From Big Data to Data Journalism
从大数据到数据新闻

华中科技大学新闻与信息传播学院
武汉大学新闻与传播学院
2014年10月17日

Jonathan Zhu 祝建華

香港城市大學
City University of Hong Kong

Department of 媒體與傳播系
Media and Communication
香港城市大學
City University of Hong Kong

Web Mining Lab
互聯網挖掘實驗室

大纲

- 大数据的真相与误解
- 数据新闻的前生与今世
- 数据新闻的善用与误用

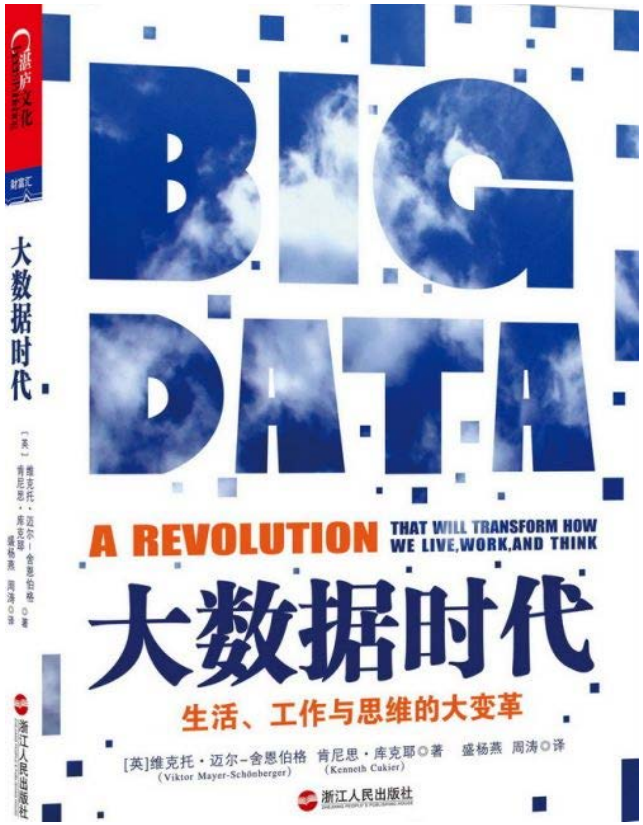
什么是大数据?

- IBM的4V定义:
  - Volume 海量
  - Velocity 快速
  - Variety 多样
  - Value 价值

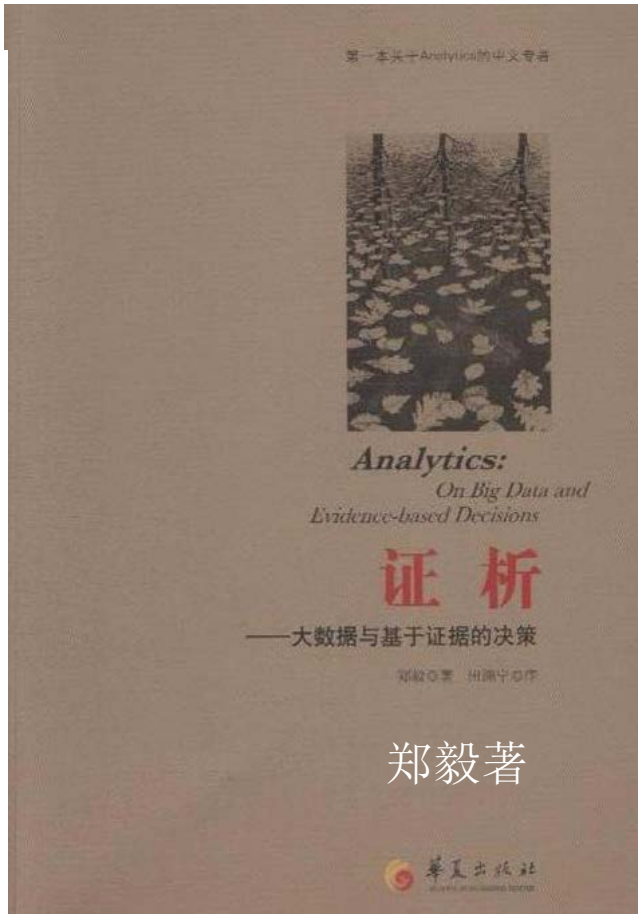
- 我的看法:

祝建华：一个文科教授眼中的大数据.《大数据中国》, 2013, V1, 10-12. http://www.china-cloud.com/dashujuzhongguo/diyiqi/2013/0517/19882.html

大数据的热门书


BIG DATA
A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK
大数据时代
生活、工作与思维的大变革
[英]维克托·迈尔-舍恩伯格 肯尼思·库克耶 著 盛杨燕 周涛 译
浙江人民出版社


The Big Data Revolution
涂子沛 著
大数据


Analytics:
On Big Data and Evidence-based Decisions
证析
——大数据与基于证据的决策
郑毅著

真相与误解

当下流行的观点	我的看法
数据量（即个案的记录数）越来越大	对，毫无疑问
数据量越大越好	对，但边际效益递减
数据信息（即个案的特征）越来越丰富	往往相反
处理大数据的技术已经成熟	言过其实

i. 为什么数据量越来越大？

传统数据来源：

- 政府统计机构
- 经济、金融
- 天文、地理
- 交通、运输
- 传统媒体
- 等等

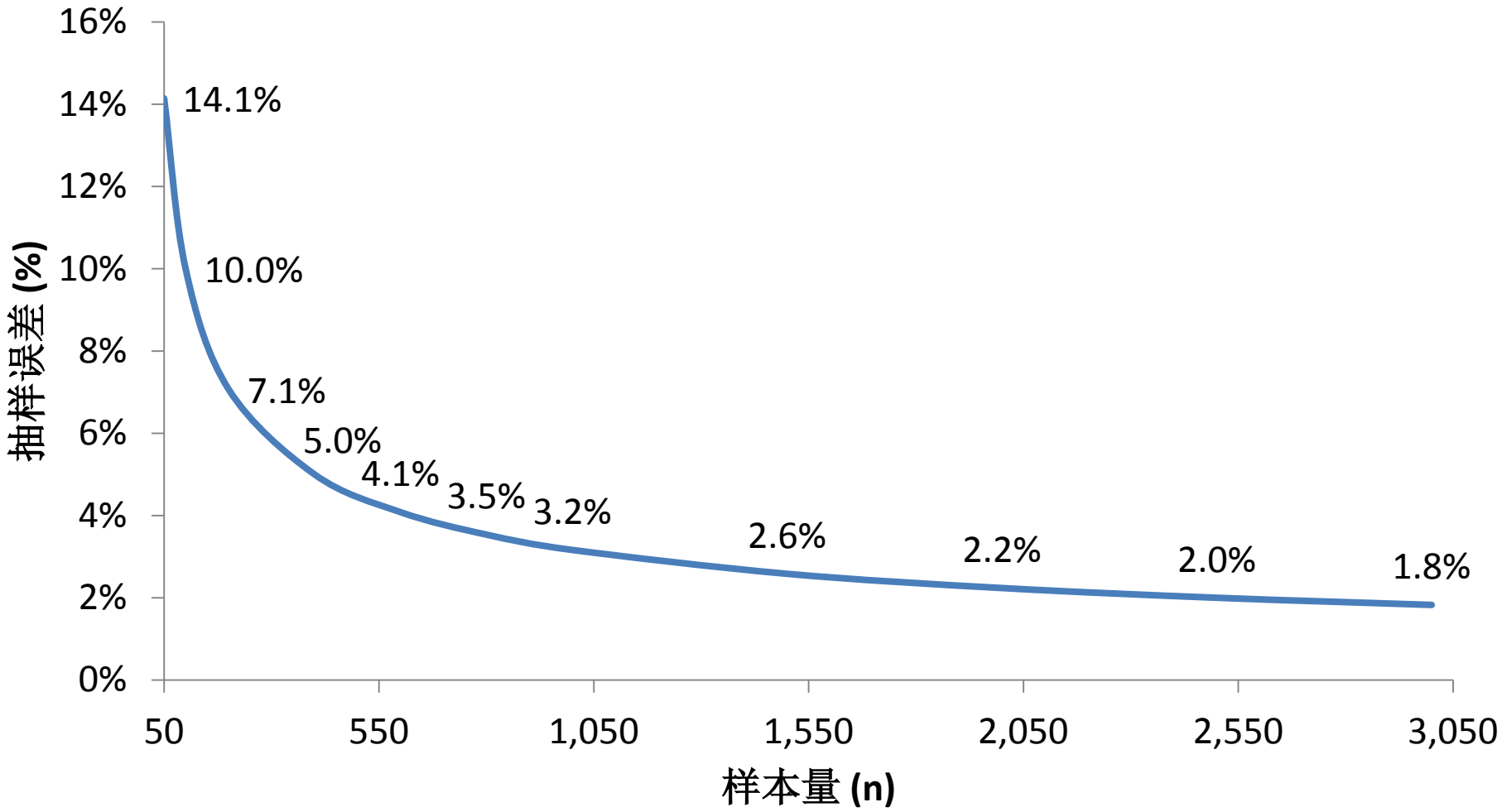
新型数据来源：

- 互联网
- 移动网
- 智能家居
- 物联网
- 生物工程/DNA
- 等等

以电视收视率数据为例(武汉或任一大城市)

