中国高校计算机大赛

第二届“中国高校计算机大赛-大数据挑战赛”（2017 年）

通　　知

“中国高校计算机大赛-大数据挑战赛”（Big Data Challenge）由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，旨在通过竞技的方式提升在校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题。

2017 年“大数据挑战赛”由清华大学和腾讯安全平台部联合承办，面向全球高校在校生，每队可有指导教师一名。竞赛数据以及处理的评价指标由大赛提供，参赛队伍通过设计算法进行数据分析与处理。以在线评测和专家评审相结合的方式进行评比。

请各校积极配合，按照通知和大赛章程做好组织工作，并在指导教师工作量认可及参赛队伍经费等相关方面给予支持。竞赛详情请登录“大数据挑战赛”网站（http://bdc.saikr.com/bdc）查询。

附件 1：2017“大数据挑战赛”规程
附件 2：2017“大数据挑战赛”组织机构名单
附件 3：赛题介绍——轨迹模式识别

教育部高等学校计算机类专业教学指导委员会
教育部高等学校软件工程专业教学指导委员会
教育部高等学校大学计算机课程教学指导委员会
全国高等学校计算机教育研究会（代章）
2017 年 4 月

附件 1：2017“大数据挑战赛”规程

2017“中国高校计算机大赛-大数据挑战赛”（Big Data Challenge）是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，清华大学和腾讯安全平台部联合承办，在腾讯 DIX 平台上开展的高端算法竞赛。大赛面向全球高校在校生开放，旨在通过竞技的方式提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校导师参与指导。

本次大赛以某人机验证产品采集的鼠标轨迹脱敏数据为基础，期望参赛队伍通过算法设计和数据分析，检测发现其中的机器轨迹，具体赛题描述见附件 3。比赛结果按照规定的评价指标使用在线评测程序进行评阅和排名，结果最优者获胜。

一、大赛组织

1. 主办单位

教育部高等学校计算机类专业教学指导委员会
教育部高等学校软件工程专业教学指导委员会
教育部高等学校大学计算机课程教学指导委员会
全国高等学校计算机教育研究会

2. 承办单位

清华大学

3. 赞助单位

腾讯安全平台部

二、赛制说明

本次大赛分为初赛、复赛和决赛三个阶段，其中初赛由参赛队伍下载数据在本地进行算法设计和调试，并通过大赛报名官网提交结果文件；复赛要求参赛者在腾讯 DIX 平台进行数据分析和处理，可使用基于 Spark、XGBoost 及平台提供的机器学习相关基础算法；决赛要求参赛者进行现场演示和答辩。具体安排和要求如下：

1. 初赛（5 月 26 日—7 月 21 日）

- 参赛队伍可从大赛官方网站下载数据，在本地进行算法设计和调试，规

定时间内在报名官网提交结果，每支队伍在一天内只能提交一次结果。

- 5 月 26 日起系统向选手开放训练样本数据 3000 条（2600 条正常轨迹样本，400 条机器轨迹样本），供参赛者下载进行建模和模型优化，同时提供正式比赛数据 10 万条供参赛者下载评测。
- 每天 10:00 AM 进行一次评测，根据参赛队伍目前为止最优成绩进行排名展示。
- 初赛截止时间（7 月 21 日 10:00AM），排名前 200 名的队伍进入复赛。

2. 复赛（7 月 22 日—8 月 11 日）

- 所有比赛数据不可下载，选手须在腾讯 DIX 平台完成数据处理、建模、算法调试、产出结果等所有环节，可使用基于 Spark、XGBoost 及平台提供的机器学习相关基础算法。
- 7 月 22 日起系统提供 200 万条正式比赛数据（对参赛选手不可见，仅供平台对参赛作品进行评测）。
- 每天 10:00 AM 按照评测指标进行一次评测，并根据参赛队伍目前为止最优成绩进行排名展示。
- 排名前 10 名的队伍将受邀参加决赛答辩会。

