



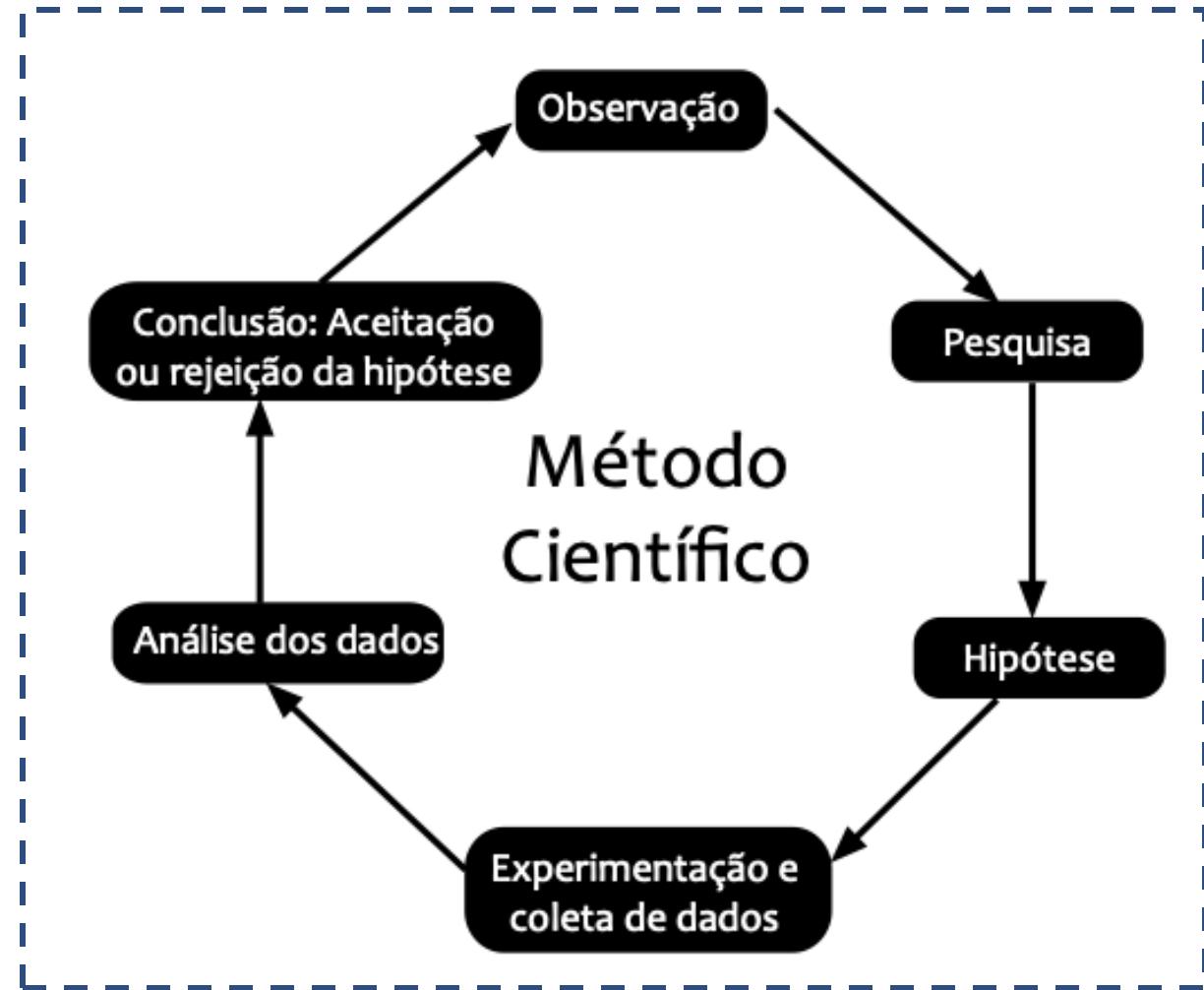
escola  
britânica de  
artes criativas  
& tecnologia

