数据科学入门 –
让数据思维成为生活的一部分

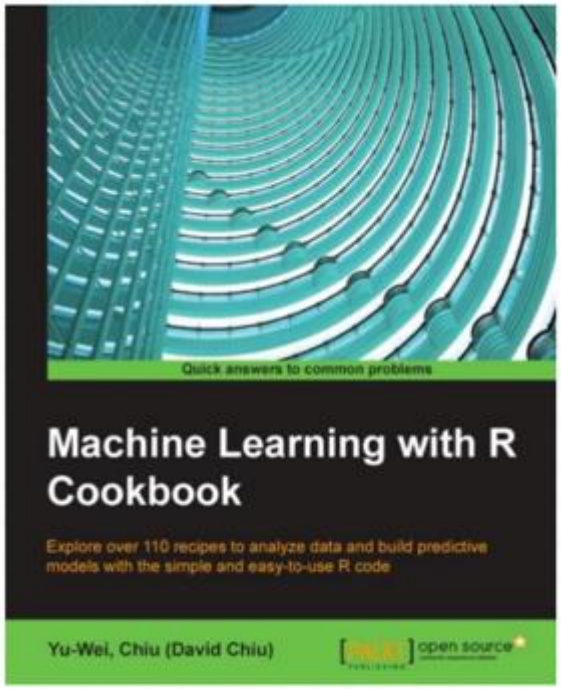
丘祐玮 – David Chiu

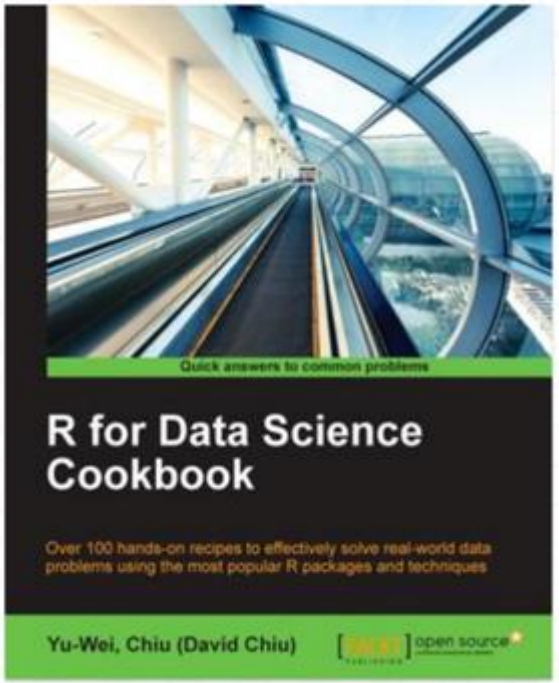
EMAIL: david@largitdata.com
网站: www.largitdta.com
电话: +886929094381

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



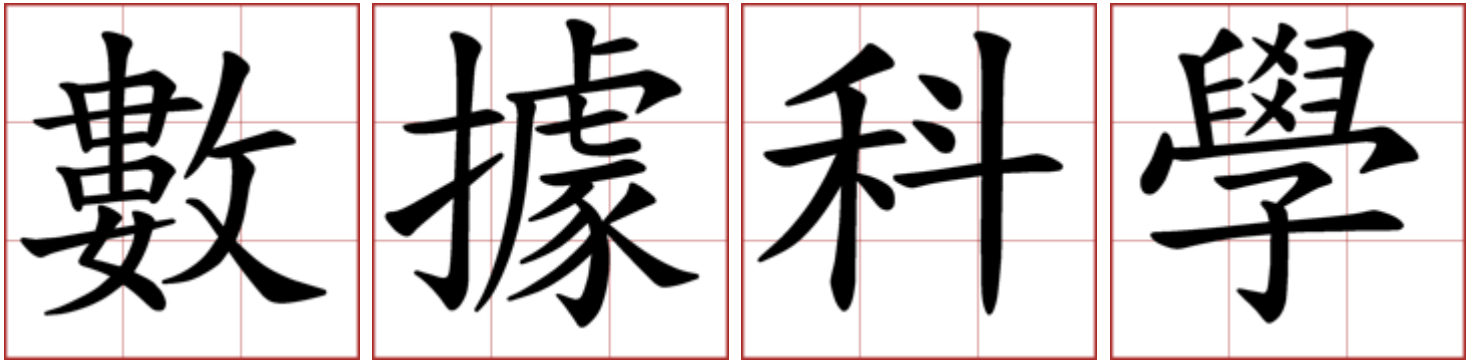


Author: Yu-Wei (David) Chiu

什么是数据科学?

数据

科学



哪家公司没有数据？

哪家公司不科学了？


靠感觉
靠经验?


