# 机器学习在社会科学中的应用：回顾及展望

**王芳 陈硕 王宣艺**

华东师范大学 复旦大学

**摘要：**随着数据的可得和计算机软硬件的发展，机器学习技术在业界及自然科学领域已经得到了广泛地应用。在社会科学领域，该技术使用虽然起步较晚，但发展也非常迅速。本文目的旨在系统介绍机器学习在社会科学中的应用。在简单介绍定义，在业界及自然科学领域的应用后，我们将从数据生成、预测以及因果识别 (DID，RD 和 IV) 三方面详细介绍机器学习在社会科学中的应用。局限于社科因果识别方法论的成熟及样本量限制，我们认为机器学习虽然拓展了社会科学研究的边界，但并不会颠覆现有研究范式。最后，本文从学界不平等及可复制性等方面讨论了该技术在应用过程中可能带来的问题。

**关键词：**机器学习、数据生成、预测、因果识别

**JEL：**C19, C55, C80

# 一、前言

机器学习 (Machine Learning，简称 ML) 指的是从数据中识别出规律并以此完成预测、分类及聚类等任务的算法总称。[1]随着数据的可得及计算机处理能力的提高，该技术在业界及自然科学领域已经得到广泛地应用。在社会科学领域，机器学习的使用虽然起步较晚，但发展也非常迅速。例如，五大经济学英文顶尖期刊中涉及到机器学习技术的文章数量在 2014 年之后以每年 74.7%的速度递增，2017 年的数量达到 16 篇。中文经济学权威期刊也有类似的趋势 (见附录图 1)。[2]本文写作目的旨在系统介绍机器学习技术在社会科学中的应用。我们首先在第二部分给出机器学习技术的定义，然后在第三和第四部分简要介绍该技术在业界及自然科学领域的应用。本文的重点在第五部分，在其中我们将从数据生成、预测以及因果识别三方面介绍机器学习在社会科学中的应用。就数据生成来说，机器学习技术可以帮助学者获得以前很难或无法获得的数据，进而对一些更具挑战性的假设进行检验；就预测来说，机器学习可以更有效地探索变量间的相关性，进而做出较为精准的预测。在这部分，我们将用公式表达的方式详细比较机器学习技术和传统基于回归方法在预测方面的异同；最后，由于机器学习在预测方面的优势，它可以被用来预测反事实进而获得因果效应。我们认为机器学习技术在上述方面的优势使其可以和社会科学现有分析工具结合，检验之前无法用传统方法检验的假设，最终会拓展现有社会科学研究的边界。同时，我们也应该对其带来的问题保持清醒认识，这些问题包括研究可复制性、过分依赖大数据及可能加剧学界不平等。本文最后一部分将对这些问题展开初步讨论。

# 二、机器学习简介

机器学习是指从数据中识别出规律并以此完成预测、分类及聚类的算法总称 (Mitchell,

[1] 和该概念相关并经常被同时提起的另外一个概念是人工智能 (Artificial Intelligence, AI)。严格意义上，机器学习应归属于人工智能研究范畴内。人工智能还包括诸如机械伦理学、自然语言处理和计算机图像识别等领域，机器学习更像是实现人工智能的手段和算法基础。为了和现有文献保持一致，我们对这两个概念不做明确区分，都称为广义的“机器学习”(Athey, 2018; Camerer, 2018)。

[2] 五大期刊指的是 *American Economic Review*、*Econometrica*、*Journal of Political Economy*、*Quarterly Journal of Economics* 和 *Review of Economic Studies*。经济学中文权威期刊是指《管理世界》、《金融研究》、《经济学（季刊）》、《经济研究》和《世界经济》。此外还有另外两个更能反映未来研究趋势的指标：NBER 工作论文在 2016–17 年的每个季度平均有 8 篇新文章是关于机器学习的；在美国前 20 名高校经济系讲座中，16–17 年每个季度平均有 2-3 个讲座涉及到机器学习。这两个指标反映出来的趋势也充分证实机器学习近年来在社会科学领域发展迅速。

1997; Athey, 2018)。[3]在操作中，机器学习方法可以根据被解释变量是否已知被分成监督 (Supervised Learning) 和非监督学习 (Unsupervised Learning)。监督学习是指被解释变量已知的机器学习。我们用$x$表示解释变量，$y$表示被解释变量，那么监督学习首先依据已有样本数据建立$y = f(x)$的函数关系。随后，当要预测解释变量为$x'$时被解释变量的取值时，只需将$x'$带入到等式右边即可获得预测值；非监督学习是指被解释变量未知的学习方式。换句话说，算法只知道解释变量$x$的取值。在这种学习方式下，算法会分析$x$的内部结构，然后根据相似性把数据聚类 (Cluster)。[4]以垃圾邮件分类为例，在监督学习下，我们告诉计算机样本中哪些是垃圾邮件哪些是正常邮件。基于该信息，算法会找出邮件内容与类别间的规律并完成对未来邮件的分类。在非监督学习下，算法并不知道某一封邮件是垃圾邮件还是正常邮件。[5]此时，算法依照邮件文本结构与相似度等指标把邮件归类。在整个过程中，算法只完成归类的工作，定义哪一类属于垃圾邮件依然依赖于人工。

在实际研究中，学者一般根据被解释变量是否可得来选择使用监督还是非监督学习。比方说我们想回答某项政策能否提高民众幸福感。如果研究者能定期走访调查并建立覆盖政策前后时期的幸福感指数的话，作为被解释变量的幸福感就是已知的。这种情况下就可以选择监督学习。当然在该数据不可得时，学者也可以从互联网上搜集网民留言并分析其幸福程度。采用人力对浩如烟海的留言进行打分显然不切实际，此时可以借助非监督学习技术：让机器自动将留言分成幸福和不幸福两类以供后续研究。[6]

## 三、机器学习在业界的应用

个体、企业及社会组织的生活及运转都会产生海量数据。针对这些数据有目的搜集及高性能计算机技术的进步使得机器学习在业界的应用成为可能 (MGI, 2016)。机器学习技术相关产业在最近几年得到快速发展。全世界人工智能领域的风投已经从 2012 年的不

[3] 计算机科学家 Mitchell (1997) 在其著作 *Machine Learning* 中对机器学习下了一个被广泛引用的定义：如果一个算法在某项任务上的表现随着经验的积累而提升，那么就称该算法为机器学习。

[4] 不同算法度量“相似性”的方法都不同。以最简单的 K-means 聚类算法为例，该方法定义的“相似”指每个类别某个解释变量 x 距离该类别样本均值比较接近。

[5] 这在直观上很容易理解。比方说一个没有生物知识的小孩 (算法)，不知道什么是猫、什么是狗 (不知道哪些邮件是垃圾邮件)。然而这个小孩仍然能通过这个动物的体型和尾巴等特征 (邮件文本的结构和相似度等信息)将猫和狗区分开来 (对邮件进行聚类)。

[6] 关于幸福感的例子见 Hills *et al*. (2016) (本文 5.1 部分)。

到 6 亿美元提高到 2016 年的 50 多亿美元。[7]本文对机器学习在业界应用的划分根据 McKinsey & Company (2017) 的行业报告，分为认知 (Cognition)、预测 (Prediction)、决策 (Decision) 和集成解决方案 (Integrated Solution) 四大类。

认知是指利用计算机收集并解释文字、图像及语音等数据信息。很多公司已经利用机器学习技术对诸如互联网信息等海量数据进行数据挖掘及分析。比如，成立于 2007 年的 Cirmson Hexagon，其主营业务就是利用人工智能技术抓取各大社交平台上顾客对客户公司产品评价及照片，然后对其进行分析进而帮助客户提高产品吸引力、识别目标受众及找出潜在竞争对手。[8]除了文本和图片，也有公司利用机器学习进行语音信息认知。科大讯飞就是该领域非常成功的公司之一。该公司目前已经能够将语音识别整合到手机端输入法中，在语音识别的同时将其转化成文本进而提高用户输入体验。[9]

对各类信息所进行的认知分析为预测提供了基础，预测是目前机器学期在业界最为广泛的应用。精准的预测能够让企业更加了解顾客偏好，改进产品并精准投放广告。这方面知名的公司包括亚马逊和淘宝。基于顾客浏览、搜索及购买等历史数据分析，机器学习预测能够让这些公司在客户端进行精准的商品推荐和广告投放。类似技术也应用到视频提供商上，比如 Netflix 就根据用户历史观影内容来预测偏好，推荐其支付意愿最高的电影 (MGI, 2016)。

这些预测结果进一步为决策提供可能。以金融业 Wealthfront 公司为例，该公司首先通过人工智能技术预测出各种投资的潜在风险及回报率，随后再结合客户资金量和风险偏好为其提供最佳投资决策参考。[10]另一个非常知名的业界决策案例是谷歌公司的 Alpha Go。该方案利用深度神经网络 (Deep Neural Networks) 和树搜索 (Advanced Tree Search) 技术预测并决策出下一步棋的落子(Silver *et al*., 2016)。

---

[7] "The 2016 AI Recap: Startups See Record High in Deals and Funding," CBINSIGHTS, January 17, 2017. https://www.cbinsights.com/research/artificial-intelligence-startup-funding/.

[8] Crimson Hexagon 运用人工智能的介绍见 https://www.crimsonhexagon.com/ai-powered-consumer-insights/。

[9] 讯飞输入法的介绍见 http://www.iflytek.com/about/index.html，该应用来自于 McKinsey & Company (2017)。

[10] 关于 Wealthfront Inc 在该领域的详细介绍见 https://www.wealthfront.com/expertise，该例子来自 McKinsey & Company (2017)。

最后，将上述各个方面应用整合并结合硬件以实现特定目标就是所谓的集成解决方案。在该领域领先的企业包括谷歌和亚马逊。谷歌旗下的无人驾驶汽车公司 Waymo 通过安装在车辆上的传感器及车载电脑对路况、信号灯、路牌及行人等信息进行认知，基于这些信息预测出下一时间道路状况并作出加速或刹车等相关决策。[11]亚马逊的集成解决方案主要在智能家居领域，将人工智能语音助手 Alexa 集成在音箱中，实现语音操控各种家具并安排日程等诸多高阶功能。[12]

## 四、机器学习在自然科学中的应用

和社会科学相比，自然科学的数据可得限制相对较少，因此较早开展相关技术的应用。本部分以物理学和医学为例，简单介绍机器学习在这两个学科中的应用现状。

### 4.1 物理学

机器学习在物理学的运用发展较快，这主要得益于机器学习在数据的分类及降维上具有的优势。下面内容将从高能物理和凝聚态物理两个领域简要阐述这些优势。

在高能物理中，粒子对撞是发现新粒子的有效方法之一。但分析对撞的实验数据涉及到非常复杂的粒子分类问题 (Signal-versus-background)：需要判断传感器接收到的信号是待发现新粒子产生的 (Signal) 还是已知粒子所产生的 (Background)。因此我们需要计算出那些仅由已知粒子所产生的信号是怎样的，然后再与实验测得信号进行对比。如果两者存在显著统计差异就能够判定是出现了未知粒子。物理学家通常采用蒙特卡洛 (Monte Carlo) 方法计算上述信号，而该方法会产生高维数据：解释变量数目很多，甚至超过样本容量 (Baldi *et al.*, 2014)。其原因在于计算信号用到的解释变量是碰撞前粒子的信息，其中包括粒子间相互作用。目前最大的大型强子对撞机 (Large Hadron Collider, LHC) 一次碰撞会用到 $10^{11}$ 个质子，因此考虑粒子间相互作用势必导致数据维度非常庞大。而机器学习有一套成熟的变量选择方法 (Feature Selection) 对高维数据进行降维。

[11] 谷歌汽车 Waymo 的介绍来自于 https://waymo.com/tech/。
[12] 亚马逊音箱及人工智能助手 Alexa 的详细信息可见其官网 https://developer.amazon.com/alexa，该应用摘自 MGI (2016) 。

