迈向大数据法律研究

左卫民*

内容提要：大数据法律研究是实证法律研究的最新发展，将会带来法学研究范式的革命性变化。当前这项研究存在若干误识，如将“大量数据”“结构化数据”等同于大数据；在如何使用大数据展开研究方面，也存在方法的科学性不足等问题。未来的大数据法律研究不仅应思考如何更好地获取法律大数据，还要探讨如何正确认识与适当使用“大量数据”，更要充分利用统计方法展开大数据法律研究，探讨如何科学使用机器学习等新方式分析法律大数据。此外，继续重视对法律“小数据”的挖掘与运用，以及加强复合型研究人才的培养，也同样重要。

关键词：法律数据　大数据法律研究　实证法律研究

一、大数据法律研究时代的来临

受惠于互联网与大数据技术的迅猛发展，数据正以前所未有的速度巨量生成，海量的数据资源由此产生。大数据资源日渐成为国家与社会的基础性战略资源，推动世界大步迈向大数据时代。因应于此，法律机关、尤其是司法机关大力推进部门信息的电子化、数字化、公开化，使得法律大数据逐渐兴起并进入公众视野。

法律大数据的出现，使得基于法律大数据的司法实践与新型实证研究成为可能，并可能带来法学研究方式的革命性变化。这种可能性源于大数据所具有的独特优势：（1）数据的“全样本性”。大数据通常是特定领域的全面数据，具有数量巨大与内容全面之特性。基于全样本数据的实证研究，能够显著减少传统抽样方法可能导致的误差，增强对研究对象的整体把握，发现传统抽样数据中难以或根本无法获取的信息，带来研究视角、研究素材、研究方法的根本性转变。（2）数据产生、收集、分析的快捷性。“数据分析的速度越来越快，经常在数据刚刚敲进去的时候就可以看到实时的分析结果”，〔1〕这有助于研究者及时有

* 四川大学法学院教授。

感谢叶燕杰同学、郭松博士、王禄生博士、刘方权教授、洪凌啸同学、詹小平博士、张潋瀚博士、朱奎彬博士对本文初稿提出的修改意见。

〔1〕［美］伊恩·艾瑞斯：《大数据思维与决策》，宫相真译，人民邮电出版社2014年版，第12页。

效地掌握相关法律实践状况的全貌，从而克服传统实证研究方法耗时、滞后的缺陷。（3）数据收集与分析技术的客观性、科学性。

与具有亲历性的传统手工作坊式实证研究“大多是自己收集、整理数据”“存在因为研究动机需要而选择性收集、运用数据”不同，〔2〕海量材料与数据远非“人工作坊时代”研究者所能亲自、逐一地审阅、统计和分析。大数据的收集和分析往往直接依托于数据技术自动处理、完成。在开源条件下，研究过程具有相当的透明度，研究结论可复盘检验，数据收集、分析的客观性、科学性明显增强。〔3〕特别是，利用不同渠道收集的数据集产生了海量数据，当这些数据聚合到一起，可以对其进行挖掘，并开展更深层次的分析，该深度分析能揭示出各种模式、相关关系，并进行有统计意义的各种预测。〔4〕这不仅能够开展历时性与变迁性的研究，也能够进行预测性研究与趋势分析，〔5〕最终促进研究科学水准的提升。

在国外，法律大数据已广泛渗透到公权力与私权利领域的法律实践。在公权力领域，法律大数据在两个方面得到较多利用：一是在警务活动中。美国、澳大利亚等国家早已开始利用法律大数据开展警务预测。在美国，法律大数据被充分运用于犯罪趋势分析、发案情况预测、警力分配以及调查工作重心的确定等。〔6〕二是在审判活动中。法律大数据已大量应用于司法管理活动和程序性司法决策。例如，法官通过对法律大数据进行分析、评估，建立“何种情况下将影响嫌疑人到庭接受审判，何种情况下容易诱发新的犯罪”的保释风险预测模型，以此决定嫌疑人能否被保释；法官利用法律大数据对罪犯是否符合假释条件进行评估，以此作为判断罪犯能否被假释的重要参考。〔7〕在私权利领域，律师（律所）和当事人也高度重视对法律大数据的利用。例如，律师（律所）利用法律大数据进行律所管理、成本控制以及诉讼（律师）费用的评估、预测，〔8〕律师、当事人利用大数据挑选对自己有利的陪审团、〔9〕进行诉讼结果预测。〔10〕在大数据法律研究方面，国外学者除开始利用大数据对具体的法律问题展开研究外，对大数据法律研究与法律实践的理论与方法问题（例如，如何确保数据本身的可靠性、公开性，如何克服算法的非透明性、非归责性以及“数据歧视”，〔11〕

〔2〕 参见左卫民：《一场新的范式革命？——解读中国法律实证研究》，《清华法学》2017 年第 3 期，第 51 页。

〔3〕 参见刘佳奇：《论大数据时代法律实效研究范式之变革》，《湖北社会科学》2015 年第 7 期，第 143 页。

〔4〕 See Timothy J. Kraft, *Big Data Analytics, Rising Crime, and Forth Amendment Protections*, 2017 University of IIIinois Journal of Law, Technology & Policy 259 (2017).

〔5〕 See Andrew Guthrie Ferguson, *The Big Data Jury*, 91 Notre Dame Law Review 960 (2016).

〔6〕 See generally Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 Mich. St. L. Rev. 948 – 949 (2016).

〔7〕 See Lyria Bennett Moses, Janet Chan, *Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools*, 37 UNSW Law Journal 643 – 678 (2014).

