



**INFODADOS**  
TRANSFORMANDO DADOS EM NEGÓCIOS

# Prevendo Valores de Automóveis

*Do web scraping ao machine learning*

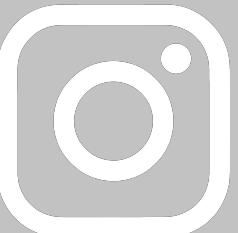
Gabriel L. Gomes  
Maio/2022

@info\_dados

@info\_dados

# APRESENTAÇÃO

- Consultor em Ciência de Dados
- Professor
- Pós-graduação em Deep Learning
- Bacharel em Ciência da Computação
- Projetos de análise de dados/aprendizado de máquina
- Treinamentos em Ciência de Dados/Desenvolvimento
- **Projetos:** financeiro, varejo, setor público P&D, segurança da informação, segurança pública e saúde.
- **Consultorias:** análise de dados(descritivas, ML), formação de equipes, capacitação/treinamentos e afins.



**@info\_dados**



**in/gabriellimagomes/**

## OBJETIVOS

- Entender como funciona a coleta de dados de páginas web (**web scraping**);
- Entender a **metodologia** de projeto de machine learning;
- Entender os principais conceitos sobre machine learning(**Rregressão**);
- Desenvolver códigos para **coleta, tratamento, análises de dados e algoritmos**;
- Entender a **capacidade analítica**;
- Entender como um projeto é desenvolvido no dia a dia;

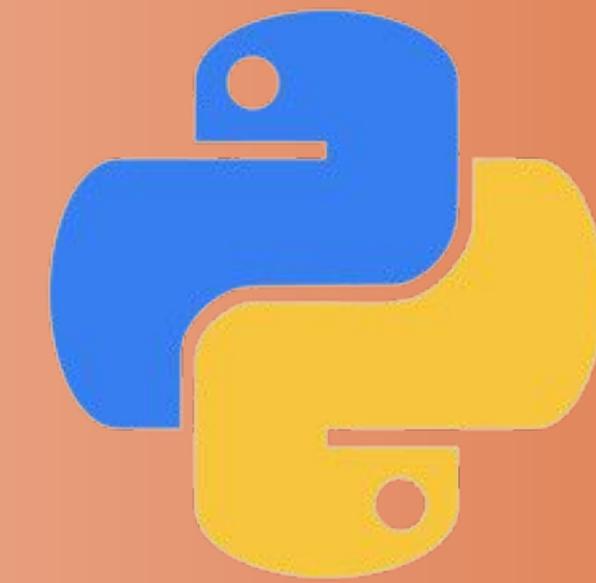
## AO FINAL...

- **Dados disponíveis;**
- **Códigos Disponíveis;**
- **Materiais teóricos disponíveis;**
- Capacidade de **desenvolver o próprio script** para coleta, análise e algoritmo;
- Inserir projeto no **portfólio**;
- Saber como documentar seu projeto de forma simples e prática;



# RECURSOS UTILIZADOS

OLX



colab



# WEB SCRAPING

# WEB SCRAPING

- **O que é web scraping?**

- Conhecido como “raspagem da web” ou “extração de dados da web”;
- É processo de coleta dos dados de páginas web de forma automatizada(semi-automatizada);
- Estes dados coletados da internet são armazenados de forma estruturada ou semi-estruturada;
- A coleta dos dados são utilizadas para diferentes objetivos como: monitoramento de preço, coleta de notícias, dados do mercado financeiro, campanhas publicitárias etc;

# WEB SCRAPING

- O que é web scraping?

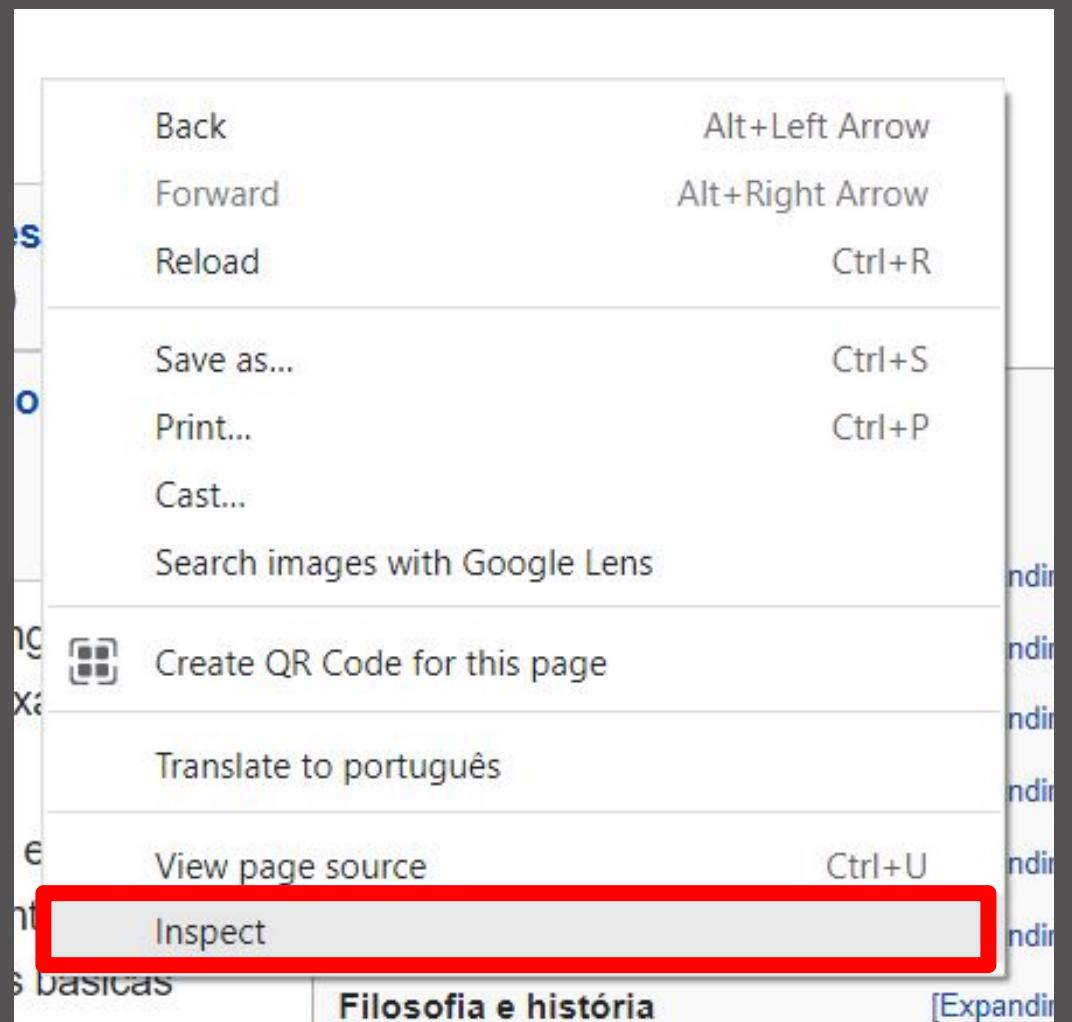


# WEB SCRAPING

- Estrutura da página html
  - O scraping acessa os dados das páginas através dos elementos(seletores) HTML disponíveis;
  - Os mais comuns e fáceis de acessar são pelos atributos: **id** e **class**;

# WEB SCRAPING

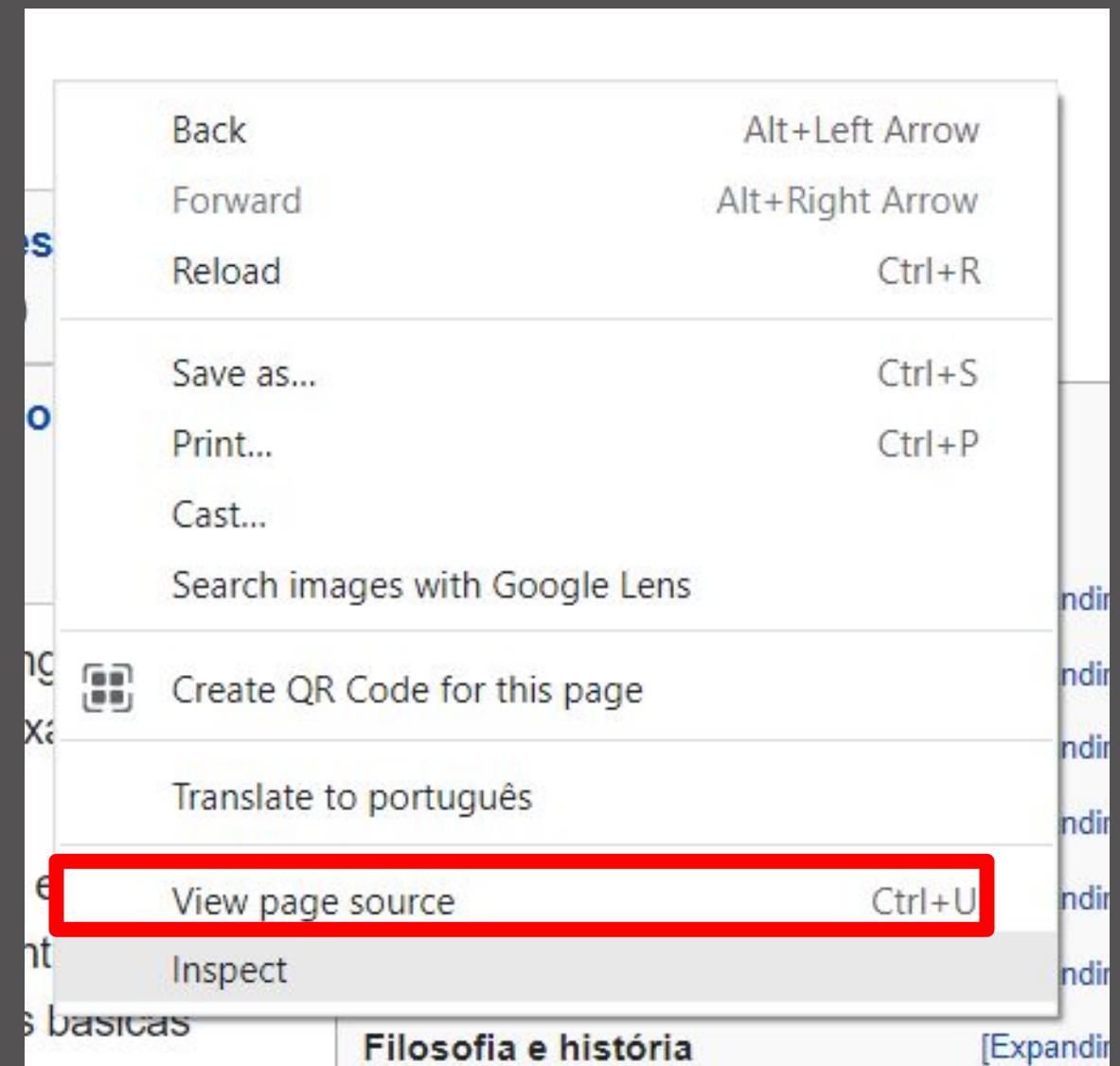
- Acessando código da página



- Você pode acessar o código da página utilizando o **botão direito do mouse** e na opção ***inspect***;

# WEB SCRAPING

- Acessando código da página
  - Em algumas situações é possível acessar em ***View page source***(ver código fonte). Porém, muitos sites hoje em dia não carrega todo o código de uma vez, assim, você não consegue visualizar o código completo nesta opção.



