机器学习与社会科学中的因果关系：一个文献综述

郭 峰 陶旭辉*

摘 要：因果识别是社会科学实证研究的焦点，而在大数据时代，机器学习为因果识别带来一些新的机遇与挑战。本文重点总结了机器学习对因果关系识别的价值：更好地识别和控制混淆因素、帮助更好地构建对照组、更好地识别异质性因果效应，以及检验因果关系的外部有效性。同时，本文还讨论了在大数据和机器学习广泛应用下，可能存在因果关系在某些情形下变得不再重要、大数据和机器学习会让因果效应识别更加困难，以及部分机器学习算法缺乏可解释性等情形。本文有助于拓展社会科学研究者的工具箱和思想库。

关键词：机器学习；因果关系；大数据
DOI：10.13821/j.cnki.ceq.2023.01.01

一、引 言

随着互联网、电子技术等对我们生产生活的日益渗透，社会科学研究中，数据来源越来越丰富，文本、图像、音频、视频、遥感等大数据都成为社会科学研究者的重要数据来源。而伴随着大数据的大规模应用，擅长处理这种大数据的机器学习方法也成为社会科学家工具箱的重要组成。目前，机器学习方法在社会科学中的用途很广，特别是在数据生成、变量预测等方面，机器学习算法已被社会科学研究者广泛接受（Mullainathan and Spiess，2017；黄乃静和于明哲，2018；王芳等，2020；洪永淼和汪寿阳，2021）。

在经济学、政治学和社会学等各类社会科学研究中，特别是最近二十年的社会科学研究中，识别因果关系已经成为重中之重（Angrist and Pischke，2009；Abadie and Cattaneo，2018）。由于在进行数据预测中，仅仅知道变量之间存在相关关系就已经足够，而因果关系并不是必须的（Kleinberg et al.，2015）。因此，很多机器学习算法也就忽略了变量间的因果关系，而只关心结果变量和特征变量之间是否存在相关关系，这造成了机器学习与社会科学主流实证方法之间存在一定的隔阂。但机器学习方法与因果关系识别之间并不全然是冲突的。机器学习方法凭借其自身优势，对因果关系识别有着非常重要的价值。一些文献已经开始应用机器学习方法来帮助识别因果关系。但在社会科学领域，使用机器学习方法识别因果关系的文献还相对较少。因此，本文聚焦于社会科学领

* 郭峰，上海财经大学公共经济与管理学院、北京大学数字金融研究中心；陶旭辉，上海财经大学公共经济与管理学院。通信作者及地址：陶旭辉，上海市杨浦区武川路 111 号凤凰楼，200433；电话：13184578885；E-mail：taoxuhui@163.sufe.edu.cn。作者感谢国家社会科学基金重大项目（18ZDA091）、国家自然科学基金青年项目（72003214）、上海市哲学社会科学规划课题（2020BJB004）的资助。感谢两位匿名审稿人提出的宝贵意见，文责自负。

域，对机器学习方法在因果关系识别中能起到的作用，以及其给因果关系识别带来的新挑战，进行一个文献综述，以期丰富社会科学研究者的工具箱和思想库。

本文与已有关于机器学习的综述性文章的主要区别在于，我们聚焦于社会科学领域内，着重介绍机器学习对因果关系识别的价值及其带给因果识别的新挑战。与本文关系密切的综述性文章包括 Varian（2016）、Guo et al.（2020）等关于机器学习与因果识别的介绍性或综述性文章。不过这些文章的目的是面向机器学习研究者介绍因果关系的基本概念和逻辑，以及如何在机器学习建模中引入因果关系的思维。值得一提的还有 Athey and Imbens（2017），她们也介绍了机器学习方法及其对因果识别策略的一些拓展。但相比之下，本文与之仍有较大差异：她们的文章主要是综述实证策略及其最新进展，机器学习的介绍并非其主要内容。而本文则更多地介绍在某些情形下，现在主流实证研究方法的不足，进而从各维度综述如何运用机器学习方法拓展因果关系识别的适用边界。相比她们对机器学习的简要介绍，本文则针对机器学习在因果识别方面的应用有比较全面和深入的梳理。

因此，本文的主要贡献是：首先，以尽可能简洁的表述全面而细致地介绍了机器学习对因果关系识别的重要价值、代表性应用场景，为相关学者拓展因果识别方法的适用场景提供新思路。其次，本文也简要讨论了大数据和机器学习的广泛应用给因果识别带来的新挑战。当前，各种因果识别方法在社会科学实证研究实践中扮演了越来越重要的角色。但越是如此，越能发现这些方法在实践中存在各种局限。这使得我们越发觉得，向社会科学研究者展示和总结如何利用机器学习方法更好地进行因果识别，拓展因果识别主流方法的适用边界，并结合具体案例勾勒利用机器学习方法进行因果识别的步骤，会给研究者提供更多的灵感。更为重要的是，我们在综述文献时深刻体会到，机器学习对因果识别的意义并不仅仅是提供新数据和新方法，还可以帮助社会科学家发现新问题。

本文余下部分安排如下：第二部分详细讨论机器学习如何帮助社会科学研究者更好地进行因果关系识别；第三部分则讨论在大数据和机器学习的广泛应用下，因果识别面临的新挑战；最后一部分是全文的总结和展望。

二、机器学习对因果识别的意义

最近几十年，因果关系识别已经成为经济学、政治学和社会学等社会科学各学科实证研究的经典范式。诸如双重差分、断点回归、随机试验等因果识别经典方法，已经被社会科学各学科广泛应用。与此同时，社会科学研究中，数据来源也变得越来越丰富。而随着社会科学研究数据和工具的逐渐丰富，社会科学研究者以及政策制定者变得越来越野心勃勃，他们不再满足于获得两个变量之间简单的因果关系。他们还希望了解如何在全新的、非结构化的、高维度的、高频率的大数据中挖掘出一些新问题，以及在探究某些问题时即便传统方法因假设无法满足而失效时，有什么其他实证工具可以备选，等等。诸如此类的问题，都是传统社会科学研究者关心却没办法很好回答的，甚至超出了传统社会科学研究范式，而回答这些问题也正是引入机器学习的必要所在。简言之，机器学习对于因果关系识别的特殊价值和意义是它可以拓宽因果识别经典方法的适用边界。关于机器学习对因果识别的意义，我们在本部分进行详细讨论。

