



# Python Data Analysis Mastery

Prof. Nicksson Freitas



# Aula 04. Como Desenvolver um projeto de Data Science

Prof. Nicksson Freitas

# Objetivos

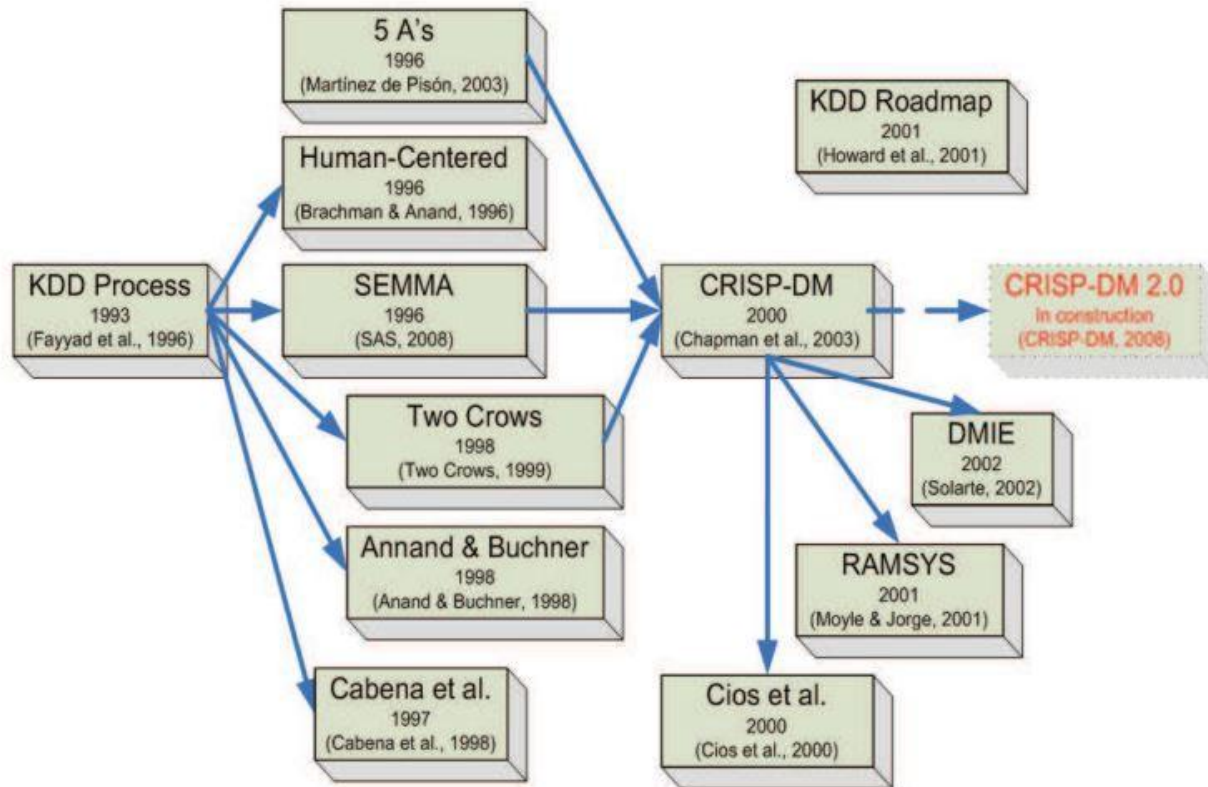
- Introduzir as metodologias para desenvolvimento de projetos de data science
- Conhecer a metodologia CRISP-DM

