大数据像中医，有疗效，别神化

——专访国家特聘专家王晓阳教授

本报记者 **曹刚**

“大数据”这个新词，近两年曝光率颇高。字面意思好像谁都看得懂，却又似懂非懂。

谷歌预测流感趋势，微软预言奥斯卡奖，多次成功后，大数据被传得神乎其神。到底怎么解释？在社会生活中有哪些应用？我国核心技术研发现状如何？大数据和保护隐私间怎样平衡？记者专访国家特聘专家王晓阳教授。

王晓阳 1992 年在美国南加州大学获得计算机科学博士学位后，留美从事 20 多年科研工作，曾任美国国家自然科学基金项目主管及佛蒙特大学计算机系 Dorothean 讲席教授，一直在和“数据”打交道，两年前受聘成为复旦大学计算机科学技术学院院长。

“大数据的魅力是全面，而非精准”

记者：能否通俗解释大数据的概念？

王晓阳：何为大数据，学界没有统一定论。按照我的理解，我想了一个比喻。

以前采集数据像西医看病，验血、测心率、查病毒，具体而直接。而大数据的采集方式不同，有点像中医问诊，望闻问切，看面色、舌苔、脉象，全方位观察病人。大数据涉及更庞杂的数据，不见得精准、专业，但是有用。它的魅力是全面，而非精准。

传统的民意调查像西医，先设计问卷，确定样本，再抽样获得信息。大数据的做法有点像中医法则，以微博数据为例，成千上万博主都在表达某些感觉或想法，用微博数据来调查民意就如同整体搭搭脉，虽有些模糊，但能很好把握社会热点和舆论倾向等。

精准数据其实是我们向往的，但它有个成本问题，不太可能全面覆盖。大数据的特征包括数量大、信息杂、传输快。在许多情况下，“大”和“杂”提供好的覆盖性，“快”则使大数据更具适时性。

大数据固然重要，但过于依赖大数据，肯定也不行，对事物的详细了解还是需要精准数据。中医看整体而粗犷，西医管局部而细致，“中西医结合”更有效。

记者：在社会生活应用中，大数据能发挥什么作用？

王晓阳：许多应用都涉及一个关键词：二次利用。数据采集之初，原为某些特定需求，但二次利用后，又产生新的奇妙结果。

搜索引擎起初只是给用户提供便利，查询知识。经过巧妙再利用，便能预测流感趋势，比政府部门的数据更快。

城市里有很多摄像头，初衷是为防盗和监控路况。大量视频数据经过分析，能全面掌握城市脉搏，帮管理者做决策。

微博一开始只为表达自我，和舆情无关。但后来发现，还能了解舆情。

交通数据也会透露其他信息——演唱会举办期间，可以从场地附近的交通拥堵程度，判断明星受欢迎程度；微软中国研究院曾为北京上千辆出租车装 GPS 设备，记录数月行动轨迹。分析这些交通数据，结合人们的交通出行习惯，能发现各城市区域的属性，是商业区、住宅区还是娱乐区。

“收敛浮躁，核心技术是长远之道”

记者：海量数据与信息爆炸早就存在，大数据的概念为何近两年才出现？

王晓阳：海量数据和信息爆炸都与因特网相伴而生。起初我们的网速很快，但没多少数据在上面跑。

现在大不一样，尤其是近十年，制造数据的人和设备大增。电脑的功能越来越强大，价格也日益亲民。智能手机、平板电脑等移动终端很容易获取。

传感器、摄像头、计算机、POS 机、天上的卫星、地下的感应线圈，都在采集数据，导致数据量大爆炸，远超出原来的规模。

软件开源运动（开放源代码）兴起后，软件也越来越便宜，容易得到。

技术进步还催生了“云计算”，就是把很多便宜的电脑串在一起，大大提高计算能力；和网络结合后，无需拥有这些机器，也能采集、处理数据。

社会上各种脉搏跳动近几年越来越多——发微博，手机签到，网上搜索，哪怕是网购一件小商品，也能反映某种需求或流行……这些都是催生大数据的因素。

记者：我国大数据应用领域很热闹，技术研发的现状如何？

王晓阳：大数据最炫的地方在应用，核心技术往往被忽视。

采集、存储和分析等核心技术，主要掌握在美国人手里。国内投资很少，大家都看重结果，所以应用领域很热闹，还在国际上发表了不少应用方面的学术论文。

大数据刚起步，我们完全有机会和美国站上同一起跑线，现在却只能跟在人家后面跑。等别人研发出新一代技术，我们马上一窝蜂地拿来用。

浮躁是原因之一，整个科技创新环境都比较浮躁。从国家层面讲，应当重视扶植大数据的核心技术研发。

我们要收敛浮躁，保持冷静，认识到核心技术才是长远之道。不要神化大数据，以为掌握了许多数据，再买来机器，就万事大吉。这是很不负责任的想法。

“数据庞杂冰冷，需人的智慧激活”

