计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500088

数据科学导论的课程设计及教学改革

朝乐门
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
中国人民大学信息资源管理学院 北京 100872

摘 要 数据科学与大数据技术、大数据管理与应用等新兴专业的建设,以及计算机科学与技术、统计学、信息资源管理等传统专业的改革,均需引入一门关键课程——数据科学导论。在调查分析哥伦比亚大学、纽约大学、哈佛大学和中国人民大学等高校开设的具有一定代表性的数据科学导论课程的基础上,作者结合自己开设的数据科学导论课程以及国家同名精品在线开放课程建设的经验,探讨了该课程的教学目的、教学内容、试验操作、考核方法、教材选择、特色与创新等课程设计问题。现阶段已开设的相关课程的教学改革重点在于培养学生的数据能力,重视数据产品的研发,加强课程建设模式的创新,加快与社会人才需求的接轨,凸显其导论类的课程特征,重视编写代码能力的培育和数据沟通能力的训练。

关键词:数据科学导论;课程设计;大数据
中图法分类号 TP391

Course Design and Redesign for Introduction to Data Science

CHAO Le-men
Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China),Beijing 100872,China
School of Information Resource Management,Renmin University of China,Beijing 100872,China

Abstract Introduction to Data Science is an intrinsic course for not only the development of emerging majors (Data Science and Big Data Technology,Big Data Management and Application,and so on),but also only the innovation of traditional ones (Computer Science and Technology,Statistics,and Information Resource Management,and so on). Course design issues for this novel course,including its objectives,contents,experiments,assessment methods,reference books,personalized design are discussed based upon conducting an in-depth for typical courses offered by Columbia University,New York University,Harvard University and Renmin University of China as well as the author's teaching experience. The redesign of exiting courses on introduction to Data Science should focus on improving the abilities of target students on full-stack data science,data product development,coding for Data Science,and communicating with non-professional users,as well as leveraging alternative course construction models,reflecting social needs,highlighting its roadmap roles for the curriculums.

Keywords Data Science,Course design,Big data

1 引言

大数据给人类带来的最主要的挑战在于新数据和老知识之间的矛盾日益凸显——数据在量级、类型、价值密度以及处理速度要求等方面均已超越了传统知识解释和解决问题的能力,而现阶段人们所掌握与具备的知识尚未随之变化。因此,如何解决新数据与老知识之间的矛盾成为了大数据教育的首要任务。作为大数据时代新兴知识的入门课程,数据科学导论课程是解决这一矛盾的关键环节。目前,数据科学导论已成为大学生的通识课程,不仅是大数据类专业(如数据科学与大数据技术、大数据管理与应用、信息资源管理类等)的核心课程,也是计算机科学、统计学、金融、医学和材料科学等多个专业领域的课程体系中亟待新增的课程。

然而,目前数据科学导论课程的建设中普遍存在两个较为严重的问题:1)对数据科学课程的教学目的、教学内容和教学方法没有任何认识和判断,不知道如何设计这门新课程;2)盲目照搬计算机科学与技术或统计学的经验与套路,将数据科学曲解为计算机科学或统计学的一个分支,这看似在开设一门新课程,实则是对传统课程体系中已有知识的重组,并没有讲清楚如何解决传统知识和新数据之间的矛盾这一核心问题。因此,国内亟待开展数据科学导论课程的设计和教学改革的大讨论。

本文结合作者自 2015 年以来从事数据科学的教学、科研和产业合作经验,尤其是 2017 年开始在中国人民大学开设数

到稿日期:2020-03-12 返修日期:2020-05-20
基金项目:2020 年度教育部人文社会科学研究规划基金(20YJA870003)
This work was supported by MOE(Ministry of Education in China) Project of Humanities and Social Sciences (20YJA870003).
通信作者:朝乐门(chaolemen@ruc.edu.cn)

据科学导论课程，主持建设国家精品在线开放课程——数据科学导论，以及参与部分其他高校的数据科学导论课程和专业建设研讨，来探讨数据科学导论的课程设计和教学改革中的若干问题，供大家参考。

2 相关工作

目前，数据科学导论课程的命名方法有多种，较常见的有数据科学导论、数据科学与大数据技术导论、大数据导论/概论、数据科学（与大数据技术）基础、数据科学（与大数据技术）通识和面向特定专业领域（如商业分析）的数据科学等。本文所探讨的数据科学导论课程包括上述不同术语命名的课程。其中，课程建设历时较长、配套教学资源齐全、社会影响较为大且有一定代表性和借鉴意义的示范课程有 4 个，如表 1 所列。

