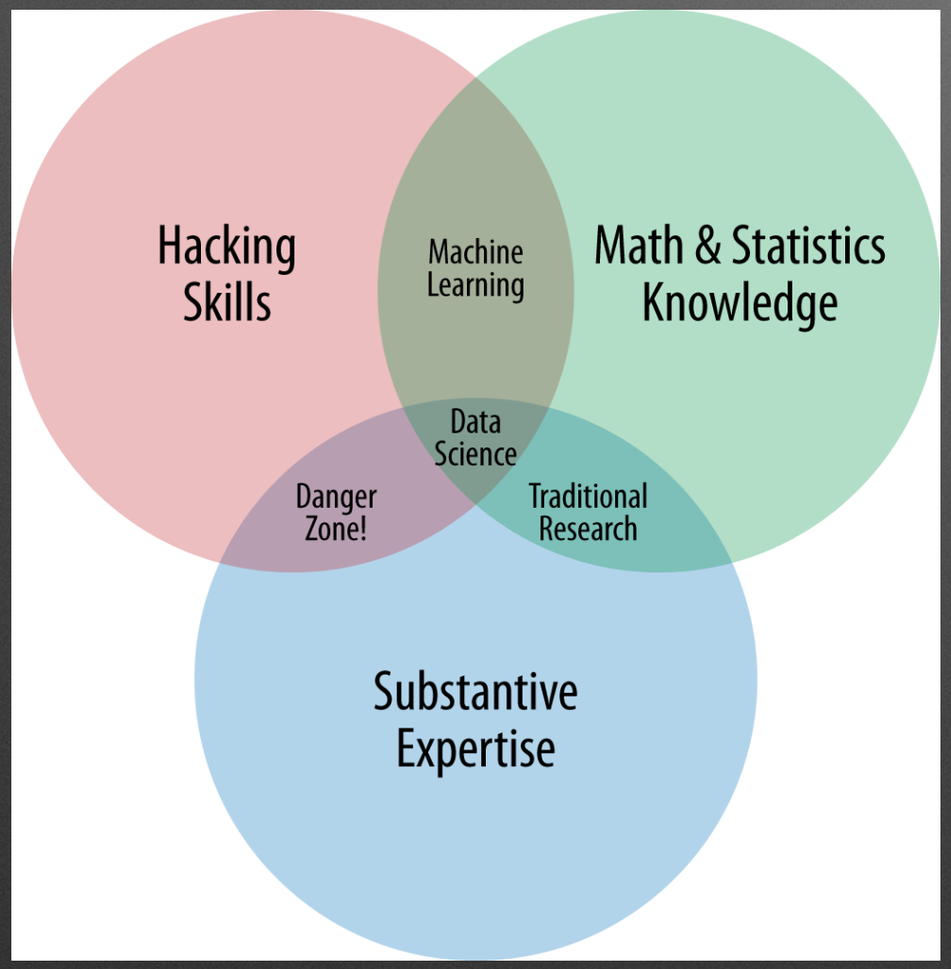
Background for R language

Qiang Shen (沈强)

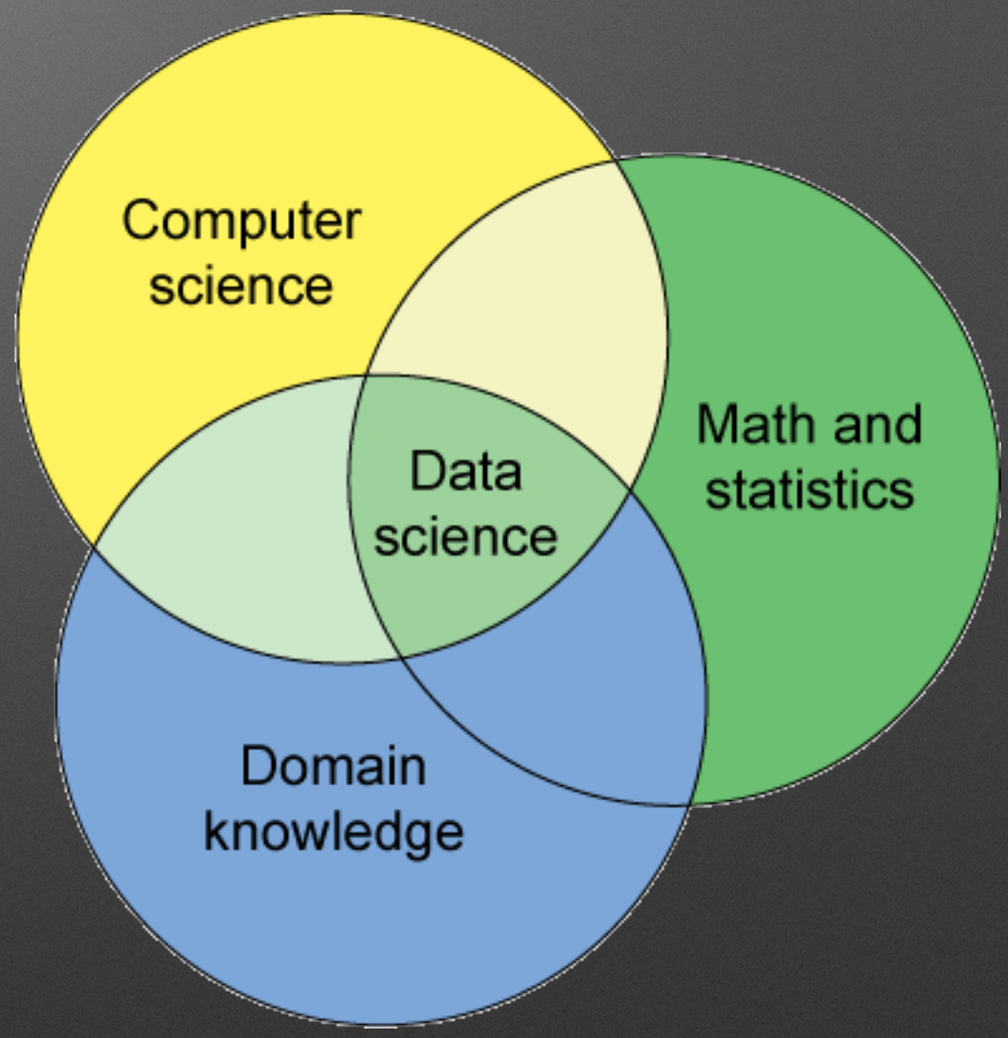
College of Economics and Management, Zhejiang University of Technology
13675883767, johnsonzhj@gmail.com

