
# 2018中国高校计算机大赛——大数据挑战赛

## ——可爱多

# 目录

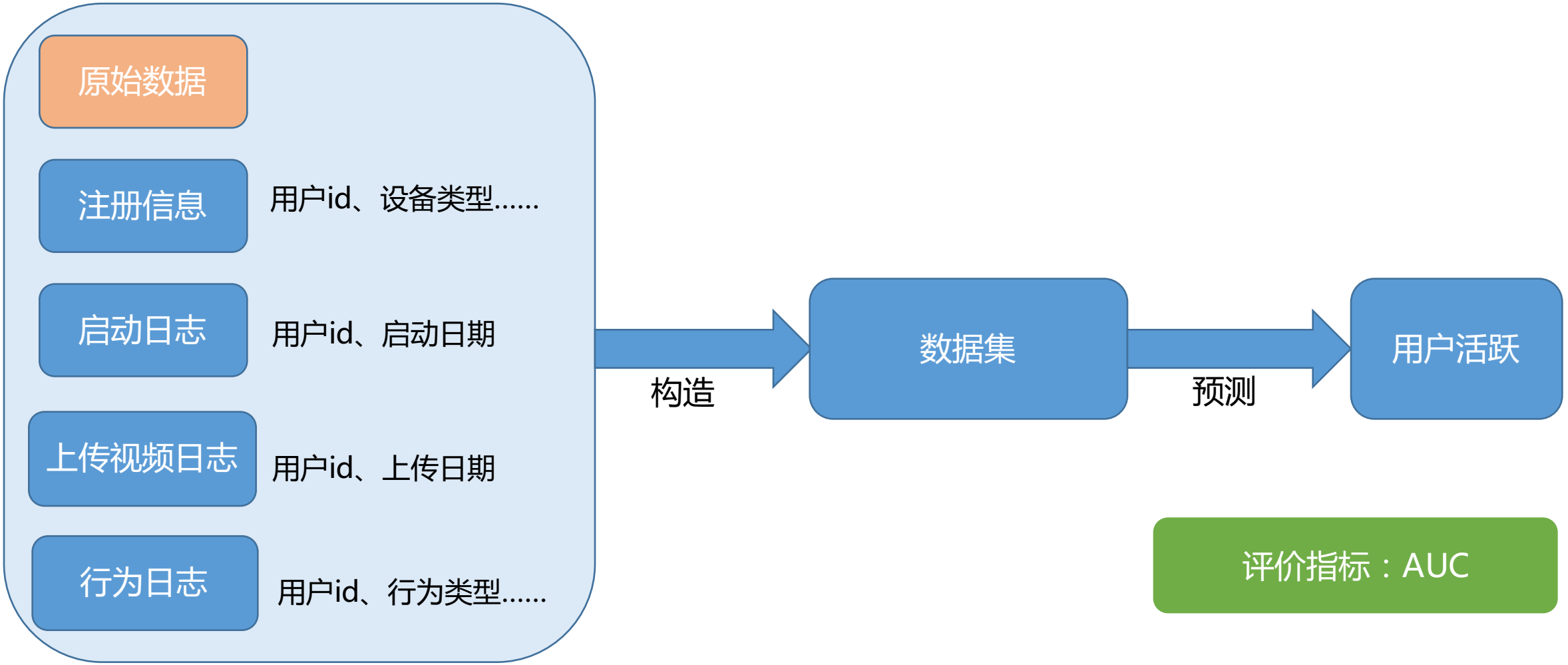
CONTENTS

# Part 1

# ·团队介绍·

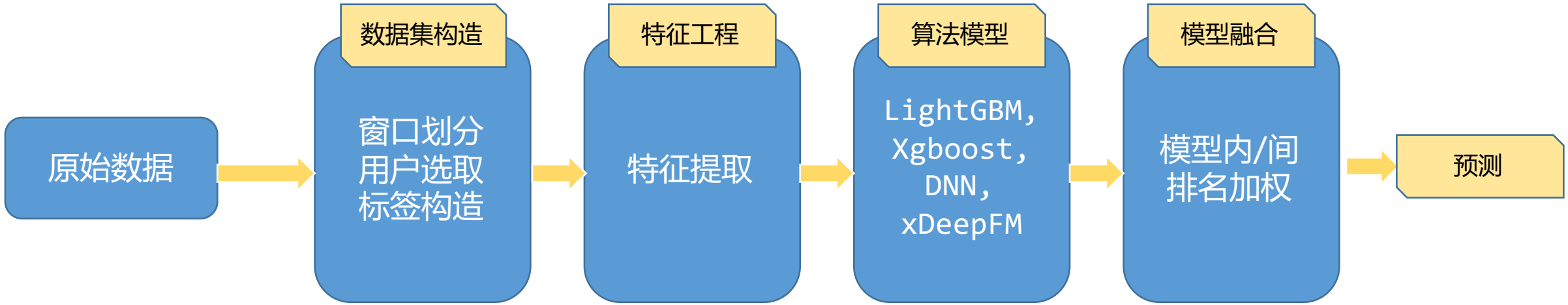

Part 2
问题简介

# 问题简介


原始数据
注册信息
用户id、设备类型……
启动日志
用户id、启动日期
上传视频日志
用户id、上传日期
行为日志
用户id、行为类型……
构造
数据集
预测
用户活跃
评价指标：AUC


Part 3
整体框架

# 整体框架


原始数据
数据集构造
窗口划分
用户选取
标签构造
特征工程
特征提取
算法模型
LightGBM,
Xgboost,
DNN,
xDeepFM
模型融合
