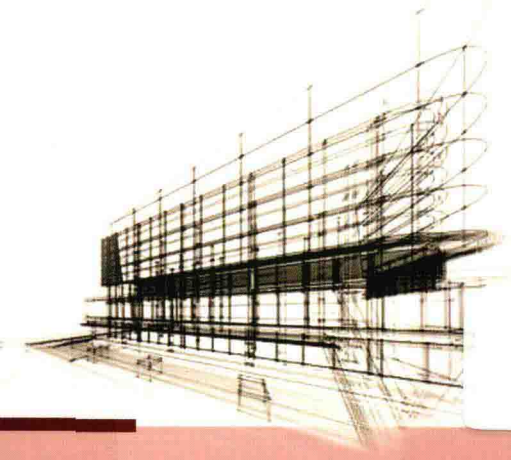
经济与管理研究文库

大数据时代的经济计量分析

李庆华 著



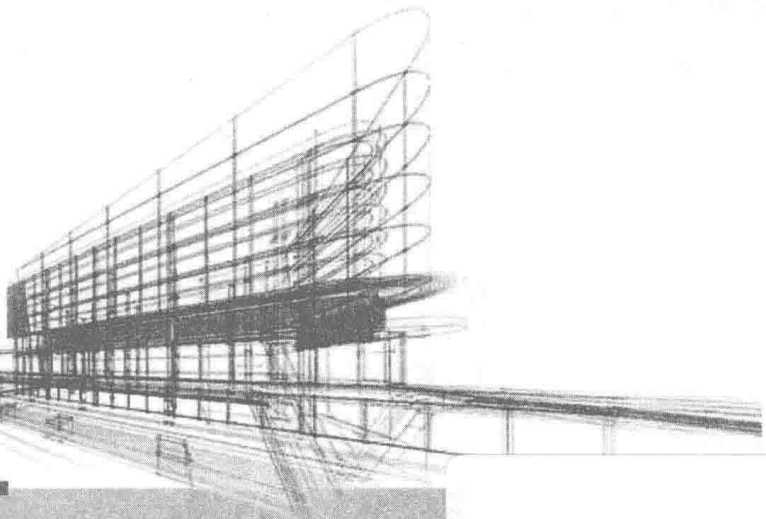
華中師範大學出版社

经济与管理研究文库

2015年湖北省学术著作出版专项资金资助项目

大数据时代的经济计量分析

李庆华 著



華中師範大學出版社

新出图证（鄂）字 10 号
图书在版编目（CIP）数据

大数据时代的经济计量分析/李庆华 著. —武汉：华中师范大学出版社，2015.12
（经济与管理研究文库）
ISBN 978-7-5622-7173-4

Ⅰ.①大… Ⅱ.①李… Ⅲ.①经济计量分析—研究 Ⅳ.①F224.0

中国版本图书馆 CIP 数据核字（2015）第 266251 号

大数据时代的经济计量分析

责任编辑：苏 睿　　责任校对：刘 峥　　封面设计：甘 英 胡 灿
编辑室：学术出版中心　　电话：027－67863220
出版发行：华中师范大学出版社有限责任公司
社址：湖北省武汉市洪山区珞喻路 152 号　　邮编：430079
电话：027－67863426（发行部）　　传真：027－67863291
网址：http://www.ccnupress.com　　电子信箱：hscbs@public.wh.hb.cn
印刷：湖北新华印务有限公司　　督印：王兴平
开本：710mm×1000mm 1/16　　印张：12.5
字数：220 千字　　版次：2015 年 12 月第 1 版
印次：2015 年 12 月第 1 次印刷　　定价：33.00 元
欢迎上网查询、购书

序

从纽约到北京，从硅谷到光谷，有关大数据时代的话题正在被快速地传播。随着智能手机和便携式信息与计算设备的出现和飞速发展，人们的行为、物理位置，乃至身体机能的变化都成为可被记录和分析的数据。以此为基础，“网络经济”、“网络金融”等新经济和新商业模式正在开始或已经形成并以闪电式的速度发展。可以毫不夸张地说，一个新时代已经来临。这个时代是以“大数据”的生产和运用为特征的。

