APRENDENDO COM OS DADOS

UMA ABORDAGEM DE CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA UTILIZANDO R

(PARTE 1)

Prof. Rafael G. Mantovani

03/09/2019

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Dois Vizinhos - PR, Brasil
Setembro, 2019

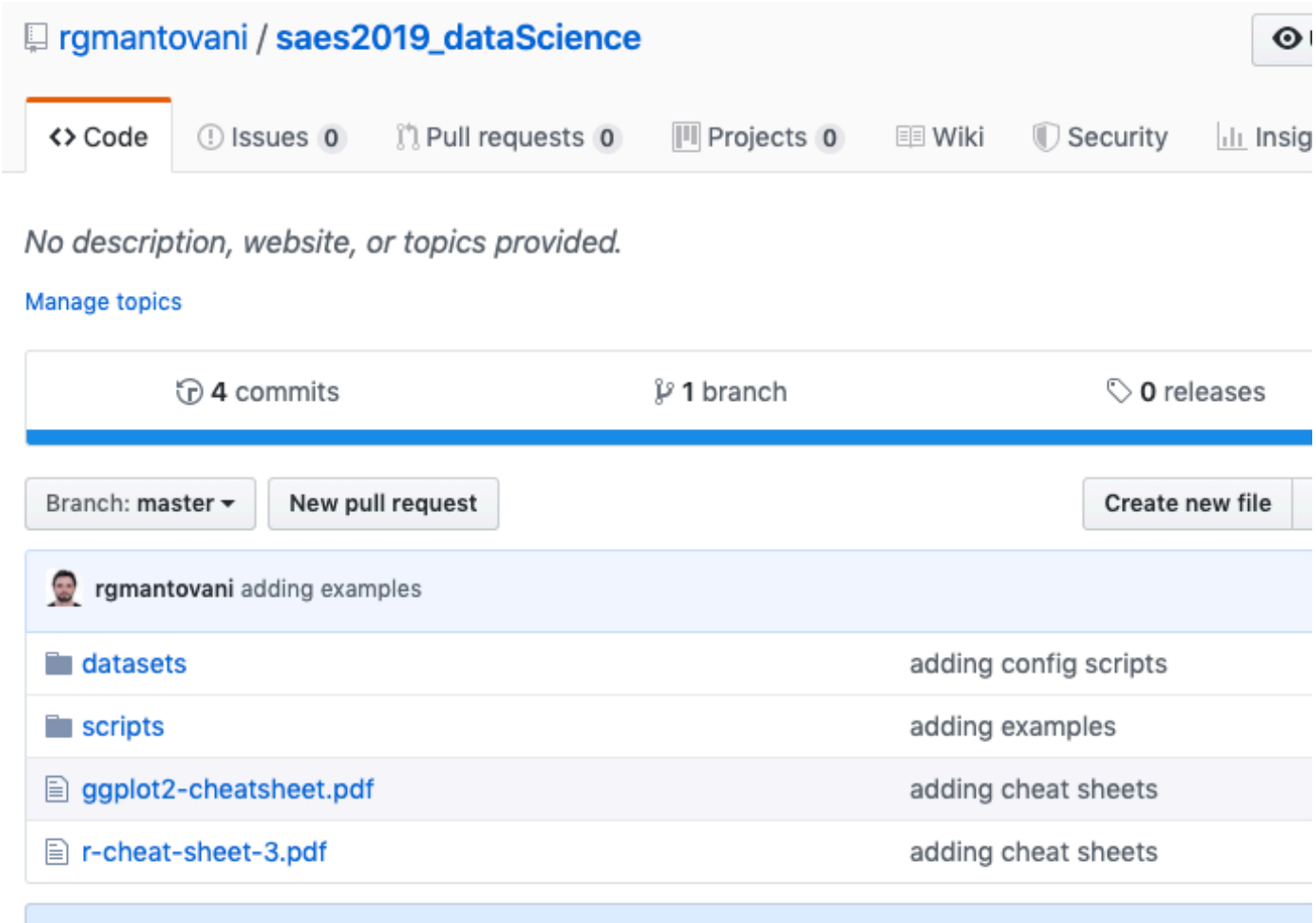
Universidade Tecnológica Federal do Paraná (UTFPR)
IV Semana Acadêmica do curso de Engenharia de Software

Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. Welcome to R :)
6. Referências

Material

Link: https://github.com/rgmantovani/saes2019_dataScience



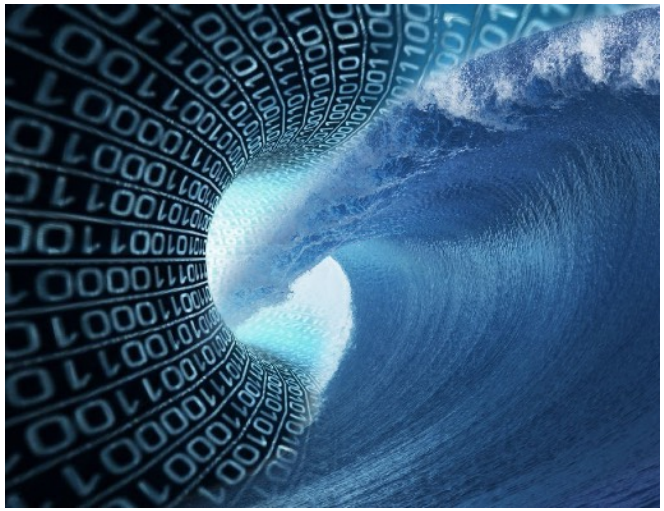
Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. Welcome to R :)
6. Referências

Introdução

poucos dados

Introdução

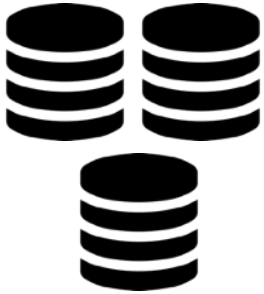


poucos dados

imensa quantidade de dados (big data)

Introdução

Pessoas

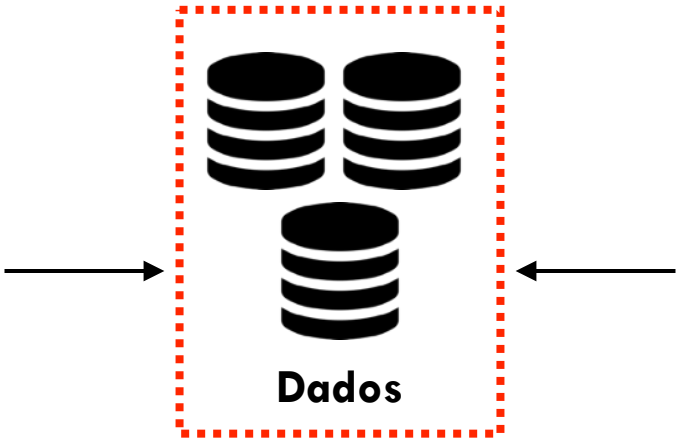


Dados

Dispositivos

Introdução

Pessoas



Dispositivos

- Dados são **continuamente**:
  - gerados, coletados, processados e transmitidos

