计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600033

数据科学平台:特征、技术及趋势

朝乐门　王　锐
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
(chaolemen@ruc.edu.cn)

摘　要　以2015年以来的《Gartner数据科学平台魔力象限系列年度报告》为线索,分析调研35种数据科学平台产品,提出数据科学平台的定义和类型。数据科学平台相关学术研究中的主要科学问题涉及数据科学平台的设计、数据科学平台的可扩展性、基于数据湖的数据科学平台研发、数据科学平台的支持团队协作能力、数据科学平台的开放策略以及数据科学平台工程方法论。数据科学平台的主要特征包括模块化开发及集成能力、开发运维一体化、重视可扩展性、强调用户体验、重视非专业级数据科学家以及重视人机协同场景;数据科学平台的实现需要的关键技术为机器学习、流处理技术、数据规整化、容器化技术和数据可视化;数据科学平台的未来发展趋势主要体现在与人工智能的融合、对开源技术的支持、对非专业级数据科学家的重视、数据治理的集成、数据湖的引入、高级分析及应用的探索、向数据科学全流水线的转型和应用领域的多样化等;数据科学平台的研发活动应遵循以激活数据价值为中心、人在环路(human-in-the loop)的设计模式、开发运维一体化、可用性和可解释性的平衡、数据科学产品生态系统的培育、强调用户体验以及与其他业务系统的集成等设计原则。现阶段的数据科学平台研发亟待在数据偏见与公平性、鲁棒性及稳定性、隐私保护、因果分析、可信任/负责任数据科学平台等方面进行理论突破。

关键词:数据科学平台;数据科学家;开发运维一体化;可解释性;可扩展性
中图法分类号　TP391

Data Science Platform:Features,Technologies and Trends

CHAO Le-men and WANG Rui
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　The concept and types of data science platform are proposed based upon in-depth studies of more than 35 data science platforms from the annual report of Magic Quadrant for Data Science Platforms since 2015. The main scientific issues in the academic research of data science platform involve the design of data science platform,the scalability of data science platform,the research and development of data science platform based on data lake,the supporting team cooperation ability of data science platform,the open strategy of data science platform and the engineering methodology of data science platform. The main features of data science platform include modular development and integration capability,DevOps,emphasis on scalability,emphasis on user experience,emphasis on citizen data scientist,and emphasis on human-machine collaboration scenario. The key technologies for the realization of data science platform are machine learning,stream processing,tidy data,containerization and data visualization. The future development trend of data science platform is mainly reflected in the integration with artificial intelligence,the support for open source technology,the emphasis on citizen data scientists,the integration of data governance,the introduction of data lake,the exploration of advanced analysis and application,the transformation to the whole pipeline of data science and the diversification of application fields. The research and development activities of data science platform should follow the design principles of activating data value as the center,human-in-the loop,DevOps,balance of usability and explainability,cultivation of data science product ecosystem,emphasis on user experience and ease of use,and integration with other business systems. At present,the research and development of data science platform needs theoretical breakthroughs in data bias and fairness,robustness and stability,privacy protection,causal analysis,trusted/responsible data science platform.

Keywords　Data science platform,Data scientist,DevOps,Explainability,Scalability

1　引言

数据平台的研发是目前数据科学领域的热点问题之一。但是,相对于工程化开发实践来说,对数据科学平台的理论研究仍未深入进行,数据科学平台的特征、技术与系统等核心问题有待进一步系统研究。数据科学平台理论研究的缺失不仅

到稿日期:2021-04-03　返修日期:2021-06-03
基金项目:国家自然科学基金项目(72074214)
This work was supported by the National Natural Science Foundation of China (72074214).
通信作者:王锐(wangrui1998@ruc.edu.cn)

严重限制了数据科学产品的工程化开发及升级优化，而且还将成为数据科学平台产业化发展的主要瓶颈。因此，对数据科学平台研究现状的调研对于数据科学理论的研究具有重要推动作用。

本文以2015年至今连续6年的7份《Gartner数据科学与机器学习平台魔力象限系列报告》为线索，分析调研了35个数据科学平台产品，给出了数据科学平台的定义和类型；在此基础上，提出了数据科学平台中的基本科学问题、主要特征、关键技术以及发展趋势；最后，提炼出了现阶段数据科学平台研发的指导原则、所面临的理论瓶颈及几点研究建议。

2 数据科学平台及其发展现状

2.1 数据科学平台的内涵

目前，数据科学平台的定义方法有两种。

(1)专门平台。将数据科学平台作为一个独立的专门平台进行定义，认为数据科学平台是支持数据科学项目生命期中绝大部分活动的工具平台。例如，Dataiku将数据科学平台定义为："数据科学平台是数据科学项目全生命期发生的结构，包含完成数据科学项目生命周期的每个阶段所需的工具和资源，汇集从开发到部署的整个数据科学生命周期中使用的人员、工具、资源以及其他必要产品。"[1]此类定义方法主要关注的是面向数据科学用户的数据科学工具平台。

(2)集成平台。将数据科学平台作为其他平台，尤其是机器学习和人工智能平台的重要组成部分的定义方法。例如，Gartner报告中将数据科学和机器学习平台集成在一起讨论，并称之为数据科学与机器学习(Data Science and Machine Learning，DSML)平台。DSML平台是核心产品及其一致集成的辅助产品、组件、库和框架(包括专有、合作伙伴来源和开源)的组合。此类平台的主要用户是数据科学专业人员，包括专业级数据科学家、非专业级数据科学家、数据工程师、应用程序开发人员和机器学习专家[2]。

本文研究中将数据科学平台定义为：从数据科学视角看，能够支持数据科学流水线的绝大部分活动的工具平台，其存在形式可以是面向数据科学家的专门性独立平台，也可以是面向包括数据科学家在内的多种数据相关工作岗位的通用性集成平台。

2.2 数据科学平台的类型

数据科学平台可以分为开源或商业平台、专业级或非专业级平台，以及企业/大规模团队级或个人/小规模团队级平台等多种类型，如表1所列。

(1)从开发与维护策略看，数据科学平台有开源产品与商业产品两种，甚至有些数据科学产品提供了开源和商业两种不同版本。例如，KNIME的平台分为开源KNIME Analytics Platform和商业KNIME Server两种版本，后者基于前者提供了更多增强或增值服务，如对数据科学流程的自动化。目前，开源技术已成为数据科学平台领域研发的主流策略，而基于开源开发策略的商业化运营成为数据科学平台的未来发展趋势之一。

(2)从目标用户定位看，数据科学平台可以分为面向专业级(expert)用户的产品和面向非专业级(citizen)用户的产品，甚至有些数据科学平台的功能分为专业级和非专业级。例如Microsoft的核心产品Azure ML为专业级数据科学家供了灵活的notebook和SDK选项，为非专业级数据科学家提供了增强机器学习和拖拽式应用。虽然非专业级用户是现阶段数据科学平台的主要关注点，但是数据科学平台功能的分层将成为数据科学平台未来发展的趋势，多数产品将会同时包括专业级或非专业级功能，并采取不同的价格和推广策略。

(3)从目标用户规模看，数据科学平台可以分为企业/大规模团队级(enterprise-grade)和个人/小规模团队级别的应用。例如SAS的核心产品Visual Data Mining and Machine Learning(VDMML)提供企业级的平台能力和支持，Dataiku的核心产品数据科学工作台(Data Science Studio，DSS)对小规模团体有不同的版本与定价方式。相对于个人/小规模团队级别，企业/大规模团队级别的平台是数据科学平台的研发难点。

表1 数据科学平台的分类

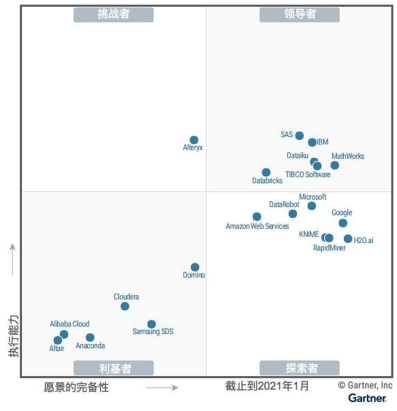
Table 1 Classification of data science platform

