数据科学与大数据技术专业特色课程研究

朝乐门[1,2] 邢春晓[3,4,5] 王雨晴[2]

(数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872)[1]

(中国人民大学信息资源管理学院 北京 100872)[2] (清华大学计算机科学与技术系 北京 100084)[3]

(清华大学信息技术研究院 北京 100084)[4] (清华信息科学与技术国家实验室(筹) 北京 100084)[5]

摘 要 目前,我国数据科学与大数据技术专业的建设已成为新的热点话题。在系统调研世界一流大学数据科学专业建设现状的基础上,从特色课程的视角重点分析加州大学伯克利分校、约翰·霍普金斯大学、华盛顿大学、纽约大学、斯坦福大学、卡内基梅隆大学、哥伦比亚大学、伦敦城市大学共8所大学的数据科学专业,提出了数据科学与大数据技术这一新专业应重视的10门特色课程,并分析了现阶段我国数据科学教育中普遍存在的8种曲解现象及对策建议。

关键词 大数据,数据科学,课程体系,专业建设

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.03.001

Unique Curriculums for Data Science and Big Data Technology

CHAO Le-men[1,2] XING Chun-xiao[3,4,5] WANG Yu-qing[2]

(Key Laboratory of Data Engineering and Knowledge Engineering(Renmin University of China),Beijing 100872,China)[1]

(School of Information Resource Management,Renmin University of China,Beijing 100872,China)[2]

(Department of Computer Science and Technology,Tsinghua University,Beijing 100084,China)[3]

(Reasearch Institute of Information Technology,Tsinghua University,Beijing 100084,China)[4]

(Tsinghua National Laboratory for Information Science and Technology (TNList),Beijing 100084,China)[5]

Abstract How to construct a novel major,called Data Science and Big Data Technology,is one of the hot topics in China. An in-depth analysis on typical universities was conducted including University of California,Berkeley,Johns Hopkins,Washington University,New York University,Stanford University,Carnegie Mellon University,Columbia University,and City University of London. And then,ten core courses of Data Science and Big Data Technology were identified and described. Finally,eight common misunderstandings in existing data science curriculum were proposed,and the solutions were provided respectively.

Keywords Big data,Data science,Curriculum system,Professional construction

2016年,教育部发布的《2015年度普通高等学校本科专业备案和审批结果》中首次增设数据科学与大数据技术专业,并批准了北京大学、对外经济贸易大学及中南大学的新增专业申请;2017年,中国人民大学等32所高校的新增专业申请获第二批次批准 。全国高校大数据教育联盟的统计数据显示,2017年申请该专业的院校多达263所,其中工学190所,理学73所[1]。从申请资料看,国内数据科学专业是一门主要以统计学和计算机科学与技术专业为基础建设的全新专业。数据科学专业已成为我国现阶段高等教育的热点问题之一。但是,建设什么样的专业以及如何建设该专业仍为各高校面临的难点问题。

在国外,数据科学(Data Science)专业是以数据分析学(Data Analytics)专业为基础发展而来的,可追溯至2007年北卡罗来纳州立大学(North Carolina State University)率先设立的数据分析硕士学位(Master of Science in Analytics)[2]。与统计学和计算机科学与技术等基础学科不同,数据分析学进一步抽象了这些底层科学中的数据问题,填补了包括统计学和计算机科学在内的基础学科与数据科学之间的空白,为数据科学这一新学科的出现奠定了直接基础。“数据分析学”向“数据科学”的实质性过渡出现在2013年左右,比较有代表性的是纽约大学于2013年新开设的数据科学硕士专业(Master of Science in Data Science)[3]。之后,包括加州大学伯克利分校、约翰·霍普金斯大学、华盛顿大学在内的多个学校设立了数据科学专业。可见,国外一流大学的数据科学专业建设至少早于国内3年。

为此,本文调查分析了世界一流大学中数据科学专业的培养方案,重点分析了数据科学专业中开设的特色课程,并探讨了其对我国数据科学专业建设的借鉴意义。

到稿日期:2017-12-05 返修日期:2018-01-03 本文受国家自然科学基金项目(91646202,71103020)资助。

朝乐门(1979—),男,副教授,博士生导师,主要研究方向为数据科学与大数据分析,E-mail:chaolemen@ruc.edu.cn(通信作者);**邢春晓**(1967—),男,教授,博士生导师,主要研究方向为云计算与大数据分析;**王雨晴**(1994—),女,硕士生,主要研究方向为数据科学与大数据分析。

1 数据调研及分析

通过 Study Portal 进行调查发现，截至 2017 年 11 月，国外数据科学专业的本科、硕士、博士学位项目分别已达到 5601，4179 和 301 项，主要分布在美国、英国、澳大利亚、加拿大、德国和意大利等国家。但是，从课程体系和人才培养定位看，能够体现国外数据科学专业教育的本质与特色的是硕士层次的教育，比较典型的学校有加州大学伯克利分校、约翰·霍普金斯大学、华盛顿大学、纽约大学、斯坦福大学、卡内基梅隆大学、哥伦比亚大学、伦敦城市大学，如表 1 所列。