Data science

Data science

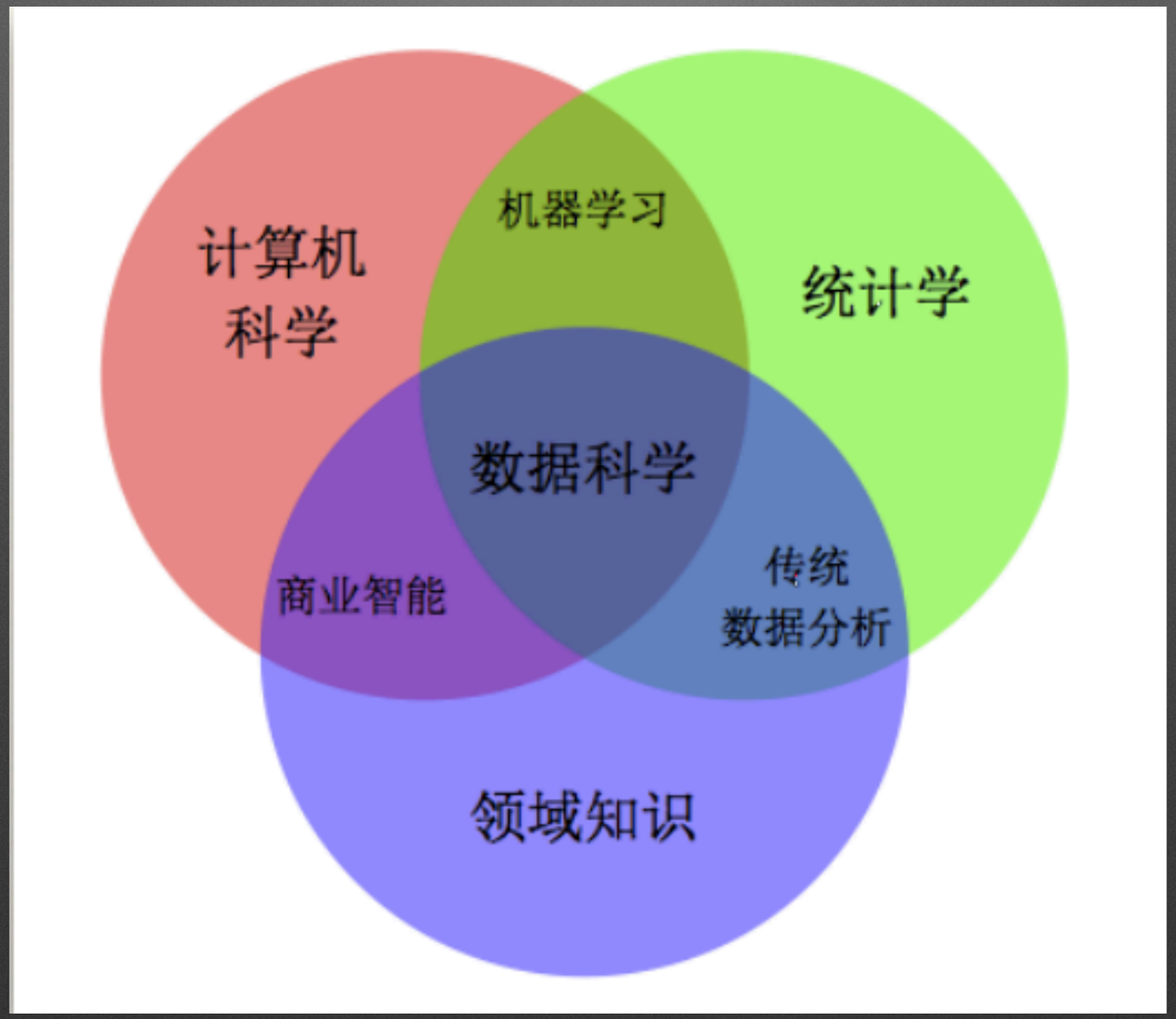


Drew Conway



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SAS	Large datasets	Expensive; outdated programming language	No	Business; Government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