	专业级(expert)	非专业级(citizen)
开源平台	1. Altair 2. Amazon Web Services 3. Anaconda 4. Cloudera 5. Databricks 6. Domino 7. Google 8. TIBCO Software	1. H2O. ai 2. IBM 3. KNIME 4. Microsoft 5. RapidMiner 6. Samsung SDS 7. SAS
商业平台	1. 阿里云	1. Alteryx 2. Dataiku* 3. DataRobot

注：带有*的平台提供个人/小规模团队级别的应用

2.3 数据科学平台的评价

2021年Gartner以数据科学与机器学习平台、收入和增长、客户数目、市场牵引力以及产品性能评分5个选择标准，确定20个平台供应商进入魔术象限进行评价。2021年Gartner数据科学及机器学习平台魔力象限(Magic Quadrant of Data Science and Machine Learning Platforms)如图1所示，其横坐标和纵坐标分别为愿景的完备性(completeness of vison)和执行能力(ability to execute)。



来源：Gartner 2021年Magic Quadrant for Data Science and Machine Learning Platforms报告

图1 2021年数据科学及机器学习平台的魔术象限

Fig. 1 Magic Quadrant of Data Science and Machine Learning Platforms in 2021

该魔术象限将数据科学平台分为领导者(leaders)、挑战

者(challengers)、探索者(visionaries)和利基者(niche players)4个象限，如表2所列。

表2 2015－2021年的数据科学平台

Table 2 2015－2021 data science platforms

年份	报告名称	领导者(leaders)	挑战者(challengers)	探索者(visionaries)	利基者(niche players)
2021	Magic Quadrant for Data Science and Machine Learning Platforms	1. SAS 2. IBM 3. Dataiku 4. MathWorks 5. TIBCO Software 6. Databricks	1. Alteryx	1. Microsoft 2. DataRobot 3. Google 4. Amazon Web Services 5. KNIME 6. RapidMiner 7. H2O. ai	1. Domino 2. Cloudera 3. Samsung SDS 4. 阿里云 5. Anaconda 6. Altair
2020	Magic Quadrant for Data Science and Machine Learning Platforms	1. SAS 2. Alteryx 3. Dataiku 4. MathWorks 5. TIBCO Software 6. Databricks	1. IBM	1. Microsoft 2. DataRobot 3. Google 4. Domino 5. KNIME 6. RapidMiner 7. H2O. ai	1. Anaconda 2. Altair
2019	Magic Quadrant for Data Science and Machine Learning Platforms	1. SAS 2. RapidMiner 3. KNIME 4. TIBCO Software	1. Dataiku 2. Alteryx	1. Microsoft 2. DataRobot 3. Google 4. IBM 5. Databricks 6. H2O. ai 7. MathWorks	1. Anaconda 2. SAP 3. Datawatch 4. Domino
2018	Magic Quadrant for Data Science and Machine Learning Platforms	1. Alteryx 2. SAS 3. RapidMiner 4. KNIME 5. H2O. ai	1. TIBCO Software 2. MathWorks	1. Domino 2. IBM 3. Microsoft 4. Databricks 5. Dataiku	1. Anaconda 2. SAP 3. Angoss 4. Teradata
2017	Magic Quadrant for Data Science Platforms	1. IBM 2. KNIME 3. RapidMiner 4. SAS	1. Alteryx 2. Angoss 3. MathWorks 4. Quest	1. Alpine Data 2. Dataiku 3. Domino Data Lab 4. H2O. ai 5. Microsoft	1. FICO 2. SAP 3. Teradata
2016	Magic Quadrant for Advanced Analytics Platforms	1. Dell 2. IBM 3. KNIME 4. RapidMiner 5. SAS	1. Angoss 2. SAP	1. Alpine Data 2. Alteryx 3. Microsoft 4. Predixion Software	1. Accenture 2. FICO 3. Lavastorm 4. Megaputer 5. Prognoz
2015	Magic Quadrant for Advanced Analytics Platforms	1. IBM 2. KNIME 3. RapidMiner 4. SAS	1. Dell 2. SAP	1. Alpine Data Labs 2. Alteryx 3. Miscrosoft	1. Angoss 2. FICO 3. Predixion 4. Prognoz 5. Revolution Analytics 6. Salford Systems 7. Tibco Software

(1)领导者：领导者在DSML市场上占据最有利地位，有很强的思想领导力和创新力，提供明确的方向和愿景，拥有较为广泛的客户群体。2021年领导者包括SAS，IBM，Dataiku，MathWorks，TIBCO Software，Databricks。

(2)挑战者：挑战者有着强大的产品能力和长期的客户关系，但需要进一步明确方向与愿景。Alteryx是2021年唯一的挑战者。

(3)探索者：探索者有着强大的愿景和坚实的支持路线图，但在产品的完整性和广度的提供能力方面仍存在差距。2021年的探索者包括Microsoft，DataRobot，Google，Amazon Web Services，KNIME，RapidMiner及H2O. ai。

(4)利基者：利基者在特定行业和细分领域拥有优势，需要增强市场执行能力，有着一定程度的远见与愿景。2021年的利基者包括Domino、Cloudera、Samsung SDS、阿里云、Anaconda、Altair。

2.4 数据科学平台相关学术研究中的科学问题

目前，数据科学平台相关学术研究主要集中在以下6个议题。

(1)数据科学平台的设计。领域无关的通用数据科学平台与领域相关的专用数据科学平台的差异化设计是数据科学平台学术研究的重要议题。目前，面向医疗[3]、材料[4]、智慧城市[5]及教育[6]等领域的数据科学平台研发成为热点问题。

(2)数据科学平台的可扩展性(scalability)。可扩展性是数据科学平台的关键技术指标之一，代表着数据科学平台对数据规模、计算资源、模型训练及调参、负载均衡等方面的弹性计算能力。数据科学平台的可扩展性需要满足适应大规模

实时数据集的分析处理需求，如创建集成的数据湖等[7]。建立集成和可扩展的数据科学平台，也有利于促进数据分析的可重复性[8]。

(3)基于数据湖的数据科学平台研发。数据湖是数据科学平台需要重视和引入的新技术。数据湖为数据科学平台提供数据存储层[9]。数据科学家需要具备数据湖的构建能力，数据湖团队中数据科学家的缺乏将影响数据存储的商业价值[10]。相对于数据库和数据仓库，数据湖将会是数据科学平台的主要数据存储技术。

(4)数据科学平台的支持团队协作能力。支持团队协作是数据科学平台，尤其是企业/大规模团队级别数据科学平台的重要属性。基于云的解决方案可以支持协作数据科学平台[11]。构建协作型数据科学云平台，支持领域专家、数据科学家和其他用户共享数据集，进行数据分析[12]。

(5)数据科学平台的开放策略。开放策略正迅速成为数据科学平台引入新功能的主流策略。开放为可重用性的核心，开源语言(如 Python 等)为实现可重用的数据分析和可视化的数据科学平台提供了基础[13]。

(6)数据科学平台的工程方法论。工程方法论的引入是确保数据科学平台研发工作的成熟度和产品质量的重要保障。因此，数据科学平台的工程化开发成为相关研究的一个热点问题，如对 Anaconda[14]，Alibaba[15] 等数据科学平台的研究。

此外，数据科学平台中的数据可视化、数据加工、知识图谱构建、大数据分析以及非专业级数据科学平台等也已成为学术界研究的热点。

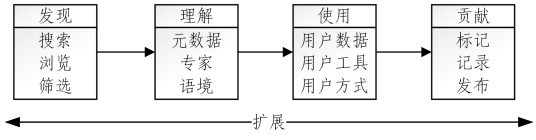
3 数据科学平台的特征

相对于其他软件平台，数据科学平台的特征主要包括 6 个方面。

3.1 模块化开发及集成能力

数据科学的流水线(life cycle)主要包括的活动有：问题和业务上下文理解、数据摄取、数据准备、数据探索、特征工程、模型创建和训练、模型测试、部署、监察、维修保养、数据和模型治理、可解释人工智能、业务价值跟踪、团队合作。从目前的数据科学平台看，目前主要重视的活动如下。

(1)数据目录(data catalogs)。数据目录是数据科学平台的关键功能之一。Microsoft 的数据目录产品 Azure Data Catalog(见图 2)主要提供基于云的服务，使管理数据的用户容易发现、理解和使用数据源，用户还可以通过标记、记录和注释已注册的数据源为数据目录提供支持[16]。



