北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

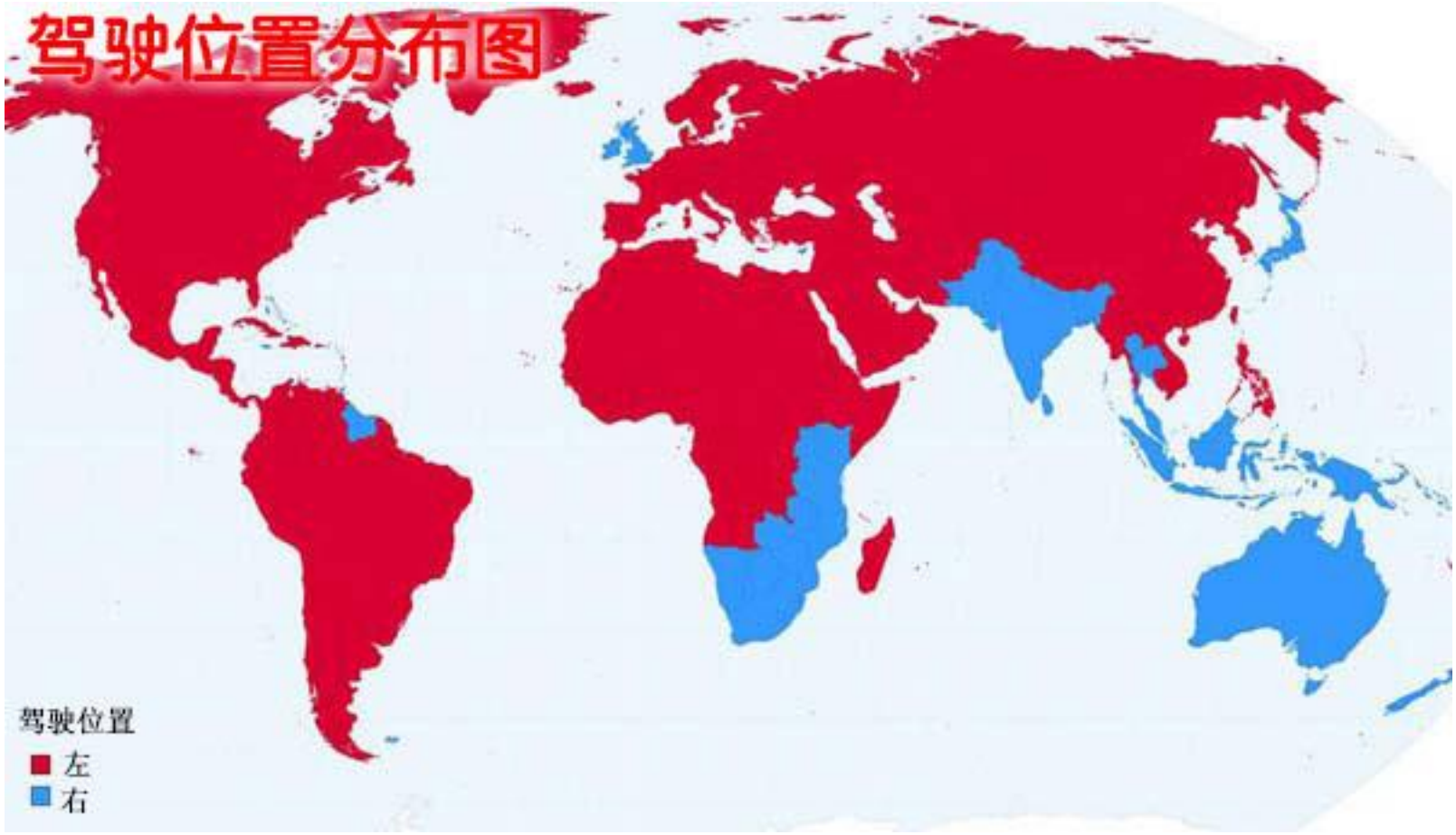
