

大数据的力量

计算能力正在推动机器学习的发展，改变着商业和金融模式

桑吉夫·兰詹·达斯

在当今世界，人们能够获取的数据量之大，是10年前所无法想象的。公司积累新数据的速度超过了它们对数据进行组织和利用的能力。现在，它们需要弄清如何利用这一规模庞大的数据，更好地制定决策，提升业绩。

数据科学这一全新的领域旨在从数据中获取可用知识，尤其是大数据——即超大规模的数据集，通过对其进行分析可发现各种数据模式、数据趋势和数据关联。数据科学从数据收集和数据组织扩大到了数据分析和数据洞察，从而最终将所学的知识运用到实践中去。该领域与所有的人类活动相关联——经济学、金融和商业无一例外。

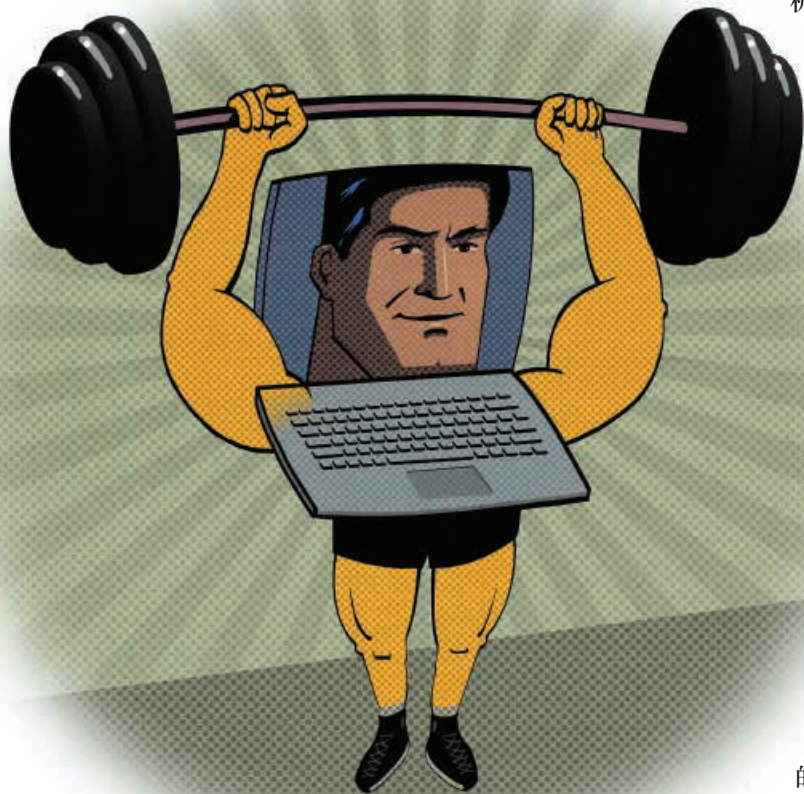
数据科学带来了机器学习的工具——机器学习是一种人工智能，它能使计算机在没有明确编程的情况下拥有学习的能力 (Samuel, 1959)。这些工具以及大量的数据具有改变整个商业管理和经济政策分析模式的潜力。

其中，所产生的一些变化为人们带来了希望。

消费者分析

由于数据科学带来了极大的经济效益，其在商业中的应用得以迅速增长，这也就不足为奇了。

在竞争激烈的市场中，买家支付的价格相同，卖家的收入等于价格乘以销售量。然而，还有许多买家愿意支付高于均衡价格的费用，所以可通过大数据进行客户分析，



挖掘这些买家的消费者剩余。

根据客户分析结果，对不同的客户制定不同的价格可以使公司获得某种商品客户愿意且能够支付的最高价格。通过大数据对差别定价或者市场细分进行优化是有利可图的，而且在一些行业中符合行业规范——如航空公司，但现在这种做法也扩大到了产品层面。

此外，通过目标定价法所获得的收益还能够使公司为消费者提供折扣，为这些无力承担商品均衡价格的消费者提供优惠，这样也会增加公司收入，扩大客户群，甚至还可能提升社会福利。通过大数据进行消费者分析是一些公司拥有较高估值的重要原因，其中包括脸书、谷歌和安客诚等，这些公司都是根据客户数据提供相关产品和服务的。

大数据在进行消费者开发的同时，还在改变着企业的经营方式，这种经营方式对消费者也同样有益。公司利用人们在社交媒体互动中产生的数据，可以更好地了解他们的信用行为。将人们过去的信用历史同他们的社交媒体记录相互关联，可以提升信用评分系统。同时，贷款方还可以通过这种方式向那些原本无法获得贷款的人提供贷款。

