第一章

缘起：范式的又一次转移，但太阳底下无新鲜事

本书的目的是深入探讨社会计算的发展历程及其所蕴含的深刻启示——社会科学在方法论与理论上的一次范式转移。

本章的副标题之所以是“太阳底下无新鲜事”，是因为“社会科学”学界的范式转移早已发生过不止一次，而这一次和上一次的转移过程有着惊人的相似之处，所以我们可以借鉴过去的历史，启发对未来发展的思考。

一　新数据、新方法扑面而来

近 20 年来，人工智能与大数据技术的飞跃发展，不仅极大地推动了科技进步，更与社会科学领域产生了前所未有的紧密联系，催生了一系列新兴现象。例如，当前国家正积极推动哲学社会科学实验室的建设，这无疑是新学术体制下的一次重要探索，其未来发展路径令人期待。

然而，值得注意的是，许多看似新颖的事物实则有着深厚的历史渊源。因此，回顾并梳理社会计算及相关领域的发展历程，对于理解并预测其未来发展具有不可估量的价值。本书的宗旨即在于追溯这段有迹可循的历史，探讨其对当前及未

来的社会科学在社会计算领域可能产生的重大影响，以及社会计算的发展如何促使这些“软科学”进行了范式转移。

本书以社会计算的具体案例为核心，但最后提及“复杂系统转型”这一关键议题，本系列的后续出版物将会深入探讨复杂系统理论与社会计算之间相辅相成的关系。范式转移不只是社会计算方法论带来的，同时也是社会科学理论的全面更新，它将化约思想变为复杂思维，从静态变成演化过程，从线性研究变成非线性的涌现研究，从封闭系统变成开放系统。我坚信，未来理论界将迎来复杂系统研究的重大突破，正如霍金所言，21 世纪将是复杂系统研究的时代。在这里，我要特别强调，本书谈到的“理论”一词，指的是用广义的社会科学解释社会现象的理论，并不指涉人工智能算法、研究方法或方法论中的理论。

在管理学领域，复杂组织系统的研究将占据重要地位；而在社会、经济、政治领域，复杂社会系统的研究也将被提上日程。这一学术范式的转型，不仅体现在方法上向社会计算转变，更源于方法论的重大突破，使得其在理论上得以向复杂系统研究转向。这一趋势值得我们高度关注并深入研究。

二　历史总是不断地在重复

在正式展开新范式转移的探讨之前，我想先谈一段历史，即结构化数据与统计学进入社会学的历程。统计学融入社会科学的历程，与我们今日所见之人工智能、大数据等技术的兴起，有着异曲同工之妙。“太阳底下无新鲜事”，当我们回顾统计学如何逐步渗透并影响社会科学领域时，不难发现，当前我们正经历着一场类似的变革。

历史总是惊人地相似，它仿佛在不断重复其既定的轨迹。前人之路可以耀今，他山之石可以攻玉。

回溯往昔，我们不追问当初是谁第一次将统计应用到社会科学中，但要举一个最有名的例子。早在 1897 年，迪尔凯姆(Emile Durkheim)就已经以群体的自杀率作为研究对象，并在《自杀论》(迪尔凯姆，1897)一书中提出了有别于个人心理因素的因果解释。在这之后，数据就常以描述性统计的方式成为很多论证的依据。学者在展开理论论述的同时，通常会补充一些统计数据，以便进行下一步的推理。

同样，我们也并不追溯是谁第一次将回归和因果结构模型用于社会科学研究。一个重要的案例是弗兰科·莫迪利阿尼(Franco Modigliani)在20世纪50年代提出的家庭储蓄“生命周期”理论，那时计算机尚未普及，因果模型与统计计量的应用显得尤为艰难。那时的经济学家或社会学家，为了完成一个简单的回归分析，往往需要组建一个由十余人构成的团队，他们手持计算尺，进行烦琐而耗时的计算。莫迪利阿尼这样努力完成消费函数的因果分析，为他在20世纪80年代赢得了诺贝尔经济学奖。回望这一过程，是不是与我们在大数据研究初期所面临的挑战颇为相似？即利用有限的技术手段、不足的算力去处理海量数据，进行初步的统计分析——这是非常困难的。

