INF 1514

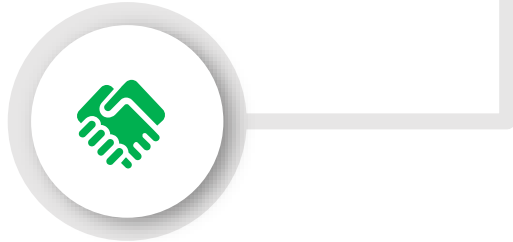
Introdução à Análise de Dados

Material 1

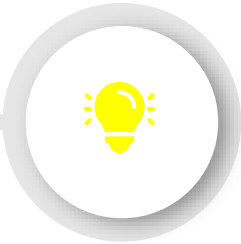
PUC
RIO

Este curso foi idealizado para **ensinar programação** para **análises básicas** e **visualizações de dados**.

INF 1514

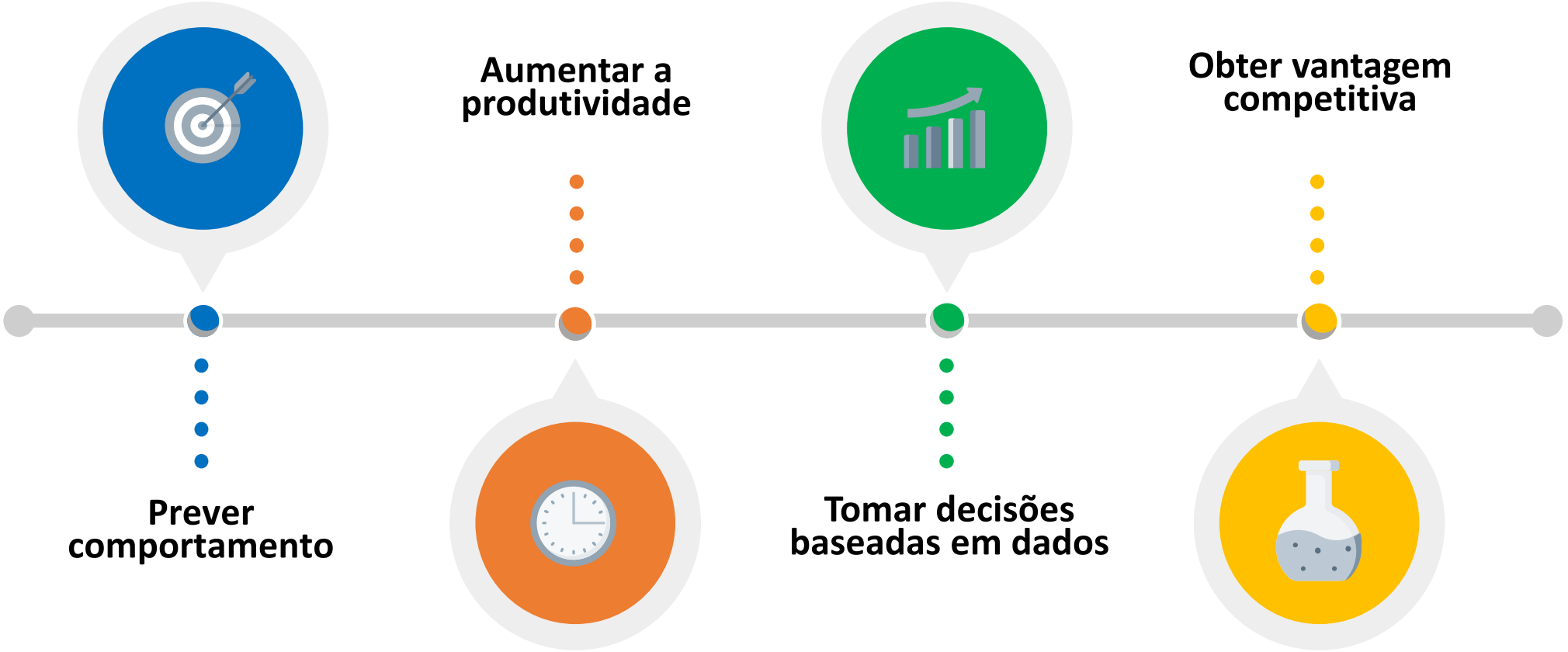


A **análise de dados** é um **processo** de inspeção, limpeza, transformação e modelagem de dados com o objetivo de **descobrir informações úteis**, **informar conclusões** e **apoiar a tomada de decisões**.

