大数据对法学研究的些许影响

白建军(北京大学法学院教授)

到底是相对封闭些，坚守自身特有的话语模式，还是适当打开自己，接受其他学科的影响，一直以来都是法学研究时不时面临的选择。比如，经济学之于法学、社会学之于法学、政治学之于法学，等等。现如今，大数据的概念来了。不管是不是情愿，法学可能又得有所准备，思考如何回应无处不在的大数据及其影响。

什么是大数据?有一本英国学者写的《大数据时代:生活、工作与思维的大变革》，从中大概得知何为大数据。所谓大数据，有三个特征:全样本、混杂性、相关性。其中，最重要的就是全样本。做经验研究的都知道，当样本等于总体时，抽样误差为零。但是，由于财力、人力、分析技术等条件的限制，人们很难拿到全样本。最早，国家为了知道纳税人的实际情况，就发展出各种消减、控制抽样误差的统计技术。而现在，随着计算机技术的发展，人们惊讶地发现，即使面对海量的信息，获取某类现象的全样本也并非完全不可能。基于这种全样本，人们可能更好地了解现实社会中的各种真实。于是，根据这些真实去预测某种现象的发生概率，就更可靠了。可见，大数据并不在于样本绝对量的大小，关键在于“全”。

比如，苹果公司的乔布斯身患癌症，尝试了许多种治疗方法，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，他支付了高昂的费用。他得到的不是一个只有一系列标记的样本，而是包括整个基因密码的数据文档。对于一个普通癌症患者，医生只能期望他或她的DNA排列同试验中使用的样本足够相似。而乔布斯的医生能够基于乔布斯的特定基因组成，按所需效果用药。尽管他仍然死于癌症，但这种获得所有数据而不仅是有限样本的方法还是将他的生命延长了好几年。从这个意义上说，某个研究的样本再大，哪怕达到上亿，如果相对总体而言只是几分之一，也只是大样本而不是严格意义上的大数据。反过来，即使对一个个体，也可能进行全样本的大数据研究。

于是，我们理解了为什么说费孝通的《江村经济》、孔飞力的《叫魂》、吉尔茨的巴厘岛人类学研究、朱晓阳的《小村故事》，尽管只聚焦某个点，但都尽最大可能收集与这个点有关的全部信息，因而也是某种意义上的大数据。例如，美国学者孔飞力是个汉学家，他研究专制权力如何凌驾于法律之上而不是受到法律的限制;官僚机制如何试图通过操纵通讯体系来控制最高统治者;最高统治者如何试图摆脱这种控制。对这样的大题目，孔飞力也是从发生在清代乾隆时期浙江的“剪辫案”这个个案着手。“叫魂”是一种民间迷信的妖术，换句话说，是一种能给他人带来不利后果的超自然的行为方式。在1768年的春天到秋天这大半年的时间里，因这种行为而引发的恐慌蔓延至大半个中国，使得整个国家陷入动荡不安。孔飞力发现，可以从小故事中发现大道理。于是，他在中国第一历史档案馆收集研究了海量的文献，至少有《朱批奏折》、《宫中上谕》、《宫中廷寄》、《附录奏折·法律·其他》、《上谕档方本》，以及图书集成局

1886年版的《刑案汇览》、薛允升的《读例存疑》、台北故宫博物院的《宫中档乾隆朝奏折》、1899年版的《大清会典事例》、光绪年编辑的《大清十朝圣训》等等，最终写出了《叫魂:1768年中国妖术大恐慌》一书。书中详细观察百姓、官僚、皇帝三个层面在叫魂案中的不同反应，发现每个群体对叫魂事件都有基于自己的利益而做的重新解释和塑造，并且这种再解释很大程度上已经远离了叫魂事件本身。可以说，叫魂事件是中国放大版的罗生门。我们从中看到的是犯罪定义者是如何从自身利益出发，千方百计对社会事实本身进行符合自身利益的再定义，从而获得有利于自己的结果。于是，犯罪定义过程就成了利益博弈过程，犯罪定义就成为一个并非纯粹客观的对于社会现实的反映，不可避免地带有浓重的定义者的主观色彩。

