

# MÉTODOS E MODELOS AVANÇADOS EM CIÊNCIA DE DADOS

Aula 08 - Fluxo de Ciência de Dados

Prof. Rafael G. Mantovani



Universidade Tecnológica Federal do Paraná (UTFPR)  
Especialização em Ciência de Dados

# Roteiro

- 1 Introdução**
- 2 Aplicações**
- 3 Conceitos gerais**
- 4 Fluxo de ciência de dados**
- 5 Síntese**
- 6 Referências**

# Roteiro

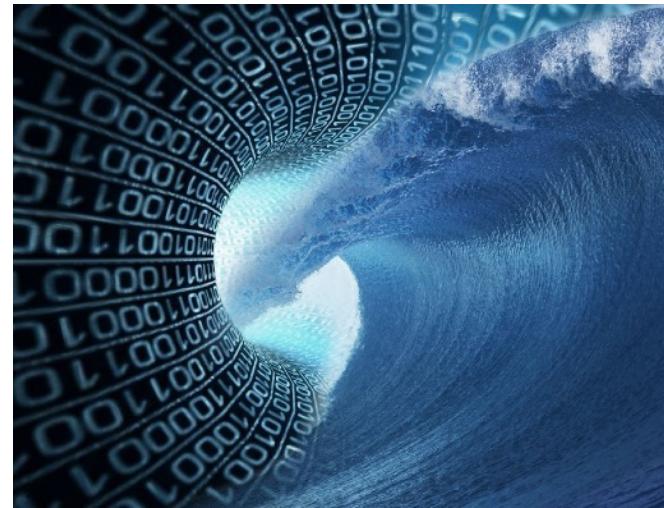
- 1 Introdução**
- 2 Aplicações**
- 3 Conceitos gerais**
- 4 Fluxo de ciência de dados**
- 5 Síntese**
- 6 Referências**

# Introdução



**poucos dados**

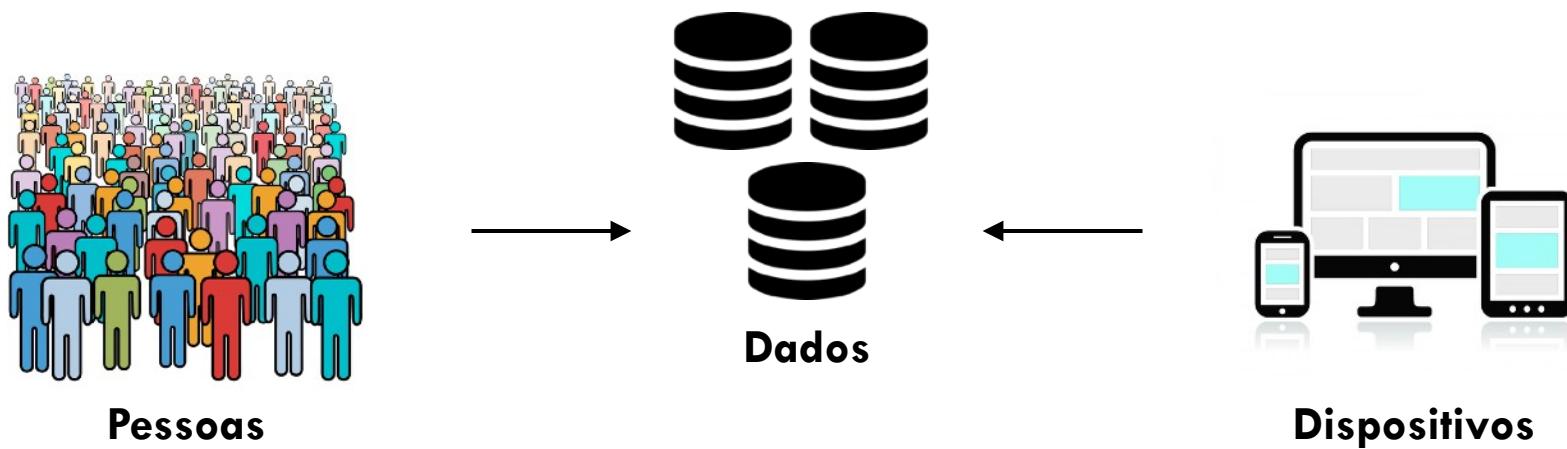
# Introdução



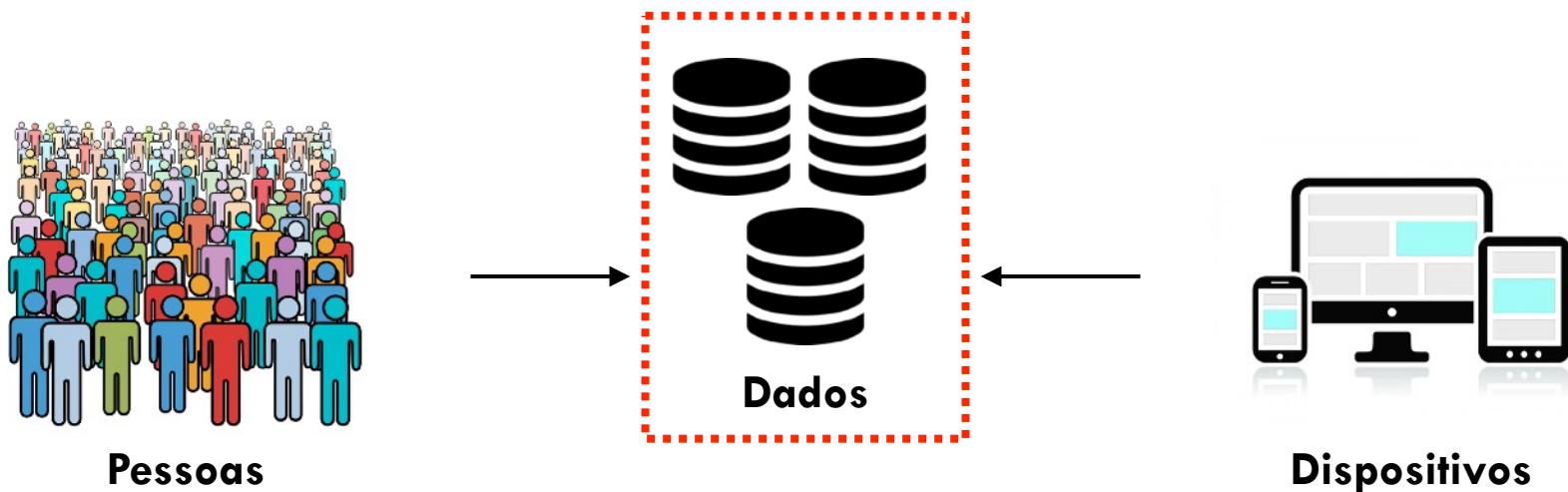
**poucos dados**

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

# Introdução



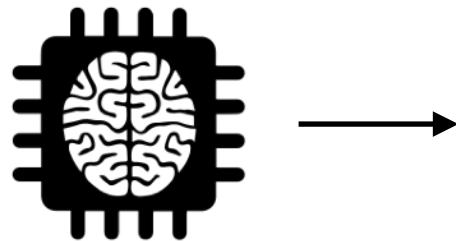
# Introdução



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

# Introdução

- Mudança de realidade



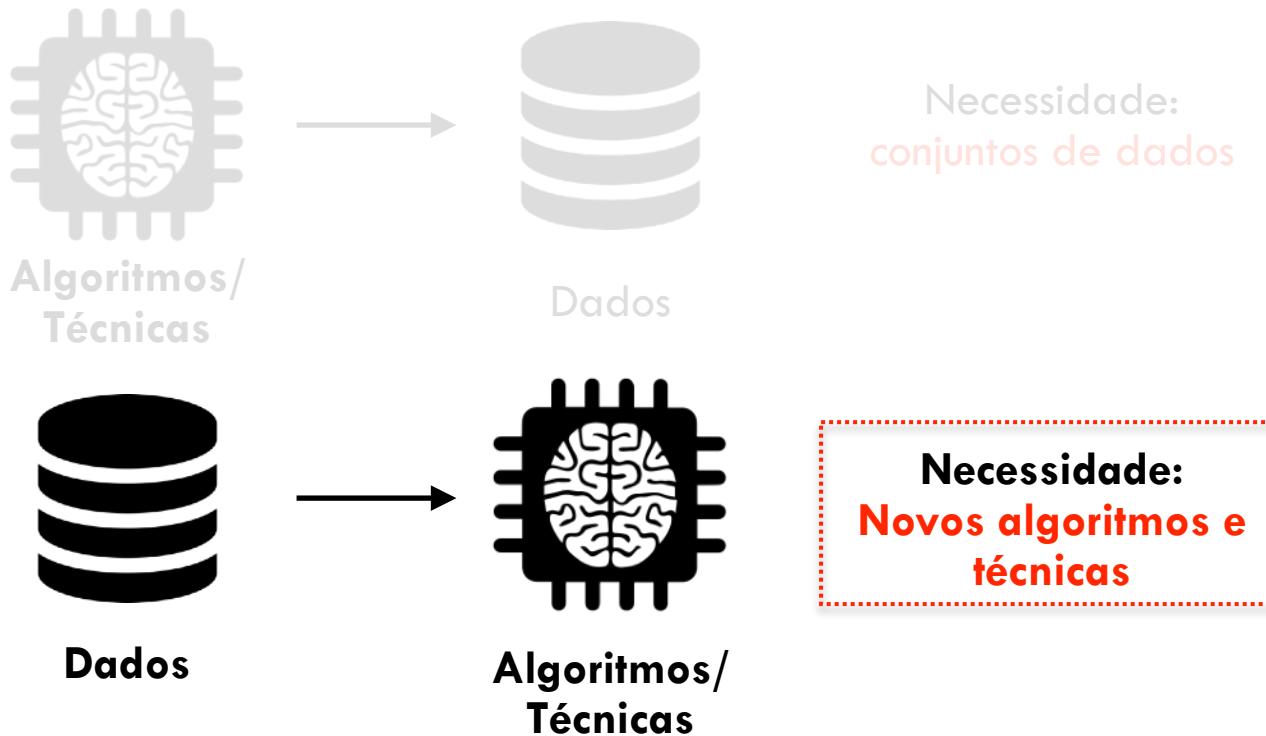
Algoritmos/  
Técnicas

Dados

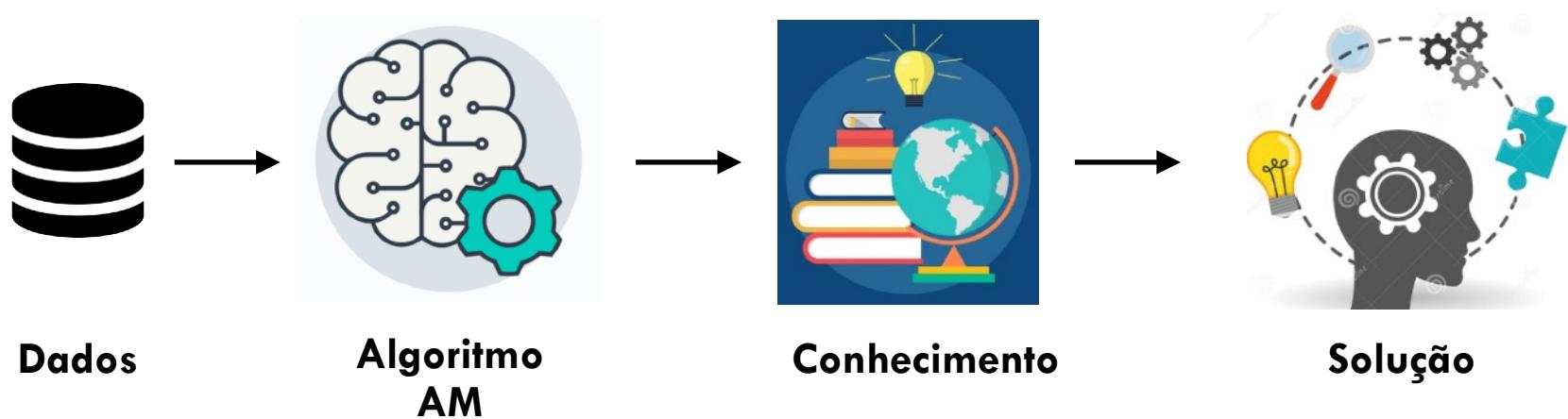
Necessidade:  
**conjuntos de dados**

# Introdução

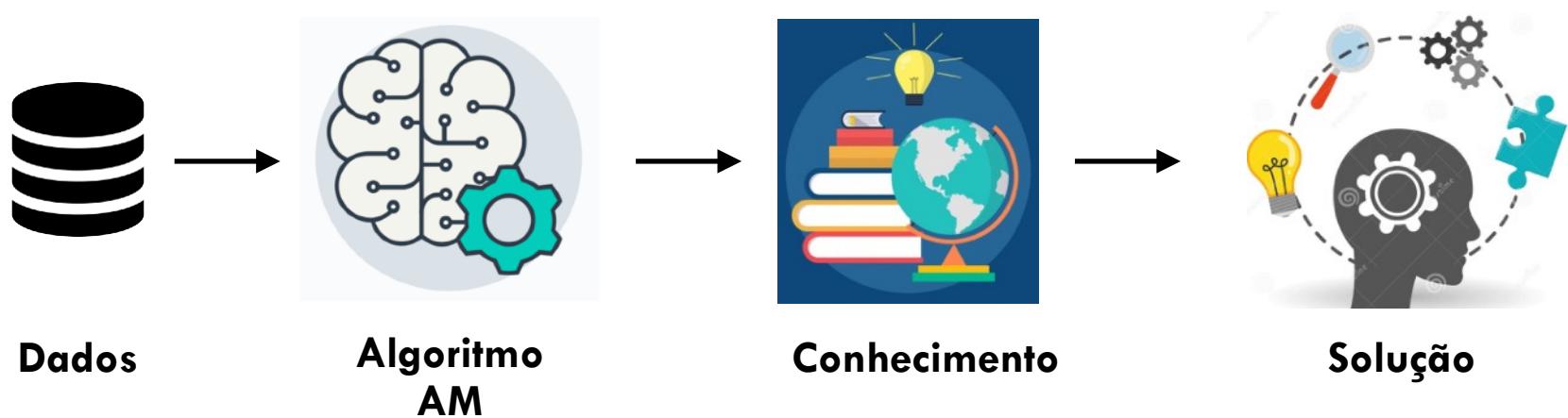
- Mudança de realidade



# Introdução



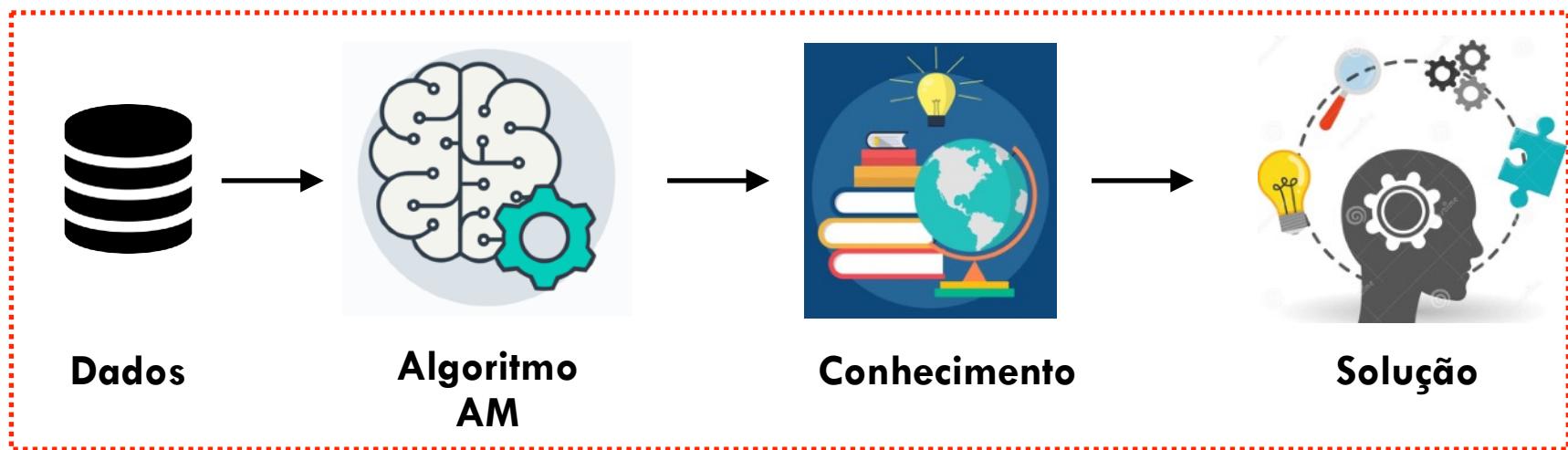
# Introdução



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

# Introdução

Pipeline  
End-to-end solution



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

# Roteiro

- 1 Introdução**
- 2 Aplicações**
- 3 Conceitos gerais**
- 4 Fluxo de ciência de dados**
- 5 Síntese**
- 6 Referências**

