大数据时代的数据科学家培养

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 200433；2.上海市数据科学重点实验室，上海 200433

摘要
大数据时代，最热门的职业是数据科学家（data scientist），而不是传统的信息科学家，也不是大数据工程师。大数据热潮促进了数据科学（data science）研究进入快速发展期，数据科学家的培养也受到广泛重视，越来越多的大学启动数据科学学位培养计划，但值得注意的是，当前数据科学家培养的基础条件缺乏，其知识结构、学科体系、人才培养计划尚未建立。结合大数据时代的人才要求，给出了科学、系统的数据科学人才知识体系，提出了超学科、多类型的培养模式。

关键词
数据科学；数据科学家；人才培养；大数据

中图分类号：TP399　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016035

Training data scientists in the era of big data

ZHU Yangyong[1,2], XIONG Yun[1,2]
1. School of Computer Science, Fudan University, Shanghai 200433, China
2. Shanghai Key Lab of Data Science, Shanghai 200433, China

Abstract
In the age of big data, data scientist has become a hot occupation, supplanting traditional information scientist and big data engineer. Big data boom has been pushing data science research into fast development phase. How to train data scientists has been paid widespread attentions. Many universities launched data science degree training plans. The current situations in data scientists training were analyzed. The achievements of training data scientists in Fudan University were summarized. A systematical data scientists training plan was proposed.

Key words
data science, data scientist, talents training, big data

1 引言

数据是网络空间(cyberspace)的唯一存在，网络空间的数据呈现出不可控、未知性、多样性、复杂性等自然界的特征，网络空间的所有数据组成了数据界(datanature)[1,2]。2008年，朱扬勇等指出“数据资源是重要的现代战略资源，其重要性越来越显现，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”[3]。数据资源作为一种基础性、战略性资源得到了空前关注，数据资源的开发利用被许多国家提高到了战略高度，纷纷出台大数据战略。

提高数据资源开发利用水平、保护国家的战略资源是增强我国综合国力和国际竞争力的必然选择[3]。对数据资源的开发利用已形成数据产业，其产业链主要包括：从网络空间获取数据并进行整合、加工和生产；数据产品传播、流通和交易[4]。代表性企业有Google、Facebook、百度、万得资讯、万方数据等。在这个新的生产链上急需数据人才。不仅如此，越来越多的领域发现数据的价值，大数据[5]对人类社会发展、科学研究、经济建设、文化生活的各个领域正在产生革命性的影响。于是，数据科学家作为一种最热门的职业在工业界已经受到追捧，例如电商、广告媒体、汽车制造行业等都在寻找数据科学家为其探寻数据价值，赢取利润高点。

早在2011年，McKinsey公司预测到2018年，仅在美国本土就可能面临缺乏19万具备深入分析数据能力人才的情况，同时具备通过分析大数据并为企业做出有效决策的数据管理人员和分析师也有150万人的缺口[6]；美国专业招聘公司罗致恒富(Robert Half)公布的《2015薪资指南(2015 salary guide)》①也已把大数据人才列为薪资涨幅最大的六大行业之一。

目前，数据人才主要来自IT、管理、与企业相关的专业领域等各个方向，通过相互合作形成数据分析决策。但是，情况正在发生变化，例如，Google Translate 团队在一次招聘中仅招收多名计算机科学家，却没有招收一名语言学家，并且其部门主管 Franz Josef Och是计算机科学家，并不精通语言学。这个案例说明，培养数据科学家并不是将几种技能的人简单地聚集成一个团队，而是应该探索一种转型模式，Google Translate团队中这些计算机背景的人才是正在向真正数据科学家转型的新型人才。

然而，目前数据科学人才培养、数据科学学科建设等刚刚起步，尚未形成持续为社会培养和输送不同层次数据人才的教育培养体系。面对大数据时代数据人才紧缺现状，大学有必要尽快研究数据科学学科构成和新型数据人才的培养体系，开设数据科学学科专业，提升人才培养和输出能力。

2 数据科学家培养现状

大数据时代，最热门的职业是数据科学家，而不是传统的信息科学家，也不是大数据工程师。在此之前，大学没有设置数据科学学科和专业。近几年，数据科学家培养开始受到大学的重视，并快速发展。