表 1 部分数据科学导论课程的比较

Table 1 Comparison of typical courses for Introduction to Data Science

	数据科学导论（Introduction to Data Science）	面向商业分析的数据科学（Data Science for Business Analytics）	数据科学导论（Introduction to Data Science）	数据科学导论（Introduction to Data Science）
开设学校	哥伦比亚大学	纽约大学	哈佛大学	中国人民大学
课程负责人	Rachel Schutt	Foster J. Provost	Pavlos Protopapas	朝乐门
学生层次	本科	硕士	本科	本科
教学目的	打通产学之间的鸿沟，帮助学生掌握实际工作中数据科学家应具备的技能	改变学生对数据及其在商业中的角色（作用）的认识	帮助学生掌握运用统计学和机器学习方法进行预测性数据分析的能力	以解决大数据时代的新数据与老知识的矛盾为出发点，讲解数据科学的新理论、方法和技术，培养学生的动手操作能力和继续学习能力
教学内容	数据科学的概述、常用统计建模方法和机器学习知识、数据可视化、社交网络、数据准备与特征工程、数据科学的未来	数据科学及数据分析工具平台概述、预测性数据分析、模型评估、统计学与机器学习的应用、数据可视化以及数据分析工具	数据爬取、数据管理、数据可视化、回归与分类、深度神经网络	大数据挑战及数据科学的核心理论；数据科学技术与工具及 Python 编程；数据产品开发、数据能力评估和数据治理
考核方法	（未找到具体方法）	作业：30%； 学期项目：30%； 期末考试：30%； 课堂表现及出勤率：10%	作业：57%； 小测试：10%； 项目：30%； 参与度：5%	考勤与课堂表现：5%； Python 小技巧演示：5%； 数据科学理论与实践调研：15%； 数据产品设计：15%； 算法与模型小组项目：20%； 期末考核：40%
推荐教材	*Doing Data Science*	*Data Science for Business*	*An Introduction to Statistical Learning*	《数据科学理论与实践》
实验平台	以 R+RStudio 为主，Python 为辅	Python+； Anaconda	Python+； Anaconda	Python+； Anaconda+； Spark+； MongoDB
特色	1. 重视产学结合； 2. 引入 Kaggle 竞赛作为期末项目	1. 面向特定领域开设数据科学导论课程； 2. 重视算法和模型的讲解； 3. 重视学生参与课程建设	1. 重视学生动手操作能力的培养； 2. 强调学生数据沟通能力的培养； 3. 基于 GitHub.io 创建课程主页	1. 以开源课程模式建设课程； 2. 引入数据产品开发、数据故事化等内容； 3. 配套资源丰富

2.1 哥伦比亚大学 Rachel Schutt 的课程

哥伦比亚大学 Rachel Schutt 的数据科学导论课程是美国最早开设的数据科学导论课程之一，对全球数据科学导论课程的建设产生了重要影响。课程负责人 Rachel Schutt 基于他在谷歌研究中心（Google Research）的工作经验，于 2011 年在哥伦比亚大学开设了一门名为 Introduction to Data Science 的课程，并基于自己的教学经历和对该课程的建设经验，与 Cathy O'Neil 合作编著了教材 *Doing Data Science*[1]。该教材于 2013 年由 O'Reilly Media 出版，目前已成为数据科学领域的经典教材。

该课程的教学目的是打通产学之间的鸿沟，帮助学生掌握实际工作中数据科学家应具备的技能。教学内容主要包括：数据科学的概述、常用统计建模方法和机器学习知识、数据可视化、社交网络、数据准备与特征工程和数据科学的未来等。课程强调以 R 语言为主，Python 语言为辅的方式培养学生的动手能力，主要开发环境为 RStudio。

该课程的主要特色在于将具有实际工作经验的数据科学家的知识和经历引入课程教学，在产学合作方面进行了有益探索。课程负责人 Rachel Schutt 曾是谷歌工程师，具有较丰富的实际工作经验。在课程教学中，Rachel Schutt 承担该课程部分内容的教学工作，而其余部分则主要邀请具有实际工作经验的数据科学家来进行讲解。该课程的另一个重要特色是引入 Kaggle 竞赛作为课程期末项目。

虽然 Rachel Schutt 于 2013 年从哥伦比亚大学离职，但是她的课程为哥伦比亚大学数据科学导论及相关课程的建设起到了较大的带动作用。目前，哥伦比亚大学开设了面向不同专业的多门数据科学导论类课程。