# Aplicações

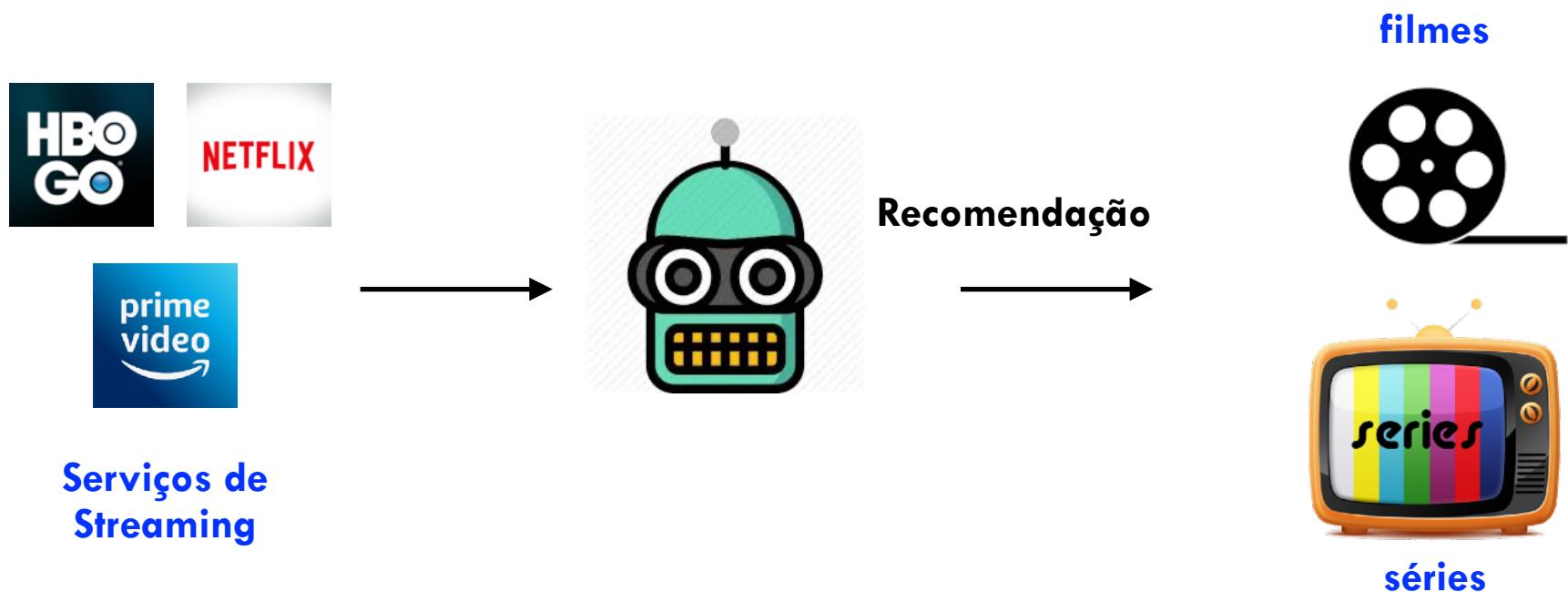
- Onde isso é usado?

# Aplicações

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

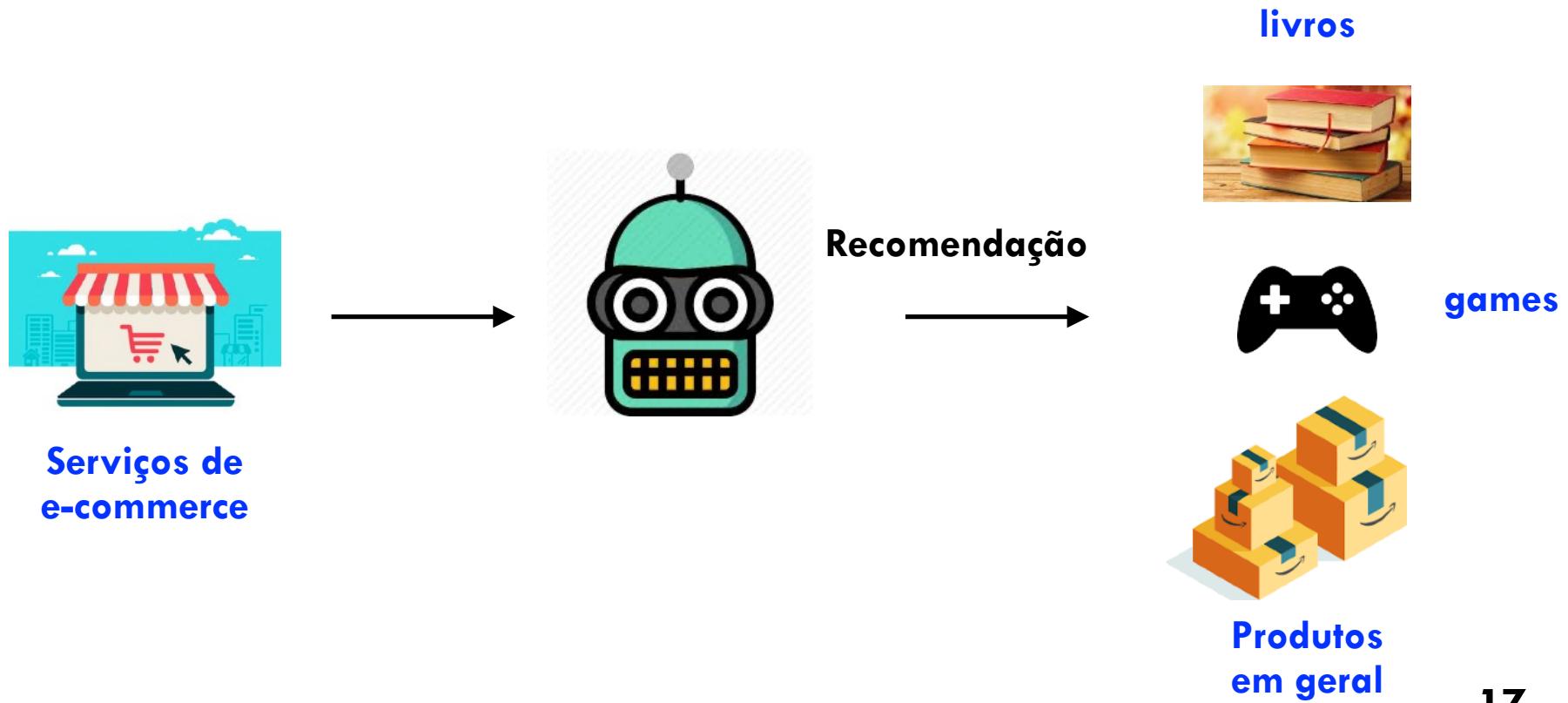
# Aplicações

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



# Aplicações

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



# Aplicações

- Onde isso é usado? **Bancos/Mercado financeiro**

# Aplicações

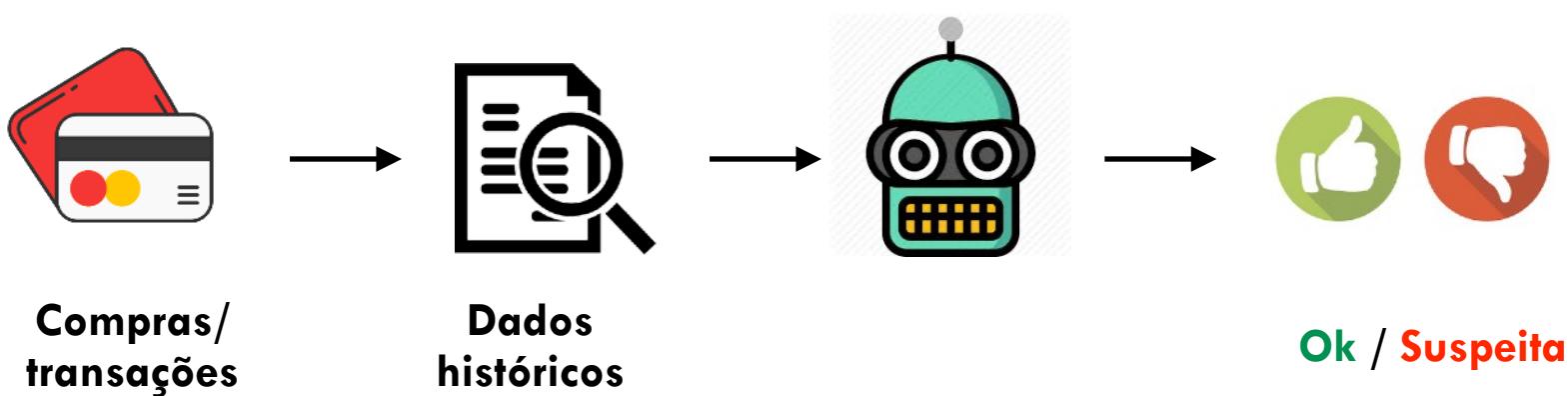
- Onde isso é usado? **Bancos/Mercado financeiro**



# Aplicações

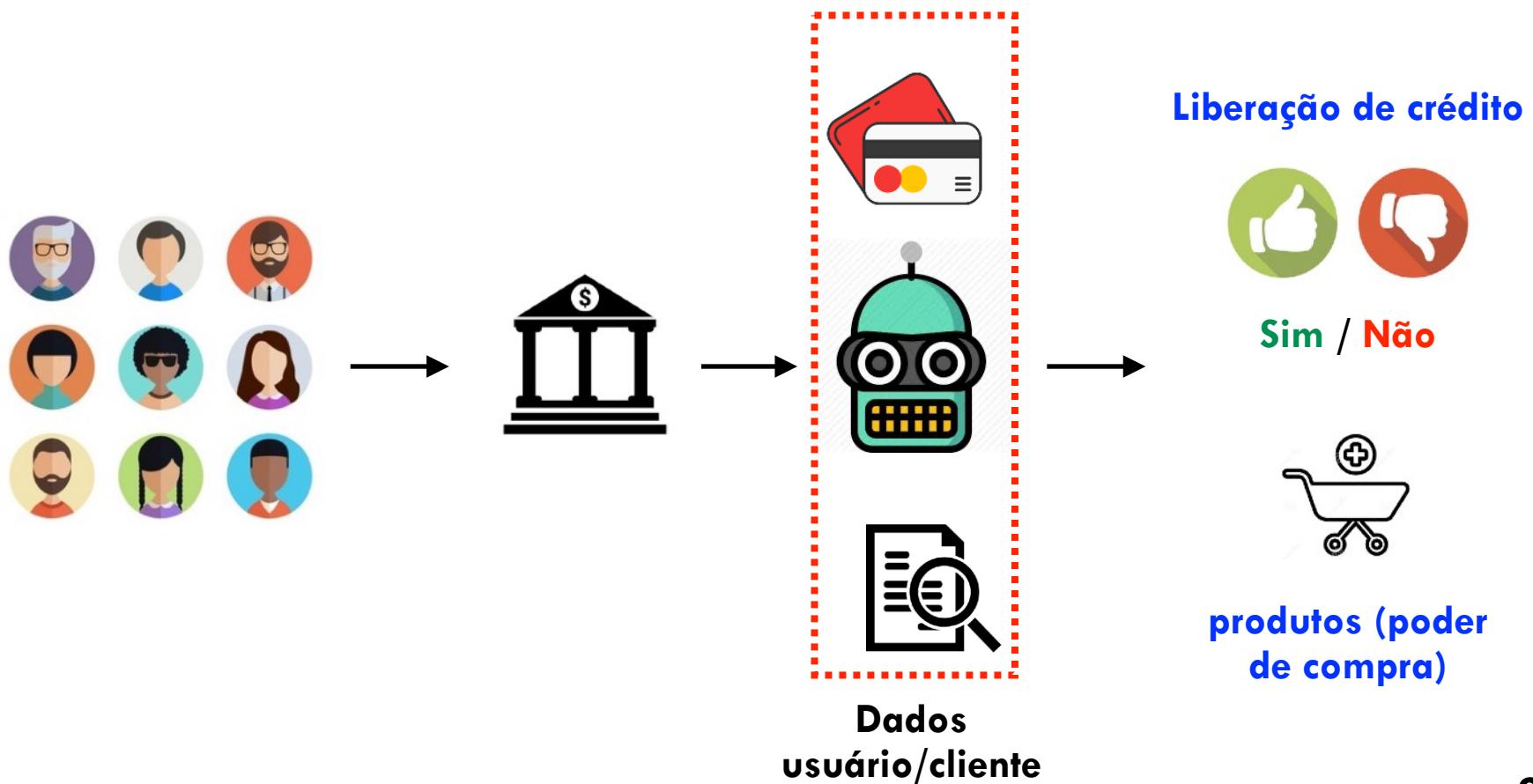
- Onde isso é usado? **Bancos/Mercado financeiro**

## Deteção de Fraudes



# Aplicações

- Onde isso é usado? **Bancos/Mercado financeiro**



# Aplicações

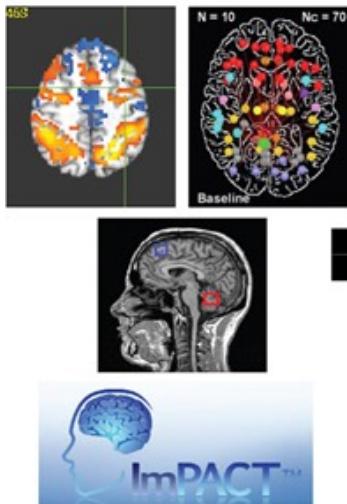
- Onde isso é usado? **Sistemas Médicos (healthcare)**

# Aplicações

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

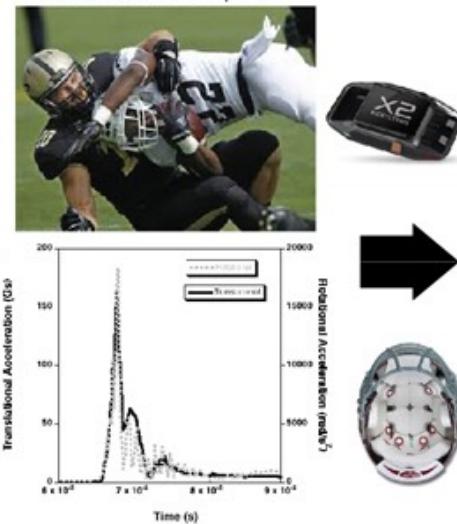
## Pre-Season Assessments

MRI, fMRI, MRS  
ImPACT



## Quantitative Head Impact Measurements

HITS, X2, Custom Systems

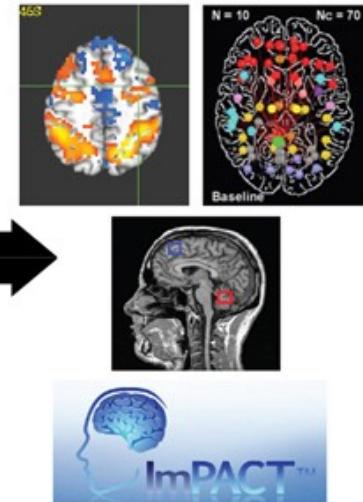


## On-Field Neurological Assessments

This section displays the "SCAT2 Sport Concussion Assessment Tool 2" form. It includes fields for Name, Date of birth, Sex, Height, Weight, and several sections of questions related to symptoms and previous concussions. A "Baseline" section is also present.

