数据科学与政治分析的融合之作

大数据与机器学习

复杂社会的政治分析

BIG DATA AND MACHINE LEARNING

Political Analysis in Complex Society

董青岭◎著

当前，大数据及其分析技术的应用已成新的社会进步驱动力量。本书主要着眼于现代社会的复杂性，重点阐述了现代政治的数据化进程与基于数据的政治分析策略。

时事出版社

据科学与政治分析的融合之作

大数据与机器学习

复杂社会的政治分析

BIG DATA AND MACHINE LEARNING

Political Analysis in Complex Society

董青岭◎著

时事出版社
北京

图书在版编目（CIP）数据

大数据与机器学习：复杂社会的政治分析/董青岭著. —北京：时事出版社，2018.2

ISBN 978-7-5195-0136-5

Ⅰ.①大… Ⅱ.①董… Ⅲ.①数据处理②机器学习 Ⅳ.①TP274②TP181

中国版本图书馆 CIP 数据核字（2017）第 201054 号

出 版 发 行：时事出版社
地　　址：北京市海淀区万寿寺甲 2 号
邮　　编：100081
发 行 热 线：（010）88547590　88547591
读者服务部：（010）88547595
传　　真：（010）88547592
电 子 邮 箱：shishichubanshe@sina.com
网　　址：www.shishishe.com
印　　刷：北京旺都印务有限公司

开本：787×1092　1/16　印张：13.25　字数：200 千字
2018 年 2 月第 1 版　2018 年 2 月第 1 次印刷
定价：80.00 元
（如有印装质量问题，请与本社发行部联系调换）

本书为对外经济贸易大学中央高校
基本科研业务费专项资金（批准号：
CXTD8－05）资助成果，特此致谢。

目录 Contents

下篇　数据原理

绪　论

争论中的大数据、机器学习与未来政治

当前，大数据和机器学习的兴起无疑已成社会科学研究的潮流范式，作为一种全新的数字化生存方式，大数据正在改变我们的生活以及我们理解世界的方式。通过对推特（Twitter）、谷歌（Google）、脸谱（Facebook）、微博等新媒体平台信息的挖掘和计算，政治研究者不仅可以跟踪大城市的抗议活动、发现恐怖主义行迹、明晰国家战略风险，还可对利益攸关人群进行精细划分、对政治态势进行整体感知、对危机进行预警和预测，从而有助于政治决策者们进行科学决策和进行高效政治沟通。总体而言，在诸多政治研究者和国际关系学者看来，大数据不仅在改变着我们社会生活的本来面貌，更在改变着我们对现实世界的认知以及如何认知这个世界的方式。但也有另外的声音认为，大数据既非“科学”也不“革命”，不过是传统统计方法的大样本化升级，花哨且复杂的分析技术并不一定会带来多少有关政治形势和外交战略的全新洞察，更不会轻易导致政治学说和外交决策模式发生革命性改变。因为在这些专家学者们看来，即使存在着大体量的结构化、半结构化和非结构化数据可供挖掘研究，但在政治领域，各个国家政府也会出于国家安全和公民个人隐私保护考量，设置重重法律门槛和技术障碍以阻止数据在不受道德伦理约

束的情形下肆意扩散，更加之以微信、微博、推特和脸谱等社交媒体为代表的流式数据体量越大，数据本身存在的噪音也就越大，数据的开发价值就越低。就此而言，在那些持谨慎态度的学者看来，大数据场景下的政治科学革命并不是那么容易发生的。

当然，还有其他观点认为，与其研究价值相比，大数据在政治学和国际关系领域的应用所带来的风险要更为突出，也更为值得关注。因为，首先大数据政治研究难以回避的一个问题就是必然会涉及数据的跨疆界流动，姑且不论一国有无权利跨越主权疆界挖掘和使用他国数据，单就技术风险而言，一旦大数据成为政治决策和外交执行的常态，则数据技术弱的国家极易为技术强的国家所窥探、掌控和摆布，数据争夺将诱发更多的“棱镜计划”。除此以外，以 Killing Robots 为代表的智能杀人技术的应用更会带来前所未有的伦理挑战和法律风险，毕竟机器学习即使训练样本中很小的误差放大至数亿人群中，也有可能导致数百人乃至数万人被错误识别为恐怖分子或暴乱分子而枉杀。再者，大规模的数据勘探在增加了社会的透明度和能见度之外，同时也存在着诱导社会走向数据极权之可能，甚至有学者担心人类的未来要么为智能机器所掌控、要么为数据精英所摆布。但不管怎么说，未来时代的政治，数据本身连同数据分析都将成为越来越严肃的社会政治问题。