模型内/间
排名加权
预测

清華大學 Tsinghua University | 快手

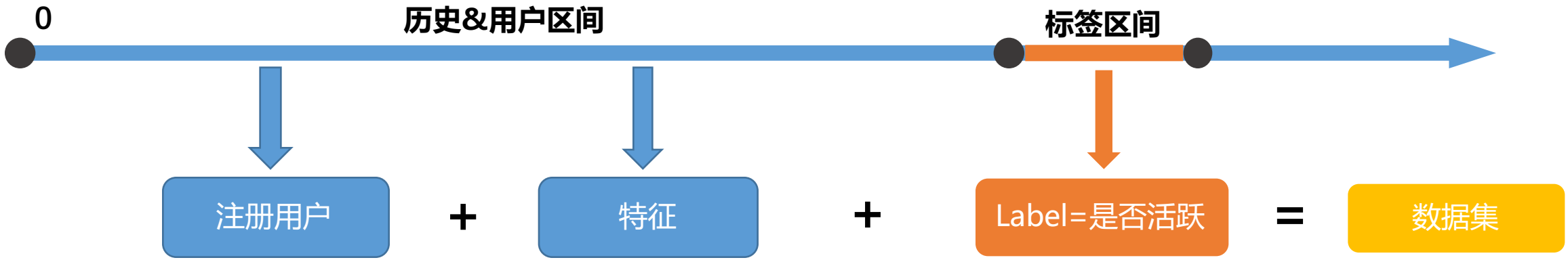
# Part 4

# 数据集构造

# 数据集构造

我们仿照线上目标，在避免窗口重叠的情况下选择以下的构造方法：

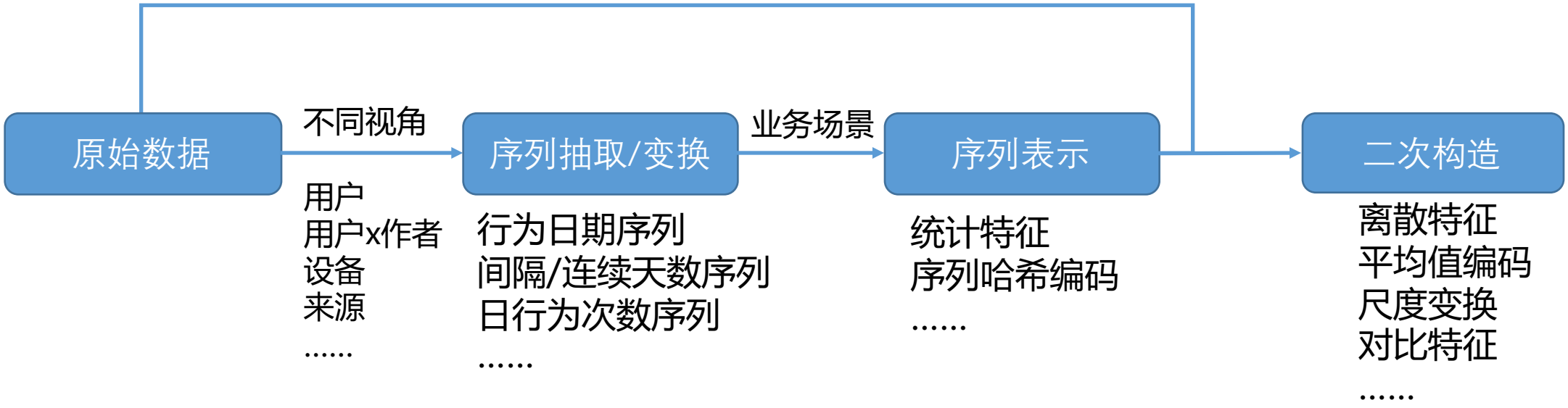
| 数据集 | 历史&用户区间 | 标签区间 | 说明 |
| --- | --- | --- | --- |
| 0 | [1,10) | [10,16] | 训练集 |
| 1 | [1,17) | [17,23] | 训练集 |
| 2 | [1,24) | [24,30] | 线下验证集 |
| 3 | [1,31) | [31,37] | 线上测试集 |



