大数据如何改变经济学研究范式?*

洪永淼　汪寿阳

摘要:本文首先从经济学视角探讨大数据给经济学实证研究所带来的范式变革,包括从理性经济人到非完全理性经济人,从孤立的经济人到互相关联的社会经济人,从代表性经济人到异质性经济主体,以及从经济分析到经济社会活动的系统分析。然后,从方法论视角讨论大数据给经济学实证研究方法所带来的变革,包括从模型驱动到数据驱动,从参数不确定性到模型不确定性,从无偏估计到有偏估计,从低维建模到高维建模,从低频数据到高频甚至实时数据,从结构化数据到非结构化数据,从传统结构化数据到新型结构化数据,以及从人工分析到智能分析等。大数据引起的经济学研究范式与研究方法变革,正在深刻重塑经济学发展方向,不但加强了经济学实证研究范式的趋势,而且还进一步突破了现代西方经济学的一些基本假设的局限性,使经济学研究日益呈现出科学化、严谨化、精细化、多元化(跨学科)与系统化的趋势,并且与社会科学其他领域在方法论上日益趋同。中国大数据资源,为从中国经济实践中总结经济发展规律,从中国特殊性中凝练可复制的经济发展模式,从而构建具有深厚学理基础的原创性中国经济理论体系,提供了一个得天独厚的"富矿"。

关键词:大数据　文本分析　机器学习　研究范式　研究方法　反身性
DOI:10.19744/j.cnki.11-1235/f.2021.0153

一、引言

在中国经济学界,绝大多数经济学家已形成高度共识,认为中国经济发展有其内在逻辑和一般规律,需要对中国经济学进行原创性理论创新,以探索中国经济发展规律(王一鸣,2017;王东京,2018;杨红丽等,2020;洪永淼、薛涧坡,2021;侯增谦,2021;刘伟、蔡志洲,2021;杨耀武、张平,2021)。但是,中国经济学家对于中国经济学应该采用什么样的研究范式与研究方法,尚未达成广泛的共识,甚至存在较大争议。例如,关于定性分析与定量分析的关系、经济思想与数学、模型的关系等,观点各异(洪永淼、汪寿阳,2020)。在研究范式方面尚未达成广泛共识,决定了中国经济学家很有必要对研究范式进行深入的学术讨论。中国经济理论创新,需要对研究范式进行深刻变革。洪永淼、汪寿阳(2021a)论述了研究范式对经济学研究的重要作用。本文首先讨论研究范式对提高经济学研究科学性的重要意义以及过去40多年经济学实证研究或经验研究(empirical study)范式革命产生的背景与特点,然后从经济学视角阐释大数据革命对现代经济学的一些基本假设和基本研究范式的深远影响,并从多个维度具体讨论大数据和机器学习如何深刻改变经济学实证研究方法。

我们的分析表明,大数据革命强化了经济学"实证革命"的研究范式,并且正在引起经济学研究范式的变革和研究方法的创新,推动交叉学科研究,促进经济学和社会科学其他领域之间的融合,促进经济学和数学、人工智能、计算机科学、统计学、认知科学等自然科学学科之间的交叉。

40多年来,中国通过改革开放,逐步建立起中国特色社会主义市场经济基本制度,主动融入世界经济体系,充分发挥比较优势,实现经济长期持续快速增长,成为了世界第二大经济体。中国经济崛起是21世纪世

*作者感谢陈丽纯、胡毅、王晓虎、谢天、许志伟、薛涧坡、杨露、杨志勇、张永山、张宇、周颖刚的建议与帮助,以及国家自然科学基金委员会基础科学中心项目"计量建模与经济政策研究"(项目号:71988101)和国家自然科学基金专项项目"经济科学发展战略研究"(项目号:71940004)的资助。汪寿阳为本文通讯作者。

界最重大的经济事件，正在深刻改变世界经济格局及其发展趋势。新时代改革开放和社会主义现代化建设的丰富实践，是理论和政策研究的“富矿”。党的十八大以来，中国及时总结新的生动实践，不断推进理论创新，在发展阶段、发展理念、发展格局、所有制、分配体制、共同富裕、市场机制、政府职能、宏观调控、产业结构、区域规划、企业治理等重大问题上提出了许多重要论断，形成了习近平新时代中国特色社会主义经济思想。如何以习近平经济思想为指导，从中国经济实践中揭示中国特色社会主义市场经济发展规律，从中国特殊性中凝练可复制的中国经济发展模式，是中国经济学家的历史机遇和时代使命。由于中国超大经济体的规模优势和数字经济的快速发展，中国在大数据资源方面与西方发达国家大致站在同一起跑线上，加上中国经济拥有多样性的所有制结构、丰富的“政策数据库”等特点，中国经济学家如果能够充分利用大数据所提供的有关中国经济实践的大量信息，与时俱进地探索科学研究范式，将能够从中国经济发展中揭示中国经济发展规律，构建具有深厚学理基础的原创性中国经济理论体系。

二、经济学研究的“实证革命”

任何学科的发展离不开其研究方法及其知识生产与积累方式的进步，而一门学科是否具有科学性或者说其科学性的程度有多高，关键在于它是否有一个与时俱进的科学研究范式。所谓研究范式，是指一个学科的学术共同体进行科学研究时所遵循的模式与框架，是学科知识生产与积累的基本研究方法的总和，这是影响经济学研究质量的关键因素。历史上自然科学每一次重大理论突破，都伴随着研究范式的革命和研究方法的创新（Kuhn，1996）。经济学的发展也是如此。自亚当·斯密《国富论》发表以来，经济学研究范式随着时代的变迁一直在变化。19 世纪 60、70 年代，经济学产生了马克思主义政治经济学以及“边际革命”；20 世纪 30 年代，出现了“凯恩斯革命”；20 世纪 50 年代，诞生了“新古典综合”。过去 40 多年来，现代经济学又出现新的范式革命，即“实证革命”（empirical revolution），也称为“可信性革命”（credibility revolution）（王美今、林建浩，2012）。实证革命是指经济学以数据作为基础，以计量经济学为主要方法研究并解释经济变量之间的逻辑关系，特别是因果关系的研究范式革命。Hamermesh（2013）发现，从 1963 到 2011 年发表在经济学顶级期刊的论文中，20 世纪 80 年代中期以前大部分论文都是理论性的，而从 80 年代中期以来，实证研究论文比例攀升到超过 70%。Angrist 等（2017）指出，从 1980 到 2015 年，国际顶尖与主流经济学期刊以数据为基础的实证研究论文数量从不到 35%上升到 55%左右，而理论性论文数量则从近 60%下降到不到 40%，实证研究成为现代经济学最主要的研究范式。40 多年来，中国经济学也从定性研究为主转变为以定量实证研究为主（李子奈、霍玲，2005；洪永淼、薛涧坡，2021；洪永淼等，2021）。

经济学实证研究之所以逐渐流行并逐渐占据主导地位，得益于计算机技术的不断发展，以及数据可获得性的不断提高，但最重要的原因在于实证研究更加符合现代科学研究范式。什么是科学研究范式？Kuhn（1996）在《科学革命的结构》一书中提出，任何理论假说都需要经过经验验证，才能证明其正确性与有效性。鄂维南（E，2021）指出，自牛顿以来，自然科学研究基本上按照开普勒和牛顿两种不同范式展开，其中牛顿范式是基于第一性原理的研究方法，其目标是发现物理世界的基本原理，如牛顿、麦克斯韦、玻尔兹曼、爱因斯坦、海森堡、薛定谔的理论物理学，主要研究方法是“思想实验”，而开普勒范式是指数据驱动的研究方法，通过对数据的分析，寻找科学规律并解决实际问题，如行星运动的开普勒定律。无论是哪一种范式，任何理论假说都需要接受经验验证，而且在相同的条件下，任何结论应该能够被独立地重复证实或发现。撤稿观察数据库（Retraction Watch Database）显示，《自然》（Nature）和《科学》（Science）从 2001 到 2020 年各撤稿 67、74 篇，其原因是这些文章的结论不能获得大多数人重复实验的验证。最近，《金融学报》（Journal of Finance）自创刊以来首次撤回获得该期刊 2020 杰出论文奖的一篇论文，主要原因是该研究的核心实证结果无法复制，研究成果可靠性不足。

可能有人会提出这么一个问题：上述实证研究范式是自然科学的研究范式，而社会科学与自然科学存在很大差别，特别是很多自然科学的主要研究对象是自然界，是物；而包括经济学在内的社会科学的主要研究对

象是人,是具有意识的人。社会存在决定社会意识,但社会意识对社会存在也有反作用,这种互动关系在社会科学被称为"反身性"(reflexivity)。这是社会科学与自然科学最显著的不同之处。社会科学与自然科学还有其他不同之处,如绝大部分经济社会现象都是非实验性的。自然科学诞生以来,其理论已被历史与实践证明了是科学理论,可精确解释与预测自然界的现象与运动规律,而这些科学理论主要是采用了科学研究范式而创建起来的。因此,社会科学可以而且应当借鉴自然科学的科学研究范式,以提升社会科学的科学性与先进性。不能因为社会科学与自然科学研究对象不同,就认为自然科学的研究范式不适合于社会科学,这实际上是以特殊性否认普遍性。同时也应强调,借鉴自然科学的研究范式与研究方法,并不是机械地照搬照抄,而是需要根据社会科学的特点(如反身性与非实验性),有所发明与创新,使之适用于研究社会科学。例如,由于社会经济系统所产生的观测数据具有非实验性的特点,经济学家与计量经济学家在识别经济因果关系时便面临所谓的内生性(endogeneity)问题,因此发展了很多可克服内生性的因果推断方法,如工具变量法、双重差分、断点回归、倾向性得分匹配与虚拟事实分析(counterfactual analysis)等,这些方法也被广泛用于定量评估各种经济社会公共政策。有关这些方法的介绍,参见 Angrist 和 Pischke(2009)。

三、大数据与经济学研究范式变革

经济学实证研究范式包含三大要素:(1)数据,包括观测数据和实验数据,大部分经济数据是观测数据;(2)分析方法与工具,包括计量经济学模型、方法、计算工具,如统计软件包和机器学习算法程序包;(3)经济理论,用于提供经济解释、经济直觉;经济理论本身也常常是受检验的对象。经济学实证研究的最主要方法论是计量经济学,这一方法论学科对推进经济学科学化发挥了重要作用(洪永淼,2007;李子奈,2008)。