算法一般从众多解释变量中挑选一部分进行运算并评估其计算结果。如果结果精度不高，那么就再挑选另一部分解释变量并重复上述操作，直至算法表现达到期望精度为止。[13] 借助此类降维算法，Baldi *et al*. (2015) 把机器学习运用在希格斯粒子发现的实验中，在给定数据量下提高了估计值的置信水平。

机器学习在物理上的另一大应用是研究凝聚态物理，这个领域旨在分析物质的微观组成和宏观性质间的联系。研究物质的微观组成和结构必然会涉及到分析微观粒子间的相互作用。但问题在于物质是由大量微粒组成的，任意两个粒子间的作用都需要被纳入考量。这意味着方程规模将特别大，就会遇到类似于经济学中一般均衡的求解问题：要考虑任意一个消费者对任意一件商品的需求，也要考虑任一厂商对任一商品的供给。由于方程组数目巨大，传统方法难以获得解析解。经济学家通常进行数值求解，物理学家解决该问题也采用类似方法。然而在利用计算机求解过程中也会产生高维数据，原因正是上文所提的方程规模过大。如果采用经典方法计算，结果很可能不收敛以致于无法获得数值解 (Kohn, 1999)。对此，机器学习同样通过降维方法获得满意数值解。基于这些优点，近期已有很多物理学家引入诸如神经网络 (Neural Network) 技术来进行凝聚态物理研究，成功地从高维数据中找出了影响物质宏观性质的序参量 (Order Parameter) (Carrasquilla and Melko, 2017; Wang, 2016; van Nieuwenburg *et al*., 2017)。

### 4.2 医学

机器学习在医学中的运用主要集中在生物医学数据及图像的分析及识别上。下文将以脑神经电信号和肿瘤基因信息分类及肠道断肠扫描图像识别这两个案例，对其应用进行简单归纳。

人脑皮层约有 $10^{10}$ 个神经元且活跃程度也彼此不同，其释放的神经电信号是较为典型的大数据。对这些电信号的解读能够让医生了解脑部活动与肢体运动间的联系，该联系对瘫痪患者康复极为重要：当患者大脑无法指挥肢体时，如果能够读懂患者神经信号对应的肢体动作，医生就可以对这个动作涉及到的肌肉进行电刺激使其完成动作。问题在于

[13] 关于 Feature Selection 与数据降维的综述可以参考 Blum and Langley (1997); Dash and Liu (1997) 及 Kohavi and John (1997)。

神经信号太过复杂，过去的统计技术无法理清如此大量的数据与肢体动作间的联系。Bouton *et al.* (2016) 的研究利用了机器学习技术试图建立了这一联系，该成果成功地让一名瘫痪患者恢复手臂运动功能。这名患者由于脊椎受伤，神经信号无法传递到手臂上。于是研究者利用电脑取代脊柱将脑神经和手臂连在一起，将神经信号“传递”到手上。为此，研究者在患者脑部植入电极以读取神经信号，并要求患者在脑中想象规定的手臂动作。随后研究者利用支持向量机 (Support Vector Machine, SVM) 方法建立神经信号数据和假象动作之间的关联。此后若是电脑读取到某个脑皮层信号，就可以根据已经建立起来的关系找到对应的动作，然后再所需肌肉进行刺激以完成对应动作。

另一类医学大数据是肿瘤基因信息。如果医生可以根据基因突变的发生位置将肿瘤划分为不同类别以施加不同方案和强度的治疗，所获疗效就更加显著。但人类的 2.5 万个基因由 30 亿对碱基对构成，找到病变发生具体位置涉及到大数据的分析和解读。科学家们已经尝试用机器学习技术对肿瘤基因信息进行细分分类，从而对症下药提高疗效 (相关研究见 Alizadeh *et al*., 2000; Shipp *et al*., 2002; Ye *et al*., 2003)。

除了医学数据的处理，机器学习的另一重要运用是识别 X 光片、造影及断层扫描图等医学图像。传统上，医生根据经验来判断图像中是否存在病变。这种方法的准确性受限于医生的经验。这就为机器学习的使用提供了机会。Halligan *et al*. (2006) 比较了人工智能和医生在诊断肠道息肉的正确率差别。作者发现机器学习技术在识别肠道 CT 图像的速度及准确率都强于人类。类似应用还包括让计算机辅助人类识别乳房相关病变 (Gilbert *et al*., 2008; Lehman *et al*., 2015; Obermeyer and Emanuel, 2016)。

## 五、机器学习在社会科学中的应用

本文把目前机器学习技术在社会科学研究中的应用分成三类：第一，数据生成 (Data Generating Process)：机器学习可以帮助学者获得以前很难或无法获得的数据；第二，预测 (Prediction)：机器学习可以更有效地探索变量之间的相关性，进而做出较为精准的预测；第三，因果识别 (Causal Inference)：社会科学、特别是经济学实证研究的核心是因果识别，而机器学习在这方面也具有一定优势。

值得注意的是，本文与 Athey (2018)的综述性文章并不相同，主要体现在以下两点：第一，务实性。Athey 将机器学习在社会科学中的影响分为政策评估和因果推论两部分，这对应于本文第二和第三类应用。但我们认为最普遍也是最重要的应用是数据生成。据我们统计，目前社会科学中关于机器学习技术应用超过 90%都是利用该技术的海量数据处理能力生成新数据或者变量。但该应用可能被 Athey 认为过于基础而没有在她的文章中提及。Athey 详细讨论的是机器学习在因果推论中的最新进展。通过综述，我们发现这方面的应用在当前研究中极其有限。本文侧重介绍其方法论基础，特别是机器学习如何与因果识别的传统方法比如 DID、RD 及 IV 的结合。因此，本文针对机器学习应用的分类更加贴近实际；第二，公式推导及受众。在预测和因果推论部分，我们使用详细的公式推导方法比较了传统线性拟合和机器学习预测间的差异，最终将差异直观地展示出来以利于一般读者理解，Athey (2018) 没有采用类似方式。我们认为本文的潜在读者群更加广泛，适用于不具备机器学习专业知识的社会科学研究者。

### 5.1 数据生成

传统社会科学实证研究基于的数据大都来自官方、问卷调查、实地调查、田野或实验室实验。最新一些研究试图利用机器学习技术拓展数据可得性。通过机器学习获得数据的主要方式是文本挖据及图像识别。

就文本信息来说，研究者关心的是文本主题。为了在海量文本数据中提取主题，学者一般使用 Latent Dirichlet Allocation (LDA) 方法。[14]例如，Hansen *et al*. (2018) 就利用该方法探究透明度政策如何影响政府内决策过程。这篇论文的研究背景是美国联邦公开市场委员会 (Federal Open Market Committee, FOMC) 在 1993 年通过决议公开了内部会议的发言记录。作者将该项政策视作自然实验以观察委员会成员的发言内容在该年前后的变化。研究基于的文本信息包含 5 万多次发言，总计 500 多万个单词，人工检索几乎不可能。作者便利用上文提到的 LDA 模型，从这些海量文本中提取 40 个不同的主题 (图 1)。任意一个成员的每一条发言都可以对应到这些主题中的一个或几个

[14] LDA 可以通过不同单词在一段文本中出现的位置和频率，推测出这段文本中含有几个主题 (Blei *et al.* 2003)。

上。每个成员发言中各个主题的占比及成员间发言的相似度等指标就可以被计算出来，作者便可以使用常规 OLS 检验透明度政策对这些被解释变量的影响了。[15]

**图 1：机器学习得到的 40 个主题及每个主题下最可能出现的单词**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Topic0 | product | increas | wage | price | cost | labor | rise | acceler | inflat | pressur | trend | compens |
| Topic1 | growth | slow | economi | continu | expans | strong | trend | inflat | will | recent | slowdown | moder |
| Topic2 | inflat | expect | core | measur | higher | path | slack | gradual | continu | remain | view | suggest |
| Topic3 | percent | year | quarter | growth | month | rate | last | next | state | averag | california | employ |
| Topic4 | number | data | look | chang | measur | use | point | show | revis | estim | gdp | actual |
| Topic5 | polici | inflat | monetarpol | need | time | can | monetari | move | tighten | view | action | believ |
| Topic6 | rate | term | expect | real | lower | increas | rise | level | declin | short | nomin | year |
| Topic7 | statement | word | chang | meet | languag | discuss | issu | want | read | sentenc | view | use |
| Topic8 | chairman | support | mr | direct | recommend | agre | asymmetr | prefer | symmetr | move | toward | favor |
| Topic9 | employ | continu | growth | job | nation | region | seem | state | manufactur | greenbook | busi | bit |
| Topic10 | dollar | unitedstates | export | countri | import | foreign | japan | growth | abroad | trade | develop | currenc |
| Topic11 | model | use | simul | shock | effect | scenario | nairu | differ | rule | chang | baselin | altern |
| Topic12 | risk | may | balanc | seem | side | uncertainti | possibl | economi | probabl | reason | upsid | much |
| Topic13 | forecast | greenbook | staff | project | differ | assumpt | littl | assum | somewhat | lower | end | period |
| Topic14 | period | committe | consist | econom | run | maintain | futur | read | slightli | stabil | expect | develop |
| Topic15 | invest | incom | spend | capit | household | consum | busi | hous | consumpt | sector | stock | stockmarket |
| Topic16 | month | report | increas | survey | expect | indic | remain | continu | last | recent | data | activ |
| Topic17 | project | forecast | year | quarter | expect | will | percent | revis | anticip | growth | next | recent |
| Topic18 | question | ask | issu | let | want | answer | rais | discuss | don | start | without | okay |
| Topic19 | peopl | talk | lot | much | comment | around | differ | number | realli | look | thing | hear |
| Topic20 | presid | ye | governor | parri | stern | vice | hoenig | minehan | kelley | jordan | moskow | mcteer |
| Topic21 | move | can | evid | signific | stage | inde | will | issu | economi | may | quit | clearli |
| Topic22 | chairman | thank | mr | time | meet | laughter | comment | let | will | point | call | may |
| Topic23 | year | panel | line | shown | right | chart | expect | project | percent | middl | left | next |
| Topic24 | district | nation | area | continu | sector | construct | manufactur | report | activ | region | economi | remain |
| Topic25 | know | someth | happen | right | thing | want | look | sure | can | realli | anyth | els |
| Topic26 | polici | might | committe | market | may | tighten | eas | risk | action | staff | possibl | potenti |
| Topic27 | year | continu | product | price | level | industri | will | sale | increas | auto | last | district |
| Topic28 | inventori | product | sale | level | order | will | sector | come | good | quarter | much | adjust |
| Topic29 | price | oil | increas | energi | effect | import | suppli | product | demand | will | market | oilprices |
| Topic30 | term | might | point | can | sens | run | short | probabl | time | longer | tri | someth |
| Topic31 | seem | may | time | certainli | bit | littl | quit | much | far | perhap | better | might |
| Topic32 | money | aggreg | borrow | seem | rang | reserv | rate | target | time | altern | suggest | million |
| Topic33 | move | market | point | will | fundsrate | rate | basispoints | need | fed | today | basi | time |
| Topic34 | report | busi | compani | year | contact | firm | sale | worker | expect | plan | director | industri |
| Topic35 | will | fiscal | ta | budget | cut | govern | effect | billion | state | spend | deficit | year |
| Topic36 | will | economi | world | rather | problem | believ | can | situat | much | seem | view | good |
| Topic37 | realli | look | side | thing | lot | problem | concern | littl | pretti | situat | kind | much |
| Topic38 | bank | credit | market | loan | financi | debt | lend | fund | concern | financ | problem | spread |
| Topic39 | economi | weak | recoveri | recess | confid | eas | neg | econom | will | turn | declin | period |

注：该图来自于 Hansen *et al*. (2018)，图中每一行代表 LDA 方法识别出的一个主题 (topic0 至 topic39)。每一行中展示出的 12 个单词为该主题下最可能出现的单词，自左往右颜色变浅表示该单词出现可能性越小。