在人类历史长河中，即使是在现代社会日新月异的发展中，人们还主要是依赖抽样数据、局部数据和片面数据，甚至在无法获得实证数据的时候纯粹依赖经验、理论、假设和价值观去发现未知领域的规律。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。大数据时代的来临使人类第一次有机会和条件，在非常多的领域和非常深入的层次上获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律，获取过去不可能获取的知识，得到过去无法企及的商机。

大数据的出现，使得通过数据分析获得知识、商机和社会服务的能力从以往局限于象牙塔之中的学术精英圈子扩大到了普通的机构、企业和政府部门。门槛的降低直接导致数据的容错率提高和成本的降低，最重要的是，人们可以在很大程度上从对因果关系的追求中，从繁琐与复杂的数学模型中解脱出来，转而将注意力放在相关关系的发现和使用上。只要发现两个现象之间存在着显著相关性，就可以创造巨大的经济效益和社会效益。

大数据之所以可能成为一个“时代”，在很大程度上是因为这是一个可以由社会各界广泛参与，八面出击，处处结果的社会运动，而不仅仅是少数

专家学者的研究对象。所以，以小样本数据为依据的传统乃至现代经济计量技术如何在大数据时代简化甚至于通俗化已经成为迫在眉睫的问题，也许还关系到相关学科的生存问题。过去那种在分析经济变量间相互关系时所需要面对的诸多问题，如异方差性问题、自相关问题，多重共线性问题等，还是问题吗？过去建立在小样本或者小数据基础上的诸如面板数据方法、空间计量经济技术等，在大数据时代，其应用价值是否会变化？等等。这些都是计量经济学学者所必须面对和回答的问题，这是时代赋予的责任。

在新时代，大数据不仅是研究社会经济活动的最重要的依据，而且它本身将成为十分重要的产品。与小数据不同，它不再具有排他性。这就是说它具有公共性，就像现代社会基础设施一样，如公路、铁路、港口、水电和通信网络。但就其价值特性而言，大数据却和这些物理化的基础设施不同，它不会因为人们的使用而折旧和贬值。例如，一组 DNA 可能会死亡或毁灭，但数据化的 DNA 却会永存。所以，很多学者认为，世界的本质就是数据。因此，大数据时代的经济学、政治学、社会学和许多学科门类都会发生巨大甚至是本质上的变化和发展，进而影响人类的价值体系、知识体系和生活方式。哲学史上争论不休的世界可知论和不可知论也许将会转变为实证科学中的具体问题。

未来将是信息技术飞速发展的大数据时代，有一位电商界巨头认为，30 年以后，iPhone 32 届时会出现，它会储存 5 000 亿首歌曲，相当于 3 亿份报纸的信息量，而依靠云存储技术和高速网络，储存和传送这些信息将不成问题；而这也将使得信息总量空前膨胀，在未来十几年，人类所创造的信息量将是过去 300 年的总和；30 年以后，所有的东西会通过物联网被连接起来，无论是手提电脑，或者是手持设备，还是眼镜、衣服、鞋子、墙等，甚至是一头牛都可以被物联网连接起来；今天每个人都有 2 个移动设备，在 2020 年时，每个人被连接的设备数量将会达到 1 000 个。

面临这样一个崭新的时代，我们准备好了吗？在从事经济学、政治学、社会学和许多科学的教学和研究工作的学者都有所准备吗？

《大数据时代的经济计量分析》正是适应这个时代需要的一部作品，是作者在从事《计量经济学》教学与研究工作 20 多年的基础上，对大数据条件下经济计量分析与小数据（小样本）时代有什么联系与区别进行的分析与探讨。