# WEB SCRAPING

- Primeiro é preciso conhecer a estrutura da página, para depois desenvolver o código;
- Acessando código da página
  - [https://pt.wikipedia.org/wiki/Intelig%C3%A3ncia\\_artificial](https://pt.wikipedia.org/wiki/Intelig%C3%A3ncia_artificial)



# WEB SCRAPING

- Acessando código da página (id)

The screenshot shows a web browser displaying the Wikipedia article on Artificial Intelligence. The page title is "Inteligência artificial". The browser's developer tools are open, specifically the "Elements" tab, which displays the HTML structure of the page. The highlighted element is the `<h1 id="firstHeading" class="firstHeading mw-first-heading">Inteligência artificial</h1>` tag, indicating it is the main heading of the page.

```
</div>
</div>
<script>
(function(){var node=document.getElementById("mw-dismissablenotice-anonplace");if(node){node.outerHTML="\u003Cdiv class=\"mw-dismissibleNotice\"\u003E\u003Cdiv c
</script>
</div>
<header class="mw-body-header">
  <nav id="p-lang-btn" class="mw-portlet mw-portlet-lang vector-menu vector-menu-dropdown" aria-labelledby="p-lang-btn-label" role="navigation">...</nav>
  ...
  <h1 id="firstHeading" class="firstHeading mw-first-heading">Inteligência artificial</h1> == $0
  <div class="mw-indicators"> </div>
  <div id="siteSub" class="noprint">Origem: Wikipédia, a enciclopédia livre.</div>
  ::after
</header>
<div id="bodyContent" class="vector-body">
  <div id="contentSub"></div>
  <div id="tocContainer"></div>
  <div id="catlinks" class="catlinks" style="display:none"></div>
</div>
```

# WEB SCRAPING

- Acessando código da página (class)



The screenshot shows a browser window displaying the Wikipedia article on Artificial Intelligence. The page title is "Inteligência artificial". On the left, there's a sidebar with various links like Ajuda, Página de testes, and Portal comunitário. The main content area has a note about the movie "A.I. - Inteligência Artificial". Below the note, there's a warning message: "Este artigo carece de reciclagem de acordo com o livro de estilo. Sinta-se livre para editá-la para que esta possa atingir um nível de qualidade superior. (Junho de 2021)". To the right, there's a sidebar for the category "Ciência" with links to Ciências físicas and Ciências da vida. At the bottom, the browser's developer tools are open, specifically the "Elements" tab, which shows the HTML structure of the page. The "hatnote" class is highlighted in the element tree, corresponding to the note at the top of the page.

# Pacotes para scraping (Python e R)

## PYTHON



Scrapy



Selenium  
Python

BeautifulSoup

## R



rvest

RCrawler

XML2

# Cuidados ao fazer coleta de página web

- Ao fazer a coleta de qualquer página web, precisamos nos atentar a algumas coisas para não ter problema. Algumas dicas:
  - Verificar as **condições de coleta** que o detentor da página impõem;
    - neste caso podemos achar essas informações em ***pagina.com.br/robots.txt*** ou os **termos de serviços**;
  - Cuidado com a **LGPD**;
  - **Não sobrecarregar o servidor** com requisições;

# Desafios do Web Scraping

- Possíveis **quebras do script** de coleta:
  - Qualquer mudança na estrutura da página;
  - Qualquer mudança na URL;
  - Surgimento de novas tecnologias de marcação de páginas web;
- Não é fácil criar testes automatizados para evitar esses possíveis erros;
- É preciso reprogramar o código.



# PROJETO PRÁTICO



<https://olx.com.br/autos-e-peças/carros-vans-e-utilitários?o=1>



# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

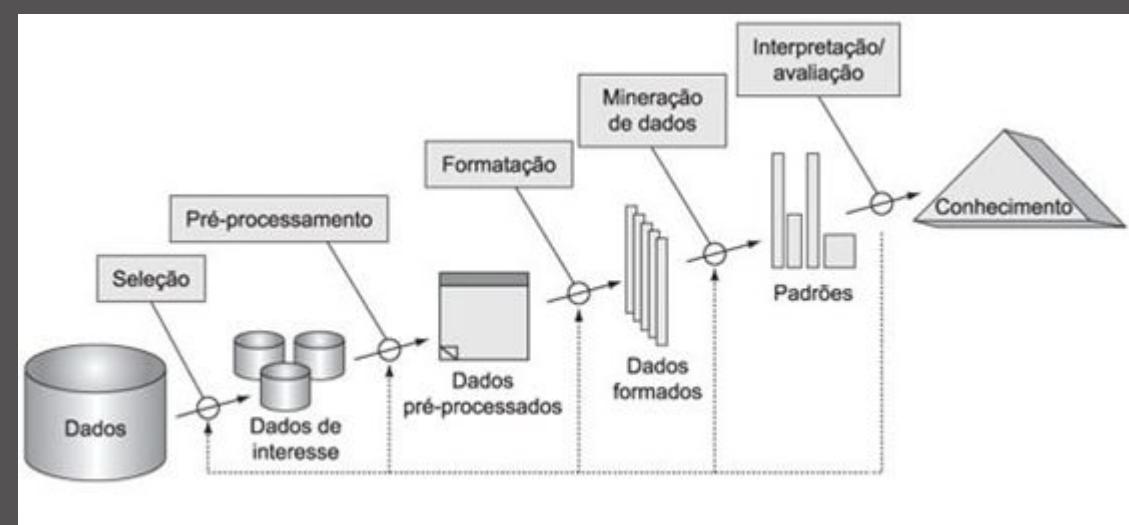
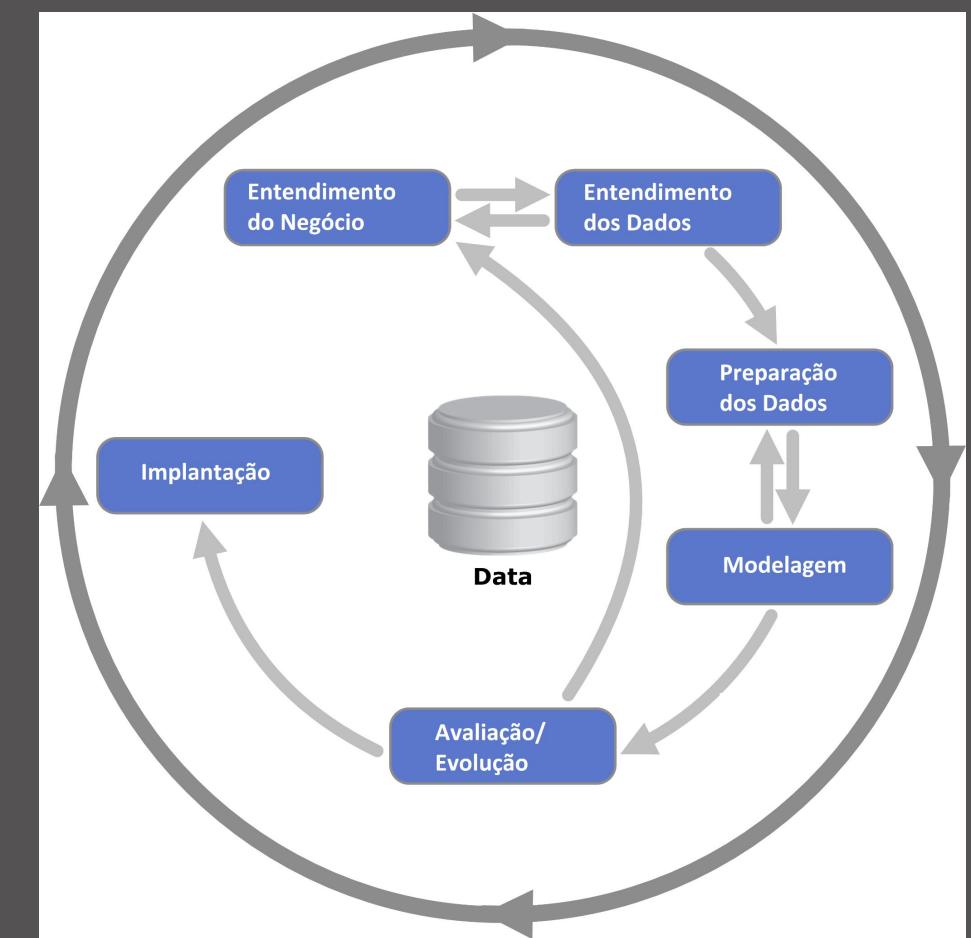


# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- São modelos de processos para descrever abordagens comuns para soluções em ciência de dados;
- Essas metodologias servem para orientar a equipe sobre o status de desenvolvimento;
- Serve também para gerenciar/contabilizar tempo gasto em cada etapa do projeto;
- Possibilita planejar melhor os prazos de entrega(s);

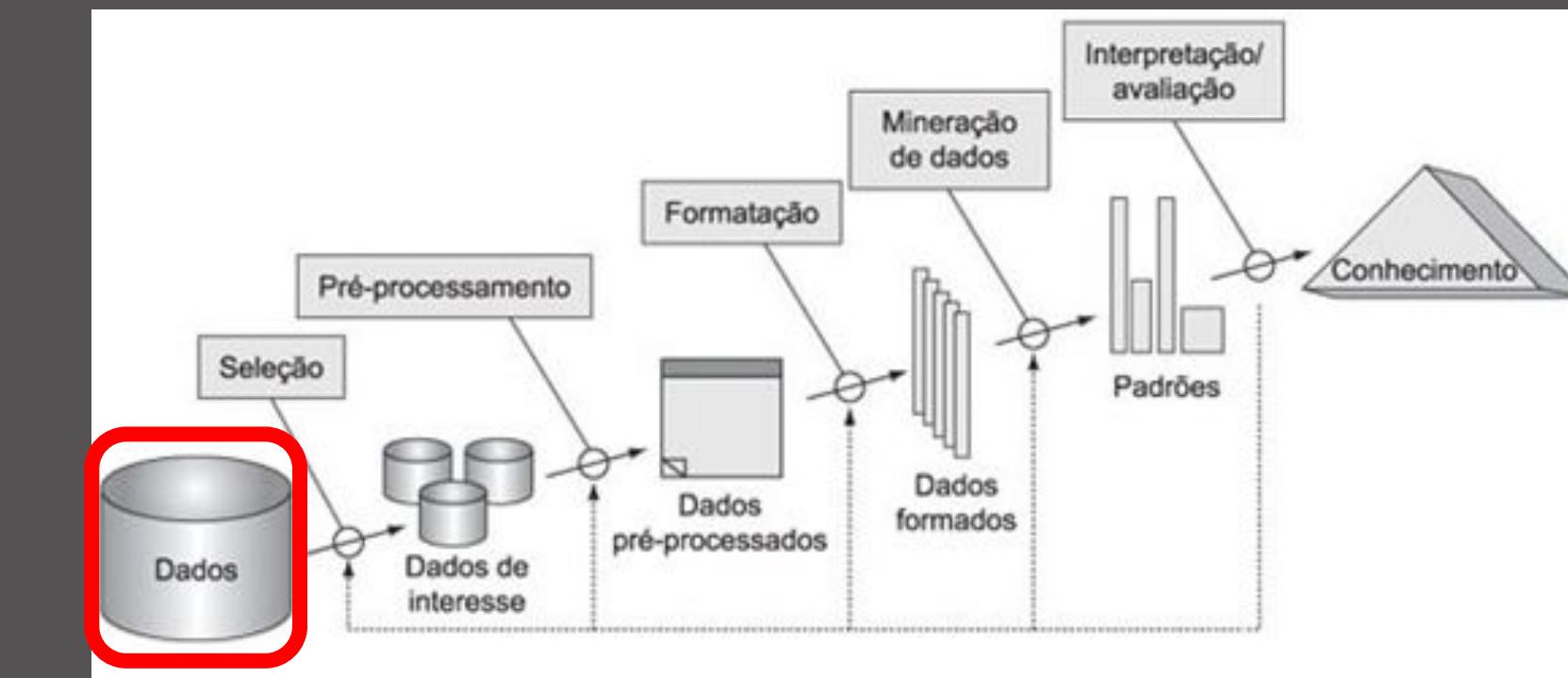
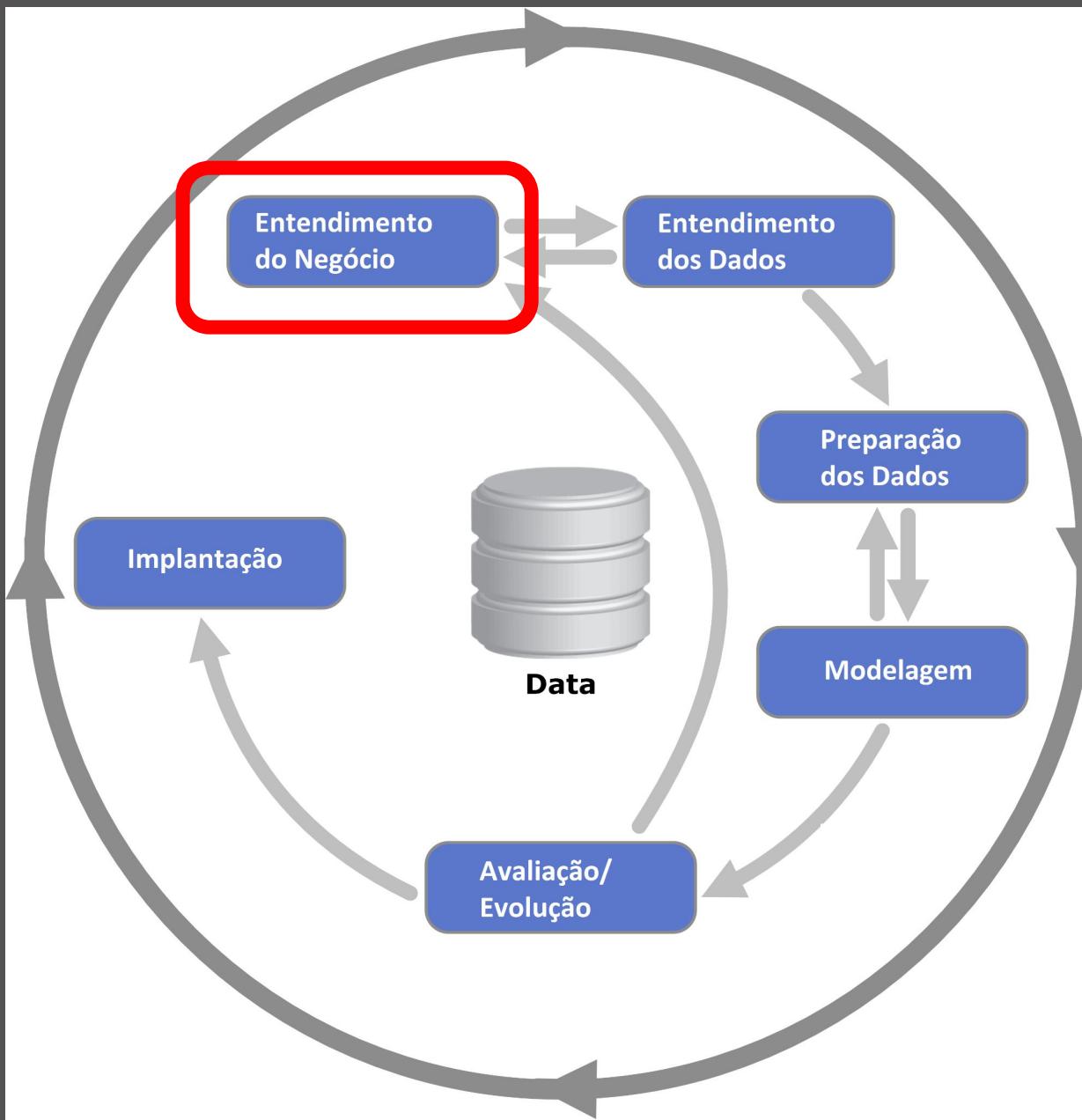
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Existem diferentes tipos de metodologia as mais populares são KDD (*Knowledge Discovery in Database*) e CRISP-DM(*Cross Industry Process for Data Mining*);
- As empresas podem desenvolver sua própria metodologia;
- Mas na prática, **as principais etapas são praticamente as mesmas em todas as metodologias:** coleta, preparação dos dados, modelagem, validação do modelo e resultados



