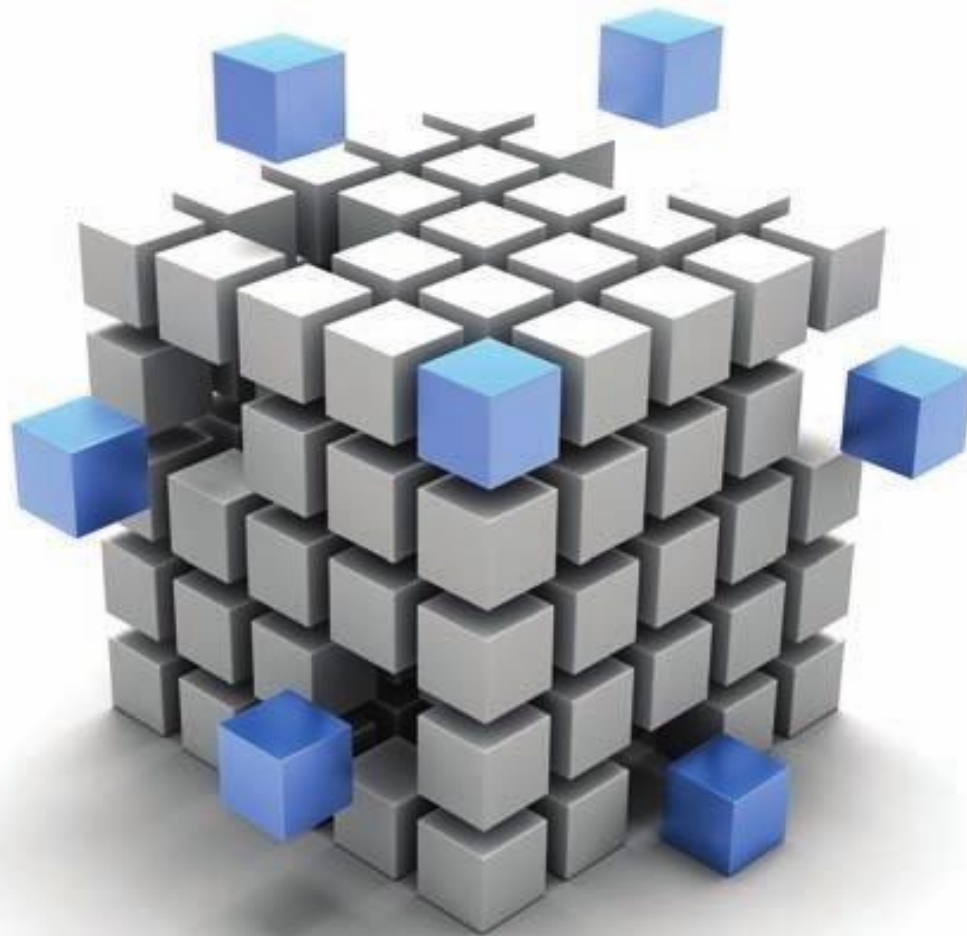


Figura - Knowledge Discovery in Databases – KDD

# Knowledge Discovery in Databases – KDD

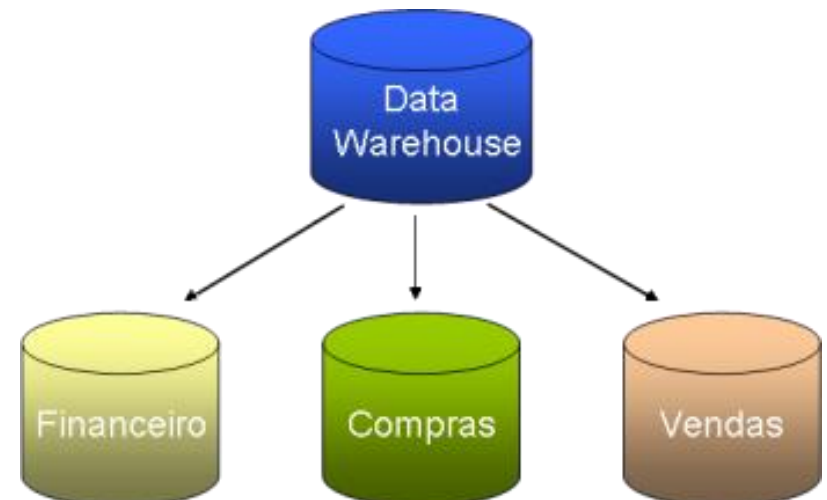
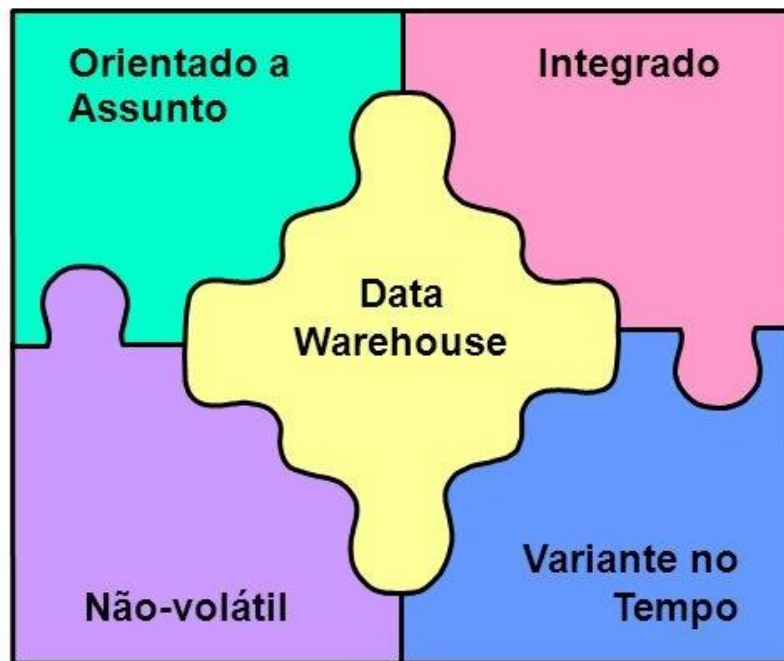


# Data Warehouse - DW



“Coleção de dados **ORIENTADOS A ASSUNTO**, **INTEGRADOS**, **NÃO-VOLÁTEIS** e **VARIANTES NO TEMPO**, utilizada para tomada de decisões.”

William H. Inmon (2005)



# Motivação para implementação de um DW

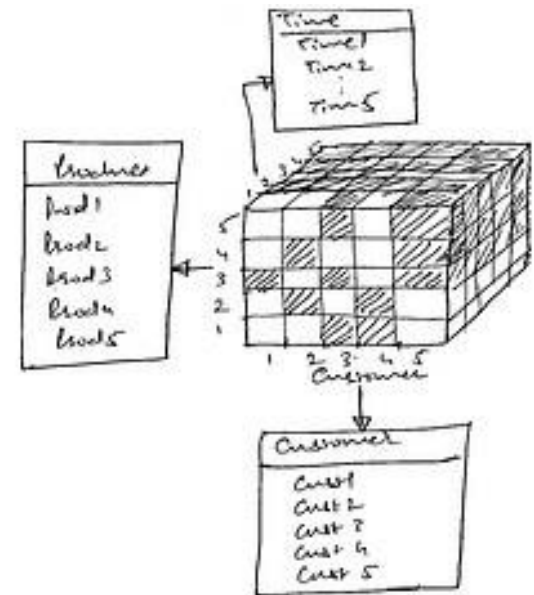
- Não causar impacto sobre o BD operacional (ambiente OLTP).
  - *OLTP - Online Transaction Processing.*
- São otimizados para aplicações analíticas.
  - *OLAP - Online Analytical Processing*



- Fornecer uma origem de dados única (centralizada) e consistente para fins de Apoio à Decisão.

