凡事预则立，不预则废。这句出自《礼记·中庸》的古训，在当今大数据时代背景下，具有更加特殊的意义。大数据时代不仅意味着数据量级和数据资源分布结构发生深刻改变，而且也意味数据处理技术的极大提升，以及数据处理模式的多样化，这为提高预测的精准性提供了较大的便利和新颖的思路。统计理论研究与实践工作天然与数据联系密切，因此廓清大数据预测的特点，摆正对大数据预测的科学态度，对加强社会经济统计预测预判工作具有重要支撑作用。

大数据预测

何　强

大数据预测的特点

首先，大数据预测将传统意义中“预测”拓展到“现测”。传统意义中的预测（forecasting）是指基于已有数据信息集，依照一定的方法和规律，对未来或未曾观测到的事情做出定性或定量的描述，主要包括样本外预测和样本内预测。而所谓现测（nowcasting）是指对极近过去的描述和极近将来的预测，简单说就是用当下预测当下。该术语最初来自气象学领域，主要针对由于信息获取困难等因素，无法准确描述目前已经发生的事情，进而采用其他可得信息进行推测的过程。比如，对于月度居民消费价格指数（CPI）而言，我国国家统计局通常在每月 9-18 日之间公布上个月的统计数据，但是上个月的数据理论上在月末当天就已经可以被统计出来，只是由于报表填报、数据审核、超级汇总等工作耽误了时间，无法在第一时间公布统计结果。但是，如果在上月刚结束时，就充分利用截止到上月末的网上电子商务交易、超市扫描结算、医院电子病历等实时数据对上月 CPI 进行预测，这种方法就可以称之为“现测”，其优势在于能够在官方统计结果公布之前帮助决策者把握趋近真实的价格走势。

美国麻省理工学院承担的“十亿价格项目”（Billion Price Project）对价格现测的研究就是典型的代表。其研究人员每天在网上抓取 50 多万条商品价格信息，测算每日网上价格指数，其中月度数据滞后期只有三天。实践证明，它为判断美国通货膨胀趋势提供了重要支撑。从方法本质上讲，现测的优势体现在它把一个非常困难的预测问题，转化为一个相对简单的描述问题，而这是传统小数据集根本无法企及的。

第二，大数据预测解决不了不确定性问题。即使将所有样本的数据都搜集完整，在绝大多数情况下也无法对未来做到精准预测，因为大数据中常常存在较多的随机因素和偶然因素信息。为了进一步减少不确定性对大

数据预测的不利影响，研究者曾经试图通过在模型中加入更多变量的方式来解决，逻辑是变量越多，提供的信息就越多，对准确预测未来的可能性就越大。但事实上，这种做法导致数据集的维度变得非常大（也即文献中常提到的高维数据）。数据维度越高，数据中的噪声对真实信号产生的干扰就越大，在模型训练时越容易带来过度拟合问题，即虽然会提高模型对历史数据的拟合程度，但却使得对样本外数据的预测效果变得更差。所以在使用机器学习方法建模的时候，要注意避免过度学习。当然，也要避免走向另外一个极端，要保证机器充分学习。这就像家长在教育孩子时，一方面要对孩子的行为加以限制，避免孩子闯祸、伤人伤己，另一方面也不能管得过死让孩子失去活泼的天性。为此，研究者通过大量的实践总结出一种便捷有效的方法，就是在模型拟合之前，事先留出一部分样本数据用于检验其余数据拟合模型在预测方面的优劣。通常预留出的样本数据量比例为20%-30%，尽管这会带来一些损失，但为了达到更好的预测效果，付出这些代价也是值得的。还有一些研究者尝试采用贝叶斯方法来处理这种不确定性问题，这种思路借助贝叶斯模型在概率推理方面的强大功能展开，目前属于较为前沿性的研究。

也是因为大数据预测无法彻底解决不确定性问题，其成功应用的案例多体现在对短期、具体事件的预测，对长期、宏观走势的预测则不具备优势。一方面这是因为时期拉长之后，不确定性也就越大；另一方面是因为具体事件的影响因素较为简单，很多时候只需更多关注因素之间的相关关系即可，而这恰好是大数据的优势。根据这种思路判断，如果试图通过大数据预测中国政府统计未来改革发展的方向，肯定是非常困难的。

第三，对于特定的预测目的而言，所需大数据的数据源通常受限。传统政府统计的数据源与大数据的数据源存在较大的差异。传统政府统计通常在获取指标数据之间，已经清楚界定了调查目的、调查范围、调查对象、调查内容、调查频率、实施方案、工作流程、上报时间和方式等，虽然数据获取的成本比较大，但所得到的数据可以直接服务于统计的目的。与之相反，大数据的数据源通常是业务工作的副产品，也即是说通常没有直接服务于大数据预测的数据源。比如，前面对CPI现测讨论中所用的网上电子商务交易数据和超市扫描结算数据，它们最初原始的目的是记录所在行业业务的交易情况，并不是直接服务于价格推算，其数据成本摊销在业务工作成本之中，但如果将其用作其他用途，基本无须再支付数据生产费用。网络搜索引擎数据、社交媒体数据等其他主要类型大数据的数据源，也具有同样的特征。退一步而言，即使专门为一项预测研究（比如价格统计）设计其所需大数据的数据源获取方案，那么其成本通常也是无法估计的，或者根本就不存

