腾讯社交广告
Tencent Social Ads

高校算法大赛

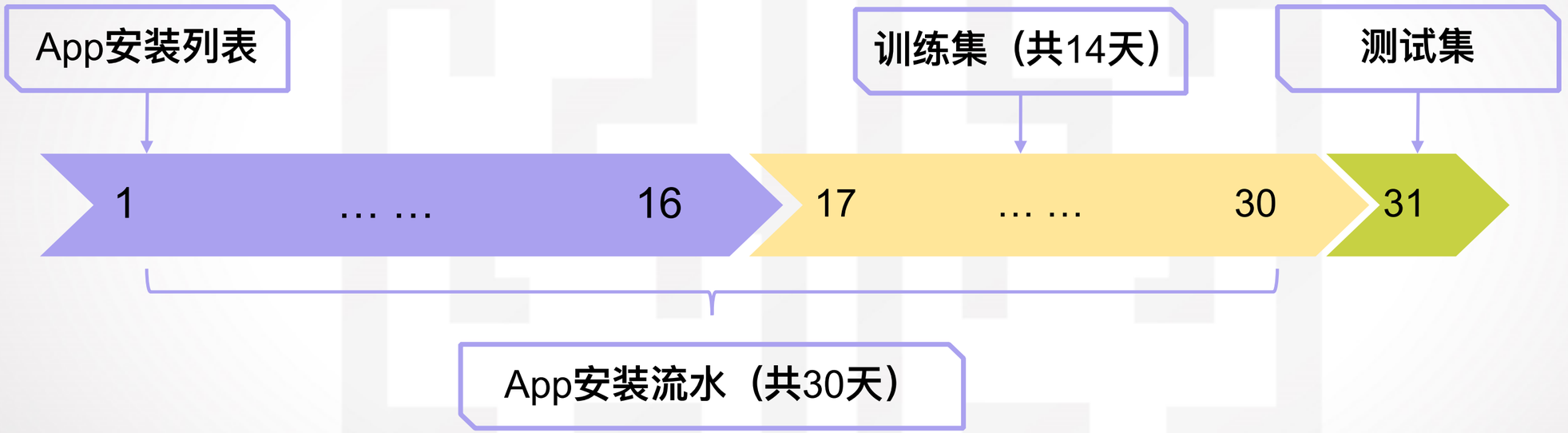
目录

Contents

1 问题描述

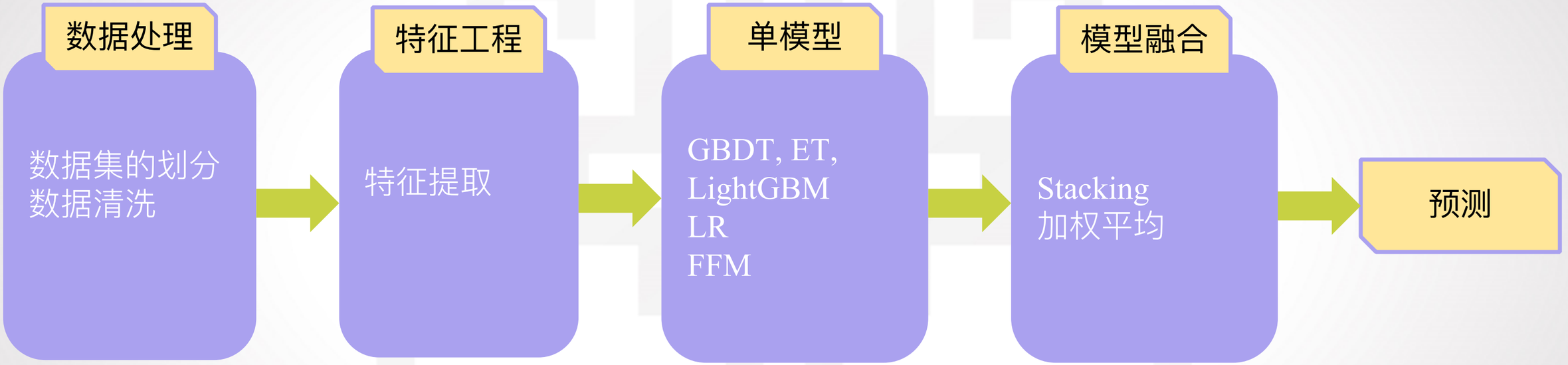
数据来源：腾讯在社交广告领域的真实数据

数据内容：广告日志数据



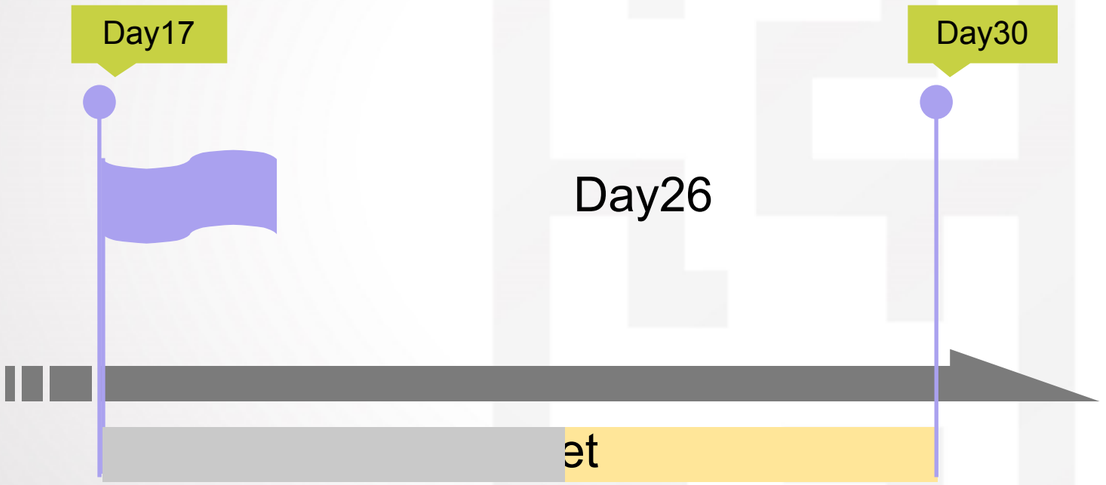
根据广告日志，预测给定广告、用户和上下文情况下广告被点击后发生激活的概率
评价指标：logloss

2 框架设计


数据处理
数据集的划分
数据清洗
特征工程
特征提取