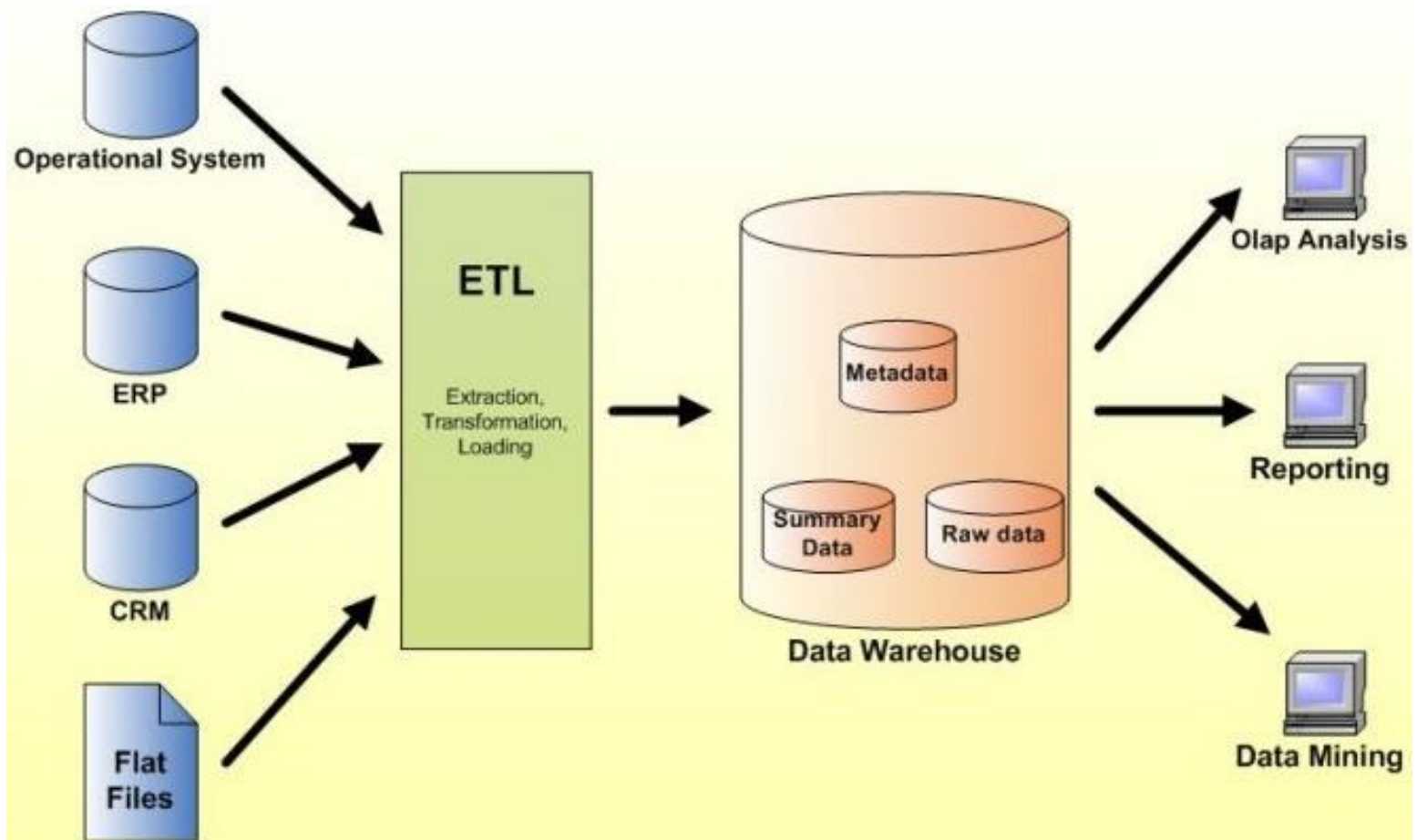
# Funcionalidades de um DW

- Eficiência no processamento de consultas, pois são livres das restrições dos ambientes transacional – Propriedades ACID.
- **São otimizados para aplicações: DSS e *Data Mining*.**
- Suporte a Modelos Multidimensionais.
  - Melhor desempenho para consultas complexas, intensivas e *ad hoc*.



# Processo ETL

- ETL: Extract → Transform → Load
  - Antes dos dados serem armazenados no DW eles passam por um processo de extração, tradução, filtragem e integração.





# BD Operacional/Transacional *versus* Data Warehouse

	BD Operacional	Data Warehouse
Usuários	Funcionários	Gerência
Utilização	Tarefas cotidianas	Decisões estratégicas
Princípios de Funcionamento	Com base em transações	Com base em análise de dados
Estruturado para:	Dados normalizados Integridade dos Dados	Dados não normalizados Facilidade de Consulta
Frequência de atualização	Em grande número	Quase inexistentes, apenas novas inserções
Padrão de Uso	Previsível	Difícil de prever
Valores dos Dados	Valores atuais e voláteis	Valores históricos e consolidados
Detalhamento	Alto	Sumarizado
Organização dos Dados	Orientado a aplicações	Orientado a assunto

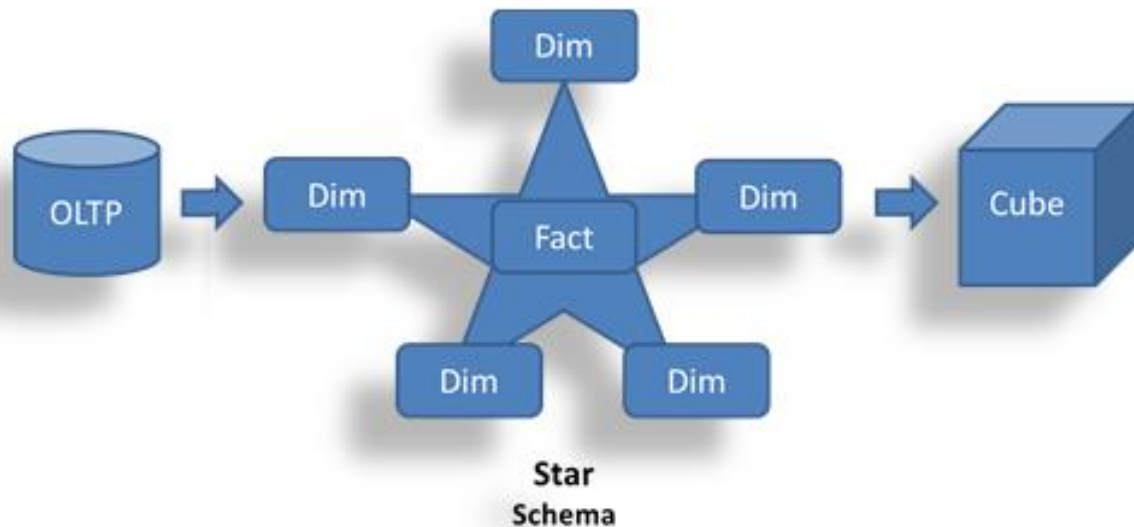


OLTP

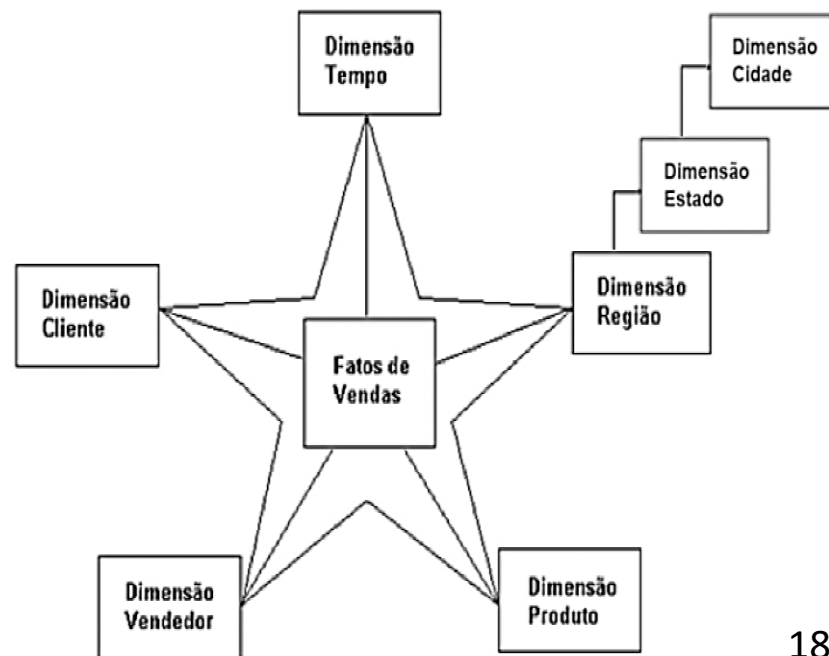
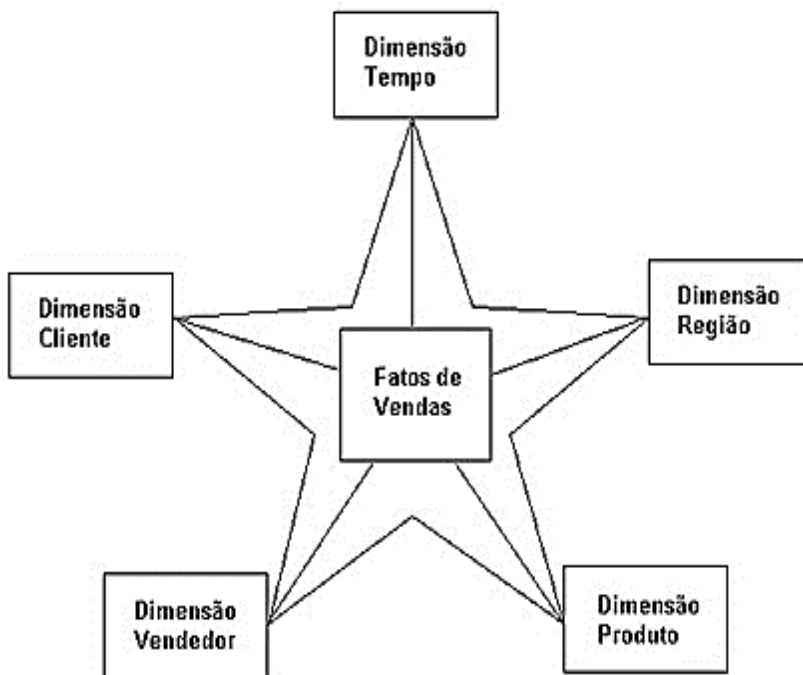


OLAP

# Modelagem Dimensional



- ✓ Esquema Estrela (*Star Schema*)
- ✓ Esquema Floco de Neve (*Snowflake Schema*)
- ✓ Esquema Constelação de Fatos (*Facts Constallation Schema*)



# Modelagem de dados para DW

- **Modelagem Dimensional**

- Tabelas de Fatos, Tabelas de Dimensões e Métricas.
- Facilita a processamento analítico e as consultas multidimensionais.

