实证社会科学

（第三卷）

主编 钟杨
副主编 吴建南 樊博

Social Science Research

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

实证社会科学

Social Science Research

（第三卷）

钟　杨　主编

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

图书在版编目(CIP)数据

实证社会科学. 第三卷 / 钟杨主编. —上海：上海交通大学出版社，2017

ISBN 978-7-313-17432-1

Ⅰ. ①实… Ⅱ. ①钟… Ⅲ. ①社会科学-文集 Ⅳ. ①C53

中国版本图书馆 CIP 数据核字(2017) 第 131735 号

实证社会科学(第三卷)

主　　编：钟　杨

出版发行：上海交通大学出版社　　地　　址：上海市番禺路 951 号

邮政编码：200030　　电　　话：021-64071208

出 版 人：郑益慧

印　　刷：上海天地海设计印刷有限公司　　经　　销：全国新华书店

开　　本：787mm×1092mm　1/16　　印　　张：8.5

字　　数：127 千字

版　　次：2017 年 6 月第 1 版　　印　　次：2017 年 6 月第 1 次印刷

书　　号：ISBN 978-7-313-17432-1/C

定　　价：42.00 元

实证社会科学

Social Science Research

主办单位：上海交通大学国际与公共事务学院

主　　编：钟　杨

副 主 编：吴建南　樊　博

CONTENTS

Introduction of Social Science Data

Instructions for Authors

目 录

学者评论

大数据与数据无关*

Gary King**

“大数据”，也就是数据科学(Data Science)，在不同领域有很多名称：在化学中，它被称为“化学计量学”；在生物学中，则被称为“生物统计”；在经济学中，是“计量经济学”；在政治科学中则是“政治学方法论”。实际上，“大数据”是媒体向社会公众报道数据科学领域的相关信息时提出的词汇。与上述专业术语相比，这一称谓出色地指明了数据科学的精髓：大数据的重点在于分析方法而非数据本身。如果没有合理的分析方法，大数据不仅不能让事情变得更容易，甚至会让问题变得更加棘手。但是，大数据为我们创造了新的机会，如果把握好这些机会，我们就能在众多领域取得丰硕的研究成果。特别是在社会科学领域，作为研究人类自身的重要方法，大数据的研究意义更是非比寻常。

一、大数据对当代社会生活的影响

问大家一个问题，对于你及你家人的生活方式影响最大的科学研究是什么？我们可以列出一长串的研究成果。例如基因革命，它研究人体的结构和运行方式并为我们治愈一些疾病提供了支持；基本粒子的发现(如希格斯粒子)；天文学对地外星系的观测和对类地行星的探索；还有就是过去一两个世纪内由于医疗条件改善而带来的人类寿命的成倍增长。这些都可以列入伟大的发明研究中。但我认为定量社会科学也是这一系列的伟大发明之一。为什么这样说？不论你称它为大数据、数据科学还是数据分析学，它都在使我们的

---

* 本文根据 Gary King 教授 2017 年 1 月 4 日在上海交通大学国际与公共事务学院所做的专题学术报告整理而成。

** Gary King，哈佛大学教授，定量社会科学研究所主任（Director of the Institute for Quantitative Social Science at Harvard University）。

生活迅速数据化。如果感兴趣的话,你会发现我们周边的一切都在生产数据。钟表、摄像机、手机都在生产数据;在公司内,大量有效信息被收集并录入人力资源或财务系统,管理人员不仅通过分析数据来制定决策,而且还通过观测反馈数据调整其运营策略。大数据不仅提升了公司的数据生产和运用能力,而且帮助大多数公司从传统运行模式转向更具效率的大数据运营模式;此外,大数据的产生和发展还催生了大量新兴产业。例如社交媒体,它创造并改变了我们的社交网络,空前提升了人们的表达能力;它还改变了竞选方式,推动了经济、司法和公共医疗等领域的诸多变革。大数据甚至改变了体育运动,《点球成金》这部电影便是一部将大数据应用于体育的案例。在当今社会,体育运动也能利用数据去分析评价。很多有趣的公共政策问题都可以通过大数据进行分析。大数据对这一领域的影响力与日俱增。尽管手机、摄像机或其他数据采集设备在推动现代社会数据化的过程中功不可没,但没有定量研究,我们就难以对数据进行有效分析,之前提到的种种产业变革便不会发生。与数据相比,数据分析方法才是未来大数据发展的重中之重。