## In-Season and Post-Season Assessments

MRI, fMRI, MRS  
ImPACT



# Aplicações

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



# Aplicações

- Onde isso é usado? **Reconhecimento de Imagens**

# Aplicações

- Onde isso é usado? **Reconhecimento de Imagens**

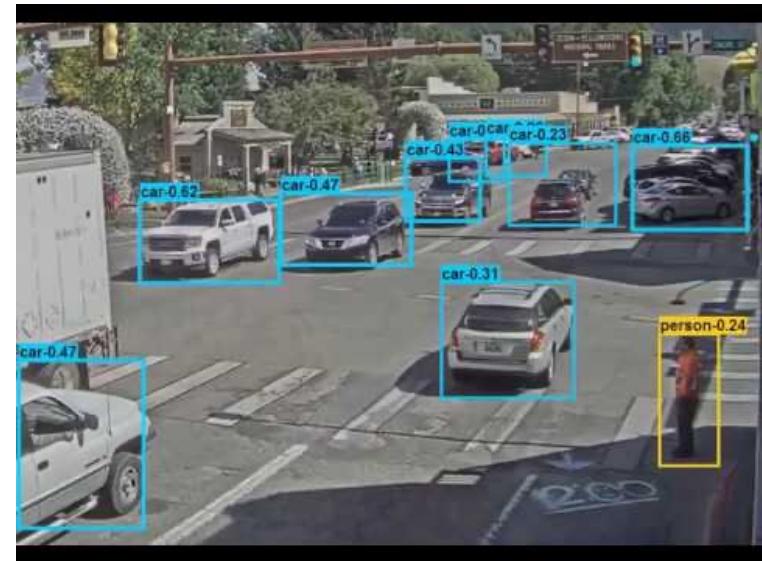
## Sistemas de Segurança



# Aplicações

- Onde isso é usado? **Reconhecimento de Imagens**

## Sistemas de Vigilância

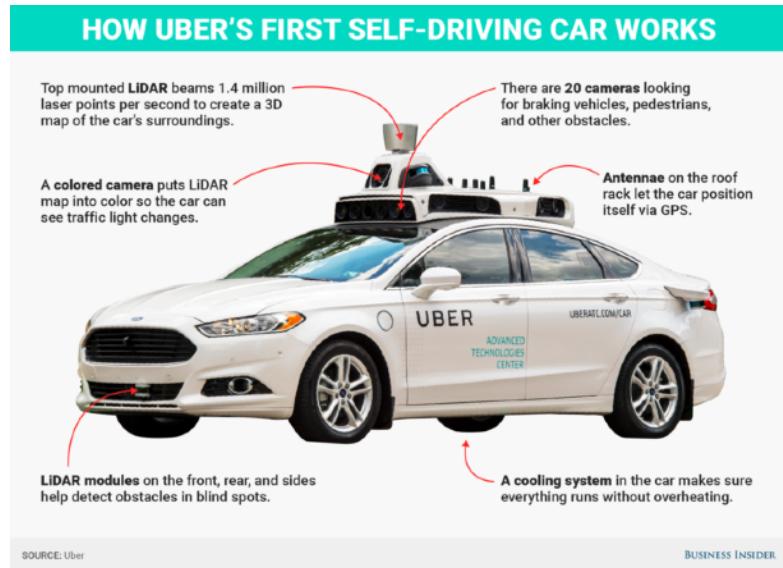


# Aplicações

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

# Aplicações

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



**Uber**



**Tesla**

# Aplicações

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

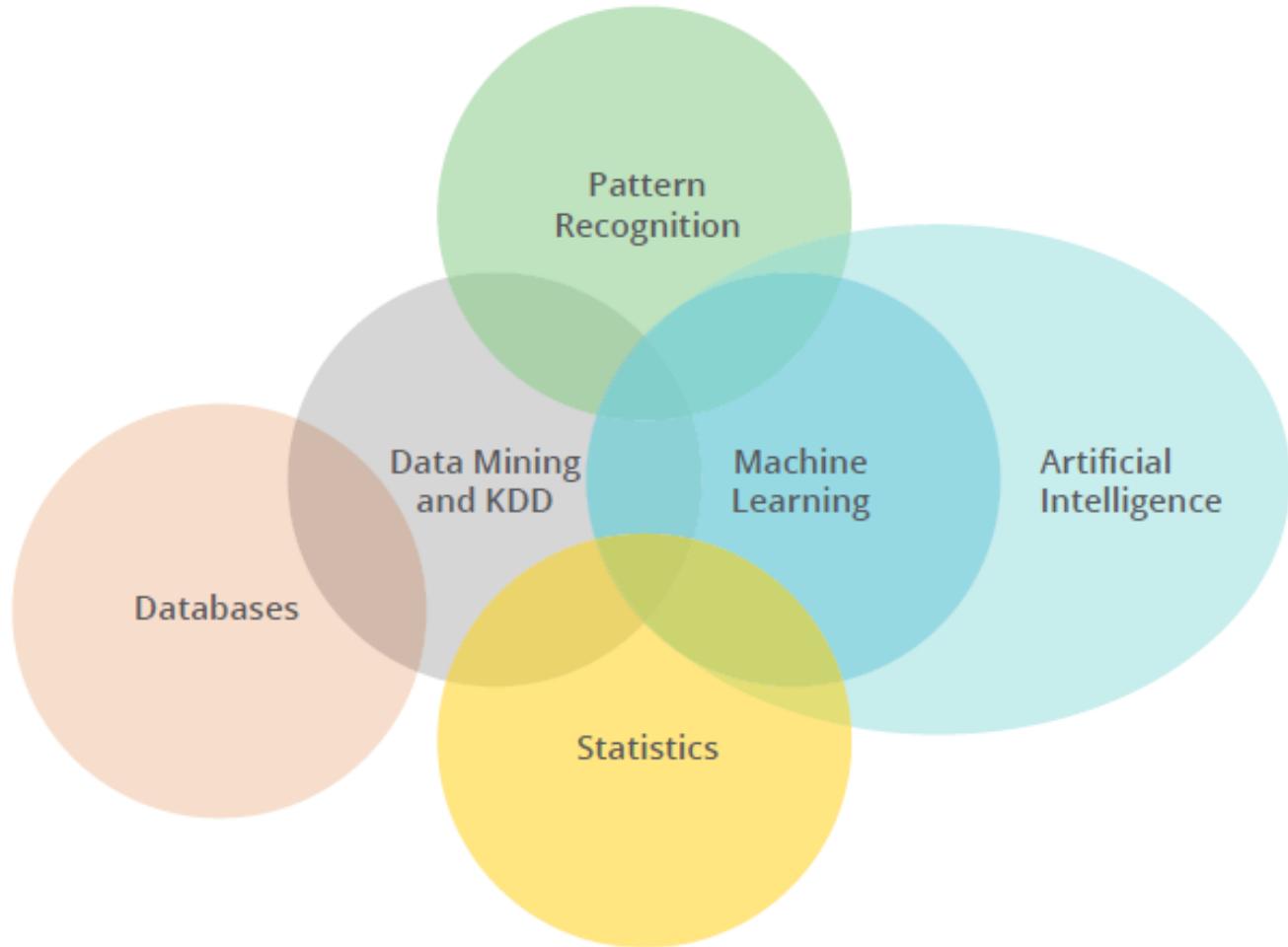


**LRM - ICMC/USP, São Carlos - SP**

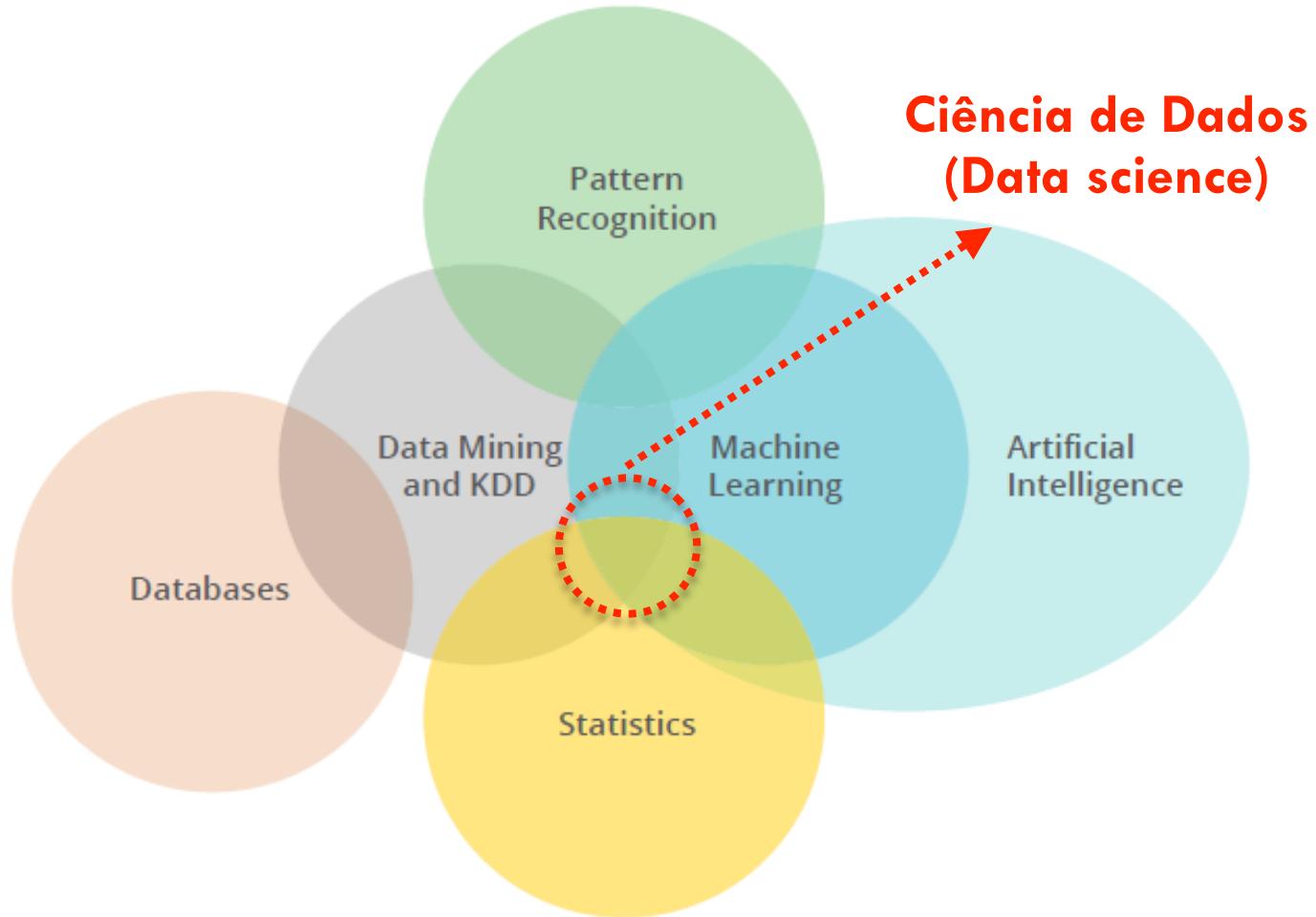
# Roteiro

- 1 Introdução**
- 2 Aplicações**
- 3 Conceitos gerais**
- 4 Fluxo de ciência de dados**
- 5 Síntese**
- 6 Referências**

# Conceitos Gerais



# Conceitos Gerais

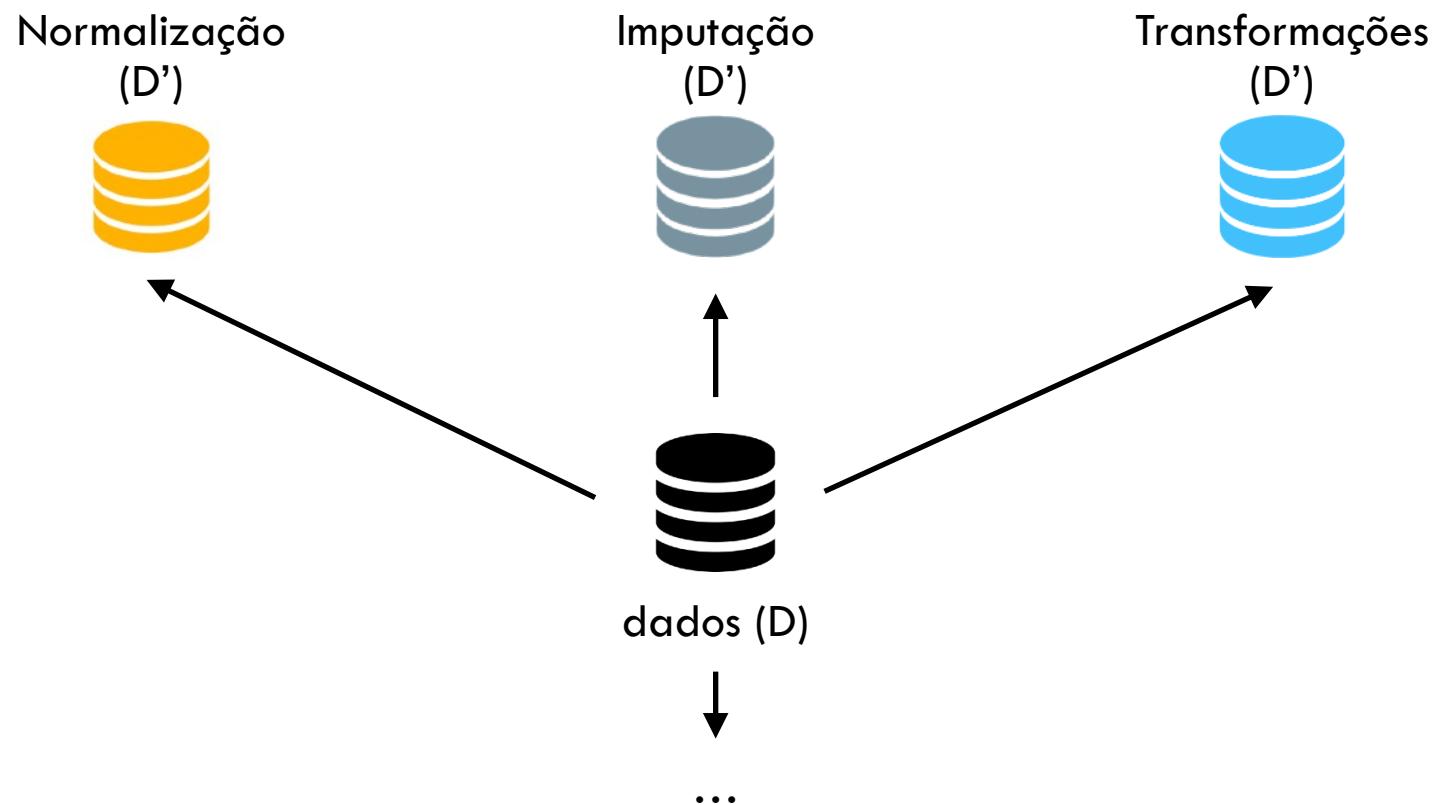


# Mineração de dados (Data Mining)

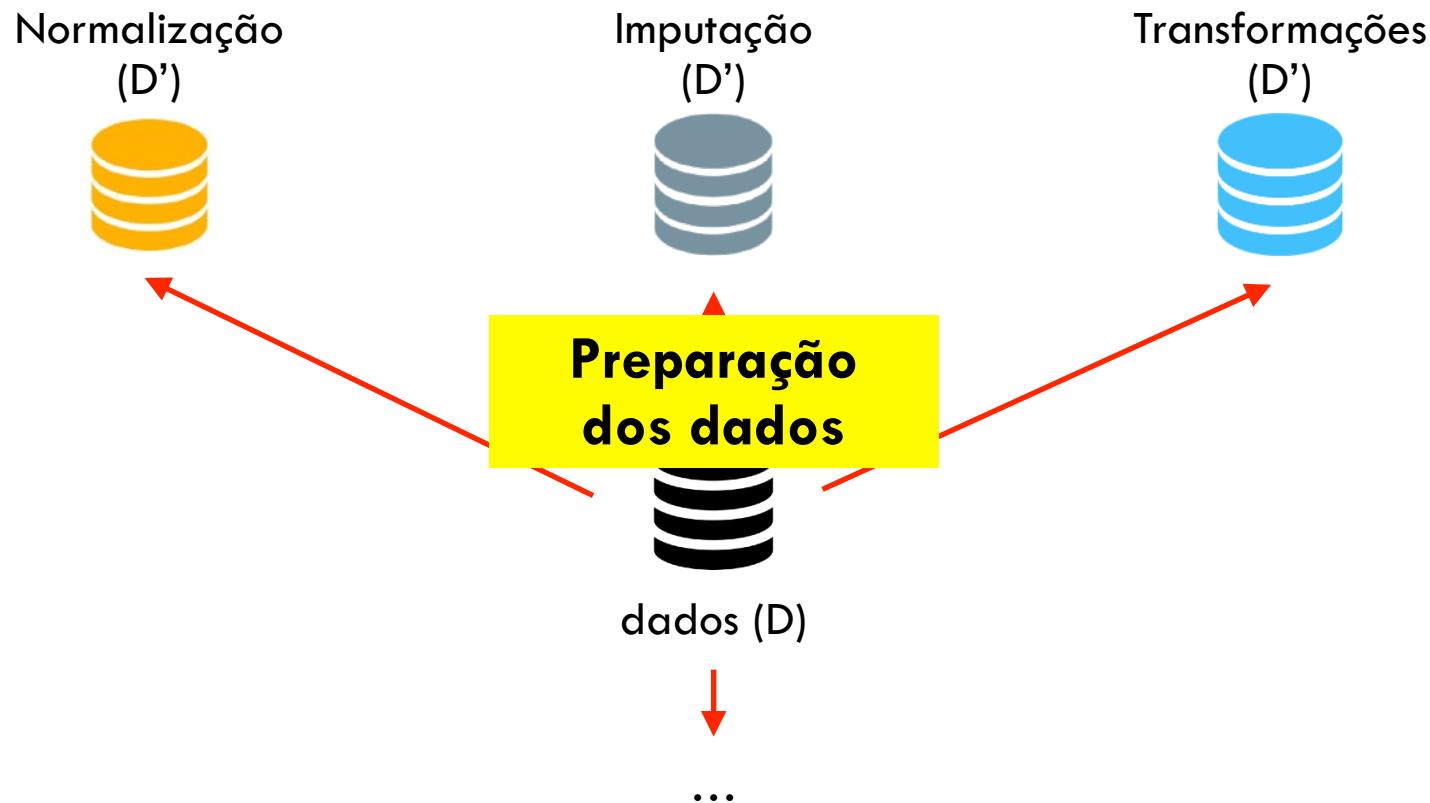


dados (D)

