编者按 进入 21 世纪以来,大数据及其分析技术迅速崛起,逐渐介入国际关系研究议程并发挥越来越大的作用。那么,应该如何看待大数据国际关系研究的缘起和发展?该研究包含哪些基本原理?我国学者在大数据国际关系研究领域有哪些尝试?这一研究的未来发展前景是怎样的?为此,本刊特约记者王海媚专访对外经济贸易大学国际关系学院董青岭教授,董教授主要从事大数据科学与国际关系的交叉研究,内容涵括大数据海外舆情监测与冲突预警、国际关系自然语言处理与社会情感挖掘、机器学习与国际关系智能分析等,代表作有《复合建构主义:进化合作与进化冲突》《大数据与机器学习:复杂社会的政治分析》《大数据安全态势感知与冲突预测》《新战争伦理:规范和约束致命性自主武器系统》等。

21 世纪以来中国的大数据国际关系研究

——董青岭教授访谈

董青岭　王海媚

一、大数据国际关系研究的缘起与发展

王海媚(以下简称"王"):董教授您好!当前,大数据及其分析技术的

崛起正在成为一种新的社会科学范式，您是如何看待大数据介入国际关系研究议程的？

董青岭（以下简称“董”）：当前，大数据及其分析技术的应用已经深入到社会生活的每一个角落。伴随着社会生活的“网络化”和“数字化”趋势的不断发展，数据体量将呈爆炸性增长、数据价值也将得到前所未有的释放，以数据公司和数据科学家为代表的数据精英正在成长为新的政治力量，新的社会结构也正围绕着数据的存储、挖掘和应用而展开。这主要体现在以下三个方面：首先，作为现代政治基础的民主选举活动正在被编程化的舆论“机器人”和各种“算法偏见”所操控，大数据及附着于数据之上的算法对决越来越显现为未来权力角逐的幕后驱动力量；其次，“数据驱动型外交”或依托于数据及其算法的“智慧型外交”正在开启人工智能时代的外交革命，尤其是在跨国政治沟通和冲突预防领域，大数据精准政治营销、大数据海外舆情监测，以及大数据反恐和早期冲突预警都已大显身手；再次，以智能决策和自主杀人为特征的军事机器人研发正在掀起新一轮军备竞赛，人类正在被自我创造的人工智能网络和漫无节制的数据使用所伤害。历史从来没有像今天这样，拥有数据便意味着主宰一切，数据即生活、数据即权力，一方面，我们越来越受益于数据革命所带来的种种生活便利、憧憬着一个高度智能化社会的到来；另一方面，我们又深刻恐惧于数据革命所带来的种种社会变革，数据的深度挖掘正使得我们的社会越来越透明化、越来越脆弱、越来越不安全。

王：您能简单介绍一下大数据与国际关系相结合这一跨学科交叉研究的起源和发展吗？

董：大数据介入国际关系研究正在受到越来越多研究者的关注，但其兴起和发展需要满足以下两方面条件：其一，有关研究对象的丰裕数据基础。随着社会生活网络化、数据化和智能化趋势的日渐增强，社会实践主体之间的高频互动每天将产生难以计数的数据痕迹，这使得国际关系研究能够获得比以往任何时期都更为丰富的数据信息。迈入大数据时代，国际关系研究存在走向“数据密集型科学研究”的可能；其二，有关数据处理的

突破性技术进步。当前，大数据分析技术的进步，特别是非结构化数据库（如MongoDB和HBase）和分布式并行计算系统（如Hahoop和Spark）的出现，不仅解决了大体量混杂数据的采集、存储和计算问题，而且还能够使国际关系研究能够抵近观察微观主体之间的互动细节。正因如此，大数据及其分析技术的介入或将打破传统国际关系研究范式，传统范式强调以群体间政治为核心观察对象、以结构主义为主导分析路径、以小样本归纳为主要知识生产方式、以传统因果律为逻辑基础。大数据或许是我们重塑现行国际关系理论、外交指导思想及冲突预测方法的历史性契机。

基于上述动因和背景，如果要粗线条地回顾一下大数据与国际关系研究的结合进程，我认为大体可以分为两个阶段：

第一阶段可称之为“数据模拟阶段”即“计算机仿真实验阶段”。它肇始于1971年哈佛大学教授托马斯·谢林（Thomas Schelling）《隔离的动态模型》一文的发表，[①]该文在计算机尚未普及甚至尚未成熟的年代，在学界率先提出未来的学者将兼具社会科学知识和编程技术，借助计算机的强大算力，学者们将就所研究问题生成随机模拟数据、建立博弈规则和形式模型，进而仿真社会互动进程、研究政治的动态演进，后世称之为“谢林模型”。简单来讲，“谢林模型”认为，计算机模拟不仅可以帮助学者解决大体量可观测样本的随机生成问题（如数十亿条电话号码、地理位置和身份信息），而且还可以使研究对象和推理过程通过编程技术动态可视化（如NetLogo软件对各种社会学、政治学和自然科学问题的可视化模拟）。在某种意义上，“谢林模型”不仅启发了后世有关社交网络和博弈论应用的模拟分析，而且还极大影响了学界有关种族、宗教、贫富和党派等对抗问题的理解，开启了计算机模拟与政治分析相结合的学术研究先河。沿着计算模拟这条路径，这一时期最引人注目的成果莫过于美国密歇根大学政治学与公共政策教授阿克塞尔罗德（Robert Axelrod）使用计算机模拟来破解“囚

① Thomas Schelling, “Dynamic Models of Segregation,” *Journal of Mathematical Sociology*, 1971, Vol.1, pp. 143-186.

