中国高校计算机大赛

中国高校计算机大赛 —— 2020 华为云大数据挑战赛

通　知

2016 年，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联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），第五届（2020 年）“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018 年被选入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2020 中国高校计算机大赛——华为云大数据挑战赛（以下简称“大赛”）是由清华大学、中国人工智能学会和华为技术有限公司联合举办，华为云和北京信息科学与技术国家研究中心提供支持，以企业真实场景和实际数据为基础，面向全球开放的高端算法竞赛。大赛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。

本次大赛面向全球高校在校学生，鼓励高校教师参与指导。参赛队伍需要根据赛题要求设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

请各学校积极配合，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方面给予支持。竞赛详情见附件（2020 华为云大数据挑战赛竞赛规程）。

全国高等学校计算机教育研究会

2020 年 4 月 12 日

2020 中国高校计算机大赛——华为云大数据挑战赛
竞赛规程

2016 年，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联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），第五届（2020 年）“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018 年被选入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2020 中国高校计算机大赛——华为云大数据挑战赛（以下简称“大赛”）是由清华大学、中国人工智能学会和华为技术有限公司联合举办，华为云和北京信息科学与技术国家研究中心提供支持，以企业真实场景和实际数据为基础，面向全球开放的高端算法竞赛。大赛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。

一、参赛对象

本次大赛开设在校学生和在职人员两个通道，其中在职人员不占用在校学生晋级和奖励名额，具体见后面小节说明。

1. 在校学生队伍要求

大赛面向中国及境外在校学生（包括高职高专、本科、研究生），具体要求如下：

- 可以自由组队参赛，每个参赛队伍人数可为 1–3 人，参赛队员必须全部为在校生，允许跨年级、跨专业、跨校组队。
- 每人只能参加一支队伍（即个人参赛后不可再与他人组队参赛，或个人参加一个队伍后不可再参加另一个队伍），每支队伍允许最多有一名指导老师，指导教师须为在职高校教师。
- 参赛者报名时应具有在校学籍，已毕业的学生不具备参赛资格。
- 参赛选手应保证报名信息准确有效，报名时应在大赛报名页“备注”栏提供学信网的教育部学籍在线验证报告编号。

2. 在职人员队伍要求

在职人员是指在单位工作的非在校生，具体要求如下：

- 可以自由组队参赛，每个参赛队伍人数可为 1–3 人，队伍中只要有一名在职人员，则整个队伍将视为在职人员队伍。
- 每人只能参加一支队伍，参赛选手应保证报名信息准确有效。

3. 禁止参赛人员

- 大赛主办和技术支持单位如有机会接触赛题和相关数据的人员不允许参赛。
- 华为技术有限公司公司在职人员（不含实习生）不允许参赛。

二、赛制说明

本次大赛分为报名&组队、热身赛（可选）、正式赛（包括初赛、复赛和决赛）等阶段，其中热身赛的目的是帮助选手进行前期训练学习以及熟悉华为云大数据平台，由选手自愿选择报名并在华为云平台进行比赛；初赛均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；复赛要求参赛者在华为云大数据平台上进行数据分析和处理，可使用平台提供的计算资源和工具包；决赛要求参赛者进行现场演示和答辩。

1. 报名&组队（4 月 17 日 – 5 月 31 日）

参赛选手须在大赛官网报名并且组队参赛（即使单人参赛也要组建单人队伍），大赛不收取任何报名费用。

- 大赛报名系统开放时间是 2020 年 4 月 17 日 10:00，截止时间是 2020 年 5 月 31 日 17:00。
- 大赛官网：https://competition.huaweicloud.com/information/1000037843/bdc2020
- 大赛官方交流 QQ 群：566353409 / 753413531 / 758344321

报名截止之后，不再允许添加或更改指导教师和任何队伍成员。如有中途退出情况，只允许在参赛队伍内部更换队长或删除队员。参赛队伍须应在决赛开始前向大赛组委会提交成员更换申请，由参赛队伍全部成员亲笔签名，经由大赛组委会审核后变更生效。