表 1 典型的数据科学专业及其特色课程

Table 1 Typical Data Science Programs and their core courses

学校	学位名称	特色课程
加州大学伯克利分校[4]	信息与数据科学硕士（Master of Information and Data Science）	面向数据科学的 Python 语言（Python for Data Science） 研究设计及数据与分析中的应用（Research Design and Application for Data and Analysis） 数据存储与检索（Storing and Retrieving Data） 应用机器学习（Applied Machine Learning） 试验与因果分析（Experiments and Causality） 大数据中的人与价值（Behind the Data：Humans and Values） （纵向扩展及真正的）大数据（Scaling Up！ Really Big Data） 数据可视化与沟通（Data Visualization and Communication） （数据科学）综合训练课程（Synthetic Capstone Course）
约翰·霍普金斯大学[5]	数据科学理学硕士（Master of Science in Data Science）	数据科学（Data Science） 数据可视化（Data Visualization） 随机优化与控制（Stochastic Optimization and Control） 数据科学家的工具箱（Data Scientist's Toolbox） 数据采集与清洗（Getting and Cleaning Data） 探索性数据分析（Exploratory Data Analysis） 可重复研究（Reproducible Research） 实用机器学习（Practical Machine Learning） 数据产品开发（Developing Data Products） 数据科学综合训练课程（Data Science Capstone）
华盛顿大学[6]	数据科学理学硕士（Master of Science in Data Science）	数据可视化与探索性分析（Data Visualization & Exploratory Analytics） 应用统计与试验设计（Applied Statistics & Experimental Design） 面向数据科学的数据管理（Data Management for Data Science） 数据科学家常用的统计机器学习（Statistical Machine Learning for Data Scientists） 面向数据科学的软件设计（Software Design for Data Science） 可扩展的数据系统与算法（Scalable Data Systems & Algorithms） 以人为中心的数据科学（Human-Centered Data Science） 数据科学综合训练课程（Data Science Capstone Project）
纽约大学[7]	数据科学理学硕士（Master of Science in Data Science）	数据科学导论（Introduction to Data Science） 大数据（Big Data） 面向数据科学的统计学与概率论（Probability and Statistics for Data Science） 推理与表示（Inference and Representation） 机器学习与计算统计学（Machine Learning and Computational Statistics） 数据科学综合训练课程（Capstone Project in Data Science） 基于优化的数据分析（Optimization-based Data Analysis） 非光滑凸优化（Convex and Nonsmooth Optimization）
斯坦福大学[8]	统计学：数据科学理学硕士学位（M. S. in Statistics：Data Science）	现代应用统计学：学习（Modern Applied Statistics：Learning） 现代应用统计学：数据挖掘（Modern Applied Statistics：Data Mining） 数据驱动型医学（Data Driven Medicine） 现代统计学与现代生物学（Modern Statistics for Modern Biology） 基于大数据的商务智能（Business Intelligence from Big Data） 基于数据的计算范式（Paradigms for Computing with Data）
哥伦比亚大学（纽约）[10]	数据科学理学硕士（Master of Science in Data Science）	数据科学导论（Introduction to Data Science） 面向数据科学的计算机系统（Computer Systems for Data Science） 探索性数据分析与可视化（Exploratory Data Analysis & Visualization） 数据科学中的因果推理（Causal Inference for Data Science） 大数据分析学（Big Data Analytics） 数据科学 Capstone 与道德（Data Science Capstone & Ethics）
伦敦城市大学[11]	数据科学理学硕士（MSc in Data Science）	数据科学原理（Principles of Data Science） 大数据（Big Data） 可视分析学（Visual Analytics） 数据可视化（Data Visualization） 神经计算（Neural Computing） 研究方法与专业问题（Research Methods and Professional Issues） 高级并发编程（Advanced Programming：Concurrency）
卡内基梅隆大学[9]	计算数据科学硕士学位（Master of Computational Data Science）	云计算（Cloud Computing） 高级云计算（Advanced Cloud Computing） 多媒体数据库及数据挖掘（Multimedia Databases and Data Mining） 移动与普适计算（Mobile and Pervasive Computing） 大数据集的机器学习（Machine Learning with Big Data Sets） 智能信息系统的设计与开发（Design and Engineering of Intelligent Info Systems） 大数据分析学（Big Data Analytics）

1）加州大学伯克利分校

该学校的数据科学专业由信息学院（School of Information）开设，专业名称为信息与数据科学，授予的学位为信息与数据科学专业硕士[12]。该专业主要侧重于培养学生的研究设计、数据清洗、存储与检索、挖掘与探索、数据可视化、道德与隐私、数据分析、沟通与呈现的能力，如图 1 所示。

图 1　加州大学伯克利分校 MIDS 专业所关注的学生能力[12]

Fig. 1　Key skill areas of MIDS at UC Berkeley[12]

为了达到上述人才培养的目的，该专业开设基础课程、高级课程和综合训练课程这 3 类课程。其中，基础课程共有 5 门：面向数据科学的 Python 语言，面向数据与分析的研究设计，面向数据科学的统计学，数据存储与检索以及应用机器学习。高级课程有 7 门：试验与因果分析，数据、人与价值，（纵向扩展及真正的）大数据，面向离散响应、时间序列和面板数据的统计方法，可扩展的机器学习，基于深度学习的自然语言处理以及数据可视化与沟通。除了基础课程和高级课程，该学校还开设了一门综合训练课程，以培养学生综合运用所学专业知识及解决现实问题的能力。

总体上看，人才培养旨在培养数据科学领域的领导者，侧重培养学生运用新工具和新方法，从现实数据中获得洞见（Insights）以及如何有效地沟通与阐释自己的研究发现，进而改变他人行动和思想的能力。该学校的数据科学专业的人才培养具有如下几个特点：

①强调数据科学的多学科交叉特点，将社会科学、计算机科学、统计学、管理学和法学等多学科知识融入具体课程中。

②凸显数据科学本身的讲解，注重提升学生的基于数据提出好问题的能力以及面向数据科学的研究设计、数据清理、存储与检索、交流与沟通、统计分析、道德与隐私、数据可视化以及数据挖掘与探索等关键技能。