类似使用机器学习技术从文本中生成变量研究还有很多。比如 Antweiler and Frank (2003) 利用朴素贝叶斯算法 (Naive Bayes) 将网络上超过 150 万股民留言分为看涨、看跌及中立三类，然后用每条留言的类别解释股票市场振幅；King *et al*. (2017) 和 Qin *et al*. (2017) 分别采用自动非参数文本分析 (Automated Nonparametric Content Analysis) 和支持向量机 (Support Vector Machine) 技术来识别微博用户或账号的身份；[16]

[15] 该文发现透明度提升后，成员更倾向于探讨经济问题并且在发言中更多地使用量化指标。造成这一变化的原因可能是官员在透明环境下更有动力探讨民众关注的经济问题并且花费更多时间去寻找量化证据来支持自己的发言。

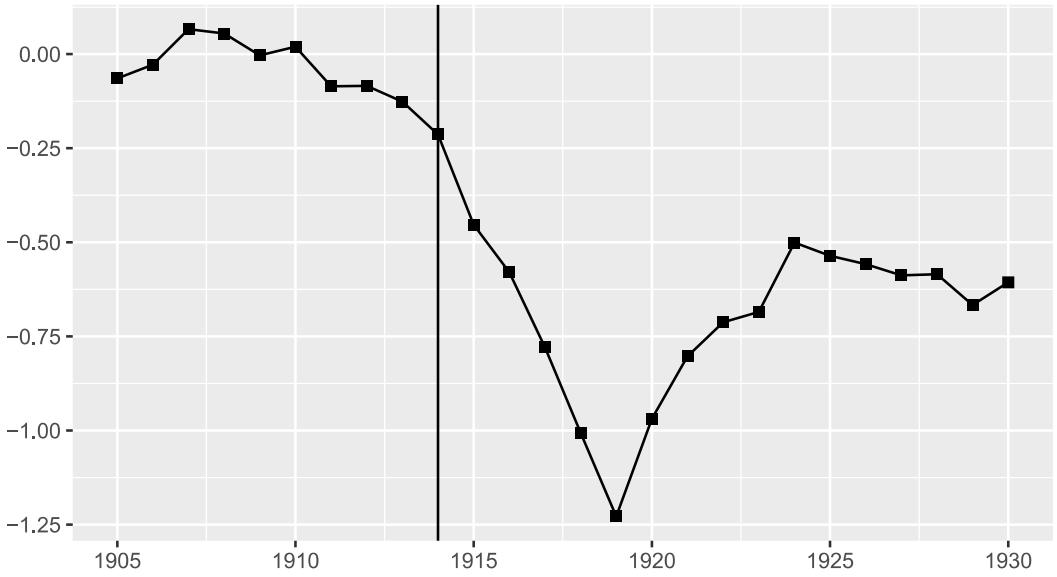
[16]自动非参数文本分析技术是 King *et al.* (2017) 一文的作者为社会科学研究中的文本分类专门开发的，读者可以进一步阅读 Hopkins and King (2010) 解其算法逻辑与实现方法。另一些研究则尝试结合人力与自然语言处理技术结合以进行“半自动化”的文本分析。这些研究包括 Benoit *et al*. (2016) 及 Carlson and Montgomery (2017)

除了文本，机器学习也可以从图像中提取变量。卫星图像就是一个被经济学家广泛研究的图像信息。[17]例如，Engstrom *et al*. (2017) 的研究试图测量一个地区的综合社会福利水平。在发达国家，研究者可以直接依赖官方数据或者调查数据。但很多落后国家由于没有足够财政维持经济普查机构的运转，其官方经济统计数据并不可得。为此，作者使用卷积神经网络 (Convolutional Neural Networks, CNN) 来识别卫星图片中建筑物、车辆及道路等固定资产，以此评估这些地区的福利水平。[18]除卫星遥感照片外，谷歌街景照片 (Google Street View) 也经常被学者用来研究诸如城市化相关习题 (Naik *et al*., 2017)。另外一个被广泛研究的图像信息是人像。比如，Edelman *et al*. (2017) 通过用机器学习技术判别 Airbnb 上的用户头像性别进而分析租房平台上是否存在性别歧视。Cao and Chen (2018) 在研究恋爱配比市场中颜值和物质条件发挥作用时，使用机器学习技术对研究对象的面貌进行打分并和人工打分比较。

上述研究主要涉及变量的“绝对”值，机器学习还可以为研究者生成“相对”意义上的变量。比较不同文本相似度是该领域的典型应用。比如，Iaria *et al*. (2018) 试图研究一战冲击是否会影响跨国学术交流合作。在这个研究中，解释变量是战争的爆发，被解释变量则是论文的相似程度。作者预期战争会降低论文相似程度，基于如下逻辑：如果两个国家的学者经常交流，那么大家的研究兴趣和方向就会比较相似，这会导致论文成果也具有相似性。战争的爆发使得国家之间进入敌对状态，跨国学术合作被迫中断。这将导致同盟国和协约国各自的论文标题相似度下降。该研究需要解决的关键问题是比较论文间的相似度：样本包含 40000 篇论文。作者采用基于机器学习的语义分析 (Latent Semantic Analysis) 来比较两两论文标题间的相似程度，实现了人工不大可能完成的工作 (图 2)。折线代表敌对阵营间论文标题的相似度。可以非常明显的看到，相似度在一战爆发后显著下降，证实了作者上述猜想。

**图 2：来自不同阵营的论文相似度**

[17] 这方面的工作和 Henderson *et al*. (2012) 的研究不同，该研究是将夜间灯光的明暗转变为数据。而我们介绍的则是从高分辨率的卫星图片中识别出房屋、汽车等和地区财富相关指标。
[18] 类似通过卫星照片来估计地区财富水平的研究还有 Jean *et al*. (2016)。



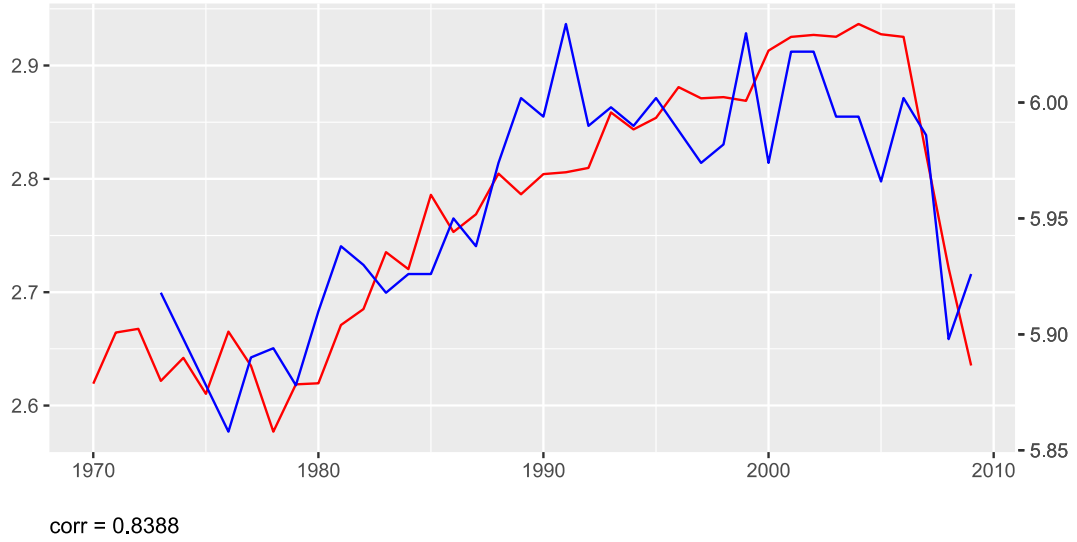

注：本图来自于 Iaria *et al*. (2018)。图中纵轴表示两个敌对国之间的论文相似度，横轴表示论文出版年份，黑色竖线为 1914 年一战爆发。同上文猜想一样，敌对国之间的论文相似度在一战爆发后极速下降。图中相似度为负是因为作者选取的对照组为本国与本国内部论文的相似度，详细讨论见原文。

其他利用文本相似度进行研究的文献包括 Bleakley and Ferrie (2016) 和 Hoberg and Phillips (2016) 等。Bleakley and Ferrie (2016) 试图研究财富增加能否增加对后代的教育投资。由于普查数据来自多个年份，造成了部分父辈与子辈无法匹配 (比如女子婚后改变姓氏)。作者使用机器学习技术并结合其他个人信息预测来自两个样本的不同个体是不是父子或父女关系以解决该问题。Hoberg and Phillips (2016) 则研究了 911 事件如何影响军火企业。由于美国传统行业划分是不随时变化的，这就导致了那些由于这一事件进入或退出军火行业的企业无法被识别出来。为解决该问题，作者同样采用机器学习技术：分析公司每年的产品描述文档并根据其相似度划分行业分类。结果发现 911 事件后进入军火行业的企业数目显著增多。

除了对海量文本进行归类和比较外，机器学习技术还可以测量文字背后的情感。比如，Hills *et al*. (2016) 试图研究历史上人们的主观幸福指数。我们可以依靠社会调查数据测量现代社会公众的幸福感。但该方法并不适用于古代。作者采用的策略是利用机器学习计算不同时期出版图书中的幸福感。研究数据来自谷歌图书 (Google Books Corpora)，该数据库收录了 1500 年以来将近 1000 万本书籍。作者首先利用语言学和心理学文献中已有的“幸福感词典”，定义出每一个词所代表的“幸福值”，然后用计算机计算出每一

本书的幸福感指数。为了验证该方法可靠性，然后，作者比较了 1970 年后分别利用该方法 (红线) 和 Eurobarometer 社会调查 (蓝线) 所建立起来的意大利公众幸福指数 (图 3)。这两个指数之间类似的发展趋势说明了上述方法的可行性。

**图 3：作者构建的意大利主观幸福指数与 Eurobarometer 的意大利民调**




注：本图来自于 Hills *et al.* (2016)。图中红色线为作者通过机器学习读取图书生成的意大利主观幸福指数，数值大小以左侧纵轴来表示；蓝色线为 Eurobarometer 在意大利进行的生活幸福感调查获得的幸福指数，可以看到两者吻合度很高。

另外一个被学者，特别是政治学者广泛关注的是文本中体现的政治立场。[19]利用机器学习技术对文本进行立场分析的相关研究包括：从各个党派党章或宣言中推测党派政策立场 (Laver *et al.*, 2003)、通过新闻报纸措辞来判断报纸党派倾向 (Gentzkow and Shapiro, 2010)、通过国会发言来测定党派分歧 (Gentzkow *et al.*, 2017) 等。[20]

## 5.2 预测

在使用机器学习之前，社会科学研究者主要依赖最小二乘回归 (OLS) 进行预测。在本小节，我们首先用公式推导的方法比较该领域被广泛使用的 Ridge (岭回归) 技术和 OLS

[19] 当然，文字能反映的感情显然不局限于幸福感和政治倾向，研究者可以根据研究所需生成想要的情感倾向变量。比如 Wu (2017) 就利用机器学习来甄别出哪些词汇被经常用来刻画男性形象，哪些词汇更可能描述女性。

[20] 除了情感和政治立场分析之外，也有一些研究利用机器学习分析主流媒体的报道来预测军事冲突。Muelle and Rauh (2018) 首先利用和正文中提到的 Hansen *et al.* (2018) 类似的 LDA 方法提取《纽约时报》、《华盛顿邮报》和《经济学人》上的新闻主题并计算每年占比，然后将占比作为解释变量以预测军事冲突的可能发生地点。

