DOI:10.13530/j.cnki.jlis.2021023

数据科学家:岗位职责、能力要求与人才培养*

朝乐门　肖纪文　王解东

摘　要　数据科学家具有区别于其他专业人才的能力要求和岗位职责,培养数据科学家是数据科学与大数据技术专业的主要使命。本文以 Indeed、LinkedIn 和百度百聘为数据来源,广泛搜集中国、美国、英国、德国、加拿大、日本、澳大利亚和韩国八个国家五种语言的数据科学家招聘公告,挑选出 206 则具有代表性的招聘公告,对其中的任职资格要求和岗位职责描述进行聚类分析和观点挖掘,提炼出数据科学家的能力要求和岗位职责。经调查发现,数据科学家的主要岗位职责有:以数据为中心提出解决方案,从海量数据中洞察有价值的信息,面向具体业务的算法模型研发,假设检验与试验设计,数据治理与数据质量控制,数据产品的研发及基于数据的传统产品的创新,数据全流程的参与以及跨部门和跨领域合作等;数据故事化、因果分析、实时流式处理、部署/生产模型等新兴业务需求将成为未来数据科学家岗位职责的新增长点;数据科学家与数据科学相关的主要能力要求包括:SQL 编程、Python/R/SAS、Hadoop MapReduce/HBase/Hive、Spark/Storm、基于 Tableau 等的可视化分析、ETL 处理、数据仓库/数据湖/BI 技术、统计学与机器学习(含深度学习)、自然语言处理及文本分析和机器视觉等。图 4。参考文献 21。

关键词　数据科学　数据科学家　岗位职责　能力要求　人才培养

分类号　G251.6

Typical Responsibilities, Key Qualifications and Higher Education for Data Scientist

CHAO Lemen, XIAO Jiwen & WANG Xiedong

ABSTRACT

A survey for collecting data scientist job announcements from Indeed, LinkedIn and Baidu Baipin is conducted, and 206 typical cases are selected for the study, which involves 5 languages and 8 countries, including China, the United States, the United Kingdom, Germany, Canada, Japan, Australia, and South Korea. Then, the key qualifications as well as typical responsibilities of data scientists are described via utilizing cluster analysis and opinion mining to provide a basis for the training of data scientists, especially the construction of data science and big data technology major. The qualifications for data scientists can be divided into two categories: data science-specific qualifications and general purpose-oriented ones. Data science-specific qualifications include SQL programming, Python/R/SAS, and Hadoop MapReduce/

* 本文系教育部人文社会科学基金项目“基于数据科学的信息资源管理研究范式创新”(编号:20YJA870003)的研究成果之一。(This article is an outcome of the project “The Paradigm-shifting Research in Information Resource Management Based upon Data Science” (No. 20YJA870003) supported by Humanities and Social Sciences Foundation, Ministry of Education of the People's Republic of China.)

通信作者:朝乐门,Email:chaolemen@pku.org.cn,ORCID:0000-0001-8218-0348(Correspondence should be addressed to CHAO Lemen, Email: chaolemen@pku.org.cn, ORCID: 0000-0001-8218-0348)

HBase/Hive, Spark/Storm, Visual Analysis with Tableau, ETL, Data Warehouse /Data Lake/BI, Statistics, Machine learning (including deep learning), Natural Language Processing, Text Analysis, and Computer Vision. General purpose-oriented qualifications mainly involve the candidate's readiness of communication and cooperation, problem-solving, 3C characteristic of data scientists, independent learning, attention to detail, stress management, and leadership skills. The main responsibilities of data scientists include designing data-centric solutions, finding valuable insights from massive data, developing algorithms/models for specific businesses, hypothesis testing and experimental design, data governance and data quality control, R&D of data products, the innovation of traditional data-based products, as well as participation in the whole data process, cross-department/domain cooperation. Besides, personal charisma, experiences of participating in big data competitions and open source communities, the quality of full-stack data scientists, mathematics and programming capabilities, user-centered design methods, and humanistic issues including big data privacy protection, have an important influence on the core competencies of data scientists. At the same time, emerging business requirements such as data storytelling, causality analysis, real-time flow processing, and deployment/production model, will become novel topics of emerging qualifications of data scientists in the future. The main implications of this study for the data science and big data technology major in China are to focus on the curriculum of data science itself, to introduce Industry-University-Research cooperation, to promote the theoretical research of data science, to develop several core courses, to leverage the capstone role of data product development in data science, to help students develop their self-learning skills, and help students master the basic knowledge and skill of data scientists. Data scientists' key qualifications and typical responsibilities are unique, and to cultivate data scientists is the main mission of data science and big data technology major. 4 figs. 21 refs.

KEY WORDS

Data science. Data scientist. Typical responsibilities. Key qualifications. Higher education.

0　引言

目前,全球数据科学家人才需求增速快且缺口较大。《2017 领英美国新兴职位报告》(*LinkedIn's* 2017 *U. S. Emerging Jobs Report*)显示,2012 至 2017 年间 LinkedIn 上公布的数据科学家招聘信息数量增长了 650%[1]。《2020 领英美国新兴职位报告》(*LinkedIn* 2020 *Emerging Jobs Report*)显示,在美国,数据科学(职位)正在蓬勃发展,并开始取代传统角色,该职位的年增长率为 37%,并表现出以下特征:所需能力涉及机器学习、数据科学、Python、R 语言和 Spark;用人单位主要在信息技术与服务、计算机软件、互联网、金融服务和高等教育等领域[2]。就目前而言,对数据科学家这一新职位的研究主要集中在两个方面。

一是以现有的数据科学家为分析对象,研究其能力和素养共性特点,比较有代表性的是:《2019 年度 Kaggle 的数据科学和机器学习现状报告》(*Kaggle's State of Data Science and Machine Learning 2019*)是基于 19 717 名 Kaggle 注册成员的调查问卷形成的,其中数据科学家占 21%[3];Mikalef 等人[4]基于 113 位 IT 高管的定量调查数据和 27 位大数据项目经理的定性访谈数据,讨论数据科学家应具备的数据技能;Baskarada 和 Koronios[5]基于澳大利亚州或联邦政府机构的九个具有丰富数据科学经验的经

