2013 年 7 月17日，习近平总书记指出：“浩瀚的数据海洋就如同工业社会的石油资源，蕴含着巨大生产力和商机。谁掌握了大数据技术，谁就掌握了发展的资源和主动权。”大数据已成为信息主权的一种表现形式，将是继边防、海防、空防之后大国博弈的另一个空间。大数据正在改变人类生活和对世界的深层理解。

第二次工业革命的爆发，导致以文字为载体的数据量约每 10 年翻一番；从工业化时代进入信息化时代，数据量每 3 年翻一番。当前，新一轮信息技术革命与人类社会活动交汇融合，半结构化、非结构化数据的大量涌现，数据的产生已不受时间和空间的限制，引发了数据爆炸式增长，数据类型繁多且复杂，已经超越了传统数据管理系统和处理模式的能力范围，人类正在开启大数据时代新航程。据国际数据公司（IDC）发布的 2017 年大数据白皮书预测，2025 年全球大数据规模将增长至 163 ZB，相当于 2016 年的 10 倍，大数据继续表现出更为强健的增长态势。中国拥有的数据在国际上举足轻重，截至 2012 年，已占全球的 13%，预计到 2020 年将产生全球 20% 的数据。

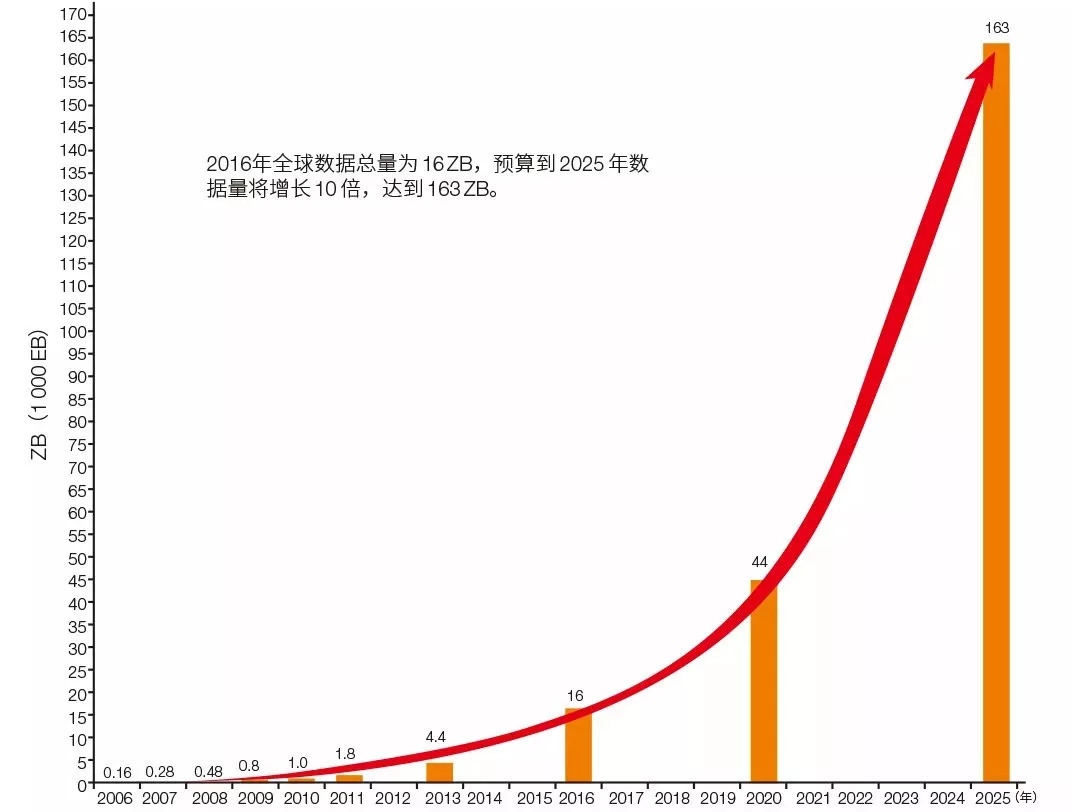


图1 2016—2025年的全球数量增长情况

从大数据搜索热度数据可清晰看出近年来全球对大数据的关注程度。国际上对大数据的关注度在 2012 年之前处于较低水平，2012—2015 年对大数据的关注度飞速增长，2016 年至今保持接近 100 的关注度。

国际上，从联合国到各国政府竞相重视大数据发展；在我国，大数据被列为国家战略后发展迅猛。全球大数据的发展方兴未艾，大数据已经开始显著地影响全球的生产、流通、分配和消费方式，它正在改变人类的生产方式、生活方式、经济运行机制和国家治理模式，它是知识驱动下经济时代的战略制高点，是国家和人类的新型战略资源。