马宝君 博士 讲师

经济管理学院
电子商务中心

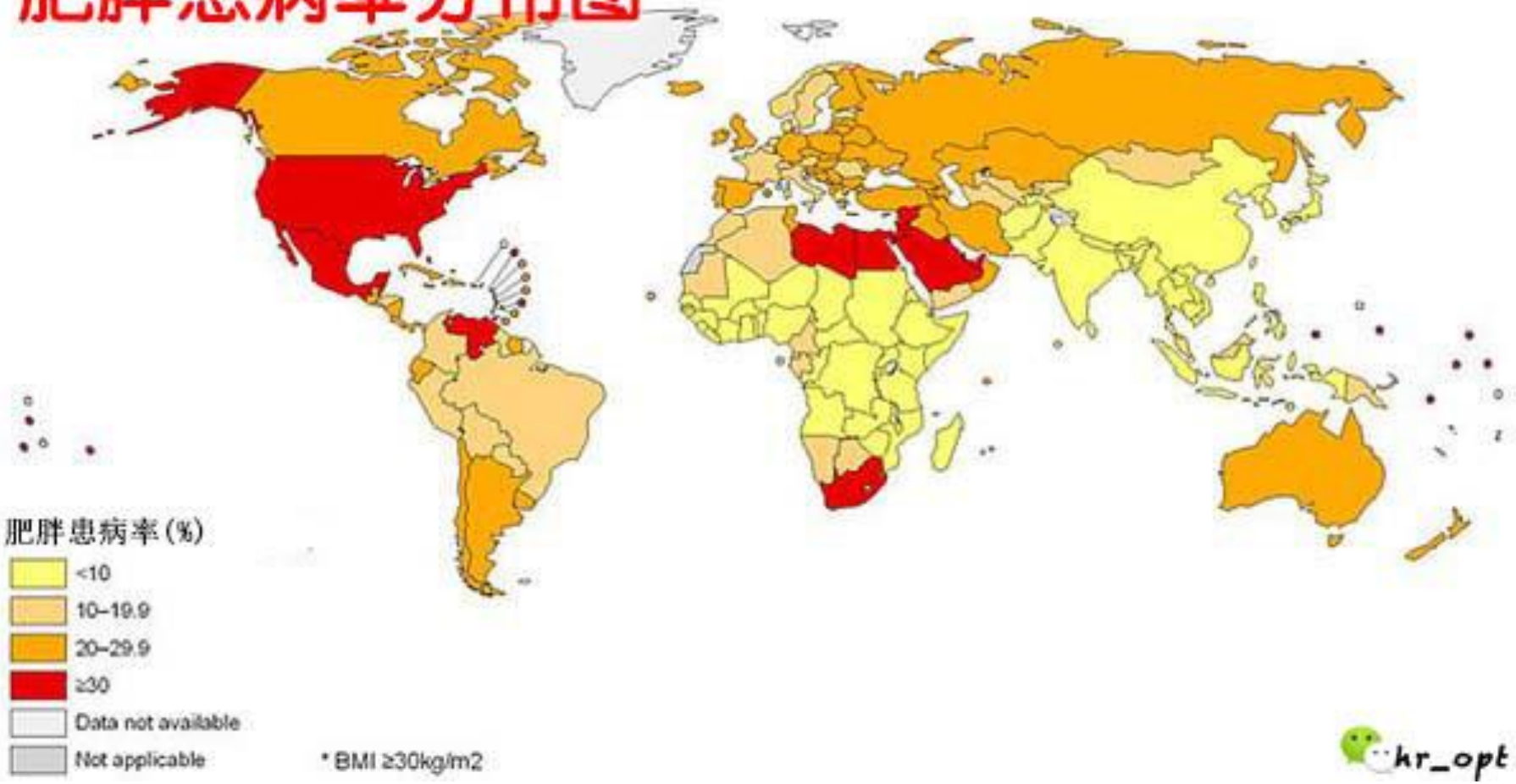
2014.10.13

分享：十张世界地图带你重新认识这个世界

