# 数据科学中的 R 语言

王敏杰

2020 年 11 月 10 日

四川师范大学

# 开场白

- 非常感谢百智享公司唐总的热情邀请，以及林晗川工程师的细致安排

# 本节课的目的

- R 能给我们生活带来什么？
  - R 是什么?
  - R 能干什么?
  - 为什么是 R？

- 五分钟上手 R 语言
  - 需要一台电脑
  - 注册一个 https://rstudio.cloud 账号
  - 登录 https://rstudio.cloud/project/1847233

# R 是什么

# R 那些事

- 1992 年，新西兰奥克兰大学统计学教授 Ross Ihaka 和 Robert Gentleman，为了方便地给学生教授统计学课程，他们设计开发了 R 语言（他们名字的首字母都是 R）。

Ross Ihaka

Robert Gentleman

# R 是什么

R 语言是用于统计分析，图形表示和报告的编程语言:

- R 是一个统计编程语言（statistical programming）
- R 可运行于多种平台之上，包括 Windows、UNIX 和 Mac OS X
- R 拥有顶尖水准的制图功能
- R 是免费的
- R 应用广泛，拥有丰富的库包
- 活跃的社区(#rstats)

# R 的前世今生

- 2000 年，R1.0.0 发布
- 2004 年，第一届国际 useR! 会议（随后每年举办一次）
- 2005 年，ggplot2 宏包（2018.8 - 2019.8 下载量超过 1.3 亿次）
- 2012 年，R2.15.2 发布
- 2013 年，R3.0.2 发布，CRAN 上的宏包数量 5026 个
- 2016 年，Rstudio 公司推出 tidyverse 宏包（数据科学当前最流行的 R 宏包）
- 2017 年，R3.4.1 发布，CRAN 上的宏包数量 10875 个
- 2019 年，R3.6.1 发布，CRAN 上的宏包数量 15102 个
- 2020 年，R4.0.0 发布，CRAN 上的宏包数量 16054 个

The History of R

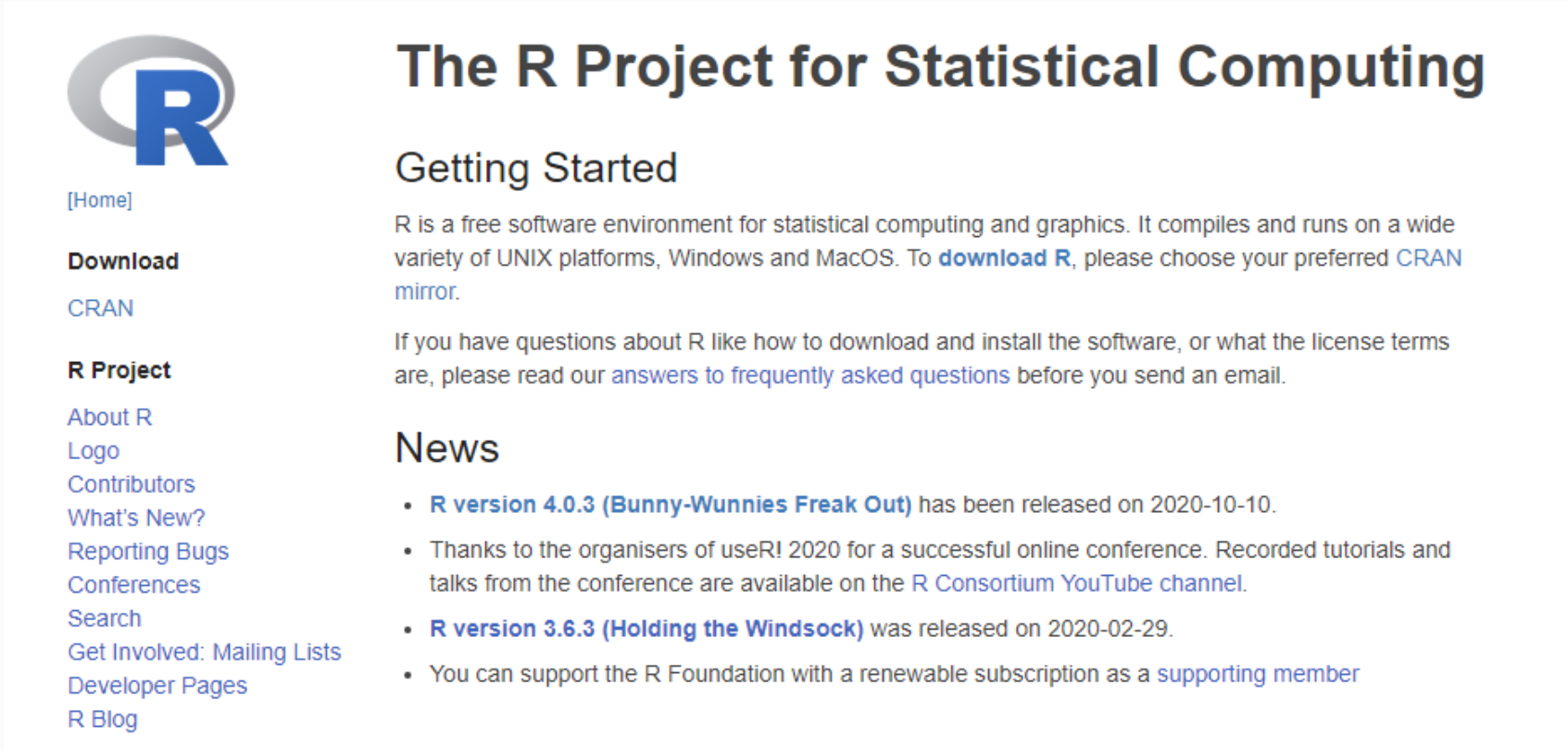
# R 语言发展趋势

| Jul 2020 | Jul 2019 | Change | Programming Language | Ratings | Change |
|---|---|---|---|---|---|
| 1 | 2 | ↑ | C | 16.45% | +2.24% |
| 2 | 1 | ↓ | Java | 15.10% | +0.04% |
| 3 | 3 | | Python | 9.09% | -0.17% |
| 4 | 4 | | C++ | 6.21% | -0.49% |
| 5 | 5 | | C# | 5.25% | +0.88% |
| 6 | 6 | | Visual Basic | 5.23% | +1.03% |
| 7 | 7 | | JavaScript | 2.48% | +0.18% |
| 8 | 20 | ⇈ | R | 2.41% | +1.57% |
| 9 | 8 | ↓ | PHP | 1.90% | -0.27% |
| 10 | 13 | ↑ | Swift | 1.43% | +0.31% |
| 11 | 9 | ↓ | SQL | 1.40% | -0.58% |
| 12 | 16 | ⇈ | Go | 1.21% | +0.19% |
| 13 | 12 | ↓ | Assembly language | 0.94% | -0.45% |
| 14 | 19 | ⇈ | Perl | 0.87% | -0.04% |