（一）更好地识别和控制混淆因素

社会科学在实证分析中的重点是考察某个处理变量对结果变量的因果效应，但处理变量往往并非随机的。在估计因果效应时，一个常规做法是假定如果控制了一些混淆因素，那么我们关心的核心变量就具有某种随机性，从而可以估计其对结果变量的因果效应。因此问题的核心就在于识别、控制这些被称为混淆因素的特征。选择控制变量在传统社会科学实证分析中的一个常规操作是依据理论分析或理论直觉，但这可能带来两个问题：一是控制变量被人为操纵，以获得统计上的显著性（Fafchamps and Labonne，2017）；二是在大数据时代，依据理论分析或理论直觉控制变量的选取有时候会变得非常困难，因为非结构化大数据的一个非常显著的特征就是高维稀疏：潜在的控制变量可能成百上千个，而最终能被用上的可能只有数个。为此，Belloni et al.（2014；2019）等提出一种称为“Post Double Selection”的数据驱动的机器学习策略：首先通过 Lasso 等附带正则项的机器学习算法，经过交叉验证等方法，识别出一组对结果变量有解释力的变量，进而重新将结果变量对这些挑选出的特征变量进行普通的线性回归。Chernozhukov et al.（2018）还提出一种较为重要的双重机器学习方法来应对传统方法中较难克服的混淆因素函数形式不确定、维度诅咒、正则化偏差的问题，其核心思路是从处理变量中过滤掉（partial out）协变量的影响。其应用包括研究在线劳动力市场垄断情况（Dube et al.，2020）、大型会计事务所更高审计质量效应问题（Yang et al.，2020）。

对于哪些控制变量应该被控制，而哪些变量不应该控制，因果图也有一套行之有效的准则。社会科学实证研究中，在控制混淆因素时，通常会借助理论谨慎地选取哪些变量必须进入模型，但很少考虑哪些变量不应该控制。这种选择是传统社会科学研究过度关注“一致性”的结果，而实际上这种倾向可能会带来样本选择偏差的问题。例如“美貌（D）→明星（X）←才华（Y）”，即一个人拥有才华会使其可能成为明星，一个人长得漂亮也会让其更有可能成为明星。但一旦控制明星（X）这个变量，分析美貌（D）对才华（Y）的影响时会得出结论：越丑的人，越有才华。而实际上，才华和美貌本身是相互独立的，导致错误结论的原因是成为明星往往是拥有才华和美貌之一即可，而控制 X 变量，即以明星为条件（明星＝1）时，反而产生了样本选择偏差。这在因果图中被称为“对撞偏倚”或“辩解效应”，即控制处理变量（D）和结果变量（Y）的共同结果而产生的偏误（Pearl and Mackenzie，2018）。这也是为什么这里的 X 不可以作为异质性变量的原因。再譬如，很多时候我们还会因错误地控制一个机制变量，导致估计偏误。以“教育→职业→收入”为例，即教育回报的研究。其中，职业实际上并不需要控制，因为教育程度往往会决定职业选择，继而影响收入。在因果图中，控制了一个这样的节点非但不能减少偏误，反而可能切断一条中间机制，从而带来估计偏误。而根据因果图，我们真正需要控制的是能够同时影响处理变量和结果变量的共同原因变量，这才是传统意义上的混淆变量。关于因果图更详细的讨论，特别是其对社会科学实证研究的启示，可以参阅 Imbens（2020）、Cunningham（2021）等文献。

上述“Post Double Selection”策略也可以使用在工具变量的挑选上。工具变量法是社会科学家解决内生性、控制不可观测混淆因素，实现因果推断时非常倚重的方法（陈云松，2012）。工具变量方法的核心是寻找一个外生的，但同时又与内生变量相关的变

量，即对预测内生变量变动有帮助的外生变量。而预测又是机器学习擅长之处，因此，工具变量方法的第一阶段分析完全可以使用机器学习方法来分析内生变量和工具变量之间的关系。而且，在某些情形下，可能并不存在一两个性能非常突出的工具变量，而是需要在众多的潜在工具变量中寻找最佳的工具变量。如果工具变量太多，可能会产生弱工具变量问题，而工具变量数量比内生变量多一至两个时才是最优的（Bollen，2012），此时可以使用机器学习算法来挑选工具变量。Belloni et al.（2012）推荐使用 Lasso 方法，在一些潜在的工具变量池中，挑选与内生变量最为相关的变量作为工具变量，然后重新进行普通的两阶段最小二乘法回归。在社会科学实证中，已经有一些文献利用 Lasso 等正则化方法来挑选工具变量，例如 Qiu et al.（2020）通过 Cluster-Lasso 方法在一组天气等变量池中，寻找各地新冠肺炎病例的最佳工具变量；Gilchrist and Sands（2016）、方娴和金刚（2020）利用 Lasso 方法，选择最优的天气与空气质量变量作为电影首映周票房的工具变量，进而考察电影首映周票房对随后几周电影票房的影响。此外，Hartford et al.（2017）还提出了一种“反事实预测＋IV”的 Deep IV 的方法，与普通的工具变量法相比，其优势在于估计时不需要满足线性假设，因此适用范围更广，而且被证明有效性更好。

（二）更好地构建对照组

反事实结果不可观测是一个因果效应识别的“根本性的问题”（Holland，1986）。但如果能够为处理组寻找到非常合宜的对照组，就可以通过对照组来构造出倘若没有处理政策的发生处理组应该具有的反事实结果，从而得到科学的因果效应估计。在传统因果识别的文献中，双重差分法、匹配法、合成控制法、断点回归法、随机试验法等都是通过为处理组构造合适对照组，进而实现反事实结果估计和因果效应识别的思路。但这些方法又各有苛刻的适用条件，因此在本小节我们分别阐述机器学习在这些方法框架下可能发挥的作用，从而拓宽这些因果识别方法的适用边界。

1. 机器学习与双重差分法

双重差分法是实证分析最为常用的方法之一，但其有一个非常重要的前提就是处理组和对照组在政策之前，保持相同的趋势，这是利用对照组构造处理组的反事实结果的前提条件，即所谓平行性趋势条件。然而，在某些情况下，处理组和对照组在处理政策前可能并非同样的线性趋势，而一旦变量间是非线性的关系，使用传统的线性回归来进行双重差分法的估计就会存在问题。而机器学习在构建模型的过程是数据驱动的，它能够通过数据信息有效地抓住变量间线性或非线性的关系，并尤其擅长处理变量间复杂的非线性关系。对此，我们以 2019 年年末暴发的新冠肺炎疫情的相关研究为例来阐述机器学习方法在双重差分法框架下的应用。在研究该次新冠肺炎疫情时，学者们会轻易地将 2020 年视为处理组，2019 年视作对照组，直接进行双重差分法分析。但由于春节等因素的干扰，2020 年疫情前后，与 2019 年疫情对应日期前后，可能并不满足相同的线性趋势，进而不满足平行性趋势。因此，Guo et al.（2022）同样利用上述双重差分法框架，但将线性回归改成了机器学习中擅长处理非线性关系的梯度提升树，利用 2020 年疫情前数据和 2019 年同期数据，成功预测了如果没有疫情发生，2020 年应该具有的反事实结果，从而获得了疫情对线下微型商户冲击的科学估计。