在预测上的差异，并简单评价这两种方法的优劣；其次介绍利用机器学习进行预测的最新文献。

### 5.2.1 OLS 与 Ridge 在预测上的差异

预测的目的在于找出两个变量间的相关关系。假设这两个变量间的真实关系是$y = f(x) + \varepsilon$。此处函数关系$f$客观存在但不为我们所知。无论依赖于机器学习还是计量经济学，研究者的目的都是找到一个与$f$尽可能接近的函数$g$，使得该函数估计值$\hat{y} = g(x)$能够非常好地吻合真实值$y$。评价一种预测方法好坏最常用的标准是均方误差 (Mean Squared Error, MSE)，也就是残差平方的期望 (Hastie *et al*.,2016)，可表述为：

$$\mathrm{MSE} = \mathrm{E}\left[(y - \hat{y})^2\right] \tag{1}$$

当解释变量取值为$x_0$时，预测值$\hat{y} = g(x_0)$与被解释变量真实值$y = f(x_0)$间的差异可被写为：

$$\begin{aligned} \mathrm{Err} &= \mathrm{E}\left[(y - \hat{y})^2 | x = x_0\right] \\ &= \underbrace{[\mathrm{E}(\hat{y}) - y]^2}_{\mathrm{Bias}^2} + \underbrace{\mathrm{E}\left\{[\hat{y} - \mathbb{E}(\hat{y})]^2\right\}}_{\mathrm{Variance}} + \underbrace{\varepsilon^2}_{\mathrm{Noise}} \\ &= \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Noise} \end{aligned} \tag{2}$$

直观的说，均方误差被分解为三部分：估计值与真实值间的偏差 (Bias)、估计值方差 (Variance) 及真实值的扰动方差 (Noise)。其中，扰动方差完全来自于随机扰动项$\varepsilon$，该部分不会消除且也不会由于预测方法的不同而存在差异。因此，不同预测方法减小均方误差的途径就是在偏差和方差间进行取舍。[21]

下面我们从偏差、方差以及最终均方误差三方面，比较 OLS 和 Ridge 在预测方面的差异。为了推导的简洁，假设$Y$与$X$的真实函数关系$f(x)$为线性且解释变量$X$为正交矩阵：[22]

[21] 偏差和方差的权衡取舍 (Bias-Variance Tradeoff) 是机器学习预测中的一个重大问题，详细讨论见 Bishop (2006)、Murphy (2012) 及 Hastie *et al*. (2016)。

[22] 对于一般的非正交矩阵情况，岭回归的预测能力也是优于 OLS 的，严格数学证明见 Theobald (1974)。

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, X^{\mathrm{T}}X = I \tag{3}$$

OLS 的预测函数$g(x)$可表示为$\widehat{Y} = X\widehat{\boldsymbol{\beta}}$，对式中$\widehat{\boldsymbol{\beta}}$的估计方法是最小化残差的平方和，表示为：

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 = \left( X^{\mathrm{T}}X \right)^{-1} X^{\mathrm{T}}Y \tag{4}$$

此时，$\widehat{\boldsymbol{\beta}}_{OLS}$的偏差是：

$$\mathrm{Bias}_{\mathrm{OLS}} = \left( X^{\mathrm{T}}X \right)^{-1} X^{\mathrm{T}}Y - \boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta} = 0 \tag{5}$$

方差是：

$$\mathrm{Var}_{\mathrm{OLS}} = \frac{\sigma^2}{X^{\mathrm{T}}X} = \sigma^2 \tag{6}$$

可知$\widehat{\boldsymbol{\beta}}_{OLS}$是真实值$\boldsymbol{\beta}$的无偏估计量。[23]

由以上可以看出利用 OLS 进行预测的优点在于估计系数偏差为0，缺点是方差可能较大。换句话说，选择若干个随机样本进行多次回归，无偏性保证所获系数的均值接近于系数的真实值。方差较大则意味着单次回归系数偏离均值较远，可能会异常大或小。当解释变量间存在多重共线性时，这一问题尤为严重。

针对此问题，Ridge 在最小化目标函数中引入估计系数平方作为惩罚项，[24]表示为：

$$\widehat{\boldsymbol{\beta}}_{Ridge} = \arg\min \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{m} \beta_j^2}_{\text{penalty term}} = \left( X^T X + \lambda I \right)^{-1} X^T Y \tag{7}$$

若读者对岭回归感兴趣，可以参考 van Wieringen (2018)。

[23] 当$X$不是正交矩阵时，OLS 得到的系数估计也是无偏的，这容易从式 (4) 中看出。

[24] 可以看到，$\lambda$反映了“惩罚力度”，当$\lambda = 0$时意味着无惩罚，此时 Ridge 和 OLS 完全一样。

公式 (7) 在直觉上非常容易理解：OLS 的缺点在于方差大，也就是估计系数的上下波动很剧烈。为了防止这种情况，机器学习在最小化过程中通过加入估计系数的平方或绝对值来“抑制”系数大小。如此便可以减小估计系数的方差使得预测更加稳定。这种思路可以理解为对系数大小的一种惩罚：过大则赋予较小权重，过小则相反。LASSO 和 Ridge 的不同就体现在惩罚系数的选取上：Ridge 惩罚项为系数的平方$\sum_{j=1}^{m}\beta_j^2$，而 LASSO 则是系数的绝对值$\sum_{j=1}^{m}\left|\beta_j\right|$。[25]引入惩罚项后，Ridge 最小化的目标函数较之 OLS 更为复杂，而 LASSO 甚至无法导出估计系数的解析表示，只能求得数值解。操作上，最小化问题往往借助机器学习技术实现。

以下将分别比较 OLS 和 Ridge 估计系数的偏差、方差和均方误差的大小。首先，根据 $X$ 是正交阵假设，由式 (7) 可得系数$\widehat{\boldsymbol{\beta}}_{Ridge}$为：

$$\widehat{\boldsymbol{\beta}}_{Ridge}=\frac{\boldsymbol{\beta}}{1+\lambda} \tag{8}$$

此时估计系数的方差是：

$$\text{Bias}_{\text{Ridge}}=\frac{\boldsymbol{\beta}}{1+\lambda}-\boldsymbol{\beta}\neq 0 \tag{9}$$

偏差是：

$$\text{Var}_{\text{Ridge}}=\frac{\sigma^2}{(1+\lambda)^2}\leq\sigma^2 \tag{10}$$

(9) 式表明Ridge系数估计是一个有偏估计量，而 (10) 式则表示其方差比OLS要更小。换一句话说，在方差和误差的权衡中，Ridge 以有偏为代价换取更小的方差。

[25] 严格说来，LASSO 和 Ridge 的“惩罚项”具有不同的含义：LASSO 进行的是 L1 范数正则化，Ridge 则是 L2 范数正则化。关于正则化详细讨论见 Bishop (2006)、Murphy (2012) 及 Hastie *et al*. (2016)。

在获得了 OLS 和 Ridge 估计系数的偏差和方差后，根据式 (2) 分别计算两者的均方误差：

$$\mathrm{Err}_{\mathrm{OLS}} = 0 + \sigma^2 = \sigma^2 \tag{11}$$

$$\mathrm{Err}_{\mathrm{Ridge}} = \left(\frac{\lambda\boldsymbol{\beta}}{1+\lambda}\right)^2 + \frac{\sigma^2}{(1+\lambda)^2} \tag{12}$$

为了比较两种方法的预测能力，我们将上两式做差：

$$\mathrm{Err}_{\mathrm{OLS}} - \mathrm{Err}_{\mathrm{Ridge}} = \left(0 + \sigma^2\right) - \left[\left(\frac{\lambda\boldsymbol{\beta}}{1+\lambda}\right)^2 + \frac{\sigma^2}{(1+\lambda)^2}\right] \tag{13}$$

若 (13) 取值为正，那么 OLS 预测误差更大，反之 Ridge 的误差更大。可以看到，式 (13) 实际是一个关于$\lambda$的函数，其正负性也依赖于λ的值。现在问题转变为：怎么样的λ会使得 (13) 取正值或者负值？我们先考察该函数极值，如果该函数极大值都小于零，OLS 的均方误差将恒小于 Ridge；反之，如果该函数的极大值大于零，意味着我们一定能找到λ使得 Ridge 的均方误差更小。为找到极值，令 (13) 的导数为零，得到一阶条件$\lambda = \frac{\sigma^2}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}}$。将该一阶条件代入，此时$\mathrm{Err}_{\mathrm{OLS}} - \mathrm{Err}_{\mathrm{Ridge}} = \frac{2(\beta^T\beta)(\sigma^T\sigma)+(\sigma^T\sigma)^2}{\beta^T\beta+\sigma^T\sigma}$。式中分子分母都为正数，因此式 (13) 大于零，这意味着 Ridge 预测能力优于 OLS。事实上，Theobald (1974) 将该条件放宽到$\lambda < 2\sigma^2\left(\beta^{\mathrm{T}}\beta\right)^{-1}$时，Ridge 的均方误差都是小于或等于 OLS 的。

我们从“无偏性”和“可解释性”两方面评价传统计量经济学方法和机器学习方法在预测方面的优劣。正如本章开头所说，任何预测方法都是在偏差和误差间进行权衡取舍。社会科学实证研究，特别是经济学研究，特别强调因果推论。基于这种考虑，计量经济学回归模型都致力于获得一致的估计系数。这意味着在这一方差–偏差权衡中，计量经济学方法宁愿付出方差较大的代价，也不能放弃无偏这一性质 (Athey, forthcoming)。比方说上面所提到的 OLS 的估计系数正体现这一思路。而机器学习的目的就是进行预测——它并不在乎用以做出预测的估计系数是否具有一致无偏性特点。这就意味着在无偏

性上，机器学习做出了“让步”：选择用偏差来换取更小的方差以提高预测性能。“可解释性”指的是从模型估计出的结果能够容易地被解释。计量经济学的目的不仅是预测，更在于解释现实中的现象以找到背后规律。从这个意义上来说，用来预测的函数形式越简单越好。因为复杂模型需要廓清模型拟合好坏的原因及解释变量与被解释变量间的互动关系等诸多问题。[26]机器学习则恰恰相反，只要这个函数能够很好地模拟现实，哪怕函数形式再复杂也无所谓。[27]在这一点上，机器学习不拘泥于“可解释性”，灵活地选择函数形式进行拟合数据，这使得其预测能力强过了计量经济学传统方法。[28]

### 5.2.2 用机器学习预测在文献中的应用

本小节将从个体和社会方面对现有利用机器学习进行预测的文献中进行简单梳理。

在个体层面上，机器学习可以帮我们更好地预测个人信息、决策或未来行为。此类研究包括 Oster (forthcoming)、Kleinberg *et al*. (2017) 、Goel *et al*. (2016) 及 Chalfin *et al*. (2016) 等。Oster (forthcoming) 尝试研究糖尿病患者在确诊后是否会改变饮食结构。由于改变饮食习惯是一件非常痛苦的事，很多糖尿病人都不愿意去节制饮食，以往文献由于缺少数据而无法测度这种“不愿意”的程度。为解决该问题，Oster 首先基于个体消费记录 (Nielsen HomeScan Panel) 中的“血糖仪”、“试纸”等关键字，利用随机森林 (Random Forest) 预测某人是否患糖尿病。得到预测结果后，再分析该个体在患糖尿病前后的食品消费记录来推断其是否改变饮食习惯。[29]该文发现患者因为患糖尿病而改变饮食习惯的幅度非常小。Goel *et al*. (2016) 同样采用随机森林方法预测哪些行人更有可能携带武器；Kleinberg *et al*. (2017) 则通过“梯度提升决策树”(Gradient Boosted Decision Trees) 预测

[26] 对于$y = x$，解释变量增加一个单位，被解释变量也增加一个单位，两者成正比。但对于$y = \sin x + \log x^2$，即便该函数对现实规律可能拟合地更好，研究者也很难对该函数进行“经济学解释”。因此在计量经济学中，学者很少使用类似复杂形式的函数模型。

[27] 很多机器学习技术都欠缺可解释性。比如，神经网络技术甚至能被视为黑箱：研究者无法得知解释变量通过何种机制影响被解释变量。一些致力于机器学习的学者也注意到了这个问题，正着手提高复杂模型的可解释性。相关讨论见 Goodfellow *et al*. (2016)。