徒困境”，并写就《合作的进化》一书。[①]

第二阶段可称之为“数据分析阶段”，也可冠之以“大数据与机器学习阶段”。20 世纪 80 年代中期后，很多学者对计算机模拟社会问题提出质疑，这些质疑点包括：其一，计算机模拟情景下的人际互动更加接近自然科学过程中的“变量控制实验”，在很大程度上刻意忽视或漠视了真实社会过程中真实数据与模拟数据的差异。在模拟过程中，不仅形式规则是由研究者主观设定，而且变量的挑选也受到研究者严格的条件限定，因为滤掉了某些至关重要的真实信息，有时模拟结果与真实社会情景相去甚远；其二，计算机模拟忽视了真实社会情景中人的相互学习和进化能力，忽略了在反反复复的社会互动中人类个体具有从实践中汲取经验教训并改进、优化行为模式的进化学习本能。就此而言，以“谢林模型”为代表的早期计算机模拟并未真正触及社会系统的开放性与复杂演进性。

直到最近，由于数据体量的爆炸性增长、数据价值的不断释放和数据处理技术的突飞猛进，有学者开始关注到，大数据及用以处理大数据的机器学习技术要比计算机模拟更适合分析国际关系问题：首先，机器学习是在非过滤、非控制信息的情形下，利用算法程序从嘈杂数据中去归纳、分类和识别模式，而不是像计算机模拟那样利用形式逻辑去演绎规则，它对数据的处理尊重社会系统的开放性、研究变量的非控制性及测量对象的相互扰动性；其次，机器学习具有环境自适应性和学习进化特征，可以根据环境的变化感知数据的细微波动，进而重构模式识别并调整预测输出结果；再次，大数据在结构化数据之外试图容纳并分析各种非结构化数据（如海量的新闻报道、社交网络对话和网页浏览痕迹等），追求数据的多样性、混杂性而非精确性。就此而言，机器学习的优势恰恰在于具有从杂乱、混合数据中寻找可辨别模式的能力，因而，大数据与机器学习较传统研究方法更容易捕捉国际社会的复杂多变性和不确定性。

王：目前，国内外学界在大数据应用于国际关系研究方面都开展了哪

① 〔美〕阿克塞尔罗德：《合作的进化》，吴坚忠译，上海人民出版社 2007 年版。

些研究、取得了怎样的成果？

董：作为一种新兴事物，由于技术门槛的限制，当前大数据在国际关系领域中的应用并不是十分普遍和活跃，现有研究主要集中在以下两个方面：

其一，面向实时数据自动采集的新型数据库建设，旨在利用大数据技术重构国际关系研究的底层数据基础。与传统的国际关系研究常用数据库，例如，战争相关因素数据库（COW）、乌普萨拉武装冲突数据库（UCDP）和全球恐怖主义数据库（GTD）不同，[①]新一代数据库的建设将着力应对当下汹涌而来的数据洪流，不仅数据体量巨大、数据产生速度快，而且数据维度和数据颗粒度也远超以前时代所能想象。在此情景下，以自动摘要和自动编码技术为核心的新一代数据库建设正在取代传统人工摘录和人工编码数据库，在这方面目前业已成型并被广泛使用的数据库如谷歌 GDELT（The Global Database of Events, Language, and Tone），[②]这是一个基于谷歌 Big Query、面向全球、免费开放的滚动型即时新闻事件数据库，由美国乔治城大学教授卡里夫·利塔鲁（Kalev Leetaru）于 2013 年创建，它不仅对新闻事件中的人物、组织、事件、语气等事件要素进行标签化提取，同时，还通过自动编码技术自动标注新闻事件的地理位置信息（即经纬度坐标），并且每 15 分钟实时更新一次。目前，该数据库已基本做到对某些国家政治新闻事件的即时监测、即时编码，其所提供的数据资料不仅包括时间序列数据，同时，还涵括地理空间信息，且每条数据都可核查、可验证，堪称真正意义上的“时空大数据”。

其二，面向特定问题解决的算法模型研发，目的在于将国际关系理论与计算机智能分析相结合改善决策质量。譬如，通过协同过滤算法（collaborative filtering）筛选恐怖嫌疑人、通过 K-Means 邻近算法进行特征聚类分析、通过 PageRank 算法进行网络链接分析，以及通过随机森林算法

① 战争相关因素数据库网址为 http://cow.la.psu.edu/；乌普萨拉武装冲突数据库网址为 https://ucdp.uu.se/；全球恐怖主义数据库网址为 https://www.start.umd.edu/gtd/about/。

② 谷歌 GDELT 数据库网址为 https://www.gdeltproject.org/。

(Random Forest)进行分类预测等。概括来说,基于算法的大数据国际关系应用重在规避数据噪音、挖掘数据关联,进而建立特征模式识别和进行分类预测。目前,大数据算法在国际关系中的应用主要集中在以下三个场景:第一,精准外交。通过抓取数据痕迹和聚类分析,精准圈定事件地域、事件人群及人群属性特征,定制化推送政治营销广告和实施精准公共外交战略;第二,冲突预防。通过数据监控和云计算,即时监测、锁定、跟进事态进展并自动生成事件报告和危机预警,动态掌控问题爆点,提前推进基于预测的预防性战略执行;第三,关联预测。通过多源数据收集和数据组合算法,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,优化决策、合理配置资源。目前,在国际关系研究中经常用到的算法模型主要分为有监督学习(supervised learning)和无监督学习(unsupervised learning)两种,其中,有监督学习最常用的训练方法主要有支持向量机、贝叶斯网络、决策树和马尔科夫链等,而无监督学习则主要包括聚类分析和模式挖掘,另外诸如主成分分析、多元线性回归以及信息熵等数值分析法也经常被用来测度数据关联关系。

二、大数据国际关系研究的基本原理

王:与传统研究路径相比,大数据正在冲击和改变现存国际关系研究的哪些方面?这些改变是否可以称得上是"革命性变革"?