TIOBE index

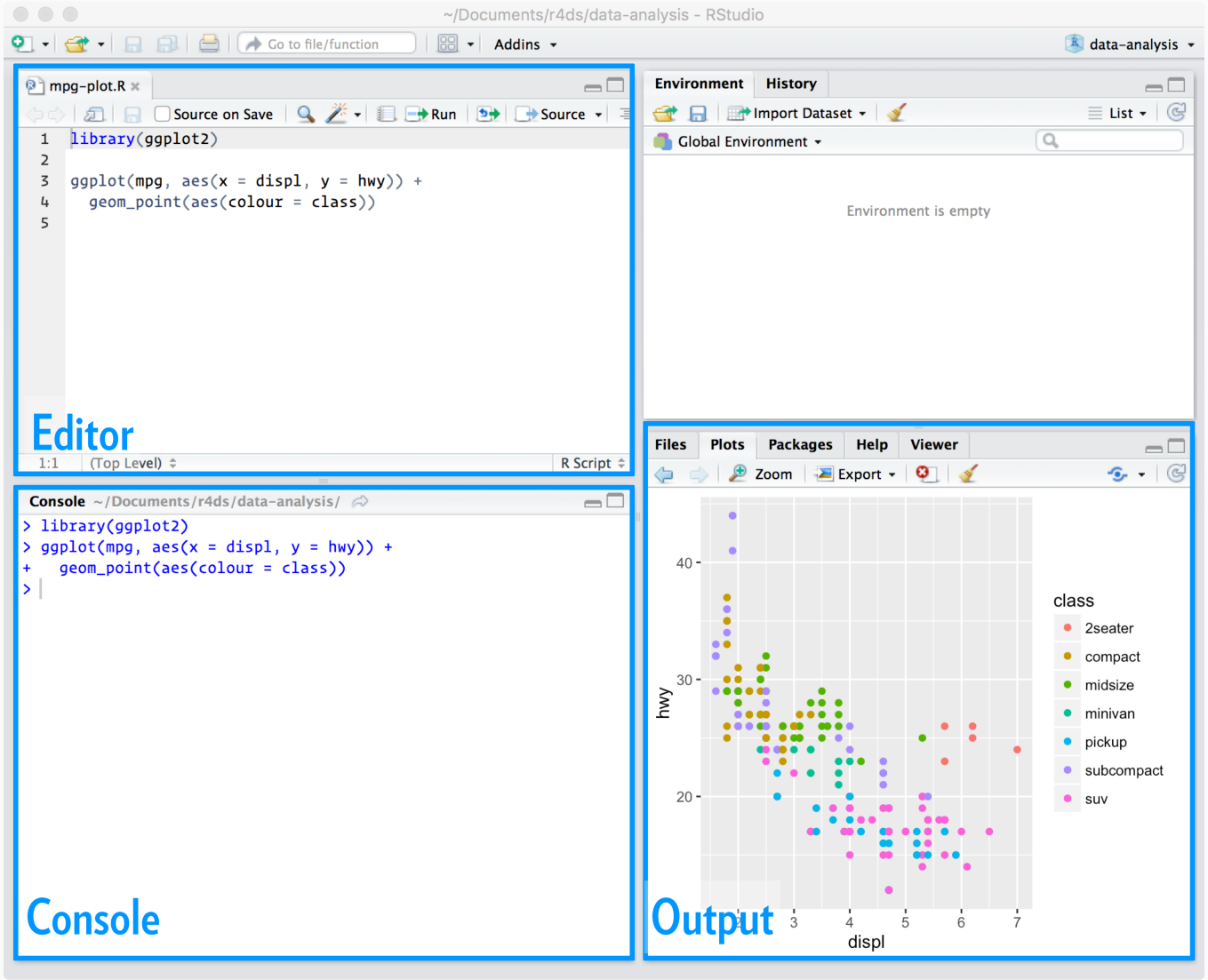
# 安装很方便

官网地址：https://www.r-project.org/



# IDE 很舒服

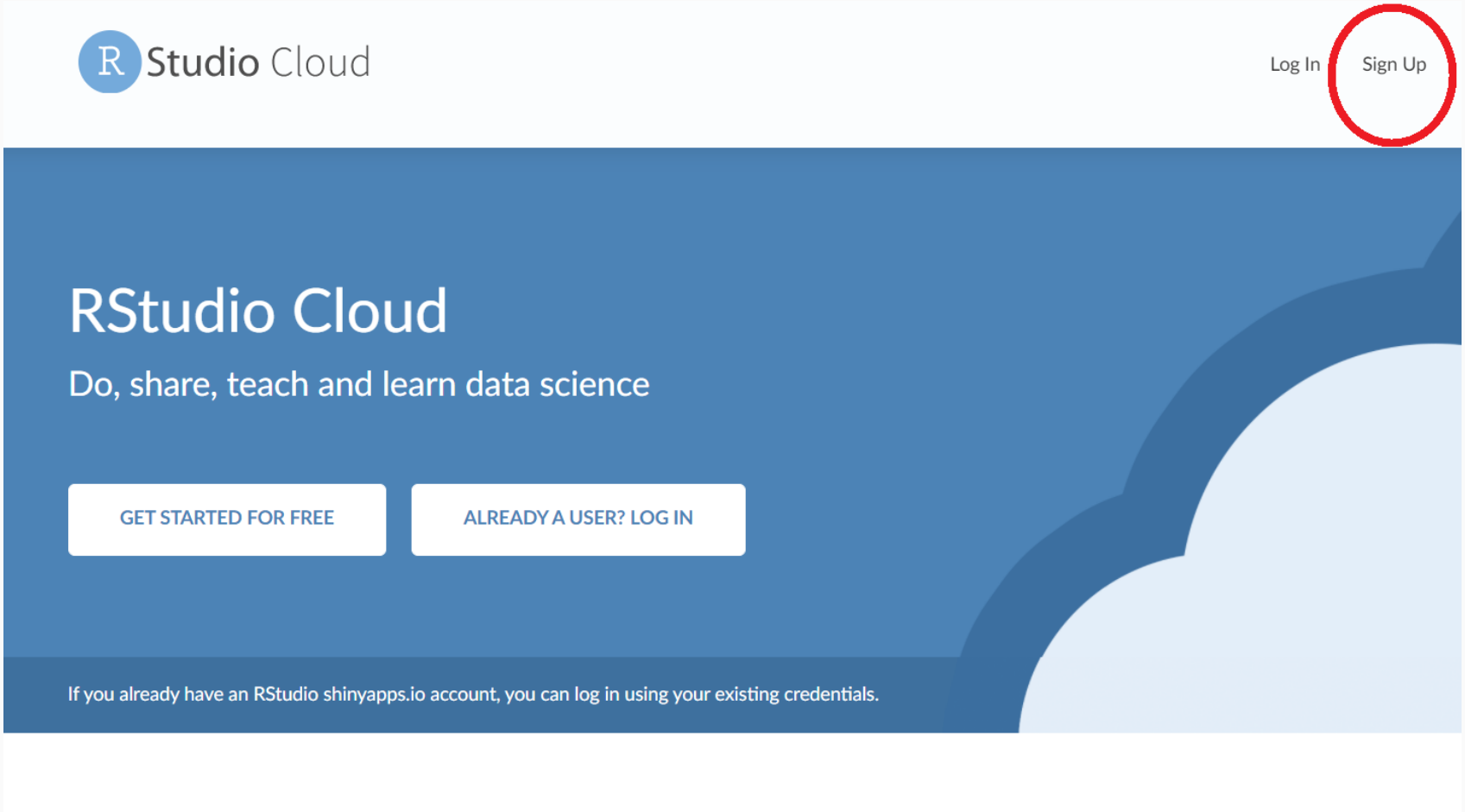
官网地址：https://rstudio.com/



# 两者的关系

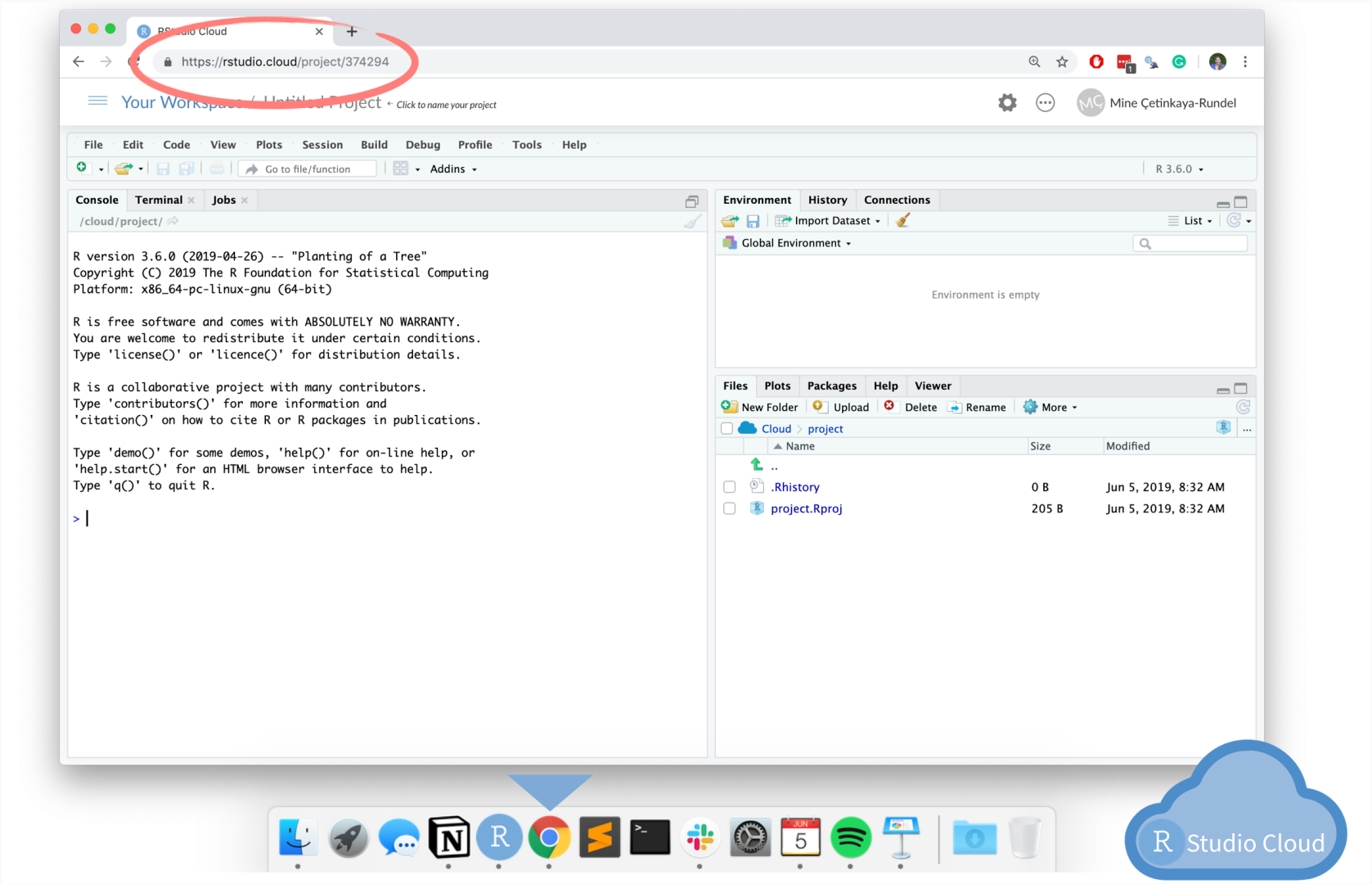
| R: Engine | RStudio: Dashboard |
| --- | --- |
|  |  |

