

# Background for R language



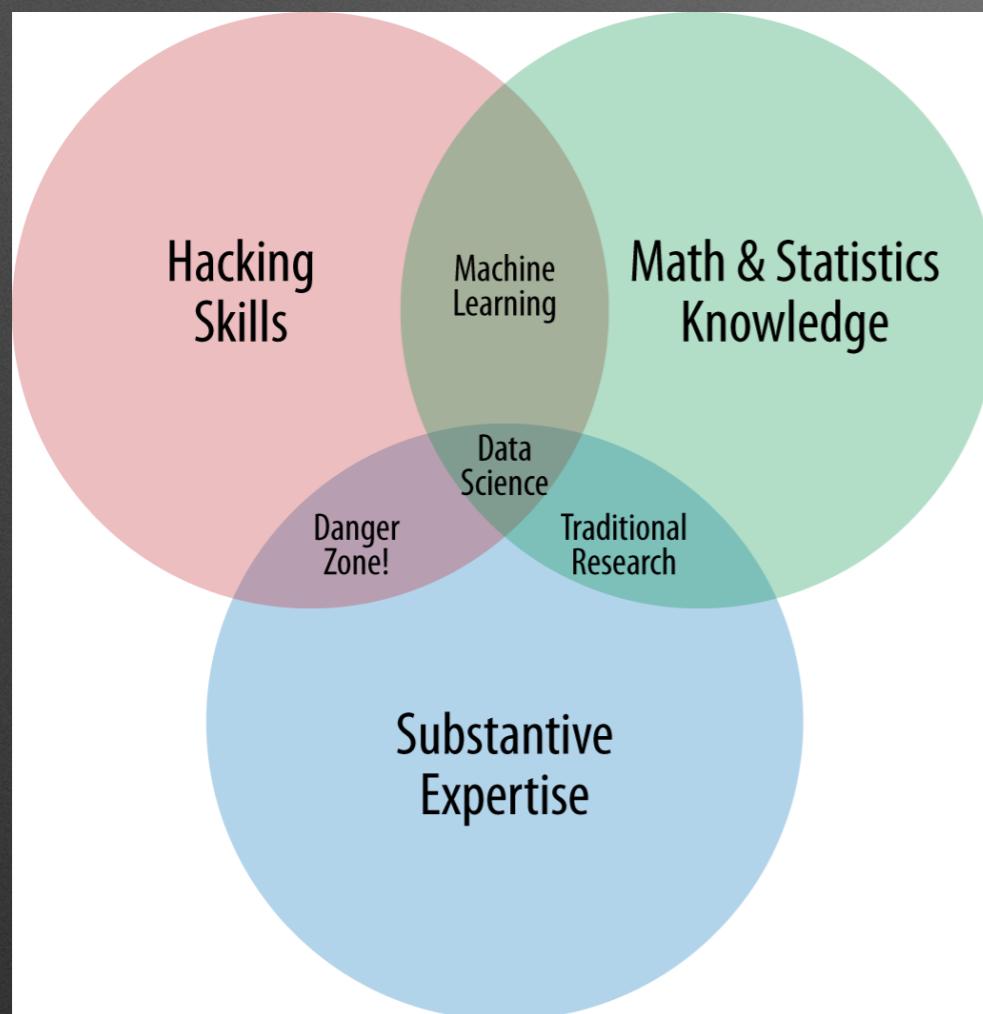
Qiang Shen (沈强)

College of Economics and Management, Zhejiang University of Technology  
13675883767, johnsonzhj@gmail.com