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 姓名 | 年龄 | 电话 |
| 123 | 21 | 121 | 1232 |
| 21 | 12 | 3123 | 131232 |

**科学大数据的认识**

作为大数据的一个分支，科学大数据正在成为科学发现的新型驱动力，引起有关国家和科技界的高度重视。欧盟提出“科学是一项全球性事业，而科研数据是全球的资产”的理念。美国的“从大数据到知识”计划、欧盟的“数据价值链战略计划”、英国的“科研数据之春”计划、澳大利亚的“大数据知识发现”项目、欧洲“地平线 2020”计划的“数据驱动型创新”课题，均聚焦于从海量和复杂的数据中获取知识的能力，深入研究基于大数据价值链的创新机制，倡导大数据驱动的科学发现模式。大数据的影响已触及自然科学、社会科学、人文科学和工程科学的各个研究领域，不同领域的大数据研究中心陆续成立。我国部署了一系列大数据科技项目，组建了不同研究方向的大数据实验室，中国科学院推出了“科学大数据工程”计划。

科学大数据具有数据密集型范式的特点，它具有数据的不可重复性、数据的高度不确定性、数据的高维特性、数据分析的高度计算复杂性等特征。利用大量数据的相关性可取代因果关系和理论与模型，基于数据间的相关性能够获得新知识、新发现。比如，早在 1609 年，第谷 ·布拉赫的助手约翰尼斯·开普勒从布拉赫对天体运动的系数观察记录中发现了行星运动定律，并发表了伟大的著作《新天文学》；又如，欧洲大型强子对撞机帮助物理学家检验关于不同粒子物理和高能物理理论的猜想，并且确定了希格斯玻色子的存在；再如，大数据使基因组学的科学发现成为可能；还如，时空大数据在全球环境研究变化中正发挥重大作用。

越来越多的科学发现证明，大科学装置是人类认识自然世界的重要手段。对地观测卫星、大型望远镜、大型强子对撞机、高通量科学仪器、传感器网络等一系列大装置的成功运行，使得科学大数据与大装置和大科学间的关系越发密切。近年来，我国的大装置诸如 500 m 口径球面射电望远镜、系列空间科学卫星等的问世，为通过科学大数据认知大自然提供了强大的基础。为满足庞大且日益快速增长的科学大数据的应用需求，迫切需要建立一些能够共享数据、算法、模型的开放系统，以此实现对已有数据的科学分析和集成应用。一个典型的例子是，2017 年 10 月，欧洲航天局“哨兵 -5P”卫星发射后，每天获取近 2 000 万条空气污染物及气体的观测数据，其数据获取量是前期任务的 10 倍以上。按照目前的处理速度，一台计算机需要 1 200 年才能处理完 300 万景全球卫星影像。而基于云计算设施，可在 45 天内完成相同处理任务，足见重大基础设施的重要性。

真正实现科学大数据的大价值尚面临着一系列技术挑战。在数据规模、数据增速、数据类型、数据质量、数据价值等方面给科学大数据处理技术与方法提出了新的科学技术问题和方向。

以上主要体现在 5 个方面：①数据存储管理方面。科学大数据本身固有的特征亟待面向海量、非结构化或半结构化数据高效存储管理的数据库。②数据分析方法方面。数据产生和数据分析过程的分离使得数据噪声增多，问题驱动的研究方式逐渐被数据驱动的研究方式所代替。③模型和算法方面。随着半结构化、非结构化数据比重的逐渐增多，针对该类数据的特征学习方法逐渐超越并取代传统的数据模型和算法。④计算体系结构方面。新型存储器件和计算器件不断涌现，使得通用处理器和单一体系结构的单机逐渐过渡为专用处理器、多核和分布式大规模异构集群。⑤计算和服务方面。以互联网为媒介的云计算模式和分布式高性能数据中心逐渐成为大数据处理的新型模式。

中国科学院正在开展科学大数据研究的一些实践。如正在进行的中国科学院战略性先导科技专项（A类）“地球大数据科学工程”，地球大数据是一种典型的科学大数据，是具有空间属性的地球科学大数据。该专项力求突破超大规模跨域分布式资源技术瓶颈问题，有效推动地球大数据技术创新、聚合多时空数据管理与关联融合以及问题导向数据挖掘与分析，以达到只要有终端和互联网，任何人在任何地点都可以享受到地球大数据提供的多样服务，实现重大科学发现和一站式全方位宏观决策支持服务的目的。

又如基于科学大数据的国际科学计划。我们于 2016 年发起的“数字丝路”（DBAR）国际计划，就是要实现大数据汇集、大数据服务、大数据分析和大数据呈现支撑，形成“一带一路”科学大数据平台。这个为期 10 年的科学计划，将为“一带一路”可持续发展、粮食安全、生态环境保护、气候变化监测、灾害风险应对，以及文化—自然遗产保护与发展等提供科学决策。

