# Aula: Introdução à Ciência de Dados: Pipeline e Modelos

**IMD1151 - Ciência de Dados**

Prof. Heitor Florencio
heitorm@imd.ufrn.br

Prof. Daniel Sabino
daniel@imd.ufrn.br

UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

metrópole DIGITAL

# Aula: Introdução à Ciência de Dados: Pipeline e Modelos

IMD1151 - Ciência de Dados

Prof. Heitor Florencio

heitorm@imd.ufrn.br

- Modelos de Processos de Ciência de Dados
- CRISP-DM
- KDD
- Comparação CRISP-DM e KDD
- Outros modelos e adaptações


Você já fez alguma análise que **não gerou algo significativo**?


Você já fez algum dashboard admirável que **ninguém (cliente) olhou**?


Existem modelos para Projetos de Ciência de Dados
INFORMATION
DATA ANALYSIS
PROFIT
01011011

# Modelos de Processos de Ciência de Dados

UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

imetrópole DIGITAL

# **REVISÃO:** Definições de **Ciência de Dados**

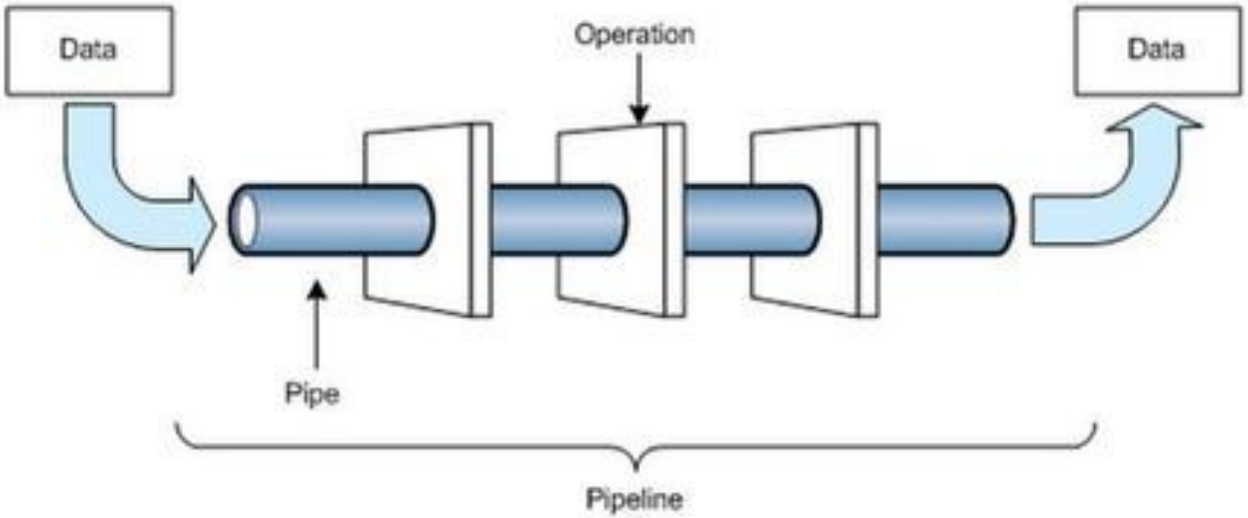
- “Combina as áreas de computação, estatística/matemática e conhecimento de negócios para **descobrir conhecimento** implícito nos dados, evidenciando padrões e associações” (DEKHTYAR, 2023)
- “Ciência de dados é um campo acadêmico interdisciplinar que utiliza estatística, computação, **processos** e sistemas para extrair conhecimento e insights de dados potencialmente ruidosos, estruturados ou não estruturados” (Wikipedia)
- A ciência de dados é a ciência que estuda as **técnicas e processos** para gerar conhecimentos (insights) a partir do conjunto de dados.
- **“A ciência de dados envolve todo o ciclo de vida do dado, da produção ao descarte” (AMARAL, 2016)**

# Modelos de Processos de Ciência de Dados

- Os **Processos de Ciência de Dados** definem as **ETAPAS** para a descoberta de conhecimento a partir do conjunto de dados.
- Os modelos de processos foram criados e vinculados ao termo mineração de dados. No entanto, mineração de dados é uma etapa do processo.
    - Mineração de dados: é o processo de extração de conhecimento em grandes quantidades de dados (HAN; KAMBER, 2006).

- O termo “Pipeline” de Ciência de Dados é usual.

# Pipeline de Dados




Um pipeline de dados — os dados de entrada são transformados em uma série de fases em dados de saída.

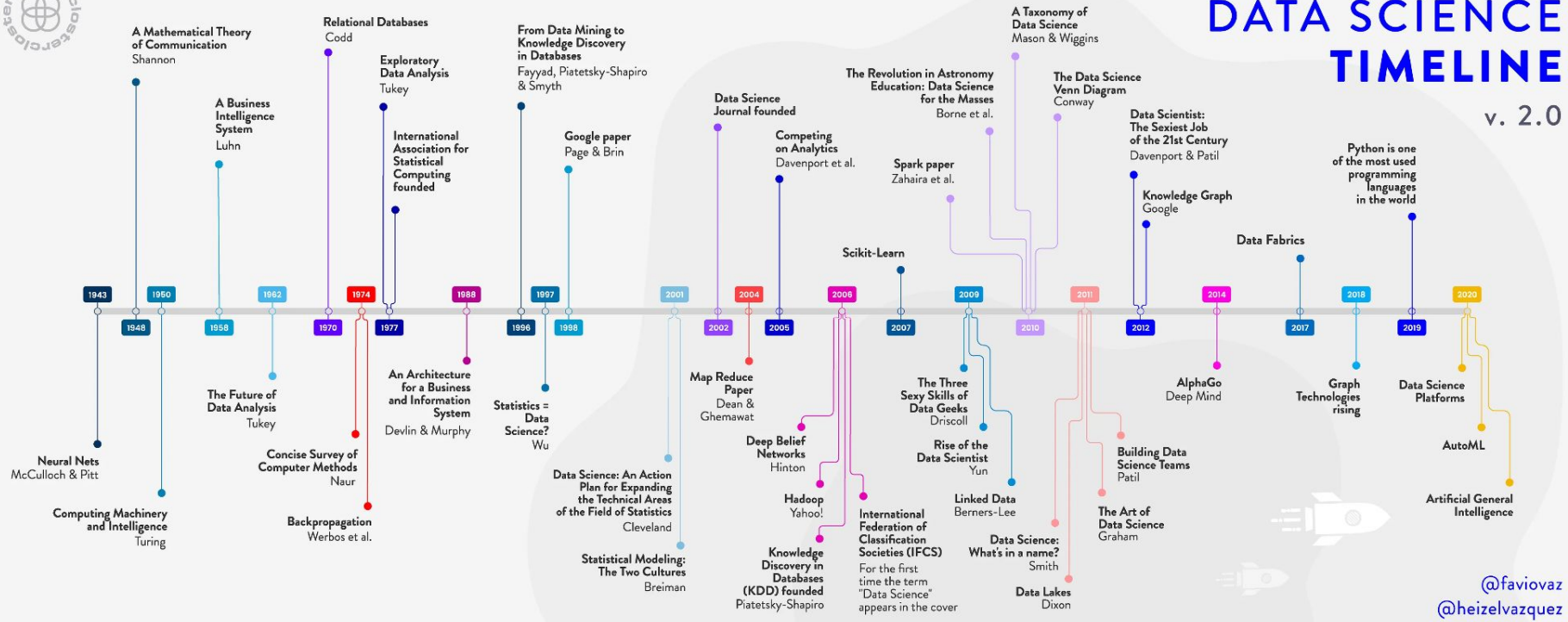
# **Histórico da Ciência de Dados** e os Modelos de Processos

UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

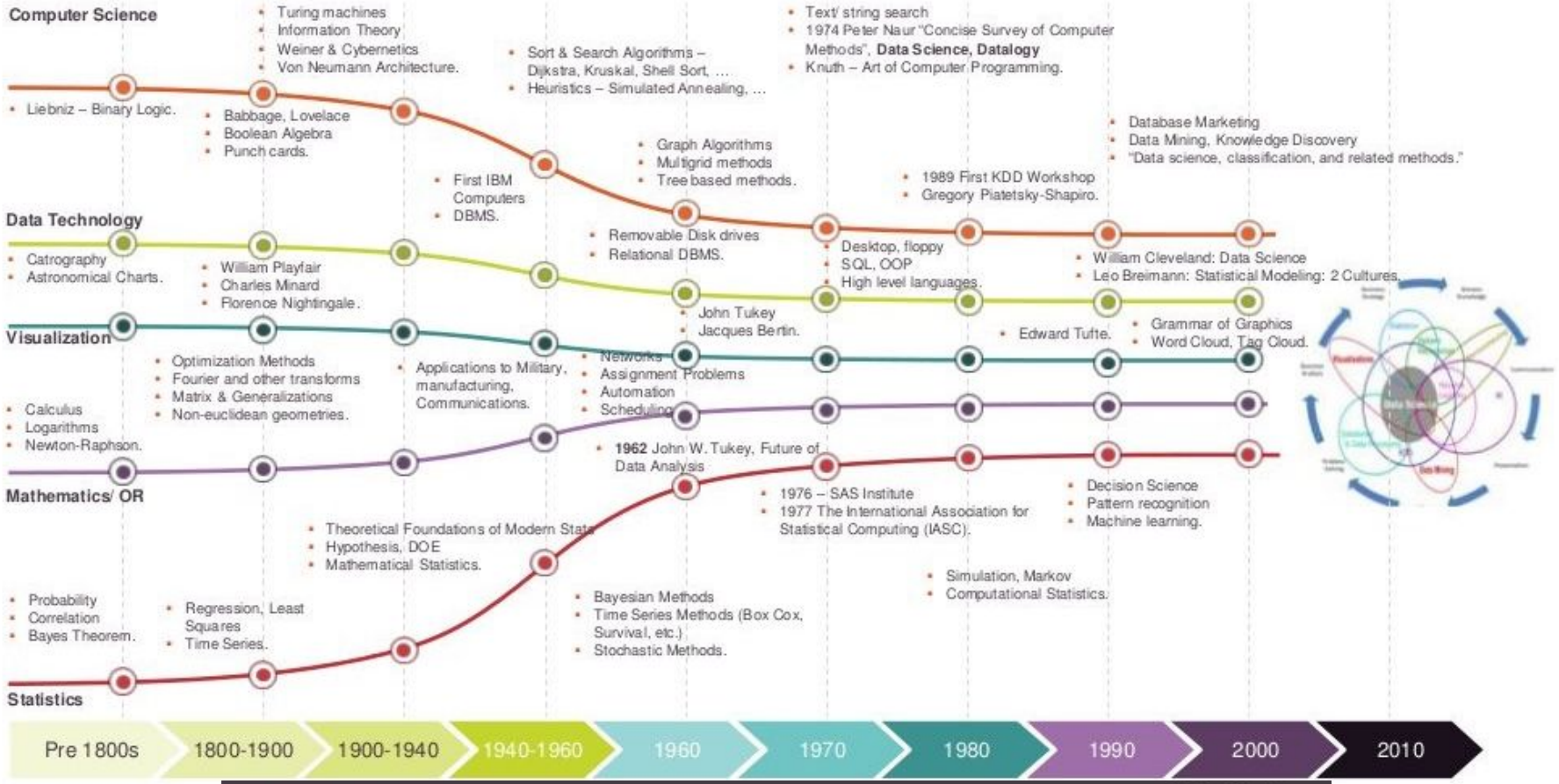
imetrópole DIGITAL

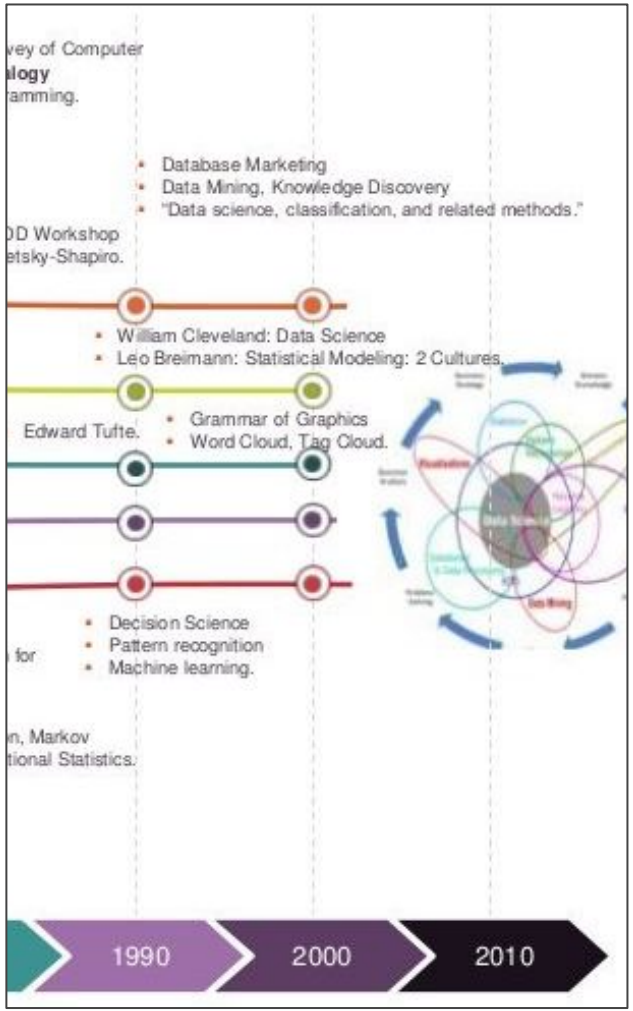
# DATA SCIENCE TIMELINE

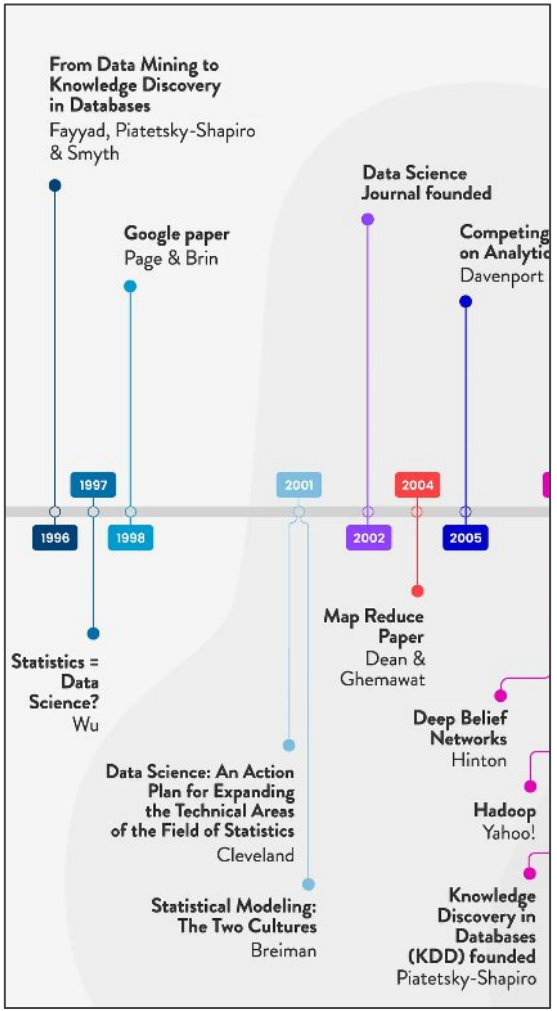
v. 2.0




@faviovaz
@heizelvazquez




vey of Computer
ology
amming.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
DD Workshop
etsky-Shapiro.
William Cleveland: Data Science
Leo Breimann: Statistical Modeling: 2 Cultures.
Edward Tufte.
Grammar of Graphics
Word Cloud, Tag Cloud.
Decision Science
Pattern recognition
Machine learning.
for
n, Markov
tional Statistics.
1990
2000
2010


From Data Mining to Knowledge Discovery in Databases
Fayyad, Piatetsky-Shapiro & Smyth
Google paper
Page & Brin
Data Science Journal founded
Competing on Analytic
Davenport
1996
1997
1998
2001
2002
2004
2005
Statistics = Data Science?
Wu
Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics
Cleveland
Statistical Modeling: The Two Cultures
Breiman
Map Reduce Paper
Dean & Ghemawat
Deep Belief Networks
Hinton
Hadoop
Yahoo!
Knowledge Discovery in Databases (KDD) founded
Piatetsky-Shapiro