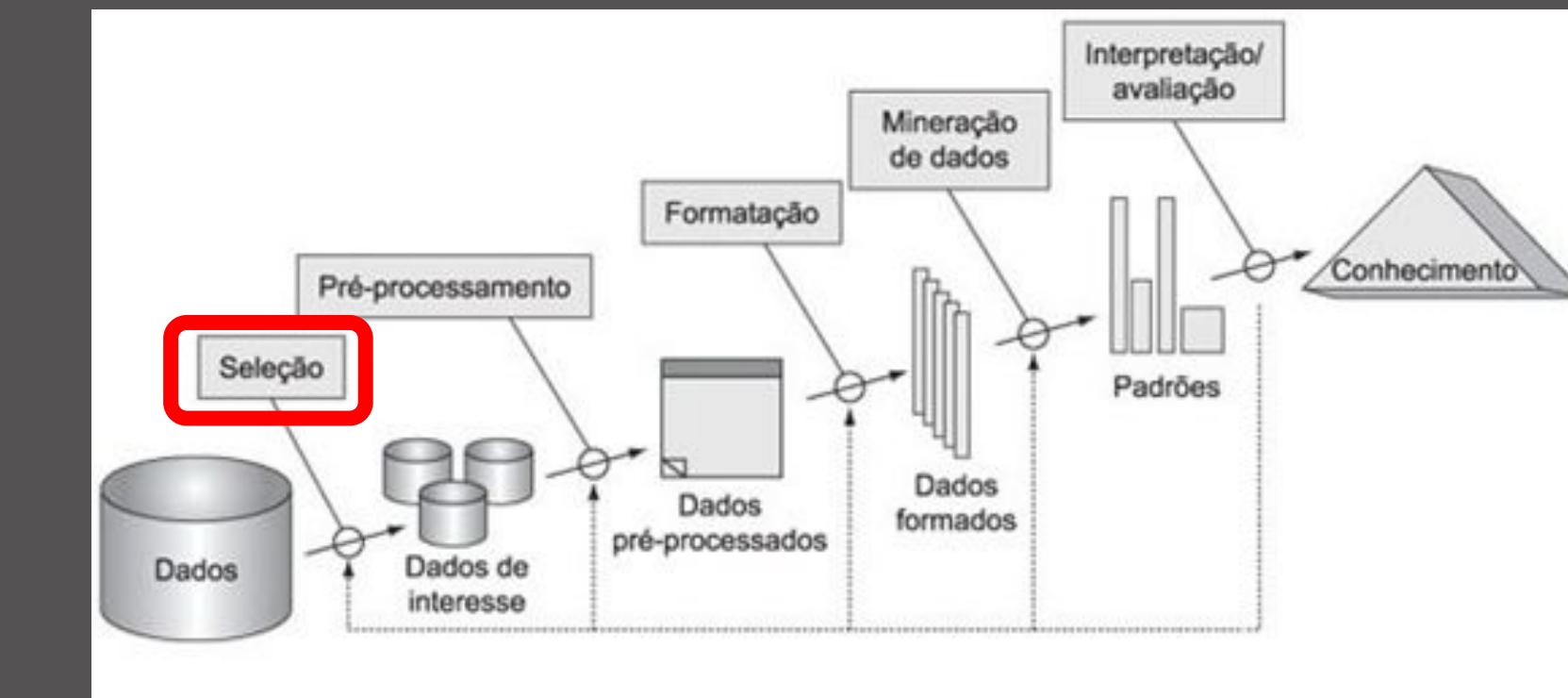
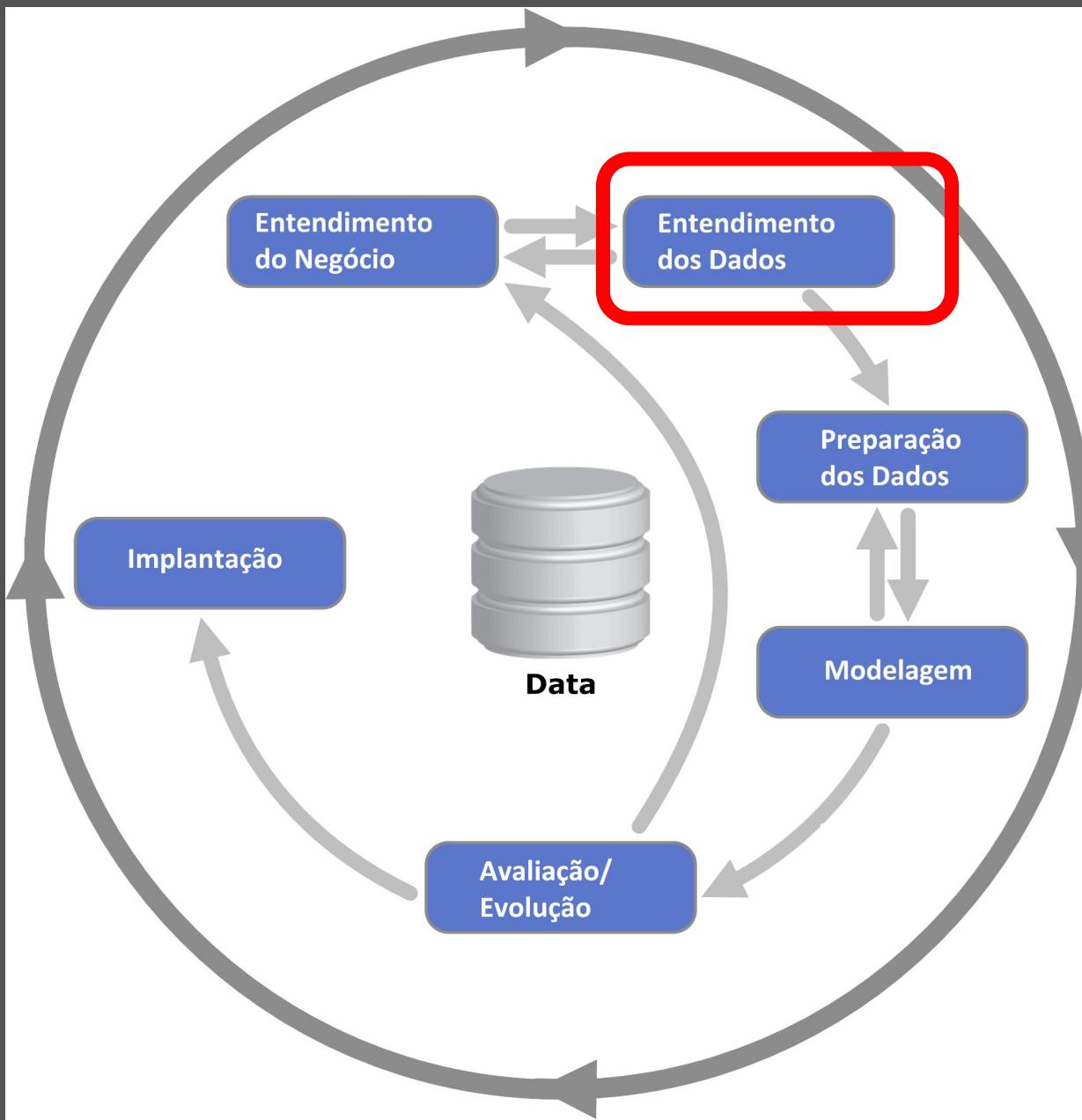
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação



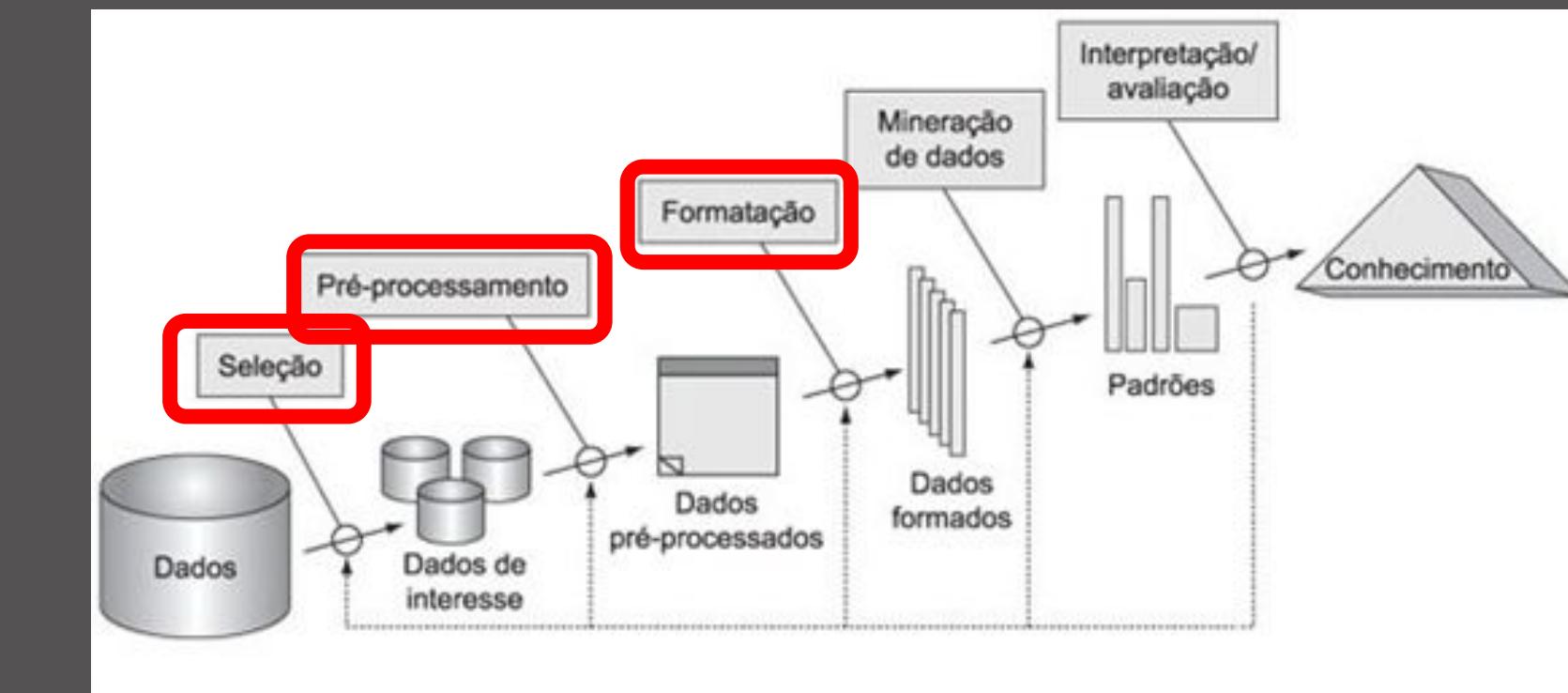
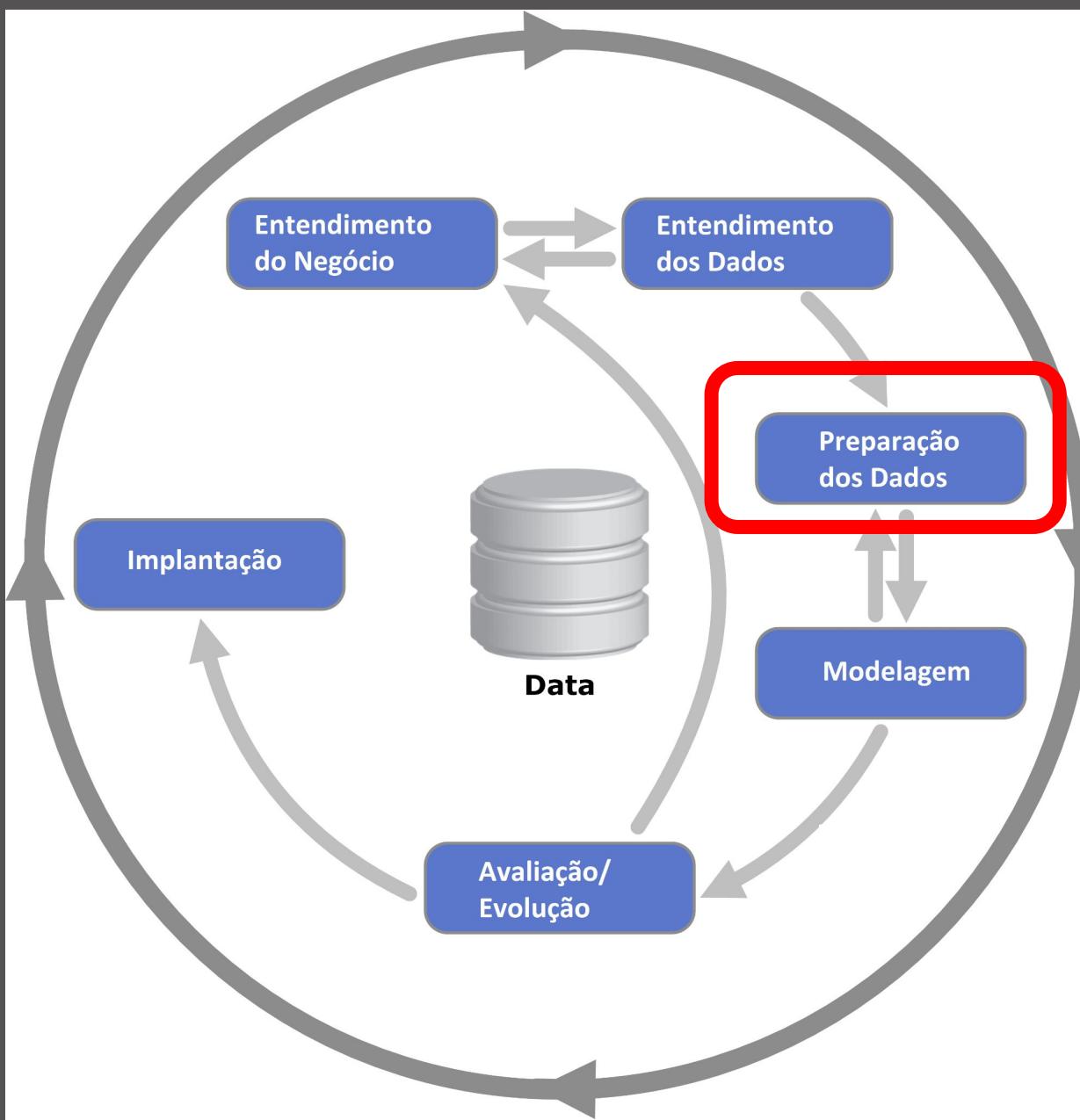
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação



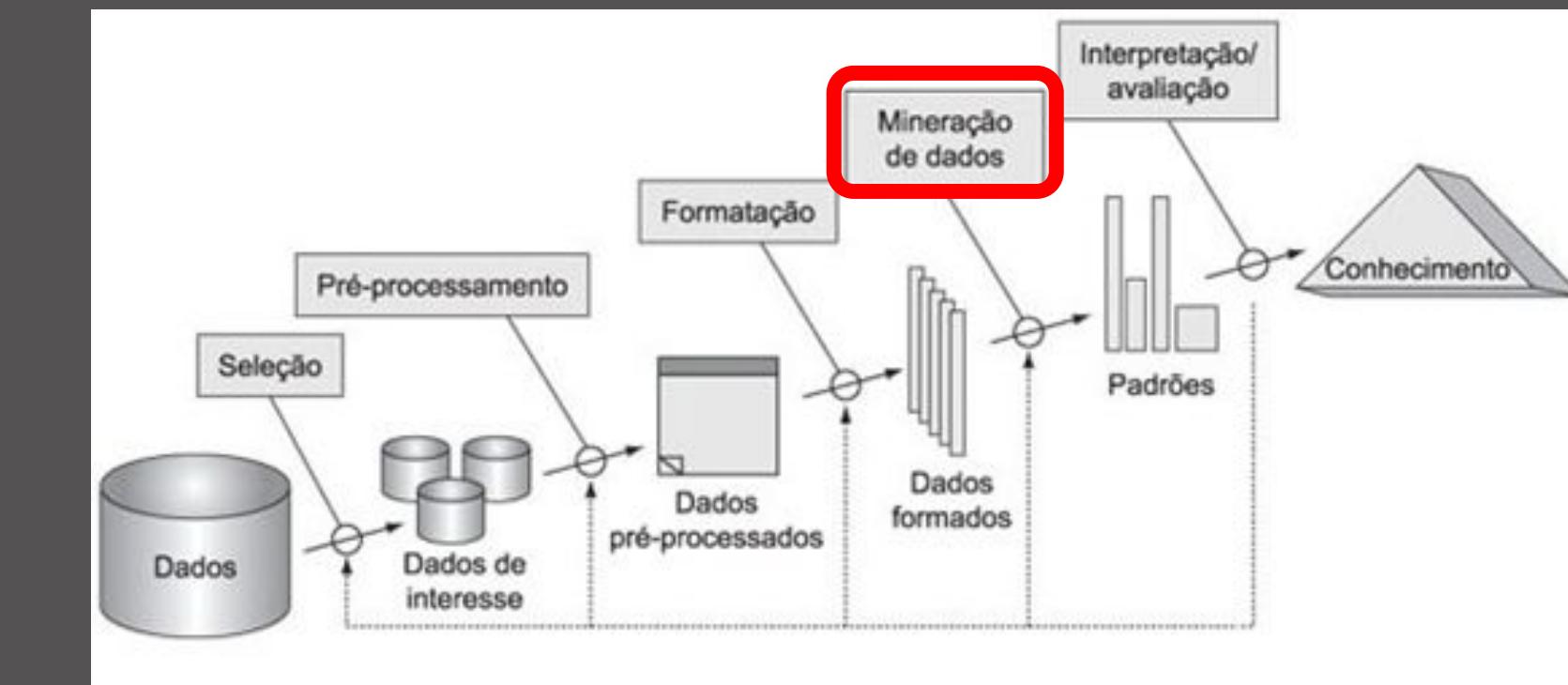
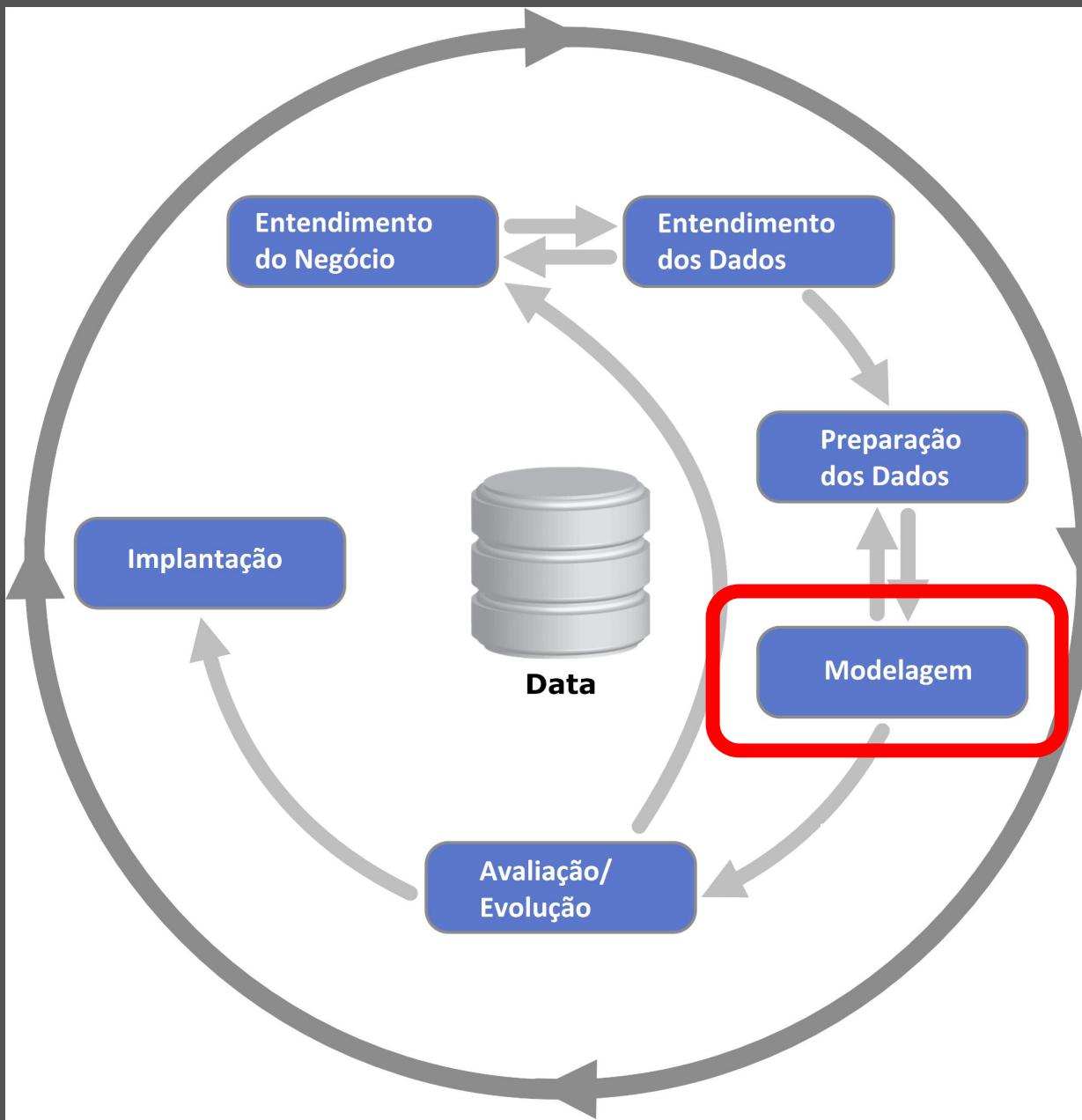
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação



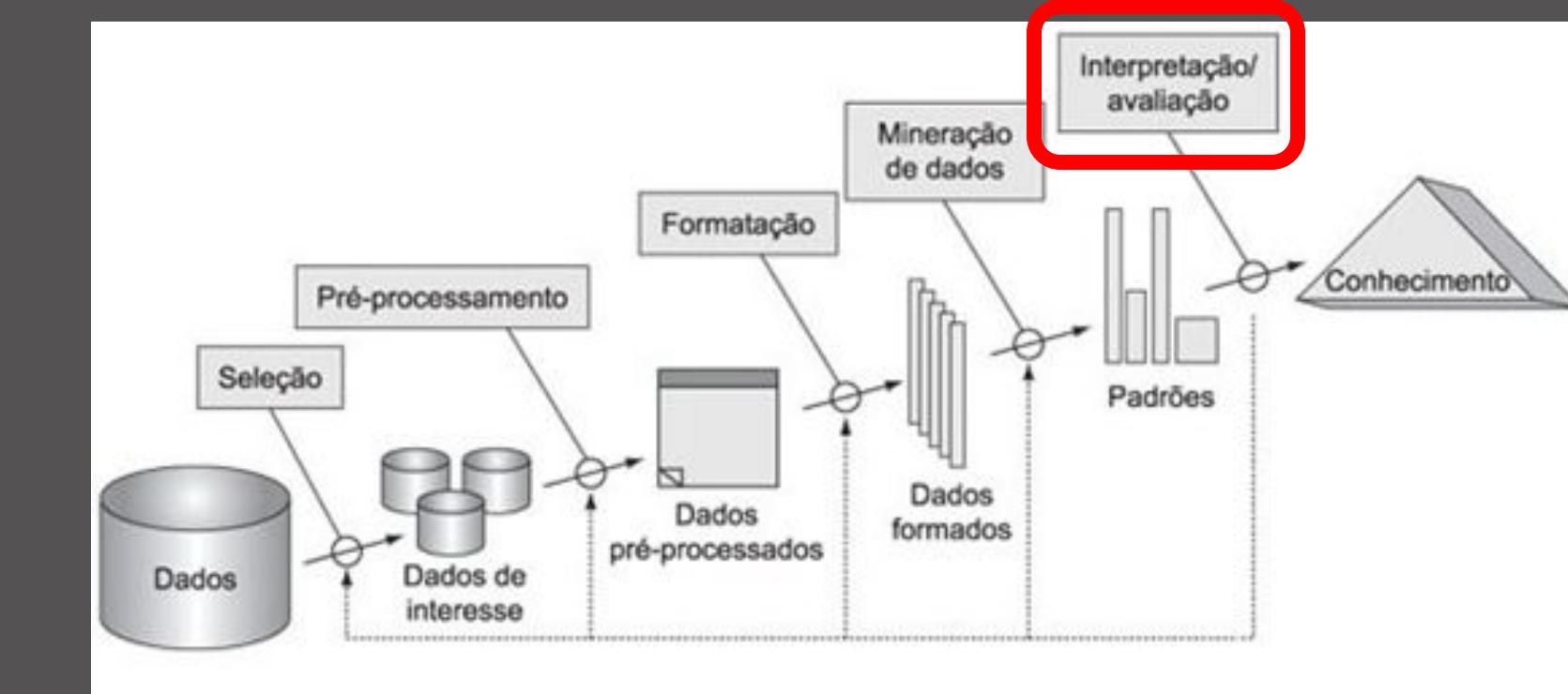
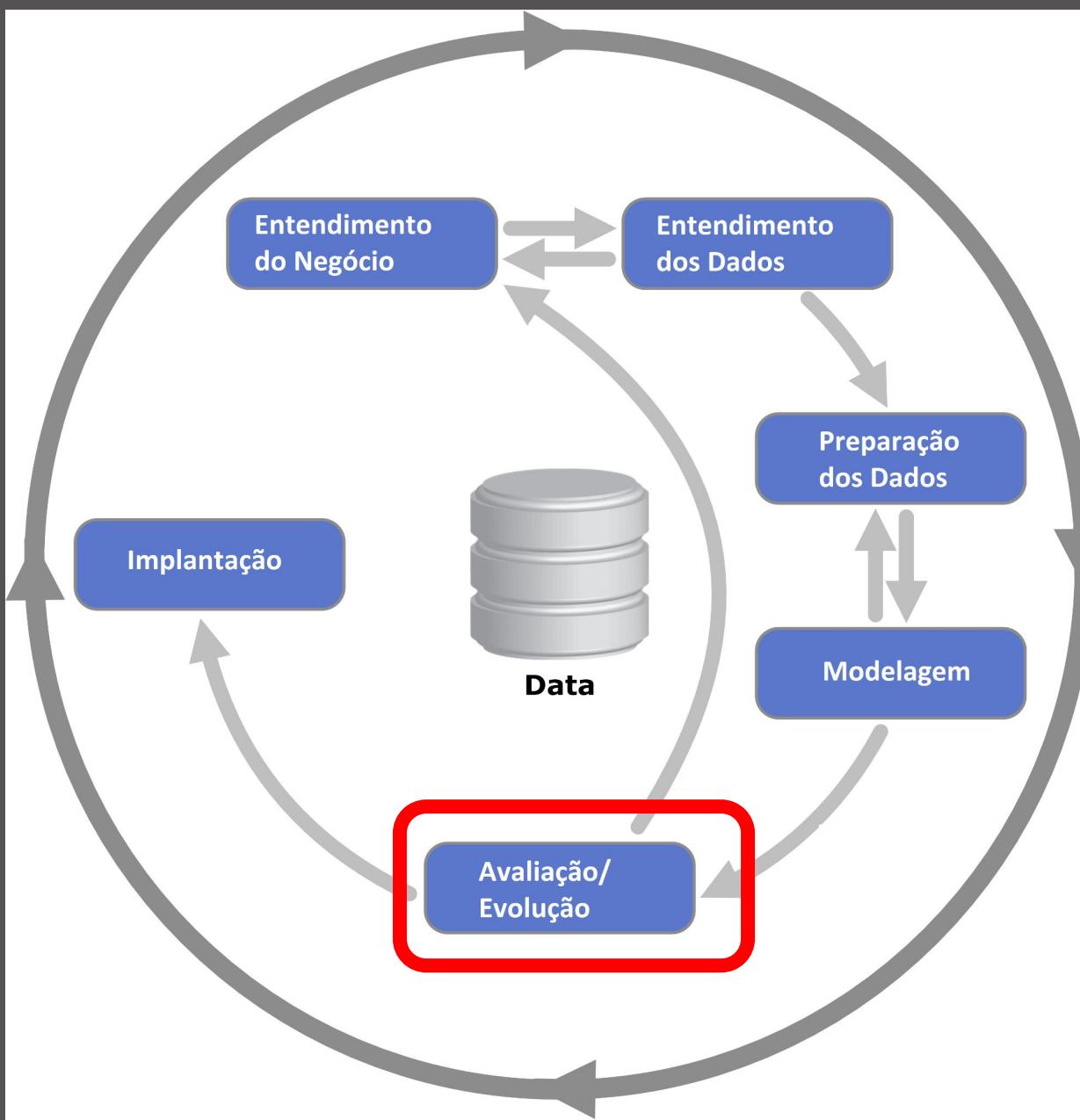
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação



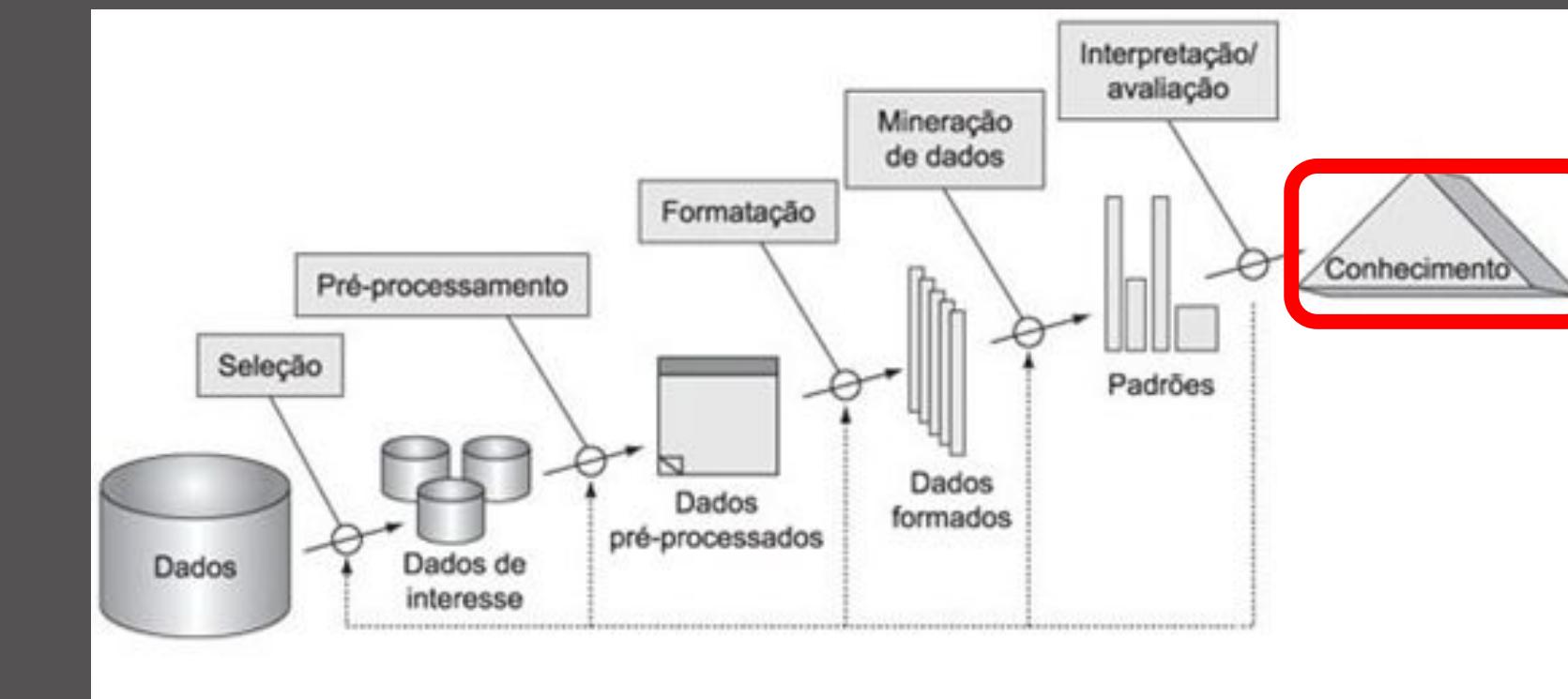
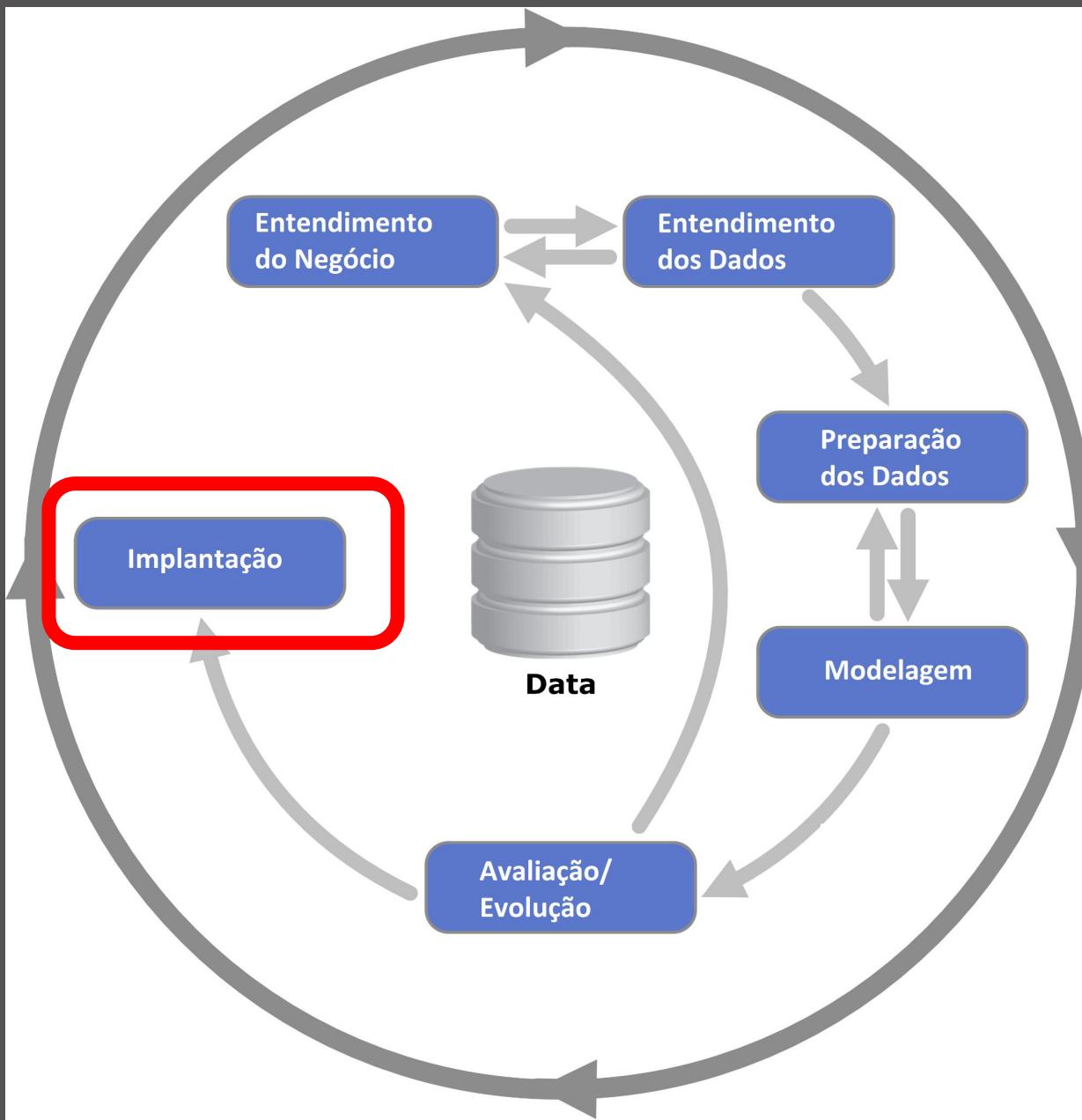
# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação



# METODOLOGIAS PARA PROJETOS DE CIÊNCIA DE DADOS

- Comparação





**Vamos abordar cada etapa já focando  
nas necessidades do case a ser  
desenvolvido**



# ETAPAS PARA MACHINE LEARNING

Entendimento  
do negócio

- Entender alguns pontos importantes da área de negócio, objetivos, dores, perguntas a serem respondidas, fazer os levantamentos de requisitos no geral;
- Fazer o mapeamento das bases de dados, relacionamentos entre as bases, onde encontrar os dados, possíveis variáveis que serão utilizadas, possíveis problemas e afins;
- **Negócio:** Comércio automobilístico.
- **Objetivos do projeto:**
  - Desenvolver um algoritmo para **prever o valor** do automóvel de acordo com suas características;
  - Fornecer alguns **insights** a partir dos dados disponíveis.

Coleta e  
entendimento  
dos dados

- Início dos trabalhos práticos e técnicos.
- Realizar a **coleta** de todos os dados necessários para atender as necessidades, com **foco nos objetivos e nas perguntas a serem respondidas**.
- Também é nesta fase que se dá o início na **exploração na base de dados**, ou seja:
  - entender como a base de dados está organizada
  - se os nomes das colunas são iguais da documentação
  - os tipos de dados que estão armazenados
  - se o tipo de dados condiz com o atributos
  - possíveis problemas a serem resolvidos, transformações
  - padronizações necessárias
  - dentre outros.

- Tabela de dados

**Coleta e entendimento dos dados**

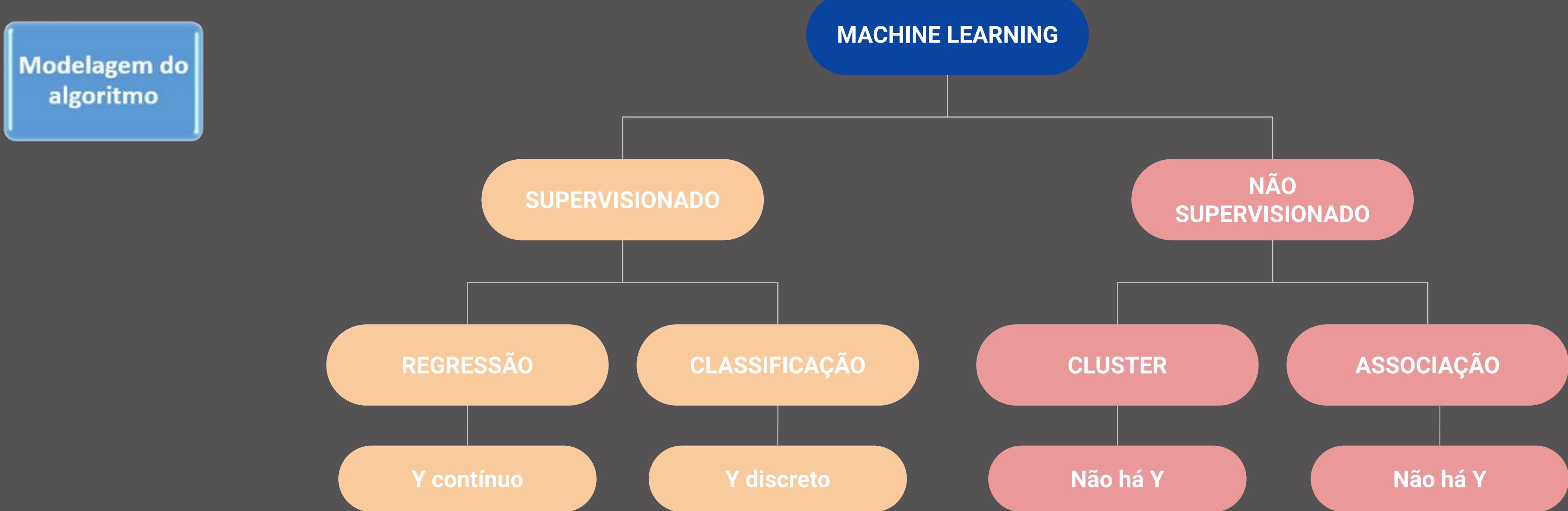
.csv: tabela com metadados dos anúncios

ATRIBUTO	Descrição	ATRIBUTO	Descrição
ID	Identificador único do anúncio	LINK	Link para anúncio completo
TITULO	Título do anúncio	DATA COLETA	Data que foi feita a extração do anúncio
INFOS	Informações básicas sobre o veículo (KM, CÂMBIO, COMBUSTÍVEL)	UF	UF do anúncio
VALOR	Valor do veículo anunciado	ANO	Ano do veículo (não tem no metadados, info complementar)
LOCAL	Localização (cidade-UF) do anúncio	DATA COLETA COMPLETA	Data que foi feita a extração do anúncio completo
DATA ANUNCIO	Data que o anúncio foi criado		

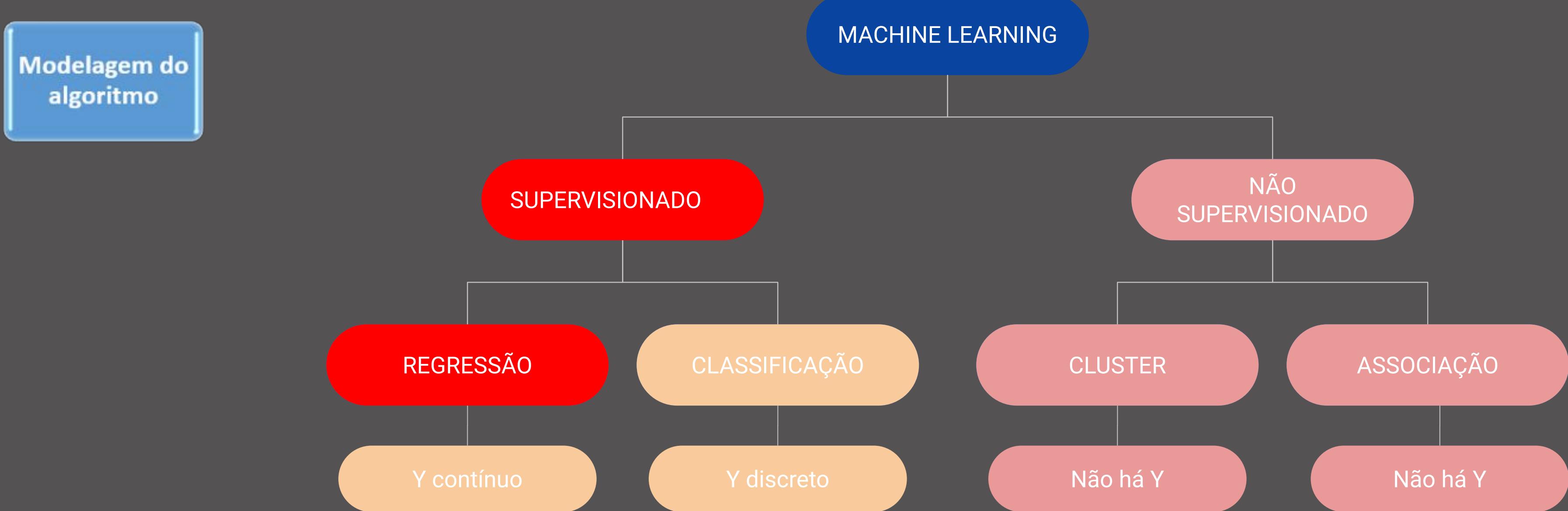
Preparação e  
exploração dos  
dados

- Procedimentos **mais importantes** de um projeto de análise de dados
- Estima-se que **esta fase pode demorar ~80%** do tempo do projeto.
- Tem-se como objetivo obter um conjunto de dados final com qualidade, e para isso realiza-se os seguintes procedimentos:
  - selecionar linhas e colunas
  - limpeza de valores inválidos
  - transformação dos dados
  - padronização dos valores (principalmente de texto)
  - merge entre base de dados (cruzamento entre tabelas)
  - limpeza de caracteres inválidos (texto)
  - dentre outras

