大数据是数据、技术，还是应用

朱扬勇[1,2]，**熊 赟**[1,2]
1. 复旦大学计算机科学技术学院 上海 201203；2. 上海市数据科学重点实验室（复旦大学） 上海 201203

摘要
通常认为大数据是一个现有技术难以处理的复杂而庞大的数据集，这将导致一个谬误的出现：大数据都不能被处理，能处理的都不是大数据。显然，如何定义大数据是一个问题。分析了已有的大数据定义和现象，发现数据、技术和应用是大数据的三要素，定义大数据是为决策提供服务的大数据集、大数据技术和大数据应用的总称。其中，大数据集是指一个决策问题所用到的所有可能的数据，而不是一个领域的所有数据。还给出了大数据应用遇到的问题及技术挑战，并指出大数据未来的研究方向。

关键词
大数据；数据科学；数据界

Defining Big Data

Zhu Yangyong[1,2], **Xiong Yun**[1,2]
1. School of Computer Science, Fudan University, Shanghai 201203, China;
2. Shanghai Key Laboratory of Data Science, Fudan University, Shanghai 201203, China

Abstract
Generally, big data is regarded as a term about data sets so large or complex that conventional data technologies cannot handle. This statement of big data leads to confusion: none of big data has been handled by existing data technologies; or none of current successful data applications can be called as big data. Therefore, what is the best way to define big data becomes a problem. Data, technology, and application were regarded as three associated key factors of big data by analyzing the state-of-the-art of big data. A comprehensive definition on big data was defined as the umbrella of big data set, big data technology, and big data application. Here, big data set means all data that can be acquired and were related to one decision-making application instead of all data in an area or an enterprise. In addition, the issues in big data applications and the main challenges in big data technologies were discussed. Finally, the future directions of big data research were presented including data science and the technologies of big data reservation and development.

Key words
big data, data science, data nature

1 引言

1997年NASA研究员Michael Cox和David Ellsworth在IEEE第8届国际可视化学术会议中首先提出了“大数据”术语[1]，但并没有引起太多重视；2008年9月《Nature》学术杂志出版了一期大数据专刊[2]，使得大数据在科学研究领域得到了高度重视；2012年3月美国政府发布《大数据研究和发展倡议》[2]，大数据引起了主要国家和全社会的重视。一场大数据引发的变革渗透到各个角落。

一个概念让政治界、商业界、学术界的各个领域都为之兴奋不已，超过了当年计算机的诞生，也超过了互联网的诞生。大数据引起政治界重视，世界强国推出大数据战略，说明大数据关系到国家竞争力、关系到国家发展、关系到国民大众；大数据引起商业界重视，跨国公司率先运用大数据，说明大数据已经实用，商业价值重大，是企业竞争的利器；大数据引起学术界重视，说明大数据科学问题众多，需要科技攻关。

然而，关于什么是大数据却众说纷纭，以至于出现一些相互矛盾的现象，最典型的矛盾现象是：技术领域说大数据是当前技术所不能解决的，而应用领域却给出了大量关于大数据成功应用的案例。“大数据都不能被处理，能够处理的都不是大数据”或者“大数据都不能用，能用的都不是大数据”这是一个谬误。事实上，到目前为止，大数据还没有一致的定义，政治界、商业界、学术界按照各自的理解推进大数据。甚至在信息技术领域，大数据概念也是争论不休的，各研究方向也都带上了大数据的帽子，似乎大数据技术将取代信息技术，这显然是有问题的。

本文探寻大数据概念的内涵、大数据问题和技术挑战，给出了一个大数据的定义，指出了大数据应用面临的6个问题（以下简称“6用问题”），分析了信息化和大数据的差异，提出了“6用问题”带来的技术挑战，并进行了展望。

2 大数据概念

严格地说，到目前为止，还没有一个明确的大数据定义，各领域按照自己的理解来研究和发展大数据。最直接的问题是大数据是数据还是技术？显然，这个问题并不容易回答。

2.1 现有定义的问题

目前，大数据有如下几个定义。

Michael Cox和David Ellsworth在提出“大数据”术语时指出：数据大到内存、本地磁盘甚至远程磁盘都不能处理，这类数据可视化的问题称为大数据[1]。