另外，我们在使用双重差分法进行政策评估时还需要注意是否存在外溢效应或再分配效应。外溢效应或再分配效应是指那些被政策干预的组别，通过其他渠道，把政策的影响也传递给了非政策干预组，进而导致政策效应低估或高估。以 Cicala（2022）为例，这项研究评估了美国将发电权从国家计划放开到市场决定所带来的收益。而逐年逐地区推行的政策，完全适用于标准的多时点双重差分的分析框架。但是，这其中就出现外溢效应，而渠道可能是影响发电的重要因素——燃料价格。例如 A 地电力实行市场决定，相邻的 B 地电力实行国家计划。当 A 地电力可以实现市场交易时，可能会影响相邻地区的燃料价格，进而影响 B 地发电。而且因为产能原因，相同燃料价格的变动对不同地区的电量供给影响也是不一样的。因而，传统的双重差分法在不满足独立性条件假设、异质性对照组的情况下往往难以识别精准。研究者会合理地想到，如果能够通过自身历史数据构建自身的反事实，便能很好地解决上述问题。此时，便转化为预测问题，而预测是机器学习的优势所在，Cicala（2022）便利用随机森林预测了假如不是市场决定时的发电量的反事实结果。

2. 机器学习与匹配法

在进行政策或项目评估时，个体在一组既定的协变量下，常常会因为自身特征的分布差异导致参与项目或政策的可能性存在差异，也就是我们通常所说的“自选择偏差”。这时候，我们就需要寻找那些与参与者有相同特征的非参与者作为可比对象，进而实现因果关系识别。传统社会科学实证方法推荐使用一种称为匹配的方法，而机器学习在匹配法上也有不少助益。

首先，在样本较少的情况下，传统匹配方法很容易因为数据不足而产生无法满足“共同支撑”的假设，也即因处理组和对照组特征差异太大而无法匹配到可比对象的情况。此时，它仅能为处在共同支撑区域的个体找到相应的对照组，当因为不满足共同支撑假设而无法为某些个体找到相应的“反事实”时，传统的做法就是丢弃这部分样本，这进一步减少了处理个体。这在样本较少的情境下，估计出的效应并不是平均处理效应，甚至无法保证样本内有效。在这种情况下，Hill and Su（2013）提出的采用贝叶斯加性回归树算法（Bayesian Additive Regression Trees，BART），能很大程度上削弱这一问题的影响。这一算法相较于倾向得分等传统匹配方法的优势在于：传统匹配法的前提假设是模型必须设定正确，而 BART 则不需要；最为关键的是，传统匹配策略更多地考虑给予能预测处理变量的协变量较高的权重，而完全忽略结果变量中关于共同支撑的信息，BART 则可以根据后验标准差分布提取这一信息。换言之，传统匹配法并不能保证选取的协变量是否对结果变量也有很好的预测能力，从而无法得知该协变量是否是有匹配价值的变量。错误地给予对结果变量没有解释力的变量更高的权重，虽不会影响样本的随机性，但是很容易在匹配时损失大量样本。更通俗来说，在进行匹配的时候，并非任何变量都适合用来作为匹配变量，这会为实现匹配导致估计的样本损失。正如在预测图片中吃鱼的是否是一只猫的时候，不应该用猫的毛发颜色特征（变量）去匹配一样，因为猫的毛发颜色并不是猫的特质。而相比传统匹配法，BART 在这一研究中样本损失要少得多，结果也更为稳定（Hill and Su，2013）。另外，一些观点认为机器学习仅适用于样本更大的大数据，但事实上，对于小样本数据，采用机器学习方法相比传统实证分析可能更能获得更好的结果，原因是机器学习擅长处理特征数大于样本数的高维稀

疏数据，如贝叶斯倾向得分匹配（An，2010）。

其次，机器学习方法可以帮助处理高维数据的匹配问题。文本、图像等高维度数据中包含大量信息，社会科学研究者试图从中挖掘新问题，此时如何匹配高维数据则成为一个迫切的需求。比如，传播学研究者可能关心在中国社交媒体中，有被审查经历是否会增加其再次被审查的概率。对此，Robertsy et al.（2020）设计了一个机器学习方法，来解决文本数据的匹配问题，进而应用在上述问题的因果识别当中。具体而言，他们使用一个合适的主题模型来表征文本，然后再用倾向得分来匹配文本主题。其实，对于文本数据的匹配，核心在于两点：如何度量文本，即如何将文本数据表征为一个低维数据；以及如何定义文本“距离”以描述文本之间的相似性。对于这两点，Mozer et al.（2020）都进行了非常详细的讨论。

最后，机器学习方法也为匹配法提供了稳健性检验。如果将是否成为处理组视作结果变量，控制变量视作特征变量，那么这就成为一个典型的分类问题，机器学习方法在分类问题上的出色表现就为匹配法提供了一个匹配验证分析。以判别分析法为例，Linden and Yarnold（2016）将分类算法用于匹配，其基本思路为：首先通过数据训练出一个最优分类算法，这个算法可以通过特征变量完美地预测哪些人应该在处理组，哪些人会在对照组；然后再将该算法应用于已经通过传统匹配法匹配好的两组样本，看通过该算法是否能成功将已经匹配好的研究对象分开，如果不能，那么我们可以认为匹配是成功的，处理组和对照组样本具有可比性。这种运用机器学习在分类方面的优势，可以实现对传统匹配方法的稳健性检验。

3. 机器学习与合成控制法

在处理组非常独特，难以寻找到合适对照组的情况下，也可以考虑利用众多对照组“合成”一个合适的对照组（Abadie and Gardeazabal，2003；Abadie et al.，2010；Abadie，2021）。合成控制法适用的前提是在处理前处理组与其他众多对照组之间的拟合关系，如果没有处理政策发生，则能够在处理政策后，仍然保持不变。

传统合成控制方法在社会科学研究中已经得到广泛应用，但也存在一些局限。例如，当合成控制法中潜在的对照组个体数量较少，以及处理组个体属于异常点时，传统的合成控制法会出现权重参数无解的情况，也即是无法合成的情境。比如我国上海，常因为其特殊的经济特征，使得传统的合成控制方法无法通过其他城市加权合成（刘甲炎和范子英，2013）。这时，Doudchenko and Imbens（2016）提出基于机器学习思想的新合成方法便适用这类存在异常点的情形。该方法在双重差分法和合成控制法基础上进一步放松为更加一般的线性组合函数来构建反事实，即可以赋予对照组个体负数权重以及权重和可以不为 1。这种合成带来两个优势：第一，因为结合双重差分方法，允许不同个体存在持久性的差异；第二，因为放松权重为非负以及权重和为 1 的约束，则当处理对象是一个异常点时，我们仍有可能找到一组权重参数的解。

此外，也可以利用机器学习方法中的正则化方法对对照组个体进行筛选，这样即便在潜在对照组非常多的情况下，这一方法依然适用，还可以防止过分外推效应。例如，Abadie et al.（2010）就指出，当面临大量对照组用于合成时，传统合成控制法会产生估计结果无效的问题。这是因为，权重约束的情况下，这一传统方法会因为无法捕捉一些权重为负数的对照组个体而产生有偏估计。Kinn（2018）就在具体实例中证明加入

Lasso 惩罚项的合成控制方法可以减少 71%的估计偏误。此外，Guo and Zhang（2019）在研究襄樊市更名为襄阳市对经济增长的影响时，也使用了机器学习算法（Lasso 和 Elastic Net）进行控制个体的筛选，以便为襄樊市合成出一个更好的对照组。