单模型
GBDT, ET,
LightGBM
LR
FFM
模型融合
Stacking
加权平均
预测

3 数据处理

训练集选取



5折交叉验证

机器条件的限制，选取26号-30号作为训练集

3 数据处理

数据清洗

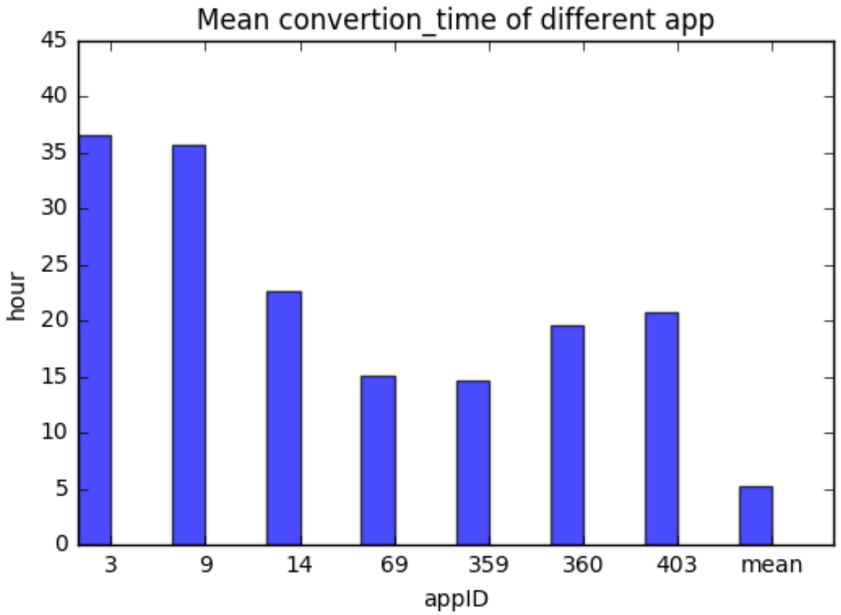
回流时间过长 → 最后五天信息不准

解决

全部删除 → 很多有用信息被丢掉，效果不好

思考

和app相关



解决办法

删除掉30号中平均转化回流时间较长的数据

4 特征工程

基础特征

User(U)

age, gender, education …

Ad(A)

