APRENDENDO COM OS DADOS

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

(PARTE 1)

Prof. Rafael G. Mantovani

30/10/2019

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Apucarana - PR, Brasil
Outubro, 2019

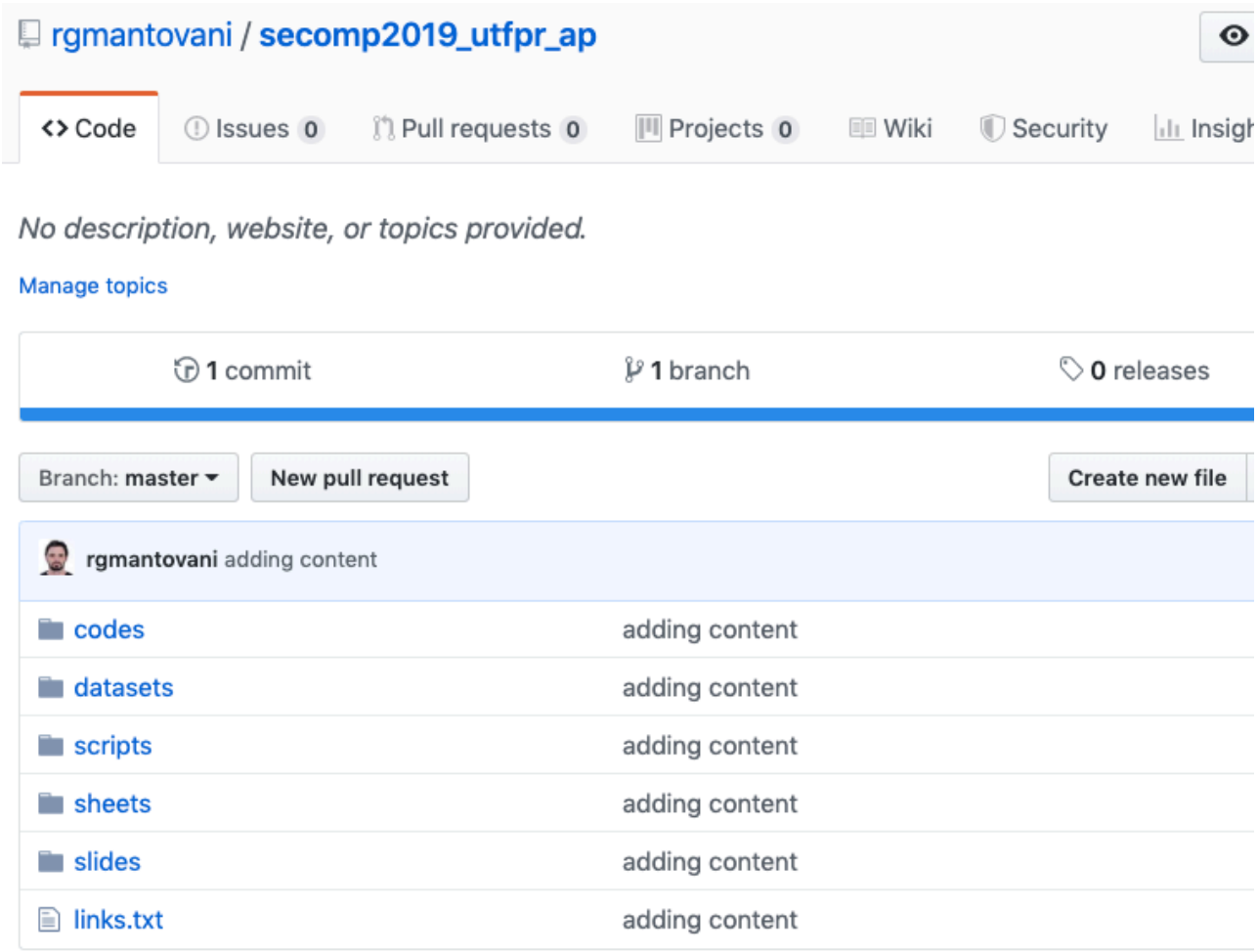
Universidade Tecnológica Federal do Paraná (UTFPR)
I Semana de Engenharia de Computação

Roteiro

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

Material

Link: https://github.com/rgmantovani/secomp2019_utfpr_ap



Roteiro

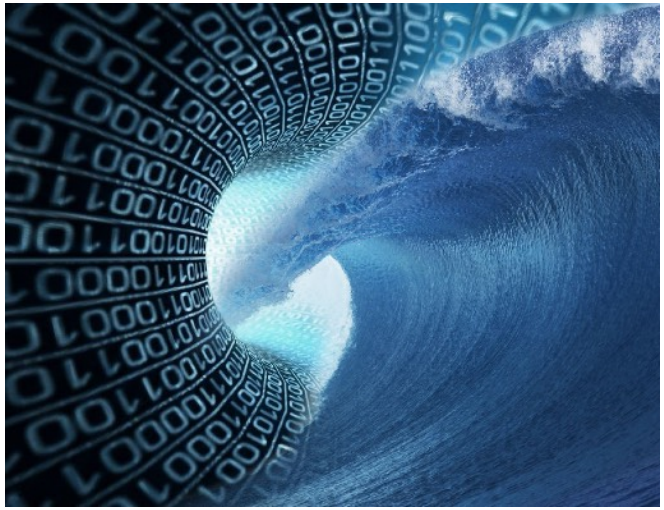
1. Introdução
2. Conceitos gerais
3. Fluxo de ciência de dados
4. Ferramentas
5. Um pouco de R :)
6. Referências

Introdução

poucos dados

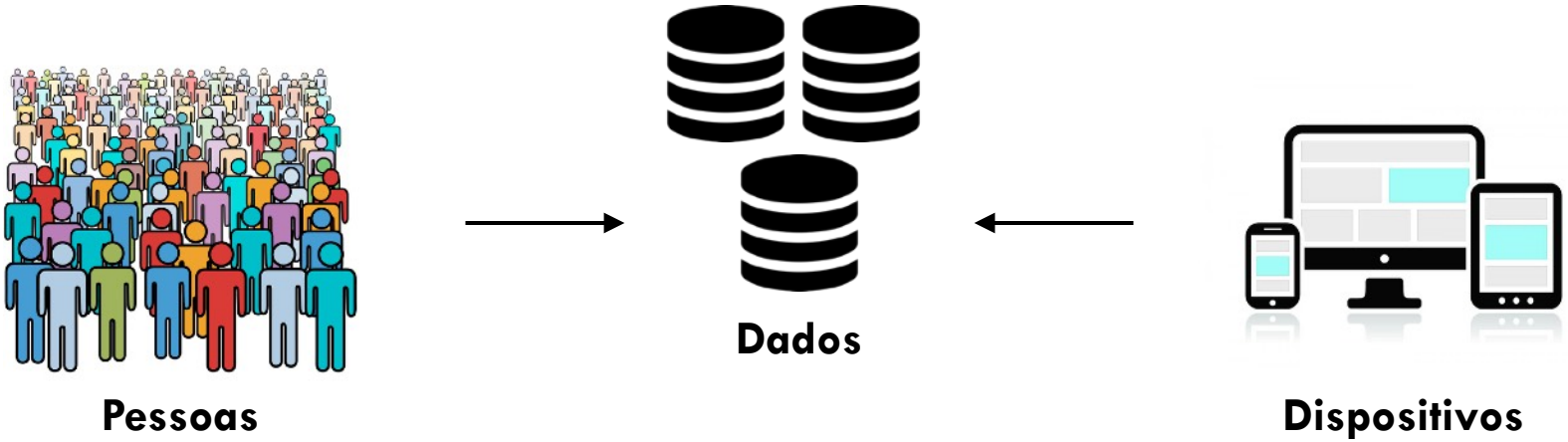
Introdução

poucos dados

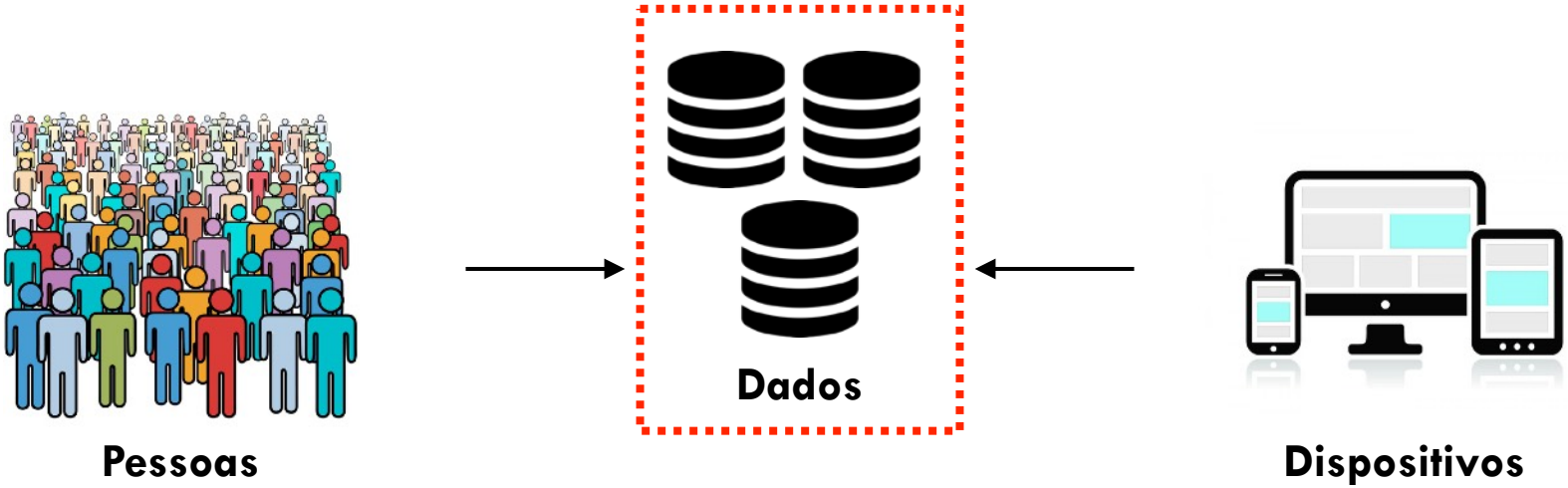


imensa quantidade
de dados (big data)

