R language and data analysis

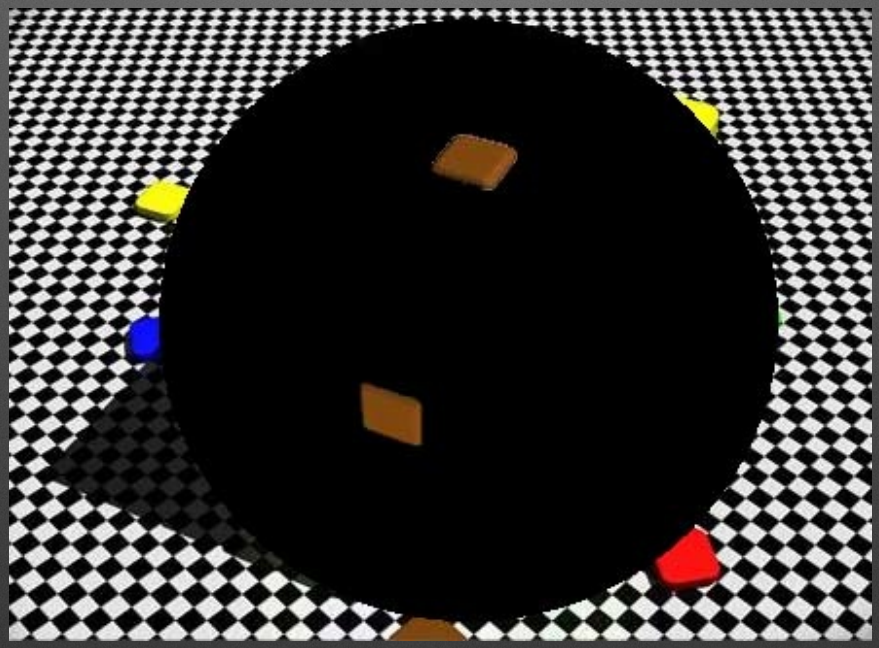
Qiang Shen (沈强)

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

Slides:https://github.com/johnsonzhj/R-course-2016
Teaching Assistant: Shenwei Wang(黄珅炜)
liutingyu2@sina.cn
18143467179

wifi: GLXY-106 psw:glxy123456789

Cognitive bias



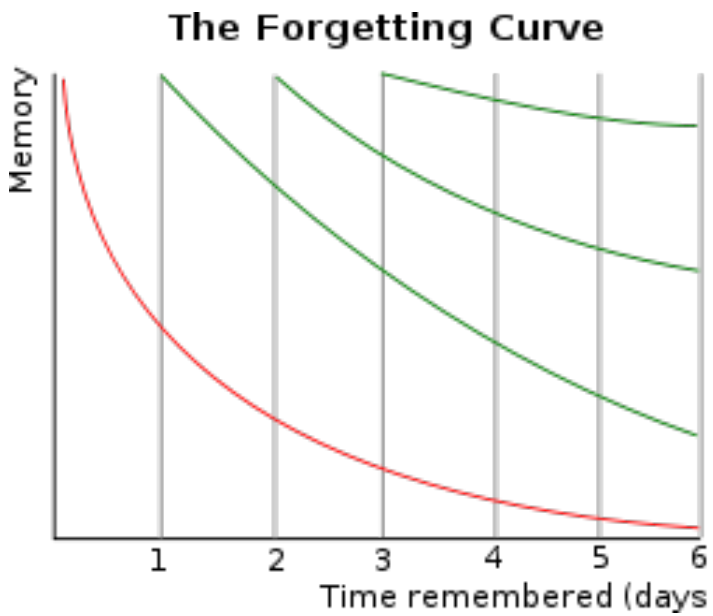
Projection bias

Projection bias is a feature in human thinking in which one thinks that others have the same priority, attitude or belief that one harbours even if that is unlikely.

Hindsight bias

This bias the inclination, after an event has occurred, to see the event as having been predictable, despite there having been little or no objective basis for predicting it.

H.Ebbinghaus



Grades

15% attendance of the course

48% homework 6*8

7% coordinate with Shenwei in advance.

（first comes first served rule）

30% Final work

Aim

- Use R to run your own analysis!
- Learning by doing.

Simple, flexible and useful!

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

Data science

Data science

浙江大学 党委组织部

首页 部门简介 组织设置 干部工作 组织工作 党校工作 文件汇编 党建知识 表格下载

您现在的位置：浙江大学党委组织部 >> 管理学院等学院下设系的公示

管理学院等学院下设系的公示

[点击数：239 | 更新时间：2016-4-1] 【字体：小 大】

经校学术委员会、校机构编制管理委员会研究，商校研究生院、本科生院，拟同意：

管理学院下设：创新创业与战略学系、数据科学与管理工程学系、服务科学与运营管理学系、市场营销学系、财务与会计学系、领导力与组织管理学系、农业经济与管理学系、旅游与酒店管理学系；