进入20世纪60年代，社会学领域迎来了两位杰出人物——奥蒂斯·邓肯(Ottis Duncan)与彼得·布劳(Peter Blau)，尤其是后者，其社会交换理论影响深远。而且彼得·布劳在中国社会学学科的重建过程中扮演了重要角色。1967年，奥蒂斯·邓肯与彼得·布劳共同撰写的《美国的职业结构》(*The American Occupational Structure*；Blau & Duncan，1967)一书，彻底摒弃了传统社会学仅凭臆想构建理论的做法，转而强调从数据中挖掘理论的重要性，开展了翔实且大规模的社会调查。同时计算机的普及，使得他们得以用较复杂的因果结构模型对美国职业分层与流动现象进行解析。这一做法不仅是对过往研究方法的深刻反思，更是对未来研究方向的明确指引，标志着社会科学研究范式的重大转变。

与此同时，20世纪50年代的社会学领域也呈现一种复杂而矛盾的态势：一方面是对旧有理论框架的批判与反思日益激烈，尤其是对宏大理论以及默顿(Robert Merton)在1949年提出的中程理论的质疑；另一方面是对新兴数据驱动研究方法的拥抱与探索。这种双重性不仅体现了学科发展的内在张力，而且预示着未来研究路径的趋势。

进入20世纪70年代，随着两个关键因素的出现，社会学研究迎来了新的变革。首先，大型资料库的建立为研究者提供了前所未有的数据。美国政府每年投入巨资进行家户收支动态的大型资料调查，以及社会学领域的类似项目，如一般社会调查(general social survey，GSS)等，均标志着数据收集与分析工作向规模化、系统化方向迈进。这些资料库不仅涵盖了广泛的社会经济领域，还包含数以千计的变量，为深入研究奠定了坚实的基础。其次，电脑的普及与软件技术的发展极大地提高了数据处理与分析的效率。20世纪70年代，个人电脑开始进入人们的日

常生活，而专业的统计分析软件，如 SPSS、SAS 等，则使得复杂的回归分析等计算工作变得轻而易举。这一技术革新不仅极大地减轻了研究者的负担，还使得数据挖掘与预测模型成为可能，从而推动了社会科学研究向大型计量模型的方向发展。自 20 世纪 70 年代以后，因果模型与较大型、多变量的联立方程组的回归模型开始逐渐流行。其中，路径模型（path model）尤为一时之盛。

但到了 20 世纪 80 年代末，一群有识之士对这样的发展提出深刻的批判。在今日看来，其实就是一群对统计因果推论有深刻了解的学者对路径模型的“探索”能不能得到因果的结论表达了强烈的质疑。数据挖掘（data mining）技术正是这一时期的代表，它能够从庞大的数据集（如上所述的 GSS 包含数以千计个变量的数据集）中，选择一个我们想研究的被解释变量（dependent variable，一般在因果模型中，以 Y 表示），比如是否离过婚，然后从其他的千余个变量中选出最佳的预测变量（predictors，这里没有因果的含义）。通过复杂的统计算法，如向前选择（forward selection，即将与被解释变量相关性最高的预测变量排序，依照顺序将最前列的变量放入模型，直到解释能力增加值变得很小时终止）、向后剔除（backward elimination，即将与被解释变量相关性最高的预测变量排序，依照顺序将最后列的变量剔出全模型，直到解释能力减少值变得较大时终止）和所有可能（all possible，即将所有可能的预测模型都列出，比如有 1000 个预测变量，就会有 2 的一千次方个模型，然后采取简约法则，选出预测变量数较少，但模型解释力较大的模型）等方法，自动筛选出关系最紧密的预测变量，从而构建预测能力最强的模型。这一过程往往能够在保留模型精简性的同时（如最终保留 18 个变量），实现对未来结果的较准确的预测（平衡之道在于当大量预测变量加入，而衡量模型准确与否的指标，比如 adjusted R^2 增加值微乎其微时，就停步于目前的模型）。