Introdução



Introdução



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

Introdução

- Mudança de realidade

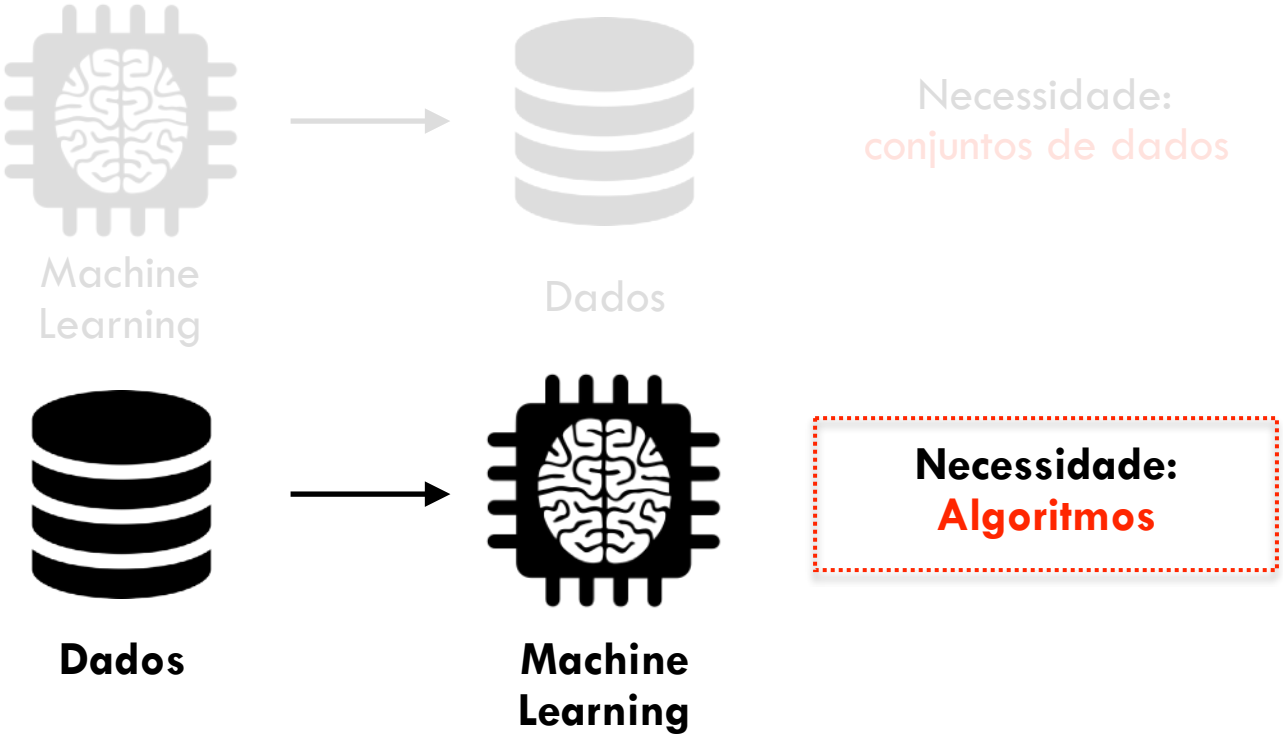
Machine Learning

Dados

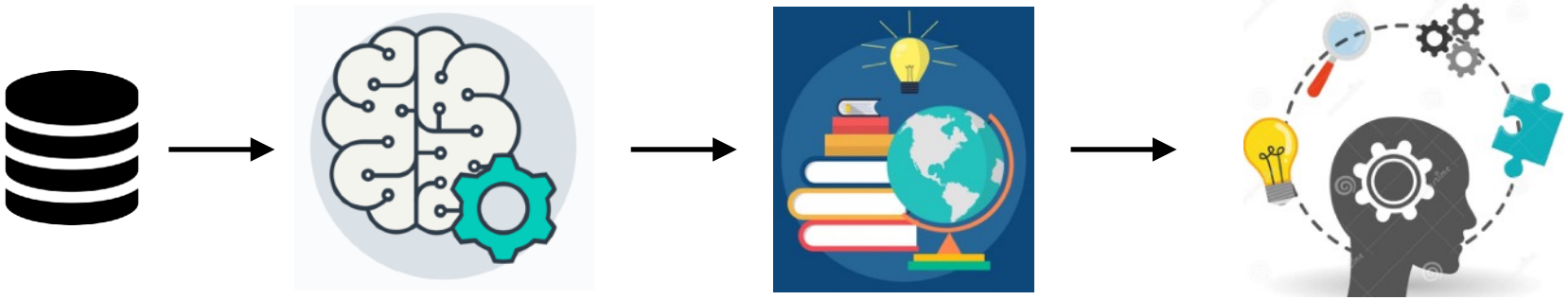
Necessidade:
conjuntos de dados

Introdução

- Mudança de realidade



Introdução



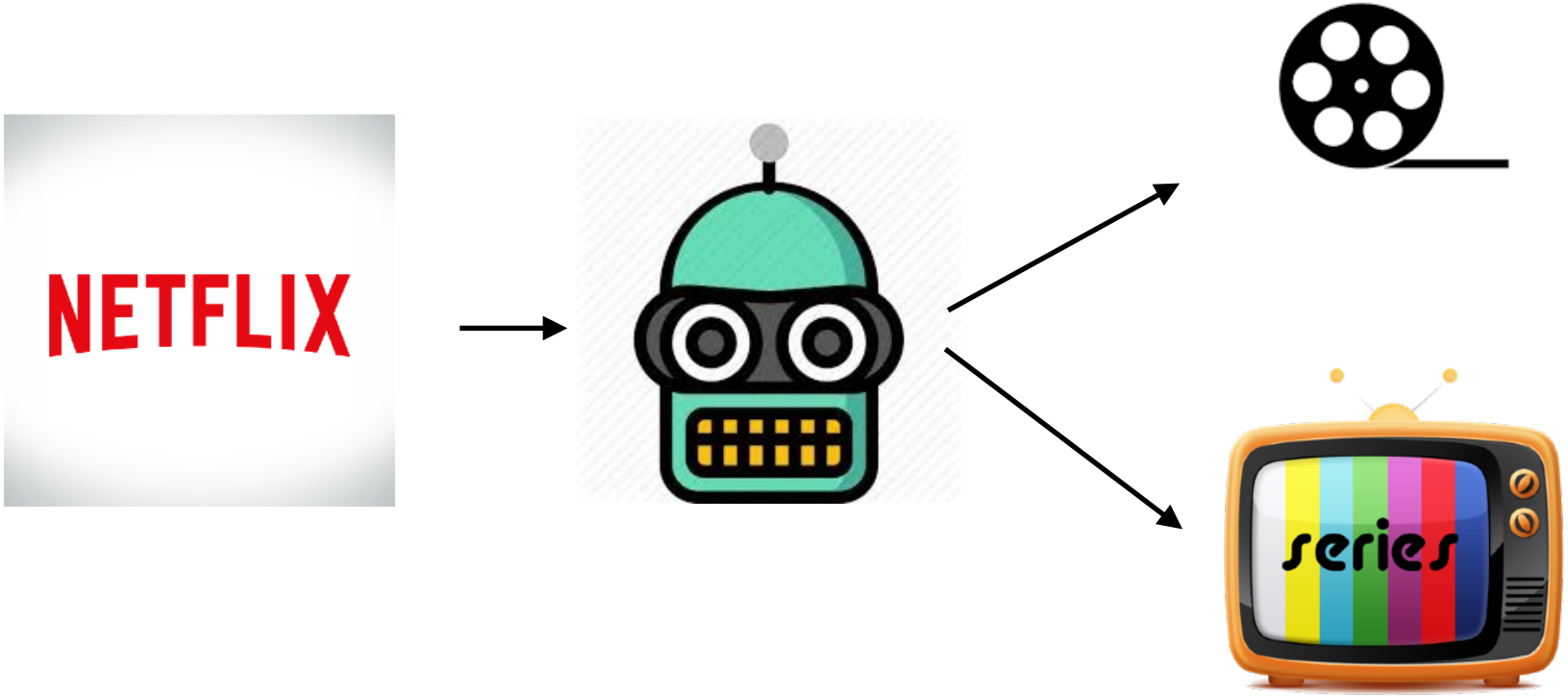
- Inteligência Artificial
- Automatiza a construção de modelos para solucionar problemas!

Introdução

- Onde isso é usado?

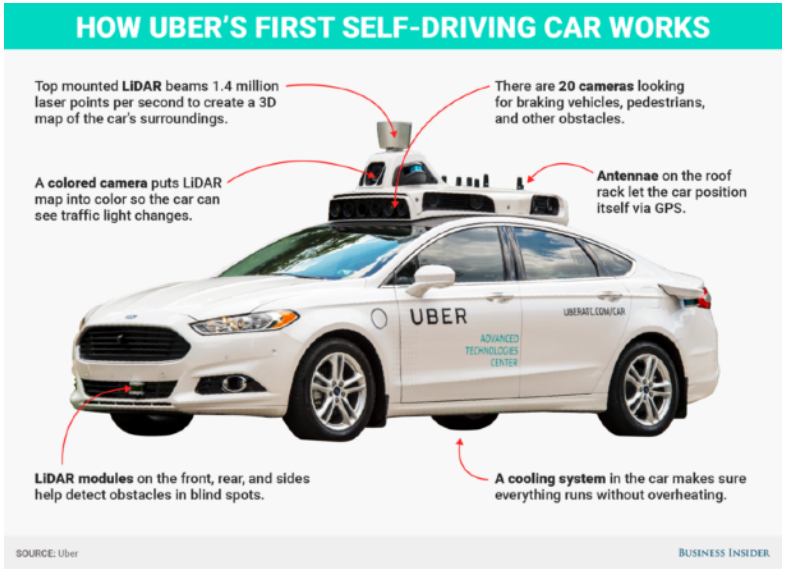
Introdução

- Onde isso é usado?



