大数据分析在宏观金融领域的文献综述
——基于中央银行的视角

肖筱林　王汉生*

摘　要：大数据和相关技术的出现，不仅形成了覆盖经济、社会运行的海量数据和大数据生态，而且也在影响和重塑公共政策的制定和实施。本文基于中央银行的视角，梳理了大数据背景下宏观金融领域中涉及大数据和相关技术的国际和国内文献，尤其是大数据分析在货币政策沟通、宏观经济预测和宏观审慎监管方面的相关研究。最后，本文深入分析了大数据分析在中国宏观金融领域的应用现状，并提出了切实可行的进一步应用建议。

关键词：大数据分析；宏观金融；中央银行；综述
中图分类号：F833
JEL 分类号：C10；E50；E58

一、引　言

大数据（Big Data）是近年来很受关注的一个领域。新的数字化工具的使用、信息系统的不断更新迭代以及数据采集技术的进步，这些因素共同导致了海量数据的出现。什么是大数据？不同学者、不同领域或有不同定义。Laney（2001）定义大数据为“高容量”“高速度”和“高多样性”的信息资产。Sagiroglu and Sinanc（2013）指出，高多样性强调大数据包含海量的结构化、半结构化和非结构化的数据（如图像、声音等）。大数据价值链依循一般数据分析的路径，包括数据生成、数据获取、数据存储和数据分析四个过程，其中数据分析包括结构化数据分析、文本数据分析、Web 数据分析、多媒体数据分析、社交网络数据分析和移动数据分析六个领域（Chen et al., 2014）。这些大数据的不同定义，其最大的共同点就是“大”（海量数据）。

而随之而来的大数据分析（Big Data Analytics），与近年来兴起的人工智能（Artificial Intelligence，简称 AI）技术既有一定的交叉（如机器学习、深度学习等），又形成互补，

* 肖筱林（通信作者），北京大学光华管理学院，E-mail: sylvia. xiao@gsm. pku. edu. cn；王汉生，北京大学光华管理学院，E-mail: hansheng@gsm. pku. edu. cn。肖筱林感谢国家自然科学基金面上项目（72073006）对本文研究的资助，也感谢陈心诺在项目初期的助研工作。作者感谢匿名审稿人和编辑部的宝贵意见，当然文责自负。

共同推动了大数据和相关技术在经济社会生活中的推广应用。而伴随着大数据、人工智能、5G、物联网等新技术的崛起以及在经济社会生活中的推广运用，经济生活和市场交换向着更加数字化、自动化、网络化及智能化的方向发展，形成了覆盖经济、社会运行各个方面的大数据生态。这一生态也在影响和重塑公共政策制定和实施的整个过程。

我们知道，中央银行（简称“央行”）的主要职能，是在一国货币政策操作框架下，通过常规货币政策工具（如公开市场业务、再贴现率和法定存款准备金率），或者非常规货币政策工具[1]，来控制利率和货币供给，以实现充分就业和价格稳定等货币政策的终极目标。在货币政策操作的事前、事中和事后等不同阶段，央行的日常工作包括：收集大量的数据，进行定期的数据分析、宏观经济预测和经济周期分析等；定期发布货币政策报告，并与公众进行沟通（传统的新闻发布会加上广泛使用的社交媒体，如国外的 Twitter、Facebook，国内的微博、微信等）；以及基于大量金融数据进行微观金融监管和宏观审慎监管等。特别要指出的是，2008 年国际金融危机以后，各国央行普遍更加注重宏观审慎监管，需要通过大数据分析密切关注影子银行、系统重要性金融机构和房地产市场等特定金融市场的实时动态。因此，基于大数据的时代背景，从央行的视角来看，我们想要通过梳理文献研究如下问题：大数据和相关技术的出现对央行的数据收集和分析，尤其是宏观金融领域的相关研究和分析，以及给宏观金融的哪些具体领域带来了新的变化？新的颗粒化的微观金融数据，伴随着新的分析工具，产生了哪些新的有趣的预测和分析结果？是否在货币政策沟通、宏观经济预测以及宏观审慎监管等领域产生了新的应用？与传统的数据和分析方法相比，大数据分析有哪些优势，是否也带来了新的问题、风险和挑战？

由于大数据及相关技术主要在近十几年兴起（Diebold，2012），而在宏观金融领域的应用和推广在大多数国家还处在早期阶段（包括中国），比较迅猛发展的时间段是从 2015 年到现在（IFC，2015，2019）。因此，本文基于中央银行的视角，尤其是基于上面提出的问题，主要分析大数据和相关技术的出现对货币政策沟通、宏观经济预测和宏观审慎监管的影响，对这些领域的前沿国际和国内文献进行综述。应该说，近年来，关于大数据和相关技术在经济金融领域的应用和研究不少，相关的综述也已出现，如沈艳等（2019）述评了文本数据分析在经济学和金融学中的应用。相比之下，本文的主要述评贡献如下：一是述评文献的研究视角和领域不同，即主要基于中央银行的视角，全面梳理大数据分析的前沿国际和国内文献在宏观金融领

1 非常规货币政策工具是指，2008 年国际金融危机以来，包括 2020 年新冠疫情全球性暴发后，利率降至零利率下限后，常规货币政策缺乏操作空间，因此各主要经济体央行相继推出的量化宽松（Quantitative Easing）、对银行准备金付息（Interest-Bearing Reserves）和前瞻性指引（Forward Guidance）等。

域的研究；二是涵盖的数据类型更多样，不仅涵盖文本数据，还包括运用电子支付、移动电话、传感器、卫星图像、在线价格和在线搜索等新型数据所进行的宏观金融领域的大数据分析；三是涉及的大数据分析方法更多样，除了文本大数据分析的方法外，还包括宏观经济实时预测和宏观审慎监管涉及的新型大数据分析方法，如采用混合频率的动态因子模型实时预测 GDP 增长，将新冠疫情期间产生的特定高频指标数据纳入贝叶斯动态因素模型以实时预测经济状况，以及使用数据可视化分析技术进行宏观审慎监管所需要的复杂、非线性和多维的金融大数据分析等；四是为大数据分析在中国宏观金融领域的应用推广和政策探讨提出了相应的建议。

另外，近些年来宏观经济学与微观实证进一步结合的研究范式兴起，这同各国央行对来自微观层面的金融大数据的重视可以说是一脉相承。具体来说，基于文献（Nakamura and Steinsson，2008，2010；Nakamura et al.，2018）的研究，中村惠美（Emi Nakamura）和合作者们构建和运用分散度更大、频率更高、时间跨度更长的微观数据库，进行了产品定价和菜单成本的研究，尤其注重强化对微观数据异质性的研究，从微观层面进一步佐证了价格黏性，并在此基础上深入分析了货币政策在短期的非中性效应、央行货币政策沟通的信息效应和财政支出的扩张效应等。大数据的出现为宏观经济预测和金融监管提供了实时的颗粒化的微观数据，减少了宏观经济预测的时滞性，因此能极大地助益宏观和微观实证结合的新研究范式。

在进入具体的文献述评之前，一些共性问题先要厘清：例如，基于央行视角的大数据，具体是指哪些数据，跟一般性的大数据来源有何不同？前文已经提及，从大数据分析涉及的内容来看，大数据包括结构化数据、文本数据、Web 数据、多媒体数据、社交网络数据和移动数据等（Chen et al.，2014）；也有文献将大数据的来源分为三大类，分别是社交媒体数据、传统商业体系数据，以及互联网相关的数据（包括机器产生的数据、智能手机产生的数据，甚至计算机的日志数据等）（IFC，2019）。而从央行的视角，大数据主要是指金融大数据（Financial Big Data），主要包含四类数据，即互联网相关的指标数据、商业数据集、金融市场指标数据和行政管理记录数据。应该说，央行能获取到的数据与大数据和相关技术的发展是密切相关的，金融大数据是大数据的一个子集。而大数据给央行带来的最大变化，应该是更加重视微观层面的数据集，这既因为海量且不断增长的涉及商业交易、金融市场和实际行政管理操作相关的数据的出现，也因为 2008 年国际金融危机之后，受到危机冲击最大的发达经济体普遍意识到收集和分析金融市场微观数据，并在此基础上进行宏观审慎监管的重要性。

余文结构安排如下：第二部分至第四部分是文献综述的主体，依次对大数据和相关技术在货币政策沟通、宏观经济预测和宏观审慎监管等方面的研究进行述评；第五部分是讨论和总结，分析大数据在中国宏观金融领域的应用现状，并提出切实可行的建议。