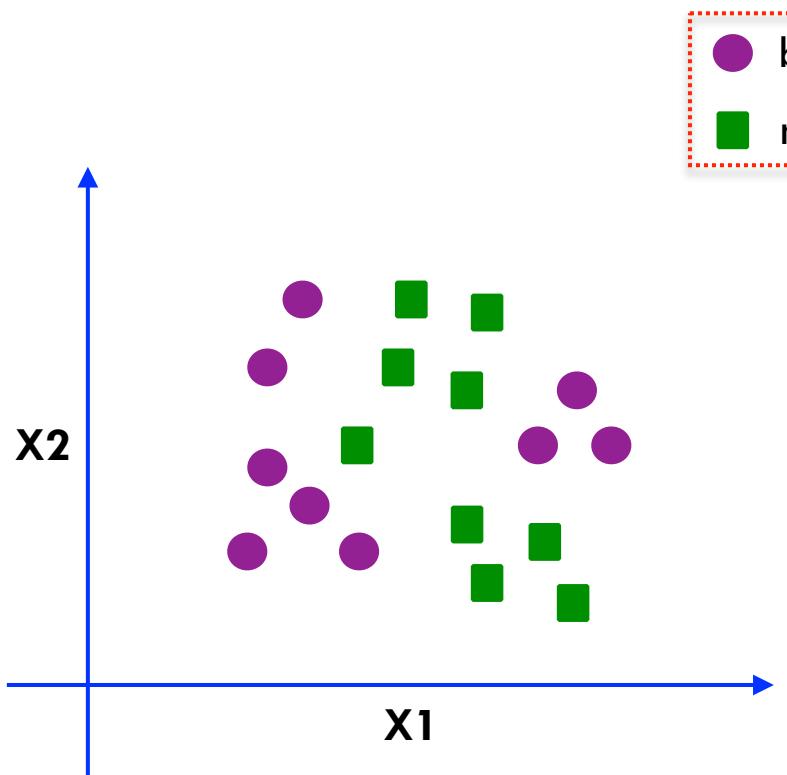
# Mineração de dados (Data Mining)



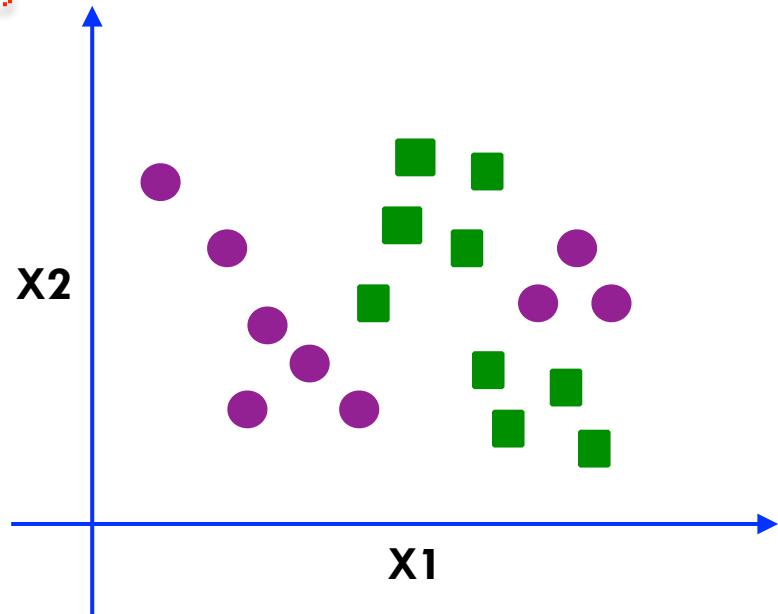
# Mineração de dados (Data Mining)



# Aprendizado de Máquina (Machine Learning)

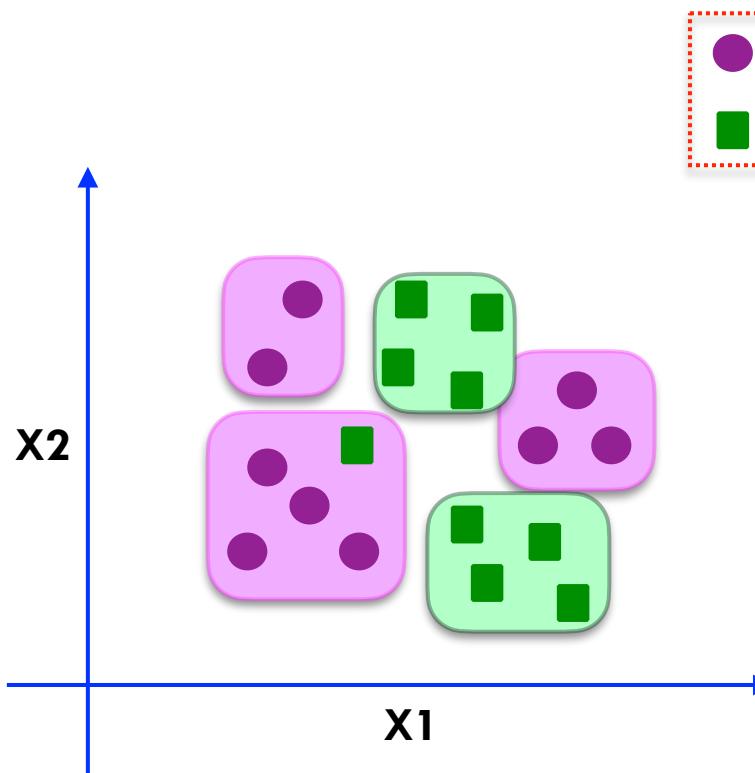


Tarefa descritiva  
(Clustering)

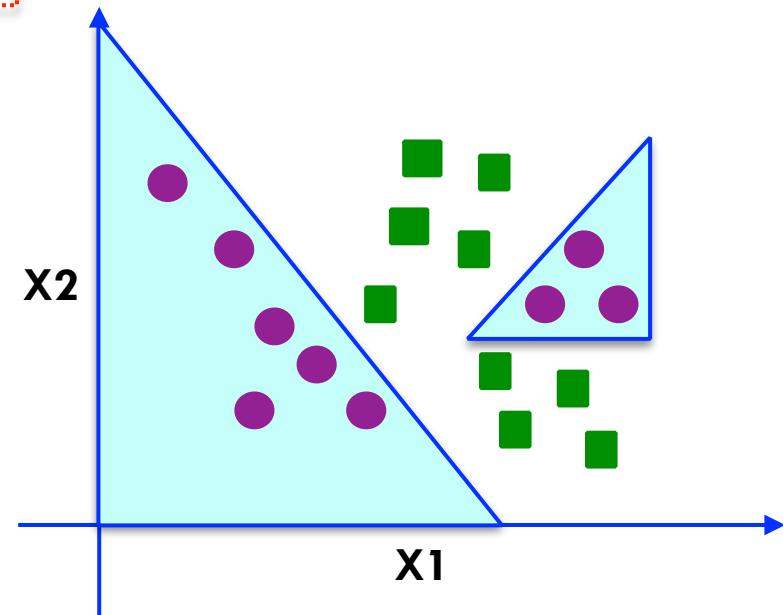


Tarefa preditiva  
(Classificação/  
Regressão)

# Aprendizado de Máquina (Machine Learning)

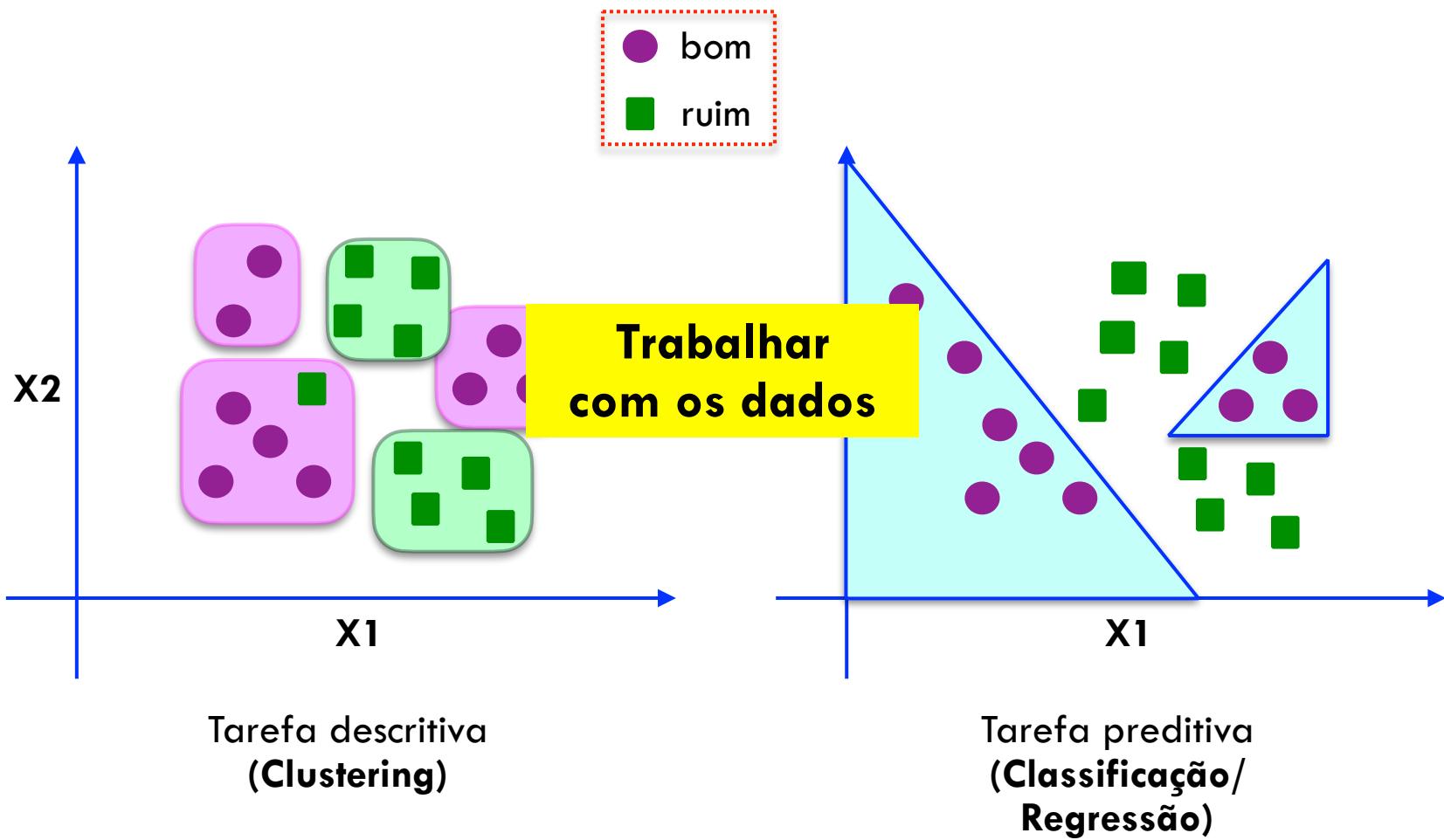


Tarefa descritiva  
(Clustering)



Tarefa preditiva  
(Classificação/  
Regressão)

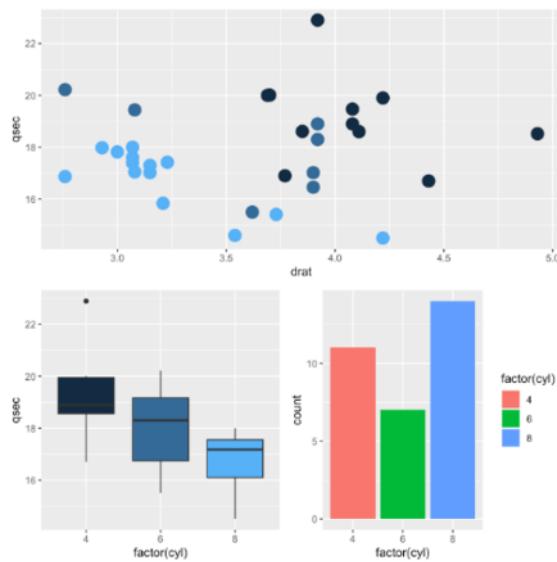
# Aprendizado de Máquina (Machine Learning)



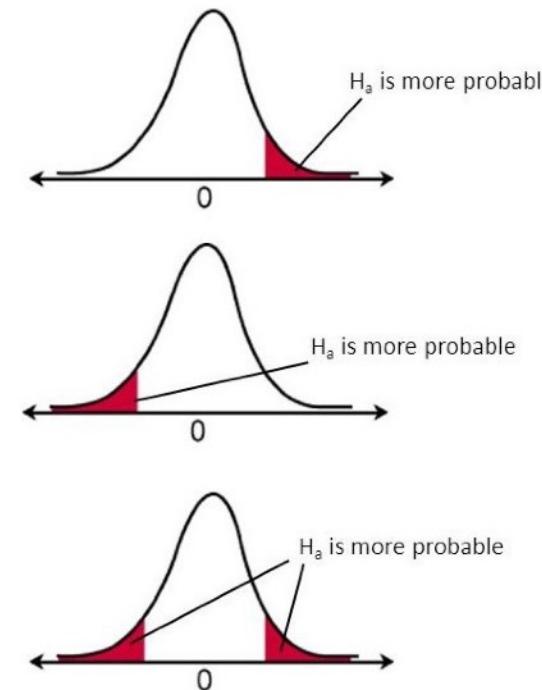
# Estatística (Statistics)



Técnicas de  
Amostragem



Estatística  
Descritiva



Testes de  
Hipótese

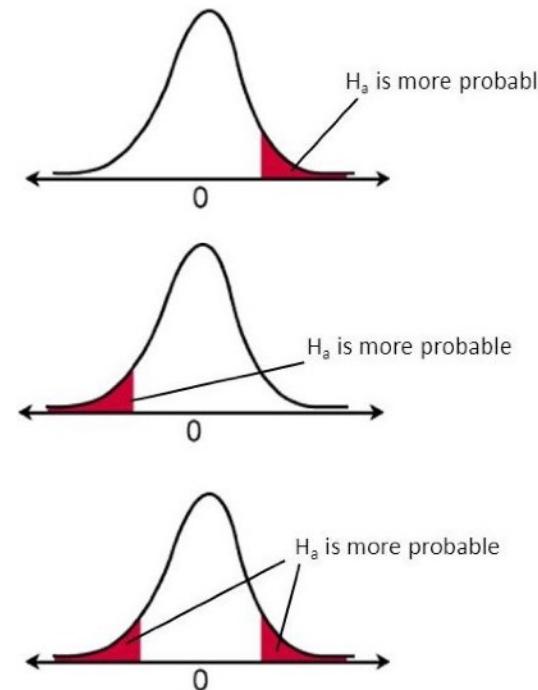
# Estatística (Statistics)



Técnicas de  
Amostragem

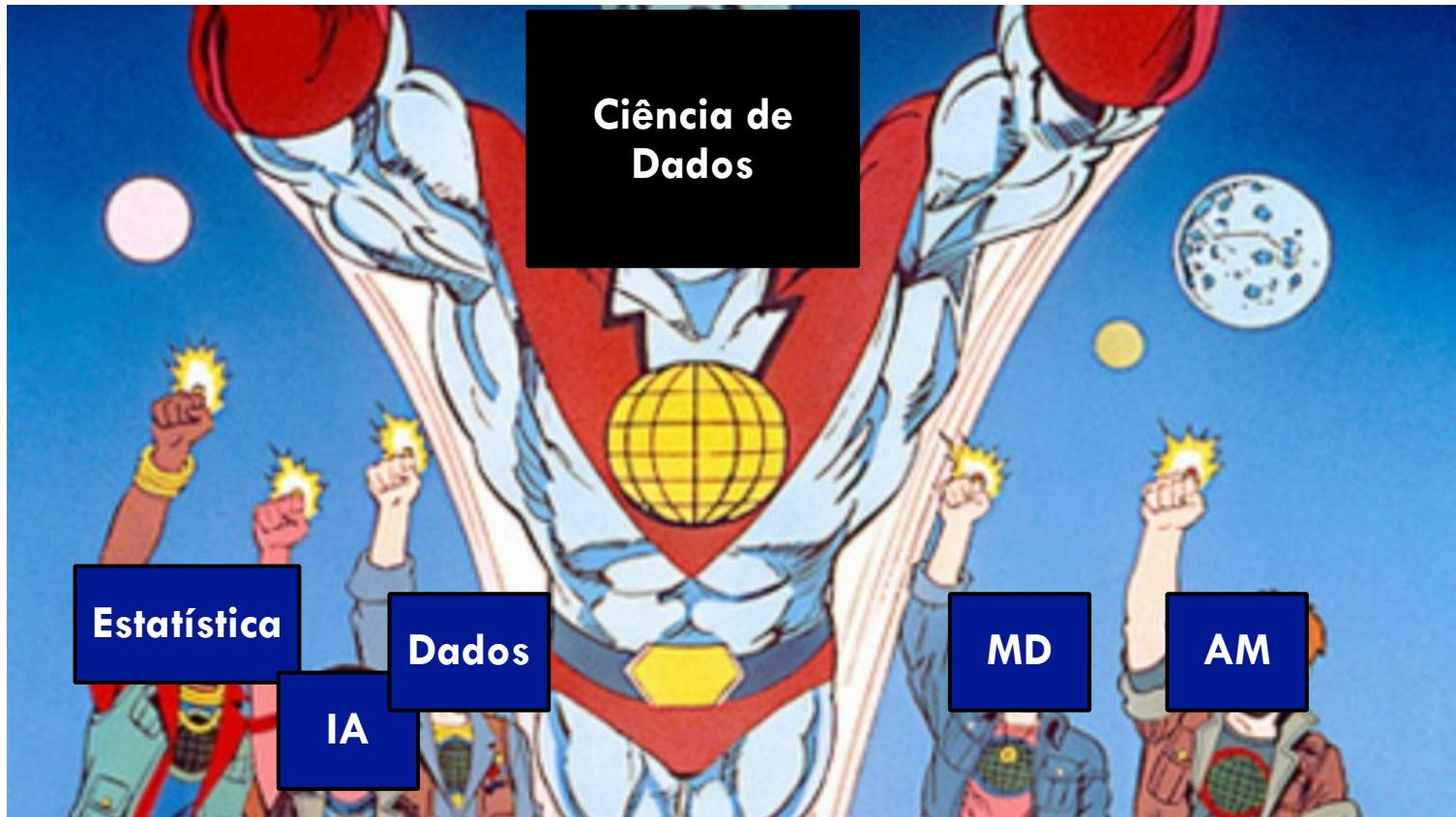


Estatística  
Descritiva



Testes de  
Hipótese

# Ciência de Dados (Data science)



# Ciência de Dados (Data science)

- Quantos algoritmos existem?

