



数据科学入门 – 让数据思维成为生活的一部分

丘祐玮 – David Chiu

EMAIL: david@largitdata.com

网站: www.largitdata.com

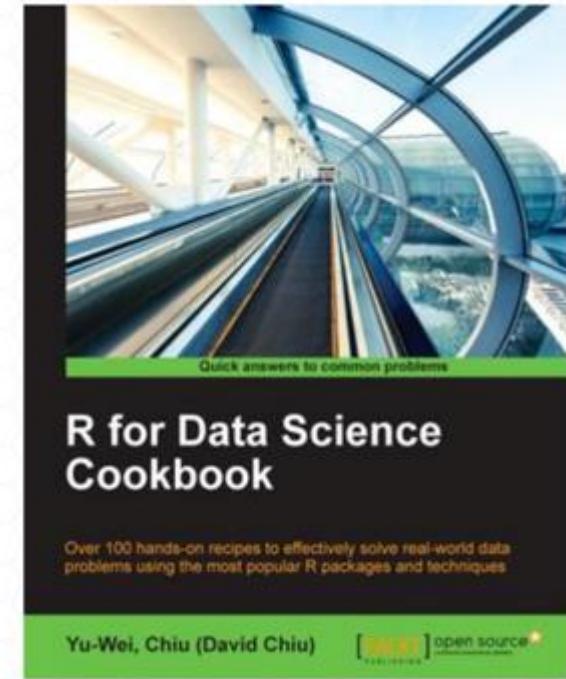
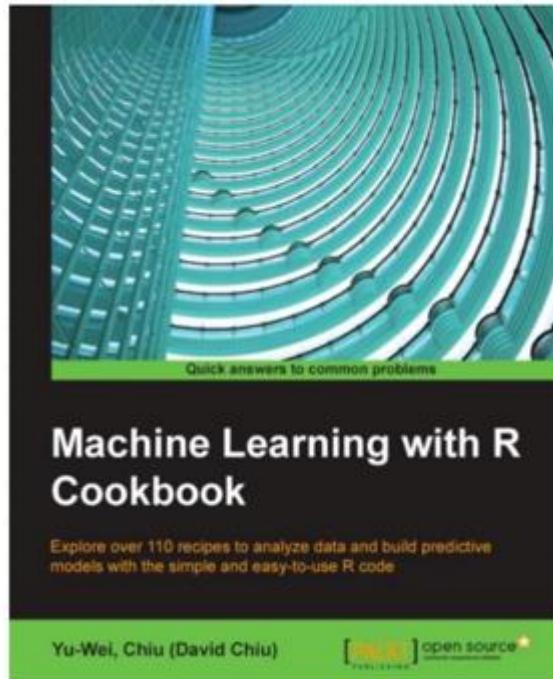
电话: +886929094381

关于我



- 大数软件有限公司创办人
- 前趋势科技工程师
- ywchiu.com
- 大数学堂
<http://www.largitdata.com/>
- 粉丝页
<https://www.facebook.com/largitdata>
- R for Data Science Cookbook
<https://www.packtpub.com/big-data-and-business-intelligence/r-data-science-cookbook>
- Machine Learning With R Cookbook
<https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r-cookbook>

Machine Learning With R Cookbook (机器学习与R语言实战) & R for Data Science Cookbook



Author: Yu-Wei (David) Chiu





什么是数据科学？



数据

科学

數據科學

哪家公司没有数据？
哪家公司不科学了？



50嵐

靠感觉

靠经验?