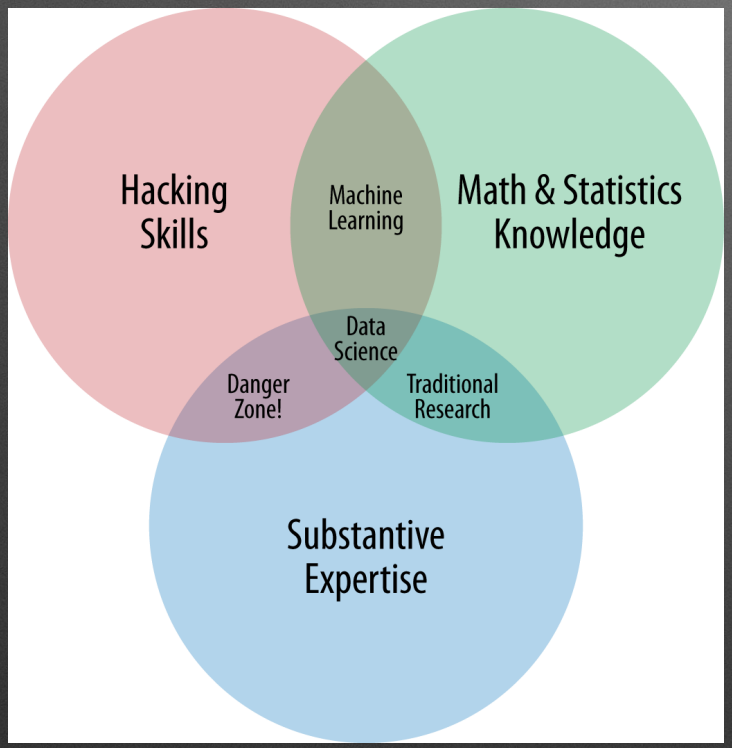
地球科学学院下设：地质学系、地理科学学系、大气科学学系、地球信息科学与技术系。

现予以公示，公示期为4月1日至4月8日。

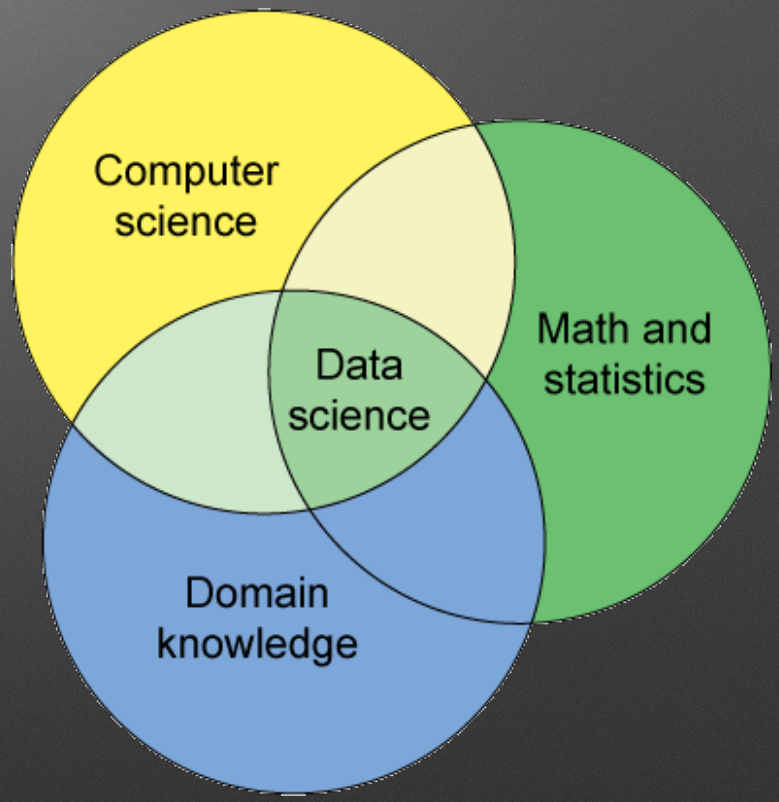
有意见或建议请联系：88981297，88981387

Email：zhangxiaojie@zju.edu.cn

Data science

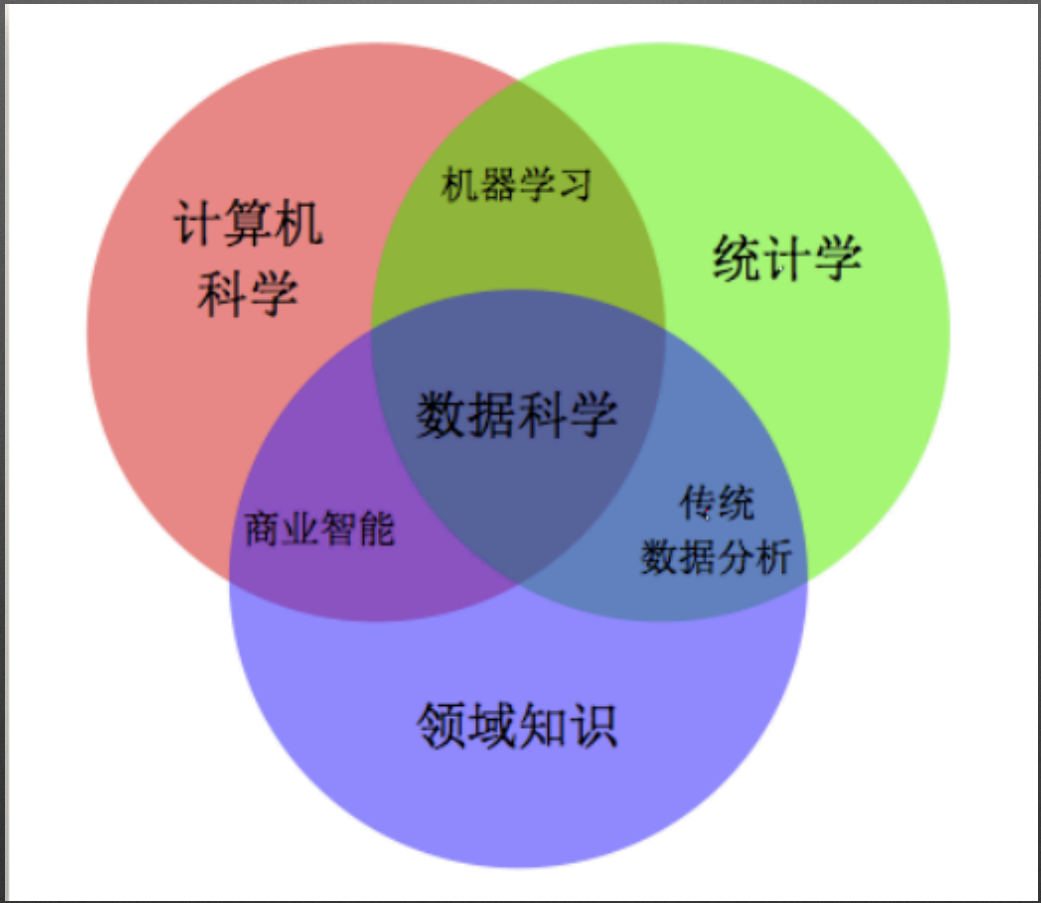


Drew Conway



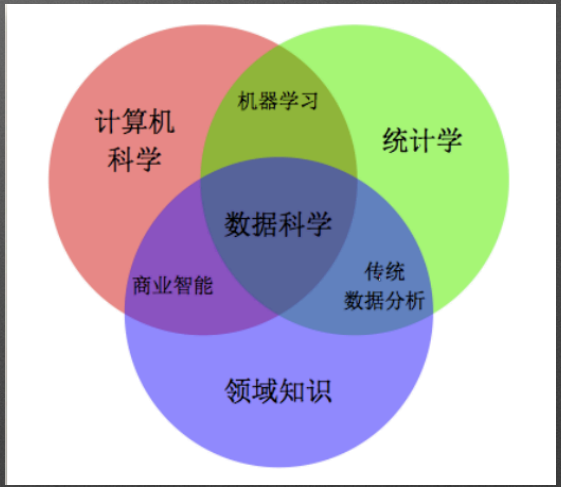
IBM

Data science



Data Science

- Statistics/Econometrics
- Computer science
- Domain knowledge
- Machine learning （S and C）
- Traditional data analysis (S and D)
- Business intelligence (C and D)



Business intelligence

“我是搞数据分析的，学会了如何从DW中用SQL对数据ETL并建立了Cube。然后算啊算啊算，得出结论：今年2月份营业收入远远小于其它月份。我试图用spss、sas中的数据挖掘模型找出原因但至今无果。扫地阿姨弱弱的说“2月份是春节，几乎所有的公司半个月无人上班…”我勒个去…… !