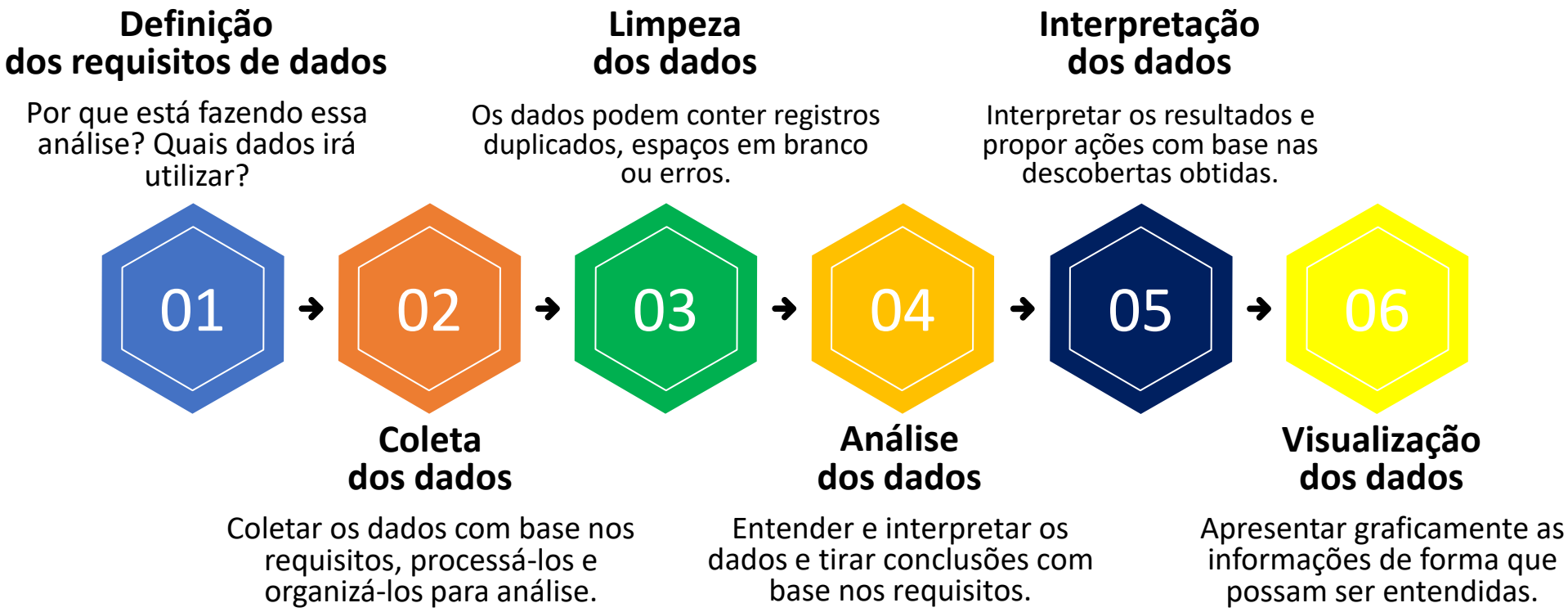


- Dados são simples observações sobre os estados do mundo.
- A análise de dados apresenta múltiplas abordagens, abrangendo diversas técnicas sob uma variedade de nomes, sendo usada em diferentes domínios de negócio, ciências e ciências sociais.
- Um exemplo simples de análise de dados pode ser visto sempre que tomamos uma decisão em nosso dia a dia, avaliando o que aconteceu no passado ou o que poderá acontecer no futuro se tomarmos uma determinada decisão.

Por que a Análise de Dados é importante?



Fases do processo de Análise de Dados



Análise de Dados

É o ponto de partida para reflexão analítica, visando responder à pergunta “O que aconteceu?”. Faz isso ordenando, manipulando e interpretando dados brutos de várias fontes para transformá-los em insights valiosos para sua organização.

O principal objetivo é explorar. Uma vez investigados os dados, a análise exploratória permite encontrar conexões e gerar hipóteses e soluções para problemas específicos. Uma área típica de aplicação é a mineração de dados.

Ajuda a obter uma compreensão contextual firme de **por que algo aconteceu**. Se você souber por que algo aconteceu e como aconteceu, será capaz de identificar as formas exatas de lidar com o problema ou desafio.

Permite olhar para o futuro para responder à pergunta: **o que vai acontecer?** Para isso, utiliza os resultados das análises descritivas, exploratórias e diagnósticas, além do aprendizado de máquina (ML) e da inteligência artificial (IA).

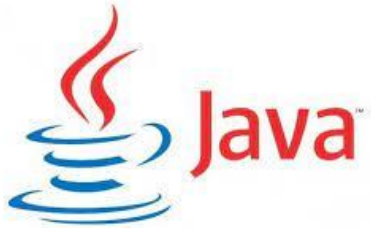
Permite determinar o melhor curso de ações a seguir em um determinado cenário. Analisa diferentes cenários é determina o resultado mais provável de cada um, permitindo assim tomar decisões com base no que espera alcançar.

Métodos para Análise de Dados

1. **Análise de Cluster**
2. **Análise de Regressão**
3. **Redes Neurais**

4. **Data e Text Mining**
5. **Árvores de Decisão**
6. **Análise de Séries Temporais**

Ferramentas para Análise de Dados



Ajudam os usuários a processar, manipular e visualizar dados, analisar as relações e correlações entre conjuntos de dados e também a identificar padrões e tendências para interpretação.