**Profissão: Cientista de Dados**  
Módulo 2: Metodologia

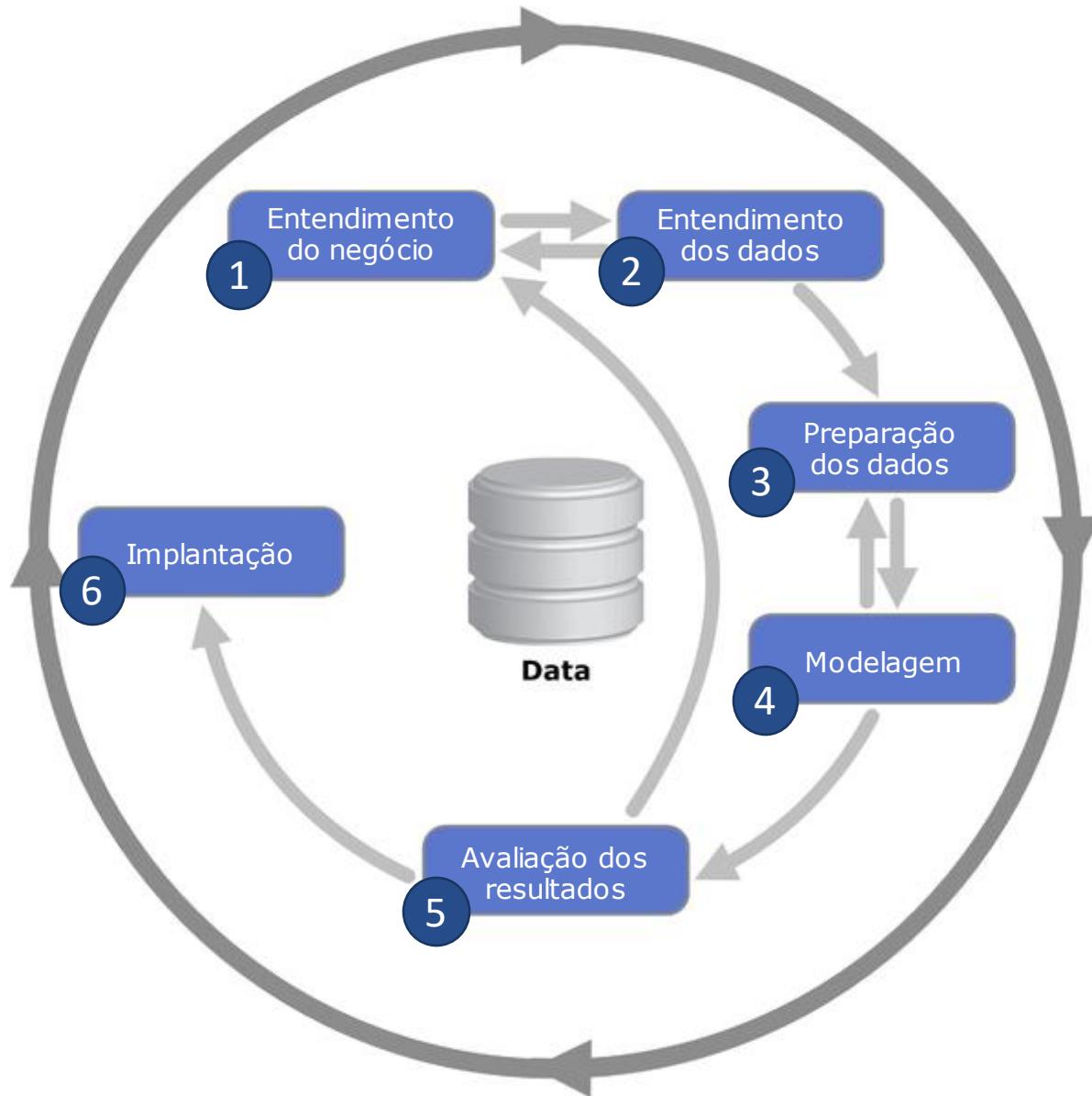
# Introdução: Ciência de Dados

# O Método Científico

É uma sequencia de passos básicos que devem ser seguidos para que um conhecimento tenha validade na comunidade científica. Os passos são:



# O que é CRISP-DM



- CRISP-DM, em tradução livre, significa "processos padrão de Data Mining para qualquer indústria" (cross industry standard processes for Data Mining em inglês).
- É um conjunto de passos recomendado para se realizar Data Mining em qualquer indústria. Tem muitas semelhanças com o método científico. A diferença mais marcante é que os resultados em geral não devem ser publicados para a comunidade científica, mas se pauta no mesmo ceticismo

