

Machine learning em saúde

Prof. Dr. Alexandre Chiavegatto Filho



POPULAR SCIENCE

THE NEW ARTIFICIAL INTELLIGENCE

SPECIAL EDITION



THE FUTURE OF
NANOBOTS

DETECTING
CANCER

INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY

The
Economist

MAY 9TH - 15TH 2015

economist.com

How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves

Artificial Intelligence

The promise and the peril



THE DECLINE OF INTERNATIONAL

FOREIGN AFFAIRS

Hi, Robot

Work and Life in the
Age of Automation

OS DOIS RITMOS DA LAVA-JATO
Curitiba: 107 condenados
Brasília: nenhum

ASSINANTE
ALUNA PEREIRA

veja

MADONNA FALA A VEJA
"Sou uma rebelde e serei
rebelde até o fim"

Edição 2015
Número 2.250 - 27 de setembro de 2015



DE MÃOS DADAS COM A INTELIGÊNCIA ARTIFICIAL

Longe dos cenários futurísticos da ficção científica, ela já faz parte do presente.
Mas em que medida pode servir ao ser humano e, ao mesmo tempo, ameaçá-lo?

The Perils of Special Counsels / This Is Your ISIS on Drugs

Newsweek

08.02.2017



THE DOCTOR WILL SEE YOU NOW

HOW AI IS GOING
TO CURE OUR SICK
HEALTH CARE
SYSTEM

What it takes to end an
AIDS epidemic p. 226

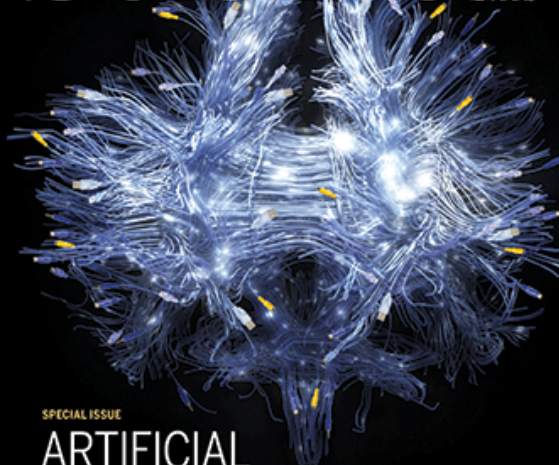
Polar bears suffer through
lean summers p. 205

Sperm produced in ovary
of mutant fish p. 208

Science

\$10
17 JULY 2015
sciencemag.org

AAAS



SPECIAL ISSUE

ARTIFICIAL INTELLIGENCE



Inteligência artificial
NÃO É HYPE CRIADO PELA MÍDIA



É CONSEQUÊNCIA DOS AVANÇOS
CIENTÍFICOS DOS ÚLTIMOS ANOS

Por que têm
ocorridos avanços
exponenciais nos
últimos 5 anos?

Por que têm
ocorridos avanços
exponenciais nos
últimos 5 anos?



Aumento da **quantidade de dados**
(importante para melhorar
performance)

Por que têm
ocorridos avanços
exponenciais nos
últimos 5 anos?



Aumento da **quantidade de dados**
(importante para melhorar
performance)



Avanços em **capacidade
computacional** (modelos de
machine learning exigem muita
memória).

Por que têm ocorridos avanços exponenciais nos últimos 5 anos?



Aumento da **quantidade de dados** (importante para melhorar performance)



Avanços em **capacidade computacional** (modelos de *machine learning* exigem muita memória).



Novos algoritmos para problemas mais complexos (*deep learning*).

The New York Times

Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent

Nearly all big tech companies have an artificial intelligence project, and they are willing to pay experts millions of dollars to help them.

the guardian

'We can't compete': why universities are losing their best AI scientists

A handful of companies are **luring away top researchers**, but academics say they are killing the geese that lay the golden eggs




A demanda por profissionais capazes de extrair informação relevante dos dados é muito grande no Brasil e no mundo



The most in-demand hard skills of 2019

- 1 - Cloud Computing
- 2 - Artificial Intelligence

- 
- A faint, light blue world map is visible in the background of the slide, centered behind the text.
- Se você trabalhar numa empresa que não está de acordo com seus padrões éticos, **saia agora.**
 - *Manipulação de sentimentos / voto.*
 - *Controle de funcionários.*
 - *Limitar acesso de pessoas a bens e serviços.*

A faint, dark blue world map is visible in the background of the slide. The text is centered over the map.

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA

*Melhorar o
mundo*

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas
grandes cidades.

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas
grandes cidades.



Desenvolvimento de
tecnologias verdes

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas
grandes cidades.



Desenvolvimento de
tecnologias verdes



Melhoria da atenção à saúde e
aumento da qualidade de vida



INTELIGÊNCIA ARTIFICIAL

Capacidade de máquinas tomarem decisões inteligentes.

Várias definições para inteligência.

Possibilidade: “capacidade de tomar a melhor decisão possível dada a informação disponível. Com a capacidade de se adaptar a novas situações.”

Segundo essa definição, inteligência é um problema de análise de dados.

MACHINE LEARNING

Inteligência artificial *clássica*

Regras para a tomada de decisão
ensinada por humanos

- Identificar spam via palavras-chave.
- Traduzir uma frase através de dicionário e regras de gramática.
- Identificar caras humanas por meio da forma de nariz, olho, boca etc.

Inteligência artificial com *machine learning*

Máquinas
aprendendo sozinhas!

Tomada de decisão via identificação de
padrões complexos nos dados.