在小数据时代（这里，将大数据时代之前的时代，或者没有考虑大数据时代的影响的时期均称为小数据时代），由于经济数据难以获得，或者获得经济数据的成本较高，经济研究者往往对经济数据中样本的多少与得失十分看重，他们更关注“更好”的方法，以最大可能解决或缓解小样本条件下估计精度问题。例如，在传统计量经济学中，为了缓解回归模型中的自相关问题，采用了广义差分法，但这个方法会导致失去一个观测数据，为了补救所失去的这一个数据，采用了温斯腾变量，以补上差分过程中所失去的这一个观测数据。学者们对经济数据样本点的多少的计较程度，由此可见一斑；又如，传统计量经济学在分析经济惯性时，是采用分布滞后模型进行的。这里，学者们对滞后长度问题、回归过程中的多重共线性问题显示了极高的关注程度，提出了多种解决方案；再如，在研究两个变量之间的关系时，多元线性回归模型告诉人们：要将解释变量放到多元回归模型中，这样才能估计出解释变量对应变量的边际影响，这样才能控制住其他变量，或者说，这样才能研究，保持其他变化不变的条件下，一个解释变量对应变量的影响，等等。

然而，在大数据时代，大数据所具有的公共属性，使数据的获得不再是困难的，大样本成了普遍现象。这时，个别样本点的得失还有那么重要吗？是研究方法的“高、大、上”重要还是对数据本身的观察重要？对普通应用者，还需要关注回归系数的最小二乘估计的分布性质吗？传统经济计量分析中的异方差性问题、自相关问题和多重共线性问题还是那么重要吗？是研究数据表面所呈现的相关关系重要，还是研究其因果关系重要？等等。这些问题，本书将会回答。

目　录

1 大数据时代的数据集特征*

1.1 更多：不是随机样本，而是超大样本，近乎是全体数据

大数据时代的数据处理技术会发生革命性变化，人们可以获得“全部”数据：“样本＝总体”。

“大数据”全在于发现和理解信息内容及信息与信息之间的关系，然而直到最近，我们对此似乎还是难以把握。IBM 的资深“大数据”专家杰夫·乔纳斯(Jeff Jonas)提出要让数据“说话”。从某种层面上来说，这听起来很平常。人们使用数据已经有相当长一段时间了，无论是日常进行的大量非正式观察，还是过去几个世纪里在专业层面上用高级算法进行的量化研究，都与数据有关。

在数字化时代，数据处理变得更加容易、更加快速，人们能够在瞬间处理成千上万的数据。但当我们谈论能“说话”的数据时，我们指的远远不止这些。

实际上，大数据与三个重大的思维转变有关，这三个转变是相互联系和相互作用的。首先，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。其次，我们乐于接受数据的纷繁复杂，而不再追求其精确性。再次，我们的思想发生了转变，不再探求难以捉摸的因果关系，转而关注事物的相关关系。

很长一段时间以来，准确分析大量数据对我们而言都是一种挑战。过去，因为记录、储存和分析数据的工具不够好，我们只能收集少量数据进行分析，

---

* 本章借鉴了赵国栋等《大数据时代的历史机遇》和涂子沛《大数据》第四章的相关内容。

这让我们一度很苦恼。为了让分析变得简单，人们会把利用有关理论找出所谓最优数据量，在经济计量分析中，这就是最优样本容量。这是一种无意识的自省：我们把与数据交流的困难看成是自然的，而没有意识到这只是当时技术条件下的一种人为的限制。如今，技术条件已经有了非常大的提高，虽然人类可以处理的数据依然是有限的，也永远是有限的，但是我们可以处理的数据量已经大大地增加，而且未来会越来越大。在某些方面，我们依然没有完全意识到自己拥有了能够收集和处理更大规模数据的能力。我们还是在信息匮乏的假设下做很多事情，建立很多机构组织。我们假定自己只能收集到少量信息，结果就真的如此了。这是一个自我实现的过程。我们甚至发展了一些使用尽可能少的信息的技术。别忘了，统计学的一个目的就是用尽可能少的数据来证实尽可能重大的发现。事实上，我们形成了一种习惯，那就是在我们的制度、处理过程和激励机制中尽可能地减少数据的使用。为了理解大数据时代的转变意味着什么，我们需要首先回顾一下过去。