理/总监的半结构化访谈数据,讨论数据科学家的能力要求;Luna-Reyes[6]主要通过文献研究,提出了计算思维、领域知识、数据管理、企业架构、利益相关者的参与等数据科学家的五项核心能力,并讨论了技能要求的多样性;秦小燕和初景利[7]通过文献调研梳理了数据科学家能力体系的研究现状。此外,还有一些研究针对特定领域的数据科学家提出所需的技能,如医疗保健[8]、教育[9]和图书馆[10,11]等。

二是以培养未来数据科学家为目的,从人才培养视角研究数据科学家的知识和能力。此类研究主要涉及四个主题:数据科学课程的建设、相关课程的教学改革、跨学科型人才培养以及女性数据科学家的培养[12]。如,Asamoah 等人[13]讨论了面向数据科学家的大数据分析课程的建设经验;Saltz 和 Heckman [14]探讨了以项目为中心的数据科学导论课程的建设经验;Volpe[15]讨论了意大利数据科学家的教育现状。

由上可见,已有相关研究主要从数据科学家的供给侧,即现有数据科学家的能力分析和未来数据科学家的人才培养两个视角进行,缺少从需求侧,即从用人单位视角对数据科学家的能力要求和岗位职责进行系统研究。本文主要从需求侧角度研究数据科学家这一新职业的能力要求和岗位职责,并进一步探讨数据科学家的需求侧研究对于其供给侧研究,尤其是对人才培养的启示,以期为进一步深入认识数据科学家的能力要求与岗位职责提供一定的借鉴价值。本文结构如下:第一部分主要讨论数据科学家的定义及本文研究的数据来源与分析方法;第二部分深入探讨数据科学家的八种岗位职责,并提出四种新趋势;第三部分主要从专业背景、与数据科学直接相关的能力要求和与数据科学无直接相关的能力要求等三个方面探讨数据科学家的能力要求,并提出五种趋势;第四部分提出对于数据科学家的供给侧研究,尤其是对数据科学与大数据技术专业建设的启示与建议。

1 数据及方法

目前,对数据科学家的能力要求和岗位职责的讨论,主要从个人经验、学术理论体系和领域未来发展趋势等视角出发,相关研究成果主要为学术论文、行业调查报告和媒体文章,研究热点集中在数据科学家的定义、能力要求、岗位职责和人才培养等方面。

1.1 术语定义

数据科学家是将现实问题映射或转换为数据问题之后,主要采用数据科学的理念、原则、理论、方法、技术、工具,通过将数据尤其是大数据转换为知识和智慧的过程中,为解决“现实世界中的问题”提供直接指导、依据或参考的高级专家[16]16。数据科学家的概念及其职业兴起有三个里程碑式的事件:一是 Patil 和 Hammerbacher 于 2008 年首次提出术语“数据科学家”(Data Scientist)[17];二是 Davenport 和 Patil 于 2012 年在《哈佛商业评论》杂志上发表题为《数据科学家是 21 世纪最性感的职业》[18]的文章,之后数据科学家受到媒体和社会的广泛关注;三是美国白宫于 2015 年首次设立数据科学家岗位,并聘请 Patil 为白宫第一任首席数据科学家[19]。

当然,目前的数据科学家只是一个岗位名称,还不是传统意义上的“科学家”。与数据工程师和数据分析师相比,数据科学家的基本知识结构和所关注的问题不同,如图 1 所示。通常,数据工程师的知识结构具有“计算机科学>数据科学>其他应用领域的知识”的特点,所解决的问题和思维模式往往基于计算机科学;数据分析师的知识结构具有“其他应用领域的知识>数据科学>计算机科学”的特点,其所关注的问题和思维模式一般侧重于解决具体应用领域的实际问题;数据科学家应具备“数据科学>计算机科学>其他应用领域的知识”的特点,同时需要掌握一定的统计学和数学知识,主要关注

数据问题,其思维模式往往基于数据科学。

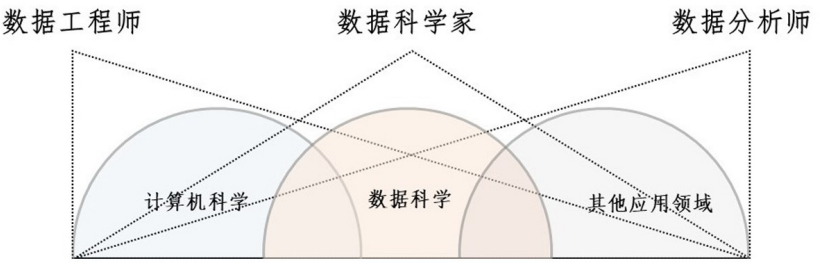

图 1　数据科学家、数据工程师和数据分析师的区别

1.2　数据来源

本文以招聘网站 Indeed、LinkedIn 和百度百聘为数据来源,在 2020 年 3 月 18 日至 5 月 28 日期间,调查分析中国、美国、英国、德国、加拿大、日本、澳大利亚、韩国八个国家五种语言的招聘公告,并按照以下六条原则最终挑选出 206 则具有代表性的招聘公告作为分析数据。

(1)在招聘岗位名称中,选择有明确的"数据科学家"(Data Scientist)字样的招聘公告,排除了虽然岗位名称中有"数据科学家"字样,但内容描述与数据科学或大数据无关的招聘信息。

(2)在招聘岗位类型上,选择全职类型的招聘公告,排除实习和兼职类型的招聘公告。

(3)保留内容齐全、对招聘人员的能力要求和岗位职责均有明确说明的招聘公告。考虑到岗位职责与能力要求的内在联系是本文研究的一个重要问题,在数据选择中排除了缺少能力要求或(和)岗位职责的招聘公告。