# Modelos de Processos (pipeline) de Ciência de Dados

- CRISP-DM
- KDD
- SEMMA
- Adaptações

# CRISP-DM
(Cross Industry Standard Process for Data Mining)

UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

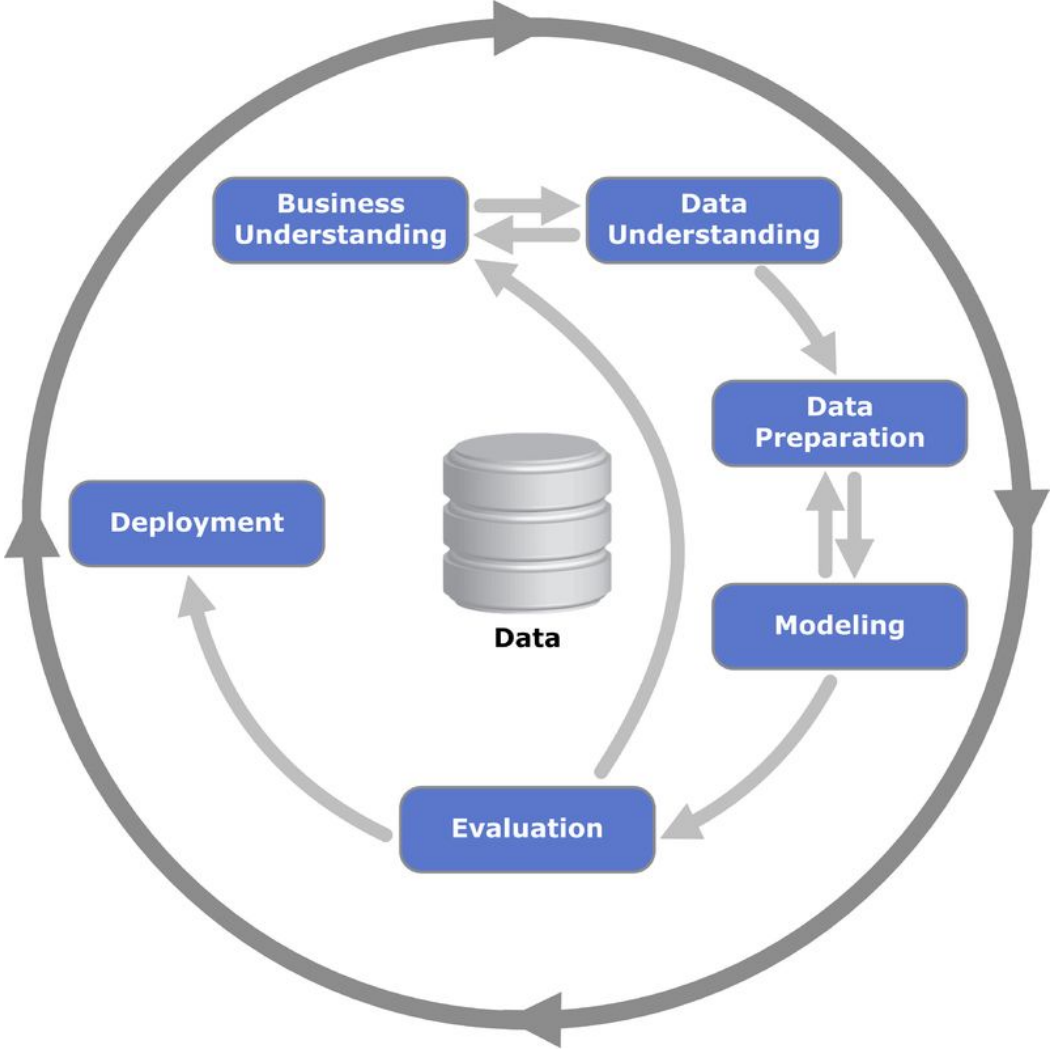
metrópole DIGITAL

# CRISP-DM

- Projeto concebido em 1996 (1999) por meio de um consórcio entre empresas (DaimlerChrysler, SPSS e NCR).
- **CRISP-DM**: Processo Padrão Inter-Indústrias para Mineração de Dados.
- Descreve o processo de análise de dados ou mineração de dados em 6 fases principais:

    1. Entendimento do negócio;
    2. Compreensão dos dados;
    3. Preparação dos dados;
    4. Modelagem;
    5. Avaliação;
    6. Implantação.

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

# CRISP-DM - Entendimento do negócio

**O que fazer?**

- Identificar o problema;
- Entender a necessidade dos dados;
- Levantar os potenciais valores da análise.


**Como fazer?**

- Faça perguntas sobre o negócio!
- Entenda a perspectiva e a realidade do cliente!

- Qual(is) o(s) objetivo(s) da análise?
- Qual(is) problema(s) queremos solucionar?

# CRISP-DM - Entendimento do negócio

- ★ O sucesso está na formulação do problema.
- ★ Empatia é essencial.
- ★ Não há receita de bolo.
- ★ Não há uma única forma de investigar o problema do cliente.

# CRISP-DM - Entendimento do negócio

- “Me diga o que você precisa ...”

- “Me explica o problema ...”

# CRISP-DM - Entendimento do negócio

- Faça entrevistas
- Faça pesquisas
- Faça brainstorm com o cliente
- Faça brainstorm com a sua equipe

# **CRISP-DM - Entendimento do negócio:** Dicas

Premissas:

- Elabore uma lista de premissas, que podem ser modificadas durante o processo de entendimento do negócio.
- Informe as premissas para o cliente desde o início.

**Exemplo que não deve ocorrer:** “Eu não sabia que você estava usando a base do sistema X. Essa base está cheia de problemas.”

# **CRISP-DM - Entendimento do negócio:** Dicas

Riscos envolvidos:

- Faça uma análise de riscos do projeto.
- Exemplos de riscos
  - Risco de trabalhar com dados de forma incorreta;
  - Risco de entregar uma solução complexa para a área de negócio;
  - Risco de não conformidade com a LGPD.
  - Risco da infraestrutura não suporte a solução.
- VARELA (2022) apresenta um estudo sobre análise de riscos em projetos de ciência de dados.

# Grupos de Riscos

Grupo: **Escopo Funcional e de Projeto**

- Projetos iniciados com perguntas erradas ou foco inadequado;
- Falta de documentação técnica.

Grupo: **Gestão de Projeto**

- Falta de alinhamento entre equipe e cliente e comunicação ineficaz das necessidades e limitações do projeto.

Grupo: **Gestão Operacional**

- Concentração no desempenho do modelo e não na usabilidade.
- Falta de conhecimento das práticas de segmentação dos dados.

Grupo: **Tecnológicos e de Qualidade:**

- Escolha inadequada de tecnologias, incompatibilidade de sistemas e problemas de migração.
- Riscos: falta de capacidade de processamento de dados.

Grupo: **Recursos e Infraestrutura**

- Riscos organizacionais como falta de apoio estrutural dos sistemas de informação.

Grupo: **Legais e de Segurança**

- Questões relacionadas à privacidade e confidencialidade dos dados.