Introdução

- Mudança de realidade

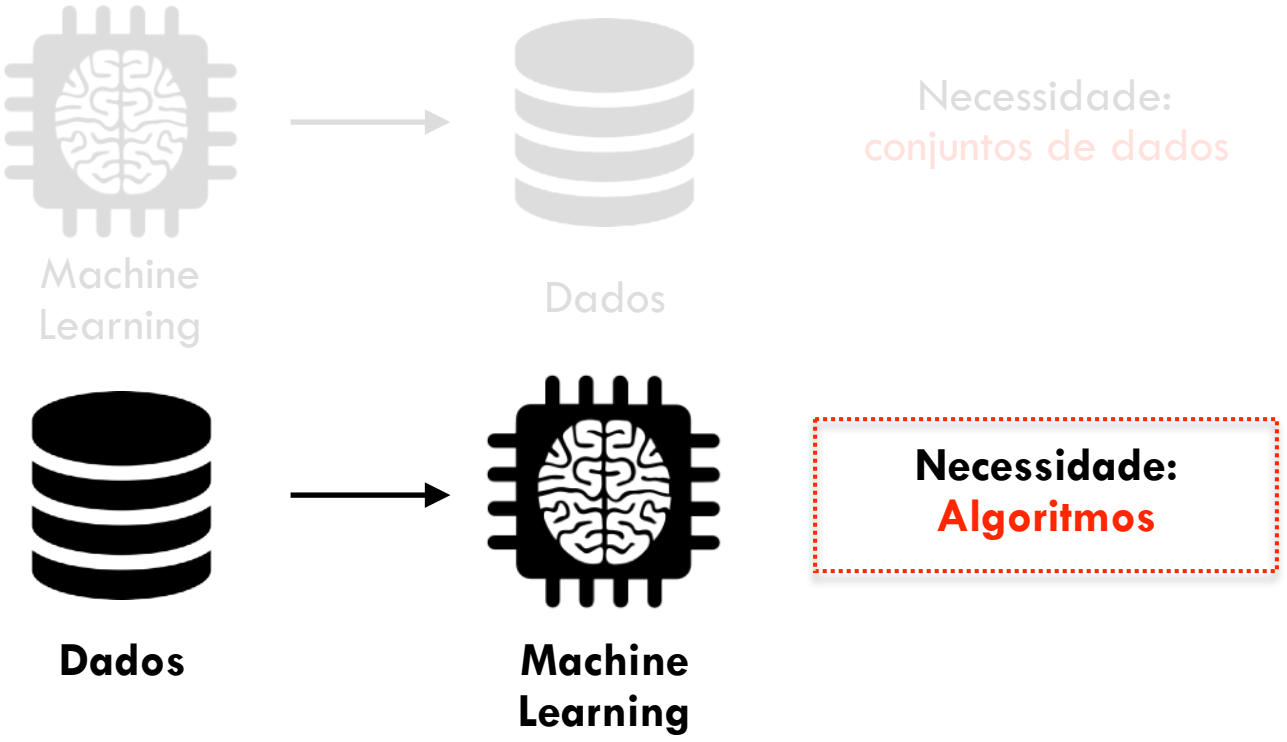
Machine Learning

Dados

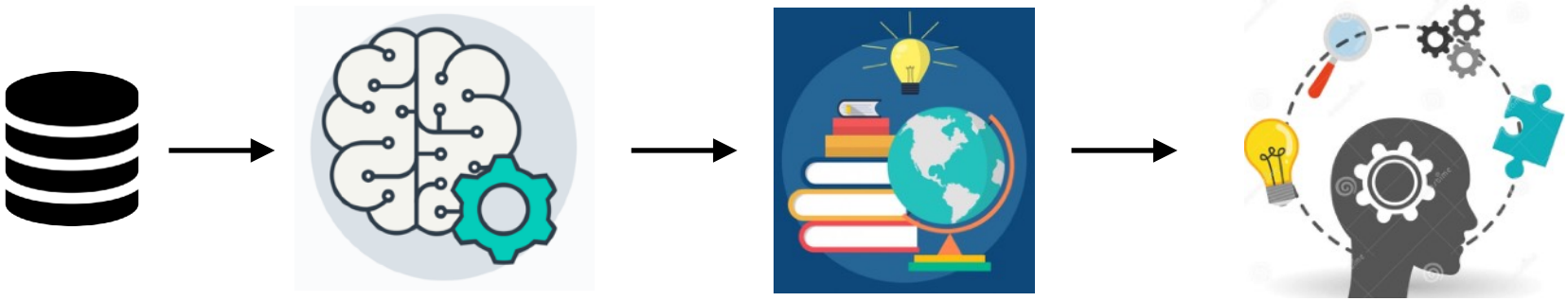
Necessidade:
conjuntos de dados

Introdução

- Mudança de realidade



Introdução



Dados → **Algoritmo AM** → **Conhecimento** → **Solução**

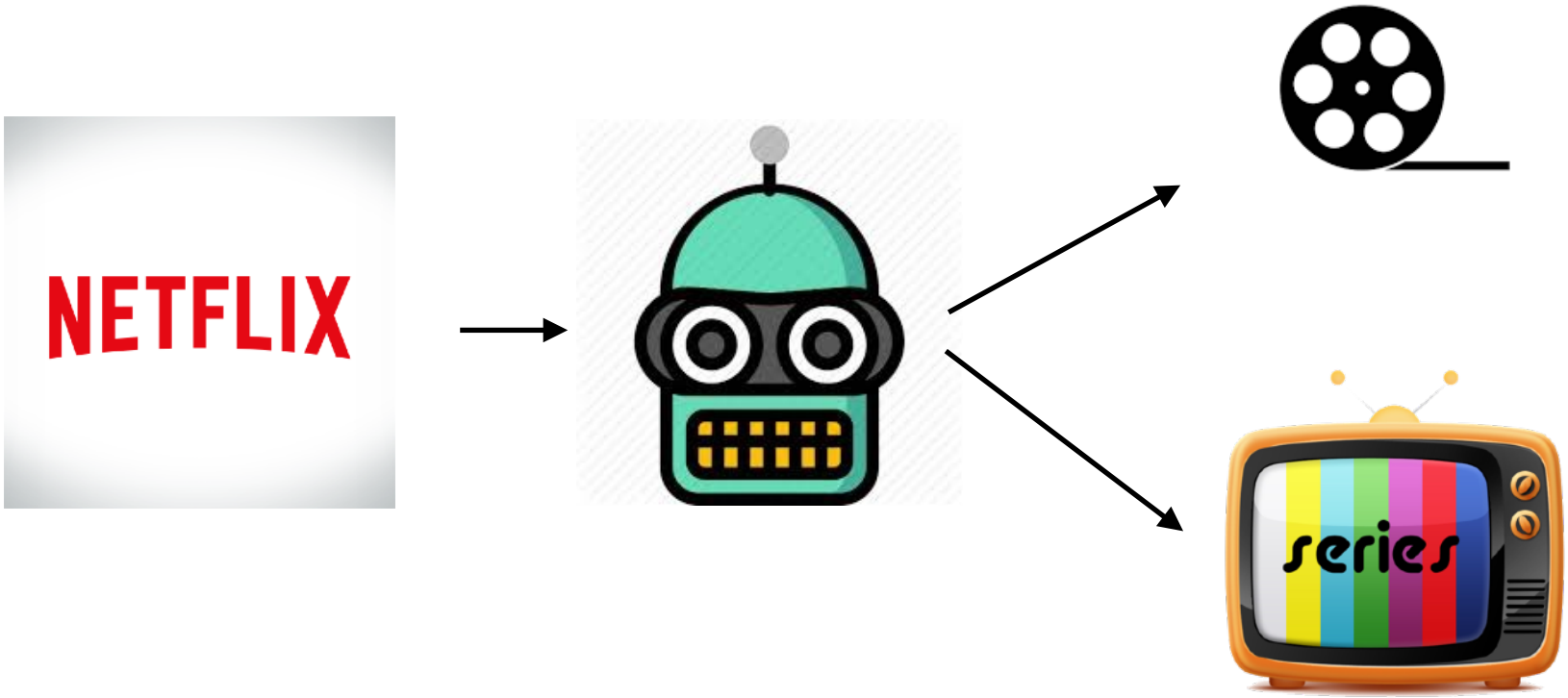
- Inteligência Artificial
- Automatiza a construção de modelos para solucionar problemas!

Introdução

- Onde isso é usado?

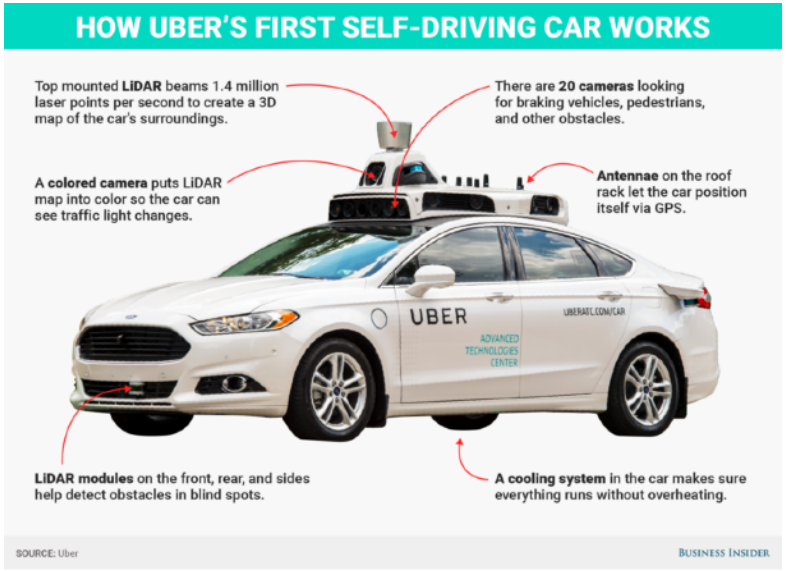
Introdução

- Onde isso é usado?