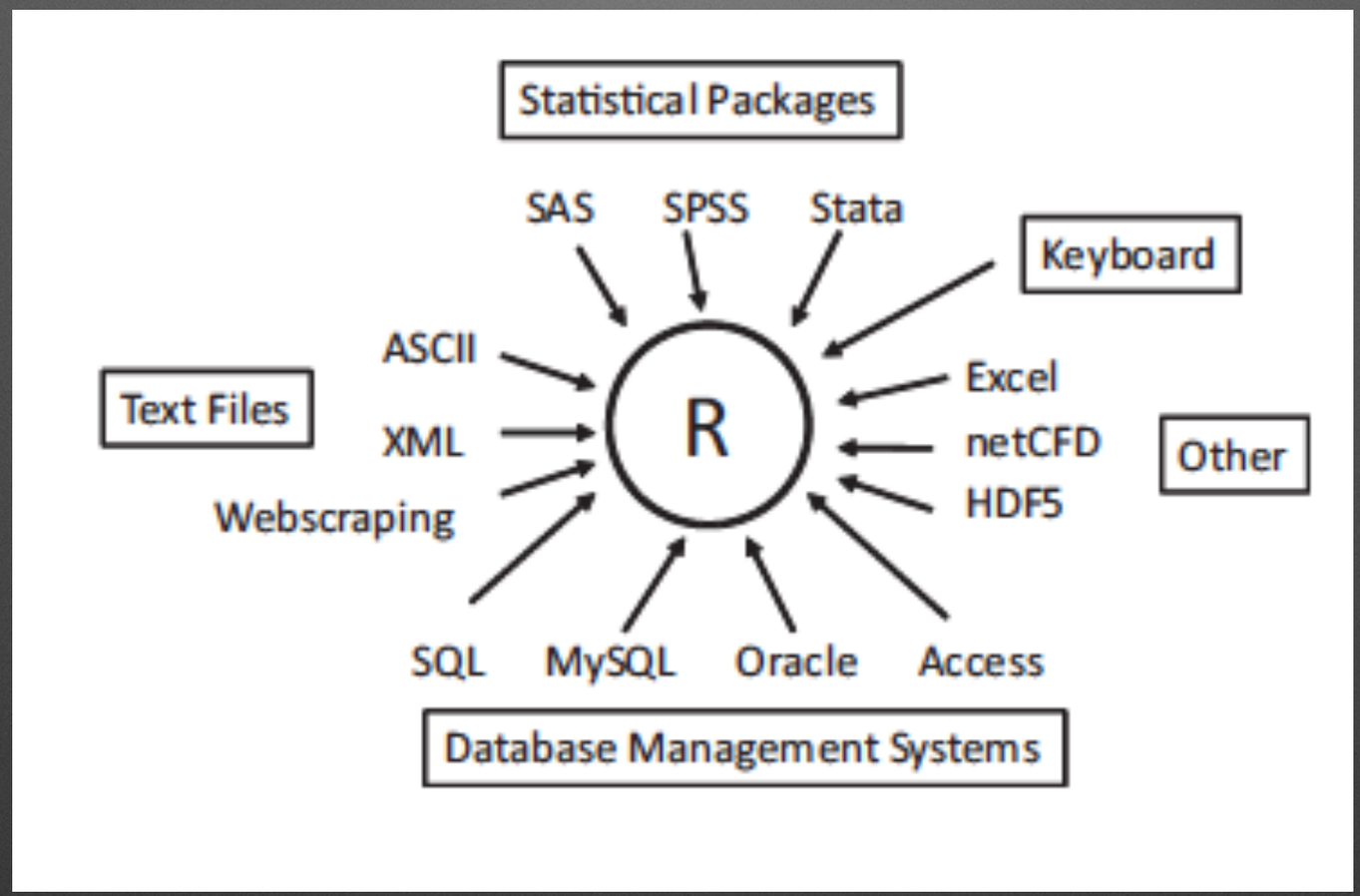
Visualization

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

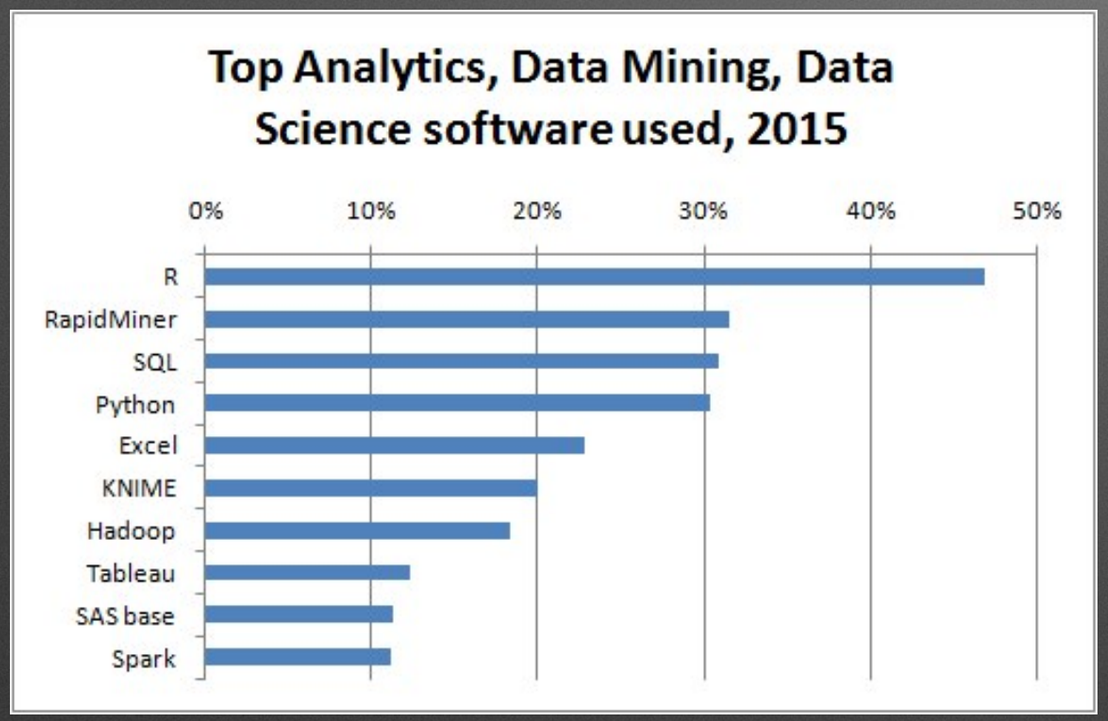
SSEC

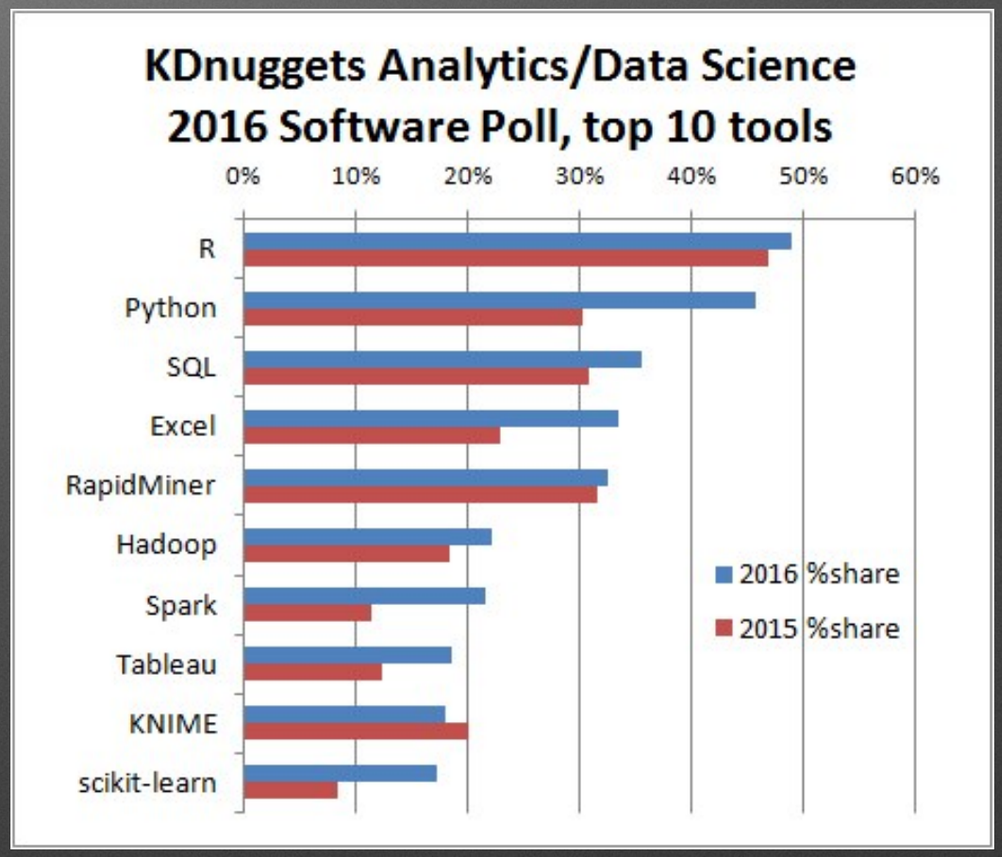

SSEC.m
[2005-01-01 08:00:00/2015-09-01 08:00:00]
Last 3114.798096
6000
5000
4000
3000
2000
1000
Volume (millions):
2,642,300
12
10
8
6
4
2
0
Jan 2005
Jan 2006
Jan 2007
Jan 2008
Jan 2009
Jan 2010
Jan 2011
Jan 2012
Jan 2013
Jan 2014
Jan 2015