# 也可以偷懒




注册就可使用：`https://rstudio.cloud/`

# 平台很友好


https://rstudio.cloud/project/374294
Click to name your project
Mine Çetinkaya-Rundel
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.6.0
Console Terminal Jobs
/cloud/project/
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Cloud project
Name Size Modified
..
.Rhistory 0 B Jun 5, 2019, 8:32 AM
project.Rproj 205 B Jun 5, 2019, 8:32 AM
Studio Cloud

# R 路上的大神

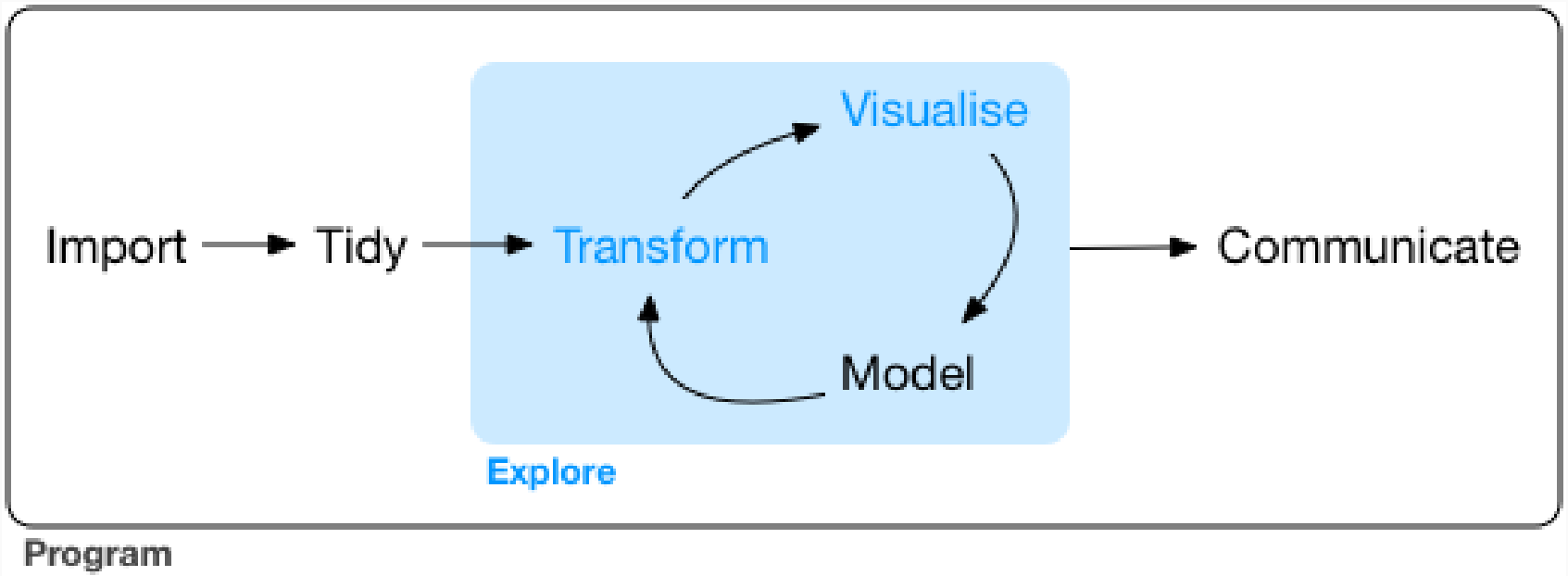
2019 年 8 月，国际统计学年会将考普斯总统奖（被誉为统计学的诺贝尔奖）奖颁给 tidyverse 的作者