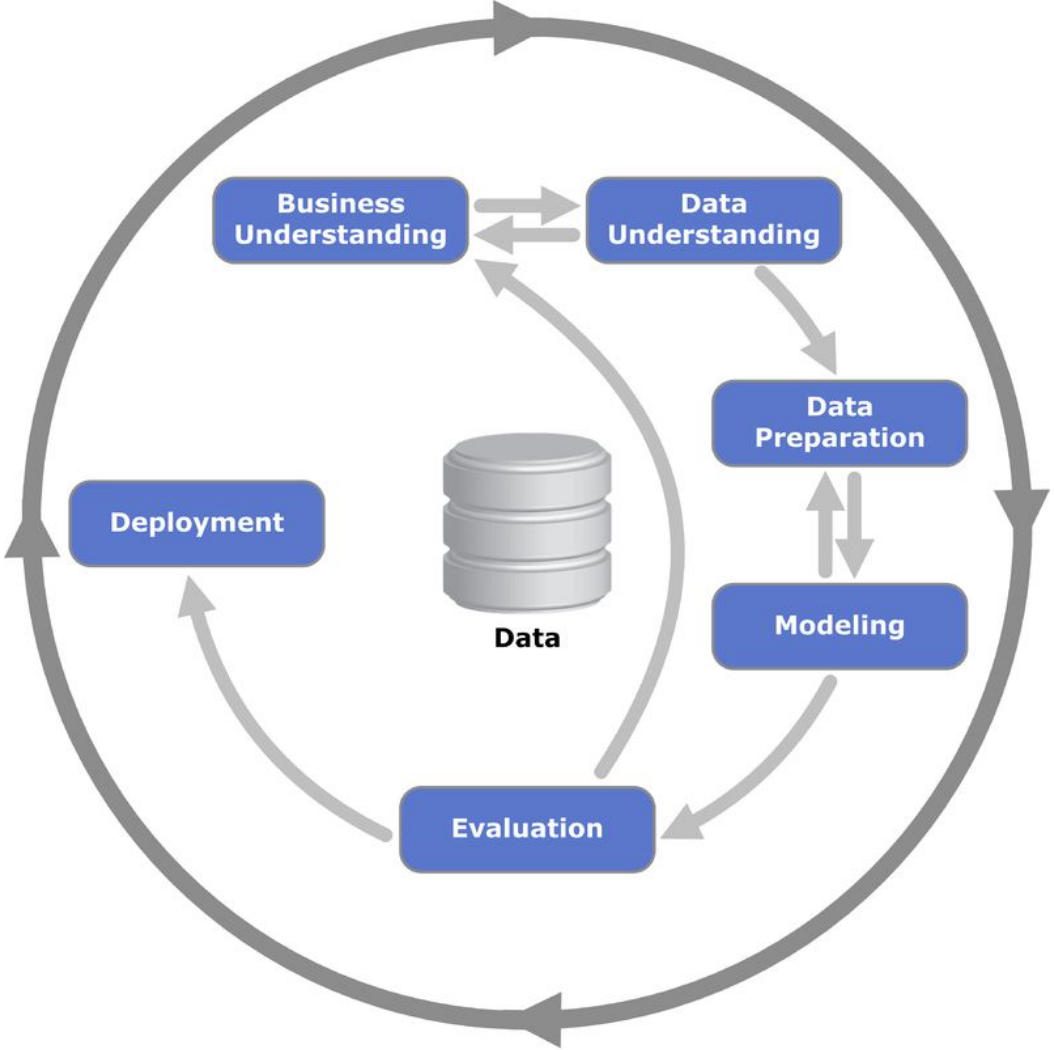
✓ Conformidade com a LGPD

# **CRISP-DM - Entendimento do negócio:** Dicas

Critérios de sucesso:

- Defina os critérios de sucesso a partir dos objetivos e premissas.
- Ele funciona como um tipo de “critério de parada”. Ou seja, quando chegarmos a esse objetivo podemos estar satisfeitos com o projeto e encerrá-lo.
- **Exemplos**:
    - **Redução esperada de X % nos custos de manutenção das máquinas.**
    - **Identificação e sugestão de correção de X% dos dados na base.**
    - **Mitigar os X riscos levantados pela área de regulação da empresa.**

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

# CRISP-DM - Entendimento dos dados

**O que fazer?**

- Identificar estrutura dos dados;
- Avaliar qualidade e disponibilidade dos dados;
- Entender a relação dos dados com os objetivos do projeto.

**Como fazer?**

- Aplicar técnicas de análise exploratória dos dados (análise descritiva);
- Aplicar técnicas para verificar a qualidade dos dados;
- Explorar os dados com o intuito de analisar a relação com os objetivos do projeto.

# CRISP-DM - Preparação dos dados

**O que fazer?**

- Escolher a granularidade temporal ou espacial dos dados;
- Corrigir inconsistências nos dados;
- Definir as técnicas de pré-processamento e formatação dos dados.

**Como fazer?**

- Aplicar técnicas de seleção de features dos dados;
- Aplicar técnicas de pré-processamento dos dados;
- Aplicar técnicas de formatação dos dados.

# CRISP-DM - Modelagem

**O que fazer?**

- Revisão de técnicas para atender aos objetivos do projeto;
- Design (desenho) do modelo;
- Estabelecer métricas para construção e avaliação do modelo.

**Como fazer?**

- Aplicar técnicas para criação do modelo;
- Construir e testar o modelo.

# CRISP-DM - Avaliação e Implantação

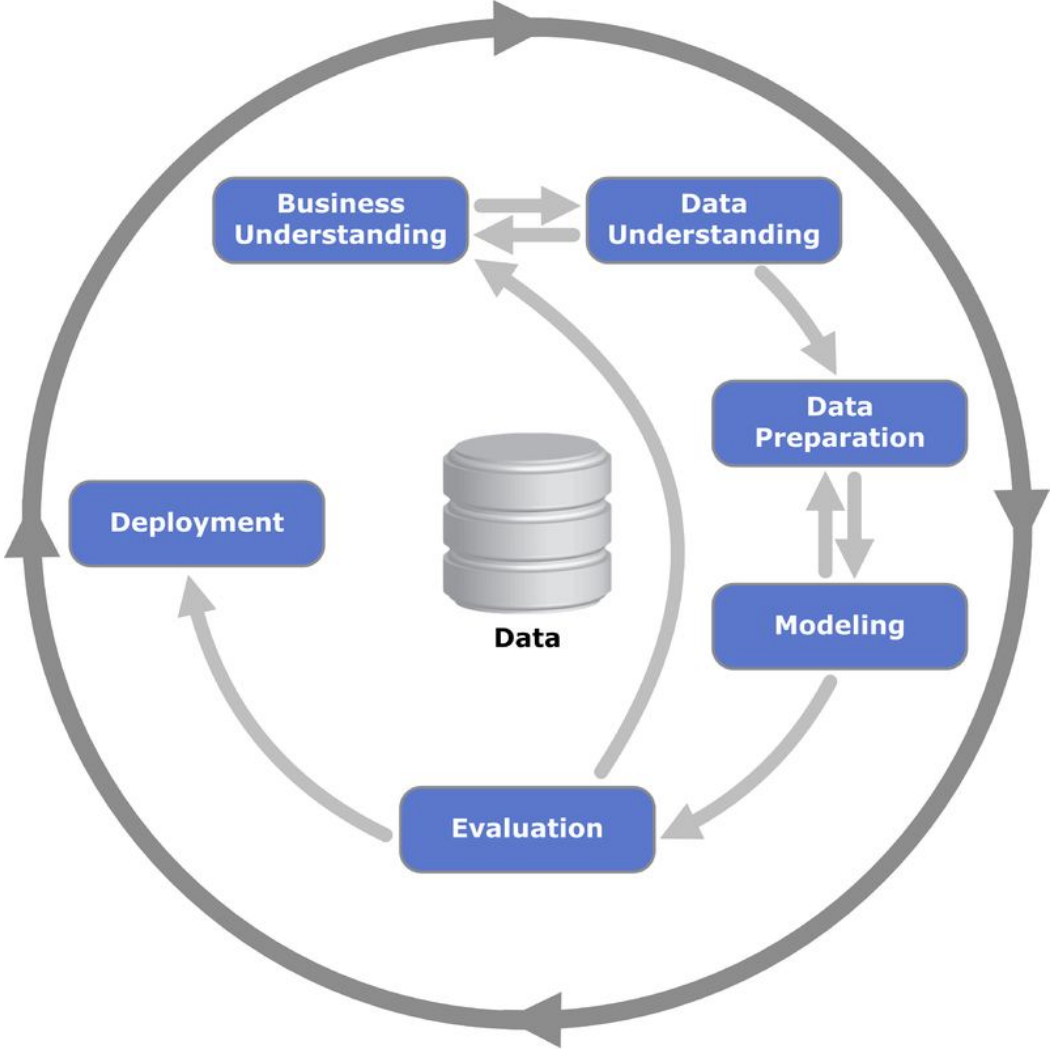
**Avaliação:**

- Interpretação dos resultados;
- Impacto dos resultados com base nos critérios de sucesso do projeto;
- Revisão do modelo.

**Implantação:**

- Elaboração da documentação para implantação e manutenção da solução;
- Implantação e validação;
- Monitoramento e manutenção da solução.