(4)同一用人单位在不同国家的招聘公告,只要内容不完全一致,按不同招聘公告对待,但对同一用人单位发出的多则相同或没有本质区别的招聘公告进行重复过滤。

(5)在搜集数据科学家的能力要求和岗位职责信息时,过滤掉了招聘单位自己对技术和产品的独有要求。例如微软公司对招聘人员的 Azure 掌握程度有一定要求。

(6)通常,招聘公告包括"岗位职责"和"任职资格"两个方面,或者招聘公告内容涉及这两方面内容。但在个别招聘信息中,对招聘人员的能力要求和岗位职责两个方面的某个(些)内容描述存在相互混淆或颠倒的情况,本研究保留了此类招聘公告,并把相关内容描述还原至所对应的方面。

1.3　分析方法

本研究的分析方法及步骤如下:①以"数据科学家"(Data Scientist)为关键字,在 Indeed、LinkedIn 和百度百聘等网站搜索并下载招聘公告网页;②按照本文"1.2 数据来源"描述的数据选择原则,过滤掉本研究无法处理或不需要处理的四类招聘公告,即实际内容与数据科学家无关、非全职类(如实习类、兼职类等)、内容不完整或多次重复发布的招聘公告;③以"语句"为单位,将招聘公告切分成若干"条款",并根据所在栏目,对每个"条款"标注"能力要求"或"任职资格"标签;④以标签为分组,采用聚类分析法将"条款"划分成若干类簇;⑤按照所包含的"条款"的数量,对类簇进行排序,选择排名相对靠前的类簇;⑥人工核对和优化上述步骤,并提炼出具有代表性的"条款";⑦结合相关文献研究和作者已有的研究基础及实践经验,提出数据科学人才培养的思路与建议。

2　数据科学家的岗位职责

岗位职责是体现数据科学家工作内容的重

要方面。应聘者入职后,通过参与用人单位的入职培训、常规培训,其知识结构会得到进一步的完善和优化。下文主要从入职后的岗位职责角度调查数据科学家的主要工作内容及其特征。

2.1 提出以数据为中心的解决方案

提出以数据为中心的解决方案是数据科学家的主要岗位职责之一。如米巴赫工程科技有限公司要求数据科学家应根据各类业务情况,提供以数据为驱动的解决方案;蚂蚁集团要求数据科学家要有非常好的产品和业务感觉,能够很好地把产品和业务问题转化成分析问题,同时也能够很好地把分析的结果转化成产品和业务决策。

将业务问题转换为数据问题是提出以数据为中心的解决方案的前提。因此,从业务问题到数据问题的转化也是很多用人单位要求数据科学家具备的主要岗位职责之一。如上海氪信信息技术有限公司要求数据科学家负责项目的需求调研,了解业务逻辑,进行数据分析从而把需要解决的业务问题转化为机器学习/数据挖掘问题。

2.2 从海量数据中发现有价值的信息

如何从大数据,尤其是海量数据中发现有价值的信息是用人单位关注的重要问题。比较有代表性的有:Omnicom Media 集团在数据科学岗位工作要求中明确提出从数据中获得洞察,并通过演示和文档将这些见解传达给非技术受众;Soomgo 公司提出数据科学家负责分析业务范围内的数据以得出见解并验证假设;北京中油瑞飞信息技术有限责任公司发布的招聘公告中,将"运用数据挖掘或统计建模的方法,从数据中发现有用信息,解决实际问题"明确列为数据科学家的岗位职责之一。

从海量数据中发现有价值的信息需要数据科学家具备较强的数据洞察能力。数据洞察(Data Insights)是指从海量数据中快速发现自己所需要的有价值信息,并将其转换为行动的能力[16]34,如图 2 所示。例如, HAYS 公司在招聘信息中,明确提出数据科学家需要识别要收集的数据集并推动数据发现。在数据洞察中,除了数据敏锐直觉和领域经验的积累外,运用数据分析与建模、数据挖掘与知识发现等手段识别潜在的隐藏模式,并将其转换成智慧和行动能力也尤为重要。以德国 Wirecard-Aschheim 公司给出的岗位职责为例,数据科学家需要利用数据科学和分析专业知识来探索和检查数据,以发现模式和以前隐藏的业务见解。

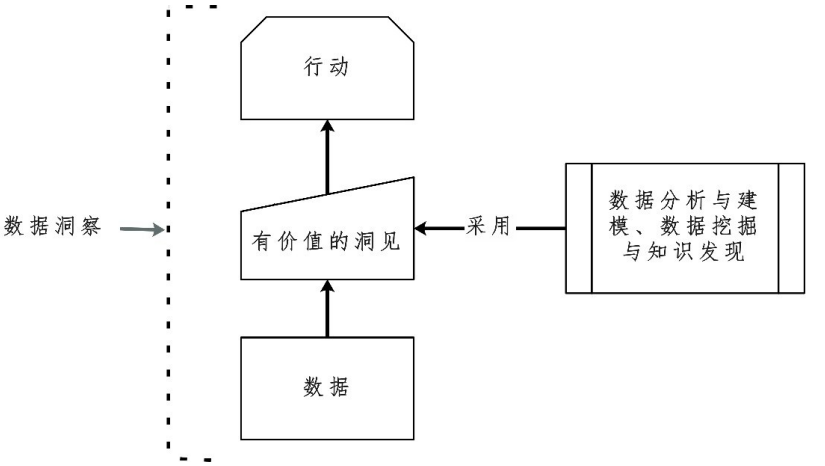

图 2 数据洞察概念的构成要素

2.3　面向具体业务的算法/模型研发

在实际工作岗位中,算法/模型的应用、评估、优化及研发是数据科学家的核心工作任务,主要涉及两方面的工作:一是决策模型或业务分析模型的训练、评估和优化。如 TechSkills Accelerator 公司要求数据科学家应根据公司要求开发新的算法和统计模型;字节跳动公司要求数据科学家负责建立和管理游戏项目管理中的核心决策模型,包括但不限于各品类游戏的留存率、收入、LTV(Lifetime Value,生命期价值)预测模型,市场渠道成本和收入模型,项目评级模型等。二是算法设计、优化与调参。MasterCard 公司要求数据科学家使用各种基于数据科学的技术来开发新算法,以优化现有和开发新的金融犯罪分析产品。Firesoft People 公司要求数据科学家负责开发和维护最先进的高级统计和机器学习模型(如广义线性模型、随机森林、GBM、xGBoost 等),以驱动营销策略和战术执行。