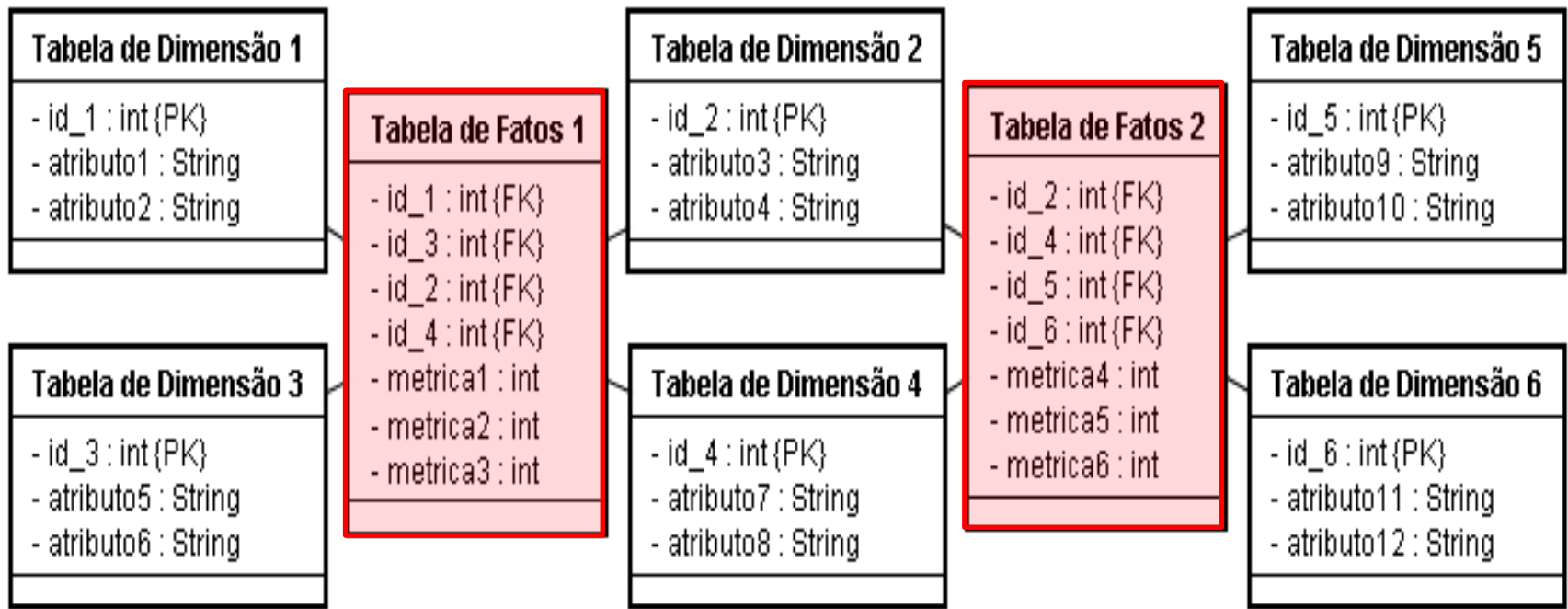
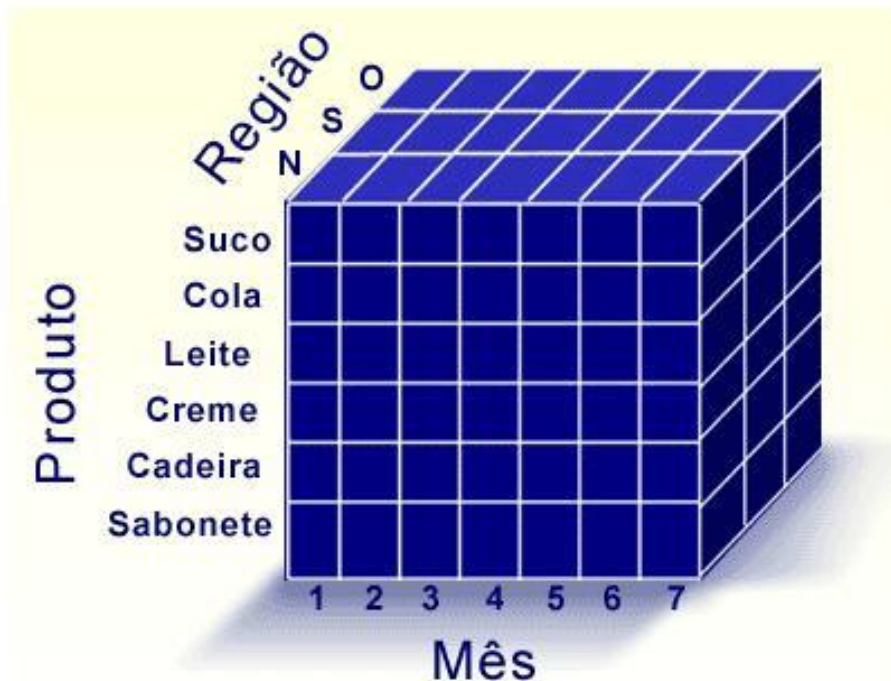


Figura - Exemplo de Modelagem Dimensional: Esquema Constelação de Fatos.

# Modelagem Dimensional

- Possibilita a **utilização de ferramentas OLAP**, cujas funções são:
  - Obter informações sumarizadas; mostrar os dados em tabelas n-dimensionais com suporte para modificações dos eixos (dimensões).
  - Favorecer análise e visualização de várias dimensões em uma única consulta.

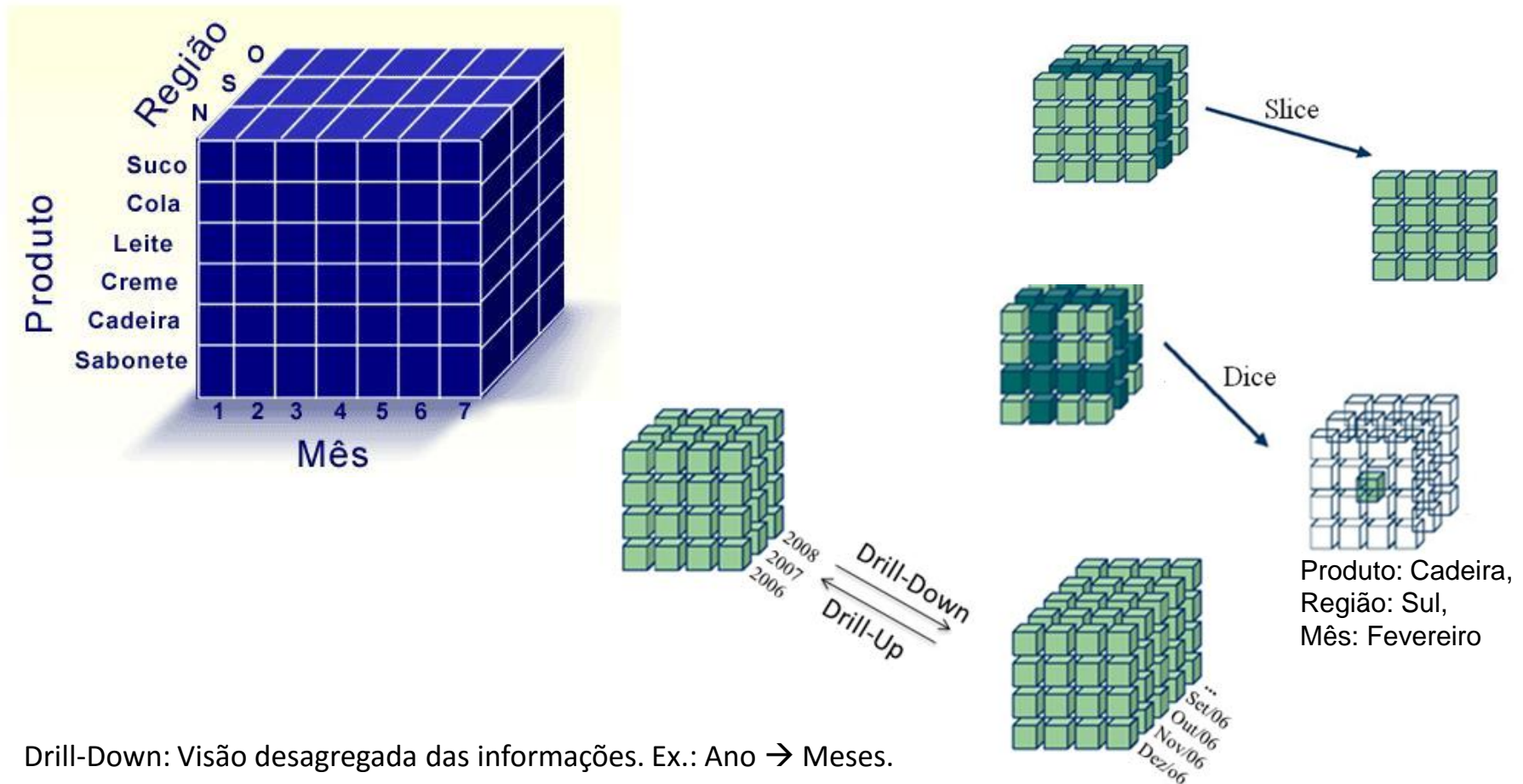


OLAP utiliza a estrutura multidimensional de **Cubo de Dados**.

As operações sobre os cubos proporcionam **múltiplas agregações**.