3. 决赛答辩（8 月 20 日）

- 决赛将以现场答辩会的形式进行，具体安排另行通知。
- 参赛队伍应提前准备现场答辩材料，包括 PPT、算法代码等。
- 组委会将根据参赛队伍的算法原理、历史成绩和评委打分，评选出整个大数据挑战赛的冠亚季军，并现场颁发奖金及证书。

三、参赛对象

本次大赛面向在校学生（包括高职高专、本科、研究生及以上），具体参赛队伍要求如下：

1. 可以单人参赛或自由组队（最多不超过 3 人，可以跨单位组队）。
2. 每人只能参加一支队伍。
3. 保证参赛队员报名信息准确有效，否则将被取消参赛资格及奖励，平台将实行实名认证。
4. 大赛主办单位和技术支持单位中有机会接触赛题相关数据的人员不允许参赛。
5. 提交的参赛作品必须是团队或个人独立完成的原创作品，不得抄袭，不得违反任何相关的法律法规，否则将取消参赛资格。
6. 大赛所提供的数据仅限于此次大赛使用，不得用于其他任何目的。若因违反此规定而给数据提供方或平台提供方造成损失的，参赛队伍所在单位和选手须承担全部责任。

四、报名方式

1. 报名方式：登录大赛官网，完成个人信息注册，即可报名参赛。
2. 报名、组队变更和实名认证截止时间均为2017年6月30日10:00 AM。
3. 大赛官方交流QQ群：628131690。

五、奖项设置

1. 初赛奖项

- **明日之星奖：**排名前 10 名的参赛队伍将获得“腾讯方舟计划明日之星奖”，颁发获奖证书，奖金2000元/队。
- **优胜奖：**排名前 11~50 名的参赛队伍将获得“腾讯方舟计划优胜奖”，颁发获奖证书和腾讯方舟计划定制礼品一份。

说明：上述奖项由平台根据测试结果计算排名而产生。

2. 复赛奖项

- **腾讯方舟计划夏令营：**排名前 10 名的参赛队伍将会受邀到深圳腾讯总部参加为期一周的腾讯方舟计划夏令营活动（往返交通费自理、食宿及其他费用均由腾讯承担）。
- **招聘绿色通道：**排名前 20 名的队伍可直接入围腾讯校招终面（即招聘流程省略简历筛选及笔试筛选阶段，直接进入面试，在校期间均有效）。
- **优秀导师奖：**排名前 10 名参赛队伍的指导教师将获得“优秀导师”证书及奖品一份。

说明：上述奖项由平台利用测试数据检测算法，并根据测试结果排名产生。

3. 决赛奖项

- **冠军：**1支队伍，奖金壹拾万元，颁发获奖证书。
- **亚军：**1支队伍，奖金伍万元，颁发获奖证书。
- **季军：**1支队伍，奖金贰万元，颁发获奖证书。

说明：上述奖项将结合参赛队伍的答辩PPT、算法原理和历史成绩综合评审而产生。答辩将在清华大学深圳研究生院举行，参加决赛队伍的往返交通费用自行承担，住宿及餐饮费用由腾讯方舟计划统一安排。

4. 周星星

自大赛公布排行榜之日起，每周榜单排名前3名的参赛队伍将成为周星星，获得腾讯方舟计划纪念公仔。

六、其他

竞赛组织委员会对本规程所有内容拥有最终解释权。

附件 2：2017“大数据挑战赛”组织机构名单

一、 竞赛指导委员会

主　任：杜小勇（中国人民大学，教育部高等学校大学计算机课程教学指导委员会副主任）
副主任：杨　勇（腾讯公司）
委　员：侯义斌（北京工业大学，教育部高等学校软件工程专业教学指导委员会副主任）
　　陈新河（中关村大数据产业联盟）
　　骆　斌（南京大学，教育部高等学校软件工程专业教学指导委员会副主任）

