数据驱动研究范式和一流高校数据服务支撑体系研究

——首届全国高校数据驱动创新研究大赛综述

□崔海媛* 罗鹏程 赵静茹 聂华 王继民 张久珍

摘要 数据驱动研究已成为全学科研究范式，数据服务也已成为高校师生最需要的新服务。通过调研国内外数据驱动研究比赛情况，介绍首届全国高校数据驱动创新研究大赛参赛情况，对组织过程、评选方法、论文选题和研究方法进行深入分析。并通过研究各学科数据驱动研究方法的特点和趋势，提出高校数据驱动研究服务支撑体系，为推动数据驱动研究范式和建设高校数据服务体系提供参考。

关键词 数据驱动研究 数据服务 数据管理 数字人文 大数据研究

分类号 G251

DOI 10.16603/j.issn1002－1027.2018.06.005

1 大赛组织

2013年被称为“大数据元年”[1]，2016年被称为“人工智能元年”[2]。在大数据和人工智能时代，数据的价值受到学术界、产业界和政府的高度重视。为了培养数据人才，很多高校与学会举办了数据大赛，如中国计算机学会的“大数据与计算智能大赛”[3]、清华大学的“中国高校计算机大赛——大数据挑战赛”等[4]。为了吸引数据人才解决热点和难点问题，大量竞赛平台在产业界涌现，Kaggle[5]、DataCastle[6]、阿里巴巴天池大数据竞赛[7]等平台吸引了数以万计的数据人才参与。为了推动研究和大数据产业发展，某些省市政府近年来也举办了一些有影响力的数据比赛，如贵州省的“中国国际大数据挖掘大赛”[8]、广东省的“广东政务数据创新大赛”[9]等。在图书情报领域，上海图书馆举办的“开放数据应用开发竞赛”，通过面向全社会征集以开放数据为基础的优秀移动应用产品原型和服务创意，以期充分释放人文数据的价值[10]。

以上数据类比赛主要是算法模型和应用创新类大赛，且大多集中在计算机科学相关领域，缺乏全学科的、研究性的比赛。在大数据与人工智能快速发展的环境下，数据驱动研究已经渗透到所有学科，需要面向全学科、研究性的数据比赛来促进高校的教学和科研。与数据受到高度重视相对应，国内研究数据本身的开放程度和管理水平还存在不足，数据管理意识也急需提升。为了促进高校学生基于数据进行研究，提高学术创新能力，促进研究数据的保存、共享和利用，北京大学图书馆、北京大学信息管理系、南海大数据应用研究院主办，联合国家信息中心等多家机构，合作组织了“首届全国高校数据驱动创新研究大赛”[11]（以下简称“大赛”）。大赛一经发布便吸引众多学生参赛报名，在全国高校范围内产生了较大影响。

本文对大赛的组织和参赛情况进行介绍；分析参赛论文选题，以及数据驱动研究方法在各学科的现状与趋势，揭示数据密集型研究范式对各学科的影响；同时，还对数据驱动研究服务支撑体系的设计提出建议，为高校数据驱动研究和服务体系建设提供参考。

1.1 实施过程

大赛实施过程主要包括以下5个部分：

①大赛策划：2017年9—11月。组织团队对国内

* 通讯作者：崔海媛，ORCID：0000-0001-5541-7100，邮箱：cuihy@lib.pku.edu.cn。

外多个重要数据比赛进行调研，对校内多个院系的16位老师访谈需求，并结合数据服务，设计大赛方案。

②宣传报名：2017年11月24日至2018年1月15日。大赛通知于2017年11月24日正式发布，并通过多种渠道广泛宣传，包括社交网络（如微信、微博）、新闻媒体（如北京大学主页和新闻网、图书馆主页、微信公众号）、各高校图书馆（通过中国高等教育文献保障系统成员馆）、合作机构（如北京大学社科调查中心领域内院校宣传）、海报展板、线下多渠道动员等。同时举办启动与培训会、大赛咨询等相关工作。这一期间，有近4万名用户访问了大赛主页，最终吸引了来自全国169所高校的1892名同学报名参赛。

③成果提交与评审：2018年1月16日至2018年3月22日。报名结束后，来自121所高校的968名同学成功提交了作品。为了保证评审的客观、公正，组委会邀请国内各领域知名专家任指导委员会委员，同时邀请来自北京大学、中国人民大学、中国科学院等近10所著名科研机构的25位专家担任评审委员会成员。评审原则以选题价值、创新性、论证严谨性、工作量、规范性与数据原创性作为标准，每篇论文需经过查重、形式审查、两位专家初审、专家复审。最终，共有13支队伍进入现场答辩，120支队伍获得优秀奖。

④现场答辩（2018年4月3日）。现场答辩在北京大学阿卜杜勒·阿齐兹国王公共图书馆分馆举行，邀请北京大学、清华大学、中国人民大学、国家信息中心、腾讯研究院等10所机构的13名专家作为答辩专家组成员。通过参赛队员的演讲展示、专家提问等环节，最终确定了特、一、二、三等奖得主。