[28] 用机器学习预测除 Ridge 之外，还可以采用非线性、半参或非参预测方法。Mullainathan and Spiess (2017) 对比了这些方法与 OLS 的差异，结果发现机器学习的预测能力普遍强于 OLS。

[29] 随机森林是一种常见的机器学习技术。该方法在技术操作上容易实现且对计算机的运算能力要求也较低，是目前文献中常用的预测方法。其原理可被看作多个“如果–那么”结构。举例来说，如果天气好且空气也不错的话，我就去打网球。此处的天气有“好”、“坏”两种，而后又进一步分出空气“好”“坏”两种情况。计算机根据每种情况计算出打网球的条件概率，这就构成是一颗决策树 (Decision Tree)。多个决策树组合到一起就是所谓的随机森林。

被保释犯人是否会出席庭审；[30]Chalfin *et al*. (2016) 采用“随机梯度提升”(Stochastic Gradient Boosting) 决策树来预测警察的工作质量。[31]作者通过警员入职申请中提及的社会经济状况、婚姻、是否服役等信息，预测其未来工作中是否偏向使用暴力。

相比之下，国内采用机器学习进行预测的相关研究起步较早且大多集中在金融领域，主要集中在对个人或企业的贷款风险信用进行预测方面 (方匡南等，2010；郭英见和吴冲，2009；吕劲松等，201；马晓君，2015；钱争鸣等，2010；徐晓萍和马文杰，2011；苏治等，2017)。限于篇幅，这里无法做过多展开，读者可参考原文。

在社会经济层面，机器学习能够帮助研究者预测经济指标。比如，Blumenstock *et al*. (2015) 试图研究发展中国家的财富分布情况。作者遇到的问题和本文第三部分提到的 Engstrom *et al*. (2017) 类似：落后地区的官方数据质量较差。和 Engstrom 依赖于卫星图像不同，Blumenstock 认为手机元数据 (mobile phone metadata，如通话历史等信息) 不光能用来推断手机使用者的财富状况，同时也具有源覆盖范围广、质量高且廉价的优势。作者先收集 856 人的手机元数据及他们的经济状况，再利用弹性网络 (Elastic Net) 建立手机数据与经济状况间的函数关系。作者发现该函数关系可以很好的预测财富分布。类似文献还有 Glaeser *et al*. (2017)，作者试图测量基层的即时经济状况。由于政府统计报表公布时效性差，作者基于网络平台 Yelp 数据并通过随机森林来预测时时微观经济活动。[32]国内学者也尝试通过机器学习完成预测目标，比如刘涛雄和徐晓飞 (2015) 及孙毅等 (2014) 研究通过互联网搜索数据分别对 GDP 及通货膨胀率进行预测。陈硕和王宣艺 (2018) 试图利用基层社会经济指标预测 GDP。

### 5.3 因果推断

[30] 梯度提升决策树可被看成决策树套入到梯度提升的框架里。梯度提升主要用来解决运算中的最优化问题。如果把运算比作爬山，每一步均沿着山坡最陡峭的方向爬，这样会使得爬到山顶的路径最短，这就是“梯度提升”的思想。

[31] 该技术的基本思想依旧是决策树，而“框架”则是“随机梯度下降”。脚标 30 提到的“上升”和此处的“下降”没有本质区别，都是最优化方法，差别仅在于根据不同的目标函数求出最大值或最小值。采用“随机”的目的是为了减少运算量：没必要每一步都精确地算出究竟哪个方向的山坡最陡峭。登山者只需要随机选取一个方向，保证每一步都是向上的，那么早晚都会爬到山顶。此时路径可能较长，但优点是节省了计算山坡每一点梯度的时间。详细讨论见 Bishop (2006)、Murphy (2012) 及 Hastie *et al*. (2016)。

[32] Yelp 是美国最大的点评网站，类似中国的大众点评。

社会科学，尤其是经济学实证研究的核心目标是获得因果推论，以探究干预 (Treatment) 措施是否导致预期结果并廓清作用发生机制。本部分将讨论机器学习技术在这方面的应用。我们首先基于著名的 Neyman-Rubin 反事实框架 (Neyman-Rubin Counterfactual Framework) (Neyman 1923; Rubin 1974) 给出“因果效应”的定义；随后结合目前应用微观计量经济学广泛使用的两种因果推论方法：双重差分 (Difference-in-Differences, DID) 及断点回归 (Regression Discontinuity, RD) 展示该技术在其中的应用。

### 5.3.1 因果关系与反事实

我们依然沿用 Rubin (1974) 中患者吃药的例子来探究药物能否“导致”疾病被治愈。在现实世界，能够被我们观察到的“药物效果”是那些头痛并吃了药的人的健康状况减去那些健康且没吃药的人的健康状况，用公式表述为：

$$\text{Observed Effect} = \mathrm{E}[Y_{1i}|D_i = 1] - \mathrm{E}[Y_{0i}|D_i = 0] \tag{14}$$

式 (14) 中，$Y$表示个体健康程度，第一个下标为1表示该个体实际吃了药，0 表示没有吃药；第二个下标$i$代表第$i$个观察对象。$D_i$取值1或0分别表示该患者是否患有头痛。因此$\mathrm{E}[Y_{1i}|D_i = 1]$就表示那些头痛且的确吃了药的患者健康程度，$\mathrm{E}[Y_{0i}|D_i = 0]$则表示健康且没有吃药的人的健康程度。

显而易见，公式 (14) 并不代表药物的因果作用：那些没吃药的人相对较为健康，与那些因为头痛而吃药的患者不可比。在这种情况下我们无法区分两个群体在吃药后身体状况的差异到底来自于药物效果还是来自于个体差异。真正的效果应该是除了吃药与否外，其他所有因素都一样 (*ceteris paribus*)。换句话说，药物作用应当是同样一群患者($D_i$均为1)，分别测量没有吃药和吃药后的健康程度的差别，公式表述如下：

$$\text{Average Treatment Effect} = \mathrm{E}[Y_{1i}|D_i = 1] - \mathrm{E}[Y_{0i}|D_i = 1] \tag{15}$$

公式 (15) 可以理解为药物作用是吃药的患者健康程度 ($\mathrm{E}[Y_{1i}|D_i = 1]$) 减去如果他没

有吃药时的健康程度 ($\mathrm{E}[Y_{0i}|D_i=1]$)。但遗憾的是，这两者在真实世界中永远无法同时被观察到的。我们把这种无法观察到的情况定义为吃药患者健康程度的“反事实”。虽然公式 (15) 在真实世界中没有操作性，但这不妨用该公式来重新组织公式 (14)：

$$\begin{aligned}\text{Observed Effect} &= \mathrm{E}[Y_{1i}|D_i=1]-\mathrm{E}[Y_{0i}|D_i=0]\\ &= \underbrace{\mathrm{E}[Y_{1i}|D_i=1]-\mathrm{E}[Y_{0i}|D_i=1]}_{\text{Average Treatment Effect}}+\underbrace{\mathrm{E}[Y_{0i}|D_i=1]-\mathrm{E}[Y_{0i}|D_i=0]}_{\text{Selection Bias}}\end{aligned} \tag{16}$$

公式 (16) 可由公式 (14) 中减掉反事实再加上反事实后获得，这依然是观察到的“因果效应”。但此时公式 (16) 由两部分组成：第一部分正是药物的真实作用 (式 15)，第二部分是$\mathrm{E}[Y_{0i}|D_i=1]-\mathrm{E}[Y_{0i}|D_i=0]$，我们将其定义为选择偏差 (Selection Bias)。该部分字面意义是那些有病没吃药的患者的健康程度减去那些没病也没吃药的个体健康程度。显而易见，该选择偏差小于0：有病但没吃药的患者的健康程度当然差于没病也没吃药的个体健康程度。自此，可以知道观察法获得的所谓“因果效应”等于真实因果效应加上一个小于0的选择偏差。换句话说，观察法获得的“因果效应”低估了真实因果效应。[33]

缺乏反事实使得观察法获得的因果效应并不等于真实因果效应。该问题被称之为因果推论的根本问题 (Fundamental Problem in Causal Inference)。而从以上分析可以发现，该问题根源在于个体差异：健康个体和病患个体存在诸多差异，因而前者无法作为后者的“反事实”。传统计量经济学所发展出来的所有分析工具，不管采用何种研究设计，其最终目的都是构建出介入组 (Treatment Group) 的反事实。一般来说，这些方法通过寻找恰当的控制组 (Control Group) 并提供证据来论证或假定该组可以作为介入组的反事实。找到适当的控制组后，它在介入后的取值即可以作为介入组的反事实，二者间差异为介入效果。

这个过程就为机器学习的应用提供了机会：与其直接计算介入组和控制组在介入后的差异，不如利用控制组中样本构建出某种函数 (比如样本的加权平均)，使得该函数的取值

[33] 在大多数情况下，我们并不清楚选择偏差是否大于还是小于零。这意味着在大多数情况下，偏差方向是未知的。

与介入组足够相似，从而便可将该函数在介入后的取值作为反事实。用公式表述如下：

$$\widehat{Y_{\mathrm{T}}}(post) = f\big(Y_{\mathrm{C}}(pre), Y_{\mathrm{T}}(pre)\big) \tag{17}$$

其中$Y_{\mathrm{C}}(pre)$是控制组没有被介入时的取值，$Y_{\mathrm{T}}(pre)$是介入组在介入前取值。而$\widehat{Y_{\mathrm{T}}}(post)$是反事实的预测值，表示介入组如果没有被介入时的取值。该函数作为介入组反事实的合理性可以通过其在介入前的值与介入组在介入前的值$Y_{\mathrm{T}}(pre)$的差值反映。如果两者差异不大，表示该函数和介入组足够相似：用它作为反事实是可靠的。一般来说，我们用该函数的预测值和$Y_{\mathrm{T}}(pre)$的差值平方来评价，公式表示如下：

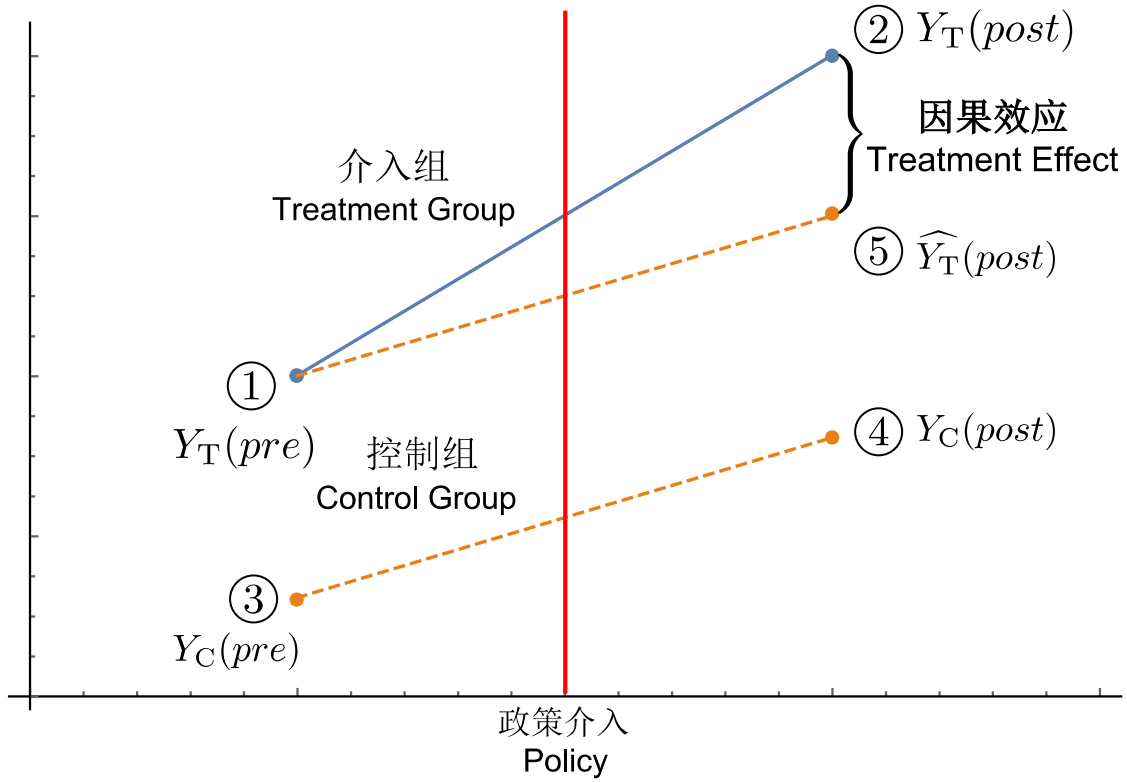
$$\text{Error} = \sum_{i \in pre} \left[Y_T(i) - \widehat{Y_T}(i)\right]^2 \tag{18}$$

