setembro de 24

Data Mining
Visão Geral

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2024

www.inf.ufsm.br/~Joaquim/

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

*A maior parte deste material foi retirado do livro: “Joaquim V. C. Assunção. **Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R**. 192 páginas. Novatec. 2021. ISBN-10 : 6586057507.”

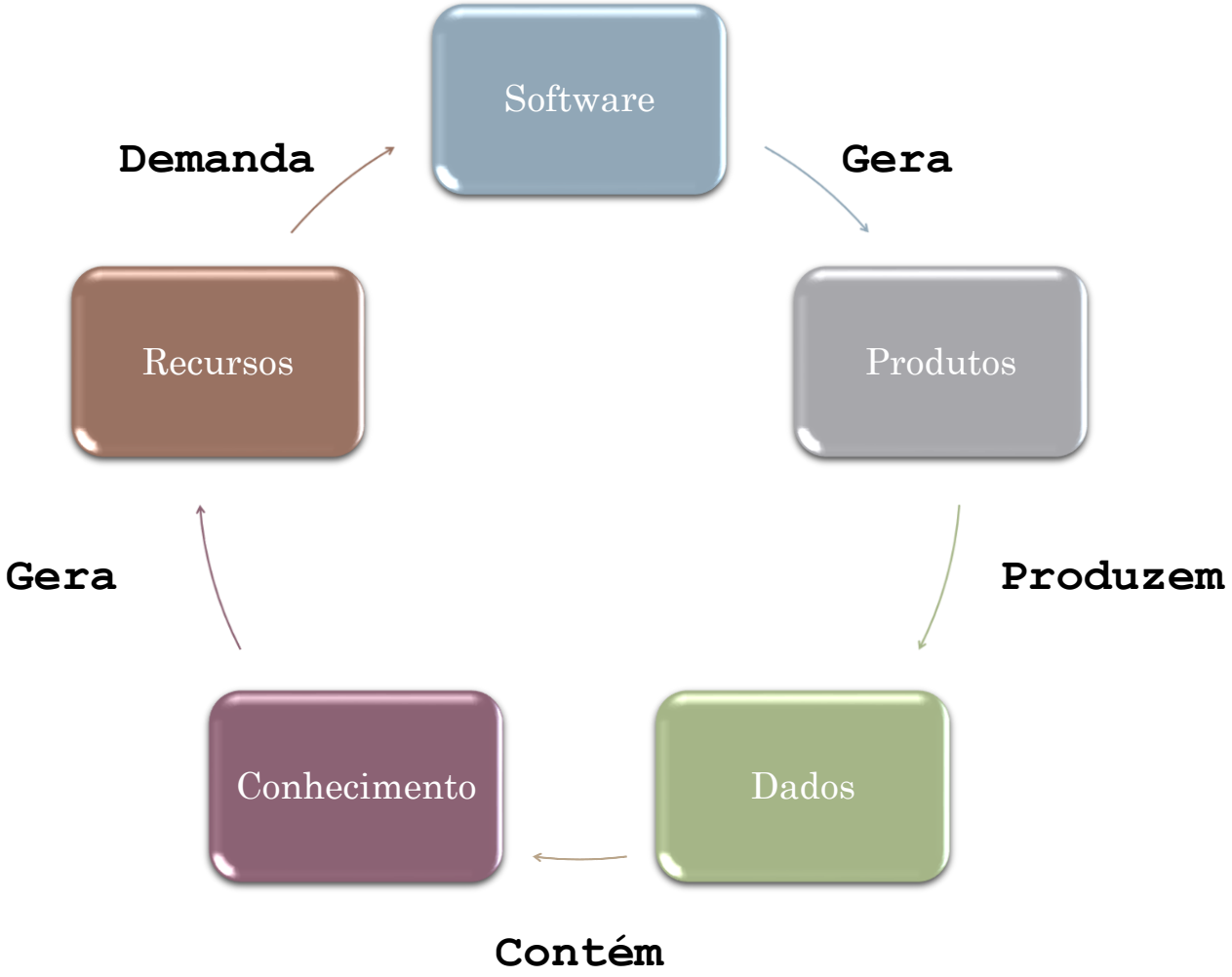
Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Prequel

