数据科学与工程：大数据时代的新兴交叉学科

周傲英，钱卫宁，王长波
华东师范大学数据科学与工程研究院 上海 200062

摘要
大数据时代的IT发展的基本特点是：应用驱动创新，开源加速创新，硬件助力创新。基于对这些特点的认识，从社会创新发展、人才需求变化、技术发展趋势等方面论述了数据科学与工程这一新兴交叉学科的发展必然性，进一步阐述了数据科学与工程学科的特点、学科内涵与知识体系，最后从科学研究、系统开发和人才培养的角度探讨了数据科学与工程学科的建设思路。

关键词
大数据；数据科学与工程；交叉学科；万众创新；人才培养

doi: 10.11959/j.issn.2096-0271.2015022

Data Sciences and Engineering: An Emerging Interdisciplinary in the Big Data Era

Zhou Aoying, Qian Weining, Wang Changbo
Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, China

Abstract
There are some characteristics for IT development in the big data era: the real-life applications are the driving force for innovation; open sourcing accelerates innovation, and the advancement in hardware lay the foundation for innovation. The data sciences and engineering was regarded as an emerging and developing interdisciplinary and discussed from the aspects such as social innovation and development, talents demand changes, and technology development. Then the features, connotations, and knowledge hierarchy of data sciences and engineering as a discipline were described. Finally, the associated research and development, talent training, and best practice were also presented.

Key words
big data, data sciences and engineering, interdisciplinary, mass innovation, talent training

1 引言

当前，“大数据”这一术语已经远远超越了当初的互联网或信息技术（IT）的技术范畴，变成了一个时代的标志。大数据时代的到来有其必然性，当计算和通信取得长足进步的时候，当传感器网络和互联网等信息采集平台日臻完善的时候，数据的存储管理和分析处理就自然成为关注的焦点。“大数据”概念的提出意味着信息技术领域的重点由“计算”转为“数据”。稍微留意一下就可以看到，许多原本在IT其他领域成就卓著的著名专家和学者都纷纷转向大数据领域。这种现象可以引发大家的思考：类比于已经发展了半个多世纪的“计算机科学”，现在是否也应该单独认真考虑一下“数据科学”这样一个学科方向？

大数据不仅仅是信息技术领域的事情，它的典型特点就是与应用密切结合。在当前阶段，大数据概念的提出和被广泛接受才不过三四年，属于发展初期。这个时期，如果离开了应用来谈大数据，相信大家都会认为是“空谈”。回顾一下，大数据在科学研究（如地球科学、生命科学、高能物理研究等）[1]和商业领域（如行为分析、趋势分析、行情预测、精准营销、商品推荐等）[2]都有成功的应用。互联网已经成为人们生活生产中不可或缺的环境和平台，正因为大数据在互联网商业领域的巨大成功，使得这一概念已经被社会各个层面广泛认可，开始从线上走到线下，越来越多的人从企业管理、社会治理、科学研究等领域探讨大数据的应用。这种来源于应用的关于大数据技术的爆发式需求，为一门新型的独立学科的形成和发展带来了挑战和机遇。因为其“应用驱动”的特点，工程实现和应用部署至关重要，“数据科学与工程”是个更贴切的学科名称。

基于以上基本认识，本文从社会需求、学科本质以及人才培养等方面进行探讨。

2 数据科学与工程学科发展背景

2.1 我国创新驱动发展的需求

我国的社会经济发展进入新常态，经济发展从高速进入中高速，生产制造从中低端转向中高端。在新常态下，如何有效促进经济结构调整，同时保证就业和经济平稳发展，这有赖于信息化。新时期的信息化还和建设生态文明、拉动消费、提高产品竞争力等密切关联。与以前的“信息化带动工业化”以及稍后的“两化融合”等信息化战略相比，新型的信息化是在移动互联网的环境下提出来的，有着深刻的云计算和大数据背景，对数据科学与工程学科的发展有重要的指导意义。

