社会计算驱动的社会科学研究方法*

周 涛 高 馨 罗家德

提要:社会计算(social computing)的方法论以社会科学理论为导引,并结合大数据与人工智能算法解决社会问题。本文从大量文献中提炼出融合大数据与社会科学理论的五种研究类型:基于大数据的探索性研究、基于大数据的验证性研究、大数据与结构化数据整合下的探索性或验证性研究、基于大型互联网实验的验证性研究和基于大数据(或结合结构化数据)先探索后验证的整合研究。本文针对上述五种研究进行了典型研究示例和若干相关研究成果展示。

关键词:大数据 人工智能 社会计算 研究方法论

一、引 言

社会科学家一般多关注因果和解释性问题,计算机学家旨在提升预测模型准确率。而随着社会计算、计算社会学、计算社会科学等交叉学科的兴起,关于"预测性和可解释性不再是权衡和竞争,而是可以相互补充"的呼吁逐渐得到重视。2021 年,霍夫曼(Jake Hofman)与瓦茨(Duncan Watts)等人在《自然》杂志上发文,依据可解释性和预测性将计算社会科学的研究方法划分到四个象限中:(1)描述性统计;(2)因果研究;(3)预测模型与预测因子分析;(4)因果与预测整合的研究(Hofman et al.,2021)。其中,第一象限与第三象限是数据驱动的探索性研究,第二象限是理论驱动的验证性研究,第四象限则是探索性和验证性结合的研究。本文以电子印迹大数据、整合的大数据与结构化数据、大型互联网实

* 感谢国家自然科学基金"基于海量企业和人才等数据的社会经济态势感知与预测"(11975071)、清华大学社会网络研究中心以及清华大学校内自主科研项目"基于通讯数据的关系强度与社会资本挖掘"(20182001706)、腾讯研究院"基于 QQ 大数据意见领袖识别研究"(20182001706)的资金支持。本文通讯作者为高馨。

验数据这三种数据来源为基础,结合单独或者整合的探索性和验证性研究方法,为这一领域的方法论做出系统的梳理和详细的案例展示。

大数据及其分析方法的出现推动了社会科学研究新范式的涌现。传统社会科学定量研究中,问卷数据存在样本规模小(兰德尔、马科夫斯基,2006)、失真(Fisher,1993)和系统误差等问题。而获得更准确且更大规模的数据,如经济社会普查数据的成本极高,同时,这类数据通常时效性也较差(高见、周涛,2016;Einav & Levin,2014)。

大量非结构化电子印迹数据(如网页搜索、社交网络互动内容、卫星遥感、视频图像、移动通信、社交媒体等)的记录和积累为社会科学研究者带来了前所未有的重大机会和挑战(舍恩伯格、库克耶,2013)。一方面,这些非结构化数据具有规模更大、实时性更强、精度更高的特点。因此,非结构化大数据的使用既可以降低小样本数据的稀疏性和偏差度,又可以增加社会现象动态发展过程的可见度,更好地描述社会经济发展态势。另一方面,理解和分析这类海量的非结构化数据,需要前沿的数据挖掘、机器学习和统计力学方法,这对以统计分析为主要工具的传统社会科学研究者提出了挑战。

近年来,大批计算机科学家和物理学家与社会科学家形成深度合作,旨在共同应对社会治理和预测问题,积极应对上述传统统计分析方法面临的挑战并提供了一些解决方案。这些具有交叉学科性质的学者一同提出了若干新的研究分支,包括计算社会科学(Lazer et al.,2009;Shah et al.,2015)、计算社会经济学(Gao et al.,2019;Zhou,2021)、社会计算(Wang et al.,2007;孟小峰等,2013;Evans et al.,2020;Evans,2020),等等。尽管这些研究分支的提法各有不同,但这些涌现出来的新研究分支具有明显的共性,即都是基于大数据,运用统计力学、动态建模和人工智能等方法和技术,来获得对社会现象和规律更精准的刻画,并提出科学解释。需要注意的是,与传统社会科学相比,二者之间的区别主要体现在研究方法的发展上,而非研究问题本身。新研究分支虽然只是在研究方法和数据上有所发展,却在解决重要社会问题上创生出巨大价值。概括来说,这些研究在与理论的对话过程中往往基于大量新数据来应用新方法,从而获得有社会科学理论观照的新发现,因此,笔者将这些研究统称为“社会计算(social computing)驱动的社会科学研究”。

国内关于社会计算研究的综述性文章,主要从数据、方法、工具或具体引入某一 方法形成的新研究范式等方面展开阐述。其中包括罗玮和罗教讲(2015)基于 2014 年美国社会学年会“新计算社会学”讨论会以及相关文献,将计算社

会学相关内容划分为“大数据的获取与分析、质性研究与定量研究的融合、互联网社会实验研究、计算机模拟研究和新型社会计算工具的研制与开发”这五类，主要从数据、方法和工具来做阐释，将这三个方面视为计算社会科学研究所需要的“原料”。陈云松等人（2020）则以瓦茨在2014年《美国社会学杂志》（*American Journal of Sociology*）上对社会科学家只重视可解释性而忽略预测性的批评为基础，重点介绍基于机器学习的方法为社会计算研究带来的新研究范式。胡安宁等人（2021）从研究方法层面介绍了机器学习模型在处理个体效应异质性中存在的优势。罗家德等人（2018，2021）主要从方法论层面阐述了理论、数据挖掘结果和预测模型间的动态三角对话的研究范式。

本文主旨是针对具体的研究问题，阐述如何使用和整合对应的社会科学理论、新数据和新方法来获得新发现、验证或修正理论，为相关研究者对上述不同要素进行组合、衔接和取舍以形成完整研究提供参考和定位。因此，我们从方法论角度作综合梳理，辅以具体案例展示，提出社会计算驱动的五类社会科学研究，分别是：（1）基于大数据的探索性研究；（2）基于大数据的验证性研究；（3）大数据与结构化数据整合下的探索性或验证性研究；（4）基于大型互联网实验的验证性研究；（5）基于大数据（或结合结构化数据）先探索后验证的整合研究。这五种方法论的提炼是笔者以萨尔加尼克（Matthew Salganik）基于大数据和调查数据提出的扩充型提问（扩展研究变量和议题）和丰富型提问（整合少数人的调查数据与大量研究对象的大数据）两种方法论（Salganik，2017）为基础，结合上述霍夫曼与瓦茨等人（Hofman et al.，2021）在《自然》杂志上提出四象限研究，进一步提炼得出的分类。在本文的划分依据下，第三类大数据与结构化数据整合下的探索性或验证性研究以及第五类基于大数据（或结合结构化数据）先探索后验证的整合研究尤其反映了萨尔加尼克（Salganik，2017）以及霍夫曼与瓦茨等人（Hofman et al.，2021）的方法论关切，展现了近年来的重要发展，标示着社会科学理论和社会计算方法在进一步深度整合。

这五大方法论划分所强调的核心内容包括以下三点。

第一，社会计算不是计算机科学（computer science）和社会数据（social data）的简单叠加，而是方法论层面的革新。社会计算扎根社会现实问题，解释和联系社会科学理论。

第二，五种方法论主要从该类研究所关注和解决的社会科学理论问题出发，以社会科学研究中对理论问题的探索性、验证性以及探索和验证的整合研究来划分。

第三,围绕所对话的理论或研究问题,社会计算驱动的社会科学研究的数据来源主要包括直接获得的电子印迹数据、电子印迹数据和结构化数据的结合,或是大规模网络实验数据。

综上所述,五种研究方法论以社会科学的问题意识为核心,以不同数据来源为基础,组织不同测量、分析方法和模型来解决问题。上述要素综合形成五种差异化的方法论演进路线。在提出上述方法论的"骨架"后,本文在每个方法论阐述下凝练和举出一个典型的案例,并简要介绍若干其他有代表性的案例,通过实例化的方式展现五种方法论下的具体的研究步骤和范式。

二、基于大数据的探索性研究

以往社会科学研究的数据往往来自问卷调查和控制实验,存在数据样本数量少、主观性高的问题。此外,当知晓自己是调查对象或实验对象,受访者会倾向于给出更易被社会接受的答案,而非真实的答案(Fisher,1993)。本文强调的电子化印迹数据是在研究对象不知情的情况下在现实生活中记录形成的,因此叫做自然数据。这类数据具有细粒度、大规模、强时序性的特点,因此,在开展大数据的探索性研究过程中发挥了重要作用。