〔8〕 See Jared D. Correia, Heidi Alexander, *Big Data, Big Problem: Are Small Law Firms Given a Sporting Chance to Access Big Data?* 37 W. New Eng. L. Rev. 144 (2014 – 2015).

〔9〕 前引〔5〕，Andrew Guthrie Ferguson 文，第 935 页以下。

〔10〕 当前美国的诉讼预测主要集中在知识产权诉讼和医疗过失诉讼领域，通过诉讼预测来决定是否诉讼与如何作出妥协。参见前引〔7〕，Lyria Bennett Moses 等文，第 644 页。

〔11〕 参见前引〔7〕，Lyria Bennett Moses 等文，第 643 页以下；前引〔4〕，Kraft 文，第 249 页以下；Kevin Miller, *Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm*, 19 Journal of Technology Law & Policy 105 – 146 (2014); Neil M. Richards, Jonathan H. King, *Big Data Ethics*, 49 Wake Forest L. Rev. 393 – 432 (2014).

大数据运用是否与美国联邦宪法第四修正案产生冲突及如何协调〔12〕）尤为关注。

目前，中国利用大数据开展的法律实践方兴未艾。例如：基于司法公开而大力推进的裁判文书上网工作；依托大数据技术建立犯罪信息判断和趋势预测；〔13〕运用大数据建设“检察大数据标准体系、应用体系、管理体系、科技支撑体系”；〔14〕利用大数据建立案件权重系数和评价指标体系，确定法官工作量，并进行科学的员额分配、案件分流；〔15〕基于大数据开展的多种法律人工智能实践，尝试如类案推荐、量刑辅助与偏离预警等应用。〔16〕其中，裁判文书大规模上网，使得中国第一次有了全国性、公开的、细节化的法律数据。但总体而言，目前国内对于法律大数据的实践性运用还相对有限，具体运用并不普遍，在一定程度上呈现出“话语热、实践冷”的现象：一方面，应用主体范围有限，主要集中在少数司法机关、法律数据公司；另一方面，应用领域相对较窄、实际运用较少，主要集中在类案检索、法律文书草拟、文书智能纠错等辅助办案方面。

近年来，国内也出现直接利用大量数据展开法学研究的探索，并已经注意到法律大数据所面临的伦理规范等问题。〔17〕其中，有学者就如何开展大数据法律研究，提出了有启发性的见解。〔18〕不过，国内的大数据法律研究整体上还处于探索阶段，一些研究缺乏对法律大数据的基本认识，研究方法和过程其实建立在某些误识上。因此，检视大数据法律研究现状，澄清若干误识，对于大数据法律研究的健康开展具有基础性意义。

二、大数据法律研究中基本问题的澄清

（一）大数据还是大量数据

大数据具备“4V”（Volume、Velocity、Variety、Value）特征，是关于某一领域（行业）全样本、能够快速流转、多样化且富价值的数据。其中，“全样本”是其最显著的特征，“全样本数据”意指相关的所有数据。然而，目前国内的法律大数据基本上只是部分的、非

〔12〕 See Sohayla M. Roudsari, *Fourth Amendment Jurisprudence in the Age of Big Data: A Fresh Look at the "Penumbers" through the Lens of Justice Sotomayor's Concurrence in United States v. Jones*, 9 Federal Courts Law Review 139 – 174 (2016); Ismail Cem Kuru, *Your Hard Drive is Almost Full: How Much Data Can the Fourth Amendment Hold*, 2016 U. of III. J. L. Tech. & Pol'y 89 – 134 (2016).

〔13〕 参见卢国强：《北京警方利用大数据预测犯罪趋势》，《科技日报》2014 年 6 月 18 日第 3 版。

〔14〕 参见曹建明：《最高人民检察院关于人民检察院全面深化司法改革情况的报告——2017 年 11 月 1 日在第十二届全国人民代表大会常务委员会第三十次会议上》，《检察日报》2017 年 11 月 2 日第 2 版。

〔15〕 参见周强：《最高人民法院关于人民法院全面深化司法改革情况的报告——2017 年 11 月 1 日在第十二届全国人民代表大会常务委员会第三十次会议上》，《人民法院报》2017 年 11 月 2 日第 1 版。

〔16〕 王禄生：《司法大数据与人工智能开发的技术障碍》，《中国法律评论》2018 年第 2 期，第 46 页以下。

〔17〕 参见徐明：《大数据时代的隐私危机及其侵权法应对》，《中国法学》2017 年第 1 期，第 130 页以下；顾理平：《大数据时代公民隐私数据的收集与处置》，《中州学刊》2017 年第 9 期，第 161 页以下；等等。

〔18〕 例如，白建军讨论了大数据时代利用大数据进行裁判预测的可能和限度问题，大数据时代如何科学取样的问题（参见白建军：《法律大数据时代裁判预测的可能与限度》，《探索与争鸣》2017 年第 10 期，第 95 页以下；白建军：《大数据对法学研究的些许影响》，《中外法学》2015 年第 1 期，第 29 页以下）；胡凌探讨了大数据时代“法学研究方法的深化”（参见胡凌：《大数据兴起对法律实践与理论研究的影响》，《新疆师范大学学报（哲学社会科学版）》2015 年第 4 期，第 108 页以下）；张吉豫研究了大数据时代法学研究如何“开展交叉学科研究和应用”（参见张吉豫：《大数据时代中国司法面临的主要挑战与机遇——兼论大数据时代司法对法学研究及人才培养的需求》，《法制与社会发展》2016 年第 6 期，第 52 页以下）。