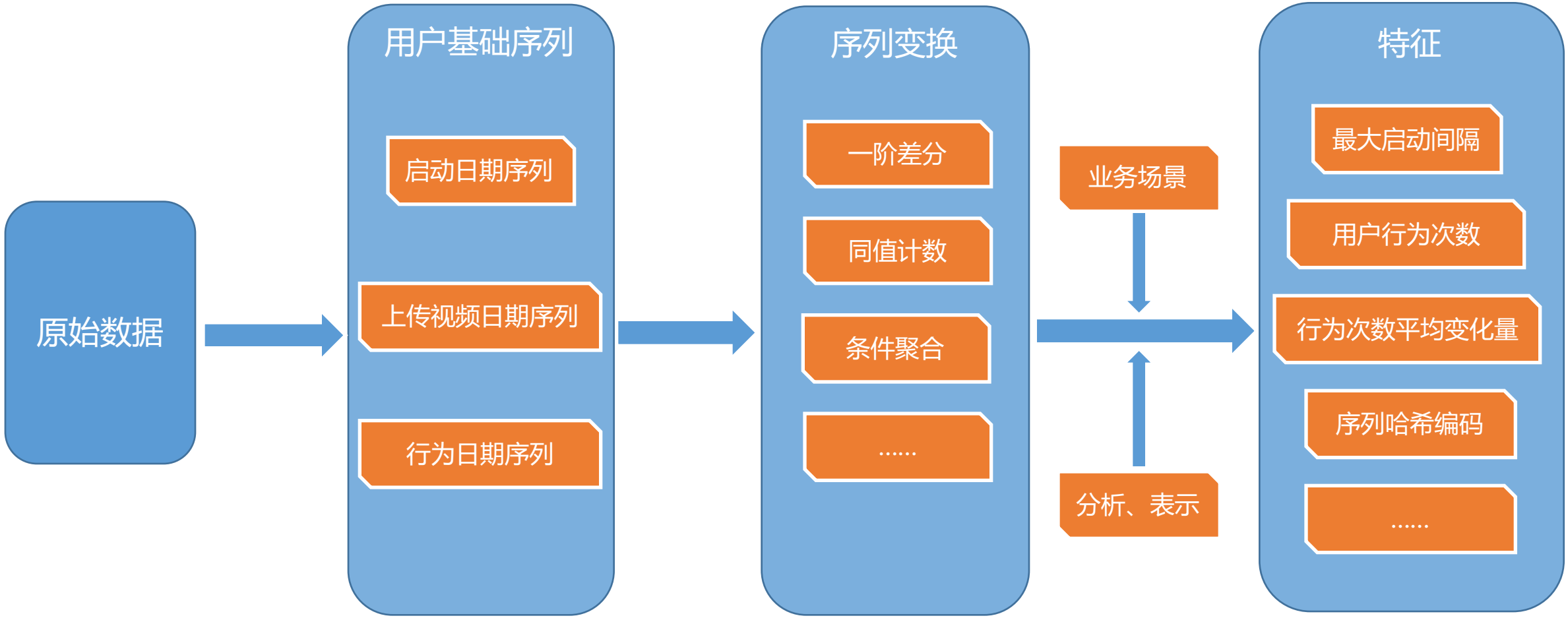

清華大學
Tsinghua University
快手
Part 5
特征工程

# 特征思路

- 用户的各类活动可以看作多个序列，结合业务场景进行序列的抽取、变换和表示，基于此思路可以全面地挖掘序列中的信息。



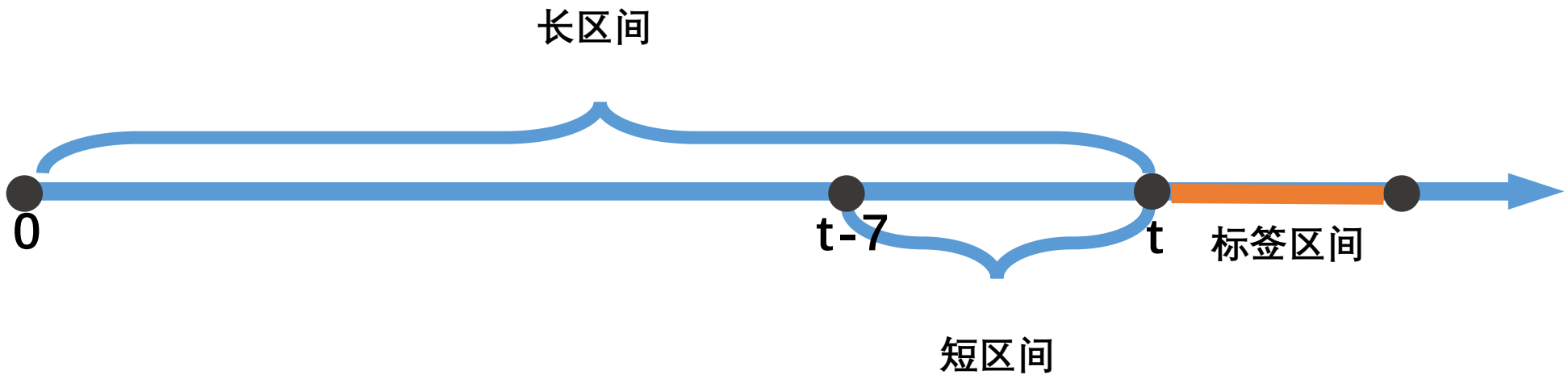