放眼未来，尽管存在种种争议，当前大数据介入政治学和国际问题研究已然是大势所趋。与传统的小样本调查、小数据分析相比，大数据具有从多样和多源数据中快速获取信息的能力，与传统统计分析不同，其技术目标主要是着眼于非结构化数据的结构化处理，这在很大程度上有助于政治研究者应对微博、微信和推特等非结构化数据的爆炸性增长，同时也意味大数据时代的政

治研究很可能不同于以往时代，基于数据驱动的政治科学研究将更加倾向那些数据密集型问题的研究，力求掌握与研究对象有关的更多数据甚至是全样本数据，着力刻画研究对象的整体特征，而非局部细节；在结构化数据之外甚至力图容纳诸多类型非结构化数据的存在，力求使用数据的混杂性，而非数据的齐整性表征，研究对象的性质属性、预测其政治行为；更加重视多源数据之间的联系性以及事物变化之间的相似性，而非仅仅因果性。有鉴于此，有越来越多的政策分析人士和政治学者认为，大数据适合分析多元混杂数据，同时又不仅仅追求因果解释，而是着力探讨看似不相关因素之间的相似性，这看似是科学的倒退，但在工程学应用上可能比传统方法更适于捕捉复杂多变政治环境的不确定性，从而也更有利于推进政治研究的科学化和技术化。

概括而言，大数据以处理非结构化数据和即时流式数据见长，并以数据可视化作为分析结果的展现形式，可以在短时间内高速处理百万维度以上的高维数据，这就使得政治形势研判和外交决策有可能确立在充沛数据分析和动态感知之上，以前那些由于技术处理水平达不到而被刻意忽视、被遗弃的信息有可能会被重新挖掘和发现，并有机会进入到政治决策过程并影响决策结果，在坚实数据支撑下，传统的政治理论和外交指导原则很可能即使不被重构，也要被重新审视和修正。可以说，作为一种新的社会生活现实和科学研究技术手段，虽然大数据能否带来政治分析的革命性变革尚未可知，但大数据辅助形势判断和外交决策已在实践领域中存在诸多应用场景，诸如数据反恐、模拟投票和社会情感分析等等，不一而足。在某种意义上，数据分析技术的进步已然将政治分析推进到了一个寻求算法解决的时代。

一、政治分析中的数据跨界流动

显而易见，作为一种新兴事物，人们对大数据的认知尚未尽知其理，甚至可以说是一知半解。但即便如此，在技术变革的驱动下，大数据对政治分析领域的渗透和介入已然势不可挡。纵览各种文献，大数据在当前政治分析和国际关系研究领域的应用，主要有以下三个趋势最为明显：第一，舆情挖掘与精准政治营销。通过数据痕迹抓取和聚类分析，大数据可以精准确定事件发生地域、事件人群以及人群属性，定制化精准推送政治营销广告和实施精准公共外交战略。第二，即时数据监控与态势感知。通过多源即时数据监控和大规模云端计算，即时监测、锁定和跟进事态发展并自动生成事件报告和危机预警，从而动态掌控问题爆点并提前推进基于态势感知的政治沟通和预防外交战略。第三，多源数据混杂建模与关联分析。通过多源数据采集和数据组合算法，在各种结构化和非结构化数据资源中发掘事件关联关系和节点因素，优化政治沟通和决策过程，从而实现最优化政治资源配置。

以上是政治分析领域大数据的三种主导应用趋势，但不管哪一种应用趋势，数据是基础，拥有数据才会拥有政治洞见。由此涉及两个跨越主权边界的法理问题：其一，一国的研究机构能否跨越主权边界到另外一国采集和挖掘数据，采集后的数据能否跨越主权疆界自由流动？其二，如果科学研究中的数据共享是必要且必须的，那么谁拥有数据的所有权、使用权和收益权？如果数据的归属权是清晰的，那么分属于不同国家的数据主体能否进行数据共享或数据自由交易？与商业领域不同，政治域内的数据采集通常会碰触到政府机密、政治稳定和国家安全等敏感问题，因而数据跨界采集的门槛极高，哪些数据可以为外国公司、民众和

