MÉTODOS E MODELOS AVANÇADOS EM CIÊNCIA DE DADOS

Aula 08 - Fluxo de Ciência de Dados

Prof. Rafael G. **Mantovani**

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

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

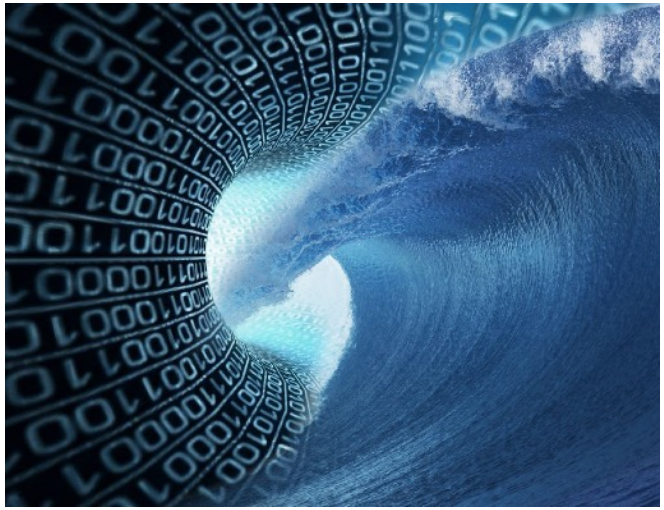
1. Introdução
2. Aplicações
3. Conceitos gerais
4. Fluxo de ciência de dados
5. Síntese
6. Referências

Introdução

poucos dados

Introdução

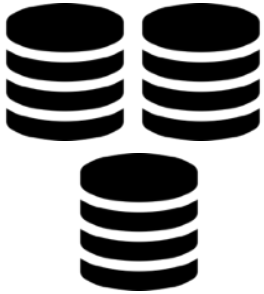
poucos dados



imensa quantidade
de dados (big data)

Introdução

Pessoas



Dados

Dispositivos

Introdução

Pessoas

Dados

Dispositivos

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

Introdução

- Mudança de realidade

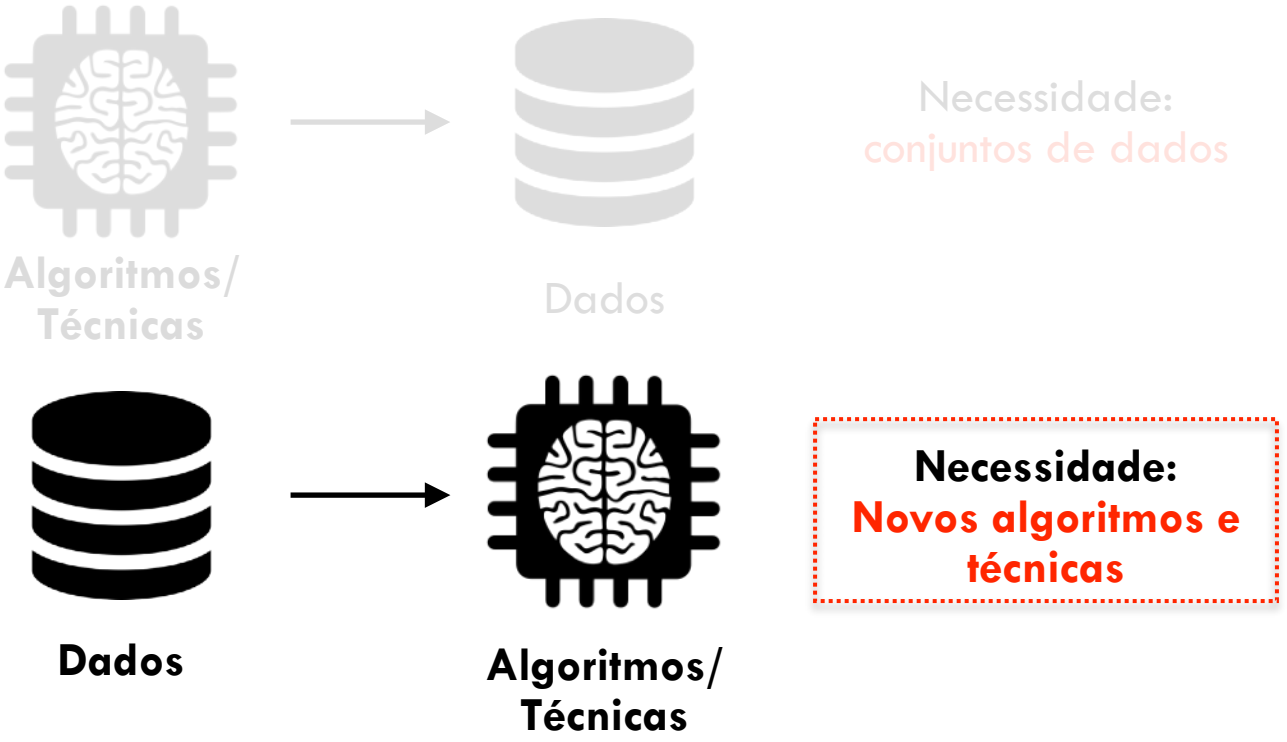
Algoritmos/ Técnicas

Dados

Necessidade:
conjuntos de dados

Introdução

- Mudança de realidade



Introdução

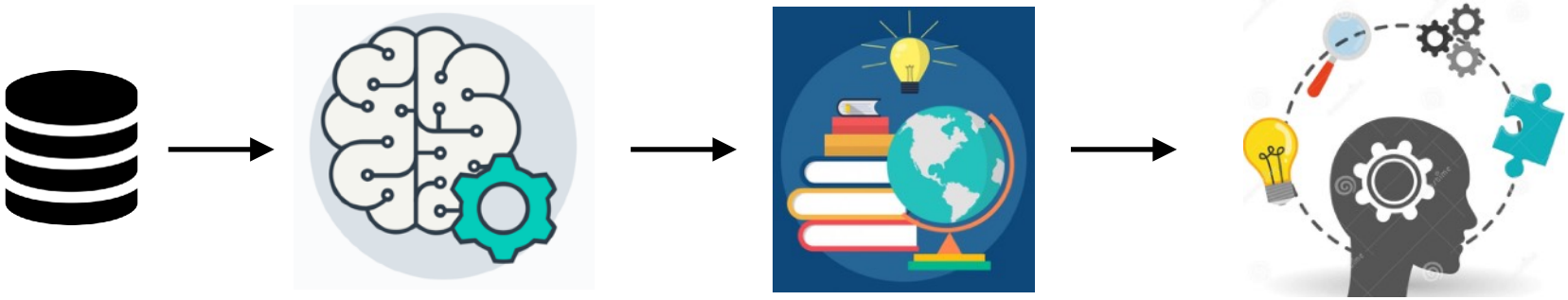
Dados

Algoritmo
AM

Conhecimento

Solução

Introdução

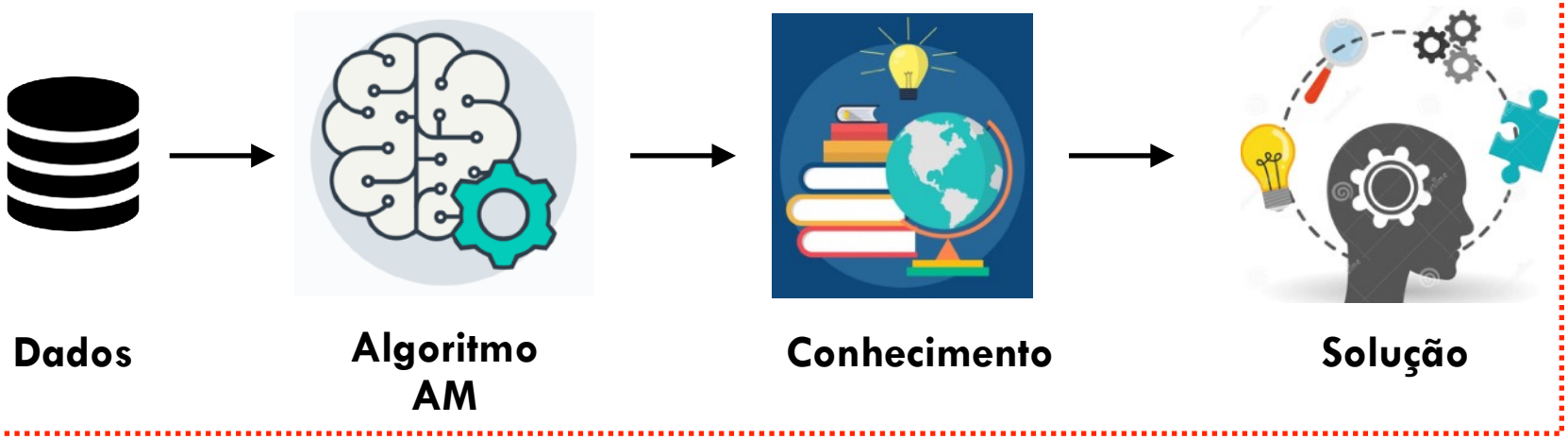


Dados → **Algoritmo AM** → **Conhecimento** → **Solução**

- 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

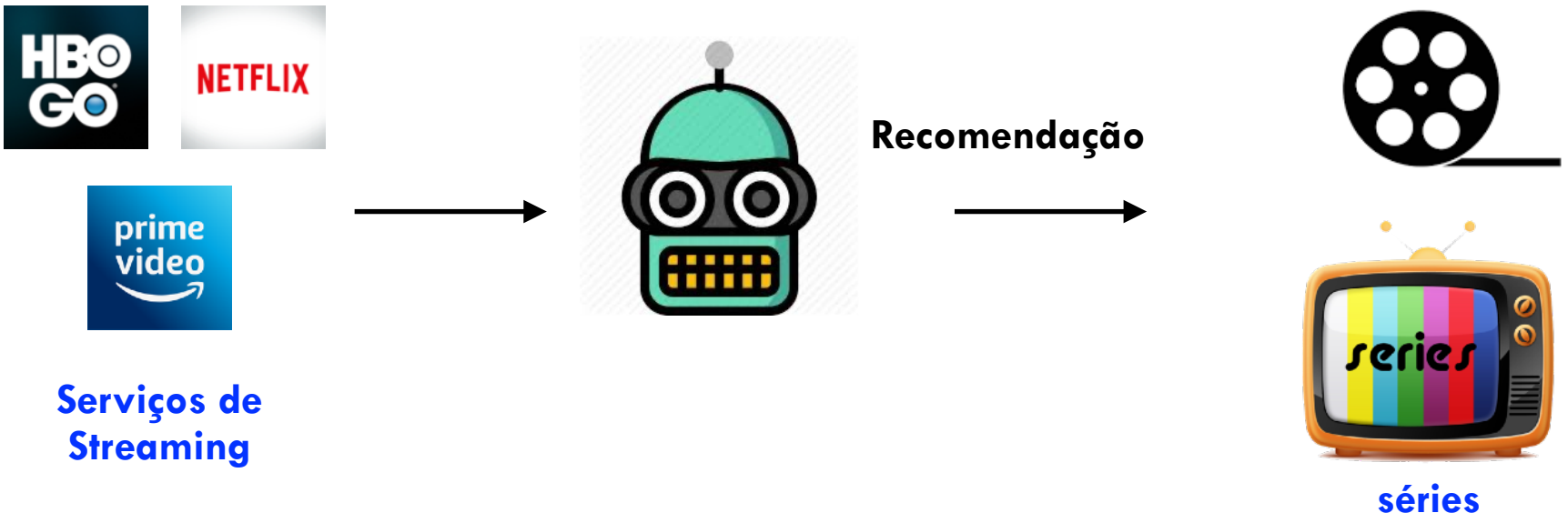
- Onde isso é usado?

Aplicações

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

Aplicações

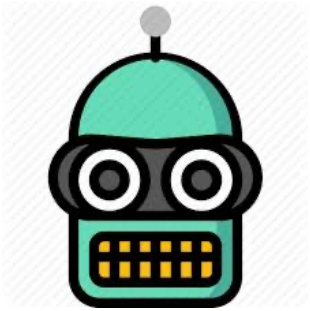
- Onde isso é usado? **Sistemas Recomendadores**



Aplicações

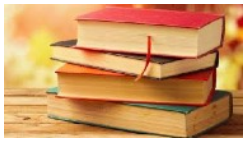
- Onde isso é usado? **Sistemas Recomendadores**

Serviços de e-commerce



Recomendação

livros



games

Produtos em geral

Aplicações

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

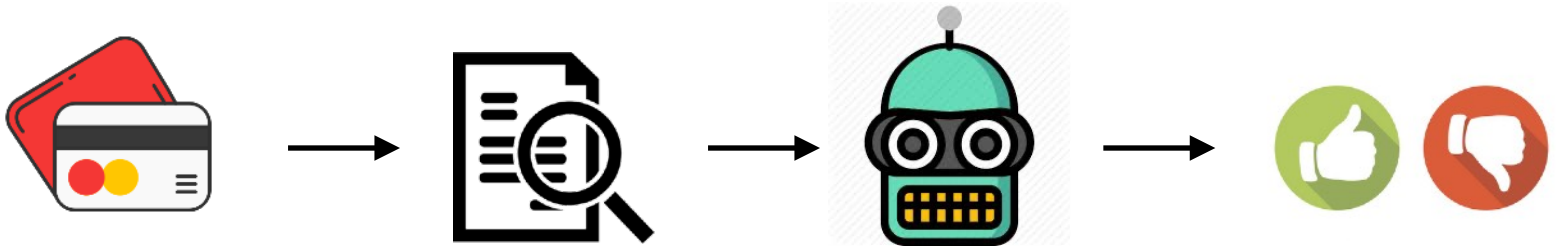
Aplicações

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

Aplicações

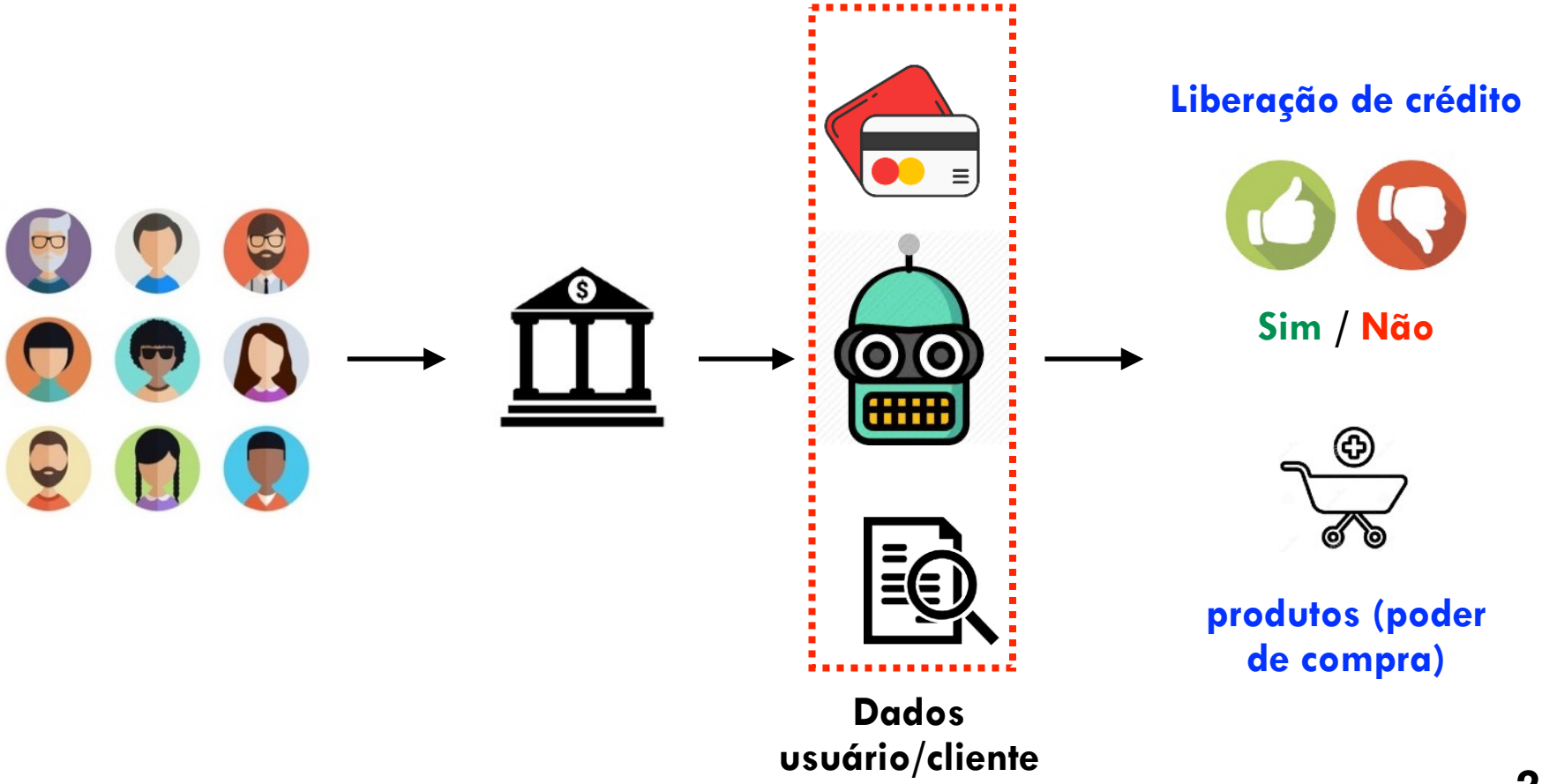
- Onde isso é usado? **Bancos/Mercado financeiro**