维基百科的定义[3]：大数据是一个复杂而庞大的数据集，以至于很难用现有的数据库管理系统和其他数据处理技术来采集、存储、查找、共享、传送、分析和可视化。

4V定义[4,5]：大数据为具有4V特征的数据集。4V特征是指：价值（value），数据价值巨大但价值密度低；时效（velocity），数据处理分析要在希望的时间内完成；多样（variety），数据来源和形式都是多样的；大量（volume），就目前技术而言，数据量要达到PB级别以上。

香山科学会议定义[6]：2013年5月召开的第462次香山科学会议给出了技术型和非技术型两个定义。

- 技术型定义：大数据是来源多样、

类型多样、大而复杂、具有潜在价值，但难以在期望时间内处理和分析的数据集。

- 非技术型定义：大数据是数字化生存时代的新型战略资源，是驱动创新的重要因素，正在改变人类的生产和生活方式。

这些定义总体来讲是从技术领域看问题的。可以看出，大数据是难以处理的数据集，即大数据是一个数据集。但是，如果大数据只是一个数据集，那么处理大数据的技术叫大数据技术吗？与之前的信息技术是否有区别？在应用方面更难说清楚。例如，是否可以说“用大数据解决问题”？显然，一个数据集是不能解决任何问题的。所以，大数据不仅仅是数据集，但也不仅仅是技术，还有大数据应用。

上述定义最大的问题是，均认为大数据是指当前技术难以（所不能）处理的数据集。但当技术改进了，能够处理了，还是不是大数据？于是，一个典型的矛盾现象出现：技术领域说大数据是当前技术所不能解决的数据集，而应用领域却给出了大量关于大数据成功应用的案例。这是对大数据的谬误：大数据是当前技术难以（所不能）处理的数据集，那么，所有能够被处理的数据集都不是大数据，所以没有大数据的成功应用，即“大数据都不能被处理，能够处理的都不是大数据”或者“大数据都不能用，能用的都不是大数据”。

另外一个现象是大数据之争，即常常有各种领域的人在一起争论什么是大数据。由于技术领域和非技术领域对大数据的理解不同，这两个领域谈论的对象其实是不同的，技术领域说的大数据是指大数据技术，而应用领域说的大数据是指大数据应用。事实上，经过长期信息化建设，几乎所有的行业和单位都积累了庞大的数据资源，所以，数据和基于数据的应用涉及几乎所有的人。可以将大数据人群分成3类：有大数据的人群、做大数据的人群和用大数据的人群，很多时候大家在谈论大数据的时候，实际上是在谈论不同的东西，即有大数据的人谈论数据资源及其规模、做大数据的人谈论大数据带来的技术挑战、用大数据的人则谈论大数据带来的决策变革，即3类人群谈论的是不同的大数据概念。

出现大数据谬误和大数据之争的现象源于大数据概念不清晰，需要一个清晰的定义来避免这些现象的发生。

2.2 数据、技术和应用是大数据的3要素

大数据到底是数据、技术，还是应用？

大数据首先是一个技术术语，来自技术领域，或者更准确一点是来自IT（information technology）领域。自Michael Cox和David Ellsworth[1]于1997年首次提出“大数据”以来，在术语发展过程中，始终提及的大数据问题是指“现有技术所不能处理的数据集”，即大数据是一个技术挑战。直到2012年3月美国政府发布《大数据研究和发展倡议》[2]，大数据一词开始在非技术领域使用。大数据在非技术领域的主要表述为：大数据是决策方式的重大变革，决策依靠数据分析而不是直觉经验，主要的内涵是“大数据改变了人类生产和生活方式，是一次大变革”[6,7]。

大数据的4V定义涵盖了所有技术型定义，也是影响最广泛的，但在具体理解和具体问题面前，还是引起了很多争论。例如，常常会争论一个数据集是不是大数据，即够不够大，是否达到了PB级别。显然，这只是问题的表面。问题的核心是：一个数据集是否有价值、是否值得去开发、能否挖掘出价值；能否在希望的时间内挖掘出价值。因此，价值和时效是大数据的核心内涵，是必须的。

