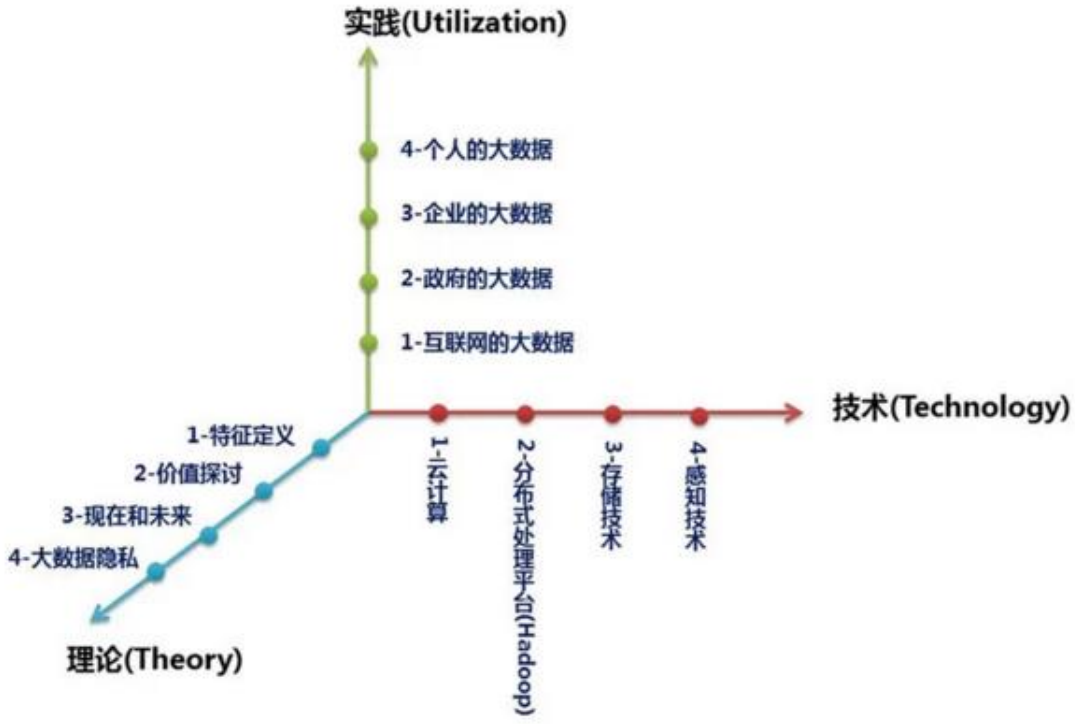
什么是大数据

目前人人都在谈大数据，但是，大数据是什么，每个人都有自己的一个看法。大数据这个概念，其实在上世纪九十年代就有人提出来了，当时希望通过将所有零散的数据归并起来，然后进行数据挖掘，以看到以前存在的问题，去预测未来几年的趋势，来指导商业决策。比如保险行业，人寿保险会通过大数据的统计计算，根据人均寿命来计算保费与回报率。这是在特定的环境，特定的时间下，对数据做一些商业化的尝试，还算不上真正的大数据。

想要系统认知大数据，必须要全面而细致的分解它，如果你听别人说大数据就是数据大，或者侃侃而谈 4 个 V，也许很有深度的谈到 BI 或预测的价值，又或者拿 Google 和 Amazon 举例，技术流可能会聊起 Hadoop 和 Cloud Computing，不管对错，只是无法勾勒对大数据的整体认识，不说是片面，但至少有些管窥蠡测、隔衣瘙痒了。下面从三个层面来展开认识：



第一层面是**理论**，理论是认知的必经途径，也是被广泛认同和传播的基线。将会从大数据的特征定义理解行业对大数据的整体描绘和定性；从对大数据价值的探讨来深入解析大数据的珍贵所在；从对大数据的现在和未来去洞悉大数据的发展趋势；从大数据隐私这个特别而重要的视角审视人和数据之间的长久博弈。

第二层面是**技术**，技术是大数据价值体现的手段和前进的基石。将分别从云计算、分布式处理技术、存储技术和感知技术的发展来说明大数据从采集、处理、存储到形成结果的整个过程。

第三层面是**实践**，实践是大数据的最终价值体现。将分别从互联网的大数据，政府的大数据，企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图。