如果将式 (18) 与第三部分式 (1) 对比，可以发现两者最小化目标函数十分相似：式 (18) 是最小化残差平方和，而式 (1) 是最小化均方误差 (MSE)。从这意义上，对反事实的估计可以被视作为一种预测 (Varian 2016; Athey 2018)。以下将依次结合 DID 及 RD 具体展示机器学习如何实现对反事实的估计。

### 5.3.2 双重差分方法

双重差分是当前应用微观计量经济学中最常用的政策评估方法之一。我们用图 4 来展示该方法识别因果的策略。

**图 4 双重差分 (DID) 识别策略示意图**




注：淡蓝色代表介入组 (地区T)，橙黄色代表控制组 (地区C)，红色竖线代表在地区T实施的某一政策。

假设某个地区 (地区T) 被政策介入，该地区某项我们感兴趣的指标在政策前 (红线左边) 的取值为①，政策实施后 (红线右边) 的取值为②。很明显，② − ①并不是政策效果：我们并不知道该地区如果没有政策的话，该指标取值是多少 (反事实)。为了解决该问题，DID 的策略是寻找另外一个没有被政策介入的地区 (地区C)，该指标在C地区政策前后的取值分别是③和④。①、②、③和④的取值分别表示为$Y_{\mathrm{T}}(pre)$、$Y_{\mathrm{T}}(post)$、$Y_{\mathrm{C}}(pre)$和$Y_{\mathrm{C}}(post)$。DID 假设⑤是T的反事实$\widehat{Y_{\mathrm{T}}}(post)$。⑤的取值可以通过①、③和④点获得：$⑤ = ① + (④ - ③)$，其中④ − ③为C地区的时间趋势，假定与T地区相同。因此，政策效果被表示为：$② - [① + (④ - ③)]$，也可以表示为$(② - ①) - (④ - ③)$，这也就是该方法被称之为双重差分的原因。当然，出于展示的便利，图 4 中仅用 4 个点表示所有样本的四种情况。在实际研究中，这四个点背后会有很有观察值。此时，$① + (④ - ③)$就被表述成：

$$\widehat{Y_{\mathrm{T}}}(post)=\frac{1}{N}\sum_{i\in Control}Y_{\mathrm{C},i}(post)+\mathrm{Constant} \tag{19}$$

其中$Y$的第一个下标仍然代表的控制组 (地区C)，而第二个下标$i$则表示样本中的第$i$个观察值。[34]式子末尾的常数项Constant为实验组与控制组在施加处理前的差异，也就是图 4 中①和③的垂直距离。因此，双重差分方法比较的是这四个点背后所有观察值的算数平均的差值。

从以上分析可得，DID 方法有效性依赖于两个地区存在相似的时间趋势。如果研究者用样本算数平均数来构建反事实，上述平行趋势假设并不容易满足。此时，采用更加一般化的加权平均方法来构建反事实的效果可能更好。这是因为构建反事实的根本在于找出与介入组T尽可能同质的控制组C，然而C中每一个观察对象与介入组的相似度可能各不相同。我们自然会想到给那些与T相似的观察对象赋予更大的权重，而非给所有对象相同大小的权重。该思路被称之为合成控制法 (Synthetic Control Method, SCM) (Abadie *et al*. 2010)。在该方法下，反事实被表示为：

$$\widehat{Y_{\mathrm{T}}}(post)=\sum_{i\in\mathrm{C}}\omega_i Y_{\mathrm{C},i}(post),\sum_{i\in\mathrm{C}}\omega_i=1 \tag{20}$$

下一步需要解决的问题是每一个观察值所被赋予的权重$\omega_i$，这可以通过最小化式 (6) 的残差平方和得到：

$$\widehat{\omega_i}=\arg\min_{\omega_i}\sum_{j\in pre}\left(Y_{\mathrm{T}}(j)-\left(\sum_{i\in pre}\omega_i Y_{\mathrm{C},i}(j)\right)\right)^2 \tag{21}$$
$$\text{s.t.}\sum_{i=1}^{N}\omega_i=1\ and\ \omega_i\geq 0, i=1,2,\cdots,N$$

受到 SCM 的启发，Doudchenko and Imbens (2016) 将加权平均进一步放松为更加一般的线性组合函数来构建反事实，这也成为了机器学习在 DID 中应用的基本思路。此时该一

[34] 为了展示的便利，此处我们假设介入组$Y_{\mathrm{T}}$中仅有一个样本。因此，$\widehat{Y_{\mathrm{T}}}(post)$没有$i$下标。该假设并不损失一般性 (Doudchenko and Imbens, 2016)。

般线性函数可被表达为：

$$\widehat{Y_{\mathrm{T}}}(post) = \sum_{i \in \mathrm{C}} k_i Y_{\mathrm{C},i}(post) + b \tag{22}$$

较之 SCM，此处并不要求$k_i$是权重，$k_i$甚至可以取负值。同时，对$\sum k_i$也不作任何要求。接下来的问题是如何找到参数$k_i$和$b$来最小化式 (6)。两位学者使用了正则化回归 (Regularised Regression)，具体细节可以参照原文。至此，我们通过机器学习技术“改善”了 DID 方法中对反事实的估计：利用控制组和介入组在政策实施前的信息建立线性函数并预测出反事实。由于该技术较为前沿，目前暂时还没有运用该方法进行政策评估的具体研究。[35]

### 5.3.3 断点回归方法

断点回归方法也是另外一种被广泛使用的因果识别方法。和双重差分方法用政策前后区分介入-控制样本组不同，样本是否被介入依赖于其中某一变量$X$ (Forcing Variable) 相对于断点 (Cut-off) 的大小，可以表述为：[36]

$$\text{Treatment} = \begin{cases} 1, X \geq \text{cut-off} \\ 0, X < \text{cut-off} \end{cases} \tag{23}$$

该式子表示当$X \geq$ cut-off时，对应的样本被定义为介入组，而当$X <$ cut-off时，对应样本被定义为控制组。介入组和控制组都有相对应的被解释变量取值。假设被解释变量在相应控制组和介入组内部是连续变化的，如果该变量在断点左右出现跳跃，我们可以将其归咎于“基于断点”的介入所导致的因果效应 (Imbens and Lemieux, 2007)。

可以通过一个例子更加直观地展示上述研究设计的逻辑并在其中指出机器学习发挥作用的地方。最常见使用 RD 方法的研究问题是大学教育对工资的影响。研究者当然不能

[35] 在 Doudchenko and Imbens (2016) 的研究中，两位作者首先用理论推导的方法表述机器学习如何与 DID 结合进行因果识别，然后以现有文献研究过的三个问题对其展示。这三个例子分别是加州的禁烟管制 (Abadie *et al*., 2010)、德国统一 (Abadie *et al*., 2015) 及马列尔偷渡 (Card, 1990; Peri and Yasenov, 2015)。

[36] 不妨假设大于断点的那部分样本被介入。

直接比较大学生和那些没上大学学生的平均工资差异：这两群人在能力上可能并不相同。这两群人在业界的工资差异除了反映出大学经历的作用外，也提取了诸如能力等个体异质性因素。而在实际操作中，能力非常难以精确度量。断点回归的目的是通过比较两组“相似”的样本进而剔除掉这些异质性因素，以下是研究设计。

假设大学录取分数线是 60 分，那么总有学生因为 1 分之差和大学失之交臂。在这种情况下，我们大都说靠 59 分的人运气太差，而非能力不行。换句话说，那些考了 59 分没能上大学的学生和恰好踩线 60 分得以进入大学的同学在能力上可能没什么太大差别，最终能否上大学完全是运气，而运气是随机的！这就使得我们在真实世界难能可贵地找到了用随机选择分配大学的机会。在该设定下，那些考 59 分的同学就无限接近为那些考了 60 分同学的反事实：考了 60 分但没读大学的同学。所以 59 分同学和 60 分同学在未来的工资差别可以视作为大学教育对收入的因果效应。

**图 5 断点回归 (RD) 识别策略示意图**

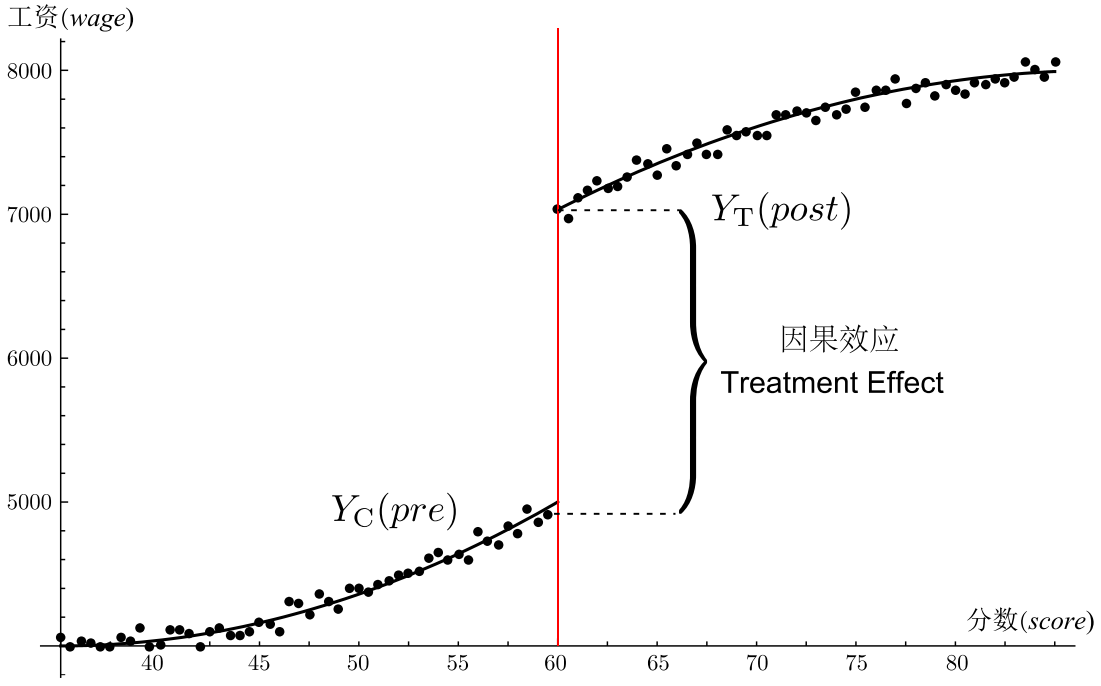



图 5 展示了上述研究设计思路，其中横坐标代表考生分数，纵坐标代表工资。图中的 60 分，也就是能否上大学的分界线，用红色竖线加以表示。可以看到，那些考上 60 分的学生们被划分到介入组T，59 分的学生们则是控制组C。在 DID 设置中，前者就是$\mathrm{Y_T}(post)$

