

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



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** Referências

Material

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

The screenshot shows the GitHub repository page for `rgmantovani / saes2019_dataScience`. The repository has 4 commits, 1 branch, and 0 releases. The main branch is `master`. The repository contains the following files and folders:

File/Folder	Description
<code>datasets</code>	adding config scripts
<code>scripts</code>	adding examples
<code>ggplot2-cheatsheet.pdf</code>	adding cheat sheets
<code>r-cheat-sheet-3.pdf</code>	adding cheat sheets

Roteiro

- 1** Introdução
- 2** Conceitos gerais
- 3** Fluxo de ciência de dados
- 4** Ferramentas
- 5** Referências

Introdução

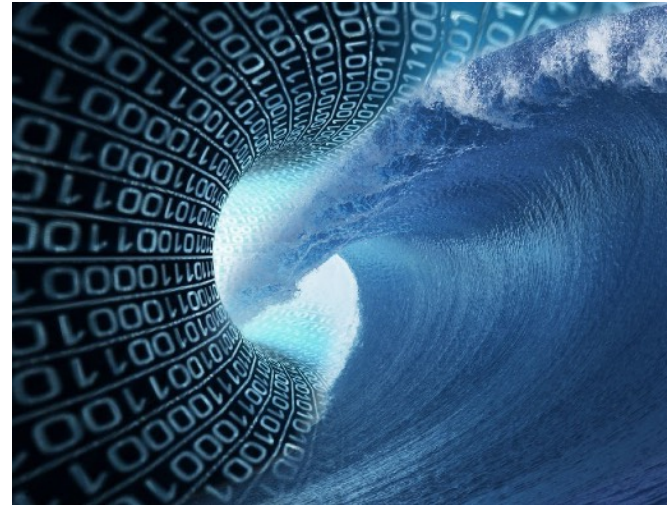


poucos dados

Introdução

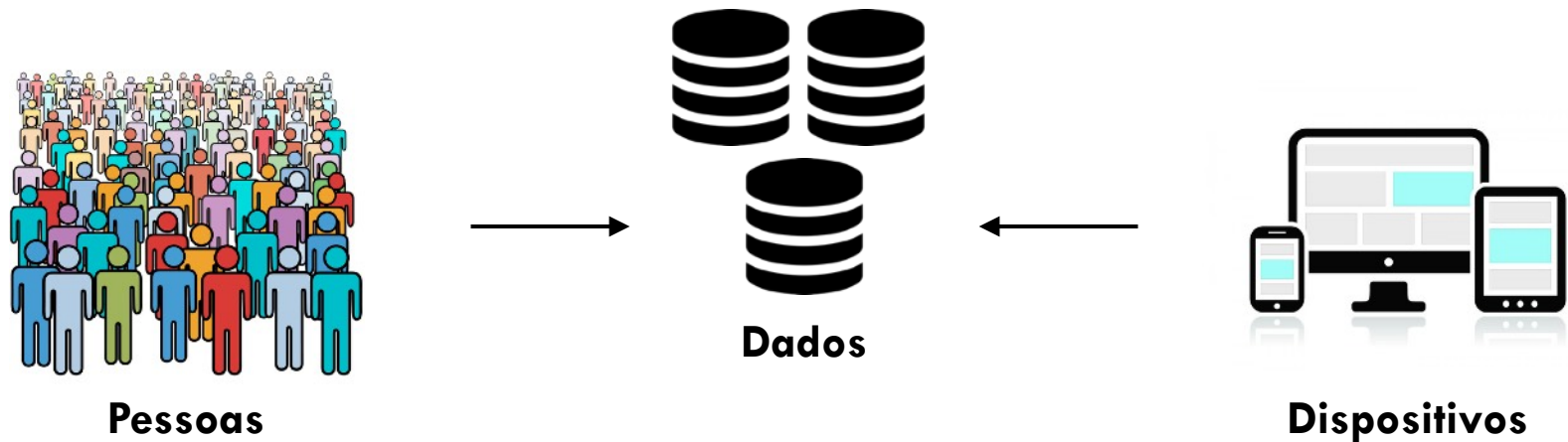


poucos dados

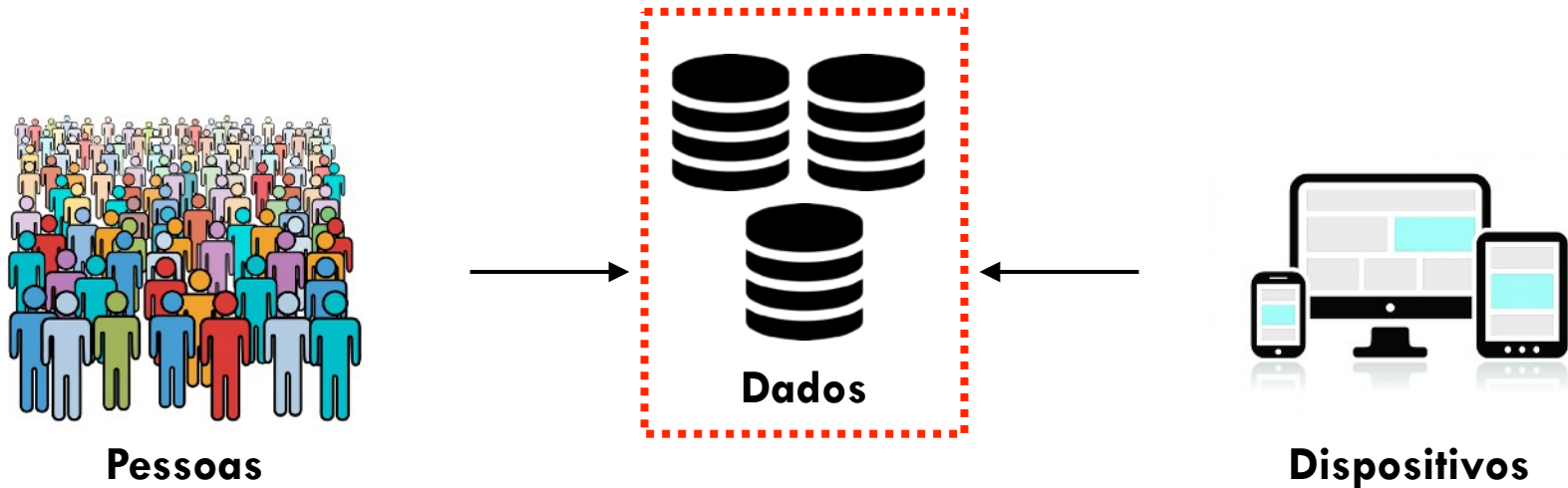


**imensa quantidade
de dados (big data)**

Introdução



Introdução



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

Introdução

- Mudança de realidade