自从斯诺登“棱镜门”事件以来，世界各国都高度重视网络（空间）安全问题。我国成立了由最高领导人担任组长的国家网络安全领导小组，负责制定和指导关键任务信息系统及其安全的规划和建设。习近平总书记提出了“没有网络安全就没有国家安全”的论断。目前，我国的核心信息系统主要还是运行在来自美国的IT垄断企业的基础系统和平台之上，摆脱这种技术依赖是IT业界和关键应用行业的当务之急。针对这种状况，互联网业界从成本考虑，提出了“去IOE”（即摆脱对IBM主机、Oracle高性能数据库以及EMC高端存储的依赖）的口号。对于国家核心信息系统，这不仅仅是成本问题，更是安全问题。因此，“技术先进、企业领先、安全可靠、自主可控”已经成为我国发展信息技术和系统的基本战略。这对从事IT研发和人才培养的

专业人士提出了很高的要求。为满足这一要求，需要与时俱进，从新的学科角度来审视面临的挑战和机遇，寻找实现“跨越式发展”和“弯道超车”的发展途径。

2015年3月5日，李克强总理在政府工作报告中发出了“大众创业、万众创新”的号召，得到了全社会的积极响应。回顾一下我国的创新发展战略，改革开放以来的30多年，大致经历了从以“星期日工程师”为标志的大学创新，到“企业是创新的主体”的企业创新，再到2011年胡锦涛总书记提出的“协同创新”，一直到当前的“大众创业、万众创新”4个阶段，创新一直被高度重视。自1996年4月江泽民总书记提出“创新是一个民族进步的灵魂”的论断以来，迄今也有20年时间，离2020年建成创新型国家的时间节点也日益迫近。党的十八大以来，随着“两个一百年”奋斗目标和实现中华民族伟大复兴的“中国梦”的提出，“创新驱动发展”作为国家的发展战略被提到前所未有的高度，凸显了新一代领导人对于创新的高度重视。从李克强总理提出的“互联网+”理念以及在各种场合对创新创业的解读来看，中国互联网企业的巨大成功是“大众创业、万众创新”最好的注解，互联网本身作为人和人之间的连接平台，为创新创业提供了崭新的环境。互联网和“互联网+”的成功与否本质上就取决于大数据技术的发展和应用。在当前的创新创业背景下，探讨数据科学与工程学科恰逢其时。

2.2 IT人才市场变化的需求

信息技术作为近年来发展最快的领域，人才市场需求的变化也最为明显。2006年是一个转折点，这个转折点的标志性事件是，百度作为国内互联网企业，第一次对国内高校的毕业生给出了比老牌的跨国IT企业更高的薪酬。在那之前，国内高校的大多数毕业生是以拿到那些著名跨国IT企业提供的职位为追求目标的。其深层次的原因在于，国内的信息系统都是架构在这些跨国IT企业的基础系统或平台之上的，国内的IT企业实际上就是系统集成商或是解决方案提供商，所有源头的核心技术都不掌握在自己手里，我国培养的IT人才要做的就是用好垄断企业的系统和平台，最多需要再做些简单的二次开发。垄断企业对优秀人才的吸引也进一步枯竭了我国自主创新和研发的能力。

近10年来，以BAT（指百度（B）、阿里巴巴（A）、腾讯（T））为代表的中国互联网企业在商业上取得了被世人认可的巨大成功，这对于我国信息技术产业以及其他相关领域的影响也同样巨大。当然，互联网企业不是IT企业，因为它不提供诸如硬件、软件或是咨询服务、解决方案等传统IT企业提供的产品，它只是第三产业中的信息服务业企业。但是，对互联网企业而言，IT能力是其核心竞争力。互联网企业的IT能力建设不依赖于传统的IT企业，这一事实有着非凡的意义：一是破除迷信，打破了IT界以往对于传统垄断性IT企业的盲目崇拜，以为那些高端的技术和系统是他们的独门秘籍，是我们望尘莫及的；二是解放思想，使得各行各业可以效仿互联网业界，针对自身的应用需求，融会贯通地利用掌握的IT知识和开源技术，从应用需求出发，从硬件体系结构到网络架构再到软件系统直至应用软件，度身定制所需要的IT系统和平台。这带来的不仅仅是成本的降低，更重要的是可以对创新型商业模式的开发提供有效的支持。商业模式是服务业企业的生命线，创新型商业模式的开发依赖于“数据科学家”，企业IT能力的建设依赖于“系统架构师”。

