Introdução

Mineração de Dados

Ronaldo C. Prati[1]

[1]Universidade Federal do ABC (UFABC), ronaldo.prati@ufabc.edu.br

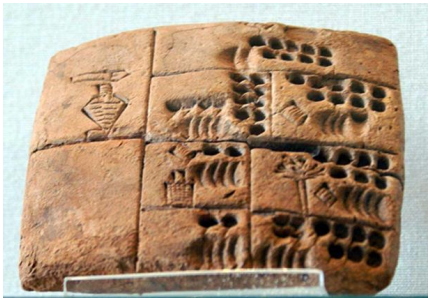
Introdução

Cultura de dados

"*In God we trust, all others bring data.*" — W Edwards Deming

- Desde os primórdios de nossa civilazação, coletamos dados

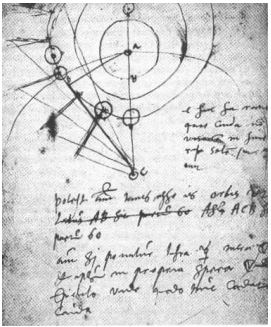
Osso de Ishango	Tábua de Kushin



Importância dos Dados

"*If we have data, let's look at data. If all we have are opinions, let's go with mine.*" — Jim Barksdale

- Grandes descobertas científicas foram feitas a partir de dados

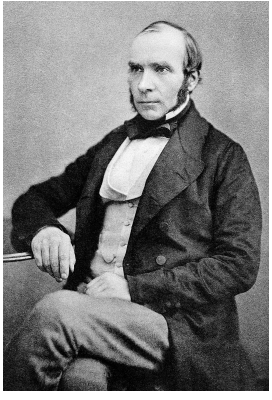
Johannes Kepler	Anotações de Tycho Brahe
	

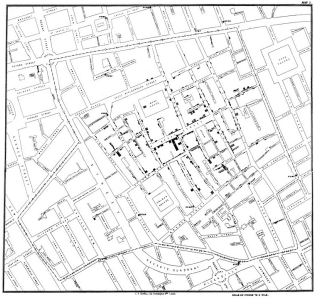
Importância dos Dados

"*If we have data, let's look at data. If all we have are opinions, let's go with mine.*" — Jim Barksdale

- Grandes descobertas científicas foram feitas a partir de dados

John Snow	Mapa de Mortes por Cólera



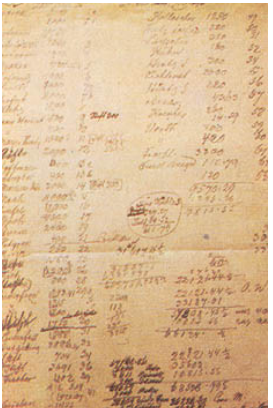


Importância dos Dados

"*If we have data, let's look at data. If all we have are opinions, let's go with mine.*" — Jim Barksdale

- Grandes descobertas científicas foram feitas a partir de dados

Gregor Mendel	Anotações de Mendel

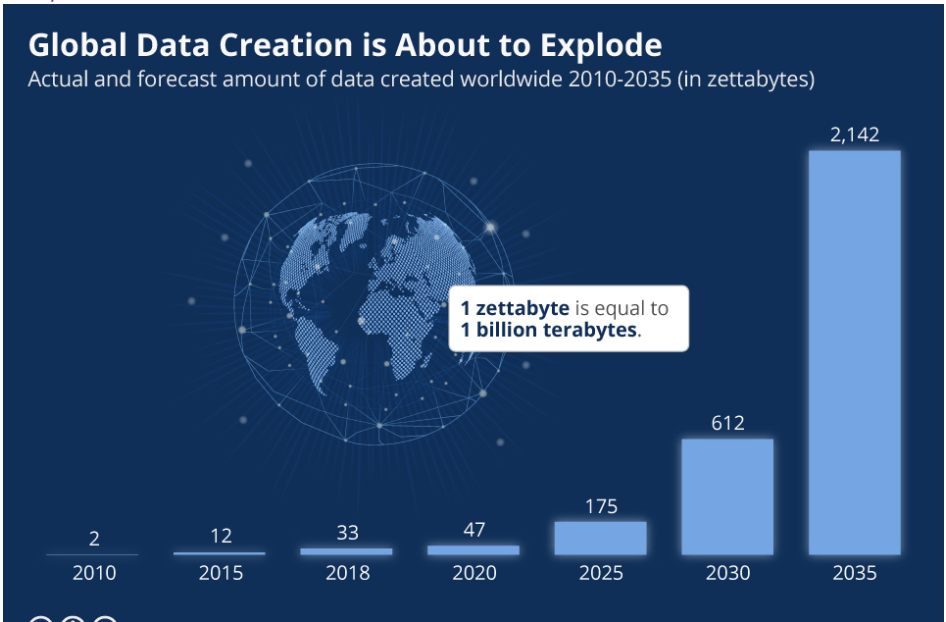


Dados, dados e mais dados

"*We are drowning in information but starved for knowledge.*" - John Naisbitt

- Progressos na coleta e armazenamento de dados tornaram comuns bases de dados enormes
- Desde o início dos anos 90
- Não há indícios de que isso irá parar tão cedo
- Como tirar o melhor proveito possível dos dados?