大数据的探索性研究可以用于分析失业情况和职场发展。在无干预的情况下,笔者通过分析企业员工在内部办公系统中留下的记录,发现员工在办公系统中的活跃程度,特别是通过回溯员工间在办公系统发布任务、领取任务、上传、分享和下载文件等行为形成的互动关系,和该员工接下来一年之内的晋升或者离职有显著关系(张琳艳等,2015;Yuan et al.,2015)。自然数据还可以用来定量刻画两性不平等的程度。笔者通过分析互联网求职者的简历数据,发现平均而言女性要比男性多读一个学位或者多工作 5 年才能获得和男性一样的预期收入(Yang et al.,2018;王军等,2019)。

以下,笔者通过一个宗教隔离的研究案例(Hu et al.,2019)来详细展示如何采集和分析自然数据,并得到有价值的结论。宗教在人类文化中扮演着重要角色,宗教信仰有正面的价值,例如促进人类的合作(Purzycki et al.,2016)、提高生活的满意度(Lim & Putnam,2010)以及精神和身体健康水平(Koenig et al.,2001),等等。与此同时,因为不同宗教之间可能存在信仰内容和观念形态的差异,宗教之间会出现隔离现象,而这种现象对于文化演进、经济发展和政治制度

往往存在负面的影响(Atran & Jeremy,2012)。

笔者利用微博的公开数据分析宗教信徒之间形成的在线社交关系,观察这种社交关系中不同宗教之间是否存在隔离现象。为了从微博用户中把有宗教信仰的用户找出来,笔者先建立了一个宗教相关关键词的列表,该词表覆盖了最常见的一些和宗教相关的词语。笔者在微博用户的自我介绍、标签和昵称中搜索这些关键词,发现 170000 多用户包含了词表中至少一个关键词,有 9000 多用户包含了至少两个关键词。笔者所在的研究团队人工标注了这 9000 多用户,确认其中 6875 名是有特定宗教信仰的用户,分别属于佛教、基督教、道教、伊斯兰教中的一类。随后,根据这些用户之间的关注行为形成宗教关注网络。通过对这一特殊社交网络的探索性分析,笔者得到以下四个主要的发现。

第一,不同宗教之间的社交隔离非常严重。事实上,98.4% 的微博关注关系出现在两个信仰同一宗教的用户间,而只有 1.6% 的连边跨越不同宗教。笔者使用"同配系数"(assortative coefficient)(Newman,2003)来比较不同类别节点间连边的比例与随机化的网络相应连边的比例,以刻画不同类别节点之间的隔离程度。最终得到同配系数 $r=0.973$(r 取值的区间是$[-1,1]$,$r=1$ 时表示完全隔离)。该结果说明不同宗教之间的社交隔离非常严重。笔者还计算了很多其他测量隔离程度的指数,包括 E-I 指数(Krackhardt & Stern,1988)、Gupta-Anderson-May 指数(Gupta et al.,1989)、优势比(Moody,2001)等,结论均相同。

第二,跨越宗教的连边对于维持整体网络的连通性起到了决定性的作用。社交网络是一种典型的具有交换和传播信息功能的网络,对于这类网络而言,连通性是非常重要的性质。要判断连边对于维持网络连通性的作用大小,最通常的办法是比较去掉这些连边前后的网络连通性的差异(Li et al.,2021)。笔者对比了移除不同类型的连边前后网络的连通性,发现相比于其他算法筛选的连边,移除跨宗教连边后网络连通性下降得最多,说明跨宗教连边在维持网络连通性方面所起到的作用比通过其他算法筛选出来的边都要显著得多。

第三,在中国,信仰不同宗教的微博用户中,佛教徒最具开放性。在排除了不同教派人数差异的影响后,笔者发现,平均而言佛教徒关注其他宗教信徒或被其他宗教信徒关注的可能性都要更大。

第四,所有的跨宗教连边中约一半都和慈善有关。笔者发现,在所有被分析的 6875 个用户中,有 309 个用户至少吸引了一个其他宗教信仰者的关注。其中有 33 个属于主要发布慈善活动消息和新闻的用户。在这 33 个用户中,有 15 人曾因为慈善活动被媒体报道过,有 12 人在微博标签中有"慈善家"这一标签。

虽然这 33 个慈善用户只占了所有用户中的 0.48%，但却吸引了 46.7% 的跨宗教连边，可见慈善是增加宗教间沟通的可能切入点。

社会计算科学驱动的探索性研究从数据挖掘（data mining）出发，在得到一些指标值与行为规律的发现后，不能止步于此，还需要与现有理论对话并进行诠释，从而可以提出命题，以丰富、修正旧理论或发展新理论。接下来，笔者将继续展示如何根据上述探索研究中发现的指标和行为规律与宗教研究的相关理论进行对话，对探索结果做出诠释并提出理论命题。

根据案例研究问题和重要发现，本研究问题与齐美尔提出的“社会距离”展开了对话。这个概念主要表征“个体之间、群体之间或者个体与群体之间的相互作用和分离程度”（孔建勋、张晓倩，2017：76）。美国芝加哥学派社会学家帕克认为“社会距离是用以描述人际、社会关系的状态，表征相互理解和亲密的程度的概念”（Park，1924；孔建勋、张晓倩，2017：77）。博格达斯（Emory Bogardus）开发了社会距离测量量表，主要用于研究种族之间的隔离（Bogardus，1925）。已有的一些宗教理论揭示了宗教之间由于受到历史、宗教文化、利益和资源分配失衡以及各国“政教分离”政策的实施等影响，最终导致宗教去中心化，造成多元文化冲突（亨廷顿，2013；Nataraj，1965）。上述发现一、二证实了在中国情境下宗教分离现象的存在，并测量了分离程度。

回顾中国情境下的宗教研究，佛教非排他性的宗教观（薛克翘，2006）解释了发现三中佛教徒更有可能关注其他宗教信仰的人的结果。另外，社会规范理论强调社会整体对于宗教的态度会影响信徒的幸福感（Eichhorn，2011；Stavrova et al.，2013）。在中国传统宗教中，由于佛教是沟通中国大陆与东亚、南亚、台湾地区和香港地区的重要载体（Ji，2011；Laliberté，2011），因此受到更多的重视和认同，例如开展世界性的佛教论坛等（Lu & Gao，2017）。因此，在中国佛教徒的幸福感水平极大可能高于其他宗教信徒。

关于发现四，信徒参加宗教活动以及个人宗教身份的认同可以提升其幸福感（Ritter et al.，2014），当幸福感提升后，这些信徒可能更愿意关注社会慈善等事务，同时也有更大可能关注其他宗教信仰的人。因此，跨宗教连边大多与慈善相关。

综上，在上述理论诠释的基础上，关于发现三、四的解释还需要因果关系的进一步验证，由此我们提出以下命题，以期在后续研究中把幸福感视作中介变量来解释宗教身份认同与参与慈善活动和关注其他宗教群体的因果关系。

命题 1：个人宗教身份的认同和参加宗教活动可以提升信徒的幸福感，使信

徒愿意关注社会慈善。

命题2:个人宗教身份的认同和参加宗教活动可以提升信徒的幸福感,使信徒愿意和其他宗教群体个人产生联系。

上述基于微博大数据的探索性研究虽然简单,但却是第一次定量化地在中国的互联网环境中测量了宗教隔离的程度。同时,在方法上可以让读者观察到如何利用自然数据对宗教和相关社会问题进行探索性研究,展示了从收集数据、指标计算、数据挖掘、发现行为规律、对话理论做出诠释、提出后续有待验证的命题的过程。

探索性研究一般到发现和诠释为止,而这里提出命题旨在展示如何在探索和理论对话中启发后续更多因果关系的验证,从而形成理论上的推论,而非仅仅止步于社会事实的发现。综上所述,以电子印迹大数据为基础的探索性研究对于理论的意义主要包括如下两点。第一,可以使用大规模(甚至是全样本)、细粒度、无干预的数据为理论提供新的测量方法和工具,定量化、科学化地揭示社会规律和事实。第二,在探索普遍性规律及变量之间可能的关联关系的基础上,启发理论上基于因果机制的发现、提出命题,助益于后续更严格的建模和验证。

三、基于大数据的验证性研究

目前大多数社会计算驱动的社会科学研究所得到的实际上只是描述与关联关系,用这种关联关系直接对社会现象进行解释是不可靠的,因为充分的解释必须建立在因果关系的前提下。瓦茨曾分析了若干社会科学研究论文,指出大量的研究都把可解释性和因果性混为一谈,同时他对因果关系的验证也提出了更高的要求,即如果变量之间存在因果关系,那么同时应具备高预测性,从而更好地用于政策干预(Watts,2014)。得益于大数据和人工智能技术的发展,预测的效果被越来越多地用于结合因果计量模型,以共同验证社会科学的理论假设,从而弥补了单单使用预测模型无法真正证明或者证伪理论假设的缺陷。事实上,关联关系也可以用于预测,所以预测精度高对于因果关系的存在是必要而非充分的(Hempel & Oppenheim,1948)。另外,应用预测方法也有很多局限性(Jasny & Stone,2017;周涛,2017)。因此,笔者倡导在使用预测模型得到结果后,仍需要结合可解释人工智能方法,如SHAP(shapley additive explanations)(Lundberg & Lee,2017)等做出理论诠释,进一步通过假设演绎、使用因果模型/实验来做验