2.2 纽约大学 Foster J. Provost 的课程

与 Rachel Schutt 的课程不同，Foster J. Provost 等在纽约大学开设的数据科学导论课程代表了另一类课程——面向特定专业领域开设数据科学导论课程。Foster J. Provost 曾担任纽约大学数据科学中心主任一职，他与 Tom Rawcett 于 2013 年合作出版的课程配套教材 *Data Science for Business: What you need to know about data mining and data analytic*

thinking[2]在数据科学领域产生了重要影响,目前已成为数据科学领域的另一部经典教材。

依据该课程公布的2020年春季教学计划,其主要教学目的为改变学生对数据及其在商业中的角色(作用)的认识。该课程教学内容选择面较广,包括数据科学及数据分析工具平台概述、预测性数据分析、模型评估、统计学与机器学习的应用、数据可视化以及数据分析工具等。其考核方式以过程性考核为主,期末考试只占总成绩的30%,且采取开卷考试方式。课程强调学生在数据挖掘方面的动手操作能力,主要采用Python语言和Anaconda作为动手操作和实验平台。

相对于其他课程,该课程的主要特色有3个。1)面向特定领域开设数据科学导论课程,该课程主要为纽约大学Stern学院的商业管理类硕士生开设,具有明确的专业方向;2)对算法和模型较为重视,教学内容设置中对机器学习算法和统计模型的讲解占比较大,还特别推荐由S Raschka和V Mirjalil合作编写的*Python Machine Learning*:*Machine Learning and Deep Learning with Python*,*scikitlearn*,*and TensorFlow*[3]作为参考书;3)重视学生参与课程建设,如该课程鼓励学生自己推荐动手操作类参考书目。

2.3 哈佛大学 Pavlos Protopapas 的课程

哈佛大学Pavlos Protopapas等的数据科学导论课程历史悠久,可以追溯至2013年[4]。目前,该课程所选用教材为James,Witten,Hastie,Tibshirani等合作编写的*An Introduction to Statistical Learning*[5],教材选择以统计机器学习为主,而不是直接使用数据科学导论类教材。依据该课程公布的2019年秋季学期教学大纲可以看出,它以提升学生运用统计学和机器学习方法进行预测性数据分析的能力为教学目的,课程内容包括数据爬取、数据管理、数据可视化、回归与分类和深度神经网络等。课程强调学生动手操作能力的培养,所选择的主要开发语言和工具分别为Python和Anaconda。课程采取全部过程性考核的方式,无期末试卷考核。

该课程主要有以下3方面特色。1)重视培养学生的动手操作能力。课程中有较大比例的内容对数据统计和常用算法及模型进行了讲解,且每周专门安排实验课程。实验任务主题与课堂教学保持同步,主要内容为Python及第三方包的应用以及基于Python的数据可视化、逻辑回归、KNN分类、PCA、决策树、增强学习和神经网络。2)强调数据呈现与沟通能力,尤其是采用可视化、故事化和可交互方式呈现数据科学项目成果的能力。3)基于GitHub.io创建课程主页,提供包括课程教学日历、PPT、程序代码和数据文件在内的丰富课程资料。

2.4 中国人民大学朝乐门的课程

朝乐门是我国第一部系统阐述数据科学理论、方法、技术、工具和实践的教材《数据科学》(清华大学出版社,2016)的作者。他于2017年开始在中国人民大学开设数据科学导论课程。该课程主要面向信息资源管理专业的本科生,同时也有来自该校多个其他专业的选课同学。课程建设结合主讲人在数据科学领域的学术研究,主持完成教育部-IBM产学合作协同育人项目"数据科学导论课程设计与教学改革研究",参与部分院校数据科学与大数据技术专业申报与建设,以及在企事业单位部门担任数据科学家/大数据顾问的实际工作经历,来主要讲解数据科学理论、方法、技术与工具以及数据产品开发等内容。其中,数据科学的理论体系、数据产品开发、数据故事化、数据规整化处理和探索型数据分析等内容在国内被率先引入到数据科学导论类课程。该课程于2019年被评为中国人民大学一流本科课程第一批建设课程。

该课程的主要特色是以开源课程方式建设课程资源,包括教学大纲、教学方案、PPT和题库等10种资源。课程负责人提出开源课程(Open-sourcing Courses)的设想,于2017年12月23日在北京召开的全国高校大数据教育联盟"数据科学与大数据技术专业核心课程建设系列研讨会"上举行启动仪式,并成立数据科学导论开源课程社区,目前已有200多所学校的教师申请加入数据科学导论课程的开源行动。课程采用主讲人自编的教材《数据科学理论与实践》[6](清华大学出版社),该专著第二版于2019年被北京市教委评为北京市高校优质本科教材。此外,课程负责人基于自己的线下数据科学导论开发出了MOOC课程——数据科学导论,其于2019年被评为国家精品在线开放课程。

总之,国内外数据科学导论课程在教学内容的选择、教学方式的设计以及教学资源的建设等方面均取得了一定的进展。但是,已开设的数据科学导论课程中普遍存在对主讲人的专业背景依赖度过高、课程内容选择面广、缺乏系统性、对数据科学本身的基础理论不够重视、对学生专业差异性不够重视及教学方法较为单一等多种问题[7]。为此,本文在对国内外数据科学导论课程调查分析的基础上,结合作者的实际教学经验,探讨数据科学导论课程的课程设计和教学改革问题。