Saber programar é importante

1

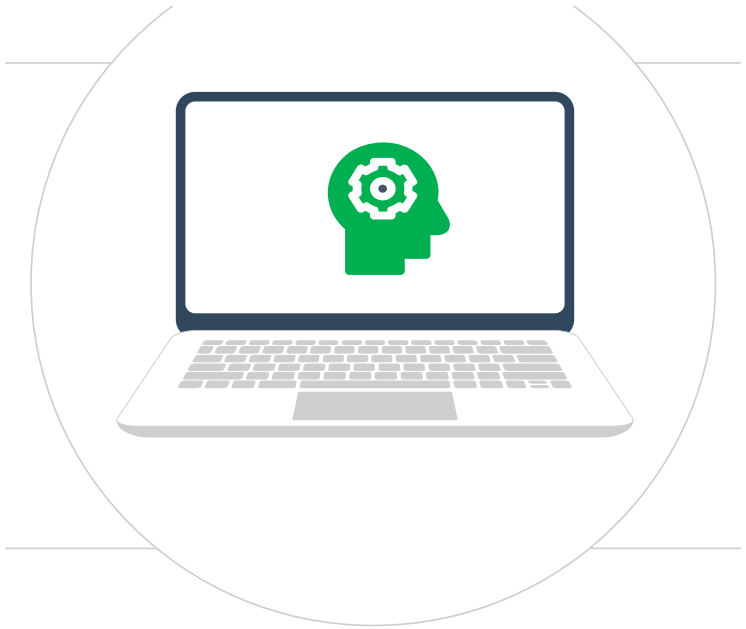
Raciocínio lógico

Ajuda no desenvolvimento do raciocínio crítico, analítico e lógico.

2

Habilidade de abstração

Melhora as habilidades de resolução de problemas e a capacidade de abstração.



Senso de organização

As habilidades que envolvem aprender a programar estão relacionadas com organização.

3

Análise de Dados

Mesmo utilizando ferramentas que auxiliam no processo, códigos precisarão ser desenvolvidos.

4

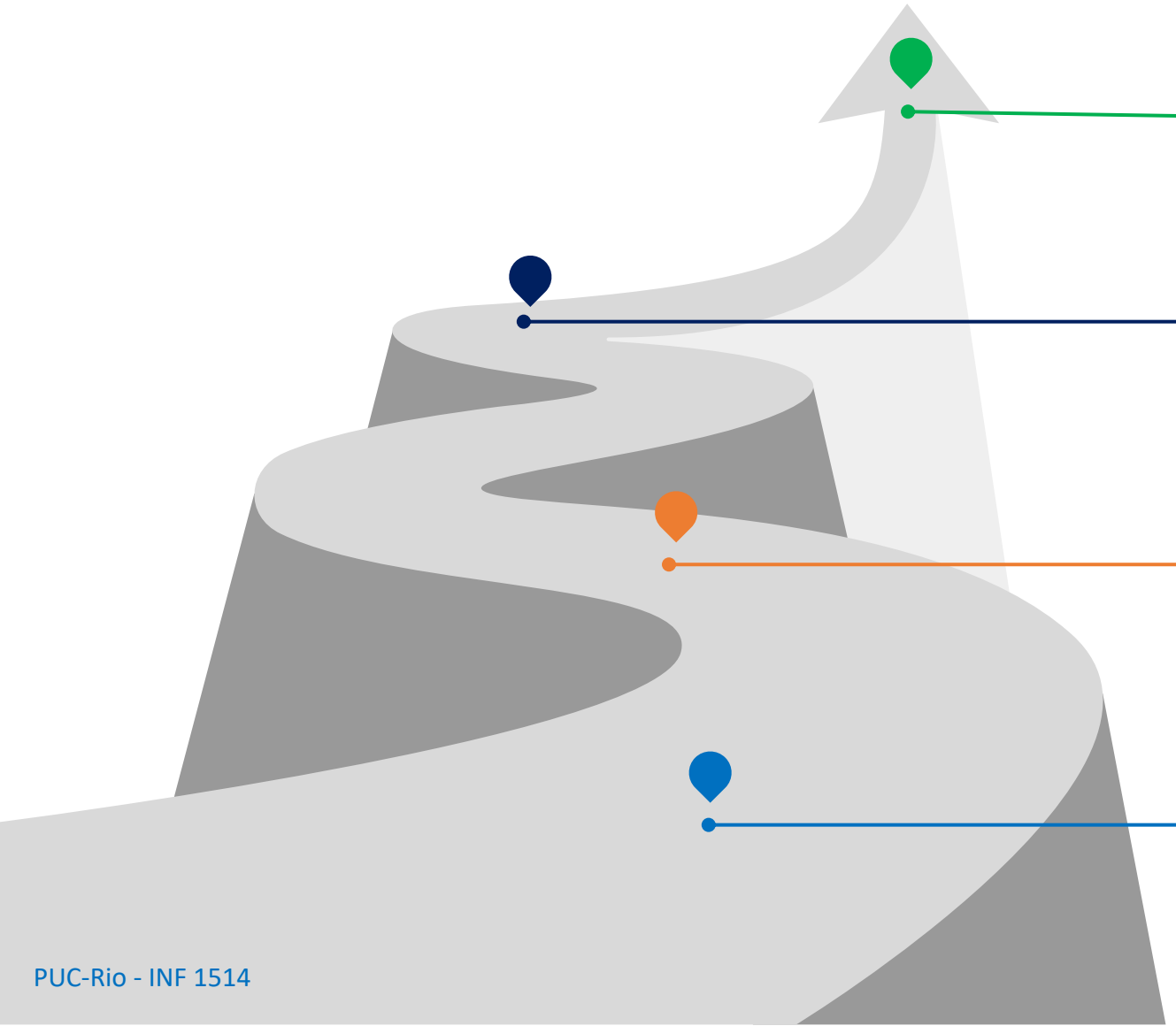
Onde estamos

Obtendo uma visão geral do que é Análise de Dados e de sua importância, bem como a sua relação com o “saber programar”.

Onde chegaremos

Ter a capacidade de implementar pequenos projetos de Análise de Dados utilizando as linguagens de programação R e Python.

