

构建大数据产业环境

专访中国工程院院士李国杰

文/本刊记者 牛禄青

日前，中国工程院院士、中科院计算所首席科学家李国杰就大数据的发展背景、重大意义、最新动向、未来趋势，以及中国的机遇与挑战等相关问题接受了《新经济导刊》记者的专访。

李国杰表示，大数据对经济社会发展和科学研究具有革命性的意义，其兴起有着内在的需求和利益驱动，因为数据里蕴藏着巨大的价值。未来将形成数据服务、数据探矿、数据化学、数据材料、数据制药等一系列战略性新兴产业。

李国杰认为，数据安全主要不是技术问题，因为数据放在哪里都有泄露的风险，它与商业模式有很大关系。中国当务之急是建立上下游相互协作、相互支撑的大数据产业环境，特别是构建有技术自主权的大数据产业链。

商业价值驱动

《新经济导刊》：继物联网、云计算、3D 打印等新技术之后，大数据已成为投资者、IT 人士以及政府部门、科研人员关注的热点，请问大数据是在什么背景下发展起来的？

李国杰：今年三月份，奥巴马宣布美国政府投资 2 亿美元启动“大数据研究和发
展计划”，这个计划可以同美国上世纪 90 年代初的“信息高速公路”相比拟。美国政
府认为，大数据是“未来的新石油”，并对将大数据的研究上升为国家意志。

表面上看“大数据热”受到美国计划的影响，但不完全是这样。过去美国副总统
戈尔呼吁的数字地球，中国也在跟进，但经济上并未成气候；有些是美国不太热，中
国反而进行得有声有色，比如物联网，这可能与中国政府对物联网的大力扶持有关。
所以层出不穷的新技术，有时候是真热，有时候是虚热。

我认为，大数据不是因为奥巴马的宣布而热起来的，它的兴起有着内在的原因。
~~过去的科学实践大数据不可能引起广泛关注，~~现在的大数据与互联网、物联网的蓬勃
发展有很大关系，特别是美国几家大型企业的大力推动，像 IBM、亚马逊、Google、

Facebook 等。

科研人员研究大数据，习惯于从数据到信息到知识再到智慧，若按照这个链条，时间太漫长，产生的经济效益也有限，大数据不会形成这么大势头。现在企业走了一条捷径，直接从数据里开发出商业价值，而不管数据中的科学规律和知识，这可以大大激发企业的兴趣。比如电子商务 eBay，它用大数据分析网络广告，发现广告里的每一个单词都与经济效益有关联，通过优化，使广告收益提高 80% 以上。所以，大数据兴起的根本原因是里面蕴藏着巨大的价值，有实实在在的经济利益驱动。

《新经济导刊》：人们对大数据的理解，见仁见智，如何界定大数据？大数据有哪些基本特征？

李国杰：一般意义上，大数据是指无法在可容忍的时间内用传统 IT 技术和硬件工具对其进行感知、获取、管理和服务的集合。大数据的特点主要表现为四个“V”：一是体量浩大 (Volume)，数据集合的规模已从 GB 到 TB 再到 PB 级，甚至已经开始以 EB 和 ZB 来计算。著名咨询公司 IDC 的研究报告称，未来 10 年全球大数据将增加 50 倍，管理数据仓库的服务器的数量将增加 10 倍。二是类型复杂 (Variety)，大数据类型包括结构化数据、半结构化数据和非结构化数据。现代互联网应用呈现出非结构化数据大幅增长的特点，到 2012 年末非结构化数据将达到整个数据量的 75% 以上。三是生成迅速 (Velocity)，大数据通常以数据流的形式动态、快速地产生，具有很强的时效性。数据自身的状态与价值也随时空变化而发生演变，数据的涌现特征明显。四是价值巨大但利用密度低 (Value)，基于传统思维与技术让人们在实际环境中面临信息泛滥而知识匮乏的窘态。

开启数据革命

《新经济导刊》：大数据对经济社会发展和科学研究有哪些重要作用？

李国杰：根据数据的来源，大数据可以分为两类：一类来自与人类社会有关的数据，特别是互联网和经济活动产生的数据，企业最感兴趣，这一块增长也最快，互联网实际上反映的是人的活动；另一类来自物理世界，通过传感器、科学观测获取，比如生物数据、脑科学数据、气象数据、野外环境保护数据、卫星遥感数据等。这类数据首先推动科学的进步，继而推动经济的发展。

大数据具有革命性的意义，作为一种重要的战略资源，不仅事关国家的数字主权和战略安全，而且可以促进我国的经济结构调整和产业升级。大数据时代，企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为大