来源：Microsoft 官网

图 2 Microsoft Azure 数据目录

Fig. 2 Microsoft Azure data catalog

(2)数据世系(data lineage)。数据世系描述了数据来自何处，如何获得以及如何随时间更新[17]，包括其来源(如基本数据集、记录仪器和仪器的操作参数)以及所有后续的应用于它的处理步骤(算法和相应的参数)[18]。数据世系是数据科学平台中数据治理的主要任务。

(3)数据加工(data wrangling/munging)。数据加工是数据科学的主要研究内容之一。数据加工作为数据分析比较耗时的部分[19]，是识别、提取、清理和集成应用程序所需的数据以生成适合于探索和分析的数据集的过程[20]。与传统数据处理不同的是，数据科学中的数据加工更强调数据处理中的增值过程[21]。数据加工作为数据科学平台的标志性功能，已成为 TIBCO Spotfire[22] 等数据科学产品主要重视的活动。

(4)数据编排(data orchestration)。数据编排是数据科学的另一个标志性活动。以数据为中心的云编排方法，可以使云资源被建模为可通过声明性语言查询的结构化数据，并使用定义明确的事务语义进行更新[23]。在基于服务的计算网络模型中，数据编排的两个功能为通过聚合数据减少数据流以及将数据转换为服务[24]。例如 KNIME 商业服务器促进了端到端分析过程的编排，其模型工厂中存在工作流编排[25]。

(5)数据刷新(data refresh)。数据刷新模块也是数据科学平台的重要组成部分，涉及数据源、数据集、数据状态和分析结果的刷新 4 种行为。例如 Microsoft Power BI[26] 采用数据刷新功能，来保证其报表和仪表板的数据最新，将数据转化为洞察力和行动[27]。

上述活动的模块化开发要求数据科学平台重视一致性(cohesion/coherence)的问题。模块化和一致性是数据科学平台研发的两个重要问题，模块化开发应以一致性为基础，确保不同功能模块的无缝集成。

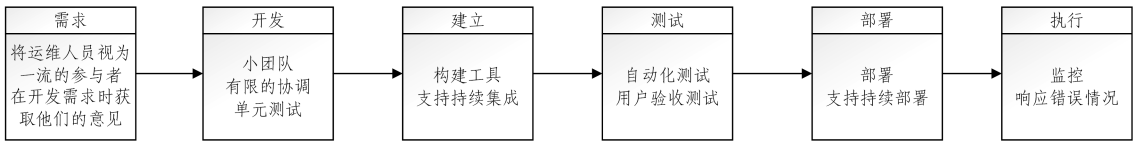
3.2 开发运维一体化

与传统软件开发不同的是，数据科学平台研发活动强调开发运维的一体化(DevOps)，需要将系统的开发、部署、运营和维护进行统筹。数据科学平台的开发运维一体化可以进一步分为以下 4 个部分。

(1)机器学习运维(MLOps)。MLOps 是机器学习的持续交付实践，虽模仿 DevOps 实践，但引入了特定于机器学习的具体活动[28]。MLOps 是 Domino 的数据科学平台中的核心功能，致力于通过部署和管理模型来推动业务发展[29]，2020 年发布的 Domino Model Monitor[30] 可进一步增强其 MLOps 能力。

(2)数据运维(DataOps)。DataOps 将对数据的集成和面向过程的观点与敏捷软件工程的自动化和方法相结合，以提高质量、速度和协作[31]。通常，数据科学平台在开发过程中需要关注 DataOps，例如 Cloudera 的共享数据体验(SDX)框架对用于 DataOps 和 MLOps 的元数据进行管理。

(3)软件开发运维的一体化(DevOps)。数据科学平台中特别强调软件开发运维的一体化。DevOps 有效地集成了开发、交付和运营。DevOps 的流程包括建立、测试和部署等，如图 3 所示[32]，Microsoft Azure DevOps 就是 Microsoft 数据科学平台的服务之一[33]。



来源：DevOps：A software architect's perspective

图 3 DevOps 生命期流程[表示法：波特的价值链]

Fig. 3 DevOps life cycle[Notation: Porter's value chain]

(4)人工智能运维(AIOps)。AIOps 旨在通过人工智能(AI)和机器学习(ML)技术授权软件和服务工程师(如开发人员、程序经理、支持工程师、站点可靠性工程师)高效地构建和操作大规模的在线服务和应用程序[34]。AIOps 目前已在在数据科学平台中应用，例如 IBM Cloud Pak 就是一种 AIOps 解决方案[35]。AIOps 支持 DevOps 的可视性并提供了自动化 IT 支持。

3.3 重视可扩展性

可扩展性(scalability)是数据科学平台的重要技术特征。数据科学平台重视可扩展性，促进平台高效运行，提升服务质量[36]。目前，云计算技术的引入成为保障数据科学平台可扩展性的重要手段。例如，Microsoft 采取的是云优先战略(cloud-first approach)运营战略，在云中存储资源。从数据科学平台的部署看，数据科学平台分为本地或私有云(on-premise)、云端或多云(multiload support)以及混合云(hybrid cloud)3 种部署方法。在混合云类解决方案中，应用程序系统的部分保留在本地，而其余部分则迁移至云端，从而保持有效的职责划分和数据流[37]。

数据科学平台重视可扩展性，根据用户的本地、混合和多云部署的需求提供不同解决方案。例如，Domino 的核心产品 Data Science Platform 使用 Kubernetes，支持复杂的混合云和多云模型的开发和部署。

3.4 强调用户体验

用户体验是多方交互的结果，即用户的内在状态(倾向、期望、需求、动机、心情等)、系统的设计特征(复杂性、目的、可用性、功能等)及交互发生环境(如组织/社会环境、活动意义、自愿使用等)共同作用的结果[38]。目前，强调用户体验是当前数据科学平台的重要特征。数据科学平台的用户体验主要表现在以下 6 个方面。

(1)易用性与学习成本低。学习曲线可以较好地刻画学习成本。易用性与学习成本低是数据科学平台的主要优势。例如，Dataiku 的主要优势是易于学习、学习曲线短，而 Google 的陡峭学习曲线提供了用户的学习成本。

(2)对用户多元化个性特征(diverse persona)的尊重。数据科学平台尊重不同文化、信仰和个性。例如，Alteryx 为不同地区的客户提供多语种支持，提供业务线(Line-of-Business，LOB)和行业解决方案模板以及快速入门套件，利用无代码和专家模式的协作，方便所有用户的使用。

(3)支持团队协作(collaborative working)。数据科学平台支持不同角色、活动和时空的合作。例如，Azure Databricks 提供交互式的协作工作区，支持数据科学家、数据工程师和业务分析师之间的协作。

(4)对稳定性(stability)的重视。稳定性是数据科学平台成功的关键因素之一。例如，SAS 的优势在于其产品的质量、稳定性和可靠性高。数据科学平台要长期发展，应加强对稳定性的重视。

(5)提供咨询和管理服务。目前，在数据科学平台开发的基础上，也出现了一些关于数据科学平台的开发、运维和利用的咨询与管理服务。例如，Altair 提供各种咨询和管理服务，以支持建立和部署模型。

(6)提供配套的社区(community)支持。数据科学平台提供配套的社区支持，可以促进协作与知识共享。例如，阿里云提供了开发者社区，方便开发者之间共享经验和交流合作。

3.5 重视非专业级数据科学家

相对于专业级数据科学家，非专业级数据科学家(citizen data scientist)一般并不具备数据科学类学科(如数据科学与大数据技术、计算机科学与技术、统计学等)的专业背景，编写代码能力较弱，不参与数据科学的全流水线活动，而是利用自己在某一应用领域的知识和经验优势，主要借助数据科学工具的方式完成数据科学流水线的某一或少数活动，如表 3 所列。近年来，非专业级数据科学家成为数据科学平台开发的主要目标用户之一。面向非专业级数据科学家的平台主要有如下 4 个特点。

(1)对拖拽式应用，尤其是 VizQL 技术的重视。拖拽式应用一般不需要写代码或编写代码量非常少。VizQL 语言描述了视图的结构以及使用数据填充该结构的查询[39]，VizQL 代数的关键技术优势在于清楚地描述了小型数据多个视图的行和列结构[40]。SAS VDMML 为非专业级数据科学家提供了拖拽式应用。