# Ciência de Dados

- Quantos algoritmos existem?



rapidminer



TensorFlow



Keras

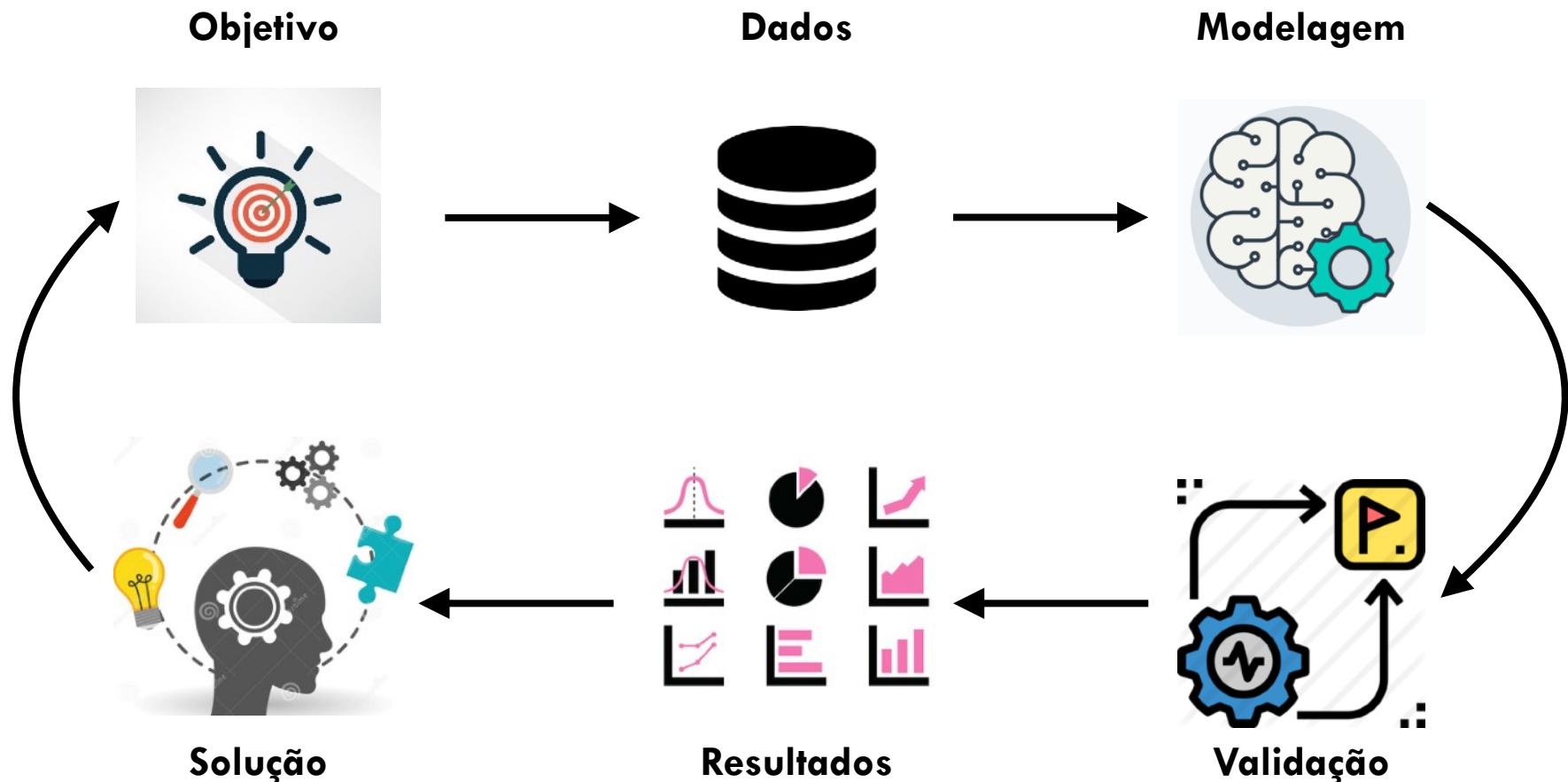


...

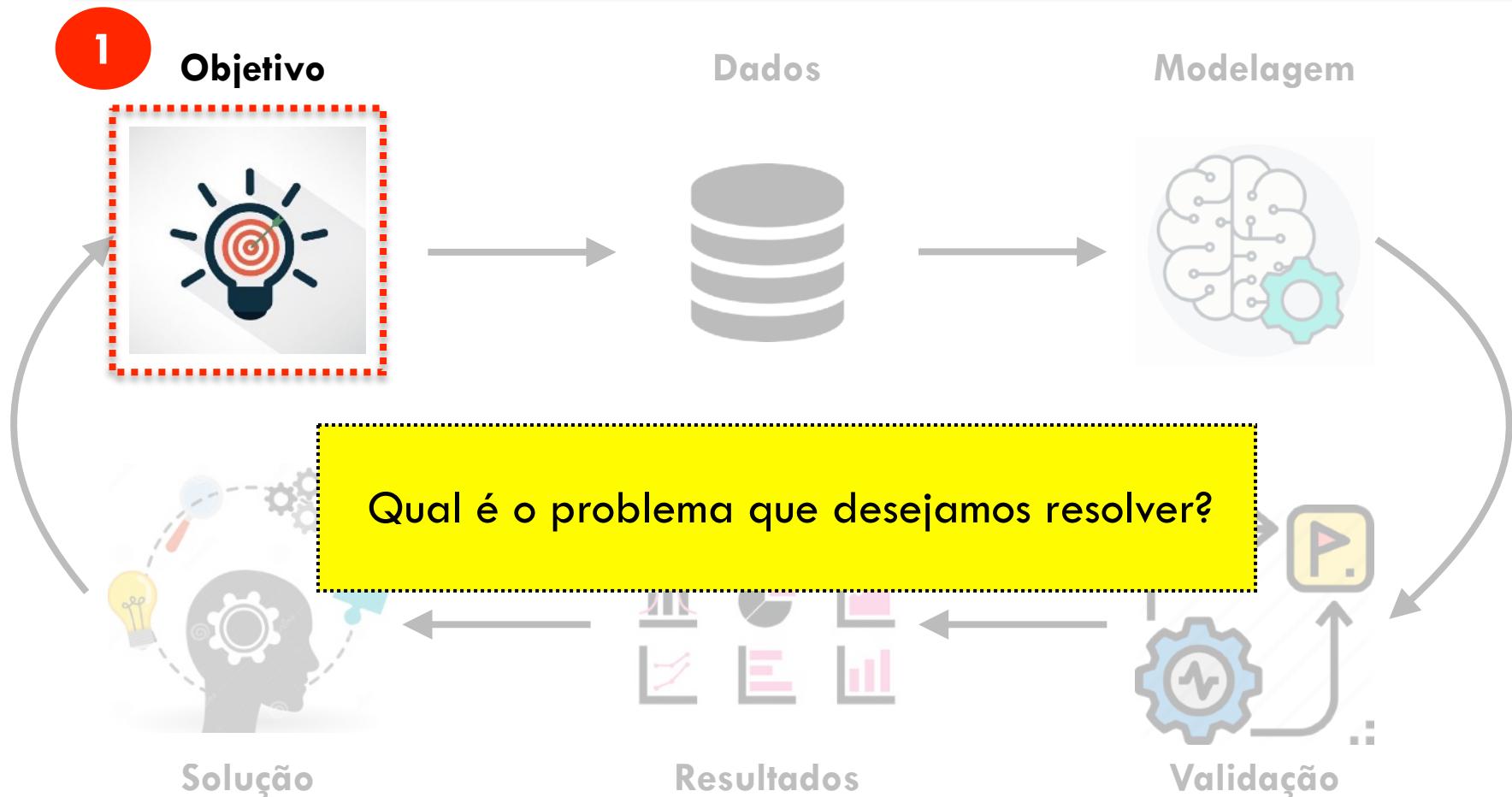
# Roteiro

- 1 Introdução**
- 2 Aplicações**
- 3 Conceitos gerais**
- 4 Fluxo de ciência de dados**
- 5 Síntese**
- 6 Referências**

