

科学大数据驱动学科发展

中国科学院院士 郭华东

上个月 IDC 的“数字宇宙”研究报告指出，全球数据总量将以每两年翻一番的速度持续增长。去年只有 20%的数据与云相关，几年之后将翻倍至 40%，这是一个非常重要的数据。如今在全球数据的分布中，中国占了 13%。6 年以后，全球 21%的数据将在中国，这将是占全球五分之一的数据，所以说中国所拥有的数据在国际上举足轻重。

大数据的持续发展

上个月 IDC 的“数字宇宙”研究报告指出，全球数据总量将以每两年翻一番的速度持续增长。去年只有 20%的数据与云相关，几年之后将翻倍至 40%，这是一个非常重要的数据。如今在全球数据的分布中，中国占了 13%。6 年以后，全球 21%的数据将在中国，这将是占全球五分之一的数据，所以说中国所拥有的数据在国际上举足轻重。

大数据的持续发展

有一位尊敬的学者曾经讲到：“信息科学的发展使人们难以预料，如果你想让哪一个人犯错误，那么你就让他预测 5 年以后信息技术的发展。”作为我们云计算、大数据实际上也是同样的原理，大数据目前在哪个位置上对我们将来的研究是至关重要的。

1997 年第一篇大数据文章的起源，2004 年里程碑式的发展，2006 年持续发挥作用作用，2011 年，大数据获得成功。

科学领域对此非常重视，6 年前在科技领域，大家最为尊崇的两大期刊，一个是《自然》，一个是《科学》。《自然》对很多入来说，它做了很多有关生命起源的专题，像大爆发，像东北鸟的起源等等，但是 6 年前《自然》杂志也出版了大数据的专刊。2009 年《第四范式》出版是一个非常重要的里程碑，后续有 2010 年《无处不在的数据》，3 年前《科学》专刊也出来了。科技界的重视、全球的重视，引起了联合国足够的重视，联合国的一个报告白皮书《大数据促发展，挑战与机遇》接踵而至，紧接着两年前高德纳的 4V 定义，同时在我们这个领域也发表了一篇文章，提出了大数据与数字地球的关系。

然而，虽然目前在大数据方面热度最高的国家是印度，但是我们国家有关决策部门对此重视程度也是相当高，特别是近一两年反映在我国部署的 973 计划、863 计划，也包括我国的经信委也在部署大数据科学计划，工业部门和一些地方也在做同样的计划。

同样在国际组织上，可以看到都在不断地推动发展和全球计划，OECD 发布了《探索数据驱动型创新》报告，指出利用数据来刺激提升生产力，充分发挥大数据的潜能，造福人类，进一步阐述了大数据未来的潜力和作用。国际科联在 2011 年之前就开始做战略规划，在 2012-2017 年的战略规划中特别强调了数据、信息对于科学驱动的着手，在数据中间特别强调了大数据的作用。

大数据在科研领域的发展现状

作为国际科联的一个委员会，CODATA——国际科联科学数据委员会也专门召开了一个大数据会议，邀请了 DWS 等重要的数据组织，同时也邀请了国科联的未来地球计划，邀请了全球综合减灾计划，以及 83 个国家和 72 个国际组织构成的一个政府间国际组织。前段时间，我在的中科院遥感数字研究所也召开了相关会议，大家讨论的就是大数据到底对我们的学科发展能带来什么，大数据为全球的科学计划起到了什么样的重要作用。

为了做这样的研讨，在全球范围内我们做了一些基础性工作。比如以 BIGDATA 为关键词进行检索，出现相关文章一千多篇，半数都是 SCI 收录的文章。从国家分布来看，如果说以 40 篇

为一个界限,超过 40 篇的有七个国家,从文章发表来看,中国有 181 篇仅次于印度,位居全球第二从中可以看到中国对大数据的研究发展是鼓舞人心的。

关于大数据的定义,仍在不断的探讨之中,其中 4V 是大家感兴趣的重要话题。大数据时代以及大数据计算的本质特征在于从模型驱动到数据驱动范式的转变以及数据密集型科学方法的确立。

如今科学范式也出现了转化,几千年前以实验为手段描述自然,数百年前利用模型和归纳方法进行论证的分支出现了,过去数十年进行计算分支仿真模拟,现在当大数据时代到来时,将理论、实验、计算仿真等统一起来形成了新的密集计算。

