

# 数据科学中的 R 语言

---

王敏杰

2020 年 10 月 26 日

四川师范大学

# R 能给我们生活带来什么？

这个问题，好比人生三大终极问题：

- R 是什么？
- R 能干什么？
- 为什么是 R ？

# 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 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/>



[Home]

[Download](#)

[CRAN](#)

[R Project](#)

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### News

- [R version 4.0.3 \(Bunny-Wunnies Freak Out\)](#) has been released on 2020-10-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- [R version 3.6.3 \(Holding the Windsock\)](#) was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

# 界面很友好

~/Documents/r4ds/data-analysis - RStudio

data-analysis

mpg-plot.R

```
1 library(ggplot2)
2
3 ggplot(mpg, aes(x = displ, y = hwy)) +
4   geom_point(aes(colour = class))
```

Environment History

Global Environment

Environment is empty

Editor (Top Level) R Script

Console ~/Documents/r4ds/data-analysis/

```
> library(ggplot2)
> ggplot(mpg, aes(x = displ, y = hwy)) +
+   geom_point(aes(colour = class))
> |
```

Output

The figure is a scatter plot titled "Output". The x-axis is labeled "displ" and ranges from approximately 2 to 7 liters. The y-axis is labeled "hwy" and ranges from 10 to 50 mpg. Data points are colored according to their vehicle class. A legend on the right side lists the classes and their corresponding colors:

- 2seater (red)
- compact (yellow)
- midsize (green)
- minivan (cyan)
- pickup (light blue)
- subcompact (purple)
- suv (pink)

The plot shows a general downward trend as engine displacement increases, with fuel economy generally decreasing. There is more variability at higher engine displacements.

# 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



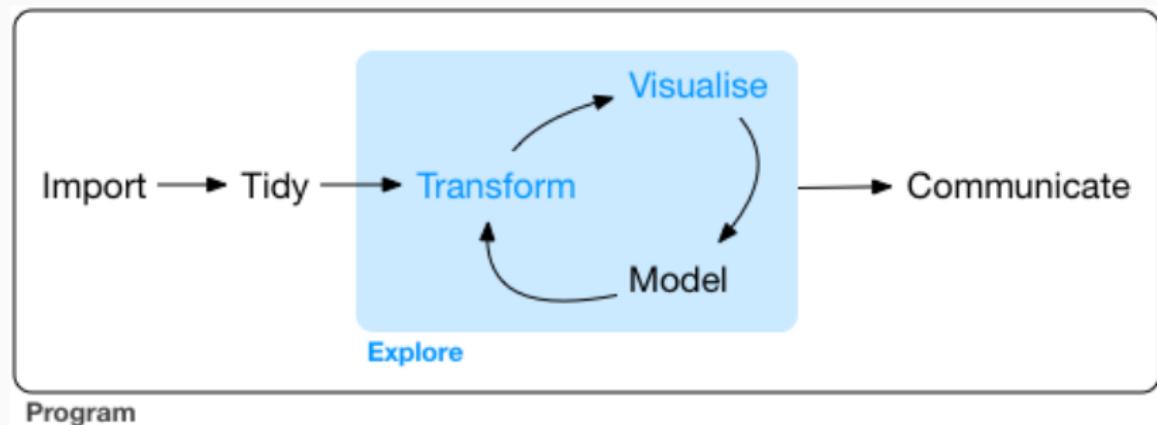
- Hadley Wickham
- R 路上的大神
- 一个改变了 R 语言的人