A importância dos dados

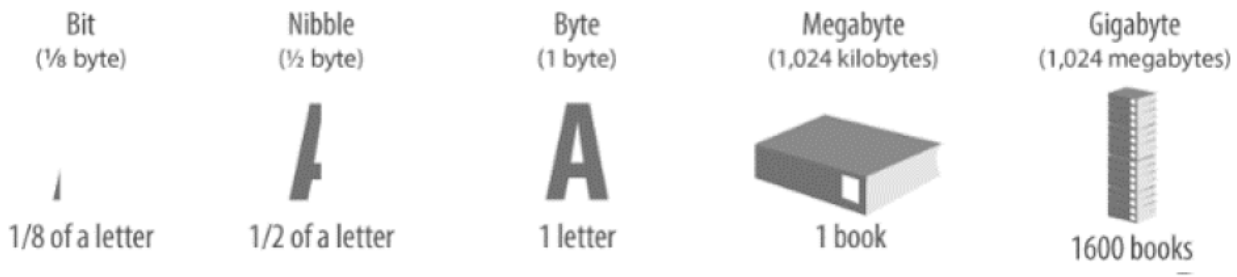
Ciclo da geração de dados



Perspectiva histórica

Qual o impacto da tecnologia?
E como a sociedade moderna é centrada em dados?

Armazenamento de Dados



Se considerarmos 1Gb para 1600 livros.

32 Gb = 51.000 Livros

256 Gb = 409.600 Livros

O poder do conhecimento

- Se Athena, a Deusa da sabedoria, lembrasse, em sua totalidade, em média 10 livros por ano, e tivesse 3000 anos de leitura, teria a memória de “apenas” 30 mil livros.

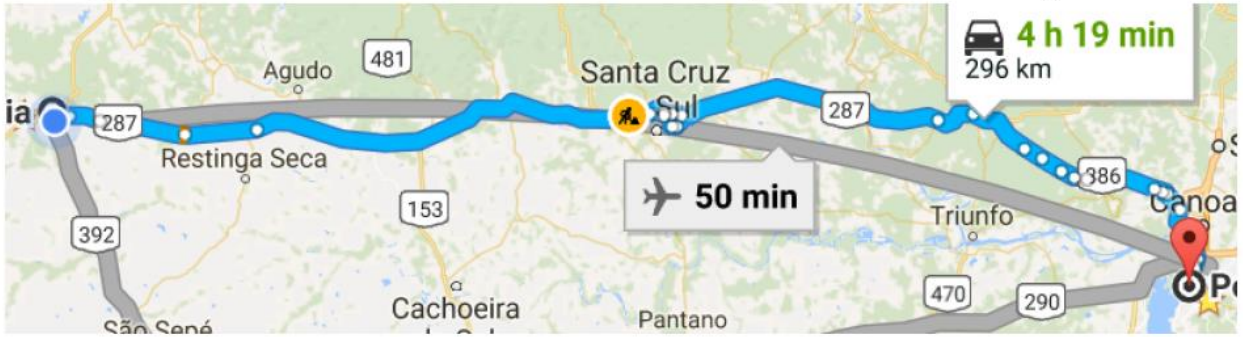
Produção de Dados

- Atualmente, estima-se que sejam produzidos cerca de 2,5 quintilhão [1] (17 zeros após o 5) de bytes de dados, todos os dias!
- Todos os livros do mundo (em cada escola, universidade, empresa etc.) correspondem a aproximadamente 6% dos dados.

[1] Levantamentos do Infosys e Network world (2007). Estimativa da IBM, em 2013, disponível em: www.ibm.com/big-data/us/en.

Produção de Dados

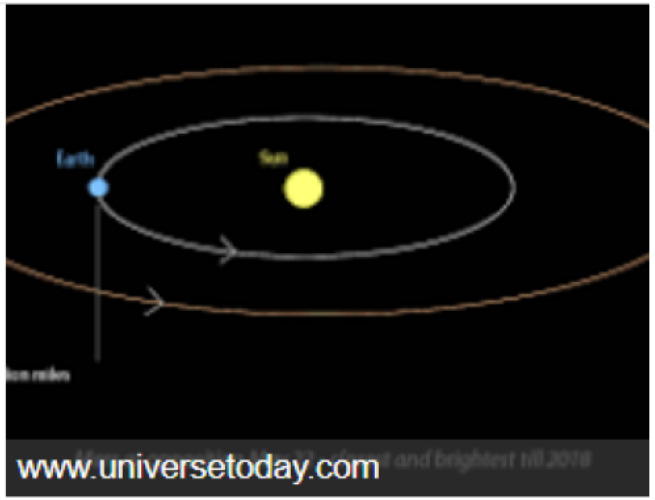
... Se pudéssemos transformar cada byte de dado em uma caixa com $1cm^2$. Poderíamos preencher a estrada: Santa Maria-Porto Alegre 4,2 bilhões de vezes (ida e volta).