数据处理能力，软件也将从编程为主转变为以数据为中心。大数据处理的兴起也改变了云计算的发展方向，使其进入以分析即服务（AaaS）为主要标志的 Cloud2.0 时代。采用大数据处理方法，生物制药、新材料研制生产的流程会发生革命性的变化，大大提高科研和生产效率，使整个行业迈入数字化与信息化的新阶段。未来将形成数据服务、数据探矿、数据化学、数据材料、数据制药等一系列战略性新兴产业。数据服务是许多企业瞄准的重要领域，华为本来是一家通信企业，现在也开拓数字医药医疗市场，打通社区医院和三甲医院，这是一个潜在的巨大市场。

此外，大数据正在引发科学研究思维与方法的一场革命。最早的科学研究只有实验科学，随后出现了以研究各种定律和定理为特征的理论科学。由于理论分析方法在许多问题上太过复杂，难以解决实际问题，人们开始寻求模拟的方法，导致计算科学的兴起。海量数据的出现催生了一种新的科研模式，科研人员只需从数据中直接查找或挖掘所需要的信息、知识和智慧，甚至无需直接接触所研究的对象。2007 年，已故图灵奖得主吉姆·格雷在他最后一次演讲中描绘了数据密集型科学研究的“第四范式”，把数据密集型科学从计算科学中单独区分开来。第四范式不仅是科研方式的转变，也是人们思维方式的大变化。现实中，许多复杂的经济社会问题无法用传统的方法进行研究，比如股市，没有大量的数据无法找到其中的规律，仅靠模型是不行的。

《新经济导刊》：大数据对地理信息和位置服务有何影响？

李国杰：地理信息和位置信息是人在物理世界的活动，是社会活动和物理世界两类数据的融合。这种数据的规模是相当大的，是未来一个新的经济增长点，而且能带动单独在物理空间和单独在社会网络领域的数据产业的飞速发展。

地理信息是国家信息基础设施的一部分，电子政务、电子商务、智能交通、智能物流等行业应用，都离不开地理信息。政府对地理信息这种公共资源应该分层管理，除了涉及国防和国家安全的信息需要保密外，其他基础信息都要免费或以较低成本提供给企业和民众，防止信息采集单位据为己有。基于基础信息上的各种应用服务，应该大量放开，让企业去做。我国的地理信息产业之所以发展不快，就是没有解决好信息的分层管理，收集地理信息的单位，理所当然地认为这些信息就是本部门的，而且互相之间也不交流和共享，导致重复建设和资源浪费。

颠覆 IOE 模式

《新经济导刊》：大数据时代，传统的数据库软件将难以满足处理海量数据的需要。您认为大数据对信息技术提出哪些新的要求？

李国杰：过去，中国对信息系统有所谓“金三角”的说法，即“IOE”，I 指 IBM 的服务器，O 指 Oracle 的数据库，E 指 EMC 的存储。这三家公司基本垄断了国内银行、证券等对计算机处理数据要求很高的行业。现在业内认为，如果某家技术公司还沿用“IOE”，那这家公司就不是大数据公司。因为你的技术建立在 ~~Oracle 的 IOE~~ 基础上，决定了你无法处理大量数据，~~仅这一条，因为~~你的基础设施就不适合做大数据 处理。

数据在基础设施层面有三类技术：存储、管理和计算。IOE 模式从上世纪 70 年代就逐步成型了，即关系数据，当时对数据的认识是线性维度。到了大数据时代，数据的巨大规模和快速变化超过了硬件能力的增长，而且数据之间的复杂关联 和再生使得非线性思维所能及无能为力。另外还有社会因素，数据与人在不断地互动，甚至人就是动态的数据集。在这种情况下，采用原有的 IOE 模式来处理大数据就难以应对了，数据中的价值也无法有效挖掘出来。现在大数据有许多应用，例如通过查询 Google，可以知道流行病在某区域的分布；通过舆情的分析，可以预测选举的结果。

大数据存储不同于传统的 EMC，它是一种高效率、低成本、多层次柔性的存储架构，不是集中到一起存储，而是把数据放在云和端。实际上，数据存储的分布很关键，哪些在主服务器，~~哪些在数据块服务器~~，哪些在客户端，这要比提高单个存储的性能更重要。虽然许多企业都在做，但还处于探索中。目前，Hadoop 开源分布式系统，已成为大数据处理的主流技术，包括资源调度、存储管理等各种数据工具。这是一种发展趋势，但也存在不少问题。