二、文本数据分析与货币政策沟通

为了向公众传达央行工作情况、保持政策公开透明和了解公众预期，各国央行经常会通过发布定期政策报告、公开演讲、发表声明和访谈等方式来向市场和公众传达货币政策信息。Blinder et al.（2008）将中央银行沟通（Central Bank Communication）定义为中央银行提供的关于货币政策目标、经济前景以及未来货币政策的信息。文本数据挖掘指的是通过信息技术从不同的文本资源中自动提取信息，将非结构化数据转为结构化数据，进而发现未知信息的过程（Hearst，2003）。常见的文本数据挖掘工作包括文本分类与聚类、信息提取、情绪分析、关键词提取、自然语言处理等（Bach et al.，2019）。而文本分析方法首先将原始文本库转化为数据矩阵，再从数据矩阵中提取信息，进而使用被提取出的信息来解释和预测某些现象（岑维等，2014）。在将文本转化为数据矩阵时，需要先将英文文本分解为单词或词组，然后将单词组成的高维矩阵进行降维处理，使其转化为数据矩阵。信息提取方法主要分为有监督学习和无监督学习两类，有监督机器学习使用有标记的样本并学习特征和标签值之间的关系（Kotsiantis et al.，2007），而无监督机器学习则对未标记的样本进行训练学习（Gentleman and Carey，2008）。

目前不少学者和央行经济学家已将文本分析技术用于货币政策沟通领域，以下从两个方面进行文献梳理：一是对中央银行的报告、采访和演讲进行文本大数据分析，量化央行货币政策沟通对金融市场和政策制定者的影响，以及从央行报告中推断央行立场；二是分析新闻媒体和公众对货币政策的评论，并剖析舆论情绪对金融市场和经济的影响。

1. 分析央行报告的影响

大部分文献通过对央行发布的政策报告、访谈和演讲等进行文本分析，研究央行发布的内容对市场利率的影响（Lucca and Trebbi，2009；Hendry and Madeley，2010）。Hubert and Labondance（2017）利用词典法调查了欧洲央行和联邦公开市场委员会在声明中表达的情绪对利率期限结构的影响，结果发现情绪冲击会在一年到两年内影响个人的利率预期，并且这种影响是非线性的。Oshima and Matsubayashi（2018）采用隐含狄利克雷分配模型的方法，集中分析了 2013—2017 年日本央行的沟通对金融市场的影响，发现在量化宽松后期日本企业资金拨备措施的信息比货币政策信息对市场影响更大。2008 年国际金融危机以后，各国央行重视并定期发布金融稳定报告，一些学者也开始用文本分析方法分析这些定期发布的报告对金融稳定的影响。Born et al.（2014）指出，金融稳定报告对股票市场收益具有显著且长期的影响，并有助于降低市场波动性；但是演讲和采访对市场回报率影响不大，只有在

金融危机期间才会降低金融市场波动。Correa et al.（2021）构建了一个专门针对金融稳定的文本分析词典，利用该词典捕获各国央行发布的金融稳定报告中的情绪，并测试了金融稳定指数对金融周期和银行危机的预测能力。

除了研究央行报告、演讲和访谈对经济和金融市场的影响，一些研究也试图通过这些文本来推断央行的立场。Picault and Renault（2017）对欧洲央行新闻发布会的句子进行分类，构建了一个领域专用词典，衡量了欧洲央行货币政策的立场（鸽派、中立派、鹰派）和欧元区经济状况（积极、中立、消极）。Shapiro and Wilson（2019）使用词典法分析了 1976—2013 年之间的联邦公开市场委员会文本、会议记录和发言，直接估计了美联储的损失函数，进而评估其货币政策目标。Dieijen and Lumsdaine（2019）将动态主题模型和隐含狄利克雷分配模型应用于美联储董事会成员 20 年来的演讲，分析了美联储的双重目标（即最大化就业和保持价格稳定）的权重随时间的推移而发生的变化，发现在金融危机后美联储对金融稳定日益重视，本质上是在双重目标基础上又增加了第三个目标。

此外，还有部分学者用文本分析方法测量了中央银行沟通的篇幅、词汇的难易程度以及可读性等。Binette and Tchebotarev（2019）指出加拿大央行的货币政策报告的复杂程度要高于加拿大人的平均理解能力。Mathur and Sengupta（2019）分析了印度央行 1998—2017 年的货币政策声明，测量了货币政策声明随时间推移的长度和可读性变化，并研究了央行沟通对股票市场的影响。他们发现，总体而言印度央行在货币政策沟通方面的语言较为复杂，但是近年来货币政策声明的可读性显著提高；而且，较低的可读性与股票市场上较高的交易量和较高的收益波动率相关，但该影响并不持久。

2. 分析媒体和舆论情绪

除了对央行发布的关于货币政策的报告进行文本分析，一些文献也使用文本分析方法研究大众对货币政策的情绪反应，进而分析舆论情绪对市场的影响。Meinusch and Tillmann（2015）收集了 Twitter 上关于货币政策讨论的数据，分析了人们对量化宽松的情绪如何显著地影响资产价格和经济发展。Azar and Lo（2016）提取了 2007—2014 年 Twitter 中有关“联邦公开市场委员会”和“美联储”的推文，通过对每条推文评分来识别文本中的情绪，发现可以通过 Twitter 来预测联邦公开市场委员会会议前后股市的反应。Bianchi et al.（2023）收集了美国前总统特朗普有关货币政策的推文，这些推文批评过去两年来美联储的货币政策操作，并持续倡议美联储降低利率。作者发现，持续的政治压力严重影响了市场参与者对美联储独立性的信任。Lüdering and Tillmann（2020）集中研究了 2013 年的“削减恐慌”（Taper Tantrum）时期，使用隐含狄利克雷分配模型从 Twitter 的讨论中提取出多个主题，然后在 VAR 框架中对所选主题的频率进行建模，进而估计冲击对资产价格的影响。结果表明，社交媒体中关于货币政策的讨论对资产价格很重要，社交媒体可以反映总体市场观点以及市场预期。

因为公众的预期最终会通过工资、储蓄、投资和消费决策等影响价格水平，所以媒体在传播货币政策信息方面起到了非常重要的作用，一些经济学者也在分析新闻媒体与央行沟通之间的相互作用。Berger et al.（2011）发现当通货膨胀率较高或者货币政策出乎公众意料时，新闻媒体对央行的报告将会较为负面。Lee et al.（2019）收集了 2005—2017 年韩国央行货币政策委员会会议前后的新闻文章、分析师报告和货币政策委员会会议记录，使用词典法对文本中的情绪进行分类并量化了“货币政策惊奇”（Monetary Policy Surprise，即出乎公众意料的货币政策），进而估计它对资产价格和长期利率的影响。Rybinski（2019）利用波兰央行在 20 年内的政策报告与相关的媒体新闻，分别基于情感词典和 Wordscores 模型来提取文本信息，并提出了有监督机器学习框架来预测经济变量，结果发现，基于 Wordscores 模型的指标预测能力要高于基于情感词典的情绪分析。Bennani（2020）利用媒体对美联储主席的报道来量化其信心和乐观程度，以构建过度自信指标（Overconfidence Indicator），发现美联储主席的过度自信与投资者情绪高涨显著相关。

三、大数据分析与宏观经济预测

传统统计方法为了保证数据的准确性，往往依靠定期的抽样调查或全面普查的方法来测量各种经济指标，具有较大的时滞性。相较于传统方法，大数据技术可以更加高频、及时和快速地收集数据并进行分析，甚至能做到“实时预测”（Nowcasting）。“实时预测”一词一开始被用于气象学，但如今该技术在预测和分析价格水平、通货膨胀率、失业率和 GDP 等央行关注的宏观经济指标上也有重要的应用。金融市场数据、电子支付数据、移动电话数据、传感器数据、卫星图像数据、在线价格数据、在线搜索数据、文本数据和社交媒体数据等，常被用于宏观经济实时预测（Buono et al.，2017）。

站在政策制定者（如央行）的角度，基于大数据技术驱动的实时预测和分析，一方面可以更及时地收集数据，涵盖更多的样本数量，从而成为传统宏观经济数据的较好补充，另一方面也可以给央行提供新的宏观经济预测视角，包括对 GDP、失业率和个人消费等的预测，从而更好地进行经济周期分析。以下我们主要聚焦大数据分析在宏观经济预测方面的应用，包括对 GDP、失业率和价格水平的实时预测，并分析外汇市场预测的相关内容（当中部分文献不属于实时预测）。

1. GDP 预测

国内生产总值（GDP）是衡量各国经济表现的最广泛使用的指标，准确且迅速地估算 GDP 对企业、个人和政府有着重要意义，而大数据技术的出现实现了这一点。目前，常见的被用于实时预测 GDP 的大数据类型包括卫星夜灯亮度数据、电子