小数据时代的随机采样，用最少的数据获得最多的信息。直到最近，私人企业和个人才拥有了大规模收集和分类数据的能力。在过去，这是只有教会或者政府才能做到的。当然，在很多国家，教会和政府是等同的。有记载的、最早的计数发生在公元前 8 000 年，当时苏美尔的商人用黏土珠来记录出售的商品，大规模的计数则是政府的事情。数千年来，政府都试图通过收集信息来管理国民。

以人口普查为例，据说古代埃及曾进行过人口普查，《旧约》和《新约》对此都有所提及。那次由奥古斯都·恺撒主导实施的人口普查，提出了“每个人都必须纳税”，这使得约瑟夫和玛丽搬到了耶稣的出生地伯利恒。1086 年的《末日审判书》(*The Doomsday Book*)对当时英国的人口、土地和财产做了一个前所未有的全面记载。英国皇家委员走遍整个国家对每个人、每件事都做了记载，后来这本书用《圣经》中的《末日审判书》命名，因为每个人的生活都被赤裸裸地记载下来的过程就像接受“最后的审判”一样。

然而，人口普查是一项耗资巨大且费时费力的事情。国王威廉一世(King William Ⅰ)在他发起的《末日审判书》完成之前就去世了。但是，除非放弃收集信息，否则在当时没有其他办法。尽管如此，当时收集的信息也只是一个大概情况，实施人口普查的人也知道他们不可能准确记录每个人的信息。实际上，“人口普查”这个词来源于拉丁语“censere”，意思就是推测、估算。

300 多年前，一位名叫约翰·格朗特(John Graunt)的英国缝纫用品商提出了一种很有新意的方法。他采用了一种新方法推算出鼠疫时期伦敦的人口数，这种方法就是后来的统计学。这个方法不需要一个人一个人地计算。虽然这个方法比较粗糙，但采用这个方法，人们可以利用少量有用的样本信息来获取人口的整体情况。但是，因为样本分析法一直都有较大的漏洞，因此，无论是进行人口普查还是其他大数据类的任务，人们还是一直使用一一清点这种“笨拙”的方法。

考虑到人口普查的复杂性以及其耗时长、耗费大量钱财的特点，政府极少进行人口普查。古罗马在拥有数十万人口的时候每 5 年普查一次。美国宪法规定每 10 年进行一次人口普查，而随着国家人口越来越多，普查人口以百万计数。到 19 世纪为止，即使这样不频繁的人口普查依然很困难，因为数据变化的速度超过了美国人口普查局统计分析的能力。

美国在 1880 年进行的人口普查，耗时 8 年才完成数据汇总。因此，他们获得的很多数据都是过时的。他们预计，1890 年进行的人口普查要花费 13 年的时间来汇总数据。即使不考虑这种情况违反了宪法的规定，它也是很荒谬的。然而，因为税收分摊和国会代表人数的确定都是建立在人口的基础上的，所以必须要得到正确的数据，而且必须是及时的数据。美国人口普查局面临的问题与当代商人和科学家遇到的问题很相似。很明显，当他们被数据淹没的时候，已有的数据处理工具已经难以应付了，所以，就需要有更多的新技术。后来，美国人口普查局就和当时的美国发明家赫尔曼·霍尔瑞斯(Herman Hollerith)签订了一项协议，用他的穿孔卡片制表机来完成 1890 年的人口普查。经过大量的努力，霍尔瑞斯成功地在 1 年时间内完成了人口普查的数据汇总工作。这简直就是一个奇迹，它标志着自动处理数据的开端，也为后来 IBM 公司的成立奠定了基础。但是，将其作为收集处理大数据的方法依然过于昂贵。毕竟，每个美国人都必须填一张可制成穿孔卡片的表格，然后再进行统计。在这么麻烦的情况下，很难想象如果不足 10 年就要进行一次人口普查应该怎么办。但是，对于一个跨越式发展的国家而言，10 年一次的人口普查的滞后性已经让普查失去了大部分意义。这就是问题所在，是利用所有的数据还是仅仅采用一部分数据呢？最明智的自然是得到有关被分析事物的所有数据，但是，当数量无比庞大时，这又不太现实。那如何选择样本呢？有人提出有目的地选择最具代表性的样本是最恰当的方法。1934 年，波兰统计学家耶日·奈曼(Jerzy Neyman)指出，这