二、大数据的价值在于分析方法
——基于不同案例的说明

那么,大数据的价值在哪里呢?大数据的价值不仅与信息技术和数据采集设备无关,而且也与数据本身无关。如前所述,数据实际上是信息技术发展的副产品。例如,学校设立新的信息系统是为了方便学生注册,但该系统在使注册工作更加方便高效的同时,也收集了大量数据。所以说,数据可能是不经意间产生的。随着大量高校引入此类信息系统,由于市场竞争,其价格也会不断下降。于是,即使高校与为其提供相应服务的公司没有去刻意收集数据,其积累的数据也会与日俱增。由此可见,数据的获取并非难事,只要付出一点努力,你的数据收集量就会不断增长。然而,实现数据的价值有赖于相应的分析方法。只有我们能够合理运用分析方法,才能从数据中有所收获,并知道如何以完全不同的方式利用这些数据。接下来,我将通过我或我的同事在研究中遇到的各种案例,来说明分析方法对大数据应用的重要性。

首先,让我们了解一下数据分析方法在提升数据运算效率中的显著作用。众所周知,根据摩尔定律,计算机的运算速度和性能每 18 个月便会翻倍。但与数据学家花费一下午的时间通过优化算法所提升的运行速度相比,它也只

能甘拜下风。我有一位同事每过几天就要收集并处理一些数据，随着时间的推移，他积累的数据越来越多。终于有一天，他的计算机已经不能处理如此庞大的数据。所以他向学校 IT 部门咨询道："告诉我，要买多大的计算机才能运行我的数据？"他得到的回答是："需要一台价值两百万美元的超级计算机。"尽管他的确可以找人来赞助这笔费用，但是我让两名研究生花一小时改进了一下算法，就使之前需要超级计算机才能完成的运算仅需要一台笔记本电脑和 20 分钟便可解决。由此可见，通过分析方法提升大数据运算效率的效费比要远高于硬件设备。随着分析方法不断改进，它对大数据的发展将产生更为显著和深远的影响。大数据令人兴奋，但是如果没有分析方法，大数据便会毫无价值。

那么，与传统社科问题相比，大数据又需要面对怎样的新问题呢？我有一位已经退休的哈佛同事，为了研究积极参与者如何影响公共政策。他随机调查了 15 000 名美国人，向他们询问诸如"你是否是一名政治积极分子？""你是否花费时间去影响政策与政治？"之类的问题。然后，他将 15 000 个调查对象缩减到 2 000 人，对他们做了更为详细的调查，并基于调查结果写了一本关于政治实践主义(Political Activism)的学术名作。其主要结论是，建设高效国家的前提，是拥有通过各种方式积极参与公共政策的公民。这本重要的政治学专著对社会科学也存在重要意义，因为它告诉我们想要了解政府和社会如何运转，就必须去与人们接触互动。然而当今社交媒体中关于政治观点和公共政策意见的信息多达兆亿。事实上，全球每天都有六亿五千万条社交媒体信息。你要如何去处理这些数据？回到家中写到卡片纸上然后叠放在你的公寓里？当数据量从六亿五千万变成七亿五千万的时候，数据会更有用吗？当然不会，这会变得更棘手。但是，如果我们能弄明白如何分析这些数据的话，其中蕴藏的机遇也是不可限量的。我们根本不需要 2 000 个访谈就能知道整个社交网络上数以亿计的用户的观点。如前所述，这其中有巨大的潜力，当然难度也是空前的。相比之下，分析 2 000 个结构化的访谈信息可比分析这六亿五千万条内容多样，语言也不尽相同的留言容易多了。但挑战越多，潜力越大。更多数据并不能让事情简化，而需要文本分析方法从旁协助。以锻炼为例，如果公共卫生人员通过询问来测量人们的运动量，例如他们上周运动的次数，但作为调查对象的我们真能如实回答这个问题吗？也许你回答的是自己的跑步次数而不是运动的次数，也许你认为自己是一个只喜欢看电视的人，所以你的回答可能不真实。那我们又如何能使用这些信息呢？所幸现在我们有诸如手

机和应用软件等现代数据收集设备,可用于记录我们的位置和运动量。即使如此,如何正确地处理这些数据中的内在联系仍然是一项挑战。同样以上文中的运动问题为例,身处高速行驶的列车上,即使我静止不动,手机中的应用也会持续记录运动里程。想厘清此类关联并不容易,但这正是我们能发挥作用的部分。下面,我会通过自己的一系列研究来向大家进一步说明这个问题。