支付数据和媒体数据等。

一些研究表明，夜间灯光数据与国家国内生产总值的数据正相关，因此可以从卫星图像中推断区域经济发展水平（Elvidge et al.，1997；Doll et al.，2006）。Henderson et al.（2011）指出国家 GDP 增长速度和夜间灯光数据存在着强烈的正相关关系，在缺乏准确官方统计数据的国家，夜间灯光数据更适合用于衡量 GDP 增长水平。Henderson et al.（2012）进一步指出，夜间灯光数据还可以衡量次国家和超国家地区的增长。

近些年，相关研究也开始使用电子支付数据来对宏观经济进行实时预测。Galbraith and Tkacz（2015）使用加拿大信用卡、借记卡以及支票交易数据来实时预测加拿大的经济情况，GDP 的实时预测效果提高了 65%。Aprigliano et al.（2017）基于混合频率动态因子模型，使用标准经济周期指标（工业产值、通货膨胀、股票市场指数、制造业指数等）以及支付系统数据（支票、借记卡等）来实时预测意大利的 GDP。

除了卫星数据和电子支付数据，也有研究使用媒体数据来预测 GDP 增长速度。Thorsrud（2016）使用隐含狄利克雷分配模型将商业报纸中的非结构化文本信息分解为每日新闻主题，并使用混合频率的动态因子模型来实时预测 GDP 季度增长。Bok et al.（2018）利用动态因子模型构建了一个自动化平台来处理实时数据，介绍纽约联储银行的实时预测模型，并用该方法对 GDP 增速进行预测。Indaco（2020）使用 2012—2013 年在 Twitter 上所有共享地理位置的推文来估算国家的 GDP，发现推文可以解释 78% 的跨国变化。

值得注意的是，在新冠疫情期间，大数据技术也发挥了巨大作用。疫情期间保持社交距离和减少接触所导致的信用卡交易、工资数据或流动性统计信息等高频指标数据，使得人们可以及时读取经济状况。Antolin-Diaz et al.（2021）就提出一种将此类指标纳入贝叶斯动态因素模型的方法，并量化这些高频数据对实时预测的贡献。

2. 失业率预测

在当前的数字化时代，求职者和潜在雇主越来越多地将互联网作为信息来源，因此会在互联网中留下他们的搜索记录。早在十几年前，就有很多学者研究了搜索引擎数据与失业率的关系（Ettredge et al.，2005；Askitas and Zimmermann，2009；Choi and Varian，2010）。近些年来，D'Amuri and Marcucci（2010）分析了采用不同领先指标的模型，比较了它们的样本外预测效果，结果发现，包含互联网求职搜索指数的模型在预测失业率方面优于传统模型。因此，他们建议使用互联网求职搜索指数（基于谷歌搜索作为预测美国每月失业率的最佳领先指标）。类似地，Pavlicek and Kristoufek（2015）指出谷歌搜索增强了捷克和匈牙利失业率的现时预测模型的准确性，而在实时预测波兰和斯洛伐克的失业率方面并没有那么好的效果。Chadwick and Sengul（2015）使用 2005 年 1 月至 2011 年 10 月的 Google 搜索数据，

采用线性回归模型和贝叶斯模型平均法改善了土耳其每月非农业失业率的实时预报表现。结果显示，Google 搜索查询数据能够成功预测样本内外的非农业失业率，以及包含 Google 搜索查询数据的模型比自回归基准模型的统计性能更好。

由于工人失业后社交和通信行为会发生显著变化，Toole et al.（2015）指出还可以使用手机通话数据来识别失业人群。他们使用贝叶斯分类模型，通过观察工厂倒闭后不同工人的手机通话行为的变化，来识别出哪些个体受到了工厂倒闭的影响，进而可以改善对宏观失业率的预测。然而，使用手机通话记录数据会导致隐私和道德问题，这也是大数据分析导致伦理问题的典型例子。

3. 价格水平预测

通货膨胀代表了价格总水平的变化，影响着家庭、投资者和政府的决策，是经济中最为重要的指标之一。在 2008 年，MIT 创办了“十亿价格项目”（the Billion Prices Project），该项目使用每天从全球数百家在线零售商处收集的价格来计算实时通胀指标，能提高预测 CPI 指数的精确性。Cavallo and Rigobon（2016）在 Cavallo（2013）的基础上[2]，介绍了“十亿价格项目”价格信息数据的采集方法，并讲述了如何使用该项目的在线价格来构建多个国家的每日价格指数。Aparicio and Bertolotto（2020）使用 PriceStats 提供的在线价格指数对德国、英国和美国等多个国家的 CPI 进行预测，结果发现在线价格指数能提前一个多月预测官方通胀趋势的变化。

除了在线价格数据以外，互联网搜索数据也常常被用于实时预测。Guzman（2011）指出 Google 通胀搜索指数比其他 37 个通胀预期指标的预测误差都要小。此外，也有学者提出可以从消费电子扫描数据创建价格指数，扫描价格变化可以揭示零售商和行业特定层面的信息。新西兰统计局从 2014 年 9 月开始，将零售交易数据或“扫描数据”纳入消费者价格指数统计中，来衡量消费类电子产品的价格变化。但是，随着越来越多的产品直接在网上销售，“扫描数据”在预测通货膨胀方面变得不如互联网在线价格数据那么有效。

4. 外汇市场预测

早在几十年前经济学家们就尝试对汇率进行预测，但很多研究者认为，任何汇率变化都遵循有效市场假说（Fama，1970，1991），所以预测汇率是不可行的。然而，行为金融学对有效市场假说提出了质疑，认为市场收益率在相当大程度上受到投资者情绪的影响（Nofsinger，2005）。因此，在大数据时代，不少研究试图从社交媒体、宏观新闻和互联网搜索指数中挖掘舆论的情绪，进而通过这些情绪来对外汇市场的走向进行预测。

---

2 Cavallo（2013）收集了 the Billion Prices Project 网站中的商品价格数据，指出阿根廷的通胀率比官方估计的通胀率高出近三倍。

新闻和社交媒体数据经常被用于外汇市场预测。Evans and Lyons（2008）尝试研究宏观新闻如何通过订单流影响汇率，结果显示，宏观新闻可以解释汇率波动的30%以上。Chatrath et al.（2014）发现9%~15%的汇率变动直接受到美国发布的新闻的影响。Ozturk and Ciftci（2014）发现Twitter情绪与USD/TRY汇率之间存在显著关系，因此利用Twitter数据可以改进对汇率变动的预测。Semiromi et al.（2020）使用文本数据分析方法分析了新闻事件如何为交易决策提供有价值的信息。他们从新闻中提取信息并建立了一个外汇市场情绪词典，发现新闻发布后汇率预测的准确性明显要高于其他时期。

在使用互联网搜索数据方面，Smith（2012）指出GARCH（1,1）模型的条件方差不能预测汇率波动，相反，以“经济危机”“金融危机”和“衰退”为关键词的谷歌搜索量的预测能力超过了GARCH（1,1）模型，因此谷歌搜索量的变化在预测外汇市场波动方面有着重要作用。Bulut（2018）将购买力平价模型、弹性价格货币模型和利率平价模型这三种传统汇率模型与基于谷歌搜索量的汇率预测相比较，发现在样本外预测中，谷歌搜索数据在预测名义汇率变化方向方面比传统汇率模型更好。

四、大数据分析与宏观审慎监管

2008年国际金融危机，以次级抵押贷款机构破产、股市剧烈动荡、系统性银行危机席卷为特征，从美国蔓延到全世界，最终造成了全球范围内的巨大损失。虽然学术界对这次国际金融危机的原因有诸多不同解释，但是达成的共识之一是，监管部门过分推崇自由主义思想，缺乏对金融机构的宏观审慎监管，是美国爆发金融危机并席卷全球的重要原因。另外，随着大数据技术的发展，数据存储成本降低、计算机处理能力提高和算法进步，我们使用机器学习技术收集和识别金融数据的能力也发生了变化。这种转变一方面带来了更快、更好、更便宜的金融产品和服务，另一方面也可能扰乱金融格局，增加金融不稳定性，给金融监管带来新的挑战。鉴于金融市场对宏观经济的巨大影响和金融危机的惨痛教训，各国普遍强化了央行在宏观审慎监管方面的职责，包括提出新的宏观审慎监管框架，或者从立法的角度明确央行这一新增的职责。