雇人拿计数器

计算不同时间点的人潮流量

用数据发现问题

用数据解决问题



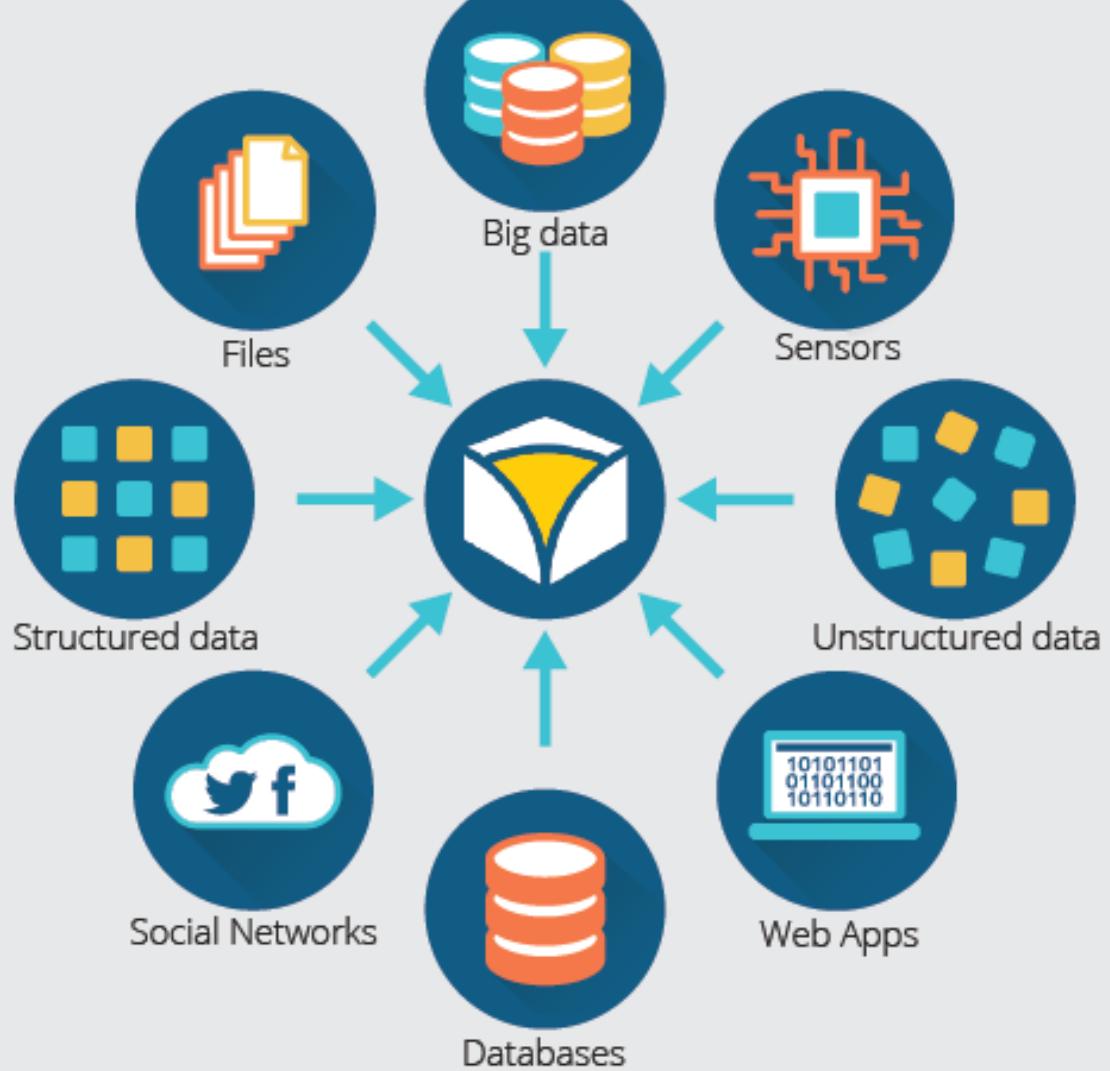
雲端計算的 興起



企業數據

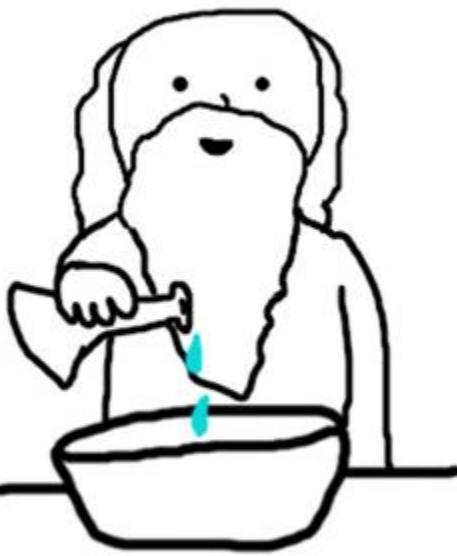
物聯網的發
展





如何快速整合分析不同的资料源？

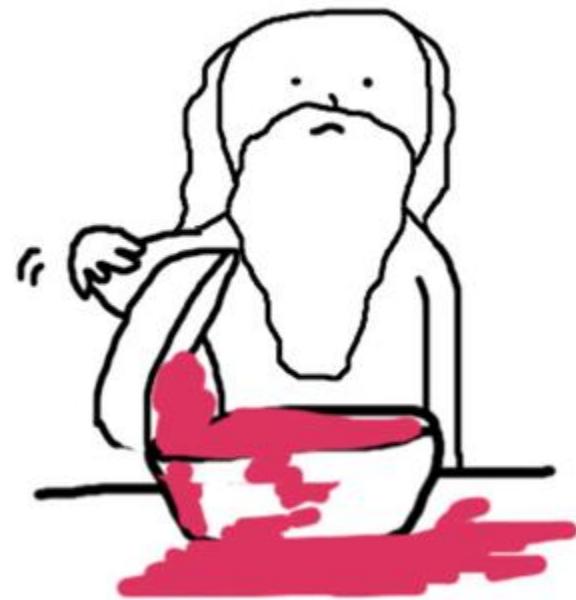
加一点数学统计



加一点工程



产生数据科学



要懂一点工程的统计学家

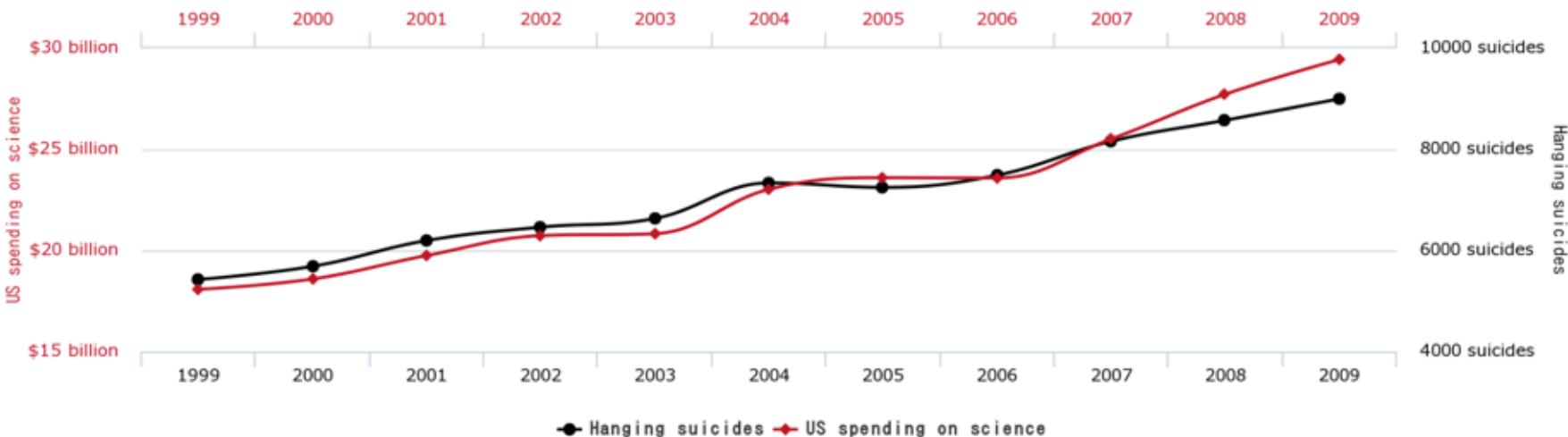
要懂一点统计的工程师



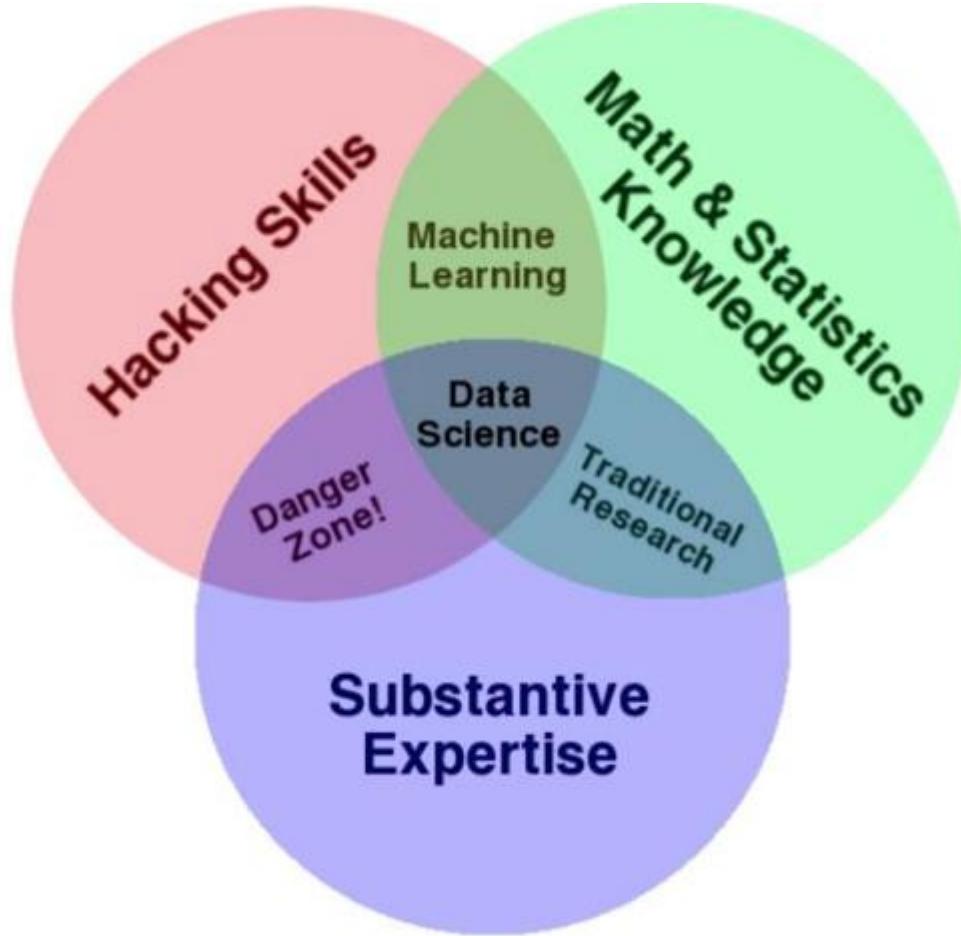
只有工程统计还不够

Spurious correlations

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



数据科学



数据科学能力

- 统计 (Statistic)

单变量分析、多变量分析、变异数分析

- 资料处理 (Data Munging)

抓取数据、清理数据、转换数据

- 数据可视化(Data Visualization)

图表、商业智能系统





WE NEED A HERO!



软件工程师

学习统计并了解
如何诠释结果

数据库管理员
如何处理
非结构化数据

统计学家

要知道如何处理
巨量数据问题

商业分析师

了解算法并知道在
不同数据量下该使
用何种工具

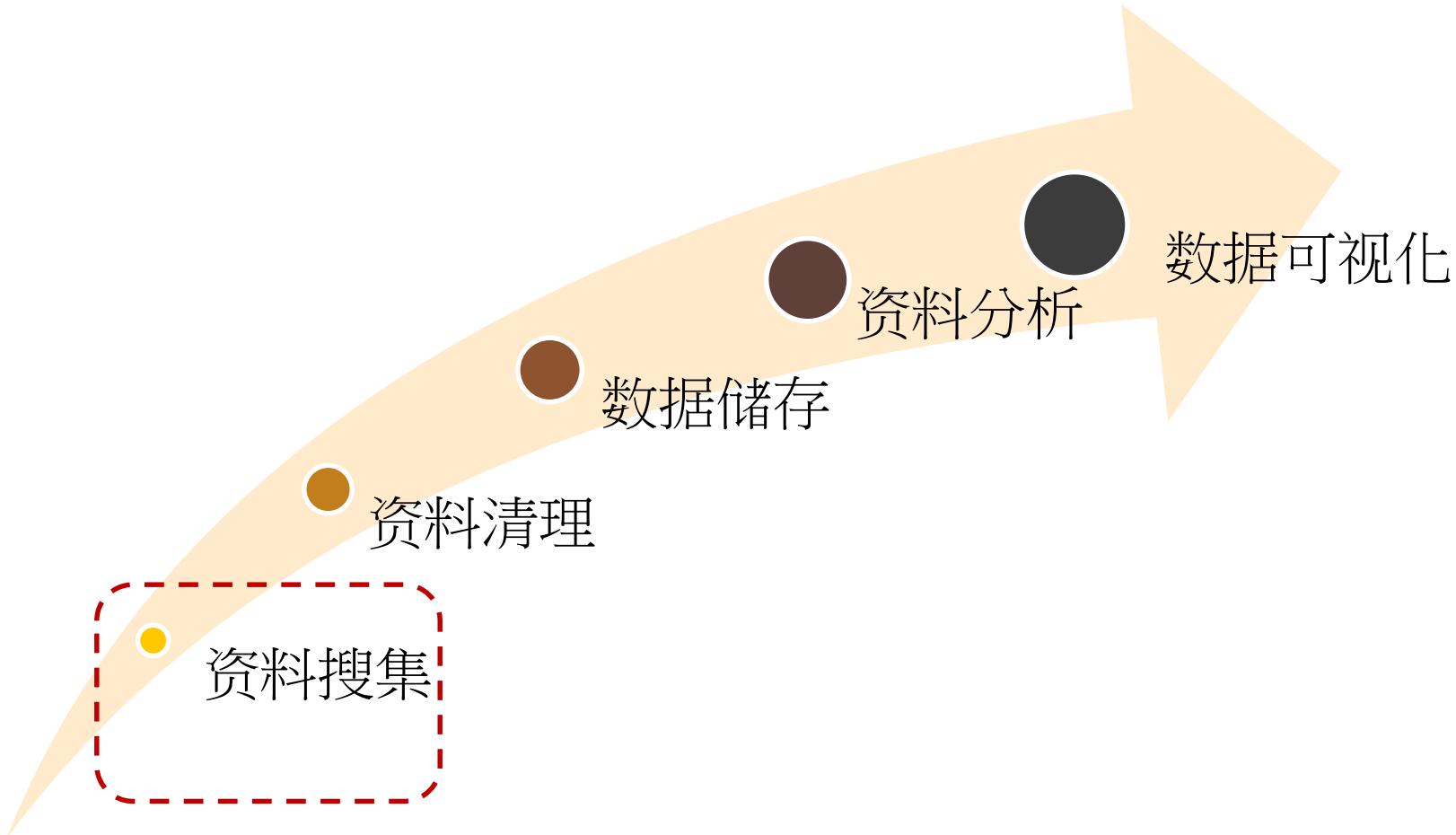


解构! 探讨数据科学流程

使用数据拟定策略



数据科学步骤



资料搜集

结构化数据

- 每笔数据都有固定的字段、固定的格式，方便程序进行后续取用与分析
- 例如：数据库

半结构化资料

- 数据介于结构化数据与非结构化数据之间
- 资料具有字段，也可以依据字段来进行查找，使用方便，但每笔资料的栏位可能不一致
- 例如：XML, JSON

非结构化资料

- 没有固定的格式，必须整理以后才能存取
- 没有格式的文字、网页数据



结构化数据

- 数据有固定的字段与格式

例如：数据库表格中所存放的资料

id	title	content	time	view_cnt	category
56	Uni-girls成立經紀公司 將選拔5位...	統一7-ELEVEn獅隊旗下Uni...	2016-06-17 16:11:00	0	體育
57	販售工業用劣質油 盡好企業判6...	盡好企業負責人吳容合被...	2016-06-17 16:10:00	0	社會
58	義大回嗆中職會長：聯盟應深切檢討	中華職棒會長吳志揚今天...	2016-06-17 16:09:00	0	體育
59	【就職滿月】禁英文滿意度4成7 ...	總統禁英文就職即將滿月...	2016-06-17 16:08:00	0	政治
60	【有片】索珠之亂有續集 洪家寶... (更新：新增影片)自稱「...	(更新：新增影片)自稱「...	2016-06-17 16:08:00	36312	社會
61	昔日老毛鬥爭工具 「中央專案組...	香港銅鑼灣書店店長林榮...	2016-06-17 16:08:00	0	國際
62	【有片】北市府給雪隧建議 林聰...	(更新:新增動新聞)台北市...	2016-06-17 16:07:00	5067	政治