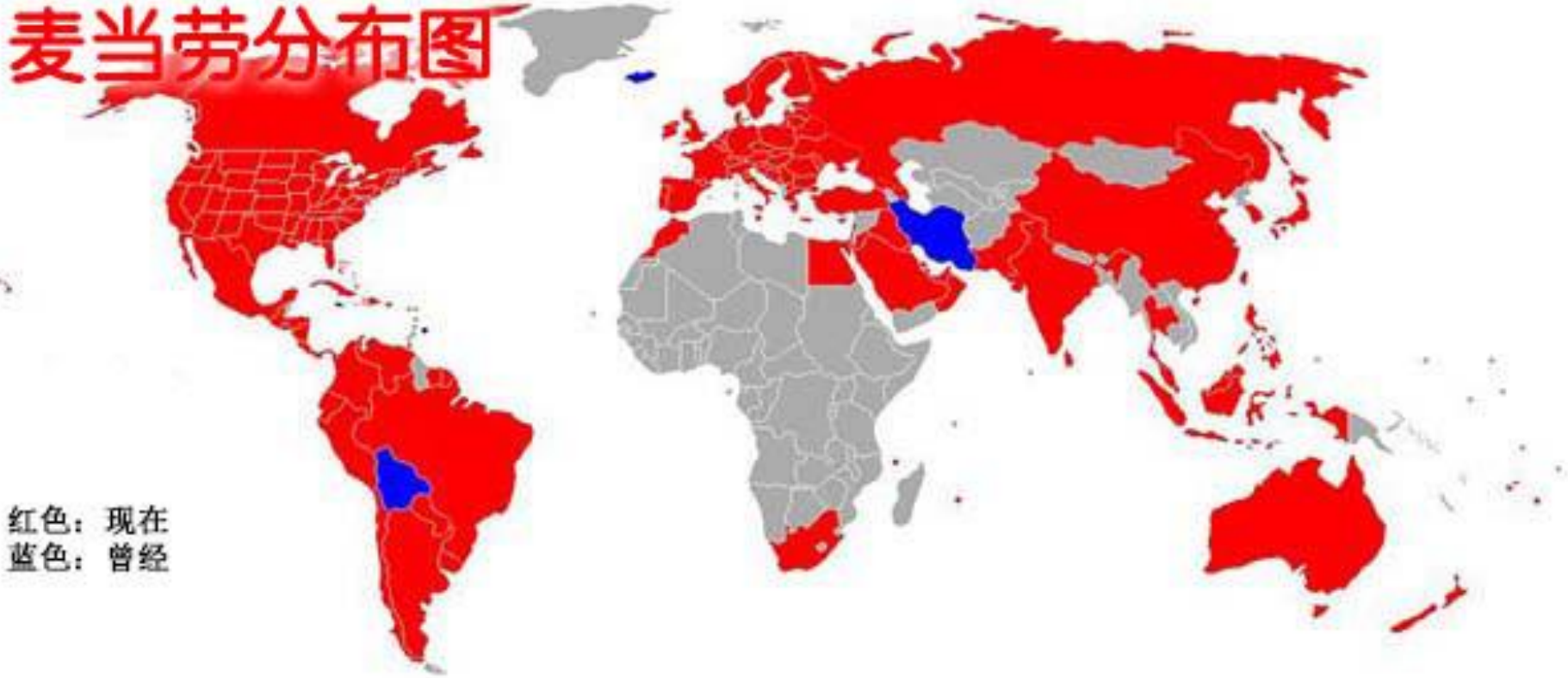
来源：http://www.199it.com/archives/280558.html