2. 热身赛（4 月 17 日 – 6 月 1 日，自愿选择参加）

热身赛要求选手基于华为云一站式 AI 开发平台 ModelArts 开发模型及提交评测，并使用华为云对象存储服务 OBS 以存储训练数据、代码、模型等文件。

热身赛的开始时间是 4 月 17 日 10:00，结束时间是 6 月 1 日 17:00，每个参赛队伍每天可以进行 3 次提交，系统立即进行评测并返回成绩。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示，实时更新排行榜。

特别说明：

- 热身赛的赛题为“交通流量预测”，具体描述大赛官网“热身赛题”栏目。
- 热身赛排名不区分在校学生队伍和在职人员队伍。
- 热身赛由选手自愿选择报名参加，其成绩不会影响正式赛成绩；选手也可以直接报名参加正式赛，最终大赛成绩以正式赛为准。

3. 初赛（6 月 2 日 – 6 月 30 日）

参赛队伍可从大赛官方网站下载数据，在本地进行算法调试，并在线提交结果。

正式赛的赛题为“船运到达时间预测”，具体描述见大赛官网“正式赛题”栏目。

初赛 A 阶段：6 月 2 日 10:00 – 6 月 26 日 22:00，每个参赛队伍每天可以进行 3 次提交，系统立即进行评测并返回成绩。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示，实时更新排行榜。

初赛 B 阶段：6 月 27 日 10:00 – 6 月 30 日 12:00，系统将在 6 月 26 日 23:00 更换测试数据，参赛队伍需再次下载数据文件。初赛排行榜将选取参赛队伍 6 月 27 日起产生的成绩进行重新排名。

初赛截止时间是 6 月 30 日 12:00，最多有 110 支队伍晋级复赛，晋级规则如下：

（1） 100 支在校学生队伍：单独进行排名之后，初赛成绩排名前 100 名队伍将进入复赛，同时前 10 名队伍还将获得初赛奖励；

（2） 10 支在职人员队伍：初赛成绩在所有队伍（包括在校学生队伍和在职人员队伍）的总排名中进入前 100 名，并且处于在职人员队伍的前 10 名；同时进入总排名的前 10 名队伍还将获得初赛奖励。

4. 复赛（7 月 13 日-8 月 10 日）

复赛参赛队伍需要在华为云大数据平台上完成数据处理、建模、算法调试、生成结果等，所有比赛数据不可下载，可使用平台提供的计算资源和工具包。

复赛 A 阶段：7 月 13 日 10:00 – 8 月 6 日 22:00，每个参赛队伍每天可以进行 3 次提交，系统立即进行评测并返回成绩。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示，实时更新排行榜。

复赛 B 阶段：8 月 7 日 10:00 – 8 月 10 日 12:00，系统将在 8 月 6 日 23:00 更换测试数据，参赛队伍应根据新的数据集提交模型（在线一键提交）。复赛排行榜将选取参赛队伍 8 月 7 日起产生的成绩进行重新排名。

复赛截止时间是 8 月 10 日 12:00，最多有 12 支队伍晋级决赛，晋级规则如下：

（1） 10 支在校学生队伍：单独进行排名之后，复赛成绩排名前 10 名队伍将进入复赛；

（2） 2 支在职人员队伍：复赛成绩在所有队伍的总排名中进入前 10 名，并且处于在职人员队伍的前 2 名；

（3） 所有晋级决赛队伍的代码需要进行审核，审核不通过的队伍将取消决赛资格，其名额顺延到其后面的下一个队伍。

5. 决赛（8 月下旬）

决赛将以现场答辩会的形式进行，晋级决赛团队需提前准备答辩材料，包括答辩 PPT、参赛总结、算法核心代码。

- 在答辩现场，每支队伍面对评委有 15 分钟的陈述时间和 10 分钟的问答时间。评委将根据选手的技术思路、理论深度和现场表现进行综合评分。
- 决赛分数将根据参赛队伍的算法成绩和答辩成绩加权得出，评分权重为复赛 B 阶段 70%，决赛答辩 30%。