*É como uma
criança aprende!*

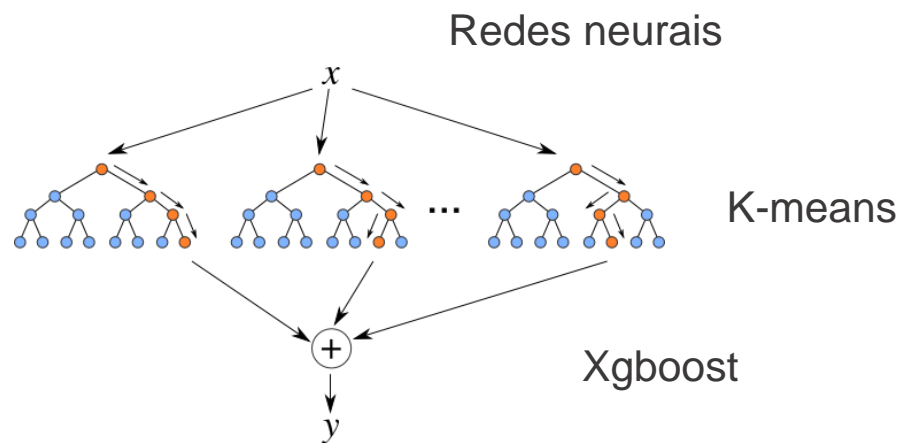


MACHINE LEARNING

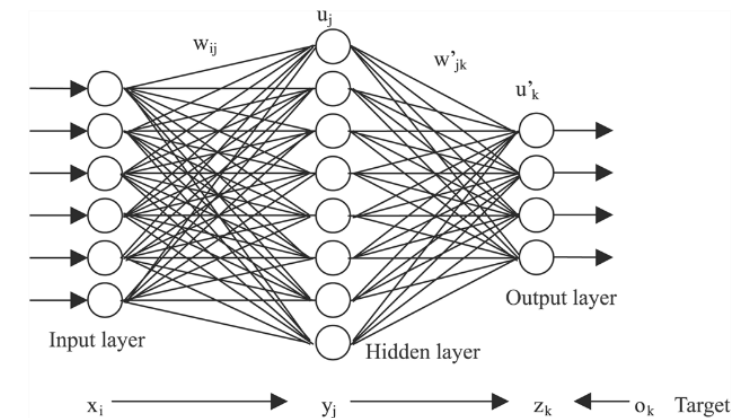
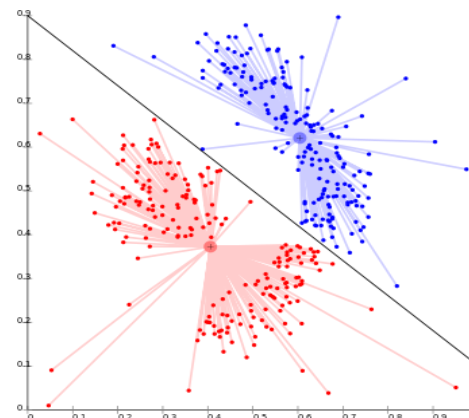
Problemas **práticos** de predição (para a tomada de decisão)

Pouco interesse em *interpretar* os modelos.

Liberdade para modelar a complexidade do mundo real



Florestas randômicas



Regressões penalizadas

How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys.

Kessler RC¹, Rose S, Koenen KC, Karam EG, Stang PE, Stein DJ, Heeringa SG, Hill ED, Liberzon J, McLaughlin KA, McLean SA, Pennell BE, Petukhova M, Rosellini AJ, Ruscio AM, Shahly V, Shalev AY, Silove D, Zaslavsky AM, Angermeyer MC, Bromet EJ, de Almeida JM, de Girolamo G, de Jonge P, Demyttenaere K, Florescu SE, Gureje O, Haro JM, Hinkov H, Kawakami N, Kovess-Masfety V, Lee S, Medina-Mora ME, Murphy SD, Navarro-Mateu F, Piazza M, Posada-Villa J, Scott K, Torres Y, Carmen Viana M.

Predizer presença de transtorno de estresse pós-traumático (TEPT).

- 24 países (incluindo Brasil) → 69.272 indivíduos.
- Prevalência de TEPT: 4,0%.

Variáveis preditoras: sociodemográficas, distúrbios mentais, tipo de evento traumático, histórico de acúmulo de evento traumático.

Algoritmo utilizado: *Super learner*.

- 10% indivíduos com maior risco de TEPT:
- 95,6% do total de casos de TEPT.



Machine Learning for Social Services: A Study of Prenatal Case Management in Illinois

Ian Pan, MA, Laura B. Nolan, PhD, Rashida R. Brown, MPH, Romana Khan, PhD, Paul van der Boor, PhD, Daniel G. Harris, MA, and Rayid Ghani, MS



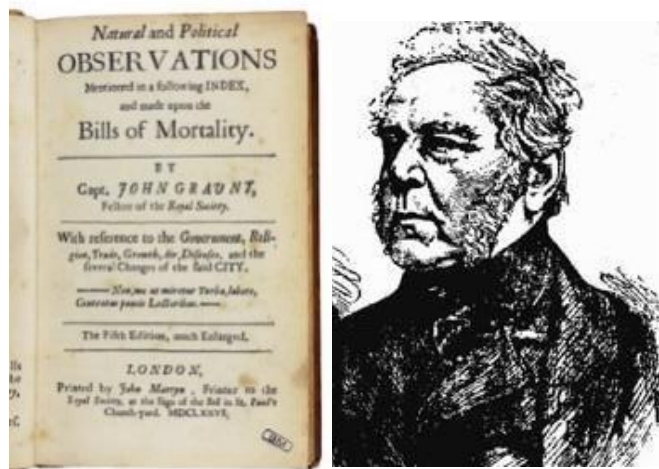
- Predizer quais mulheres grávidas têm maior risco de parto adverso (terem filho prematuro, baixo peso, internação na UTI ou morte no primeiro ano).
 - Objetivo: orientar a inscrição no *Better Birth Outcome*.
 - Seleção atual: presença de 2 fatores de risco (de 17).
 - Quatro algoritmos de machine learning (random forest, linear discriminant analysis, regressão penalizada e naive Bayes) para predizer risco de parto adverso.
- Os 4 algoritmos identificaram melhor mulheres com maior risco de parto adverso.
- A cada 2.000 mulheres, o algoritmo incluiria mais 170 mulheres que teriam parto adverso.

Grandes questões da saúde pública podem ser respondidas com *machine learning*

- É possível prever quem vai morrer em 5 anos (e por qual doença)?
 - É possível prever a qualidade de vida futura de pacientes com doença grave?
 - Identificação de boas políticas públicas em saúde, bons hospitais, bons profissionais.
-

É possível prever quem vai morrer em breve (e por qual causa) ?

Uma das grandes questões da epidemiologia, desde as tábuas de mortalidade de John Graunt (1662).



Problema resolvível:

Resultados promissores em amostras pequenas (tutorial em revisão).

Acesso a um grande estudo longitudinal (10 anos de acompanhamento).

Machine learning to predict
30-day quality-adjusted
survival in critically ill patients
with cancer.

*dos Santos HG, Zampieri FG,
Normilio-Silva K, Cavalcanti AB,
Chiavegatto Filho ACF.*