以互联网、移动互联网、云计算、人工智能为代表的信息科技革命和第四次工业革命正在深刻改革人类的生产与生活方式,催生了数字经济这一新的经济形态。人类很多经济社会活动与行为轨迹都以数字化的形式记录下来,形成了各种形式的大数据,这些大数据包含着大量互相关联的微观经济主体行为动态信息。早在2010年美国加州举办的科技经济会议(Techonomy Conference)上,谷歌总裁施密特(Eric Schmidt)就曾表示:"当今世界每2天产生的数据相当于2003年以前人类历史中产生的所有数据的总和。"相对于传统数据,大数据具有什么特点?大数据对经济学研究,特别是经济学研究范式与研究方法有什么影响?众所周知,大数据有以下4个特征:(1)规模性(volume),即样本容量大,变量个数多。若样本容量大于变量个数,称为高大数据;若变量个数大于样本容量,称为胖大数据。大部分经济大数据均是大量互相关联的微观经济主体(如消费者、生产者、投资者等)的动态行为大数据。(2)高速性(velocity),即可获得高频数据甚至实时数据。(3)多样性(variety),即具有结构化数据,又有各种形式的非结构化数据,包括文本、图形、音频、视频等。即使是结构化数据,也有新型的数据,如矩阵数据、函数数据、区间数据、符号数据等。(4)准确性(veracity),即噪声大、信息密度低。这些特征是传统数据所不具备的。

在很多情景下,大数据包含传统数据所没有的信息。例如,高频微观行为大数据提供了大量互相关联的经济主体的互动关系如何随时间演变的信息,而类似于一次性快照的传统微观调查数据则不包含这些动态信息。又如,社交媒体平台的文本数据包含了经济主体(如投资者、消费者)丰富的情绪、情感等心理信息,这也是传统数据所没有的。情绪、情感是人类的非理性现象,但可从文本数据中提取并定量测度。新型数据需要新的分析方法与工具,例如对文本数据的情感分析需要用到自然语言处理技术与包括机器学习在内的分析方法,如词典方法(dictionary methods)、主题模型(topic models)、词向量模型(word embedding models)(关于自然语言处理的介绍,参见 Manning et al.,2008;Jurafsky and Martin,2009)。大数据的可获得性和机器学习的应用,不可避免地引起经济学实证研究范式与研究方法的变化(胡毅等,2019)。那么,大数据和机器学习如何改变经济学的研究范式与研究方法呢?大数据是开辟新的研究领域、研究方向、研究命题,还是以更新颖更有启发性的方式来回答传统问题?大数据是带来一次研究范式的变革,还是仅仅只是渐进式范式变化的延续?以下,我们首先从经济学视角来讨论这些重要问题。

（一）从完全理性到非完全理性

长期以来，新古典经济学假设理性经济人在完全竞争市场环境下进行经济决策，优化配置稀缺资源，但理性经济人这一新古典经济学的最基本假设与实验经济学、社会心理学的经验发现并不兼容。随着经济理论的发展，完全竞争市场假设拓展为垄断与寡头垄断，完全信息假设拓展为信息不对称假设，而完全理性经济人假设也通过实验经济学得以放松，如假设有限理性。宏观经济学的理性预期学派也研究认知偏差（expectations bias）对经济运行所带来的影响（崔丽媛、洪永淼，2017）。这些研究均取得了丰硕的理论成果，如产生了信息经济学、规制经济学、实验经济学、行为经济学、行为金融学等新兴学科。

社会科学和自然科学一个最大的不同之处是自然科学的主要研究对象是自然界，是没有意识的物，而社会科学的主要研究对象是有心理意识的人，存在情绪、情感、价值判断等心理现象。比如，新冠肺炎疫情大流行，给人类社会经济带来了巨大的不确定性，经济主体对于这种不确定性给现在与未来经济造成的可能影响会形成一定的心理预期，这种预期反过来会影响经济主体当下的消费与投资行为，从而影响整个经济运行。经济学家早就认识到心理因素在经济学中的重要性，19 世纪 70 年代的“边际革命”首先通过效用这个概念将心理因素引入经济学的分析框架中，宏观经济学从凯恩斯革命到理性预期学派，都非常注重经济主体（如消费者、投资者等）的心理预期对宏观经济的影响，如所谓“流动性陷阱”就是指投资者对前景极其悲观，因此不管利率有多低也不愿意借贷去投资。但是，很多经验事实表明，人的决策并不都是完全理性的，常常受到情绪、情感、情景以及偶然因素的影响（Shiller，2000，2019）。要精确研究经济主体的心理因素（如投资者的情绪、情感，消费者的幸福感、满意度等）及其对经济的影响，需要对经济主体的心理进行测度。由于传统数据很少包含经济主体的心理信息，以往很难开展关于经济主体的心理如何影响经济的定量实证研究。如今，大数据特别是文本数据，提供了很多消费者、投资者的情绪、情感、价值判断等信息，这些心理信息可通过自然语言处理技术与人工智能方法从文本数据中提取出来（Tetlock，2007）。因此大数据使经济学家能够采用定量实证研究方法，精确研究社会心理对经济的影响。诺贝尔经济学奖获得者罗伯特·席勒（Shiller，2019）在《叙事经济学》一书中，倡导重视研究社会情感及其传染对重要经济事件的影响。众所周知的抢购、银行挤兑、线上直播、羊群效应、资产泡沫、金融传染病等，都是社会情感及其传染影响经济行为的例子。2021 年初，美国股市大量散户投资者在与机构投资者博弈时取得了胜利，让人们见证了散户投资者通过社交网络平台的情感传染所爆发出来的巨大影响力。同样地，作为一种长期形成的社会心理与行为习惯，文化也可定量刻画。例如，荷兰社会心理学家霍夫斯泰德（Hofstede，1984，1991）基于跨国调查数据提出了一个文化维度理论，从 6 个维度定量测度不同国家的文化差异。另外，可从企业财务报表和工作报告等文本数据中提取刻画文化元素的有用信息，构建并测度文化变量，这样便能精确研究企业文化对企业经营的影响（Goldberg et al.，2016；Li et al.，2021）。

（二）从孤立经济人到社会经济人

新古典经济学所假设的理性经济人在微观层面上是一个孤立的经济人，这与现实生活中的人完全不同，这是新古典经济学最突出的一个缺陷。在《〈政治经济学〉导言》中，马克思批判了从孤立的个人出发来研究财富与生产的错误做法。马克思强调人的社会性，注重研究人与人之间的生产关系。现实中，人是社会人，人与人之间具有千丝万缕的直接或间接的联系。特别是随着互联网技术的广泛使用和经济全球化的深入发展，人与人、企业与企业、行业与行业、群体与群体、国家与国家之间等各个层面的联系更加紧密。这些联系所构成的各种社会网络（如地理网络、行业网络、平台网络、数字网络等）会深刻影响微观经济主体的行为与心理。以前，绝大多数的微观调查数据相当于一次性快照的数据，不包含人与人之间互相联系的信息，因此很难将经济人当做社会经济人加以研究。现在，大量微观行为高频大数据，如脸书（Facebook）、推特（Twitter）、领英（LinkedIn）、微博、QQ、知乎、豆瓣、贴吧等社交媒体平台上的各种文本数据，可提供大量、丰富的人与人之间的动态联系信息，这使经济学家可将经济人视为社会人，研究他们之间的经济社会关系及其动态演变。习近平总书记在 2020 年 8 月召开的经济社会领域专家座谈会上指出，“我国社会结构正在发生深刻变化，互联网深刻改变

人类交往方式,社会观念、社会心理、社会行为发生深刻变化。”大数据可用于精确刻画与研究这些社会变化及其影响,以适应社会结构、社会关系、社会行为方式、社会心理等的深刻变化。

(三)从代表性经济人到异质性微观主体

20世纪30年代的凯恩斯革命宣告宏观经济学的诞生,对世界各国经济政策特别是货币政策与财政政策的制定产生了深远影响。宏观经济学主要研究总产出(如GDP)、价格水平(如CPI)、失业率、汇率等宏观经济变量之间的数量关系,如奥肯定律(Okun's law)、泰勒规则(Taylor's rule)等。在20世纪70年代之前,宏观政策分析主要使用简约联立方程组刻画宏观经济变量之间的数量关系,其本质是通过观察经济主体对既往政策变化的反应,对其行为方程进行估计,从而预测新政策的效果。但这种方法没有考虑到政策变化后经济主体通过预期改变自身行为,从而导致政策失效的可能性(Lucas,1976)。“理性预期革命”后,宏观经济学逐渐发展出动态一般均衡模型,通过引入理性代表性经济人内生跨期最优决策来解决“卢卡斯批判”问题,其本质是假设经济主体的偏好等结构参数(structural parameters)不会随政策而改变,通过估计代表性经济主体的结构参数,而非其行为参数,并结合经济主体跨期优化的理论结果,来预测政策效果。但在单一代表性经济人假设下,宏观模型仍然缺乏对微观主体决策行为的深入刻画,特别是刻画宏观经济变量之间数量关系的方程并不是在众多互相关联的微观主体行为的假设基础上推导出来的。现实中的经济主体,如消费者、生产者、投资者、地方政府等,存在显著的异质性(heterogeneity),具有不同的结构参数以及不同的经济行为。例如,低收入和高收入家庭受新冠肺炎疫情的影响程度不同,他们应对疫情的行为也不一样。在中国,不同所有制的企业,其行为也有很大差别。宏观经济总量通常是由加总(aggregation)获得的,由于存在异质性,加总可能导致信息失真。由异质性很强的不同群体所构成的宏观经济动态趋势,可能与代表性经济主体假设下的宏观经济趋势有显著差别,甚至相反。比如,通过效用最大化推导出来的个人消费函数(即个人消费与个人收入之间的关系),在加总后并不能得到相同函数形式的宏观消费函数,除非每个人的效用函数均属于齐序函数(homothetic function)(Varian,1999)。Granger(1980)通过一个例子说明,具有“短记忆”(short memory)性质的个人消费时间序列,在加总后,宏观消费变量将变成具有“长记忆”(long memory)性质的时间序列。微观主体的异质性使得为宏观经济理论奠定微观基础的尝试更加困难。然而,大量高频微观经济主体行为大数据的出现,如消费者在线消费数据与企业投资数据,可用于识别外生经济或者政策冲击对不同行业、不同部门、不同微观主体产生的分布效应(distributional effects),刻画这些冲击在经济系统内的传导路径,从而更好理解宏观经济政策传导机制,帮助政府制定精准有效的宏观经济政策。