1. 大数据相关的理论

1.1. 特征定义

业界(IBM 最早定义)将大数据的特征归纳为 4 个“V”(量 Volume，多样 Variety，价值 Value，速 Velocity)：第一，数据体量巨大。大数据的起始计量单位至少是 P(1000 个 T)、E(100 万个 T)或 Z(10 亿个 T);第二，数据类型繁多。比如，网络日志、视频、图片、地理位置信息等等。第三，价值密度低，商业价值高。第四，处理速度快。最后这一点也是和传统的数据挖掘技术有着本质的不同。

1.2. 价值探讨

大数据是什么?投资者眼里是金光闪闪的两个字：资产。比如，Facebook 上市时，评估机构评定的有效资产中大部分都是其社交网站上的数据。如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

Target 超市以 20 多种怀孕期间孕妇可能会购买的商品为基础，将所有用户的购买记录作为数据来源，通过构建模型分析购买者的行为相关性，能准确的推断出孕妇的具体临盆时间，这样 Target 的销售部门就可以有针对的在每个怀孕顾客的不同阶段寄送相应的产品优惠卷。

Target 的例子是一个很典型的案例，这样印证了维克托·迈尔-舍恩伯格提过的一个很有指导意义的观点：通过找出一个关联物并监控它，就可以预测未来。Target 通过监测购买者购买商品的时间和品种来准确预测顾客的孕期，这就是

对数据的二次利用的典型案例。如果，我们通过采集驾驶员手机的 GPS 数据，就可以分析出当前哪些道路正在堵车，并可以及时发布道路交通提醒；通过采集汽车的 GPS 位置数据，就可以分析城市的哪些区域停车较多，这也代表该区域有着较为活跃的人群，这些分析数据适合卖给广告投放商。

从大数据的价值链条来分析，存在三种模式：

- 手握大数据，但是没有利用好;典型的是金融机构，电信行业，政府机构等。
- 没有数据，但是知道如何帮助有数据的人利用它;比较典型的是 IT 咨询和服务企业，比如，埃森哲，IBM，Oracle 等。
- 既有数据，又有大数据思维;比较典型的是 Google，Amazon 等。

未来在大数据领域最具有价值的两种事物：

- 拥有大数据思维的人，这种人可以将大数据的潜在价值转化为实际利益;
- 还未有被大数据触及过的业务领域。这些是还未被挖掘的油井，金矿，是所谓的蓝海。

1.3. 现在和未来

先看看大数据在当下有怎样的杰出表现：

- 大数据帮助政府实现市场经济调控、公共卫生安全防范、灾难预警、社会舆论监督；
- 大数据帮助城市预防犯罪，实现智慧交通，提升紧急应急能力；
- 大数据帮助医疗机构建立患者的疾病风险跟踪机制，帮助医药企业提升药品的临床使用效果，帮助艾滋病研究机构为患者提供定制的药物；
- 大数据帮助航空公司节省运营成本，帮助电信企业实现售后服务质量提升，帮助保险企业识别欺诈骗保行为，帮助快递公司监测分析运输车辆的故障险情以提前预警维修，帮助电力公司有效识别预警即将发生故障的设备；
- 大数据帮助电商公司向用户推荐商品和服务，帮助旅游网站为旅游者提供心仪的旅游路线，帮助二手市场的买卖双方找到最合适的交易目标，帮助用户找到最合适的商品购买时期、商家和最优惠价格；
- 大数据帮助企业提升营销的针对性，降低物流和库存的成本，减少投资的风险，以及帮助企业提升广告投放精准度；
- 大数据帮助娱乐行业预测歌手，歌曲，电影，电视剧的受欢迎程度，并为投资者分析评估拍一部电影需要投入多少钱才最合适，否则就有可能收不回成本；
- 大数据帮助社交网站提供更准确的好友推荐，为用户提供更精准的企业招聘信息，向用户推荐可能喜欢的游戏以及适合购买的商品。

其实，这些还远远不够，未来大数据的身影应该无处不在。比如，Amazon 的最终期望是：“最成功的书籍推荐应该只有一本书，就是用户要买的下一本书。”Google 也希望当用户在搜索时，最好的体验是搜索结果只包含用户所需要的内容，而这并不需要用户给予 Google 太多的提示。

未来的大数据除了将更好的解决社会问题，商业营销问题，科学技术问题，还有一个可预见的趋势是以人为本的大数据方针。人才是地球的主宰，大部分的数据都与人类有关，要通过大数据解决人的问题。