Introdução

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



Uber

Tesla

Introdução

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

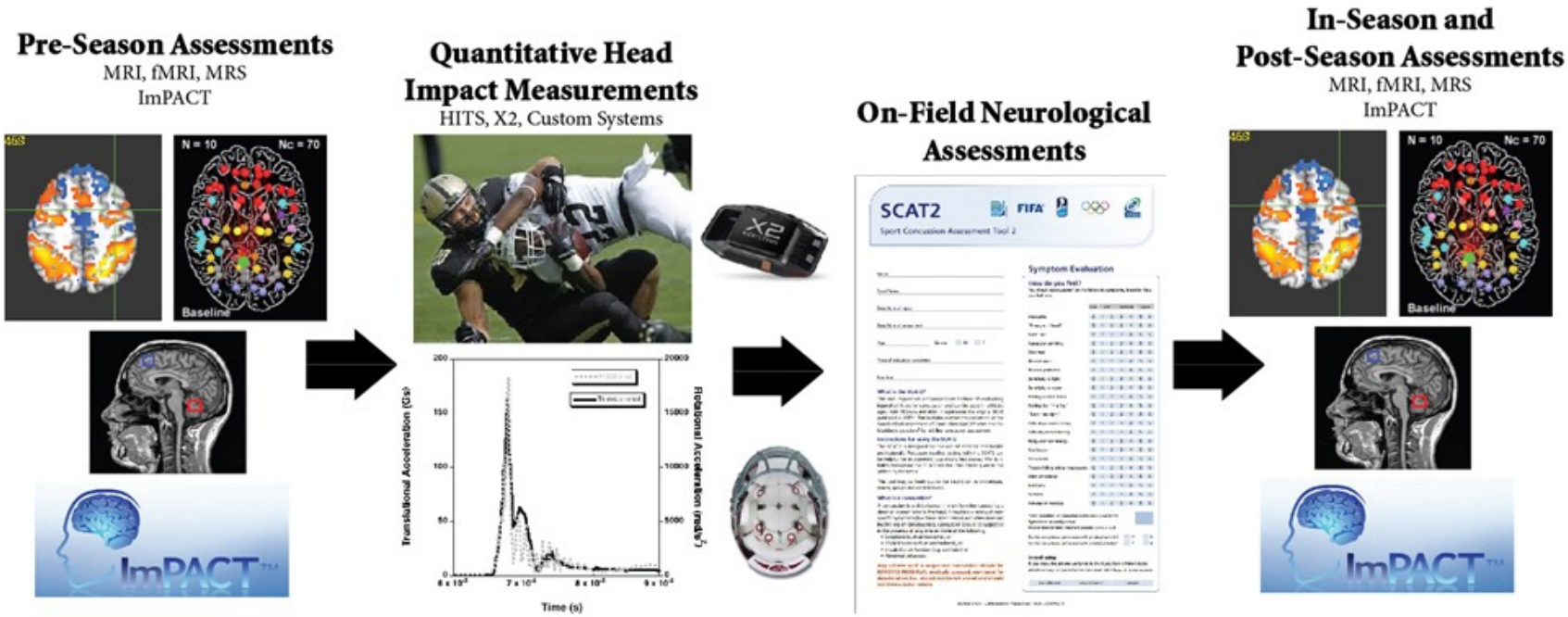
LRM - ICMC/USP, São Carlos - SP

Introdução

- Onde isso é usado? **Bancos**

Introdução

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



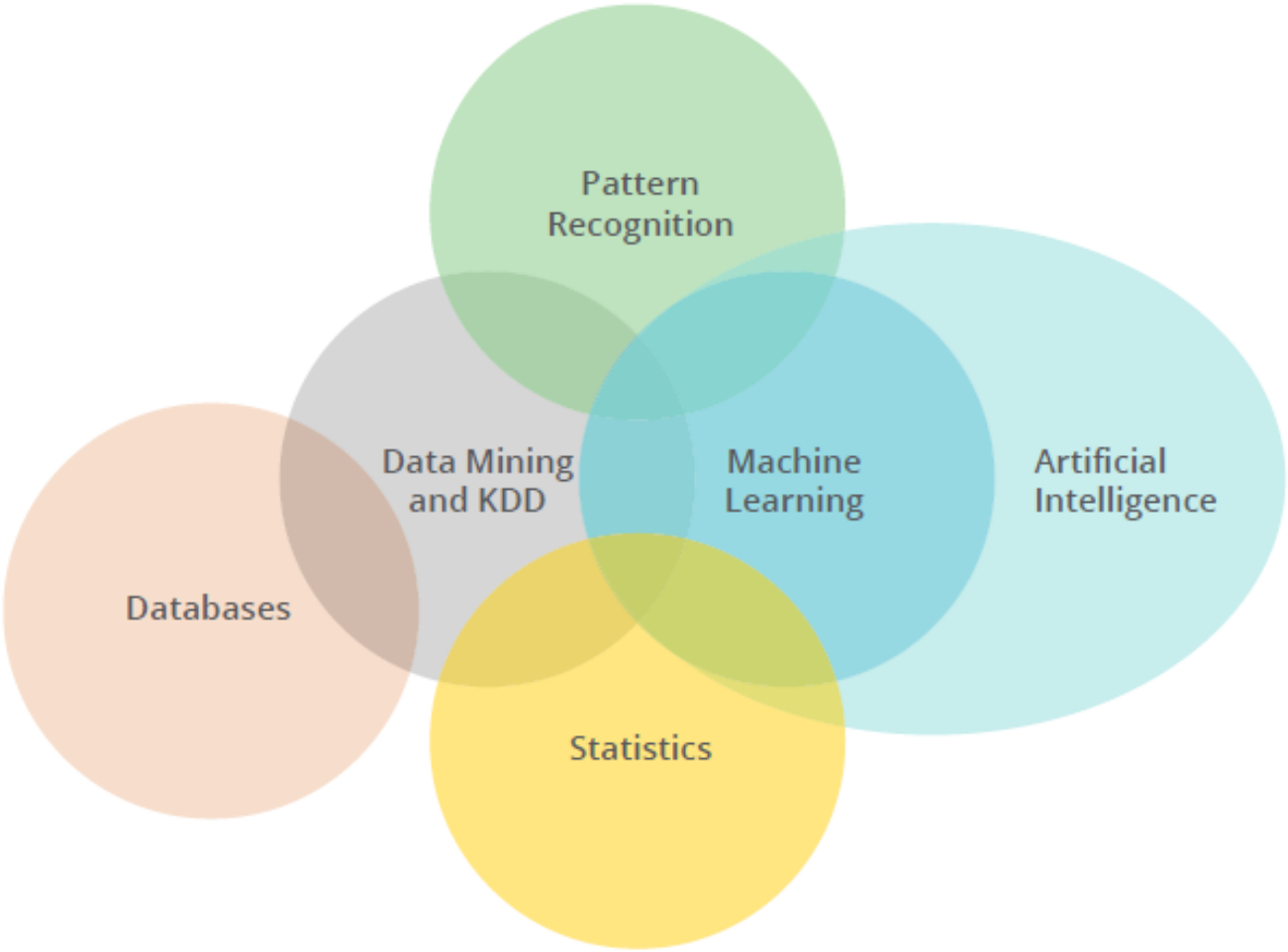
Introdução

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

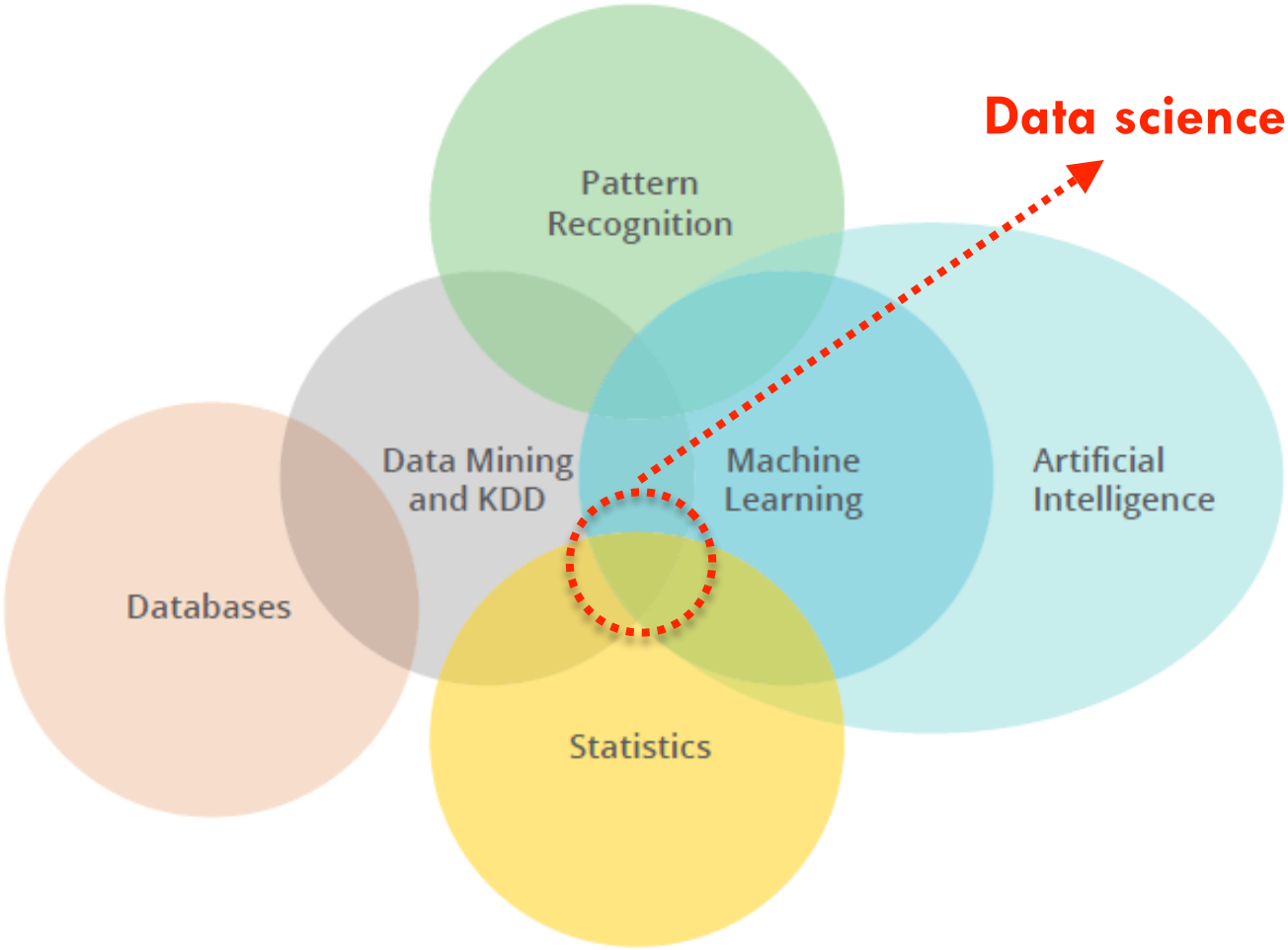
Roteiro

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

Conceitos Gerais



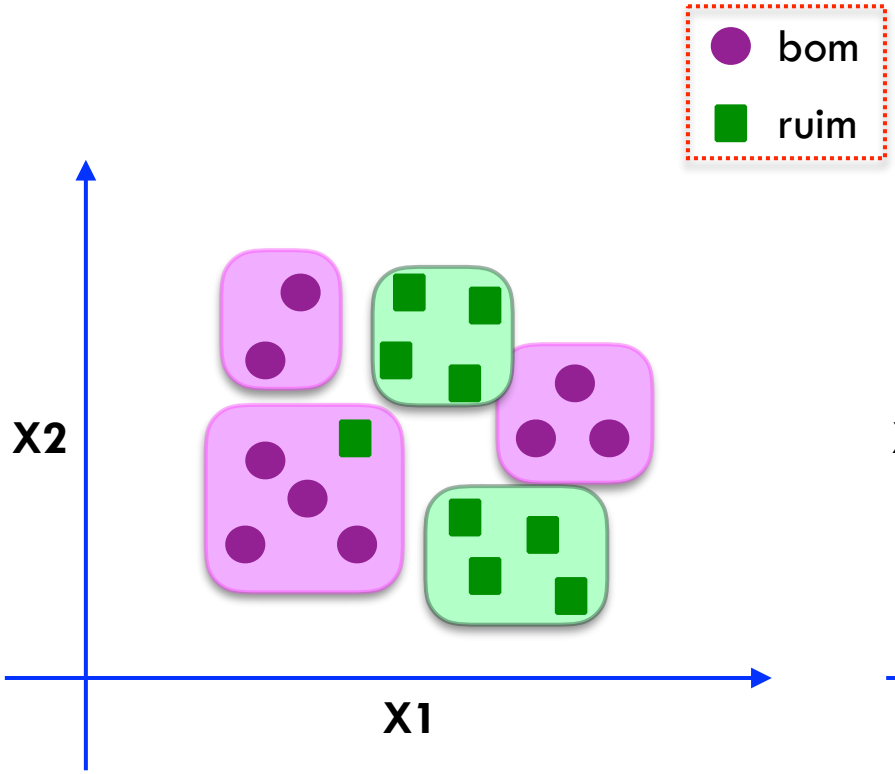
Conceitos Gerais



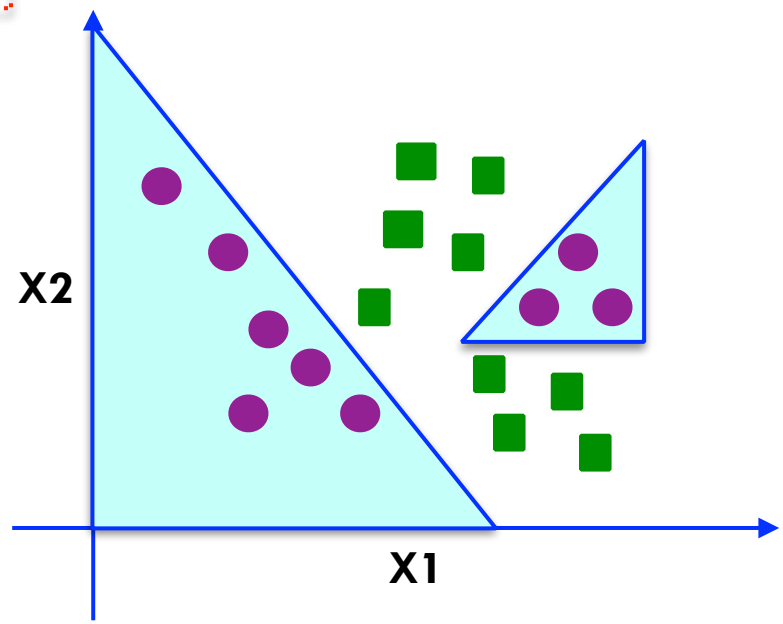
Mineração de dados

- Preparação dos dados:
  - imputação
  - normalização
  - transformações
  - …

Aprendizado de Máquina



Tarefa descritiva
(Clustering)



Tarefa preditiva
(Classificação/ Regressão)

Matemática / Estatística

- Amostragem
- Estatística descritiva - visualização
- Testes de Hipótese
- …

Ciência de Dados

- Quantos algoritmos existem?