⑤京陵峰会展示（2018年5月5日至2018年5月6日）。大赛特等奖、一等奖得主受邀在“第二届京陵大数据高峰论坛”主论坛介绍研究成果，二等奖得主在“数字中国”分论坛介绍研究成果。大赛参赛作品受到与会专家、企业代表的广泛好评，为大赛赢得良好口碑。

1.2 参赛情况

本次大赛在全国高校范围内引起广泛关注，共吸引来自北京大学等全国169所高校的1892名同学报名参赛，共有593组队伍，其中本科生392组、研究生201组，平均每组队员3.19人。在参赛报名学校中，北京大学、武汉大学人数众多，分别为190人、146人。在参赛报名的学科中，共涉及了56个一级学科，其中图书馆、情报与档案管理的报名队伍最多，达到102组，其次是应用经济学，达84组。

最终，来自121所高校的968人（共289支队伍）成功提交了参赛作品，在成功提交作品的学校中，北京大学、武汉大学、中山大学人数最多，分别达到139人、83人、38人。图1给出了成功提交作品人数最高的前20所高校，从图中可以看出，至少有20所高校成功提交参赛作品的人数在10人以上，部分高校在报名阶段人数众多，但最终成功提交作品的人数却很少。在成功提交作品的学科中，共涉及45个一级学科，图2给出了成功提交作品最多的前20个一级学科。从图中可以看出，图书馆、情报和档案队伍最多，达到59组；其次是应用经济学、社会学、管理科学与工程、统计学、计算机科学，也分别达到50组、27组、23组、19组、19组。

1.3 数据使用情况

大赛期间，北京大学开放研究数据平台（以下简称“开放数据平台”）[12]的日均访客量、页面浏览量提升10余倍，分别达到549人、3793页，注册用户量与数据下载量均增加了4倍多。参赛队员在平台中提交了约170个数据集。

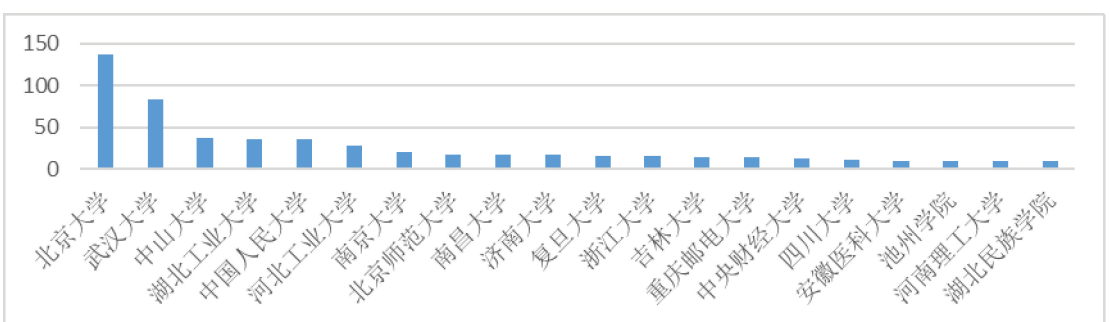

图1 成功提交作品的参赛队员高校分布情况（TOP 20）

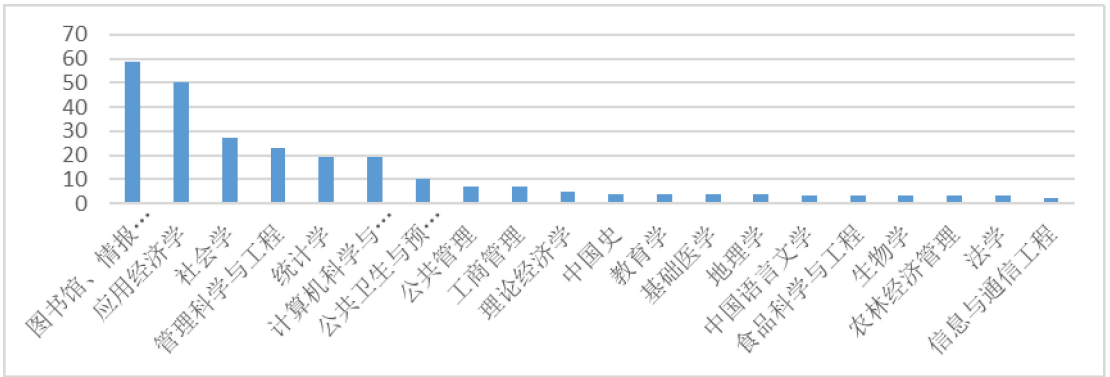

图 2　成功提交作品的参赛团队一级学科分布情况(TOP 20)