完整的数据，远非“相关的所有数据”，称其为“大量数据”或更合适。基于这些大量数据展开的研究，似乎很难视为严格意义上的大数据法律研究。

从某种意义上讲，中国的法律大数据肇始于裁判文书统一集中上网；在裁判文书上网之前，中国并没有法律大数据研究，法律实证研究基本上是基于“小数据”，即研究者自己在局部范围或特定领域所收集的数据，而展开的“手工作坊式”研究。裁判文书网的诞生与发展，使得丰富的全国性数据第一次制度性涌现，其与既有实证研究所使用的数据在数量级、广泛性上大不相同。然而，裁判文书网已经公布的裁判文书数据整体上并不完全具备全样本特征：公布文书数量与实际结案数量相差较大，数据缺失问题相当严重。根据全国法院 2014 年和 2015 年的裁判文书上网统计显示：按省份看，上网裁判文书占实际结案文书比重最高的达 78.14%（陕西），最低的仅为 15.17%（西藏）；最高人民法院在这两年的上网裁判文书仅占其实结案件量的 46.13%，这一比重与全国的总体情况大体持平。〔19〕截至 2017 年 7 月 11 日，四川省的法院在 2012－2016 年间的裁判文书上网 1134249 份，而根据四川省高级人民法院工作报告，2012－2016 年全省共审结案件 3865125 件，〔20〕上网量不足审结量的 1/3。此外，上网裁判文书所涉及的案件类型并不全面，特别是一些重大职务犯罪类案件，其裁判文书往往并不上网。

概括起来，刑事案件的公开比率优于民事案件，一般刑事案件的公开比率优于敏感刑事案件。裁判文书上网的数量、地域、案件类型等方面的局限，使得相关数据往往并非全数据，远离标准的大数据，这容易导致一些基于裁判文书的实证研究存在支撑证据不足，甚至观点可能错误的问题。此外，部分地区法院在公开裁判文书时还对文书内容进行了删减，其删减往往并非对当事人身份信息的屏蔽处理，而是对文书特定段落的删除。这也会使得某些依靠从裁判文书网获取的文书对特定问题的分析，存在不同程度的数据偏差。因此，尽管特定领域、特定区域的分类数据可能较为齐全，但从整体上看中国当下的法律大数据，虽然数据量可能较多，许多领域均可能有 20－70% 左右的全国性或全局性数据，但其实仍多是大量数据。

如何认识大量数据的学术研究价值？一方面，完美的法律大数据往往难以强求。作为官方化的数据，公开与不公开往往并存，法律、政治、传统的各种因素都会影响法律和司法数据的公开程度。欧洲国家地方法院裁判文书的公开度往往不如中国，美国法院刑事审判中同样少有关于裁判心证的公开信息。无论中外，法律数据都均非丰富、完整，难以完全反映法律和司法实践。由此，有缺失的大量数据往往可能是“现实中的大数据”。另一方面，大量数据不仅在数据量、丰富性方面远超小数据，而且经过清洗后可以具有相当的全局代表性。在求全不得的条件下，如果能够正确清洗数据，正确把握数据缺失的程度、特别是有无系统性缺失，大量数据就具有不可替代的学术研究价值。

（二）法律数据的官方性、结构化

相比于商业、社会领域的大数据，法律大数据具有自身的独特性：商业、社会领域的大数据往往是非官方的机构收集并使用的，而法律领域大数据则具有“官方化”的特征；

〔19〕 参见马超等：《大数据分析：中国司法裁判文书上网公开报告》，《中国法律评论》2016 年第 4 期，第 208 页。

〔20〕 根据四川省高级人民法院的工作报告，四川法院 2012－2016 年间年审执结案件总量分别为 685300 件、738857 件、750254 件、821285 件、869429 件，五年合计 3865125 件。

这种差异深刻影响数据的生成和使用。官方化特征不仅使得法律数据的公开程度受到影响，也影响到法律数据的内容、类型及格式。基于法律机关的政策考虑，相关法律数据的内容多表现出格式化、预设性与法律化特征，据此向社会公开的法律数据其实是按照司法机关的管理目标所生产的内容，而非公众所欲知晓的有关法律实践的充分、真实数据。这与商业、社会领域的大数据颇不相同，后者常常是更为自然的非结构性数据。

比较典型的结构化数据，主要是来源于司法机关工作报告与法律统计年鉴的数据。此类数据都经过“精细加工”，数据发布主体自身的价值偏好也潜藏其中。目前，“公开的司法统计数据不完整，许多应当公开的数据并未公开，公开比例也难以令人满意”，〔21〕诸如刑事案件律师辩护率、民事案件律师代理率等数据难以获得；数据的统计口径往往也不一致，甚至同一主题在不同年份的统计口径也会出现变化，以致数据的连贯性较差。这些结构化或半结构化特征明显的大量数据，对司法管理具有一定的参考意义，也有相当的研究资料价值，但由于其生产目的的特定性，整体上并不充分和全面，尤其是中观、微观层面数据的缺失，使得它并不完全具备大数据的特征。对于此类数据，或许视作“重要和宏观的司法数据”更恰当。〔22〕而裁判文书的结构性则要弱一些，或可称为半结构化的数据。裁判文书的事实认定与法律适用的表述思路和风格，是由众多风格各异的法律实践者个人或集体完成的，但其基本写作逻辑和格式仍然受到制度与实践层面的严格规范，大体上还是半结构化的。