Ciência de Dados

- Quantos algoritmos existem?

rapidminer

scikit learn

TensorFlow

Keras

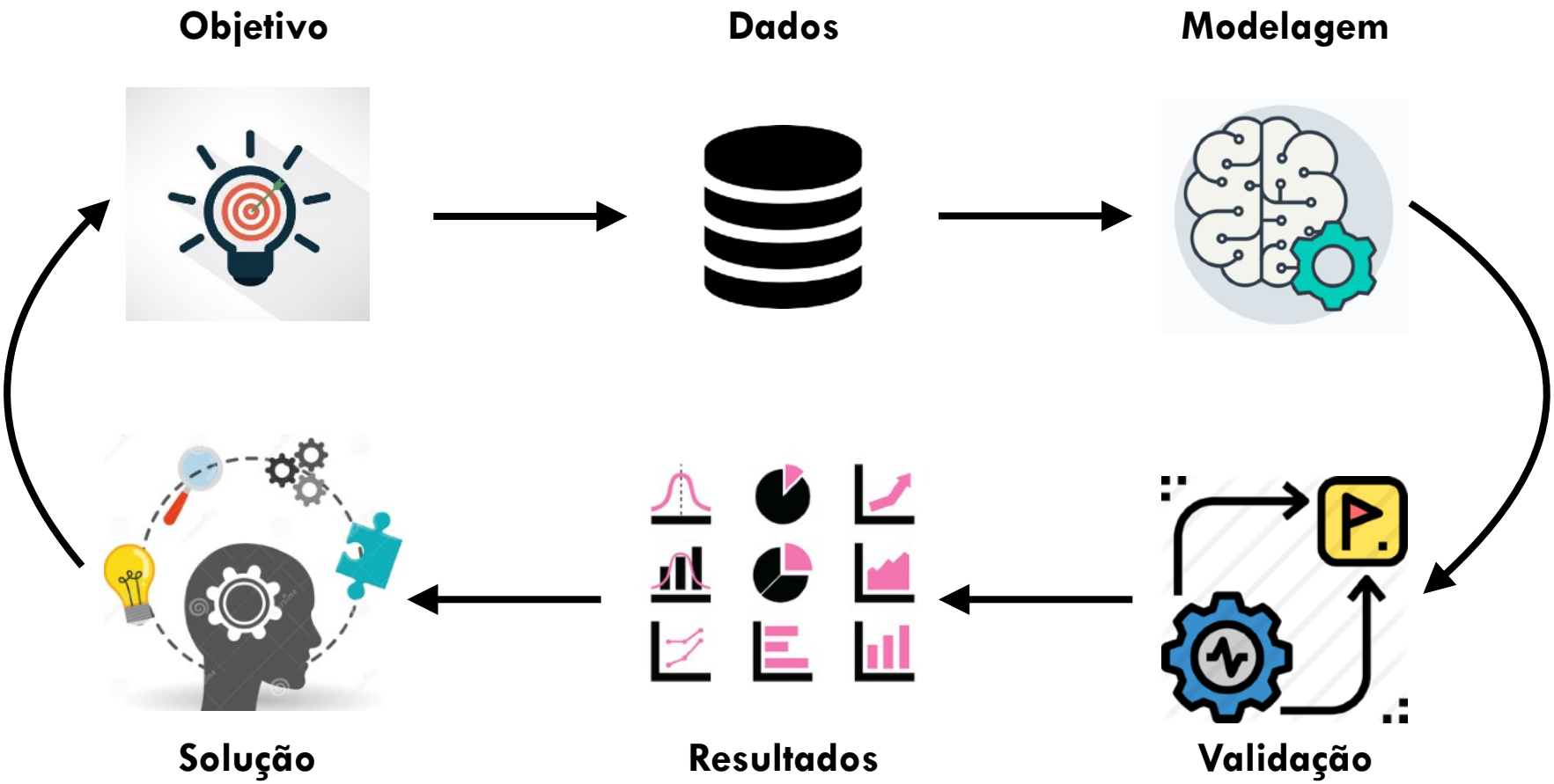
WEKA The University of Waikato

...

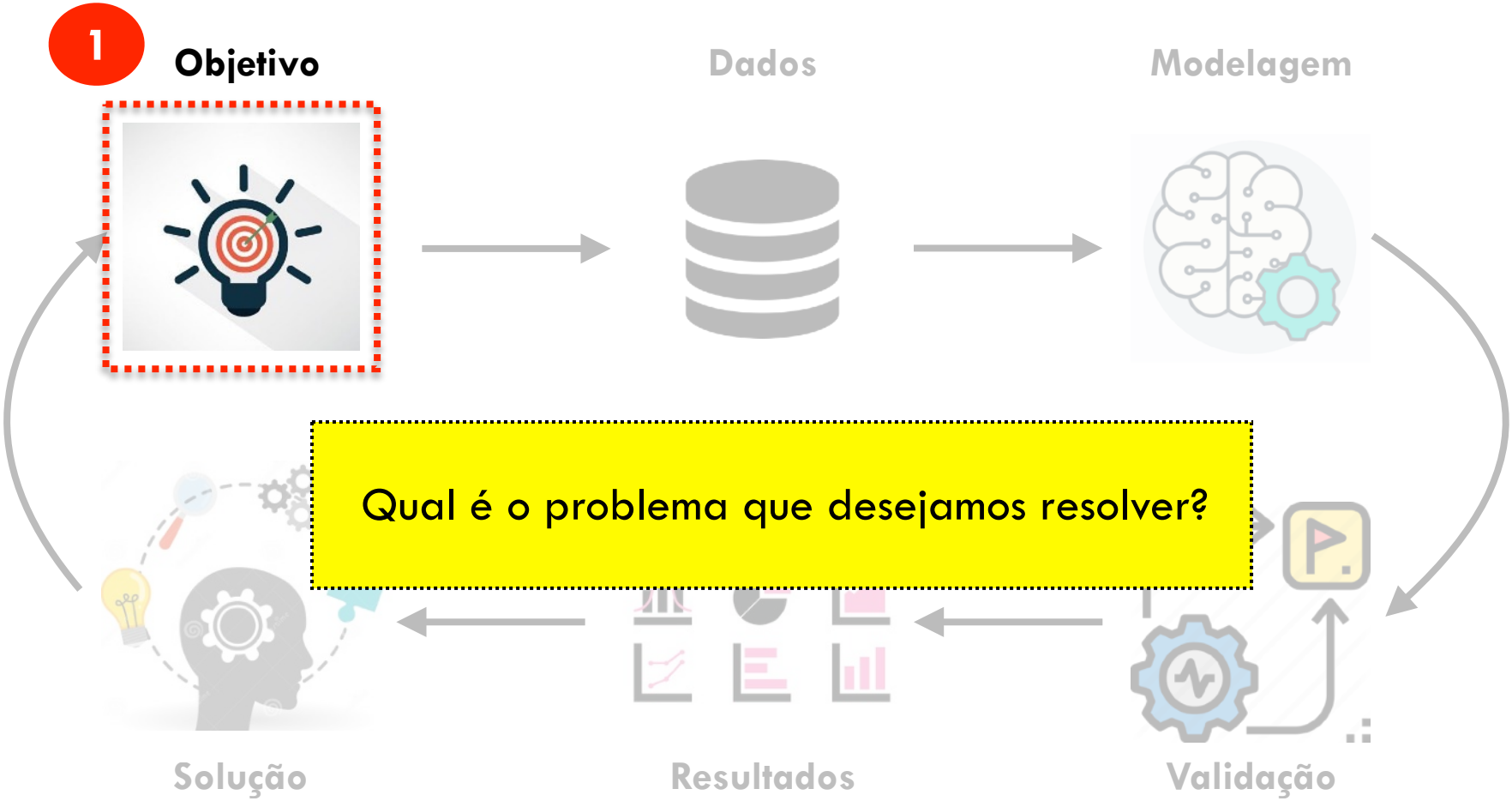
Roteiro

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

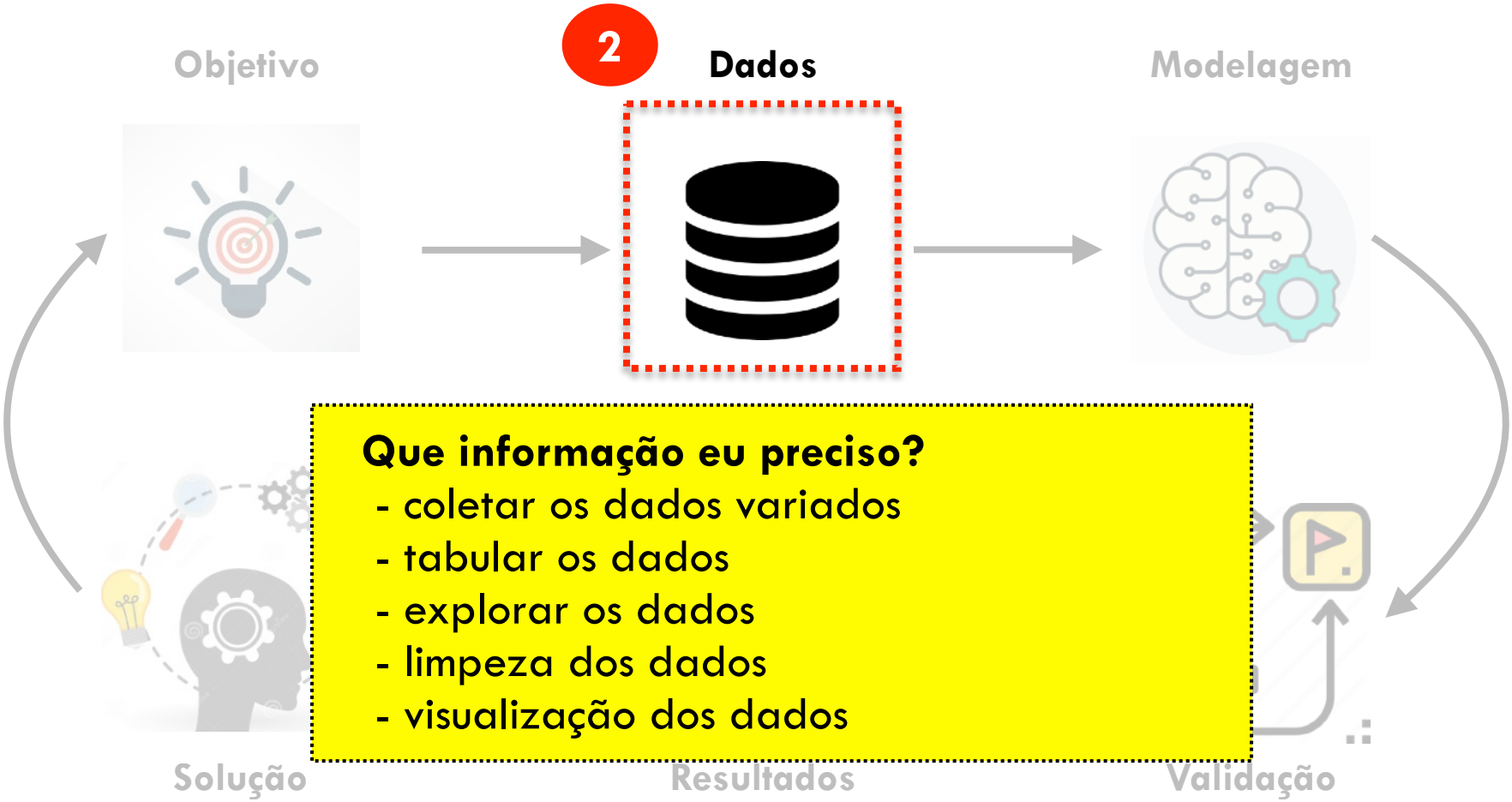
Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



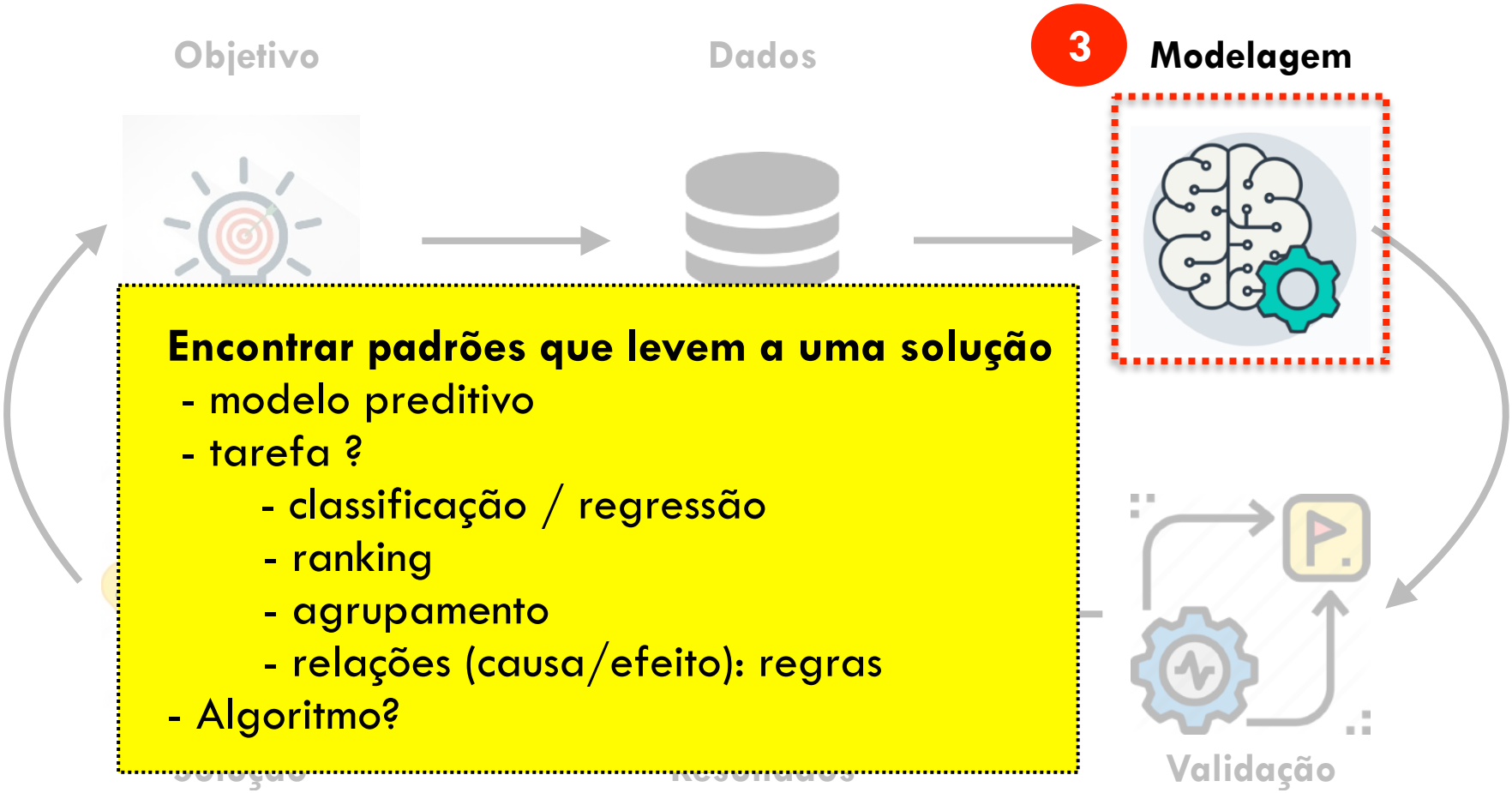
Fluxo de Ciência de Dados



Que informação eu preciso?