**Machine
Learning**



Dados

**Necessidade:
conjuntos de dados**

Introdução

- Mudança de realidade



Machine
Learning

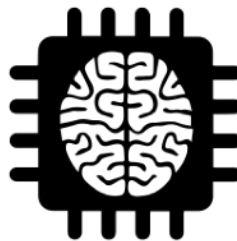


Dados

Necessidade:
conjuntos de dados



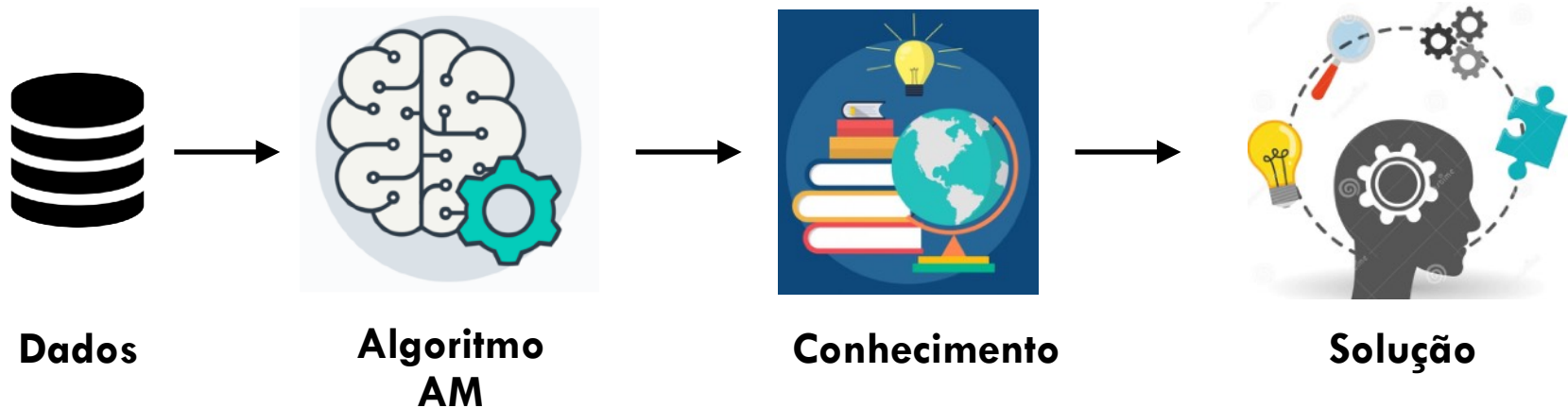
Dados



Machine
Learning

Necessidade:
Algoritmos

Introdução



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

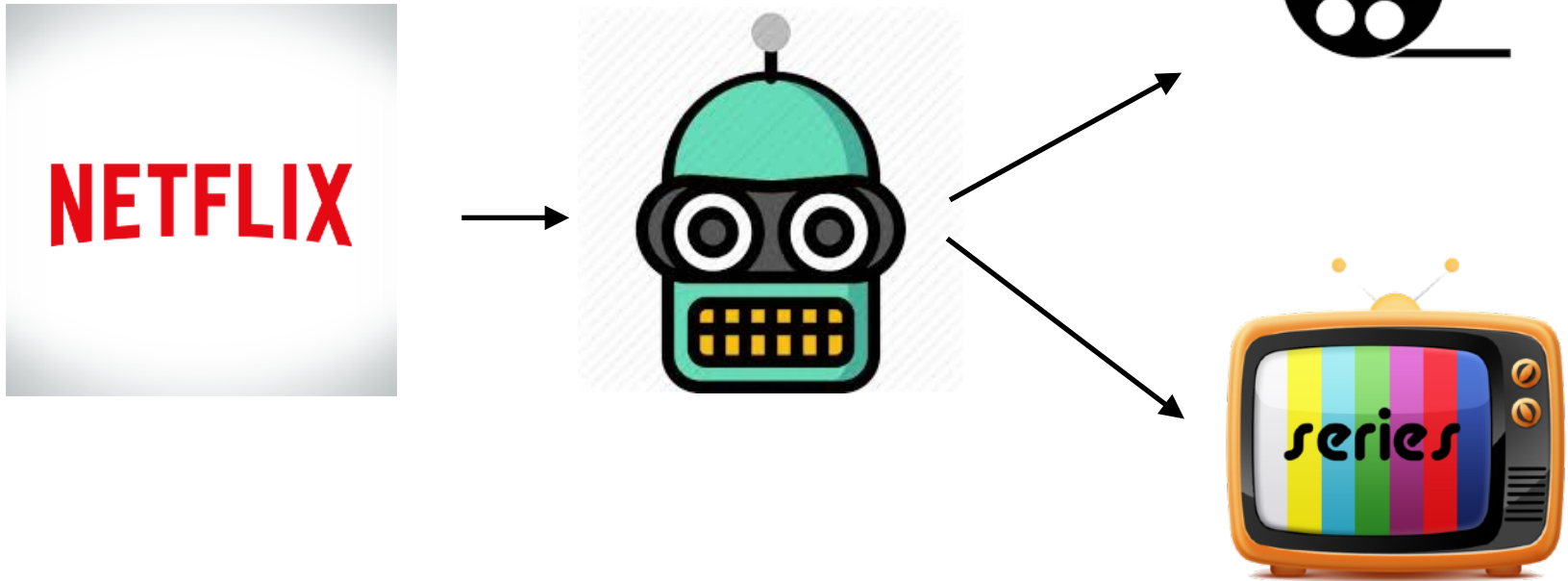
Introdução



- Onde isso é usado?

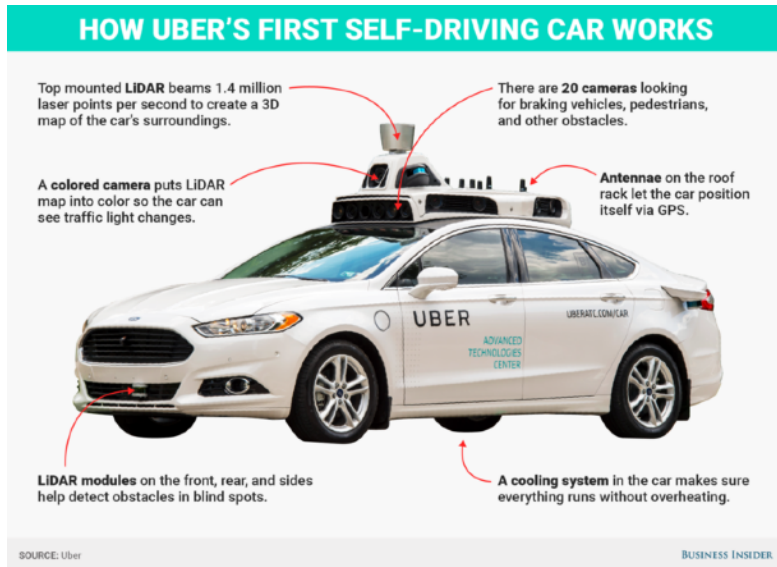
Introdução

- 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**

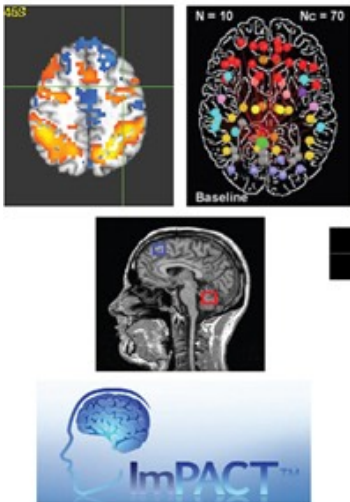


Introdução

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

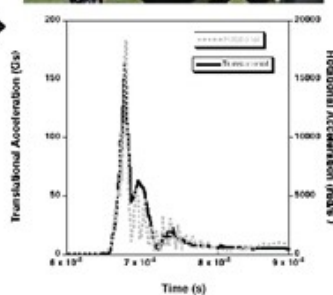
Pre-Season Assessments

MRI, fMRI, MRS
ImPACT



Quantitative Head Impact Measurements

HITS, X2, Custom Systems

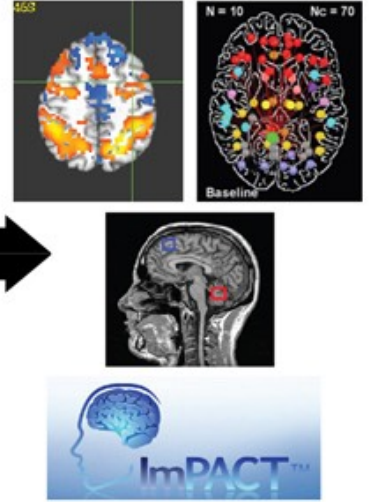


On-Field Neurological Assessments

On-Field Neurological Assessments: SCAT2 Sport Concussion Assessment Tool 2. The figure shows a screenshot of the SCAT2 form, which includes sections for 'Symptom Evaluation' and 'Neurological Examination'.