Nossa trilha



04 Junho

Implementar algoritmos mais complexos relacionados à Análise de Dados.

Segunda Avaliação da disciplina.

03 Maio

Implementar algoritmos mais complexos para manipulação, tratamento e visualização de dados utilizando fontes de dados conhecidas.

Primeira Avaliação da disciplina.

Terceiro e Quarto Testes da disciplina.

02 Abril

Implementar em R e Python algoritmos básicos para manipulação e tratamento de dados.

Segundo Teste da disciplina.

01 Março

Entender o processo de Análise de Dados e conceitos básicos de programação em R.

Primeiro Teste da disciplina.

O que é linguagem de programação de alto nível?

Linguagens de programação de alto nível

- São compostas de símbolos mais complexos, inteligível pelo ser humano e não-executável diretamente pela máquina.

```
print("Bom dia!")
```

Linguagens de programação de baixo nível

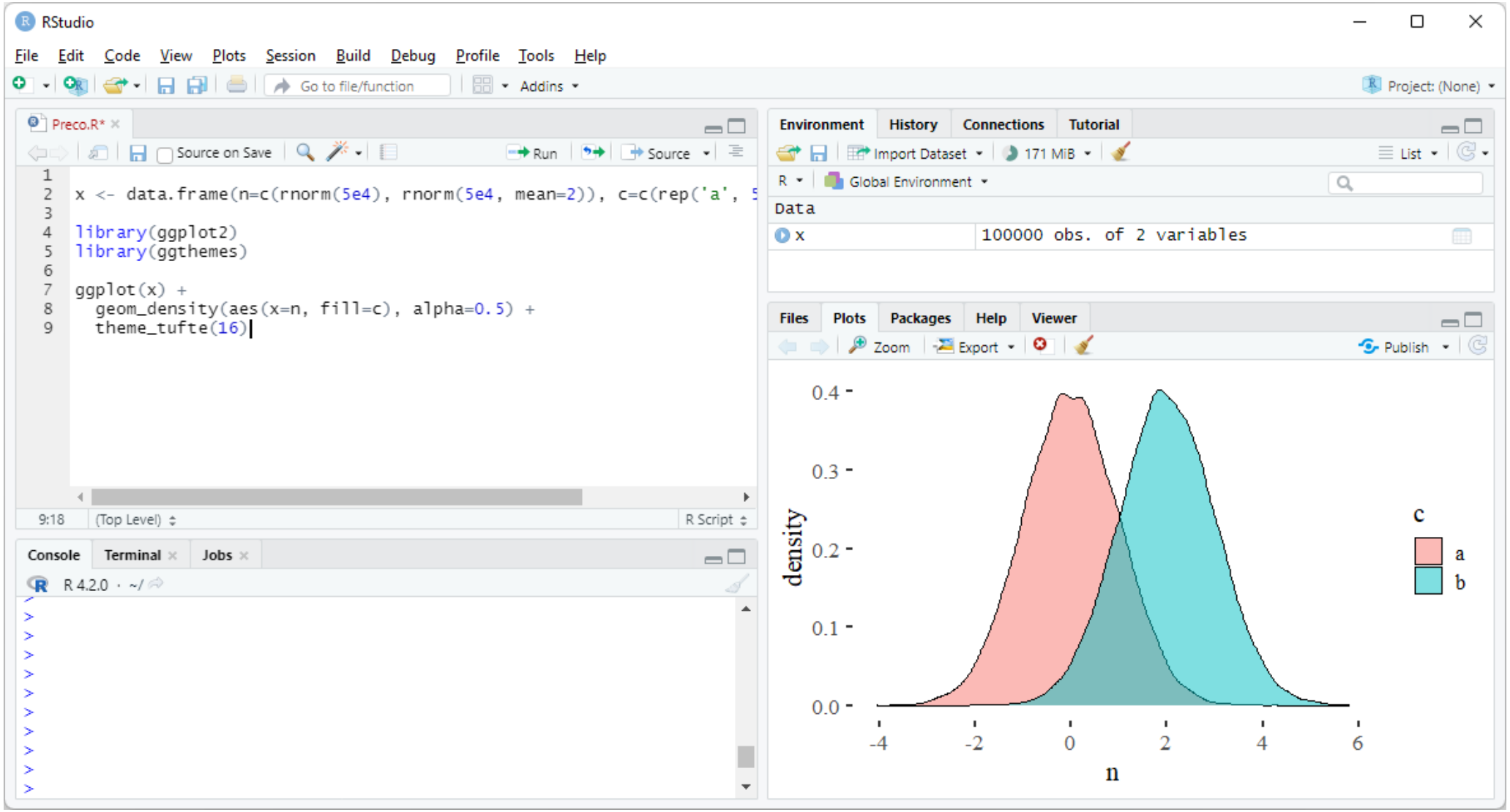
- Seus símbolos são uma representação direta do código de máquina.

```
section .data
        hello:     db 'Bom dia!',10
        helloLen:  equ $-hello

section .text
        global _start

_start:
        mov eax,4
        mov ebx,1
        mov ecx,hello
        mov edx,helloLen
        int 80h
        mov eax,1
        mov ebx,0
        int 80h;
```

R

- R é uma **linguagem** e um **ambiente de desenvolvimento** integrado.
- Foi criada originalmente por Ross Ihaka e por Robert Gentleman na universidade de Auckland, Nova Zelândia, e foi desenvolvida por um **esforço colaborativo** de pessoas em vários locais do mundo.
- A sua estrutura de **código aberto** e de **software público** e **gratuito** atraiu um grande número de desenvolvedores, sendo que há hoje inúmeros pacotes para o R.
- O R disponibiliza uma **ampla variedade de técnicas estatísticas e gráficas**, incluindo modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento e outras.
- Tem uma **história longa e confiável** e uma **forte comunidade** de suporte no setor de dados.