2010年起，各国大学开始了数据科学人才培养工作。美国哥伦比亚大学从2011年起开设《数据科学导论》课程，

① www.roberthalf.com/salary-guides

2013年起开设《应用数据科学》课程以及“数据科学专业成就认证”培训项目，并从2014年起设立硕士学位，2015年起设立博士学位②；美国加州大学伯克利分校从2011年起开设《数据科学导论》课程③，并从2012年起开设《数据科学和分析》课程；美国伊利诺伊大学香槟分校从2011年起举办“数据科学暑期研究班”④；美国纽约大学从2013年秋季起设立“数据科学”硕士学位⑤；美国华盛顿大学从2013年5月起开设《数据科学导论》课程，并对修满数据科学相关课程学分的学生颁发数据科学证书⑥；美国芝加哥大学开设夏季数据科学培训课程⑦；美国南加州大学设立“数据科学”硕士学位⑧；美国雪城大学也提供数据科学高级研究证书培训项目⑨；英国邓迪大学从2013年起设立“数据科学”科学硕士学位⑩。

在中国，复旦大学从2007年起开设数据科学讨论班，2010年开始招收数据科学博士研究生，并从2013年起开设研究生课程《数据科学》，2014年开始举办数据科学家训练营，2015年开始正式招收数据科学专业研究生以及本科第二专业学位⑪；香港中文大学自2008年起设立了“数据科学商业统计”科学硕士学位⑫；清华大学于2014年推出大数据硕士项目，并于2014年9月开始招收研究生。

尽管数据科学的学位项目大量出现，但是，对数据科学家的培养还缺少统一的认识，具体表现在两个方面。

（1）数据科学缺少统一的认识，研究机构发展迅速，但学科体系还没有建立

事实上，数据科学已经发展了很多年，远比大数据早，1966年，Peter Naur建议计算机科学应该被称为Datalogy，即“关于研究数据使用和本质科学”[7]。2009年，朱扬勇等对数据科学进行了定义，并引入Dataology一词[1]。2008年《Nature》、2011年《Science》都出版了关于数据研究的专辑，随后Microsoft、IBM、Google等公司都开始了大数据技术研究，大数据热潮促进了数据科学研究进入快速发展期。国内外纷纷成立数据科学研究机构，例如，美国哥伦比亚大学数据科学和工程研究院、美国纽约大学数据科学研究中心、英国帝国理工学院数据科学研究院、中国科学院虚拟经济和数据科学研究中心、上海市数据科学重点实验室、清华大学数据科学研究院、中山大学数据科学学院、华东师范大学数据科学与工程研究院等。

然而，数据科学还缺少统一的认识。当前，数据科学概念和观点出现在科学数据处理领域、计算机科学领域、统计学领域、商业智能应用等方面。这些概念和观点的基本思想是：认为数据科学是“从领域数据中获取知识，为现有的科学研究、管理决策提供服务”。这些工作还不足以形成一个新的科学，因为它们的研究对象仍然是现实中的事物，并且相应的科学问题也都是现有科学领域的问题，数据科学学科体系尚未建立。

（2）数据科学家的知识结构还没有形成统一框架

信息化是一个生产数据的过程。目前，几乎所有领域都已经或正在信息化，都或多或少地使用计算机来解决遇到的数据存储和数据计算问题，计算机科学与技术无疑成为数据科学家的基本技能。现有的数据科学家很大一部分来自于计算机学科，具备计算机科学相关专业背景，掌握处理大数据所必需的Hadoop、Spark、Mahout等大规模并行处理技术、数据挖掘与机器学习知识。但是，数据科学的研究对象、目的和方法等都与计算机科学、信息科学和知识科学有本质的不同[1,2]，仅仅具备这些计算机技能并不能被称为一个真正的数据科学家。