# R 能干什么

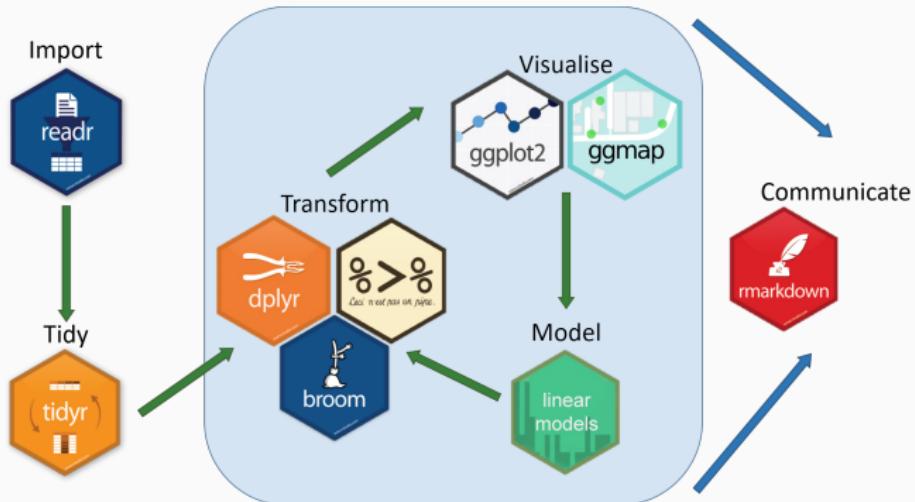
---

# 数据科学的流程

Hadley Wickham 定义了数据科学的工作流程



# tidyverse 套餐



<https://www.tidyverse.org/>

# R 能干什么

序号	内容	代码演示
1	统计	<a href="#">1_stats.R</a>
2	可视化	<a href="#">2_visual.R</a>
3	探索性分析	<a href="#">3_eda.R</a>
4	可重复性报告	<a href="#">4_reproducible.R</a>