In-Season and Post-Season Assessments

MRI, fMRI, MRS
ImPACT



Introdução

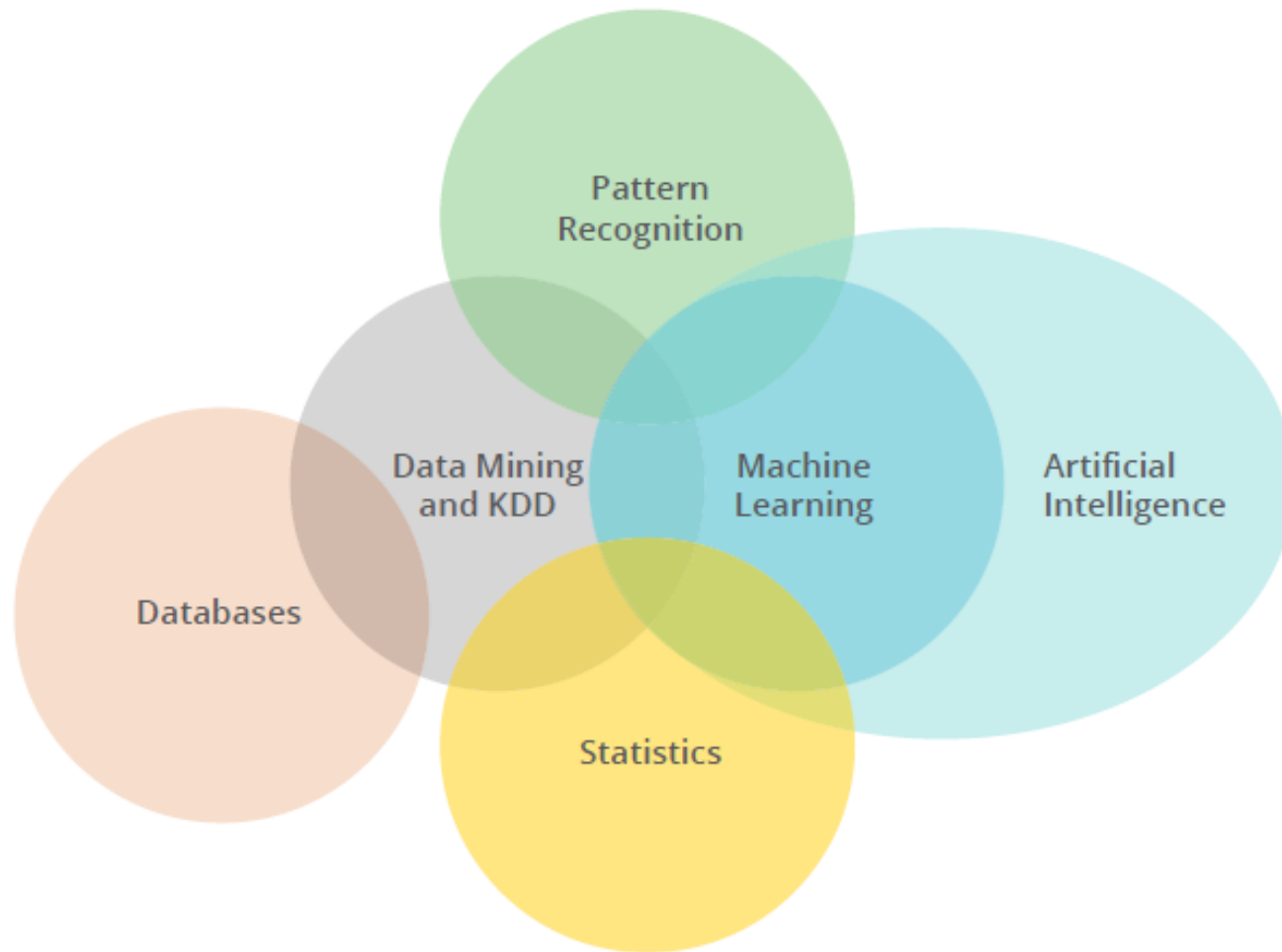
- Onde isso é usado? **Sistemas de Segurança**



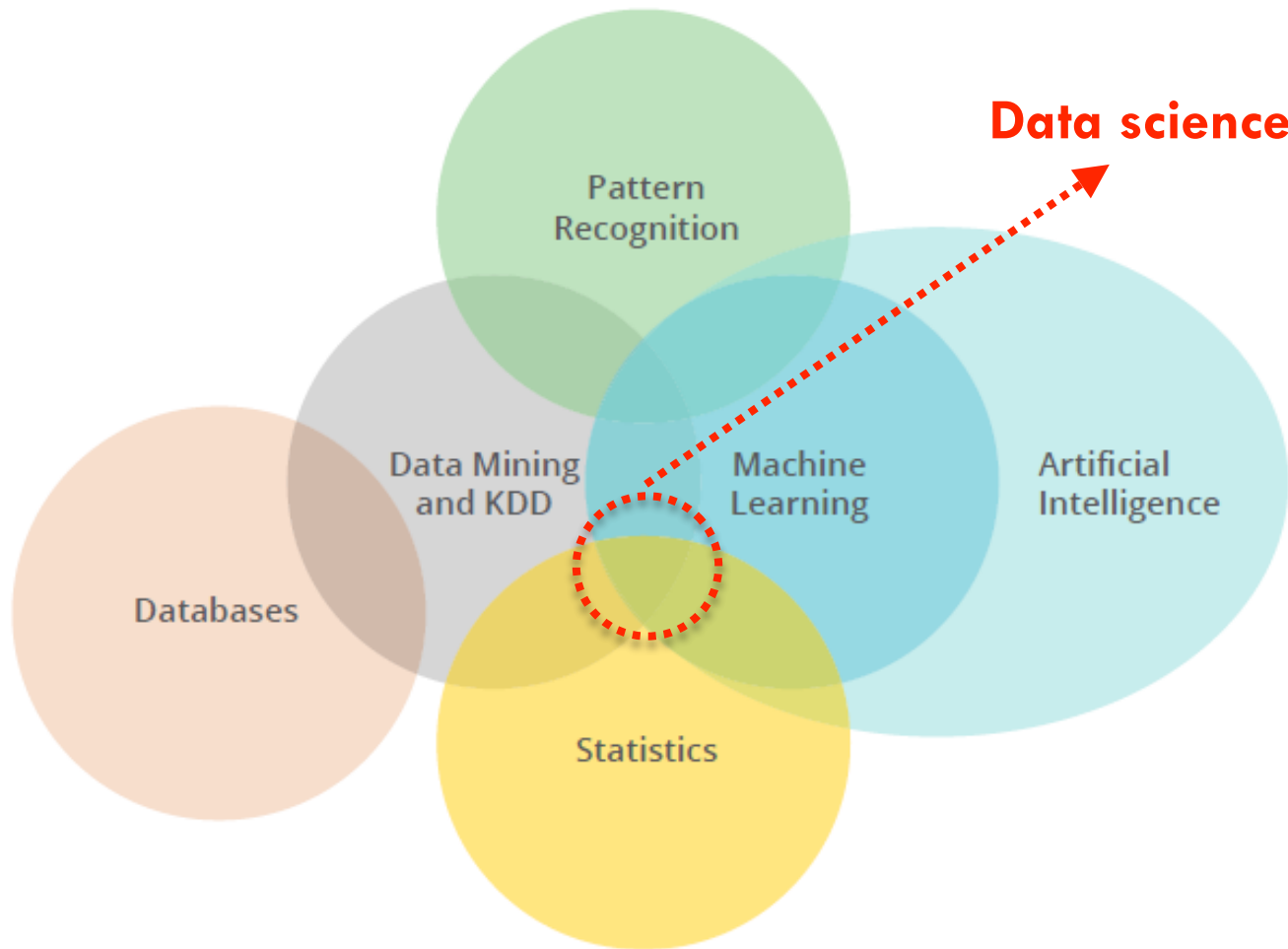
Roteiro

- 1 Introdução
- 2 Conceitos gerais
- 3 Fluxo de ciência de dados
- 4 Ferramentas
- 5 Referências

