首届全国高校数据驱动创新研究大赛

1. 大赛介绍

随着大数据时代和数据密集型研究范式的到来，基于数据进行研究，对数据进行管理、共享和再利用，成为学术研究的新趋势。为了鼓励各学科领域学子基于数据进行创新研究，促进研究数据的保存和共享，由国家信息中心大数据发展部、北京市信息资源管理中心作为行业指导单位，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，联合北京大学中国社会科学调查中心、北京大学计算语言学研究所、重庆市仙桃大数据与物联网职业培训学院，面向全国高校在读学生，开展首届全国高校数据驱动创新研究大赛。

本次大赛将于 2017 年 12 月至 2018 年 4 月期间举行，欢迎各学科领域优秀学子提交论文参与竞赛。

大赛最新信息请参见官网 http://opendata.pku.edu.cn/competition-2018.xhtml。

1.1. 参赛对象

全国高校本科、硕士、博士在读学生。

1.2. 参赛形式和内容

数据驱动创新研究大赛要求包括：总体要求、论文要求、数据要求。

1.2.1. 总体要求

（1） 以 1~5 人组队报名(每人只能参与一支队伍)；

（2） 要求有指导教师；

（3） 研究内容为各学科领域相关学术问题；

（4） 需要基于数据进行研究，包含针对数据的相关分析和结论；

（5） 参赛成果提交的形式为研究论文，同时提供所使用研究数据；

（6） 入围决赛的参赛团队，要求参加现场答辩；

（7） 参赛者允许组织方对参赛作品汇集成册、展示和宣传,并可推荐发表。

1.2.2. 论文要求

（1） 研究内容需要具有一定的创新性；

（2） 论文字数在 8000~15000 之间；

（3） 论文格式需要遵循“全国高校数据驱动创新研究大赛-论文模板.doc”的要求；

（4） 参赛者允许提交的研究论文收录在北京大学机构知识库，论文在一定禁锢期后公开，不影响论文向期刊投稿发表。

1.2.3. 数据要求

使用的数据需要满足如下条件之一：

（1） 北京大学开放研究数据平台中的数据。

参赛团队可使用北京大学开放研究数据平台（http://opendata.pku.edu.cn）中数据，平台中包含社会科学、计算机、历史等学科领域的 100 多个数据，如中国家庭追踪调查、中国健康与养老追踪调查等。平台及数据介绍见第 4 节。

（2） 自己收集整理的、具有一定原创性的研究数据。

研究数据需要具有一定的原创性。即以为研究目的，自己收集整理了相关数据资源，对数据进行采集、清洗、预处理等加工步骤。数据的原创性将作为评分标准之一。例如，如下为具有一定原创性的研究数据：①为了研究微博用户行为而自己收集的微博博文数据；②为了研究大学生海洋意识而自己收集的调查问卷数据。

数据需要整理并提交至北京大学开放研究数据平台。对数据进行整理，并提供数据文档，说明数据的来源、采集和处理方法、数据格式及使用等。在成果提

交时，数据也需要提交至北京大学开放研究数据平台的首届全国高校数据驱动创新研究大赛数据空间（http://opendata.pku.edu.cn/dataverse/contest），即在该数据空间下创建一个新的数据集。在成果评审时，管理员将对数据进行审核，并公开发布。

研究数据需要遵循北京大学开放数据平台使用政策。提交的数据不应：侵犯他人或其他实体的专利权、商标权、商业秘密权、著作权、公开权或其他权利的内容；包含非法、威胁、辱骂、骚扰、诽谤、中伤、欺骗、欺诈、侵犯他人隐私、侵权、淫秽、攻击或亵渎性质的内容；非授权广告、推送广告、垃圾或批量电子邮件（俗称“垃圾邮件”）；包含软件病毒或任何其他计算机代码、文件或有意破坏、损害、限制或干扰任何软件、硬件或通讯设备正常功能的程序，或者意图破坏或非授权访问 PKU Opendata 或其他第三方系统、数据或其他信息的程序。

1.3. 赛程赛制

大赛的时间安排与组织形式如下：

（1） 启动与培训。时间：2017-11-30~2017-12-01。举行大赛启动与培训会，介绍大赛的基本情况和要求，介绍北京大学开放研究数据平台的使用及相关数据的情况。方式：现场培训与网络直播，详情见附录 1：