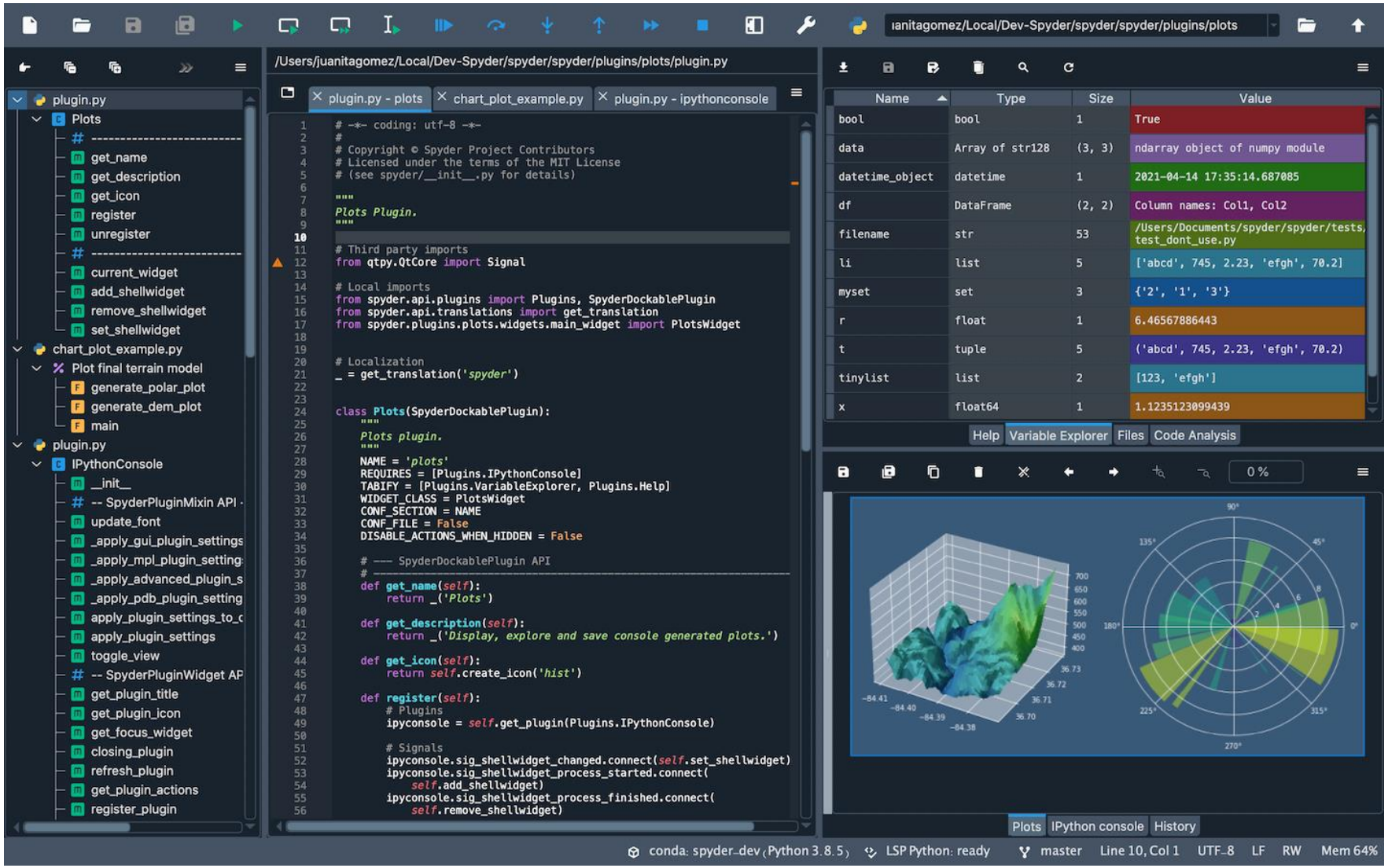
R Studio


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Preco.R*
Source on Save
Run
Source
1 x <- data.frame(n=c(rnorm(5e4), rnorm(5e4, mean=2)), c=c(rep('a',
2
3 library(ggplot2)
4 library(ggthemes)
5
6 ggplot(x) +
7 geom_density(aes(x=n, fill=c), alpha=0.5) +
8 theme_tufte(16)
9
9:18
(Top Level)
R Script
Console
Terminal
Jobs
R 4.2.0 · ~/
Environment
History
Connections
Tutorial
Import Dataset
171 MiB
List
R
Global Environment
Data
x
100000 obs. of 2 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
6
n
c
a
b

Python

- Foi lançada por Guido van Rossum em 1991 e, atualmente, possui um **modelo de desenvolvimento comunitário**, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation.
- É uma **linguagem de programação multiplataforma** que permite desenvolver aplicações para games, desktops, web e dispositivos móveis sendo também muito utilizada em aplicações que lidam com processamento de texto e machine learning.
- Combina uma sintaxe concisa e clara com os recursos poderosos de sua **biblioteca padrão** e por **módulos e frameworks** desenvolvidos por terceiros.
- A **facilidade de integração** com outras linguagens também é um fator de destaque.