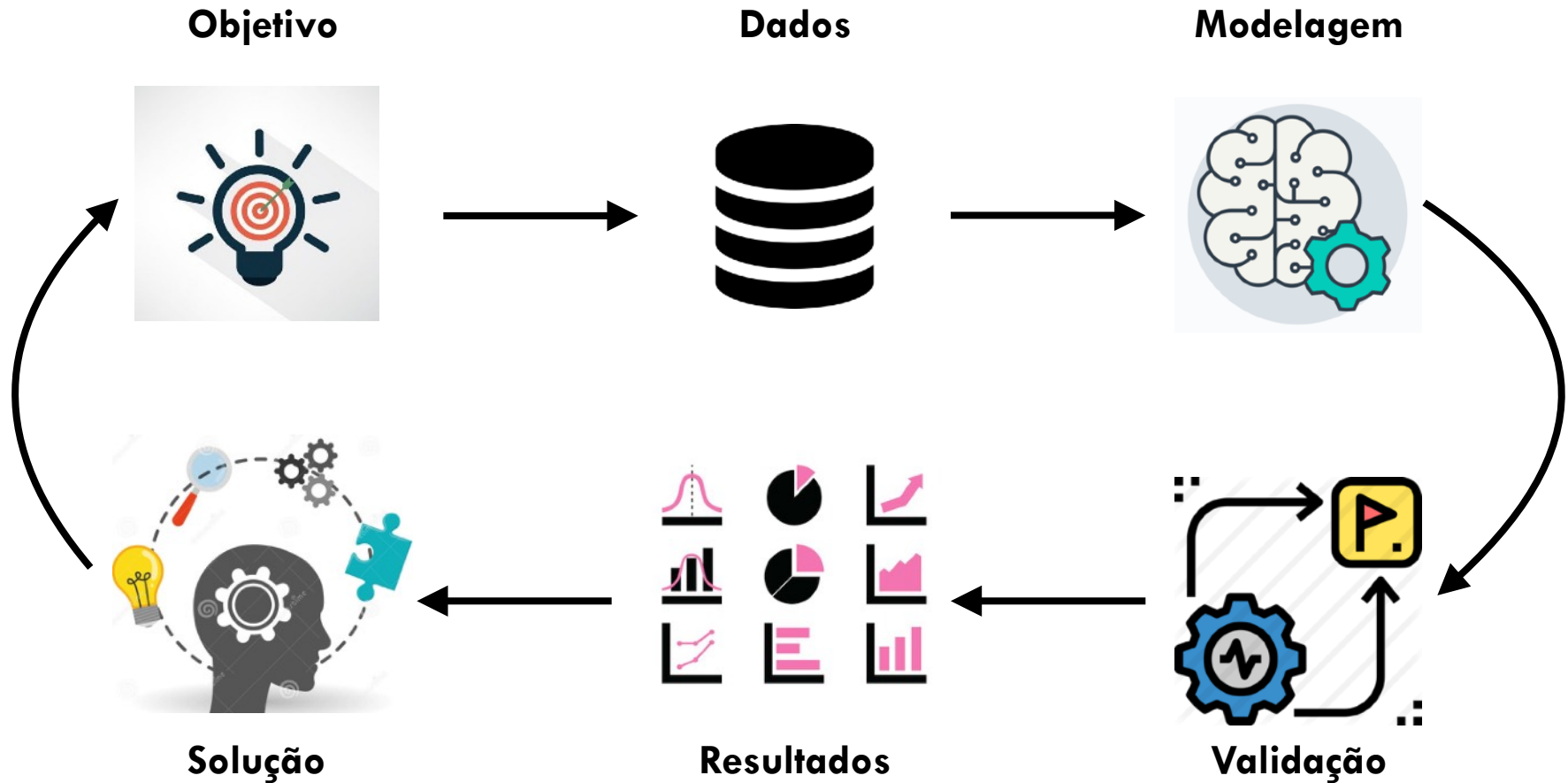
Conceitos Gerais



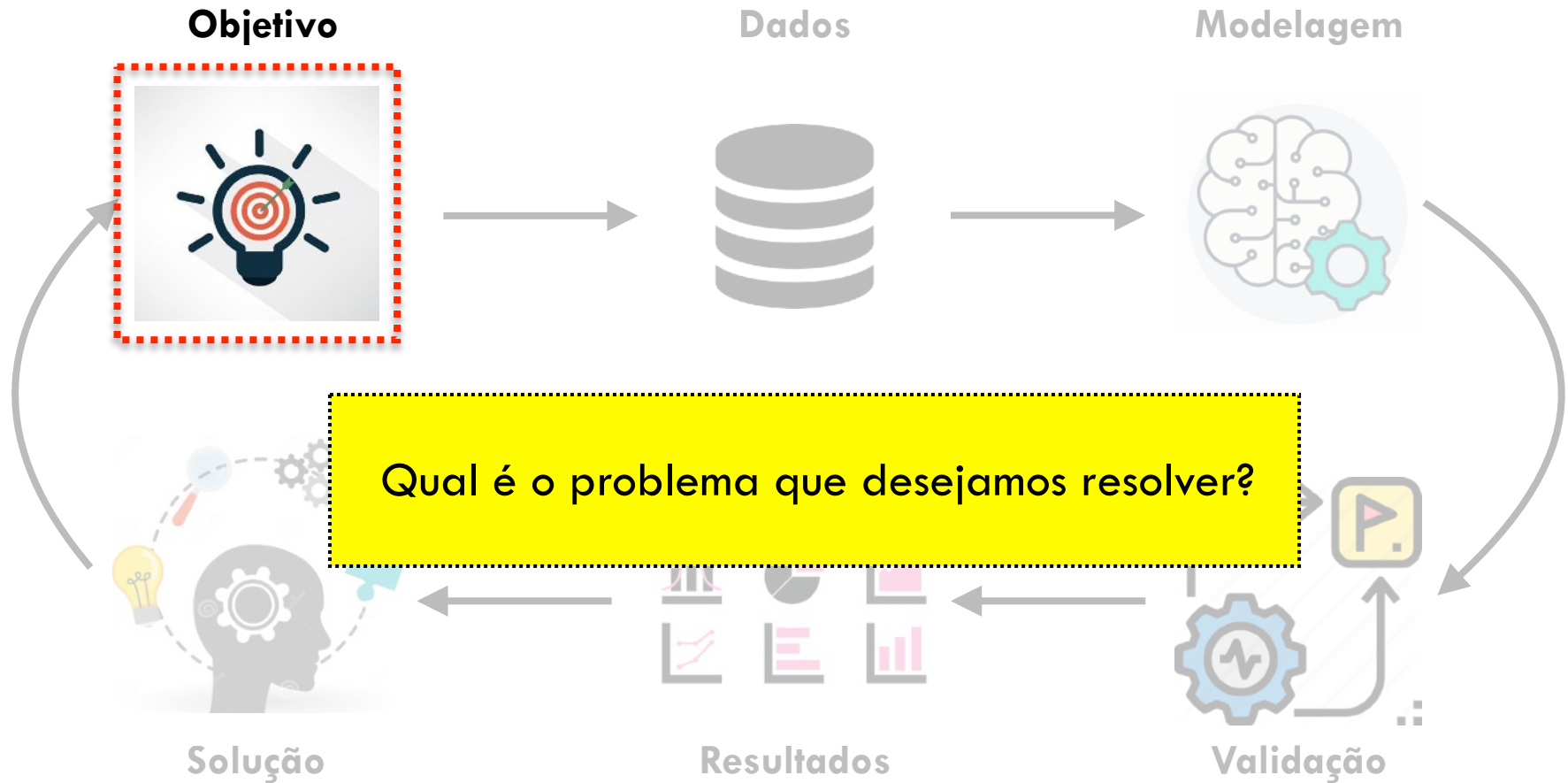
Conceitos Gerais



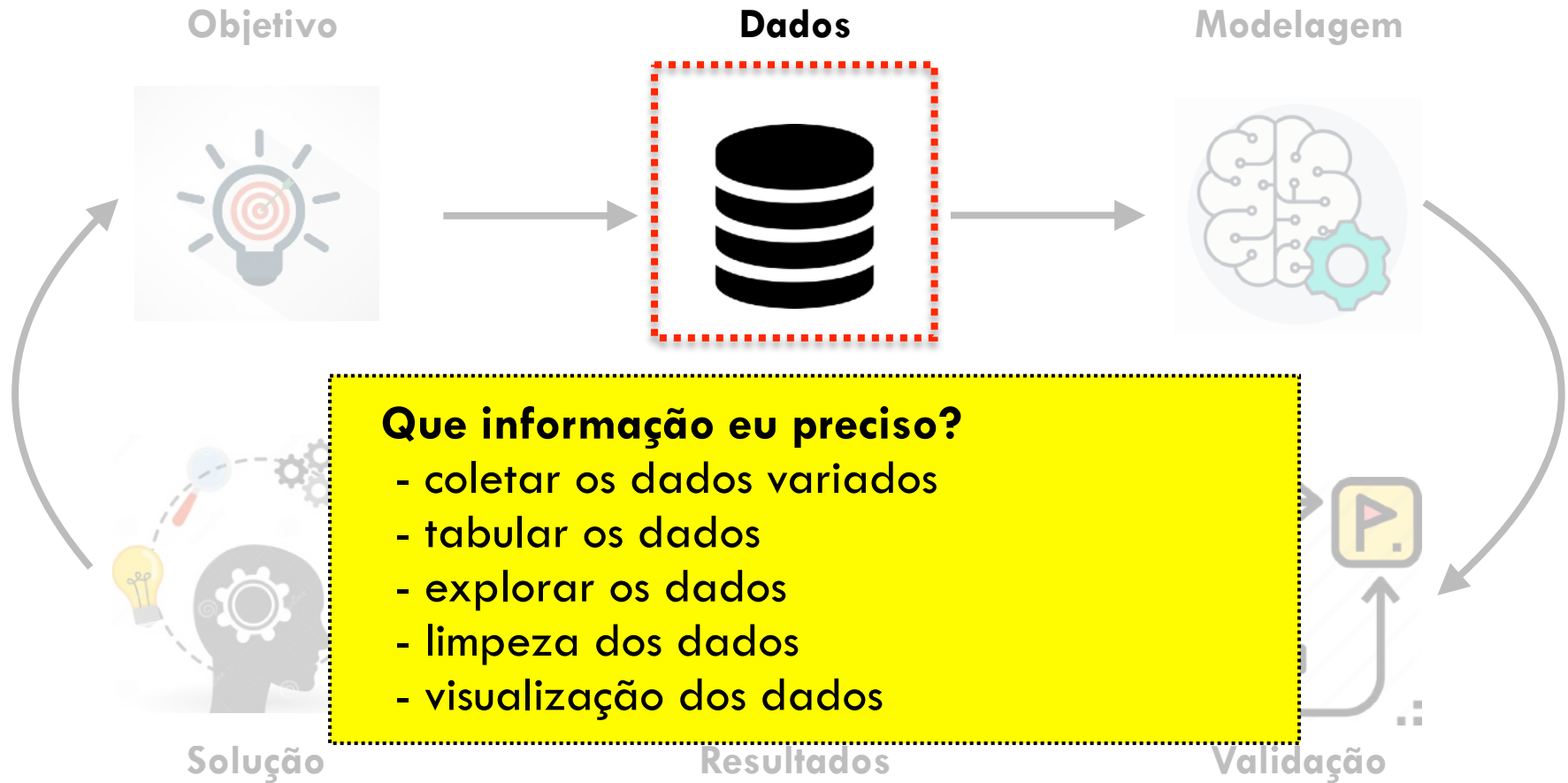
Ciência de Dados



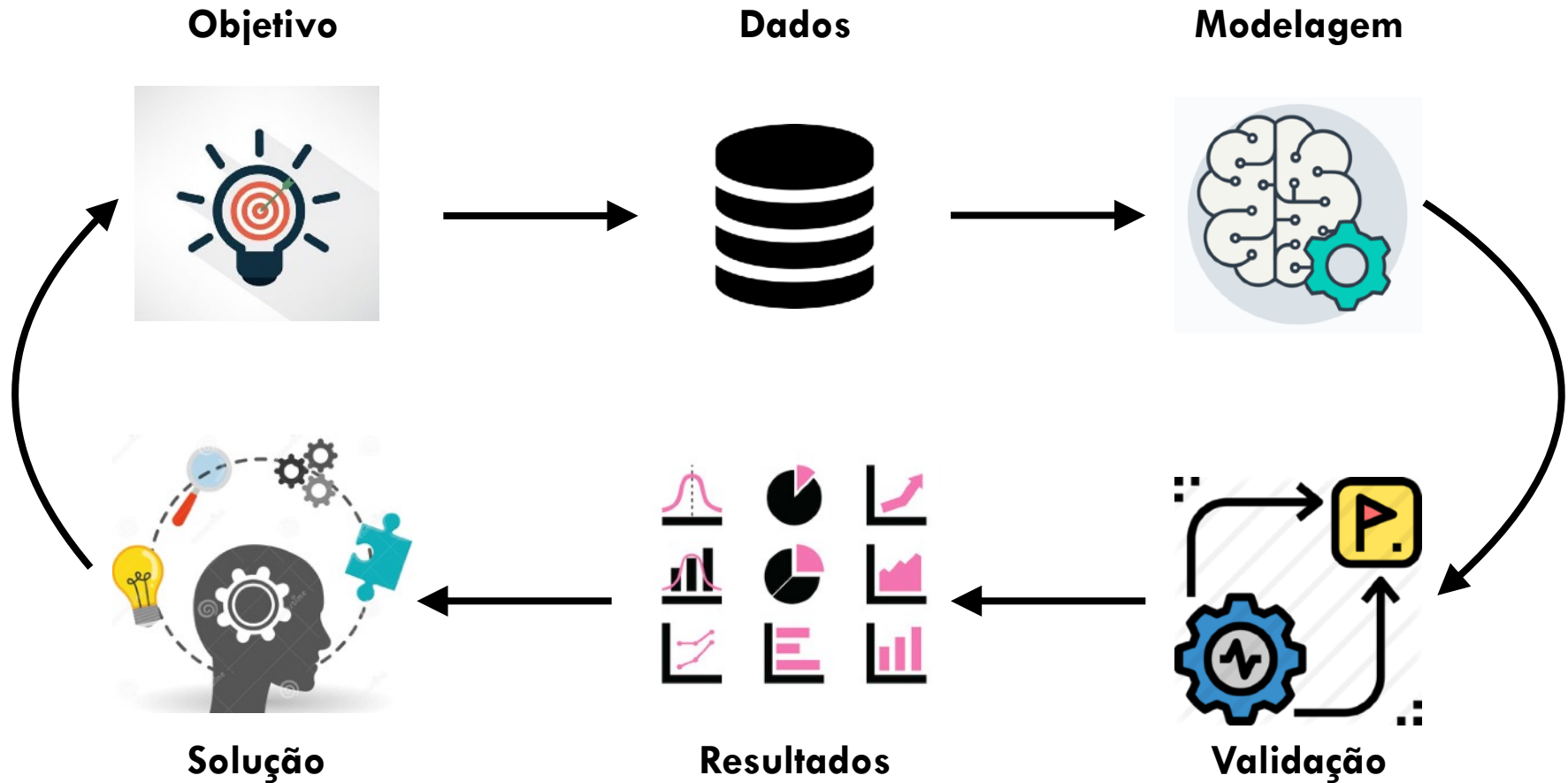
Ciência de Dados



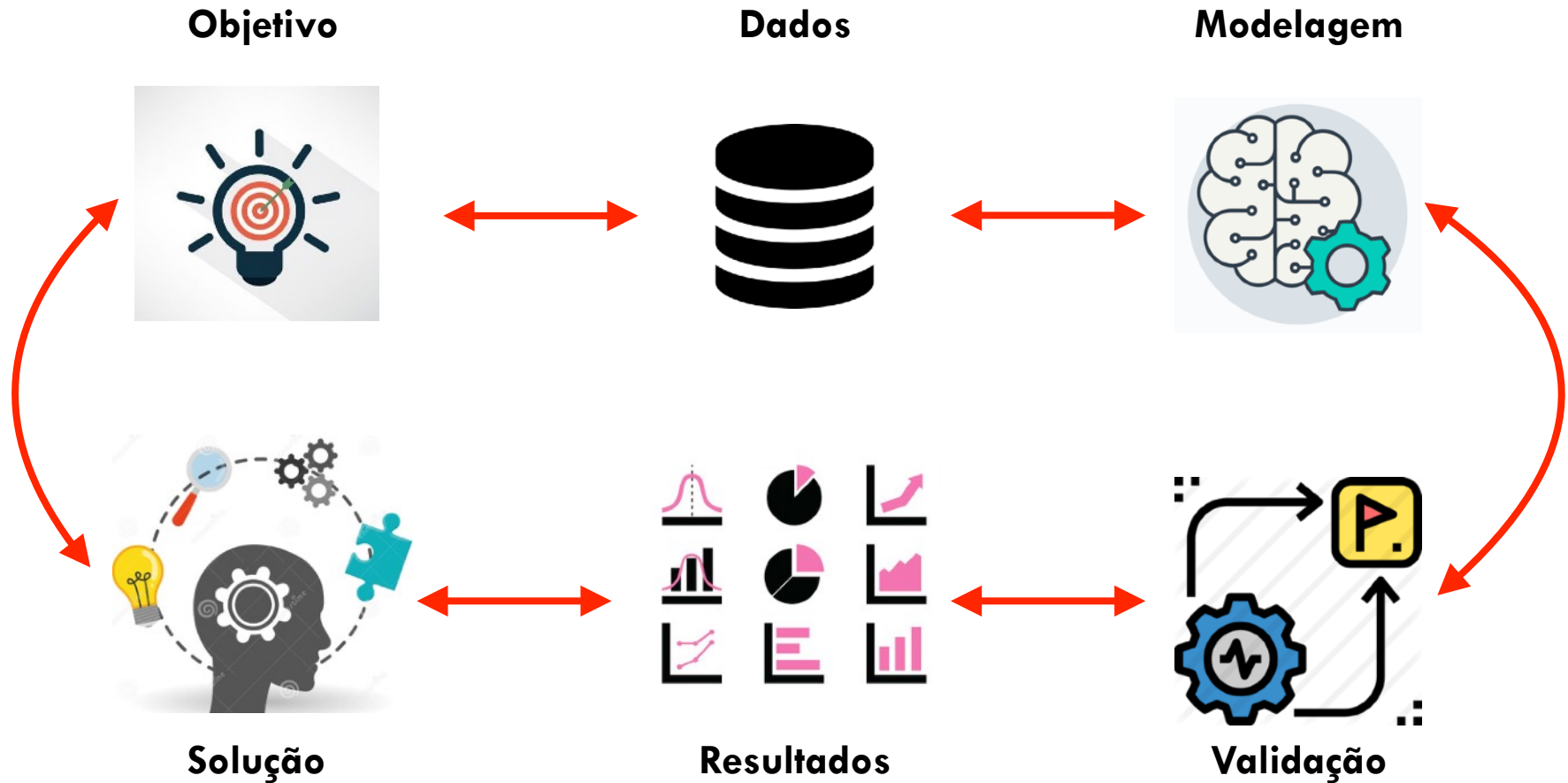
Ciência de Dados



Ciência de Dados



Ciência de Dados



Aprendizado de Máquina



- Quantos algoritmos existem?

Aprendizado de Máquina

- Quantos algoritmos existem?



...

Roteiro

- 1 Introdução
- 2 Conceitos gerais
- 3 Fluxo de ciência de dados
- 4 Ferramentas
- 5 Referências

Roteiro

- 1 Introdução
- 2 Conceitos gerais
- 3 Fluxo de ciência de dados
- 4 Ferramentas
- 5 Referências

Ferramentas



OpenML

RStudio

ggplot2

mlr

Ferramentas



OpenML

RStudio

ggplot2

mlr

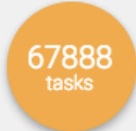
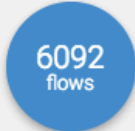

OpenML / Dados

Search



OpenML^{beta}

Machine learning, better, together