真正丰富的法律大数据应兼具大数据的自然特征与法律特征，主要由各种法律主体参与生产、制作并发布，具有全样本、即时性、多样化特征。现阶段中国法律大数据整体上是以裁判文书网为主要来源的官方化、结构化或半结构化的大量数据，实质上只是法律领域中的有限数据，也是角度特定的数据。

（三）数据在研究上的应用：方法和目的

作为实证研究的一种新形式，大数据法律研究应当遵从实证研究的一般范式，即利用大数据分析、发现经验现象，并基于经验现象提出、证实或证伪假设，最终发展和创新理论。同时，大数据与小数据的分析方式在研究模式方面有着共性：都应用数理统计的一般规律，采用统计学的许多方法，尤其是回归分析。当然，实践中“大小数据”研究的界限时常有所模糊。一些小数据并不小，特别是一些区域性数据研究涉及的样本可能高达十几万甚至几十万个，其研究方式可能与大数据研究并无二致，甚至有的小数据研究已经在使用复杂的机器学习。〔23〕

尽管如此，大数据法律研究有其独特性，与小数据研究存在诸多不同：（1）研究者的亲历性不同。由于小数据的有限性，研究者一般亲自、逐一收集、审阅和分析每一个研究样本，具有很强的亲历性。然而，面对全国性的裁判文书或者某个领域的裁判文书时，研

〔21〕 易霏霏等：《我国司法统计数据的公开：现状与建议》，《中国应用法学》2017 年第 2 期，第 68 页。

〔22〕 参见倪寿明：《充分挖掘司法大数据的超凡价值》，《人民法院报》2017 年 6 月 23 日第 2 版。

〔23〕 例如，有学者研究了新奥尔良地检署包含 145000 个被告人的 280000 起案件，通过机器学习的方法建立了被告人的再犯可能性模型。该模型可以降低 5－9% 的再犯危险，并有效区别出人类检察官所蕴藏的主观因素。See Daniel L. Chen (TSE－IAST), *Algorithms as Prosecutors: Lowering Rearrest Rates Without Disparate Impacts and Identifying Defendant Characteristics 'Noisy' to Human Decision-Makers*, The 11th Digital Economics Conference, Toulouse, January 11, 2018.

究者便无力如此操作了。对于此类研究，如果没有好的数据收集、分析方式与技术，研究根本不可能有效地开展。因此，小数据研究中的判断一般是亲历性、实感化的判断，大数据研究中的判断往往依赖计算机软件，是一种间接性的判断，实感性较弱。（2）数据量的差异使得大数据研究更依赖诸如机器学习等新方式。面对海量数据，应用计算机软件和机器学习在所难免。巨大的数据量使得精细梳理变量间关系的研究受到挑战：大数据本身既可能粗糙，也可能信息过载，干扰因素与各种相关变量较多，研究者往往难以有效把握。这或许也是很多实证研究者依然致力于小数据研究的重要原因。

就当前的研究现状来看，虽然中国的大数据法律研究已经开始使用爬虫软件等抓取数据，但内容分析仍以描述性的数据分析为主，很少有研究者能够使用统计软件与统计学分析方法对数据资料进行精确的定量分析。〔24〕对于如何整理与分析大数据，法学研究者大多“还不能科学、熟练地运用数理统计等分析手段与方法对问题展开统计学意义上的定量分析，更遑论在研究中进行数理模型的建构，从而在定量研究的方法上与统计学、社会学、经济学等其他学科展开对话”。〔25〕如果不得不采取数据科学方法，研究者往往也只能依靠统计学家和数据科学家进行数据收集、挖掘、统计与分析。但技术专家经常不能把握法学研究者的真正意图，对基本法律问题也缺乏相应判断，这无疑增加法学研究人员与统计学家、数据科学家之间的沟通成本。或许不得不承认，当前“对大数据的收集、研究和应用还处在一个比较粗浅的层面上，司法大数据可能具有的超凡价值远远没有得到挖掘”。〔26〕

对于经验性法律现象，如律师辩护率、刑民事案件二审的改判率等，基于法律大数据的描述性分析可能是适当的。然而，法律实证研究毕竟是一种可量化的社会科学研究，需要归纳出法律运作过程的规律，并对其背后的因果关系进行深度阐释，或至少指出需进一步探究的相关性。一旦需要进行更多的因果关系或相关性研究，描述性分析则明显力有不逮。例如，通过大数据来分析家庭经济收入、父母受教育程度、父母情感关系、同辈朋友中的犯罪情况、未成年人的学习情况等，是否对未成年人犯罪具有直接影响以及影响的强弱时，传统的描述性统计分析可能就难以胜任。更加深入的法律大数据研究，还涉及机器学习与算法应用，尤其在对法律大数据进行应用研究时更是如此。例如，对于通过数据关联分析在大量散乱的数据中如何发现数据之间的相关性，并将这些数据形成一个数据集，从而描绘出某个事物或事件的发展规律或趋势，传统的统计学方法往往力不从心，需要通过机器学习实现研究目标。Jon Kleinberg 等人利用决策树、迭代算法等机器学习算法，分析了美国 15 万余件重罪案件的法官假释决定，认为机器学习算法的预测要优于人类法官的判断。〔27〕

实际上，大数据法律研究是一项综合性、系统性工程，研究者掌握与运用相关研究方法的能力在很大程度上决定了研究的深度与层次。法律大数据研究的核心在于对海量数据的价值挖掘、处理，这就涉及上述数据的获取、清洗与使用。以典型的裁判文书大数据分析为例，由于目前上载的裁判文书达到 4000 万以上的量级，传统人工下载的方式远远无法

〔24〕 参见前引〔2〕，左卫民文，第 51 页。

〔25〕 同上。

〔26〕 参见前引〔22〕，倪寿明文。

〔27〕 机器学习在法学研究中的展开与运用，See Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan, *Human Decisions and Machine Predictions* (February 2017). 113 (1) The Quovterly Tournal of Economics 237 – 293 (2018).