Produção de Dados

... Se pudéssemos traçar uma linha usando um byte como $1cm$, daria 44 mil linhas até Marte.

The minimum distance from the Earth to Mars is about **54.6 million kilometers**. The farthest apart they can be is **about 401 million km**. The average distance is **about 225 million km**. Feb 29, 2012



A velocidade de uma mensagem/dados

Zeus, o Deus dos Deuses, enviava Hermes para entregar suas mensagens.

- Se Hermes viajasse a incríveis 3000 Km/h, levaria mais de 3h para entregar uma mensagem de São Paulo até Paris.

Produção de Dados

Perspectiva

Em torno de 90% dos dados no mundo foram criados nos últimos dois anos![2]

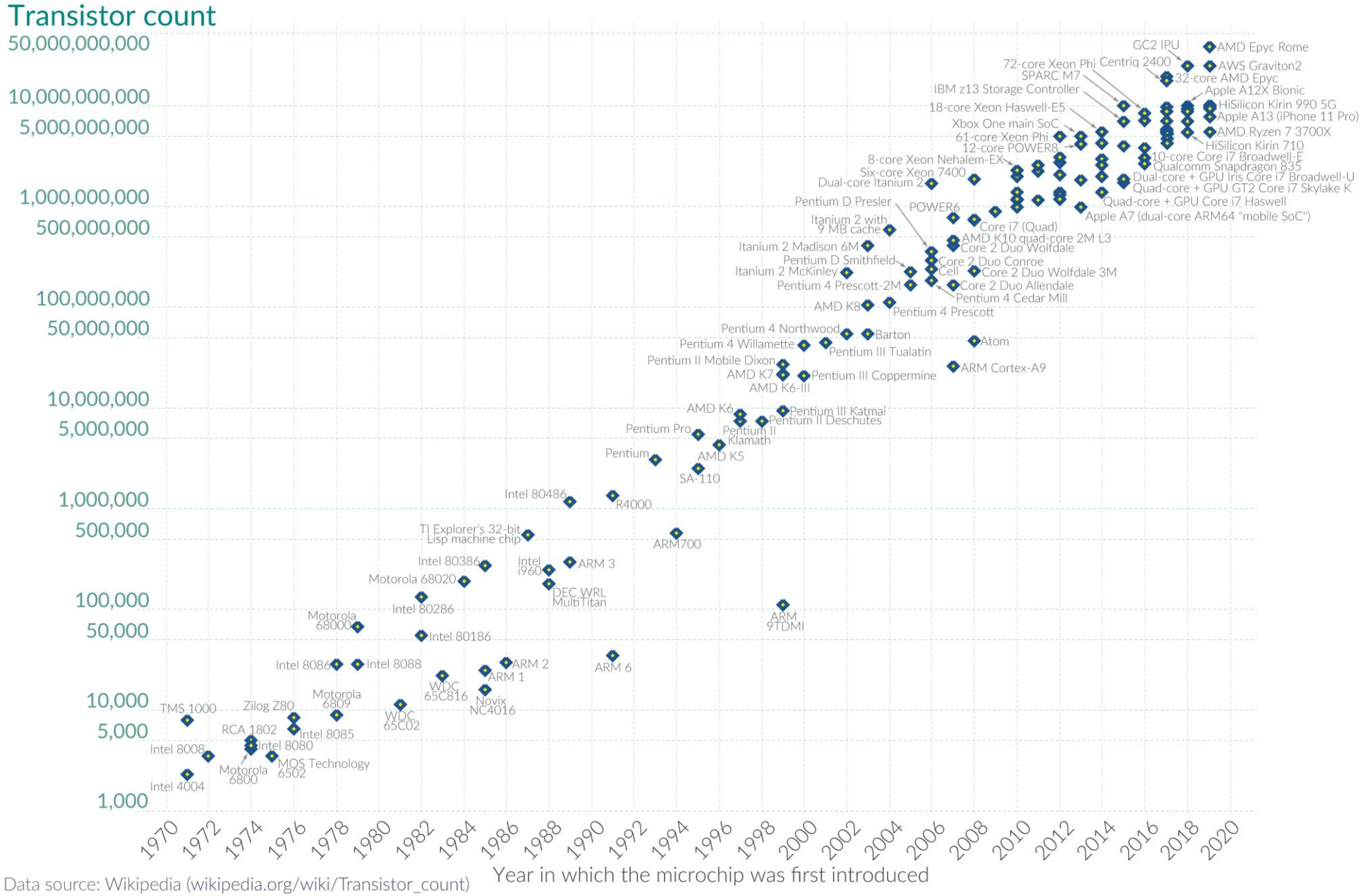
Do diferencial competitivo a necessidade. Empresas de grande porte contratam cada vez mais profissionais para análise e descoberta de conhecimento em Big Data.