# 序列抽取/变换


原始数据
用户基础序列
启动日期序列
上传视频日期序列
行为日期序列
序列变换
一阶差分
同值计数
条件聚合
……
业务场景
分析、表示
特征
最大启动间隔
用户行为次数
行为次数平均变化量
序列哈希编码
……

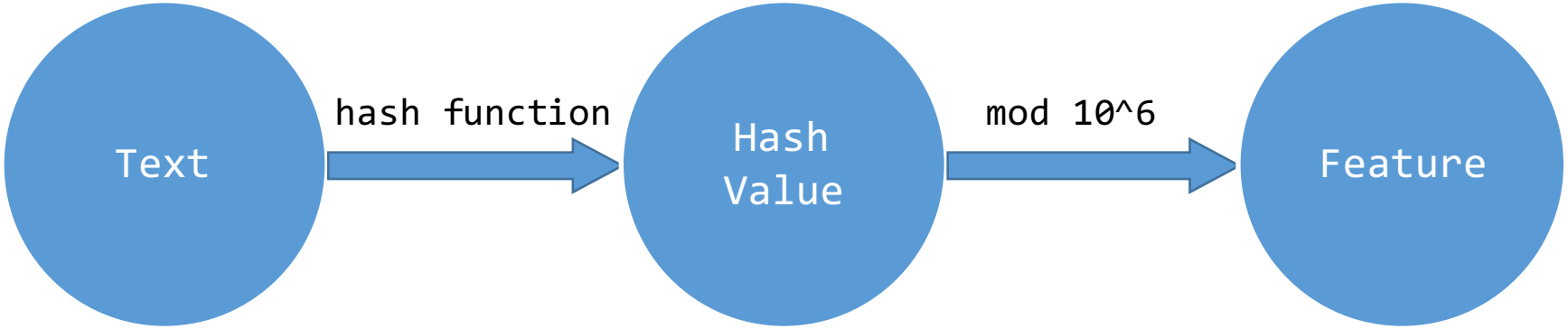
# 统计特征

- 统计特征包括：最大最小值、均值、方差、中位数、极差.....