面向具体业务的算法/模型研发不仅要求数据科学家具备较强的机器学习和统计学的知识基础,而且还需要具备一定的动手实践操作能力。具体而言,面向具体业务的算法/模型研发工作内容包括:特征选择与数据准备,算法的选择与超参设置,模型的训练、评估与优化,预测结果的解释等。

2.4　假设检验与试验设计

假设检验是数据科学家的主要岗位职责之一。例如,Warner Bros 集团要求数据科学家负责设计多变量测试以检验、调整和测量自己的假设;Amazon. com Services LLC 要求数据科学家检验多重假设。

与其他专业人才不同的是,数据科学家所提出的假设是基于数据的,侧重于提出数据密集型问题。与计算密集型不同的是,数据密集型问题的主要挑战来自于数据,而不是计算。计算机科学家关注的是计算密集型问题,而数据科学家的主要职责是解决数据密集型问题。假设检验的重点在于将科学研究方法应用到实际问题解决之中,通常需要进行数据试验设计。例如,微软公司要求数据科学家应具有 5 年以上的试验设计经验,并具备将科学方法应用于业务问题和假设检验的能力。

2.5　数据治理与数据质量控制

数据治理是数据科学家的主要岗位职责之一。例如,腾讯公司明确提出数据科学家需要推动数据治理,倡导和共建数据支持、数据驱动业务发展的文化;Alberta 汽车协会要求数据科学家负责数据治理,包括在分析、模型开发和部署中利用版本控制;Bupa 公司要求数据科学家负责支持模型的实施、监控和治理。

数据质量的控制和审计是数据治理的重要工作内容。CGI 要求数据科学家参与处理、清洗和验证用于分析的数据的完整性,创建自动异常检测系统,跟踪性能并根据需要执行其他即时分析;Koch Industries, Inc. 要求数据科学家负责处理、清洗和验证用于分析的数据的完整性。

2.6　数据产品的研发及基于数据的传统产品的创新

与软件产品开发在计算机科学领域的重要地位类似,数据产品的研发是数据科学家对用人单位的主要贡献之一。因此,用人单位期待数据科学家进行数据产品研发。数据产品研发有两种,如图 3 所示。

(1)基于数据的新产品和服务的研发,包括以数据为中心设计的新产品与服务,如数据(Data)、信息(Information)、知识(Knowledge)、理解(Understandings)和智慧(Wisdom)类产品,其中理解类产品有很多种,如模型和算法的解释以及对预测结果的解读等。腾讯公司明确提出数据科学家的基本职责之一是主导数据产品的开发,包含数仓的建立,数据的挖掘、清洗,以及建立智能决策引擎;再如,Pro Search Partners Pty Ltd 要求数据科学家构建从概念验证到生产的数据科学产品。

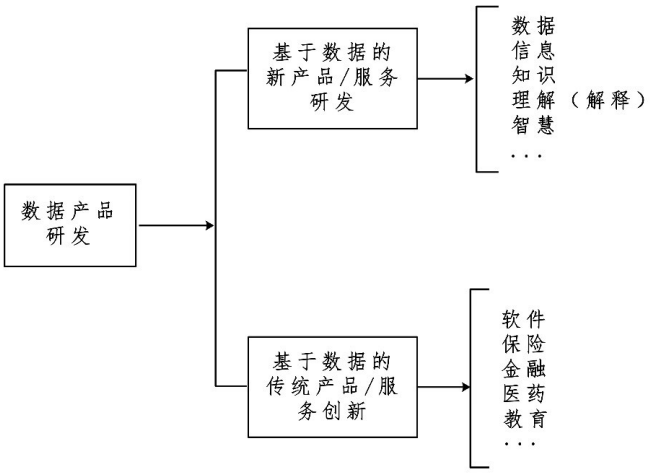

图 3　数据产品研发的类型

(2)基于数据的传统产品和服务的创新,如将数据思维应用于软件、保险、金融、医药、教育等具体行业领域,进而实现传统产品和服务的创新。例如,Roche 公司要求数据科学家利用自己专业特长识别、推荐和开展医疗保健和 PHC 的推进方法,进而改善患者护理活动及其效果;SmartNews,Inc. 要求数据科学家与产品经理、工程师等紧密合作,以确保将数据见解实际转化为具体的产品改进或行动。

2.7　数据全流程的参与

数据科学家通常需要参与数据全流程处理。除了上述六个主要活动之外,数据科学家还可能需要完成以下工作。

(1)数据准备,尤其是数据源的识别与 ETL(Extract, Transform and Load,抽取、转换和加载)。例如,巴克莱银行在招聘信息中要求数据科学家负责与业务、技术和控制组一起定义/完善范围、数据源、统计元素、系统集成等工作,收集业务/功能要求并为 ETL 团队制定技术规范。

(2)数据分析工具选择。例如,Babcock 要求数据科学家负责定义、定制、实施、评估、测量、自动化及改进方法和工具,确保方法和工具在整个组织中得到有效使用。

(3)数据呈现,尤其是数据可视化和故事化呈现。如 StreetLight Data 公司明确提出数据科学家需要创建可视化和仪表板(Dashboards)以及用数据讲故事。