在大数据管理方面，过去的关系数据管理在冗余、一致性、复杂查询优化上解决得非常好，在 Hadoop 体系下，数据规模、增量速度和灵活性上远远超过 Oracle，但在数据价值挖掘和分析方面 还不能提供较好的技术支持。所以大数据管理是打碎了旧的模式，新的模式还没有规范化和体系化。

《新经济导刊》：中国与国外大数据产业相比，有哪些优势和劣势？国外互联网企业在大数据处理方面有哪些经验可供我们借鉴？

李国杰：不管是 Google、亚马逊还是 Facebook、Twitter，肯定不是用 IOE 这种模式，都是重新设计和建立新的系统。比较领先的应该是 Google，但 Google 把自己的技术包裹起来，不告诉别人。所以，后来形成了以 Hadoop 为代表的一系列开源技术。虽然 Hadoop 借鉴了 Google 的一套大数据处理思路：GFS 存储、MapReduce 计算、BigTable 管理，但因为是全球人的贡献，所以全球 60%-70% 的大型互联网企业都在使用 Hadoop 开源技术。

国内的大数据代表性企业是百度、腾讯、阿里巴巴。由于大数据原创体系和基础

技术的话语权在国外（开源组织和大企业），所以国内企业在这方面有一定差距，但应用上不比国外企业差。需要注意，国内大型互联网企业是自己确实有需求和价值驱动才改进原来的技术架构，而不是一味的跟风。比如淘宝网，每天的日志按照传统方法只能存储一周，数据量太大，不可能为了存储数据再盖大楼和机房，这就逼迫它必须优化存储、提高效率和节约成本。

由于中国人口多、市场大、数据量大，所以中国大数据发展的动力非常强劲，大数据的应用需求丝毫不亚于国外。同时，我们也要看到，中国发展新兴产业是一个后来者，以前经常讲，我们有后发优势，能够避免走弯路。实际上，前面所走过的二十年，更多暴露出后发的劣势和壁垒。比如桌面计算机(PC)，技术掌握在因英特尔和微软手里，这就形成了它们的一个事实标准，必须用它们的平台来做，而我们争取创新的空间很小，联想在奋起直追，毛利率也是 15%以下，净利率只有 2%。通信也是这样，2G 和 3G 的无线通信专利掌握在高通公司手里，虽然我们可以做 4G 等新技术，但必须与 2G 和 3G 兼容，只要兼容就无法摆脱高通的控制。这两个大产业，虽然我国花了很大力气，但在平台上受制于人，创新空间一直很小。

大数据也不是没有限制，但大数据是在比较开放的环境下发展的，尽管 Google 不很开放，但相比 PC 和无线通信领域还是有利的。中国科研机构和企业研发的大数据开源软件，也可以加入到国际标准中去，成为世界开源组织大家庭的一部分。比如中科院计算所牵头的 Hadoop in China 大会就, 现在转移给中国计算机学会大数据专家委员会承办, 获得子就一直得到世界开源组织的认可支持。大数据时代，不是国际大数据企业愿意主动贡献 Hadoop 等数据处理技术，而是大数据产业特征所决定的了必须协作共享。大数据无所不在，不是几个大企业就能垄断得了。

《新经济导刊》：在 IT 领域，包括大数据在内，为什么新的技术都是国外先有，然后再引入到国内？什么时候“中国创新”能引领世界呢？

李国杰：总体来讲，中国的信息技术水平处于世界第二军团的前沿，与美国、日本等发达国家相比还有一定差距。如果中国的信息技术水平已经与发达国家并驾齐驱，那中国还是一个发展中国家吗？信息技术是当代的一个特征技术，一个国家是不是经济强国，不能仅拿制造业来衡量，关键看信息技术和生物技术。中国到 2020 年才能全面建成小康社会，到 2050 年才能实现现代化。科技不可能在国家综合实力还落后的情况下一枝独秀，当然不排除个别技术脱颖而出，但整体上还是受制于经济和社会发展的状况。

我国在发表论文方面，个别领域已经位居前列，国外的引用也较多，最典型的是

材料科学。世界前十位材料科学论文引用率最高的作者，60%-70%都来自于中国大陆，按说中国应该是一个材料强国了，但实际情况并非如此，80%-90%的尖端材料全部是进口。所以我们对中国的科技水平期望太高，毕竟中国还是一个发展中国家，科技与经济是相辅相成的，必须有市场需求，才有科技创新，信息技术同样如此。