Deteção de Fraudes



Aplicações

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

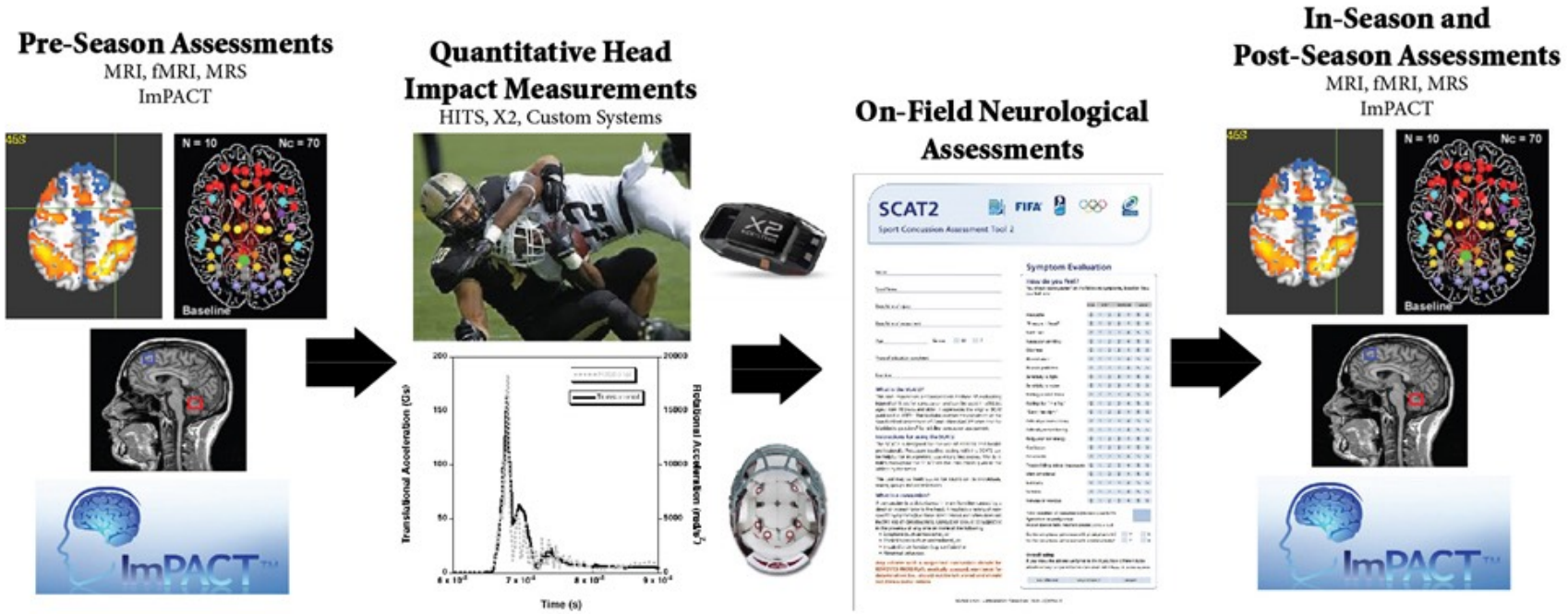


Aplicações

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

Aplicações

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



Aplicações

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

Aplicações

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

Aplicações

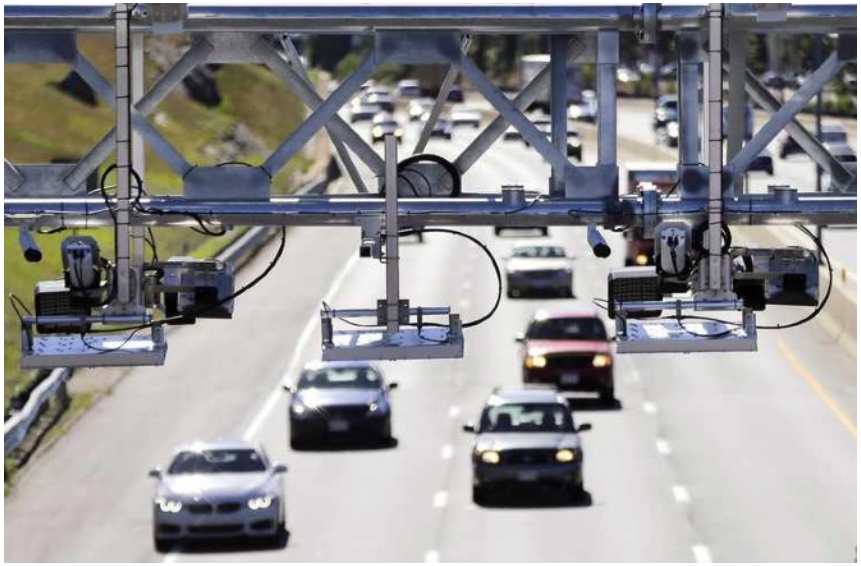
- Onde isso é usado? **Reconhecimento de Imagens**

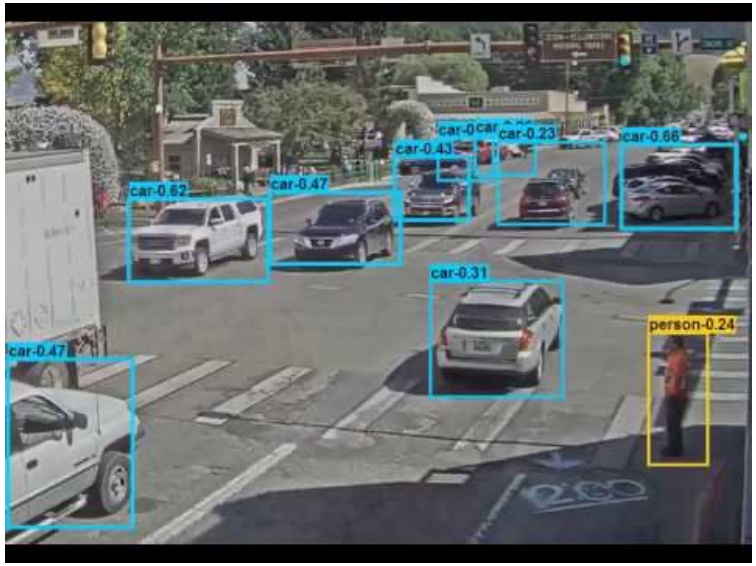
Sistemas de Segurança

Aplicações

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

Sistemas de Vigilância



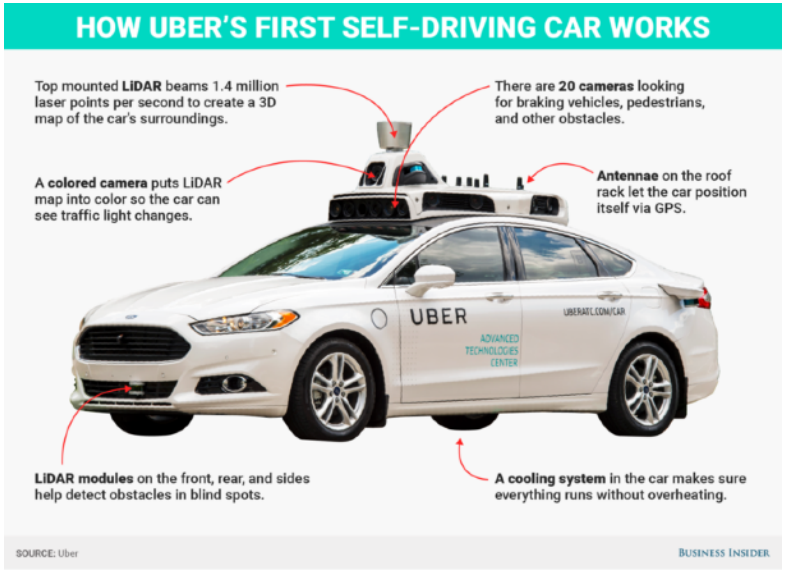


Aplicações

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

Aplicações

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



Uber

Tesla

Aplicações

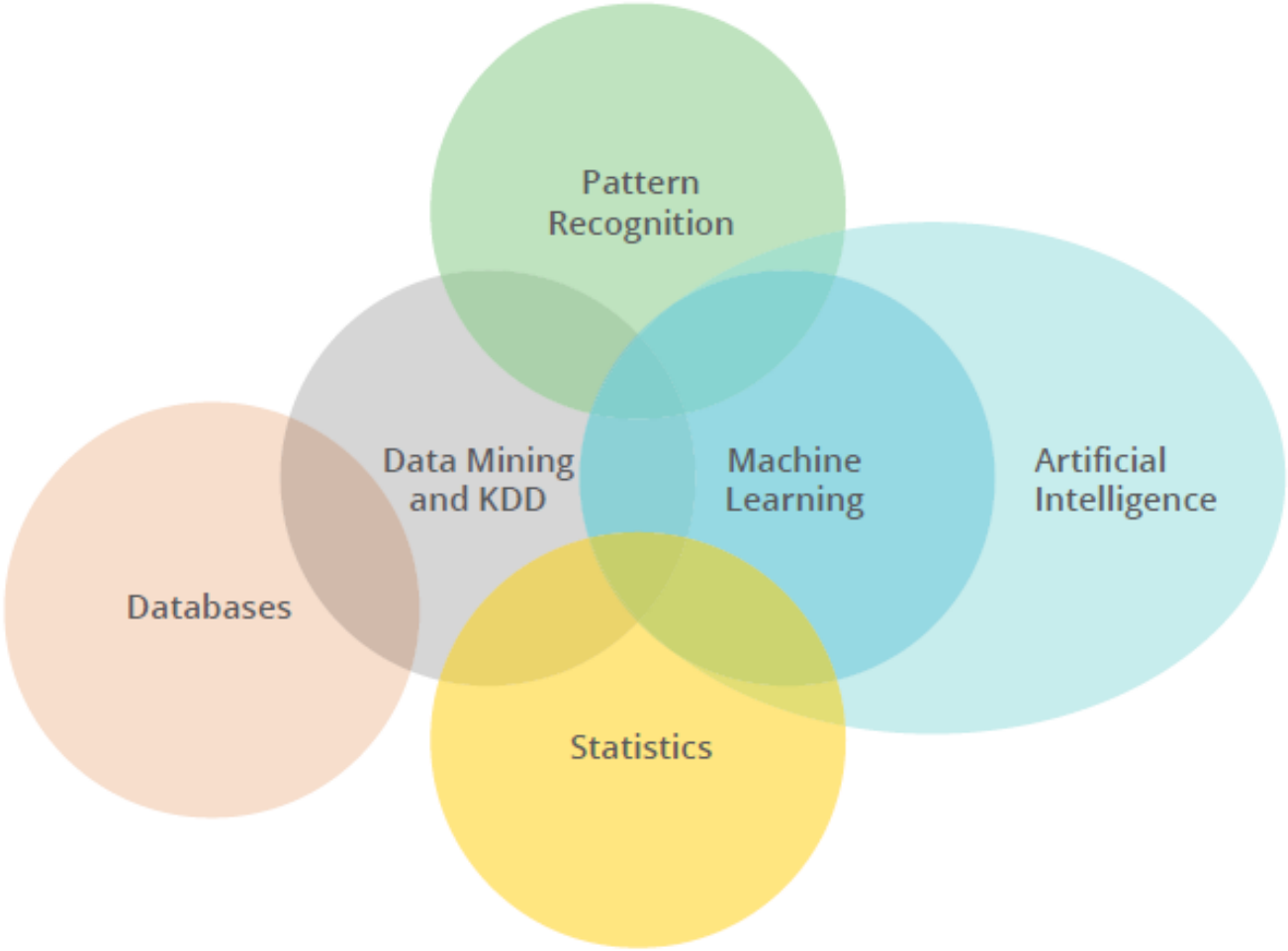
- Onde isso é usado? **Veículos Autônomos**