再者，Abadie（2021）还进一步强调当结果变量波动较大或不稳定时，传统方法没法测度真实的政策效果，因为经常会为了匹配政策前的趋势而产生过拟合的问题。事实上，我们常使用的日度、月度等高频数据都具有这种季节性波动。以 Cole et al.（2020）为例，其在研究武汉封城对空气污染和健康的影响时，就采用了随机森林的方法用于剔除这种季节性或时间趋势噪音。除此之外，由于武汉市的独特性，传统的合成控制方法也无法很好地合成武汉市封城以前的趋势，为此他们采用 Ben-Michael et al.（2021）提出的带有岭回归的增强合成控制方法，可以放松对照组权重必须为正的假设，进而相比传统的合成控制方法能实现在封城前对武汉市较好地拟合，并通过对增强合成控制下的权重与传统合成控制下的权重距离的惩罚，来调整合适的泛化水平。

最后，在合成控制法应用中，我们还可能面临政策干预前的期数相对较少，传统的合成控制很难完全复制处理地区的经济特征与干预前的结果变量的情形，这时合成估计会产生“内插偏差”（interpolation bias）的问题。而放松约束条件并施以正则的机器学习方法是可以减少偏差存在的。具体实践中，Kumar and Liang（2019）在考察美国得克萨斯州 1998 年房地产信贷制度改革对经济增长的影响时，就利用 Doudchenko and Imbens（2016）提出的机器学习算法对该州进行了合成控制，而这篇文章的问题就在于数据中政策干预前期的期数相对较少。

4. 机器学习与断点回归法

断点回归方法典型的应用场景是将配置变量低于（或高于）某一阈值的个体作为处理组，而在另一侧的个体作为对照组。断点回归的逻辑是如果我们仅关注断点附近的个体，则可以很大程度上保证处理组和对照组之间可比，从而估计因果效应。

但是，断点回归在具体应用时，一般会假定结果变量与配置变量之间在断点附近是线性、多项式（如二次项、三次项）分布，或者局部线性分布，但这些往往依赖于研究者对函数形式的事先设定，对于变量间关系是否是非线性并没有检验。而对于传统的断点回归而言，函数形式的错误设定会导致估计存在一定偏差。为此，Branson et al.（2019）提出了一种贝叶斯非参的方法结合 Gaussian 过程回归，这种方法的优势在于估计时不会过分依赖函数形式的设定。因此，即便结果变量与配置变量在断点附近存在非线性关系，也可以获得一致的估计。进一步，他们将这个方法应用于研究 NBA 第一轮选秀对篮球运动员的表现和上场时间的影响，对比传统断点回归的局部线性回归和稳健局部线性回归，他们发现 Gaussian 过程回归的结果与实际情况更为吻合。

不仅如此，断点回归运用的首要任务是寻找合适的配置变量及其断点，在传统的断点回归设计中，配置变量一般都是一维的，研究者可以通过对研究问题背景的分析，明确断点的位置。但如果考虑多维的配置变量，则断点的具体位置就变得不直观了，甚至无法通过人工观察而确定，此时可以使用机器学习的方法来自动判别断点的具体位置（Herlands et al.，2018）。

更为重要的是，断点回归有着较为苛刻的假设条件，如个体分配概率在阈值左右存在跳跃、配置变量必须是连续的、个体不能精确控制或操纵配置变量使之超过阈值等。

而在实际应用中，这些条件往往很难被满足。譬如，个体会通过一系列标准判定阈值位置，并主动跨越阈值而使得断点回归失效。Narayanan and Kalynam（2020）结合机器学习的断点回归设计给我们提供了解决上述问题很好的案例应用。他们利用客户已有的行为数据，评估采用定制型营销干预是否会影响客户消费行为。具体而言，基于高维的消费者行为特征数据，如浏览记录、购买历史等，通过一个复杂的机器学习算法将特征合成一个倾向得分作为配置变量，企业依据商业目标设定门槛。这样的优势在于：基于复杂算法和大量行为特征得到的分数是一个连续值，个体分配概率在阈值左右也存在跳跃，更为重要的是得分和阈值无法被个体所观测从而个体无法操纵自己的行为。

最后，考虑到断点回归的核心逻辑在于估计出如果没有政策冲击的发生，那么结果变量在断点右侧应该具有的分布（反事实结果），因此便可以利用断点左侧的数据，对结果变量与配置变量（以及其他变量）之间的关系进行建模，进而将建模参数泛化到断点右侧，从而估计断点右侧如果没有处理政策的话应该具有的“反事实结果”，而确保泛化能力则正是机器学习算法的优势所在（Imbens and Wager，2019；王芳等，2020），但我们尚未在社会科学文献中发现这一思路的具体应用。

5. 机器学习与随机试验法

随机试验被认为是因果识别的黄金法则，而随着试验条件和试验方法的进步，越来越多的社会科学研究者通过随机试验的方法来进行因果识别。在进行随机试验分析时，我们通常采用多元线性回归，而不是简单的报告处理组和对照组之间的均值差异，目的是为了通过调整协变量来减少因果效应的方差。但当协变量个数和样本数量相当的时候，回归的结果可能会因为过拟合而影响其结论的外部有效性。在这种情况下，Bloniarz et al.（2016）指出基于 Neyman-Rubin 模型下的 Lasso 方法能够保证估计量更有效，获得一个渐近方差的保守估计以及更为紧凑的置信区间。在社会科学具体研究中，为获得更为一致的估计，擅长处理高维数据的双重机器学习法，以及擅长处理非线性、多值处理变量的支持向量机也被运用到随机试验中（Chernozhukov et al.，2018；Imai and Ratkovic，2013）。

虽然随机试验是识别因果关系的非常好的办法，但因其成本高昂，在社会科学中其实运用较少，而在互联网科技公司中，随机试验（A/B test）则已经成为评估一项产品（政策）商业效果的重要途径。有时候一些公司为了考察多个变量之间的因果关系，例如不同维度的产品组合的市场效果，可能会发起很多次随机试验。比如，互联网视频公司可能希望了解哪一种类型的节目（如搞笑 vs. 严肃，短节目 vs. 长节目）会影响收视者的行为（如增加观看时间，订阅这个频道）。但是一旦发起多次随机试验，即调整节目类型，观众可能会产生比较大的反应，例如不再观看该频道，导致客户流失。加之传统 A/B test 很依赖于统计显著性，即使当处理组明显优于对照组时，其仍需要保证对照组有足够的样本，才能获得统计显著性。这样在花费大量时间的同时，又很可能造成用户流失。特别是在面临多个处理变量的时候，该缺陷更为明显。为此，Athey and Imbens（2019）提到一种改进随机试验的多臂老虎机（multi-armed bandit）方法。这一方法的基础是贝叶斯更新：当观测到某一处理组明显更好时，就可以将更多的用户增加到这个处理中，从而更快地找到因果效果最好的处理政策组合。其优势在于它识别最佳臂（最佳处理）所需要的试验次数和需要的样本远低于简单的多重 A/B test。更进一步，对于