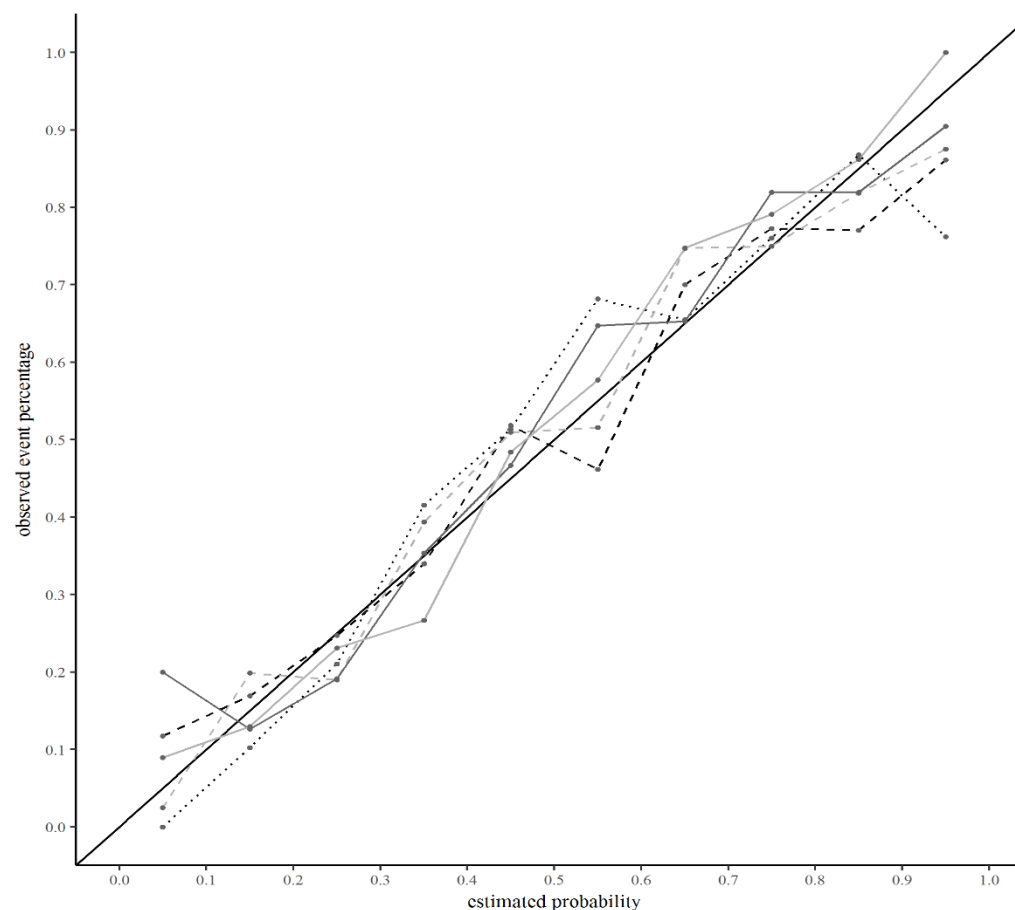
- Predizer < 30 dias QALY em 777 pacientes com câncer internados em UTIs.
 - 27 variáveis: cinco comorbidades (falência renal crônica, falência pulmonar crônica, falência cardíaca crônica, falência cardíaca e diabetes), quatro fatores de risco (IMC, alcoolismo, uso de esteroides e tabagismo), idade, sexo...
 - Seis algoritmos de machine learning: redes neurais artificiais, árvores de decisão, regressão logística e logística penalizada, random forests, gradient boosted trees.
-

Machine learning to predict 30-day quality-adjusted survival in critically ill patients with cancer.

dos Santos HG, Zampieri FG, Normilio-Silva K, Cavalcanti AB, Chiavegatto Filho ACF.

Resultados surpreendentes

É possível prever com muito boa performance quais os pacientes que têm menos de 30 dias de qualidade de vida pela frente (AAC 0,82 para redes neurais, gradient boosted trees e random forests).



Overachieving municipalities in public health: a machine learning approach

Chiavegatto Filho ADP, dos Santos HG, do Nascimento CF, Massa K, Kawachi I.

Publicado na *Epidemiology*

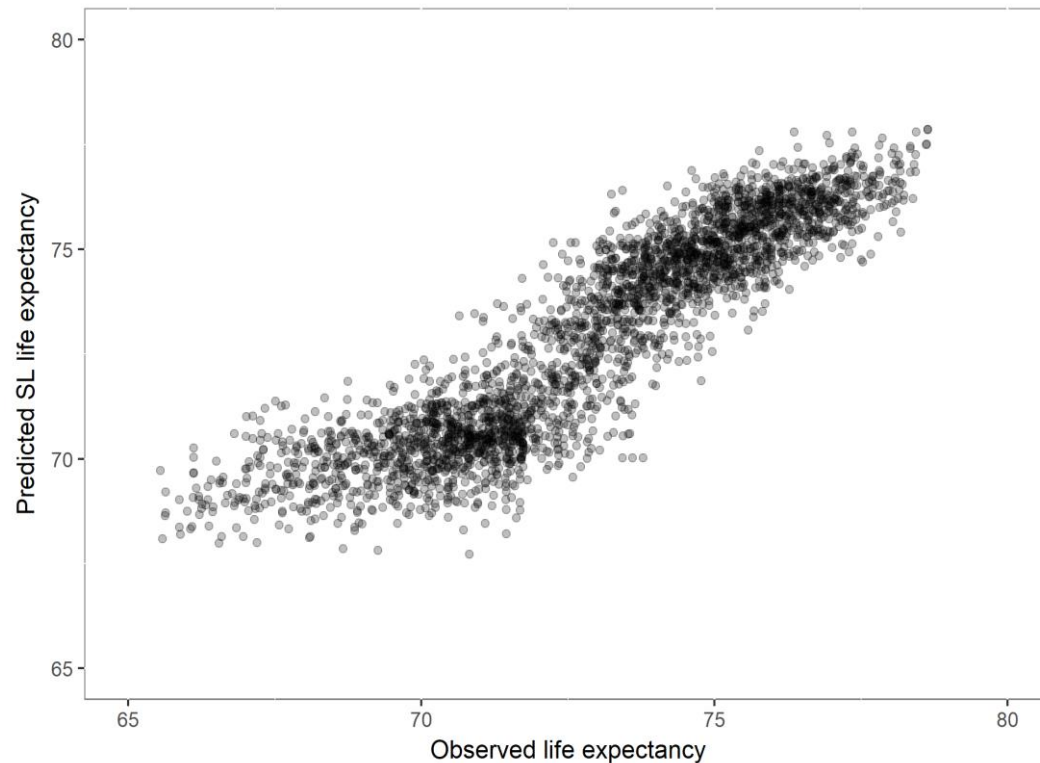
Dificuldade de identificar boas práticas de políticas públicas: importância dos fatores socioeconômicos.

Objetivo

- Predizer a expectativa de vida dos municípios sem usar características de saúde (apenas socioeconômicas e demográficas).
- Identificar *overachievers*: o que eles estão fazendo certo?

Overachieving municipalities in public health: a machine learning approach

Chiavegatto Filho ADP, dos Santos HG, do Nascimento CF, Massa K, Kawachi I.



Resultados:

Nossos modelos predizem bem a expectativa de vida dos municípios brasileiros.

Mesmo assim, há alguns *overachievers*. e alguns *underachievers*.

- O que faz um município ir melhor do que o esperado?
 - Investimento em atenção primária.

Overachieving municipalities in public health: a machine learning approach

Chiavegatto Filho ADP, dos Santos HG, do Nascimento CF, Massa K, Kawachi I.

Predizer para identificar *overachievers*

- Quais hospitais têm uma menor taxa de reinternação (ou de mortalidade ou infecção hospitalar ou...) do que o esperado dado o tipo de pacientes que atendem?
- Quais escolas têm melhor resultados dado o tipo de alunos que estudam lá?

Educação

As cinquenta escolas com os melhores resultados no Enem

Nome da escola (cidade) - rede de ensino - média nas provas objetivas

- 1) Colégio Objetivo Integrado (São Paulo) - privada - 751,29
- 2) Colégio Etapa III (São Paulo) - privada - 736,34
- 3) Colégio Vértice Unidade II (São Paulo) - privada - 710,68
- 4) Colégio Móbile (São Paulo) - privada - 706,69
- 5) Colégio Objetivo Integrado (Mogi das Cruzes) - privada - 690,85
- 6) Colégio Santa Cruz (São Paulo) - privada - 688,12
- 7) Liceu de Artes e Ofícios (São Paulo) - privada - 683,99
- 8) Colégio Bandeirantes (São Paulo) - privada - 683,50
- 9) Colégio Porto Seguro Unidade II (Valinhos) - privada - 681,99
- 10) Colégio Anglo Leonardo da Vinci (Carapicuíba) - privada - 681,77

DESAFIOS PREDITIVOS EM SAÚDE

DESAFIOS PREDITIVOS EM SAÚDE

Predizer para identificar overachievers.