在我国，虽然经济下行没有影响IT的

就业形势，但是市场上对IT人才的需求与高校能够提供的人才相比还是有很大的差距，这表现在企业需要的合格的“系统架构师”和“数据科学家”很难直接从学校招到。这一点在高校表现尤为明显，课堂和实验室学的东西远离市场需求，厌学频发。

2.3 技术和产业发展的需求

现有的计算机或IT技术和系统是基于三四十年以前的硬件技术水平而研发的。最近十几年以来，硬件技术产生了突飞猛进的发展。CPU从多核走向众核、万兆以太网等网络连接技术的成熟、新型存储设备和非易失存储介质的研发成功、计算机新型体系结构的探索，这在很大程度上打破了大多数沿用至今的IT技术和系统的假设前提。表1展现了硬件技术近40年以来的迅猛发展，也说明了其发展的不均衡性。如何充分发挥硬件技术发展的潜力，是传统的IT企业在考虑其优势产品升级换代时重点考虑的问题，但由于基本假设前提的不吻合以及本质上的不适应，想做到和硬件发展与时俱进几乎是不可能的。

为了充分利用硬件技术的发展，也为了降低成本和契合现实应用的实际需求，人们开始了围绕应用进行定制式的系统研发和部署。也就是说，针对应用进行垂直式的系统架构设计和功能模块开发，从计算平台搭建和系统软件开发，直到应用的开发都是为解决目标应用而做的。相对这种垂直式的技术研发，传统的IT系统发展是水平式的，从计算机系统到系统软件再到中间件都是通用或相对通用的，应用开发人员要做的工作就是选型、系统集成，然后再进行应用层的开发和部署。GFS[3]和MapReduce[4]就是这种垂直开发的典型例子，为了解决Google公司的PageRank问题，内部人员开发了存储网页数据和日志数据的文件系统GFS以及其上的便于分布并行处理数据的MapReduce编程界面。如果说“one size fits all”是传统的理念，那么垂直式定制化的研发就是秉承“one size fits a case”理念[5]。这种探索，不仅可以充分利用硬件技术的最新成果，更能体现IT领域“应用驱动创新”的基本精神。

Hadoop的成功开源以及对以后数据管理领域产生的巨大影响昭示了新的技术发展趋势，那就是开源社区和技术生态的重要性[6]。这和我国时下倡导的“万众创新”也非常吻合。正是通过开源，吸引更多的人致力于技术的研发或是应用，反过来又贡献于开源社区，产生创新的正循环。Hadoop开源的成功也给了开源技术鼻祖的美国加州大学伯克利分校以有益的启示，AMP实验室的开源系统Spark成为来源于大学实验室的成功开源系统[7]。通过开源，可以把来源于应用的垂直式定制化的技术和系统推广到其他应用领域，并吸引广大技术人员参与研发和创新。把一种只适合于某一个具体应用的技术和系统变成适合于一类应用，这就是

表1 40多年来硬件技术发展对比

	内存				处理器核数	网络带宽
	寻址时间	带宽	容量	价格		
20世纪70年代	850 ns	300 MB/s	128 KB	1美元/KB	1	2.94 Mbit/s
2010年至今	100 ns	32 GB/s	1 TB	8美元/GB	192	100 Gbit/s
提高程度	8.5倍	100倍	800 万倍	12 万倍	192倍	34 013倍