在实施的可能性。这就为利用大数据进行预测提出了一个重要问题，即如何低成本地寻找潜在可用的数据源问题。这个问题非常重要，因为如果数据源的样本数据不充足、数据质量不高，无论多么科学的预测模型都会无能为力。

第四，当前大数据预测模型应用的主流模式是将传统的统计学、计量经济学与机器学习等分析手段充分融合。所谓机器学习，是指计算机根据预先设定的算法，反复通过训练在输入变量与输出变量之间建立某种拟合较好的匹配关系，主要包括决策树、支持向量机、人工神经网络、自组织映射网络、遗传算法等。在建模数据使用方面，通常以结构化数据为主建立模型，然后通过加入以文本、图像、声音等代表的非结构化数据来改善模型的预测能力，单纯以非结构数据为主的预测结果往往并不理想。这种现象背后的机制在于一方面非结构化数据的来源和形式都十分多样，无法用数字或统一的结构表示，往往包含大量噪声，数据质量较低，无法取代传统的结构化数据；另一方面，非结构化数据常常涵盖传统统计调查无法覆盖的信息，尤其是一些实时、高频的信息，因此可以作为结构化数据的有益补充。

此外，需要重点指出的是，大数据预测常常需要面临很多高维数据模型，这就需要对模型进行降维处理，对此目前比较流行的方法是使用惩罚整合分析模型和动态因子模型（这两种模型包含很多衍生模型），其处理的基本思路是通过引入惩罚项或提取变量公共因子的方式，将那些对预测不重要的变量（或其衍生变量）的系数赋予0，从而实现模型的去冗降噪，部分模型还可以同时实现对混频数据的处理（比如数据集中同时包含月度、季度等不同频率的数据）。

第五，大数据预测无法避免“卢卡斯批判”。虽然利用大数据模型有助于提高预测的精度，但如果将模型预测的结果用于某种政策制度，实际的结果往往不如模型预测的那样理想，因为政策制度的变化会影响到被管理者的预期，预期的改变会使模型的参数发生变化，但这种参数的变化难以在这种大数据预测模型中衡量。这种思想是由诺贝尔经济学奖得主卢卡斯提出，被称为“卢卡斯批判”。它最初用于批判新古典宏观经济学理论和凯恩斯主义理论，认为这两种理论的政策分析没有充分考虑到政策变动对人们预期的影响。

对利用大数据预测应秉持的科学态度

尽管利用大数据预测存在很大的艺术成分，但其中的科学色彩更浓，这就像莎士比亚四大悲剧之一《哈姆雷特》中一句台词所说的那样，“虽然这很疯狂，但依然有法可循”。大数据预测的结果通常存在误差，因为它假定事物的行为存在惯性和可预测性，但数据量级的增大和结构的多样化并不能彻底消除事物行动轨迹中的偶然性。不过，只要预测的准确率超过了毫无依据的猜测，那么即使预测的结果没有那么精确，也远胜于在黑暗中摸索，这是大数据预测价值的核心所在。进一步而言，对预测分析的态度将成为人类文化认知变迁的重要组成部分。

因此，在利用大数据进行预测时，要时刻铭记大数据本身只是一种工具，有用则用，没用就可以放弃，选择别的科学方法去尝试，不能让这种工具成为累赘，更不能将大数据作为最后一根救命稻草，事先假定利用大数据肯定能解决自己的问题，只是自己目前利用大数据的智慧还没有达到相应的高度。进一步而言，利用大数据预测未来本身是为了更好地把握未来，如果能主动利用大数据的强大优势去影响预测本身，就像管理学大师德鲁克所说的“预测未来的最好方式是创造未来”这句话所体现的境界一样，那么对大数据的驾驭能力，或许就可以趋近于我国历代大儒们所孜孜以求的那种“知行合一，止于至善”的高度。

作者单位：国家统计局科研所

参考文献

[1] 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？[J]. 经济研究，2015（12）：68-83.

[2]Carriero,A.,Clark,T.E.,and Marcellino, M. Real-time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility[EB/OL]. Federal Reserve Bank of Cleveland Working Paper(No.1227), 2012.

[3]Kopoin,A.,Moran,K.,and Paré,J.P. Forecasting Regional GDP with Factor Models: How Useful are National and International Data?[J]. Economics Letters, 2013, 121(2):267-270.

[4]Varian,H.R.Big Data: New Tricks for Econometrics[J].Journal of Economic Perspectives, 2014, 28(2):3-27.