# Operações OLAP em Cubo de Dados

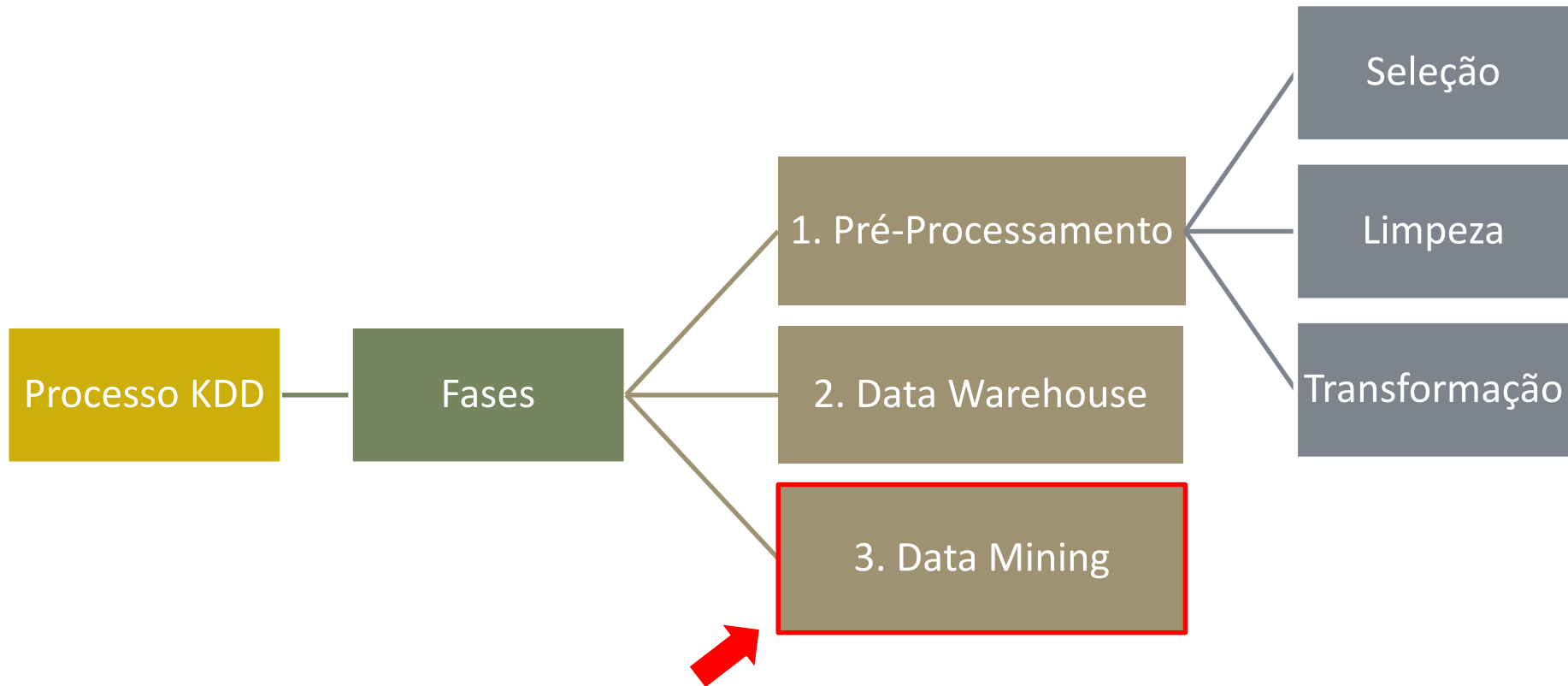
- A função **Slice** faz restrição de um valor ao longo de uma dimensão.
- A função **Dice** faz restrições de valores em várias dimensões.



Drill-Down: Visão desagregada das informações. Ex.: Ano → Meses.

Drill-up: Visão agregada das informações. Ex.: Meses → Ano.

# Knowledge Discovery in Databases – KDD



# Multidisciplinaridade



- ✓ Quer conhecer melhor os clientes?
- ✓ Deseja encontrarem tendências úteis, tais como o comportamento dos consumidores?
- ✓ Pretende agregar valor (\$) com as técnicas de análise de dados?
- ✓ Almeja tornar o marketing mais eficiente?
- ✓ Pretende fazer sua empresa prosperar?

## Então, prepare-se para Minerar seus dados!

Detectar regras, hábitos e padrões de comportamento.





# Áreas de Aplicações Potenciais





O que **cerveja** tem a ver com **fraldas**?

- Suposições:
  - Tem o mesmo número de letras?
  - Cerveja no presente, fraldas no futuro?
  - Aumenta o consumo de fraldas, diminui o consumo de cerveja? (...)

# Exemplo 1 (clássico)



- Constatou-se que muitos homens casados, entre 25 e 35 anos, compravam fraldas e cervejas às sextas-feiras à tarde/noite (no caminho do trabalho para casa).
- Walmart otimizou as prateleiras nos pontos de vendas, colocando as fraldas ao lado das cervejas.

- Resultado: o consumo cresceu ainda mais.



## Exemplo 2



Target, uma grande rede de varejo dos EUA, descobre gravidez de adolescente antes dos pais!



# Exemplo 2

**Forbes** - **New Posts** +41 posts this hour | **Most Popular** Highest-Paid Models | **Lists** Top-Earning Tennis Stars | **Video** Veg...

TECH | 2/16/2012 @ 11:02AM | 2,149,798 views

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

38.5k

Share

15.4k

Tweet

5.8k

Share

2.4k

reddit

4.7k

+1

363

Submit

322 comments, 169 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. **Target**, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the [New York Times](#) how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the [NYT](#):

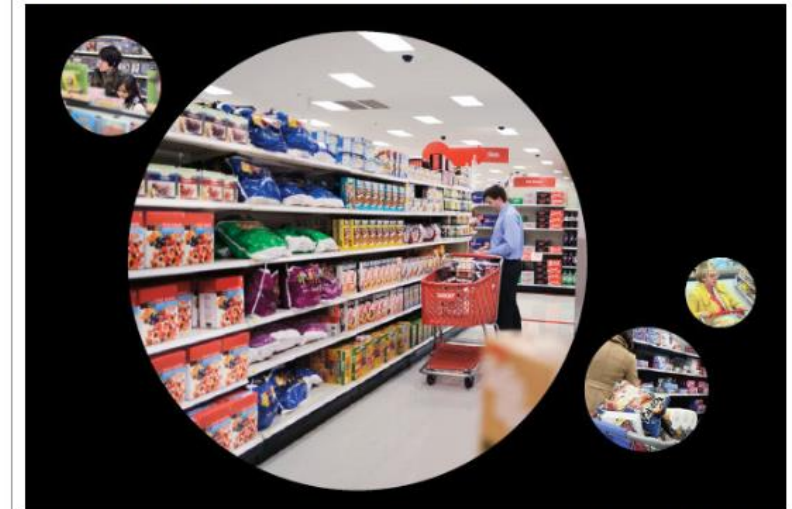


Target has got you in its aim

<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>



## How Companies Learn Your Secrets



Antonio Ballo/Reportage for The New York Times

By CHARLES DUHIGG  
Published: February 16, 2012

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?"

FACEBOOK

TWITTER

GOOGLE+

SAVE

EMAIL

<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&module=Search&mabReward=relbias%3As&r=0>

# Exemplo 3 - Banco Itaú



- Enviava mais de 1 milhão de malas diretas, para todos os correntistas.
  - No máximo 2% deles respondiam às promoções.
- Hoje, com a mineração dos dados, as cartas são enviadas apenas a quem tem maior chance de responder.
  - A taxa de retorno subiu para 30%.
  - A conta do correio foi reduzida a 1/5.

# Exemplo 4 - SERPRO



- Investiu milhões no seu projeto de DW e DM, desenvolvido em parceria com a Oracle.
- Consolidou apenas 5% de suas informações, mas atualmente já é possível fazer em 5 minutos cruzamentos de dados que antes demandavam dias de trabalho.

# Visão Geral

- DM refere-se à descoberta de novas informações em função de **padrões** ou **regras** em (grandes) bases de dados.
- Metas que podem ser alcançadas pela DM:

## Previsão

- Antecipar os valores de variáveis desconhecidas. Indica as chances de uma ação ocorrer.

## Descrição

- Procurar por padrões que descrevem os dados e que sejam de entendimento dos usuários.