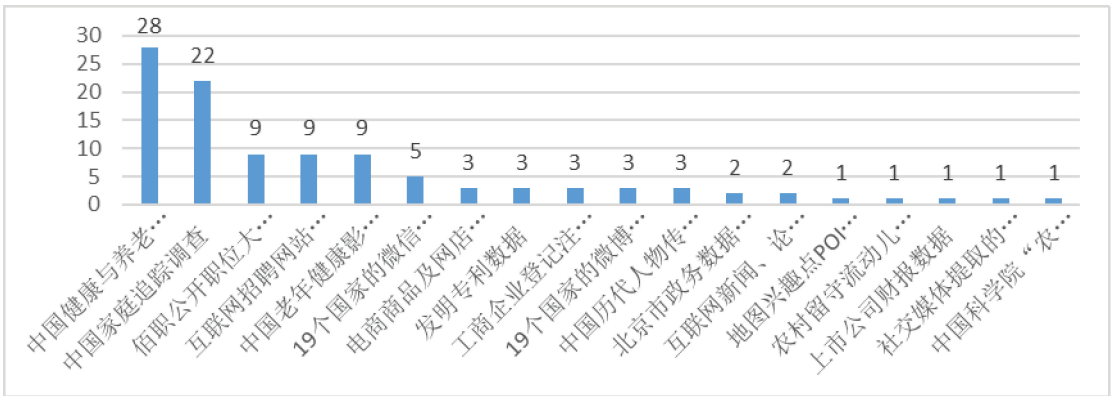

图 3　平台已有数据使用情况

在通过形式审核的 233 篇论文中，有 103 篇论文使用平台已有数据，135 篇论文使用自带的原创性数据，同时使用平台已有数据和自己原创数据的共有 5 篇。图 3 统计了平台已有数据在本次大赛中的使用情况。从图 3 可以看出，参赛队伍共使用了 18 个平台的已有数据集。使用量最多的为中国健康与养老追踪调查，达到 28 队；其次是中国家庭追踪调查，达到 22 队；佰职公开职位大数据信息、互联网招聘网站数据、中国老年健康影响因素跟踪调查使用量也较多，均达到 9 队。

参赛队伍使用数据情况与笔者前期调研情况表明：数据已是全学科研究的基础，研究者需要高质量的大数据提交、发布和研究服务平台；高质量的调查数据、招聘数据与社交媒体数据仍是人文社科经济学领域研究者重点关注的研究数据。

2　参赛论文分析

以 289 篇参赛论文为研究对象，通过数据分析，从总体上研究各学科基于数据驱动研究的选题。针对 133 篇获奖论文[13]，依据参赛作品数量和学科类别，将提交的论文归属到计算机、理工农医、人文社科、经济管理与图书情报 5 个领域，分别分析其研究主题与方法，研究各学科在数据驱动研究领域的研究现状与数据需求特点。

2.1　研究主题分析

利用百度 AI 开放平台语言处理基础技术[14]中词法分析接口，对 289 篇参赛论文的标题进行分词处理，统计各标题中名词和动词出现的频次，结果如表 1、表 2 所示。从中可以发现，“数据”一词出现的次数最多，高达 68 次，充分反映出参赛队伍紧扣“数据驱动”的大赛主题。对表中数据进行总结，可以发现各参赛队伍注重数据分析，关注于“网络”“信息”“老年人”“家庭”“教育”“招聘”“企业”等主题的“影响”“因素”“行为”“特征”等方面的“分析”“挖掘”“预测”“实证”研究等。

利用百度 AI 开放平台语言处理基础技术中短文本相似度接口，计算参赛论文标题两两之间的相似度，再利用 scikit-learn[15]中谱聚类算法对论文标题进行聚合(尝试了不同参数，最终选择 8 个类)。从聚类结果中发现两个类有明显含义：一个是关于

“老年人”和“养老”主题的研究(共有 42 篇论文)，另一个是关于“招聘”和“人才”主题的研究(共有 24 篇论文)。从表 1 和表 2 中相关名词、动词的出现频率可以观察到，这两个主题确实涉及较多论文。随着我国人口老龄化的加速，老年人的心理健康、生活满意度、养老和社会保障等问题日益突出，通过大赛中关于这一主题的众多成果可以看出，学术领域十分重视对这一社会问题的研究。招聘和人才一直是社会关注的焦点，同时也是同学们以后会面临的问题，参赛作品涉及这一主题的论文较多，反映出同学对此的关心。

表 1　标题中名词出现次数排名(频次 5 以上)

名词	频次	名词	频次	名词	频次	名词	频次	名词	频次	名词	频次
数据	68	中	13	方法	8	大数据	7	技术	6	模式	5
因素	33	模型	12	CHARLS	8	用户	7	互联网	6	聚类	5
例	26	需求	12	系统	8	情况	7	领域	6	收入	5
下	23	文本	11	热点	8	微博	7	视角	6	学生	5
实证	21	中老年	10	人才	8	知识	7	科学	6	CFPS	5
国	21	信息	10	状况	7	环境	6	农村	5	情感	5
老年人	19	企业	9	城市	7	机制	6	子女	5	机器学习	5
家庭	16	社会	9	经济	7	心理	6	资本	5	金融	5
网络	16	特征	9	人	7	关系	6	市场	5		
行为	16	现状	8	行业	7	职位	6	政策	5		

表 2　标题中动词出现次数排名(频次 5 以上)

动词	频次	动词	频次	动词	频次	动词	频次	动词	频次	动词	频次
研究	104	预测	19	基于	13	可视化	8	创新	6	追踪	6
分析	96	挖掘	16	发展	11	养老	8	消费	6	探究	6
影响	59	调查	15	教育	9	评价	7	参与	6	应用	5
为	29	招聘	14	服务	8	识别	7	利用	6	相关	5
										构建	5