（1）关于价值：如果一个数据集没有价值，就不需要关注；如果一个数据集的价值密度高，即大部分数据都是有价值的，直接读取数据集就能获得价值，可以成功应用，没有技术难度。然而，通常情况是价值巨大但价值密度低，像大海捞针，因此大数据是一个很难的技术挑战。

（2）关于时效：所有的大数据处理和分析都应该在希望的时间内做完，如果过了希望的时间就没有意义了，这也是一个技术挑战。

从上述定义中可以看出：首先，所有的定义都谈到了数据，一个庞大的数据集；其次，技术方面强调了大数据是当前技术所不能的，这里的“不能”是指“不能在希望的时间内”做到，是技术问题；第三，大数据是用来解决决策应用问题的，是一个基于数据集和数据技术的决策应用，改变着生产和生活中的决策方式。因此，数据、技术和应用是大数据的3个要素，数据隐含价值、技术发现价值、应用实现价值。

2.3 定义大数据

应该如何定义大数据呢？首先，不能把一个技术挑战定义为大数据，否则，一旦技术挑战解决了，就不是大数据了，而且挑战本身不是一个事物，不能命名；其次，也不能把一个数据集定义为大数据，数据集本身只是隐含价值，不能直接发挥作用；最后，更不能将一个数据应用定义为大数据，那样会导致所有基于数据的系统都是大数据。可以采用如下描述定义大数据。

大数据是指为决策问题提供服务的大数据集、大数据技术和大数据应用的总称。其中，大数据集是指一个决策问题所用到的所有可能的数据，通常数据量巨大、来源多样、类型多样；大数据技术是指大数据资源获取、存储管理、挖掘分析、可视展现等技术；大数据应用是指用大数据集和大数据技术来支持决策活动，是新的决策方法。

大数据能否为一个决策问题提供服务的关键是：是否能在决策希望的时间内有效完成所有的任务。由于数据增长的速度远快于技术进步的速度，因此就出现大数据问题。

大数据问题是指不能用当前技术在决策希望的时间内处理分析的数据资源开发利用问题。大数据问题的关键技术挑战在于：找到隐含在低价值密度数据资源中的价值；在希望的时间内完成所有的任务。

根据这个定义，大数据谬误和大数据之争就可以避免。

首先，给定一个大数据集，当没有大数据技术能够在希望的时间内开发其价值，那么该大数据是一个技术挑战，否则就是一个大数据应用。需要注意的是，一个大数据应用可能会转化成大数据的技术挑战。例如，无人驾驶汽车在道路上行驶时，需要综合分析汽车自身的工作数据（行驶速度、油量、引擎工作状态等）、地图及实时路况数据、道路管理数据（红绿灯、限速等）等，快速做出驾驶决策。假设汽车10 km刹车距离为45 m，那么当汽车时速小于60 km/h时，发现50 m外车道上有行人后，经过2 s的数据分析得出需要刹车的结论是可以接受的，因此是一个成功的大数据应用；但当车速提高到100 km/h时，数据分析的时间就得小于0.18 s，这就变成了技术挑战。反之，一个大数据挑战也同样可以变成一个大数据应用。上述例中，在高速公路上数据分析的时间小于0.18 s，这是一个大数据技术挑战，但是，如果市内汽车限速为小于50 km/h，那么2 s的数据分析技术就可以使用，就会有成功的大数据应用。

其次，有数据的、做数据的、用数据的人群谈论的大数据分别是大数据集、大数据技术和大数据应用，所以不同人群谈论

的大数据只是大数据的不同侧面，分析清楚后就可以避免无谓的争论。

2.4 信息化与大数据

信息化的本质是生产数据的过程，数据被大量生产而形成了数据资源。数据资源的开发利用逐渐成为人类的新需求，从早期的数据仓库和数据挖掘技术的提出，到决策支持系统和商业智能的应用，都是在进行数据资源的开发利用工作。直到大数据的出现，数据资源的开发利用工作从量变发展到了质变：数据开发发展成为一个新的领域或行业，信息技术发展出新的技术分支——大数据技术，并迅速壮大，对数据界的探索发展成为一个新的科学——数据科学[8~11]。图1展示了信息化和大数据的差异。