决赛地点和时间安排另行通知，受邀参加决赛的选手在决赛期间的食宿由大赛组委会安排，往返交通费及其他费用自理。

三、奖项设置

1. 热身赛奖项

大赛将提供 100 元华为云资源代金券，成功报名的参赛者可点击页面上方“领取”获得代金券（每位参赛者仅可领取一次）。另外，热身赛将设置以下奖项和奖品：

奖项名称	数量	对象	奖励办法
一等奖	5	热身赛前 5 名队伍	荣耀魔法系列手表
二等奖	15	热身赛 6-20 名队伍	华为 AI 音箱
三等奖	30	热身赛 21-50 名队伍	荣耀手环 5i 标准版

2. 初赛奖项

初赛最终排行榜的前 20 名队伍颁发初赛名次证书，初赛名次奖不区分在校学生队伍和在职人员队伍。

另外，对于初赛排行榜前 20 名中使用华为云一站式 AI 开发平台的参赛队伍，经过对其代码、模型和结果进行审核确认后，还将同时获得华为云颁发的“大数据之星”奖励证书。

3. 复赛与决赛奖项

大赛奖金池总额为 30 万元人民币，所有奖金均为税前金额。

奖项名称	数量	对象	奖励办法
一等奖	3	在校学生队伍 决赛前 3 名	证书以及奖金： - 第 1 名奖金 15 万元 - 第 2 名奖金 5 万元 - 第 3 名奖金 2 万元
二等奖	7	在校学生队伍 决赛 4-10 名	证书，奖金 1 万元

名次奖	2	在职人员队伍	证书，奖金 5000 元
三等奖	20	在校学生队伍 复赛 11-30 名队伍	证书

另外，入围决赛参赛队伍的指导教师获得优秀指导教师奖，颁发获奖证书。

4. 周周星

自大赛公布排行榜之日起，正式赛的每周榜单排名前三名的参赛队伍将获得周周星。周周星以每周五中午 12 点的评分为准，取前三名，发放精美纪念礼品；对于前面已经获得周周星的队伍，不重复发放，名额按名次顺延。

5. 其他激励

招聘绿色通道：复赛排名前 50 的在校学生队伍可直接获得华为校招面试直通卡（即招聘省略简历筛选及笔试筛选阶段，直接进入面试阶段，2021 年底前有效）。

四、违规处理

参赛者应本着诚实、公平的态度参加比赛，如在以下情况出现违规，大赛组织委员会（简称“大赛组委会”）有权取消参赛者所在队伍的参赛资格，情节严重者将通报参赛者所在高校并追究其违法责任。

1. 账号使用：参赛者有义务保证账号信息的真实性和有效性，且账号仅限于参赛者本人使用；参赛者禁止使用多账号参赛，同一参赛者不可使用多个账号进行提交、刷分操作；如根据判断认为参赛账号存在异常或违背正常使用条例，组委会可以单方面暂停或终止该账号登录大赛平台。
2. 比赛成果：严禁参赛队伍之间相互抄袭。如不同参赛队伍提交结果高度相似，经判定存在抄袭行为的，组委会将取消相关参赛队伍的参赛资格，相关参赛成绩无效。另外，参赛者应保证其在比赛过程中所产出的所有成果未侵犯任何第三方的知识产权、商业秘密及其他合法权益。如第三方因为参赛者侵权行为提出索赔、诉讼等，参赛者应承担由此产生的全部责任及损失。
3. 数据使用：对于大赛提供的数据（数据集），参赛者须仅在比赛场景下使用，同时不得以任何形式使用比赛之外的任何数据参赛。对于不提供下载的比赛数据，参赛者不得以任何形式擅自复制、下载或获取。参赛者如发现任何出现数据未授权访问的可能，应立即通知组委会并积极提供相关信息。
4. 代码分享：在大赛举办期间，未经组委会同意，参赛者禁止公开分享与赛事相关的数据、模型和代码；大赛结束之后，参赛者可以在拥有模型和代码的知识产权的情况下自行选择公开分享，但需要确保此类公开共享不会侵犯任何第三方的知识产权、商业秘密及其他合法权益。