但最好的预测变量就一定有因果意义吗？那时我们常说的一个“笑话”就是：“预测模型告诉我，下雨前我的膝盖会痛，但我的膝盖痛真的会造成下雨吗？”对于具备基本理论知识的人来说，很容易便能识别出某些现象之间的共变效应，如雨天与膝盖疼痛之间的关联，很可能仅是由于湿气增加与风湿病症的共变关系所致。如果我换一个地方住，或在气候条件改变的情况下，或过一段时间我的风湿病治好了，那么这个最好的“预测变量”就可能失效了。基于常识或理论基础的快速判断，凸显了那个时代单纯以统计方式进行数据挖掘在理解复杂现象时的局限性。

同样地，以今日观之，路径模型也是一个用以探索因果结构的方法，亦即各因

素之间如何形成联立的方程组从而共同对被解释量造成影响。预测变量中有两个因素，比如 A 与 B，在没有理论说明它们之间的关系时，A 与 B 之间可能存在四种关系：A 是 B 的因，B 是 A 的因，A 与 B 互为因果，A 与 B 互不影响。如果一个模型中有 18 个因素就形成 153 组双元关系，每一组又有 4 种可能。另外，任一预测变量和 Y 之间除了互不影响之外，其他 3 种因果关系亦都可能存在，这就形成了一个极大的因果关系路径网络。路径模型往往在多数因果关系的理论讨论中缺位，或只有简单几句推理的情况下，就得出一张因果结构图，然后在得到较好的估计结果之后，就径自宣布这些变量间的因果结构。对此我们也有一个玩笑话，就是："如果你对数据折磨得够久，它一定会招供的。"(If you torture data long enough, it will confess.)言下之意，在如上述的成千上万种可能的因果模型中，其中你想估计或不想估计的参数亦多如牛毛，只要你有耐性，一个一个地去试，资料总会给你想要的结果。

20 世纪 80 年代末 90 年代初，一群社会统计学领域的顶尖学者，如唐纳德·鲁宾(Donald Rubin；详见鲁宾模型，Rubin Model；Rubin，1974)、克利夫·克洛格(Cliff Clogg)、史密斯·赫伯特(Smith Herbert)、保罗·霍兰德(Paul Holland)以及他们的统计学盟友如大卫·弗里曼(David Freeman)等看到了这种统计分析的限制，发起了"统计如何才能得到因果推论"的大讨论。过去把"社会科学"理论排除在外，只从资料挖掘得到"因果"的想法被否定。随着方法论大论战的深入，学者开始致力于构建更为严谨的理论模型，并探索如何在鲁宾模型等实验性框架下，通过精确的实验方法揭示因果关系。这一过程涉及对实验收集资料与调查资料的比对，旨在模拟现实世界中复杂的因果关系，从而推动因果推论研究的深入发展。至 20 世纪 90 年代初，因果推论的大论战达到了高峰，并在此后逐渐平息。在这一过程中，上述基于相关研究的探索方法逐渐被更为严谨、注重每一对变量间因果理论的发展的研究范式所取代。在当今的社会科学领域，构建理论模型、明确模型与数据规范、排除虚假相关效应及其他潜在干扰因素，已成为确定因果关系的主流方式。

简言之，没有不由理论出发的实证，也没有不能被实证证明的科学理论。在因果模型与统计分析不断对话的情况下，终于建立了今天我们常见的演绎法加理论验证的"主流定量论文范式"——确定社会科学议题，穷尽相关理论回顾，通过逻辑完整的理论推演得到假设，导出因果模型，并完成研究设计，在时间或空间维度识

别中介或干扰变量，进而明确模型的核心要素，包括变量选取、测量方法、因果机制、因果函数形式、资料来源，等等。由此，我们确立了以演绎法与理论验证性研究为主的研究范式：命题构想→理论演绎（可能辅以定性研究）→提出假设（hypothesis）→研究设计（解释模型指定、变量指定、参数指定以及资料指定）→验证解释模型→理论肯证或否证→提出研究不足并导出下一轮研究。没有理论对研究设计的每一步给予指定（specifications），任何可以任意设定的参数，都可能让我们“操纵”研究的结果，所以理论回潮，回到了定量实证研究的核心位置上来。