	50-70年代	80-00年代	10年代+
采集手段	日记填写	人员记录仪	DTV
样本家庭数	1,000	1,000	3,000,000
时间单位	15分钟	15秒钟	10毫秒
数据量(条/天)	96,000	日记的60倍	日记的2.7亿倍; 人员仪的450万倍

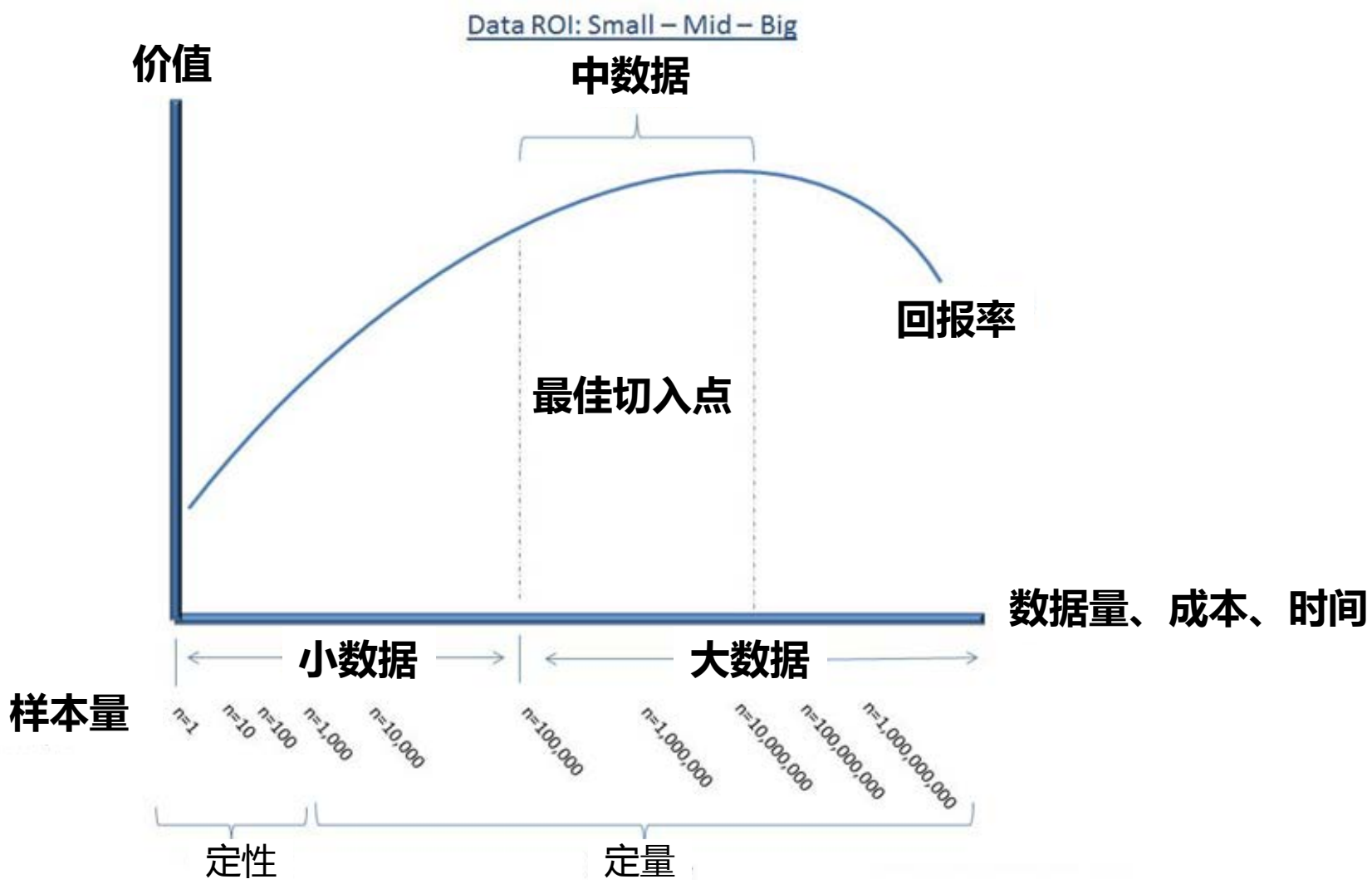
ii. 为什么数据量越多越好？



还以电视收视率数据为例

样本(家庭数)	人员仪 (1,000)	DTV总体 (3百万)	DTV样本 (100,000)	DTV样本 (10,000)
抽样误差	3.2%	0%	0.3%	0.4%
原始时间单位	15秒钟	10毫秒	10毫秒	10毫秒
数据量(条/天)	576万	人员仪的450万倍	人员仪的15万倍	人员仪的1.5万倍
抽样时间单位	--	1秒	1秒	1秒
数据量(条/天)	--	人员仪的45,000倍	人员仪的1,500倍	人员仪的150倍

“中数据”最优化



Tom Anderson (2013) Forget big data, think mid data. http://blog.odintext.com/?p=300

iii. 为什么大数据的信息反而不丰富?

传统小数据: 个案不多但变量很多

ID	X1	X2	…	X_j	Y1	Y2	…	Y_k
1	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
n	…	…	…	…	…	…	…	…