③引入基于项目的学习方法，借鉴本校信息学院其他专业的培养经验，通过基于项目的教学方式，鼓励学生综合运用多种不同的工具和方法来解决复杂问题。

④强调学生动手实践能力的培养，为其提供亚马逊 Web 服务和 IBM 大数据平台等实践平台。

2）约翰·霍普金斯大学

该学校的怀廷工程学院（Whiting School of Engineering）开设了名为数据科学的新专业，授予的学位为数据科学理学硕士。

该专业的课程体系包含先修课程（Prerequisite Courses）、基础课程（Foundation Course）、必修课程（Required Courses）、选修课程（Electives）以及独立学习（Independent Study）课程等近 60 门课程[13]。基础课程有两门，即算法基础（Foundations of Algorithms）和统计方法与数据分析（Statistical Methods and Data Analysis）。必修课程包括数据库系统原理、数据科学、数据可视化、优化导论（Introduction to Optimization）、统计模型与回归、计算统计学。选修课程分为机器学习和统计学两个大方向，共有 14 门主要课程，均为较常见的课程。值得一提的是，该专业另提供了近 30 门扩展选修课程（Additional Selections）供学生置换同一个领域的必修/选修课程，这些扩展课程均为统计学和计算机科学与技术专业的常见课程。独立学习课程主要包括独立动手实战（Capstone 项目）和独立学术研究。

总体上看，该学校的数据科学专业的人才培养具有如下几个特点：

①从人才培养的目的看，该专业旨在培养“有竞争力”的数据科学家，要求学生具备 3 方面的能力：综合运用计算机科学和应用数学的知识来分析与处理大规模数据集的能力；从复杂数据中快速洞察有价值的信息和从信息中发现相关关系的能力；基于规范的技术和抽象的方法以及面向现实世界中的具体问题的建模能力[14]。

②强调学生对数据科学的理论基础的掌握程度，突出了 3 个主要领域：计算机科学和技术、统计学和应用数学。其中，对应用数学的重视是该学校数据科学专业的一大特色。

③从课程设计及内容选择看，该专业鼓励在每一门课程中引入来自现实世界的具体问题作为例题和主要关注点。例如，独立学术研究课程中强调对具体行业中实际问题的处理能力。

④强调培养学生的数据全生命期管理、统计分析和故事化描述能力。

3）华盛顿大学

该学校整合自己的应用数学系、生物统计学系、Paul G. Allen 计算机科学与工程学院、以人为本的设计与工程系、统计系、信息学院 6 大院系以及电子科学研究所的资源，开设了一个面向在职人员的夜大类数据科学专业项目，所授予的学位为数据科学理学硕士。该专业的课程设计得较为简洁，包括 8 门核心课程以及 1 个 Capstone 项目。其中，8 门核心课程分别是统计与概率论（Introduction to Statistics & Probability）、信息可视化（Information Visualization）、应用统计与试验设计、面向数据科学的数据管理、数据科学家常用的统计机器学习、面向数据科学的软件设计、可扩展的数据系统与算法和以人为中心的数据科学。Capstone 项目要求学生自己组队，并自主完成项目的选题、研究设计和研究过程等工作，侧重于培养学生对大规模数据集的处理能力、从数据中获得洞见的能力以及与其他人分享自己所发现的洞见（Insights）的能力[15]。

从整体上看，该学校的数据科学专业主要定位于应用型人才的培养，尤其注重培养数据分析师和应用型数据科学家。人才培养的主要特点如下：

①面向在职人员。该专业主要针对刚毕业的学生或在职人员开设，一般在业余时间上课，允许学生脱产或在职学习。

②重视团队协作能力的培养。多数课程的作业均需要以团队的方式完成。

③强调动手操作能力,加强学生运用 Python 和 R 进行数据分析的能力,部分作业还需要进行 Java 编程。

④突出以人为中心的数据科学与可视化(Human-centered Data Science and Visualization)能力,开设有专门的以人为中心的数据科学课程。

4)纽约大学[16]

该学校的数据科学专业由数据科学中心(Center for Data Science)开设,授予的学位为数据科学理学硕士。主要必选课程有数据科学导论、面向数据科学的统计学与概率论、机器学习、大数据以及 Capstone 项目。此外,该专业还要求学生从以下 6 门课程中任选一门作为选修课:推理与表示、深度学习、基于表示学习的自然语言处理、自然语言理解与计算语义、基于优化的数据分析、优化与计算线性代数。值得一提的是,该学校的数据科学专业设有多个培养方向(Track)。

①大数据方向(Big Data Track),开设有自然语言理解与计算语义、信息可视化、大规模可视化分析、数据库导论、高级数据库系统等课程。

②数学与数据方向(Mathematics and Data Track),开设有基于优化的数据分析、推理与表示、数据科学中数学:图与网络(Mathematics of Data Science:Graphs and Networks)以及非光滑凸优化等课程。

③自然语言处理方向(Natural Language Processing Track),开设有基于表示学习的自然语言理解、自然语言理解与计算语义、统计自然语言理解、推理与表示、深度学习、文本数据(Text as Data)、自然语言处理以及高级语言学等课程。

④物理学方向(Physics Track),开设的主要课程有推理与表示、实验物理研究(Experimental Physics Research)、理论物理研究(Theoretical Physics Research)、研究式阅读(Research Reading)、计算物理(Computational Physics)、统计物理、生物物理(Biophysics)、专题研讨课、天体物理学专题(Special Topics in Astrophysics)以及相变与临界现象(Phase Transitions and Critical Phenomena)。