3 课程设计

通常,课程教学方案的设计涉及一门课程的教学目的、教学内容、实验部分、考核方法以及创新与特色等具体问题。

3.1 教学目的

从根本上看,大数据类专业的人才培养,尤其是数据科学导论课程的教学重点应放在大数据时代新数据和老知识之间的矛盾的解决上。作为数据科学的入门性、导论类课程,数据科学导论课程需要讲解大数据时代及其背后的科学问题——数据科学的理念、理论、方法、技术、工具、应用及数据产品开发等内容,为学生进一步学习数据科学理论奠定基础。具体而言,课程教学的目的是帮助学生:

(1)了解大数据挑战及大数据背后的新科学问题;

(2)理解数据科学中的新理念、理论、方法、技术、工具及其应用;

(3)掌握数据产品研发理论以及数据能力评估模型和数据治理方法;

(4)提升运用机器学习算法和统计学模型解决实际业务问题的能力,提高动手操作能力以及继续学习数据科学知识的兴趣和信心。

当然,数据科学导论的课程设计还需要考虑该课程在不

同专业领域的差异性。为此，我们可以将现有人才培养专业领域大致分为两大专业领域进行讨论。1)大数据类专业，主要特点为将数据从其应用领域分离出来后，专注于对数据本身的获取、存储、计算、管理、分析、挖掘和治理问题进行研究，主要包括数据科学与大数据技术、大数据管理及应用、统计学、计算机科学和信息资源管理类等专业；2)非大数据类专业，即除了上述大数据类专业之外的专业，如材料学、医学、化学、农学、经济学、新闻学等，此类专业对数据问题的研究侧重点往往仅限于其专业领域，因此数据研究有鲜明的领域特色。在课程教学目的上，大数据类专业和非大数据类专业的区别体现在数据科学导论课程在整个课程体系中的定位不同：大数据类专业中开设数据科学导论，主要强调其“导论”作用，因为该专业的后续课程中会开设大数据技术、统计学、机器学习、数据分析、数据加工等多门课程，数据科学相关的课程体系较为完整，数据科学导论课程所涉及的主要内容还会在其他课程中深入讲解。非大数据类专业中开设数据科学导论，应强调其“方法和技术”作用，原因在于非大数据类专业中开设的数据科学类相关课程较少，并不成体系。因此，非大数据专业中开设数据科学导论课程应更加强调如何将数据科学应用于所在专业领域，凸显其方法、技术与工具作用。

值得一提的是，保护学生对数据科学的兴趣以及培养信心也是数据科学导论课程中不可忽视的问题。通常，保护学生的学习兴趣和培养学习信心是课程教学最基本的目的。然而，这一最基本目的却是当前数据科学导论课程教学中面临的主要挑战之一，主要原因在于两个方面：1)数据科学建立在统计学和机器学习等大多数学生觉得较有“难度”的课程之上，学生对先修课程的学习效果或第一印象直接影响数据科学导论课程的教学质量；2)数据科学导论课程的设计存在内容过多或过深的现象，课程内容的广度与深度之间失去平衡，丢掉了“导论类课程”的定位和特征，存在要么以点盖面，要么试图在一门课来讲数据科学这一新学科的全部知识等问题。

3.2 教学内容

在实际教学活动中，数据科学导论课程的教学内容可根据学校所开设专业的特色灵活设置，但其教学内容应至少涉及以下4个方面。

(1)大数据挑战及数据科学的核心理论，主要包括数据科学提出的背景、定义、发展简史、理论体系、术语与本体、基本原则、数据科学项目、数据科学家、如何跟踪数据科学的最新动态、数据科学的典型应用讲解。

(2)数据科学的技术与工具，主要包括大数据的采集、存储、计算、管理和分析的新技术以及常用工具的基本介绍。作为导论类课程，数据科学导论的教学内容的重点并不在于详细讲解上述技术与工具的实现细节，而在于通过对技术和工具的讲解加深学生对数据科学理论的认识并提升其动手操作能力。

(3)数据产品研发，主要包括数据产品的设计、数据治理、数据能力评估以及数据安全、伦理、偏见和道德问题。数据产品研发的教学目的在于培养学生(以数据为中心)的产品设计理念、数据驱动型决策支持能力以及数据密集型解决方案的设计能力。

(4)动手操作实践，主要运用Python或R等数据科学开源工具以及机器学习算法与统计分析方法，练习数据科学的基本流程、主要方法、注意事项和常用技能，进而提升学生解决实际问题的能力。

3.3 实验操作

数据科学是一门实践性很强的学科。从数据科学的韦恩图(Data Science Venn Diagram)[8]可以看出，数据科学课程不仅要有理论知识的学习，还应强调动手操作能力。数据科学导论课程的实验操作通常涉及以下2个层面。