董:作为一种新型数据、新型方法或新的思维方式,大数据确实给当前的国际关系研究带来不小的冲击,但是,截止目前,这些冲击尚未引起国际关系研究根本层面上的变化,大量的相关研究文献目前多以前瞻性、实验性探索为主。确切地讲,大数据对现存国际关系研究的影响才刚刚拉开帷幕,一些颠覆性的研究议程或研究结论尚未涌现,当下很难判断大数据的终极影响是否将是"革命性的"。仅就已有的研究议程和研究成果来看,改变主要发生在以下几个方面:

第一,数据类型、密度和颗粒度的变化。以前,国际关系研究主要以分

析低密度、小体量的结构化数据为主，数据颗粒度主要集中在以国家、政党和集团为统计单位，很少有研究文献会以政治行为体个人等微细数据为观察对象；在统计时序上，大多数国际关系数据库也多以年、季度和月为统计时段，以周和天为时段的即时追踪研究更是少之又少。即便是当下，为了印证或构建一种理论，研究者们通常会东拼西凑一些非连续性的碎片化数据来支撑论证。在这种情形下，数据通常是稀疏且大颗粒度的，很难具体到每一个微观主体在给定观察期内每一天的数据变化，研究者即使可以观察到问题的轮廓，也很难触及政治互动的内在细节，所得出的结论、所提出的对策因而也多是方向性和战略性的，多不能对现实问题的解决提供迅速且及时的纾困方案。而现在大数据及其分析技术的出现使得析取、观察和计算超大规模的混杂数据成为可能，即在结构化数据之外诸如传感信号、搜索痕迹和视频声像等非结构化数据都可以被一一纳入分析视野，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、影响决策结果。在高密度、连续性和细颗粒度数据支撑下，传统的国际关系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。

第二，国际关系研究前提假设的重设。传统上，主流国际关系分析多是以世界彼此分割、社会稀疏互动为前提假设的，社会信息传递并不像今天这么迅捷且及时，人际扰动性几乎可以忽略不计，国际政治现象的变动更像是各个政治行为体在互不干扰的情形下独立决策、理性选择的结果。然而，大数据的出现却正在挑战这一理性假说，它认为：未来世界是一个深度互联、全球一体的网络化社会，全球看似不相关的各个政治行为体拥有复杂的多重关联、彼此学习并相互影响，进而产生社会压力和政治规范；即使那些在以前政治观察中经常被忽略或被漠视的弱小行为体，在网络化社会也会因不经意的微小举动如随手拍、点赞或网络发帖而扰动整个系统的平衡和稳定，甚至重建新的社会结构和新的人际互动模式。譬如，突尼斯网络上一个不起眼的“城管打人”帖子居然会颠覆掉中东许多国家政权、韩国梨花女子大学一个普通的女生炫富事件居然会影响当下东亚政局走向，

国际政治中越来越多的“蝴蝶效应”事件正频频出现，这预示着我们所生活的世界正在加速走向一体化、系统联动性和高复杂不确定性，人们的决策越来越倾向于相互学习、相互影响和相互扰动。就此而言，肇始于人际稀疏关联和低频互动时代的现代政治分析，于当下恰恰忽略和低估了全球数以百亿计的各个政治微观主体之间的深广社会联系，更漠视了人类社会的复杂演进与不确定性，大数据时代的国际关系研究应该更多关注“微观主体之间的互动和相互扰动究竟是如何造就和再造宏观体系的”。①

第三，国际关系研究兴趣旨向的分化。围绕着国际关系研究应该优先重视“因果性”还是优先重视“相关性”，学界长期争论不已。当前，大数据的兴起无疑再度强化了这一争论。一种观点认为，国际关系学科的起源本就是为应对反复出现的战争与和平问题而设立，国际关系学本质上具有面向现实问题解决的价值追求，其终极目标应该是朝向社会工程应用而不是自我限定为一门战争诠释学，即该学科应通过技术手段来为现实政治问题的解决提供应对方案和战略指导，因而“相关性比因果性更重要”“建立在相关关系之上的预测分析是大数据的核心”“相关关系能够帮助我们更好地了解这个世界”，②譬如，国家反恐怖活动中的人脸识别技术和基因检测技术，它们本质上是一种相似性检测而无关因果性。当然，也有另外一种观点针锋相对，认为国际关系研究不宜偏离因果理论探索，一旦“放弃了对因果性的追求，就是放弃了人类凌驾计算机之上的智力优势，是人类自身的放纵和堕落”，“认为相关重于因果，是某些代表性的大数据分析手段（譬如机器学习）里面内禀的实用主义魅影，绝非大数据自身的诉求”。③ 此外，还有折中主义的观点主张，国际关系学科应是一门基础理论与技术应用并重的新型交叉学科，“相关关系是对因果派生关系的描述”“相关关系根植于因果性”，二者不是相互对立的。但不管持哪一种观点，可以肯定的是，在大数据时代，围绕着“相关性”和“因果性”的争论仍将继续，相关与因果、