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

# KDD
# (Knowledge Discovery in Databases)

UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

imetrópole DIGITAL

# KDD (Knowledge Discovery in Databases)

- KDD (Descoberta de Conhecimento em Bases de Dados) proposto por **Fayyad et al (1996)** (na Microsoft).
- KDD é descrito como metodologia, já que envolve uma série de passos para a descoberta de conhecimento útil a partir de grandes volumes de dados.
- Etapas no KDD:
  - Entendimento do negócio.
  - Seleção dos dados;
  - Pré-processamento dos dados;
  - Transformação dos dados;
  - Mineração dos dados:
    - Seleção da função de mineração;
    - Seleção do algoritmo de mineração;
  - Interpretação/Avaliação

# KDD (Knowledge Discovery in Databases)

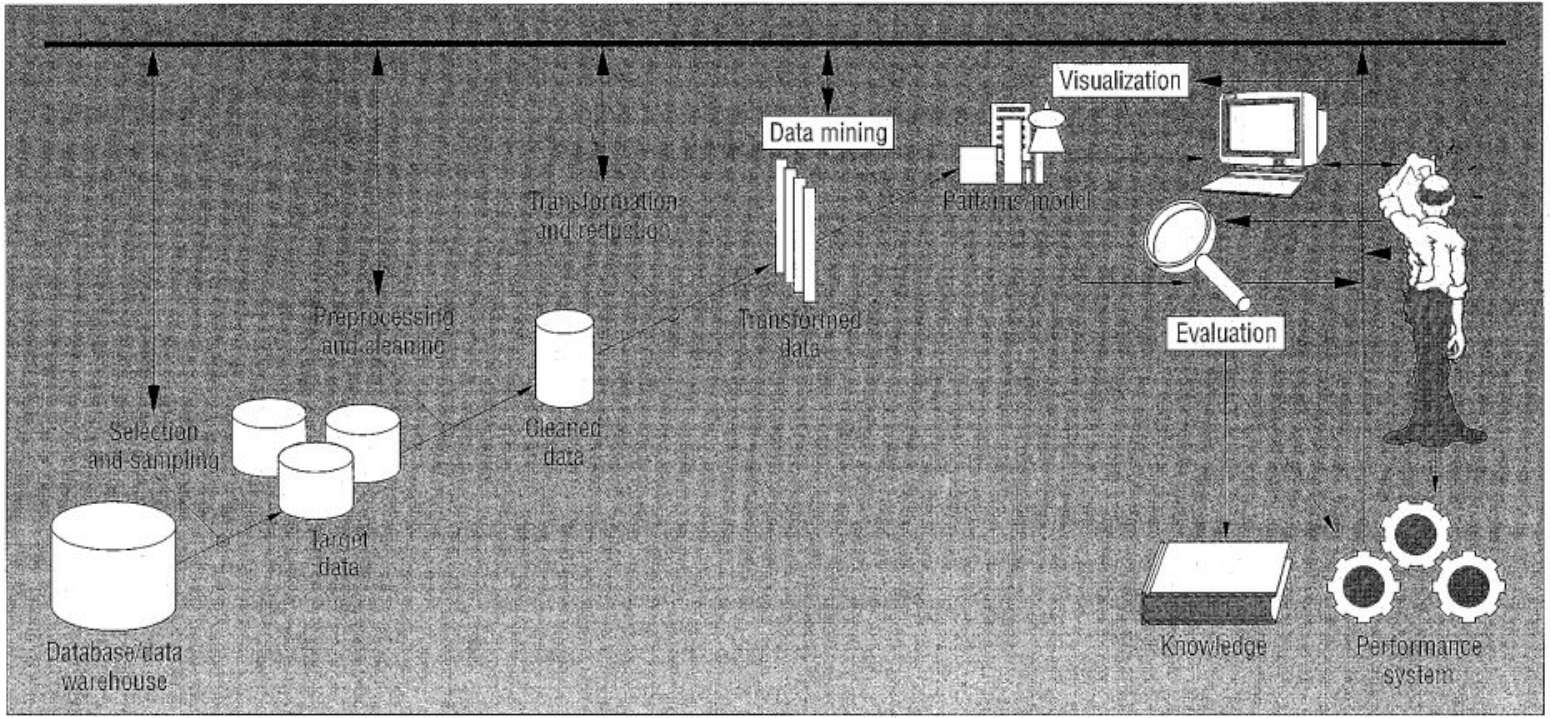
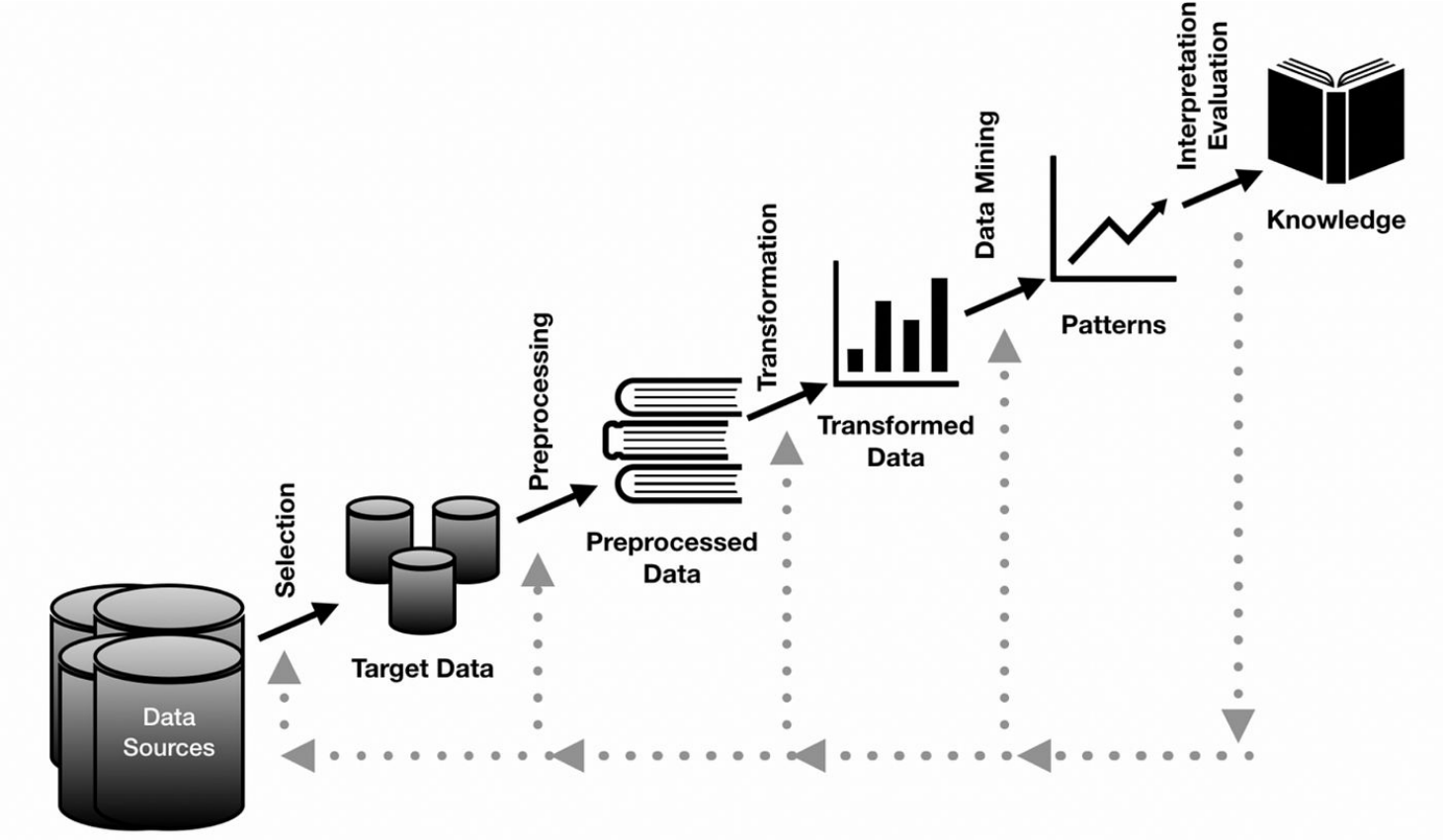