(1)编写代码能力的训练，尤其是运用Python或R等数据科学语言编码代码的能力。代码编写能力是大数据人才的基本功。数据科学导论应包括Python基本语法的熟悉、基于Python的可视化、统计分析和机器学习实践的训练。

(2)数据实验的设计、实施、评价与优化。利用所学知识完成特定专业领域的数据分析和数据产品开发，具体涉及基于数据的研究假设的提出与检验、探索性数据分析、实验设计(Design of Experiment)、数据项目结果的可视化及解释。此类实验的目的在于帮助学生对数据科学项目流程进行全面认识，为后续课程中分别学习云计算、数据工程、大数据分析、深度学习和数据可视化等课程及理解其内在联系奠定基础。

3.4 考核方法

数据科学导论课程适合采取以过程性评价为主的考核方式。从用人单位对大数据人才，尤其是招聘公告中对数据科学家的能力要求来看，数据科学导论课程的考核应重视对学生数据能力、自学能力和沟通能力的培养。然而，这些能力的培养效果无法通过传统的以结果性评价为主的考核方式(如期末试卷)来衡量。因此，应在数据科学导论课程的考核中加大过程性评价的占比，突出学习过程中的能力提升及继续学习兴趣的培养，避免课程成绩对应试能力的高度依赖。

以作者开设的数据科学导论课程为例，该课程平时成绩占总成绩的60%，设有针对大数据领域新理论、技术、工具、产品和应用的调研类作业2次，个人展示自己的Python小技巧作业1次，数据产品设计类作业1次，以及基于机器学习算法和统计学模型的数据分析项目1次。其中，基于机器学习算法和统计学模型的数据分析项目以小组形式合作完成，一般每个小组不超过3名学生。

3.5 教材选择

教材质量是影响课程学习效果的主要因素之一。近年来，我国数据科学和大数据导论类课程的教材如雨后春笋般增长，但优质教材却屈指可数。一本优质教材应体现当前及未来一段时间内社会对人才的需求，全面、准确地阐述学科专业的基本理论、基础知识、基本方法和学术体系，做到理论联系实践；同时，必须结构严谨、逻辑性强、体系完备，能反映教学内容的内在联系、发展规律及学科专业特有的思维方式，凸显创新性和学科特色，富有启发性，有利于激发学生的学习兴趣及创新潜能。在数据科学导论课程的教材选择中应注意以下几个方面的问题。

(1)避免选择过于简单、没有介绍任何实质性知识的教

材。目前,个别教材所讲解的内容没有实质性知识,要么是一些所谓大数据案例,要么是关于什么是大数据的思维辩解,其实质就是一些网络文献的简单拼凑。选择此类教材不仅不能明显提升学生在数据科学方面的知识和能力,还会导致学生对数据科学的曲解,容易使其产生“数据科学和大数据是炒作概念,没有实质性内容”的错觉,给学生造成负面影响。

(2)避免用传统知识,尤其是计算机科学和统计学中的老知识拼凑成数据课程导论的教材。此类教材表面上介绍的是有用的、实质性的知识,但是这些内容并非数据科学的核心理论。选择此类教材会导致学生对数据科学产生另一种曲解,无法使其掌握数据科学领域的理论框架和实践特色,削弱数据科学导论课程在整个课程体系和人才培养过程中的重要地位及开设数据科学导论课程的意义。

3.6 创新与特色

每一门数据科学导论课程都需要有其自己的特色与创新。下面介绍作者所开设的数据科学导论课程的主要特色和创新点。

(1)在教学活动的组织上,有效结合线上课程——国家精品在线开放课程数据科学导论和线下课程——中国人民大学线下课程数据科学导论,引入线上线下混合式教学新模式,将一些容易掌握的显性知识放在线上课程,而将学习难度较大的隐性知识安排在线下课堂,充分发挥二者的互补作用。

(2)在教学内容的选择上,广泛借鉴国外一流大学开设相关课程的教学经验,不仅讲解数据科学的核心理论——数据科学提出的背景、定义、发展简史、理论体系、术语与本体、基本原则、数据科学项目、数据科学家、如何跟踪数据科学的最新动态、数据科学的典型应用;还以“数据产品的研发”为抓手,讲解了数据科学的应用领域,包括数据产品、数据柔术、数据战略、数据能力、数据治理、开发技术及典型案例。

(3)在教材选择上,编写配套教材《数据科学理论与实践(第二版)》(清华大学出版社)。该教材先后受到陈国良院士、IBM 全球战略合作总监与数据科学社区首席技术官 Leon Katsnelson 等国内外专家的一致好评,并于 2019 年被北京市教育委员会评为“北京高等学校优质教材课件”。