⑤生物学方向(Biology Track),重点讲解基础生物学、健康与疾病等基础知识,并要求选修生物学的课程。

纽约大学将数据科学专业的人才培养定位在“下一代数据科学家”,为具备数学、计算机科学和应用统计学基础的学生提供了多个可选的培养方案,其主要特点如下:

①设有多个培养方向,如大数据、数学与数据、自然语言处理、物理学和生物学等,其人才培养特别强调数据科学与其他专业的深度融合。

②重视对优化论的学习,在课程体系中设置了多个与优化论相关的课程,如基于优化的数据分析、优化与计算线性代数、非光滑凸优化。

③强调实践操作能力,重视对现实世界中具体问题的处理能力。

5)斯坦福大学

该学校的数据科学专业由统计系(Department of Statistics)和计算与数学学院(Institute for Computational and Mathematical Engineering)联合开设,授予的学位为数据科学方向的统计学理学硕士。该学校共开设 29 门课程[17],分为以下 5 个模块。

①基础课程模块,开设有数值线性代数(Numerical Linear Algebra)、离散数学与算法、优化论、工程中的随机方法(Stochastic Methods in Engineering)以及随机算法与概率分析(Randomized Algorithms and Probabilistic Analysis)等课程。

②数据科学模块,开设有统计推理导论、回归模型及方差分析导论、统计模型导论、现代应用统计学:学习以及现代应用统计学:数据挖掘等课程。

③高级科学编程及高性能计算(Advanced Scientific Programming and High Performance Computing)模块,涉及的课程有高级科学编程(Advanced Scientific Programming)、并行计算导论、分布式算法与优化论、数值分析的并行方法、并行计算、并行计算机的架构及编程以及高级多核系统。

④专业选修(Specialized Electives)模块,开设的课程有计算分子生物学中的表示与算法(Representations and Algorithms for Computational Molecular Biology)、数据驱动型医学、面向现代生物学的统计学、社会与信息网络分析、机器学习、面向视觉认知的卷积神经网络(Convolutional Neural Networks for Visual Recognition)、海量数据集的挖掘、计算机图形学、地理统计学(Geostatistics)、大数据商务智能、人类神经影像学方法(Human Neuroimaging Methods)和基于数据的计算范式。

⑤实战(Practical Component)模块,包括 Capstone 项目和独立学习项目。

斯坦福大学的数据科学专业主要侧重于培养统计学家,而并非数据科学家。其最突出的特点是将数据科学作为统计学的一个方向,从而培养出面向数据科学的统计学家。因此,与其他学校的数据科学专业不同,该大学的数据科学专业强调的是数据科学与统计学的深度融合。

6)哥伦比亚大学(纽约)

该学校的数据科学专业由数据科学学院(Data Science Institute)开设,授予的学位为数据科学理学硕士。课程体系可分为导论类课程、计算机科学、统计学、选修课程和 Capstone 课程这 5 大类。

①导论类课程被定位为计算机科学和统计学的交叉课程,课程名称为数据科学原理。

②计算机科学类课程包括面向数据科学的计算机系统、数据科学中的机器学习、数据科学中的算法。

③统计学类课程包括概率论、面向数据科学的概率统计(Probability & Statistics for Data Science)、探索性数据分析及可视化、统计推理与建模。

④选修课程为跨专业课程,旨在鼓励学生跨专业选修哥伦比亚大学的其他专业的课程。比较受欢迎的选修课程包括翻译生物信息学(Translational Bioinformatics)、应用机器学习、数据科学中的因果推理、数据科学的要素、面向数据科学的机器学习产品、社会意义的计算模型(Computational Models of Social Meaning)、数据科学项目、大数据分析学、面向计算机可视化、语音和语言的深度学习(Deep Learning for Computer Vision, Speech, and Language)、金融大数据(Big Data in Finance)和可持续技术与智慧城市的演化(Sustainability Technology and the Evolution of Smart Cities)。

⑤Capstone 项目的名称为数据科学 Capstone 与道德，该项目综合运用所学知识来解决产业、政府和非盈利部门的实际数据和具体问题[18]。

该专业的人才培养目标定位于数据科学家。主要特点有两个：

①专业教育与专业认证相结合。该学校不仅开设有数据科学硕士专业，而且还提供一项专业认证——数据科学专业成就认证（The Certification of Professional Achievement in Data Sciences），融合二者的课程设置。

②专业教育与在线免费课程相结合。作为线下专业课程的重要补充，在线开放课程——数据科学与分析 X 系列课程（Data Science and Analytics XSeries）同时开设，该课程主要介绍数据科学的最新工具及其在金融、健康医疗、产品开发、市场营销等领域中的应用。目前，已开设的在线课程有：数据科学与分析学中的统计思维（Statistical Thinking for Data Science and Analytics）、数据科学与分析学中的机器学习（Machine Learning for Data Science and Analytics）、数据科学与分析学中的驱动技术：物联网（Enabling Technologies for Data Science and Analytics：The Internet of Things）。

7）伦敦城市大学

该学校的数据科学专业由数学、计算机与工程学院（School of Mathematics，Computer Science & Engineering）和计算机系（Department of Computer Science）联合开设，授予的学位为数据科学理学硕士。其课程体系由核心模块、选修模块和综合训练课程 3 部分组成，每个模块包括实验教程和课程作业。其中，核心模块包括数据科学原理、机器学习、大数据、神经计算、可视分析学、研究方法与专业问题等课程；选修课程包括高级并发编程、高级数据库、信息检索、数据可视化、数字信号处理及音频编程（Digital Signal Processing and Audio Programming）、云计算、计算机视觉、软件代理（Software Agents）等。综合训练课程与其他学校不同，不以小组形式完成，而是要求学生在指导教师或合作企业的指导下独立完成，且要求必须使用来自实际部门的真实数据来解决现实问题[19]。