- coletar os dados variados
- tabular os dados
- explorar os dados
- limpeza dos dados
- visualização dos dados

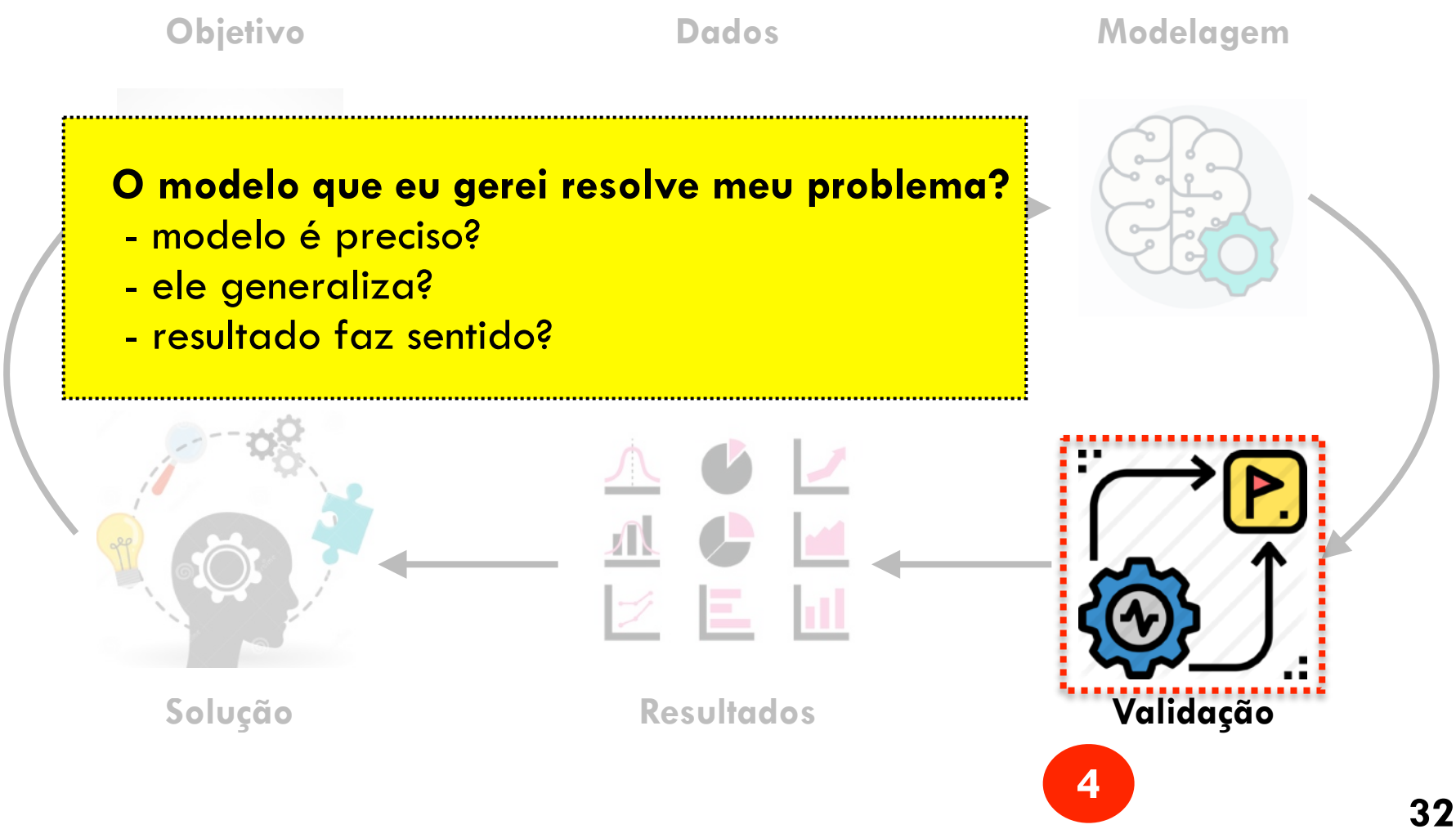
Fluxo de Ciência de Dados



Encontrar padrões que levem a uma solução

- modelo preditivo
- tarefa ?
  - classificação / regressão
  - ranking
  - agrupamento
  - relações (causa/efeito): regras
- Algoritmo?

Fluxo de Ciência de Dados



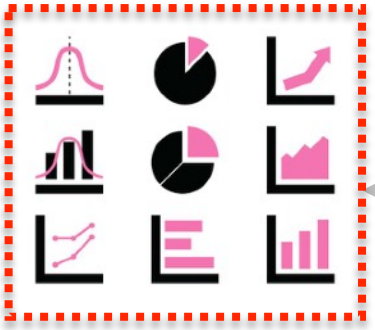
O modelo que eu gerei resolve meu problema?
- modelo é preciso?
- ele generaliza?
- resultado faz sentido?