比如，建立个人的数据中心，将每个人的日常生活习惯，身体体征，社会网络，知识能力，爱好性情，疾病嗜好，情绪波动……换言之就是记录人从出生那一刻起的每一分每一秒，将除了思维外的一切都储存下来，这些数据可以被充分的利用：

- 医疗机构将实时的监测用户的身体健康状况；
- 教育机构更有针对的制定用户喜欢的教育培训计划；
- 服务行业为用户提供即时健康的符合用户生活习惯的食物和其它服务；
- 社交网络能为你提供合适的交友对象，并为志同道合的人群组织各种聚会活动；
- 政府能在用户的心理健康出现问题时有效的干预，防范自杀，刑事案件的发生；
- 金融机构能帮助用户进行有效的理财管理，为用户的资金提供更有效的使用建议和规划；
- 道路交通、汽车租赁及运输行业可以为用户提供更合适的出行线路和路途服务安排；

1.4. 大数据隐私

用户隐私问题一直是大数据应用难以绕开的一个问题。在大数据的背景下，很多人都在积极的抵制无底线的数字化，这种大数据和个体之间的博弈还会一直继续下去……

当下，很多人依然没有建立对于信息隐私的保护意识，让自己一直处于被滋扰，被精心设计，被利用，被监视的处境中。

2． 大数据相关的技术

2.1． 云技术

大数据常和云计算联系到一起，因为实时的大型数据集分析需要分布式处理框架来向数十、数百或甚至数万的电脑分配工作。可以说，云计算充当了工业革命时期的发动机的角色，而大数据则是电。

云计算思想的起源是麦卡锡在上世纪 60 年代提出的：把计算能力作为一种像水和电一样的公用事业提供给用户。

如今，在 Google、Amazon、Facebook 等一批互联网企业引领下，一种行之有效的模式出现了：云计算提供基础架构平台，大数据应用运行在这个平台上。

那么大数据到底需要哪些云计算技术呢？

这里暂且列举一些，比如虚拟化技术，分布式处理技术，海量数据的存储和管理技术，NoSQL、实时流数据处理、智能分析技术（类似模式识别以及自然语言理解）等。

2.2． 分布式处理技术

分布式处理系统可以将不同地点的或具有不同功能的或拥有不同数据的多

台计算机用通信网络连接起来，在控制系统的统一管理控制下，协调地完成信息处理任务—这就是分布式处理系统的定义。

举个实际的例子，虽然这个例子有些陈旧，但是淘宝的海量数据技术架构还是有助于我们理解对于大数据的运作处理机制:

淘宝的海量数据产品技术架构分为五个层次，从上至下来看它们分别是：数据源，计算层，存储层，查询层和产品层。

数据来源层。存放着淘宝各店的交易数据。在数据源层产生的数据，通过 DataX，DbSync 和 Timetunel 准实时的传输到下面第 2 点所述的“云梯”。

计算层。在这个计算层内，淘宝采用的是 Hadoop 集群，这个集群，我们暂且称之为云梯，是计算层的主要组成部分。在云梯上，系统每天会对数据产品进行不同的 MapReduce 计算。

存储层。在这一层，淘宝采用了两个东西，一个是基于 MySQL 的分布式关系型数据库的集群 MyFox，Prom 是基于 Hadoop Hbase 技术的一个 NoSQL 的存储集群。

查询层。在这一层中，Glider 是以 HTTP 协议对外提供 restful 方式的接口。数据产品通过一个唯一的 URL 来获取到它想要的数据。同时，数据查询即是通过 MyFox 来查询的。

最后一层是产品层，这个就不用解释了。

2.3. 存储技术

大数据可以抽象的分为大数据存储和大数据分析，这两者的关系是：大数据

存储的目的是支撑大数据分析。到目前为止，这是两种截然不同的计算机技术领域：大数据存储致力于研发可以扩展至 PB 甚至 EB 级别的数据存储平台;大数据分析关注在最短时间内处理大量不同类型的数据集。

2.4. 感知技术

大数据的采集和感知技术的发展是紧密联系的。以传感器技术，指纹识别技术，RFID 技术，坐标定位技术等为基础的感知能力提升同样是物联网发展的基石。全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，都会产生海量的数据信息。