- 双统计区间：
  - 长区间：所有历史数据，变长，长期，保证用户信息完整性
  - 短区间：最近7天的数据，定长，近期，突出用户近期行为



# 序列哈希编码




0/1序列 ：1001010110
频次序列：2001040320

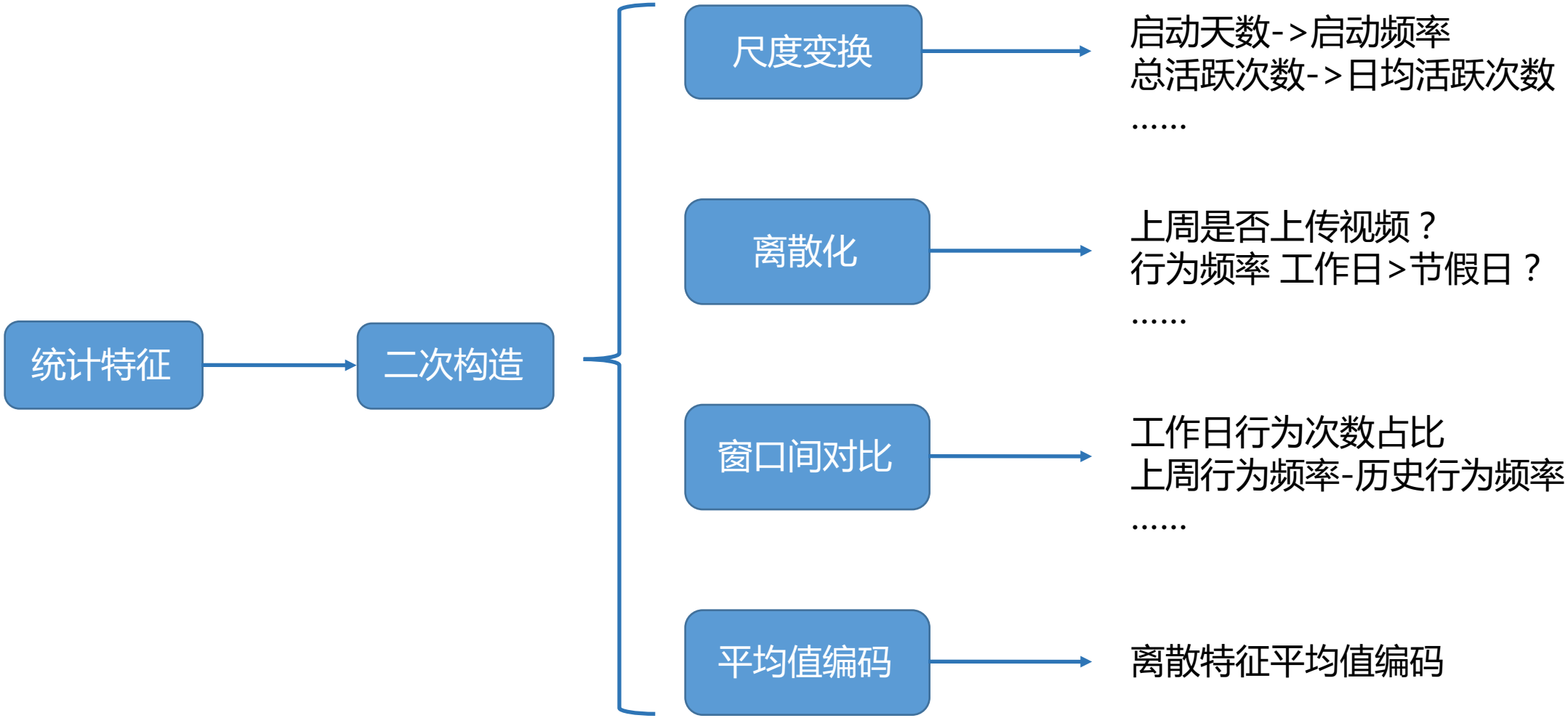
760236474605507633215 3
705046781904503284860 9

332153
848609

使用Hashing Trick离散化用户行为序列，是CTR中的常用技巧，有以下优点：

① 特征构造速度快、耗时少；
② 实现降维，内存友好；
③ 数据一致性强，适合在线学习；

# 二次构造

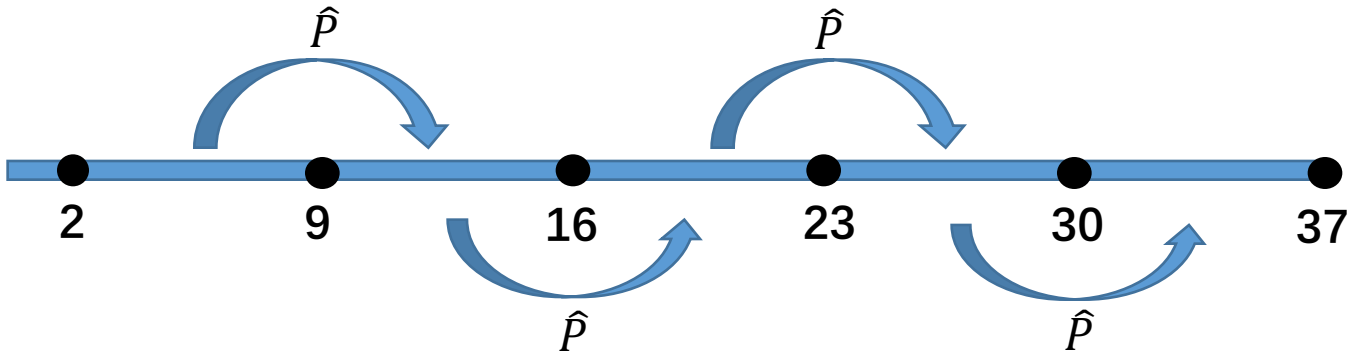

统计特征
二次构造
尺度变换
启动天数->启动频率
总活跃次数->日均活跃次数
……
离散化
上周是否上传视频？
行为频率 工作日>节假日？
……
窗口间对比
工作日行为次数占比
上周行为频率-历史行为频率
……
平均值编码
离散特征平均值编码