难道说“大数据信息更丰富”的人没有见过真实的大数据是什么模样的？

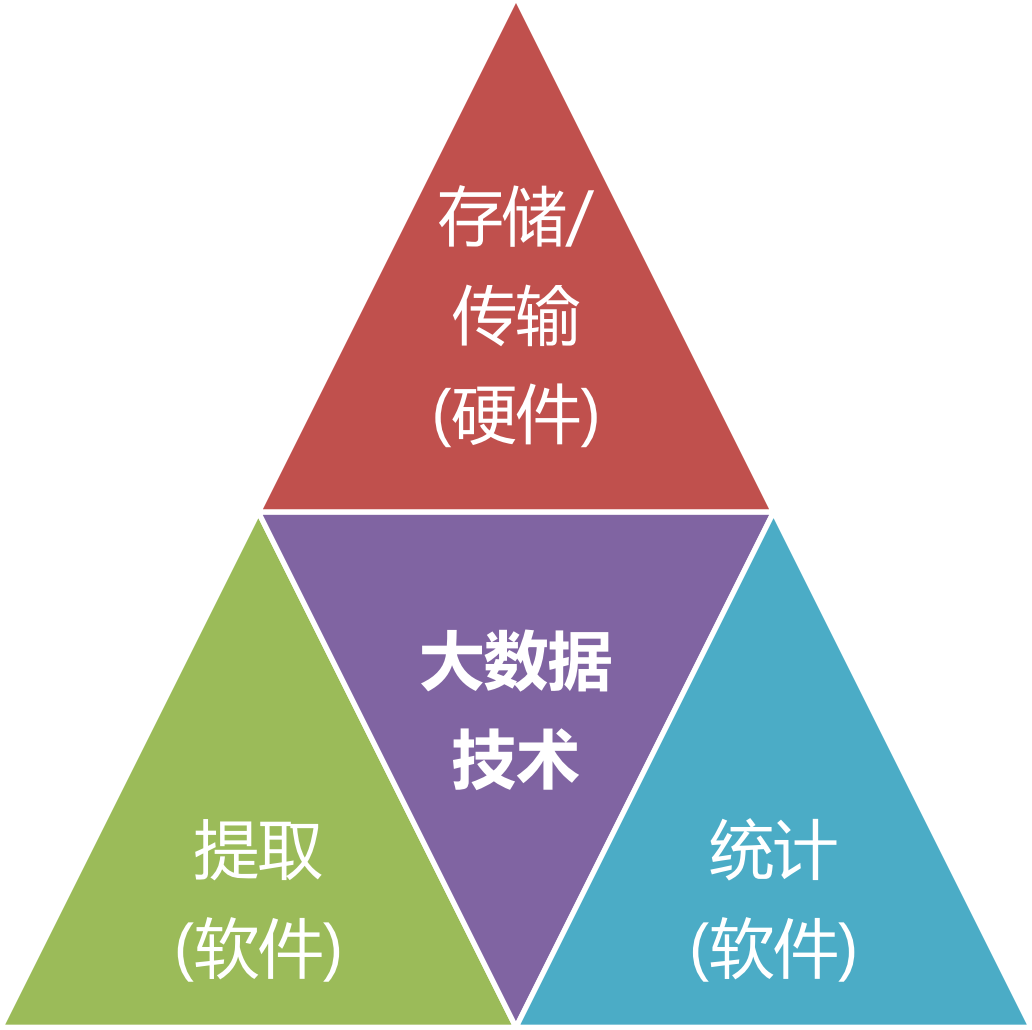
理想的大数据：海量个案海量变量

ID	X1	X2	…	X_j	Y1	Y2	…	Y_k
1	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
n	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
∞	…	…	…	…	…	…	…	…