雇人拿计数器
计算不同时间点的人潮流量

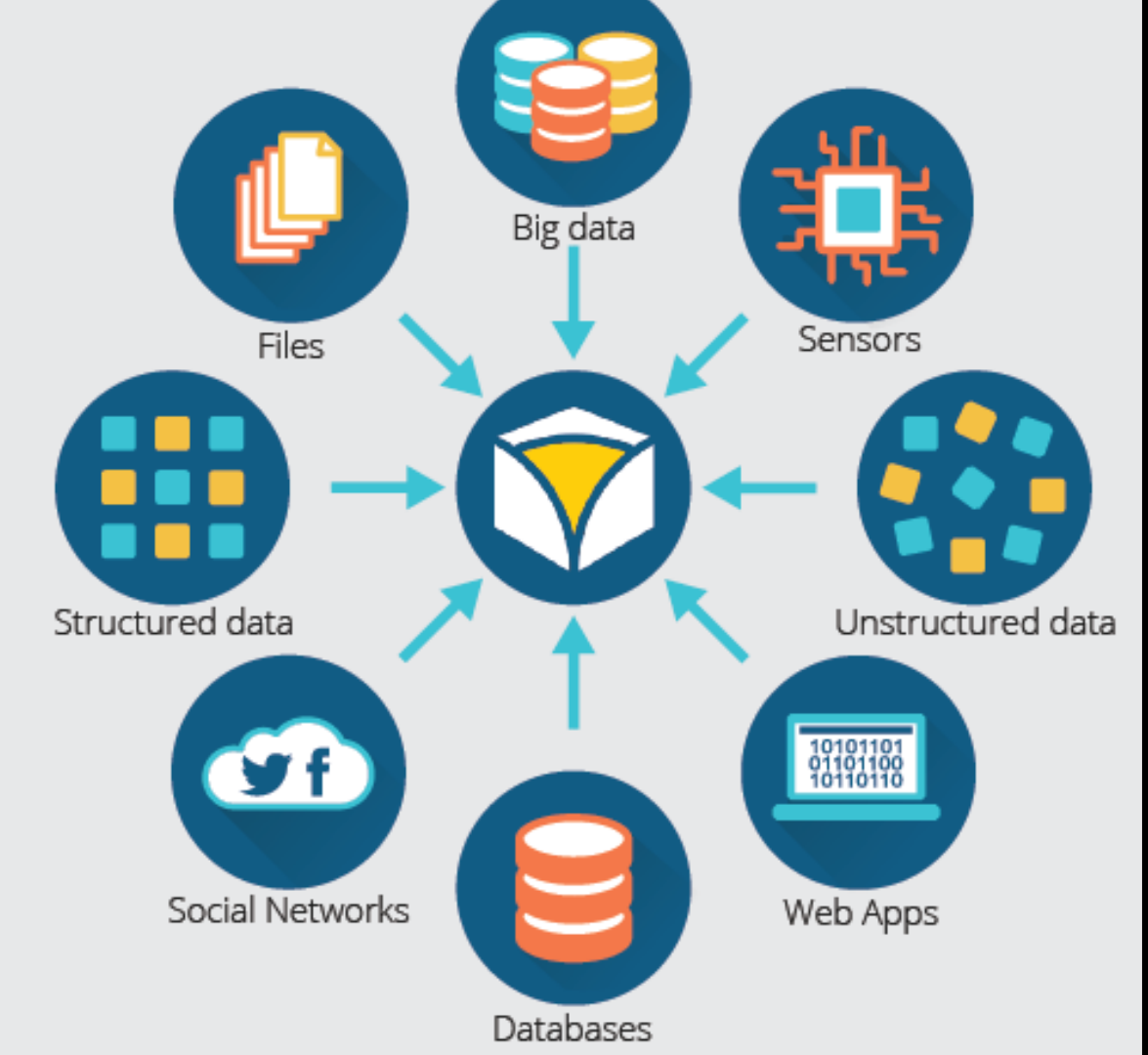
用数据发现问题

用数据解决问题

雲端計算的興起

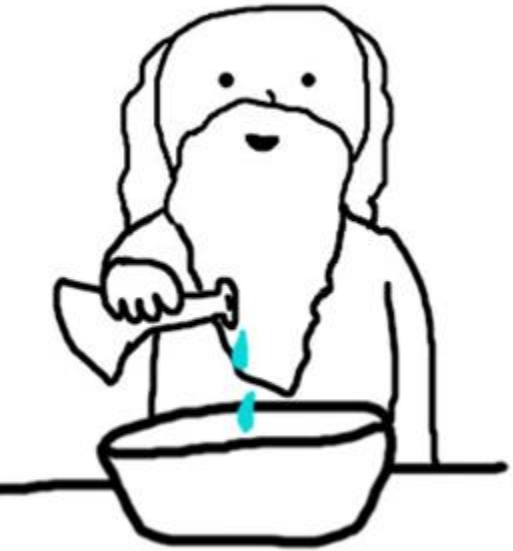
企業數據

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

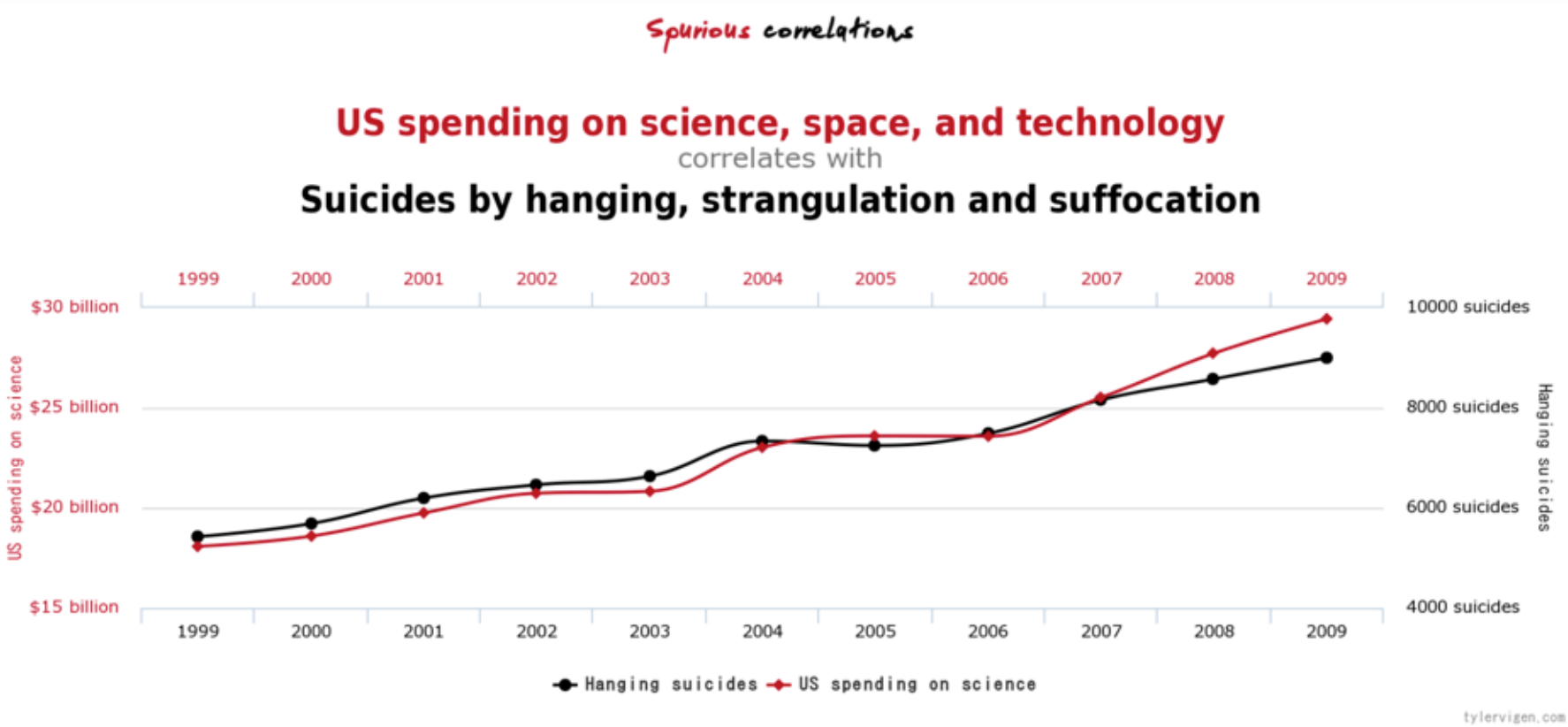


加一点工程

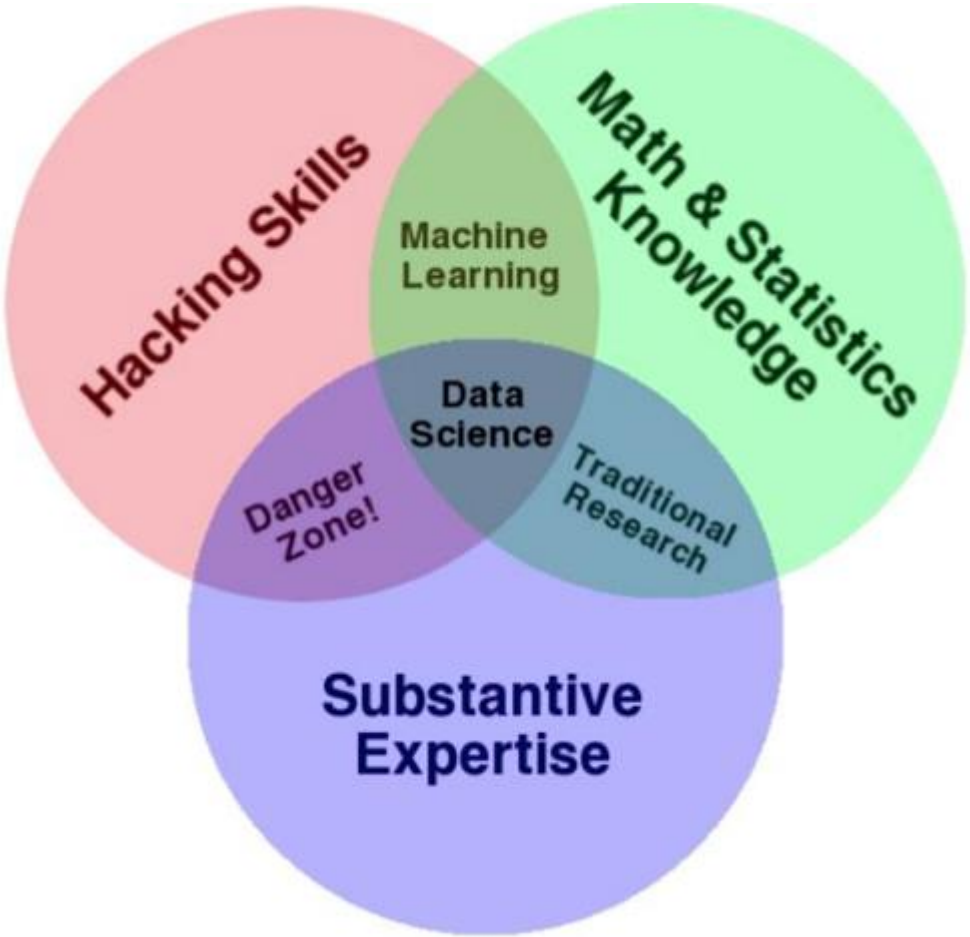
产生数据科学


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
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Hanging suicides
US spending on science
tylervigen.com