二、 竞赛专家委员会

主　任：王建民（清华大学）
副主任：臧斌宇（上海交通大学）
　　胡　珀（腾讯公司）
委　员：陈恩红（中国科学技术大学）
　　胡学钢（合肥工业大学）
　　李雁翎（东北师范大学）
　　滕桂法（河北农业大学）
　　王宏志（哈尔滨工业大学）
　　吴中海（北京大学）
　　吴黎兵（武汉大学）
　　肖　侬（国防科学技术大学）
　　于　炯（新疆大学）
　　张瑞生（兰州大学）
　　张玉志（南开大学）

三、 竞赛组织委员会

主　任：刘　强（清华大学，教育部高等学校软件工程专业教学指导委员会秘书长）
副主任：甘　祥（腾讯公司）
委　员：陈　梦（腾讯公司）
　　洪　玫（四川大学）
　　舒　坚（南昌航空大学）
　　王树良（北京理工大学）
　　袁　坤（腾讯公司）
　　杨永健（吉林大学）
　　张　莉（北京航空航天大学）
　　赵文耘（复旦大学）
　　左保河（华南理工大学）

附件 3：赛题介绍——轨迹模式识别

【赛题描述】

鼠标轨迹识别当前广泛运用于多种人机验证产品中，不仅便于用户的理解记忆，而且极大增加了暴力破解难度。但攻击者可通过黑产工具产生类人轨迹批量操作以绕过检测，并在对抗过程中不断升级其伪造数据以持续绕过同样升级的检测技术。本次大赛期望用机器学习算法来提高人机验证中各种机器行为的检出率，其中包括对抗过程中出现的新的攻击手段的检测。

【比赛数据】

本题目数据来源于某人机验证产品采集的鼠标轨迹，经过脱敏处理，数据分为 3 部分（数据量分别为 3000 条、10 万条、200 万条）。

赛事分为三个阶段（初赛、复赛、决赛答辩）：5 月 26 日起，初赛提供 3000 条数据作为训练样本，供参赛者下载进行建模和模型优化，同时提供 10 万条正式比赛数据供下载评测，识别结果为初赛得分；复赛提供 200 万条比赛数据（不可下载，数据不可见，仅供评测），识别结果为复赛得分；决赛将以现场答辩会的形式进行。

训练数据：

训练数据表名称：dsjtzs_txfz_training

字段	类型	解释
a1	bigint	编号 id
a2	string	鼠标移动轨迹(x,y,t)
a3	string	目标坐标(x,y)
label	string	类别标签：1-正常轨迹，0-机器轨迹

测试数据：

初赛测试表名称：dsjtzs_txfz_test1

复赛测试表名称：dsjtzs_txfz_test2

字段	类型	解释
a1	bigint	编号 id
a2	string	鼠标移动轨迹(x,y,t)
a3	string	目标坐标(x,y)

【测评标准】

选手将识别为机器行为的编号id提交到计算平台 ,需要提交的结果表只包含一个字段：编号 id。

- 初赛提交表名：dsjtzs_txfzjh_preliminary
- 复赛提交表名：dsjtzs_txfzjh_semifinal

设定 Precision 为 P，Recall 为 R，白样本为正常轨迹，黑样本为机器轨迹。其中：

- P = 判黑的数据中真正为黑的数量/判黑的数据总量
- R = 判黑的数据中真正为黑的数量/真实黑数据总量

例如对于 10w 条数据，其中 8w 条为白样本，2w 条为黑样本，参赛者将 1w 条数据判断为黑样本(其中真正的黑样本有 8000 条 ,2000 条白样本被错误判黑),则 P=8000/10000 = 80%；R=8000/20000=40%。

参赛队伍最终得分 F = 5PR/(2P+3R)*100。

最终排名按照 F 值评判，F 值越大，代表结果越优，排名越靠前。