History of Data science

Concise Survey of
Computer Methods,1974

Peter Naur,1960

History of Data science

Statistics = data science?

C.F. Jeff Wu,1997

History of Data science

2008, DJ Patil and Jeff Hammerbacher used the term "data scientist" to define their jobs at LinkedIn and Facebook, respectively.

Data Scientists

Data science

- What is data science?
- Data Science is the extraction of **knowledge** from **large volumes of data** that are structured or unstructured which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD).
- 使用科学的方法从数据中获取知识的学科

Data science

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Big Data: The Management Revolution

Making Advanced Analytics Work for You

The Sexiest Job of the 21st Century is Tedious, and that Needs to Change

境界

第一重境界：昨夜西风凋碧树。独上高楼，望尽天涯路；

第二重境界：衣带渐宽终不悔，为伊消得人憔悴；

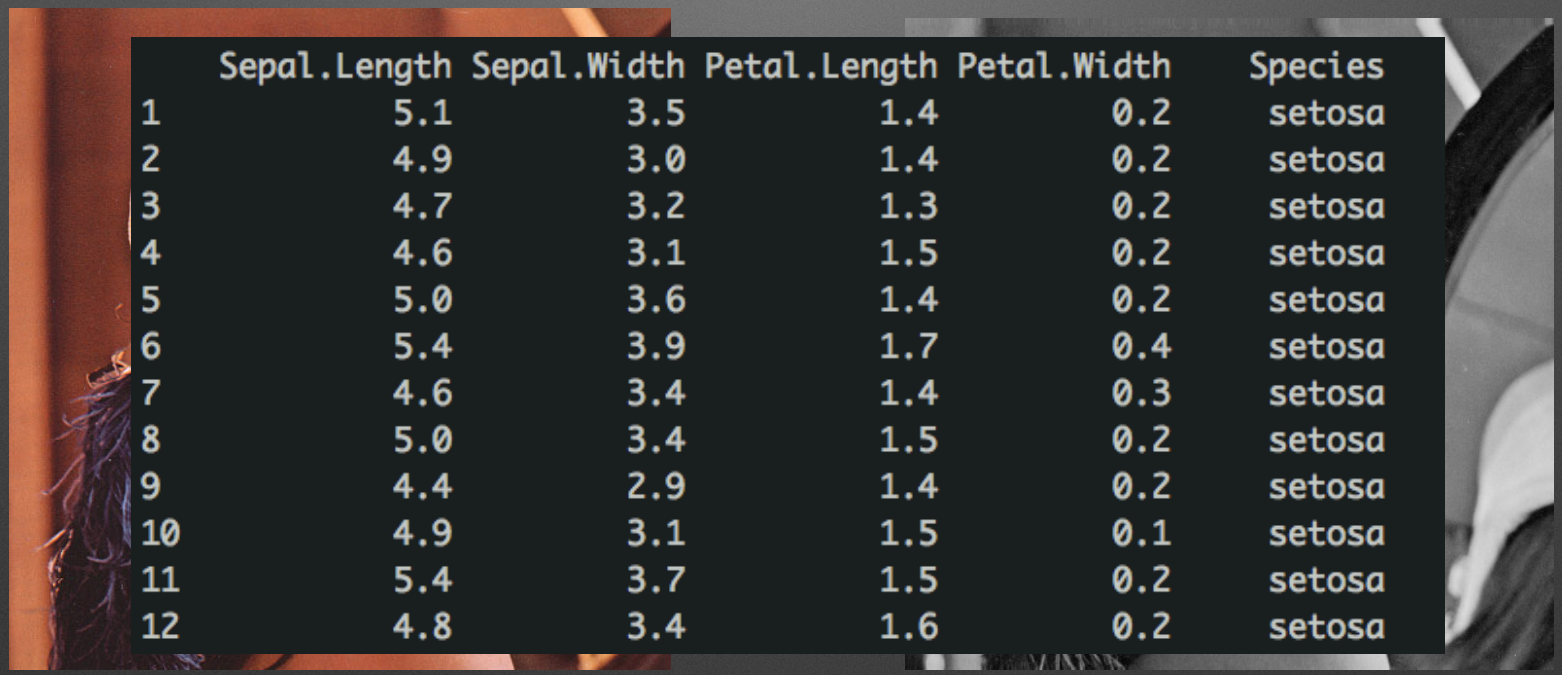
第三重境界：众里寻他千百度，蓦然回首，那人却在灯火阑珊处。

--王国维《人间词话》

境界

1. Run the analysis
2. Understand the principle
3. Create your own analysis： tourism

What is data?



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa

the First Lady of the Internet:Lenna

50th Anniversary IS&T conference

R YOU Ready?

- What?
- Why?
- How?

The very beginning mind itself is the most accomplished mind of true enlightenment

—-Avatamsaka Sutra（华严经）

What is R?

- R letter
- R movie ratings ?

Robert Gentleman and Ross Ihaka

What’s R?

R language

- R is a language and environment for statistical computing and graphics

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

Why R?

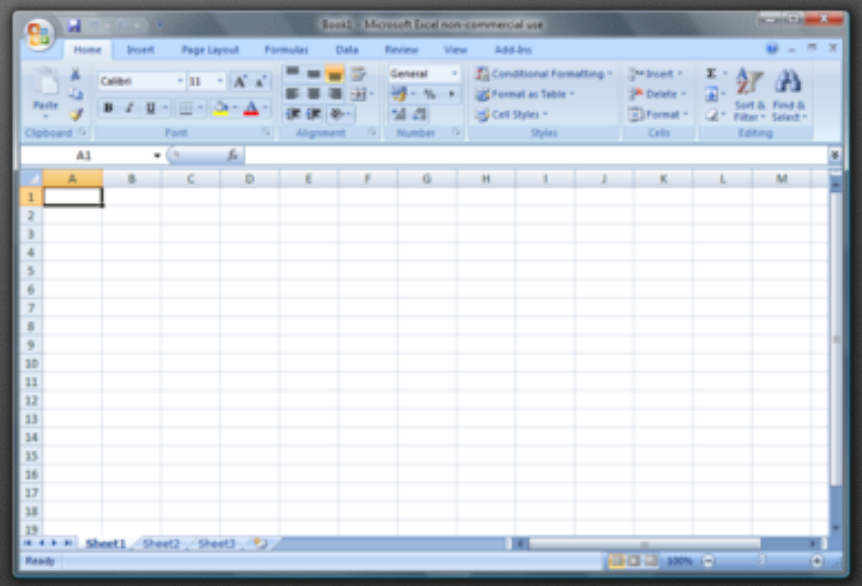
- Statistics: SAS
- Computer science: Fortran
- Domain knowledge：Excel (VBA)
- Machine learning: Python(scikit-learn)
- Traditional data analysis (SPSS)
- Business intelligence (RapidMiner,SAP, ORACLE, IBM)

Why R?