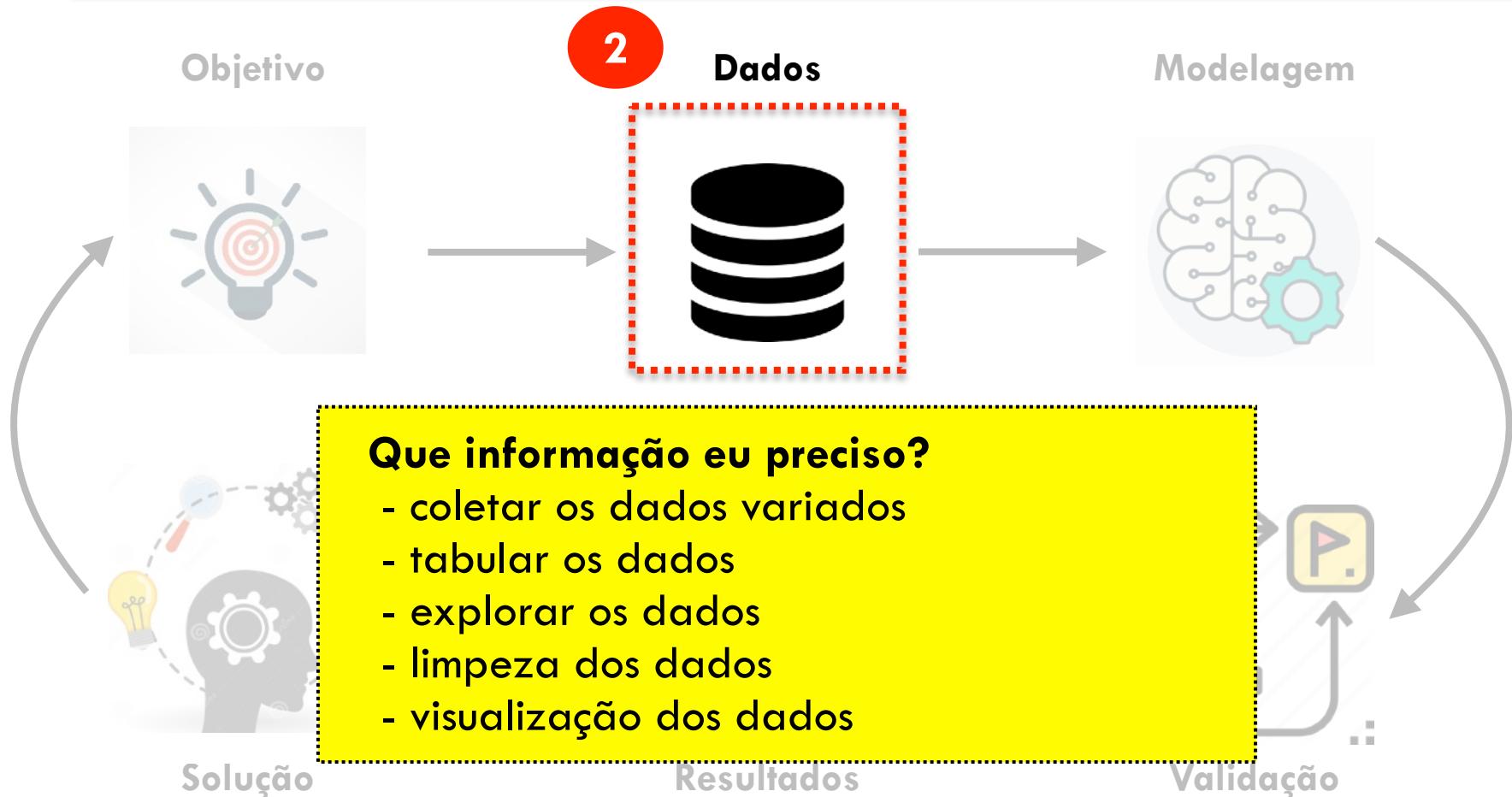
# Fluxo de Ciência de Dados



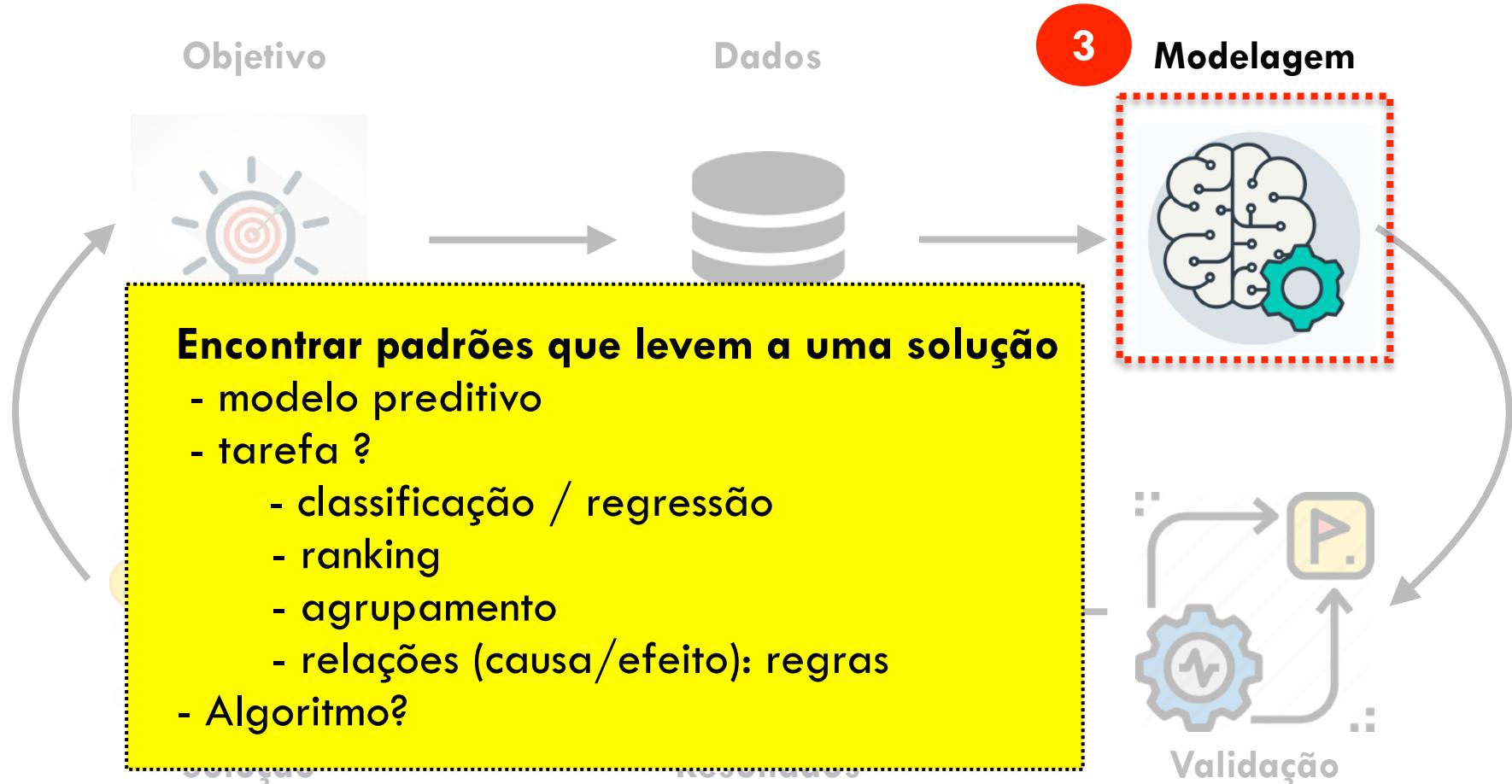
# Fluxo de Ciência de Dados



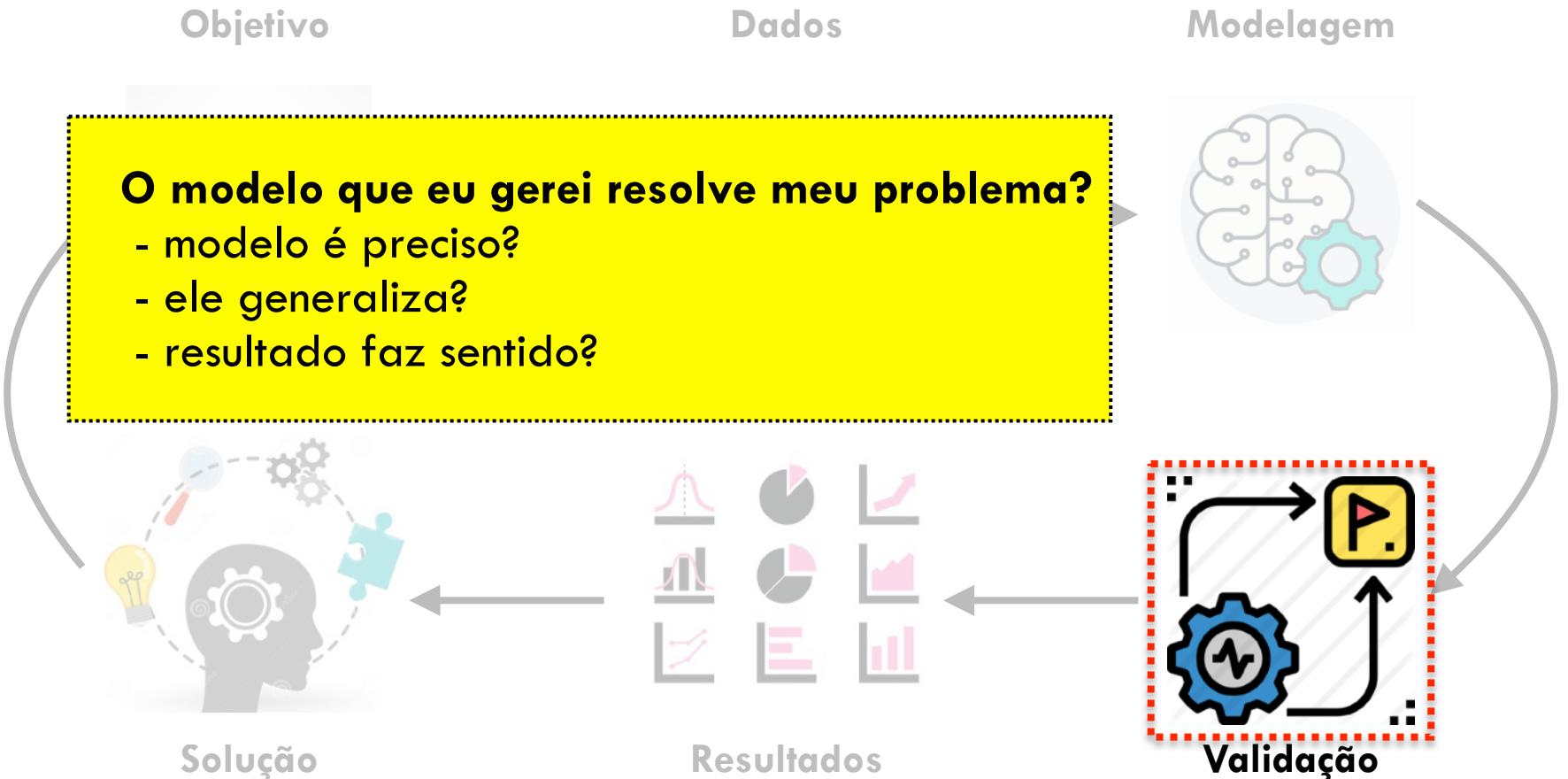
# Fluxo de Ciência de Dados



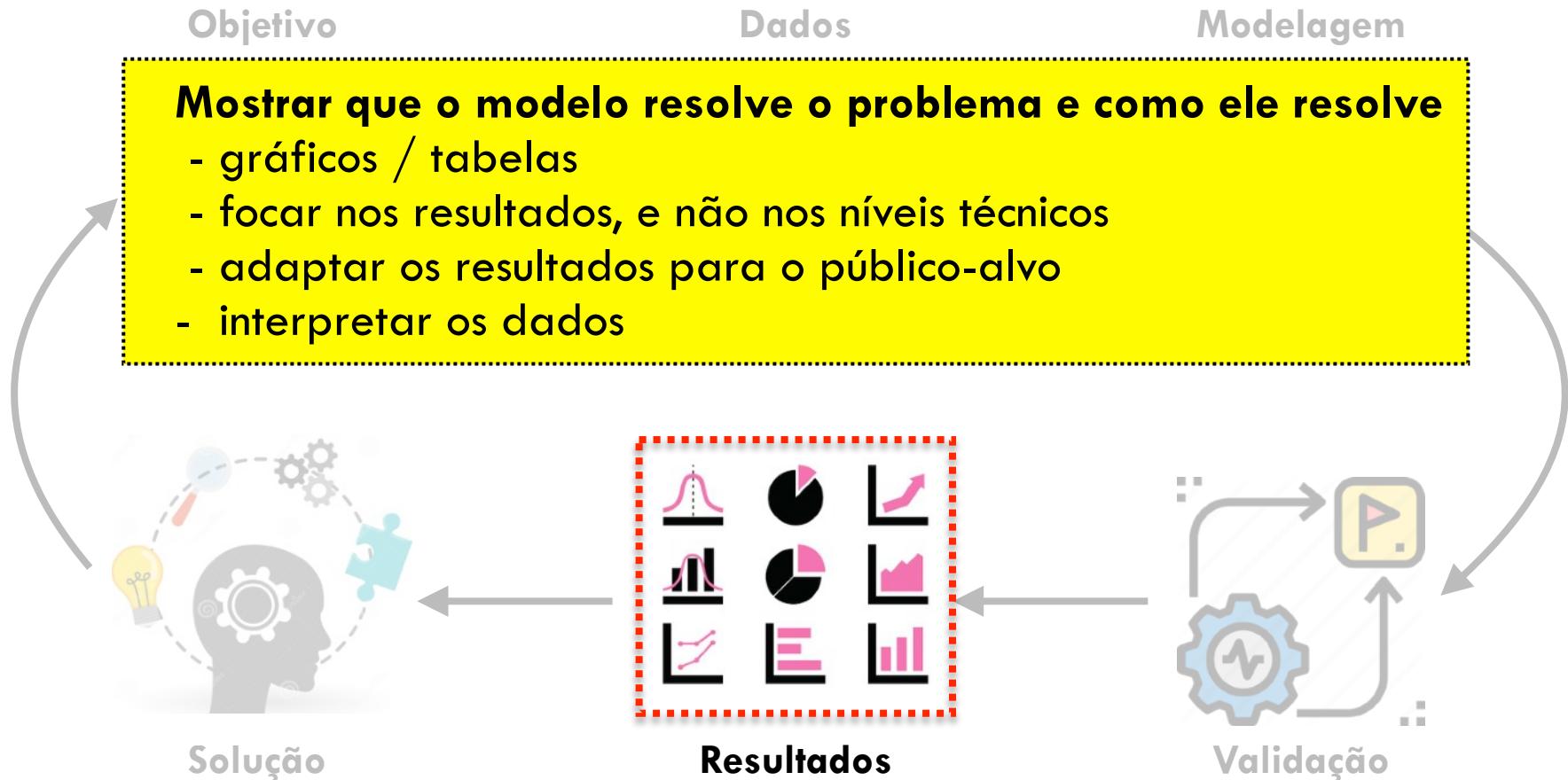
# Fluxo de Ciência de Dados



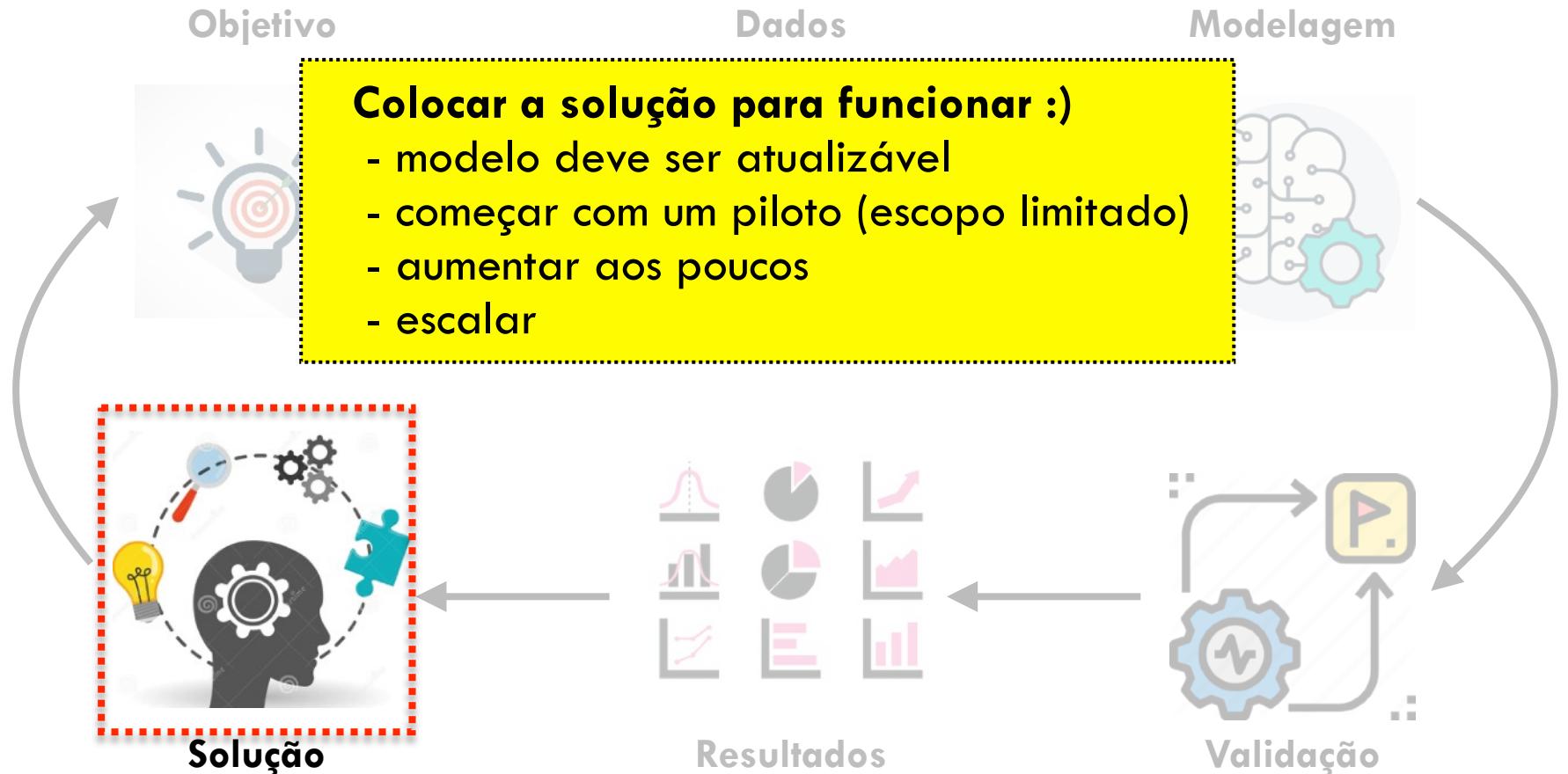
# Fluxo de Ciência de Dados



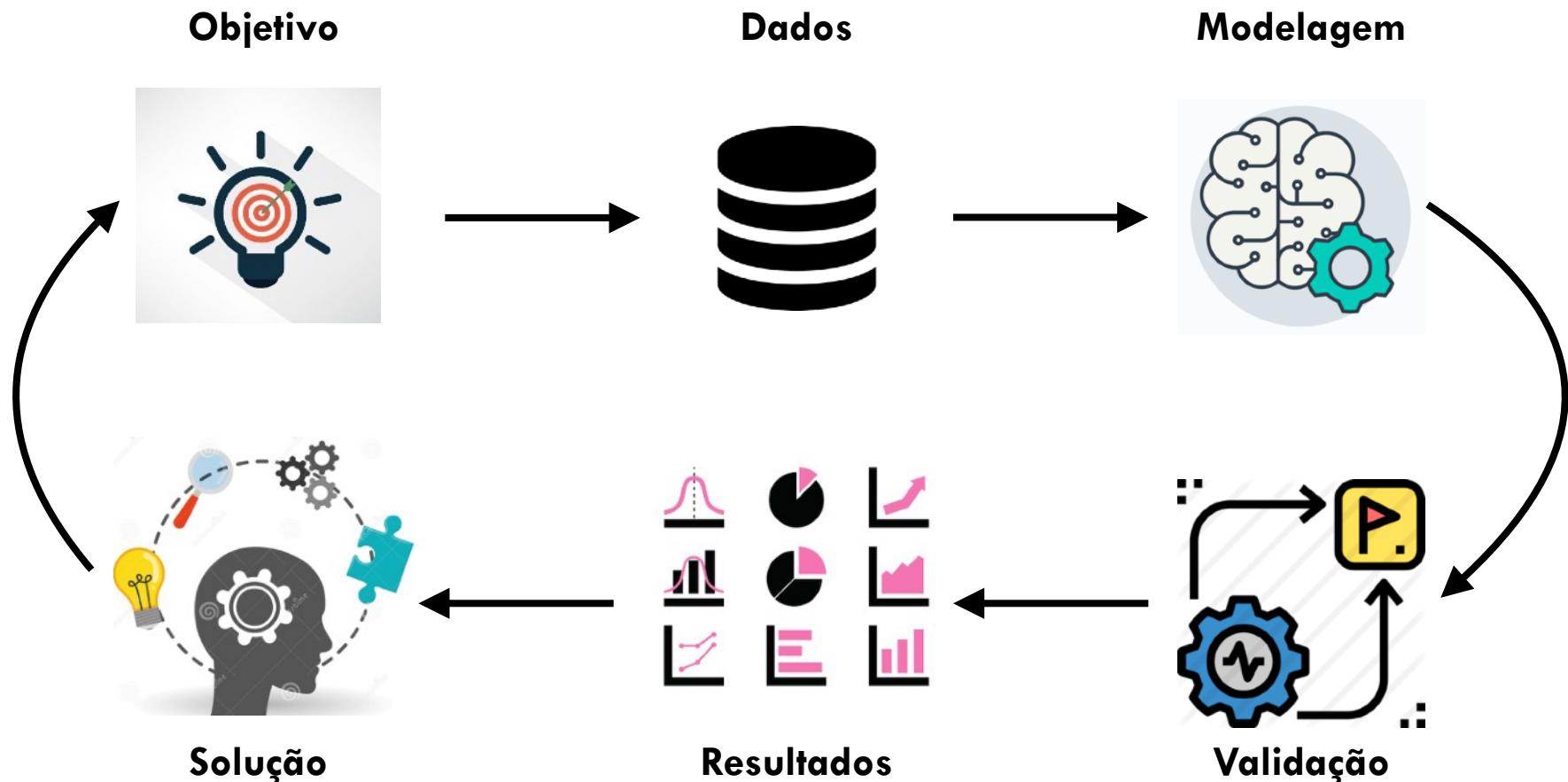
# Fluxo de Ciência de Dados



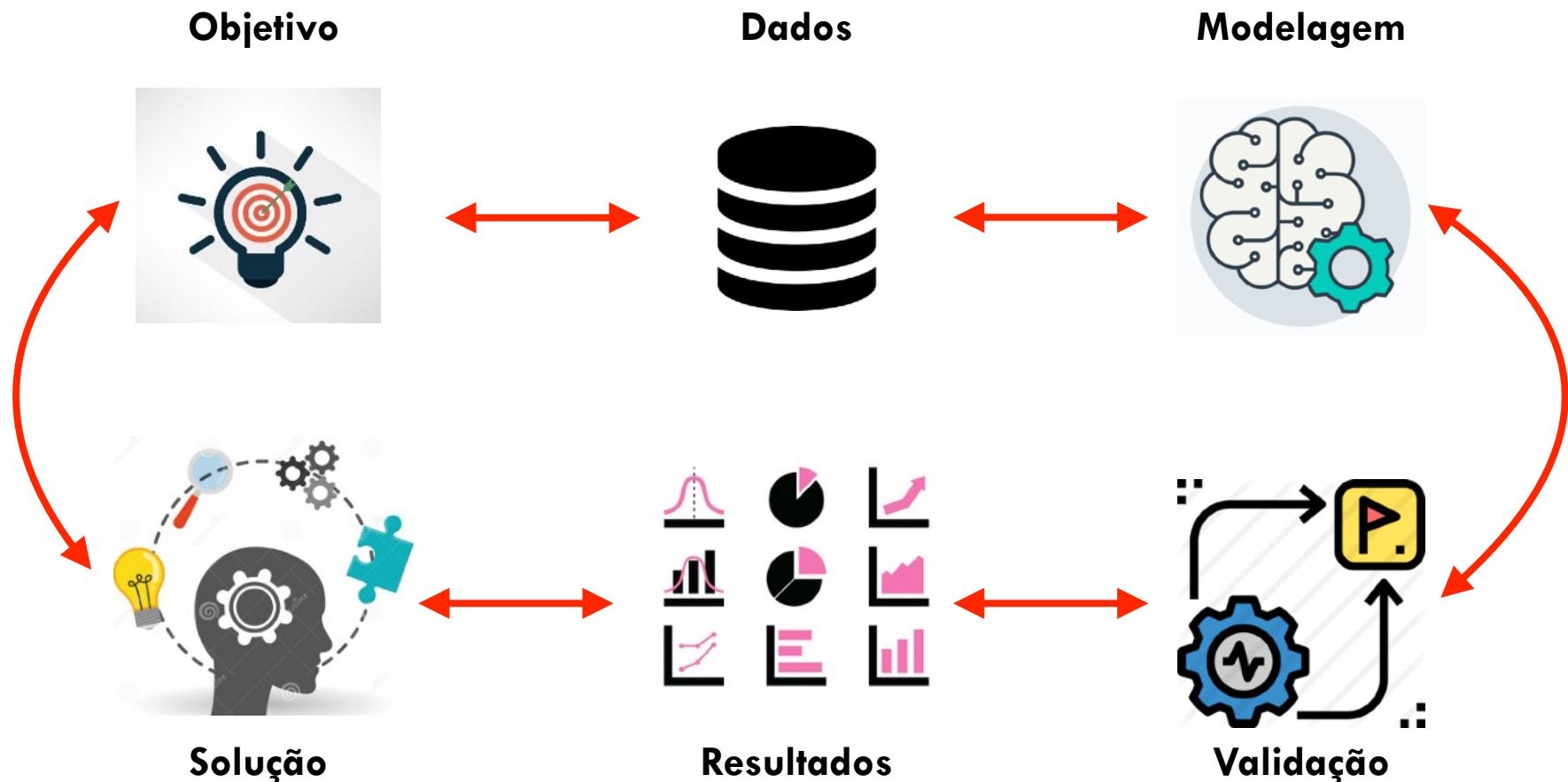
# Fluxo de Ciência de Dados



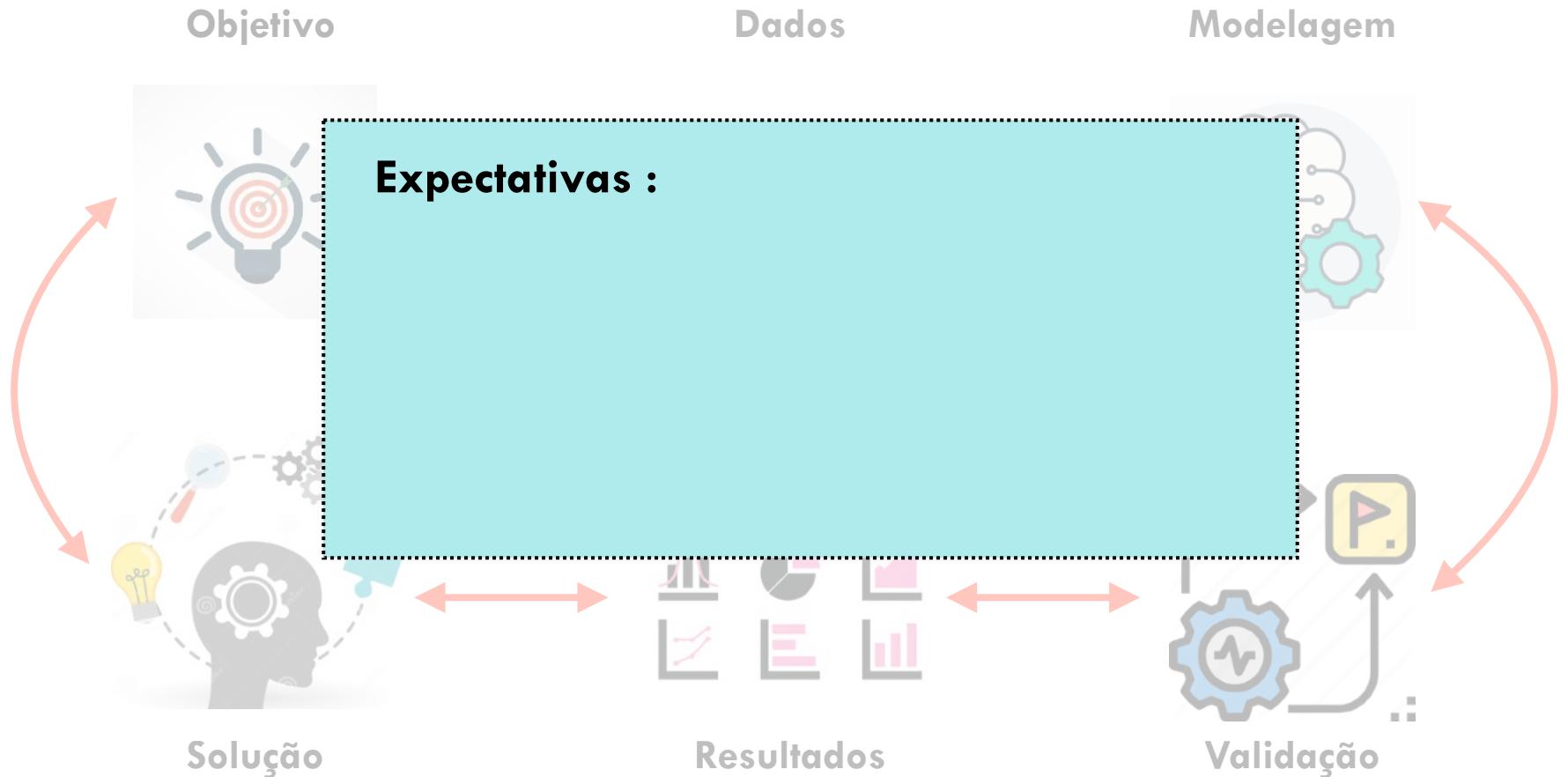
# Fluxo de Ciência de Dados



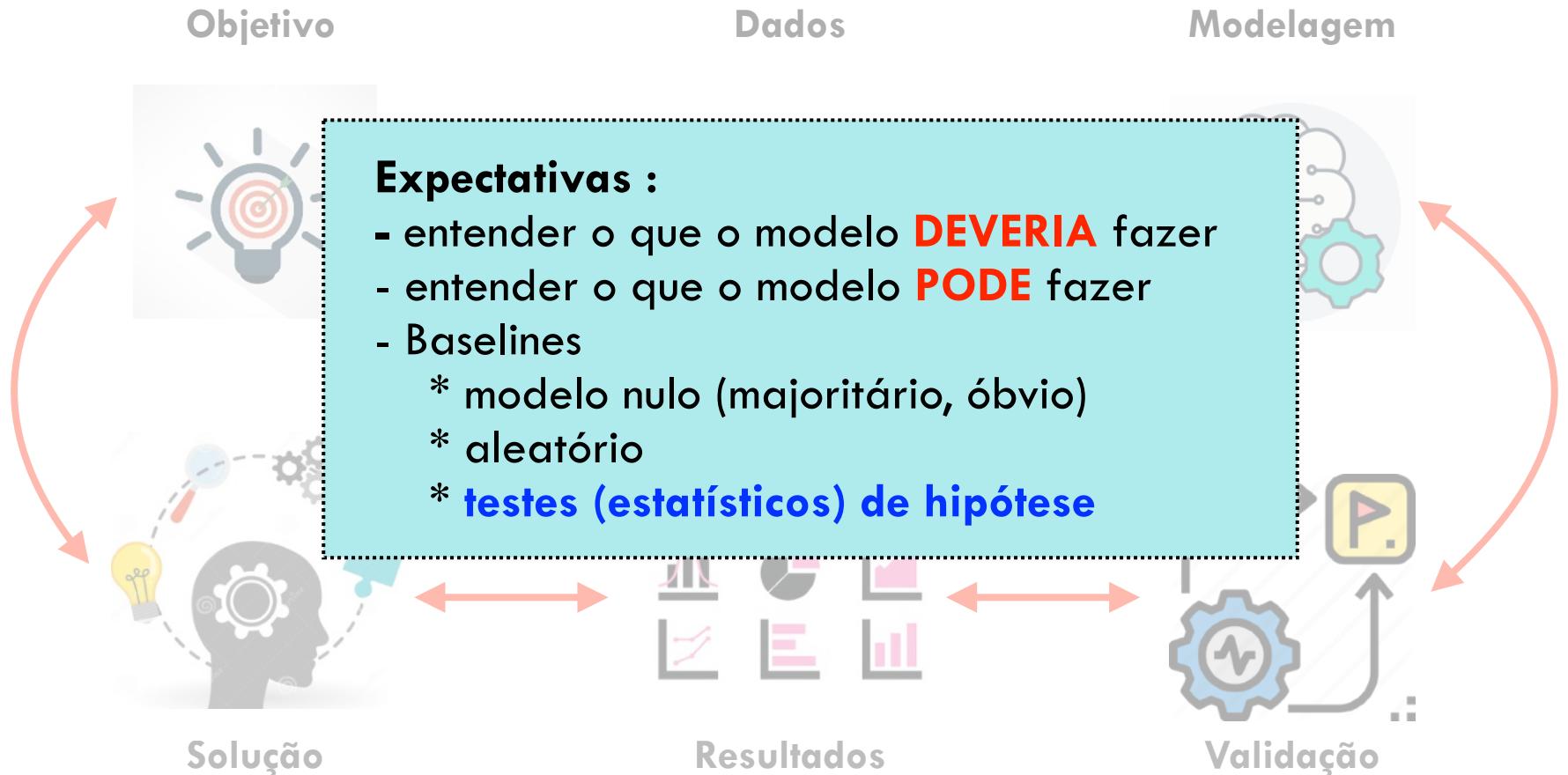
# Fluxo de Ciência de Dados



# Fluxo de Ciência de Dados



# Fluxo de Ciência de Dados



# Roteiro

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

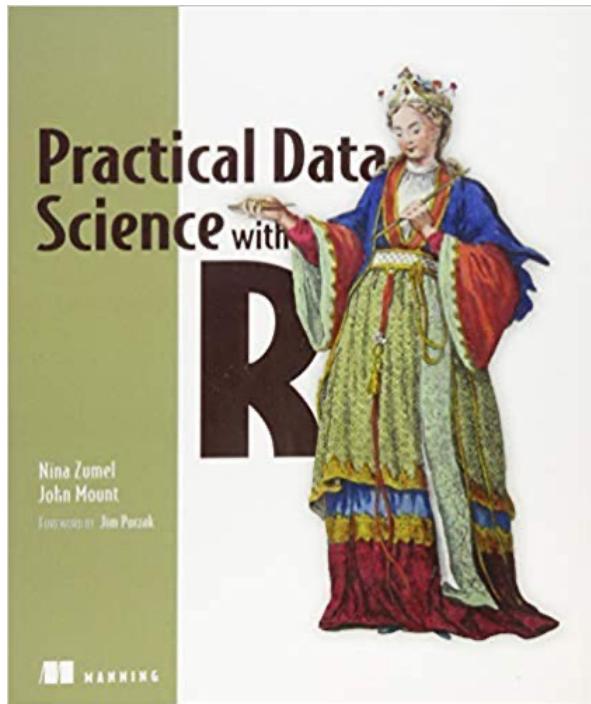
# Síntese

- Ciência de Dados (CD) é um tema atual (*hot topic*)
- Cada vez mais usada em soluções ao nosso redor
  - quantidade de dados gerados/disponíveis
  - automação/compreensão de processos
- CD é uma área interdisciplinar
  - Aprendizado de Máquina, Mineração de Dados, Estatística
- Várias ferramentas *open source* para CD
  - R / Python, mas não apenas esses ambientes

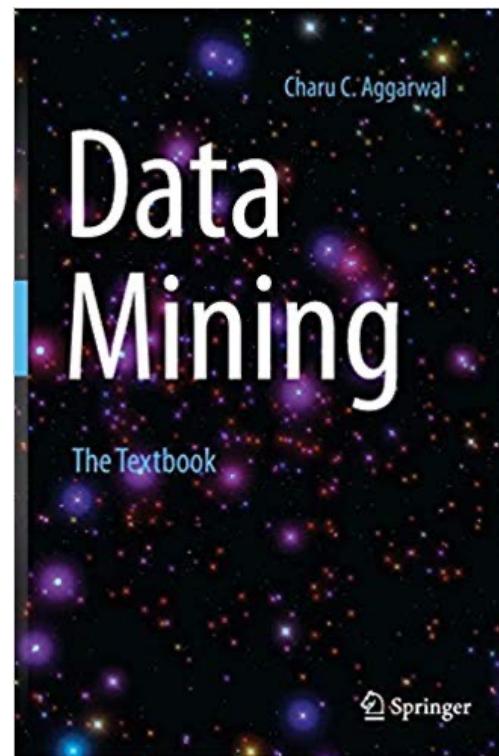
# Roteiro

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

# Referências

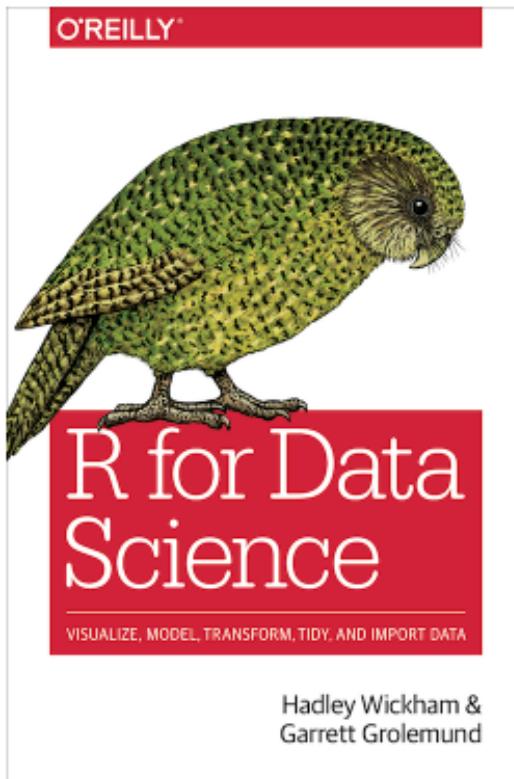


[Zumel & Mount, 2014]

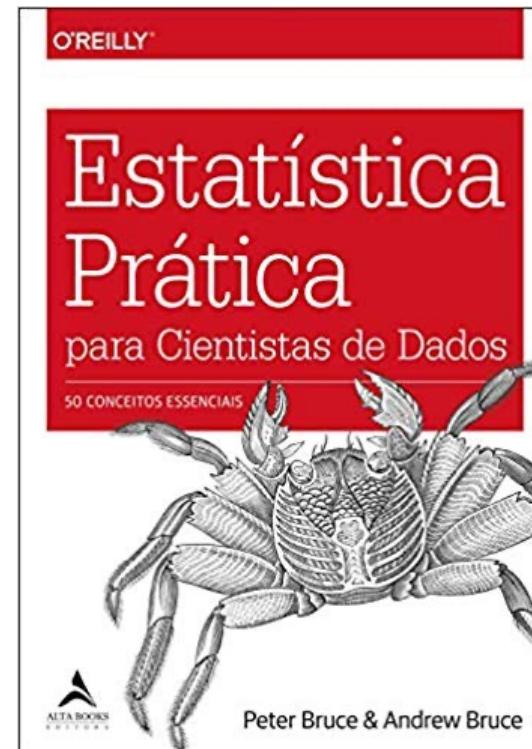