- Hadley Wickham
- R 路上的大神
- 一个改变了 R 语言的人

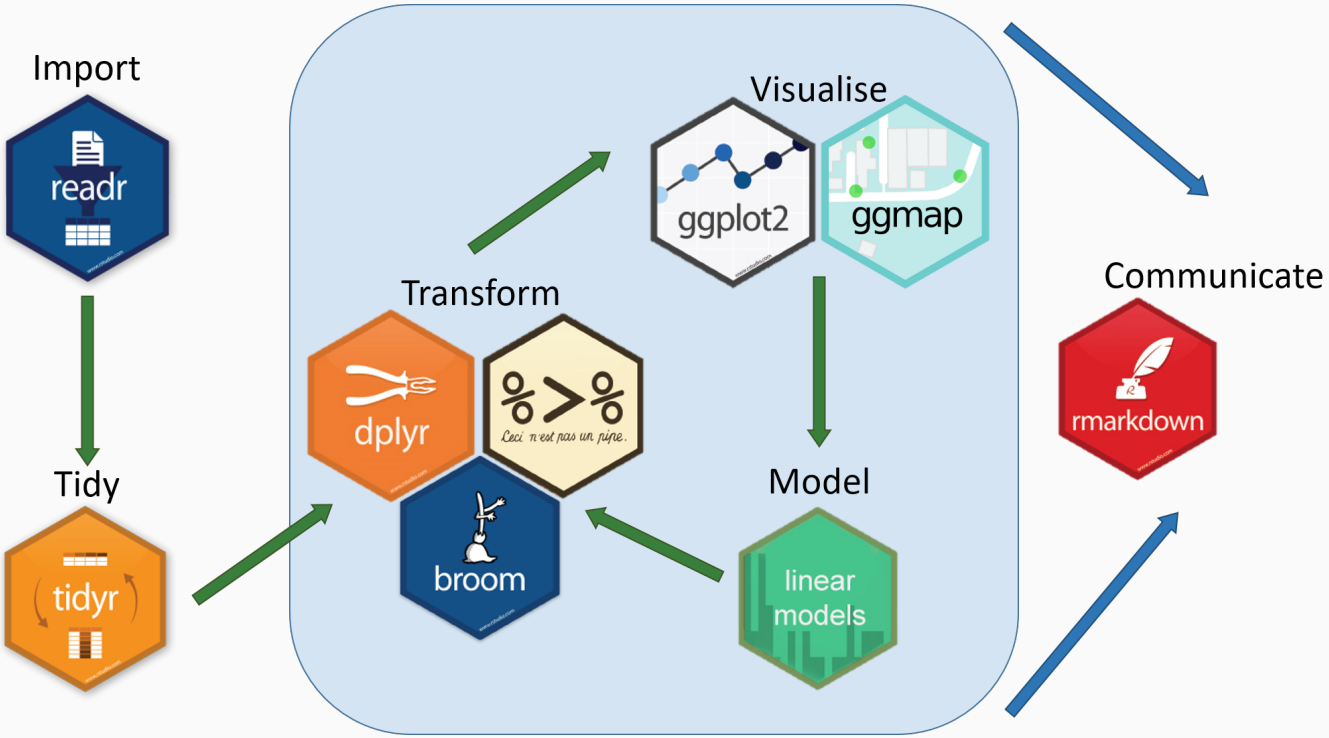
# R 能干什么

# 数据科学的流程

Hadley Wickham 将数据科学流程分解成 6 个环节



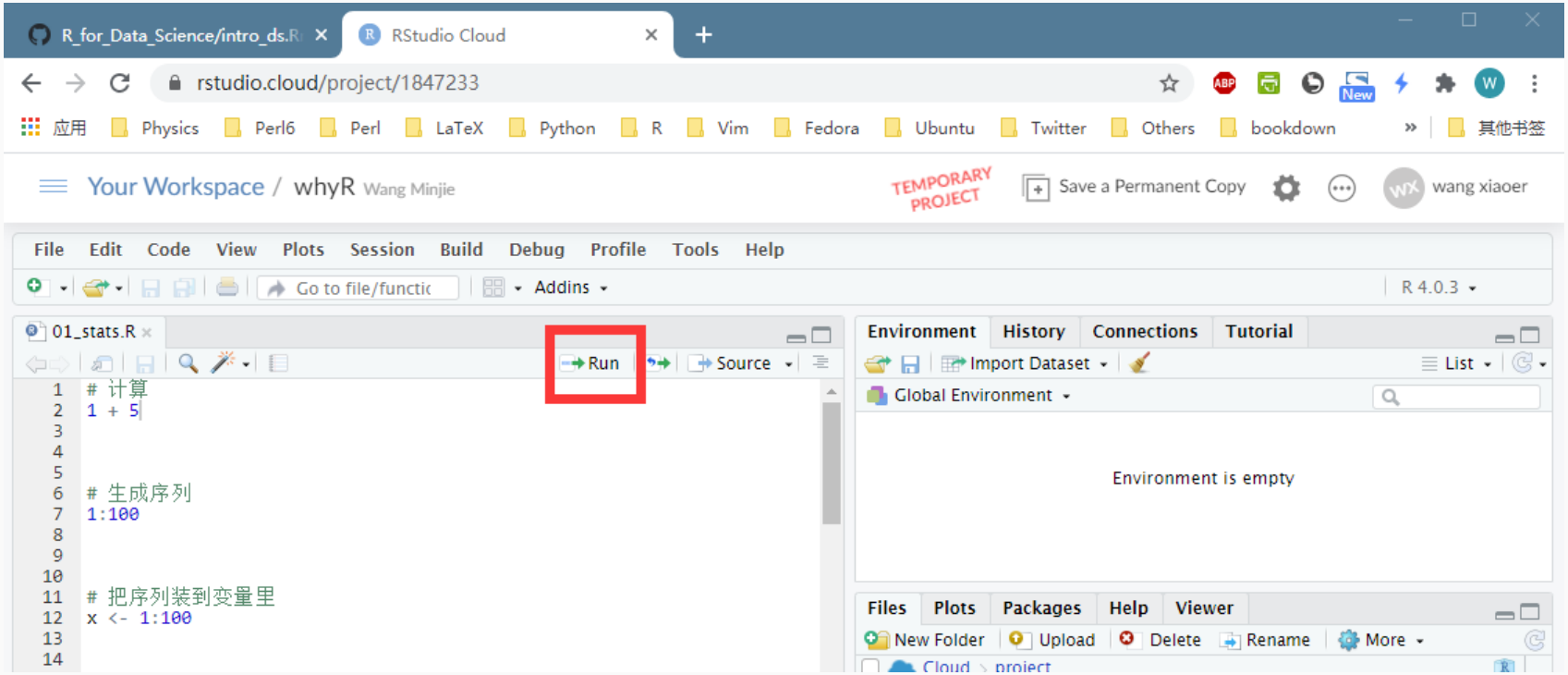
# tidyverse 套餐



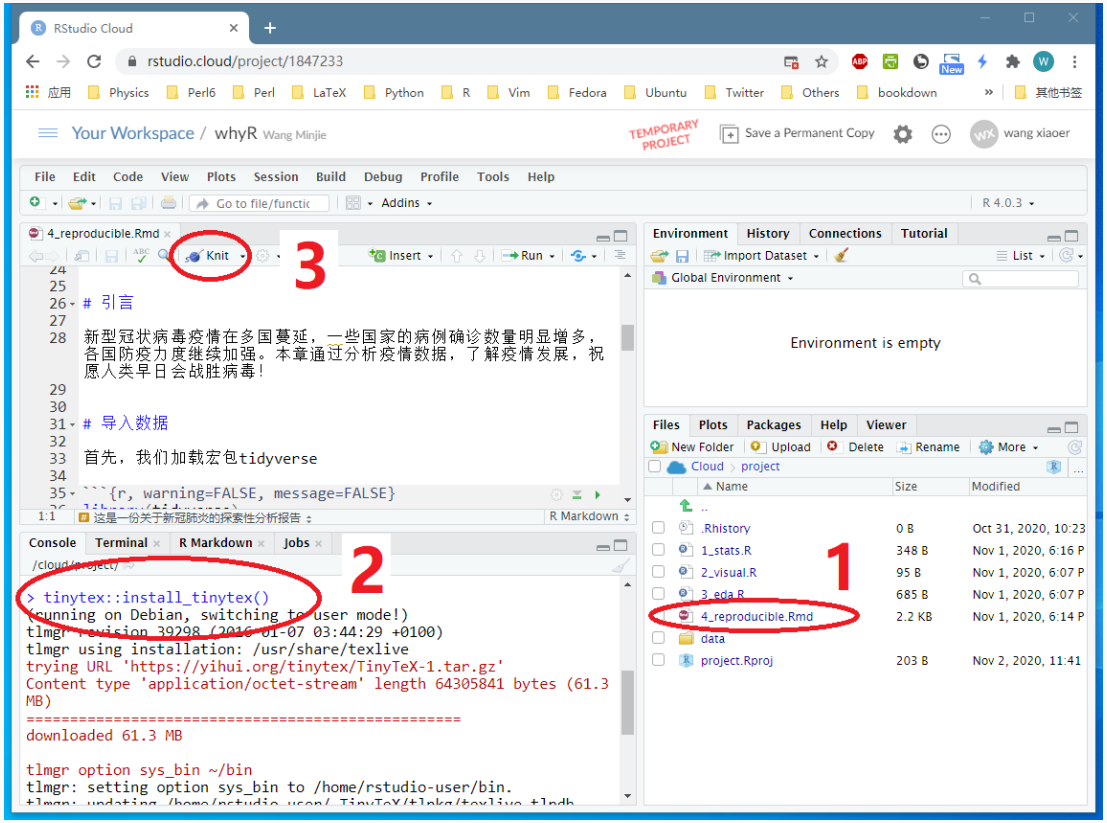

https://www.tidyverse.org/

# 以最快的速度获得一次成功的经验

- 登录 https://rstudio.cloud
- 打开链接 https://rstudio.cloud/project/1847233
- 运行代码（点击右上角的 Run）



# 生成 pdf




- Console 中输入 `tinytex::install_tinytex()` 回车
- 等待 2 分钟，然后重新点击 `Knit` 就可以看到 pdf 了