只会导致更多更大的漏洞。事实证明，问题的关键是选择样本时的随机性。统计学家们证明，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。虽然这听起来很不可思议，但事实上，一个对1 100人进行的关于“是否”问题的抽样调查有着很高的精确性，其精确度甚至超过了对所有人进行调查时的97%。这是真的，不管是调查10万人还是1亿人，20次调查里有19次都能猜对。为什么会这样？原因很复杂，但是有一个比较简单的解释就是，当样本数量达到某个值之后，我们从新个体身上得到的信息会越来越少。就如同经济学中的边际效应递减一样，认为样本选择的随机性比样本数量更重要，这种观点是非常有见地的。这种观点为我们开辟了一条收集信息的新道路。通过收集随机样本，我们可以用较少的花费做出高精准度的推断。因此，政府每年都可以用随机采样的方法进行小规模的人口普查，而不是只能每10年进行一次。事实上，政府也这样做了。例如，除了10年一次的人口大普查，美国人口普查局每年都会用随机采样的方法对经济和人口进行200多次小规模的调查。当收集和分析数据都不容易时，随机采样就成为应对信息采集困难的办法。

很快，随机采样就不仅仅应用于公共部门和人口普查了。在商业领域，随机采样被用来监管商品质量。这使得监管商品质量和提升商品品质变得更容易，花费也更少。以前，全面的质量监管要求对生产出来的每个产品进行检查，而现在只需从一批商品中随机抽取部分样品进行检查就可以了。本质上来说，随机采样让大数据问题变得更加切实可行。同理，它将客户调查引进了零售行业，将讨论焦点引进了政治界，也将许多人文问题变成了社会科学问题。

随机采样取得了巨大的成功，成为现代社会、现代测量领域的主心骨。但这只是一条捷径，是在不可收集和分析全部数据的情况下的选择，它本身存在许多固有的缺陷。它的成功依赖于采样的绝对随机性，但是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。最近，以固定电话用户为基础进行投票民调就面临了这样的问题——采样缺乏随机性。因为没有考虑到只使用移动电话的用户——这些用户一般更年轻和更热爱自由；没有考虑到这些用户，自然就得不到正确的预测。2008年在奥巴马与麦凯恩之间进行的美国总统竞选中，盖洛普咨询公司、皮尤研究中心(Pew)、美国广播公司和《华盛顿邮报》社这些主要的民调组织都发现，如果他们不把移动用户考虑进来，民意测试结果就会出现3个点的偏差，而一旦考虑进来，偏差就只有

1 个点。鉴于这次大选的票数差距极其微弱，1 个点的偏差已经是非常大的偏差了。更糟糕的是，随机采样不适合考察子类别的情况。因为一旦继续细分，随机采样结果的错误率会大大增加。这很容易理解。倘若你有一份随机采样的调查结果，是关于 1 000 个人在下一次竞选中的投票意向。如果采样时足够随机，这份调查的结果就有可能在 3%的误差范围内显示全民的意向。但是，如果这个 3%左右的误差本来就是不确定的，却又把这个调查结果根据性别、地域和收入进行细分，结果是不是越来越不准确呢？用这些细分过后的结果来反映全民的意愿，是否合适呢？设想一下，一个对 1 000 个人进行的调查，如果要细分到“东北部的富裕女性”，调查的人数就远远少于 1 000 人了。即使是完全随机的调查，倘若只用了几十个人来预测整个东北部富裕女性选民的意愿，还是不可能得到精确结果。而且，一旦采样过程中存在任何偏见，在细分领域所做的预测就会大错特错。