以下相关文献从宽泛意义上都属于大数据技术应用于宏观审慎监管方面的分析，但可以进一步细分为宏观审慎监管、金融危机预警和股票市场预测这三个具体方面。

1. 宏观审慎监管

随着金融业混业的趋势越发明显，也随着数据的准确性和可靠性在宏观审慎监管中发挥着越来越重要的作用，各国央行要求各类金融机构提交更多的数据。2010

年美国颁布的《多德－弗兰克法案》就要求大型货币基金每月都必须提交交易数据，授权美国财政部金融研究办公室收集金融机构的头寸数据。美联储每月都在收集由包括抵押贷款、房屋净值产品和信用卡交易在内的个人贷款数据，如今，这些大数据已被用于分析各种零售金融产品。为了方便收集金融市场的微观数据，美国还大力推进“法人实体识别编码”（Legal Entity Identifier，简称 LEI）系统的建设，参与金融交易的机构必须严格按照标准及时提交相关信息。但是，Buch（2017）指出，金融危机发生 10 年后，各国在一些重要的数据收集或者监管上依然存在短板，例如收集商业地产价格的信息、衍生金融产品市场的信息，以及对跨国金融控股公司的监管等。以影子银行为代表的金融机构反过来也可以使用大数据和机器学习来规避金融监管（Jagtiani et al.，2018）。

除了需要创建新的数据源和分析方法之外，金融危机还表明需要有更大的能力来整合和理解大量、动态和异构的金融数据，而数据可视化分析就是这样的一种技术。Cook and Thomas（2005）将可视化分析定义为由交互式可视化界面促进的分析推理科学，它使用统计图形、图表、信息图表、动画等工具对海量数据进行分析。在 2008 年金融危机前已有学者使用可视化技术对金融机构进行分析，例如 Soramäki et al.（2007）对美国商业银行支付网络的拓扑结构进行了分析，指出“9·11 事件”显著改变了网络的拓扑结构，网络中的节点数量减少，节点之间的平均路径长度显著增加。2008 年国际金融危机后，Heijmans et al.（2016）使用荷兰货币市场交易数据，以动态、非交互的方式对荷兰银行间市场的模型进行了可视化分析，展示了如何对不同时间的同一市场或同一时间的不同市场进行比较。Flood et al.（2016）指出，由于宏观审慎监管日益受到大量动态和异构数据的支配，可视化技术在分析复杂、非线性和多维的金融数据中发挥着重要作用。

另外，特别值得关注的是，当前全球范围内超过 90% 的央行正在研发和试点的央行数字货币（Central Bank Digital Currency，简称 CBDC），未来可能对央行的数据收集和宏观审慎监管带来的改变。CBDC 即数字化法币，虽然各国研发、试点或正在发行 CBDC 的目的各不相同，但最常见的共同目的，一是替代现实世界中使用逐渐减少甚至趋于消亡的现金，二是应对比特币等私人加密货币带来的冲击（Kosse and Mattei，2022；肖筱林等，2023）。中国的 CBDC 就是数字人民币（之前称 DC/EP，现已改名为 e-CNY），自 2019 年末以来一直在进行密集试点，自 2022 年末起已正式加入中国 M0 的相关统计。世界上其他国家，诸如非洲大国尼日利亚和一些加勒比海的小国已经正式发行 CBDC（来源：Atlantic Council 网站）。目前学术界、政策制定者和国际组织对 CBDC 的相关研究聚焦的维度各不相同（Auer et al.，2022；肖筱林等，2023），但跟大数据相关的一个维度是，CBDC 作为一种数字化形态的法币，极有可能助力央行方便地收集关于商业银行、非银行金融机构乃至整个金融体系的金融大数据，并更好地实施宏观审慎监管（Keister and Monnet，2022）。中国数字人民

币的白皮书中也指出，“人民银行还为数字人民币建立大数据分析及风险监测预警框架，以提高数字人民币管理的预见性、精准性和有效性”（中国人民银行数字人民币研发工作组，2021）。CBDC 是一种崭新的数字货币，在全球范围内正式发行的国家或地区还不多，但其对金融大数据和宏观审慎监管可能带来的巨大改变，值得持续关注。

2. 金融危机预警

金融危机会给金融系统和实体经济造成重大损失，建立一套可靠的危机预警系统对决策者来说非常重要。传统计量方法无法解释变量之间复杂的交互作用和非线性，也很难处理样本量较大的数据集，相比之下，支持向量机和随机森林等机器学习算法显得更加有效。

Ravisankar et al.（2011）使用西班牙、土耳其、英国和美国银行的数据集，介绍了三种神经网络架构来预测银行破产，包括数据分组处理方法、反向传播神经网络方法和模糊自适应共振理论，结果发现数据分组处理方法的性能优于其他技术。Erdogan（2013）根据土耳其银行系统数据，使用支持向量机分析银行财务比率，研究表明，当参数选取得当时，模型对试验数据集的误差为 0.05，灵敏度为 0.92，这说明高斯核函数支持向量机能够从金融数据中提取有用的信息。Joy et al.（2017）研究了 1970—2010 年 36 个发达经济体银行业在金融危机爆发前的经济和金融状况，使用分类和回归树方法及其随机森林扩展来预测危机，发现银行业的低净利息倒转或收益率曲线变平是银行业危机的短期前兆。为了提高银行危机预警模型的能力，Ristolainen（2018）使用区域数据集构建了一个基于人工神经网络的预警系统，结果表明，使用区域数据集进行估计可以大大提高预测准确性，基于神经网络的预警系统模型的预测效果明显优于 Logit 回归等常用模型。Bluwstein et al.（2023）使用 1870—2016 年 17 个国家的金融数据来预测金融危机，将信贷增长和收益率曲线的斜率作为主要预测指标，结果表明机器学习模型在预测金融危机方面优于 Logit 模型。类似地，在不远的将来，随着 CBDC 在更多国家和地区的发行，可能也会更有利于各国央行收集金融危机的预警性指标，以及应对金融危机（Keister and Monnet，2022）。

3. 股票市场预测

股票市场受许多因素的影响，例如政治事件、总体经济状况和交易员的预期等。预测股票市场走势一直是投资者和研究者最广泛研究但也是最具挑战性的问题之一。行为金融学理论充分证明，投资者的情绪会影响其行为，进而对金融市场产生影响（Baker and Wurgler，2007；Mian and Sankaraguruswamy，2012）。目前大部分国际文献通过社交媒体数据、新闻数据和互联网搜索数据来分析投资者的情绪与关注度，进而对金融市场进行分析和预测。

国外学者多使用 Twitter 和 Facebook 这两个社交软件作为预测股市走向的数据来源（Bollen et al., 2011；Vu et al., 2012；Karabulut, 2013）。Yasir et al.（2020）为英国、土耳其、中国、墨西哥等国家或地区构建了一个预测利率的深度学习模型，并将英国脱欧、加沙攻击、墨西哥大选、美国大选等事件的 Twitter 情绪数据输入，当事件情绪被考虑时，深度学习模型的误差会显著减小。金融新闻文本是另一种经常用于股市预测的数据源（Mahajan et al., 2008；Lupiani-Ruiz et al., 2011；Vargas et al., 2017）。Koppel and Shtrimberg（2006）将新闻报道标记成正面新闻与负面新闻，然后使用支持向量机方法提出了一种基于新闻文章的股票预测模型，发现该模型准确性高达 70%。互联网搜索数据反映着大众对某个事件的关注度，常常被用于流行病（Ginsberg et al., 2009；Desai et al., 2012）、房地产市场（Wu and Brynjolfsson, 2015）和消费者信心预测（Choi and Varian, 2012）等诸多领域中，近些年来也开始用于股票市场预测（Bank et al., 2011；Da et al., 2011）。Preis et al.（2013）发现当金融市场有低价卖出的趋势时，投资者会更多地收集有关金融市场状况的信息，进而导致谷歌搜索量增加。Kim et al.（2019）得出了不一样的研究结果：谷歌搜索数据不能预测股票收益率，但是可以用于预测股票交易量和波动性。但是他们也指出，可能是因为样本量规模较小，才导致谷歌搜索数据不能预测股票收益率。

五、大数据分析在中国宏观金融领域的应用和建议

行文至此，我们已对大数据分析在宏观金融领域，尤其是在货币政策沟通、宏观经济预测和宏观审慎监管三个主要方面的国际文献进行了述评。总体而言，近年来大数据分析在宏观金融领域的应用发展迅速，相关研究不断涌现。相比之下，国内近年来对宏观经济大数据的关注也在不断升温，不少学者也进行了相关研究，但同国际研究相比，还有提升空间。接下来，我们先梳理国内相关文献和研究，进而对大数据分析在中国宏观金融领域的应用提出建议。

1．大数据分析在中国宏观金融领域的应用现状

使用文本大数据对中国人民银行的政策沟通效果进行分析的文献较少。McMahon et al.（2018）指出，中国人民银行的政策沟通主要通过货币政策执行报告、讲话和新闻发布、货币政策委员会会议新闻稿和公开市场操作公告四个渠道进行。该文指出，虽然中国人民银行有较为丰富的政策沟通方式，但由于中文语言结构的特点，尤其是中国人民银行的政策沟通语言往往经过反复推敲等原因，使用文本大数据技术对中国人民银行的货币政策进行分析遇到了较大的障碍。姜富伟等（2021）发现，货币政策报告的文本情绪的改善会引起股票市场价格上升，报告文本相似度的增加会引起股票市场波动性下降，报告可读性对股票市场的波动性影响不显著。

