# 商业智能选址预测课题

## 课题概述

本课题旨在探索数据驱动的商业选址决策方法。通过分析品牌历史开店的地理网格分布数据，建立预测模型，以预测品牌下一家门店的最佳位置。这一研究不仅具有重要的学术价值，也有广泛的商业应用前景。同学们将有机会运用机器学习、深度学习和空间分析等知识，解决一个真实世界中的复杂问题。

## 问题背景

在现代商业环境中，选址决策是品牌扩张过程中最关键的环节之一。一个成功的门店位置可以最大化客流量、品牌曝光和销售潜力，而选址失误则可能导致巨大的经济损失。

传统的选址决策往往依赖于专家经验和简单的人口统计学分析，这些方法存在主观性强、难以量化评估等缺点。随着大数据和机器学习、深度学习技术的发展，基于历史数据的选址预测模型正成为一个极具潜力且值得深入探索的研究方向。

本课题提供了多个品牌在某个城市中的门店分布数据，以网格(Grid)为基本地理单位记录了门店的空间分布。我们的核心假设是：品牌的历史选址行为反映了其内在的扩张策略和偏好，通过学习这些模式，可以预测其未来的选址决策。

## 3. 问题定义

给定一个品牌的历史门店网格序列，预测该品牌下一家门店最可能选择的网格位置。

形式化表示：

* **输入**：品牌A的历史开店网格序列 G = [g₁, g₂, ..., gₙ]，其中gᵢ表示该品牌第i家店所在的网格ID，以及每个网格的地理属性特征（经纬度范围等）
* **输出**：品牌A的第n+1家店最可能选择的k个网格位置及其概率/得分
* **目标**：最大化预测准确率，即预测结果与实际选址的匹配程度

## 4. 数据集介绍

本课题提供了三个数据文件：

### 4.1 训练集（train\_data.csv）

包含多个品牌的历史门店分布数据，主要字段包括：

* brand\_name：品牌名称
* brand\_type：品牌类型
* longitude\_list：该品牌所有门店的经度列表
* latitude\_list：该品牌所有门店的纬度列表
* grid\_id\_list：该品牌所有门店所在的网格ID列表

### 4.2 测试集（test\_data.csv）

结构同训练集，用于评估模型性能。

### 4.3 网格坐标信息（grid\_coordinates.csv）

提供了每个网格的地理范围信息，主要字段包括：

* grid\_id：网格唯一标识符
* grid\_lon\_min：网格最小经度
* grid\_lat\_min：网格最小纬度
* grid\_lon\_max：网格最大经度
* grid\_lat\_max：网格最大纬度

**说明**：

1. longitude\_list、latitude\_list和grid\_id\_list中的元素顺序对应，即同一索引位置的经度、纬度和网格ID描述的是同一个门店位置。注意：提供的数据中可能存在同一个网格中存在该品牌的多家门店的情况，可以根据需要进行去重。
2. 数据已经按7:3的比例随机划分为训练集和测试集。
3. 网格是等大小的地理单元，覆盖了整个研究区域。研究区域坐标范围为：

* lat\_min, lat\_max = 36.6035, 36.7640
* lon\_min, lon\_max = 116.8315, 117.3673

## 5. 任务与挑战

本课题的核心任务是开发一个能够准确预测品牌下一家门店位置的模型。这一任务面临以下挑战：

1. **空间依赖性**：门店位置不是独立的，存在空间自相关性，需要合理建模空间关系。
2. **数据稀疏性**：某些品牌的门店数量可能较少，带来数据稀疏性挑战。

## 6. 评估标准

模型性能将通过以下指标进行评估：

1. **命中率（Hit Rate@k）**：预测的前k个位置中包含实际位置的比例
2. **平均排名（Mean Rank）**：实际位置在预测排名中的平均位置

## 7. 提交要求

同学们需要提交以下内容：

1. **技术报告**：详细描述问题分析、方法设计、实验结果和讨论
2. **代码和数据**：完整的代码实现和必要的中间数据
3. **预测结果**：在测试集上的预测结果
4. **演示文稿**：用于展示项目的PPT或其他形式的演示材料（比如网页demo)

## 8. 可能的研究方向

同学们可以（但不限于）探索以下研究方向：

1. **基于序列的方法**：将门店开设视为时间序列，利用RNN、LSTM或Transformer等模型捕捉时间依赖关系
2. **基于图的方法**：构建空间关系图，应用GNN等图学习方法建模空间关系
3. **多模态融合**：结合地理信息、品牌特性和其他辅助信息的多模态模型
4. **迁移学习**：利用品牌间的相似性，通过迁移学习解决数据稀疏问题