LRM - ICMC/USP, São Carlos - SP

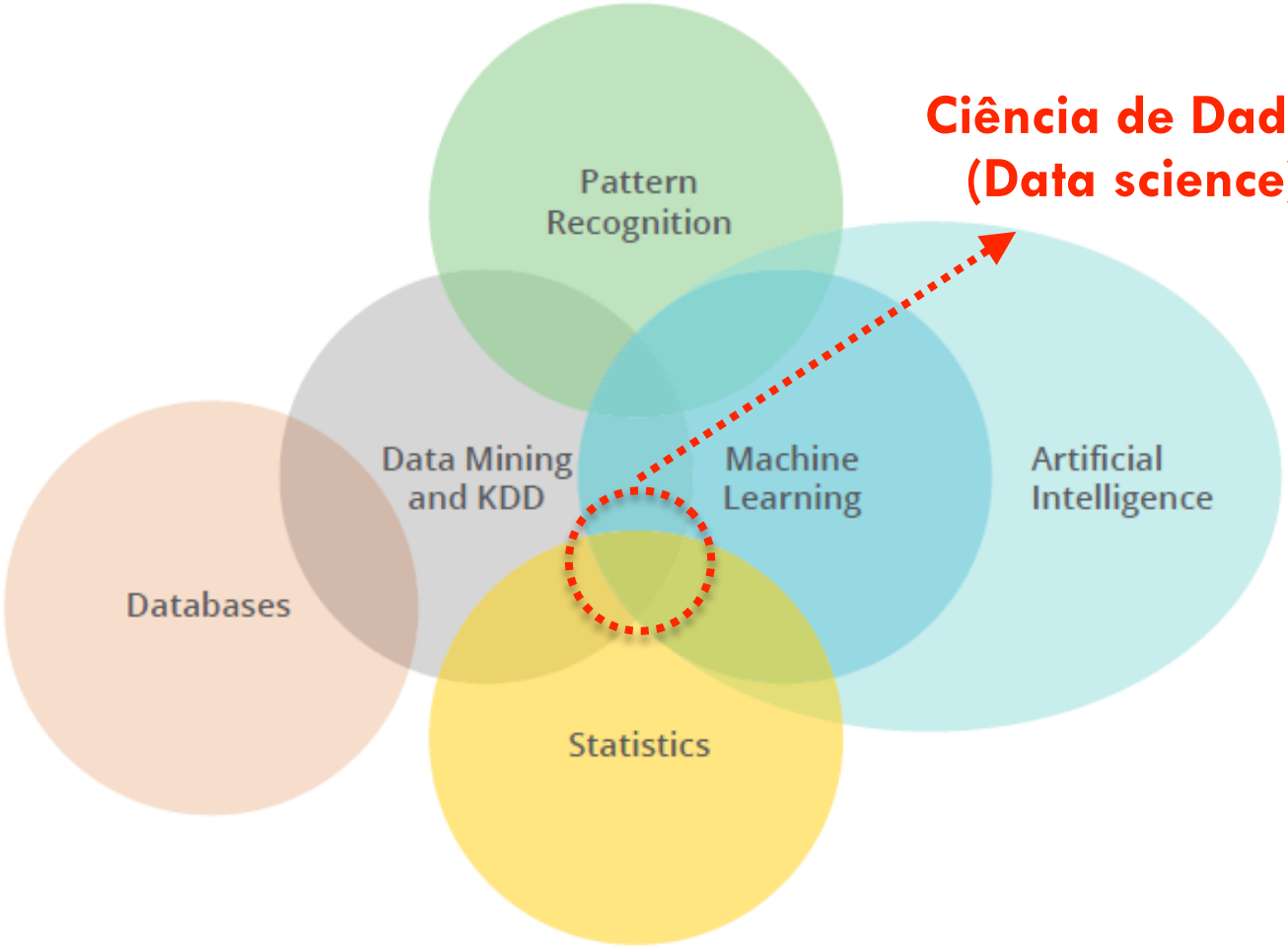
Roteiro

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

Conceitos Gerais



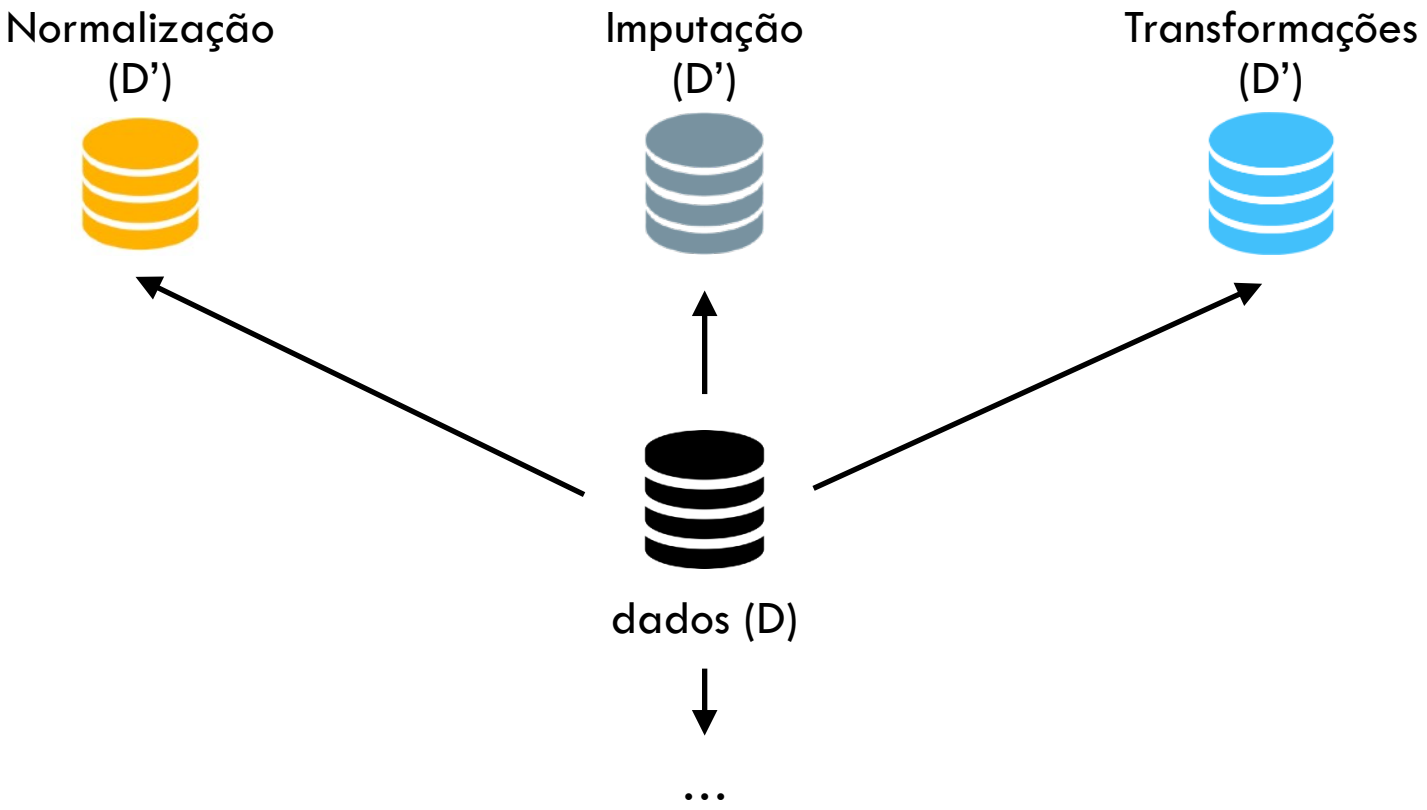
Conceitos Gerais



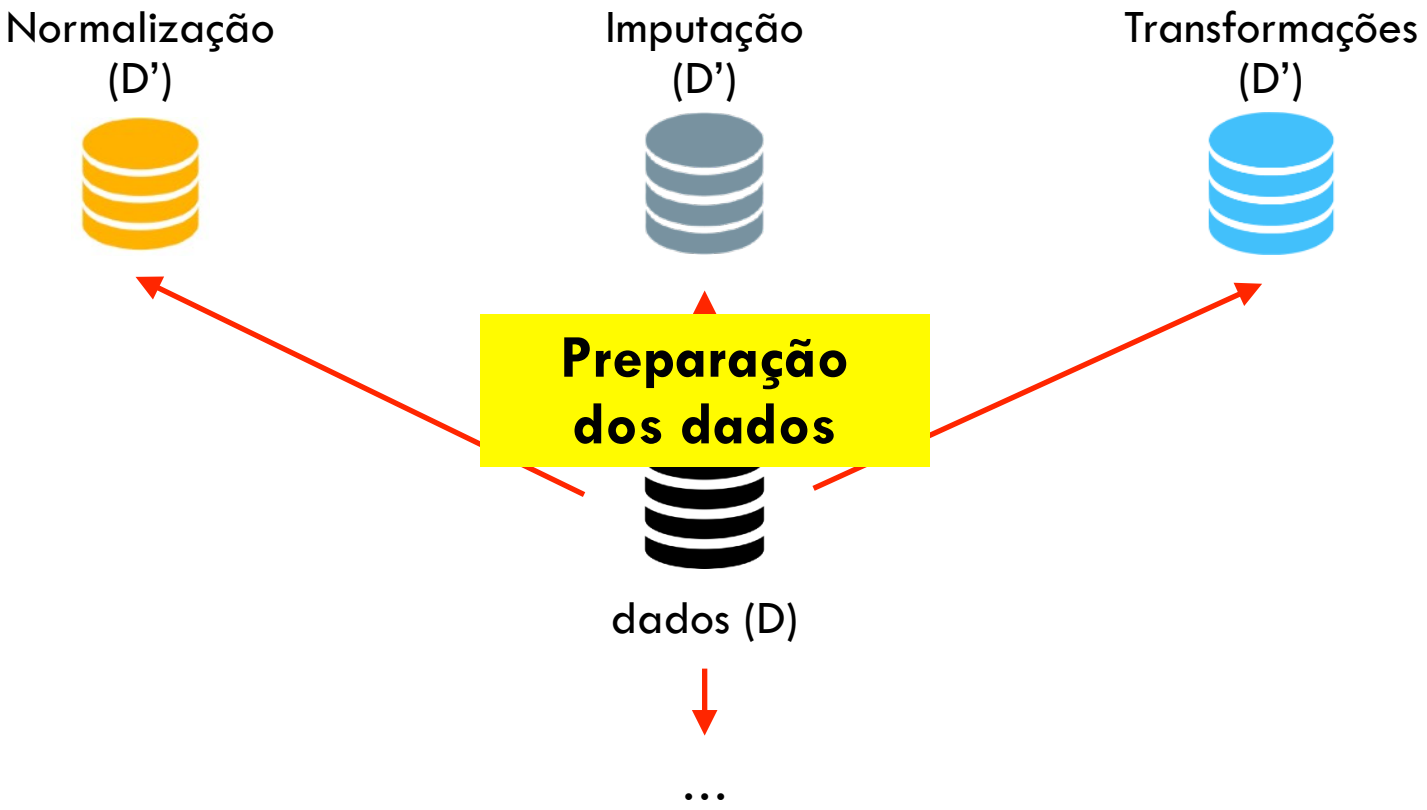
Mineração de dados (Data Mining)

dados (D)

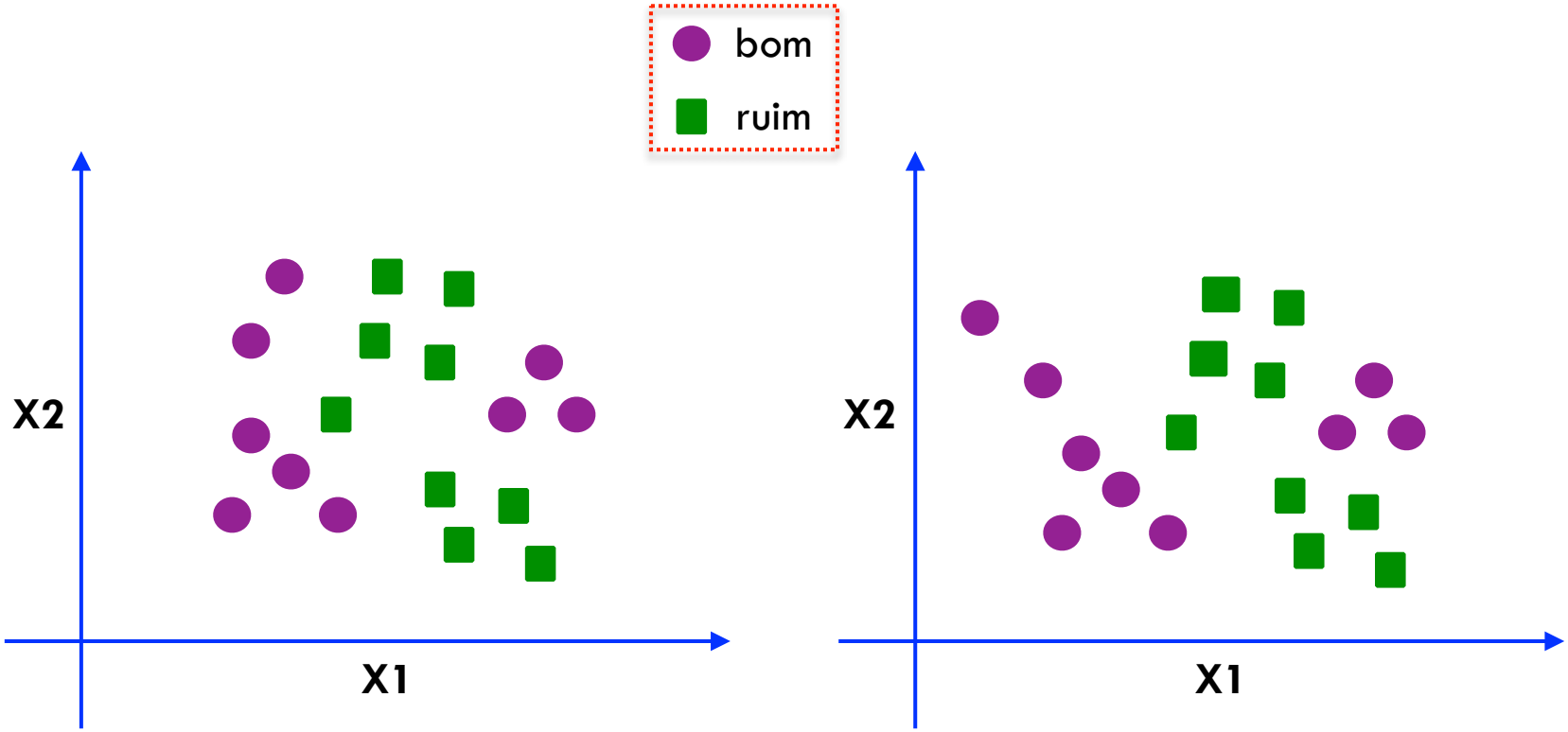
Mineração de dados (Data Mining)



Mineração de dados (Data Mining)



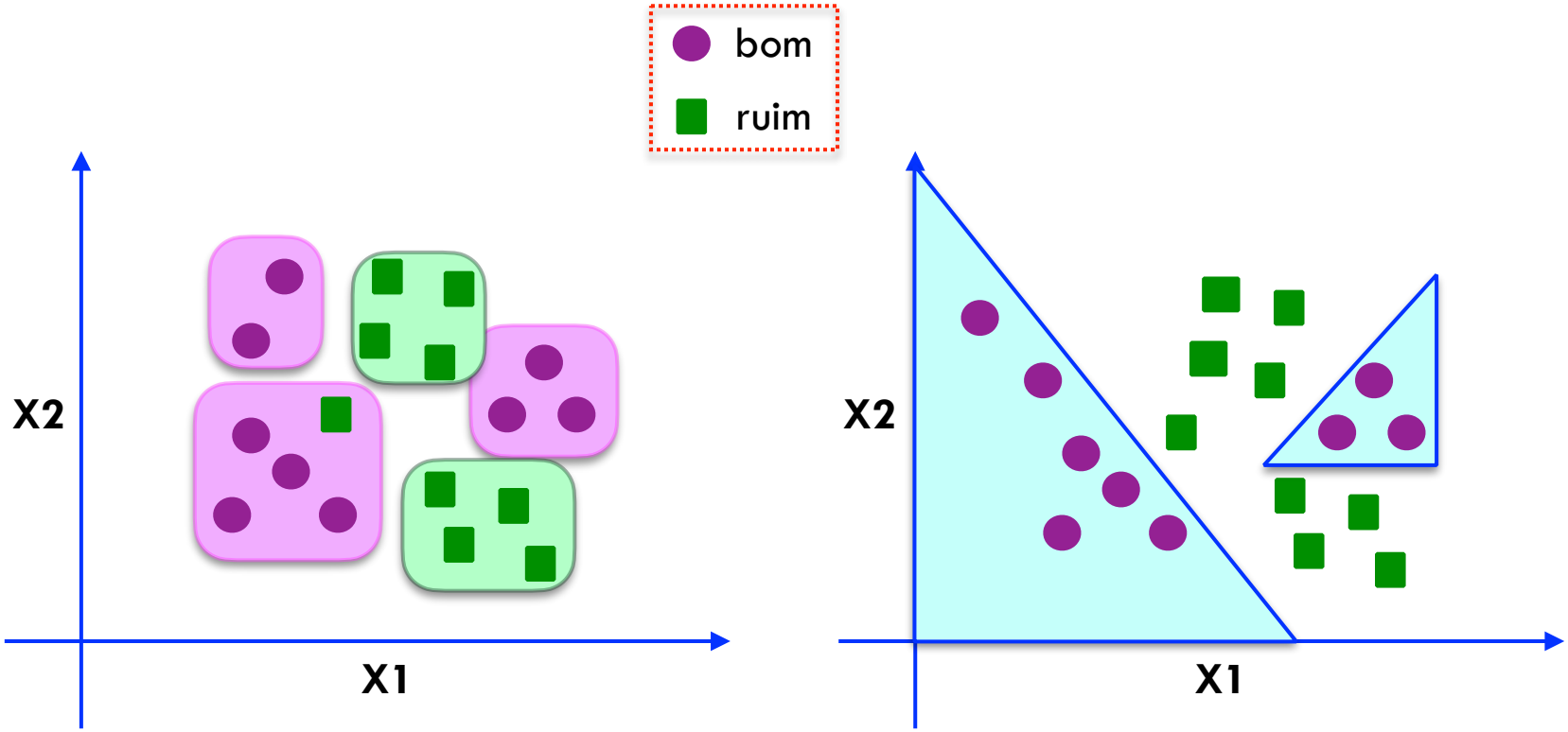
Aprendizado de Máquina (Machine Learning)



Tarefa descritiva
(Clustering)

Tarefa preditiva
(Classificação/ Regressão)