Dados, dados e mais dados



Dados, dados e mais dados

Figure 1: Comércio Eletrônico

- O que você procura
- O que você compra
- O que você critica

Dados, dados e mais dados

Figure 2: Twitter

- O que você escreve
- De onde você escreve
- Com quem você conversa
- Quem você lê

Dados, dados e mais dados

Figure 3: Facebook

- Quem é você
- O que você faz
- Com quem e quando você faz
- O que você acha sobre o que os outros fazem

Dados, dados e mais dados

Figure 4: *Smartphones*

- Com quem você fala
- Por onde você anda
- Como você se locomove
- Sem entrar em detalhes sobre *apps*

Dados, dados e mais dados

Figure 5: Internet das coisas

- Acompanhamento 24/7
- Sensores em chão de fábrica, plantações, cidades, rodovias, etc.

Dados, dados e mais dados

- Esses são casos extremos, mas a situação é comum:
  - Concessionárias
  - Hospitais
  - Escolas
  - Bancos
  - ...
- Empresas ainda tomam decisões importantes considerando apenas em intuição
- "HiPPO" *the highest-paid person's opinion*
  - Pessoas se apoiam muito em experiência e intuição mas pouco em dados

Mineração de Dados

Mineração de Dados

- Amadureceu conforme as bases de dados cresceram em tamanho e complexidade.
- Disciplina interdisciplinar relacionada a:
  - Estatística
  - Bancos de dados/Data warehousing
  - Métodos de modelagem e visualização de dados
  - Reconhecimento de Padrões
  - Sistemas Especialistas e Aquisição de Conhecimento
- Fronteiras entre as áreas **não** são rígidas

Mineração de Dados

- Recursos
  - Fundamentos teóricos/matemáticos
  - Aprendizado de Máquina
  - Inferência Lógica
  - Estatística e sistema dinâmicos
  - Sistemas gerenciadores de bases de dados

Mineração de Dados

- Áreas relacionadas
  - Business Intelligence
  - Data Science
  - Big Data
  - Predictive Analytics ...

Mineração de Dados

- Inicialmente era conhecida como a etapa de extração de padrões dentro do Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*)
- Com o tempo alguns pesquisadores começaram a utilizar como sinônimos

Mineração de Dados

Definição (Hand, Manilla & Smith) *Mineração de dados é a análise de bases de* ***dados observacionais*** *(frequentemente grandes) para encontrar* ***relações desconhecidas*** *e para sumarizar os dados em formas que sejam* ***compreensíveis*** *e* ***úteis*** *para o dono dos dados.*

- Cervejas e fraldas (*fun fact*): https://www.kdnuggets.com/news/2000/n14/8i.html

Mineração de Dados

- Alguns termos na definição chamam atenção:
  - dados observacionais
  - relações desconhecidas
  - formas compreensivas
  - formas úteis

Processo de Mineração de dados

- Minerção de processos é um processo amplo de encontrar conhecimento nos dados.
- Diferentes abordagens foram propostas para formalizar esse processo

Processo KDD (Fayyad et. al)

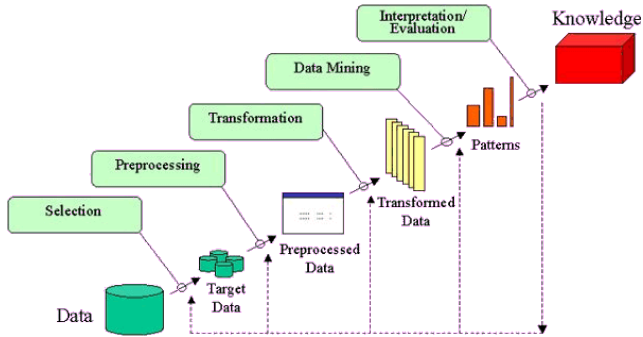


Figure 6: KDD

Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

Processo KDD (Fayyad et. al)

- Processo com diferentes fases:
  1. Seleção – Criação de um conjunto de dados alvo, focando em um conjunto de variáveis e exemplos em que o processo será aplicado
  2. Limpeze e pré-processamento, com o intuito de obter um conjunto consistente de dados
  3. Transformação: preparação dos dados de acordo com a técnica de modelagem a ser utilizada
  4. Mineração de dados: busca por padrões de interesse, dependendo dos objetivos do processo
  5. Interpretação/avaliação: análise dos padrões obtidos