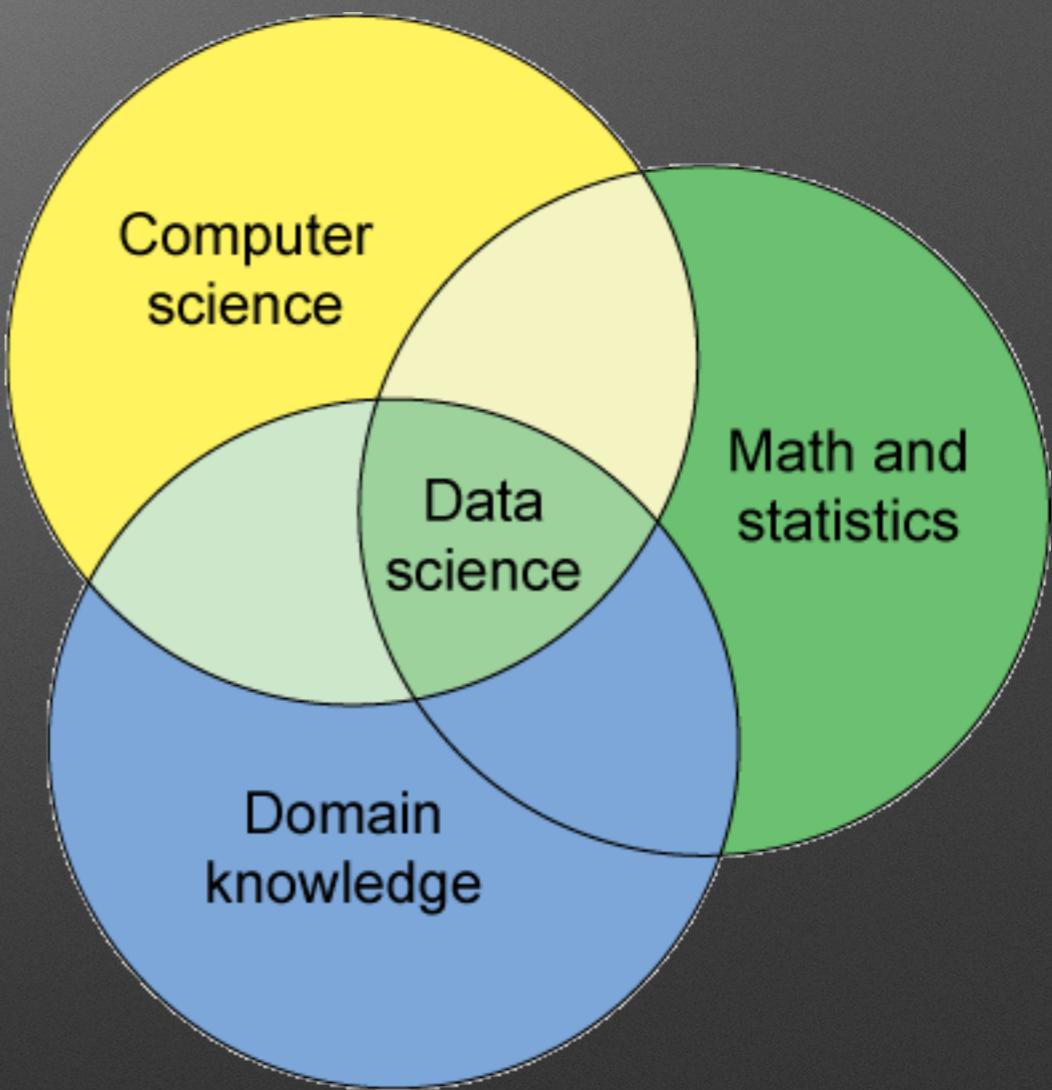
# Data science



# Data science

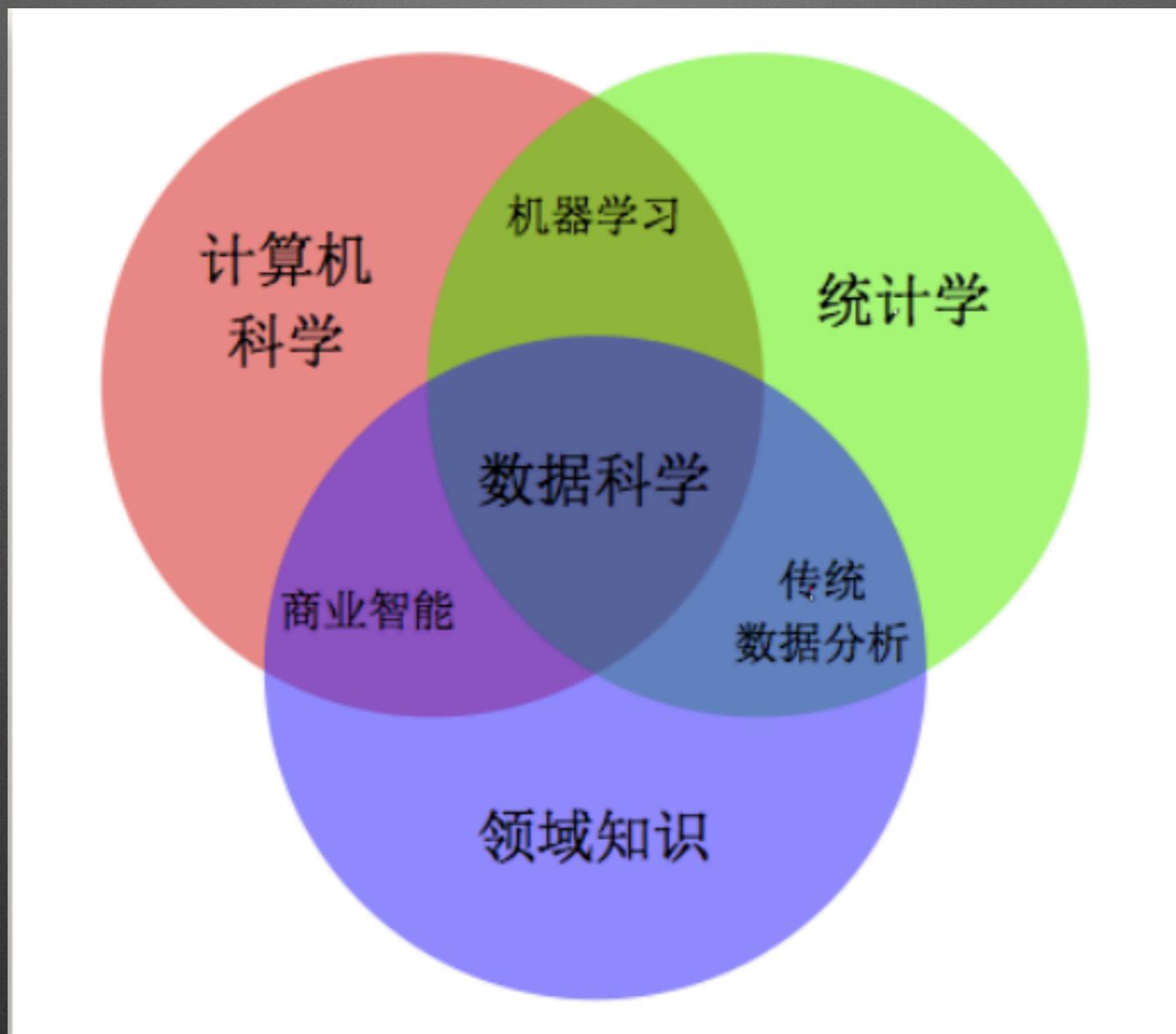


Drew Conway

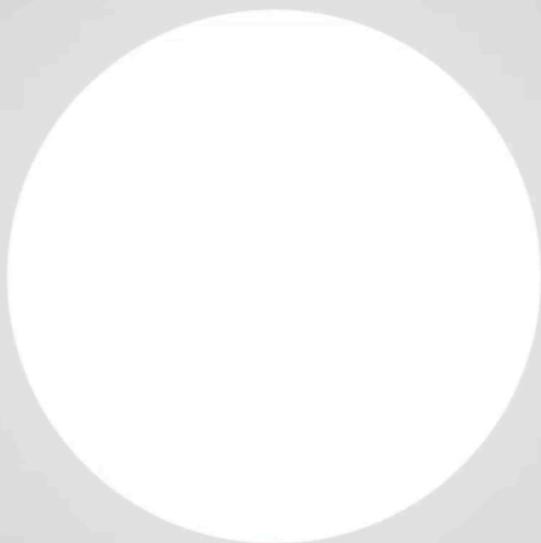


IBM

# Data science



# What's R?



# Why R?

- Statistics: SAS
- Computer science: Fortran
- Domain knowledge : Excel (VBA)
- Machine learning: Python(scikit-learn)
- Traditional data analysis (SPSS)
- Business intelligence (RapidMiner,SAP, ORACLE, IBM)



disdain chain

Name	Advantages	Disadvantages	Open source?	Typical users
R	Library support; visualization	Steep learning curve	Yes	Finance; Statistics
Matlab	Elegant matrix support; visualization	Expensive; incomplete statistics support	No	Engineering
SciPy/NumPy/Matplotlib	Python (general-purpose programming language)	Immature	Yes	Engineering
Excel	Easy; visual; flexible	Large datasets	No	Business
SAS	Large datasets	Expensive; outdated programming language	No	Business; Government
Stata	Easy statistical analysis		No	Science
SPSS	Like Stata but more expensive and worse			