科学研究的对象也信息化了，变成了计

② idse.columbia.edu/certification-professional-achievement-data-sciences

③ datascienc.es

④ mias.illinois.edu/

⑤ datascience.nyu.edu/academics/programs/

⑥ www.pce.uw.edu/certificates/data-science.html

⑦ www.ci.uchicago.edu/blog/summer-data-science-fellowship

⑧ www.cs.usc.edu/academics/masters/msdata.htm

⑨ ischool.syr.edu/future/cas/datascience.aspx

⑩ www.computing.dundee.ac.uk/study/postgrad/degreedetails.asp?17

⑪ www.datascience.cn

⑫ www.sta.cuhk.edu.hk/mscdbs/DBSdefault.asp

算机中的数据，并且需要处理的数据越来越多，形成了专门的科学数据处理领域，于是有了生物信息学、地理信息学、行为信息学等。科学家可以通过研究数据来研究自然和行为，数据科学为科学研究提供了数据方法。于是，数据科学家的培养逐步发展为多领域联合培养。在培养过程中，领域专家重点是学习如何将领域业务需求转化为数据问题交给数据分析人员，并不关注数据处理细节；而数据分析人员注重对领域专家所给的数据集进行处理，缺乏对领域知识的理解。这是目前数据科学家培养的常见方式，但却缺乏系统性。

3 如何培养数据科学家

自计算机发明以来，人们一直在处理和使用数据，主要工作是将现实的东西用计算机数据表示，存储在计算机中，然后管理这些数据，并在需要时使用它们。随着数据量的不断、快速增长，对数据的处理分析变成了科学研究、商业应用的一个重要环节，而这样的数据分析工作往往依靠人的创造性，于是从事商业数据分析、科学数据分析的人被称为数据科学家。近年来，对数据分析理论和技术的一些共性需求导致对数据本身的研究，例如，分析数据本身的现象和规律；研究数据每年的增长规律；预测10年后网络空间数据的规模等，这样就出现了专门研究数据自身规律和现象的数据科学家。

上述“从事商业（业务）数据分析的人”、“从事科学数据分析的人”、“研究数据的人”是目前被称为数据科学家的3类人。但在解决一个大数据分析问题时，常常是由来自于数学与统计、计算机和业务领域的一个数据科学家团队完成。这说明，目前在大学没有什么专业具备了数据科学家所需要的知识，这是一个新问题。

下面，以精准营销与数据相关的业务为例，讨论数据科学家做什么工作。

简单地将一个互联网精准营销描述为：“将商品推荐给可能购买的人群”。其具体实施则涉及工程、技术和科学3个层次的工作，见**表1**。

（1）工程实施

作为互联网广告，当用户上网登录页面时，需要在不到100 ms的时间内将广告弹出，这主要是一个工程实施的问题。

（2）业务模型和技术手段

精准营销的业务模型包括商品分类和人群分类，对应的技术手段主要是聚类分析等数据挖掘技术。

（3）科学研究

聚类分析的核心是相似性及其计算。如何确定两个客户是相似的，这是一个科学问题，需要科学家创造性地劳动。不同的相似性设计会导致不同的聚类结果，不同的聚类结果会导致不同的营销精准性，最后导致营销效果，即商品销售。

“精准营销”的例子也说明了为什么叫做数据科学家。因为这些人是在从事创造性的工作，是在发现数据的现象和规律，而不是从事制造性工作，他们的工作结果会有不确定性，因此是一项科学工作，所以他们是数据科学家。

2014年，Cleveland W S提出了一个数据科学行动计划，指出了数据科学需要发展的重要方面（跨领域数据分析能力、数据建模和方法、数据计算能力、学科规划、工具、基础理论）[8]。笔者认为，

表1 数据科学家做什么（以精准营销为例）

互联网精准营销的需求描述	数据科学家的工作
在希望的时间内完成	工程实施
构建精准营销模型和聚类分析	业务模型、技术手段
定义和计算相似性	科学研究

数据科学是研究数据界的科学或关于数据的科学[1,2,9,10]，主要由两部分组成⑬：一是研究数据本身的规律和现象，解决关于数据界的科学问题，这部分研究工作并不考虑数据的现实含义，只研究数据自身的现象和规律，包括数据的历史、进化和迁移等；二是研究数据表示的现实含义的现象和规律，即通过研究数据来研究现实，是指数据科学为传统科学研究提供了方法，其目的在于揭示自然界和人类行为的现象和规律。相应地，数据科学的主要研究内容包括：数据科学基础理论研究，如数据相似性、数据测度、数据代数、数据实验、数据分类、数据百科全书等；数据界探索，如数据界有多大、全球数据如何增长等科学问题；科学研究的数据方法，如数据方法的框架；数据技术研究，如数据分析、数据探索、数据挖掘、数据伪装和辨伪、领域驱动的数据技术（如生物信息学、业务智能（business intelligent，BI）和社会计算等）。