# 1. Entendimento do negócio

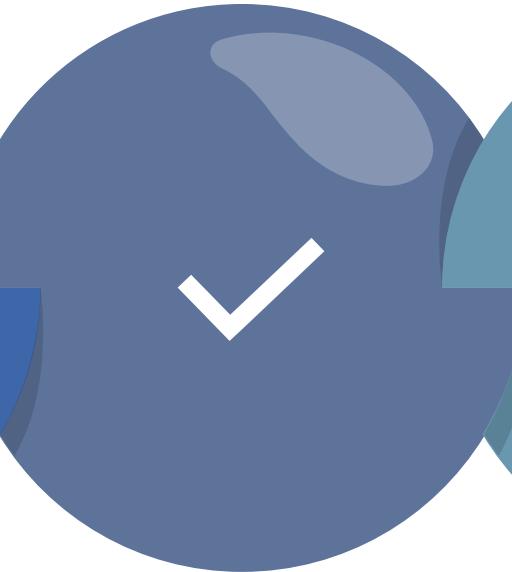
---



Entender os  
objetivos do  
negócio



Avaliar a situação



Entender os  
objetivos de  
modelagem



Realizar o  
planejamento do  
projeto

## 2. Entendimento dos dados

---



Fontes



Coleta



Descrever



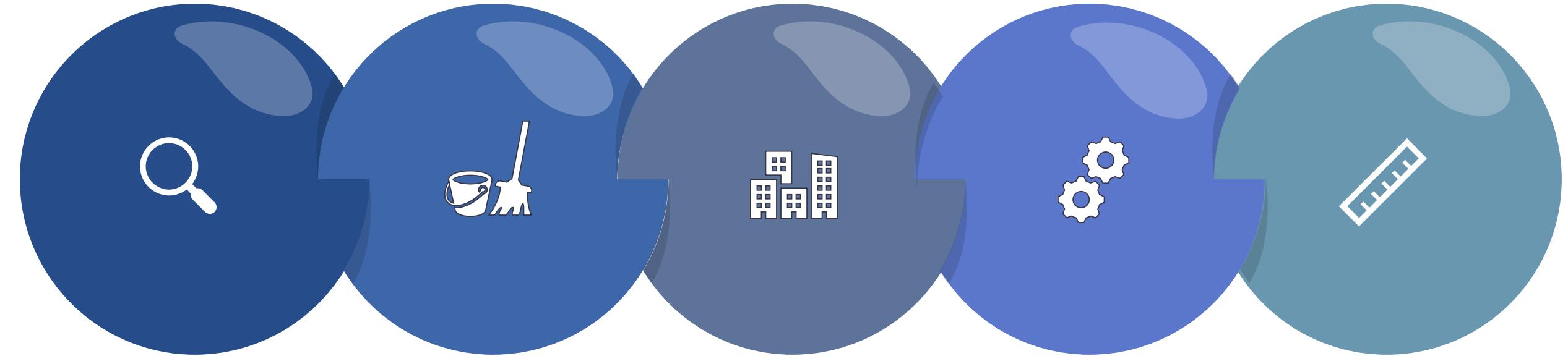
Explorar



Qualidade

# 3. Preparação dos dados

---



Seleção

Limpeza

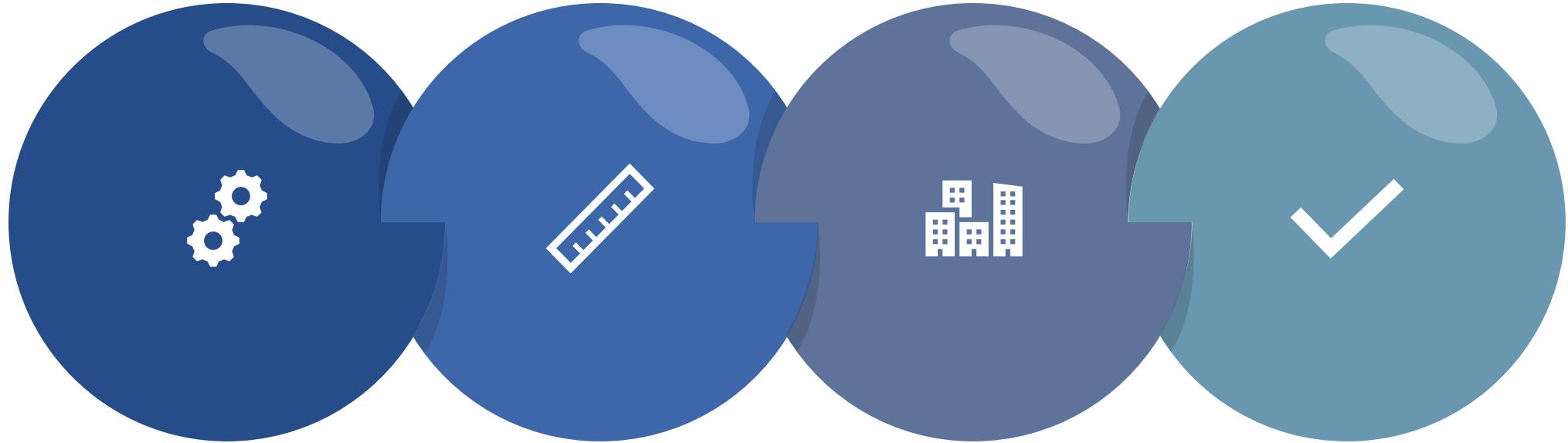
Construção

Integração

Formatação

# 4. Modelagem

---



Técnica

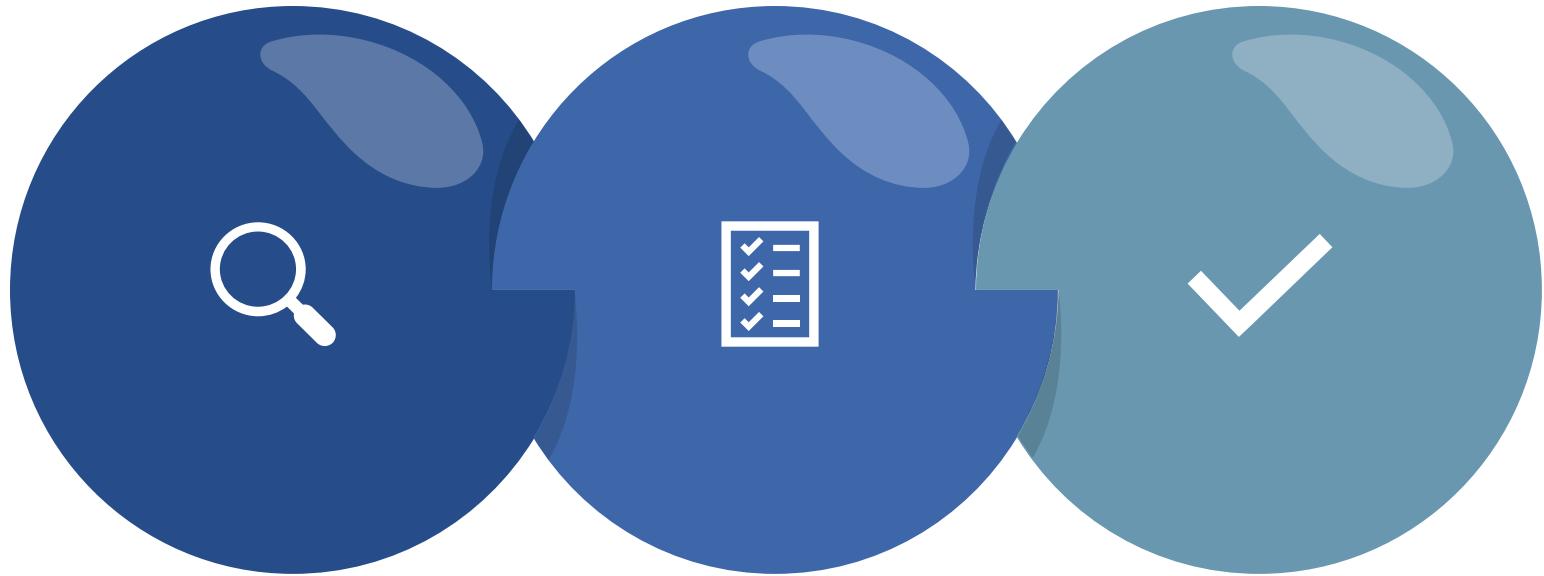
Design

Construção

Avaliação

# 5. Avaliação dos resultados

---



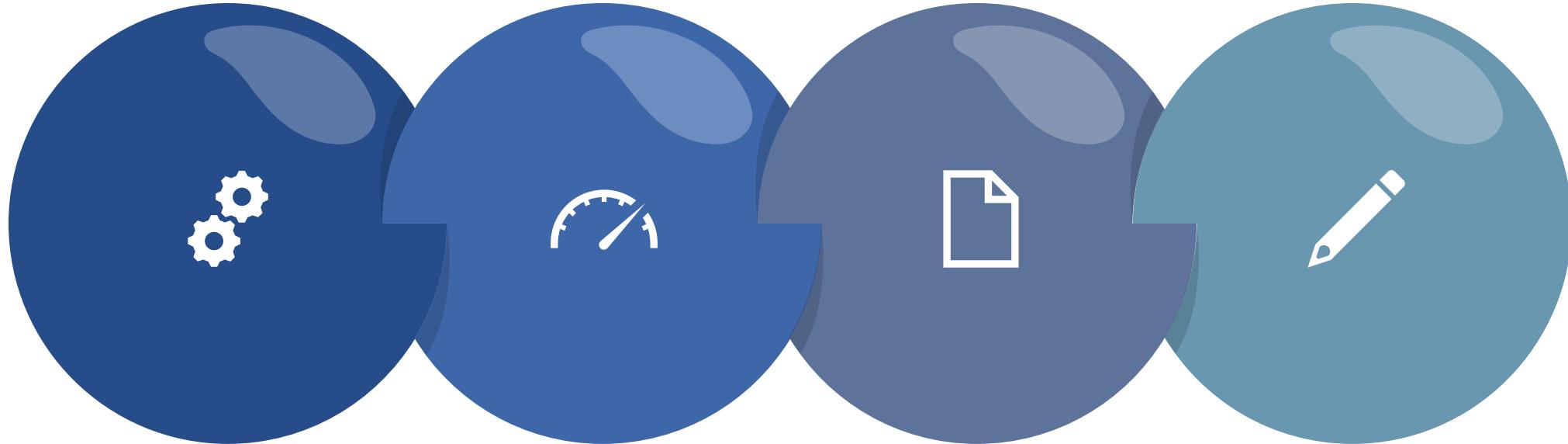
Avaliar

Rever processos

Próximos passos

# 6. Implantação

---



Implantação

Monitoramento