# Tarefas da Mineração de Dados

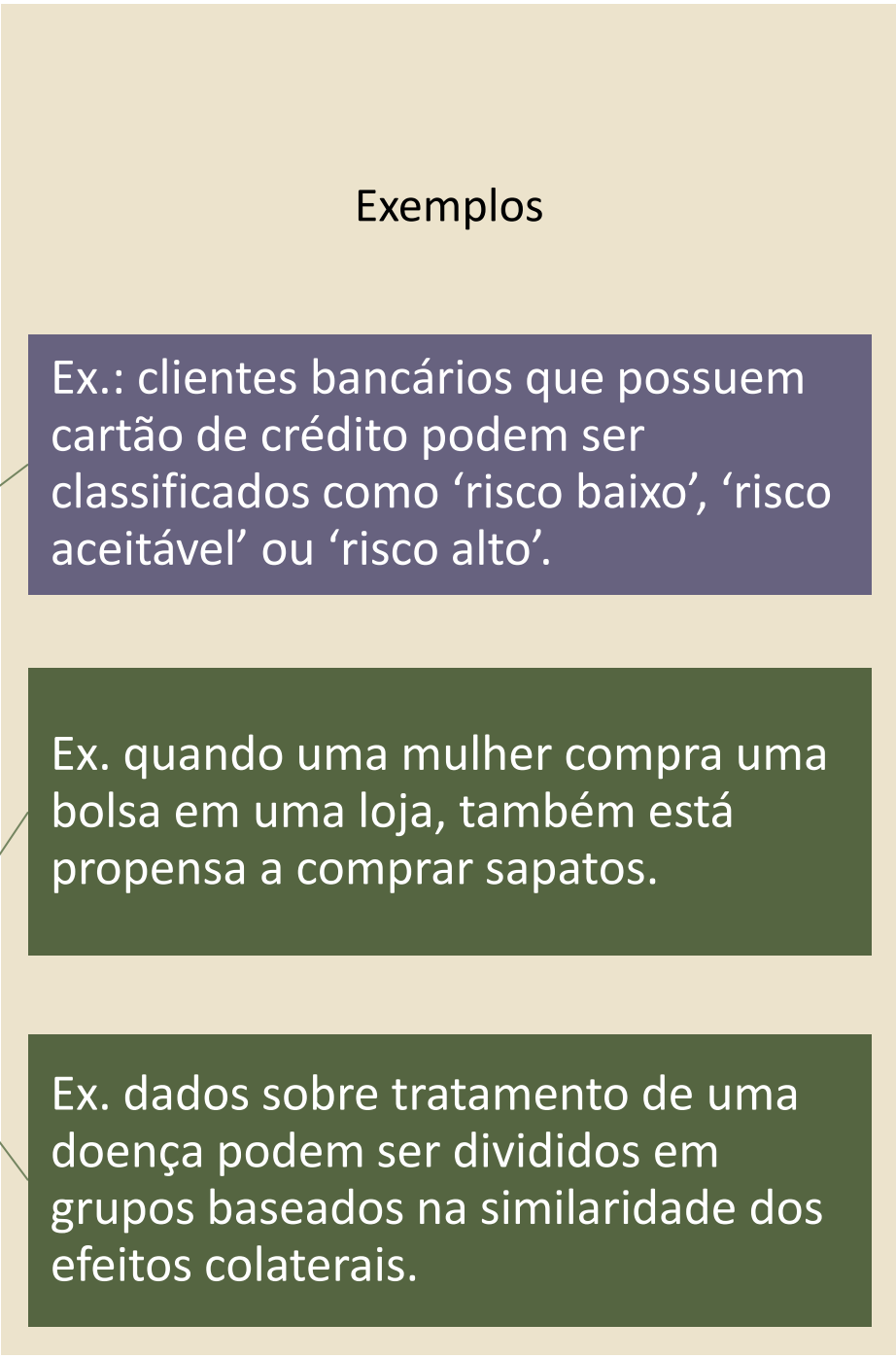
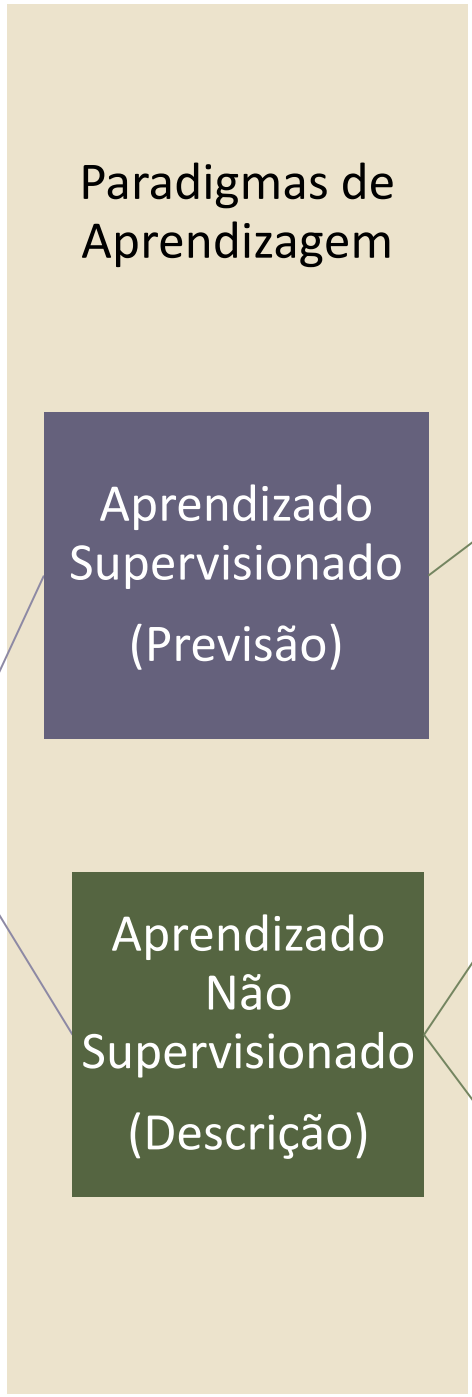
## Previsão

- Classificação
- Regressão

## Descrição

- Associação
- Agrupamento/Clustering
- Sumarização

# Mineração de Dados

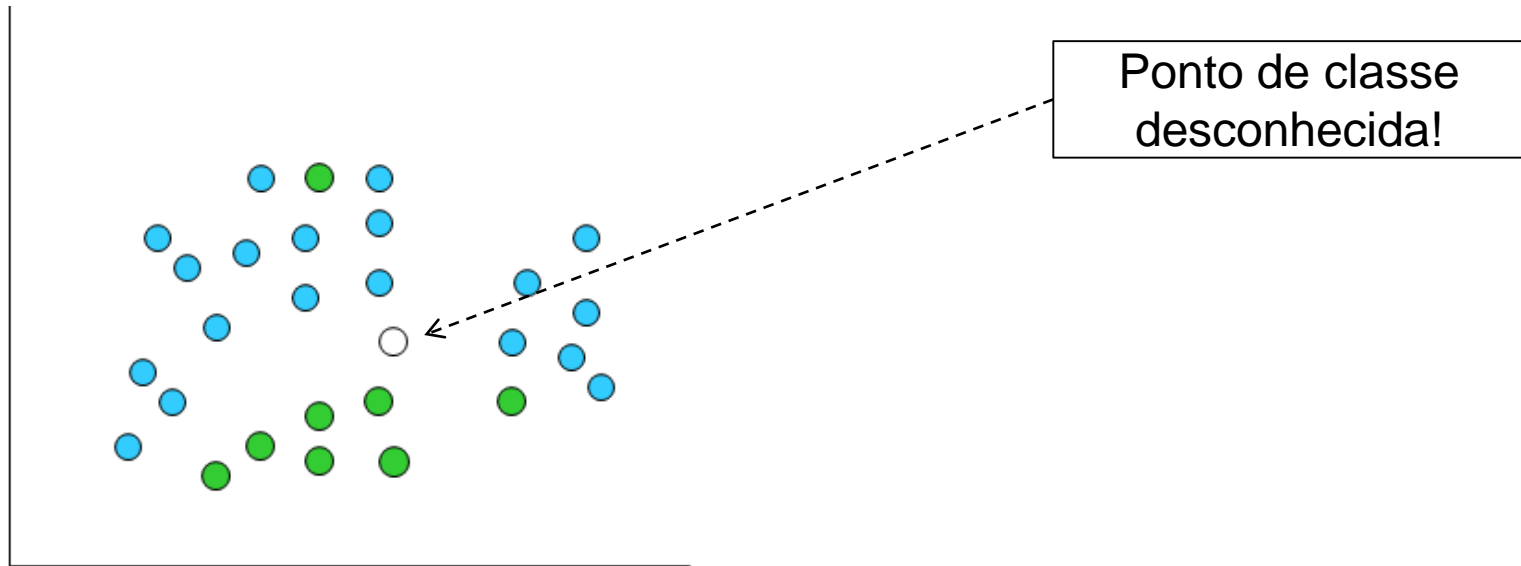


# Exemplos de Previsão

Encontrar um método para prever a classe de uma nova instância a partir de instâncias pré-classificadas.

Ex. Dado um conjunto de pontos das classes Verde e Azul.

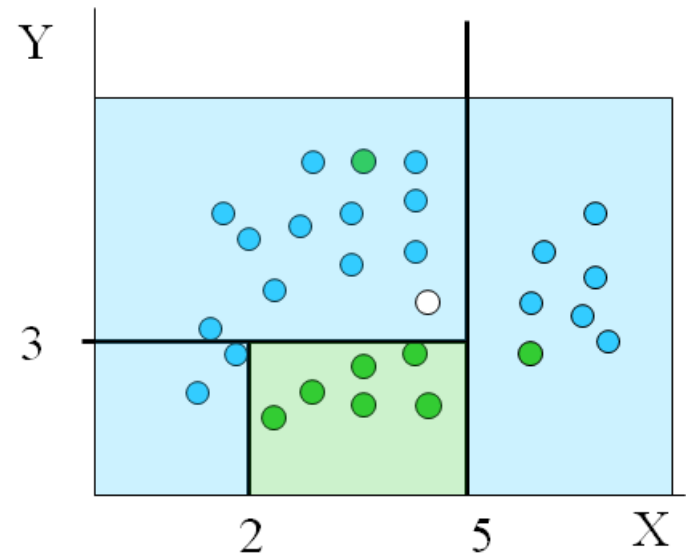
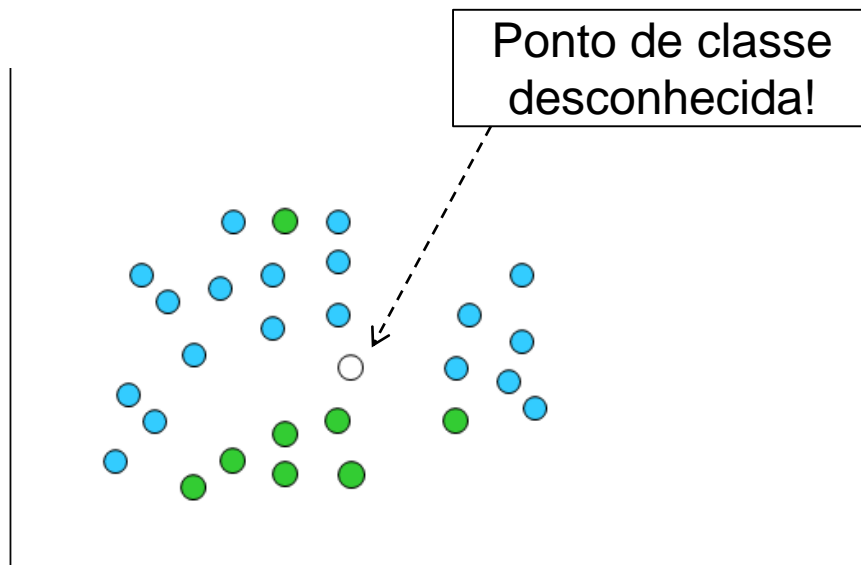
Qual é a classe para o novo ponto desconhecido? **Verde** ou **Azul**?



