大数据人才培养的基础条件初探

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
人才短缺是发展大数据的主要障碍，越来越多的大学启动了大数据人才培养计划。大数据人才培养的基础条件有哪些？首先要有师资，但这是一个矛盾的基础条件，人才短缺意味着师资更短缺；其次要有数据，且是“大”的数据，没有数据的人才培养是纸上谈兵；有了“大”数据，就需要相应的计算条件。探索了大数据人才培养所需的师资、数据和计算条件问题，提出超学科创新培养模式解决师资条件问题、建立大数据试验场解决数据和计算条件问题。

关键词
大数据；人才培养；数据分析师；数据科学家

中图分类号：TP3 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016059

On prerequisites for cultivating big data talents

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
The shortage of big data talents becomes a global concern, which restricts the development of big data. Cultivating big data talents has been paid attention widely and increasing universities have launched big data talents training plans. It is important and necessary to discuss the prerequisites for cultivating big data talents, including qualified teachers, data resources, computing capabilities. Building qualified teachers team is the first element. It is impossible to discuss cultivating talents if there is no qualified teacher. However, this is a contradiction, because the shortage of big data talents means the shortage of the qualified teachers for big data training. The second one is data resource, especially big data. If there is no data, the big data talents training will not make sense. Correspondingly, the third one is computation capability for big data. Three main prerequisites for big data talents training were discussed including qualified teachers, data resources and computation capabilities.Two solutions were presented: one was to develop an innovation talents training pattern, namely transdisciplinary, for the shortage of qualified teachers, the other was to establish big data arena for innovation and advance to supply the data resource and computation capability.

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

1 引言

从应用的视角来看，大数据是运用来自多个领域的数据解决问题，数据的交叉意味着领域的交叉，领域的交叉意味着知识的交叉，知识的交叉意味着方法的交叉，从而产生新的科学研究方法、新的治理管理方法、新的经济增长方式、新的社会发展方式等。交叉导致了在实施一个大数据应用时，常常由来自于统计、计算机和业务领域的一个数据科学家团队完成[1]。然而，这些专业在大学里是分别设置的，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。事实上，大数据人才短缺是全球性问题[2]，大数据人才被《2015薪资指南（2015 salary guide）》列为薪资涨幅最大的六大行业之一①。面对大数据人才紧缺现状，大学纷纷启动了数据科学或大数据专业培养计划，提升人才培养和输出能力。到2016年，美国有包括哈佛大学、斯坦福大学、芝加哥大学等数10所知名大学开设了数据科学或大数据学位计划，中国有清华大学、复旦大学、南京大学等10余所大学开设了数据科学或大数据学位计划。

尽管大数据人才的培养已经起步，但值得注意的是，当前的数据科学家培养的基础条件缺乏。首先，师资是人才培养的主体，师资结构要与专业适应，然而，大数据人才短缺意味着大数据师资的短缺，因此，这成为一个非常矛盾的基础条件；其次，大数据人才培养的核心是要有数据，而且是“大”的数据，因为人才培养需要得到基础研究和应用研究的训练，具有扎实的基础理论知识和实验技能，没有数据的大数据人才培养就像是纸上谈兵；最后，有了“大”的数据，就需要有相应的计算条件，提供实践锻炼的基础环境。围绕大数据人才培养的师资条件、数据条件和计算条件三大基础问题，本文提出两个解决方案：一是，利用超学科创新培养模式解决师资条件问题；二是，建立大数据试验场解决数据条件和计算条件问题。

① http://www.roberthalf.com/salary-guides

2 大数据人才培养基础条件

2001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）[3]。这是最早的关于数据学科的研究，但长期来没有引起重视，直到大数据热潮到了，大数据人才培养才引起广泛重视。

2.1 大数据人才及其培养

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。数据资源开发利用是实现大数据价值的关键，而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[4]。因此，大数据人才主要包括：能够用数据解决业务领域的问题的人和解决数据自身的问题的人这两大类，可细分为：领域大数据应用人才、大数据工程师、大数据分析师和数据科学家等。

用一个应用例子来说明这些人才在大数据中扮演的角色。

以RTB（real time bidding，实时竞价）精准广告为例。

设有一个网络平台（NP），一批广告商$AS(as_1, as_2, \cdots, as_i, \cdots, as_m)$，一批广告和广大网民$X(x_1, x_2, \cdots, x_j, \cdots, x_n)$，精准广告按照如下步骤运行：