政府采集和处理，哪些数据不可以可以跨界出境，都通常会成为各国政府数据开放战略（OPEN DATA）首要关注的问题。当然，即使是在商业领域，数据的跨界采集与跨界流动也通常是受到严格限制的，诸如印度和俄罗斯就规定从事互联网运营和服务的数据采集器、服务器，必须处于业务主管国的主权管辖之下，无论是商业银行交易信息还是私人手机通讯信息都不得置于境外存储，毕竟数据一旦流出，谁也不能保障商业信息或手机通讯信息不被用作恐怖活动或国家谍报行为。换言之，不少学者主张，数据是天然固有主权属性并从属于国家主权管辖的。这一观点的主要立论理由包括：

首先，数据的产生依附于特定的人格主体，数据所承载并展现的是特定人格主体的基本属性和行为轨迹，因而数据的所有权、使用权和处置权应优先归属于产生它的特定人格主体。关于这一点，不少学者做如下场景想象：假定人们到某一电商网站购物（无论是外国的还是本土的），人们通常会登记自己的基本属性信息，诸如姓名、性别、电话、邮箱以及家庭住址和银行账号等，那么在购物和服务结束时，人们所登记的这些信息是不是就归属于电商网站所有呢？如果人们发现这些个人信息在他们毫不知情的情形下被出售给其他商家，进而每天收到各种广告骚扰，那么人们是否有权利要求商家删除这些信息或停止出售这些信息呢？再进一步讲，这些信息的出售和再次出售不断产生着经济收益，人们是否有权利要求利益分割或损害补偿呢？显而易见，人们在注册购物信息时明确点击同意了电商网站的数据采集合同条款，但这并不等于人们同步转让了与自身信息相关的数据所有权。人们之所以同意电商数据收集条款那是因为只有提供准确信息才能方便电商投递商品并顺利完成此次交易，人们明确意思表

示并实质授权的是个人数据有限次数（一次或多次）、有限目的（为顺利完成交易）的特定数据使用权，而不是数据所有权、处置权和收益权。在逻辑上，数据因人们自身属性和行为而产生并从属于人格个体，当一次或N次交易完成以后，人们当然有理由要求商家删除涉及自身隐私的数据以及因此产生的网络痕迹（Cookies）。与此同理，人们去医院看病时，医院并不能因为为患者提供了服务而将患者的个人信息占为己有，更不得擅自将患者的个人信息和病例擅自公开、转让或移交第三方。

其次，特定的人格主体总是附着于特定国籍，而特定国籍恰恰是确立国家主权管辖的基本依据。换言之，如果数据因人的属性和行为而所产生，而人又从属于某一特定国家主权，那么是否可以说数据本身也是有主权或受主权管辖呢？更加具象一点：如果某农户A的绵羊啃食了B户的庄稼，B农户会要求A农户赔偿，只因为那是A的绵羊，A当然负有责任管理好绵羊的行为并对绵羊的行为后果承担责任；与之相类比，如果数据从属于我们个人，而我们个人又从属于国家，那么国家是否对我们的个人数据及其使用负有管辖责任呢？诸如服务器等硬件设施从属于某个公司所有并受公司管辖支配那样，如果主权当局认为个人数据流出境外有伤国家安全或有碍公民个人隐私，那么数据当局是否有权利禁止数据出境呢？显而易见，大数据时代的政治分析，最大的难题是数据的可获取与数据的跨界流动性。如果数据是有主权的，那么任何一个国家都可以随时随地以危害国家安全或侵犯个人隐私为由阻止数据的跨境获取与跨境流动，从而使得大数据政治分析和国际关系研究的前景并不那么乐观。

