大数据时代，统计学还有用吗？清华大学统计学研究中心主任刘军做客《文化讲坛》

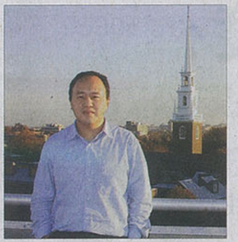
大数据是“原油”，不是“汽油”

易 林 葛亮亮

文化讲坛

在数据“爆炸”的时代，大数据常常被寄予厚望。到底，什么样的数据才算大数据，怎样才能用好大数据，传统统计学还有用武之地吗？清华大学统计学研究中心前不久成立，著名统计学家、哈佛大学终身教授刘军担任主任。日前，刘军做客人民日报、人民网《文化讲坛》，分享他的思考。

——编 者



让大数据区别于数据的，是其海量积累、高增长率和多样性

什么是数据？数据(data)在拉丁文里是“已知”的意思，在英文中的一个解释是“一组事实的集合，从中可以分析出结论”。笼统地说，凡是用某种载体记录下来的、能反映自然界和人类社会某种信息的，就可称之为数据。古人“结绳记事”，打了结的绳子就是数据。步入现代社会，信息的种类和数量越来越丰富，载体也越来越多。数字是数据，文字是数据，图像、音频、视频等都是数据。

什么是大数据呢？量的增多，是人们对大数据的第一个认识。随着科技发展，各个领域的数据量都在迅猛增长。有研究发现，近年来，数字数据的数量每3年多就会翻一番。

大数据区别于数据，还在于数据的多样性。正如高德纳咨询公司研究报告指出的，数据的爆炸是三维的、立体的。所谓的三维，除了指数据量快速增大外，还指数据增长速度的加快，以及数据的多样性，即数据的来源、种类不断增加。

从数据到大数据，不仅是量的积累，更是质的飞跃。海量的、不同来源、不同形式、包含不同信息的数据可以容易地被整合、分析，原本孤立的数据变得互相联通。这使得人们通过数据分析，能发现小数据时代很难发现的新知识，创造新的价值。

通过数据来研究规律、发现规律，贯穿了人类社会发展的始终。人类科学发展史上的不少进步都和数据采集分析直接相关，例如现代医学流行病学的开端。伦敦1854年发生了大规模的霍乱，很长时间没有办法控制。一位医师用标点地图的方法研究了当地水井分布和霍乱患者分布之间的关系，发现有一口水井周围，霍乱患病率明显较高，借此找到了霍乱暴发的原因：一口被污染的水井。关闭这口水井之后，霍乱的发病率明显下降。这种方法，充分展示了数据的力量。

本质上说，许多科学活动都是数据挖掘，不是从预先设定好的理论或者原理出发，通过演绎来研究问题，而是从数据本身出发通过归纳来总结规律。近现代以来，随着我们面临的问题变得越来越复杂，通过演绎的方式来研究问题常常变得很困难。这就使得数据归纳的方法变得越来越重要，数据的重要性也越发凸显出来。

大数据是非竞争性资源，有助于政府科学决策、商家精准营销

大数据时代，数据的重要作用更加凸显，许多国家都把大数据提升到国家战略的高度。

政府合理利用大数据，引导决策的将是基于实证的事实，政府会更有预见性、更加负责、更加开放。中国古代治国就已经有重数据的思想，如商鞅提出，“强国知十三数……欲强国，不知国十三数，地虽利，民虽众，国愈弱至削”。大数据时代，循“数”治国将更加有效。小数据时代，政府做决策更多依凭经验和局部数据，难免头痛医头、脚痛医脚。比如，交通堵塞就多修路。大数据时代，政府做决策能够从粗放型转向集约型。路堵了，利用大数据分析，可以得知哪一时间、哪一地段最容易堵，或在这一地段附近多修路，或提前预警引导居民合理安排出行，实现对交通流的最佳配置和控制，改善交通。

对于商家来说，大数据使精准营销成为可能。一个有趣的故事，是沃尔玛超市的“啤酒、尿布”现象。沃尔玛超市分析销售数据时发现，顾客消费单上和尿布一起出现次数最多的商品，竟然是啤酒。跟踪调查后发现，有不少年轻爸爸会在买尿布时，顺便买些啤酒喝。沃尔玛发现这一规律后，搭配促销啤酒、尿布，销量大幅增加。大数据时代，每个人都会“自发地”提供数据。我们的各种行为，如点击网页、使用手机、刷卡消费、观看电视、坐地铁出行、驾驶汽车，都会生成数据并被记录下来，我们的性别、职业、喜好、消费能力等信息，都会被商家从中挖掘出来，以分析商机。