满足研究的需要。这就必须借助爬虫软件自动从互联网上下载海量文书。然而，文书的获取只是大数据研究的第一步。由于爬虫软件抓取的文书是典型的无标签非结构化数据，其中包含重复文书、空白文书等“脏数据”，此时就必须借助数据清洗手段处理这些文书，添加案号、案由、审级等常规标签。在数据清洗的基础之上，才可能进行数据挖掘。由于人工统计无法完成数据挖掘的任务，因而需要运用正则表达式等数据挖掘方法。可见，大数据本身为法学实证研究设定了先天的技术门槛。

舍恩伯格等认为，“大数据时代绝对不是一个理论消亡的时代，相反地，理论贯穿于大数据分析的方方面面”，“大数据不会叫嚣‘理论已死’”，反而会“从根本上改变我们理解世界的方式”。〔28〕目前有关大数据的法律研究，在研究取向上偏重于实践型、应用型，而非学理性、抽象性，侧重数据的调查与描述，过度沉迷于让数据“自己说话”甚至“自己思考”，疏于开展深度的理论剖析与建构。很多冠以“大数据”的实证研究不过是运用大数据或大量数据对某个法律现象或问题的简单描述，各种法律数据的简单归类统计，以及在此基础上提出问题与解决对策。对大数据所呈现的普遍现象进行深度剖析与理论解读的研究还较为缺乏，更遑论相关理论建构。

（四）作为方法的大数据法律研究

基于大数据的法律研究对法律研究方法到底意味着什么？这是否一场新的研究范式革命？法学界目前更多只是将之看作一种实践现象。大数据法律研究的一些基本理论问题，如它的内涵、特征、优势与局限，大数据法律研究与社科法学、实证法律研究之关联，如何适当运用、科学展开等，至今尚未得到充分讨论。如果说基于小数据法律实证研究的理论图景已日渐清晰，那么基于大数据法律实证研究的理论问题似乎未昭未揭。这可能会使研究者陷入“过分关注技术分析，忽视创新思维和思辨分析”〔29〕的窠臼中。有论者在谈到大数据对社会学研究的影响时指出，“‘大数据’概念的广泛应用和巨大影响，对社会学研究的冲击更为直接。这种冲击涉及数据来源、研究方法、社会测量等诸多重要领域”。〔30〕事实上，这种冲击和影响甚至已经开始波及法学研究。从研究对象看，大数据法律研究扩展了法学研究的问题域，使法学研究不再拘泥于传统的研究对象和素材，从而拓展了法学研究的领域和格局。从研究范式看，大数据法律研究可能推动实证研究的跨越式发展，特别是机器学习方式的引入，会使法学研究从法教义学、社科法学和实证法律研究等范式转向数据科学式的法学研究，形成“数据驱动 + 理论假设驱动”的范式革命，最终重构传统法律实证研究。〔31〕就此而言，或许可以将其视为法律实证研究的新阶段。

大数据法律研究应当具有什么样的问题意识？当前，一些大数据法律研究缺乏必要的问题意识，主要是描述式研究，沦为“调查报告式”的数据展示。针对法律实证研究，曾有论者提出“受众是谁”的问题，〔32〕大数据法律研究同样应重视此问题。从某种程度上讲，“受众”不仅是指知识生产所面向的市场，也意味着知识生产者与消费者之间的互动。

〔28〕［英］维克托·迈尔—舍恩伯格、肯尼思·库克耶：《大数据时代》，盛杨燕、周涛译，浙江人民出版社 2013 年版，第 94 页。

〔29〕参见孙建军：《大数据时代人文社会科学如何发展》，《光明日报》2014 年 7 月 7 日第 11 版。

〔30〕参见孙秀林、陈华珊：《互联网与社会学定量研究》，《中国社会科学》2016 年第 7 期，第 119 页。

〔31〕参见刘涛雄、尹德才：《大数据时代与社会科学研究范式变革》，《理论探索》2017 年第 6 期，第 29 页。

〔32〕参见唐应茂：《法律实证研究的受众问题》，《法学》2013 年第 4 期，第 28 页。

一方面，大数据法律研究应该面向司法实践与司法改革，从司法机关与公众真正关心、急欲解决的现实问题中寻找研究灵感，从而使大数据法律研究具备较强的实践性。例如，最高人民法院和最高人民检察院正围绕“智慧法院”“智慧检察”，深入推进法院与检察院办案、办公的阳光化、网络化、智能化，人工智能开始运用于司法改革推进和司法建设之中，而司法领域人工智能的运用显然无法离开法律大数据的支撑。因此，大数据法律研究应该回应“智慧司法”“智慧检察”的实践需要，并藉此获得更大的致用空间，在理论与实践的互动过程中形成问题意识，推进大数据法律研究的发展。另一方面，大数据法律研究也应该直面一些传统的法学理论命题，借助法律大数据分析工具对其进行检验或创新发展。当然，基于大数据发现新的法律现象、提出新的理论命题，并利用大数据分析技术进行论证，亦是大数据法律研究的应有之义。