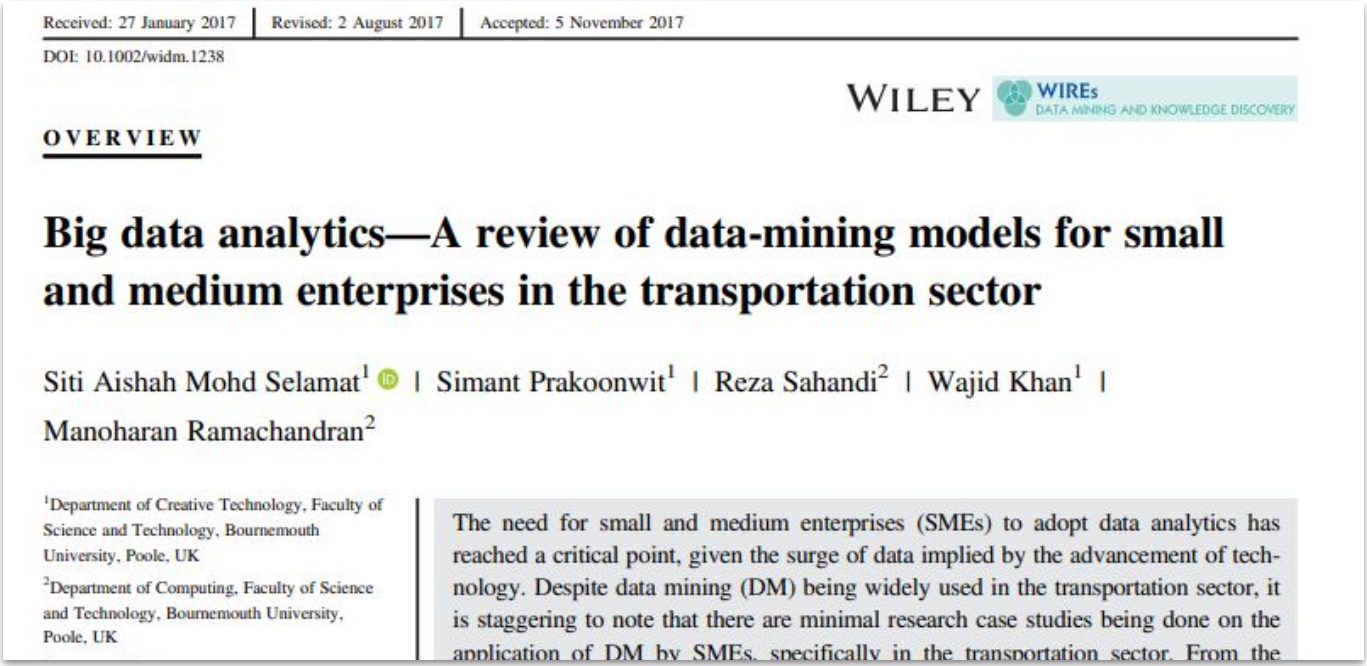

Figure 1. An overview of the KDD process.[2] (For simplicity, the illustration omits arrows indicating the multitude of potential loops and iterations.)

# KDD (Knowledge Discovery in Databases)



# Comparação entre KDD e CRISP-DM

- MOHD SELAMAT (2018) compara alguns modelos de processos KDD, CRISP-DM e SEMMA:



Received: 27 January 2017 | Revised: 2 August 2017 | Accepted: 5 November 2017

DOI: 10.1002/widm.1238

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

OVERVIEW

## Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector

Siti Aishah Mohd Selamat[1] | Simant Prakoonwit[1] | Reza Sahandi[2] | Wajid Khan[1] | Manoharan Ramachandran[2]

[1]Department of Creative Technology, Faculty of Science and Technology, Bournemouth University, Poole, UK

[2]Department of Computing, Faculty of Science and Technology, Bournemouth University, Poole, UK

The need for small and medium enterprises (SMEs) to adopt data analytics has reached a critical point, given the surge of data implied by the advancement of technology. Despite data mining (DM) being widely used in the transportation sector, it is staggering to note that there are minimal research case studies being done on the application of DM by SMEs, specifically in the transportation sector. From the

# SEMMA (Sample, Explore, Modify, Model and Assess)

- Amostrar, Explorar, Modificar, Modelar e Avaliar
- ano: 1997


**FIGURE 3** Sample, Explore, Modify, Model, Assess methodology. The figure represents the overall SEMMA methodology process. Source: Mariscal et al. (2010)

# Comparação entre KDD e CRISP-DM

**TABLE 4** Model comparison—KDD, CRISP-DM, and SEMMA

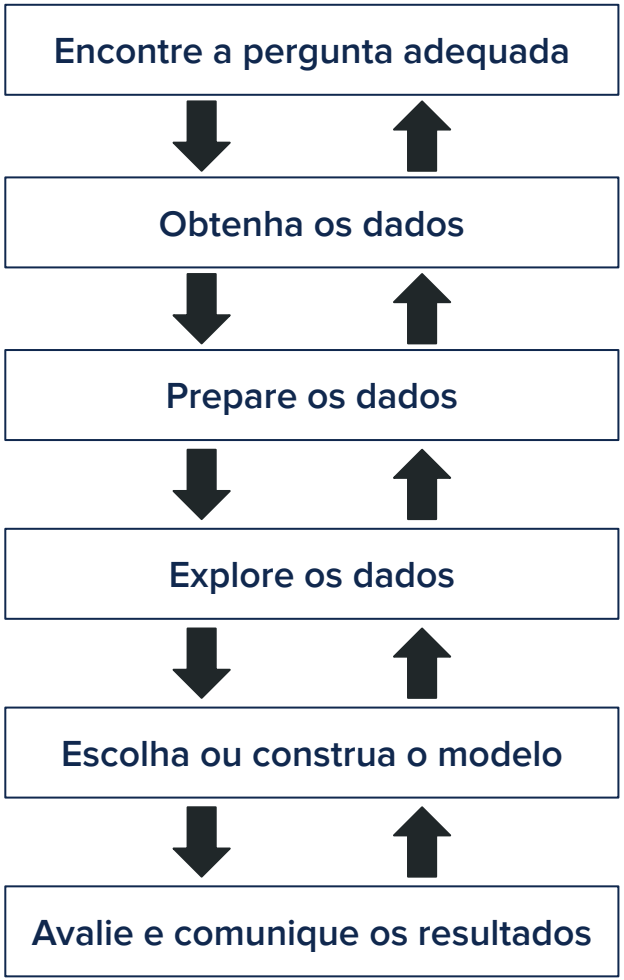
| MODEL | KDD | CRISP-DM | SEMMA |
|---|---|---|---|
| Developed by | Fayyad et al. | CRISP-DM Consortium | SAS Institute |
| Year of model introduced | 1996 | 1996 (officially released in 2000) | 1997 |
| Functions | 1. Uncover new and unique insights<br>2. Generate predictions | | |
| Total steps | 9 | 6 | 5 |
| Phase | 1. Application domain understanding | 1. Business understanding | - |
| | 2. Creating a target data set | 2. Data understanding | 1. Sample |
| | 3. Data cleaning and | | 2. Explore |
| | 4. Data transformation | 3. Data preparation | 3. Modify |
| | 5. Data-mining method selection | 4. Modeling | 4. Model |
| | 6. Data-mining algorithm selection | | |
| | 7. Data-mining application | | |
| | 8. Discovered patterns interpretation | 5. Evaluation | 5. Assessment |
| | 9. Using discovered knowledge | 6. Deployment | - |

# Comparação entre KDD e CRISP-DM

**TABLE 4** Model comparison—KDD, CRISP-DM, and SEMMA

| MODEL | KDD | CRISP-DM | SEMMA |
|---|---|---|---|
| Requires background knowledge in DM | Yes | Yes | Yes |
| Software tool support | Yes | Yes | Yes |
| | Mineset | SPSS Clementine | SAS |
| Documentation | No | Yes | Yes |
| Open-source tool support | Yes | Yes | No |
| Total case studies in the transportation sector count | 2 | 6 | 2 |
| Total case studies in the SME context count | 3 | 6 | 1 |
| Overall case studies count | 5 | 12 | 3 |
| Application areas | Aviation, rail, tourism, financial, manufacturing | Logistic, cargo, aviation, rail, public transport, software, financial, marketing and sales, trading | Software, public transport, street taxis |
| Kdnuggets poll results for 2007 (200 votes total) Kdnuggets (2014) | 7.3% | 42% | 13% |
| Kdnuggets poll results for 2014 (200 votes total) Kdnuggets (2014) | 7.5% | 43% | 8.5% |

# Outros pipelines e adaptações




- O que deve ser aprimorado?
- Qual decisão será tomada?
- Que benefício isso traz para o cliente?

- Esse é realmente o dado que preciso?
- Esses dados são suficientes?
- Existem problemas de privacidade?

- Está no formato que preciso?

- Existem padrões?
- Existem anomalias?