① 董青岭：《大数据安全态势感知与冲突预测》，《中国社会科学》2018 年第 6 期，第 172—182 页。

② 〔英〕维克托·迈尔·舍恩伯格、肯尼斯·库克耶：《大数据时代》，盛杨燕、周涛译，杭州：浙江人民出版社 2013 年版。

③ 周涛：《译者序》，载〔英〕维克托·迈尔·舍恩伯格、肯尼斯·库克耶：《大数据时代》，第 IX 页。

工程应用与理论诠释或将成为国际关系学科未来发展的两大趋向。

第四，国际关系研究伦理规范的争议。放眼未来，大数据和机器学习技术的兴起无疑将成为国际关系研究新的潮流范式。通过对推特（Twitter）、谷歌、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，而且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，从而辅助政治科学决策和高效政治沟通。但也有观点认为，与其研究价值相比，大数据在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。首先，大数据国际关系研究难以回避的一个问题是数据的跨疆界流动，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，一旦大数据成为国际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“监控门”和“棱镜计划”。其次，以“杀手机器人”（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几亿人群中，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被枉杀，这为目前人类的道德准则和法律秩序所不容。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的可能，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，需要通过严肃的法律制度和研究规范加以约束。

王：鉴于上述担忧，目前学界围绕大数据国际关系分析是否存在争论？

董：与传统数据不同，大数据不仅体量大、产生速度快，而且噪音也大、获取有价值信息难。但即便如此，仍然有不少学者支持将大数据引入国际关系研究，当然反对声音也不容小觑，争议主要集中在以下三个方面：

第一，大数据究竟是不是一种新的研究方法？一种观点认为，大数据主要处理的是高密度混杂数据，颠覆了以前国际关系研究所赖以支撑的数据类型、数据维度与数据颗粒度，目标是是驱动国际关系研究朝向“数据密

集型科学研究”演变，大数据国际关系研究的重心是决策智能化或半智能化，因而基本可以看作是一种新的方法或研究范式。而另外一种声音则认为，大数据不过是传统统计方法的优化和升级，本质上仍属于统计学和计量学的范畴，至少到目前大数据尚未提出任何具有颠覆性的理论。就此而言，大数据未必会带来有关国际形势和外交战略的全新洞察，谈及诱导政治决策模式发生革命性变革更是为时尚早，因为即使存在着较以前更为丰富的大体量数据可供技术性挖掘，各国出于安全忧虑和隐私保护也会设置种种障碍阻止数据跨境流动，更加之数据体量越大、噪音越多，导致数据开发价值极低。

第二，大数据是不是体量越大、洞察力越强？目前，大数据在社会科学研究各领域炙手可热，但大数据真的能够比小数据提供更多的理论洞察与政治洞见吗？一种观点认为，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”或“大而无当”。譬如，针对脸谱的大数据分析是不是就代表了美国的主流民意呢？也许使用脸谱的核心用户是 18—40 岁之间的美国年轻人，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、几百或几千个用户是活跃用户，他们的政治表达是否又能够代表整个国家的民意呢？在此情形下，大数据版的民调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？此外，还有观点认为，大数据会说谎且更容易作假，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的话，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。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。

第三，大数据究竟是有助促进和平还是更易诱发战争？一些主张数据开放的学者认为，数据的价值在于开放、流动与不断被使用，如果通过数据分析可以挖掘更多决策信息、精准定位外交对象和客观预测竞争对手的战略意图，则基于大数据的研究发现将会极大改善外交决策质量、提高资源

配置效率并最终促进和扩大国家利益。然而，反对的观点也非常鲜明和尖锐，这部分学者认为，一旦大数据分析变成对外政策的工具，由机器人参与或操控的战争决策很可能会使全球武装冲突“常态化”，科学家、编程人员和游戏玩家等各色人员都有可能利用无人系统参与到武装冲突中来，机器杀人程序的“开源”传播和低廉制作成本，行将塑造人人都将拥有相互伤害能力的冲突进程。就此而言，“如果未来有一天，机器和计算完全接管了世界，那么，这种放弃就是末日之始”。[①] 正是由于上述认知分歧，欧陆学者多主张数据主权学说，并试图以此为理论支撑，主张建立更为严密的数据收集法案和数据治理体系，而美英学者则更多支持“数据自由流动说”和“数据开放论”。

王：我关注到您最近的研究兴趣主要是使用大数据进行冲突预测，您能简要介绍一下大数据预测武装冲突的主要原理吗？这种大数据预测和传统冲突预测有何本质区别？

董：好的。概括来讲，传统冲突预测主要是以“工具理性人”为前提假定、以“结构分析”为主要研究路径，重在因果分析和理论推演。冲突行动通常被认为是特定社会结构压力下，作为理性行为体的冲突各方理性博弈的结果。一方面，冲突中的各行为体理性且自私，每个冲突群体或个体都将冲突行动视为实现自我利益的工具手段，从自身利益最大化出发计算成本与收益，考虑利弊、权衡得失；另一方面，冲突行动主要不是微观主体之间突发性的情绪发泄或盲目的从众行为，而是基于特定社会条件、特定资源约束的审慎考量与理性选择。基于上述假设，传统冲突预测的目标主要聚焦于找寻那些有可能诱发冲突的结构性社会条件，并做出符合行为体利益最大化的理性推测。简单来说，传统冲突预测主要是以结构理性为主线来构建预测预测逻辑的，其主要适用于预测群体间冲突的中长期态势，但难以预测冲突的时空节点、也难以即时评估冲突的可能影响。