驾驶位置分布图
驾驶位置
左
右

肥胖患病率分布图

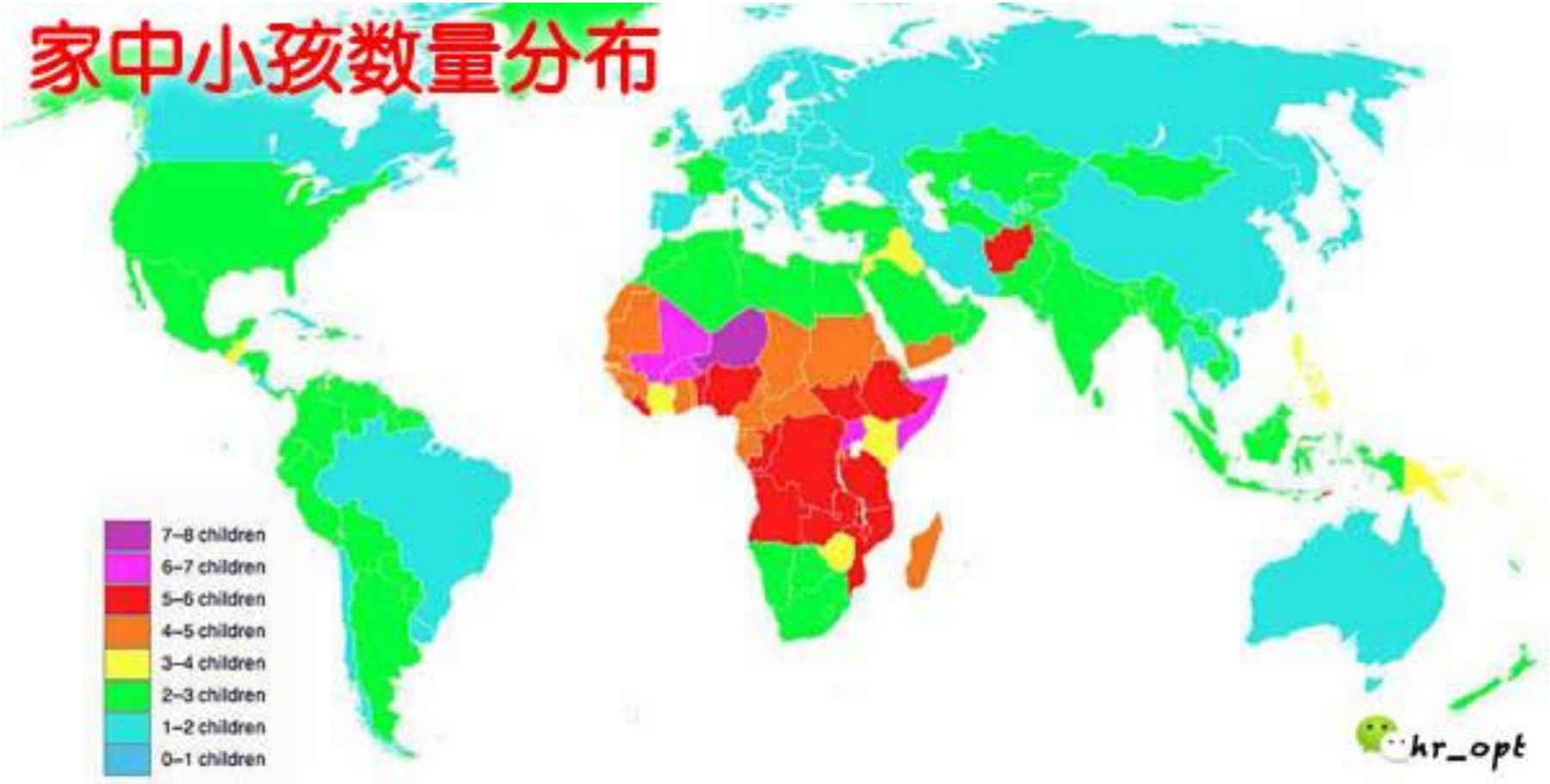

肥胖患病率(%)
<10
10–19.9
20–29.9
≥30
Data not available
Not applicable
* BMI ≥30kg/m2
hr_opt

麦当劳分布图




英格兰曾入侵过的国家
LIECHTENSTEIN
LUXEMBOURG
ANDORRA
MONACO
VATICAN CITY
BELARUS
MONGOLIA
UZBEKISTAN
KYRGYZSTAN
TAJIKISTAN
MALI
CHAD
GUATEMALA
IVORY COAST
CENTRAL AFRICAN REPUBLIC
BURUNDI
REPUBLIC OF CONGO
BOLIVIA
PARAGUAY
MARSHALL ISLANDS