2.2　计算机学科

通过形式审核的计算机类论文共有 19 篇，其中 1 篇获得特等奖，1 篇获得二等奖、另有 9 篇获得优秀奖。这些论文主要以现实需求为背景，利用机器学习、数据挖掘等方法解决问题。中国人民大学史嘉彤等的论文《AASC：基于 Logistic 曲线与深度回归的众筹项目建议自动生成系统》，以众筹项目的成功性预测和资金募集持续时长的建议为目标，收集 Kickstarter 中众筹项目数据，利用深度回归和四变量逻辑回归(Logistic)构建了预测模型。论文立意新颖、论证合理、分析深入、数据规范，具有较好的应用前景，得到初审、复审、答辩专家的一致认可，最终获得特等奖。中山大学张驰俊等的论文《基于 GBRT 的电影日票房预测建模》，以电影日票房预测为目标，收集电影资讯网站中的影片属性、用户评论以及演员百度指数等数据，利用梯度渐进回归树构建了预测模型。论文思路清晰、数据收集得当、特征分析详实，具有创新性和应用价值，经过现场答辩获得二等奖。

其他参赛队伍有从校园生活入手，基于校园大数据构建学生画像，并利用机器学习算法实现奖助学金预测、失联预警、成绩预测等；有以可穿戴设备人机交互需求为背景，利用智能手环收集运动数据、构建特征，使用机器学习分类算法实现手势识别，并应用于游戏控制。还有使用支持向量机算法对股票进行预测，使用关联规则、K 均值聚类等数据挖掘方法对招聘数据进行分析，使用主题模型、情感分析等文本挖掘算法对微信公众号、学生评教文本进行分析。也有算法设计与实现类的论文，如社交网络中意见领袖识别算法。

通过对计算机类的参赛论文的分析可知，目前计算机学科领域数据驱动的研究集中于深度学习等机器学习类算法、关联规则等数据挖掘算法的研究与应用。在大数据驱动下，深度学习是当前计算机领域的热门研究。

2.3　理工农医学科

通过形式审核的理工农医类论文共有 35 篇，其中 1 篇获得一等奖，1 篇获得二等奖，1 篇获得三等奖，另有 17 篇获得优秀奖。这些论文均从各自学科领域的问题出发，基于数据进行问题分析、模型改进等工作。北京大学吴瑶等的论文《中国人群的心血管发病风险及预测》，通过利用“中国健康养老追踪调查”中的居民健康相关信息，对三种心血管发病风险预测模型进行了评估。并在评估的基础上，筛选出相关危险因素，建立新的预测模型并进行相关检验。该研究选题具有较大的社会意义和应用价值，通过现场答辩得到专家认可，获得一等奖。浙江大学陈万成等的论文《基于数据挖掘方法的 HEDONIC 房屋价格评估模型——以美国城市西雅图为例》，以基于数据的自动房屋价格评估为目标，将随机森林、神经网络、K 最近邻三种数据挖掘算法与 HEDONIC 模型相结合，得到了效果更优的模型。该研究选题具有一定的应用价值，在模型算法

上有一定的创新，经过现场答辩获得二等奖。首都医科大学张莹等的论文以北京市顺义区妇幼保健院临床实验数据为基础，针对霍夫曼(Hoffmann)方法选择数据方面存在主观性的不足进行了分析研究，经过现场答辩获得三等奖。

其他参赛队伍有以保健食品本体构建为目标，通过获取中药材网站、百度百科等数据，综合利用自然语言处理等方法半自动地构建本体。有从提升高速公路管理运营效率为出发点，构建绿色通行车辆画像，利用机器学习算法设计实现假冒绿通车识别系统。有以南京紫金山西麓中山植物园内南方红豆杉为对象，采集相关数据，量化南方红豆杉与邻近木之间的竞争状况，分析种群的更新方式以及林分结构与植株生长之间的关系。还有基于统计、GIS 等分析方法对老龄人口常见疾病发病率和环境因素进行时空探索分析；有基于电子鼻技术和机器学习方法识别中药种类和产地；也有结合国家战略，利用大数据方法研究海上丝绸之路的石油运输发展。

通过对理工农医类参赛论文的分析可知，目前数据驱动的研究聚焦于本学科研究问题，从各种途径采集并分析数据，研究分析方法主要来自各领域研究提出的模型和统计学方法。此外，部分研究论文也开始使用计算机领域的最新研究方法，将机器学习、数据挖掘、自然语言处理等方法应用于本学科问题解决。

2.4 人文社科学科

通过形式审核的人文社科类论文共有 52 篇，其中 2 篇获得三等奖，另有 28 篇获得优秀奖。北京大学周丽玮等的论文《大数据视阈下 2015－2017 年大气污染治理政策对“雾霾”网络舆论的影响》，基于 PM2.5 观测数据、新浪微博数据，通过统计分析、情感分析等技术手段，比较了近年来雾霾治理的成效以及与网络舆论间的关系。该研究选题较为新颖，有一定应用价值，经过现场答辩，获得三等奖。南京大学贺鲲羽和中南大学彭圣钦的论文《教育获得的多代传递：基于中国家庭追踪调查 2010 年数据的分析》，基于开放数据平台“中国家庭追踪调查”的数据，运用偏比例优势模型估计祖代的教育程度对孙代学习进度的影响效应。该论文方法严谨、撰写规范，经过现场答辩，获得三等奖。