Introdução

- Onde isso é usado? **Veículos Autônomos**



Uber

Tesla

Introdução

- Onde isso é usado? **Veículos Autônomos**

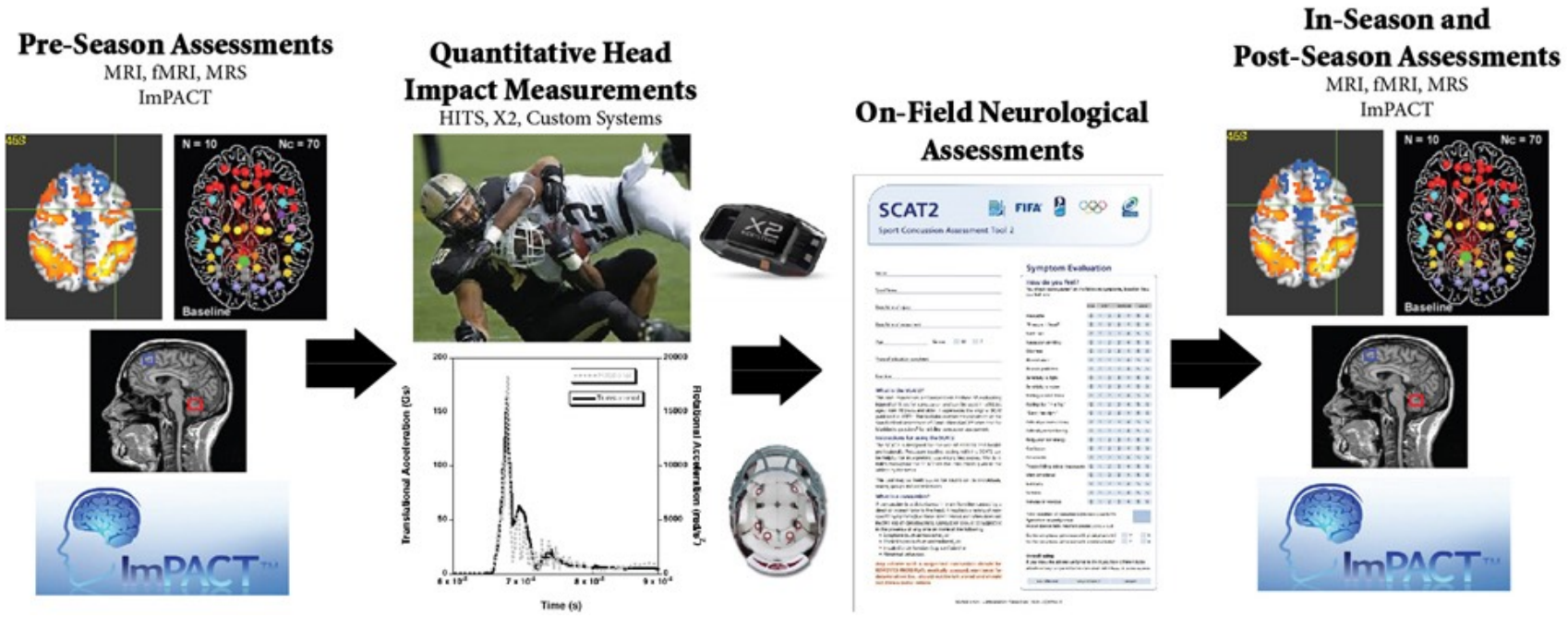
LRM - ICMC/USP, São Carlos - SP

Introdução

- Onde isso é usado? **Bancos**

Introdução

- Onde isso é usado? **Sistemas Médicos**



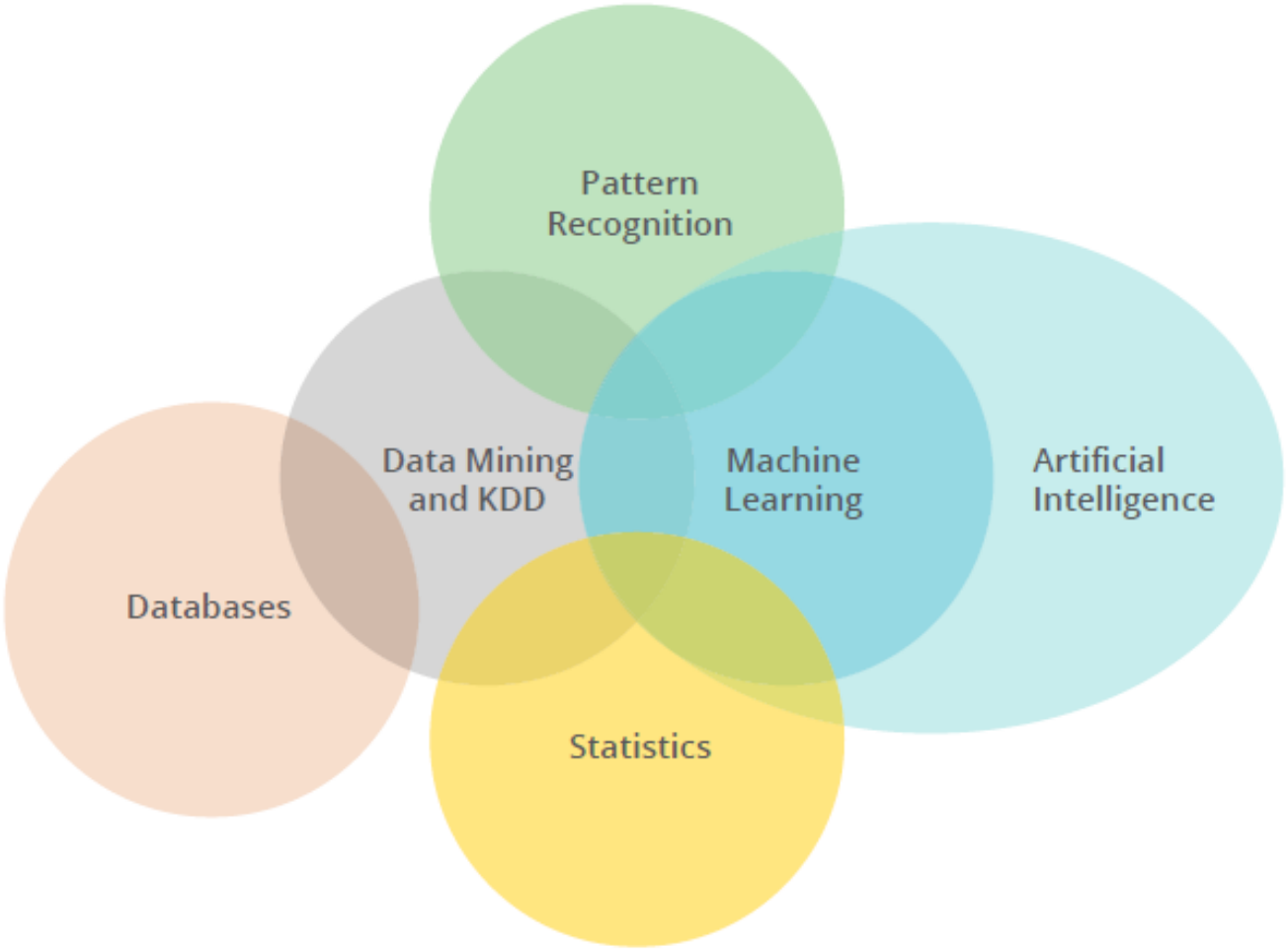
Introdução

- Onde isso é usado? **Sistemas de Segurança**

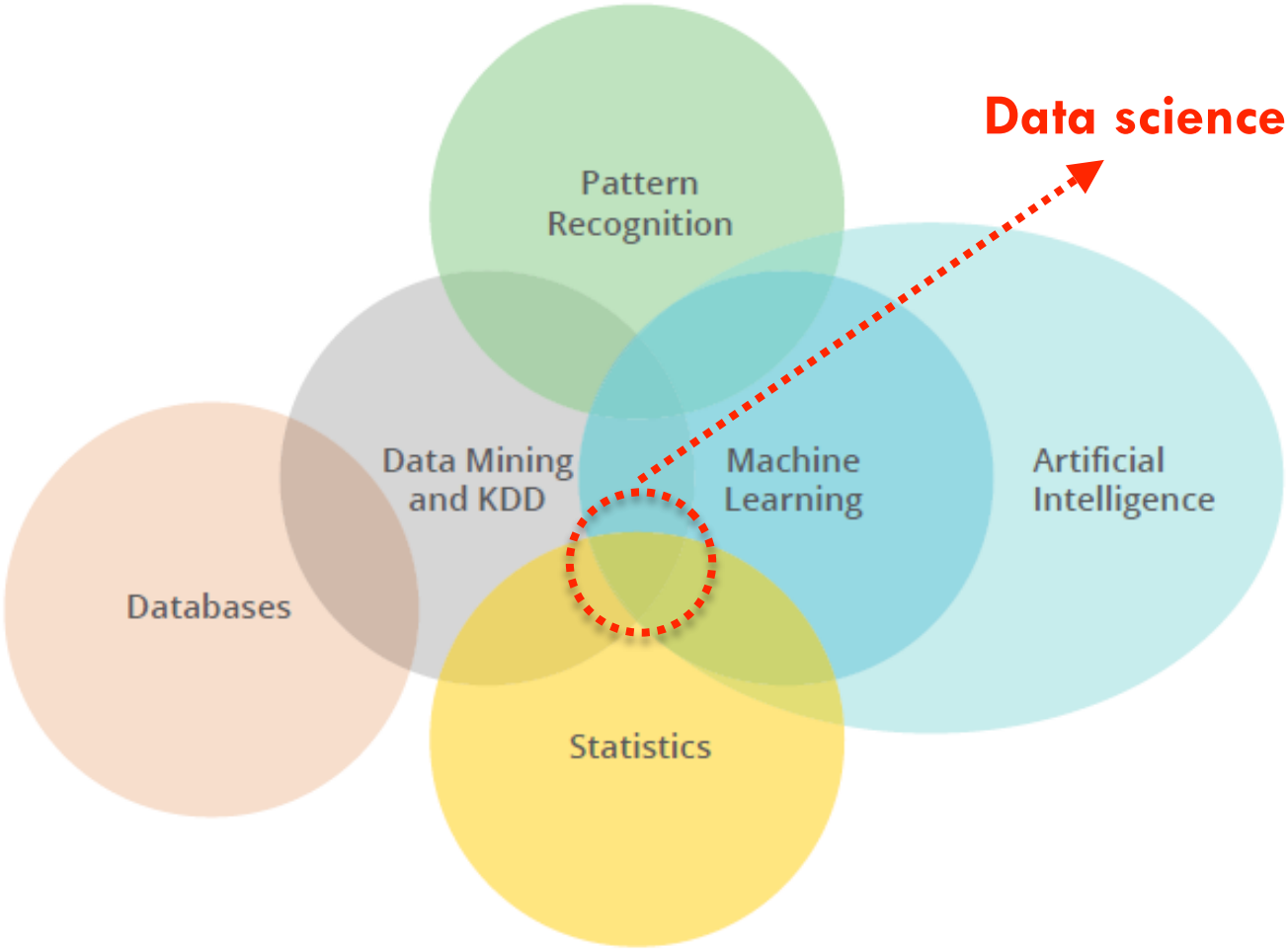
Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. Welcome to R :)
6. Referências

Conceitos Gerais



Conceitos Gerais



Mineração de dados

Aprendizado de Máquina

Matemática / Estatística

Fluxo de Ciência de Dados

- Quantos algoritmos existem?

Fluxo de Ciência de Dados

- Quantos algoritmos existem?

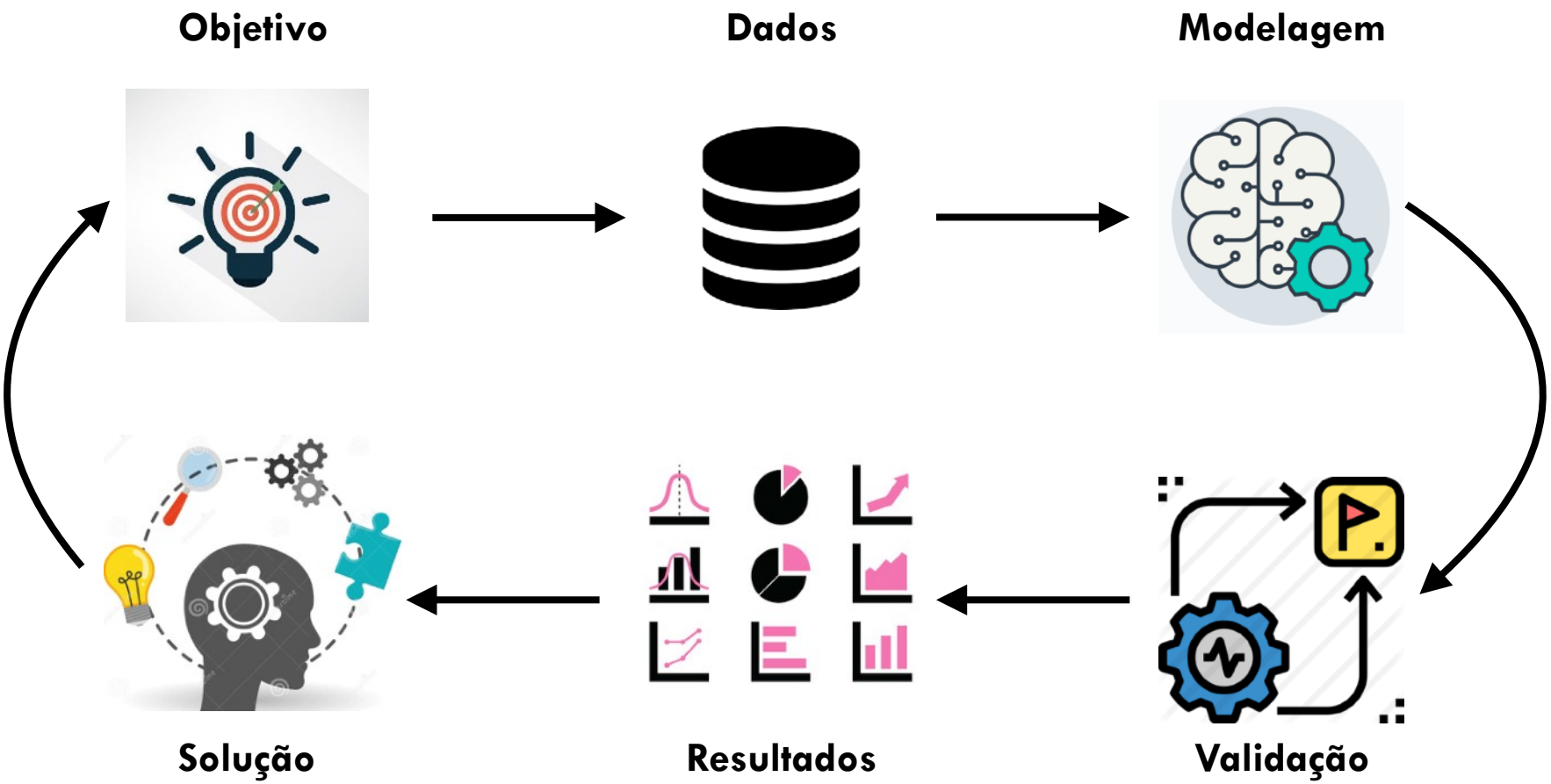
TensorFlow

...