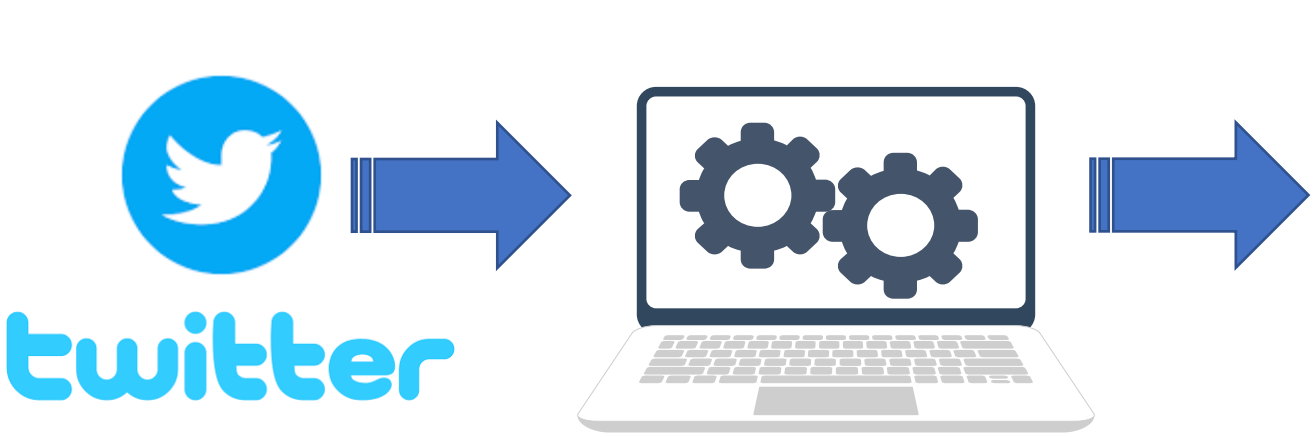
Spyder


/Users/juanitagomez/Local/Dev-Spyder/spyder/spyder/plugins/plots/plugin.py
plugin.py - plots
chart_plot_example.py
plugin.py - ipythonconsole
Name
Type
Size
Value
Help
Variable Explorer
Files
Code Analysis
Plots
IPython console
History
conda: spyder-dev (Python 3.8.5)
LSP Python: ready
master
Line 10, Col 1
UTF-8
LF
RW
Mem 64%

Por que vamos ver R e Python?

- De acordo com as características do projeto uma “linguagem” pode ser **mais indicada do que a outra** ou mesmo podem ser **utilizadas em conjunto** aproveitando-se o melhor de cada uma.
- **Traçar paralelos** (prós e contras) entre linguagens demonstrando de forma prática diferentes abordagens, conceitos e recursos que podem ser encontrados em outros ambientes e linguagens.
- Apesar de existirem muitas outras possibilidades, estas duas linguagens tem **polarizado as discussões** sobre que ferramenta utilizar para Análise de Dados.
- As duas linguagens **são simples** (e gratuitas) para instalar e relativamente fáceis de começar a usar servindo como excelente ponto de partida para se aprender a programar para Análise de Dados.
- Vamos **começar por R** por apresentar menor curva de aprendizado.