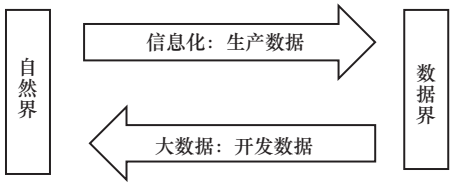

图1 信息化与大数据

3 大数据应用

大数据应用是决策应用，即给定一个决策需求，然后获取数据，分析数据，形成决策依据。很早期的关于沃尔玛公司的“尿布和啤酒”的故事，决策需求是“哪些商品最容易被同时购买”。其他如Google预测流感[12]、亚马逊推荐图书[13]、科学家发现“上帝粒子”[14]等，都是解决决策应用的。

2008年《Nature》大数据专刊说明：科学研究领域率先遇到大数据决策问题[2]。决策可以发生在任何场合，大到国家宏观决策、科学研究，小到选择一家合适餐馆、确定一条行车路线。由于决策的复杂性、困难性，大数据集通常是数据量巨大、来源多样和类型多样的数据集，这样大数据应用通常具有跨界数据、跨界应用的特点，打破原有行业领域界限，是决策方式的质变。

3.1 决策依靠数据

从古到今，无论在战场战争、商业竞争、科学研究、日常生活中，取胜的重要因素是比别人知道更多、比别人更快地做出正确的决策。计算机出现之前的决策是采用人工方式：依靠手工收集和分析信息、依靠决策者的经验和直觉做出决策。后来有了计算机决策支持系统（decision support system，DSS），再后来有商业智能（business intelligence，BI），这个时候就可以利用自身信息化积累的数据来开展决策[15]。然而，自身的数据积累是一个漫长、费钱和困难的工作，只有大型企业和政府有能力这样做。

随着技术进步和互联网的普及应用，不论是政府、组织、企业还是个人都越来越有能力获得决策需要的各种数据，这些数据来源多样、类型多样，甚至超过早期大型企业自身的积累，并且数据分析技术也取得了长足进步，人们可以通过分析这些数据得到决策依据。这样，一种新型的决策方式产生了，这就是大数据决策。由于这是一个从量变到质变的过程，不能简单地说之前的BI不是大数据，也不能简单地说BI是大数据。

大数据形成决策依据的3种重要方式是：从精确分析到近似分析、从样本分析到总体分析、从因果分析到关联分析[16]。大数据决策主要体现在“通过分析不同来源的各种可能的数据来支持决策活动”。由于大数据过于庞大和复杂，难以弄清数

据之间的因果，所以大数据决策常常表现出“知其然就可以做出决策，而可以不知其所以然”[15]。

那么如何来实施大数据决策呢？首先，需要获取数据，并进行数据清洁和整合，形成大数据集；然后，使用大数据技术分析大数据集；最后，解释和展示大数据开发的结果，实现大数据决策。

3.2 6用问题

给定一个大数据应用需求，通常会遇到以下6个方面的问题，即“6用问题”。

（1）数据不够用

获取尽可能多的数据（决策素材）是一种直觉上的追求，即数据越多对决策越有利，或者至少比别人知道的更多，虽然实际情况可能不是数据越多越好，但这很难判断。所以，大数据应用的第一个问题是“数据不够用”。

（2）数据不可用

在数据够用的情况下，还会遇到数据不可用问题。数据不可用是指拥有数据，但访问不到数据。例如，某个公共决策需要用到民政局、公安局、人力资源和社会保障局、税务局的数据，这些数据在各部门都有，但是数据不在一个系统里，是数据孤岛，并不能用来做大数据决策；又如，一些交易系统只保留活跃用户数据，不活跃用户的数据被备份到了备份系统中，访问备份系统数据是一件费时、费力的工作，甚至是不可能的工作。

（3）数据不好用

面对足够可用的数据资源，下一个问题是数据不好用问题，即数据质量有问题。例如，信用判定应用中，发现一些持卡人的登记信息缺失（如没有职业数据）或不正确（如收入数据不对），这些问题直接影响了决策依据的获得；又如，在战场环境中，由于敌方的有意伪装和干扰，获得的数据质量更差。