(介入组被介入后，②)，后者就是$\mathrm{Y_C}(pre)$ (控制组在介入前，③)。和 DID 的区别在于，RD 的研究设置中不存在$\mathrm{Y_T}(pre)$和$\mathrm{Y_C}(post)$——前者表示考上 60 分的学生 (介入组) 没有上大学的工资 (①)，后者表示 60 分以下的学生 (控制组) 上了大学的工资 (④)。由于缺乏时间维度带来的这两个信息，RD 无法使用 DID 中采用的①③及④的信息构建出反事实⑤。在这种情况下，RD 采用的策略是绕开①和④的信息，直接假设③等同于⑤。该假设在 RD 中之所以成立的原因在于上文所说的考 59 分的学生和考 60 分的学生高度同质。此时，用公式表述大学的因果效应如下：

$$\begin{aligned}\text{Treatment Effect} &= \lim_{\varepsilon\to 0^+}\mathrm{E}[Y_i|score_i = 60+\varepsilon] - \lim_{\varepsilon\to 0^-}\mathrm{E}[Y_i|score_i = 60+\varepsilon] \\ &\approx \mathrm{E}[Y_i|score_i = 60] - \mathrm{E}[Y_i|score_i = 59] \\ &= Y_{\mathrm{T}}(post) - Y_{\mathrm{C}}(pre)\end{aligned} \tag{24}$$

式中的$score_i$代表第$i$个学生的分数，$Y_i$表示其工资。式中的第一行利用了左、右极限刻画了真实的因果效应：同样考 60 分的一群学生仅仅由于是否上大学而导致的工资差。随后依据上文假设，RD 近似地认为 59 分与 60 分的学生同质，从而得到了后两行的结论。

即使如此，挑战依然存在。考 59 分和 60 分的学生在能力上相同的假设可能太强：1 分只差也是能力的差别！此时，RD 效果依然会提取能力作用进而导致大学对工资的影响被高估。机器学习可以帮助研究者消除这一差异 (Varian, 2016)。通过各种预测模型，机器学习技术能够通过以下步骤构建出考到 60 分但却没有读大学的同学的未来工资 (⑤)。首先，机器学习可以利用③的信息 (小于等于 59 分的样本) 归纳出没有上大学人群中工资与分数间的关系：

$$Y_C(score) = f(score), score \in pre \tag{25}$$

然后扩大函数的“定义域”：将 60 分作为解释变量带入 (14) 右端的函数中。此时函数

的取值就是那些“倘若”考到 60 分但却没有接受大学教育的同学未来的工资，将此作为 60 分且读大学同学的反事实 (⑤)：

$$\widehat{Y_{\mathrm{T}}}(60) = \hat{f}(60) \tag{26}$$

此时，比较机器学习得到的反事实$\hat{f}(60)$与传统 RD 方法得到的反事实$Y_{\mathrm{C}}(59)$，我们可以发现机器学习已经将 59 到 60 分之间的能力剔除在外了，进而获得更加精确的因果效应。

现在问题转变为怎样的机器学习预测函数$f$能够达成推测因果效应的目的。第一，$f$应当具有较好的预测性能，即尽可能减小均方误差。第二，$f$给出的估计量应当具有良好的统计性质：包括在大样本下渐进一致性以及较窄置信区间等。针对这些要求，Imbens and Wager (forthcoming) 利用凸优化的数值方法 (Numerical Convex Optimization Method) 来进行断点回归的因果推断，超越了传统上用来进行 RD 识别的局部线性回归 (Local Linear Regression)。[37]

### 5.3.4 工具变量方法[38]

除了双重差分及断点回归方法之外，应用微观计量经济学者也经常使用工具变量方法 (Instrumental Variable Approach) 来识别因果关系。和以上两种方法依赖于寻找同质样本的思路不同，工具变量方法试图寻找外生变量来克服异质性与样本是否被介入间的关系。实际操作采用两阶段最小二乘法 (Two-stage Least Squares, 2SLS) 实现。在第一阶段通过 OLS 线性估计用外生工具变量“替代”内生解释变量 (是否介入或者介入的程度)，从而获得内生解释变量的预测值。该预测值的方差都是由于外生工具变量所解释，与异质性之间的关系便不再存在。在第二步中，用解释变量预测值和被解释变量回归，获得

[37]传统方法有两大不足：首先，传统方法的前提假设是变量的连续性，在上文例子中该假设意味着分数必须是连续的，然而很多实证研究里的变量都是离散的。其次，传统方法虽然具有一致性，但它并不是最小化均方误差的估计量。两位作者的研究利用了机器学习中常用的凸优化数值方法改进了这两个问题，具体细节请参考原文。局部线性回归的相关研究见 Armstrong and Kolesár (2018) 及 Kolesár and Rothe (forthcoming)。

[38] 机器学习在工具变量方法中的应用本质上仍然属于预测部分，而并不涉及反事实的估计。处于内容的相关性，我们将这部分放在因果推论部分论述。

解释变量的一致性估计系数。

我们在这里仍然采用教育对收入作用来展示工具变量方法的操作方法。该例子来自 Angrist and Krueger (1991)，作者试图估计教育时长对工资的作用，估计公式如下：

$$wage_i = \beta_0 + \beta_1 edu_i + u_i \tag{27}$$

上文提到，不管是否上大学还是大学教育时间长短均和个体异质性有关系，这就意味着上述因果关系中存在内生性问题：$\mathrm{Cov}(edu, u) \neq 0$。这导致研究者无法区分观察到的收入差异到底来自于教育还是个体异质性。为了应对该内生性问题，两位作者采用工具变量方法，他们认为出生时间$z$是一个很好的工具变量。对该变量的评价需要了解一下美国的义务教育制度：义务教育法律规定学童在年满 6 周岁时要入学读书，年满 16 周岁后才可以离开学校。[39]法律规定的“年满 6 周岁”指的是当年 1 月 1 日年满 6 周岁。该一刀切的规定会导致出生月份不同的学童实际接收教育时长存在差别。举一个极端例子，一个 12 月 31 日出生的学童，在 6 年后的 1 月 1 日时恰好 6 周岁多一天。按照法律规定，该学童符合入学条件。而另一个在 1 月 2 日出生的学童在入学日时却只有 5 周岁 364 天。虽然之有 1 天之差，但依然不能入学，必须等到下一年 1 月 1 日。那时他已经 6 周岁 364 天。由于离校都是 16 周岁的那天，这会导致 1 月 2 日出生的学童比 12 月 31 日出生的学童少接受 364 天教育。当然大部分学童受到的教育时常都小于该极端值。可以看出，上述制度设置所导致的教育时长差异是由于出生月份导致，如果我们假设能力和出生月份无关的话，那么该变量就是教育时长的有效工具变量。可用公式表示为：

$$\mathrm{Cov}(edu, z) \neq 0, \mathrm{Cov}(u, z) = 0 \tag{28}$$

满足上述条件之后，研究者便可以用两阶段最小二乘法估计教育对收入的影响作用。在实际操作上，先将教育$edu$与出生月份$z$进行回归 (第一阶段)：

$$edu = Z\delta + v \tag{29}$$

[39] 有些州规定的离校最低年龄是 17 岁，这不会影响我们下面的讨论。

该阶段的目标是获得教育的预测值：

$$\widehat{edu} = Z\hat{\delta} = Z\left(Z^T Z\right)^{-1} Z^T edu = P_Z edu \tag{30}$$

接下来，把$\widehat{edu}$作为解释变量，工资$wage$作为解释变量，再进行回归（第二阶段）：

$$wage = \widehat{edu}\beta + u \tag{31}$$

最终得到的系数估计$\beta_{2\text{SLS}} = \left(edu^T P_Z edu\right)^{-1} edu^T P_Z wage$，该系数是教育作用的一致性估计值。

工具变量方法的实施关键在于第一阶段，不光需要给出证据证明工作变量具有外生性，还要通过统计指标说明该工具变量和内生解释变量之间存在足够强的相关关系。在这篇研究中，作者给出一些证据比如$Z$估计值的显著性来说明出生季节的确和教育时长之间存在相关关系，但后续许多学者认为该相关关系并不强以至于影响最终的估计结果(Bound *et al*., 1995; Staiger and Stock, 1997; Card, 1999)。[40]该问题本质上仍然是外生$Z$对内生$edu$的预测能力，而这正是机器学习最擅长的地方 (Varian, 2016; Mullainathan and Spiess, 2017; Athey, 2018)。因此，工具变量方法的第一阶段完全可以采用机器学习技术预测内生解释变量。这一领域已经积累起了较多的理论计量文献：有些学者采用正则化回归，比如 LASSO 和 Ridge 等方法来构建第一阶段的估计 (Belloni, *et al*. 2012; Carrasco, 2012; Hansen and Kozbur, 2014)；另一些学者则采用神经网络等非线性方法来进行第一阶段的估计(Hartford *et al*., 2016)。

## 六、展望及结论

和业界及自然科学领域中机器学习技术的应用相比，社会科学中该技术的应用近几年

[40] 限于篇幅，这里我们无法给出弱 IV 为什么会使得估计结果产生偏差或产生不一致的系数估计。关于 IV 影响估计结果的讨论,见 Bound *et al*. (1995); Staiger and Stock (1997) 及 Chao and Swanson (2005)。

也获得了长足发展，但整体来说仍然处于较为初步的阶段。不管数据生成、预测还是因果识别，我们都认为机器学习技术的引入对整个社会科学研究范式的冲击有限。就数据生成来说，机器学习仅提高了数据搜集和整理的生产率，将以前通过人力难以获得的数据变为可得。但这些由机器学习生成的数据依然以变量形式进入到传统社会科学研究框架内，本质上没有改变社会科学的研究方法；就预测来说，目前社会科学在该领域的应用在很大程度上是对业界已经成果的复制。引领这一领域发展的驱动力依然是商业应用；就最有可能产生颠覆意义的因果识别来说，虽然利用机器学习的预测优势构建处理组的反事实在方法论上行得通，但目前并没有被研究者所广泛接受和使用。本文认为其原因有两个：第一，很大程度上在于社会科学，特别是经济学在识别因果上已经发展出非常成熟的范式。除非能够带来颠覆性的边际贡献，研究没有理由抛弃传统因果识别方法。我们认为目前一个较为务实的做法是将机器学习识别因果的相关证据作为稳健性检验方式放入原有框架；第二，充分发挥机器学习的预测能力依赖于海量数据，当前社会科学研究的样本量远没有达到能够让其获得精准预测的下限。[41]

机器学习使得研究者获得了以前通过人工投入无法获得的海量数据，检验了一些依靠传统方法无法有效的假设，这在一定程度上拓展了社会科学研究的边界。我们相信未来几年会有越来越多的学者会在研究中尝试机器学习技术。但我们也必须对该技术在应用过程中可能带来的问题也要有清醒的认识，这主要涉及学者间不平等及数据可复制性问题。机器学习依赖海量数据，这些数据的产生者主要来自业界和政府组织。可以想象，获得这些数据的主要方式并不是团队劳动投入，而是学者通过个人和组织的网络关系获得使用许可。这无疑给大部分学者设置了进入障碍，进而造成赢者通吃并可能加剧学界内部的不平等；机器学习带来的另外一个问题是研究的可复制性。学者通过公布数据及程序代码可以让其他学者和学生复制论文结论。但基于大数据的研究，学者虽然可以公布代码，但数据的公开必须获得数据提供方的许可。和一般数据相比，业界和政府可能更不情愿公布这些海量数据。这可能导致基于大数据研究的可复制性降低。我们对此的建议是，学者在获得数据的同时一并争取获得在未来公布其中的若干部分 (比如数据量的万分之一) 的权利：随机取样的子样本依然具有重复复制

[41] 2017 年《美国经济学评论》(*American Economic Review*) 和《经济研究》上面实证研究样本量的中位数分别是 16068 和 4557。

的价值。

## 参考文献


Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105 (490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. “Comparative Politics and the Synthetic Control Method.” *American Journal of Political Science* 59 (2): 495–510.

Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, *et al*. 2000. “Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling.” *Nature* 403 (6769): 503–11.