其他参赛队伍有利用贝叶斯统计方法的分层负二项回归模型，考察高速公路建设与制造业企业进入退出之间的因果关系。有基于中国历代人物传记资料库，利用社会网络社群发现方法分析宋朝元祐年间文人和政治人物之间关系；也有分析北宋熙宁变法中变法派与保守派代表人物的社会关系，探究熙宁变法与朋党政治之间的关联。有以北京市近年的第三产业比重为研究对象，运用多元线性回归和神经网络进行分析和预测。还有利用中国健康与养老追踪调查数据，采用回归模型分析老年人主观幸福感的影响因素，研究数字鸿沟对老年人生活满意度的影响。

通过对人文社会科学类参赛论文的分析可知，目前数据驱动的研究方法主要以统计学为基础，运用描述性统计、相关性分析、回归分析等手段分析问题。部分参赛论文也会融合自然语言处理、机器学习、社会网络分析方法，总体来看，这类方法使用的还不是很广泛。这表明人文社科领域的数据驱动研究方法，仍需要进一步与计算机学科更深度交叉融合，应用计算机最新研究技术，产生更具创造性的研究方法和研究成果。

2.5 经济管理学科

通过形式审核的经济管理类论文共有 74 篇，其中 1 篇获得二等奖，2 篇获得三等奖，另有 39 篇获得优秀奖。东北财经大学张梦吉等的论文《引入新闻短文本的个股走势预测模型》，选取与股价有强相关性的资金流向、公司财务指标等定量数据，同时引入深度学习模型自动提取新闻事件数据，建立个股走势预测模型。该论文选题具有实际应用价值，数据分析方法科学且有一定的难度，经过现场答辩获得二等奖。武汉大学王宠霖等的论文《数据驱动下的股票市场短期预测》，通过多种途径采集数据，比较分析了传统统计分析方法、逻辑分类以及深度学习对一分钟高频交易股价变化趋势预测的效果。该论文选题具有实际应用价值，经过现场答辩获得三等奖。南开大学温旭东等的论文《捕捉“隐形的篮子”——微观经济智能决策系统》，利用大数据、机器学习相关方法，融合劳动力市场、金融市场、行业发展、投资情绪、天气因素、地理特征等多样化数据，构建了微观经济决策系统，并对其中的工资波动预测、股票智能推荐、投资组合三个模块进行分析。该研究使用多种数据源、工作量较大，具有一定的市场转化价值，经过现场答辩获得三等奖。

其他参赛队伍，有采用社会网络方法对招聘求

职数据进行分析，研究互联网招聘职位与求职者的优化匹配。有采用文本聚类、情感分析、机器学习等手段分析微博数据，研究北京市群租房问题。有基于电商商品及网店数据，使用数据可视化、回归模型挖掘商品和商家之间的关系，帮助商家寻找适合的盈利模式。有基于重型货车北斗车联网数据，通过地图可视化及聚类方法对危险报警行为进行区域分析。有基于“一带一路”沿线国家区位特征与投资项目信息，用机器学习模型推荐适合投资的国家。还有通过构建知识图谱的方式对互联网行业招聘数据集进行挖掘，分析人才需求并为应聘者提出建议。

通过对经济管理类参赛论文的分析可知，目前数据驱动研究方法中机器学习、数据挖掘等方法应用得较多。在经济管理中，有许多可以根据历史信息预测未来发展变化的研究问题，这些问题非常适合机器学习方法的应用。因此，许多经济管理类的参赛论文都或多或少使用了最近较热门的深度学习等机器学习方法。

2.6 图书情报学科

通过形式审核的图书情报类论文共有 53 篇，其中 3 篇获得三等奖，另有 27 篇获得优秀奖。武汉大学周莉娜等的论文《中文诗歌知识图谱构建与服务》，以中国历代人物传记资料库、古代史知识库等数据为基础，进行数据抽取和知识融合，构建中文诗歌知识图谱，在此基础上实现诗歌知识查询、语义检索、智能问答等功能。知识图谱是图情领域的研究热点，经过现场答辩，该论文获得三等奖。合肥工业大学江仔玲等的论文《社会化阅读环境下阅读内容计量研究——以阅读推广类微信公众号推送文章为例》，基于清博指数大数据平台采集阅读推广相关微信公众号数据，通过统计分析、信息计量等方法对公众号文章特征进行深入分析，获得三等奖。北京大学李洪伟等的论文《基于时序狄利克雷过程主题模型的数据挖掘分析——以知乎求职话题为例》，以知乎“求职”话题下用户原创文章为数据源，通过基于时序狄利克雷过程的主题模型，对话题热点和时序演化情况进行了剖析，获得了三等奖。