而随着智能手机的普及，感知技术可谓迎来了发展的高峰期，除了地理位置信息被广泛的应用外，一些新的感知手段也开始登上舞台，比如，“iPhone”在 home 键内嵌指纹传感器，新型手机可通过呼气直接检测燃烧脂肪量，用于手机的嗅觉传感器面世可以监测从空气污染到危险的化学药品，微软正在研发可感知用户当前心情智能手机技术，谷歌眼镜 InSight 新技术可通过衣着进行人物识别。这些感知被逐渐捕获的过程就是世界被数据化的过程，一旦世界被完全数据化了，那么世界的本质也就是信息了。就像一句名言所说，“人类以前延续的是文明，现在传承的是信息。”

3. 大数据相关的实践

3.1. 互联网的大数据

互联网上的数据每年增长 50%，每两年便将翻一番，而目前世界上 90%以上的数据是最近几年才产生的。据 IDC 预测，到 2020 年全球将总共拥有 35ZB 的数据量。互联网是大数据发展的前哨阵地，随着 WEB2.0 时代的发展，人们似乎都习惯了将自己的生活通过网络进行数据化，方便分享以及记录并回忆。互联网上的大数据很难清晰的界定分类界限，我们先看看 BAT 的大数据：

百度拥有两种类型的大数据：用户搜索表征的需求数据;爬虫和阿拉丁获

取的公共 web 数据。搜索巨头百度围绕数据而生。它对网页数据的爬取、网页内容的组织和解析，通过语义分析对搜索需求的精准理解进而从海量数据中找准结果，以及精准的搜索引擎关键字广告，实质上就是一个数据的获取、组织、分析和挖掘的过程。

阿里巴巴拥有交易数据和信用数据。这两种数据更容易变现，挖掘出商业价值。除此之外阿里巴巴还通过投资等方式掌握了部分社交数据、移动数据。如微博和高德。

腾讯拥有用户关系数据和基于此产生的社交数据。这些数据可以分析人们的生活和行为,从里面挖掘出政治、社会、文化、商业、健康等领域的信息，甚至预测未来。

简要归纳一下，在互联网大数据的典型代表性包括：

- 用户行为数据(精准广告投放、内容推荐、行为习惯和喜好分析、产品优化等)
- 用户消费数据(精准营销、信用记录分析、活动促销、理财等)
- 用户地理位置数据(O2O 推广，商家推荐，交友推荐等)
- 互联网金融数据(P2P，小额贷款，支付，信用，供应链金融等)
- 用户社交等 UGC 数据(趋势分析、流行元素分析、受欢迎程度分析、舆论监控分析、社会问题分析等)

3.2. 政府的大数据

国内来说，政府各个部门都握有构成社会基础的原始数据，比如，气象数据，金融数据，信用数据，电力数据，煤气数据，自来水数据，道路交通数据，客运数据，安全刑事案件数据，住房数据，海关数据，出入境数据，旅游数据，医疗数据，教育数据，环保数据等等。这些数据在每个政府部门里面看起来是单一的，静态的。但是，如果政府可以将这些数据关联起来，并对这些数据进行有效的关联分析和统一管理，这些数据必定将获得新生，其价值是无法估量的。

具体来说，现在城市都在走向智能和智慧，比如，智能电网、智慧交通、智慧医疗、智慧环保、智慧城市，这些都依托于大数据，可以说大数据是智

慧的核心能源。

3.3. 企业的大数据

企业的 CXO 们最关注的还是报表曲线的背后能有怎样的信息，他该做怎样的决策，其实这一切都需要通过数据来传递和支撑。

哪些传统企业最需要大数据服务呢?举几个例子：

1) 对大量消费者提供产品或服务的企业(精准营销);

2) 做小而美模式的中长尾企业(服务转型);

3) 面临互联网压力之下必须转型的传统企业(生死存亡)。

对于企业的大数据，还有一种预测：随着数据逐渐成为企业的一种资产，数据产业会向传统企业的供应链模式发展，最终形成“数据供应链”。

这里有两个明显的现象：

1) 外部数据的重要性日益超过内部数据。在互联互通的互联网时代，单一企业的内部数据与整个互联网数据比较起来只是沧海一粟;

2) 能提供包括数据供应、数据整合与加工、数据应用等多环节服务的公司会有明显的综合竞争优势。

3.4. 个人的大数据

个人的大数据概念很少有人提及，简单来说，就是与个人相关联的各种有价值数据信息被有效采集后，可由本人授权提供第三方进行处理和使用，并获得第三方提供的数据服务。

未来，每个用户可以在互联网上注册个人的数据中心，以存储个人的大数据信息。用户可确定哪些个人数据可被采集，并通过可穿戴设备或植入芯片等感知技术来采集捕获个人的大数据。

