美国数据科学专业教学大纲调查

付希善　曲国庆
（山东理工大学科技信息研究所，淄博 255000）

摘要：数据科学已成为美国高等学术机构提供的常见课程之一，我国高校也纷纷开始建立数据科学专业，对美国数据科学专业教学大纲进行调查可以为我国培育数据科学人才提供借鉴。在考察美国42所高等教育机构数据科学教学大纲中最常开设的课程、最常使用的参考书目及师生交流平台等内容的基础上，将其与大数据科学教学大纲进行比较。美国高校数据科学教学大纲内容清晰而丰富，为学生提供更直接的承诺，更强调学习目标，更具包容性。最后，结合调查结果，对如何形成我国独具特色的数据科学课程体系和教学内容提出建议。

关键词：数据科学；教学大纲；大数据
中图分类号：G251
DOI：10.3772/j.issn.1673-2286.2020.05.007

数据是影响高等教育和图书馆未来最重要的因素[1]。当前，美国许多图书情报机构都将数据科学作为一个重要的学术领域，占据着突出地位。我国教育部为了响应社会发展需要，也于2016年开始正式开设“数据科学与大数据技术”本科专业，随后，全国形成申报与建设数据相关专业的热潮，到目前为止，已有40余所高校获教育部批准开设该专业。随着专业建设的深入，大家发现一个共同的难题：没有系统的数据科学教材，人才培养定位模糊，“数据科学”与“大数据科学”难以区分，也缺乏来自图书情报领域、图书馆等教学一线的知识支持和实践指导。

通过对相关研究梳理，我们发现国内学者对数据科学教育的研究已取得一定的成果，如曹高辉等[2]调研了美国18所高校数据科学硕士专业的培养要求、课程结构和课程设置，提出针对我国数据科学硕士专业建设的建议；陶俊等[3]通过对国外5所具有图书情报背景的iSchool院校在数据科学专业上的培养目标、学分学制和课程结构的调查与分析，得出图书情报学科专业嵌入数据能力的发展建议；陈沫等[4]对国内外大数据相关专业的培养目标和课程设置模式进行调研，设计了情报学取向的大数据专业人才培养计划；苏日娜等[5]选取开设数据科学研究生项目的15所iSchool院校作为调研对象，从专业学科优势、学科体系划分、课程目标、核心课程设置、课程制度等方面研究数据科学课程体系，讨论图书馆与信息科学（LIS）学科下的数据科学定位、数据科学与传统LIS课程结合、LIS人才培养等问题。

从上述研究可以看出，许多研究人员讨论数据科学教育一般专注于数据科学的课程结构和课程设置，却很少关注其教学大纲的内容，而教学大纲是教学内容合乎逻辑的构架，是教学顺序、教学工作的指南，以及师生对该专业掌握程度的反映。有鉴于此，本文对美国高等教育机构的数据科学教学大纲进行调研，考察领先的学术机构和教师是如何组织数据科学教学大纲的。这项研究主要关注五个方面的问题：数据科学教学大纲涉及的课程和参考书目的内容，数据科学专业课程的共同趋势，数据科学教学大纲中教学互动的形式与效果，数据科学教学大纲与大数据科学教学大纲的区别，以及对我国数据科学和大数据专业课程建设的有益启示。

1　调查对象和研究方法

教学大纲既是制作课程和教授学生的标准步骤，也是课程的具体操作路线图。在实践中，教学大纲会

列出课程目标和教学目的、先决条件、课程名称、课程目录编号、课程等级、课程时间和地点、教师及联系方式、评分/评估方案、使用的材料（教科书、软件）及参考书目等常见内容。

笔者针对提供数据科学课程的机构和学院进行了调研，在美国高校的网站中找到有关数据科学的68个教学大纲，发现大多数数据科学教学大纲都来自信息技术领域，包括计算机科学、工程学和信息科学等，在剔除无关的、信息量不足的教学大纲后，选择其中最具代表性的42个教学大纲作为研究对象。

2 研究发现

2.1 数据科学专业教师概况

在调查数据科学入门课程教师时，发现了3种类型的教师，即全职教师、访问教师和兼职教师。调研的40名教授数据科学概论的教师中有32名是全职教师，说明数据科学课程教学有较为稳定的教师队伍，这有力地推动了数据科学学科的发展。在所调研的教师中，助理教授讲授大部分课程（65%），副教授排名第二（20%），教授排名第三（15%）；助理教授一般具有博士学位，有较扎实的理论基础和较丰富的实践经验，而且精力充沛，成为教师的主力，这也与其他学科教师队伍的比例相一致。此外，提供数据科学教学大纲最活跃是加利福尼亚州，有5个教学大纲，佛罗里达州、爱荷华州、马萨诸塞州、纽约州均有3个。显然，美国经济发展较好的地区对数据科学教学大纲的重视程度和信息化建设水平高于经济欠发达地区。