再如基于科学大数据的研究项目。联合国设立了一项名为“全球脉动”的计划，其使命之一是用大数据应对气候挑战。2014 年，在联合国气候变化峰会召开之际，来自 46 个国家的大数据应对气候变化项目参加了奖项竞争“挑战”，最终 9 个项目获得不同的奖励。我们的“对地观测大数据应对全球变化”研究项目入选其中，显示了空间对地观测大数据在气候变化研究中的作用和价值。

科学大数据正深刻改变传统的科研模式，正驱动现代科学研究的迅猛发展。科学大数据正在为科技创新带来大机遇。作为少量依赖因果关系，而主要依靠相关性发现新知识的新模式，科学大数据已成为继经验、理论和计算模式之后的数据密集型科学范式的典型代表。

**科学大数据的思考**

随着数据积累和计算能力的提升，直接从大数据中获取知识已经成为可能。2013 年 9 月，笔者及团队提出“科学大数据”概念，并于 2014 年 1月以“科学大数据与数字地球”为题发表于《科学通报》。我们认为，科学大数据与互联网大数据、商业大数据等存在本质属性和特点上的区别，具有自己独特的科学内涵和特点。

整体看来，科学大数据具有如下外部特征：从数据内容来讲，科学大数据一般表征自然客观对象和变化过程；从数据体量来讲，科学大数据在不同学科中存在较大的差异；从数据增长速率来讲，科学大数据依学科不同其数据增长速率也变化较大；从数据获取手段来讲，科学大数据一般来自观测和实验的记录以及后续加工；从数据分析手段来讲，科学大数据的知识发现一般需要借助科学原理模型。

通过归纳科学大数据的外部特征，其内部特征也变得相对清晰，主要概括为：数据内容的不可重复性。正如哲学家赫拉克利特的名言“人不能两次踏进同一条河流”，对于一般自然与物理的客观过程的观测具有一定的不可重复性。数据的高度不确定性。由于采用的直接或非直接观测方式、采样手段和记录技术，往往引入系统观测误差及数据记录误差。数据的高维特性。由于观测对象和采样方法本身的时间、空间属性以及观测传感器的多通道特征，科学大数据往往具有时空连续性和谱段多维性，导致维数灾难。数据分析的高度计算复杂性。数据的高度不确定性、高维特性，以及与科学数据分析相伴随的原理模型的复杂性，导致了科学数据处理分析的计算复杂性。总之，科学大数据具有不同于一般大数据的特征，其内在机理及如何应用于知识发现需深入研究。

2014 年 6 月，在我们的倡议和主持下，“国际科学计划大数据研讨会：挑战与机遇”在北京召开。该会议由国际科学和技术数据委员会（CODATA）主办，7 个国际组织共同主办。会议发表的声明强调科学研究要加强对大数据的理解，通过发展与大数据有关的研究、政策和框架来强化国际大数据科学合作，促进社会发展。尽管这在当时只是一个起点，但这份声明是人们关注大数据潜力迈出的实质性一步。声明要点包括：响应大数据对国际科学计划的重要性；开发大数据为社会服务的潜力；通过国际合作来增进对大数据的理解；通过全球研究基础设施促进大数据的普及；探索和应对大数据管理工作带来的挑战；鼓励大数据科学能力建设；促进政策制定，最大限度地利用大数据。

自那时起，我们主办或共同主办了一系列关于科学大数据的会议，其中包括“科学大数据前沿香山科学会议”“中国科学院学部空间地球大数据科学与技术前沿论坛”“自然科学与人文科学大数据前沿探索圆桌会议”“地球大数据香山科学会议”等。有关部门和单位相续组织召开了不同的与科学大数据有关的会议，进行深入研讨。

特别重要的是，在中国科学院的组织下，我们提出发展“科学大数据”的建议，上报后受到政府的重视。2015 年《国务院关于印发促进大数据发展行动纲要的通知》中把科学大数据作为纲要的一部分，提出“发展科学大数据：积极推动由国家公共财政支持的公益性科研活动获取和产生的科学数据逐步开放共享，构建科学大数据国家重大基础设施，实现对国家重要科技数据的权威汇集、长期保存、集成管理和全面共享。面向经济社会发展需求，发展科学大数据应用服务中心，支持解决经济社会发展和国家安全重大问题”。

科学大数据是国家大数据战略的有机组成，这使得深入开展科学大数据的研究具备了良好的政策支撑和理论基础。科学大数据是国家大数据战略的基石，科技界和科学家肩负重大的使命——推进科学大数据的全面系统发展。