证——这样才是具有说服力的社会计算研究方法。

下面着重介绍一个基于高校学生校园行为的大数据研究学生行为和成绩之间关系的案例(Cao et al.,2017)。回顾相关理论和实证研究。第一,生活规律的学生往往有更好的自控能力,而自控能力和成绩表现是正相关的(Poropat,2009)。第二,更规律的生活,例如规律地吃饭、打热水、洗澡等活动,往往意味着更好的健康状况,而健康状况与学生表现有直接的关系(Santana et al.,2017;Hoffmann et al.,2018)。实证表明,规律的吃饭行为与学业表现具有很强的相关性(Valladares et al.,2016)。尤其是吃早餐对于学生的认知能力、心理健康和幸福感具有正向的影响。第三,社会发展理论表明学生习得行为就是来自其在社会化过程中个体行为以及与他人互动的一致性(Catalano et al.,2009),因此有规律的生活有助于个体的发展和增强其学校连结(school conneetedness)。而且,已有研究显示,较强的学校连结有利于学生的健康,可以提高其学业表现(Basch,2011;Sampasa-Kanyinga & Hamilton,2017)。

因此我们形成了一个理论假设:有规律的生活会正向影响学生成绩。

笔者及其研究小组分析了中国某大学校园一卡通中 18960 名本科生的匿名数据,覆盖了五个学期,包括 3380567 次洗澡、20060881 次吃饭、3466020 次进出图书馆和 2305311 次在教学楼打水的记录等——这些也是第二部分强调的"自然数据"。笔者通过计算学生洗澡和吃饭时间的真实熵[①]来定量刻画生活的规律性。之所以采用真实熵而非香农熵[②],是因为衡量学生吃饭是否规律不仅要看时间分布是否集中,还要看是否有序,比如吃了早餐吃中餐再吃晚餐,第二天同样吃早餐、中餐、晚餐,这是有规律的。而如果第一天吃了早餐不吃中餐,直接吃晚餐,第二天不吃早餐,吃中餐和晚餐,规律性相较于第一种情况有所降低。笔者用真实熵来度量集中度和周期性(Song et al.,2010;Xu et al.,2019),最终证明了假设,即生活规律的程度和学生学习成绩显著相关(Cao et al.,2017)。

为排除其他因素的影响,笔者同时控制了其他行为指数(例如努力程度)对上述相关关系的影响,同时控制了单纯的学生行为数据等变量,最终得到生活规律性仍然对学习成绩有显著影响并且可以显著提升预测准确率(Cao et al.,2019;Yao et al.,2019)。因此本案例展示从理论对话、提出假设、验证假设、稳健性检验的验证性研究过程,这一方法论已经广泛用于社会计算驱动的社会科学研究的方方面面,例如对劳动力市场的分析(张琳艳等,2015;Yuan et al.,2015)

① 用于度量信息的序列性和不确定性程度,这里主要用于度量学生生活的有序性和规律性。

② 用于度量信息的不确定性程度,这里主要用于度量学生生活的规律性。

和家庭财富情况的分析(Blumenstock et al.,2015)等。

此外,大数据结合网络动态模型可以为探索复杂系统理论的动态演化机制提供解决方案。风险投资领域普遍存在的联合投资现象可以给投资者带来更广阔的市场机会(Hochberg et al.,2010)和更高的市场声望(Poldolny,2001;Milanov & Shephere,2013),帮助其抵御不确定的市场环境和投资风险。中国风险联合投资中存在"主投—跟投"的现象,这些"主投"机构一般为产业领袖。产业领袖会建立自己的圈子,圈子中存在其他跟投机构,这些产业领袖同时充当着沟通不同圈子的"桥"的角色,导致小世界网络结构的形成。但不同圈子的其他跟随者之间则存在很少的联系,因此,这些产业领袖之间形成了一个互相联系紧密的"精英俱乐部"。笔者对中国 2000 年到 2013 年风险投资产业投资事件进行整理,将两个机构在同一时间投资同一家公司的行为视为一次联合投资,从而形成整个风险投资产业的联合投资网络(Gu et al.,2019)。笔者基于社会网理论中伙伴选择机制(partner-selection mechanism)和嵌入性理论(embedding theory)建立关于联合投资网络两种嵌入性的合作策略——关系性嵌入和结构性嵌入的假设(Granovetter,2017),使用基于多主体建模的方法,从网络的初始情况出发,预测网络发展演化的过程。在与真实风险投资网数据对比后发现,与随机选择模型相比,这两个机制下的模拟模型在全局和局部网络特征方面更接近真实的风险联合投资网络,并随时间变化,逐渐涌现出"精英俱乐部型小世界网络"的结构。该研究从大数据与基于多主体的模型为中国风险投资联合网络结构的涌现提供了理论上的解释,属于验证性研究,而过去调查问卷几乎不可能收集这类长时段、多时间戳、全产业的社会网络数据。由于大数据的积累,这类动态演化的问题才可以得到解释。

在社会计算驱动的验证性研究中,无论是预测模型还是模拟模型,强调的都是使用大数据、利用因果模型或者理论指导下的模拟模型来验证因果关系和理论机制,最终获得变量间因果关系上的可解释性。

四、大数据与结构化数据整合下的探索性或验证性研究

非结构化大数据不能取代传统的问卷调查或是档案数据库中的结构化数据。这两类数据的整合一方面可以测量更丰富的理论构念,增强理论发展、修正和探索,另一方面可以增强社会预测,助力社会治理,解决单一数据源不能解决

的问题。

大数据与结构化数据整合可以推断更多较难获得的调查数据。大数据技术的发展让我们有机会采集和处理与一个国家总人口规模相当(同一数量级)的数据,例如 Facebook、Twitter、微信、微博等社交媒体和智能手机覆盖总人口中占据相当比例的用户。因此,这类数据不再被看作是人口数据的一个小规模抽样,而是可以直接表达总体的统计性质。一方面,获得这些大数据的成本远远低于经济或人口普查,但另一方面,这些在社交媒体和手机通信中表现出来的行为本身往往不能直接回答我们亟须了解的有关家庭收入、就业情况、身心健康等重要社会问题。

将社交媒体和手机通信大数据与部分社会调查数据相结合,可以为上述难题提供可行的解决方案。例如,当我们需要分析大量个体的家庭收入时,就会面临以下两个方面的困难:一是很多较贫困的国家和地区不进行全民经济普查,二是这类数据往往不轻易向科研机构开放。在这种情况下,一方面,我们可以利用传统问卷调查的方式获得少量家庭收入的调查数据,由于这部分数据和研究问题非常相关且获得成本较高,我们不妨称其为"难获得的直接数据"。另一方面,大量社交媒体和手机通信大数据可称为"易获得的间接数据"。当二者结合,我们将"难获得的直接数据" 作为扎根真相(ground truth),利用机器学习的方法,建立通过"易获得的间接数据"来预测扎根真相的模型。先基于这些少量样本训练优化模型,待达到相当精确度后,原则上就可以推论出所有"易获得的间接数据"样本的"难获得的直接数据",如使用社交媒体或智能手机用户的家庭收入。尽管预测得到的数据不完全准确,但是其精确度对于分析宏观问题往往是足够的。

布卢门施托克(Joshua Blumenstock)等人利用上述方法尝试绘制了卢旺达全国范围的财富分布情况并识别最贫困的人口(Blumenstock et al., 2015; Blumenstock,2016;Steele et al.,2017)。通过和运营商合作,该研究团队可以分析匿名处理后的卢旺达 150 万手机用户数十亿次电话和短信的频率数据。他们在卢旺达手机用户中招募了 856 名志愿者,收集了这些人非常详细的社会经济状况的问卷调查数据,内容包括财产所有权、住宅特征、福利情况,等等。根据这些志愿者每人平均数千次电话和短信记录,建立了机器学习模型,可以利用志愿者的手机记录预测他们的财富指数,预测得到的财富指数和真实财富指数之间的皮尔逊相关系数高达 0.68。尽管预测数值和真实数值还存在不小的偏差,但这个预测模型能够通过手机数据估计 150 万卢旺达家庭的财富情况,并描绘覆