Predizer para descobrir quais desfechos são predizíveis

*Ranking: por exemplo: câncer é
mais predizível que infarto?*

Predizer para comparar o que deveria acontecer vs. o que de fato aconteceu com uma intervenção (projeto cesáreas).

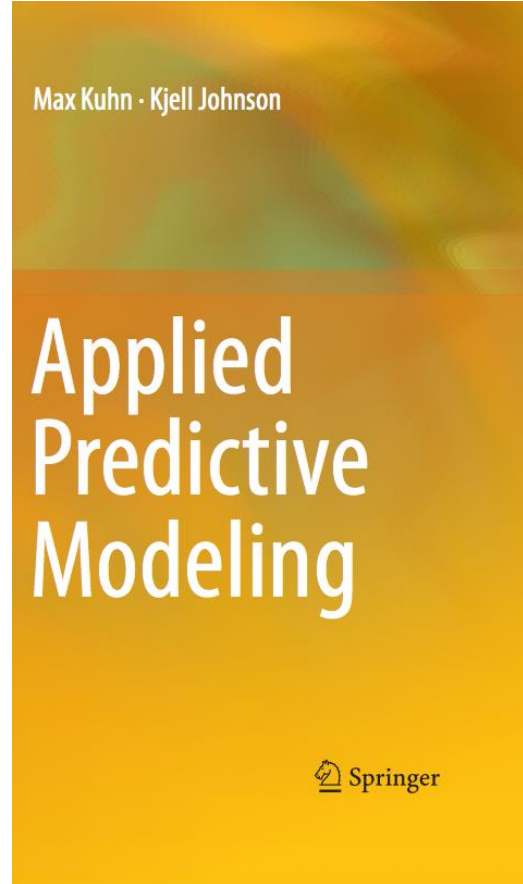
Predizer para comparar com regras de decisão atual (SAPS...).

OUTLINE DO CURSO

- Introdução ao R e ao Python.
- Tipos de modelos machine learning.
- Pré-processamento: seleção de variáveis, vazamento de dados, padronização, redução de dimensão, colinearidade, estratificação, valores missing, one-hot encoding.
- Sobreajuste.
- Medição de performance.
- Regressões penalizadas.
- Mínimos quadrados parciais.
- Support vector machines.
- Redes neurais.
- Árvores de decisão, random forests e gradient boosted trees.
- Deep learning.
- Importância preditora das variáveis.
- Considerações finais.

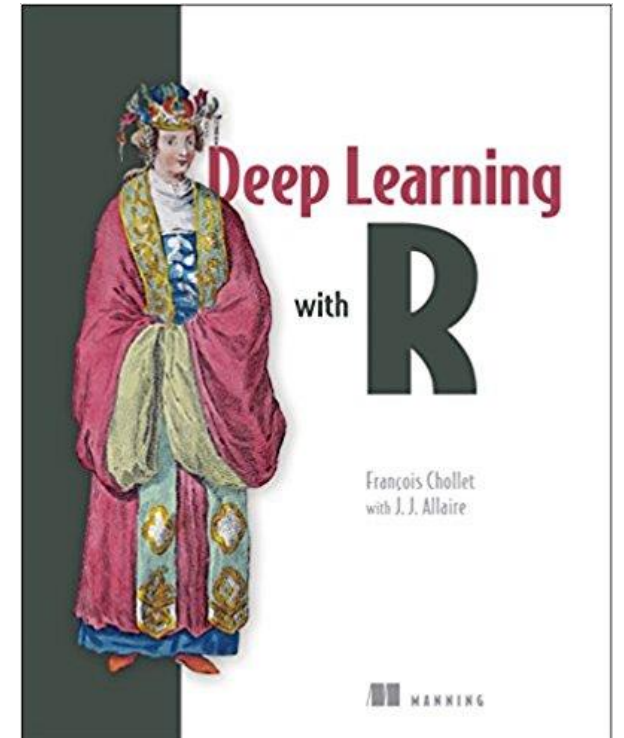


LIVROS-TEXTOS



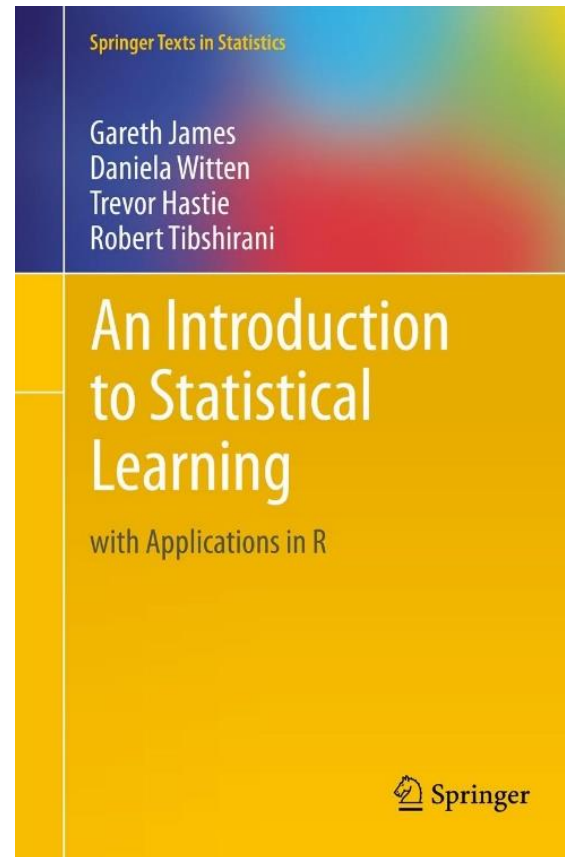
Applied Predictive Modeling (2013)

Deep Learning with R (2018)