对价格水平、GDP、失业率和汇率的传统统计较为准确，但是往往具有较大的时滞性，因此近些年一些国内学者开始使用大数据对中国宏观经济指标进行预测。在预测中国的价格水平和通货膨胀方面，大部分文献都使用搜索引擎数据（张崇等，2012；孙毅等，2014；徐映梅和高一铭，2017）和购物网站数据（袁铭，2015；韩胜娟和张敏，2017）来实现对价格水平的实时预测。曾嘉和李洁（2020）使用电力大数据来对 GDP 进行精准预测，提高了国内生产总值统计的时效性。彭赓等（2013）、王勇和董恒新（2017）分别使用谷歌搜索数据和百度搜索数据来实时预测中国的失业率。王轩和杨海珍（2017）通过构造的互联网搜索指数来改进传统时间序列模型，对人民币兑美元汇率进行了较为有效的预测。

至于大数据分析在宏观审慎监管方面的应用，在宏观审慎监管、金融危机预警方面的文献较少，运用于股票市场分析和预测的文献相对较多。王达（2015）分析了美国收集金融微观数据的方法，提议中国可以参考美国的做法，制定金融数据报送机制，并为金融机构和金融产品进行数字化编码。苗子清和张卓群（2020）从金融机构、股票市场、宏观经济、影子银行、房地产市场和政策干预等九个维度，构建了中国系统性金融风险的指标体系。在运用大数据技术对股票市场进行分析方面，相关研究则较为丰富。国内研究者最常使用微博数据（程琬芸和林杰，2013；张书煜等，2015；陈云松和严飞，2017；孙明璇和李莉莉，2020）、网络论坛数据（岑维等，2014；杨晓兰等，2016；石勇等，2017）和搜索引擎数据（宋双杰等，2011；俞庆进和张兵，2012；刘锋等，2014）来分析股票市场情绪和投资者关注度，进而对金融市场走向进行预测。

最后，从实践来看，中国人民银行从 2010 年起，就开始了金融统计大数据方面的探索，包括建立理财与资金信托统计、服务交叉性金融产品监测，以及建立标准存贷款统计，服务利率监测（阮健弘，2021）。更全面的工作框架是 2018 年国务院办公厅印发的《关于全面推进金融业综合统计工作的意见》。在此框架的指引下，由中国人民银行牵头建立的国家金融基础数据库投产使用，统一管理的金融统计数据采集系统和大数据智能分析平台已完成部署，并与中国人民银行各分支机构及 4 600 余家金融机构互联互通，实现数据智能化、整体化的采集和使用。国家金融基础数据库和主体信息库的建设和使用，不仅方便进行微观审慎监管，还能通过数据关联实现对金融机构、金融市场、交易对手和金融活动的全方位刻画，可用于金融风险传染研究，服务系统性风险防控，从而助力宏观审慎监管。

2. 大数据分析在中国宏观金融领域应用的建议

关于大数据分析在宏观金融领域的研究，对比国内外文献，可以发现，国内文献主要集中在股票市场分析，而用于货币政策沟通、宏观审慎监管、金融危机预测等方面的研究较少。而这些也正是机器学习、深度学习、文本数据分析等大数据分析技术能够充分发挥作用的领域。从中国人民银行牵头的金融统计大数据方面的软

硬件建设进程来看，宏观金融领域的颗粒化微观数据的收集和整合，以及相关数据库的建设进展很快。但是，从国内研究和实践的现状来看，还有不少可以提升的空间。为此，我们提出如下三方面的建议：

第一，强化宏观金融领域的舆情分析，关注和引导公众预期。在当前社交媒体当道，资讯迅猛传播的大数据时代，央行也要与时俱进，考虑使用最新的大数据分析技术和方法，对涉及货币政策和金融市场相关的舆论和公众情绪进行及时的“捕捉”和处理。另外，及时“捕捉”和处理宏观金融领域重要的舆情，其实也是在关注和引导公众预期，是货币政策沟通和传导中的重要环节，也属于前瞻性指引（Forward Guidance）的范畴。前瞻性指引是 2008 年国际金融危机后，发达经济体普遍实行零利率下限，为此各国央行不仅加大了常规沟通的力度，还通过低利率承诺引导公众预期，从而成为中央银行沟通和预期管理的重要方式。中国虽然没有零利率下限的困扰，但可以借鉴发达经济体央行通过各种社交媒体发布和传播货币政策资讯的做法，进而可以通过大数据来进行货币政策沟通事前、事中和事后的相关分析。例如，中国人民银行在 2013 年开设微博账号，2019 年开设微信公众号，通过社交媒体进行货币政策沟通已经有了一定的经验，也为相关的大数据分析提供了基础，但目前相关研究几乎是空白。这方面值得持续关注，今后可以通过文本数据分析或者最新的大数据分析方法来开展相关研究。

第二，对宏观经济重要指标进行实时预测，与传统统计形成良好互补。利用大数据技术进行实时预测（Nowcasting），在发达经济体央行的相关研究中已经大量使用。前文提及，金融市场数据、电子支付数据、移动电话数据、传感器数据、卫星图像数据、在线价格数据、在线搜索数据、文本数据、社交媒体数据等，都可用于宏观经济实时预测，并能与传统统计形成良好互补。截至 2023 年 6 月的数据，中国互联网络信息中心（CNNIC）发布的报告显示，中国互联网网民规模已达 10.79 亿人，互联网普及率达 76.4%。再加上中国目前领先全球的移动支付产业，位居世界前列的数字经济，这些都给大数据分析用于宏观经济指标的实时预测提供了坚实的基础。尤其是与电子支付、移动支付、互联网和社交媒体相关的数据，都是中国大数据的优势所在。目前虽然已有一些国内研究（见前文），但还有更多的研究值得进一步推进。

第三，依托数字人民币将来的发行，进一步完善金融统计大数据。中国人民银行前行长易纲在《建设现代中央银行制度》一文中提到，中国人民银行应该“统筹规划金融业综合统计、反洗钱以及金融市场登记托管、清算结算、支付、征信等金融基础设施，推动境内外各类金融基础设施互联互通，构建适应金融双向开放的金融基础设施管理体系”（易纲，2022），这其实是对央行在金融统计大数据方面应起的作用进行的纲领性概括。前文也提及了国内在这方面的最新进展。进一步联想到，近年来中国密集试点和推广中的 CBDC，即数字人民币，具体运营采用双层架构，即

中国人民银行作为央行向处在第二层的商业银行和其他指定运营机构发行数字人民币，目前十家指定运营的商业银行不仅负责数字人民币的兑换和流通服务，还要承担反洗钱、反恐怖融资等义务，在收集客户重要信息的同时，也肩负着保护客户商业机密、个人隐私、个人信息和交易记录的重任。具体来说，个人和企业通过一家作为指定运营机构的商业银行开设数字人民币账户，相关信息和交易记录只能被该指定银行获取，而不能被同处于第二层的其他运营机构所获取；而中国人民银行作为第一层，则能收集到来自全部指定运营机构收集的个人信息和交易记录，最终能够形成数字人民币的大数据中心。如前所述，数字人民币的大数据中心在中国人民银行发布的白皮书中也已明确提及（中国人民银行数字人民币研发工作组，2021）。总之，在不远的将来，数字人民币的全面发行以及所依托的双层运营架构，将给中国建设中的金融统计大数据提供更全的数据，也将更便利数字人民币使用的大数据分析以及任何使用数字人民币的相关金融交易的分析。

参考文献

[1] 岑维，李士好，童娜琼，2014. 投资者关注度对股票收益与风险的影响——基于深市“互动易”平台数据的实证研究[J]. 证券市场导报，(7)：40－47.

[2] 陈云松，严飞，2017. 网络舆情是否影响股市行情？基于新浪微博大数据的 ARDL 模型边限分析[J]. 社会，37(2)：51－73.

[3] 程琬芸，林杰，2013. 社交媒体的投资者涨跌情绪与证券市场指数[J]. 管理科学，26(5)：111－119.

[4] 韩胜娟，张敏，2017. 大数据时代官方价格指数与非官方价格指数的融合——基于 aSPI 与 CPI、RPI 比较的视角[J]. 价格理论与实践，(4)：84－87.

[5] 姜富伟，胡逸驰，黄楠，2021. 央行货币政策报告文本信息、宏观经济与股票市场[J]. 金融研究，(6)：95－113.