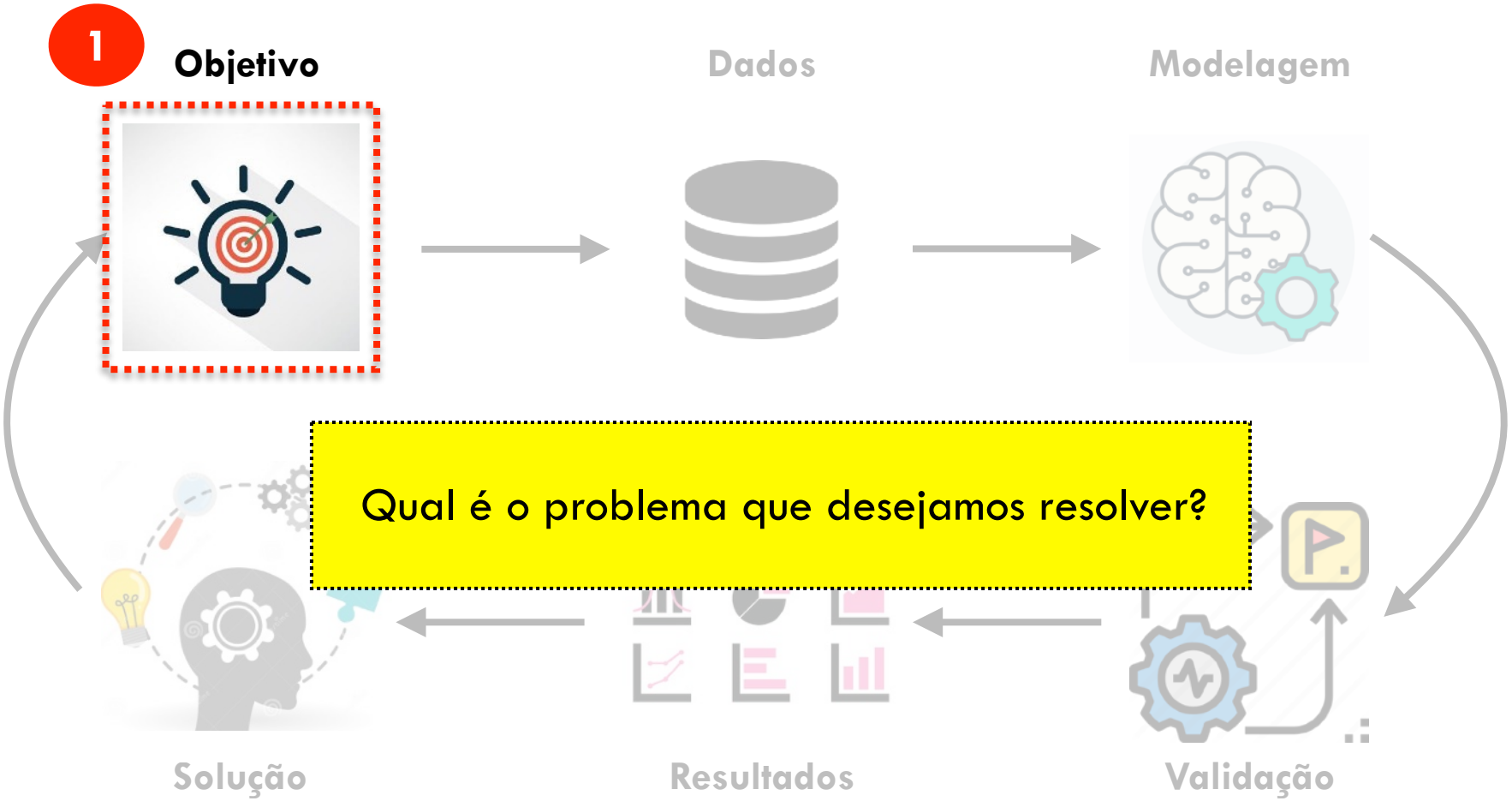
Roteiro

1. Introdução
2. Conceitos gerais
3. **Fluxo de ciência de dados**
4. Ferramentas
5. Welcome to R :)
6. Referências

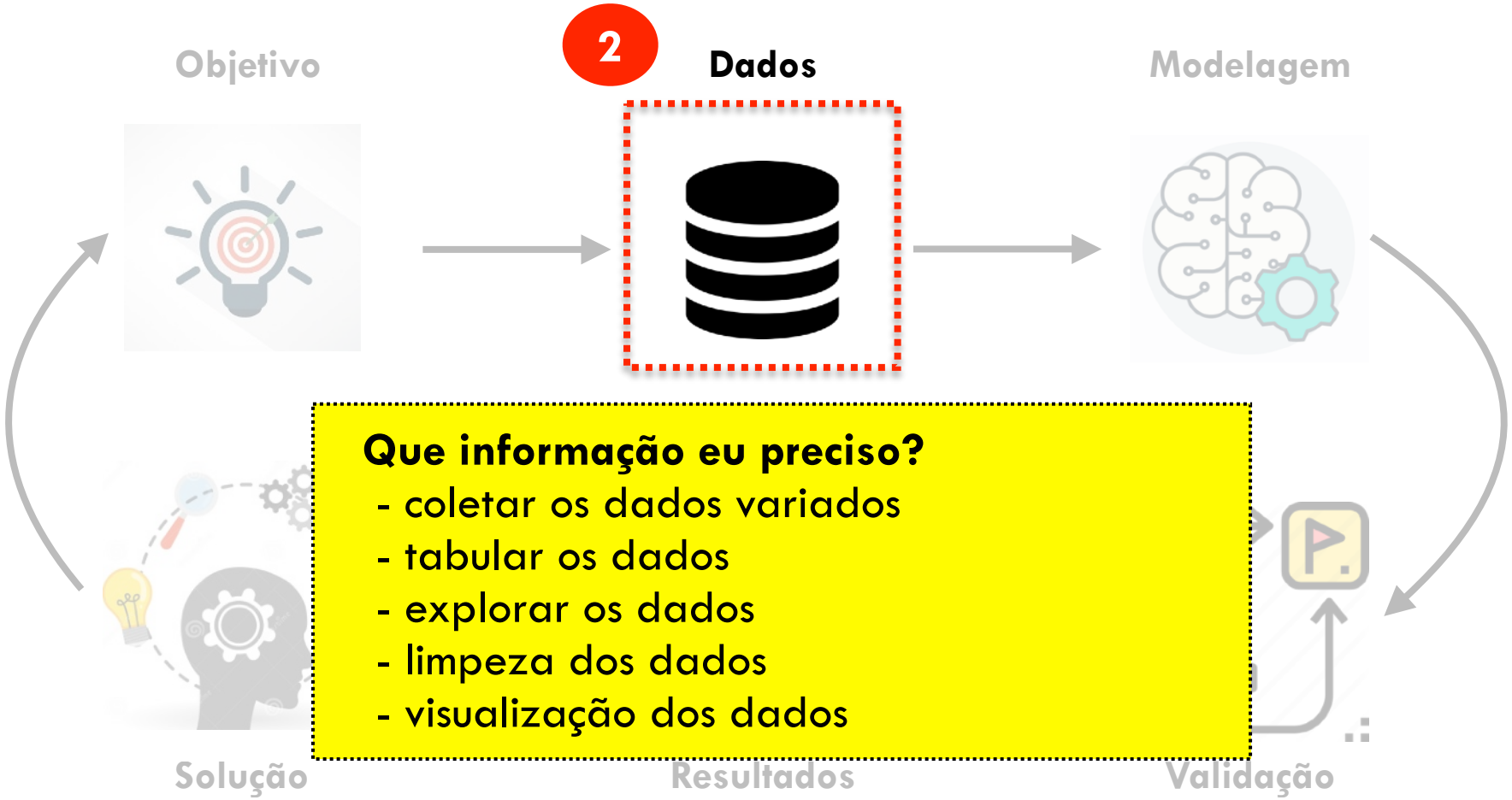
Fluxo de Ciência de Dados



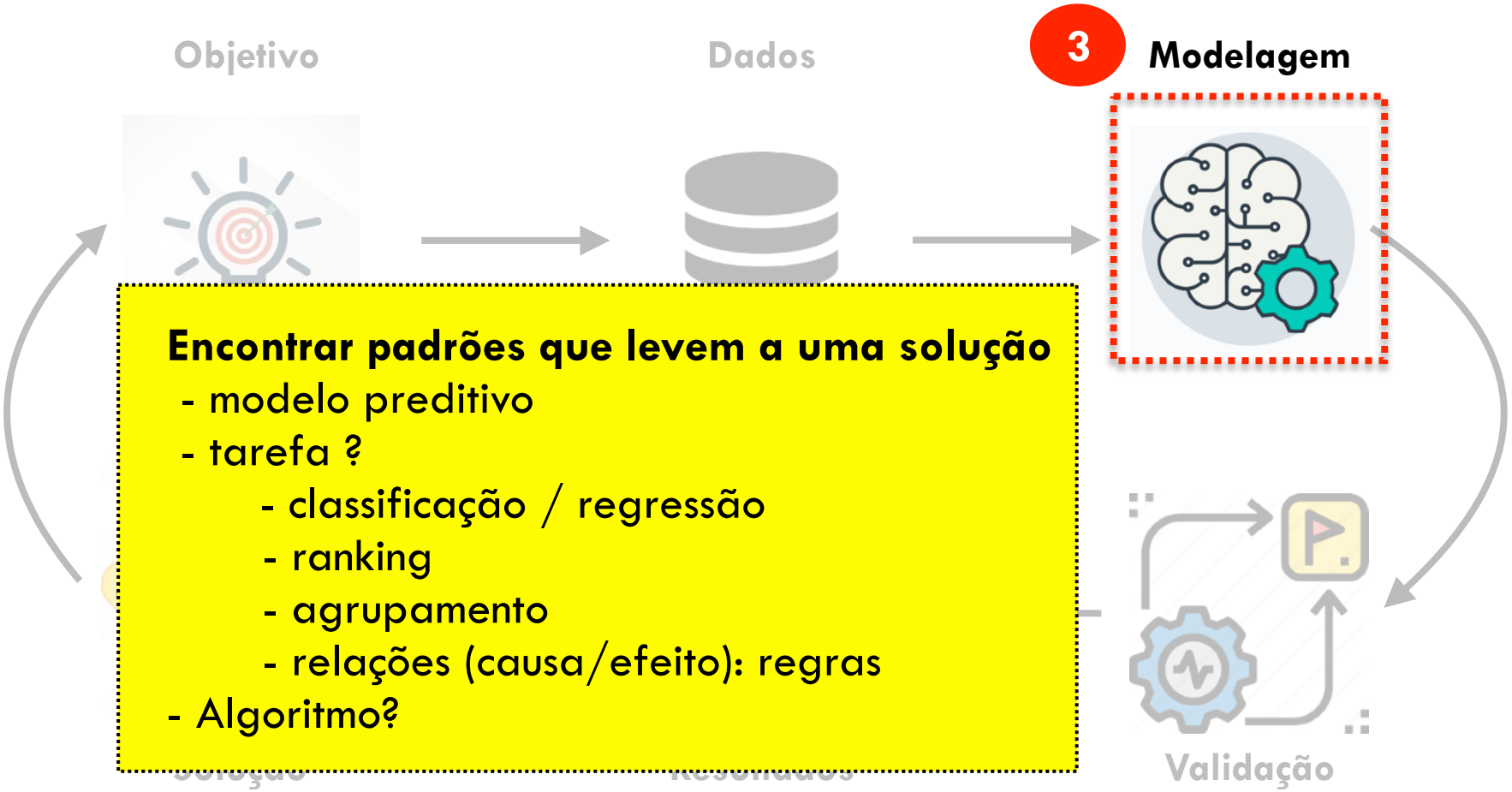
Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



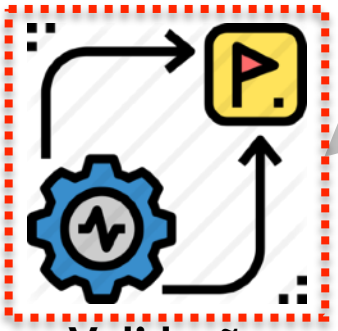
Fluxo de Ciência de Dados

Objetivo

Dados

Modelagem

O modelo que eu gerei resolve meu problema?
- modelo é preciso?
- ele generaliza?
- resultado faz sentido?



Validação

4

Resultados

Solução

Fluxo de Ciência de Dados

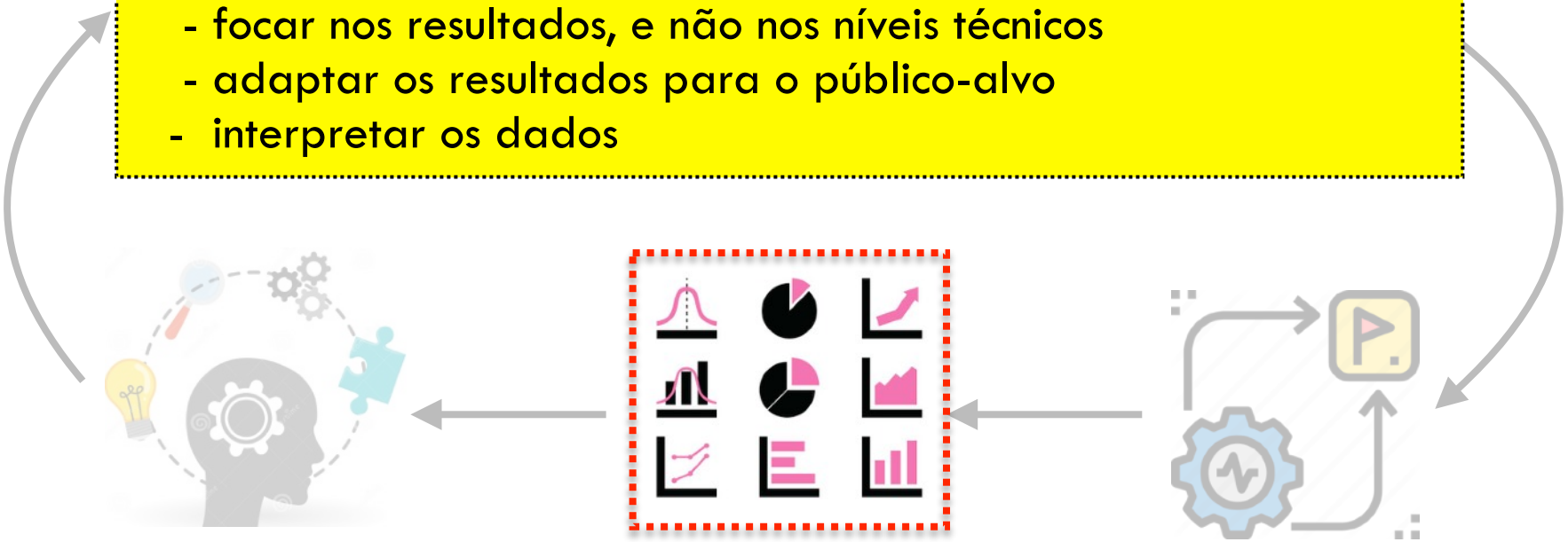
Objetivo

Dados

Modelagem

Mostrar que o modelo resolve o problema e como ele resolve

- gráficos / tabelas
- focar nos resultados, e não nos níveis técnicos
- adaptar os resultados para o público-alvo
- interpretar os dados