（4）数据不会用

数据不会用问题是指不懂大数据分析技术、不会将业务问题转化为数据分析问题，而这正是大数据决策的核心。由于数据分析技术门槛很高，能够使用大数据分析技术的人很少，而将业务问题转化为数据分析问题，更需要数据科学家创造性的劳动。例如，在网站上做精准广告是一个业务问题，在理解业务问题的基础上，用大数据技术实现对用户的购买喜好和需求进行聚类分析，将广告和用户簇进行对照，好的精准广告可以针对每个用户来做。数据科学家极其短缺，使得数据不会用的问题在实际中表现非常严重。

（5）数据不敢用

数据不敢用是指因为怕担责任而将本该用起来的数据束之高阁。很多政府数据资源之所以没有很好地开发利用，其中一个主要原因是数据拥有部门不愿意将数据用于非本部门业务，怕丧失数据安全（如所有权和数据秘密）。

（6）数据不能用

数据不能用有两个方面，一个是数据权属问题，即数据不属于使用者；另一个是社会问题，即隐私、伦理等问题。首先，没有使用权的数据不能用；其次，涉及隐私的数据需要脱敏处理，或者只做总体分析，不做个体分析，例如人口统计数据就只能做总体分析，不能做个体分析；第三，涉及伦理等社会问题的数据也不能用，例如信用评分中的种族、民族、性别等数据就不能用。

4 大数据集

根据本文的定义，大数据集是指解决

一个决策应用问题所用到的所有数据，但不是全球的数据总和，也不是一个行业的数据总和，也不是一个组织的数据总和。但由于决策问题的复杂性，一般来讲，大数据集的数据量巨大、来源多样、类型多样。一个决策问题用到的数据由具体的决策问题决定，有些可能数据量大但不复杂，有些可能复杂但数据量不大。

4.1 从数据界获取大数据集

数据作为一种资源已经获得广泛认识。早在2008年，笔者提出：数据资源是重要的现代战略资源，其重要程度将越来越显现，在本世纪有可能超过石油、煤炭、矿产，成为最重要的人类资源；2012年，Amazon前首席科学家Andreas Weigend表示：数据是原油，但石油需要加以提炼后才能使用，从事海量数据处理的公司就是炼油厂；2012年瑞士达沃斯召开的世界经济论坛上，大数据是讨论的主题之一。这个论坛上发布的一份题为《大数据，大影响》（big data，big impact）的报告[7]宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。

大数据是数据资源开发利用的一种当前表现形式，即数据资源已经存在于网络空间，大数据是对网络空间数据资源的开发利用。网络空间的所有数据构成数据界[8,9]，因此，大数据可以看成用数据界的数据来解决决策问题，大数据集应该是从数据界获取，而不是从自然界获取，从自然界获取数据是信息化。

各种大数据的定义都在说大数据是数据集、资源、资产，说明数据已经存在于网络空间。前面提到“随着技术进步和互联网的普及应用，不论政府、组织、企业还是个人都越来越有能力获得决策需要的各种数据，这些数据来源多样、类型多样，甚至超过早期大型企业自身的积累”，也说明数据来自数据界。

4.2 大数据集的要求

大数据使决策者从看到局部数据转变为看到全局数据、从样本分析转变为总体分析。从局部数据到全局数据要求数据集尽量全面，从各种来源获取所需要的数据；从样本分析到总体分析要求数据集足够大。因此，大数据集的要求应该是数据量大或者复杂。

（1）大数据集应该有来源多样、类型多样的数据

由于决策的复杂性、困难性，为满足决策需求，大数据集通常由来源多样和类型多样的数据构成，使用跨界数据，开展跨界应用。数据来源多样的一个要点是来源于决策者/决策机构自身积累之外，这会给数据获取、数据分析技术带来挑战，来源多样通常也意味着类型多样。例如，环境生态研究是进化论、基因组学、地理学、海洋学、气候学、流行病学和经济学的综合研究，其研究工作需要有来源多样的数据[17]。2010年位于墨西哥湾的“深水地平线（deepwater horizon oil）”钻井平台爆炸溢油长达80 mile（约128 km）。对溢油带来的生态影响（如对海岸、海平面、海底的影响，对鱼、虾、昆虫、植物、鸟类、鲸鱼、海龟的影响等）的研究是一个重要课题，需要深海浮游生物（planktonic）和远洋生物（pelagic organisms）、化学（油和分散剂）、毒理学（toxicology）、海洋学（oceanography）和天文学等多源数据支持。灾难发生后，美国国家海洋和大气管理局派出科学考察船，对污染海域进行取样；美国宇航局利用卫星上的中解析度成像光谱仪对海上石油污染进行监测；科学家们还在陆上收集相关数据；英国石油公