[http://stanfordphd.com/Statistical\\_Software.html](http://stanfordphd.com/Statistical_Software.html)

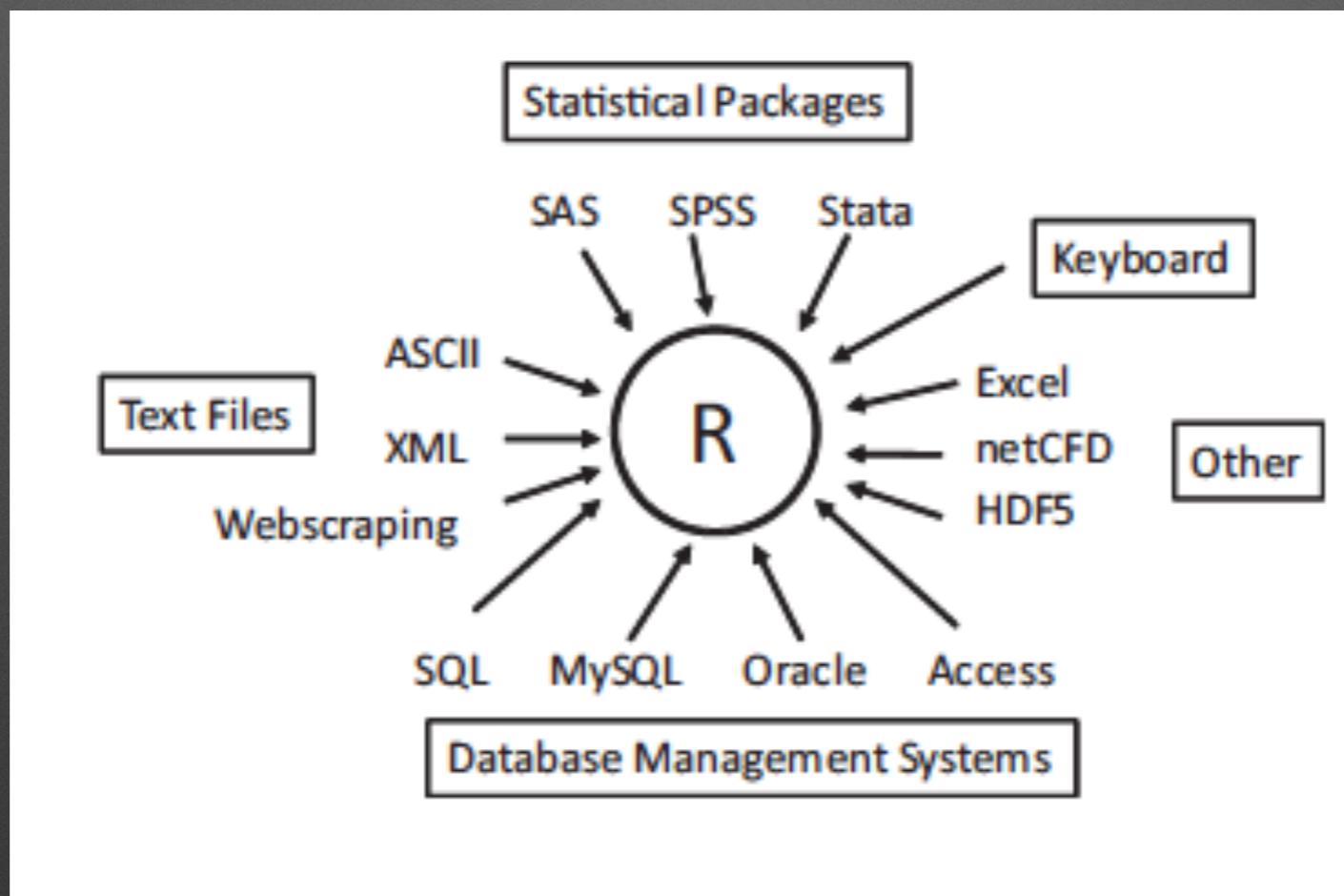
# Visualization

```
require(quantmod)
sse<-getSymbols("^SSEC",from = "2005-01-01"
,to = Sys.Date(),src = "yahoo")
SSEC.m<-to.monthly(SSEC)
candleChart(SSEC.m,theme="white")
```