Relatório

Revisão

# Conclusões

---

Regra geral

Uso do modelo

Ciclo de vida

Múltiplos modelos

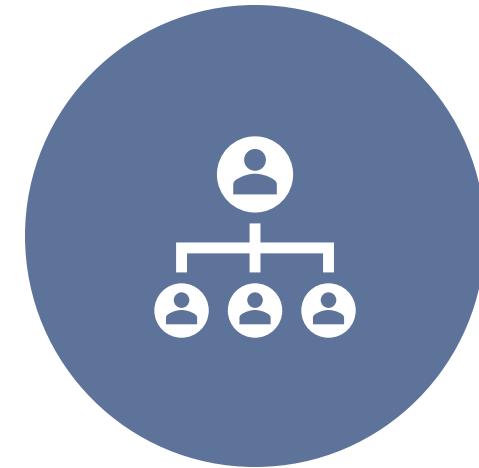
# Tutorial

# Setup

---



PYTHON



GITHUB

# Instalando o Anaconda no Windows

---

- A distribuição do Python que vamos usar é o Anaconda – mais detalhes nas próximas lições.
- Para instalar, visite a página do Anaconda: <https://docs.anaconda.com/anaconda/install/windows/>
- Faça o download do instalador ([Download the Anaconda installer](#))
- Clique duplo no instalador para iniciar
- Clique "próximo"
- Leia os termos e condições e clique em "aceito"
- Se você quiser instalar para todos os usuários vai precisar de permissões de administrador da máquina. De outra forma clique em "para mim apenas"
- Selecione um local de instalação e clique "próximo"
- Não adicione o Anaconda no ambiente "PATH" - pode interferir com outros programas. Inicie o Anaconda pelo Anaconda Navigator.
- Escolha a opção de tornar o Anaconda o seu interpretador de Python padrão.
- Ao final você vai ver a caixa de diálogos "obrigado por instalar o Anaconda".



# Instalando o Anaconda no Mac

---

- A distribuição do Python que vamos usar é o Anaconda – mais detalhes nas próximas lições.
- Para instalar, faça o download do macOS installer
- Clique duplo no instalador, clique "Continuar"
- Responda aos termos introdutórios e de licença
- Clique "instalar" para instalar o anaconda no seu diretório ~ /opt (recomendado)
- Clique "continuar"
- Você vai ver a tela de uma instalação bem sucedida.