由此我想起，一位学者曾经计划深入到某个县法院，收集该法院全部文革前的判决书进行观察，看看在没有法律的情况下，法院是如何处理纠纷的。这无疑是一个极有价值的想法，尽管样本范围只限于一个县，但在这个范围内，如果做到全样本研究，那也是标准的经验研究，也是法律大数据研究。只可惜，这个计划未能成行。

可见，我们对大数据来袭的恐惧或反感，可能与我们对大数据的误解有关。形式上，大数据好像意味着大量的数据运算、统计甚至大型计算机的运用。其实，大数据的核心是尊重经验真实，敬畏经验真实，在乎经验的代表性。哪怕从一个小故事切入，只要收集足够的信息，也可能得到大张力、大格局的结论，用来解释、预测较大时间跨度和空间跨度的社会现象。正是由于不懂得这一点，我们一方面会排斥大样本经验研究，同时会夸大、轻信个案甚至只是几经裁剪的教学案例的可推论性，以为理解了这种个案，也就理解了所有个案。可是，天下没有两片一样的树叶，法律现象的异质性越大，某片树叶的代表性以及某个案件的可推论性就越有限。除非你坚持认为，天下所有的麻雀都没有任何差异，那你只解剖一只麻雀当然可以认为知道了所有麻雀。而在法律世界中，如果说所有案例都一个样，你自己信吗?

说到样本与数据，还有一点需要特别说明:大数据与大样本的区别其实也是相对的。当样本大到一定程度，即使不完全等于总体，只要其代表性和可推论性已经基本上不是问题了，就是近似的大数据。比如，谷歌基于5000万条最频繁出现的检索词条进行分析推算，成功地早于官方两周准确预测到流感的传播。那么，这个5000万是全样本吗?未必，只能说是最大的样本，其预测的可靠性其实也来自于这个样本的巨大。所以，当我们接受大数据时，切忌走到另一个极端，放弃所有大样本研究，一味地追求全样本。

用大样本做研究，结论不一定是科学的;科学的结论也不都出自大样本研究。但我还是偏好大样本研究，也常常受益于大样本研究。因为我相信，真理藏在大量事件背后。有人常会说，不用大样本，不是一样能得出你现在得出的结论吗?只用一两个故事，不是一样能表达你想表达的思想吗?没错。我不否认，幸运的淘宝者一伸手就能抓到个金娃娃。从一两个案例中，也可以提炼出某些宏大理论、原则或者规则。我不知道我有没有这个运气，但我知道我没这个勇气。不论怎样，多观察一些现象，得出结论所冒的犯错误的风险总会小一点。

一次，一个学生想写篇论文，题目是“从贪污罪看……犯罪学原理”。下面是我和这位学生的对话:

提问:贪污罪的确可以反映出……犯罪学原理。不过，刑法规定有几百个犯罪，何以见得某某犯罪学原理可以从A罪中抽象出来，因而也一定能从B罪、C罪……等其他各种犯罪中抽象出来呢?换句话说，你为什么对几百分之一的个罪足以代表所有犯罪抱有如此的自信或把握呢?

答辩:据我所知，著名社会学家费孝通先生的博士论文《江村经济》就是以一个乡村的材料为样本对中国农村状况的研究。

(好厉害!一个问题就惹出了费先生，要再问一个问题，恩格斯还不举着《英国工人阶级状况》出来帮他理论?按照他的意思，费先生可以用一个江村代表中国农村，我为什么不能用一个犯罪代表所有犯罪?)