再次，如果数据是可以跨界流动的，人们又焉知跨界流动的数据是政府开源数据还是个人隐私数据？数据出境以后，数据分

析是否会被用于诸如攻击水坝、交通设施或银行系统等非法目的呢？显见，数据的安全性直接关乎数据的跨界流动性，如果数据的安全性不能确保，则数据的跨界流动性就会极大受阻。在现实的数据科学研究中，即便学者也要考虑数据的安全性与流动性平衡。对此，一部分专家提出，数据脱敏可以塑造数据安全，脱敏以后的数据是群体画像而非个人信息，因而是可以自由交易、自由流动和自由研究的。但问题是，数据脱敏到什么程度才可以不会析出个人隐私或国家安全问题？个人数据被隐去姓名是不是就可以成为脱敏数据随意使用了？搜索引擎将用户的 IP 地址隐去以后，将用户的搜索行为信息转手给广告公司是不是也算数据脱敏了？显而易见，数据脱敏并不是一个可以确保数据安全的好的战略，因为何谓敏感数据以及数据脱敏到什么程度，到目前为止根本未存在一个通行标准。另外还有学者和商业公司认为，加密数据是安全的、可交易的，因为数据加密以后，只有买家和卖家知道数据交易的内容，数据交易平台只是提供数据交易的场所而不窥探数据交易的内容，从而使得数据交易两头可见而中间不可见，由此可以确保数据传输过程和交易过程的安全性。基于这一观点，数据可以像股票和期货一样在交易所内挂牌交易。但问题来了，加密以后的数据对使用者、研究者以及监管者而言，因为加密又怎么知道该次交易不涉及国家安全或个人隐私呢？如果对数据交易的中间过程不加以监管，又怎知数据不会被非法传输、非法迁移和非法使用呢？如果加以监管，又有多少数据是可以在不伤害国家安全或个人隐私的前提下而被科学家们随意使用呢？这是一个科学进步与数据伦理的二重悖论，同时也是数据科学研究所无法会回避的现实问题。

二、争议中的“大数据”概念

传统上，由于数据采集和数据分析手段的限制，很多时候政治决策和外交政策执行是建立在经验感知的基础之上的，无论是数据体量、数据真实性、数据生成方式还是数据获取渠道都是极为有限的，小样本抽样调查、历史经验知觉感悟与因果逻辑推演是人们洞察这个纷繁芜杂世界的主要决策基础，透过小样本调研和结构化数据分析，人们所获得的知识多是关于我们所生活之世界的线性因果逻辑，以至于很多时候世界的复杂性和不确定性被刻意忽略了。在有限获取数据和有限提取信息的条件下，政治分析和外交决策执行通常只能专注于问题的一个侧面而无法顾及大局，有时甚至无法洞悉和还原政治事件的本来面目。在此情形下会经常性地出现理论与现实相脱节的决策偏差。有鉴于此，很多学者认为，大数据具有出色的非结构化数据处理能力，而用以处理数据的机器学习方法又多以非线性模型为主，二者在当前时代的完美结合或许可以使传统政治研究步入新的知识发现。

然而，大数据真的能够提供比小数据更多的理论洞察与政治洞见吗？答案众说纷纭，目前争论异常激烈：其一，大数据的“大”未必真的“大”。当前，绝大部分的大数据分析存在“大而失真”。以当前较为盛行的大数据网络收视率调查为例，抗日神剧的高吐槽率和高点击率真的代表该剧最受欢迎、最火爆吗？显而易见，这不仅仅是一种错觉，更是一种对大数据的误解。在某种意义上，大数据版的网络收视率调查是以网络吐槽等非结构化文本数据为分析对象的，然而一部神剧之所以吐槽率高，很大程度上不是因为人们喜欢该剧而是因为人们厌恶它，因为厌恶才吐槽并不等于该剧收视率高或到户率高；再退一步讲，纵然吐槽的人成千上万，又焉知吐槽的人占了点击该剧人群的几成呢？在

通常情况下，绝大数人观而不语，在点击该剧的人群中究竟是吐槽的人多还是不吐槽的人多呢？显见，如果有1万人观看了该剧，有1千人吐槽，在吐槽的人中喜欢该剧的占了51%，由此得出抗日神剧是受欢迎的，如此大数据版的收视排名是不是严重失真呢？类似的大数据应用误解同样存在于政治分析领域，针对脸谱的大数据分析是不是就代表了美国民众的主流民意呢？也许使用脸谱的主要用户是美国18—40岁之间的年轻人，而这群年轻人又占美国民众整体的几成呢？即使是单就这群年轻人而言，又有多少用户乐于表达自己的生活意见和政治见解呢？也许，只有几十个或几个百用户是活跃用户，他们的政治表达是否就能够代表整个国家的民意呢？在此情形下，大数据版的民意调查有没有考虑剔除重复数据和高频无效数据呢？有没有考虑数据的失真问题呢？相对于传统民意调查，大数据在政治分析领域同样面临数据的清洗、整理和交叉验证等问题。