与之相对照，基于大数据的冲突预测则不以“工具理性人”为前提假

① 周涛：《译者序》，载〔英〕维克托·迈尔·舍恩伯格、肯尼斯·库克耶：《大数据时代》，第IX页。

设,也不以“结构主义”为分析路径。相反,它假定现代社会是以信息交换为主导特征的复杂巨系统,在这样一个复杂社会中,由于各个行为体之间是彼此关联、相互传递信息、相互扰动的,一切冲突现象的爆发、持续和终止都会对应着一系列信息,映射上的变化,通过观察这些作为冲突表征的信息映射上的关联性变化,基于大数据的冲突预测就可以感知冲突临近与否以及即将到来的冲突烈度如何。具体来讲,大数据冲突预测假设:在现代社会中,作为政治体系的基本构成单元,人是一种高度重视自我利益保护和规避风险的感性动物,且极易受人际关系网络中信息流动之影响。一般来说,如果不受突发事件的影响,人与人之间的互动频度与互动方式是相对稳定的,因此,人与人之间的信息传递内容、速度和方式也是相对稳定的,并由此决定了人的行为轨迹及其交际内容在日常实践状态下通常也是高度结构化可循的。但是,一旦某些数据在特定地区的大多数人群中突然发生同步异变,则很可能是该地区正在遭受经济危机、自然灾害、疾病传播、政治骚乱、武装冲突或恐怖袭击等异常事件之侵扰。

具体而言,大数据态势感知预测的操作逻辑非常接近自然科学中的地震预测、医学领域中的“并发症”研究,以及声学领域中的信号识别。具体到国际关系应用场景,当一个地区安全环境恶化时,作为微观主体的个人因身处危险最前沿会率先感受到威胁,继而将采取预防性规避措施,并将危险信息和切身感受沿社会网络传递到与之互动的其他个人和群体,由此可能导致越来越多的人改变日常行为。例如,当人们凭直觉感到骚乱或动荡临近时,商人会为规避损失而提前另谋出路、投资者会表现出种种抽逃资金迹象、旅行者会减少出游频次、留学生可能会提前回国、居民们会因相互传染恐慌而囤积生活用品,进而导致当地食品和医用品大幅涨价、物价指数全面飙升等。在现代信息分享机制的促动下,数以亿计的个体微观感知很容易汇集为有关冲突临近的整体性画面。研究者如果凭借大数据手段观测到多重数据信号的同步异变,就可以做出较传统因果性分析不一样的冲突预测。理论上,大数据分析观测到的同步异变特征向量越多,冲突预测结果越准确。

总体而言，相比于传统冲突预测研究所推崇的结构主义路径，大数据感知预测更加强调将国家想象为由数以亿计微观主体互动所构成的系统集合、将国际社会看作是跨越国界而又彼此关联的人际关系之网，冲突预测重在监测考察微观主体之间的互动对宏观结果的影响和塑造，其分析着力点在于捕捉网络化社会中微观主体之间的复杂关联与即时信息流动。

王：除冲突预测以外，您认为大数据在国际关系领域还有哪些值得关注的应用场景，能否简单介绍一下其操作原理？

董：除冲突预测以外，我认为大数据文本分析、情感倾向分析和社会网络分析都与国际关系研究有着较高的契合度，颇为值得关注，且已取得不少成果。

其一，文本分析，又称“意见挖掘”。其首要目标就是要解决非结构化数据的结构化问题，即利用自然语言处理技术（NPL）将看起来乱糟糟的“文本”转换为整齐划一的“数据”，在具体操作中常常会涉及词频分布、模式识别、关联分析、信息提取、可视化，以及预测分析等多个应用场景。通过文本的拆解和重组，非结构化数据的关键特征以向量的形式得以表现，并最终以词频矩阵的形式转化为可定量衡量的结构化数据，政治研究者可以通过深究被拆解后文本中词的分布和聚类，以及统计词和词的组合在文中的出现频率，来推断文本含义、理解文本提供者的政治意图。当前，文本分析最前沿的应用主要是聚焦网络舆情观测和民意挖掘，新闻网站和社交网络是其关注重点。

其二，情感分析，又被称为“情感倾向分析”。它主要是利用自然语言处理技术对带有情感色彩的主观性文本进行分析、处理、归纳和推理，其中，情感分析又细分为情感极性分析、情感程度分析和主客观分析等。例如，对于“喜爱”和“厌恶”这两个对立词的词频统计与概率分布观察，就属于典型的情感极性分析。程度分析主要是以量化指标的形式对情感极性进行细分度量，以描述该极性的强度，例如“喜爱”和“敬爱”都是褒义词，但是“敬爱”相对来说褒义的程度更加强烈一些；主客观分析则主要是指对文本中哪些部分是客观称述、哪些部分是带有情感的主观定性的一种统计

推断。

一般来说，情感分析主要有基于词典的分类和基于机器学习的分类两种操作原理。其中，基于词典的方法主要是将分析对象看作是词性标注，即通过制定一系列的情感词典和规则，对文本进行段落拆分、句法分析，根据情感词搜索数进行标注、赋予不同权值，进而按照句子、段落和篇章顺序的递进分类计算情感值，最后通过情感值来作为文本的情感倾向依据；而基于机器学习的方法大多将情感值计算转化为一个分类问题来看待，根据情感极性“正向极性”还是“负向极性”划分目标类别，进而对文本内容进行结构化处理，输入到给定分类算法（如朴素贝叶斯、随机森林和支持向量机等）中进行训练，并使用测试数据来预测分类结果。可以说，基于机器学习的方法是分类预测问题，而不是全样本统计问题。总体而言，短文本分析采用基于词典的情感分析效果更加好，而长文本则更加适合机器学习来处理。