但这套方法论并不反对在理论形成的过程中使用资料挖掘的方法进行探索性研究，所以两者是对话的、兼容的。只是在以调查为主的定量研究中，大多数情况下样本量都太小，不太能支持溯因法（abductive method）。溯因法又称作反绎方法，把探索成果发展成理论并加以验证。令人欣喜的是，鲁宾及其跟随者中不乏后来荣获诺贝尔经济学奖的杰出人物，他们的研究成果为我们理解因果关系奠定了更为坚实的理论基础。

统计学家鲁宾在 1974 年提出的“数据驱动模型”，即我们常说的预测模型，与“准实验”（quasi-experiment）方法得出的因果推论差距多大？准实验就是我们经常看到的，尤其在医疗领域中分“实验组”与“对照组”的那种实验，它奉行“能控制的就控制，不能控制的就随机抽样”（control what you can control，randomize what you cannot）。换言之，以一个药物对病情的作用效果为例，病人在医院中的环境是可控制的，如光照、营养、医疗照顾等；不可控制的，如个人的修养、之前的人生轨迹与健康史等。随机抽出两群足够多的人，分成两组。然后一组给药，另一组却给“让病人以为是药”的维生素，观察一段时间，药物与病情的因果关系会在实验中显现出来。

调查资料相对于实验资料被称为“脏资料”（messy data），因为它是在完全不受控的“开放环境”中访查到的众多变量，在复杂社会系统中，变量间交叉影响，形成网络。而且我们的认知有限，设计问题项时一定还有大量的重要因素没涉及，更何况复杂社会系统是动态演化的，除非该调查是针对一群人的多年期追踪调查，否则动态因素很难被考虑进去。这些都造成了我们的因果模型估计出来的回归系数与准实验估计出来的有一定的差距。尽管这些预测模型在预测能力上表现出色，但它们往往难以直接揭示变量之间的因果关系。换句话说，模型能够告诉我们哪些因素与结果高度相关，但并不能直接证明这些相关是导致结果的原因。因此，在享受数据挖掘带来的便利与高效的同时，我们需要对因果关系保持审慎态度。

三　另一种声音的呐喊

另一个有趣的“太阳底下无新鲜事”是“传统的”人文、社科学者对这种新发展的极端否定。1959 年，剑桥知名学者斯诺(C. P. Snow)在一个讲座上提出了“两种文化”理论[Snow，1959(2001)]。斯诺担忧的是，随着统计学等自然科学方法逐渐渗透至人文、社科研究领域，其实早在 19 世纪 30 年代社会物理学概念的提出便预示着这一趋势，他担心这种融合会削弱社会科学的独特性与人文价值。从某种角度看，笔者认同斯诺的部分担忧，社会科学确实蕴含了丰富的人类智慧、情感及价值观，这些元素是“不同的文化”，与自然科学存在本质差异，虽然社会科学是一个融合两者发展出来的学科，但其中确实有一些方法是自然科学难以触及或替代的。

同样地，面对来势汹汹的大数据与人工智能方法，芝加哥大学教授安德鲁·阿伯特(Andrew Abbot)发表文章，提出了人类认知方式中，“人工智能算法式思维”无法取代“联想式思维”，指出有些认知方式是不可逆的，但有些是周期性的，“社会物理学”至少来了三次(今天的社会学就曾在 19 世纪 30 年代被社会学创始人孔德以社会物理学命名)，如今再度回潮[①]。同样，“科学主义”与“数据驱动”都是来了又去，去了又来，尤其科学主义在 19 世纪末 20 世纪初盛极一时，但仍无法“统一”人类认知方式。这让我立刻想起斯诺讲的“两种文化”，当原子弹惊天爆炸，当电脑、资料库进入学者书房，当科学主义当道，总有思想家当头棒喝，告诉世人人文社科的核心是人文思想。阿伯特的文章也是如此。

笔者基本上支持安德鲁·阿伯特的想法。“联想性思维”一定是社会科学的核心，一时的“人工智能算法万能”“数据驱动”的呼声总会过去，社会科学研究还是会以理论为中心，而这些理论还是需要人文思想来浸润。