---自：36 大数据

时代的变革

古语云：三分技术，七分数据，得数据者得天下。先不论谁说的，但是这句话的正确性已经不用去论证了。维克托·迈尔-舍恩伯格在《大数据时代》一书中举了百般例证，都是为了说明一个道理：在大数据时代已经到来的时候要用大数据思维去发掘大数据的潜在价值。书中，作者提及最多的是 Google 如何利用人们的搜索记录挖掘数据二次利用价值，比如预测某地流感爆发的趋势；Amazon 如何利用用户的购买和浏览历史数据进行有针对性的书籍购买推荐，以此有效提升销售量；Farecast 如何利用过去十年所有的航线机票价格打折数据，来预测用户购买机票的时机是否合适。

4. 思维变革

4.1. 需要全部数据样本而不是抽样

当数据处理技术已经发生了翻天覆地的变化时，在大数据时代进行抽样分析就像在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，“样本=总体”。

我们要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本。小数据时代的随机采样，源自记录、存储、和分析数据的工具不够好，用最

少的数据获得最多的信息。然而随机采样毕竟有随机性，而且会丧失一些微观细节的信息，甚至还会失去对某些特定子类别进行进一步研究的能力。而现在，因为有了大数据存储，处理的能力，我们开始关注整体数据中价值。“大”是相对意义的大，也就是相对所有数据来说的。拥有全部或者几乎全部的数据，我们就能够从不同的角度，更细致地观察研究数据的方方面面。

4.2. 关注效率而不是精确度

数据量的大幅增加会造成结果的不准确，与此同时，一些错误的数据也会混进数据库。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。因为收集的信息量比较少，所以我们必须确保记录下来的数据尽量精确。因为收集信息的有限意味着细微的错误会被放大，甚至有可能影响整个结果的准确性。“大数据”时代，我们需要与各种各样的混乱做斗争。混乱，简单地说就是随着数据的增加，错误率也会相应增加。混乱还可以指格式的不一致性，因为要达到格式一致，就需要在进行数据处理之前仔细地清洗数据，而这在大数据背景下很难做到。

“大数据”通常用概率说话，而不是板着“确凿无疑”的面孔。整个社会要习惯这种思维需要很长的时间。其中也会出现一些问题。但现在，有必要指出的是，当我们试图扩大数据规模的时候，要学会拥抱混乱。

大数据时代要求我们重新审视精确性的优劣。大数据不仅让我们不再期待精确性，也让我们无法实现精确性。接受数据的不精确和不完美，我们反而能够更好地进行预测，也能够更好地理解这个世界。值得一提的是，错误并不是大数据固有的特性.而是一个亟需我们去处理的现实问题，并且有可能长期存在。

4.3. 关注相关性而不是因果关系

相关关系的核心是量化两个数据值之间的数理关系。相关关系强是指当一个数据值增加时，其他数据值很有可能会随之增加。比如谷歌流感趋势:在一个特定的地理位置，越多的人通过谷歌搜索特定的词条，该地区就有更多的人患了流感。相反，相关关系弱就意味着当一个数据值增加时，其他数据

值几乎不会发生变化。例如，我们可以寻找关于个人的鞋码和幸福的相关关系，但会发现它们几乎扯不上什么关系。

在小数据世界中，相关关系也是有用的，但在大数据的背景下，相关关系大放异彩。通过应用相关关系，我们可以比以前更容易、更快捷、更清楚地分析事物。**关联物，预测的关键。**

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。如果 A 和 B 经常一起发生，我们只需要注意到 B 发生了，就可以预测 A 也发生了。这有助于我们预测 A 可能会发生什么，即使我们不能直接测量或观察到 A。更重要的是，它还可以帮助我们预测未来可能发生什么。当然，相关关系是无法预知未来的，他们只能预测可能发生的事情。

建立在相关关系分析法基础上的预测是大数据的核心。它告诉你的是会发生什么，而不是为什么发生。事实上，就是因为不受限于传统的思维模式和特定领域里隐含的固有偏见，大数据才能为我们提供如此多新的视野。

5. 商业变革

5.1. 数据化：一切皆可“量化”