其二，大数据偏重相似性比较，但仅有相似性是不够的。目前，学界有一种观点认为，大数据并不追求因果逻辑而是旨在探索看似不相关要素之间的相关关系，这是一种极大的误解。因为大量的不同维度、不同类型的数据聚合在一起，的确可以发现很多数据的相似之处，但仅有这种数据相似性是不可以直接构成相关解释的，在很大程度上不同维度数据的相似性是需要进一步多角度去伪存真检验的。举个例子，从空间位置大数据来看，世界石油储量的地理分布与世界伊斯兰教的地理分布几乎是重合的，无论是中东、中亚、北非、东南亚还是中国新疆，石油储量普遍存在于信仰伊斯兰宗教的地理区域，两组数据的相似度高达80%以上，但我们依然很难判断石油与伊斯兰宗教是不是存在必然的相关关系，更难以确定究竟是石油争端导致宗教冲突还是宗

教冲突影响石油争端？同样的情形，即使通过大规模的数据相似性训练，谷歌翻译也依然难以判断“china”这个词此处意为“瓷器”还是指涉“中国”，“war”仅仅只是一种激烈争吵状态的言语修辞还是真的战争状态？显而易见，纵然欧洲的沃尔玛超市发现尿布的销售与啤酒的销量存在数据相似性，但也很难判断究竟是不是买婴儿尿布的那些顾客顺手买走了旁边的啤酒。简单的来说，大体量的数据相似性比较有助于研究者在社会科学研究中发现研究奇点，但不一定会产生关于研究对象之间关系的最真实解释。很多时候，数据量越大、噪声也就越大，没有经过去噪声处理的大数据分析是很容易偏离事物本质的。

三、走向大数据时代的政治分析

传统上，政治形势的感知、冲突预防战略的部署乃至外交政策的执行主要依赖于智囊团队和政治精英的经验感知，这些人拥有丰富的历史、政治、军事和战略知识，并对形势的走向和冲突的爆发有着敏锐的直觉。事实上，在目前的很多情况下，一个经验丰富的决策者的直觉也许要比客观数据要可靠得多。数据是会说谎的，一个数据的可靠性跟统计方法、样本选择等一系列因素有关，初期的差之毫厘如果没有被察觉或者纠正，运算到最后很可能会导致结果与真相谬之千里，这无关乎数据体量的大与小、数据维度的多与少。即便数据都是客观真实的，但如果选择了错误的数据类型和算法模型来做判断依据，那么决策仍然会出现失误，严重的甚至会出现南辕北辙。总体而言，大数据时代的到来并不意味着传统经验感知和逻辑分析方法的终结；恰恰相反，数据与直觉需要相关支撑、相互印证，方能对政治判断和政策执行产生更具启发性的指导。

当然，传统的逻辑经验分析在大数据时代也确实存在诸多劣势和需要改进之处。单就传统的经验判断而言，同样需要数据的收集和分析，但由于技术水平和技术条件的限制，人们只能使用采样的方法，力求用最少的数据得到最多的信息，数据采集需要层层下发指令，再经过收集筛选层层上报，不仅数据收集时间长，而且数据失真度大，当数据真正变成决策参考变量时往往已错过事件最佳决策期。而如今，进入互联网和大数据时代，很多问题便可以借助大数据来解决，诸如恐怖分子的日常联络、通知发布、经济往来和地点定位等信息传递行为，或多或少都会在网络上留下蛛丝马迹，只要能够通过有效途径获取到这些信息，经过大数据和机器学习技术处理，便可以不同程度地掌握恐怖分子的动向，对恐怖组织的下一步活动进行态势感知，提前做好冲突预防准备。与传统政治决策相比，大数据决策的优势在于信息的海量获取和高速处理，从而政策响应更迅速、更及时。而这对于我们这个越来越纷繁复杂、信息极度泛滥世界而言，高速的大体量信息处理与快速的政策反应当然是必要的。

综上所述，在大数据时代，大数据介入政治分析不仅是可行而且也是必要的。但首先我们要正确理解大数据的含义、性质与缺陷，大数据并非体量越大越好，也非算法越复杂分析越精确。数据的质量决定着分析预测的准确与否。数据在采集、传输、分类、整理和储存的过程中，总会掺杂进去某些干扰因素；大数据固然体量大，但其特点是价值密度低，海量的数据同时意味着数据噪音多。因此，被收集而来的数据不是每一项都会被用到，大数据使用首先要进行数据清洗和去噪声。其次，数据跨境采集需要正视法律和政治障碍，特别是国家安全与个人隐私忧虑。概括来讲，数据隐私与安全问题主要存在两个领域：一是数据收集前