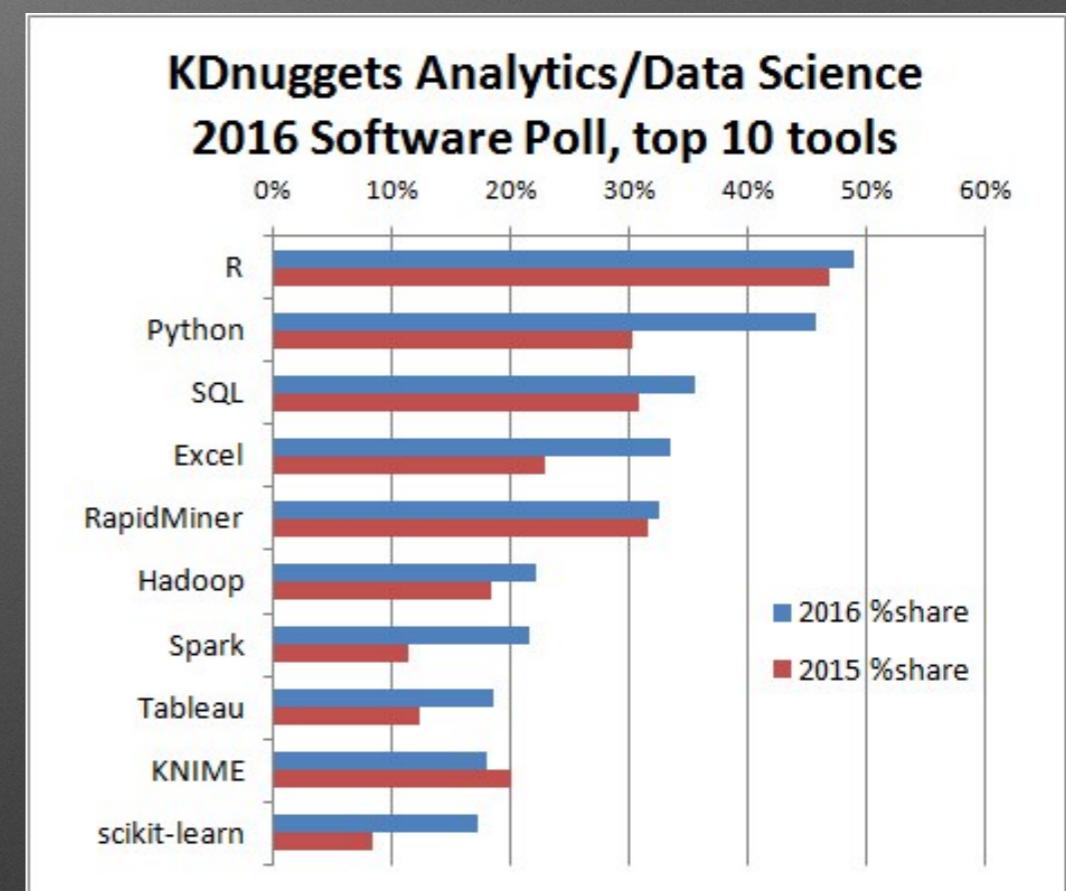
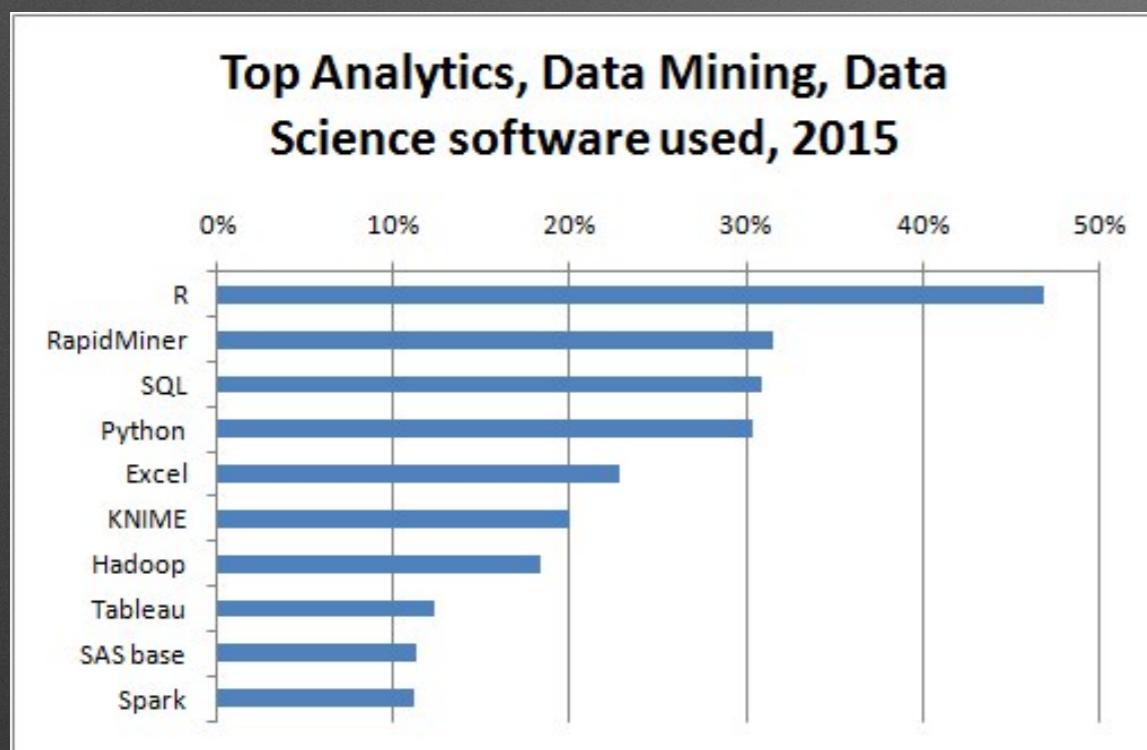
# SSEC



