大数据时代——读书笔记

一、引论

1. 大数据时代的三个转变：
    1. 可以分析更多的数据，处理和某个现象相关的所有数据，而不是随机采样
    2. 不热衷于精确度
    3. 不热衷与寻找因果关系
2. 习惯：用来决策的信息必须是少量而精确的。实际：数据量变大，数据处理速度变快，数据不在精确
3. 危险：不是隐私的泄露而是未来行动的预判

二、大数据时代的思维变革

1. 原因：没有意识到处理大规模数据的能力，假设信息匮乏，发展一些使用少量信息的技术（随机采样）
    1. 1086 年 末日审判书 英国对人的记载
    2. 约翰·格朗特：统计学，采样分析精确性随着采样随机性上升而大幅上升，与样本数量关系不大
    3. 1890 年，穿孔卡片制表机，人口普查
    4. 随机采样有固有的缺陷
        1. 采样过程中存在偏差
        2. 采样不适合考察子类别
        3. 只能得出实现设计好的问题的结果
        4. 忽视了细节考察
2. 全数据模式：样本=总体
    1. 通过异常量判断信用卡诈骗
    2. 大数据分析：不用随机抽样，而是采用所有数据。不是绝对意义而是相对意义。（Xroom 信用卡诈骗，日本相扑比赛）
    3. 多样性的价值（社区外联系很多》社区内联系很多）
3. 混杂性而非精确性
    1. 葡萄树温度测量：数据变多，虽然可能有错误数据，但总体而言会更加精确。
    2. 包容错误有更大好处
    3. word 语法检查：语料库》算法发展
    4. google 翻译：让计算机自己估算对应关系，寻找成千上万对译

    结论：大数据的简单算法好过小数据的复杂算法

    5. 大数据让我们不执著于也无法执着于精确
    6. MIT 的通货紧缩软件：即时的大数据
    7. 标签：不精确
    8. 想要获得大规模数据的好处，混乱是一种标准途经
    9. 新的数据库：大部分数据是非结构化的，无法被利用
    10. Hadoop：与 mapreduce 系统相对的开源式分布系统，输出结果不精确，但是非常快

    结论：相比于依赖小数据和精确性的时代，大数据因为更强调数据的完整性和混杂性，帮助我们进一步接近事情的真相。“部分”和”确切“的吸引力是可以理解的。但是当我们的视野局限在我们可以分析和确定的数据上时，我们对世界的整体影响就会产生偏差和错误。不仅失去了尽力收集一切数据和活力，也失去了从不同角度观察时间的权利。

三、不是因果是相关

1. 知道是什么就够了，不需要知道为什么。
    1. 亚马逊放弃书评组，使用大数据预测人们的未来购书需求
    2.
2. 在小数据世界，相关关系有用，但是大数据背景，相关关系大放异彩。通过找关联物，相关关系可以帮助我们捕捉现在和预测未来
    1. A 和 B 经常一起发生，那么 A 发生时可以预测 B 发生
    2. 例子：沃尔玛把飓风用具和蛋挞放在一起
    3. 过时的寻找关联物的方法
        a) 原因：数据少且收集花时间
        b) 在建立，应用假想和选择关联物时容易犯错误
        c) 结论：我们不需要人工选择关联物
3. 大数据的相关分析法更准确，更快
    1. 例子：FICO 我们知道你明天会做什么
    2. 伊百丽：根据个人信用卡交易记录预测个人收入，防止逃税
    3. Aviva：根据生活方式数据预测疾病
    4. 美国零售商 target：通过购买习惯预测是否怀孕
4. 通过找出新种类数据的相互联系解决日常需要：找到关联物并监控，我们可以预知未来
    1. 例子：UPS 与汽车修理预测
    2. 新生儿健康监测：肉眼看不到，但是计算机能看到
5. 当收集分析和储存数据的成本较高时，应当适当丢弃一些数据
6. 数据的非线性关系
    1. 幸福的非线性关系
7. 快速思维模式使人们偏向于用因果关系看待周围的一切，因此经常对世界产生错误认识。这也使大脑为了避免辛苦思考而产生的捷径。大数据会经常被用来证明我们习惯的思维方式是错误的。
8. 证明因果关系的实验开销大，难于操作；相关关系很有用，不仅是因为能为我们提供新的视角，而且提供的视角都很清晰。一旦我们考虑因果关系，这些视角会被蒙蔽。
9. 大数据并非是理论消亡的时代。