提问:很好，你读了不少书。的确，费先生的博士论文在伦敦大学通过的当晚，他的导师就将其介绍给英国Routledge书局出版。书局的编辑拿到书稿后，还建议把书名《开弦弓，一个中国农村的经济生活》中的“开弦弓”(村)和“一个”去掉，直接称作《中国农民的生活》呢!不过，我们现在看到的该书中文版，书名仍是《江村经济》，而不是“中国农村经济”什么的。这是为什么呢?当然，费先生能不能帮得了你，要看你怎么回答这样一个问题:江村的确是中国农村的一部分，贪污罪也的确是犯罪的一部分。问题是，江村与中国其他乡村之间的关系，和贪污罪与其他犯罪之间的关系一样吗?

(我暗想，这可是第一个陷阱，看他怎么办。为了证明用一个犯罪代表所有犯罪的合理性，他很可能回答说，两个关系之间没什么根本区别，都是部分与整体的关系。正因此，江村可以代表中国农村，贪污罪也可以代表所有犯罪。言外之意，费先生做得，我为什么做不得。不过，他要真这么答就惨了，因为这将使他自己陷于一个被动境地，他没办法把“乡村”与“个罪”这两个分析单位完全不同的事物做简单类比。这显得多不严谨呀!果然，他非常审慎地绕开了这个陷阱。既没有说两者具有可比性，又没有说两者不具有可比性。)

答辩:这个，不一定，两者既有相同点，又有不同之处。不过，费先生可是社会科学大家，娴熟运用实证分析的研究方法研究许多社会问题，是我们每个学者的榜样。

(你看，博士就是博士。不仅绕开了我设下的陷阱，还用费先生堵我的嘴———意思是别在费先生面前摆弄实证研究!不过，该窃喜的还指不定是谁呢。他已经走近另一个陷阱。)

提问:你说的很好。也就是说，我们没有根据说，江村与其他乡村之间的关系，等同于贪污罪与其他个罪之间的关系。是，亦或不是?

答辩:嗯，是。

(因为真的太聪明，所以他已经意识到被套牢，可怜的学生一脸的沮丧。)

提问:既然没有足够的根据，从江村与其他农村之间的关系直接推论贪污罪与其他犯罪之间的关系，那你凭什么从一个贪污罪就抽象出那么大一个犯罪学理论呢?

……

我用这个例子是想说明，有的研究者对大样本、大数据的偏见，源自于并不真正理解小样本及个案研究。结果，在误解大样本研究的同时，也在误用小样本研究。

其实，我们生活中也常常见到缺乏样本意识的例子。一个城市中有一家大医院和一家小医院。根据记录，大医院三天来每天接生的新生儿中，男女各占约50%。而那家小医院三天来每天接生的新生儿中，恰巧60%是男孩，40%是女孩。这时，一对年轻父母尽管每天都梦想着生男孩，也不会仅仅根据这个统计数据就做出决定，到那家小医院产子。因为谁都知道，出生率的性别比是大约男女各占50%。大医院每天接产数量大，所以样本性别比更可能接近实际比例。但是，可以设想，如果这对夫妇并不知道这个一般的统计数据，或者说，如果他们脑子里没有这个先验概率，我们还敢肯定他们不选择小医院产子吗?这样提问有点可笑，因为他们不会蠢到分不清怀孕在先还是产子在先。但很难说类似的低级错误不以高级的形式发生在我们中间。

当然，要想证明一种理论，人们随时可以找到一两个事例作为支持这种理论的证据，这种个别事例也是一种意义上的真实。但严格地说，个别事例作为证据，不仅可能随时遭遇反例，而且其误差是不可控的。因此，只有一两个事例作为证据的所谓理论，很可能只能是一种意见、猜想或者判断，无法作为规律性认识为人所接受，更不能作为社会政策制定过程的决策基础。因为个别事件可能处在正态分布中的任何一个位置上，既可能碰巧代表大量同类事件的集中趋势，也可能只是极端事件。从这个意义上说，实证分析所追求的客观真实来自符合科学抽样程序性、规模性和可重复性要求的样本。