在此之前,我们先来通过一些医学和文本分析方面的案例了解一下现行分析方法的局限以及改进方向。在医疗注册系统健全的国家,公共卫生部门可以通过尸检来确定逝者的死因,并统计不同死亡原因的死亡人数与比例,进而采取措施预防疾病或疫情。然而,世界上大多数欠发达国家并没有相关记录。一种解决方法就是口头验尸,找到死亡现场的其他目击者询问他们一系列简单问题,例如,病患死亡前是否肚子疼痛?是否在流血?然后将这些问题的答案交给医生,医生就会对死亡原因作出判断。但这种做法的问题在于,不同医生可能会对同一病例产生分歧和误判,而这种谬误可能会因为各种外在因素而放大。例如,上海的医生未必能准确诊断坦桑尼亚等地的常见病例,因为上海几乎没有疟疾等当地常见病;而坦桑尼亚甚至缺乏足够的医疗人员。我们用相同的方法再来看看另一个案例。苹果公司收集了社交媒体中对"iPhone"的所有公开评价,并统计了与苹果手机相关的积极词汇和消极词汇的词频,如"苹果手机真棒,太好了。"或"苹果手机烂死了,不如扔到厕所里。"但这种看似高端的做法实际上和"口头验尸"别无二致。几个月前我在新加坡,当地政府希望了解人们对哪些公共政策感兴趣,所以他们检索了社交媒体中关于公共政策的关键字,发现人们对于教育尤为关注,如学校教育(schooling)、上课、教科书等词汇出现非常多。据此,政府官员认为有必要在教育领域投入更多的精力。但他们忽略了一点,彼时正值夏季奥运会期间,新加坡产生了第一位奥林匹克金牌获得者:Joseph Schooling。他的姓氏与网络搜索的关键词 schooling 恰好重叠,但是政府官员并没注意到这一点,相应的分析结果自然也宣告无效。类似的案例不胜枚举,由此可见,语句分析中的词频统计在很多情况下存在明显局限。回到之前关于"口头验尸"的问题,我们在上海重新讨论了这一问题并得出结论:公共卫生领域的工作人员无须专注于个体病例,而是要将逝者的死因进行归类。若 90%的死亡都是由于心脏病引起的,剩下 10%是其他疾病,当你在为某一位逝者进行死因归类时,选择心脏病的正确率便有 90%。此时,多数人会追求 90%的正确率。尽管正确的目标应该是实事求是地分类,但是少数错误的分类带来的偏差并不会产生明显

的负面影响。所以，我们的方法不是要准确分类，而是要提升整体分类准确率。实际上，我们不仅用以上观点在世界卫生组织上海会议中说服了与会者，而且还将其应用到社交媒体的文本分析中。毕竟，我们不必关注每个人在推特或微信上说了什么，而是要设法弄清作为整体的“人们”在社交媒体中关注的议题类型。由我协助创办的 Chimson Hexagon 公司便是一家使用这套方法的媒体数据分析公司，该公司目前在全球十大创新公司排名第七。同一台电脑，同一套编码，同样的分析方法，虽然数据不同，但效果依然不错。

那么这种分析方法有没有进一步改进的空间呢？当然有。接下来我将通过电脑辅助阅读与国会议员发言记录分析的案例向各位说明这一点。大家是否还记得几年前的波士顿马拉松爆炸案？我们对 43 名本科生做了一个实验，让他们从一万条关于波士顿马拉松爆炸案的推特和微博中筛选出与凶手相关的关键词。学生们筛选出的关键词数的中位数是 8 条，那学生们一共总结出多少个不同的关键词呢？149 条。这意味着如果我们只让一名学生去完成这项工作，那么他遗漏其他人找到的关键词的概率几乎高达三分之二！人类在关键词分类这个问题上是靠不住的。尽管我们每时每刻都在使用搜索引擎，在网上搜索关键词，但实际上我们并不擅长这件事情，这不是很奇怪么？尽管我们能判断某些关键词是否具有实际意义，但我们却很难记住所有的词汇。然而，我们可以让程序帮助我们弥补这项缺陷，而这正是其优势所在。我的团队研发了名为“Concilience”的系统，该系统可以使用技术推荐 50 或 100 个关键词，并由研究人员人工判定保留的词汇。这个系统可以阅读 100 亿条文件，并迅速将大量的文件分成少数几类。这一系统在针对美国国会议员发言记录的文本分析中发挥了明显作用。众所周知，美国的政客们会通过宣扬政绩（Credit Claiming），如申明政治立场、为选取做出的贡献来争取选票。一直以来，政治学都在对政客们的发言进行研究，并试图对其中的议题进行归类。但是面对浩如烟海的新闻文本，即使最勤勉的学者也无能为力。然而，我们使用上文中提到的新系统分析了 64 000 篇参议院议员的新闻稿，并发现了一个独立于经济和外交政策等传统议题的新议题：党派嘲讽（Partisan Taunting），即一个政党总是会拿另一个政党开玩笑。例如，国会参议员 Lautenberg 用一幅画着身着军服的鸡的漫画炮轰共和党人是“鸡鹰”，或共和党人在奥巴马的国情咨文会议上站起来说：“他在撒谎。”尽管这些内容看上去似乎别有深意，但它们的确与经济、政治还有外交等传统议题无关，只是为了取笑对方。我甚至认为 2016 年的美国大选基本上就是一场党派嘲讽。实际上，分析结果表明，