- Excel
- SPSS
- Stata
- SAS

disdain chain

Excel: Domain knowledge

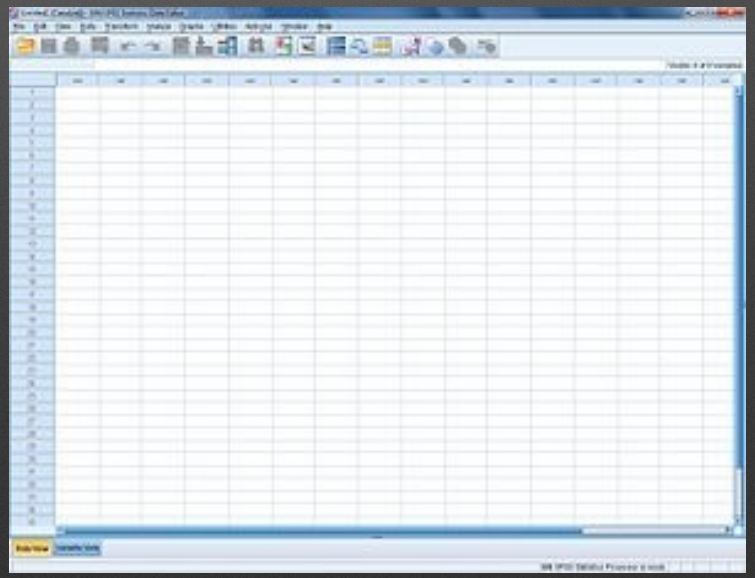


VBA
Regression: Analysis ToolPak
Mac:StatPlus

SPSS:Traditional data analysis

SPSS （Statistical Product and Service Solutions)

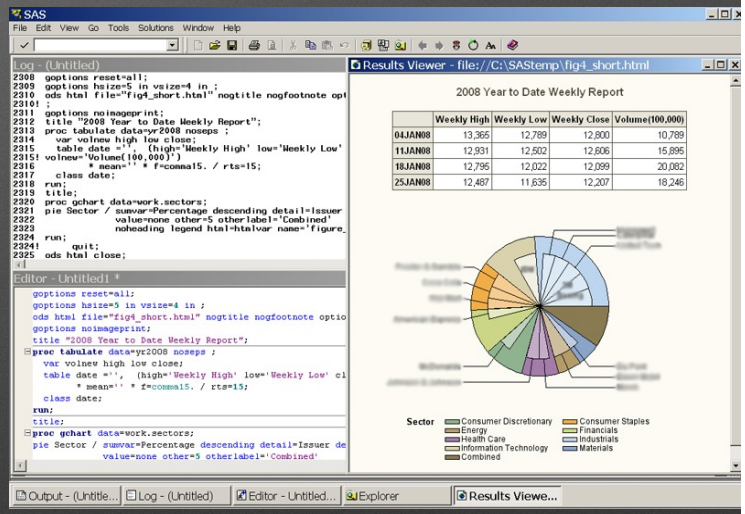
medicine, psychology and social sciences



SAS:Statistics

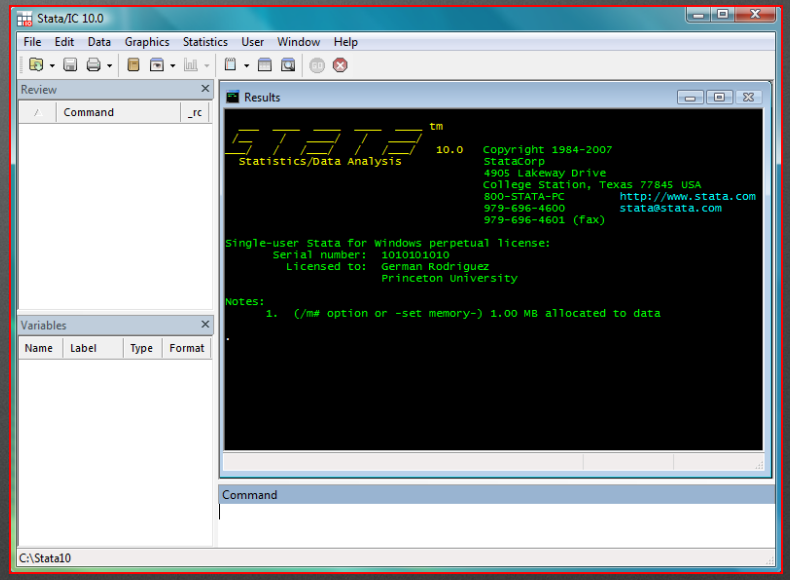
SAS is short form Statistical Analysis System。

权威、许多美国公司使用。
价格不菲、占用空间大
尽管现在已经尽量“傻瓜化”，仍然需要一定的训练才可以进入。对于基本统计课程则不那么方便。



Stata: econometrics

Stata



Python: machine learning

scipy
numpy
Pandas
Matplotlib
scikit-learn

anaconda

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

What is R?

R is a dialect of S

History of S

S version 1	John Chambers 1976:Fortran
S version 2	1978
S version 3	1983-1992 rewrite in C 1988
S version4	1995

S won the Association for Computing Machinery(ACM)'s Software System Award in 1998.

History of S

Mathsoft:S-plus	1993
↓	
Insightful: Lucent	2004
TIBCO:insightful	2008

Philoshphy of S

- *“We wanted users to be able to begin in an interactive environment,where they did not consciously think of themselves as programming.Then as their needs became clearer and their sophistication increased,they should be able to slide gradually into programming, when the language and system aspects would become more important.”*

--John Chambers

History of R

- 1991: Created by Ross Ihaka and Robert Gentleman.
- 1993: First announcement of R to the public.1995: GNU General Public License to make R free software.
- 1997: The R Core Group is formed (20 people).
- 2000: R version 1.0.0 is released.
- 2016.9: R version 3.3.1

Features of R

- free and open source
- visualisation
- multiple data source
- fast upgrade and packages
- cross platform
- Reproducible
- Interact with other software

Free and open source!

Reliable?