有学者就指出:大数定律保证非常大的样本确实能高度代表它从中抽出的总体。而如果一个研究人员信守小数律，就会对在小样本基础上得出的结论的有效性抱有夸大的自信。因为小数律的信徒是这样从事科学研究的:①在检验研究假设时，他把赌注放在小样本上，而未意识到他的失败机会非常之高。他高估了检验力。②他对于初期的趋势(如最早的几个被试的数据)以及观察到的模式的稳定性(如显著结果的数量和属性)，有过分的自信。他高估了结果的显著性。③在评价自身或别人的重复实验的时候，他对显著结果的可重复性，抱有非分的高预期。他低估了置信区间的范围。④他很少将实际结果与预期间的偏离归结为样本的变异性，因为对于任意的偏差，他都能发现因果“解释”。总之，人们对样本的直觉往往会产生不适当的后果。

当然，也许有人会说，这里所说的是发现真理的过程，而不是叙述真理的过程。发现真理时，当然要多观察些现象，得到更多个案的数据支持。而叙述真理时，样本就不需要太多。当你在课堂上讲授故意杀人罪的概念时，没必要历数几百个故意杀人案甚至穷尽所有个案后再告诉学生什么是故意杀人罪。没错，这其实正是我要说的。研究性论文或专著不是教科书，更不是学习心得或者综述。在教科书中，可以例举少量故事说理。但通过一项研究，你要告诉人们你发现了什么，而不是告诉人们你认为怎样。既然如此，怎么能刚看见一棵树就宣告说，我发现了一片森林?

由此还可以看出，就是对定量研究而言，样本规模不同，研究结果也可能不同。关键不在

于定量不定量，而在于是否对经验(集体经验、群体经验)心怀敬畏。我们可以掰着手指做样本，把十个手指的特征输入SPSS，照样可以运行交互分析、T检验、方差分析、多元线性回归、降维分析等几乎所有量化分析过程，然后用图表、饼图、线图等形式热热闹闹地表现出来。我们还可以上街随便找来三个路人，问他们是否赞成废除死刑。然后我们照样可以报告说，有66．6666%的民众赞成或反对废除死刑。这都是在做量化分析，但都是对经验的亵渎，是对现实生活的亵渎，是对科学的亵渎，也是对学者这个称谓的亵渎。换个角度看，我们不能说，一百个样本中的经验才是经验，一个样本中的经验就不是经验。更不能说，我的经验才是经验，你的经验就不是经验。关键在于，谁报告的经验相对更加接近生活现实的总体。

这样想问题便不难理解大样本研究的几个好处:第一，只要抽样过程符合随机性要求，样本越大，抽样误差就越小，由此所得结论偏离现实世界的可能性就越小。理论上说，当样本等于总体时，误差为零。第二，样本越大，所含信息、类型就越丰富，所研究的对象就能以更多的方式展现自己。通常，人们对定量分析有一个误解，认为量化过程对现象进行压缩处理，脱水后的研究对象失去了生气，面无血色。的确，这正是小样本量化分析可能有的效果。但随着样本的增大，人们可以灵活运用各种观察手段，看到事物更多的侧面。大样本用得好，可以让研究对象表情丰富，百般风情;而用极端个案说事，展现的往往是说故事者自己。极端个案的确有血有肉，生动具体。但是，由于无法控制某个极端个案在多大程度上代表了总体，因此，也无从知道这种用极端个案说故事的方法是否掩盖、侵吞甚至扭曲了多少客观真实。第三，样本越大，可供选择的分析工具也就越多，其结论也越可信。如果只有二、三十个样本，就算用上多元线性回归，统计软件也会报告结果，但这样的结果连你自己都不信。换句话说，样本越大，可选的分析工具越多，你就越自由。难道，你不想要这种自由吗?