[2] Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications. Wiley, 2014.

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data

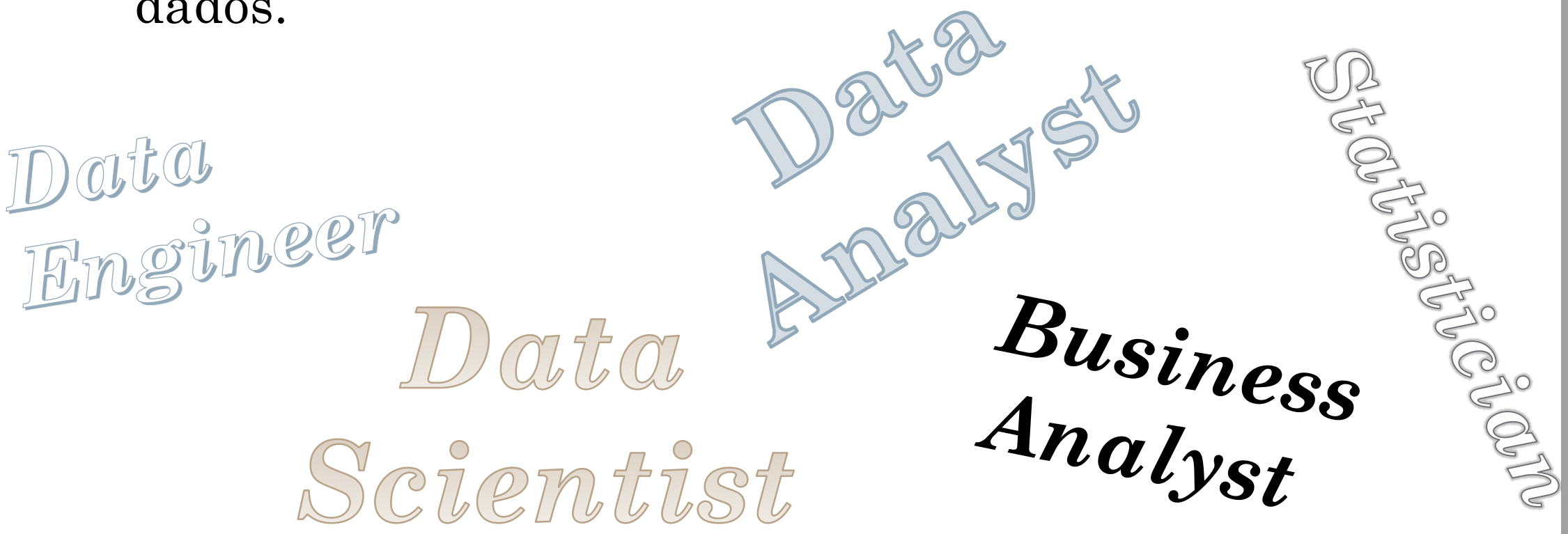


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Algumas Profissões

- Profissões foram criadas para se trabalhar com dados.



Uma disciplina base

- Minerar dados é cada vez mais comum nestas profissões.

Data Mining

O que é mineração de dados?

- Termo (*data mining*) criado nos anos 90 para a confluência de ideias de **estatística** e **ciência da computação** (aprendizagem de máquina e métodos de banco de dados) aplicados em grandes bancos de dados em ciência, engenharia e negócios.

* Primeiro workshop que usou Data Mining no nome foi em 1995

** Figura de retirada de Han *et. al., Data Mining, concepts and techniques*. 2011.

Wal-Mart. Fraldas & cervejas.

- Correlação entre fraldas e cervejas…

* Figuras retiradas de https://blog.wiser.com/offline-dynamic-pricing/ e https://blog.patternbuilders.com/2011/03/02/tales-of-beers-and-diapers/ em 2018.

Target e propagandas

- Com base nos dados da empresa, a Target envia para seus clientes cupons e propagandas personalizadas. Em um destes casos, a empresa enviou roupas de maternidade, mobília de berçário e fotos de bebês sorridentes para uma adolescente...