因此，当人们想了解更深层次的细分领域的情况时，随机采样的方法就不可取了。在宏观领域起作用的方法在微观领域失去了作用。随机采样就像是模拟照片打印，远看很不错，但是一旦聚焦某个点，就会变得模糊不清。随机采样也需要严密的安排和执行。如果人们只能从采样数据中得出事先设计好的问题的结果，那么，千万不要奢求采样的数据还能回答你突然意识到的问题。所以，虽说随机采样是一条捷径，但它终究也只是一条捷径。随机采样方法并不适用于一切情况，因为这种调查结果缺乏延展性，即调查得出的数据不可以重新分析以实现计划之外的目的。

我们来看一下 DNA 分析。由于技术成本大幅下跌以及在医学方面的广阔前景，个人基因排序成了一门新兴产业。2012 年，基因组解码的价格跌破1 000美元，这也是非正式的行业平均水平。从 2007 年起，美国硅谷的新兴科技公司 23 andme就开始分析人类基因，价格仅为几百美元。这可以揭示出人类遗传密码中一些会导致其对某些疾病抵抗力差的特征，如乳腺癌和心脏病。23 andme 希望能通过整合顾客的 DNA 和健康信息，了解到用其他方式不能获取的新信息。23 andme对某人的一小部分 DNA 进行排序，标注出几十个特定的基因缺陷。这只是该人整个基因密码的样本，还有几十亿个基因碱基对未排序。最后，23 andme只能回答其标注过的基因组表现出来的问题。发现新标注时，该人的 DNA 必须重新排序，更准确地说，是相关的部分必须重新排列。只研究样本而不是整体，有利有弊：能更快更容易地发现问题，但不能回答事先未考虑到的

问题。

苹果公司的传奇总裁史蒂夫·乔布斯在与癌症斗争的过程中采用了不同的方式，成为世界上第一个对自身所有 DNA 和肿瘤 DNA 进行排序的人。为此，他支付了高达几十万美元的费用，这是23 andme报价的几百倍之多。所以，他得到的不是一个只有一系列标记的样本，他得到了包括整个基因密码的数据文档。对于一个普通的癌症患者，医生只能期望他的 DNA 排列同试验中使用的样本足够相似。但是，史蒂夫·乔布斯的医生们能够基于乔布斯的特定基因组成，按所需效果用药。如果癌症病变导致药物失效，医生可以及时更换另一种药，也就是乔布斯所说的，“从一片睡莲叶跳到另一片上”。乔布斯开玩笑说：“我要么是第一个通过这种方式战胜癌症的人，要么就是最后一个因为这种方式死于癌症的人。”虽然他的愿望都没有实现，但是这种获得所有数据而不仅是样本的方法还是将他的生命延长了好几年。

在数据时代，人们分析数据的模式应该是全数据模式，在这种模式下，样本＝总体。

在信息处理能力受限的时代，世界需要数据分析，却缺少用来分析所收集数据的工具，因此随机它应运而生，它也可以被视为那个时代的产物。如今，计算和制表不再像过去一样困难。感应器、手机导航、网站点击和 Twitter 被动地收集了大量数据，而计算机可以轻易地对这些数据进行处理。采样的目的就是用最少的数据得到最多的信息。当我们可以获得海量数据的时候，它就没有什么意义了。数据处理技术已经发生了翻天覆地的改变，但我们的方法和思维却没有跟上这种改变。采样一直有一个被我们广泛承认却又总有意避开的缺陷，它忽视了细节考察，现在这个缺陷越来越难以忽视了。虽然我们别无选择，只能利用采样分析法来进行考察，但是在很多领域，从收集部分数据到收集尽可能多的数据的转变已经发生了。如果可能的话，我们会收集所有的数据，即“样本＝总体”。正如我们所看到的，“样本＝总体”是指我们能对数据进行深度探讨，而采样几乎无法达到这样的效果。上面提到的有关采样的例子证明，用采样的方法分析整个人口的情况，正确率可达 97%。对于某些事物来说，3%的错误率是可以接受的。但是你无法得到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。我们不能满足于正态分布一般中庸平凡的景象。生活中真正有趣的事情经常藏匿在细节之中，而采样分析法却无法捕捉到这些细节。