# 难吗？

感觉很难吗？

如果是，那说明你认真听了

# 看了这些代码，可能第一眼感觉是这样的

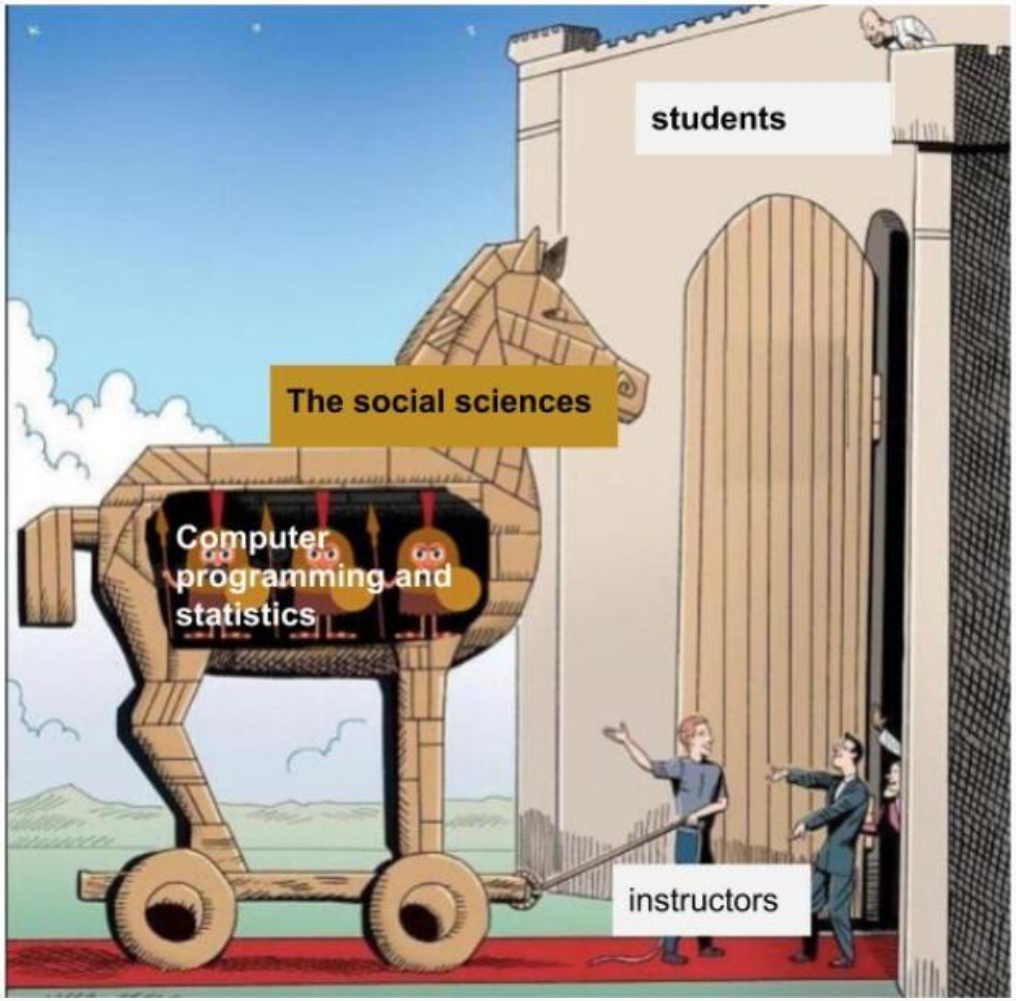
图 1: 图片来自电影《降临》

# 但我更希望学完后

图 2: 图片来自美剧《权利的游戏》

# 为什么是 R

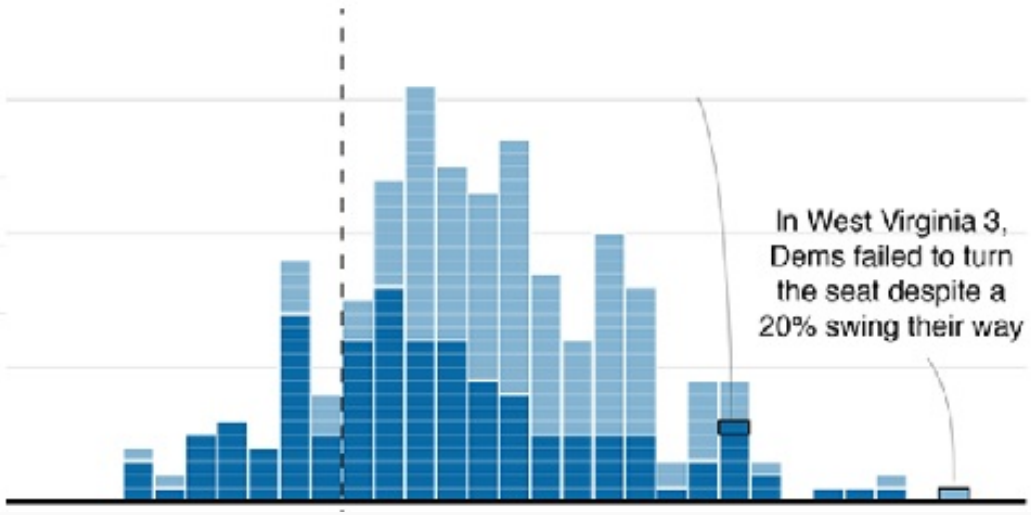
# 社会科学需要统计




往往用的统计的，都不是学统计的

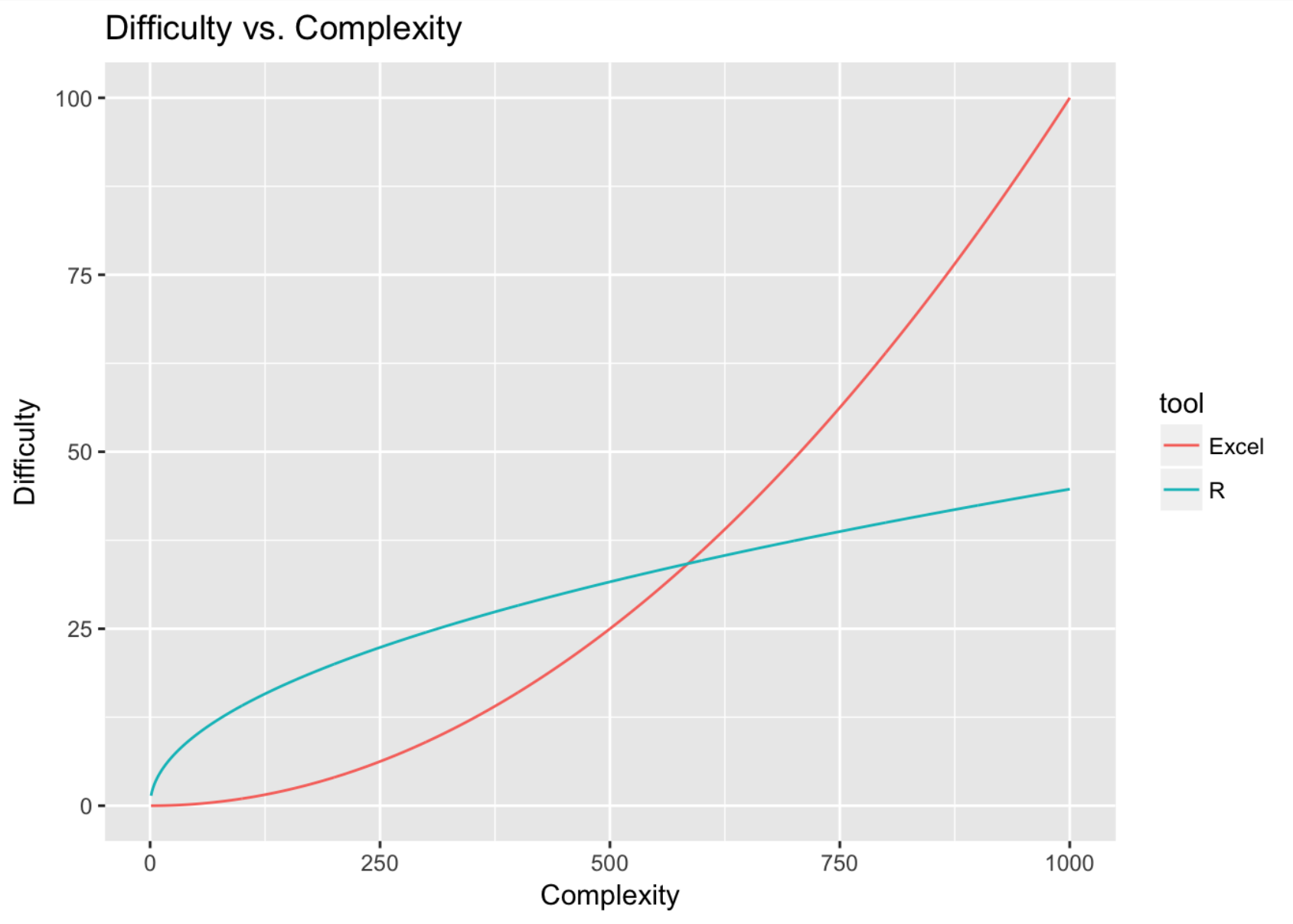
# 社会科学需要可视化


| | table.Q1b.-Awareness - % | table.Q4.Brand-.Attitude.T2B - % | table.Q3.Pre-ferred.cola - % |
|---|---|---|---|
| Coca-Cola | 100.0 | 73.8 | 42.8 |
| Diet Coke | 100.0 | 34.5 | 10.3 |
| Coke Zero | 92.7 | 42.7 | 17.3 |
| Pepsi | 100.0 | 45.2 | 9.2 |
| Diet Pepsi | 82.3 | 17.5 | 2.7 |
| Pepsi Max | 91.0 | 40.7 | 15.7 |
| NET | 100.0 | 96.0 | 100.0 |