（1）当一个网民x_j打开网络平台网页；

（2）网络平台就会向所有的广告商发布一条信息，说有一个网民x_j现在正打开网页，谁要在RTB窗口出广告；

（3）广告商分析网民x_j的个人行为信息，在他的广告库中搜寻匹配广告，如果有，就找出一个匹配度最高的广告，然后向网络平台发出竞价报价；

（4）网络平台开始启动各广告商的报价竞争，在规定的时间内，以价高者得的原则，将广告位卖给某个广告商as_i；

（5）获得广告位的广告商as_i从广告库中将最高匹配的广告推送到网络平台网页的RTB窗口中；

（6）一次RTB广告结束，整个过程耗时不超过100 ms。

精准广告是大数据应用最成功的领域，从中可以看到大数据人才从事的具体工作：领域大数据应用人才要给出精准广告的业务模型，根据可能得到的数据，设计业务逻辑；大数据工程师要进行软件开发、工程实施、优化系统，确保100 ms内完成所有工作；数据分析师要运用各种数据分析工具对广告和网民进行聚类、分类等数据分析工作；数据科学家则需要根据RTB精准广告业务和潜在的网民、广告内容等设计相似性函数、数据分析算法、建立分析模型等。

从上述分析，可以大致看出如下几类大数据人才的情况。

（1）领域大数据应用人才

他们是各领域中的数据人才，之前，他们中的大部分从事本单位的信息化工作，现在开始从事本单位的数据资源开发工作。其中，主要代表是一些之前的首席信息官（chief information officer，CIO）正试图转变为首席数据官（chief data officer，CDO）。调查机构IDC预测，到2020年，全球组织中将有60%的首席信息官被首席数据官取代②。

（2）大数据工程师

他们是掌握Hadoop、MapReduce、Spark、HBase等大数据开发环境和工具的工程师，善于在数据规模和系统配置、软件优化方面进行调优，使得大数据系统得以在用户希望的时间内完成相应的工作。

（3）大数据分析师

他们掌握了MATLAB、R、Python语言之类的大数据分析工具，具备良好的数理统计知识背景，通常是统计学家，能理解业务需求并应用工具开展数据分析的人。

（4）数据科学家

他们掌握数据分析算法原理、善于发挥个体能力和经验，创造性地设计数据分析算法，尤其是设计相似性函数之类的创造性劳动。他们是发现数据规律和现象、探寻数据本质的科学家。

针对不同的大数据人才，可以设计针对性的培养方案，即大数据人才的培养是多类型的。复旦大学上海市数据科学重点实验室就建立了系统化的大数据培养体系，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，是目前国际上最为系统化的数据科学家培养计划[1]。

从大的范围看，大数据人才培养就是学位培养和应用培训两类。学位培养需要设置完整的培养体系，包括：培养方案、课程体系、师资力量、实验条件等；应用培训相对比较简单，主要注重的是技能培训，掌握大数据分析工具，例如Hadoop、MapReduce、Spark、Mahout等，熟悉大数据应用案例等。

② http://www.forbes.com/sites/gilpress/2014/10/30/idc-to-cios-60-percent-of-you-will-be-supplanted-by-chief-digital-officers-by-2020/#676fc280313c

2.2 师资条件

师资条件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。大数据师资建设需要优化知识结构、教材和教师队伍，培养在大数据领域具有影响力的学术带头人，形成大数据学术创新团队。

从知识结构看，大数据人才的知识体系结构主要由科学的基础理论和方法、大数据计算技术、领域业务知识3方面构成[1]。大数据人才应该是具备多种能力的跨界人才，数据科学人才培养体系应该是多层次多类型的。

目前，关于大数据、数据科学方面的书籍大多是零散的大数据技术的介绍，系统化地适用于大数据、数据科学人才培养方面的教材尚未出现，这是大数据师资队伍建设的源头，需要尽快组织相关教材的编撰；此外，大数据师资队伍的建设，不能在现有的单个专业或学院中拥有大部分课程和教师，需要根据数据科学的知识结构进行合理配置，设置大数据专业课程。