但笔者也认为安德鲁·阿伯特把人工智能算法式思维绝对化了(当然，也确实有一些社会计算学者是如此主张的)，立了个“稻草人”(与联想式思维两极对立的绝对化后的人工智能算法式思维)，再猛打“稻草人”。每一次“数据驱动”“科学主义”当道，最后都证明了思想与理论还是重要的，但每一次这样的变化又会促成社

① 可参见：亚历克斯·彭特兰(Alex Pentland)2014 年的著作《社会物理学：社会网络如何让我们更聪明》(Pentland，2014)。

会科学研究方法的范式革命，因为方法的改变会促成理论的重大更新。从纯粹只是孔德式的譬喻所构成的“社会学”（更确切地说是一种社会思想），到社会研究需要田野调查、历史文献，以至于“没有调查就没有发言权”，学者因此要“上穷碧落下黄泉，动手动脚找材料”，但其间也出现了韦伯、迪尔凯姆等思想理论大家，在方法和理论上都有了新范式。

20 世纪 60 年代，正如上述方法论大论战前因后果所述，又再次出现“数据驱动”“科学主义”。大型数据库的建立、问卷调查的推广、统计方法的进步、电脑和统计软件的普及、社会学界 20 世纪七八十年代流行的路径模型，导致方法论大论战的爆发，但最后还是回归到理论的重要性。这段时期的新数据、新方法同样地促成了理论的大发展，社会网络学派、理性选择学派、新制度学派带来了诸多新理论。

历史总是不断地重复，好像每隔五六十年，就会来一次“数据驱动”“科学主义”，而现在的这次是无结构化电子印迹数据加上人工智能算法以及动态模型电脑模拟。同样地，又一次会有人高举思想与理论重要性的大旗，强调人类智慧与认知方式的文理殊途，反击这些“新生事物”。就像阿伯特强调很多事情是“周期性的”，笔者完全同意。十余年来，笔者参加社会计算的研究，并在国内致力推广，一路走来更加坚信，这一轮的“社会物理学”浪潮仍会指向理论的回归，并发展出更多、更精微的理论，解释更多过去解释不了的现象。以古鉴今，历史总是一再重复这样的故事，从新数据到新方法出现，再到丢弃理论只要“数据驱动”，最后又是“正反合”，理论回潮，思想发光，而且新数据与新方法带来了新理论。

所以笔者反对将两种文化截然对立。

一边的“走极端”在于把理论丢掉了，这里再次强调本书讲的理论是社会科学针对社会经济现象提出解释的理论（本书所用的范例集中在社会科学研究领域，但也包括了其他相类似的研究，如教育、传媒等），不包括方法或方法论的理论，也不包括诸如宗教之类关乎意识形态的理论，以及人生哲学、美学这类理论。因为大数据数量庞大，人工智能算法诸如机器学习、深度学习提供了前所未有的预测能力，但其内部还是一个“黑箱”，我们不得窥知其堂奥，总会有人认为人工智能终将取代人类的“联想式思维”与灵光一现的智慧。虽然今天似乎对 AI 到底能走多远还见仁见智，但我相信，人类的理论思维是难以被替代的，“数据驱动”的研究在复杂、开放、动态的社会系统解释上是有止境的，只是笔者既好奇这个止境在哪里，又持开放态度，而且参与了相关研究，想探究这个止境的边界。

另一边的“走极端”则是看到了这个“止境”，于是索性连新数据与新方法都不要了。尽管社会科学理论中存在不可替代的人类的情感、思想、大理论与联想思维，但这并不意味着我们应当全盘否定科学方法在社会科学领域中的应用。我们反对把科学无限扩张为无所不能的“科学主义”，但也没必要把科学方法能拓展的空间也一起忽视掉。相反地，随着时代的发展，特别是随着结构化数据加统计学到大数据加人工智能算法、电脑模拟模型等技术的广泛引入，社会科学领域正经历着前所未有的变革与融合。这并不意味着我们要放弃社会科学研究的独特性与人文价值，而是要在尊重其本质的同时，探索自然科学方法如何能够为我们服务，推动社会科学理论的进步与发展。