盖整个卢旺达的财富地图和贫困人口分布图。相比大规模的经济普查或问卷调查,这种“从间接数据推断直接数据”方法的性价比在绘制地区经济状况画像和时事政策干预中具有显著优势。

大数据与调查数据相结合的方法还可以进一步修正或探索理论。邓巴提出以功能划分五种不同互动模式的理论——亲族支持团体、共情群体、共宿群体、社群或族系和部落群体(Dunbar,1993;Dunbar & Spoors,1995;Hill & Dunbar,2003;Zhou et al.,2005;Pollet et al.,2011)。黄光国提出了中国语境下的三种不同的社会关系行为原则(Hwang,1987)——需求法则、人情交换法则和公平法则。基于此,笔者尝试探索关于中国人的人脉圈层到底可以划分为几层(罗家德等,2021;Gao et al.,2020)。笔者以问卷调查收集到的用户之间的关系强度作为扎根真相,再从这些用户在社交软件中互动的电子印迹化数据整理出指标,结合扎根真相,建立五层、四层、三层探索性预测分类模型,辅以解释模型,不断寻找准确率最高的划分方式,从而得到最合适的圈层结构。初步探索结果发现,家人、亲密熟人、一般熟人、认识之人四层模型解释力最强,预测模型最逼近扎根真相。

综上所述,融合大数据与社会调查的探索性或验证性研究主要强调的是使用大数据和预测模型来提出新的测量模型,建立基于理论构念或需要通过问卷调查和定性访谈获得的“扎根真相”,形成新的测量模型,从而通过易取得的大数据和预测模型去估计不易获得的扎根真相的过程,最大限度地展现大数据与调查数据结合后的价值,解决各类社会科学问题,极大地满足了社会治理、营销等多个场景中对于获得更具实时性、全局性、低成本性的扎根真相的应用需求。

五、基于大型互联网实验的验证性研究

除了电子印迹收集到的“自然数据”以及与社会调查和结构化数据整合得到的数据外,另外一种重要的大数据来源就是网络实验收集的数据。这类数据具有规模较大、成本较低、代表性较高的特点,为以实验为基础的验证性研究带来了新的机遇。

社会实验是在实验室的环境中抽象和模拟真实社会情景,并通过受试者在实验室中的反馈推断真实人群的社会心理和社会行为的研究方法,也是最近半个世纪以来社会科学研究中被越来越多使用的手段之一。与观察行为不同,研究人员开展实验,以期系统性地干预世界,获得因果性的验证(Salganik,2017)。

在实验设计中,研究者可以设计随机对照实验以排除混杂因素,这就要求研究者要先提出理论假设,然后把假设转化成研究设计。具体来说,包括设计实验方式、确定混杂因素、具体设计实验、招募参与者、随机分组、实施干预、测量结果、验证假设、得出结论等步骤(陈晓萍等,2012)。因此,实验是一种严格以理论为指导的研究类型,是一种验证性研究。

虽然传统的线下实验研究是因果研究中非常重要的研究范式,但也存在如下局限。第一,由于招募志愿者和实施实验的成本较高,因此受试者数量往往很少,由此造成统计显著度和可信度降低。第二,为了节省成本和实施方便,很多研究人员直接在其工作的校园内招募志愿者,这些被招募的学生并不能充分代表广泛人群,因此实验结果的普适性常被质疑。最近开放科学合作组织对 100 项心理学实验进行了重复,结果发现,成功复现的实验还不到 40%(Open Science Collaboration,2015)。尽管对于这个结论还存在争议(Gilbert et al.,2016),但是目前越来越多的证据显示小样本的实验室心理行为研究的可信度远低于预期。

运用互联网的手段设计和实施大规模社会科学实验有望克服原有实验设计成本过高、样本数量较少、样本代表性不足等问题。例如,在米尔格拉姆(Stanley Milgram)著名的线下小世界实验中,其中一次,他让受试将发出的 296 封信件通过熟人关系送达随机选择的千里之外的陌生人(Milgram,1967)。事实上只有 64 封信件送到了收件者手中,米尔格拉姆仅从这 64 封信件要经历多少次转手中得到了著名的"六度分离"理论,即两个陌生美国人之间只隔了五个中间的熟人就可以相互连接起来。与之相对,多德(Peter Dodds)等人利用互联网上电子邮件重做了米尔格拉姆的实验(Dodds et al.,2003)。来自 168 个国家和地区的 98865 人参加了这次实验,数据可信度和推论范围远超过米尔格拉姆的线下实验。实验的结果显示,在当时普遍使用的互联网通信网络中,人们连接更加紧密,美国大陆的"六度分离"演变成世界范围的"四度分离"。

另外一个具有代表性的利用大规模互联网实验研究社会科学问题的工作是邦德(Robert Bond)等人针对美国议会大选做的涉及 6100 多万人的政治动员实验(Bond et al.,2012)。他们假设个人的投票行为会受到朋友投票行为的影响。通过与 Facebook 合作,他们在 2010 年 11 月 2 日美国议会大选日当天,以所有 18 岁以上登录了 Facebook 网页的用户为实验对象并将其划分为三类:(1)社会组——实验对象的 Facebook 页面"新闻推荐"栏目的顶部会出现一个鼓励用户去投票的通告,并提供帮助用户找到附近的投票站的链接。这个通告下面有一

个写着“我已经投过票了”的按钮,用户通过点这个按钮来表达政治选择,该信息可以被 Facebook 好友获取。此外,用户还可以看到好友中已经点了那个按钮的数量,系统还会随机展示 6 个报告投过票的朋友的头像。(2)信息组——用户除了看不到任何投过票朋友的头像外,能够得到和社会组完全相同的信息。(3)控制组——用户在自己的 Facebook 主页没有收到任何相关信息。

这个研究最终发现,社会组有 20.04% 的人会点击“我已经投过票了”按钮,而信息组只有 17.96%,相差 2.08%。这证明了社会化的场景会大幅度提高人们政治表达的意愿。此外,通过对比真实的投票记录来分析这些用户是否真正会去投票,研究者发现,社会组和信息组的真实投票率相同,都比没有得到信息的用户高 0.39%,进一步证明利用人际关系网络的政治动员是有效果的。另外,社会组和信息组的真实投票率相同说明社会场景更多是让人们愿意表达和分享,而不是立刻和直接地改变人的行为。

网络大规模科学实验在很大程度上解决了传统线下实验样本量小、代表性不足等缺陷,也给社会科学理论中因果关系的验证带来了很多机会。下面,笔者将简要介绍几个典型案例,以便读者可以从中更加全面地了解这类研究的方法论优势。

2009 年 10 月,美国国防高级研究计划局(DARPA)组织了一个旨在探索“互联网和社交网络如何在解决一个紧急的跨区域问题上发挥重要作用”的竞赛,竞赛的目标是要参赛队在最短的时间内找到美国大陆上 10 个红色的气象气球。与其他团队设置的只奖励最后找到气球的人的奖励机制不同,麻省理工学院彭特兰(Alex Pentland)等人基于 Kleinberg-Raghavan 机制(Kleinberg & Raghavan, 2005)提出了一种层层递推的激励机制。参与者领取待解决的任务后,通过社交媒体或者其他方式找到自己朋友中可能会解决问题或者对解决问题有帮助的人参与进来,最终形成一个链条。如果某个人 A 最终解决了问题,A 是由 B 引入的,B 是由 C 引入的,C 是由 D 引入的,那么 A、B、C、D 共同分享奖金。彭特兰小组利用这个机制的吸引力在比赛前就招募到了 4400 人帮助寻找气球,最终也在正式比赛中以最短的时间找到了 10 个气球,夺取了比赛的冠军(Pickard et al., 2011)。这一实验奖励机制也启发了一系列后续研究(Li et al.,2017;顾勤、周涛,2021)。

另外一个是关于马太效应理论的网络实验。萨尔加尼克等人(Salganik et al., 2006)利用互联网招募了 14341 名青年志愿者参加一个音乐鉴赏任务。他们被要求对 48 首新歌进行从优到差的排序。这些人被分为 9 个组,其中控制组