对大数据法律研究规范与伦理问题的讨论亦很重要。大数据法律研究涉及海量数据的收集、整理和分析，这对数据收集、分析主体的专业水平，特别是研究规范提出了较高甚至是苛刻的要求。如果大数据研究人员的专业水平有限，对研究规范不够重视，很有可能导致收集的数据失真、分析结果失准，不仅不能对法律现象进行客观量化，甚至可能形成错误结论，以致谬以千里。在数据本身有限且运用相对简单的条件下，其他主体可以对研究结论进行重复性或经验式验证，但在大数据背景下，这种验证无疑困难得多，成本也更为高昂。特别是，在当前数据公司与研究者普遍不愿公布数据来源、内容、收集工具、分析标准的情况下，如果研究者不能对大数据法律研究抱持严谨态度并恪守必要的研究规范，甚或不遵守相关学术伦理，基于功利动机而突破学术底线，将导致相应的大数据法律研究存在研究标准不科学、研究结论荒谬甚至数据造假等问题。此外，大数据法律研究表面上是围绕数据展开，但由于很多数据其实指涉或反映主体的行为、意识与习惯等，这使得大数据法律研究不可避免地牵涉“人”的问题，甚至可能会将作为研究对象的“人”置于相对危险的境地。这样的伦理问题目前似乎并未引起研究者的注意，对此更应有意识地采取相应的技术处理策略。

三、迈向大数据法律研究

（一）致力于获取全面、多样的法律数据

第一，尽力获取全面化的法律数据。数据的官方化，是法律大数据不同于商业性、社会性数据的特殊之处，这就决定了法律大数据的获取与应用水平在相当程度上取决于法律机构是否充分、及时公开其收集、掌握的法律信息。所以，法律机构基于共享理念推行数据公开机制是法律大数据获取和应用的关键之一。前已述及，裁判文书网所收集、整理、统计的数据往往并不能称作“法律大数据”，甚至有时数据质量可能还不如抽样调查中的“代表性数据”。虽然 2016 年最高人民法院修订了《关于人民法院在互联网公布裁判文书的规定》（法释［2016］19 号），但由于其约束力不强、操作性较弱，事实上并未实现“（除不予公布的）其他裁判文书一律在互联网公布”〔33〕的目标。为了提高司法的公开水平，促

〔33〕 参见最高人民法院编：《中国法院的司法改革：2013－2016》，人民法院出版社 2017 年版，第 38 页。

进法学实证研究的发展，法学界需要呼吁最高人民法院进一步健全裁判文书发布的责任机制，加强对裁判文书不上网的审查力度，大力推动并真正实现裁判文书网络发布的“应上尽上”原则，〔34〕促进裁判文书网不断由大量数据平台向大数据平台转变。

第二，努力扩展法律数据的来源。数据是大数据法律研究展开的基础，“只有具备足够的数据源才可以挖掘出数据背后的价值”。〔35〕然而，在法律大数据的来源上，目前过度依赖官方尤其是法院的主动发布，内容、渠道存在单一性、有限性等问题。中国法治和中国司法的整体样貌不可能据此充分展现，司法决策信息更不是裁判文书所能充分显示的。当前，除了对外公布的法律裁判文书，法律决策过程中的关键行为，如形成决策的内部讨论，往往是高度非文字化、非数据化的。“一方面，‘庭审笔录不是一种公开的法律证明文书’，其亦未实现充分的数据化；另一方面，大量的程序过程如警察的侦查过程、检察院的起诉过程、法院庭审前后的过程也没有公开的、正式或非正式的文字记录，更遑论在此基础之上的数据化了。”〔36〕为此，首先要拓宽法律大数据的领域。只有将检察机关、公安机关、司法行政机关所收集和制作的、符合公开条件的数据全面纳入公开范围，才可能“推动形成有利于平台互联互通、信息共享共用、业务衔接联动的体制机制”，〔37〕进而实现公、检、法、司的相关数据接驳、联通共享，提高法律大数据的集成化水平。其次，要丰富法律大数据的类别与内容。其他类型的诉讼文书（或材料），如庭前会议笔录、庭审笔录、案卷材料等，尚未成为法律大数据的来源。检察机关虽然公布了部分诉讼文书，但存在数量少、不全面以及可获得性较差等问题。一些相当重要的司法数据，如检察机关的批捕、公诉与抗诉等数据，并未充分公开。随着语音识别、文本抓取等人工智能技术在实践中的运用得到普及，对更多诉讼活动进行电子记录、数据提取，将一些“僵尸数据”转化为可计量、可使用的统计数据已成为可能。因此，未来应将视野拓展到裁判文书和司法统计之外的信息，更加重视对起诉书、庭审笔录等记录诉讼活动与程序的结构化数据、非结构化数据的收集和整理，尤要思考如何将实践中大量的非结构化数据、半结构化数据转化为有价值、可运用的结构化数据，确保大数据的有效性、有用性。另外，电子卷宗的推广、证据标准判断的数据化，也为更多地使用裁判文书以外的其他数据来源提供了重要机遇。这有助于我们获得裁判文书网以外的丰富材料，进而助力开展更为多元的研究。只有当法官乃至所有法律行动者的行为模式与决策信息充分数据化时，法律大数据才能真正被称为“大数据”。