 <p>20072 data sets</p>	 <p>67888 tasks</p>	 <p>6092 flows</p>	 <p>9012316 runs</p>
Find or add data to analyse	Download or create scientific tasks	Find or add data analysis flows	Upload and explore all results online.

OpenML / Dados

7983 results

▼ FILTERS

FOR SEARCH OPTIONS, SE



iris (4)

This is perhaps the best known database to be found in the pattern

★ 0 runs ♥ 0 likes 📄 0 downloads 🌐 0 reach ⚡ 3 impact

150 instances - 5 features - 3 classes - 0 missing values



iris (1)

This is perhaps the best known database to be found in the pattern

★ 8030 runs ♥ 5 likes 📄 82 downloads 🌐 87 reach ⚡ 31 impact

150 instances - 5 features - 3 classes - 0 missing values



Subgroup Discovery on iris

★ 1298 runs ♥ 0 likes 📄 0 downloads 🌐 0 reach ⚡ 2 impact

uploader_id : 1 - quality_measure : Information gain - target_feature : class - target_value : Iris-setosa - reus



Learning Curve on iris-test

OpenML / Dados

7983 results

▼ FILTERS

FOR SEARCH OPTIONS, SE



iris (4)

This is perhaps the best known database to be found in the pattern

★ 0 runs ♥ 0 likes 📄 0 downloads 🌐 0 reach ⚡ 3 impact

150 instances - 5 features - 3 classes - 0 missing values



iris (1)

This is perhaps the best known database to be found in the pattern

★ 8030 runs ♥ 5 likes 📄 82 downloads 🌐 87 reach ⚡ 31 impact

150 instances - 5 features - 3 classes - 0 missing values



Subgroup Discovery on iris


★ 1298 runs ♥ 0 likes 📄 0 downloads 🌐 0 reach ⚡ 2 impact


uploader_id : 1 - quality_measure : Information gain - target_feature : class - target_value : Iris-setosa - reus





Learning Curve on iris-test


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	<div><div>50</div><div>50</div><div>50</div></div> <div><div>Iris-setosa</div><div>Iris-versicolor</div><div>Iris-virginica</div></div>
sepallength	numeric	35 unique values 0 missing	<div><div></div><div></div><div></div><div></div></div>