实现所谓的“one size fits a bunch”[5]。开源和技术生态建设是当前技术发展的重要趋势之一。

2.4 国内外现状分析

在开设数据科学和工程相关课程方面，美国的加州大学伯克利分校、伊利诺伊大学香槟分校、哥伦比亚大学、纽约大学等从2011年开始就进行了卓有成效的尝试。纽约大学、华盛顿大学等著名高校已经开始设置硕士学位培养计划。在我国，从2012年开始，清华大学、中国人民大学、复旦大学、北京航空航天大学等高校也开始设置了学术型或专业型硕士学位培养计划。

在本科专业设置方面，上海纽约大学从2015年4月份开始就在内部讨论设置一个数据科学的学士学位，除了计算机系的教授外，商学院、设计学院等教授也参与其中，并计划于2015年9月开始招收本科生。

2015年6月7日，中山大学宣布成立“数据科学与计算机学院”，整合了与计算机相关专业的优势资源。2015年5月27日，复旦大学在其110周年校庆日宣布筹建“大数据科学与技术学院”。相信未来会有更多的学校在学科设置和学院建制方面进行新的探索。

华东师范大学从2007年成立海量计算研究所以来，一直致力于培养海量数据处理领域的人才，探索数据科学与工程领域的协同创新和人才培养道路；2012年，华东师范大学在国内外伙伴企业和兄弟高校的支持下，成立了云计算与大数据研究中心；2013年，华东师范大学宣布成立国内第一个数据科学与工程研究院，重申协同创新的理念，聚焦中国式应用，进行大数据技术和系统研发以及创新人才培养。

3 数据科学与工程学科特点

3.1 应用驱动创新

虽然互联网是推动大数据热的始作俑者，但广泛来说，大数据不仅仅局限于互联网数据。要讨论这林林总总的数据，从认识论的观点来看，首先就是要对大数据进行分类，这非常必要，它是确保大家在同一论域进行讨论的前提。按照笔者的理解，大数据大致可以分为Web数据、决策数据、科学数据三大类。顾名思义，Web数据是与Web相关的数据，包括网页、链接、日志等具体类型，门户网站、搜索引擎、社交网络、电子商务等以Web形式呈现或以Web为载体的新型信息服务系统产生的数据大多可以归纳为此类型。决策数据主要是指由传统数据库和数据仓库管理的、在生产过程中产生的数据，是用于决策的数据，也可称为商务智能（business intelligence，BI）数据。科学数据实际上是最早的一类大数据，包括科学实验数据、科学观测数据、科学文献数据、设计数据等，这类数据与科学领域密切相关，品种最多，研究最难，若没有领域专家的参与，IT专家难以胜任科学数据的管理和分析任务。

关于大数据研究的认识，笔者也有一个3个层次的观点。大数据的研究全景可以看作一个倒立的三角形，如图1所示。这个倒立三角形分为3层：第一层代表形形色色的各种应用，这些应用是数据的来源，也是数据的应用场所；第二层（中间一层）代表模型和算法，是指把对应用进行理解、抽象、建模，然后在底层的计算平台上予以实现[8]；第三层（最下面的一层）就代表IT计算系统或平台，这是传统信息技术行业

关心和擅长的领域[9~11]。这3个层次中，第一层中每一类应用有各自对应的学科去深入研究；第二层是有关模型和算法的；第三层对应的学科就是计算机或IT学科。

第一个层次是大数据应用层次，大数据应用是一个从科学研究、企业管理到电子商务、搜索引擎的完整谱系。这个层次涉及的人员来自各个领域，包括领域专家、用户和客户等。在理解现实应用的基础上进行建模，再选定合适的技术和系统予以实现，这体现了应用驱动创新的特点。