司也展开了对该地区空气、水质等方面的测试。

（2）大数据集应该有PB级别的数据规模

就目前技术水平而言，引发技术挑战的大数据集的规模应该有PB级别。PB级别的数据规模是传统数据库管理系统（DBMS）软件所不能有效存放的，因此，PB级别数据规模需要新型的数据管理技术，于是出现分布式文件系统（HDFS）。这只是初步解决了数据存储问题，数据计算、数据分析、数据展现等方面还有很多技术问题。

2008年《Nature》大数据专刊的封面中，除了醒目的“big data”外，还有一句话“science in the Petabyte era（科学处在PB时代）”，这个封面有两层意思：第一层意思是科学研究已经到了大数据时代；第二层意思是PB级数据是大数据规模的一个基本标志，数据量足够大，使用时有技术难度。

在实际中，很多成功的大数据应用的数据集规模都没有超过PB级别，但是，由于决策者所处的计算环境、资金支持所限，很多小于PB级别的数据集已经构成了技术挑战。《Science》杂志于2011年对许多数据相关研究人员（他们都是国际、交叉领域的科学研究团队的负责人）进行了调查，收到了1 700份回应，其中，20%的人回应一般使用和分析的数据集超过了100 GB，7%的科学家使用和分析1 TB以上的数据。一半的科学家认为他们一般仅使用存储于自己实验室的数据，但这不是一个理想的解决方案。国际千人基因组计划（1 000 genomes project）自2008年启动以来，短短4年间已获得1 092人的基因组数据[18]，产生的数据量已达到50 TB。

但在可以预见的未来，PB级别的数据量是科学研究领域进行一项科学研究的常态，也是很多领域的决策应用的常态。例如，2013年3月14日，通过对大约200 PB的数据用150个计算中心进行长达3年的计算分析，欧洲核子研究组织宣布确认希格斯玻色子[14]。又如，美国斯坦福线性加速器中心（SLAC）国家加速器实验室（National Accelerator Laboratory）计划建造的大型综合巡天望远镜（large synoptic survey telescope，LSST）将每晚获取数据5~10 TB（而目前的SDSS仅有每晚200 GB），计划获取60 PB影像数据[19]。

5 大数据技术

面对“6用问题”，大数据技术面临很多挑战。

针对数据不够用问题，需要研究、使用数据获取技术：如何获取足够的数据，是大数据的第一个技术挑战。大数据需要从数据界获取跨领域行业、多类型的数据，而不是从自然界获取数据，因此网络空间的哪些地方有所需的数据、如何拿到数据等是主要的技术挑战，搜索、爬取、下载等是常见的数据获取技术。

针对数据不可用问题，需要研究、使用数据储备和管理技术：数据不可用问题对技术的挑战是巨量数据存储与管理、跨地域数据访问与计算。分布式文件系统、Hadoop是当前被较多采用的技术。

针对数据不好用问题，需要研究、使用数据质量技术：数据不好用问题对技术的挑战是数据质量判定、数据质量提升、数据质量修复。数据清洁是当前采用的数据质量技术，但效果有限。

针对数据不会用问题，需要研究、使用数据分析技术：数据不会用问题需要既能

理解业务需求又懂数据分析技术的数据科学家，其技术挑战是数据挖掘算法的设计和实现、在可接受的时间完成计算。面对PB以上级别的复杂数据，还缺少有效的数据挖掘算法和软件工具。

针对数据不敢用问题，需要研究、使用数据开放共享技术：如果技术做得好，这个问题是有希望解决的。例如，在传统数据管理系统软件中，数据管理员管理整个数据库，但是他并不具备访问具体数据的权限，因此他并不能知晓数据秘密。之前，大部分数据都不开放，所以相应的技术研究有很多空白。数据不敢用的技术挑战是在保护数据安全（所有权和数据秘密）的前提下实现数据开放共享。