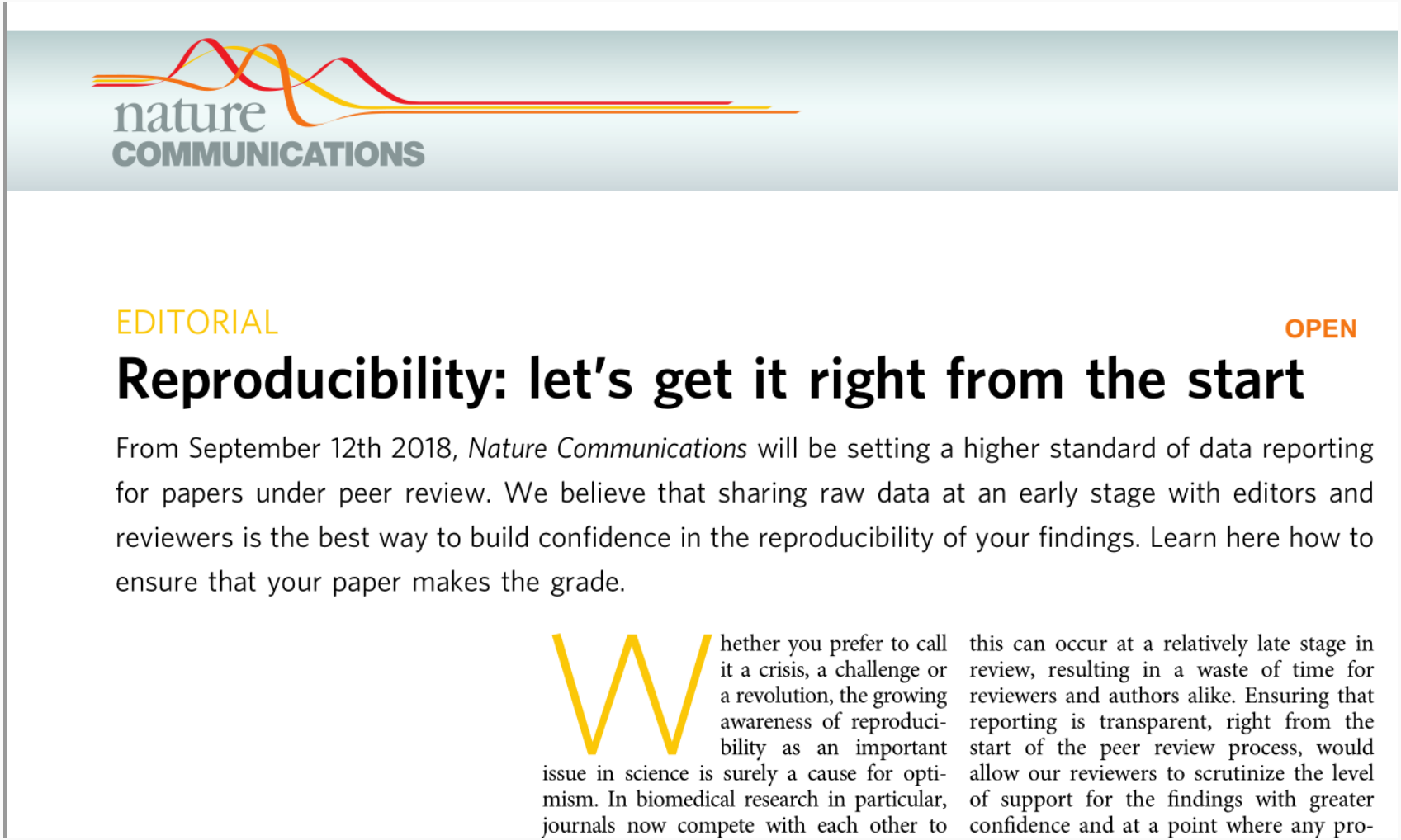



没有好看的皮囊，没人愿意了解你的灵魂

# 社会科学需要编程


Difficulty vs. Complexity
Difficulty
100
75
50
25
0
0
250
500
750
1000
Complexity
tool
Excel
R

# 社会科学需要可重复性



nature COMMUNICATIONS

EDITORIAL

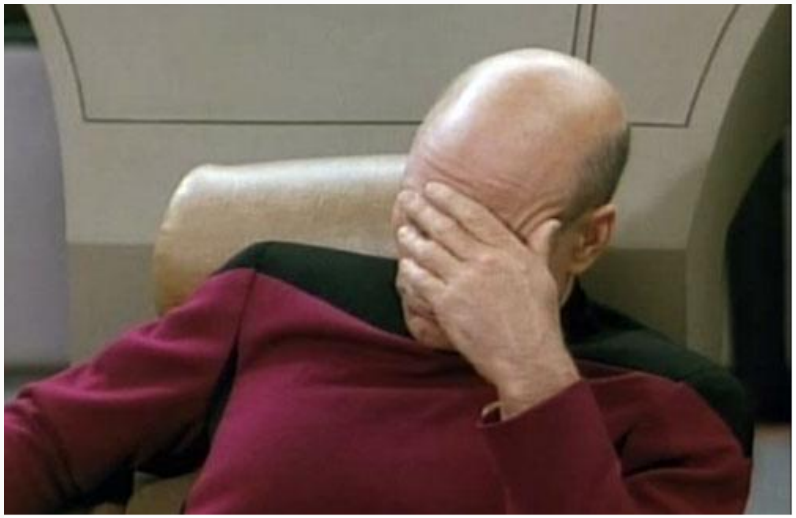
OPEN

## Reproducibility: let's get it right from the start

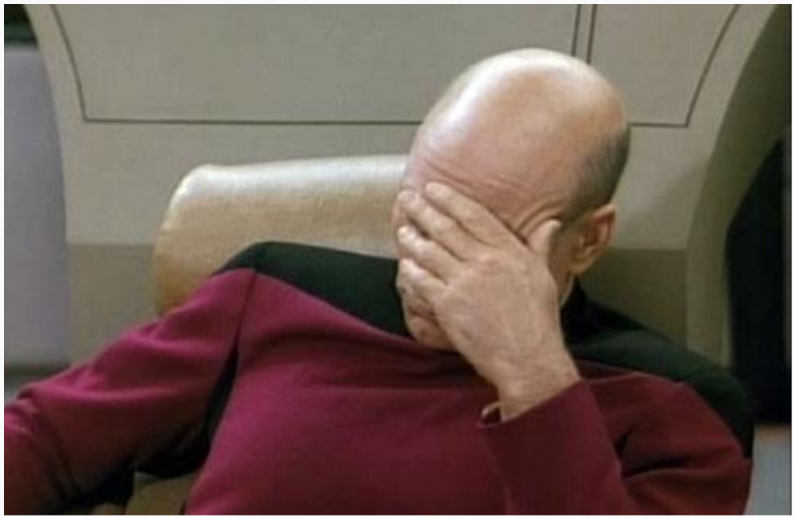
From September 12th 2018, *Nature Communications* will be setting a higher standard of data reporting for papers under peer review. We believe that sharing raw data at an early stage with editors and reviewers is the best way to build confidence in the reproducibility of your findings. Learn here how to ensure that your paper makes the grade.

Whether you prefer to call it a crisis, a challenge or a revolution, the growing awareness of reproducibility as an important issue in science is surely a cause for optimism. In biomedical research in particular, journals now compete with each other to this can occur at a relatively late stage in review, resulting in a waste of time for reviewers and authors alike. Ensuring that reporting is transparent, right from the start of the peer review process, would allow our reviewers to scrutinize the level of support for the findings with greater confidence and at a point where any pro-