# multiple data source

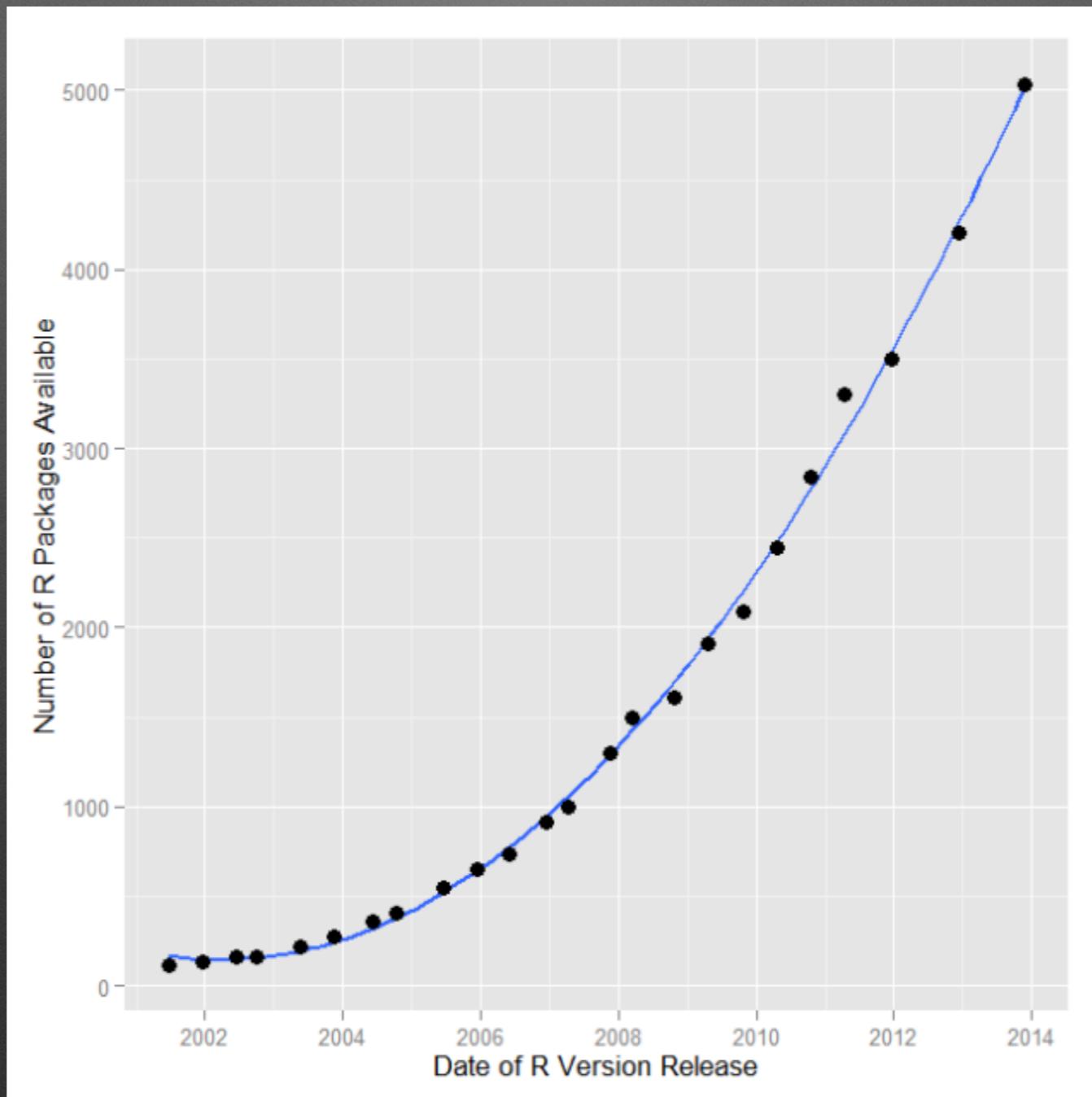


# The trend for R

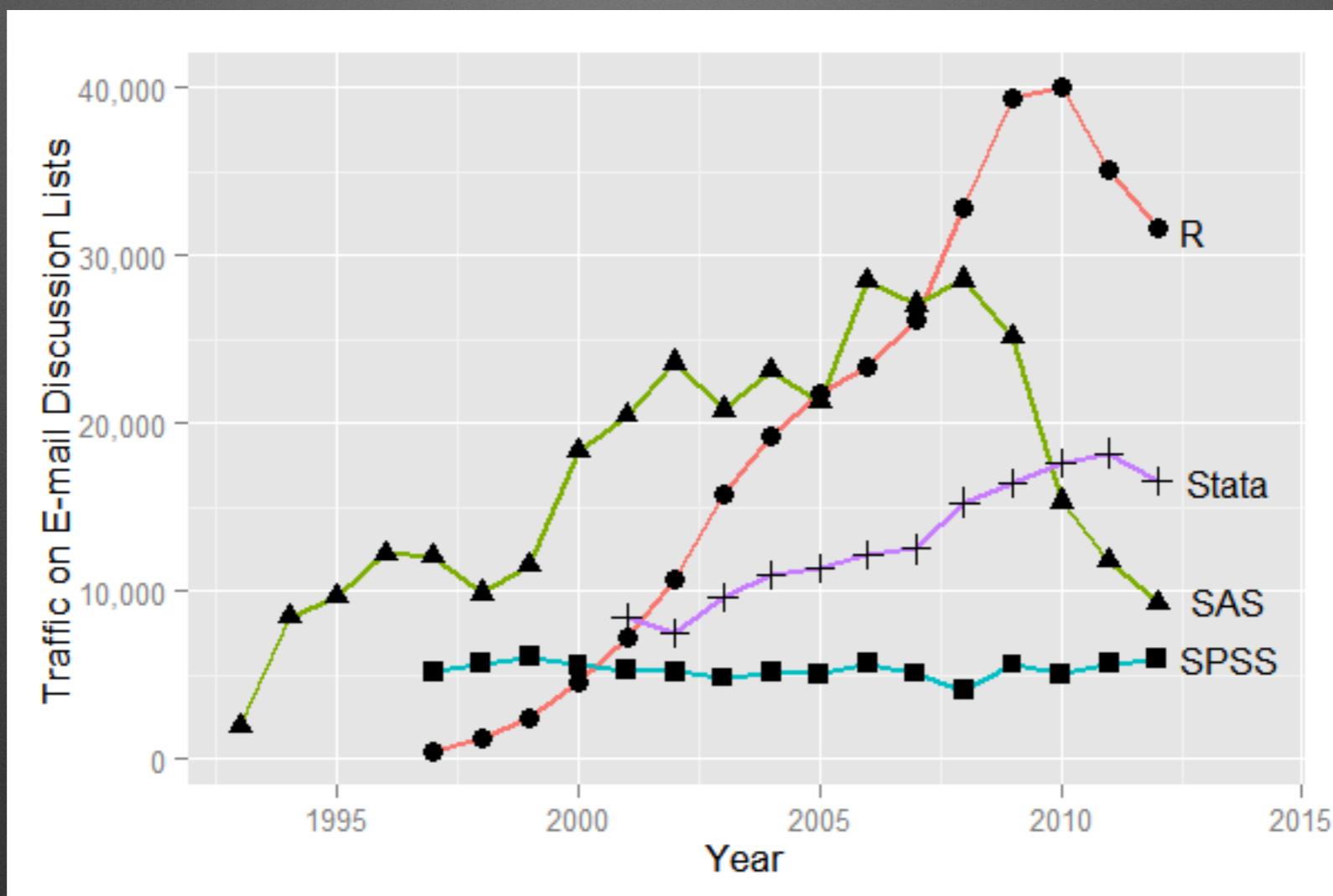


kdnuggets

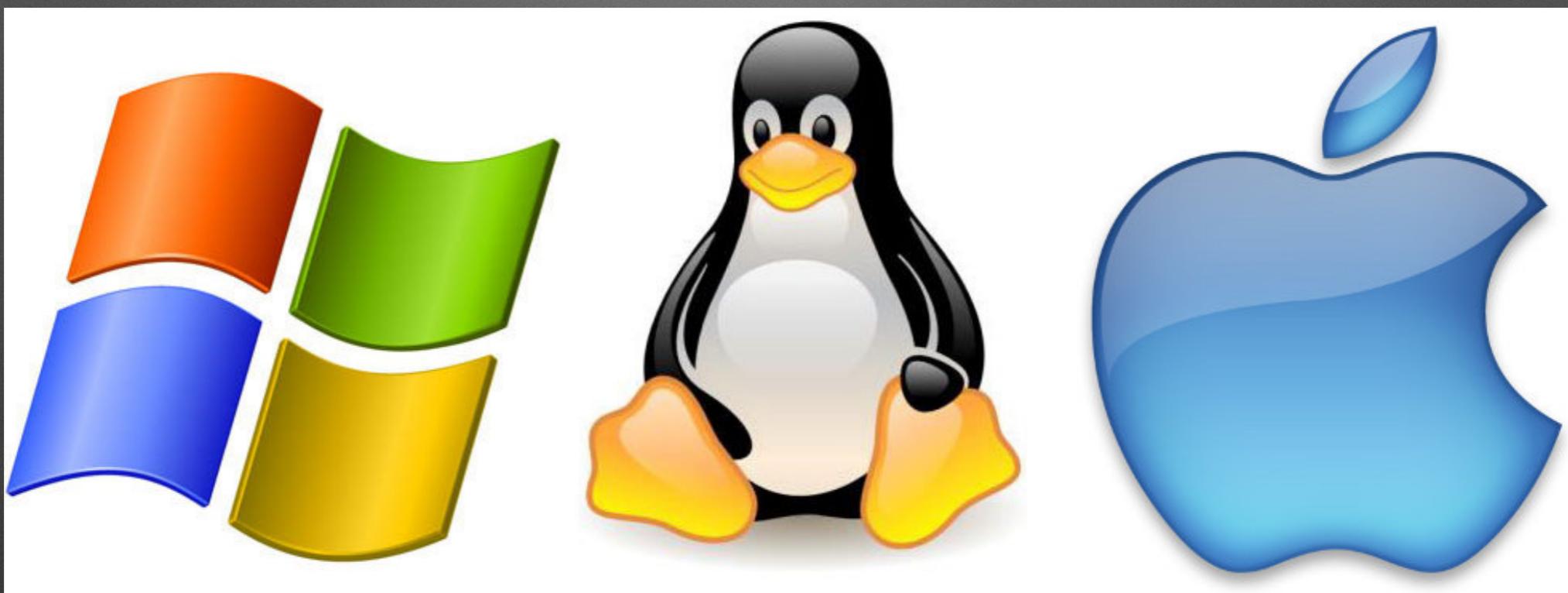
# R packages



# R users



# Cross