5. 参赛者若在参赛过程中发现相关规则漏洞或技术漏洞，有义务及时告知资委会相关漏洞的信息，组委会将对提供相关信息的参赛者表示相关感谢；若参赛者利用相关漏洞进行参赛，经判断查证后，成绩将会被判断为无效成绩。

五、申诉与仲裁

1. 参赛团队或选手对不符合大赛规定的设备、工具和软件，有失公正的评判和奖励以及工作人员的违规行为等，均可向大赛组委会提出申诉。组委会负责受理比赛中提出的申诉并进行调解仲裁，以保证大赛的顺利进行和大赛结果的公平公正。组织委员会作出的仲裁结果为终局决定。
2. 申诉报告应明确申诉内容，指定一名成员作为联系人，并要有参赛队伍成员的签名，否则申诉将不予以受理。
3. 组织委员会将在收到申诉之日起 5 个工作日之内受理，并认真核查和处理。

六、其他

本大赛规程的最终解释权归“中国高校计算机大赛——华为云大数据挑战赛”组织委员会所有。

“中国高校计算机大赛——华为云大数据挑战赛”组织委员会

2020 年 4 月

附件：热身赛题——交通流量预测

随着电子信息和移动通信技术高速发展和不断融合，人工智能在各个领域都相继取得了巨大的突破，城市智能体也应运而生，而城市交通又是城市智能体的核心。交通流量数据既是城市交通中的基础数据，又是反应交通状况的重要指标之一，准确预测交通流量对城市交通具有重大意义。本题以交通流量预测为目标，邀请各个队伍以历史交通流量数据建立对应的算法模型，预测目标流量数据，通过预测值和真实值之间的对比得到预测准确率，以此来评估各队伍所提交的预测算法。

一. 赛题说明

本次比赛任务是利用历史数据并结合地图信息，预测五和张衡交叉路口未来一周周一（2019 年 2 月 11 日）和周四（2019 年 2 月 14 日）两天的 5:00-21:00 通过 wuhe_zhangheng 路口 4 个方向的车流量总和。

要求模型输出格式如下：

```
{"data":{"resp_data":{"wuhe_zhangheng":[1,4,5,6,4...]}}}
```

从 5:00 开始每 5min 的预测数据，第一个数据为 5:00-5:05 的流量值，最后一个数据为 20:55-21:00。两天的数据按时间先后放在一起，总共有 384 个数据。

小提示：如果不考虑天气、周边活动、节假日等因素，预测结果可能不准确哦。

二. 数据说明

本次比赛提供 4 周（2019.1.12 – 2019.2.8）深圳龙岗区坂田街道交通流量历史数据。车流数据格式如下:

time	cross	direction	leftFlow	straightFlow
2018/10/12 20:00:00	wuhe_zhangheng	east	1	0
2018/10/12 20:05:00	wuhe_zhangheng	west	2	1

其中，time 为上述格式时间字符串，cross 为路口名，direction 为车流起始方向，leftFlow 是左转车流，straightFlow 是直行车流。

说明：

1) 十字路口包含四个方向车流数据，此处未全部列出。
2) 路口名称分别为：五和路、张衡路、稼先路、隆平路、冲之大道。可以通过但不限于百度地图等地图软件获取地图路网信息。
3) 因为右转车流不受信号灯控制，因此未做统计。

获取竞赛数据集：

从 OBS 拷贝竞赛数据集，首先登录 OBS 管理控制台，在**华北-北京四**创建您的 OBS 桶；然后登录 ModelArts 管理控制台，在**华北-北京四**创建 Notebook，**将如下代码中的 my_bucket/my_folder 替换成您自己的 OBS 桶**；最后运行代码，将竞赛数据集拷贝至您的 OBS 桶中。

```
import moxing as mox
mox.file.copy_parallel('s3://obs-bdc2020-bj4/traffic_flow_dataset',
's3://my_bucket/my_folder')
print('Copy procedure is completed !')
```

