大数据内涵的挖掘角度辨析

文/刘晓葳　朱建平

“大数据”现象诞生于人们数据搜集能力、数据传输能力、数据存储能力以及数据处理能力的极大发展，这也自然得益于技术进步的大环境引致信息生产方式与速度的变革。自2011年下半年起，该名词在业界与学界被广泛提及，无论这体现出的是真实繁荣抑或泡沫，大数据绝非仅仅是一种商业炒作；就其信息生产与使用方式而言，大数据环境将呈现出信息生产廉价化与泛化，披露民主化及载体数字化等特点，可以预见，借由大数据的发展，人们身处的信息环境将会得到根本性改变，数据将作为资本直接产生价值，而非仅提供参考意见。

人们倾向于从表象出发谈论大数据，如数据体量大（Volume）、多样化（Variety）、产生速度快（Velocity）、真实性（Veracity）与价值密度低（Value）。援引国际数据公司（Internet Data Center，IDC）的监测统计，预计到2020年全球将总共拥有35ZB（1ZB=1万亿GB）数据量，较2011年增长近20倍，通常这种数量化的表示被认为是大数据时代开启的标志，一个类似的比喻是，大数据如同数据的“群体像素”变大，使得其中体现的信息“图像”更为清晰。但注意对于信息量而言，“大”在具体技术语境下不尽相同：若以统计学概念厘定，大小样本量仅以30为界；若以处于特定环境下的个体感觉界定，大、小则只能是相对概念，十年前人们对信息体量的感觉显然与现今不可同日而语。可见大数据不能简单等同于大样本，而是具有确然语义的范畴，以表象层面的若干特点概括其内涵难免有失偏颇。可以预见，大数据愈益迫近由概念论说与尝试应用进入到大规模业务流程和产生巨大商业价值阶段，既然如此，避免误读大数据内涵就显得十分重要。至少迄今，大数据仍是既有莫大诱惑又充满神秘的“灰箱”，彰显和揭示大数据潜质以及蕴含意义必须借助于某种能够读懂、解构隐藏于大数据深层秘密的“工具眼”，即如对于一种新的艺术样式、新的卓尔不群的艺术作品，领略其中甘味需要借助类似立体眼镜等工具介质。笔者以为，目前适于读懂与解构大数据的工具眼（至少是其一）即数据挖掘。当然，也有可能大数据出现能够推进数据挖掘向新的更有价值的方向发展。

大数据背景下的数据价值产生，可视为由分布式数据架构、机器学习(Machine Learning，ML)及结果分析三部分完成。捕捉客户行为特征即数据获取主要属于数据架构部分，机器学习和结果分析则共同构成对已有数据的挖掘流程，已知数据挖掘是知识发现（Knowledge Discovery in Databases，KDD）行为，挖掘过程直接决定从已有数据中能得到哪些有用的知识。从这个特点出发，以数据架构为基础，大数据的价值实际落脚于挖掘流程，所以从挖掘角度出发解读大数据内涵，符合两者内在逻辑一致性的基本特征。

本文将从数据挖掘角度出发尝试对大数据内涵做一辩证探讨，其中关涉大数据的一系列热点问题。大数据就简单地等同于数据规模大？挖掘算法与直觉的重要性如何，大数据要求业务驱动还是数据驱动？目前大数据行业究竟是短缺还是过热？大数据最终导向的是客户便利还是隐私侵犯？借由辨析以上问题，大数据的内涵可处之昭然。

一、“大”“小”之辨

如前文所述，以数据量作为判断“大”、“小”的准则并不稳定，会随技术环境的发展而变化。同对高清制式、分辨率等图形技术指标的要求日臻苛刻类似，二十多年前 1.44MB 软盘（Floppy Disk）是 PC 的标准传输制式，人们觉得这个边长 3.5 英寸的正方形磁盘足够完成资料存储与搬运工作，而今以 GB、TB 为单位的移动数据存储介质都不再会令普通使用者感到惊叹。单纯从数据量解读大数据内涵的另一个危险在于，如果数据架构及挖掘技术不断进步，使得原先难以驾驭的海量数据变得易于处理，那么数据量大小就不成问题，而大数据概念也将昙花一现，最终被归为传统数据分析的一个发展阶段而已。可见理解大数据之大不能仅将目光聚焦在数据规模这样的特征性描述。大数据可以被理解为一种思维方式，这种思维方式要求人们“以数据为大”，即重视数据本身价值，将数据直接视作价值生成的资本，而非决策辅助工具或直觉的验证工具。就数据性质而言，大不仅仅是数据量的问题，更是分析逻辑和数据结构的变化，不能说有了应对的手段，大数据分析就和传统意义上的数据分析一致，从而使得数据挖掘无意义化。