的志愿者看不到任何其他人的信息,其余志愿者被分为 8 个组,他们在评价歌曲的同时可以看到每一首歌被他们所在小组人员下载的次数。萨尔加尼克等人发现,控制组不同歌曲下载的次数比较接近,但是另外 8 个组中歌曲下载次数的差异远远大于控制组,出现了“赢者通吃”的现象。这暗示了马太效应的存在:初始下载次数占优的歌曲会影响其他志愿者下载,从而使得初始的差距被进一步放大。在一个后续实验中,研究团队颠倒了受欢迎和不受欢迎的歌曲下载的初始排名,最终得到了完全不同的结果(Salganik & Watts,2008)。范德赖特(van de Rijt)研究小组做了另外一个揭示初始条件设置导致马太效应发生的网上实验,并完整地观察到这一过程(van de Rijt,2014)。他们在众筹网站 Kickstarter 上随机选择了 200 个新的众筹项目,这些项目被选择时的筹款总额都是 0 美元。然后,研究组随机选择 100 个项目(控制组)不做任何操作,另外 100 个项目给一笔随机选定额度的初始捐助。尽管在选择要给哪些项目进行初始捐助的时候研究组是盲目的,但是这些获得了少量初始捐助的项目最终成功募集到所需资金的概率是另外 100 个控制组项目的两倍,初始的 24.52 美元平均能吸引 191 美元的额外捐助。

综上,实施大规模互联网实验可以解决以往线下实验样本量小、代表性不足等问题,从而通过较低的成本形成或验证以往无法验证的理论。

六、基于大数据(或结合结构化数据)先探索后验证的整合研究

基于大数据或大数据与结构化数据的整合数据还可以开展先探索后验证的整合性研究。一方面,探索性研究可以获得定量化的社会现象之间可能的联系的规律,在与理论对话中获得对现象的解释或提出待验证的理论命题。另一方面,验证性研究可以在已有命题的基础上做出理论验证的研究设计,提出假设,用计量工具/实验来验证假设。这种整合性的研究兼具理论上的推论性和应用上的可预测性。

笔者以组织管理中某大型互联网公司中“影响团队创新能力的因素”为例来简述这一过程(Luo & Gao,2021)。我们收集了该公司两万余名员工形成的三千多个团队从 2014 年到 2018 年的五年期资料。本案例采用的研究思路如下:第一,在探索性研究部分,回顾已有文献中影响团队创新的因素,整合大数据与

结构化数据并计算表征上述影响因素的指标,形成指标体系,针对“团队是否具有高创新能力”这一问题建立预测模型,依据预测模型作可解释机器学习模型(interpretable AI)分析,对特征重要性、特征之间的关系进行诠释。第二,在探索性研究启发验证性研究部分,针对探索分析得到的、但已有理论未提供解释的新的发现,使用反溯推理(abductive reasoning)(Peirce,1893)提出假设,验证假设并得出结论,从而对预测模型的黑箱作出进一步解释,由此完成先探索后验证的整合研究。

本案例数据来源主要有两部分,一部分是全公司员工参与项目和培训课程的带有时间戳的印记化记录,包括其参加项目和课程名称的文本,还包括团队内面试官对新招聘员工拥有的知识和技能的评价文本等非结构化数据。另一部分为这些团队的创新奖励记录、员工个人信息等结构化数据。已有影响团队创新的因素主要包括:团队在合作网络中的位置(团队对外获取异质知识、资源的能力)、知识与技能、团队内网络密度(团队内成员的互动特征)、团队规模、团队成员组成(相似性和异质性)等。因此笔者对团队创新能力、团队合作网络位置以及员工知识进行定义和测量,并从数据中得到相关的指标。

为了测量团队在合作网络中的位置,笔者根据员工参加项目记录,将员工作为节点,如果两个员工在同一时间段共同参与同一个项目,则二者之间就形成了一条项目合作的连边。由于员工本身嵌入在正式的团队中,最终可以形成嵌入在同一或不同“团队”中的员工与员工之间项目合作网络,进而计算相关网络指标来表征团队在合作网络中的位置以及团队内的互动特征。

员工培训课程中的知识多样性主要通过员工参加培训课程的记录来计算。因为培训课程知识为文本数据,需要通过人工标注的方式对其中涉及的知识作分类,形成每一个团队中“员工—获取知识分类”的矩阵,并使用信息熵计算团队中员工通过参加培训课程获得知识的多样性。

招聘员工知识的多样性则基于面试官对新招聘员工知识和技能的评价文本,通过自然语言处理技术,提取出新聘员工技能的实词,再利用词向量和词嵌入的方法表示出文本特征。最终所有员工知识和技能点在嵌入空间中的平均距离即可表示出员工知识的多样性。

通过对结构化数据库的分析,可以得到团队的创新能力、团队成员组成的性别、职级、工作类型多样性,以及平均年龄、任期、团队规模等。

基于上述建立的指标体系,随机筛选 50% 的样本建立针对“团队是否具备高创新能力”的预测模型,使用十折交叉验证的方法以保证结论的稳健性,剩下

50%的样本用于后续验证性分析。以50%作为训练集,50%作为测试集建立预测模型,使用多种预测模型的实验中得到XGBoost预测模型准确率最高,平均准确率为76%。接下来作预测模型的可解释AI分析,使用SHAP模型,对指标重要性进行排序并筛选重要指标,得到除团队成员构成的其他指标外,“参加培训课程知识多样性”“合作网络E-I指数”①“新员工知识与上一年入职新员工知识差异”“团队内合作网络密度”(不分先后)这些指标对于团队创新能力的提高影响最大。在此基础上,笔者对这些变量之间的交互关系与团队高创新能力关系做了进一步的探索,发现培训知识多样性和新员工与上一年入职员工知识差异的交互作用有利于提高对团队创新能力的预测准确率,而在已有理论中却缺少知识和网络之间的交互关系对团队创新能力影响的阐述。

综上,在建立预测模型和可解释模型的探索中,得到三个结论:(1)合作网络对外开放程度对创新能力的积极效果(Burt,2004;Carnabuci & Diószegi,2015)。(2)团队员工参加培训知识多样性是影响创新的重要指标(Tannenbaum & Yuki,1992;Brown & Charliez,2013)。(3)招聘知识多样性高,或与上一年入职员工知识异质性高的员工进入团队,可以为团队带来更异质的想法,增强团队创新能力。

针对上述几个因素之间的交互关系对于团队创新能力的影响在已有理论中缺少解释的现状,启发我们在理论上来进行推理(reasoning),发展理论并开展验证性研究(Shrestha et al.,2021)。基于此,在验证性分析部分,笔者提出并验证了团队在合作网络中的位置和团队员工培训课程学习知识的多样性分别受到新注入的知识的多样性(即招聘知识与上一年入职员工知识的差异)的调节,对提高团队创新产生积极的影响。

首先提出假设。由于在网络中多样化的知识对于创新的积极效果受到传播过程异质信息快速同质化的影响,学习多样化知识很难长期维持团队的高创新能力。而“注入”知识多样性的员工会给网络带来一个积极的“震动”,改变团队在网络中组织知识的方式,使得团队可以更好地利用网络中异质的结构和内容的创新优势,带来更多的想法,改变已有成员的认知结构(Perretti et al.,2006),进而调整成员固有的看待已有知识的方式,塑造他们新的认知,为创新带来新的活力。因此,我们提出了假设1。

假设1:新加入员工与上一年入职员工知识的差异正向调节团队培训知识

① 合作网络E-I指数=(与其他团队成员连接数量-团队内连接数量)/(与其他团队成员连接数量+团队内连接数量)。

多样性对提高团队创新能力的正向影响。

一个团队中的员工有更多的团队外的合作者有利于团队成员与更多的不同工作模式和知识特征的团队进行交流、资源共享,因此有助于团队创新。但是,由于长期稳定合作的团队之间已深谙彼此交互的术语,这种例行的合作模式会导致团队之间在合作中变得僵化(Morrison,2002;Perretti & Negro,2006)。因此,团队中有新的成员加入,会对团队之间的合作者重新思考合作方式和互动模式产生影响。因此,笔者认为,与已有成员相比,具有异质知识的新成员的加入不仅会让已有合作网络中的成员重新思考他们与新成员的工作配合方式,旧的合作关系下的成员之间也会重新调整和思考他们合作的策略与模式。带有异质知识的新成员加入合作网络会增强团队在跨团队合作网络中的优势地位。因此提出假设2。

假设2:新的带有异质知识的员工的加入正向调节团队在合作网络中的开放度(E-I指数来衡量,越大代表开放度越大)对团队创新的积极影响。

使用上述随机划分的剩余50%的数据集,笔者使用面板数据进行回归分析,得到在控制历史创新能力和其他影响团队创新因素的基础上,合作网络E-I指数和新入职员工与上一年入职员工知识差异的交互作用对维持团队创新能力存在积极的效果($\beta=0.025$,$P<0.001$),员工培训课程知识多样性和新入职员工与上一年入职员工知识差异的交互作用同样对维持团队创新能力存在积极的效果($\beta=0.019$,$0.001<P<0.01$)。最终假设1、假设2均被证明。

综上,第一阶段探索性分析中二者之间的特征交互作用对团队创新能力影响的探索结果在理论上获得了解释,后续研究可以根据探索性分析得到的其他洞见来形成更多值得验证的理论假设。