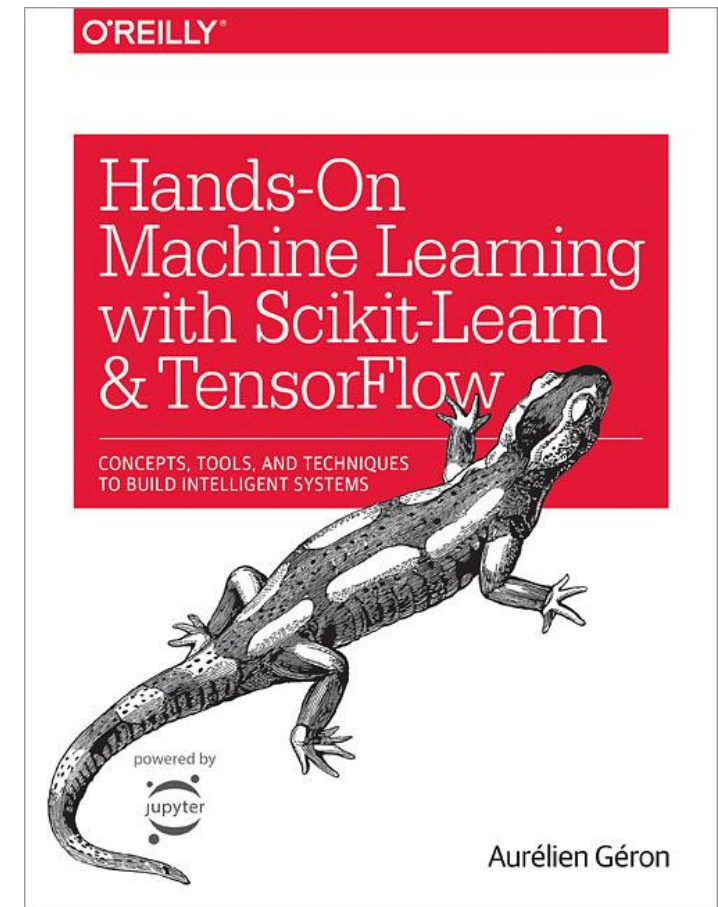


LEITURA SUPLEMENTAR



An Introduction to Statistical
Learning (2013)

Hands-On Machine
Learning (2017)



Sonho

Consenso sobre qual linguagem/software utilizar em ciência de dados!

Evitar ter de aprender várias linguagens de programação.

The STATA logo, featuring the word "STATA" in a bold, blue, sans-serif font.The Python logo, consisting of two interlocking snakes (one blue, one yellow) followed by the word "python" in a lowercase, sans-serif font.The R logo, featuring a stylized "R" inside a grey oval.The JMP logo, with the letters "jmp" in a stylized font and the tagline "Statistical Discovery™ From SAS" below it.The SAS logo, featuring a blue stylized "S" followed by the letters "sas" in a bold, sans-serif font.

Realidade

Forte aumento no uso do R e/ou Python em ciência e nas empresas.

Estamos na torcida para que virem consenso.
Estamos fazendo a nossa parte!



KDNUGGETS ANNUAL POLL 2017

2.900 cientistas de dados

“Quais softwares você usou para
analisar dados nos últimos 12
meses?”

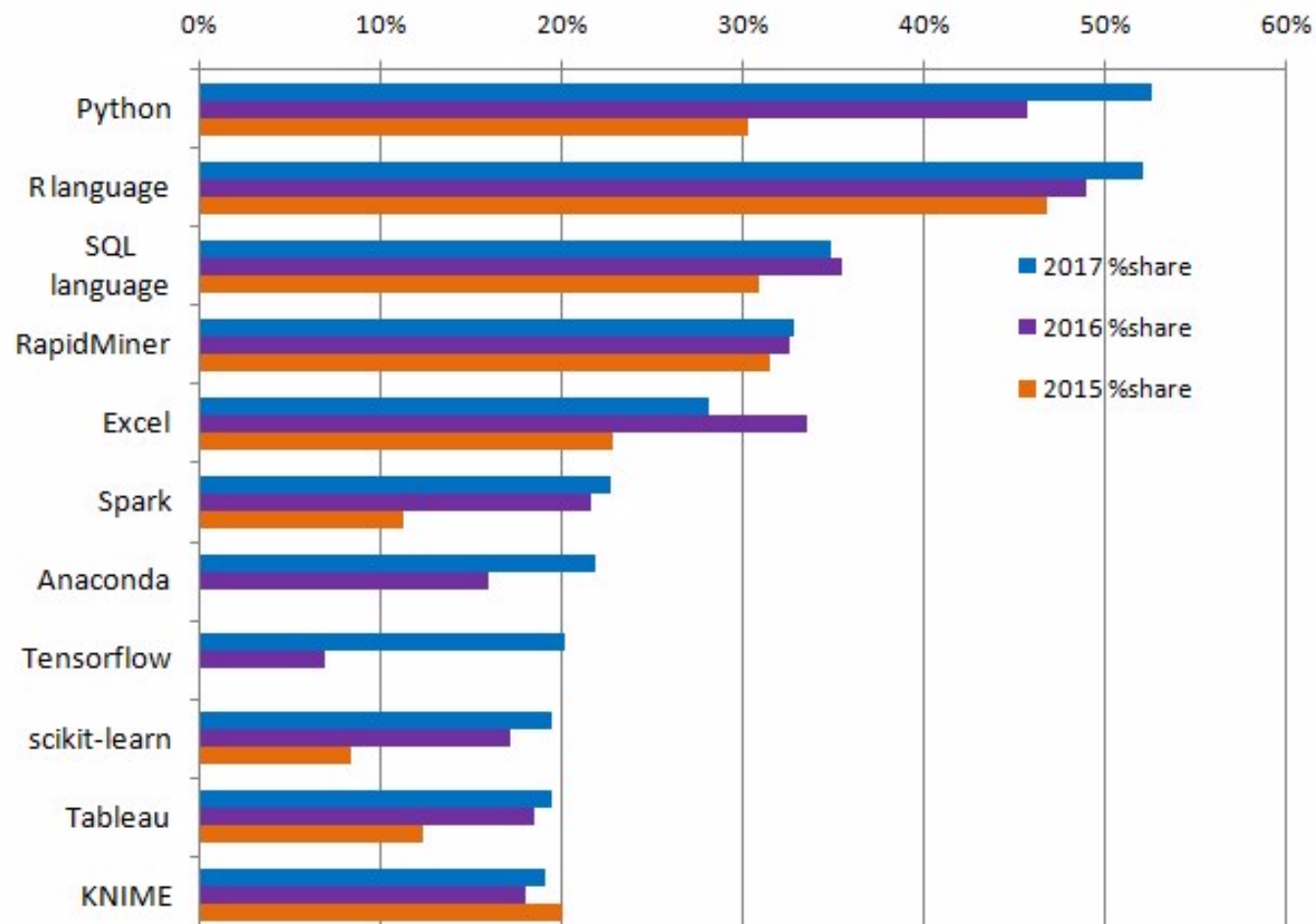


52,1%



python™ 52,6%.

KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017





Linguagem de programação

Muito fácil fazer alterações nas análises
(1 linha de código)

Facilita análises colaborativas

Garante reprodutibilidade dos resultados em
novas amostras

Gratuito

Comunidade ativa de programadores

Foco na análise de dados



Linguagem de programação

Muito fácil fazer alterações nas análises
(1 linha de código)

Facilita análises colaborativas.

Garante reprodutibilidade dos resultados em
novas amostras.