（2） 参赛报名。时间：2017-12-01~2018-01-15。参赛同学在大赛网站中组队报名，提交团队成员信息、指导教师、论文题目、简要介绍等。报名网址为：http://opendata.pku.edu.cn/registry-competition.xhtml。

（3） 成果提交。时间：2018-01-16~2018-02-28。参赛同学在大赛网站中提交研究论文，原创数据上传至北京大学开放研究数据平台。成果提交网址为：http://opendata.pku.edu.cn/registry-competition.xhtml。

（4） 成果评审。时间：2018-03-01~2018-03-22。对论文进行形式审查、专家评审，评审结果于 2018-03-22 在大赛官网公布。

（5） 现场答辩。时间：2018-04-03,具体时间待通知，地点北京大学。现场答辩，决出特等奖、一等奖、二等奖、三等奖。

（6） 海南颁奖。时间：2018 年 4 月中旬，特等奖、一等奖、二等奖获奖代

表，将受邀参加在海南陵水举办的“第二届京陵大数据峰会”，进行成果展示和颁奖。

2. 评审办法

参赛团队将分组评比，包括：本科生组、研究生组（含硕士、博士）。参赛团队类型由该团队中成员最高学历决定，即本科生组的队员均为本科生，研究生组的成员至少有一位是硕士或者博士。

（1） 形式审核。在研究成果征集阶段，主办方对提交作品进行形式审核，审核的标准包括：论文是否书写规范、是否基于数据进行了研究、数据是否符合要求、论文查重等，符合要求的成果进入书面评审。

（2） 书面评审。主办方邀请学科领域相关专家对成果进行评价，评价标准包括：论文成果的创新性、数据的原创性和规范性等。根据专家评分结果选择排名前 13 位的参赛团队进入决赛，并现场答辩，排名第14~133 位的参赛团队将获得优秀奖。

（3） 现场答辩。排名前 13 位的参赛团队需要进行现场答辩，由专家进行评审，决出特等奖、一等奖、二等奖、三等奖。如不参与答辩，视为放弃决赛资格，依次由排名第 14~133 位的团队进行替补。

3. 奖项设置

（1） 特等奖（1 组），奖金 20000 元

（2） 一等奖（1 组），奖金 10000 元

（3） 二等奖（3 组），奖金 5000 元

（4） 三等奖（8 组），奖金 3000 元

（5） 优秀奖（120 组），奖金 1000 元

4. 北京大学开放研究数据平台

（1） 平台简介

北京大学开放研究数据平台的由北京大学图书馆、国家自然科学基金-北京

大学管理科学数据中心、北京大学科研部、北京大学社科部联合主办和推出。平台以“规范产权保护”为基础，以“倡导开放科学”为宗旨，鼓励研究数据的发布、发现、再利用和再生产，促进研究数据引用的实践和计量，并探索数据长期保存，培育和实现跨学科的协同创新。