通过对大数据的充分利用,本案例使用了综合的、扩展的测量指标,建立了具有较好预测准确率的预测模型,对影响团队创新解释机制做出了理论上的贡献。同时,通过这个研究案例还可以看出,非结构化与结构化数据的融合可以形成较大规模的样本量,探索性和验证性研究中使用不同的数据集进行探索和验证,避免使用同一批数据既做拟合又做验证,能够在很大程度上检验以往研究中理论可重复性低的问题(Nosek et al.,2015)。

以上案例展示了从探索性研究到验证性研究的一条演进道路。此外,还有研究先基于理论假设作验证,再使用预测模型探索重要特征(Christoph et al.,2021),也有研究同时建立预测模型和可解释模型,二者之间不断对话,启发理论创新。例如在2018年发表在《自然》杂志上的一篇文章(Awad et al.,2018)收集

了223个国家4000万参与者对于自动驾驶汽车决策选择的数据,通过探索性和验证性整合研究来启发新的心理学理论,发现了以往忽略的自动驾驶汽车的一些伦理规范问题,例如人类决策过程的内在冲突、人际冲突、伦理道德的文化差异等。在此基础上,后续研究(Agrawal et al.,2020)使用该数据对预测模型和决策心理模型进行了对话和相互的修正,针对预测模型和心理模型预测结果差异较大的样本进行分析,获得在具体决策情境下,一些在心理模型中未被注意到的几种因素的交互关系对决策的影响,启发作者提出理论假设并做出实验验证,使得心理可解释模型更加完善。最终得到了仅由22个参数组成的心理理论模型,相比于未考虑交互项但具有超过3000个参数的深度学习模型,该模型兼备高预测精度和可解释性。正如鲁丁(Charles Rudin)所强调的,在高风险决策中仅仅依靠预测模型及其特征重要性分析是非常危险的,应该辅以严格的验证性分析,结论才具有可靠性和推论性(Rudin,2019)。

综上,大数据(或与结构化数据整合)的探索性和验证性相结合的研究可在理论解释模型上提供新的洞见,具体包括:(1)获得新的理论的测量指标、测量方法,或者对原有缺少可解释性的指标进行拆解或重新划分。(2)获得不同指标之间的交互关系与因变量可能存在的因果关系。(3)获得在社会治理、知识决策、政策干预上新的洞见和启发。将这种探索性和验证性的研究综合起来,可以获得理论上的可解释性和较高的预测性。

七、总结与讨论

融合大数据与社会科学理论的研究方法开始见诸学术期刊不过二十年左右的时间,大量的文章则在最近几年才出现。相比社会科学漫长的历史来说,社会计算驱动的社会科学研究方法论方兴未艾,整体上来说还处于不断摸索前进的阶段,本文主要在方法论上做了一个阶段性小结。本文对社会科学理论的探索性研究、验证性研究以及数据取得的三种不同方法(收集电子印迹的大数据、整合大数据与结构化数据库或问卷调查的数据以及互联网实验数据)划分了五类研究方法,并对每一类研究方法给出了一个梗概的介绍。不排除将来还有更多的研究进路被发现和使用。毋庸置疑的是,社会计算驱动的社会科学研究新方法论,深刻地改变了整个社会科学的理论发展与研究范式。

需要强调的是,与早期的大数据研究仅强调归纳而忽略因果、演绎推理不

同,社会计算更多强调用理论指导计算机技术探索并解释社会规律和模式,与社会学理论形成密切的对话,最终用于启发、验证或修正理论。

当然,如前文所述的方法还存在很多的缺陷和挑战,需要有志于此的学者作更多的贡献。

首先,将社会计算研究用于政策干预和指导存在较大的挑战。邦德等人研究中涉及6100万人的实验并形成干预,这类研究不仅仅立足于解释和预测,更重要的目标是达成有利于社会发展、降低不平等(Bond et al.,2012;Blumenstock,2016)的干预,但此类研究还是凤毛麟角。

其次,几种研究方法可能带来一系列法规、道德和伦理的问题,需要研究人员谨慎对待。大数据和人工智能的研究本身就带来了一系列的科学伦理问题(Poldolny,2001)。具体来说,第一,要特别注意保护被分析对象的隐私。在使用"自然数据"时,一部分数据并非来自公开网站(例如智能手机数据),一部分数据虽然来自公开网站但不等于用户希望别人看到自己被分析的结果——例如一个人愿意在Facebook上向好友公开他的信息,但不等于他愿意公开通过其Facebook数据预测到他罹患抑郁症(de Choudhury et al.,2014)或者是男同性恋(Kosinski et al.,2013)的结论。虽然研究论文使用和报道的数据经过匿名化处理,但是最近一些研究发现可以通过这些匿名数据反推到个体(de Montjoye et al.,2013,2015)。所以在报告研究结果和共享研究数据时要非常谨慎,避免其他研究人员通过技术手段反向识别数据对象的身份。第二,在开展互联网实验时,有些时候为了实验效果,受试者并不知道自己处于实验环境下,研究人员必须充分评估实施实验对受试者的情绪和心理造成的影响。第三,对于分析结果和结论的公开也要谨慎。譬如数据分析和生物、物理实验可能会揭示不同种族的人群因为基因或者其他原因导致的智力、体力和心理的差异,这些差异的公开可能反而会将弱势群体置于更不利的位置。

结合本文所给出的几种研究方法类型,在社会科学理论指导下的社会计算研究可总结为验证性研究和探索性研究,或是两者的结合。固然在单一论文中这五类研究方法多是单独使用的,但在系列研究中,探索性与验证性研究却应在如图1所示的理论、数据挖掘和模型的三角对话中一轮又一轮地交互进行。伴随着理论、数据挖掘和模型的三角对话,这一过程中同时存在着演绎法和归纳法。

一方面,社会科学理论可以为大数据挖掘提供指导,选择更适合刻画研究对象的指标,实例化或修正算法。另外,理论还可以为机器学习模型或者动态模拟

模型的建立提供灵感或直接支持。得到模型后也需要继续与理论进行对话,判断是否与已有理论一致,如果一致则为理论的验证,如果不一致则对模型进行影响准确率因素的分析,以不断修正模型。

另一方面,随着数据驱动的预测模型的建立,数据挖掘结果和机器学习模型同样可以启发探索新的理论方向,验证或挑战已有理论。当理论与机器学习模型不一致时,通过对模型作可解释人工智能算法、定性调查、分错误样本溯因,综合判断是否需要对理论进行修正和重新阐述,可提出相应的命题(Evans et al., 2020)。同样,因数据驱动而建构的模型后续还可继续根据提出的命题来完成验证性研究。这些例子可以在本文第二部分和第四部分中找到。

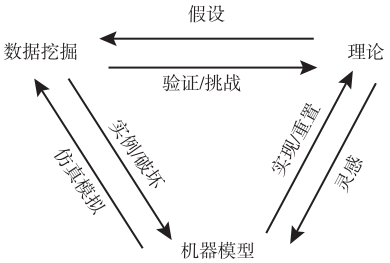

图 1　理论验证、数据挖掘、机器模型的三角对话(Evans et al.,2020)

本文所讨论的验证性研究可以粗略地分为三类。第一类是利用数据挖掘结果和机器学习或动态模拟模型,提供理论修正或重建洞见。这只是完成了弱模型(weak model)的建立,接下来可根据探索性研究提出的命题,继续结合理论提出严格的理论假设,再收集大数据、建立因果模型、验证假设,最终得出更具推论性的结论。第二类是利用大数据与结构化数据的整合进行理论验证。结构化数据库或问卷调查可以提供扎根真相,因果模型可以验证理论假设,大数据则可以测量和计算更多相关指标,从而极大地丰富传统社会科学理论研究议题。第三类则是根据理论设计大规模互联网实验。

如本文第四部分所描述的探索性研究案例,可以在一轮又一轮的三角对话中,与某一研究相关的各类型数据整合在一起,预测模型被建立且在修正中准确度不断提高,由此社会科学的理论可以得到更深入的发展。从探索性到验证性的完整研究在第六部分案例中有所展示。

很多学科都曾因为理论和方法的突破而产生重大转折。这个转折期往往群星璀璨,硕果累累。20 世纪初期量子力学理论对物理学的影响、20 世纪后期基

因测序技术对生物学的影响就是非常典型的例子。笔者认为,社会科学这一历史悠久的学科正在因为大数据和人工智能技术的引入而经历一次重大的转折。希望在可见的未来可以有更多新鲜的血液,在一个学科发生重大转折的开始时期就注入进来,从而将此一新方法深植在社会科学研究之中。

参考文献:

