# 数据挖掘课程期末项目：【数据创新应用】介绍

**本数据创新应用共提供三个数据集，请选择其中一个作为课程期末项目方向，学生可以参考每个方向给出的参考提示，具体题目不限。**

**方向一：出租车数据可视化与热门区域挖掘（考察重点：时空数据；数据聚合；静态/动态可视化展示）**

数据：深圳市一天的出租车 GPS 数据(可从 Blackboard 下载)，出租车数据字段为 Id, lat, lon, time, speed, is_passenger。

做法参考提示：考虑使用地图可视化软件（如 Deck.GL）将轨迹显示出来。也可以尝试将原始数据转化为订单 OD 格式（初始出发点 origin 与到达目的地点 destination）。可自行寻找更多的数据，比如与路网，社区，POI，空气质量，防疫等数据进行结合，并产生更深度的分析。

**方向二：足球数据可视化以及胜率预测（考察重点：数据收集；预测算法；建模）**

数据：statbombs(可从 https://github.com/statsbomb/open-data 下载)

做法参考提示：可以使用 socceraction 库对现有数据进行建模分析，将原始数据转换为 SPADL 格式并计算 VAEP Score。后续也可寻找或手动记录更多的足球赛数据，并使用模型进行胜率预测与 MVP 挖掘等分析。也可使用其他的足球模型进行分析。

**方向三：电子商务数据分析（考察重点：数据处理；聚类算法；推荐算法；图网络）**

数据：淘宝用户行为数据集(可从 https://tianchi.aliyun.com/dataset/dataDetail?dataId=649 下载)

做法参考提示：可以尝试通过用户的行为对用户进行分类以及商品推荐，也可以针对商品挖掘潜在用户，因为我们无法进行 A-B 测试，所以可以针对此数据集用前 6 天预测第 7 天，也可寻找额外的电子商务数据集进行更广泛的分析与分类。