第三，重视和利用好当下的大量数据，包括区域性的全样本数据。受制于各种客观条件，大量数据而非大数据可能是研究者在很长一段时间内所面临的窘境。但大量数据也是法学研究的重要材料，值得高度重视与充分利用。为此，一方面，要避免数据样本带来的数据偏误，特别是系统性偏差。了解现有数据公布的偏差情况，是利用好已有的数据材料，

〔34〕 最高人民法院最新发布的《最高人民法院司法责任制实施意见（试行）》中，再次重申“裁判文书送达后7个工作日内，承办法官应当督促指导法官助理或书记员完成拟公开裁判文书的技术处理和裁判文书上网公开工作”。裁判文书的上网公开有望更具规范性。

〔35〕 刘鹏主编：《大数据》，电子工业出版社2017年版，第4页。

〔36〕 关于法律数据尤其是法律大数据如何在法律人工智能中进行运用，可以参见左卫民：《关于法律人工智能在中国运用前景的若干思考》，《清华法学》2018年第2期，第108页以下。

〔37〕 参见孟建柱：《主动拥抱新一轮科技革命，全面深化司法体制改革，努力创造更高水平的社会主义司法文明》，《贵州日报》2017年7月12日第1版。

尤其是裁判文书网所公布的裁判文书的前提所在。例如，由于不同案件公开的比例差异，在进行裁判文书的数据挖掘时，刑事类的分析所包含的数据偏误就天然小于民事类；一般刑事案件的分析就优于贪腐类犯罪；离婚纠纷由于大量采用调解的方式结案，而调解文书通常不予公开，这就决定了有关婚姻类的大数据挖掘报告需要谨慎对待。基于数据本身的局限性，在利用裁判文书网进行研究时，可以适当缩小研究范围，并限定研究对象，确保在有限的数据条件下尽可能地收集、获取某领域或某类别相对完整、具有一定代表性的真实数据。此外，还可运用诸如“贝叶斯方法”和“大数定律”等数理统计方法对现有数据进行推断，从而正确识别并验证数据的代表性。另一方面，重视区域性的全样本法律大数据。我国疆域辽阔，不同地域之间的人文、地理环境差异巨大，收集全国范围内的全样本（或近似于全样本）数据无疑具有相当难度，如果转而收集若干具有代表性的区域性全样本数据，则可以提高数据收集的成功率。

（二）探索并深入展开大数据法律研究的科学方式

第一，探索新型、专门的大数据获取、分析技术，并充分运用于大数据法律研究。“基于大数据技术而获取的数据，已经不同于社会科学研究中普遍使用的随机数据”，因而，“在统计推断等方面需要因应调整”。〔38〕在大数据挖掘、整理、分析方面，目前已经有较为成熟的统计方式和数据科学方式，而与统计学相关但又颇为不同的机器学习方法也已崛起并运用于大数据分析之中。如何甄别大数据的有效性、真实性，如何分析、判断数据之间的相关性与因果关系，还应有更多的方法与技术。在目前的大数据法律研究领域，数据挖掘依然主要通过正则表达式的方式。该方式在处理高度规整的文书表达时具有很强的准确性，如从海量文书中自动提取辩护人的数量、身份等表述高度一致的数据。然而，正则表达式在面临高度多元化的表达时，由于无法穷尽表述，就多少显得力不从心。例如，“自首”也许在文书中并不会以“自首”的关键词出现，而是以“家属扭送”等诸多样态的语词呈现，此时就需要用自然语义识别技术（NLP）。这类技术在法学领域才刚刚起步，主要出现在大数据与人工智能的司法实践中，还较少被应用于法律大数据的研究中。

专门的数据分析机构具有得天独厚的技术与人才优势，法律研究者和司法部门必须思考如何更好地借助专门数据分析机构和人工智能科技公司的优势，充分挖掘、分析与利用数据。中国电子信息产业发展研究院在 2017 年发布的《中国大数据产业发展水平评估报告》中指出，“我国大数据产业发展将迎来‘黄金期’，产业聚集将进一步特色化发展，技术融合创新将更加深入”。〔39〕法律研究者也应搭乘大数据发展的“快车”，充分发挥专门的大数据获取、分析技术的作用。这些技术往往既非传统法律实证研究的方法，也不全是当下分析小数据所运用的统计方法，而是数学与计算机内容交叉、不断发展进化的、以机器学习为主的新型方法。当然，研究者也需要注重对技术的深度学习与直接使用，努力做到自己掌握、使用现有技术工具进行数据收集、挖掘与分析。

第二，充分利用数据进行深度分析。简单的描述性统计分析方法在面对海量数据时显得力不从心，特别是当大数据获取的信息本身就“漫无边际”“支离破碎”而“根本不可能

〔38〕 马亮：《实证公共管理研究日趋量化：因应与调适》，《学海》2017 年第 5 期，第 199 页。

〔39〕《首个年度大数据产业评估报告发布，将为我国大数据产业健康发展提供有力支撑》，《信息技术与标准化》2017 年第 9 期，第 7 页。

直接用于任何量化分析时”更是如此。〔40〕为了提升大数据的利用水平与分析效能，需要将小数据社科研究中已普遍运用和相对成熟的数据分析方法，如列联表分析、相关性分析、回归分析与统计学中处理高维数据的方法等，运用到大数据分析中，熟练运用 SPSS、SAS 等统计分析软件深度挖掘隐藏在法律大数据之中的宝藏。“只有通过对数据的大量输入并加上复杂运算，让数据不断产生又不断拆分、整合，融合生成新的产品，然后输出、使用，才能形成‘数据生产信息，信息改善决策’，这正是大数据发挥作用的基本原理。”〔41〕考虑到大数据分析的重要性，必须思考如何在中短期内提升大数据深度分析水平。当然，大数据法律研究在多大程度上真正需要运用以及如何运用统计学之外的其他分析方法，还有待进一步思考与探索。此外，面对法律大数据在数量、内容上的急遽增加，特别是面对“来源更加广泛，数据粒度更小，记录单元更加碎片化，结构更加多元化”的大数据，〔42〕现有的分析工具和统计手段可能无法满足处理需求，此时就要借助人工智能。通过将人工智能与法律大数据结合，对巨量数据进行智能筛选与算法分析，从而提升海量数据的分析效能。