multiple data source



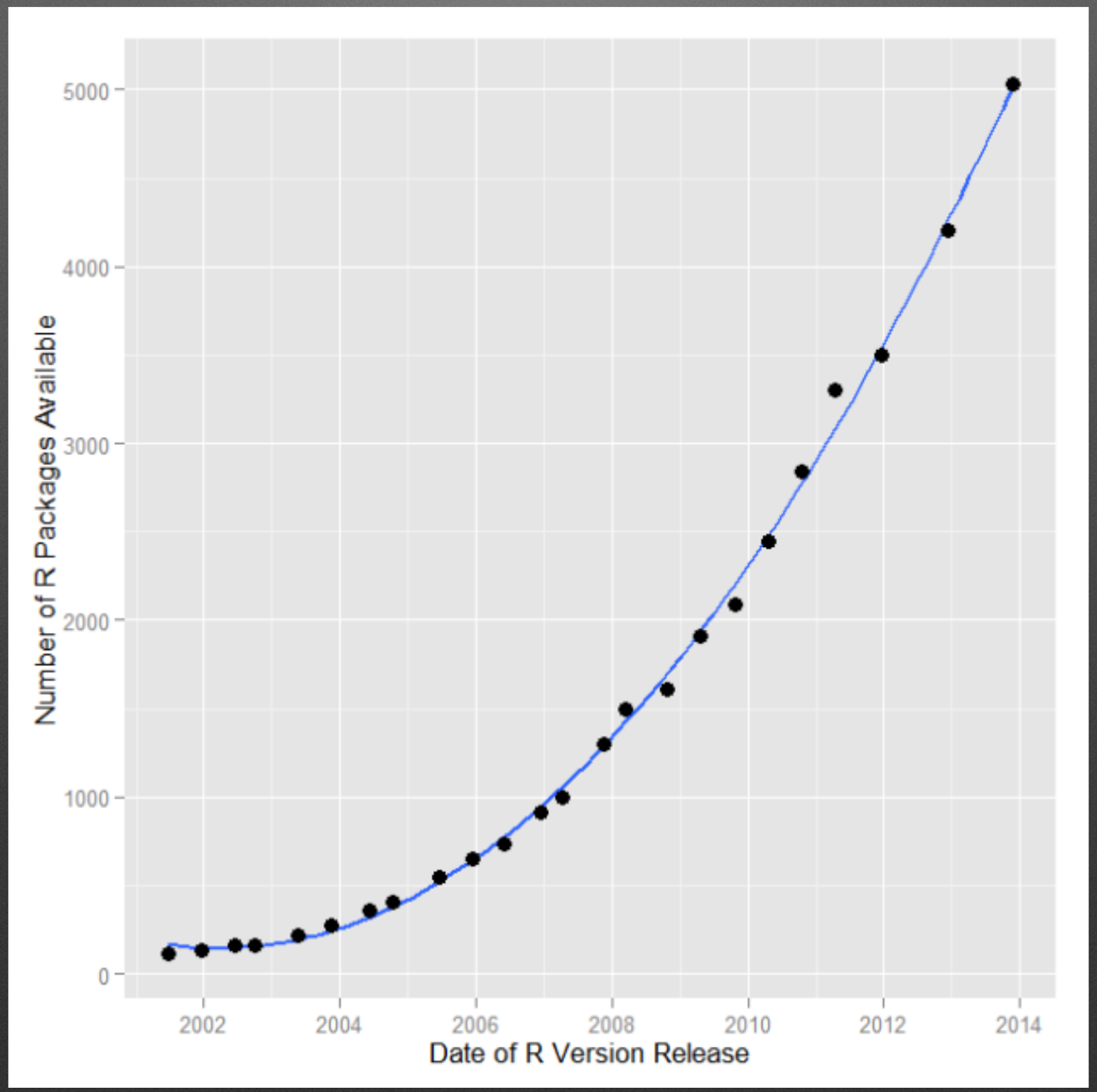
The trend for R



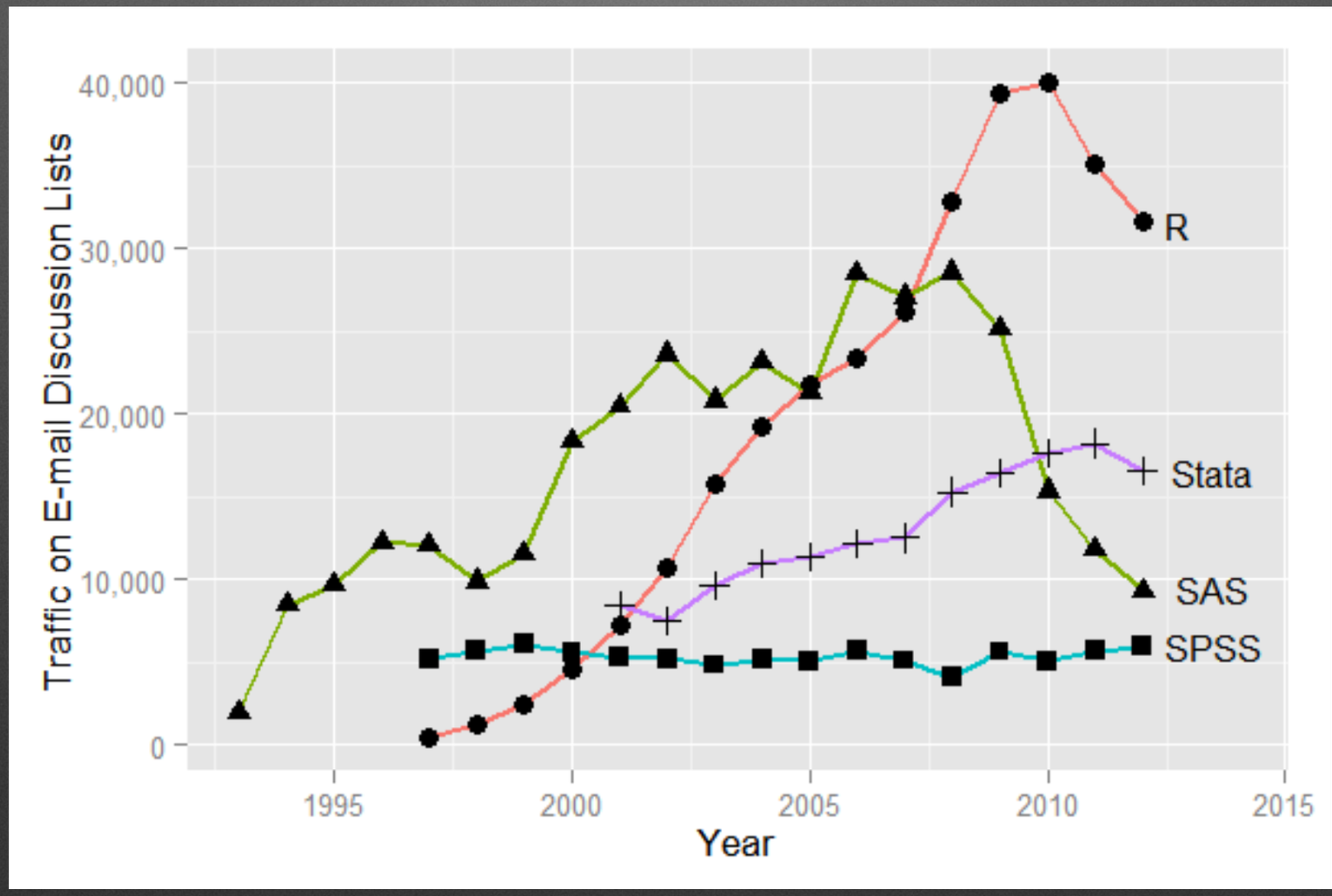


kdnuggets

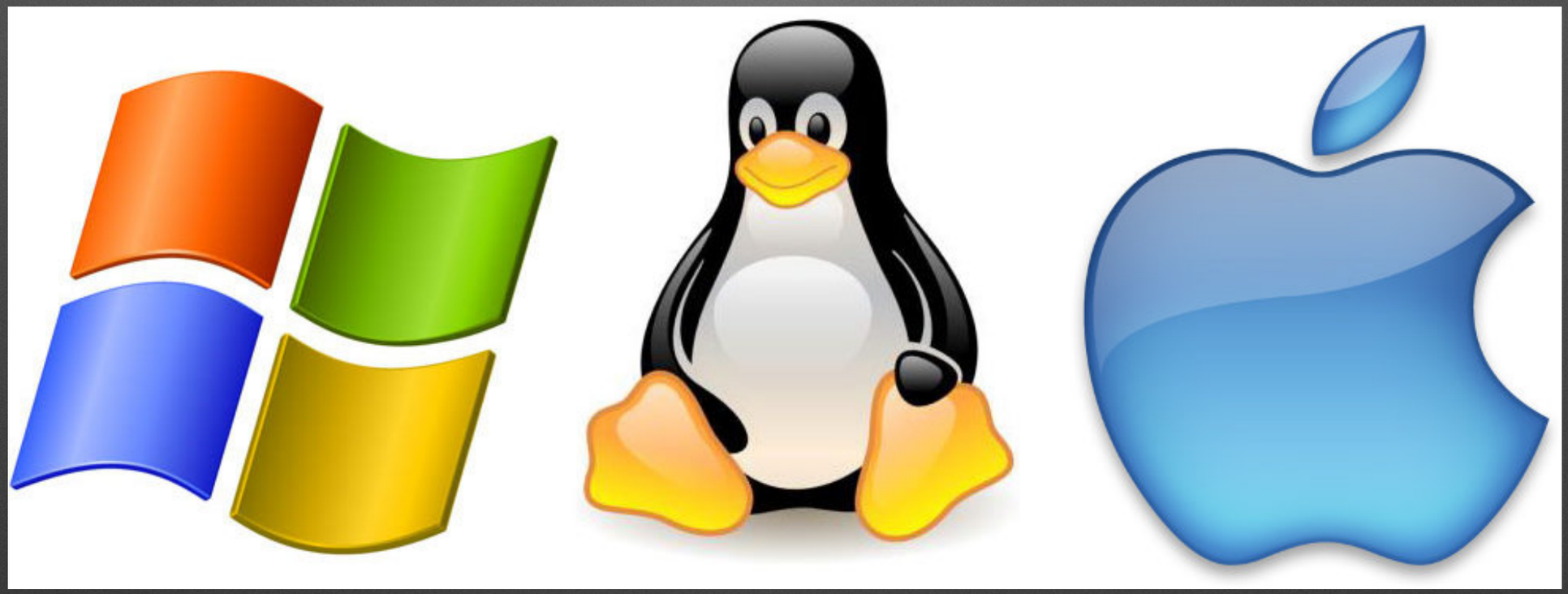
R packages


5000
4000
3000
2000
1000
0
Number of R Packages Available
2002
2004
2006