(四)从经济分析到经济社会系统研究

人类社会是一个复杂系统,由经济、科技、政治、法律、社会、历史、文化、地理气候、生态环境等诸因素共同组成,而且经济与其他因素交织在一起。经济学家早就认识到这一点,因此除了政治经济学外,还出现法与经济学、经济史学(包括量化经济史学)、生态经济学、环境经济学、气候变化经济学、教育经济学、健康经济学、文化经济学等交叉学科。新一代信息技术的快速发展与广泛应用,除了记录大量微观经济行为大数据外,还产生很多关于生态环境、医疗健康、政治法律、公共政策、历史文化等领域的大数据。这些大数据的可获得使经济学家能够在一个统一的经济社会框架中,以系统方法研究经济与其他因素或其他子系统之间的互动关系(洪永淼、汪寿阳,2021a)。在大数据背景下,经济学的跨学科交叉融合研究的趋势因此日益加强,经济学与社会科学其他领域之间的界限越来越模糊,特别是社会科学各个领域以大数据为基础的定量实证研究范式与研究方法日益趋同。近年来,由于大数据在社会科学各个领域的可获得性与广泛使用,认知科学、实验心理学、人工智能、计算机编程、数据科学等方法论学科的知识与方法,如机器学习、深度学习、文本分析、社会网络分析以及模拟仿真等,已被广泛应用于社会科学各个领域的研究中。事实上,经济学与社会科学其他领域一个共同的主要目的是识别因果关系与定量评估经济社会公共政策,又都面临经济社会系统的非实验性特点,因此所使用的很多定量实证方法具有共性。例如,经济学家和计量经济学家所熟悉的很多因果推断和定量政策评估方法,包括工具变量、双重差分、断点回归、倾向积分匹配、虚拟事实分析等,也日益广泛应用于社会学、政

治学、历史学、教育学等社会科学其他领域。

2009年，美国15位学者（Lazer et al.，2009）在《科学》上提出“计算社会科学”（computational social science）这个新兴学科的概念。社会科学的最主要研究对象是人，它是关于人类如何思考（心理学）、如何处理财富（经济学）、如何互相联系（社会学）、如何治理人类自己（政治学）以及如何创造文化（人类学）等的科学。2012年，14位欧美学者（Conte et al.，2012）联合发表《计算社会科学宣言》，呼吁计算社会科学通过结合信息技术、人工智能和社会科学理论来解决新时代社会科学面临的重要问题。目前，计算社会科学进入了基于大数据的实证研究范式：数据驱动（data driven）的研究方法将算法和计算工具应用于复杂数据，以揭示社会现象的本质。计算社会科学的研究范式蕴含着交叉学科方法，需要包括经济学家在内的社会科学家、认知科学家、计算机科学家、数学家、统计学家、物理学家等各领域学者的通力合作。

综上所述，大数据的可获得性，特别是大量互相关联的异质性微观经济主体行为（包括心理）高频大数据，使经济学实证研究有望突破现代西方经济学中一些经常受到批判的重要缺陷，如假设孤立的理性经济人，忽略经济主体的社会联系（即社会性），忽略经济主体进行经济决策时所处的历史、文化、心理、情景等因素的影响。大数据特别是文本数据使得测度社会心理变量（包括情感、情绪、价值判断）和文化变量成为可能，使经济学的实证研究能够将社会科学的“反身性”特点纳入定量实证研究框架，即所谓的文本回归（textual regression）分析框架，从而将原来只能进行定性分析的问题转变为严谨的定量分析，并且通过跨学科交叉研究，将经济置于一个更大的人类经济社会系统之中，以系统的观念与方法研究经济与人类社会系统中其他子系统的互动关系。此外，利用大量互相关联的微观主体行为高频大数据，可让经济学家更好识别外生冲击（如新冠肺炎疫情、中美地缘政治冲突）或政策冲击对不同微观主体的分布效应、识别这些冲击的传导机制，从而奠定宏观经济学的微观基础。毫无疑义，历史上对经济学发展有重要影响的哲学、政治学、法学、社会学、历史学、心理学等学科将继续产生重要影响，与此同时，因大数据分析而需要的数学、统计学、计算机科学、数据科学、认知科学等学科也将发挥重要的方法论作用，所有这些学科将极大推进经济学和人文社会科学之间以及经济学和数学与自然科学之间的交叉融合。

四、大数据与经济学研究方法变革

新型数据需要新的分析方法与工具。Einav和Levin（2014）讨论了大数据，特别是美国政府部门行政大数据和私人部门大数据如何改变经济学实证研究的统计方法。Varian（2014）和洪永淼、汪寿阳（2021a，2021b）分析了大数据与机器学习给计量经济学与统计学带来的机遇与挑战。Mullainathan和Spiess（2017）和Athey（2019）讨论了机器学习对计量经济学理论与方法的影响。这里，我们从多个维度具体说明大数据如何深刻改变经济学实证研究方法。

（一）从模型驱动到数据驱动

首先是从模型驱动（model driven）转变为数据驱动。从广义上说，经济学以数据为基础的定量实证研究可视为数据驱动的研究。从狭义上说，大数据背景下的模型驱动研究和数据驱动研究有其特殊含义：两者都是以数据为基础的研究，但前者通常是指使用一个低维参数模型（如线性回归模型），这样的模型存在误设的可能性，从而导致模型证据（model evidence）和数据证据（data evidence）出现差异；而后者是指直接使用机器学习算法分析数据，机器学习算法本质上是一种正则化（regularized）非参数统计方法，不假设具体的函数形式，因此具有较大的灵活性，比较接近数据证据（洪永淼、汪寿阳，2021c）。随机森林提出者里奥·布瑞曼（Breiman，2001）详细讨论了这两种研究范式。以下，我们在经济学框架中分析这两种研究方法的优劣性与异同点。

在现代经济学中，很多经济理论都是基于一些关于制度、技术、经济主体偏好与行为等基本假设上通过数学模型建立起来的。这种理论建模方法是对复杂经济系统的一种高度简化与抽象，聚焦于主要经济变量之间的因果关系，以揭示经济运行的内在本质，但由于数学模型的高度简化与抽象，现实中的很多其他因素没有被

考虑进来。因此，当经济模型用于解释现实观测数据时，可能会出现模型误设的情形，从而对经济实证研究的结论造成不可忽略的影响（洪永淼，2021）。这是模型驱动的实证研究的一个主要弊端。当然，并非模型误设就不能使用。例如，分析文本数据的自然语言处理方法（如词典方法、主题模型、词向量模型）都是文本语言的误设模型，但这些误设模型在提取文本数据中的信息时非常有用（Grimmer and Stewart，2013）。

很多经济学理论假说与模型无关（model-free）。比如经典的有效市场假说定义为：

$$E(Y_t|I_{t-1})=E(Y_t)$$

其中，Y_t是某个资产在一个时期的收益率，I_{t-1}是历史信息集合，$E(Y_t)$是无条件期望收益率，$E(Y_t|I_{t-1})$是基于历史信息的预期未来收益率。有效市场假说成立时，历史信息对将来的收益率没有任何预测力。如果要用观测数据验证这一假说，通常需要假设一个预测模型，如线性自回归模型，

$$Y_t=\beta_0+\sum_{j=1}^{p}\beta_j Y_{t-j}+\varepsilon_t$$

然后验证该模型所有滞后项的系数都等于零的统计假设：

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0$$

通过这样的方式将经济假说转变为统计假说，从而可使用计量经济学方法来检验经济假说。但这种方法存在局限性，即如果发现所有滞后项的系数都为零，并不能证明有效市场假说是正确的。因为线性自回归模型只是预测收益率的一种方式，还有无穷多的非线性预测方式。有可能线性自回归模型没有预测能力，但非线性模型有一定的预测能力（Hong and Lee，2003）。因此当不能拒绝统计假说时，只能说线性模型没有发现拒绝有效市场假说的证据，而不能说证实了有效市场假说，除非能穷尽所有的预测模型，但这是做不到的。这就是通常所说的实证研究只能“证伪”，不能“证实”。因此模型证据与数据证据两者之间存在差异。大数据的出现，使我们可采用机器学习的方法，不假设具体的模型或函数形式，而是让数据本身告诉真实的函数关系是什么，从而突破传统低维参数模型的局限性，挖掘更多的数据证据，缩小模型证据和数据证据之间的差异。对大多数传统数据来说，线性模型常比非线性或复杂模型在预测时表现更好，但在大数据条件下，样本容量、变量维度以及噪声都大幅度提高，线性模型无法刻画大数据的非线性、异质性、动态性、离散性等重要特征，而机器学习则能够有效刻画它们并进行精准预测。比如，决策树和随机森林可有效捕捉交互效应等非线性特征。

在宏观计量经济学，以韩德瑞（David Hendry）为代表的计量经济学家，曾提出了“伦敦政经学院计量经济学方法论”，即“LSE Econometric Methodology”（Campos et al.，2005），强调从一般到特殊（from general to specific）的建模方法，即从一个复杂、高维、与数据相吻合的计量经济学模型出发，再利用经济理论与统计推断方法来降维简化模型，以提升模型的经济可解释性和样本外预测能力。这里，经济理论可视为对模型参数的约束，例如在线性自回归模型中，有效市场假说意味着所有滞后项系数为零。这样，便可从一个高维统计模型中得到一个具有经济含义的简约计量经济学模型。也有计量经济学家主张从特殊到一般（from specific to general）的建模方法，即从一个简单的模型开始，逐渐放入新的解释变量，并考虑是否存在非线性关系，通过模型诊断和模型设定检验，最后得到一个适用的计量经济学模型。因为大数据的容量大、变量多，从一般到特殊的方法在大数据情景下可能更有科学性，特别是可减少因为模型误设而产生的系统偏差。需要强调，从一般到特殊的方法仍需要经济理论的指导，特别是在降维和经济解释时。如何将数据驱动方法与经济理论相结合，是数据驱动方法增强其经济可解释性的必由之路。