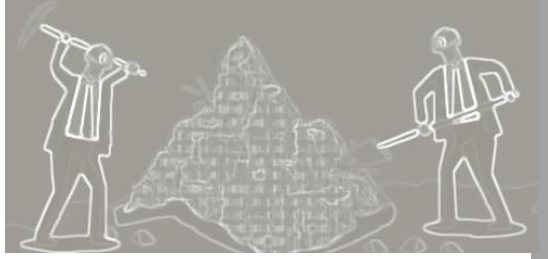
TARGET

* Figura retirada de https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#3e4f5d256668 em abril de 2018.

Target e propagandas

- Um pai foi falar com o gerente da Target em Minneapolis. → "Minha filha pegou isso no correio" ... "Ela ainda está no ensino médio e você está enviando cupons para roupas de bebê e berços? Você está tentando incentivá-la a engravidar?"

* Figura e texto retirados de https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#3e4f5d256668 em abril de 2018.



O que é mineração de dados?

- Constante debate e confusão entre termos (*knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology*).
- Terminologia não padrão. E.g., *feature = independent variable*, *target = dependent variable*, *case = row*

Falamos em "Mineração de ouro" não em "Mineração de rochas". ... Dai o termo "*knowledge mining from data*"

O que é mineração de dados?

- “Processo de descoberta automática de informações úteis em grandes depósitos de dados” [Tan et. al., 2006]
- Google “*data mining definition*” -> “*the practice of examining large databases in order to generate new information*.”
- **“*Data mining* *is the application of specific algorithms for extracting patterns from* *data*”** [Fayyad et. al., 1996]

O que é mineração de dados?

"*The nontrivial extraction of implicit, previously unknown, and potentially useful information from data**".

O processo, não trivial, de extrair informação implícita, potencialmente útil e previamente desconhecida de dados.

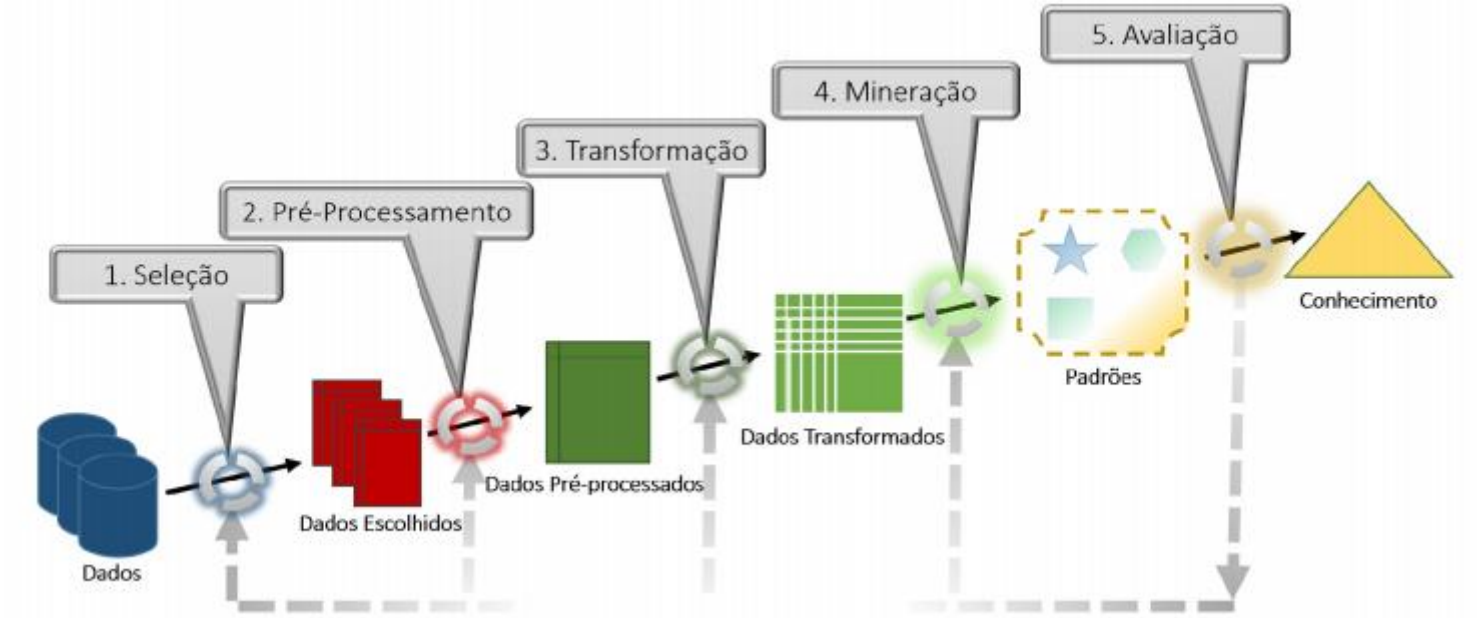
*G. Piatetsky-Shapiro & W. J. Frawley, Knowledge Discovery in Databases, 1991.