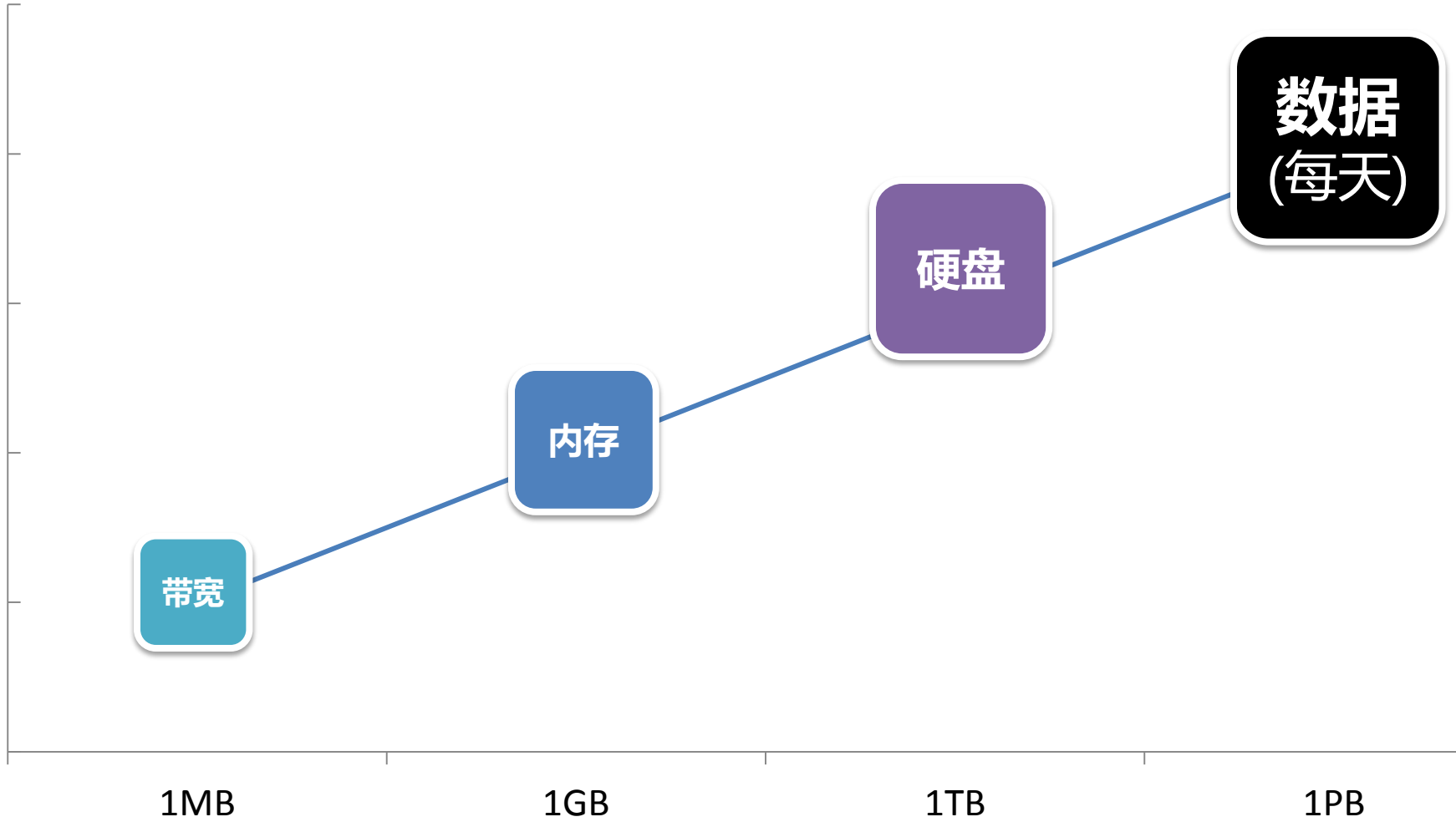
现实中的大数据：海量个案极少变量

ID	X1	X2
1	…	…
…	…	…
n	…	…
…	…	…
∞	…	…

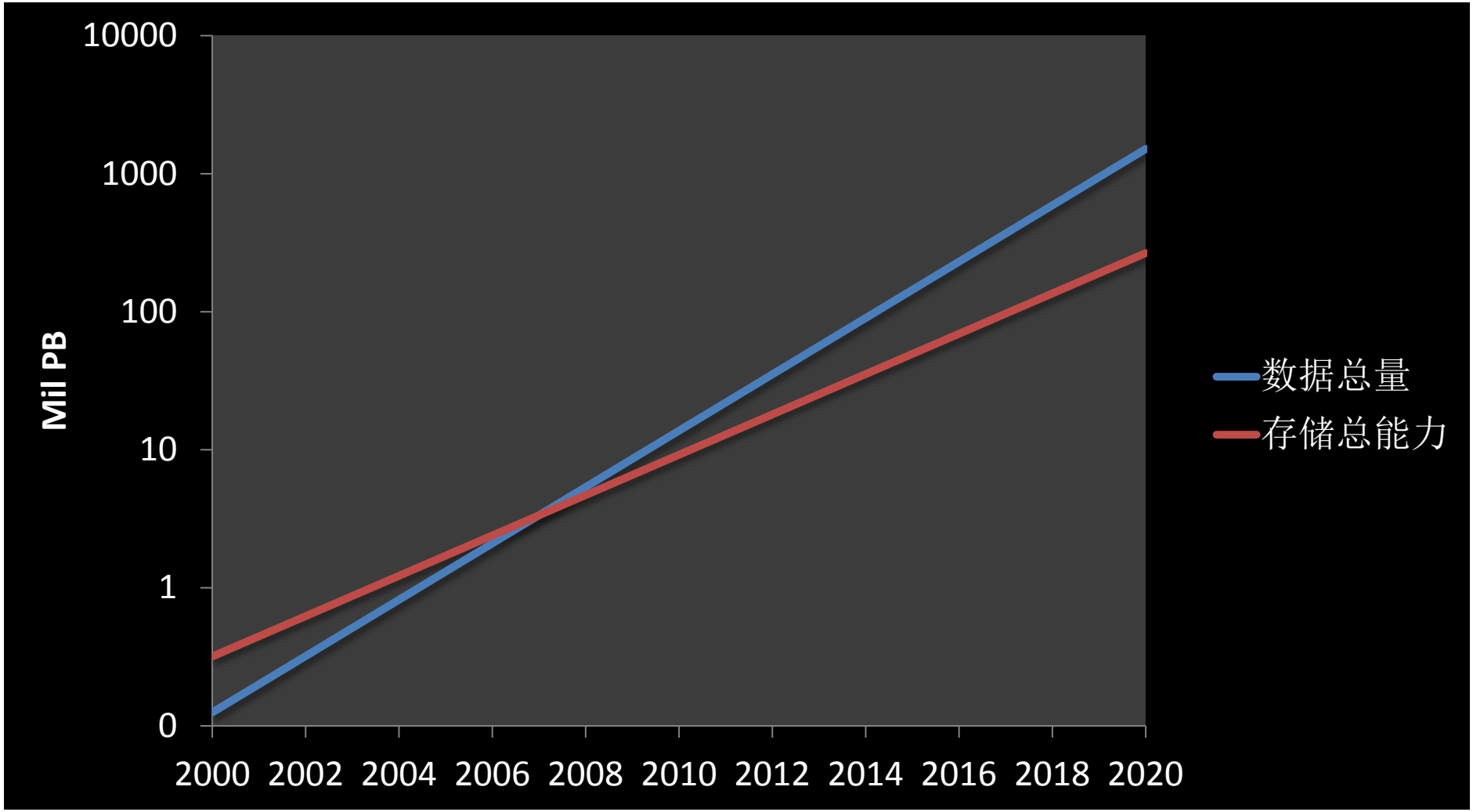
iv. 为什么大数据的技术并不成熟？



数据规模与硬件规模的竞赛



全球数据量与存储能力的剪刀差



适用于大数据的统计分析方法还在襁褓之中

现在用于分析大数据的所有统计方法，都是100余年前发明的、或在此基础上而改进的。这些经典方法以小数据的正态分布（或t、F和卡方分布）为前提，对大数据（往往是极度偏差的幂律分布）并不合适。这是大数据技术尚不成熟的一个最重要标志。

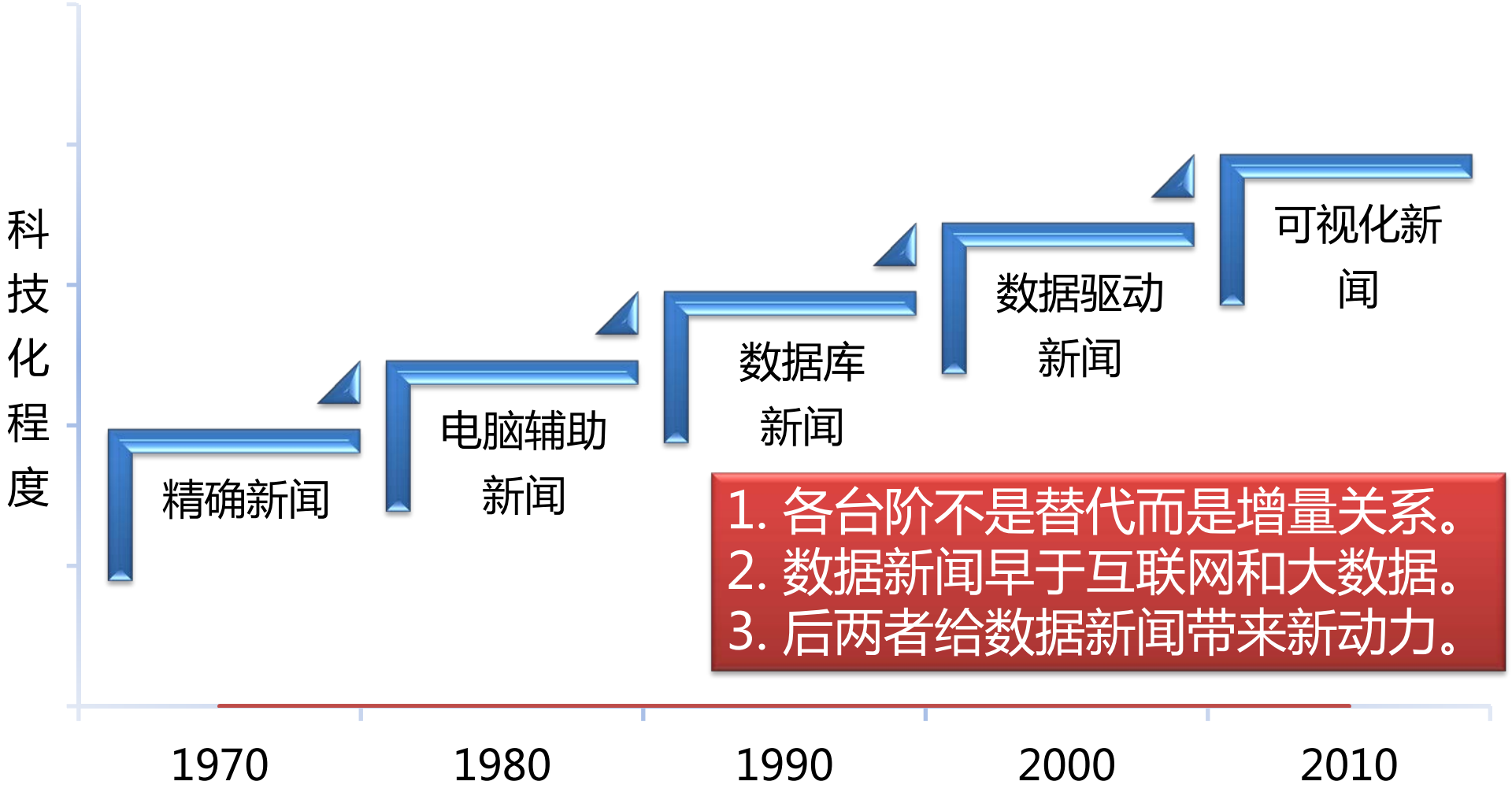
《科学》2011年发表的论文
David Reshef, Yarik Reshef, et al. Detecting novel associations in large data sets using maximal information coefficient (MIC)

相当于1890年皮尔森发明的经典相关系数。

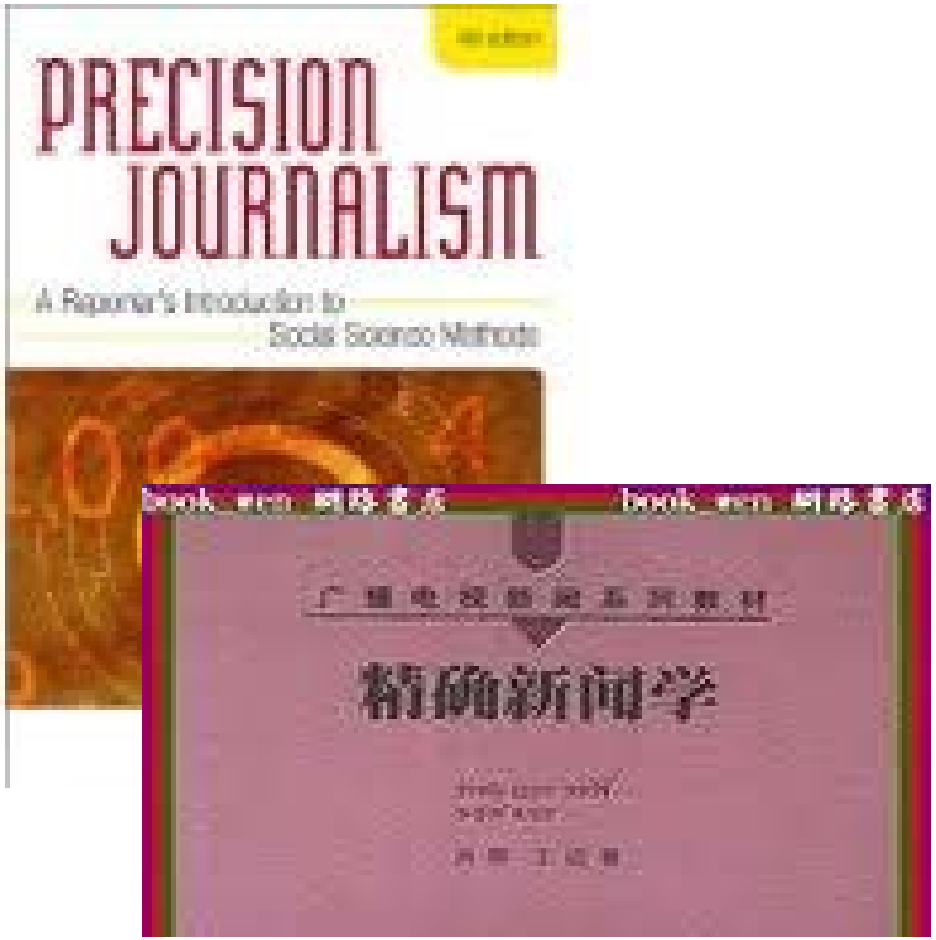
大纲

- 大数据的真相与误解
- 数据新闻的前生与今世
- 数据新闻的善用与误用

数据新闻的演化路径


科技化程度
精确新闻
电脑辅助新闻
数据库新闻
数据驱动新闻
可视化新闻
1. 各台阶不是替代而是增量关系。
2. 数据新闻早于互联网和大数据。
3. 后两者给数据新闻带来新动力。
1970
1980
1990
2000
2010