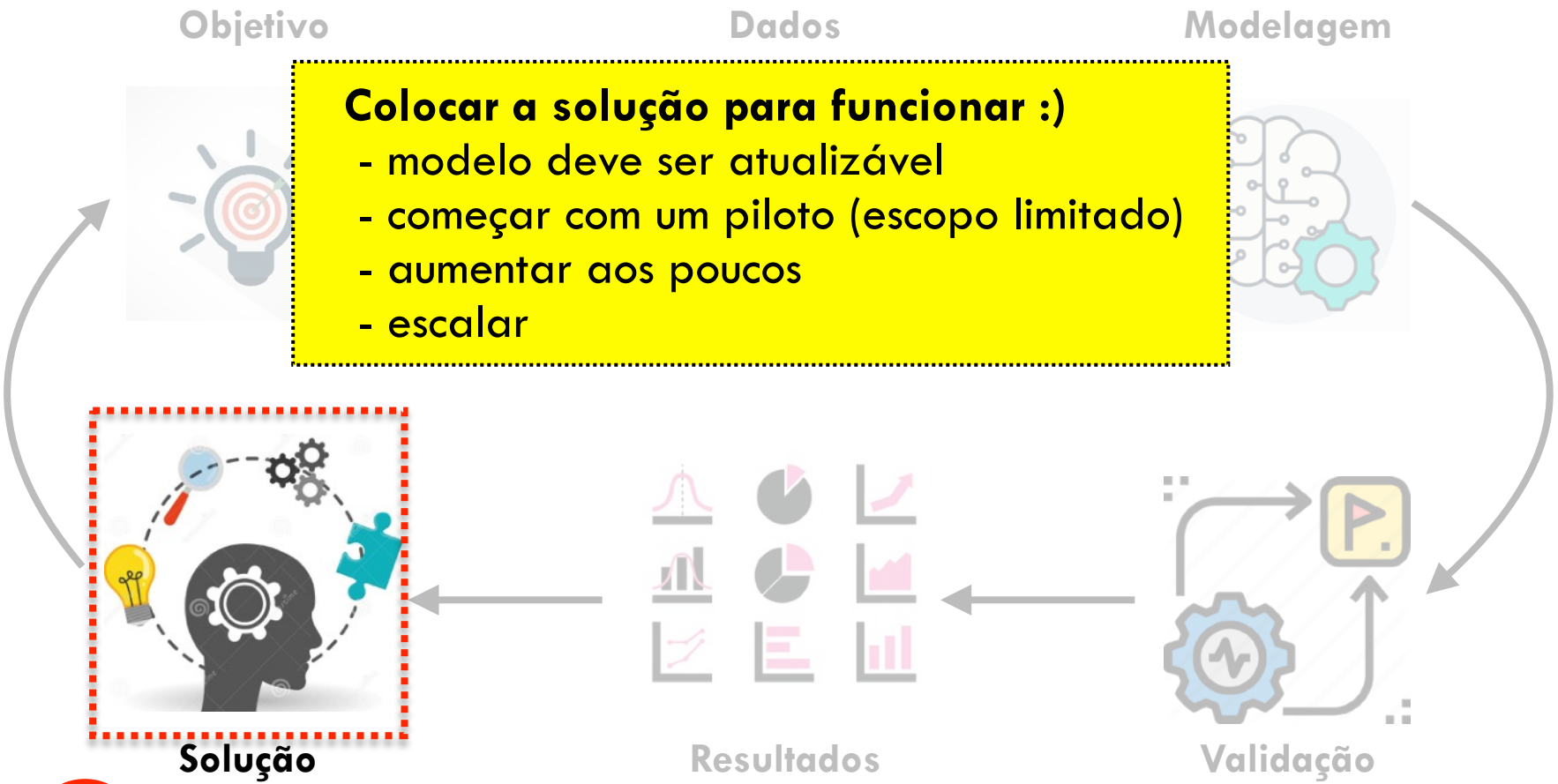
Solução

Resultados

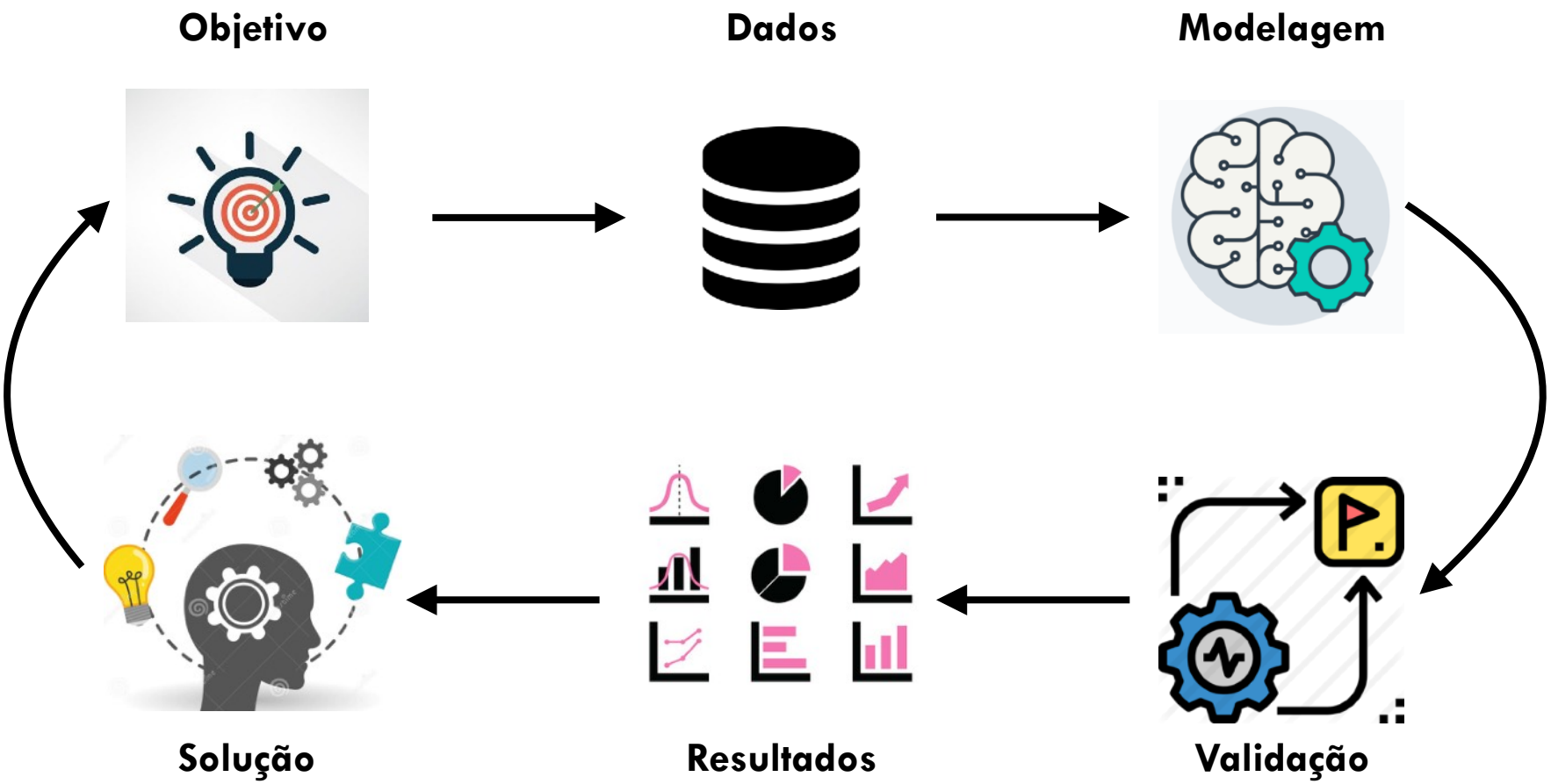
Validação

5

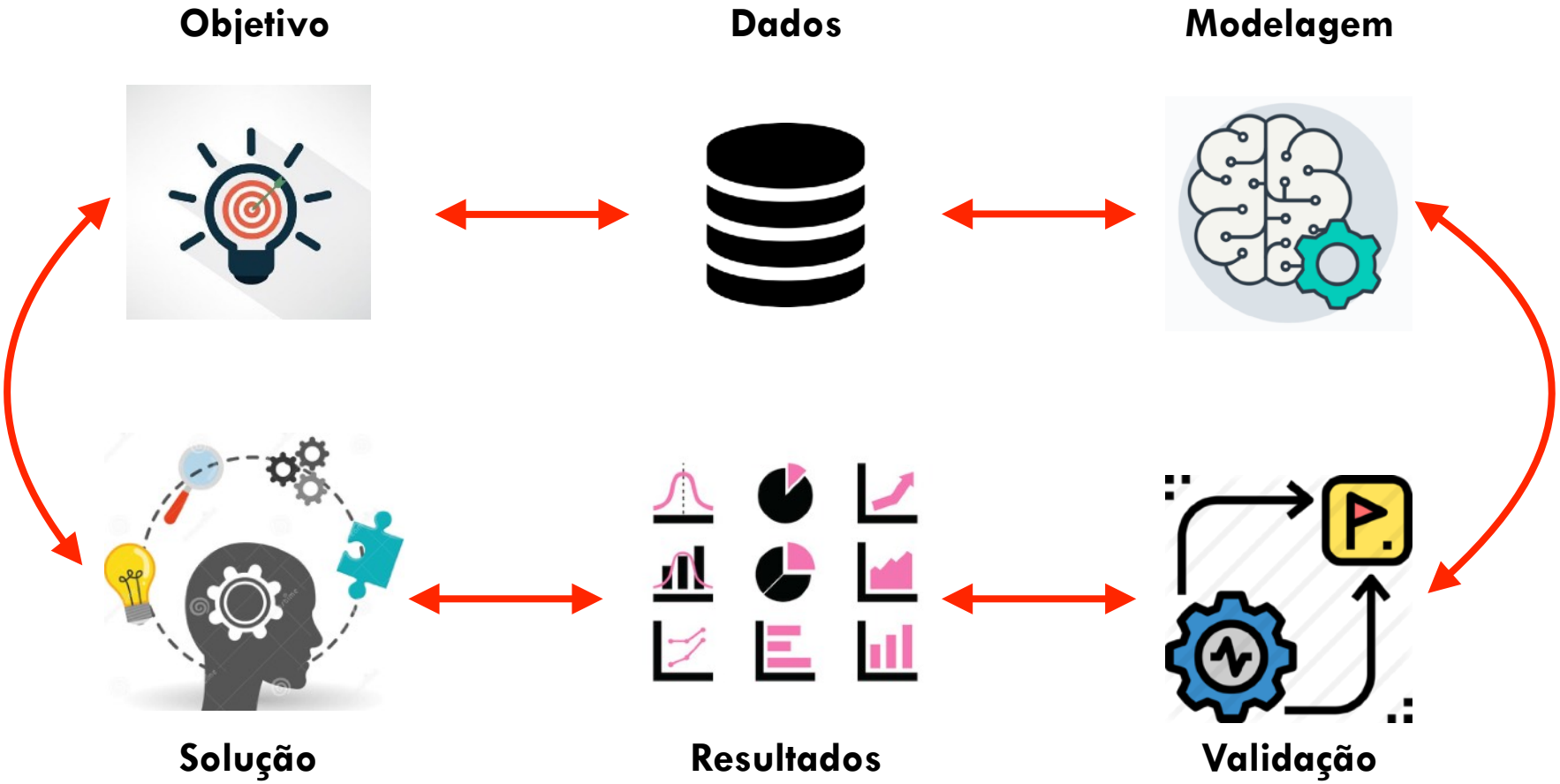
Fluxo de Ciência de Dados



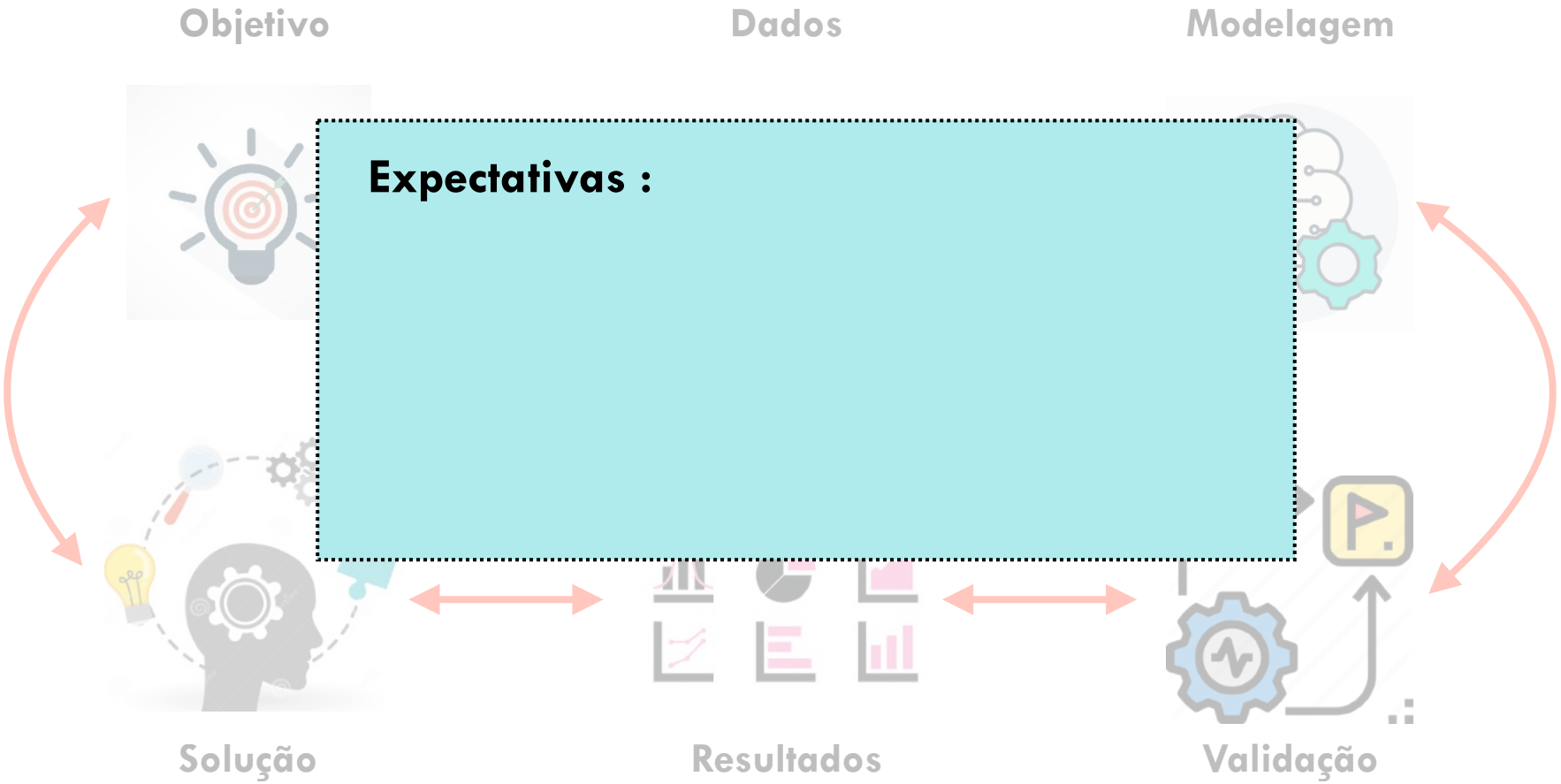
Fluxo de Ciência de Dados



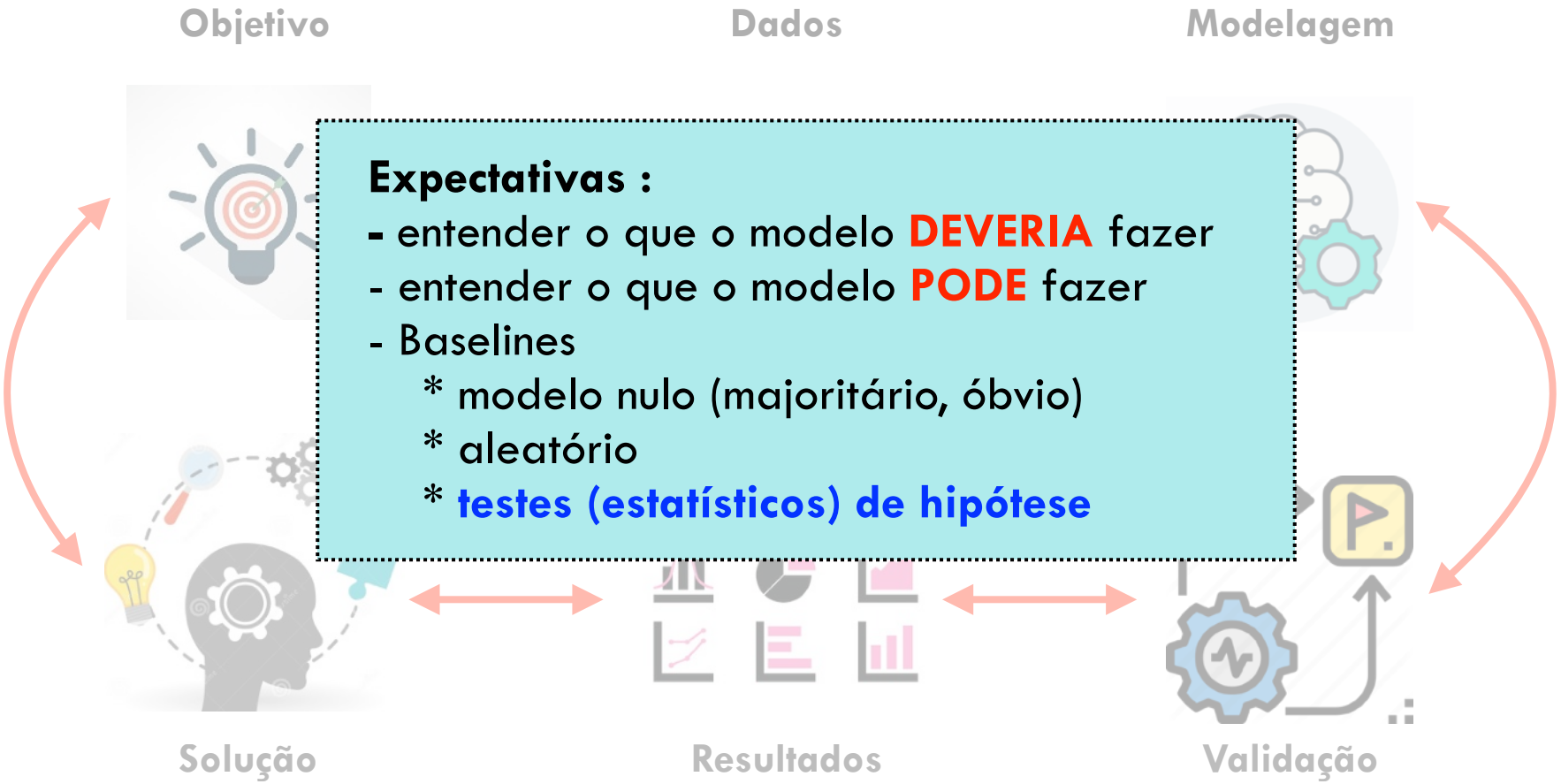
Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



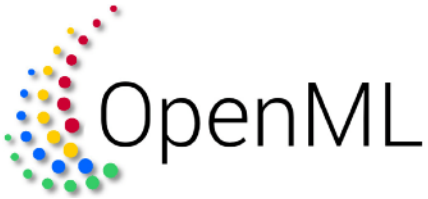
Fluxo de Ciência de Dados



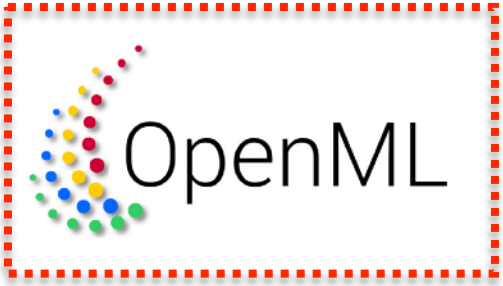
Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. **Ferramentas**
5. Welcome to R :)
6. Referências