(2)对自动化处理，尤其是自动化机器学习(AutoML)的支持。在机器学习流水线上，能够减少数据预处理、模型选择、超参数优化和模型解释等流程上的人工投入和精力的研究领域称为自动化机器学习[41]。例如，Altair Knowledge Studio 提供了 AutoML 功能。

(3)支持领域应用，尤其是基于数据科学的应用领域创新。数据科学平台例如 DataRobot 在银行、保险、金融服务、制造、零售、生命科学和医疗保健等领域拥有相关业务。

(4)对端到端(end-to-end)的应用的重视。相对于专业级数据科学平台，非专业级数据科学平台提供端到端的 ModelOps 功能(End-to-end ModelOps Capabilities)。ModelOps 是机器模型和 DevOps 的集成，是一种用于 AI 应用程序工件的端到端生命期管理的新框架和平台[42]。例如，TIBCO 的端到端的 ModelOps 功能中包括 TIBCO 工件管理服务器的功能，以及对其 ML 流水线功能的改进。

表 3 专业级与非专业级数据科学家的区别

Table 3 Difference between expert and citizen data scientists

	非专业级数据科学家 (Citizen data scientist)	专业级数据科学家 (Expert data scientist)
专业背景	不具备数据科学类学科的专业背景	具备数据科学类学科的专业背景
编写代码能力	较弱	较强
知识结构	应用领域知识>数据科学知识	应用领域知识<数据科学知识
岗位职责	数据科学流水线的某一或少数活动	数据科学流水线的全部或多数活动
数据科学平台	上层应用为主	底层研发为主

3.6 重视人机协同场景

重视人机协同场景是数据科学平台需要重视的重要产品属性。增强分析(augmented analytics)与增强人工智能(augmented AI)是人机协同场景的主要趋势。增强分析指在整个分析周期中应用人工智能[43];增强人工智能是将人的认知能力或人的认知模型引入 AI 系统[44]。DataRobot 将增强分析的功能纳入数据科学平台中,支持开发人员、数据科学家、统计人员和业务分析师之间的协作。因此,数据科学平台应增强人的作用,重视人机协同场景。

4 数据科学平台的关键技术

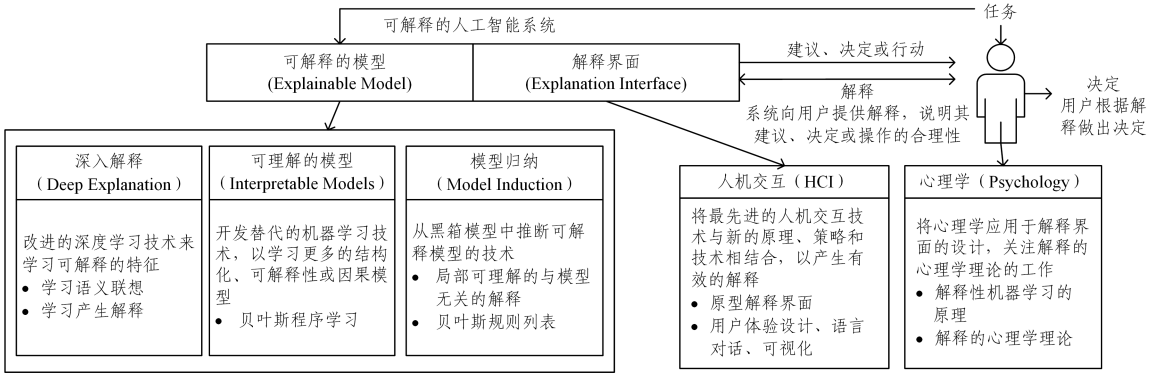
数据科学平台的实现需要的关键技术为机器学习、流处理技术、数据规整化、容器化技术和数据可视化。

4.1 机器学习

目前,机器学习是数据科学平台的内核,需要突破的关键技术如下。

(1)模型训练与再训练(model training and retraining)。模型训练与再训练是模型治理中的技术[45],也是面向数据科学的机器学习活动的关键所在。RapidMiner 的模型治理模块包括模型训练、模型使用、模型再训练等功能。

(2)可解释机器学习及可理解性人工智能(XAI)。可理解性人工智能系统由可解释的模型和解释界面两部分组成,如图 4 所示。可理解性人工智能系统向用户提供解释,说明其建议、决定或操作的合理性,用户根据解释做出决策[46]。H2O.ai 支持 XAI 功能,针对机器学习的整个流水线提供解释功能。



来源:DARPA(2017)的 Explainable artificial intelligence (xai)

图 4 可理解性人工智能

Fig. 4 Explainable AI

(3)深度学习(deep learning)。通常,深度学习是一种端到端的学习方法,即特征不是由人给出的,而是通过自动学习得到输入数据的方式[47],允许由多个处理层组成的计算模型学习具有多个抽象级别的数据表示[48]。Amazon Web Services 平台支持机器学习和深度学习框架优化选项。深度学习作为人工智能框架中的功能,是现阶段数据科学平台的热门技术之一,尤其是在数据科学平台中的自然语言处理、图像处理等领域得到广泛应用。

(4)高性能机器学习(high performance ML)。目前数据科学平台普遍重视高性能机器学习,例如 H2O.ai 提供高性能机器学习组件,其开源机器学习组件为行业建立标准,并与其他平台集成在一起,该组件针对 CPU 多核和多节点配置进行优化和并行处理。

(5)对抗性机器学习(adversarial machine learning)。对抗性机器学习是数据科学的新兴研究领域。对抗性机器学习指针对对抗性对手的有效机器学习技术的研究。目前,针对机器学习系统的攻击主要包括影响(因果关系攻击、探索性攻击)、违反安全规定(完整性攻击、可用性攻击、隐私攻击)、特异性(有针对性攻击、普遍攻击)等[49]。

4.2 流处理技术

流处理技术是数据科学平台的关键技术之一,主要解决的是数据的实时采集和分析处理活动。目前,Spark 和 Storm 等流处理技术在数据科学平台中被广泛采用[50]。流处理技术需要解决的问题在于其输入数据必须在不完全存储的情况下进行处理[51]。

流处理技术的引入对实时决策至关重要。例如 Databricks 数据科学平台的主要竞争力源自基于 Spark 的流处理能力。

4.3 数据规整化

数据规整化是将数据集的含义映射到其结构的标准方法。规整数据遵循 3 个基本原则:每个观察占且仅占一行,每个变量占且仅占一列,每一类观察单元构成一个关系表[52]。

数据科学主要从数据形态视角关注质量问题,重视的是数据是为规整数据(tidy data)还是混乱数据(messy data)。因为混乱数据的结构无法分析,在进行数据分析时,通常需要将混乱数据规整化为规整数据。

数据规整化是数据科学平台的关键技术之一,例如 SAS communities Library 提出了 3 种常见的混乱数据问题以及如

何在 SAS 中进行数据规整[53]。

4.4 容器化技术

Docker 和 Kubernetes 的容器化(containerization with Docker & Kubernetes)是目前数据科学平台领域普遍采用的关键技术。

容器化技术是数据科学平台的关键技术。Kubernetes 和 Docker 技术可以互补,Kubernetes 提供 Docker 容器的编排、调度并自动部署它们跨越 IT 环境,以确保高可用性。目前数据科学平台重视容器化技术的应用,例如 IBM Cloud Kubernetes 服务和 RapidMiner 使用 Docker 和 Kubernetes 进行容器化,透明地运行和扩展模型。

4.5 数据可视化

可视化是数据科学不可或缺的一部分,对于实现数据的探索性分析至关重要[54]。数据可视化能够简化和转换复杂的信息和数据的细节,进行数据可视化前需要进行数据收集、数据理解、数据过滤、数据挖掘、数据表示的数据预处理。目前使用较多的数据可视化工具包括 Tableau,Zoho Reports,Infogram 等。Python,R 与 JavaScript 也提供了用于数据可视化的工具和包,如 Python 中的 Pandas,Matplotlib 和 Seaborn,R 中的 ggplot2 和 Lattice 等[55]。

目前,数据科学平台所支持的数据可视化较为丰富。例如,KNIME Analytics Platform 支持视觉工作流程的连续性,构建了包括自动化机器学习、数据可视化、交互式应用程序和部署模型在内的平台。数据可视化功能的优劣正成为数据科学平台是否具有竞争力的一个重要标志。