[Aggarwal, 2015]

# Referências



[Wickham & Grolemund, 2018]



[Bruce & Bruce, 2019]

# Referências

- [Zumel & Mount, 2014] ZUMEL, N.; MOUNT, J. **Practical Data Science with R**. 2nd edition. Manning, 2014.
- [Aggarwal, 2015] AGGARWAL, C. C. **Data Mining**. Springer, 2015.
- [Wickham & Grolemund, 2018] WICKHAM, H; GROLEMUND, G. **R for Data Science**. O'Reilly, 2018.
- [Bruce & Bruce, 2019] BRUCE, A.; BRUCE, P. **Estatística Prática para Cientistas de Dados**. O'Reilly, 2019.



# Obrigado :)

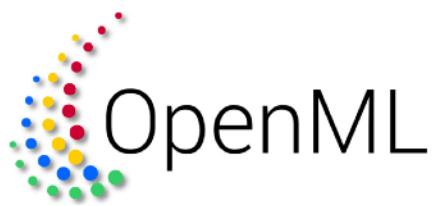
Rafael G. Mantovani

[rafaelmantovani@utfpr.edu.br](mailto:rafaelmantovani@utfpr.edu.br)

# Links Interessantes :)

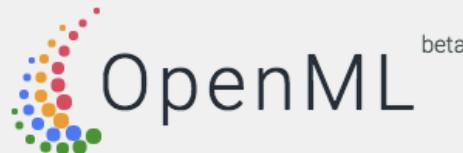
- R for Data Science: <https://r4ds.had.co.nz>
- Tidyverse: <https://www.tidyverse.org>
- mlr: <https://mlr.mlr-org.com>
- mlr3: <https://mlr3.mlr-org.com>
- Skicit learn: <https://www.tidyverse.org>
- matplotlib: <https://matplotlib.org>
- OpenML: <https://www.openml.org>
- UCI: <https://archive.ics.uci.edu/ml/index.php>
- RStudio: <https://rstudio.com>
- Spyder: <https://www.spyder-ide.org>

# Ferramentas



# OpenML / Dados

Search



Machine learning, better, together

20072  
data sets

Find or add **data** to analyse

67888  
tasks

Download or create scientific  
**tasks**

6092  
flows

Find or add data analysis **flows**

9012316  
runs

Upload and explore all **results**  
online.