Banco de Dados vs Mineração de Dados

Banco de Dados	Mineração de Dados
Consulta	
• Bem definida	• Fraca em definições
• SQL (maioria)	• Sem linguagem definida
Saída	
Subconjunto de um BD	Não é um subconjunto de um BD

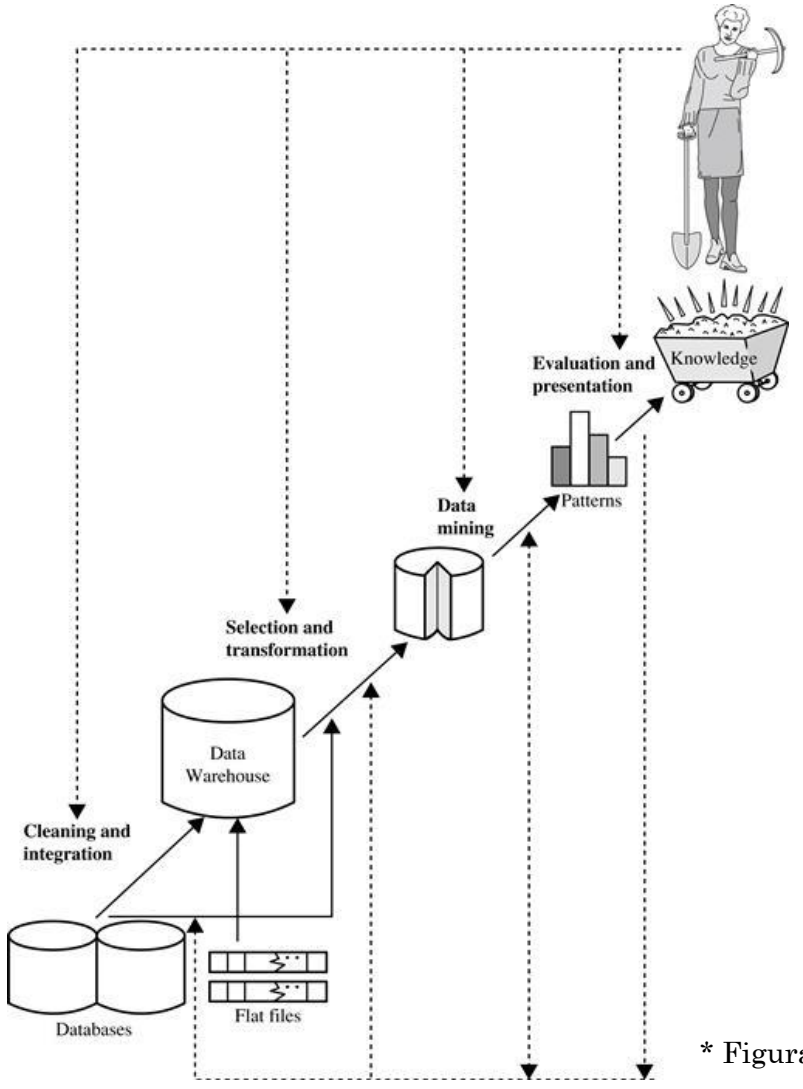
KDD (*Knowledge Discovery from Data)*

- KDD é um processo que compreende os passos comuns desde a coleta de dados em um banco de dados até a obtenção de padrões úteis e previamente desconhecidos.



* Figura mostrando o processo de KDD, proposto por Fayyad, 1996.

KDD (*Knowledge Discovery from Data)*



- KDD é um processo que compreende os passos comuns desde a coleta de dados em um banco de dados até a obtenção de padrões úteis e previamente desconhecidos.

- Jiawei Han e seus coautores julgaram adequado incluir um *data-werehouse* no processo de mineração.

* Figura de retirada de Han *et. al., Data Mining, concepts and techniques.* 2011.