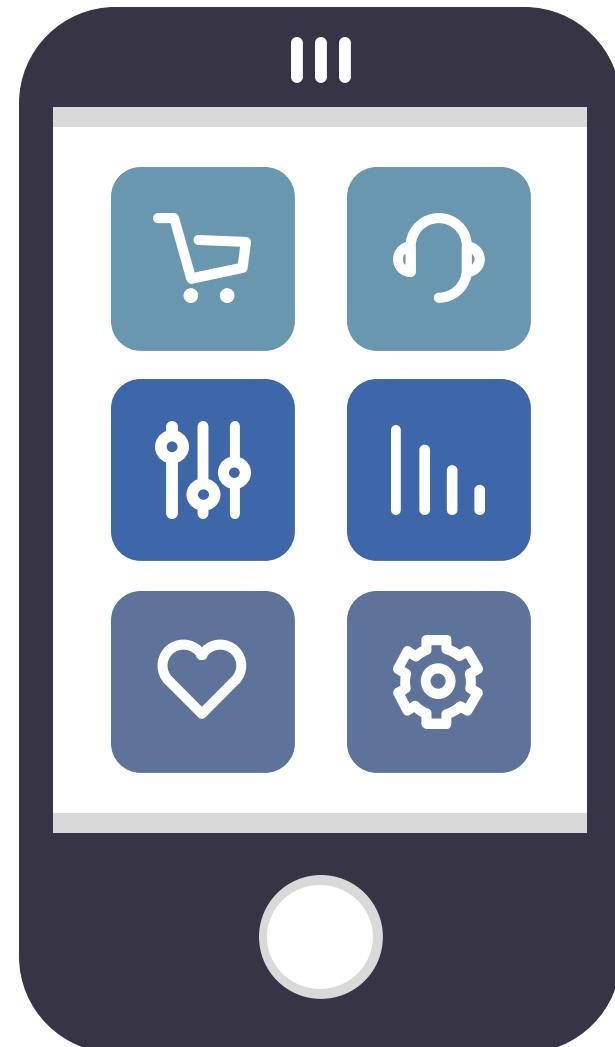
# Instalando o Anaconda no Linux

- A distribuição do Python que vamos usar é o Anaconda – mais detalhes nas próximas lições.
- Para instalar, verifique os pré-requisitos da instalação em <https://docs.anaconda.com/anaconda/install/linux/>
- Faça o download do instalador (Anaconda installer for Linux)
- Digite o código seguinte no terminal para instalar:  
`bash ~/Downloads/Anaconda3-2020.02-Linux-x86_64.sh`
- Faça o scrolldown nos termos de licenciamento e clique em "Aceito" para seguir.
- O instalador vai perguntar: "Do you wish the installer to initialize Anaconda3 by running conda init?" Responda "Yes".
- O instalador vai mostrar uma mensagem de instalação concluída.



# Abrindo uma conta no GitHub

- Acesse a página do GitHub (<https://github.com/>)
- Clique em *'Sign Up'*
- Preencha os seus dados e siga as instruções do site
- Preencha o *captcha* (pra provar que você não é um robô ?!?)
- Clique em criar conta
- Você receberá um e-mail para confirmar a sua conta.  
Confirme ☺
- Prontinho!



# Primeira Prática

---

## O conceito de MVP

Data Jiu-Jitsu: a arte de usar elemento de dados de forma inteligente para solucionar problemas interativos, de forma que, quando combinados, solucionam problemas que de outra forma são intratáveis. Jiu-jitsu é a arte ou técnica de manipular a força do oponente contra ele mesmo ao invés de confrontá-lo com a sua própria força.



# Atividade do módulo

- Baixe o arquivo Proj01.ipynb
- Nele você vai ver as instruções em azul
- Você vai alterar o código para:
  - Reproduzir o gráfico de barras para outras variáveis
  - Reproduzir o histograma
  - Fazer uma bivariada
  - Criar uma célula de Markdown e escrever suas conclusões
  - Fazer alterações na Random Forest e checar a acurácia
- Em seguida você vai subir o notebook no GitHub (ou fazer o upload na plataforma EBAC).