- 可以使用SQL 处理与捞取数据

select title, content from newsmain;

半结构化数据

XML

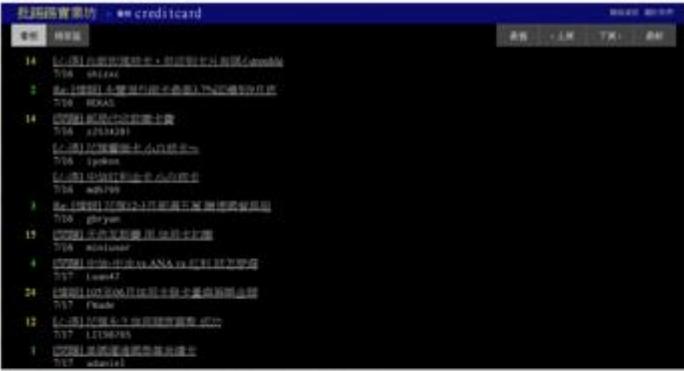
```
<users>
  <user>
    <name>QOO</name>
    <gender>M</gender>
    <age>12</age>
  </user>
  <user>
    <name>Mary</name>
    <gender>F</gender>
  </user>
</users>
```

JSON

```
[ user:{ name:QOO, gender:M, age:12 },
  user:{ name:Mary, gender:F } ]
```



网页资料

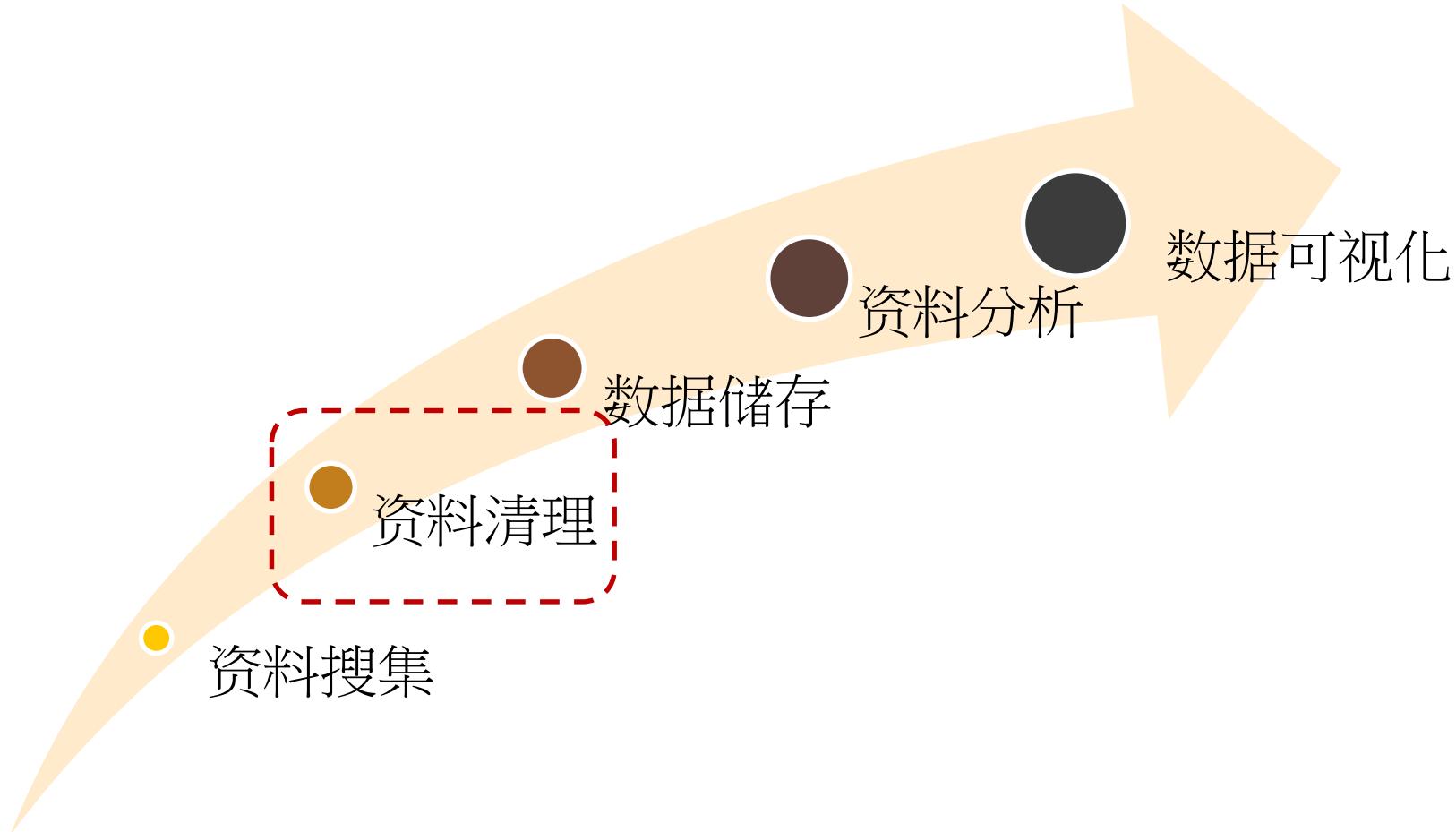


將非結構化的網頁資料
轉成結構化資訊

time	title	category
1 16:55	印度天橋坍塌 已知2死多人被困(0)	國際
2 16:52	異議人士高喊北京住進 聖母遭強拆(0)	國際
3 16:50	【更新】50元偽幣流竄 台南管查獲6萬枚(157224)	社會
4 16:50	【特企】一份用愛傳遞的禮物 當愛生命發約(2911)	特企
5 16:50	【法廣RFI】安倍參加核峰會欲與習近平會... (0)	國際
6 16:48	喬治亞州死刑犯 國際注射毒針(45)	國際
7 16:46	【更新】銷情藥物卻作劇梵化籤員工 188...(2428)	社會
8 16:43	飭令取緝 學者批評者：會動聽的結構式... (166)	生活
9 16:40	【TOMO要語錄】「這裡台北號」斷彩(藏身... (253)	國際
10 16:40	駁打蓮將標誌刪除 湘雲嘆「努力還是不夠」(2171)	娛樂
11 16:39	【有片】身體因癌法返台？ 痘敏憲以色列... (38642)	政治
12 16:35	【10種可測體力大！】花錢都買不到(562)	財經
13 16:33	【法廣RFI】富家女被拐賣 港府被指扶手... (330)	國際
14 16:32	檢討新聞發展辦公室：經典：審時注必要(135)	財經