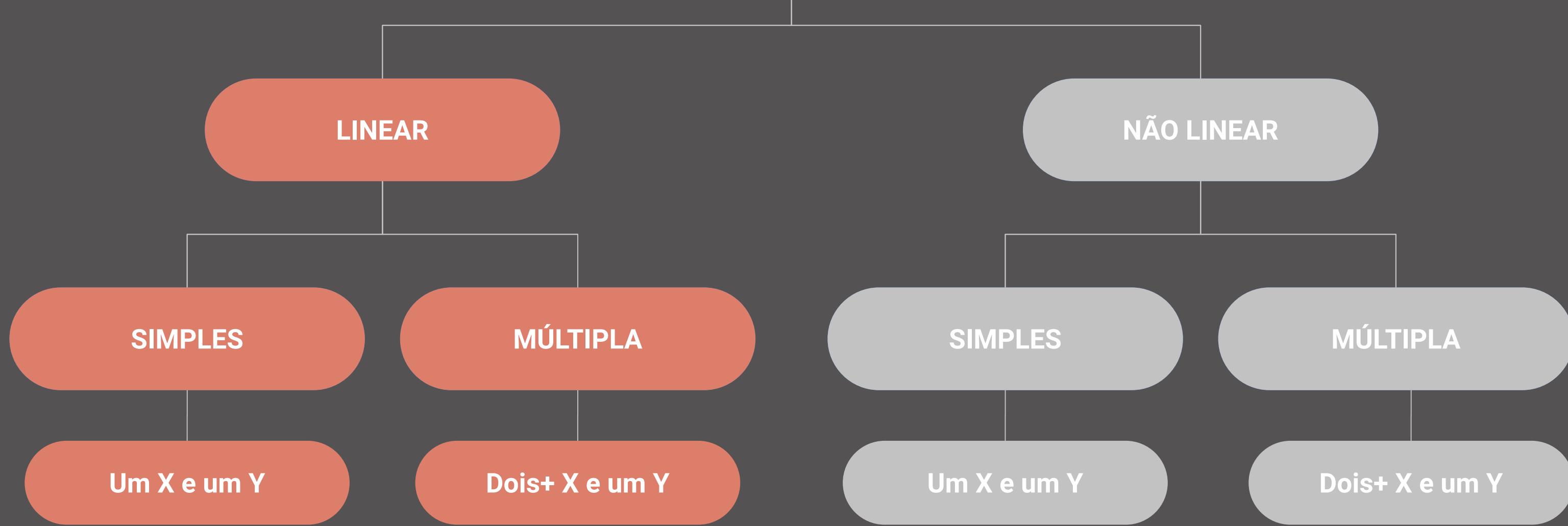
- A escolha dos algoritmos será feita com base no problema proposto.



- A escolha dos algoritmos será feita com base no problema proposto.



## REGRESSÃO

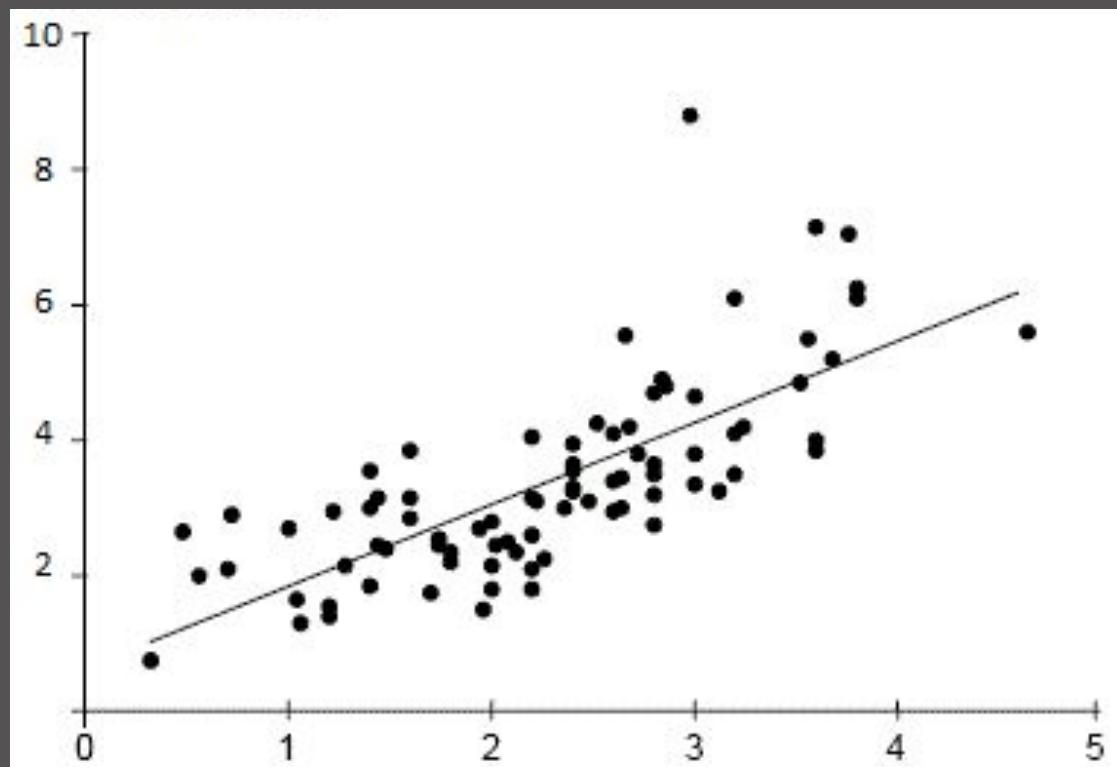


- **Regressão Linear:** Quando há uma reta inclinada entre os valores de X e Y;
- **Regressão Não Linear:** Quando não há uma reta entre os valores de X e Y.

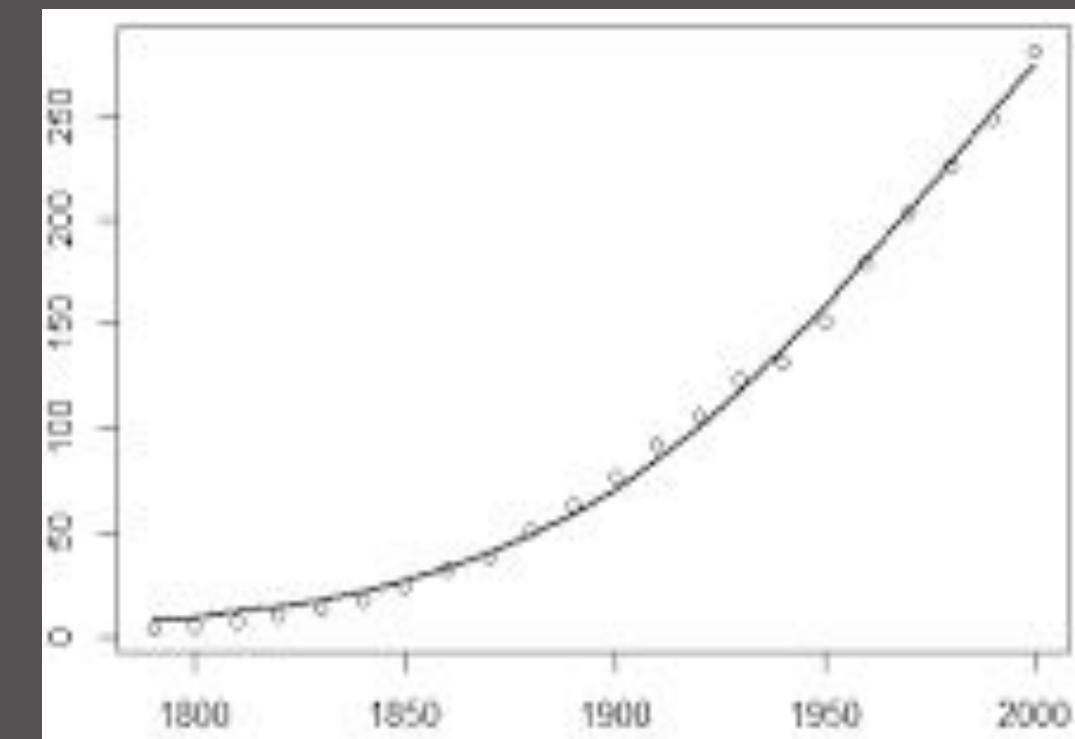
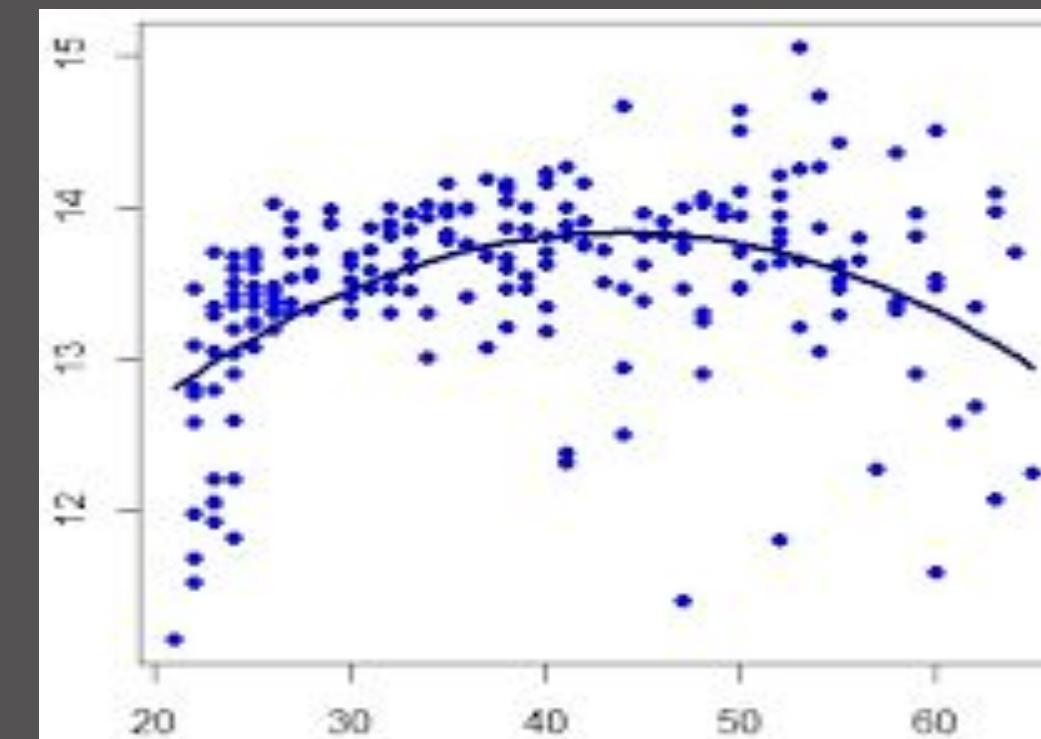
- **Regressão Simples:** Tem apenas **uma variável X**(preditora);
- **Regressão Múltipla:** Tem **duas+ variáveis X**(preditoras);