针对数据不能用问题，需要研究使用数据权属及保护技术：之前，大部分数据都是自己生产，自己保管，问题不严重，所以相应的技术研究有很多空白。数据不能用的范围广泛，主要的技术挑战包括数据权属的认证和判别技术、隐私保护技术等。

长期以来，信息技术主要是用于信息化的，即生产数据，而大数据是用于开发数据的，如图1所示。面对大数据决策的“6用问题”，之前的技术在数据获取、数据存储与管理、数据质量保障、数据安全与隐私保护等方面遇到了一系列新的技术挑战，需要开发大数据技术来应对这些挑战，而以数据分析技术为核心的数据开发技术正逐步形成独立的技术分支。表1展示了生产数据和开发数据的技术差异。

6 结束语

长期的信息化实践，从数据生产、数据积累、数据资源形成到数据开发，从量变到质变，数据开发发展成为一个新的领域或行业，信息领域发展出新的分支——大数据。大数据是指为决策问题提供服务的大数据集、大数据技术和大数据应用的总称。大数据问题是指不能用当前技术在决策希望的时间内处理分析的数据资源开发利用问题。大数据引发了决策方式的质变，对政治界、商业界、学术界都产生重大影响。

数据的增长给技术带来了挑战，所谓“当前技术所不能”；随着技术的进步，成功的大数据应用不断出现，大数据正是在

表1 生产数据与开发数据的技术差异

6用问题	数据技术	信息化（生产数据技术）	大数据（开发数据技术）
数据不够用	数据获取	从自然界获取数据：通过数字化设备和计算机I/O设备获得数据	从数据界获取数据：购买数据或从各数据源通过下载、爬虫、分发等技术手段获得数据
数据不可用	数据存储管理	开发各种存储技术，包括存储设备、DBMS等各种存储技术	数据已经存在网络空间的某个地方，主要技术包括数据搜索和访问技术、异地计算技术、适合数据分析的存储技术
数据不好用	数据质量保障	内部数据：数据质量技术	有大量外部数据，数据质量问题较严重，需要新的数据质量技术
数据不会用	数据挖掘分析	数据挖掘分析技术被分离出来，形成数据开发技术的核心	数据融合、统计分析、数据挖掘、深度学习等是数据开发的核心技术，还有数据勘探、可视化等
数据不敢用	数据开放共享	数据开放不多，技术有限	新技术，如保护数据安全（所有权和数据秘密）的前提下实现数据开放共享技术
数据不能用	数据安全隐私	内部数据：技术有限	有大量外部数据，数据权属的认证和判别技术、隐私保护技术等

“数据增长”和“技术进步”之间交替前行，成就了当今的大数据热潮。从理论上讲，大数据的技术挑战在摩尔定律的作用下可以自行解决，但数据增长的速度远快于技术进步的数据，所以今天出现了大数据问题。除非出现革命性技术，否则大数据问题不可能被解决。这就需要关注数据本身的变化发展规律，发展数据科学。

对大数据和数据科学的发展展望如下。

（1）大数据储备技术需求迫切

数据作为资源，建立数据储备将是重大需求，因此，数据获取、数据储备设计、数据储备管理、数据搬运、异地数据计算、数据主权保护等数据储备技术有望快速发展。

（2）大数据开发技术快速发展

数据生产技术相对成熟，并形成稳步发展。数据开发技术即将进入快速发展期，包括数据分析技术、大数据软件工程、决策应用技术等。

（3）数据科学稳步前行

从科学研究、学科发展和人才培养角度来看，数据科学将会快速发展。近3年，在美国有包括哥伦比亚大学、纽约大学、加州大学、卡耐基梅隆大学等许多高校建立数据科学研究机构或开设数据科学专业研究生培养项目。

参考文献

[1] Cox M, Ellsworth D. Application-controlled demand paging for out-of-core visualizetion. Proceedings of the 8th Conference on Visualization, Phoenix, AZ, USA, 1997: 235~244

[2] U. S. Government. Big data research and development initiative. http://www.whitehouse.gov/ sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf, 2012