# 平均值编码

对离散特征进行平均值编码[1]，当前标签区间的概率估算由上一个定长区间确定，计算公式如下:

$$\hat{P} = \lambda * prior + (1-\lambda) * posterior$$
$$= \lambda * \hat{P}(y=1) + (1-\lambda) * \hat{P}(y=1|var=k)$$

其中，$\lambda$为权重函数，一般表示为$\lambda(n) = \frac{1}{(1+e^{(n-k)/f})}$， n为一个特征类别的出现次数

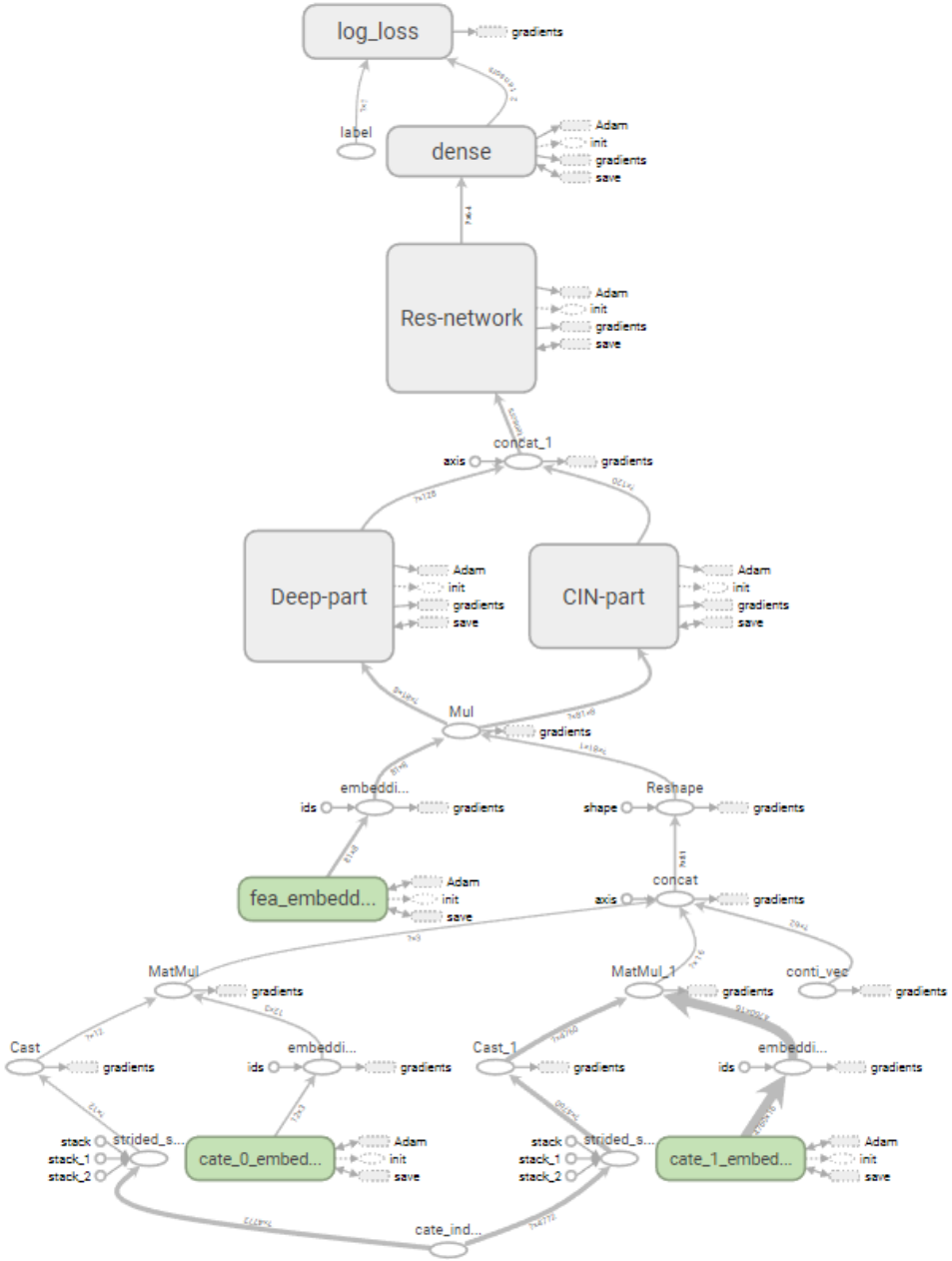



[1] Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems[J]. ACM SIGKDD


Part 6
算法模型

# xDeepFM

- 两层embedding：更好的提取隐藏特征
- he_normal+BatchNormalization+relu：优化梯度下降
- CIN内部使用卷积和部分连接层：解决过拟合问题
- res-network：更快收敛、更好的拟合目标函数