（二）从参数不确定性到模型不确定性

大数据将实证研究的关注点从参数估计不确定性（parameter estimation uncertainty）转变为模型不确定性（model uncertainty）。传统计量经济学模型常包含低维解释变量与低维未知参数，研究者主要关注未知参数的一致性估计，然后通过t-统计检验量或P-值判断参数估计的统计显著性，进而推测其经济重要性，特别是当某个参数估计值在统计上显著不为零时，研究者将下结论说相应的解释变量是“重要的”。从统计学角度看，t-统计检验量或P-值刻画了参数估计不确定性，这种估计不确定性主要是样本容量有限等原因造成的。在大数据条件下，由于样本容量大，参数估计值十分接近真实的参数值或其概率极限，因此标准误差很小。哪怕真

实参数值非常接近零，以至没有多大的经济重要性，其t值在统计意义上也是非常显著的。换言之，经济重要性与统计显著性不是一回事（洪永淼、汪寿阳，2021b）。在数据容量不大的情形下，实证研究者通常没有区分经济重要性和统计显著性，但在大数据条件下，区分这两者就显得特别重要，因为任何参数估计不确定性在样本容量很大时将大大降低，甚至在实际中可忽略不计。

另一方面，由于大数据特别是胖大数据包含大量潜在的解释变量，可能存在共线性（multicollinearity）或近似共线性，从而导致估计模型出现不确定性。模型不确定性是指当数据出现“微扰”（perturbation），即增加或减少一小部分数据时，基于某一准则（可以是统计准则，也可以是经济准则）的最优估计模型会出现显著变化，比如重要或显著的解释变量集合突然改变了，显示模型对数据的微小扰动具有高度的敏感性。因此，在大数据情形下，需要将注意力从（给定模型下）参数估计不确定性转移到模型不确定性。Varian（2014）指出，很多经济学实证研究包含所谓的“敏感性分析”（sensitivity analysis），即通过假设不同模型设定来检验实证发现的稳健性，实际上是在检验模型不稳定性的影响。从经济预测视角看，当出现模型不确定性时，可将不同的模型进行线性组合或模型平均，以提升样本外预测的稳健度（Bates and Granger，1969；Sun et al.，2021）。从经济学的角度看，可能存在不同的经济理论或模型可解释同一个经济现象，但因为样本数据不多等原因没有办法拒绝其中错误的模型，或者有可能每一个模型可解释现象的一部分，但就像日本20世纪50年代著名电影《罗生门》那样，每个人对于同一个案件都有合乎逻辑的解释，法官则由于证据不足而无法判断谁是真正的杀人凶手。模型不确定性也会影响经济主体的决策行为。Hansen和Sargent（2001）研究了当经济主体对数据生成过程（即产生数据的真实模型）存在一定程度的不确定性判断时，这种模型不确定性或模型模糊性（model ambiguity）如何影响经济主体的决策行为。

（三）从无偏估计到正则化估计

经济学实证研究主要是识别与推断经济因果关系，很多传统的统计推断方法均基于无偏估计。一个例子是经典的低维线性回归模型：

$$Y_i = X'_i \beta^0 + \varepsilon_i ,\ i = 1, 2, \cdots, n$$

其普通最小二乘法（OLS）估计量以及相应的残差方差估计量均为无偏估计。常用的统计推断方法，如经典的t-检验和F-检验，均基于这些无偏估计量。但无偏估计不一定是最优估计。随着大数据的广泛使用，可能出现很多解释变量，当解释变量维数较高时，有较大概率会存在近似共线性，导致OLS估计不稳定，即OLS估计量的方差很大。如果对参数施加一定约束，通过牺牲无偏性质，换取估计方差的显著减少，这将显著减少均方误差，提高预测精准度。一个例子是Hoerl和Kennard（1970）提出的岭回归（ridge regression），其损失函数定义为：

$$\min_{\{\beta_j\}} \sum\nolimits_{i=1}^{n} \left(Y_i - X'_i \beta\right)^2 + \lambda \sum\nolimits_{j=0}^{p} \beta_j^2$$

参数估计量为$\hat{\beta}=(X'X+\lambda I)^{-1}X'Y$，其中，$I$是单位矩阵。这个估计量不是无偏估计量，但其解存在且比较稳定。从本质上说，岭回归通过约束未知参数值的大小，以牺牲无偏性换取方差的显著减少，从而改进预测效果。在大数据时代，经常使用机器学习进行预测（包括分类），其基本思想是将数据分成两个子集，一个是训练数据（training data），用于训练算法；一个是测试数据（test data），用于测试算法的样本外预测（out-of-sample prediction）能力或泛化（generalization）能力。为了获得较好的泛化能力，机器学习通常引入一个惩罚项，限制算法的复杂度，这实际上是在算法预测的方差与偏差之间，取得一个适当的平衡。因此，算法预测大多是有偏估计。目前，统计学家与计量经济学家正在将机器学习应用于政策评估等统计推断中（Athey and Imbens，2019）。关于基于有偏估计量的统计推断方法，需要系统地建立一套新的统计学与计量经济学理论（Lee et al.，2016）。

（四）从样本内拟合到样本外预测

任何一种经济理论的生命力取决于其对经济现实的解释力，特别是其所揭示的因果关系的解释力。经济

学传统建模与经验解释大多基于样本内拟合(in-sample goodness of fit)。然而,任何一种科学理论或假说,必须能够在同样的条件下,独立地重复通过经验验证。因此,一种科学理论或模型不但需要能够解释已经发生的现象,更重要的是能够进行精准的样本外预测,即拥有良好的泛化能力。在实际应用中,样本内拟合和样本外预测之间也存在一个权衡的问题。一般而言,一个模型越复杂,其样本内拟合越好。但是,一个模型的样本外预测能力如何,取决于它是否能够捕捉不同数据中的共同特征(即通常所说的"信号")。不同数据的共同特征越多,或模型捕捉共同特征的能力越强,其样本外预测能力越好。例如,机器学习依靠非参数统计方法,具有强大样本内拟合的能力,但这并不能保证样本外精准预测。一种高度灵活的机器学习算法,不但能够捕捉数据中的"信号",而且还会捕捉数据中无助于样本外预测的"噪声",从而导致样本内过拟合。为了改善样本外预测精准度,必须限制模型复杂度,这就需要对模型进行正则化(regularization)。

正则化通过限制参数值或参数维度或模型复杂性,减少捕捉训练数据中的"噪声",避免算法的过拟合,以获得良好的样本外预测。大部分经济决策(如消费、投资)是在不确定市场条件下所做的决策,均基于样本外预测,因此良好的样本外预测能力十分重要。由于经济结构常常具有时变性,以前表现优越的模型不一定能够继续精准预测未来。此外,经济主体的理性预期使经济主体会随政策变化而改变其行为,从而导致政策失效(Lucas,1976)。因此,精准的样本外预测具有很大的挑战性。在实证研究中,经常看到一些模型具有很显著的样本内证据(如预测变量的参数估计值很显著),但样本外预测能力则很弱。但是,任何科学理论或假说,都必须建立在可靠、可重复验证的实证基础之上。可重复验证意味着在相同的条件下,任何科学理论或假说都应该有很好的样本外预测能力,而不仅仅是有很好的样本内拟合。Varian(2014)指出,随着大数据可获得性的增强,经济学的实证研究在检验经济理论的有效性时,将会更多地从样本内拟合转变到样本外预测。Hofman 等(2021)提出了在计算社会科学领域兼顾解释与预测的整合建模(integrative modelling)思想。

(五)从低维建模到高维建模

传统计量经济学模型大多是低维模型,即解释变量维数小,未知参数维数也小。低维模型存在模型误设的可能性,如遗漏重要的解释变量。而大数据特别是胖大数据提供了大量潜在的解释变量,其维数甚至比样本容量更大,这给计量经济学建模带来很大挑战,但也提供了巨大的灵活性,可显著减少因模型误设而引起的系统偏差,避免遗漏重要的解释变量。事实上,很多经济金融问题涉及高维潜在的经济变量。高维建模将所有潜在的解释变量放进模型中,再用统计方法排除不重要的解释变量,实现有效降维,从而达到识别重要解释变量、增强模型可解释性、提升预测稳健性与精准度等目的。

高维建模思想可用于金融学中的高维投资组合选择问题,比如假设要从标准普尔500中选择30只股票进行投资,如何在每个时期选择最重要的30只股票并决定其最优组合权重,是一个降维问题。再以异质性资本资产定价模型为例:

$$Y_{it}=\alpha_i+\beta'_i X_t+\gamma'_i Z_{it}+\varepsilon_{it}, i=1,2,\cdots,n; t=1,2,\cdots,T$$

其中,Y_{it}是资产i在时期t的回报率,X_t是影响所有资产价格的共同风险因子,而Z_{it}是特质风险因子,只与资产i密切相关。一般情形下,X_t和Z_{it}的维度都不高,但不同的资产i有不同的特质风险因子。如何从包括所有潜在的共同风险因子和所有资产特质风险因子的高维风险因子集合中,识别出共同风险因子和每个资产的特质风险因子,是一个降维问题。再以多元波动率模型估计(Cui et al.,2021)为例,假设有p个资产,则刻画其时变波动率与相关性的条件方差—协方差矩阵的维数为$p\times p$,需要估计的未知参数个数可高达$3p^2+3p$个。自Engle 和 Kroner(1995)以来,如何在保证条件方差—协方差矩阵半正定性的前提下,有效估计多元波动率模型的未知参数值,一直是金融计量学的一个难题。

如何对高维模型进行降维,解决所谓的"维度灾难"(curse of dimensionality)问题?岭回归没有降维功能,但 Tibshirani(1996)提出的统计学习方法 LASSO 可用于选择重要的解释变量,达到降维目的。假设存在稀疏性,即在大量潜在的解释变量中,只有少数变量的系数不为零。在这种情形下,可考虑如下最小化问题:

$$\min_{\{\beta_j\}} \sum_{i=1}^{n}\left(Y_i - X'_i\beta\right)^2 + \lambda\sum_{j=0}^{p}\left|\beta_j\right|$$

这个方法称为LASSO，由于对未知参数值的约束从原来岭回归的L2范数（参数平方和约束）改变为L1范数（参数绝对值加总约束），LASSO会令数值很小的系数直接为零，从而达到降维的目的。当样本容量足够大时，LASSO将以大概率正确识别重要的解释变量，同时排除所有其他不重要的解释变量。机器学习的基本思想类似于LASSO，但有两个显著不同。首先，机器学习一般不用线性回归模型，而是采用非参数分析方法，即让数据挑选最优的函数关系，因此具有很大灵活性，可避免模型误设而导致的系统偏差。其次，由于非参数方法的灵活性，存在对数据过拟合的可能性。为了改进样本外预测精确度，机器学习将数据分为训练数据和测试数据，其中训练数据用于决定算法结构，而测试数据用于检验样本外预测效果。