精确新闻 Precision Journalism



Philip Meyer (1973): *Precision Journalism: A Reporter's Introduction to Social Science Methods* (精确新闻学：记者的社会科学方法入门)

- 抽样原理
- 调查方法
- 数据统计
- …

报道调查数据时必须提供的技术细节

- 调查赞助者（如果有，必须报告）
- 调查日期和地点
- 调查对象（如成年居民、常住居民、选民、等等）
- 抽样方法（随机还是便利，具体如何抽取）
- 样本人数（及其对应的抽样误差）
- 访问成功率（$=\frac{\text{成功访问人数}}{\text{合格被访人数}}$，一般按美国民意研究协会的具体公式计算和报告）
- 问题与答案的原话
- 等等

透明、公开、防误导、免操控

电脑辅助报道与数据库新闻

电脑辅助报道 Computer-Assisted Reporting (CAR)

- 采用电脑软件帮助新闻采访、编辑与写作
- 美国全国电脑辅助报道研究所 (NICAR)
- 丹麦国际分析报道中心 (DICAR)

IRE INVESTIGATIVE REPORTERS & EDITORS

数据库新闻 Database Journalism (DBJ)

- 采用数据库挖掘新闻、整合不同来源信息、建设结构化新闻系统
- 华盛顿邮报 Fixing DC's School
- Adrian Holovaty EveryBlock ($1.1m grant by Knight-Ridder)
- Chrinon Ltd. OpenCorporates

数据驱动新闻(DDJ)

Data-driven journalism is a journalistic process based on analyzing and filtering large data sets for the purpose of creating a new story. Data-driven journalism deals with open data that is freely available online and analyzed with open source tools. Data-driven journalism strives to reach new levels of service for the public, helping consumers, managers, politicians to understand patterns and make decisions based on the findings. As such, data driven journalism might help to put journalists into a role relevant for society in a new way. (http://en.wikipedia.org/wiki/Data_driven_journalism)

数据驱动新闻是一个通过分析和过滤大型数据而制作新闻故事的过程。数据驱动新闻采用网上免费的开放数据，并用开源工具进行分析。数据驱动新闻旨在为公众提供新层次的服务，帮助消费者、管理者、决策者理解(现象)模式并根据数据结果而作出决策。数据驱动新闻因此而将记者推到了一种与社会相关的新角色。

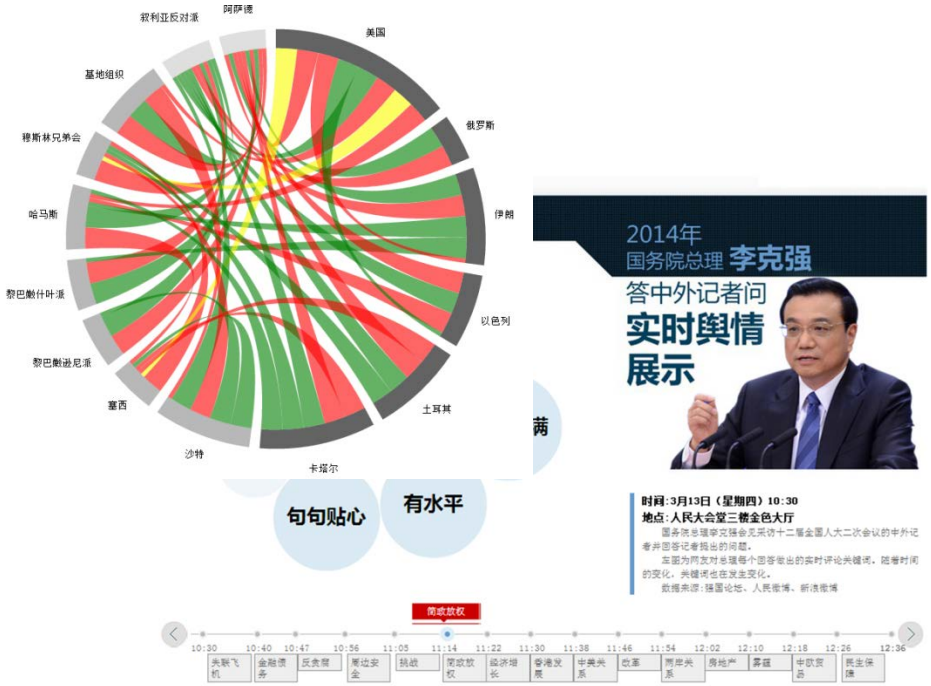
可视化新闻分类

1. 作为新闻**主体**：可视化就是新闻，一张图告诉你所有的故事
2. 作为新闻**主题**：可视化是故事的框架或流程，文字围着转
3. 作为新闻**导语**：可视化引发出故事，先图后文
4. 作为新闻**插图**：可视化配合文字，提供背景，帮助理解

新闻主体vs.新闻主题

可视化作为新闻主体：

- 中东地区的敌友关系
- 李克强记者会舆论反馈



可视化作为新闻主题：

- 大陆土壤重金属污染史

数据新闻三大范式之比较

	精确新闻	CAR-DBJ-DDJ	可视化新闻
表述手段	文字	分析	图像
关键词	准确、严谨、透明	探秘、深入、确凿	简化、形象、互动
局限	应用面狭隘（如与财经新闻隔离）	依赖现存数据库和记者定量分析能力	片面追求形式、喧宾夺主、游离主题

大纲

- 大数据的真相与误解
- 数据新闻的前生与今世
- 数据新闻的善用与误用

数据与新闻的关系

马金馨(数据新闻网联合创始人)：

可视化与数据的关系

Jonathan Zhu (2014)

- Data visualization differs from the general graphic design in that it is of the data, by the data, and for the data.
  - Of the data: an integrated phase of the discovery rather than a post-analysis phase to decorate the findings
  - By the data: guided primarily by data results rather than esthetical considerations
  - For the data: to tell accurate, informative, and understandable quantitative stories

祝建华:

- 数据可视化与一般艺术可视化的不同之处，是数据可视化
  - **发源于**数据(of the data)
  - **听命于**数据(by the data)
  - **服务**于数据(for the data)

可视化新闻的常见误用或滥用

症状：

- 表达不准确、失真
- 漏报或瞒报关键的方法细节（详见精确新闻部分）
- 为形式而形式、喧宾夺主、游离主题、没有真正的故事
- 过分复杂、难以理解

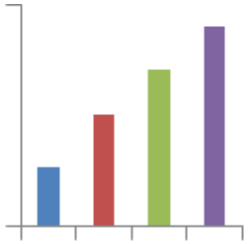
成因：

- 数据本身的复杂性、抽象性、多维性
- 误将信息可视化等同于艺术可视化，前者追求准确而后者追求夸张和戏剧性效果

数据可视化的九个主打形式

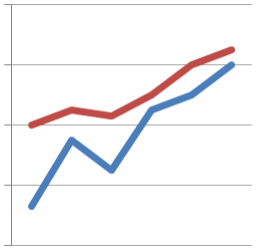
- 直方图 (比较)
- 饼图 (比例、份额)
- 线图 (趋势)
- 散点图 (相关关系)
- 时间轴 (演化进程)
- 甜圈图 (多维比例)
- 雷达图 (多维比较)
- 地图 (地理位置)
- 热力图 (强度)
- 等等