2.3 数据条件

大数据人才是解决大数据问题的，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。为了训练大数据人才，就需要各种各样的数据环境，在实践中总结经验，训练发现问题和解决问题的能力。数据环境是要有来源多样、类型多样的数据集合，并且数据规模要足够大。

首先，数据来源多样、类型多样造成了数据复杂性。一是，数据来源于不同的数据采集设备或由专用数字设备产生，例如传感器、医疗设备、GIS、多媒体等，这产生了多种数据类型；二是，数据由不同的数据库及其管理系统存储和管理，例如Oracle、HBase、MongoDB等，这形成多种数据结构；三是，业务数据分析需要来自多个相关领域的数据辅助，例如精准医疗中除了来自医院的电子病历数据，还需要生物组学数据，甚至需要有环境、社交等数据。为实现不同领域的数据的融合，需要分析数据在格式、类型、来源等方面的复杂性。异质数据网络[5]是大数据环境下的一种主要数据组织方式[6]，是一种复杂数据类型。异质数据网络具有多种类型对象（节点）和多种类型连接（边）的数据网络，网络中的不同路径代表了对象间的不同关系，具有不同的语义信息。

其次，数据规模足够大，意味着超出了当前技术能力。随着数据规模的增大，数据处理的能力也在不断地发展，当前已经产生了大量满足大规模数据分析能力的挖掘算法和计算技术，例如K-means++[7]、K-means Ⅱ[8]等聚类算法对经典K-means算法进行了改进，实现了大规模数据的高效聚类；又如特异群组挖掘算法[9]的提出，实现了不同于簇或孤立点的特异群组这样一类高价值低密度的大数据分析；同时，一系列大数据计算框架也发展迅速，包括Hadoop、HDFS、MapReduce、NoSQL、Hive、Storm、Spark等，这些框架中的功能也存在差异。

大数据人才培养需要有足够多的数据作为基础条件。如果数据量、数据种类有限，目前已有的信息技术能够很好地进行处理，那么研究的技术、应用是否真的适用于大数据，是否真的是大数据将无法保证；没有数量足够多、种类足够多的数据作为研发的支撑，很难真正开展大数据技

术研究与应用研发。此外，需要足够多的数据也意味着需要有能够存储管理大量、多种类数据的设备和能力。

那么，到底多大规模的数据才是足够的数据呢？就目前技术水平，引发技术挑战的大数据集，其规模应该要有PB级别。PB级别的数据计算、数据分析、数据展现等方面有很多技术问题。虽然，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，数据的复杂度相对较高。

2.4 计算条件

面对以上的数据条件，需要相应的计算条件，需要有能够分析处理这些数据的软硬件环境。有了足够多的数据之后，若要分析挖掘这些数据，就需要具有足够计算能力的计算环境。以深度学习为例，Hinton G E于2006年在《Science》上发表的论文[10]提出数据降维方法deep autoencoder，这成为深度学习开创性标志算法之一。然而，其却并没有成为广泛关注和使用的方法，而是随着数年后计算条件和计算能力的提升，在大数据的热潮下，深度学习方法开始发挥更为重要的应用价值。

传统的独立服务器（或小规模服务器集群）是无法直接处理大数据的。然而，建立一套可用的大数据分析处理环境需要投入大量的硬件设备和构建复杂的软件环境，这使得开展大数据研发需要有足够的资金投入。

3 超学科人才培养模式

由于大数据的知识结构还没有统一认识、学科体系还没有建立，目前还没有单个学院或专业具备培养大数据人才的能力，多学科的课程和师资队伍共同培养大数据人才是一种可行的培养模式，称为超学科人才培养模式。其内涵是：在大数据学科还不成熟的情况下，不将大数据作为单个学科来看待。事实上，大数据的广泛交叉性（不是两个、三个之类的简单交叉）决定了其人才培养的广泛交叉性。在人才培养方面将打破原有的学科限制，大数据人才所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，甚至超越并取代它们，是一种新的视角和一种新的学习体验，即超学科[11]。

在超学科概念下，可以组织各学科（包括数学、计算机、金融、医疗、生物、管理、经济、新闻等多学科领域）的科学家，围绕大数据人才所需要的数学基础、计算机技能、分析基础、领域知识和实践经验，设置课程、编写教材、安排实验，使学生对数据科学的基本原理、方法、技术及领域应用具有深入的理解。