其他参赛队伍有利用情感分析、文本聚类、主题模型等手段分析网络大数据，探究江南古镇旅游形象的同质化问题。有以科学引文索引（Web of Science）网络版中特定领域文献数据为基础，通过文献计量方法分析跨学科领域的演化情况。有利用共引文献数量、引文位置和顺序等信息设计论文查重算法，提高查重检测效果。有利用地理信息系统等工具，实现古代京官籍贯地理分布的可视化呈现与诠释。有利用情感分析、聚类和主题模型分析双十一微博话题热点和情感倾向的变迁。有以微博中“祈福锦鲤”的转发数据为基础，综合运用统计分析、社会网络分析、自然语言处理等手段分析吉祥物转发祈福的用户行为特征。还有以国内外健康问答社区的数据为基础，运用社会网络分析、内容分析等手段识别用户的信息需求。

通过对图书情报类参赛论文的分析可知，目前数据驱动研究方法以自然语言处理、文本挖掘、数据挖掘、社会网络分析等计算机类方法为主，同时也包含统计分析、信息计量等手段。在图书情报领域，无论是网络用户行为分析、知识组织与管理、网络文本和科学文献分析等，都离不开计算机信息处理手段。因此，相比其他非计算机类学科，图情领域应用计算机类方法最为突出。

3 科学研究与支撑服务变革

3.1 全学科数据驱动研究

通过对前期调研、大赛参与情况、数据统计与参赛作品的内容分析，可以发现，在学术交流新生态环境下，呈现全学科数据驱动研究的态势。

3.1.1 数据资源已是全学科研究的基础，价值日益重要

数据正成为科学研究的基础，对数据本身的研究、建设、保存和共享成为各学科领域的重要工作。在自然科学与工程领域，数据的重要性不言而喻，国内一些机构很早便开始了科学数据的建设工作，例如中国科学院从 1986 年开始启动科学数据库工程，现已建成科学数据云[16]；国家科技部也从 2002 年开始实施“国家科学数据共享工程”，在资源环境、农业、人口与健康、基础与前沿等领域开展了科学数据共享工作[17]。在人文社会科学领域，数据资源的建设在近年来获得突飞猛进的发展。从国家社科基金项目立项情况[18]可以发现，以数据库建设为目标的项目（题名包含“数据库”）在 2010 以前每年均在 5 项左右，2011 年达到 17 项，之后开始快速增长，在 2017 年达到历史最高的 39 项，其中重大项目就有 24 项。

在大赛策划阶段，我们对北京大学多个院系的

十多位教师进行了访谈调查。在访谈中，教师都表示需要数据支撑教学和科研，并希望能够有更加方便的途径获取数据资源，提升教学和科研效率。数据资源是学术交流新生态系统中的重要基础，已经超越文献资源的价值，成为教师、研究者、学生学习和研究的支撑性资源。正是因为数据本身的价值受到如此广泛的重视，在大赛策划发起后，短短2月内就获得了校内外众多机构的合作与支持。

3.1.2 数据驱动成为全学科研究方法、学科的交叉创新凸显

当前，大数据和机器学习的不断发展使得数据处理能力得到极大提升，数据驱动逐渐成为几乎所有学科可以利用的研究方法。本次大赛的成果来自于数十个一级学科，涉及理、工、农、医、人文、社科等领域，数据驱动研究方法不再局限于对计算分析要求较多的理工科，已经扩散至社会科学，甚至是人文科学，已经成为全学科研究方法。从参赛作品来看，图书情报、经济管理应用机器学习、数据挖掘、自然语言处理、社会网络分析等数据驱动类研究方法较多，人文社科类主要还是以统计方法为主，理工农医也主要以各学科自己的模型和统计方法为主，少数研究会使用最新的计算机类方法。

学科交叉融合能够做出更多创新性研究，在本次大赛中存在许多跨学科研究。例如，有信息管理学科的同学对历史、计算机、经济学问题进行研究，也有理工科专业的同学研究图书情报、经济管理领域的问题，还有许多跨学校、跨院系的研究团队。数据驱动研究范式，需要多学科交叉融合，碰撞出更多的创新火花。

3.1.3 对数据驱动研究方法的关注与学习是高校师生的普遍需求

大赛得到的高关注度和广泛参与度，表明数据驱动研究已经成为高校师生的普遍需求。应对大数据发展趋势，高等教育已经在变革，在学习门户网站(Studyportals)统计中，美国有104所高校设置“数据科学与大数据”(Data Science & Big Data)本科专业[19]。在中国，2016年2月，北京大学、对外经济贸易大学、中南大学首次成功申请到“数据科学与大数据技术”本科新专业。2017年3月，第二批32所高校获批。2018年3月，教育部最新公布的高校新增专业名单中，有248所学校获批，是过去两次审批通过额度的近8倍[20]，数据科学已成为当下高校最热门的专业领域。除数据科学专业、信息科学专业外，量化社会科学、数字人文研究也成为人文社科领域的研究热点，表明数据驱动研究方法需求广泛。设计与提供数据驱动研究服务体系，满足师生对数据研究与学习的需求，成为高校教学科研服务支撑体系的迫切需求。