数据科学学科结构布局与数据科学的研究内容是对应的：将数据科学基础理论研究为基础，尤其是数据相似性理论研究是数据研究的关键和基础，这是第一类数据科学家——“研究数据的人”的基础知识结构；数据界探索作为数据科学的科学问题的探索，并且与社会科学、自然科学形成差异和支持，突出数据科学学科特色，这是第一类数据科学家——“研究数据的人”必备的知识结构；科学研究的数据方法是对现有科学研究创新研究方法，是数据科学学科的重点内容，涉及各个科学研究领域方向，这是第二类数据科学家——“从事科学数据分析的人”必备的知识结构；数据技术研究是数据科学学科的技术支撑和应用体现，这是第三类数据科学家——“从事商业（业务）数据分析的人”必备的知识结构。因此，数据科学在人才培养方面将打破原有的学科限制，数据科学家需要的知识结构是涵盖和横跨不同学科，融合多学科的研究方法，甚至取代并超越它们，是一种新的视角和一种新的学习体验，即超学科[11]。

数据科学家培养应该是多类型的，包括学位培养、科研人员培养和应用人才培训。学位培养和科研人员培养的主要是在数据上做科学研究的人以及研究数据的人；而应用人才培养的主要是从事商业数据分析的人。并且，不同类型人才的培养在整个知识体系结构中的侧重是不同的，其重点掌握的知识层次是有所划分的，具体如下。

（1）学位培养

针对未来从事研究数据本身的人的学位培养，应该注重数据基础理论的训练，要求掌握各种数据技术；针对未来从事在数据上做科学研究的人的学位培养，则应该注重学生对专业领域知识的掌握以及对领域数据学的培养，提升在专业领域的数据能力。

（2）科研人员培养

主要是指获得数据科学学位后，继续从事科学研究活动的人。这里指的从事科学研究活动，包括从事数据科学研究和从事社会科学或自然科学研究。他们已经具备了学位培养期间的专业训练，需要进一步提升他们的数据创新能力。

（3）应用人才培训

主要针对从事商业数据分析的人才，这里包括获得数据科学学位后从事商业数据分析的人以及未接受数据科学学位培养的社会人才，需要注重的是技能培训，掌握大数据分析工具，例如Hadoop、Spark、MapReduce、Mahout等，熟悉大数据应用案例。以开展数据科学家训练营或社会技能培训的方式开展。

尽管国内数据科学家的培养已经起步，但值得注意的是，当前数据科学家培养中遭遇的主要问题是：独立培养、缺乏交叉。在技能培训方面，更多的是让受训者

⑬ http://arxiv.org/ftp/arxiv/papers/1501/1501.05039.pdf

掌握数据分析工具，却缺少数据科学家思维。总体而言，数据科学家培养的基础条件缺乏，需要重视数据科学人才培养的基础条件建设，具体如下。

- 计算条件：建设数据科学人才培养所需的计算能力，包括软硬件环境。
- 数据条件：数据是资源，也是数据科学人才培养的核心，需要建设丰富的数据资源环境。
- 师资条件：这是目前相当缺乏的数据科学人才培养资源，也是影响未来数据科学人才培养成果的关键。

4 复旦大学探索实践

上海市数据科学重点实验室（依托复旦大学）在数据科学家培养方面起步早，主要思路是强调数据基础、数据分析能力，注重超学科特色教育。目标是培养具有深度的数据探索能力、扎实的数据挖掘技能以及掌握数据分析工具的数据人才，能够将数据技术、理论和方法与实际应用结合，实现数据驱动决策。

复旦大学数据科学家培养体系建设已初见成效，主要的探索成果如下。

（1）系统化的培养体系

包括青年数据科学家交流计划、数据科学家博士后计划、数据科学家研究生计划、数据科学家本科第二专业计划、软件工程硕士大数据方向培养计划和数据科学家训练营计划、数据科学FIST课程计划，涵盖了数据科学家培养的各个方面，是目前国际上领先的系统化的数据科学家培养计划。