# Exemplos de Previsão

Ex. Dado um conjunto de pontos das classes Verde e Azul.

Qual é a classe para o novo ponto desconhecido? Verde ou Azul?



**Árvore de Decisão**

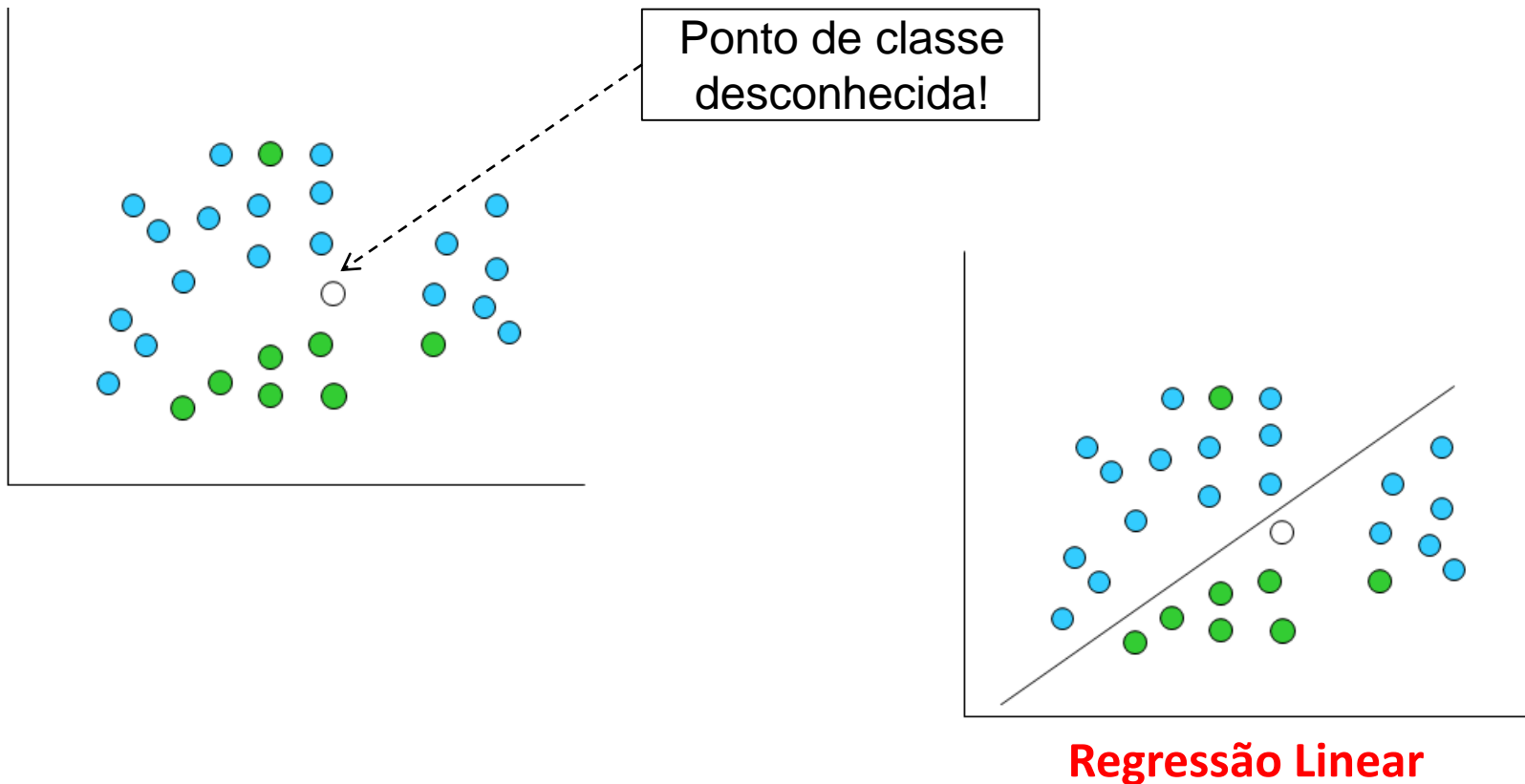
```
IF X > 5
  THEN AZUL
  ELSE IF Y > 3
    THEN Azul
    ELSE IF X > 2
      THEN VERDE
      ELSE AZUL
```

Novas instâncias são classificadas seguindo o caminho que leva da raiz até a folha.

# Exemplos de Previsão

Ex. Dado um conjunto de pontos das classes Verde e Azul.

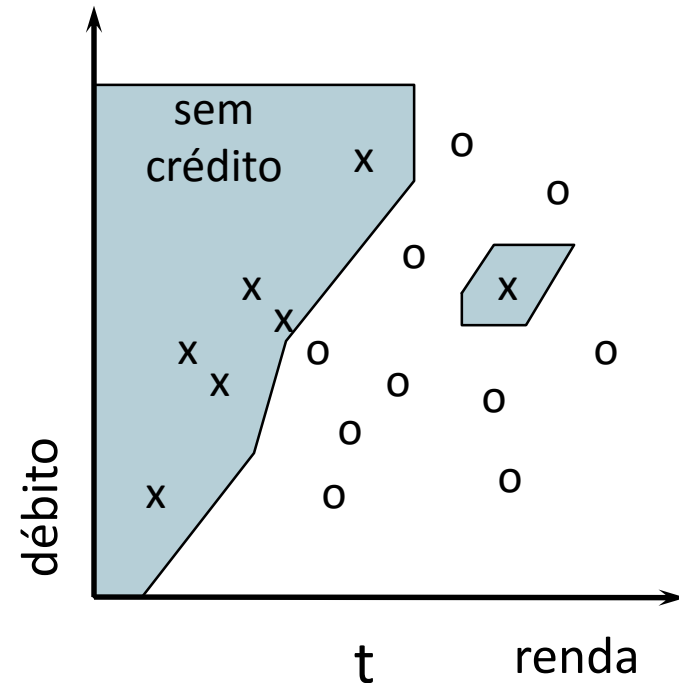
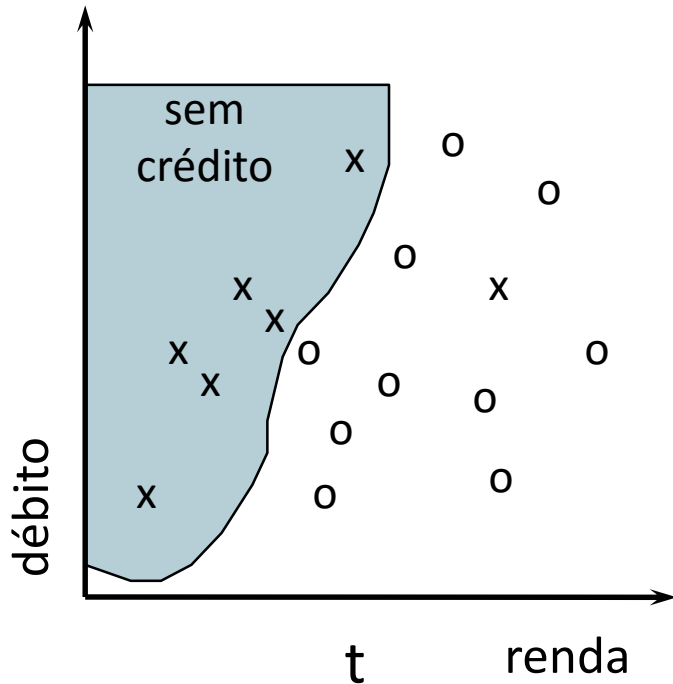
Qual é a classe para o novo ponto desconhecido? Verde ou Azul?



# Exemplos de Previsão

## Análise de Crédito

x: crédito recusado  
o: crédito aceito



Superfície não linear: melhora o poder de classificação.

Exemplo: **regressão não-linear**.

Métodos baseado em exemplos.

Exemplos: **k-vizinhos mais próximos**.

# Exemplos de Descrição

- Uma **Regra de Associação** é da forma  **$X \Rightarrow Y$** 
  - Onde  $X = \{x_1, x_2, \dots, x_n\}$  e  $Y = \{y_1, y_2, \dots, y_m\}$  são conjuntos de itens com  $x_i$  e  $y_j$  sendo distintos para todo  $i$  e todo  $j$ .
- Essa associação estabelece que, **se um cliente comprar X**, ele também estará **propenso a comprar Y**.
  - Ex. 98% dos consumidores que adquiriram pneus e acessórios de automóveis, também se interessaram por serviços automotivos (balanceamento, alinhamento).
- A regra de associação precisa satisfazer duas medidas de interesse:
  - Liminares mínimos de **SUPORTE** e **CONFIANÇA**.