Ferramentas



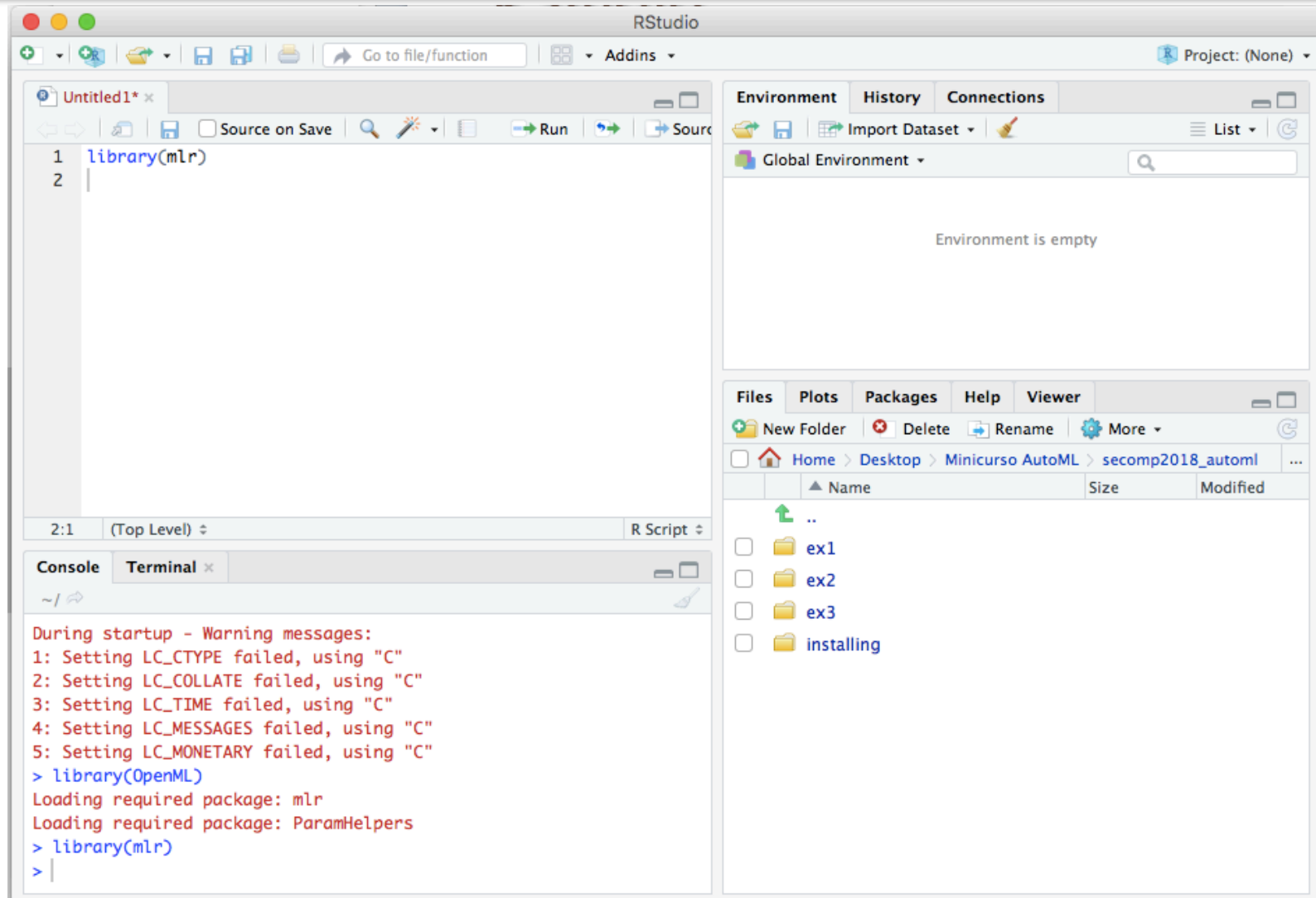
OpenML

RStudio

ggplot2

mlr

Studio / IDE para R



Ferramentas



OpenML

RStudio

ggplot2

mlr

mlr / framework em R

Machine Learning in R



build failing build failing CRAN 2.13 downloads 7732/month stackoverflow mlr

- [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



OpenML

RStudio

ggplot2

mlr

ggplot2



ggplot2 part of the [tidyverse](#)
3.2.1

Overview

ggplot2 is a system for declaratively creating graphics, based on [The Grammar of Graphics](#). It makes it easy to map variables to aesthetics, what graphical primitives to use, and it takes care of