四、一切皆可量化

1. 莫里的信息交换计划：总结所有船只的航海日志已获得好的航线，为第一根大西洋电缆奠定基础
2. 坐姿研究与汽车防盗系统
3. 数据化
    1. 把现象转变成可指标分析的量化形式的过程
    2. 计量和记录促成了数据：
        1. 阿拉伯数字
        2. 计数板
        3. 复式记账法
    3. 数字化与数据化的区别
        1. 例子：google 的数字图书馆：开始使用扫描-》数字化，进而光学识别-》数据化。Google 借此改进自己的翻译
        2. 文化组学：定量分析揭示人类行为
    4. 文字变成数据：人可以阅读，机器可以分析

5. 方位变成数据：需要一套标准的标记系统和收集，记录数据的工具。
    1. 始于古希腊
    2. 1884 年，国际子午线会议
    3. 1978 年，全球定位系统
    4. 英国汽车保险
    5. UPS 的最佳行车路线：减少左转
    6. 收集用户地理位置数据，以便进行忠诚度计划。或者可以预测交通情况
6. 现实挖掘
    1. 处理大量手机数据，发现并预测人类的行为。
    2. 例子：预测流感隔离区域
    3. 例子：通过非洲预付费用户的位置信息和他们账户的资金，发现贫民窟是经济繁荣的跳板
7. 沟通变成数据
    1. FaceBook：社交关系数据化
    2. 推特：情绪数据化。对冲基金正在分析微博的文本，以作为股市投资的信号。新推特频率可以预测电影票房
    3. 例子：微博与疫苗：人们对于疫苗的态度与他们实际注射预防流感药物的可能性呈现正相关
8. 万物数据化
    1. 触觉地板：适时开关灯，确定身份，某人摔倒之后是否站起来
    2. 人体传感器：监控健康状态

4. 结论：世界的本质是信息和数据，大数据提供新视角。

五、大数据的潜在价值

1. 例子：captcha（验证码，全自动区分人类和电脑的图灵测试）与数据再利用。作者使用了新的验证码 recaptcha，人们从计算机光学字符识别程序无法识别的文本扫面项目中读入单词并输出，知道他们都输出正确后才确定（用来破译数字化文本中不清楚的单词）

2. 大数据时代，所有的数据都是有价值的。现在，我们能够以较低成本获取并存储数据。数据的真实价值就像漂浮在海洋中的冰山，绝大部分隐藏在表面之下。

3. 不同于物质性的东西，数据的价值不会随它的使用而减少，而且可不断被处理。意味着数据的最终价值远远大于它的最初价值。在基本用途完成后，数据的价值仍然存在，数据的价值是其所有可能用途的总和。

4. 例子：IBM 与电力汽车动力系统的优化预测：大数据预测模型，甚至考虑天气预报

5. 数据再利用：
    1. 搜索关键词，搜索结果预测夏天流行色
    2. google 保存语音翻译记录，开发自己的语音识别技术
    3. 移动运营商长期使用大数据微调网络性能
    4. 有些公司可能会收集到大量的数据，但是他们并不急需使用，也不擅长使用数据，但是别的公司可以借此探寻数据的潜在价值

8. 重组数据
    1. 例子：丹麦癌症协会与手机致癌调查：使用所有的手机用户信息和所有的中枢神经系统肿瘤信息。

    随着大数据的出现，数据的总和比部分更有价值，当我们将多个数据集的总和重组在一起，重组总和本身的价值也比单个总和更大

9. 可拓展数据
    1. Google 街景和 GPS 采集，不仅将其用于基本用途，而且进行了大量的二次利用。例如，对 Google 自动驾驶汽车的运作
10. 数据的折旧值
    1. 随着时间的推移，大多数数据都会失去一部分基础用途，不应用此破坏新数据
    2. 挑战：如何得知某些数据不再有价值
    3. 并非所有数据都会贬值。例子：Google 希望得到每年的同比数据

    结论：组织机构应收集尽可能多的使用数据并保存尽可能长的时间。同时也应该与第三方分享数据
11. 数据废气：用户在线交互的副产品，包括浏览哪些页面，停留多久，输入信息等
    1. 数据再利用的方式很隐蔽
    2. 例子：Google 的拼写检查：搜集每天处理的查询中数据搜索框的错误拼写
    3. 例子：Google 的过滤噪音技术：如果用户点击搜索结果靠后的链接，说明这个结果更加有相关性，Google 会把这个页面的排名相应提升。
    4. 当用户指出了各种自动化程序的错误，实际上是训练了系统
    5. 例子：巴诺与数据快照，电子书阅读器捕捉人们阅读书籍的习惯
    6. 例子：Coursera 通过捕捉学生犯的错误来提示未来犯错误者

    结论：数据废气可以成为公司的巨大竞争优势，和对手的强大进入堡垒
12. 开放数据
    1. 最大的数据收集者：政府，可以强迫人们提供信息，但是信息利用效率低下。最好允许私人运营部门和社会大众访问
    2. 例子：FlyOnTime 网站，通过开放的数据分析航班延误可能性。
    3. 给数据估值：从数据持有人在价值提取上所采取的不同策略入手，将数据授权给第三方