5 数据科学平台的发展趋势

数据科学平台的未来发展趋势主要包括以下 8 个方面。

5.1 与人工智能的融合

目前,数据科学平台的发展呈现出与机器学习和人工智能高度融合的趋势,具体包括:

(1)Gartner 的数据科学与机器学习魔力象限系列报告。2014—2016 年,该报告的名称为《高级分析平台的魔力象限》(Magic Quadrant for Advanced Analytics Platforms);2017 年,该报告名称被调整为《数据科学平台的魔力象限》(Magic Quadrant for Data Science Platforms);2018—2021 年,该报告名称修改为《数据科学与机器学习平台的魔力象限》(Magic Quadrant for Data Science and Machine Learning Platforms)。从报告名称的变化可以看出数据科学平台与机器学习的融合式发展趋势。

(2)可理解性人工智能。目前,机器学习模型及数据分析结果的可解释性已成为数据科学及相关领域的主要关注点之一。可理解性人工智能的出现为解决数据科学中的可解释性问题提供了新思路。因此,可理解性人工智能的研究也受到数据科学领域的高度关注,RapidMiner,H2O.ai,Google 等数据科学平台都关注可理解性人工智能。

(3)对抗性机器学习。对抗性机器学习是未来的发展趋势,数据科学平台接受对抗性机器学习,保护平台免受对抗性攻击[56]。目前,IBM[57],Microsoft[58]等数据科学平台加强了对对抗性机器学习的重视。

(4)复合人工智能(composite AI)在数据科学中的应用。复合人工智能是将不同 AI 技术结合起来以达到最佳效果的方法[59]。目前,数据科学平台开始关注复合人工智能。例如 IBM Watson Studio 平台的复合人工智能的愿景、MathWorks 的主要产品 MATLAB 的复合人工智能的能力。

5.2 对开源技术的支持

开源技术正成为数据科学平台研发的主要趋势,许多数据科学平台采用开源技术推动创新。具体表现有 3 种。

(1)开源软件模式建设和维护数据科学平台。例如,以 R 和 Python 语言为基础的数据科学平台通常采用开源软件模式的建设和维护策略。

(2)基于开源软件开发数据科学平台。例如,H2O.ai 通过 Wave(一种用于构建 AI 应用程序的开源产品)扩展其产品功能。

(3)支持调用开源工具包。例如,Anaconda 提供了第三方开源工具包的扩展接口,进而实现数据科学平台的可扩展性。

5.3 对非专业级数据科学家的重视

非专业级数据科学家正成为各数据科学平台争夺的目标市场用户。Gartner 预测,非专业级数据科学家产生的高级分析和商业价值的数量将超过数据科学家[60]。目前,在数据科学平台的研发中重视非专业级数据科学家的主要动因在于:

(1)使用数据科学平台的非专业级数据科学家数量显著增加。越来越多的非专业级数据科学家正在构建 DSML 模型,生成使用高级诊断分析或预测和说明功能的模型。

(2)面向专业级数据科学家的供应商也正在调整产品策略,以吸引非专业级数据科学家。例如 SAP,RapidMiner 等针对非专业级数据科学家提供相应产品,如拖拽式应用、自动化处理等。

(3)部分数据科学平台由于缺乏对非专业级数据科学家的支持而在数据科学平台评估中表现不佳。例如,Anaconda 的目标受众是专业级数据科学家,缺乏针对非专业级数据科学家的功能设计,导致其整体评价受到影响。

5.4 数据治理的集成

数据治理成为数据科学平台功能设计的重要趋势之一。具体包括:

(1)模型治理(model governance)。模型治理主要涉及模型设计、实现、部署、运维、验证和调参等多个活动。作为数据治理的重要组成部分,模型治理也成为数据科学平台需要解决的必要功能之一。例如,DataRobot 提供了模型的验证、版本控制、访问权限设置等模型治理功能。

(2)人工智能治理(AI governance)。人工智能治理研究人类如何最好地向先进人工智能系统过渡,重点关注政治、经济、军事、治理和伦理层面[61]。可见,相对于模型治理,人工智能治理的涉及面广,所讨论的问题并不仅限于模型或技术层面。

(3)负责任的人工智能(responsible AI)。负责任的人工智能指在真实组织中以公平、模型可解释性和问责制为核心的大规模实施人工智能方法的方法[62],是数据科学平台未来发展的重要趋势。目前,Google,IBM,KNIME 等数据科学平

台在人工智能解释能力和责任领域占据了领导地位，为数据科学平台研发提供了新的指导原则。

(4)访问治理(access governance)。访问治理是一个包括策略、控制、激励措施和管理用户对信息资源访问流程的集成系统，其目标是确保信息系统在正确的时间向正确的人提供正确的信息，同时保护信息不被滥用[63]。目前访问治理已经引起数据科学平台如 H2O. ai 等的关注，将成为数据科学平台研发中不可忽略的细节性问题。

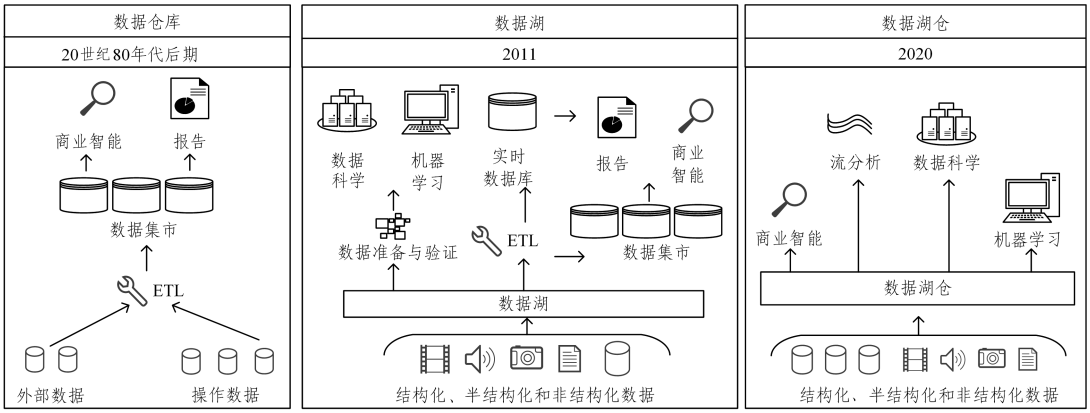
(5)可信赖与弹性的平台(trusted and flexible platform)或可验证且可靠的机器学习(verifiable and reliable ML)。通常，支持经过身份验证操作的硬件软件平台称为受信任的开放系统[64]。目前可信任的平台是未来发展的趋势之一。例如，Anaconda 在编码社区中提供了一个灵活和可信任的平台，为初学者和专家提供了多种选择；MathWorks 提供可验证和可靠的机器学习。

(6)人工智能向善(AI for good)。人工智能向善是在更大的人工智能领域内开展的一项运动，旨在发展和使用人工智能方法，以进一步朝着可持续性、健康、人道主义援助和社会正义的目标迈进[65]。未来数据科学平台在开发上会更加关注人工智能向善，例如 H2O. ai 人工智能向善的愿景。

5.5 数据湖的引入

数据湖正成为数据科学平台的新的基础设施。例如，Databricks Delta Lake 是一个统一的数据管理系统，支持数据湖的生命期管理，加快高质量数据进入数据湖的速度。数据湖解决了数据格式错误、数据删除、合规性删除或修改数据以捕获数据等问题。

图 5 为 Databricks 的数据管理架构——数据湖仓(Lakehouse)的示意图[66]，它结合了数据湖和数据仓库，在数据湖的低成本存储上实现与数据仓库中类似的数据结构和数据管理功能，支持从非结构化数据到结构化数据的多种数据类型的统一管理。



来源：The Delta Lake Series Lakehouse

图 5 Databricks Delta Lake 数据管理架构

Fig. 5 Databricks Delta Lake data management architecture

5.6 高级分析及应用的探索

(1)诊断性分析和规范性分析。相对于描述性分析和预测性分析而言，诊断性分析和规范性分析是数据分析的高级应用。对高级应用的支持是数据科学平台的竞争力所在。例如 IBM 的数据科学平台中集成了预测性和规范性的能力，对于数据科学平台的研发具有一定的借鉴意义。

(2)图分析(graph analytics)。通过图分析和可视化，可以方便用户浏览和可视化数据[67]。例如 Dataiku 改进了图分析和时间序列分析，提供高级分析。图分析已经成为数据科学平台高级应用探索的方向之一。