该专业的人才培养目标定位于数据科学家，特别强调学生“洞察”能力，尤其是从大规模数据中快速洞见对自己有价值的信息，并将其转换为实际行动的能力的培养。主要特色如下：

①重点培养学生的 3C 精神，尤其是好奇心，通过掌握新技术来提升自己的职业竞争力。该专业的学员主要来自于经济学、统计学和计算机科学等专业。

②强调数据科学的 3 个要素，突出数据科学的跨学科性。开设的课程涉及计算机科学、统计学、机器学习及实战应用。此外，该学校特别强调机器学习在数据科学中的重要地位，重视学生掌握和应用机器学习及数据可视化的能力。

③强调实习的重要性。开设有 PLU（Professional Liaison Unit）资助的专业实习项目（Professional Internships Program），将学生派送到 NHS、Facebook、亚马逊、BBC 的实际工作部门进行为期 6 个月的实习。

④重视学生基于产业真实数据来处理现实问题的能力。该学校设有个人大作业（The Individual Project），该作业要求学生综合运用所学知识，选择来自产业、学术或政府的真实数据来解决现实世界中存在的具体问题。

8）卡内基梅隆大学

该学校的数据科学专业人才培养分散在多个专业中，如表 2 所列，其中直接用数据科学命名的专业为计算数据科学（Computational Data Science）[20]。计算数据科学专业由计算机学院开设，课程体系设有分析和系统两个方向，学生必须完成 5 门核心课程、3 门选修课和 1 个 Capstone 项目。分析方向的核心课程为智能信息系统、机器学习、大规模数据集的机器学习、搜索引擎和可扩展分析学；系统方向的核心课程为操作系统的实现、数据库应用、并行计算机架构及编程、分布式系统、数据库系统、高级存储系统、云计算及高级云计算、数据库系统的前沿问题及多媒体数据库。

表 2　卡内基梅隆大学在数据科学专业的硕士培养情况[20]

Table 2　Cultivation of master in the major of data science at Carnegie Mellon University[20]

学院		学位	时间	类型	背景要求	未来工作去向
海因茨学院		公共政策硕士（政策分析方向）	2 年	专业硕士	商业、科学或技术学位	政府、咨询公司、智库
		信息系统管理硕士（商务智能与数据分析方向）	1.5 年	专业硕士	具有工科学位和工作经验	金融服务公司、科技公司、初创企业
泰伯商学院		工商管理硕士（商务分析方向）	2 年	专业硕士	不同的背景（见正文）	咨询公司、IT 公司、财务数据分析公司等
计算机科学学院	语言技术研究院	计算数据科学硕士	1.5 年	专业硕士	计算机科学或其他相关专业	高科技公司的软件工程职位
		智能信息系统硕士	1 年	专业硕士	计算机科学或其他相关专业	高科技公司的软件工程职位
		语言技术硕士	2 年	专业硕士	计算机科学或其他相关专业	软件工程工作、博士项目
	人机交互研究院与心理学系	教育技术硕士	1 年	专业硕士	心理学、教育学、计算机科学等专业	各种相关工作
	机器学习系	机器学习硕士	1.5 年	专业硕士	计算机科学、统计或其他相关专业	软件工程、财务工作、博士项目
迪特里希人文社会科学学院		统计实践硕士	1 年	专业硕士	数学或统计数据相关专业	咨询公司、金融公司、市场营销公司等

该学校侧重于培养专业中的数据科学家，强调与具体专业学科高度融合的人才的培养。其主要特点有：

①侧重于融合式教育及专业中的数据科学家的培养。与上述其他学校不同，该学校的数据科学专业分散在多个学位项目中，如公共政策、信息系统管理、工商管理、计算数据科学、智能信息系统、语言技术、教育技术、机器学习和统计实践等。其中，以数据科学命名的专业只有一个，即计算数据科学。

②强调跨学科方法(Interdisciplinary Approach)。重视统计学、计算机科学和具体应用领域之间的深入融合,所涉及的具体应用领域有公共政策、信息系统管理、商务分析、智能信息系统、语言技术、教育技术与应用学习。

2 特色课程

特色课程是一个新专业存在的标志之一。进一步对上述8个学校的培养方案进行深入调研发现,数据科学与大数据技术专业的特色课程有以下10种。

1)"理论基础"类课程。该类课程主要讲解正式学习数据科学之前必备的知识,而对数据科学本身的介绍较少,是数据科学专业的先修课程,为学生学习数据科学课程奠定基础。常见的理论基础类课程有统计学、机器学习以及Python语言(或R语言)。

①"统计学"类课程。该类课程主要讲解面向数据科学的应用统计学的知识,为学生深入学习数据科学理论奠定基础。例如,华盛顿大学的应用统计与试验设计课程[21]主要学习离散和连续随机变量的推理统计方法,包括手段和比例差异的测试、线性和逻辑回归、因果关系以及重采样方法等。再如,斯坦福大学开设了两门统计学类课程,即现代应用统计学:学习和现代应用统计学:数据挖掘[22]。

②"机器学习"类课程。该类课程主要讲解面向数据科学的应用机器学习的知识,为学生深入学习数据科学理论奠定基础。例如,加州大学伯克利分校开设的应用机器学习[23]课程认为机器学习是计算机科学与统计学交叉点上发展迅速的领域,强调的是寻找数据中的模式。类似的课程还有华盛顿大学的数据科学家常用的统计机器学习[24]和纽约大学的机器科学与计算统计学等课程。