Exemplo em R

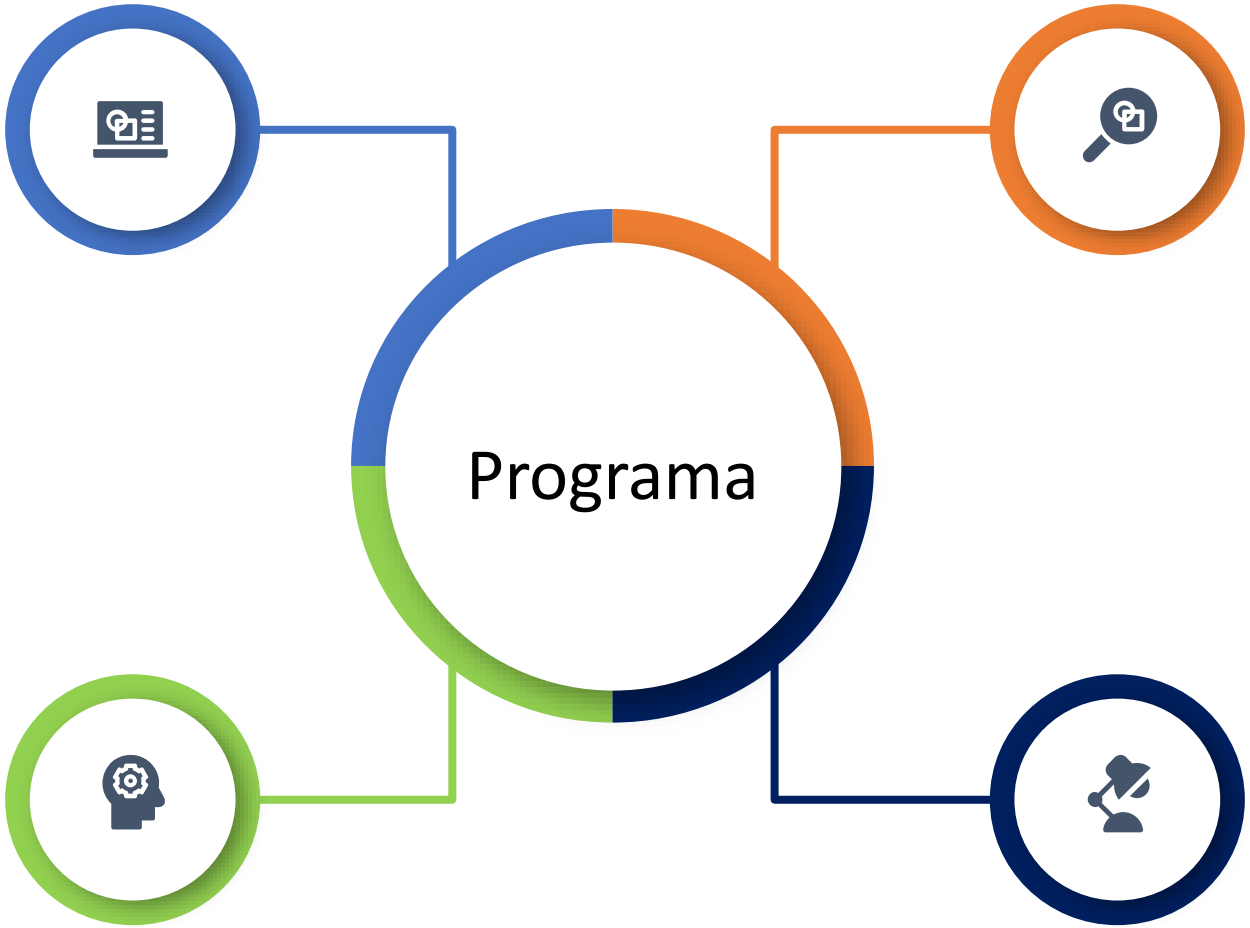

twitter


melhor
website
loja
simples
lista
clientes
contato
empresa
fontes
cidade
pessoas
oficiais
forma
escolha
cliente
através
ser
todos
confiáveis
informações
produtos
filtro
formulário
informe
sobre
Internet
mensagem
endereço
hoax
tempo
medida
apenas
Clique
Onbiz
e-mail
Newsletter
Google
serviços

INF 1514

Base de programação

Conceitos básicos; importação e exportação de dados; repetição; condicionais; funções.



Visualização de dados

Dados temporais; espaciais; espaço-temporais.

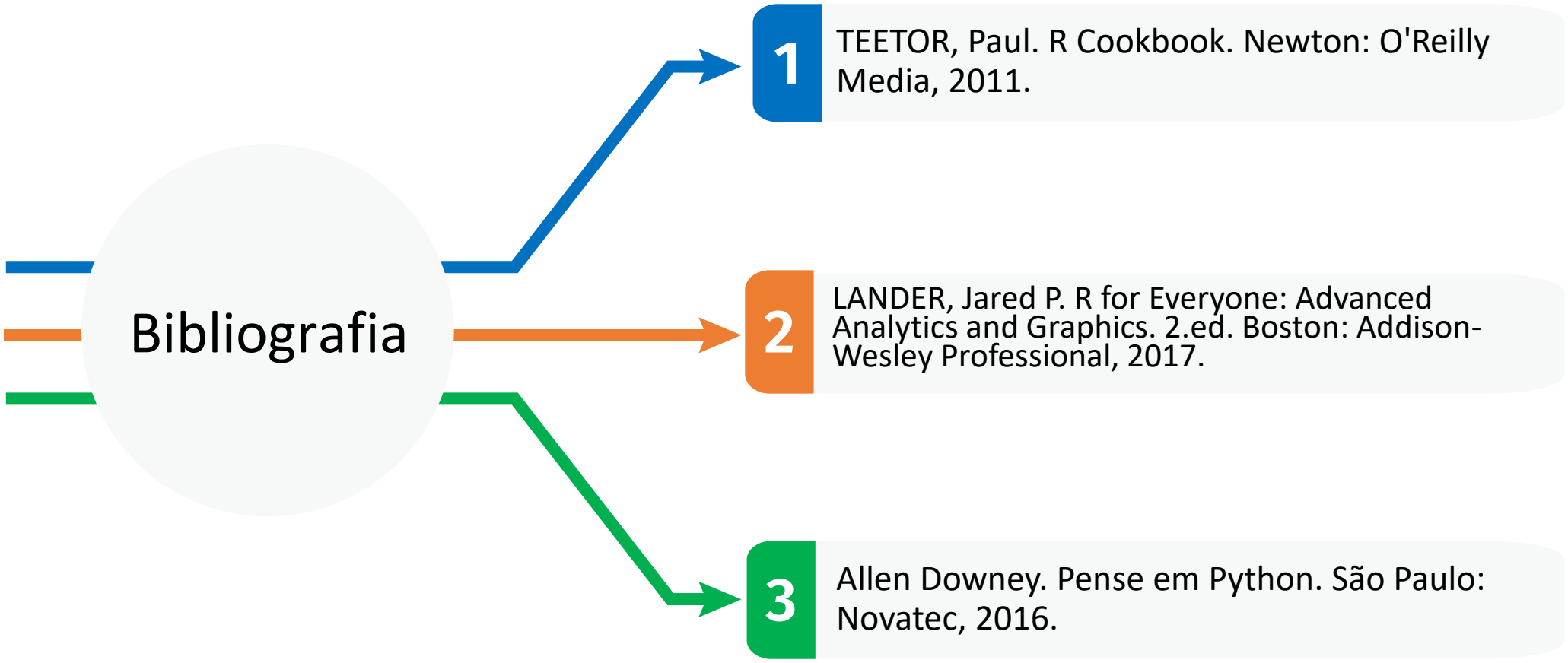
Manipulação de dados

Subconjuntos; ordenação; transformações; merge; fontes de dados.