数据规模之大不构成大数据的充分条件，但却是其分析的必备属性。同小体量数据相比，大的数据规模提供了确凿的归纳结果——这是随机性较强的少量数据无法比拟的。大数据允许我们通过在未经处理的数据集中逐条比对来发现微小但有说服力的线索，从而进行分析研究——这为深入的数据挖掘分析和精确的决策提供了可能性，而小规模数据分析通常得不到精确的结论，数据的潜在规律难以被发现。要注意到，并非全部行业都已有应用大数据的基础，这取决于行业是否能够在架构层面上提供恰当规模的数据。若以大数据思维分析小数据集，模型训练无法达到预期精确度，分析结果缺乏验证环境，可能效果反不如直觉判断。

大数据之“大”，根本上是一种思维方式的转变，这种思维方式又要以大规模数据为基础产生价值，在大数据实践中，数据与思维，二者紧密结合，缺一不可。

二、算法与直觉之辨

算法与直觉的角色转变应该是大数据时代的另一热点，两者定位不同直指在数据的价值产生过程中应以数据还是业务目标为根本驱动力。支持以业务目标为核心的理由显而易见：数据挖掘直接为业务目标服务，挖掘中新发现的知识内容很难通过理论进行有效性检验，投入实践是检验挖掘结果的必要路径，业务驱动似乎理所当然。但问题在于客户的真实需求往往难以确定——消费者时常无法准确表达真实需求，这就表示可能存在一些无法被原有业务目标所涵盖的有价值的知识；消费者的潜在行为习惯和真实需求往往是寓于某一特定的数据集中，而大数据环境为挖掘这些习惯提供了便利，使得“啤酒与尿布”一类经典案例成为可能。

不妨回归数据挖掘的定义来讨论此议题。对比计量经济学与理论经济学的关系，计量经济学遵循波普尔的证伪主义线路：提出假说，检验假说，然后过渡到下一个问题和猜想；一个模型的建立和验证以及经济意义说明，实际上是对某种预先的理论假设加以验证的过程，这在经典计量经济学发展阶段尤为突出。与此不同，数据挖掘是典型的知识发现过程，与传统数据分析方法的最大区别在于可能自主的发现一些出人意料的新知识。该特性决定了数据挖掘并非对某些假设的验证，而是一种提出假设的过程，例如上文谈及的“啤酒与尿布”案例，数据挖掘的作用流程是对数据进行预处理与挖掘（从数据出发）——某类消费者通常同时购买啤酒与尿布（获得待验证的知识）——以此知识为基础观察发现，该购买特征为年轻父亲所具备 (对知识验证说明)。而计量经济学则是通过大量观察或天才的内省猜测啤酒与尿布似乎通常被同时购买，即经验与猜想、理论假设；进而分析两者数量相关性，或基于理论假设建模；最终检验两者是否具备相关性，或对模型进行检验与解释，最终支持或推翻假设；进一步可能调整模型重新验证原假设或转向其他猜想。数据挖掘的思路更接近数据生成过程本身，具备先验性质，挖掘所得不依赖于预先假设，且可能超过预期业务目标。从这点出发，数据驱动模式似乎能够产生更多的价值，但对该模式的质疑随之而来：一是数据可靠性问题。数据可能具备大量随机性，数据真实性难以保证，尤其是非原始数据，其加工过程可能对结果有效性有所影响；二是目前挖掘数据的算法仍然不够智能，不足以处理含有大量人类行为、兴趣和偏好的数据。

大数据环境下，大规模原始数据变得易于获取，这使得第一点疑虑迎刃而解，大数据集降低了随机偏误，人们在分析时可自主对原始数据进行清洗（Data Cleaning），数据可靠性得到增强。而对算法不够智能化的担忧依然是目前的热点议题，2013 年 1 月在麻省理工学院召开的关于大数据的业界会议就以算法与直觉为主题，提出“大数据很重要，直觉也不可或缺”的观点。诚然，目前的机器学习的可靠性似乎还离 1950 年图灵试验的设想有一定距离，倡导抛弃模型，更多依靠数据分析师的能力与经验判断的呼声不绝于耳。好在随着深度学习（Deep learning）的提出与不断完善，近些年机器学习领域出现了突破性进展，算法的智能化大大增强。实际上，机器学习算法的发展空间巨大，只有将算

法作为超越直觉的驱动力，才能称得上真正的“以数据为大”，达到大数据时代人们所期望的高智能化水平、高生产效率、高服务质量和高决策精确度。当然，数据分析师的直觉亦不可忽视，具体挖掘操作需要业务理解的支持，使数据挖掘过程不至陷入“数据陷阱”。

三、短缺与过热之辨

著名市场研究公司 Gartner 曾于 2011 年上半年展望大数据前景，称该行业将在几年内创造 440 万 IT 岗位，最终跻身传统行业之列。振奋人心的预测言犹在耳， Gartner 却又于 2013 年 1 月起呼吁给大数据降温，称业界对大数据关注过热，存在泡沫。