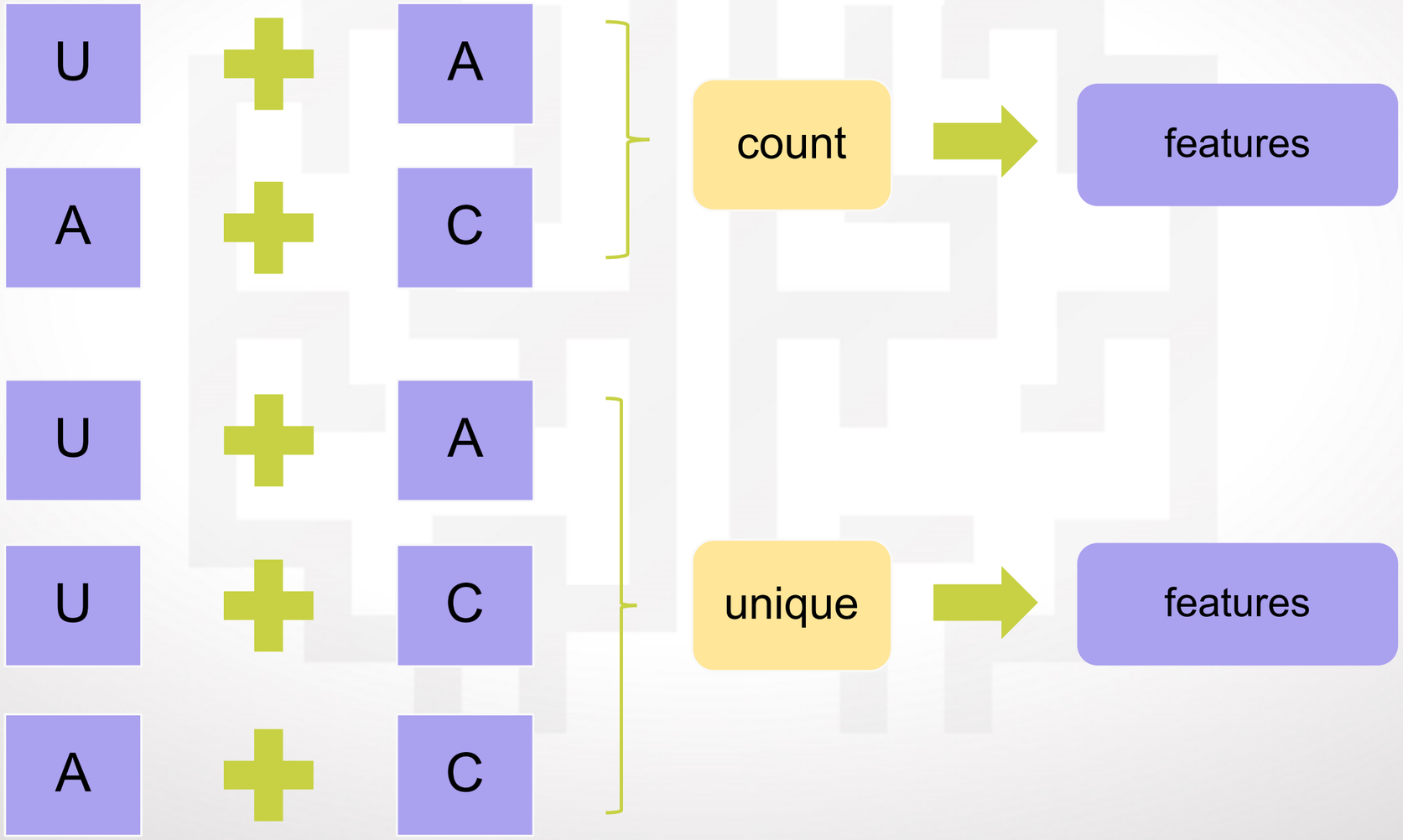
creativeID,advertiserID, appID …

Context(C)