## Resumo

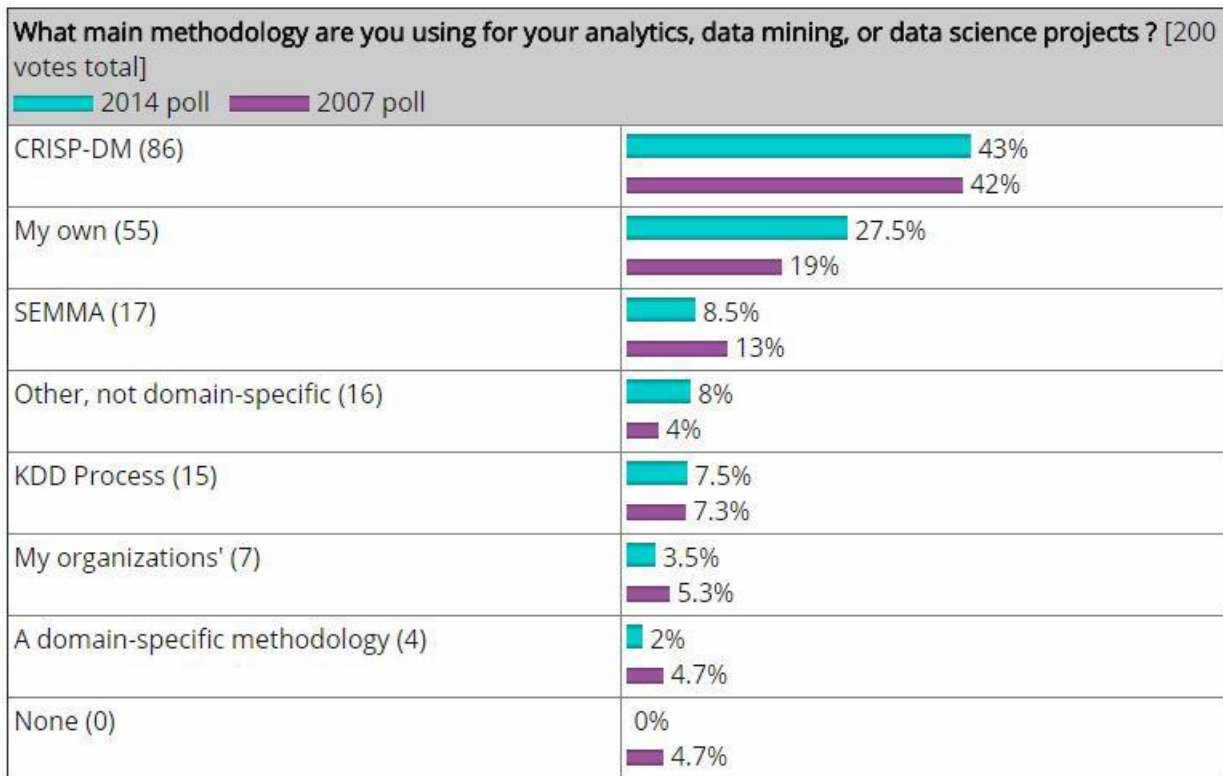
1. Metodologias para projetos de dados
2. CRISP-DM