尤其是，大数据可以避免人们在信息量有限的情况下做出带有偏见的决策。缺乏详尽的个人数据会导致申请贷款时受到歧视，这种现象自 20 世纪 30 年代以来就一直存在。抵押贷款人会将地图上的一些区域划上红线，表明由于种族或者民族构成的原因，他们不会给那些地方的人放贷。这种带有成见的刻板做法使得一些社会阶层无法获得信用贷款。

然而，大数据消除了这种刻板的做法。粗糙、主观的数据现在被更为精细、更加个性化的数据所取代。信用评级公司除了利用典型的人口统计数据，如收入、年龄和住址等信息，还能通过人们的社交媒体互动、文本流、微博、信用卡类别和分析数据了解客户异质性 (Wei 等人, 2014)。使用更为细化的数据，有助于更好地根据信用情况对客户进行分类。

预测和风险分析

有了数据科学方法，经济预测发生了巨大的转变。在传统的预测方法中，像季度 GDP 报告这样的重要经济统计数据，必须要等待很长时间才能发布。而数据科学能够根据经常上报的信息，如失业率、工业订单量，甚至是新闻媒体情绪等，对较少报道的那些变化的数据进行预测，可以及时发布统计报告，避免延迟问题。

这一系列的方法被称为“临近预报”——也叫做“对当前的预测”，但最好将其理解为“实时预测”（见“数据女王”，《金融与发展》2014 年 3 月号）。

数据科学在系统性金融风险方面也取得了进展。现在世界比以往任何时候都要互联得更为紧密，通过衡量这些关系，可以在进行经济决策时获得全新的视角。

透过网络的镜头观察系统性风险是一种有效的方法。现在，数据科学家们利用丰富的数据绘制出了银行、保险公司、经纪人等各方之间的相互联系。这样可以明显地了解哪些银行间的相互关联更为紧密。通过基于特征值的计算方法，也可以了解哪些银行的影响力最大。这些网络一旦建立，数据科学家们便可以衡量出个体金融机构产生的风险在整体风险中的占比，同时还能衡量出某个金融体系中存在的风险程度，从而为监管部门提供一种全新的分析方法，以便对系统性风险进行管理。见 Espinosa-Vega 和 Solé (2010)；IMF (2010)；Burdick 等人 (2011)；Das (2016)。

这些方法广泛地借鉴了在社会学中所开发的社交网络数学运算法，并被应用于使用高级计算机科学模型的大型网络中，与多门学科实现了卓有成效的结合。

不止于“词”

文本分析在数据科学领域发展非常迅速，也是金融和经济领域定量数据的一个有趣补充（见本期《金融与发展》，“变革的两面性”）。基于文本挖掘的商业应用程序数量众多：比如，像 iSentium 这样的公司通过推特从社交媒体上提炼长短期的情绪数据；StockTwits 公司通过一款手机网络应用程序提供投资情绪指标。

大数据可以避免人们在信息量有限的情况下做出带有偏见的决策。

现在根据每年向美国证券交易监督委员会递交的公司财务业绩年度报告，即企业年度报表 (10-K) 中的季度收益对企业进行排名已经成为可能。季度报告中和风险相关的词汇汇集可以准确地对收入预测进行排名。那些季度报告内容难以让人理解的公司，其收益通常比较差——很可能是因为它们想在报告中用含糊其辞的表达来汇报坏消息（见 Loughran 和 McDonald, 2014）。利用迷雾指数这一古老的可读性衡量标准，可以很容易地对财务报告的可读性进行评分，像消费者金融保护局这样的监管部门正在探索设置可读性标准。

研究甚至发现仅根据季度报告的长度就足以察觉到是否有坏消息（较长的报告预示着公司收入下降），同样也是由于含糊其辞而导致报告内容冗长；最后，仅仅根据公司上传给美国证券交易监督委员会文件的大小也能看出公司的季度收入情况。在这个发展非常迅速的工作领域，有望出现更多的信息。

一个名为“新闻分析学”的新领域也在挖掘数据相关的信息。像 RavenPack 这样的公司所提供的服务正在不断增长。这些服务包括情绪评分、对交易进行预测性分析以

及宏观经济预测等。RavenPack 公司从新闻和社交媒体中挖掘出大量的非结构化数据，将其转化为细化的数据和指标，为金融公司进行资产管理、做市、风险管理以及确保合规性提供支持。