(4)在教学方法上,强调启发式教育,提倡以学生为中心的教学模式,重视培养学生对数据科学和大数据技术的学习兴趣和信心;从大数据时代的“新数据”和“旧知识”之间的矛盾入手,重点讲解大数据时代的新理念、理论、方法、技术、工具与实践,以提升学生的数据科学 3C 素质——创造性设计、批判性思考和好奇性提问能力。例如,该课程设有学生个人 Python 小技巧展示环节,每个同学展示 3 个自己认为比较好且有用的小技巧,这既较好地达到了学生的复习或自学数据科学导论课程所需的 Python 知识的目的,又避免了过多介绍 Python 语言而占用数据科学导论课程教学时间的现象。

(5)在解决数据科学本身的快速发展和变化对课程教学带来的问题时,考虑到数据科学理论发展较为快速,对课程内容,尤其是每个专题之后的作业和推荐读书目录进行适当更新。此外,每一专题之后设有“如何继续学习本专题”的讨论环节,用于重点讲解学习方法与注意事项,推荐解读国内外经典图书及重要学术论文,拓展学生的视野,提升学生继续学习和主动学习的能力。

(6)在课程资源的建设方面,提出开源课程的设想,并将本课程的全套资源共享在 GitHub 上,建立数据科学导论开源课程社区,目前已有 200 多所学校的教师申请加入了数据科学导论课程的开源行动中。通过本课程的建设,探讨“MOOC+开源课程+线上线下混合式教学”的教学改革,将课程全套资源分享给全国同行教师,推进教师之间的协作与交流,与全国高校大数据教育联盟合作召开专题研讨会,以总结经验,推动创新。

(7)在实验环节的设置上,课程采用边写代码边讲解理论的方式,改变了传统的“先理论课程后安排上机课程”的做法,学生自带笔记本电脑,用 Anaconda 作为主要实验平台,进行现场操作和现场讲解,不另安排上机课时。通过基于机器学习算法和统计模型的小组项目,锻炼学生合作和沟通的能力。此外,本课程的部分案例中引入了 IBM Data Scientist Workbench(现更名为 Cognitive Class Labs)平台,方便学生进行大数据编程与大数据技术的学习。

(8)在产学研相结合方面,承担教改项目“数据科学导论课程设计与教学改革研究”,获得“教育部-IBM 产学合作协同育人项目”的资助,并因此荣获“IBM 全球卓越教师奖”。此外,任课教师还积极参与国家电网等多家企业的委托项目,并担任多个企事业单位的首席数据科学家或数据顾问,积极参与相关课程的论证和建设研讨会,旨在努力提升自己的教学能力和水平。

4 教学改革

目前,部分数据科学导论课程亟待进行教学改革,而其重点在于明确数据科学导论课程的教学目的与定位,将教学内容回归到数据科学这一新学科,凸显其自身独有的理论与实践,避免基于统计学和计算机科学中的传统知识拼凑出“新课程”的现象。从作者在中国人民大学负责数据科学导论课程的建设以及协助其他兄弟院校进行数据科学导论课程建设的经验看,现阶段数据科学导论课程的教学改革重点在于以下 7 个方面。

4.1 数据能力的培养

数据能力的培养数据科学导论课程建设的核心任务。通常,数据能力指数据人才应具备的独特素质,具体表现在以数据为中心思维模式、数据密集型问题解决能力、数据驱动型决策(支持)能力以及数据产品开发能力这 4 个方面。目前,数据科学导论课程中普遍存在的问题是过于关注计算密集型问题,而忽略了对数据密集型问题的重视,导致数据科学导论课程与其他课程的区别不鲜明,甚至出现了以“统计学+机器学习”来代替数据科学导论课程内容的现象。

脱离于数据能力的培养,数据科学专业,尤其是数据科学导论课程难以准确定位自身的教学目的,无法发挥其在专业课程体系及人才培养中应有的作用。因此,培养学生的数据能力是数据科学导论课程改革的立足点。

4.2 数据产品研发的重视

数据产品研发是目前国内数据科学类导论课程中普遍缺失的重要内容。数据产品不仅限于数据模态的产品，而泛指基于数据实现用户某一目的的产品。从数据科学理论体系看，数据产品开发是数据科学的重要组成部分[9]，正如软件和算法类产品在计算机科学中的重要地位，数据产品才是数据科学对人类作出的重要贡献所在。

数据产品研发活动具有以数据为中心、数据驱动和数据密集型3个特征。常见的数据产品类型包括数据类、算法类、模型类、可视图表、仪表板(Dashboard)、数据故事以及基于数据的传统产品的功能创新和用户体验的优化。数据产品开发的主要内容涉及数据产品的设计、优化与验证、数据治理和数据能力评估。数据科学导论课程应以数据产品作为抓手，重点培养学生的数据产品的设计、优化和评价能力。