谷歌对“流感趋势”的预测并不是依赖于对随机样本的分析，而是分析了整个美国几十亿条互联网检索记录。分析整个数据库，而不是对一个小样本进行分析，能够提高微观层面分析的准确性，甚至能够推测出某个特定城市的流感状况，而不只是一个州或是整个国家的情况。Farecast 的初始系统使用的样本包含 12 000 个数据，所以取得了不错的预测结果。随着奥伦·埃齐奥尼不断添加更多的数据，预测的结果越来越准确。最终，Farecast 使用了每一条航线整整1 年的价格数据来进行预测。埃齐奥尼说：“这只是一个暂时性的数据，随着你收集的数据越来越多，你的预测结果会越来越准确。”所以，我们现在经常会放弃样本分析这条捷径，选择收集全面而完整的数据。我们需要足够的数据处理和存储能力，也需要最先进的分析技术。同时，简单廉价的数据收集方法也很重要。过去，这些问题中的任何一个都很棘手。在一个资源有限的时代，要解决这些问题需要付出很高的代价。但是现在，解决这些难题已经变得简单容易得多。曾经只有大公司才能做到的事情，现在绝大部分的公司都可以做到了。

通过使用所有的数据，我们可以发现如若不然则将会在大量数据中淹没掉的情况。例如，对信用卡诈骗是通过观察异常情况来识别的，只有掌握了所有的数据才能做到这一点。在这种情况下，异常值是最有用的信息，你可以把它与正常交易情况进行对比。这是一个大数据问题。而且，因为交易是即时的，所以你的数据分析也应该是即时的。

Xoom 是一个专门从事跨境汇款业务的公司，它得到了很多拥有大数据的大公司的支持。它会分析每一笔交易的所有相关数据。2011 年，它注意到用“发现卡”从美国新泽西州汇款的交易量比正常情况多一些，系统于是启动报警。Xoom 公司的首席执行官约翰·孔泽(John Kunze)解释说：“这个系统关注的是不应该出现的情况。”单独来看，每笔交易都是合法的，但是事实证明这是一个犯罪集团在试图诈骗。而发现异常的唯一方法就是，重新检查所有的数据，找出样本分析法错过的信息。然而，使用所有的数据并不代表这是一项艰巨的任务。大数据中的“大”不是绝对意义上的大，虽然在大多数情况下是这个意思。谷歌对流感趋势的预测建立在数亿的数学模型上，而它们又建立在数十亿数据节点的基础之上。完整的人体基因组有约 30 亿个碱基对，但这只是单纯的数据节点的绝对数量，并不代表它们就是大数据。大数据是指不用随机分析法这样的捷径，而采用所有数据的方法。谷歌对“流感趋势”的预测和乔布斯的医生们采取的方法就是大数据的方法。日本国民体育运动“相扑”中非法操纵比赛结果

的发现过程，就恰到好处地说明了使用“样本＝总体”这种全数据模式的重要性。消极比赛一直被极力禁止，备受谴责，很多运动员深受困扰。芝加哥大学的一位经济学家斯蒂夫·列维特(Steven Levitt)在《美国经济评论》上发表了一篇研究论文，其中提到了一种发现这种情况的方法——查看运动员过去所有的比赛资料。他的畅销书《魔鬼经济学》(*Freakonomics*)中也提到了这个观点，他认为检查所有的数据是非常有价值的。列维特和他的同事马克·达根(Mark Duggan)使用了11年中超过64 000场摔跤比赛的记录，来寻找异常性。他们获得了重大的发现。非法操纵比赛结果的情况确实时有发生，但是不会出现在大家很关注的比赛上。冠军赛也有可能被操纵，但是数据显示消极比赛主要还是出现在不太被关注的联赛的后几场中。这时基本上没有什么风险，因为很多选手根本就没有获奖的希望。相扑比赛的一个比较特殊的地方是，选手需要在15场赛事中的大部分场次取得胜利才能保持排名和收入。这样一来就会出现利益不对称的问题。当一名7胜7负的相扑选手碰到一名8胜6负的对手时，比赛结果对第一个选手来说极其重要，对他的对手而言则没有那么重要。列维特和达根发现，在这样的情况下，需要赢的那个选手很可能会赢。这看起来像是对手送的“礼物”，因为在联系紧密的相扑界，帮别人一把就是给自己留了一条后路。有没有可能是要赢的决心帮助这名选手获胜的呢？答案是：有可能。但是数据显示的情况是，需要赢的选手的求胜心也只能把胜率提高25%。所以，把胜利完全归功于求胜心是不妥当的。对数据进行进一步分析可能会发现，与他们在先前比赛中的表现相比，当他们再相遇时，上次失利的一方要拥有比对方更高的胜率。因为在相扑界，你的付出总会有所“回报”，所以第一次的胜利看上去更像是一名选手送给另一名选手的礼物。这个情况是显而易见的。但是如果采用随机采样分析法，就无法发现这个情况。而大数据分析通过使用所有比赛的极大数据捕捉到了这个情况。这就像捕鱼一样，开始时你不知道是否能捕到鱼，也不知道会捕到什么鱼。