目前，数据科学研究机构人员组成一般来自多个学科交叉领域，下面以中国复旦大学、美国哥伦比亚大学、美国纽约大学为例。

（1）复旦大学

复旦大学上海市数据科学重点实验室③的师资力量包括复旦大学各学院教师形成的固定人员团队以及复旦大学外部的国外高校和企业形成的流动人员团队，其专业方向分别来自计算机、数学、生命科学、管理、经济等多学科，见**表1**。

③ http://www.datascience.cn/

（2）哥伦比亚大学

哥伦比亚大学数据科学研究院（Data Science Institute，Columbia University）划分为多个分研究中心，分别研究数据科学基础、智慧城市、新媒体等，每个中心的研究人员均来自多学科领域，其人员④结构情况见**表2**。

④ http://datascience.columbia.edu/people/all

（3）纽约大学

纽约大学数据科学研究中心（NYU

表1 复旦大学上海市数据科学重点实验室人员学科结构

人员类别	人数/人	学院或专业
校内固定人员	56	计算机、数学、社会学、管理学、经济学、生命科学、新闻学
校外流动人员	60	计算机、数学、社会学、管理学、经济学、生命科学、金融学、天文学、法学、信息工程、物理学

表2 哥伦比亚大学数据科学研究院人员学科结构

研究院分中心	学院或专业
数据科学基础	统计学、计算机、工业工程、化学工程、电子工程、商业、应用物理、应用数学、地球科学、计算学习系统、政治学、环境工程、能源、决策风险、生物学、生物医疗工程、放射学、生物医学信息学、机械工程
智慧城市	土木工程、工程力学、电子工程、计算机、国际公共事务、地球和环境工程、建筑规划、计算学习系统、应用数学、应用物理、地球和气象科学、机械工程、化学工程
新媒体	新闻学、计算学习系统、社会学、计算机、历史学、电子工程、建筑规划、教育、商业、决策风险、经济、英语、文学
健康分析	生物医学信息学、计算生物学、生物信息学、流行病学、生物工程、计算机、电子工程、应用物理、应用数学、生物学、商业、决策风险、麻醉学、神经学、国际公共关系
金融分析	工业工程和操作、统计学、商业、决策风险、金融、经济、应用数学、计算机、化学工程
Cyber安全	计算机、法学

Center for Data Science, New York University)的人员[5]结构情况见表3。

[5] http://cds.nyu.edu/people/

4 大数据试验场

大数据试验场是邬江兴和朱扬勇于2014年提出的概念，目前已经写入上海市大数据相关规划，上海市正在推进建设大数据试验场。众所周知，大数据最先是作为技术问题或技术挑战提出来的。就是说，现阶段还没有适合大数据分析的计算机及集群、计算框架和软件系统，但大数据应用需求迫切，因此，边使用、边探索是好的方式。这包含两个方面：一个方面用现有的技术解决各类数据应用问题、建立应用模型（如精准广告、精准医疗等）；另一方面，对于现有技术不能解决的问题，探索新型技术。把拥有大规模数据及其相应的计算分析能力的试验环境称为大数据试验场。

开展大数据人才培养，需要做大量的大数据试验，需要一个大数据试验场，以解决大数据人才培养的数据条件和计算条件。数据条件和计算条件是相辅相成的，良好的数据条件需要良好的计算条件支撑，良好的计算条件需要良好的数据条件来实践。针对当前大数据状况，1 PB的数据规模应该是开展大数据研究、训练的基础要求。然而，在1 PB规模的数据上做大数据分析，则需要5 PB以上的存储空间以及相应的计算能力，需要5 000万元左右的投资。显然，这样的投资规模，对于一般的大学都是难以承受的，因此，需要建设公共的大数据人才培养大数据试验场。

一个用于大数据人才培养的大数据试

表3 纽约大学数据科学研究中心人员学科结构

人数	学院或专业
106	计算机、神经科学、电子工程、生物信息学、病理学、政治学、数学、物理学、商业分析、应用统计学、管理学、心理学、社会学、经济学、艺术、生物学、统计学、医学、药学、气象学、海洋学等