举一个例子,是一个很久远的故事,喜欢天文的先生们都知道第谷,他的助手开普勒建立了一个伟大的理论。这中间我们描述大数据与大科学是一个什么样的关系,然后大科学中我们要有系列的知识发现,讲方法论、讲挑战,这种挑战我们规定为三个 H。

大科学装置要催生带领一系列的大数据,以中国科学院遥感与数字地球研究所为例,每天都在诞生大量的卫星和航空数据,这样的大科学投资强度大、学科交叉,产生的大数据为大科学的诞生奠定了坚实的基础。例如在高能物理领域,每年投入很多,重要的是产生了很多数据,这些数据为科学家带来了重要机遇。因为是需要通过分析和推断“上帝粒子”是否存在,从数据和图形的分析中来揭示这些现象,所以“上帝粒子”的产生和发展与大数据有着重要的、密切的关系。

其次,人类基因组是一个伟大的计划,如果没有大数据,如今人类基因组可能很难有如此丰富的成果。现在有很多人在做多年来产生序列的数据,这需要很多的存储、分析和计算能力去完成。我们正面临全球变化的严重挑战,解决之道就是数据。不仅仅是搜集数据,而是通过对现有数据的分析,预测未来的模拟数据,预测未来的全球变化。一百年之后全球是变暖了,还是变冷了,十万年之后,人类是否还能在地球上居住,其实都要靠数据去建立,可以说这些数据为气候的长期变化都可以提供坚实的基础。

再者,大数据与天文方面的例子更是不胜枚举,我们从地面对天文的观测也是建立在大数据的基础之上,其重要性体现在有了大数据,也就相当于拥有了各种探测设备一样。

简而言之可以看出科学大数据一些非常重要的特征,复杂性、综合性、全球性等多种特点融于一身。大数据特别是科学大数据也有若干外部特征,从内容、体量、速率、手段等等来讲,具备着五大特征。它的内部特征包括超高维度、高度计算复杂性、高度的不确定性。大数据研究方法论一个量、一个是质、一个是用,三个方面互为因果、互相促进。

数字地球驱动的发展情况

1999 年中国科学家发起召开全球国际第一次数字地球会议,从此数字地球在全球范围内广泛的传播。2006 年国际数字地球学会诞生,创建了《国际数字地球学报》等刊物。当前发展势头良好,数字地球研究在全球范围内硕果累累,过去 15 年来长足的发展,提出的数字地球理念的 85% 已经实现。

数字地球的建立基于各方面的数据,其中观测数据非常重要。过去半个世纪以来,国际的对地观测卫星的发展一浪高过一浪,从最早全球 14 个卫星发展到现在每年都有十几颗卫星上天。可以看到中国未来十年对地观测卫星的发展,未来中国遥感卫星的数量要居世界前列,还要发射一系列遥感卫星和其他卫星,不仅发射中国的卫星,还要有强大的接收能力,所在的遥感地球所的地面接收系统能覆盖全国,也能接收覆盖亚洲将近 70% 的数据。最近还要在北极建站,实现一个理念,国家的利益延伸到哪,我们的空间信息保障能力就建立在哪。

15 年前有记者问我什么是数字地球?我说我不懂,但是要让我讲给中学生听,数字地球就是把我们的地球装进我的计算机。15 年前过去了,现在从科学角度来讲,你问我什么是数字地球?我可以从一个角度来讲,数字地球就是地球大数据。

依据这个新的观点,数字地球一些理念出来了,并且研究探讨了数字地球的一系列科学问题,包括大数据的动态汇聚,数据密集型的地球计算,要素的呈现理论与方法等等。我们现在所做的

理论研究、高技术发展，特别对我们这样一个发展中国家来说，最终还是要用在经济社会发展的一些领域，那么将来新一代数字地球构造起来是做什么呢?在此特别提炼五个重要的数字地球构造发展方向:一是全球变化；二是减灾防灾；三是新能源开发利用；四是农业食品安全；五是数字城市和智慧城市的发展。

因此，从大数据的理念来构造新一代的数字地球，其实还是建立一个虚拟地球，虚拟地球事实上服务于未来的地球，使全球环境可持续发展。我们生活在这个地球上，都有责任和理念来建立未来的数字地球，让人类的生活更加美好。

(以上内容系根据郭华东院士在“2014 第六届中国云计算大会”上的演讲整理而成)