**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

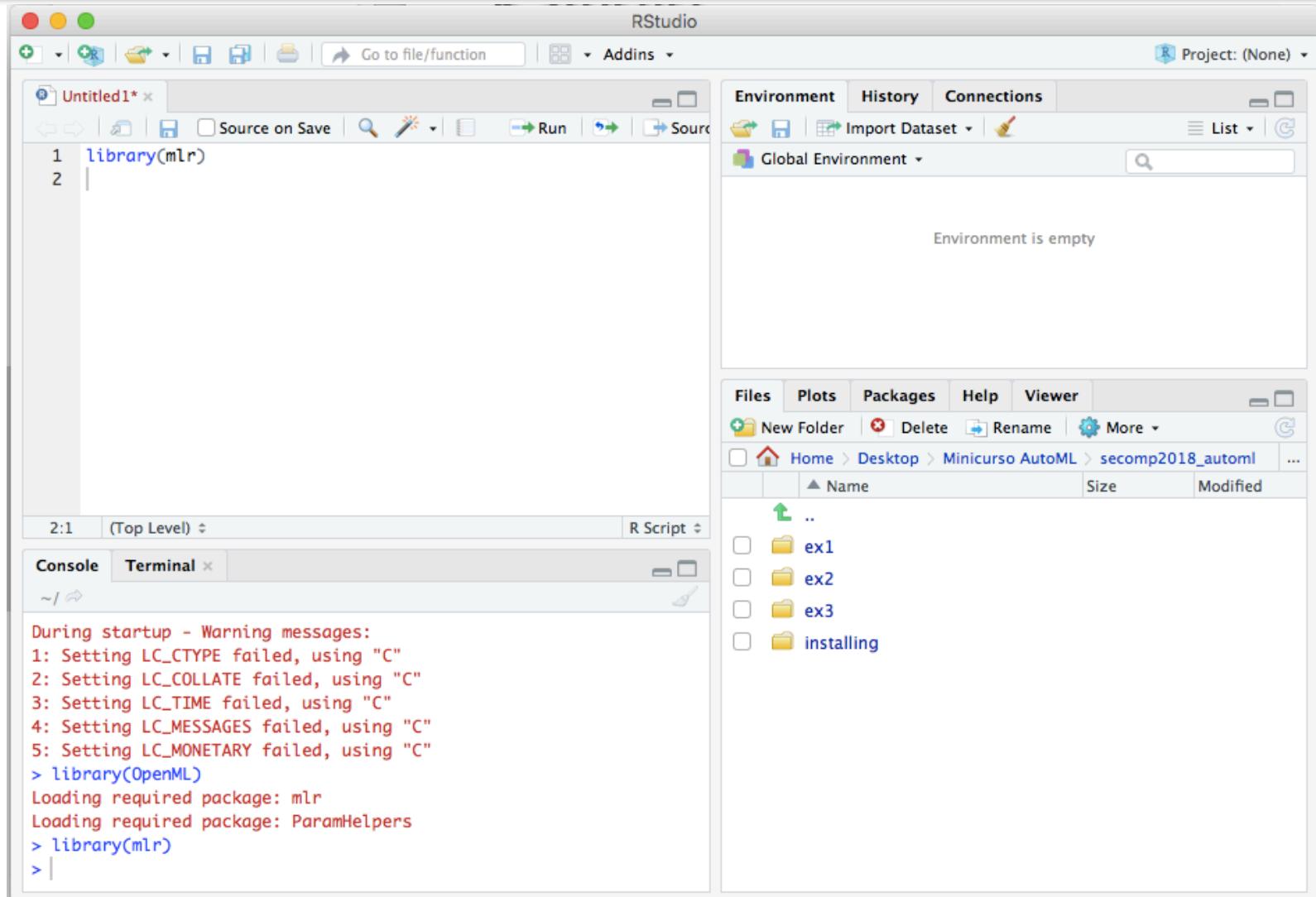
This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris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

[▼ Show all](#)

## 5 features



# RStudio / IDE para R



# Spyder / IDE para Python

The screenshot shows the Spyder IDE interface with the following components:

- Project explorer:** Shows the file structure of the current project, including files like `temp.py`, `interpolation.py`, and `__init__.py`.
- Code editor:** Displays the `temp.py` script with Python code for generating data, performing calculations, and plotting.
- Variable explorer:** Lists variables and their values, such as `bars` (a `BarContainer` object), `df` (a `DataFrame` object), and `rgb` (an `RGBArray` object).
- Python console:** Shows the execution of code to calculate a shaded surface and display it as a 3D plot.
- 3D plot:** A 3D surface plot generated from the data in the code editor.

```
6
7 import pylab
8 from numpy import cos, linspace, pi, sin, random
9 from scipy.interpolate import splprep, splev
10
11 # XX Generate data for analysis
12
13 # Make ascending spiral in 3 space
14 t = linspace(0, 1.75 * 2 * pi, 100)
15
16 x = sin(t)
17 y = cos(t)
18 z = t
19
20 # Add noise
21 x += random.normal(scale=0.1, size=x.shape)
22 y += random.normal(scale=0.1, size=y.shape)
23 z += random.normal(scale=0.1, size=z.shape)
24
25
26 # XX Perform calculations
27
28 # Spline parameters
29 smoothness = 3.0 # Smoothness parameter
30 k_param = 2 # Spline order
31 nests = -1 # Estimate of number of knots needed (-1 = maximal)
32
33 # Find the knot points
34 knot_points, u = splprep([x, y, z], s=smoothness, k=k_param, nests=-1)
35
36 # Evaluate spline, including interpolated points
37 xnew, ynew, znew = splev(linspace(0, 1, 400), knot_points)
38
39
40 # XX Plot results
41
42 # TODO: Rewrite to avoid code smell
43 pylab.subplot(2, 2, 1)
44 data = pylab.plot(x, y, 'bo-', label='Data with X-Y Cross Section')
45 fit, = pylab.plot(xnew, ynew, 'r-', label='Fit with X-Y Cross Section')
46 pylab.legend()
47 pylab.xlabel('x')
48 pylab.ylabel('y')
```

# mlr3 / framework em R

mlr3



Package website: [release](#) | [dev](#)

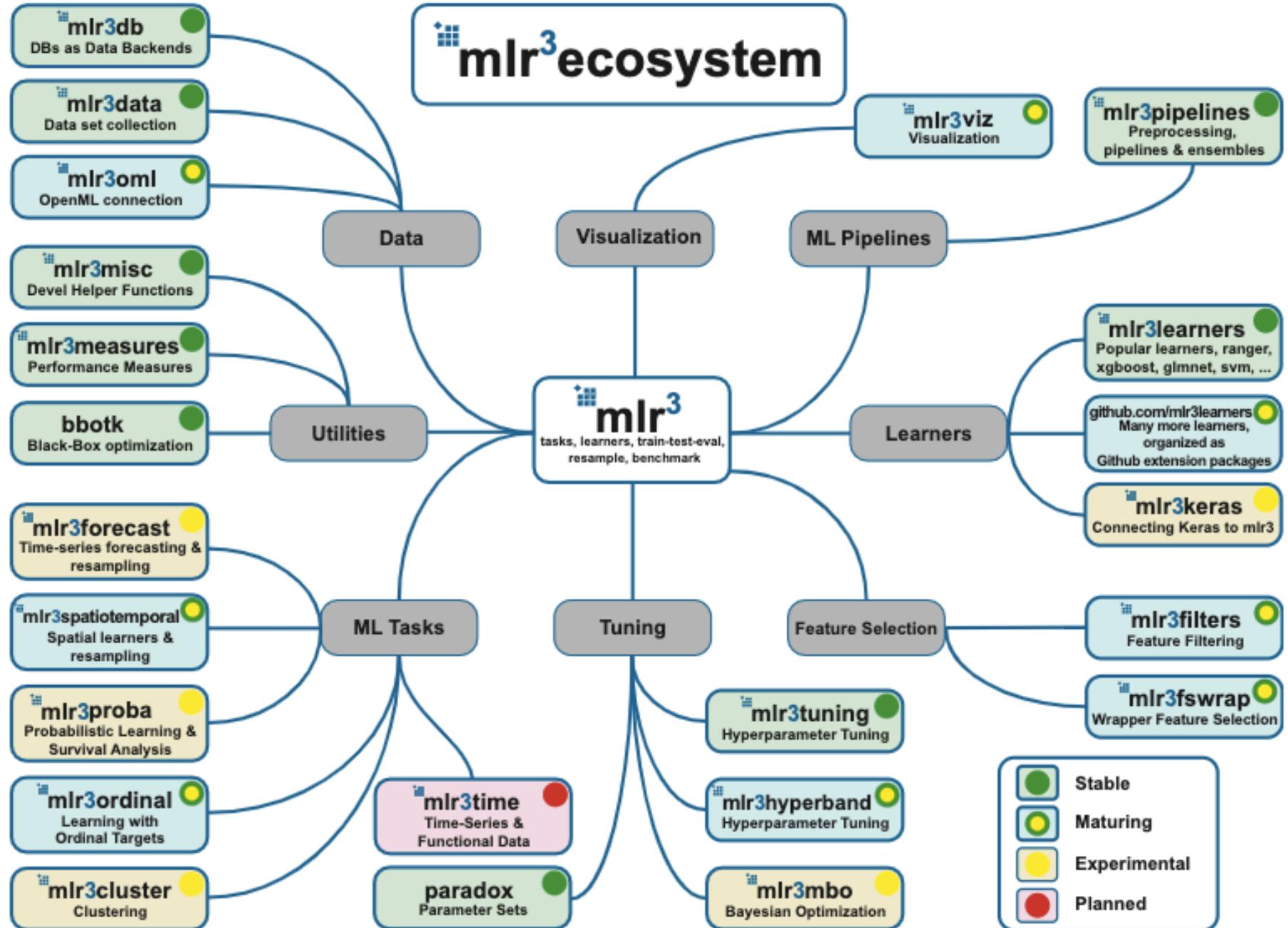
Efficient, object-oriented programming on the building blocks of machine learning. Successor of [mlr](#).

[tic](#) passing   [JOSS](#) [10.21105/joss.01903](#)   CRAN [0.5.0 – 17 days ago](#)   CRAN OK   [codecov](#) 92%   stackoverflow [mlr3](#)  
 dependencies [12/15](#)

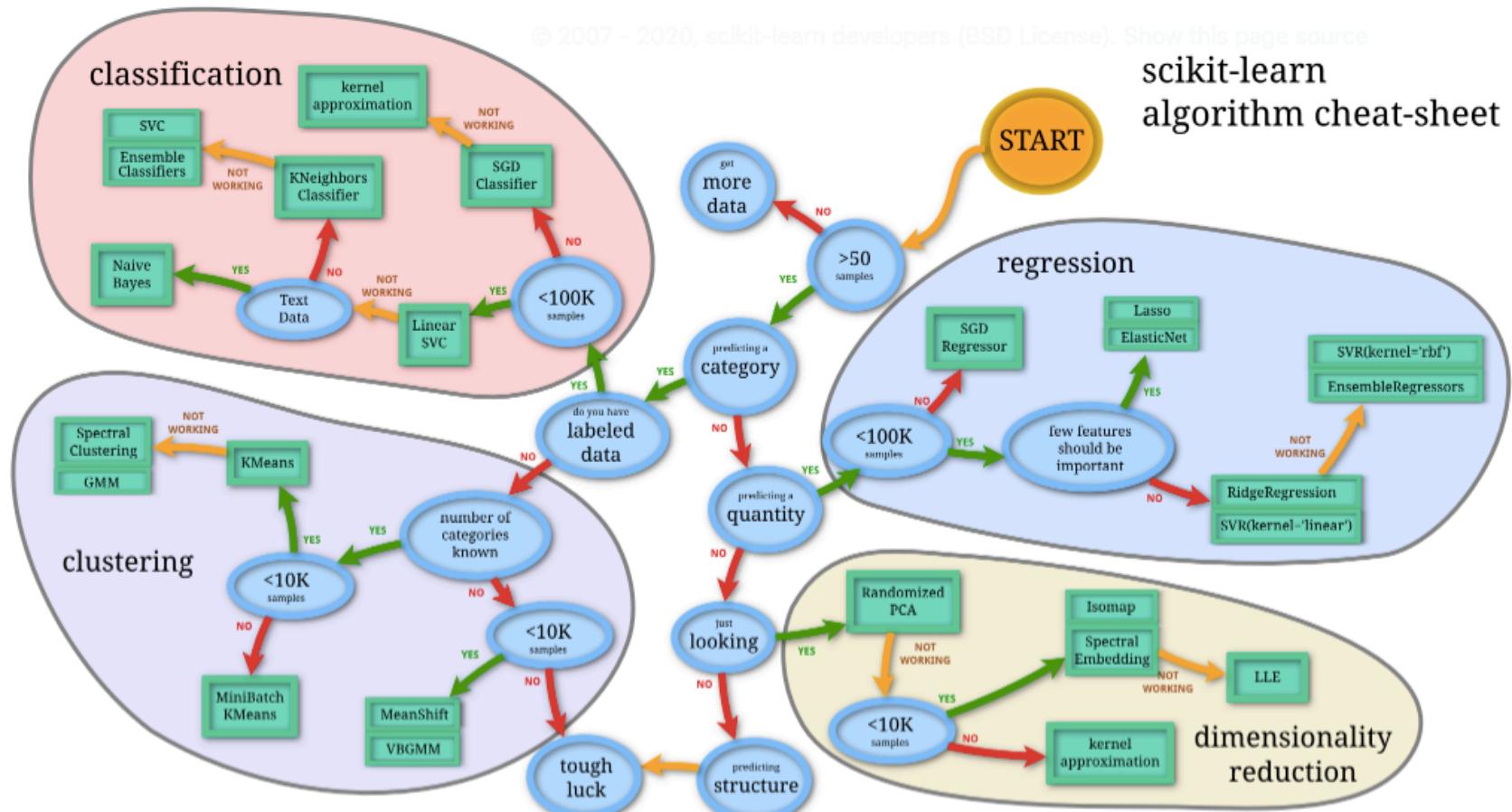
## Resources (for users and developers)

- We *started* writing a [book](#). This should be the central entry point to the package.
- The [mlr3gallery](#) has some case studies and demonstrates how frequently occurring problems can be solved. It is still in early days so stay tuned for more to come.
- [Reference manual](#)

# Extension Packages

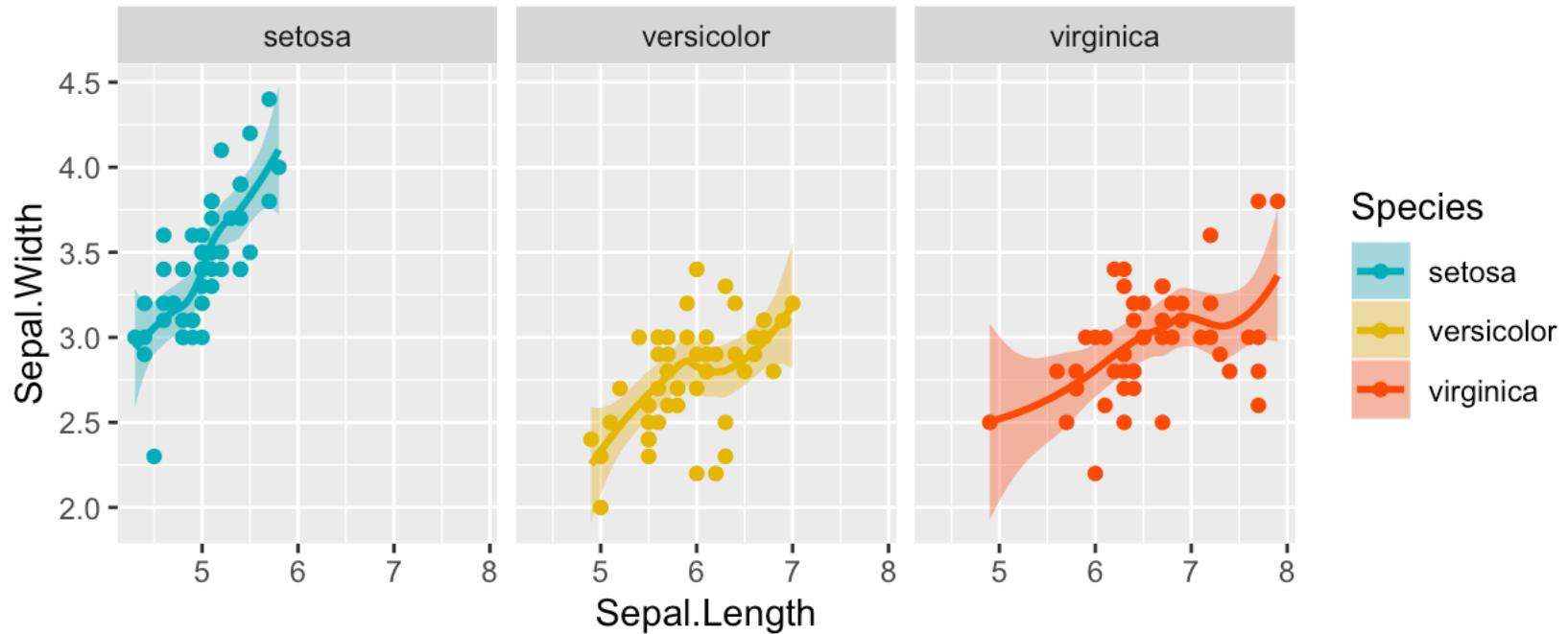


# scikit-learn / framework em Python



# ggplot2

- Visualização dos dados :)



# ggplot2