研究甚至发现仅根据季度报告的长度就足以察觉到是否有坏消息。

在这一类别中，新闻流分析非常有趣。对冲基金每天挖掘出数千条动态消息，从中找出排名前五，甚至前十的消息，然后每天对这些话题的占比情况进行追踪，以发现市场条件所发生的可交易转变。与此相似的分析法对于决策者和监管机构同样有用，如央行行长。例如，当新闻中所讨论的某些话题占比突然发生变化时（如通化膨胀、汇率或者增长），可能就是重新审视利率政策的时候了。

话题分析始于创建一个巨大的词汇频率表，即“术语—文档矩阵”，对数千条新闻文章进行分类。术语（词汇）是表格的行向量，每篇新闻文章在表格中单独占一行。这个巨型的矩阵可以通过数学分析法分析词汇和文章之间的关系，从而发现相关主题。对词汇进行索引，通过利用潜在语义索引和潜在狄利克雷分布之类的机器学习方式查找主题。潜在狄利克雷分布分析可在这些主题中生成一组话题以及单词列表。

虽然在此讨论这些建模方法太过于技术性，但这些不只是一些统计技巧，通过这些技巧可在一系列文档中（如新闻流）中找到主要的词群。而这些语言线索可能会被经济政策制定者和政治决策者所广泛采用——例如，对某个政治运动中的信息进行重新定义。

人工智能及未来

计算机功能变得比以往任何时候都要强大，其处理大批量数据的能力促进了人工智能领域的发展。受生物神经网络的启发，一种被称为“深度学习网”的新型算法经证实模仿人脑运行机制方面表现非常优秀，在人工智能方面有许多成功的案例。

深度学习是一种统计学方法论，可利用人工神经网络映射出大量输入变量和输出变量——即进行模式识别。通过基于硅和软件的神经网络可对信息进行分析。通过数据强化神经元之间的联系，这一点很像人类学习经验的过程。深度学习获得很大成功的原因有两方面：一是有大量数据供机器学习，二是计算功能呈现指数式增长，这些主要是受到深度学习应用程序专用计算机芯片的发展所驱动。

深度学习促进了很多被人们认为是理所当然的现代科

技的发展，如机器翻译、无人驾驶汽车和图像识别及标记等。这种科技可能很快就会使经济和政策发生改变。信用评级机构已经利用深度学习功能，在无人类干涉的情况下生成报告了。大型的深度学习神经网络可能很快就可以预测并识别经济变量之间的关系，而且效果比标准的统计方法要好。

很难预测在这个沉闷的科学中哪些领域通过使用机器学习会获得最大的增长，但这个新的时代确实已经到来。正如科幻作家威廉·吉布森（William Gibson）所说的，“未来已经在这儿了，只是表现得不是很均衡罢了。”■

桑吉夫·兰詹·达斯（Sanjiv Ranjan Das）是圣塔克拉拉大学利维商业学院的教授。

参考文献：

- Billio, Monica, Mila Getmansky, Andrew W. Lo, and Liorana Pelizzon, 2012, “Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors,” *Journal of Financial Economics*, Vol. 104, No. 3, pp. 535–59.
- Burdick, Douglas, Mauricio A. Hernandez, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Shivakumar Vaithyanathan, and Sanjiv Das, 2011, “Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study,” *IEEE Data Engineering Bulletin*, Vol. 34, No. 3, pp. 60–7.
- Das, Sanjiv, 2016, “Matrix Metrics: Network-Based Systemic Risk Scoring,” *Journal of Alternative Investments*, Vol. 18, No. 4, pp. 33–51.
- Espinosa-Vega, Marco A., and Juan Solé, 2010, “Cross-Border Financial Surveillance: A Network Perspective,” *IMF Working Paper 10/105* (Washington: International Monetary Fund).
- International Monetary Fund (IMF), 2010, “Systemic Risk and the Redesign of Financial Regulation,” *Global Financial Stability Report, Chapter 2* (Washington, April).
- Lin, Mingfen, Nagpurnanand Prabhala, and Siva Viswanathan, 2013, “Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending,” *Management Science*, Vol. 59, No. 1, pp. 17–35.
- Loughran, Tim, and Bill McDonald, 2014, “Measuring Readability in Financial Disclosures,” *Journal of Finance*, Vol. 69, No. 4, pp. 1643–71.
- Samuel, A.L., 1959, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, Vol. 3, No. 3, pp. 210–29.
- Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas, 2015, “Credit Scoring with Social Data,” *Marketing Science*, Vol. 35, pp. 234–58.