2008
2010
2012
2014
Date of R Version Release

R users


40,000
30,000
20,000
10,000
0
Traffic on E-mail Discussion Lists
1995
2000
2005
2010
2015
Year
R
Stata
SAS
SPSS

Cross platform



Interact with other software

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

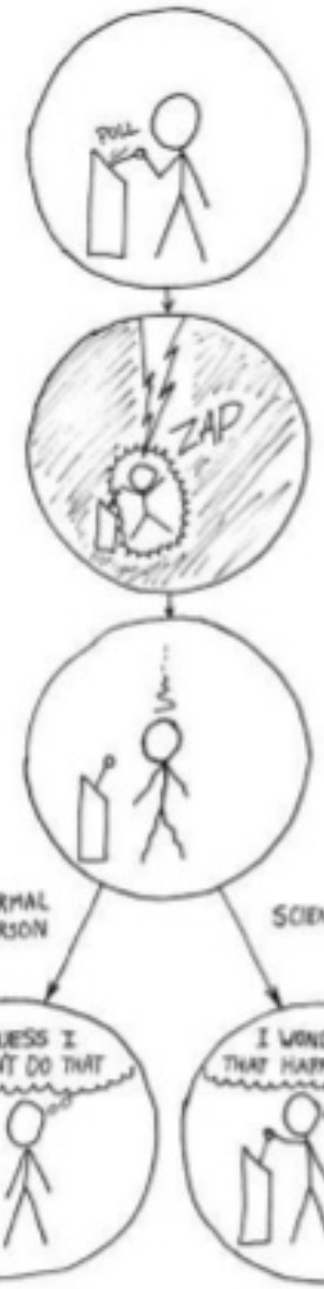
Reproduciblity

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

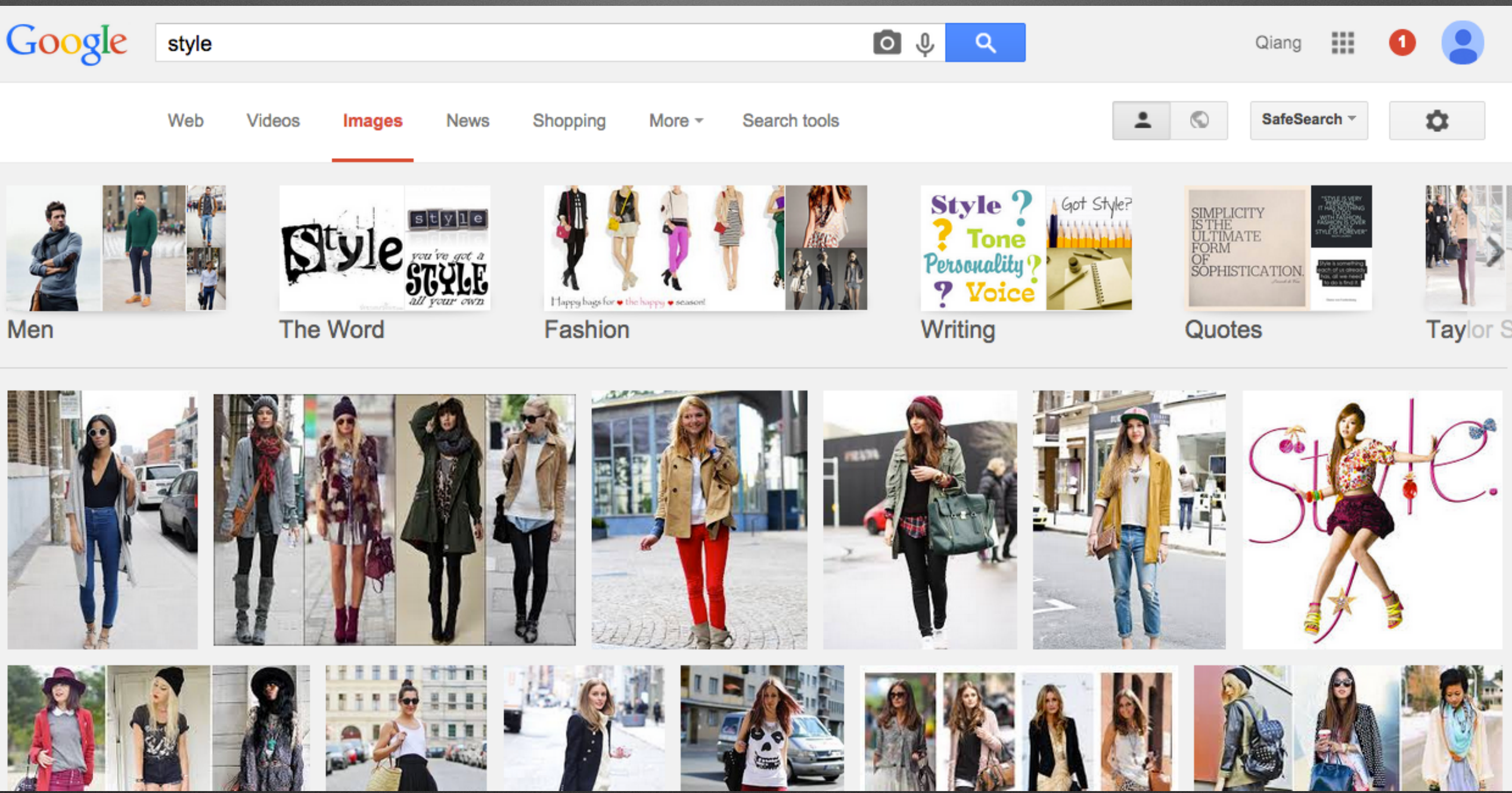
- A **process** for:
  - Sharing the method
  - Describing the environment
  - Recreating the results