# **1. Metodologia para projeto de Dados**

# Metodologias para projeto de dados



# Metodologias para projeto de dados

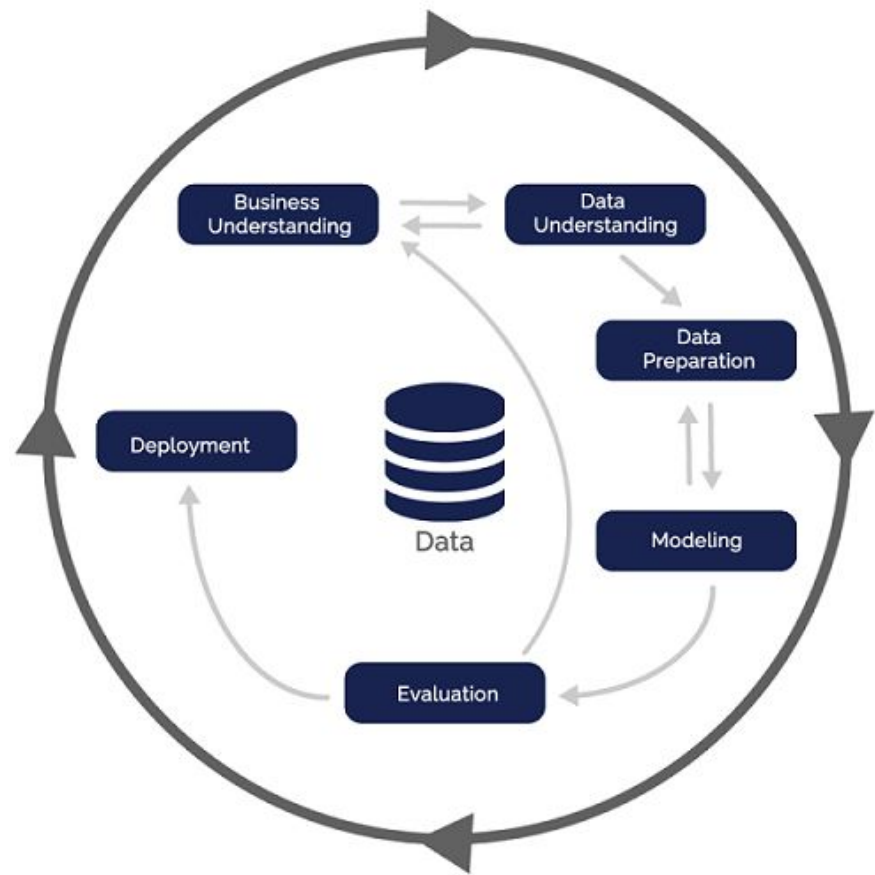


## 2. CRISP-DM



# CRISP-DM

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- Uma metodologia publicada em 1999 para padronizar os processos de mineração de dados em todos os setores da indústria
- Possui seis fases



# Visão geral do CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<b>Determine Business Objectives</b> Background Business Objectives Business Success Criteria	<b>Collect Initial Data</b> Initial Data Collection Report  <b>Describe Data</b> Data Description Report	Data Set Data Set Description  <b>Select Data</b> Rationale for Inclusion / Exclusion	<b>Select Modeling Technique</b> Modeling Technique Modeling Assumptions  <b>Generate Test Design</b> Test Design	<b>Evaluate Results</b> Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models	<b>Plan Deployment</b> Deployment Plan  <b>Plan Monitoring and Maintenance</b> Monitoring and Maintenance Plan
<b>Assess Situation</b> Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits	<b>Explore Data</b> Data Exploration Report  <b>Verify Data Quality</b> Data Quality Report	<b>Clean Data</b> Data Cleaning Report  <b>Construct Data</b> Derived Attributes Generated Records	<b>Build Model</b> Parameter Settings Models Model Description	<b>Review Process</b> Review of Process  <b>Determine Next Steps</b> List of Possible Actions Decision	<b>Produce Final Report</b> Final Report Final Presentation
<b>Determine Data Mining Goals</b> Data Mining Goals Data Mining Success Criteria		<b>Integrate Data</b> Merged Data  <b>Format Data</b> Reformatted Data	<b>Assess Model</b> Model Assessment Revised Parameter Settings		<b>Review Project</b> Experience Documentation
<b>Produce Project Plan</b> Project Plan Initial Assessment of Tools and Techniques					

# Entendimento do negócio

- Tarefas importantes
  - Estruturação da equipe de desenvolvimento
  - Modelagem do problema
    - Reuniões com especialistas
  - Definição das ferramentas e KPIs
  - Gestão dos riscos
  - Definição dos prazos
  - Finalizar o checklist do negócio

# Entendimento do negócio

- Identificar o tema a ser abordado e fazer um levantamento bibliográfico
- Perguntas importantes:
  - Qual o problema que vamos resolver?
  - Por que ele é importante?
  - Existe uma solução atual para o problema?
  - Dada uma solução hipotética para o problema afetaria a empresa?
  - Qual o tempo estimado para desenvolver a solução?
  - Como vamos medir a solução ou quais são os KPIs?