一个数据库并不需要有以太字节计的数据。在这个相扑案例中，整个数据库包含的字节量还不如一张普通的数码照片包含的多。但是大数据分析法不只关注一个随机的样本。这里的“大”，取的是相对意义上的“大”，而不是绝对意义上的“大”，也就是说这是相对所有数据来说的。很长一段时间内，随机采样都是一条好的捷径，它使得数字时代之前的大量数据分析变得可能。但就像把一张数码照片或者一首数码歌曲截取成多个小文件似的，在采样分析的时候，

很多信息就丢失了——你能欣赏一首歌的抽样吗？拥有全部或几乎全部的数据，我们就能够从不同的角度，更细致地观察和研究数据的方方面面。

我们可以用 Lytro 相机来打一个恰当的比方。Lytro 相机是具有革新性的，因为它把大数据运用到了基本的摄影中。与传统相机只可以记录一束光不同，Lytro 相机可以记录整个光场里所有的光，达到 1 100 万束之多。具体生成什么样的照片则可以在拍摄之后再根据需要决定。用户没必要在一开始就聚焦，因为该相机可以捕捉到所有的数据，所以之后可以选择聚焦图像中的任一点。整个光场的光束都被记录了，也就是收集了所有的数据，“样本＝总体”。因此，与普通照片相比，这些照片就更具“可循环利用性”。如果使用普通相机，摄影师就必须在拍照之前决定好聚焦点。同理，因为大数据是建立在掌握所有数据，至少是尽可能多的数据的基础上的，所以我们就可以正确地考察细节并进行新的分析。在任何细微的层面上，我们都可以用大数据去论证新的假设。是大数据让我们发现了相扑中的非法操纵比赛结果、流感的传播区域和对抗癌症需要针对的那部分 DNA，它让我们能清楚分析微观层面的情况。当然，有些时候，我们还是可以使用样本分析法，毕竟我们仍然生活在一个资源有限的时代。但是更多的时候，利用手中掌握的所有数据成了最好也是可行的选择。

社会科学是被“样本＝总体”撼动得最厉害的学科。随着大数据分析取代了样本分析，社会科学不再单纯依赖于分析实证数据。这门学科过去曾非常依赖样本分析、研究和调查问卷。当记录下来的是人们的平常状态，也就不用担心在做研究和调查问卷时存在的偏见了。现在，我们可以收集过去无法收集到的信息，不管是通过移动电话表现出的关系，还是通过 Twitter 信息表现出的感情。更重要的是，我们现在也不再依赖抽样调查了。

艾伯特·拉斯洛·巴拉巴西(Albert-László Barabási)和他的同事想研究人与人之间的互动，于是，他们调查了 4 个月内所有的移动通信记录——当然是匿名的，这些记录是一个为全美国五分之一人口提供服务的无线运营商提供的。这是第一次在全社会层面用接近于“样本＝总体”的数据资料进行网络分析。通过观察数百万人的所有通信记录，我们可以产生也许通过任何其他方式都无法产生的新观点。有趣的是，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，这个关系网会变得没那么高效但却不会解体；但如果把一个与所在社区之外的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多小块。这个研究结果非