3.1.4 提升师生数据素养、建设数据学习空间(平台)日益重要

通过评审参赛作品，发现部分参赛团队的选题创新性和新颖性还有不足，对数据清洗和数据规范描述不够清楚，数据分析方法相对单一，某些学科领域以描述和统计方法居多，数据挖掘深度不够，样本量不足，论文的学术规范性问题较多，对以往相关研究的综述不够深入。这些问题说明，高校学生对数据驱动研究的热情和关注度很高，但对数据使用、数据分析方法和数据挖掘技术掌握不够。反映出高校需要加强师生数据素养培训和指导，提升数据意识。新的研究方法和领域需要数据资源、研究案例和工具软件支持，需要尽快建设数据学习空间(平台)，支持数据驱动的教学和科研需要。普遍存在的论文学术规范性问题说明需要进行学术研究和论文写作规范的培训和指导。

3.2 数据驱动研究服务支撑体系设计

大数据和人工智能影响全球各行各业，数据驱动研究已是全学科研究范式，高校的教学、科研也随之转型，一流的教学和科研需要一流的服务体系支撑，设计与建设高水平的数据驱动研究服务支撑体系成为高校建设一流教学和科研服务的重要内容。欧洲、美国、澳大利亚等国家和地区的一流高校已经建设与提供数据服务[21][22][23]。根据对国内外高校数据服务的调研和对大赛的研究，建议数据驱动研究服务支撑体系设计应包括数据素养、学习支持、研究支持、数据长期保存和数据政策等五方面的内容，见表3。高校应确定由信息化建设机构、图书馆等相关职责部门研究和制定数据政策，开展数据素养培训，为教学和科研过程提供数据服务，为数据完整保存、复用提供长期保存支持。

①数据素养。

数据素养培训：开展相关的数据素养培训和咨询，包括数据的概念、价值、数据生命周期，以及国家、资助机构、研究机构数据政策。数据能力培训：开展数据清洗、规范描述和分析标准规范、方法培

训。工具软件：数据分析和挖掘等工具软件（R、Python 等）的使用培训。研究培训：研究过程中数据管理和数据方法等培训。学术规范培训：开展论文写作技巧与数据引用、出版合同签署、出版媒介选择、数据版权和发布等服务支持。数据资源培训：提供数据资源搜索发现服务培训，了解和使用数据资源。

表 3　数据驱动研究服务支撑体系

数据素养	学习支持	研究支持	数据长期保存	数据政策
数据素养培训	开放数据资源平台	项目管理私有云服务	研究数据保存服务	数据政策制定与发布
数据能力培训	数据资源导航	研究过程数据管理服务	元数据标准	政策服务（国家、资助机构、机构、期刊）
工具软件	数据学习空间	创新研究孵化空间	存储保存	知识产权服务
研究培训	教学案例库	工具软件	安全备份	
学术规范培训	工具软件		源代码保存	
数据资源培训				

②学习支持。

开放数据资源平台：提供高质量研究数据的发布与检索服务，为研究者提供数据发布和检索入口服务。数据资源导航：提供国家、国际组织、资助机构等主体的数据研究最新发展动态（网站、手册等），提升研究者的数据素养，帮助研究者了解最新研究进展。数据学习空间：以开放数据平台为基础，建设数据学习空间，提供数据、案例、工具软件、在线学习、数据竞赛等服务，为学习者提供从初学者到研究者的发展过程中的学习空间。教学案例库：以开放数据平台为基础，提供教学案例资源（数据、研究成果），不断迭代，以案例支持教学提升。工具软件：数据软件、工具与服务平台，为学习者提供支持。

③研究支持。

项目管理私有云服务：提供研究过程中数据管理私有云服务。研究过程数据管理服务：研究过程中项目管理平台和服务支持，以专职数据服务人员、平台和工具软件等多种服务，为研究者提供研究过程中的数据服务支持。创新研究孵化空间：提供数据创新研究、教学与产业孵化服务，以活动、比赛、技术转化等方式，推动校企合作，建设创新研究和产品孵化。工具软件：构建研究过程中的数据管理软件、工具与服务平台，为项目团队提供数据管理支持。

④数据长期保存。

研究数据保存服务：提供数字资源长期保存体系服务，保障数据资源长期保存。元数据标准：制定数据资源元数据标准规范。存储保存：提供数据存储保存基础设施、平台和服务。安全备份：制定数据安全备份制度。源代码保存：源代码保存平台，为研究者提供项目源代码长期保存服务。

⑤数据政策。

数据政策制定与发布：研究制定与推动机构研究数据管理和服务政策，推进开放学术交流。政策服务（国家、资助机构、机构、期刊）：数据政策培训，帮助研究者了解国际、国际组织、资助机构、机构和期刊数据政策，顺利完成项目申请和成果发表。知识产权服务：提供关于数据、成果等的知识产权培训与服务，帮助研究者了解知识产权法律法规。