③Python语言(或R语言)课程。该类课程主要讲解面向数据科学的数据分析语言及开源工具。例如,加州大学伯克利分校开设的面向数据科学的Python语言[25]侧重讲解数据科学工作所必需掌握的Python知识——Python的基本语法及数据科学常用包的应用。

2)"基础理论"类课程。该类课程主要讲解数据课程本身的术语、理念、理论、方法、技术、工具和最佳实践应用,属于数据科学专业的入门性、导论类课程。例如,约翰·霍普金斯大学的数据科学[26]课程涵盖数据科学领域的核心概念和技能,包括问题识别和通信、概率、统计推断、可视化、提取/变换/加载、探索性数据分析、线性和逻辑回归、模型评估以及常用机器学习算法等。该课程以有效沟通和可重复分析为指导思想,认为数据科学并不等同于统计学和机器学习的简单拼接,强调对数据科学自身新知识的讲解。

3)"领域应用"类课程。该类课程主要讲解数据科学对某一学科领域的影响及其应用方法论或最佳实践。例如,斯坦福大学开设的数据驱动型医学[27]和基于大数据的商务智能[28]课程分别探讨如何将数据科学的理念、理论方法和技术应用于医学和商务智能领域。

4)"数据呈现和沟通"类课程。该类课程主要讲解数据呈现与沟通能力在数据科学中的重要地位以及数据科学中常用的可视化表示与故事化描述方法。例如,加州大学伯克利分校的数据可视化与沟通[29]、约翰·霍普金斯大学的数据可视化[30]以及伦敦城市大学的可视分析学[31]课程讲解了可视化方法在数据科学专业中的应用。此外,数据的故事化描述也是数据科学家的基本能力之一。杜克大学的交叉数据科学硕士专业认为数据的故事化描述与可视化表示同等重要,并开设课程数据逻辑、可视化表达与故事化描述(Data Logic,Visualization,and Storytelling)[32]。

5)"数据计算"类课程。该类课程主要讲解大数据环境下计算模式的变化及新的算法、技术、工具与平台。例如,华盛顿大学的可扩展的数据系统与算法主要讲解面向大规模数据的可扩展算法。卡内基梅隆大学的云计算[33]课程不仅介绍云计算模式,还涉及数据中心、虚拟化、云存储和编程模型等主题。斯坦福大学也同样开设了关于数据计算的基于数据的计算范式[34]课程。

6)"数据管理"类课程。该类课程主要讲解数据管理,尤其是大数据时代的数据管理新挑战、新理念、新方法、新技术和新工具。例如,华盛顿大学开设的面向数据科学的数据管理主要讲解数据模型、查询语言、数据库调优和优化、数据仓库以及并行处理等内容。加州大学伯克利分校开设的数据存储与检索[35]课程的涉及面很广,鼓励学生综合运用Python、关系数据库、Hadoop、Map reduce、Spark和云计算(AWS)等多种技术来完成分布式数据处理、流式数据分析、图计算和大数据架构设计等工作。

7)"数据分析"类课程。该类课程主要讲解数据分析,尤其是大数据分析的方法和技术。例如,卡内基梅隆大学开设的多媒体数据库及数据挖掘、华盛顿大学开设的大数据分析学(Big Data Analytics)以及哥伦比亚大学(纽约)开设的大数据分析学[36]均强调了大数据分析的主要挑战和新方法。值得一提的是,根据Gartner数据分析价值扶梯模型(Gartner's analytic value escalator),因果分析是大数据分析中的重要组成部分。例如,哥伦比亚大学开设的数据科学中的因果推理课程重点讲解因果分析在数据科学中的应用。此外,探索性数据分析成为数据科学专业的重要课程之一,如约翰·霍普金斯大学和哥伦比亚大学均开设有探索性数据分析课程。

8)"数据产品开发"类课程。该类课程主要讲解数据产品的开发方法、试验设计和优化论等知识。其中,数据产品开发是数据科学专业教育的重要抓手之一。例如,卡内基梅隆大学的智能信息系统的设计与开发。在数据产品开发中,试验设计和优化论是必不可少的支撑课程,如华盛顿大学和纽约大学分别开设了有关试验设计(Design of Experiment)和优化论(Optimization)的课程。

9)"人文"类课程。该类课程主要讲解数据科学的研究与实践中的非技术和工程类问题,主要涉及与大数据和数据分析相关的道德、隐私、法律、经济和社会影响。例如,华盛顿大学开设的以人为中心的数据科学[37]课程涉及数据道德与隐私、算法偏倚、法律框架和知识产权保护、数据溯源和再现、数据管理与长久保存、大数据的用户体验和可用性测试、大规模协同中的道德问题、数据沟通以及数据科学的社会影响。

10)"综合训练"类课程。该类课程主要讲解如何综合运用数据科学专业中学习的理论、方法、技术和工具来解决具体行业中的实际问题,重点培养学生的实战能力。加州大学伯克利分校、约翰·霍普金斯大学、华盛顿大学、纽约大学、哥伦比

亚大学（纽约）的综合训练课程称为数据科学综合训练课程（Data Science Capstone），均强调学生以团队工作的形式，基于具体行业中的真实数据来解决实际问题，从而提升数据洞见、数据产品开发和综合动手能力。

3　启示与建议

目前，我国数据科学与大数据技术专业的建设仍处于起步阶段，课程体系的设计存在不足之处，甚至存在曲解现象。我国大数据教育中存在的常见曲解以及以上分析的借鉴意义主要体现在以下几个方面。