（2）多学科的课程和师资队伍

利用实验室多学科团队优势，组织数据科学家培养课件的编写，内容涵盖数学、计算机、金融、医疗、生物、管理、经济、新闻等多学科领域，围绕数据科学家所需要的数学基础、计算机技能、领域知识和实践经验，设置课程和配置老师，使学生对数据科学的基本原理、方法、技术及应用进行深入的理解。

（3）雄厚的基础设施

建设形成了近200 TB的各类数据资源，主要涵盖：常用的科研实验数据集；世界主要语种语料库；交通、医疗、生物、证券期货、社交网络与舆情、互联网营销、公共设施安全、天文和遥感等应用领域的数据资源；208个CPU核心，4 032 GB内存；1 081 TB的数据存储能力；48个万兆以太网口、144个吉比特以太网口接入能力；30个公网地址。

5 结束语

数据的生产、存储、管理和分析已成为常态工作。大数据催生了数据科学人才的需求，数据科学为各行各业革命性的变革提供数据方法。掌握数据科学的理论基础、数据技术的研发和科学研究的数据方法，有助于科学研究的方法创新和能力提高，有助于将数据技术与应用结合产生经济效益，有助于数据产业的培育和发展。通过分析数据科学人才培养现状，指出数据科学并不是简单的学科交叉，应该基于并和所有学科相关；分析数据科学学科构成，给出数据科学系统知识结构，提出超学科数据人才培养体系，实现以团队培养为主的数据人才培养模式向培养具有数据能力的人（而非团队）为目标的培养模式转变。

参考文献：

[1] ZHU Y Y, ZHONG N, XIONG Y. Data explosion, data nature and dataology[C]// International Conference on Brain

Informatics, October 22-24, 2009, Beijing, China. New York: Springer, 2009: 147-158.

[2] 朱扬勇, 熊赟. 数据学[M]. 上海: 复旦大学出版社, 2009.
ZHU Y Y, XIONG Y. Dataology and data science[M]. Shanghai: Fudan University Press, 2009.

[3] 上海市信息化专家委员会. 专家论城市信息化[M]//朱扬勇, 熊赟. 数据资源保护与开发利用. 上海: 上海科技文献出版社, 2008: 133-137.
Shanghai Informationalization Expert Committee. Expert forum on urban informationalization[M]//ZHU Y Y, XIONG Y. Protection and utilization of data resources. Shanghai: Shanghai Scientific & Technical Publishers, 2008: 133-137.

[4] 朱扬勇. 数据科学与数据产业[J]. 科技促进发展, 2014, 10(1): 72-75.
ZHU Y Y. Data science and data industry[J]. Science & Technology for Development, 2014, 10(1): 72-75.

[5] 朱扬勇, 熊赟. 大数据是数据、技术, 还是应用[J]. 大数据, 2015007.
ZHU Y Y, XIONG Y. Defining big data[J]. Big Data Research, 2015007.

[6] McKinsey Global Institute. Big data: the next frontier for innovation, competition, and productivity[R]. [S.l]: McKinsey Global Institute, 2011.

[7] NAUR P. The science of datalogy[J]. Communications of the ACM, 1966, 9(7): 485.

[8] CLEVELAND W S. Data science: an action plan for expanding the technical areas of the field of statistics[J]. International Statistical Review, 2001, 69(1): 21-26.

[9] ZHU Y Y, XIONG Y. Towards data science[J]. Data Science Journal, 2015, 14(8): 1-7.

[10] CODATA中国全国委员会. 大数据时代的科研活动[M]// 朱扬勇, 熊赟. 数据科学发展与展望. 北京: 科学出版社, 2014: 188-198.
Chinese National Committee for CODATA. Scientific discovery in big data era[M]// ZHU Y Y, XIONG Y. Research progress and prospect for data science. Beijing: Science Press, 2014: 188-198.

[11] BASARAB N. Transdisciplinarity: theory and practice[M]. Cresskill: Hampton Press, 2008.

作者简介

朱扬勇(1963-), 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(1980-), 女, 博士, 复旦大学计算机科学技术学院教授。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, 出版著作3本。目前研究兴趣为数据科学、大数据。

收稿日期: 2016-04-06
基金项目: 上海市科技发展基金资助项目(No.13511504300, No. 14511107302)
Foundation Items: Shanghai Science and Technology Development Fund (No.13511504300, No. 14511107302)