Bar

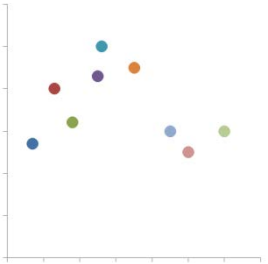


Pie

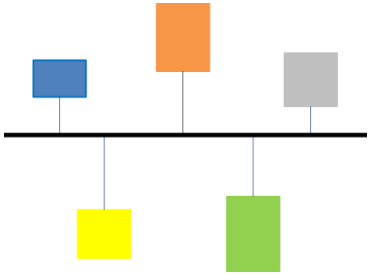
Line



Scatter Plot

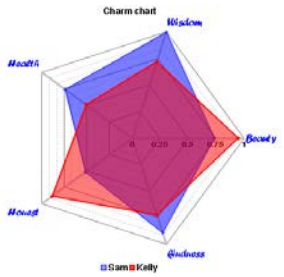


Time Line

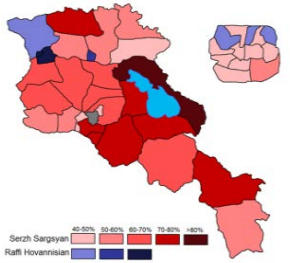


Doughnut

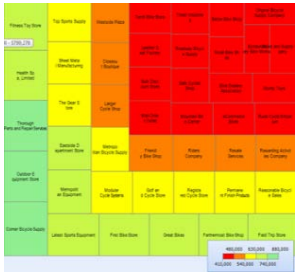
Radar



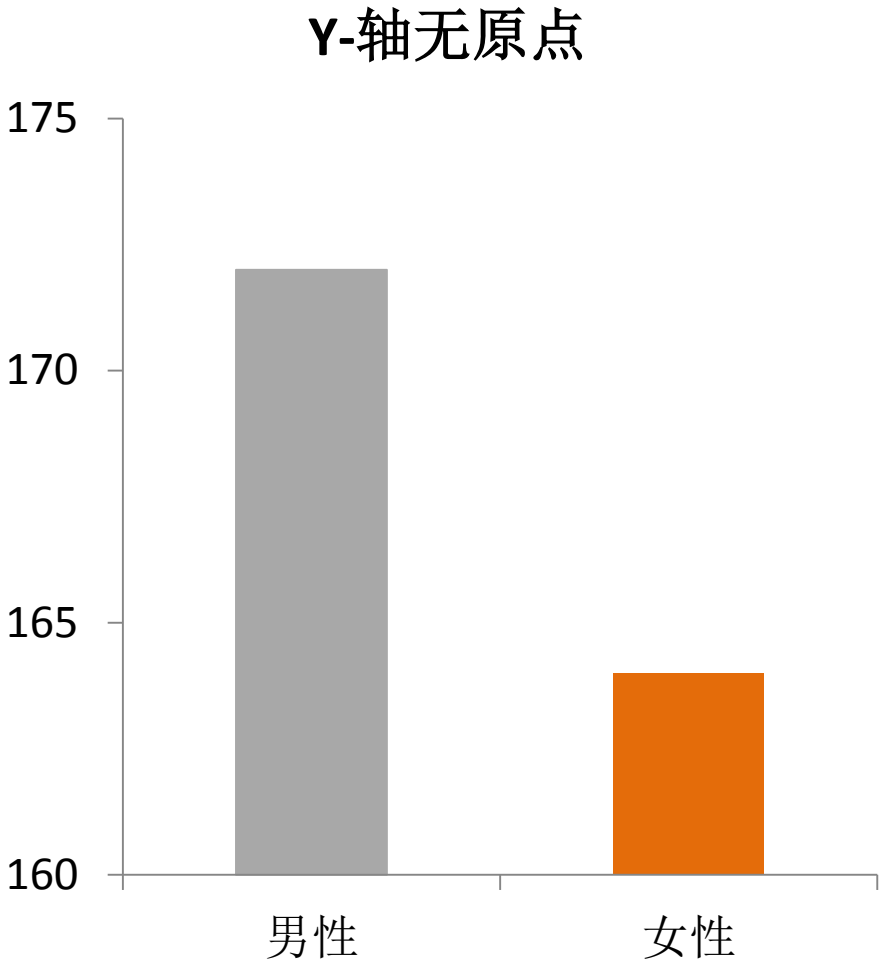
Map

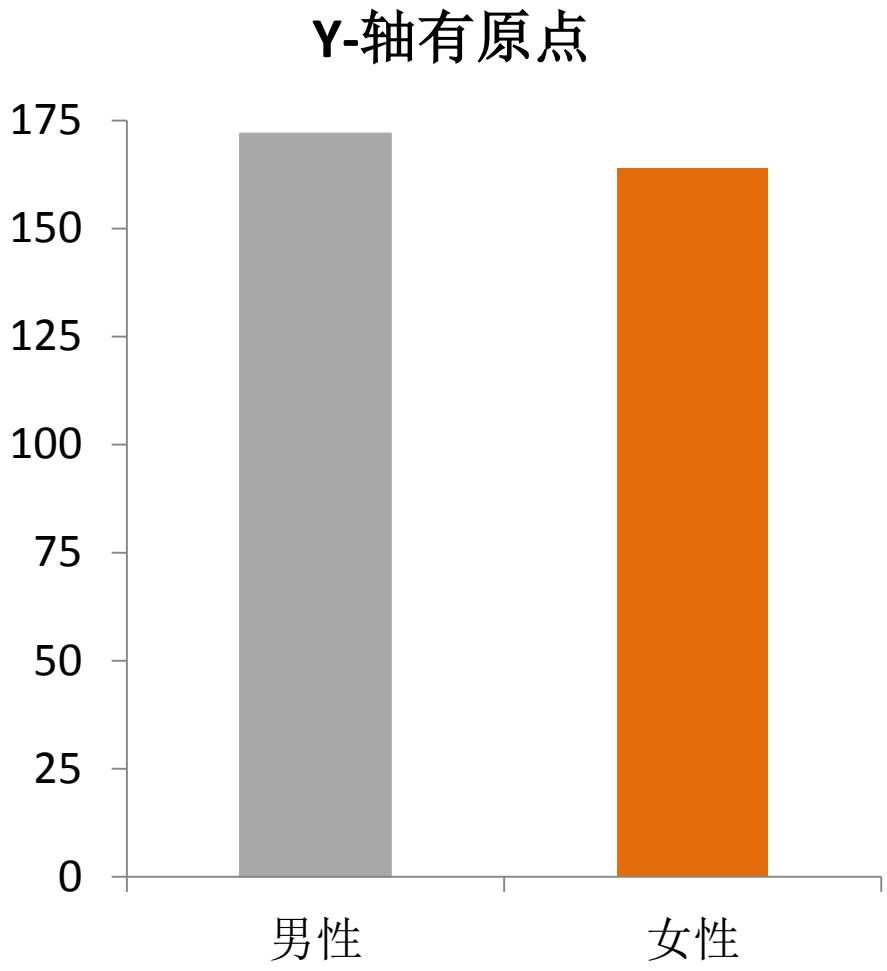