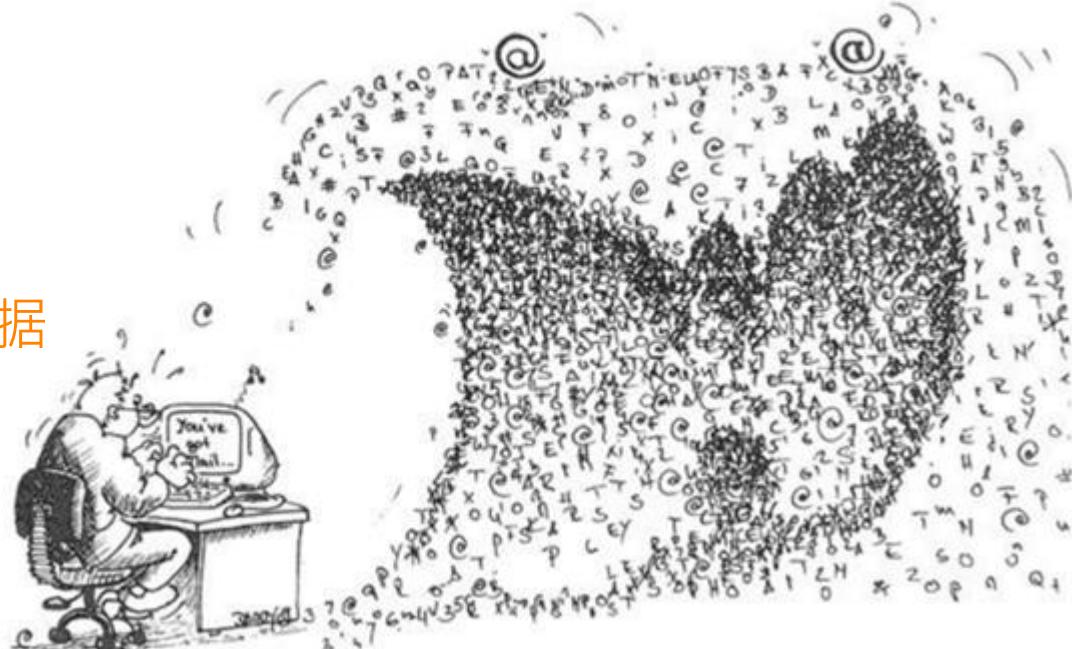
数据科学步骤



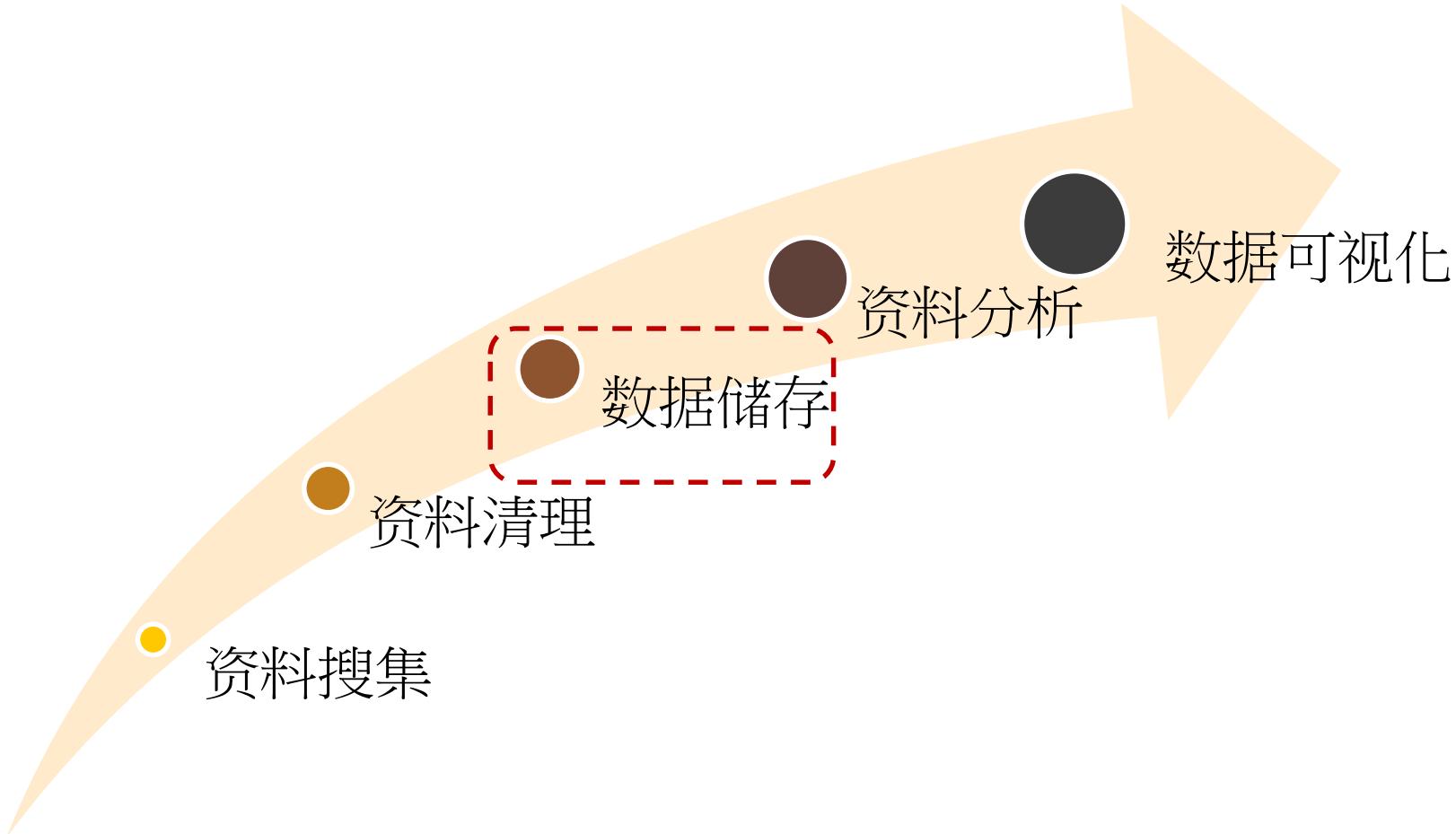
资料清理二三事

80% 的时间都在清理数据

- 资料筛选
- 侦测遗失值
- 补齐遗失值
- 资料转换
- 处理时间格式数据
- 重塑资料
- 学习正规运算式
- ...



数据科学步骤



关系数据库

安全存储、管理数据

- 有效管理磁盘上的数据

保持数据的一致性

可以透过标准模型整合数据

- 使用SQL 操作数据



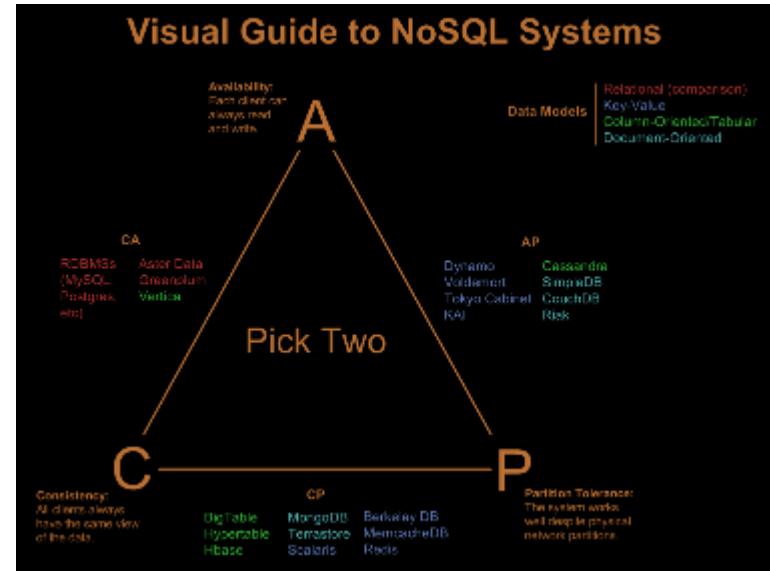
非关系数据库

NoSQL (Not Only SQL)

或是指相对于关系数据库而言(MySQL 或 PostgreSQL) 的非关系数据库

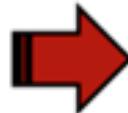
特性

- 可以将数据平行分散于丛集中
- 不使用关联性模型 (Schema Less)
- 极具成本效益 (Cost Effective)
- 开源 (Open Source)



NoSQL 应与关系数据库机制并行

- Linear Scalability
- Schema flexibility
- High Performance



NoSQL

- Multi-document transactions
- Complex security needs
- Complex joins
- Extreme compression needs



RDBMS

- Both / depends on the data



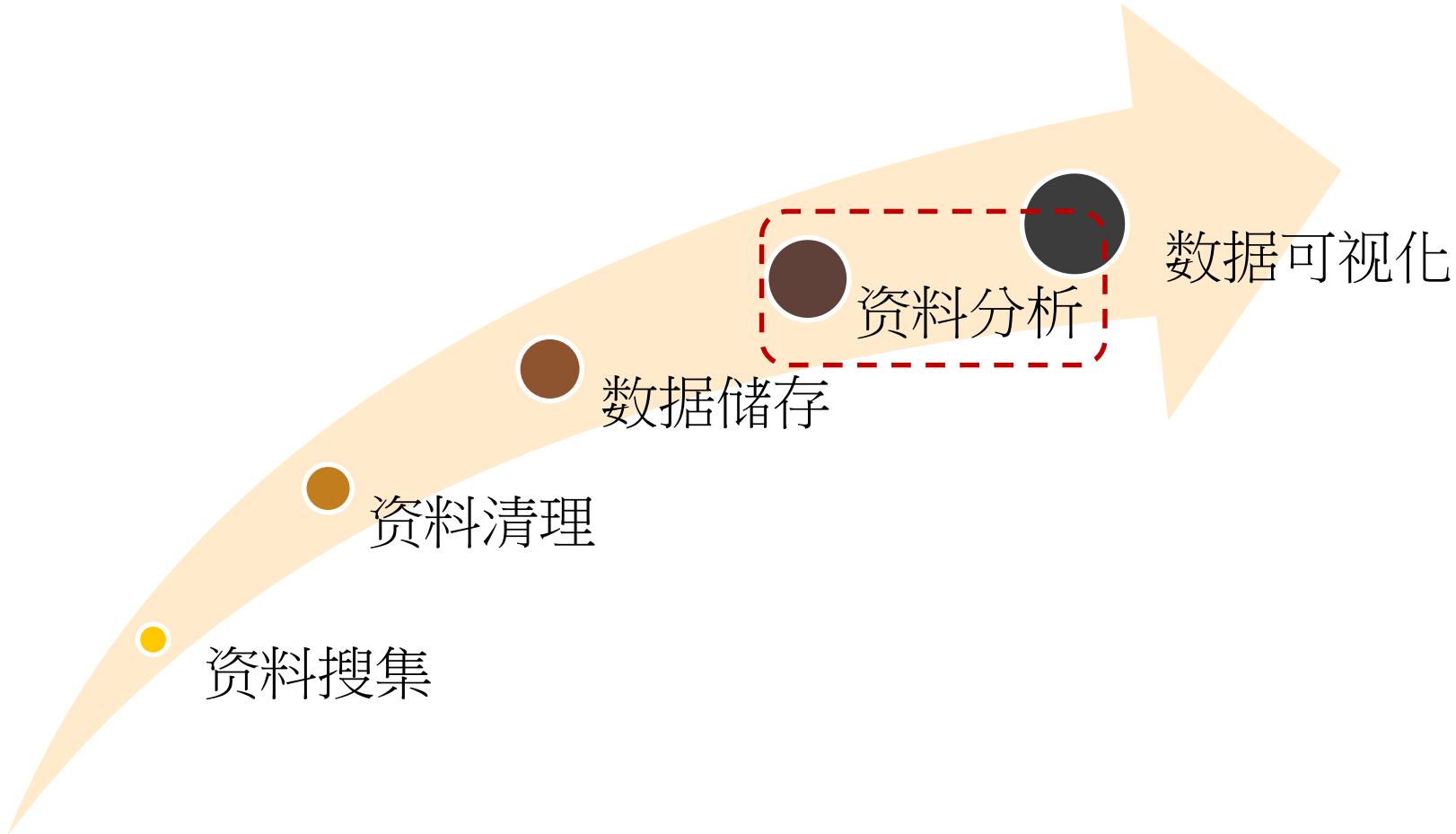
RDBMS



NoSQL



数据科学步骤



统计

日常生活中常需要根据不完整的信息做决定

统计可以把**不确定的程度量化**，用精确的方式来表达，掌握不确定的程度

统计学的目的

- 分析数据，将数据做出摘要
- 做出更好的决定
- 辨识出能提升做每件事的效果
- 评估决策或事项的功用



叙述性统计 v.s. 推论性统计

叙述性统计

- 有系统的归纳数据，了解数据的轮廓
- 对数据样本做叙述性陈述，例如：平均数、标准偏差、计次频率、百分比
- 对数据资料的图像化处理，将数据摘要变为图表

推论性统计

- 资料模型的建构
- 从样本推论整体资料的概况
- 相关、回归、单因子变异数、因素分析



机器学习

机器学习的目的是：归纳 (Induction)