说明：详细操作请查看大赛官网交流论坛的相关文档。

三. 评分标准

第一部分（分类问题）

分类问题评价标准：预测的评价还是通过每一个 5min 预测车流和真实通过车流对比，看看趋势是否一致（比如 10 月 19 日的 5：00 到 5:05 的真实车流是 4，10 月 20 日的 5：00 到 5:05 的真实车流为 5，那么只要车流预测值大于 4，就得 100 分，最后得分为所有得分求加权平均（权重为该时间段所在小时的车流量占 16 小时总车流的比重））。

$$grade = \sum_{i=1}^{16}(\frac{1}{12} * \sum_{j=1}^{12}(w_i * (0, 100)))$$

第二部分（回归问题）

回归问题评价标准：预测的评价还是通过每一个 5min 预测车流和真实通过车流通过 grade 公式计算最后得分，加权细则与第一部分相同：

$$grade = \sum_{i=1}^{16}(\frac{1}{12} * \sum_{j=1}^{12}(\omega_i * sigmoid(\frac{30}{(x_j - \bar{x}_j)^2 + \varepsilon}) * 100))$$

其中 wi 为权重，xj 为真实车流数据，xj 拔为预测车流数据，ε为 e^{-9}。

最后将两部分分数做归一化处理，第一部分占比 40%，第二部分占比 60%。

四. 模型规范

1) 所提交的模型必须请满足赛题说明中的模型输出格式，且要符合 ModelArts 模型包规范。
2) 评分系统使用 ModelArts 批量服务加载参赛者所提交的模型，批量服务的输入目录中为一个 batchin.csv 文件，文件内容为预测时间（2019-2-11,2019-2-14）。**建议参赛者在提交模型之前，先通过 ModelArts 的“批量服务”验证模型的可用性和准确性。**

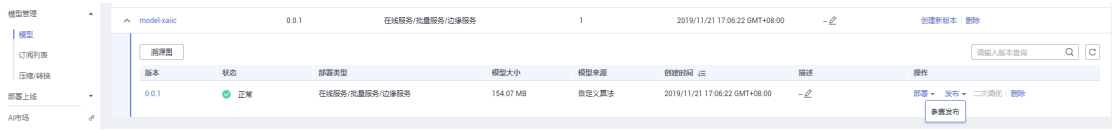
3) ModelArts 模型管理中的模型创建后，不会自动更新，如果您有了更好的模型需要提交判分，要重新导入模型，然后再将重新导入的模型提交判分。

五. 提交说明

所有参赛者需使用华为云一站式 AI 开发平台 ModelArts 来开发模型，且将模型部署为在线服务或批量服务验证其正确性。确认模型输出无误后，在 ModelArts 平台上将开发好的模型提交判分，最后在竞赛平台上查看分数及排名。

提交方法：

1) 在 ModelArts 左侧导航栏中选择“模型管理>模型”，单击模型名称左侧“v”，然后单击页面右侧操作栏中的“发布>参赛发布”。



2) 在弹出的“参赛模型提交”对话框中，选择比赛项目、比赛阶段，然后单击确定。

点击确定后，即成功提交模型判分。在如下界面中可点击“现在加入”，也可以点击“以后再说”或直接点击右上角关掉该对话框。

说明：模型提交判分后，需等待一定时间（判分系统进行判分需一定时间，运行时长与选手提交的模型有关），判分系统完成判分后，可在竞赛平台“提交作品”中查看得分，其中“提交作品”页面需报名比赛后才会显示。

评分说明：

1) 本次比赛榜提交时间段为：4 月 17 日 10:00 - 5 月 22 日 14:00。
2) 每个团队每天有 3 次评测机会，所提交的模型得分可在大赛平台页面“提交作品”中查询。
3) 排行榜每 6 个小时刷新一次。