# Entendimento dos dados

- Coletar dados e extrair informações
- Perguntas importantes:
  - Todos os dados estão disponíveis?
  - Como os dados são armazenados?
  - Quais sistemas ou quem produzem os dados?
  - Quais os tipos de dados disponíveis?
  - Qual a dificuldade para extrair os dados?
  - Existem custos dos dados? Comprar? Gerar?
  - Existem dados sensíveis (LGPD)?
  - Existe uma variável alvo no problema em questão (por exemplo, supervisionado ou não supervisionado)?

# Preparação dos dados

- Tratar os dados e realizar mineração de dados;
- Questões importantes
  - Como os dados serão processados?
  - Qual pipeline ideal para os dados?
  - Como iremos lidar com dados ausentes?
- O que será feito?
  - Analisar dados nulos
  - Analisar dados duplicados
  - Analisar dados inconsistentes

# Modelagem de Dados

- Definição das técnicas a serem usadas
  - É comum usar mais de um modelo para medir seu desempenho e performance computacional
- Questões importantes
  - Quais algoritmos podem resolver meu problema?
  - Quanto tempo durará o treinamento dos algoritmos?
- O que será feito?
  - Divisão dos dados para treino e teste
  - Seleção dos algoritmos
  - Treinamento
  - Otimização de parâmetros dos algoritmos

# Avaliação

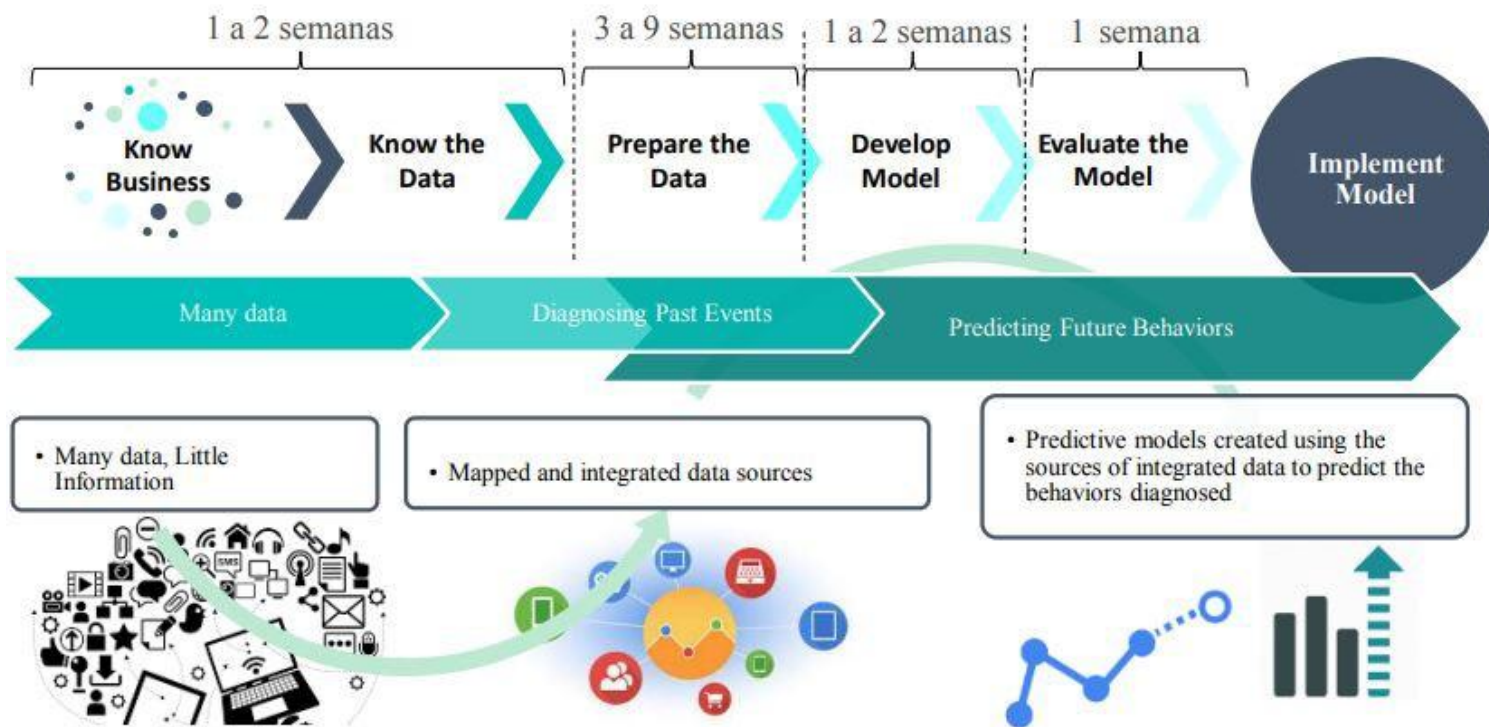
- Analisar os resultados junto com um especialista;
- Se os resultados forem bons?
  - O modelo pode ser enviado para produção e implementado em um protótipo.
- Se os resultados forem ruins?
  - Volta a etapa de entendimento do negócio