- 从详细事实到一般通论

A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E -- Tom Mitchell (1998)

找出有效的预测模型

- 一开始都从一个简单的模型开始
- 藉由不断喂入训练数据，修改模型
- 不断提升预测绩效



机器学习

监督式学习 (Supervised Learning)

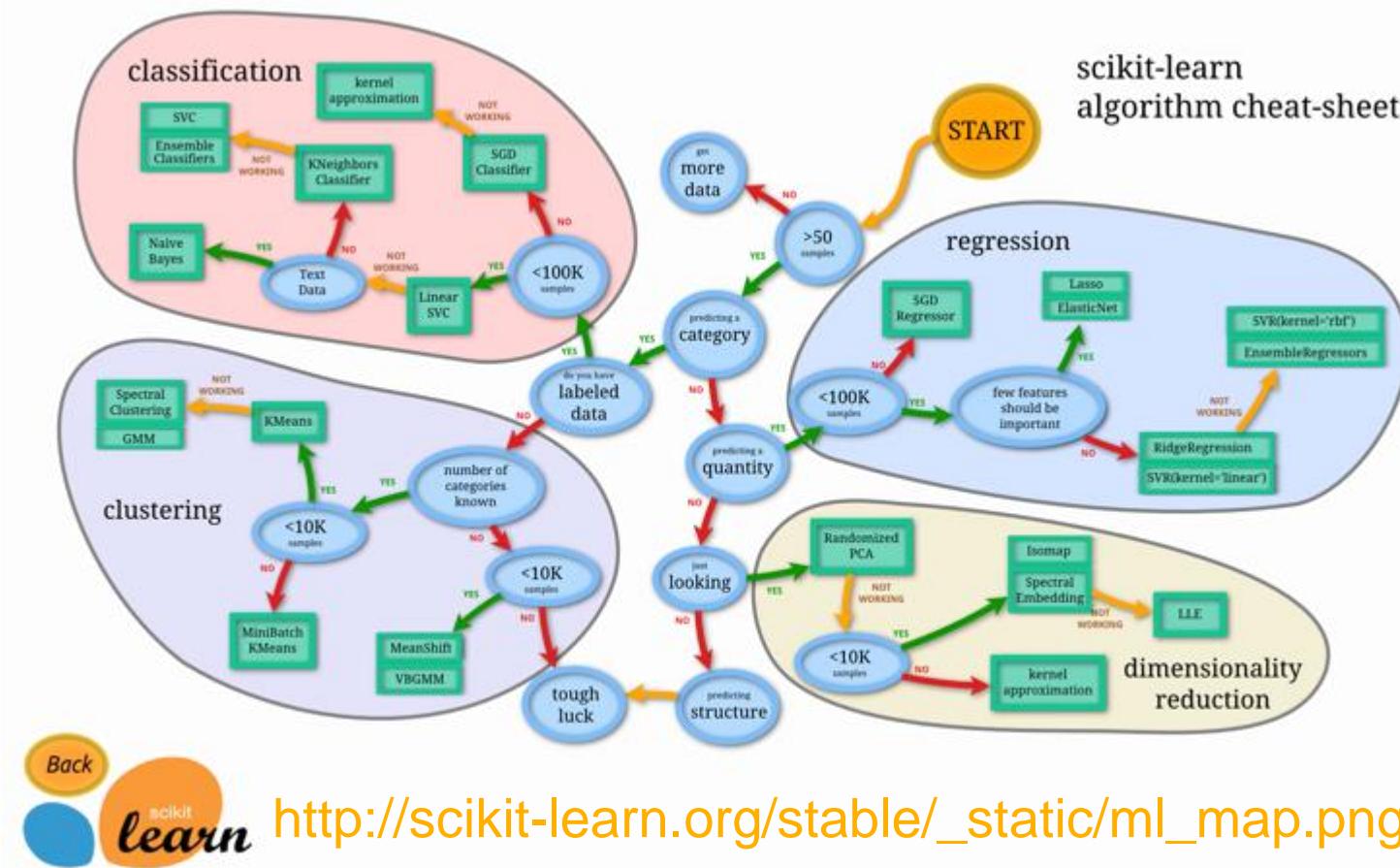
- 回归分析 (Regression)
- 分类问题 (Classification)

非监督式学习 (Unsupervised Learning)

- 降低维度 (Dimension Reduction)
- 分群问题 (Clustering)



机器学习判别方法



机器学习 v.s. 统计

统计学家

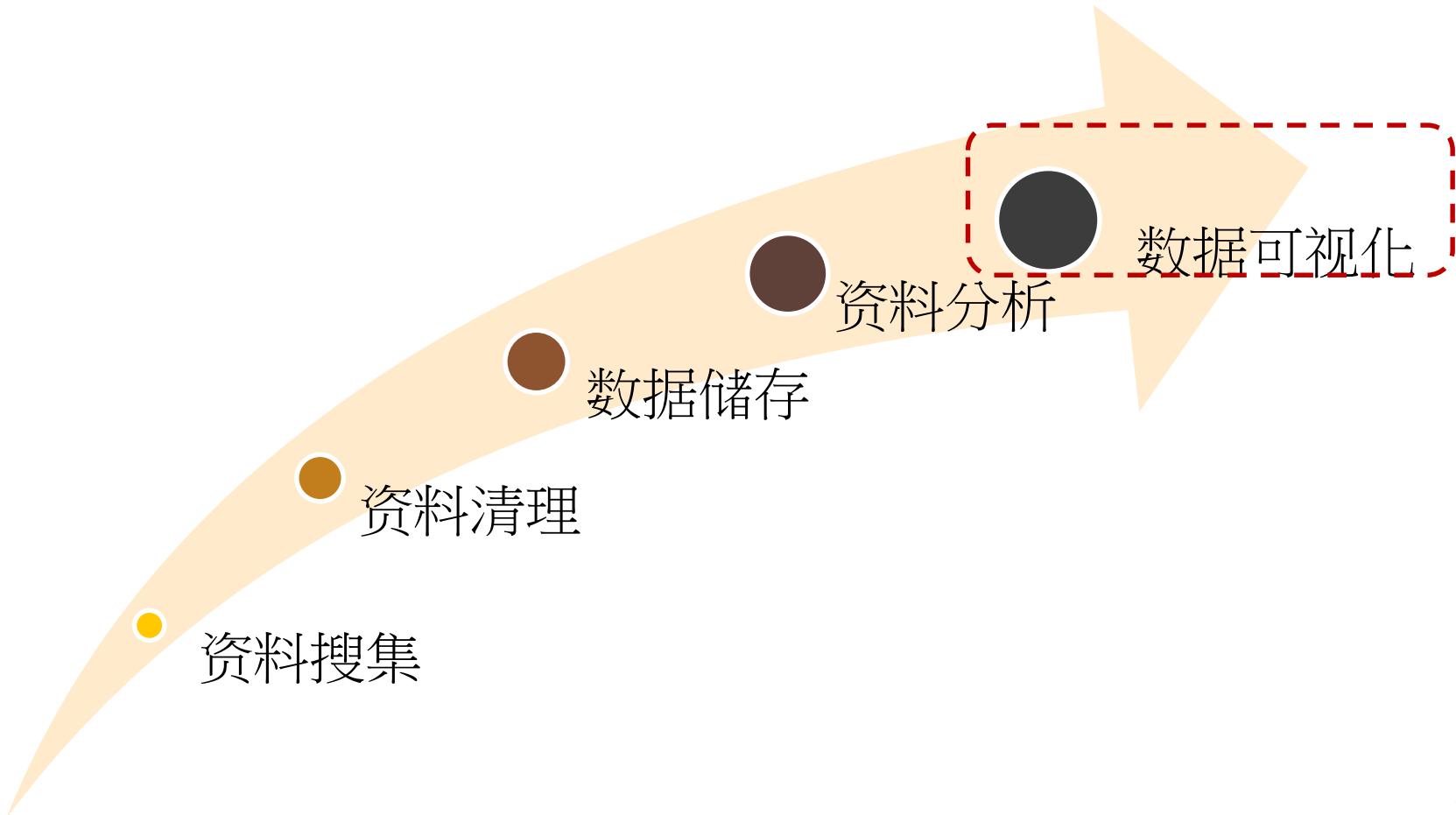
- 专注于模型甚于数据
- 可以用小量数据便可进行预测
- 建立预测模型时已考虑随机误差

机器学习学家

- 专注于数据甚于模型
- 用大量数据校正模型



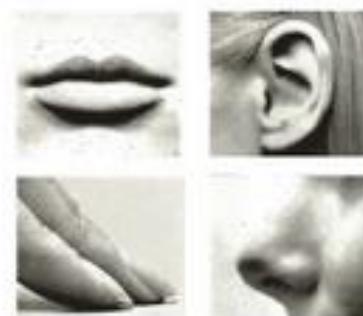
数据科学步骤



人是视觉性的动物



70%



30%

这里面有多少个 9?