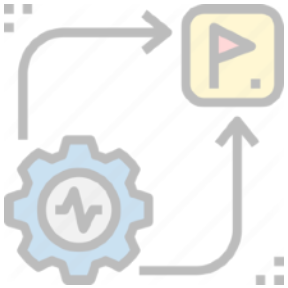
Fluxo de Ciência de Dados

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

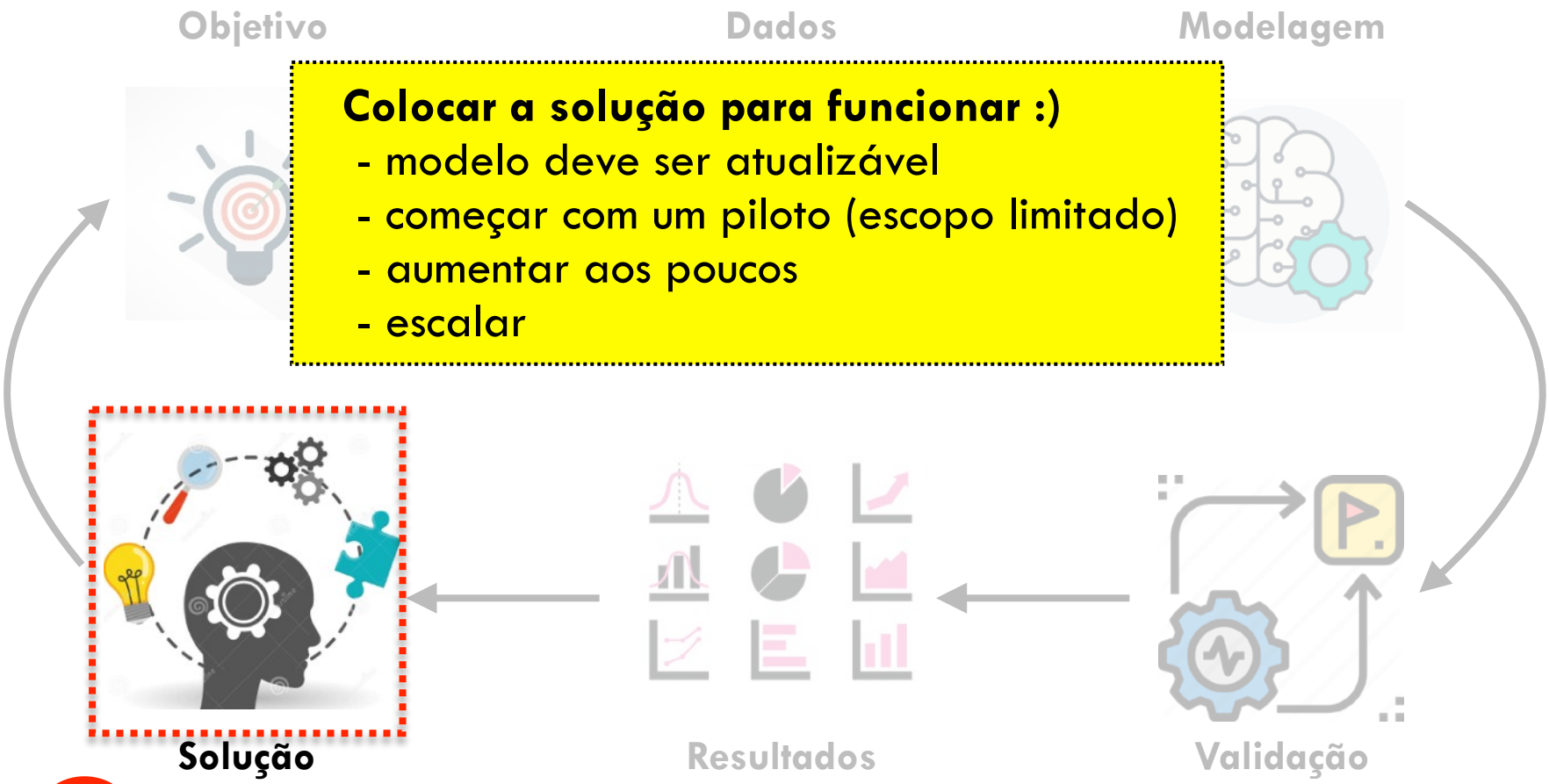




Solução **Resultados** Validação

5

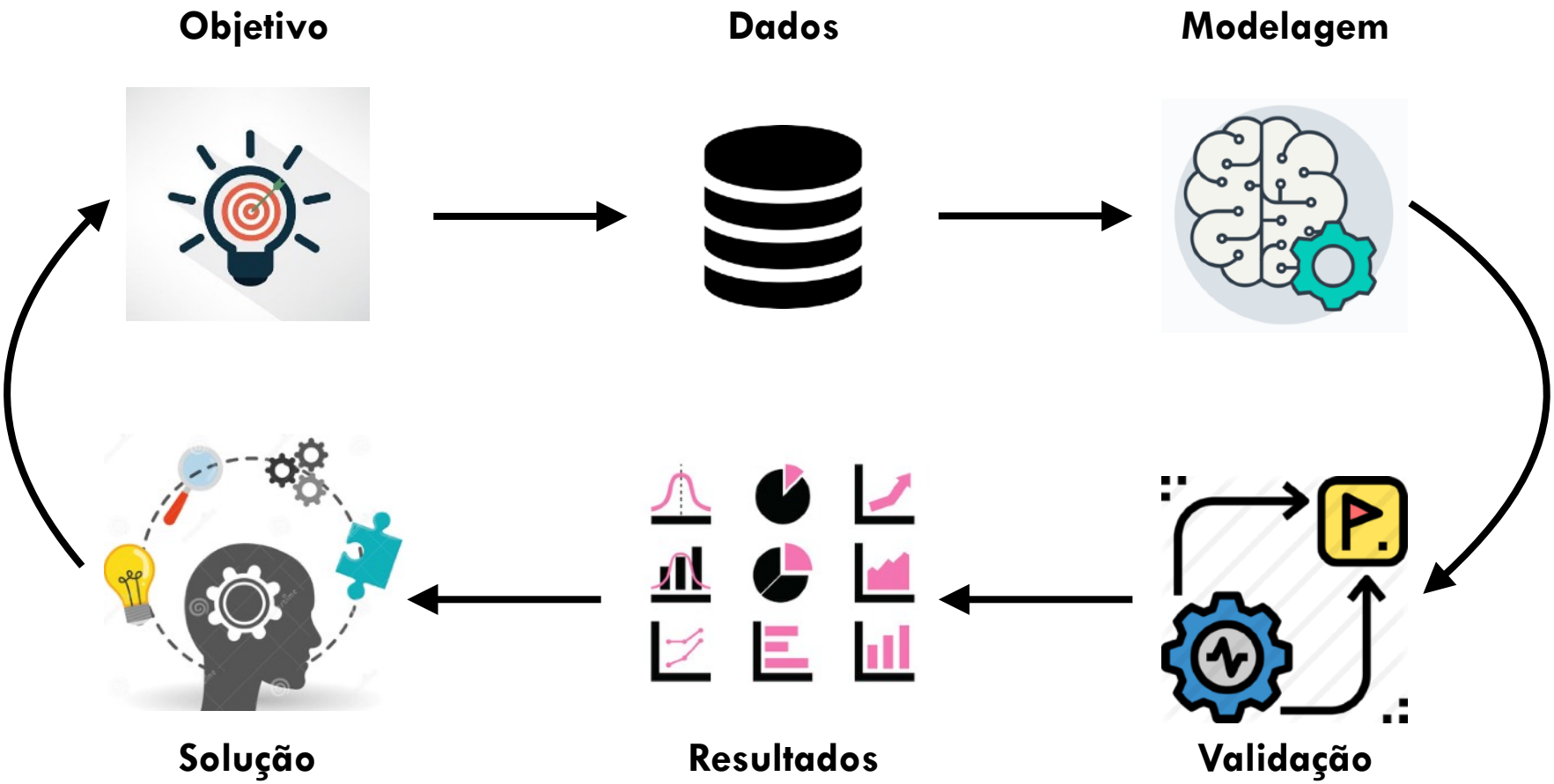
Fluxo de Ciência de Dados



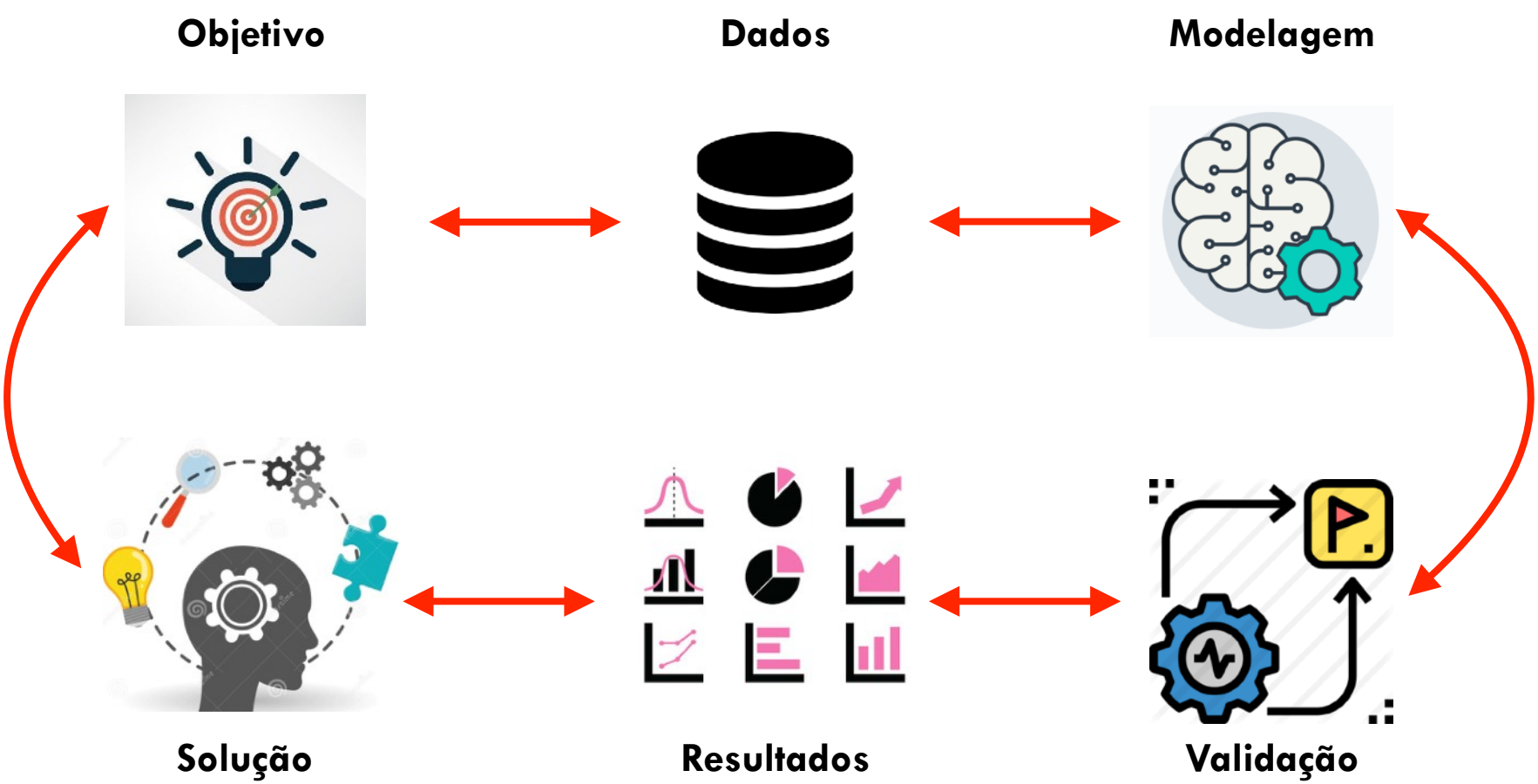
Colocar a solução para funcionar :)

- modelo deve ser atualizável
- começar com um piloto (escopo limitado)
- aumentar aos poucos
- escalar

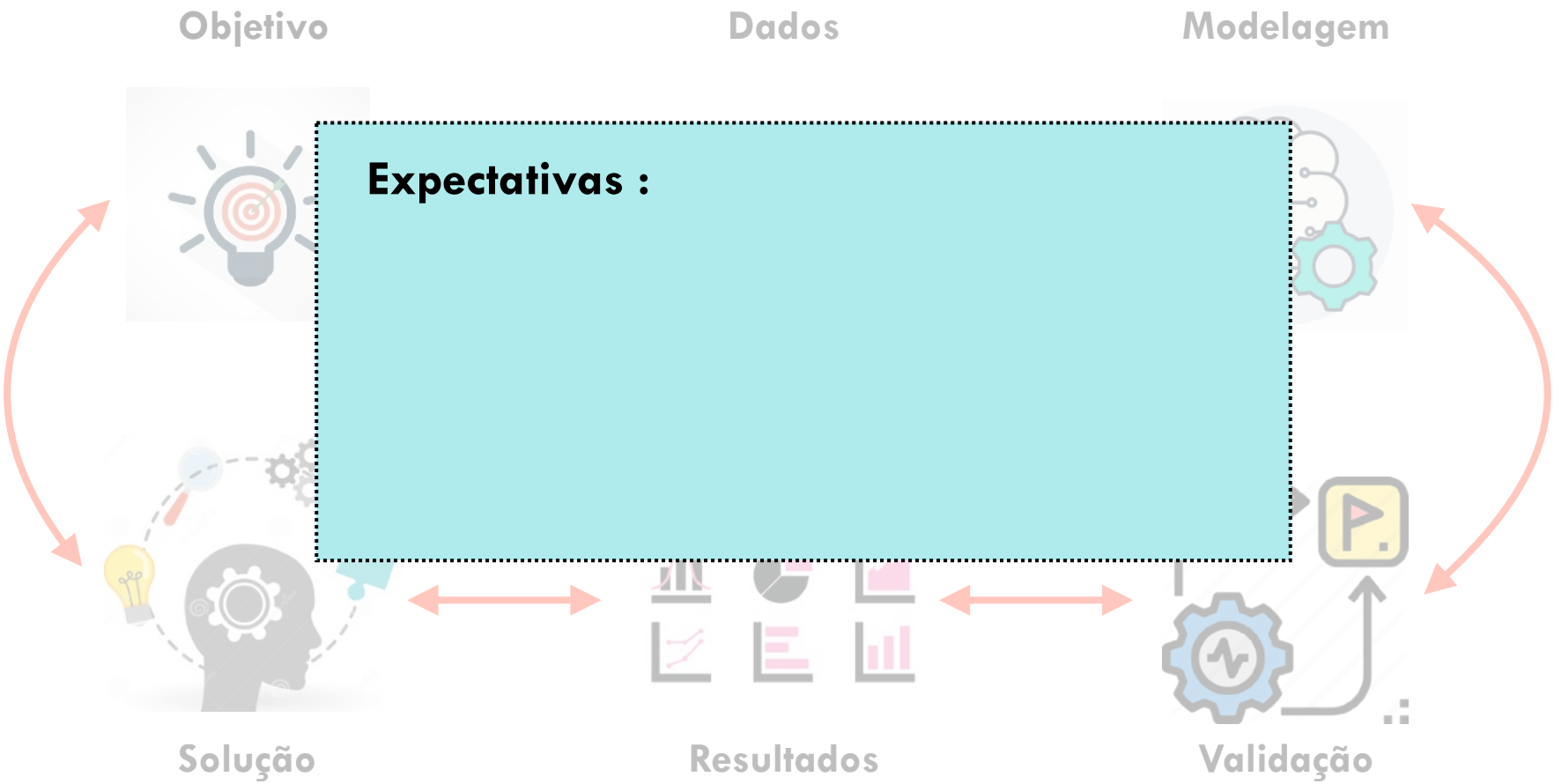
Fluxo de Ciência de Dados



Fluxo de Ciência de Dados

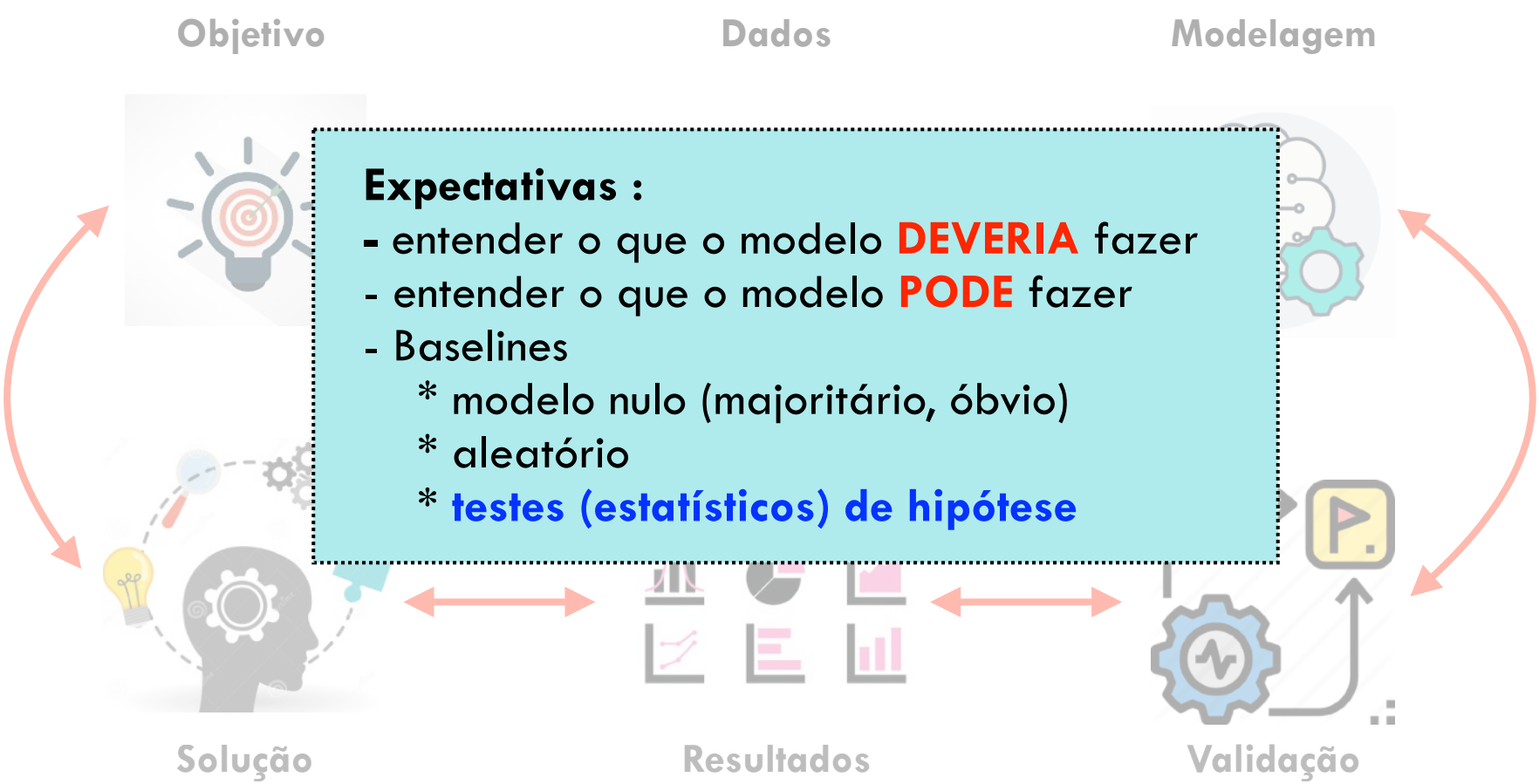


Fluxo de Ciência de Dados



Expectativas :