1）曲解之一：数据科学＝计算机科学＋统计学。从目前国内部分高校的培养方案可看出，数据科学专业课程体系主要由计算机科学和统计学两大学科领域的主干课程组成，而对数据科学本身的关注不够，并没有开设数据科学专业的特色课程。需要注意的是，计算机科学和统计学是数据科学的理论基础，并非是数据科学特有的知识[38]。从世界一流大学的数据科学课程设置来看，数据科学专业并非是计算机科学和统计学的简单拼凑，而是更加突出数据科学本身——数据科学的基础理论、数据加工、数据分析、数据计算、数据管理及数据产品开发。本次调查分析发现，数据科学专业中应重视的新课程有：

①数据科学导论或数据科学原理；

②数据可视化或可视分析学；

③数据产品开发；

④探索性数据分析；

⑤大数据分析；

⑥试验设计；

⑦优化论；

⑧因果分析；

⑨数据科学综合训练课程。

2）曲解之二：照搬传统统计学和计算机科学专业的课程。从国内部分高校公布的数据科学专业课程体系看，它们一般均设有两门基础课程——统计学和机器学习，并直接将计算机科学和统计学专业的两门课程照搬到数据科学这一新专业中，甚至教学大纲都没有做任何改动。但是，从上文一流大学的课程设置来看，数据科学专业中讲解统计学和机器学习的方式与统计学和计算机科学等传统学科不同。以机器学习为例：

①加州大学伯克利分校和约翰·霍普金斯大学开设的应用机器学习和实用机器学习课程，强调从应用角度讲解机器学习；

②华盛顿大学开设数据科学家常用的统计机器学习课程强调从数据科学的视角来讲解统计学，突出了数据科学与机器学习之间的关联；

③卡内基梅隆大学的大数据集的机器学习课程强调面向大数据的机器学习；

④纽约大学开设的机器学习与计算统计学课程强调机器学习与统计学并非是简单拼凑，而是深层融合。

3）曲解之三：大数据教育的重点是相关性分析。鉴于数据在相关性分析领域的应用案例和故事较多，部分高校的大数据教育过分强调相关分析，而忽略了因果分析，甚至认为大数据或数据科学不善于或不包括因果分析。因此，因果分析的课程在国内数据科学与大数据技术专业中极其罕见。与之不同，国外数据科学专业中的因果分析课程较为常见，体现了数据分析的多样性以及因果分析在数据科学中的重要地位。例如：

①加州大学伯克利分校开设有试验与因果分析课程；

②哥伦比亚大学的数据科学专业开设有数据科学中的因果推理课程。

4）曲解之四：数据科学与大数据技术专业关注的是数据本身的管理。部分学校的数据科学专业的人才培养方案与数据工程、数据仓库、商务智能等其他专业或方向的区别并不明显，课程设置仍以培养数据工程师为目标，关注的科学问题是数据本身的管理。但是，与数据工程专业不同，数据科学专业侧重于“基于数据的管理”，而并非“数据本身的管理”，其培养目标为数据科学家和数据分析师。例如：

①斯坦福大学开设有数据驱动型医学；

②纽约大学开设有基于优化的数据分析；

③约翰·霍普金斯大学开设有数据产品开发。

5）曲解之五：数据科学与大数据技术专业的课程名中应有“大数据”字样。从部分学校的大数据专业课程体系看，为了区分和凸显新专业的特殊性，它们在每个课程的名称中简单机械地增加了“大数据”字样，如大数据系统与算法等。但是，从国外高校的课程设置中可以看出，数据科学专业的课程不一定要打“大数据”的旗号，例如：

①华盛顿大学开设的可扩展的数据系统与算法课程虽没有“大数据”字样，但充分体现了大数据系统与算法的核心需求和主要矛盾——可扩展性（Scalability）；

②斯坦福大学开设的基于数据的计算范式课程虽然没有“大”字样，但抓住了数据科学的核心问题——基于数据的计算范式。

6）曲解之六：数据科学与大数据技术专业亟待标准化。目前，国内多所高校的数据科学专业的培养方案趋于同质，相互参照得过多，并没有体现各高校自身的优势。从国外课程体系的设计看，不同学校的数据科学与大数据技术专业的人才培养方案并非趋同，在自身的学科优势和人才培养的定位上存在主要区别。例如，斯坦福大学结合自己在统计学、医学、生物学和商务智能上的优势，开设了一些特色课程：

①现代应用统计学：学习；

②现代应用统计学：数据挖掘；

③数据驱动型医学；

④现代统计学与现代生物学；

⑤基于大数据的商务智能。

7）曲解之七：数据科学专业纯属理工科。目前，国内多数学校的数据科学专业的课程设计中仅强调技术和工程问题，忽略了人文和管理问题。但是，从国外数据科学专业课程的设计来看，数据科学不仅是技术和工程的问题，还涉及道德和法律的范畴。例如：

①加州大学伯克利分校开设的大数据中的人与价值课程；

②华盛顿大学的以人为中心的数据科学课程；

③哥伦比亚大学的数据科学 Capstone 与道德课程。

8)曲解之八:数据科学专业的主要受众学生来自计算机科学、统计学或数据科学专业。目前,国内数据科学专业的课程是专门为计算机科学、统计学或数据科学专业的学生设计的,忽略了其他专业学生的需求。但是,从国外大学数据科学专业或课程的选修情况看,该类课程的主要生源并非来自上述3个专业,反而是其他专业的学生占大多数。纽约大学的数据科学专业的多个培养方向也证明了这一点。因此,在数据科学专业的课程设计中应适当考虑学生的来源和去向,加强数据科学与领域知识的高度融合。

参 考 文 献

[1] 全国高校大数据教育联盟.2017申报"数据科学与大数据技术"专业本科院校数量再创新高[OL].http://www.sohu.com/a/168748806_589639.