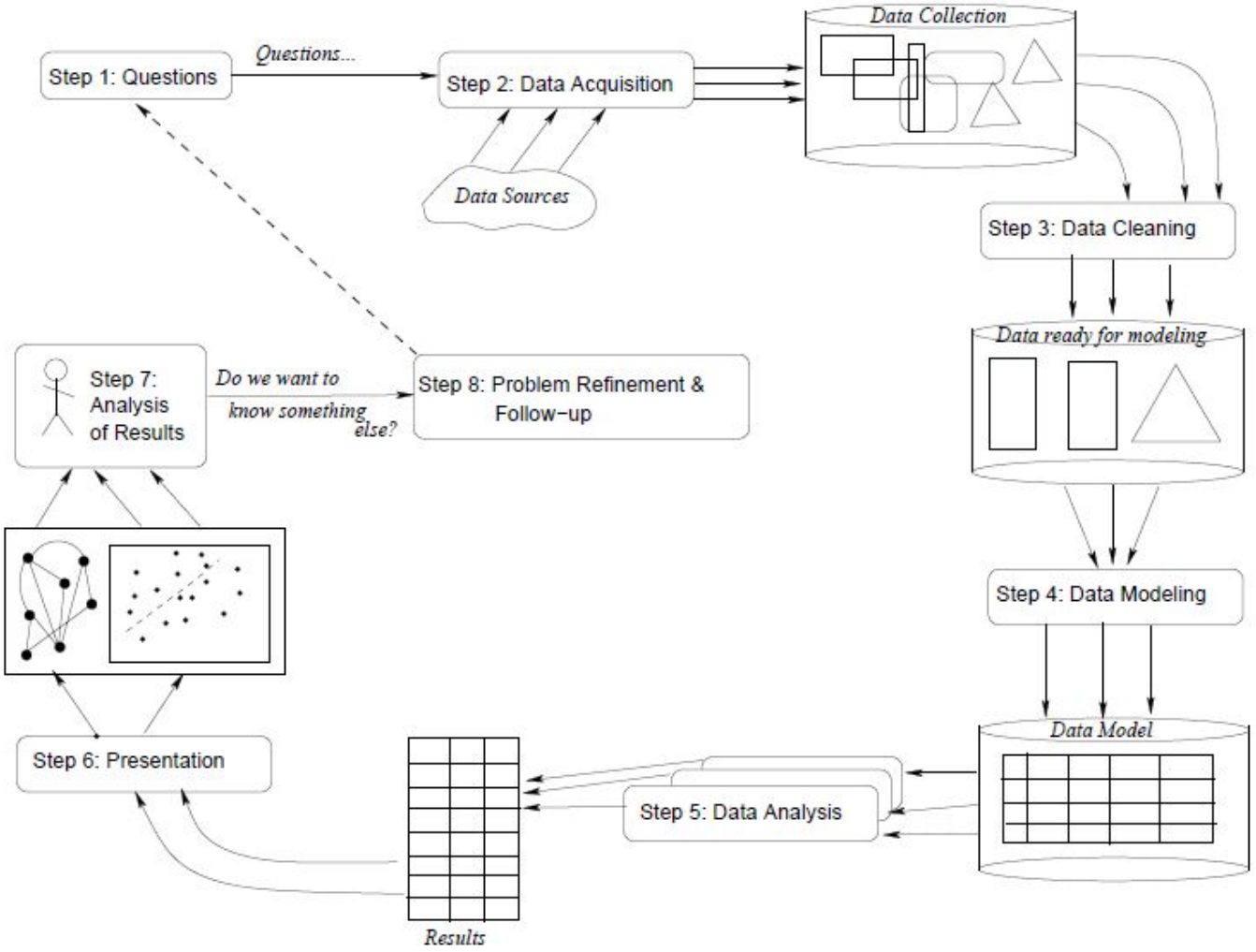
- Os modelos existentes são suficientes?
- Quais são os melhores modelos para esse problema?

- Os resultados estão realmente corretos?
- Os resultados permitem a tomada de decisão?
- Como interpretar os dados?

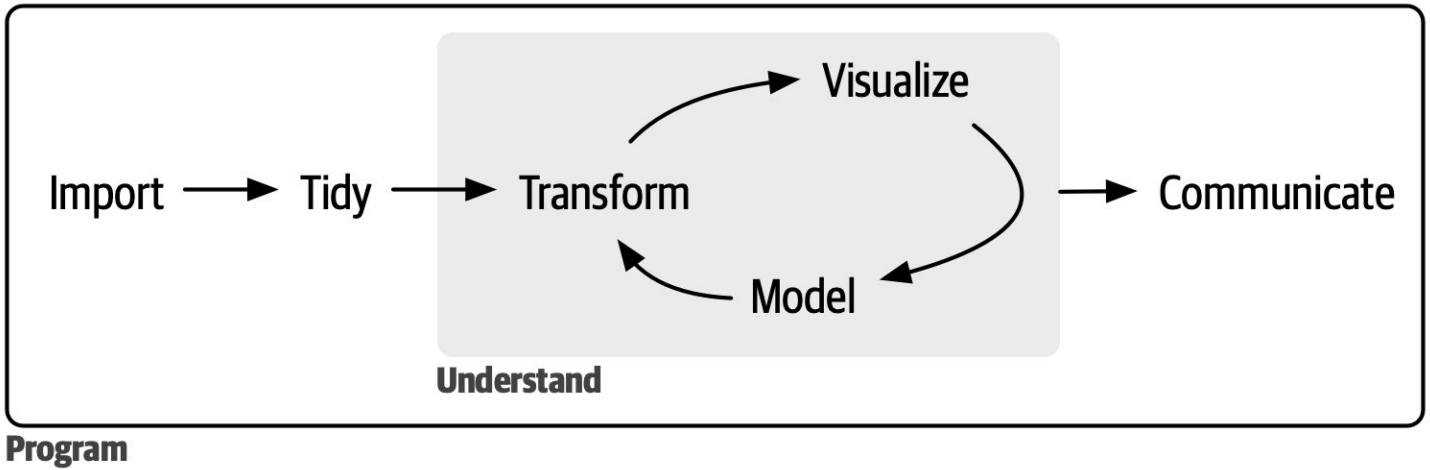
Adaptado de Maurício Seiji Rezende (2022)

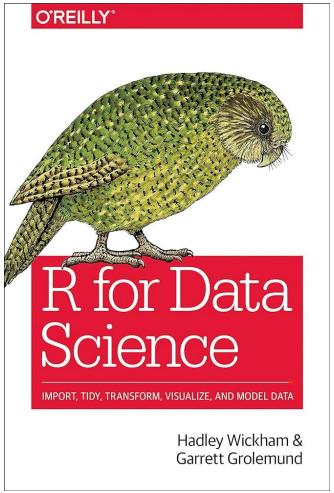
# Processo de Ciência de Dados como um Ciclo

por Alexander Dekhtyar



# **Etapas de Ciência de Dados** por GROLEMUND and WICKHAM (2023)

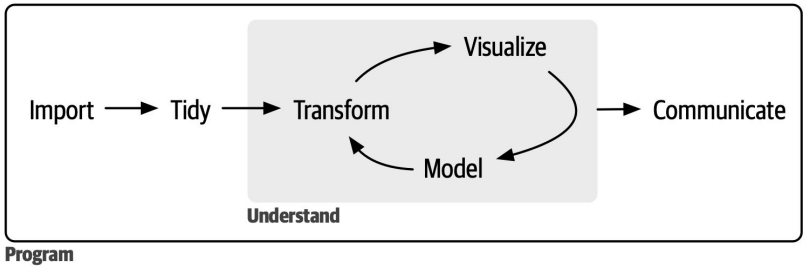



Fonte: GROLEMUND and WICKHAM (2023).

# **Etapas de Ciência de Dados** por GROLEMUND and WICKHAM (2023)

- ***import* (importar)**: coletar os dados das estruturas de armazenamento dos dados.
- ***tidy* (limpar)**: organizar os dados de uma forma consistente que corresponda à semântica do conjunto de dados, i.e., organizar as variáveis em colunas e as observações em linhas.
- ***transform* (transformar)**: selecionar dados de interesse, criar novas variáveis e calcular dados de estatísticos do conjunto de dados.
- ***visualize* (visualizar)**: utilizar técnicas de visualização para levantar novas questões sobre os dados e sugerir respostas aos problemas especificados no início do projeto.
- ***model* (modelar)**: implementar modelos, ferramentas matemáticas e computacionais, para fazer suposições com base nos dados.
- ***communicate* (comunicar)**.