支持过热说的最重要证据是，大数据兴起以来的典型成功案例并不多。追溯一个无法被忽视的事实：数据的体量与其总体价值并不呈线性增长关系，数据规模不断变大，但其价值密度相对减少，这对业务专家的个人直觉、数据科学家的分析能力以及算法的效率与精确度提出了更高要求。另一个事实是，数据挖掘方法所获取的新知识缺乏证实与证伪的逻辑一致性，它是依靠精确度判断的。抛开应用很难对新知识的有效性加以验证，更无法加入其后的知识再创造过程。但若只关注应用层面，大数据获得的知识便只是业务验证性的，难以产生新价值，即使获得了新知识，投入商业实践检验后，能否成为一种成功的范式加以推广也未可知。目前业界对大数据的关注过热之处，就是急于产生价值、试图占领大数据时代制高点。在这种情况下，数据沦为知识消费品而非可产生二次价值的资本，信息冗余现象越发常见，而为数不多的有用数据被迅速消耗失去再利用价值，所以目前还很难讲大数据为我们的决策环境提供了更多的资源还是制造了更严重的信息污染。

过热真实存在，大数据暂时的泡沫化反映的是真实与期望的差距，是思维和技术的相对短缺。此后大数据的进一步发展，分布式数据架构当然是资本；从挖掘角度讲，大数据环境对数据科学家和业务领域专家的需求空间巨大；两者之间传递的畅通化也需要注意，机器学习和数据分析领域应向模板化界面发展，以更为友好的使用界面为业务专家提供便利的渠道，使其直接参与数据分析过程，打破数据与业务的沟通壁垒。采取适当思维与技术运行大数据，是克服目前大数据受关注过热而创造价值不足的唯一手段。

四、便利与侵害之辨

维克托·迈尔·舍恩伯格在《删除：大数据取舍之道》一书中提到，现今世界上 90% 以上的信息以数字形式呈现，能够毫不费力地被存储、加工、操作和发送。这些数字信息塑造了大数据的可行性，也带来个人隐私被侵犯的可能性，这个结论符合直觉——毕竟数据处处都暗含隐私，数据越多，隐私的线索越容易被串联。数据构筑的“圆形监狱”使人举步维艰，即使隐私侵害没有即刻发生，我们也好像时时在受着监视：数据是否自愿上传？数据的使用是否得到客户授权？数据整合由谁完成？很可惜这几个问题的答案目前尚且暧昧不明。大部分的“行为数据”不是自愿上传，而是不经意或“不得不”透露的，如分布在医疗机构与金融机构服务器的个人资料。我们很少有对自己的数据进行授权声明的机会，或将授权当作无关紧要；很多数据看似是不重要的，不涉及安全隐私问题，但利用数据挖掘技术可能会从中找到线索；已有授权也缺乏时限，随着信息越来越多，原来的安全信息可能会被联系起来暴露个人隐私变成不安全信息。随着一些信息售卖事件的曝光，信息使用者的诚信也被怀疑。

但数据挖掘本身并不侵犯隐私，视频网站追踪单一客户浏览记录推荐视频，网商抓取购买历史推荐商品，这种涉及具体客户的推荐形式不是挖掘，只是传统销售手段的数字化表现。数据挖掘关注的是“典型人”特征和生成规则集，想要发现的也非这样显而易见的知识，相对于“你需要这个，所以你应该也需要那个”的模式，数据挖掘的模式实际上是“如果一个人需要这个，那可能他也需要那个”这样的“大众化”规则。所以买菠萝推荐沙拉酱（菠萝可能被用于制作水果沙拉）比买菠萝推荐菠萝削皮器这样的直线推介形式更接近数据挖掘的本意。理论上数据挖掘过程不涉及隐私，它不是个体特征描述，而是对数据集共性的抽取。类似统计学思想，规律的总结来自于平均，而精确地按照函数化拟合数据点则需要一系列的基函数——这些来自数据集外部的基函数替代了数据点中含有的全部个体标识信息，从而预测和分类工作都避开了客户隐私。

想要利用大数据创造价值，又要避免陷入数字化圆形监狱的处境，保密合约、有时限的数据上传与使用授权及挖掘过程中机密属性保护和身份标识模糊化是必要的，这需要法律和技术的双重保障。对数据信息加以保护之后，实在没有必要对大数据避之不及，毕竟它能够带来的便利无可比拟：世界越来越智能，工作效率越来越高，决定越来越准确——以数据为大，将数据作为直接产生价值的资本，可以将我们带向这样一个美好时代。CS

参考文献

[1] 高济等编 . 人工智能基础 [M]. 高等教育出版社，2002.

[2] 李子奈等著 . 计量经济学模型方法论 [M]. 清华大学出版社，2011.

[3] 魏瑾瑞 . 数据挖掘的“行情”[J]. 中国统计，2012（2）.

[4] 王珊等 . 架构大数据：挑战、现状与展望 [J]. 计算机学报，2011（10）.

[5] 朱建平著 . 数据挖掘的统计方法及实践 [M]. 中国统计出版社 .2005.

[6] 郑毅 . 大数据时代的特点 [J]. 新金融评论 .2012(1).

[7] Viktor Mayer-Sch！nberger 著 . 删除：大数据取舍之道 [M]. 袁杰译 . 浙江人民出版社 ,2013.

作者单位：厦门大学统计系
厦门大学数据挖掘研究中心