论文，要公布原始数据和如何分析的代码

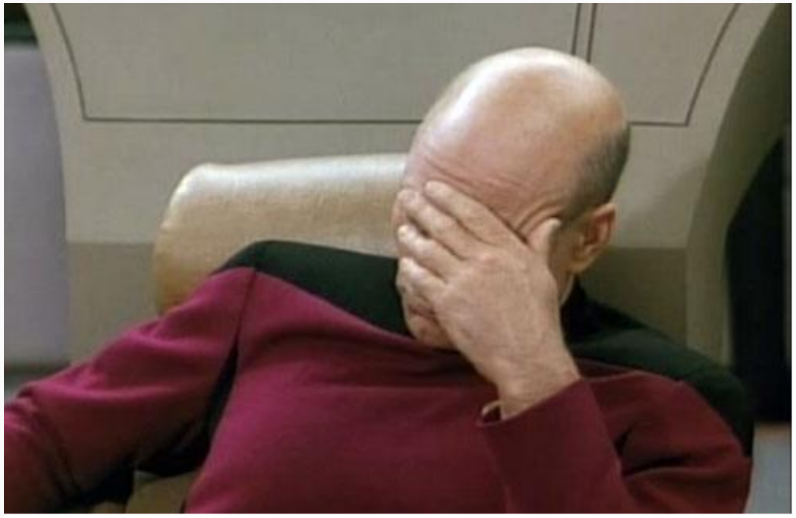


# 残酷的现实



# 残酷的现实

何以解忧，唯有撸 R



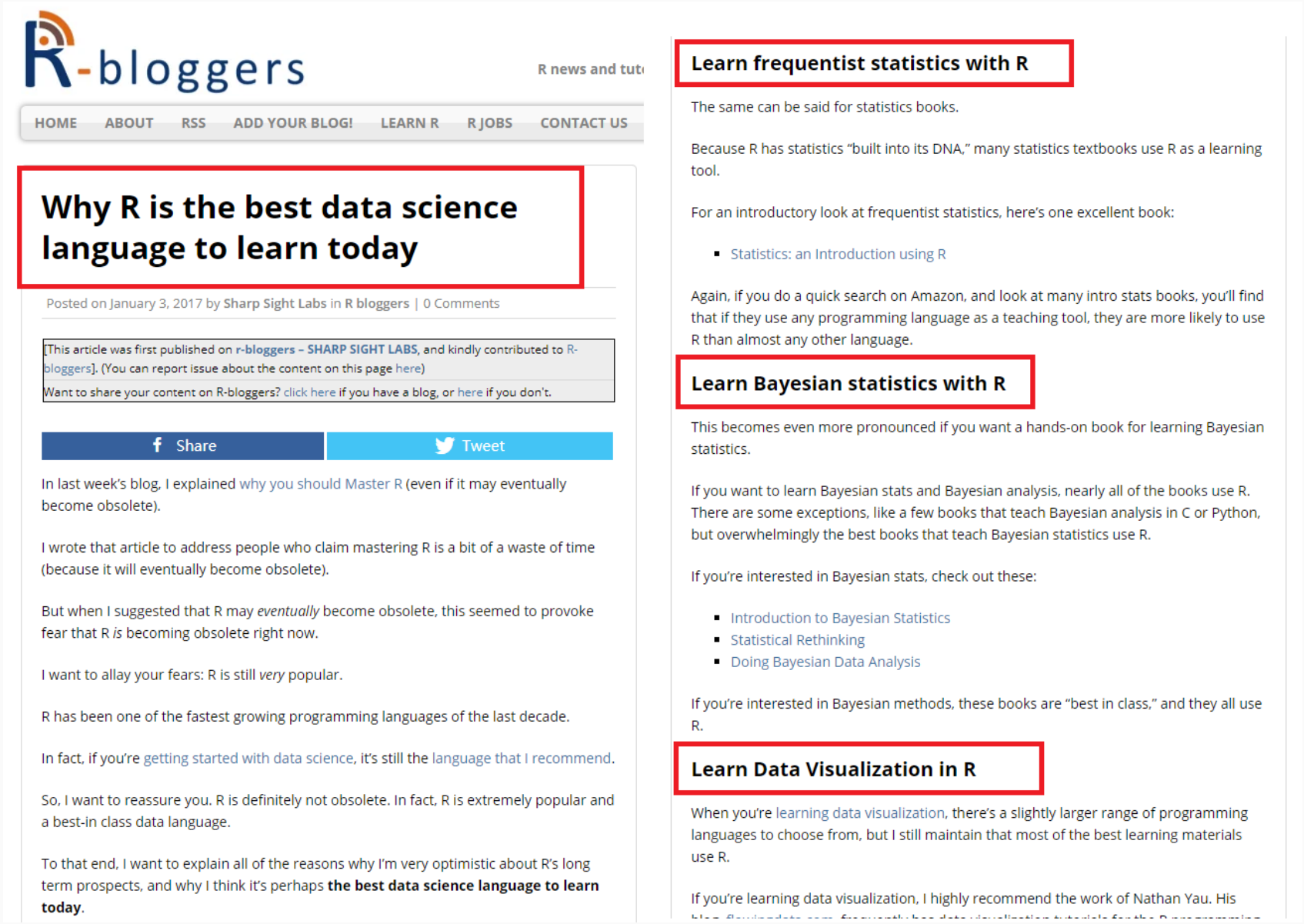
# 残酷的现实

何以解忧，~~唯有~~ 可以撸 R

# R 语言之美，你值得拥有

| 序号 | 内容 | 特性 | 评价 |
|---|---|---|---|
| 1 | 统计分析 | 看家本领 | 好用 |
| 2 | ggplot2 画图 | 颜值担当 | 好看 |
| 3 | tidyverse 语法 | 人类语言 | 好学 |
| 4 | 可重复性报告 | 方便快捷 | 好玩 |

# 当今最值得学习的数据科学语言



R-bloggers

R news and tut

HOME ABOUT RSS ADD YOUR BLOG! LEARN R R JOBS CONTACT US

## Why R is the best data science language to learn today

Posted on January 3, 2017 by **Sharp Sight Labs** in **R bloggers** | 0 Comments

[This article was first published on **r-bloggers – SHARP SIGHT LABS**, and kindly contributed to R-bloggers]. (You can report issue about the content on this page here)

Want to share your content on R-bloggers? click here if you have a blog, or here if you don't.

Share Tweet

In last week's blog, I explained why you should Master R (even if it may eventually become obsolete).

I wrote that article to address people who claim mastering R is a bit of a waste of time (because it will eventually become obsolete).

But when I suggested that R may *eventually* become obsolete, this seemed to provoke fear that R *is* becoming obsolete right now.

I want to allay your fears: R is still *very* popular.

R has been one of the fastest growing programming languages of the last decade.

In fact, if you're getting started with data science, it's still the language that I recommend.

So, I want to reassure you. R is definitely not obsolete. In fact, R is extremely popular and a best-in class data language.

To that end, I want to explain all of the reasons why I'm very optimistic about R's long term prospects, and why I think it's perhaps **the best data science language to learn today**.

### Learn frequentist statistics with R

The same can be said for statistics books.

Because R has statistics "built into its DNA," many statistics textbooks use R as a learning tool.

For an introductory look at frequentist statistics, here's one excellent book:

- Statistics: an Introduction using R

Again, if you do a quick search on Amazon, and look at many intro stats books, you'll find that if they use any programming language as a teaching tool, they are more likely to use R than almost any other language.

### Learn Bayesian statistics with R

This becomes even more pronounced if you want a hands-on book for learning Bayesian statistics.

If you want to learn Bayesian stats and Bayesian analysis, nearly all of the books use R. There are some exceptions, like a few books that teach Bayesian analysis in C or Python, but overwhelmingly the best books that teach Bayesian statistics use R.

If you're interested in Bayesian stats, check out these:

- Introduction to Bayesian Statistics
- Statistical Rethinking
- Doing Bayesian Data Analysis

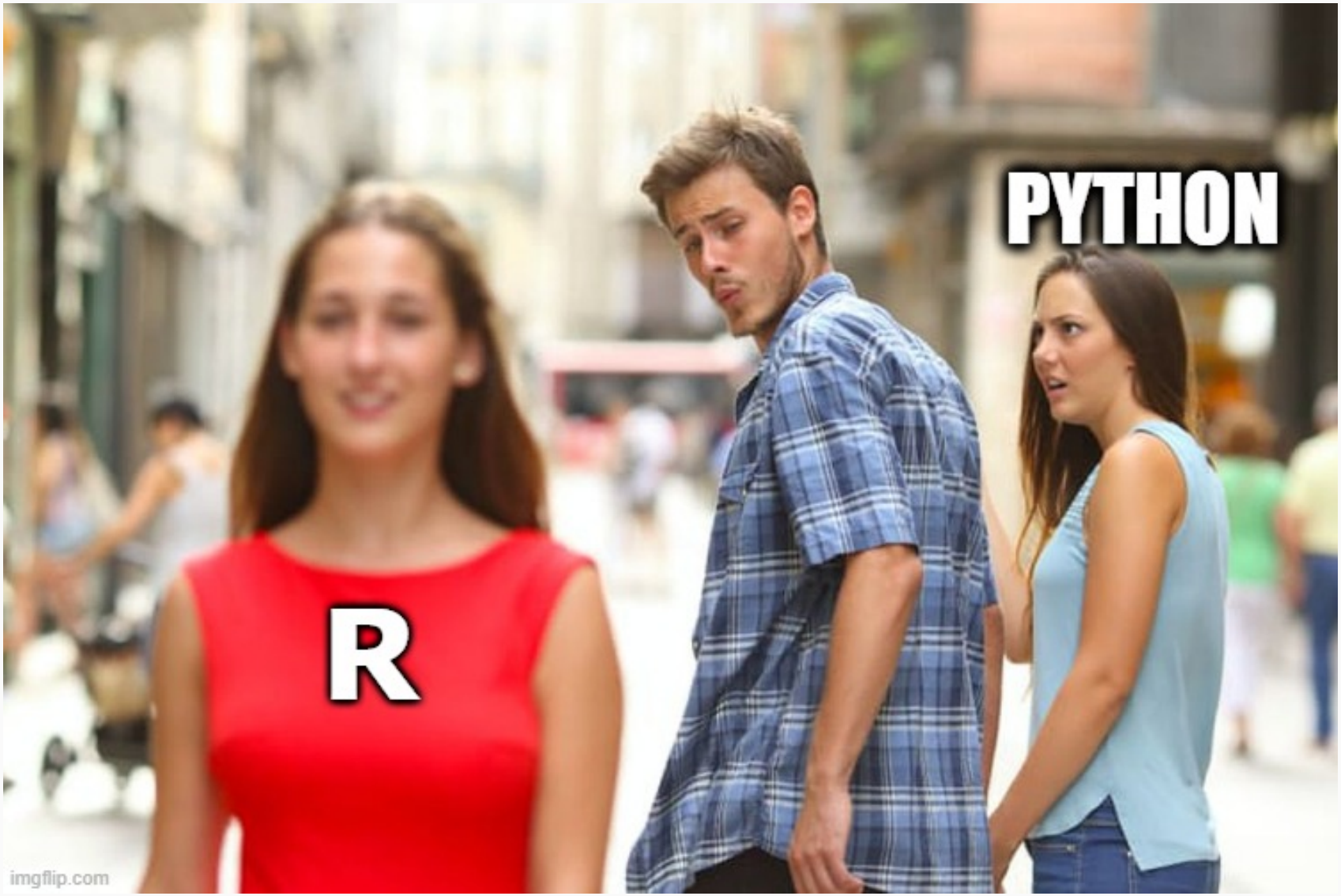
If you're interested in Bayesian methods, these books are "best in class," and they all use R.