4.3 课程建设模式的创新

作为一门近几年才出现的新课程，数据科学导论课程的建设应吸取传统课程尤其是计算机专业和统计学专业中导论类课程建设的经验与教训，尝试一种新的建设模式。其中，开源课程模式是朝乐门提出的一种课程建设新模式，其主要思想是以开源社区模式维护课程资源，集中优势力量打造几门精品课程作为其他课程建设的基础，鼓励课程建设中教师之间知识与经验的转移，将教师从找素材、文字输入和教案排版等可重复性低级体力劳动中解放，进而提升教师备课工作的效果与效率。

除了开源课程行动，数据科学导论课程中还应探索其他新的课程建设模式。例如，课程资源的模块化建设与分享、教学兴趣小组的建设与即时沟通机制、线上线下混合式教学方式的探讨、产学研协同育人、与国际同行合作以及开展学生参与式教学等。

4.4 与社会人才需求的接轨

数据科学导论课程中应关注社会用人单位对大数据人才的实际需求及其变化趋势，并将其作为课程设计和教学改革的重要依据之一。作者在自己的数据科学导论课程建设中进行了以下3项调研工作。

(1)数据科学课程建设现状的调研。2016年，课题组在调查分析数据科学专业特色课程[10]和全球数据科学课程建设现状的基础上，总结出数据科学课程的共性特点、主要挑战及解决对策[7]。

(2)数据科学领域岗位面试题的收集与整理分析。在课程支撑平台——微信公众号“数据科学 DataScience”上公布了数据科学领域中常见的500道面试题。

(3)数据科学领域招聘信息的搜集和分析。通过对大数据岗位招聘公告的收集分析发现，特定算法与模型的应用、沟通能力、A/B测试、探索型数据分析，是数据科学相关招聘公告中常见的岗位要求，因此将其放入数据科学导论课程的教学和配套教材中。

4.5 凸显其导论类课程特征

作为一门导论课程，数据科学导论应强调其“导论”性。作为导论课程，数据科学导论应强调学生对数据科学的核心理论和代表性实践的系统性认识，帮助学生正确掌握数据科学的特殊思维模式、方法论和技术工具，培养学生继续学习数据科学的信心和兴趣，为后续学习其他课程以及自学大数据相关知识奠定基础。但是，目前数据科学导论课程的建设中存在以下两种问题。

(1)课程在专业人才培养方案及课程体系中的定位不明确，与其他课程的内容重复过多。目前，部分院校的数据科学导论课程中存在一个突出问题——该课程的教学目标在整个专业人才培养方案及课程体系中的定位不合理，教学内容的选择主观随意，与后续课程的重复过多，缺乏与其他课程之间的有效衔接。作为导论课程，应做到难易适中，避免片面强调数据科学的某一个(些)方面，导致学生对数据科学的曲解。

(2)课程内容过于简单或陈旧，通过课程学习，学生对数据科学的认识水平和动手能力并没有任何实质性提高。目前，部分院校的导论类课程中存在的另一个问题是课程内容的选择没有任何难度与挑战，导致学生缺乏学习活动的积极性。造成此类问题的主要原因有很多，如所选择的教材过于简单且缺乏实质性内容，所讲解的内容已过时或为众所周知的科普性知识，课程内容与大学低年级甚至高中阶段学习的知识重复过多。

4.6 代码编写能力的培育

目前，代码编写能力已不再是计算机学科领域的特有能力，越来越多的其他领域，尤其是数据科学领域，均需要较强的代码编写能力。目前，数据科学导论课程在编码能力的培育方面存在两个问题。

(1)教学过程中采用高度抽象的“傻瓜式”实验平台，缺少必要的代码编写环节，虽然从表面上看起来这是在做大数据实验，但是此类实验对学生深入认识数据科学没有实质性帮助。数据科学导论课程的教学中不宜采用这种过于抽象和封装的“傻瓜式”数据科学实验平台，应重视关键技术和活动的体验；数据科学本身该有的安装、参数配置、实验设计、代码调试、运行过程的监测、解读的结果、优化实验过程等，是学生掌握知识必不可少的重要环节。

(2)直接照搬计算机专业编写代码的思路，过于关注代码的计算密集型问题，而忽略了数据科学应侧重关注的数据密集型问题，导致数据科学类人才的竞争力和优势不明显。例如，部分学校的大数据类专业中开设的Python课程的教学内容和编程思维与C语言、Java语言等软件开发语言之间没有区别，并未给学生强调或讲清楚面向数据科学的Python编程与面向软件工程或计算机的Python编程的差异性。另外，有的Python编程课程仅以学生能够编写可以执行的源代码为教学目标，并没有注意到PEP8和Pythonic Coding在此课程中的重要地位。