The New York Times

Business Computing

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Data Analysts Captivated by R's Power

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html

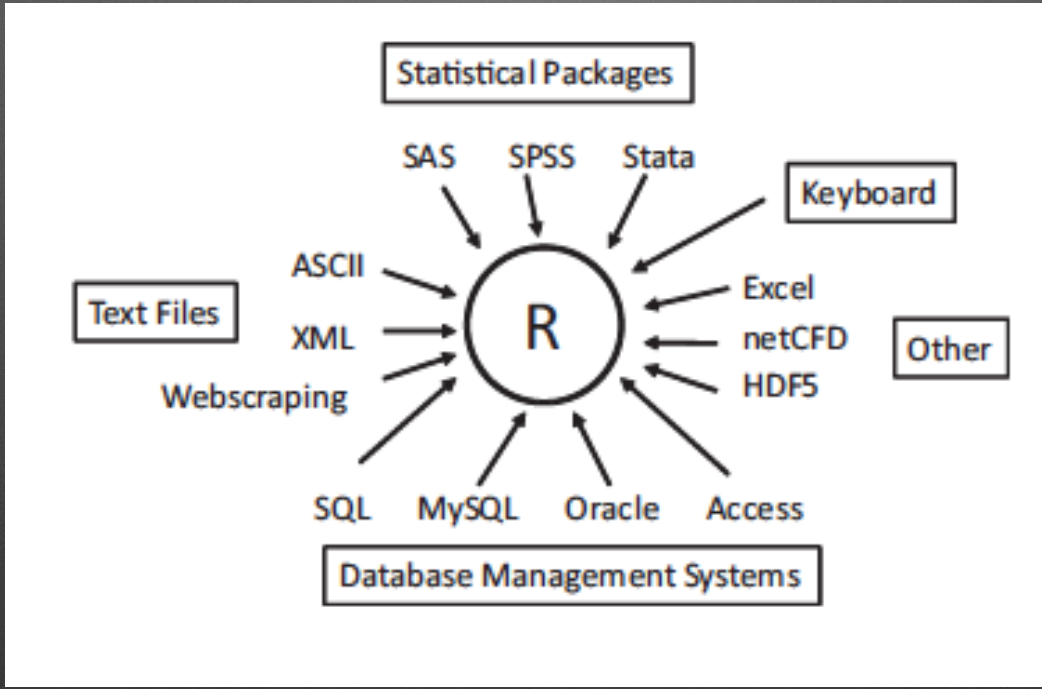
Visualization

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

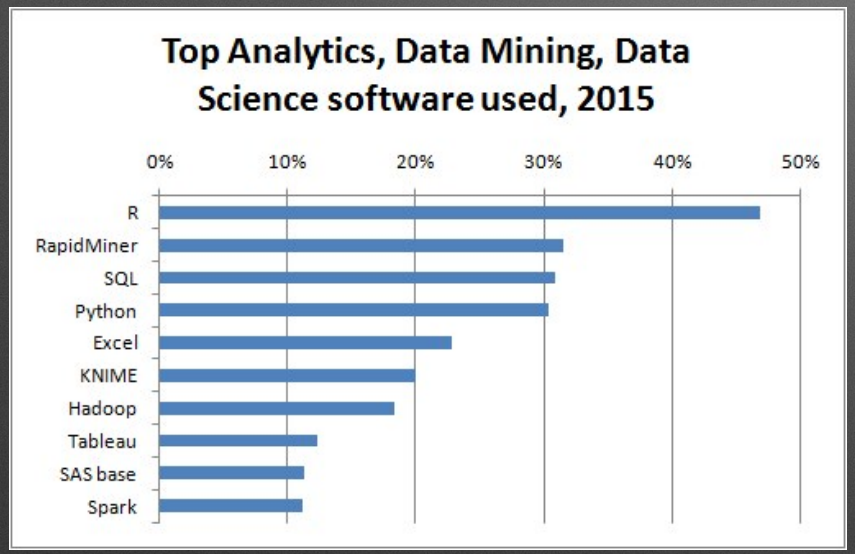
SSEC

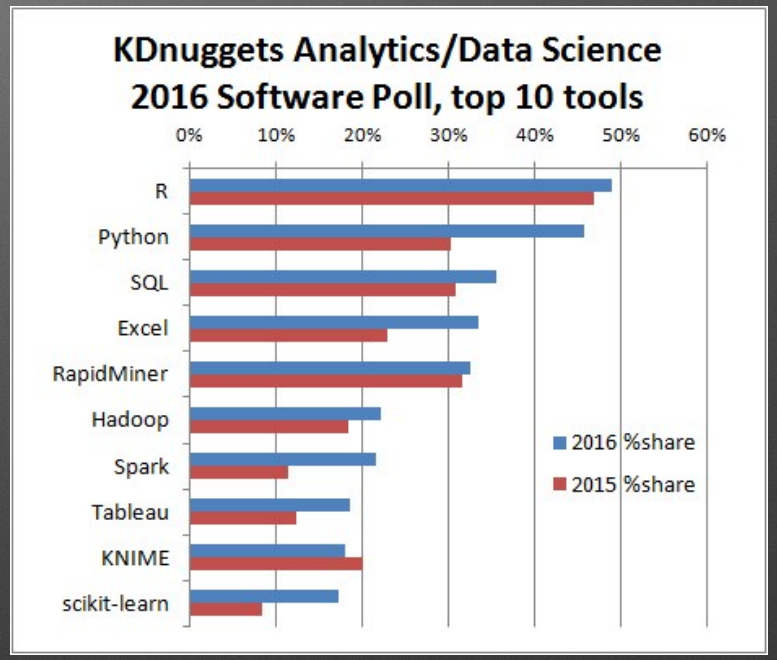
SSEC.m
[2005-01-01 08:00:00/2015-09-01 08:00:00]
Last 3114.798096
6000
5000
4000
3000
2000
1000
Volume (millions):
2,642,300
12
10
8
6
4
2
0
Jan 2005
Jan 2006
Jan 2007
Jan 2008
Jan 2009
Jan 2010
Jan 2011
Jan 2012
Jan 2013
Jan 2014
Jan 2015