其三，社会网络分析。社会网络分析以数学图论为基础，核心思想是把“关系”置于政治研究的中心位置，旨在计算茫茫人海中的人际关系网络的大小和人际距离的远近。它常常将数以亿计的政治行为体（个人或组织）抽象为社会的节点，而节点与节点之间的连接则被定义为人际关系的“边”，而信息和能量则主要是沿着“节点”和“边”循环流动并由此构成一个复杂社会关系网络。简单来说，社会网络分析主要是对社会网络的关系结构加以分析，它致力于研究：（1）政治体系中个体或关键组织的权力与声望，通过在定义网络中节点的度数（与一个人有关系的人数的多少）、介数（一个人在社会网络中的位置）和接近度（一个人与其他所有人的平均距离），来揭示社会系统中的权力集聚和政治威望；（2）政治系统的组织关系与结构变迁，特别是通过测量社会网络的平均距离（反映社会中信息传递的速度）、聚簇因数（反映社会关系的传递性）和密度（反映社会交往的频繁程度），来理解政治政治系统的“子群”分布和整个社会的政治稳定性。在某种意义上，社会网络分析所要分析的是由不同社会单位（个体、群体或社会）所构成的社会关系，而不是抽象的个体。通过研究网络关系，政治学家

们试图理解微观主体之间的互动究竟是如何建构并改变宏观结构的。

换言之,社会网络分析认为,人本质上是社会关系网络中的人,任何政治行为体都不是孤立的社会存在,而是附着于社会关系网络上接收信息、吸取资源和发挥权力效应,一旦离开了社会关系网络的支撑,不仅政治权力的投射将不可能,就连人本身的存在也将失去本体论意义。具体到现实中的政治关系,人类政治不过是各种关系的组合,国家、阶级、政党、利益集团乃至政治行为体个人都可作为网络中的节点,节点与节点之间的互动构成政治网络,每个政治行为体都是通过自身在网络中的位置来发挥影响的,越是处于中心位置的节点越是彰显权力和资源的控制能力,越是拥有较多的社会联系,政治的上不受其他权力控制的回旋空间就越大。就此而言,政治本质上是人际关系网络中信息的流动和资源的分配,权力是由关系所界定的,而不单单是由行为体的属性所决定的,越是接近社会关系的中心越是接近社会权力的中心。

三、大数据国际关系研究的前景

王:*您认为,当前大数据国际关系研究在发展中遇到哪些障碍?应当如何克服?*

董:作为一个新兴交叉领域或新的知识增长点,大数据国际关系研究在某些方面确实正在产生不同于以往的知识发现,但其研究议程推进也遇到不少麻烦,当前核心障碍主要有两个:

其一,数据障碍。首先,大数据国际关系研究的前提基础就是需要有数据,不仅数据体量要大、数据来源要广,而且数据类型要多样化、数据颗粒度要尽可能小。然而,目前无论是市场上还是高校科研院所均缺乏专业的国际关系与外交大数据平台,由此导致大数据研究的前端基础"数据"严重缺失。在此情形下,"巧妇难为无米之炊",即使是最为高明的数据算法也难以进行数据挖掘,因而,也就难以产生目前常规数据分析之外的知识创见;其次,大数据采集和整理是一个高难度且又枯燥冗长的体力劳动过

程，不仅需要大数据挖掘和应用算法方面的专业人才、技术参与，更需大量的人力、物力乃至资金投入，在目前数据采集成本高、数据分析技术非普及状态下，有限的文科科研经费很难产生真正意义上的大数据国际关系研究议程，但拥有雄厚资金支持的研究计划就另当别论；再次，精通国际关系同时又熟悉数据科学的跨学科交叉人才严重缺失。放眼环球，各国争相拟定数据人才培养计划并投入巨额资金，以期抢占大数据时代国际关系竞争的制高点，美国的核心国际关系院校如哈佛大学贝尔福中心、乔治城大学爱德蒙·沃尔什外交学院都已开设相关课程并设立相关学位，而我国大学和科研院所里的大数据国际关系专业人才的培养却尚未正式起步，目前已知对外经济贸易大学国际关系学院、清华大学国际关系研究院已在本科、硕士研究生培养方面开设了相关课程，并拟订了大数据相关人才培养和储备计划。

其二，制度障碍。不同于商业、娱乐和其他社会生活领域，国际关系领域的数据采集和存储将会受到严格的制度约束。首先，这一学科的数据采集关涉国家安全、国防机密和公民隐私等问题，虽然当前很多国家都提倡数据开放和政府开放，然而，即使数据是开源和开放的，一国外交机构对他国基础人口数据、基本经济数据和精细国防开支数据的收集会不会演变为两国间的情报间谍活动，在没有任何国际性条约约束的前提下也着实令人忧虑。进而，由此引发一个很严肃的数据安全问题：即使是出于善意竞争（指的是非恶意进攻、或胁迫或勒索）的需要，一国对他国进行数据收集和分析，哪些数据可以收集、哪些不可以收集？哪些数据国家间可以自由交换、哪些不可以？其次，任何精细化的数据收集最终都将触及个人隐私问题，对于外交决策和外交执行来说，数据收集的颗粒度越小，知识发现和政策洞察的可能性就越大。换言之，对一个国家人口数据的采集到省级的颗粒度显然不如到市县级的颗粒度洞察力强；对一个恐怖分子地理位置知悉相差数十里和数米的价值显然是不一样的。然而，数据采集通过什么方式、采集到什么程度？又用于何种目的才不算侵犯个人隐私，才能为各国隐私保护法案所允许？这将是一个严肃的国内法律问题，同时，这也是一个跨越主权疆界到的国际法问题，更是涉及科学技术是否可以无边界、无