Ferramentas

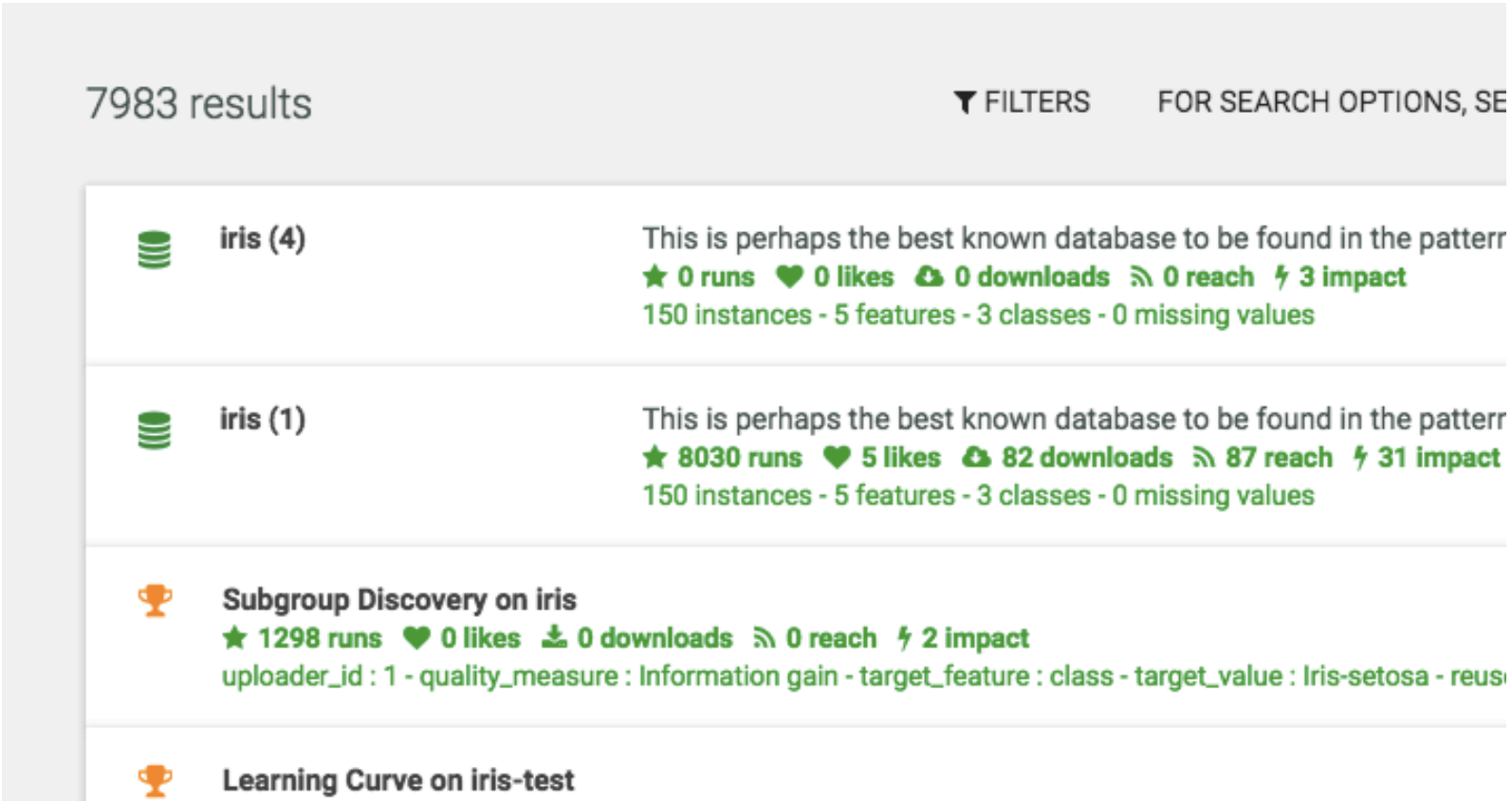


Ferramentas

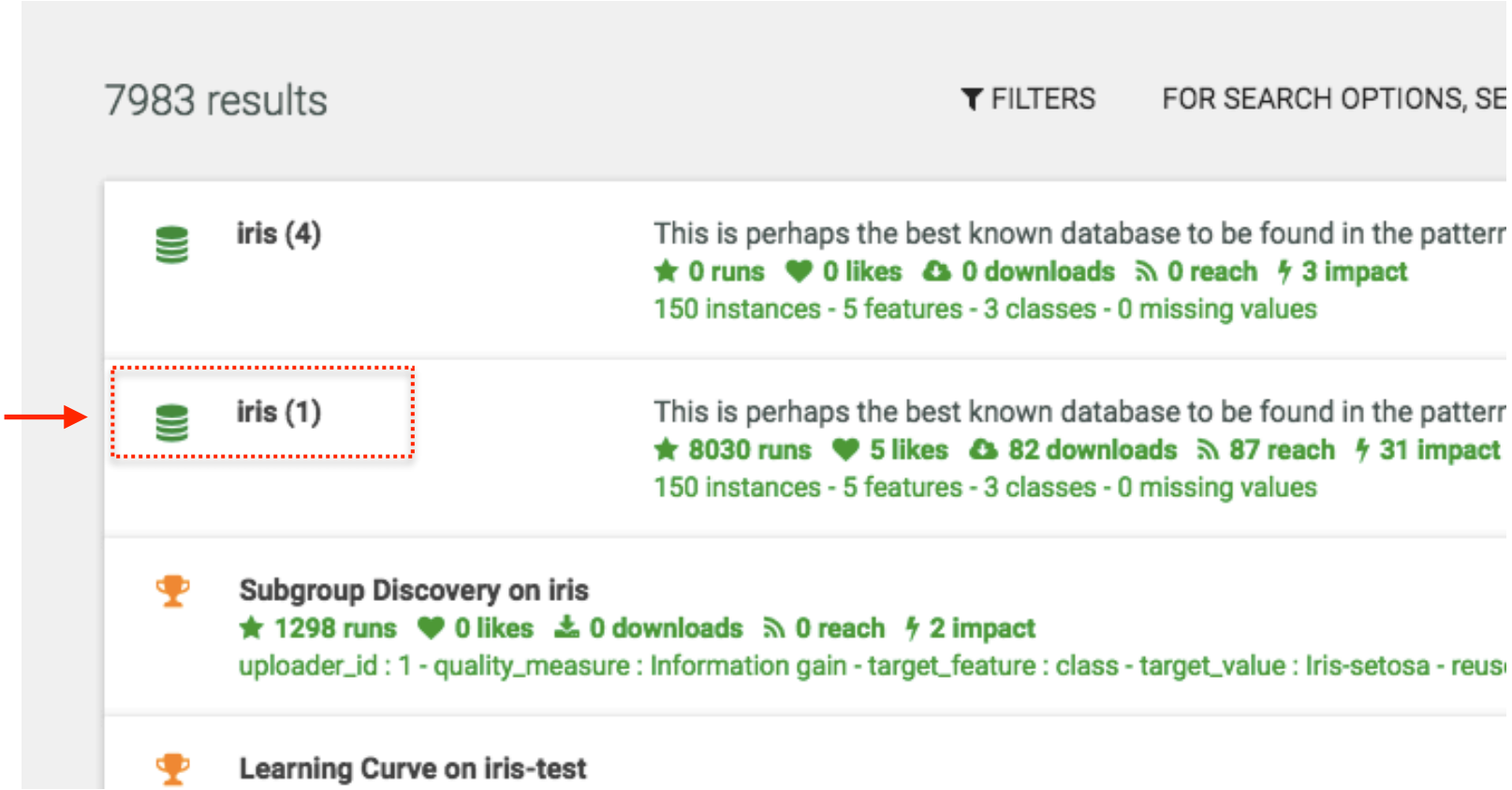


OpenML / Dados

OpenML / Dados



OpenML / Dados



iris

ARFF CSV JSON XML RDF V. 1

active ARFF Publicly available Visibility: public Uploaded 06-04-2014 by Jan van Rijn

5 likes downloaded by 82 people , 104 total downloads 0 issues 0 downvotes

study_1 study_25 study_4 study_41 study_50 study_52 study_7 study_86 study_88 study_89 uci + Add tag

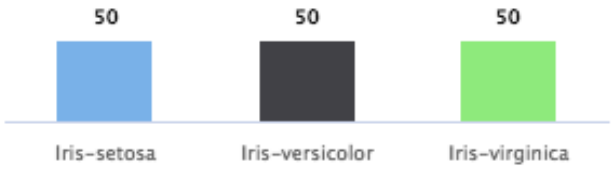
Help us complete this description → Edit

Author: R.A. Fisher
Source: UCI - 1936 - Donated by Michael Marshall
Please cite:

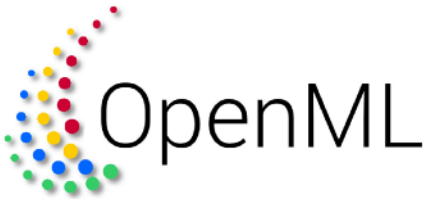
Iris Plants Database
This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly

▾ Show all

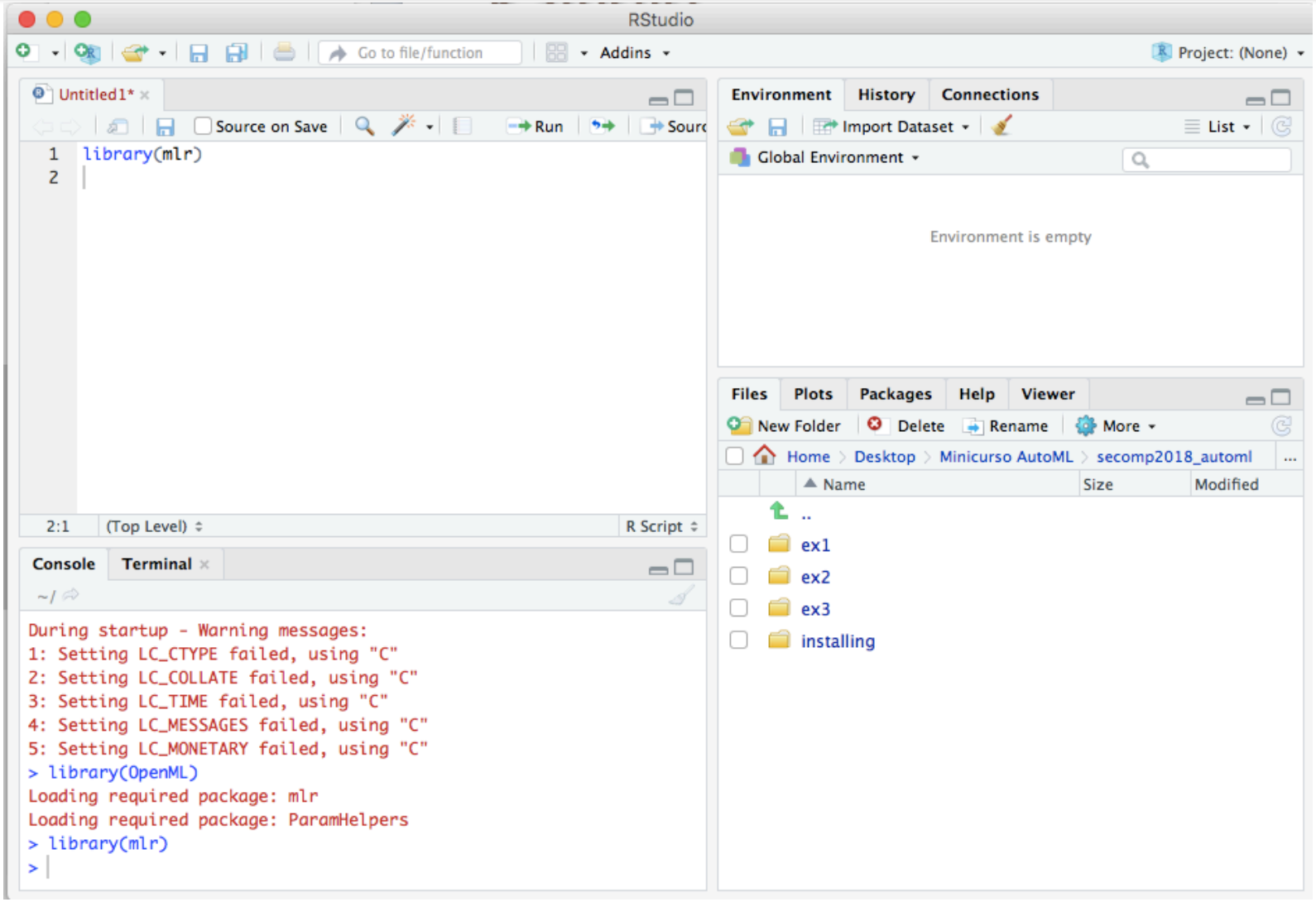
5 features

class **(target)**	nominal	3 unique values 0 missing	
sepallength	numeric	35 unique values 0 missing	

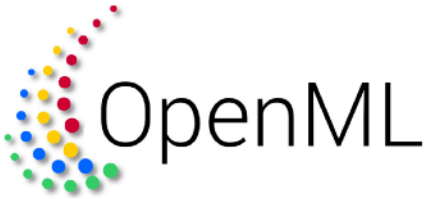
Ferramentas



Studio / IDE para R



Ferramentas



mlr / framework em R

Machine Learning in R

- CRAN release site
- Detailed Tutorial: Online as HTML
- mlr cheatsheet
- Install the development version

```
devtools::install_github("mlr-org/mlr")
```

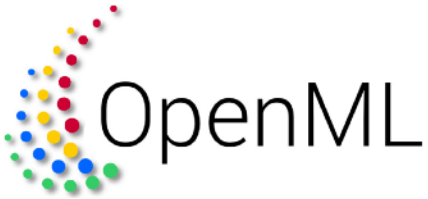
- Further installation instructions
- Ask a question about mlr on Stackoverflow
- We are on Slack (Request invitation: code{at}jakob-r.de)
- We have a blog on mlr
- A list of possible enhancements to mlr is available on the wiki - contributors welcome!
- We are in the top 20 of the most starred R packages on Github, as reported by metacran.

mlr / framework em R

- Página principal:
  - https://github.com/mlr-org/mlr

- Tutoriais:
  - https://mlr-org.github.io/mlr/
  - https://mlr-org.github.io/mlr/articles/wrapper.html
  - https://mlr-org.github.io/mlr/articles/integrated_learners.html
  - https://mlr-org.github.io/mlr/articles/measures.html
  - https://mlr-org.github.io/mlr/articles/advanced_tune.html