Heat Map

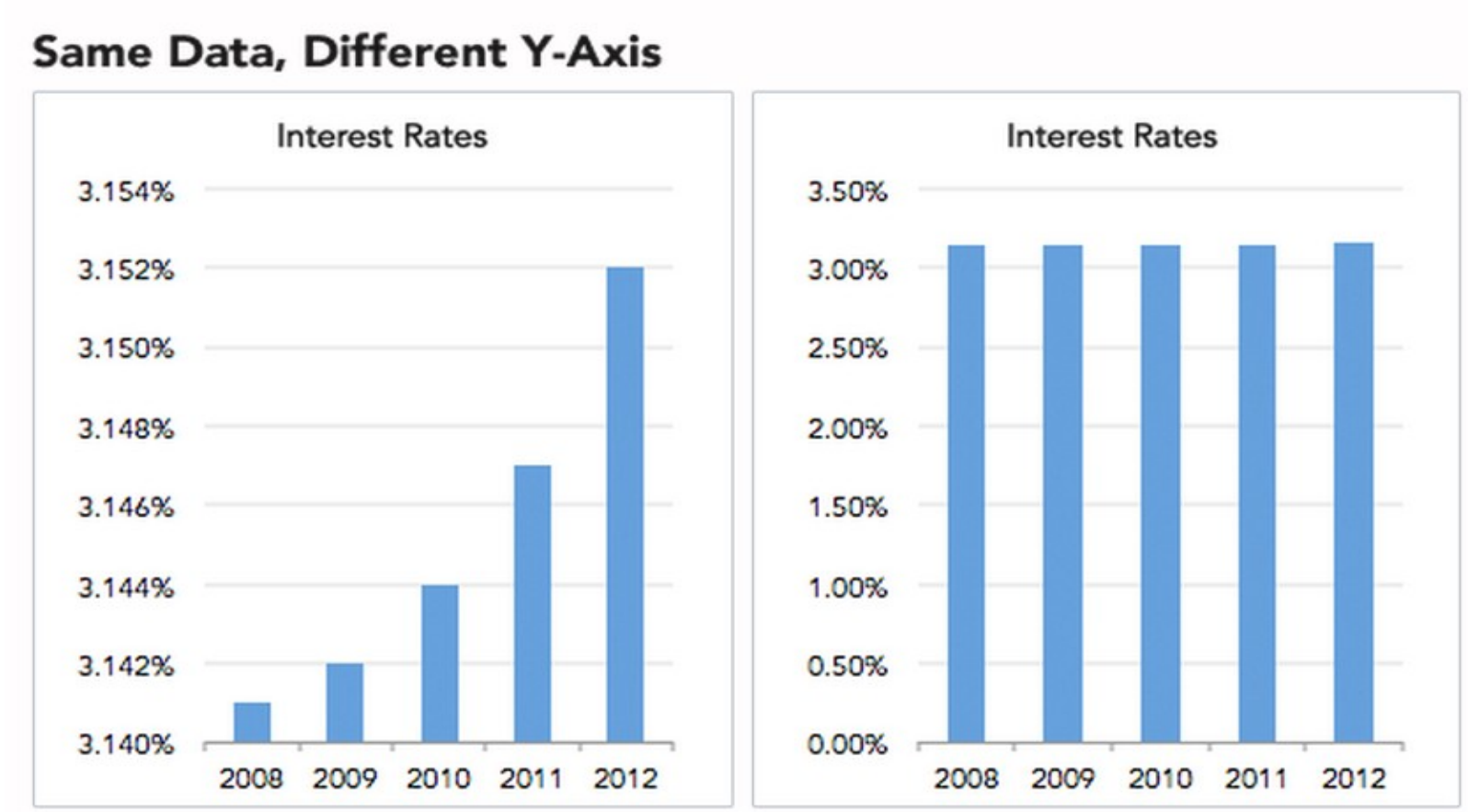


Y-轴必须含原点(即0值)



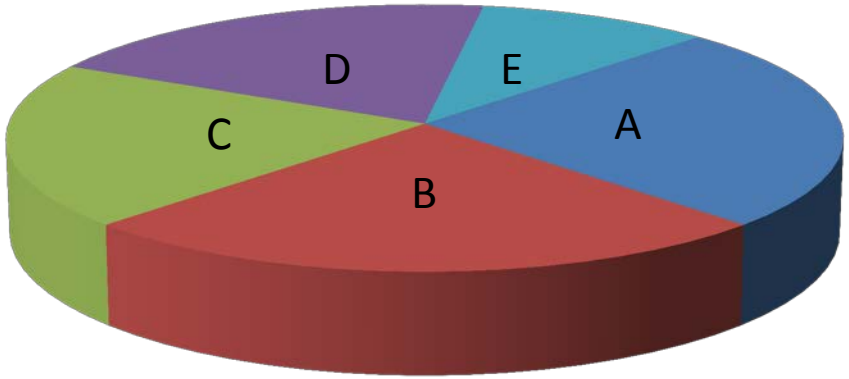


Y-轴缺原点：夸大比较对象之间的差别



Source: http://data.heapanalytics.com/how-to-lie-with-data-visualization/

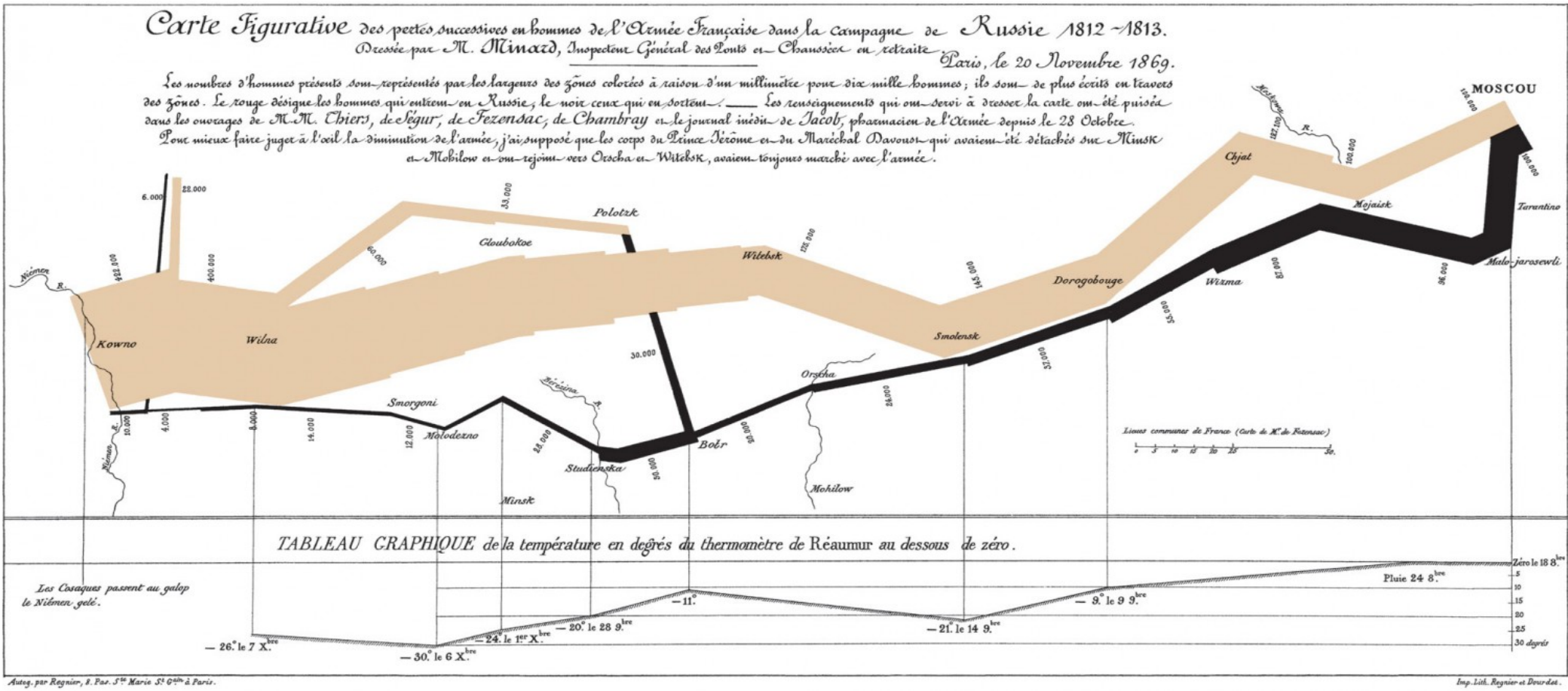
不要用3D饼图以免数据失真



谁大谁小？

谁大谁小？

数据可视化经典之作：拿破仑在1812
(经度、纬度、方向、时间、温度、人数的六维度展示)