约束使用的道德伦理问题。

王：大数据介入国际关系研究似乎已是大势所趋，并且伴随数据科学技术的普及，对这一研究议程感兴趣的国际关系学者也越来越多，您是如何看待大数据国际关系研究的前景？

董：的确，作为一种全新的数字化生存方式，大数据不仅正在改变着我们的生活，同时也在重塑我们观察和理解世界的方式。在现实政治运行中，数据力量已然开始解构我们社会的传统组织形式、人际互动方式和信息传递模式。曾几何时，数据公司和数据精英已然步入社会权力竞争的中心舞台，数据型权力正在崛起为新的权力表现形态；而与此同时，在政治科学研究中，数据分析技术的进步也已然开始冲击我们对国际关系和外交决策的传统认知，新的知识生产正卓然滋生于数据密集型科学研究议程之上。可以说，伴随着社会生活的网络化和数据化趋势的纵深发展，无论一个政治行为体身处全球哪个角落，人与人之间的相互扰动性正在显著增强，全球社会连带政治运行正在加速进入一个真正意义上的复杂巨系统时代，传统政治研究的属性分析正加速让位于数字时代的关系主义，新研究方法的引入将更多着力于社会工程应用而不是用于验证科学假设，数据科学与政治科学的结合时代虽尚未真正到来，但这一跨学科进程的帷幕已然开启。

放眼未来，大数据国际关系研究或将沿着以下路径进行拓展和扩延：其一，部分学者可能基于共同的抱负和情怀聚集为“数据学派”。确切地讲，界定为“国际关系研究领域内的数据库建设学派”或更为贴切，该派学者的目标非常清晰，主要是致力于国际关系研究底层数据的采集、清洗、整理和存储工作，目标是向国际关系研究者和外交决策者提供研究素材即数据公共产品，以便为智能时代的国际关系研究奠定坚实的数据基础，在这方面美国GDELT数据库的建设堪称典范，我国学者近期也有积极动作，譬如清华大学的孟天广团队致力于收集新闻报道的官员落马数据、试图建立一个开源腐败信息数据库，[①]再如，对外经贸大学与《国际安全研究》编辑

① 李莉、孟天广：《公众网络反腐败参与研究：以全国网络问政平台的大数据分析为例》，《中国行政管理》2019年第1期，第45—52页。

部联合团队，试图依托多线程定时爬虫和有监督学习半自动编码技术构建一个“国际安全大数据平台”，目标是向学界提供与国际安全研究相关的开源数据包，目前这一平台建设已略见雏形。①

其二，部分学者可能基于技术比拼、技术交流与合作凝聚为“算法学派”，即形成致力于开发算法模型，或使用机器学习技术研究大体量、高密度数据来提供国际关系洞察的技术共同体。该派学者的发展方向不是累积数据，而是改进和优化算法模型，目的通过技术研发和技术应用直接解决现实政治问题所面临的种种棘手难题。在这方面，较为典型的大数据应用如美国统计学者纳特·西尔弗（Nate Silver）所开发的“538”选举预测网站、亚历山大·尼克斯（Alexander Nix）所掌舵的剑桥分析公司、爱德华·斯诺登（Edward Snowden）所揭发的美国“棱镜计划”，以及美加澳等国情报机构所主导的“五眼联盟”，这些都属于较大也较为成功的国际关系大数据工程应用。而在我国学界如此规模或如此成功的大数据国际关系研究项目尚不多见，不过，已有不少学者开始关注大数据算法对国际关系研究的积极意义，相关成果譬如复旦大学唐世平团队基于复杂系统对选举预测方法的优化、清华大学庞珣团队使用社交网络分析对国家社会性权力的度量，以及对外经贸大学大数据国际关系研究中心（也就是我所在的团队）利用新闻摘录数据、神经网络和机器学习对武装冲突爆发时空点的概率性预测。此外，我和我的团队目前正在进行的另外一个项目也与大数据算法紧密相关，该项目名称是“基于自动摘要和自动编码技术的全球新闻情绪指数”，项目设计目标是通过大规模提取并度量滚动性新闻数据中的隐含社

① 目前，《国际安全研究》杂志已陆续发布了七类国际安全研究开源大数据，分别是联合国难民署：《国际安全研究开源大数据·全球难民统计（2009—2014 年）》，《国际安全研究》2016 年第 1 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·世界各国军费统计（2009—2014）》，《国际安全研究》2016 年第 2 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·世界环境保护开支统计（2009—2014 年）》，《国际安全研究》2016 年第 3 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·全球外交开支统计（2009—2016 年）》，《国际安全研究》2016 年第 5 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·国际安全态势感知指数（1995—2015 年）》，《国际安全研究》2016 年第 6 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·全球网络安全事件报告频次统计（2009—2016 年）》，《国际安全研究》2017 年第 2 期；对外经济贸易大学大数据国际关系研究中心：《国际安全研究开源大数据·世界各国文化交流频次统计（2009—2016 年）》，《国际安全研究》2017 年第 3 期。