从调研的结果看，超过90%的教学大纲中列出授课教师的姓名和详细信息，如教师的学科背景、教育程度、电子邮件地址或者现场答疑的具体地点等。部分教学大纲未提供教师详细信息，其主要的原因是授课教师的流动性较大，不同教师会根据自身的专长和喜好来制定教学内容，从而造成教学大纲在一定程度上的虚置。

2.2 课程内容

笔者调查了教学大纲中涵盖的主要课程，研究发现出现最多的课程是《数据科学基础》《统计推理》《预测分析》《数据挖掘模型》《机器学习》《可视化》《社交媒体中的数据驱动应用》《数据科学编程语言》和《数据科学伦理问题》等。根据课程出现的次数对其排名发现，最受欢迎的课程是《统计推理》（有78%的教学大纲包括至少一个关于该主题的课程），其次是《预测分析》（72%）和《机器学习》（53%）。调查发现，统计学和计算机科学是数据科学的重要支柱，而机器学习是基于统计学的预测分析，三门课程相互依赖、相互影响，共同构成数据科学的骨干课程。除了上述基础理论课程外，技术类课程也占比较高，显示出数据科学是技术性很强的专业。调查显示，一般学生在入学时会被要求具备统计学知识、数据分析基础和一定的编程能力，如果不具备相关知识，学校会为其补充专业背景知识。而数据科学的编程语言和数据驱动的应用等内容则与该机构的研究情况有关，如有研究人员或教师从事该方向的研究工作，就会增加该方面内容的描述，体现了研究与教学统一的原则。值得一提的是，数据使用的伦理道德问题也被视为教学大纲中的重要课程，说明隐私泄露、数据安全和数据鸿沟等问题已经引起美国学界的高度重视。

研究还调查了这些课程处于教学大纲课程安排的哪个阶段。发现这些课程处于调查的42个教学大纲的不同时段/模块。在美国，一学期一般为14周（不包括假期和考试季）。数据科学教学大纲中最常提到的统计推理和预测分析在14周内平均出现6次，排在第三位的数据挖掘模型平均出现了4次，同时它也会在数据驱动和社交媒体应用的课程中出现。同时，我们调研了各院校数据科学课程的课时，平均课时数为57.6，最长学时数为92，最短学时数为44。从整体上看，学时越长，该机构在数据科学研究和教学方面的实力越强。

2.3 常用的参考书目、论文和博客

教师使用关于课程主题的参考书目是常见的做法。作为课程的补充，参考书目为学生提供了更多的支持。在此次调查中，我们发现比较常用的参考书是O'Neil and Schutt所著的《数据科学实践》、Saltz和Stanton的《数据科学导论》，最受欢迎的参考书是James的《统计学导论》。大部分教学大纲还提供了推荐书籍和论文的清单。这些清单提供了更为多样化的资源，涵盖从一般到具体的数据科学主题。最常提到的论文是Dhar的《数据科学与预测》，其次是Diggle的《统计学：21世纪的数据科学》和Aha等的《基于实例的学习算法》。

教师还经常使用自己的博客来供学生进行学习，除教师自己的博客以外，教学大纲中最常推荐的博客是Data Science Central、KDnuggest、DataScience-101和Revolutions。表1列出了教学大纲中经常提到的参考书、学术论文和博客。

