第二届全国高校数据驱动创新研究大赛

1. 大赛介绍

随着大数据和人工智能时代，以及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研究，创新性地利用新方法、新技术分析发掘数据潜在价值，促进数据的流通和共享，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、北京大学中国社会科学调查中心、重庆西部大数据前沿应用研究院主办，中国高等教育文献保障系统、重庆大学图书馆协办，面向高校、研究院（所）在读学生，开展数据驱动创新研究大赛。

大赛于 2018 年 11 月至 2019 年 4 月期间举行，欢迎各学科领域优秀学子提交作品参与竞赛。

大赛最新信息请参见官网（http://opendata.pku.edu.cn/competition-2019.xhtml）和微信公众号“第二届数据驱动创新研究大赛”。

大赛官网

微信公众号

1.1. 参赛对象

全国高校、研究院（所）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(每人最多只能参与2支队伍，且最多只能作为1支队伍的第一作者)；

（2） 要求有指导教师；

（3） 需要基于数据进行研究，包含针对数据的相关分析和结论；

（4） 参赛成果提交的形式为研究论文，同时提供所使用的研究数据；

（5） 入围决赛的参赛团队，要求参加现场答辩；

（6） 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；

（7） 满足如下之一选题要求

a) 不限主题：各学科领域相关学术问题；

b) 限定主题：选择以下给定专题之一进行研究，专题详情及附加要求见附录1“ 专题选题”，如下为专题列表。

专题一 基于“中国家庭追踪调查”（CFPS）的数据发现和挖掘

专题二 社会经济调查的职业和行业自动编码模型构建

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，可从大赛官网下载；

（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）中的数据，平台中包含社会科学、计算机、历史等学科领域的200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。平台及数据介绍见附录2。

（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“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研究数据平台使用政策。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北京大学开放研究数据平台或其他第三方系统、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） 培训与讲座。时间：2018-11-19~2018-11-23。举行大赛培训，介绍大赛的基本情况和要求，同时举办数据相关的讲座。方式：现场培训与网络直播，详情见附录 3。

（2） 参赛报名。时间：2018-11-20~2019-01-10。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：2019-01-11~2019-03-17。参赛同学在大赛网站中提交研究论文，原创数据、源代码（如果选择专题）需要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：2019-03-18~2019-04-10。对论文进行形式审查、专家评审，评审结果于 2019-04-11 在大赛官网公布。

（5） 现场答辩。时间：2019 年 4 月下旬，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。

（6） 赛后活动。时间：2019 年 4 月起，组委会将围绕大赛成果开展相关活动，提升作品的影响力。如：论文推荐发表、论文转写为数据新闻等，后续活动详情见大赛官网通知。

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、专题中算法模型的效果等。根据专家评分结果选择排名前 8 位的参赛团队进入决赛，并现场答辩，排名第 9~16 位获得三等奖，其他排名靠前的参赛团队将获得优秀奖。其中，不限主题和限定主题的获奖名额根据作品比例和质量确定。

（3） 现场答辩。排名前 8 位的队伍，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。如不参与答辩，视为放弃决赛资格，按排名依次替补。