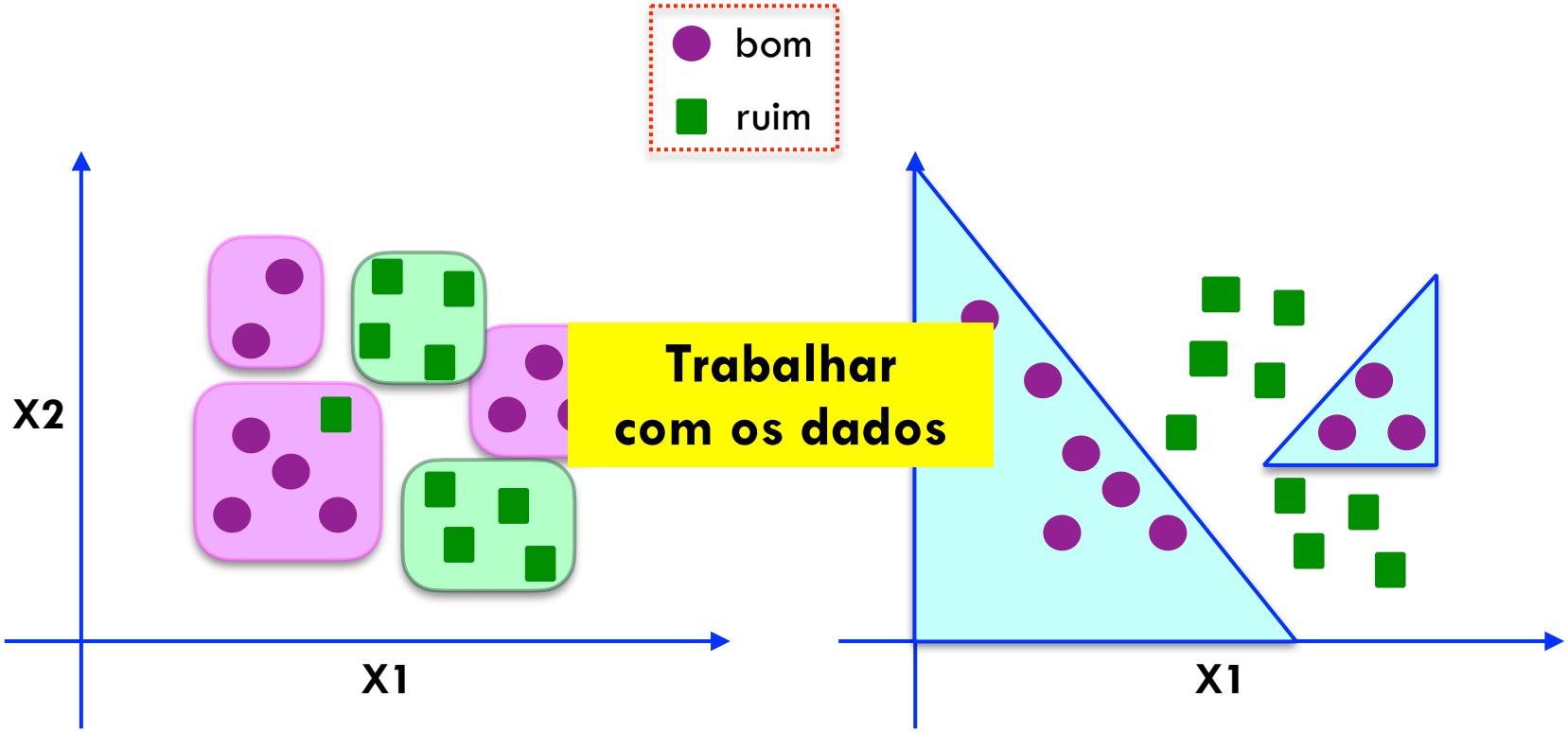
Aprendizado de Máquina (Machine Learning)



Tarefa descritiva
(Clustering)

Tarefa preditiva
(Classificação/ Regressão)

Aprendizado de Máquina (Machine Learning)

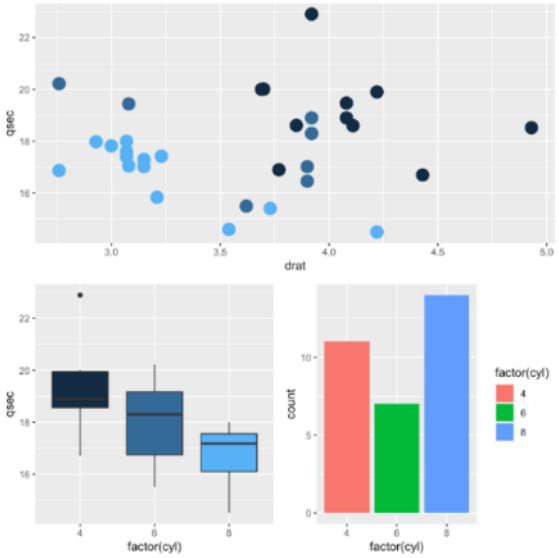


Tarefa descritiva
(Clustering)

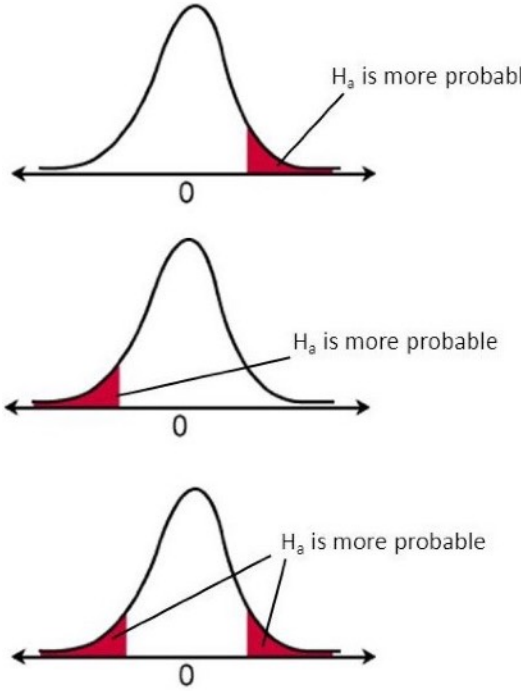
Tarefa preditiva
**(Classificação/
Regressão)**

Estatística (Statistics)

Técnicas de Amostragem

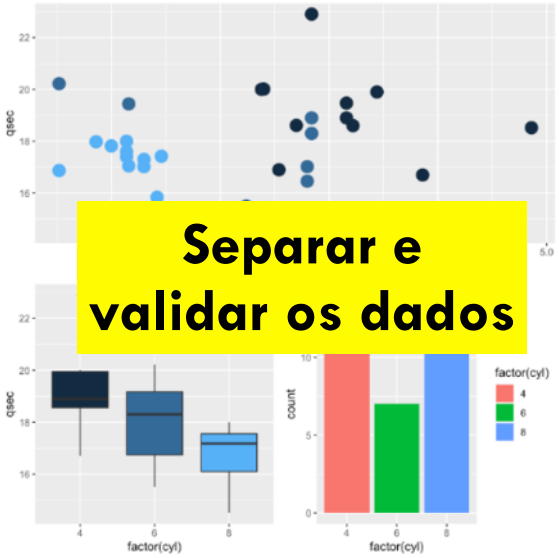


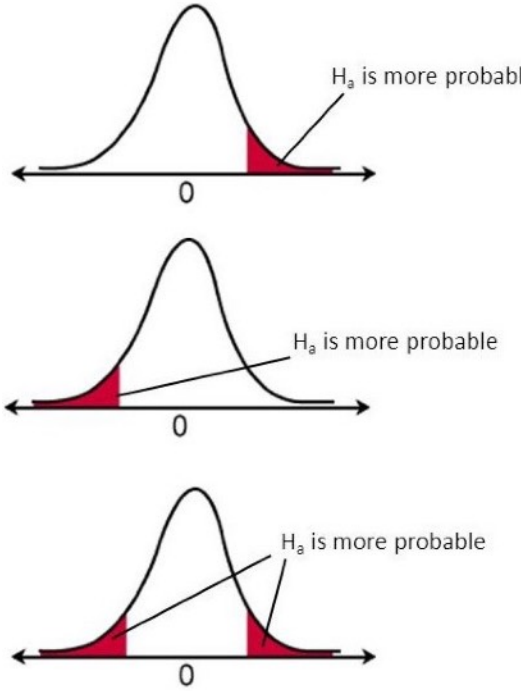
Estatística Descritiva



Testes de Hipótese

Estatística (Statistics)



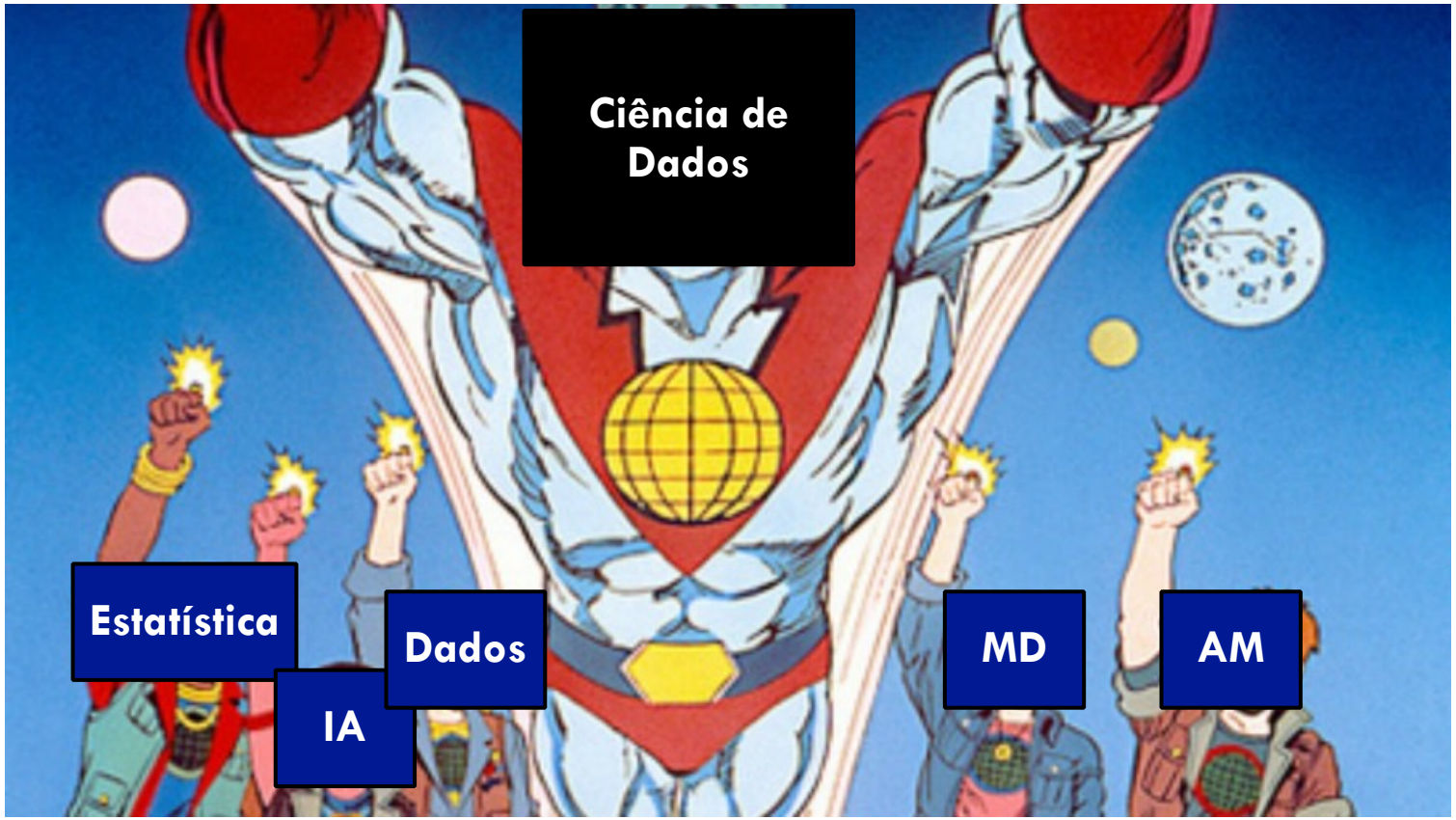


Técnicas de Amostragem

Estatística Descritiva

Testes de Hipótese

Ciência de Dados (Data science)



Ciência de Dados (Data science)

- Quantos algoritmos existem?

Ciência de Dados

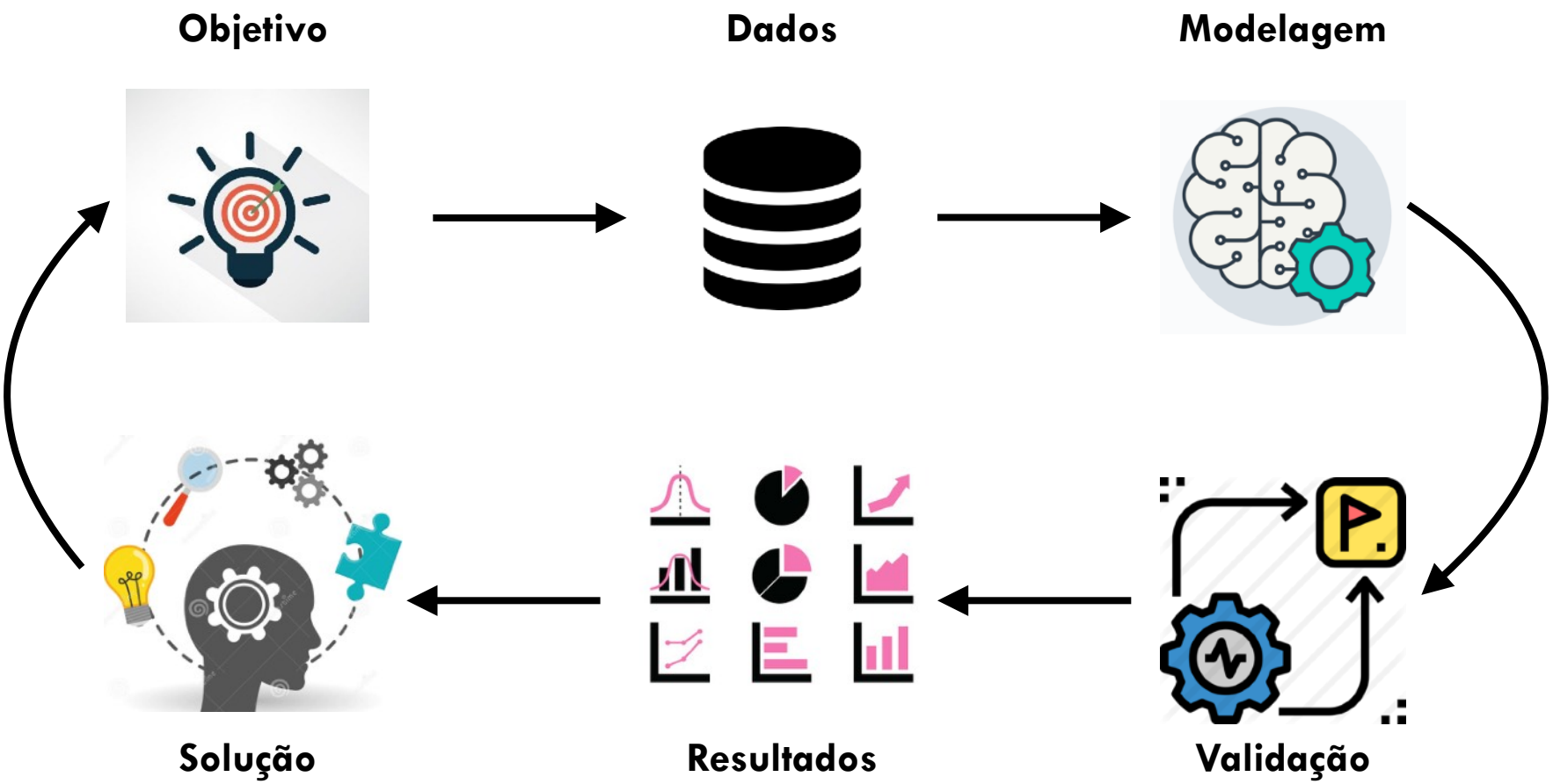
- Quantos algoritmos existem?