表1 美国数据科学教学大纲中最受欢迎的参考书、学术论文和博客

最常用的参考书		占比/%	最常引用的文章		占比/%	最受欢迎的博客	占比/%
书 名	作 者		题 名	作 者			
《数据科学实践》（*Doing Data Science: Straight Talk from the Frontline*）	O'Neil and Schutt	34	《数据科学与预测》（*Data Science and Prediction*）	Dhar	27	Data Science Central	36
《数据科学导论》（*An Introduction to Data Science*）	Saltz and Stanton	21	《统计学：21世纪的数据科学》（*Statistics: A Data Science for the twenty-first Century*）	Diggle	21	KDnuggest	25
《统计学导论》（*An Introduction to Statistical Learning*）	James et al.	17	《基于实例的学习算法》（*Instance-based Learning Algorithms*）	Aha et al.	16	DataScience-101	14
《数据科学：基本概念》（*Introduction to Data Science: Essential Concepts*）	Morgan	9	《Orange：从实验机器学习到交互式数据挖掘》（*Orange: From Experimental Machine Learning to Interactive Data Mining*）	Demšar et al.	15	Revolutions	9
《数据分析概论》（*A General Introduction to Data Analytics*）	Moreira and Carvalho	6	《机器学习和数据挖掘》（*Machine Learning and Data Mining*）	Mitchell	11	Algorithms Weekly by Petr Mitrichev	7
《统计和数据分析：用R语言练习、解决和应用》（*Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*）	Heumann and Schomaker	5	《数据挖掘中的十大算法》（*Top 10 Algorithms in Data Mining*）	Wu et al.	9	Quantopian Blog	3
《计算材料科学导论：应用基础》（*Introduction to Computational Materials Science: Fundamentals to Applications*）	LeSar	4	《知识发现与数据挖掘：走向统一的框架》（*Knowledge Discovery and Data Mining: Towards a Unifying Framework*）	Fayyad et al.	6	Analyticsvidhya blog	2
《基于R语言的数据分析》（*Data Analysis with R*）	Fischetti	2	《数据挖掘和知识发现中的信息可视化》（*Information Visualization in Data Mining and Knowledge Discovery*）	Grinstein and Wierse	5	Chris Albon-Data Science, Machine Learning, and Artificial Intelligence	2

可以看出，《数据科学实践》是美国高校数据科学专业最受欢迎的参考书，同时该书也是亚马逊数据科学类受欢迎的图书之一。该书基于哥伦比亚大学的数据科学课程，结合案例，由Google、Microsoft和eBay等公司的数据科学家通过展示案例研究和使用的代码来分享新的算法、方法和模型，让读者可以从多个视角打开对数据科学的认知，因此被认为是数据科学入门的必读书目。另外，参考书目中R语言仍然是数据科学最受欢迎的语言，很多文献对于数据科学基本概念（如概率论、统计学、贝叶斯、机器学习）的解释都是基于R语言的，

因此在数据科学领域R语言仍然是最好的工具，不会被其他语言取代，即使是势头正猛的Python[6]。

从最常引用的论文来看，数据挖掘算法类的论文占据了很大部分，说明数据资源的开发利用和技术研究仍然是数据科学中最重要的内容。同时也说明理论类与技术类课程仍然是数据科学专业教育的主体，应用类和实践类课程还未引起足够的关注。此外，众多数据科学家开通博客展示他们的工作，分享他们的研究和感悟，并与对数据科学有兴趣的人交流实践并建立联系。美国高校教学大纲将博客学习作为一种教育活动，因为博客不仅可以促进数据科学知识的扩散，还可以交流反馈，能促生很多新的想法和合作项目。

2.4 教学互动的形式与效果

学习涉及两种类型的互动，即内容互动和人际互动[7]。因此本研究还调查了教学大纲中用于促进学生参与的师生互动平台类型。研究发现，在线论坛是学生与教师最常用的互动形式，在41%的教学大纲中提到，即时通信工具排名第二（29%），电子邮件排名第三（21%），主要的师生互动平台见表2。可见，在信息技术的推动下，师生的交互行为已经从线下转为线上，各类同步、异步交互工具在教学实践中得到广泛应用。由于在线交互具备不受时空限制、多样化的交互形式、实时记录交互内容等特征，使得在线交流已经成为师生互动的主要方式，并取得良好的效果。

表2 用于促进学生参与的教学互动平台

教学大纲中推荐的交互平台	占比/%
在线论坛	41
即时通信工具	29
电子邮件	21
电话	5
面对面会议	4

2.5 数据科学教学大纲和大数据科学教学大纲的比较

美国南佛罗里达大学坦帕信息学院教授Friedman[8]的研究显示，在其所调研的35个美国大数据科学教学大纲中，没有找到任何共同的教科书或参考书，只有一些相关的主题涉及大数据基础、数据源、数据挖掘、机器学习、可视化、统计分析、预测分析、数据清理、数据驱动的应用程序等。显然，在课程主题方面，大数据科学和数据科学并无明显的区别，但二者依然不同，因为大数据是支持分析的基础设施，而分析才是数据科学。“数据科学”一词源于统计学领域，该术语强调数据的统计和发现，而不仅仅是采集[9]。也就是说，我们可以不用分析就使用大数据，如简单的存储日志或媒体文件的位置。我们也可以在没有大数据数据库的情况下使用分析，如使用Excel[10]。