大选期间27%的新闻稿都是关于党派嘲讽的,即每四篇新闻中就有一篇属于这一议题。然而,这一议题甚至是之前我们都没有考虑到的类型。所以说,分析方法不仅能帮助我们解决问题,甚至可以帮助我们发现还没有意识到的问题。

那么除学术研究以外,数据分析与大数据对社会和民生又产生了何种影响呢?现在让我通过大数据在提升学生阅读能力和美国社保政策决策辅助方面的案例说明它对宏观政策和人们的日常生活带来的影响。首先,让我们来看一看大数据在现代高等教育中的应用。请问在座的各位学生,你们有多少人会花钱购买教材?又有多少人会按时完成老师布置的课前阅读作业?在西方国家,前者的数量低于50%,而后者只有20%到30%。这不仅使随堂测验变成了学生的梦魇,而且还使教学质量变得惨不忍睹。为什么学生群体难以独立完成他们的学习任务呢?这是因为教育仍然是一个集体性经历的过程,而不是一个独自学习的过程。打个比方,尽管iTunes上的音乐质量很好,但人们在自身集体行动的本能的驱使下,还是会花大价钱去听音乐会。有鉴于此,我们发明了一款具有分析功能的电子阅读器——Perusall。该系统不仅具有文献阅读、重点标记和成绩记录等传统功能,而且还与我们创立的集体学习平台相关联。在平台上,不仅学生们可以提出疑问或与其他同学交流观点,而且教授也可以根据系统对学生讨论记录的分析结果,发现教学过程中存在的普遍问题,并在课堂教学时直入重点。这一系统在提升学生的学习效率、增加师生互动之余,也有效改善了现代高校的学习环境和教学质量。然后,让我们了解一下大数据和数据分析在公共政策制定过程中的辅助决策功能。

当前,社会保险是美国政府的一项主要开支。它不仅帮人们脱离贫困,而且也为退休和残障人士等社会弱势群体提供保障。美国社保基金实行现收现付制度,也就是说,你缴纳的费用被支付给已经退休的人们,而当你退休时,彼时的年轻人也会为你买单。这意味着如果人们的预期寿命增长或退休人口激增,社保基金就会面临失衡的风险。在过去的85年里,社保管理机构每年都会通过质性预测方法预测资金需求量,但收效甚微。然而,在大数据时代,算法的改进使得更精准的预测成为可能。我们通过测算发现,社保基金的收支平衡转折点大约出现在2000年。尽管相关部门矢口否认并宣称整个系统运作良好,但实际上社保失衡的问题正是从那时起不断恶化,并为政府带来了不少麻烦。由此可见,当前大多数政策中都存在诸多不可预期的变量,所以及时评估政策并采取措施还是非常必要的。回到社保问题上,我们的进一步预测

显示，当前社会基金大约存在 80 亿美元的赤字。这次，政府管理部门就开始据此调整政策了。这就是运用大数据分析的意义。

三、终结定性与定量之争

尽管随着信息技术和统计方法的进步，大数据为我们处理之前不曾或者不能处理的数据提供了可能，但我们仍然需要意识到，定性研究不仅离我们并不遥远，而且比我们想象的更重要。定性与定量研究的对立存在于各个科研领域，而且这两种研究往往是彼此交融的。实际上，不仅定性研究需要定量研究来帮助其验证各种观点，定量研究也需要定性研究为其提供数据量化的理论依据。得益于现代科技，文本、影像等大量研究资料既可以用于定性研究，也可以用于定量研究，这意味着二者的联系变得更加紧密而非疏离。在未来的大数据时代，定性研究和定量研究唯有携手共进，才能有所突破。如果我们将数据比作一辆汽车，那定量研究便是车辆的引擎，而定性研究则是汽车的方向盘。总之，唯有两种研究通力合作，才能研制出人类可控的计算机技术，让人们在信息高速公路上纵横驰骋。这就是我今天想展示给你们的内容。