大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。数据，会从最不可能的地方提取出来，很多从不被认为是数据、甚至不被认为和数据沾边的事物转化成了可以用数值来量化的数据模式。比如日本的越水重臣研究的坐姿和汽车防盗系统，很少有人会认为一个人的坐姿能表现什么信息，但是它真的可以。当一个人坐着的时候，他的身形、姿势和重量分布都可以量化和数据化。这样根据人体对座位的压力差异识别乘坐者的身份。

新工具和开放的思维促进了测量事物和记录数据的繁荣。计算机的出现带来了数字测量和存储设备，这样就大大提高了数据化的效率。计算机也使得通过数学分析挖掘出数据更大的价值变成了可能。有了大数据的帮助，我们不会再将世界看做是一连串我们认为或是自然或是社会现象的事件，我们会意识到本质上世界是由信息构成的。将世界看作信息，看作可以理解的数据的海洋，为我们提供了一个从未有过的审视现实的视角。它是一种可以渗透到所有生活领域的世界观。

5.2. 价值：“取之不尽用之不竭”的数据创新

在数字化时代，数据支持交易的作用被掩盖，数据只是被交易的对象。而在大数据时代，事情再次发生变化。数据的价值从它最基本的用途转变为未来的潜在用途。这一转变意义重大，它影响了企业评估其拥有的数据及访问者的方式，促使甚至是迫使公司改变他们的商业模式，同时也改变了组织看待和使用数据的方式。

数据的价值并不仅限于特定的用途，它可以为了同一目的而被多次使用，也可以用于其他目的。要了解大数据时代究竟有多少信息对我们有价值，后面这一点尤其重要。过去，一旦数据的基本用途实现了，我们便认为数据已经达到了它的目的，准备将其删除让它就此消失。毕竟，数据的首要价值已经得以提取。而在大数据时代，数据是就像一个神奇的钻石矿，在其首要价值被发掘之后仍能不断给予。数据的价值是其所有可能用途的总和。

数据价值的关键是看似无限的再利用，即它的选择价值。收集信息固然至关重要，但还远远不够，因为大部分的数据价值在于它的使用，而不是占有本身。

5.3. 角色定位：数据、技术与思维的三足鼎立

如今，我们正处在大数据时代的早期，思维和技术是最有价值的，但是最终大部分的价值还是必须从数据本身来挖掘。来自于电子商务网站和互联网的公开数据很多，每个人都可以利用。技术上，谁也并没有无可替代的技术人才。虽然数据和技术是不可或缺的，但是真正使得某家公司取得成功的是拥有大数据的思维观念。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。

6. 管理变革

6.1. 风险：让数据主宰一切的隐忧

进行大数据分析的人可以轻松地看到大数据的价值潜力，这极大地刺激着他们进一步采集、存储、循环利用我们个人数据的野心。随着存储成本继续暴跌而分析工具越来越先进，采集和存储数据的数量和规模将突飞猛进地增长。大数据时代正在加深我们隐私的威胁。毕竟，大数据的核心思想就是用规模剧增来改变现状。

大数据的价值不再单纯来源于它的基本用途，而更多源于它的二次利用。更重要的是，大数据时代，很多数据在收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。所以，公司无法告知个人尚未想到的用途，而个人亦无法同意这种尚是未知的用途。但是只要没有得到许可，任何包含个人信息的大数据分析都需要向个人征得同意。但是这又是何其之难的事啊？

在大数据时代，不管是告知与许可、模糊化还是匿名化，这三大隐私保护策略都失效了。如今很多用户都觉得自己的隐私已经受到了威胁.当大数据变得更为普遍的时候，情况将更加不堪设想。大数据大大地威胁到了我们的隐私和自由，这都是大数据带来的新威胁。但是与此同时，它也加剧了一个旧威胁:过于依赖数据，而数据远远没有我们想象的那么可靠。

在由“小数据”时代向大数据时代转变的过程中，我们对信息的一些局限性必须给予高度的重视。数据的质量可能会很差;可能是不客观的;可能存在分析错误或者具有误导性;更糟糕的是，数据可能根本达不到量化它的目的。尽量避免

收到数据的统治。

6.2. 掌控：责任与自由并举的信息管理

随着世界开始迈向大数据时代，社会也将经历类似的地壳运动。在改变我们许多基本的生活和思考方式的同时，大数据早已在推动我们去重新考虑最基本的准则，包括怎样鼓励其增长以及怎样遏制其潜在威胁。大数据时代，对原有规范的修修补补已经满足不了需要，也不足以抑制大数据带来的风险，我们需要全新的制度规范，而不是修改原有规范的适用范围。