3.2 多学科交叉融合

随着大数据成为当前的热点，信息技术发展的重点从计算转向数据，数据的有效应用变得至关重要。数据科学就是在这一背景下产生和发展起来的。数据科学通常指基于计算机科学、统计学、信息系统等学科的理论和技术，研究数据的收集整理以及从海量数据中分析处理，获得有效知识并加以应用的新兴学科；数据工程是指利用工程的观点进行数据管理和分析以及开展系统的研发和应用。

数据量的爆炸式增长不但改变了人们的生活方式、企业的运营模式，也改变了科学研究的基本范式。数据科学和工程可以作为支撑大数据研究与应用的交叉学科，其理论基础来自多个不同的学科领域，包括计算机科学、统计学、人工智能、信息系统、情报科学等。数据科学与工程学科的目的在于系统深入地探索大数据应用中遇到的各类科学问题、技术问题和工程实现问题，包括数据全生命周期管理、数据管理和分析技术和算法、数据系统基础设施建设以及大数据应用实施和推广。培养具有扎实理论功底和大数据思维的数据科学与工程方面的高层次专门人才，推动与大数据相关的理论体系的建设和技术的进步，为解决各行各业中遇到的大数据管理和应用问题提供人才和技术储备。因此，多学科交叉融合也是数据科学与工程学科的另一个特点。

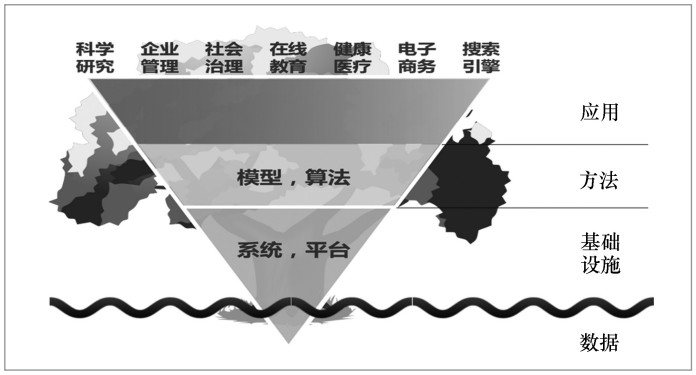


图1 大数据研究全景

3.3 学科的基础内涵

与传统计算机和软件工程等学科相比，数据科学与工程学科具备独特的学科基础和内涵。数据科学与工程学科的理论基础涉及统计分析、商务智能以及数据处理基础，具体包括以下几个方面。

- 大数据表达理论方面：包括大数据的生命周期、演化与传播规律，数据科学与社会学、经济学等之间的互动机制以及大数据的结构与效能的规律性。
- 在大数据计算理论方面：研究大数据的表示以及大数据的计算模型及其复杂性。
- 在大数据应用基础理论方面：研究大数据与知识发现，大数据环境下的实验与验证方法以及大数据的安全与隐私。

相比较而言，计算机科学学科是研究算法的科学，而数据科学不局限于此，其研究对象是数据，随着计算机应用从以计算为中心逐渐向以数据为中心的迁移，数据科学与工程学科的内涵和外延更加宽泛。软件工程学科中的相关技术提供了数据分析处理的工具以及具体开发时的范式。数据处理技术是数据研究领域的一种

重要的研究方法，用于研究和发现数据本身的现象和规律。

数据科学与工程也不同于传统的商业智能和统计学，商业智能主要从商业模式、经济管理的角度对数据应用进行研究，而统计学提供具体的数据分析处理的方法论，但是面对PB级以上的海量数据，大数据的分析不能停留在获得概率分布结果，也不能满足于对细节问题的数据挖掘，而是需要更简单、有效的问题求解方法，争取从大数据中获得新的知识，构建新的应用范式。