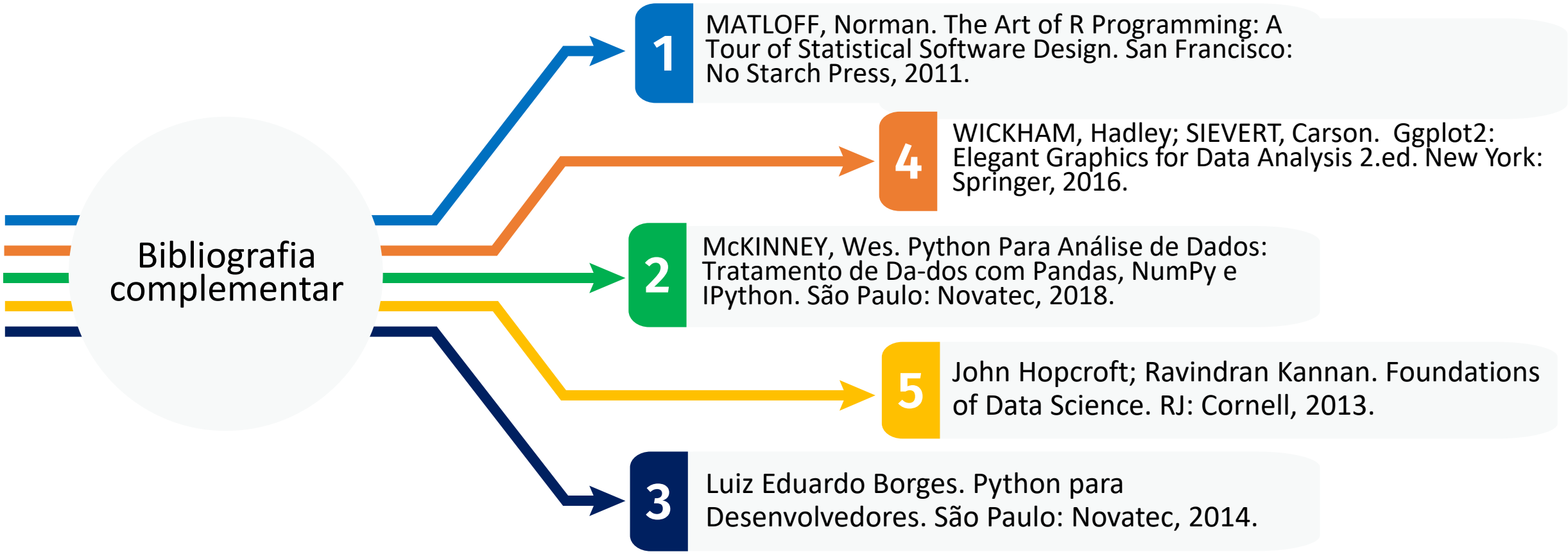
Outros tópicos

Web scraping; organização de código; documentação; versionamento de código.

INF 1514


Bibliografia
1
TEETOR, Paul. R Cookbook. Newton: O'Reilly Media, 2011.
2
LANDER, Jared P. R for Everyone: Advanced Analytics and Graphics. 2.ed. Boston: Addison-Wesley Professional, 2017.
3
Allen Downey. Pense em Python. São Paulo: Novatec, 2016.

INF 1514


Bibliografia complementar
1
MATLOFF, Norman. The Art of R Programming: A Tour of Statistical Software Design. San Francisco: No Starch Press, 2011.
4
WICKHAM, Hadley; SIEVERT, Carson. Ggplot2: Elegant Graphics for Data Analysis 2.ed. New York: Springer, 2016.
2
McKINNEY, Wes. Python Para Análise de Dados: Tratamento de Da-dos com Pandas, NumPy e IPython. São Paulo: Novatec, 2018.
5
John Hopcroft; Ravindran Kannan. Foundations of Data Science. RJ: Cornell, 2013.
3
Luiz Eduardo Borges. Python para Desenvolvedores. São Paulo: Novatec, 2014.

Critério de Avaliação

- Critério 3.
- Nota da G1:
  - Teste 1, Teste 2 e Prova.
  - $G1 = \left(\frac{Teste\ 1 + Teste\ 2}{2}\right) * 0{,}4 + Prova * 0{,}6$
- Nota da G2:
  - Teste 1, Teste 2 e Trabalho.
  - $G1 = \left(\frac{Teste\ 1 + Teste\ 2}{2}\right) * 0{,}4 + Trabalho * 0{,}6$
- Média Final (MF):
  - $MF = \left(\frac{G1 + G2}{2}\right)$
  - $MF \geq 5$, aprovado.