# Exemplos de Descrição - Regras de Associação

- **Suporte** para uma regra  $X \Rightarrow Y$  refere-se a frequência com que ela acontece no BD.
- A **Confiança** da regra  $X \Rightarrow Y$  é calculada da seguinte forma:

$$\text{Confiança} = \frac{\text{Suporte } (X \cup Y)}{\text{Suporte } (X)}$$



# Exemplos de Descrição - Regras de Associação

- Exemplo: dados do carrinho de supermercado (itens que um consumidor comprou em um supermercado durante 4 visitas distintas (4 transações no BD))

<b>Id-Transação</b>	<b>Tempo</b>	<b>Itens-Comprados</b>
101	6:35	leite, pão, bolachas, suco
792	7:38	leite, suco
1130	8:05	leite, ovos
1735	8:40	pão, bolachas, café

# Exemplos de Descrição - Regras de Associação

- Considerando as regras:

1. Leite => Suco

2. Pão => Suco

Id-Transação	Tempo	Itens-Comprados
101	6:35	leite, pão, bolachas, suco
792	7:38	leite, suco
1130	8:05	leite, ovos
1735	8:40	pão, bolachas, café

- Suporte de {Leite, Suco} é 50%
  - Das 4 transações, a regra é satisfeita em duas delas
- Suporte de {Pão, Suco} é 25%
  - Das 4 transações, a regra é satisfeita em apenas uma delas

# Exemplos de Descrição - Regras de Associação

- Considerando as regras:

1. Leite => Suco

2. Pão => Suco

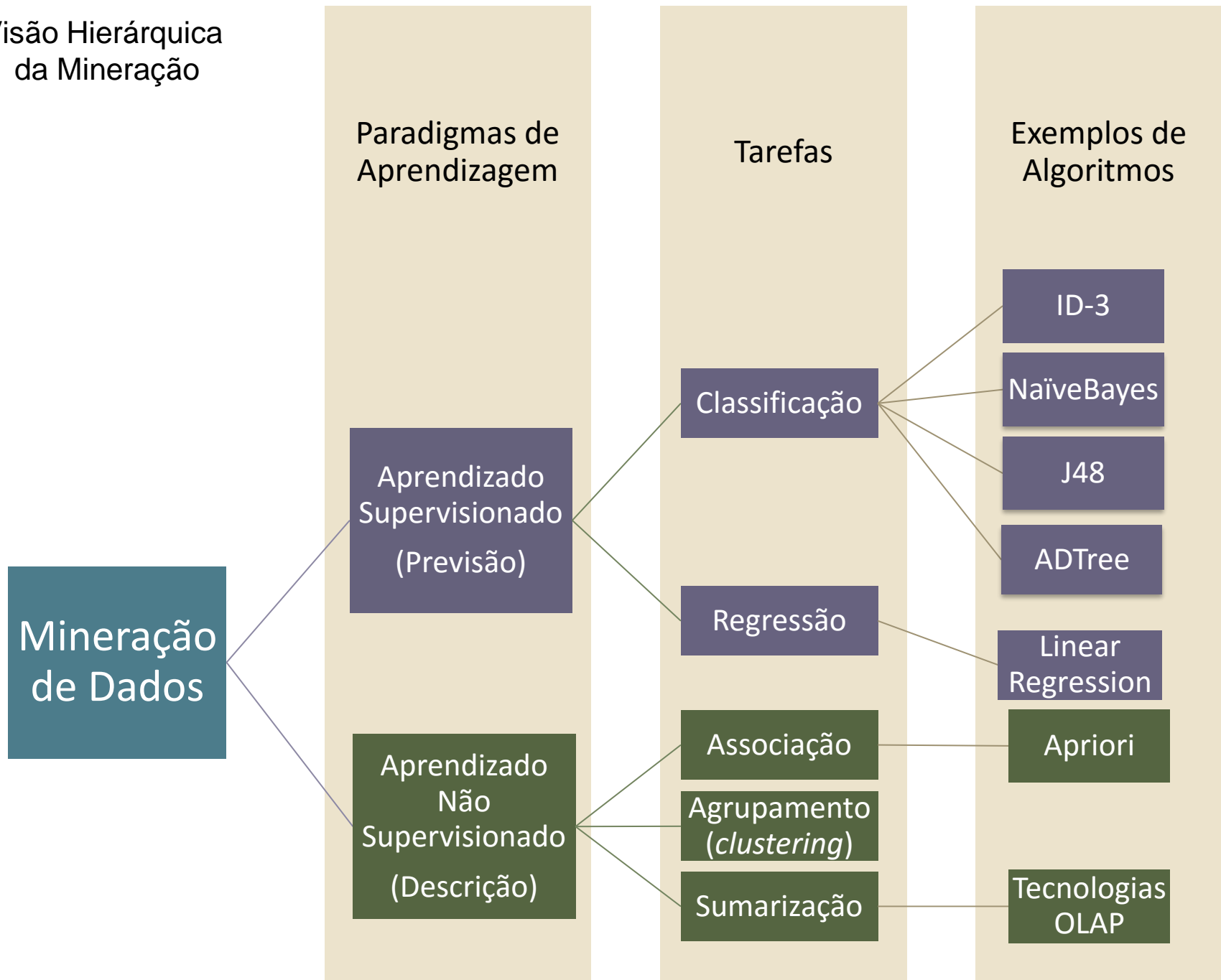
Id-Transação	Tempo	Itens-Comprados
101	6:35	leite, pão, bolachas, suco
792	7:38	leite, suco
1130	8:05	leite, ovos
1735	8:40	pão, bolachas, café

- Confiança de “Leite => Suco” é 66,7%
  - Das três transações nas quais Leite ocorre, duas contêm Suco
- Confiança de “Pão => Suco” é 50%
  - Das duas transações nas quais Pão ocorre, uma contêm Suco

# Exemplos de Descrição - Regras de Associação

- **CONCLUSÃO:** a MD gera todos os conjuntos de itens que estejam acima dos limites estabelecidos.
  - Ou seja, **suporte** e **confiança** precisam estar acima dos limites definidos.
- Exemplo: suponha que o limite definido para **Suporte** e **Confiança** da regra de associação seja  $\geq 50\%$ .
  - A regra **Pão**  $\Rightarrow$  **Suco** não é válida, pois o **Suporte** da regra de associação **Pão**  $\Rightarrow$  **Suco** foi apenas 25%.