Angrist, Joshua D., and Alan B. Keueger. 1991. “Does Compulsory School Attendance Affect Schooling and Earnings?” *Quarterly Journal of Economics* 106 (4): 979–1014.

Antweiler, Werner, and Murray Z. Frank. 2004. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” *Journal of Finance* 59 (3): 1259–94.

Armstrong, Timothy B., and Michal Kolesár. 2018. “Optimal Inference in a Class of Regression Models.” *Econometrica* 86 (2): 655–83.

Athey, Susan. 2017. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355 (6324): 483–85.

Athey, Susan. Forthcoming. “The Impact of Machine Learning on Economics.” In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press.

Athey, Susan, and Guido W. Imbens. 2017. “The State of Applied Econometrics: Causality and Policy Evaluation.” *Journal of Economic Perspectives* 31 (2): 3–32.

Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. 2014. “Searching for Exotic Particles in High-Energy Physics with Deep Learning.” *Nature Communications* 5: 4308.

Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. 2015. “Enhanced Higgs Boson to $\tau+ \tau-$ Search with Deep Learning.” *Physical Review Letters* 114 (11): 111801.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain.” *Econometrica* 80 (6): 2369–2429.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. “Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data.” *American Political Science Review* 110 (2): 278–95.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. 1sted. New York: Springer-Verlag New York.

Bleakley, Hoyt, and Joseph Ferrie. 2016. “Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations.” *Quarterly Journal of Economics* 131 (3): 1455–95.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3: 993–1022.

Blum, Avrim L., and Pat Langley. 1997. “Selection of Relevant Features and Examples in Machine Learning.” *Artificial Intelligence* 97 (1–2): 245–71.

Blumenstock, J., Gabriel Cadamuro, and Robert On. 2015. “Predicting Poverty and Wealth from Mobile Phone Metadata.” *Science* 350 (6264): 1073–76.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak.” *Journal of the American Statistical Association* 90 (430): 443–50.

Bouton, Chad E., Ammar Shaikhouni, Nicholas V. Annetta, Marcia A. Bockbrader, David A. Friedenberg, Dylan M. Nielson, Gaurav Sharma, *et al*. 2016. “Restoring Cortical Control of Functional Movement in a Human with Quadriplegia.” *Nature* 533 (7602): 247–50.

Camerer, Colin F. Forthcoming. “Artificial Intelligence and Behavioral Economics.” In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press.

Cao, Yiming and Shuo Chen. 2018. “Bad Romance.” School of Economics, Fudan University Working Paper.

Card, David. 1999. “The Causal Effect of Education on Earnings.” In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 1801–63. Amsterdam: North-Holland.

Card, David. 1990. “The Impact of the Mariel Boatlift on the Miami Labor Market.” *ILR Review* 43 (2): 245–57.

Carlson, David, and Jacob M. Montgomery. 2017. “A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts.” *American Political Science Review* 111 (4): 835–43.

Carrasco, Marine. 2012. “A Regularization Approach to the Many Instruments Problem.” *Journal of Econometrics* 170 (2): 383–98.

Carrasquilla, Juan, and Roger G. Melko. 2017. “Machine Learning Phases of Matter.” *Nature Physics* 13 (5): 431–34.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. “Productivity and Selection of Human Capital with Machine Learning.” *American Economic Review* 106 (5): 124–27.

Dash, Manoranjan, and Huan Liu. 1997. “Feature Selection for Classification.” *Intelligent Data Analysis* 1 (1–4): 131–56.

Doudchenko, Nikolay, and Guido Imbens. 2016. “Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis.” NBER Working Paper 22791.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. “What Drives Media Slant? Evidence from U.S. Daily Newspapers.” *Econometrica* 78 (1): 35–71.

Gentzkow, Matthew, Jesse Shapiro, and Matt Taddy. 2016. “Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech.” NBER Working Paper 22423.

Gilbert, Fiona J., Susan M. Astley, Maureen G.C. Gillan, Olorunsola F. Agbaje, Matthew G. Wallis, Jonathan James, Caroline R.M. Boggis, and Stephen W. Duffy. 2008. “Single Reading with Computer-Aided Detection for Screening Mammography.” *New England Journal of Medicine* 359 (16): 1675–84.

Glaeser, Edward L., Hyunjin Kim, and Michael Luca. 2017. “Nowcasting the Local Economy:

Using Yelp Data to Measure Economic Activity." NBER Working Paper 24010.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *Annals of Applied Statistics* 10 (1): 365–94.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge: MIT Press.

Halligan, Steve, Douglas G. Altman, Susan Mallett, Stuart A. Taylor, David Burling, Mary Roddie, Lesley Honeyfield, Justine McQuillan, Hamdan Amin, and Jamshid Dehmeshki. 2006. "Computed Tomographic Colonography: Assessment of Radiologist Performance With and Without Computer-Aided Detection." *Gastroenterology* 131 (6): 1690–99.

Hansen, Christian, and Damian Kozbur. 2014. "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE." *Journal of Econometrics* 182 (2): 290–308.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *Quarterly Journal of Economics* 133 (2): 801–70.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." ArXiv: 1612.09596.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nded. New York: Springer.

Hills, Thomas, Eugenio Proto, and Daniel Sgroi. 2015. "Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books." CESifo Working Paper 5906.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy* 124 (5): 1423–65.

Huff, Connor, and Joshua D. Kertzer. 2018. "How the Public Defines Terrorism." *American Journal of Political Science* 62 (1): 55–71.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*." *Quarterly Journal of Economics* 133 (2): 927–91.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide

to Practice." *Journal of Econometrics* 142 (2): 615–35.

Imbens, Guido W., and Stefan Wager. forthcoming. "Optimized Regression Discontinuity Designs." *Review of Economics and Statistics*.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–88.

King, Gary, Christopher Lucas, and Richard A. Nielsen. 2017. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference." *American Journal of Political Science* 61 (2): 473–89.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111 (3): 484–501.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Kohavi, Ron, and George H. John. 1997. "Wrappers for Feature Subset Selection." *Artificial Intelligence* 97 (1–2): 273–324.

Kohn, Walter. 1999. "Nobel Lecture: Electronic Structure of Matter—wave Functions and Density Functionals." *Reviews of Modern Physics* 71 (5): 1253–66.

Kolesár, Michal, and Christoph Rothe. forthcoming. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review*.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311–31.

Lehman, Constance D., Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, and Diana L. Miglioretti. 2015. "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection." *JAMA Internal Medicine* 175 (11): 1828.

McKinsey Global Institute. 2016. "The Age of Analytics: Competing in a Data-Driven World." https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-

of-analytics-competing-in-a-data-driven-world.

McKinsey Global Institute. 2017. “Artificial Intelligence – The Next Digital Frontier?” https://www.mckinsey.com/mgi/overview/2017-in-review/whats-next-in-digital-and-ai/artificial-intelligence-the-next-digital-frontier.

McKinsey & Company. 2017. “中国人工智能的未来之路.” http://www.mckinsey.com.cn/中国人工智能的未来之路.

Mitchell, Tom M. 1997. *Machine Learning*. New York: McGraw Hill.

Mott, Alex, Joshua Job, Jean-Roch Vlimant, Daniel Lidar, and Maria Spiropulu. 2017. “Solving a Higgs Optimization Problem with Quantum Annealing for Machine Learning.” *Nature* 550 (7676): 375–79.

Mueller, Hannes Felix, and Christopher Rauh. 2018. “Reading Between the Lines: Prediction of Political Violence Using Newspaper Text.” *American Political Science Review* 112 (2): 358–75.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.

Neyman, Jersey. 1923. “Sur Les Applications de La Théorie Des Probabilités Aux Experiences Agricoles: Essai Des Principes.” *Roczniki Nauk Rolniczych* 10: 1–51.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine.” *New England Journal of Medicine* 375 (13): 1216–19.

Oster, Emily. Forthcoming. “Diabetes and Diet: Purchasing Behavior Change in Response to Health Information.” *American Economic Journal: Applied Economics*.

Peri, Giovanni, and Vasil Yasenov. 2015. “The Labor Market Effects of a Refugee Wave: Applying the Synthetic Control Method to the Mariel Boatlift.” NBER Working Paper 21801.

Qin, Bei, David Strömberg, and Yanhui Wu. 2017. “Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda.” *Journal of Economic Perspectives* 31 (1): 117–40.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and

Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Shipp, Margaret A., Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C.T. Aguiar, Michelle Gaasenbeek, *et al*. 2002. "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning." *Nature Medicine* 8 (1): 68–74.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, *et al*. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557.

Theobald, C. M. 1974. "Generalizations of Mean Square Error Applied to Ridge Regression." *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (1): 103–6.

Nieuwenburg, Evert P. L. van, Ye-Hua Liu, and Sebastian D. Huber. 2017. "Learning Phase Transitions by Confusion." *Nature Physics* 13 (5): 435–39.

Wieringen, Wessel N. van. 2015. "Lecture Notes on Ridge Regression." ArXiv: 1509.09169.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Varian, Hal R. 2016. "Causal Inference in Economics and Marketing." *Proceedings of the National Academy of Sciences* 113 (27): 7310–15.

Wang, Lei. 2016. "Discovering Phase Transitions with Unsupervised Learning." *Physical Review B* 94 (19): 195105.

Wu, Alice H. 2017. "Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum."

Ye, Qing-Hai, Lun-Xiu Qin, Marshonna Forgues, Ping He, Jin Woo Kim, Amy C. Peng, Richard Simon, *et al*. 2003. "Predicting Hepatitis B Virus–positive Metastatic Hepatocellular Carcinomas Using Gene Expression Profiling and Supervised Machine Learning." *Nature Medicine* 9 (4): 416–23.

方匡南、吴见彬、朱建平、谢邦昌：《信贷信息不对称下的信用卡信用风险研究》，《经济研究》2010 年增刊 1：97–107。

方意:《主板与中小板、创业板市场之间的非线性研究:“市场分割”抑或“危机传染”?》,《经济学(季刊)》2015 年第 4 期:373–402。

郭英见、吴冲:《基于信息融合的商业银行信用风险评估模型研究》,《金融研究》2009 年第 1 期:95–106。

李静:《文化创意产业与乡村旅游产业的融合发展研究》,《管理世界》2017 年第 6 期:182–83。

刘键、戴俭:《我国工业设计产业竞争力分析与发展对策研究》,《管理世界》2017 年第 5 期:182–83。

刘涛雄、徐晓飞:《互联网搜索行为能帮助我们预测宏观经济吗?》,《经济研究》2015 年第 12 期:68–83。

吕劲松、王志成、隋学深、徐权:《基于数据挖掘的商业银行对公信贷资产质量审计研究》,《金融研究》2016 年第 7 期:150–59。

马晓君:《基于数据挖掘的新标准客户信用风险管理规则的构建——以央企中航国际钢铁贸易公司为例》,《管理世界》2015 年第 3 期:184–85。

钱争鸣、李海波、于艳萍:《个人住房按揭贷款违约风险研究.》,《经济研究》2010 年增刊 1:143–52。

沈艺峰、王夫乐、黄娟娟、纪荣嵘:《高管之“人”的先天特征在 IPO 市场中起作用吗》,《管理世界》2017 年第 9 期:141–54。

苏治、卢曼、李德轩:《深度学习的金融实证应用:动态、贡献与展望》,《金融研究》2017 年第 5 期:111–26。

孙毅、吕本富、陈航、薛添:《大数据视角的通胀预期测度与应用研究》,《管理世界》2014 年第 4 期:171–72。

徐晓萍、马文杰:《非上市中小企业贷款违约率的定量分析——基于判别分析法和决策树模型的分析》,《金融研究》2011 年第 3 期:111–20。

于焕杰、杜子芳:《基于随机森林的企业监管方法研究》,《管理世界》2017 年第 9 期:180–81。

张自然:《寿险公司的财务评估——基于主成分分析的 RIDIT 方法.》,《管理世界》2016 年第 3 期:182–83。