数据科学


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

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

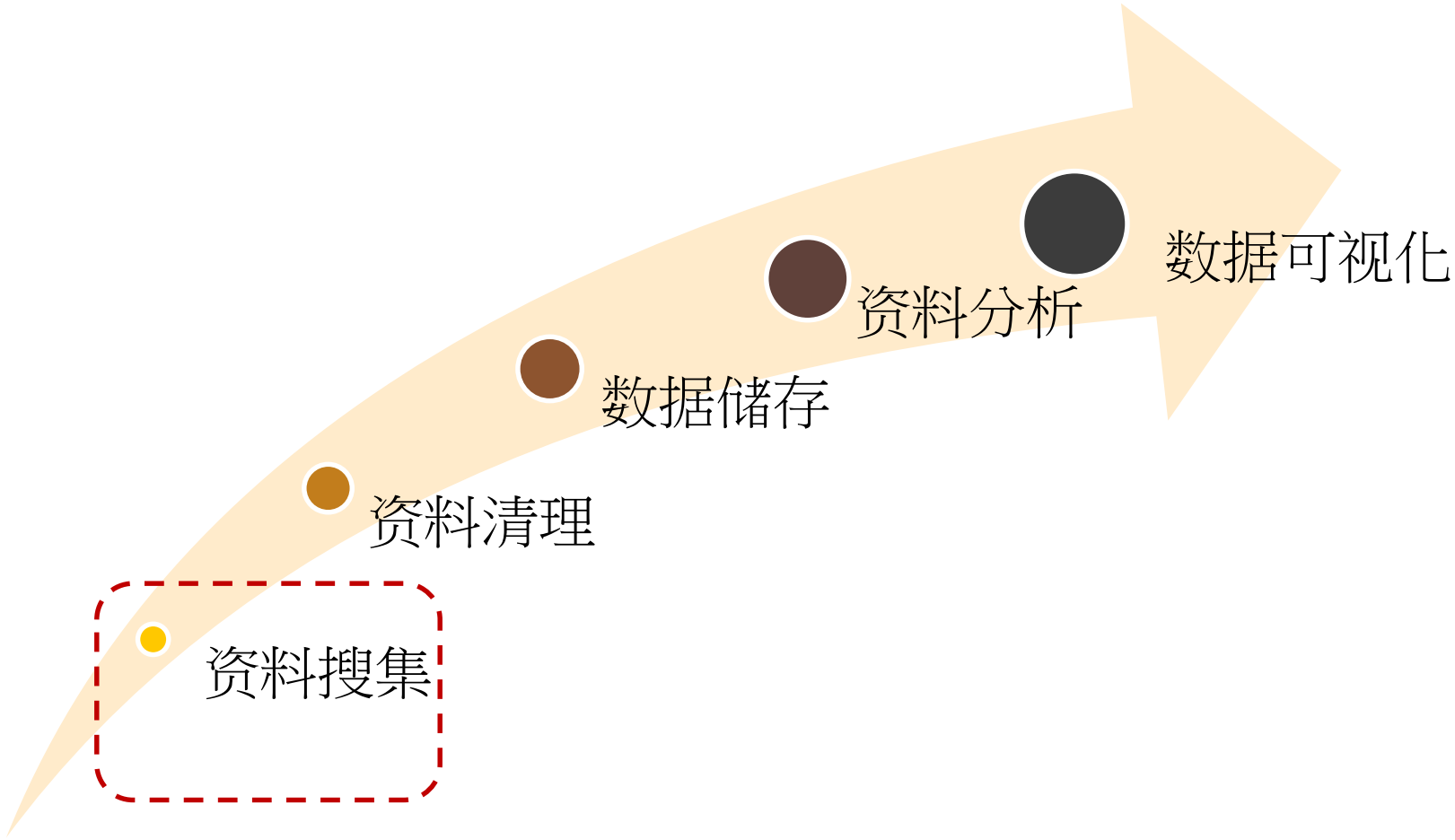
商业分析师
了解算法并知道在
不同数据量下该使
用何种工具

解构! 探讨数据科学流程

使用数据拟定策略


去年有多少客戶流失？
為什麼去年有客戶流失比例增加？
哪個客戶最有可能流失？
該如何避免客戶流失？
敘述
診斷
預測
處方
資訊
洞察
決策

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　鑫好企業判6...	鑫好企業負責人吳容含被...	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】蔡英文滿意度4成7　...	總統蔡英文就職即將滿月...	2016-06-17 16:08:00	0	政治
60	【有片】索珠之亂有續集　洪家賣...	(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

- 可以下SQL 处理与捞取数据

  select title, content from newsmain;

半结构化数据

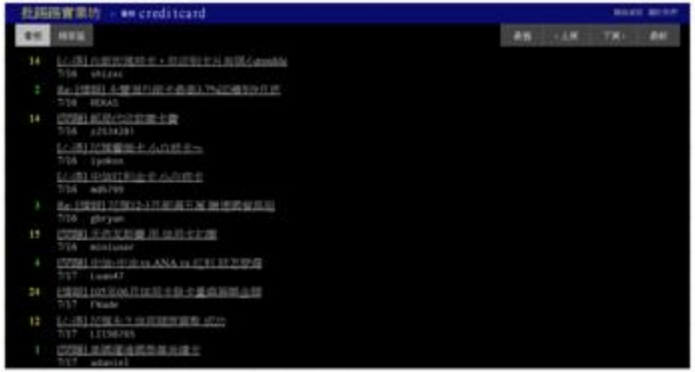
XML

```
<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
   }
]
```

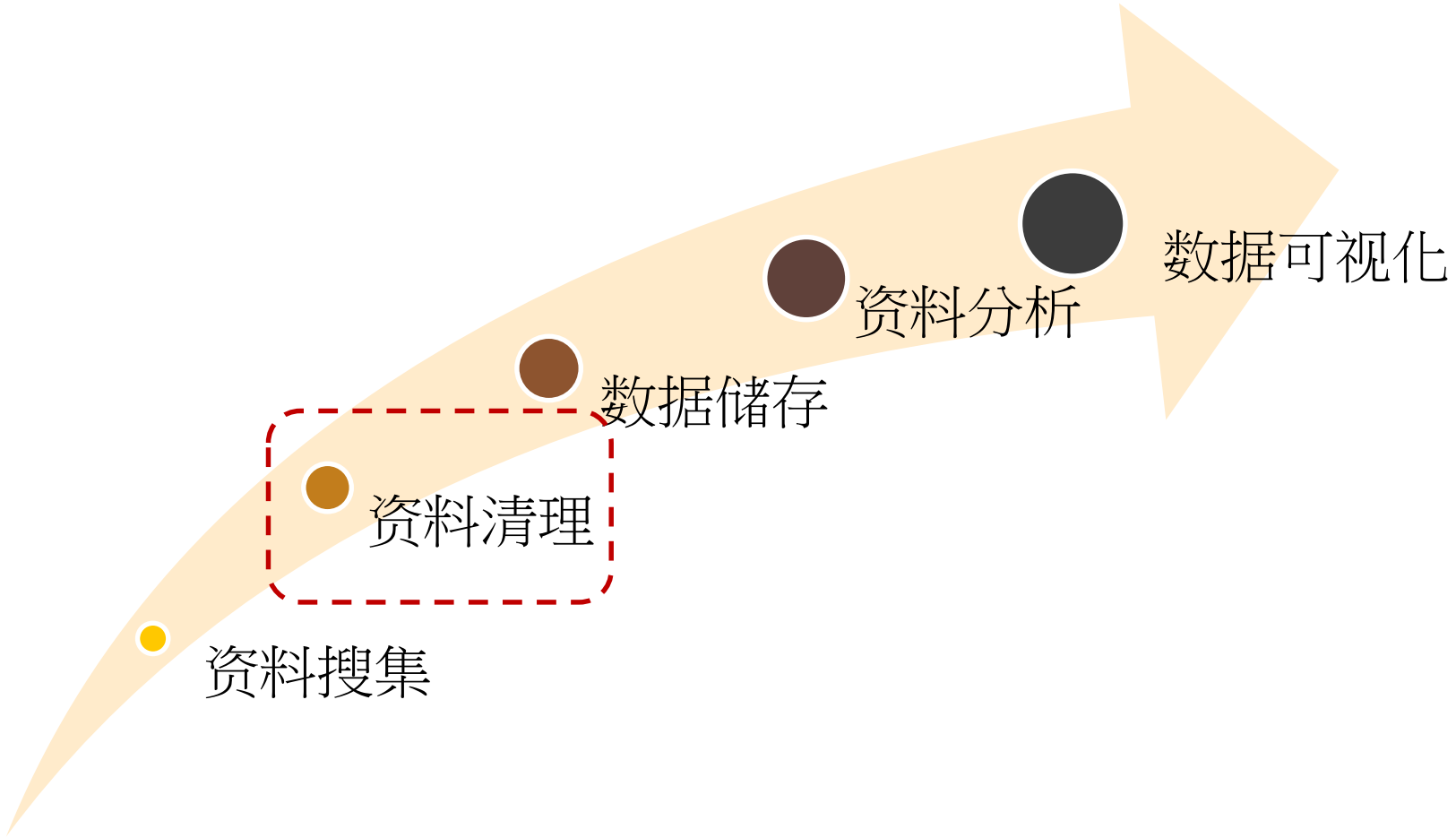
网页资料



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	異議人士高瑜北京住處　驚傳遭強拆(0)	國際
3	16:50	【更新】50元偽幣流竄　台南警查獲6萬枚(157224)	社會
4	16:50	【特企】一份用愛傳遞的禮物高安生命契約 (2911)	特企
5	16:50	【法廣RFI】安倍參加峰會欲與習近平會...(0)	國際
6	16:48	喬治亞州死刑犯　周四注射毒針(45)	國際
7	16:46	【更新】疑爆裂物品炸傷焚化爐員工　188...(2428)	社會
8	16:43	超商敬悲　學者嗆批評者：會虧錢的結構式...(166)	生活
9	16:40	【TOMO愛語傳】「德翔台北輪」斷裂漏油...(253)	國際
10	16:40	毆打運將標題翻轉　湘瑩嘆「努力還是不夠」(2171)	娛樂
11	16:39	【有片】身體因素無法返台？　高敏慧以色列...(38642)	政治
12	16:35	【10樓可樂魅力大！】花錢都買不到(582)	財經
13	16:33	【法廣RFI】富家女被綁案 港府被指拱手...(330)	國際
14	16:32	檢討新聞發展條例？局長：暫時沒必要(135)	財經