记者：谷歌预测流感；微软预言奥斯卡，都靠大数据，怎么做到的？

王晓阳：谷歌 5 年前推出“流感趋势”服务，通过分析关键词，准确预测美国各地流感爆发情况。

这是大数据应用的典型案例。谷歌综合分析大量实时数据和历史数据，了解不同历史阶段流感爆发情况，以及对应时期的关键词、频率，判断哪些因素的相关性大，反复试错、调整，直到建立可靠模型。

微软猜准多项奥斯卡奖，也是相同套路——分析大量历史数据，找准多个相关因素，不断调试，建模型。说来简单，实际操作很复杂。

记者：最近中国股市行情惨淡，能借大数据之力预测涨跌吗？

王晓阳：按我的理解，股市可预测，但有效期很短，超出几秒甚至几毫秒，就不准了。天气预报也类似，报 7 天后的，基本没谱；报明天的，命中率相当好；如果报几分钟后的，更准。

华尔街有些公司，专在毫秒间买进卖出。时间短，对机器运行速度要求很高。公司就设在交易所旁边，以保证数据线尽可能短。

但对绝大多数股民来说，股市不能预测，因为等你买到手时，早过预测有效期了。

记者：大数据时代，机器的性能越来越强，人的作用还剩多少？

王晓阳：业内有观点认为人不重要，交给机器处理就行。我不赞同。数据庞杂而冰冷，需要人的智慧去激活。

美国人 Nate Silver 擅长预测，有“大数据时代的巫师”之称。去年美国总统大选，他预言奥巴马有 90.9%的机会获胜，对全美 50 个州投票结果的预测全对。2008 年总统大选，他也猜对了。

Silver 不是靠蒙，而是收集众多民调数据，综合分析多方面因素。为什么准？因为他有独特洞察力，在纷繁复杂的数据里，能找准关键变量，建立靠谱模型。这是机器不能代替的。

大数据对人提出了更高要求。以前做抽样调查，预设问卷，再有针对性地分析结果，套路清楚、易学。现在数据铺天盖地，选哪些来支持预测和判断，如何建模型，都比原来难。

我想强调“概率”。Silver 预测再准，也从不讲绝对，而是用百分比说话。有不确定性，所以更需要依靠人的判断。假如今天降水概率 70%，你出门带不带伞？

“要彻底保护隐私，数据就没用了”

记者：社交媒体在大数据时代扮演什么角色？是否不可替代？

王晓阳：微信、微博、论坛等社交媒体，是给社会搭脉的重要途径。因为在这些地方，人们能充分表达情感。“我在哪，买了一杯咖啡，好不好喝，天气如何，心情怎样……”与态度、情绪有关的数据，只有这里能采集到。

媒体在街头做民意调查，问路人“你幸福吗”，获得的数据可能不准，倒不如去微博里寻找人们自发表达的内容。

记者：大数据正被越来越多商家利用，对消费者是好是坏？

王晓阳：如果大数据被商家正确使用，消费者确有需求，当然觉得方便；反之就成了骚扰。一些购物网站常根据消费行为推荐商品。比如有人买了一本孕期保健书，网站几个月后便频频推荐尿布和奶粉。其实她是替朋友买的，与个人需求无关。

这不是大数据本身的问题，而是对数据的分析出了错，脉搏没搭清楚，就会带来负面影响。这也反映了人的智慧在分析数据时所起的重要作用。

记者：大数据时代对个人隐私的侵犯似乎防不胜防，如何规避数据开放带来的风险？

王晓阳：保护隐私确实是大数据面临的难题。有人曾想证明一个命题：能否既保护隐私，又保证数据有用？结果发现做不到，要彻底保护隐私，数据大多就没用了，两者之间没有平衡点。如果完全关闭手机位置信息，确实保住了隐私，但导航等软件就不能用了。

保护隐私不是不能做，而是要在采集数据的环节及采集到数据后同时做。采集数据时应透明，要让用户知道什么数据在被采集，并征得用户同意。其实最严重的问题出在采集数据之后，谁可以用这些数据？如何用？现状是，采集者掌握众多隐私数据后，随意贩卖，给许多人带来困扰，甚至对大数据产生恐慌。

我国这方面法律不健全，必须加快立法。采集到隐私数据后，谁能用，能怎么用，如何告知消费者，违规了怎么处罚……都需法律规范。

●国内邮发代号3–5／国外发行代号D694／全国各地邮局均可订阅／广告经营许可证号：3100020050030／社址：上海市威海路755号／邮编：200041／总机：021–52921234转各部
●本报印刷：文新集团印务中心等，在国内外9个印点同时开印／上海沪太、上海龙吴、上海金桥、上海界龙、崇明／北京、深圳、香港、美国洛杉矶
●本报在23个国家地区发行海外版／美国、澳大利亚、加拿大、俄罗斯、西班牙、泰国、菲律宾、日本、法国、巴拿马、意大利、荷兰、南非、匈牙利、新西兰、罗马尼亚、尼日利亚、印度尼西亚、阿联酋、英国、德国、希腊等
24小时读者服务热线 962288
本报零售价 每份 1.00元