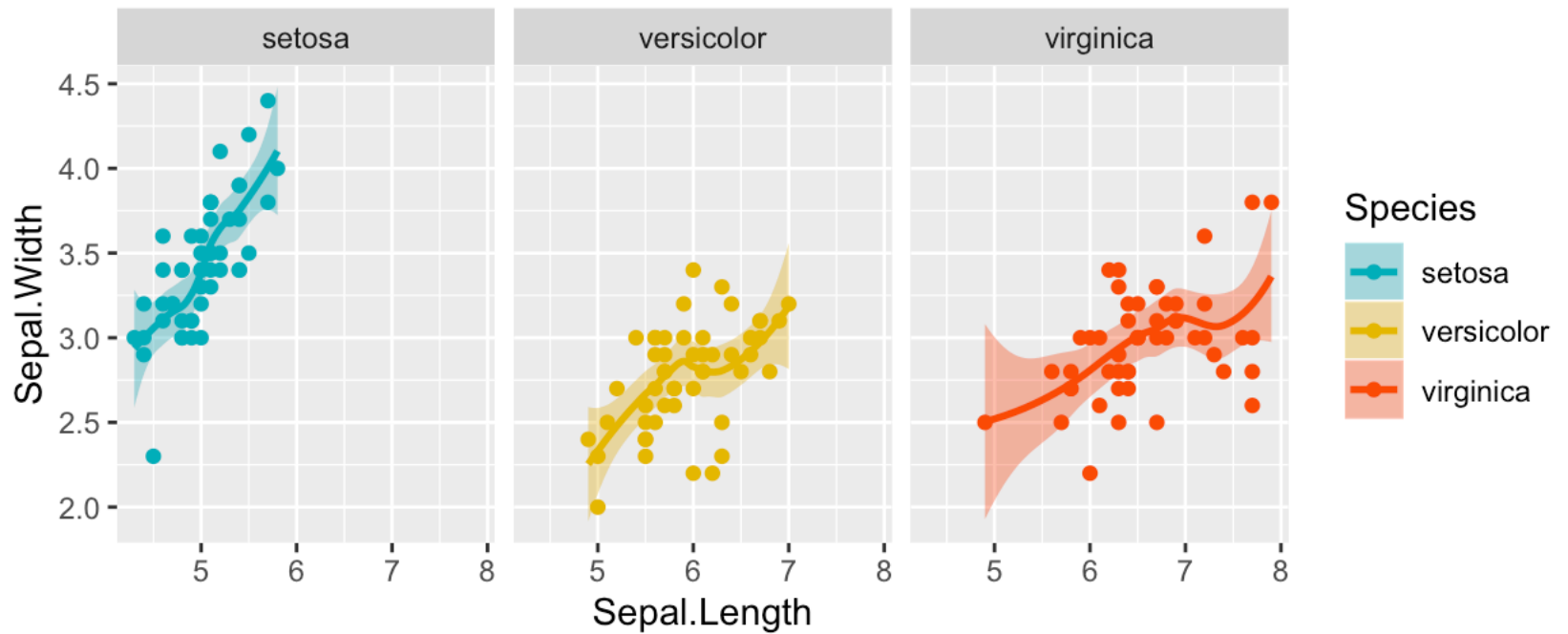
Installation

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

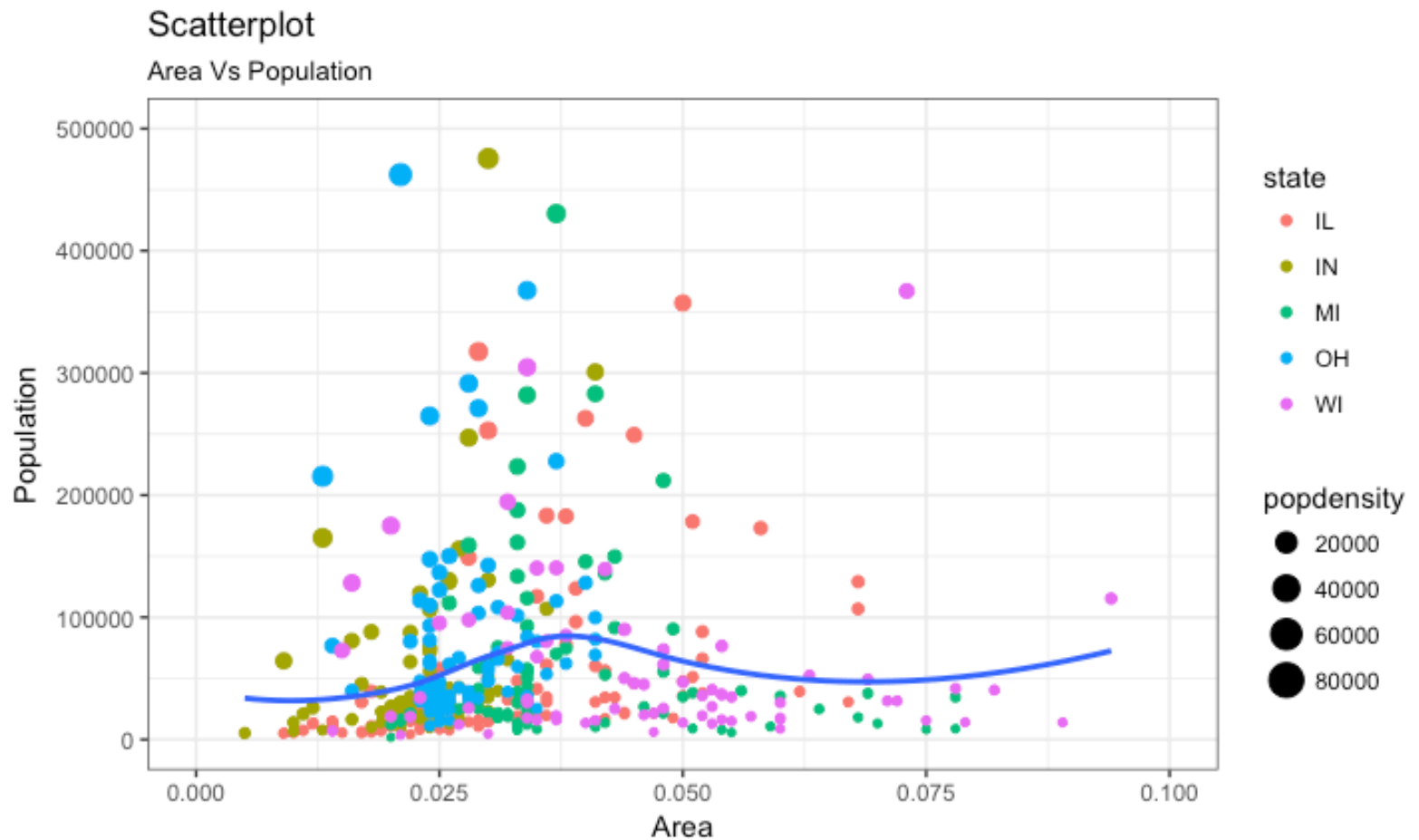
```
# Alternatively, install just ggplot2:  
install.packages("ggplot2")
```

ggplot2

- Visualização dos dados :)



ggplot2

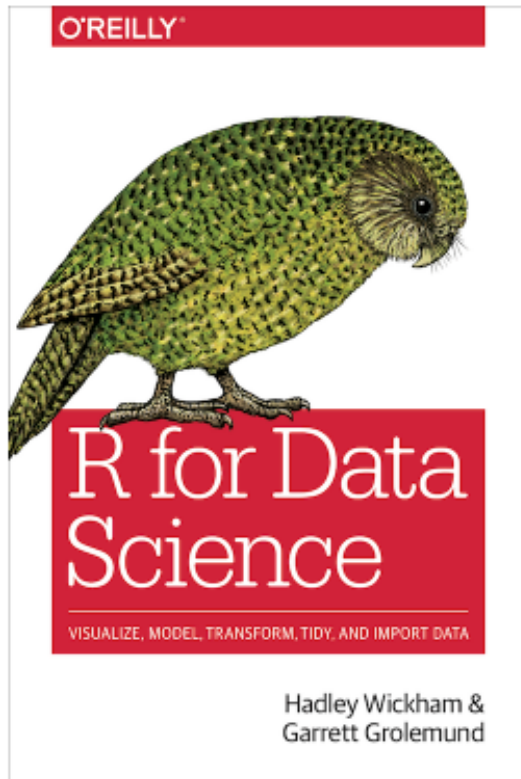


Source: midwest

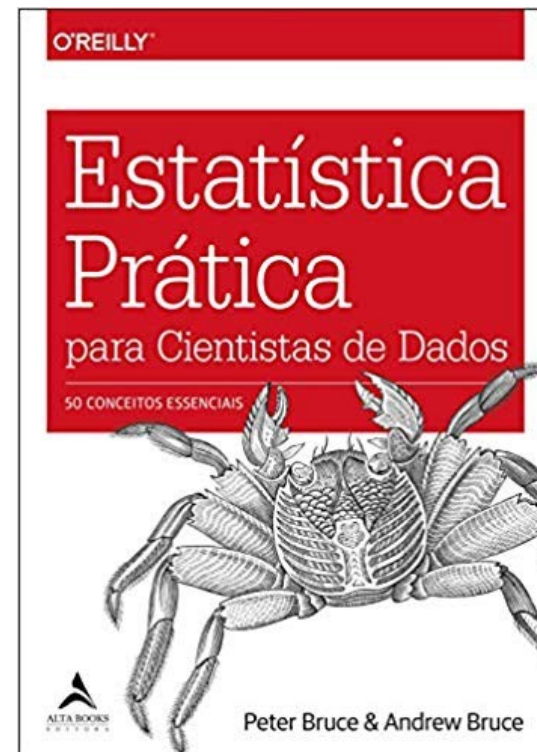
Roteiro

- 1 Introdução
- 2 Conceitos gerais
- 3 Fluxo de ciência de dados
- 4 Ferramentas
- 5 Referências

Referências

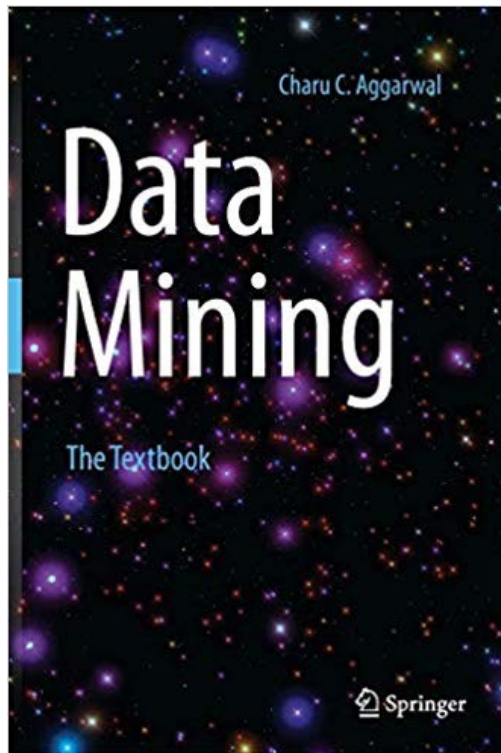


[Wickham & Grolemund, 2018]

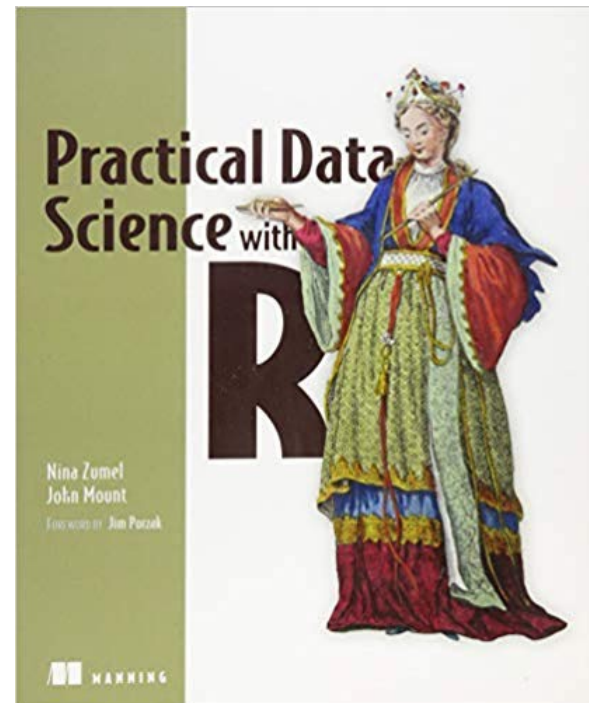


[Bruce & Bruce, 2019]

Referências



[Aggarwal, 2015]



[Zumel and Mount, 2014]

Perguntas?

Prof. Rafael G. **Mantovani**

rafaelmantovani@utfpr.edu.br