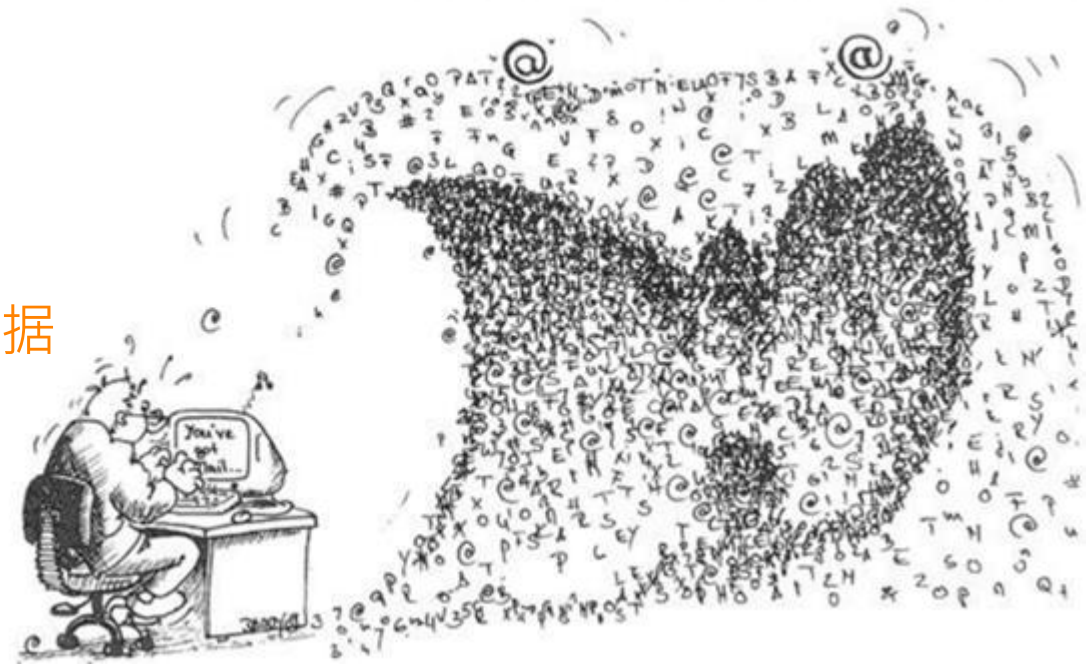
数据科学步骤



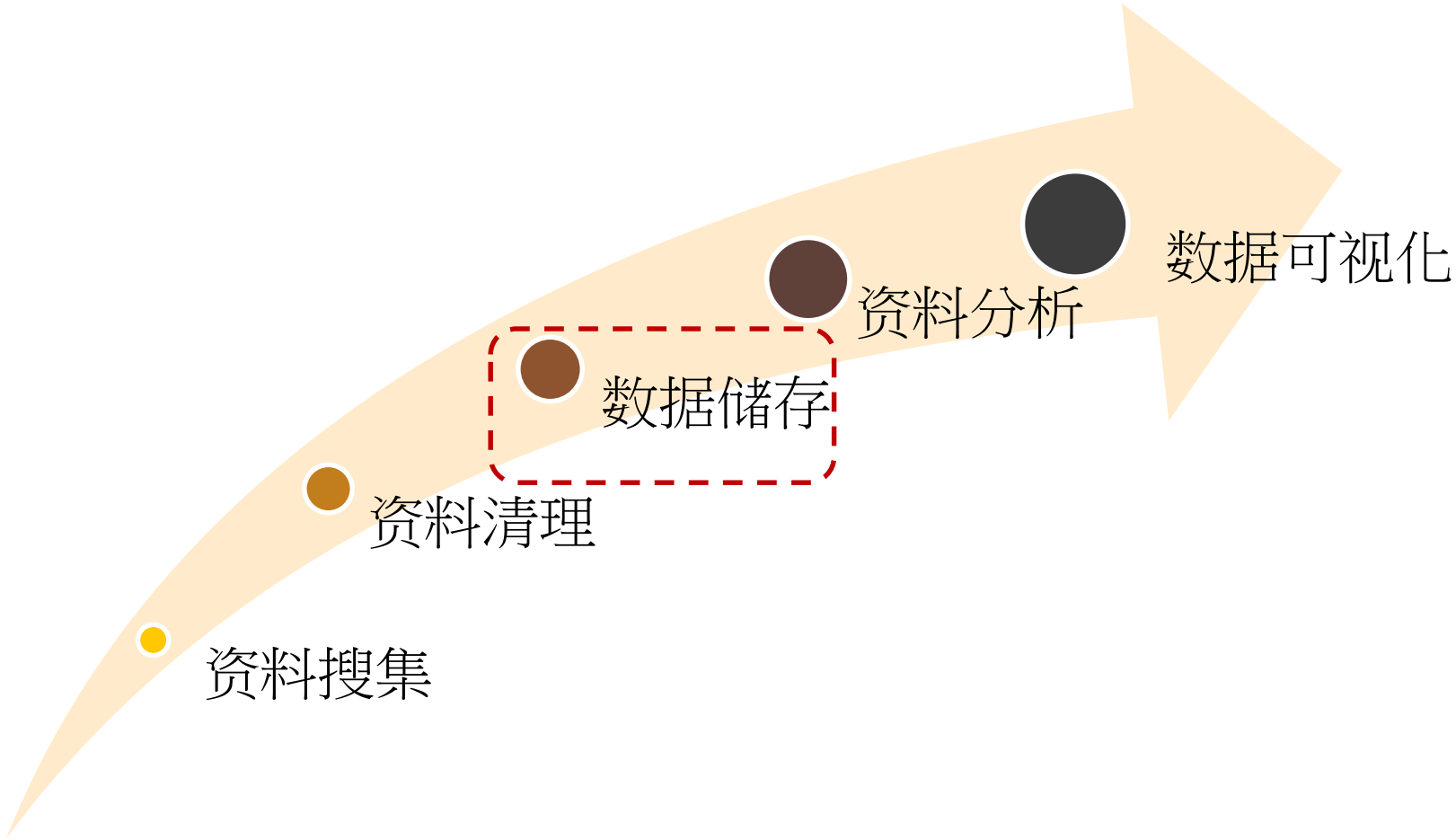
资料清理二三事

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

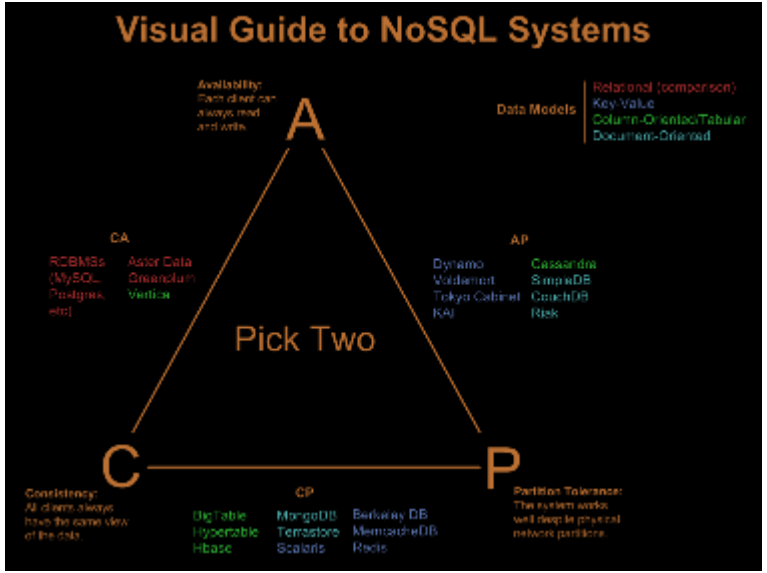
可以透过标准模型整合数据

- 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

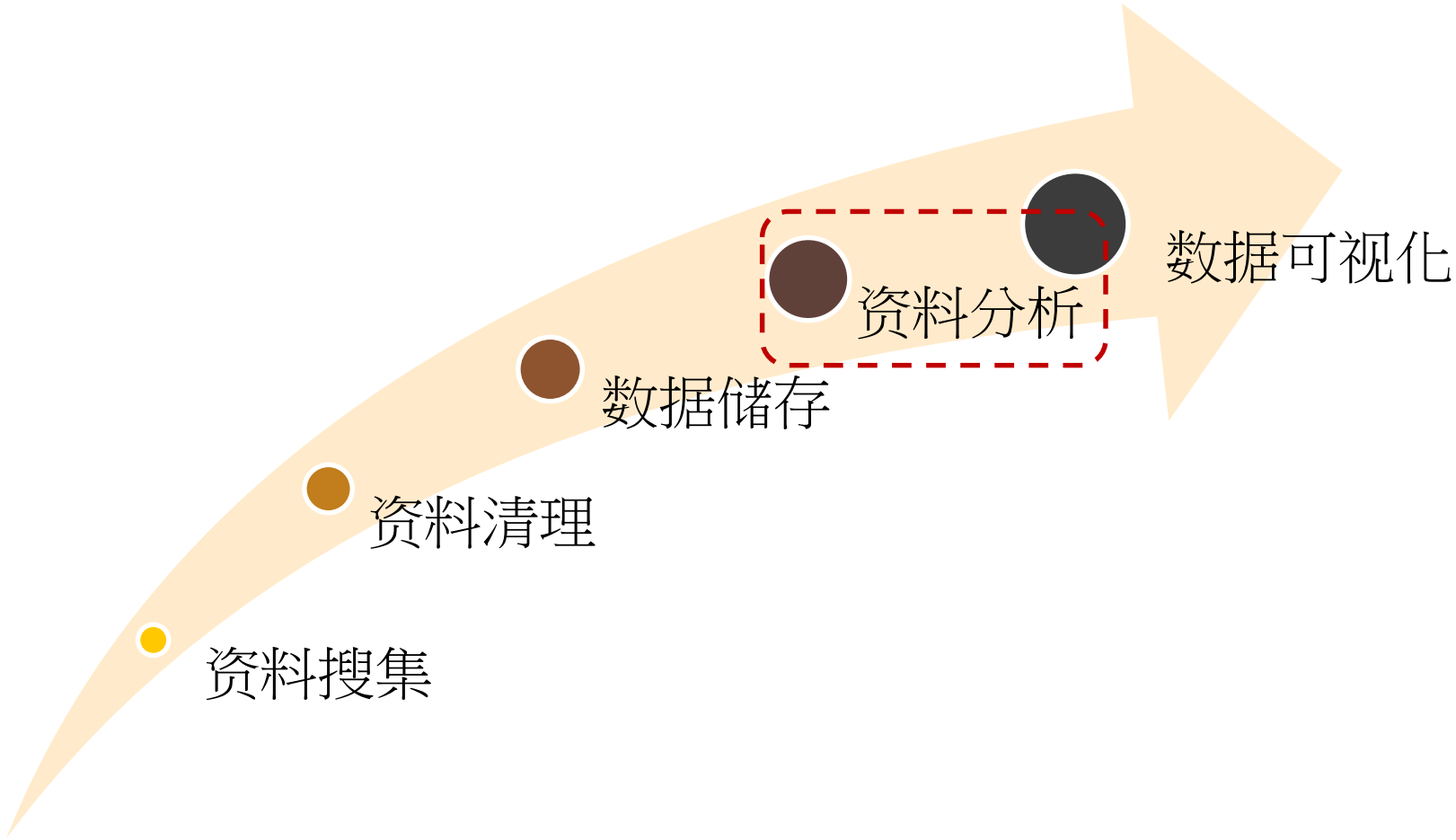
- Multi-document transactions
- Complex security needs
- Complex joins
- Extreme compression needs

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
- 评估决策或事项的效用

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

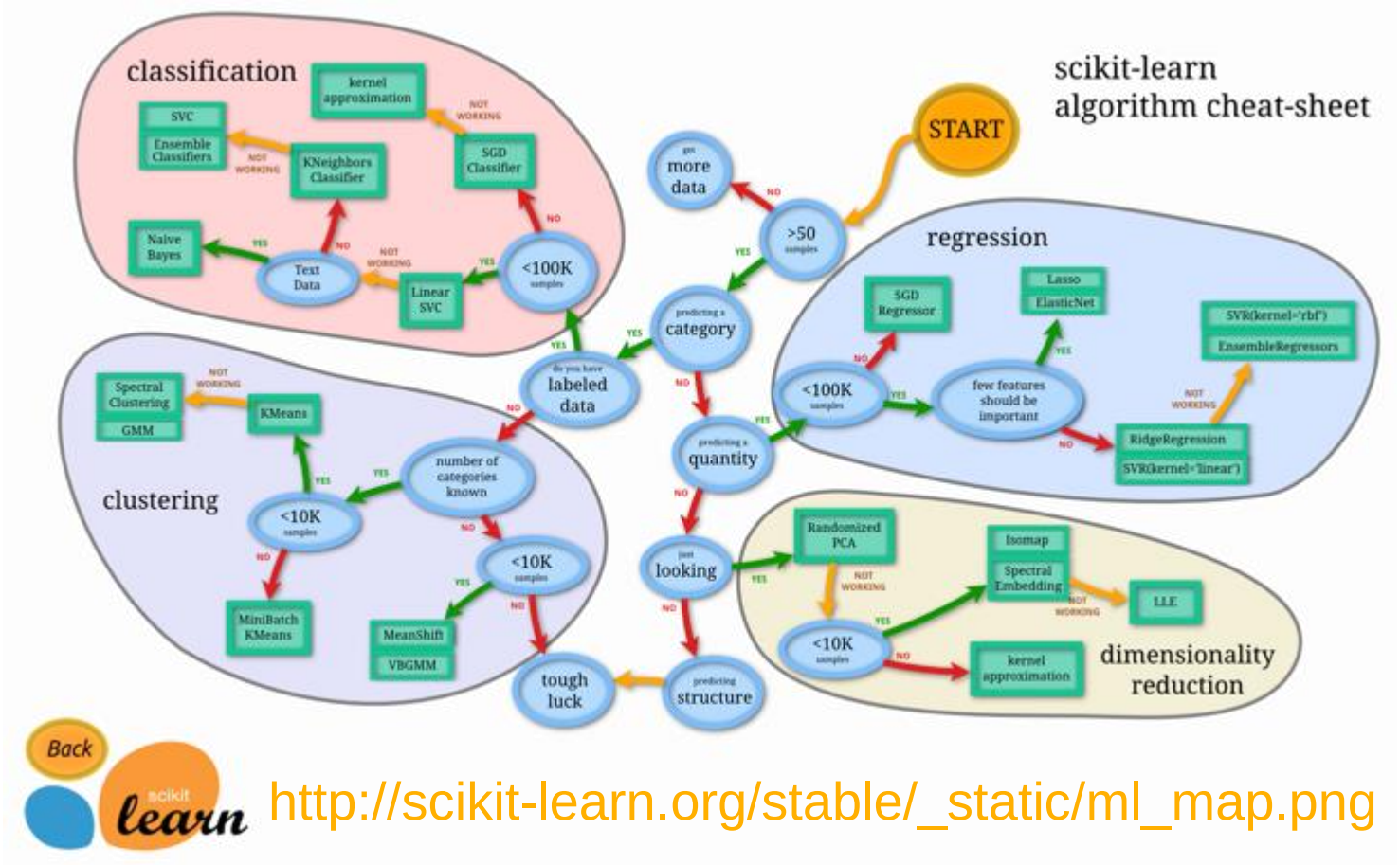
监督式学习 (Supervised Learning)

- 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


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K
more data
>50
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K
few features should be important
RidgeRegression
SVR(kernel='linear')
predicting a category
labeled data
predicting a quantity
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K
MiniBatch KMeans
<10K
MeanShift
VBGMM
tough luck
just looking
predicting structure
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K
kernel approximation
Back
scikit learn
http://scikit-learn.org/stable/_static/ml_map.png