multiple data source



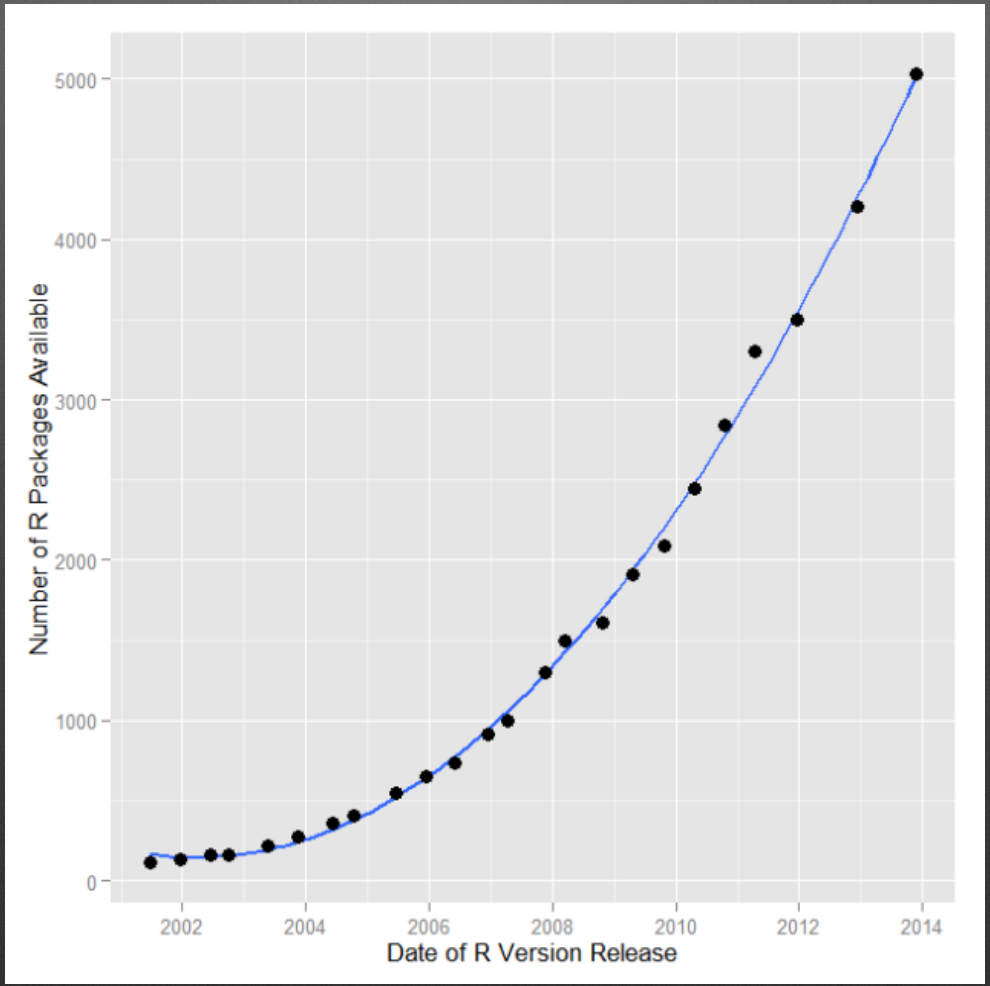
The trend for R



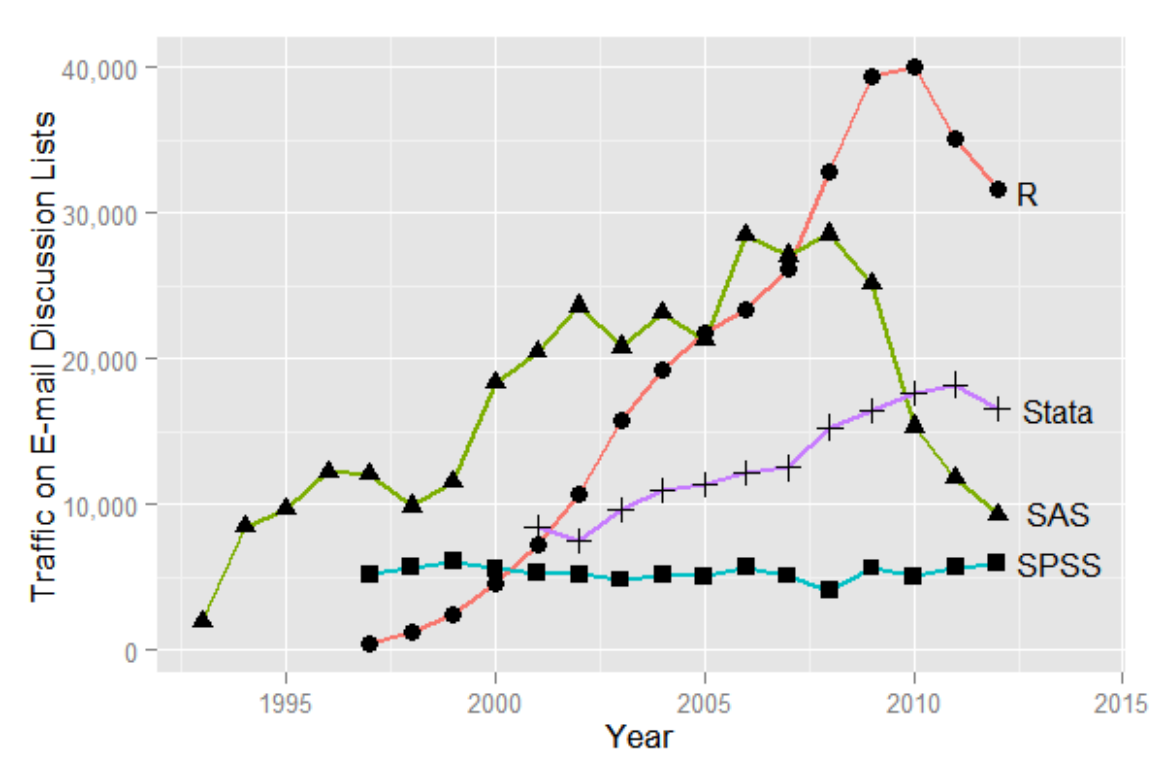


kdnuggets

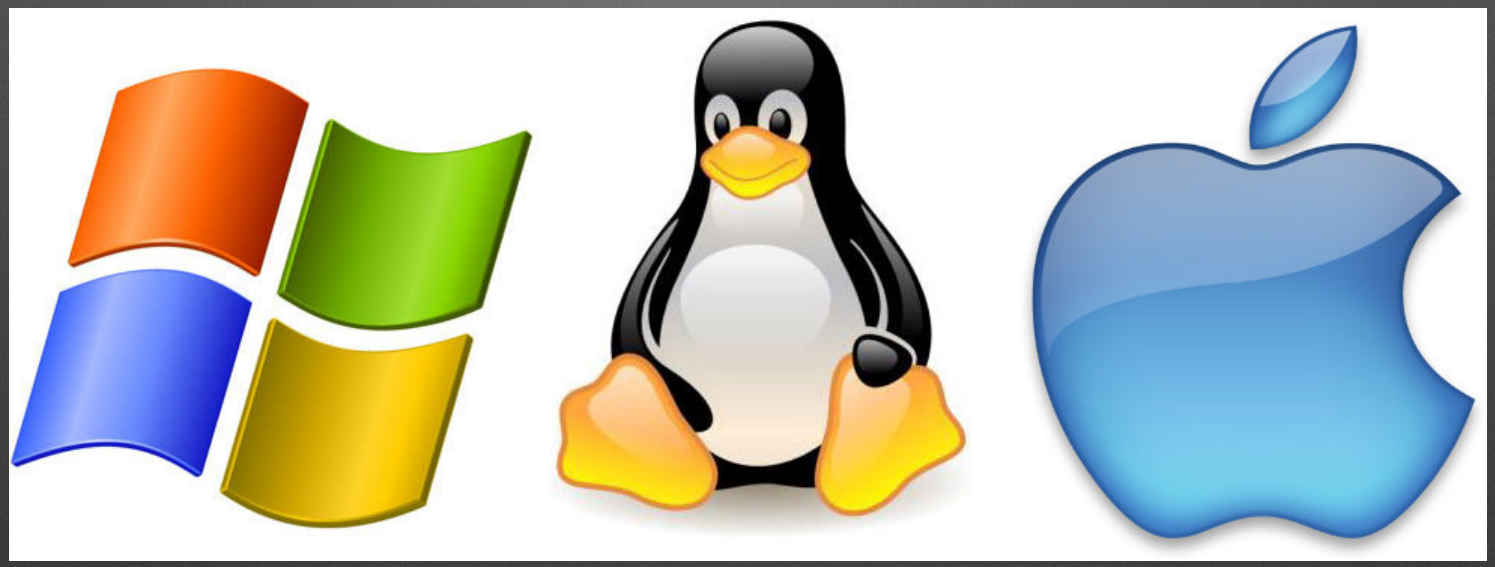
R packages


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

R users


Traffic on E-mail Discussion Lists
40,000
30,000
20,000
10,000
0
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

