# big data《大数据时代》

# 精华观点和核心语句

不再追求精确度，不再追求因果关系，而是承认混杂性，探索相关关系。
如同工业革命要开放物质交易、流通一样，开放、流通的数据是时代趋势的要求。开放所带来的改变远远大于拥有权和隐私性保护所带来的问题。
要全体不要抽样，要效率不要绝对精确，要相关不要因果。
作者认为相关关系比因果关系重要，译者表示反对，认为放弃因果等于放弃人类的智力优势，是末日之始。导致相关关系比因果关系重要的原因在于，我们机器学习和以结果为导向的研究思路误导人类。
公共医疗：Google 通过分析 03 到 08 的流感相关搜索词条，将 45 中词条组合输入一个数学模型之后，得到的流感预测数据和官方统计数据有 97%吻合。09 年判断准确，及时预报流感。
商业： farecast 利用十万亿条飞机票价记录，预测飞机票价准确度高达 75%，利用 farecast 购买机票的旅客平均每张机票节省 50 美元。
不再需要一致性的数据库和僵化的层次结构，不再需要结构化查询语言 sql，最新的数据库为非关系型数据库 nosql。
美国股市每天成交量高达 70 亿股，其中三分之二都是由数学模型和算法之上的计算机程序自动完成的，这些程序利用海量数据来预测利益和降低风险。
数据爆炸式增长，绝大部分为数字信息，极少部分为模拟数据。数据每三年多翻一番。
数据规模的量变产生质变，就比如万有引力对生物体大小的关系，纳米技术对现实生活物质的性质有所改变一样，空气阻力和重量和形状关系一样。
大数据的核心在于预测，把数学算法运用到海量数据中来预测事情发生的可能性。
不再依赖于随机采样，不在热衷于追求精确度。并非完全放弃精确度，只是不再沉迷于此。不在热衷于寻找因果关系，而是寻找事物之间的相关性。
数据化意味着从一切事物中汲取数据，甚至包括我们以前认为和“信息”搭不上边的事情。比方说，一个人所在的位置、引擎的振动、桥梁的承重等等。
如同电影《点石成金》中，棒球球探们在统计学家面前相形见绌——直觉的判断被迫让位于精准的数据分析。

正文：
第一章：样本 =全体
统计学家证明，采样分析的精确性随着采样随机性的增加而大幅度提高，但与样本数量的增加关系不大。随机采样取得了巨大的成功，但是他的成功利亚与采样的绝对随机性，实现采样的随机性非常困难，一旦采样过程中存在任何偏见，分析结果就会相去甚远。
搜集的数据越来越多，分析和预测结果就会越来越准确，并发现一些细节和微乎其微的重要问题。
有些情况下，异常值才是重要的信息，大数据的处理方法就不会错过这个异常值。商务是即时的，因此数据分析也应该是即时的。
《魔鬼经济学》
大数据是指不用随机分析法这样的捷径，而是通过采用所有数据的方法。数据量不一定很大，但需要全部，包含了所有的信息。

Lytro 相机记录整个光场的信息，搜集了所有的数据，拍摄完之后再对焦，而且有“可循环利用性”。
《爆发》

第二章：混杂性。
只有 5%的数据是结构化的，可以适用于传统数据库，如果不接受混乱，剩下 95%的非结构化数据都无法被利用。
少量数据下运行最佳的算法，可能在大数据下可能会表现差强人意，在少量数据下表现差的算法，可能在大数据下惊呆小伙伴们。大数据的简单算法比小数据的复杂算法更有效，混杂是关键。
谷歌翻译之所以好，除了数据量庞大以外，还接受了有错误的数据，即来自互联网的废弃内容。
Hadoop 超大量数据下的分布式处理，假设系统瘫痪而建立数据副本，假定数据量巨大无法移动，人们必须在本地进行数据分析。它的输出结果不想关系型数据库那般精确，无法用于卫星发射、开具银行账户明细，但是运行却快很多。

第三章 不是因果关系，而是相关关系
通过数据推荐产品所增加的销售远远超过书评家的贡献。计算机可能不知道为什么喜欢海明威作品的客户会购买菲茨吉拉德的书，但是他只要通过算法统计分析，得知这个结果就可以了。
沃尔玛领导了零售链的革命，让供应商监控销售速率、数量、以及存货情况。这个数据库不仅包含了每一个顾客的购物清单以及消费额，还包括购物篮中的物品、具体购买时间，甚至购买当天的天气。
在大数据时代，通过建立在人的偏见上的关联物检测法已经不再可行，因为数据库太大而且需要考虑的领域太复杂。幸运的是，许多迫使我们选择假想分析法的限制条件也逐渐消失了。现在我们拥有如此多的数据，这么好的机器计算能力，因而不再需要人工选择一个关联物或者一小部分相似的数据来逐一分析了。大数据的相关关系分析法，取代了基于假想的易出错的方法。大数据的相关关系法更准确、更快，而且不易受偏见的影响。
塔基特公司在完全不合准妈妈对话的前提下预测一个女性会在什么时候怀孕。她们会光顾以前不会去的商店，渐渐对新的品牌建立忠诚。
ups 与汽车修理预测，车辆处故障后，造成延误和在装载的负担，消耗大量人力物力。通过检测汽车的每个部位，及时更换需要更换的零件，免除了可能会造成的困扰。同样的方法也可以用在人的身上，，检测病人的即时信息。