3.4 学科的知识体系

数据科学与工程作为一个大数据时代的新兴交叉学科，主要的知识结构来源于计算机科学、应用数学以及信息系统和信息管理3个学科，但是也和这3个学科分别都有很大的不同。在当前大数据时代，从知识结构和人才培养角度来看计算机、软件工程学科，不难得出，其知识结构过于老化，教材和课堂上传授的知识基本属于“博物馆”和“百科全书式”的内容，还是服务于垄断企业的IT产品和系统，对于知识的融会贯通和综合应用不够重视。这也导致学校教育无法满足人才市场的需求，出现学生厌学、老师厌教的现象。而综合应用和融会贯通是互联网企业和开源社区最为重视的方面，也是一个“系统架构师”必须具备的能力和素养。应用数学学科也很强调与信息学科和产业的结合，这一点从“计算数学”专业的更名历史就可略见一斑，计算数学1987年更名为“计算数学及其应用软件”，1998年教育部将其更名为“信息与计算科学”专业。但是，这一专业在招生和就业方面频频亮起红灯。究其根本原因，想必就是没有真正和现实应用相结合，也许是因为我国单纯的数学背景的院系缺少这方面的基因。信息系统和信息管理专业非常重视企、事业单位的应用，关注需求和机构组织，这是实现应用系统至关重要的因素。但因为在管理学院或商学院，数学和计算机的训练相对薄弱，在针对应用的数学建模和信息系统的工程实现方面就难以胜任。

根据前面所描述的大数据全景图（如图1所示），数据科学与工程学科的知识体系构建的基本原则是：针对不同的应用，本学科培养的人才可以充分理解应用需求，利用合适的数学工具进行建模，同时能够根据具体的应用搭建计算环境和平台，并进行有效的算法实现。

在计算机学科方面，主要包括新型的专用型计算平台的搭建，这涉及互联网计算架构、新硬件的应用以及开源系统的使用等。由此倒推，需要对计算机学科的现有知识体系进行裁剪，舍弃那些与系统和平台搭建无关的知识。在应用数学方面，着重于对数学建模工具的灵活掌握，具体而言，就是对概率论、数理统计以及矩阵计算（计算方法）等工程数学能活学活用，既能利用这些数学工具来抽象具体的现实应用，又能进行有效的算法实现。在信息系统学科方面，需要培养数据全生命周期管理的基本理念，从数据的生成和收集，到数据的存储和管理，再到数据的使用和共享，实现数据的价值。

4 数据科学与工程学科建设

信息技术和互联网是创新创业的最前沿，在专业教学和人才培养中践行创新创业教育。“万众创新”其实就是“草根创新”，“草根创新”的本质就是立足应用，解决应用中遇到的现实问题。我国成功的互联网公司就是典型的“草根创新”，其基

本的途径是通过商业模式设计，着重用户体验，利用开源技术，搭建服务平台，部署应用，收集反馈信息，再进行完善和优化，形成一个完整的创新链条。如果说“草根创新”是从应用出发，以追求商业价值为驱动力，那么还需要“精英创新”配合进行概念抽象和应用推广。大学的师生作为有学术情怀的“精英”阶层，需要主动对接创业企业的“草根创新”，这样才能把在实际应用中获得的创新固化下来，并广为传播，同时也能养成学生对创新创业的深入理解。

设计思维对于践行“大众创业、万众创新”有着重要的参考意义。设计思维的本质就是尽一切可能站在用户的角度看问题，设身处地地体验用户需求，进行社会化思考，通过原型设计和试用，经过反复迭代完善产品设计。这是互联网上的服务产品的典型开发过程，应用设计思维进行工业产品设计是当前的趋势，在国际顶级的商学院和设计学院成为必修课程，会很快渗透到各个学科的人才培养计划中。破除迷信，解放思想，需要克服传统思维定势，从思想观念上主动对接当前提倡的创新创业理念。