(4)结果解读与模型解释。如 AstraZeneca 公司要求数据科学家为机器学习模型的开发、解释和应用提供专业支持。

2.8　跨部门和跨领域合作

跨部门和跨领域合作是数据科学家岗位职责的一个重要特点。在实际工作中,数据科学项目并非脱离于业务独立存在,而是依附在具体业务项目之中。因此,数据科学家的工作通常需要与来自不同部门和不同领域的利益相关者和专家合作。具体而言,可以分为两大类型。

(1)与来自企业内外部的利益相关者合作,而这些利益相关者往往并非业务专家。例如,TechSkills Accelerator 公司在招聘信息中将“与内部和外部的所有利益相关者合作”明确列在数据科学家的工作职责范围之中。

(2)与不同领域的业务专家团队合作。例如,亚马逊韩国(Amazon Web Services Korea LLC)发布的数据科学家招聘信息中,数据科学家的主要角色与责任之一是与解决方案架构

师、销售、业务开发和 AI 服务团队合作,以加速客户采用和提高收入。

除此之外,部分企业还要求数据科学家具备与学术界有效合作的能力。例如,麦肯锡公司要求数据科学家与学术界建立并保持紧密联系,并不断分享想法,并在最新方法上保持领先;AstraZeneca 公司则要求数据科学家通过同行评审的出版物、在会议上的演讲等,在公司内外推进计算安全的科学研究。

除上述八种主要岗位职责之外,本次调查提示我们还应注意以下 4 种新的岗位职责。

(1)数据故事化是数据呈现的一种新方法,相对于数据可视化,数据故事化具有易于记忆、易于认知和易于体验的特点[20]。数据故事化能力是数据科学家应具备的能力。Oracle 公司要求数据科学家负责分析复杂的数据集,提取见解,得出结论,并能够基于数据讲述影响企业内外部业务决策的引人入胜的故事; Root 保险公司要求数据科学家应具备出色的沟通和讲故事的能力。

(2)因果分析是数据科学领域不可忽视的重要活动。NEC(中国)有限公司要求数据科学家负责综合因果技术和其他相关 AI 技术,解决具体领域的行业问题。

(3)流处理及实时计算是大数据分析的关键所在。MasterCard 公司要求数据科学家具备使用流技术的实践经验,包括流平台(如 Kafka)、在线算法(如随机梯度下降)和固定内存数据结构(如 Bloom Filters)。

(4)部署/生产模型(Deployment/Productionizing Models)主要关注将模型部署到实际应用场景中,并投入生产。部署/生产模型将会是数据科学家的重要职责之一,目前已在个别公司的招聘公告中被明确列为岗位职责,如 TechSkills Accelerator 公司要求数据科学家负责设计、开发和部署分析模型,联邦银行要求数据科学家将机器学习和 AI 驱动的模型部署到生产中(包括实时处理)。

3　数据科学家的能力要求

通过本项调研可以看出,数据科学家的任职要求可以从专业限制、工作经验要求、与数据科学直接相关的知识技能以及与数据科学无直接相关的要求四个方面进行讨论,如图 4 所示。

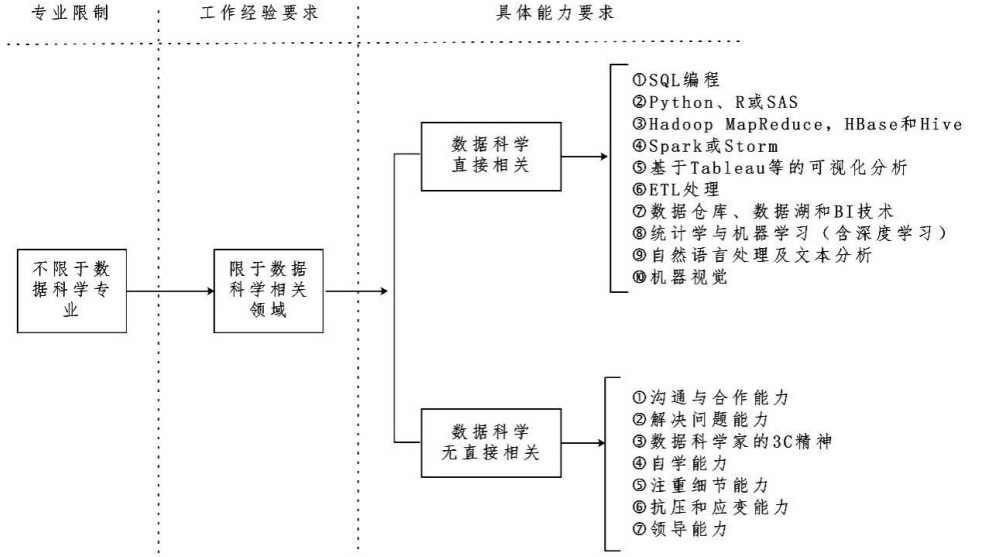

图 4　数据科学家的任职资格要求

3.1 专业背景要求

从调查结果看,数据科学家岗位并不仅限于数据科学、计算机和统计学等众所周知的数据科学相关专业,而是特别强调"定量(研究类)领域"(Quantitative Field)。主要原因有两个:一是数据科学、计算机科学、统计学等专业领域对数据科学人才的培养仍处于起步阶段,人才的数量和质量并不理想;二是数据科学本身具有学科交叉性,对用人单位而言,跨学科性更具有现实意义。

在数据科学家岗位招聘公告中,常见的定量研究类专业为计算机科学、统计学、工程学、应用数学和计量经济学。除上述专业外,信息系统、运筹学、生物统计学、物理和化学等相关学科也较为常见。通常招聘公告对专业要求还会特别注明"以及相关专业"(Related Fields)字样。