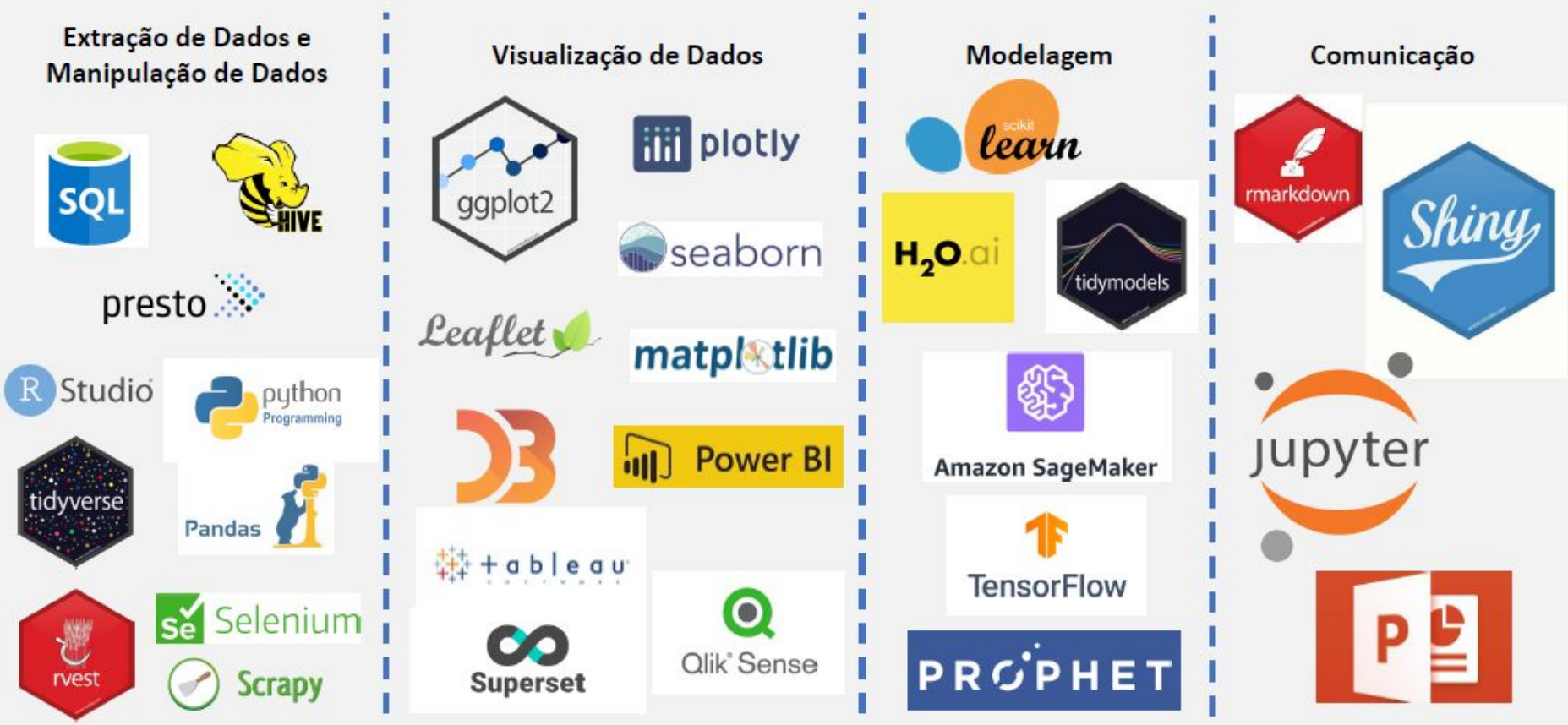


# Modelos de Processos (pipeline) de Ciência de Dados

**Dúvidas?**

- CRISP-DM
- KDD
- SEMMA
- Adaptações

# Ferramentas para os **Cientistas de Dados**


Extração de Dados e Manipulação de Dados
SQL
HIVE
presto
R Studio
python Programming
tidyverse
Pandas
rvest
Selenium
Scrapy
Visualização de Dados
ggplot2
plotly
seaborn
Leaflet
matplotlib
D3
Power BI
tableau
Superset
Qlik Sense
Modelagem
scikit learn
H2O.ai
tidymodels
Amazon SageMaker
TensorFlow
PROPHET
Comunicação
rmarkdown
Shiny
jupyter

Fonte:

**Cuide** dos dados,
**Entenda** o negócio,
**Conheça** as técnicas,
**Aplique** as ferramentas/modelos e
**Comunique-se** bem!

# O que vimos hoje?

Aula: Introdução à Ciência de Dados: Pipeline e Modelos

IMD1151 - Ciência de Dados

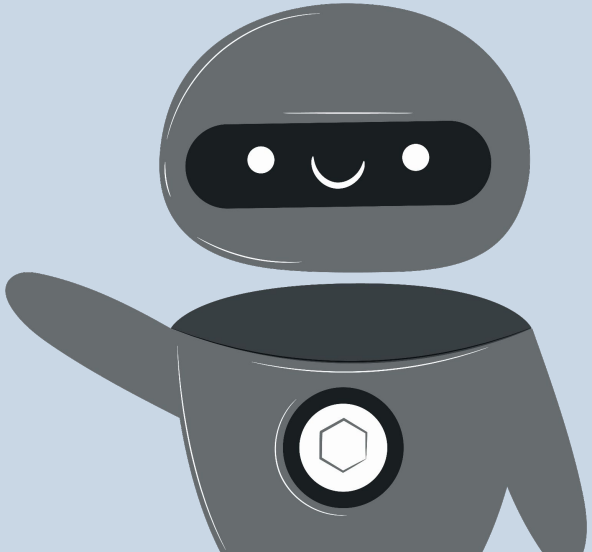
Prof. Heitor Florencio

- Modelos de Processos de Ciência de Dados
- CRISP-DM
- KDD
- Comparação CRISP-DM e KDD
- Outros modelos e adaptações

# Dúvidas?

**Prof. Heitor Florencio**
Sala 103 - nPITI/IMD
heitorm@imd.ufrn.br

**Prof. Daniel Sabino**
Sala A226 - CIVT/IMD
daniel@imd.ufrn.br



## Referências


- AMARAL, Fernando. **Introdução à ciência de dados: mineração de dados e big data.** Alta Books Editora, 2016.
- GROLEMUND, Garrett; WICKHAM, Hadley. **R for data science**. O'Reilly Media, Inc., 2nd edition, 2023. Disponível em: https://r4ds.hadley.nz/ . Acesso em 11 de março de 2024.
- JIAWEI, Han; MICHELINE, Kamber. **Data mining: concepts and techniques.** Morgan kaufmann, 2006.
- FEYYAD, U. M. **Data mining and knowledge discovery:** Making sense out of data. IEEE expert, v. 11, n. 5, p. 20-25, 1996.
- VARELA, Cristina; DOMINGUES, Luísa. **Risks of Data Science Projects-A Delphi Study**. Procedia Computer Science, v. 196, p. 982-989, 2022.
- SOUZA, Vinícius. **Entenda o CRISP-DM**, suas etapas e como de fato gerar valor com essa metodologia. PREDITIVA.AI. Disponível em: https://www.preditiva.ai/blog/entenda-o-crisp-dm-suas-etapas-e-como-de-fato-gerar-valor-com-essa-metodologia. Acesso em: 19 mar. 2024.

## Referências


- CRISP-DM. In: WIKIPEDIA, a enciclopédia livre. Disponível em: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining . Acesso em: 19 mar. 2024.
- SILVEIRA, Juliano Gomes. **Pré-processamento de Dados**. 2022. Notas de aula.
- FREITAS, Rômulo. **Fundamentos de Estatística para Ciência de Dados**. 2022. Notas de aula.
- MOHD SELAMAT, Siti Aishah et al. **Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector.** Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 8, n. 3, p. e1238, 2018.
- DEKHTYAR, Alexander. **Introduction to Data Science**. 2023. Notas de aula. Disponível em: https://users.csc.calpoly.edu/~dekhtyar/ . Acesso em: 11 mar. 2024.
- VAZQUEZ, Favio. **The Roots of Data Science**. Towards Data Science, 2020. Disponível em: https://towardsdatascience.com/the-roots-of-data-science-77c71115229 . Acesso em: 19 mar. 2024.