互联网科技公司，在无法精准地定位客户的情况下，无差别地投放广告或播放节目可能不仅是浪费，甚至会激怒消费者，进而产生负面影响。而上文提到的机器学习与断点回归设计的结合在互联网市场营销干预领域的探索是一类很有启发性的替代性试验研究（Narayanan and Kalynam，2020）。

在社会科学简约模型（reduced form）中，随机试验被认为是政策评估最好的识别设计。但是它依然没办法很好地回答什么样的政策是最优的，或者说什么样的政策是收益-成本最大化的。而在某些情形下，机器学习方法可以提供一些解决思路。Knittel and Stolper（2021）就将因果森林算法应用于一个大规模行为干预试验的评估，讨论发送家庭能源报告是否有助于推动家庭节能，这一研究对传统随机试验有很好的借鉴意义。具体而言，能源报告的主要内容是告知用户相比其邻居的能源使用情况，以及相应的节能建议。使用因果森林进行高维异质性分析，发现有些群体接收到家庭能源报告后其能源消耗下降，但是有些特征的群体反而上升。进一步，研究者依据因果森林获得的异质性因果效应的结果，再次针对性地设计干预政策，即仅对具有正向处理效应的家庭发送家庭能源报告，发现社会效益会额外提升 12%—120%。这篇文献的贡献不仅在于细微粒度因果效应的发现，还在于针对因果效应结论有目标地进行“多轮干预”的试验思路，这一点值得随机干预试验实操时借鉴。

（三）更好地识别异质性因果效应

如上文所述，因果效应往往只能在总体样本上取得，即平均因果效应。然而，因果效应在不同的群体，甚至在不同成员之间都很可能有所不同。因此，异质因果效应分析对于社会科学实证研究而言意义重大。通过分析异质性因果效应，可以得到关于稀缺社会资源在非平等社会中的分布和关于社会政策的重要见解（Brand and Thomas，2013）。

在异质性分析中，我们很多情况下是不清楚处理个体的哪些特征是存在异质性的。传统的做法是将交互项逐个加入，或者一次性将交互项放入模型中。如果采用前一种做法，则必然耗费大量精力，并且可能有遗漏变量的风险；如果选择后者，则每增加一个交互项，就会带来一个假设检验，进而增加推断错误的概率（Davis and Heller，2017）。如何在成百上千的协变量中筛选变量并且保证计算的可行，这是机器学习的优势所在。例如，Knaus et al.（2022）发现一个奇怪的现象，他们通过数据发现瑞士花费巨大的就业培训项目对就业的影响竟然是负面的，这显然违背了政策制定的初衷。他们进一步用 Post-Lasso 算法在 1 268 个特征变量中进行筛选，对几乎所有可能的样本分组进行异质性分析，发现原因是培训项目分配给了那些处理效应为负的人，也即那些已经掌握了就业技巧的人更多地参加了该项目，反而造成其错过就业最佳时期。文章依据异质性因果效应的估计结果认为，将培训机会更多分配给那些具有更高的处理效应的群体，可以减少就业率的消极影响大约 60%。

传统方法分组比较异质性除了面临选择分组变量的问题，还存在连续变量如何切分的问题。而机器学习的树模型在变量如何选取、连续变量在哪里切分，以及如何避免过拟合问题等方面优势明显。Athey and Imbens（2015；2016）将机器学习中常用的分类

回归树引入传统的因果识别框架，用它们来考察异质性因果效应。这一方法的优势还在于，数据驱动的树模型算法可以处理多重异质性的问题，也即是交互或分组变量很可能是多变量的交互等非线性形式。这一点，对异质性分析尤为重要。例如，Seungwoo et al.（2018）就发现地铁开通产生的房屋增值效应，是房屋大小、房间数量、厕所面积等诸多异质性特征搭配产生的结果。他们借助机器学习中的回归树算法，在双重差分法的框架下，讨论了首尔某地铁开通对周边房地产市场的异质性因果效应：在地铁周边住房的 142 个特征组合中，有 89 个特征组合会带来住房价格增值，53 个特征组合会使得房价减值。这篇文章还发现了一些有趣现象，即自地铁开通后，地铁附近新建的公寓基本都按照正向因果效应的特征建造。这一理论与现实的吻合进一步凸显多重异质性因果效应分析的政策和商业价值。

传统方法中还有 Xie and Wu（2005）和 Zhou and Xie（2019）提出的以倾向值为导向的异质性处理效应分析方法，其核心思想是考察处理效应如何随着处理倾向值变化而变化。这一方法的做法是将很多协变量降维为一个个体接受处理变量某个取值水平影响的概率，因此也可以处理高维数据。但是，该方法缺陷也很明显：首先，具体采用哪些变量估计倾向值仍然不确定，这使得倾向值模型可能设定错误（胡安宁，2017；胡安宁等，2021），而机器学习中的一些集成算法在最优模型设定方面表现很好。其次，传统的倾向值导向的方法将变量降维为倾向值，但是让我们损失了很多信息，即我们仍然不知道究竟哪个变量有异质性特征。因此，在不损失异质性信息的前提下，机器学习处理高维数据的方式仍然是最佳的。最重要的是，这种倾向值方法无法让我们评估每个人对政策的反应究竟如何，即个体处理效应。个体处理效应的研究在医疗领域开展较多，尤其是在精确医疗方面。在机器学习领域，Wager and Athey（2018）提出的因果森林方法，可以用来进行个体处理效应的估计。在社会科学中，也有具体应用。例如，Davis and Heller（2017）就将因果森林应用于随机干预试验，研究讨论暑期给一些无业青少年工作培训对其行为的影响。他们发现工作培训后，青少年暴力性犯罪明显减少。他们试图分析是否是因为改善这些人的就业进而使得这些人都变好了。这个问题的回答非常依赖于更细微粒度的异质性处理效应分析，即到底哪些人在项目中获得就业改善，哪些人没有。

（四）更好地检验因果关系的外部有效性

在传统因果推断的社会科学实证研究中，一般都缺乏结论是否能够外推的考察，很少去强调模型的验证问题，似乎默认根据理论指导而得来的便是一个“正确”的实证模型。给定这一假定，研究人员的任务是去估计模型中的参数，而不是验证整个模型对研究问题和情形的适用性。而机器学习方法特别强调模型的泛化能力，即整个模型和参数在更多数据中的预测能力。因此为提升泛化能力，机器学习方法有很多特别的做法，例如训练集和测试集的划分、交叉验证，等等。这些方法也可以应用在因果识别中，以提高传统因果识别的外推能力（Wager and Athey，2018；Fafchamps and Labonne，2017；Chernozhukov et al.，2018）。