3. 奖项设置

（1） 特等奖：奖金 20000 元，1 组

（2） 一等奖：奖金 10000 元，2 组

（3） 二等奖：奖金 5000 元，5 组

（4） 三等奖：奖金 3000 元，8 组

（5） 优秀奖：奖金 1000 元，若干组，不少于成功提交作品参赛队伍的 30%

4. 组织单位

主办单位：北京大学图书馆、北京大学信息管理系、北京大学中国社会科学调查中心、重庆西部大数据前沿应用研究院

协办单位：中国高等教育文献保障系统、重庆大学图书馆

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：企研数据（杭州古德科技有限公司）

数据支持单位：北京国信宏数科技有限责任公司、企研数据（杭州古德科技有限公司）、成都数联铭品科技有限公司、同方知网（北京）技术有限公司、重庆泛语科技有限公司

5. 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，感谢您对大赛的关注与支持！

邮箱：data-research@lib.pku.edu.cn

电话：张老师 010-62753907

附录1 专题选题

专题一：基于“中国家庭追踪调查”（CFPS）的数据发现和挖掘

分主题 1 ：预测家庭样本的流失。参赛者在 CFPS 2016 年发布的家庭关系库（数据下载地址为：http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/45LCSO）中近 15000 个 fid16 中选出 1000 个最有可能在 2018 年流失的家庭。CFPS 2018 实地工作结束后我们根据执行的最后结果选出命中率最高的参赛作品。

分主题 2 ：收入数据的插补。由于收入数据较为敏感，在抽样调查中会出现一定比例的缺失情况。参赛者针对缺失以及可疑的收入数据提出插补方案并给出插补结果。我们将组织相关方面专家对方案的合理性以及最终结果进行评估。

以上两个主题均不限研究方法，传统的统计模型或机器学习方法均可。

附加要求：①需要提交参赛源代码至北京大学开放数据平台，代码需要为 Python、R 或其他编程类语言代码；②需要有说明文档描述代码的运行环境和使用方法；③代码结构清晰，有适当的注释。

专题二：社会经济调查的职业和行业自动编码模型构建

社会经济调查中通常会采集职业和行业信息，为方便数据用户使用这些信息，一般会事先基于国家标准化管理委员会发布的《职业分类与代码》对上述信息进行编码。组委会将在竞赛平台上提供部分社会经济调查中采集得到的职业和行业的具体描述信息，以及相应的已经编码成功的代码。要求参赛者基于上述数据（数据下载地址为：http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/PEMXPX），构建自动编码模型。组委会将利用该模型，应用于其他已人工编码成功的数据。基于模型预测的准确度，评判模型的优劣。

附加要求：①需要提交参赛源代码至北京大学开放数据平台，代码需要为 Python、R 或其他编程类语言代码；②需要有说明文档描述代码的运行环境和使用方法；③代码结构清晰，有适当的注释。

附录2 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 200 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

附录3 培训与讲座

直播地址(培训和讲座都通过以下两种方式直播,培训结束后在官网提供录播):

直播地址 1(北京大学多媒体平台):

http://162.105.138.115/index.php?m=live&c=index&a=lists

直播地址 2(目睹平台):

见各次培训右侧二维码

(1) 第一次:大赛培训

时间: 2018 年 11 月 19 日 上午 9:00~11:00

现场培训地点: 北京大学图书馆 304 教室

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。
	数据预处理方法(30 分钟):介绍数据预处理的一般技术与方法,主要包括数据清理、数据集成、数据变换、数据	**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

	归约、数据离散化等。	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。
	北京市政务数据资源网开放数据介绍（30 分钟）：对北京市政务数据统一开放平台——北京市政务数据资源网所开放的数据进行介绍	**高文飞 高级工程师（北京市信息资源管理中心）** 北京市信息资源管理中心数据开放部项目主管，北京市大数据工作推进小组办公室成员，长期致力于政务数据资源管理及应用相关研究和实践工作，目前主要负责政务数据开放、数据汇聚共享、大数据应用等工作。
	现场答疑（30 分钟）	

（2） 第二次：数据讲座

时间：2018 年 11 月 19 日 下午 14:00~15:00

现场培训地点：北京大学图书馆 304 教室

主要内容	讲座老师
人文社科数据管理与服务的研究与实践（60 分钟）：介绍国内外人文社会科学数据管理与服务，以及所参与政府和高校科学数据中心建设的一些数据平台，如上海人口数据实验室、复旦大学人文社会科学数据研究所、上海市慧源共建共享平台、国家卫计委流动人口数据平台等等，并展示一些项目的数据可视化。	**殷沈琴 副研究馆员（复旦大学人文数据研究所）** 复旦大学人文社会科学数据研究所科学数据中心主任、硕士生导师，上海人口数据实验室副主任、“开放数林”政府开放数据专家委员会委员。主要从事科学数据管理、政府开放数据、社会管理与社会政策等领域的研究。承担和参与十余个国家级和省部级的项目，并多次负责数据平台的规划和部署实施工作，有丰富的数据研究和落地实践经验。