4　结语

在大数据和人工智能时代，数据的价值日益凸显，数据驱动研究方法渗透到各个学科领域。为了促进高校学生基于数据进行研究，提高学术创新能力，促进研究数据的保存、共享和利用，北京大学图书馆和信息管理系面向全学科领域，推出研究型的数据驱动创新研究比赛，获得了全国高校和研究机构的广泛关注，促进了教学和科研。当前，世界各国均将数据视为战略资源，纷纷抢占发展先机，制定和推动发展战略规划。中国高校正处于“双一流”建设的发展机遇中，高校需要建设一流的数据驱动研究服务体系，服务国家战略，支持各学科前沿研究，以数据服务为凝聚力，提供基于数据的教学与研究的支撑平台，开展数据创新研究、教学与产业孵化等服务，推动数据教学与科研创新。服务转型中的高校图书馆，需要适应新的发展趋势，在数据驱动研究服务中，找到角色定位，发挥作用，提高图书馆的服务影响力。

参考文献

1　熊建，黄碧梅，林琳等. 2013 大数据元年[N].人民日报，2013－12－25(010).
2　卫人.中国人工智能历史元年，一切才刚刚开始[N].中国经济导报.2016－12－14(B01).
3　第五届 CCF 大数据与计算智能大赛启航[EB/OL]. [2018－07－11]. http://www.ccf.org. cn/c/2017－09－29/615097.shtml.
4　2018 中国高校计算机大赛——大数据挑战赛[EB/OL]. [2018－07 － 11]. https://www. kesci. com/apps/home/competition/5ab8c36a8643e33f5138cba4.

5 Kaggle: your home for data science[EB/OL]. [2018－07－11]. https://www.kaggle. com/.
6 DataCastle 大数据竞赛平台[EB/OL]. [2018－07－11]. http://www.pkbigdata.com/.
7 阿里巴巴天池大数据竞赛[EB/OL]. [2018－07－11]. https://tianchi.aliyun.com/.
8 云上贵州 数据之都[N]. 经济日报，2017－05－28(08).
9 广东政务数据创新大赛. [EB/OL].[2018－07－11]. https://tianchi.aliyun.com/ markets/tianchi/gov20172.
10 开放数据应用开发竞赛 2018[EB/OL]. [2018－07－11]. http://opendata.library.sh.cn/.
11 首届全国高校数据驱动创新研究大赛[EB/OL]. [2018－07－11]. http://opendata.pku.edu.cn/competition－2018.xhtml.
12 北京大学开放研究数据平台[EB/OL]. [2018－07－11]. http://opendata.pku.edu.cn/.
13 首届全国高校数据驱动创新研究大赛成果展示[EB/OL]. [2018－07－13]. http://opendata. pku. edu. cn/competition-product.xhtml.
14 语言处理基础技术——百度 AI[EB/OL]. [2018－07－13]. http://ai.baidu.com/tech/ nlp/.
15 Scikit-learn: machine learning in python[EB/OL]. [2018－07－13]. http:// scikit-learn.org/stable/.
16 中国科学院数据云[EB/OL]. [2018－07－13]. http://www.csdb.cn/aboutus/585.jhtml.
17 张先恩. 国家科学数据共享工程[J]. 科学中国人，2004(9)：11－13.
18 国家社科基金项目数据库[EB/OL]. [2018－07－13]. http://fz.people.com.cn/skygb /sk/.
19 Searchbachelor's programmes worldwide [EB/OL]. [2018－07－13]. https://www. bachelorsportal.com/search/#q=di-282|lv-bachelor,preparation|tc-EUR&start= 0&order=relevance.
20 教育部最新:283 所高校获批数据科学与大数据专业[EB/OL]. [2018－07－13]. http://36kr.com/p/5125134.html.
21 Bryant R, Lavoie B. A tour of the research data management (RDM) service space [EB/OL]. [2018－07－13]. https://www.oclc.org/content/dam/research/ publications/2017/oclcresearch-research-data-management-service-space-tour-2017-a4.pdf.
22 Tenopir C, Talja S, Horstmann W, et al. Research data services in European academic research libraries[J]. Liber Quarterly, 2017, 27(1): 23－44.
23 Yoon A, Schultz T. Research data management services in academic libraries in the US: a content analysis of libraries' websites [J]. College & Research Libraries, 2017, 78(7): 920－933.

作者单位:北京大学图书馆,北京,100871
收稿日期:2018 年 7 月 30 日

The Current Status and Development Trend of Data Driven Research

—Review of the First National University Data Driven Research Contest

Cui Haiyuan Luo Pengcheng Zhao Jingru Nie Hua Wang Jimin Zhang Jiuzhen

Abstract: Data-driven research has become new research paradigm. Data services have also become one of the most important services of academic library. Based on the review of the First National University Data Driven Research Contest, this paper analyzes the situation and trend of the Data Driven Research. By reviewing the contest and generalizing the model of data driven Research, it also provides the framework of Data Driven Service to Academic Universities. It is hoped to help academic libraries to promote the development of Data Driven Research as well as open access and research innovation in China.

Keywords: Data Driven Research; Data Service; Data Management; Digital Humanities; Big Data Research

(接第 92 页)

Research on the Archival Science of Peking University Library Science Special Course from 1949 to 1952

Guo Peng Han Juanjuan

Abstract: Through literature reviews, the paper clarifies the exact time when Peking University Library Science Special Course begin to carry out the modern archival science education and dig out the background of the establishment, content of the curriculum and so on. Peking University Library Science Special Course should be the starting point of China's modern archival science education.

Keywords: Library Science Special Course; Archival Science; Beginning