positionID,sitesetID, positionType …

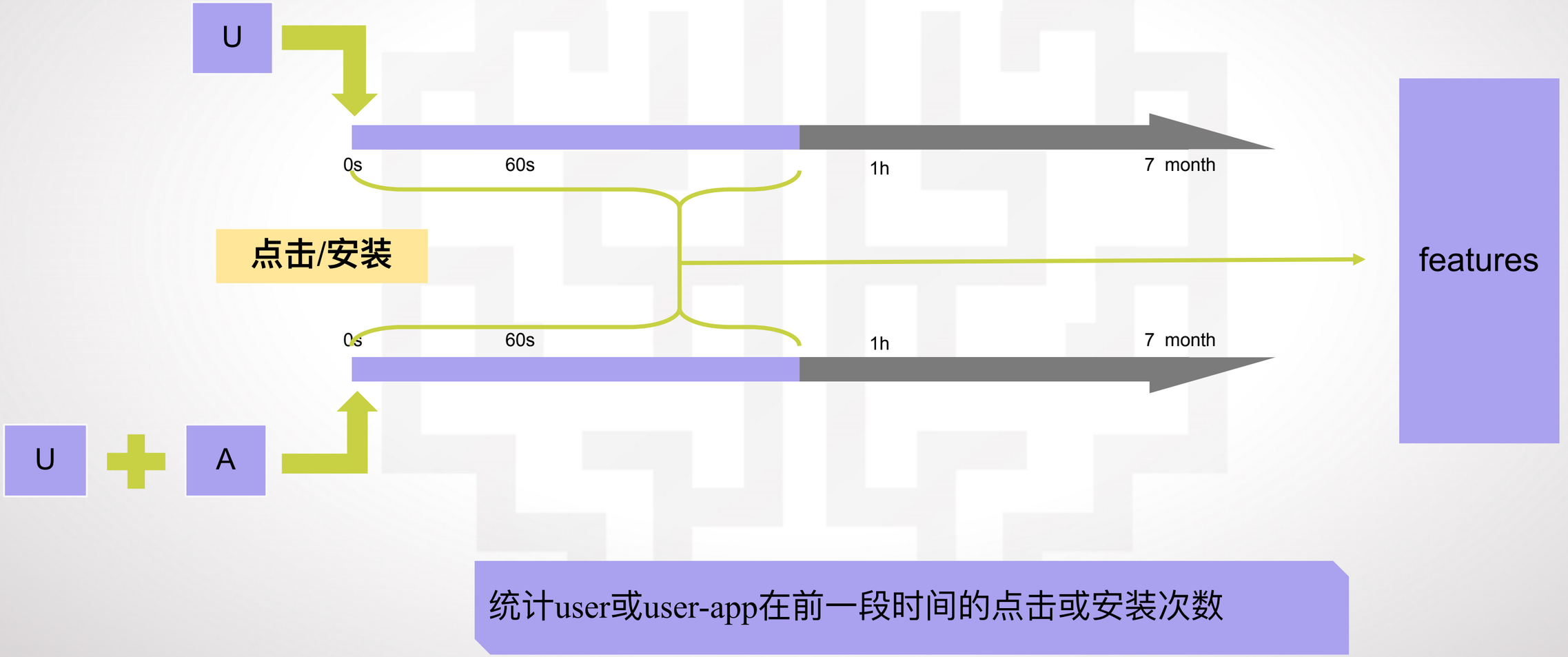
4 特征工程

统计特征



4 特征工程

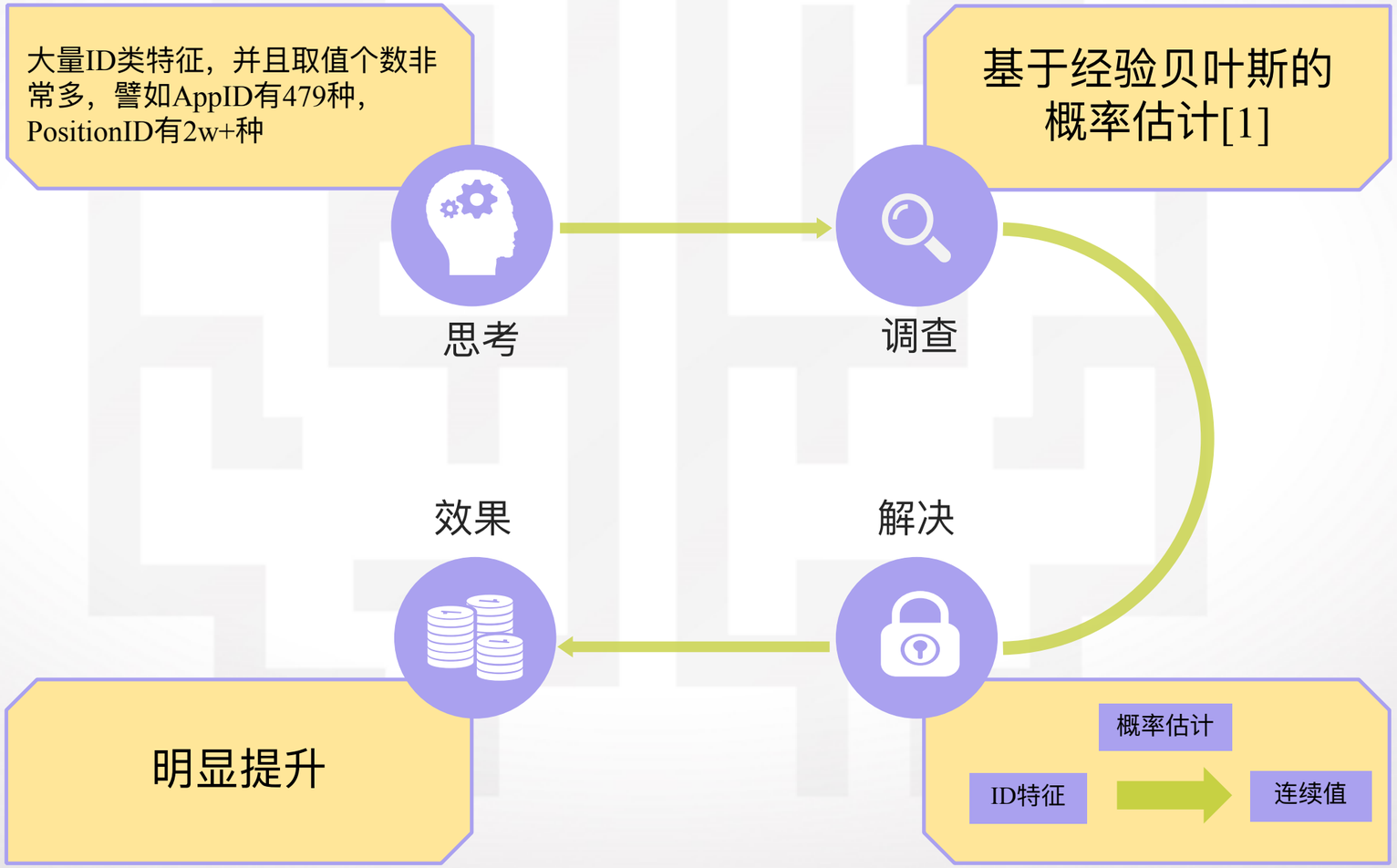
时间相关特征



统计user或user-app在前一段时间的点击或安装次数

4 特征工程

概率估计特征



[1]A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. SIGKDD

算法实现及模型融合

前期(单模型)