# Visão Hierárquica da Mineração

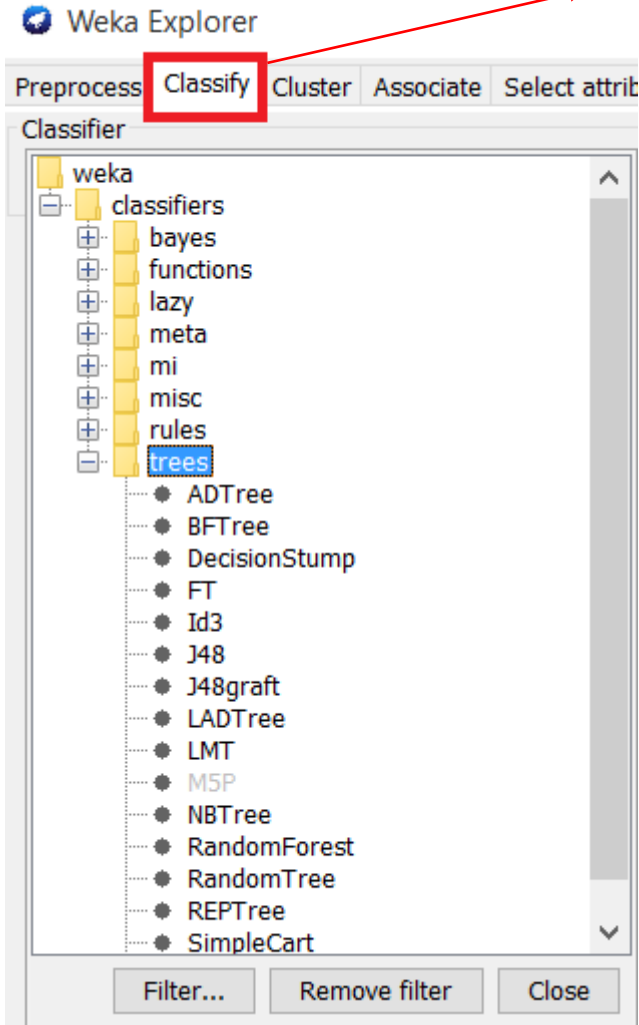


# Feramentas de Mineração de Dados

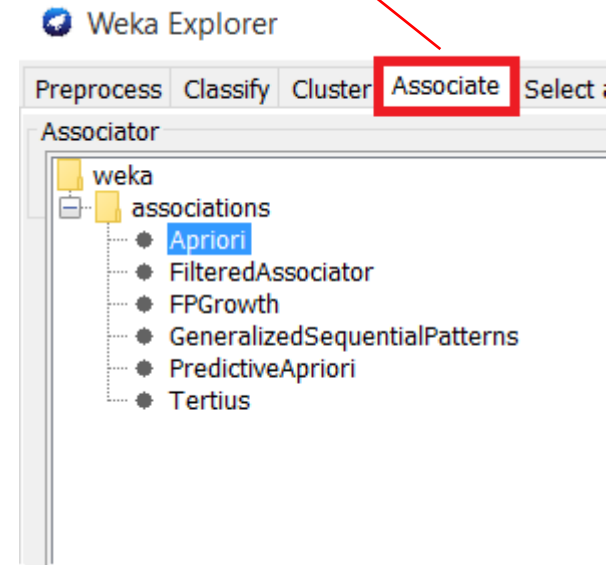
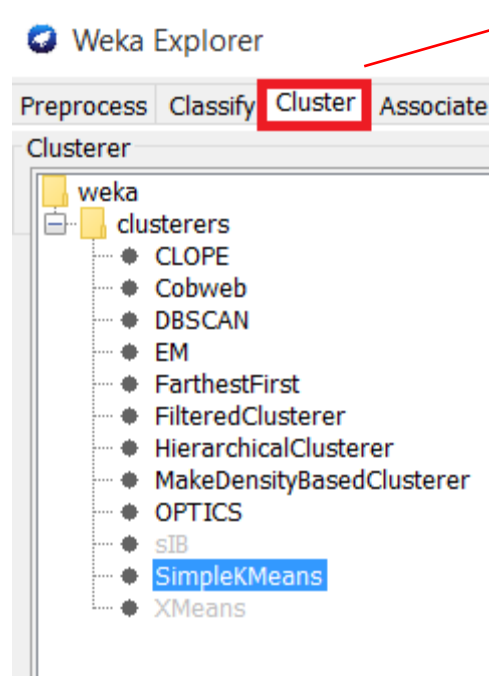
Ferramenta	Modelos Implementados	Fabricante
<b>Intelligent Miner</b>	Classificação, Regras de Associação, Clusterização e Sumarização.	IBM Corp. <a href="http://www.ibm.com">www.ibm.com</a>
<b>Weka</b>	Classificação, Regressão e Regras de Associação.	University of Waikato <a href="http://www.cs.waikato.ac.nz">www.cs.waikato.ac.nz</a>
<b>Oracle Data Mining</b>	Classificação, Regressão, Associação, Clusterização e Mineração de Textos.	Oracle <a href="http://www.oracle.com">www.oracle.com</a>
<b>SAS Enterprise Miner</b>	Classificação, Regras de Associação, Regressão e Sumarização.	SAS Inc. <a href="http://www.sas.com">www.sas.com</a>
<b>SPSS/Clementine</b>	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.	SPSS Inc. <a href="http://www.spss.com">www.spss.com</a>

# WEKA - Waikato Environment for Knowledge Analysis - Ferramenta open source de DM.

Algoritmos do Aprendizado Supervisionado



Algoritmos do Aprendizado Não-Supervisionado





## Data Mining with Weka

Everybody talks about Data Mining and Big Data nowadays. [Weka](#) is a powerful, yet easy to use tool for machine learning and data mining. This course introduces you to practical data mining.

The 5-week course is currently closed.

The course features:

- online access to chapters from [Data Mining \(3rd Edition\)](#)
- [CC-BY](#) videos & slides (see the [materials site](#))
- online assessment leading to a Statement of Completion ([example](#))
- English & Chinese captions on YouTube

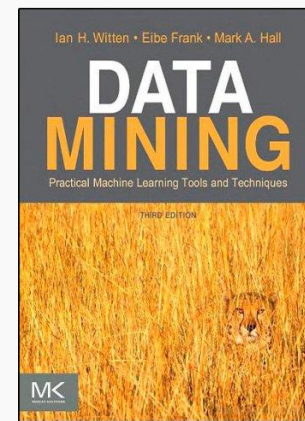
Subscribe to the [Announcements forum](#) for updates and reminders.

Please read the [Terms of Service](#) and [Participant Information Sheet](#) before registering.

Prof Ian H. Witten  
Department of Computer Science  
University of Waikato



YouKu version: [English subtitles](#); 中文字幕





*“Big Data é um tsunami ainda em alto mar.”*



*Data Scientist*

*Tendências em BD*



*Data Analyst*

Big Data Analytics in Cloud  
Armazenamento em Nuvem  
*Open Data*  
NoSQL

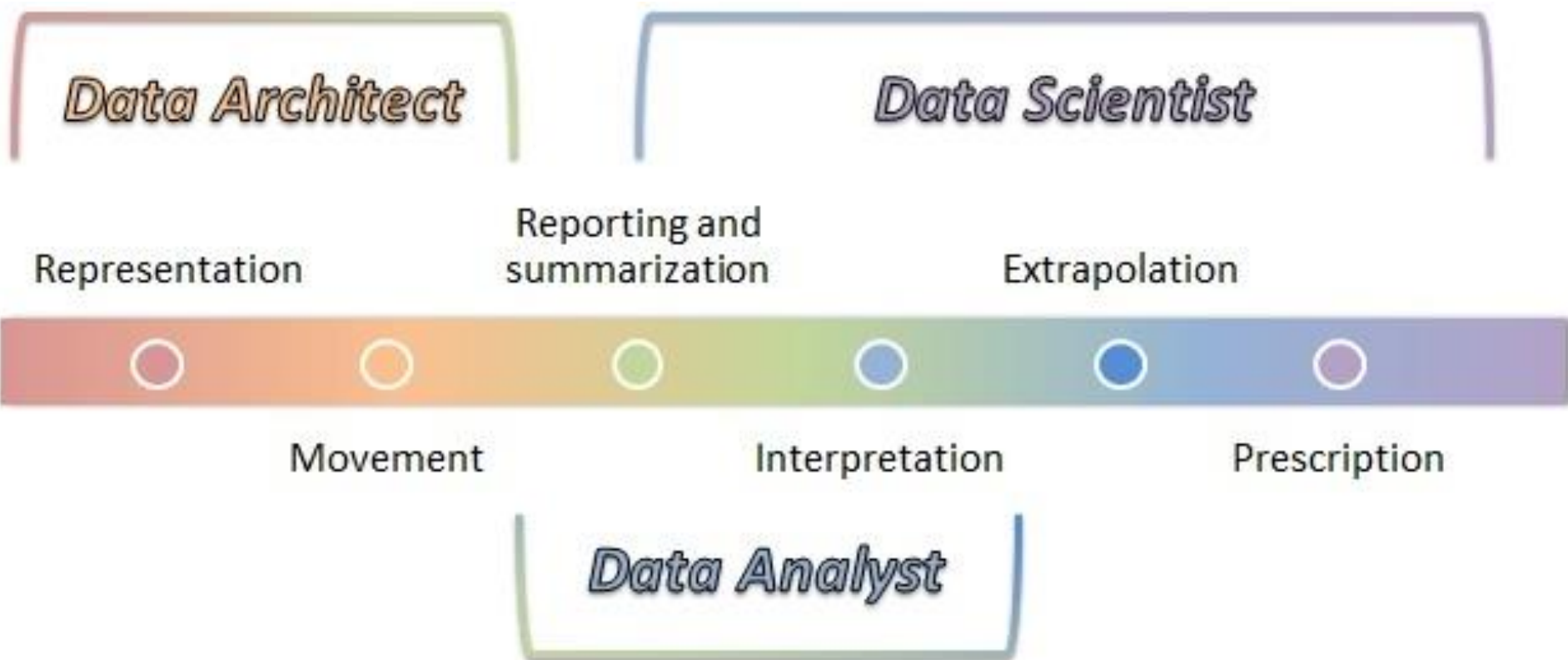


Figura – *Data Mining Lifecycle.*



Fonte: <http://youtu.be/tfaYKbbYnXU>

# Principais eventos/conferências da área

- SBBD - Simpósio Brasileiro de Banco de Dados.
- ACM SIGKDD - Conference on Knowledge Discovery and Data Mining
- ACM SIGMOD/PODS Conference
  - MOD - Management of Data / PODS - Principles of Database Systems.
- IEEE International Conference on Data Mining (ICDM)
- SIAM International Conference on Data Mining.
- VLDB - International Conference on Very Large Data Bases.
- DMIN - International Conference on Data Mining.
- DMKD - Workshop on Research Issues on Data Mining and Knowledge Discovery.

# Projetos



# Projetos



- **PESQUISA:**

- Processo de Descoberta de Conhecimento em Ambientes Virtuais de Aprendizagem da Educação a Distância (FACEPE/CNPq) - Até 2017.
  - 1 Bolsista de IC.
- Data Mining em Ambientes Virtuais de Aprendizagem para Educação a Distância (PIBITI/CNPq).
  - 1 Bolsista de IC – 2014-2015.
- Mineração de Dados Educacionais em Ambientes B-learning de Instituições Federais de Ensino Superior (PIBIC/UFRPE)
  - 2 estudantes de IC (PIC/UFRPE) – 2015-2016.

- **EXTENSÃO:**

- Inclusão Socioambiental nas Comunidades de SUAPE Utilizando Design Computacional Centrado no Humano (PRAE/UFRPE)
  - 1 Bolsista de extensão - 2015.

# Projeto com início em 2016



- PESQUISA/EXTENSÃO:
  - Sistema Computacional Colaborativo utilizando Dados Abertos dos Programas Sociais do Governo Federal.
    - 1 Bolsa
    - Seleção de bolsista (recebendo histórico escolar até 11/11/2015).



Profa. Roberta Macêdo M. Gouveia  
[robertammg@gmail.com](mailto:robertammg@gmail.com)

05/11/2015