本研究还对Friedman调研的大数据教学大纲和本研究中的数据科学大纲进行了比较。研究发现，使用Palmer等[11]的评估标准，比较数据科学教学大纲与大数据教学大纲得到了不同的分数。在课程目标维度上，数据科学教学大纲的课程目标更加清晰明确，与课程等级、班级规模、授课地点和学生特征更为匹配；在评估维度上，数据科学教学大纲提供了更直接的学习评估、学生评估信息和及时的反馈。因此，数据科学教学大纲在教学目标和评估活动的描述方面得分高于大数据科学教学大纲。数据科学教学大纲提供了更直接的承诺，其语气在整体学习环境和评估活动方面更具包容性。

3 结论与建议

当前，学术界有关数据科学的研究正如火如荼地开展，然而，有关数据科学教育的研究，特别是教学大纲在数据科学教育中的重要作用的研究还未引起足够的重视。在这项研究中，我们发现美国数据科学常见的课程是《统计推理》《预测分析》和《机器学习》等，而受欢迎的参考书是《数据科学实践》《数据科学导论》《统计学导论》等，师生互动的主要平台是在线论坛。研究还发现，美国数据科学教学大纲的评估分较高，在教学目标和评估活动的描述方面得分高于大数据教学大纲，更多地强调了课堂上的学习目标。

数据科学与大数据技术是一个新兴的热门专业，其建设工作还需要不断地发展与完善。我国数据科学教学大纲，要在满足社会现实需求和保持课程内容相对稳定的基础上，形成独具特色的课程体系和教学内容。

首先，要充分吸收国外的优秀成果和教学经验，结合自身的优势和特征，构建本土特色的数据科学教学大纲。既不摒弃我们已有的优良传统和宝贵经验，也不全盘照搬国外的做法，而是在吸收美国等发达国家优

秀成果的基础上，传承优良文化，逐步有序建构中国特色的数据科学课程体系。虽然有学者已经结合我国本土实际，提出了图书馆与LIS学科下的数据科学定位、数据科学与传统LIS课程结合的方案，但我们要重视学生对数据科学概念与技能的掌握，也要强调教师教学和学生学习的过程，同时不能忽视学习环境的建设，要让学生感受到更加直接的承诺、更友善的氛围和对数据科学专业更好的期待。

其次，要从多个视角和多个维度加强对数据科学专业的理解，突出数据科学课程解决实际问题的地位。要成为优秀的数据科学专业教育工作者，必须对数据科学有正确而充分的理解，并将其传达给学生，这是数据科学教育的首要任务。而要理解数据科学，就需要了解数据科学的内涵、原理、理论、方法，把握好数据科学知识间的多元联系和逻辑意义，挖掘其中蕴含的价值。尤其要掌握数据科学和大数据之间的联系和区别，既要加强二者的深度融合，也要厘清边界和范围。另外，教师在讲授完基础内容后，应结合自身的研究方向，紧跟前沿，拓宽学生知识面，为学生推荐各种参考书目、论文等，这样一方面可以让学生充分认识所学内容的重要价值，另一方面可以激发学生学习研究的兴趣。

再次，要实施差别化课程，因需设教、因材施教。数据科学与大数据技术专业旨在培养全面掌握数据科学与大数据基础理论、基本技术与常用工具，了解应用领域大数据，能胜任大数据分析与挖掘、大数据处理系统开发与构建等工作的专门性科学技术人才，其课程体系应全面覆盖所涉基础学科以及大数据对象引发的专门性问题[12]。数据科学既需要理论人才，也需要技术人才，更需要应用与实务方面的人才，尤其对于图书馆来说，数据科学人才极度缺乏，已经严重影响到科学数据管理和服务的正常开展[13]。因此，在具体实践中，相关高等教育机构可以结合自身特征、优势和人才需求来调整相应的课程。学校教育只是人才培养的一个重要环节，强调理论基础和通用技术的系统培养，而只有在真实的工作环境中，才能生成更多的领域知识与实践知识。因此，除了要有稳定的教师队伍、足够的课时、与时俱进的课程内容和参考文献外，与相关企业合作办学，聘请具有丰富实战经验的兼职教师教学，加强应用与实践，也是培养数据科学专业人才的重要途径。

最后，要强调课程评估，保障数据科学课程以学生的学习为中心。课程评估是教学交互的重要环节，是持续的信息反馈，即教师通过采集学生学习信息来告知学生如何改善其学习，同时也提醒教师需要做出何种努力来改进其教学。数据科学的课程评估包括总结性评估、形成性评估和诊断性评估。课程评估与教学相结合，可以有效地推动更多学生对数据科学课程的积极学习和参与，使得学生的学习自主权得到充分保证。教学交互过程可以使用多种平台和方式进行，如在线论坛、即时通信工具和电子邮件等。