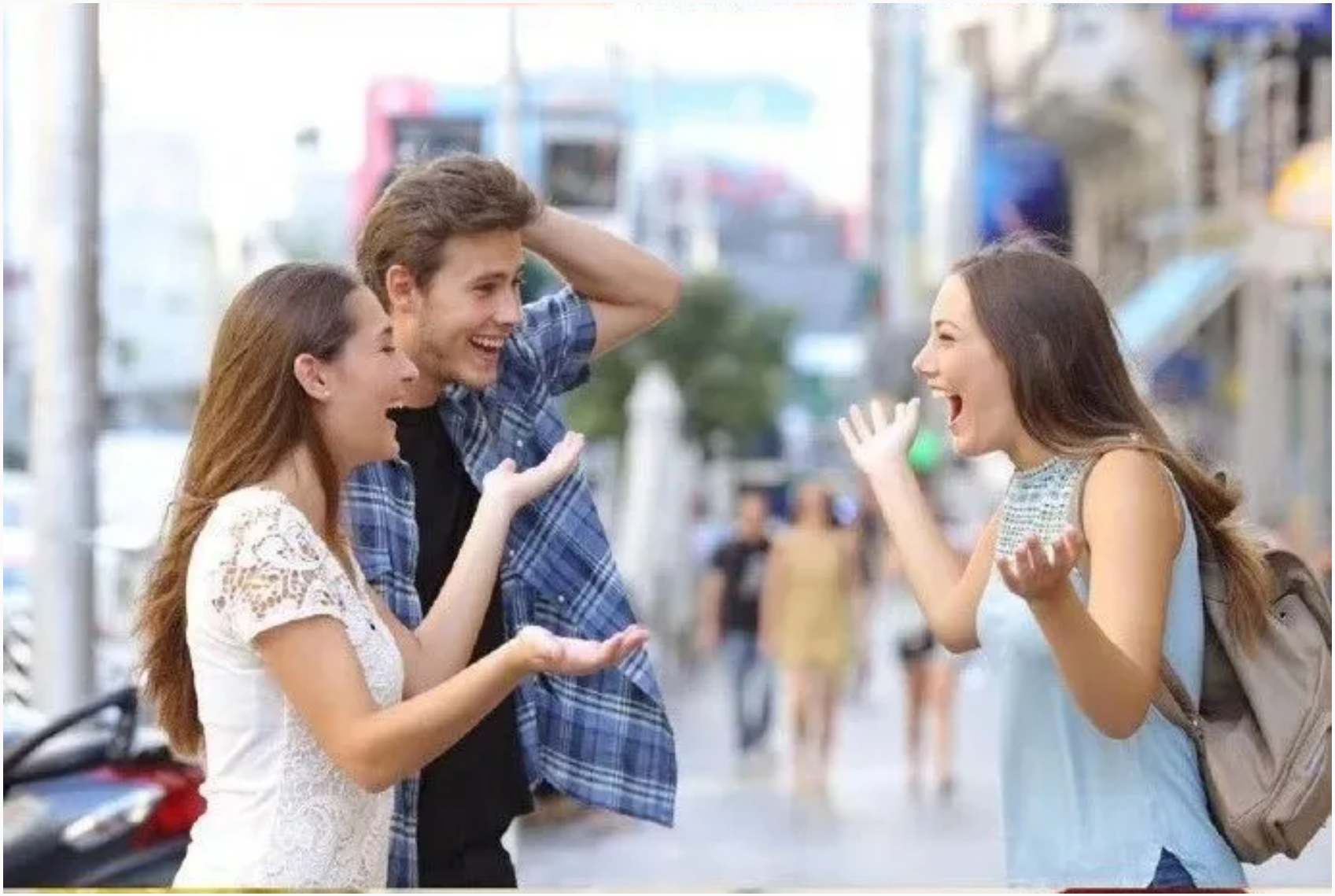
### Learn Data Visualization in R

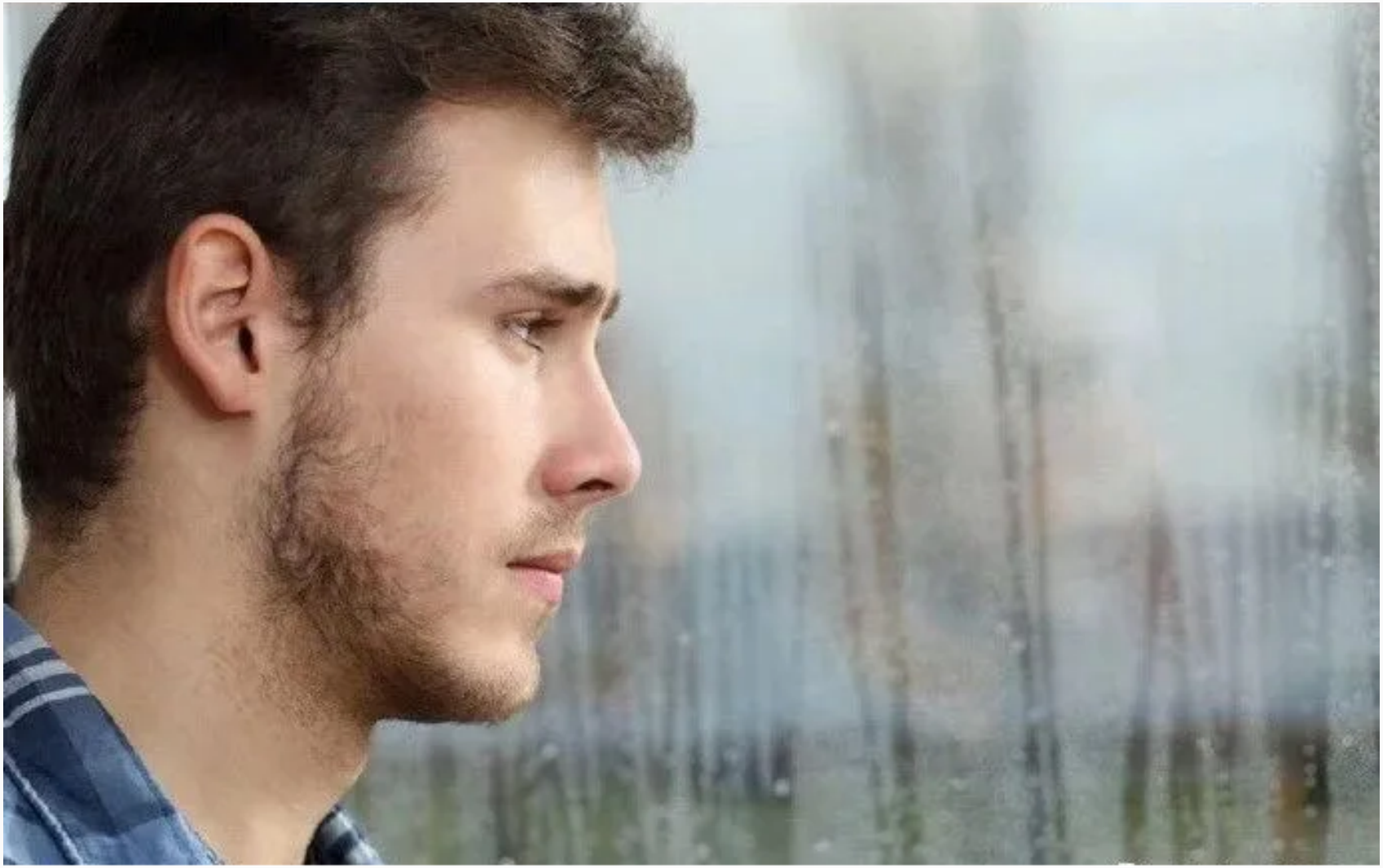
When you're learning data visualization, there's a slightly larger range of programming languages to choose from, but I still maintain that most of the best learning materials use R.

If you're learning data visualization, I highly recommend the work of Nathan Yau. His

# 一见钟情，还是相见恨晚？







# Welcome to R, RStudio, and the tidyverse!