platform



# Interact with other software

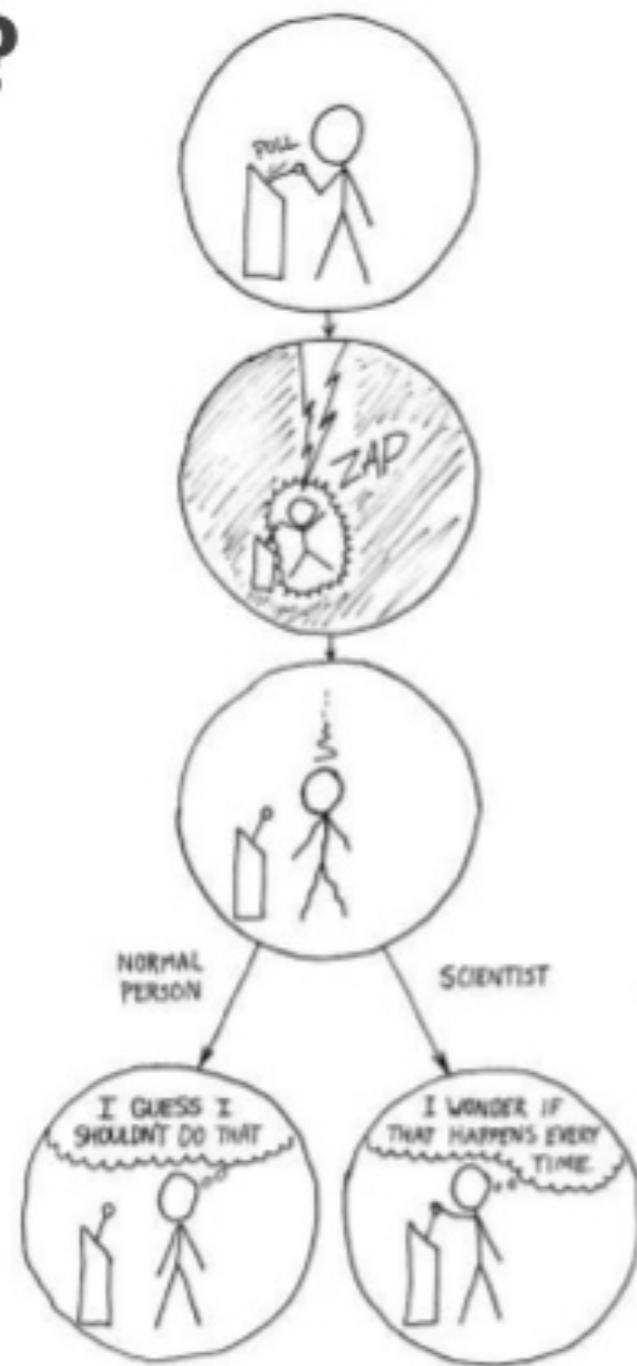
- **Database:** ORACLE, DB2, MySQL
- **Language:** Java, C, C++, Python
- **Web**
- **Statistical software:** SAS, SPSS, Statistica