CRISP-DM

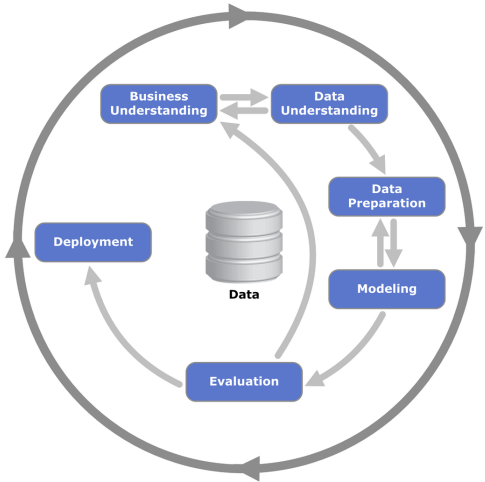


Figure 7: Cross-industry standard process for data mining

Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining." Proceedings of the 4th international conference on the

CRISP-DM

- "Equipara" o processo de KDD e Mineração de Dados

1. Entendimento do problema – foco na fase de entendimento dos objetivos e requisitos do processo, a partir de uma pespectiva do negócio. O resultado é a formalização como um problema de mineração de dados, e um plano inicial para atingir esses objetivos.
2. Entendimento dos dados - faz uma análise exploratória dos dados, para identificar problemas e descobrir os primeiros insights ou detectar
3. Preparação dos dados - incorpora as fases de seleção, limpeza e transformação do processo de KDD
4. Modelagem - similar a fase de mineração de dados do processo de KDD
5. Avaliação - verifica se os resultados obtidos são compatíveis com os elencados na etapa de entendimento do problema
6. Implantação - busca incorporar os resultados obtidos nos processos decisórios.

TDSP

- Recentemente, um grupo de estudos da Microsoft propos o Processo de Ciência de Dados em Equipe (do inglês *Team Data Science Process* - **TDSP**)
  - Processo ágil baseado em SCRUM
  - Reconhece a necessidade de uma equipe, com diferentes atores
    - Arquiteto de solução
    - Gerente de projeto
    - Engenheiro de dados
    - Cientista de dados
    - Desenvolvedor de aplicativos
    - Líder do projeto

TDSP

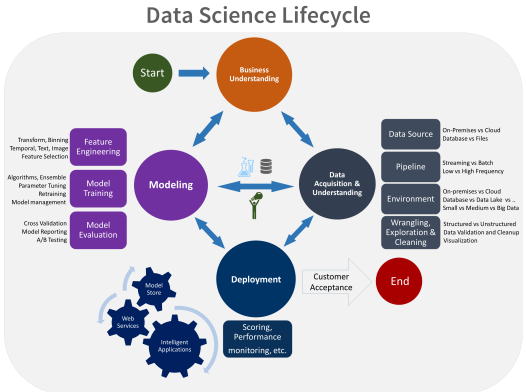


Figure 8: Processo de Ciência de Dados em Equipe

TDSP helps improve team collaboration and learning. It contains a distillation of the best practices and structures from Microsoft and others in the industry that facilitate the successful implementation of data science initiatives." – Microsoft, 2020

TDSP

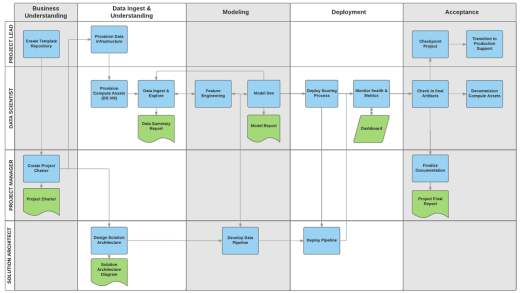


Figure 9: Processo de Ciência de Dados em Equipe

O que não iremos cobrir

- Extração de dados
  - Crawlers
  - APIs
- Armazenamento e organização de dados
  - Data Warehouse
  - Online Analytic Processing (OLAP)
- Inplantação
  - Servições
  - Aplicações
  - Integração

Referências

- D. Hand, H. Manilla e P. Smith. Principles of Data Mining. **Capítulo 1**
- S. Rezende. Sistemas Inteligentes: Fundamentos e Aplicações. **Capítulo 12**
- P. Tan, M. Steinbach e V. Kumar, Introduction to Data Mining. **Capítulo 1**
- Breiman, Leo. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (2001), no. 3, 199–231. doi:10.1214/ss/1009213726. https://projecteuclid.org/euclid.ss/1009213726
- Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34
- Saltz, Jeffrey S., and Nancy W. Grady. "The ambiguity of data science team roles and the need for a data science workforce framework." 2017 IEEE international conference on big data (Big Data). IEEE, 2017.