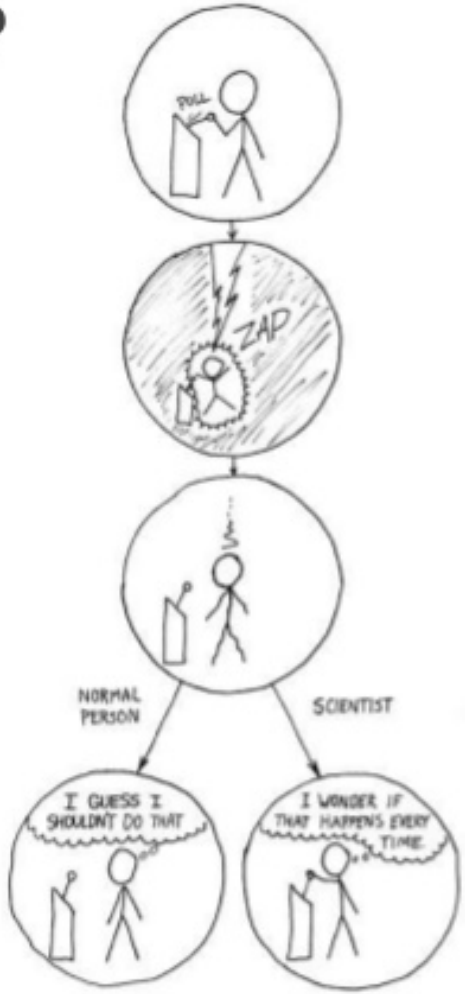
Reproduciblity

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/

Disadvantage of R

- **Not for dummies: Need to program!!!**

black box

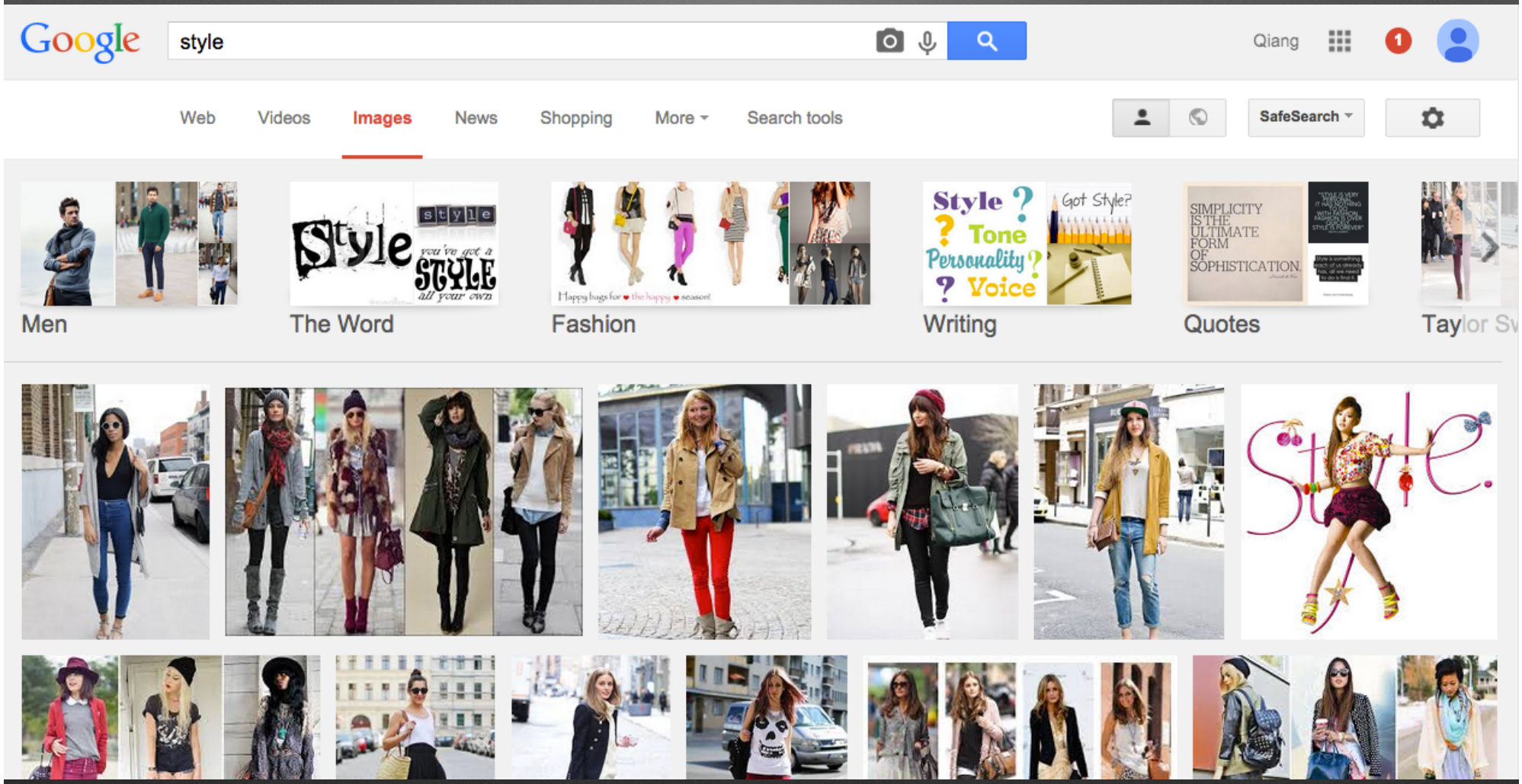
import———————> export

Disadvantage of R

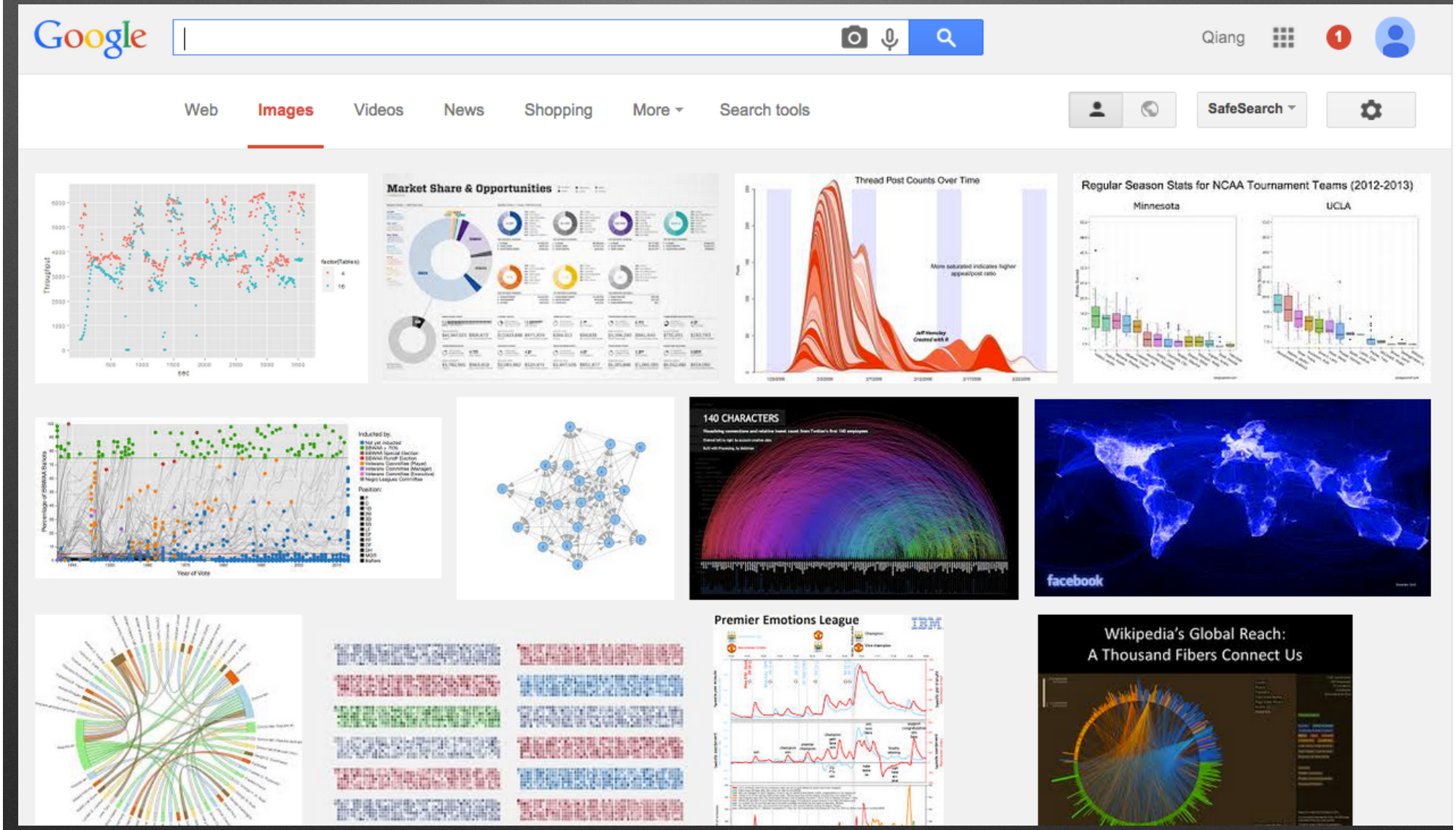
- Efficiency(non-parallel): interpreted programming language,single threaded
- Power (vs. SAS): memory bound
- As a Language, it is not perfect

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
140 CHARACTERS
facebook
Premier Emotions League
IBM
Wikipedia's Global Reach:
A Thousand Fibers Connect Us

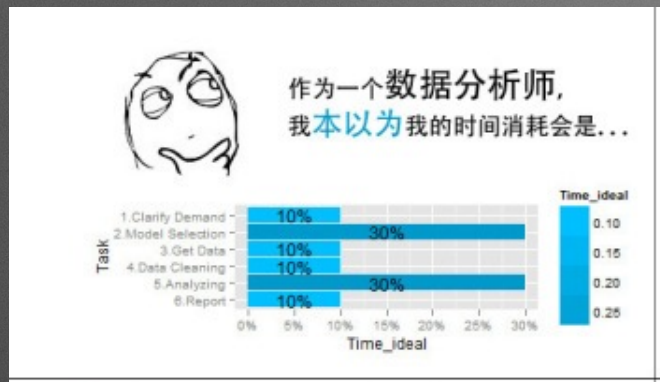
The reality is…

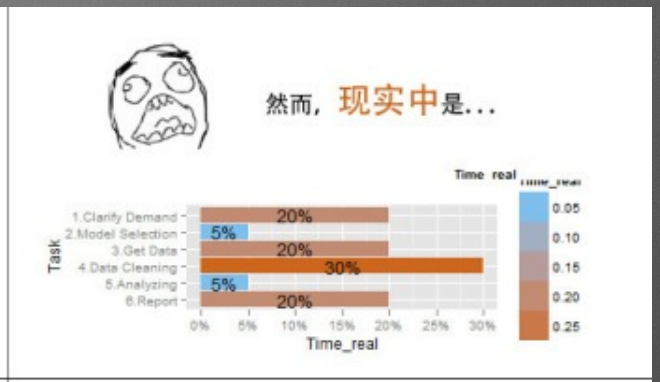
script

<functions> Help search

```r
# chapter6: Summarized Data Distributions
## 6.1.Making a Basic Histogram

library(ggplot2)
ggplot(faithful, aes(x=waiting)) + geom_histogram()

head(faithful)

# Store the values in a simple vector
w <- faithful$waiting

ggplot(NULL, aes(x=w)) + geom_histogram()

# Set the width of each bin to 5
ggplot(faithful, aes(x=waiting)) +
    geom_histogram(binwidth=5, fill="white", colour="black")

# Divide the x range into 15 bins
binsize <- diff(range(faithful$waiting))/15
ggplot(faithful, aes(x=waiting)) +
    geom_histogram(binwidth=binsize, fill="white", colour="black")

h <- ggplot(faithful, aes(x=waiting))  # Save the base object for reuse

h + geom_histogram(binwidth=8, fill="white", colour="black", origin=31)

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 You Ready?!

Practice make perfect！

R
www.r-project.org

The engine

Rstudio
www.rstudio.org

The pretty face

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

- Statistical software: R user
- Programming language: R developer