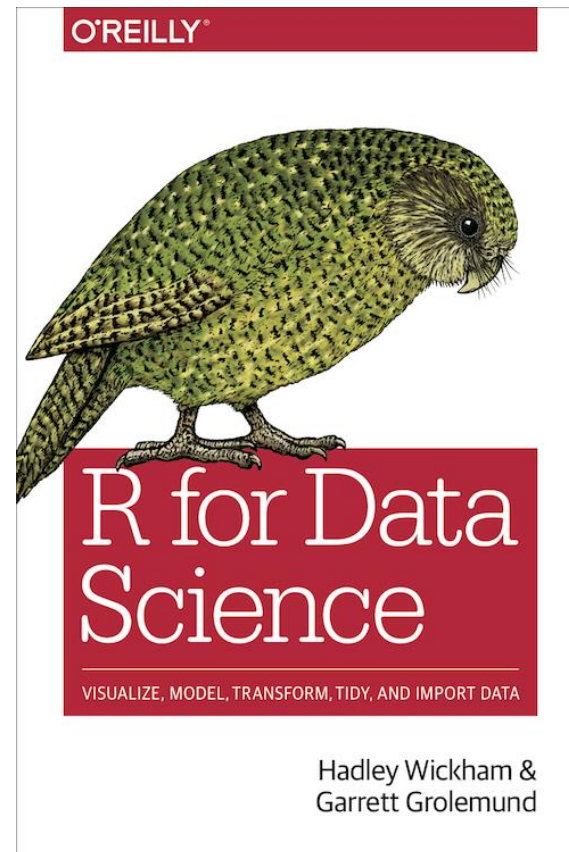
Gratuito

Comunidade ativa de programadores

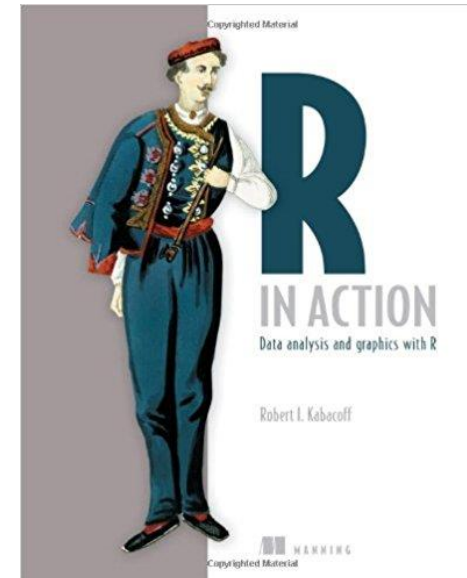
Linguagem de programação geral

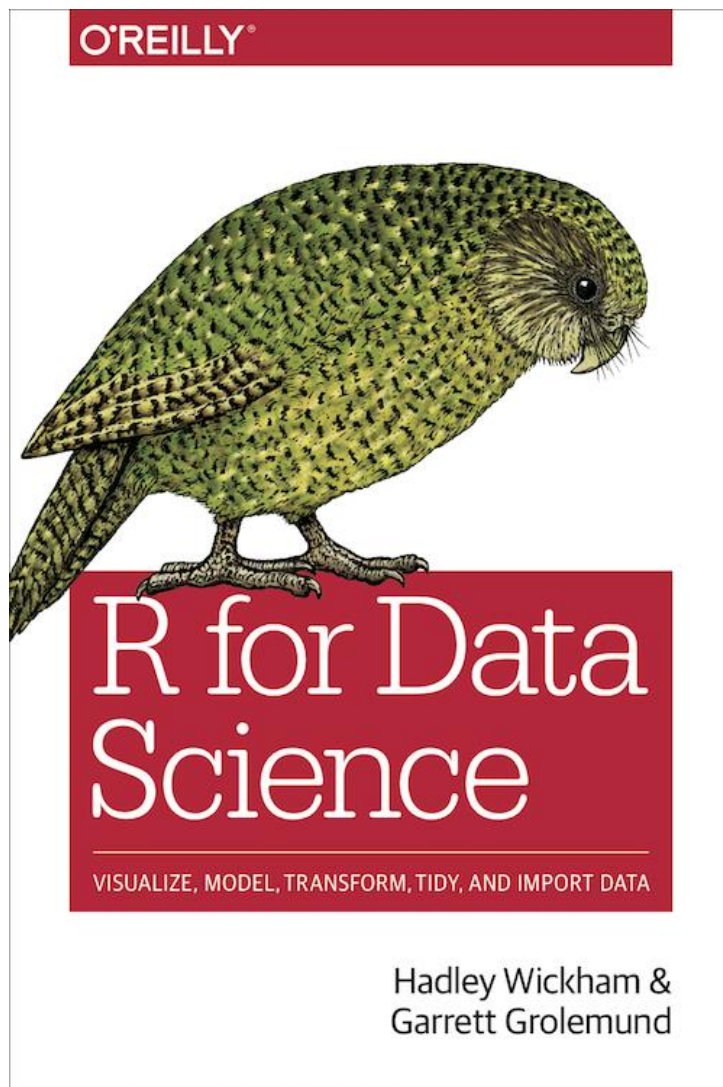


Introdução ao R



Leitura complementar





- Disponível completamente online (recomendo a compra)
- Utiliza o tidyverse (anteriormente conhecido como hadleyverse).



<http://r4ds.had.co.nz>



Hadley Wickham

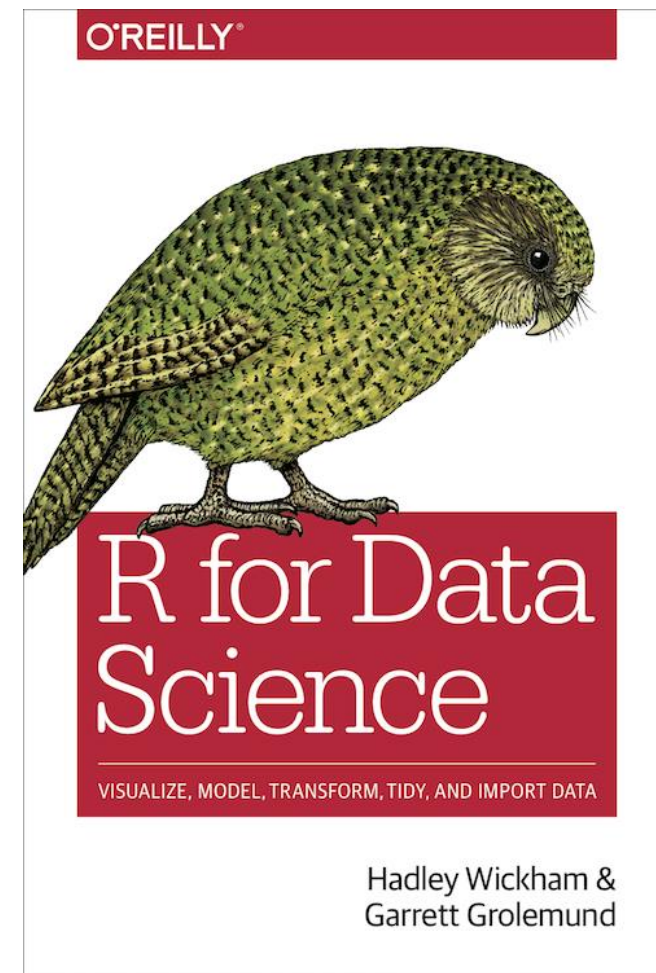


- Tidyverse: conjunto de pacotes que permitem a importação, limpeza e visualização de dados.
- Resolvem um problema importante do R: pacotes com estruturas de comandos diferentes.

```
install.packages("tidyverse")
```

Instala automaticamente:

- readr: importar dados.
- tidyr: limpeza de dados.
- dplyr: manipulação dos dados.
- ggplot2: visualização de dados.
- Entre outros...