4.1 科学研究和系统开发

数据科学与工程学科是一个面向应用的综合交叉型学科，学科交叉和协同创新是开展科研开发的基本途径。立足中国式应用，瞄准国际研究前沿，通过与企业或用户的合作，提高研发能力和应用能力，研发具有中国特色的大数据技术和系统，为大数据应用中的数据采集、整理、存储、维护、分析等管理任务提供全方位的支持，提供公共技术平台、大数据应用部署咨询服务等。

通过和企业合作伙伴的密切合作，落实应用驱动研发的战略。研发工作根据与企业合作的成熟程度，切实做到科学研究与生产实践相结合，克服科研和生产“两张皮”的现象，闯出一条我国数据管理技术和系统研发的可持续发展的新路。

立足“数据科学与工程”学科特色，发挥高校在技术综述、核心技术研发、原型设计与开发上的优势，秉承“one size fits a bunch”的理念，面向行业应用，充分了解需求，在应用抽象的基础上，从核心技术研发出发，通过原型系统开源，逐步从理论与技术验证走向系统试用和最终应用。在这一过程中，营造或融入以开源社区为中心的技术生态圈，催生技术型初创公司或促成研发成果的技术转化，在人才培养的同时，实现科研成果的推广应用。

4.2 数据科学与工程学科人才培养

围绕计算机、应用数学和信息系统设计从本科生到博士生的人才培养方案，结合开源技术与时俱进地更新计算机教学，结合应用实践加强数理统计和矩阵计算等建模和算法训练，培养“系统架构师”和“数据科学家”，这也是当前最需要的两类人才。

基于以上培养目标，针对本科生、硕士研究生、博士研究生各自的学制和教学特点，专业的课程设计遵循以下指导思想。

- 突出数据科学基础课程教学：结合统计、应用数学等学科的优势，在强调概率论教学的同时，将数理统计、数值计算与优化、机器学习、数据挖掘、信息检索、自然语言处理等课程作为重要的专业必修课或选修课在本科教学阶段进行讲授，为研究生阶段讲授统计学习理论、概率图模型、语言模型、信息抽取与集成、海量数据分析与挖掘等高阶课程打下扎实的基础。
- 裁剪传统计算机和信息系统类课程，适应新技术发展和应用场景：在操作

系统、数据库系统、计算机体系结构、编译原理、分布式系统等传统计算机课程中，弱化历史性材料的讲授（作为课外阅读作业），补充相关系统与应用的最新进展。例如，补充云计算系统中的资源调度、大数据系统（如Hadoop）、集群搭建与实践、新型编程范型（如MapReduce）及其编译执行等内容，弥补近年来技术与应用快速发展造成的传统课程教学内容和教材与时代的脱节。

● 强调数据管理与处理的全生命周期：结合情报和信息管理等学科的优势，课程覆盖数据从获取、整理、存储、索引，到查询与检索、分析与挖掘、加工与展现的整个生命周期的基础理论、技术方法以及系统。在课程设置上，通过在本科低年级设置计算机系统、信息管理与信息系统等专业必修课，在高年级开设开源软件、大数据系统等专业选修课，达到从宏观角度介绍数据全生命周期、联接相关课程的目的。

● 充分利用企业和行业力量，强调设计思维（design thinking），提升课程实用性：开设计算广告、智慧城市、社会计算、推荐系统等具有较强实用性的选修课，由企业兼职教师单独或与专职教师联合讲授，突出应用场景抽象、问题建模、案例分析、原型系统搭建、结果评测等环节的教学，将学生所学的基础理论和方法与应用联系起来，同时培养学生针对实际应用的发现问题、分析问题、解决问题的能力。