[6] 刘锋，叶强，李一军，2014. 媒体关注与投资者关注对股票收益的交互作用：基于中国金融股的实证研究[J]. 管理科学学报，17(1)：72－85.

[7] 苗子清，张卓群，2020. 基于大数据方法的中国系统性金融风险预警研究[J]. 经济论坛，(12)：5－17.

[8] 彭赓，苏亚军，李娜，2013. 失业率预测研究——基于网络搜索数据及改进的逐步回归模型[J]. 现代管理科学，(12)：40－43.

[9] 阮健弘，2021. 金融统计大数据服务宏观调控的探索与实践[EB/OL]. (2021－03－03)[2023－08－01]. http://www.pbc.gov.cn/redianzhuanti/118742/4122386/4122510/4199818/index.html.

[10] 沈艳，陈赟，黄卓，2019. 文本大数据分析在经济学和金融学中的应用：一个文献综述[J]. 经济学(季刊)，18(4)：1153－1186.

[11] 石勇，唐静，郭琨，2017. 社交媒体投资者关注、投资者情绪对中国股票市场的影响[J]. 中央财经大学学报，(7)：45－53.

[12] 宋双杰，曹晖，杨坤，2011. 投资者关注与 IPO 异象——来自网络搜索量的经验证据[J]. 经济研究，46(S1)：145－155.

[13] 孙明璇，李莉莉，2020. 基于数据挖掘的投资者情绪对股市波动影响研究[J]. 燕山大学学报(哲学社会科学版)，21(1)：68－77.

[14] 孙毅，吕本富，陈航，薛添，2014. 大数据视角的通胀预期测度与应用研究[J]. 管理世界，(4)：171－172.

[15] 王达，2015. 宏观审慎监管的大数据方法：背景、原理及美国的实践[J]. 国际金融研究，(9)：55－65.

[16] 王轩，杨海珍，2017. 基于互联网搜索指数的多因素集成下人民币汇率预测[J]. 系统工程学报，32(3)：360－369.

[17] 王勇，董恒新，2017. 大数据背景下中国季度失业率的预测研究——基于网络搜索数据的分析[J]. 系统科学与数学，37(2)：460 - 472.
[18] 肖筱林，黄益平，龚六堂，2023. 数字货币研究综述：2009 - 2022[R]. 工作论文，北京：北京大学.
[19] 徐映梅，高一铭，2017. 基于互联网大数据的 CPI 舆情指数构建与应用——以百度指数为例[J]. 数量经济技术经济研究，34(1)：94 - 112.
[20] 杨晓兰，沈翰彬，祝宇，2016. 本地偏好、投资者情绪与股票收益率：来自网络论坛的经验证据[J]. 金融研究，(12)：143 - 158.
[21] 易纲，2022. 建设现代中央银行制度[J]. 中国金融，(24)：9 - 11.
[22] 俞庆进，张兵，2012. 投资者有限关注与股票收益——以百度指数作为关注度的一项实证研究[J]. 金融研究，(8)：152 - 165.
[23] 袁铭，2015. 基于网购搜索量的 CPI 及时预测模型[J]. 统计与信息论坛，30(4)：20 - 27.
[24] 曾嘉，李洁，2020. 基于工业电力大数据的 GDP 数据精准测算实证分析[J]. 哈尔滨工业大学学报(社会科学版)，22(1)：133 - 140.
[25] 张崇，吕本富，彭赓，刘颖，2012. 网络搜索数据与 CPI 的相关性研究[J]. 管理科学学报，15(7)：50 - 59 + 70.
[26] 张书煜，王瑶，范婷婷，赵珵，王旭泽，2015. 基于社交媒体的投资者情绪对股市收益影响的大数据分析[J]. 中国市场，(25)：65 - 68.
[27] 中国人民银行数字人民币研发工作组，2021. 中国数字人民币的研发进展白皮书[EB/OL]. (2021 - 07 - 16)[2023 - 08 - 01]. http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/4293590/index.html.
[28] ANTOLIN-DIAZ J, DRECHSEL T, PETRELLA I, 2021. Advances in nowcasting economic activity: secular trends, large shocks and new data[R]. CEPR Discussion Paper, No. 15926.
[29] APARICIO D, BERTOLOTTO M I, 2020. Forecasting inflation with online prices[J]. International Journal of Forecasting, 36(2): 232 - 247.
[30] APRIGLIANO V, ARDIZZI G, MONTEFORTE L, 2017. Using the payment system data to forecast the Italian GDP[R]. Bank of Italy Temi di Discussione (Working Paper), No. 1098.
[31] ASKITAS N, ZIMMERMANN K, 2009. Google econometrics and unemployment forecasting[J]. Applied Economics Quarterly, 55(2): 107 - 120.
[32] AUER R, FROST J, GAMBACORTA L, MONNET C, RICE T, SHIN H S, 2022. Central bank digital currencies: motives, economic implications and the research frontier[J]. Annual Review of Economics, Forthcoming.
[33] AZAR P D, LO A W, 2016. The wisdom of Twitter crowds: predicting stock market reactions to FOMC meetings via Twitter feeds[J]. The Journal of Portfolio Management, 42(5): 123 - 134.
[34] BACH M P, KRSTIĆ Ž, SELJAN S, TURULJA L, 2019. Text mining for big data analysis in financial sector: a literature review[J]. Sustainability. DOI: 10.3390/su11051277.
[35] BAKER M, WURGLER J, 2007. Investor sentiment in the stock market[J]. Journal of Economic Perspectives, 21(2): 129 - 152.
[36] BANK M, LARCH M, PETER G, 2011. Google search volume and its influence on liquidity and returns of German stocks[J]. Financial Markets and Portfolio Management, 25(3): 239 - 264.
[37] BENNANI H, 2020. Central bank communication in the media and investor sentiment[J]. Journal of Economic Behavior & Organization, 176: 431 - 444.
[38] BERGER H, EHRMANN M, FRATZSCHER M, 2011. Monetary policy in the media[J]. Journal of Money, Credit and Banking, 43(4): 689 - 709.
[39] BIANCHI F, GÓMEZ-CRAM R, KIND T, KUNG H, 2023. Threats to central bank independence: high-frequency identification with twitter[J]. Journal of Monetary Economics, 135(4): 37 - 54.
[40] BINETTE A, TCHEBOTAREV D, 2019. Canada's monetary policy report: if text could speak, what would it say?[R]. Staff Analytical Notes.
[41] BLINDER A S, EHRMANN M, FRATZSCHER M, DE HAAN J, JANSEN D J, 2008. Central bank communication and monetary policy: a survey of theory and evidence[J]. Journal of Economic Literature, 46(4): 910 - 945.

[42] BLUWSTEIN K, BUCKMANN M, JOSEPH A, KAPADIA S, SIMSEK Ö, 2023. Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach[J]. Journal of International Economics. DOI: 10. 1016/j. jinteco. 2023. 103773.

[43] BOK B, CARATELLI D, GIANNONE D, SBORDONE A M, TAMBALOTTI A, 2018. Macroeconomic nowcasting and forecasting with big data[J]. Annual Review of Economics, 10: 615 – 643.

[44] BOLLEN J, MAO H, ZENG X, 2011. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2(1): 1 – 8.

[45] BORN B, EHRMANN M, FRATZSCHER M, 2014. Central bank communication on financial stability[J]. The Economic Journal, 124(6): 701 – 34.

[46] BUCH C, 2017. Data needs and statistics compilation for macroprudential analysis[G]//BANK FOR INTERNATIONAL SETTLEMENTS. Data needs and statistics compilation for macroprudential analysis, volume 46.

[47] BULUT L, 2018. Google Trends and the forecasting performance of exchange rate models[J]. Journal of Forecasting, 37(3): 303 – 315.

[48] BUONO D, MAZZI G L, KAPETANIOS G, MARCELLINO M, 2017. Big data types for macroeconomic nowcasting[J]. Eurostat Review on National Accounts and Macroeconomic Indicators, (1): 93 – 145.

[49] CAVALLO A, 2013. Online and official price indexes: measuring Argentina's inflation[J]. Journal of Monetary Economics, 60(2): 152 – 165.

[50] CAVALLO A, RIGOBON R, 2016. The billion prices project: using online prices for measurement and research [J]. Journal of Economic Perspectives, 30(2): 151 – 178.

[51] CHADWICK M G, SENGUL G, 2015. Nowcasting the unemployment rate in Turkey: let's ask Google[J]. Central Bank Review, 15(3): 15 – 40.

[52] CHATRATH A, MIAO H, RAMCHANDER S, VILLUPURAM S, 2014. Currency jumps, cojumps and the role of macro news[J]. Journal of International Money and Finance, 40: 42 – 62.

[53] CHEN M, LIU Y, MAO S, 2014. Big data: a survey[J]. Mobile Network and Applications, 19(2): 171 – 209.