[2] Steve Pierson. Master's Programs in Data Science and Analytics [OL]. (2017-12-03). http://magazine.amstat.org/blog/2017/06/01/masters-programs2.

[3] New York University. Academics[OL]. https://cds.nyu.edu/academics.

[4] UC Regents. Data Science (DATASCI)[OL]. (2017-11-21). http://guide.berkeley.edu/courses/datasci.

[5] Johns Hopkins Engineering for Professionals. Data Science [OL]. (2017-11-21). https://ep.jhu.edu/programs-and-courses/programs/data-science.

[6] University of Washington | Seattle, WA. Courses & Curriculum [OL]. (2017-11-21). https://www.datasciencemasters.uw.edu/details/courses.

[7] New York University. MS in Data Science Courses[OL]. (2017-11-21). https://cds.nyu.edu/academics/ms-in-data-science/ms-courses.

[8] Stanford University, Stanford, California 94305. M. S. in Statistics: Data Science[OL]. https://statistics.stanford.edu/academics/ms-statistics-data-science.

[9] Carnegie Mellon University. Data Science Overview[OL]. (2017-11-21). https://www.cmu.edu/graduate/data-science.

[10] Graduate Curriculum. Columbia University[OL]. (2017-11-21). http://datascience.columbia.edu/course-inventory.

[11] City, University of London. Data Science[OL]. (2017-11-21). https://www.city.ac.uk/courses/postgraduate/data-science-msc.

[12] UC Regents. Master of Information and Data Science[OL]. (2017-11-21). https://www.ischool.berkeley.edu/programs/mids.

[13] Johns Hopkins Engineering for Professionals. Courses. Courses [OL]. (2017-11-21). https://ep.jhu.edu/programs-and-courses/programs/data-science.

[14] Johns Hopkins Engineering for Professionals. Courses. About [OL]. (2017-11-21). https://ep.jhu.edu/programs-and-courses/programs/data-science.

[15] University of Washington|Seattle, WA. Career Outlook[OL]. (2017-11-21). https://www.datasciencemasters.uw.edu/details.

[16] New York University. MS in DATA SCIENCE[OL]. (2017-11-21). https://cds.nyu.edu/academics/ms-in-data-science.

[17] Stanford University, Stanford, California 94305. M. S. in Statistics: Data Science [OL]. (2017-11-21). https://statistics.stanford.edu/academics/ms-statistics-data-science.

[18] Columbia University. Mission[OL]. (2017-11-21). http://datascience.columbia.edu/columbia-data-science.

[19] City, University of London. Objectives[OL]. (2017-11-21). https://www.city.ac.uk/courses/postgraduate/data-science-msc.

[20] Carnegie Mellon University. Overview: Carnegie Mellon's Interdisciplinary Approach to Data Science[OL]. (2017-11-22). https://www.cmu.edu/graduate/data-science.

[21] 2017 University of Washington | Seattle, WA. Course Descriptions[OL]. (2017-11-21). https://www.datasciencemasters.uw.edu/details/courses/course-descriptions/#DATA557.

[22] Stanford University, Stanford California 94305. STATS315B-Modern Applied Statistics: Data Mining[OL]. (2017-11-21). http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId=1164541.

[23] UC Regents. Info 251Applied Machine Learning[OL]. (2017-11-21). https://www.ischool.berkeley.edu/courses/info/251.

[24] 2017 University of Washington | Seattle, WA. Course Descriptions[OL]. (2017-11-21). https://www.datasciencemasters.uw.edu/details/courses/course-descriptions/#DATA557.

[25] 2017 UC Berkeley School of Information. Python for Data Science[OL]. (2017-11-21). https://datascience.berkeley.edu/academics/curriculum/python-for-data-science.

[26] Johns Hopkins Engineering for Professionals. 605.448 - Data Science [OL]. (2017-11-21). https://ep.jhu.edu/programs-and-courses/605.448-data-science.

[27] Stanford University. BIOMEDIN 215:Data Driven Medicine [OL]. (2017-11-21). http://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&academicYear=20172018&q=+Data+Driven+Medicine&collapse.

[28] Stanford University. OIT 367:Business Intelligence from Big Data[OL]. (2017-11-21). http://explorecourses.stanford.edu/search?q=OIT%2b367&academicYear=20172018.

[29] UC Berkeley School of Information. Data Visualization[OL]. (2017-11-21). https://datascience.berkeley.edu/academics/curriculum/data-visualization.

[30] Johns Hopkins Engineering for Professionals. 605.462 - Data Visualization [OL]. (2017-11-21). https://ep.jhu.edu/programs-and-courses/605.462-data-visualization.

[31] City, University of London. Core modules[OL]. (2017-11-21). https://www.city.ac.uk/courses/postgraduate/data-science-msc.

[32] Duke University. MIDS-Program Overview[OL]. (2012-12-03). https://datascience.duke.edu/content/course-schedule.

[33] Master of Computational Data Science. Masters - CDS - Curriculum[OL]. (2017-11-21). https://mcds.cs.cmu.edu/masters-cds-curriculum.

[34] Stanford University, Stanford, California 94305. Paradigms for Computing with Data[OL]. (2017-11-21). https://statistics.stanford.edu/courses/2014-2015-stats-290.

[35] UC Berkeley School of Information. Storing and Retrieving Data [OL]. (2017-11-21).https://datascience.berkeley.edu/academics/curriculum/storing-retrieving-data.