(3)时空数据分析(spatiotemporal data analysis)。地理空间分析往往与时间数据分析联系在一起，形成时空数据分析，并将成为数据科学平台的一个重要组成部分。

(4)离散事件仿真或基于主体的仿真(discrete event or agent-based simulation)。仿真指在计算机上对现实的假设情况进行建模的过程，以便对其进行研究以了解系统的工作原理[68]。离散事件仿真是一种将系统操作建模为时间上离散事件序列的仿真技术[69]。基于主体的仿真是一种对由个体、自治的、交互的“主体”组成的系统进行建模的方法[70]。由于模拟和仿真是诊断性分析和规范性分析的主要技术手段，数据科学平台需要加强在离散事件仿真或基于主体的仿真等高级应用方面的探索。

(5)优化实验与决策管理的设计(design of optimization experiments & decision management)。优化实验和决策管理的设计是数据科学平台的未来发展趋势之一。目前，数据科学平台需要在优化实验和决策管理的设计等高级应用方面加强探索。

5.7 向数据科学全流水线的转型

目前，数据科学平台支持数据科学生命期中的多项任务，已经从数据科学的单一活动向全流水线发展。

(1)Alteryx：于 2016 年通过提供数据准备和高级分析来补充不断增长的数据发现市场，到 2017 年提供基于云的分析库，用于工作流(workflow)的协作、共享和版本控制。

(2)RapidMiner：支持端到端数据科学功能，完成数据科学项目从创建到模型构建再到生产的全流水线管理。

(3)SAS：加大对全流水线的支持。SAS 提供从数据获取到模型部署的系列功能，并推出称为“统一洞察(unified insights)”的生命期产品，以降低许可的复杂性，提供从探索到建模和部署的全生命期支持。

(4)Databricks：提供从数据工程(data engineering)到端

到端分析全生命期功能、混合云环境以及对各种用户的可访问性的支持。

除了上述数据科学平台,像阿里云、Oracle、Samsung SDS等大多数数据科学平台的关注均从数据科学的某个活动转向全流水线。

5.8 应用领域的多样化

目前,数据科学平台主要应用于生命科学、生产制造、电子商务、银行、保险和其他金融服务、医疗保健、教育和政府、通讯、媒体和服务等领域。

(1)DataRobot:为银行、保险、其他金融服务、制造、零售、生命科学和医疗保健等多个领域提供应用服务。

(2)Domino:在银行、金融服务、制造业和生命科学领域有自己的业务。

(3)Altair:在银行和其他金融服务领域有自身的业务,提供各种模拟和高性能计算解决方案,以吸引汽车、航空航天和制造部门的客户。

(4)RapidMiner:在制造业、生命科学、银行业、保险业、能源、商业服务、政府及教育等领域得到较多的应用。

目前,数据科学平台的跨行业客户数量逐渐增多,数据科学平台的应用领域越来越广是数据科学平台发展的趋势之一。

6 讨论与结论

数据科学是一个实践领先于理论研究的领域,通过对数据科学平台进行特征、技术以及发展趋势等的研究,得出数据科学平台的指导原则,研究挑战及对数据科学理论研究的建议。

6.1 数据科学平台的指导原则

通过对数据科学平台的调查分析,可以看出数据科学平台研发的指导原则如下:

(1)以激活数据价值为中心。以激活数据价值为中心是数据科学平台的指导原则之一。随着工业互联网的发展,数据转化为洞察力和行动,从分析中获得战略洞察力是运营成功的关键[71]。例如,Databricks致力于为客户创造价值,H2O.ai有着创造价值的愿景,Alteryx也强调分析内容的创造和从洞察到行动的发展。数据科学平台有着明确的愿景,以价值创造为中心,追求愿景的完备性。

(2)人在环路(human-in-the loop)的设计模式。强调增强人工智能在人在环路中需要发挥循环中人的作用[72]。例如Amazon重视人在环路的能力,其增强人工智能(Amazon A2I)辅助构建工作流,以人工审核已部署的模型。

(3)开发运维一体化。目前大部分数据科学平台都对开发运维一体化给予较高的关注,包括MLOps,DataOps,AIOps,DevOps等。例如,Dataiku有着统一XOps的愿景,开发运维一体化是数据科学平台开发的指导原则之一。

(4)可用性和可解释性的平衡。可用性和可解释性之间的矛盾是现阶段数据科学平台的主要矛盾之一。Altair,Google,H2O.ai,RapidMiner等数据科学平台都增强了可理解性人工智能方面的功能。数据科学平台需要注意可用性和可解释性的平衡。

(5)数据科学产品生态系统的培育。数据科学产品生态系统的培育是数据科学平台研发和运维的最终目的。例如,Dataiku扩大其联盟、合作伙伴和经销商的生态系统——Dataiku生态合作者系统(Dataiku partner ecosystem);Samsung SDS有着综合生态系统愿景,提供整体解决方案,将Brightics AI与其他三星SDS产品相互补充。

(6)强调用户体验和易用性。用户体验是数据科学平台研发活动成功与否的重要标志之一。目前,数据科学平台从支持团队协作、重视稳定性、关注易用性与学习成本等多方面强调用户体验和易用性。

(7)与其他业务系统的集成。Cloudera,Databricks,TIBCO Software等数据科学平台的客户横跨多个行业和不同的业务功能。数据科学平台的应用领域越来越多,与领域业务系统集成是数据科学平台的重要指导原则。

6.2 数据科学平台的研究挑战

未来数据科学平台的发展将面临着以下的研究挑战:

(1)可解释性与可理解性。机器学习和人工智能的可解释性和可理解性是数据科学平台的研究挑战之一。虽然当前的机器学习方法具有良好的预测性能,但其有效性将受到机器无法向用户解释其决策和动作的限制[73]。可解释性不仅对证明决策具有重要意义,它还可以防止问题出现,提供对未知漏洞和漏洞的更大可见性,帮助开发者开发更有用的工具[74]。数据科学是可解释性过程中的核心要素,未来数据科学平台的开发需要加大对可解释性与可理解性的关注。

(2)数据实验设计及规范性分析。数据实验设计及规范性分析是数据科学平台的高级功能,目前一些数据科学平台例如IBM等在这方面进行了探索,但大部分数据科学平台如H2O.ai还是存在欠缺。而数据实验设计及规范性分析作为数据科学平台的未来发展趋势之一,在未来开发数据科学平台时需要重点研究。

(3)数据故事化。数据可视化是数据科学平台的关键技术之一,数据故事化可以被看作是数据可视化处理的必要补充[75],数据可视化主要解决的是数据感知问题,而数据故事化更关注的是如何将数据感知转换为数据认知[76]。数据故事化是数据科学的主要研究内容之一,未来数据科学平台应该关注如何将数据故事化融入其中,这是数据科学平台的主要研究挑战。

(4)对非专业级数据科学家的支持。对非专业级数据科学家的重视是数据科学平台的重要特征与未来发展趋势。但目前在实践中许多数据科学平台例如Domino、Anaconda、阿里云等均缺乏对非专业级数据科学家的支持,在未来研究中需要加强对非专业级用户的重视。

(5)学习曲线。数据科学平台在设计时强调用户体验,易用性和学习成本低是吸引用户使用的主要原因。目前数据科学平台的研究挑战之一在于其陡峭的学习曲线,Cloudera,Google,H2O.ai等数据科学平台都存在学习成本高的情况,需要进一步改进,以方便用户使用。

(6)可复现(repeatable/reproducible)。复现是判断科学主张的最终标准[77],是科学方法不可或缺的一部分[78]。可重现性也是数据科学平台的研究挑战之一,是数据科学平台

设计的指导原则之一[79]。目前数据科学平台在建设中也关注可重复性，例如 Cloudera 维护可按需扩展的可重复的集装箱化工作流，支持构建重复 DSML 管道；Microsoft 支持简化创建可复制的机器学习管道等。

6.3 对数据科学理论研究的建议

数据科学平台研发在数据科学理论，尤其是以下几个方面亟待突破性研究。

(1)数据偏见与公平性。目前数据科学平台进行公平性管理，例如 Google 在机器学习生命周期中应用公平分析，来提高机器学习模型的公平性。IBM 为解释能力、偏见、公平性、准确性和监控、合成数据和隐私提供广泛的支持。近年来，在建立和部署机器学习和数据科学系统时，处理偏见和公平性问题受到研究界越来越多的关注，但大多数研究都集中在理论方面，应用领域和数据集非常有限[80]。数据科学平台的理论研究要关注减少偏见与公平度量。