大数据也将使个人受益。从生物学、医学上讲，以前生物学家只是通过对单个或几个基因的操控来观察其对生物体的影响，很难发现整体的关联。现在由于技术的发展，可以分析很多，如遗传信息、全体基因的表达量信息、蛋白质族谱信息、全基因组甲基化信息、表观遗传信息等。同时还有个人健康指标、病历、药物反应等数据。如果真能达成生物学上多维多向数据的有机融合，就能够把个人完整地描述出来，从而实现精准医疗的目的。

大数据时代，审核数据的真实性也有了更有效的手段。大数据的特征之一是多样性，不同来源、不同维度的数据之间存在一定的关联度，可以交叉验证。例如，某地的工业产值虚报了一倍，但用电量和能耗却没有达到相应的规模。这就是数据异常，很容易被系统识别出来。发现异常后，相关部门再进行复核，就能更有针对性地防止、打击数据造假。

数据是一种资源，但数据又跟煤、石油等物质性资源不一样。物质性资源不可再生，你用多了，别人就用少了，因而很难共享。数据可以重复使用、不断产生新的价值。大数据资源的使用是非恶性竞争的，共享的前提下，更能够制造双赢。从另一个角度来说，数据如果不被融合、联系在一起，也不能称之为大数据。

大数据不能被直接拿来使用，统计学依然是数据分析的灵魂

现在社会上有一种流行的说法，认为在大数据时代，“样本=全体”，人们得到的不是抽样数据而是全数据，因而只需要简单地数一数就可以下结论了，复杂的统计学方法可以不再需要了。

在我看来，这种观点非常错误。首先，大数据告知信息但不解释信息。打个比方说，大数据是“原油”而不是“汽油”，不能被直接拿来使用。就像股票市场，即使把所有的数据都公布出来，不懂的人依然不知道数据代表的信息。大数据时代，统计学依然是数据分析的灵魂。正如加州大学伯克利分校迈克尔·乔丹教授指出的，“没有系统的数据科学作为指导的大数据研究，就如同不利用工程科学的知识来建造桥梁，很多桥梁可能会坍塌，并带来严重的后果。”

其次，全数据的概念本身很难经得起推敲。全数据，顾名思义就是全部数据。这在某些特定的场合对于某些特定的问题确实可能实现。比如，要比较清华、北大两校同学数学能力整体上哪个更强，可以收集到两校同学高考时的数学成绩作为研究的数据对象。从某种意义上说，这是全数据。但是，并不是说我们有了这个全数据就能很好地回答问题。

一方面，这个数据虽然是全数据，但仍然具有不确定性。入校时的数学成绩并不一定完全代表学生的数学能力。假如让所有同学重新参加一次高考，几乎每个同学都会有一个新的成绩。分别用这两组全数据去做分析，结论就可能发生变化。另一方面，事物在不断地发展和变化，同学入校时的成绩并不能够代表现在的能力。全体同学的高考成绩数据，仅对于那次考试而言是全数据。“全”是有边界的，超出了边界就不再是全知全能了。事物的发展充满了不确定性，而统计学，既研究如何从数据中把信息和规律提取出来，找出最优化的方案；也研究如何把数据当中的不确定性量化出来。

所以说，在大数据时代，数据分析的很多根本性问题和小数据时代并没有本质区别。当然，大数据的特点，确实对数据分析提出了全新挑战。例如，许多传统统计方法应用到大数据上，巨大计算量和存储量往往使其难以承受；对结构复杂、来源多样的数据，如何建立有效的统计学模型也需要新的探索和尝试。对于新时代的数据科学而言，这些挑战也同时意味着巨大的机遇，有可能会产生新的思想、方法和技术。

看人民映像，品百味人生

我科学家揭示葡萄糖转运蛋白的工作机理

本报北京7月19日电 （记者**赵永新、赵婀娜**）继去年在世界上第一次解析出人源葡萄糖转运蛋白GLUT1的三维晶体结构后，清华大学颜宁研究组又获得重大进展：获得人源葡萄糖转运蛋白GLUT3处于不同构象的3个高分辨率晶体结构，并通过与GLUT1的结构比对，完整揭示出葡萄糖转运蛋白底物识别与转运的分子机理，为基于结构的小分子肿瘤药物设计提供了直接依据。