3	3	0	3	0	1	8	7	6	8	2	1	4	0	3	8	3	7	7	2	0	5	2	3	2	7	0	2	0
7	1	4	6	0	2	1	3	2	7	6	0	2	5	6	3	2	5	7	6	3	3	0	2	0	3	0	7	2
8	7	5	7	2	8	3	8	7	7	8	2	0	7	7	5	2	3	1	1	5	6	3	8	4	7	8	2	0
0	5	0	5	1	6	1	7	5	6	8	0	4	4	6	7	4	7	1	4	0	0	8	4	4	3	0	3	2
2	4	3	1	3	5	4	9	5	0	7	6	0	7	4	3	1	8	2	7	3	4	6	0	2	4	8	2	3
8	6	2	2	6	5	4	6	7	0	7	6	0	0	3	9	0	2	4	7	1	7	2	3	3	5	8	7	0
0	8	4	5	1	3	1	7	6	4	5	4	1	2	4	5	3	3	5	4	9	6	7	7	6	3	4	2	5
4	7	7	0	2	2	0	1	1	7	7	7	0	2	6	6	4	7	5	8	6	1	4	3	7	8	5	4	6
4	3	6	6	4	6	6	2	8	4	8	5	3	7	8	8	1	3	8	5	4	5	7	4	0	3	2	8	4
5	5	0	3	5	3	5	3	8	3	2	3	8	2	3	1	6	2	7	2	4	6	3	6	4	4	3	2	5
4	4	0	2	1	7	2	4	4	7	4	1	9	2	4	5	2	5	0	4	0	0	5	3	6	3	3	6	7
7	4	6	6	8	7	5	7	9	2	0	2	8	8	8	3	2	4	2	6	4	0	4	6	3	7	2	1	
0	1	7	1	5	9	1	4	2	8	7	3	7	1	4	5	1	8	7	8	0	5	1	7	0	5	8	8	1
2	8	5	2	1	2	8	7	7	6	2	5	6	2	6	4	1	5	1	6	1	2	1	1	0	5	6	4	0
2	1	1	7	7	2	0	0	1	8	7	0	2	9	0	2	8	5	7	8	4	6	0	6	5	0	7	1	2
0	5	2	4	1	5	3	3	1	5	5	1	4	0	1	6	4	3	3	9	8	8	3	4	6	8	4	8	6
7	3	7	5	2	4	0	2	7	6	3	8	5	5	4	5	8	8	7	5	5	6	5	6	7	9	7	7	4
0	3	2	8	1	4	4	6	0	8	2	3	0	1	3	4	6	2	0	5	7	7	3	6	1	8	7	3	5
4	4	8	3	3	3	5	0	1	0	3	8	6	3	2	0	5	0	6	1	3	3	4	3	6	1	5	8	6
1	0	2	2	7	6	3	3	0	8	8	0	3	1	8	8	1	2	1	7	5	2	9	3	5	8	3	2	5

数据可视化的重要性

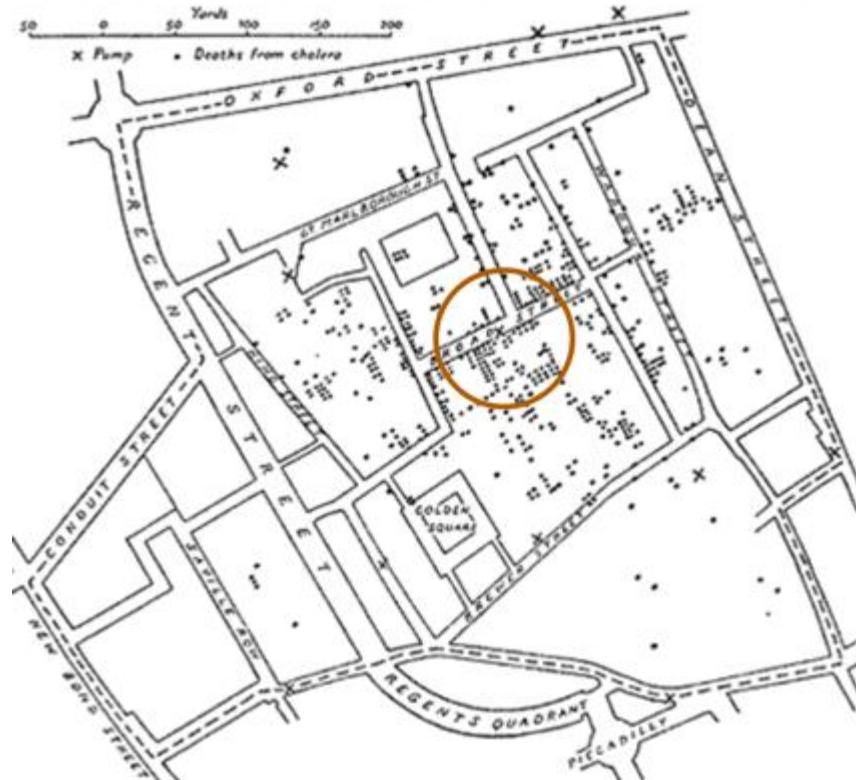
3	3	0	3	0	1	8	7	6	8	2	1	4	0	3	8	3	7	7	2	0	5	2	3	2	7	0	2	0
7	1	4	6	0	2	1	3	2	7	6	0	2	5	6	3	2	5	7	6	3	3	0	2	0	3	0	7	2
8	7	5	7	2	8	3	8	7	7	8	2	0	7	7	5	2	3	1	1	5	6	3	8	4	7	8	2	0
0	5	0	5	1	6	1	7	5	6	8	0	4	4	6	7	4	7	1	4	0	0	8	4	4	3	0	3	2
2	4	3	1	3	5	4	9	5	0	7	6	0	7	4	3	1	8	2	7	3	4	6	0	2	4	8	2	3
8	6	2	2	6	5	4	6	7	0	7	6	0	0	3	9	0	2	4	7	1	7	2	3	3	5	8	7	0
0	8	4	5	1	3	1	7	6	4	5	4	1	2	4	5	3	3	5	4	9	6	7	7	6	3	4	2	5
4	7	7	0	2	2	0	1	1	7	7	7	0	2	6	6	4	7	5	8	6	1	4	3	7	8	5	4	6
4	3	6	6	4	6	6	2	8	4	8	5	3	7	8	8	1	3	8	5	4	5	7	4	0	3	2	8	4
5	5	0	3	5	3	5	3	8	3	2	3	8	2	3	1	6	2	7	2	4	6	3	6	4	4	3	2	5
4	4	0	2	1	7	2	4	4	7	4	1	9	2	4	5	2	5	0	4	0	0	5	3	6	3	3	6	7
7	4	6	6	8	7	5	7	9	2	0	2	8	8	8	3	2	4	2	6	4	0	4	6	3	7	2	1	
0	1	7	1	5	9	1	4	2	8	7	3	7	1	4	5	1	8	7	8	0	5	1	7	0	5	8	8	1
2	8	5	2	1	2	8	7	7	6	2	5	6	2	6	4	1	5	1	6	1	2	1	1	0	5	6	4	0
2	1	1	7	7	2	0	0	1	8	7	0	2	9	0	2	8	5	7	8	4	6	0	6	5	0	7	1	2
0	5	2	4	1	5	3	3	1	5	5	1	4	0	1	6	4	3	3	9	8	8	3	4	6	8	4	8	6
7	3	7	5	2	4	0	2	7	6	3	8	5	5	4	5	8	8	7	5	5	6	5	6	7	9	7	7	4
0	3	2	8	1	4	4	6	0	8	2	3	0	1	3	4	6	2	0	5	7	7	3	6	1	8	7	3	5
4	4	8	3	3	3	5	0	1	0	3	8	6	3	2	0	5	0	6	1	3	3	4	3	6	1	5	8	6
1	0	2	2	7	6	3	3	0	8	8	0	3	1	8	8	1	2	1	7	5	2	9	3	5	8	3	2	5



数据可视化的重要性

1854 年，霍乱疫情爆发，造成十天之内死了五百多人

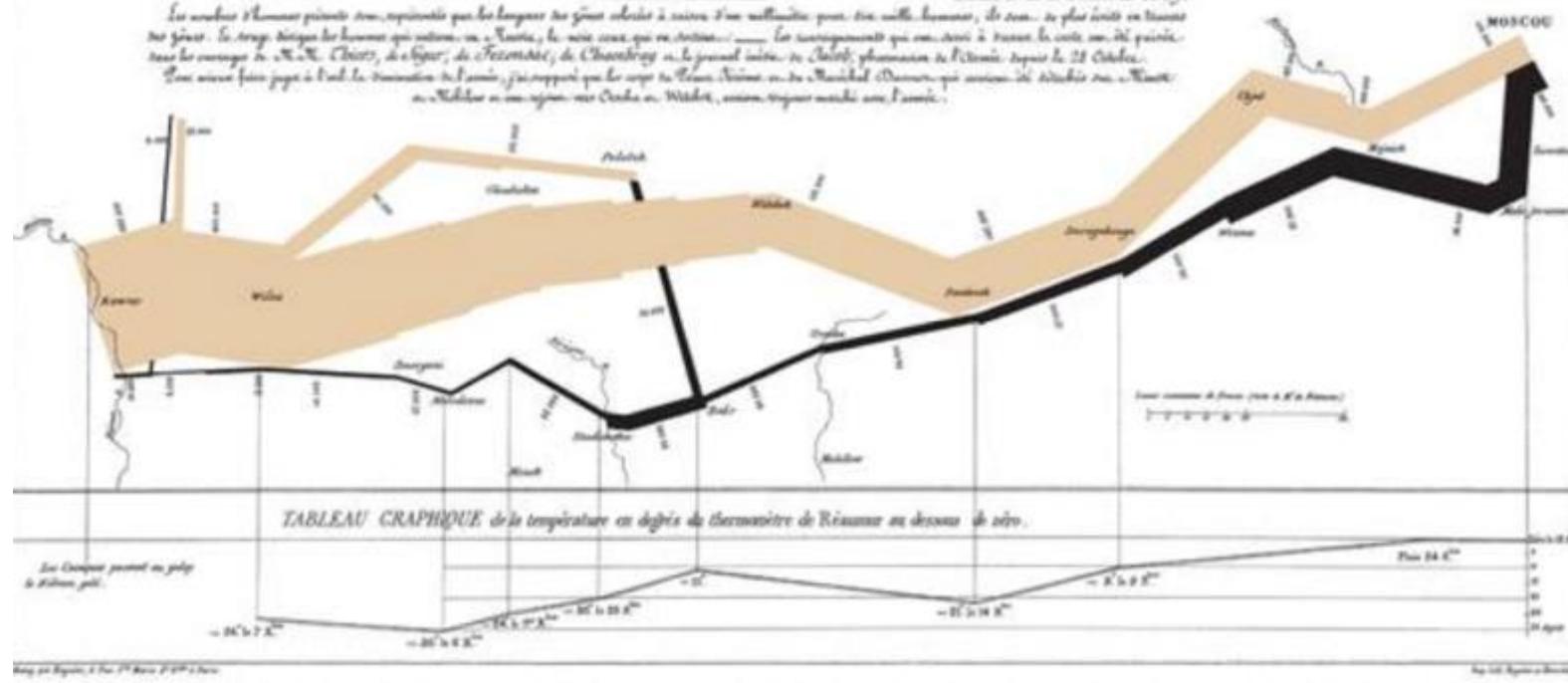
Dr. John Snow 将所有病患的住家位置点在地图上，发现病例聚集在一口井附近