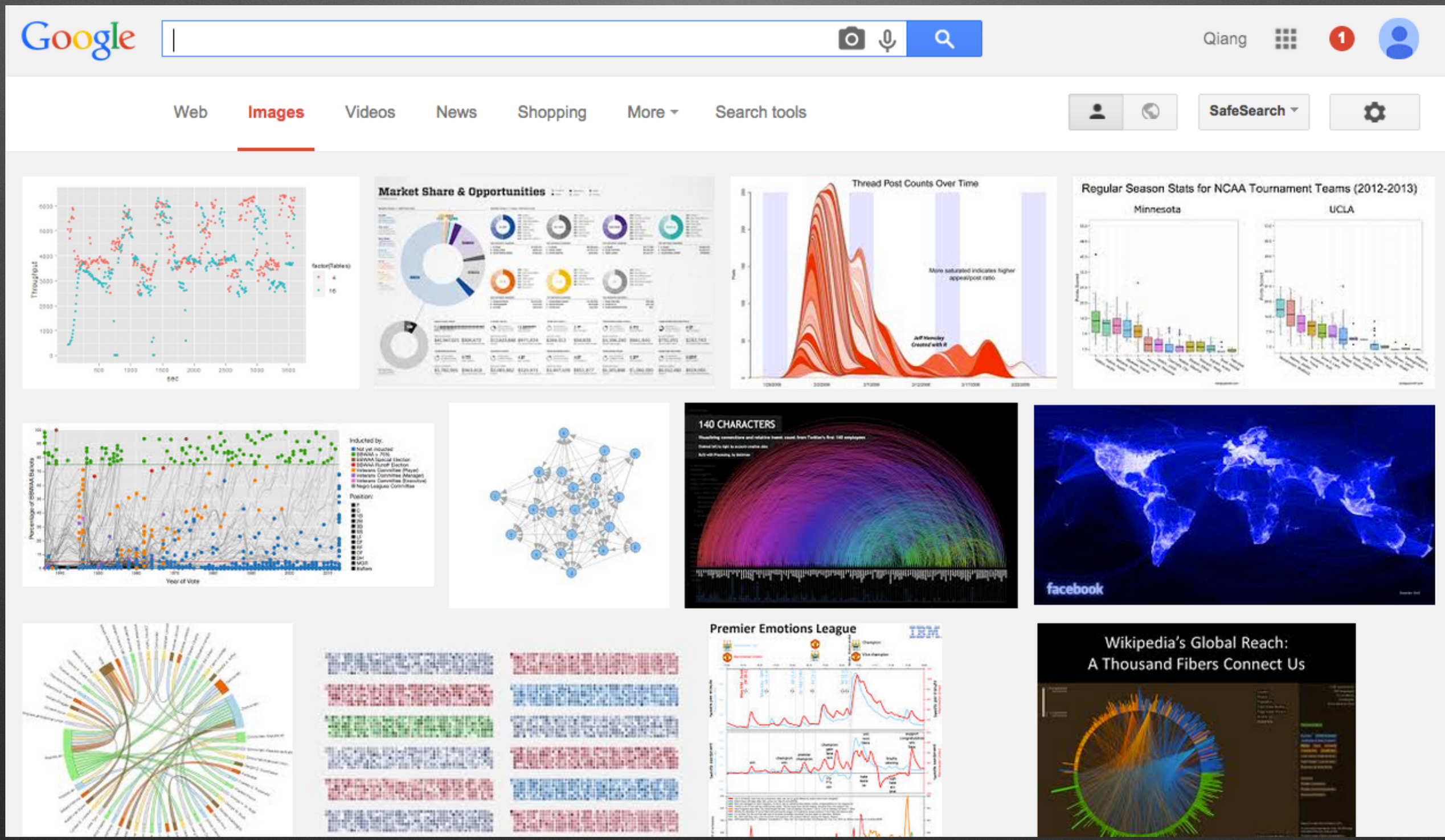
xkcd.com/242/

Style

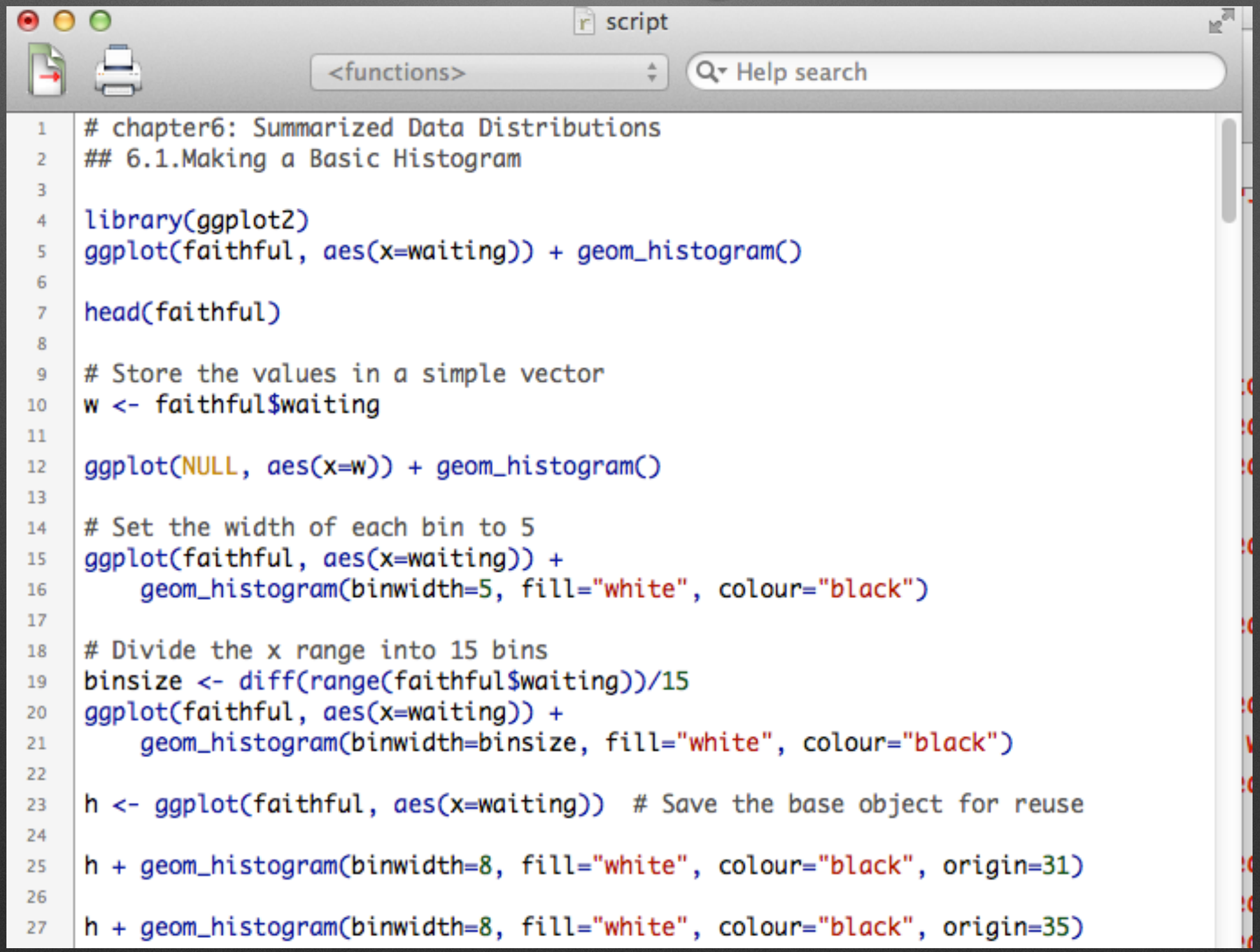
Not the style like this...



like this?

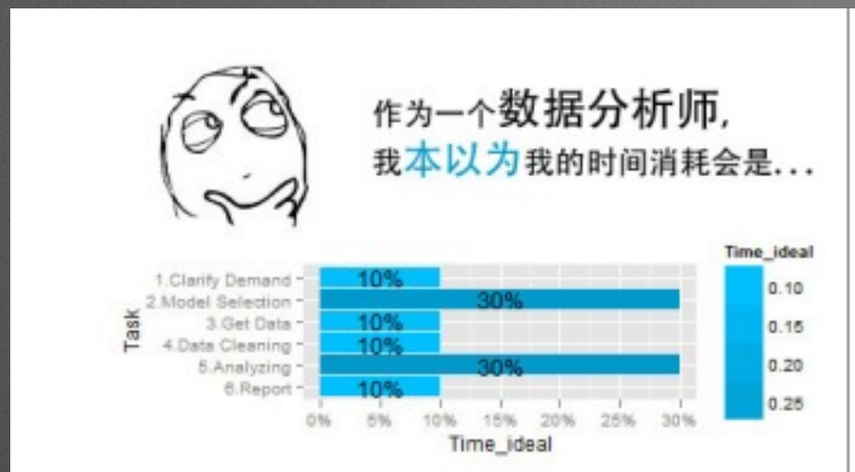

Google
Qiang
Web
Images
Videos
News
Shopping
More
Search tools
SafeSearch
Market Share & Opportunities
Thread Post Counts Over Time
Regular Season Stats for NCAA Tournament Teams (2012-2013)
Minnesota
UCLA
Inducted by:
Position:
Year of Vote
140 CHARACTERS
facebook