当然，我们不能无条件地说，样本越大越好。我们把某个省的全部案件都拿来分析，有几十万，够大了吧?但我们还是不能把结论直接推论到全中国。样本是否具有代表性，还要看抽样程序是否规范。

抽样是从研究总体中抽取部分单位加以研究，并用所得结果推断总体特征的方法，是实证研究的基本功之一。之所以需要抽样，首先因为样本与总体是个别与一般的关系。研究总体，没有必要对总体中每个单位进行逐一调查。只要符合统计要求，可以认为样本特征近似于总体特征。第二，由于需要研究的总体巨大，受人力、财力所限，除国家实施的大规模人口普查以外，不可能逐一调查所有研究对象的个体。所以，不仅可以借助样本观察总体，也只能借助样本观察总体。第三，被研究的总体本身具有程度不同的异质性，只抽取其中一个单位，不可能代表总体中其他未被抽取单位的情况。因此，用来观察总体的样本尽管不可能太多，但也不能过少。过多的样本耗费调查资源，过少的样本可能产生过大的抽样误差。

具体来说，抽样分为随机抽样(概率抽样)和非随机抽样(非概率抽样)两种。在随机抽样中，总体中的每个单位都有同等机会被抽取成为样本。其特点有四:第一，按随机原则抽取而非随意抽取。第二，每个单位被抽取的概率是已知的，而非未知的。第三，由样本推论到总体的可靠程度可计算，可控制。第四，抽样前，对总体边界已知。随机抽样分为简单随机抽样、分

层抽样、系统抽样、聚类抽样等等。与随机抽样不同，非随机抽样是无法精确给出抽样误差因而无法将研究结论直接推论到研究对象的总体的抽样方法。非随机抽样包括方便抽样、立意抽样等等。抽样技术的关键，就在于尽可能减少误差，控制误差，抽出真正代表总体的样本。

作为社会现象的一部分，法律现象与自然现象之间有着显著区别。法是由人制定的，法是由人实施的，法是由人违反的。所以，法律现象有着太多的异质性和不确定性。但另一方面，法律现象的总体又往往巨大无比，每年法院处理的各类案件几百万件，每个达到一定责任年龄的公民都是潜在的违法者，所有公民都是潜在的被害人。那么，法学研究该如何迎接大数据的到来，至少做出一些像样的的大样本研究呢?

首先，全样本选题。在法律现象的研究中，并不是所有问题的对象总体都是十三亿人或者百万、千万计的案件。比如，截止到2006年6月《刑法》修正案(六)通过颁布以前，中国《刑法》规定有425个罪名，截止到2003年12月23日，最高司法当局发布的刑事司法解释共有1233个，某一笔专项资金总额400亿元，涉及该项资金的全部职务犯罪案件共几百件。这些，都是力所能及的全样本选题。此外，某个行业的行业性规范、某个部门的执法活动等等，也都可以成为全样本研究的选题。除了这些以全国范围为总体的选题以外，还可以将有代表性的某个省、某个市、某个地区，甚至某个县、乡的全部某类案件、某些司法文书、判决结果、政策文件等确定为全样本研究的对象。此类全样本虽非全国范围的全样本，但为什么研究对象及其结论一定要能推论到全国才算是科学呢?为什么学术活动一定要左右于一个中心才算是触摸到了真理呢?其实，这本身就是一种关于学术研究的误解，一种盲目追求宏大叙事而不屑于细微具体研究的浮躁。既然如此，法律实证研究中丰富的全样本选题，是尽可能降低抽样误差的一个较好对策。

其次，合理确定抽样框架。所谓抽样框架，就是一份与总体非常相似的用来选取具体样本的名单。例如，1936年是美国的选举年，民主党竞选人是竞选连任的总统富兰克林·罗斯福，共和党的竞选人是来自堪萨斯州的阿尔弗·兰登。为了预测谁将在选举中获胜，美国的《文摘》杂志进行了一次美国历史上规模最大的民意测验，它调查了240万美国人的选举倾向。根据调查结果，《文摘》杂志宣布，兰登将以57%对43%击败罗斯福。而实际的选举结果却是，罗斯福以62%对38%获得大胜。预测失败的问题就出在抽样框架上。《文摘》杂志总共寄出了1000万份调查表，地址与姓名大都取自于电话簿与汽车俱乐部会员名单。但在1936年，大多数美国人没有安装电话，很多人也没有汽车。这样，低收入的穷人就被完全排斥在调查之外，而正是这部分穷人支持了罗斯福，造成了同样是美国历史上规模最大的抽样误差。这个例子中的抽样框架就是《文摘》所选定的电话簿和汽车俱乐部会员名单。从抽样原理来看，这个抽样框架与美国全体选民这个总体之间的相似性程度不大，所以才会预测失败。