拿破仑远征俄国

Carte Figurative des pertes successives en hommes de l'Armée française dans la campagne de Russie 1812-1813.
Dessiné par M. Félixard, Ingénieur Général des Ponts et Chaussées en état-major.

Les nombreux échanges présentés nous rappellent que les langues des groupes identifiés à ce stade d'un millénaire passé, sont aussi homogènes, de sorte... le plus évident en termes des groupes. Le temps distingue les hommes qui naissent, en « Australie », le reste ceux qui se transforment... Les enseignements qui sont donnés à travers la culture sont de préférence les enseignements de l'Australie, de l'Asie, de l'Europe, de l'Amérique du Sud, et également de l'Amérique, depuis le 25 Octobre. Ces œuvres font partie à l'ordre du moment, qui rappelle que les œuvres de Paul Valéry ou de Marcel Proust, qui concernent le siècle dans son « Australie » ou au contraire, aux Océans ou à l'Asie.

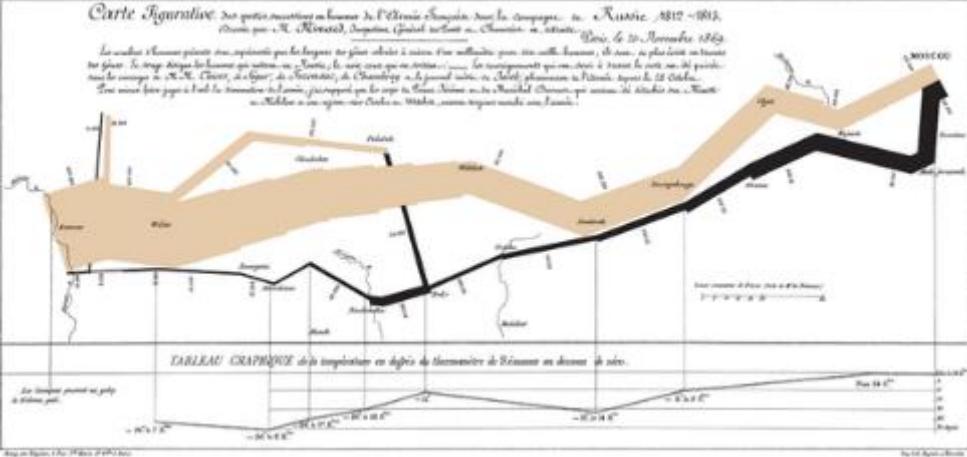


信息图表的功能

Carte figurative des pertes humaines de l'Armée française dans la campagne de Russie, 1812-1813.

(Créée par M. MORÉAU, Ingénieur Général sortant en Commission, à Paris, le 20 Novembre 1848.)

Le résultat de cette guerre est représenté par les lignes de plus éloigné à celles d'un millionne pour les deux batailles de Rostov, le plus court en temps que plus long. Le temps depuis les batailles qui suivent un point, le plus court que ce dernier, est le temps nécessaire pour venir à bout du plus court ou du plus long. Le temps depuis la bataille de Malakof, de Novorossiisk, de Chouvaloff, de Koubane, de Tchelkatchevsk, de Piatigorsk, jusqu'à la bataille de Malakof. Le temps depuis la bataille de Malakof jusqu'à la bataille de Malakof. Le temps depuis la bataille de Malakof jusqu'à la bataille de Malakof. Le temps depuis la bataille de Malakof jusqu'à la bataille de Malakof.



Hans Rosling's 200 Countries, 200 Years, 4 Minutes

<https://www.youtube.com/watch?v=jbkSRLYSoho>





善用手边的开源工具 -
让数据科学变得超简单

数据分析语言

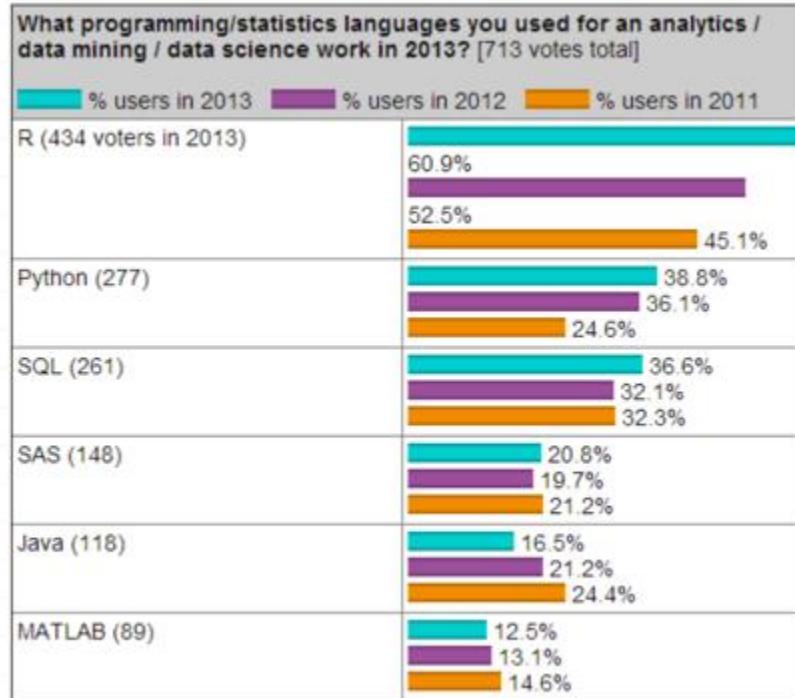


 python

The Python logo consists of two interlocking snakes, one blue and one yellow.

数据分析语言

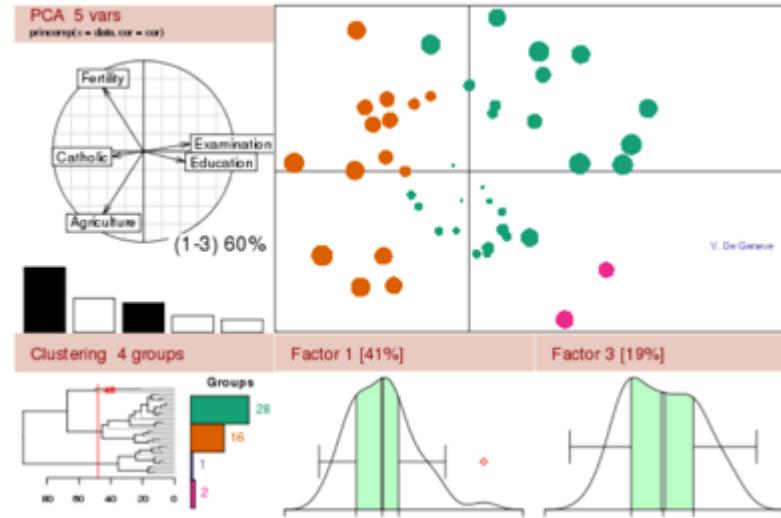
最受歡迎的語言持續為 **R, Python (39%)**, 及 **SQL (37%)**. SAS 大約在 20% 上下.



By Gregory Piatetsky, Aug 27, 2013.

R 语言

- AT&T贝尔实验室暨S语言所发展出来的GNU 专案
- 提供统计分析与图形可视化功能的开源程序语言
- 使用C, Fortran 编程的函式语言



R 语言

- S 语言的方言 (分支)
- 受到函数式编程语言Scheme 的启发，因而想将该功能加入到 S 语言当中
- 1992年Ross Ihaka 与 Robert Gentleman 为了教授统计，因此开发出了 R语言
- 除了R 以外，还有S-Plus，但两个分支走向不同，一个走向社群，一个走向商业



R 语言

立即完成统计分析

- 数据处理
- 资料分析
- 报表制作



内建许多数学函式及图形套件(也可安装第三方套件)

- 可以结合其他语言：如Java, C++
- 免费且开源 <http://cran.r-project.org/src/base/>

容易扩充和客制化

Python 语言

动态语言 (Dynamic Language)

- 于执行时期(Runtime)执行程序代码 (不用编译)
- Dynamic Type: 函数与变量都不需要宣告类型

直译式语言 (Interpreted Language)

每次执行后可以直接看到结果

物件导向语言 (OOP)



可执行于多平台 (Python VM)

Guido van Rossum – Python 之父



Guido van Rossum

+ 加入社群

59,938 位追随者 | 4,444,554 次点赞

Guido Van Rossum
(<https://goo.gl/2kul7Y>)

Guido van Rossum
公關分享 - 2013年9月19日

Do not send me email like this:

Hi Guido,

I came across your resume in a Google web search. You seem to have an awesome expertise on Python. I would be glad if you can reply my email and let me know your interest and availability.