# Implantação

- O modelo enviado para produção e implementado em um protótipo
- É necessário acompanhar a performance do modelo em um período alinhado com o especialista, caso haja algum problema ou oportunidade de melhoria;

# Qual a duração de cada fase?



## Curiosidades

- ML project requires lots of data
- **ML requires clean data.**
- More than 50% of project time is spent gathering, cleansing, and visualizing data.
- 36% see dirty data as the #1 challenge.

# Considerações Finais

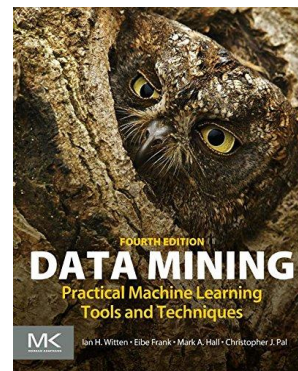
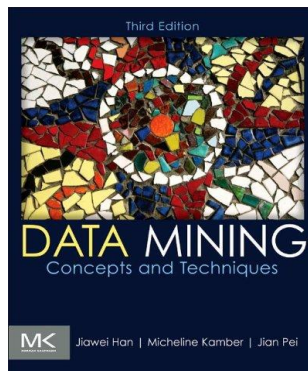
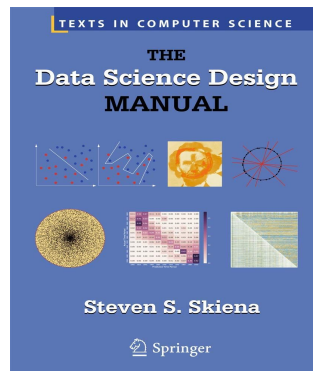
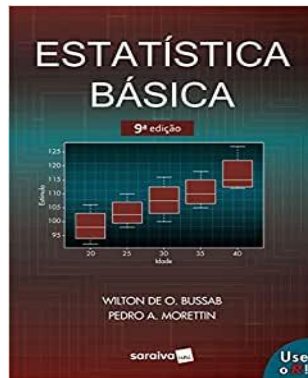
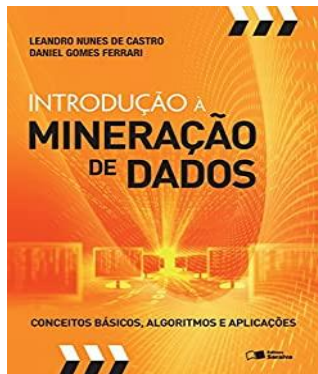
# Considerações finais

- Você conheceu diversas metodologias para desenvolver projetos de Ciência de Dados
- Você aprendeu como usar a metodologia CRISP-DM em um projeto de Ciência de Dados.