另外，由于数据科学与大数据技术的飞速发展，其理论方法和技术工具日新月异，教学大纲中的某些课程可能会不合时宜或过时陈旧。在美国数据科学教学大纲中，通常会依据教师对数据科学的理解，使用不同的知识将数据科学的思想分解成小的内容块。因此，在这种情况下，自编教材将成为一种有效途径。需要强调的是，自编教材对教师提出了很高的要求，需要教师能够全面深入地理解所讲授的课程，掌握教学过程。同时自编教材也对学生学习提出了较高要求，需要学生能够自主学习和主动探究。

总之，本研究的结果可以为我国数据科学和大数据技术专业的课程建设提供一定的借鉴。但教学大纲是一个持续改进的过程，教育工作者需要不断完善教学大纲来吸引学生，改善教学。同时，数据科学的教育工作者需要更好地分析教学大纲的每项内容，以加深对该专业的认识和理解。未来的研究，我们将进一步关注学生对这些课程的期望。

参考文献

[1] LONG P, SIEMENS G. Penetrating the fog：analytics in learning and education [J]. Italian Journal of Educational Technology, 2014, 22（3）：132-137.

[2] 曹高辉，胡紫祎，郭家乐，等. 美国数据科学硕士专业培养要求与课程设置研究 [J]. 数字图书馆论坛，2018（5）：38-45.

[3] 陶俊，何晓东. 面向图书情报的数据科学专业课程结构比较研究 [J]. 图书馆学研究，2019（6）：10-16.

[4] 陈沫，李广建，陈聪聪. 情报学取向的“数据科学与大数据技术”专业人才培养 [J]. 图书情报工作，2019，63（12）：5-11.

[5] 苏日娜，杨沁. LIS学科中数据科学课程体系设置研究——以iSchools高校课程调研为中心 [J]. 图书馆论坛，2019，39（4）：40-49.

[6] Sharp Sight. Why You Should Master R（Even If It Might Eventually Become Obsolete）[EB/OL]. [2020-03-21]. https://www.sharpsightlabs.com/blog/master-r-obsolete/.

[7] LUNE H，BERG B L. Qualitative research methods for the social sciences [M]. Boston：Allyn & Bacon Pearson Higher Ed，2016：145.

[8] FRIEDMAN A. Measuring the promise of big data syllabi [J]. Technology，Pedagogy and Education，2018，27（2）：135-148.

[9] HAYASHI C. What is data science? Fundamental concepts and a heuristic example [M]. Tokyo：Springer，1998：40-51.

[10] Walker Rowe. Big Data vs Analytics vs Data Science：What's The Difference? [EB/OL]. [2019-12-10]. https://www.bmc.com/blogs/big-data-vs-analytics/.

[11] PALMER M S，BACH D J，STREIFER A C. Measuring the promise：a learning-focused syllabus rubric [J]. To Improve the Academy，2014，33（1）：14-36.

[12] 贺文武，刘国买. 数据科学与大数据技术专业核心课程建设的探索与研究 [J]. 教育评论，2017（11）：31-35.

[13] 李金建. 高校图书馆科学数据服务研究现状与趋势分析 [J]. 图书馆工作与研究，2019（12）：86-91.

作者简介

付希善，男，1989年生，硕士研究生，通信作者，研究方向：知识网络与科学评价，E-mail：fuxishan1989@126.com。
曲国庆，男，1962年生，教授，硕士生导师，研究方向：测量数据处理理论与方法、变形监测与数据分析技术、GIS理论及应用、知识素养、知识网络与科学评价。

Survey of Data Science Syllabus in American

FU XiShan QU GuoQing
(Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255000, China)

Abstract: Data science has become one of the most common courses offered by higher academic institutions in the United States. Chinese universities have also begun to establish data science majors, the investigation of the teaching syllabus of data science in the United States can provide reference for the cultivation of data science talents in China. This paper examines the most frequently offered courses, bibliographies and teacher-student communication platforms in the data science syllabuses of 42 higher education institutions in the United States, and compares the data science syllabuses with the big data syllabuses. The research shows that the content of the syllabus is clear and rich, it provides more direct commitment, more emphasis on learning objectives, more inclusive. Finally, combined with the results of the survey, some suggestions are put forward on how to form the unique curriculum system and teaching contents of data science in China.

Keywords: Data Science; Syllabus; Big Data

（收稿日期：2020-03-16）