其实，实证结论可能缺乏外部有效性这一问题之所以会经常出现是因为其跟目前社

会科学领域的研究和学术发表惯例相关。在经济学、管理学等社会科学领域，学术发表中非常重要的一点就是讨论研究发现的统计显著性：p 值。但根据 Gerber and Malhotra (2008) 和 Brodeur et al. (2016) 的研究，社会科学领域学术论文中，p 值明显存在一个围绕习惯门槛 (0.05) 的异常聚集。上述证据说明在社会科学学术界存在有意识或无意识地选择更好的模型方法、分析样本和控制变量等现象。根据机器学习的启示，我们可以轻松知悉，这一精巧地、反复地敲打给定数据得出的显著的“因果效应”，泛化能力很可能会很差。换言之，在这个给定的数据集上得到的显著回归结果和因果关系结论，在新的数据集上很可能就不存在了。没有外部有效性的因果关系，不能称为真正的因果关系。而使用机器学习实践中的标准做法，可以更好地验证模型的外部有效性。例如，参考机器学习方法，可以将待分析样本分成训练集和测试集两部分，训练集数据用于建模分析，估算因果效应，测试集则用来评估上述因果效应结论的稳健性和泛化能力，这是一个在样本量充足的研究设计中非常值得推荐的方法。Anderson and Magruder (2017) 对于分割样本以避免错误结论做法的背后思想、技术路径等都给予了详细的介绍，我们这里不再详细展开。

三、大数据和机器学习对因果识别的冲击

在大数据时代，机器学习为因果关系的识别提供了很多新的方法和新的场景，拓宽了其适用边界，但同时大数据和机器学习方法的广泛应用，也给因果关系识别带来一些新的冲击，本部分对此进行简要的讨论。

(一) 因果关系在某些情形下变得不再重要

因果关系在我们社会科学研究中，确实具有非常重要的价值，但是，我们也应该承认，在某些情形下，并不是非要识别出因果关系才算是一个有价值的研究，这一点在大数据时代更加凸显。当要研究某些重要的经济社会问题，而又缺乏直接的数据时，有时放弃对因果关系的执着，改为只关注经济社会变量的相关关系，会为我们开辟一个新的“脑洞”。例如，在财富和贫困问题的讨论上，有文献利用手机使用记录数据推断一个人的社会经济地位，并进一步预测整个国家社会财富区域分布 (Blumenstock et al., 2015)，或个人的贷款违约概率 (Björkegren and Grissen, 2020)。这类研究的意义在于对于那些存在资源约束或者缺少普查和调查数据的地方，可以通过类似方法为某些重要的经济社会特征寻找替代性的统计指标：不能认为手机使用习惯“导致”了该个体的信用状况，但又不能否认这种大数据征信的巧妙之处。再比如，战争的危害无需过多强调，但战争正在进行时，对战争实况的了解和针对性的人道主义援助存在很大挑战。Li and Li (2015) 就独辟蹊径，利用夜间灯光数据来预测叙利亚战争状况和后果。

社会科学一直作为“解释性”的学科，往往在实际应用中受到很多限制，根据解释提出的政策建议有时候也是差强人意。相比自然科学，社会科学的成果转化率也较弱，这致使社会科学置于尴尬的境地。但随着统计工具的发展，特别是在计算机技术的帮助下，社会科学实证方法得到了极大的进步。在大数据中通过预测、分类等手段得出的结

论，即便不是因果关系，也往往具有很高的政策和商业价值，能够为研究者和政策实施者提供重要启示。对此，社会科学研究者应该保持开放的心态。

（二）大数据和机器学习让某些情形下因果关系识别更困难

关于为什么大数据和机器学习可能让因果关系识别变得更加困难，我们以文本数据为例来进行说明。文本大数据一个鲜明特征是其为非结构化的高维数据。在因果识别中，不管文本数据是处理变量，还是结果变量，都需要将文本数据通过某种人工和机器学习的方法，映射到一个低维的结构化数据上，例如将一段文本数据映射到它体现的政治态度、情感或主题上。由于这种映射函数并不是唯一的，从而可能会产生识别问题或过拟合问题（Egami et al.，2018）。

具体而言，在这当中之所以会出现识别问题，主要在于映射函数的不稳定上。上述方法实质上相当于从文本大数据中构造出一些指标，然后纳入传统的因果关系识别框架进行分析，但这些构造出的指标，准确率其实并不太高，这样就可能产生测量误差，而且这种测量误差，很多情形下又跟处理变量相关，从而也就成为因果关系识别中一个新的挑战（Wood-Doughty et al.，2018）。而产生过拟合的原因则主要看是由于为了更好地获得处理变量和结果变量之间的关系，在映射函数的选择设计上，可能会穷尽训练集数据中的各种细节和噪音，从而在训练集得到的处理变量和结果变量之间的关系，无法泛化到一般化的情形当中，从而产生虚假的因果关系。

（三）部分机器学习算法缺乏可解释性

如上文所述，机器学习确实可以拓展传统因果识别方法的适用边界，而因果识别是现代社会科学家理解经济社会运行规律的重要手段，因此机器学习对我们社会科学研究者更好地认清社会科学规律本源，具有一定的帮助。但同时，机器学习在拓展因果识别适用边界的同时，确实也带来一些挑战，特别是其存在可解释性差的问题。具体而言，为了提高预测能力，机器学习方法的很多创新，在社会科学研究者看来都是“离经叛道”的，如前文提到过的神经网络，其计算过程类似一个“黑箱”，可解释性很差，这给因果关系识别带来很大干扰（Steinkraus，2018）。学者们仅知道“黑箱”的输入和输出，而“黑箱”内部的运算是如何进行的，往往不得而知，这是很多社会科学研究者对机器学习方法持批评态度的一大原因。例如，在使用 IV 的两阶段回归的一阶段分析中，如果使用神经网络等机器学习算法进行内生变量与工具变量（以及协变量）之间的建模，很可能无法解释清楚工具变量与内生变量之间到底有什么关系。再比如，利用机器学习方法进行因果识别时，一个常见的方法是利用处理组被处理之前的数据和对照组数据等来预测如果没有发生处理政策，处理组应该具有的“反事实结果”，而很多机器学习算法在进行这样的预测中，仍然是一个“黑箱”，无法解释清楚这个反事实结果到底是如何预测出来的，这对于注重理论分析的社会科学家而言是非常尴尬的。总之，机器学习在这一方面仍然存在很多问题，这是机器学习方法在社会科学实证应用，特别是因果关系识别当中进一步拓展的重要瓶颈。