5 结束语

互联网改变了一切，也改变了信息技术的发展范型。IT领域当前的热点无疑是云计算和大数据，是互联网企业而非传统的IT企业推动了云计算和大数据的发展。这一现象的意义在于，IT的发展范型发生了改变，“应用驱动创新”成为IT领域创新链上的重要环节。互联网企业IT能力建设的巨大成功，破除了“迷信”；硬件技术的飞速发展为新一代IT技术的发展奠定了基础；“安全可靠、自主可控”的国家安全战略的提出和落实对我国IT界而言是挑战，更是机遇。基于以上3点，再加上我国经济社会发展提出的丰富、迫切而又极具特色的信息化应用需求，可以看出，当前我国IT界处在充满机遇的窗口期。如何利用这个难得的时间窗口实现跨越式发展和弯道超车，不仅需要认真分析和清晰认识现实的创新机遇，更需要适时定义和发展新的学科方向，探索学科实质内涵，明确知识结构，开展人才培养，从而进行持续、有效的“万众创新”行动，全面激发创新活力。

参考文献

[1] Hey T, Tansley S, Tolle K M. The Fourth Paradigm: Data-Intensive Scientific Discovery. USA: Microsoft Rr, 2009

[2] Manyika J, Chui M, Brown B, *et al*. Big Data: the Next Frontier for Innovation, Competition, and Productivity. USA: McKinsey Global Institute, 2011

[3] Ghemawat S, Gobioff H, Leung S T. The Google file system. Proceedings of the ACM Symposium on Operating Systems Principles(SOSP), Lake George, NY, USA, 2003: 29~43

[4] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Proceedings of the 6th Symposium on Operating System Design and Implementation, San Francisco, USA, 2004: 137~150

[5] Stonebraker M, Cetintemel U. One size fits all: 10 years later. Proceedings of International Conference on Data Engineering, Seoul, Korea, 2015

[6] White T. Hadoop – The Definitive Guide: Storage and Analysis at Internet Scale (4. ed., revised & updated). USA: O'Reilly Media, 2015

[7] Stoica I. A berkeley view of big data: algorithms, machines & people. Proceedings of Berkeley EECS Annual Research Symposium, California, USA, 2011

[8] 美国国家学术院国家研究委员会. 海量数据分析前沿. 华东师范大学数据科学与工程研究院译. 北京: 清华大学出版社, 2015
National Research Council of the National Academies. Frontiers in Massive Data Analysis. Translated by Data science and Engineering Research Institute of East China Normal University. Beijing: Tsinghua University Press, 2015

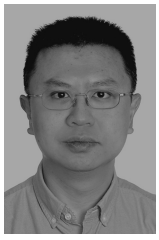
[9] 李战怀, 王国仁, 周傲英. 从数据库视角解读大数据的研究进展与趋势. 计算机工程与科学. 2013, 35(10): 1~11
Li Z H, Wang G R, Zhou A Y. Research progress and trends of big data from a database perspective. Computer Engineering & Science, 2013, 35(10): 1~11

[10] Abadi D J, Agrawal R, Ailamaki A, *et al*. Proceedings of The Beckman Report on Database Research, California, USA, 2014: 61~70

[11] Jagadish H V, Gehrke J, Labrinidis A, *et al*. Big data and its technical challenges. Communications of the ACM, 2014, 57(7): 86~94

作者简介

周傲英, 男, 华东师范大学长江学者、特聘教授、数据科学与工程研究院院长, 主要研究方向为Web数据管理、数据密集型计算、内存集群计算、分布事务处理、大数据基准测试和性能优化。



钱卫宁, 男, 华东师范大学数据科学与工程研究院教授、博士生导师, 主要研究方向为互联网环境下的数据管理、大数据管理系统评测基准、社交媒体数据分析、知识图谱构建与应用等。

王长波, 男, 华东师范大学教授、博士生导师、软件学院常务副院长, 主要研究方向为信息可视化、大数据可视分析、计算机图形学。

收稿日期: 2015-06-28
论文引用格式: 周傲英, 钱卫宁, 王长波. 数据科学与工程: 大数据时代的新兴交叉学科. 大数据, 2015022
Zhou A Y, Qian W N, Wang C B. Data sciences and engineering: an emerging interdisciplinary in the big data era. Big Data Research, 2015022