非参数方法可有效刻画非线性关系，如边际递减或递增效应、交互效应等，但也存在“维数灾难”，特别是当存在高维潜在的解释变量时。为了解决这个问题，机器学习采用了类似LASSO的惩罚项，实现有效降维和避免过拟合。这种带有约束的统计优化问题称为正则化，通过限制模型复杂性，在偏差与方差之间取得适当平衡，以提升预测精准度。这种思想广泛应用于决策树、随机森林、人工神经网络、深度学习等机器学习方法中。需要强调，正则化并不一定都对高维参数施加稀疏性假设。例如，在估计多元波动率模型时，直接假设参数稀疏性并不能保证时变方差—协方差矩阵的半正定性，在这种情形下，可假设未知参数矩阵是低秩的（low rank），即假设很多参数行可表示为少数参数行的线性组合，这样既可实现降维估计，又能保证矩阵的半正定性（Cui et al.，2021）。

高维问题或“维数灾难”并不是统计学与计量经济学所特有的现象。例如，在微观经济学中，包含大量经济主体（或博弈者）的超大型博弈（large games）问题的求解也面临维数灾难问题。在宏观经济学中，当状态变量维数变大或服从非马尔科夫过程时，刻画随机动态最优规划的贝尔曼方程（Bellman equation）的数值求解也存在维数灾难问题。其他学科如物理学和应用数学，多元偏微积分方程的数值求解在变量维数增加时也面临同样的难题（E，2021）。如果拥有大数据，机器学习特别是深度学习将是解决上述高维求解难题的一个有效方法。

（六）从低频数据到高频数据

大数据的一个显著特点是其动态性，即产生高频数据甚至实时数据。高频与超高频金融数据的可获得性催生了高频金融计量学（Engle and Russell，1998；Engle，2000）和高频微观金融学（如市场微观结构（market microstructure）金融学，参见O’Hara，1995）。20世纪90年代，Engle和Russell（1998）基于高频与超高频金融交易数据，提出了一个自回归条件久期（autoregressive conditional duration，ACD）模型，用于刻画资产价格变动或交易的时间间隔与历史信息之间的动态关系，这类模型的产生得益于高频金融数据的可获得性。

由于不能实时监测GDP等宏观经济变量，宏观经济学研究长期以来受到低频数据的限制。实时预测（nowcasting）原是气象学的一个术语。Giannone等（2008）提出了利用大数据实时预测当期GDP的方法，即在季度GDP数据发布之前，利用实时更新的数据预测当期GDP，其基本思想是将大量的异质数据（如失业率、工业销售、贸易差额等）作为信息源，在传统季度GDP数据发布前从中提取出有关当期GDP变化的信息。美联储每天都在利用高频大数据预测当期季度的GDP增长率和通货膨胀率，这对美联储制定货币政策可提供很大帮助。

随着高频微观经济数据的产生，很多宏观经济指标都能实现高频化甚至实时化，比如，可用互联网消费价格大数据构建日度CPI数据。一个例子是美国麻省理工学院（MIT）的研究项目（Billions Price Project）所构建的美国和阿根廷的日度CPI指数（Cavallo，2012，2013）。Scott和Varian（2014，2015）使用谷歌搜索数据构建了重要宏观经济变量的高频数据，包括失业人数、消费零销额、消费者情感指数等，以往这些变量只能通过统计调查构建低频数据。预计高频宏观经济数据的可获得性将催生一门新兴学科——高频宏观经济学。宏观实体经济与金融市场高度相关。金融市场有高频数据，但长期以来宏观经济指标数据的获得相对滞后，因此研

究者没有办法研究实体经济与金融市场之间的即时互动关系。如果宏观经济变量能够高频化，那么这种研究将成为可能。除了用于构建高频宏观经济指标之外，高维大数据在识别外生经济或政策冲击对不同行业、不同经济主体的分布效应，以及宏观经济政策的传导机制等方面具有天然优势。

比如，可用高频金融市场大数据精准识别货币政策冲击。针对特定的货币政策工具（如利率），利用“高频”数据（以日为频率）估计货币政策执行前后金融市场价格（反映了市场对政策的预期）的变化，并利用胖大数据控制其他高维因素，识别没有预期到的外生政策冲击（Gertler and Karadi，2015）。较之宏观计量经济学的结构向量自回归模型，上述方法能够更精准识别外生货币政策对金融市场的冲击。

再比如高频微观行为大数据（如家庭在线消费和企业的投资），可用于识别宏观经济政策对家庭消费与企业投资的分布效应。异质性主体新凯恩斯（heterogeneous-agent new Keynesian）理论认为货币政策冲击会对面临不同约束（如信贷约束）的微观家庭产生异质性影响，从而导致政策具有分布效应并影响其传导机制。分析微观层面的家庭消费与投资在货币政策实施前后的动态变化，可精准刻画货币政策对不同家庭冲击的分布效应及其背后的市场摩擦机制。同样地，企业投资大数据可用于刻画宏观经济政策（如信贷供给）对微观层面的异质性企业投资行为的分布效应，从而为制定精准信贷政策提供科学依据。

基于高频的企业生产与销售数据，可估计重大外生冲击（如新冠疫情、中美贸易冲突）发生后，同一产业内不同企业之间的动态关联，以及不同产业之间的动态关联，刻画重大冲击的产业网络或产业链传导机制，特别是对系统性重要产业和核心企业的识别，这将有助于制定科学的定向经济复苏政策（如定向信贷供给和政策补贴），提升产业链的稳定性与韧性，有效降低系统性风险，增强扩张性政策的有效性。

（七）从结构化数据到非结构化数据

大数据包括结构化数据和非结构化数据，后者不能以传统的行—列格式表示。非结构数据包括文本、图像、视频、音频等，可用于定量刻画结构化数据无法描述的社会经济活动与现象，如群体心理、企业文化、经济政策不确定性等。非结构化数据一般是高维的。例如，从统计学视角看，文本数据是一种高维的复杂数据。假设一个文件包含10000个汉字，每个汉字从500个最常用的中文字库中提取，则完全表示这个文件的维度将高达$10^{4\times500}$！如果去掉最常用和最不常用的汉字以及标点符号，假设共剩下3000个汉字以及每个汉字在文件中出现的频率，则需要用一个3000×2维度的矩阵来表示，维数还是很大。因此，分析非结构化数据的第一步通常是借助深度学习等人工智能方法，例如，利用自然语言处理技术获取文本中的语义学信息，利用语音识别（speech recognition）确定声音和音频中的声调，以及通过计算机视觉（computer vision）提取图像和视频蕴含的地理信息等。

以文本数据为例，各种政府工作报告与政策文件、各类新闻报道、社交媒体平台的各种评论等都是文本数据。文本数据的现代统计分析可追溯到Mosteller和Wallace（1963）。他们通过分析《联邦党人文集》（The Federalist Papers）中每篇文章中的冠词（如“an”、“of”、“upon”）出现的频率，并基于每个人写作习惯不会轻易改变的假设，分辨出《联邦党人文集》中一些原来作者不明的文章的作者是詹姆斯·麦迪逊（James Madison），而非亚历山大·汉密尔顿（Alexander Hamilton）。在计量经济学史上，对谁发明工具变量法，计量经济学界有过争议。关于工具变量估计的推导最早出现在Wright（1928）所著的《动物油与植物油关税》一书的附录，但附录的写作风格与正文完全不同。Stock和Trebbi（2003）对文本数据进行主成分分析，并使用前4个主成分作为预测变量，最终得出结论，即工具变量估计的提出者是Philip Wright而非他的儿子Sewall Wright。在中国，也早有学者基于《红楼梦》文本数据所包含的常用副词，用统计学两样本均值检验方法研究《红楼梦》前80回的作者和后40回的作者是否为同一个人。

文字语言是人类表达思想、情感，进行沟通、交流的最主要工具，因此可从文本数据中提取有用信息，测度各种社会心理变量，如金融学中的投资者情感指数（Tetlock，2007；García，2013）、福利经济学中的国民幸福感指数（张兴祥等，2018）、市场营销学中的顾客满意度指数（He et al.，2013；Homburg et al.，2015）、经济学中的经济政策不确定指数（Brogaard and Detzel，2015；Baker et al.，2016；Gulen and Ion，2016；Baker et al.，2020）、教育

学中的学生学习压力指数(Munezero et al.,2013)以及新闻传播学中的社会舆情指数等。

还可基于文本数据构建与测度文化变量。文化是人类社会相对于经济、政治而言的精神活动及其产物,分为物质文化和非物质文化,非物质文化是长期形成的社会心理与行为习惯,可通过文本数据进行刻画。例如,可测度诸如创新(innovation)、正直(integrity)、质量(quality)、敬畏(respect)和团队协作(teamwork)之类的企业文化(Li et al.,2021)。在Graham等(2017)的访谈研究中,企业高管们推荐了11个度量文化的数据来源,其中大多数是非结构化数据,如财报电话会议记录。Li等(2021)通过自然语言处理技术对企业文化进行研究,他们使用5个标准普尔500公司网站中最常提到的词汇作为"核心价值词汇",包括"创新"、"正直"、"质量"、"敬畏"、"团队协作",并借用Guiso等(2015)所提供的与各个"核心价值词汇"相关的"种子词汇",将财报会议记录中的词语与"种子词语"联系起来,建立异质性的企业"文化字典",并在每一财务年度为每个企业文化指标赋值,其中每个文化指标的得分是其相关词语的加权计数占总词数的比例。Li等(2021)突破以往企业文化研究主要使用代理变量或采用调查访谈的做法,使用词向量模型度量文化。词向量模型突破传统的词袋模型将字词视为相互独立符号的假设,避免或减少了忽视上下文语境而导致的偏差,将语法表达层面的定量方法推进到语义层面。测度好各种文化指标后,可将这些指标代入回归模型中,使原来的定性分析转变为定量分析。