4.7 数据沟通能力的训练

数据沟通能力主要指如何向其他用户，尤其是非专业用户(如数据科学项目中的投资人、管理方和最终用户等)，介绍和解释数据产品的能力。可视化呈现和故事化描述是两种重要的数据沟通能力。

目前，数据科学导论类课程中普遍存在的问题是仅关注

技术和算法上的现象,而忽略了对学生数据沟通能力的训练。数据可视化能力的培养被越来越多的数据科学导论课程所重视,但对另一种重要的数据能力——数据故事化能力的培养仍为空白。为此,作者在自己的数据科学导论课程及其配套教材《数据科学理论与实践》中均对数据故事化理论与实践进行了一定的描述。

结束语 从文中的讨论和分析可以看出,数据科学导论的课程设计和教学改革应该以培养学生的数据能力为中心,避免照搬计算机科学与技术、统计学等相关学科的传统知识和教学经验,积极探索数据科学独有的知识体系和教学方法,正确认识数据科学导论在所在专业人才培养方案中的地位,使数据科学导论课程回归数据科学本身的核心问题,防止课程内容边缘化及空洞化,为学生进一步学习大数据知识提供路线图和方法论,培养学生学习数据科学的兴趣与信心。

数据科学导论的课程建设需要教学、科研、产业合作等3个维度上的综合能力。对于多数教师而言,数据科学是一门全新的学科,仅重视课堂教学、实践操作、科学研究和产生合作中的部分环节,容易导致对数据科学导论课程的偏见。此外,数据科学的教学不仅要重视教师的教学能力,更要关注学生的接受能力,将教学模式从以教师为中心转变为以学生为中心,做到因材施教。

在近几年的亲身教学过程和与相关教师的交流中发现,教师个人魅力和号召力在数据科学导论课程的教学中也是不可忽略的问题。当然,教师个人魅力主要取决于教师本人的长期积累与投入,尤其是在数据科学领域的教学、科研、产业合作的经历。数据科学导论课程的教师魅力主要体现在教师应具备的数据科学家素质——对大数据的批判性思考、好奇性提问和创造性设计能力。

数据科学导论的课程设计与教学改革是一个动态的、持续改进的过程。目前,数据科学本身在快速发展之中,其理论体系尚未完善,人们对数据科学的认识尚未统一。因此,数据科学导论课程的建设需要重视核心知识的讲解、正确知识体系的构建、学生数据能力的提升以及学习兴趣与信心的建立。同时,也应关注社会对数据人才需求的变化。例如,全栈数据科学家(Full Stack Data Scientist)是近几年,尤其是2019年以来的职场热词。全栈数据科学家是相对于"偏科型伪数据科学家"的概念,更准确地说,就是初级数据科学家的别称。全栈数据科学家的兴起进一步表明了用人单位在重新审视数据科学的学科交叉性和全流程性后,为数据科学家这一新的岗位提出了基本要求——数据科学家的知识结构不仅要有深度,更要有广度。数据科学导论是全栈数据科学家培养过程中的核心课程,其课程设计与教学改革应体现这一变化。

参 考 文 献

[1] O'Neil C, SCHUTT R. Doing data science: Straight talk from the frontline[M]. O'Reilly Media, Inc., 2013.

[2] PROVOST F. Data Science for Business Analytics Syllabus—Spring 2020[OL]. http://web-docs.stern.nyu.edu/ioms/SYLLABI/PROVOST_TECHGB2336_Spring20.pdf.

[3] RASCHKA S, MIRJALILI V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2[M]. Packt Publishing Ltd, 2019.

[4] Institute for Applied Computational Science. CS109a: Introduction to Data Science[OL]. https://harvard-iacs.github.io/2019-CS109A/.

[5] JAMES G, WITTEN D, HASTIE T, et al. An introduction to statistical learning[M]. New York: springer, 2013.

[6] CHAO L M. Data science: srinciples and practices[M]. Beijing: Tsinghua University Press, 2017.

[7] CHAO L M, YANG C J, WANG S J, et al. Data Science Curriculums Around the World: An Empirical Study [J]. Data Analysis and Knowledge Discovery, 2017, 1(6): 12-21.

[8] CONWAY D. The Data Science Venn Diagram [OL]. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

[9] CHAO L M, XING C X, ZHANG Y. Data science: the state of art and trend[J]. Computer Science, 2018, 45(1): 1-13.

[10] CHAO L M, XING C X, WANG Y Q. Unique Curriculums for Data Science and Big Data Technology[J]. Computer Science, 2018, 45(3): 3-10.

CHAO Le-men, born in 1979, Ph.D, associate professor, is a member of Technical Committee of Information System of China Computer Federation. His main research interests include data science, big data analytics, and knowledge processing on the semantic Web.