7月15日，《自然》杂志在线发表了相关文章《葡萄糖转运蛋白识别与转运底物的分子基础》，文章共同第一作者为清华大学生命科学联合中心博士后邓东、一年级博士生孙鹏程，通讯作者为颜宁。

葡萄糖是地球上各种生物最主要的能源物质，不仅为生长代谢提供能量，还参与合成其他生命组成大分子。葡萄糖分子进出细胞，需要依靠膜上的“搬运工”——葡萄糖转运蛋白GLUTs，人体中的GLUTs共有14种，目前研究比较清楚的是GLUT1、2、3、4（简称GLUT1—4）。过去8年来，颜宁研究组一直致力于系统揭示GLUTs的结构与工作机理。2014年初，研究组解析出人源GLUT1的三维结构，成为世界上第一个人源次级转运蛋白晶体结构。

据介绍，GLUT1的结构处于向胞内开放的状态，只是GLUTs蛋白在行使转运功能过程中众多构象的一种。为揭示整个转运过程，获得GLUTs处于其他不同转运状态下的结构信息至关重要；为理解GLUTs识别底物（参与生化反应的物质）的机理，则需获得它们与底物结合的复合物结构。颜宁研究组由此精心设计实验，利用膜蛋白脂立方相结晶和微聚焦X—射线衍射，最终解析出GLUT3处于3种不同状态的高分辨率晶体结构。其中GLUT3与底物分子葡萄糖的复合物晶体结构处于向胞外闭合的状态，分辨率高达1.5埃（1埃等于0.1纳米），是目前为止分辨率最高的转运蛋白结构。这一超高分辨率首次清晰地表明，GLUT3可以识别葡萄糖α和β两种异构体。

为进一步得到GLUT3蛋白向胞外开放的构象，研究组通过加入竞争性抑制分子麦芽糖或者纤维二糖，并最终获得了GLUT3与麦芽糖结合的向胞外开放和向胞外闭合两种状态下的晶体结构。至此，通过进一步比较GLUT1向胞内开放的结构，GLUTs家族蛋白通过交替开放转运底物的整个过程已基本清晰。

用火遗物遗迹密集出现

“北京人”用火再添力证

本报北京7月19日电 （记者**余荣华**）记者近日从北京市文物局获悉：“北京人”有控制的用火行为再添力证。在北京周口店北京人遗址第一地点（猿人洞）2009年—2014年的清理发掘过程中，古人类用火遗物、遗迹密集出现。

据悉，火塘、原地烧结土、烧石、烧骨等古人类用火遗物、遗迹的密集出现，为“北京人”用火行为的研究提供了[illegible]全炭化，内外都呈黑色，可判断是火烧的结果。另外还发现有石灰岩被烧成白色的石灰，而这只有在高温且持续燃烧的情况下才会发生。同时，还发现可以被称为火塘的两三处集中用火的部位，在火塘内部及周围，有大量燃烧过的物质，其灰烬经分析测试，因燃烧而形成的富集的元素碳、钾和植硅体在灰烬样品中有足够数量的存在，证明曾发生过原地燃烧（用火）行为。上述发现表明，“北京人”已经会用火，且能保留火种，是在有控制地用火。

暑期电影票房连破多项纪录

本报北京7月19日电 （记者**刘阳**）近日，中国电影票房连破多项纪录——10日至19日突破11亿元，刷新周末票房最高纪录；18日票房突破4.17亿元，刷新单日票房最高纪录；18日全国观影人次突破960万次，刷新单日观影人次纪录；国产影片《捉妖记》首日票房突破1.6亿元，刷新国产片首日票房纪录。在以国产片为主导力量的暑期档中，《捉妖记》等多部国产片的口碑不断发酵，票房持续走高。业内人士表示，好口碑带来高票房在这个暑期档的国产影片中得到充分体现。

反法西斯电影研讨会举行

本报北京7月19日电 （**巩育华、刘启民**）由中国人民大学文学院、《当代电影》杂志社联合主办的“可见的左翼”——纪念夏衍逝世20周年暨30年代反法西斯电影研讨会18日在北京拉开帷幕。在为期两天的会程中，黄会林、戴锦华、张建勇等学者就夏衍研究、左翼文化和反法西斯电影等3个议题展开热烈讨论，回顾了夏衍波澜壮阔的一生、中国左翼电影人卓越而艰苦的奋斗历程。

上个世纪30年代，在左翼社会思潮的背景下，中国共产党团结领导电影界进步人士，协同摄制了《渔光曲》等一批宣扬反帝反封建的影片，反映大众疾苦、倡导社会公正、追求时代进步。以夏衍为代表的中国左翼电影人，不仅将反法西斯的阵地扩展到新兴的电影领域，同时拓展了电影这一表现形式的精神和艺术追求，使左翼电影成为中国电影史上的一座丰碑。