需要指出,中文文本数据的定量分析难度高于英文文本数据。例如,与能够自动分词、断句的英文文本数据相比,中文文本数据的分词、断句的位置不同可能产生截然不同的含义,一个经典的例子是:"下雨天,留客天。天留我不留。"与"下雨天,留客天。天留我不?留。"另外,一些中文关键词的词性在上下文中会发生变化,如"领导"可以是名词,也可以是动词。因此,中文词性的判断往往需要一定程度的深度学习和较为庞大的训练数据。还有,中文是不断进化的语言。完全相同的词汇,可能在短短数年间,其含义便发生巨大变化,特别是大量网络语言不断涌现,这些词汇往往代表强烈的感情色彩,但无法按照常规的中文语句含义进行分析。

文本回归分析不仅使经济学与人文社会科学的跨学科交叉研究成为可能,也使系统性的人类经济社会研究成为可能。众所周知,经济只是人类社会的一个组成部分(当然,是重要组成部分),除了经济因素的影响外,人类的经济活动还受到政治、法律、科技、历史、文化、社会与自然环境等因素的深刻影响,并且反过来影响这些因素。习近平总书记指出,"系统观念是具有基础性的思想和工作方法。"经济学研究也需要坚持系统分析方法。跨学科跨领域的大数据特别是文本数据,可为人类经济社会的系统研究提供很多新的洞见和发现。可以预见,基于大数据的文本回归分析将成为经济学与人文社会科学一个基本的定量实证研究方法(洪永淼、汪寿阳,2021a)。Grimmer和Stewart(2013),Evans和Aceves(2016),Loughran和McDonald(2016)以及Gentzkow等(2019)分别介绍了文本数据的一些基本分析方法及其在政治学、社会学、会计学与金融学,以及经济学实证研究中的应用。

(八)从传统结构化数据到新型结构化数据

除了非结构化数据外,大数据还包括新型结构化数据。新型结构化数据例子包括矩阵数据(matrix data)、函数数据(functional data)、区间数据(interval data)以及符号数据(symbolic data),其中向量数据是矩阵数据的一个特例,区间数据是符号数据的一个特例,而面板数据则是函数数据的一个特例。长期以来,很多经济金融数据所包含的信息没有得到充分利用。比如,在金融波动率建模时,人们通常只使用金融资产每天的收盘价数据,而由金融资产每天的最高价和最低价所组成的价格区间数据,或者其每天从开盘到收盘的函数价格数据,所包含的信息要比每天的收盘价丰富得多,但却长期没有得到有效利用。作为一个实际应用的例子,股市投资中的K线预测可视为部分利用区间数据进行交易的技术投资策略。K线反映了各种股票每日、每周、每月的开盘价、收盘价、最高价、最低价等涨跌变化情况(Xie et al.,2021)。Chou(2005)提出一个基于范围(range,即最高价减最低价)数据的条件自回归范围(conditional autoregressive range)模型,发现基于范围数据的波动率预测优于基于收盘价的GARCH波动率模型预测。而He等(2021)和Zhu等(2021)使用自回归区间模型(Han et al.,2021)和门框自回归区间模型(Sun et al.,2018),分别发现在预测月度原油价格波动率和每天外汇

市场波动率时,区间模型预测优于范围模型,而范围模型又优于基于点数据的GARCH模型,展现了有效利用区间数据信息可显著改进波动率预测的信息优势(区间数据既包含范围信息,也包含中点价(midpoint)和收盘价信息)。关于区间数据建模与预测的更多讨论,参见洪永淼和汪寿阳(2021a)。

新型结构化数据比传统点数据提供更加丰富的信息,但新型结构化数据建模需要新的分析方法与工具,比如一个区间是无穷多点的集合,因此需要构建随机集合的计量经济学模型,而不是点数据的计量经济学模型(Han et al.,2021;Sun et al.,2018)。对新型结构化数据建模需要新的数学工具,这将给计量经济学研究带来范式变革。

(九)从人工分析到智能化分析

由于大数据的海量性和复杂性(如不同结构、不同频率、不同来源、噪声等),由人工收集、储存、处理与分析大数据是极其困难甚至不可能的。人工智能,特别是机器学习,也因此应运而生,并得到了空前大发展。机器学习,如深度学习,是分析大数据的最主要工具,已广泛应用于各种现实经济活动中,如高频算法交易。MIT最近开发了一个PClean数据清洗系统(Lew et al.,2021),可自动清洗脏数据,如错误、数值缺乏、拼写错误和数值不一致等常见的数据问题。据报道,在中国杭州市余杭区,"统计机器人"正在帮助及时收集各个部门、各个单位的统计数据报送。机器学习也正在应用于经济学研究中,特别是基于大数据的经济学实证研究,例如文本数据的情感分析需要使用各种自然语言处理方法与技术。人工智能可应用于自然语言处理、计算机视觉、语音识别以及商业智能分析。计量经济学家正在发展一些新的基于机器学习的因果识别与政策评估方法,用于精确评估经济社会公共政策效应(Athey and Imbens,2019)。中国人工智能之父吴文俊曾长期研究如何用机器来证明数学定理。机器人现在还可以帮助科学家做科学实验和写学术论文。

大数据与人工智能的发展对经济学者的编程能力和数据分析素养带来了新的挑战。比如,为处理海量大数据和及时获取最新算法,经济学家需要掌握一些难度较高的开源可编译软件(如Python,R,Java,C++等),并熟悉诸如GitHub、码云等代码共享平台。再比如,若数据量超过一定规模,在单独服务器上使用计算软件进行数据分析将变得不再可行,这时需要进行分布式计算,将庞大的工作量分散到多个节点服务器分别进行,最后再进行汇总。因此,研究人员也需要熟练掌握如Hadoop、Storm等分布式计算软件。

五、结束语

本文的分析表明,大数据正在深刻改变经济学的研究范式与研究方法。由于大数据包含大量互相关联的异质性微观主体的行为(包括心理)信息,使经济学家能够从实证研究的视角出发,突破现代西方经济学的一些基本假设的局限性,如假设完全理性经济人而忽视非理性行为因素,忽视经济人的社会性与社会心理的反作用,忽视宏观经济学的微观基础,忽视以系统观点将经济活动放在更广泛的人类社会系统中来研究经济等重要缺陷,同时,大数据也促进了经济学与认知科学、人工智能、计算机编程学、数据科学等相关领域之间的交叉,特别是促进了这些新兴方法论学科在经济学与社会科学其他领域中的应用,从而推动了经济学与社会科学其他领域之间以及经济学与数学、自然科学之间的融合。经济学与社会科学其他领域的实证研究范式正呈现出科学化、严谨化、精细化、多元化(跨学科)、系统化与趋同化(方法论)的趋势。一个新兴方法论学科,即大数据与机器学习计量经济学正在兴起。需要重视和学习交叉学科和跨学科的理论与方法,包括各种大数据分析方法、技术与工具。

应该强调,不是使用了定量实证研究方法,经济学研究便自动具有科学性。任何定量实证研究方法,都有其适用的前提条件,如果这些前提条件不满足,相应的方法便不适用。例如,不管样本容量有多大,经典的t-检验和F-检验在条件异方差情形下便会失效(洪永淼,2021)。此外,与任何其他研究方法一样,定量实证研究方法也有其缺点。例如,当使用文本数据测度社会心理变量和经济政策不确定性时,不仅所使用的自然语言处理方法均基于语言的误设模型,而且还可能有不同的构建方式(如赋予不同权重),存在一定的随意性。由于从文本数据构造的变量大多是解释变量,自然语言处理方法所用的误设语言模型会产生变量误差(errors

in variables),导致估计偏差,因此需要使用工具变量等方法加以矫正(洪永淼,2011)。另一方面,在实证研究中,通常是研究者事先提出一个理论假说,然后设计一个实验或选择一个方法来检验该假说。不管是拒绝或接受理论假说,研究人员不会事先预知结果。但是,如果为了获得某个预期结果而提出适合该结果的理论假说,并且反复从数据中寻找"证据"支持,这将可能导致数据窥视(data snooping)偏差(Campbell et al.,1997)。例如,研究人员可能会对一种算法的不同版本在同一数据进行反复试验,直至获得某种符合预期结果的版本。这不是科学的态度与方法。但是,不能因此就放弃定量分析而退回到定性分析;相反地,应该研究如何改进测量社会心理变量的方法,如何减少或避免数据窥视偏差。事实上,10年来,分析文本数据的自然语言处理统计方法已显著地变得更加精准,并且还在不断完善中。

另一方面,也不能说不用定量方法就没有科学性。逻辑分析、历史分析不一定非用数学和其他定量方法不可。但是,在大数据时代,海量大数据包含很多传统数据所没有的信息,特别是大量互相关联的微观主体行为信息,这些信息可用于揭示个人与群体的行为,个人之间与群体之间的关系,以及宏观经济运行的规律。在这种情况下,不采用定量方法是不可想象的。定量分析并不意味着一定要使用高深的数学和复杂的模型,而且需要注意模型的可解释性(特别是经济解释)与数据分析的可视化。实证研究特别是定量实证研究是现代经济学最主要的研究范式,但也只是一类研究范式。不同的研究范式或研究方法都有其合理性和局限性,需要兼容并包。应当鼓励使用多元的研究范式和研究方法,互相补充、互相交叉、互相促进、共同提高中国经济学研究的科学性与先进性。

中国经济是中国特色社会主义市场经济,以公有制为主体、多种经济成分并存,市场在资源配置上发挥决定性作用,同时政府发挥重要作用。中国经济经过40多年持续快速增长,成为世界第二大经济体、最大制造业国家、最大货物贸易国、全球三大主要供应链中心之一,并且即将成为全球最大消费国,中国经济崛起是21世纪上半叶世界最重要的经济事件,已经并且正在深刻影响世界经济格局的发展趋势。从中国经济实践中揭示中国经济发展规律,凝练可复制的中国经济发展模式,构建具有深厚学理基础的原创性中国经济理论体系,是中国经济学家的历史机遇与时代责任。由于超大经济体的规模优势,以及中国政府"互联网+"政策,中国数字经济发展迅速,在某些领域(如移动支付)领先全球,中国在大数据资源方面与西方主要发达国家处于同一起跑线,并且拥有巨大潜力。海量大数据资源,加上中国数字经济的快速发展、中国经济所有制的多样性以及全球最具特色的"政策数据库"等得天独厚的优势,为中国经济学家开展以大数据为基础的定量实证研究,探索中国经济发展规律、数字经济运行规律、政府与市场之间关系等重要理论与现实问题,提供了一个可以产生重大理论创新成果的"富矿"(陈国青等,2021)。