(2)鲁棒性及稳定性。健壮的模型学习的表示形式往往能够与突出的数据特征和人类感知更好地吻合[81]。数据科学平台在实践中注重鲁棒性及稳定性，例如，SAS Enterprise Miner(EM)具有稳健性，从数据提取和准备到模型生产和部署，该平台持续提供可靠的结果。数据科学在发展中应注重鲁棒性及稳定性。

(3)隐私保护。随着数据科学的发展，引发了关于隐私的争议。出于保护隐私的目的，专家和政策制定者制定隐私保护措施，将隐私价值纳入数据科学[82]。数据科学平台例如 Databricks 支持一般数据保护条例(General Data Protection Regulation，GDPR)和加利福尼亚消费者隐私法(California Consumer Privacy Act，CCPA)，并嵌入了减少偏见和解释性的开放源码技术。

(4)因果分析。数据科学来自传统研究领域，必须了解因果分析的基本原理[83]。数据科学平台注重可用性和可解释性的平衡，可理解性人工智能是数据科学生命期的任务之一。数据科学注重可解释性，了解其背后的因果关系，注重因果分析。相关关系可以帮助我们预测未来，而因果关系有助于我们进一步理解和把握未来，在数据科学的发展中应重视因果分析。

(5)信任/负责任数据科学平台。信任(trusted)或负责任(responsible)的数据科学平台是数据科学平台的未来发展趋势之一，数据科学平台在实践中致力于建设可信任的数据科学平台。建设可信赖的数据科学系统是组织和研究人员的首要任务[84]，数据科学的发展中要注重信任。

(6)快速响应能力。数据科学在实践中要有应对极端情况的能力，面对逆境也可以保持创新、持续发展。COVID-19 新冠疫情爆发后，H2O.ai 等利用人工智能与主要的医疗保健组织合作创新，以应对全球新型冠状病毒疾病流行期间的挑战[85]。数据科学家对于像新冠肺炎疫情这样的极端情况需要积极贡献自身的力量[86]，增强应对及处理极端情况的能力。

参 考 文 献

[1] What Is a Data Science Platform ? [EB/OL]. (2021-03-23) [2021-05-22]. https://blog.dataiku.com/what-is-a-data-science-platform.

[2] IDOINE C, KRENSKY P, BRETHENOUX E, et al. Magic Quadrant for data science and machine-learning platforms[R]. Gartner, Inc, 2021.

[3] MARUNGO F, ROBERTSON S, QUON H, et al. Creating a data science platform for developing complication risk models for personalized treatment planning in radiation oncology[C]// 2015 48th Hawaii International Conference on System Sciences. IEEE, 2015: 3132-3140.

[4] WARD L, DUNN A, FAGHANINIA A, et al. Matminer: An open source toolkit for materials data mining[J]. Computational Materials Science, 2018, 152: 60-69.

[5] DOBRE C, XHAFA F. Intelligent services for big data science [J]. Future Generation Computer Systems, 2014, 37: 267-281.

[6] MIAO K, LI J, HONG W, et al. A Microservice-Based Big Data Analysis Platform for Online Educational Applications[J]. Scientific Programming, 2020, 2020: 1-13.

[7] MCPADDEN J, DURANT T J S, BUNCH D R, et al. Health care and precision medicine research: analysis of a scalable data science platform [J]. Journal of Medical Internet Research, 2019, 21(4): e13043.

[8] TOROUS J, KIANG M V, LORME J, et al. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research [J]. JMIR Mental Health, 2016, 3(2): e16.

[9] NARGESIAN F, ZHU E, MILLER R J, et al. Data lake management: challenges and opportunities[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1986-1989.

[10] FANG H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem[C]// 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). IEEE, 2015: 820-824.

[11] ESPOSITO C, CASTIGLIONE A, TUDORICA C A, et al. Security and privacy for cloud-based data management in the health network service chain: a microservice approach[J]. IEEE Communications Magazine, 2017, 55(9): 102-108.

[12] PATTERSON E, MCBURNEY R, SCHMIDT H, et al. Dataflow representation of data analyses: Toward a platform for collaborative data science[J]. IBM Journal of Research and Development, 2017, 61(6): 9: 1-9: 13.

[13] POLDRACK R A, GORGOLEWSKI K J, VAROQUAUX G. Computational and informatic advances for reproducible data analysis in neuroimaging[J]. Annual Review of Biomedical Data Science, 2019, 2(1): 119-138.

[14] KADIYALA A, KUMAR A. Applications of Python to evaluate environmental data science problems [J]. Environmental Progress & Sustainable Energy, 2017, 36(6): 1580-1586.

[15] CHEN J, TAO Y, WANG H, et al. Big data based fraud risk management at Alibaba[J]. The Journal of Finance and Data Science, 2015, 1(1): 1-10.

[16] Microsoft Azure Data Catalog[EB/OL]. (2019-08-01) [2021-05-22]. https://docs.microsoft.com/en-us/azure/data-catalog/overview.

[17] IKEDA R, WIDOM J. Data lineage: A survey[R]. Stanford InfoLab, 2009.

[18] WOODRUFF A, STONEBRAKER M. Supporting fine-grained data lineage in a database visualization environment[C]// Proceedings 13th International Conference on Data Engineering. IEEE, 1997: 91-102.

[19] KANDEL S, HEER J, PLAISANT C, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data[J]. Information Visualization, 2011, 10(4): 271-288.

[20] FURCHE T, GOTTLOB G, LIBKIN L, et al. Data Wrangling for Big Data: Challenges and Opportunities[C]// EDBT. 2016, 16: 473-478.

[21] CHAO L M, XING C X, ZHANG Y. Data Science Studies: State-of-the-art and Trends[J]. Computer Science, 2018, 45(1): 1-13.

[22] Data Wrangling with Spotfire[EB/OL]. [2021-05- 22]. https://www.tibco.com/products/tibco-spotfire/data-wrangling.

[23] LIU C, MAO Y, VAN DER MERWE J, et al. Cloud resource orchestration: A data-centric approach[C]// Proceedings of the biennial Conference on Innovative Data Systems Research (CIDR). 2011: 1-8.

[24] LIU X, LIU Y, SONG H, et al. Big data orchestration as a service network[J]. IEEE Communications Magazine, 2017, 55(9): 94-101.

[25] The KNIME Model Process Factory [EB/OL]. (2017-05-08) [2021-05-22]. https://www.knime.com/blog/the-knime-model-process-factory.

[26] What is Power BI [EB/OL]. [2021-05- 22]. https://powerbi.microsoft.com/zh-cn/what-is- power-bi/.

[27] Data refresh in Power BI [EB/OL]. (2021-05-07) [2021-5-22]. https://docs.microsoft.com/en-us/power-bi/connect-data/refresh-data.

[28] MÄKINEN S, SKOGSTRÖM H, LAAKSONEN E, et al. Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOpsHelp? [J]. arXiv: 2103.08942, 2021.

[29] Platform Component: Model Ops [EB/OL]. [2021-05-22]. https://www.dominodatalab.com/product/model-ops/.

[30] Domino Model Monitor[EB/OL]. [2021-05-22]. https://www.dominodatalab.com/product/domino-model-monitor/.

[31] ERETH J. DataOps-Towards a Definition[J]. LWDA, 2018, 2191: 104-112.

[32] BASS L, WEBER I, ZHU L. DevOps: A software architect's perspective[M]. Addison-Wesley Professional, 2015.

[33] What is Azure DevOps? [EB/OL]. (2021-01-22)[2021-05-22]. https://docs.microsoft.com/en-us/azure/devops/user-guide/what-is-azure-devops?view=azure-devops.

[34] DANG Y, LIN Q, HUANG P. AIOps: real-world challenges and research innovations[C]// 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion). IEEE, 2019: 4-5.

[35] IBM Cloud Pak for Watson AIOps [EB/OL]. [2021-05-22]. https://www.ibm.com/cloud/cloud-pak-for-watson-aiops?lnk=STW_US_STESCH&lnk2=learn_CloudPakAIOps&pexp=DEF&psrc=NONE&mhsrc=ibmsearch_a&mhq=AIOPS.