三、角色定位：数据，技术与思维

1. 例子：decide.com 广泛收集数据，用来发现不正常，不合理的价格高峰。
2. 思维转变的重要性
3. 三种大数据公司
    1. 基于数据本身的公司：twitter

    大数据最值钱的是他本身，所以应该优先考虑数据拥有者

        例子：机票预订系统 ITA 不直接使用数据：担心暴露利润
        例子：MasterCard 通过大数据预测客户的消费习惯
    2. 基于技能的公司：咨询公司，技术供应商或者分析公司：Teradata
        例子：埃森哲公司利用大数据检测汽车零件并节省费用
        例子：微软分析公司利用大数据降低病人的再入院率
    3. 基于思维的公司：创新思维
        例子：FlightCaster 飞机晚点预测
        例子：prismatic 分析新闻并排序
4. 大数据先驱者一般有跨学科的知识
5. 例子：google 和 amazon 三者兼备
6. 全新的数据中间商：从各个地方搜集数据，提取有用的信息进行利用，并不威胁数据拥有者的利益
    1. 社会需要定向广告

    例子：Inrix：分析各种汽车制造者的数据和用户的数据，提供卫星导航服务

汽车制造商们本身数据量不够，自身也没有技术利用大数据，也并不介意数据会被中间商利用。同时可以提供失业率等相关数据

例子：Quantcast：收集用户访问信息来测评用户年龄等，之后发定向广告

例子：HCCI 收集医疗保单，分析美国医疗费用上涨是否合理

结论：

1. 数据价值的转移：从技术到数据本身和大数据思维
2. 传统商业模式颠覆：交易数据而不是交易技术
3. 传统专家的光芒会被统计和数据学家取代，因为后者只关心数据
    1. 例子：谷歌翻译团队的工程师都不会说出翻译的语言
    2. 真正的专家不会消亡，但是主导地位会改变
    3. 专业技能只适用于小数据时代，因为那是需要依靠直觉和经验指导，但是遭遇海量数据时，可以通过数据挖掘得到更多
4. 数据和统计学知识将成为现代工厂的基础，人类的价值体现在交流上，以进行广泛而深刻的传播
    1. 例子：交互式游戏，会根据用户来改良，以数据为基础运作
    2. 例子：The-numbers.com 通过大数据来预测电影票房
5. 大数据决定企业核心竞争力
    1. 数据规模决定价值
    2. 例子：劳斯莱斯通过大数据监测引擎，预测可能出问题的引擎
    3. 例子：苹果进军手机
    4. 大数据为小公司带来了机遇：能享受非固有资产规模的好处，低成本传播创新结果，只需要创新思维
    5. 大数据拥有者会想办法增加数据存储量
    6. 消费者成为数据拥有者并与中间商交易
    7. 大数据对中等规模的公司帮助不大：既没有灵活性也没有规模效应
6. 大数据撼动国家竞争力：西方世界优势减少

四、大数据时代的管理

1. 大数据会带来很多危险，因为其核心思想是用规模剧增来改变现状。
2. 滥用大数据的力量会伤害人身安全
3. 大数据的二次利用颠覆了隐私保护法：无法征得个人同意
4. 如果所有人的信息在数据库里，有意识地避免就是此地无银三百两
5. 匿名化：交叉检验会检验出来
6. 大数据预测：罪责判定基于对个人未来行为的预测。大数据可能会否定人的自由意志
7. 数据有其局限性，数据的质量可能会很差，有误导性。
8. 卓越的才华并不依赖数据：Apple 乔布斯的才能

五、掌握大数据

1. 个人隐私保护：从个人许可到让数据使用者承担责任，因为将责任从民众转移到数据使用者很有意义因为数据使用者比其他人更明白他们想怎么样使用数据，也因为他们是最大利益获得者：监管机制可以决定不同种类的个人数据必须删除的时间
2. 信息模糊处理
3. 个人应该为他们的行动而非倾向负责
4. 打破大数据的黑盒子：大数据算法师：评估数据源，分析数据工具，解读运算结果
    1. 外部算法师：审计大数据的准确程度和有效性

2. 内部算法师：监督大数据的运转

5. 反数据垄断

六、结语

没有什么是上天注定的，因为我们总能就手中的信息制定出相应的对策。大数据的预测结果也并非铁定而只是一种可能性，也就是说，只要我们愿意，结果可以改写，我们可以判定出迎接未来的最佳方式，也无需理解宇宙的奥秘或者神的存在，因为大数据帮我们做好了。

更大的数据来源于人本身，大数据所不能预测的，正是人类的直觉，勇气，探索精神和独创性。使用大数据的时候，我们应该怀有谦卑之心，铭记人性之本