3.2 与数据科学直接相关的能力要求

在本研究中,与数据科学直接相关的知识和技能是指所涉及的知识和技能只在数据科学及其相关专业(如计算机学科与技术、统计学等)中进行重点学习,而其他非相关专业中不学习或不会深入掌握的知识与技能。从调查结果看,数据科学家必须掌握数据科学及相关专业中的如下知识和技能,按出现频次从高到低排序为:①掌握 SQL 编程是数据科学家岗位能力要求中最为常见的知识和技能要求;②Python、R 或 SAS 等数据科学语言;③Hadoop,尤其是 Hadoop MapReduce、HBase 和 Hive;④Spark 或 Storm;⑤可视化方法及基于 Tableau、PowerBI 和 QlikView 的可视化分析;⑥数据的 ETL(抽取/转换/加载)处理;⑦数据仓库、数据湖和 BI(Business Intelligence,商务智能)技术;⑧统计学与机器学习(含深度学习),尤其是预测和时间序列分析(指数平滑、ARIMA)、回归(线性和非线性:GLM、SEM、贝叶斯等)、分类(随机森林、SVM、KNN、神经网络)、优化/模拟和聚类(K 均值、DBScan 等)和异常检测(长短期存储网络、一类 SVM 等)。此外,要求掌握 scikit-learn、TensorFlow、Keras、PyTorch 和 PySpark 等常用包的熟练调用;⑨自然语言处理及文本分析法,如文本聚类、LSTM、SVM、关联分析、神经网络、朴素贝叶斯、TF-IDF 和 SVD 等;⑩机器视觉,如 OpenCV 编程。除上述知识和技能之外,还有 A/B 测试、试验设计、探索型数据分析、Lambda 架构、Git、MATLAB、JAVA/C++/Scala。

3.3 与数据科学无直接相关的能力要求

除上述与数据科学直接相关的知识和技能外,数据科学家的招聘信息中还会经常提到如下几种能力。

(1)沟通与合作能力,主要包括:一是与非技术类受众(Non-technical Audience)沟通的能力,即通过演示和文档将复杂的数据分析结果讲述给非技术类受众,尤其是领导层和客户群体;二是与媒体和客户打交道的能力,尤其是营销类数据科学家需要具备与客户和媒体密切合作的能力;三是即时报告能力(Ad-hoc Reporting),即根据用户需求和数据分析结果的变化给出动态有效的描述与解释;四是与来自企业内外的利益相关者有效沟通的能力,尤其是在跨学科和跨文化团队合作中需要具备提出数据科学解决方案的能力。此外,部分招聘信息中还特别强调了数据科学家与学术界之间的团队合作能力。

(2)解决问题的能力,即采用工程化方法,基于可用的数据提出问题解决方案的能力。解决问题的能力是多数专业人才的一个通用要求,但是数据科学家解决问题的能力具有其特殊性,即基于数据的解决问题的能力、数据驱动型业务决策支持能力、以数据为中心的思维模式和能够解决数据密集型问题。从招聘信息看,数据科学家的问题解决能力强调的是提出解决方案和决策支持的能力,并非解决方案的执行能力。

(3)数据科学家的 3C 精神。3C 精神包括解决问题的原创性(Creative)、思考问题的批判

性(Critical)和提出问题的好奇性(Curious)[21]。部分招聘公告将数据科学家的3C精神称为"喜欢有挑战的工作",且特别提到上述三种精神的"天生"特点,强调应聘人才对数据问题的热爱和天生才华。如Stanley Black & Decker公司要求数据科学家要有天生的好奇心以及对实证研究和解决问题的强烈热情;Loblaw Companies Limited要求数据科学家应具备精湛的分析和批判性思维能力。

(4)自学能力。本次调研发现,自学能力在数据科学岗位任职要求中较为常见。背后的主要原因有两个:一是数据科学的知识体系尚未成熟,数据科学家需要及时学习最新知识,不断更新和完善自己的专业知识体系;二是目前数据科学及其相关专业中所学知识不是深度不够就是广度不足,在实际工作中数据科学家还需要结合业务需求不断学习特定领域知识。例如,Tessella公司要求数据科学家应具备快速学习新领域和新技术的渴望和能力。

(5)注重细节(Detail-oriented)的能力。调研发现,注重细节的素质在数据科学家岗位认知要求中出现频次不少。例如,苹果、加拿大皇家银行(RBC)、Source Systems、SentiLink和Loblaw Companies Limited的数据科学家岗位招聘信息中均要求应聘者具备注重细节的能力或素质,尤其是注重数据质量以及程序代码的可读性、可维护性和健壮性。

(6)抗压与应变能力。一些公司在数据科学家招聘公告中特别提到,应聘者要有适应不断变化或模糊/复杂问题的能力以及承受压力的能力,强调应聘者不仅要有团队合作精神,而且还要具备较强的独立工作能力。例如,Oracle要求数据科学家能够适应快节奏的工作环境和不断变化的工作需求;Omnicom Media集团要求数据科学家应具备在快节奏的增长环境中工作的能力;Welab公司要求数据科学家应能够接受及适应快节奏、测试驱动、持续迭代的工作环境。

(7)领导能力。数据科学家的任职通常都会要求应聘者有一定的领导经验,特别强调领导大型项目的经验以及将复杂问题转化为简单问题的能力。例如,腾讯公司在数据科学家招聘公告中明确提出具有两年以上领导、管理和团队指导经验更佳;TechSkills Accelerator公司要求数据科学家最好具有担任小型项目的项目经理或大型项目团队领导的经验。领导经验在数据科学岗位任职要求中较为常见的主要原因是领导能力是一种综合素质,可以较好地集中体现上述六种能力。

3.4　其他要求

数据科学家岗位均有一定的工作经验要求,而工作经验的时长主要取决于应聘者的学历水平和应聘职位的高低,即初级/中级/高级数据科学家。通常,最高学历为学士学位和硕士学位的应聘者分别要求至少有五年和两年的工作经验。而对已获得博士学位者没有特别的工作经验年限要求,但部分生物医疗和健康医疗类企业对博士学位获得者的工作经验会给出限制。例如,赛默飞世尔科技(Thermo Fisher Scientific)和联合健康集团(United Health Group)对博士学位获得者申请高级数据科学家分别要求有三年和五年的工作经验。