...

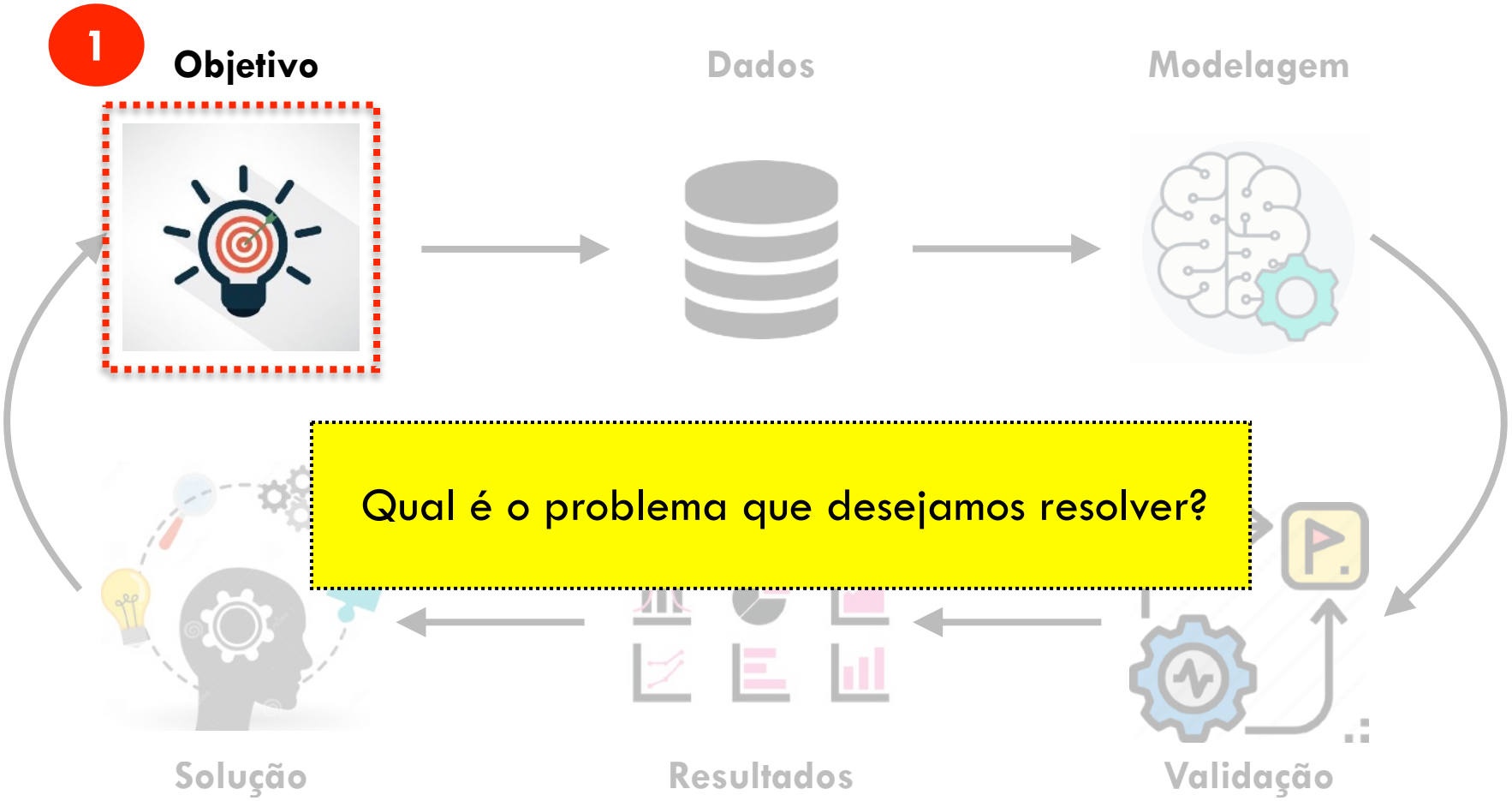
Roteiro

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

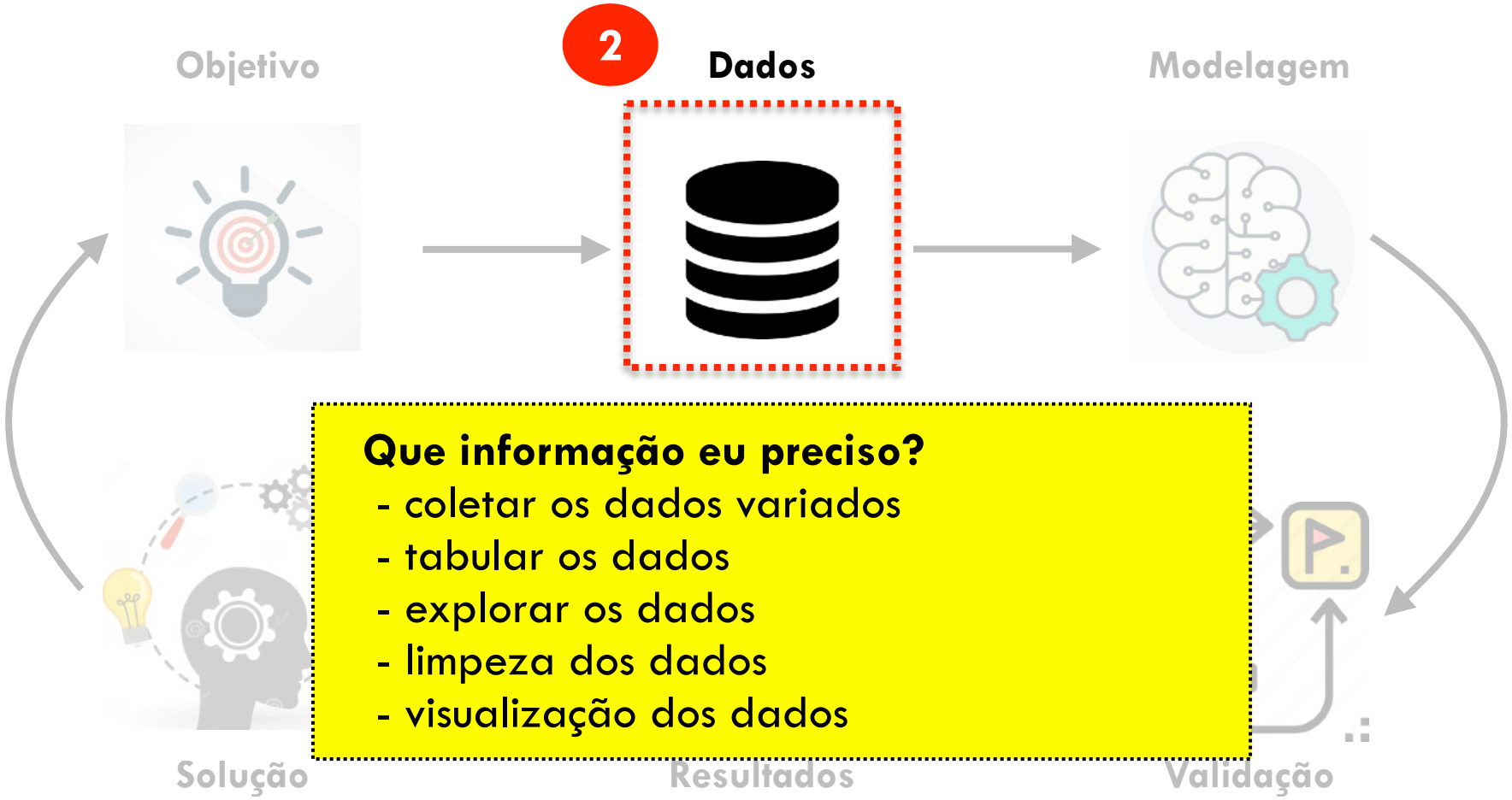
Fluxo de Ciência de Dados



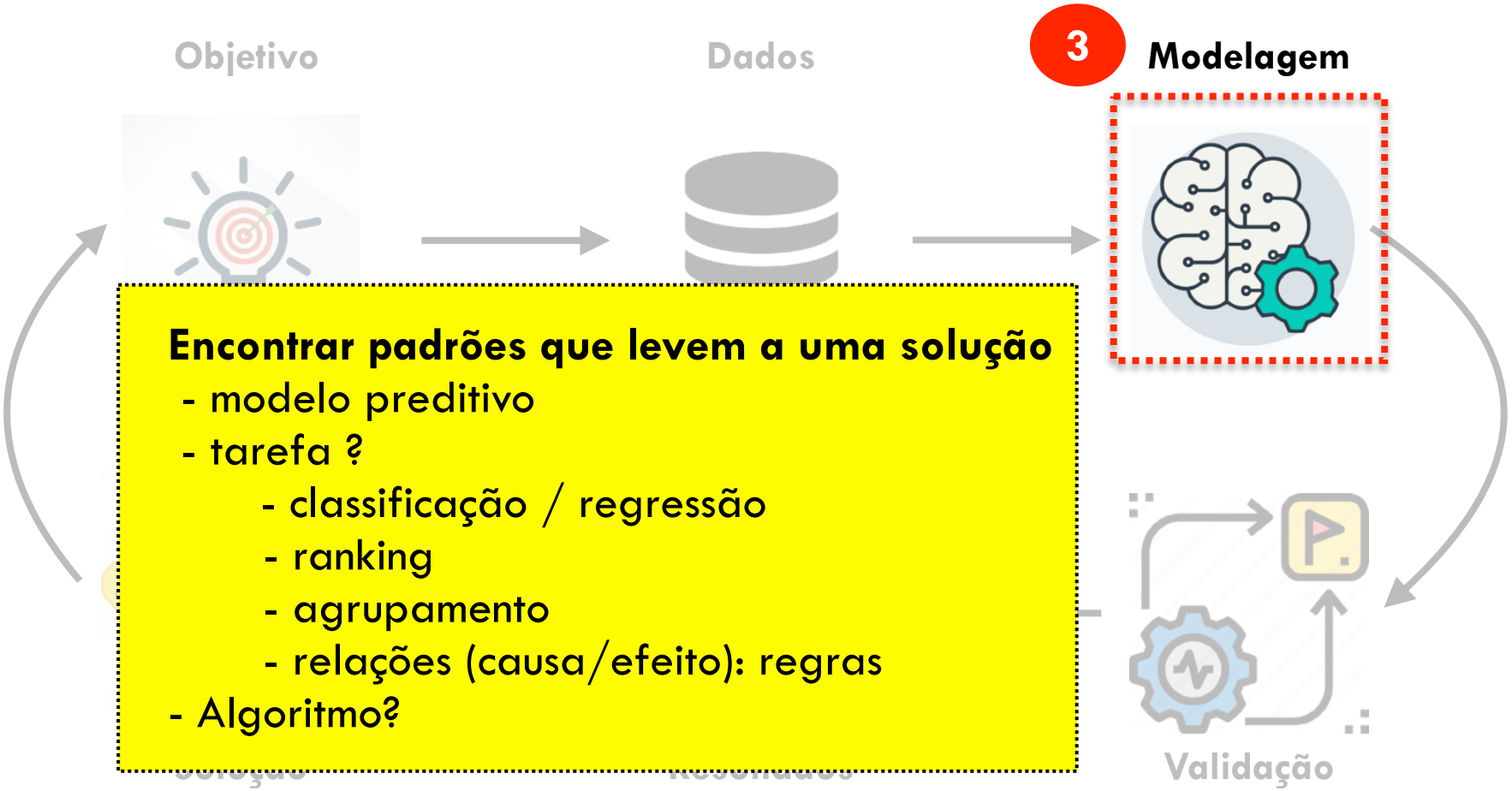
Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