建立数据市场

《新经济导刊》：面对纷繁复杂、无处不在的数据，中国在大数据管理和应用过程中如何确保数据安全呢？

李国杰：大数据时代，安全是一个基础保障，但如果建立一个竞争有序的大数据交易市场，将大数据打包成产品依法进行交易，那所谓的数据隐私问题就可以规范化了。现在数据市场还未成型的情况下，那从顶层设计上要注意保障数据安全，包括隐私权、执行权、防范数据篡改和崩溃、可信度等一系列问题。

但数据安全主要不是技术问题，因为数据放在哪里都有泄露的风险，它与商业模式有很大关系，中国迫切需把数据市场、数据产业、数据产品的形态和交易模式清晰化。这些问题解决了，数据安全也就迎刃而解。现在最大的问题是，拥有原始数据的机构和企业，不知道如何把这些数据变成产品。

中国数据市场的建立可以借鉴金融衍生品市场的模式，一是政府出台优惠措施加以扶持；二是建立透明公开的交易平台；三是加强创新，突破关键技术；四是发挥资本市场的作用。这样就可以把各个载体的数据开发出一系列数据产品。目前产业界和投资界走在前面，科技界紧跟其后，政府还没有认识清楚。

实际上，大数据现在仅仅是冰山一角，它的巨大价值还远远没有挖掘出来，人们也确实不知道它的价值到底有多大。国家要创造一个支持新兴业态的环境，让新产品和新服务能够喷薄而出。有些东西应该冒出来而未能冒出来的原因就是，被负面的东西和可能产生的负作用吓怕了，比如安全和隐私。任何事物都是一把双刃剑，关键看你的着眼点和出发点，如果只是除弊而不兴利，那就越除越小，最终抑制它的成长。所以要正确看待新兴事物，并在发展中解决存在的问题，而不是一棒子打死。例如，互联网和手机，十年前谁也无法预料到能发展成现在这样，都是在市场竞争中不断优胜劣汰发展起来的。当然，新技术和新兴产业刚开始都有一个炒作过程，然后逐渐挤掉泡沫，沉淀下来，从而步入正常发展轨道。

《新经济导刊》：请介绍一下中国的大数据产业链建设情况？

李国杰：IT 产业在发展过程中已经形成了一些层次分布，有做服务器和底层系统

的，有做软件的，有做应用的，大数据也需要在原有的架构上加以发展。原来做基础设施的企业，如联想、华为，也要向大数据转型，提供低成本、低能耗的大型存储器，这是大数据产业的基础。中间层是类似 Hadoop、MapReduce 的数据分析软件，原有的软件产业也要转型，由卖软件转为以数据为中心。再往上就是百度、腾讯、阿里巴巴等大数据应用服务公司。

中国大数据产业的整体实力与国外相比有很大差距。应用企业由于服务中国市场，具有民族和语言等方面的优势，所以市场占有率较高，但需要走向国外，提高国际竞争力。软件实力较弱，基础设施更弱，芯片和操作系统还在成长中。所以，国家要针对大数据的需求，尽快出台政策措施，全面提高大数据产业的竞争实力。企业要抓住机遇，加大自主创新力度，实现转型发展，扭转国际竞争中的不利地位。科研机构要瞄准国际前沿，大胆尝试，积极探索。中科院计算所已经在研发五年以后的服务器和计算机了，我们如果研发成功，将会缩小国际差距。

《新经济导刊》：您认为中国大数据产业的难点和瓶颈是什么？

李国杰：当务之急是建立上下游相互协作、相互支撑的大数据产业环境，特别是构建有技术自主权的大数据产业链，避免核心技术受制于人，重蹈 PC 和通信产业的老路。发展大数据产业，还需要突破一个瓶颈，那就是宽带网络的滞后。没有高速的宽带网络做支撑，大数据将成为“空中楼阁”。所以，信息产业的“短腿”要赶紧补上。

新的应用需求呼吁新的人才，但我们的教育落后于经济和社会的发展。大数据时代，需要从学校 and 实践中培养各类数据人才，如数据科学家、首席数据官、数据咨询师、数据分析师、数据工程师等。特别是数据咨询人才，要大力培养，加快数据咨询产业发展。另外，培养大数据人才，要打破专业限制，取长补短，除了传统的计算机、电子信息专业，还应该更多从各行业中培养熟悉本行业的数据人才，教会他们从行业数据中挖掘价值。学计算机的人要放下身段，甘当配角，主角由行业人才来担当。避免懂数据分析的没有数据，不懂数据分析的却拥有大量数据。（中科院计算所副总工程师、网络数据科学与技术实验室主任程学旗研究员对本文也有贡献）