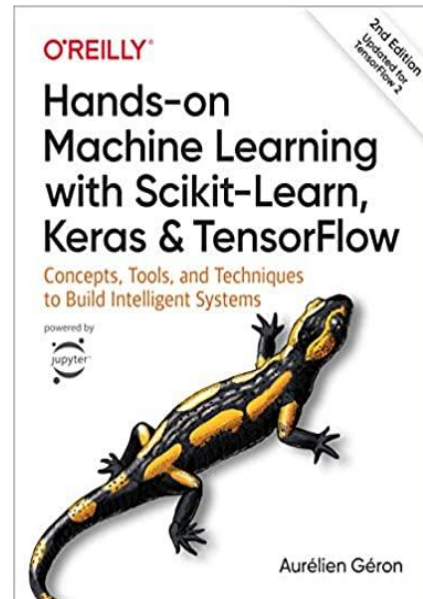
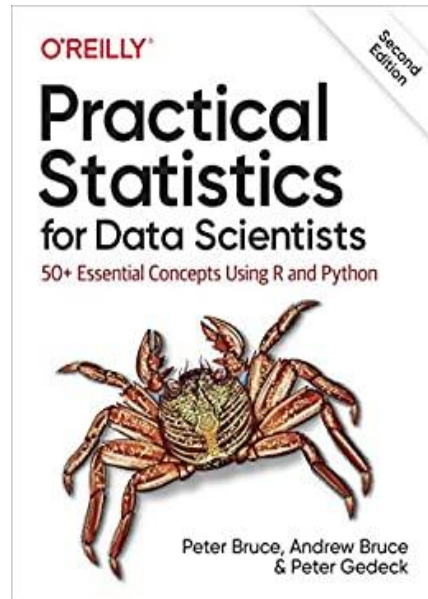
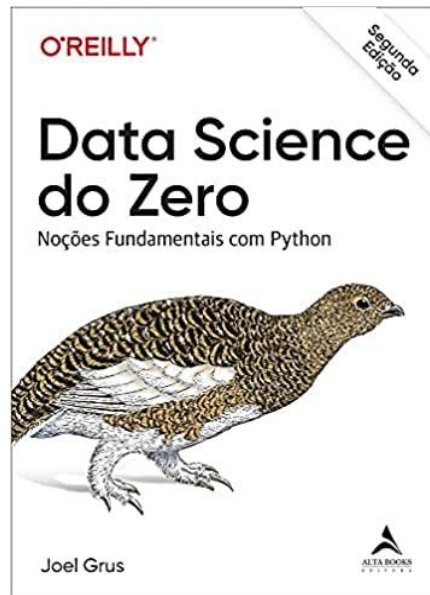
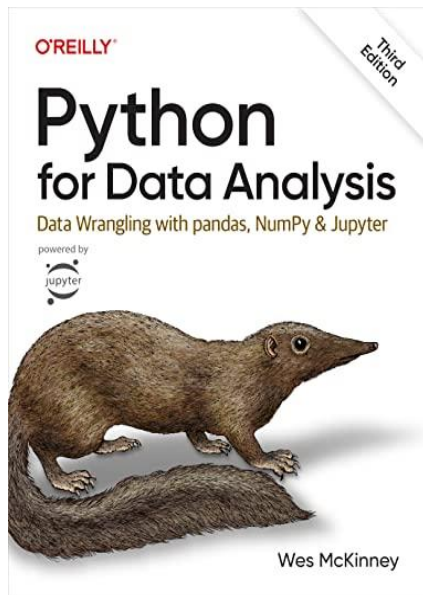
# Referências



# Bibliografia fundamental



# Bibliografia técnica





**OBRIGADO**