Fluxo de Ciência de Dados



Encontrar padrões que levem a uma solução

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

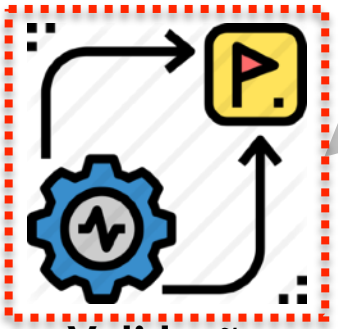
Fluxo de Ciência de Dados

Objetivo

Dados

Modelagem

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



Validação

Resultados

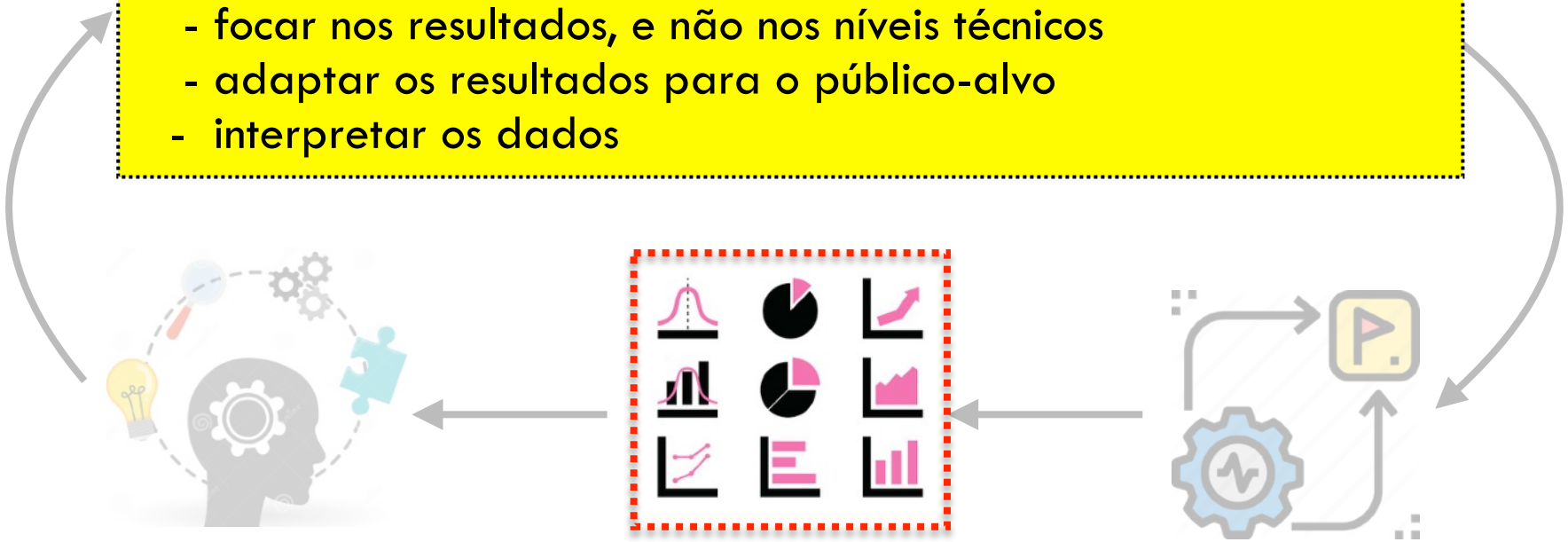
Solução

4

Fluxo de Ciência de Dados

Objetivo Dados Modelagem

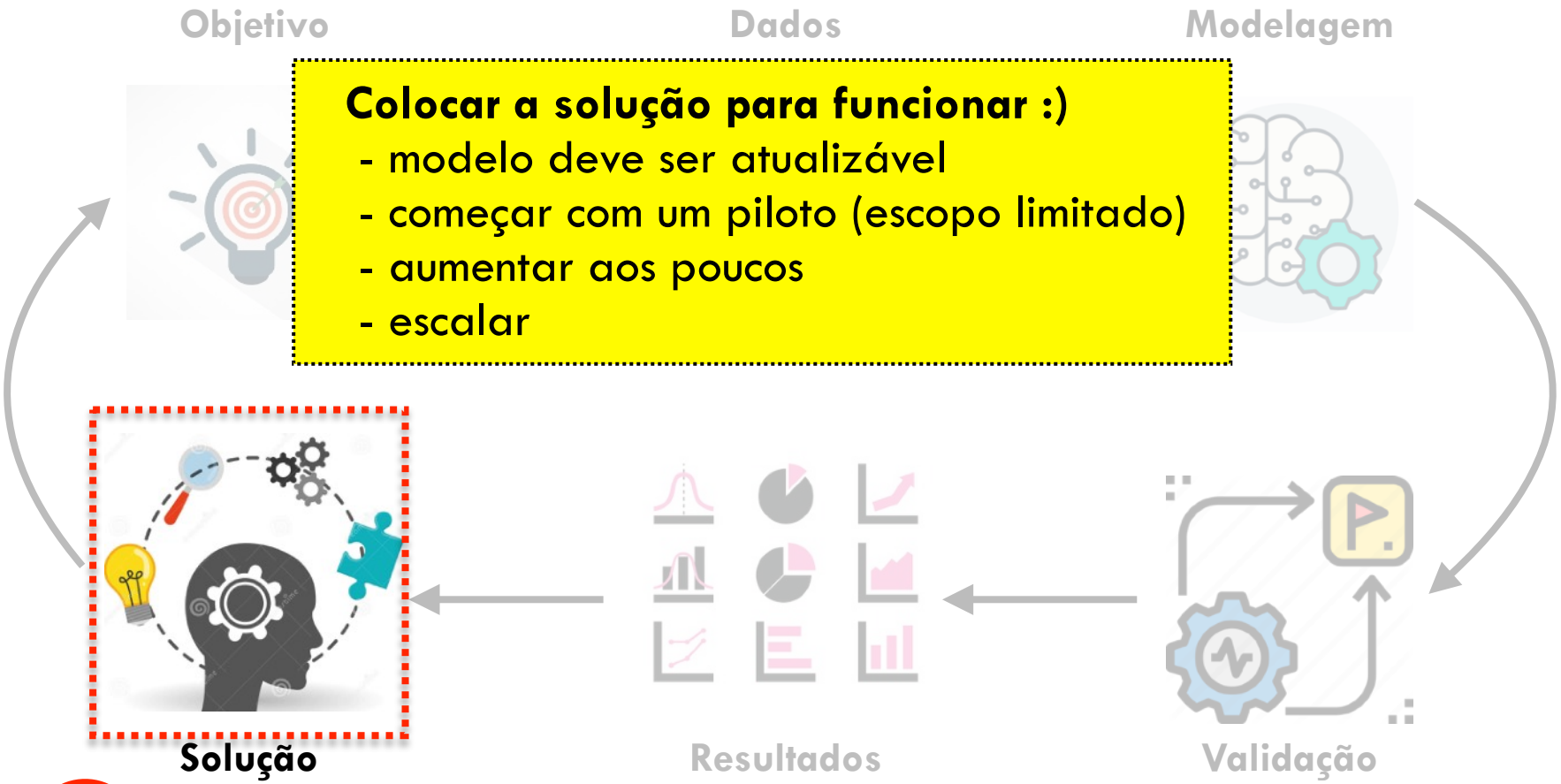
Mostrar que o modelo resolve o problema e como ele resolve
- gráficos / tabelas
- focar nos resultados, e não nos níveis técnicos
- adaptar os resultados para o público-alvo
- interpretar os dados