Fluxo de Ciência de Dados



Expectativas :

- entender o que o modelo **DEVERIA** fazer
- entender o que o modelo **PODE** fazer
- Baselines
  * modelo nulo (majoritário, óbvio)
  * aleatório
  * **testes (estatísticos) de hipótese**

Roteiro

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

Ferramentas

OpenML

R Studio

ggplot2

mlR

Ferramentas

OpenML

RStudio®

ggplot2

mlR

OpenML / Dados

OpenML / Dados

7983 results

FILTERS FOR SEARCH OPTIONS, SE

iris (4)

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

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 patterr

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 patterr

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 patterr

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

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

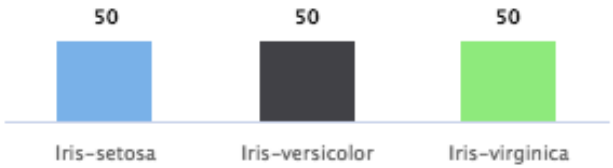
Help us complete this description → Edit

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

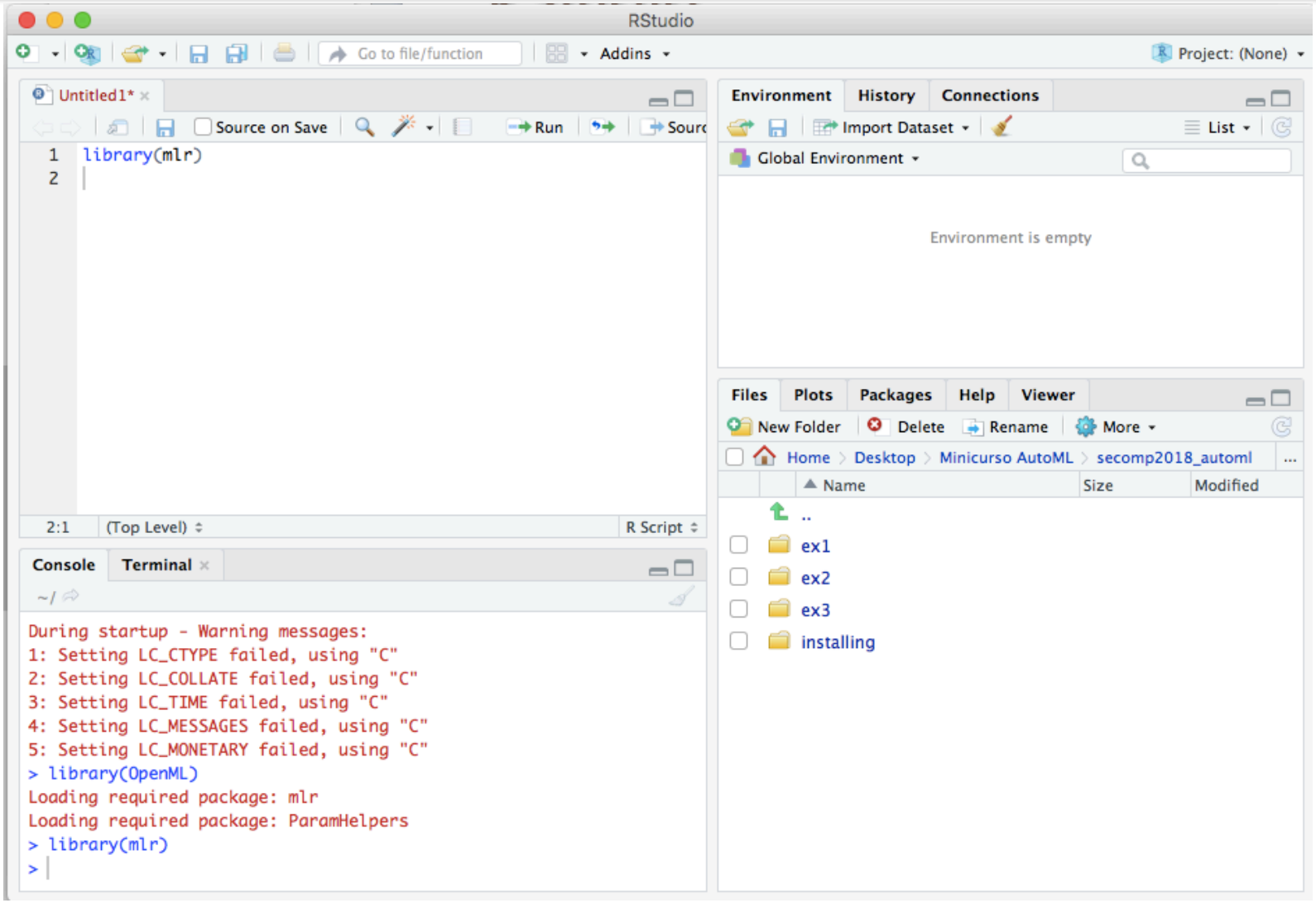
OpenML

RStudio

ggplot2

mlR

Studio / IDE para R



Ferramentas

OpenML

R Studio

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

R Studio

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 Graphi

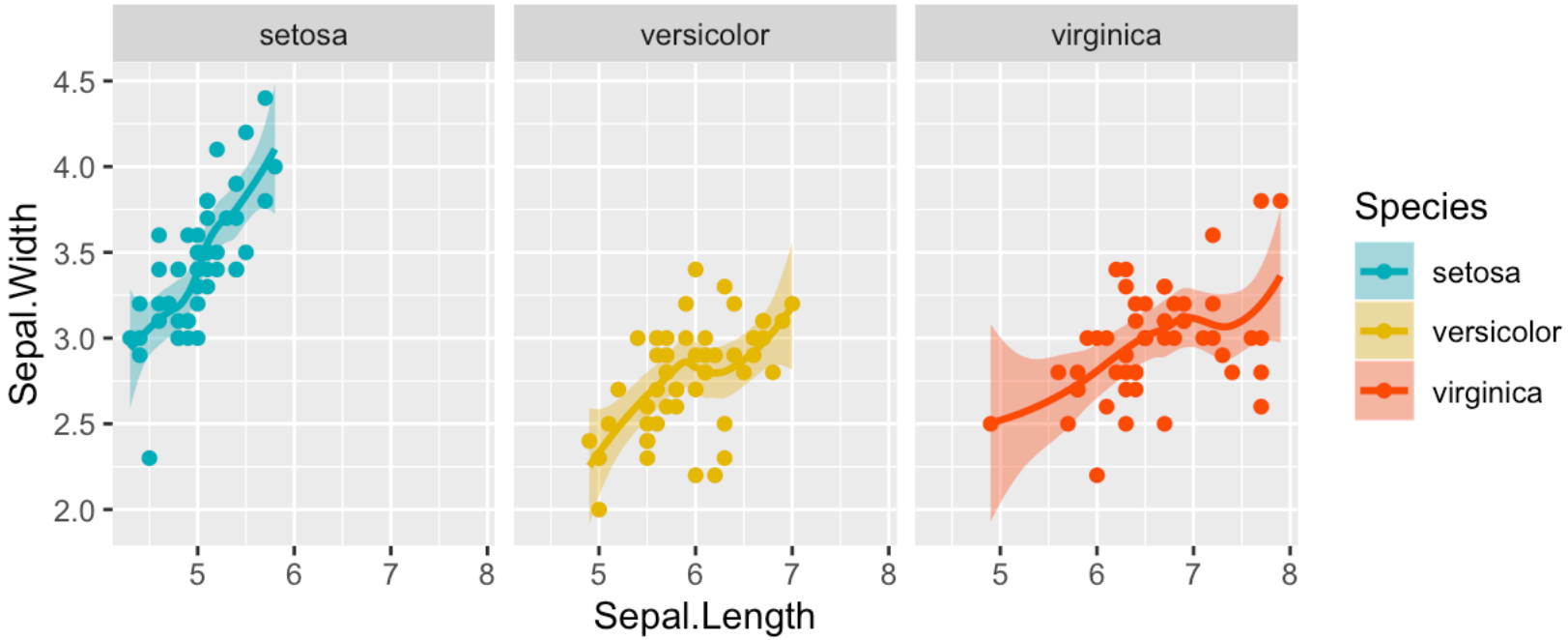
how to map variables to aesthetics, what graphical primitives to use, and it takes care of