Ferramentas



ggplot2

ggplot2 part of the tidyverse

3.2.1

Overview

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphi

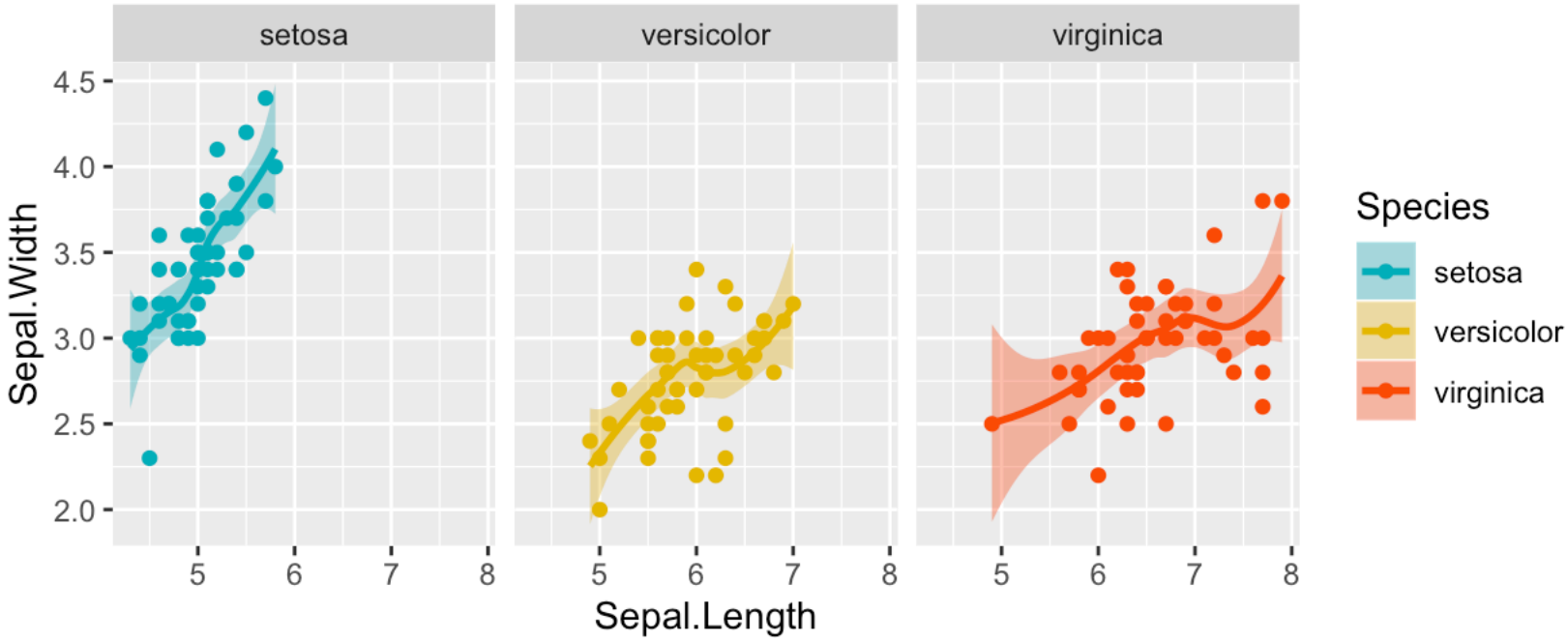
how to map variables to aesthetics, what graphical primitives to use, and it takes care of

Installation

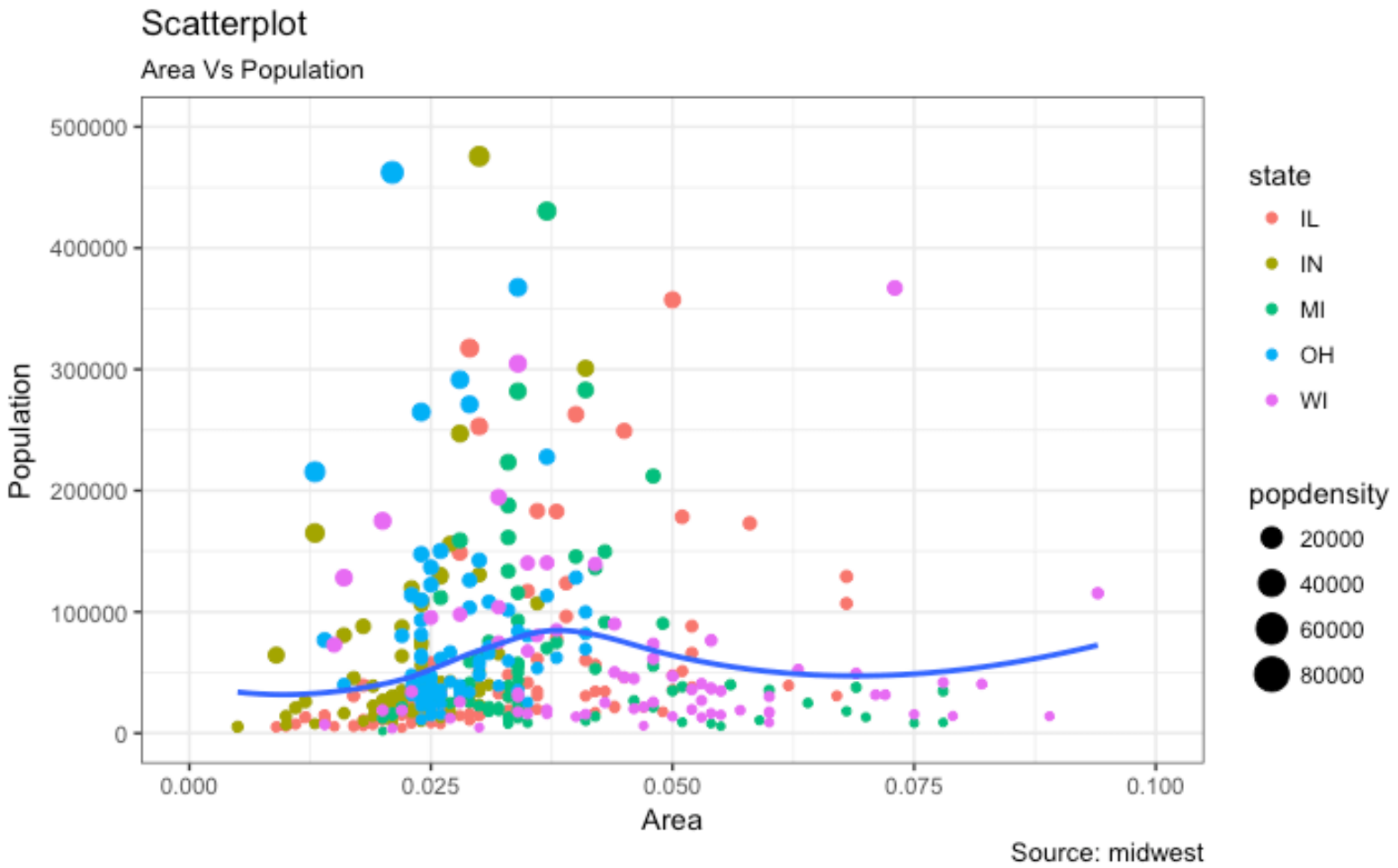
```
# The easiest way to get ggplot2 is to install the whole tidyverse:
install.packages("tidyverse")

# Alternatively, install just ggplot2:
install.packages("ggplot2")
```

ggplot2

- Visualização dos dados :)



ggplot2


Scatterplot
Area Vs Population
Population
500000
400000
300000
200000
100000
0
0.000
0.025
0.050
0.075
0.100
Area
Source: midwest
state
IL
IN
MI
OH
WI
popdensity
20000
40000
60000
80000

Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. **Welcome to R :)**
6. Referências

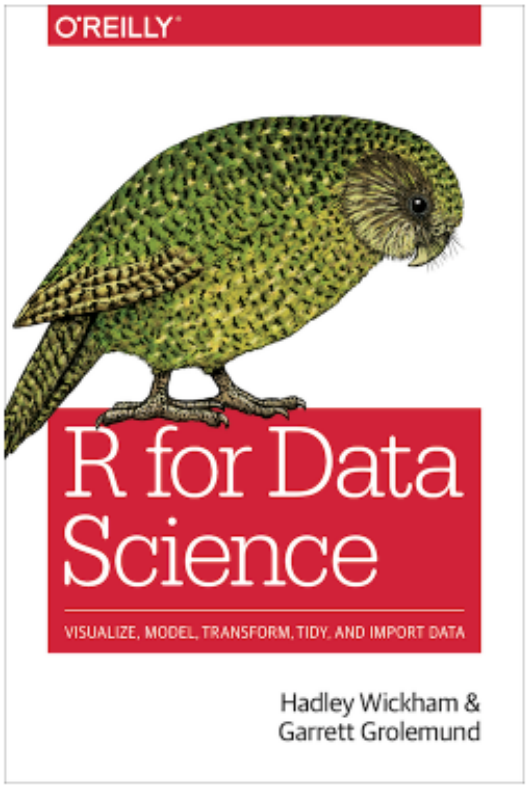
Welcome to R

- TODO: fazer hello words
- data frame
- read.table / read.csv
- summary
- class, table
- data.frame

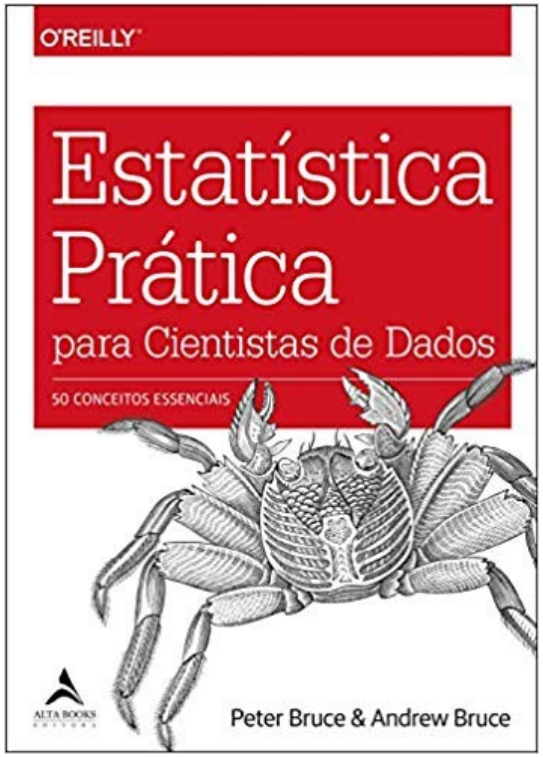
Roteiro

1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. Welcome to R :)
6. Referências

Referências

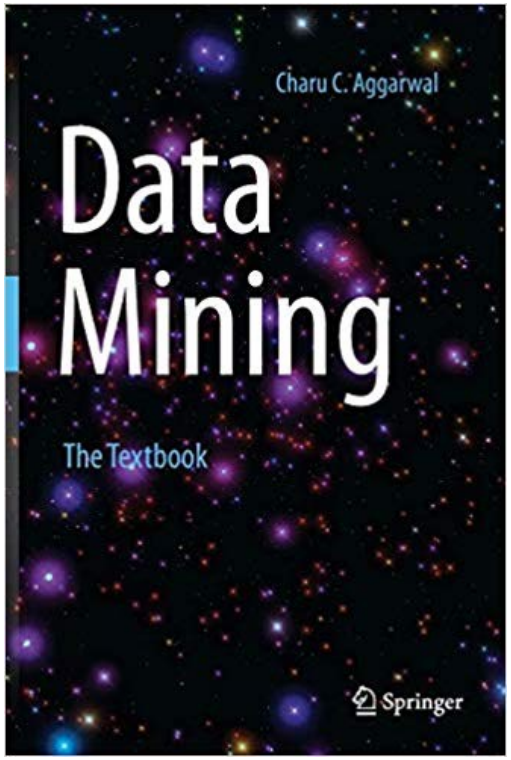


[Wickham & Grolemund, 2018]

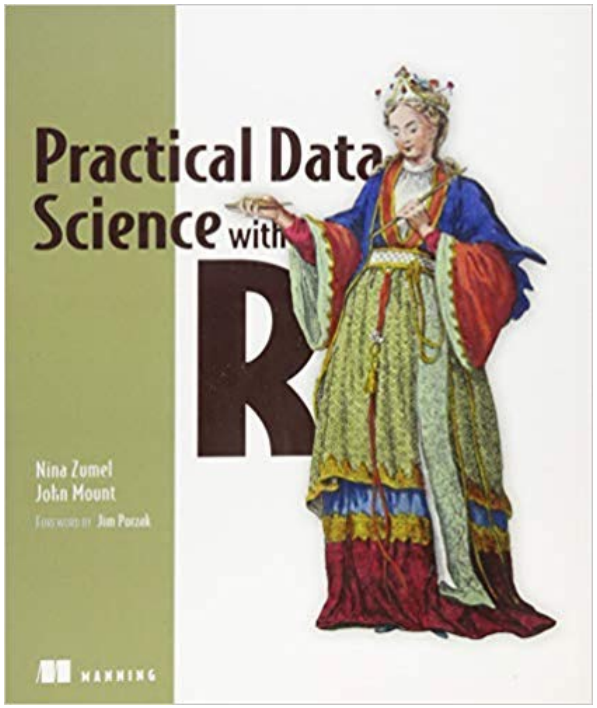


[Bruce & Bruce, 2019]

Referências



[Aggarwal, 2015]



[Zumel and Mount, 2014]

Perguntas?

Prof. Rafael G. **Mantovani**

rafaelmantovani@utfpr.edu.br