由此也可以看出，关键不在于样本的数量大小，也不在于抽样框架是出于何种目的确定的，而在于根据某个框架所获得的样本与总体之间是否相似。而所谓是否相似，其实又有多个可能的侧面:年龄、性别、职业、文化，还是社会地位?只要对既定研究目的而言，抽样框架与总

体之间具有相似性即可，而两者不可能在所有方面都满足相似性要求。调查者所以选定电话簿和俱乐部名单，也是因为他们真的相信这个框架的选举意向能代表总体。否则，他们为什么要有意制造自己的预测失败呢?所以，当无力于全国普查时，我们可以根据研究目的的要求确定一个抽样框架，假定这个抽样框架可能代表总体，然后或者基于这个框架进行全样本研究，或者在这个框架内进行随机抽样。这样，研究结论能否推论到总体首先可以基本上排除主观偏好或者其他人为因素对样本获取过程的影响，而剩下的问题只是人们在多大程度上相信这个框架与总体之间的相似性，或者说两者之间的差异在多大程度上可能对研究结论向总体推论构成根本性影响。

例如，我们不可能首先获得全国所有刑事案件的名单，然后据此进行随机抽样，但我们可以把来自最高法院各业务庭、研究机构、出版单位、网站等权威机构公开发布、发表的全部真实判决设定为抽样框架，并称其为“示范性案例”，然后抽取其中的某类案件进行全样本研究。这种案例的代表性在于:第一，由于这些案件来自全国各地，由各地各级法院选送，具有对全国总体的代表性;第二，由于是最高法院各权威机构认可并公开的案件，因而具有对司法实践的指导性;第三，由于其中绝大部分案件属于生效判决，因而具有一定的有效性;第四，由于各地选送案件以及最高法院各单位选取案件时充分考虑到案件类型和性质的多样化，因而对学术研究而言具有一定的标志性;第五，由于是公开发布的案件，因而对公民行为而言具有相当的规范性、模范性和可预测性;最后，由于提取了这个范围内的几乎全部某类案例，将抽样误差降低为零，因而具有研究依据上的准确性。其实，如果可能将总体的所有特征一模一样地微缩到某个随手可得的抽样框架中的话，无异于对总体完成了一次严格的随机抽样，并以其结果为抽样框架进行二次抽样，其实这已经不是在选择抽样框架而是进行多段抽样了。

再次，避免盲目放大样本容量。一般而言，研究总体本身的异质性程度越大，需要分析的变量的个数越多，则所需要的样本规模就越大。但是，一个占总体5%的样本，未必要比一个只占总体1%的样本要好上5倍。有研究证明，在总体小于1000的情况下，如果样本占总体的比例低于30%，那么，样本误差将会很大。但是，当总体的规模增加时，样本比例的作用趋向于越来越小，当总体为10000时，我们只需有10%的样本比例，当总体为150000时，1%的样本比例就已经足够。当总体为1000万或者以上时，样本比例的增加实际上已经不起作用。换言之，样本规模绝对数值的重要性大大超过样本占总体比例的重要性。

最后需要说明，最高法院已经从2013年起开通了裁判文书网，公开了几乎全部司法判决文书。尽管在技术上还有待改进，但这件事的意义之大，超出了许多人的想象。至少今后我们不能再说，拿不到全样本，所以无法做大数据。现在的问题是，司法当局已经为法律大数据研究提供了相应的条件，学界能跟上吗?