# Regressão

Regressão Linear

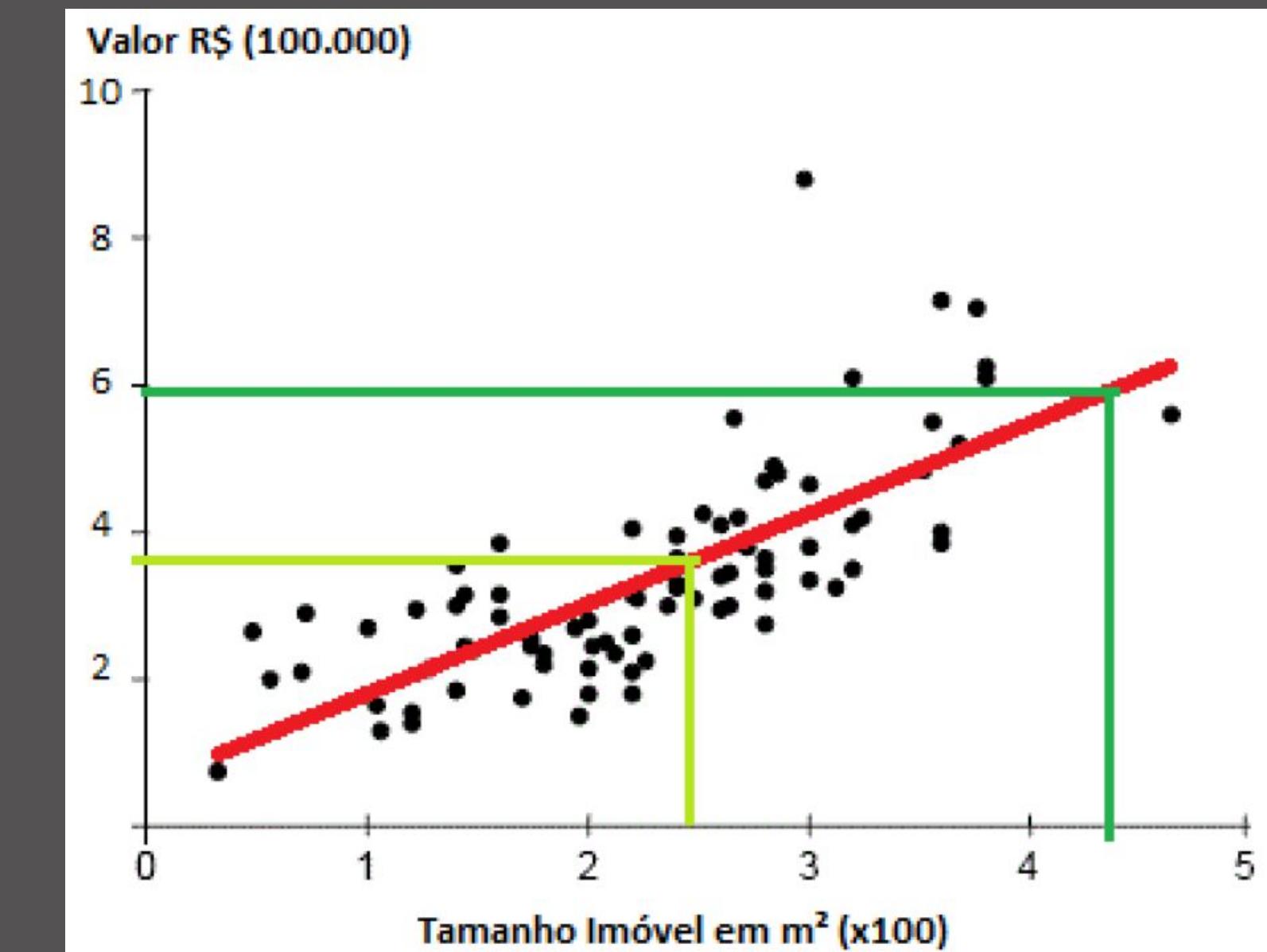
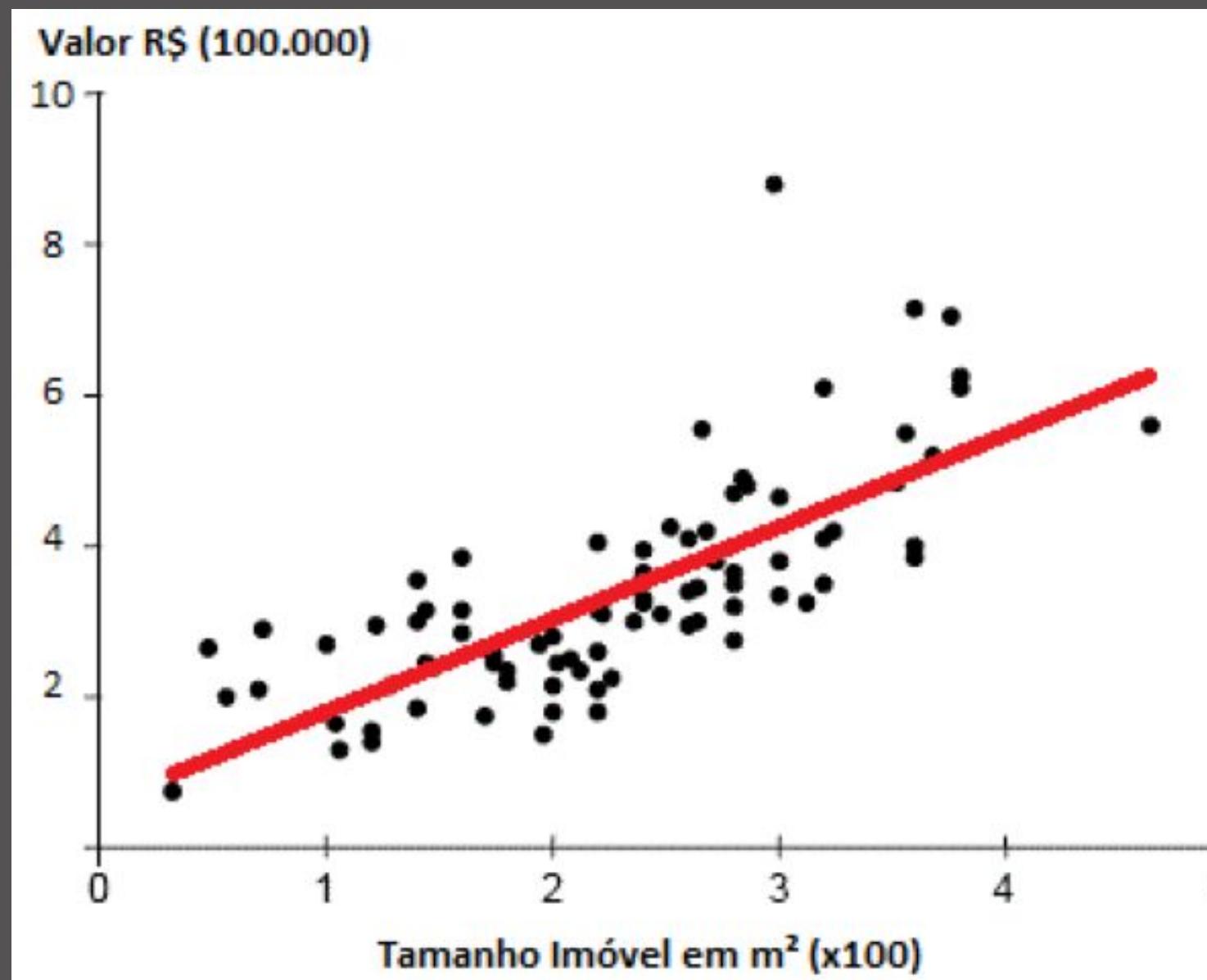


Regressão Não Linear





## Previsão de valores com regressão



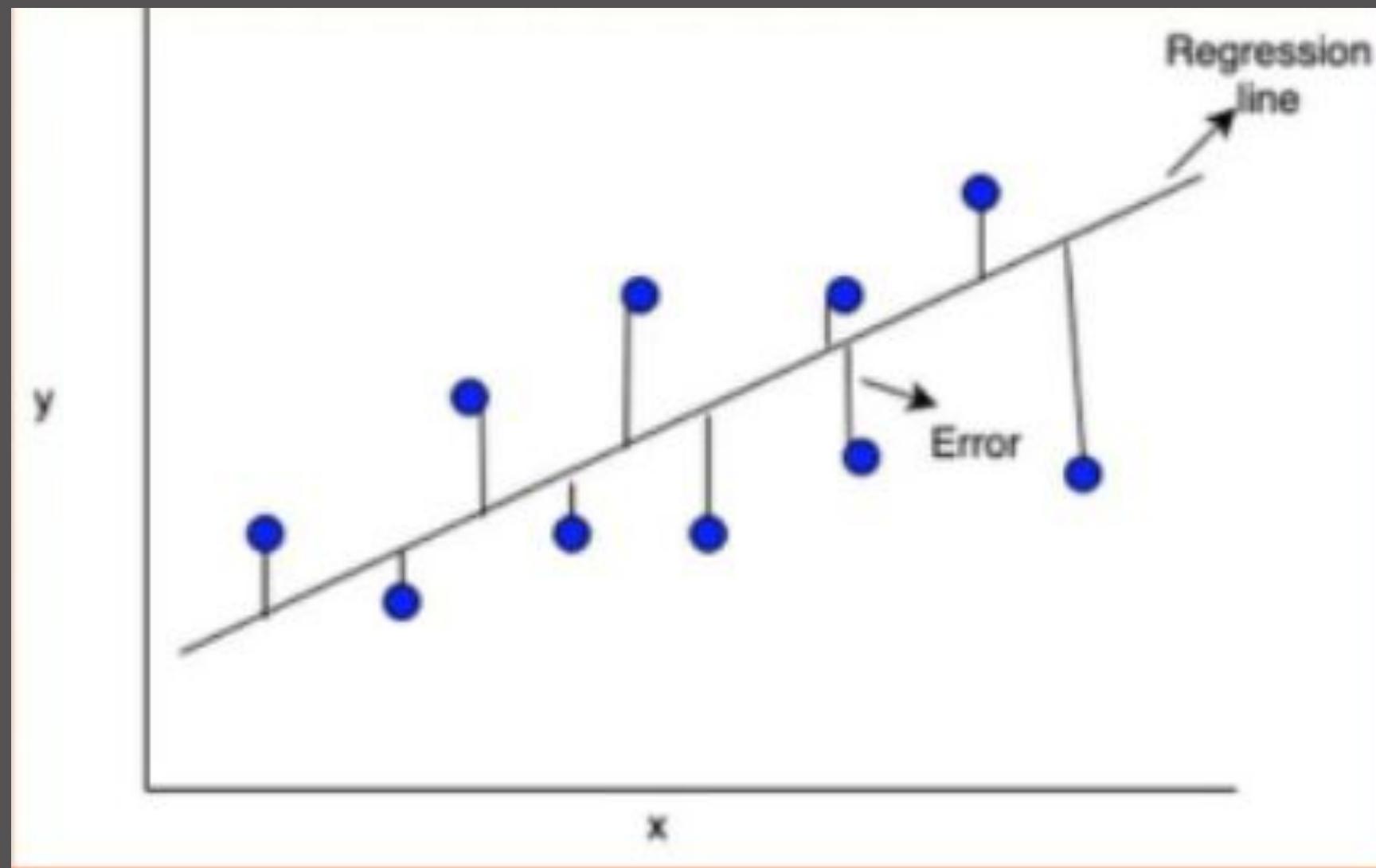
## Trabalhos Correlatos(semelhantes)

- <https://sungsoo.github.io/2018/04/11/predicting-car-prices.html>
- <https://github.com/suhasmaddali/Car-Prices-Prediction>
- <https://www.youtube.com/watch?v=v-C2zV6939c>
- <https://www.youtube.com/watch?v=6qxUcdKd43I>
- <https://marcusnunes.me/posts/predizendo-o-preco-de-carros-usados-com-machine-learning/>
- [https://colab.research.google.com/drive/1\\_hTGhFCkQ\\_NDcv9oQ5cgxZE81N523dwc?authuser=1#scrollTo=MgKthV0ftNri](https://colab.research.google.com/drive/1_hTGhFCkQ_NDcv9oQ5cgxZE81N523dwc?authuser=1#scrollTo=MgKthV0ftNri)
- <https://www.slideshare.net/HARPREETSINGH1862/predicting-model-for-prices-of-used-cars>
- <https://towardsdatascience.com/predicting-car-price-using-machine-learning-8d2df3898f16>
- <https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/>
- <https://rpubs.com/Zetrosoft/lbb-rm>

Validação do  
algoritmo

- Etapa para validar se o algoritmo realmente atende as necessidades
- Se não há overfitting/underfitting
- Aplicar métricas de validação de algoritmo mais detalhados:
  - MAE (erro absoluto médio);
  - MSE (erro quadrático médio);
  - RMSE (raiz do erro quadrático médio);
  - $R^2$ ;
  - $R^2$  Ajustado;
  - Análise de resíduos;
  - outros.

Validação do  
algoritmo





Resultados

- Relatórios;
- Tabelas/arquivos excel;
- Modelos/algoritmos;
- Dashboard online;
- APIs;
- Integração a sistema existente;
- Sistema completo;
- dentre outras.

# PESQUISA/FEEDBACK

<https://bit.ly/3wM8zFn>