附件：正式赛题——船运到达时间预测

在企业全球化业务体系中，海运物流作为其最重要的一项支撑。其中，船运公司会和数据供应公司进行合作，对运输用的船通过 GPS 进行定位以监控船的位置；在运输管理的过程中，货物到达目的港的时间是非常重要的一项数据，那么需要通过船运的历史数据构建模型，对目的港到达时间进行预测，预测时间简称为 ETA（estimated time of arrival），目的港到达时间预测为 ARRIVAL_ETA。

本次大赛提供历史运单 GPS 数据、历史运单事件数据、港口坐标数据，预测货物运单的到达时间，对应“历史运单事件”数据中 EVENT_CODE 字段值为 ARRIVAL AT PORT 时 EVENT_CONVOLUTION_DATE 的时间值。

一. 比赛数据

大赛提供脱敏后的训练数据及测试数据，训练数据集包括：历史运单 GPS 数据、历史运单事件数据、港口坐标数据，这些数据主要用于参赛队伍训练模型，制定预估策略；测试运单数据为不同运单、运输过程中的不同位置所构成，供选手测试对应的 ETA 时间。

货物运单在船运过程中，会产生大量的 GPS 运单数据，记录为“历史运单 GPS 数据”；货物运单在船运过程中离开起运港、到达中转港、到达目的港等关键事件，记录为“历史运单事件数据”；“港口的坐标数据“为与运单船运相关的港口坐标信息。

允许选手合理增加与题目相关的外部数据进行纠正，如大赛提供的港口坐标数据存在偏差时可自行补充数据纠正。

1. 历史运单 GPS 数据

历史运单 GPS 数据描述每个运单在船运的过程中，所在船产生的 GPS 位置的相关信息。

列名	类型	说明
loadingOrder	VARCHAR2	脱敏后的主运单，货物的运单编号，类似快递单号
carrierName	VARCHAR2	脱敏后的承运商名称，类似快递公司名称
timestamp	DATE	时间，格式为：yyyy-MM-dd'T'HH:mm:ss.SSSZ，如 2019-09-05T16:33:17.000Z
longitude	NUMBER	货物在运输过程中，当前船舶所处的经度坐标，如 114.234567
latitude	NUMBER	货物在运输过程中，当前船舶所处的纬度坐标，如 21.234567

vesselMMSI	VARCHAR2	脱敏后的船舶海上移动业务识别码 MMSI， 唯一标识，对应到每一艘船
speed	NUMBER	单位 km/h，货物在运输过程中，当前船舶的瞬时速度，部分数据未提供的可自行计算。
direction	NUMBER	当前船舶的行驶方向，正北是 0 度，31480 代表西北方向 314.80 度，900 代表正东偏南 9 度。
vesselNextport	VARCHAR2	船舶将要到达的下一港口，港口名称可能不规范，如 CNQIN、CN QIN、CN QINGDAO 都代表下一站为中国青岛港口。
vesselNextportETA	DATE	船运公司给出的到“下一个港口”预计到达时间，格式为：yyyy-MM-dd'T'HH:mm:ss.SSSZ，如 2019-09-12T16:33:17.000Z
vesselStatus	VARCHAR2	当前船舶航行状态，主要包括： moored under way using engine not under command at anchor under way sailing constrained by her draught
vesselDatasource	VARCHAR2	船舶数据来源（岸基/卫星）：Coastal AIS，Satellite
TRANSPORT_TRACE	VARCHAR2	船的路由，由“-”连接组成，例如 CNSHK-MYPKG-MYTPP。由承运商预先录入，实际小概率存在不按此路由行驶（如遇塞港时），但最终会到达目的港口。

数据说明：

每个运单表示一次运输的运输单号，不会重复使用，一次运输过程中的多条 GPS 数据拥有相同的运输单号。船号为运单货物所在的船编号，会重复出现在不同次运输的 GPS 数据中。需要注意的是 GPS 数据中可能会有异常的 GPS，可能且不限于如下问题：

- GPS 坐标在陆地，或者有些港口是内陆的港口。
- GPS 漂移：两点距离过大，超过船的行驶能力。
- GPS 在部分地区的比较稀疏（比如南半球、敏感海域）。
- 最后的 GPS 点可能和港口的距离较远（比如塞港时，或者临近目的港时已无 GPS 数据）。
- speed 字段之后数据可能会有少量缺失（如 GPS 设备短暂异常）。