KDD (*Knowledge Discovery from Data)*

- "*Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.*"

KDD – Passos

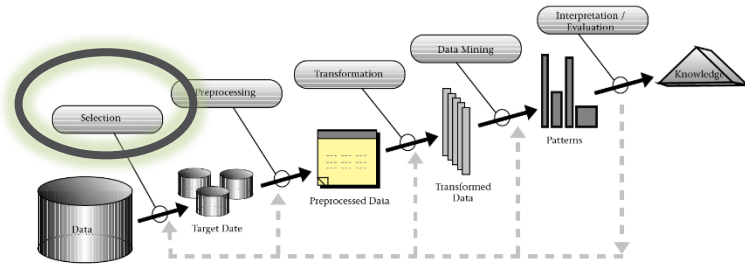
1. ***Data Selection***
2. *Data Preprocessing*
3. *Data Transformation*
4. *Data Mining*
5. *Interpretation / Evaluation / Presentation*

KDD – Passos

1. ***Data Selection***

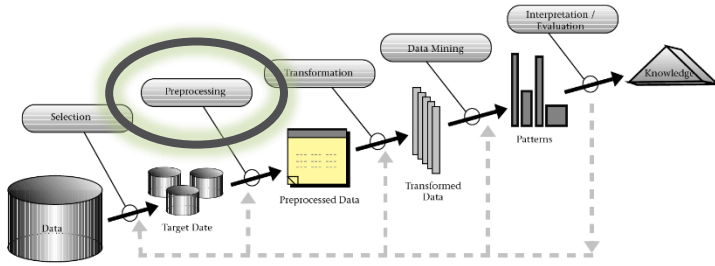
   Uma pré-seleção dos dados deve ser feita para obter os dados desejados, isto inclui selecionar dados de diferentes fontes (BD) para obter aquilo que será potencialmente relevante para a análise.

2. *Data Preprocessing*
3. *Dat Transformation*
4. *Data Mining*
5. *Interpretation / Evaluation / Presentation*



KDD – Passos

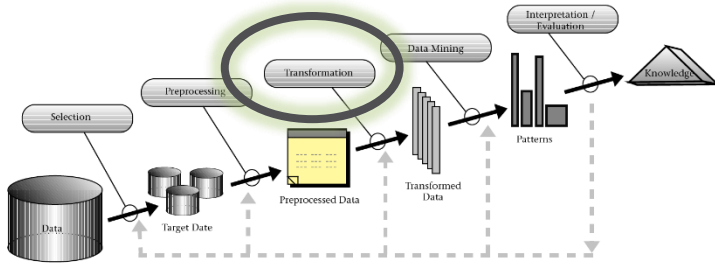
1. *Data Selection*
2. ***Data Preprocessing***, inclui:
   I. **limpeza de dados** (***cleaning***), que consiste em remover ruídos e dados inconsistentes.
   II. **Integração dos dados** (***Integration***), que consiste em integrar múltiplas fontes de dados para serem combinadas no arquivo pretendido.
   III. **Seleção** (***selection***), que inclui escolher os dados propícios a serem analisados (Versão de J.Han).
3. *Data Transformation*
4. *Data Mining*
5. *Interpretation / Evaluation / Presentation*



KDD – Passos

1. *Data Selection*
2. *Data Preprocessing*
3. ***Data Transformation***

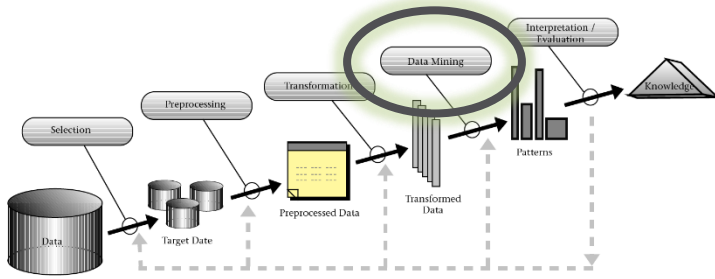
   Aqui os dados devem ser transformados para os formatos apropriados para a mineração. Isto depende do tipo(s) de algoritmo(s) que será rodado.
4. *Data Mining*
5. *Interpretation / Evaluation / Presentation*



KDD – Passos

1. *Data Selection*
2. *Data Preprocessing*
3. Data Transformation
4. ***Data Mining***

   Nesta parte, métodos inteligentes são aplicados com o objetivo de extrair padrões ou quaisquer informações que sejam potencialmente úteis e previamente desconhecidas.
5. *Interpretation / Evaluation / Presentation*

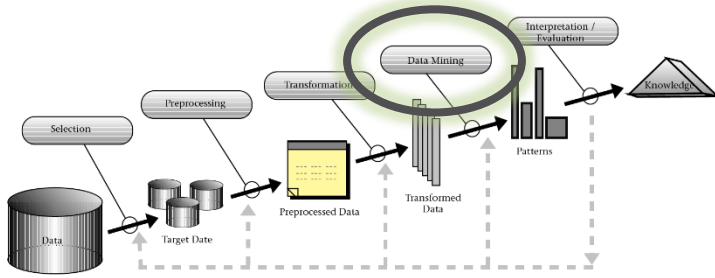


KDD – Passos

1. ***Data Mining***

   Nesta parte, métodos inteligentes são aplicados com o objetivo de extrair padrões ou quaisquer informações que sejam potencialmente úteis e previamente desconhecidas.