[54] CHOI H, VARIAN H, 2010. Predicting initial claims for unemployment benefits[J]. Social Science Electronic Publishing.

[55] CHOI H, VARIAN H, 2012. Predicting the present with Google Trends[J]. The Economic Record, 88(S1): 2 – 9.

[56] COOK K A, THOMAS J J, 2005. Illuminating the path: the research and development agenda for visual analytics [M]. Richland: Institute of Electrical and Electronics Engineers.

[57] CORREA R, GARUD K, LONDONO J M, MISLANG N, 2021. Sentiment in central banks' financial stability reports[J]. Review of Finance, 25(1): 85 – 120.

[58] D'AMURI F, MARCUCCI J, 2010. 'Google it!' Forecasting the US unemployment rate with a Google job search index[J]. SSRN Electronic Journal. DOI: 10. 2139/ssrn. 1594132.

[59] DA Z, ENGELBERG J, GAO P, 2011. In search of attention[J]. The Journal of Finance, 66(5): 1461 – 1499.

[60] DESAI R, LOPMAN B A, SHIMSHONI Y, HARRIS J P, PATEL M M, PARASHAR U D, 2012. Use of internet search data to monitor impact of rotavirus vaccination in the United States[J]. Clinical Infectious Diseases, 54(9): e115 – e118.

[61] DIEBOLD F X, 2012. On the origin (s) and development of the term 'Big Data'[R]. Working Paper, No. 12 – 037. Philadelphia: University of Pennsylvania.

[62] DIEIJEN M, LUMSDAINE R L, 2019. What say they about their mandate? a textual assessment of Federal Reserve speeches[J]. Social Science Electronic Publishing. DOI: 10. 2139/ssrn. 3455456.

[63] DOLL C N, MULLER J P, MORLEY J G, 2006. Mapping regional economic activity from night-time light satellite imagery[J]. Ecological Economics, 57(1): 75 – 92.

[64] ELVIDGE C D, BAUGH K E, KIHN E A, KROEHL H W, DAVIS E R, DAVIS C W, 1997. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption[J]. International Journal of Remote Sensing, 18(6): 1373 – 1379.

[65] ERDOGAN B E, 2013. Prediction of bankruptcy using support vector machines: an application to bank bankruptcy

[J]. Journal of Statistical Computation and Simulation, 83(8): 1543 – 1555.

[66] ETTREDGE M, GERDES J, KARUGA G, 2005. Using web-based search data to predict macroeconomic statistics [J]. Communications of the ACM, 48(11): 87 – 92.

[67] EVANS M D, LYONS R K, 2008. How is macro news transmitted to exchange rates?[J]. Journal of Financial Economics, 88(1): 26 – 50.

[68] FAMA E F, 1970. Efficient capital markets: a review of theory and empirical work[J]. The Journal of Finane, 25(2): 383 – 417.

[69] FAMA E F, 1991. Efficient capital markets II[J]. The Journal of Finance, 46(5): 1575 – 1617.

[70] FLOOD M D, LEMIEUX V L, VARGA M, WONG B L W, 2016. The application of visual analytics to financial stability monitoring[J]. Journal of Financial Stability, 27(12): 180 – 197.

[71] GALBRAITH J W, TKACZ G, 2015. Nowcasting GDP with electronic payments data[R]. Statistics Paper, Series 10. European Central Bank.

[72] GENTLEMAN R, CAREY V J, 2008. Unsupervised machine learning [M]//HAHNE F, HUBER W, GENTLEMAN R, FALCON S. Bioconductor case studies. Springer: 137 – 157.

[73] GINSBERG J, MOHEBBI M, PATEL R, BRAMMER L, SMOLINSKI M, BRILLIANT L, 2009. Detecting influenza epidemics using search engine query data[J]. Nature, 457: 1012 – 1014.

[74] GUZMAN G, 2011. Internet search behavior as an economic forecasting tool: the case of inflation expectations[J]. Journal of Economic and Social Measurement, 36(3): 119 – 167.

[75] HEARST M, 2003. What is text mining[EB/OL]. (2003 – 10 – 17)[2023 – 08 – 01]. https://people. ischool. berkeley. edu/ ~ hearst/text-mining. html.

[76] HEIJMANS R, HEUVER R, LEVALLOIS C, VAN LELYVELD I, 2016. Dynamic visualization of large transaction networks: the daily Dutch overnight money market[J]. SSRN Electronic Journal, 2(2): 57 – 79.

[77] HENDERSON J V, STOREYGARD A, WEIL D N, 2011. A bright idea for measuring economic growth[J]. American Economic Review, 101(3): 194 – 199.

[78] HENDERSON J V, STOREYGARD A, WEIL D N, 2012. Measuring economic growth from outer space[J]. American Economic Review, 102(2): 994 – 1028.

[79] HENDRY S, MADELEY A, 2010. Text mining and the information content of Bank of Canada communications [J]. Staff Working Paper 10 – 31, Bank of Canada.

[80] HUBERT P, LABONDANCE F, 2016. Central bank sentiment and policy expectations[J]. Working Paper 2016 – 07, CRESE.

[81] IFC, 2015. Central banks' use of and interest in"big data"[R]. IFC Bulletin, No. 3.

[82] IFC, 2019. The use of big data analytics and artificial intelligence in central banking[R]. IFC Bulletin, No. 50.

[83] INDACO A, 2020. From twitter to GDP: estimating economic activity from social media[J]. Regional Science and Urban Economics. DOI: 10. 1016/j. regsciurbeco. 2020. 103591.

[84] JAGTIANI J, VERMILYEA T, WALL L D, 2018. The roles of big data and machine learning in bank supervision [J]. Banking Perspectives, Forthcoming.

[85] JOY M, RUSNÁK M, ŠMÍDKOVÁ K, VAŠÍŠEK B, 2017. Banking and currency crises: differential diagnostics for developed countries[J]. International Journal of Finance & Economics, 22(1): 44 – 67.

[86] KARABULUT Y, 2011. Can facebook predict stock market activity?[R]. Working Paper, University of Frankfurt, Germany. DOI: 10. 2139/ssrn. 2017099.

[87] KEISTER T, MONNET C, 2022. Central bank digital currency: stability and information[J]. Journal of Economic Dynamics and Control. DOI: 10. 1016/j. jedc. 2022. 104501.

[88] KIM N, LUČIVJANSKÁ K, MOLNÁR P, VILLA R, 2019. Google searches and stock market activity: evidence from Norway[J]. Finance Research Letters, 28(3): 208 – 220.

[89] KOPPEL M, SHTRIMBERG I, 2006. Good news or bad news? let the market decide[M] // SHANAHAN J G, QU Y, WIEBE J. Computing attitude and affect in text: theory and applications. Dordrecht: Springer, 297 – 301.

[90] KOSSE A, MATTEI I, 2022. Gaining momentum-Results of the 2021 BIS survey on central bank digital currencies

[R]. BIS Paper, No. 125.

[91] KOTSIANTIS S B, ZAHARAKIS I, PINTELAS P, 2007. Supervised machine learning: a review of classification techniques[J]. Emerging Artificial Intelligence Applications in Computer Engineering, 160(1): 3 – 24.

[92] LANEY D, 2001. 3D data management: controlling data volume, velocity and variety[R]. Technical Report, META Group.

[93] LEE Y J, KIM S, PARK K Y, 2019. Measuring monetary policy surprises using text mining: the case of Korea[J]. SSRN Electronic Journal. DOI: 10. 2139/ssrn. 3347429.

[94] LUCCA D O, TREBBI F, 2009. Measuring central bank communication: an automated approach with application to FOMC statements[R]. Working Paper, No. 15367. National Bureau of Economic Research.

[95] LUPIANI-RUIZ E, GARCÍA-MANOTAS I, VALENCIA-GARCÍA R, GARCÍA-SÁNCHEZ F, CASTELLANOS-NIEVES D, FERNÁNDEZ-BREIS J T, CAMÓN-HERRERO J B, 2011. Financial news semantic search engine [J]. Expert Systems with Applications, 38(12): 15565 – 15572.

[96] LÜDERING J, TILLMANN P, 2020. Monetary policy on twitter and asset prices: evidence from computational text analysis[J]. The North American Journal of Economics and Finance. DOI: 10. 1016/j. najef. 2018. 11. 004.

[97] MAHAJAN A, DEY L, HAQUE S M, 2008. Mining financial news for major events and their impacts on the market[C]. Sydney: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 423 – 426.

[98] MATHUR A, SENGUPTA R, 2019. Analysing monetary policy statements of the Reserve Bank of India[J]. Social Science Electronic Publishing. DOI: 10. 2139/ssrn. 3383869.