随着数据来源以惊人的速度扩展，人们会逐渐加深对大数据的依赖，〔43〕也需要保持对大数据及其相关技术的超脱。一方面，大数据的分析手段如人工智能的算法本身就面临诸多“技术陷阱”，甚至被一些研究者认为是在黑箱中运作，〔44〕因此必须警惕其潜在风险。另一方面，特别“要防止为技术所裹挟，避免成为简单的技术主义者”。只有如此，才能保持“人文社会科学工作者的思想高度、理论品格和价值定位”，〔45〕进而产出更有温度的优秀成果。

第三，推动研究的团队化与多学科的交叉融合，并致力于培养复合型大数据法学人才。以往的法律实证研究注重研究者的专业性和个体性，表现为研究者独自收集资料、分析问题、撰写文章，个人的冥思与独创发挥着主要作用。在小数据研究中，这种模式能基本胜任。但大数据法律研究时常所处理的是海量杂乱数据，这“意味着人类的记录范围、测量范围和分析范围在不断扩大，知识的边界在不断延伸”。〔46〕大数据时代的到来，“提供了人文社会科学学者大规模协作的可能”，〔47〕也使之成为一种必要。在大数据法律研究及相关人工智能应用研究中，无论是数据的收集、整理，还是其分析、运用，都需要研究者具有多学科的知识与经验，如数据挖掘就涉及数据库技术、机器学习、模式识别、知识库工程、神经网络、数理统计、信息的可视化等众多领域，〔48〕知识结构单一的研究者甚至研究团队，往往难以应对。为了更好地开展大数据法律研究，法学研究者需要通过加强团队建设，特别是加强与计算机科学、软件科学、统计学等相关学科的专业人士以及大数据、人工智能科技公司之间的合作，以更好地应对大数据法律研究带来的机遇与挑战。同时，大数据

〔40〕 参见前引〔1〕，艾瑞斯书，第 12 页。
〔41〕 参见前引〔22〕，倪寿明文。
〔42〕 参见前引〔29〕，孙建军文。
〔43〕 参见前引〔8〕，Correia 等文，第 144 页。
〔44〕 参见前引〔7〕，Lyria Bennett Moses 等文，第 646 页。
〔45〕 欧阳康：《大数据与人文社会科学研究的变革与创新》，《光明日报》2016 年 11 月 10 日第 16 版。
〔46〕 涂子沛：《大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活》，广西师范大学出版社 2015 年版，第 57 页。
〔47〕 参见前引〔29〕，孙建军文。
〔48〕 参见前引〔35〕，刘鹏主编书，第 35 页。

法律研究者自身更需要突破知识储备、学术理念、价值方面的障碍，学习、掌握和应用统计学、数据科学的知识与研究方法。就此而言，面向未来的大数据法律研究复合型人才培养极为重要。目前，国内一些高校已经相继开设了大数据课程。未来，具备资源优势和技术条件的院校可以制定“大数据——法学复合型人才培养计划”，系统培养既懂技术又懂法律的复合型人才，为大数据法律研究与实践储备更多生力军。

第四，长期以来，基于小数据的法律实证研究一直是主流研究方式，有其重大学术价值。大数据研究在数据不足、方法把握不够的情况下，如何确保研究的科学性呢？对此，将小数据研究和大数据研究相结合应该是重要思路。具体而言，一是要在同一研究中就同一问题既应用大数据研究，也开展小数据研究，共同验证研究结论；二是在大数据研究中适当使用小数据研究的精细化思路与方法，把大数据研究细致化，从而提升大数据法律研究的科学性。

总之，身处大数据时代，我们正无时不刻不受到大数据广泛而深刻的影响。这不仅为大数据法律研究的发展提供了空前机遇，也是传统法律实证研究乃至法学研究范式升级转型的一个重要契机。立足眼下，更为要紧的工作可能是正视并努力突破大数据法律研究所面临的困境与羁绊。要正确理解法律大数据，科学、有效地开展大数据法律研究，开发大数据法律研究独特的技术与方法，提升数据获取与分析技术，注重培养复合型的研究人才。

Abstract: Big data based legal research is the latest development of empirical legal research, which will bring revolutionary changes to the paradigm of legal research. At present, there are some misunderstandings among Chinese legal scholars about big data based legal research, such as equating “a large amount of data” or “structured data” with big data. More importantly, there is also a lack of scientific method for using big data to carry out legal research. In the future, Chinese legal scholars should not only consider the question of how to get better access to legal big data, but also discuss the question of how to correctly understand and appropriately use “a large amount of data”. Moreover, they should make full use of statistical methods to conduct big data based legal research and explore ways of scientific use of machine learning and other new methods to analyze legal big data. Besides, it is equally important to pay continued attention to the mining and application of “small data” in order to support big data based legal research, and to strengthen the cultivation of inter-disciplinary research talents.

Key Words: legal data, big data based legal research, empirical legal research