陈云松、吴晓刚、胡安宁、贺光烨、句国栋,2020,《社会预测:基于机器学习的研究新范式》,《社会学研究》第3期。

陈晓萍,徐淑英,樊景立,2012,《组织与管理研究的实证方法(第2版)》,北京:北京大学出版社。

高见、周涛,2016,《大数据揭示经济发展状况》,《电子科技大学学报》第4期。

顾勤、周涛,2021,《数据要素流通的分账机制研究》,《电子科技大学学报》第3期。

亨廷顿,塞缪尔,2013,《文明的冲突》,周琪等译,北京:新华出版社。

胡安宁、吴晓刚、陈云松,2021,《处理效应异质性分析——机器学习方法带来的机遇与挑战》,《社会学研究》第1期。

兰德尔,柯林斯、迈克尔·马科夫斯基,2006,《发现社会之旅——西方社会科学思想述评》,李霞译,北京:中华书局。

孔建勋、张晓倩,2017,《当前缅甸不同宗教群体之间的社会距离及其影响因素》,《世界宗教文化》第1期。

罗家德、刘济帆、杨鲲昊、傅晓明,2018,《论社会学理论导引的大数据研究——大数据、理论与预测模型的三角对话》,《社会学研究》第5期。

罗家德、高馨、周涛、刘黎春、傅晓明、刘知远、苏毓淞,2021,《大数据和结构化数据整合的方法论——以中国人脉圈研究为例》,《社会学研究》第2期。

罗玮、罗教讲,2015,《新计算社会学:大数据时代的社会学研究》,《社会学研究》第3期。

孟小峰、李勇、祝建华,2013,《社会计算:大数据时代的机遇与挑战》,《计算机研究与发展》第12期。

舍恩伯格,维克托·迈尔、肯尼思·库克耶,2013,《大数据时代:生活、工作与思维的大变革》,盛杨燕、周涛译,杭州:浙江人民出版社。

王军、高见、杨枭、刘金虎、周涛,2019,《在线数据揭示预期薪金的影响因素》,《电子科技大学学报》第2期。

薛克翘,2006,《佛教与中国文化》,北京:昆仑出版社。

张琳艳、高见、洪翔、周涛,2015,《大数据导航人力资源管理》,《大数据》第1期。

周涛,2017,《预测的局限性》,《大数据》第4期。

Agrawal, M., J. C. Peterson & T. L. Griffiths 2020, "Scaling Up Psychology via Scientific Regret Minimization." *Proceedings of the National Academy of Sciences* 117(16).

Atran, S. & G. Jeremy 2012, "Religious and Sacred Imperatives in Human Conflict." *Science* 336(6083).

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. Bonnefon & I. Rahwan 2018, "The Moral Machine Experiment." *Nature* 563.

Basch, A. E. 2011, "Breakfast and the Achievement Gap among Urban Minority Youth." *Journal of School Health* 81(10).

Blumenstock, J. 2016, "Fighting Poverty with Data." *Science* 353(6301).

Blumenstock, J., C. Gabriel & O. Robert 2015, "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264).

Bogardus, E. S. 1925, "Social Distance and its Origins." *Journal of Applied Sociology* 9(216).

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle & J. H. Fowler 2012, "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489(7415).

Brown, K. G. & S. D. Charlier 2013, "An Integrative Model of E-Learning Use: Leveraging Theory to Understand and Increase Usage." *Human Resource Management Review* 23(1).

Burt, R. S. 2004, "Structural Holes and Good Ideas." *American Journal of Sociology* 110 (2).

Cao, Y., J. Gao, D. F. Lian, Z. H. Rong, J. T. Shi, Q. Wang, Y. F. Wu, H. X. Yao & T. Zhou 2017, "Orderliness Predicts Academic Performance: Behavioral Analysis on Campus Lifestyle." *Journal of The Royal Society Interface* 15(146).

Cao, Y., J. Gao & T. Zhou 2019, "Orderliness of Campus Lifestyle Predicts Academic Performance: A Case Study in Chinese University." In C. Montag & H. Baumeister (eds), *Digital Phenotyping and Mobile Sensing: Studies in Neuroscience, Psychology and Behavioral Economics.* Basel: Springer, Cham.

Carnabuci, G. & B. Diószegi 2015, "Social Networks, Cognitive Style, and Innovative Performance: A Contingency Perspective." *Academy of Management Journal* 58(3).

Catalano, R. F., K. P. Haggerty, S. Oesterle, C. B. Fleming & J. D. Hawkins 2009, "The Importance of Bonding to School for Healthy Development: Findings from the Social Development Research Group." *Journal of School Health* 74(7).

Christoph, R., Y. J. Kim, P. Gupta, T. W. Malone & A. W. Woolley 2021, "Quantifying Collective Intelligence in Human Groups." *Proceedings of the National Academy of Sciences* 118(21).

de Choudhury, M., S. Counts, J. H. Eric & A. Hoff 2014, "Characterizing and Predicting Postpartum Depression from Shared Facebook Data." Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. Baltimore Maryland, February 15.

de Montjoye, Y. A., C. A. Hidalgo, M. Verleysen & V. D. Blondel 2013, "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 3(1376).

de Montjoye, Y. A., L. Radaelli & V. K. Singh 2015, "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata." *Science* 347(6221).

Dodds, P. S., R. Muhamad & D. J. Watts 2003, "An Experimental Study of Search in Global Social Networks." *Science* 301(5634).

Dunbar, R. I. 1993, "Coevolution of Neocortical Size, Group Size and Language in Humans." *Behavioral and Brain Sciences* 16 (4).

Dunbar, R. I. & M. Spoors 1995, "Social Networks, Support Cliques, and Kinship." *Human Nature* 6(3).

Eichhorn, J. 2011, "Happiness for Believers? Contextualizing The Effects of Religiosity on Life-Satisfaction." *European Sociological Review* 28(5).

Einav, L. & J. Levin 2014, "Economics in The Age of Big Data." *Science* 346 (6210).

Evans, J. 2020, "Social Computing Unhinged." *Journal of Social Computing* 1(1).

Evans, J., X. M. Fu & J. D. Luo 2020, "Inaugural Message from Editors-in-Chief." *Journal of Social Computing* 1(1).

Fisher, R. J. 1993, "Social Desirability Bias and The Validity of Indirect Questioning." *Journal of Consumer Research* 20 (2).

Gao, J., Y. C. Zhang & T. Zhou 2019, "Computational Socioeconomics." *Physics Reports* 817.

Gao, X., J. D. Luo, K. Yang, X. M. Fu, L. C. Liu & W. W. Gu 2020, "Predicting Tie Strength of Chinese Guanxi by Using Big Data of Social Networks." *Journal of Social Computing* 1(1).

Gupta, S., R. M. Anderson & R. M. May 1989, "Networks of Sexual Contacts: Implications for The Pattern of Spread of HIV." *AIDS* 3(12).

Gilbert, A. T., G. King, S. Pettigrew & T. D. Wilson 2016, "Comment on 'Estimating the Reproducibility of Psychological Science'." *Science* 351(1037).

Granovetter, M. 2017, *Society and Economy: Framework and Principles*. Cambridge: Harvard University Press.

Gu, W., J. D. Luo & J. Liu 2019, "Exploring Small-World Network with an Elite-Clique: Bringing Embeddedness Theory into the Dynamic Evolution of a Venture Capital Network." *Social Networks* 57.

Hempel, C. G. & P. Oppenheim 1948, "Studies in the Logic of Explanation." *Philosophy of Science* 15.

Hill, R. & R. I. Dunbar 2003, "Social Network Size in Humans." *Human Nature* 14(1).

Hochberg, Y. V., A. Ljungqvist & L. U. Yang 2010, "Networking as a Barrier to Entry and the Competitive Supply of Venture Capital." *Journal of Finance* 65(3).

Hoffmann, I., C. Diefenbach, C. Gräf, J. König, M. F. Schmidt, K. Schnick-Vollmer, M. Blettner, M. S. Urschitz & ikidS Study Group 2018, "Chronic Health Conditions and School Performance in First Graders: A Prospective Cohort Study." *PloS one* 13(3).

Hofman, J. M., D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani & T. Yarkoni 2021, "Integrating Explanation and Prediction in Computational Social Science." *Nature* 595.

Hu, J. T., Q. M. Zhang & T. Zhou 2019, "Segregation in Religion Networks." *EPJ Data Science* 8.

Hwang, K. 1987, "Face and Favor: The Chinese Power Game." *American Journal of Sociology* 92(4).

Jasny, B. R. & R. Stone 2017, "Prediction and its Limits." *Science* 355(6324).

Ji, Z. 2011, "Buddhism in the Reform Era: A Secularized Revival?" In A. Y. Chau (eds.), *Religion in Contemporary China: Revitalization and Innovation*, London: Routledge.