[3] Wikipedia. Big data. http://en.wikipedia.org/wiki/Big_data, 2015

[4] Mark B. Gartner says solving ‘big data’ challenge involves more than just managing volumes of data. http://www.gartner.com/newsroom/id/1731916, 2011

[5] Villanova University. What is big data. http://www.villanovau.com/resources/bi/what-is-big-data/, 2015

[6] 数据科学与大数据的科学原理及发展前景. 第462次香山科学会议, 北京, 中国, 2013
The scientific principle and prospect of data science and big data. Proceedings of the 462nd Xiangshan Science Conference, Beijing, China, 2013

[7] World Economic Forum. Big data, big impact: new possibilities for international development.http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_ 2012.pdf, 2012

[8] Zhu Y Y, Zhong N, Xiong Y. Data explosion, data nature and dataology. Proceedings of International Conference on Brain Informatics, Beijing, China, 2009: 147~158

[9] 朱扬勇, 熊赟. 数据学. 上海：复旦大学出版社, 2009
Zhu Y Y, Xiong Y. Dataology and Data Science. Shanghai: Fudan University Press, 2009

[10] CODATA中国全国委员会. 大数据时代的科学活动. 北京：科学出版社, 2014
CODATA China National Committee. Scientific Discovery in Big Data Era. Beijing: Science Press, 2014

[11] Zhu Y Y, Xiong Y. Defining data science. http://arxiv.org/ftp/arxiv/papers/1501/1501.05039.pdf, 2015

[12] Google. Google flu trends. http://www.google.org/flutrends, 2008

[13] Greg L, Brent S, Jeremy Y. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76~80

[14] Adrian C. Higgs boson positively

identified. http://news.sciencemag.org/sciencenow/ 2013/03/higgs-boson-positively-identifie.html?ref=hp, 2013

[15] 吴俊伟, 朱扬勇. 汇计划在行动. 上海: 上海科学技术出版社, 2015
Wu J W, Zhu Y Y. Shanghai Big Data in Action. Shanghai: Shanghai Scientific and Technical Publishers, 2015

[16] Schonberger V M, Cukier K. Big Data: A Revolution That Will Transform How We Live Work and Think. London: Hodder Export, 2013

[17] Reichman O J, Jones M B, Schildhauer M P. Challenges and opportunities of open data in ecology. Science, 2011, 331(6018): 703~705

[18] McVean G A, Abecasis D M. An integrated map of genetic variation from 1092 human genomes. Nature, 2012, 491(7422): 56~65

[19] Feigelson E D, Babu G J. Big data in astronomy. http://astrostatistics.psu.edu/2012Significance.pdf, 2012

作者简介

朱扬勇, 男, 博士, 复旦大学计算机科学技术学院教授、学术委员会主任, 上海市数据科学重点实验室主任。1989年起从事数据领域研究, 2008年提出数据资源保护和利用, 2009年发表了数据科学论文 “Data explosion, data nature and dataology”, 并出版专著《数据学》, 对数据科学进行了系统探讨和描述。2010年创办了 “International Workshop on Dataology and Data Science”, 2014年和石勇、张成奇共同创办了 “International Conference on Data Science”。第462次香山科学会议 “数据科学与大数据的理论问题探索” 的执行主席, 《大数据技术与应用丛书》主编。目前研究兴趣为数据科学、大数据。

熊赟, 女, 博士, 复旦大学计算机科学技术学院副教授。2004年起从事数据领域方面的研究工作, 作为项目负责人主持国家自然科学基金、上海市科委发展基金以及企业合作项目。相关研究成果在本领域国际权威期刊或会议发表论文30余篇, 出版专著2本。目前研究兴趣为数据科学、大数据。

收稿日期: 2015-04-21; **修回日期**: 2015-05-05
基金项目: 国家自然科学基金资助项目(No.61170096, No.71331005), 上海市科技发展基金资助项目(No.13dz2260200, No.13511504300, No.14511107302)
Foundation Items: The National Natural Science Foundation of China (No.61170096, No.71331005), Shanghai Science and Technology Development Fund (No.13dz2260200, No.13511504300, No.14511107302)
论文引用格式: 朱扬勇, 熊赟. 大数据是数据、技术, 还是应用. 大数据, 2015007
Zhu Y Y, Xiong Y. Defining big data. Big Data Research, 2015007