TOP3

LightGBM

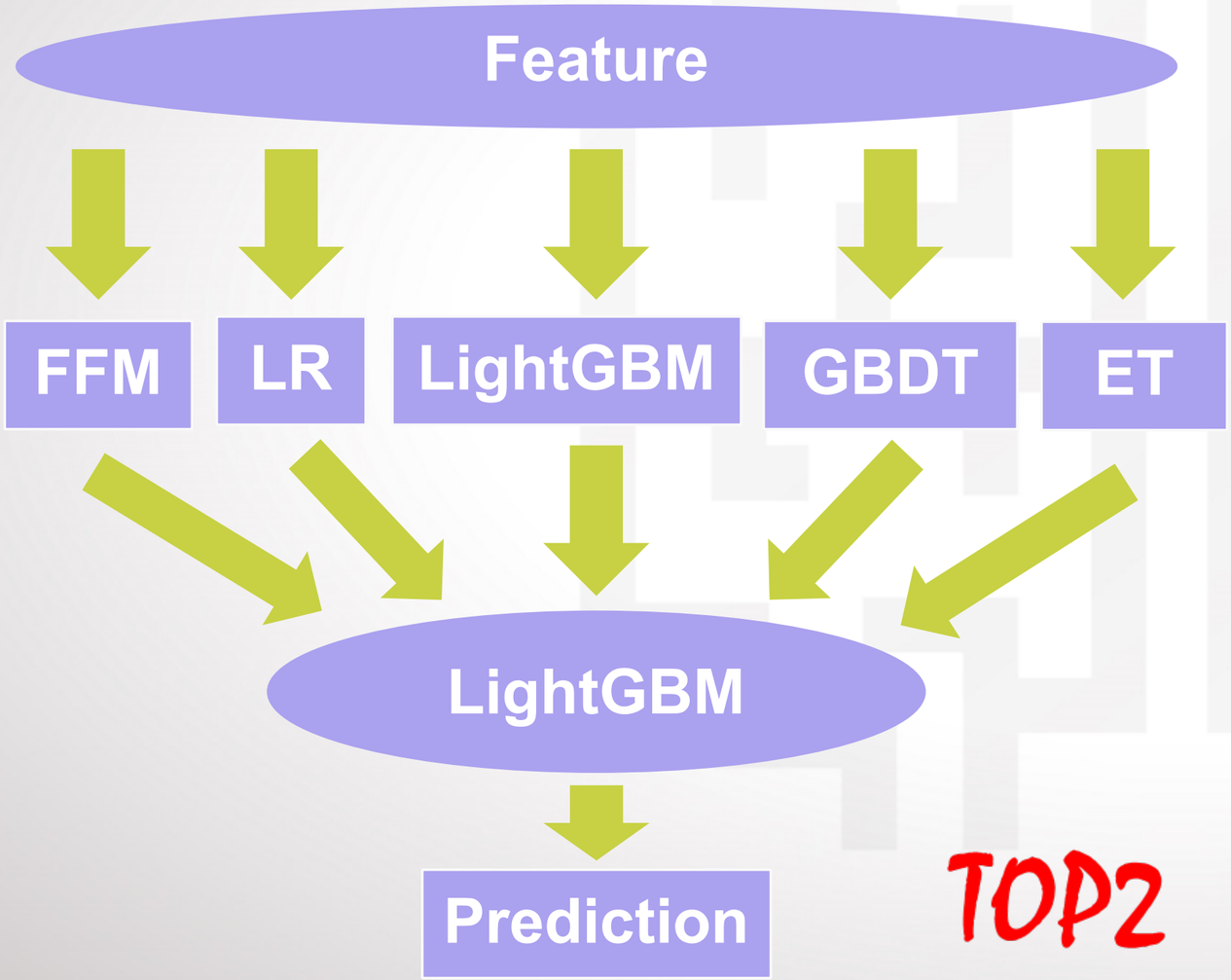
Vs

XGBoost

1. 二者准确率相当；
2. LightGBM更快(优选)

5 算法实现及模型融合

★ Stacking



TOP2

★加权融合

□直接对概率预测进行加权平均
□ 0.5 * XGB + 0.5 * LightGBM
□效果并不理想

6 总结

比赛中遇到的问题和解决方法

1

最后几天label可能不准确

过滤30号转化回流时间较长的app

2

数据量大，代码运行慢

换个思维方式，缩短至几小时

3

特征分析时遇到瓶颈

研究赛题的实际意义，发现position信息的重要性，并由此构造一些列特征

6 总结

优点

1. 对数据进行了充分的分析
2. 充分挖掘了positionID信息
3. 模型多样性、特征多样性
4. 记录笔记，代码多而不乱

遗憾

1. GBDT生成特征送到LR或FFM
2. 对用户历史安装挖掘不充分

6 总结

致谢

- 感谢所有参赛队伍
- 感谢各个周冠军、进步最快队伍的耐心总结
- 感谢大家的耐心聆听
- 感谢腾讯为在校生举办了这样一次成功的比赛，给了我们锻炼和展示自己的机会

腾讯社交广告
Tencent Social Ads

高校算法大赛

THANKS