Installation

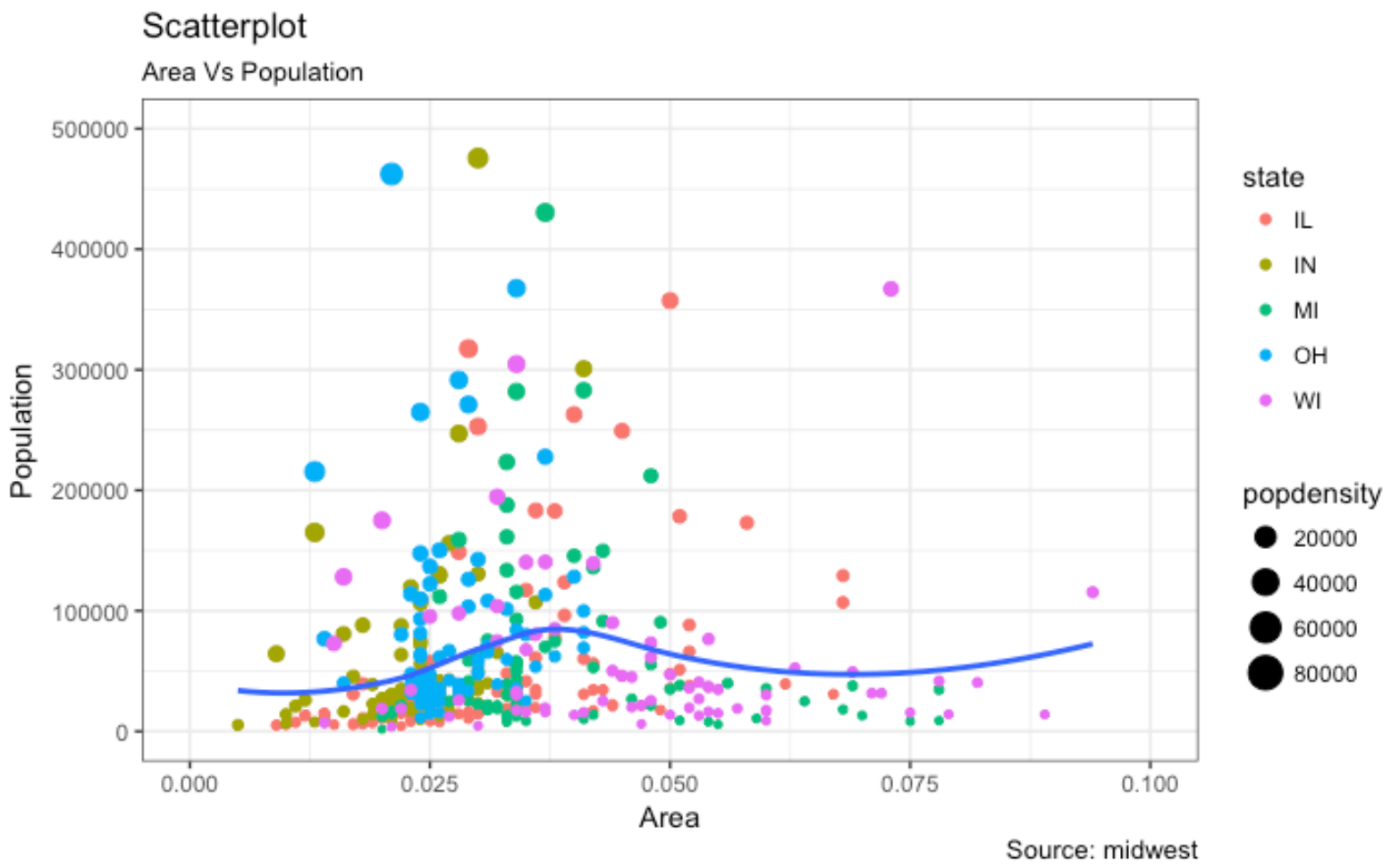
```
# The easiest way to get ggplot2 is to install the whole tidyverse:
install.packages("tidyverse")

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

ggplot2

- Visualização dos dados :)



ggplot2

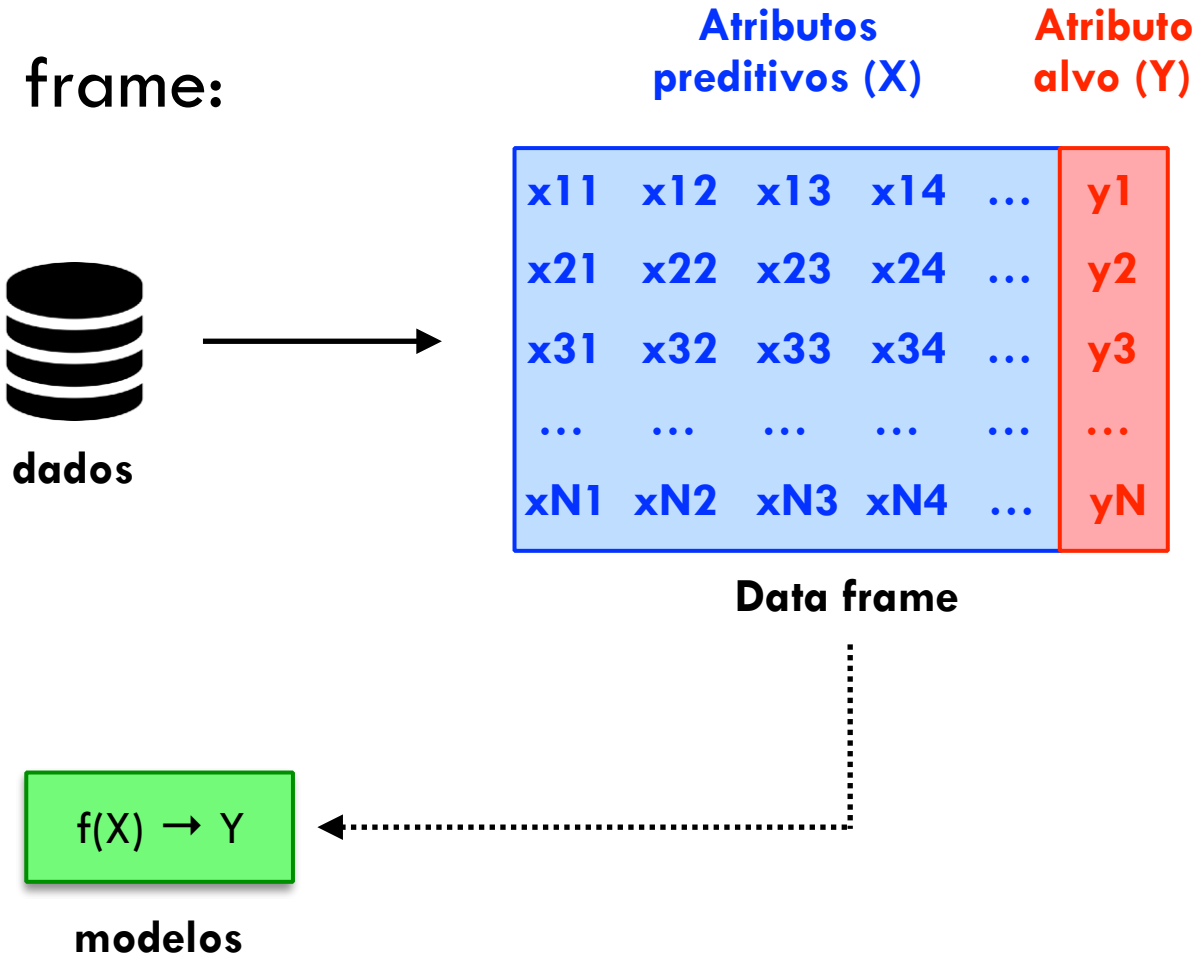

Scatterplot
Area Vs Population
Population
500000
400000
300000
200000
100000
0
0.000
0.025
0.050
0.075
0.100
Area
Source: midwest
state
IL
IN
MI
OH
WI
popdensity
20000
40000
60000
80000

Roteiro

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

Um pouco de R :)

- Data frame:



Um pouco de R :)

- Data frame:

Iris

Atributos preditivos (X)

Atributo alvo (Y)

	sepallength	sepalwidth	petallength	petalwidth	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa

Showing 1 to 16 of 150 entries

Hello world

- O que faremos?
  - ler dados no R
  - ver características dos dados
  - plotar

Hello world

- O que faremos?
  - ler dados no R
  - ver características dos dados
  - plotar

helloWorld.R

helloWorld.R

Branch: master secomp2019_utfpr_ap / codes / 01_initialCodes / helloWorld.R

rgmantovani adding content

1 contributor

49 lines (33 sloc) | 823 Bytes

```
# carregando o pacote ggplot2
library(ggplot2)

# acessando o dataset
mpg
View(mpg)

# contando numero de linhas do dataset
nrow(mpg)
```

Exercício 01

- Fazer o mesmo com o dataset: iris
  - visualizar par a par as coordenadas
  - o que pode ser visto?

- **Usar:**
  - **library**(ggplot2)
  - plot( )
  - geom_point( )

Exercício 02

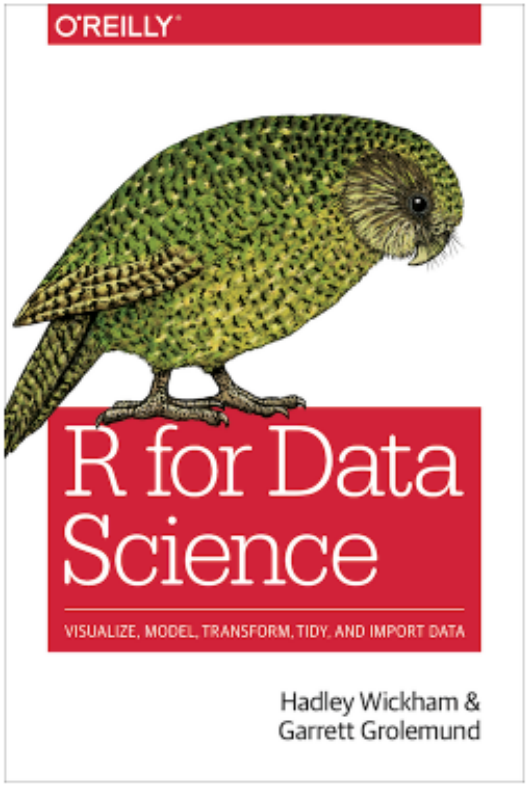
- Encontrar outro dataset no OpenML
  - baixar e ler no R
  - visualizar informação
  - Qual informação o dataset nos mostra?

- **Usar:**
  - library(ggplot2)
  - library(OpenML)
  - read.csv ( )
  - read.table ( )
  - read.arff ( )
  - getOMLDataSet(data.id = <id>)

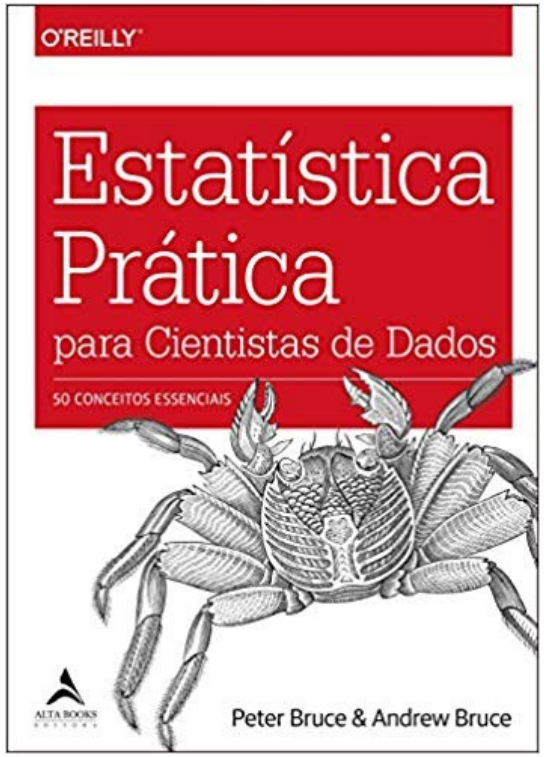
Roteiro

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

Referências

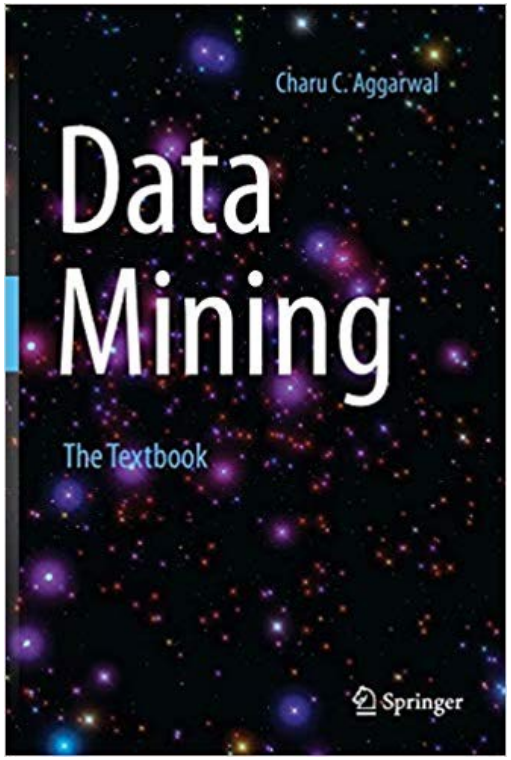


[Wickham & Grolemund, 2018]

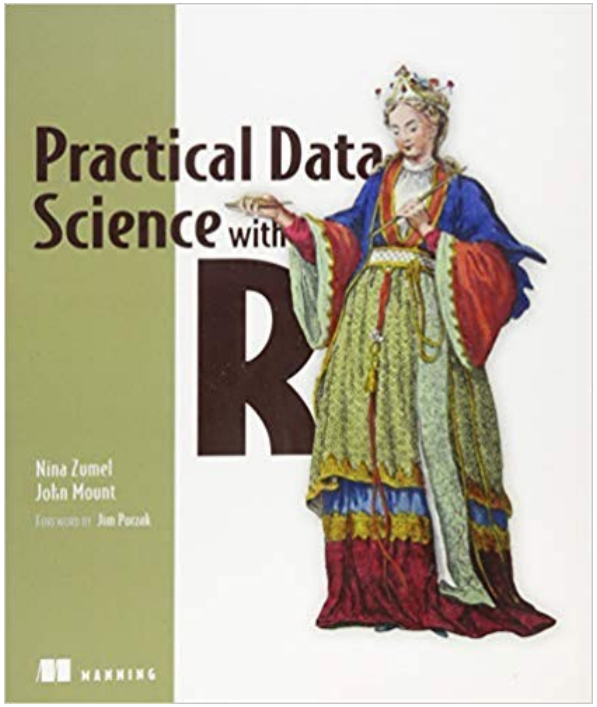


[Bruce & Bruce, 2019]

Referências



[Aggarwal, 2015]



[Zumel and Mount, 2014]

Perguntas?

Prof. Rafael G. **Mantovani**

rafaelmantovani@utfpr.edu.br