# Reproducibility

## What is Reproducibility?

*"The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better understood and verified."*  
CRAN Task View on Reproducible Research (Kuhn)

- Method + Environment  
-> Results
- A **process** for:
  - Sharing the method
  - Describing the environment
  - Recreating the results

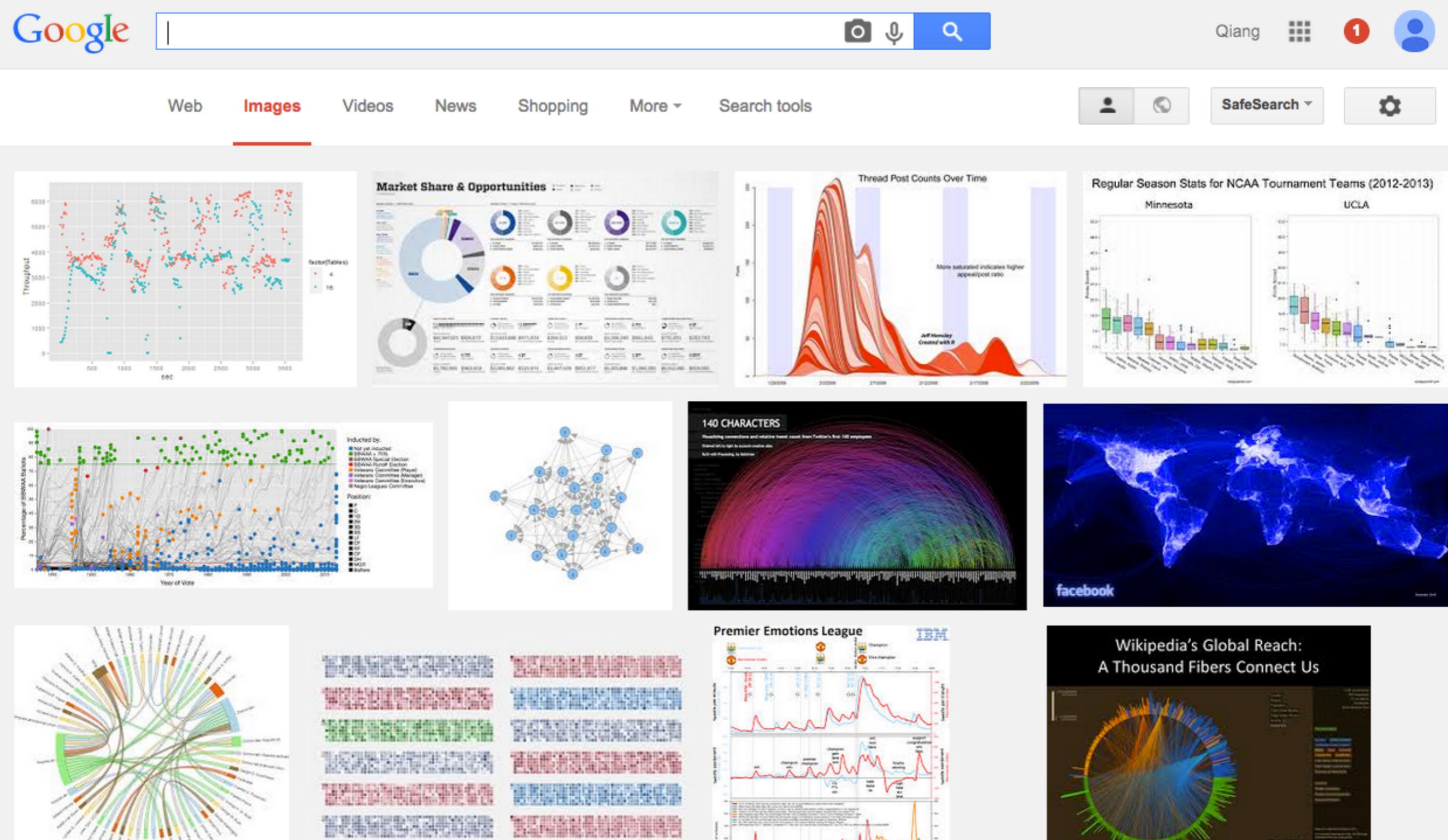


# Style

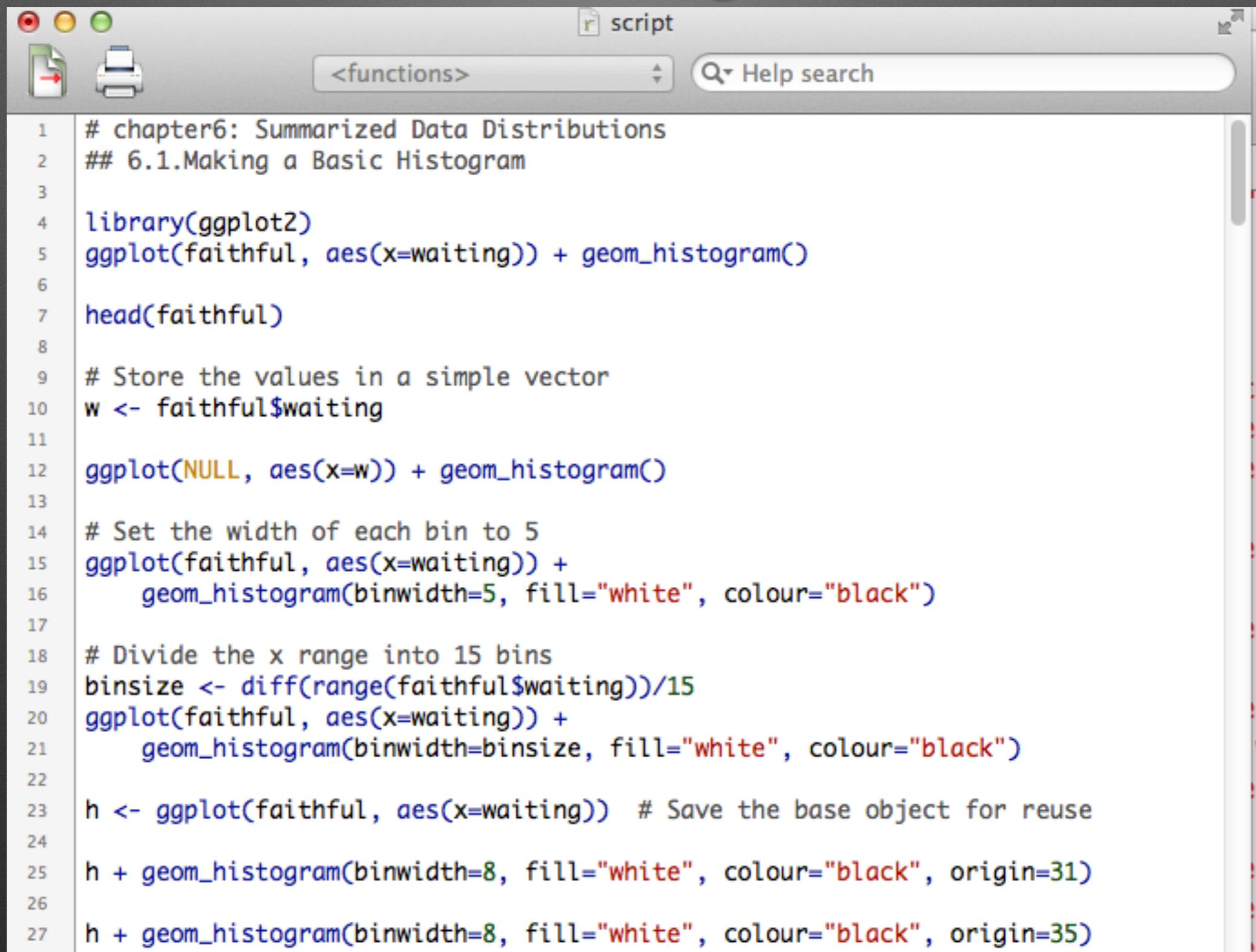
Not the style like this...

A screenshot of a Google Images search results page. The search bar at the top contains the query "style". Below the search bar, there are navigation links for "Web", "Videos", "Images" (which is highlighted in red), "News", "Shopping", "More", and "Search tools". To the right of these are user profile icons for "Qiang" and a notification badge with the number "1". Further right are "SafeSearch" and "Settings" buttons. The main content area displays five main categories with sub-images: "Men" (three male models in casual attire), "The Word" (two logos for "Style" and "style"), "Fashion" (a grid of women's fashion looks), "Writing" (a graphic about style as tone, personality, and voice), and "Quotes" (a quote by Oscar de la Renta). Below these are two rows of more fashion-related images, including a large graphic of a woman in a colorful outfit with the word "Style" written in pink.

# like this?



# The reality is...



The screenshot shows the RStudio interface with a script editor window titled "script". The window contains R code for generating histograms using the ggplot2 package. The code includes loading the ggplot2 library, plotting the "waiting" variable from the faithful dataset, and creating multiple histograms with different bin widths and colors.

```
1 # chapter6: Summarized Data Distributions
2 ## 6.1.Making a Basic Histogram
3
4 library(ggplot2)
5 ggplot(faithful, aes(x=waiting)) + geom_histogram()
6
7 head(faithful)
8