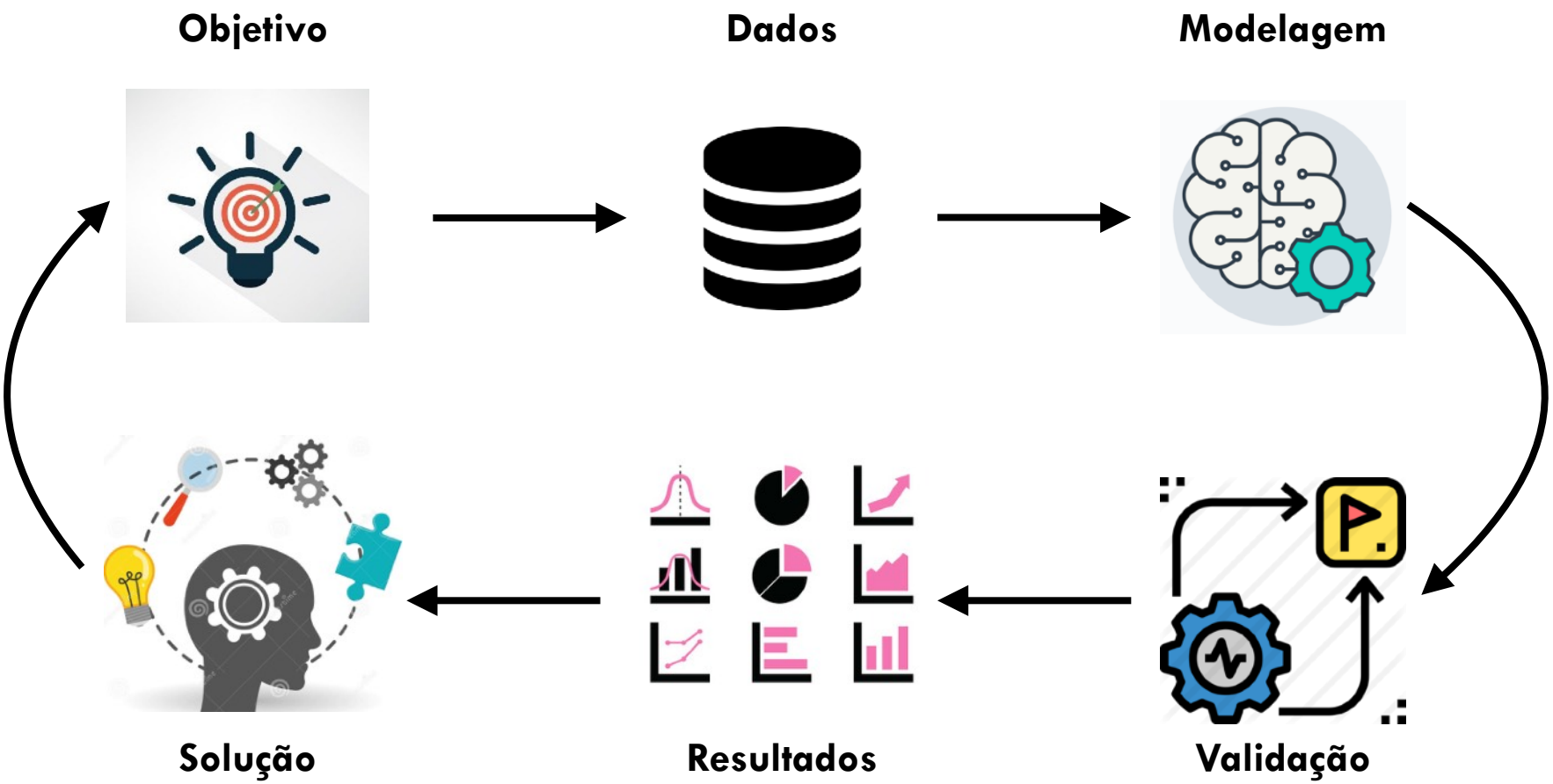
Solução **Resultados** Validação

5

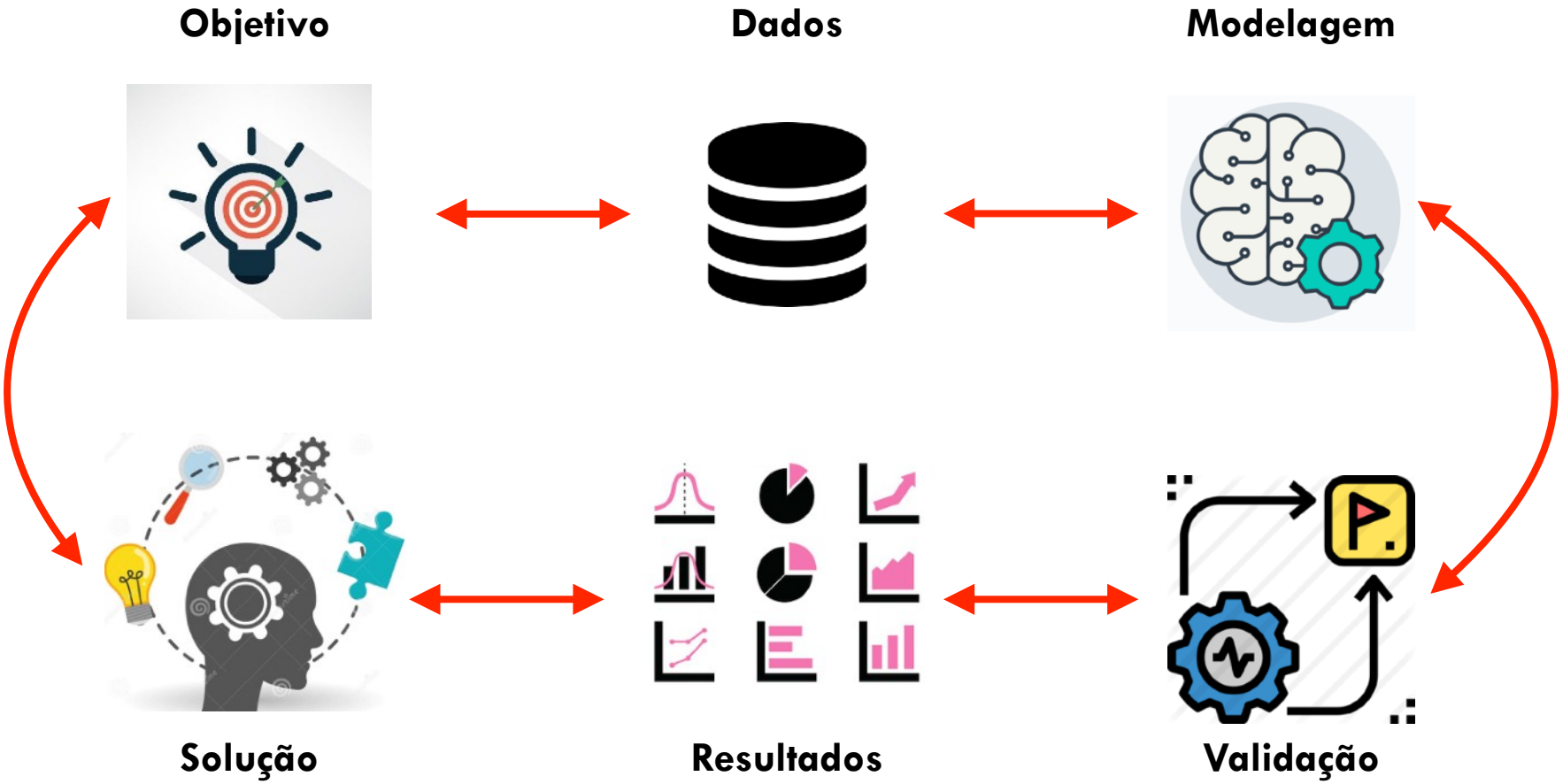
Fluxo de Ciência de Dados



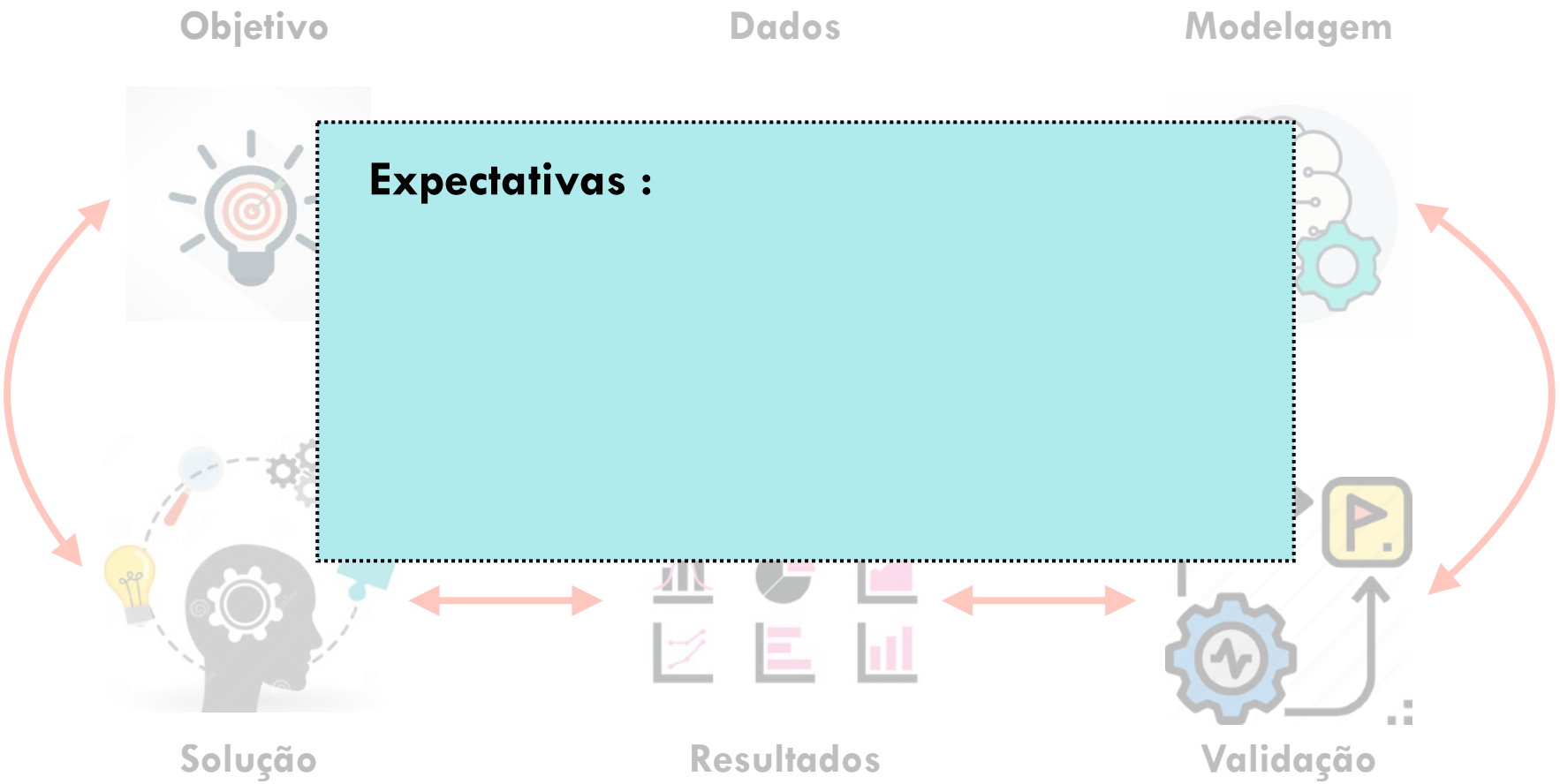
Fluxo de Ciência de Dados



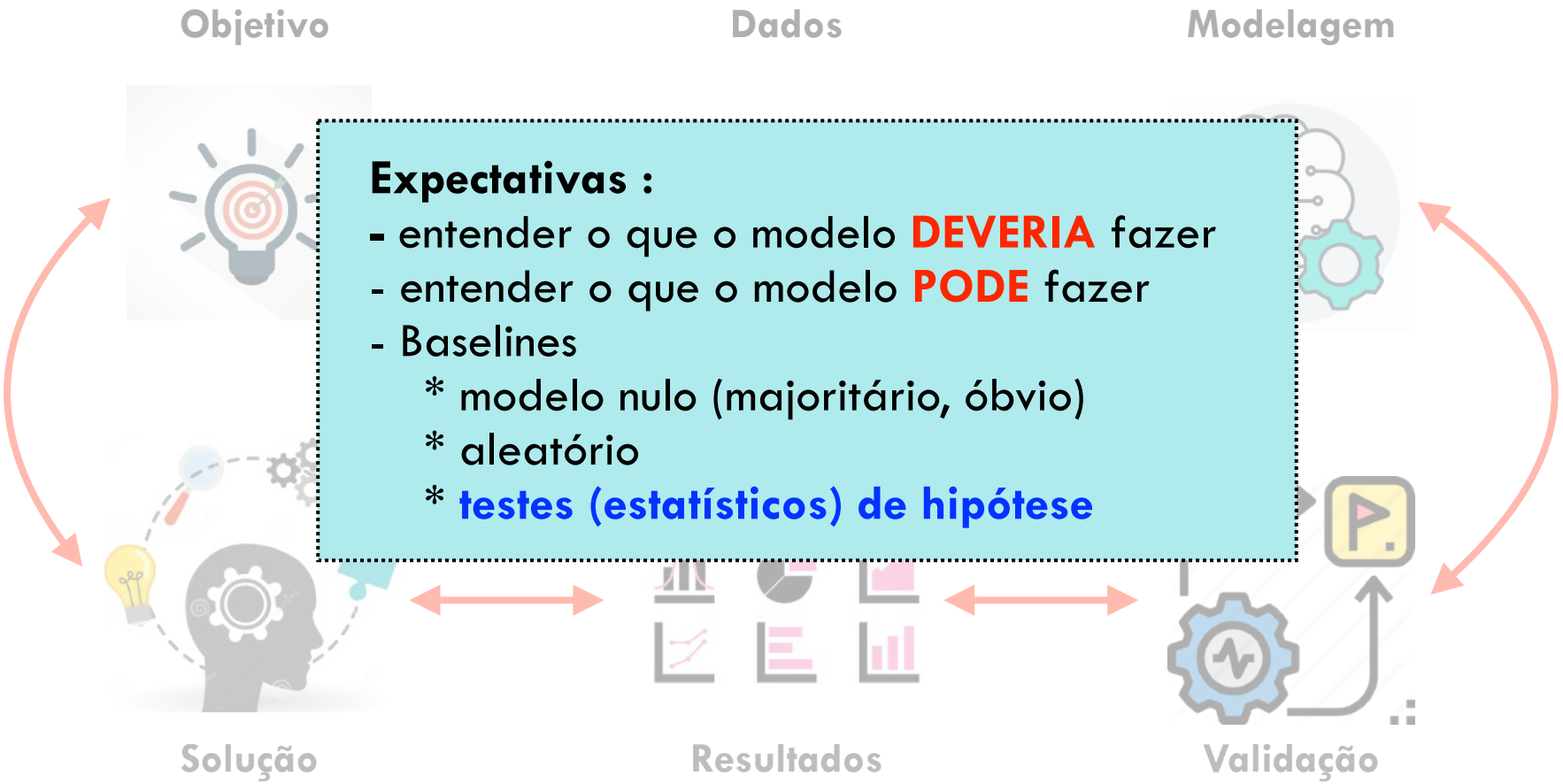
Fluxo de Ciência de Dados



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. Síntese
6. Referências

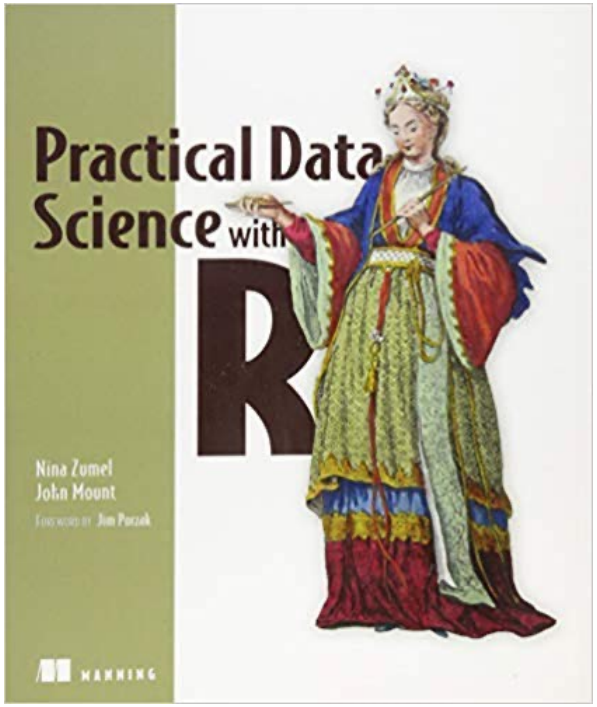
Síntese

- 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

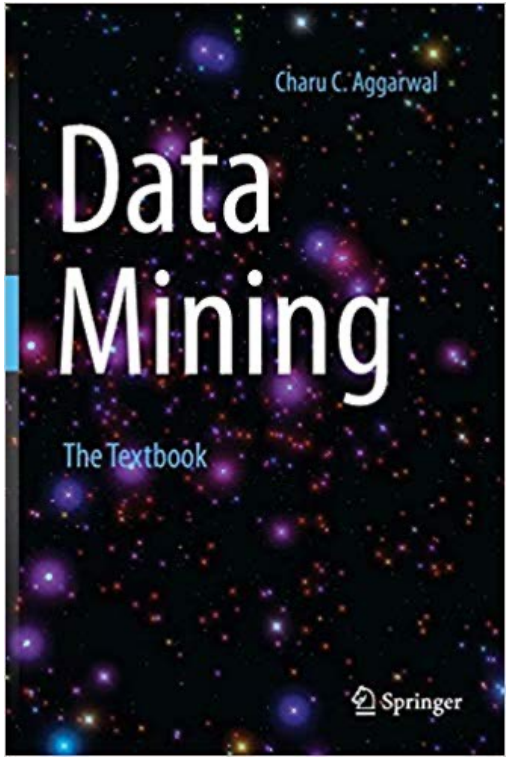
Roteiro

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

Referências

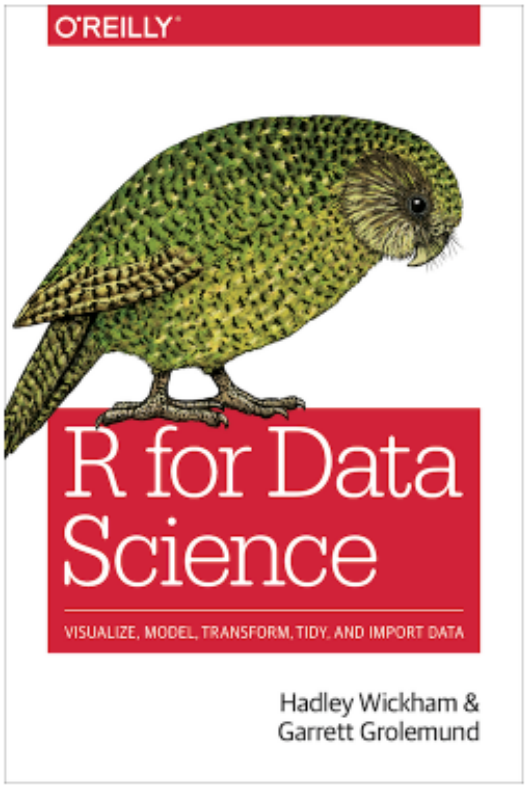


[Zumel & Mount, 2014]

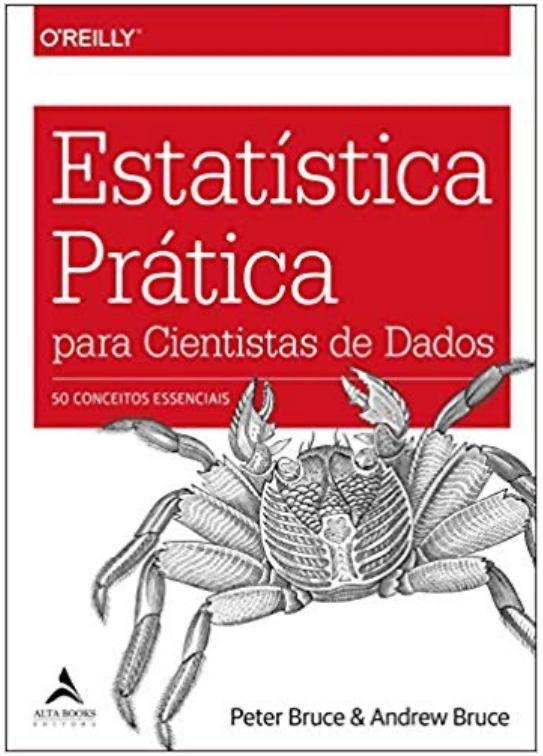


[Aggarwal, 2015]

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.

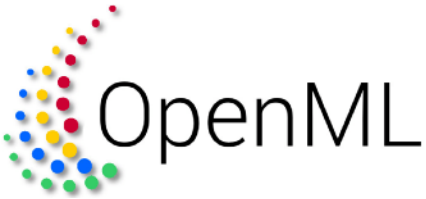
Obrigado :)

Rafael G. Mantovani
rafaelmantovani@utfpr.edu.br

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

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

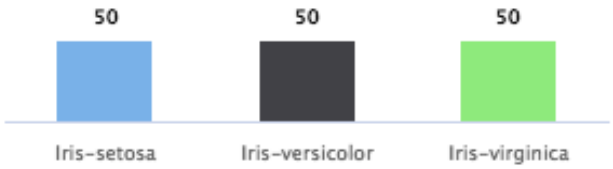
Help us complete this description → Edit

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

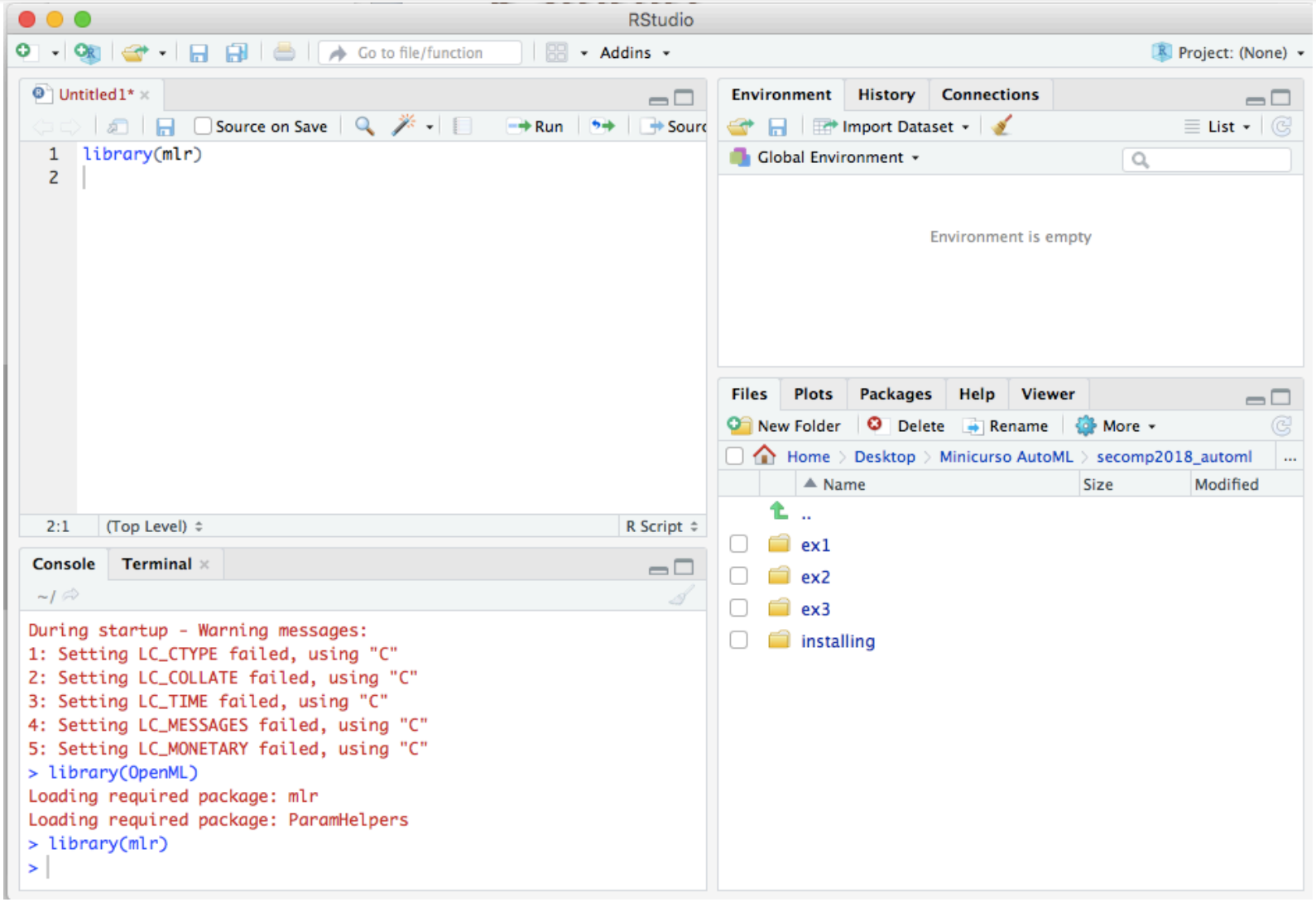
Iris Plants Database
This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers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