[99] MCMAHON M, SCHIPKE A, XIANG L, 2018. China's monetary policy communication: frameworks, impact, and recommendations[R]. Working Paper, IMF. DOI: 10. 5089/9781484385647. 001.

[100] MEINUSCH A, TILLMANN P, 2015. Quantitative easing and tapering uncertainty: evidence from Twitter[R]. MAGKS Paper on Economics, Philipps-Universität Marburg.

[101] MIAN G M, SANKARAGURUSWAMY S, 2012. Investor sentiment and stock market response to earnings news [J]. The Accounting Review, 87(4): 1357 – 1384.

[102] NAKAMURA E, STEINSSON J, 2008. Five facts about prices: a reevaluation of menu cost models[J]. The Quarterly Journal of Economics, 123(4): 1415 – 1464.

[103] NAKAMURA E, STEINSSON J, 2010. Monetary non-neutrality in a multisector menu cost model[J]. The Quarterly Journal of Economics, 125(3): 961 – 1013.

[104] NAKAMURA E, STEINSSON J, SUN P, VILLAR D, 2018. The elusive costs of inflation: price dispersion during the US great inflation[J]. The Quarterly Journal of Economics, 133(4): 1933 – 1980.

[105] NOFSINGER J R, 2005. Social mood and financial economics[J]. The Journal of Behavioral Finance, 6(3): 144 – 1460.

[106] OSHIMA Y, MATSUBAYASHI Y, 2018. Monetary policy communication of the bank of Japan: computational text analysis[R]. Discussion Paper, No. 1816. Kobe: Kobe University.

[107] OZTURK S S, CIFTCI K, 2014. A sentiment analysis of twitter content as a predictor of exchange rate movements [J]. Review of Economic Analysis, 6(2): 132 – 140.

[108] PAVLICEK J, KRISTOUFEK L, 2015. Nowcasting unemployment rates with google searches: evidence from the visegrad group countries[J]. PloS one. DOI: 10. 1371/journal. pone. 0127084.

[109] PICAULT M, RENAULT T, 2017. Words are not all created equal: a new measure of ECB communication[J]. Journal of International Money and Finance, 79: 136 – 156.

[110] PREIS T, MOAT H S, STANLEY H E, 2013. Quantifying trading behavior in financial markets using Google Trends[J]. Scientific Reports, 3(1): 1 – 6.

[111] RAVISANKAR P, RAVI V, RAO G R, BOSE I, 2011. Detection of financial statement fraud and feature selection using data mining techniques[J]. Decision Support Systems, 50(2): 491 – 500.

[112] RISTOLAINEN K, 2018. Predicting banking crises with artificial neural networks: the role of nonlinearity and heterogeneity[J]. The Scandinavian Journal of Economics, 120(1): 31 – 62.

[113] RYBINSKI K, 2019. A machine learning framework for automated analysis of central bank communication and media discourse: the case of Narodowy Bank Polski[J]. Bank i Kredyt, 1: 1 – 20.

[114] SAGIROGLU S, SINANC D, 2013. Big data: a review[C] // 2013 International Conference on Collaboration Technologies and Systems. IEEE. DOI: 10. 1109/CTS. 2013. 6567202.

[115] SEMIROMI H N, LESSMANN S, PETERS W, 2020. News will tell: forecasting foreign exchange rates based on news story events in the economy calendar[J]. The North American Journal of Economics and Finance. DOI: 10. 1016/j. najef. 2020. 101181.

[116] SHAPIRO A H, WILSON D J, 2019. Taking the Fed at its word: direct estimation of Central Bank objectives using text analytics[J]. Working Paper, Federal Reserve Bank of San Francisco. DOI: 10. 24148/wp2019 – 02.

[117] SMITH G P, 2012. Google Internet search activity and volatility prediction in the market for foreign currency[J]. Finance Research Letters, 9(2): 103 – 110.

[118] SORAMÄKI K, BECH M L, ARNOLD J, GLASS R J, BEYELER W E, 2007. The topology of interbank payment flows[J]. Physica A: Statistical Mechanics and its Applications, 379(1): 317 – 333.

[119] THORSRUD L A, 2016. Nowcasting using news topics. Big data versus big bank[R]. Working Paper, Norges Bank. DOI: 10. 2139/ssrn. 2901450.

[120] TOOLE J L, LIN Y R, MUEHLEGGER E, SHOAG D, GONZÁLEZ M C, LAZER D, 2015. Tracking employment shocks using mobile phone data[J]. Journal of the Royal Society Interface, 12(107). DOI: 10. 1098/rsif. 2015. 0185.

[121] VARGAS M R, DE LIMA B S, EVSUKOFF A G, 2017. Deep learning for stock market prediction from financial news articles[C] // 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, 60 – 65. DOI: 10. 1109/CIVEMSA. 2017. 7995302.

[122] VU T T, SHU C, HA Q T, COLLIER N, 2012. An experiment in integrating sentiment features for tech stock prediction in twitter[C] // International Conference on Computational Linguistics.

[123] WU L, BRYNJOLFSSON E, 2015. The future of prediction: how Google Searches foreshadow housing prices and sales[M]. Chicago: University of Chicago Press, 89 – 118.

[124] YASIR M, AFZAL S, LATIF K, CHAUDHARY G M, SONG O Y, 2020. An efficient deep learning based model to predict interest rate using twitter sentiment[J]. Sustainability, 12(4). DOI: 10. 3390/su12041660.

Literature Review on Big Data Analytics in Macro Finance
—A Central Bank Perspective

Sylvia Xiaolin Xiao*　Hansheng Wang
(Guanghua School of Management, Peking University)

Summary: Big data and relevant technologies have not only provided unprecedented amounts of data related to the macroeconomy and the whole society, formalizing a big data "ecology", but also influenced and reshaped the process of public policy making and operation. Meanwhile, intensive attention from the field of economic research, particularly from the perspective of central banks all over the world, has been paid to big data and relevant analytical methodologies.

We know that the main functions of a central bank are, within the framework of a country's monetary policy operations, to use conventional monetary policy tools (such as open market operations, discount-window loans, and required reserve ratios), or unconventional monetary policy tools, to adjust interest rates and money supply, to achieve the mandates of monetary policy, such as full employment and price stability. In different stages of monetary policy operations, including before, during, and after, the central bank's daily work includes: collecting a large amount of data, conducting regular data analysis, macroeconomic forecasting, and economic cycle analysis; releasing regular monetary policy reports and communicating with the public (traditional press conferences along with widespread use of social media such as Twitter and Facebook overseas, and Weibo and WeChat in China); and conducting micro-financial supervision and macro-prudential supervision based on a large amount of financial data, and so on. It's worth noting that, especially after the 2008 Great Recession, central banks around the world have paid more attention to macro-prudential supervision, closely monitoring real-time dynamics in specific financial markets, such as shadow banking, systemically important financial institutions, and the real estate market, through big data analysis. Therefore, in the current big data era, from the perspective of central banks, we want to address the following questions through a review of the literature: With the emergence of big data and related technologies, what new changes have occurred in data collection and analysis by central banks, particularly in the field of macro finance research and analysis, and in which specific areas of macro finance? Alongside new granular micro financial data and new analytical tools, what interesting new predictions and analysis results have emerged? Have new applications arisen in the fields of monetary policy communication, macroeconomic forecasting, and macro-prudential supervision? In comparison to traditional data and analytical methods, what advantages do big data analytics have, and has it also brought new problems, risks, and challenges?

* Corresponding Author: Sylvia Xiaolin Xiao, Guanghua School of Management, Peking University, E-mail: sylvia. xiao@gsm. pku. edu. cn.

This paper conducts a comprehensive review of recent heuristic efforts in applying big data analytics to macro finance, offering contributions of review as follows. Firstly, the review focuses on a central bank's perspective. Secondly, it covers diverse data types such as textual data and emerging economic indicators (e.g., electronic payments, mobile data, satellite images). Thirdly, it employs varied analytics like Bayesian dynamic factor models for real-time economic trend estimation. Lastly, it provides insightful suggestions for future research and application, particularly concerning China. The review identifies three key literature domains. First, researchers extract structural insights from various central bank communication channels, including press releases and media sentiment. Second, big data enables more accurate macroeconomic forecasting, even narrowing the gap between the current and most recent data (nowcasting). Third, big data bolsters macro-prudential policies by offering indicators for policy framework enhancement, supervision, crisis prediction, and market trends. While big data's potential is acknowledged, unexplored avenues persist, especially in China. Recommendations include analyzing media sentiment regarding monetary policies, leveraging China's data-rich environment for better nowcasting, and exploring central bank digital currency (CBDC) applications in big financial data collection and analysis.

Keywords: Big Data Analytics; Macro Finance; Central Bank; Review

JEL Classification: C10; E50; E58