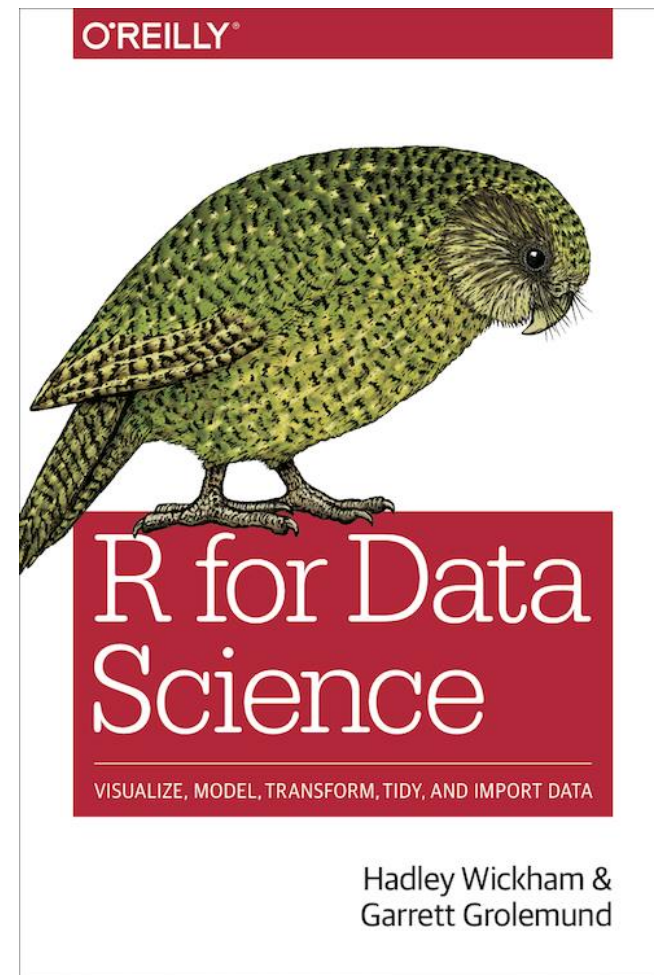
Existia uma grande demanda.

Positiva e rápida absorção pela comunidade de R.

Lançado em
janeiro 2017.

Bestseller na
Amazon em
ciência de
dados.

Cursos da USP:
verão do IME,
veterinária, FSP...



DOWNLOADS



Baixar o R.

<https://cran.r-project.org/>



Baixar o RStudio.

<https://www.rstudio.com/products/rstudio/download/>

```
R Console

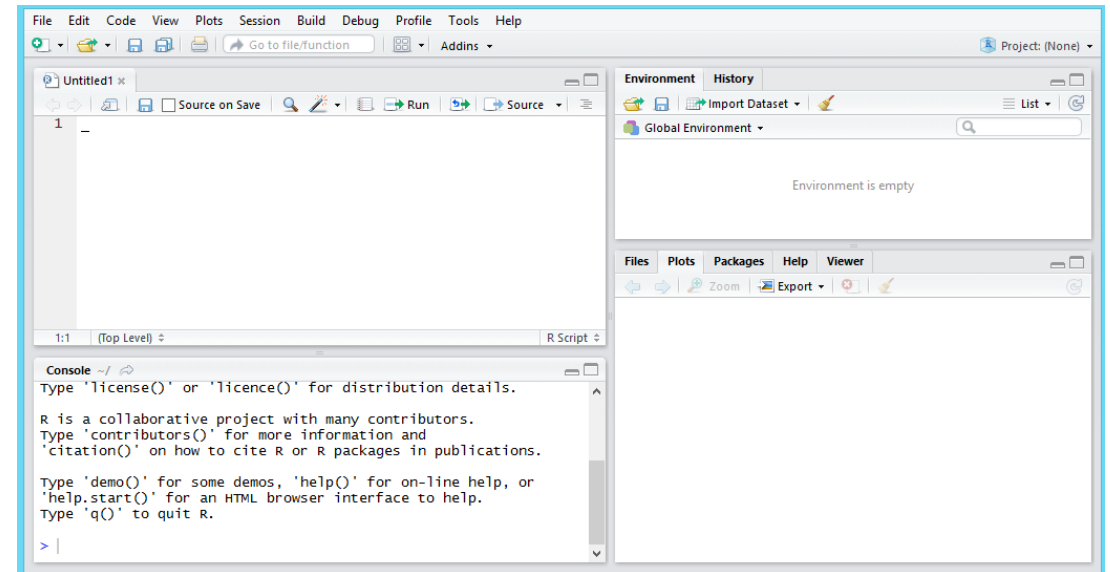
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> |
```





Linguagem de programação.

Software para rodar códigos.

Software composto por uma única janela ("Console").

O software do R é pouco utilizado na prática.

A screenshot of the R Console window. The window has a title bar that says "R Console" and standard Windows window controls. The text inside the console is as follows:

```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)


R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

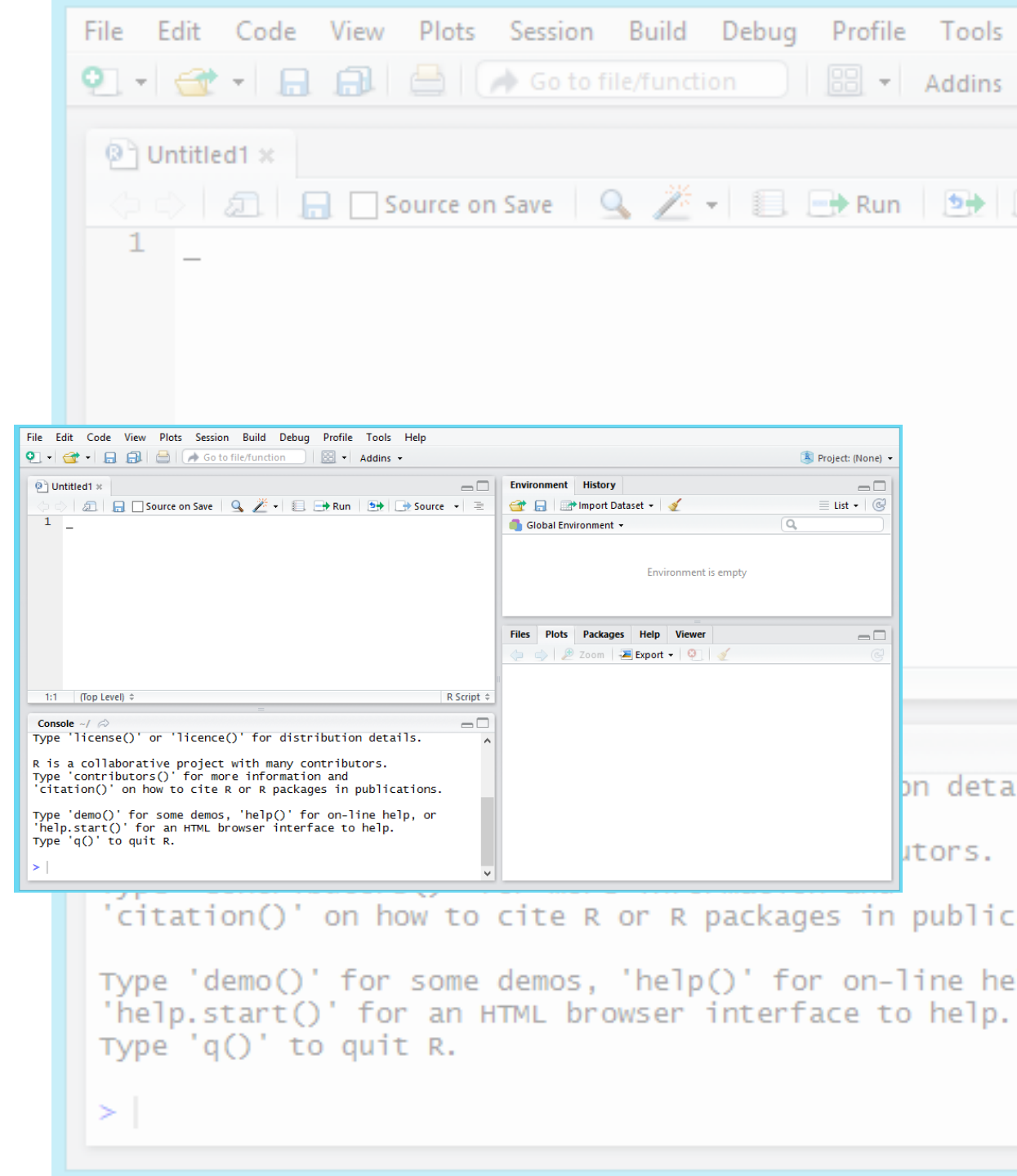
R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> |
```