第四章 数据化 一切皆可量化
莫里整合美国海军的航海日志，绘制更安全和快速的航海图表，其他商船需要使用图表，必须（病毒式传染）按照要求撰写航海日志并提交给莫里。将海上的船只都变成一个个科学站和天文台。
数据化不是数字化，数字化只是把模拟数据变成 1 和 0 来表示。
gps 全球定位系统的地理定位能精确到米，实现了自古以来无数航海家、制图家和数学家的梦想。
airsage 每天通过处理上百万手机用户的 150 亿条位置信息，为超过 100 个美国城市提供实时交通信息。

facebook，twitter 等社交网络将我们的关系、经历和情感进行数据化。他们不仅提供我们寻找和维持朋友、同事关系的场所，也将我们日常生活中的无形元素提取出来，转化为可用作新用途的数据。华尔街的数学奇才们将数据传输到他们的算法模式当中，寻找能被有效利用的关系模式当中。社交网络分析之父写了一个程序，能通过监听新微薄的发布频率，预测一部电影的成败，比其他传统方法还要准确。
自我量化是一项由一群健身迷、医学疯子以及技术狂人发起的运动，通过测量身体每一个部位和每一件事来让生活更美好。

第五章 价值 取之不尽用之不竭
验证码输入时，一个用于证明对方是人类，另一个则是图书扫描时计算机无法识别的模糊单词，由网络上大量用户帮忙识别，节省了大量人力物力财力。
随着购物平台、设计平台、金融等的出现，我们的人脉关系、想法、喜好和日常生活模式也逐渐被加入到巨大的个人信息库中。
数据的价值不会随着它的使用而减少，而是可以不断被处理，个人的使用不会妨碍其他人的使用。
ibm 搜集汽车电量和路线、充电站插槽、天气等等信息，开发了复杂的预测模型，确定充电的最佳时间和地点，揭示充电站的最佳设置点。
google 推出语音识别服务，借助 nuance 的技术，但是自己储存语音识别记录，依靠此记录重新创建了一个新的语音识别系统。
搜集数据是必须确保数据具有再利用性、重组能力、可拓展能力。
有部分数据价值会随之时间推移失去价值，比如在亚马逊上购买一本书，数月后对这方面的书完全失去了兴趣，则这个数据就失去了价值。但并非所有的数据都会贬值，大数据下鼓励储存所有数据并试图挖掘其中的价值。
google 拥有世界上最完整的拼写检查器，涵盖世界上每一种语言，依据是每天处理的 30 亿查询中输入搜索框中的错误拼写。
"数据废气"——他是用户在线交互的副产品，包括浏览了那些页面、停留了多久、鼠标光标停留的位置、输入了什么信息等。比如 google 如果发现用户搜索之后再重复搜索，则表明搜索结果不满意，或者发现用户点击后面的选项，则算法自动将后面的选项调前。是搜索引擎的自我训练。
电子阅读器捕捉大量关于文学喜好和阅读人群的数据，贩卖给出版社。比如阅读一页或一节需要多长时间，读者是略读还是直接放弃阅读，是否划线强调还是在空白处做了标记，这些信息都是出版商和作者之前不会知道的信息。
在线课程跟踪学生的 web 交互来寻找最佳的教学方法，比如多次看一个课程，说明该课程没有讲清楚。
政府是最大规模信息的原始采集者。美国、欧盟等政府已经公开了很多信息，除了一些机密的信息。flyontime.us 航班时间预测，搜集交通运输局的历史航班延误数据、美国联邦航空管理局的机场信息，以及美国国家海洋和大气管理局的以往天气报告、国美气象服务的实时状态等。
给数据估值——facebook 更具会计准则计算出的价值为 63 亿美元，但市场估值却为 1040 亿美元，为什么差距这么大？公司账面价值和市场价值之间的差额被记为"无形资产"。二十世纪八十年代中期，无形资产在美国上市公司市值中约占 40%，而在 2002 年，这一数据已经增长为 75%。无形资产早期包括品牌、人才和战略这些应计入正规金融会计制度的非有形资产部分。但渐渐地，公司所持有和使用的数据也渐渐纳入了无

形资产的范畴。几乎肯定数据的价值将显示在企业的资产负载表上，成为一个新的资产类别。
催生了一大批倒卖数据的公司和机构，纷纷给数据定价，数据在不断被转手和利用，共同挖掘其中的价值。