中国微电影工作委员会成立

本报北京7月19日电 （记者**任姗姗**）中国电影家协会微电影工作委员会成立揭幕仪式日前在北京举行。微电影工作委员会会长谷国庆在成立大会上表示，由陈凯歌担任名誉会长的微电影工作委员会，将在中国影协领导下，秉承规范、交流、合作、创新、发展、服务的宗旨，广泛联系国内外微电影创作人员、相关学者，建设、交流微电影学术理论，建立行业规范，引导创作方向，为中国影视创作不断注入新鲜血液。

北京市本科一批录取工作启动

高分落档现象大为减少 考生志愿梯度较为合理

本报北京7月19日电 （**赵婀娜、朱可然**）记者在北京市高招录取开放日上获悉：北京市2015本科一批录取工作已于7月17日启动，将于7月22日结束。此前，北京市已完成零批次、本科提前批次和特殊类型录取工作。

据悉，今年共有167所院校在京参加一本招生，计划招生14861人。其中，文史类计划2748人，理工类计划12113人。目前已经投档15610人，其中文科2845人，理科12765人。

今年北京市首次实施大平行志愿，减少了考生志愿填报风险，增加了录取机会，改革取得预期成效。截至目前，招生计划完成情况良好，本科一批仅剩余计划160多个；高分落档的现象大为减少；考生志愿梯度较为合理，在6个平行志愿中，前4个志愿高校的满足率接近80%；招生高校对生源情况表示满意。

截至目前，北京市普通高校招生零批次和本科提前批次录取工作已经完成，242所院校参加录取，共录取考生8889人，其中零批次录取319人，本科提前批次录取8570人。今年北京市还首次在市属高校实施了“双培计划”和“外培计划”招生，一本高校（专业）、城六区的计划完成情况良好，二本高校（专业）以及郊区县的生源略显不足。

北大、清华两所高校也对外公布了在京录取结果。记者从北京大学招办了解到，今年的高分考生不再扎堆所谓热门专业，比如今年北京文科第一名、来自北师大二附中的蔡雨珐心仪“外国语言与外国历史”专业，高考取得710分、来自北京师范大学附属实验中学的王佳萌首选专业是“心理学”。记者从清华大学招办了解到，今年的北京考生在选择专业时大多从自身兴趣出发，眼光长远，不盲目跟从社会潮流，专业填报分布比较合理。

“木绘拼花”手工技艺让故宫藏画走近大众

本报北京7月19日电 （记者**时圣宇**）19日，故宫博物院下属的北京故宫文化传播有限公司与河北百年巧匠手工艺品股份有限公司签订合作协议，双方将在故宫博物院专家指导下，运用“木绘拼花”手工技艺，共同研发出品故宫藏画木制工艺品，使深藏故宫数百年之久的皇家藏画在新工艺、新载体上焕发新的生机，让更多公众能把故宫文化带回家。

签约仪式上，故宫珍藏乾隆皇帝御笔《岁朝图》组画复制版正式亮相。据了解，本次双方共同研发、联合出品《岁朝图》木绘版产品，既是京津冀文创产业协同发展的一次重要成果，又是故宫文创产品研发的一次有益尝试。

油画里的新疆

19日，在新疆博物馆的展厅内，一位老人正带着孙子观看油画作品。

当日，由新疆文化厅主办、新疆油画学会承办的第四届中国新疆国际民族舞蹈节系列活动之“十年·新疆油画学术展”在新疆博物馆开展。本次展出新疆本地51名油画家的54幅油画作品，呈现了10年来新疆油画创作的丰硕成果与发展历程。

新华社记者 王 菲摄

万名中学生入驻高校科学营

本报北京7月19日电 （记者**余建斌**）19日，由中国科协和教育部主办的2015青少年高校科学营活动北京分营在北京大学开营，标志这项将有全国1万多名中学生参加的科学活动全面展开。

2015年青少年高校科学营活动以“走进科学殿堂，放飞科学梦想”为口号，包括由高校承办的常规营50个，高校与企业、科研院所联合承办的专题营15个及西部试点营1个，将为来自31个省、市、自治区和新疆生产建设兵团以及港澳台地区的10210名在校高中生提供全新科学探索之旅，将在全国范围为青少年举办300场名师科学讲座，开放300个重点实验室，开展各领域科技实践活动上千次。