- É um ambiente de desenvolvimento integrado (IDE, em inglês).
- Software que possibilita um ambiente didático para programar e visualizar resultados.
- Necessita ter o  baixado





Vem apenas com um conjunto básico de pacotes.

Pacote: conjunto de funções, dados e códigos com um objetivo em comum (análise espacial, testes psicológicos...).

É necessário instalar os pacotes específicos de interesse.

R Console

```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> |
```



1

Baixar o pacote (apenas uma vez)

`install.packages("tidyverse")`

2

Chamar o pacote (ao início de toda sessão)

`library(tidyverse)`

Ver todos baixados: `library()`

Ver todos chamados: `(.packages())`



R Console

```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.
```

```
R Console

R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> |
```

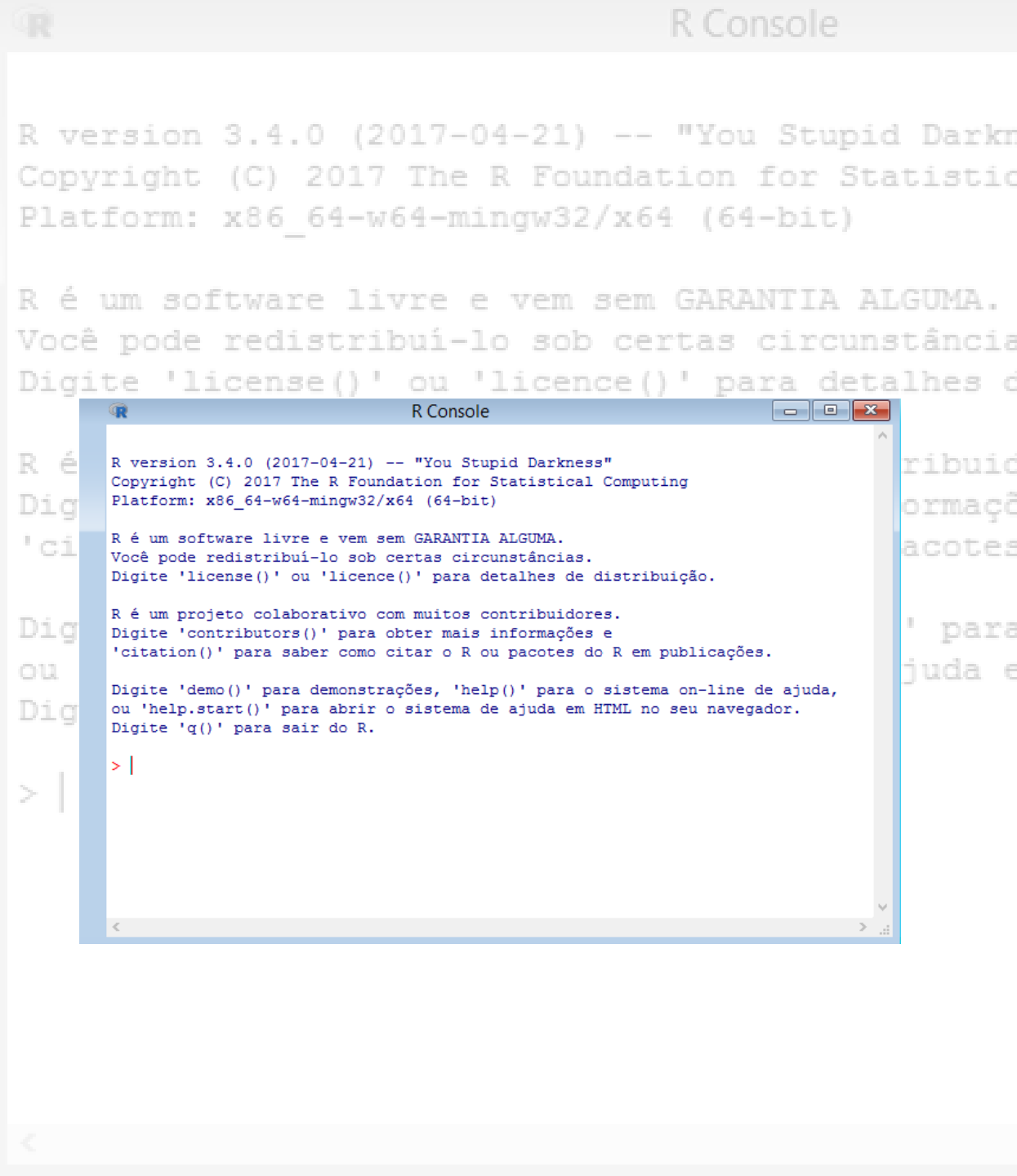


A linguagem do R diferencia maiúscula de minúsculas.

Missing no R: NA

- Stata: .
- MIWin: *

Objetos criados em uma sessão são armazenados apenas temporariamente.





Dataframe: conjunto de vetores com o mesmo número de observações.

- Podem ter vetores de diferentes tipos.
- Equivalente à planilha do Excel.

Listas: não precisam ter o mesmo tamanho.

Função mais simples do R: calculadora.

- Digitar após ">"
- Calcular IMC de pessoa com 89kg e 1,76m: $89 / 1.72^2$
- Se "+" é porque faltou alguma coisa