家中小孩数量分布


7–8 children
6–7 children
5–6 children
4–5 children
3–4 children
2–3 children
1–2 children
0–1 children
hr_opt


最受欢迎的运动分布图
足球
爱尔兰式橄榄球
美式橄榄球
冰球
篮球
棒球
板球
英式橄榄球
乒乓球
自由搏击
摔跤
滑雪
前术
未知


IQ分布图
Greenland
Canada
United States
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Russia
Mongolia
China
Australia
India
Brazil
TARGET MAP
National IQ Scores
Average IQ score
99 - 108
91 - 99
83 - 91
74 - 83
67 - 74
59 - 67
hr_opt

未采用公制单位的国家

CUP SIZE 分布图


Greenland
Canada
United States
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
Russia
Kazakhstan
Mongolia
China
Brazil
Australia
New Zealand
TARGET MAP
Average Breast Cup Siz
Breat Cup Size
5
1
Breast Cup Size.xls
SUM of Value
Larger than D Cup
D Cup
C Cup
B Cup
A Cup

上次课程小结

- **大数据时代的特征**

- **大数据时代的思维变革**

大数据时代的思维变革

- 更多
  - 不是随机样本，而是**全体数据**
- 更杂
  - 不是精确性，而是**混杂性**
- 更好
  - 不是因果关系，而是**相关关系**

（1）不是随机样本，而是全体数据

- **利用所有的数据，而不再仅仅依靠一小部分数据**
- **统计学**
  - 用尽可能少的数据来证实尽可能重大的发现
- **小数据时代的随机采样，最少的数据获得最多的信息**
  - 人口普查(Census / Censere 推测，估算)的可行性、时效性
  - 选取最具有**代表性**的样本：**随机性**
  - 采样分析的精确性随着采样随机性的增加而大幅提高，但与样本数量的增加关系不大
  - **社会科学**过去的很多研究依赖于样本分析、调查问卷
  - 随机采样是在不可收集和分析全部数据的情况下的选择，**存在许多固有缺陷**

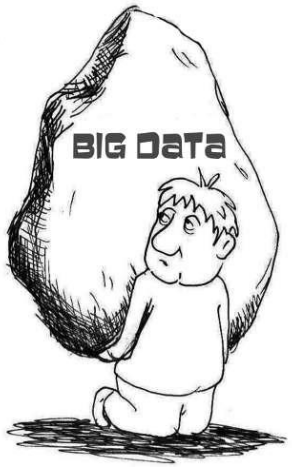
随机采样的固有缺陷

- **成功依赖于采样的绝对随机性：非常困难**
  - 以固定电话用户为基础的投票民调/春节联欢晚会满意度调查
- **无法得到一些微观细节的信息**
- **随机采样不适合考察子类别的情况**
  - 一旦继续细分，随机采样结果的错误率会大大增加

- **随机采样就像是模糊照片打印，远看很不错，但是一旦聚焦某个点，就会变得模糊不清**
- **随机采样只是一条捷径，但并不适用于一切情况，因为随机采样结果缺乏延展性，调查出的结果不可以重新分析以实现计划之外的目的**

全数据模式，样本 = 总体

- 足够的数据处理和分析能力
- 最先进的分析技术
- 简单廉价的数据收集方法
  - 信用卡诈骗识别：只有掌握了所有的数据才能做到



- 大数据中的“大”不是绝对意义上的大，而是相对意义
- 大数据是指不用随机分析法这样的捷径，而**采用所有相关数据**的方法

“样本 = 总体”模式的例子

- Albert-Laszlo Barabasi教授的研究团队
- 想研究人与人之间的互动
- 调查了4个月内美国某无线运营商的所有移动通信记录（覆盖近1/5全美人口）