Premier Emotions League
IBM
Wikipedia's Global Reach:
A Thousand Fibers Connect Us

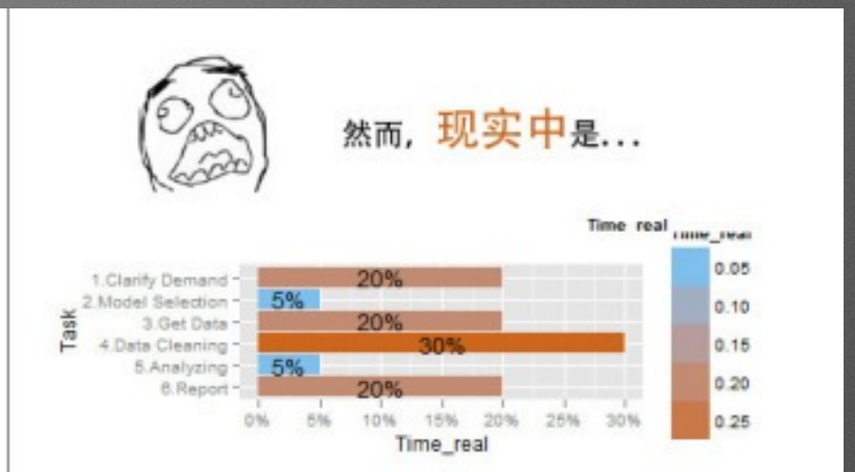
The reality is…



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

The engine

Rstudio
www.rstudio.org

The pretty face

Microsoft R Open
mran.microsoft.com/open/
Enhanced R distribution

Books

- **R in action. Robert Kabacoff. Manning Publications, 2015, 2nd 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