会情绪来建立社会“晴雨表”，然后，借助贝叶斯定理和马尔可夫链来映射预测特定时空位置点的社会动荡与经济涨落。

其三，还有部分学者将集中研究国际关系中的数据使用伦理问题和人工智能挑战，或可冠之以“数据规范学派”。概括起来，目前有三大研究趋向成果颇为丰硕：第一，数据的跨境流动、数据安全与数据主权问题研究。一些学者提出，当前数字化的国际生存环境已经使得大数据成为国际关系竞争的新领域，并使得主权概念开始与地理要素脱离，数据主权正成为国家主权概念的新要素甚至是核心要素，数据安全正成为国家安全的新风险、新挑战。较有代表性的研究成果如复旦大学美国研究中心蔡翠红的《云时代的数据主权概念及其运用前景》、国防科学技术大学国际问题研究中心杜雁芸的《大数据时代国家数据主权问题研究》、复旦大学国际关系与公共事务学院沈逸的《网络时代的数据主权与国家安全：理解大数据背景下的全球网络空间安全新态势》等。[①] 第二，基于数据驱动的外交决策模式创新与外交形态演变研究。很多学者关注到，大数据及其分析技术的崛起正在挑战传统外交决策的方方面面，现代外交不仅要越来越疲于应付汹涌而来的数据洪流并努力加快政策响应速度，同时，在大数据时代，外交的思维方式也正在变得和以前传统思维迥然不同，基于数据驱动的智慧型外交或将发展成为未来外交形态。相关研究如原天津师范大学政治与行政学院王存刚和赵阳的《大数据与中国外交决策机制创新：基于组织决策理论的视角》、广州大学公共管理学院沈本秋的《大数据支持下的对外政策决策过程：优化与局限》、华东师范大学政治学系陆钢的《大数据时代下的外交决策研究》等。[②] 第三，人工智能时代的国际竞争、数据使用规范和机器伦

① 蔡翠红：《云时代的数据主权概念及其运用前景》，《现代国际关系》2013年第12期，第58—65页；杜雁芸：《大数据时代国家数据主权问题研究》，《国际观察》2016年第3期，第1—14页；沈逸：《网络时代的数据主权与国家安全：理解大数据背景下的全球网络空间安全新态势》，《中国信息安全》2015年第5期，第59—61页。

② 王存刚、赵阳：《大数据与中国外交决策机制创新：基于组织决策理论的视角》，《外交评论》2015年第4期，第1—18页；沈本秋：《大数据支持下的对外政策决策过程：优化与局限》，《国际论坛》2016年第5期，第32—37页；陆钢：《大数据时代下的外交决策研究》，《社会科学》2014年第7期，第3—15页。

理研究。伴随着数据的累积和深度学习算法的突破，一些学者认为，人工智能正在成为新的赋权技术并将最终改变全球权力结构，人工智能时代的到来将使人类进入一个变革且不平等的世界，数据的滥用和误用甚至有可能会挑起新的冲突和战争并伤及人类自身，因而，即便人工智能尚未真正成为现实，有关人工智能的规约研究也应未雨绸缪。遵循这一研究路径下的典型文献如国防科技大学文理学院刘杨钺的《全球安全治理视域下的自主武器军备控制》、上海国际问题研究院国际战略研究所封帅《人工智能时代的国际关系：走向变革且不平等的世界》、华东政法大学政治学研究院高奇琦《人工智能：驯服赛维坦》、清华大学国际战略与安全研究中心傅莹的《人工智能对国际关系的影响初析》、中国人民大学国际关系学院保健云的《大数据、人工智能与超级博弈论：新时代国际关系演变趋势分析》，以及暨南大学国际关系学院王悠和陈定定的《迈向进攻性现实主义世界？——人工智能时代的国际关系》等。①

总体而言，大数据和人工智能技术的崛起已是大势所趋。一方面，传统社会服务和公共决策过程正在被大规模数据使用和深度学习技术所重塑，人们将享受到越来越多智慧化服务所带来的社会生活便利，如无人驾驶汽车、无人快递机、精准医疗和加密货币等都已开始渗入到人们的社会生活；另一方面，伴随着数据挖掘和智能分析技术的进步，数据滥用和算法误用也正给国家安全、人们的隐私保护带来前所未有的困扰。数据和算法或将重新定义未来人与物（机器）的关系，一些具有自主决策能力的智能机器人或将取代人类决策的中心地位，届时人的尊严和价值、人的生存与安全都或将遭受前所未有之挑战。简言之，在数据技术应用之外，大数据国际关系规范研究也将成为一个新的学术拓展空间，并将产生丰硕成果。

① 刘杨钺：《全球安全治理视域下的自主武器军备控制》，《国际安全研究》2018 年第 2 期，第 49—71 页；封帅：《人工智能时代的国际关系：走向变革且不平等的世界》，《外交评论》2018 年第 1 期，第 128—156 页；高奇琦：《人工智能：驯服赛维坦》，上海交通大学出版社 2018 年 3 月版；傅莹：《人工智能对国际关系的影响初析》，《国际政治科学》2019 年第 1 期，第 1—18 页；保健云：《大数据、人工智能与超级博弈论：新时代国际关系演变趋势分析》，《国家治理》2019 年第 11 期，第 19—33 页；王悠、陈定定：《迈向进攻性现实主义世界？：人工智能时代的国际关系》，《当代世界》2018 年第 10 期，第 22—26 页。