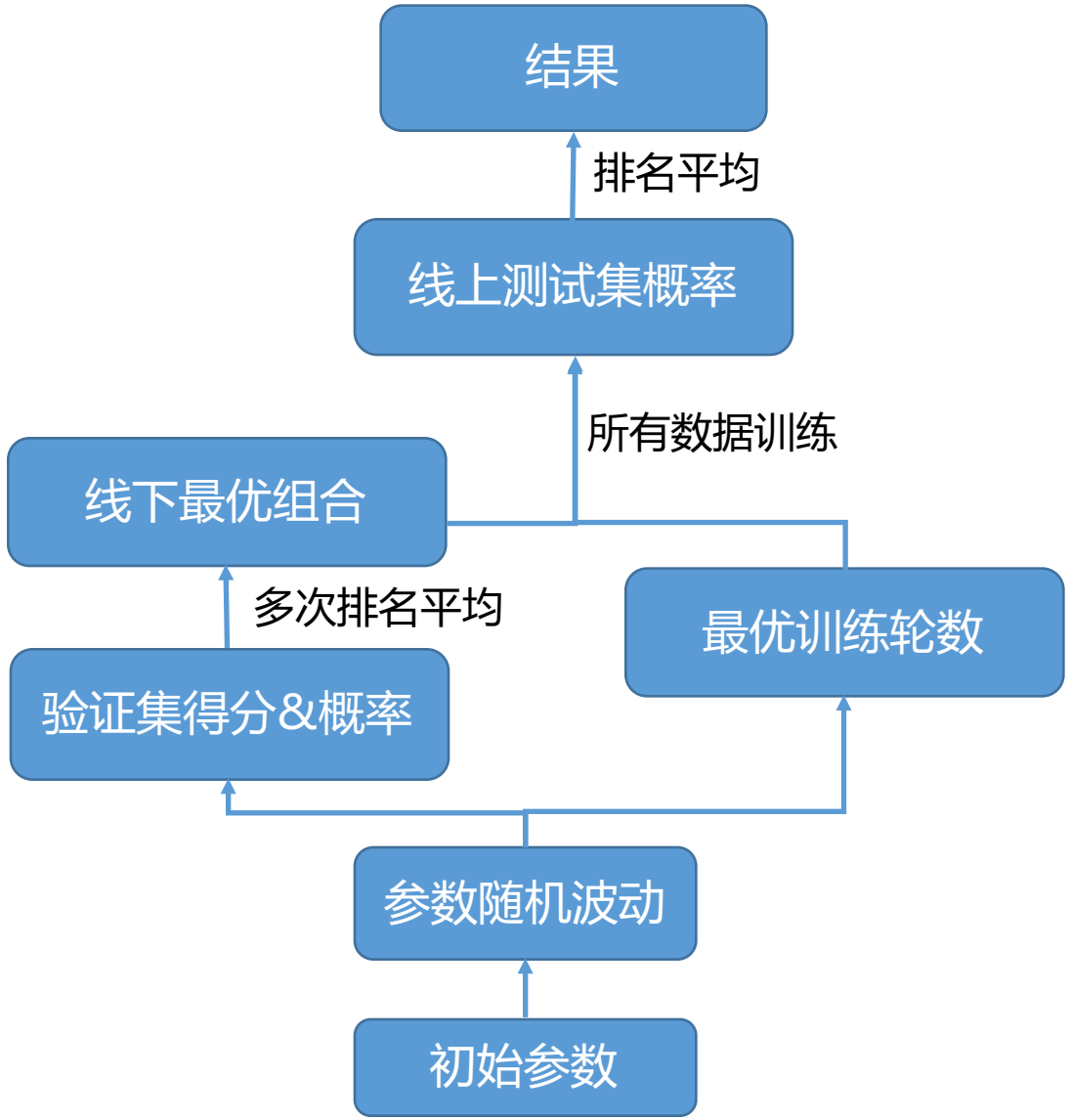

[1]Lian J, Zhou X, Zhang F, et al. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems[J]. 2018.