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》

但我更希望这门课结束后

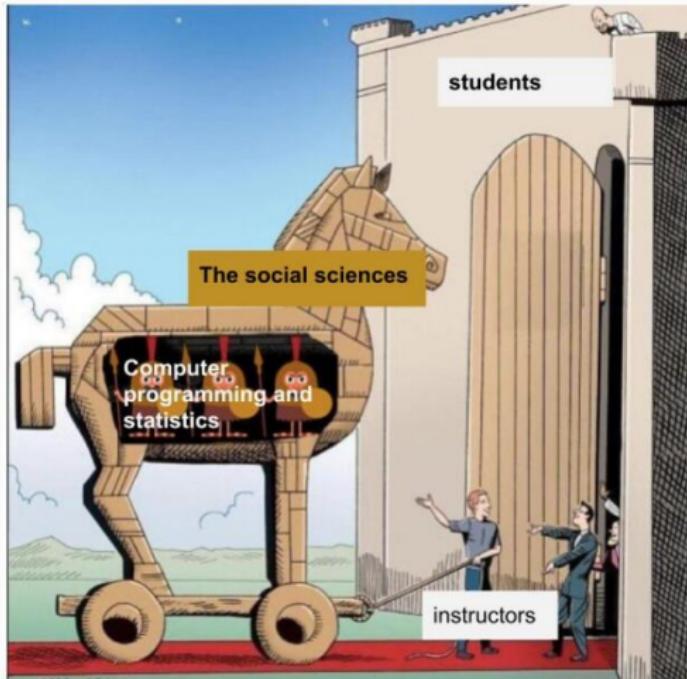


图 2：图片来自美剧《权利的游戏》

# 为什么是 R

---

# 社会科学需要统计



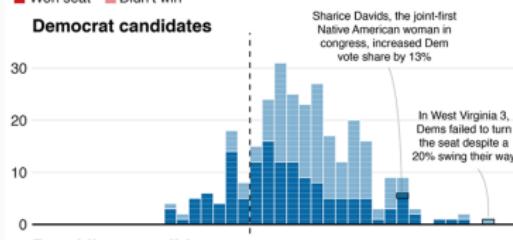
我们不是学统计的，但需要统计

# 社会科学需要可视化

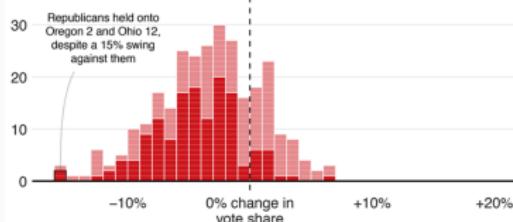
## Blue wave

■ Won seat ■ Didn't win

### Democrat candidates



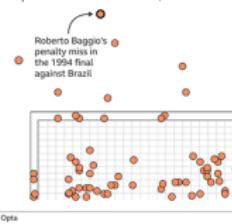
### Republican candidates



Source: AP, 19:01 ET

## Where penalties are saved

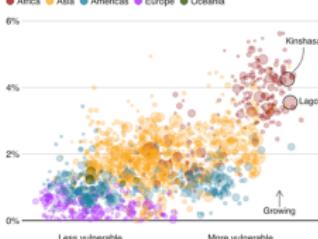
World Cup shootout misses and saves, 1982-2014



## Fast-growing cities face worse climate risks

Population growth 2018-2035 over climate change vulnerability

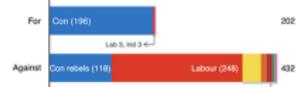
■ Africa ■ Asia ■ Americas ■ Europe ■ Oceania



BBC

Source: Verisk Maplecroft. Circle size represents current population.

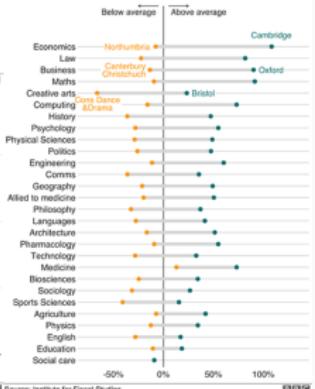
## MPs rejected Theresa May's deal by 230 votes



Source: Commons Votes Service. Excludes Ministers, the Speaker and deputies

Earnings vary across units even within subjects

Impact on men's earnings relative to the average degree



Source: Institute for Fiscal Studies

我们不是学美术的，但要可视化

# 社会科学需要编程



我们不是学计算机的，但需要编程

# 社会科学需要可重复性



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-

社会科学研究，是需要可重复性的

# 挖掘机技术到底哪家强？

你有需求，而



很专业

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	简单易懂	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



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.

f Share

t 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

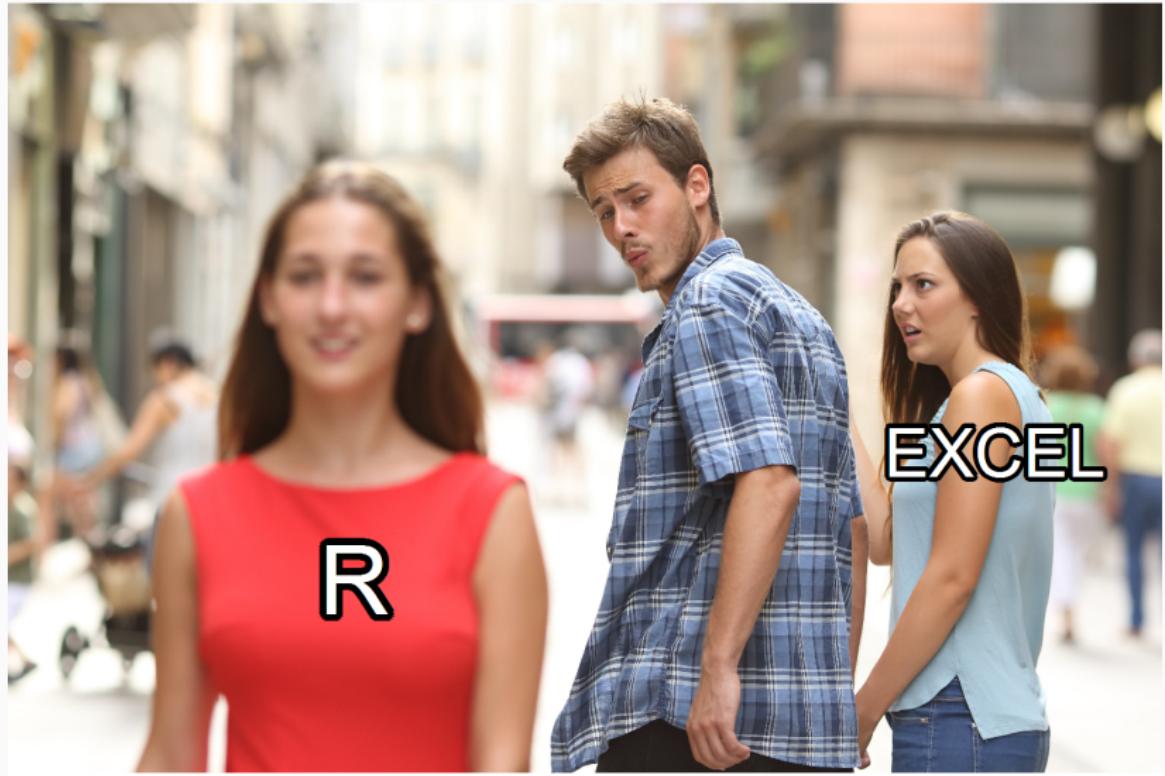
If you're interested in Bayesian methods, these books are "best in class," and they all use R.