更重要的是,大数据的出现,使中国经济学家可以克服现代西方经济学研究范式的一些根本性缺陷,并从中国经济实践中提炼出新的带有普遍性的经济知识体系,为当代世界经济学的发展做出中国经济学家应有的贡献。同时,新型数据需要新的研究方法与工具,需要不断创新基于大数据的实证研究方法,并应用于研究各种现实经济问题,包括以证据为基础精准评估经济社会公共政策,提升政策制定的科学性、精确性、时效性与协同性,从而更好支持政府科学决策。

在构建原创性中国经济理论过程中,还应坚持国际学术交流与合作,批判性借鉴现代西方经济学中有益的理论成分与研究方法,以科学研究范式分析中国经济问题,用国际语言讲述中国经济故事,不断加强中国经济学的国际学术影响力。

(作者单位:中国科学院数学与系统科学研究院、中国科学院大学经济与管理学院、中国科学院预测科学研究中心)

参考文献

(1)陈国青、张瑾、王聪、卫强、郭迅华:《"大数据—小数据"问题:以小见大的洞察》,《管理世界》,2021年第2期。

(2)崔丽媛、洪永淼:《投资者对经济基本面的认知偏差会影响证券价格吗?——中美证券市场对比分析》,《经济研究》,2017年第8期。

(3)洪永淼:《计量经济学的地位、作用和局限》,《经济研究》,2007年第5期。

(4)洪永淼:《高级计量经济学》,高等教育出版社,2011年。
(5)洪永淼:《理解现代计量经济学》,《计量经济学报》,2021年第2期。
(6)洪永淼、汪寿阳:《数学、模型与经济思想》,《管理世界》,2020年第10期。
(7)洪永淼、汪寿阳:《大数据革命和经济学研究范式与研究方法》,《财经智库》,2021年a第1期。
(8)洪永淼、汪寿阳:《大数据、机器学习与统计学:挑战与机遇》,《计量经济学报》,2021年b第1期。
(9)洪永淼、汪寿阳:《非参数统计学与机器学习:基本思想、方法及相互关系》,工作论文,2021年c。
(10)洪永淼、汪寿阳、任之光、薛涧坡、钟秋萍、钟锃光:《"十四五"经济科学发展战略研究的背景与论证思想》,《管理科学学报》,2021年第2期。
(11)洪永淼、薛涧坡:《中国经济发展规律与研究范式变革》,《中国科学基金》,2021年第3期。
(12)侯增谦:《研究中国经济发展规律,促进经济高质量发展》,《中国科学基金》,2021年第3期。
(13)胡毅、陈海强、齐鹰飞:《大数据时代计量经济学的新发展与新应用——第二届中国计量经济学者论坛(2018)综述》,《经济研究》,2019年第3期。
(14)李子奈:《计量经济学应用研究的总体回归模型设定》,《经济研究》,2008年第8期。
(15)李子奈、霍玲:《从〈经济研究〉与AER发文比较分析看计量经济学教学与研究》,《21世纪数量经济学》,2005年第6期。
(16)刘伟、蔡志洲:《中国经济发展的突出特征在于增长的稳定性》,《管理世界》,2021年第5期。
(17)王东京:《中国经济体制改革的理论逻辑与实践逻辑》,《管理世界》,2018年第4期。
(18)王美今、林建浩:《计量经济学应用研究的可信性革命》,《经济研究》,2012年第2期。
(19)王一鸣:《中国经济新一轮动力转换与路径选择》,《管理世界》,2017年第2期。
(20)杨红丽、刘志阔、陈钊:《中国经济的减速与分化:周期性波动还是结构性矛盾?》,《管理世界》,2020年第7期。
(21)杨耀武、张平:《中国经济高质量发展的逻辑、测度与治理》,《经济研究》,2021年第1期。
(22)张兴祥、钟威、洪永淼:《国民幸福感的指标体系构建与影响因素分析:基于LASSO的筛选方法》,《统计研究》,2018年第11期。
(23)Angrist, J., Azoulay, P., Ellison, G., Hill, R. and Lu, S., 2017, "Economic Research Evolves: Fields and Styles", *American Economics Review*, 107(5), pp.293~297.
(24)Angrist, J. D. and Pischke, J. S., 2009, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.
(25)Athey, S., 2019, "The Impact of Machine Learning on Economics", in Agrawal, A., J. Gans and Goldfarb, A., eds: *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, Chicago.
(26)Athey, S. and Imbens, G. W., 2019, "Machine Learning Methods That Economists Should Know About", *Annual Review of Economics*, 11, pp.685~725.
(27)Baker, S., Bloom, N. and Davis, S. J., 2016, "Measuring Economic Policy Uncertainty", *Quarterly Journal of Economics*, 131(4), pp.1593~636.
(28)Baker, S., Bloom, N., Davis, S. J. and Terry, S., 2020, "COVID-Induced Economic Uncertainty", Working Paper.
(29)Bates, J. M. and Granger, C. W., 1969, "The Combination of Forecasts", *Journal of Operational Research Society*, 20(4), pp.451~468.
(30)Breiman, L., 2001, "Statistical Modeling: The Two Cultures", *Statistical Science*, 16(3), pp.199~215.
(31)Brogaard, J. and Detzel, A., 2015, "The Asset-Pricing Implications of Government Economic Policy Uncertainty", *Management Science*, 61(1), pp.3~18.
(32)Campbell, J. Y., Lo, A. W. and MacKinlay, A. C., 1997, *The Econometrics of Financial Markets*, Princeton: Princeton University Press.
(33)Campos, J., Ericsson, N. R. and Hendry, D. F., 2005, "General-to-Specific Modeling: An Overview and Selected Bibliography", FRB International Finance Discussion Paper, No.838.
(34)Cavallo, A., 2012, "Scraped Data and Sticky Prices", MIT Sloan Working Paper.
(35)Cavallo, A., 2013, "Online and Official Price Indexes: Measuring Argentina's Inflation", *Journal of Monetary Economics*, 60(2), pp.152~165.
(36)Chou, R. Y., 2005, "Forecasting Financial Volatilities with Extreme Values: The Conditional Autoregressive Range (CARR) Model", *Journal of Money, Credit and Banking*, 37(3), pp.561~582.
(37)Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M. and Helbing, D., 2012, "Manifesto of Computational Social Science", *European Physical Journal Special Topics*, 214(1), pp.325~346.
(38)Cui, L., Hong, Y. and Li, Y., 2021, "Solving Euler Equations via Two-Stage Nonparametric Penalized Splines", *Journal of Econometrics*, 222(2), pp.1024~1056.
(39)E, W., 2021, "The Dawning of a New Era in Applied Mathematics", *Notice of American Mathematical Society*, 68(4), pp.565~571.
(40)Einav, L. and Levin, J., 2014, "Economics in the Age of Big Data", *Science*, 346(6210), Article ID: 1243089.
(41)Engle, R. F., 2000, "The Econometrics of Ultra-High-Frequency Data", *Econometrica*, 68(1), pp.1~22.
(42)Engle, R. F. and Kroner, K. F., 1995, "Multivariate Simultaneous Generalized ARCH", *Econometric Theory*, 11(1), pp.122~150.
(43)Engle, R. F. and Russell, J. R., 1998, "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data",

Econometrica, 66(5), pp.1127~1162.

(44)Evans, J. A. and Aceves, P., 2016, "Machine Translation: Mining Text for Social Theory", *Annual Review of Sociology*, 42, pp.21~50.

(45)García, D., 2013, "Sentiment during Recessions", *Journal of Finance*, 68(3), pp.1267~1300.

(46)Gentzkow, M., Kelly, B. and Taddy, M., 2019, "Text as Data", *Journal of Economic Literature*, 57(3), pp.535~574.

(47)Gertler, M. and Karadi, P., 2015, "Monetary Policy Surprises, Credit Costs, and Economic Activity", *American Economic Journal: Macroeconomics*, 7(1), pp.44~76.

(48)Giannone, D., Reichlin, L. and Small, D., 2008, "Nowcasting: The Real-Time Informational Content of Macroeconomic Data", *Journal of Monetary Economics*, 55(4), pp.665~676.

(49)Goldberg, A., Srivastava, S. B., Manian, V. G., Monroe, W. and Potts, C., 2016, "Fitting in or Standing out? The Tradeoffs of Structural and Cultural Embeddedness", *American Sociological Review*, 81(6), pp.1190~1222.

(50)Graham, J. R., Campbell, R. H., Jillian, G. and Shivaram, R., 2017, "Corporate Culture: Evidence from the Field", NBER Working Paper, No.w23255.

(51)Granger, C. W. J., 1980, "Long Memory Relationships and the Aggregation of Dynamic Models", *Journal of Econometrics*, 4(2), pp.227~238.

(52)Grimmer, J. and Stewart, B. M., 2013, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", *Political Analysis*, 21(3), pp.267~97.

(53)Guiso, L., Sapienza, P. and Zingales, L., 2015, "The Value of Corporate Culture", *Journal of Financial Economics*, 117(1), pp.60~76.

(54)Gulen, H. and Ion, M., 2016, "Policy Uncertainty and Corporate Investment", *Review of Financial Studies*, 29(3), pp.523~564.

(55)Hamermesh, D. S., 2013, "Six Decades of Top Economics Publishing: Who and How?", *Journal of Economic Literature*, 51(1), pp.162~72.

(56)Han, A., Hong, Y., Wang, S. and Sun, Y., 2021, "Conditional Autoregressive Models for Interval-Valued Time Series Data", Working Paper, Center for Forecasting Science, Chinese Academy of Sciences.

(57)Hansen, L. and Sargent, T. J., 2001, "Robust Control and Model Uncertainty", *American Economic Review*, 91(2), pp.60~66.

(58)He, Y., Han, A., Hong, Y., Sun, Y. and Wang, S., 2021, "Forecasting Crude Oil Price Intervals and Return Volatility via Autoregressive Conditional Interval Models", *Econometric Review*, 40(6), pp.584~606.

(59)He, W., Zha, S. and Li, L., 2013, "Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry", *International Journal of Information Management*, 33(3), pp.464~472.

(60)Hoerl, A. E. and Kennard, R. W., 1970, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12(1), pp.55~67.

(61)Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A. and Yarkoni, T., 2021, "Interpreting Explanation and Prediction in Computational Social Science", *Nature*, 595, pp.181~188.

(62)Hofstede, G., 1984, *Culture's Consequences: International Differences in Work-Related Values*, Beverly Hills: Sage Publications.