四、结论与展望

随着机器学习方法在社会科学定量分析中的大规模应用，我们仍然有必要重新思考 Angrist and Pischke（2009）提出的“方法是否有必要如此复杂”，以及“它们是否是有害”这两个问题。关于社会科学领域的因果推断，Holland（1986）认为其“根本问题”是个体的“反事实”状态无法同时观察到。因此，传统的社会科学实证分析工具，从控制变量到固定效应回归，从工具变量、双重差分、倾向得分匹配、合成控制、断点回归等时髦的新方法，到随机干预试验等，无一不是向寻找“可比”对象靠近的过程。这些方法都有其科学性，但在某些特殊情形下，又存在一定的局限。当那些近乎苛刻的条件无法满足时，使用这种方法得出的结论可能就与真理更加背道而驰了。而机器学习在获得变量间的非线性关系、控制混淆因素、应对高维数据等方面的优势，以及其在分类、降维、预测等方面的成功，可以帮助我们获得一个更为置信和稳健的结论；可以在一些非结构化、高维的大数据和领域中发掘出一些有价值的新问题；可以在传统方法因假设无法满足而失效时依然有方法可以备选；可以让我们的结论在样本外也有预测能力；可以帮助制定最优的政策设定以实现收益-成本最大；甚至帮助我们从更细微的维度了解每一个人的处理效应，等等。因此，虽然机器学习对社会科学来说是一个全新的领域和全新的分析工具，说其复杂也并不为过，但是如果结合机器学习能够帮助这个学科进一步靠近真理，将是非常值得尝试和进一步探索的工作。

当然，对于其是否有害的问题，我们上文也总结了机器学习给社会科学因果关系识别带来的几个挑战。这也是我们社会科学研究者在利用机器学习和其他大数据方法时，需要保持警惕的地方。倘若我们过分地追求机器学习方法，也可能会与社会科学目标相悖。社会科学的目标是能够回答一个实质性的问题，提升人们对社会状况的理解，最终带来理论上的进步。如果过分关注机器学习的预测能力而忽视社会科学的解释功能，无疑也会本末倒置。未来，我们应该如何处理大数据、机器学习和因果推断的关系，本文认为应该如 Grimmer（2015）倡导的，我们首先是一个社会科学家，其次才是一个数据分析人员，我们只是在利用大数据和机器学习的工具，来帮助我们更好地理解这个社会。

参 考 文 献

[1] Abadie, A., “Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects”, *Journal of Economic Literature*, 2021, 59 (2), 391-425.

[2] Abadie, A., and M. D. Cattaneo, “Econometric Methods for Program Evaluation”, *Annual Review of Economics*, 2018, 10, 465-503.

[3] Abadie, A., A. Diamond, and J. Hainmueller, “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program”, *Journal of the American Statistical Association*, 2010, 105 (490), 493-505.

[4] Abadie, A., and J. Gardeazabal, “The Economic Costs of Conflict: A Case Study of the Basque Country”, *American Economic Review*, 2003, 93 (1), 113-132.

[5] An, W., “Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference”, *Sociological Methodology*, 2010, 40 (1), 151-189.

[6] Anderson, M., and J. Magruder, "Split-Sample Strategies for Avoiding False Discoveries", *National Bureau of Economic Research* Working Paper, 2017.

[7] Angrist, J. D., and J. S. Pischke, *Mostly Harmless Econometrics*. New Jersey: Princeton University Press, 2009.

[8] Athey, S., and G. W. Imbens, "Machine Learning Methods for Estimating Heterogeneous Causal Effects", *Stat*, 2015, 1050 (5), 1-26.

[9] Athey, S., and G. W. Imbens, "Recursive Partitioning for Heterogeneous Causal Effects", *Proceedings of the National Academy of Sciences*, 2016, 113 (27), 7353-7360.

[10] Athey, S., and G. W. Imbens, "The State of Applied Econometrics-Causality and Policy Evaluation", *Journal of Economic Perspectives*, 2017, 31 (2), 3-32.

[11] Athey, S., and G. W. Imbens, "Machine Learning Methods Economists Should Know About", *Annual Review of Economics*, 2019, 11 (1), 685-725.

[12] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen, "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain", *Econometrica*, 2012, 80 (6), 2369-2429.

[13] Belloni, A., V. Chernozhukov, and C. Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls", *The Review of Economic Studies*, 2014, 81 (2), 608-650.

[14] Belloni, A., V. Chernozhukov, and C. Hansen, "Estimation of Treatment Effects with High-Dimensional Controls", *AEA Papers and Proceedings*, 2019.

[15] Ben-Michael, E., A. Feller, and J. Rothstein, "The Augmented Synthetic Control Method", *Journal of the American Statistical Association*, 2021, 116 (536), 1789-1803.

[16] Björkegren, D., and D. Grissen, "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment", *The World Bank Economic Review*, 2020, 34 (3), 618-634.

[17] Bloniarz, A., H. Z. Liu, C. H. Zhang, J. S. Sekhona, and B. Yu, "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments", *Proceedings of the National Academy of Sciences*, 2016, 113 (27), 7383-7390.

[18] Blumenstock, J., G. Cadamuro, and R. On, "Predicting Poverty and Wealth from Mobile Phone Metadata", *Science*, 2015, 350 (6264), 1073-1076.

[19] Bollen, K. A., "Instrumental Variables in Sociology and the Social Sciences", *Annual Review of Sociology*, 2012, 38 (1), 37-72.

[20] Brand, J. E., and J. S. Thomas, "Causal Effect Heterogeneity", In: Morgan, S. L. (ed.), *Handbook of Causal Analysis for Social Research*. Netherlands: Springer, 2013.

[21] Branson, Z., M. Rischard, L. Bornn, and L. Miratrix, "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs", *Journal of Statistical Planning and Inference*, 2019, 202, 14-30.

[22] Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg, "Star Wars: The Empirics Strike Back", *American Economic Journal: Applied Economics*, 2016, 8 (1), 1-32.

[23] 陈云松，"逻辑、想象和诠释：工具变量在社会科学因果推断中的应用"，《社会学研究》，2012 年第 6 期，第 192—216 页。

[24] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased Machine Learning for Treatment and Structural Parameters", *The Econometrics Journal*, 2018, 21 (1), 1-68.

[25] Cicala, S., "Imperfect Markets versus Imperfect Regulation in U. S. Electricity Generation", *American Economic Review*, 2022, 112 (2), 409-41.

[26] Cole, M. A., R. J. R. Elliott, and B. Liu, "The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach", *Environmental and Resource Economics*, 2020, 76 (4), 553-580.

[27] Cunningham, S., *Causal Inference: The Mixtape*. Yale University Press, 2021.

[28] Davis, J. M., and S. B. Heller, "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous

Effects of Summer Jobs", *Review of Economics and Statistics*, 2017, 1-47.

[29] Doudchenko, N., and G. W. Imbens, "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis", *National Bureau of Economic Research* Working Paper, 2016.

[30] Dube, A., J. Jacobs, S. Naidu, and S. Suri, "Monopsony in Online Labor Markets", *American Economic Review: Insights*, 2020, 2 (1), 33-46.

[31] Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart, "How to Make Causal Inferences Using Texts", Working Paper, 2018.

[32] Fafchamps, M., and J. Labonne, "Using Split Samples to Improve Inference on Causal Effects", *Political Analysis*, 2017, 25 (4), 465-482.

[33] 方媚、金刚，"社会学习与消费升级：来自中国电影市场的经验证据"，《中国工业经济》，2020 年第 1 期，第 43—61 页。

[34] Gerber, A., and N. Malhotra, "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals", *Quaterly Journal of Political Science*, 2008, (3), 313-326.

[35] Grimmer, J., "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together", *Political Science & Politics*, 2015, 48 (1), 80-83.