2. 历史运单事件数据

历史运单事件数据描述每个运单在船运的过程中，与港口相关的关键信息，如离开起运港、到达目的港等。

列名	类型	说明
loadingOrder	VARCHAR2	运单号，与历史运单 GPS 数据中的 loadingOrder 字段一致
EVENT_CODE	VARCHAR2	事件编码，主要事件包括： TRANSIT PORT ATD 实际离开中转港 SHIPMENT ONBOARD DATE 实际离开起运港 TRANSIT PORT ATA 实际到达中转港 ARRIVAL AT PORT 实际到达目的港 注：部分船可能没有中转港
EVENT_LOCATION_ID	VARCHAR2	港口名称，对应“港口坐标数据”表中的字段 TRANS_NODE_NAME
EVENT_CONVOLUTION_DATE	DATE	事件发生的时间，格式为：yyyy/MM/dd HH:mm:ss（dd 与 HH 之间为两个空格）。 例如 Event_code 为“SHIPMENT ONBOARD DATE"时，此字段表示船从起运港出发的时间。 EVENT_CODE 为“ARRIVAL AT PORT"时，此字段表示船到达目的港的时间。

3. 港口坐标数据

港口坐标数据描述每个运单在船运的过程中涉及的港口位置信息。

列名	类型	说明
TRANS_NODE_NAME	VARCHAR2	港口名称，如：WAREHOUSE_TURKEY，MOSCOW_RUSSIAN FEDERATION，CHIWAN(44)，SHEKOU，深圳蛇口港等
LONGITUDE	VARCHAR2	港口的经度坐标
LATITUDE	VARCHAR2	港口的纬度坐标
COUNTRY	VARCHAR2	国家
STATE	VARCHAR2	省|州
CITY	VARCHAR2	城市
REGION	VARCHAR2	县|区
ADDRESS	VARCHAR2	详细地址。
PORT_CODE	VARCHAR2	港口编码，即港口的字母简码，如 CNSHK 代表中国蛇口港

数据说明：

部分地址信息可能已脱敏、缺失或有偏差，选手可自行补充或修正。

4. **测试运单数据**

测试运单数据为运单运输过程中的不同位置点所构成，供选手测试对应的 ETA 时间。测试运单数据如下表描述。

列名	类型	说明
loadingOrder	VARCHAR2	运单的运单号
timestamp	DATE	运单当前所处位置（经度、维度）的时间，格式为：yyyy-MM-dd'T'HH:mm:ss.SSSZ，如 2019-09-05T16:33:17.000Z
longitude	VARCHAR2	运单承运船舶的当前经度：114.234567
latitude	VARCHAR2	运单承运船舶的当前纬度：21.234567
speed	NUMBER	货物在运输过程中，当前船舶的瞬时速度，部分数据未提供的可自行计算。
direction	NUMBER	当前船舶的行驶方向，正北是 0 度，31480 代表西北方向 314.80 度，900 代表正东偏南 9 度。
carrierName	VARCHAR2	承运商名称，类似快递公司名称
vesselMMSI	VARCHAR2	脱敏后的船舶海上移动业务识别码 MMSI， 唯一标识，对应到每一艘船
onboardDate	DATE	离开起运港时间，格式为：yyyy/MM/dd HH:mm:ss（dd 与 HH 之间为两个空格），如 2019/09/05 16:33:17
TRANSPORT_TRACE	VARCHAR2	船的路由，由“-”连接组成，例如 CNSHK-MYPKG-MYTPP。由承运商预先录入，实际小概率存在不按此路由行驶（如遇塞港时），但最终会到达目的港口。