(63)Hofstede, G., 1991, *Cultures and Organizations: Software of the Mind*, New York: McGraw Hill.

(64)Homburg, C., Ehm, L. and Artz, M., 2015, "Measuring and Managing Consumer Sentiment in an Online Community Environment", *Journal of Marketing Research*, 52(5), pp.629~641.

(65)Hong, Y. and Lee, T. H., 2003, "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models", *Review of Economics and Statistics*, 85(4), pp.1048~1062.

(66)Jurafsky, D. and Martin, J. H., 2009, *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing* (2nd edition), Upper Saddle River: Prentice Hall.

(67)Kuhn, T., 1996, *The Structure of Scientific Revolutions* (3rd edition), Chicago: University of Chicago Press.

(68)Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Alstyne, M. V., 2009, "Computational Social Science", *Science*, 323, pp.721~723.

(69)Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E., 2016, "Exact Post-Selection Inference, with Application to the LASSO", *Annals of Statistics*, 44(3), pp.907~927.

(70)Lew, A., Agrawal, M., Sontag, D. and Mansinghka, V., 2021, "PClean: Bayesian Data Cleaning at Scale with Domain-Specific Probabilistic Programming", in Banerjee, A. and Fukumizu, K., eds: *Proceedings of 24th International Conference on Artificial Intelligence and Statistics*, AAAI Press, Palo Alto.

(71)Li, K., Liu, X., Mai, F. and Zhang, T., 2021, "The Role of Corporate Culture in Bad Times: Evidence from the COVID-19 Pandemic", *Journal of Financial and Quantitative Analysis*, Accepted Manuscript, pp.1~68.

(72)Loughran, T. and McDonald, B., 2016, "Textual Analysis in Accounting and Finance: A Survey", *Journal of Accounting Research*, 54, pp.1187~1230.

(73)Lucas, R. E. Jr., 1976, "Econometric Policy Evaluation: A Critique", in Brunner, K. and Meltzer, A. H., eds: *The Phillips Curve and Labor Markets, Carnegie-Rochester Conference Series on Public Policy*, North Holland, Amsterdam.

(74)Manning, C. D., Raghavan, P. and Schütze, H., 2008, *Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

(75)Mosteller, F. and Wallace, D. L., 1963, "Inference in an Authorship Problem: A Comparative Study of Discrimination"(下转第 72 页)

pp.120~145.

(54)Holmstrom, B., 1989, "Agency Costs and Innovation", *Journal of Economic Behavior & Organization*, Vol.12, pp.305~327.

(55)La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R., 1998, "Law and Finance", *Journal of Political Economy*, Vol.106, pp.1113~1155.

(56)La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R., 1997, "Legal Determinants of External Finance", *Journal of Finance*, Vol.52, pp.1131~1150.

(57)Leff, N., 1964, "Economic Development through Bureaucratic Corruption", *American Behavioral Scientist*, Vol.8, pp.8~14.

(58)Long, C., Yang, J. and Zhang, J., 2015, "Institutional Impact of Foreign Direct Investment in China", *World Development*, Vol.66, pp.31~48.

(59)Lui, F., 1985, "An Equilibrium Queuing Model of Bribery", *Journal of Political Economy*, Vol.93, pp.760~781.

(60)Mauro, P., 1995, "Corruption and Growth", *The Quarterly Journal of Economics*, Vol.110, pp.681~712.

(61)Park, S. and Luo, Y., 2001, "Guanxi and Organizational Dynamics: Organizational Networking in Chinese Firms", *Strategic Management Journal*, Vol.22, pp.455~477.

(62)Paul, C. and Wilhite, A., 1991, "Rent-seeking, Rent-defending and Rent Dissipation", *Public Choice*, Vol.71, pp.61~70.

(63)Peng, M. and Luo, Y., 2000, "Managerial Ties and Firm Performance in a Transition Economy: The Nature of a Micro-macro Link", *Academy of Management Journal*, Vol.43, pp.486~501.

(64)Putnam, R., 1993, "The Prosperous Community: Social Capital and Public Life", *The American Prospect*, Vol.13, pp.35~42.

(65)Treich, N., 2010, "Risk-aversion and Prudence in Rent-seeking Games", *Public Choice*, Vol.145, pp.339~349.

(66)Tullock, G., 1980, "Rent Seeking as a Negative-Sum Game", in Buchanan et al.(eds): *Toward a Theory of the Rent Seeking Society*, College Station: Texas A&M University Press.

(上接第55页)Methods Applied to the Authorship of the Disputed Federalist Papers", *Journal of American Statistical Association*, 58(302), pp.275~309.

(76)Mullainathan, S. and Spiess, J., 2017, "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31, pp.87~106.

(77)Munezero, M., Montero, C. S., Mozgovoy, M. and Sutinen, E., 2013, "Exploiting Sentiment Analysis to Track Emotions in Students' Learning Diaries", *Koli Calling International Conference on Computing Education Research*, 13, pp.145~152.

(78)O'Hara, M., 1995, *Market Microstructure Theory*, Cambridge: Blackwell Publishing.

(79)Scott, S. and Varian, H., 2014, "Predicting the Present with Bayesian Structural Time Series", *International Journal of Mathematical Modelling and Numerical Optimisation*, 5, pp.4~23.

(80)Scott, S. and Varian, H., 2015, "Bayesian Variable Selection for Nowcasting Economic Time Series", in Goldfarb, A., Greenstein, S. M. and Tucker, C. E., eds: *Economic Analysis of Digital Economy*, University of Chicago Press, Chicago.

(81)Shiller, R., 2000, *Irrational Exuberance*, Princeton: Princeton University Press.

(82)Shiller, R., 2019, *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*, Princeton: Princeton University Press.

(83)Stock, J. H. and Trebbi, F., 2003, "Retrospectives: Who Invented Instrumental Variable Regression?", *Journal of Economic Perspectives*, 17(3), pp.177~194.

(84)Sun, Y., Han, A., Hong, Y. and Wang, S., 2018, "Threshold Autoregressive Models for Interval-Valued Time Series Data", *Journal of Econometrics*, 206(2), pp.414~446.

(85)Sun, Y., Hong, Y., Lee, T. H., Wang, S. and Zhang, X., 2021, "Time-Varying Model Averaging", *Journal of Econometrics*, 222(2), pp.974~992.

(86)Tetlock, P. C., 2007, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *Journal of Finance*, 62(3), pp.1139~1168.

(87)Tibshirani, R., 1996, "Regression Shrinkage and Selection via the LASSO", *Journal of the Royal Statistical Society: Series B(Methodological)*, 58(1), pp.267~288.

(88)Varian, H. R., 1999, *Intermediate Microeconomics: A Modern Approach*(5th edition), New York: WW Norton & Company.

(89)Varian, H. R., 2014, "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28(2), pp.3~28.

(90)Wright, P. G., 1928, *The Tariff on Animal and Vegetable Oils*, New York: Macmillan Company.

(91)Xie, H., Fan, K. and Wang, S., 2021, *Candlestick Forecasting for Investments: Applications, Models and Properties*, London and New York: Routledge.

(92)Zhu, M., Hong, Y. and Wang, S., 2021, "Can Interval Data Help Improve Volatility Forecasts? Evidence from Foreign Exchange Markets", Working Paper, Center for Forecasting Science, Chinese Academy of Sciences.

How Is Big Data Changing Economic Research Paradigms?

Hong Yongmiao and Wang Shouyang

(Academy of Mathematics and System Science, Chinese Academy of Sciences; School of Economics and Management, University of Chinese Academy of Sciences; Center for Forecasting Science, Chinese Academy of Sciences)

Summary: With the emerging new generation of information technology revolution and the Fourth Industrial Revolution, massive interrelated and high-frequency Big data of microeconomic behaviors and activities are being generated and recorded, examples being data from mobile smartphones, satellites and sensors, scanning machines, digital business platforms, digital social medias, government and private institution websites, and digital libraries. The Big data revolution is profoundly changing the ways of human production and life as well as the research paradigms and methodologies in economics. In this paper we first investigate the paradigm shifts brought by Big data to the mainstream economic empirical research, particularly the relaxation of some fundamental assumptions, including from a rational economic agent to a non-completely rational economic agent, from an isolated economic agent to socially connected economic agents, from a representative economic agent to heterogeneous economic agents, and from economic analysis to the systematic analysis of econ-social behaviors and activities. These changes and shifts make economic modeling and analysis much closer to economic reality and are expected to offer new knowledge discoveries about and insights into the economy.

Next, we discuss the methodological changes brought by Big data to economic empirical research, including from a model-driven approach to a data-driven approach, from focus on the impact of parameter estimation uncertainty to focus on the impact of model uncertainty, from the use of unbiased estimators to the use of regularized (biased) estimators, from a low-dimensional modeling strategy to a high-dimensional modeling strategy, from the use of low-frequency data to the use of high-frequency and even real-time data, from the use of structured data to the use of unstructured data (such as texts, graphs, photos, audio and video), from the use of traditional structured data to the use of new kinds of structured data (such as matrix data, functional data, interval-valued data and symbolic data), and from human data analysis to intelligence data analysis. Among many other things, Big data offers new information not available in traditional data, making it possible to conduct cutting-edge interdisciplinary research and to transform qualitative analysis to quantitative analysis. Leading examples include the textual regression analysis on the impact of investor sentiments on asset pricing due to the availability of text data, and analysis of heterogeneous distributional effects of macroeconomic policies and external shocks on different sectors or groups of economic agents due to availability of microeconomic high-frequency time series data.

The revolution in economic research paradigms and methodologies, brought by Big data, is reshaping the directions of modern economics, especially strengthening the trend of economic empirical research and breaking the restrictions of some basic assumptions of modern Western economics to various extents. As a result, economic research becomes increasingly scientific, rigorous, elaborate, interdisciplinarily diverse, and systematically integrated, and is increasingly similar to other fields of social science in terms of the use of quantitative methodologies. Given the scale advantages of mega economy, the rapid development of digital economy, and the rich experience of government reforms and policy-makings, China´s Big data resources provide a unique "rich ore" for Chinese economists to build an innovative theoretical system of Chinese economics by summarizing China´s economic development rules and developing a replicable China´s economic development model from the past 70 years of Chinese economic practice, particularly from the past 40 years of economic reforms and open-door policies to the outside world.

Keywords: Big data; textual analysis; machine learning; research paradigms; methodologies; reflexivity

JEL Classification: B4