（2） 平台数据

北京大学开放研究数据平台现有 100 多个数据集，数据被 Web of Science 数据引用索引数据库收录。如下给出了一些典型的研究数据集：

中国家庭追踪调查，http://opendata.pku.edu.cn/dataverse/CFPS

中国健康与养老追踪调查，http://opendata.pku.edu.cn/dataverse/CHARLS

中国老年人健康长寿影响因素调查，http://opendata.pku.edu.cn/dataverse/CHADS

中国历代人物传记资料库，http://opendata.pku.edu.cn/dataverse/crach

北京社会经济发展年度调查，http://opendata.pku.edu.cn/dataverse/BAS

国家信息中心大数据发展部提供的数据，

http://opendata.pku.edu.cn/dataverse/contest_official

5. 组织单位

主办单位： 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

协办单位： 北京大学中国社会科学调查中心、北京大学计算语言学研究所、重庆市仙桃大数据与物联网职业培训学院

支持单位： 海南省陵水黎族自治县人民政府

行业指导单位： 国家信息中心大数据发展部、北京市信息资源管理中心

赞助单位： 圣智学习集团 Gale 公司、腾讯科技（北京）有限公司

数据支持单位： 北京国信宏数科技有限责任公司

北京清博大数据科技有限公司

北京麒麟心通网络技术有限公司

大连瀚闻资讯有限公司

中国电信股份有限公司云计算分公司

百职科技(北京)有限公司

广东和诚信息技术有限公司

6. 联系方式

大赛最终解释权归主办方所有。如果您对大赛有任何问题，可以通过邮箱、电话与我们联系。非常感谢您对大赛的关注与支持！

邮箱：data-research@lib.pku.edu.cn

电话：010-62751062-22

附录1 培训计划

（1） 第一次培训

时间：2017 年 11 月 30 日 下午 3:00~4:30

现场培训地点：北京大学图书馆 304 教室

网络直播地址：http://162.105.138.115/index.php?m=live&c=index&a=lists

录播地址：

http://162.105.138.115/index.php?m=content&c=index&a=show&catid=33&id=6641

表 1 第一次培训内容

主持人	主要内容	培训老师
刘雅琼（北京大学图书馆）	**大赛基本情况介绍**（30 分钟）：介绍大赛的基本情况，包括大赛要求、赛制赛程、注册和成果提交流程、北京大学开放数据平台等。	**罗鹏程 馆员（北京大学图书馆）** 北京大学图书馆信息化与数据中心馆员，负责北京大学开放研究数据平台的建设工作，曾参与国家自然科学基金委基础研究知识库、北京大学科研管理系统等平台的建设。参与负责本次大赛的相关组织工作。
	数据挖掘方法介绍（30 分钟）：简要介绍数据挖掘的基本流程和方法。	**王继民 教授（北京大学信息管理系）** 教授，博士生导师，北京大学信息管理系副主任。研究领域包括：搜索引擎、Web 数据挖掘、科学评价学、信息可视化等。近几年主持国家社科基金、国家“核高基”重大科技专项子课题、以及国家发改委、教育部、北京市科委等科研课题 30 余项。发表学术研究论文 50 余篇；出版专著或合著《搜索引擎原理技术与系统》、《Web 用户查询日志挖掘与应用》、《中国人文社科类一级学科数据分析报告》、《“一带一路”沿线国家五通指数报告》、《国民海洋意识发展指数研究报告（2016）》等 6 部。获得发明专利 2 项；获得省部级科研奖励 2 项。
	现场答疑（30 分钟）	

（2） 第二次培训

时间：2017 年 12 月 01 日 下午 3:30~5:00

现场培训地点：北京大学图书馆 304 教室

网络直播地址：http://162.105.138.115/index.php?m=live&c=index&a=lists

录播地址：

http://162.105.138.115/index.php?m=content&c=index&a=show&catid=33&id=6659

表 2 第二次培训内容

主持人	主要内容	培训老师
赵飞（北京大学图书馆）	**中国家庭追踪调查及分析方法**（30 分钟）：对中国家庭追踪调查数据（CFPS）进行介绍，并简要介绍相关的分析方法。	**吴琼 副研究员（北京大学社会科学调查中心）** 美国宾州州立大学教育与心理测量学博士、统计学硕士。现任北京大学中国社会科学调查中心副研究员，“中国家庭追踪调查”（CFPS）项目办公室主管，主要负责 CFPS 数据管理、数据服务、与问卷设计和执行相关的数据支持工作。加入调查中心之前，她就职于哈佛大学人口与发展研究中心，作为该中心的量化分析师，她的主要职能之一是分析大型调查数据。主要研究领域包括测量学方法、认知功能的影响因素、少儿发展等，已发表 SSCI、SCI 论文 20 余篇。
	中国健康与养老追踪调查及分析方法（30 分钟）：对中国健康与养老追踪调查数据（CHARLS）进行介绍，并简要介绍相关的分析方法。	**陈欣欣 副研究员（北京大学社会科学调查中心）** 浙江大学管理学博士，现任北京大学中国社会科学调查中心副研究员，中国健康与养老追踪调查（CHARLS）项目主管，曾在斯坦福大学师从 Scott Rozelle 教授从事博士后研究。2008 年以来参与了 CHARLS 的实地执行工作，并组织实施了中国中老年人生命历程调查、CHARLS 第三轮追踪调查和共和国初期基层经济史调查。研究兴趣集中在微观发展经济学和老年经济学。
	国家信息中心大数据发展部数据介绍（30 分钟）：介绍国家信息中心大数据发展部的开放数据。	**廖尚围 项目经理（国信宏数公司）** 国信宏数公司数据采集项目经理。曾任蓬天公司 CTO，负责陕西省、江西省地税征管系统技术架构，具有丰富的 J2EE 项目开发实施经验。目前主要负责国信宏数公司数据采集工作，通过设

		计分布式采集平台，实施互联网结构化、非结构化数据的采集、清洗、存储。