[36] Guo, F., Y. P. Huang, J. Y. Wang, and X. Wang, "The Informal Economy at Times of COVID-19 Pandemic", *China Economic Review*, 2022, 71, 101722.

[37] Guo, R., L. Cheng, J. Li, R. Hahn, and H. Liu, "A Survey of Learning Causality with Data: Problems and Methods", *ACM Computing Surveys* (CSUR), 2020, 53 (4), 1-37.

[38] Guo, J., and Z. Zhang, "Does Renaming Promote Economic Development? New Evidence from a City-renaming Reform Experiment in China", *China Economic Review*, 2019, 57, 101344.

[39] Gilchrist, D. S., and E. G. Sands, "Something to Talk About: Social Spillovers in Movie Consumption", *Journal of Political Economy*, 2016, 124 (5), 339-1382.

[40] Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A Flexible Approach for Counterfactual Prediction", *Proceedings of the 34th International Conference on Machine Learning*, 2017, 70, 1414-1423.

[41] Herlands, W., E. McFowland, A. Wilson, and D. Neil, "Automated Local Regression Discontinuity Design Discovery", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, 1512-1520.

[42] Hill, J., and Y. S. Su, "Assessing Lack of Common Support in Causal Inference Using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive Outcomes", *The Annals of Applied Statistics*, 2013, 7 (3), 1386-1420.

[43] Holland, P. W., "Statistics and Causal Inference", *Journal of the American Statistical Association*, 1986, 81, 945-970.

[44] 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 年第 1 期，第 17—35 页。

[45] 胡安宁，"统计模型的'不确定性'问题与倾向值方法"，《社会》，2017 年第 1 期，第 186—210 页。

[46] 胡安宁、吴晓刚、陈云松，"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》，2021 年第 1 期，第 91—114 页。

[47] 黄乃静、于明哲，"机器学习对经济学研究的影响研究进展"，《经济学动态》，2018 年第 7 期，第 115—129 页。

[48] Imai, K., and M. Ratkovic, "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation", *The Annals of Applied Statistics*, 2013, 7 (1), 443-470.

[49] Imbens, G. W., "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics", *Journal of Economic Literature*, 2020, 58 (4), 1129-1179.

[50] Imbens, G., and S. Wager, "Optimized Regression Discontinuity Designs", *Review of Economics and Statistics*, 2019, 101 (2), 264-278.

[51] Kinn, D., "Synthetic Control Methods and Big Data", Working Paper, 2018.

[52] Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer, "Prediction Policy Problems", *American Economic Review*, 2015, 105 (5), 491-95.

[53] Knaus, M. C, and M. Lechner, and A. Strittmatter, "Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach", *Journal of Human Resources*, 2022, 57 (2), 597-636.

[54] Knittel, C. R., and S. Stolper, "Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use", *AEA Papers and Proceedings*, 2021, 111, 440-44.

[55] Kumar, A., and C. Liang, "Credit Constraints and GDP Growth: Evidence from a Natural Experiment", *Economics Letters*, 2019, 181, 190-194.

[56] Li, D., and X. Li, "Applications of Night-Time Light Remote Sensing in Evaluating of Socioeconomic Development", *Journal of Macro-quality Research*, 2015, (3), 1-8.

[57] Linden, A., and P. R. Yarnold, "Using Machine Learning to Assess Covariate Balance in Matching Studies", *Journal of Evaluation in Clinical Practice*, 2016, 22 (6), 844-850.

[58] 刘甲炎、范子英，"中国房产税试点的效果评估：基于合成控制法的研究"，《世界经济》，2013 年第 11 期，第 117—135 页。

[59] Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos, "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality", *Political Analysis*, 2020, 28 (4), 445-468.

[60] Mullainathan, S., and J. Spiess, "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 2017, 31 (2), 87-106.

[61] Narayanan, S., and K. Kalyanam, "Behavioral Targeting, Machine Learning and Regression Discontinuity Designs", Working Paper, 2020.

[62] Pearl, J., and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York : Basic Books, 2018.

[63] Qiu, Y., X. Chen, and W. Shi, "Impacts of Social and Economic Factors on the Transmission of Coronavirus Disease 2019 (COVID-19) in China", *Journal of Population Economics*, 2020, 33, 1127-1172.

[64] Robertsy, M., B. Stewartz, and R. Nielsen, "Adjusting for Confounding with Text Matching", *American Journal of Political Science*, 2020, 64 (4), 887-903.

[65] Seungwoo, C., M. E. Kahn, and M. H. Roger, "Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach", *Real Estate Economics*, 2018, 48 (3), 1-29.

[66] Steinkraus, A., "Rethinking Policy Evaluation-Do Simple Neural Nets Bear Comparison with Synthetic Control Method?", Working Paper, 2018.

[67] Varian, H. R., "Causal Inference in Economics and Marketing", *Proceedings of the National Academy of Sciences*, 2016, 113 (27), 7310-7315.

[68] Wager, S., and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests", *Journal of the American Statistical Association*, 2018, 113, 1228-1242.

[69] 王芳、王宣艺、陈硕，"经济学研究中的机器学习：回顾与展望"，《数量经济技术经济研究》，2020 年第 4 期，第 146—164 页。

[70] Wood-Doughty, Z., I. Shpitser, and M. Dredze, "Challenges of Using Text Classifiers for Causal Inference", *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, 2018, 4586-4598.

[71] Xie, Y., and X. Wu, "Reply to Jann: Market Premium, Social Process, and Statisticism", *American Sociological Review*, 2005, 70 (5), 865-870.

[72] Yang, J. C., H. C. Chuang, and C. M. Kuan, "Double Machine Learning with Gradient Boosting and Its Application to the Big N Audit Quality Effect", *Journal of Econometrics*, 2020, 216 (1), 268-283.

[73] Zhou, X., and Y. Xie., "Marginal Treatment Effects from a Propensity Score Perspective", *Journal of Political Economy*, 2019, 127 (6), 3070-3084.

Machine Learning and Causal Relationship in Social Science: A Literature Review

GUO Feng
(Shanghai University of Finance and Economics; Peking University)
TAO Xuhui*
(Shanghai University of Finance and Economics)

Abstract: Causal identification is the core of empirical research in social sciences. In the era of big data, machine learning brings some new opportunities and challenges to causal identification. This research focuses on the value of machine learning in causal inference: identification and control of confounding factors, better designation of control group, better identification of heterogeneous treatment effects, and more assurance of the external validity. Several challenges of causal inference under application of big data and machine learning are also discussed: circumstances when causality is no longer important, possibilities that big data and machine learning makes it more difficult to identify causal effects, and some results are not interpretable. This article can help researchers expand their toolboxes and think tanks.

Keywords: machine learning; causality; big data
JEL Classification: B41, C55, C80

* Corresponding Author: Tao Xuhui, School of Public Economics and Administration, Shanghai University of Finance and Economics, Shanghai 200433, China; Tel: 86-13184578885; E-mail: taoxuhui@163.sufe.edu.cn.