目前,虽然数据科学家岗位的专业背景要求并不仅限于数据科学及其密切相关的专业,但是数据科学家的工作经验要求往往明确限制在数据科学及其相关领域,较为常见的业务领域有:数据科学、数据分析、商务智能、数据挖掘、预测分析、统计建模、机器学习(包括深度学习)、软件开发、定量分析、工程开发、自然语言理解和机器视觉等。

不同于其他岗位,一些数据科学家招聘公告中还会强调要有领导经历,如有担任项目管理、产品经理和部门领导的经历。这与数据科学家的另一个任职要求——沟通、协同和合作能力密切相关。例如,微软公司要求数据科学家具有与利益相关者进行有效互动、协作和协商的能力,包括有效的优先级排序、清晰的沟通

以及可操作的数据派生见解的传递,而在这些能力的培养过程中领导经历是必要的途径。

除上述能力要求外,调研工作中还发现一些能力要求的出现次数虽不多,但在一定程度上能够表明未来趋势,具有较高的借鉴意义。

(1)个人威望和影响力。Amazon Web Services 要求数据科学家应是能够令人信服地影响和建立受众心智的天才。

(2)参加竞赛及开源社区的经历。例如,普源精电科技股份有限公司要求数据科学家应具备良好的工程能力,遵循技术规范,能将机器学习模型应用于实际生产中,掌握 Git、SVN 等源代码管理工具及 Maven、sbt 等编译工具。金山公司的数据科学家应聘要求中设有"加分项"一栏,并注明"参加过机器学习与数据挖掘相关竞赛(如 Kaggle KDD Cup 等),能够熟练使用常用算法和数据结构,对算法有较强的实现能力,参与过商业化的模型落地项目"。

(3)全栈数据科学家(The Full Stack Data Scientist)的素质。全栈数据科学家是近几年,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,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重新审视数据科学的学科交叉性和全流程性,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。Booz Allen Hamilton 公司在发布的招聘公告中将数据科学家岗位称为全栈数据科学家(Full Stack Data Scientist)。

(4)数学和编程能力。很多用人单位要求应聘者需具有数学专业背景或优先考虑有数学专业背景的应聘者。例如,HAYS 公司要求数据科学家应具备较强的数学技能,如统计和代数;Loblaw Companies Limited 要求数据科学家应具备 Python、PySpark、Scala 或 R 的专业编程技能,精通机器学习框架和库,且至少有 3 年使用高级 SQL 的经验。

(5)数据科学中的人文与管理问题,尤其是以用户为中心的数据产品设计方法和数据隐私保护。如 BAE Systems Applied Intelligence 公司要求数据科学家应了解以用户为中心的设计(User Centric Design,UCD)、数据隐私和数据安全准则;WooliesX 公司要求数据科学家负责开发、维护和增强各种客户细分模型,以发现客户见解。

4 讨论与结论

大数据类专业,尤其是数据科学专业需要重视数据科学家的培养。培养数据科学家是数据科学与大数据技术专业人才培养不可忽视和不可推卸的责任,也是数据科学专业与计算机科学、统计学等其他相关学科在人才培养上的根本区别所在。但是,目前我国部分院校数据科学与大数据技术专业建设盲目追求"特色化",建设思路过分依赖于主观判断,缺少必要的调查与论证工作,将"主观"当作"特色",偏离了数据科学本身的学科定位和人才培养目标。本次调研对数据科学家的人才培养,尤其是数据科学与大数据技术专业的建设具有以下启示。

(1)数据科学与大数据技术专业的人才培养需要回归数据科学这一新学科本身。我国数据科学专业建设亟待开展对数据科学的本质特征,尤其是数据科学和其他相关学科之间区别的大讨论。在数据科学人才培养中,应避免将数据科学专业简单等同于统计建模、机器学习、数据分析、数据可视化或数据工程,而应将数据科学专业的人才培养聚焦于数据科学家的培养。从调研中可以看出,数据科学专业的人才培养需要探索自己的新模式,突出培养学生的基于数据的问题解决能力、数据驱动型业务决策支持能力、以数据为中心的思维模式和解决数据密集型问题的能力。

(2)数据科学与大数据技术专业建设需要进一步加强产学研相结合。目前,数据科学虽然是热门话题,但对其专门的深入研究较少,数据科

学的理论体系尚未形成,其核心理论有待深入研究。同时,数据科学家的供给侧——人才培养和需求侧——用人单位之间存在较大鸿沟,导致一方面社会对数据科学家的需求越来越多,但另一方面高等院校培养的学生难以满足用人单位的人才需求。数据科学专业建设不仅要引入和借鉴产业用人需求和工作经验,更重要的是需要加强数据科学本身的理论研究。

(3)数据科学与大数据技术专业建设中应开设核心课程。目前,不同院校在数据科学类专业课程设置上差异较大,部分学校甚至没有将数据科学的核心内容纳入课程体系中。数据科学类专业核心课程的内容应包括数据科学导论、应用统计学、应用机器学习、大数据分析、Hive 或其他 NoSQL 数据库、Spark、SQL 数据库(关系数据库)、Python/R 编程、数据可视化、人工智能、深度学习及应用、数据科学的人文与管理问题等。

(4)数据产品研发是数据科学人才培养的抓手,也是数据科学人才区别于计算机科学和统计学人才的核心竞争力所在。但是,国内绝大部分院校尚未将数据产品开发作为数据科学类专业人才培养的侧重点,导致数据科学专业与其他相关专业(如计算机科学、统计学等)的差异不明显,数据科学类专业缺乏核心竞争力。因此,数据科学与大数据技术专业的人才培养应重视数据产品研发,包括以数据为中心的新产品(或服务)以及基于数据对传统产品(或服务)的创新。

(5)数据科学与大数据技术专业建设应重视学生自学能力的培养。数据科学是一个快速发展的学科,学生需要通过不断学习最新知识以跟进社会用人需求。学生自学能力的培养关键在两个方面:一是厚基础、重应用,加强学生对基础知识的学习,尤其是数学与统计学为主的定量分析、数据科学基础理论、大数据技术等基础理论,为今后自学奠定基础;二是保持学生学习数据科学的兴趣,避免由于课程设置或教学方式的缺陷导致学生失去信心。