机器学习 v.s. 统计

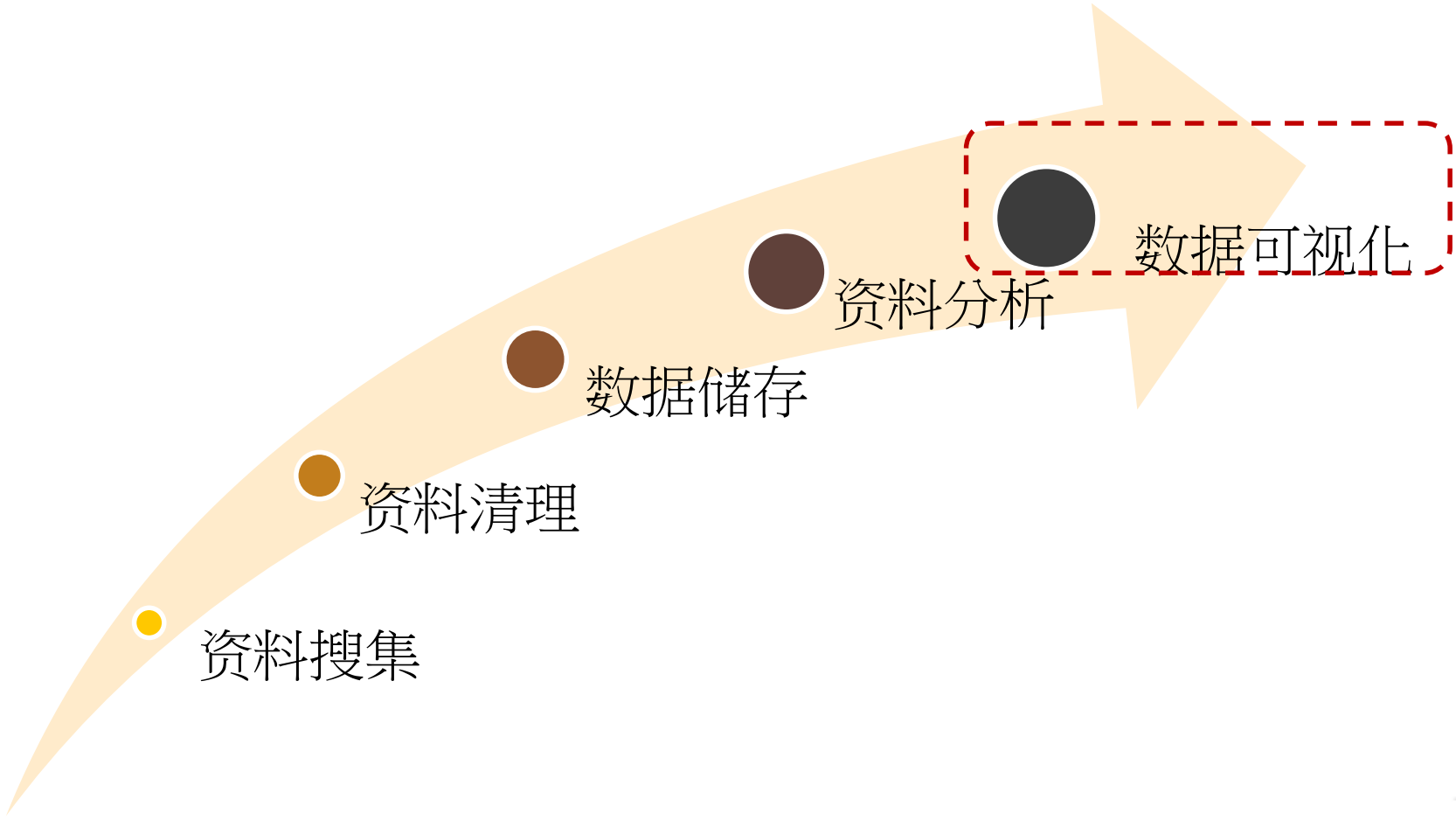
统计学家

- 专注在于模型甚于数据
- 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

- 专注在于数据甚于模型
- 用大量数据校正模型

数据科学步骤

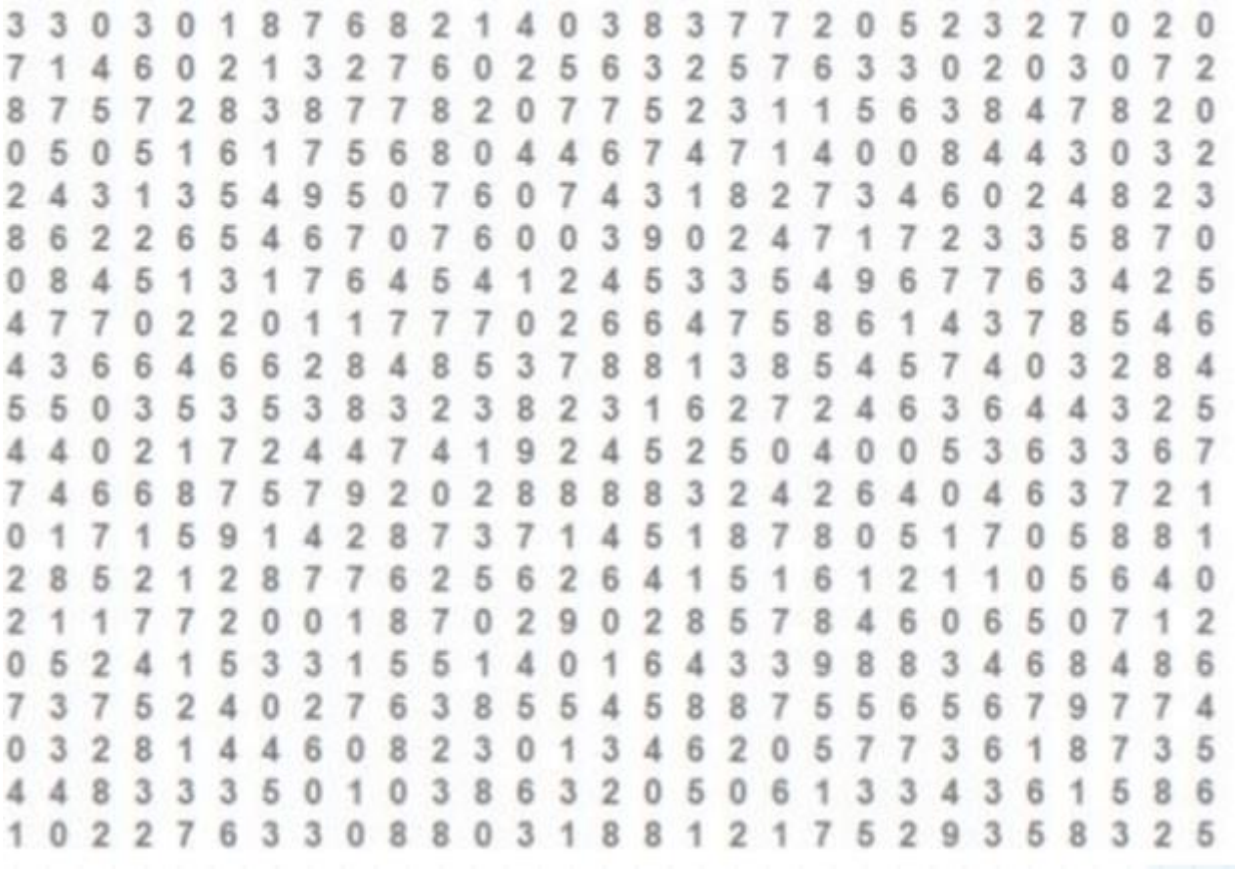


人是视觉性的动物

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 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

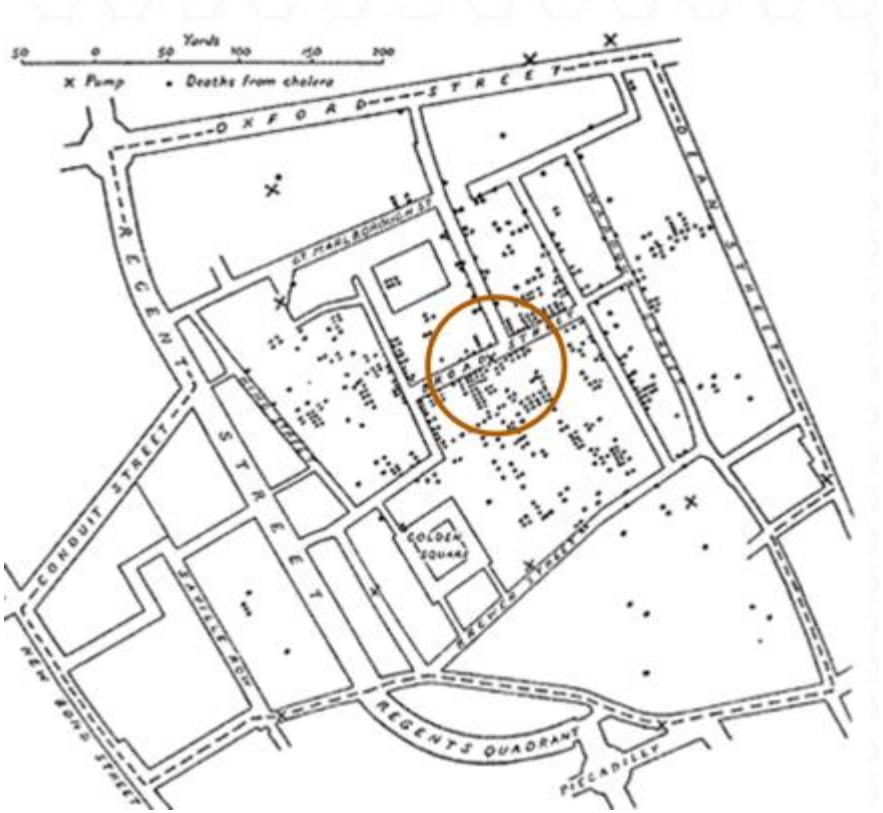
数据可视化的重要性


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 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