- 新发现
  - 与小规模的研究相比，这个团队发现：如果把一个在社区内有很多连接关系的人从社区关系网中剔除掉，这个社会关系网会变得没那么高效但却不会解体；（强关系）
  - 但如果把一个与所在社区之外很多人有着连接关系的人从这个关系网中剔除，整个关系网很快就会破碎成很对小块；（弱关系）
  - 这说明：一般来说，无论针对一个小团体还是整个社会，多样性都是有额外价值的，这个发现促使我们重新审视一个人在社会关系网中的存在价值

（2）不是精确性，而是混杂性

- **对“小数据”而言，最基本、最重要的要求就是减少错误，保证质量和精确性，但物理学在1927年提出了量子力学的基本原理：“测不准原理”，永远粉碎了“测量臻于至善”的幻梦。**
- **在不断涌现的新情况里，允许不精确的出现已经成为一个新的亮点，而非缺点**
- **混乱（混杂性）**
  - 随着数据量的增加，错误率也会相应增大
  - 数据格式的不一致，是否需要仔细清洗数据？

- **牺牲精确性的好处**

适量错误

  - 例子：葡萄园温度测量
  - 更广泛的数据：可能会看到如若不然无法观察到的细节
  - 高频率数据：可能观察到一些本可能被错过的变化

大数据通常用“概率”说话

- “大数据”通常用概率说话，而不是板着“确凿无疑”的面孔
- 大数据的简单算法比小数据的复杂算法更有效！
- 例子：无所不包Google翻译系统（2006年开始）

- “Google的翻译系统不会像Candide一样只是仔细地翻译300万句话，它会掌握用不同语言翻译的质量参差不齐的数十亿页的文档”
- 尽管输入源很混乱，但教其他翻译系统而言，Google翻译的质量相对而言还是最好的，而且可翻译的内容更多
- 效果好，不是因为它拥有一个更好的算法机制，而是它增加了很多各种各样的数据

（3）不是因果关系，而是相关关系

- **亚马逊推荐系统**
  - 内部书评家的个人建议和评论 vs 机器生成的个性化推荐
  - 如今，据说亚马逊销售额的三分之一来自于个性化推荐系统
- **相关关系**
  - 核心：量化两个数据值之间的数理关系
  - 通过识别有用的关联物来帮助分析现象，而不是揭示内部运作机制
- **通过给我们找到一个现象的良好的关联物，相关关系可以帮助我们捕捉现在和预测未来**
  - 沃尔玛：季节性风暴来临时，蛋挞和飓风用品（例如：手电筒）销量同时增加
- **大数据的相关分析分析法更准确、更快，而且不容易受到偏见的影响**
- **建立在相关关系分析法基础上的预测是大数据的核心！**

“是什么”，而不是“为什么”

- 在小数据时代，由于计算机能力的不足，大部分相关关系分析仅限于寻求“线性关系”
- 事实上，随着数据量的增加，实际情况比我们想象的要复杂，应当可以发现数据的“非线性关系”
- 例子：“收入水平”和“幸福感”的非线性关系
  - 政策影响！

因果关系的悖论

paradox

- **因果关系真的存在吗？**
  - 在哲学界，对于因果关系是否存在的争论已经持续了几个世纪
  - 如果凡事皆有因果，那么我们就没有决定任何事的自由了
  - 与“自由意志”相对立
- **日常生活中，因果关系思维成为习惯**
  - 并非浅而易寻
  - 即使使用数学这种比较直接的方式，因果联系很难被轻易证明
  - 即使我们慢慢思考，想要发现因果关系也是很困难的
  - 例：狂犬疫苗救治
- **因果关系被完全证实的可能性几乎没有，只能说：某两者之间很有可能存在因果关系**

从“相关关系”到“因果关系”

- 相关关系分析本身意义重大，同时它为研究因果关系奠定了基础

- 因果关系只是一种特殊的相关关系
- 相关关系分析通常情况下能够取代因果关系起作用，即使不可取代的情况下，它也能指导因果关系起作用
- 例子：曼哈顿沙井盖（下水道检修口）的爆炸

小结：大数据时代的思维变革

- 更多
  - 不是随机样本，而是**全体数据**
- 更杂
  - 不是精确性，而是**混杂性**
- 更好
  - 不是因果关系，而是**相关关系**