二. 选手提交结果

所有参与竞赛的选手登录到大赛平台，提交结果数据，具体提交格式要求：

列名	类型	示例
loadingOrder	VARCHAR2	测评提交的 test 运单号

timestamp	DATE	运单当前所处位置（经度、维度）的时间，格式为：yyyy-MM-dd'T'HH:mm:ss.SSSZ，如 2019-09-05T16:33:17.000Z
longitude	NUMBER	运单承运船舶的当前经度：114.234567
latitude	NUMBER	运单承运船舶的当前纬度：21.234567
carrierName	VARCHAR2	承运商名称，类似快递公司名称
vesselMMSI	VARCHAR2	脱敏后的船舶海上移动业务识别码 MMSI， 唯一标识，对应到每一艘船
onboardDate	DATE	离开起运港时间，格式为：yyyy/MM/dd HH:mm:ss（dd 与 HH 之间为两个空格），如 2019/09/05 16:33:17
ETA	DATE	到达目的港口的 ETA，格式为：yyyy/MM/dd HH:mm:ss（dd 与 HH 之间为两个空格），如 2019/09/18 22:28:46
creatDate	DATE	当前表创建时间，格式为：yyyy/MM/dd HH:mm:ss（dd 与 HH 之间为两个空格），如 2020/05/05 16:33:17

其中，ETA 为选手评估的时间值；creatDate 为该表或该 CSV 文件创建时间，用于区别多次提交数据。对于未提交的运单 ETA，后台统一取 timestamp 时间计算。

1. 初赛

大赛初赛提供：

- 训练数据：1.5 万量级运单对应的历史运单 GPS 数据、历史运单事件数据和港口坐标数据，用于模型的训练
- 测试数据：2 万量级测试数据，由不同运单的不同 GPS 位置所组成。

选手通过训练数据训练好的模型，对测试数据进行预测和提交结果，初赛排行榜以选手的提交结果评分为准。

2. 复赛

大赛复赛提供：

- 训练数据：1.8 万量级条运单对应的历史运单 GPS 数据、历史运单事件数据和港口坐标数据，用于模型的训练。
- 测试数据：2.5 万量级条测试运单数据，由不同运单的不同 GPS 位置所组成。

选手通过训练数据训练好的模型，对测试数据进行预测和提交结果，复赛排行榜以选手的提交结果评分为准。

三. 评估标准

选手提交结果的评估指标是 MSE，即 ARRIVAL AT PORT 预测时间 ETA 与真实时间 ATA 的差距的平方和，计算如下：

$$MSE = \frac{\sum_{i=1}^{ETA\_NUM}(hETA_i - hATA_i)^2}{ETA\_NUM}$$

其中：

- hETA 为同一个货物运单到达目的港口的预测所需时间。选手提供 DATE 时间，评测程序转换为单位所需时间，单位：小时。
- hATA 为同一个货物运单到达目的港口的实际所需时间，大赛测评程序后台保存，用于测评运算。
- ETA_NUM 为预测的 ETA 数量，测评程序后台运算，大赛测评程序后台保存，用于测评运算。

最终使用 MSE 值作为参赛选手得分，MSE 值越小，排名越靠前。

示例说明：

如某一货物运单路由 CNSHK-MYPKG-MYTPP，已离开起运港 CNSHK，SHIPMENT ONBOARD DATE 为 2019/09/05 16:33:17，通过经纬度等信息判断船位置在 CNSHK 与 MYPKG 之间，根据,预测目的港口 MYTPP 的时间，提交的 ETA："2019/09/18 22:28:46"。

	SHIPMENT ONBOARD DATE	ETA	ATA
CNSHK	2019/09/05 16:33:17	-	-
MYTPP	-	2019/09/18 22:28:46	2019/09/18 13:28:46

ARRIVAL AT PORT 实际到达目的港 MYTPP 为 2019/09/18 13:28:46；选手预测为 2019/09/18 22:28:46；

目的港 MYTPP 的 hETA1 为(2019/09/18 22:28:46) – (2019/09/05 16:33:17) = 317.9925（单位：小时），hATA1 为(2019/09/18 13:28:4) – (2019/09/05 16:33:17) = 308.9925 （单位：小时）。

只算此条路径 MSE：

$$MSE = \frac{(317.9925 - 308.9925)^2}{1} = 81$$