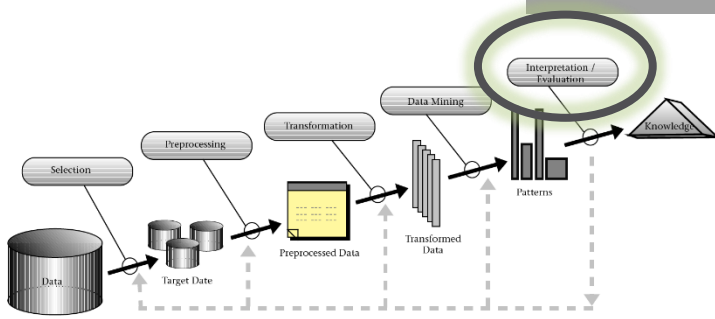
Figura 1.3: As quatro categorias de algoritmos mais usadas em mineração de dados.



KDD – Passos

1. *Data Selection*
2. *Data Preprocessing*
3. Data Transformation
4. Data Mining
5. ***Interpretation/Evaluation/Presentation***

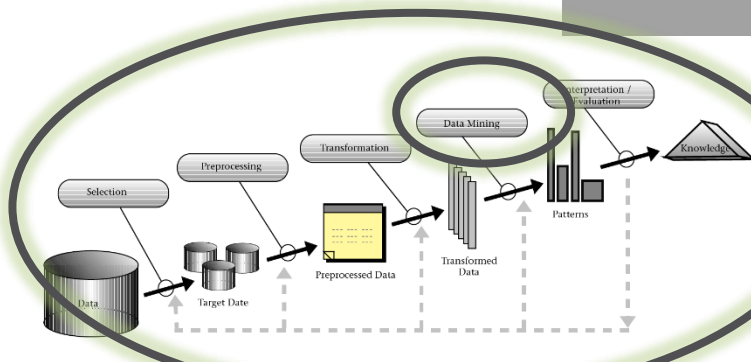
Normalmente, ter informações não é suficiente. Precisamos ter maneiras de apresentar o conhecimento obtido com essas informações, ou simplesmente expor a saída de modo claro o suficiente para que outras pessoas consigam obter algum conhecimento sobre elas.



?!!

Talvez pelo efeito de marketing ou pela simplicidade que o termo possui em relação a sigla, o termo "*data mining*" virou sinônimo de *KDD*.

Então quando fala-se em "data mining", frequentemente se referencia ao KDD.



O que pode ser Minerado?

- Quase todo conjunto de dados que:

1. Possua volume minimamente significativo;
2. Seja passível de se questionar algo a respeito;
3. Não seja trivial.

Funcionalidades/tipos de mineração

- As funcionalidades da mineração de dados são divididas de acordo com o propósito da mineração e o tipo de algoritmo de *machine learning* a ser usado. Alguns exemplos:
- Caracterização e discriminação.
- Associação.
- Correlação.
- Classificação e regressão.
- Análise de Grupos.

Categorias

- As funcionalidades da mineração são usadas para especificar um objetivo geral.
- Estes objetivos podem ser agrupados em duas categorias: **Descritivo** e **Preditivo**, ou mineração descritiva e mineração preditiva.

Categorias

- Mineração de cunho **Descritivo** serve para descrever características de variáveis ou objetos em um conjunto de dados.
- Mineração de cunho **Preditivo** usa os dados para fazer predições.

Exemplo

- Um gerente de relacionamento com clientes da EmprezaXYZ pode solicitar a seguinte tarefa de mineração de dados: Resuma as características dos clientes que gastam mais de R$ 9000 no ano passado na EmprezaXYZ . ... O resultado obtido é um perfil geral desses clientes, com 40 a 50 anos de idade, empregados e com excelente classificações de crédito.

Exemplo adaptado de Han *et. al., Data Mining, concepts and techniques*. 2011.