新数据、新方法固然会带来方法的范式转移，但其真正的内核是带来新理论，如复杂系统理论。笔者相信这就是这一波理论的关键发展，至少是关键的理论发展之一，这也是我在最后一章想要探讨的问题，而且，还会有另外一本书专门讨论如何结合社会计算与复杂社会系统研究。不可避免地，方法和理论的范式都会转移，新方法与新理论的新范式才是不可逆的核心，就好像我们今天不会再回到 20 世纪 50 年代的方法以及只讲之前的理论了。因此，笔者认为，在社会科学领域，我们既应谨慎地了解新数据、新方法也是有止境的，又应积极迎接自然科学方法带来的新机遇与挑战。两者绝对不是非此即彼的关系，而是可以相互补充、共同促进的。怕的是不相互理解，不相互包容，走两极对立，把原本多元宽阔的研究变成了意识形态的你死我活的对立，把新知识、新议题、新理论的康庄大道越走越窄，直入死巷绝境。

只有相互学习与包容，我们才能在保持社会科学独特性的同时，不断推动其向更加科学化的方向发展。

四　避免走入极端

在本书的一开始笔者不厌其烦地回顾了一段历史，因为知古鉴今，在 30 多年后回望那一段历史，发现今天社会计算的发展好像依然跟随其后，亦步亦趋。太阳底下无新鲜事，我们再以历史宏大的眼光看新数据、新方法的发展，才不会迷失在社会计算每天推陈出新的人工智能算法丛林中，可以从过去范式转移的历史中得

到很多启发。

一是，随着统计学的广泛应用，越来越多的学者开始采用定量研究方法，使得这一领域的研究面貌焕然一新。时至今日，从社会学的整体发展趋势来看，以美国为例，一些主要核心期刊（如 ASR《美国社会学评论》等权威期刊）所发表的论文中，有七成甚至更多采用了定量研究方法。这一数据充分说明，统计学在社会学界的普及程度已达到了一定的高度，同时也反映了当前学术界对于量化研究的重视与认可。

二是，从方法论大论战得到的结果来看，问卷调查、结构化数据以及统计方法，从简单的描述性统计，发展到探索性的研究，并一度以为那些方法可以得到因果关系的结论，但最终得到的还是理论的回归。所以新数据、新方法的目的不只是方法的范式转移，更重要的是促成新理论的发展，让我们对社会现象有更深刻的了解。比如 20 世纪 90 年代最主要的社会学理论流派，如社会网络学派、新制度学派以及理性选择学派，就不再是 20 世纪四五十年代结构功能学派那样的大理论，而是呈现出大理论启发中程理论的发展路径，比如社会网络理论（social network theory，Granovetter，1973；White，1963）就发展出了结构洞理论（structural hole theory，Burt，1992）、社会资本理论（social capital theory，Nan Lin，1981；Coleman，1990），在中国则发展出关系社会学理论等。这些中程理论能够明确变量间的因果指向，即哪些变量可以作为因变量，哪些社会现象成为其对应的结果，再落实到一个具体聚焦的研究议题时，就可以发展出相应的假设和理论模型，从而指定研究设计，取得资料，并对理论加以证实或证伪。

三是，新数据、新方法不是用来取代老理论与老方法的，而是会产生融合，比如，定性研究并不会因为统计方法的大行其道而消失，反而不断改进，并与定量研究相辅相成，因此众多学术著作均致力于将两者融合产生混合方法（mixed method；Small，2011），以提供更全面、更深入的分析视角。展望未来，我们可以预见一个趋势：尽管大数据分析、人工智能算法与系统动态建模等方法会在可预见的未来渐成主流，但传统的定性与定量研究方法不会消亡，而且诸多方法会熔一炉而冶之，产生新的“混合方法”。随着理论的精进、解释现象的有效、预测模型的精准、应用范围的扩大，研究人员的需求呈数倍甚至十数倍的增加，反而促成了传统定性、定量研究人员可能在某些领域保持规模甚至增加其从业者数量。

然而，值得注意的是，任何一个科学理论与方法都有止境，只是有的来得快，有