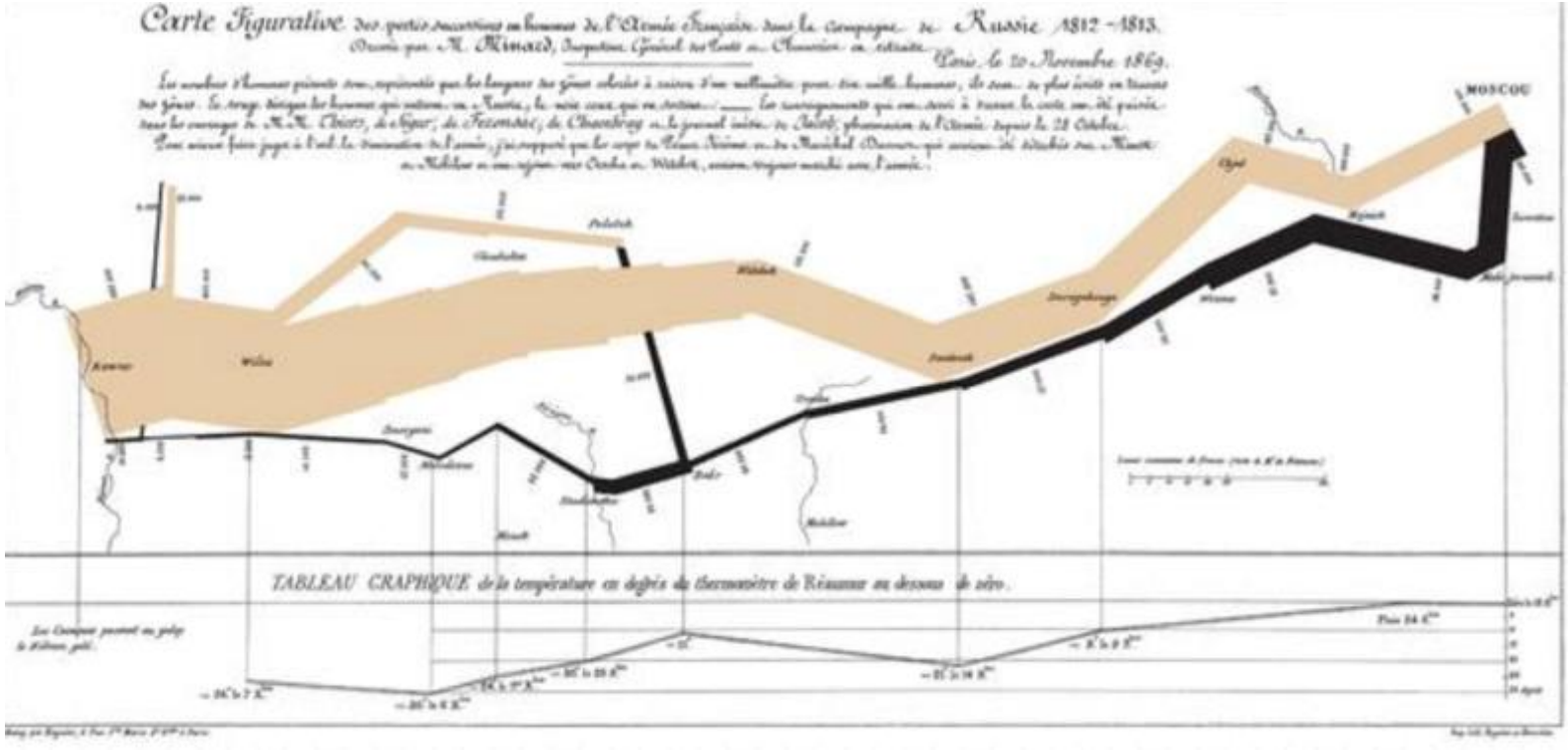
数据可视化的重要性

1854 年，霍乱疫情爆发，造成十天之内死了五百多人

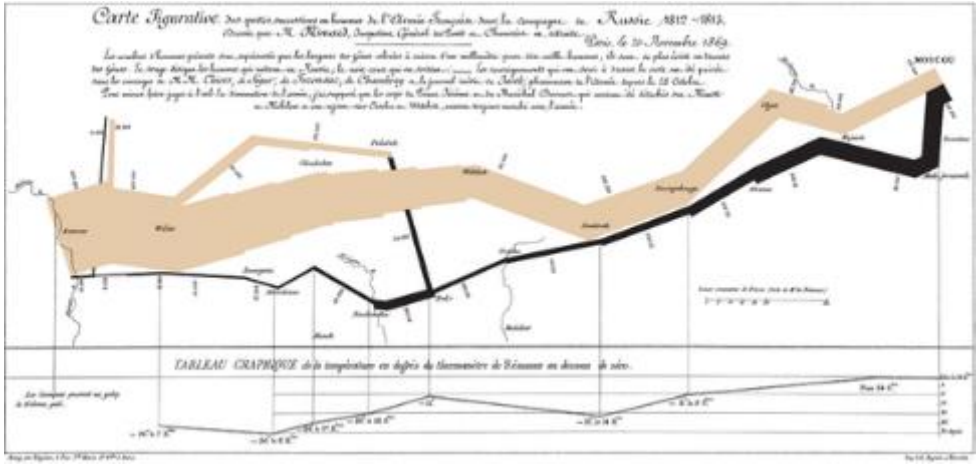
Dr. John Snow 将所有病患的住家位置点在地图上，发现病例聚集在一口井附近



拿破仑远征俄国


Carte Figurative
MOSCOU
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

信息图表的功能



Storytelling
向听众沟通已知信息

Exploration
从资料中发现背后的事实

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

https://www.youtube.com/watch?v=jbkSRLYSojo

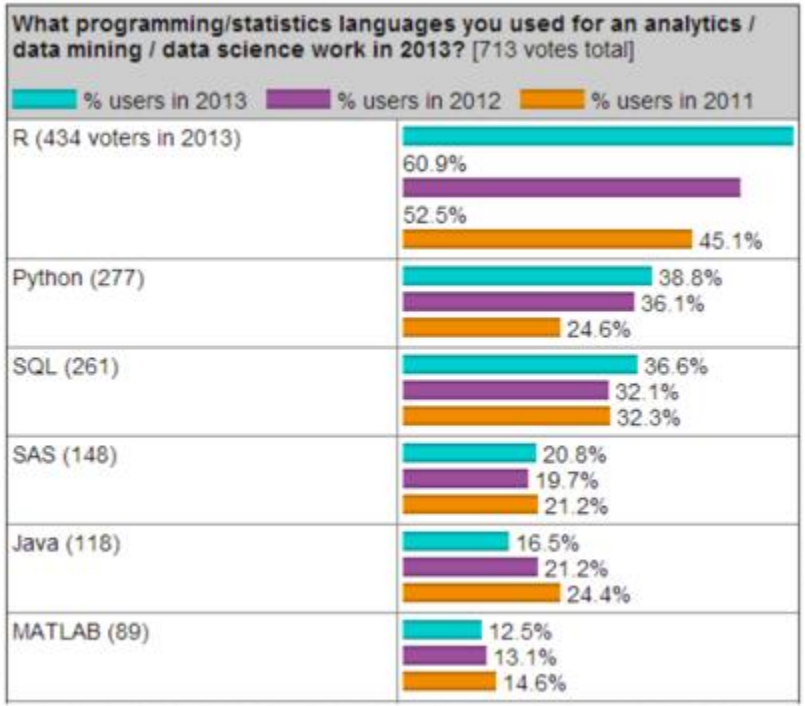
善用手边的开源工具 - 让数据科学变得超简单

数据分析语言

数据分析语言

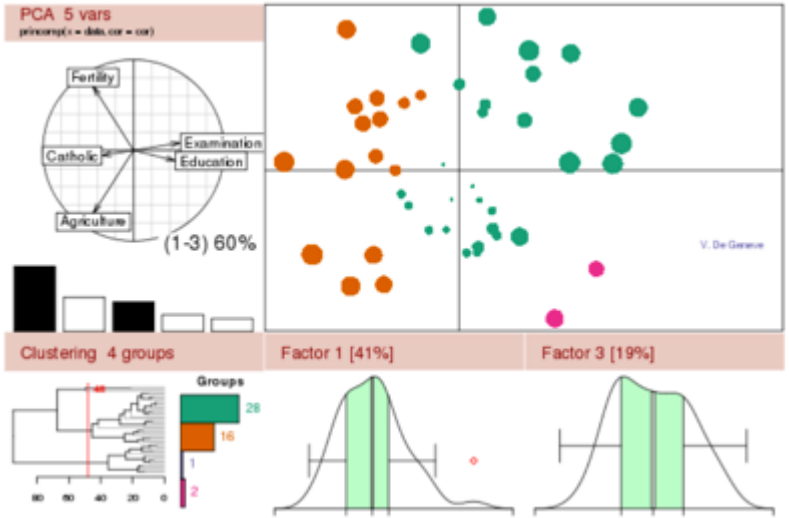
*最受歡迎的語言持續為 **R, Python (39%), 及 SQL (37%). SAS** 大約在 **20%**上下.*

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: 函式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
(https://goo.gl/2kul7Y)

Guido van Rossum
公開分享 - 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.