[36] JOGALEKAR P, WOODSIDE M. Evaluating the scalability of distributed systems[J]. IEEE Transactions on parallel and distributed systems, 2000, 11(6): 589-603.

[37] PAHL C, XIONG H, WALSHE R. A comparison of on-premise to cloud migration approaches[C]// European Conference on Service-Oriented and Cloud Computing. Berlin, Heidelberg: Springer, 2013: 212-226.

[38] HASSENZAHL M, TRACTINSKY N. User experience—a research agenda[J]. Behaviour & Information Technology, 2006, 25(2): 91-97.

[39] STOLTE C, TANG D, HANRAHAN P. Polaris: A system for query, analysis, and visualization of multidimensional relational databases[J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 52-65.

[40] MACKINLAY J, HANRAHAN P, STOLTE C. Show me: Automatic presentation for visual analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1137-1144.

[41] TSIAKMAKI M, KOSTOPOULOS G, KOTSIANTIS S, et al. Implementing AutoML in educational data mining for prediction tasks[J]. Applied Sciences, 2020, 10(1): 90.

[42] HUMMER W, MUTHUSAMY V, RAUSCH T, et al. Modelops: Cloud-based lifecycle management for reliable and trusted AI[C]// 2019 IEEE International Conference on Cloud Engineering (IC2E). IEEE, 2019: 113-120.

[43] PRAT N. Augmented analytics [J]. Business & Information Systems Engineering, 2019, 61(3): 375-380.

[44] ZHENG N, LIU Z, REN P, et al. Hybrid-augmented intelligence: collaboration and cognition[J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(2): 153-179.

[45] SRIDHAR V, SUBRAMANIAN S, ARTEAGA D, et al. Model governance: Reducing the anarchy of production ML[C]// 2018 {USENIX} Annual Technical Conference. 2018: 351-358.

[46] GUNNING D. Explainable artificial intelligence (xai)[R]. Defense Advanced Research Projects Agency (DARPA), 2017.

[47] MIAO H, LI A, DAVIS L S, et al. Towards unified data and lifecycle management for deep learning[C]// 2017 IEEE 33rd International Conference on Data Engineering (ICDE). IEEE, 2017: 571-582.

[48] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[49] HUANG L, JOSEPH A D, NELSON B, et al. Adversarial machine learning[C]// Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence. 2011: 43-58.

[50] NAYAK S, GOURISARIA M K, RAUTARAY P M. Recent Dimensions of Data Science: A Survey[M]// Advances in Data and Information Sciences. Singapore: Springer, 2020: 465-476.

[51] SHAHRIVARI S. Beyond batch processing: towards real-time and streaming big data[J]. Computers, 2014, 3(4): 117-129.

[52] WICKHAM H. Tidy data[J]. Journal of statistical software, 2014, 59(10): 1-23.

[53] 3 common messy data problems and how to tidy them in SAS [EB/OL]. (2016-06-02) [2021-05-22]. https://communities.sas.com/t5/SAS-Communities-Library/3-common-messy-data-problems-and-how-to-tidy-them-in-SAS/ta-p/272165.

[54] PERER A, LIU S. Visualization in data science[J]. IEEE Computer Graphics and Applications, 2019, 39(5): 18-19.
[55] PATHAK S, PATHAK S. Data Visualization Techniques, Model and Taxonomy[M]// Data Visualization and Knowledge Engineering. Springer, Cham, 2020: 249-271.
[56] KUMAR R S S, NYSTRÖM M, LAMBERT J, et al. Adversarial machine learning-industry perspectives[C]// 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020: 69-75.
[57] Adversarial Machine Learning [EB/OL]. [2021-05-24]. https://researcher.watson.ibm.com/researcher/view_group.php?id=9571.
[58] Threat Modeling AI/ML Systems and Dependencies[EB/OL]. (2019-11-11)[2021-05-24]. https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml.
[59] Hype cycle for artificial intelligence [EB/OL]. (2020-07-27)[2021-05-24]. https://www.gartner.com/en/documents/3988006/hype-cycle-for-artificial-intelligence-2020.
[60] Gartner Says More Than 40 Percent of Data Science Tasks Will Be Automated by 2020 [EB/OL]. (2017-01-16)[2021-05-24]. https://www.gartner.com/en/newsroom/press-releases/2017-01-16-gartner-says-more-than-40-percent-of-data-science-tasks-will-be-automated-by-2020.
[61] DAFOE A. AI governance: a research agenda[R]. Governance of AI Program, Future of Humanity Institute, University of Oxford, 2018.
[62] ARRIETA A B, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 2020, 58: 82-115.
[63] ZHAO X, JOHNSON M E. Access governance: Flexibility with escalation and audit[C]// 2010 43rd Hawaii International Conference on System Sciences. IEEE, 2010: 1-13.
[64] ENGLAND P, LAMPSON B, MANFERDELLI J, et al. A trusted open platform[J]. Computer, 2003, 36(7): 55-62.
[65] KSHIRSAGAR M, ROBINSON C, YANG S, et al. Becoming Good at AI for Good[J]. arXiv: 2104.11757, 2021.
[66] The Delta Lake Series-Lakehouse [EB/OL]. [2021-05-24]. https://databricks.com/p/ebook/the-delta-lake-series-lakehouse.
[67] ROSSI R, AHMED N. The network data repository with interactive graph analytics and visualization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015.
[68] BANKS J. Discrete event system simulation[M]. Pearson Education India, 2005.
[69] SHARMA P. Discrete-event simulation[J]. International Journal of Scientific & Technology Research, 2015, 4(4): 136-140.
[70] MACAL C, NORTH M. Introductory tutorial: Agent-based modeling and simulation[C]// Proceedings of the Winter Simulation Conference 2014. IEEE, 2014: 6-20.
[71] WHITE P. The power of the industrial internet: turning data into insight and action[J]. Journal of Petroleum Technology, 2014, 66(11): 90-93.
[72] PRETLOVE J, SKOURUP C. Human in the loop[J]. ABB Review, 2007, 1: 6-10.
[73] SARKAR S, WEYDE T, GARCEZ A, et al. Accuracy and interpretability trade-offs in machine learning applied to safer gambling[C]// CEUR Workshop Proceedings. 2016: 1773.
[74] ADADI A, BERRADA M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)[J]. IEEE access, 2018, 6: 52138-52160.
[75] KOSARA R, MACKINLAY J. Storytelling: the next step for visualization[J]. Computer, 2013, 46(5): 44-50.
[76] CHAO L M, ZHANG C. Data Storytelling: From Data Perception to Data Cognition[J]. Journal of Library Science in China, 2019, 45(5): 61-78.
[77] PENG R D. Reproducible research in computational science[J]. Science, 2011, 334(6060): 1226-1227.
[78] MUNAFÒ M R, NOSEK B A, BISHOP D V M, et al. A manifesto for reproducible science[J]. Nature Human Behaviour, 2017, 1(1): 1-9.
[79] WEIßGERBER T, GRANITZER M. Mapping platforms into a new open science model for machine learning[J]. it-Information Technology, 2019, 61(4): 197-208.
[80] SALEIRO P, RODOLFA K T, GHANI R. Dealing with bias and fairness in data science systems: A practical hands-on tutorial[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3513-3514.
[81] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[J]. arXiv: 1805.12152, 2018.
[82] MULLIGAN D K, KOOPMAN C, DOTY N. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, 374(2083): 20160118.
[83] PROVOST F, FAWCETT T. Data science and its relationship to big data and data-driven decision making[J]. Big data, 2013, 1(1): 51-59.
[84] PASSI S, JACKSON S J. Trust in data science: Collaboration, translation, and accountability in corporate data science projects[J]. Proceedings of the ACM on Human-Computer Interaction, 2018, 2(CSCW): 1-28.
[85] H2O.ai + COVID-19 [EB/OL]. [2021-05-24]. https://www.h2o.ai/covid-19/.
[86] LATIF S, USMAN M, MANZOOR S, et al. Leveraging data science to combat covid-19: A comprehensive review[J]. IEEE Transactions on Artificial Intelligence, 2020, 1(1): 85-103.

CHAO Le-men, born in 1979, Ph.D, associate professor, Ph.D supervisor. His main research interests include data science and big data analysis.

WANG Rui, born in 1998, postgraduate. Her main research interests include data science and big data analysis.