(6)数据科学与大数据技术专业的人才培养不仅需要帮助学生掌握与本专业相关的知识和技能,而且还需要重视培养与专业无直接关联的基本能力和素质。从调研可以看出,与数据科学专业直接相关的主要知识和技能有:SQL 编程,Python、R 或 SAS 等数据科学语言,Hadoop、Spark 或 Storm、数据可视化、ETL、数据仓库与 BI、统计学与机器学习、自然语言处理和机器视觉等。与数据科学专业无直接关联的知识和技能包括:沟通与合作能力、解决问题的能力、数据科学家的 3C 精神、自学能力、注重细节的素质、抗压与应对变化的能力以及领导能力。

目前,声称自己是“数据科学家”的人越来越多,但是他们的知识与能力往往参差不齐,有的只会一点儿机器学习或统计学甚至是 Excel 等简单工具的使用。在职场上,用人单位也开始注意到这种仅仅掌握“冰山一角”的知识或经验的人根本无法胜任“数据科学家”这一新兴的职位。数据科学家并不是传统数据人才在大数据时代的新提法,而是应具备数据科学全流程的大部分知识与技能,包括从数据感知到数据产品开发、从问题提出到解决方案执行、从数据试验设计到编程实现以及从模型训练到部署/生产的全栈知识与能力。

参考文献

[1] LinkedIn's 2017 U.S. emerging jobs report[R/OL]. [2020-05-01]. https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report.

[2] LinkedIn 2020 emerging jobs report[R/OL]. [2020-05-01]. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf.

[3] Kaggle. Kaggle's state of data science and machine learning 2019[R/OL]. [2020-05-01]. https://www.kaggle.com/kaggle-survey-2019.

[4] Mikalef P, Giannakos M N, Pappas I O, et al. The human side of big data: understanding the skills of the data scientist in education and industry[C]//2018 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2018:503-512.

[5] Baskarada S, Koronios A. Unicorn data scientist: the rarest of breeds[J]. Program Electronic Library and Information Systems, 2017,51(1):65-74.

[6] Luna-Reyes L F. The search for the data scientist: creating value from data[J]. ACM SIGCAS Computers and Society, 2018,47(4):12-16.

[7] 秦小燕,初景利. 国外数据科学家能力体系研究现状与启示[J]. 图书情报工作,2017,61(23):40-50. (Qin Xiaoyan, Chu Jingli. Research and enlightenment on data sciectist competency systems abroad[J]. Library and Information Service,2017,61(23):40-50.)

[8] Meyer M A. Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings[J]. Journal of the American Medical Informatics Association, 2019,26(5):383-391.

[9] Agasisti T, Bowers A J. Data analytics and decision making in education: towards the educational data scientist as a key actor in schools and higher education institutions[M]//Ohnes G, Johnes J, Agasisti T, et al. Handbook of contemporary education economics. Cheltenham, UK: Edward Elgar Publishing, 2017:184-210.

[10] Ekstrom J, Elbaek M K, Grigorov I, et al. The research librarian of the future: data scientist and co-investigator [EB/OL]. [2020-05-01]. https://blogs.lse.ac.uk/impactofsocialsciences/2016/12/14/the-research-librarian-of-the-future-data-scientist-and-co-investigator.

[11] Erdmann C. Data scientist training for librarians[J]. Open Science: At the Frontiers of Librarianship, 2014, 492: 31.

[12] 朝乐门,邢春晓,张勇. 数据科学研究的现状与趋势[J]. 计算机科学,2018,45(1):1-13. (Chao Lemen, Xing Chunxiao, Zhang Yong. Data science studies: state of the art and trends[J]. Computer Science, 2018,45(1):1-13.)

[13] Asamoah D A, Sharda R, Hassan Zadeh A H, et al. Preparing a data scientist: a pedagogic experience in designing a big data analytics course[J]. Decision Sciences Journal of Innovative Education, 2017,15(2):161-190.

[14] Saltz J, Heckman R. Big Data science education: a case study of a project-focused introductory course[J]. Themes in Science and Technology Education, 2016,8(2):85-94.

[15] Della Volpe M, Esposito F. How universities fill the talent gap: the data scientist in the Italian case[J]. African Journal of Business Management, 2020,14(2):53-64.

[16] 朝乐门. 数据科学理论与实践[M]. 第2版. 北京:清华大学出版社,2019. (Chao Lemen. Data science theory and practice[M]. 2nd edition. Beijing: Tsinghua University Press,2019.)

[17] Patil D J. Building data science teams[M]. O'Reilly Media, Inc., 2011:12.

[18] Davenport T H, Patil D J. Data scientist:the sexiest job of the 21st century[J]. Harvard Business Review, 2012,90(5):70-76.

[19] Smith M. The White House names Dr. DJ Patil as the first US chief data scientist[EB/OL]. [2020-05-01]. https://obamawhitehouse.archives.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist.

[20] 朝乐门,张晨. 数据故事化:从数据感知到数据认知[J]. 中国图书馆学报,2019,45(5):61-78. (Chao Lemen, Zhang Chen. Data storytelling: from data perception to data cognition[J]. Journal of Library Science in China,2019,45(5):61-78.)

[21] 朝乐门. 数据科学[M]. 北京:清华大学出版社,2016:28. (Chao Lemen. Data science[M]. Beijing:Tsinghua University Press, 2016:28.)

朝乐门 中国人民大学信息资源管理学院副教授,数据工程与知识工程教育部重点实验室(中国人民大学)研究员,博士生导师。北京 100872。

肖纪文 中国人民大学信息资源管理学院硕士研究生。北京 100872。

王解东 中国人民大学信息资源管理学院硕士研究生。北京 100872。

(收稿日期:2020-06-01;修回日期:2021-04-15)