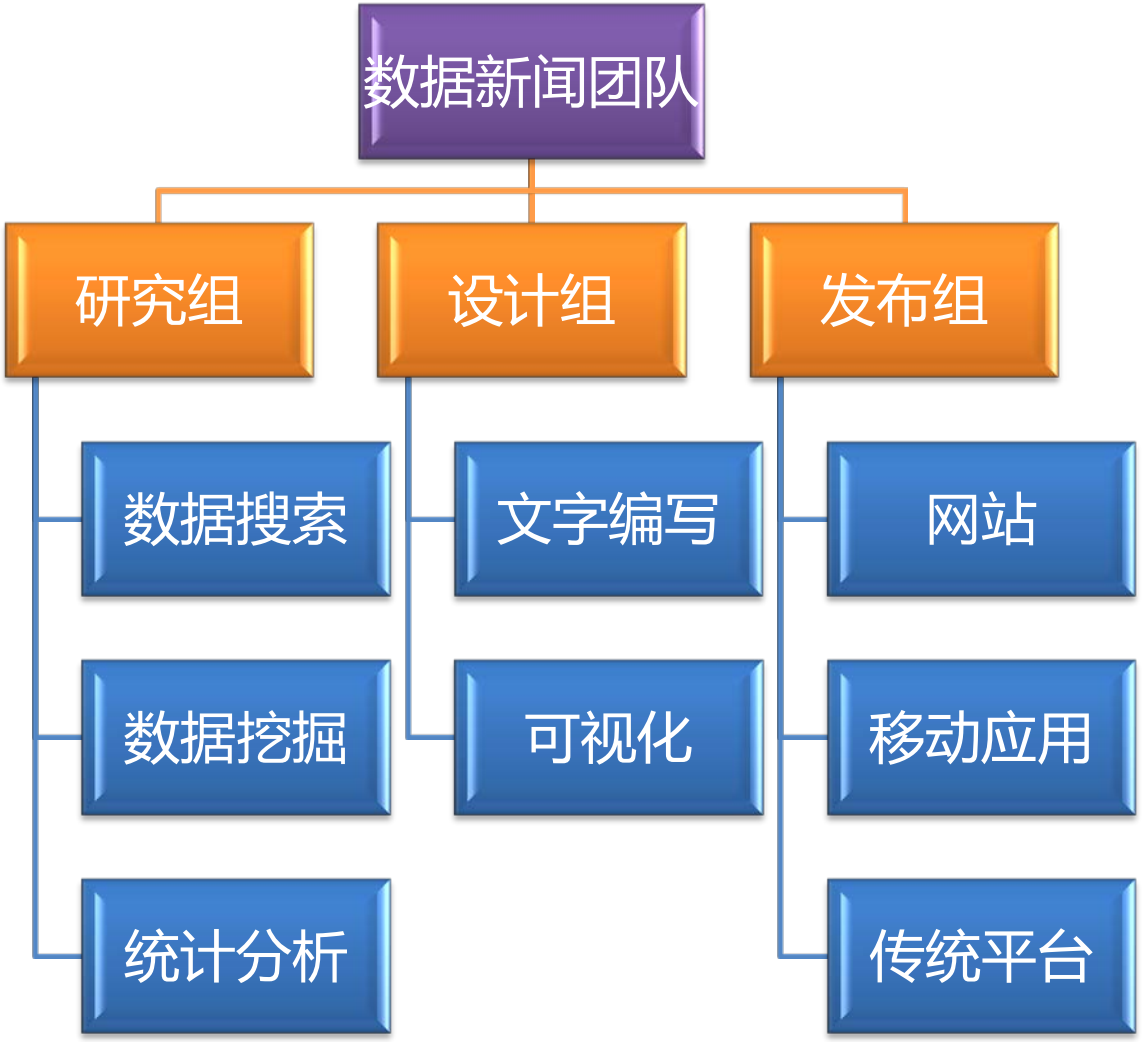


Source: http://thumbnails.visually.netdna-cdn.com/napoleons-march-to-moscow-the-war-of-1812_50290b656ab82_w1500.png

多维数据与互动可视化

- 数据的维度：
  - 1维：对1个指标（如产量）的单独分析或展示
  - 2维：对2个指标（如产量与年份）的交叉分析或展示
  - 多维：对3+个指标（如不同产品与年份）的嵌套交叉分析或展示
  - 对高维数据的分析不难；只在一个2维空间里展示高维数据则是一个大难题
- 互动可视化
  - 互动媒体（光碟、互联网等）是实现高维数据可视化的有效途径
  - “互动”可视化是指用户可在媒体界面上选择第3个或更多个指标来改变展示内容（即逐层钻取）
  - 与此相比，平面和视频媒体无法提供逐层钻取，因此展示的高维数据势必庞杂难懂

数据新闻团队的结构



数据新闻团队应该由谁主导？

数据故事化

理想的数据新闻：讲述准确而形象的数据故事

新闻艺术化

新闻数据化

数据新闻记者是如何练成的？

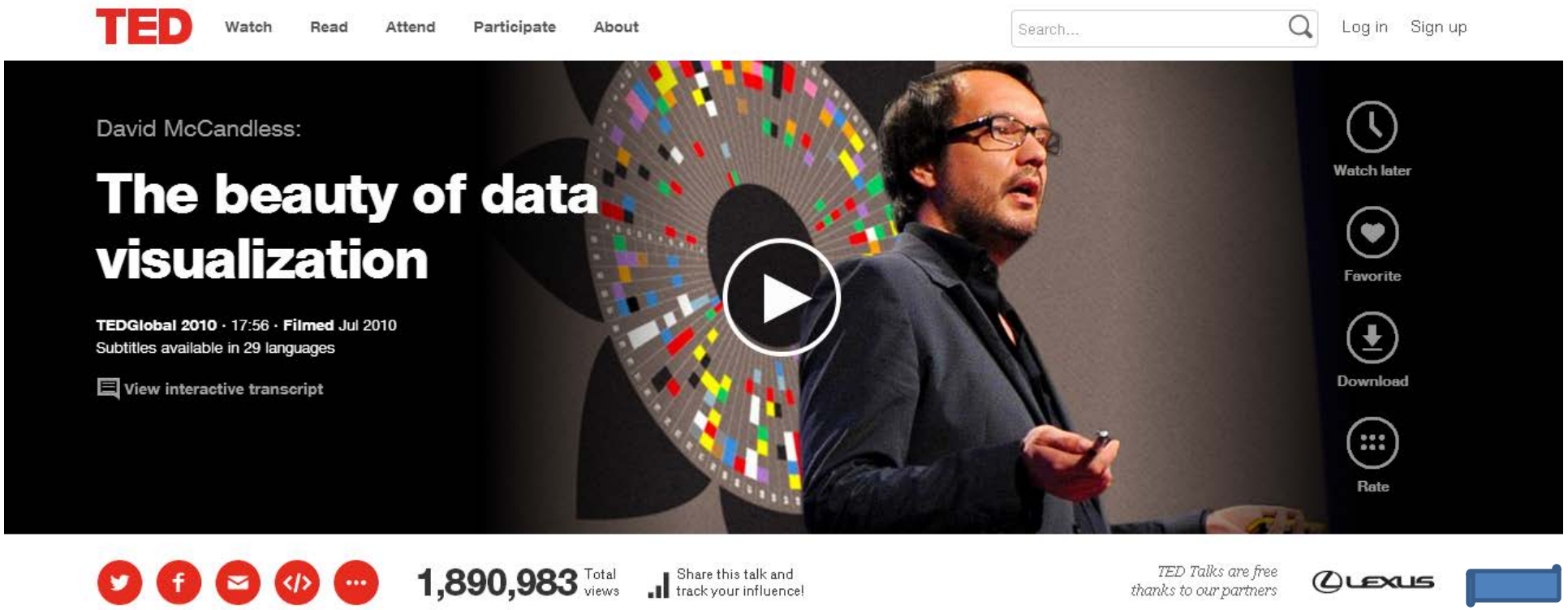
- 首先要懂数据，有“数据意识”
  - 学习和掌握统计学基本概念，
  - 能告诉码农或美工按什么主题或故事分析或展示数据
  - 能从别人的分析结果或可视化中看出问题
- 其次要学一点技术（编程或使用软件包）
  - 以便对数据做粗略的探索，寻找或检验自己初始的猜想
- 最后要略懂一点视觉艺术（构图、色彩等）
  - 以便对美工的炫丽作品有基本判断能力和与他们沟通的共同语言

David McCandless: 数据可视化之美

一位从码农到记者的可视化设计师经验之谈：

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization

http://v.youku.com/v_show/id_XNDYxMjUxMTk2.html



多谢聆听、欢迎交流

j.zhu@cityu.edu.hk
weblab.com.cityu.edu.hk
weibo.com/weblabcityu




Web Mining Lab
互聯網挖掘實驗室