▾ Show all

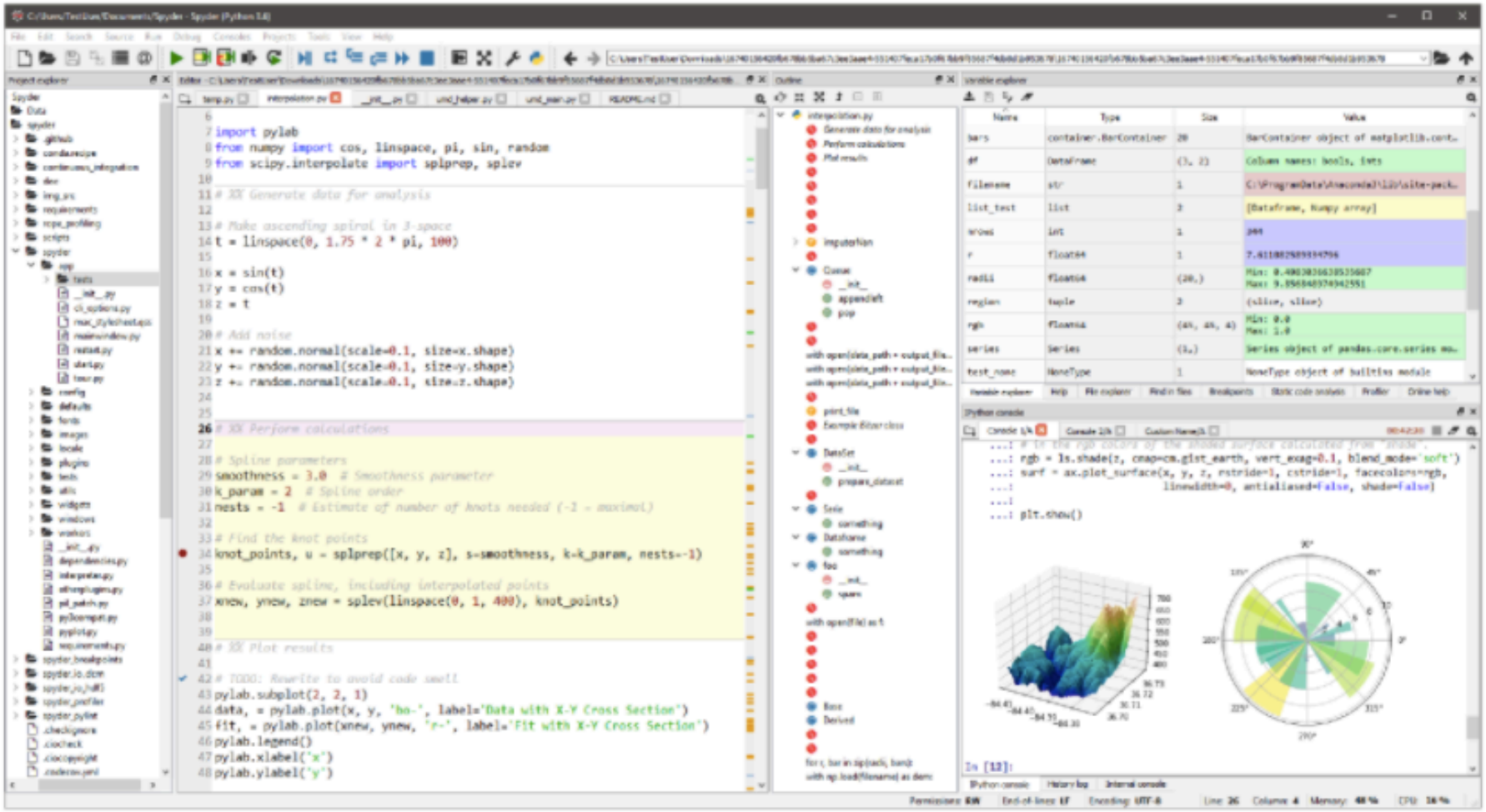
5 features

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

RStudio / IDE para R



Spyder / IDE para Python

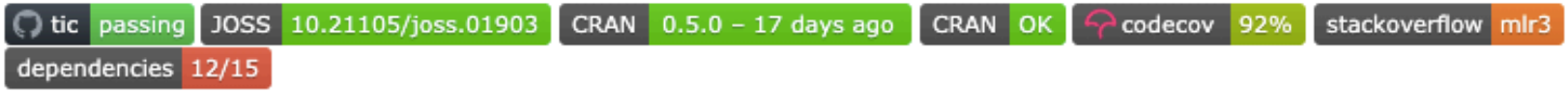


mlr3 / framework em R

mlr3

Package website: release | dev

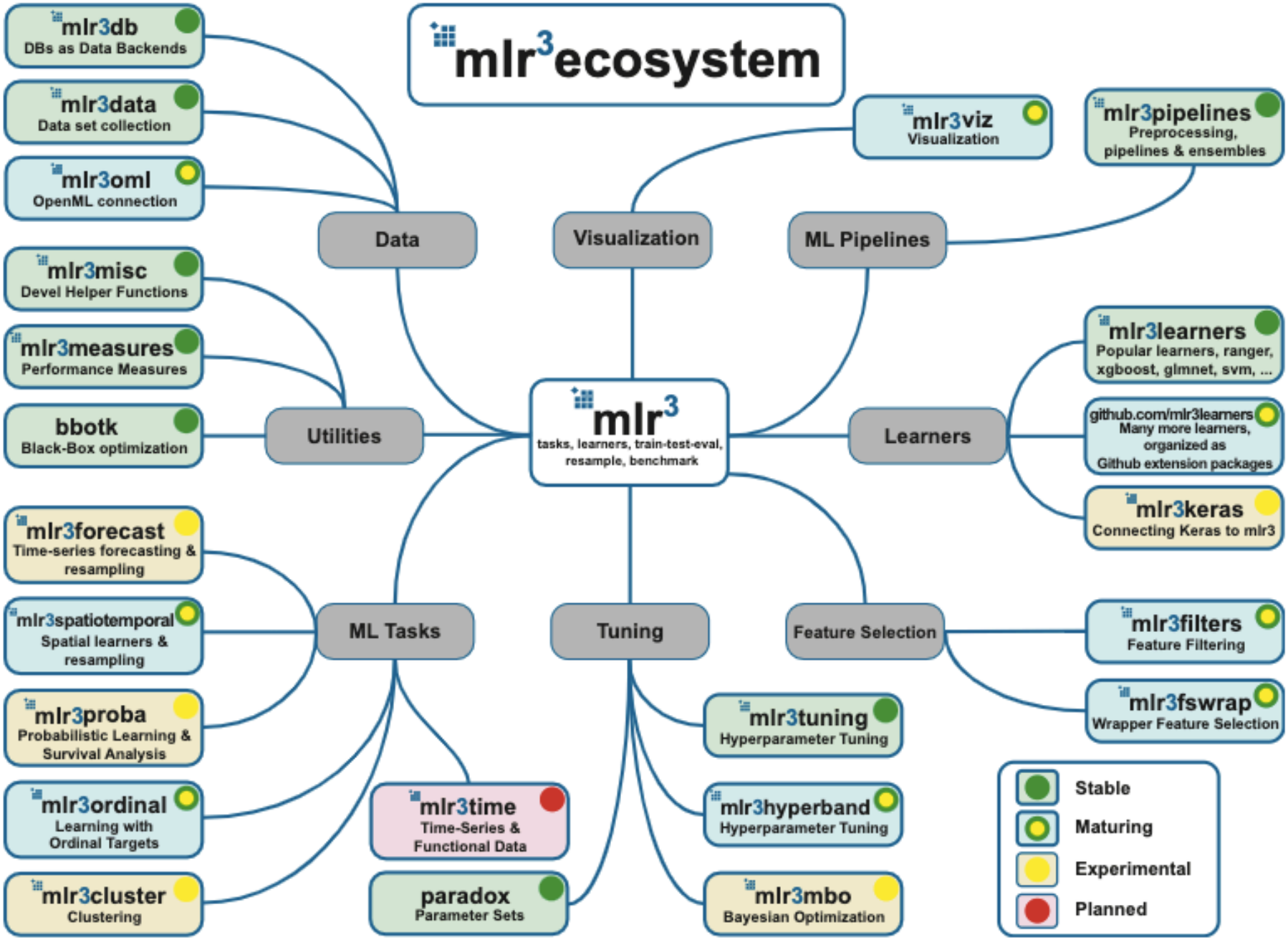
Efficient, object-oriented programming on the building blocks of machine learning. Successor of mlr.



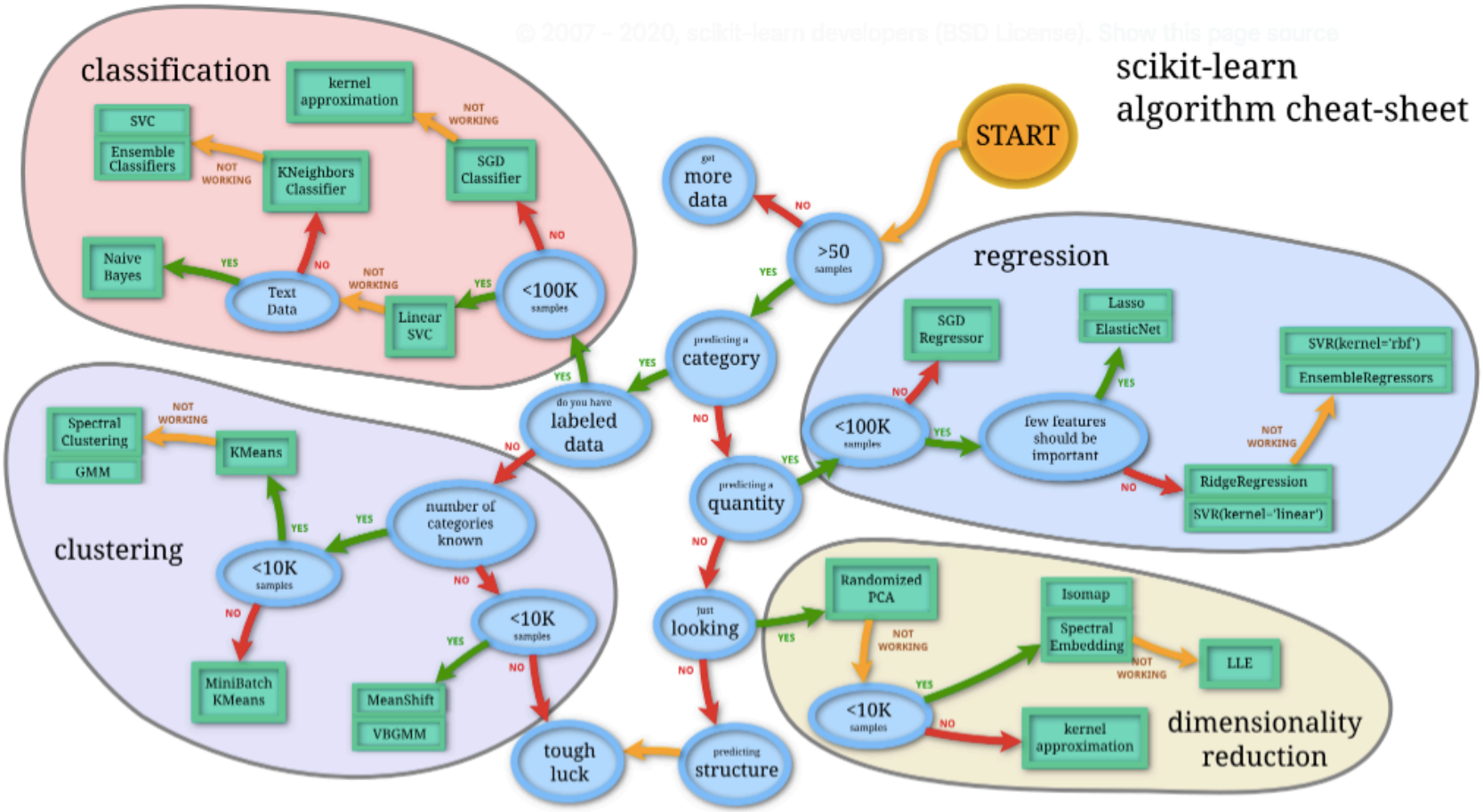
Resources (for users and developers)

- 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

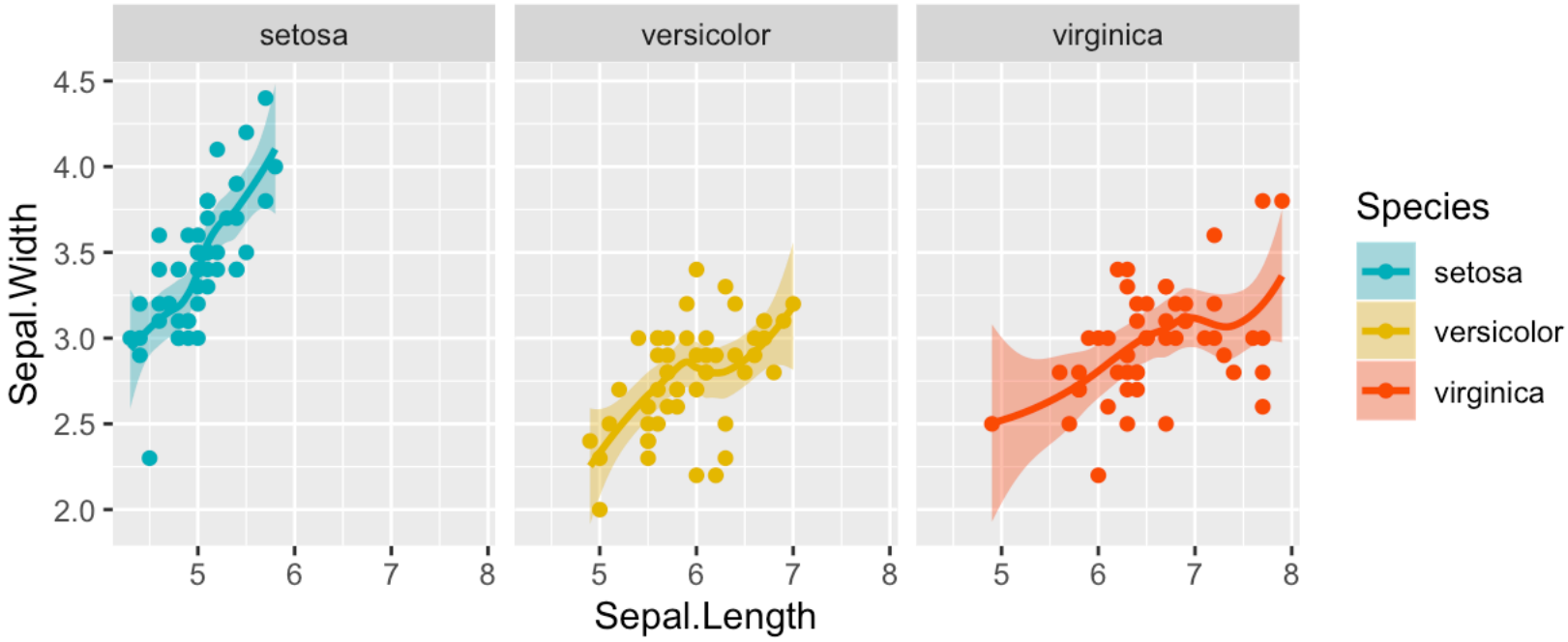


scikit-learn / framework em Python

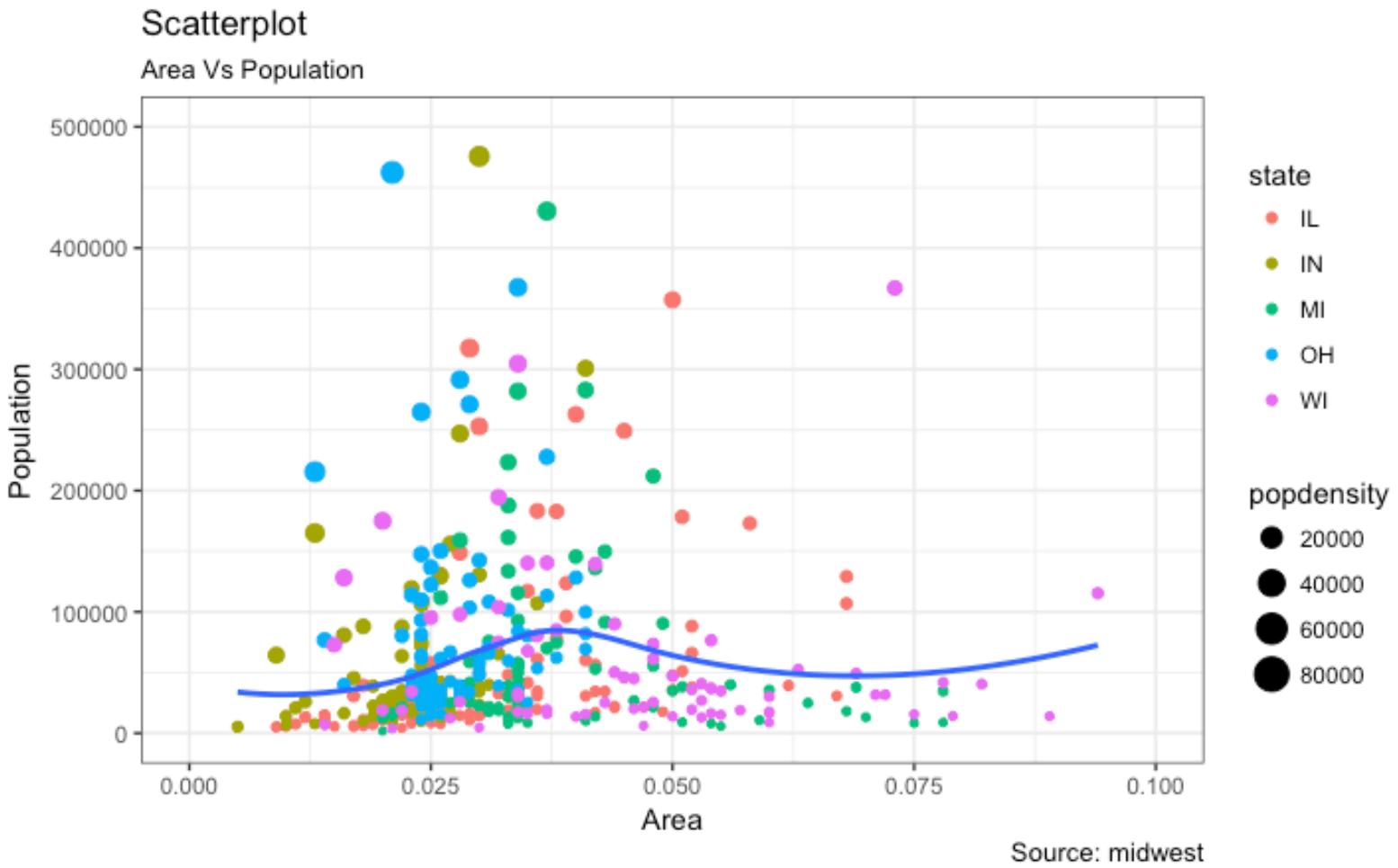


ggplot2

- Visualização dos dados :)



ggplot2


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