Our client immediately needs a PYTHON Developers at its location in *, NJ. Below are the job details. If interested and available, kindly fwd me your updated resume along with the expected rate and the availability.

[...]

I might reply like this:

I'm not interested and not available.

Monty Python
(<https://goo.gl/dTjCxR>)



Python 无所不在

物聯網
(<http://goo.gl/2j45Nk>)



網頁製作
(<https://goo.gl/2304w7>)



資料分析
(<https://goo.gl/2304w7>)

Python 简单易用

JAVA

```
class test{  
    public static void main(String args[]){  
        System.out.println("Hello World");  
    }  
}
```



PYTHON



```
print("Hello World")
```

R v.s. Python

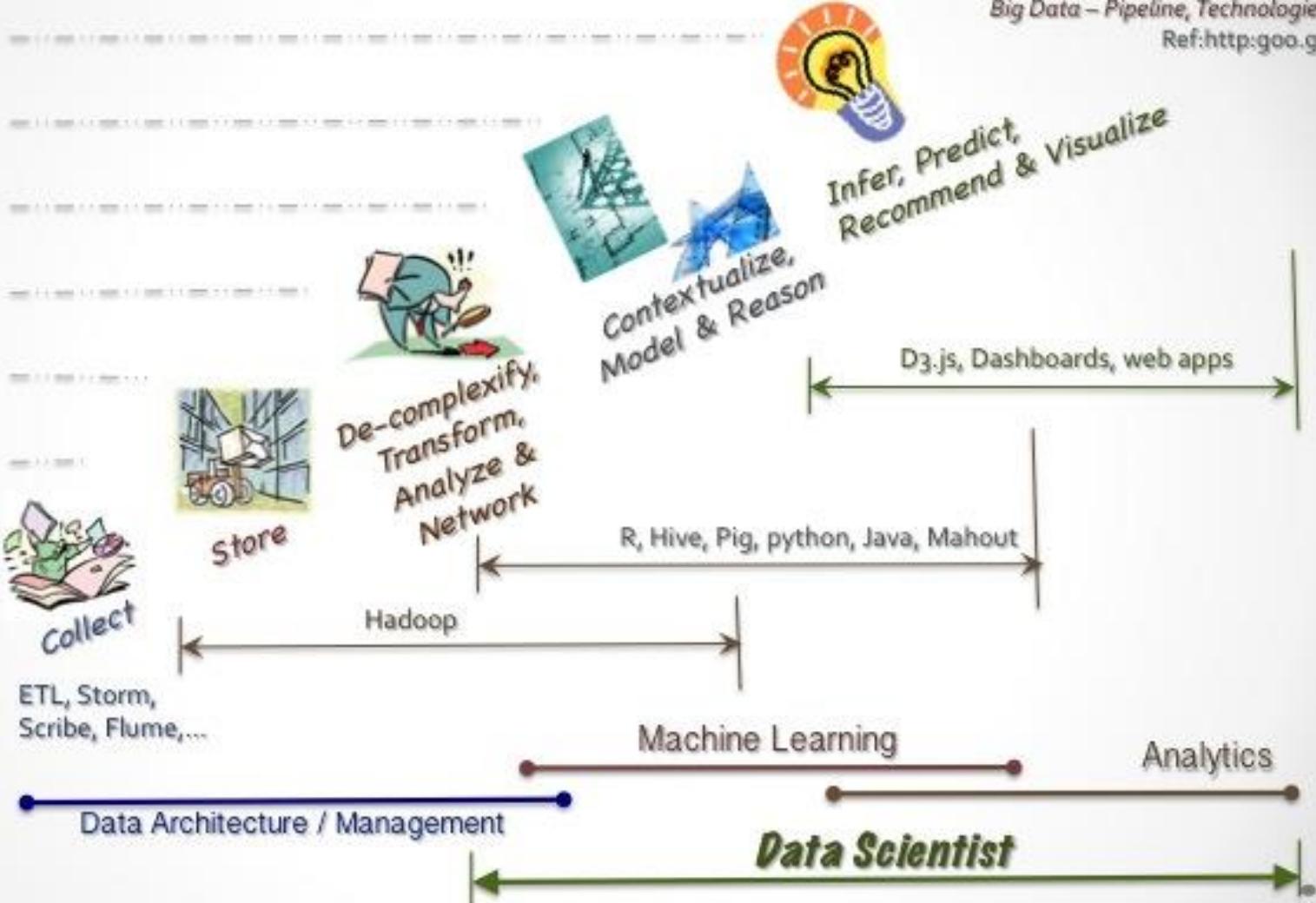
Analysis Tool	Similar Superhero	Super Powers in Common
R 	Batman 	<ul style="list-style-type: none">• Detective Work• Intelligence• Cunning• Usage of Tools• More Brain than Muscles
Python 	Superman  <small>DC Comics</small>	<ul style="list-style-type: none">• Muscle Power• Super Strength• Elegance• Wide Range• More Muscles than Brain





数据科学的学习资源与建议

Model
Context
Connectedness
Variety
Variability
Velocity
Volume
Technologies
Roles



选择你的武器

资工背景、写过程序的人、对工程面不排斥者
Python 是你们的好选择

数学统计背景者、或只想碰分析的人
R 是你们的好选择

懂商业逻辑者，但完全不想写程序的人
SQL 与 可视化工具(e.g. Tableau) 是好选择



Coursera

<https://zh-tw.coursera.org/>

The screenshot shows the Coursera website for the Data Science specialization. At the top, there's a navigation bar with the Coursera logo, a search bar, and links for '合作夥伴', '登錄', and '註冊'. Below the navigation is a large banner featuring a man in a plaid shirt. The banner text reads 'Launch Your Career in Data Science' and 'A nine-course introduction to data science, developed and taught by leading professors.' To the left of the banner is a sidebar with a menu for the specialization, including '專項課程介紹', '課程', '製作方', '常見問題解答', and 'Data Science 專項課程'. A blue button labeled '註冊' with the text '於1月9開始' is also visible. The main content area below the banner includes a section titled '專項課程介紹' with the text 'Ask the right questions, manipulate data sets, and create visualizations to communicate results.' and a detailed description at the bottom.

coursera

目錄 搜索 目錄

合作夥伴 登錄 註冊

分享

Launch Your Career in Data Science

A nine-course introduction to data science, developed and taught by leading professors.

專項課程介紹

課程

製作方

常見問題解答

Data Science
專項課程

註冊 於1月9開始

專項課程介紹

Ask the right questions, manipulate data sets, and create visualizations to communicate results.

Financial Aid is available for learners who cannot afford the fee. Learn more and apply.

This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone

Large Data

Kaggle

<https://www.kaggle.com/>

12 active competitions

Sort By: Prize ▾

Active All Entered All Categories Q

Competition	Description	Prize	Teams
Data Science Bowl 2017  Can you improve lung cancer detection? Featured · 3 months to go · 17 kernels	\$1 million 381 teams		
The Nature Conservancy Fisheries Monitoring  Can you detect and classify species of fish? Featured · 3 months to go · 205 kernels	\$150,000 1,117 teams		
Dstl Satellite Imagery Feature Detection  Can you train an eye in the sky? Featured · 2 months to go · 78 kernels	\$100,000 117 teams		
Two Sigma Financial Modeling Challenge  Can you uncover predictive value in an uncertain world? Featured · 2 months to go · 145 kernels	\$100,000 1,284 teams		
Outbrain Click Prediction  Can you predict which recommended content each user will click? Featured · 8 days to go · 380 kernels	\$25,000 964 teams		
Santa's Uncertain Bags 			



大數學堂

<http://largitdata.com/>

The screenshot shows the homepage of the Large Data Academy (大數學堂). At the top, there is a navigation bar with a logo, the text "大數學堂", and a "跳到頁面" (Jump to page) link. Below the navigation bar is a banner featuring a laptop displaying a landscape image, a smartphone, and a pen holder with pens and a pair of scissors. The banner has text overlay: "只要三分鐘 打造明日的競爭力" (Achieve tomorrow's competitiveness in just 3 minutes) and "完全免費的線上教材" (Free online courses). At the bottom of the page, there are three blue call-to-action boxes: "實用課程" (Practical Courses) with text about learning from small data to big data and logos for ABC; "InfoMiner" with text about web data monitoring services; and "聯絡我們" (Contact Us) with text about connecting to Large Data and a building icon.

大數學堂

[跳到頁面](#)

課程列表 INFOMINER商情分析平臺

只要三分鐘 打造明日的競爭力

完全免費的線上教材

實用課程

從小數據到大數據，從R與Python到Hadoop, Spark。我們讓資料分析的技能不再是樞口號，更要落實到實際的工作與學習之中。

A
B
C

InfoMiner

「InfoMiner」網頁資料監控服務，可以快速、精準擷取同業／競爭對手於網路上公開的資料，並探勘出網路趨勢，協助企業客戶迅速掌握市場動態。

聯絡我們

大數軟體 X 大數學堂，我們期望打造的是真正的數據生態，歡迎有任何想法的同好聯絡我們，期待能擦出不一樣的火花。

LargeData

实践是检验真理的唯一标准

- 数据科学的目的是让你开始以数据做决策的依据，不是万灵丹！
- 数据科学是工程 + 数学 + 领域知识的学问，**没办法速成，也没有每个领域都通的数据科学家**
- 放下你的微积分、高微课本，懂点数学是好的！但过于专注算式，你可能会难以踏出第一步！
- 大数据的确很潮，但分析方法如果没有改变，**只是换汤不换药！**更应该关注的是**问题**本身，而不是**技术**！
- 数据科学有很多面向，你不用全能，挑一个你喜欢（擅长）的位置即可，**只有团队，没有英雄！**





微信扫码，立即参团



THANK YOU

EMAIL: david@largitdata.com

网站: www.largitdta.com

电话: 0929094381