### 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？



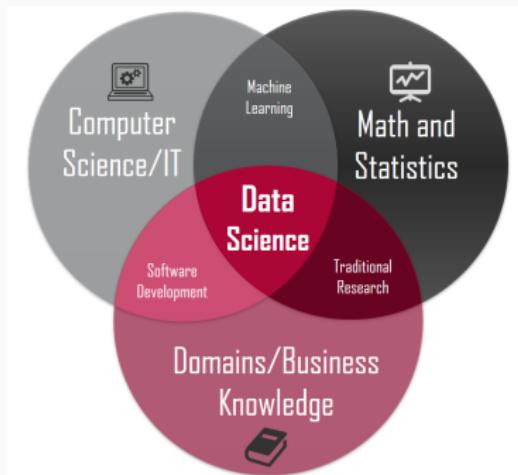
EXCEL

R

# 关于学习

---

# 我们的课程不会枯燥



- 数据科学是为社会科学服务的，我们会有很多案例
- 编程是工具，统计是灵魂，专业是核心

# 关于学习

我很少使用

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

更多的是

```
library(tidyverse)
summary_monthly_temp <- weather %>%
  group_by(month) %>%
  summarize(mean = mean(temp),
            std_dev = sd(temp))
```

# 关于学习

## 课程目标

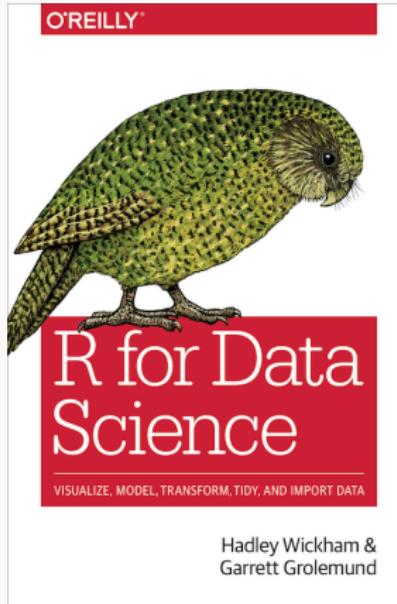
- 训练数据思维，提升编程技能，培养创新能力

## 学习方法

- 问题驱动型学习
- 多实践（光看李小龙的电影，是学不会功夫的）
- 不是 learning R，而是 learning with R
- 把 R 看做是知识学习的脚手架



# 参考书目



- R for Data Science
- <https://bookdown.org/wangminjie/R4DS/>