第六章 角色定位
decide.com 收集电子商务网站上所有的电子产品的价格数据和产品信息，告知用户何时才是购买电子产品的最佳时机。预测准确率高达 77%。他和 farecast 都出自奥伦之手。
大数据价值链三大构成：基于数据本身的公司，基于技能的公司，基于思维的公司。
google 和亚马逊幸运地同时拥有这三个方面。
数据科学家是统计学家、软件程序员、图形设计师和作家的结合体，通过搜寻数据库来得到新的发现。
信用卡发行商搜集消费信息。
微软和医院合作，分析多年来的匿名医疗记录，发现出现压抑的病人再次入院的概率更高，因此出院以后的医学干预必须以解决病人的心理问题为重心，降低再入院率和医疗成本。
所谓大数据思维，是指一种意识，认为公开的数据一旦处理得当就能为千百万人急需解决的问题提供答案。
金矿产业链中，金子最珍贵，因此数据的价值胜过算法技术和大数据思维。
inrix 搜集全美和欧洲的汽车交通信息，并提供 app 给司机，供司机查询交通情况，同时司机自身的交通数据也上传分享了出来。他同时发现一些价值点，比如一个商场周围车辆很多，说明商场的销量增加。上下班高峰时期的交通状况变好了，这就说明失业率增加了，经济状况变差了。
行业专家和技术专家的光芒都会被统计学家和数据分析家的出现而变暗，因为后者不受旧观念的影响，能够聆听数据发出的声音。
人们把专业人才看的比全才更重要，深度才是财富。
苹果公司与运营商签订合约的时候规定，运营商提供给它大部分的有用数据。
普通消费者愿意免费提供这些数据来换取更好的服务，比如亚马逊的图书推荐、博客、twitter，维基百科等等。

第七章 风险
大数据时代，很多数据在搜集的时候并无意用作其他用途，而最终却产生了很多创新性的用途。
无处不在的信息泄露，侵犯了人们的隐私，一个可能的途径是匿名化，但是匿名化对大数据是无效的，因为搜集的数据越来越多，我们会结合越来越多不同来源的数据。
“蓝色粉碎”为警员提供情报，关于哪些地方更容易发生犯罪事件，什么时候更容易带到罪犯。帮助执法部门更好的分配资源，使犯罪发生率下降了 26%。
过分依赖数据，而数据远远没有我们所想的那么可靠。美国国防部长衡量越战成果用死亡人数，但只有 2%的美国将军认为死亡人数对战争成果是有意义的，美国很多部门一层一层将数字夸大化。
其实，卓越的才华并不依赖于数据。乔布斯依靠的是直觉，他的第六感，记者问他做了多少市场调研时，“没做！消费者没有义务去了解自己想要什么。”

第八章 掌控 责任与自由并举的信息管理

管理改革 1：个人隐私保护，从个人许可到让数据使用者承担责任。
新的隐私保护模式，着重于数据使用者为其行为承担责任，而不是将重心放在收集数据之处取得个人同意上。监管机制可以决定不同种类的个人数据必须删除的时间。再利用的时间框架则取决于数据内在风险和社会价值观的不同。公司可以利用数据的时间更长，但相应必须为其行为承担责任以及富有特定时间之后删除个人数据的义务。或者故意将数据模糊黑醋栗，促使大数据库的查询不能显示精确地结果，而只有相近的结果。
管理改革 2：个人动因 vs 预测分析。犯罪评定必须根据过去发生的事实评定，对未来的预测即使准确，但有失公平性和说服力。
管理改革 3：击碎黑盒子，大数据算法师的崛起。
管理改革 4：反数据垄断大亨。反垄断法遏制了权利的滥用，促进了大数据平台的良性竞争，世界上一些大型数据拥有者和政府都在逐步公布其数据。

结语 正在发生的未来
大数据为我们提供的不是最终答案，只是参考答案，帮助是暂时的，而更好的方法和答案还在不久的未来。
佛劳尔成为纽约市的"分析主人"，利用城市尚未开发的数据库开展分析和研究，提取价值。佛劳尔对经验丰富的统计学家没有兴趣，，他担心他们不愿意采取这种新方法来解决问题。"我想要可执行的洞察力"。挑选了五个毕业一两年的经济学专业学生组成团队。一起专注处理"非法改建问题"，将一套住房隔出很多个小房间，容纳多十倍的人，带来巨大的火灾隐患。纽约市每年会收到 25000 起非法改建的投诉，但只有 200 名检察院在处理这些事情。没有好办法区分简单的滋扰问题和严重的爆炸起火事件。——佛劳尔用大数据来解决。
输入来自 19 个机构的数据集，房产税、公用设施使用异常、建筑类型、修建时间、救护车访问次数、犯罪率和啮齿动物投诉等信息，数据形式都不可用，不一致，很凌乱，但他们整合以后，忽略精确度，将巨大的混杂数据库与火灾数据严重性排名进行对比并得到一个模型，预测投诉迫切度。
现场考察发现新的数据集线索，比如装修、砖工等，而且让检查员来测试他们的模型。最终让检查准确度提高了五倍。"我对因果关系不感兴趣，除非他用行动说话。"
大数据提示我们接受类似的不准确，因为不准确正是我们之所以为人的特征之一，就像我们处理混乱数据一样，毕竟混乱构成了世界的本质，也构成了人脑的本质，学会接受和运用他们才会得益。
**西方谚语——预测未来最好的办法就是创造未来。**