（3） 第三次：数据讲座

时间：2018 年 11 月 19 日 晚上 19:00~20:00

现场培训地点：北京大学图书馆 304 教室

主要内容	讲座老师
科研常用开放数据资源的查找与获取（60 分钟）：介绍网络上开放获取的各类数据资源，并结合具体案例介绍查找各类数据资源的技巧。	**朱玲 馆员（北京大学图书馆）** 北京大学图书馆信息化与数据中心数据管理服务主管，负责北京大学图书馆多个系统平台项目：资源发现系统“未名学术搜索”、电子资源导航系统、北京大学开放研究数据平台、电子资源使用监控与统计系统等。已发表图情类 CSSCI 论文 9 篇。

（4） 第四次：大赛培训

时间：2018 年 11 月 21 日 上午 10:00~11:20

现场培训地点：北京大学图书馆 304 教室

主持人	主要内容	培训老师
张元俊（北京大学图书馆）	**中国家庭追踪调查介绍**（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介绍（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，现任北京大学中国健康与养老追踪调查（CHARLS）执行主任，相继组织实施了中国中老年人生命历程调查、共和国初期基层经济史调查、京津居民家庭结构和生命历程调查、CHARLS 第三轮和第四轮常规追踪调查。研究兴趣集中在微观发展经济学和老年经济学。
	专题介绍（20 分钟）：对专题一和专题二进行详细解读和介绍。	**吴琼 副研究员（北京大学社会科学调查中心）** **赵银霞 软件工程师（北京大学社会科学调查中心）** 毕业于郑州大学计算机科学与技术专业，目前从事 PHP 软件开发，WEB 前端开发。

（5） 第五次：大赛培训+数据讲座

时间: 2018 年 11 月 23 日 上午 8:30~10:00

现场培训地点: 阿卜杜勒·阿齐兹国王公共图书馆北京大学分馆地下一层报告厅

主要内容	培训（讲座）老师
【培训】**中国创新创业数据库介绍**（30 分钟）：对中国企业创新创业数据库进行介绍，该数据库集合了微观企业在创业、投资，以及企业创新产出等方面的信息。	**杨奇明 董事长（古德科技有限公司）** 杨奇明拥有浙江大学管理学博士学位，北京大学应用经济学博士后。主持博士后基金项目和国家社科青年项目，在《经济研究》、《管理世界》等学术期刊发表论文 10 余篇。自北京大学从事博士后研究期间（2013.7-2016.1）开始，一直从事企业大数据的融合与开发工作，协助合作导师组织构建了朗润-龙信创新创业指数，作为联合主编出版了《中国区域创新创业报告》。
【讲座】**Making your data count! - Why you need to care about sharing data and**	**Arend Küster / 柯安德（Springer Nature 开放科研大中**

how we can help you（60 分钟）：Following the recent Research Data Survey to find out more about Data Management and attitudes, we will introduce the importance of data for research purposes, why efficient data management is important for research to have an impact and how we can provide help.	**华区总监）** Arend 负责 Springer Nature 旗下所有开放获取业务在中国的发展，包括 Nature Communications、BMC 系列期刊、Springer Open 系列期刊等，他在科技和医学出版领域拥有非常丰富的经验，致力于促进中国学者与国际出版业科研理念的交流。 **Grace Baynes（Springer Nature 研究数据与新产品研发副总裁）** Grace 在 Springer Nature 领导着开放数据与优秀数据实践事业。她还负责广受赞誉、高速发展的 Nature 旗下期刊 Scientific Data 的出版。 她还是国际开放科研政策与实践的专家。最近，Grace 领导着开放数据团队，与 Wellcome Trust 和德国马普学会签署了开拓性的合作。