# 模型内融合

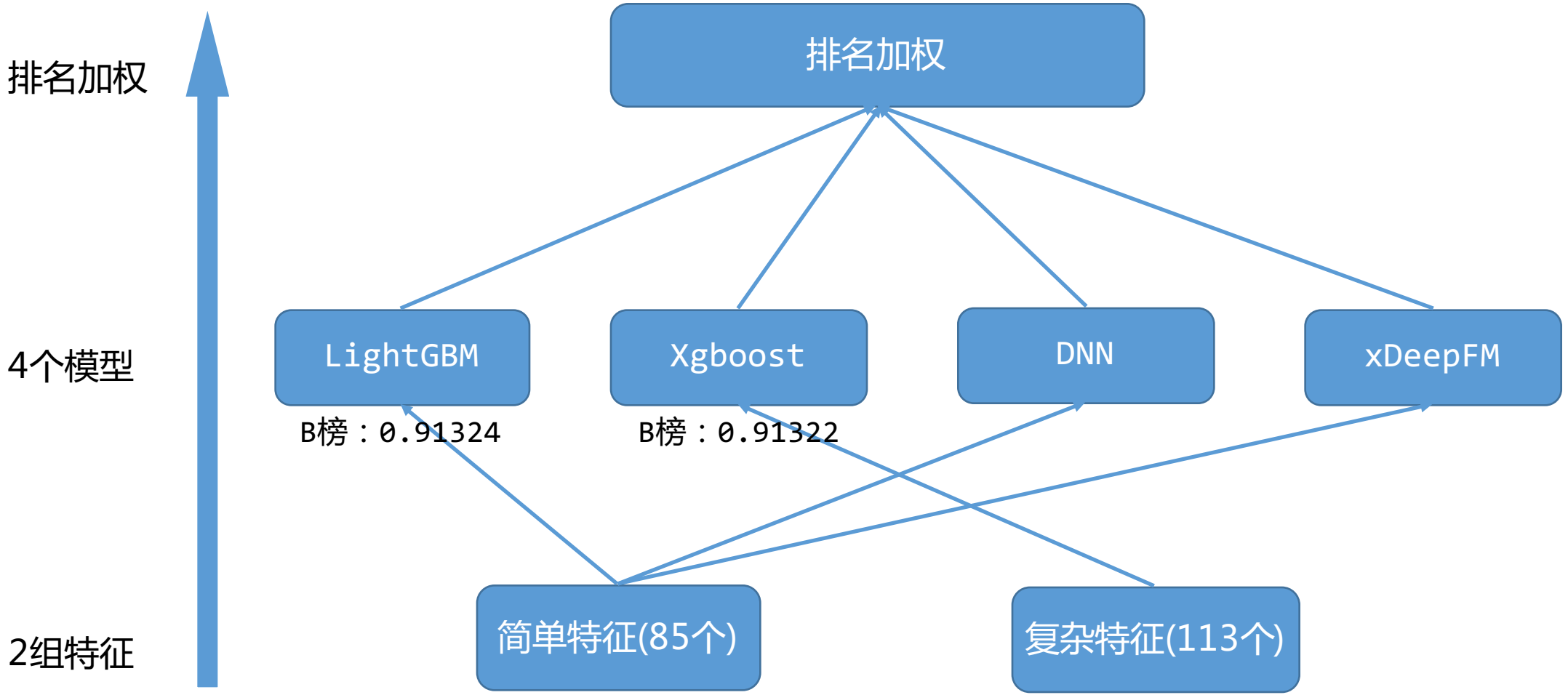
- 在模型内，我们使用右侧策略提高模型稳定性

- 基于验证集调参，得到初始参数
- 在此基础上进行小范围随机波动
- 根据验证集确定训练轮数，得到验证集得分和概率
- 根据排名平均得分，确定最优组合
- 使用所有数据训练，对测试集概率进行排名加权

LightGBM B榜 0.91324 Rank3



# 模型间融合


排名加权
4个模型
2组特征
排名加权
LightGBM
Xgboost
DNN
xDeepFM
B榜：0.91324
B榜：0.91322
简单特征(85个)
复杂特征(113个)

# 其他问题

1. 线上/线下不一致
    - 降低模型复杂度
    - 分析不同数据窗口中特征的均值/方差/取值范围等，将差异过大的去除。
    - 检查特征是否泄漏

2. 特征复现慢/内存不足
    - 代码重构
    - 多kernel并行
    - 存储中间结果

3. 复赛特征文件存储
    - 将特征文件制成压缩包，可以节省大量持久化空间

# 致谢

- 感谢所有的参赛队伍
- 感谢每周周星星的总结和分享
- 感谢举办方、科赛平台对于此次比赛的支持


清華大學
Tsinghua University
快手
谢谢聆听