翻譯

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 	• Detective Work • Intelligence • Cunning • Usage of Tools • More Brain than Muscles
Python 	Superman 	• Muscle Power • Super Strength • Elegance • Wide Range • More Muscles than Brain

数据科学的学习资源与建议

Infer-ability

Model

Context

Connectedness

Variety

Variability

Velocity

Volume

Technologies

Roles

Infer, Predict, Recommend & Visualize

Contextualize, Model & Reason

De-complexify, Transform, Analyze & Network

D3.js, Dashboards, web apps

Store

R, Hive, Pig, python, Java, Mahout

Collect

Hadoop

ETL, Storm, Scribe, Flume,...

Machine Learning

Analytics

Data Architecture / Management

Data Scientist

选择你的武器

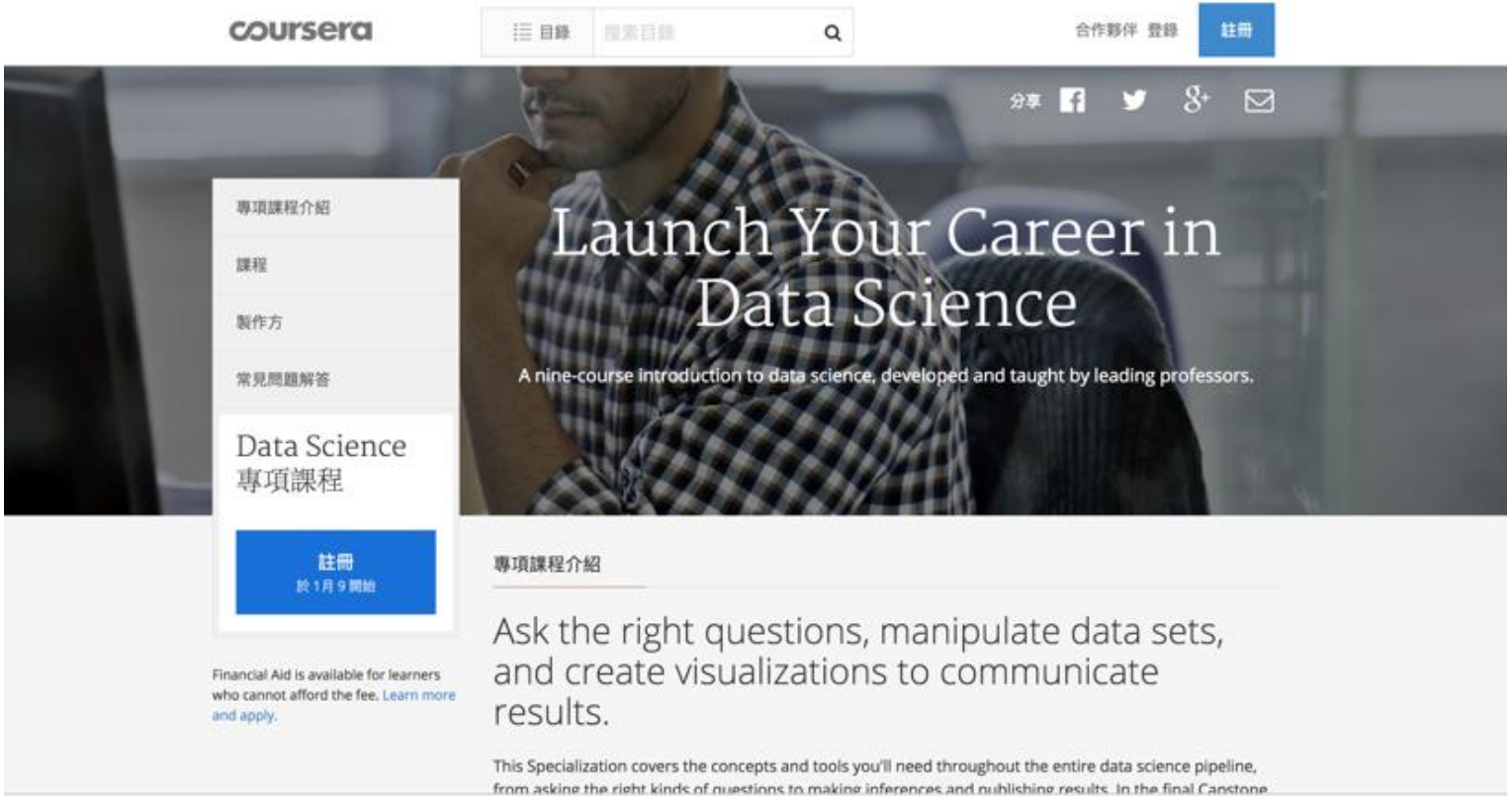
资工背景、写过程序的人、对工程面不排斥者
Python 是你们的好选择

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

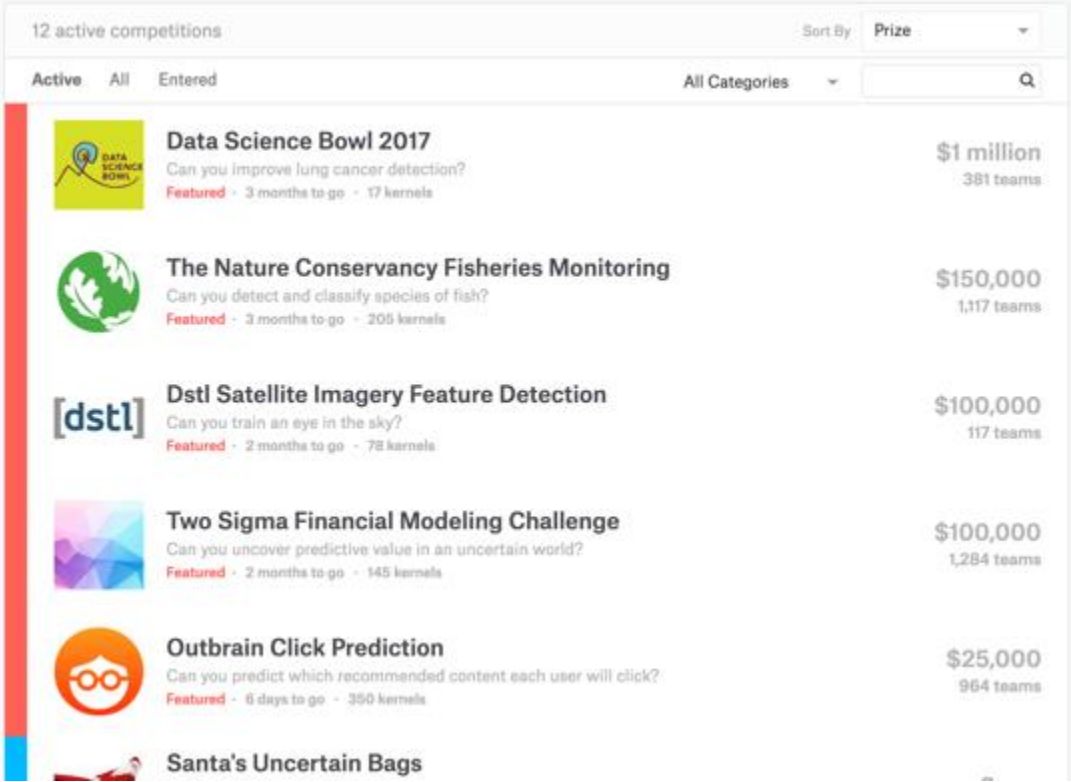
Cousera

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



Kaggle

https://www.kaggle.com/



大数学堂

http://largitdata.com/

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！


10:17
4G
49%
天善智能的团
讲师
丘祐玮
人人都爱数据科学家
Python数据科学精华实战课程
人人都爱数据科学家！Python数据科学精华实战课程
拼团价￥299.00
￥899.00
113 人参团
团长:天善智能
113 人参团/1人成团
剩余时间 13天内
参团
商品详情
火爆拼团
我的订单
更多

微信扫码，立即参团

THANK YOU

EMAIL: david@largitdata.com
网站: www.largitdta.com
电话: 0929094381