验场，其数据条件和计算条件如下。

（1）数据条件

首先，要求大数据试验场要能够存储1 PB的待处理数据，可以采用两种形式：一种是单体数据规模达到1 PB，用于探索、训练和试验大规模数据的移动、管理、分析等方面的快速方法；另一种是多类型可关联的多学科数据，总规模是1 PB，用于探索、训练和试验复杂数据的关联和分析方法。同时要配置相应的存储设备。考虑到主流的大数据平台（如Spark或基于Hadoop的各发行版本等）的数据自动备份、多副本并行处理等因素，因此至少需要3倍的数据存储空间，即实际用于存储数据的容量大于3 PB。另外，还需要2 PB的存储空间用于数据副本或虚拟化工作以及数据分析工作。因此，1 PB数据规模的大数据试验场至少要达到5 PB的物理存储能力。

（2）计算条件

从低成本出发，采用单台主流的PC服务器（8个CPU内核）单次任务处理4 TB数据，1/3的数据需要同时处理估算，需要近100台PC服务器，相当于采用虚拟化技术后达到每内核处理约0.5 TB以上数据的并行处理规模。再加上作为集群管理、任务调度等专门用途的服务器，共需要约130台服务器。另需要一批网络设备。由于大数据处理对服务器间的网络通信压力巨大，需要能够快速传输GB级甚至TB级的数据，因此，整个服务器间的网络至少应达到10 Gbit/s（按80%线速传输计算，约为每秒传输1 GB数据），试验场内网的骨干交换机之间应达到至少40 Gbit/s的数据交换能力。

5 结束语

虽然大数据是新生事物，大数据人才的知识结构、培养计划还需要较长时间的探索，当前还没有一个获得广泛认可的大数据或数据科学学科计划，但是，各种人才培养方式都需要师资、数据和计算这3个基础条件。本文通过分析大数据人才培养现状，指出大数据并不是简单的学科交叉，而是和所有学科相关，提出用超学科人才培养方法解决大数据师资短缺问题；提出建设公共的大数据人才培养试验场来解决数据条件和计算条件。建议政府出资建设大数据人才培养大数据试验场，支持跨校、跨学科的大数据综合人才培养，支持大数据市场培训机构。

参考文献：

[1] 朱扬勇，熊赟. 大数据时代的数据科学家培养[J]. 大数据，2016，2(3)：106-112.
ZHU Y Y，XIONG Y. Training data scientists in the era of big data[J]. Big Data Research，2016，2(3)：106-112.

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

[3] 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.

[4] 朱扬勇，熊赟. 大数据是数据、技术，还是应用[J]. 大数据，2015，1(1)：71-81.
ZHU Y Y，XIONG Y. Defining big data[J]. Big Data Research，2015，1(1)：71-81.

[5] SUN Y，HAN J. Mining heterogeneous information networks：principles and methodologies[J]. ACM Sigkdd Explorations Newsletter，2010，14(2)：439-473.

[6] 熊赟，朱扬勇，陈志渊. 大数据挖掘[M]. 上海：上海科学技术出版社，2016.
XIONG Y，ZHU Y Y，CHEN Z Y. Big data mining[M]. Shanghai：Shanghai Scientific and Technological Literature Press，2016.

[7] BAHMANI B，MOSELEY B，VATTANI A，et al. Scalable k-means++[J]. Proceedings

of the VLDB Endowment, 2012, 5(7): 622-633.
[8] ARTHUR D, VASSILVITSKII S. K-means++: the advantages of careful seeding[C]//Eighteenth ACM-SIAM Symposium on Discrete Algorithms, January 7-9, 2007, New Orleans, USA. [S.l.:s.n.], 2007: 1027-1035.
[9] 熊赟, 朱扬勇. 特异群组挖掘: 框架与应用[J]. 大数据, 2015020.
XIONG Y, ZHU Y Y. Abnormal group mining: framework and applications[J]. Big Data Research, 2015020.
[10] HINTON G E, SALAKHUDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[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"。担任第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。目前主要研究方向为数据科学、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。目前主要研究方向为数据科学、大数据。

收稿日期: 2016-08-10
基金项目: 上海市科技发展基金资助项目(No. 16JC1400801)
Foundation Item: Shanghai Science and Technology Development Fund (No. 16JC1400801)