9 # Store the values in a simple vector
10 w <- faithful$waiting
11
12 ggplot(NULL, aes(x=w)) + geom_histogram()
13
14 # Set the width of each bin to 5
15 ggplot(faithful, aes(x=waiting)) +
16   geom_histogram(binwidth=5, fill="white", colour="black")
17
18 # Divide the x range into 15 bins
19 binsize <- diff(range(faithful$waiting))/15
20 ggplot(faithful, aes(x=waiting)) +
21   geom_histogram(binwidth=binsize, fill="white", colour="black")
22
23 h <- ggplot(faithful, aes(x=waiting)) # Save the base object for reuse
24 h + geom_histogram(binwidth=8, fill="white", colour="black", origin=31)
25 h + geom_histogram(binwidth=8, fill="white", colour="black", origin=35)
```

# aspiration

# reality



原来, 现实中数据分析师也是苦逼的码农的一种...



You know, it is also an alias for  
PERMENANT HAED DAMAGE.



the dream is plump while the  
reality is skinny



梦想还是要有的  
万一实现了呢

*We'd better have a dream, in case it  
comes true someday.*      *-Jack Ma?*





R  
[www.r-project.org](http://www.r-project.org)  
The engine



Rstudio  
[www.rstudio.org](http://www.rstudio.org)  
The pretty face



Microsoft R Open  
[mran.microsoft.com/o  
pen/  
Enhanced R  
distribution](https://mran.microsoft.com/open/)

# Books

- R in action. Robert Kabacoff. Manning Publications, 2015, 2<sup>nd</sup> edition  
<http://www.statmethods.net/> textbook
- R语言实战（第2版） 人民邮电出版社, 2016
- Learning R (影印版), Richard Cotton, 东南大学出版社, 2014
- R Graphics Cookbook, Winston Chang, O'Reilly Media, Inc, 2013
- Using R for introductory statistics (second edition), John Verzani, Chapman & Hall/CRC The R Series, 2014
- The Art of R Programming, Norman Matloff No Starch Press, 2011
- R for Everyone: Advanced Analytics and Graphics, Jared Lander, Addison-Wesley Professions, 2013
- 数据科学中的R语言 李舰 肖凯, 西安交通大学出版社, 2015
- R数据分析——方法与案例详解. 方匡南 电子工业出版社. 2015

# websites

- English, Google
- <http://www.r-bloggers.com>
- <http://stackoverflow.com/>
- 统计之都