Kleinberg, J. & P. Raghavan 2005, "Query Incentive Networks." Proceedings of 46th Annual IEEE Symposium on Foundations of Computer Science. Pittsburgh, October 23 – 25.

Koenig, H., E. M. Michael & B. L. David 2001, *Handbook of Religion and Health*. New York: Oxford University Press.

Kosinski, M., D. Stillwell & T. Graepel 2013, "Private Traits and Attributes are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences* 110 (15).

Krackhardt, D. & R. N. Stern 1988, "Informal Networks and Organizational Crises: An Experimental Simulation." *Social Psychology Quarterly* 51.

Laliberté, A. 2011, "Buddhist Revival under State Watch." *Journal of Current Chinese Affairs* 40(2).

Lazer, D., A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy & M. V. Alstyne 2009, "Life in the Network: The Coming Age of Computational Social Science." *Science* 323(5915).

Li, B., D. Hao, D. J. Zhao & T. Zhou 2017, "Mechanism Design in Social Networks." Proceedings of 31st AAAI Conference on Artificial Intelligence. New York, February 13.

Li, M., R. R. Liu, L. Lü, M. B. Hu, S. Xu & Y. C. Zhang 2021, "Percolation on Complex Networks: Theory and Application." *Physics Reports* 907.

Lu, J. & Q. Gao 2017, "Faith and Happiness in China: Roles of Religious Identity, Beliefs, and Practice." *Social Indicators Research* 132.

Lundberg, S. & S. I. Lee 2017, "A Unified Approach to Interpreting Model Predictions." Proceedings of the International Conference on Neural Information Processing Systems (Nips). Long Beach, May 22.

Luo, J. D., X. Gao, 2021, "How can We Improve Team Innovation in Organization Changes?", Paper Presented at the 3rd International Conference of Social Computing, China, Beijing.

Milanov, H. & D. A. Shephere 2013, "The Importance of The First Relationship: The Ongoing Influence of Initial Network on Future Status." *Strategic Management Journal* 34(6).

Milgram, S. 1967, "The Small World Problem." *Psychology Today* 2(1).

Moody, J. 2001, "Race, School Integration, and Friendship Segregation in America." *American Journal of Sociology* 107(3).

Morrison, E. W. 2002, "Newcomer's Relationships: The Role of Social Network Ties During Socialization." *Academy of Management Journal* 45(6).

Nataraj, P. 1965, "Social Distance Within and Between Castes and Religious Groups of College Girls." *Journal of Social Psychology* 65(1).

Newman, M. 2003, "Mixing Patterns in Networks." *Physical Review E* 67(2).

Nosek, B. A., G. Alterg & C. Banks et al. 2015, "Promoting an Open Research Culture." *Science* 348.

Open Science Collaboration 2015, "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

Park, R. E. 1924, "The Concept of Social Distance as Applied to the Study of Racial Attitudes and Racial Relations." *Journal of Applied Sociology* 8(6).

Peirce, C. S. 1893, *Fallibilism, Continuity, and Evolution.* MS.

Perretti, F. & G. Negro 2006, "Filling Empty Seats: How Status and Organizational Hierarchies Affect Exploration Versus Exploitation in Team Design." *Academy of Management Journal* 49 (4).

Pickard, A., W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan & A. Pentland 2011, "Time-Critical Social Mobilization." *Science* 334(6055).

Poldolny, J. M. 2001, "Networks as the Pipes and Prisms of the Market." *American Journal of Sociology* 107 (1).

Pollet, T. V., S. G. Roberts & R. I. Dunbar 2011, "Use of Social Network Sites and Instant Messaging Does Not Lead to Increased offline Social Network Size, or to Emotionally Closer Relationships with Offline Network Members." *Cyberpsychology, Behavior, and Social Networking* 14(4).

Poropat, A. E. 2009, "A Meta-Analysis of The Five-Factor Model of Personality and Academic Performance." *Psychological Bulletin* 135(2).

Purzycki, B. G., C. Apicella, Q. D. Atkinson, E. Cohen, R. A. McNamara, A. K. Willard, D. Xygalatas, A. Norenzayan & J. Henrich 2016, "Moralistic Gods, Supernatural Punishment and The Expansion of Human Sociality." Nature 530(7590).

Ritter, R. S., J. L. Preston & I. Hernandez 2014, "Happy Tweets: Christians Are Happier, More Socially Connected, and Less Analytical Than Atheists on Twitter." *Social Psychological and Personality Science* 5(2).

Rudin, C. 2019, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1.

Salganik, M. J. 2017, *Bit by Bit: Social Research in the Digital Age.* Princeton: Princeton University Press.

Salganik, M. J. & D. J. Watts 2008, "Leading the Herd Astray: An Experimental Study of Self – Fulfilling Prophecies in An Artificial Cultural Market." Social Psychology Quarterly 71(4).

Salganik, M. J., P. Sheridan & D. J. Watts 2006, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311(5762).

Sampasa-Kanyinga, H. & H. A. Hamilton 2017, "Eating Breakfast Regularly Is Related to Higher School Connectedness and Academic Performance in Canadian Middle- And High-School Students." *Public Health* 145.

Santana, C., J. O. Hill, L. B. Azevedo, T. Gunnarsdottir & W. L. Prado 2017, "The Association Between Obesity and Academic Performance in Youth: A Systematic Review." *Obesity Reviews: An Official Journal of The International Association for the Study of Obesity* 18(10).

Shah, D. V., J. N. Cappella & W. R. Neuman 2015, "Big Data, Digital Media, and Computational Social Science: Possibilities and Perils." *The Annals of the American Academy of Political and Social Science* 659(1).

Shrestha, Y. R., V. F. He, P. Puranam & G. von Krogh 2021, "Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?" *Organization Science* 32(3).

Song, A. M., Z. H. Qu, N. Blumm & A. L. Barabasi 2010, "Limits of Predictability in Human Mobility." *Science* 327(5968).

Stavrova, O., D. Fetchenhauer & T. Schlösser 2013, "Why Are Religious People Happy? The Effect of the Social Norm of Religiosity Across Countries." *Social Science Research* 42(1).

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, Tomas J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem & L. Bengtsson 2017, "Mapping Poverty Using Mobile Phone and Satellite Data." *Journal of The Royal Society Interface* 14(127).

Tannenbaum, S. I. & G. Yuki 1992, "Training and Development in Work Organizations." *Annual Review of Psychology* 43(1).

Valladares, M., E. Durán, A. Matheus, S. Durán-Agüero, A. M. Obregón & R. Ramírez-Tagle 2016, "Association Between Eating Behavior and Academic Performance in University Students." *Journal of the*

American College of Nutrition 35(8).

van de Rijt, A., S. M. Kang, M. Restivo & A. Patil 2014, "Field Experiments of Success-Breeds-Success Dynamics." *Proceedings of the National Academy of Sciences* 111(19).

Wang, F. Y., K. M. Carley, D. Zeng & W. Mao 2007, "Social Computing: From Social Informatics to Social Intelligence." *IEEE Intelligent Systems* 22(2).

Watts, D. J. 2014, "Common Sense and Sociological Explanations." *American Journal of Sociology* 120(2).

Xu, P. H., L. K. Yin, Z. T. Yue & T. Zhou 2019, "On Predictability of Time Series." *Physica A: Statistical Mechanics and its Applications* 523.

Yang, X., J. Gao, J. H. Liu & T. Zhou 2018, "Height Conditions Salary Expectations: Evidence from Large-Scale Data in China." *Physica A: Statistical Mechanics and its Applications* 501.

Yao, H. X., D. F. Lian, Y. Cao, Y. F. Wu & T. Zhou 2019, "Predicting Academic Performance for College Students: A Campus Behavior Perspective." *ACM Transactions on Intelligent Systems and Technology* 10(3).

Yuan, J., Q. M. Zhang, J. Gao, L. Y. Zhang, X. S. Wan, X. J. Yu & T. Zhou 2015, "Promotion and Resignation in Employee Networks." *Physica A: Statistical Mechanics and its Applications* 444.

Zhou, T. 2021, "Representative Methods of Computational Socioeconomics." *Journal of Physics: Complexity* 2.

Zhou, W., D. Sornette, R. A. Hill & R. I. Dunbar 2005, "Discrete Hierarchical Organization of Social Group Sizes." *Proceedings of the Royal Society B: Biological Sciences* 272 (1561).

作者单位:电子科技大学计算机科学与工程学院、大数据研究中心(周涛)
清华大学社会科学学院(高馨)
清华大学社会科学学院、公共管理学院(罗家德)

责任编辑:徐宗阳